Chapter 1: Introduction & overview.

1.1 Introduction.

Within mathematical statistics, there are many conflicting views about what constitutes information and how to extract it from data. Different approaches, applied to the same data, may produce radically different interpretations, so identifying good methods is very important for the advancement of the social sciences, medicine, economics and other areas.

The existing theories can be classified into two general groups. The first, which we will call 'frequentist[1]', is derived principally from the theories of Fisher and of Neyman and Pearson. While different philosophically and (to some extent) in presentation, these methods are similar in some important respects, notably, that they are in breach of the likelihood principle. This is because the densities of outcomes not actually observed influence the results, for example, via the tail area probabilities that define the *p-value* or *significance level*. In this work, we will contrast frequentist inference with a form of inference consistent with the likelihood principle. The best known such theory is Bayesianism (in its various forms), which is usually seen as the main rival to frequentist inference. However, one may adhere to the likelihood principle without necessarily using prior probabilities, and we will contrast frequentist inference with a non-Bayesian approach in which the likelihood ratio of the observed data is interpreted as a measure of the evidence for one hypothesis relative to another.

When and whether to *condition* upon an observed feature of the data is a point of contention within frequentist inference, closely related to the issue of finding an appropriate *reference set*. Fisher advocated conditioning on ancillary statistics; this (and his related rejection of *power*) was one of the main points where his theory differed from that of Neyman and Pearson. Welch (1939) discussed the issue, as it applied to a particular case, and argued that Neyman and Pearson were correct in not

[1] 'Frequentist inference' is equivalent to 'sampling theory inference' – a more correct term that is not as widely used. It does not imply any particular view about the nature of probabilities.

conditioning on ancillary statistics. However, in 1958, Cox used a different example to argue that conditional inferences are superior – because more relevant – and that Neyman-Pearson inference should be modified to take this into account. His argument, although not fundamentally different from that of Fisher, was much better received so that within a few years it had been widely accepted (at least in theory) that inferences should be carried out conditional on ancillary statistics.

Conditional frequentist inferences are no more consistent with likelihood theory than unconditional inferences. However, in 1962, Birnbaum showed that an *unrestricted conditional principle* (not advocated by Cox but consistent with his core argument) in combination with the *sufficiency principle* is equivalent to the *likelihood principle*. This means that two principles, well regarded by supporters of classical theory, imply a third principle, completely at odds with it. Despite the standing of the two component principles, Birnbaum's theorem did not lead to a widespread conversion from frequentist inference to likelihood-based inference. The fact that, in the type of examples used by Cox, conditioning does not produce inferences that are *closer* to likelihood inferences may help explain why Birnbaum's theorem was not more influential among frequentist statisticians. Some have criticised Birnbaum for basing his conditional principle on an ancillary statistic that is not part of the minimal sufficient statistic for the parameter of interest.

Traditionally, ancillary statistics have been defined with respect to a conventional parameter space, such as the set of real numbers, and this has meant that in many common cases no exact ancillary statistic exists.[2] On the other hand, the argument for conditioning on a statistic that is *very nearly* ancillary is almost as compelling as the argument for conditioning on ancillary statistics. In such a case it is necessary to ascertain that the extra information gained by conditioning outweighs that which has been lost because of the lack of strict ancillarