Inferences from observations to simple statistical hypotheses

Inferences from observations to simple statistical hypotheses

Jason Grossman MA MPH

submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy, University of Sydney, 2005

I am impressed also, apart from prefabricated examples of black and white balls in an urn, with how baffling the problem has always been of arriving at any explicit theory of the empirical confirmation of a synthetic statement.

(Quine 1980, pp. 41–42)

Typeset in Belle 12/19, using Plain TEX by Donald Knuth . . .

英語の文字は日本語の文字ほど優美ではないので、
英語で植字するのに恥るのは馬鹿らしいのですが。

(Thanks to Koji Tanaka for illustrating this point.)

CONTENTS

PART II: FOR AND AGAINST THE LIKELIHOOD PRINCIPLE

ACKNOWLEDGEMENTS

Max Parmar, who first got me interested in this topic and thus ruined my relatively easy and lucrative careers in computing and public health. Thanks Max.

Geoffrey Berry, Peter Lipton, Neil Thomason, Paul Griffiths, Huw Price and Rachel Ankeny, each of whom has given extraordinary support to my career. Alison Moore, Jackie Grossman, Justin Grossman, Tarquin Grossman, Nancy Moore and Alan Moore, for massive long-term support.

The three examiners — Mark Colyvan, Alan Hájek and Nicholas J.J. Smith — who read this whole thesis amazingly carefully and corrected many mathematical mistakes and other infelicities. I would especially like to thank Alan Hájek for dividing his 154 corrections into no less than six categories, from "bigger things" to "nano things".

David Braddon-Mitchell, Stephen Gaukroger and Huw Price, for patient thesis supervision. Fiona Mackenzie, Sue Dodds and Rachel Ankeny, for letting me drag my heels on our joint research projects while I wrote this thesis. The Center for Philosophy of Science, University of Pittsburgh, for a Visiting Fellowship during which most of this thesis was written.

Jim Bogen, David Braddon-Mitchell, Mike Campbell, Mark Colyvan, Stephen Gaukroger, Ian Gordon, Dave Grayson, Alan Hájek, Allen Hazen, Matthew Honnibal, Claire Hooker, Kevin Korb, Claire Leslie, Alison Moore, Erik Nyberg, Max Parmar, Huw Price, John Price, Denis Robinson, Daniel Steel, Ken Schaffner, Teddy Seidenfeld, David Spiegelhalter and Neil Thomason, for discussions which helped me directly with the

ideas in this thesis, and many others for discussions which helped me with indirectly related topics.

David Hume and James O. Berger, neither of whom has given me any help on a personal level; but many of the ideas in this thesis are (it seems to me) more or less implicit in their work. I've tried to write the thesis that those two would have written if they had been one PhD student.

STATEMENT OF ORIGINALITY

All of the work presented in this thesis is my own original research except as clearly stated otherwise. None of it has been submitted for another degree, except that the case study in chapter 15 is based partly on my Master's thesis (Grossman 1993).

— 1 —
Prologue

Statistical inference is in a mess. As a result of the considerations briefly sketched in this prologue and discussed at length in the rest of the thesis, not just some but the vast majority of inferences made by applied statisticians are seriously questionable. Jokey topics with which deductive logicians while away an idle hour, like what science would be like if most of our inferences were wrong, are not funny to philosophers of statistics. Science probably *is* like that for us. In the cases in which people's decisions depend crucially on statistical inferences — which is primarily in the biomedical sciences — it seems very likely that most of our decisions are wrong, a state of affairs which leads to major new dietary recommendations annually, new "cures for cancer" once a month and so on.

Statisticians would be fixing this situation if only they could agree on its cause. What is hindering them is nothing merely technical. It is the absence of rational ways to agree on what counts as a good inference procedure. We need to do something about this, much more urgently than we need further work on the details of any particular inference method.

Consequently, this thesis investigates statistical inference primarily by investigating how we should *evaluate* statistical inference procedures. I will use considerations about the evaluation of statistical inference procedures to show that there is an important constraint which statistical inference procedures should be bound by, namely the likelihood principle. This principle contradicts ways of understanding statistics which philosophers

of science have been taking for granted, as I will show in the final section of this prologue. Later in the thesis, I will use the likelihood principle to suggest that almost everything that applied statisticians currently do is misguided.

1. EVALUATING INFERENCE PROCEDURES

Statistical inference is the move from beliefs and/or statements about observations to beliefs and/or statements about what cognitive states and/or actions we ought to adopt in regard to hypotheses.[1] Since this thesis focuses on statistical inference, it does not discuss everything that statisticians do (not even everything they do at work). Firstly, the most important thing it ignores is what statisticians do before they have observations to work with. Most of that activity comes under the title *experimental design.* It is important to bear in mind throughout this thesis that the methods which I criticise for being inadequate to the task of inference may be very useful for experimental design. Secondly, although the problem of inference *from* data *to* hypotheses is the main problem of inference these days, for historical reasons it is sometimes called the problem of *inverse* inference, as if it were a secondary problem. The opposite problem, which is to infer probabilities of data sets from mathematically precise hypotheses, is called *direct* inference. Eighteenth- and nineteenth-century mathematics made direct inference relatively easy, and it has always been relatively straightforward philosophically, so I will be taking it for granted. Again, the methods I criticise for being inadequate for inverse inference may be adequate for

1. "And/or" is meant to indicate lack of consensus. As we will see, some say that statistical inference is only about actions, others that it is only about beliefs, and so on.

direct inference. The take-home message of this paragraph is that I will only be discussing inference *from* data *to* hypotheses, and when a method fails to be good for that I will be calling it a bad method, even if it is good for something else.

There is one unsolved problem about statistical inference which is both more important and more urgent than any other. The problem is how to *evaluate* statistical inference procedures.[2] Experts in this area cannot agree, even roughly, on what makes one statistical inference procedure better than another, as the survey of theories of statistical inference which makes up the bulk of Part I of the thesis will show.

It is instructive to compare statistical inference to deductive inference. Everyone agrees that a sine qua non of deductive inference procedures is that they should lead from true premises to true conclusions. There are many ambiguities in that statement, leading to active disagreements about modal logics, relevant logics, higher-order logics, paraconsistent logics, intuitionistic logics and so on, but — and this is a big but — deductive logic is being successfully developed and applied even in the absence of agreement on these questions. This is possible because the basic idea of deductive inference as truth-preserving means more or less the same thing to everybody.

In contrast, there is no equivalent agreed sine qua non for statistical inference. Statistical inference procedures cannot be evaluated by whether they lead from truths to truths, because it is in the very nature of statistical inference that they do not . . . at least, unlike deductive inference, they do not lead from truths about the first-order subject matter of scientific

2. Exactly what I mean by "inference procedures" is explained in chapter 2. Almost any algorithm which makes probabilistic inferences from data to hypotheses will qualify.

investigation (objects and events) to other truths about that subject matter. They *may* lead from truths about the subject matter to truths about what we ought to believe about relative frequencies or some such; but what we ought to believe is not something that can ever be verified in the direct way that first-order claims can (sometimes) be verified.

There is a similar contrast between the problems of simple induction, such as Goodman's (1983) paradox, and the problems of statistical inference.[3] Simple induction asks questions like, "1, 1, 1, 1, 1: what next?" A plausible answer is "1", and this answer can be tested by subsequent experience. The statistical problem of induction, in contrast, asks questions like, "1.1, 0.9, 1.0, 1.1, 1.1: what next?" There is no first-order answer to this; by which I mean that there is no answer such as "1.1". The answer has to be something more like "*Probably* something *in the region* of 1.1." This answer can be explicated in various ways but clearly, however it is cashed out, it is not something that can be tested directly by subsequent experience. (As Romeyn 2005, p. 10, puts the point, "statistical hypotheses cannot be tested with finite means".) Any possible test is dependent on a theory of statistical inference. Consequently, the ability of a statistical inference procedure to pass such tests cannot (by itself) justify the theory behind the procedure, on pain of circularity.

In the next two sections, I will consider two different things which we might want to do when we evaluate a statistical inference procedure: we might want to count the number of times (in different situations) it is right, on the assumption that some hypothesis or other is true; or we might want to compare what it says about various hypotheses in the same

3. See also (Teller 1969) for a plausible but arguably incomplete attempt to solve Goodman's paradox using Bayesian statistical inference.

situation. Then I will use an example to discuss the conflict between these two modes of evaluation.

ONE OPTION: FREQUENTISM

One option is to evaluate statistical inference procedures by seeing *how often* an inference procedure leads from truths to truths. This method for evaluating statistical inference procedures is prima facie closest to the truth-preservation test which we use to evaluate deductive inference procedures.

This might mean that we should work out the number of times we should expect a given inference procedure to get the right answer, in some hypothetical set of test cases. If we do this in the same way we would for a deductive inference procedure, we will start with some known true premises and see how often the inference procedure infers true (and relevant) conclusions from them. Now, before we can embark on such an evaluation, we have to decide what types of conclusions we want the statistical inference procedure to infer. Perhaps, if it is going to be a useful procedure, we want it to infer some general scientific hypotheses. We might then evaluate it by asking how often it correctly infers the truth of those hypotheses, given as premises some other general hypotheses and some randomly varying observational data. We can imagine feeding into the inference procedure random subsets of all the possible pieces of observational data, and we can calculate the proportion of those subsets on which it gets the right answer.[4]

4. Such a method of evaluation requires the inference procedure to produce a determinately true or false answer, which might or might not be a desideratum for the procedure independently of the need to evaluate the procedure.

This method is referred to as the "frequentist" or "error-rate" method. Unfortunately, both terms are misnomers. I will explain why in chapter 4; see also chapter 2 for an alternative meaning of the word "Frequentist" and for the reason why I give it a captial letter.

I hope it seems plausible that the Frequentist method might be the best way to evaluate statistical procedures, as almost all applied statisticians currently take it to be, because Frequentism will be the foil for most of my arguments. In particular, one of the main goals of this thesis, and an essential preliminary to arguing for the likelihood principle, is to show that despite its popularity the Frequentist method is *not* a sensible way to evaluate statistical procedures.

ANOTHER OPTION: FACTUALISM

It might even seem as though the Frequentist method were the *only* way of finding something analogous to the logician's method for testing deductive inferences. In order to see whether it is, consider what information is available to us when we are getting ready to use a statistical inference procedure. Some of our premises at that time will be general statements about the way the world is, of the nature of scientific hypotheses. The rest of our premises will be statements about specific observed phenomena. The distinction between these two — fuzzy though it inevitably is — is fundamental to stating the nature of statistical inference. The most common epistemological goal of science is to make inferences from the latter to the former, from observations to hypotheses. (Not that this is the only possible goal of science.) And in order for this to be *statistical* inference, none of the hypotheses must be deductively entailed by the premises. In

other words, when we need a statistical inference procedure it is because we have collected some data and we want to infer something from the data about some hypotheses.

What we want to know in such a situation is how often our candidate statistical inference procedure will allow us to infer truths, and we want to calculate this by comparing its performance to the performance of other possible procedures in the same situation, with the *same* data as part of our premises. The idea that *this* is what we want to know when we are evaluating statistical procedures has no name. I will call it the **factual** theory, because it ignores counterfactual statements about observations we haven't made. (More on such statements later.) I will also refer to **factualism**, meaning the doctrine that we should always apply the factual theory when doing statistical inference.[5]

The factual method is the one recommended by **Bayesians**, and it is the only one compatible with the **likelihood principle** (defined at the end of this chapter and again, more carefully, in chapter 8). Indeed, when made precise in the most natural way it turns out to be logically equivalent to the likelihood principle, as I will show.

If the Frequentist method agreed with the factualist method then we would have a large constituency of people who agreed on how to evaluate statistical inference procedures. Perhaps they would be right, and if so we could pack up and go home. But no: the Frequentist method is deeply incompatible with the factualist method. The Frequentist method is to

5. Factualism is a normative methodological doctrine. It is *not* a metaphysical doctrine; it must not be confused with (for example) actualism. To see clearly the difference between factualism and actualism, note that unless the factualist calculates the result of every possible alternative procedure, he may not be trading in *observations* he might make but has not, but he is still trading in *calculations* he might make but hasn't: hence, factualism does not rule out the use of counterfactuals. What factualism rules out is any dependence of statistical conclusions on counterfactuals whose antecedents are false *observation* statements.

evaluate the performance of an inference procedure *only* on (functions of subsets of) all the possible pieces of observational data, while the factualist method is to evaluate its performance *only* on the data actually observed. Total conflict.

STATISTICAL INFERENCE IS IN TROUBLE

What we have just discovered is that the very concept of "the performance of an inference procedure" is a completely different animal according to two competing theories of how to evaluate inference procedures. We are not used to this situation — it *can* arise in non-probabilistic inference, when competing ways of measuring success are on offer, but it rarely does — and so we do not always notice it; but we are hostage to it all the time in statistical inference.

The comparison I have been making with methods of deductive reasoning might seem to suggest a nice solution to the problem of how to evaluate statistical methods. In deductive reasoning, as I've mentioned, one wants to go from true statements to true statements; and, helpfully, the meaning of "true", although contentious, is to some extent a separate issue from the evaluation of logical procedures; and hence logicians of differing persuasions can often agree that a particular inference does or doesn't preserve truth. In statistical methods, one wants to go not from true statements to true (first-order) statements but from probable statements to probable (first-order) statements. Statisticians of differing schools often fail to agree whether a particular inference preserves probability. But if they were at least to agree that it *should*, then that in itself would seem to rule out many methods of statistical inference. In particular, it would

seem to rule out methods which restrict attention to a single experiment in isolation, because we know that doing that can lead from probable premises to improbable conclusions. (This is because the conclusions drawn from an experiment in isolation can be rendered improbable by matters extraneous to that experiment — by, for example, a second, larger experiment.)

Sadly, this line of argument does not work. The problem with it is that all methods of statistical inference *sometimes* lead from the probable to the improbable. We might amend the principle we're considering, to say that a good method of reasoning is *likely* to generate probable statements from probable statements. But then the principle becomes ambiguous between (at least!) the Frequentist and factualist interpretations described above, which interpret "likely" differently: we are back in the impasse we have been trying to escape.

If I can clarify this problem and give a clear justification for a solution, even though my solution is only partial and only partially original, I will have achieved something.

2. A SIMPLE EXAMPLE

Although the questions I am asking are entirely scientific questions, at the level of abstraction at which I will be dealing with them very few of the details of applied science will matter. *Some* of the details of applied science will matter in various places, especially in the final chapter, but most of the minutiae of applied statistics will be irrelevant. It is therefore possible to conduct most of the discussion I wish to conduct in terms of a simple example table of numbers, which I construct as follows.

Suppose we have precise, mutually exclusive probabilistic hypotheses which tell us the probabilities of various possible observations. Suppose further that we observe one of the possible observations that our hypotheses give probabilities for. No doubt this sounds like an ideal situation. Let's make it even more ideal by making there be only finite numbers of hypotheses and possible observations. Then we can draw a table:

	actual observation	possible observation 1	possible observation 2	...
hypothesis I	$p_{1,a}$	$p_{1,1}$	$p_{1,2}$	...
hypothesis II	$p_{2,a}$	$p_{2,1}$	$p_{2,2}$	...
⋮	⋮	⋮	⋮	⋱

Table 0

Now let's get concrete. A vomiting child is brought to a Rwandan refugee camp. The various possible diagnoses give rise to various major symptoms with known frequencies, as represented in Table 1 below which says, for example, that only 1% of children with PTSD (Post-Traumatic Stress Disorder) have diarrhoea. It ought to be easy to tell from Table 1 whether the child is likely to be suffering primarily from one or the other of the two dominant conditions among children in the camp: PTSD (in which case they need psychotherapy and possibly relocation) or late-stage dehydration (in which case they need to be kept where they are and urgently given oral

rehydration therapy). The possibility of the child suffering from both PTSD and dehydration is ignored in order to simplify the exposition. The possibility of the child suffering from neither PTSD nor dehydration is considered but given a low probability.

	possible symptoms			
hypotheses	vomiting (observed in this case)	diarrhoea (not observed in this case)	social withdrawal (not observed in this case)	other symptoms & combinations (not observed in this case)
dehydration	0.03	0.2	0.5	0.27
PTSD	0.001	0.01	0.95	0.029
anything else	0.001	0.001	0.001	0.997

Table 1

The table is to be read as follows. Each hypothesis named at the left hypothesises or stipulates some probabilities.[6] The hypothesis that the child has dehydration stipulates that the probability that a dehydrated Rwandan child's main symptom will be vomiting is 3%, the probability that its main symptom will be diarrhoea is 20%, and so on.

6. We might wonder how such sets of hypotheses are selected for consideration. That question, of course, precedes the main question of this thesis, which is how to evaluate a procedure which chooses *between* the given hypotheses. I do not agree with Popper that the provenance of a hypothesis is irrelevant to philosophy, and yet this thesis does not aim to discuss the issue of hypothesis selection in any detail. It will not matter for my purposes where these hypotheses come from as long as they include all the hypotheses which some set of scientists are interested in at some time.

In case there is any doubt about the meaning of the table, it can be expanded as follows:

p(data = vomiting | hypothesis = dehydration) = 0.03
p(data = diarrhoea | hypothesis = dehydration) = 0.2
p(data = withdrawal | hypothesis = dehydration) = 0.5
p(data = other symptoms | hypothesis = dehydration) = 0.27
p(data = vomiting | hypothesis = PTSD) = 0.001
p(data = diarrhoea | hypothesis = PTSD) = 0.01
p(data = withdrawal | hypothesis = PTSD) = 0.95
p(data = other symptoms | hypothesis = PTSD) = 0.029

Note that there is a catch-all column, to ensure that all possible symptoms are represented somewhere in the table.

The types of analysis that have been proposed for this sort of table, and for infinite extensions of it, do not agree *even roughly* on how we should analyse the table or on what conclusion we should draw. In particular, Frequentists and factualists analyse it differently.

Let's look briefly at a standard analysis of this table, as would be performed by practically any applied statistician from 1950 to the present. A statistician would run a statistical significance test in SPSS or one of the other standard statistical computer packages, and that would show that we should clearly reject the hypothesis that the child is dehydrated (p = 0.03, power = 97%). The reasoning behind this conclusion is Frequentist reasoning. It goes like this. If the statistician ran that same test on a large number of children in the refugee camp it would mislead us in certain specific ways only 3% of the time. This has seemed to almost all designers of statistical computer programs, who are the real power-brokers in this situation, to be an admirable error rate. I will show later that the exact

ways in which running the test on a large number of children would mislead us 3% of the time are complicated and not as epistemically relevant as one might hope: so it is misleading (although true) to say that the analysis in SPSS has a 3% error rate.

I will champion the factualist analysis of Table 1, which is opposed to the Frequentist reasoning of the previous paragraph. The factualist says that the rate at which the applied statistician's inference procedure would make mistakes if he used it to evaluate a large number of dehydrated children is *totally irrelevant*, and so are a number of other tools of the orthodox statistician's trade, including confidence intervals and assessment of bias (in the technical sense). The reasoning is simple. We should not care about the error rate of the statistician's procedure when applied to many children who are in a known state (dehydrated), because all we need to know is what our observations tell us about this child, who is in an *unknown* state, and that means we should not take into account what would have happened if — counterfactually — we had applied this or that inference method to other children.

One might reasonably suspect that this factualist reasoning is flawed, because one might suspect that even if the error rate is not something we want to know for its own sake it is nevertheless epistemically relevant to the individual child in question. One of the main jobs of this thesis will be to show that the factualist is right — the error rate is not epistemically relevant to the individual child — given what else we know (and with some exceptions).

The counterfactual nature of the error-rate analysis is the primary source of the disagreement between Frequentists and factualists. This

is what makes resolving the disagreement a task for a philosopher. Not only are Frequentist methods irreducibly dependent on the evaluation of counterfactuals, but moreover they will often reject a hypothesis which is clearly favoured by the data not just *despite* but actually *because* the hypothesis accurately predicted that events which did not occur would not occur: in other words, they will reject a hypothesis on the grounds that it got its counterfactuals *right*. (See chapter 4 for more details.) Perhaps even more surprisingly, I will show that this defect in orthodox methodology cannot be fixed piecemeal. The only way to get rid of it is to show that counterfactuals of this sort are irrelevant to statistical inference, and then to give them the boot. Or rather, to be more precise and less polemical, the only way to fix the problem is to delineate a clear, precise class of cases of statistical inference in which such counterfactuals are irrelevant; and that is what I will do. This task will take up most of Part III of this thesis.

The alternative to using these counterfactuals is to restrict our attention to the single column of the table which represents the observation we actually made, as the factualist advises us to do:

	actual symptoms vomiting
hypotheses	
dehydration	0.03
PTSD	0.001
others	0.001

Table 2: the only part of Table 1 that a factualist cares about

It would be nice if the two sides in this disagreement were just different ways of drawing compatible (non-contradictory) conclusions about the child. I will show in detail that they are not that. To show this just for Tables 1 and 2 for the moment, a look at the probabilities given by the hypotheses shows that the observed symptoms are much more likely on the hypothesis of dehydration than they are on all the other hypotheses. So according to the factualist way of proceeding we should think that the child probably *is* dehydrated, despite the result of the significance test which suggested otherwise (unless we have other evidence to the contrary, not represented in the table). We will see in chapter 3 and chapter 5 that this reasoning is too simple, because there are various competing factualist positions, but all of them would be likely to draw the same conclusion from Tables 1 and 2.

So we have a disagreement between what practically any applied statistician would say about the table and an alternative conclusion we might draw from the table if we restrict ourselves to considering only the probabilities that the various hypotheses assign to the actual observation. I will show that this disagreement generalises to more or less any table of hypotheses and observations; it even generalises to most tables (as it were) with infinitely many rows and columns. Thus, the simple table above illustrates a deep-seated disagreement about probabilistic inference: the disagreement between Frequentism and factualism. The table shows that sometimes (and, as it happens, almost always) these two views are fundamentally incompatible.

3. WHAT THIS THESIS WILL SHOW

The main purpose of this thesis is to consider principles of statistical inference which resolve the debate about counterfactual probabilities presented above and hence tell us something about which conclusion we would be right to draw from the Table 1 and other such tables.[7] These principles will turn out to be extremely powerful normative constraints on how we should do statistical inference, and they will have implications for almost everything applied statisticians do and hence for most of science.

I will defend the factualist school of thought in the form of the likelihood principle, which I introduce here very briefly.

My discussion will suggest that when we have made any observation in any scientific context, it is good to consider what each of our current

7. Of course Table 1 is only an example. My conclusions will hold in much more generality than that. But not in complete generality, unfortunately: there will be various caveats, which will be presented in chapter 2 and chapter 8.

competing hypotheses says the probability of that result was or is. We should, for example, take into account all the numbers in Table 2. To say the same thing in more technical language, it is good to consider the probability of that observation *conditional* on each of our current competing theories. (To do so is known as *conditioning*.) I will claim that these conditional probabilities — the numbers in a single column — should form the basis of any inference about which hypotheses we should accept, retain or follow. This claim is known as the likelihood principle. Of all the principles in the literature which have been considered important enough to merit their own names, the likelihood principle is the closest thing to a precise statement of factualism.

There is one important caveat to my advocacy of the likelihood principle which I must cover straight away. It is not that the likelihood principle is ever wrong. It is that sometimes it fails to answer the most important question. I have been blithely talking about "evaluating" an inference procedure as if that meant something univocal. But in fact there are (at least) two reasons why one might want to evaluate an inference procedure: reasons which seem compatible at first sight but which, in fact, may pull in different directions.

- Firstly, one might want to decide which of two competing inference procedures to use.
- Secondly, one might want to calculate some number which describes in some sense how good an inference procedure is.

I will be claiming, without hesitation, that the likelihood principle always gives the right answer to the first question (if it answers it at all; in some instances it is silent), while Frequentism is misleading at best and

downright false at worst. But I would like to say something different in answer to the second question, because the second question is ambiguous in a way in which the first is not. When we ask the first question, we are (we *must* be) imagining ourselves in possession of a token observation from which we want to make one or more inferences about unknown hypotheses. We must, roughly speaking, be in the situation which I will describe in full detail in chapter 2. In that situation, Frequentism is a very bad guide, as I will spend most of this thesis showing, while the likelihood principle is our friend, as I will suggest throughout and show fairly definitively in chapter 13. In contrast, when we ask the second question, we may want either of two things: we may want to know how well our inferences are likely to perform, in which case again Frequentism will be misleading and the likelihood principle will be helpful; or, we might want to know how well this *type* of inference would perform on repeated application in the presence of some *known true* hypothesis and variable data, without any interest at all in how in performs on any particular token data. In that case, it is not immediately clear which of the arguments I present against Frequentism in this thesis still apply, or which of the arguments in favour of the likelihood principle still apply. In fact, some of my arguments against standard forms of Frequentism in chapter 4 do still apply, but not all of them; and my arguments in favour of the likelihood principle, based as they are on the framework from chapter 2, are rendered irrelevant. Consequently, I will not attempt to reach any conclusions about how Frequentism fares when we are attempting to evaluate the long-run performance of inference procedures in the presence of known true hypotheses. To do so would be interesting, but it would take another whole thesis.

This thesis is one of very few lengthy discussions of the likelihood principle. It is the first extended treatment of the likelihood principle to take non-experimental observations (observations made without deliberate interference in the course of nature) as seriously as experimental observations. This huge widening of scope turns out to make practically no difference to the validity of the arguments I will consider; and that very absence of a difference is a noteworthy finding of my investigation.

Part I of this thesis deals with preliminary material. In chapter 2 I lay out a number of useful, relatively uncontentious idealisations carefully and explicitly but with the bare minimum of argument.[8] Then, in chapters 3 to 5, I survey the methods of statistical inference which have been proposed in the literature to date.

In Part II I motivate the likelihood principle and show that objections to it fail. I start, in chapter 7, by discussing criticisms of Frequentist analyses of Table 1. In chapter 8 I introduce the literature on the likelihood principle and begin to compare it to Frequentism. In chapters 9 to 12 I discuss criticisms of the likelihood principle.

In Part III I present proofs of the likelihood principle and a brief case study of its use. In chapter 13 I offer proofs of a new version of the likelihood principle, a version which overcomes the objections which have been voiced against previous versions, while in chapter 14 I discuss objections raised by the proof itself. At the risk of spoiling the dénoument, here is the version of the principle I will prove.

---

8. Elsewhere, I have worked on a much more critical discussion of one part of this framework of idealisations: the part involved in supposing that credences are represented by single, precise real numbers (Grossman 2005). I do not include this work here, because it would distract from the main thrust of my arguments.

The likelihood principle

> **Under certain conditions** outlined in chapter 2 and stated fully in chapter 8, **inferences from observations to hypotheses should not depend on the probabilities of observations which have not occurred**, except for the trivial constraint that these probabilities place on the probability of the actual observation under the rule that the probabilities of exclusive events cannot add up to more than 1.

The consequences of this principle reaches into many parts of scientific inference. I give a brief theoretical discussion of such consequences, and one detailed practical example, in chapter 15.

This thesis may seem to have a Bayesian subtext, because it attacks some well-known anti-Bayesian positions. This pro-Bayesian appearance is real to a certain extent: the likelihood principle does rule out many anti-Bayesian statistical procedures without ruling out very many Bayesian procedures. But that is a side effect: the likelihood principle is intended to cut across the Bayesian/non-Bayesian distinction, and may turn out to be more important than that distinction.

4. WHY PHILOSOPHERS NEED TO READ THIS THESIS

Throughout history, it has become clear from time to time that philosophy has to stop taking some aspect of science at face value, and start placing it under the philosophical microscope. To pick only the most exciting examples, the philosophical community was forced by Hume and Kant to turn its attention to the scientific notions of space, time and causality; it

was forced by Bolzano, Russell and Gödel to problematise proof; and it was forced by the founders of quantum theory to look at the determinacy of physical properties. Jeffreys, Keynes, Ramsey and de Finetti forced a re-evaluation of the philosophy of probability in the 1920s, and since then it has become standard to acknowledge that the definition and use of probability concepts needs careful thought. But this interest in the philosophy of probability has not been extended sufficiently carefully to statistical inference. It is common for even the best-educated philosophers of science to write critically, and at length, about the many ways in which probability can be understood, and yet to take statistical notions entirely at face value. I will discuss Bayesian philosophers as a particularly clear example.

Bayesianism currently enjoys a reasonable degree of orthodoxy in analytic philosophy as a theory of probability kinematics (a theory of rational changes in probability). Of course there are detractors, but among philosophers of probability and statistics there are not many. I will give reasons later for thinking that the more extreme detractors — those who decry Bayesianism even in the limited contexts in which I suggest using it — are wrong; but even if you are one of them (and a fortiori don't agree with all of my arguments) you will agree that to speak to philosophical Bayesians, as I will in this section, is to speak to a large audience.

It is almost universal for Bayesian philosophers to espouse Bayesianism in a form which entails the likelihood principle, and yet many of them — perhaps almost all of them — simultaneously espouse error rate Frequentist methodology, which is incompatible with the likelihood principle. In symbols:

B $\Longrightarrow$ likelihood principle is true

F $\Longrightarrow$ likelihood principle is false

B + F $\Longrightarrow$ contradiction

where B is almost any Bayesian theory of probability kinematics, and F is any Frequentist theory of statistical inference.

A very fine philosopher who has found himself in this position is Wesley Salmon. I use Salmon as an example because my point is best made by picking on someone who is universally agreed to be clever, *and* well versed in the literature on scientific inference including probabilistic scientific inference, *and* well versed in at least some aspects of science itself. Salmon is unimpeachable in all three respects. Many further examples from the work of other philosophers could be given, but for reasons of space I hope a single example will be enough to illustrate my point.

> When . . . scientists try to determine whether a substance is carcinogenic, they will administer the drug to one group of subjects (the experimental group) and withhold it from another group (the control group). If the drug is actually carcinogenic, then a higher percentage in the experimental group should develop cancer than in the control group. [So far, so good.] If such a difference is observed, however, the results must be subjected to appropriate statistical tests to determine the probability that such a result would occur by chance even if the drug were totally noncarcinogenic. A famous study of saccharine and bladder cancer provides a fine example. The experiment involved two stages. In the first generation of rats, the experimental group showed a higher incidence of the disease than the control group, but the difference was judged not statistically significant (at a suitable level). In the second generation of rats, the incidence of

> bladder cancer in the experimental group was sufficiently higher than in the control group to be judged statistically significant.
>
> (Salmon 2001a, p. 70)

This quotation shows one of the most important champions of Bayesianism among philosophers give a startlingly anti-Bayesian account of an experiment, even though the purpose of the paper from which this quotation is taken is to exhort us to *accept* Bayesianism. In the quoted passage, he does not quite say that Frequentist significance tests are always the best tool for drawing statistical conclusions, but he does identify the judgement of statistical significance (a Frequentist judgement) as an "appropriate statistical test", and commends work which uses statistical significance testing as a "fine example" of what is required. In doing this, he endorses the use of significance tests to draw conclusions about hypotheses; but that is counter to the likelihood principle and hence counter to Bayesianism.

This, I think, illustrates how philosophers understand Bayesianism accurately in simple probabilistic situations but have not internalised its consequences for statistical inference. From the point of view of Bayesian philosophers, it is the incompatibility of these positions which calls for the work presented in this thesis.

On with the show.

Part I

The state of play in statistical inference

— 2 —
Definitions and Axioms

1. INTRODUCTION

In this chapter I present definitions of the fundamental terms I will be using in the rest of the thesis and axioms governing their use, along with just enough discussion to establish why I have made the choices I have made.

Since this chapter mainly takes care of terminological issues, and since terminological issues tend to have relatively few deep links to each other, this chapter is more like a collection of short stories than a long narrative. I beg the reader's indulgence. The short stories include basic notation, basic axioms, an exciting (to me at least) new way of describing exchangeability, and a variety of small controversies related to terminology.

One disclaimer: the reader will notice that I attempt to resolve only a very few of the many pressing problems in philosophy of probability. I hope to show by example that it is possible to achieve a good deal of insight into statistics without first giving the final word on probability. In this chapter I define my probabilistic terminology fully but say very little about the interpretation of probability and almost nothing about its ontology. A few further issues in the philosophy of probability will intrude into later chapters — most importantly, a discussion of epistemic probability in chapter 4 — but we will see that many issues in the philosophy of

probability do not need to be discussed. For example, qua philosopher of probability I would like to know whether objective chance is inherent in the world or is a Humean projection (or something else); but qua philosopher of statistics I can achieve a lot without that question arising.

2. THE SCOPE OF THIS THESIS

The exact range of applicability of the conclusions of this thesis is simply the cases in which we can uncontentiously draw a table such as Table 1 (finite or infinite). In other words, it is the cases in which we have an agreed probabilistic model which says which hypotheses are under consideration and what the probability of each possible observation is according to each hypothesis.

This thesis is about *inference procedures* in science. One of my claims will be that the study of the philosophy of statistics (and hence, derivatively, the philosophy of most of the special sciences) can be clarified tremendously by analyses of inference procedures, largely (although of course not entirely) independently of analyses of more primitive concepts (such as "evidence", for example). I will therefore give an explicit definition of "inference procedures", at the risk of stating the obvious.

> An **inference procedure** is a formal, or obviously formalisable, method for using specified observations to draw conclusions about specified hypotheses.[9]

9. Throughout this thesis, important terms are set in **bold text** where they are defined, while *italic text* is used both for the definitions of relatively unimportant terms and for general emphasis; except that within quotations from other authors bold text is my emphasis while italic text is the original authors' emphasis.

I sometimes refer to inference procedures as *methods*; sometimes I do this just for variety and sometimes I do it because I want to ephasise the operational nature of inference procedures.

I am discussing ways to evaluate inference procedures, not ways to evaluate individual inferences. Does this mean that I can't draw any conclusions about individual inferences? It almost does. I cannot conclusively infer from the deficiencies of an inference procedure that any given inference is a bad one. This admission may seem rather weak, but it is the best anyone can do at such a general level of analysis. Indeed, it is the best anyone can do not only in statistical inference but even in better developed fields of inference such as deductive logic. Deductive logic confirms an individual inference as valid when it instatiates a valid procedure . . . regardless of whether it also instantiates an invalid procedure (which in fact it always does, since any non-trivial argument instantiates the argument form $p \vdash q, p \neq q$). This does not deter us from working out which deductive inference procedures are invalid. Finding invalid inference procedures has proved to be useful, despite the fact that not all instances of invalid inference procedures have token invalidity. We should expect the same to be true of inductive inference procedures: it will be useful to know which are invalid, even though arguments constructed using invalid inference procedures may occasionally be good arguments.

HYPOTHESES

I will be concentrating on *statistical* inference procedures, and so it will be useful to restrict the use of the word "hypotheses" in the above definition, in two ways.

Firstly, my interest in hypotheses will mostly be restricted to hypotheses which specify *precise* probabilities for all possible outcomes of a given experiment or of a given observational situation. (I mean "possible" in the sense of forseeable, of course, since my topic is entirely epistemological. Metaphysical possibility is irrelevant. A third type of possibility, logical possibility, is factored in to my work via the axiomatic probability theory which I will state later in this chapter.) This type of hypothesis is known in the literature as a "simple hypothesis". I will use the qualification "simple" often enough to remind the reader that I am discussing precise hypotheses, but for the sake of a bit of grammatical elegance I will only use it when the distinction between simple and compound (non-simple) hypotheses is directly relevant, not every time I use the word "hypothesis". Many parts of the literature use the terminology in the way I am suggesting or, compatibly, restrict the word "hypothesis" to simple hypotheses.[10]

---

10. Thus, "if a distribution depends upon l parameters, and a hypothesis specifies unique values for k of these parameters, we call the hypothesis *simple* if $k = l$ and *composite* if $k < l$" (Stuart et al. 1999, p. 171), although unlike Stuart et al. I will not generally assume that hypotheses are characterised by parameters. Similarly, "By hypotheses we mean statements which specify probabilities." — Barnard, in (Savage & discussants 1962, p. 69).

Some authors use the word "theory" interchangeably with "hypothesis", but I will need to use the word "theory" to mean theory of statistical inference, so I will never use it to mean scientific hypothesis.

A disadvantage of my stipulation that hypotheses must specify probabilities is that it forces me to restrict the meaning of the word "hypothesis" to exclude statements which are functions of the observations which we wish to use to make inferences about those very statements (hypotheses h_i such that $h_i = f(x_a)$ for some f, in the notation which I will introduce below). Let me briefly (just for the duration of this paragraph) introduce the term "hyperthesis" to refer to such a statement, and "metathesis" to refer jointly to hypotheses and hypertheses. Now, were I to measure the heights of a random sample of two philosophers, and then to wonder whether the taller of the two people in my sample was cleverer than the shorter one, assertions about their relative braininess *based on knowledge of who was in the sample* would be hypertheses, not hypotheses. The problematic aspect of such hypertheses is that their meanings change when the observation is made: beforehand they are general (or, if you like, variable) assertions about the whole population of philosophers, but afterwards they are assertions about two particular, known philosophers, say Hilary Putnam and Ruth Anna Putnam. Consider whether Hilary is cleverer than Ruth Anna. It is, I hope, obvious that the likelihood principle applies to this question if it applies anywhere: if only the probability of the observation according to various hypotheses is relevant to

The practical advantages of discussing statistical inference in terms of inference procedures will become clear as we go. There is also a theoretical advantage: discussing inference procedures is (I claim) exactly what we need to do in order to abstract away from unimportant details of specific contexts of application of inference methods without losing the details that matter. A discussion of the concept of evidence, to take my example of a more primitive concept that I could have started with instead of inference procedures, is extremely important — indeed, I have written on that topic (Moore & Grossman 2003, Grossman & Mackenzie 2005) — but it requires a discussion of sociological and political issues surrounding the use of the

---

inference about those same hypotheses (as the likelihood principle asserts) then surely it is also the case that the probabilities of the observation according to various hypotheses *plus* the probabilities of the observation according to various hypertheses is sufficient for inference about metatheses. To illustrate with the Putnams, if only the probabilities according to various hypotheses of observing the Putnams are relevant to inference from the observation to any hypothesis, then surely those same probabilities *plus* the probabilties according to various hypertheses of observing the Putnams are sufficient for inference about all metatheses. Thus, if the arguments of this thesis in favour of the likelihood principle for hypotheses narrowly construed have any weight, then the likelihood principle will also be true for metatheses in general. However, dealing with hypertheses would considerably complicate some of the arguments in this thesis, because many of my arguments use the fixed nature of hypotheses as a simplifying assumption; so I do not attempt to give detailed arguments in favour of the view that the likelihood principle applies to hypertheses as well as to hypotheses.

The problem which I have just described is known in the literature as "the prediction problem" (Dawid 1986, p. 197), even though most problems which we might non-technically call prediction problems *do not* have this form and *do* fall within the scope of this thesis — for example, the question of how clever I ought to expect a *third* randomly-sampled philosopher to be, given information from a sample of two random philosophers, or the question of how clever I ought to think the population of philosophers as a whole, again given information from a sample of two, are common-or-garden prediction problems in which the hypotheses do not depend on the observation for their meanings, and such hypotheses are well within the scope of this thesis. (Any such problem *could* be stated in terms of hypotheses which are functions of the observation, by taking "the observation" to include hypothetical future observations of the third philosopher or of the whole population of philosophers, but although it *could* be stated in such a form it *need* not be.)

It is possible in principle to incorporate the so-called prediction problem into the framework presented here. Dawid (1986, p. 197) sketches a proof that the stopping rule principle, which he rightly calls "the most controversial of all the consequences of the likelihood principle", is true even in prediction problems. However, for simplicity of exposition of the likelihood principle (which is not so easily proved to apply to prediction problems as the stopping rule principle is), I restrict the meaning of "hypothesis" so as to exclude prediction problems. The only exception is at the end of chapter 13, where I state a mathematical result about the prediction problem, without proof, in order to show that it is at least plausible that the likelihood principle is true even in prediction problems (as technically construed; I emphasize again that common-or-garden prediction problems are unproblematic).

word "evidence" which have very little bearing on the normative task undertaken in this thesis.

Having said that, I will discuss several specific contexts, for illustrative purposes and to check my assertion that I am abstracting the important aspects of statistical inference. This will be especially clear in chapter 15, in which I will discuss an urgent problem in applied statistical inference with enough scientific and social context to test the accuracy and relevance of my theorising.

THEORIES OF THEORY CHANGE

Why do I restrict my conclusions to only part of science, so that they cannot give us a complete theory of theory change? Recall that the range of applicability of the conclusions of this thesis is the cases in which we have an agreed probabilistic model which says which hypotheses are under consideration and what the probability of each possible observation is according to each hypothesis. This is an extremely common situation in science: indeed, it covers the vast majority of scientific experimentation, especially in the biomedical sciences. However, the reader can easily think of examples that are not covered by this sort of model. That is because the *atypical* cases that are not covered are some of the most *interesting* cases for philosophers and historians of science. Cases in which theories are only vaguely described but are nevertheless in active competition with each other, as was the case with theories of the shapes of the continents in the 1960s, are of extreme interest to all of us, especially to those of a Kuhnian disposition. The reason I do not discuss these cases in this thesis is probably obvious: they raise the problem of how to make a mathematical

model describing the theory. That problem is of course important and interesting, but the considerations which it brings into play hardly overlap at all with the considerations needed to work out how to *analyse* a *given* mathematical model. It therefore makes no sense to attempt both in one thesis; and I will attempt only the latter.

Fortunately, most of science is not like 1960s theories of continental drift. In the vast bulk of scientific work the hypotheses under active consideration are extremely clearly described, to the point where the probabilities involved are stated explicitly by the hypotheses. For example, in all clinical trials of treatments for life-threatening diseases, there is a continuum of hypotheses stating that the life expectancy (expressed as relative risk of death adjusted for measurable predictive factors such as age) of subjects who are given the experimental treatment is x, for all x between 0 and 1. Each of these hypotheses has sub-hypotheses describing the possible side-effects of the treatment, but we can ignore those sub-hypotheses for simplicity — they are just additional rows in the table and make no difference to the principles of analysis. What's more, this clarity of hypotheses is observed not *just* during periods of Kuhnian normal science (if indeed there are any) but during periods of conflict between rival theories as well. It is very common (although not, I admit, universal) for rival theories to each have well defined hypotheses which are considered to be workable and precise (although false and perhaps unimportant) even by their opponents. In other words, most of science is stamp-collecting, and this thesis, I hope, describes stamp collecting rather well.

What forms can these hypotheses take? In assuming that they define probabilities of possible outcomes, I am assuming that they are partly

mathematical, so it might be expected that I would have to say something about their mathematical form. But thankfully that isn't necessary. A philosopher can state a statistician's model of nature as simply

$$p(X = x|h) = f_h(x)$$

where x represents possible data, X is a *random variable* (statisticians' jargon for a function from the structured set of possible events to the set of possible observation reports), p denotes probability and f_h is the probabilistic model according to hypothesis h.

In general, x is a vector, often of high dimension — typically several dimensions for each observed data point, which means that in a large medical study, for example, the dimensionality of x is in the hundreds of thousands or millions (although the dimensionality can often be reduced by summarising the data using sufficient statistics, which I discuss in chapter 13 when I come to the sufficiency principle).

There are various questions we *must* ask about f and x for philosophical purposes, but the functional form of f (log-Normal, Cauchy or whatever) is not one of them, or at least is not foremost among them, as we will see from the amount of work we can do without it. This should come as a great relief to those of us who are not mathematicians.

Among the questions which we *cannot* ignore, for reasons which will become apparent later, are:

- whether f is discrete or continuous,
- whether x is multidimensional

- and if so whether the dimensions of x are commensurable (in the mathematical sense of being multiples of each other, not in any subtle Kuhnian sense).

I will say more about the problems of multidimensional data in chapter 15.

Very occasionally I will assume that f is either continuous or discrete (finite); but mostly I will assume nothing about it at all except that it takes values between 0 and 1 inclusive and integrates to 1.

3. BASIC NOTATION

I use small letters in $p(x|y)$ as shorthand for $p(X' = x|\Upsilon' = y)$, where X' and Υ' are random variables. And similarly $p(F(x)|G(x))$ is shorthand for $p(F(X') = F(x)|G(\Upsilon') = G(y))$.

Random variable is standard terminology in discussions of statistics, but it is slightly misleading. Fortunately, I will be able to do without discussing random variables most of the time; but not quite all the time. A random variable such as X' is (famously) neither random nor a variable: it is a function which associates a real number with each possible observation into real numbers (typically, subject to the constraint that $(\forall x \in R)$ the set $\{y : X'(y) \leq x\}$ is measurable according to a standard measure on R).

Although X', a random variable, is not a variable, x, a possible value of X', *is* a variable, and may in some cases need to be treated as random (although only rarely in this thesis). I write the set of possible values of x — in other words, the range of the random variable X' — as X. Elsewhere in the literature, plain capitals (X, Υ) usually stand for random variables, not for sets of possible outcomes, but for my purposes the range of each random variable is more important than the random variable itself, and it

is well worth reserving the simpler notation (X rather than X') for the more important concept.

The following terms have meanings that are more or less specific to this thesis.

A *doxastic agent* is the epistemic agent from whose point of view a probabilistic or statistical inference is meant to be a rational one. As we will see, some theories of statistical inference require such an agent, while others (notably Frequentism) do not.

X is a space of possible observations.

x_a is an actual observation ("a" for "actual") — either the result of a single experiment or observational situation, or the totality of results from a set of experiments and observational situations which we wish to analyse together. When x_a is the only observation (or set of observations) being used to make inferences about a hypothesis space H, I will often refer to x_a as *the* actual observation. Presumably (human fallibility aside) it includes all the relevant data available to the agent making the inferences, even though it is not necessarily the only observation relevant to H which has ever been made by anyone.

H is the set of hypotheses under active consideration by anyone involved in the process of inference.

Θ is a set (typically but not necessarily an ordered set) which indexes the set of hypotheses under consideration. I will always treat θ as an index on the *whole* set of hypotheses.[11] Very occasionally, in quotations from other

11. In other words, $(\forall h \in H)\,(\exists \theta \in \Theta : H_\theta = h)$.

authors, it will be just a partial index on H. In this rare case, θ will be one of several parameters in a parametric model.

AN OBJECTION TO USING X

Although the above set of quantities is the usual starting point for discussions of statistical inference, Lindley (1990a) complains that the sample space, X, is irrelevant to discussions of statistical inference (although not, of course, to discussions of experimental design, which are more or less identical with discussions of the value of alternative choices of X). I will quote Lindley at length, because his views about X will help to motivate the main contentions of this thesis:

> The [Bayesian] objects to the . . . use of an arbitrary sample space [X] . . .
>
> Since the arbitrariness of the sample space is not often appreciated, it might be worth discussing it. The practical reality is the data [x_a] (not X), the parameter space Θ and the likelihood function $p([x_a]|\cdot)$ for fixed x_a and variable θ. The sample space X is, to use Jeffreys' vivid description, the class of observations that might have been obtained but weren't. Both in practice and in theory, this class can be hard to specify. . . .
>
> Let me digress [from Lindley's topic, not from mine] to answer a point raised by two referees . . . to the effect that the sample space X and its associated densities are the primary entities from which the likelihood is derived. This need not be so. Although it is customary for any paper in probability to begin with the triplet $(X, [H], p)$ . . . this complete specification is not necessary and often extends beyond the bounds of the reality. Why, when discussing probabilities, is it necessary to have them defined for more sets than those of interest? . . . The $(X, [H], p)$-introduction is a useful starting point for many problems [such

> as experimental design] but not [for statistical inference] when the data are to hand.
>
> (Lindley 1990b, p. 46)

I agree with Lindley's claim that X is not essential to statistical inference. His use of the phrase "the bounds of reality" is not due to an interest in ontology, but to a recognition that the hypothetical repetitions of an experiment on which the sample space is based are often, for technical reasons, not the same as any repetitions of the experiment which could conceivably be expected.[12] In any case, the most important foundations of Lindley's complaint are not whether or not X is in any sense real but two less difficult issues: whether or not it is arbitrary and whether or not it is "of interest". I will discuss in chapter 7 and chapter 15 the extent to which X is arbitrary. As for being of interest, it is clearly of only subsidiary interest at most, compared to H; if we could do statistical inference without it, there would be, at the very least, an argument from parsimony for doing so.

Further details of the reasons for which I agree with Lindley do not belong here; they will be discussed abundantly in Part II of this thesis.

Lindley's point is equivalent to a version of the likelihood principle, which I will champion in Part II. But my aim is not only to explain how the likelihood principle works but also to show that it is an improvement over competing principles of statistical inference; and in order to discuss these competing principles I must be able to talk about a complete sample space, X, no matter how much I agree with Lindley that such a thing is

12. Two examples are the fixing of the marginal totals in contingency tables, and the fixing of the "independent" variable in bivariate regression analysis: in each of theses cases, the sample space which is used in the analysis is given constraints which need not be expected to apply to repetitions of the experiment (Lindley 1990b, p. 47). Lindley is uncontentiously right about these examples; but the details are unimportant for our purposes.

an unnecessary invention. Consequently, I will assume throughout this thesis that X has been specified, whether arbitrarily or not and whether unnecessarily or not.

NON-PARAMETRIC STATISTICS

I will generally assume the existence of an index set on H, and in a loose sense this index set will give us a parameter on H; but this does not restrict my work to what statisticians call "parametric" models. As the authoritative *Kendall's Advanced Theory of Statistics* explains the terminology,

> [When] no parameter values are specified in the statement of the hypothesis; we might reasonably call such a hypothesis *non-parametric*. . . . [When the hypothesis] does not even specify the underlying form of the distribution [it] may reasonably be termed *distribution-free*. Notwithstanding these distinctions, the statistical literature now commonly applies the label 'non-parametric' to test procedures that we have just termed 'distribution-free'[.]
>
> (Stuart et al. 1999, p. 171)

In other words, a parametric hypothesis is one which not only is indexed by a parameter(s) but also *mentions* its parameter(s). This point is relevant to the long shelves of books in the mathematics library with titles such as "non-parametric statistics". These books are (almost exclusively) about finite collections of arbitrary hypotheses which do not mention any parameters (or which, for a variety of idiosyncratic reasons, are to be analysed as if they did not mention any parameters). Since they are finite, they can be indexed; and since they can be indexed, the work presented in this thesis is directly applicable to them, even when the work presented here appears to be parametric. (See also Salsburg 1989 for a very general argument

to the effect that all interesting theories apply equally to parametric and non-parametric models.)

Under what general circumstances can we be sure that there is an index set on H? First of all, if H is finite or countably infinite then of course it can be indexed. (Strictly speaking it should be provably countably infinite by a constructive proof, but this is an unimportant detail.) Interestingly, if there is a countable number of observables, each with a countable number of possible states, H can be the set of *all* probability distributions and thus the arguments to be given here can be *completely* general (in terms of H at least) and H can still be indexed. Alternatively, if we can fully describe an uncountable but continuous distribution (either in natural language or in mathematics) then we can still count it as being indexed by parameters, the parameters in this case being whatever lexical tokens are used to describe the function (possibly an infinite number of them, if the definition contains terms like $(\forall i \in Z)$). So H can be indexed in the discrete case and in all describable continuous cases. In most systems of pure mathematics there are, provably, indescribable functions; but as philosophers of applied mathematics we need not worry about them too much.

I will assume in most of this thesis that *all* variables in a model should be considered as parameters. This is one of the places in which the likelihood principle is open to re-formulation, and it is an issue on which Bayesians have interestingly diverging opinions, as I will discuss in chapters 9 to 12.

4. PROBABILITY AXIOMS

It will be useful to have a mathematical axiomatisation of probability. I will give axioms based on a set of axioms by Harold Jeffreys (1961, pp. 16-25) (not to be confused with Richard Jeffrey).

Jeffreys's axioms take conditional probability, $P(a|b)$ (which I write henceforth with a lower-case p, $p(a|b)$), to be primitive. In such a system, probability is relative to background knowledge even though it might seem that it shouldn't be — just as time is relative to a reference frame even though it might seem that it shouldn't be. In a moment I will argue that Jeffreys is right to take this stance, although I will conclude that it need not make much difference. From among the various axiomatisations which take conditional probability to be primitive, there is no important reason to prefer Jeffreys's, but he does perhaps profit by paying particular attention to the epistemological context within which his axioms are to operate.[13]

CONDITIONAL PROBABILITY AS PRIMITIVE

One advantage of taking conditional probability to be primitive is that this avoids the problems raised by trying to find a *definition* of conditional probability. The definition of conditional probability used by standard theories that take non-conditional probability as primitive is the equation:

13. The many sets of probability axioms (including the most famous sets due to Kolmogorov, Renyi and Carnap and the seminal set due to Keynes) are in agreement on most points except for ontology and hence more or less interchangeable with each other as far as applied mathematics is concerned. Many other sets of axioms would have done almost as well for my purposes as Jeffreys's. There might have been some advantage to using axioms which allowed non-real-valued probabilities. I explore some such possibilities in (Grossman 2005), and I find there that the extra complications, although valuable in their own right, seem likely to add very little to our understanding of theories of statistical inference.

$$p(a|b) = \frac{p(a\&b)}{p(b)}, \text{ provided } p(b) \neq 0.$$

This definition proves awkward for such theories, because sometimes $p(b)$ *is* 0 (Hájek 2003). This problem of zero divisors could be avoided, at least when talking about epistemic probabilities, by stipulating that no rational agent should ever, strictly speaking, hold the probability of anything to be quite zero. My own preferred way of doing this would be to construct a probabilistic equivalent of David Lewis's (1996) theory of knowledge. In Lewis's theory, we know everything which we are justified in believing when we're ignoring possibilities that may be "properly ignored" — and what may be properly ignored is contextual. Similarly, one could argue, we call things zero-probability when we're ignoring possibilities that may properly be ignored . . . and that is contextual too. Nothing is ever zero-probability simpliciter. However, this does not save the proposed definition of conditional probability from a second problem. Hájek suggests substituting the variables in the above equation as follows (2003, slightly paraphrased):

- a = I get heads
- b = I toss a coin

Then the supposed definition of conditional probability would give us:

$$p(\text{I get heads} \mid \text{I toss a coin}) = \frac{p(\text{I toss a coin and get heads})}{p(\text{I toss a coin})}.$$

But this is a lousy definition of the left-hand side, argues Hájek, because the left-hand side can be well defined (e.g., equal to ½ for a fair coin) even if the

right-hand side is left hopelessly vague. I consider this second argument of Hájek's to be a knock-down argument in favour of making conditional probability primitive.[14]

Despite these arguments, the conditional nature of probability can be ignored in many cases. Even though I will be treating all probabilities as conditional in my *axioms*, I will usually be talking about cases in which the question of exactly what a particular probability is conditional on is not interesting. And although I deny that we can always reduce a doxastic agent's total belief state to a simple description, it seems clear that all we need in order to enable us to do so is the caveat that we are interested only in belief states which can be shared by members of an epistemic community. (I am especially interested in scientific communities, of course, because they are the paradigm users of statistical inference.) So I will be using categorical (non-conditional) probabilities freely in this thesis after all, as do many other authors who take conditional probability to be primitive,

---

14. A third argument, just in case one is needed, is an epistemic argument which applies (at least) to epistemic probabilities rationally ascribed by epistemic agents. This argument shows that an epistemic agent cannot rationally hold a categorical (non-conditional) probability, except in a trivial way. Suppose, for the purposes of reductio, that an agent holds that the categorical probability of j is k: in symbols, $p(j) = k$. Suppose the agent also believes m, where m is distinct from j. Then either m is probabilistically irrelevant to $p(j)$ — in other words, $p(j|m) = p(j)$ — or m is probabilistically relevant to $p(j)$. In the latter case, it is irrational to say that $p(j)$ is the agent's probability of j. $p(j|m)$ may still not be a rational ascription of probability, since there may be further epistemic factors to take into account — call them $y, z, w, t, \ldots$ — but these factors only cause the rational ascription of probability to be as it were more conditional. If $x, y, z, w, t, \ldots$ are all relevant to the probability of j then the only probability that the agent can rationally claim to be *her* probability of a is $p(j|B)$, where B is her whole belief state. (Note that my argument uses only pairwise comparisons of parts of belief states, and therefore applies even if agents' belief states cannot be fully decomposed into summable components.)

This third argument leaves me with the conclusion that the only cases in which conditional probabilities can be avoided, even were we not to buy Hájek's arguments, would be:

- when we are definitely and only considering non-epistemic probabilities; and
- in the trivial case in which, for all m in the domain of things that can occupy the position after the "|" in a conditional probability ascription, $p(j|m) = p(j)$.

including Jeffreys and Lindley. This is the purpose of my addition of Axiom 8 to Jeffreys's axioms in the next section.

STATEMENT OF PROBABILITY AXIOMS

The following axioms are from (Jeffreys 1961, pp. 16-25), except for Axiom 0, which I have added. Axioms 1-7 and Conventions 1-3 are reproduced here almost verbatim (my changes are in square brackets), but omitting Jeffreys's interspersed comments.

Axiom 0. The domain of the propositions mentioned in the following axioms is a fixed set of sentences, H.[15]

Axiom 1. Given [the truth of a proposition] p, q is either more, equally, or less probable than [p], and no two of these alternatives can be true.

Axiom 2. If p, q, r, s are four propositions, and, given p, q is more probable than r and r is more probable than s, then, given p, q is more probable than s.

---

15. The purpose of my added Axiom 0 is to resolve an ambiguity in Jeffreys's axioms: he leaves it unclear whether their contentful primitives are *sentences* or *propositions*. Jeffreys's use of quotation marks in some phrases seems to suggest that they are sentences, while his failure to use quotation marks in other phrases such as pq seems to suggest that they are propositions (especially since one does not make a logical union of sentences by concatenating them, and even more especially since he *calls* them propositions). The main point at issue is that sentences, unlike propositions, are only meaningful if they are produced (or at least imagined to be produced) by specific epistemic agents. In this way, sentences suit the philosophy of statistics better than propositions. (Earman (1992, p. 35) seems to agree, although he does not emphasize the point.) The sentence-producing doxastic agents in question are the members of scientific communities studying the specific problems for which statistical models are produced. But this will not *quite* do: we will need a way to treat synonymous sentences as being identical (whereas synonymous propositions simply *are* identical, by definition). For my limited purposes, I can do this with a wave of the hand. No subtle problems about synonymy will crop up in this thesis. I will only be discussing cases in which it is perfectly clear to a given scientific community which sentences are to all intents and purposes synonymous with which other sentences. Any situation in which a dispute about synonymy is sufficiently heated to have short-term scientific consequences is a case of prima facie (at least) Kuhnian incommensurability, and such cases I have already foresworn. I do not doubt that in more subtle ways synonymy is *always* in dispute, but if the dispute has no short-term scientific consequences then it is irrelevant to statistical inference.

Axiom 3. All propositions deducible from a proposition p have the same probability on data p; and all propositions inconsistent with p have the same probability on data p.

Axiom 4. If, given p, q and q' cannot both be true, and if, given p, r and r' cannot both be true, and if, given p, q and r are equally probable and q' and r' are equally probable, then, given p, 'q or q'' and 'r or r'' are equally probable.

Convention 1. We assign the larger number on given data to the more probable proposition (and therefore equal numbers to equally probable propositions).

Convention 2. If, given p, q and q' are exclusive, then the number assigned on data p to 'q or q'' is the sum of those assigned to q and to q'.

Axiom 5. The set of possible probabilities on given data, ordered in terms of the relation 'more probable than', can be put into one-one correspondence with a set of real numbers in increasing order.

Convention 3. If p entails q, then $P(q|p) = 1$.

Axiom 6. If pq entails r, then $P(qr|p) = P(q|p)$.

Axiom 7. For any propositions p, q, r, $P(qr|p) = P(q|p)\ P(r|qp)\ /\ P(q|qp)$.

5. EXCHANGEABILITY AND MULTISETS

In what I hope is a useful terminological innovation, I would like to draw attention to the similarity between the notions of **exchangeable sequence**

(invented by statisticians) and **multiset** (invented by computer scientists and used by logicians but not — until now — by other philosophers).

EXCHANGEABILITY

When we do statistics, we usually think of ourselves as observing *sequences* of events, and so, not surprisingly, mathematical statistics tends to be worked out in terms of the mathematics of sequences. But we will see that sequences are not quite the best mathematical tools for the job.

A **sequence** is a set of items occurring in some order. Sequences are usually written in angle brackets. By definition, the order of terms in a sequence is an essential property of that particular sequence; so when we distinguish sequences from each other, we count $\langle A, B, C, A\rangle$ as being different from $\langle A, A, B, C\rangle$, as well as being different from $\langle A, B, C\rangle$. We will be discussing sequences of observations; or, speaking more strictly, we will be considering sequences of opportunities to make observations.[16]

Statisticians sometimes talk about *sets* of events instead of sequences of events. In a statistical context, this is really talk of sequences after all, because statisticians always take each event to be implicitly labelled (indexed) by its spatiotemporal location; hence the set is really an ordered set; in other words, a sequence. This equivalence in statistical language between sets of events and sequences of events is made explicit whenever necessary in the statistical literature.

16. The distinction between observations on the one hand and observation opportunities on the other is that for the latter we don't need to know the outcomes of the observations, or even be confident that the observations will be made. Statisticians, parsimoniously but confusingly for us philosophers, allow themselves to use the word "event" ambiguously to cover both observation opportunities and their outcomes.

In 1937, Bruno de Finetti made a breakthrough in mathematical statistics by inventing (or, for the Platonists among us, discovering) the tremendously useful property of *exchangeability*. De Finetti's definition:

> [Events] are said to be exchangeable if they play a symmetrical role with respect to every problem of probability[.]
>
> (de Finetti 1980, p. 195)

A more careful definition:

> Two or more observation opportunities are **exchangeable** iff we have the same information about them; and two or more observation outcomes A and B are **exchangeable** iff it does us no good, either before or after the fact, to distinguish between the outcome sequences $\langle A, B\rangle$ and $\langle B, A\rangle$. Exchangeability for events means one or the other of these according to context. In all cases, assignments of exchangeability, when properly made, are relative to some specified purpose; in this thesis, the purpose will always be the purpose of making inferences about a set of hypotheses.

For example: I ask you to use a pin to pick two words at random from "Two Dogmas of Empiricism" (Quine 1980), in order to estimate the number of times the word "Carnap" appears in that paper. You pick the words "Ptolemy" and "Lewis" (meaning C. I. Lewis). I should consider your two observations to be exchangeable, because it would be irrational of me to make any inferences about Quine's prose on the basis of the order in which you came across the words . . . unless, that is, I have some reason to believe that you were using the pin to sample words in a strange and time-dependent manner.

The judgement that your observations are exchangeable is a synthetic judgement (pace Quine!) which rests on an understanding of the mechanics of the situation and does not have any logical justification (as far as I know). It is not, for example, meant to be justified as an application of the principle of indifference. Although often difficult to justify, the judgement is often easy to make. In particular, a method of random sampling from a population that would disturb exchangeability would have to be so strange that it would hardly deserve the name "sampling". One such method is poking the pin at the words of Quine's text in alphabetical order if you already know that "Carnap" appears more than twice, and in reverse alphabetical order otherwise. I have no reason to think you've done anything of the sort, and I declare that I am ignoring the possibility when I assign exchangeability to the experimental outcomes. The sort of get-out clause that I would invoke later if I found out you had sampled like that is always implicit in any claim that an agent *ought* to make an assignment of exchangeability. Broadly speaking the get-out clause is a ceteris paribus clause: it says that if things I haven't thought of yet turn out not to be equal then I may be forced to reconsider my epistemic position. I must acknowledge here that ceteris paribus clauses are rarely if ever easy to analyse (Earman et al. 2002). But recall that I am not claiming that it is easy to *justify* any assignment of exchangeability, which would (it seems) require a rather vague ceteris paribus clause; only that it is easy to *make* such an assignment. *Making* such an assignment does not require a ceteris paribus clause.

Exchangeability can be given a more mathematical definition (below) which makes clear some of its nice mathematical properties. One very nice mathematical property, for example, is that every finite exchangeable

sequence of events defined on a discrete event space (i.e., a set of possible outcomes with the topological characteristics that we would normally associate with a subset of the whole numbers) can be modelled by the probabilist's favourite experimental setup, drawing balls from urns (Diaconis & Freedman 1980, p. 234). For this and other reasons, modern statisticians usually define exchangeability using the sort of mathematical gloss which makes contentious assumptions about probability theory. For example:

> **Exchangeability** is the property that a sequence of events $\langle A1, A2, \ldots \rangle$ has when "the subjective joint distribution over $\langle A1, A2, \ldots \rangle$ is unchanged by any (finite) re-labelling of the events." (Dawid 1977, p. 218)

Note that Dawid's definition does not really apply to a sequence of events simpliciter but rather applies to a sequence of events *with respect to some probability distribution.* The idea of a subjective distribution is a worryingly partisan concept. We can avoid this difficulty in most epistemic contexts by saying that when a sequence of events is exchangeable with respect to all the probability distributions in play (those defined by the set of hypotheses H) it is exchangeable simpliciter. This clarification makes Dawid's definition equivalent to my definition above, in which it was required that "it does us no good" (implicitly: according to any probability distribution under consideration) to distinguish between the memers of a sequence. A similar clarification should be applied to Gelman et al. 's definition below.

The idea of exchangeability is central to almost all statistical analysis. In non-Bayesian statistics, the word "exchangeability" is rarely used, but instead exchangeability takes the form of the assignment of a set of identically distributed variables, together with an analysis that uses only

test statistics which are blind to the order in which the variables occur or are observed — typically, the test statistic is the sum or the product of the variables, or a function of their sum or product. In the Bayesian statistics literature, exchangeability is used directly. To quote from a fairly comprehensive treatment of applied Bayesian statistics,

> The usual starting point of a statistical analysis is the (often tacit) assumption that the n values y_i may be regarded as exchangeable, meaning that the joint probability density $p(y_1, \ldots, y_n)$ should be invariant to permutations of the indexes. . . . The idea of exchangeability is fundamental to statistics[.]
>
> (Gelman et al. 1995, p. 6)

De Finetti also invented a weaker notion, known as partial exchangeability:

> A probability assignment P on sequences of length n is partially exchangeable for a statistic T if
>
> $$T(x) = T(y) \vdash P(x) = P(y)$$
>
> where x and y are sequences of length n.
>
> (Diaconis & Freedman 1980, p. 238)

I will explain the relationship between exchangeability and partial exchangeability shortly.

MULTISETS

I have been surprised to discover (thanks to Allen Hazen) than an idea that turns out to be interchangeable with exchangeability was invented by computer scientists, completely independently of the statistical literature, in the 1960s. This idea is the multiset:

> A **multiset** is a collection of items in which multiple appearances of the same item are significant but order is not.

For example, the multiset $[A, B, C, A]$ is the *same* multiset as $[A, A, B, C]$, but it is not the same as the multiset $[A, B, C]$. I write multisets using square brackets, following (Meyer & McRobbie 1982). The concept seems to have been invented by Knuth, who credits the terminology to N. G. de Bruijn (Knuth 1968, p. 551).

Meyer and McRobbie, who find a use for multisets of premisses in relevant logic, explain the intuitive appeal of the concept with the following diagram:

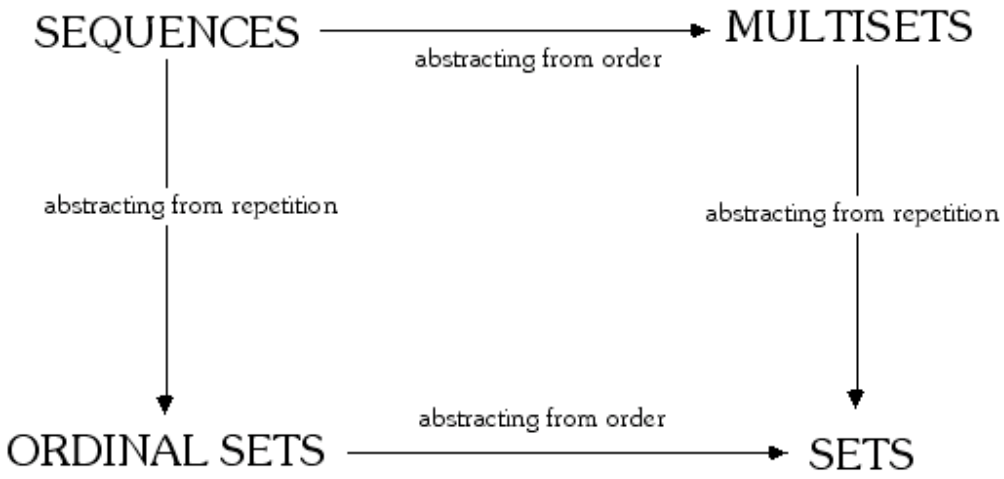

Figure 1: Multisets
(adapted from Meyer & McRobbie 1992)

Multisets can equally be seen as things like sequences in which order does not matter or as things like sets in which repetition is allowed. As this suggests, multisets are easily axiomatised in terms of sets (Knuth 1968) or in terms of sequences. This is only useful in order to check that no problems crop up in the transfinite case, and I will not take the space to reproduce an axiomatisation here.

Multisets give us a natural way to express many ideas that have been with us for some time. For example, every whole number can be expressed as the multiset of its prime divisors: 6 is equal to 2×3, 12 is equal to $2 \times 2 \times 3$, and so on; and a large part of contemporary number theory relies on this decomposition. But this is not a decomposition into *sets* of factors, because to say that the factors of 12 are the members of the set $\{2, 2, 3\}$ would be quite wrong, given that that set is identical to the set $\{2, 3\}$, the members of which do not multiply to give 12. The factors of 12 are $2, 2$ and 3, or they are 2 (twice) and 3; they are not merely 2 and 3. And nor is it a decomposition into *sequences* of factors, because no sequence represents the factors of 12 uniquely: the factors can be represented by $\langle 2, 2, 3\rangle$, or $\langle 3, 2, 2\rangle$, or $\langle 2, 3, 2\rangle$, but each of these representations implies a specificity that is not there, and is therefore misleading, just as it is misleading to represent a measurement of 6. 1 inches as "6. 10000 inches". What we should say instead is that the factors of 12 are represented by the *multiset* $[2, 2, 3]$.

Some programming languages refer to multisets as "bags" (Lewis 1995, p. 71).

It might be worth remarking that it is uniformly agreed that it is essential to the concept of a set that it is extensional. Indeed, extensionality

(the property of depending only on its members, without counting repetitions) is the only feature of sets that is agreed on by all set theorists. It is similarly essential to the concept of a sequence that it is ordered. So we should not try to capture the new notion of a multiset by merely extending the meanings of the words "set" or "sequence".

A small payoff of the terminology of multisets is that it allows us to express the relationship between exchangeability and partial exchangeability in a neat way: the two are equivalent when the function T in the definition of partial exchangeability above is the function that takes a sequence to the multiset of its members.

The big payoff, for me, of multiset terminology is that for any epistemic purpose any *exchangeable sequence* of observations is equivalent to a *multiset* of observations; and once the switch from sequence to multiset has been made, there is no need for the exchangeability assumption to be given explicitly any more. On de Finetti's definition of exchangeability the set of multisets is exchangeable with the set of exchangeable sequences . . . but it is perhaps more perspicuous simply to say that every exchangeable sequence is equivalent to a multiset.[17]

The reason this counts as a payoff is that simply by talking about multisets instead of sequences we can avoid talking explicitly about exchangeability assumptions. The assumption of exchangeability will still

17. To see this, consider any sequence under the assumption of exchangeability. Permutations in the members of the sequence will not be epistemically relevant to probabilistic inferences, by definition (using any of the three definitions of exchangeability given above), and so the sequence can be replaced by the multiset containing the same members. Conversely, consider any multiset of events. It can be replaced by any sequence of events containing the same members, each with the same number of repetitions as it had in the multiset. A complication in this case is that in order to be sure that a sequence containing the same members as the multiset exists we have to assume that the members of the multiset can be put into some order, which is a non-trivial assumption if the multiset is infinite. Fortunately, events (in statistics) are observation opportunities, and multisets of observation opportunities are always orderable, e.g. spatiotemporally.

be present, but our terminology will be immensely simplified. The idealisations involved in using exchangeable events are not idealisations that we constantly need to be reminding ourselves about; consequently, the terminology of multisets will be more perspicuous, as well as more efficient, than the terminology of sequences and exchangeability.

6. MERRIMENT

Let us use the phrase **statistical measurement** to mean a report of one or more observations made of any physical circumstances (possibly very loosely defined, and possibly horribly disjunctive) considered as evidence about the hypotheses of a fixed statistical model. A statistical measurement is thus a report of the act of observing a particular physical situation, and not just a decontextualised measurement report such as "6 cm". (This distinction is useful in deflating an objection of Lane to the likelihood principle — an objection which I discuss in chapter 10.)

A statistical measurement need not be part of an experiment. I take it that an experiment is a premeditated manipulation of the world and observation of the consequences. There are many differences between experiments and non-experimental observations; the difference which will matter particularly for my purposes is that a non-experimental observation need not be considered as a sample from any particular sample space, whereas an experimental observation, at least according to the Frequentist theories which we will meet in chapter 4, is always considered to be a sample from the sample space consisting of the possible outcomes of the experiment . . . or so the purest form of the theories recommend, although as we will see this recommendation is not one which is always followed.

I will give a unified treatment of non-experimental measurements and experimental measurements; or, at least, I will recommend such a treatment. As we will see, the likelihood principle guarantees that such a treatment is possible. It is only when I discuss rival theories (such as Frequentism) and objections to the likelihood principle that I will have to depart from even-handedness between experiments and mere observations.

Bayesians occasionally mention this even-handed character of the likelihood principle when criticizing the ways in which Frequentist analyses force us to take into account the intentions which experimentalists had when collecting data (Berger & Wolpert 1984, passim). As we will see in later chapters, Frequentist analyses of experiments take into account the intentions of the data-collectors, including those of their intentions which were never actualised (Grossman et al. 1994) — a point which often amazes and confuses non-statisticians, including the medical and financial decision-makers who run clinical trials. Chapter 7 and chapter 15 discuss this issue in more detail.

In contrast to the treatment I give here, most writers on the foundations of statistics believe that statistical methods are meant to apply to experiments, by which they mean that they're meant to apply *only* to experiments. This is true on both sides of the Bayesian/non-Bayesian divide. For example, of all the versions of the likelihood principle which I have culled from the literature for chapter 8, only I. J. Good's versions are stated in a form which applies both to experiments and to observations (and perhaps also Lindley's. Lindley's is ambiguous, since he sometimes uses the word "experiment" "in a wide sense to cover cases where no planned experiment has been performed but merely some results have been observed" (Lindley

1953, p. 31)). But all of the rest of the definitions could have been stated in terms of merriments without any loss of plausibility.

Many of the most important scientific observations do not take place as parts of experiments: one need only think of astronomy to realise this. One problem with the standard experiment-driven development of the foundations of statistics is that it forces on us a strict *epistemological* distinction between an experiment and any other type of observation. For this reason alone we should start by talking about observations in general, regardless of whether the observations are performed as part of an experiment. We can still specify that we are talking about experiments, controlled experiments, ideal experiments or whatever when we actually need to. Additionally, we should try to avoid needing to. If statistics is a branch of epistemology, then the more narrowly we define its raw materials the harder it is going to be to put it together with the rest of epistemology.[18]

I will often have to talk about "experiments" rather than merriments, but only in order to accurately reproduce other people's ideas. In particular, I will have to talk about experiments a good deal in order to discuss censoring in chapter 4 and the closely related topic of stopping rules in chapters 9 to 12 and chapter 15.

---

18. It might seem to a statistician that restricting herself to experimental situations is going to give a benefit in terms of efficiency of exposition, because it will minimise the epistemological assumptions that need to be stated, since presumably the epistemological assumptions that need to be stated when discussing only experimental data are a subset of the assumptions that need to be stated when discussing observational data in general. That may be the case to some extent . . . but the hidden cost is that if she ever wants to embed her statistical theory in a general theory of epistemology then she will have to work out exactly which assumptions were minimised, and the supposed efficiency benefit of starting with a restriction to experimental data will be lost. My preference is therefore for setting off to theorise about statistics applied to observations in general, and then restricting ourselves to the subset of observations that are called "experimental results" only if we get stuck. And we won't get stuck.

The fact that the philosophy of statistics does not (I claim) need to distinguish between experiments and observation opportunities means that for maximum clarity I should refer to a statistical measurement as something more long-winded, such as "measurement from an experiment/non-experiment". For maximum clarity and minimum length (a sort of minimax procedure), I will abbreviate this to "merriment" (**me**asurement f**rr**om an exper**iment**/non-exper**iment**), defined thus:

> A **merriment** is a reasonably well-specified situation in which a doxastic agent makes and reports an observation which will be considered as evidence about the hypotheses of a fixed statistical model. If a merriment is set up by deliberate control of one or more variables which are believed to be directly relevant to the hypotheses in question then it is also known as an experiment.

7. JEFFREY CONDITIONING

I will assume throughout that we have observed something relevant to our hypotheses and that we know what it is. A clever alternative, developed by Richard Jeffrey (not to be confused with Harold Jeffreys), is to assign a probability $p(x_i)$ to the *veracity* of each possible observation x_i. On the assumption that we are performing Bayesian inference, Jeffrey then proposes that we update our probabilities using the formula

$$p(h) = \sum_i p(x_i)p(h|x_i).$$[19]

19. This formula, since it is stated in terms of $p(h)$, assumes a Bayesian ontology: non-Bayesians, by and large, do not admit that hypotheses have probabilities. Jeffrey's idea of taking into account the less-than-certain nature of observation has not yet been adapted to non-Bayesian methods of statistical inference. But I see no reason why it should not be. Non-Bayesians are opposed to giving probabilities to objective parameters in general, but I

Jeffrey's theory of probabilistic observations seems to me to be one of the cleverest contributions to epistemology since 1950. Having said that, I do not intend to adopt it in this thesis. A price I will have to pay is my assumption that we know for sure what we have observed. This does *not* mean that we have to know theory-independently what we have observed: on the contrary, I will be allowing that our observations are as theory-dependent as you like. All I will be assuming is that each observation has an agreed value according to each theory.

I do not have space to discuss in detail, with examples and mathematics, whether Jeffrey conditioning is ever strictly necessary, but I would like to suggest that it is not. For suppose that we have a set $\{x_i\}$ of putative observations, observed with probabilities $\{p(x_i)\}$, and a set of hypotheses $\{h_j\}$. Then instead of using Jeffrey conditioning we can replace the set $\{h_j\}$ with the set $\{h'_ij\}$, where each h'_ij says that hypothesis h_j is true and that observation x_i was made. It would be very time- and space-consuming to show that this gives the same results as Jeffrey conditioning, firstly because such a proof would have to be repeated for each different way of making inferences from the data (and as we will see, there are many) and secondly because Jeffrey conditioning is not even well defined for most of them, so I would have to not only apply it but also invent its method of application in most cases. But in the one case in which the application of my suggestion is well defined, Bayesian inference, it does agree completely with Jeffrey conditioning.

---

imagine that at least some of them could be convinced to give probabilities to observations. If so, it would be possible to adapt the major non-Bayesian schools of mathematics to Jeffrey's reasoning: indeed, complicating P-values and confidence intervals (defined in chapter 4) by adding a term corresponding to a previously unrecognised type of uncertainty is exactly the sort of work that mathematical statisticians love to do.

Howson and Urbach, in response to a similar claim by Skyrms (1986), complain that although this "will do the trick from the purely logical point of view, it hardly seems a solution to the problem of finding a statement describing the content of the [vague] experience which *caused* the change in belief" (Howson & Urbach 1993, pp. 105-106). It may be true that there is something pragmatically unsatisfactory about the description of the experience if my suggestion is followed, but my claim is a purely logical one: I am suggesting that Jeffrey conditionalisation can in principle be brought within my framework, and hence that the conclusions I draw in this thesis apply even when Jeffrey conditionalisation is used (at least my logical conclusions, if not my pragmatic ones). For this purpose Howson and Urbach's objection is not relevant. Be that as it may, I will play safe and retain the assumption that we know what we have observed for the rest of this thesis, to make sure that my conclusions are definitely applicable to at least the restricted domain defined by that assumption.

8. THE WORDS "BAYESIAN" AND "FREQUENTIST"

The assumption that there is a rational way to make defeasible inferences from observations to theories when part of our knowledge is probabilistic (i.e., always) is central to almost the whole of philosophy of science (pace Feyerabend 1993), much of epistemology and parts of metaphysics. In the philosophical analysis of such inferences, the battle lines have traditionally been drawn between "Bayesians" and "non-Bayesians": the former believe in the ubiquity of Bayes's Theorem in the social and historical processes guiding the scientific community's choices between competing scientific theories, and the latter don't.

A leading philosophical anti-Bayesian, Clark Glymour, wrote in his much-anthologised chapter "Why I am not a Bayesian",

> It is not that I think the Bayesian scheme or related probabilistic accounts capture nothing. On the contrary, they are clearly pertinent where the reasoning involved is explicitly statistical [but are less pertinent] so far as understanding scientific reasoning goes[.]
>
> (Glymour 1981, chapter XII)

The distinction which Glymour draws between Bayesians and non-Bayesians — the distinction in terms of which he is not one — is *not* the same as the distinction between Bayesians and non-Bayesians as drawn in terms of statistical theory. Statistical Bayesians believe in the ubiquity of Bayes's Theorem *only* in the correct application of statistical models to fully interpreted observations. In other words, statistical Bayesians believe in applying Bayes's Theorem within a set of precise alternative theories, not within any set of theories whatsoever. This is a crucial distinction.

Leaving aside the issue that the philosophical Bayesian's position is often primarily descriptive while the statistical Bayesian's position is fully normative, the main difference between the two is due to the fact that the adequacy of both statistical models and observation reports can be disputed whether or not one disputes the methods used to draw inferences from one to the other. Statistical Bayesians deal with models and observations that for one reason or another can be assumed to be unproblematic, at least tentatively and temporarily; philosophical Bayesians by and large do not. So what statistical Bayesians say, no matter how correct, barely even begins to satisfy the demands of philosophical Bayesians.

The terminological distinction that I am drawing between philosophical and statistical Bayesians is not commonly noticed; rather, the philosophical and statistical literatures have simply defined the term independently of each other. The theory's namesake, Thomas Bayes, was arguably not a Bayesian at all, so neither use of the word has clear etymological pride of place. For what it is worth, though, the earliest campaigner for either of the modern uses of the word, I. J. Good, used it primarily in the statistical sense.[20]

It is statistical Bayesianism that I will be discussing in this thesis. Within these four walls I will therefore refer to it simply as "Bayesianism". If it were necessary to pick one to be labelled Plain Vanilla Bayesianism in the world at large then perhaps that should be the statistical version too, but I leave that for others to decide.

The word "frequentist" has its problems too. It is used by most authors to denote statistical methods which evaluate results according to the frequency with which certain hypothetical long-run outcomes occur according to a "null" hypothesis. Other authors reserve the word "frequentist" for the frequency view of *probability*. The frequentist view of probability says that *probabilities* are defined as limits of sequences of long-run outcomes. There are both historical and normative links between the frequentist view of probability and the frequentist view of statistics, but neither the normative nor the historical links are straightforward. Not all supporters of frequentist statistical methods, no matter how clear-thinking, hold the frequentist view of probability, while some of the staunchest critics

20. See (Fienberg 2006) for the history of the emergence of statistical Bayesianism. Fienberg pays very little attention to the philosophical literature, even though he is in a research group with Glymour: this nicely illustrates the independent development of the two types of Bayesianism.

of frequentist statisics *do* hold the frequentist view of probability (such as A. W. F. Edwards (1972, p. xv)). A very small number of works on the foundations of statistics, notably (Seidenfeld 1979), restrict the word "frequentist" to its probability sense so clearly and explicitly that they avoid confusion, but this is at the expense of being unable to discuss the statistical literature in its own terms.

A number of alternatives to the term "frequentist" are in use. Some authors use "classical" to mean what I mean by "frequentist". But that is no better than "frequentist" from the point of view of contradicting the terminology of the philosophy of probability, because many other authors use "classical" probability to mean probability assigned according to simple symmetry principles as exemplified by pre-Bayesians such as Laplace; so "classical" in that context is close to meaning the *opposite* of frequentist. Some authors use "orthodox" to mean what I mean by frequentist; but that is no good either, partly because it implies that there is only one such position, while I discern at least two frequentist positions, and also because what is orthodox is subject to rapid change.

Yet other authors use the phrase "error rate procedures" to describe what I am calling "frequentist statistical procedures". A common rationale for this is that frequentist procedures guarantee that their results will be in error at most a certain fixed proportion of the time (see chapters 3–5 and chapter 4 for details). This rationale is mistaken, though, as the more thoughtful proponents of frequentist procedures are quick to admit (Casella & Berger 2002). The actual error rate of a procedure depends on facts which are typically unknown and which are almost never part of the statistical model even if they are known. This dependence of error

rates on unknown properties of the world is masked by the fact that the primary "error rate" of a frequentist procedure is calculated conditional on one privileged hypothesis (the "null hypothesis"): thus, it is not the expected rate of errors at all, but only the rate of errors to be expected *if* the null hypothesis is true (and, additionally, if the statistical model is totally accurate, including accounting for all possible sources of error in measurements). But we do not know whether the null hypothesis is true (if we did, we would not need to make any statistical inference), so this "error rate" is not an expected rate of errors at all.[21]

A secondary error rate which is often calculated, the "power" of a frequentist statistical test, can have either one of two meanings. It may be the *minimum*, or infimum (greatest lower bound) in the continuous case, of the proportion of errors that would be expected if any one of the non-null hypotheses were true: in other words, the minimum of a set of expected error rates, not actually an error rate itself. Alternatively, the power of a test may be the proportion of errors on the assumption that one particular alternative hypothesis is true, in which case my criticism of assuming the null hypothesis applies (except in the vanishingly rare case in which the alternative hypothesis chosen is very likely to be exactly true). Applied statisticians generally take the latter tack, and so will I.

---

21. To take a realistic example, many surgical procedures have been believed to work for hundreds of years, but there is a recent fashion for evaluating them statistically just in case their apparent effectiveness is illusory (Grossman & Mackenzie 2005). The statistical evaluations which are used are themselves evaluated by calculating an error rate conditional on the null hypothesis (in other words, assuming the null hypothesis to be true), and the null hypothesis (for legal and ethical reasons) is always that the surgical procedure is *completely* without effect. As a result, epidemiologists often find themselves evaluating surgical procedures which we are practically certain have *some* effect using statistical procedures chosen on the basis of characteristics which they only have on the assumption that the surgical procedures have *no* effect.

So "error rate" procedures do not give us expected error rates; nor do they give us actual error rates. So the term really is a misnomer.

For this reason, I believe it is important to continue to use "frequentist" rather than "error rate" to describe such procedures, even though taken out of context the word is ambiguous. In any case, I am working on the foundations of statistics, and the literature on this topic mostly uses the word "frequentist" to refer to statistical procedures based on error rates, so it is safe for me do the same. I will capitalise it from now on to emphasise that I am using it in a way which some may find idiosyncratic.

While discussing terminology, I have already produced an argument against the use of Frequentist procedures, namely that we tend to think they have guaranteed error characteristics when in fact they do not. But there are much more substantive arguments to come — arguments which do not depend on whether we are misled by terminology. I will lay the groundwork for these in chapter 4 and give some of them in detail in chapter 7.

9. OTHER PRELIMINARY CONSIDERATIONS

My loose talk about inferences from observations to hypotheses may have ignited a worry about the theory-dependence of observation. Happily, the theory-dependence of observation, although real, is not a problem I need to take into account explicitly. The nature of the epistemic framework within which statistical inference takes place is that we are interested in the observation as interpreted by the various hypotheses. That is all we can possibly mean when we say that each hypothesis assigns a precise probability to the observation. So either we implicitly but fully embrace

any theory-dependence of observation that may crop up, or we give up on doing statistical inference.

I assume a classical logic, although I will note a point in chapter 13 in which to do otherwise would make an interesting difference to my conclusions.

In the statistical literature, statistical inference is claimed to license partial beliefs or credences (according to most members of the Bayesian school) or beliefs (according to practically all authors before 1920) or actions (according to the Neyman-Pearson school). In the rest of this thesis, I will present results which are very nearly neutral between these options.

Even when I discuss credences in detail, I will not be dealing with all of the tricky questions about belief which are important to the philosophy of mind. In particular, I will avoid worrying about what sorts of systems can have beliefs. It will be enough if I can say something useful about the probabilistic beliefs of adult humans in a numerate Western culture. But there is no need to assume that what I have to say is only applicable to humans, and so instead of calling my protagonists "humans" or "people" I will call them "doxastic agents" (agents with partial, or probabilistic, beliefs). I will also sometimes call them "epistemic agents", to fit the language in which some of the issues I discuss are traditionally debated; in particular, from chapter 4 onwards I will talk a lot about "epistemic probabilities", which are never called "doxastic probabilities", and then I will use "epistemic agent" to match. In all cases I mean "epistemic agent" and "doxastic agent" to be semantically interchangeable.

— 3 —
Survey I: Bayesianism

1. INTRODUCTION

The next three chapters survey the existing normative theories of inference from data to probabilistic hypotheses. While the main purpose of these chapters is simply to expound what theories there are, a subsidiary purpose is to show that not all of the possible theories of statistical inference have yet been stated — not even all the possible theories which use the framework set out in chapter 2. (This is, unsurprisingly, easy to show; see the section *Other pure likelihood methods* in chapter 5 for an example.)

Surprisingly, there are only five such theories by a rough classification[22] and only fifteen even by a more precise classification[23]. Classifications vary, since one man's theory is another man's amendment, but by any classification there are very few such theories . . . indeed, classifications in the literature often stop at two, typically Subjective Bayesianism and Neyman-Pearson Frequentism. If any other theories have ever been invented, they have gone unnoticed, and not just by me.

---

22. Bayesianism, Frequentism, pure likelihood methods, pivotal inference and plausibility inference.

23. Subjective Bayesianism, Restricted Bayesianism, Empirical Bayesianism, conjugate ignorance prior Bayesianism, Robust Bayesianism, Objective Subjective Bayesianism, Neyman-Pearson Frequentism, Fisher Frequentism, Structural Inference, the method of maximum likelihood, the method of support, fiducial inference, pivotal inference, plausibility inference and, arguably, Shafer belief functions — all defined in the following three chapters.

There are of course many more than fifteen theories of *probability*, and many theories concerning the mathematics of statistical distributions and error rates. When I say there are only fifteen theories of statistical inference, I mean that there are only fifteen distinct, more or less complete answers to the following question:

> Given that a doxastic agent is considering some precise, probabilistic scientific hypotheses and comes into the posession of some relevant scientific data, how should she alter the beliefs or actions which follow from taking one or another of the hypotheses as true, assuming that she has enough time, patience and computing power to be fully rational?

Each of the fifteen theories is an answer to this guiding question.

Tell a layperson that the answer to this question is contentious — that there is more than *one* theory — and he will be surprised. And yet there is no consensus on the answer either in philosophy or in theoretical statistics; and the current consensus in applied statistics is a bad one, as we will see in chapter 7.

The meaning of each of the theories which I will survey is well operationalised and therefore relatively uncontroversial (by the standards of philosophical theories). What's controversial is which theory is right, if any. Perhaps none of these theories is right. Although it seems to me to be a requirement of the possibility of rationality that there must be some optimal theory of statistical inference, the search for such theories is only a little over a hundred years old, and (as I will show) all of the existing theories have prima facie flaws, so it is plausible that we have not yet come up with a good theory. For this reason among others, the overall

conclusion of this thesis will be not that we ought to use one particular theory, but only that we ought to obey principles which rule out some of the competing theories.

I will assume for the purposes of this survey that a full utility function for all members of H is not available (or is available but disputed). I make this assumption only to save space. If a full utility function were available we would be in the realms of statistical decision theory. The various precise statistical decision theories which have been proposed each imply a certain precise theory of statistical inference, and as it happens (perhaps for historical reasons) all of these imply either one of the Bayesian theories detailed in this chapter (for maximum expected utility decision theories) or one of the Frequentist theories (for Wald-style minimax decision theories). Hence the range of theories of statistical inference would not be altered if I took decision theory into account. A *full* account of statistical inference would certainly include some discussion of decision theory; but to include that discussion here would use considerable space at the expense of clarity.[24]

I included a caveat about time, patience and computing power in my guiding question in order to sidestep the recently developed theory of *bounded rationality*, a theory which studies the consequences of the fact that real-world epistemic agents have to cut short the computations they might have liked to perform in order to (for example) get out of the way

24. To take a random example of how clarity would suffer: a statistical procedure which does not lead to incoherence under maximum expected utility theory is known as "admissible", and it can be shown that only Bayesian procedures are admissible; to define the terminology required to describe this result precisely would take some time, but would not give us any new methods of statistical inference; nor would it help us to see which methods are preferable, unless I could also give a convincing defence of the universal applicability of maximum expected utility theory, which in turn would require assumptions which would distract from the main thrust of my work. In later chapters I will show that the likelihood principle can be defended without requiring any assumptions about utility.

of sabre-toothed tigers, or get a company report in on time (Simon 1982, Gigerenzer et al. 1999, Gigerenzer & Selten 2002).

We do not need to worry that bounded rationality might have supplanted the approach I take in this thesis, for at least three reasons.

Firstly, I am dealing primarily with *scientific* inference, in which it is possible to spend a great deal of time and computational power on statistical inference. The issues that crop up in the bounded rationality literature are primarily (although admittedly not exclusively) about the constraints involved in cases in which an epistemic agent has *extremely* little time available to make a decision, and the evidence taken to support the importance of bounded rationality is evidence of single biological organisms solving personal decision problems.

Secondly, I bite a bullet and admit that the type of rationality investigated in this thesis is an idealisation.

Thirdly, and most importantly, the bounded rationality literature is almost entirely descriptive: it consists almost entirely of descriptions of how epistemic agents actually behave, and of psychological and philosophical consequences of those descriptions; it is therefore no threat to my views on how inferences *ought* to be made.

In response to my third point, the bounded rationality literature is sometimes taken to be normative as well as descriptive. It is hard to be absolutely sure whether this reading is right or wrong, but I think it is wrong. Consider the following quotations from a representative paper in the bounded rationality literature. On the one hand, the bounded rationality position is set up in clear opposition to both the statistical view

and Kahneman and Tversky's "heuristics and biases" view (which describes actual departures from a probabilistic norm). It describes these views thus:

> [A] discrepancy between the dictates of classical rationality and actual reasoning is what defines a reasoning error in this program. Both views accept the laws of probability and statistics as normative, but they disagree about whether humans can stand up to these norms.
>
> (Gigerenzer & Goldstein 1996, p. 650)

. . . and *criticises* such views for

> lead[ing] us to believe that humans are hopelessly lost in the face of real-world complexity[.]
>
> (Gigerenzer & Goldstein 1996, p. 651)

Passages such as this might be (and often are) read as saying that the normativity of the old program is wrong and is to be replaced by a new normativity, that of bounded rationality. On the other hand, and on the very same page, Gigerenzer and Todd say:

> [bounded rationality] algorithms are designed to be fast and frugal without a significant loss of inferential accuracy[.]
>
> (Gigerenzer & Goldstein 1996, p. 651)

This implies that there is some *other* standard of inferential accuracy, apart from bounded rationality, which bounded rationality approximates to. It is implicit in the experimental work of the bounded rationality school that there is such a standard and that that standard of inferential accuracy is Bayesian — see (Gigerenzer & Goldstein 1996) again and, especially, (Gigerenzer et al. 1999).

Hence, the normativity of the statistical view of inference is both denied and taken for granted. It seems most likely to me that the authors of this literature are not in the least confused, and that they regard their view as *not* normative, but find it hard to say so because the normative character of the view they oppose has lent it stature. Be that as it may, as long as bounded rationality measures its correctness on the basis of its approximate *agreement* with inferential statistics it will be impossible for it to be any better justified than inferential statistics is, and hence it poses no normative threat.

So, a fully normative version of bounded rationality is no better justified than inferential statistics is, at present; but maybe it is plausible that it will be better justified in the future.

Does this mean that the conclusions I draw here are hostage to the possibility that bounded rationality gives us the correct description of our epistemic constraints? Not at all, for two reasons:

(1) Of course we are all bound by computational constraints, just as bounded rationality supposes. But in a world of large scientific research budgets and fast computers, what happens when we try to estimate the size of those constraints? It may well be that *zero* is a better estimate of our constraints than any particular finite number. In that case, there is nothing particularly inaccurate about my theory at all (even though it is not perfectly accurate), because any other estimate of what our constraints are would also be inaccurate to some degree. In other words, unbounded rationality may be the most rational estimate of what type of bounded rationality we employ in science.

(2) My discussion of unbounded rationality is a limiting case of bounded rationality: it is what bounded rationality becomes as the bound tends to infinity. Even if normative bounded rationality turns out to be what we will all end up studying one day, it will be useful (and, at a guess, probably essential) to have a well worked out theory of what happens in the limit.

I see these as decisive reasons to ignore the fact that the gurus of bounded rationality ask us to reject "the laws of probability and statistics as normative" (Gigerenzer et al. 1999).

For each theory described in the next three chapters, I will give its main originators, a summary of its tenets and a brief comment. I will not attempt to give a full justification of any of these theories, because I am not sure that any of them is right, and because for the purposes of defending the likelihood principle it does not matter if they are all bad theories. The only thing I need to show conclusively is that the theories which conflict with the likelihood principle are bad. In later chapters I will discuss foundational issues which reflect (badly) on the Frequentist methods and (well) on methods compatible with the likelihood principle.[25]

2. BAYESIANISM IN GENERAL

The Bayesian position emerged extremely gradually from the work of

25. Methods compatible with the likelihood principle are given by: the pure likelihood theory, the maximum likelihood theory, and all the Bayesian theories with minor exceptions such as some ill-named versions of the Empirical Bayesian theory.

a number of 18th-century authors including Thomas Bayes (1763).[26] Bayesianism was not formulated clearly until around 1920.[27]

Bayes's Theorem says:

$$p(a|b) = \frac{p(b|a)\,p(a)}{p(b)}, \text{ provided } b \neq 0.$$

Bayes's Theorem (unlike Bayesianism) is uncontroversial. Whenever $p(a)$, $p(b)$, $p(a|b)$ and $p(b|a)$ all exist and $p(b)$ is non-zero, Bayes's Theorem can be proved from any of the standard axioms of probability. My favourite proof is as follows:

$$\begin{aligned} p(a\&b) &= p(a|b)\,p(b) \\ &= p(b|a)\,p(a) \text{ by symmetry of "\&".} \end{aligned}$$

$$\text{Rearranging, } p(a|b) = \frac{p(b|a)\,p(a)}{p(b)}, \text{ provided } b \neq 0.$$

For example, what is the probability of drawing the Ace of Hearts, given that I've drawn an unknown Ace from a pack of cards? Letting a stand for

---

26. Bayes's own work, although pioneering, did not clearly state either its epistemic or its mathematical assumptions, and so it is arguable that Bayes did not found Bayesianism, despite the popularisation of Bayes's work by his contemporary Price, who immediately saw that it would be a cornerstone of future attempts to quantify the scientific practice of induction (Bayes 1763, p. 371).

27. Bayesianism probably developed from unpublished work of Johnson (Jeffreys 1961, p. i), after which it was quickly developed into a mature theory. Wrinch and Jeffreys have some claim to having invented modern Bayesianism in their (1919), but their work did not at first attract the attention it deserved — perhaps because it was ahead of its time, or perhaps because its objectivist foundations (described below) failed to distinguish it sufficiently from other statistical methods available at the time.

The modern meaning of the term "Bayesian" is probably due to I. J. Good (Smith 1995). A thorough overview of the history of Bayesian inference has yet to be written, and many questions about the relationships of early Bayesian views to each other are not settled, although some aspects of its separate development into two non-equivalent systems by Jeffreys and de Finetti are well documented. From the 1950s onwards a large number of authors produced systems which extended on one of those two early systems or which proposed various compromises between them.

the event of drawing a Heart and b stand for the event of drawing an Ace, we can use Bayes's Theorem to confirm that the answer is $\frac{1/13 \times 1/4}{1/13} = \frac{1}{4}$.

Bayesians believe that the quantities mentioned in Bayes's Theorem always exist (or, in de Finetti's system, can always be treated as if they exist), for any stateable a and b. Hence, the probability of any hypothesis conditional on any set of observations can be calculated, by setting $a = h$ for any $h \in H$ and $b = x_a$ (in the terminology of chapter 2), so that

$$\forall h \in H, \quad p(h|x_a) = \frac{p(x_a|h)\, p(h)}{p(x_a)}, \text{ provided } p(x_a) \neq 0.$$

(Henceforth, I take the proviso that the denominator of the right-hand side not be zero as implicit.)

The term $p(h)$, considered either as a single probability or as a function over the hypothesis space (H), is known as a *prior probability* or *prior probability distribution* respectively. The single word "prior" is often used ambiguously (or, rather, polymorphically) to refer to either a prior probability or a prior distribution. The word "prior" need not carry any temporal weight: in some versions of Bayesian theory, the prior is meant to be known before x_a is known, but in most versions this is not required. Many writers have bemoaned the choice of the word "prior". A word with no temporal connotations, such as "ulterior", would be better, but unfortunately it seems too late to change this.

The term $p(h|x_a)$, considered either as a single probability or as a function over H, is known as a *posterior probability* or *posterior probability distribution* respectively. (Just as the word "prior" need not carry any temporal weight, nor need the word "posterior".)

The term $p(x_a|h)$, considered as a function over a hypothesis space, is the *likelihood function* which I introduced in chapter 1. Likelihood functions are discussed in detail in chapter 8 and chapter 13.

The denominator of the above equation, $p(x_a)$, does not need to be calculated, provided that H exhausts the hypotheses considered possible in the merriment in question. (Recall from chapter 2 that a merriment is an experiment or a non-experimental observation.) This is because $p(x_a)$, unlike $p(x_a|h)$, does not depend on h, so that instead of using Bayes's Theorem in full for inferences about h one can use

$$\forall h \in H, \quad p(h|x_a) \propto p(x_a|h)p(h)$$

instead; and if the constant of proportionality is needed one can calculate it simply by dividing by a factor which makes the function $p(x_a|h)p(h)$ add up to 1.[28]

Even better, one can get rid of the normalising factor completely when comparing two hypotheses, h_1 and h_2, by writing

$$p(h_1|x_a) = \frac{p(x_a|h_1)\, p(h_1)}{p(x_a)}$$

and

$$p(h_2|x_a) = \frac{p(x_a|h_2)\, p(h_2)}{p(x_a)}$$

and then dividing the two equations to get

28. Of course, this factor is always $\sum_{h\in H} p(x_a|h)p(h)$ if H is finite and $\int_{h\in H} p(x_a|h)p(h)$ otherwise.

$$\frac{p(h_1|x_a)}{p(h_2|x_a)} = \frac{p(x_a|h_1)}{p(x_a|h_2)} \frac{p(h_1)}{p(h_2)}$$

with no mention of $p(x_a)$ on the right-hand side. ($p(x_a|h_i)$, in contrast to $p(x_a)$, is always known: that it is known is a corollary of the fact that h_i is fully specified.) This method of getting rid of $p(x_a)$ has been known since at least the 1920s by followers of Jeffreys and de Finetti, and was independently discovered by Salmon in 1996 in an important paper in which he replies to criticisms of Bayesianism by Glymour (Salmon 1996).

Thus, Bayesianism allows us to compare simple hypotheses in the light of data in an appealingly straightforward way.

Bayesianism is, I think, unavoidable for those who believe that uncertainty is always best described by the use of probabilities. Lindley (1990, p. 214), for example, has described Bayesianism as the natural result of "adopting probability as *the* language of science (unlike a classical statistician who only uses it as part of the language, denying its validity for hypotheses or parameters)". Whether or not one accepts Lindley's view, it is hard to deny that probabilities are the only way to quantify epistemic uncertainty[29] in many scientific situations, and this leads to the natural use of Bayesianism in at least those cases, provided its drawbacks can be swallowed.

The only philosophical drawback of Bayesian methods — but it's a biggie — is that many authors dispute the existence of the term $p(h)$ for interesting hypotheses h, while some dispute its existence for *any* hypothesis h. Until recently there were also computational drawbacks

29. Thanks to Mark Colyvan for pointing out that this point does not apply to the *semantic* uncertainty introduced by vague and ambiguous language.

to Bayesianism, but nowadays the necessary calculations can be done by brute force on a cheap computer in most cases.

Recall that $p(h|x_a)$ is known as the "posterior probability" of the hypothesis "on" or "given" the observation: "posterior" because it is calculated after the observation has been made (except in cases in which the whole analysis is hypothetical, in which case x_a will not be an actual observation, but will instead be a possible observation treated as actual within a fictional story). According to the Bayesian view of statistical inference, the set of posterior probabilities for all hypotheses of interest gives us the probability that each hypothesis is true. (This is the probability relative to the inference situation; for the Bayesian there is no such thing as probability simpliciter, unless it is merely convenient shorthand for one or another type of indexical or conditional probability — one example of such a scheme of convenient shorthand is given in chapter 2.) The posterior probability, according to Bayesians, therefore tells us everything we can possibly learn from an inference situation (conditional on the model being reasonable). One might want to add as a rider to this that sometimes one can draw *no* conclusions in a given situation: this gives a form of Restricted Bayesianism (see below).

While setting up an appropriate mathematical model of a merriment is roughly as difficult for a Bayesian as it is for anyone else, the calculation of posterior probabilities given a model is easy. There are no ad hoc or debatable steps in this procedure (although there are many ad hockeries typically involved in choosing the mathematical model in the first place . . . and fortunately so, or Bayesian mathematicians would have nothing to write research papers on).

The fact that posterior probabilities can virtually always be calculated gives Bayesians a virtually complete theory of inference — much more complete, indeed, than any of the competing theories (except, arguably, for the Neyman-Pearson theory, the only non-Bayesian theory which is ever claimed to apply to absolutely *all* statistical inferences).

To summarise, Bayesian statistical inference is based on two premises:

(i) Bayes's Theorem is applicable to any statement of conditional probability, provided that the relevant probabilities are defined. (We will see below that, according to some but not all versions of Bayesian inference, all relevant probabilities are always defined.)

(ii) The set of posterior probabilities tells us everything we can know about uncertain propositions, given the mathematical model and observations available to a particular doxastic agent at a particular time.

BAYESIAN CONFIRMATION THEORY

Two recent schools of thought raise questions about the standard terminology of Bayesianism.

Firstly, a number of philosophers, relatively recently, have proposed a school of inference called "Bayesian confirmation theory" which we need to distinguish from Bayesianism simpliciter. According to these recent Bayesian confirmation theorists, there is some function of the data which tells us to what extent data confirm a hypothesis and, crucially, this "confirmation" function need not be (and often is not) a function of the posterior probability distribution. Steel (2003) has recently shown that some such functions are incompatible with Bayesianism as defined above. Since Bayesian confirmation theorists see a need for a theory of confirmation

not based on the posterior distribution, they are generally not Bayesians according to the bulk of the literature on Bayesianism, despite their name.[30]

Most Bayesians see no need for a single confirmation function separate from the posterior probability function. These Bayesians — for example, Howson and Urbach (1993, pp. 117–118) — note that if the posterior $p(h|e)$ is greater than $p(h)$ then e confirms h, but they do not claim to have told us to what numerical *extent* e confirms h. The basis of Bayesianism is that the posterior probabilities tell us the probabilities of all the hypotheses in H conditional on the observation. If we like we can compare these probabilities to the probabilities of the same hypotheses before we conditioned on the observation (in other words, we can compare the posterior distribution to the prior distribution). There is more than one way to compare these two functions — one can subtract them, divide them, perhaps take more complicated functions of them — but the question of how one should compare them has not seemed a very interesting question to most Bayesians who, after all, already hold, in the posterior distribution, the answer to much more interesting questions. Bayesians say, in effect (and sometimes in actuality), "Look: here's exactly, quantifiably, what we thought before we saw the data; here's exactly what we thought after we saw the data; now are you telling us we haven't shown you the effect of the data on our beliefs?"[31]

---

30. To make things even more confusing, they *may* be Bayesians on anybody's definition, if their beliefs about confirmation functions happen not to contradict orthodox Bayesianism. And as if that wasn't confusing enough, some authors use "Bayesian confirmation theory" to mean simply Bayesianism, with no mention of confirmation functions at all (Strevins 2004).

31. See (Hawthorne 2005) for a counter-argument to this position. I do not wish to give space to counter- or counter-counter-arguments, since the question at issue is tangential to the main work of this thesis.

There is a small literature on confirmation functions written by orthodox Bayesians. These works argue that the confirmation function obtained by dividing the posterior by the prior (resulting in the likelihood ratio — essentially the same quantity as is foregrounded by the likelihood principle) is the most desirable. I will discuss the rationale for this while discussing Barnard's views in chapter 8.[32] Orthodox Bayesians, regardless of whether they accept this argument or not, rarely give it any importance. Consequently, the school of thought that says that Bayesianism as formulated by premises (i) and (ii) above is already complete is proceeding almost independently of the school of thought that says that Bayesianism needs to be supplemented by a confirmation function.

There is no good reason for this schism in the use of the term "Bayesian".[33] One resolution of the problem would be to use the words as they are currently used by both schools of thought, on the clear understanding that "Bayesian confirmation theory" is not always Bayesian; but I believe this resolution is not feasible. We can no more expect people to bear in mind that "Bayesian confirmation theory" is not always Bayesian than we can expect people to remember that George W. Bush's "environmental" legislation is not environmental. Instead, it would be best if the "Bayesian confirmation theorists" would drop the tag "Bayesian". I would like to emphasise that to say that "Bayesian confirmation theory" is misnamed, as I do, is not to disparage it; it is only to wish on it a separate existence.

---

32. Good, who is rather proud of being prolific, counts 33 publications in which he has made this point, and says that "[w]hat I say thirty-three times is true" (Good 1983, p. 159).

33. Indeed, such a schism, to the extent that it exists, is counterproductive, because it confuses our reading of the literature. Worse, it *retrospectively* confuses our reading of the pre-confirmation-theory Bayesian literature.

I propose to restrict the use of the word "Bayesian" in this thesis to its only unconfusing meaning, which is the one given to it by the founders of Bayesian theory and the *vast* majority of its practitioners: namely, the meaning given by the two premises above.

Secondly, Bayesianism, as defined above, is about the inferences available to a single doxastic agent. Since some schools of Bayesians (the Subjective Bayesians, as defined below) see these inferences as depending on subjective judgements, it is not clear that there is any reason for even two doxastic agents to agree about inferences, never mind for a whole scientific community to agree. This has always been seen as a problem (Lindley 1980), and it is beginning to be addressed in detail (Kadane et al. 1999). Some, in a pragmatic frame of mind, address it by giving reasons to think that agents' subjective views in fact agree closely enough that their scientific inferences will be effectively the same — J. O. Berger has a sustained research program showing from a mathematical point of view that this will often be the case, and Freedman and Spiegelhalter among others have demonstrated it very successfully in practice with oncologists. These pragmatists are clearly right about at least *some* situations, since whenever a large amount of data is available (large relative to whatever initial disagreements the doxastic agents have) initial disagreements will be swamped by the data in the sense that for any fixed hypothesis, any fixed initial level of disagreement about that hypothesis and any fixed small number ε, there is an amount of data that will cause all the epstemic

agents to agree about the probability of the hypothesis to within ε. This theorem is an easy consequence of the Bayesian premises.[34]

But the pragmatic school of thought which says that all is well is seen as wishful thinking by others, who have proved theorems showing that in some circumstances it is *impossible* for a group of doxastic agents to reach joint inferential conclusions (Kadane et al. 1999), either because insufficient data is available or because it is necessary to take into account the doxastic agents' utilities as well as their probabilities (or, in old-school language, their desires as well as their beliefs). If this second group of theorists is right (and I am afraid it is), we need a new theory which tells us how to make joint inferences when individual judgements vary in a way which makes strict Bayesian methods ineffective. This new theory is likely to be based on Bayesian theory, with the traditional Bayesian theory as a limiting case in the large-data, low-disagreement case. The new theory may or may not be called "Bayesian". (Seidenfeld, for one, is agnostic about whether it should be.) I will not be discussing joint decision-making in this thesis, and I will be consistently agnostic about Bayesianism, so I do not urgently need to decide these questions.

3. SUBJECTIVE BAYESIANISM

All forms of Subjective Bayesianism hold that the Bayesian equations given above are applicable to all cases in which a doxastic agent is uncertain about what inferences to make from an observation. To the best of my

34. It may be worth noting the order of the main quantifiers here: $\forall$ required levels of agreement, $\exists$ an amount of data which will produce such agreement. It would be nice if $\forall$ amounts of data greater than some value, $\neg\exists$ a level of disagreement too great to withstand the irenic power of the data; but sadly this is not so.

knowledge, all Subjective Bayesians agree that may be very little objective evidence on which to base a prior probability distribution. It follows from these two points that all Subjective Bayesians hold that there is a non-objective component to any fully specified prior probability distribution. This non-objective component may be subjective or intersubjective. If it is subjective then it depends on an individual's cognitive state in a way which cannot be fully justified by that individual to another rational doxastic agent; if it is intersubjective then it depends on a position jointly reached by a community of doxastic agents which cannot be fully justified by that community to an external rational doxastic agent. The notion of justification in play may be vague or strict. If strict, then there is a fundamental epistemic difference between Subjective Bayesians and Objective Bayesians. If vague, then there is no such fundamental difference, but there remains a major methodological difference, for very few prior probability distributions are actually held to be objective by communities of scientists; and hence, in the absence of practical, comprehensive ways of determining whether a prior probability distribution *ought* to be accepted by a community, there remains a great divide between those who think that statistical analysis need not wait on such agreement (subjectivists) and those who think it must wait. In summary, Subjective Bayesians analyse situations in which there is incomplete agreement about prior probability distributions by using an individual's or community's choice of prior, however established, while Objective Bayesians analyse such situations by waiting for agreement or (as we shall see below) by proposing general methods of producing priors which can force such agreement.

There are no differences between the various schools of Subjective Bayesianism that matter for the limited purposes of this thesis (although of course there are many differences which are interesting in their own right), except for disagreements between individual theorists about the definition and role of the likelihood principle, which I will discuss in detail in chapter 8 and chapters 9 to 12.

THE UNIQUENESS PROPERTY OF SUBJECTIVE BAYESIANISM

A number of authors have produced sets of axioms under which Subjective Bayesianism can be proved to be the unique rational way to conduct statistical inference. The founder of this school of thought — and, indeed, the founder of the modern school of subjective probability — was Keynes (1921).[35] Savage (1954), extending the work of Keynes (1921) and Ramsey (1978), laid the foundations for modern statistical decision theory (perhaps best exemplified by Raiffa & Schlaifer 2000) by showing that if both precise but subjective prior probability distributions and precise utility functions are assumed known for a given doxastic agent then natural axioms of rationality and mathematics require that agent to be a Subjective Bayesian. Philosophical objections to this claim have concentrated on the existence of prior probability and utility functions. Since my goal is not to defend Bayesianism, I will not attempt to show that Savage's axioms are reasonable in general. However, there are some *sub*-domains of statistical inference in which both exact prior probability distributions and exact utilities are

35. It is sometimes said that Keynes believed only in objective or logical probability, but this is not the case. As he summarises his own work, "The method of this treatise has been to regard subjective probabliity as fundamental and to treat all other relevant conceptions as derivative from this" (1921, p. 282). He may not have intended "subjective" in the same way as modern subjectivists, who mostly follow de Finetti (de Finetti 1972), but he certainly did not intend "subjective" to mean either objective or logical.

uncontentiously available (for example, statistical puzzles in which priors can be set by a godlike adjudicator and utilities can be set by fiat), and in these sub-domains Savage's theory is uncontentious, at least for a single doxastic agent.

4. OBJECTIVE BAYESIANISM

Objective Bayesianism includes any school of thought which holds that Bayes's Theorem can be used ubiquitously or almost ubiquitously but which, unlike Subjective Bayesianism, does not call for any subjective prior probability distributions. There is an important ambiguity here: an Objective Bayesian theory may hold that subjective priors are *never* required (equivalently, may silently render them unnecessary), or it may hold that subjective priors are *sometimes* not required. Both of these cases contrast with Subjective Bayesianism, which holds that subjective priors are *always* required. Of the theories I discuss in this section, Restricted Bayesianism and Empirical Bayesianism are of the former type (subjective priors never required), while Jeffreys's and Jaynes's theories are of the latter type (subjective priors sometimes not required).

Objective Bayesian methods are thus more objective than Subjective Bayesian methods, but apart from that I make no positive claims for their objectivity. They do not have all the features which some might think are necessary to a completely objective system.[36]

---

36. To take a trivial example of such a feature, all the theories in this whole thesis make reference to mind-dependent entities in the form of statements of hypotheses. A less trivial example, but perhaps also less of a barrier to objectivity, is that none of the theories in this section can be stated without epistemic probabilities — as opposed, especially, to Neyman's theory in which, as we will see later, the probabilities are remarkably non-epistemic.

RESTRICTED BAYESIANISM

Restricted Bayesianism is my own term for the type of Objective Bayesianism which holds that we should use Bayesian methods only when a prior probability distribution has been given to us by an objective process separate from the merriment we are analysing. Perhaps surprisingly, there are many applications for it. For example, almost the whole of clinical epidemiology (the study of the determinants of medical success and failure) can use Restricted Bayesian techniques, because an objective prior probability distribution is given by the rates to date (not counting the merriment at hand) with which clinicians have achieved certain medical outcomes with patients in a certain epistemic equivalence class (patients with certain medically relevant characteristics).[37]

Restricted Bayesianism is open to the objection that determining the relevant equivalence class is often essentially ad hoc. (This is known as "the problem of the reference class" in the literature on probabilistic inference (Hájek 2003).) This objection is well-founded, although there is some disagreement on this point in the literature; moreover, even those who agree with the objection are often able to agree that a particular

37. Consider, for example, a doctor who wants to know the probability that a patient who has a positive test result actually has the disease that the test tests for. (This probability is usually much less than one, and indeed often less than ½, so a patient with a positive test result is often most likely *not* to have the disease. The low typical value of this probability causes many misunderstandings of test results, but these need not concern us here.) Let a be the event of having the disease and b be the event of receiving a positive test result. The quantity $p(b|a)$ is a property of the test, and is provided by the company which markets the test kit; it is usually approximately independent of the characteristics of the individual patient, and can therefore be established once and for all. The prior, $p(a)$, is harder to come by but nevertheless is usually claimed to be objective: it can be calculated (according to clinical epidemiologists) by finding out what proportion of the population of the geographical area in which the patient lives has the disease. This information is often readily available. $p(b)$ can be calculated as a normalisation factor, as described above. It is then a simple matter to use Bayes's Theorem to calculate the required probability that the patient with a positive test result has the disease, $p(a|b)$. Were all of the uses of Bayes's Theorem of this type, very few of the objections to Bayesianism which I mention above would crop up.

application of Restricted Bayesianism is no more ad hoc than any of its rival theories would be in the same situation.

Restricted Bayesianism is also open to the objection from subjectivist Bayesians that it only allows *some* epistemically relevant types of knowledge to affect the prior probabilities, not *all* epistemically relevant types (some of which may be subjective).

Despite these criticisms, Restricted Bayesianism has very few active opponents. Surprisingly (in my view), it also has very few *pro*ponents: even in areas in which its worth is undisputed it is rarely applied. It is because it has been so rarely acknowledged that I have had to invent a name for it. Its lack of generality reduces its philosophical interest somewhat, but it ought to be of very great scientific interest indeed. A recent paper by Daniel Goodman (2004) promoting this method presages such an interest, I hope.

The main barrier to the use of Restricted Bayesianism in a wide class of problems (apart from social inertia) is that the relevant equivalence class, even when it exists and is objective, may not be large enough to provide prior believable probabilities.

EMPIRICAL BAYESIANISM

Empirical Bayesian methods are radically different from other Objective Bayesian methods; indeed, although they are Bayesian in the letter of the law (at least according to my definition, although not according to the definition of Deely and Lindley (1981)) they are very far from its epistemic spirit. They "estimate" (a weasel word in this context) any probabilities which are unknown — especially prior probabilities — from the very

same observational data that are to be used for inference to hypotheses (Breslow 1990, Morris 1983, Lindley 1983). They are importantly different from Restricted Bayesianism in that Restricted Bayesianism uses prior probability distributions which are antecedently (or at least independently) justified.

I have not been able to find any theoretical justification for Empirical Bayesianism in the literature. According to standard Bayesian theory it is clearly unjustifiable. All justifications of the use of Bayes's Theorem assume that the likelihood function which is used to update the prior distribution of probabilities contains entirely new information. The numerical degree of updating recommended by Bayes's Theorem depends on this assumption. Empirical Bayesianism violates the assumption by using the *same* data to construct probabilities *and* to perform updating of probabilities. This is similar in intent, and often similar in effect, to considering the very same set of data twice as if it had been collected on two separate occasions. Empirical Bayesianism, by using the same set of data to set the prior as it uses to set the likelihood function, gives the data a demonstrably, and quantifiably, larger role in the analysis than it should have according to Bayes's Theorem. It is quantifiably illicit double-dipping (Deely & Lindley 1981).

Empirical Bayesianism is also disreputable according to the Subjective Bayesian school of thought for an additional reason, namely that, unlike Subjective Bayesianism (and also unlike Jeffreys's and Jaynes's schools, and unlike Pivotal Inference for that matter), it does not enable probabilities which are not "estimated" from the data to be taken into account at all, even when they are probabilities agreed by a whole community.

Despite these immense philosophical drawbacks, the method is popular among scientists (see, for example, Bernardinelli & Montomoli 1992), because it is extremely objective: not only does it not require any subjective judgements of probability, it also does not require the use of Jeffreys/Jaynes ignorance priors, to which I now turn.

CONJUGATE IGNORANCE PRIORS I: JEFFREYS

Another strand of Bayesianism is similar to Restricted Bayesianism in using frequency information to construct a prior probability distribution when such information is available, but otherwise uses a prior probability distribution constructed using principles of symmetry of an abstract sort. A theory of this type was historically the first type of Bayesianism, and arguably the first statistical theory of any sort, to be given a reasonably comprehensive treatment covering its philosophy, its mathematics and some of its practicalities (Jeffreys 1931). So far, two theories of this type have been developed, giving alternative views of the principles of symmetry responsible for determining the prior: one is mainly due to Jeffreys and one mainly due to Jaynes.

Both Jeffreys and Jaynes present their principles for constructing prior distributions as ways of *representing ignorance* about H; this way of thinking about their theories makes it clear that they are compatible with Restricted Bayesianism (compatible in the sense that the methods will agree whenever frequency information suitable for constructing a prior probability distribution is available and uncontentious). It is also, roughly speaking, the traditional approach to statistical inference, discussed by Bayes, Laplace and their followers in the 18th Century. Thus, for example,

Jeffreys's principle for a *finite* set of hypotheses is the same as Laplace's, namely to give each member of the set the same probability:

> If there is no reason to believe one hypothesis rather than another, the [prior] probabilities are equal.
>
> (Jeffreys 1961, p. 33)

. . . and Jeffreys is explicit in saying that this is a way of representing ignorance:

> The rule that we should take them equal is not a statement of any belief about the actual composition of the world, nor is it an inference from previous experience; it is merely the formal way of expressing ignorance.
>
> (Jeffreys 1961, pp. 33–34)

However, as we will see below, the literature on Subjective Bayesianism contains criticisms of the idea that we can be ignorant about a set of hypotheses. Since Bayesianism is not my main topic, I will not attempt to resolve this dispute.

Jeffreys's theory assigns prior probability distributions to sets of hypotheses either on the basis of objective (frequency) information relevant to H or on the basis of conjugacy. Conjugacy is an algebraic concept which guarantees that a Bayesian analysis will be mathematically neat, in the following sense. A family of probability distributions P is *conjugate* to a family of likelihood distributions L when it has the property that using a member of P as the prior probability distribution guarantees that the posterior distribution will have the same mathematical form as the prior. (More precisely, P is conjugate to L iff when a prior distribution is in P and a likelihood function is in L the Bayesian posterior distribution

is guaranteed to be in P.) A general principle of Jeffreys's theory is that a prior probability function representing ignorance should be chosen from a family of distributions conjugate to the likelihood function (more precisely, conjugate to a reasonably narrow family containing the likelihood function) whenever possible. Since this guarantees that the prior and posterior will have a similar mathematical form, it has great practical advantages, especially when the posterior distribution becomes the prior distribution for a subsequent analysis (as it often does). Note, though, that this is purely a mathematical criterion, and one whose only clear advantage is simplicity of calculation. No philosophical justification for choosing conjugate priors has been suggested, and most writers on this principle — even its supporters — see it as ad hoc from the epistemological point of view.

In order to construct conjugate priors for likelihood functions indexed by $\theta \in \Theta$, we need to know what type of parameter θ is. "Type" here is meant to distinguish primarily between location and scale parameters, as follows.

In the simplest two-dimensional case, a location parameter is one which we can vary to move a distribution left or right along the x-axis, while a scale parameter is one which we can vary to compress or expand a distribution towards or away from its centre. Typically the mean of a distribution is a location parameter while its variance and standard deviation are scale parameters. A formal and reasonably general definition of parameter types is:

> [L]et X and Θ be scalar random variables. If the conditional distribution of $X - \Theta$ given $\Theta = \theta$ is the same for all θ, then Θ

> is called a *location parameter* for X. If $\Theta > 0$, and the conditional distribution of X / Θ given $\Theta = \theta$ is the same for all θ, then Θ is called a *scale parameter* for X.
>
> . . . Now, let X be a vector, and let Θ be a scalar. Let $\mathbf{1}$ denote the vector of the same length as X with every coordinate equal to 1. Then Θ is a *location parameter* for X if the conditional distribution of $X - \Theta\mathbf{1}$ given $\Theta = \theta$ is the same for all θ. If $\Theta > 0$ and the conditional distribution of X / Θ given $\Theta = \theta$ is the same for all θ, then Θ is a *scale parameter* for X.
>
> . . . Next, let X be a vector, and let Θ be a vector of the same dimension. Then Θ is a *location parameter* for X if the conditional distribution of $X - \Theta$ given $\Theta = \theta$ is the same for all θ. If Θ is a nonsingular matrix parameter [i.e., if Θ^{-1} exists] and the conditional distribution of $\Theta^{-1}X$ given $\Theta = \theta$ is the same for all θ, then Θ is a *scale parameter* for X.
>
> (Schervish 1995, p. 345)

A more intuitive and equally general (although less precise) definition is as follows. θ is a location parameter iff the likelihood $p(x|\theta)$ depends on θ only via $\theta - x$: an example is the mean of a Normal distribution. θ is a scale parameter iff $p_\theta(x)$ depends on θ only via θ / x: an example is the variance of a Normal distribution. For Bayesian analysis (which we are considering here), these conditions need only hold at the value $x = x_a$.

Not all parameters are either location or scale parameters; further definitions can be made without limit to take account of other possibilities for the algebraic role of θ.

Given a classification of parameters, Jeffreys's rules for priors representing ignorance are as follows. I state these without discussion because I have no intention of either criticising or defending them; my comments

about Jeffrey's theory rest on more general considerations than whether his rules for priors are plausible when taken individually.

- If Θ is finite, we have the obvious (although not obviously right!) Laplacian principle of indifference: each of the $\|\Theta\|$ possibilities is assigned an equal probability, so $p(\theta) = \frac{1}{\|\Theta\|}$, where $\|\Theta\|$ is the size of the set Θ.
- If θ is a location parameter, or a scale parameter which runs from $-\infty$ to ∞, then $p(\theta) = k$, for any constant k. (The choice of k does not matter, as it cancels out when the posterior distribution is derived using the Bayesian machinery described above.) This is an *improper* prior: it does not integrate to 1.
- If θ is a scale parameter which runs from 0 to ∞ then $p(\theta) = \frac{1}{\theta}$ (not to be confused with $\frac{1}{\|\Theta\|}$).

. . . and so on. The classification of θ into location parameters, scale parameters and so on is necessarily incomplete, and hence so is Jeffreys's theory. In itself I cannot see this as a criticism: I know of no argument to the effect that our theory of statistical inference *can* be complete, apart from specific suggestions that it should be this complete theory or that complete theory, none of which is without its drawbacks.

A more important criticism of Jeffreys's theory is that it is ad hoc. This criticism is often made, and rightly so. In many parts of Jeffreys's theory his justifications for his choices of priors are subtle, complicated and easily missed; so the ad hocness charge is not always easy to make stick. But in other places the ad hocness charge is clearly right. This is illustrated nicely by the following exchange between Jeffreys and Haldane. For $\Theta = (0, 1)$, Haldane suggests the alternative prior

$$p(\theta) \propto \frac{1}{\theta(1-\theta)}.$$

Jeffreys's response is to dismiss this as giving "too much weight to the extremes of Θ"; but he has no principled discussion of what "too much weight" might be (in contrast to the care he takes over those parts of his theory which he himself considers to have philosophical importance). In short, Jeffreys as good as admits the charge of ad hocness. This point is ad hominem, but that is the way with charges of ad hocness: the burden of proof absolutely has to be on the defenders of a putatively ad hoc theory, not on its attackers. Since I can neither find in the literature nor see for myself a principled defence of the whole of Jeffreys's theory of ignorance priors, I conclude that it is ad hoc. (And so is Haldane's suggested replacement.)

There is another component of objectivity in Jeffreys's work (as compared to the Subjective Bayesian schools), emphasised especially in his (1973): this is that we should order hypotheses according to their simplicity, which in turn can be measured by the number of parameters needed to state the theory. (Strictly speaking, it is only the number of "adjustable" parameters which is taken to be relevant: an adjustable parameter is one which analysis may attempt to estimate, as opposed to one which is known for sure.) Howson and Urbach rightly object to this, pointing out that Newton's theory, for example, has very few adjustable parameters as usually written "[b]ut as applied, say, in the kinetic theory of gases, it contains of the order of 10^{23} undetermined parameters, and when further degrees of freedom are added, the number rises correspondingly" (Howson & Urbach 1993, p. 418). In any case, merely *ordering* hypotheses is not enough to give us an objective theory of statistical inference: the claim to objectivity

of Jeffreys's theory rests on its ignorance priors, which are not just ordered but fully specified. (This is very clear from Jeffreys's (1961,1973), and is admitted even by his strongest supporters (Jeffreys & Berger 1991).) But in the work of Jaynes, to which I now turn, considerations related to simplicity provide not only an ordering of hypotheses but also actual prior distributions.

CONJUGATE IGNORANCE PRIORS II: JAYNES

Jaynes has written about the genesis of his theory in a way which also serves nicely to introduce the content of the theory:

> In 1965 it occurred to me that one very reasonable interpretation of 'complete ignorance' was group invariance. . . . I found immediately a much deeper understanding of the Jeffreys prior . . . in the location-scale parameter problem. This rule had been rejected [by me] because Jeffreys' argument in favor of it seemed ad hoc and arbitrary. But now it was clear that the point was not merely that σ was positive, the rationale that Jeffreys had given [sic: in fact, Jeffreys gave more rationale than that, albeit perhaps still not enough]. The point was that σ was a scale parameter, complete ignorance of which meant invariance under the group of scale changes. I immediately became an advocate, rather than a critic, of the Jeffreys rule . . . with the sanction of a clear rational justification.
>
> This work, which was for me a major advance in thinking [and which has subsequently become a widely studied theory] suffered the standard fate. It was submitted to a well-known statistical journal in 1966, and was indignantly rejected. The editor (whom I had thought to be a Bayesian) took the trouble to

> write me a letter requesting that I never again send him anything like it.
>
> (Jaynes 1983, p. 115)

Unlike Jeffreys's principles for choosing conjugate priors, Jaynes's principle of group invariance has an epistemological basis. One starts by "finding the group of transformations on the parameter space which convert the problem into an equivalent one" (Jaynes 1968, p. 227, reprinted as Jaynes 1983, p. 116), where by "equivalent" Jaynes means *epistemically* equivalent. Jaynes does not give a complete characterisation of the ways in which we might decide that a transformation of a problem leaves it equivalent; he argues merely by paradigm examples. Possibly he is only able to get away with this because the notions of location parameter and scale parameter (defined above) cover the vast majority of the uses of statistical inference; and so examples which seem to cover those cases adequately have been found convincing, despite the obvious lack of a full justification for Jaynes's theory. In the case of location parameters, Jaynes argues that the appropriate group of transformations is the infinite group formed by the real numbers under addition, $(\mathbb{R}, +)$, which transforms the variable x into $x + a$ for any fixed a, without affecting the results of the analysis. Similarly, in the case of scale parameters, the appropriate group of transformations is said to be $(\mathbb{R}, \times)$, which transforms the variable x into bx for any fixed b.

These groups of transformations serve as a way of tightening up Jaynes's older principle of *maximum entropy*, which says that "the prior probability assignment should be the one with the maximum entropy consistent with the prior knowledge" (Jaynes 1968, p. 229), where by "entropy"

is meant the following generalisation of Shannon entropy (Jaynes 1968, p. 235):

$$H = -\int p(x)\ log\left[p(x) / m(x)\right] dx$$

The function $m(x)$ is generally underdetermined; Jaynes's group-theoretic considerations or some such are needed to fix $m(x)$. When $m(x)$ is fixed as recommended by Jaynes, Jeffreys's theory is recovered, with some minor exceptions which, from the point of view of Jaynes's theory, can be counted as mistakes on Jeffreys's part. Unlike Jeffreys's theory, though, Jaynes's gives us some idea of how to extend the theory beyond the classification of parameters so far produced: the extension will depend on finding transformation groups which leave the problem epistemically equivalent.

Jaynes's theory is remarkably complete, attractively simple and relatively objective, but it has its problems. One is that some of his (and Jeffreys's) "prior probability distributions" are not strictly speaking probability distributions at all: they do not integrate to 1, as a probability distribution must. This presents no immediate mathematical problem: the posterior distribution *is* guaranteed to be a proper probability distribution on Jaynes's theory, so all the inferences about hypotheses drawn from his theory are straightforwardly probabilistic. However, Stone and others have proved that any theory which uses Bayes's Theorem with *improper* priors (priors not integrating to 1) can be fed examples which lead them into strict, logical internal inconsistency (Stone 1976). This inconsistency is decision-theoretically acceptable because it cannot be used to ensure a sure loss in a betting scenario (see Hill's comments in Berger & Wolpert 1988, pp. 167–171); nevertheless, from the philosophical point of view any

inconsistency is a high price to pay. Some Jaynesians hope that this inconsistency can be eliminated by finding some way of approximating the improper prior distribution by proper prior distributions (much as the inconsistencies in Greek calculus were eliminated by Weierstrass's method of taking limits).

A common criticism of Jeffreys's and Jaynes's use of priors representing ignorance is that there is no such thing as a probabilistic representation of ignorance. For example:

> it [is not] a tenable claim that the distribution which maximises entropy is "the one which is maximally noncomittal with regard to missing information" (Jaynes, 1957, p. 623). Any distribution, in our opinion, is as informative as any other insofar as it supplies a definite probability to every Borel set.
>
> (Howson & Urbach 1993, p. 417)

The Borel sets are just the mathematically well-behaved subsets of X; so Howson and Urbach are saying that every prior which has no holes in it is equally informative. Indeed, both subjective and ignorance priors seem similar in terms of first-order properties such as assigning probabilities to the same sets of possible data. But it is second-order (and higher) properties which tell us (if anything does) how well a prior represents ignorance: in particular, Jaynes's claim is that a particular measure of the *spread* of a distribution (namely its entropy) measures the extent to which it represents ignorance. Howson and Urbach ignore such second-order properties in their criticism of Jaynes.

Jaynes concedes that his method does not represent *complete* ignorance, but claims that to reject his method on these grounds "would be just as

absurd as to reject Euclidean geometry on the grounds that a physical point does not exist" (Jaynes 1968, p. 236). This may be right, but it is hardly a full defence. Euclidean geometry has been defended, for most of its history, only as an axiomatic system based on agreement about Euclid's axioms; whether a physical point exists is irrelevant to such a defence. In contrast, the properties of ignorance (ignorance itself, not a formalist's primitive concept such as a Euclidean point) are essential to the justification of Jaynes's method. On the other hand, modern defences of Euclidean geometry often do depend on the properties of physical space, so Jaynes's theory is in a similar position to these modern theories of geometry. But such theories *fail* to defend the truth (simpliciter) of Euclid's theory! And this is so not only because physical space is not, in fact, Euclidean, but also because in order to adequately represent physical space Euclid's theory would have to be dramatically recast in a more synthetic mould, giving physical justifications for its axioms. Similarly, Jaynes's theory requires either a defence of the existence of complete ignorance or a substantial argument showing that partial ignorance is necessarily best represented in the way he suggests. Such an argument does not yet exist. As with Jeffreys's theory, my criticism does not show that it is flawed; only that it is (as yet) insufficiently justified.

ROBUST BAYESIANISM

Robust Bayesianism is a type of Bayesianism which is based on Subjective Bayesianism, both philosophically and mathematically, but which avoids drawing subjective conclusions by using the fact that conclusions about hypotheses can sometimes be drawn without assuming that any particular

subjective prior is correct (Edwards et al. 1963). Instead, one finds conclusions which come out true on *any* reasonable prior, where "reasonable" is operationalised in terms of constraints which are either objective or (at least) uncontentious. The conclusions are often phrased as if they were approximations, but when that is the case they are precisely delineated approximations: a Robust Bayesian analyst will typically say, "my posterior distribution is such-and-such; and for any prior in the precise class so-and-so the posterior distribution differs from mine by at most the function $f(\theta)$, where f is defined by the equation such-and-such".[38]

From an epistemological point of view, Robust Bayesianism is similar to supervaluation — the theory that a sentence containing a vague term is true iff all reasonable precisifications of the vague term make the sentence true. The idea is more plausible as a theory of statistical inference than as a general theory of truth. As a general theory of truth, it leads to counter-intuitive truth-values. For example, a typical supervaluationist has to admit that the sentence "My height is vague" is false (where "my height" is a vague term) because, under each possible precisification of my height, my height is precise. This problem essentially depends on the second-order nature of the sentence: only sentences that both use and mention vague terms fall into the trap. This situation *can* arise in Robust

38. For example, the Robust Bayesians William O. Jeffreys (not to be confused with either Harold Jeffreys or Richard Jeffrey) and James O. Berger (not to be confused with the Roger Berger of (Casella & Berger 2002)) show that Einstein's theory of General Relativity is more probable than a certain Newtonian theory (remember that hypotheses in this thesis must be simple hypotheses, so we cannot use a vague term such as "Newtonian theory" simpliciter) under the constraint that the prior for the epihelion of Mercury is symmetric with a peak at the observed value, and monotonic (constant or decreasing) on either side of that value (Jeffreys & Berger 1991). It could be argued that the objectivity of that constraint can be established by considerations about the mechanism which was used to measure the perihelion. (Such an argument would have to be long and detailed. I do not claim that it is obviously right, only that it is not obviously wrong.) Jeffreys and Berger's argument allows for priors of any degree of vagueness.

Bayesianism, since one could be evaluating a set of hypotheses *about* one's own prior probability distribution; but it is not of the nature of a typical scientific problem, so the problem is not widespread (and, to the best of my knowledge, has not even been noticed before). There are sophisticated versions of supervaluation which are not subject to this problem, and which could perhaps be used to create a more complicated version of Robust Bayesianism which could handle hypotheses about priors.

A much more obvious and widespread problem with Robust Bayesianism is that it is not objective unless the constraints are objective and — even worse — there is no general theory showing that such constraints always exist, even subjectively. However, the progress of Robust Bayesianism is exciting to watch, both because large classes of problems *can* be shown to have such constraints (sometimes objective constraints, sometimes just very plausible subjective constraints) and because these classes of problems *might*, for all we know to date, be able to be extended without limit.

Mayo (1996, pp. 359–360) uses the term "robust Bayesianism" to describe methods which use Bayesian mathematics but assess their procedures in terms of Frequentist error rates. I have not seen this use of the term elsewhere; it is certainly not standard in the Bayesian literature. Mayo rightly classifies such procedures as Frequentist.[39]

OBJECTIVE SUBJECTIVE BAYESIANISM

All types of Objective Bayesianism have the same mathematical form as Subjective Bayesianism, and this allows any Objective Bayesian method to be easily converted into a Subjective Bayesian method when epistemic

39. In Grossman et al. (1994) my colleagues and I describe a method of this sort, calling it a "unified" method because the same mathematics can be given either a Frequentist or (by ignoring the error rates) a Bayesian interpretation.

circumstances permit. This happens when the prior probability distribution calculated by an Objective Bayesian method is the same as the prior probability distribution assigned by a subjectivist investigator.[40] So token Subjective Bayesianism need not be in conflict with token Objective Bayesianism.

For example, Breslow (1990) has argued that Empirical Bayesian techniques are acceptable only in cases in which subjectivist Bayesian techniques would have given the same answers (to a high degree of approximation). There are such cases, despite the way in which Empirical Bayesianism uses the same data twice, because it is possible — and in fact fairly likely in typical scientific situations — that the prior produced by illicit Empirical Bayesian methods is roughly the same as would have been produced by subjectivist methods: in other words, Empirical Bayesianism, using purportedly objectivist means, often happens to duplicate the subjective degrees of belief of the scientists involved. (This is especially likely when the amount of data is large; and in any case the error caused by double-dipping on the data tends to zero as the amount of data tends to infinity.) In that special case, the fact that the same set of data is used to calculate the likelihood function is not a problem, at least for the subjectivist, who a fortiori believes that the two techniques, Subjective Bayesianism and Empirical Bayesianism, in giving the same answers, must be giving the right answers.

---

40. Both schools of Bayesian thought are concerned with mathematical simplicity to a certain extent when they formulate prior probability distributions, and because of this it happens reasonably often that practitioners of the two schools of thought can agree *exactly* on a prior distribution. When this happens, all their conclusions must be identical; even their interpretations of their conclusions are if not the same then at least easily translatable into each other.

Perhaps this point is not as philosophically interesting as the conflict between Subjective Bayesianism and Objective Bayesianism which remains in principle no matter how often they agree in practice. The existence of this conflict is sometimes denied by those Subjective Bayesians who see the objectivity of the priors in Objective Bayesianism as completely illusory:

> *No prior probability or probability-density distribution expresses merely the available factual data; it inevitably expresses some sort of opinion about the possibilities consistent with the data.*
>
> (Howson & Urbach 1993; italics in the original)

So Howson and Urbach consider Jaynes's system to be no more objective than Subjective Bayesianism. But this is missing a point (albeit a small one). Certainly Jaynes's system "expresses" (so to speak) a limited set of possibilities consistent with the data, as do all the systems here and as must any system falling within the framework I set out in chapter 2. To *this* extent it is just as badly off as any other system. But Jaynes's claim is that it expresses ignorance objectively *given* the constraints of being a system of statistical inference suitable for doxastic agents. As I noted above, this claim remains less than fully justified, but it may yet be shown to be justifiable. If it is justifiable then there is a substantial difference between Objective Bayesianism and Subjective Bayesianism.

— 4 —

Survey II: Frequentism

In this chapter I survey Frequentist theories of statistical inference. (See chapter 3 for an introduction to this survey as a whole.) Here I will expound Frequentist theories as they are expounded by their proponents; and I will raise some issues which, prima facie, make the interpretation of the results of Frequentist analyses difficult. In particular, I will discuss the general impossibility of interpreting Frequentist probabilities epistemically (i.e., as directly relevant to what a rational doxastic agent ought to conclude).

I have chosen to separate the uncontentious aspects of Frequentism (this chapter) from the contentious aspects (chapter 7). All of the issues raised in this chapter are universally acknowledged aspects of Frequentist reasoning. None of them is seen as an objection to Frequentism by proponents of Frequentism, and I will argue that as far as I take the issues in this chapter they are right not to see them as objections: they are, for the moment, merely peculiarities. But in chapter 7 I will develop these issues further and show that, after all, they entail fundamental problems with Frequentism.

1. DEFINITION OF FREQUENTISM

The defining characteristic of **Frequentist** procedures is that they base all their conclusions on functions averaged over the sample space X. The ra-

tionale for this is the following principle, sometimes known in the literature as the *Repeated Sampling Principle*:

> A procedure for making inferences from data to hypotheses must have good average properties on repeated application in similar situations with different data.

The best way to think about what makes Frequentist methods of statistical inference special is to think in terms of Table 1:

	possible symptoms			
hypotheses	vomiting (observed in this case)	diarrhoea (not observed in this case)	social withdrawal (not observed in this case)	other symptoms & combinations (not observed in this case)
dehydration	0.03	0.2	0.5	0.27
PTSD	0.001	0.01	0.95	0.029
anything else	0.001	0.001	0.001	0.997

Table 1

A Frequentist method of inference is one which requires at least one whole row of the table. Thus, Frequentist methods are very different from Bayesian methods and from all other methods compatible with the likelihood principle, which only use the column in the table corresponding to the actual observation. In other words, Frequentist methods fix a hypothesis

and compare various hypothetical data sets under the assumption that that hypothesis is true, as opposed to what the likelihood principle tells us to do, namely to fix a data set and in some way compare hypotheses.

In the set of all possible methods of statistical inference complying with the framework given in chapter 2, Frequentist methods are almost the exact complement of methods compatible with the likelihood principle: the likelihood principle forbids a method of inference to use values of X other than x_a, while Frequentist methods *must* use values of X other than x_a.

I say that Frequentist methods are *almost* the exact complement of the methods compatible with the likelihood principle. It is possible for a method to be incompatible with the likelihood principle without being Frequentist. This is because a method might require values of X other than x_a, thus contravening the likelihood principle, and yet might not require a *whole* row of the table, thus making it not quite Frequentist. But as far as I can see, and as far as the literature goes, there are no useful methods of inference which fit into this chink: all the methods of inference you will come across here or elsewhere are either Frequentist or factualist.

Although the above distinction is the one to keep in mind to see the most important differences between Frequentist methods and others, it needs to be fleshed out a bit before we can see how Frequentist methods operate. I will first do the fleshing out in an abstract way, in the rest of this section, and then in the rest of this chapter I will give definitions and brief discussions of the most prominent Frequentist methods. In chapter 15 I will give concrete examples of how certain specific instances of these Frequentist methods differ from alternative, Bayesian methods.

A Frequentist method of inference (as I will use the term — see chapter 2 for my reasons) first fixes:

(I) a reference class of real or hypothetical experiments to be presumed similar to the experiment to be analysed,

(II) a set of hypotheses to be assumed false (the "error set"), and

(III) a mathematical form for the analysis, which varies from one Frequentist method to another.[41]

Given these ingredients, Frequentist statistical inferences are made by considering all possible equations of the chosen mathematical form (each of which has exactly one vector-valued variable, representing a possible observation) and choosing the one which minimises, subject to constraints which may vary according to different Frequentist schools of thought, the proportion of experiments in the reference class which cause hypotheses in the error set to be inferred to be true. This minimisation picks out a single equation; this equation is then applied to the actual observation, x_a, and the result is reported as the analysis of the actual experiment.

Frequentist methods are only applicable to experiments, not to merriments in general. In order to apply Frequentist methods to a non-experimental observational study, the study can be turned into an experiment by adding fictional experimental structure. For example, the accidental observation of a surprising supernova can be turned into an experiment by imagining that it was obtained from a random sample of observations of the whole sky. This extra step required to use a Frequentist

41. In addition to varying between methods, there is some variation within each method, if methods are grouped coarsely (counting, for example, Neyman-Pearson confidence intervals as defined below as a single method). This latter variation comes from the fact that all Frequentist methods use a "test statistic", $T(X)$, which is a real-valued function of the data chosen at the discretion of the analyst. See below and chapter 7 for more on how $T(X)$ is chosen.

method in a non-experimental situation is importantly arbitrary: different ways of turning an actual observation into an imaginary experiment yield different results. The supernova might be imagined to have been obtained from a sample from different *times* instead of different places, or from a sample of galactic clusters instead of the sky as a whole, or . . . and each such imagining gives different Frequentist analyses. Such methods of turning observations into experiments are extremely controversial, but I will not describe all the dimensions of the controversy here; I need only note that they can only lead to an enormous decrease in clarity compared to methods which can analyse observational studies directly, treating any epistemic ambiguities as part of the analysis rather than as a swept-under-the-carpet prerequisite for the analysis (Good 1976).[42]

2. THE NEYMAN-PEARSON SCHOOL

By far the most influential philosophy of Frequentism has been Neyman's, developed in the late 1930s in competition to Fisher's theories of maximum likelihood (see chapter 5) and hypothesis testing (see below). Although Neyman's theory is old, its modern forms as used by hundreds of thousands of researchers, as discussed by recent authorities such as (Barnett 1999) and (Stuart et al. 1999), and as currently championed by philosophers such as Mayo (1996) are, remarkably, unchanged from Neyman's original theory, except in a number of areas which I will discuss below (notably the shift in emphasis from actions to inferences, and the amalgamation of

42. Such methods do exist: Subjective Bayesianism is one. But I will not pursue this line of thought further here, because I do not have space to discuss such methods in detail, nor to discuss how Frequentist methods might best avoid this problem by making the relationships between non-experimental and experimental studies explicit.

Neyman's theory of hypothesis testing with Fisher's theory of P-values, defined below) and except in one area which is not relevant to my discussion (the decreasing popularity of frequentist theories of *probability*, largely in favour of propensity theories of probability within a Frequentist statistical framework).[43]

3. NEYMAN'S THEORY OF HYPOTHESIS TESTS

Neyman's theory of hypothesis tests starts with a reference class. Every observation from which an inference is to be drawn must be considered as part of a reference class. ("Reference class" is modern terminology for what Neyman himself called a "fundamental probability set".) This reference class may be constructed in one of only two ways: via random samples or via "random experiments".

REFERENCE CLASS 1: RANDOM SAMPLES

Firstly, the reference class may be a population from which the observation is a random sample (originally, one of a number of samples of equal probability, but the equiprobability requirement is superfluous and was soon relaxed in favour of known probabilities, not necessarily equal).

---

43. Neyman's theory was expounded in his (1937), reprinted as (Neyman 1967). Neyman and E. S. Pearson later jointly proved theorems which made it plausible that Neyman's theory could be applied in a wide variety of cases, and which helped to make the choice of test statistics ($T(x)$, as described below) within Neyman's theory less ad hoc. Because of these additions by Pearson, the theory is often called the Neyman-Pearson theory. Pearson's work was mathematically very important, but not philosophically. Therefore, when I am discussing the theory in a way which does not rely on Pearson's embellishments I will refer to it as Neyman's theory. When I am referring to essentially the same theory as interpreted by authors who do not distinguish between the 1937 theory and other versions, I will refer to it as the Neyman-Pearson theory. This distinction is only very rarely important: generally, what is true in one is also true in the other.

In Neyman's theory, the reference class must be one which "cannot be studied exhaustively" (Neyman 1967, p. 250). If it is "possible, though it might involve great practical difficulty," to study the population exhaustively, then "any character of this population will be a constant" and hence cannot be made the subject of statistical inference, at least not on Neyman's own version of the theory (Neyman 1967, p. 256). For example:

> [C]onsider a specified population, say the population π_{1935} of persons residing permanently in London during the year 1935 . . . . In the sense of the terms used here, there will be no practical meaning in a question concerning the probability that the average income, say I_{1935}, of the individuals of this population is, say, between £100 and £300. As the fundamental probability set consists of only one element, namely I_{1935}, the value of this probability is zero or unity, and to ascertain it we must discover for certain whether £100 $\leq I_{1935} <$ £300 or not. . . . Any calculation showing that $P\{$ £100 $\leq I_{1935} <$ £300 $\}$ has a greater value than zero and smaller than unity must be either wrong or based on some theory of probability other than the one considered here.
>
> (Neyman 1967, p. 256)

In other words, Neyman's theory does not allow probability questions to be asked of determinate propositions. Neyman used to cite Jeffreys's Bayesian theory (described in chapter 3) as an alternative which could be used instead of his own in the determinate case. But his advice has not been taken seriously; instead, later developers of Neyman's theory have made the theory universal. The constraint that the population from which random samples are taken be one which cannot be studied exhaustively has been abandoned, along with the restriction that one is not allowed to

place a probability on the average income being between £100 and £300. Perhaps confusingly, it is still accepted that the average income has a fixed value, and it is normal to ask whether the interval [£100, £300] *encloses* this value instead of whether the value is *in* the interval, thus emphasising that the value cannot move. Sadly, I have to say that this is but lip service to Neyman's point. Nevertheless, all of the rest of Neyman's constraints on sampling theory have been kept, including those which Neyman himself took to follow from the constraint which has now been abandoned; so whether the abandonment of the constraint itself ought to count as a major blow to the theory is rendered irrelevant, at least for my somewhat ahistorical purposes.

REFERENCE CLASS 2: "RANDOM EXPERIMENTS"

Secondly, the reference class may be a (typically infinite, possibly imaginary) series of experiments each of which gives results with certain known probabilities, not necessarily equal. Such an experiment is called a "random experiment". The phrase "random experiment" is meaningless when applied to an experiment in isolation. It assumes a meaning only when applied to a series of experiments together with either a description of what all the experiments have in common or, better, "a definition of the measure appropriate to the fundamental probability set and its subsets" (Neyman 1967, p. 254) — i.e., a description of the known probabilities for each hypothesis and each possible outcome (together with an assurance that they obey the probability calculus), just like the requirements laid out in chapter 2 and illustrated in Table 1.

The distinction between the two types of reference classes is clearly "only superficial" (Neyman 1967, p. 252), since random samples are generated by random experiments; thus, random experiment reference classes are the most general type and subsume population reference classes; so it is unnecessary, from the philosophical point of view, to discuss sampling specifically, as long as we adequately discuss random experiments.

PROBABILITIES FIXED ONCE AND FOR ALL

The series of experiments which makes up Neyman's reference class, and this series *alone*, gives the probabilities used throughout any ensuing Frequentist inference. It does this by giving "a definition of the measure appropriate to the fundamental probability set and its subsets" (Neyman 1967, p. 254), which is a table like Table 1 or an infinite version thereof. This is why Neyman's theory requires usually the whole table and at least a whole row.

Crucially, the "known probabilities" are fixed by the model for the *whole* analysis. This means, at least, that the probabilities given by a model are fixed for a doxastic agent or community for the duration of an inferential or decision-making episode. Neyman makes it very clear indeed that the probabilities are fixed even after relevant data comes to hand, despite the fact that the reference class is generally not homogeneous (uniform), and that it often happens that an event falls into a part of the reference class which is known to be special. Thus, if an event ε has a frequency of 1 in 4 in the reference class but it is discovered (for certain) that a single experiment uses only a particular part of the reference class in which the

event is known to have a frequency of 1 in 2, its probability does not become ½; it remains fixed at ¼.

It is not clear whether, according to the Neyman way of thinking, these probabilities can *ever* be changed, even between analyses. Very few authors address this question. In the writings of Neyman's school, so little is said about the option of changing these probabilities that one gets the impression that to do so would be beyond the pale of Neymanism: it would be not only to change the analysis but to change the *method* of analysis. To change the underlying probabilities is clearly to change reference classes; but the choice of reference class is essentially arbitrary (as Neyman is happy to admit), so this consideration is inconclusive.[44] In any case, regardless of whether probabilities can *ever* change, it is clear from the writings of Neyman and his followers, and from the practice of Frequentist applied statistics, that probabilities cannot be changed in typical situations in which new information comes to light.

FREQUENTIST PROBABILITY IS NOT EPISTEMIC

This fixity of probabilities in Neyman's theory even when relevant information comes to light is central to all of its inferential calculations, and has not been relaxed in its descendants. It entails that probability is not epistemic. An epistemic probability is one which represents the beliefs of a rational doxastic agent. I do not have a clear analysis of epistemic probability to offer; the idea suggests various ambiguities (e.g., perhaps

44. The only solution to this problem I am aware of is due to Seidenfeld (1979, p. 36): "the probabilities must *not* become altered because of knowledge available about some specific trial, e.g. the next one, which is *not* true of all trials in the repeated trial sequence." Presumably this is meant to imply that one may change the probabilities when information comes to hand which applies to all trials in the sequence. I do not think that Seidenfeld's solution is widely accepted; and, in any case, the condition he mentions is only rarely met, so his solution is not sufficiently general to yield a new theory of statistical inference.

the beliefs of an idealised agent, in some sense) which I do not have an opinion about. But no matter how we resolve such ambiguities, Neyman's theory cannot be epistemic, as I will now show.

In Neyman's theory it is possible for an event to have a low probability even if an agent employing the theory rationally expects the event to happen. I will show this using an example taken from the game of bridge. (Readers who do not care about the details may skip the footnotes to the rest of this paragraph.) Consider a bridge player who is asked, prior to a deal, the probability that he can defeat a contract.[45] Bridge players will quickly recognise that this probability is greater than the probability that the player's partner has the King of Spades.[46] The player might choose the obvious reference class (namely, all four players' hands), according to which the probability of his partner being the one of the four players to have the King of Spades is ¼. He might calculate error rates, including a P-value or a confidence interval as defined below, for the hypothesis that his partner has the King of Spades using this reference class, and he might promise to apply these error rates in his future decisions about the hand of cards, secure in the knowledge that he has a guaranteed low rate of error. Now suppose that the player's left-hand opponent does bid Seven Spades, after which the dummy's hand is made public.[47] Suppose,

---

45. Let us say that the contract is Seven Spades bid by his left-hand opponent, where the opponent claims to be able to win all the tricks with Spades as trumps, given that our proponent knows that the opposing bidder has the Ace of Spades but with no other information.

46. I apologise to those to whom examples drawn from bridge are gobbledegook. For those to whom the terminology of bridge brings happy memories but whose grasp of the inferential structure of bridge needs refreshing: our protagonist's partner having the King of Spades would render the opponents unable to win at least one of the tricks. If our player has the King himself, it will probably fall under the Ace, but if his partner has it it will probably win a trick, thus defeating the contract.

47. Now the player can see whether either he himself or the dummy has the King of Spades. Neither does, so his partner or the other opponent must have it.

moreover, that his partner's bidding indicates that he has a strong hand, most likely with some strength in Spades.[48] Now he *expects* his partner to have the King of Spades, and it is rational, in anybody's book, for him to act according to a probability of at least ½ that his partner has the King of Spades. Anybody, even a Frequentist statistician, would agree that the epistemic probability has increased as information has come to hand. But on Neyman's definition the probability has not changed, because the reference class has not changed. I can be sure that the reference class has not changed for two reasons. Firstly, the reference class *never* changes during an analysis, according to Neyman himself, no matter how much new information comes to light. Secondly, Frequentist statisticians do not change their reference classes as new information comes to light, not only because to do so would be to violate their (Neyman's) theory but also because they know that if they changed their error rates in this sort of case they would no longer be able to quote guaranteed error rates for their procedures. Now, a practising Frequentist statistician would ignore Neyman's definition for some purposes. He would be unlikely to say with a straight face that the probability of his partner having the King of Spades remained exactly ¼. But he would certainly (precisely by virtue of being a Frequentist) stick to the error rates which he had calculated before the new information came to light. In his calculations, even if not in his off-the-cuff statements, he would be willing to give a low probability to an event which he expects to happen.

My claim that practising Frequentists use Neyman's definition of probability (or, at worst, something operationally equivalent) can easily

48. He has bid Spades himself, perhaps.

be confirmed by looking at any applied science journal: P-values and confidence intervals are calculated on the original reference class, and are not updated when it becomes apparent that a variable has a value which is unlikely according to the reference class. In other words, the probabilities which a practising Frequentist statistician uses in his statistical analyses do not change as new information comes to light.

I have established that a Frequentist can give a low probability to something which he expects to happen (e.g. for the bridge partner to play the King of Spades). This would entail an internal contradiction if probability were epistemic, because it cannot be rational to expect something and yet to give it a low probability. Hence, Neyman probability is not epistemic.

Neyman clearly acknowledges that his notion of probability is non-epistemic, and is happy for it to be so. This is perhaps confusing, perhaps reasonable, according to one's position on the metaphysics of probability, but in either case it is certainly not self-contradictory. However, it is immediately fatal to a certain conception of inference. A hypothesis under consideration might state simply that the event ε occurs. Neyman's theory may give that hypothesis a low probability (¼), even though (in the absence of other evidence apart from the fact that ε occurs in ½ of the cases in a subclass to which, we may imagine, it happens to be known to belong) we should think that the hypothesis is probably true. Hence, on Neyman's theory, probability can no longer have any close connection with the credence due to a hypothesis. In this respect, Neyman's theory, and Fisher's slightly earlier theory reviewed below, were decisive breaks with most, although not all, earlier theories of statistical inference.

The whole point of statistical inference is to "move from beliefs and/or statements about observations to beliefs and/or statements about what cognitive states and/or actions we ought to adopt in regard to hypotheses" (to quote from chapter 1). In Bayesian theory, as in informal reasoning about probability prior to the modern schools of thought, we work out what to infer by calculating the probabilities of hypotheses. Since probability is non-epistemic in Neyman's theory, he cannot use probabilities directly to model what we ought to do or think; so his inference procedures cannot be as simple as calculating the probabilities of hypotheses.[49] Since the probabilities of hypotheses can no longer do this job in Neyman's theory, Neyman needs something else that can. The problem which this sets for him is well known. Birnbaum, for example, in the paper in which he first proves the likelihood principle, quotes Savage on this problem:

> Rejecting both necessary and personalistic views of probability [by which Savage means to encompass all epistemic views of probability] left statisticians no choice but to . . . seek a concept of evidence, and of reaction to evidence, different from that of the primitive, or natural, concept that is tantamount to application of Bayes' theorem. Statistical theory has been dominated by the problem thus created.
>
> (Birnbaum 1962, p. 277, quoting Savage)

Savage's point does not quite cover the subtleties of Neyman's own position in the 1930s since, as we have seen, he believed then that Bayesian inference had a role to play; but it does accurately cover the school of thought

49. In addition to this reason why Neyman cannot rely on the probabilities of hypotheses for inference, the discussion of the income of Londoners above shows that hypotheses generally do not *have* probabilities in Neyman's system; but we need not dwell on that issue because, even if hypotheses did have probabilities, they would be non-epistemic probabilities and hence not able to tell us directly what we should infer.

which Neyman founded, intentionally or not. Neyman's seminal (1937) addressed itself to the problem of Frequentist inference alone, and thus required a principle of evidence different from the probabilities of hypotheses. Neyman's solution to this problem is, of course, to make inferences in the Frequentist way, by calculating error rates. The following sections describe such methods.

NEYMAN-PEARSON HYPOTHESIS TESTING

A **Neyman-Pearson hypothesis test** is a function from a variety of parameters to a *rejection region*. The parameters of the test are:

- a hypothesis space, H,
- a single, simple hypothesis, $h_0 \in H$ (the *null hypothesis*),
- a sample space, X,
- a test statistic, $T(x_a)$, where x_a is an actual observation and T is a function which converts the observed value x_a into a simplified form (typically a single real number),
- a probability function defined on h_0 alone, p_{h_0}, and
- a desired probability (relative to p_{h_0}) of a type I error (defined below).

If the test falls into the rejection region then the null hypothesis is rejected. If the test does not fall into the rejection region then the null hypothesis in not rejected. Under no circumstances is the null hypothesis accepted: the procedure is essentially falsificationist. At least, the purest form of the Neyman-Pearson hypothesis test is essentially falsificationist. Not surprisingly, another form has evolved in which the null hypothesis is either accepted or rejected, according to the result of the test; and often

the two forms are used interchangeably by members of the same research group.

The resulting theory, according to Neyman, requires a mathematical model whose relevance to the world is something which needs to be tested empirically in specific cases, not something which can be assumed on epistemic grounds, and the results of which — acceptance and rejection — are actions, not epistemic outcomes. This helps him to avoid the problem of non-epistemic probabilities noted above: if only actions are considered then epistemology is not relevant (at least, not directly), so the counter-intuitive consequences of using non-epistemic probabilities are unimportant (or so it was argued). However, by the 1950s a variety of versions of the theory had evolved which disagreed on these points. These versions were all described interchangeably by the same small set of names (most notably: "statistics", "Neyman-Pearson statistics", "hypothesis testing"), and there remains quite some work to be done to differentiate the versions from each other and to delineate the versions which became clearly epistemic from those which remained (or at least attempted to remain) non-epistemic. Fortunately, the topics treated in this thesis are neutral between these alternatives . . . not because epistemology is unimportant, but because *all* versions of the theory have epistemic consequences, as I will show in chapter 7. So it will not be me who has to do the work of inventing new terminology to distinguish between the mutually incompatible stances which modern versions of the Neyman-Pearson theory take on questions of epistemology versus action.

Neyman-Pearson hypothesis tests are usually constructed to ensure that if h_0 is true then the probability of rejecting h_0 is 5%. This probability

is known as the *size* or *probability of a type I error* (or sometimes just *type I error*) of the test. The meaning of "probability" at work in this statement is Neyman's meaning: the probability of a type I error cannot change, even if in the process of making the test information comes to hand which tells us that h_0 is more or less than 5% likely to be true. As I mentioned above, this is not an inconsistency in the theory: it is merely a particularly anti-epistemic way of using the word "probability". Whether this marks a flaw in the theory in some sense weaker than inconsistency is something which I will discuss in chapter 7.

The modern Neyman-Pearson theory requires the statistical analyst to have an alternative hypothesis in mind (possibly a composite one), although Neyman himself did not always require this. The alternative hypothesis may be simply H with h_0 omitted $\left(H \setminus \{h_0\}\right)$; or it may be a hypothesis h_1 introduced especially for the purposes of a single Neyman-Pearson analysis. Here I take the former approach, for consistency with the framework I set out in chapter 2 and for easier comparison with other methods of inference.[50]

If H consists of only two simple hypotheses then the probability that a Neyman-Pearson test will fail to reject h_0 if h_0 is false is known as the *probability of a type II error* of the test. If H has a more complicated structure then the probability of a type II error is the maxiumum or supremum (least upper bound) of the probability of type II error as h_i varies, where $h_i \in H(h_i \neq h_0)$. The *power* of the test is generally thought of as one minus the probability of a type II error; and Neyman-Pearson theory holds that

50. If necessary, the latter approach can be accommodated within this framework by setting $H = h_0 + h_1$. In the context of this thesis such a move is unproblematic, although I would have to tell a longer story if I were trying to fully describe the pragmatics of constructing hypothesis spaces.

the size of a test is to traded off against its power. This is a substantive requirement; it may not seem so if probability is treated as epistemic, but remembering that Neyman's probability is not epistemic makes it far from trivial.

I said that the power of a test is thought of as one minus the probability of a type II error; but a little thought shows that the type II error cannot be calculated from the ingredients available to the statistician, because the type II error rate depends on the actual values of the unknown variables or, in other words, the true h. (This problem does not arise for the type I error, because the type I error is calculated on the assumption that h_0 is true. That assumption tells us the values of the unknown variables.) Frequentists sidestep this problem by defining the type II error — and hence the power — in terms of an estimate of the unknown variables. There is no principled Frequentist way to make this estimate: it is subjective in exactly the same way that Subjective Bayesian priors are subjective.

4. NEYMAN-PEARSON CONFIDENCE INTERVALS

In addition to testing hypotheses, Neyman, like most statisticians, wished to be able to estimate the value of a parameter. Being able to do one of these two things does not necessarily mean being able to do the other, because of the merely dichotomous nature of hypothesis testing: for example, knowing that a hypothesis passes a dichotomous test does not tell us whether it gives us a unique reasonable estimate of a parameter or one of many reasonable estimates (or perhaps, on some theories, no reasonable estimate at all). As we will see, in Neyman's theory there is a close link between the calculation of some significance tests and the calculation of some

confidence intervals, but that link is not a fundamental part of the theory; and the link between hypothesis tests and confidence intervals is broken in certain important cases, including the case of clinical trials described in chapter 15. Consequently, I must describe Neyman's confidence intervals from scratch.

Like all theories of estimation but unlike his theory of hypothesis testing, Neyman's theory of confidence intervals depends on H being indexed by a parameter θ. The definition of a Neyman-Pearson confidence interval is:

> If there exist functions of x, $T\downarrow$ and $T\uparrow$, both statistically independent[51] of θ, such that
>
> $$(\forall\theta) \qquad p(T\downarrow(x) \leq \theta \leq T\uparrow(x)) = 1 - \alpha$$
>
> then the interval $[T\downarrow(x_a), T\uparrow(x_a)]$ is a $1-\alpha$ **confidence interval** for θ.
>
> (adapted from Kendall & Stuart 1967, volume II, p. 99)

$(1 - \alpha)$ is then known as the *coverage probability* of the interval.

$T\downarrow$ and $T\uparrow$ may be chosen in a variety of ways. In general, the choice is ad hoc, but a number of additions to Neyman's theory (mostly due jointly to Neyman and Pearson) make it less so.

The primary criterion used to pick a Neyman-Pearson confidence interval is to choose the "shortest" interval by choosing $T\downarrow$ and $T\uparrow$ such that

$$(\forall[T\downarrow, T\uparrow])(\forall\theta' \neq \theta) \quad p_{h0}\left(\theta' \in [T\downarrow(x), T\uparrow(x)]\right) \leq p_{h_0}(\theta' \in [T\downarrow(x), T\uparrow(x)])$$

51. A is statistically independent of B iff $p(A\&B) = p(A)p(B)$.

where θ represents the unknown true value of the parameter(s) of H. However, this primary criterion is not generally enough to pick out a single confidence interval. For a start, there need not *be* a unique shortest interval, so this supposed definition hides a degree of arbitrariness. And even when there is a unique shortest interval according to the definition above, the choice of interval is not invariant under change of variables — i.e., the interval which is shortest for θ will not be the shortest for $f(\theta)$ for arbitrary f, even when f is a bijection (a one-to-one correspondence).

There are various other criteria for choosing confidence intervals; unfortunately, none of the others is guaranteed to apply either. There is no principled theory which assigns priority to one criterion over another. There is some consensus that choosing shortest intervals is most important (Stuart et al. 1999); beyond that, the next priority is usually to choose an interval using a method which has a "probability of covering the true value of the parameter . . . greater than the probability of covering any false value, no matter what the true value be" (Seidenfeld 1979, p. 54) (where "probability" has its non-epistemic sense, as usual in a Frequentist method). The third priority is usually to insist on a form of mathematical invariance (Seidenfeld 1979, p. 55) which falls short of complete invariance under parameter transformations. (Most likelihood-based methods of inference, in contrast, have complete invariance under parameter transformations: for example, Bayesian inference using proper priors has this property.)

Now that we have used a probability statement involving a fixed parameter for the first time, this is a good moment to revisit Neyman's stricture about such things. It may appear that the probability statement used in the definition of shortest intervals contradicts Neyman's insistence

that fixed values such as θ' cannot have probabilities; but in this case x is a random variable (see chapter 2), and it is this which licenses probabilities. If x were replaced by x_a (an actual observation), the formula would no longer have a meaning.

It is instructive to compare this case with the commonly-seen $p\big(\theta \in [T\downarrow(x_a), T\uparrow(x_a)]\big) = 95\%$ — "my confidence interval has a 95% chance of containing the true value of its parameter" — in which there are no random variables, and which Neyman would not permit. Statements of that impermissible form are often seen in the scientific literature, even from committed Neyman-Pearsonites, because the above distinction is often lost even on the faithful.

An interesting objection to choosing the shortest confidence interval has been raised by Howson and Urbach: that the *only* justification of confidence intervals available within Neyman-Pearson theory — namely, the usefulness of intervals with known coverage probabilities — is a justification which gives equal validity to both long and short intervals (Howson & Urbach 1993, p. 245) (and, indeed, to unions of disjoint intervals). In considering this objection, it is important to realise that it is not something which can be overcome by a small adjustment to Neyman's theory. The objection follows from Neyman's insistence on evaluating methods of inference *only* according to their rate of errors on repeated applications of a fixed rule. This restriction is fundamental to Neyman's theory (by which I mean that if other ways of evaluating inference methods were allowed to have any force the theory would be totally different, both philosophically and practically). In terms of inference methods evaluated in this restricted way, all confidence intervals complying with the above equation

are equally correct, just as Howson and Urbach claim. Therefore the decision to choose the shortest is not a decision based on the principles of the theory.

This objection is right; but it is not an objection to the *validity* of the theory. Neyman's theory claims that any confidence interval will do; Neyman's and Pearson's extensions to the theory tell us to choose the shortest interval; this choice is not justified by the theory, but that does not make it wrong. On the other hand, it certainly makes the theory ad hoc. I will return to this point in chapter 7.

Although Howson and Urbach's argument does not invalidate Neyman's theory, it does invalidate the almost universally held belief among Frequentists that the shorter interval can be adequately justified by the fact that it is more "accurate" or "precise" and therefore gives us better information about θ than the longer interval does. Howson and Urbach quote Mood giving this argument in 1950 ("in comparing two 95 per cent confidence intervals, he stated that one of them was 'inferior' because of its greater length, for 'it gives less precise information about the location' of the parameter." (Howson & Urbach 1993, p. 245), quoting (Mood 1950, p. 222)), and they could just as easily have quoted almost any statistics textbook from 1940 up to now. The shorter interval *is* more precise, but a longer interval is equally well justified by its Frequentist characteristics, so it makes no sense to argue that the precision of the shorter interval gives us better information about θ. Consider:

- Fred tells me that a standard London bus is exactly 16 metres long.
- Jenny tells me, more vaguely, that a standard London bus is somewhere between 10 and 30 metres long.

– Fred and Jenny are equally reliable judges of length.

If I have to pick one of their two estimates, should I should pick Fred's on the grounds that it is more precise? Not necessarily. If I were planning to jump a London bus on a motorcycle, I would be much better advised to pick Jenny's. There is no rule of rationality that says we should use the most precise estimate when less precise estimates are equally well justified.

Another criterion for choosing a confidence interval from among the infinite number of intervals with equal coverage probabilities is to make sure that the centre of every confidence interval is a point estimate defined using an estimator function with certain supposedly desirable properties such as consistency and unbiasedness. The main purpose of this is to narrow down the set of acceptable estimates in order to make the theory less ad hoc, rather than to make the estimates themselves better justified. Making the theory less ad hoc is important, because it reduces the scope for an individual who wants to see a particular result to conduct an analysis which favours his preferences; but there are many such ways of making the theory less ad hoc, and none of them seems to be particularly central to the theory. It is therefore not clear to me whether this requirement should be seen as a central part of Neyman's theory. Certainly neither Neyman nor his successors defend it in the rigorous way in which they defend the parts of the theory presented above. In any case, I will discuss these supposedly desirable requirements, especially unbiasedness, in chapter 11.[52]

52. As a side issue, the choice of confidence intervals can be made less ad hoc if a utility function is available (Wald 1947, Lindley 1990b, p. 46). Such a decision-theoretic situation is outside the scope of this thesis, as explained in chapter 2. But it can be argued that when a utility function is available Bayesian decision theory is more attractive than Frequentist decision theory, and for this and other reasons (foremost of which is probably the unwillingness of the vast majority of Frequentists to countenance anything that smacks of subjectivity to the extent that a utility function does) utility functions are very rarely used in Frequentist inference.

5. INFERENCE IN OTHER DIMENSIONS

The Frequentist theories which I have discussed above concern hypothesis testing, which we might characterise as zero-dimensional inference, since it results in either rejecting or failing to reject a point null hypothesis, and confidence intervals, which we might characterise as two-dimensional inference, since the result is a pair of real numbers. A part of the Neyman-Pearson theory of inference which I have not presented here is the theory of estimating a parameter by a single integer or real number, which we might characterise as one-dimensional inference. I see no need to deal with one-dimensional inference separately, for three reasons:

- because the single-number inference problem is treated by the literature as a less important problem than the zero- and two-dimensional cases;[53]
- because the single-number inference problem is usually subsumed into the theory of confidence intervals (since the best single-number estimate of a parameter is usually considered to be the centre of a confidence interval for that parameter); and
- because the single-number inference problem raises no philosophical issues other than those which have already been raised by hypothesis tests and confidence intervals.

What, though, about inferences resulting in measures of *more* than two dimensions? Such things simply do not arise in the literature on the foundations of Frequentist inference. They do very occasionally arise

53. The grounds for this, when grounds are given at all, are that a single-number estimate ought always to be accompanied by a confidence interval lest we assign it too much weight (Armitage & Berry 1994).

in the Bayesian literature, in the decision theory literature, and in the literature on the mathematics of probability, but none of those need trouble us here, because the Bayesian theory on higher-dimensional inferences is identical to the Bayesian theory of lower-dimensional inferences (both philosophically and, in the most important respects, mathematically as well), and hence has already been treated in chapter 3, while decision theory and measure theory lie outside the scope of this thesis.

I know of no theoretical reasons why the literature on the foundations of Frequentist inference should have ignored higher-dimensional inferences, but there are obvious practical reasons. For example, one might wonder about the properties and usefulness of a confidence interval whose bounds were pairs of real numbers (perhaps representing complex numbers) rather than single real numbers; but there would be little use for such a thing in the traditional domains of inferential statistics, which are the biological sciences very broadly construed (including agriculture). Consequently, there is no standard theory of such things for me to present in this survey chapter. Instead, I turn to alternatives to the Neyman-Pearson theory which remain within the Frequentist canon. In the following sections I will present three such theories: Fisher's, Fraser's (structural inference), and a mishmash theory which has no name and yet is the most commonly used of all.

6. FISHER'S FREQUENTIST THEORY

By far the most influential Frequentist theory which does not stem from Neyman's theory is Fisher's. Fisher's foundational work mostly predates Neyman's and, of course, influenced Neyman. Despite this, I have discussed

Neyman's program first because Neyman's is more important, both because its philosophy is much more internally coherent than Fisher's and because modern Frequentism owes much more to Neyman's philosophy than it does to Fisher's; and so it is more important for us to be clear about Neyman Frequentism than about Fisher Frequentism. (For a brave attempt to make Fisher's Frequentism coherent enough to rival Neyman's, see Seidenfeld 1979.)

Fisher's work is notoriously plagued by internal philosophical and mathematical contradictions — contradictions which are sometimes attributed to his attempts to overcome fundamental problems in all preceding theories of statistical inference and sometimes, less charitably, to Fisher's personal inability to admit to having been wrong (Savage 1976). The details of Fisher's Frequentist program are particularly tangled. His proposal for Frequentist confidence intervals, in particular, were intimately connected with his program for fiducial inference (described in chapter 5), although some strands of his Frequentist program make sense without assuming the prerequisites of fiducial inference (Seidenfeld 1979). In this section, I will discuss the two completed parts of his Frequentist program: his theory of significance tests and (very briefly) his theory of confidence intervals. I will discuss the the non-Frequentist parts of his program, maximum likelihood estimation and fiducial inference, in chapter 5. (It has been argued that fiducial inference is Frequentist (Seidenfeld 1979), but whether it is or not will make no difference to my appraisal of it.)

Fisher importantly disagreed with Neyman about the construction and relevance of reference classes. Fisher believed that reference classes could not be dependent on ancillary statistics (see the definition in chapter

5), and he had a more epistemic notion of probability than Neyman. Despite this, Fisher agreed with Neyman's insistence that a probability based on a reference class could not be changed once an event was known to lie within a smaller subclass, as described above. All Frequentists seem to be agreed on this point.

Fisher's signficance tests require the same ingredients as Neyman's except that they do not use H (except for h_0). In other words, Fisher's significance tests do not require an alternative hypothesis. They do still require a test statistic (which, recall, is a function, usually designated T, which simplifies members of the sample space X, usually converting them to real numbers).

The basis of Fisher's significance test is the set of hypothetical outcomes

$$\{T(x) \geq T(x_a)\}$$

for a "one-sided" test and

$$\{\,|T(x)| \geq |T(x_a)|\,\}$$

for a "two-sided" test. x_a enters into the significance test only through these sets.

How to make the choice between one-sided and two-sided tests is controversial, and it has been argued that no principled choice is possible (Salsburg 1989). I will not discuss this question here, since I will argue in chapter 7 that any choice of significance test is ad hoc, regardless of whether there is a principled means of choosing between one-sided and two-sided tests.

The *significance level* or **P-value** of the observed outcome, x_a, is calculated as

$$\mathcal{P} = p_{h_0}\big(T(x) \geq T(x_a)\big)$$

or

$$\mathcal{P} = p_{h_0}\big(\,|T(x)| \geq |T(x_a)|\,\big)$$

for one-tailed and two-tailed tests respectively.

If $\mathcal{P} < 5\%$ (or some other fixed number) then the outcome is considered *statistically significant* and h_0 is rejected. According to Fisher, h_0 should also be rejected if $\mathcal{P}$ is close to 1, but this stipulation has not survived in descendants of his theory.

Clearly, for every choice of H, h_0, X and x_a, statistical significance occurs in less than 5% of repeated trials in some Neyman reference class. This fact corresponds to the fact that every Fisher significance test is mathematically equivalent to a Neyman-Pearson hypothesis test (or rather to a family of tests with fixed size but varying power depending on the alternative hypothesis chosen). Fisher's methods are *constructed* in a way which guarantees this property. However, Fisher insisted that his test should not be given a Neyman-style *justification*. Fisher's preferred justification was the following:

> The force with which such a conclusion [rejection of h_0 on the basis that $\mathcal{P}$ is very low] is supported is logically that of the simple disjunction: *Either* an exceptionally rare chance has occurred, *or* the theory of random distribution is not true.
>
> (Fisher 1973, p. 39, quoted in Edwards 1972, p. 177

I will discuss this justification further in chapter 7. I count Fisher's significance tests as Frequentist, which means (in my terminology, at least) that they must have a Frequentist justification. Fisher distanced himself from the Neyman Frequentist justification, but the justification he offered instead is still Frequentist: the reference to "an exceptionally *rare* chance" (my emphasis) is an appeal to the frequency with which such a chance will occur in a population of experiments yielding different data, which is precisely what separates Frequentist from non-Frequentist methods of inference. In other words, Fisher's methods are not only constructed by fixing a hypothesis but are also *justified* by fixing a hypothesis and comparing various hypothetical data sets under the assumption that that hypothesis is true, as opposed to fixing a data set and in some way comparing hypotheses.

Fisher also developed a theory of interval estimates, somewhat like Neyman's confidence intervals. Fisher's theory agrees with Neyman's numerically in most cases (although not all). From all points of view — philosophical, mathematical and practical — Fisher's and Neyman's theories of interval estimates have become merged, so that their descendants cannot be cleanly distinguished from each other, while descendant theories take their justification from Neyman's epistemology (or lack of epistemology, as described above) or, very occasionally, from Fisher's fiducial argument (chapter 5).

7. STRUCTURAL INFERENCE

A relatively recent Frequentist development is the theory of structural inference, which combines mathematical methods essentially the same as

pivotal inference (described in chapter 5) with an orthodox Neyman philosophy of inference (Fraser 1996, Fraser 1968). It has been criticised for not being applicable to all cases (Barnett 1999, p. 318). Since it shares all the epistemic features of Neyman's theory we need not consider it in detail. Mutatis mutandis, it is subject to all the objections to Neyman's theory which I will give in chapter 7.

8. THE POPULAR THEORY OF P-VALUES

Disagreements between Fisher and Neyman about reference classes and probability were of no interest to the vast majority of the buying public. In the late 1940s and early 1950s a plethora of textbooks was produced to satisfy the postwar boom in applied statistical inference, and the most popular of these books made mincemeat of the careful distinctions invented by Neyman and Fisher (Gigerenzer 1993). The distinctions were not lost from the more theoretical parts of the literature, of course, but in the practically oriented textbooks a third Frequentist method was born, like something from Minoan mythology with the head of Neyman and the body of Fisher.

This popular theory, which has no name as far as I can tell, can be found in almost any elementary introduction to statistics from 1960 to the present day. It uses Neyman's theory of reference classes, the philosophy of Fisher's theory of P-values combined with the mathematics of Neyman's theory of hypothesis tests, and Neyman's theory of confidence intervals. Gigerenzer argues that this combination of theories is a misrepresentation of both Fisher and Neyman (Gigerenzer 1993). Gillies (1973, pp. 206–215) has shown that Neyman himself sometimes used a mishmash of his own

and Fisher's theories (although not the same mishmash as the popular one; specifically, Neyman calculated P-values without an alternative hypothesis, which is in accord with Fisher's theory but contrary to his own). Fisher, on the other hand, never settled on a clear account of his own preferred methodology. So we need not feel too bad on Neyman's or Fisher's behalf.

Since the popular theory contains no new ingredients, it needs no further description; but it needs to be mentioned because it is by far the most used statistical theory of all time and the basis of almost all contemporary experimental science.

In addition to the published popular theory, there is a mostly unpublished popular folklore of statistics which uses epistemic terms to describe Neyman's non-epistemic probabilities. For example, "Oakes (1986, p. 82) found that '96% of academic psychologists erroneously believed that the level of significance specifies the probability that the hypothesis under question is true or false.' " (Gigerenzer 1993, p. 330).

— 5 —
Survey III: Other Theories

In this chapter I survey the remaining theories of statistical inference. (See chapter 3 for an introduction to this survey as a whole.) This chapter is a mishmash. One of the theories covered here, the pure likelihood theory, has a sound theoretical justification but is rather incomplete in comparison with the major theories — Bayesianism and Frequentism — presented in the previous two chapters. The other theories presented in this chapter, apart from Shafer's, are all speculative in the sense that they have been invented without any discernable justification. They have, to date, been examined by only a small number of theorists and, except for the fiducial method, no serious effort has been made to give any of them a philosophical basis. Possibly for this reason, or possibly coincidentally, none of the methods which I label speculative has ever been in frequent use. Moreover, there is no reason I can find to think that any of them, including the fiducial method, *has* a philosophical basis. I mention them mainly for the sake of being exhaustive. Consequently, I will deal with these theories relatively briefly.

1. PURE LIKELIHOOD INFERENCE

The pure likelihood school of thought, the philosophical development of which is due largely to Hacking, holds that one should operate with the minimum of ingredients. In particular, this school holds that it is impossible

to do statistical inference without considering the likelihood function, $(\forall h \in H)\, p(x_a|h)$, but that one can do statistical inference while considering pretty much *only* the likelihood function. (Hacking himself does not present this parsimony as an important desideratum of his theory. I do so because it is what distinguishes it from the other methods presented in this survey.) Methods of statistical inference which attempt to rely only on the likelihood function I call **pure likelihood methods**.

THE METHOD OF MAXIMUM LIKELIHOOD

By far the oldest pure likelihood method is the **method of maximum likelihood**, also known as **maximum likelihood estimation**, which says that we should accept (in some sense) the hypothesis h which maximises $p(x|h)$. The origin of the theory is lost in the mists of time; Neyman (1967, p. 260) credits it to Karl Pearson, while Fisher (1930, p. 531) credits it to Gauss.

The method of maximum likelihood was the historical precursor of the likelihood principle. It is not the same as the likelihood principle (although the two are sometimes confused), but it furnished the conceptual tools that the likelihood principle uses.

Fisher (1921) gave the first clear statement of the method of maximum likelihood. The method starts by considering the likelihood function $p(x_a|h)$ with x_a fixed (at whatever was actually observed) and h variable. The maximum likelihood method then estimates h by picking the value of h which maximises $p(x_a|h)$. Fisher was to champion this method many times over the next four decades, and gave many examples of its use, although he

also advertised his other methods as superior to the method of maximum likelihood for certain purposes.

The method of maximum likelihood has two major problems. The first has been picked up by Hacking (1965) among others. The problem is this. Suppose the likelihood function following an experiment has the following shape:

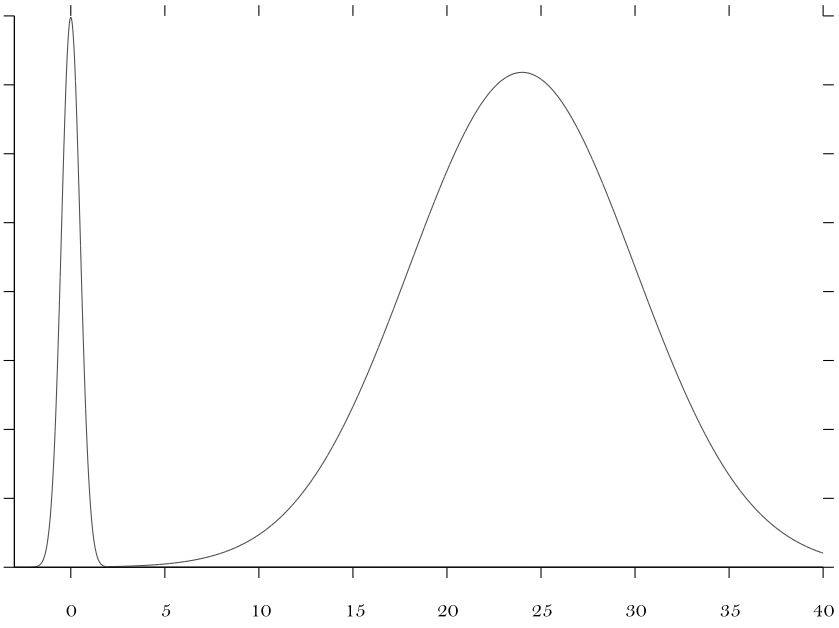

Figure 3: $p(x_a|\theta)$

Then the method of maximum likelihood will pick $\theta = 0$ as the best estimate of θ. In one sense this *is* the best estimate: it is where the likelihood function attains its highest value. But in another, probably more important sense, it is more reasonable to expect θ to be somewhere between 20 and 30. A Bayesian analysis would express this by saying that (given a reasonably flat prior probability distribution) the most important *single* value of θ is 0, but the range of values of θ between 20 and 30 is, in toto, much more likely

than $\theta = 0$. The same sentiment is harder to express for a non-Bayesian, but the fact remains, Bayesian or not, that the best estimate of θ may not be 0; and yet that is the estimate that the method of maximum likelihood would always recommend.

More specifically, if some hypothesis A is the most probable in the face of the evidence but hypothesis B is more likely to be near the truth (as might be the case if other reasonable hypotheses cluster around B but not around A, as illustrated by the choice of A = 6 and B = 16 in Figure 3 above) then A is not necessarily a better hypothesis than B. As Hacking puts a similar point,

> Speaking very intuitively for a moment, an estimate is good if it is very probable that the true value is near it. But an [sic] hypothesis is not best supported [in Fisher's sense] as it is probable or not that the truth lies near the hypothesis. To take a crude but instructive example, suppose there are six hypotheses about the value of A, namely $A = 0.90$ or 0.12 or 0.11 or 0.10 or 0.09 or 0.08. Suppose that the last five are equally probable—in any sense you care to give 'probable'—and that the first is slightly more probable. Then, if one may infer that the most probable hypothesis is best supported, $A = 0.9$ [sic] is best supported. But it is much more probable that A is near to 0.1, and so 0.1 may be the best estimate [contrary to the method of maximum likelihood. . . . This suffices] to establish a difference between 'best-supported' and 'best-estimate'.
>
> (Hacking 1965, p. 29).

This is not quite the same as my example, because Hacking uses probabilities of hypotheses, $p(h)$, where Figure 3 uses likelihoods, $p(x_a|h)$. But the spirit of the example is the same, and so is the conclusion to be drawn:

that the maximum of a probability distribution is not necessarily the best estimate.[54]

Hacking's own conclusion from this argument is that:

> [t]he best-supported hypothesis [the maximum likelihood estimate] is necessarily the most reasonable one, but that doesn't mean that one should behave as if it were true.
>
> (Hacking 1965, p. 28)

In other words (but still his words), Hacking claims that 0.9 is "the most probable value [of A] in the face of the evidence" and yet 0.11 is "more likely to be near the truth" (Hacking 1965, p. 29) and hence is a better estimate for many purposes.

This issue is very much clarified if we drop our quest for a single number to summarise the likelihood function, and instead use the *whole* likelihood function as our representation of what an observation tells us about a set of hypotheses. Hacking does not seriously consider this option, but Edwards (1972), following otherwise very much in Hacking's footsteps, does consider it. I describe this option further under *the method of support*, below.

Although graphs of the shape shown above are rare, essentially the same problem can occur with many other likelihood functions. For example, consider the old chestnut of estimating the maximum value of the numbers representing bus routes in a town, given only the information that one such route is numbered 75. Numbers lower than 75 have a likelihood

54. Hacking says only that the maximum likelihood estimate (0.9 in his example, 6 in mine) *may* not be the best estimate. As he correctly adds later (p. 62), "whether or not an estimate is good or bad may depend on the purpose to which it will be put". But even this is enough to establish that Fisher's method of maximum likelihood is not always the best inference procedure to follow.

of zero of being the maximum, while on any reasonable model numbers higher than 75 have progressively lower and lower likelihoods. (Two assumptions which are adequate to demonstrate this are (i) that the likelihood is monotonic and (ii) that no bus route is numbered 1,000,000,000 since that number would not fit on a bus's display panel.) It follows that the maximum likelihood estimate of the maximum bus number is 75; and yet that is not a good estimate, because we know for sure that it lies right at the bottom of the range of reasonable estimates.

A second major problem is that unrestricted maximum likelihood methods cannot be right, because they fail in the presence of a suitable evil demon hypothesis. Let H contain the hypothesis h_Ξ that an evil demon has magicked up my observation report from pure malice. Evil demons are infallible, so the probability of the observation result, conditional on h_Ξ, is 1. Hence h_Ξ is the maximum likelihood estimate. (It may not be the unique maximum likelihood estimate, but if the rest of H is scientifically plausible then it will be.) This is a reductio, because the method of maximum likelihood estimation is meant to determine which hypothesis we should infer the truth of, or act according to, or some such, and yet we do not want to take evil demons seriously. The problem is very easily solved, but the only obvious solution — namely, to disallow implausible hypotheses — is not only ad hoc but also subjective in precisely the sense in which Subjective Bayesianism is subjective. Indeed, the *most* obvious solution to the evil demon problem is to weight the members of H according to their plausibility — in other words, to adopt Subjective Bayesianism.

A third problem with the method of maximum likelihood is that it ignores information which is available but not presented as part of X,

x_a or H. This is also true of every other method in this survey except for Bayesianism and pivotal inference, and creates problems in every case, but the problems are particularly noticeable in the method of maximum likelihood because only this method tells us unambiguously to prefer a single hypothesis over all others. The problem is best illustrated by this example:

> If maximum likelihood is the only criterion the inference from the throw of a head would be that the coin is two-headed.
>
> (Jeffreys 1961, p. 383, citing Wilson 1952)

Any method of inference which counsels us to ignore information which is not presented as part of X, x_a or H will run into related problems; and all methods except for Bayesianism, pivotal inference and Shafer belief functions are of this type. I will illustrate this point further using the (more complicated) problems of Frequentist methods in chapter 7.

The method of maximum likelihood, as I have presented it here, is compatible with the likelihood principle. However, it is often used in conjunction with Frequentist methods which are not compatible with the likelihood principle. For example, Frequentist methods are often used to "understand and evaluate the precision of the maximum likelihood estimate" by seeing how it varies in hypothetical repetitions of an experiment (Basu 1975, p. 23). Such combinations of methods are occasionally referred to as "maximum likelihood estimation" in the literature (but not in this thesis).

THE METHOD OF SUPPORT

The **method of support** was first developed, as part of Bayesian statistics, by Jeffreys and Good, and was made into an independent school of statistical inference by Hacking (Hacking 1965), with notable refinements by Edwards (1972) and Royall (1997, 2004). It has been linked to fundamental issues in physics by Hilgevoord and Uffink (1991). I give Hilgevoord and Uffink's characterisation here as it is accurate and succinct and uses the same terminology as my chapter 2.

> The basic principles [of the method of support] are:
>
> a. All the information provided by the data x about the value of θ is contained in the [likelihood] function
>
> $$L_x(\theta) \equiv p_\theta(x)$$
>
> b. [Any strictly monotonic function of t]he ratio $L_x(\theta_0) / L_x(\theta_1)$ can be interpreted as a degree of relative support, in the sense that the data provide stronger support for θ_0 than for θ_1 if, and in so far as, this ratio exceeds unity.
>
> (Hilgevoord & Uffink 1991)

Principle b is a version of the law of likelihood. I discuss this principle further in chapter 8, where I show how it differs from the likelihood principle.

The method of support is a type of confirmation theory: that is, it is a theory about the extent to which the observation x supports the members of H, not a theory about which members of H are most likely to be true after we have observed x. The latter may depend on what we thought about h before we observed x while the former, at least according to the method of support, does not. As Hacking explains this dichotomy:

There are two quite distinct questions:

(1) Which hypothesis about the true value is best supported by current data?

(2) In the light of the data, which is the best estimate of the true chance?

(Hacking 1965, p. 28)

The method of support answers only the first question. The other methods discussed in this thesis answer the second question: they tell us which hypothesis we should approve (or, in Neyman's original theory, act on), using considerations beyond merely which hypothesis is best supported by the data x_a. One possible position is that such considerations are not germane; someone who took such a view would say that the method of support answers Hacking's question (1) and, a fortiori, his question (2). But none of the prominent exponents of the method of support take this view; they say that question (2) is beyond the scope of a completely general method of inference, either because it is badly posed in one way or another or because it requires subjective elements which are best added by individual consumers of statistical analyses rather than by statistical analysts (Edwards 1972, Berger & Wolpert 1988). If the beliefs of the consumers are filled out numerically in the most natural way then this latter view becomes equivalent to Subjective Bayesianism with the added constraint that the analyst must not report his own priors.

A tricky problem for the method of support is how to present the results of an analysis. Unlike the Bayesian, the likelihoodist cannot give the probabilities of the various hypotheses. Instead, a pure likelihood statistician can present the full likelihood function if the number of dimensions is

not too great, or she can present a projection of the full likelihood function onto a smaller number of dimensions if the number of dimensions would otherwise be too large. A likelihood function with two dimensions (one dimension of probability and one dimension for a parameter in the hypothesis space) can be drawn as a two-dimensional graph, like this:

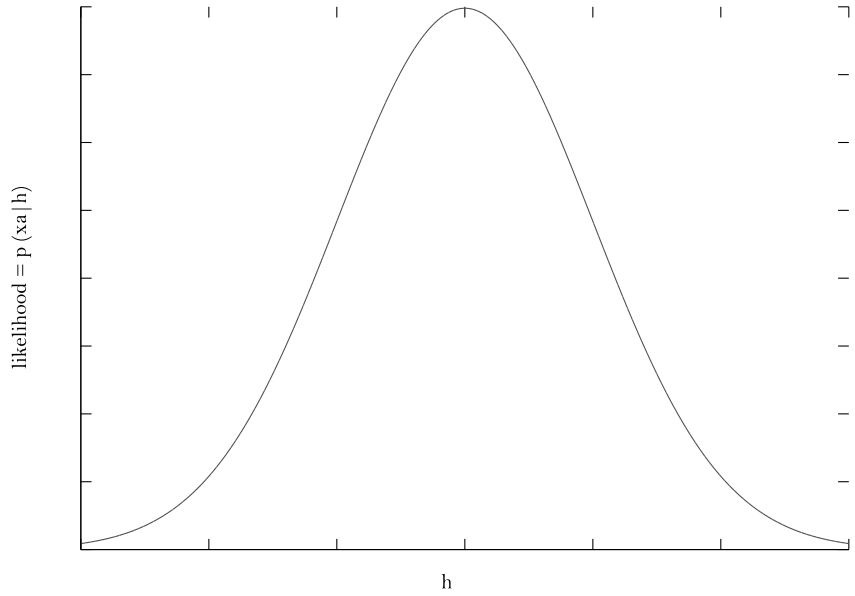

Figure 4: A two-dimensional likelihood function

Functions with three dimensions can be drawn in projection using tricks of perspective (although not by me!). But higher numbers of dimensions than that are tricky to draw. However, there is no theoretical limit to how many dimensions can be represented on a piece of paper. In a specific case, it is hard to know whether it can be represented in two dimensions or not. Much depends on the amount of data, the extent of redundancy in the data, the intended audience and the ingenuity of the analyst.

In the example below, a particularly ingenious analyst (Charles Joseph Minard) crams five dimensions into two without even using colour. The dimensions represented are: size of Napoleon's armies in their trip from Poland to Moscow and back again (width of main line — hatched to represent outgoing armies and solid to represent homecoming armies), position of army (two dimensions), time and temperature.[55]

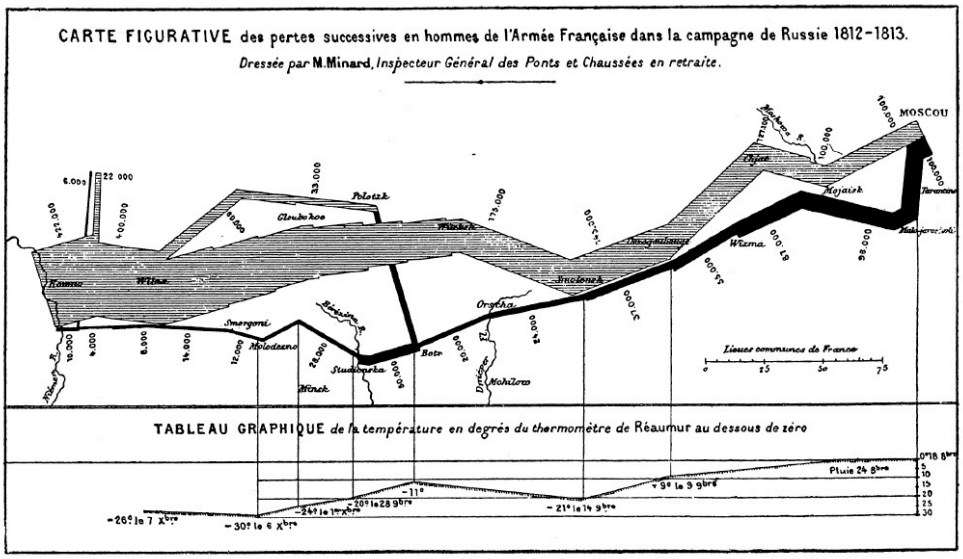

Figure 5: Napoleon's Moscow campaign (Tufte 2001, p. 40)

Still, the practical limit is three dimensions in most scientific applications, one of which must be reserved for the probability axis: therefore, it is very common that the complete likelihood function cannot be shown in a graph.

In any case, conveying a lot of information to an audience that prefers to receive only a little bit is refusing to answer the question of what the most relevant conclusions to be drawn from a piece of research are. The

55. Although the folly of war is beyond the scope of this thesis, it is remarkable that Napoleon managed to kill 75% of his own army by cold and starvation (Winterson 1988) before they even got to the city they had set out to attack.

audience reading a report of a piece of statistical inference wants to be given a simple answer not just as a summary of a complicated answer, but as an additional piece of information: "In addition to all the things you can tell me about the hypothesis space," they cry, "what is the major take-home lesson that ought to most strongly shape my future decisions?"

For these two reasons — inability to convey complicated information on paper, and the need to highlight the most relevant aspects of research results — it is often the case that the full likelihood function will not satisfy all audiences. But that does not put paid to the method of support. The best simple answer its adherents can give is still much better than a bare dichotomy: it is usually a comparison of the likelihoods of a small number of particularly salient hypotheses. Each comparison is simply a number: one likelihood divided by the other, in accordance with Hilgevoord and Uffink's rule b (Hilgevoord & Uffink 1991). Often this tells us everything we could reasonably want to know. Alternatively, high-likelihood regions (typically intervals) of the hypothesis space can be quoted. Other methods of reducing the likelihood function to something understandable are also possible (Basu 1975, pp. 23–25).

FISHER'S FIDUCIAL INFERENCE

No-one is quite sure what fiducial inference is, even though it is frequently mentioned. Its difficulty is due to the fact that its inventor said many interesting things about it but not enough of them to amount to a definition — not even an operational or pragmatic definition. Instead, he gave paradigm examples which seem to determine its use in some cases but not in others. This criticism of the "theory" of fiducial inference is widely

acknowledged — for example, by Seidenfeld (1979, p. 3) and by Barnett (1999, p. 299).

The fiducial theory *appears* to be as follows. (What follows is consistent with Fisher's most direct statements about the theory and, as far as I can tell, with his examples. It is contradicted by some of his more ambitious claims about the power of the theory.) Consider the maximum likelihood estimator of a parameter θ. This is a function of the possible observations X: $f(X,\theta) = \theta^{\max}(X,\theta) \stackrel{\text{def}}{=} \max_\theta\big(p(X|\theta)\big)$. Now we can invert this function: $\theta = f^{-1}(\theta^{\max}, X)$, and can use the value of the inverted function at the actual observation, $\theta = f^{-1}(\theta^{\max}|x_a)$, as the basis for probability statements about θ (Fisher 1930). The same method can also be applied to a sufficient statistic for θ, instead of to $\theta^{\max}$ (Fisher 1973). (See chapter 13 for a definition of "sufficient statistic".)

So the fiducial method is equivalent to the pure likelihood methods discussed above except for one unimportant difference (using a sufficient statistic of θ instead of using θ directly) and one important but naughty addition: the fiducial method regards the likelihood function as a probability function, which no other method does (see below for why not).

The relationship between the fiducial method and Bayesianism is instructive. A prior probability function is needed to normalise f after inversion (to make f to integrate to 1 by dividing it by a constant) to make sure it is still a probability distribution. So it is natural for a Bayesian to ask: what is the prior in Fisher's method? The answer is that there is none. The function is not normalised after inversion. This makes fiducial inference in most cases (not all) mathematically and epistemically equivalent to Objective Bayesian inference constrained to use a flat prior probability

distribution. Using a Bayesian method makes no sense to opponents of Bayesianism; and constraining the prior probability distribution to this extent makes no sense to Bayesians, for both philosophical and pragmatic reasons. Philosophically, the prior distribution is meant to represent something — uncertainty in most cases (Savage & discussants 1962), epistemic inertia (known in the literature on this topic as "conservatism") in others (Grossman et al. 1994, Freedman et al. 1983) — and a flat prior distribution has no degrees of freedom with which to represent any variation at all in these things from situation to situation. Pragmatically, a prior which is flat when measured according to one set of measures turns out not to be flat when measured against another set of measures such as the squares of the variables of the first part — this is essentially Bertrand's paradox (Jaynes 1973). As Rosenkrantz writes in a different context, "if we are ignorant of θ, the argument runs, then, equally, we are ignorant of $T(\theta)$. But if T is a non-linear function, like $T(\theta) = \theta^k$, a uniform distribution of $T(\theta)$ induces [is both mathematically and epistemically equivalent to] a non-uniform distribution of θ, and we have an obvious contradiction" (Jaynes 1983, p. xiv).

Thus the apparently flat prior probability distribution is only flat on a choice of measure, which is often ad hoc. This makes the whole fiducial procedure underdetermined by the epistemic and physical aspects of the situation it is meant to model. And as if that weren't bad enough, Stone has proved that using flat functions as priors leads to strict logical incoherence in some cases (Stone 1976). Many modern Bayesians avoid the use of flat functions to represent ignorance, for these and other reasons (especially since Stone's proof became known), but fiducial inference cannot avoid it.

In addition to these philosophical criticisms, there are mathematical reasons why we should not regard the likelihood function as a probability function:

- It does not integrate to 1.
- In some cases it cannot be made to integrate to 1 even by normalisation (division by a constant) because in some cases its area is infinite.
- In some cases it cannot be standardised by Fisher's preferred method, which was to subtract a constant from the likelihood function so that its maximum value becomes 1, because in some cases it has no maximum value (and, even worse, it may be unbounded).

To summarise, most authors disdain fiducial inference on the grounds that pure likelihood methods and flat-prior Bayesian methods embody conceptual mistakes (according to them); and the rest disdain fiducial inference on the grounds that it embodies mathematical mistakes (according to everyone except perhaps Fisher).

Since *nobody* approves of the fiducial method under its standard interpretation, I have to briefly discuss possible reasons for it having ever been taken seriously, or else it would seem as though I were hiding something. Good suggests that the reasons are Fisher's overwhelming personality, combined with the lack of clarity in the exposition of the theory:

> [I]f we do not examine the fiducial argument carefully, it seems almost inconceivable that Fisher should have made the error which he did in fact make [treating the likelihood function as if it were a probability function]. It is because (i) it seemed so unlikely that a man of his stature should *persist* in the error, and (ii) because, as he modestly says . . . his 1930 'explanation left a good deal to be desired', that so many people assumed for so

> long that the argument was correct. They lacked the *daring* to question it.
>
> (Good 1971, quoted in Barnett 1999, p. 306)

A third possible explanation for interest in the fiducial method is the belief that something fascinating and subtle lies buried in the method. All three explanations are reminiscent of explanations given for the popularity of Wittgenstein's writings, although no hidden fascinating subtlety has yet been found in the fiducial argument, whereas in Wittgenstein's writings many such have been found (albeit some of them mutually contradictory).

Insofar as fiducial arguments obey the likelihood principle they are either pure likelihood methods or Bayesian methods (if *any* of the current understandings of Fisher's arguments are correct!), so there is no need to consider them as a separate category in the rest of this thesis, in which I concentrate on the likelihood principle.

OTHER PURE LIKELIHOOD METHODS

Other pure likelihood methods, such as estimation using the mean of the likelihood function instead of its maximum, are possible but have never developed in detail or evaluated. I see no reason to think that any of them could fare better than the method of support.

2. PIVOTAL INFERENCE

Pivotal inference is mathematically fairly similar to Objective Bayesianism, but unlike any of the Bayesian schools of thought discussed above it is not guaranteed to obey the likelihood principle, even though it was invented

by the inventor of the likelihood principle, G. A. Barnard. So far it has proved of interest only to a very small number of theoretical statisticians.

Pivotal inference assumes that H is indexed by a parameter θ, and also assumes that we have a function $P(x, \theta)$ whose distribution as a function of x is independent of θ (a function with this property is called a *pivotal*) and that we have under consideration a family $\boldsymbol{D}$ of distributions for P. The choice of P is generally underdetermined. In particular, in models which have only location and scale parameters — which is the vast majority of the models currently used in science — "there are lots of pivotal quantities. . . . In general, *differences* are pivotal for location problems, while *ratios* (or products) are pivotal for scale problems" (Casella & Berger 2002, p. 427).

We then find a function of this pivotal which is an **ancillary statistic**. An ancillary statistic is any function of x which is not a function of θ or, more formally:

> $h = h(x)$ is called an *ancillary* statistic if [the probability distribution] f admits the factored form
>
> $$f(x, \theta) = g(h)f(x|h, \theta)$$
>
> where $g = g(h) = \text{Prob}(h(X) = h)$ is independent of θ.
>
> (Birnbaum 1972, p. 858)

With appropriate restrictions on $\boldsymbol{D}$, this ancillary is guaranteed to be maximal in the sense that any other ancillary statistic is a function of it. Call this maximal ancillary statistic $a(x)$. We then calculate the unique function $q(x, \theta)$ such that $a(x)q(x, \theta) = P(x, \theta)$ (Barnett 1990, p. 320; Barnard 1985, p. 58). If we have any prior information about θ, we state this in the form of a prior probability function $b(\theta)$. This ability to bring in prior information

gives pivotal inference some of the advantages of Subjective Bayesianism, although not the optimality property noted in chapter 3.

We then base our inferences about θ on the joint distribution of q and b conditional on a. When b is a full prior distribution for θ, this procedure is a form of Bayesianism, but when b is absent or only partial, pivotal inference depends on averages taken over parts of the sample space which were not observed, which is contrary to the likelihood principle and hence contrary to Bayesianism.

3. PLAUSIBILITY INFERENCE

Plausibility inference was invented and developed by Barndorff-Nielsen (1976) and (independently, but with less mathematical development) by Gillies (1973). Gillies recommends that we use plausibility inference together with Frequentist inference, while Barndorff-Nielsen recommends that we use it together with maximum likelihood inference; neither recommends that we use it on its own, so it is not clear that it deserves a section in this survey. It is unclear, also, whether it is intended to be a method of inference from data to hypotheses. Barndorff-Nielsen (1976, p. 116) says that it "pertains to the predictability of the data on the various hypotheses", and explicitly not "to how well the hypotheses explain the data". Nevertheless, I include it in this survey, to be on the safe side.

Plausibility inference compares the probability of an observation to the probability of the same observation under different hypotheses. So far, this is the same as maximum likelihood estimation. But, unlike maximum likelihood estimation, plausibility inference standardises these probabilities as follows:

$$\Pi_h(x_a) = \frac{p_h(x_a)}{sup_{x \in X} p_h(x)}.$$

Inference is then based on the maximum plausibility estimator, which is:

$$\check{h} = \{h : \Pi(h) = sup_{\theta \in \Theta} \Pi_\theta(x_a)\}$$

In terms of Table 1, the plausibility of each row is the probability which the hypothesis for that row assigns to the actual observation divided by the largest number in the row; and the maximum plausibility estimator is the row with the largest plausibility, or the set of rows which tie for first place if there is more than one with equal top plausibility.

Plausibility inference suffers from essentially all of the criticisms I make of Frequentist inference in chapter 7, as a result of its dependence on an unobserved part of the sample space X, except that it does not suffer from the ad hoc choice of test statistic which plagues Frequentist inference. On the other hand, it fails to have what most supporters of chapter 4 see as Frequentism's main advantage: it does not give us fixed "error rates".

4. SHAFER BELIEF FUNCTIONS

Glenn Shafer has proposed a non-Bayesian subjectivist theory of belief updating (Shafer:1976, Howson & Urbach 1993, pp. 424–426). This is a theory of personal beliefs, and has not been extended to a theory of applied statistical inference: unlike any of the other theories described in these survey chapters (with the possible exception of fiducial inference) it does not offer any *recipes* for moving from scientific data to scientifically

useful inferences about hypotheses. As Aickin writes, "one does not see applications of Dempster-Shafer theory directed toward practical problems of parametric inference" (Aickin 2000, p. 348). But many of the refinements necessary to make it into an applicable statistical theory could perhaps be borrowed, with some adjustments, from Bayesianism, in areas such as the construction of prior belief functions, the possibility of robustness theorems, and ideas about how to summarise a posterior distribution. Natural choices of these adjustments are liable to make Shafer's theory a form of Bayesianism (Aickin 2000), but it need not necessarily be so. So I will treat Shafer's theory here as if it were a theory of statistical inference, although I will not have anything to say about it elsewhere.

In Shafer's theory, the doxastic agent starts with a hypothesis space H, which (unlike H in any other theory except for some forms of Subjective Bayesianism) is taken to be exhaustive not only of the doxastic agent's partial beliefs but of all possibly true (or perhaps possibly believable) hypotheses, and for this reason Shafer refers to it as a "frame of discernment". The agent assigns a subjective *basic probability* to each set of hypotheses in H. These "basic probabilities" must sum to 1, and the empty set of hypotheses must receive "basic probability" 0, but otherwise they need not obey the probability calculus. Given these basic probabilities, a *belief function* Bel is then constructed on each subset $s \subset H$ by taking the sum of the basic probabilities assigned to s and all proper subsets of s. Bel is not required to obey the probability calculus either. The range $[\mathrm{Bel}(s), 1-\mathrm{Bel}(H \setminus s)]$ is known as the *belief interval* for s.

Given basic probabilities m_1 and m_2, Shafer calculates an overall belief function using a rule due to Dempster:

$$m \quad \propto \sum_{A,B \in H, A \cap B \neq \emptyset} m_1(A) m_2(B).$$

m_1 might describe hypotheses simpliciter and m_2 might somehow describe evidence: this provides a way to turn Shafer's theory of belief functions into a theory of statistical inference.

Aickin (2000) notes that Shafer's theory is (inappropriately) sensitive to the order in which beliefs are updated, and suggests additional axioms for the theory which fix this problem. Kyburg (1987) and Howson & Urbach (1993, p. 424–430) give a number of objections to Shafer's theory, notably the following:

> If you have equal degrees of belief in each of the numbers from 0 to 10 being called, then . . . you should not . . . have equal degrees of belief in the propositions '0 will be called' and 'A non-zero number will be called'. But in Shafer's theory you can[.]
>
> (Howson & Urbach 1993, p. 430)

So, Shafer's theory arguably does not contain sufficient constraints on beliefs to give them plausible identity conditions.

5. THE TWO-STANDARD-DEVIATION RULE (A NON-THEORY)

A method of statistical inference widely used by sciences in which observations are cheap (notably, large parts of physics) is to tentatively reject hypotheses according to which an observed data point is more than two standard deviations from its population mean. No theory of statistical inference can justify such a simple procedure, except as an approximation to

more complicated procedures. It survives nevertheless because, precisely in those sciences in which observations are cheap, tentatively rejecting a hypothesis only means collecting more data: the lack of major consequences of such an inference mean that justification can be treated more lightly than it can in the other sciences.

This procedure can be given an approximate Frequentist justification in many circumstances, and an approximate Bayesian justification in other (overlapping) circumstances. It therefore belongs in chapter 4 or chapter 3. I mention it separately here because it is often considered separately from its justification, and in that guise it belongs in neither chapter 4 nor chapter 3; but in that guise there is nothing philosophical to say about it.

6. POSSIBLE FUTURE THEORIES

It is tempting to treat these survey chapters as a menu from which we should choose the best form of statistical inference available, and many authors have done just that (although usually picking from a smaller menu, concentrating, quite reasonably, on the theories with the most detailed philosophical underpinnings, namely Subjective Bayesianism and Frequentism). It is tempting, but it is not what I will be doing; partly because I have another agenda, and partly because the criticisms I have mentioned of each theory strongly suggest that none of them is right as it stands, and it is possible (for all I can show) that none of them is right even in outline.

In this thesis I wish to show particularly that there are no good theories of statistical inference which do not obey the likelihood principle, so I will devote a chapter (chapter 7) to making it plausible that there are *insuperable* drawbacks to all the theories which both (a) have to date

been given some theoretical justification and (b) contradict the likelihood principle; all such theories fall into the Frequentist camp. It may seem a little unfair to have a whole chapter on objections to Frequentism while dwelling hardly at all on the objections to its main rival, Bayesianism; but it is not as unfair as it seems, because I do not see Frequentism and Bayesianism as exhaustive alternatives. My main claim (the likelihood principle) is in conflict with Frequentism, but that does not mean it supports Bayesianism: while compatible with Bayesianism it does not show it to be correct. The truth of Frequentism would imply that the likelihood principle is false, and so I pursue the criticisms of Frequentism to some sort of conclusion; in contrast, the truth or otherwise of Bayesianism does not imply the truth or otherwise of the likelihood principle, so I need not attack Bayesianism in detail.

The drawbacks of the other theories I leave as objections which may or may not be overcome in future versions of the theories. In some cases the objection is simply that no epistemic justification for the theory has been given. I cannot say much more about these existing theories of inference; but I can say something more about all future theories of statistical inference: I can classify them, ahead of time, according to whether they are compatible with the likelihood principle or not. This gives us the following picture:

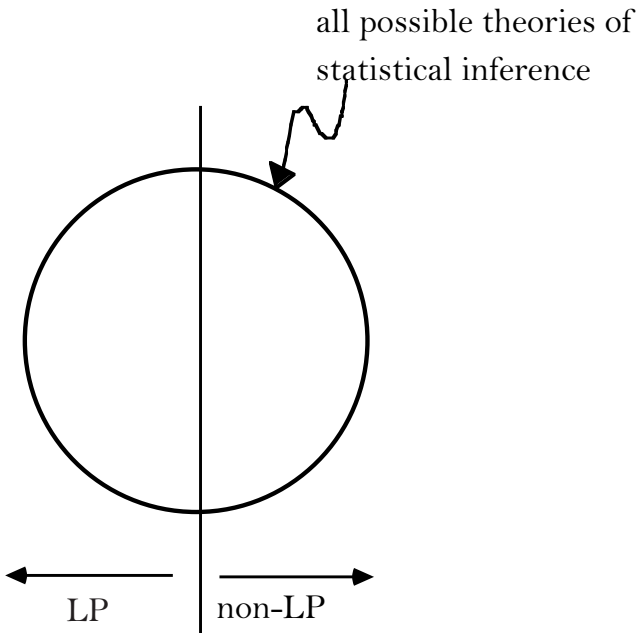

Figure 6

It seems to most authors, including me, that there is a useful dichotomy between Frequentist statistics on the one hand and Bayesian statistics (both subjectivist and objectivist) on the other:

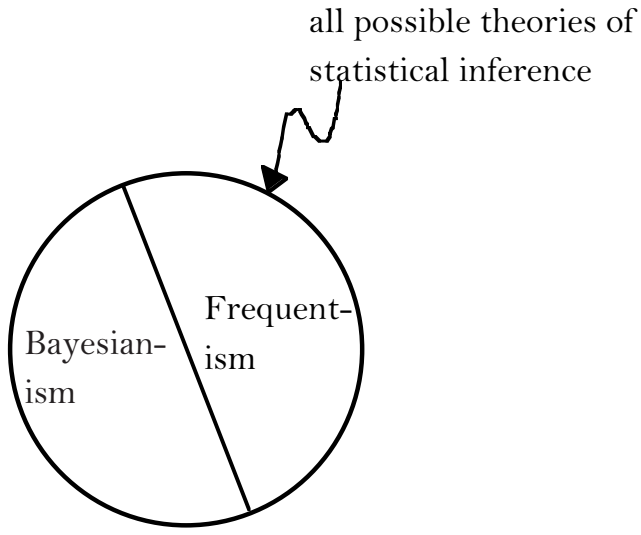

Figure 7

However, it would make more sense to show *gaps* between Frequentist and Bayesian statistics, since the diagram is meant to contain all possible theories of statistical inference, and there is no reason to rule out the invention of new theories which are neither Frequentist nor Bayesian. In other words, although Bayesianism contradicts Frequentism (as we will see in more detail later) it is not the logical contrary of Frequentism. The resulting diagram, incorporating these gaps, is as follows:

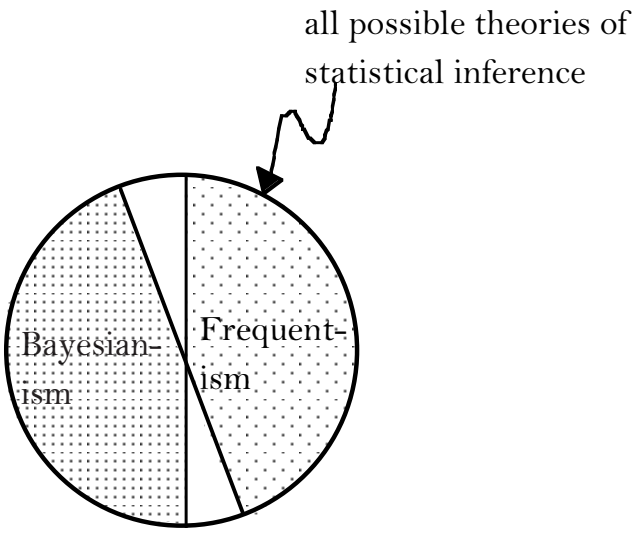

Figure 8

This can be put together with my first diagram in the obvious way:

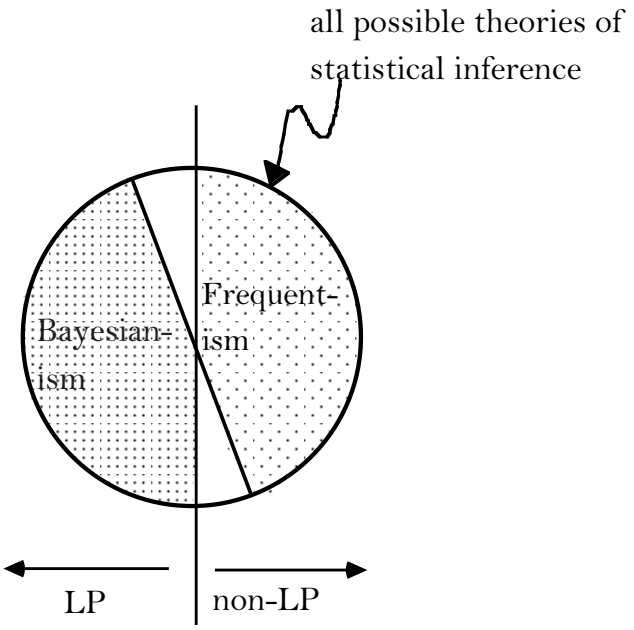

Figure 9

Note the very important point that the LP/non-LP line does not align with the Bayesian/Frequentist line. There are both Bayesian and non-Bayesian possible theories that obey the likelihood principle. There are also both Frequentist and non-Frequentist possible theories that do not obey the likelihood principle. There are, however, no Frequentist theories that obey the likelihood principle, and no Bayesian theories that do not (with the exception of some forms of Empirical Bayesianism, as discussed in chapter 3).

One of the main conclusions of this thesis will be that the best theory of statistical inference — a theory we do not yet have — may lie in the asterisked portion of the following diagram. It should obey the likelihood principle, and yet it need not be Bayesian.

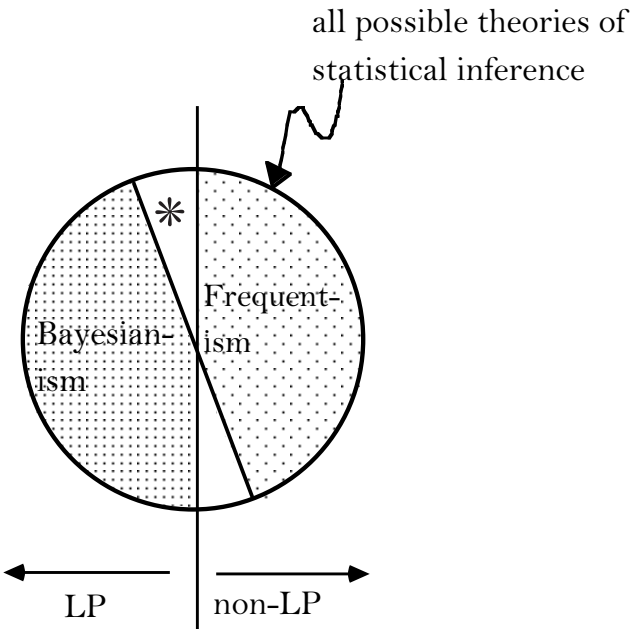
all possible theories of
statistical inference
*
Frequent-
ism
Bayesian-
ism
LP
non-LP

Figure 10

Part II

For and against the likelihood principle

— 6 —
Prologue to Part II

In Part II I will motivate, present and defend the likelihood principle, a principle which I defined roughly in chapter 1 and will define in more detail in chapter 8. This prologue to Part II outlines roughly how I will do that.

Recall Table 1 from chapter 1:

	possible symptoms			
hypotheses	vomiting (observed in this case)	diarrhoea (not observed in this case)	social withdrawal (not observed in this case)	other symptoms & combinations (not observed in this case)
dehydration	0.03	0.2	0.5	0.27
PTSD	0.001	0.01	0.95	0.029
anything else	0.001	0.001	0.001	0.997

Table 1

It is time to consider the meanings of the rows and columns of Table 1 in more detail.

The horizontal lines in Table 1 are meant to indicate that the probabilities are conditional on the hypotheses. Thus, the probabilities in the

first row are all conditional on the hypothesis that the child is dehydrated, and the probabilities in the second row are conditional on the hypothesis of PTSD. So each row is normalised (the probabilities add up to one). Although the rows need to be normalised, the columns don't, since the numbers in the table are conditional on the row headings (hypotheses) but not on the column headings (possible observations).

The symptom observed in a particular child is vomiting. Which hypothesis about that particular child does this observation support?

The Frequentist answer is that if we take dehydration to be the null hypothesis (which we should, as I will explain later) then we should reject this hypothesis in a falsificationist fashion, and hence either accept the alternative hypothesis that the child has PTSD or accept neither hypothesis. As we will see, a Frequentist stastitician would support this inference by quoting a type I error of 3% and a power of 97%.

Now recall the likelihood principle:

> **The likelihood principle**
>
> **Under certain conditions** outlined in chapter 2 and stated fully in chapter 8, **inferences from observations to hypotheses should not depend on the probabilities of observations which have not occurred**, except for the trivial constraint that these probabilities place on the probability of the actual observation under the rule that the probabilities of exclusive events cannot add up to more than 1.

The likelihood function of the actual observation is given by the first column in Table 1. So the likelihood principle entails that the falsificationist method which rejects the diagnosis of dehydration, which uses numbers

which do not appear in that first column, is not a good inference procedure. And although the likelihood principle does not *directly* address the question of evidential support, standard ways of applying the likelihood principle (insofar as there are such things yet — it is early days in this field) are likely to support a diagnosis of dehydration.

I will assume that we plan to analyse the table either by rows or by columns. What I mean by analysing "by rows" is restricting our attention to one or more rows. A typical Frequentist method does this by picking one particular hypothesis as being of interest (such a hypothesis is a row heading, known as the "null hypothesis"), and considering the values given by the table for various possible observations, only one of which is the one we have actually observed. I discussed this option in chapter 4, and will return to it in chapter 7. By analysing "by columns" I mean restricting our attention to the column corresponding to our known observation (which is represented by a column heading).

Analysing by rows is considering $p(e|h)$, the probability of evidence given a hypothesis, with h (hypothesis) fixed and e (evidence) variable, while analysing by columns is considering the same formula, $p(e|h)$, but this time with e fixed and h variable.

Why can we not analyse in some third way, perhaps with e and h both variable? There are two types of reason for not considering such methods here. Firstly, no well-worked-out version of statistics does so. But that is a bad reason — perhaps as philosophers we should consider possibilities that scientists have not got around to yet. Secondly, there is a better reason. We are forced to consider at least the two possibilities of analysing by rows and analysing by columns: forced to consider analysing by rows because that

is what the vast majority of statistical analysis does, and forced to consider analysing by columns by the proof which I will give later which shows that, on rather mild assumptions, it is the *only* rational analysis. So we have two ways of looking at the data, one of which we must discuss as philosophers of science because it is how scientists actually behave and the other of which we must discuss because it is how they should behave, at least some of the time. Of course I do not think I have demonstrated either of these points yet; I am foreshadowing the fact that I will be demonstrating them later merely to show that they are the two methods of analysis that we should be concentrating on.

Despite these points, there *are* other ways of analysing the table, and perhaps future work should have a look at the possibilities in the light of the fact that the assumptions under which I will prove an analysis by columns to be optimal are not always satisfied. In particular, when considering vague hypotheses — something, remember, which this thesis does not claim to do — looking at rows and columns simultaneously may make more sense than the proposal I will develop here.

I start Part II by motivating the need for statistical inferences about simple hypotheses to use probabilities which are conditional on the observation actually made (which is roughly equivalent to analysing Table 1 by columns). I do this, in chapter 7, by showing the various problems which Frequentist statistics encounter as a result of ignoring such conditional probabilities. Chapter 7 has a dual function: by showing the importance of conditioning, it motivates the likelihood principle; and at the same time, it disposes of the main rival to the likelihood principle, by showing that all well developed methods incompatible with the likelihood principle (all

of which happen to be Frequentist) are subject to major and insuperable problems. It is because of the neatness of this dual function that I delay any detailed discussion of the meaning of the likelihood principle until chapter 8.

Having motivated the likelihood principle by discussing the importance of conditioning on the actual observation, I present various versions of the likelihood principle (all of which entail that we should analyse Table 1 by columns) in chapter 8. I then give arguments against the likelihood principle, with counter-arguments (chapters 9 to 12). All of this will eventually be followed, in Part III, by an argument in favour of the likelihood principle and a case study on its application.

— 7 —
Objections to Frequentist Procedures

> *[Frequentist] theory is arbitrary, be it however "objective," and the problems it solves, however precisely it may solve them, are not even simplified theoretical counterparts of the real problems to which it is applied.*
>
> (Pratt 1961, p. 164)

In chapter 4, I defined Frequentist inference procedures. In this chapter I will say much more about how they work.

This chapter serves two functions for the thesis as a whole:

- Its primary purpose is to show that we should not look to Frequentist theories to provide the best theory of statistical inference, and thus that they do not provide good alternatives to the likelihood principle.
- Along the way, it will motivate the idea that the problem with Frequentist theories is that they are insufficiently *conditional*: that is, that they fail to fully condition on, or take into account, the fact that out of all of X only x_a has occurred. Having motivated this idea here, I will formalise it as the likelihood principle in chapter 8.

My tactics will be:

— first of all, to show that specific Frequentist methods are plausible but, despite their plausibility, both ad hoc and inferentially useless;

— to give examples of the failures of Frequentist methods;

— to question the objectivity of Frequentist methods, although I will conclude that their objectivity is not totally illusory;

— and then to diagnose the problem with Frequentism in terms of

  — over-reliance on counterfactuals and

  — the failure to condition on x_a.

1. FREQUENTISM AS REPEATED APPLICATION OF A PROCEDURE

Recall that the defining characteristic of Frequentist procedures is that they base all their conclusions on functions averaged over the sample space X. The rationale for this is the following principle (with the exact wording varying between authors, of course):

> A procedure for making inferences from data to hypotheses must have good average properties on repeated application in similar situations with different data.[56]

In a moment I will show how a suitable error set can be constructed; this will lead to the definition of the P-value, the commonest type of Frequentist statistic. I will then criticise the use of the P-value in statistical inference. Then I will state the definition of the confidence interval, the only other common type of Frequentist statistic, and criticise that.

Many of my criticisms will not rely on specific features of P-values and confidence intervals but, rather, will apply to Frequentist procedures

---

56. We might elucidate this definition by adding that a good frequentist procedure must have a low error rate, where an error rate is the proportion of times the procedure produces a conclusion which is incorrect in the sense of falling in some pre-specified error set, conditional on the truth of some hypothesis. But this does not really add anything to the definition, since there is no general definition of an error set.

in general. (I relate them to specific types of Frequentist procedures mainly for clarity of exposition, and to show that my criticisms are directly applicable to the types of Frequentist procedures in common use.) In the remainder of the chapter, I will diagnose two problems which underlie all the various criticisms: namely, firstly the inability of Frequentist methods to take into account all the information which is available at the time of analysis, and secondly an overreliance on hypothetical data which takes the place of the neglected actual data.

In subsequent chapters, I will evaluate a remedy to these problems: the likelihood principle. In order to begin to motivate this principle at the same time as critiquing Frequentism, I must briefly discuss why we should contrast the set of Frequentist procedures with the set of procedures which obey the likelihood principle. That is the task of the next section.

GENERAL FEATURES OF FREQUENTIST PROCEDURES

I will argue that the principle on which Frequentism rests (that a procedure for making inferences from data to hypotheses must have good average properties on repeated application in similar situations with different data) is misguided. It will not *immediately* follow from this that Frequentist statistical inference is wrong. It will, however, immediately follow that its distinguishing characteristic is no virtue; and from there it will be but a short step to seeing that other theories of inference are more rational.

A Frequentist inference procedure must incorporate functions of averages (possibly weighted averages) over the sample space (the space of possible observations, X) . . . or provably give the same result as one which

does. For example, recall from chapter 4 that the definition of a confidence interval is:

> If there exist functions of x, $T\downarrow$ and $T\uparrow$, both statistically independent of θ [see chapter 13 for a definition of statistical independence], such that
>
> $$(\forall\theta) \qquad p(T\downarrow(x) \leq \theta \leq T\uparrow(x)) = 1 - \alpha$$
>
> then the interval $[T\downarrow(x_a), T\uparrow(x_a)]$ is a $1-\alpha$ **confidence interval** for θ.
>
> (adapted from Kendall & Stuart 1967, volume II, p. 99)

Note that the probability statement in this definition uses statistics defined in terms of the members of X: that is why the observation is written as x (a variable, denoting hypothetical observations) rather than x_a (a constant, denoting the actual observation). Once we have found functions $T\downarrow$ and $T\uparrow$ which satisfy the probability statement, we switch our attention from averages over possible values of x to the actual value, x_a. This switch makes it a bit difficult to see what the probabilities are probabilities of: they are in fact probabilities of the required relationship holding between $T\downarrow$, $T\uparrow$ and θ in hypothetical repetitions of the merriment. That is why it is a Frequentist definition.

Frequentist inference procedures can be contrasted with **likelihood procedures**, by which I mean those which obey the likelihood principle. Plausible likelihood procedures always have some justification other than their behaviour on repeated application.

It may appear from this discussion as if the main difference between Frequentist and likelihood procedures is that Frequentist procedures retain their properties in long runs of experiments while non-Frequentist

procedures do not. This is a good way to think about the difference for most purposes, but it is not entirely accurate. Non-Frequentist procedures can be repeated just as easily as Frequentist procedures can. The difference is not whether these properties can or can't be evaluated, or whether they are or are not important. The difference is more subtle than that. It is that if a Frequentist inference procedure is to be acceptable on the basis of its Frequentist justification then it must be evaluated according to and *only* according to its properties when repeated with imaginary random data (plus, for pragmatic reasons, the mathematical tractability of its equations). A Frequentist procedure *must* be evaluated in this way, while a likelihood procedure need not be (depending on the intended audience).[57]

I said earlier that Frequentist procedures can be contrasted with likelihood procedures. This is because there are no generally applicable types of statistical inference procedure which both obey the likelihood principle and have good Frequentist properties. There are, however, certain token statistical inference procedures which can be considered to have both a reasonable Frequentist justification and a reasonable likelihood justification (Deely & Lindley 1981). Such procedures only keep this confusing property on some values of their parameters. For example, the procedures used to calculate P-values are not generally compatible with the likelihood principle. This can be shown in many ways; for example, by Lindley's proof

57. Although any likelihood procedure *can* be evaluated according to its Frequentist properties, in order to find out whether it can please a Frequentist audience, likelihood procedures are very rarely evaluated in this way in the literature. I suspect that this is for the following rather strange reason. According to the most vociferous non-Frequentist school of thought, Bayesianism, there is a *unique* optimum inference procedure for any given (fully-specified) statistical model. Bayesians never vary from this optimum inference procedure (except for small variations made for mathematical convenience). Since their optimality (in Bayesian terms) is preordained by their method of construction, it is rarely necessary to evaluate their actual properties on repeated use. For example, the Bayesian method of conducting pharmaceutical trials had not been evaluated according to Frequentist criteria until (Grossman et al. 1994).

that "for any classical significance level for rejecting the null hypothesis (no matter how small) and for any likelihood ratio in favour of the null hypothesis (no matter how large), there exists a datum significant at that level and with that likelihood ratio" (Edwards et al. 1963, p. 219). And yet some exceptional P-value calculations are compatible with the likelihood principle.

Such exceptional instances are only both good qua Frequentist procedures and good qua non-Frequentist procedures if they have coincidentally suitable values of the hypothesis space *h and* the sample space *X and* other parameters of the non-Frequentist procedure such as (for a Bayesian analysis) the prior distribution and the utility function. Even then they not only have different *justifications* considered as Frequentist or likelihood procedures, they also have different *interpretations*, and hence scientific consequences, considered in these two ways.[58]

USES OF ERROR RATES: EXPECTANCY VERSUS INFERENCE

Hacking's (1965) work on statistical inference suggests that we should distinguish between two very different uses of the error rates which characterise Frequentist statistical procedures. One use is in calculating what our *expectations* of the average performance of a statistical procedure should be. The other use — the one I am concerned with in this thesis — is the use of error rates to perform statistical *inference*, by which I mean inference from observations to hypotheses.

Hacking equates these two uses of error rates with uses of error rates respectively *before* and *after* an experiment has been conducted. This

58. I will give an example of the different interpretations afforded to extensionally equivalent Frequentist and non-Frequentist inference procedures in chapter 15.

makes some sense: before we have data, we are likely to want to calculate the average performance of a statistical procedure which we are planning to use, whereas once we have data we should ignore such average figures in favour of evaluations of the actual performance of the procedure on the actual data. However, I see two problems with equating the expectation-versus-inference dichotomy with the before-versus-after-experiment dichotomy.

The first problem is that, as usual in epistemology, time is not an important factor in its own right; it is a proxy for what order an epistemic agent learns things in. Thus, instead of talking about *before* and *after* collecting the data we should be talking about whether the collected data is available at the point when the statistical procedure is evaluated. This translates into on the one hand taking the data into acount in its own right (as x_a) and on the other hand taking the data into account merely as a representative of some function of the data space X. The former option translates directly into the likelihood principle;[59] the latter option is the definition of Frequentism.

A second problem with the pre- and post-experimental dichotomy is that, as I argued in chapter 2, there is no need to assume that all data which leads to a statistical inference comes from experiments.

For these two reasons, I will not be using Hacking's insight in its raw form, but rather in the guise of the likelihood principle.

---

59. I will return to the likelihood principle at the end of this chapter, where I will offer factualism (acceptance of the likelihood principle) as an alternative to Frequentism.

2. CONSTRUCTING A FREQUENTIST PROCEDURE

My main criticism of Frequentist inference, namely that it is countermanded by the likelihood principle, is completely general: it does not depend on the specific features of any particular Frequentist method. Despite this generality, we will need examples of Frequentist inference procedures, for clarity. There is more than one important subtype of Frequentist inference procedure. I will give a philosophical exposition of a class of inference procedures which is and has always been by far the dominant form in both theoretical and applied Frequentist statistics: the P-value. I will also briefly discuss the second-most-influential form of Frequentist inference, "confidence" intervals (whose name is misleading, as we will see). Between them, these two types of Frequentist inference procedure make up most of the work of contemporary applied statisticians. I will make no attempt to discuss any other Frequentist inference procedures in specific terms, but I will give my criticisms of the procedures I do explicitly discuss in a form which applies to all Frequentist inference procedures as far as possible, and the final conclusions of this chapter will be stated in a form which uses only those parts of my argument which do apply to all Frequentist inference procedures.

In the next few sections, I will offer for consideration a function which will stand as a candidate for use in Frequentist analysis. Rather than starting with one of the procedures defined in chapter 4, I will construct such a function from scratch. In this way, we will see clearly what issues of justification arise.

PRIVILEGING A HYPOTHESIS

By Frequentist lights, a procedure is evaluated according to its performance on repeated application with different observations. In order to give it fixed properties on such repetitions, we start by fixing our attention on a single hypothesis of interest and comparing the probabilities in the row of Table 1 (see chapter 1 or insert) designated by that hypothesis.

But now we are in trouble already. Fixing a single hypothesis can, in general, be criticised for being ad hoc. This ad hockery is side-stepped — or SUTC, for "swept under the carpet", according to Good (1976) — by noting that in most cases there is one hypothesis which it would be particularly disastrous to believe were it false.[60] So in these cases it is less than totally ad hoc to single out a particular hypothesis. But this is not a convincing justification for always doing so. In contrast, analysis in line with the likelihood principle does not require (or, indeed, allow) us to privilege a particular hypothesis. Instead, it requires us to privilege a particular one of the possible vectors of observations . . . but that is easy: we privilege the actual one, x_a. It practically privileges itself.

---

60. In the case of clinical trials, for instance, it would be particularly disastrous for a drug company or a regulatory authority to believe that a drug worked when in fact it did not, for obvious reasons (displacement of better drugs, side-effects, litigation). Believing that a drug was inefficacious when in fact it was efficacious, on the other hand, is much less damaging from everyone's point of view, especially when we bear in mind that similar chemicals are likely to be tested later and correctly found to be efficacious. (Drug companies always test many related chemicals when they sniff any possibility of being on to a good thing.) Similarly, in Table 1 it is more important to make sure not to miss dehydration than to make sure not to miss PTSD, since a child with undiagnosed PTSD will probably live to be rediagnosed another day whereas a child with undiagnosed dehydration will almost certainly die quickly.

CALCULATING A FREQUENTIST ERROR RATE

Having picked a privileged hypothesis, h_0, we need a suitable way of calculating an *error rate*:

> **Error rate**: The proportion of times an experiment gives an answer that falls into some predefined error set, if repeated infinitely or indefinitely.

If such a number is small then we have an observation which is unlikely according to h_0. That in turn seems to speak against h_0. This last step appears obvious to most writers on statistics. It is vitally important, so I will give it a name:

> **The unlikely events principle**:
> A hypothesis which assigns a low probabilities to an event is disconfirmed by the occurrence of that event to the extent that, if a hypothesis says that an event is unlikely, and yet that event occurs, it is reasonable to conclude, at least tentatively, that the hypothesis is probably false.

This principle is related to Cournot's principle (sometimes attributed to Kolmogorov or Popper), which says that "certain events [those with low probabilities] are so unlikely as to be 'essentially impossible' " (Sorkin 1983). (Thanks to Alan Hájek for this point.) According to this principle, a hypothesis can be falsified by predicting that an actual event is improbable. Cournot's principle is not the only way to justify the unlikely events principle, and it may be better to see the unlikely events principle as primitive. In any case, the unlikely events principle is essential to the

standard justification of Frequentist procedures. Recall that Fisher, when defending Frequentist tests, wrote:

> The force with which such a conclusion [rejection of h_0] is supported is logically that of the simple disjunction: *Either* an exceptionally rare chance has occurred, *or* the theory of random distribution is not true.
>
> (Fisher 1973, p. 39)

The disjunction itself is trivially true; but it would not have any connection to the first part of Fisher's statement, about drawing conclusions, were it not for the unlikely events principle.

I will return to this principle later. First, let us find a plausible way of calculating an error rate.

First candidate error rate

The first thought about how to calculate an error rate, both historically and, perhaps, in the mind of the reader, is to calculate the probability that we would have seen the observation if the null hypothesis were correct — $p(x_a|h_0)$. Using this as an error rate means using $\{x : p(x|h_0) < 5\%\}$ as the error set. (As is well known, the figure 5% is ad hoc and could just as well be replaced by some other figure, so there is an ad hoc element in this suggestion.) This is my first candidate error rate. It is the most straightforward reading of Popper's idea of subjecting a hypothesis to a severe test (Mayo 1996). It is also what some prominent philosophers of science, including Mayo (2000, p. 181) at least sometimes, think statisticians do.

But statisticians *never* use $p(x_a|h_0)$ to calculate an error rate, and for a very good reason. Suppose for the moment that only a finite number of observations is considered possible. The probabilities of all the observations put together must be 1, so if there are a lot of possible observations — and there usually are — then most of them must have low probabilities. If their probabilities are not too wildly different from each other then it follows that *all* of them must have low probabilities. In that case, every hypothesis will be falsified by *any* piece of evidence. This is true *whenever* there is a large number of possible observations of roughly equal probability, and also in many of the cases typically found in applied statistics, including some cases in which the observations are not of roughly equal probability. (There is a trade-off between the variability of the observations and their number.)

For example, I toss a coin twenty times and record the exact sequence of heads and tails, and then consider $p(x_a|h_0)$ where h_0 is the hypothesis that the coin is fair. Then *no matter what* x_a *is*, $p(x_a|h_0)$ is less than 0.000001. (For example, $p(HTTTTHTHTTTHTHHTHHTH) = (\frac{1}{2})^{20} < 0.000001$.) This will not do as an error rate with which to evaluate the procedure (even though technically speaking it is a perfectly valid error rate), because it cannot be large no matter what x_a is.

Now let us turn to the case in which infinitely many possible observations are under consideration. This case is also very common: it typically occurs when we want to estimate some parameter θ which takes values from the real numbers (or, more generally, from $\mathbb{R}^n$). For example, we may want to test the hypothesis that a particular foetus is at normal weight

for its age from the abdominal measurements of its mother. The hypotheses will have the form $p_{h_0}(\theta|\text{abdominal size}) = f_{h_0}(\text{abdominal size})$, where f_{h_0} is some appropriate probability density function such as a log-Normal distribution. Now if we calculate the formula $p(x_a|h_0)$, we find that it is always zero: in order to account for the fact that infinitely many foetus sizes are possible for each abdominal size, the formula assigns a value of zero to each. (Some would say that it *has* to assign a value of zero to each; others only that it typically does. The dispute there turns on whether probability density functions may incorporate delta functions to represent "lumps" of probability. Berger & Sellke (1987) argue that they may and often should; a more orthodox Bayesian viewpoint is that they must not, since delta functions are not strictly functions. Either way is fine for my argument.) The reason why foetus sizes are assigned probabilities of zero is that probabilities of hypotheses correspond to areas on the following graph, and the probability of a point hypothesis (the hypothesis that the foetus is exactly θ long) is the area of an infinitely thin slice of the graph. (Of course we could avoid this problem by restricting ourselves to discrete values of abdominal size instead of all the values in $\mathbb{R}$, as we could with any infinite sample space; then we are back in the position discussed in the previous paragraph.)

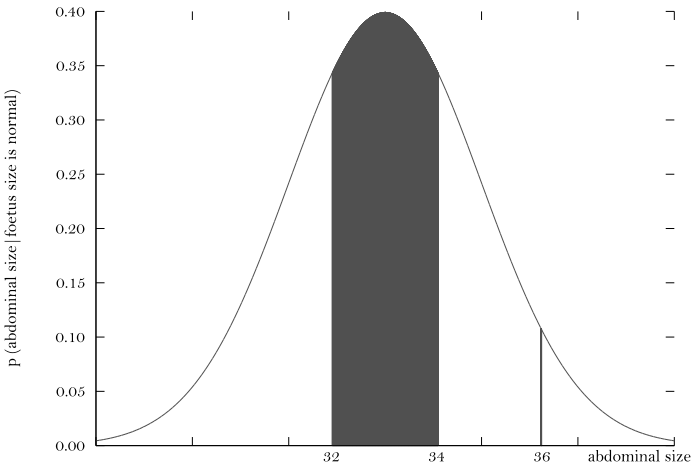

Figure 11: probability of abdominal size under h_0

This function never has a value of zero, and yet each slice (e.g. at 36) has zero area. This is what we mean by calling the function a probability *density* function.

Hence, our first candidate for calculating an error rate does not work when the number of possible observations is large or infinite.

Second candidate error rate

Why not use the *height* of the graph at the relevant place, instead of the area of the slice, to construct an error rate? This is our second candidate error rate. A mathematician's immediate thought would be that this candidate makes no sense, because the height of the graph (in the continuous case) does not represent the probability of anything — the graph is a representation of a function designed to integrate to a probability, not to represent a probability directly. Let us put such squeamishness to one side, and take the height seriously for a moment to see what happens. We cannot simply take the height as a probability, because the height of the graph may be more than 1 in places, but we can think of various proposals

to fix that problem: perhaps, for example, we can reduce all the heights by subtracting or multiplying by a constant. This has been suggested by Fisher (1973, p. 76) and Edwards (1972).

If we implement some such strategy to make sure the height of the graph never exceeds 1, we get a procedure which, as far as I can see, is satisfactory according to the logic of falsificationism. But it is not satisfying simpliciter. Firstly, the answers we get in typical situations are still counter-intuitive: the coin-tossing example gives a graph with a constant *height* of 0.000001 as well as with a constant *probability* of 0.000001. A second problem with this proposal which is more severe, although less ubiquitous, is that sometimes the graph has no maximum value. (The graph of $ln(x)$ from $x = 0$ to ∞ has this property, for example.) Then there is no such way of preventing the height of the graph from exceeding 1 (Bayarri et al. 1987). So there is no natural way of turning the height of the graph into a probability; and hence there is no natural way to use it to calculate an error rate. This is not a conclusive argument against using the height of the graph in some way, but it is a reason to look elsewhere for our error rate.

Grouping possible observations

The obvious next move is to avoid having to consider the case of the large hypothesis space which has proved so difficult, and to do this by grouping the large number of possible observations into a small number of clumps. This makes complete sense from the purely mathematical point of view, but it is worryingly dependent on a choice of grouping strategy, and different grouping strategies give very different conclusions. If we group the possible observations in large clumps then we do not distinguish

adequately between importantly different pieces of data. If we group in small clumps then there are too many clumps, so again the answers we get for typical hypotheses are unacceptably counter-intuitive (again, we find that many perfectly reasonable hypotheses are falsified by *any* observation). Moreover, we may get different results for different clumpings. What we need is a grouping strategy which is in some sense natural and which does not give counter-intuitive results.

Such a grouping strategy is available, at least in most cases. It is this: group the observation which was actually seen together with all possible observations that are in some mathematical sense more extreme than it. Typically this is a tail area, $p(x \geq x_a|h_0)$, as shown below. This is the grouping strategy which is, and always has been, used in almost all Frequentist analyses.

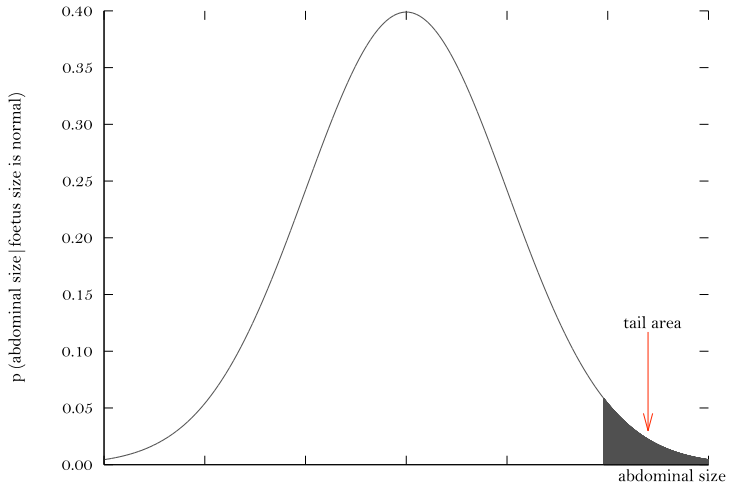

Figure 12: a tail area representing $p(x \geq x_a|h_0)$

If you teach statistics to bright undergraduates, you find that occasionally a student asks, "Yes, but *why* do we calculate the tail area?" I know of only two justifications for this grouping strategy, and only one of them

makes it seem more than ad hoc. The first justification is that it works: it gives answers which, by and large, are not counter-intuitive. The other justification — the less ad hoc one — is that it can be shown mathematically that this grouping strategy gives the same results as a Bayesian analysis of the data in a variety of moderately common cases (Deely & Lindley 1981). Accepting the correctness of a Bayesian procedure is the only honest way I know to answer the student's question. Should the student not want to accept Bayesianism then I can see no answer to her question (and should the student want to accept Bayesianism, she would presumably see no reason for calculating Frequentist error rates at all).

Despite its ad hocness, this is the path that applied statistics has taken: one calculates tail areas. Since no-one (to the best of my knowledge) has suggested any other completely general way to instantiate error rates statistically (apart from confidence intervals, discussed below), we will have to take this as a given for the moment.

This solution can be applied in the discrete case too: again, one takes tail areas — $p(x \geq x_a|h_0)$.

A better candidate error rate: the P-value

The previous section has motivated the use of tail areas in calculating error rates. It remains to find a formula for doing this.

Recall the definition of an error rate:

> **Error rate**: The proportion of times an experiment gives an answer that falls into some predefined error set, if repeated infinitely or indefinitely.

My tentative candidate for an error rate is as follows. First of all we calculate a number called $\mathcal{P}$, thus:

> $\mathcal{P}(h_0, T(x_a))$ is defined as the proportion of an infinite sequence of hypothetical experiments, each duplicating the experiment we have actually conducted, on the assumption that the hypothesis h_0 is true, that would result in a value of $T(x_i)$ greater than or equal to $T(x_a)$, where x_i is the observation made in each hypothetical experiment, x_a is the observation made in the actual experiment, and T is an arbitrary function from the space of possible observations to the real numbers.

Then we reject h_0 if $\mathcal{P}$ is less than some fixed value p_0.

According to the falsificationist thinking of Neyman, if $\mathcal{P}$ is sufficiently small then we should reject the hypothesis h_0.[61] Interestingly, what we should do if the number is *large* (close to 1) is not quite as widely agreed. It is standard practice among scientists nowadays to take a large number as evidence in favour of h_0, but Neyman, who was a very orthodox falsificationist (and who was responsible for promoting falsificationism in science, and who by the way was much more influential in this project than Popper was) believed that the size of $\mathcal{P}$ should make absolutely no difference, except that we should note whether it was on one side or the other of the agreed cut-off. Meanwhile, Fisher took the view that a large value of $\mathcal{P}$ was evidence *against* h_0! Fortunately, all Frequentist schools of thought agree that finding a small value of $\mathcal{P}$ should lead us to reject h_0, so

---

61. This is a convenient simplification. What he actually says is that we should reject h_0 if and only if $p(x_a|h_0)$ falls into some small predefined error set. My discussion above explains why this error set was and is always taken to be a tail area, even though according to Neyman's theory it need not be. Salsburg (1989) gives arguments showing that defining the error set in any other way will not help; in chapter 13 I give more general arguments for the same conclusion by proving the likelihood principle.

I will concentrate on that eventuality for the moment and deal with large values of $\mathcal{P}$ later.

CHOOSING A TEST STATISTIC (T)

Why do we need the function T in the definition of $\mathcal{P}$? The important point to remember is that x_i are *vectors*, and typically high-dimensional vectors at that. A typical x_i in medical research is a very large, complexly structured vector part of which might look like this:

<age of subject 1: **801 months**,
initial tumour histology for subject 1: **t35**,
initial treatment for subject 1: **radiotherapy**,
size of subject 1's tumour at 6 months: **unknown**,
side-effects at 6 months: **unknown**,
size of subject 1's tumour at 13 months: **11 mm**,
side-effects at 13 months: **unknown**,
adjuvant treatment for subject 1: **chemotherapy**,
site of subject 1's secondary tumours: **leukemia**,
. . .
age of subject 2: **684 months**,
initial tumour histology for subject 2: **q+**,
initial treatment for subject 2: **none**,
. . .
age of subject 3: **787 months**. . . >

We have already seen that we need something like a notion of one value of X being more extreme than another. In other words, we need the

set X of possible outcomes to be *ordered*. But there is no natural sense in which this whole n-tuple is bigger or smaller than another one (except in the vanishingly rare, trivial case in which one has bigger numbers in every dimension than the other).

There is no general, non-arbitrary notion of one vector being bigger than another (or more extreme in any other sense). If one thinks of vectors in Euclidean three-space, there is an obvious sense in which one vector *is* bigger than another, namely when the natural metric $||x|| = \sqrt{x_x^2 + x_y^2 + x_z^2}$ assigns one vector a greater length than the other; but that relies on the fact that dimensions in Euclidean space are commensurable with each other, in the sense of being merely rotations of each other. Statistical sample spaces are not at all like this (at least, not usually). x_i above is a vector in a very general sense: it is an n-tuple of observations, each of which can be of any observable type at all.

To make matters even more complicated, the vectors in the sample spaces of clinical trials don't even have the same number of dimensions as each other, since the various vectors represent different possible outcomes in which different numbers of subjects have been recruited and followed up. (Thus, the vectors are not even in the same vector space, unless we artificially extend some of them with zeros.) Traditionally the sample space in Frequentist inference is restricted to samples of a fixed size (although I am aware of no philosophical justification for this — indeed, it seems inconsistent with Neyman's basic theory), but this maneuvre is not possible in large clinical trials, in which the results are analysed as they are collected (for scientific, ethical and legal reasons — more on this in chapter 15) and in which conseqently there can be no fixed sample size.

So, how can we compare vectors which are not naturally comparable to each other? Only by allowing the statistician to introduce an arbitrary function T which reduces each vector to some ordered quantity (almost always a single real number). $T(X)$ is known as a *statistic* or *test statistic*.

The raw data can fail to be commensurable with each other even if x_i is scalar (not a vector). Consider the set of possible observational outcomes $X = \{\text{a sheep, a cow, a goat}\}$. For the sake of argument, suppose we observe a sheep. How are we going to obey $\mathcal{P}$'s requirement that we consider the probability of observing the actual observation or something more extreme? Is a goat more extreme than a sheep? Perhaps this is a bad example, since a goat *is* more extreme than a sheep, but the moral is clear: these comparisons are artificial. In order to analyse the result of this experiment using $\mathcal{P}$, we need to explicitly introduce a function T which maps the set X of possible outcomes to an ordered set.

Another use for T is to allow adjustments to be made for data which have not been measured for one reason or another, such as because a trial subject cannot be contacted. These adjustments are known as "censoring", and the mathematical problems they cause are of major concern in the literature. I discuss censoring in chapters 9 to 12.

Historically, a more important function of T used to be to simplify the data for computational purposes. But since the 1980s computers have been fast enough to alleviate the need for this in most cases. T is still considered important by statisticians for four reasons: (i) for the philosophical reasons given above; (ii) because computers are not fast enough to analyse unreduced data in *all* cases; (iii) because humans like to be able to make

pencil-and-paper approximations to the calculations their computers are making; and (iv) as a matter of historical inertia.

For all we have seen so far, T could be completely ad hoc; and in many cases it is. But there are principles which constrain the choice of T to some extent. In particular, in many cases it is possible to choose a "uniformly most powerful" statistic (one with the highest possible power for every value of θ). Despite the name, uniformly most powerful statistics are not always the best statistics to choose. One reason for this is that "[i]t is possible for an outcome to be significant at one level but not at a less extreme level by uniformly most powerful tests" (Pratt 1961, p. 166, citing Lehmann 1959, p. 116). But I do not need to insist on that point, because in some situations a uniformly most powerful statistic cannot be chosen because there isn't one.

That T suffers from an ad hocness problem should not be surprising, given the example of the sheep and the goat. The fundamental nature of the problem is very simple: use of the procedure $\mathcal{P}$ relies on our possible observations being ordered; but there is no general reason to think that our possible observations should be ordered, in any sense that has any epistemic importance, and often they patently are not; so in order to make $\mathcal{P}$ work at all, we have to *artificially* order the elements of X. Now, if we were doing that merely for some presentational reason — for example, to print them out into a readable table — this would be a very minor problem. But the ordering of X is buried deep in the analysis. $\mathcal{P}$ does not first tell us that h_0 is (or isn't) believable and then present the results in terms of an ordering of X; it tells us that h_0 is (or isn't) believable *simpliciter*. But to decide whether it is (or isn't), $\mathcal{P}$ uses an ordering of X.

To translate this into concrete terms, someone who orders goats above sheep will come to one conclusion about the farmyard experiment, while someone who orders sheep above goats will come to a different conclusion; and their ad hoc assumptions about farmyard hierarchies are typically hidden from each other. So the ordering of X is important but is not available for perturbation analysis. (Perturbation analysis is the testing of mathematical variations on assumptions to see what difference that makes to conclusions.) If their assumptions were not hidden from each other, the source of their disagreement would at least be clear; but there would still be no natural way to resolve it.

Approaches to statistical inference based on the likelihood principle (such as Bayesian statistical analysis) have no general need for test statistics. This difference between likelihood and Frequentist approaches is often misunderstood or misleadingly described in the literature, when it is noted at all. For example, Barndorff-Nielsen writes (about his own methods, which are not strictly Frequentist, but which do require test statistics; see chapter 5 for details):

> In most cases [my methods] should be applied not to the original data x and the model for x, but to some suitable statistic T of x, and a derived model for T, and the [question] is that of which T and derived model are to be considered. Stated briefly the answer is that, before estimates and tests are computed, the inference problem should be purged for [sic] irrelevant features by such means as margining to sufficient and conditioning on ancillary statistics. It must be emphasised that likelihood inference is subject to a similar qualification.
>
> (Barndorff-Nielsen 1976, p. 105)

This seems to suggest that a scientist, whether using Frequentist or likelihood methods, must purge her model of irrelevant features before statistical inference can be conducted, using difficult procedures such as conditioning on ancillary statistics (defined in chapter 5, and again below). But in fact that is the case only for Frequentist methods (and closely related methods such as Barndorff-Nielsen's own), not for methods obeying the likelihood principle. Barndorff-Nielsen continues:

> There is the difference though that in likelihood inference often part or even all of the necessary purging is taken care of automatically because factors, of the original likelihood function, depending only on the observations and/or on possible incidental parameters do not influence estimation and testing pertaining to the parameter of interest.
>
> (Barndorff-Nielsen 1976, p. 105)

In other words, in likelihood inference there need be no extra step of purging the model of extraneous features: it will be done automatically by the main analysis itself. And even this is still a little misleading. In fact, extraneous features in Barndorff-Nielsen's sense can be kept in the model all the way through the analysis, if required: any analysis which is in accord with the likelihood principle will automatically cope with them, in the sense of giving the same answer as if they had not been present.

Apart from the ad hocness introduced by h_0 and T, $\mathcal{P}$ is a very straightforward function. It is one way of capturing the idea of analysing Table 1 by rows. I will consider another way later; but first I will evaluate $\mathcal{P}$.

T'S LACK OF INVARIANCE

It is sometimes the case that two rival test statistics produce different results — one leading to the rejection of a hypothesis and the other not — even though the two statistics, $t_1(x_a)$ and $t_2(x_a)$, say, are merely bijective transformations of each other (i.e., the value of each fully determines the value of the other) (Howson & Urbach 1993, pp. 191–192). In such a case, the two statistics clearly embody exactly the same information about the observation, and yet they give contradictory results. Not only may t_1 and t_2 give different results, often one will have good Frequentist properties while the other has bad Frequentist properties: thus, the property of being a good Frequentist procedure is itself not invariant under bijective transformations. (Jos Uffink has brought this point to my attention.)

But such a bijection is merely a Cambridge change: it is merely a change in description which leaves the thing described unaltered, without even changing the extent of our knowledge about it. Our epistemic inferences *ought* to be similarly unaltered. So an inference procedure that uses a quantity which is not invariant under a bijective change of variables is irrational. Dawid calls this requirement the "transformation principle":

> **Transformation Principle** (TP). Let $\xi : [x_a] \in [X]$, and let $t : [X] \to \mathcal{Y}$ be one-to-one [bijective]. Then TP requires: (i) $\xi \in \Xi$ [the set of all possible experiments] $\Rightarrow \xi^T \in \Xi$ and (ii) $I(\xi, [x_a])$ [any inference drawn from x_a] $= I(\xi^T, t([x_a]))$ [the analogous inference drawn from $t(x_a)$].
>
> (Dawid 1977, p. 248)

A further argument in favour of the transformation principle is as follows. "[I]f the inference made in a given experiment [or merriment] depends only

on a certain function of the raw data, then the same inference should be made if only that function is made available" (Dawid 1977, p. 250). This can be stated formally as follows:

> **Reduction Principle** (RP). Let $\xi \in \Xi, \xi : [x_a] \in [X]$. Consider an [inference procedure] I, and let $T = t([x_a])$ be a statistic satisfying the following definition.
>
> > Definition. T is **reductive** for I in ξ if $I(\xi, x_1) = I(\xi, x_2)$ whenever $t(x_1) = t(x_2)$. (Thus I depends on the data only through the value of T).
>
> Then RP requires: (i) $\xi^T \in \Xi$ and (ii) $I(\xi, [x_a]) = I(\xi^T, t([x_a]))$ [with notation as in Dawid's transformation principle above].
>
> (Dawid 1977, p. 250)

This reduction principle entails the transformation principle, as the following simple argument shows. Let t be any bijection on X. Then t is necessarily reductive in the sense of the above definition, because if $t(x_1) = t(x_2)$ then $x_1 = x_2$ and so of course $I(\xi, x_1) = I(\xi, x_2)$. Thus the transformation principle depends only on the almost undeniable reduction principle, even though the latter may seem much weaker at first sight.

The transformation principle is satisfied by many statistical methods obeying the likelihood principle. For example, likelihood ratios (the basis of almost all Bayesian inference and much of pure likelihood inference, as discussed in chapter 3 and chapter 5 respectively) are invariant under any transformation of θ and x. Maximum likelihood inference, however, need not satisfy the transformation principle (Dawid 1977, p. 250, citing an example due to Pratt). Nor need Bayesian inference with an improper prior (a prior not integrating to 1) (Stone & Dawid 1972): this failure is

related to Stone's (1976) proof that Bayesian inference with an improper prior can be internally inconsistent. (I come back to this issue several times in chapters 9 to 12.)

Dawid (1977, p. 248) makes the point that the transformation principle is "often violated in practice. A statistician may be tempted to assume normality, for example, for the data as actually presented, unless there is sufficient evidence to the contrary. If the data were to be transformed before presentation, he might well end up with a different inference". Such behaviour amounts to changing $p_h, h \in H$, and hence H, after seeing x_a. Dawid seems to be referring to cases in which the statistician has very little guidance on how to set H and therefore uses x_a to help him to take a punt. This violates the spirit of the likelihood principle but not the letter, because such a statistician is operating outside the framework of chapter 2, which took it for granted that H was fixed independently of (and typically prior to) the observation of x_a. Regardless of what we ought to think of this sort of case, it is clearly different in principle from the violation of the transformation principle which Frequentist inference entails, which occurs whether or not the framework of chapter 2 applies and whether or not the statistician has a good grasp of H. One way to express this difference is to note that in Dawid's example the statistician may be willingly risking incoherence, in desperation (since ex hypothesi he has very little idea of how to set H), whereas in the Frequentist case he is forced to be incoherent (in the sense of violating the transformation principle) whether he likes it or not.

The best that can be said for $\mathcal{P}$ is that it has the property that if the same analysis is repeated on a long sequence of experiments which are identical except for random variation it will correctly fail to reject h_0 in 95% of cases in which h_0 is true, assuming that the model is correct (in particular, that all measurement error is entirely represented in the model). I have already argued that $\mathcal{P}$ is ad hoc; I will argue in later sections that it is not as informative about H as it appears to be; and in this section I will argue that it effectively fails to have the attractive theoretical property which I have just cited.

The reason why $\mathcal{P}$ *effectively* fails to have this property is that practically no experiment calculates a single P-value. When more than one P-value is calculated, each one has a chance of being in error, so the statistical analyst faces a dilemma:

- give each P-value an error rate of 5%, in which case the analysis as a whole will have an error rate greater than 5%; or
- adjust each P-value so that the overall error rate of the analysis remains 5%.

Since the whole point of Frequentist theory is to make mistakes at most a known proportion of the time, a fully Frequentist theory must take the second fork of the dilemma and adjust each P-value (Neyman 1937, Kendall & Stuart 1967, Stuart et al. 1999, Mayo 1996). This is usually done using a Bonferroni correction, in which the cut-off for attributing statistical significance "at the 5% level" becomes (5% / n), where n is the number of P-values (or equivalent measures, such as confidence intervals) being calculated. Such a correction is called a correction for *multiplicity* of

analyses. Not all Frequentist analyses use such a correction, but practically all works on Frequentist theory say that they should, and Frequentist analysts who omit to use a correction typically amend their ways when taken to task.

But then each P-value has a probability of error (in the Neyman sense of "probability") which depends on how many other P-values are being calculated as part of the same analysis. This has the following bad consequences:

- It means that when we read a P-value in a scientific paper we cannot tell what its error rate is unless we have been told how many analyses the experimenters made. If they do not tell us that or if we see the P-value quoted out of context there is *no way to tell* what its error rate is.
- When a single experiment is analysed by two or more analysts who calculate different numbers of P-values, they reach different inferential conclusions, despite their analyses being based on the same data.[62]

I have stated this problem in terms of P-values, but it should be clear that it arises for any method of statistical inference based on error rates — i.e.,

62. I do not have space for detailed examples here, but I should mention that this effect is responsible for much of the confusion surrounding statistical analyses of rare events such as brain tumours in cell phone users or around power lines: looking at the same data, some statisticians have calculated many P-values in order to turn up whatever health problems may be there; these statisticians have used Bonferroni corrections with large values of n, which makes their P-values statistically insignificant. Hence, these statisticians conclude that cell phones or power lines or whatever do not cause cancer, and their conclusions are quoted by companies with an interest in continuing to sell cell phones and overhead power cables. Meanwhile, statisticians with a particular interest in one phenomenon — say, cell phone use — calculate a single P-value, which is then much more likely to be statistically significant: its cutoff for statistical significance is n times as large as the generalist statisticians'. These statisticians are much more likely to conclude that cell phones *do* cause cancer, and their conclusions are quoted by shock journalists and health compaigners. Both sets of statisticians are looking at exactly the same data (necessarily so, in this case, since there is only one set of cancer data) and both are using a 5% significance level.

for any Frequentist method. I will give a more detailed example in chapter 15.

A related problem is that there is no way to combine P-values from separate merriments to produce a valid P-value for the overall data (without re-analysing the individual data points from both merriments, which is usually impossible for reasons of logistics and intellectual property). Importantly, this is in contrast to likelihood methods, which allow us to very easily combine likelihoods from separate merriments without needing to look at individual data points: we simply multiply the likelihoods. This is because if $T(x_1)$ and $T(x_2)$ are independent statistics from the same statistical model (i.e., x_1 and x_2 do not depend on each other) then the product of their likelihoods is the same as the likelihood of a combined observation consisting of the data from x_1 and x_2 put together; but there is *no* function of two Frequentist statistics alone which gives the statistic that would be calculated if the data were pooled. For example, there is no function of two confidence intervals which gives a confidence interval for the combined data. Perhaps this is a less deep criticism of Frequentist methods than my other criticisms, since it could, in principle, be overcome by the (drastically impractical) method of always publishing all the raw data on which every analysis is based.

ARE P-VALUES INFORMATIVE ABOUT H?

A first suggestion that, in addition to being ad hoc, P-values do not give us useful information about H comes from a point which I made in the prologue: that Frequentist methods often reject a hypothesis which is clearly

favoured by the data not just *despite* but actually *because* the hypothesis accurately predicted that events which did not occur would not occur. They reject a hypothesis because it got its counterfactuals *right*. This claim was first made by Jeffreys, in what has become one of the most quoted passages in the philosophy of statistics literature:

> $\mathcal{P}$ . . . gives the probability of departures, measured in a particular way, equal to *or greater than* the observed set, and the contribution of the actual value is nearly always negligible. *What the use of $\mathcal{P}$ implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.* This seems a remarkable procedure. On the face of it the fact that such results have not occurred might more reasonably be taken as evidence for the law, not against it. The same applies to all the current significance tests based on $\mathcal{P}$ integrals [which includes the rejection of hypotheses on the basis of confidence intervals — see below].
>
> (Jeffreys 1961, p. 385)

Jeffreys's argument, taken more slowly, is as follows. First of all, the probability of the actual observation, $p(x_a|h_0)$, is almost irrelevant to the value of $\mathcal{P}$ . . . and, in the common continuous case, it is literally irrelevant to $\mathcal{P}$. The probabilities which make up $\mathcal{P}$ are the combined probabilities of observations *greater than* x_a. *These* observations did not occur, and if h_0 assigns a low probability to them then it is correctly failing to predict them (or retrodict them). Now suppose that the calculation of $\mathcal{P}$ leads to the rejection of h_0. Then the aggregate probability of the values of x greater than x_a must be small. In particular, h_0 is rejected if this aggregate probability is less than some critical value, typically 5%. Since it is the correctness of h_0's prediction that the observations in question did not

occur which leads to the small value of the aggregate probability, and since it is the smallness of the aggregate probability which leads to the rejection of h_0, it is the correctness of one of h_0's predictions which leads to its own rejection.

The above argument is not a mere logical trick. Compare Figures 13 and 14:

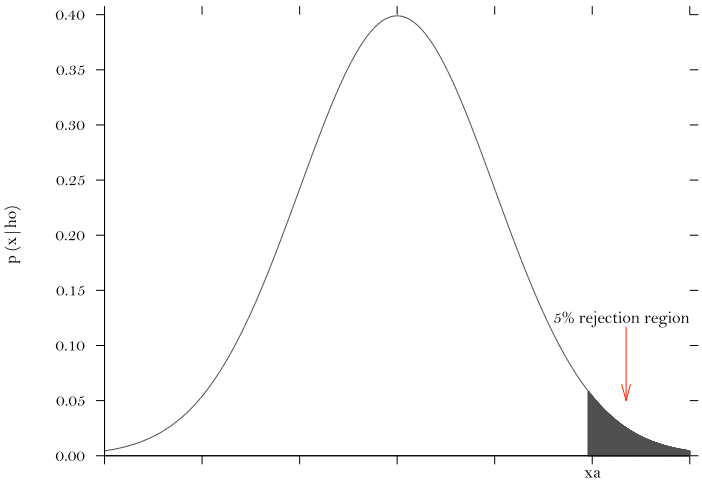

Figure 13

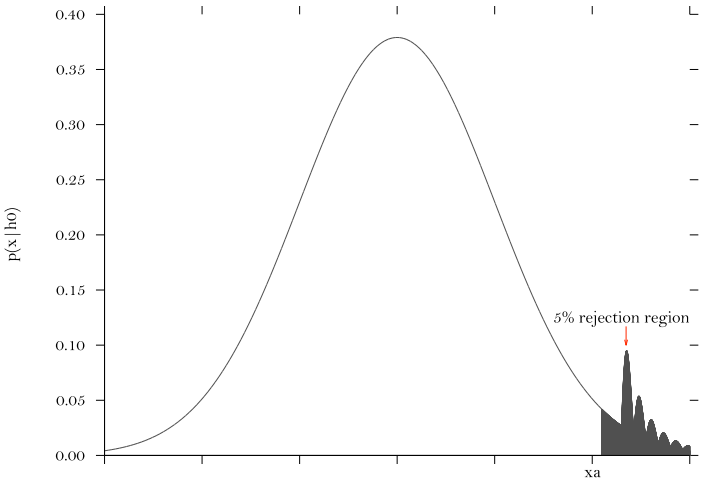

Figure 14

In Figure 13, a significance test based on $\mathcal{P}$ rejects h_0, and the rejection may seem appropriate. But if we change Figure 13 to Figure 14, in which h_0 no longer gives such a low probability to observations larger than x_a, we have a situation in which h_0 will not be rejected, although the probability of the observed result stays the same as it was in Figure 13 (and, incidentally, the probabilities near the centre of the graph also remain almost unchanged).

Jeffreys's argument can be illustrated without resorting to the squiggly distribution shown in Figure 14. It is hard to illustrate the point on a continuous distribution without resorting to squiggles, because small, smooth changes in continuous distributions are hard to notice on a graph; so to show how the point applies to non-squiggly distributions I will quote an illustration with a discrete sample space, adapted from (Berger & Wolpert 1988, p. 106):

	$x = 0$	$x = 1$	$x = 2$	$x = 3$	$x = 4$
$P(x\|h_0)$	.75	.14	0.4	.037	.033
$P(x\|h'_0)$	.70	.25	0.4	.005	.005

Table 3

Suppose that $x = 3$ is observed. A Frequentist statistical test of h_0 will not reject h_0 at the 5% level, because the probability of seeing what was seen or something more extreme — the sum of the entries in the first row to the right of the column $x = 2$ — is greater than 5%.[63] But h'_0 will be

63. This is a rare case in which the choice of the test statistic T is not a problem, because in a simple unidimensional example such as this we can set $T(x) = x$.

rejected, because the sum of the entries to the right of the column $x = 2$ in the second row is less than 5%. h_0' is rejected because it fails to predict the results which have not occurred ($x = 3$ and $x = 4$). h_0, which predicts those results more strongly than h_0 does, escapes rejection. The contrast between h_0 and h_0' makes it clear that, just as Jeffreys said, a hypothesis (h_0) has been rejected because it failed to predict (assigned low probability to) results which did not occur.

Jeffreys's (correct) conclusion from his argument is to endorse the likelihood principle, although without naming it:

> Yates . . . recommends, in testing whether a small frequency n_r is consistent with expectation, that $\chi^2\left[T(x)\right]$ should be calculated as if this frequency was $n_r + \frac{1}{2}$ instead of n_r, and thereby makes the actual value contribute largely to $\mathcal{P}$. This is also recommended by Fisher . . .. It only remains for them to agree that nothing but the actual value is relevant.
>
> (Jeffreys 1961, p. 385 footnote)

The upshot of this argument is that $\mathcal{P}$ does not help us to reach the right conclusions about H.

Can this conclusion be over-ruled by some justification for the use of $\mathcal{P}$ in inference? So far I have given such a justification only implicitly. The best explicit justification I can make for $\mathcal{P}$ is as follows. We want to avoid believing h_0 unless h_0 is true. So we should not believe h_0 unless we have good reason; and we might reason that if $\mathcal{P}$ is small then it seems to be telling us that what we have observed is unlikely, according to h_0. Unlikely things mostly do not happen; but it is only according to h_0 that what we have observed is unlikely (for all that $\mathcal{P}$ tells us); so the unlikeliness of our observation under h_0 is good reason to reject h_0.

Unfortunately, this reasoning is fallacious. It makes no difference whether $T(x_a)$ is unlikely on h_0 unless it is *more* likely on some other hypothesis in H. It is true that *for all* $\mathcal{P}$ *tells us* observing $T(x_a)$ is unlikely according to h_0 and therefore, plausibly, more likely according to other hypotheses; but remember that $\mathcal{P}$ was constructed precisely to exclude consideration of other hypotheses. In order to see whether what we have observed is really more likely according to other hypotheses, we have to examine the table by columns . . . which is incompatible with Frequentism, and, in any case, is something $\mathcal{P}$ manifestly does not do.

An informal diagnosis of how we have got into this mess is that the unlikely events principle, although plausible, is false. Recall:

> **The unlikely events principle**:
> A hypothesis which assigns a low probabilities to an event is disconfirmed by the occurrence of that event to the extent that, if a hypothesis says that an event is unlikely, and yet that event occurs, it is reasonable to conclude, at least tentatively, that the hypothesis is probably false.

In cases in which the outcome is unlikely not only according to the hypothesis under consideration but also according to all competing hypotheses, we should not follow this rule. And, indeed, outside statistical inference we do not follow this rule in such cases. Unlikely events happen all the time, and very rarely do they or should they cause us to reject any hypotheses. To take a coin-tossing example again, consider the hypothesis that a coin is fair. Now toss it twenty times. *Whatever* the outcome is, it is unlikely according to that hypothesis, as we have already seen. Even an outcome

with an equal number of heads and tails — one that intuitively seems to fit the hypothesis best — is extremely unlikely. What *should* cause us to reject the hypothesis is consideration of the probability of the outcome according to the hypothesis we have in mind in comparison with other hypotheses. But this cannot be calculated by an analysis of probability tables by rows.

In summary, our candidate procedure for analysing probability tables by rows suffers from the following flaws:

- The choice of h_0 is generally ad hoc.
- The choice of T is generally ad hoc, and invariant under even bijective transformations of variables.
- A hypothesis may be rejected for *correctly* assigning a low probability to $T(x)$.
- The problem of multiplicity means that the calculation of a P-value does not have an inherent error rate: its error rate depends on what other analyses were conducted at the same time.
- The use of a P-value to reject or fail to reject h_0 makes no sense unless it contains an illicit implicit appeal to other hypotheses.

These are criticisms of the use of a small value of $\mathcal{P}$ to reject h_0. I promised earlier to return to the subject of how we should use a large value of $\mathcal{P}$. I can now clear up that issue very quickly. I hope it is obvious by now that any use of a large value of $\mathcal{P}$ either to accept or reject h_0 is going to suffer from exactly the same problems as our candidate use of a small value of $\mathcal{P}$. The remaining possibility is that a large value of $\mathcal{P}$ should cause us to refrain from saying anything about h_0. That possibility is far from innocuous. It

would have us refrain from saying anything about observing a quiet child if we were taking h_0 to be PTSD (see Table 1). That would not be wise.

Almost all scientific statistical inferences depend on P-values in one way or another. Using P-values in the raw, as it were, is currently out of fashion among statisticians, for an excellent reason, namely that P-values do not give enough information about the data for most scientific purposes. This criticism is right, but it is not my main criticism of P-values. My main criticism is not that they give too little information but that they give misleading information. The current fashion for disliking P-values does not take any of my points into account. The new orthodoxy says that although P-values should not be quoted in the results sections of scientific papers they *should* be used to calculate confidence intervals. From the point of view of my criticisms of Frequentist inference the new fashion is no better than the old. In principle we can see this simply by noting that the lower end-point of a symmetric confidence interval is simply the value of x which would give some fixed P-value (typically 2.5%). That is enough to ensure that the criticisms I have already given apply to confidence intervals. But rather than merely relying on that relationship between confidence intervals and P-values it will be more illuminating, and more fun, to explore confidence intervals in their own right.

3. CONFIDENCE INTERVALS

At the beginning of this chapter, I promised to define both of the commonly used types of Frequentist inference procedere, P-values and confidence intervals. To spare the reader's patience, I will not motivate confidence intervals in the detailed way in which I motivated P-values. Instead, I will

start with the definition of confidence intervals and move straight on to criticisms of their use in statistical inference.

"Confidence interval" sounds as if it denotes an interval in which an unknown parameter is likely to lie; but it does not. Recall its definition from chapter 4:

> If there exist functions of x, $T{\downarrow}$ and $T{\uparrow}$, both statistically independent of θ, such that
>
> $$(\forall\theta) \qquad p(T{\downarrow}(x) \leq \theta \leq T{\uparrow}(x)) = 1 - \alpha$$
>
> then the interval $\left[T{\downarrow}(x_a), T{\uparrow}(x_a)\right]$ is a $1-\alpha$ **confidence interval** for θ.
>
> (adapted from Kendall & Stuart 1967, volume II, p. 99)

$1-\alpha$ is known as the **coverage probability** or the **confidence level** of the confidence interval.[64]

The definition given above (which is the standard definition) hides the fact that $T{\downarrow}$ and $T{\uparrow}$ are functions not only of x_a (the observed data) but also of H (which in this context is, effectively, the set of probabilities that the various possible values of θ assign to the elements of X). The apparently innocuous statement that $(\forall\theta)\, p(T{\downarrow}(x) \leq \theta \leq T{\uparrow}(x)) = 1 - \alpha$ is strongly dependent on the probabilities that *non-actual* values of θ assign to *non-actual* values of X. This is why the use of confidence intervals for inference about θ is a *Frequentist* inference procedure. And it is clear from the definition of a confidence interval, and from the formulae used to calculate confidence intervals, and from applied statisticians' actual practice, that

64. The coverage probability could just as well be denoted α instead of $1 - \alpha$, but the tradition prefers that the letter α should be reserved for an error rate, while $1 - \alpha$ is something more like a success rate.

confidence intervals must be calculated using the whole of the sample space X.

To bring out the dependence of α on the whole of X, it would be better to write the defining equation as:

$$(\forall\theta) \qquad p(T{\downarrow}(x, X, H) \leq \theta \leq T{\uparrow}(x, X, H)) = 1 - \alpha.$$

ARE CONFIDENCE INTERVALS INFORMATIVE ABOUT H?

Typically, functions $T{\downarrow}$ and $T{\uparrow}$ are found such that $(\forall\theta)\ p(T{\downarrow}(x) \leq \theta \leq T{\uparrow}(x)) = 95\%$. Such functions can be calculated from P-values. (In most cases, the endpoints of a 95% confidence interval are simply the values of θ which give the observed data a P-value of 2½%.) Frequentists claim, implicitly or explicitly, that this probability gives us a basis for inference about θ. Usually the claim is made explicitly, and often the inferential usefulness is made part of the definition. To take an example from an influential health policy document:

> **Confidence interval:** the computed interval with a given probability e.g. 95%, that the true value of a variable such as a mean, proportion or rate is contained within the interval.
>
> (Liddle et al. 1996, p. 39)[65]

---

65. Liddle et al. are, of course, implying that probability is relevant to health policy, and hence are making an epistemic claim, even though they *calculate* probability in Neyman's way.

Neyman himself, with his avowedly non-epistemic notion of probability, did not claim that confidence intervals could be used for epistemic inference, and yet both he and his followers did so on a daily basis. This apparent contradiction is explained by the fashionability of a behaviouristic form of falsificationism at the time when he developed his theory (in the 1930s). This made it seem reasonable to say that confidence intervals give us a basis for *action* without having any epistemic consequences at all. Such a view is no longer popular.

The question of who claims that confidence intervals are relevant to epistemic inferences and who does not is contested, but the contest is unimportant for my purposes. I will attack the claim *non ad hominem*. This will be an important building block for my attack on Frequentist inference procedures. Insofar as my opponents are divided about whether their procedures really are inference procedures, so much the better for my position, which is not that Frequentist methods have no place but that they have no place in inferences from x_a to H.

The claim that confidence intervals are relevant to inferences from x_a to H is questionable because it assigns a probability on the basis of a single multiset of observations without taking into account what else is known about θ. This is a point continually stressed by Bayesians, the only people to date who have a comprehensive methodology for taking such information into account . . . which is not to say that their methodology is right (this thesis does not claim to judge that issue) but only that if ulterior knowledge about θ is not taken into account in *any* way then we have no right to make such probability statements. It is an obvious and widely acknowledged fact that probability statements are nonsensical, epistemically speaking, if known information is ignored.[66]

More specifically, the claim that confidence intervals are relevant to inferences from x_a to H is false because once we know x_a (which we must know in order to calculate a confidence interval) we typically know for sure that the probability that θ is in the interval is not $1 - \alpha$.

In a moment I will give an example (involving bonobos) in which an interval C is a bona fide 75% confidence interval calculated in a perfectly standard (and optimal) way but in which we know *for sure* that C contains θ. In this example, the claim that statement that $p(T{\downarrow}(x, X, H) \leq \theta \leq T{\uparrow}(x, X, H)) = 75\%$ is true if interpreted according to its definition, but *only because it is part of the definition of that probability that we ignore some relevant evidence* (as discussed in chapter 4). If interpreted in accordance with the principle of total evidence (contrary to its definition), as a statement

66. Thus, for example, the probability that I am a rock, given that I am an Earthbound physical object but ignoring what else we know about me, is rather high; but to conclude from that that I am a rock would be irrational.

Or, the probability that the Senator for Pennsylvania is the Democrat Joe Hoeffel, ignoring the fact that I know that the Senator is a Democrat, is maybe a half; but it makes no sense to state that probability when I know for sure that the Senator is a Democrat.

about the probability of θ holding such-and-such a value in the light of the available evidence, it is no longer true.

This argument against the use of confidence intervals for purposes of statistical inference is widely admitted, although its *importance* is widely disputed. Thus, for example, Kendall and Stuart, in what is probably the most authoritative single work on the technicalities of the major twentieth-century theories of statistics, agree that we cannot say that there is a probability of $(1 - \alpha)$ that θ lies in its $(1 - \alpha)$ confidence interval. But they go on to dispute the importance which I attach to this fact, on the following grounds:

> Note, in the first place, that we cannot assert that the probability is $1-\alpha \ldots$ [but] If we assert that [θ lies in its confidence interval] in each case presented for decision, we shall be right in a proportion $1 - \alpha$ of the cases in the long run. . . . This idea is basic to the theory of confidence intervals which we proceed to develop, and the reader should satisfy himself that he has grasped it.
>
> (Kendall & Stuart 1967, volume II, p. 99)

If this is intended to be a rationale for the use of confidence intervals (and I believe it is), it must be read as saying that making a probability statement that is known to be wrong is OK provided we bear in mind that if we used the same inference procedure it would turn out correct in a known proportion of *other* cases!

A CLEARLY USELESS CONFIDENCE INTERVAL

I will now back my abstract arguments up with an example adapted from (Berger & Wolpert 1988).

I am studying bonobo chimpanzees in the wild. Researchers further up the river have told me that three new bonobos have moved into my study area. Two of them, Adam and Colin, are indistinguishable apart from size; the third, Bertie, is unusually pale and exactly intermediate in size betwen Adam and Colin. Because my contacts saw them only their heads and upper bodies above the ground cover, they cannot tell me the actual heights of any of them, but they can tell me that Colin is two metres taller than Adam. My task is to estimate x_B, the height of the unusual one, Bertie.

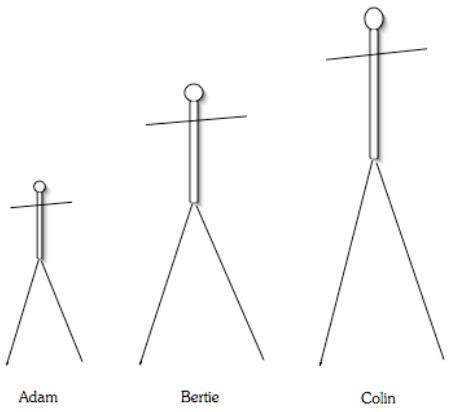

Figure 15: Bertie the bonobo

What makes this a statistical problem is that the bonobos, being new to my area, are likely to move off again if I disturb them. I estimate that I can only afford to get close to one of them at a time, and only twice during my study period; so those two observations will become my merriment. When I observe an ape from close up, I can peer over the ground cover to see his exact size, so if I could observe Bertie this way there would be no problem. However, Bertie is particularly shy, so the chance of observing him from

close up is negligible. All I can get is two height observations of the others: two of Adam, two of Colin, or one of each. Call these observations x_1 and x_2.

A "shortest" (see chapter 4) 75% confidence interval for x_B is:

$C = (x_1 - 1, x_1 - 1)$ if $x_1 = x_2$

$= (\frac{x_1+x_2}{2}, \frac{x_1+x_2}{2})$ if $x_1 \neq x_2$.[67]

I am comforted to find that with only two observations I can estimate Bertie's height with 75% confidence.

What makes this a paradox is that it makes no sense at all for me to plan to give the interval C as my estimate of Bertie's height, or indeed to plan to use it for any other inferential purpose.

Suppose $x_1 = x_2$ (I have observed the same ape twice). Then should I give C as my 75% confidence interval? No, because I know that there is only a 50% chance, not a 75% chance, that C contains Bertie's height. That's bad enough. Now suppose $x_1 \neq x_2$. Should I give C as my 75% confidence interval? Hardly, because this time I'm 100% sure that Bertie's height is $\frac{x_1+x_2}{2}$. So, no matter what I observe, it makes no sense to report 75% confidence in my 75% confidence interval.

The calculation above which seemed so pleasant for a moment has turned out to be completely useless . . .well, almost completely useless. I might have wanted to know in advance what my chances were of finding

---

67. To verify that this is a 75% confidence interval, imagine that the experiment is repeated a large number of times. In half of these repetitions I will observe both Adam and Colin, giving me an accurate reading of x_B; in the other half, I will observe Adam *or* Colin, make a wild guess of which one I'm observing, and estimate Bertie's height correctly half of those times, or one quarter of all times. So in all I will get Bertie's height right 75% of the time. The other quarter of the time I will get it wildly wrong.

If we change my information slightly, so that for example I do not know exactly how much taller Colin is than Adam, it should be clear that the coverage probability of my confidence interval is still roughly 75%. Although such a change would complicate the calculation, the required adjustment would be small. So not too much hangs on the details of my rather contrived example.

out Bertie's height. For that, the same calculation would have been right, but the correct interpretation of it in that case would certainly not be as a confidence interval. Hacking (1965) makes much of the occasional usefulness of such calculations in quality control in factories; Backe (1999) makes a similar point, while Seidenfeld says:

> the N-P [Neyman-Pearson] theory is plausible as a theory of inference *before seeing the actual evidence* (on the 'forward' look), but fails as a theory of inference *after seeing the data* (on the 'backward' look).
>
> (Seidenfeld 1979, p. 15)

This view is compatible with everything I claim in this thesis. I do not argue either for or against it. It certainly does often turn out that the mathematics used to construct confidence intervals is useful in *designing* experiments. But note not only that confidence intervals are not useful in *analysing* experiments, if my example and the likelihood principle are to be believed, but note also that this fact is clearly known in advance. I know right now that my observations of Adam and Colin will not — *cannot*, under any circumstances — lead me to have 75% confidence in the interval C.

This problem with confidence intervals is so bad that people who are aware of it use alternative terminology to designate non-Frequentist estimates which look like confidence intervals but are non-paradoxical: they call them either "interval estimates" (a term which is meant to be neutral as to how the intervals are calculated) or "credible intervals" (a term usually reserved for intervals calculated in a Bayesian way).

Inference procedures which obey the likelihood principle are not subject to problems of this type, because they take into account the observed data by making all the probabilities used in calculating the intervals fully conditional on x_a. In particular, it can be proved that Bayesian credible intervals cannot suffer from nasty examples such as the bonobo example.

Bayesian credible intervals also have some advantages from the point of view of Frequentist criteria (criteria based on long-run averages of performance on hypothetical data). For example, Frequentist confidence intervals, but not Bayesian credible intervals, are subject to the problem of biased relevant subsets, a problem which bothers some Frequentist theorists just as much as it bothers me.

BIASED RELEVANT SUBSETS

This section owes much to (Leslie 1998).

> A **biased relevant subset**, B, is a subset of the sample space . . . such that P(B) is strictly positive (for all θ), and *within which*, for some *positive* value ε, either:
>
> i) The [Frequentist] long run success rate for θ lying within the confidence interval is greater than or equal to $(1-\alpha)+\varepsilon$, for *all* possible θ, or
>
> ii) The [Frequentist] long run success rate for θ lying within the confidence interval is less than or equal to $(1-\alpha)-\varepsilon$, for *all* possible θ.
>
> Relevant biased subsets of form (i) are called positive relevant biased subsets; those of form (ii) are called negative relevant biased subsets.
>
> (Leslie 1998, p. 48)[68]

68. Despite Leslie's use of the variable ε, which mathematicians sometimes use to denote a small quantity, the discrepancy can be very large. In some simple problems it is 30% (Robinson 1975), and it can be even larger.

If the guaranteed coverage probability of a confidence interval is $1-\alpha$, how can the long-run success rate be $1-\alpha\pm\varepsilon$? The answer is that it is only within B that it is $1-\alpha\pm\varepsilon$. The overall success rate is still $1-\alpha$. But this is no cause for comfort. Consider first that when it is time to make inferences about θ based on x_a we *know* whether the result (x_a) is in the subset B or not. This on its own has been enough to cause the Frequentist community to accept the necessity of taking biased relevant subsets into account, at least in some cases, even though to do so breaks the fundamental principles of Frequentist analysis by compromising overall long-run properties. Thus: "Today it is widely accepted by adherents of confidence interval theory that they should perform their analyses conditional on the value of ancillary statistics" (Robinson 1975, p. 155).

Frequentists can achieve some relief from this problem by conditioning on ancillary statistics. Recall from chapter 5 that an **ancillary statistic** is a function T such that $p(T|\theta)$ is independent of θ. Such an ancillary is a function T of x_a such that observing the value of T tells us nothing about θ. Conditioning only on ancillary statistics gives some but, as we will see, not all of the advantages of conditioning on x_a.[69]

Moreover, and far worse from the point of view of any statements made before x_a has been observed, it is sometimes the case that the *whole* of the set of possible observations is made up of biased relevant subsets. Such is the case in the bonobo example above. Sometimes such examples cannot be adjusted in the usual Frequentist way (by performing analyses conditional on ancillary statistics), as illustrated by Robinson, who has

69. See (Leslie 1998) for the history of the conversion of the Frequentist orthodoxy from refusing to condition on any function of x_a to conditioning on ancillary statistics, and for further discussion of the examples in this section.

constructed a confidence interval with a bona fide overall coverage probability of 50% with the property that every possible observation is in one of two biased relevant subsets, one with coverage less than or equal to 20% and one with coverage greater than or equal to 80%, so that no matter what is observed the Frequentist coverage probability of 50% is wildly misleading. In Robinson's example, unlike the bonobo example, the biased relevant subsets do not go away upon conditioning on any ancillary statistic. They *would* go away on conditioning on x_a, as the likelihood principle recommends, but Frequentists cannot do that, since then they would be failing to report the long-run characteristics of the procedure.[70]

The existence of biased relevant subsets (although not the name) has been known since 1939, only a few years after the invention of the confidence interval, when Welch came up with the following example. Suppose we draw a sample of size n from a uniformly distributed population with (unknown) population mean θ and spread 1:

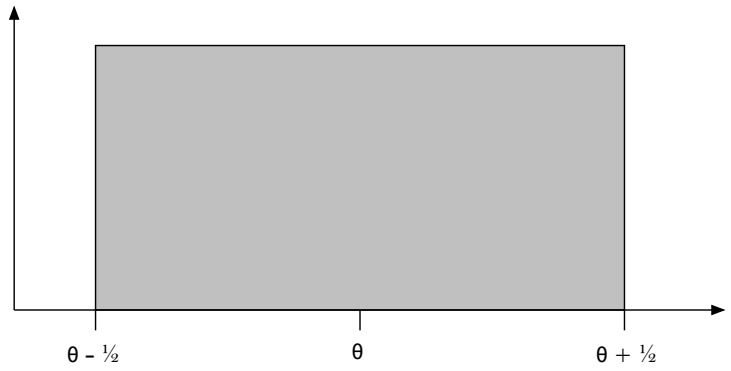

Figure 16

70. In fact Frequentists worthy of the name cannot condition on ancillary statistics either, but they do it anyway and count it as merely a venial sin, whereas conditioning on x_a would be to go all the way with the likelihood principle and therefore would be a mortal sin.

Then the Neyman-Pearson 95% confidence interval for θ is

$$\left[\max\left(\min(x)+d, \max(x)-½\right), \min\left(\min(x), \max(x)+d+½\right)\right],$$

where $2d^n - (2d-1)^n = 0.95$ if $n < 1 - log_2(0.95)$ and $2d^n = 0.95$ otherwise.

Welch did not analyse this example in detail (see (Leslie 1998) for some notable omissions in Welch's analysis), but he did note a disagreement between Fisher's non-Frequentist analysis of the case and Neyman and Pearson's Frequentist analysis. It turns out that the confidence interval has the following properties:

- When $\max(x)-\min(x) > d$ (as is bound to happen sometimes, of course), the 95% confidence interval is *guaranteed* to contain θ.
- When $\max(x)-\min(x) < 2d-1$, the 95% confidence interval *cannot* contain θ (Leslie 1998, p. 38).

Wallace (1959) has shown that Bayesian credible intervals do not suffer from the same problems as Frequentist confidence intervals, and in particular that Bayesian credible intervals cannot have biased relevant subsets (unless they are based on an improper prior distribution — one which is not a probability function). This should come as no surprise (and indeed came as no surprise in 1959), since Bayesian credible intervals obey the likelihood principle and hence condition on x_a.

In chapter 15, I will show other ways in which Bayesian credible intervals can (sometimes) be preferable to confidence intervals even from some Frequentist points of view. The example in chapter 15 will be less impressive than the examples in this chapter but arguably much more important, because it is drawn directly from the actual practice of large

pharmaceutical trials and has direct implications for how such trials should be run in the future.

The problems discussed above suggest that confidence intervals are *invalid* as bases for statistical inferences. In addition, confidence intervals inherit all the ad-hockery of P-values, and add some of their own. Recall, for example, my discussion in chapter 4 of Howson and Urbach's argument that choosing to quote the shortest of the infinite number of valid confidence intervals is ad hoc.

The lesson to be learned from my discussion of confidence intervals is: good riddance to them. But the above problems of confidence intervals (except for *some* of the ad hockery) are entirely attributable to the problems of error rates in general. Since the end-points of symmetric confidence intervals are P-values, for every paradox of P-values it is trivial (although perhaps unenlightening) to generate a dual paradox of confidence intervals, and vice versa. So, instead of further critiquing confidence intervals separately from the task of critiquing other error rates, I will move on to other issues.

4. IN WHAT WAY IS FREQUENTISM OBJECTIVE?

Now that I have shown that Frequentist procedures have many ad hoc elements, I need to ask whether the *objectivity* of Frequentism makes up for its ad hocness.

First, note that any theory can be made objective (in one sense) by revising it so that the theory (rather than any particular application of the theory) defines which choices are to be made whenever an arbitrary decision is called for. This is a completely general point. For example, the

theory of Roschach tests, which is often criticised for being subjective, can be made objective by eliminating any subjectivity on the part of the psychiatrist analysing the pictures; this can be done very easily, by classifying the possible pictures in any determinate way and providing, within the revised theory, determinate rules which the psychiatrist must follow in reaching conclusions. As it is with Freudian analysis, so it is with statistical analysis. Any theory of statistical inference can be made objective by revising it so that what were subjective choices become determinate. A slight modification of Jeffreys's theory, for example, can be seen as an objectification of Subjective Bayesianism: the decisions about prior probability distributions which a Subjective Bayesian makes subjectively a neo-Jeffreys Bayesian can make by using one of the priors which Jeffreys specifies (see chapter 3).[71]

Of course there is something wrong with making a subjective theory objective by modifying it in an arbitrary way. I imagine everyone agrees that there is no virtue in doing so. Or rather, there is no *epistemic* virtue. There is a very great pragmatic virtue in doing so when otherwise the choices allowed by the subjective theory will be used in a pernicious way. For example, one might argue that a legal system (or an electoral system) in which arbitrary choices are enforced by a constitution is better than one in which arbitrary choices are made by individuals, on the hoof. This point is under-emphasised by those Bayesians who denigrate Frequentist theory as being just as subjective as Bayesian theory (Howson & Urbach

71. This is not what Jeffreys's theory actually is: his theory is meant to prescribe prior probabilities only in cases in which the agent doing the analysis is actually ignorant about the parameters in question. Also, it would be a misreading of history to think of Jeffreys as attempting to make Subjective Bayesianism more objective, at least initially, since early versions of his theory predate any statement of Subjective Bayesianism. But we need not worry about Jeffreys's actual theory for the moment. All I am doing is borrowing his mathematics in order to devise a Bayesian theory which is completely objective in the sense I am currently dealing with — the sense of not allowing any subjectivity in its *application*.

1993, p. 12 & passim). In this one important respect Frequentist theories are more objective than Subjective Bayesianism.[72]

A disclaimer: I am not pronouncing on whether the type of objectivity obtained in this way really ought to count as objectivity. I have no theory of what ought to count as objectivity. I do not need one for present purposes; what I am doing instead is discussing the value of something specific which many people count as a type of objectivity.

Frequentist theories of statistical inference have exactly this type of objectivity. Thus, they have an appearance of virtue which is real (as I have argued above) but non-epistemic. Consequently, there is no more reason to believe that the products of Frequentist statistical analysis are right than there would be if the choice of significance level or the choice of confidence interval were set by dice throws. There is more reason to believe that they're right than if they were set by malicious interested parties; that is why I think they have some virtue. But that does not make them epistemically defensible.

---

72. For example, by making the significance level required to reject a hypothesis always 5%, the form of Neyman's theory which has become standard in biomedicine has stopped experimenters from using the arbitrariness of that cut-off to reject any hypotheses they happen not to like. Similarly, although Neyman's theory of confidence intervals does not adequately *justify* any particular choice of interval from among the infinite number with a given coverage probability, the arbitrary rules which he and Pearson and their successors have developed for choosing confidence intervals almost always prevent a statistician from making a *personal* decision about which interval to quote. As a result, a drug company statistician and a public health advocate will virtually always agree on the correct application of this part of the theory; and this has the very beneficial effect of leaving them time to argue about more important matters such as whether the drug company's analysis is being interpreted correctly, is being advertised in misleading and illegal ways, is being ignored in the company's Third World marketing policy, and so on.

So much for the political advantages of putting all the arbitrariness into the theory rather than in the hands of the practitioners of the theory. There is clearly no *epistemic* advantage to doing so; at least, not for an individual epistemic agent. A practitioner of a theory cannot be said to have objective knowledge on the basis of arbitrary decisions made by the inventors of the theory, any more than they could be said to have objective knowledge on the basis of arbitrary decisions which they made themselves. This is why the idea of turning a subjective theory into an objective theory merely by fixing all the arbitrary decisions in advance is a straw man: nobody advocates it.

The arbitrary basis of the objectivity of Frequentist procedures is easily missed, especially since the vast majority of scientists learn the principles of Frequentist inference from texts which concentrate on teaching them how to operate computer algorithms embodying Frequentist methodology. Sadly, the programs in question make it easy (although not absolutely compulsory) to run modules which have all the arbitrary components of the theory hard-coded into them, and the texts take full advantage of this fact. Consequently, it is very hard for a scientist or an applied statistician to find out what is arbitrary in the theory or even that the theory has any arbitrary components at all (with the sole exception of the 5% cut-off, which is *obviously* arbitrary).

Frequentist theories of statistical inference also have other types of objectivity: they study only intersubjectively verifiable phenomena, they use mathematics rather than numerology, and so on. But these types of objectivity are shared by all the theories studied in this thesis.

I conclude that although Frequentist theories are more objective than their main contemporary rivals, the Subjective Bayesian theories, their objectivity is of a sort which confers no epistemic virtue.

5. FUNDAMENTAL PROBLEMS OF FREQUENTISM

In the remainder of this chapter, I will give two very general, related criticisms of Frequentist methods which, I claim, represent Frequentism's most fundamental problems. The first is that Frequentism methods are over-reliant on probabilities assigned to non-actual observations; the second is that Frequntist methods are under-reliant on the information carried by actual observations.

As I foreshadowed earlier, the calculation of a Frequentist error rate is strongly dependent on the probabilities that the *non-actual* values of θ assign to *non-actual* values of x. (This follows immediately from the definition of a confidence interval, for example.)

That this is deeply problematic has been noted many times (mainly in the literature which contrasts Frequentism with Bayesianism), and I can do no better by way of an example than to quote and analyse the following famous passage by Pratt.

Pratt's example

> An engineer draws a random sample of electron tubes and measures the plate voltages under certain conditions with a very accurate voltmeter, accurate enough so that measurement error is negligible compared with the variability of the tubes. A statistician examines the measurements, which look normally distributed and vary from 75 to 99 volts with a mean of 87 and a standard deviation of 4. He makes the ordinary analysis, giving a confidence interval for the true mean.
>
> Later he visits the engineer's laboratory, and notices that the voltmeter used reads only as far as 100, so the population appears to be "censored". This necessitates a new analysis, if that statistician is orthodox.
>
> (Pratt 1962, pp. 314–315)

Censoring is the real or hypothetical lack of potential observations — i.e., observations which might have occurred but didn't; in this case those over 100.

The reason that censoring necessitates a new analysis is that the statistician is performing a Frequentist statistical procedure and therefore needs to be able to report the proportion of an imaginary series of experiments whose results are in some error set — in this case, whose confidence intervals fail to contain the true value of the average plate voltage of a population of tubes from which the sample is drawn. In the merriment actually performed (which the Frequentist statistician must treat as an experiment), none of the tubes had a plate voltage above 99. We can be sure of this, because the voltmeter reads accurately up to 100. But in the imaginary series of experiments which the Frequentist uses in his calculations some of the tubes will have plate voltages over 100.[73]

Some of these unobserved (and very likely non-existent) tubes with plate voltages over 100 would lead to different results in some of the imaginary series of experiments which the Frequentist uses in his calculations (or rather which his computer program uses — I will examine this distinction shortly), since the voltmeter would — hypothetically — incorrectly assign those tubes a value of 100, and this error would have to be corrected (as far as possible) in the analysis. Frequentist statistical methods embody corrections for such errors: in fact, the mathematics used to calculate Frequentist results takes into account the entire probability distribution on possible outcomes of the experiment in a way which guarantees automatic correction for censoring errors provided the censoring is fully described in

73. Note that this is so even if none of the tubes which the engineer actually owns has a plate voltage over 100, and even if none of the tubes which have ever existed or will ever exist have plate voltages over 100! It is guaranteed by the assumption that the plate voltages vary according to a statistical distribution with long tails (in this case the "normal" or Gaussian distributon, but any similar distribution would have the same effect). A non-Frequentist statistician would no doubt make the same assumption about the distribution of plate voltages, but since she does not have to imagine a non-actual series of experiments the assumption is innocuous for her.

the statistician's mathematical model. This is why the Frequentist statistician does a new analysis when he finds out that the engineer's voltmeter only reads up to 100.[74]

One problem resulting from the counterfactual nature of these examples is that evaluating the counterfactuals involved may bring in theories which are quite extraneous to the problem at hand. For example, to decide whether the voltmeter would or would not read above a certain number might require an understanding of how the circuitry near the dial behaves at relatively high temperatures, which in turn might require a theory of the behaviour of doped semiconductors at high temperatures, which is a difficult problem. But that theory seems to be irrelevant to the case at hand, since the voltages applied never exceed 99 and hence the circuitry never gets hot. In other words, the Frequentist statistician's analysis depends on inventing a sufficiently complete context for hypothetical eventualities to enable him to evaluate his counterfactuals. Moreover, this context often has to include factors — psychological, social and even political — which go beyond what one normally thinks of statisticians as taking into account.

An alternative way of stating this problem with Frequentist counterfactuals is that whereas all the statistical inference procedures discussed in this thesis require the statistician to establish a statistical model linking hypotheses to the actual observations, only Frequentist inference procedures require the statistician to model the whole experiment within which the observations are made. Where the non-Frequentist needs only a merriment, the Frequentist needs a fully-modelled experiment.

---

74. One can imagine weirder illustrations of this problem, such as a superstitious experimenter who never reports a result of 13. If this experimenter's observations are, for example, 3, 9, 18 and 20, then his superstitions never come into play, and a non-Frequentist statistician would have no need to take them into account or even to find out about them. But a Frequentist statistician would.

Pratt illustrates this problem nicely in the continuation of his example, in which the statistician is forced to model the experimental situation in much more detail than seems warranted:

> However, the engineer says he has another meter, equally accurate and reading up to 1000 volts, which he would have used if any voltage had been over 100. This is a relief to the orthodox statistician, because it means the population was effectively uncensored after all.

Phew.

> But the next day the engineer telephones and says, "I just discovered my high-range voltmeter was not working the day I did the experiment you analyzed for me." The statistician ascertains that the engineer should not have held up the experiment until his meter was fixed, and informs him that a new analysis will be required. The engineer is astounded. He says, "But the experiment turned out just the same as if the high-range meter had been working. I obtained the precise voltages of my sample anyway, so I learned exactly what I would have learned if the high-range meter had been available. Next you'll be asking about my oscilloscope."
>
> (Pratt 1962, p. 315)

And the engineer is right: if there is a non-negligible chance that the oscilloscope is broken, the statistician *does* have to ask about it. Otherwise, the statistician would not be correctly analysing his imaginary series of experiments.

It is important to note that the analysis of this imaginary series of experiments is not a choice which the Frequentist statistician can take or leave. It is what he does every time, in order to calculate his error rates

(or rather, what his computer program does for him, whether he realises it or not, based on the model he supplies). This is what Howson and Urbach call "the essential weakness of the classical [Frequentist] principle that an estimate must be evaluated relative to the method by which it was derived" (Howson & Urbach 1993, p. 233).

Berger and Wolpert extend Pratt's example (although without mentioning Pratt) to make it clear that the facts which determine a Frequentist calculation may be sociological or political:

> suppose [a scientist conducts an experiment with 200 observations in which] significance has been [narrowly] obtained. . . . the statistician asks what the scientist would have done had the results not been significant. Suppose the scientist says, "If my grant renewal were to be approved, I would then take another 100 observations; if the grant renewal were to be rejected, I would have no more funds and would have to stop the experiment in any case." The advice of the [Frequentist] statistician must then be: "We cannot make a conclusion until we find the outcome of your grant renewal; if it is not renewed, you can claim significant evidence against H_0 [because there will be no need to adjust your existing results], while if it is renewed you cannot claim significance [as explained below] and must take another 100 observations." The up-to-now honest scientist has had enough, and he sends in a request to have the grant renewal denied[.]
>
> (Berger & Wolpert 1988, p. 74.2, and Berry 1988, p. 31–32: exactly the same words are used in both papers!)

Again, as in Pratt's example, the statistician is right to argue the way he does. Frequentist theory demands an overall error rate of 5% (or whatever) for hypothetical repetitions of the experiment, and this error rate can only

be calculated by taking into account what the scientist would have done had his results been different. The necessity for this calculation (within Frequentist theory) is the problem of multiplicity described earlier in this chapter. This calculation is made by applying a Bonferroni correction to each part of the experiment: in other words, once it is known how many observations the scientist is going to make in the future, the Bonferroni correction can be applied to his existing observational results, without waiting to see what the future results are. If the Bonferroni correction is large enough (which it will be if the number of future observations is large enough, relative to the size and statistical significance of the scientist's existing results), the observed results which, on their own, would be judged significant, will become non-significant.[75]

The scientist's grant application might be one he would not even *submit* unless, counterfactually, his first experiment was non-significant. When we once start using such counterfactual considerations, we ought to take into account all relevant counterfactuals which have non-negligible probability. One could argue that if the scientist's first experiment fails there is a small but non-negligible probability that he will apply, and be funded, to perform any number of experiments without bound — enough experiments, that is, to call for a Bonferroni correction large enough to turn the actual significant result into a non-significant result. So Berger

75. This example, as Berger and Wolpert state it, depends on the initial results being only barely significant, so that the Bonferroni correction (for multiplicity) for the hypothetical 100 additional observations changes the P-value by enough to make the results become non-significant; but note that this is the case *whatever* the criterion for significance is (provided only that it is Frequentist; if it is not Frequentist then no Bonferroni correction is needed). And of course the example can be made to apply to cases in which the experiment is as highly significant as you like, by increasing the number of additional observations which the scientist might be funded to perform.

and Wolpert's example is relevant to *every* case of Frequentist statistical inference in which such probabilities are non-negligible.

It may seem implausible that Frequentist statisticians behave as Pratt, Berger and Wolpert claim, so I give a real example of how Frequentist statisticians take into account the context needed to evaluate counterfactuals in chapter 15.

Is it problematic to consider exprimenters' intentions?

Mayo (1996, p. 349) claims that Frequentists have no problem with experimenters' intentions: "the error [Frequentist] statistician has a perfectly nonpsychologistic way of taking account of the impact of . . . experimental plans. The impact is on the error probabilities (operating characteristics) of a procedure." There is no problem, Mayo implies, because that impact is objective. But this is wrong. Certainly the error probabilities are objective, given the "procedure". It is the "procedure" which is not objective. In the "procedure" Mayo includes not just what the experimenter does but all of the things the experimenter might have done had the results turned out differently — which leads straight to Pratt's problem, which remains as "psychologistic" as ever.

Mayo also has a tu quoque argument, claiming that *every* method of statistical inference, Frequentist or not, takes into account subjective intentions, since

> Any and all aspects of what goes into specifying an experiment could be said to reflect intentions—sample size, space of hypotheses, prediction to test, and so on[.]
>
> (Mayo 1996, p. 347)

Mayo's example of sample size misses the point dramatically: the sample size is decidedly not just in the head of the experimenter, but is objectively available to everyone. The sample space and prediction to test are more arguably subjective (although in fact the prediction to test does not figure anywhere in my positive arguments for the likelihood principle — likelihood advocates rarely test hypotheses, preferring to estimate parameters instead), but Mayo's point still fails to go through for at least two reasons. Firstly, even if there were a tu quoque argument it would not give us any reason to add *more* subjectivity to the analysis in the form of propositions about the hypothetical behaviour of broken voltmeters and oscilloscopes. Secondly, and more importantly, no matter how subjective the hypothesis space and prediction to test are, they are available to the *analyst*, to the agent who is making conclusions. In extreme contrast, the point of Pratt's problem is that the hypothetical behaviour of broken equipment is known only by the *experimenter*, if it is known by anyone at all; and if there is more than one subjective view from more than one experimenter, the Frequentist analyst has no way — not even a subjective way — to decide how to interpret the results.

This contrast between experimenter and analyst becomes particularly clear if there is more than one experimenter and more than one analyst, all of whom have different counterfactual beliefs about the broken equipment. It makes some sort of sense for the analysts to disagree — after all, they are doxastic agents with different beliefs. But the *experimenters* enter into the picture not as doxastic agents but as instrumental agents: they set up

the equipment, but there is no reason why their beliefs ought to affect the analyst.[76]

Nor should the analyst be perturbed if the results came about as non-experimental observations (as they do in observational astronomy). A likelihood analyst can cope with non-experimental observations without trouble, while a Frequentist cannot use them at all (in principle; of course this issue is fudged by actual Frequentists by using imaginary experiments, as described under Neyman's theory in chapter 4).

Mayo notes the epistemic force of mere observations. She sees it as a drawback of the Bayesian theory that "it permits [us] to draw conclusions from whatever data and whatever features one happens to notice" (Mayo 1996, p. 350, quoting Le Cam). But it seems to me that being able to draw conclusions from whatever one happens to notice is an essential and valuable part of the life of an epistemic agent.

The above discussion of counterfactuals gives us reason to be wary of Frequentist methods used in the analysis of merriments. It is also useful fodder for the distinction Hacking draws between what I am calling inference and expectancy uses of error rates (as discussed at the beginning of this chapter). When designing inference procedures, the Frequentist needs to evaluate complex social counterfactuals while the non-Frequentist does not. But when designing expectancy procedures — procedures for describing the likely results of merriments not yet undertaken — *both* the Frequentist and non-Frequentist statistician have to evaluate such counterfactuals. Prior to the engineer measuring any tubes, for example, both the Frequentist and non-Frequentist statistician would have to take

76. My point only makes sense if the experimenter and the analyst are different people, of course. My argument is best read by imagining the experimenter to be a mere lab technician or a machine.

into account the brokenness of his low-range voltmeter and the availability of his high-range voltmeter, because they would not know whether the brokenness would affect his results or not. After the fact, they both know that the brokenness has not affected the results, but the Frequentist statistician ignores this knowledge while the non-Frequentist statistician uses it to his advantage. So Hacking's distinction is useful in distinguishing between the cases in which the non-Frequentist statistician has a major epistemic advantage over the Frequentist statistician and those in which she doesn't.

CONDITIONING ON NEW INFORMATION

The following paradox adapted from (Cox 1958) will tell us something about the root causes of the problems I have been discussing.

Suppose that we are doing an experiment to test a hypothesis h_0 and that we decide — unwisely — to go along wtith the Frequentist idea that we should imagine repetitions of the experiment and make sure that at most 5% of them give the wrong answer, on the assumption that h_0 is true (where "wrong answer" is defined in the rather ad hoc way it was for $\mathcal{P}$). An almost realistic thought experiment which sheds light on our options involves sending blood to one of two pathology labs according to which of the labs sends the next pick-up courier, or according to the toss of a coin. One laboratory is known to send back an estimated haemoglobin count with a large amount of random error; the other lab always sends back a count that's almost exactly correct. To achieve an overall 5% error rate as defined above we need to take into account *both* error rates. So if the blood

actually went to the accurate laboratory, we need to increase the error rate on the grounds that it could have gone to the inaccurate one!

This is unsatisfactory, and of course it is not what any practising statistician would actually do. What she would actually do is take into account only the characteristics of the laboratory the blood actually went to. (Recall from chapter 1 that this process of taking into account information which was not available at the time the experiment was designed is called *conditioning*.) Cox himself, in his (1958), came to the conclusion that we should perform a conditional calculation (i.e., take into account only the characteristics of the laboratory the blood actually went to) in this particular case. He did not have available to him the proof of chapter 13 which shows (on very mild assumptions) that this case generalises to practically all statistical analyses. Cox — and all non-Bayesian statisticians at the time — held that one should condition only under special circumstances but was unable to work out exactly what those circumstances were. His example was therefore seen as a paradox.

As discussed above, Frequentist methods do sometimes condition on new information. However, there is not — and *cannot be* — a general Frequentist method for deciding which new information to condition on. One way of seeing this is to take seriously Neyman's theory in which probabilities are based on the reference class and are therefore fixed (see chapter 4). Conditioning would necessarily change these probabilities. Thus, there is no *principled* theory of conditioning available to followers of Neyman. We can, however, imagine a new, anti-Neyman theory which conditions

on available information while retaining other aspects of Neyman's Frequentism.[77] But all such efforts are doomed to failure. I will show this in chapter 13, by showing that the necessity of conditioning in Cox's example, combined with a very plausible axiom of sufficiency, is enough to prove rigorously that one should follow the likelihood principle, which in turn entails that one should always condition on *all* available data.

I will demonstrate that all Frequentist theories must fail in this way by proving the likelihood principle:

> **The likelihood principle**
>
> **Under certain conditions** outlined in chapter 2 and stated fully in chapter 8, **inferences from observations to hypotheses should not depend on the probabilities of observations which have not occurred**, except for the trivial constraint that these probabilities place on the probability of the actual observation under the rule that the probabilities of exclusive events cannot add up to more than 1.

In the light of the proof of the likelihood principle, Cox's example seems more like a reductio of the Frequentist theory of error rates than a true paradox.

---

77. We have seen, for example, that many Frequentist statisticians recommend conditioning on ancillary statistics. This recommendation may sound general, but in fact it is not, because ancillary statistics (a) often don't exist, and (worse) (b) when they do exist are often not unique, with different choices of ancillary statistic on which to condition leading to different conclusions. There are other, arguably more sophisticated conditional Frequentist theories (see, for example, Casella & Berger 1987), but none gives any justification for conditioning on only part of the available information.

6. CONCLUSION

I have shown that the use of Frequentist statistical procedures suffers from the following problems. (I omit from this list ad-hockeries specific to confidence intervals, as discussed above.)

- The choice of h_0 is generally ad hoc.
- The choice of T is generally ad hoc, and invariant under even bijective transformations of variables.
- A hypothesis may be rejected for *correctly* assigning a low probability to $T(x)$.
- The problem of multiplicity means that Frequentist statistics do not have inherent error rates, even though error rates are their raison d'être.
- The use of error rates to reject or fail to reject a hypothesis makes no sense unless it contains an illicit implicit appeal to other hypotheses.

It seems to me that every one of these problems affects the vast majority of statistical analyses currently popular in the sciences, although to justify this claim I would have to survey every science and that is, of course, well beyond the scope of a single thesis.[78]

In addition to those problems, which are severe but which for all I have shown might have partial solutions, the problems of experimenters' intentions revealed by an analysis of the counterfactual nature of Frequentist procedures, and Cox's example, both show that there are *fundamental* problems with Frequentist procedures. In chapter 13, I will show more

78. Some work towards part of such a survey is undertaken in (Grossman & Mackenzie 2005).

formally that coherent inferential procedures must obey the likelihood principle and thus cannot be Frequentist.

I conclude that we should not be looking to Frequentist theories to provide the best theory of statistical inference, and thus that they do not provide good alternatives to the likelihood principle, no matter how popular they currently are with scientists.

In subsequent chapters, I will introduce and prove a carefully worded version of the likelihood principle, using normative axioms which are extremely weak and plausible.

— 8 —
The Likelihood Principle

1. INTRODUCTION

This chapter describes the likelihood principle in detail. The main aim of the chapter is to construct versions of the principles which will withstand all the objections that have been levelled at earlier versions, without losing the spirit of those earlier versions.

Here is a first, approximate definition of the likelihood principle, the rough from which I will attempt to facet a shining gem:

> In certain situations the only permissible contribution of a space of observations X to inferences about a set of hypotheses $\{h_i\}$ is via the likelihood function of the actual observation, $p(x_a|h_i)$.[79]

Or, in terms of tables:

---

79. Recall that the *likelihood* of a given observation is defined as the function from hypotheses to numbers specified by the column in Table 1 representing that observation (and in an analogous way in the infinite case). Thus, there is a separate likelihood function for each possible observation. For example, the likelihood function for the symptom of vomiting in Table 1 is the following function:

dehydration $\longmapsto$ 0.03

PTSD $\longmapsto$ 0.001

etc.

The identity conditions of likelihood functions are not the same as those of functions in general. Two likelihood functions are considered the same iff they are proportional: i.e., iff $L_1(h) = c \times L_2(h)$ for some $c > 0$.

> We should analyse Table 1, and any similar table, using only the numbers in the single column corresponding to the observation result which actually obtained in a given merriment.

The new, more precise version which I will develop in this chapter will be very similar to the first group of definitions given below, especially to Good's (1983), Hill's (1987) and Berger and Wolpert's (1988) versions. The main difference will be a more comprehensive statement of the conditions under which the principle is applicable.

I motivated the likelihood principle in chapter 7. We saw there that Frequentist methods of statistical inference produce unacceptable results because of a failure to make their probabilities conditional on known facts, which suggested that statistical inference ought always to use probabilities which are conditional on the fact that x_a has been observed. This is, roughly, the likelihood principle. I will make this idea precise, and discuss its relationship to previously published versions of the likelihood principle.

THE IMPORTANCE OF THE LIKELIHOOD PRINCIPLE

The likelihood principle tells us something about what it means to be a good theory of statistical inference. It does not (unfortunately) tell us what the best theory of statistical inference is, but it rules out many theories by showing them to contradict axioms which are all but indisputable (the WSP and WCP axioms, presented in the next chapter). We need such principles, because statistical theorists not only can't agree on which statistical procedures are best; they can't agree even in outline on what it *means* for one statistical procedure to be better than another. Now, clashes of epistemic values are the sort of problem that philosophers are usually good at

identifying: noticing, for example, that evolutionary biologists don't agree on what it means for something to be a gene, and noticing the importance of understanding this disagreement prior to trying to assess the divergent theories of evolutionary genetics (Falk 1986, Griffiths & Neumann-Held 1998). We can perform the same job for statistical inference. The likelihood principle principle will help us to do this.

I have already discussed the direct applicability of the likelihood principle to philosophy of science in chapter 1. I showed there that prominent philosophers such as Salmon accept principles which entail the likelihood principle, while the very same philosophers exhort us to do science in a way which contradicts the likelihood principle.

2. CLASSIFICATION

It is helpful to classify the versions of the likelihood principle in the literature into three groups:

I First, there is a group of fairly precise claims about when two different statistical measurements give the same evidence about a set of hypotheses. I place these first in the list of versions of the likelihood principle below, and I will have the most to say about them.

II Then there is a group of claims about the incoherence of averaging over the sample space. These claims are saying that we should not analyse Table 1 (or anything like it) by rows. The claims in this group give us the best way of understanding the practical consequences of the likelihood principle. In chapter 13 I will give an explicit argument for considering the versions in group II to be logically equivalent to

the versions in group I (modulo the vagueness inherent in some versions).

III The third group contains stronger claims which hope to tell us not only when the evidence from two different statistical measurements is the same but, further, to exactly what *extent* a well defined observation supports any relevant hypothesis more than another observation does.[80]

There is another way of classifying versions of the likelihood principle which gives exactly the same groups as above — it is equivalent extensionally to the previous classification, although perhaps it is not quite equivalent intensionally as it is somewhat vaguer.

CLASSIFICATION 2

I STRONG VERSIONS: Inferences about θ may be functions of $p(x_a|h_\theta)$ but should not be functions of $p(x|h_\theta)$ where $x \neq x_a$.

II WEAKER VERSIONS: Inferences about θ must not be functions of x where $x \neq x_a$.

. . . where θ is an index on the set of hypotheses under consideration, X is a space of possible observations, and x_a is the actual observation, as elaborated in chapter 2.

III ANOMALOUS VERSIONS: As group III above.

---

80. This third group is very much the odd one out, historically. It represents an enormous increase in ambition over the other two groups, which are historically prior. Generally the authors who work on principles in the third group do *not* use the term "likelihood principle" for their rules; but three or four of them do, so it is worth saying expliticly that I will not be discussing this third group in this chapter except to note its existence.

In addition to these groups of definitions of the likelihood principle, the literature on the foundations of statistics contains a group of principles recommending that we accept the hypothesis which has the highest likelihood on the observed data — in other words, that we use maximum likelihood estimation as defined in chapter 5. Versions of these principles are sometimes called "the maximum likelihood principle". It is important to note that these principles are not corollaries of the likelihood principle, although they do represent one way of applying the likelihood principle (along with Bayesianism, the method of support, and others). I will not dwell on maximum likelihood principles in this chapter, preferring instead to discuss directly the more fundamental principles which they instantiate.[81]

I now present all of the definitions of the likelihood principle which I have been able to find in the literature, with the exception of many almost word-for-word copies of Jeffreys's definition (see below). Within each of the three groups, I give the definitions chronologically. Some of the definitions I comment on extensively; others, not at all. When I do not comment, it is because the definition in question is relatively imprecise and raises no new issues. At the end of the section on group I, and again at the end of the section on group II, I give a new definition which encapsulates the best of the previous definitions.

---

81. I must, however, mention the importance of maximum likelihood principles to the theory of inference to the best explanation (IBE). I obviously do not have space to discuss this connection in detail, but I should mention that the likelihood principle is compatible with inference to the best explanation and indeed offers some support for it, provided that IBE is defined, as Lipton (2005) defines it, as advising us to take explanatory loveliness as a guide to what we should infer, rather than as the narrower principle that says that we should infer only the most explanatory single hypothesis. See also (Salmon 2001a, Salmon 2001b, Lipton 2001) for a discussion of the relationship between IBE and the Bayesian version of the likelihood principle.

3. GROUP I: THE LIKELIHOOD PRINCIPLE

THE LIKELIHOOD PRINCIPLE: BARNARD'S VERSION (1947)

This was the first statement of the likelihood principle:

> The connection between a simple statistical hypothesis H and observed results R is entirely given by the likelihood, or probability function $L(R|H)$. If we make a comparison between two hypotheses, H and H', on the basis of observed results R, this can be done only by comparing the chances of, getting R, if H were true, with those of getting R, if H' were true.
>
> (Barnard 1947, p. 659)

Barnard immediately gave an argument about the use of the likelihood principle. Since this book touches only briefly on the *use* of the likelihood principle this argument will not figure large, but it is worth quoting for the simplicity it imposes on the mathematical structure of inferences based on the likelihood principle, and for its relevance to confirmation theory (discussed briefly in chapter 3):

> Mathematically, if $L(R|H) = L$, and $L(R|H') = L'$, then our decision about H and H', in the light of data R, must depend on the value of some function $f(L, L')$. Furthermore, this function f must be a function of the ratio, L' / L, only. (Because, intuitively, we can imagine that in addition to observing R, we might have observed some irrelevant event, such as the fall of a coin, whose probability is p, independent of R. Then the likelihoods on H and H' would become pL and pL' [because these are the probabilities, under each of the two hypotheses, of observing both the actual result of the coin toss *and* the data x], and since such an irrelevant

> observation could not affect our decision about H and H', we must have $f(pL, pL') = f(L, L')$ [and so the factor p must cancel out, which is only possible if f is a function of L' / L alone].)
>
> (Barnard 1947, pp. 659–660)

The force of this argument is best seen by imagining an inference procedure which uses the likelihood function $p(x_a|h)$ but which is not a function of $L'/L \equiv p(x_a|h_1)/p(x_a|h_2)$ (where x_a represents the observed data, as usual). Such a procedure might use, for example, $p(x_a|h_1) - p(x_a|h_2)$. Instead of considering a separate coin toss, consider that part of the observed data which is clearly irrelevant to our inferences about hypotheses. When there is no such part of the data, the argument will have to fall back on Barnard's coin toss; but there almost always is some such part of the data. Most commonly, the order in which the data points were collected is exactly such a part of the data: provided the data points are exchangeable (as defined in chapter 2), information about order can be used or neglected, as we prefer, without making any difference to what we know about the hypotheses. In this case, we should be able to use our inference procedure two ways, to calculate both $p(x_a|h_1) - p(x_a|h_2)$ and $p(y_a|h_1) - p(y_a|h_2)$, where x_a is a sequence of observations and y_a is a multiset representing the same observations but without order information. $p(x_a|h)$ will generally be very different from $p(y_a|h)$ — for example, the probability of getting the sequence of die rolls $\langle 1, 2, 2\rangle$ is 1 / 216 on the hypothesis that the die is fair, but the probability of getting the multiset $[1, 2, 2]$ on the same hypothesis is 1 / 72. Similarly, $p(x_a|h_1) - p(x_a|h_2)$ will generally be very different from $p(y_a|h_1) - p(y_a|h_2)$. This is a reductio of the use of these formulas in inference, given our assumption that the order of data points is irrelevant

to inferences about hypotheses. But the formula $p(x|h_1) / p(x|h_2)$ escapes this problem: $p(x_a|h_1) / p(x_a|h_2)$ is just the same as $p(y_a|h_1) / p(y_a|h_2)$, as can be rigorously proved, provided that x_a is statistically independent from y_a.[82] So, Barnard argues (and I agree), whenever we compare pairs of hypotheses our inferences must be based on functions of $p(x_a|h_1) / p(x_a|h_2)$.

What function of $p(x_a|h_1) / p(x_a|h_2)$ we should use might depend on general theoretical considerations outside the scope of this book, or it might depend on prior probabilities or utilities. Maximum likelihood estimation, Bayesian statistical inference procedures and Bayesian decision theory — the only well-developed general methods compatible with the likelihood principle to date — all obey Barnard's restriction that we make inferences from data to hypotheses using *some* function of $p(x_a|h_1) / p(x_a|h_2)$ alone (the main differences being that the third takes into account utility functions and prior probabilities, the second takes into account prior probabilities but not utility functions and the first takes into account neither utility functions nor prior probabilities).

THE LIKELIHOOD PRINCIPLE: JEFFREYS'S VERSION (1961)

> The prior probability of the hypothesis has nothing to do with the observations immediately under discussion, though it may depend on previous observations. Consequently the whole of the information contained in the observation that is relevant to the posterior probabilities of different hypotheses is summed up in the values that they give to the likelihood

---

82. My reductio is dependent on the assumption that x_a is statistically independent from y_a; Barnard's argument involving a separate coin toss is not, so his conclusion is more general. I offer my example even though Barnard's is more general because it is perhaps not obvious that a thought experiment in which we consider the addition of new irrelevant information can be the basis of a tenable argument, whereas my version in which we consider the same analysis with and without information which is already part of the dataset is more obviously realistic.

(Jeffreys 1961, p. 57)

This sums up nicely the orthodox Bayesian view of the likelihood principle: only the posterior distribution matters for inference (this is not explicit in the quotation above, but it is made plentifully clear in context), and the only contribution of observations to a posterior is via the likelihood function (and via $p(x_a)$, which in turn is a function of the likelihood function).

Similar statements are found both implicitly and explicitly in many other works on Bayesian inference.

THE LIKELIHOOD PRINCIPLE: BIRNBAUM'S VERSION (1962)

> *The likelihood principle (L):* If E and E' are any two experiments with the same parameter space $[H]$, represented respectively by density functions $f(x, \theta)$ and $g(y, \theta)$; and if x and y are any respective outcomes determining the same likelihood function; then $\mathrm{Ev}(E, x) = \mathrm{Ev}(E', y)$ [where Ev is a placeholder for a measure of evidence; Birnbaum gives it no precise definition]. That is, the evidential meaning of any outcome x of any experiment E is characterised fully by giving the likelihood function $cf(x, \theta)$ (which need be described only up to an arbitrary positive constant factor), without other reference to the structure of E.
>
> (Birnbaum 1962, p. 283)

This has been the most influential single definition of the likelihood principle, coming as it did in a paper which proved the likelihood principle from plausible axioms for the first time. (I give a similar proof in chapter 13.) However, it is an awkwardly vague definition: some commentators find the notion of an evidence function opaque. My own definition of the likelihood principle will not use the notion of an evidence function.

THE LIKELIHOOD PRINCIPLE: SAVAGE'S 1962 VERSION

> According to Bayes's Theorem, $\Pr(x|\lambda)$ [$p(x_a|h)$, in my terminology], considered as a function of λ, constitutes the entire evidence of the experiment, that is, it tells all that the experiment has to tell. More fully and precisely, if y is the datum of some other experiment, and if it happens that $\Pr(x|\lambda)$ and $\Pr(y|\lambda)$ are proportional functions of λ (that is, constant multiples of each other), then each of the two data x and y have exactly the same thing to say about the values of λ. For example, the probability of seeing 6 red-eyed flies in a randomly drawn sample of 100 is proportional to $\lambda^6(1-\lambda)^{94}$, where λ is the frequency of red-eyed flies in the population, whether the experiment consisted in counting the number of red-eyed flies in a random sample of 100, or of sampling flies at random until 6 with red eyes are observed, or countless other sequential [analysed while in progress] variations of these experiments. I, and others, call this important principle the likelihood principle.
>
> (Savage & discussants 1962, p. 17)

Savage's 1962 definition is part of a defence of Bayesianism and is therefore presented in terms of Bayes's Theorem, a theorem which non-Bayesians believe applies only in unusual circumstances, so it is inappropriate for my purposes. But Savage's example (as opposed to his definition) is relevant to everyone, Bayesian and non-Bayesian alike. Edwards, Lindman and Savage comment on a similar example (where 20 successes have been obtained out of 100):

> What is the datum, and what is its probability for a given value of the frequency p? We are all perhaps overtrained to reply, "The datum is 20 successes out of 100, and its probability, given p, is $C_{20}^{100}p^{20}(1-p)^{80}$." Yet it seems more correct to say, "The

> datum is this particular sequence of successes and failures, and its probability, given p, is $p^{20}(1-p)^{80}$." The conventional reply is often more convenient, because it would be costly to transmit the entire sequence of observations; it is permissible, because the two functions $C_{20}^{100}p^{20}(1-p)^{80}$ and $p^{20}(1-p)^{80}$ belong to the same likelihood; they differ only by the constant factor C_{20}^{100}.
>
> (Edwards et al. 1963, p. 238)

In other words, the likelihood principle explains something which everyone agrees on: that we can transmit the results of such an experiment using a *sufficient statistic* (see chapter 13) which describes only the number of successes and sample size. It is important to note that this explanation is not trivial, especially in the light of the fact that the probability of the sufficient statistic is not the same as the probability of the actual data: the former is $C_{20}^{100}p^{20}(1-p)^{80}$, while the latter is $p^{20}(1-p)^{80}$. So in reporting only the sufficient statistic we are reporting an event which is much more probable than the event which actually occurred (500,000,000,000,000,000,000 times more probable, in fact). This discrepancy between the probability of the event and the probability of the reported summary of the event is prima facie at odds with the Frequentists' insistence that the probability of an event is paramount in deciding what inferences to draw from it. So the *Frequentists'* agreement that it makes sense to quote only the sufficient statistic (and they do all agree on this) very definitely requires explanation. The likelihood principle does the job of explaining this nicely . . . at a cost, for a Frequentist, since the likelihood principle contradicts the basic tenets of Frequentism (see chapter 7 and chapter 15); but as we will see in chapter 13 that is something which they will have to deal with in any case.

THE LIKELIHOOD PRINCIPLE: EDWARDS, LINDMAN AND SAVAGE'S VERSION (1963)

> Two possible experimental outcomes D and D'—not necessarily of the same experiment—can have the same (potential) bearing on your opinion about a partition of events H_i, that is, $P(H_i|D)$ can equal $P(H_i|D')$ for each i. Just when are D and D' thus evidentially equivalent, or of the same import? . . .
>
> $$P(D'|H_i) = kP(D|H_i).$$
>
> . . .the likelihood principle: Two (potential) data D and D' are of the same import if [this equation] obtains.
>
> (Edwards et al. 1963, p. 237)

Edwards, Lindman and Savage's paper in *Psychological Review* was somewhat influential at the time, although its influence seems to have faded a little.

THE LIKELIHOOD PRINCIPLE: LINDLEY'S BAYESIAN VERSION (1965)

> If two sets of data, x and y, have the following properties: (i) their distributions depend on the same set of parameters; (ii) the likelihoods of these parameters for the two sets are the same; (iii) the prior densities of the parameters are the same for the two sets; then any statement made about the parameters using x should be the same as those made using y. The principle is immediate from Bayes's Theorem because the posterior distributions from the two sets will be equal.
>
> (Lindley 1965, p. 59)

Lindley's statement of the likelihood principle, like Jeffreys's and Savage's statements above, explicitly addresses a Bayesian audience. To a non-Bayesian, and indeed to a Restricted Bayesian, Lindley's stipulation that two prior densities (prior probability functions) should be the same is nonsensical except in the situations in which uncontentious, objective prior probabilities exist. More importantly, restricting the likelihood principle to Bayesian analyses is unnecessary, as the many non-Bayesian versions of the principle suggest and as shown conclusively in my proof of a non-Bayesian version in chapter 13.

Nevertheless, Lindley's definition of the likelihood principle is important for the clarity with which it presents the conditions of applicability of the principle: in particular, his condition (i), that likelihoods should only be compared if they refer to the same set of parameters, is often overlooked. This condition is notably lacking from Edwards, Lindman and Savage's definition, for example. I include Lindley's condition in my own definition of the likelihood principle, when I state (below) that two likelihood functions can only be considered the same if all their variables have the same meanings within the theories represented by each hypothesis.

THE LIKELIHOOD PRINCIPLE: SAVAGE'S 1976 VERSION

> The likelihood principle . . . says that the likelihood function for the datum that happens to occur is alone an adequate description of an experiment without any statement of the probability that this or another likelihood function would arise under various values of the parameter.
>
> (Savage 1976, p. 474)

When I am in the mood for a concise but sloppy definition of the likelihood principle, this one is my favourite. But it is much too vague about the conditions under which the likelihood applies to meet the objections of its opponents, so we may as well move straight on to other versions.

THE LIKELIHOOD PRINCIPLE: EDWARDS'S VERSION (1972)

> Within the framework of a statistical model, *all* the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on [given] the data.
>
> (Edwards 1972, p. 30)

Again, this is too vague to meet objections about the exact range of applicability of the likelihood principle. Edwards (unlike Savage) may believe that the likelihood principle *always* applies; but I do not, for reasons which will become clear in chapters 9 to 12.

THE LIKELIHOOD PRINCIPLE: BASU'S VERSION (1975)

> By the term 'statistical data' we mean . . . a pair $(\mathscr{E}, x)$ where $\mathscr{E}$ is a well-defined statistical experiment and x the sample generated by a performance of the experiment. . . . To begin with, let us agree to the use of the notation
>
> $$\mathrm{Inf}(\mathscr{E}, x)$$
>
> only as a pseudo-mathematical short hand for the ungainly expression: 'the whole of the relevant information about [the world] contained in the data $(\mathscr{E}, x)$'.
>
> (Basu 1975, pp. 1–2) . . .

> *(The weak likelihood principle)* : $\mathrm{Inf}(\mathscr{E}, x') = \mathrm{Inf}(\mathscr{E}, x'')$ if the two sample points x' and x'' generate equivalent likelihood functions
>
> (Basu 1975, p. 10) . . .

> *(The likelihood principle)* : If the data $(\mathscr{E}_1, x_1)$ and $(\mathscr{E}_2, x_2)$ generate equivalent likelihood functions on Ω, then $\mathrm{Inf}(\mathscr{E}_1, x_1) = \mathrm{Inf}(\mathscr{E}_2, x_2)$.
>
> (Basu 1975, p. 11)

I see no need for the distinction between weak and strong versions of the likelihood principle. (Basu does not justify the distinction; nor do Barnett or Stuart, Ord and Arnold, who give the same distinction below.) In my framework, the set of hypotheses under consideration is, if necessary, the union of two sets of hypotheses considered as part of two merriments. (Such a move is explicitly countenanced in some although not all discussions of the likelihood principle in the literature, such as (Berger 1985, p. 35).) So my framework encompasses both Barnett's weak and strong versions of the principle. My framework is sufficient to prove the likelihood principle and sufficient to discuss its consequences — indeed, it is better at that than Basu's, since my merriments encompass non-experimental observations as well as experiments. Basu may have in mind Hill's point that two-experiment applications of the likelihood principle can produce certain counter-intuitive results if the principle is not stated carefully, while one-experiment applications cannot. But the necessary care in stating the principle is implicit in the way I set up the mathematics of statistical inference in chapter 2, and in any case is explicit in my own version of the principle below. So there does not seem to be any need for me to explore Basu's more restricted framework.

I will avoid using Basu's terminology of "information", for three reasons: because it has seemed to some authors to be unclear in the same way as Birnbaum's possibly problematic terminology of an "evidence function", because it is ambiguous (other, incompatible notions of information having been defined by Fisher, Shannon and others), and because I do not need it.

Basu also gives a paraphrase of the principle without the contentious word "information":

> We are debating about the basic statistical question of how a given data $d = (\mathscr{E}, [x_a])$, where $\mathscr{E} = (X, \Omega, p)$ is the model and x is the sample, ought to be analysed. ... the likelihood principle ... asserts that if our intention is not to question the validity of the model $\mathscr{E}$ but to make relative (to the model) judgements about some parameters in the model, then we should not pay attention to any characteristics of the data other than the likelihood function generated by it.
>
> (Basu 1975, p. 62)

This definition brings out nicely the negative character of the likelihood principle: it tells us what *not* to do in statistical inference.

THE LIKELIHOOD PRINCIPLE: GOOD'S VERSIONS (1976 AND 1983)

> (I) [T]o me the likelihood principle means that the likelihood function exhausts all the information about the parameters that can be obtained from an experiment or observation, provided of course that there is an undisputed set of exhaustive simple statistical hypotheses such as is provided, for example, by a parametric model.
>
> (Good 1976, reprinted in Good 1983, pp. 35–36)

This is a much more precise version of the likelihood principle than its predecessors, making explicit as it does the restriction of the likelihood principle to simple hypotheses. (Recall from chapter 2 that a simple hypothesis is one which specifies *precise* probabilities for all possible outcomes of a given experiment or of a given observational situation.) This restriction was not made explicit by earlier Bayesian commentators, probably because most early Bayesians were subjectivists in the school of de Finetti, according to whom all hypotheses are simple (i.e., all hypotheses state probabilities for every possible observation). This follows from de Finetti's insistence that prior probabilities can be found for any statement; this tells us that the probability of data x on of any compound hypothesis $h = h_1 \cup h_2$ can be calculated as $\sum_i p(x|h_i)p(h_i)$. This calculation will not be convincing to non-Bayesians, since they deny the guaranteed existence of the prior $p(h_i)$. Among the earlier commentators who were not Bayesians, Barnard probably considered Good's restriction to simple hypotheses to be implicit, while Birnbaum, Hacking (who quotes Birnbaum's version of the principle in his (1965)) and Edwards certainly considered it implied by their claim that the likelihood ratio between two hypotheses exists, forgetting only to make that claim an explicit part of the principle.

Good's definition of the likelihood principle requires that the set of hypotheses be exhaustive. This seems to me to be ambiguous. It could mean that all hypotheses to be considered must be included in the likelihood function — a good idea, which I adopt in my own definition of the likelihood principle. Alternatively, it could mean that all possible hypotheses must be included in the likelihood function. This option is unnecessary . . . and fortunately so, because it is prima facie impossible. Subjective Bayesians

include in their analyses all hypotheses with non-negligible probability, but even they do not include all possible hypotheses.

Good's second go is even better, not because it is more mathematical but because it is more comprehensive:

> (II) Let E and E' be two distinct experimental results or observations. Suppose that they do not affect the utilities (if true) of hypotheses $H_1, H_2, \ldots H_n$ under consideration. Suppose further that E and E' provide the same likelihoods to all the hypotheses, that is, that $\mathrm{P}(E|H_i) = \mathrm{P}(E'|H_i)(i = 1, 2, \ldots n)$. Then E and E' should affect your beliefs, recommendations, and actions concerning $H_1, H_2, \ldots H_n$ in the same way.
>
> (Good 1981, reprinted in Good 1983, p. 132)

The fixed-utilities clause

Good's second version of the likelihood principle is noteworthy because it takes into account the possible complicating factor of utilities (which, in this context, means judgements of how bad it would be to make various inferential errors). Taking utilities into account means that the likelihood principle can tell us something about rational *actions* as well as beliefs and statements. This is a major bonus, because it allows the likelihood principle to go head-to-head with those opposing theories of statistical inference, notably Neyman and Pearson's, which are phrased in terms of actions rather than beliefs or statements.[83]

---

83. The *modern use* of Neyman and Pearson's methods consists of a theory about which scientific statements we should accept; but Neyman and Pearson famously denied that they were giving the foundations for an epistemic theory of any kind; and so, in opposing Neyman and Pearson, I must oppose their original, behaviourist theory as well as the modern pastiche of it. (Some of the most salient history of the mis-quoting of Neyman and Pearson's theory is given in (Gigerenzer 1993).)

I will call Good's restriction of the likelihood principle to cases in which the data does not affect utilities the **fixed-utilities** clause.[84]

THE LIKELIHOOD PRINCIPLE: BERGER'S VERSION (1980)

> In making inferences or decisions about θ after $[x_a]$ is observed, all relevant sample information is contained in the likelihood function.
>
> (Berger 1980, p. 25)

At the risk of quoting Berger too often (a problem exacerbated by the fact that this book quotes the work of two Bergers, James O. and Roger, who hold opposing views on the role of the likelihood principle), I include this relatively imprecise definition for the emphasis it places on inferences made *after* x_a becomes known. The likelihood principle is solely about inferences from *known* data to hypotheses, unlike Frequentist methods (described in chapter 4), some of the important properties of which, such as type I and type II error, can be determined from features of the sample space X before x_a is known. This is an appealing characteristic of Frequentist methods which methods based on the likelihood principle cannot match. But in chapter 7 we saw that the names "type I error" and "type II error" are misleading, and that some of their appeal is illusory.

---

84. Note that even under the fixed-utilities clause the data we observe will affect the utilities of believing in or acting on the various hypotheses; what the data may not do is affect the utilities of the various hyotheses "if true": in other words, the data will only affect our utilities *via* changes in our beliefs about the truth of the hypotheses, not in any other way.

THE LIKELIHOOD PRINCIPLE: BERGER AND WOLPERT'S VERSIONS (1984 AND 1988)

> [E]ssentially . . . all the evidence, which is obtained from an experiment, about an unknown quantity θ, is contained in the likelihood function of θ for the given data[.]
>
> (Berger & Wolpert 1988, p. 1)

To translate into table-talk: all the evidence which is obtained from an experiment about a set of hypotheses is contained in the column of the table which corresponds to the data actually observed.

Berger and Wolpert also provide a more careful version of the likelihood principle which incorporates important caveats:

> Two likelihood functions for θ (from the same or different experiments) contain the same information about θ if they are proportional to one another [i.e., the same as each other] . . . [where] θ represents only the unknown aspect of the probability distribution of X. . . . A second qualification for the LP is that it only applies *for a fully specified model* $\{f_\theta\}$. If there is uncertainty in the model, and if one desires to gain information about which model is correct, that uncertainty must be incorporated into the definition of θ. . . . A third qualification is that, in applying the LP to two different experiments, it is imperative that θ be the same unknown quantity in each.
>
> (Berger & Wolpert 1988, pp. 19–21.2)

Berger and Wolpert say that the likelihood principle does not apply when there is "uncertainty in the model". What they mean by this is perhaps not immediately clear: within their mathematical framework, "model" does not carry any of the important heuristic connotations that it does in some other

parts of the literature. Berger and Wolpert's meaning is well illustrated by the following example, which Forster and Sober (2004) attribute to Popper (1959). Suppose we have information about two variables, x and y, and we want to know whether y is a parabolic function of x or a linear function. The likelihood principle may tell us a lot about the various competing values of the parameters describing the slope of the x–y graph once we have decided whether we are looking at a straight line or a parabola; but it does not tell us anything about whether the graph is a straight line or a parabola. Why not? Because those two hypotheses are *composite hypotheses*: the straight line hypothesis, for example, is the union of various sub-hypotheses — the various straight lines, plus some noise function describing probabilistically how the data depart from the perfect line — each of which gives an exact probability to any given data set; but the composite hypothesis which states merely that the relationship is *some* straight line gives no precise probability to any data set. The likelihood principle does not apply to these hypotheses, because it only applies to what Berger and Wolpert call a "fully specified model", which is what in chapter 2 I called a hypothesis space (H) containing only simple hypotheses.

Later in their (1988), Berger and Wolpert recommend that uncertainty about composite hypotheses ("uncertainty in the model") should be encoded in θ "if one desires to gain information about which model is correct". (I find this phrasing unfortunately euphemistic. We practically always desire to gain information about which model is correct!) To apply Berger and Wolpert's solution to Forster and Sober's example, we would add to the mathematical model a *separate* binary parameter indicating whether the data were best fitted by a straight line or a parabola. We would then find

some way of attributing probabilities to the data conditional on the value of this binary parameter. To apply Berger and Wolpert's recommendation in terms of tables, we would first draw two rows corresponding to the two composite hypotheses and then we would fill in the corresponding cells in the table, at least in the column corresponding to the data actually observed. (For proponents of the likelihood principle, the other columns are not needed for inference from the data to hypotheses . . . not even for inference about composite hypotheses.) We would then analyse the table using our preferred likelihood-principle-compatible analysis.

This is, in outline, a complete solution to the problem of composite hypotheses, but it elides a very large problem: there is no general method for filling in the table when the rows represent composite hypotheses. The whole of my discussion in this book so far has been (and in most of the rest will be) predicated on the premise that we are considering hypotheses which are sufficiently specific to assign probabilities to the various possible vectors of data. The problem is that no matter how specific are the simple hypotheses we start with, there is no guarantee that composite hypotheses formed from them will be specific enough to give probabilities to possible data. Even if we combine just two simple hypotheses, the combination may not assign any probabilities.[85]

---

85. Consider, for example, the hypothesis that you will eat oysters tomorrow. This gives a nice specific probability to the possibility that you will feel woozy and sick (at least, it does if you live in an area where the matter has been studied carefully, as it has where I live). Suppose it is 0.00001. And the hypothesis that you will smoke too much cannabis tomorrow similarly gives a specific probability to the possibility that you will feel woozy and sick — say, 0.5. The combined hypothesis — the union of the two hypotheses — is that tomorrow you will *either* eat oysters or smoke too much cannabis. This combined hypothesis is perfectly clear, but it does not give a probability to feeling sick and woozy. Why not? Because the probability of that possibility depends not only on the facts that the individual probabilities depend on, but also on the relative probability that you will smoke too much cannabis compared with the probability that you will eat oysters. If you are much more likely to smoke cannabis than to eat oysters (and a fortiori not very likely to do both) then the probability of the outcome

When there are competing models which do not give precise probabilities to the possible observations, either because they are composite or because they are vague (Schaffner 1993, chapter 5), the likelihood principle does not apply to the choice between those hypotheses, although it does apply to all the simple *sub*-hypotheses of the composite/vague hypotheses. This point is not made clear by Berger and Wolpert. It is incorporated into my more careful statement of the likelihood principle below.

Berger and Wolpert's final qualification of the likelihood principle (that θ be the same unknown quantity in each observation) will have a role in my proof of the likelihood principle in chapter 13; but as far as I can see it is not needed in the statement of the likelihood principle itself — it is used during my proof, but not in its premises nor in the conclusion.

Elsewhere, Berger and Wolpert give a third definition of the likelihood principle which omits to mention that the principle only applies to conclusions *about* h (or, equivalently if h is indexed, about θ):

> If $E = (X, \theta, \{f_\theta\}$ is an experiment, then $\mathrm{Ev}(E, [x_a])$ [i.e., evidential conclusions drawn from E and x_a] should depend on E and x_a only through $l_x(\theta)$.
>
> (Berger & Wolpert 1988, p. 27)

They note later (p. 41.5) that this version of the principle is false, giving counterexamples in which inferences about certain sub-variables which comprise θ (so-called nuisance parameters) cannot be made without taking into account the whole of θ.

---

conditional on the composite hypothesis is close to the probability conditional on the cannabis hypothesis — 0.5. But if you are very unlikely to smoke too much cannabis, relative to your probability of eating oysters, then your probability of feeling sick and woozy relative to the composite hypothesis is about 0.00001. Relevant information about your habits might be available, but there is nothing in the statement of either simple hypothesis to suggest that it is. If it is not then we simply cannot fill in the table as Berger and Wolpert suggest.

THE LIKELIHOOD PRINCIPLE: HILL'S VERSION (1987)

> Consider two experiments $E_1 = (X_1, q, \{f_1 q\})$ and $E_2 = (X_2, q, \{f_2 q\})$, where q is the same quantity in each experiment. Suppose that for the particular realizations x_1 and x_2 from experiments E_1 and E_2, respectively, $L_{x_1}(q) = c.\, L_{x_2}(q)$ [where L_{x_i} is the likelihood function $p(x_i|q)$], for some positive constant c, and also that the choice of experiment is uninformative with regard to q. Let P be any proposition concerning the value of q and nothing else, i.e., that q lies in some specified set. Then P should be regarded as equally valid whether x_1 is observed in E_1 or x_2 is observed in E_2; and in any decision problem where the loss function depends only upon q and the act taken, the same post-data preference for acts should obtain whether x_1 is observed in E_1, or x_2 in E_2.
>
> There are two differences between [this] and the likelihood principle of Birnbaum and of Berger and Wolpert. The first is that their conclusion . . . has been replaced by a weaker conclusion, that rules out joint statements [see below] . . . The second difference is that the qualification that the choice of experiment not be informative as to the parameter has been added.
>
> (Hill 1987)[86]

Hill makes essentially four changes to Berger and Wolpert's version of the likelihood principle. The first three changes introduce caveats which are so important that I will give them names, while the fourth change fixes a mistake in Berger and Wolpert's statement of the principle which results from a difference between their notation and mine, and which therefore holds no danger for me.

---

86. Hill wrote his version before the edition of Berger and Wolpert's monograph which I quoted a moment ago, (Berger & Wolpert 1988), but after the first edition (Berger & Wolpert 1984), upon which Hill comments.

The no-inference-without-conditioning rule

Let us call Hill's first caveat the **no-inference-without-conditioning** restriction on the likelihood principle. The restriction is straightforward: in applying the likelihood principle, one can only use it for inferences from the data to the hypotheses which have been used to calculate the conditional probabilities one is using. In terms of our table: one can only apply the likelihood principle when one is making inferences from observed data to the hypotheses laid out in the table. In the context of a fixed table, this is obvious: it just says that when the likelihood principle authorises us to use the column of the table containing the observed data, it does not authorise us to make inferences about any hypotheses not contained in the table. The reason for my choice of name is that the restriction says that we may only use the likelihood principle to make inferences about the hypotheses on which we have conditioned to get the probabilities in the table. "Conditioning" here merely means that the probabilities are the probabilities given by the hypotheses in question.

The no-inference-without-conditioning qualification is also obvious in *most* real-life applications of the likelihood principle. We will see the importance of stating it explicitly when we come to discuss objections to the likelihood principle in chapters 9 to 12.

The uninformative-choice-of-merriment rule

Hill's second caveat is one we might call the **uninformative-choice-of-merriment** rule. It concerns cases in which we apply the likelihood principle to a pair of experiments (which, according to Berger and Wolpert's statement of the likelihood principle, which Hill implicitly accepts, must

give information about the same set of hypotheses). It says that our choice of experiments (and, in my extension, merriments) must not in itself give us any information about which hypothesis is likely to be right . . . or, if it does, that must be taken into account, in which case the likelihood principle does not apply simpliciter. The logical importance of this caveat is easy to see: for example, an experimenter might know that one particular hypothesis is particularly plausible, and base her choice of merriments on that knowledge. Hill gives the example of a statistician choosing an experiment which is efficient only if a parameter is small, thus giving us evidence that the statistician believes the parameter to be small. We would be irrational not to take such knowledge into account. The rhetorical importance of this caveat will become clear in later chapters, when we will see that the uninformative-choice-of-merriment rule is a generalisation of the distinction between informative and uninformative stopping rules, a distinction which is prominent in the literature on the merits of the likelihood principle.

A third caveat which Hill introduces is that "the loss function" must depend only on the hypothesis and "the act taken". These are decision-theoretic words. They are important for people who believe that the conclusions we may draw about hypotheses depend on our utilities. Hill's caveat about loss functions is logically equivalent to Good's fixed-utilities clause. A loss function is, by definition, a utility function multiplied by -1.

Hill's example

Hill's fourth change to Berger and Wolpert's version of the likelihood principle is to require that the inferences which are drawn using the principle

are inferences about "the value of q and nothing else", where q is a parameter which distinguishes between the competing hypotheses. This fourth caveat serves to rule out a case which would otherwise be a counterexample to Hill's statement of the likelihood principle. I will describe this case; and we will see that it does not threaten any statement of the likelihood principle based on the framework which I set out in chapter 2, whether it includes Hill's fourth caveat or not.

Hill constructs a case in which Berger and Wolpert's formulation of the likelihood principle is mistaken. He does this by arranging matters so that the parameter q encodes inferentially important information about $h \in H$ which cancels out in the likelihood ratio of q, $\frac{p(E_1,x_1|q)}{p(E_2,x_2|q)}$. His example is as follows (verbatim, except for some abbreviation, and except that I write r where Hill writes θ, in order for the example to match the terminology of both Hill's version of the likelihood principle and my chapter 2). The reader may skip the details of the example if he is willing to trust that the likelihood ratio of q can be made independent of $h \in H$, for a specific choice of E_i and x_i, even if E_i and x_i remain inferentially relevant to h.

> We consider two experiments, E_1 and E_2, which are to be as in the definition of the likelihood principle [above]. If E_i is performed then we will observe the value of the random variable X_i . . . Let $p_i(x; q)$, $i = 1, 2$, be two different probability mass [density] functions for the data, that depend only upon the parameter q. If experiment E_1 is performed then let the probability mass function for the random variable X_1 that will be observed be $p_1(x; q)$, given that H_1 is true, and let it be $p_2(x; q)$, given that H_2 is true. If E_2 is performed then let the probability mass function for the random variable X_2 that will be observed be $p_2(x; q)$, given that H_1 is true, and let it be $p_1(x; q)$, given that H_2 is true. [Note

that this specific choice of probability functions, in which $p_i(x; q)$ depends on H_i, means that the value of q carries information about H_i.] . . .

We shall assume . . . that $Pr\{H_1|E_i, q\} = \text{p}$, a known constant . . . Hence p is simply the unconditional probability of H_1, and similarly, 1-p is the unconditional probability of H_2. . . . We also assume that the choice of experiment is itself uninformative, i.e., that $Pr\{E_i|q\}$ does not depend upon q . . .

Suppose now that the observation x_1 in the experiment E_1 is taken as the data. Then the likelihood function for q . . . is

$$\begin{aligned} L(q; E_1, x_1) &= Pr\{X_1 = x_1, E_1|q\} \\ &= Pr\{X_1 = x_1|q, E_1\} \times Pr\{E_1|q\} \\ &\propto Pr\{X_1 = x_1, H_1|q, E_1\} + Pr\{X_1 = x_1, H_2|q, E_1\} \\ &\ldots \\ &= p_1(x_1; q) \times p + p_2(x_1; q) \times (1 - p). \end{aligned}$$

Similarly, if x_2 is observed in E_2, then the likelihood function for q is

$$\begin{aligned} L(q; E_2, x_2) &\propto Pr\{X_2 = x_2|q, E_2\} \\ &= p_2(x_2; q) \times p + p_1(x_2; q) \times (1 - p). \end{aligned}$$

. . . We now make the further (and last) assumption that there exists a value x_1 of the random variable X_1, for which $p_1(x_1; q) = 0$ for all q, while $p_2(x_1; q) > 0$ for all q; and that there exists a value x_2 of the random variable X_2, for which $p_1(x_2; q) = 0$ for all q, while $p_2(x_2; q) > 0$ for all q.

(adapted from Hill 1988, p. 121)

This construction ensures that the likelihood function for q is the same whether x_1 is observed in E_1 or x_2 is observed in E_2; and yet one should

draw different inferences about $h \in H$ in those two cases, for in the first case H_1 has probability zero and in the second case H_2 has probability zero. This may appear to violate the likelihood principle. In fact it does not, unless the likelihood principle is stated particularly sloppily, because the apparent counterexample trades on calculating the likelihood function of q but then making inferences about a different variable, h.

It is essential to the workings of this example that q encodes important information about h which cancels out in the likelihood ratio $\frac{p(E_1,x_1|q)}{p(E_2,x_2|q)}$. If the likelihood function of q is used for inferences about q alone no problem arises, but if it is used for joint inferences about q and h then the information about h is relevant to the relationships between possible values of q, even though it cancels out in the likelihood ratio; and yet, because it cancels out in the likelihood ratio, (E_1, x_1) and (E_2, x_2) apparently entail the same inference about q. (This cancelling out is a feature of this particular example; it is not something that is bound to happen whenever q encodes information about h.)

If, instead of calculating the likelihood function of q, we calculate the joint likelihood function of q and h, we can safely make joint inferences about q and h after all. It is impossible for any inferentially relevant information which the *joint* likelihood function holds about h to cancel out. This safe procedure — calculating the joint likelihood function of q and h — is just the same thing as calculating the likelihood function of the whole of H, or of calculating the likelihood function of $\theta \in \Theta$, where Θ is any index on the whole of H, since in my terminology (see chapter 2), the hypothesis space H consists of the possible values of H_1, H_2 and q, not of H_i or q alone. (An index on H is a bijective function of $h \in H$,

also known as a one-to-one correspondence with H. The simplest index, if H is finite, is of course a function which simply counts through the members of H.) Hill's parameter q is not an index on H. What makes Hill's example particularly clever is that q is an index of H in experiment E_1 separately or in experiment E_2 separately, but it is a *different* index on H in the two experiments; so, taking Hill's example as a whole, q is not a bijective function of H and hence not an index on H.

In summary, provided we either restrict our inferences to inferences about the parameter whose likelihood function we calculate, or use an index on the whole of H (or H itself) as the parameter whose likelihood function we calculate, we need not worry about Hill's counterexample. The former option tells us more than the latter, so it is preferable epistemically. Hill himself comments,

> Of course one might very well also be interested in the H_i, in which case one might want to include the hypothesis as part of the overall parameter [i.e., use an index θ instead of an arbitrary q], but my point is that there is nothing in the conventional statements of the likelihood principle or in the conventional view of statistical inference that would *force* us to do so.
>
> (Hill 1988, p. 124)

In my own version of the likelihood principle below, inferences are always and only about hypotheses: in Hill's terminology, the hypothesis is always part of the overall parameter. When the hypothesis is represented by a numeric variable θ, θ must be simply an index on the hypothesis space, and cannot serve a dual purpose by encoding any relationships between the hypotheses as q does in Hill's example. (This is all in accord with chapter

2.) Thus, Hill's clever counterexample will not apply to my version of the likelihood principle.

THE LIKELIHOOD PRINCIPLE: BERRY'S VERSION (1987)

> *Likelihood Principle.* The likelihood function $L_x(\theta)$ contains all of the information in an experiment relevant for inferences about θ, where x stands for the observed data.
>
> (Berry 1987, p. 118)

This is not quite the same as any other version, but it adds nothing of philosophical interest to previous versions. I include it because I like it (because it is concise without being extremely misleading) and for completeness.

THE LIKELIHOOD PRINCIPLE: STUART, ORD AND ARNOLD'S VERSION (1999)

> the *likelihood principle* . . . comes in weak and strong forms. The weak principle . . . states that all the information about θ obtained from statistical experiment, E, is contained in the [likelihood function], $L(x|\theta)$. If two replications, yielding observations x_1 and x_2, lead to proportional likelihoods:
>
> $$L(x_1|\theta) = c(x_1, x_2)L(x_2|\theta),$$
>
> where the function c is independent of θ, x_1 and x_2 provide the same information about θ, or
>
> $$\mathrm{Ev}(E, x_1) = \mathrm{Ev}(E, x_2).$$
>
> The strong form . . . extends the principle to include two different experiments, E_1 and E_2, so that

$$\mathrm{Ev}(E_1, x_1) = \mathrm{Ev}(E_2, x_2).$$

(Stuart et al. 1999, p. 438)

Note that c may depend on x_1 and x_2. (It is obvious that c must not depend on θ.)

THE LIKELIHOOD PRINCIPLE: CASELLA AND BERGER'S VERSION (2002)

> LIKELIHOOD PRINCIPLE: If x and y are two sample points such that $L(\theta|x)$ is proportional to $L(\theta|y)$, that is, there exists a constant $C(x, y)$ such that
>
> $$L(\theta|x) = C(x, y)L(\theta|y) \qquad \text{for all } \theta,$$
>
> then the conclusions drawn from x and y should be identical.
>
> (Casella & Berger 2002, p. 291)[87]

THE LIKELIHOOD PRINCIPLE: ROYALL'S VERSION (2004)

> Two instances of statistical evidence are equivalent if and only if they generate the same likelihood function. This proposition is called the likelihood principle[.]
>
> (Royall 2004, p. 126)

87. Note that this Berger, who is also the Berger of (Casella & Berger 1987) is a different person from the Berger of (Berger 1980, Berger & Wolpert 1984, Berger 1985, Berger & Sellke 1987, Berger & Wolpert 1988, Berger & Berry 1988, Berger 1993).

THE LIKELIHOOD PRINCIPLE: BARNETT'S VERSION (1999)

> **The Likelihood Principle.** We wish to draw inferences about a parameter θ in a parameter space Θ. If two sets of data x_1 and x_2 have likelihood functions that are proportional to each other, then they should support identical inferential conclusions about θ. . . .
>
> There are really two versions of the principle—the *weak version* where x_1 and x_2 arise under a common probability model . . . and the *strong version* where the models differ but relate to a common parameter and parameter space.
>
> (Barnett 1999, p. 188)

4. GROUP II: COROLLARIES OF GROUP I

The following four statements which are claimed by their authors to be versions of the likelihood principle might be better seen as rather immediate *corollaries* of the versions given above. They are logically weaker than Group I versions of the likelihood principle, because they do not say that one must use the likelihood function; only that one must not do what Frequentists do, namely to base our inferences on averages over unobserved possible values of observed variables (see chapter 4).[88]

88. It is worth noting that the most vocal opponents of the likelihood principle, such as Mayo, base their position precisely on their view that it is advisable to base all statistical conclusions on averages over unobserved possible values of observed variables: in other words, their alternative to the likelihood principle (although not, of course, the only possible alternative) is to adopt precisely its converse.

THE LIKELIHOOD PRINCIPLE: BIRNBAUM'S COROLLARY (1962)

> [The likelihood principle] may be described informally as asserting the "irrelevance of outcomes not actually observed."
>
> (Birnbaum 1962, p. 271)

THE LIKELIHOOD PRINCIPLE: BERLINER'S COROLLARY (1987)

> One should not base final conclusions or confidences on criteria involving averages over unobserved possible values of observed variables.
>
> (Berliner 1987)

This is the group II version closest to my heart. There are two important subtleties in Berliner's position. He does not say that we should not take averages of unobserved possible values of variables. He says only that we should not *base our inferences* on averages of unobserved possible values of variables *of observed variables*. First of all, we can and should consider averages over unobserved values of variables when we are doing things other than making inferences about our set of hypotheses. In particular, it seems to me and, probably, to Berliner, that we should take such averages when we need to work out the expected (average) properties of a merriment that we have not conducted yet. And secondly, it is only *after* we have observed values of the variables available to us that we should stop taking into account the unobserved values. In terms of my table, we should only restrict ourselves to a single column once we have observed something actual on which to base our choice of which column to look at.

THE LIKELIHOOD PRINCIPLE: BERGER'S COROLLARY (1993)

> The LP states that an estimator should be dependent only on the observed data, rather than the data not seen
>
> (Berger 1993)

5. GROUP II IS LOGICALLY EQUIVALENT TO GROUP I

The principles in Group II entail that we can and should ignore counterfactuals of the following form:

> Had we observed members of the sample space X which we did not in fact observe, they would have made some contribution to the error rate of our inference procedure.

As I showed in chapter 7, all Frequentist procedures necessarily rely on counterfactuals of this form. The error rates defined by Frequentist procedures are, by definition, affected by such counterfactuals; and I suggested towards the end of chapter 7 that it is precisely this property of Frequentist error rates which makes them unsuitable for inference about hypotheses.

It is clear that the likelihood principle as defined by Group I entails the principle as defined by Group II (modulo the vagueness of some of the above definitions), because the Group I principles say that only the likelihood function may be used to draw inferences from data to hypotheses while the Group II principles say that only functions of the actual observation may be so used.[89] The entailment of Group II from Group I follows directly from

89. I am using informal statements of the two groups of principles here, as befits an argument about two vague categories of more or less vague principles. Consequently, the present argument about the equivalence of the two groups needs to be taken with a pinch of salt. I do not think it is important to make these groups more precise. It is, of course, important to make the likelihood principle more precise, and I do that in a later section of this chapter.

the fact that the likelihood function *is* such a function. The entailment in the other direction, from Group II to Group I, is less obvious. It can be shown as follows. Consider an inference procedure which satisfies the requirements of Group II. Let us label the use of the actual observation made by the procedure $f(x_a)$, without loss of generality. Whatever $f(x_a)$ is, it can be algebraically decomposed into three components: a component which does not depend on the probability of x_a, which we can label f_1; a conditional probability component $f_3\big(p(x_a|f_2(H))\big)$, where f_2 is an arbitrary function of the hypothesis space H; and an unconditional probability component $f_4\big(p(x_a)\big)$. Taking these component functions in turn:

> f_1, trivially, is irrelevant to whether the inference procedure satisfies the principles in Group I.
>
> No matter what f_2 is, f_3 is a function of $p(x_a|h)$ where h is a free variable; in other words, f_3 is a function of the likelihood function of x_a.
>
> f_4 may appear not to be a function of the likelihood function of x_a, but Bayes's Theorem (the theorem itself, not the more controversial claims of Bayesianism) ensures that it is, as follows. Since p is a probability function, the integral of $p(h|x_a)$ over $h \in H$ must be 1 (provided that no relevant hypotheses are omitted, a condition which I make explicit in my definition of the likelihood principle below). By Bayes's Theorem,
>
> $$p(h|x_a) = \frac{p(x_a|h).\,p(h)}{p(x_a)}.$$
>
> Integrating both sides over H, and noting that the left-hand side must integrate to 1 as just mentioned, we get
>
> $$1 = \int_h \frac{p(x_a|h).\,p(h)}{p(x_a)}$$

$$\text{so } 1 = \frac{1}{p(x_a)} \int_h p(x_a|h).\, p(h)$$

$$\text{so } p(x_a) = \int_h p(x_a|h).\, p(h)$$

which shows that $p(x_a)$ is a function of the likelihood functon of x_a. (In fact, it is best seen as just a normalisation constant.) This proof is valid even if (as Frequentists sometimes suggest) there is no epistemological or statistical meaning to be attached to probabilties of hypotheses. The proof uses only the mathematical properties of such probabilities, not any interpretation of them.

So each of the components $f_{1\ldots4}$ is a function of the likelihood function. By construction, then, $f(x_a)$ itself is also a functon of the likelihood function. So the inference procedure in question satisfies the Group I principles merely by virtue of satisfying the Group II principles. This completes the (informal) proof that the two groups are equivalent (modulo the vagueness inherent in the definitions).

6. GROUP III: THE LAW OF LIKELIHOOD

I am aware of only four authors who define the likelihood principle in a way which does not fit into Groups I or II; and at least three of these four authors (all except Miller) do so only sometimes, and in other work define it in a way which fits into Group I.

THE LIKELIHOOD PRINCIPLE: BARNETT'S RESTATEMENT (1999)

As we saw above, Barnett (1999, p. 188) defines the likelihood principle in an orthodox way. But later in the same book, while discussing the views of Barnard (who has an orthodox definition of the likelihood principle), Barnett writes:

> for present purposes [the likelihood principle] may be restated as follows in two parts.
>
> (i) If the ratio of the likelihoods for two sets of data is constant for all values of a relevant parameter θ, then inferences about θ should be the same whether they are based on the first, or the second, set of data. This implies that the likelihood function conveys all the information provided by a set of data concerning the relative plausibility of different values of θ.
>
> (Barnett 1999, p. 309)

So far this is orthodox and more or less agrees with Barnett's other definition. But then he adds a statement of the law of likelihood, calling it part (ii) of the likelihood principle:

> (ii) The ratio of the likelihoods, for a given set of data, at two different θ values is interpretable as a numerical measure of the strength of evidence in favour of the one value relative to the other.
>
> (Barnett 1999, p. 309)

THE LIKELIHOOD PRINCIPLE: MILLER'S VERSION (1987)

> This might be called the "likelihood principle": the strength with which a body of data supports a hypothesis as against rivals is the greater as the data are more likely should the hypothesis be true and less likely should the rivals be true.
>
> (Miller 1987, p. 270)

This is the same idea as Barnett's second definition (presented in a more confusing way) except that, unlike Barnett, Miller does not say that the likelihood *is* a strength of evidence but only that it increases as the strength of evidence increases (i.e., it is a monotonic function of a strength of evidence).

THE LIKELIHOOD PRINCIPLE: FORSTER AND SOBER'S VERSION (2004)

Forster and Sober (2004), claiming to quote Hacking (1965) and Royall (1997), give a two-part version of the likelihood principle:

> There is first of all the idea . . . which we will call the qualitative Likelihood Principle:
> (QUAL) O favors H_1 over H_2 if and only if $\Pr(O|H_1) > \Pr(O|H_2)$.
>
> [And then there is the idea] that the likelihood ratio measures the degree to which the observations favor one hypothesis over the other:
> (DEGREE) O favors H_1 over H_2 to degree x if and only if O favors H_1 over H_2 and $\Pr(O|H_1) / \Pr(O|H_2) = x$.
>
> (Forster & Sober 2004a, p. 3)

Elsewhere in the literature (including in Royall's book, contrary to what Forster and Sober *say* Royall says) this principle is called the law of likelihood and is clearly distinguished from the likelihood principle (Royall 1997, p. 3; Hacking 1965, p. 65). Forster's later work steps back from the above definition, and instead defines the likelihood principle along the lines of Group I above (Forster, personal communication).

In this book, I will not be concerned with DEGREE. I will discuss the differences between QUAL and the likelihood principle proper in chapter 9.

As I have already emphasised, the statements in group III are really statements of the law of likelihood, a principle which is logically much stronger and therefore potentially more contentious than the likelihood principle. I believe it is extremely important for clarity in the discussion of statistical inference that criticisms of the law of likelihood do not rub off on the likelihood principle. Similarly, there are dangers in confusing arguments for one principle with arguments for the other. An argument which supports the likelihood principle need not be, and generally is not, an argument for the law of likelihood. This is easy to see if we recall that the former is a only principle about *when*, not *how*, we should use the whole likelihood function, while the latter is a principle about numerical measures of relative strength of evidence.[90]

---

90. Does the law of likelihood imply anything at all about the truth of the likelihood principle? I am not sure, because of ambiguities in the statement of the law of likelihood (ambiguities which I do not need to resolve for the main work of this book). It is not clear to me whether the law of likelihood, as stated above, implies that any other adequate measure of relative strength of evidence must be equivalent to the measure suggested by DEGREE. If so then the law of likelihood implies the likelihood principle. This has been the view of the some prominent writers on the law of likelihood. (Royall is one such. He does not make this point explicitly, but it is fairly clear from the discussion at (Royall 1997, pp. 22–24).) But alternatively one could read the principle as saying, more agnostically, that the suggested measure is only one among many non-equivalent measures, in which case the law of likelihood implies nothing about the likelihood principle.

Incidentally, in 1965 Hacking took the view that the law of likelihood does not imply the likelihood principle because the law of likelihood allows the likelihood function to be changed when the statistical model is reappraised, while the likelihood principle, he takes it, does not (Hacking 1965, pp. 219–220). In my version, and many others, it does, but in 1965 that was not as clear as it is now.

Unlike Groups I and II, Group III is not sufficiently important to my investigation to merit a careful rewording.

7. A NEW VERSION OF THE LIKELIHOOD PRINCIPLE

The following version of the likelihood principle states the main body of the principle precisely and incorporates all of the assumptions which other authors have stated piecemeal. It is the only version of the principle to date to incorporate *all* the necessary assumptions: namely, all the assumptions required by previous versions of the principle, except for the specifically Bayesian assumptions suggested by Lindley and others and except for the assumptions required by Basu's distinction between weak (intra-experiment) and strong (inter-experiment) versions (which I have shown to be unnecessary).

My version of the principle incorporates all the assumptions which are necessary to the proof which I will give in chapter 13, but it does not incorporate all the assumptions thought to be necessary by all authors. For example, Barnard, Jenkins and Winsten (1962) suggest that the likelihood prinicple fails to apply when the sample space or the hypothesis space "are provided with related ordering structures, or group structures, or perhaps other features". Barnard et al. do not argue explicitly in favour of this

restriction of the likelihood principle, while Basu explicitly argues against it, essentially by arguing that the burden of proof is on Barnard et al. to justify this "blank cheque against all violations of [the likelihood principle]" (Basu 1975, p. 20). My own position is simply that neither the informal arguments for the likelihood principle which I have given so far nor my proof of chapter 13 require or even suggest a restriction of the sort Barnard et al. recommend. ("Hypothesi non fingo.")

Terminology

i By "inferences" I mean any beliefs and partial (probabilistic) beliefs which are held or followed and any actions which are taken, as deliberate results of an observation.

ii x_a denotes a vector representing all observations considered relevant to any of the hypotheses in some set H. x_a can be purely observational: it need not result from one or more deliberately constructed experiments. [Discussion: the likelihood principle is only *useful* if the set H contains all the hypotheses of interest; but this need not be made an explicit condition of its applicability, provided inferences about hypotheses not in H are avoided, as formalised in the following point.]

iii By "inferences about hypotheses" I mean any inferences about the hypotheses in H: such inferences must not mention any hypotheses not contained in H except that they may (trivially) mention any hypotheses whose truth is not in doubt and any hypotheses on which x_a has no bearing. [Discussion: in the absence of this condition, the likelihood principle could require that we treat two observations as evidentially equivalent even though one supports an important but unmentioned

hypothesis more strongly than the other one does. Detailed examples illustrating the need for this condition are given in (Berger & Wolpert 1988, pp. 36–38) and elsewhere in the statistical literature, but really the need is completely general and detailed examples are not necessary to show its importance.]

iv Two likelihood functions are considered equal if all their variables have the same meanings within the theories represented by each hypothesis, and if the two functions are proportional (iff $(\exists c > 0)\ (\forall h)\ (L_1(h) = c.\, L_2(h))$. [Discussion: the caveat that the variables must have the same meanings is what I called "Lindley's condition" above. It meets an objection by Pratt which I consider in chapter 13.]

Conditions of applicability

1. We cannot infer anything about the relative importance of the various possible inferential errors from the observation (i.e., the loss function, or equivalently the utility function, is either independent of the observation or unimportant). [This caveat replaces Good's fixed-utilities clause.]
2. The *choice* of observation is not informative about the hypotheses, only its outcome. [This replaces Hill's uninformative-choice-of-merriment clause.]
3. The Well Defined Likelihood Function condition: For each hypothesis h under consideration in a statistical analysis, $p_h(x_a) \equiv p(x_a|h)$ must be a well defined function (i.e., have a single value). [This incorporates the no-inference-without-conditioning clause, discussed above, which says that one can only use it for inferences from the data

to the hypotheses which have been used to calculate the conditional probabilities one is using.]

The likelihood principle

Inferences from observations to hypotheses should not depend on the probabilities of observations which have not occurred, except for the trivial constraint that these probabilities place on the probability of the actual observation under the rule that the probabilities of exclusive events cannot add up to more than 1.

The likelihood principle, as I define it, does not entirely deny that inferences can be based on unobserved or counterfactual outcomes, and nor does it deny the importance of modal considerations in general. It is not in any sense a disguised form of actualism (the metaphysical notion that only the actual exists). It is only certain specific non-actual probabilities which the lilkelihood principle holds to be irrelevant . . . and, even then, it only holds them to be irrelevant to inferences about simple hypotheses *after* observations have been made, and not to (for example) the design of experiments.

Note in particular that the likelihood principle allows inferences about hypotheses to depend on beliefs about merely possible outcomes as long as those beliefs are not probabilistic. (Thanks to Alan Hájek for this point.) In this and other ways, the likelihood principle does not rule out the use of modal claims in general in statistical inference. It only rules out the use of a very specific type of modality. It would be interesting to investigate whether it could be extended to cover any other types of modality without

becoming equivalent to metaphysical actualism. I do not attempt such an investigation in this thesis.

8. OTHER USES OF THE LIKELIHOOD FUNCTION

In this section, I will discuss an important body of work on the epistemology of the likelihood function. This body of work was established by books by Hacking and Edwards appearing in 1965 and 1972 respectively, and was extended by a book by Royall in 1997. All three books promote the method of support, which I have already described briefly in chapter 5. I was not able to discuss the method of support then in as much detail as I would have liked, because at that point we did not yet have a detailed understanding of the likelihood principle.

The work I will discuss in this section is concerned with the question of whether (and if so how) the likelihood principle can be applied to scientific inference without adding anything, such as prior probabilities, to the usual scientific description of the situation. The question is whether we can perform substantive inferences about hypotheses given *only* the ingredients shown in Table 1 — a set of possible observations X and a set of hypotheses H (remembering that p is incorporated into H as described in chapter 2).[91]

The logical relationship between this work and the likelihood principle is not completely straightforward. Pure likelihood methods are useful for

91. Prior to Hacking's and Edwards's books, the literature uniformly took it that there was no way to do this. In particular, almost the only supporters of likelihood methods prior to Hacking were Fisher, who, as we have seen, supported them only intermittently, and Bayesians, who (without exception) only make inferences about hypotheses once they have determined a prior probability distribution for the parameters of interest — and a prior probability distribution, although it may be objective in some cases, is certainly not part of either X or H. It is for this reason that I call Hacking's and Edwards's methods "pure" likelihood methods: they use *purely* the likelihood function and nothing else.

inference only if the likelihood principle is true, so in supporting pure likelihood methods the authors I discuss in this section are supporting the likelihood principle at least implicitly and, in fact, explicitly. But the converse is not true: the likelihood principle does not imply that any pure likelihood method exists that will give epistemologically valid inferences (except for trivial ones, of course). The likelihood principle entails that inference procedures in a given situation must supervene on the likelihood function (no difference in inference from x_a to H without a difference in the likelihood function), but it does not entail that the likelihood function *alone* can tell us anything non-trivial.[92] Consequently, the supporters of pure likelihood methods are claiming much more than the likelihood principle claims.

The law of likelihood ≠ the likelihood principle

It is vital to distinguish between two principles with confusingly similar names. One is the likelihood principle, for which we have already had definitions ad nauseam. The other is the law of likelihood, also very occasionally called the likelihood principle (see group III above), which says something superficially similar but actually very much more ambitious. The *least* ambitious version of the law of likelihood in the literature is this:

> If h and i are simple joint propositions [something slightly more specific than what I have been calling hypotheses] and e is a joint

92. This is easy to see if we recall that Bayesianism is compatible with the likelihood principle. Bayesians agree that our conclusions supervene on the likelihood function *in a particular epistemic situation*, but only because *in a particular epistemic situation* the prior probability function is held fixed. The idea that supervenience on the likelihood function may not guarantee the existence of any non-trivial pure likelihood methods is not restricted to Bayesians, although they are the most prominent writers on this point (Berger & Wolpert 1988, chapter 5, for example). In principle, one might believe that all sorts of information about a particular observational situation or about the variables in question is necessary before one can use the likelihood principle to make inferences. (D. A. S. Fraser is one non-Bayesian who has made this point repeatedly.)

> proposition [an observation], and e includes [is compatible with] both h and i, then e supports h better than i if the likelihood of h exceeds that of i."
>
> (Hacking 1965, p. 59)

Or, in less idiosyncratic terminology,

> *Law of likelihood:* If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$, and the likelihood ratio, $p_A(x) / p_B(x)$, measures the strength of that evidence (Hacking, 1965).
>
> (Royall 1997, p. 3)

If the defence of the law of likelihood in (Hacking 1965), (Edwards 1972) and (Royall 1997) is successful then its success is inherited by the likelihood principle, because the law of likelihood entails the likelihood principle (provided the two principles are stated with the same conditions of application). But if the law of likelihood falls, that does not necessarily reflect badly on the likelihood principle, because the likelihood principle is much weaker. The important difference between the two principles is that the law of likelihood talks about an observation supporting one hypothesis *to a greater extent* than another, while the likelihood principle makes no mention at all of the *extent* to which an observation supports a hypothesis. This may seem an unproblematic difference, or even no difference at all, since the likelihood principle talks about the conditions under which an observation supports two hypotheses equally. The appearance that the two principles are logically equivalent may arise from the fact that both tell us that the

following statement describes a function Ev which in *some sense* tells us the evidential support that x_a provides for h_1 and h_2:

$$\text{If } p(x_a|h_1) = p(x_a|h_2), \text{ then } \text{Ev}(h_1|x_a) = \text{Ev}(h_2|x_a).$$

For proponents of the law of likelihood, this statement is straightforwardly true. For proponents of the likelihood principle it is true but misleading, because they need not hold that Ev is a number (as the statement seems to suggest it is).

The bulk of the proponents of the likelihood principle, being Bayesians, hold that Ev should be equated with the posterior distribution which is the result of a Bayesian analysis, perhaps together with a utility function. The posterior distribution is typically a continuous function, and often highly multidimensional. It most certainly is not a number. So much for Bayesians. Other proponents of the likelihood principle do not have to say that Ev is any sort of mathematical entity at all. They assert that our conclusions about h_1 and h_2 should be the same in some circumstances, but they need not say that those conclusions must have any *formal* structure. Now, whatever Ev is, *perhaps* it can be reduced to a number for some purposes. The law of likelihood asserts that it always can (or, of course, that Ev actually *is* a number). The likelihood principle does not. That is why the likelihood principle is much weaker than the law of likelihood.

Recently, a number of authors writing on confirmation theory (the theory which treats Ev as a number) have begun to investigate the constraints which this approach places on the nature of Ev (Fitelson 2001, Steel 2003). It is to be hoped that their investigations are fruitful; but those who do not treat Ev as a number do not have to abide by those

constraints — at least, not as far as anyone has shown to date. Of course it is possible that mathematical results in any field of enquiry may impact on any other field, and so results from confirmation theory *could* impact on the likelihood principle; but so far they have not done so.[93]

9. THE LIKELIHOOD PRINCIPLE IN APPLIED STATISTICS

Historically, one source of opposition to methods that comply with the likelihood principle has been intellectual, but there has been another source of opposition as well: namely, that statistical methods which comply with the likelihood principle were not feasible in many areas of science until the advent of computers.

Although the likelihood principle itself is very simple, and the procedures developed to date which comply with it (almost all of them Bayesian) are also, conceptually, very simple, as we will see when we meet examples of them in chapter 15, their very simplicity causes a calculational problem. Non-likelihood methods require ad hoc manipulations of the data, some of which are summarised in the test statistic $T(x)$ which I discussed in chapter 7. If the question we are asking is whether we should believe the results of an inference procedure, the ad hocness of $T(x)$ counts strongly against it. On the other hand, if the question we are asking is whether an inference procedure lends itself to easy calculation then we need to look at $T(x)$ more favourably, because it can be chosen so as to simplify the

93. Steel has suggested in print (Steel 2003) that his results have considerable force for Bayesians, but acknowledges in personal communication that as long as Bayesians are committed only to the likelihood principle and not to the law of likelihood his results do not affect them.

calculations enormously. Now, Bayesian procedures never need to introduce any components that are ad hoc.[94] But this leads to a need to evaluate integrals of (typically) very high dimensions and arbitrary shapes; and that has hindered the use of such procedures in real-world problems.

This problem rather rapidly became less important in the 1980s and 1990s, when cheap computers became available which could evaluate such large integrals — at least in most cases. Sadly, the methods used by applied statisticians, and sanctioned by regulatory agencies and funding bodies, became ossified just a couple of decades before computers were powerful enough to provide a good menu of alternatives to Frequentist procedures.

It is hard to be sure whether the slowness of computers can be read into the history as a really important factor in the decisions that have been made by the statistical community; but one suggestive piece of supporting evidence is that recently some important regulatory agencies have begun to liberalise the inference procedures which they sanction among the scientists whose work they are asked to endorse. For example, in mid 2004 the US Food and Drug Administration, possibly the most influential arbiter of statistical methods in the world, advertised posts for fifteen statisticians with PhDs in Bayesian methods (all of whom will a fortiori be experts on methods of inference compatible with the likelihood principle). This liberalisation coincides with the widespread availability of computers fast enough to implement Bayesian methods of the type necessary for pharmaceutical research. Possibly this timing is not coincidental. If not, that

94. Bayesian procedures may have some components which are subjective, which is perhaps problematic, but even the subjective components are not ad hoc: they are chosen to accurately reflect some agent's belief state. By and large, Bayesians have chosen to hold on to this pristine nature of their inference procedures by not introducing any unnecessary ad hoc simplicifation of the model.

suggests (perhaps tentatively) that the lack of availability of computationally feasible procedures has always been a barrier to the acceptance of the likelihood principle. Such speculations have no bearing on the truth of the main arguments of this book; but they do have a bearing on their practical importance.

— 9 —
Misreadings of the Likelihood Principle

This chapter, and the following three, consider objections to the likelihood principle. I will show that none of them is convincing. I will defer objections to my proof of the likelihood principle to chapter 14. There I will show that none of those objections is convincing either.

In this first objections chapter, I will get out of the way a number of objections which are based on accidental misreadings of the likelihood principle. All of the objections I consider in this chapter have been made by authors who quote from the same pool of versions of the likelihood principle as I give in chapter 8; so they are not intending to refer to some different principle. They are objecting to what they see as essentially the same principle as the one which this thesis supports; however, they have misread the principle, and are actually, in their various ways, attacking something which I do not defend.

In the three chapters which follow this one, I will consider objections which apply to the likelihood principle as I have stated it.

1. OBJECTION 9.1
THE LIKELIHOOD PRINCIPLE IMPLIES THAT WE SHOULD TAKE NO CARE OVER EXPERIMENTAL DESIGN

One of the claims [of the Bayesian approach] is that the experi-

> ment matters little, what matters is the likelihood function after experimentation. . . . It tends to undo what classical statisticians have been preaching for years: think about your experiment, design it as best you can to answer specific questions, take all sorts of precautions against selection bias and your subconscious prejudices.
>
> Le Cam, quoted in (Mayo 1996, p. 337)

It may be that Le Cam did not mean to attack the likelihood principle, only some orthogonal part of Bayesianism; but regardless of what Le Cam meant, Mayo gives this quotation as an objection to the likelihood principle.[95]

This thesis is entirely about the problem of statistical inference, not about experimental design, as I said at the outset; but if I reached conclusions which had *blatantly* false implications for how we should think about experimental design, that would be no good; so I must respond to Le Cam's objection. Since I am not defending Bayesianism, but only the likelihood principle, it will suffice to show that the likelihood principle does not imply that we should ignore experimental design. This is a trivial task: my version of the likelihood principle, and also all other versions when read in context, say that *given* that observations have been made in a certain inferential context certain consequences follow. So the likelihood principle simply does not say anything about experimental design, except what we can infer from it by very indirect means, after considering possible frameworks of experimental design with which it could be used. Such possible frameworks are infinitely varied, and neither I nor (as far as I can see in

95. Mayo's implication that Le Cam intended to attack the likelihood principle, although unimportant here, is probably correct given his other views.

this or in their other work) Le Cam nor Mayo believe that the likelihood principle must be used only with experiments designed by Bayesians.

That on its own is enough to show that the likelihood principle does not imply that we should ignore experimental design; I do not additionally need to show that Bayesians do not ignore experimental design. But it is easy to show this too, at least in a sketchy way, so I will do so. Experimental design, including all of the aspects which Le Cam cites as important, is discussed in the following Bayesian works among many, many others: (Gelman et al. 1995), (O'Hagan 1994), (Berger 1980), (Jaynes 1983), (Raiffa & Schlaifer 2000), (Savage 1954), (Lindley 1965), (Jeffreys 1973), (Good 1965), (Good 1965), (Bernardo & Smith 1994), (Spiegelhalter et al. 1986), (Freedman & Spiegelhalter 1989), (Freedman et al. 1983).

There may be versions of Bayesianism which incite us not to care about Le Cam's concerns, but if so I have never heard of them; and (Gelman et al. 1995) and (O'Hagan 1994), which seem to currently be the dominant references for Bayesian scientists, are clearly in agreement with Le Cam about the importance of experimental design.

2. OBJECTION 9.2
IN A WIDE RANGE OF CASES, THE LIKELIHOOD PRINCIPLE FORCES US TO PREFER A COMPLEX MODEL TO A SIMPLE ONE

In order to fully state this objection to the likelihood principle, I need to reiterate Forster and Sober's definition of the principle (already discussed briefly in chapter 8).

FORSTER AND SOBER'S DEFINITION

In a recent paper on the likelihood principle, Forster and Sober (2004) redefine the principle as follows.

> There is first of all the idea . . . which we will call the qualitative Likelihood Principle:
> (QUAL) O favors H_1 over H_2 if and only if $\Pr(O|H_1) > \Pr(O|H_2)$.
>
> [And then there is the idea] that the likelihood ratio measures the degree to which the observations favor one hypothesis over the other:
> (DEGREE) O favors H_1 over H_2 to degree x if and only if O favors H_1 over H_2 and $\Pr(O|H_1) / \Pr(O|H_2) = x$.
>
> (Forster & Sober 2004a, p. 3)

As I have shown (with numerous citations to back my definition), the likelihood principle does not include QUAL or DEGREE. Forster and Sober have confused the likelihood principle with a totally different (although admittedly confusingly named) principle called "the law of likelihood" (Boik 2004). *None* of the definitions I can find in the literature agree with Forster and Sober's or significantly disagree with mine, with the sole exceptions of Miller's and one (but not the other) of Barnett's.

Since, as I have argued, the likelihood principle is important, it is vital to be clear about its meaning and clear about which arguments count against *it* as opposed to counting merely against the law of likelihood. I will therefore pursue this terminological issue a litttle further. Forster and Sober cite Royall's (1997) for their definition of the likelihood principle.[96]

96. They are explicit about this. They write: "Royall follows Hacking in construing the likelihood principle as a two-part doctrine. There is first of all the idea, noted above, which we will call the qualitative Likelihood Principle", followed by the definitions QUAL and DEGREE given above.

But this attribution to Royall is simply a mistake. Royall's definition is essentially the same as mine. And while it is true that Royall's definitions come from Hacking, Hacking is even more explicitly opposed to a definition like Forster and Sober's than Royall is.

Since attribution is at issue, I quote at length instead of paraphrasing. First of all, Royall:

> **The likelihood principle**
> Suppose two simple hypotheses for the distribution of a random variable X assign respective probabilities $f_1(x)$ and $f_2(x)$ to the outcome $X = x$, while two different hypotheses for the distribution of another random variable Y assign respective probabilities $g_1(y)$ and $g_2(y)$ to the outcome $Y = y$. If $f_1(x)/f_2(x) = g_1(y)/g_2(y)$ then the evidence in the observation $X = x$ regarding f_1 vis-à-vis f_2 is equivalent to that in $Y = y$ regarding g_1 vis-à-vis g_2. If a third distribution, f_3, is considered for X, and a third, g_3, for Y, then the two outcomes, $X = x$ and $Y = y$, are equivalent evidence concerning the respective collections of distributions, $\{f_1, f_2, f_3\}$ and $\{g_1, g_2, g_3\}$, if all of the corresponding likelihood ratios are equal: $f_1(x) / f_2(x) = g_1(y) / g_2(y)$, $f_1(x) / f_3(x) = g_1(y) / g_3(y)$, etc. This fact is called the **likelihood principle**
>
> . . . The likelihood principle asserts that two observations that generate identical likelihood functions are equivalent as evidence; in Birnbaum's (1962) words, 'the "evidential meaning" of experimental results is characterized fully by the likelihood function'. "
>
> (Royall 1997, p. 24, quoting Birnbaum 1962)

Royall does give a principle identical to QUAL, but (quite rightly, in view of the previous literature) he calls it the law of likelihood:

> *Law of likelihood:* If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$, and the likelihood ratio, $p_A(x) / p_B(x)$, measures the strength of that evidence
>
> (Royall 1997, p. 3)

Moreover, Hacking, after giving essentially the same definitions as Royall (and citing Barnard 1947, Savage 1961 and Birnbaum 1962 in support), says:

> The likelihood principle does not entail the law of likelihood[.]
>
> (Hacking 1965, p. 65)

So attacking the law of likelihood does not attack the likelihood principle.

Given that Forster and Sober have made an error in citation, it remains to see what we *should* mean by "the likelihood principle". Perhaps there is some reason to prefer Forster and Sober's definition to mine. But I think not. The considerations which we normally use to decide on the meanings of technical terms include etymological precedence and some notion of pragmatism or suitability for purpose. Both of these considerations come out against Forster and Sober's definition.

Firstly, the etymological point can be won either historically or through sheer weight of numbers. The weight of numbers are clear from my earlier citations of many versions of the principle. Historically, Forster and Sober's definition has appeared only in recent years. By contrast, the use of the likelihood principle by other authors is based on the following observation of G. A. Barnard in 1947:

> The connection between a simple statistical hypothesis H and observed results R is entirely given by the likelihood, or probability function $L(R|H)$. If we make a comparison between two hypotheses, H and H', on the basis of observed results R, this can be done only by comparing the chances of, getting R, if H were true, with those of getting R, if H' were true.
>
> (Barnard 1947, p. 659)

This is clearly the likelihood principle as I define it, not the law of likelihood. This definition immediately gained currency, and was widely popularised by a much-cited paper in 1962 (Birnbaum 1962). So it has very clear precedence in the literature.

Secondly, we can ask whose version of the principle is most acceptable on pragmatic grounds. Forster and Sober agree that the likelihood principle is a better foundation for statistical inference than either of the dominant two schools of thought ("Neyman-Pearson-Fisher statistics and . . . Bayesianism", (Forster & Sober 2004a, p. 152)). It follows that it would be extremely useful to have a name for this principle which shows it in the best light possible. After all, we all want to avoid attacking a straw man. Forster and Sober's definition is clearly not the best, since it falls to at least one of their criticisms while alternative definitions do not (as I show below). Moreover, elsewhere Sober himself has accepted that the likelihood principle can validly be defined as I define it.[97] Of course we *could* accept Forster and Sober's definition despite these objections, but I have shown that that would be expensive.

---

97. Sober has written: "the Likelihood Principle, taken at its word, does not rule out the possibility that one can talk about the evidence for or against a given hypothesis without reference to alternative hypotheses. True, advocates of 'likelihoodism' have endorsed the Likelihood Principle and have *also* insisted that evidence is essentially comparative . . ." (Sober 2002b) (my emphasis). In (Sober 2002b) Sober does not say what he takes the likelihood principle to be, but he does say that it need not take evidence is essentially comparative, from which it follows that it cannot be the same as the law of likelihood.

On the same lines of reasoning, DEGREE should not be taken to be part of the likelihood principle. I agree with Forster and Sober's criticisms of DEGREE — for example, that it uses an ad hoc confirmation function, which Fitelson 2001 has shown to be problematic. But again, this is not a criticism of the likelihood principle.

OBJECTION 9.2 CONTINUED

According to Forster and Sober, QUAL cannot generally be used to compare hypotheses which have different numbers of parameters. They use an example to show this. They consider the set containing all hypotheses of the form $y = a + bx + u$ *and* all hypotheses of the form $y = a + bx + cx^2 + u$. Note that the former subset consists of all the straight lines in the plane, and that it is nested inside the latter subset, the parabolas, since every straight line is also a parabola.

At this point in their argument, Forster and Sober say that to compare the likelihoods of composite hypotheses we may do one of two things. We may take an average of the likelihoods of the members of each — a move taken from Bayesian theory, in which likelihoods may be combined using weighted averages, with the weights being ascribed by a prior probability function. Or we may compare the likeliest member of one subset with the likeliest member of the other. They dismiss the former possibility with the following argument:

> Because LIN [the set of straight-line hypotheses] is nested inside of PAR [the set of parabolic hypotheses], it is impossible that Pr(LIN|Data) > Pr(PAR|Data), no matter what the data say. When scientists interpret their data as favoring the simpler

> model, it is impossible to make sense of the judgement within the framework of Bayesianism."
>
> (Forster & Sober 2004a, p. 159)

No argument is given for this, and it is mathematically wrong, as is probably obvious; if not, just slot in the following prior probability density function:

$$\Pr(\text{straight line with slope } a \text{ and intercept } b) \propto a$$
$$\Pr(\text{parabola with coefficients } a, b \text{ and } c) = 0.^{98}$$

Forster and Sober's hasty rejection of the Bayesian option for evaluating composite hypotheses leads them to accept the following principle (which they do not name, so I name it for them):

> [COMPARE] Composite hypotheses are to be compared "by comparing their likeliest special cases."
>
> (Forster & Sober 2004a, p. 159)

COMPARE tells us how to compare the likelihoods of composite hypotheses — hypotheses, like the subsets above, which do not themselves assign probabilities to individual possible observations but which contain sub-hypotheses which do.

The fact that the straight line subset is nested in the subset of parabolas, along with COMPARE, entails that the maximum likelihood must be found within the set of parabolas. Forster and Sober say that "when models are nested, it is almost certain that more complex models will fit the data better than models that are simpler. However, scientists don't take this as

---

98. This function is known as an improper prior because it does not integrate to 1. Some Bayesians allow improper priors. For the others, the improper function can be replaced by the proper but less perspicuous prior $\Pr(\text{parabola with coefficients } a, b \text{ and } c) = 0$ if $c \notin [0, \varepsilon]$, $\Pr(\text{parabola with coefficients } a, b \text{ and } c) = f(c)$ otherwise, where $\varepsilon \ll \Pr(\text{PAR}|\text{Data})$ and f is any function which integrates to $1/\varepsilon$ — a Random Bessell function would do, for example.

a reason to conclude that the data always favor" the more complex models. Therefore QUAL, which says that they should favour the more complex models in this case, is implausible. Thus, Forster and Sober conclude, the likelihood principle is implausible. This ends the statement of the current objection.

RESPONSE TO OBJECTION 9.2

Strictly speaking, I have already defused this objection by showing that it applies only to the law of likelihood, not to the likelihood principle. Nevertheless, let us see whether criticism 2 tells us anything at all. I will conclude that it does have some force, but not enough to rule out the likelihood principle.

There are two logical errors in criticism 2. Firstly, the principle COMPARE is not part of the likelihood principle even as it is stated by Forster and Sober; it requires an additional argument, and the additional argument given is mathematically flawed, as shown above. Secondly, even if we allow COMPARE, we still have a small non sequitur: it follows from Forster and Sober's argument that the likeliest hypothesis must be a parabola, but it may *also* be a straight line. (If c is 0 then $y = a+bx+cx^2+u$ is both a parabola and a straight line, according to mathematical convention.) Hence it is not "almost certain" that the best parabola will be more likely than the best straight line, although it is certain that it will be at least as likely.

Despite those quibbles, we should take some note of criticism 2. After all, some (although I think very few) proponents of the likelihood principle do accept COMPARE (separately from the likelihood principle). Also,

the "almost certain" clause does apply in some situations. So criticism 2 does have some force; but it is not a criticism of the likelihood principle and, as I have shown, the likelihood principle is important and should not be maligned merely because it bears a superficial resemblance to the law of likelihood. Thus, the likelihood principle is saved; but to avoid a Pyrrhic victory I will conclude my response to Forster and Sober by saying something about how the likelihood principle can be *used* in the absence of COMPARE.

CAN WE DO INFERENCE IN THE ABSENCE OF COMPARE?

Forster and Sober might argue that COMPARE is indispensable for inferring anything about composite hypotheses, and hence that there is no point in promulgating a likelihood principle which is incompatible with it. In this section I will briefly discuss the alternatives to COMPARE. I will also give an exceedingly brief case study.

The most obvious way to choose between models is a Subjective Bayesian one. The problem which COMPARE addresses is the assignment of likelihoods to composite hypotheses. The Subjective Bayesian uses her beliefs about the particular matters at hand to decide how composite hypotheses relate to their simple components. Specifically, she uses these beliefs in the form of a prior probability function, as a way of determining a weighted average over the various member hypotheses of each model.

Forster and Sober reject this Bayesian move, but their rejection is based on a mathematical error, as I mentioned above. In any case, one need not be a Subjective Bayesian in order to have alternatives to COMPARE. A non-subjectivist can still take a leaf out of the Bayesian's book

and use domain-specific synthetic considerations to decide how to assign likelihoods to composite hypotheses. Examples of this way of working include the many applications of Bayesian mathematics by objectivist non-Bayesians, especially so-called "Empirical Bayes" methods in biostatistics (Breslow 1990, Morris 1983).[99]

Let me sketch a case study of how one might apply the above reasoning to choose between a parabola and a straight line in the absence of COMPARE. Suppose that amateur astronomers observe a small body moving through the upper atmosphere. Suppose that their observations are not well enough calibrated to determine the body's speed nor its exact trajectory, but that its path appears to be roughly but not exactly a straight line. Then an analysis of the data could model the body's trajectory as a straight line or as a parabola, just as in Forster and Sober's example. The body could be a large meteor, in which case it is best modelled using hypotheses consisting of various straight lines (because meteors move very fast and hence in approximately straight lines). In this case, deviations from straightness would be best modelled as observational error. Alternatively, it could be a ballistic missile moving more slowly, in which case its path is best modelled as a parabola. A non-COMPARE consideration might be that amateur astronomers are unlikely to notice something as small as a missile, while they are much likelier to notice a large meteor. The notion of "unlikely" at work here can be an informal one, taking into account any cogent but non-mathematical considerations: for example, it might be thought that there are no missiles big enough to be widely noticed

99. Empirical Bayes methods are so-called because they use mathematical tools superficially identical to Bayesian mathematics. They are, arguably, not Bayesian in the philosophical sense, and are certainly not subjectivist (Deely & Lindley 1981, Bernardinelli & Montomoli 1992). See chapter 3 for further discussion.

unless a new missile has been developed in secret, and that possibility might be thought unlikely, for no formally specifiable reason. Alternatively, the measure of unlikeliness might depend on the *average* likelihood of noticing a small body, averaged over the various linear hypotheses (meteors) and judged to be large compared to an average over the various parabolic hypotheses (missiles). Either way, this is not an application of COMPARE, which would tell us to take into account only the single likeliest hypothesis on each model.

— 10 —
Is the Likelihood Principle Unclear?

This chapter and the next two examine objections which apply to the likelihood principle as I have stated it (as opposed to the objections which apply only to misreadings of the likelihood principle which I dealt with in the previous chapter). This chapter looks at objections which claim that the likelihood principle does not make sense. The next chapter, chapter 11, considers objections which are based on conflicts between the likelihood principle and other principles and practices, including cases in which the likelihood principle seems at first sight to lead to incorrect analysis of specific statistical models (models which are known in the literature as counter-examples to the likelihood principle) which, supposedly, can be analysed better by applying other methods. Finally, in chapter 12, I will consider miscellaneous other objections to the likelihood principle. The division between these three chapters is not meant to be important; I make it primarily to keep each chapter short.

Some of the criticisms of the likelihood principle presented here are valid when applied to earlier versions of the principle. I will not try the reader's patience by listing for each objection the versions of the principle for which it succeeds and the versions for which it fails, because when earlier versions have been defeated by objections it has been in relatively uninteresting ways. Each of the earlier versions has some degree of sloppiness in its statement of its conditions of applicability, and it is this which has made it vulnerable. I have already catalogued the dimensions of this

sloppiness in chapter 8, and it would be redundant to revisit each dimension in this chapter. Instead, I will compare each objection to my new version of the likelihood principle. This version (which I concocted in chapter 8) agrees in spirit with all earlier versions of the principle (except for the Group III version of Forster and Sober) while tightening up the conditions of applicability. I wish to show that this new version survives all the objections which have been levied at earlier versions.

Some of the objections which I will consider turn on the incompatibility of the likelihood principle with some currently standard Frequentist method of statistical inference. Such objections, in the forms in which they appear in the literature, tend to *hide* the fact that they turn on the mutual incompatibility of the likelihood principle and Frequentism; as we will see, it is often silently assumed that if the likelihood principle were true it would be compatible with Frequentist inference. Typically, this is put together with substantive considerations of some other sort, which are displayed as the ostensible subject of the objection, and inconsistencies are seen to follow, from which it is concluded that the likelihood principle is false.

In organising the discussion of objections of this sort, I have had to choose between **A** stating all such arguments as a single objection and **B** stating each separately. I have chosen route **B**, partly because it follows the organisation of the literature, but mainly because route **A** would have been unenlightening. It is well known that the likelihood principle is incompatible with Frequentist methods; this claim is neither surprising nor helpful. Frequentist methods correspond to analysing Table 1 by rows, while likelihood methods correspond to analysing it by columns. I show in detail, in many places in this thesis (especially chapter 7 and chapter 15),

that these two approaches to statistical inference arise from fundamentally different motivations and are fundamentally incompatible. Demonstrating a specific incompatibility with Frequentism cannot therefore be considered an objection to the likelihood principle. Blaming such incompatibilities on the likelihood principle simply begs the question of which we should prefer: Frequentist inference or the likelihood principle. This thesis as a whole (especially chapter 7 and chapter 13) is an answer to that question. It would be unhelpful to give a necessarily condensed version of the whole thesis as the answer to the objection that the two are incompatible (route **A**). On the other hand, route **B** will prove to be interesting, as I uncover the fundamental objection that the likelihood principle is incompatible with Frequentism from apparently unrelated objections. Once I have uncovered this as the sole source of the objection (by considering the substantive issues in each case), I will of course repeat that the likelihood principle is not shown to be false by its incompatibility with a method (Frequentism) which I have already devoted a whole chapter (chapter 7) to undermining.

This chapter is mostly concerned with the objection that the likelihood principle does not make sense, because either the hypothesis space or the likelihood function (or both) is not well defined. There is no comparable objection that the sample space may not be well defined, because the likelihood principle does not rely on the existence of a sample space: the reader may recall that this is one of its advantages over Frequentist principles.

1. OBJECTION 10.1
THE HYPOTHESIS SPACE IS NOT WELL DEFINED

Lane (Berger & Wolpert 1988, pp. 176–178) objects that there are three possible definitions of h, each of which leads to major problems for the likelihood principle. The following list is quoted from (Berger & Wolpert 1988, pp. 176–178), but with Lane's variables relabelled to match mine, so that his $(X, \Theta, \{P_\theta\})$ becomes my (X, H, p).

The possible definitions of h which Lane canvasses are:

1. h *is* the distribution p;
2. H is an abstract set and h merely indexes the distribution p_h;
3. h is a possible value for some 'real' physical parameter, and p is to be regarded as the distribution of the random quantity X should h be the true value of that parameter.

These options require some explanation; and then we will see that Lane's objections to each of them is right but that he has missed a better option.

1. **Option**: "h *is* the distribution p"

   Explanation: Each hypothesis h consists solely of a probability distribution.

   Objection: Proofs of the likelihood principle (including mine) make use of "mixed experiments" consisting of one merriment followed by another. But these mixed experiments do not have the same probability distribution as either of the simple experiments that they're constructed from. A typical mixed experiment has observations of the form (j, x_j), a cross-product which cannot be described by the same probability distribution as the one that describes observations

of the form x_i, if only because there are more possible observations in the mixed case as in the simple case (if we are dealing with discrete distributions; otherwise the mismatch is not in size but in dimensionality). So the weak conditionality principle, which describes mixed experiments of the general form $(X \otimes 2, h, p)$, is using h in a context in which it cannot possibly be applied.

This objection is correct, strictly speaking. The obvious riposte is that h is meant to apply just to *part* of the mixed experiment. That too fails to work, strictly speaking, so long as the mixed experiment is described as $(J \otimes X, h, p)$. When we get to my final solution to Lane's objections, I will give a precise way to avoid this problem; but already we are nearly there.

2. **Option**: "H is an abstract set and h merely indexes the distribution p_h"

   Explanation: H is the set $\{1, 2, 3, \ldots n\}$, telling us which member of a set of distributions $\{p_1, p_2, p_3, \ldots p_n\}$ we should use.

   Objection: In that case, we could apply the likelihood principle to *any* two merriments so long as they had the same number of possible outcomes. But then the principle would no longer be plausible. For example, we consider a merriment in which we examine a human blood sample, and another in which we examine the Soviet flag. We might then ask what colour our object is. On receiving the answer "red", which has likelihood 1 in both experiments, the likelihood principle would tell us to draw the same conclusions about the world from either one.

This objection is also correct. h cannot be merely an index applicable equally to any set of probability distributions.

3. **Option**: "h is a possible value for some 'real' physical parameter, and p is to be regarded as the distribution of the random quantity X should h be the true value of that parameter."

   Explanation: h is the probabilistic equivalent of a truthmaker for the probability distribution p.

   Objection: It is unclear what these underlying properties might be. Lane gives the example of a coin toss: in using the likelihood principle in such a case, one would have to be a realist, non-pragmatist believer in propensities in order to think that there is an underlying real physical parameter.

This objection is almost certainly correct, bearing in mind the great variety of uses of the likelihood principle. Even if the world contains propensities, there would have to be a separate propensity underlying every useful probabilistic statement that a scientist can make. Since that is contrary to standard propensity theories, I will not investigate that possibility any further; instead, I now come to my proposed alternative to which none of Lane's objections apply.

4. h is none of the above. h is any one of the various equivalent *statements* of a hypothesis by the community of people who understand that hypothesis.

One can construct cases in which this idea is difficult to apply; and this problem forms one of the main limitations of the likelihood principle. But

in the sort of Kuhnian normal science which I have primarily set out to examine, this option is completely unproblematic. Moreover, it is very similar to Lane's option 1: according to my option 4, h directly tells us which distribution p to use; but h does not actually *consist* of p. This distinction serves as a precise way to avoid the problem I noted under Lane's objection 1. It is true, as Lane says, that if h was p it would be wrong to use it directly in the mixed experiment $O^* = (J \otimes X, h, p^*)$, if only because it would have the wrong size (or, in the continuous case, dimensionality). But if h is a scientific hypothesis which tells us what p is, it is perfectly straightforward for it to also tell us what p^* is in the mixed experiment.

It is not surprising that the distinction between options 1 and 4 has escaped Lane's notice, because in many cases the most natural way for a scientist to write down a hypothesis h is actually to write down p! But she does not mean h as a purely mathematical object, equal to everything it's isomorphic to[100] (or, if she does, I don't, when I use it in an instance of the likelihood principle). Rather, she means h as a linguistic direction for obtaining probabilities — in which guise it is easily used in mixed experiments, with a flexibility only limited by the natural language in which the scientist is working.

Earman, in his book on Bayesianism (p. 35), comes to the conclusion that this is how Bayesian statisticians operate: they assign "probabilities

100. If the reader has any doubt about this, consider a scientist who says, "I have the most excellent hypothesis about temperature inversions over Los Angeles. It is . . . [some convoluted equation]." What would we make of a scientist who replies by saying, "My hypothesis about the distribution of temples in Angkor Wat is also . . . [the same convoluted equation]." We would say that these two hypotheses are *related* in an interesting way; but we would not say that they were the *same hypothesis*. We would distinguish between my option 4 and Lane's option 1, and we would prefer option 4.

to objects that express propositions, namely sentences". Earman comes to this conclusion not as a solution to Lane's problem (which he does not consider) but as the most reasonable explication of how scientists actually work. My own point is normative rather than descriptive, but a little descriptive support does it no harm.

2. OBJECTION 10.2 THE LIKELIHOOD FUNCTION IS NOT WELL DEFINED

Some Bayesians have argued that Bayesianism does not imply the likelihood principle, on the grounds that there is no such thing as an isolated likelihood function (Bayarri et al. 1987). They argue that in a Bayesian analysis there is no principled distinction between the likelihood function and the prior probability function. A related possible assertion which I will consider at the same time is that the likelihood function fails to be well defined for non-Bayesians also, although this latter form of the objection does not appear in the literature.

This objection is motivated in the literature by the fact that Bayesians generally reject the idea that the likelihood principle is useful on its own, because (they say) we need prior probabilities in order to apply the likelihood principle; and once we have admitted the universal necessity of using prior probabilities (they say) we will no longer need to separate the likelihood function from the prior (Bayarri et al. 1987, Berger & Wolpert 1988). Thus, they accept proofs of the likelihood principle, *conditional* on the assumption that a likelihood function has been specified; but they deny that specifying a likelihood function is necessary, and they deny that it is

possible to do so in a principled way. Thus, they believe that the likelihood principle is true, if stated carefully, but not straightforwardly applicable.

Despite decrying the applicability of the likelihood principle in this way, Bayesians in this school see it as a useful weapon with which to combat Frequentism. I like to think of this view as Bayesian Hegelianism, as it sees the likelihood principle as an important part of a historical dialectic which will inevitably lead to a synthesis in which it is no longer required. Such a prediction has been beautifully summarised by Bayarri, DeGroot and Kadane, following a metaphor proposed by Butler (1987, p. 21):

> The [Frequentist] Cheshire Cat vanished quite slowly, first the tail and then the body of frequentist methods. The last visible part was the likelihood [principle] grin, "which remained some time after the rest of it had gone". But that, too, disappeared.
>
> (Bayarri et al. 1987, p. 27)

To return to the objection itself: the claim is that there is no principled definition of the likelihood function because there is no principled way of deciding what should be labelled x (data) and what should be labelled h (hypothesis) in the definition of the likelihood as $p(x_a|h)$.

Bayarri, DeGroot and Kadane's examples all involve the following set-up. (Throughout this section I replace Bayarri, DeGroot and Kadane's y by x_a, x by y, θ by ψ, and f by p, in order to remain consistent with the terminology of chapter 2.)

> Suppose that the random variable Y is not observed but another random variable X is observed with conditional density $p(x_a|y, \psi)$. [Then] it is irrelevant which of the factors on the right-hand side [of

$$p(\psi, y|x_a) = \frac{p(x_a|y)p(y|\psi)p(\psi)}{\int p(x_a, y, \psi)}]$$

> are regarded as part of the [likelihood function] and which are regarded as part of the prior distribution.
>
> (Bayarri et al. 1987, pp. 6–7)

This is correct: although elsewhere I have presented the Bayesian method as if it distinguished between the likelihood function and the prior probability function, mathematically speaking such a distinction is not needed once the above equation has been specified. In contrast, we do have to distinguish the likelihood function in order to apply the likelihood principle. Three natural choices are $p(x_a|\psi)$, $p(x_a, y|\psi)$ and $p(x_a|y, \psi)$ [$\equiv p(x_a|y)$], but there is no natural way to choose between these three possibilities . . . or so Bayarri, DeGroot and Kadane claim.

The problem for the likelihood principle, as thus stated, is very easily solved. One need merely specify what one means by "likelihood function". I have already done this, in chapter 2: for me, the likelihood function is always $p(x_a|y, \psi)$. As Berliner (1987, p. 19) correctly notes, the likelihood principle "applies equally well, though separately, in each of the potential cases [which Bayarri, DeGroot and Kadane] enumerate", so my solution is perfectly adequate, as would be any other solution which serves to disambiguate the term "likelihood function".

However, it may appear that a problem remains, since others may disambiguate the likelihood function differently from me. For example, Bayarri, DeGroot and Kadane imagine a case in which two doxastic agents see the same observation, and analyse it using the same mathematical model except that one of them introduces an unobserved variable y into

the model while the other does not. This leads the two agents to define the likelihood function in different ways, following which they cannot use the likelihood principle to compare their results.

To see that my version of the likelihood principle still applies, we have only to note that these two agents are using different hypothesis spaces H: for one of them, H includes a specification of an unobserved variable, while for the other it does not. Given a fixed H (which is an explicit precondition of my version of the likelihood principle), only one likelihood function is possible, namely $p(x_a|h \in H)$. (Note that they agree on x_a; otherwise no joint analysis of any sort would be possible.) A merely practical problem remains if neither of the two agents accepts the other's parameterisation of the hypothesis space, but there is no reason why this should happen, since the two parameterisations essentially agree with each other (more precisely, one parameterisation is easily reducible to the other by taking a marginal distribution with respect to y).

My reply to this objection is essentially the same as a reply due to Berliner. He states his definition of the likelihood function as follows:

> The [likelihood function] is that function of the quantities of interest which is the carrier of the information concerning those quantities provided by the observed variables.
>
> (Berliner 1987, p. 19)[101]

101. Berliner's definition, unlike mine $\left(p(x_a|h \in H)\text{, for some fixed } H\right)$, makes the problem of specifying the likelihood function seem worse than it is. We do not need the apparently vague term "carrier of the information"; all we need is a unique specification of the hypothesis space.

Berger and Wolpert agree that definitions of the likelihood function such as mine and Berliner's solve the problem posed by Bayarri, DeGroot and Kadane.[102]

Berger and Wolpert give a different reply to the current objection. This reply functions as a wicket-keeper for my purposes: I think it is right but less helpful than the replies I have given above, and hence best ignored by those who find the replies I have already given convincing. For the as yet unconvinced, the third reply runs as follows:

> [We] view [this point] as tangential to the LP [likelihood principle]. The LP leaps into action *after* [the likelihood function has] been defined, and $X = [x_a]$ observed. The process of getting to this point is inherently vague and rather arbitrary; but that doesn't alter the fact that, having reached this point and assuming that the model is correct, all information about θ . . . is contained in [the likelihood function] for the given data.
>
> (Berger & Wolpert 1988, p. 39)

To see that Berger and Wolpert's reply is right, it is only necessary to look at the likelihood principle as I have worded it in chapter 8. The assumption

---

102. These authors note (correctly) that such solutions may be misleading, since my definition of the likelihood function for the purposes of applying the likelihood principle is not always the best definition for the purposes of maximum likelihood estimation (defined in chapter 5) (Bayarri, DeGroot & Kadane 1987, pp. 7–8; Berger & Wolpert 1988, p. 39). This is of course irrelevant to the work of this thesis since the only use I make of my definition is to defend the likelihood principle, but it should be borne in mind in the unlikely event that my definition is adopted widely. It is also relevant to anyone who thinks that the method of maximum likelihood is uniquely defined. This may be a problem for advocates of inference to the best explanation.

An alternative reply to Bayarri, DeGroot and Kadane's problem, due to Butler (1987, p. 21), is to define the likelihood function relative to the "model and inferential aim" of the agent. In some cases this may yield a different likelihood function from mine, but this raises no inconsistencies because the likelihood principle applies (separately) to both likelihood functions. Butler's definition is better suited than mine to maximum likelihood estimation, but I do not adopt it because it is open to a charge of excessive subjectivity (at least prima facie).

that the likelihood function has been defined before the likelihood principle becomes applicable is simply my Well Defined Likelihood Function assumption.

3. OBJECTION 10.3
THE LIKELIHOOD PRINCIPLE IS UNIMPORTANT BECAUSE IT DOES NOT TELL US HOW TO PERFORM STATISTICAL INFERENCE

This objection is suggested by (Berger & Wolpert 1988, p. 2).

It is true that the likelihood principle does not tell us how to perform statistical inference; it only tells us how not to. However, since the ways in which it tells us not to include almost all of the commonest statistical methods (namely, Frequentist methods), it is important. In addition, my case study in chapter 15 shows how the likelihood principle can at least suggest, if not mandate, promising statistical methods.

In the next chapter, I move on from objections to the clarity of the likelihood principle to objections based on conflicts between the likelihood principle and other principles, including cases in which the likelihood principle seems at first sight to lead to incorrect analysis of specific statistical models.

— 11 —
Conflicts With the Likelihood Principle

This chapter discusses objections which are based on conflicts between the likelihood principle and other principles and practices.

1. OBJECTION 11.1
THE LIKELIHOOD PRINCIPLE UNDERMINES STATISTICS AS CURRENTLY PRACTISED

By far the most influential argument against the likelihood principle is hinted at more often than stated, and is rather unphilosophical in nature. This most influential of arguments is that statisticians successfully make inferences from data to hypotheses using Frequentist methods which contradict the likelihood principle. Thus, it is claimed, regardless of what is wrong with the likelihood principle, *something* must be, for it rules out the use of exactly the methods that seem to be most successful. A rare explicit statement of this objection is in (Mayo 1996, p. 362).

I have three and a half answers to this objection.

Firstly, the likelihood principle does not entail that the conclusions drawn by Frequentist methods are wrong; it only entails that statisticians can do better than to choose methods on the basis of their Frequentist properties. It therefore does not rule out any token statistical procedures, only methods for choosing procedures. Incorrect methods for choosing procedures may, as it happens, have chosen good procedures. I do not

have space to discuss whether this is really plausible; I offer it in order to question where the burden of proof lies rather than as a knock-down answer to the objection.

Secondly, and more importantly, I deny that Frequentist methods are generally successful. The reasons why we might think they are successful are twofold: that they are successful in their own terms, instantiating as they do guaranteed low error rates; and that applied science, which rests on Frequentist methods, produces successful technology.

But if Frequentist methods are successful in their own terms that proves nothing about whether the Frequentist way of evaluating inference procedures is the one we should use.[103] And that Frequentist methods produce successful technology, while it shows that Frequentist methods are not *sufficiently* bad to entirely disrupt technological progress, does not show that they are generally successful, nor that they are more successful than the alternatives.

This brings me to my final answer to the objection, namely that it only succeeds if the alternatives to Frequentist methods are *un*successful. I am not aware of any empirical reasons to think that Bayesian methods (for example) are unsuccessful. On the contrary, in chapter 15 I present a prima facie successful use of Bayesian methods to solve a problem to which Frequentist methods offer only an impractical solution. Moreover, in the

103. It is also *false* that Frequentist methods are generally successful in their own terms, as they guarantee that both type I error and type II error will be small only if sample sizes are large and measurement error is fully modelled, neither of which caveats is commonly observed. The most obvious cases in which these caveats are broken are in psychometric research, in which sample sizes of under 20 are the norm and in which questionnaires which are known to correlate very badly with the mental states which they purport to measure are treated as if they had no measurement error at all. But I do not claim to have conclusively demonstrated the falsity of the claim that Frequentism is successful in its own terms: that would require an unmanageably large survey of the uses of statistical inference. So I count this as only half an answer to the objection.

few areas in which Bayesian methods are the norm (for example, analysis of noisy digital images, and email spam filtering) they appear to be admirably successful.

2. OBJECTION 11.2
THERE ARE COUNTER-EXAMPLES TO THE LIKELIHOOD PRINCIPLE

A counter-example to the likelihood principle is, of course, any case in which two likelihood functions are derived from a situation fitting within the conditions of applicability of the likelihood principle, and are proportional, and yet ought to lead to different conclusions. I deny that there are any such cases. I present the supposed counter-examples which have appeared in the literature and explain why the two likelihood functions in question need not lead to different inferences.

OBJECTION 11.2.1
FRASER'S EXAMPLE

A form of this example was first suggested in (Fraser 1963, pp. 642–643). I will give an example from (Evans et al. 1986, pp. 186–187) which is essentially similar but which has been discussed more widely in the literature.

> Consider $\lceil X \rceil = \{1, 2, \ldots\}$, and let the distribution for $\lceil X \rceil$ be uniform on $\{\lfloor \theta/2 \rfloor, 2\theta, 2\theta + 1\}$, where $\lfloor s \rfloor$ is the greatest-integer function except that $\lfloor \frac{1}{2} \rfloor$ is taken to be 1.
>
> (Evans et al. 1986, pp. 186–187)

In other words, for some parameter θ, the probability of observing anything except $\lfloor \theta / 2 \rfloor$, 2θ or $2\theta + 1$ is zero, while the probability of observing each of those three options is $\frac{1}{3}$.

Probably the only way to understand this example is to draw the following table of values of $p(x|\theta)$.

	x=1	x=2	x=3	x=4	x=5	x=6	x=7	x=8	x=9	...
$\theta = 1$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0	0	0	0	$\cdots$
$\theta = 2$	$\frac{1}{3}$	0	0	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0	0	$\cdots$
$\theta = 3$	$\frac{1}{3}$	0	0	0	0	$\frac{1}{3}$	$\frac{1}{3}$	0	0	$\cdots$
$\theta = 4$	0	$\frac{1}{3}$	0	0	0	0	0	$\frac{1}{3}$	$\frac{1}{3}$	$\cdots$
$\theta = 5$	0	$\frac{1}{3}$	0	0	0	0	0	0	0	$\cdots$
$\theta = 6$	0	0	$\frac{1}{3}$	0	0	0	0	0	0	$\cdots$
$\theta = 7$	0	0	$\frac{1}{3}$	0	0	0	0	0	0	$\cdots$
$\theta = 8$	0	0	0	$\frac{1}{3}$	0	0	0	0	0	$\cdots$
$\theta = 9$	0	0	0	$\frac{1}{3}$	0	0	0	0	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Table 4

Note that the table is symmetrical in x and θ.

Evans, Fraser and Monette continue:

> For a given $\lfloor x \rfloor$, the likelihood function is flat on three possible θ values$\lfloor . \rfloor$

This is clearly right. In any given column, there is nothing to choose between the three values of θ which have non-zero probability. This is the

only conclusion compatible with the likelihood principle, which says that our inferences can depend on x only via the likelihood function. Hence, from the likelihood point of view, once x_a is observed each of three values of θ is equally well supported. But:

> an examination of the probability matrix shows that choosing the smallest of the three possible θ-values provides a confidence procedure at level $\frac{2}{3}$, that is, one of the three θ-values (each with the same likelihood) is a 2-to-1 favourite.
>
> (Evans et al. 1986, p. 187)

The table shows that this is right too. Suppose we fix θ at 2, for example. The values $x = 4$ and $x = 5$ are twice as likely, put together, as $x = 1$. So any policy which gets θ right when we observe $x = 4$ and when we observe $x = 5$ is twice as good as one which only gets θ right when we observe $x = 1$. And the same is true for *any* value of θ: for any value of θ, a policy which gets θ right when we observe one of the two larger x values compatible with the θ in question is twice as good as one which only gets θ right when we observe the smallest of the x values. Now, the proposed policy of choosing the smallest plausible θ for a given x is just such a policy. Suppose once again that θ is actually 2. If we follow the proposed policy, we will get θ right in $2/3$ of the plausible cases (when we observe $x = 4$ or $x = 5$); and similarly for any value of θ. Since one value of θ is right (according to the model), even though we do not know which one, and since this policy is apparently such a good policy for *any* fixed θ, we should (Evans, Fraser and Monette imply) adopt this policy. Once we have observed x_a, we should estimate θ as the smallest of the three values compatible with the observation.

And yet estimating θ according to this policy contradicts the likelihood principle which, as we have already seen, says that for any observed x_a all three plausible values of θ are equally well supported. (We may not choose the best-supported θ, because of our prior probabilities or other considerations, but that is orthogonal to what is at issue in this example.)

An easy solution to this problem would be to note that Evans, Fraser and Monette's analysis begs the question of whether we should take any notice of Frequentist evaluations of the proposed procedure. After all, we already know that the likelihood principle is incompatible with Frequentist analysis in many cases, and strictly speaking this is all that the example tells us. It is not news. However, the example shows a case in which our intuitions are particularly likely to pull both ways. It may well seem to the reader that in this particular case we ought to opt for the policy which Evans, Fraser and Monette recommend. In order to show that even in this sort of case — perhaps the worst possible case for the likelihood principle, from the point of view of clashes of intuitions — the likelihood principle is still clearly right, I will criticise Evans, Fraser and Monette's proposed method of estimating θ directly, instead of relying on the general criticisms I have already made of Frequentist methods.

I would like to open my criticism of Evans, Fraser and Monette's analysis of this example with a story:

> the teacher asked her to imagine she was an Eskimo walking across the North Pole when she was suddenly attacked by a huge polar bear.
>
> 'What would you do?' the teacher asked.
>
> 'I'd throw a spear at him,' the girl answered.

'And what would you do if a second polar bear appeared?' the teacher asked.

'I'd throw another spear at him.'

'And what if a third and a fourth and a fifth bear attacked?'

'I'd throw three more spears,' the girl answered.

Then the teacher said, 'Hang on, where are you getting all the spears from?'

And the girl said, 'The same place you're getting all the polar bears.'

(Ball 2001, pp. 18–19)

The moral of this parable is that we should ask where Evans, Fraser and Monette are getting their infinite list of values of x from. (The same place as the spears?) Each x must be finite since it is a member of the real numbers and, moreover, if the list comes from any physically describable source then the length of the list must be finite and hence x must have an upper bound. This is so even if the list comes from a physical source which is in principle unbounded, because epistemic agents such as humans can only explicitly list a finite number of quantities before dying of old age. One could argue that we may well not know what the bound on $x$is, and hence that the table above is a reasonable representation of our state of knowledge about x. That seems fair enough. But since nevertheless x *is* bounded, albeit at a possibly unknown bound, let us represent the bound by B and see whether we can draw any conclusions which are valid on any (finite) value of B. It will turn out that we can, because Evans, Fraser and Monette's analysis turns on the possible values of x being literally unbounded, not merely bounded by an unknown bound.

Have another look at the table, this time cutting it off at $x = B$. I will draw this as if B were 3, although we can imagine it to be as big as we like

just as long as it is finite. Omitting rows in which all the values are zero (i.e., rows in which $\theta > 7$), the table looks like this:

	x=1	x=2	x=3
$\theta = 1$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$\theta = 2$	$\frac{1}{3}$	0	0
$\theta = 3$	$\frac{1}{3}$	0	0
$\theta = 4$	0	$\frac{1}{3}$	0
$\theta = 5$	0	$\frac{1}{3}$	0
$\theta = 6$	0	0	$\frac{1}{3}$
$\theta = 7$	0	0	$\frac{1}{3}$

Table 5

Now the policy of choosing the smallest θ for the observed x_a is no longer sensible. Recall that the justification for the policy was that it was on to a good thing for any fixed θ. But this is no longer the case. If θ is 1 then the policy is still good, for it gets θ right on any of the three values of x. But if θ is anywhere from 2 to 7, the policy is guaranteed to get θ wrong.

This argument works for any value of B: imagining the table to be larger shows that for any finite value of B the Frequentist error rate of the proposed procedure is $\frac{1}{3}$ or less for most values of θ. So this supposed counter-example to the likelihood principle fails, provided only that x has some bound, however large.

An analysis similar to this is given by Hill in (Berger & Wolpert 1988, pp. 167–171), although Hill uses decision-theoretic analysis where I stick strictly to an inferential analysis with no mention of utilities or

loss functions. Other authors such as Berger & Wolpert (1988) give an analysis of the example which solves the problem on the assumption that Bayesianism is right, which seems to me to miss the point of the objection.

OBJECTION 11.2.2
EXAMPLES WHICH RELY ON IMPROPER PRIORS

Recall that improper priors are prior probability functions which do not sum to 1 over the hypothesis space. The likelihood principle is incompatible with such functions, in the sense that the joint use of the likelihood principle and improper priors can lead to inconsistent inferences, as I will illustrate in a moment.

In the literature, this objection is sometimes phrased in a much more aggressive way, by saying that the likelihood principle is wrong simpliciter and by supporting that claim with examples which demonstrate that the likelihood principle leads to incoherence in plausible inference scenarios. I collect here a number of such scenarios which depend on the use of Bayesian methods with improper prior probability functions. I will admit that these scenarios lead to incoherence; but I will exonerate the likelihood principle by arguing that improper priors are illicit.

The following supposed counter-example to the likelihood principle is adapted from (Stein 1962).

Suppose a statistical experiment has two possible measurements, x and y, with $x \in X = (-\infty, \infty)$ such that

$$X \sim \text{ Normal } (\theta, \sigma^2)$$

and $y \in \Upsilon = (0, b\theta)$ such that

$$p(y|\theta) = \frac{c}{y} e^{-\frac{1}{2}d^2(1-\frac{\theta}{y})^2}$$

where σ is known, c is the normalising constant

$$c = \frac{1}{\int \frac{1}{y} e^{-\frac{1}{2}d^2(1-\frac{\theta}{y})^2} dy},$$

$d = 50$ and $b = 10^{10^{1000}}$.

Now suppose that we observe either $x_a = \sigma d$ or $y_a = \sigma d$. Then for all θ, $p(x_a|\theta) \propto p(y_a|\theta)$ except for a term in y / b, which is negligible since b is so large. In other words, x_a generates practically the same likelihood function as y_a. So, according to the likelihood principle, we must draw the same conclusions about an experiment which observes x_a as about one which observes y_a.[104]

Stein observes that the following interval is a 95% Neyman-Pearson confidence interval for θ:

$$\left(x_a - 1.96\sigma, x_a + 1.96\sigma\right)$$

and so, by the likelihood principle, the following interval must also be a 95% Neyman-Pearson confidence interval for θ:

$$\left(y_a - 1.96\frac{y}{d}, y_a + 1.96\frac{y}{d}\right).$$

And yet the (Frequentist) probability of y falling into that interval on repetitions of such an experiment is less than $\frac{1}{10^{100}}$. So the likelihood principle has caused us to produce an unsatisfactory Frequentist interval.

104. Or so Stein claims. I do not concede that the likelihood principle is always applicable to likelihood functions which are merely approximately proportional to each other; but for the sake of argument let us go along with Stein's claim that it applies in this particular case.

This only shows once again that the likelihood principle is incompatible with Frequentism, and so it is no real objection to the likelihood principle. However, there is a troubling extension of Stein's example due to Basu. Suppose a Bayesian endorses the likelihood principle, and also holds a flat prior probability function for θ. Then she must calculate the same results as the Frequentist (numerically speaking; their interpretations of the results may differ): she must give each of the above intervals a 95% probability, and must also give a probability of less than $\frac{1}{10^{100}}$ to the second interval (Basu 1975, p. 50, translated into the terminology of Berger & Wolpert 1988, p. 134).

The Bayesian prior which leads to this difficulty is an improper prior (one which does not integrate to 1), as recommended by Jeffreys (see chapter 3). In order to fully defend the likelihood principle, I must therefore give some independent reason for being wary of improper priors. I do this in the following section.

Are improper priors satisfactory idealisations?

Commenting on the Stein example discussed above, Basu says:

> Mathematics is a game of idealizations. We must however recognize that some idealizations can be relatively more monstrous than others. . . . the super-idealization of a uniform prior over the infinite half-line $(0, \infty)$ is really terrifying in its monstrosity. Can anyone be ever so ignorant to begin with about a positive parameter θ that he is (infinitely) more certain that θ lies in the interval (C, ∞) than in the interval $(0, C)$ — and this for all finite C however large?! Naturally, everything goes completely haywire when such a person, with his . . . all-consuming belief in $\theta > C$ for any finite C, is asked to make an inference about θ

> by observing a variable Υ which is almost sure to be at least 10 times larger than θ itself!
>
> (Basu 1975, p. 52)

I would not like to endorse Basu's piece of philosophy-by-exclamation-mark as it stands, because it does not give us a clear reason to disallow improper priors (only a reason to be unsurprised when they cause trouble), but I would like to take on board Basu's suggestion that Stein's improper priors are unsatisfactory idealisations of any epistemic agent's situation.

Basu's point is that an epistemic agent whose mental state is represented by an improper prior is one who believes that the probability of θ falling in any finite region is zero; consequently (and unlike an agent with a vague but proper prior) she must believe that θ has probability zero of being around the same size as C; consequently, we should not be surprised if her belief state cannot be rationally updated to take account of an event which she counts as essentially impossible[105], such as the observation of $\Upsilon > \theta$.

Berger and Wolpert claim that it is rational to use improper priors as an approximation "[w]hen prior opinions are . . . reflected by a locally noninformative prior (in the region of Θ for which the likelihood function is significant)" (Berger & Wolpert 1988, pp. 135–1366). This is tantamount to saying that improper priors are reasonable whenever they are likely (according to the model in use) to give similar results to a proper prior, because regions of X in which the likelihood function is small are unlikely to

105. One might respond to Basu that events with probability zero can occur, and hence are not impossible. This may be true, but such events are only anticipated by an epistemic agent when they fall in regions of a probability density function which have a non-zero measure, the zero probability of the events themselves being an artifact of our representation of continuous probability on a real axis. In contrast, the event which Basu's epistemic agent cannot cope with not only has zero probability itself but also occurs in a large region of zero probability.

be observed. However, sometimes such a region *is* observed, and it would not make sense to apply Berger and Wolpert's reasoning retrospectively in such a case; and therefore it would be dangerous to hold it as a general principle. The Stein example rests on assuming that such a case is observed, so the Stein example shows that Berger and Wolpert's suggestion can lead to contradictory inferences.

Hill (Berger & Wolpert 1988, p. 167–171) argues in more generality that improper priors can be used to approximate flat (but bounded) proper priors whenever there is a physical limit on (the absolute value of) the size of the possible observations which, arguably, is always.[106] Hill's argument is not essential to my discussion, since I have no need to support the use of improper priors; I mention it only to show, as a matter of separate interest, that the debate on this issue is still open.

I conclude that Stein's improper priors (in the mathematical sense) are improper (in a normative sense). This disposes of any remaining worries about Basu's version of Stein's supposed counter-example to the likelihood principle.

The same reply as I have given to Stein serves to deal with other examples in which an improper prior is shown to introduce difficulties into a Bayesian analysis, such as a number of variations on Fraser's example (Goldstein & Howard 1991) and a well-known example due to Stone (1976) which is occasionally proposed (although never by Stone himself, according to his (1991)) as a putative counter-example to the likelihood principle. No new philosophical issues are raised in these other examples.

---

106. As Hill points out, even if there are no limits on the sizes of parameters in nature there certainly are limits on the sizes of physical quantities which finite epistemic agents can report.

3. OBJECTION 11.3
AKAIKE'S UNBIASED ESTIMATOR IS PREFERABLE TO THE LIKELIHOOD PRINCIPLE

Forster and Sober claim that a major goal of statistical inference is to produce a model which is "predictively accurate", in the sense that it makes predictions which are good at predicting as-yet-unseen data. They note that this is a big ask: "the predictive accuracy of a model depends on what the true underlying distribution is. In making an inference, we of course don't know in advance what the truth is. [So] maximizing predictive accuracy . . . so far . . . appears to be epistemologically inaccessible." (Forster & Sober 2004a, p. 160)

They then state that, despite this apparently knock-down argument against the accessibility of predictive accuracy, "Akaike has shown that predictive accuracy is epistemologically accessible" after all (and this claim is repeated in (Forster 2002, Sober 2002a)) by demonstrating, under some fairly mild statistical and epistemic assumptions,

> that an unbiased estimate of a model's predictive accuracy can be obtained by taking the log-likelihood of its likeliest case, relative to the data at hand, and correcting that best-case likelihood with a penalty for complexity:
>
> > An unbiased estimate of the predictive accuracy of model M = Log Pr[Data$|L(M)$] $- k$ [where] k is the number of adjustable parameters in the model
>
> (Forster & Sober 2004a, p. 161)[107]

107. In fact, it is Akaike's *estimator* (the function), not his *estimate* (a realised value of the function) which is unbiased. There is no such thing as an unbiased estimate, as we will see. The distinction between estimator and estimate will be particularly important in my reply to objection 11.4.

Foster and Sober do not give any further justification for caring about Akaike's estimator, so presumably they see its unbiasedness as the property which should recommend it to us. Their reply to earlier criticism seems to confirm this (Forster & Sober 2004b).

It then transpires that the use of Akaike's estimator contradicts the likelihood principle in some cases. This completes Forster and Sober's objection: in cases in which the two conflict, they say, we should prefer Akaike's estimator to the likelihood principle, and hence the likelihood principle is false.

I will not attempt to show that Akaike's criterion is unimportant,[108] but I will rebut Forster and Sober's reasons for thinking it can be used to overrule the likelihood principle.

Forster and Sober's criticism amounts to citing a Frequentist principle which gives a different result from the likelihood principle. Of course that is going to contradict the likelihood principle — Frequentist methods do, as is well known. Why, though, should we just assume that the Frequentist approach is right, as Forster and Sober do? I will give reasons to think that the Frequentist approach is wrong in the particular use of it that Forster and Sober make, which is to support the criterion of unbiasedness in an estimator used for inference after the data have been observed.

As we saw earlier, Forster and Sober at first thought that "the predictive accuracy of a model depends on what the true underlying distribution is" and hence was not something we could know at the time of doing a

108. Having said that, it is easy to show that Forster and Sober's use of Akaike's criterion cannot be the final word on statistical inference. This is because it cannot be right to imply that the best estimate of a model's predictive accuracy depends *only* on the properties of the model's likeliest case. This would mean that a maximally vague model which contains a true case would count as predictively accurate even if the true case were effectively swamped in the model by many dreadfully inaccurate cases.

statistical inference. This is because the only way in which we can know anything about predictive accuracy, over and above what the likelihood function tells us about it (and over and above a Bayesian prior distribution, for those who believe in such things), is if we already know the parameters we are trying to estimate, in which case the statistical inference in question is completely superfluous. But Forster and Sober were able to pull a rabbit out of a hat: they discovered an unbiased estimator of the predictive accuracy of an estimate which *does* tell us something over and above the likelihood function. I will put the rabbit back into the hat. Forster and Sober were right in the first place: we cannot know the predictive accuracy of our methods unless we know the truth about the parameters we are trying to estimate.

Forster and Sober's choice of Akaike's criterion rests on the fact that it is an unbiased estimator,[109] but they do not give any reason for preferring unbiased estimators. A response which is obvious to anyone familiar with the literature on Bayesian statistical inference is that lack of bias (in the technical sense) gives us no reason to approve of an estimator. In itself this perhaps does not bother Forster and Sober, because they are not Bayesians, presumably because they distrust the prior probability distributions required for Bayesian inference; but if we look at the reasoning about unbiasedness which is commonplace in the Bayesian literature, and which I outline below, we will see that the reasoning makes no use of prior distributions, and that one need not be Bayesian to accept it.

The need to investigate unbiasedness will make my reply to Forster and Sober rather long-winded. I will examine what unbiased estimators

109. In fact, contra (Forster & Sober 2004a), Akaike's estimator is not generally an unbiased estimator (Boik 2004, Forster & Sober 2004b), but it sometimes is, and to simplify the argument I will pretend it always is.

are, briefly look at how they are discussed in the literature, give arguments against relying on them, and then give a theory tentatively explaining their spurious appeal. (Of course this theory is not essential for my argument, but it does make my conclusion more plausible: without it, it would seem as if I were saying that the world had gone mad.)

THE DEFINITION OF AN UNBIASED ESTIMATOR

Any function $\hat{\theta}$ which is used to estimate an unknown parameter θ is known as an *estimator* of θ.

An estimator $\hat{\theta}$ is an *unbiased* estimator of θ if and only if

$$\int \hat{\theta}(x)p(x|\theta)dx = \theta$$

where the integration is, of course, taken over the space of observations, X.

I cannot state my view of unbiasedness any better than it was stated by Hacking in 1965, although I will give more detailed arguments for the view than Hacking did.

> It has quite often been proposed that estimators should be unbiased, or at any rate that the best estimators are in fact unbiased. The thesis is no longer as fashionable as it once was, probably because no good reason for it has ever been given. Notice that there is not only no reason for believing that, in general, an unbiased estimator will give a better individual estimate than some biased estimator. There is also no reason for believing that in general unbiased estimators are better on the average than biased ones. For an estimator can on the average be persistently awful, but as long as its errors are of opposite sign, it may still be unbiased, and have an average estimate equal to the true value.

> . . . it might be true that some very good estimators are unbiased, but this would be an incidental fact. We cannot use unbiasedness as a criterion of excellence.
>
> (Hacking 1965, pp. 182–183)[110]

Applied statisticians do often expect estimators to be unbiased. One reason for this is that restricting attention to unbiased estimators is a convenient way to cull an otherwise overwhelming field of possibilities. This is a pragmatic consideration in the most superficial sense of the term. (Not that this consideration is *bad*; it is merely unimportant.) I will discuss later another, psychological, reason why statisticians might prefer unbiased estimators; but first I will consider whether there is some less pragmatic, more strongly normative reason.

Neyman claimed that:

> [t]he advantage of the unbiased estimates and the justification of their use lies in the fact that in cases frequently met the probability of their differing very much from the estimated parameters is small.
>
> (Neyman 1967, p. 259)

There are two problems with this justification. Firstly, it is simply false, if "cases frequently met" is meant to include all the cases in which Neyman and his successors recommend that we use unbiased estimators. Secondly, it is not strong enough to justify Forster and Sober's argument, which requires that unbiased estimators are *always* desirable.

---

110. Hacking was writing several decades after the invention of unbiasedness, so the failure of statisticians to provide a rationale for it was not a temporary oversight; nor was it an oversight which has since been corrected, as we will see in a moment. Hacking's view that unbiasedness is no criterion of excellence was not new in the 1960s, and arguably the most influential statistician ever, R. A. Fisher, saw no use at all for unbiasedness, despite supporting almost every other criterion for statistical inference.

Apart from Neyman, theoretical statisticians do — often — say that unbiasedness is a desirable property in and of itself, but without ever saying why. I am genuinely perplexed by this. The theoretical statisticians I am thinking of are authors who are in masterful command of the mathematics behind their assertions, so they are not omitting to mention any putatively desirable properties of unbiased estimators through failure to understand them . . . and yet they simply do not mention any such properties. I cannot, of course, survey here the hundreds of books on the topic by reputable statisticians, but I will quote briefly from two authorities on statistical inference to give the flavour of the literature.

Kendall and Stuart on unbiasedness

> Consider the sampling distribution of an estimator t. If the estimator is consistent, its distribution must, for large samples, have a central value in the neighbourhood of θ. We may choose among the class of consistent estimators by requiring that θ shall be equated to this central value not merely for large, but for all samples.
>
> If we require that for all n and θ the mean value of t shall be θ, i.e. that
>
> $$E(t) = \theta,$$
>
> we call t an *unbiased* estimator of θ. This is an unfortunate word, like so many in statistics. The mean value is used, rather than the median or the mode, for its mathematical convenience. This is perfectly legitimate, but the term should not be allowed to convey non-technical overtones.
>
> (Stuart et al. 1999, pp. 4–5)

I give arguments below for thinking that unbiasedness, $E(t) = \theta$, is not really a desideratum, followed by a speculative argument suggesting why people might think it is. Note that Stuart et al. do not explicitly disagree with me; they only say that this is *one* way of narrowing the class of consistent estimators, which would otherwise be inconveniently large, and they give no other justification for it at all. Indeed, they later give a reason to think that unbiasedness is *not* a desideratum in general:

> Our discussion in 17.8 shows that consistent estimators are not necessarily unbiased. We have already (Example 14.5) encountered an unbiased estimator that is not consistent. Thus neither property implies the other. . . . In certain circumstances, there may be no unbiased estimator (cf. Exercise 17.12). Even if there is one, it may be forced to give absurd estimates at times, or even always.
>
> (Stuart et al. 1999, p. 5)

When Stuart et al. discuss censoring (unavailable data), they note that censoring makes it hard to preserve unbiasedness, and comment in this context:

> A user of statistical methods must decide upon the properties considered desirable in an estimator and, for example, an overly rigid insistence upon unbiasedness may lead to difficulties.
>
> Nevertheless, the notion of unbiasedness has considerable intuitive appeal and many would be reluctant to abandon it.
>
> (Stuart et al. 1999, pp. 432)[111]

111. Stuart et al. also mention the concept of an "unbiased estimating equation", due to Godambe, which has some of the properties of unbiasedness but which is not affected by censoring. Since this new concept is much more general than unbiasedness, I do not expect that it would be able to play the role of unbiasedness in an alternative version of Forster and Sober's criticism of the likelihood principle. In any case, Stuart et al. present no justification for the new concept but do say, rather inconclusively, that "[i]t may be argued that this revised concept of unbiasedness gives away too much" (Stuart et al. 1999, pp. 432). So their discussion on this point is unhelpful: I mention it only for completeness.

Why does unbiasedness have intuitive appeal? We are not told here, nor anywhere else that I can find in the literature. (The statistical literature is far too vast to permit an exhaustive search, but I have searched hard.) I will suggest an answer in a later section.

Casella and Berger on unbiasedness

> [A] comparison of estimators based on MSE [mean squared error] considerations may not yield a clear favorite. Indeed, there is no one "best MSE" estimator. Many find this troublesome or annoying, and rather than doing MSE comparisons of candidate estimators, they would rather have a "recommended" one.
>
> The reason that there is no one "best MSE" estimator is that the class of all estimators is too large a class. . . . One way to make the problem of finding a "best" estimator tractable is to limit the class of estimators. A popular way of restricting the class of estimators, the one we consider in this section, is to consider only unbiased estimators.
>
> (Casella & Berger 2002, p. 334)

> The *bias* of a point estimator W of a parameter θ is the difference between the expected value of W and θ; that is, $\text{Bias}_\theta W = E_\theta W - \theta$. An estimator whose bias is identically (in θ) equal to 0 is called *unbiased* and satisfies $E_\theta W = \theta$ for all θ.
>
> Thus, MSE incorporates two components, one measuring the variability of the estimator (precision) and the other measuring its bias (accuracy). An estimator that has good MSE properties has small combined variance and bias.
>
> (Casella & Berger 2002, p. 330)

Casella and Berger partition the mean squared error of an estimator into two components. One component measures the variability of the estimator;

this component is *independent* of the bias. The other component measures its bias. Nowhere do Casella and Berger say why the bias is important, and nowhere do they justify calling it "bias" or its inverse "accuracy". They might as well have called them "squiggle" and "squoggle".

UNBIASEDNESS IS NOT A VIRTUE

So much for what the authorities say about unbiasedness. What should *we* say about unbiasedness? We should say that, despite having a nice mathematical symmetry, unbiased estimators can have wildly unacceptable epistemological properties. Later I will explain how the nice symmetry is compatible with the horrible epistemological properties; but first, let us see what the horrible properties are.

Bernardo & Smith (1994) list four well-known epistemological reasons against requiring an estimator to be unbiased, of which I quote two. Of the other two, one, that sometimes there are no unbiased estimators, is not relevant to Forster and Sober's use of Akaike's theory, while the other, that "the unbiasedness requirement [makes] the answer dependent on the sampling mechanism", would make no sense without a long discussion to disentangle various different types of dependence on experimental design, since biased estimators suffer from the same problem unless we distinguish cases very carefully.

> (ii) . . . unbiased estimators may give nonsensical answers, and no theory exists which specifies conditions under which this can be guaranteed not to happen. For example, . . . if θ is the mean of a Poisson distribution, $\mathrm{Pn}(x|\theta) = e^{-\theta}\theta^x \,/\, x!$, $x = 0, 1, \ldots$, then the *only* unbiased estimator of $e^{-\theta}$, a quantity which [cannot be 1 or 0], is 1 if x is even and 0 if it is odd . . . but—even more

> ridiculously—the *only* unbiased estimate of $e^{-2\theta}$ is $(-1)^x$, leading to the estimate of a probability as -1 (for all odd x)!

So the only unbiased estimator available is often a value which the parameter cannot take according to the definition of the model; and sometimes it is a value which the parameter cannot take on *any* model, as we can see from the fact that a probability cannot be -1. These facts alone should be enough to convince us that we need not feel pressured into using unbiased estimators to estimate anything.

> (iv) ... unbiased estimators may well be unappealing if they lead to large mean squared errors, so that an estimator with small bias and small variance may be preferred to one with zero bias but a large variance.

Point (iv) ought to be particularly telling for Forster and Sober, since they refer to Akaike's estimator as a measure of predictive *accuracy*. Its unbiasedness does not protect it from being an extremely inaccurate measure of accuracy; and whether it is actually accurate or not in a given situation is impossible to assess since (again) we do not know θ.

A further argument against unbiased estimators is that the mean of a posterior probability distribution cannot be an unbiased estimator of any unknown parameter (Casella & Berger 2002, pp. 368–369). For Bayesians this is a knock-down argument. It is also a persuasive argument for those of us who are not entirely subjectivist Bayesians but who believe that the Bayesian mathematical machinery applies in at least some cases.

A typical response to the assertion that unbiased estimators are often bad estimators is to accept the instances but reject the generalisation that we should not care about bias. Here is a typical example of this move:

> Unbiasedness may not be a compelling property of an estimator: there are certainly examples in which the best unbiased estimator is terrible, and other examples where biased estimators more than compensate for their bias through reduced [error rates]. However, substantial bias in the absence of such considerations seems like a bad thing[.]
>
> (Martinsek 1988, p. 58)

This move is always made (as far as I can find) without any good reason being given for why we should care about bias. Martinsek is unusual in giving any reason at all: he cites his own intuition, and the fact that many laypeople agree with him. These arguments from authority might well give us pause, but in the absence of any better arguments for unbiasedness anywhere in the literature they do not bear much weight against Bernardo and Smith's substantial arguments.

So I claim that unbiasedness is no indication of a good estimator. It follows that the unbiasedness of Akaike's estimator is a bad card with which to try to trump the likelihood principle. Since Forster and Sober *are* claiming precedence for a particular use of Akaike's unbiased estimator over the likelihood principle, the fact that unbiased estimators often come unstuck is sufficient to shift the burden of proof away from the defenders of the likelihood principle and onto its attackers: they will have to attack the likelihood principle with something stronger than an unbiased estimator chosen merely because it is unbiased.

I forsee two objections to my claim that the unbiasedness of an estimator is unimportant. (Thanks to Huw Price, Alan Hájek and others for enunciating these objections for me. I have borrowed Alan Hájek's wording of the objections.)

Objection 1: Although unbiased estimators may be bad estimators, they are not bad *in virtue* of their unbiasedness. Their unbiasedness is a good thing, admittedly outweighed by other bad things.

Response: Assuming as most do that mathematical properties are not *causally* related to each other, what can it mean to say that it is not in virtue of unbiasedness that unbiased estimators can have unacceptable properties? (Compare for example the claim that it is not in virtue of being odd that 17 is a prime number.) Presumably it must mean that unbiased estimators need not have unacceptable properties. (Compare: odd numbers need not be prime.) This is true. But in certain particular situations it ceases to be true. As the thesis clearly shows, there are cases in which an estimator, if it is to be unbiased, must have unacceptable properties such as being negative despite representing a probability. (Compare: if we are only considering numbers which are single digits, as sometimes we do, then odd numbers must be prime; so, in such a situation, numbers are prime by virtue of being odd.) I conclude that unbiased estimators *can* have unacceptable epistemological properties in virtue of their unbiasedness, insofar as such a claim means anything.

Objection 2: To be sure, unbiasedness can be trumped by other considerations — e.g. high variance, or inconsistency, or intractability. But if all other things are equal then unbiasedness is a desideratum of an estimator.

Response: I can only take this to mean that if other two estimators are equal in their other desirable properties but one is biased while the other is unbiased then the unbiased one should be chosen. I have two replies to this. Firstly, I do not see any argument in favour of it apart from the bad argument which I will outline below. Secondly, it cannot be argued

in the absence of a complete list of the desirable properties of estimators; for if one could argue, in the absence of such a list, that any property of an estimator is desirable then it is impossible for two such estimators to exist (since they cannot differ in any property and hence must be the same estimator). Both of these arguments could perhaps be challenged by suitable counter-arguments, but I cannot find any such counter-arguments in the literature to date.

I now turn to possible reasons why, despite the apparent nonexistence (to date) of explicit reasons for caring about unbiasedness, people still do care.

AN EXAMPLE OF TALK ABOUT BIAS

When we think of bias with our layman's hat on, we might think of a darts player who tends to hit the board to the left of the bull's eye. (Thanks to Alan Hájek forthis example.)

What does "tends to" mean here? If it means *has a propensity to*, then it illustrates almost perfectly that what people (very reasonably) think "bias" refers to is *not* what it currently refers to in statistics. A player who tends to hit the board to the left of the bull's eye, indeed one who almost always does that, may still be unbiased.

Now let us consider a player whose throws are mostly to the left of the bull's eye *and* are biased to the left in the technical sense. Whether that is bad depends on what, quantitatively, the examiner means by saying that the thrower "tends to" hit the board to the left. It also depends on what counts as bad in a particular context — on the game's scoring system, whether and if so how bets are placed, and so on. Let us compare two

example players, A and J. J tends to hit the board to the left, in the sense that almost all of his throws end up to the left of the bull's eye. However, when he misses the bull he always hits the semi-bull or whatever that ring thing around the bull's eye is called. J is biased and therefore inaccurate in the statistical senses. In lay parlance, however, he is reasonably accurate. A throws to the left exactly as much as he throws to the right, so he is perfectly unbiased. Moreover, for the sake of argument, I will imagine (although I do not need to concede this much) that A also hits the semi-bull whenever he misses the bull, so he is also reasonably accurate in the lay sense. If you like, we can even arrange that A has the same mean squared error, measured from the centre of the bull, as J. However, A hits the bull's eye less often than J does. (This is perfectly consistent with everything else I am stipulating.) Who should we expect to win games of darts: the biased J, or the equally lay-accurate and perfectly unbiased A? The answer is J. So I continue to maintain that the fact that player J is biased is totally unimportant.

If we were to rewrite the example so that J is biased in certain particular ways while A is a good player, we might expect A to win. That would show that some (token) biased estimators are bad estimators. That is consistent with everything I claim. In particular, it is consistent with the claim that we should not care about bias. My claim is that we can see that a particular biased estimator is a bad estimator without calculating its bias. I can, for example, calculate instead how far away from the bull's eye the thrower throws, on average. Or I can calculate his or her expected score.

WHY IS UNBIASEDNESS CONSIDERED GOOD?

The very strength of Bernardo and Smith's arguments may make the reader suspicious. If they are right, why does anybody ever look for unbiased estimators? I have already suggested a pragmatic answer to this question, but now I would like to suggest a more plausible, psychological answer.

A property which it really would be nice for an estimator $\hat{\theta}$ to have is

$$\hat{\theta} = E(\theta) \tag{1}$$

where $E(\theta)$ is the *expected value* (average value) of θ. But a non-Bayesian statistician cannot calculate such a thing, because θ (as defined) is an unknown fixed parameter whose expected value is itself and is, ex hypothesi, unknown. Only Bayesians have a solution to this problem and can calculate $E(\theta)$ in a useful and non-trivial way, using prior probabilities.

Any statistician, Bayesian or non-Bayesian, can, however, consider the equation

$$E(\hat{\theta}) = \theta, \tag{2}$$

which is really shorthand for $E(\hat{\theta}|\theta) = \theta$.

It is generally possible to evaluate (2) without knowing θ, because it is (almost always) possible to calculate $\mathrm{E}(\hat{\theta})$ for each possible value of θ, and it usually eventuates that a suitable choice of $\mathrm{E}(\hat{\theta})$ is equal to θ in all of these possible cases. A pleasant feeling then ensues. This pleasant feeling

is my explanation for the apparent value of seeking unbiased estimators: if we expand equation (2), we get

$$\int \hat{\theta}(x)p(x|\theta)dx = \theta,$$

which is the definition of unbiasedness I gave above. So, in finding unbiased estimators, we are finding estimators which satisfy (2); and in doing that, we feel as though we have satisfied (1).[112]

But we have not. And I submit that it is *only* the superficial similarity between equations (1) and (2) which makes equation (2) seem important. After any amount of data has been collected, $\hat{\theta}$ is still going to have a single value. Knowing that its *unknown* expected value (its average expected value over hypothetical repetitions of the data-gathering process) is equal to the *also unknown* value of θ does us no good at all.

In conclusion, I do not claim that unbiasedness is a *bad* thing, but I do claim that I can find only bad reasons for preferring it in an estimator. Unbiased estimators are like estimators which use only even numbers: they are neither here nor there, inferentially speaking. In the absence of any good reason for preferring unbiasedness, it cannot play a substantive role in objections to the likelihood principle.

4. OBJECTION 11.4
WE SHOULD USE ONLY CONSISTENT ESTIMATORS

An estimator t is **consistent** iff $p(t \longrightarrow \theta$ as the sample size tends to $\infty) = 1$.

112. Equation (1) does not have a name, since most people believe it can't be calculated and all the other people — Bayesians — can show that within their theory it is trivially easy to satisfy.

Unlike unbiasedness, which, as we have seen, is not considered to be important by canonical works such as (Stuart et al. 1999), consistency is taken to be a virtue in almost all thorough presentations of Frequentist inference.

Estimators produced using the likelihood principle are not guaranteed to be consistent. So it would be open to an objector to frame an argument against the likelihood principle by producing a consistent estimator and claiming that that estimator trumps the likelihood principle in certain cases.

Howson and Urbach (1993) give two excellent responses to such an objection. Firstly:

> A corollary of this [objection] is that an estimate's worth depends on who derived it. For suppose statistician *A* employed the sample mean to estimate a population mean, while *B* used some non-consistent . . . function of the sample mean; and imagine that they each arrived at identical estimates from the same sample. [It is perfectly possible to arrive at the same *estimate* (value) from different *estimators* (functions). What it requires in this particular case is that the function of the mean which statistician *B* uses is equal to the mean at the particular sample size that was actually collected.] According to classical [Frequentist] ideas, since these identical estimates have different pedigrees, they must be differently evaluated: one would be 'good', the other 'bad'! This, of course, contradicts the difficult-to-gainsay assumption that logically equivalent statements are equally 'good'.
>
> (Howson & Urbach 1993, p. 233)

There is nothing wrong with evaluating a statement according to who made it (consider indexicals, for a start), so Howson and Urbach's complaint is a little misleading. It is best rephrased as follows. The *only* way to

find out whether an estimate is consistent is to find out which estimator (function) the estimate (value) is taken from. Thus, there is no such thing as an estimate being consistent simpliciter. It can only be consistent relative to some choice or other of estimators. *That* is Howson and Urbach's complaint, more properly stated.

Secondly, Howson and Urbach respond to such an objection by citing

> an example in which an 'inconsistent' method of estimation yields a perfectly satisfactory and confidence-inspiring estimate. Let the goal of the estimation be the mean of some population [parameter] and imagine a scientist eccentrically selecting $\bar{x}+(n-100)\bar{x}^2$ as the estimating statistic, where . . . $\bar{x}$ and n are the sample mean and sample size, respectively. Clearly this odd statistic is not consistent (in the statistical sense), for it diverges ever more sharply from the population mean as the sample is enlarged. Nevertheless, for the special case where $n = 100$, the statistic is just the familiar sample mean, which on intuitive grounds gives a perfectly satisfactory estimate.
>
> (Howson & Urbach 1993, p. 233)

Since the objection we are considering is that we should use *only* consistent estimators, this counter-example is decisive. It is no good to reply that of course we can use inconsistent estimators provided that they give an estimate which coincides with that given by a consistent estimator (as the counter-example obviously does) because, as I have already mentioned, the fact that the definition of consistency is asymptotic ensures that *all* inconsistent estimators *always* give an estimate which coincides with that given by some consistent estimator. In principle it remains open to the objector to the likelihood principle to say that we can use inconsistent estimators provided that they give an estimate which coincides with that

given by some *particular* consistent estimator, but that objection would need a separate justification having nothing to do with consistency.

This concludes my responses to objections to the likelihood principle on the basis of conflicts with other principles and practices. In the next chapter I consider a miscellany of further objections to the likelihood principle.

— 12 —
Further Objections to the Likelihood Principle

This chapter continues my examination of objections to the likelihood principle. A general introduction to these objections is given in chapter 10.

1. OBJECTION 12.1
THERE ARE NO ARGUMENTS IN FAVOUR OF THE LIKELIHOOD PRINCIPLE

Mayo writes:

> Apparently, the LP is regarded by some as so intrinsically plausible that it seems any sensible account of inference should obey it. Bayesians do not seem to think any argument is necessary for this principle, and rest content with echoing Savage's declaration in 1959: "I can scarcely believe that some people resist an idea so patently right". However much Savage deserves reverence, that is still no argument.
>
> (Mayo 1996, pp. 345–346)

If all of the detailed argument in favour of the likelihood principle given in this thesis were entirely original then perhaps this objection would at least have been right in 1996, when Mayo made it. In fact, my work is not that original, as my citations show. (And almost all of the most relevant citations were published before the above objection was made.) In any case, this thesis as a whole shows that the objection is not currently tenable.

Incidentally, Mayo gives a reason for concentrating on Bayesian advocates of the likelihood principle:

> The LP is regarded as having been articulated by non-Bayesian statisticians, principally George Barnard (1947) and R. A. Fisher (1956). [I believe this reference to Fisher is a confusion of the likelihood principle with the method of maximum likelihood — see chapter 5.] But, as it is *their* principle now, I will let the Bayesians do the talking.
>
> (Mayo 1996, p. 339)

This is odd, because of the four book-length monographs to date which discuss the likelihood principle in detail, three are by non-Bayesians, and two of these predate Mayo's claim ((Hacking 1965), (Edwards 1972); (Royall 1997) postdates (Mayo 1996), and (Berger & Wolpert 1988) alone is by Bayesians). It is true that the set of statisticians who *tacitly accept* the likelihood principle is dominated by Bayesians, but the set of authors who *discuss* the likelihood principle is not, so it makes no sense to let only the Bayesians "do the talking".

2. OBJECTION 12.2
THE LIKELIHOOD PRINCIPLE IS LESS WIDELY APPLICABLE THAN I CLAIM

I will discuss two versions of this objection: that the framework of chapter 2 is importantly incomplete (i.e., fails to capture important problems of statistical inference), and that the likelihood principle does not endorse any reasonable methods of statistical inference.

OBJECTION 12.2.1
MY FRAMEWORK IS SERIOUSLY INCOMPLETE

In discussing (Birnbaum 1962), Barnard criticises Birnbaum's assertion that the likelihood principle is widely applicable, on the grounds that Birnbaum's framework, which is similar to mine, fails to capture many important problems of statistical inference.

> [The likelihood principle] applies to those situations, and essentially only to those situations, which are describable . . . in terms of the sample space S, and the parameter space Ω and a probability function f of x and θ defined for x in S and θ in Θ. If these elements constitute the whole of the data of a problem, then it seems to me the likelihood principle is valid. But there are many problems of statistical inference in which we have less than this specified, and there are many other problems in which we have more than this specified. In particular, the simple tests of significance arise, it seems to me, in situations where we do not have a parameter space of hypotheses; we have only a single hypothesis essentially, and the sample space then is the only space of variables present in the problem.
>
> (Barnard 1962, p. 308)

It is certainly true that in such a case we cannot usefully apply the likelihood principle, because the likelihood function will consist of a single point; having nothing to compare that point to, the likelihood principle tells us nothing. In a moment I will argue that such a case need not arise.

Barnard continues:

> The fact that the likelihood principle is inconsistent with significance test procedures in no way, to my mind, implies that significance tests should be thrown overboard; only that the

> domain of applicability of these two ideas should be carefully distinguished.
>
> (Barnard 1962, p. 308)

I believe I have stated the domain of applicability of the likelihood principle more carefully than ever before; so, having taken Barnard's advice about that, I am in a good position to consider his objection.

Barnard's objection is compatible with everything I claimed in chapter 8; but it is incompatible with my claim that the likelihood principle renders Frequentist inference invalid. In particular, Barnard believes that P-values are often required (P-values justified by Fisher's theory in which no alternative hypothesis is required, not by Neyman's).

Barnard's premise is that often "we want to have a single hypothesis with which to confront the data [and ask:] Do they agree with this hypothesis or do they not?" (Savage & discussants 1962, p. 75). The likelihood principle does not help us with this question, as far as either Barnard or I can see, because it says that we must base our inferences on the sample space X only via the observed data x_a. If, in addition to considering only one point in the sample space, we are considering only one point in the hypothesis space, there seems to be only one number on which we can base inferences, namely $p(x_a|h)$, and nothing to which we can compare it. But considering $p(x_a|h)$ raw, as it were (not in comparison to anything, just in terms of its absolute magnitude) does not give sensible inferences, because if X is large and h assigns probabilities anywhere near uniformly then $p(x|h)$ will be approximately zero; and if X is infinite (as it often is) and p is a genuine probability function (integrating to 1) then $p(x_a|h)$ will be precisely zero.

Barnard's question, "Do [the data] agree with this hypothesis. . .?", is one which statisticians often ask (and often answer), but I do not believe it is a meaningful question. I will argue that the notion of "agreement" is not a useful notion in a probabilistic situation. In a non-probabilistic situation, the idea of a hypothesis agreeing with data is straightforward, apart from Duhem-Quine problems: the two things agree if and only if they are relevant to each other and do not contradict each other. Both relevance and non-contradiction are symmetrical notions, so it does not matter whether we ask whether the hypothesis agrees with the data or vice versa. But what could agreement mean in a probabilistic situation? The usual statistical answer is that data and hypothesis agree if and only if p(data|hypothesis) is large. But there is an alternative definition: that data and hypothesis agree if and only if p(hypothesis|data) is large. These two numbers are generally very different from each other. As we saw in chapter 4, Frequentist statistical methods use functions of the former number, while the likelihood principle is usually applied by using the latter number. So there is no univocal answer to Barnard's question as stated.

As Barnard knows, the commonest methods for answering his question, namely Neyman-Pearson hypothesis tests, require at least two hypotheses to be specified. I take it, from the quotation, that this is something which he believes he would not be willing to do in some cases, but I do not know why. If we have an arbitrary hypothesis h (and even Barnard is willing to assume that there is always at least one hypothesis available) then the other hypothesis that we need in order to apply the likelihood principle can be the catch-all hypothesis. Barnard may have in mind the fact that the *general* catch-all hypothesis — the logical negation of h, or

equivalently the set-theoretic negation of h within the set of all possible hypotheses — is often undefined. But a more local catch-all hypothesis, the set-theoretic negation of h within the universe of models under consideration (Lipton 1993), is always well defined in the types of mathematical models that statisticians use (see chapter 2). For example, if we take h to be the hypothesis that the effect of AZT on HIV in Australia is a decrease in death rate characterised by a rate ratio of approximately 0.4, we can produce an alternative hypothesis by considering the local catch-all hypothesis h' that the rate ratio is not approximately 0.4 (as opposed to the logical negation of h, which is that it is not the case that the rate ratio is approximately 0.4). All we need to do to counter the argument I am imputing to Barnard is produce a relevant alternative hypothesis of this sort. We may have nothing to say about the *logical* negation of h, because probabilities based on the logical negation of h depend on the probabilities of all sorts of strange possibilia such as (inter alia) the probability that there is no such thing as HIV and therefore no such thing as the rate ratio of AZT in reducing it; but this is not a problem, because h' is perfectly adequate for our current epistemic need.

So I dispute Barnard's claim that we often need Fisherian (single-hypothesis) significance tests. As far as I can see, we never do.

Barnard also claims that sometimes we have *more* information than my framework allows for, and that in these cases too we have to use procedures contrary to the likelihood principle. I admit that when we have more information than my framework allows for — something which is clearly logically possible — I cannot show that the likelihood principle still applies. However, I cannot see any such cases in inference from data to

hypotheses. Prima facie we should think that they are rare or perhaps even nonexistent, bearing in mind that the hypothesis space and sample space in my framework can encode *any* amount of structure. Barnard has given (in various publications) many cases in which he believes pivotal inference (defined in chapter 5) takes advantage of the mathematical structure of a problem in a way which is not part of the *usual* construction of hypothesis and sample spaces; but I cannot find any such case which cannot be covered by my framework.

To the best of my knowledge, Barnard does not suggest any particular such case as a counter-example to the likelihood principle. He does give two general examples of types of structure which may be added to a problem, but neither of these presents any difficulty for the likelihood principle, at least in the form in which I have presented it in this thesis. These two general examples are as follows.

> (1) We may have properties of invariance, and such things, which enable us to make far wider, firmer assertions of a different type; for example, assertions that produce a probability when these extra elements are present.
>
> (Barnard 1962, p. 308)

But I see no argument against incorporating such things into the hypothesis space of my framework, *especially* when they produce a probability.

> (2) And then, of course, there are the decision situations where we have loss functions and other elements given in the problem which may change the character of the answers we give.
>
> (Barnard 1962, p. 308)

Decision problems which specify loss functions (or, equivalently, utilities) are outside the scope of this thesis, so I admit Barnard's charge that my framework is not all-encompassing. However, it is worth noting that the principles of the most prominent version of decision theory, Bayesian decision theory, entail the likelihood principle (Berger 1980, Raiffa & Schlaifer 2000).

OBJECTION 12.2.2
THERE ARE NO ADEQUATE THEORIES OF INFERENCE WHICH OBEY THE LIKELIHOOD PRINCIPLE

The following existing theories of statistical inference obey the likelihood principle (see chapter 3 and chapter 5 for definitions): all forms of Bayesianism except Empirical Bayesianism, and all pure likelihood methods including maximum likelihood estimation and the method of support. Of these, Subjective Bayesianism, Restricted Bayesianism and maximum likelihood estimation are in active use (although much less so than Frequentism). The objection therefore cannot be that there are no theories which obey the likelihood principle; it must be that the theories in question are inadequate in some way.

Objectors to the likelihood principle can argue that only a very general method for applying the principle will do, because a principle which cannot be demonstrably applied in every case does not deserve its name. Neither Restricted Bayesianism, nor the method of support, nor maximum likelihood estimation is as widely applicable as Frequentism, as I showed in chapter 3 and chapter 5. This leaves Subjective Bayesianism as the best competitor to Frequentism.

The likelihood principle is widely associated with Subjective Bayesianism, as demonstrated both by several of the definitions in chapter 8, notably Lindley's, and by many of the attacks on its rationality, notably Mayo's. Subjective Bayesianism is difficult to defend precisely because it is such a complete theory: to defend it properly, a large number of examples would need to be discussed, in addition to a good deal of basic epistemology. (For just some of the details, see (Howson & Urbach 1993).) I cannot rehearse these arguments here, so the question of whether Subjective Bayesianism is an adequate champion for the likelihood principle will have to remain open.

The objection is doing quite well up to this point: I have conceded that there are no widely applicable, practical methods of statistical inference which can easily be demonstrated to be rational and which have decent histories of practical application. However, having a history of practical application is almost irrelevant to a theory's adequacy. What the objector needs to show — or at least make plausible — is not that the *current* theories of statistical inference which obey the likelihood principle cannot *easily be shown to be* rational, but that at no point in the future will there be a theory of statistical inference which obeys the likelihood principle and is rational. This is a hard task; such a hard task that I cannot see how it could be attempted except by attacking the likelihood principle directly, as the other objections I discuss do. But the burden of proof is on the objector since I have motivated the likelihood principle, and will prove it, in ways which do not depend on the existence of a general theory of statistical inference which implements it; so as long as this objection remains an open question, the likelihood principle remains unscathed by it.

To give a bit more flesh to this response, note that I have conceded that Objective Bayesianism does not have much of a history of practical application; but I have certainly not conceded that there is anything wrong with all possible forms of Objective Bayesianism, and thus Objective Bayesianism remains a contender as a practical implementation of the likelihood principle. Nor do I concede that Subjective Bayesianism has been defeated. Moreover, I showed in chapter 5 that there are methods of statistical inference which have not yet been enunciated, and one of these may turn out to be just what we need.

3. OBJECTION 12.3
THE LIKELIHOOD PRINCIPLE ALLOWS SAMPLING TO A FOREGONE CONCLUSION

Mayo examines the problem of sampling to a foregone conclusion by discussing the following example:

> [W]e will imagine that the researchers have an effect they would like to demonstrate, and that they plan to keep experimenting until the data differ statistically significantly, say at the .05 level, from the null hypothesis of "no effect."
>
> (Mayo 1996, p. 338)

As Mayo rightly says, such a procedure is problematic, because one can be sure that such a procedure will achieve statistical significance, regardless of which hypothesis is true . . . not in literally any case, as Mayo goes on to imply, but certainly in many cases.

The literature on the likelihood principle, including books and papers which Mayo herself cites, is full of passages which emphasise that

statisticians using the likelihood principle should not also use significance (P-value) tests. In this case, the point is that Mayo is envisaging a Frequentist rule being used to determine what counts as sampling to a foregone conclusion: namely, the rule that says that sampling to a foregone conclusion has occurred iff a certain P-value is less than 0.05. If, on the contrary, a rule compatible with the likelihood principle is used to determine what counts as sampling to a foregone conclusion, then sampling to a foregone conclusion is no longer inevitable. I will illustrate this later by showing that if we use posterior odds to fashion such a rule then there is no problem. This result is well known in the Bayesian literature.

Mayo comments on one passage which mentions this result, by Savage, and argues against the point in two ways: (a) with an unsupported rhetorical question — "Why should we accept the likelihood principle?" (Mayo 1996, p. 345) — a question to which there are a number of published answers which I have summarised elsewhere, and (b) by saying that the person who convinced Savage of the truth of the likelihood principle, Barnard, has now "changed his mind" (Mayo 1996, p. 345). Although I realise that there is little point in replying to an ad hominem argument with another ad hominem argument, it is interesting to note that the author who Mayo describes as "the most forthright error [i.e., Frequentist] statistician at the 1959 Savage forum", Armitage, later changed his mind to a much greater extent than Barnard did. Armitage moved from roughly Mayo's anti-Bayesian position in 1959 to a pro-Bayesian position in his (1989). Mayo (1996, pp. 343–334) describes Armitage further as "a leader in the development of sequential trials, having devoted whole books to their use and interpretation within the error statistical framework", apparently

not realising that by the time she writes Armitage no longer supports her anti-Bayesian interpretation of his work. Even as early as 1969 Armitage was sufficiently fond of certain Bayesian methods to make a public call for an assessment of their error rate characteristics (Armitage et al. 1969). When such an assessment was made for the first time, in (Grossman et al. 1994), Armitage (the same Armitage) cited it approvingly in his textbook (Armitage & Berry 1994, p. 506; Armitage et al. 2002, p. 622).

A REPLY TO THE OBJECTION

Besides such ad hominem arguments, there is, of course, a more direct way to clear the likelihood principle from the charge which Mayo incorrectly attributes to him.

The reply is straightforward. It is to note that the likelihood principle applies to analyses of *observations*, not to analyses of *significance tests*. (This was made abundantly clear in chapter 2 and again in chapter 8.) Consequently, a correct application of the likelihood principle to the case Barnard discusses would be as follows: ignore the significance tests and conduct a new analysis. So the point which Mayo sees Barnard as making is simply irrelevant to the likelihood principle. Thus, the likelihood principle does *not* allow sampling to a foregone conclusion; at least, certainly not in the way in which Mayo claims it does and, as far as I can tell, not at all.

4. OBJECTION 12.4
THE LIKELIHOOD PRINCIPLE IMPLIES A COUNTER-INTUITIVE STOPPING RULE PRINCIPLE

First, a rough definition to get our bearings. A *stopping rule* is an agreement by experimenters and statistical analysts to execute an experiment in parts, with each part being subjected to a pre-agreed type of statistical analysis as soon as possible after its completion, and with the series of sub-experiments guaranteed to terminate "early" (before some pre-agreed maximum sample size has been reached) if one of the analyses has some pre-agreed outcome. The sequence of sub-experiments which results from applying a stopping rule is called a *sequential* experiment. Typically the only outcome which is allowed to cause early termination of a sequential experiment is a pre-agreed rate of events (e.g. deaths) among the experimental subjects. The outcome required to trigger early termination of the experiment is typically, but not necessarily, worked out by requiring a pre-agreed level of significance against some pre-agreed null hypothesis. I will discuss the use of stopping rules in much more detail in chapter 15.

Birnbaum's version of the likelihood principle entails the following principle:

> In a sequential experiment E^{τ}, with observed final data $[x_a]$, $\mathrm{Ev}(E^{\tau}, [x_a])$ should not depend on the stopping rule τ.
>
> (Berger & Wolpert 1988, p. 76)

Revising this to avoid the undefined term "$\mathrm{Ev}(E^{\tau}, [x_a])$", we get the following version of the SRP, which follows from my version of the likelihood principle:

> The **stopping rule principle (SRP)** (Grossman version): Inferences about hypotheses made on the basis of experimental data should not depend on the stopping rule which was either planned or actually used in the experiment in which the data were collected, provided that the conditions of applicability of the likelihood principle are satisfied.

Mayo claims that the stopping rule principle is false, and hence that the likelihood principle is false.

Mayo's main argument against the stopping rule principle is that it allows sampling to a foregone conclusion. I have already defended the likelihood principle against the allegation that it allows sampling to a foregone conclusion. An exactly parallel argument defends the stopping rule principle against the same allegation.

OBJECTION 12.4.1
THE LIKELIHOOD PRINCIPLE IMPLIES A FALSE STOPPING RULE PRINCIPLE

There are other versions of the stopping rule principle, and the various versions are easily confused. Consider, in particular, the following:

> The **universal stopping rule principle**: in any experiment, the stopping rule is always irrelevant to inferences from the experimental data to conclusions about hypotheses.

A possible objection to the likelihood principle is that it entails the universal stopping rule principle. I acknowledge that the universal stopping rule principle is counter-intuitive; in fact, it is false, as I will show in a moment. So, to defend the likelihood principle I must show that it does not entail the universal stopping rule principle. Mayo, the most prominent opponent

of the SRP, acknowledges that the likelihood principle does not entail the universal stopping rule principle (Mayo 1996, p. 342, footnote). In this section I will show that she is right: the universal SRP is false.

Howson and Urbach give a nice illustration of the falsity of the universal SRP. They note that an experiment's stopping rule would be relevant to one's conclusions about hypotheses

> . . . if one were relying upon a random sample to measure the mean height of a group of cooks who happened to be preparing lunch at the same time as the experiment was in progress [and if we also knew] that tall chefs cook faster than short ones and that the trial was concluded as soon as lunch was ready. . . . Ignoring the stopping rule in such a case would be overlooking relevant information. . . .
>
> This concession should not be misunderstood. It does not mean that the scientist's *intention* to stop the trial at a particular point is of any inductive significance; hence, our position is quite different from that of the classical [Frequentist] statistician. We are simply claiming that in estimating a parameter, one normally would derive all one's information from the composition of a suitable sample, but that sometimes events attending the sampling process also have significance as evidence.
>
> (Howson & Urbach 1993, p. 366)

It is fortunate for the likelihood principle, in the face of this decisive criticism of the universal SRP, that it (the likelihood principle) does not entail the universal SRP. This can be seen by noting that the stopping rule in this example is naturally seen as part of x_a and hence may be used in a likelihood analysis. If, perversely, Howson and Urbach's stopping rule is not made part of x_a then the SRP might seem to apply, but in fact it does not (at least, my version does not), because it is a precondition of the application

of my version of the likelihood principle, and hence of my version of the SRP, that x_a represents "all observations considered relevant to any of the hypotheses" (chapter 8). A last gasp objection might be that Howson and Urbach's stopping rule might not be *considered* relevant to the hypotheses, even though it *is* relevant. In answer to this I can only say that when a mistake of this kind is made it is rational (although unfortunate) to accept an unsatisfactory analysis until the mistake is discovered and corrected, at which point the analysis can be amended. Consequently, the fact that an analysis which leaves out part of the observation can give bad results is no criticism of the likelihood principle.

A stopping rule which is naturally seen as part of x_a is known as an *informative* stopping rule. As we have just seen, a stopping rule can be informative even if the agent doing the analysis doesn't know that it's informative (or falsely believes that it isn't). This holds the solution to a puzzle about *deliberate* misleading of the analyst by an experimenter. For example, a pollster employed by the Evil Bayesian Party might start his poll in the suburbs most likely to vote for his party, and might stop when the proportion of support for his party went above say 90%. A statistician analysing the results but unaware of the order in which the experimenter had sampled would, rightly from her point of view, ignore the stopping rule. The stopping rule is informative (because of the ordering of the data; otherwise it would not be), but the analyst does not know this. She is performing an unsound analysis, but only because of ignorance. It is hard not to be worried about the impact of accepting the stopping rule principle on this sort of example — it even worries me — but the situation is no different in principle from any other withholding of information to mislead

an epistemic agent. *Of course* an evil experimenter can mislead a statistical analyst. He can always do so, by withholding information or by lying. The stopping rule principle, because it is a powerful analytic tool, gives him one more way to do so; but that should not count against the stopping rule principle. At most, it means that the stopping rule should be put to one side if the analyst believes the experimenter to be evil, just as various other rules of inference from testimony need to be suspended in such a case. Alternatively, a Bayesian analyst can put a prior probability distribution on the behaviour of the evil experimenter, and then the analysis becomes unproblematic.

OBJECTION 12.4.2
THE STOPPING RULE PRINCIPLE IS FALSE EVEN WHEN NO FREQUENTIST METHODS ARE USED

This objection is certain to be a remaining niggle in the minds of readers: that Mayo's argument above is irrelevant, because the likelihood principle and the stopping rule principle do not apply when significance tests (Frequentist methods) are used; but that it does not follow that everything is OK when Frequentist methods are not used. Perhaps sampling to a foregone conclusion is possible anyway.

It is impossible to certify that this cannot happen in complete generality, since the likelihood principle does not specify exactly how a statistical analysis is to be done, and there is no limit to the (rational and irrational) ways in which it can be used. But it *is* possible to certify that sampling to a foregone conclusion cannot happen when the likelihood principle is used as part of a Bayesian analysis (subject to the constraints of chapter 2), and similar arguments are in principle available for other reasonable ways of

using the likelihood principle. The Bayesian issue has been dealt with a number of times in the literature. A particularly succinct version is given in the forum from which Mayo quotes:

> Dr P. ARMITAGE: I should like Professor Savage to clarify a point he made in Part I. He remarked that, using conventional significance tests, if you go on long enough you can be sure of achieving any level of significance; does not the same sort of result happen with Bayesian methods? The departure of the mean by two standard errors corresponds to the ordinary five per cent level. It also corresponds to the null hypothesis being at the five per cent point of the posterior distribution. Does it not follow that by going on sufficiently long one can be sure of getting the null value arbitrarily far into the tail of the posterior distribution?
>
> SAVAGE: The answer is surely no, under any interpretation. It is impossible to be sure of sampling until the data justifies an unjustifiable conclusion, just as surely as it is impossible to build a perpetual-motion machine. After all, whatever we may disagree about, we are surely agreed that Bayes's Theorem is true when it applies. But to understand this impossibility let us examine first a simple case.
>
> Consider an urn that contains three red balls and a black one or three black balls and a red one. To convince you of the first hypothesis as opposed to the second, for some given purpose, would mean to make the likelihood ratio in favour of the first sufficiently large, say at least 10. Suppose that I, in my zeal, decide to keep sampling (with replacement) until the likelihood ratio, which in this particular case is $3^{(r-b)}$, exceeds 10. This will happen if and only if I sometimes succeed in drawing three more red balls than black ones; if there are really three black balls and a red one, it is quite probable that I never will succeed until the end of time. In fact, the probability of failure in this

> unfavourable circumstance is at least 9 / 10, as it ought to be on general principles; the exact value is 26 / 27.
>
> As I understand it, Dr Armitage is particularly interested in the following sort of example. The prior distribution of a parameter μ is rather broadly distributed around 0, and observations of μ with unit standard deviation are sequentially available. From 'your' point of view, that is, the point of view summarized by the assumed prior distribution, what is the probability P that I should succeed in sampling until your posterior odds that μ is positive are at least 10 times your initial odds that μ is positive, if μ is in fact negative? There can be no escape from the simple formula that P is at most a tenth.
>
> (Savage & discussants 1962, pp. 72–75)

Since this is an important example, I will expand on it a little. Suppose there are N balls in the urn: either $N-1$ reds and one black or $N-1$ blacks and one red. We will look at one ball at a time, with replacement, to decide which of those hypotheses to believe. I will allow you, the experimenter, to have whatever stopping rule you like, including "continue sampling until I have falsely proved that the balls are mostly black". I am willing to make a bet at even odds that there are more red than black balls, on the condition that you use a straightforward likelihood method to evaluate the evidence, namely: I lose the bet if p(observed balls conditional on there being mostly blacks in the urn) / p(observed balls conditional on there being mostly reds in the urn) exceeds some ratio, say 10. (If I were instead to agree that the Frequentist rule $p < 0.05$ was an adequate test of a hypothesis, you would be able to deceive me with probability 1.)

If there are N balls of which 1 is red, and if the alternative hypothesis is that all except 1 are red, then the likelihood ratio p(mostly blacks)/p(mostly

reds) is $(N-1)^{(b-r)}$. So I will lose the bet if at *any* time before you choose to stop the experiment there have been $log_{10}(N-1)$ more blacks than reds. And you are completely in charge of the stopping rule. Still I should expect to win the bet, as a little algebra will show. This is a nice illustration of the fact that we may intuitively expect to be able to sample to a foregone conclusion even when in fact we cannot.

Mayo (1996, p. 352) quotes the above question from Armitage about sampling to a foregone conclusion but not the response from Savage, and adds:

> Although Savage wants to deny Armitage's implication, he appears to grant it, though fuzzily, and moves on to another example
>
> (Mayo 1996, p. 353)

and

> Savage [is] plainly uncomfortable with Armitage's result
>
> (Mayo 1996, p. 356)

Apparently "The answer is surely no, under any interpretation" is a way of granting a proposition and shows discomfort. (Nowhere does Savage qualify his "no" to a "yes" or even a "perhaps".) In the face of a claim that an author says almost the exact opposite of what he actually says I am at a loss for words.

In addition to misrepresenting her opposition on this point, Mayo claims to give an example in which a Bayesian would sample to a foregone conclusion. Her example involves the assumption that "In certain cases, rejecting a null hypothesis H_0, say at level of significance .05, corresponds

to a result that would lead a Bayesian to assign a low (e.g., .05) posterior probability to H_0" (Mayo 1996, p. 352). It follows that, since the Frequentist would sample to a foregone conclusion if he ignored the stopping rule, the Bayesian (who does ignore the stopping rule, in line with the likelihood principle) will also sample to a foregone conclusion. The stated assumption is prima facie plausible, but in fact it is not true, unless the Bayesian uses prior probabilities which do not form a probability distribution (an improper prior). Mayo claims that "the kind of prior that leads to the trouble [is] a commonly acceptable one" (Mayo 1996, p. 356). It is true that *some* Bayesians use such priors, while other Bayesians have been attacking them for doing so since the 1920s (e.g. Hill in (Berger & Wolpert 1988, p. 162)). In any case, the probability calculus forbids improper priors, so they are ruled out by the framework I set out in chapter 2.[113]

In an example like Armitage's, if we sample until the posterior odds that μ is positive are at least k times the initial odds that μ is positive, the probability of sampling to the "foregone" conclusion that μ is positive

113. At the worst, Mayo has convicted those Bayesians who *both* use improper priors *and* take the stopping rule principle seriously of inconsistency; but it has been known since the 1970s that Bayesians who use improper priors can be shown to be inconsistent with or without the stopping rule principle (Stone 1976). If Bayesian methods based on improper priors are inconsistent with or without the stopping rule principle then Mayo's proof that they are inconsistent with it proves nothing important about the stopping rule principle and hence nothing about the likelihood principle. It does perhaps prove that Bayesians ought to be even more careful about using improper priors than some of them realise, but that point does not reflect badly on the stopping rule principle.

Incidentally, although sampling to a foregone conclusion is possible with an improper prior it is still, usually, not feasible: a typical realistic Bayesian analysis, conducted in a reasonable amount of time, cannot reach a foregone conclusion even if it uses an improper prior (Berger & Wolpert 1988, pp. 81–82). Hill argues that this result generalises to a proof that it is impossible for an improper prior to result in a betting loss (Berger & Wolpert 1988, p. 167–171), contra Mayo's claim. An intuitive understanding of why this is so must consider that Mayo's claim about improper priors is true only if an infinite number of analyses is possible: any less than infinite time changes the outcome substantially. This is why some Bayesians still feel free to use improper priors. A related point is that even a prior which differs only slightly from Mayo's improper prior can make sampling to a foregone conclusion impossible.

when in fact it is negative is at most $\frac{1}{k}$; and hence it is not really foregone at all.

One final remaining worry might be that we ought to be able to make the probability of making this mistake even lower, perhaps by breaking the likelihood principle. The answer to this is that, yes, it can be made lower. By breaking the likelihood principle and taking the stopping rule into account in a Frequentist manner we can make this particular false conclusion as rare as we like. We can do this, for example, by refusing to document a given proportion of red balls, or by using a non-ignorance prior, or by introducing a utility function. But if we do these things we make sampling to the *opposite* false conclusion (that μ is negative when in fact it is positive) more frequent. I conclude that there is nothing particularly unsatisfactory about the repeated-sampling properties of the Bayesian use of the likelihood principle.

This completes my responses to all the objections to the likelihood principle of which I am aware. In the next chapter, I give a proof of the likelihood principle, followed in chapter 14 by responses to objections to the proof.

Part III

Proof And Pudding

— 13 —
A Proof of the Likelihood Principle

1. INTRODUCTION

I hope that Part II of this thesis has made the likelihood principle plausible. Part III completes my story in two ways: first by giving proofs of the likelihood principle, and then by offering a serving suggestion showing how it can best be eaten, by discussing a case study of its application.

In this chapter I present a proof of the likelihood principle, for discrete statistical distributions. Mathematically speaking, my proof is only slightly different from a number of previous proofs of the likelihood principle, all of which follow the general strategy of (Birnbaum 1962); I have borrowed especially from (Berger & Wolpert 1988, pp. 27-28). My proof differs from its predecessors mostly in the careful wording of its premises, which for the first time incorporate all the necessary conditions of applicability. In the following chapter I will present and refute objections to my proof.

The main premise from which the proof proceeds is essentially the uncontentious conclusion which I drew from Cox's example in chapter 7. Recall Cox's example: if we send blood to one of two non-equivalent laboratories, basing the decision as to which on the toss of a coin, it is reasonable to take into account which laboratory it actually went to when making inferences from the results the laboratory sends back, even though that makes it impossible to fix the overall error rate for the experiment

at any predetermined level (counting both the coin toss and the laboratory results as part of the same experiment). The conclusion I drew from Cox's example was that inferences about the blood must take into account properties of the laboratory that was actually used, and must ignore properties of the one that wasn't. In other words, one should condition on the coin toss. In this particular case at least one should treat the coin toss and the laboratory measurement as two separate merriments, regardless of whether they were planned together. I have never come across anyone who disagrees with this conclusion. There are many authors (such as Mayo) who believe that one should not *always* condition on observed data, but there are none who believe that one should not condition on the coin toss in Cox's example.

I promised earlier that we would be able to generalise this conclusion, using very mild assumptions indeed, to a fully-fledged principle which makes precise the idea that we should analyse Table 1 by columns instead of by rows. This principle is, of course, the likelihood principle. One of the most stunning results in twentieth-century applied mathematics, due to Birnbaum (possibly based on a sketch of a proof in (Pratt 1961, p. 166), and also independently discovered by Barnard in 1962), is that agreement about what to do with the results of an experiment chosen by a coin toss is very nearly enough to support a proof of the likelihood principle. No assumptions about experiments which are chosen in another way — perhaps a deterministic way, or even a deliberate way — are required, even though the likelihood principle can be applied to such experiments once it has been proved. Once the Cox example is formalised as the weak conditionality principle (below), all that needs to be added is a small set of

conditions limiting the domain of applicability in line with the framework of chapter 2 and a weak sufficiency principle which is fairly uncontentious (although I consider some objections to it in chapter 14).

The idea of *proving* a normative principle may seem strange. What makes it possible is that the conditioning premise drawn from the Cox example is normative (although very, very weak): it says that we *must* take into account the properties of one laboratory and *must* ignore the properties of the other. The weak sufficiency principle is also normative. The chain of reasoning from these premises to the likelihood principle is purely mathematical but slightly difficult, and therefore takes the form of a deductive proof.

By proving the likelihood principle from a premise about conditioning which mentions only a single coin toss, I will show in this chapter that authors who are squeamish about conditioning on observed data in some cases must bite the bullet and disown conditioning in all cases, even in Cox's case; because if they give me Cox's case then they give me the likelihood principle, and that in turn entails that conditioning is *always* necessary (when we are doing inference within the framework of chapter 2).

2. PREMISES

FORMAL DEFINITION OF A LIKELIHOOD FUNCTION

A likelihood $L(h)$ is a function $p(x_a|h)$, where p is a probability function or a probability density function, h ranges over a set of hypotheses H, and x_a is some observed data considered as a constant.

Recall that two likelihood functions are the same if and only if they're proportional to each other. In other words:

$L_1(h) = L_2(h)$ iff $L_1(h) \propto L_2(h)$

— i.e., iff $(\exists c > 0)\ (\forall h)\ L_1(h) = cL_2(h)$.[114]

THE WELL DEFINED LIKELIHOOD FUNCTION CONDITION

The concept of a statistical measurement is only useful under a condition which is not always made explicit but which, if made explicit at this stage, will save a lot of trouble later. It is what I call the **Well Defined Likelihood Function** condition (**WDLF**):

> For each hypothesis h under consideration in a statistical analysis, $p_h(x_a) \equiv p(x_a|h)$ must be well defined (i.e., have a single value).

The WDLF is an explicit condition of applicability of my version of the likelihood principle. Although I state it explicitly here for maximum clarity, it really follows from the framework which I set up way back in chapter 2.

---

114. If the two functions L_1 and L_2 both have finite integrals then this condition is the same as saying that they are the same likelihood function iff they reduce to the same function when normalised. To normalise a function is to ensure that it integrates to 1, without changing its shape. This is easily done by replacing the function f with f / $\int$(f).

But likelihood functions need not have finite integrals; and functions without finite integrals cannot be normalised in this way.

There I stated that the hypothesis space H must be fixed for the duration of the analysis.[115] Since the likelihood function, $p(x_a|h)$ with x_a fixed and h variable, obviously supervenes on H, the assumption that H is fixed for the purposes of the analysis of a statistical measurement implies that the likelihood function is also fixed.

In situations in which merriments are described formally, the WDLF or something very like it is often made explicit, especially when disagreements about the likelihood function would have legal ramifications, as in clinical trials. In any case it is a sensible assumption. Even in an informal situation in which the likelihood function is not explicitly agreed by all parties to an analysis, my discussion is still relevant: it simply applies to the statistical model which is being used after all personal disagreements have been ironed out.

The following important sub-premises are implicit in the WDLF.

Sub-premise A: H takes into consideration all the hypotheses we're going to consider, no matter what the data turn out to be.

In most scientific cases this is very easy to ensure. It may be a problem for the representation of our actual psychological processes by statistical models, but in this thesis I am only worrying about normative considerations.

Sub-premise B: All factors that are considered epistemically relevant to the problem are included in the model.

---

115. H need not be held fixed in any temporal sense; it merely needs to be held fixed across possible observations for the purposes of any single analysis of the data. Specifically, the model need not be fixed in *advance* of the data collection. Also, as far as the likelihood principle is concerned those parts of the model representing possible observations which did not occur are irrelevant and therefore need not be held fixed, nor even exist, at any time.

Usually all the factors of interest to the agents who are making the statistical model are represented as components of hypotheses of interest, which reduces Sub-premise B to Sub-premise A; but some may be unknown factors which may not be of interest on their own account but may still be epistemically relevant. Such factors are called "nuisance parameters".

When Assumption B is broken, things can become very confusing. The most interesting example for us is probably the case of stopping rules, as discussed in chapter 11. We saw there that stopping rules are relevant to an analysis if and only if they are proxies for important information (e.g., sample size) that was not mentioned in the stated model. It is hard to think of reasons why this situation should be allowed to arise: it is hard to think of reasons why any acknowledged important aspects of the epistemic situation should not be included in the model. (It is hard to think of a reason why sample size should ever be ignored, for example.) Note in this context that we can add any type of data we like to the model: we are not restricted to parameters of a distribution.

Sub-premise C: For the purposes of *this* statistical measurement we have decided not to change H in an ad hoc way as a result of seeing the data. (By ad hoc I mean without a principled reason.)

The literature tends not to worry about this issue. Dawid's version of the likelihood principle, for example, allows for ad hoc inference procedures but not for *unpredictable* ad hoc inference procedures: "an [inference procedure] may be entirely ad hoc, so long as it specifies the particular inference to be made in every relevant situation." (Dawid 1977, p. 247)

Sub-premise D: We have decided not to change H in a *non*-ad hoc way.

We could combine Sub-premises C and D, of course. But they are clearer if kept separate.

The obvious argument in favour of Sub-premise D is that if we contradicted it by deciding to change $p_h(d)$ to $q = f(p_h, d)$ on seeing data d, apparently contradicting Sub-premise D, we could and should replace p_h by q in the initial description of the statistical measurement, which would result in a well-behaved probability distribution (modulo a normalisation) for which Sub-premise D would hold. In that case we might as well accept Sub-premise D after all. Any non-ad hoc decision to change H should be forseeable, so Sub-premise D is reasonable in every case. The only weak point of this argument is that it assumes a rational doxastic agent. While there is something to be said for assuming that a *single* doxastic agent ought to be rational, it may be impossible — for logistical reasons, quite apart from the theoretical problems inherent in collective decision-making (problems which are well discussed in (Kadane et al. 1999)) — for a group of agents to be rational to the extent required by Sub-premise D. Consequently, I cannot pretend to deal fully with the problem of multiple doxastic agents.

A possible argument against Sub-premise D is that statisticians, even lone statisticians, do not always behave in accordance with it: they sometimes accept rules which require them to change H after seeing data. Two examples will illustrate this.

Firstly, Bayesians of many schools are willing to make small changes to their priors if the data suggest that the likelihood function ought to have a different functional form — for example, after seeing the outcome of a merriment, a statistician might change the functional form of the likelihood

fuction from a Normal (Gaussian) distribution to a log-Normal distribution of a very similar shape and size. This can be shown to have very little effect on the conclusions the statistician will draw in a very wide range of cases, but obviously there are cases in which changing the functional form of the likelihood function will make a difference to some conclusion. We cannot doubt that Sub-premise D is broken by such scientists.

Secondly, some authors (Basu 1975, p. 19; Gelman et al. 1995) advocate changing the likelihood function in a more substantial way if the data turn out in certain ways. The most interesting issue is whether such changes are merely matters of convenience, in which case agreement that one could validly apply the transformation $f(p_h, d)$ would make Sub-premise D still valid in principle.

Despite the plausibility of these objections in certain circumstances, two points need to be made in favour of Sub-premise D. The first is that the objections of Gelman et al. only apply in unusual circumstances. In typical scientific uses of the likelihood principle, such as fixed-size biomedical experiments, the WDLF (and therefore Sub-premise D) is uncontested. The second point to be made is how extremely *little* Sub-premise D claims, even in contentious circumstances. The availability of the transformation $f(p_h, d)$ means that changing the likelihood function after seeing the data is fine, provided that the participants in the analysis *either* agree on the change and hence agree to re-analyse the data *or* agree to report their results separately (agree to disagree); and this latter situation is no different in principle from what would have happened if the participants had not been able to agree on a statistical model in the first place.[116]

116. Incidentally, it is rare for scientists to fail to agree on a statistical model, perhaps

Then the simplest definition of sufficiency is as follows:

> **sufficiency definition 1**
> $T(x)$ is a sufficient statistic for h iff $p(x|T(x))$ is algebraically (functionally) independent of h.

A sufficient statistic for h, $T(x)$, typically contains much less information about the world than X does, but the same amount of information (in a sense which I will make precise in the rest of this section) about h.[117]

The reason for the name "sufficient" is that if $T(X)$ is sufficient for h (in the technical sense above) then it is all we need to know about X, if our sole purpose is to infer things about h, and so it is sufficient information in the lay sense. Anything else we know about X, over and above $T(X)$, is epistemically redundant. For example, if we're sure that all we care about is the average height of a population (a big if), there is no point in recording more than the average height of the test sample; any other information about the test sample can be thrown away.

There is a problem with the above definition of sufficiency: it can only be applied by people who are willing to talk about $p(h)$. According to Neyman and many others, including some proponents of the likelihood principle, such as Hacking (1965) and Edwards (1972), $p(h)$ is meaningless in many circumstances. Happily, there is an alternative definition of sufficiency which agrees with the first version whenever $p(h)$ exists but which

---

because their reputations as productive members of their community depend on being able to conclude statistical analyses quickly and without fuss.

117. For example, if x is a vector of the heights of a sample of people then, under the Normal or log-Normal models most often used for human heights, the average (mean) height of the sample, $T(x) = \sum_{x_i=1}^{n} x_i / n$, is a sufficient statistic for the average (mean) height of the population.

does not require $p(h)$ to exist, and which has the same epistemological properties as the first version. This more widely applicable definition is:

sufficiency definition 2
$T(X)$ is a sufficient statistic for h iff we can find a function T' which allows p to be factorised in the following way:

$$(\forall h \in H) \quad p(x|h) = T'(T(x), h) \times p(x|T(x)).$$

A theorem known as the factorization theorem shows that this definition is equivalent to the earlier definition.[118]

Seeing that statistical sufficiency implies epistemic sufficiency is even easier using definition 2 than it was using definition 1. For definition 2 shows that all functions of $p(x|h)$ can be calculated from $T(x)$ and h, when T is sufficient for h. This point may look superficially as though it assumes the likelihood principle, but it does not. That all inferences about h depend on $p(x_a|h)$ is, more or less, the likelihood principle; but that all such inferences depend on $p(x_i|h)$ for some set $\{x_i\}$, is an unrelated, trivial claim.

For example, in most experiments on coin tossing, the number of heads and the number of tails are jointly sufficient for all inferences; the order in which we observe the heads and tails is irrelevant. ($x_1, x_2, \ldots$ are jointly sufficient iff the ordered tuple $\langle x_1, x_2, \ldots \rangle$ is sufficient.) The only assumption we need to make in order to be sure that we have a

118. One part of the factorization theorem is easy to prove. It is easy to see that if this equation holds then $T(X)$ is sufficient for h on definition 1, thus: if we know $T(x)$ then we know the right-hand side as a function of h (bearing in mind that we can calculate $p(x|T(x))$, because it does not depend on h); hence, we know the left-hand side, which establishes that $T(X)$ is sufficient for h. The converse is more long-winded to prove, and I will not be relying on it (since my working definition of sufficiency will be the second version, and all I need show is that whatever fits my definition also fits the other one, not vice versa), so I will omit the proof.

non-trivial sufficient statistic in the coin-tossing case is the assumption of exchangeability explained in chapter 2. As we have seen, this requirement is trivially satisfied if the results take the form of a multiset.

All statistical models have sufficient statistics, trivially, since h itself is a sufficient statistic for h according to the above definition. Of course, such trivial sufficient statistics are not very useful. In addition, a model may have *many* sufficient statistics.

PREMISE: THE WEAK SUFFICIENCY PRINCIPLE (WSP)

> **The weak sufficiency principle**: If $T(X)$ is a sufficient statistic for h, and if $T(x_1) = T(x_2)$, then inference procedures should not derive different inferences about h from x_1 and x_2.
>
> (adapted from Basu 1975, p. 9)

The weak sufficiency principle was named thus by Dawid (1977) because it is weaker (claims less) than other similar principles. I will not be considering rival principles, but I have retained the name, partly for consistency with the literature but mostly because it is useful to be reminded how modest the principle's claims are.

No statistician knowingly breaks the WSP. If a conflict with the WSP ever arises, the only reasonable conclusion is that $T(X)$ is not a sufficient statistic for h after all. I would like to give four arguments for this. I do not claim that the four arguments are independent of each other; only that one might convince where the others fail.

Firstly, the WSP follows directly from the claim (defended above) that statistical sufficiency entails epistemic sufficiency.

A second argument for the WSP, adapted from (Basu 1975, p. 9), is as follows. Let us imagine that we have observed x_a in a two-step procedure: we have first conducted an experiment with sample space X but noted only the value of $T(x)$, not the precise value of x. Then we conduct a further, separate experiment with sample space $T(x)$, noting this time the exact value x_a obtained. Since T is sufficient for h, the second experiment is "statistically trivial" (Basu's term) and tells us nothing about h. Hence, the outcome of the second experiment can make no difference to our inferences about h. Hence values x_1 and x_2 which are possible outcomes of the second experiment (i.e., such that $T(x_1) = T(x_2)$) should lead to the same inferences about h.

Thirdly, here is a Bayesian argument for the WSP. We can prove that if $T(X)$ is sufficient for h, as defined above, then $(\forall x)\, p(h|T(x)) = p(h|x)$. So knowing $T(x)$ allows us to know the entire function $p(h|x)$ (as a function of h).[119]

---

119. This Bayesian argument will only be helpful for those who believe that $p(h|x)$ is meaningful — some do not — but since the proof is simple it is worth presenting.

Proof: $(\forall a, b, c)\quad p(a|b) = p(a|c)p(c|b) + p(a|\bar{c})p(\bar{c}|b)$, either from the definition of conditional probability or (better, since I take conditional probability as primitive) from the probability axioms of chapter 2.

So $p(h|x) \equiv p(h|X = x)$

$= p(h|T(X) = T(x)).\, p(T(X) = T(x)|X = x) + p(h|T(X) \neq T(x)).\, p(T(X) \neq T(x)|X = x)$

$= p(h|T(X) = T(x)) \times 1 + p(h|T(X) \neq T(x)) \times 0$

$= p(h|T(x))$.

Interestingly, the above proof assumes a classical logic; otherwise, the fact that knowing parts of x above and beyond $T(x)$ cannot affect the fact that we know the whole of $p(h|x)$ might not imply that it cannot affect our conclusions about h in any way. In a paraconsistent logic (one in which some but not all contradictions are true (Priest 1987)), knowing $T(x)$ might enable one to know $p(h|x)$, but finding out more about x might enable one to find out that some of the truths previously discovered were false (as well as true). This would invalidate the weak sufficiency principle, and thus the likelihood principle (since the converse of the likelihood principle can be proved from the converse of the weak sufficiency principle given the weak conditionality principle, as I prove below when I show that the likelihood

PREMISE: THE WEAK CONDITIONALITY PRINCIPLE (WCP)

Informal statement:

> If one of two possible statistical measurements is chosen by the toss of a fair and indeterministic coin, no inference procedure should require information about the merriment that was *not* performed.

Formal statement:

> **The weak conditionality principle**: Consider two statistical measurements $M_1 = (X_1, H, p_1)$ and $M_2 = (X_2, H, p_2)$. (By this I mean that M_1 has sample space X_1, hypothesis space H and probability function p_1, and similarly for M_2.)
>
> Note that the set of hypotheses is the same for each. This is a deliberate restriction which entails that this principle does not apply to hypotheses about alchemy compared with hypotheses about chemistry, although it does apply to comparing statistical models that each consider both alchemy and chemistry. This is in accordance with the WDLF.
>
> Now consider an observation from a new merriment, M^*, which consists of using a fair, epistemically indeterministic coin to select one of M_1 and M_2 at random with probability 1/2 each. M^* still falls within our definition of a statistical measurement: formally, $M^* = (\ (\mathrm{J}, X_J), \mathrm{H}, p_j\ (\mathrm{J}, X_J)\ )$. (By "epistemically indeterministic" I mean simply that no deterministic pattern in the behaviour of the coin has been noted or is expected. Some people are known to be able to toss a coin so as to yield a predetermined outcome. Our coin-tosser must not be one of those people.)

---

principle is logically equivalent to the union of the two weaker principles, and since the weak conditionality principle is not thrown into doubt by paraconsistency). As far as I know this is an original point, and perhaps worth following up . . . but not here.

> Suppose M^* is performed, and turns out to consist of M_1. Then any inference procedure should derive the same inference from this instance of M^* as it would have derived from M_1 alone.

The weak conditionality principle is called "weak" for the same reasons as the weak sufficiency principle: for consistency with the literature and to remind us how modest it is. Stronger versions of the conditionality principle replace the coin with an arbitrary ancillary statistic, thus:

> The *conditionality principle* is given as follows:
> **C**: cont(I_1) = cont(I_2) if I_2 is the conditional inference base given the value of an ancillary for $I_1[.]$
>
> (Evans et al. 1986, p. 185)

where "cont(I)" refers to "what the model and data in I . . . say concerning the unknown θ" (Evans et al. 1986, p. 184). I will not dwell on this more general principle, because I do not need it. I only need the Weak Conditionality Principle, which is so similar to Cox's example given above that it is barely even a generalisation of it. The coin is still a coin. All that has changed is that the two laboratory measurements have been replaced by two arbitrary measurements relevant to H.

ALTERNATIVE PREMISES

The likelihood principle is not just entailed by the the WSP and the WCP, it is actually logically equivalent to their conjunction. So it is impossible to weaken or remove either principle without strengthening the other one, unless a new principle is added.

The WSP can be replaced, in the proof of the likelihood principle, by a principle saying that if an ancillary statistic exists then it is acceptable

to condition on it, provided that we add an additional axiom saying that certain types of structural information can be ignored (types of structural information which only turn up in the structural theory of Fraser and the pivotal theory of Barnard). A proof using these alternative axioms is given in (Berger 1985, pp. 37–39).

Evans, Fraser and Monette (1986) have a proof that depends only on a version of the conditionality principle, but it is a stronger version than the one given here, and it makes additional assumptions about ancillarity. It seems to most authors, and to me, that the weak sufficiency principle is already so uncontentious that it is better to leave it in the proof, in return for being able to use such a weak conditionality principle.

Birnbaum pulls a similar trick: he proves the likelihood principle from a conditionality principle slightly stronger than mine plus the following "principle of mathematical equivalence":

> *Mathematical equivalence* (M): If $f(x, \theta) = f(x', \theta)$ for all $\theta \in \Omega$, then $\mathrm{Ev}(E, x) = \mathrm{Ev}(E, x')$.
>
> (Birnbaum 1972, p. 858)

Dawid (1977, p. 249) gives a proof of the principle of mathematical equivalence from the transformation principle which we met in chapter 7 plus the assumption that the description of a statistical observation such as the one given in chapter 2 in terms of X, x_a and H is complete: this second assumption rules out structural inference (chapter 4) and pivotal inference (chapter 5).

According to Pratt (1962, pp. 314–315) and Birnbaum (1972, p. 861) it is also possible to eliminate the conditionality principle by replacing it

with a Weak Relabelling Principle implied by Pratt's voltmeter example which I discussed in chapter 7. Pratt's relabelling principle is:

> A relabelling of possible outcomes which does not affect the outcome actually observed surely should not change an inference or decision.
>
> (Pratt 1961, p. 166)

Pratt sketches a proof of the likelihood principle from this relabelling principle:

> However, there are almost always such relabellings which change the P-value and hence may change an inference or decision based on a significance test. Suppose, for instance, an experiment has possible outcomes $a, b, \cdots, z$. Suppose Meter 1 tells the outcome, while Meter 2 tells only whether the outcome was or was not d. If in fact the outcome is d, you would learn this from reading either meter and would want, therefore, to make the same inference or decision; yet the result of a significance test would ordinarily depend on which meter you were reading. A direct continuation of this argument shows an inference or decision should depend on the probability under the possible hypotheses of the outcome observed only (and on this only up to multiplication by a constant). The use of the probabilities of other outcomes also, as in the Neyman-Pearson formulation, inevitably leads to inconsistencies.
>
> (Pratt 1961, p. 166)

Why does Pratt say that a Frequentist analysis "would *ordinarily* depend on which meter you were reading"? Because using Meter 1 forces a Frequentist to use a detailed ordering of possible outcomes — either the obvious numerical ordering or some other fixed ordering, say O — in order

to calculate a P-value (for reasons explained in detail in chapter 7). Meter 2 prevents us from having such a detailed ordering: the only possible orderings based on Meter 2 are $< d$, not-$d >$ and $<$ not-d, $d >$. Hence — and this is where the caveat "ordinarily" comes in — we will get different P-values from Meter 1 and Meter 2 unless it happens that the detailed ordering and the coarse ordering coincidentally give the same result . . . which is unlikely unless d happens to be at the very top or the very bottom of the ordering used with Meter 1. This is not "ordinarily" the case, as we can see from the fact that the value of d is completely arbitrary: some particular values of d are at the top or bottom of the ordering O of $a, b, \cdots, z$, but most are not. Hence Pratt's example is extremely general. All that remains to turn it into a proof of the likelihood principle is a bit more precision (especially a plausible definition of "inconsistency" which covers this case — a role played by the WCP in my proof), plus an extension from the theory of P-values to Frequentist theories in general. Indeed, Pratt's example is sometimes cited as the first proof of the likelihood principle (Berger & Wolpert 1988), although this seems to me to be stretching the notion of proof a little.

Of course such a relabelling principle is almost identical to the Weak Conditionality Principle: the main difference between them is that in the Weak Relabelling Principle the choice (as it were) of a censored or a non-censored observation may be part of a single run of a single experiment, whereas in the Weak Conditionality Principle it is a separate coin toss. It seems to me that the latter is even more clearly irrelevant to what we ought to infer than the former, and consequently in the next section I will

present a proof of the likelihood principle from the Weak Conditionality Principle rather than from the Weak Relabelling Principle.

I call Pratt's relabelling principle the Weak Relabelling Principle (this is my own terminology, unlike "Weak Conditionality Principle") because Pratt himself objects to a stronger relabelling principle suggested by Birnbaum. Birnbaum's relabelling principle, also known as a principle of mathematical equivalence, is as follows:

> let (E, x) and (E', y) be any two instances of statistical evidence, with E and E' having possibly different mathematical structures but the same parameter space $\Omega = \{\theta\}$. Suppose that there exists a one-to-one transformation of the sample space of E onto the sample space of $E' : y = y(x), x = x(y)$, such that the probabilities of all corresponding (measurable) sets under all corresponding hypotheses are equal: $\text{Prob}(Y \in A'|\theta) = \text{Prob}(X \in A|\theta)$ if $A' = y(A)$. Then the models E and E' are *mathematically equivalent*, one being a relabelling of the other. If respective outcomes x of E and y of E' are related by $y = y(x)$, they also are mathematically equivalent, and the two instances of statistical evidence (E, x) and (E', y) may be said to have the same evidential meaning: $\text{Ev}(E, x) = \text{Ev}(E', y)$. A simple concrete example is that of models of measurements which differ only in the units in which measurements are expressed.
>
> (Birnbaum 1962, pp. 277–278)

This version of the relabelling principle is much stronger than Pratt's, because in Pratt's the relabelling cannot assign different values to the outcome which actually occurred (x_a), while in Birnbaum's it can. Pratt's objection to the stronger principle is as follows:

> I believe there is more to relabelling than meets the eye when the framework is left abstract [as it is in Birnbaum's relabelling

> principle]. It is not merely a matter of interchanging the labels attached to the states of nature. What is really involved is interchanging the distributions attached to the states of nature. An example, over-simplified to bring it into a two-state framework, would be this. If a certain drug has no effect, it helps the same proportion of patients as a placebo, which, let us say, is 25 per cent; if it has an effect, it helps 40 per cent. Relabelling does not mean that the two states are called 2 and 1 rather than 1 and 2 respectively. It seems to me relabelling gives a situation where no treatment effect means 40 per cent are helped and effect means 25 per cent are helped, instead of no effect meaning 25 per cent are helped and effect 40 per cent. This makes no physical sense to me, and accordingly I don't feel compelled to accept equal prior probabilities in Jeffreys' framework. . . . two samples giving the same likelihood on the same parameter space need not logically have the same evidential meaning unless the physical interpretations of the parameters are identical in the two cases.
>
> (Pratt 1962, pp. 315–316)

This amounts to the complaint that a relabelling may incorrectly over-ride prior knowledge of the sort which a Bayesian would incorporate using a probability distribution. (I say this because the fact that there is any lack of symmetry between the phrases "effect" and "no effect", which is the basis of Pratt's complaint, is a piece of prior, non-mathematical knowledge.) Pratt ought to observe that this same complaint applies to his Weak Relabelling Principle. However, had he noticed that, he could have replied that it has less force against his own principle than against Birnbaum's, because in his own principle we can at least be sure that we are not interchanging an *actual* state of nature with a non-actual one.

Be that as it may, this discussion is only indirectly relevant to the issue of alternative proofs of the likelihood principle: it has no direct impact on

the rest of this thesis, since I do not use either relabelling principle in my proof of the likelihood principle. Moreover, I make it part of my statement of the likelihood principle that Pratt's requirement that "the physical interpretations of the parameters are identical in the two cases" is met. I do this by insisting that two likelihood functions are considered equal only if all their variables have the same meanings.

We might wonder whether either relabelling principle is implied *by* the likelihood principle, in which case of course I am committed to it. The answer is that Pratt's weak principle is implied by the likelihood principle but Birnbaum's stronger principle is not, as we can see from the fact that orthodox Bayesian inference is compatible with the likelihood principle and yet can take into account prior knowledge such as the difference between "effect" and "no effect" (when such a difference exists).

3. A PROOF OF THE LIKELIHOOD PRINCIPLE FROM THE WSP AND THE WCP

PROOF OF THE LIKELIHOOD PRINCIPLE

Recall the likelihood principle:

Terminology

i By "inferences" I mean any beliefs and partial (probabilistic) beliefs which are held or followed and any actions which are taken, as deliberate results of an observation.

ii x_a denotes a vector representing all observations considered relevant to any of the hypotheses in some set H. x_a can be purely observational: it need not result from one or more deliberately constructed experiments.

iii By "inferences about hypotheses" I mean any inferences about the hypotheses in H: such inferences must not mention any hypotheses not contained in H except that they may (trivially) mention any hypotheses whose truth is not in doubt and any hypotheses on which x_a has no bearing.

iv Two likelihood functions are considered equal if all their variables have the same meanings within the theories represented by each hypothesis, and if the two functions are proportional (iff $(\exists c > 0)\ (\forall h)\ (L_1(h) = c.\, L_2(h))$.

Conditions of applicability

1. We cannot infer anything about the relative importance of the various possible inferential errors from the observation (i.e., the loss function, or equivalently the utility function, is either independent of the observation or unimportant).
2. The *choice* of observation is not informative about the hypotheses, only its outcome.
3. The Well Defined Likelihood Function condition: (perhaps trivially) for each hypothesis h under consideration in a statistical analysis, $p_h(x_a) \equiv p(x_a|h)$ must be well defined (i.e., have a single value).

The likelihood principle

Inferences from observations to hypotheses should not depend on the probabilities of observations which have not occurred, except for the trivial constraint that these probabilities place on the probability of the actual observation under the rule that the probabilities of exclusive events cannot add up to more than 1.

Consider two statistical measurements $M_1 = (X_1, H, p_1)$ and $M_2 = (X_2, H, p_2)$.

Next, consider the mixed merriment M^* (which was defined in statement of the weak conditionality principle as follows: a fair, epistemically indeterministic coin is tossed; according to its outcome, one of the merriments M_1 and M_2 is performed). Now suppose that whichever merriment hasn't been performed yet is also performed. At this stage we have an outcome x_1 from M_1, an outcome x_2 from M_2, an outcome j indicating which experiment was performed first ($j = 1$ for M_1 and $j = 2$ for M_2), and an outcome from M^*. The outcome from M^* is $J = 1$ or 2 and $x^* = x_1$ or x_2. The possible outcomes are denoted (j, x_j).

Then let $t_0 = (0, 0)$ and consider the statistic

$T(j, x_j) = t_0$ if $(j, x_j) = (1, x_1)$ or $(2, x_2)$

$= (j, x_j)$ otherwise.[120]

Is T a sufficient statistic for h? Generally, no. Recall that T is a sufficient statistic for h iff p can be factorised as

120. The *otherwise* clause can never represent an actual outcome, since I have defined the indexing of the merrimens in such a way that only $(1, x_1)$ can occur if M_1 is performed first and only $(2, x_2)$ can occur if M_2 is performed first. The fact that other outcomes cannot occur need not stop us considering their mathematical properties. According to the likelihood principle we should not base inference procedures on such properties; but we are currently proving the likelihood principle and so cannot assume it to be true. And even if we were able to assume it to be true at this point we could still *consider* such properties, even though we would have to refrain from endorsing inference procedures based on them.

$$(\forall h \in H) \qquad p(j, x_j|h) = T'(T(j, x_j), h)\,.\, p(j, x_j|T(j, x_j)).$$

There need not, in general, exist a suitable T' to match our choice of T. But suppose that the likelihoods of x_1 and x_2 are equal (i.e., $p_1(x_1|h) \propto p_2(x_2|h)$, or $(\exists k > 0)(\forall h \in H)(p_1(x_1|h) = k.(p_1(x_1|h))$. To prove the likelihood principle, we require to show that T is now sufficient for h. Let T' be as follows:

$$\begin{aligned} T'((j, x_j), h) &= \tfrac{1}{2}p_1(X = x_1|h) + \tfrac{1}{2}p_2(X = x_2|h), \qquad \text{if } (j, x_j) = t_0 \\ &= p(j, x_j|h) \text{ otherwise.} \end{aligned}$$

Then

$$\begin{aligned} T'(T(j, x_j), h) &= \tfrac{1}{2}p_1(X = x_1|h) + \tfrac{1}{2}p_2(X = x_2|h), \qquad \text{if } (j, x_j) = (1, x_1) \text{ or } (2, x_2) \\ &= p(j, x_j|h) \text{ otherwise.} \end{aligned}$$

To calculate $p(j, x_j|T(j, x_j))$ (the final term in the sufficiency equation), note:

$$\begin{aligned} p((1, x_1)|T = t_0, h) &= p^*(J = 1|T = t_0, h)\,.\, p_1(X_1 = x_1|T = t_0, h) \\ &= \tfrac{1}{2}p_1(X_1 = x_1|T = t_0, h) \\ &= \frac{\frac{1}{2}p_1(X=x_1|h)}{\frac{1}{2}p_1(X=x_1|h) + \frac{1}{2}p_2(X=x_2|h)} \end{aligned}$$

$$p((2, x_2)|T = t_0, h) = \frac{\frac{1}{2}p_2(X=x_2|h)}{\frac{1}{2}p_1(X=x_1|h) + \frac{1}{2}p_2(X=x_2|h)}, \text{ by symmetry}$$

$$p((j, x_j)|T = (j, x_j), h) = 1, \qquad (j, x) \neq t_0.$$

Now we can check the sufficiency equation:

If $J = 1, X_1 = x_1$ then

$$\begin{aligned} &T'(T(j, x_j), h)\,.\, p(j, x_j|T(j, x_j)) \\ &= \tfrac{1}{2}p_1(X = x_1|h) + \tfrac{1}{2}p_2(X = x_2|h) \times \frac{\frac{1}{2}p_1(X=x_1|h)}{\frac{1}{2}p_1(X=x_1|h) + \frac{1}{2}p_2(X=x_2|h)} \\ &= \tfrac{1}{2}p_1(X = x_1|h) \\ &= p(j, x_j|h). \end{aligned}$$

By symmetry, if $J = 2, X_2 = x_2$ then

$$T'(T(j, x_j), h) \,.\, p(j, x_j | T(j, x_j)) = p(j, x_j | h).$$

And for all other (J, X_J),

$$T'(T(j, x_j), h) \,.\, p(j, x_j | T(j, x_j)) = p(j, x_j | h) \times 1.$$

This establishes that T is sufficient for h.

It follows the sufficiency of T for h and from the weak sufficiency principle $\big($applied to the fact that $(T(1, x_1) = T(2, x_2)\big)$ that no inference about h is valid on observation $(1, x_1)$ in the mixed experiment unless it is also valid on $(2, x_2)$.

Now recall that j is chosen by a fair, indeterministic coin toss. Consequently, the weak conditionality principle applies. It tells us that no inference about h is valid on $(1, x_1)$ unless it is also valid on x_1 alone. (x_1 corresponds to M_1 in my formal statement of the weak conditionality principle above.) In other words, the observations $(1, x_1)$ and x_1 are equivalent in terms of the inferences they license. Similarly, the observations $(2, x_2)$ and x_2 are equivalent in terms of the inferences they license. And we determined in the previous paragraph that the observations $(1, x_1)$ and $(2, x_2)$ are equivalent in the same sense. Hence, the observations x_1 and x_2 license the same observations as each other.[121] Consequently, no inference is valid on x_1 (regardless of the value of j) unless it is also valid on x_2.

We have proved this for *any* x_1 and x_2 with equal likelihoods under the models under consideration ($p_1(x_1|h) \propto p_2(x_2|h)$). It follows that *any* two observations which share likelihood functions must share inferences about any unknown parameters mentioned by their statistical models, provided

121. To summarise this paragraph: if we write "$\equiv$" for "license the same observations as each other", we have just shown that $x_1 \equiv (1, x_1) \equiv (2, x_2) \equiv x_2$. The relation $\equiv$ is transitive, so $x_1 \equiv x_2$.

only that those unknown parameters index the same set of hypotheses for both.

The proof so far is sufficient to prove some versions of the likelihood principle, including Barnard's (1947) and Birnbaum's (1962) (see chapter 8). I am grateful to Daniel Steel for pointing out to me that the proof so far is not sufficient to prove all versions of the likelihood principle, because it leaves one important question ambiguous. Steel distinguishes (in personal communication) between two statements:

(1) If $p(x_1|h) = p(x_2|h)$ then x_1 and x_2 have the same evidential impact on h.

(2) If $p(x|h_1) = p(x|h_2)$ then x has the same evidential impact on h_1 as on h_2.

I have already proved (1), by proving that if $p_1(x_1|h) \propto p_2(x_2|h)$ then no inference about H (and hence about any h) is valid on x_1 unless it is also valid on x_2. But many versions of the likelihood principle, including mine, also imply (2). So it is necessary to extend the proof to handle this issue.

Consider any merriment $M = (X, H, p)$, label the outcome of M x (as usual), and based on M and x define a new merriment $M^{\odot} = (\Upsilon, H, p_\Upsilon)$ where Υ is 1 or 0 according to whether $X = x$ or not, thus:

$$\Upsilon = \begin{cases} 1 & \text{if } X = x \\ 0 & \text{if } X \neq x \end{cases} \tag{1}$$

so that:

$$p(\Upsilon = 1|h) = p(x|h) \quad \text{and} \quad p(\Upsilon = 0|h) = 1 - p(x|h). \tag{2}$$

So $(\forall h)\ p_{\mathcal{Y}}(\mathcal{Y} = 1|h) \propto p(x|h)$: the observations $\mathcal{Y} = 1$ and x share likelihood functions. Hence no inference is valid on the observation of x in M unless it is also valid on the observation of $\mathcal{Y} = 1$ in $M^{\odot}$. (All this is a trivial consequence of the part of the likelihood principle already proved above.) So we should ask which inferences are valid on the observation of $\mathcal{Y} = 1$ in $M^{\odot}$.

If all we know is that $\mathcal{Y} = 1$, whatever we can infer about H from X and x must be a function of the functions of X and x that appear in the description of $M^{\odot}$. But the only such functions are $p(x|h)$ and $1 - p(x|h)$ (from (2), or directly from (1) if you prefer). But these are just the likelihood function of x and 1 minus the likelihood function of x. And, in particular, no mention of any part of X except x is made in the description of $M^{\odot}$. So all inferences from $M^{\odot}$ and hence from M must depend functionally on x only via the likelihood function, and in particular no inferences from M may use probabilities of any part of X except the part which was actually observed.

This establishes the full likelihood principle.

To see that it establishes (2) along the way, note that I have proved that inferences from x to h must depend on x only via $p(x|h)$. So inferences from x to h_1 and h_2 must depend on x only via $p(x|h_1)$ and $p(x|h_2)$. When these are the same, as in (2), the inferences must be the same.

The premises used in this proof are exactly the minimum needed to prove the likelihood principle, as can be proved by proving the premises *from* the likelihood principle. I will do this by showing that it implies each of them separately. (Since it is implied by both of them jointly, it must then

be equivalent to the union, since if $(a \wedge b) \Rightarrow c$ and $c \Rightarrow a$ and $c \Rightarrow b$ then $c \Leftrightarrow (a \wedge b)$.)

It follows directly from the likelihood principle that the correct conclusion in Cox's example is to ignore the characteristics of the laboratory not used. To prove the the weak conditionality principle given above (the formal version of the obvious solution to Cox's paradox), we note that in merriment M^*,

$$p(j, x_j|h) = \frac{1}{2} p_j(x_j|h) \propto p_j(x_j|h).$$

So M^* and M_j have proportional likelihood functions, where M_j is the measurement chosen by the coin toss. Hence M^* and M_j licence identical inferences. Hence only M_j matters.

To prove the weak sufficiency principle from the likelihood principle, note that if T is sufficient for H then (by definition) $p(X|T(X))$ is independent of h. If $T(x_1) = T(x_2)$ (as in the premises of the weak sufficiency principle) then $p(x_1|T(x_1), h) = p(x_2|T(x_2), h)$. Then $p(x_1|h) = p(x_2|h)$ — x_1 and x_2 have identical likelihood functions. So, by the likelihood principle, any inference procedure should draw the same different conclusions from x_1 as from x_2.

This completes the proof that the likelihood principle is logically equivalent to the conjunction of the weak conditionality principle and the weak sufficiency principle.

HOW THE PROOF ILLUSTRATES THE LIKELIHOOD PRINCIPLE

The only functions we needed to consider in the proof of the likelihood principle were $p_1(x_1|h)$, $p_2(x_2|h)$, $p_Y(y|h)$ etc., all considered as functions of h:

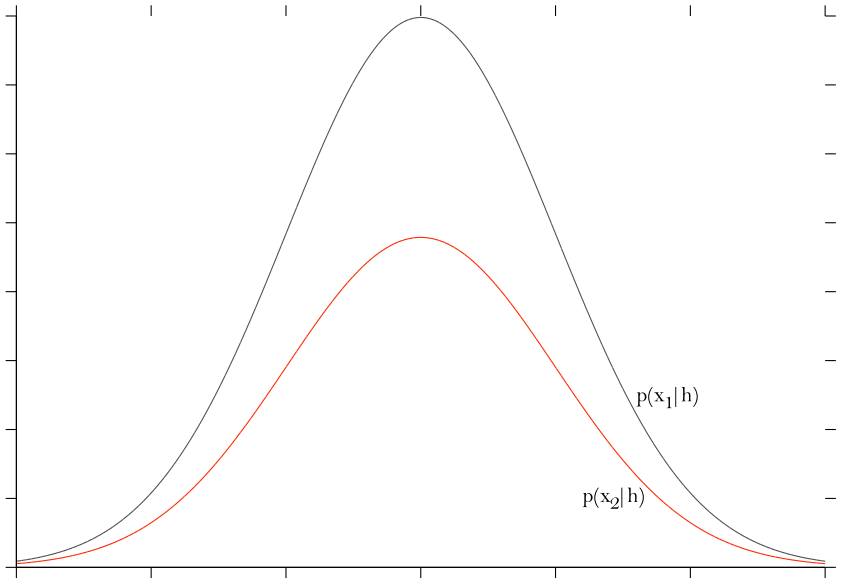

Figure 17

The exciting thing about the likelihood principle is that only $p(x_a|h)$ is relevant to any inference about H. In stark contrast, as we saw in chapter 4 almost all statistical methods in common use rely on $p(x|h_0)$, considered as a function of x (fixing a single hypothesis and imagining the observation varying). This function is, in general, quite unrelated to any of the above functions:

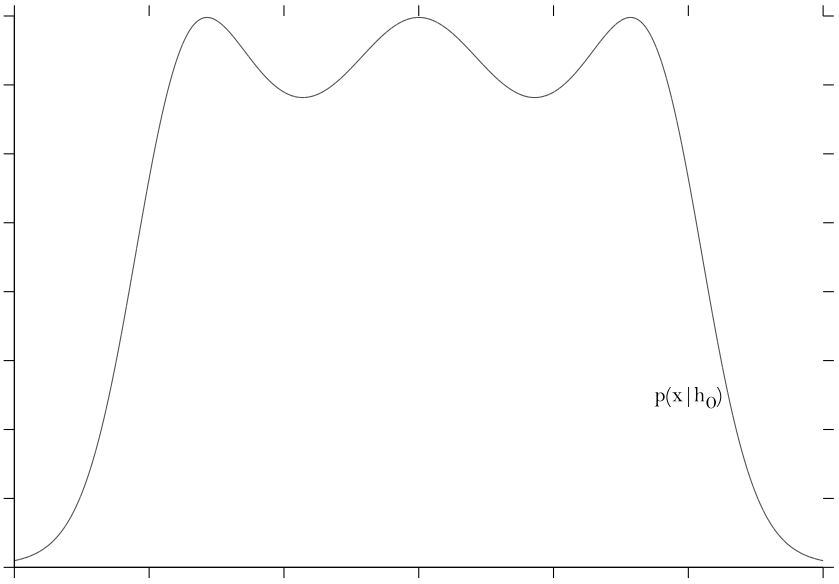

Figure 18: The shape of the graph $p(x|h_0)$ doesn't matter

The likelihood principle says that the shape of this graph is totally irrelevant to inferences from any actual observation (or set of observations — recall that x is generally a vector).

THE LIKELIHOOD PRINCIPLE FOR INFINITE HYPOTHESIS SPACES

The proofs above do not go through for arbitrary probability density functions, because of ambiguities in the notion of sufficiency (Basu 1975) (Evans et al. 1986) (Berger & Wolpert 1988, pp. 28-30). But they do go through (with very minor modifications) for *continuous* functions, where a continuous function is one such that the preimages[122] of topologically open sets are topologically open sets (Berger & Wolpert 1988, p. 30) (Bjørnstad 1996).[123] Any function whose graph can be drawn without taking the

122. The preimage of a set A is the maximal set whose members map onto members of A by the function in question.

123. In the real numbers, topologically open sets are those which have the form of the union of a finite number of intervals (a, b), not including the endpoints a and b.

pen off the paper is continuous. The topological definition is needed to make this idea precise and to generalise it to arbitrary spaces. All of the functions commonly used as probability density functions in applied statistics are continuous.

For the sake of completeness, and because the notation is pretty, I quote the following theorem of Berger and Wolpert, which is proved from measure-theoretically more sophisticated versions of the weak sufficiency principle and weak conditionality principle. The theorem shows that the likelihood principle applies to some (in a sense, most) non-continuous infinite hypothesis spaces.

> Let $\phi : U_1 \to U_2$ be a Borel bimeasurable one-to-one mapping from $U_1 \subset \aleph_1$ onto $U_2 \subset \aleph_2$, and suppose there exists a strictly positive function c on U_1 such that for all $\theta \in \Theta$,
>
> $$p_\theta(A) = \int_{\phi^{-1}(A)} \frac{1}{c(x_1)} p_\theta(dx_1), \quad A \subset U_2.$$
>
> Then if an inference can only be drawn from the observation x if it can also be drawn from the observation $\phi(x)$, for all x except for a set of probability zero (regardless of the value of θ). If it is agreed to ignore the possibility of events of probability zero then inferences about Θ may depend on $\aleph_1$ and $\aleph_2$ only via x and $\phi(x)$.
>
> (Adapted from Berger & Wolpert 1988, pp. 33-34)

Taken together with my proof above, this shows that the likelihood principle is true for any *finite* set of hypotheses and for any *parametric* infinite set of hypotheses and for many non-parametric infinite sets of hypotheses.

BJØRNSTAD'S GENERALISATION OF THE LIKELIHOOD PRINCIPLE

Bjørnstad has proved a version of the likelihood principle which applies even when the hypotheses to be examined depend on the observed data. This is a case which is excluded by my general framework, but I will briefly state Bjørnstad's theorem because it holds the prize as the most general version of the likelihood principle to have been proved to date. In particular, Bjørnstad's theorem is applicable even to instances of the prediction problem (see chapter 2), in which θ is a function of x.

> Define M as (X, h, p) as previously, but this time let the quantity about which we wish to draw inferences be not h but λ, and let λ be a function of x, thus:
>
> $\lambda = \lambda(y, \psi)$, where ψ represents the unknown quantities which are being treated as variables. Let θ represent the unknown quantities which are being treated as parameters.
>
> Let $h = (\psi, \theta)$.
>
> Then inferences from $x \in X$ to λ must be depend on x only via the ordered pair $\langle \lambda, p(x|\lambda, \theta) \rangle$. The first term in this pair is new. The second term is a likelihood function, but not the same likelihood function as in the simpler case proved above (in which it was $p(x|h)$).

See (Bjørnstad 1996) for a proof of this principle.

In real scientific cases $\lambda(x)$ is often independent of x, and in this common case Bjørnstad's likelihood principle reduces to the simpler likelihood principle proved above.

This concludes my discussion of proofs of the likelihood principle. In the next chapter, I give, and answer, objections which have been raised to proofs similar to mine.

— 14 —
Objections to Proofs of the Likelihood Principle

As yet the literature contains no objections specifically directed at my proof of the likelihood principle. But my proof is similar enough to proofs given by (Birnbaum 1962, Birnbaum 1972) and (Berger & Wolpert 1988) that objections to those proofs, if they succeed, may plausibly defeat my proof as well.

The relationship between my work and the work of other authors in this chapter is a little different from the situation with regard to objections to the likelihood principle. I believe that my version of the likelihood priciple is immune to many criticisms which were valid criticisms of earlier versions of the principle; but I do not believe my proof is such a big improvement on earlier proofs that serious objections to the earlier proofs fail to apply to my proof. This is why I do not trouble the reader's patience by reproducing the earlier proofs.

I will show that none of the objections to earlier proofs that I am aware of succeed either as criticisms of my proof (as I demonstrate explicitly) or (implicitly) as criticisms of Birnbaum's or Berger and Wolpert's proofs.

1. OBJECTION 14.1
THE WSP IS FALSE

Evans, Fraser and Monette claim that the Weak Sufficiency Principle is false, on the following grounds. I must quote at length, because the

essential part of their objection — the third paragraph below — is not phrased as an objection. The fact that it is intended to be an objection only becomes clear in context.

> [T]he general deficiency of the ordinary statistical model provides the mechanism for the proofs giving the paradoxical results.
>
> Given the disturbing consequences of Birnbaum's formulation of the common principles [the WSP and the WCP, defined in chapter 13], we examine more closely the meaning and uses of a principle. We recall that $\text{cont}(I_1) = \text{cont}(I_2)$ means that I_1 and I_2 contain the same information concerning the parameter θ. We . . . question to what degree a statistical principle is merely the statement of [such] an equivalence.
>
> Consider the sufficiency principle. . . . the sufficiency principle as described above [my WSP] asserts that [x and a sufficient statistic $T(x)$] contain the same information. Operationally, however, the principle . . . seems to imply more: that we should replace [x] by [$T(x)$] for purposes of inference. For associated with any inference base is a wealth of inference procedures that can commonly be invoked, and in replacing [x] by [$T(x)$] we are restricting this class, unless of course [T] is trivial. In this sense *sufficiency* can be viewed as an operational step towards cont, and would be more than a mere statement of equivalence.
>
> . . . Birnbaum did not address these aspects of the principles, only treating them as equivalence relations. Accordingly his proofs . . . allow the use of the principles in contexts where the justification for the principles is violated. Such applications are clearly inappropriate and indicate at least that some clarification is needed of the principle, or of the application context
>
> (Evans et al. 1986, pp. 191–192)

It seems to me that Birnbaum's proofs are not subject to this objection, but I cannot show that without a considerable aside on Birnbaum's work. Instead, I will merely defend my own proof.

There is a little confusion in this objection. Neither I nor other writers on the likelihood principle say that one must replace x with an arbitrary sufficient statistic $T(x)$. It would clearly be daft to say that, since there are many such sufficient statistics. Nor do I assert that the WSP says that one must do this.

Perhaps Evans, Fraser and Monette believe that the WSP is obviously false. However, they do not give any arguments against it apart from the above argument which, as I have just shown, misses its target. So, although I would like to support the WSP against objections, there is nothing explicit for me to argue against. Instead, I refer to the arguments I gave in favour of the WSP in chapter 13, and hope that the considerations I used there outweigh whatever considerations lie behind Evans, Fraser and Monette's objection.

2. OBJECTION 14.2
THE WCP IMPLIES THAT IT SHOULD BE IRRELEVANT WHICH MERRIMENT OCCURS

In addition to objecting to the WSP, Evans, Fraser and Monette object to a class of proofs of which mine is one. I will translate the objection into the terminology of my proof, in square brackets. I do not quote the terms from which I am translating, because if I did so the objection would become unreadably messy.

> consider a context in which a statistician who accepts [the Weak Conditionality Principle and the Weak Sufficiency Principle] is presented with the mixture inference base $[M^*]$. Conditionality indicates that the relevant model for inference about θ is given by $[M_1]$. On the other hand, application of [the WSP] establishes—as is clearly seen via the sufficient statistic $[T(j, x_j) = t_0$ if $(j, x_j) = (1, x_1)$ or $(2, x_2)$, which I proved above to be sufficient for h provided the preconditions of the likelihood principle are met]—that the information as to which model has occurred is irrelevant information for inferences about $[h]$. The statistician is presented with contradictory recommendations from these principles.
>
> (Evans et al. 1986, p. 190)

This objection is mistaken, because the Weak Conditionality Principle does not say that the information as to "which model has occurred" (which way the coin toss came out and hence which merriment has been conducted) is irrelevant to inferences about h. It says that the information as to which model has occurred is irrelevant to inferences about h *given* the likelihood function $p(j, x_j)$. This is just as it should be: we need to know which merriment has been made, since otherwise we could not interpret the results. But we don't need to know which merriment has been made once we have the likelihood function, since ex hypothesi it is the same for each. Of course these recommendations are only *correct* if the likelihood principle is correct, but even if the likelihood principle is wrong the recommendations are not (as Evans, Fraser and Monette claim) contradictory.

3. OBJECTION 14.3
THE PROOF FAILS IF THE WCP IS LIMITED TO CONDITIONING ON A MINIMAL SUFFICIENT STATISTIC

A minimal sufficient statistic is a sufficient statistic $t(x)$ such that any other sufficient statistic $T(x)$ depends on x only via $t(x)$. Minimal sufficient statistics do not always exist, and when they do they are not always unique.

Durbin (1970) shows that if the WCP is restricted to conditioning on a minimal sufficient statistic it can no longer play the role it is required to play in (Birnbaum's, or my) proof of the likelihood principle.[124]

Why might one possibly think that the WCP should be restricted in this way? Durbin only sketches an answer:

> Birnbaum's sufficiency principle [similar to the WSP] implies that, as a function of the observations, evidential meaning depends only on the minimal sufficient statistic, where this exists . . . Since evidential meaning depends only on the minimal sufficient statistic it would seem reasonable to require that any analysis or interpretation of the results of the experiment should depend only on the value of the minimal sufficient statistic. This leads naturally to the requirement that the domain of applicability of (C) [roughly, the WCP] should be restricted to components of the minimal sufficient statistic.
>
> (Durbin 1970, pp. 395–396)

124. I am not sure that this point is intended by Durbin himself to be an objection to a proof of the likelihood principle, but it has been cited as an objection of this sort, without any elaboration, by a number of other authors.

In this passage, Durbin talks about conditioning only on "components of" minimal sufficient statistics. He does not say what he means by "components of" a statistic, and he does not use this qualification anywhere else in his paper, so we must consider the possibility that it is a typographical error; however, his argument is much more plausible if it is not a typographical error but is, rather, what he means to say throughout his paper. I therefore consider both possibilities.

OBJECTION 14.3.1
THE WCP SHOULD BE LIMITED TO CONDITIONING ON A MINIMAL SUFFICIENT STATISTIC

For the duration of this section, I ignore the idea of "components of" a minimal sufficient statistic, and take it that Durbin is asserting that the WCP should be restricted to conditioning on a minimal sufficient statistic itself. Then Durbin's argument may be paraphrased as follows: in some cases, evidential meaning depends only on the minimal sufficient statistic (assuming the WSP is true; if not, then of course any proof based on it is unsound); but one should only condition on evidentially important variables; hence one should condition only on minimal sufficient statistics in general and, a fortiori, in applications of the WCP.

It is true that according to Birnbaum's proof only sufficient statistics (and hence only minimal sufficient statistics, where such exist) are evidentially important to inferences about H. This follows from Birnbaum's sufficiency principle, which is as follows:

> *Principle of Sufficiency (S):* Let E be any experiment, with sample space $\{x\}$, and let $t(x)$ be any sufficient statistic (not necessarily real-valued). Let E' denote the derived experiment, having the

> same parameter space, such that when any outcome x of E is observed the corresponding outcome $t = t(x)$ of E' is observed. Then for each x, $\mathrm{Ev}(E, x) = \mathrm{Ev}(E', t)$, where $t = t(x)$.
>
> (Birnbaum 1962, p. 278)

But there is no reason to restrict conditioning to "evidentially important variables". This is one place where Birnbaum's proof is significantly different from mine, for my premises say nothing at all about what is "evidentially important", while Birnbaum's do. Birnbaum's terminology of evidential importance, as encapsulated in the sufficiency principle above, may suggest that only evidentially important variables are important simpliciter. But of course that need not be the case, and in applying the WCP it certainly is not the case. For although the variables in question are (according to Birnbaum) the only evidentially important variables *for inferences about H*, they are not the only important variables for analysis of the structure of the merriment. A variable can be vital for the latter purpose while, on its own, carrying no information at all about H and hence not being a sufficient statistic, never mind a minimal sufficient statistic.[125]

I have shown that Durbin has no clear argument against Birnbaum's proof, even if we allow the terminology of "evidentially important" variables on which Durbin's argument relies. If we avoid that terminology, as I do in my proof, Durbin's argument becomes even weaker. Recall that the only

---

125. Cox's example illustrates this point nicely. Recall that in Cox's example the toss of a coin determines which laboratory receives a sample of blood. In the context of the example, H is a set of hypotheses about the blood. The result of the coin toss on its own carries no information about the blood, and so in Birnbaum's limited sense it is not evidentially important for H. It is, however, prima facie reasonable to condition on it — indeed, any statistician, even a Frequentist statistician, would condition on it (in the Frequentist's case, using the rationale that it is an ancillary statistic), and Durbin offers no argument against doing so except for his incorrect assertion that the WSP tells us not to.

Note, for future reference, that the pair (coin toss, laboratory result) is a minimal sufficient statistic for Cox's example, and so the coin toss is in some sense a *component* of a minimal sufficient statistic; I will show the importance of this in the next section.

version of the sufficiency principle which I use in my proof, namely the WSP, says: "If $T(X)$ is a sufficient statistic for h, and if $T(x_1) = T(x_2)$, then any procedure that derives different inferences about h from x_1 and x_2 is incoherent." My WSP does not say that only sufficient statistics are evidentially important, never mind that one should condition only on them. Durbin's argument could perhaps be recast in terms of the incoherence of inferences based on other than sufficient statistics, but then the fact that it is only inferences about h which are so constrained would be even more obvious than it was in the previous paragraphs, and hence again Durbin's argument would fail.

OBJECTION 14.3.2
THE WCP SHOULD BE LIMITED TO CONDITIONING ON THE COMPONENTS OF A MINIMAL SUFFICIENT STATISTIC

The possibility remains that Durbin's reference to the components of a minimal sufficient statistic was not a typographical error. Perhaps Durbin's assertion is best understood as being that one should only condition on the components of a minimal sufficient statistic. This assertion is plausibly true, if "components of" is taken to mean "functions which are part of a factorisation of" or, more weakly, "functions which are functions of", as Birnbaum (1970, p. 402) suggests Durbin's phrase should be interpreted. But the use I make of the WCP is compatible with this interpretation of Durbin's assertion, except for one special case which I will deal with in the next paragraph.

Recall that the WCP says nothing more than that we may condition on the result of a coin toss in a mixture experiment such as the Cox example.

As I mentioned in the previous section, the coin toss is a component of the pair (coin toss, laboratory result) which in turn is a minimal sufficient statistic for Cox's example; similarly, in the terminology of my proof, the result of the coin toss, j, is a component of the pair (j, x_j), which is minimal sufficient for the mixed merriment M^*. Hence Durbin's assertion is compatible with my use of the WCP, which is merely to condition on j, except for one special case which must be dealt with separately.

Berger & Wolpert (1988) interpret Durbin as I do in this section — that is, as saying that we may condition on any part of a factorisation of a sufficient statistic — and perhaps it was stupid of me to consider any other interpretation. Berger & Wolpert also note that a special case arises when the two experiments which are performed as a result of the coin toss of the WCP happen to give results x_1 and x_2 which have proportional likelihood functions $\big((\exists c)\,(\forall h)\,p(x_1|h) = c.\,p(x_2|h)\big)$. In this case alone, the coin toss is not part of any (non-trivial) factorisation of the minimal sufficient statistic, since the minimal sufficient statistic in this case is the (shared) likelihood function. So it follows from Durbin's assumption that the likelihood principle as it applies to this particular case cannot be considered proved. However, considering such a case makes it particularly clear why we should not accept Durbin's assertion. Berger and Wolpert illustrate this by applying Cox's example to two laboratories, one in California and one in New York:

> by Durbin's argument, whether or not one chooses to condition on the actually performed California experiment with observation $[x_a = x_1$, say$]$ would depend on the existence, or lack thereof, of an observation $[x_2]$, in the unperformed New York experiment, having a likelihood function proportional to that of $[x_1]$. Such

> dependence of conditioning on the incidental structure of an *unperformed* experiment would be rather bizarre.
>
> (Berger & Wolpert 1988, p. 46)

Berger and Wolpert are content to let their case rest there. To clarify why, note that although Frequentist theory says that the possible outcomes of unperformed parts of mixed experiments are relevant because they form part of the sample space X, Durbin's assertion entails that something much more complicated: that the whole likelihood function $p(x_2|h)$ of an unobserved part of X is relevant to inference *if* it happens to be proportional to $p(x_1|h)$ but not otherwise. I cannot see any reason to accept this and, as far as I know, no argument in its favour has ever been presented. (Certainly Durbin presents none.) Savage (1970) points out that the inherent implausibility of this idea is exacerbated by the fact that it makes $p(x_2|h)$ relevant only if it is *exactly* proportional to $p(x_1|h)$, but otherwise completely irrelevant; hence, if there is any doubt at all about the exact value of any part of $p(x_2|h)$, Durbin's argument becomes impossible to apply. It seems to me that this is a sufficient argument against Durbin's assertion in its second interpretation.

— 15 —
Consequences of Adopting the Likelihood Principle

In this chapter I first of all give a case study which I hope will make clear the importance and urgency of the likelihood principle. I then consolidate a number of theoretical conclusions which I have drawn in the thesis as a whole, and present some of their general implications for applied statistics and hence for most of science.

1. A CASE STUDY

INTRODUCTION

It is now time to consider an example more realistic than that of Table 1. In this case study, I will describe an area of scientific enquiry — namely, large clinical trials — which has been extensively studied but in which no consensus has been achieved on the best method of statistical inference. I will sketch the history of the study of inference methods in this area. The history will show particularly clearly the ad hockery of Frequentist methods: Frequentist methods in this case are *so* ad hoc that not even the most committed Frequentists have been able to claim that there is any unique optimal Frequentist solution to the inference problem (at least, not to date). Literally dozens of Frequentist methods are available, no two of which give equivalent results (with trivial exceptions), and no method for choosing between them is available to Frequentists.

As a matter of pragmatics — and the applied statistics community is nothing if not pragmatic — the statisticians who design clinical trials would benefit enormously from standardising on a single method, so that their results could not be challenged by regulatory authorities, drug companies or consumers.[126] Therefore, clinical trials centres have attempted to standardise on a single Frequentist method; but the methods are so ad hoc, and there is so little to choose between them, that they have not been able to do so. Of course the failure to standardise has depended on social issues as well as technical issues; but the technical issues have not been irrelevant. The result, to date, is that two or three Frequentist methods have become popular but none has become dominant. This lack of standardisation in itself represents a major scientific problem, in addition to the further problem that Frequentist methods are (as I have argued) often uninformative about H.

In this case study I will describe an obvious solution to the problem which is compatible with the likelihood principle. This will be a Subjective Bayesian solution. This solution does not suffer from any of the ad hocness of the Frequentist solutions, but it has not been acceptable to regulatory bodies for two reasons:

1. its subjective nature is fundamentally unacceptable to public regulators (rightly or wrongly); and

2. its Frequentist error rates have not been known until recently.

126. The ability for a plaintiff to challenge a scientific result depends almost entirely on whether the result was achieved using standard methods, and practically not at all on whether the result was achieved using rational methods.

I will describe a modification of the Subjective Bayesian solution which remains compatible with the likelihood principle but which avoids the above two objections: it is not subjective (it does use prior probabilities, but they are not set subjectively), and its Frequentist error rates are known and, moreover, are excellent.[127]

This case study will show that the likelihood principle can be used to formulate methods which are less ad hoc, as well as (as the main part of this thesis has argued) more epistemically coherent, than the orthodox Frequentist methods. It will thus demonstrate that the arguments of this thesis have practical importance. My main aim in giving this case study is to show the beneficial effects that have accrued to the sections of the statistical community that accept the likelihood principle, and the otherwise intractable problems that have been faced by the sections of the statistical community that do not accept it.

I present this study in some historical detail. I concentrate on the philosophical aspects of the history, but not to the exclusion of the scientific details. There is a rationale for this. Philosophers' toy examples of scientific practice, of the sort I have used up to this point, are unsafe: one can never be sure to what extent the lessons learned from them are relevant to what scientists actually do, unless one checks them against a real example which is sufficiently complicated to have some hope of being a fair representative of science as it is practised. This case study therefore uses a reasonably complicated example of scientific practice, alluding (although necessarily

127. Mayo objects to non-Frequentists citing the good Frequentist error rates of likelihood methods as a reason to use those methods, but I do not see the force of her objection. It is true that a non-Frequentist does not believe that it is *rational* to care about error rates, but it is extremely rational to want to use a method which one's opponents consider to be rational, both for social reasons and just in case one's philosophy turns out to be wrong.

briefly) to a number of complexities which bear on the importance of the likelihood principle and which would not be evident in a simpler example.

SEQUENTIAL CLINICAL TRIALS

My case study is on stopping rules for large clinical trials.[128] Meier has nicely introduced the importance of such stopping rules by comparing clinical trials to the agricultural trials with which Fisher was familiar when he developed the methods described in chapter 4:

> The planning, execution, and analysis of an agricultural field experiment are all well separated in time. The intended design, if properly executed, will be the framework for the final analysis. Long-term clinical trials, by contrast, are still recruiting patients when the findings of analysis begin to emerge. These findings may quite properly cause the design to change in radical ways — even, on occasion, leading to early termination of the study.
>
> For a time it was possible to consider such decision making as outside the domain of statistical analysis and to regard it rather as the intrusion of extrastatistical humane considerations that caused us on occasion to terminate or alter an ongoing study.
>
> More recently it has become clear that the possibility of changes in the study brought about by early findings is not a rare incursion by extrascientific elements but rather a necessary and typical feature of this type of clinical experimentation.
>
> (Meier 1981, p. 340)

128. I introduced the idea of a stopping rule in chapter 12: recall that a stopping rule is an agreement by experimenters and statistical analysts to execute an experiment in parts, with each part being subjected to a pre-agreed type of statistical analysis as soon as possible after its completion, and with the series of sub-experiments guaranteed to terminate "early" (before some pre-agreed maximum sample size has been reached) if one of the analyses has some pre-agreed outcome. Typically the only outcome which is allowed to cause early termination of the experiment is a pre-agreed death rate among the experimental subjects. (There is some ambiguity in "allowed to", but that need not concern us here.)

The outcome (typically, the number of deaths) required to trigger early termination of a clinical trial is typically, but not necessarily, worked out by calculating a pre-agreed level of significance against some pre-agreed null hypothesis. If the null hypothesis is considered refuted then the treatment is considered to have been proved efficacious, and the trial stops. If the null hypothesis is not considered refuted then the trial continues until some pre-agreed sample size is reached.[129]

If a statistical analysis is performed after each trial subject has yielded an outcome (typically, either dying, or living for a pre-agreed period) then the experiment is called a *fully sequential* trial. If a new statistical analysis is performed every time a new group containing a pre-agreed number of subjects has yielded an outcome (or, equivalently, if the experiment is analysed up to a pre-agreed number of times before it reaches its maximum sample size), it is called a *group sequential trial*.[130]

Group sequential methods are intended to provide statistically legitimate methods for monitoring accumulating data, with the possibility of stopping a trial before it has reached its maximum size. For various economic reasons, group sequential theory concentrates on *phase III* trials: that is, large, randomised, controlled trials on a more or less representative

---

129. The rationale for proceeding in this way is that an experiment on human subjects is only considered ethical (by the bulk of the medical community) if there is *equipoise*: that is, if and only if the treatment given to the experimental subjects is neither confidently believed to be efficacious (in which case it ought to be given to the control group too, thus making a clinical trial impossible) nor confidently believed to be inefficacious (in which case it ought not to be given to anybody, including the trial subjects). Early termination is often desirable, either because equipoise has been lost or because it comes to seem unlikely that the trial will reach any conclusion. The marginal cost of recruiting new subjects to a trial is high, so trials which will probably be inconclusive are to be abandoned as soon as possible. Various ethical arguments can be made against each aspect of this view, but the fact that it has been the accepted ethics of the medical research community since the second world war is enough to make it a sine qua non of the statistical methods considered in this case study, regardless of whether it is right.

130. I will sometimes use the more general term *sequential trial* to refer indiscriminately to fully or group sequential trials.

population which are conducted to elucidate the best therapy for a given condition. Group sequential clinical trials gradually gained in popularity over the period covered by my case study, especially for large trials. Nowadays they are almost the only method used for large drug trials (at least, prior to government approval of drugs for population-wide use; after such approval, different methods are used, such as comparisons of individuals who have side effects with groups of individuals who don't — these are so-called *case-control studies*).

The case study I have chosen is typical in many ways of the problems encountered in twentieth-century applied statistics, although it is special in the degree to which the scientists involved have discussed issues bearing on the likelihood principle. Indeed, the likelihood principle was first stated as a contribution to the debate I will present (Barnard 1947), and this first statement of the likelihood principle was immediately followed (within a few words) by the stopping rule principle (defined in chapter 12).

In this case study, in keeping with the rest of the thesis, I treat the statistical analysis of clinical trials as a problem of inference about hypotheses, as opposed to treating it as decision theory. There have been many attempts to use decision theory to formulate sequential methods, but they are relatively unimportant, because the main question which is asked in a medical context is not "how can we maximise benefit?" (although this may be asked occasionally, for example when a drug is prohibitively expensive). It is usually something much simpler, which can be answered without recourse to decision-theoretic assumptions: "how effective is this intervention in this population?"

All of the statistical methods which have been used in phase III trials of drugs (pharmaceuticals) and clinical implants, if they have been used as they were intended to be used, have been Frequentist and hence have approximately fixed the overall type I error of the experiment. As I have already discussed in chapter 7, this leads to epistemic paradoxes. These problems are worse than usual in the case of sequential trials, and they are joined by some brand new problems.

As we saw in chapter 4, the essence of Frequentist statistics is that a probabilistic choice should be made in such a way as to do as well as possible in an arbitrarily long sequence of repetitions of the situation which led to the choice. This has been formalised in several ways, most notably by Neyman in his theory of hypothesis tests. To recap a little, Neyman's theory has passed on to modern Frequentist statistics a controversial feature: the only admissible pieces of evidence about a statistical procedure are the properties of the procedure averaged over the sample space (X). Properties of the procedure conditional on the occurrence of particular events in the sample space are not relevant except as part of such an average. This includes properties conditional on the event which actually occurs (x_a). Therefore, evidence from the experiment itself is not permissible in characterising the procedure.

This criterion is not so very strong, in the ordinary run of things, because if one wants to condition on an event x_a which happens after an experiment, A, has been defined, one only has to start a new experiment, B; the design of this new experiment can then depend on x_a in any way one pleases. This is a typical Frequentist statistician's (partial) solution to

the problems raised in chapter 4: a Neyman non-epistemic probability can be made to seem much more rational if this trick is executed ad libitum.[131] This trick is circumvented, though, when it is not feasible to perform a new experiment to take into account the new data — particularly in sequential applications, where x_a is recorded in an interim analysis. This point makes sequential analysis a field of enquiry in which the differences between Frequentist and likelihood inference are brought into particularly sharp relief.

In addition to such basic epistemological problems of Frequentist methods, some new mathematical problems arise in sequential trials. The worst of these is the incompatibility of sequentially calculated P-values and confidence intervals. In the absence of multiplicity (a concept which is explained in chapter 7 and again below), the endpoints of standard confidence intervals are P-values, but in the presence of multiplicity they generally are not. Since sequential trials are always subject to multiplicity, fixing the type I error of a sequential trial generally (i.e., except in trivial cases) causes it to provide point and interval estimates which are incompatible with each other.

This problem is best understood by first considering a simpler problem which makes P-values problematic in their own right:

> We need to be extra careful with the term *statistically significant difference* in the optional stopping case. Here, one keeps taking more and more samples until the observed difference is *computed* to be statistically significant . . . The computed significance level with an optional stopping plan refers to the significance level

131. Whether this trick is allowed by Neyman's own theory is neither important nor clear: as I mentioned in chapter 4, Neyman does not tell us exactly when, if ever, a reference class can be changed, although he does strongly imply that it should not happen during a statistical analysis.

> that would be calculated under a fixed sample size plan . . . Say it took k tries to achieve a difference computer to be .05 statistically significant. The *actual* or overall significance level is the probability that out of k tries at least one would be computed to be .05 statistically significant, even if the null hypothesis is true.
>
> (Mayo 1996, p. 343)

The sequential ("optional stopping") problem which Mayo describes for Frequentists is very simple. Frequentists, by definition, consider it necessary to design trials such that they have a predetermined overall error rate, and the error rate which they consider it most important to fix is the type I error (which they usually set at 5%). Analysing the results more than once gives an overall type I error for the trial greater than the type I error of each analysis.[132] The Frequentist's sequential problem is how best to adjust the individual analyses in order to fix the overall type I

132. Recall that the type I error is the probability (in the non-epistemic, Neyman sense) of rejecting the null hypothesis, conditional on the null hypothesis being true; and in repeated analyses there is more than one opportunity to do so, so the overall type I error is greater than the individual type I error of any of the analyses.

Why is the null hypothesis considered so important? In biostatistics the null hypothesis is generally taken to be the statement that a treatment has no more effect than standard treatment or a placebo (whenever such a statement is meaningful, which it always is in a large drug trial, at least in rich countries, in which treatments are standardised). Given this, type I error is particularly close to the hearts of clinical trial designers because of the overriding principle of *nonmaleficence*: above all, the trial must not mistakenly report a new treatment as effective. It is considered much better to risk perpetuating an inferior standard treatment. It is almost universally held that the principle of nonmaleficence is represented in statistical terms by a very small overall type I error. Consequently, maintaining a small type I error is of supreme importance to trial designers.

In evaluating this position, it is important to realise that the type I error is defined as the *Neyman* (pre-trial, relative to a fixed reference class) probability *under the null hypothesis* of falsely claiming a positive result. This quantity is often treated, by clinicians and statisticians alike, as though it were equal to the probability of falsely claiming a positive result — in other words, the two italicised phrases tend to be used when the type I error is calculated but ignored when it is interpreted. The statement that the type I error represents the principle of nonmaleficence is an example of such a confusion. It is easy to construct artificial examples in which it leads to absurdities, along the lines of the examples of chapter 7. Whether it leads to absurdities in real life is a different question, and one which has been examined very little. Certainly, one would have thought that if confusion about type I error was going to be a problem anywhere it would be in sequential medical trials, because of the relevance of type I error to the Frequentist's sequential problem. We will see that this is indeed the case.

error. It is easy to find a way to do this; what is hard is to choose the best way from among the many (in fact, infinitely many) alternatives. It turns out that several of the various methods of adjustment which have been proposed to fix the overall type I error are equally acceptable to the statistical community, which leads to an embarrassing problem of deciding which to use — embarrassing because the various methods give different answers. A drug company does not want to have to say to the regulatory agencies (or to a court, in the event of litigation), "our drug is acceptable according to statistical method 1 but not according to method 2, and we have no way of choosing between these methods".

Multiplicity

Recall the general problem of multiplicity in Frequentist methods which I described in chapter 7. The best that can be said for Frequentist methods is that they have the property that if the same analysis is repeated on a long sequence of experiments which are identical except for random variation they will make errors in a known proportion of cases, conditional on the null hypothesis. Even this is not true in practice, because measurement error is generally not included in the model; but let us leave that issue to one side. Still, Frequentist methods fail to have this attractive property in typical applications because practically no experiment calculates a single Frequentist statistic. When more than one is calculated, each one has a chance of being in error, so the statistical analyst faces a dilemma, which I present here in terms of P-values for simplicity but which could be described in terms of any Frequentist measure including the coverage of confidence intervals. The Frequentist's dilemma is that he must either:

- give each P-value an error rate of 5%, in which case the analysis as a whole will have an error rate greater than 5% and, in many cases, approaching 100%;

or

- adjust each P-value so that the overall error rate of the analysis remains 5%.

Since the whole point of Frequentist theory is to limit overall error rates, a fully Frequentist theory must take the second fork of the dilemma and adjust each P-value (Neyman 1937, Kendall & Stuart 1967, Stuart et al. 1999, Mayo 1996), even though this means that the error rate of each interim analysis is changed in an arbitrary fashion to suit the context in which the interim analysis happens to take place.

I noted in chapter 7 that the correction for multiplicity usually takes the form of a Bonferroni correction: that is to say, the cut-off for attributing statistical significance "at the 5% level" becomes (5% / n), where n is the number of P-values being calculated. The Bonferroni method can also be applied to the endpoints of confidence limits. However, the Bonferroni correction does not give the right answer (an overall error rate of 5%) in sequential trials.[133] Once the Bonferroni correction is abandoned, we get a problem even worse than ad hocness: it becomes impossible to find a

133. The Bonferroni correction only works, in the Frequentist's sense of "works", when the P-values being combined are independent (in the statistical sense of "independent" explained in chapter 13); but the repeated measurements on the same subjects which are made in sequential trials are not independent. This is for two reasons. A single subject's health at one point of time is, of course, not independent of the same subject's health at a previous point of time; and even if it were, the measurements which are subjected to analysis in a Frequentist analysis are cumulative, so that the analysis at time t includes all the data from times $< t$, and necessarily so or else important information would be being discarded in the later analyses. Hence the measurements cannot be independent of each other, not merely as a point of biology but also as a point of mathematics. So the Bonferroni correction does not apply. Consequently, a Frequentist analysis of clinical trials is saddled with the problem of finding a *mathematically* valid correction for multiplicity (i.e., one which gives an overall error rate of 5%), in addition to the epistemological problems raised by any such method.

method of correcting P-values and confidence intervals which leaves them compatible with each other (in the obvious sense that the P-values are the endpoints of confidence intervals).

Are these problems bad enough to warrant abandoning Frequentist methods for designing clinical trials?

The two groups of commentators on this question — the yessers and the noers — have, since the 1950s, drawn up battle lines along the great divide between proponents of the likelihood principle and Frequentists. As we will see, the yessers have often been drawn into prodigious complexities in trying to solve the sequential problem, while the noers have usually been content to rest on their laurels, have not, until (Grossman et al. 1994), published in detail the statistical procedures which they recommend, and until very recently have been roundly ignored by practising statisticians.

I will give my own answer to this question gradually. I have already presented many problems with Frequentist methods (in chapter 7) — enough to answer the question peremptorily — but rather than assume that my arguments there have been successful in showing that we should not use Frequentist methods, I propose to approach the issue from a different angle in this case study. I will arrive in essentially the same place as I did in chapter 7, but with more of an emphasis on the problem of multiplicity which I briefly introduced there and the ad hocness which that particular problem introduces, and less of an emphasis on the fundamental epistemological incoherence of Frequentism. In order to give this different angle a chance to shed fresh light on the problem, I will put the objections of chapter 7 to one side for the time being, returning to them only at the end of the chapter when I sum up the whole thesis.

In this case study (unlike the rest of the thesis), all of the supporters of the likelihood principle of any importance have been Bayesians. That Bayesianism and Frequentism would come into particular conflict over sequential clinical trials has been noted often in the literature. For example:

> At the heart of this debate are two conflicting fundamental principles for assessing the meaning of experimental data. These are the Likelihood Principle and the [Frequentist] Repeated Sampling Principle [that only the properties of a procedure on repeated application are important]. If we accept the Likelihood Principle, it follows that all inferences should be based on the experiment that was was actually performed, the data that was actually obtained, and the relative probabilities of obtaining the observed results under various plausible alternative hypotheses. If we accept the Repeated Sampling Principle, it follows . . . that strength of evidence should be quantified by sequentially adjusted P values or other probabilities. In most of statistics, these two principles lead to remarkably similar inferences. However, in sequential clinical trials, they come into sharp contradiction.
>
> (Dupont 1984, p. 277)

Conditioning and the likelihood principle

The only system compatible with the likelihood principle which has been applied to sequential analysis is the Bayesian system. Although I do not wish to defend any form of Bayesianism in all its details, I hope to show in this case study that Bayesianism fares better than Frequentism. It will then follow that the best of the methods compatible with the likelihood principle, whether that be Bayesianism or not, will fare at least as well as Bayesianism and hence better than Frequentism.

The Bayesian system, of course, has the drawback of demanding that every problem is analysed using an expression of belief which is ulterior ("prior") to the experiment. Bayesians who have worked on sequential analysis have developed an interesting approach to this problem. It is well known to clinical epidemiologists that although the likelihood principle allows one to calculate likelihood ratios for diagnostic tests, it is impossible to give the rate at which a test for a disease gives false positive results without specifying the prevalence of disease in the population to which it will be applied. In a way which is mathematically exactly analogous, these Bayesians claim that the false positive rate of clinical trials can be specified by finding the prevalence of false positives in similar trials. This false positive rate can then (they argue) be used to construct a reasonable prior probability distribution for the trial's main parameters.

Although such Bayesians clearly have epistemological problems of their own, they easily avoid the ad hoc choice between many different methods of adjusting type I errors which faces Frequentists. Bayesians need not care about type I error rates. The only constraints on their procedures are the prior probability distribution and the probability calculus; from these two ingredients, a Bayesian can directly calculate the probabilities of hypotheses, as we saw in chapter 3, and Bayes's Theorem guarantees that there is only one way to do this.

I will give the mathematical details of this Bayesian approach later. First, here is a concise history of Frequentist approaches to the problem.

A BRIEF HISTORY OF GROUP SEQUENTIAL PROCEDURES

The first important publication on sequential analysis was (Wald 1947). This book established general Frequentist methods for fully sequential analysis. Bross was the first to apply sequential methods to medical research, in his (Bross 1952). The work of Wald, Bross and others was followed by (Armitage 1960), which adapted Wald's methods to clinical trials. These books established the importance of adjusting P-values in Frequentist sequential analysis.

The first statement of the likelihood principle (and one which I quoted at length in chapter 8) was by Barnard in (1947, p. 659). An advocate of Barnard's new likelihood principle, Anscombe, reviewed Wald's book in 1954 and Armitage's in 1963. In the latter review, he asserts that

> 'Sequential analysis' is a hoax. The correct statistical analysis of the observations consists primarily of quoting the likelihood function. So long as all observations made are fairly reported, the sequential stopping rule that may or may not have been followed is irrelevant.
>
> (Anscombe 1963, p. 381)

One direct result of Anscombe's 1963 article was that a group of prominent clinical trials statisticians at the U.S. National Institutes of Health held a seminar to discuss the role of hypothesis testing from a practical point of view. This seminar lead to the first published suggestion of *group* sequential trials, by Shaw (Cutler et al. 1966).

In 1969, Armitage, McPherson and Rowe (1969) first discussed group sequential methods in detail. They introduce there what was later to become known as the Pocock stopping rule based on uniformly spaced

analyses and a uniformly distributed correction for multiplicity (the closest thing possible to a Bonferroni correction). They give approximate values for the necessary Frequentist adjustment for 5, 10, 15, 20, 50, 100 and 200 analyses. They also discuss the appropriateness of repeated Frequentist significance testing, and note the lack of error rates for Bayesian group sequential methods:

> The exchanges of opinion on these matters have been remarkable for the lack of quantitative information about the optimal stopping effect. It has not, for example, been possible to answer questions such as the following.
>
> (a) What is the probability of obtaining a result "significant" at a certain nominal level, within the first 50 tests?
>
> (b) Does the enhancement of the probability of obtaining a significant result reach a noticeably high level only after a very large number of tests?
>
> (c) What is the effect of repeated tests when the null hypothesis is not true?
>
> (Armitage et al. 1969)

Such unanswered questions dominated the debate for three decades. These error rates were soon quantified for repeated significance tests, but not for Bayesian methods. They were quantified for Bayesian tests for the first time (albeit only for a simple family of priors, and considering only up to 10 tests) in (Grossman et al. 1994).

Subsequently, there was an extraordinary proliferation of mutually incompatible Frequentist group sequential methods.[134]

---

134. Let me present a few prominent examples from the literally dozens of methods. A reader who is willing to take my word for the fact that the various Frequentist methods differ substantially from each other can skip this footnote.

- Haybittle (1971) suggests an extremely simple group sequential procedure: stop the trial if the difference between the treatment and control groups at an interim analysis exceeds three standard deviations (on some parameter of interest).
- Elfring and Schultz (1973) present a very advanced group sequential plan for binary outcomes. It incorporates features which were not rediscovered until much later: for example, it gives a stopping rule for incorrectly *failing* to reject the null hypothesis as well as one for incorrectly rejecting it, an idea reinvented by Emerson and Fleming 16 years later (1989). Elfring and Schultz also allow for interim analyses to be conducted on variably-sized groups of trial subjects (not permitted by any other method until 1983). But as far as I know, Elfring and Schultz's method has never been used in a large phase III trial. A major drawback was that their stopping rule is not given explicitly but has to be separately computed by simulation for each trial.
- Peto and nine colleagues (1976) propose considering a P-value significant at interim analyses only if it reaches some arbitrary very extreme value, followed by an unadjusted P-value test in the final analysis.
- Pocock (1978) suggests a simple P-value test with a significance level of 1% for every analysis (including the final analysis) in a trial with up to 11 analyses. This gives an overall significance level of *less* than 5%. (This is not the same as the standard "Pocock" test described above and tabulated below.)
- O'Brien and Fleming (1979) suggest a stopping rule which becomes less conservative (more likely to reject the null hypothesis) as time passes. Formally, an O'Brien and Fleming trial stops iff $\chi^2(n) > k / n$, where k is a constant. This increases the power of the study (decreases the type II error), but it also increases the discrepancy between unadjusted and adjusted significance levels. This is currently the most popular stopping rule in large trials, along with the original Pocock method (the one described in (Armitage et al. 1969), not the one described in (Pocock 1978)).
- DeMets and Ware (1980) suggest a one-sided design, which means that one chosen arm of the trial (usually the placebo arm) cannot be found to be better than the other. Such a design has higher power (lower type II error) than the more usual two-sided design. The issue of whether we should use one-sided or two-sided designs can be generalised to the issue of whether the null hypothesis should be no treatment difference, a small treatment difference, or a small or negative treatment difference. All of these possibilities have associated Frequentist group sequential methods (O'Brien & Fleming 1979, Pocock 1983, Freedman et al. 1983). It has even been suggested that we should test for one chosen arm being better than the other in interim analyses but then only test for the opposite effect in the final analysis (Chi et al. 1986). The rationale seems to be that one might want to stop the trial as soon as possible if the new treatment is worse than the standard treatment or placebo, but continue the trial if the new treatment is better, in order to look at its safety. (At least, this is the way the argument is presented in English. The algebra in (Chi et al. 1986) reads the other way around — that one only stops the trial early if the new treatment is better than the standard treatment — but I presume this is a typographical error.) This suggestion requires yet another Frequentist group sequential method. Similar choices must be made in a Bayesian analysis, but for a Bayesian they need not be made once and for all; different approaches can be tried, without worrying about the effect of such extra analyses on the type I error.
- Rubinstein and Gail (1982) suggest that data which accrue after the trial has been

A problem which has been faced by all these Frequentist attempts at a stopping rule is that the sample space at the ith interim analysis is not one-dimensional: it is the space of vectors consisting of the main response variable evaluated at each interim analysis to date. (I gave an example of such a vector in chapter 7.) The necessity of averaging over more extreme data which could have occurred (see chapter 4) means that this i-dimensional space must have an ordering imposed on it. The most radical attempt to solve this problem within a Frequentist framework would be to simplify the sample space by taking it to consist of the main response variable at the end of the trial plus the time at which the trial was stopped. But even with such an extreme simplification, there seems to be no natural ordering for this pair of numbers, and nor will any ordering give confidence intervals which are always consistent with P-values from the same trial. This consideration is what makes the problem of multiplicity even worse in sequential analysis than it is in other domains of Frequentist inference.

---

formally stopped (as some data almost always do, because of delays in postage and so on) should be included in the analysis. Such a move invalidates all the above Frequentist methods, which are not specifically designed to take such data into account. (It makes no difference to a method compatible with the likelihood principle.)

- Falissard and Lellouch (1991) suggest stopping a trial early only if a series of r interim analyses all give unadjusted P-values under 5%, where r is an arbitrary number (typically set to 2 or 3). This rule has mathematical advantages: in particular, it avoids the possibility of the adjusted P-values becoming more and more significant even while the unadjusted P-values become less and less significant, a problem which affects all other Frequentist methods. However, Falissard and Lellouch's plan has severe practical disadvantages. Under their rule, a scientist who is planning seven analyses would be forbidden to stop the trial at the first or second analysis, no matter what the sample size was and no matter how many trial subjects had been killed. This nicely illustrates the desperate measures which some statisticians have felt forced to advocate in struggling with the problems of Frequentist sequential analysis.
- Koepcke (1989) recommends half of an O'Brien and Fleming stopping rule combined with half of a Pocock stopping rule. There is nothing particularly interesting about this suggestion except that it illustrates the increasingly obviously ad hoc nature of the Frequentist stopping rule enterprise.

This list includes only about a tenth of the Frequentist group sequential methods which have been advocated to date.

To solve this problem without abandoning Frequentism, one must find some natural way of reducing the number of dimensions of the ordering of the sample space. Jennison and Turnbull (1984) do this by stipulating (without justification) that the confidence intervals are to be symmetrical and centred on the unadjusted point estimate at each interim analysis. However, it is clear that this point estimate will be biased by the stopping rule, and Hughes and Pocock (1988) show this to be a severe problem in particular, realistic cases. Moreover, because Jennison and Turnbull's point estimates are naive (unadjusted), the only way to secure control over the confidence interval error rates is to force a large increase in the width of the intervals, leading to some very counterintuitive results. The width of the intervals is of more concern than the bias to those of us who do not consider bias to be a problem (see chapter 11). Primarily for these reasons, when Jennison and Turnbull presented their work to the Royal Statistical Society in 1989 they attracted a lot of criticism.

The good news is that an ordering of the sample space is only necessary so long as we insist on fixing the coverage probability averaged over the sample space. Any procedure which follows the likelihood principle avoids this problem entirely, since unobserved points in the sample space are no longer relevant.

Thus, there are four options:

(1) Specify a complete ordering of the sample space.

(2) Use very wide confidence intervals centred on estimates which may be wildly biased.

(3) Reduce the number of dimensions of the problem in some other way — but a plausible way has not been found.

(4) Refuse to average over the sample space, in accordance with the likelihood principle.

I now turn to what happens if we take option 4.

A SUBJECTIVE BAYESIAN SOLUTION

In the 1950s and 1960s, Anscombe suggested separating the stopping rule from the analysis, on the grounds that the stopping rule ought to be sequential but the analysis oughtn't (Anscombe 1954, Anscombe 1963). This idea is fundamentally incompatible with Frequentist reasoning, since the overall type I error of a trial analysed in this way could not be fixed at any particular value. However, it was revived in 1983 by Dupont, with the following motivation:

> it is hard to see why decisions that would have been made in response to outcomes that did not occur should have any bearing on the strength of evidence that can be attributed to the results that were actually observed.
>
> (Dupont 1983, p. 3)

This is of course a statement of the likelihood principle (at least, on a loose interpretation of "decisions" it is).[135]

The first suggestions of a fully Bayesian solution to this problem were published, by Anscombe (1954) and Cornfield (1966, 1976), substantially before it became obvious that Frequentist solutions were unacceptably ad hoc. Neither Anscombe nor Cornfield described the Bayesian solution in

135. There is an unsatisfying inconsistency in Dupont's approach if the P-values in both the sequential interim analyses and the non-sequential final analysis are given the same interpretation; so Dupont suggests that the non-sequential P-value should be read as an approximation to the likelihood ratio (whereas Anscombe preferred to forswear P-values altogether). This problem is not relevant to my own proposed Bayesian analysis, so we need not consider it further.

any detail, probably because the details of a solution are so obvious to a Bayesian, and so mathematically simple, that they felt there was no need to do so. But of course a statistical method with no published details was not going to be widely adopted, so in 1985 Berry published for the first time the mechanics of an explicit Bayesian method for analysing group sequential trials (Berry 1985, Berry 1987, Berger & Berry 1988, Berry 1991). In such a system, all of the problems which plague sequential methods disappear: at any stage in a trial, point and interval estimation are easy to do, need no adjustment, and are automatically compatible with each other and easy to interpret.

The simplest Bayesian method for analysing a group sequential trial is as follows. $2n$ subjects enter a trial with two treatment arms (one of which may be a placebo). The primary outcome is represented by a real-valued variable, and the unknown true difference in outcome between the two treatments is represented by δ.[136]

Suppose further, without much loss of generality (Pocock 1977), that the differences in paired samples from the two arms of the trial are Normally (Gaussianly) distributed with variance σ^2. Our task, given n and σ, is to estimate δ.

Let Υ_t be the mean treatment difference in block t (the block of subjects recruited between analysis t and analysis $t-1$). Then the Υ_t are independently distributed as $N(\delta, \sigma^2 T / n)$.

A prior probability distribution $p(\theta)$ is ascertained. For simplicity, we can consider this distribution to be $N(\mu_0, \sigma_0^2)$. Nothing much rests

136. This can be given a population interpretation, as the difference in outcome which would be seen in the population from which the samples were drawn if all members of the population were given the treatment, or it can be given a subjective interpretation. Which interpretation it should be given is an interesting issue, but the choice does not affect anything I have to say.

on the shape of the distribution (provided that it is mathematically well behaved); only its low-order moments (mean, standard deviation and skew) have much effect on the results, and it is reasonable in most trials to set the skew to zero. Moreover, since we are considering a clinical trial which aims to convince regulatory authorities to overcome a natural conservatism, it makes sense for μ_0, the mode (most probable value) of the prior probability function, to be set at 0 (no treatment difference), so that the prior becomes $N(0, \sigma_0^2)$. Strictly speaking this is sullying the Subjective Bayesian method with an element which is supposedly set according to the belief state of a doxastic agent but actually set with its effect in influencing another doxastic agent in mind. The reason that the epistemic sleight-of-hand involved in setting μ_0 to zero does not bother me is that it is almost always the case that the two treatments are believed by all the doxastic agents involved to be approximately equivalent. This is because an agent who thought otherwise would find the trial unethical. (See the discussion of equipoise above.) If this argument fails in a particular case then a non-zero value of μ_0 can be worked into the mathematics below without difficulty.

At each analysis, we construct the statistic

$$S_t = \sum_{i=1}^{t} \frac{Y_i}{t + k}$$

where $k \equiv \frac{1}{n}\sqrt{\frac{\sigma_0}{\sigma}}$.

S_t is our point estimate of δ. It is the mean of the Bayesian posterior distribution $N(\delta, \sigma^2 / n_t)$, where $2n_t$ is the number of subjects seen to date. (This follows from the definition of Bayesianism which I gave in chapter 3.)

We can also construct Bayesian credible intervals for δ. For example, a 95% credible interval is

$$CI_{95\%}(S_t) = S_t \pm 1.96\,\sigma\sqrt{\frac{t+k}{t}}.$$

Again, this follows from the definitions in chapter 3.

The most natural Bayesian stopping rule is to use a two-sided test which stops the trial if $CI_{95\%}(S_t)$ excludes 0; in other words, if

$$\left|\sum_{i=1}^{t} Y_i\right| > 1.96\,\sigma\sqrt{\frac{t+k}{t}}.$$

Since this is a Bayesian method and therefore one which observes the likelihood principle, no adjustment is made for the number of analyses.

This simple Bayesian method can easily be generalised to more complicated trials such as those with more than two arms, if necessary; but many complications can be ignored.[137]

This Subjective Bayesian solution is of course compatible with the likelihood principle, and hence with the likelihood principle. It does not suffer from any of the ad hocness of the Frequentist solutions, but it has not been acceptable to regulatory bodies for two reasons:

1. its subjective nature is fundamentally unacceptable to public regulators (whether or not it ought to be);
2. its Frequentist error rates have not been known until recently.

137. For example, the simple procedure gives approximately the same results as the obvious generalisations even when the group sizes are unequal, up to about a 20% difference in size, and even when adjustments are made for the differing health states of individuals ("prognostic factors") (Jennison & Turnbull 1989).

A MORE OBJECTIVE SOLUTION

A breakthrough in the acceptability of Bayesian methods came in 1988, when Hughes and Pocock, luminaries of the Frequentist school which I considered in the historical section above, embraced a Bayesian method. Hughes and Pocock recommended using a fixed, relatively objective prior distribution (Hughes & Pocock 1988, Pocock & Hughes 1989, Pocock & Hughes 1990). I will not consider their method in detail, because it suffers from a major flaw: although it uses a Bayesian to calculate its results, it uses Frequentist considerations to choose the stopping rule. It is thus open to objections from both Frequentists and Bayesians.

Until 1989, despite previous interest in Bayesian methods of parameter estimation, there was nothing to show whether a Bayesian group sequential method could have reasonable error probabilities. Freedman and Spiegelhalter (1989) took the first step towards finding this out by showing that for certain reasonable priors[138] a Bayesian trial would have a stopping rule very similar to those in common use — specifically, those of Pocock and O'Brien and Fleming. My own work has made these results more precise, as I describe below.

Only a couple of other non-subjective Bayesian solutions to the group sequential problem have been proposed. Mehta and Cain (1984) and Goldman (1987) consider using a flat (improper) prior probability function; but this gives an extremely radical stopping rule (one which stops the trial extremely easily), and inherits the problems of improper priors which I discussed in earlier chapters. Gharraf and Al-Nassar (1990) propose a

138. similar to those proposed by McPherson (1982) and Hughes and Pocock (1988)

prior which distributes the probability on just two points in the hypothesis space; they do not justify this choice.

The relatively objective procedure which I propose is similar, mathematically, to the Subjective Bayesian procedure discussed in the previous section. The details are as follows.

Up to $2n$ subjects enter a trial with two treatment arms (one of which may be a placebo) in up to T groups, with an analysis planned after each group. (The number of subjects receiving each treatment is then n / T per group.) As before, the true difference in primary outcome between the two treatments is represented by δ. Again, let Y_t be the mean treatment difference in block t. Then the Y_t are independently distributed as $N(\delta, \sigma^2 T / n)$.

Unlike the Subjective Bayesian method, the objective method sets a prior distribution in terms of a *handicap* which the data must overcome in order to overturn the null hypothesis. This handicap is mathematically equivalent to a set of $f \times n$ outcomes distributed according to the null hypothesis, for some f yet to be determined. (f is no longer defined as $\frac{1}{n}\sqrt{\frac{\sigma_0}{\sigma}}$ as it was in the Subjective Bayesian analysis.) This yields the prior probability function of $N(0, \sigma^2 / f.n)$.

Now I play a dirty trick. I assert, without detailed justification, that $\frac{1}{3}$ is a reasonable value of f. This value can be supported in three ways:

1. It is, approximately, the value given by clinicians asked what degree of conservatism in favour of the null hypothesis is required in particular trials (Freedman et al. 1983).

2. It is, very approximately, the degree of conservatism in favour of the null hypothesis required by regulatory authorities (in so far as we can estimate such a thing).
3. It gives the objective likelihood method which I am proposing here excellent Frequentist properties. Although the Frequentist properties of a procedure are not important from the point of view of a rational doxastic agent's own belief revisions (or so I have argued), they are important in gaining acceptance for a procedure in a predominantly Frequentist world. Taking this issue seriously is part of my claim that this case study is a *realistic* application of the likelihood principle.

I do not claim that any of these reasons for choosing $f = \frac{1}{3}$ is conclusive, only that they make the choice plausible and not entirely ad hoc.

At each analysis, we construct the statistic

$$S_t = \sum_{i=1}^{t} \frac{Y_i}{t + fT}.$$

S_t is our point estimate of δ. It is the mean of the Bayesian posterior distribution $N(\delta, \sigma^2 / n_t)$, where $2n_t$ is the number of subjects seen to date.

Just as in the Subjective Bayesian case, we can construct Bayesian credible intervals for δ. A 95% credible interval is

$$CI_{95\%}(S_t) = S_t \pm 1.96\, \sigma \sqrt{\frac{t + fT}{t}}.$$

And just as before, the most natural Bayesian stopping rule is to use a two-sided test which stops the trial if $CI_{95\%}(S_t)$ excludes 0; in other words, if

$$\left| \sum_{i=1}^{t} Y_i \right| > 1.96\,\sigma \sqrt{\frac{t+fT}{t}}.$$

If the amount of conservatism, f, is chosen so as to fix the Frequentist error rates at some prespecified levels, the procedure contravenes the likelihood principle, as is inevitable for any general Bayesian procedure constrained in such a way (Sweeting 2001, p. 658). If, however, it is chosen so as to include a reasonable amount of conservatism, it abides by the likelihood principle but still has excellent Frequentist error rates, as my colleagues and I show in (Grossman et al. 1994). We show there that the procedure given above has a type I error between $2\frac{1}{2}\%$ and 5% for any number of interim analyses between 0 and 10. We also show that its stopping rule is similar numerically to the Frequentist rules most commonly used (those of O'Brien and Fleming (1979) and Pocock (1997)). More surprisingly, we show that our method has a lower expected sample size than the standard Frequentist methods in many cases, as the following table shows.

power	Grossman et al	Armitage, McP. & R.	O'Brien & Fleming
50%	16.0	14.3	14.7
75%	23.4	23.4	22.8
90%	28.8	31.5	29.1
95%	31.3	36.2	32.5

Table 6

Expected sample size for a trial with 4 interim analyses

To interpret this table, multiply each entry by σ^2 / δ^2, where δ is an estimate of the treatment difference between groups.[139] "Armitage, McP. &R." stands for the method of (Armitage et al. 1969), which is essentially the same as the method of Pocock (1997); "O'Brien & Fleming" stands for the method of (O'Brien & Fleming 1979). The expected sample sizes of the Armitage et al. and O'Brien and Fleming methods are taken from (Geller & Pocock 1987).

It is possible to investigate the expected sample size of the new method in more detail. Pocock (1982) has calculated the smallest sample size attainable for a given power. There is no such thing as a sequential procedure which is optimal (in this sense) at *every* power; but one might hope for a procedure which is optimal for a reasonable power (not too high, since very high powers require unfeasibly large study sizes, and not too low, since Frequentist audiences find low-powered trials unconvincing).

Table 7, below, gives the cut-off to which P-values are compared in various published procedures and compares them to the optimal levels for various powers as given in (Pocock 1982). It shows that the new design is remarkably close to optimal for a power of between 75% and 80%. For my method, the values tabulated are not strictly P-values, since it is not a Frequentist method; but they are sufficiently cognate to P-values that a power calculation based on them is correct. For details see (Grossman 1993, Grossman et al. 1994).

---

139. The power of the procedures is defined in terms of such an estimate of treatment difference. Recall that the power of a procedure is $1 - \beta$, where β is the type II error rate. This rate depends on the values of the unknown parameters, as explained in chapter 4. Factualists do not care about the power of a test, of course, since the power is an average over the sample space.

No. of analyses	Analysis	Armitage, McP. & R.	O'Brien & Fleming	Grossman et al	Optimal for power of 75%	Optimal for power of 80%
2	1st	0.029	0.005	0.024	0.023	0.025
2	2nd	0.029	0.048	0.035	0.036	0.034
3	1st	0.022	0.0005	0.011	0.012	0.014
3	2nd	0.022	0.014	0.024	0.021	0.021
3	3rd	0.022	0.045	0.031	0.033	0.030
4	1st	0.018	0.0001	0.006	0.006	0.008
4	2nd	0.018	0.004	0.016	0.016	0.017
4	3rd	0.018	0.019	0.024	0.020	0.020
4	4th	0.018	0.043	0.029	0.032	0.029
5	1st	0.016	0.00001	0.003	0.003	0.004
5	2nd	0.016	0.0013	0.011	0.011	0.013
5	3rd	0.016	0.008	0.018	0.016	0.017
5	4th	0.016	0.023	0.023	0.019	0.018
5	5th	0.016	0.041	0.027	0.031	0.028

Table 7

P-value cut-offs compared to optimal cut-offs

The fact that the new method is approximately optimal in the sense of Table 7 means that it has approximately the lowest possible expected sample size for a given power. Lowering expected sample sizes not only saves money, it also protects trial subjects from receiving inferior treatments and gets good drugs to market earlier. It saves lives.

Having said that, I have only shown that the new method saves lives, on average, for a given power; and I know of no cogent reason for a non-Frequentist to wish to assign or fix a given power. So the small expected

sample size of this procedure ought to be convincing to Frequentists only. Non-Frequentists, on the other hand, may not see it as optimal in any sense but *will* see my method as better than any Frequentist method, provided they accept the likelihood principle.

It is possible that likelihood-principle-based methods which are better than my own will appear in the future. I do not claim that my method is the best possible. I only claim that it is better in every way than any method so far available in the literature, including any Frequentist method.

What has this case study shown about the likelihood principle? It has shown that in a practical, important, non-toy case study the likelihood principle, despite being incompatible with standard methods, is compatible with what turns out to be the best method available so far. This speaks in its favour.

2. GENERAL CONCLUSIONS

I will now sum up the most important conclusions from the thesis as a whole.

I started this thesis by claiming that the study of the philosophy of statistics (and hence, derivatively, the philosophy of most of the special sciences) could be clarified tremendously by analyses of inference procedures. I promised that I would delineate a clear, precise class of cases of statistical inference in which Frequentist error rates are irrelevant. I have done that. By analysing inference procedures, and especially by considering how inference procedures are evaluated, I have shown that — subject to caveats which I trust I have made clearer than they have been made

by other authors — the sample space of a merriment, and hence any Frequentist error rate, is irrelevant to the conclusions which should be drawn from that merriment about a hypothesis space. This, when the caveats are spelled out, is what I have called the likelihood principle. Indeed, I have shown that a merriment need not even have a sample space, and hence that accidental observations can be analysed in the same way as designed experiments.

I have shown, in a case study, that the likelihood principle can lead us to statistical inference procedures which are better (in every sense, even, sometimes, the Frequentist sense) than standard non-likelihood procedures, in realistic (non-toy) situations.

The rest of my conclusions are somewhat negative, because the direct consequences of the likelihood principle is entirely negative. By ruling out certain inference procedures, it tells us what *not* to do. As Basu puts this point,

> It is best to look upon [the likelihood principle] as a sort of code of conduct that ought to guide us in our inference making behaviour. In this respect it is analogous to the unwritten medical code that . . . disallows a Doctor to include a symmetric die or a table of random numbers as a part of his diagnostic gadgets.
>
> (Basu 1975, p. 22)[140]

140. Alan Hájek has pointed out to me that the current unwritten code of conduct for statisticians, if not for doctors, *does* in fact allow them to do something very similar to throwing a die as part of a diagnosis. The procedure in question is one which is used when it is desired to get a certain significance level (say, 5%). If the P-value of a given experiment is bound to be either strictly greater than or strictly less than 5%, as is sometimes the case, a random number generator is used to determine which of the possible levels to use as the level at which significance will be proclaimed. For example, if the possible outcomes include P-values of 4% and 6% but not 5%, one might randomly, according to the toss of a coin, use the 4% level as one's cut-off for significance half of the time and use 6% the other half of the time. However, I think it is uncontroversial among philosophers that such a procedure is irrational. If not, I will have to assert that the likelihood principle is even *less* rational than such a procedure.

My main negative conclusions about the consequences of the work presented in this thesis come in the form of two consequences of the likelihood principle, as follows:

1. The likelihood principle invalidates almost all Frequentist methods of applied statistics at least mildly.
2. The likelihood principle grossly invalidates some Frequentist methods of applied statistics.

1. MILDLY INVALIDATING ALMOST ALL FREQUENTIST METHODS

> With [the likelihood principle] as the guiding principle of data analysis, it no longer makes any sense to investigate (at the data analysis stage) the 'bias' and 'standard error' of point estimates, the probabilities of the 'two kinds of error' for a test, the 'confidence-coefficients' associated with interval estimates, or the 'risk functions' associated with rules of decision making.
>
> (Basu 1975, p. 16)

Basu's claim amounts to the assertion that the likelihood principle invalidates the criteria by which Frequentist procedures are selected. I will argue here for an even stronger version of this claim: that the likelihood principle alone is enough to invalidate those criteria, even though the factual principle is weaker than the likelihood principle. I will take each of the criteria which Basu mentions in turn.

I have argued at length against caring about the bias of estimates in chapter 11. Part of my argument was that bias is something which the likelihood principle cautions us to avoid in statistical inference, since it depends on averages over the sample space. (This was a rather small part of my argument, since it is obvious; I spent more time establishing independent

reasons for being wary of bias.) As we saw, the bulk of authors continue to care about bias not because they have any positive arguments in favour of doing so but because it helps to reduce the otherwise unmanageable number of Frequentist procedures available for most estimation problems. But the likelihood principle helps with this problem too: by saying that *none* of these methods is coherent, it leaves us with a smaller number of non-Frequentist methods from which to choose.

The standard error of an estimator is a measure of how much that estimator is expected to vary on repeated applications of a procedure. It is clearly something else which we should not use in statistical inference, according to the likelihood principle, because it requires averages over the sample space (as does any measure which depends on taking averages over imaginary repeated applications of a procedure).

The claim that "it no longer makes any sense to investigate . . . the probabilities of the 'two kinds of error' for a test" formed the main argument of the second part of chapter 7.

The claim that "it no longer makes any sense to investigate . . . the 'confidence-coefficients' associated with interval estimates" formed the main argument of the third part of chapter 7.

Basu's final claim, that "it no longer makes any sense to investigate . . . the 'risk functions' associated with rules of decision making", is outside the scope of this thesis, since a risk function is a type of utility function, something which statistical inference per se may not have available to it. However, it is, at least, compatible with my claims, since some forms of decision theory which are compatible with the principle, including standard

Bayesian decision theory, have no use for risk functions (Raiffa & Schlaifer 2000).

2. GROSSLY INVALIDATING SOME FREQUENTIST METHODS

The claim that the likelihood principle grossly invalidates *some* Frequentist methods can be proved by example. I gave an example in chapter 7, where we met a Frequentist 75% confidence interval which we could be certain contained the true value of the parameter (the height of a bonobo chimpanzee). I gave another, more detailed example in the sequential trials case study above, where we saw that Frequentist methods with equal plausibility are drastically at odds with each other. We saw that factualist analyses avoid both the epistemic incoherence of the bonobo's confidence interval and the ad hockery of the case study. Hence, if the likelihood principle is right, some of Frequentist applied statistics is grossly wrong.

FINAL CONCLUSIONS

I have argued that we should accept the likelihood principle, but I have come to no firm conclusions about how it should be applied (as opposed to how it should not be applied).

As we have seen, the only likelihood method which has yet been worked out in detail and shown to be applicable to a wide variety of problems is Bayesianism. Bayesianism appears to most people to be necessarily more subjective than the standard methods, and consequently the standard methods are in no danger of disappearing quickly. But there are some contexts in which we don't have to wait for complete agreement on the

subjectivity of Bayesian methods. Areas in which likelihood methods could replace non-likelihood methods with no loss of objectivity include:

- problems in which the priors are so important and so obviously subjective that they can and should be provided separately by each decision participant, leaving the statistician free to publish only the likelihood function;
- problems in which inferences can be drawn from raw likelihood functions or raw likelihood ratios;
- problems in which the priors correspond to empirically known frequencies — these are fairly common, covering, for example, almost all of clinical epidemiology, as I showed in chapter 3.

This thesis has argued that the likelihood principle should be applied in all areas of statistical inference. We now see that it can be applied uncontentiously in at least some areas. The exact extent of these areas is a matter for further research.

REFERENCES

Aickin, Mikel (2000) "Connecting Dempster-Shafer belief functions with likelihood-based inference". *Synthese*, 123:347–364.

Anscombe, F. J. (1954) "Fixed-sample-size analysis of sequential observations". *Biometrics*, 10:89–100.

——— (1963) "Sequential medical trials [review]". *Journal of the American Statistical Association*, 58:365–383.

Armitage, Peter (1960) *Sequential medical trials.* Oxford: Blackwell.

——— (1989) "Inference and decision in clinical trials". *Journal of Clinical Epidemiology*, pp. 42: 293–299.

Armitage, Peter & Geoffrey Berry (1994) *Statistical Methods in Medical Research.* Oxford: Blackwell, 3rd edition.

Armitage, Peter, Geoffrey Berry & J. N. S. Matthews (2002) *Statistical Methods in Medical Research.* Oxford: Blackwell, 4th edition.

Armitage, Peter, C. K. McPherson & C. Rowe (1969) "Repeated significance tests on accumulating data". *Journal of the Royal Statistical Society, Series A*, 132:235–244.

Backe, Andrew (1999) "The likelihood principle and the reliability of experiments". *Philosophy of Science*, 66:S354–S361.

Ball, Duncan (2001) *Selby's Selection.* Sydney: Harper Collins.

Barnard, G. A. (1947) "Review of Wald's 'Sequential analysis'". *Journal of the American Statistical Association*, 42:658–669.

——— (1962) "Discussion of 'On the foundations of statistical inference' by Allan Birnbaum". *Journal of the American Statistical Association*, 57(298):308–309.

Barnard, G. A., G. M. Jenkins & C. B. Winsten (1962) "Likelihood inference and time series". *Journal of the Royal Statistical Society Series A*, 125:321–372.

Barnard, George (1985) "Discussion of 'In defense of the likelihood principle: axiomatics and coherency'". In J. M. Bernardo, M. H. DeGroot, D. V. Lindley & A. F. M. Smith (eds.), *Bayesian Statistics*, volume 2, pp. 57–60. Elsevier.

Barndorff-Nielsen, O. (1976) "Plausibility inference". *Journal of the Royal Statistical Society Series B*, 38(2):103–131.

Barnett, Vic (1999) *Comparative Statistical Inference*. New York: John Wiley, 3rd edition.

Basu, Debabrata (1975) "Statistical information and likelihood (with discussion)". *Sankhyā Series A*, 37:1–71.

Bayarri, M. J., M. H. DeGroot & J. B. Kadane (1987) "What is the likelihood function? (with discussion)". In Shanti S. Gupta & James O. Berger (eds.), *Statistical decision theory and related topics. IV. Vol. 1.*, pp. 3–27. New York: Springer-Verlag.

Bayes, Thomas (1763) "An essay towards solving a problem in the doctrine of chances". *Philosophical Transactions of the Royal Society of London*, 53:370–418.

Berger, J. O. (1985) "In defense of the likelihood principle: axiomatics and coherency". In J. M. Bernardo, M. H. DeGroot, D. V. Lindley & A. F. M. Smith (eds.), *Bayesian Statistics*, volume 2, pp. 33–66. Elsevier.

Berger, James (1993) *An Overview of Robust Bayesian Analysis. Technical Report 93-53C*, Purdue University.

Berger, James O. (1980) *Statistical Decision Theory: Foundations, Concepts, and Methods*. New York: Springer–Verlag.

Berger, James O. & Donald A. Berry (1988) "The relevance of stopping rules in statistical inference (with discussion)". In S. S. Gupta & J. O. Berger (eds.), *Statistical decision theory and related topics*, volume 1, pp. 29–72. New York: Springer-Verlag.

Berger, James O. & Thomas Sellke (1987) "Testing a point null hypothesis: the irreconcilability of p values and evidence". *Journal of the American Statistical Association*, 82(397):112–122.

Berger, James O. & Robert L. Wolpert (1984) *The Likelihood Principle*. Hayward, California: Institute of Mathematical Statistics, 1st edition.

——— (1988) *The Likelihood Principle*. Hayward, California: Institute of Mathematical Statistics, 2nd edition.

Berliner, Mark (1987) "Discussion of 'What is the likelihood function?'". In Shanti S. Gupta & James O. Berger (eds.), *Statistical decision theory and related topics. IV. Vol. 1.*, pp. 17–20. New York: Springer-Verlag.

Bernardinelli, Luisa & Cristina Montomoli (1992) "Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk". *Statistics in Medicine*, 11:983–1007.

Bernardo, José M. & Adrian F. M. Smith (1994) *Bayesian Theory*. Chichester: John Wiley.

Berry, DA (1985) "Interim analyses in clinical trials: classical vs. Bayesian approaches". *Statistics in Medicine*, 4:521–526.

——— (1987) "Interim analysis in clinical trials: the role of the likelihood principle". *The American Statistician*, 41:117–122.

Berry, Donald A. (1991) "A case for Bayesianism in clinical trials". .

Birnbaum, Allan (1962) "On the foundations of statistical inference". *Journal of the American Statistical Association*, 57(298):269–306.

——— (1970) "On Durbin's modified principle of conditionality". *Journal of the American Statistical Association*, 65(329):402.

——— (1972) "More on concepts of statistical evidence". *JASA*, 67:858–861.

Bjørnstad, Jan F. (1996) "On the generalization of the likelihood function and the likelihood principle". *Journal of the American Statistical Association*, 91(434):791–806.

Boik, Robert J. (2004) "Why likelihood? commentary". In Mark L. Taper & Subhash R. Lele (eds.), *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, pp. 167–180. Chicago and London: University of Chicago Press.

Breslow, Norman (1990) "Biostatistics and Bayes". *Statistical Science*, 5:269–298.

Bross, I. (1952) "Sequential medical plans". *Biometrics*, pp. 8: 188–205.

Butler, R. W. (1987) "A likely answer to 'What is the likelihood function?'". In Shanti S. Gupta & James O. Berger (eds.), *Statistical decision theory and related topics. IV. Vol. 1.*, pp. 21–26. New York: Springer-Verlag.

Casella, George & Roger L. Berger (1987) "Reconciling Bayesian and Frequentist evidence in the one-sided testing problem". *Journal of the American Statistical Association*, 82:106–111.

——— (2002) *Statistical Inference.* Pacific Grove: Duxbury, 2nd edition.

Chi, P. Y., R. Bristol & J. V. Castellana (1986) "A clinical trial with an interim analysis". *Statistics in Medicine*, 5:387–392.

Cornfield, J. (1966) "Sequential trials, sequential analysis and the likelihood principle". *American Statistician*, 20:18–23.

——— (1976) "Recent methodological contributions to clinical trials". *American Journal of Epidemiology*, 104:408–421.

Cox, D. R. (1958) "Some problems connected with statistical inference". *The Annals of Mathematical Statistics*, 29(2):357–372.

Cutler, S. J., S. W. Greenhouse, J. Cornfield & M. A. Schneiderman (1966) "The role of hypothesis testing in clinical trials". *Journal of Chronic Disease*, 19:857–882.

Dawid, A. P. (1977) "Conformity of inference patterns". In J. R. Barra et al (ed.), *Recent Developments in Statistics*, pp. 245–256. Amsterdam: North-Holland.

——— (1986) "Discussion of 'On principles and arguments to likelihood' by M. J. Evans, D. A. S. Fraser and G. Monette". *The Canadian Journal of Statistics*, 14(3):196–197.

de Finetti, Bruno (1972) *Probability, Induction and Statistics: The art of guessing*. London: Wiley.

——— (1980) "On the condition of partial exchangeability". In R. C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*, volume II, pp. 193–205. Berkeley: University of California Press.

Deely, J. J. & D. V. Lindley (1981) "Bayes empirical Bayes". *Journal of the American Statistical Association*, 76:833–841.

DeMets, D. L. & J. H. Ware (1980) "Group sequential methods for clinical trials with a one-sided hypothesis". *Biometrika*, 67:651–60.

Diaconis, P. & D. Freedman (1980) "Generalizations of exchangeability". In R. C. Jeffrey (ed.), *Studies in Inductive Logic and Probability II*, volume II, pp. 234–249. Berkeley: University of California Press.

Dupont, W. D. (1984) "Bias, maximum likelihood estimators, and sequential clinical trials [letter]". *Controlled Clinical Trials*, 5:275–277.

Dupont, WD (1983) "Sequential stopping rules and sequentially adjusted p values: does one require the other?" *Controlled Clinical Trials*, 4:3–10.

Durbin, J. (1970) "On Birnbaum's theorem on the relation between sufficiency, conditionality and likelihood". *Journal of the American Statistical Association*, 65(329):395–398.

Earman, John (1992) *Bayes or Bust?*. Cambridge, Massachusetts: MIT Press.

Earman, John, John Roberts & Sheldon Smith (2002) "Ceteris paribus lost". *Erkenntnis*, 57:281–301.

Edwards, A. W. F. (1972) *Likelihood*. London: Cambridge University Press.

Edwards, W., H. Lindman & L. J. Savage (1963) "Bayesian statistical inference for psychological research". *Psychological Review*, 70:193–242.

Elfring, G. L. & J. R. Schultz (1973) "Group sequential designs for clinical trials". *Biometrics*, 29:471–477.

Emerson, S. S. & T. R. Fleming (1989) "Symmetric group sequential test designs". *Biometrics*, 45:905–923.

Evans, M., D. A. S. Fraser & G. Monette (1986) "On principles and arguments to likelihood". *Canadian Journal of Statistics*, 14:181–199.

Falissard, B. & J. Lellouch (1991) "Some extensions to a new approach for interim analysis in clinical trials". *Statistics in Medicine*, 10:949–957.

Falk, Raphael (1986) "What is a gene?" *Studies in the History and Philosophy of Science*, 17(2):133–173.

Feyerabend, Paul (1993) *Against Method*. London and New York: Verso, 3rd edition.

Fienberg, Stephen E. (2006) "When did Bayesian inference become "Bayesian"?" *Bayesian Analysis*, 1(1):1–40.

Fisher, R. A. (1921) "On the probable error of a coefficient of correlation deduced from a small sample". *Metron*, 1:3–32.

——— (1930) "Inverse probability". *Proceedings of the Cambridge Philosophical Society*, 26:528–535.

——— (1973) *Statistical Methods and Scientific Inference*. New York: Hafner Press, 3rd edition.

Fitelson, Branden (2001) *Studies in Bayesian Confirmation Theory*. Ph.D. thesis, University of Wisconsin–Madison.

Forster, Malcolm & Elliott Sober (2004a) "Why likelihood?" In Mark L. Taper & Subhash R. Lele (eds.), *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, pp. 153–167. Chicago and London: University of Chicago Press.

——— (2004b) "Why likelihood? rejoinder". In Mark L. Taper & Subhash R. Lele (eds.), *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, pp. 181–190. Chicago and London: University of Chicago Press.

Forster, Malcolm R. (2002) "Predictive accuracy as an achievable goal of science". *Philosophy of Science*, 69:S124–S134.

Fraser, D. A. S. (1963) "On the sufficiency and likelihood principles". *Journal of the American Statistical Association*, 58(303):641–647.

——— (1968) *The Structure of Inference*. New York: Wiley.

——— (1996) "Some reparks on pivotal models and the fiducial argument in relation to structural models". *International Statistical Review*, 64:231–235.

Freedman, L. S., D. Lowe & P. Macaskill (1983) "Stopping rules for clinical trials". *Statistics in Medicine*, 2:167–174.

Freedman, L. S. & D. J. Spiegelhalter (1989) "Comparison of Bayesian with group sequential methods for monitoring clinical trials". *Controlled Clinical Trials*, 10:357–367.

Geller, N. J. & S. J. Pocock (1987) "Interim analyses in randomized clinical trials: ramifications and guidelines for practitioners". *Biometrics*, 43:213–223.

Gelman, Andrew, John B. Carlin, Hal S. Stern & Donald B. Rubin (1995) *Bayesian Data Analysis*. London: Chapman and Hall.

Gharraf, M. K. & S. A. Al-Nasser (1990) "A bayesian group sequential analysis procedure". *Communications in Statistics - Theory and Methods*, 19:977–985.

Gigerenzer, Gerd (1993) "The superego, the ego, and the id in statistical reasoning". In Gideon Keren & Charles Lewis (eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*, pp. 311–339. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Gigerenzer, Gerd & Daniel G. Goldstein (1996) "Reasoning the fast and frugal way: Models of bounded rationality". *Psychological Review*, 103(4):650–669.

Gigerenzer, Gerd & R. Selten (eds.) (2002) *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press.

Gigerenzer, Gerd, P. M. Todd & the ABC Research Group (eds.) (1999) *Simple Heuristics That Make Us Smart*. Oxford: Oxford University Press.

Gillies, D. A. (1973) *An Objective Theory of Probability*. London: Methuen.

Glymour, Clark (1981) *Theory and Evidence*. Chicago: University of Chicago Press.

Goldman, A. I. (1987) "Issues in designing sequential stopping rules for monitoring side effects in clinical trials". *Controlled Clinical Trials*, 8:327–337.

Goldstein, M. & J. V. Howard (1991) "A likelihood paradox, with discussion". *Journal of the Royal Statistical Society Series B*, 53(3):619–628.

Good, I. J. (1965) *The estimation of probabilities*. Cambridge, Mass.: MIT Press.

——— (1976) "The Bayesian influence, or how to sweep subjectivism under the carpet". In C.A. Hooker & W. Harper (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Sience*, volume 2, pp. 125–174. Dordrecht: D. Reidel.

——— (1981) "Some logic and history of hypothesis testing". In Joseph C. Pitt (ed.), *Philosophical Foundations of Economics*, University of Western Ontario Series on the Philosophy of Science, pp. 149–174. Dordrecht: D. Reidel.

——— (1983) *Good Thinking*. Minneapolis: University of Minnesota Press.

Goodman, Daniel (2004) "Taking the prior seriously: Bayesian analysis without subjective probability". In Mark L. Taper & Subhash R. Lele (eds.), *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, pp. 379–399. Chicago and London: University of Chicago Press.

Goodman, Nelson (1983) *Fact, Fiction and Forecast*. Cambridge, Mass.; and London: Harvard University Press, 4th edition.

Griffiths, Paul E. & Eva M. Neumann-Held (1998) "The many faces of the gene". *BioScience*.

Grossman, Jason (1993) *Unified hypothesis testing, point estimation and interval estimation for group sequential trials*. Master's thesis, University of Sydney.

——— (2005) "What's wrong with the easy route to the definition of subjective probability?" In peer review.

Grossman, Jason & Fiona J. Mackenzie (2005) "The randomised controlled trial: Gold standard, or merely standard?" *Perspectives in Biology and Medicine*, 48:forthcoming.

Grossman, Jason, M. K. B. Parmar, D. J. Spiegelhalter & L. S. Freedman (1994) "A unified method for monitoring and analysing controlled trials". *Statistics in Medicine*, 13:1815–1826.

Hacking, Ian (1965) *Logic of Statistical Inference*. London: Cambridge University Press.

Hájek, Alan (2003) "Conditional probability is the very guide of life". In Henry Kyburg & Mariam Thalos (eds.), *Probability is the Very Guide*

of Life: The Philosophical Uses of Chance, pp. 183–203. Peru, Illinois: Open Court.

Hawthorne, James (2005) "Degree-of-belief and degree-of-support: why Bayesians need both notions". *Mind*, 114(454):277–320.

Haybittle, J. L. (1971) "Repeated assessment of results in clinical trials of cancer treatment". *British Journal of Radiology*, 44:793–797.

Hilgevoord, J. & J. Uffink (1991) "Uncertainty in prediction and in inference". *Foundations of Physics*, 21:323–341.

Hill, Bruce M. (1987) "The validity of the likelihood principle". *American Statistician*, 41(2):95–100.

——— (1988) "On the validity of the likelihood principle". In S. S. Gupta & J. O. Berger (eds.), *Statistical decision theory and related topics*, volume 1. New York: Springer-Verlag.

Howson, Colin & Peter Urbach (1993) *Scientific Reasoning: The Bayesian Approach*. Open Court, 2nd edition.

Hughes, M. D. & S. J. Pocock (1988) "Stopping rules and estimation problems in clinical trials". *Statistics in Medicine*, 7:1231–1242.

Jaynes, E. T. (1968) "Prior probabilities". *IEEE Transactions on Systems Science and Cybernetics*, SSC-4:227–241.

——— (1973) "The well-posed problem". *Foundations of Physics*, 30:477–493.

——— (1983) *Papers on Probability, Statistics and Statistical Physics*. Dordrecht: D. Reidel.

Jeffreys, Harold (1931) *Theory of Probability*. Oxford: Oxford University Press, 1st edition.

——— (1961) *Theory of Probability*. Oxford: Oxford University Press, 3rd edition.

——— (1973) *Scientific Inference*. Cambridge: Cambridge University Press, 3rd edition.

Jeffreys, William H. & James O. Berger (1991) *Sharpening Ockham's Razor on a Bayesian Strop. Technical report*, Purdue University.

Jennison, C. & B. W. Turnbull (1984) "Repeated confidence intervals for group sequential clinical trials". *Controlled Clinical Trials*, 5:33–45.

——— (1989) "Interim analysis: the repeated confidence interval approach". *Journal of the RSS, Series B*, 51:305–361.

Kadane, Joseph B., Mark J. Schervish & Teddy Seidenfeld (1999) *Rethinking the Foundations of Statistics*. Cambridge Studies in Probability, Induction, and Decision Theory. Cambridge: Cambridge University Press.

Kendall, M. G. & A. Stuart (1967) *The Advanced Theory of Statistics*. London: Charles Griffin and Co.

Keynes, John Maynard (1921) *A Treatise On Probability*. London: Macmillan.

Knuth, D. E. (1968) *The Art of Computer Programming, Volume 2 (Seminumerical Algorithms)*. Reading, Massachusetts: Addison-Wesley, 1st edition.

Koepcke, W. K. (1989) "Analyses of group sequential clinical trials". *Controlled Clinical Trials*, 10:222S–230S.

Kyburg, Henry E. (1987) "Bayesian versus non-Bayesian evidential updating". *Artificial Intelligence*, 31:271–293.

Lehmann, E. L. (1959) *Testing Statistical Hypotheses.* New York: John Wiley.

Leslie, Claire F. (1998) *Lack of Confidence: A study of the suppression of certain counter-examples to the Neyman-Pearson Theory of Statistical Inference with particular reference to the Theory of Confidence Intervals.* Master's thesis, University of Melbourne.

Lewis, David (1996) "Elusive knowledge". *Australasian Journal of Philosophy*, 74:549—67.

Lewis, Simon (1995) *The Art and Science of Smalltalk.* London: Prentice Hall.

Liddle, Jeannine, Margaret Williamson & Les Irwig (1996) *Method for Evaluating Research Guideline Evidence. Technical report*, NSW Department of Health, Sydney.

Lindley, D. V. (1953) "Statistical inference". *Journal of the Royal Statistical Society Series B*, 15(1):30–76.

——— (1980) "L. J. Savage—his work in probability and stastitics". *The Annals of Statistics*, 8(1):1–24.

——— (1983) "Response to 'parametric empirical Bayes inference' by Morris". *Journal of the American Statistical Association*, p. 381?

——— (1990a) "Good's work in probability, statistics and the philosophy of science". *Journal of Statistical Planning and Inference*, 25:211–213.

Lindley, David V. (1965) *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge: Cambridge University Press.

Lindley, Dennis V. (1990b) "The 1988 Wald memorial lectures: The present position in Bayesian statistics". *Statistical Science*, 5(1):44–65.

Lipton, Peter (1993) "Is the best good enough?" *Proceedings of the Aristotelian Society*, pp. 89–104.

——— (2001) "Is explanation a guide to inference? a reply to wesley c. salmon". In G. Hon & S. Rakover (eds.), *Explanation: Theoretical Approaches and Applications*, pp. 93–120. Amsterdam: Kluwer.

——— (2005) *Inference to the Best Explanation*. Routledge, 2nd edition.

Martinsek, Adam T. (1988) "Discussion of 'The relevance of stopping rules in statistical inference' by Berger and Berry". In S. S. Gupta & J. O. Berger (eds.), *Statistical decision theory and related topics*, volume 1, pp. 57–61. New York: Springer-Verlag.

Mayo, Deborah (1996) *Error and the Growth of Experimental Knowledge*. Chicago and London: University of Chicago Press.

——— (2000) "review of the third edition of 'what is this thing called science?'". *Metascience*, 9(2):179–188.

McPherson, Klim (1982) "On choosing the number of interim analyses in clinical trials". *Statistics in Medicine*, 1:25–36.

Mehta, C. R. & K. C. Cain (1984) "Charts for the early stopping of pilot studies". *Journal of Clinical Oncology*, 2:676–692.

Meier, P. (1981) "Jerome Cornfield and the methodology of clinical trials". *Controlled Clinical Trials*, 1:339–345.

Meyer, R. K. & M. A. McRobbie (1982) "Multisets and relevant implication". *Australasian Journal of Philosophy*, 60(2):107–139.

Miller, Richard W. (1987) *Fact and method: Explanation, confirmation and reality in the natural and the social sciences*. Princeton, New Jersey: Princeton University Press.

Mood, A. M. (1950) *Introduction to the Theory of Statistics*. New York: McGraw-Hill.

Moore, Alison & Jason Grossman (2003) "Exploring the meaning of 'assignments of meaning in epidemiology'". In Ian Kerridge, Chris Jordens & Emma-Jane Sayers (eds.), *Restoring Humane Values to Medicine: a Miles Little reader*, pp. 99–102. Annandale: Desert Pea Press.

Morris, Carl N. (1983) "Parametric empirical Bayes inference: theory and applications". *Journal of the American Statistical Association*, 78:47–65.

Neyman, Jerzy (1937) "Outline of a theory of statistical estimation based on the classical theory of probability". *Philosophical Transactions of the Royal Society of London, Series A*, 236(767):333–380.

——— (1967) "Outline of a theory of statistical estimation based on the classical theory of probability". In *A Selection of Early Statistical Papers of J. Neyman*. Berkeley and Los Angeles: University of California Press.

O'Brien, Peter & Thomas Fleming (1979) "A multiple testing procedure for clinical trials". *Biometrics*, 35:549–556.

O'Hagan, Anthony (1994) *Kendall's Advanced Theory of Statistics Vol. 2B: Bayesian Statistics*, volume 2B. London: Edward Arnold.

Peto, R., M. C. Pike, P. Armitage, N. E. Breslow, D. R. Cox, S. V. Howard, N. Mantel, K. McPherson, J. Peto & P. G. Smith (1976) "Design and analysis of randomised clinical trials requiring prolonged observations of each patient. i. Introduction and design". *British Journal of Cancer*, 34:585–612.

Pocock, S. J. (1977) "Group sequential methods in the design and analysis of clinical trials". *Biometrika*, 64:191–9.

——— (1978) "Size of cancer clincal trials and stopping rules". *British Journal of Cancer*, 38:757–766.

——— (1982) "Interim analyses for randomized clinical trials: the group sequential approach". *Biometrics*, 38:153–162.

——— (1983) "Clinical trials: a practical approach". *John Wiley, Chichester*, pp. John Wiley, Chichester.

Pocock, S. J. & M. D. Hughes (1989) "Practical problems in interim analyses, with particular regard to estimation". *Controlled Clinical Trials*, 10:209S–211S.

——— (1990) "Estimation issues in clinical trials and overviews". *Statistics in Medicine*, 9:657–671.

Popper, Karl (1959) *The Logic of Scientific Discovery*. London: Hutchinson.

Pratt, J. W. (1961) "Review of Lehmann's 'Testing Statistical Hypotheses'". *Journal of the American Statistical Association*, 56:163–166.

——— (1962) "Discussion of 'On the foundations of statistical inference' by allan birnbaum". *Journal of the American Statistical Association*, 57:314–315.

Priest, Graham (1987) *In Contradiction*. Dordrecht: Martinus Nijhoff.

Quine, W. V. (1980) "Two dogmas of empiricism". In *From a Logical Point of View*, pp. 20–46. Harvard University Press, 2nd, revised edition. URL `http://www.ditext.com/quine/quine.html`.

Raiffa, Howard & Robert Schlaifer (2000) *Applied Statistical Decision Theory*. New York: Wiley, 2nd edition.

Ramsey, Frank Plumpton (1978) "Truth and probability". In D. H. Mellor (ed.), *F. P. Ramsey: Philosophical Papers*, pp. 58–100. London and Henley: Routledge and Kegan Paul.

Robinson, G. K. (1975) "Some counterexamples to the theory of confidence intervals". *Biometrika*, 62(1):155–161.

Romeyn, Jan-Willem (2005) *Bayesian Inductive Logic*. Ph.D. thesis, Rijksuniversiteit Groningen, Groningen.

Royall, Richard (1997) *Statistical Evidence: A likelihood paradigm*. London: Chapman and Hall.

——— (2004) "The likelihood paradigm for statistical evidence". In Mark L. Taper & Subhash R. Lele (eds.), *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, pp. 119–138. Chicago and London: University of Chicago Press.

Rubinstein, L. V. & M. H. Gail (1982) "Monitoring rules for stopping accrual in comparative survival studies". *Controlled Clinical Trials*, 3:325–343.

Salmon, Wesley (1996) "Rationality and objectivity in science or Tom Kuhn meets Tom Bayes". In C. Wade Savage (ed.), *Minnesota Studies in the Philosophy of Science*, volume XIV, pp. 175–204. Minneapolis: University of Minnesota Press.

Salmon, Wesley C. (2001a) "Explanation and confirmation: A bayesian critique of inference to the best explanation". In G. Hon & S. Rakover (eds.), *Explanation: Theoretical Approaches and Applications*, pp. 61–92. Amsterdam: Kluwer.

——— (2001b) "Reflections of a bashful Bayesian: A reply to Peter Lipton". In G. Hon & S. Rakover (eds.), *Explanation: Theoretical Approaches and Applications*, pp. 121–126. Amsterdam: Kluwer.

Salsburg, David (1989) "Use of restricted significance tests in clinical trials: Beyond the one- versus two-tailed controversy". *Controlled Clinical Trials*, 10:71–82.

Savage, L. J. (1954) *The Foundations of Statistics*. New York: Wiley.

——— (1961) "The foundations of statistics reconsidered". In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and*

Probability, volume 1. Berkeley, California: University of California Press.

——— (1970) "Comments on a weakened principle of conditionality". *Journal of the American Statistical Association*, 65:399–401.

——— (1976) "On rereading R. A. Fisher (with discussion)". *Annals of Statistics*, 42:441–500.

Savage, L. J. & discussants (1962) *The Foundations of Statistical Inference*. London: Methuen.

Schaffner, Kenneth F. (1993) *Discovery and explanation in biology and medicine*. Chicago: University of Chicago Press.

Schervish, Mark J. (1995) *Theory of Statistics*. New York: Springer–Verlag.

Seidenfeld, Teddy (1979) *Philosophical Problems of Statistical Inference: Learning from R. A. Fisher*. Dordrecht: D. Reidel.

Simon, H. A. (1982) *Models of Bounded Rationality*. MIT Press.

Skyrms, Brian (1987) "Dynamic coherence and probability kinematics". *Philosophy of Science*, 54:1–20.

Smith, Adrian (1995) "A conversation with Dennis Lindley". *Statistical Science*, 10(354):305–319.

Sober, Elliott (2002a) "Instrumentalism, parsimony, and the akaike framework". *Philosophy of Science*, 69:S112–S123.

——— (2002b) "Intelligent design and probability reasoning". *International Journal for the Philosophy of Religion*, 52:65–80.

Sorkin, Rafael D. (1983) "Quantum measure theory and its interpretation (gr-qc/9507057)". Http://physics.syr.edu/ sorkin/some.papers/83.drexel.ps, URL `http://physics.syr.edu/ sorkin/some.papers/83.drexel.ps`.

Spiegelhalter, D. J., L. S. Freedman & P. R. Blackburn (1986) "Monitoring clinical trials - conditional or predictive power?" *Controlled Clinical Trials*, pp. 7: 8–17.

Steel, Daniel (2003) "A Bayesian way to make stopping rules matter". *Erkenntnis*, 58:213–227.

Stein, Charles (1962) "A remark on the likelihood principle". *Journal of the Royal Statistical Society Series A*, 125(4):565–568.

Stone, M. (1991) "Discussion of 'A likelihood paradox' by Goldstein and Howard". *Journal of the Royal Statistical Society Series B*, 53(3):628.

Stone, M. & A. P. Dawid (1972) "Un-Bayesian implications of improper Bayes inference in routine statistical problems." *Biometrika*, 59:369–375.

Stone, Mervyn (1976) "Strong inconsistency from uniform priors". *Journal of the American Statistical Association*, 71:114–116.

Strevins, Michael (2004) "Bayesian confirmation theory: inductive logic, or mere inductive framework?" *Synthese*, 141(3):365–379.

Stuart, Alan, J. Keith Ord & Steven Arnold (1999) *Kendall's Advanced Theory of Statistics Vol. 2A: Classical Inference and the Linear Model*. London: Arnold, 6th edition.

Sweeting, Trevor J. (2001) "Coverage probability bias, objective Bayes and the likelihood principle". *Biometrika*, 88(3):657–675.

Teller, Paul (1969) "Goodman's theory of projection". *British Journal of the Philosophy of Science*, 20:219–238.

Tufte, Edward R. (2001) *The Visual Display of Quantitative Information.* Cheshire, Connecticut: Graphics Press, 2nd edition.

Wald, A. (1947) *Sequential Analysis.* New York: John Wiley.

Wallace, D. L. (1959) "Conditional confidence level properties". *Annals of Mathematical Statistics*, 30:864–876.

Wilson, E. B. (1952) *An Introduction to Scientific Research.* McGraw-Hill.

Winterson, Jeanette (1988) *The Passion.* Atlantic Monthly Press.

Wrinch, Dorothy & Harold Jeffreys (1919) "On some aspects of the theory of probability". *Philosophical Magazine*, 38:715–731.

Insert

	possible symptoms			
hypotheses	vomiting (observed in this case)	diarrhoea (not observed in this case)	social withdrawal (not observed in this case)	other symptoms & combinations (not observed in this case)
dehydration	0.03	0.2	0.5	0.27
PTSD	0.001	0.01	0.95	0.029
anything else	0.001	0.001	0.001	0.997

Table 1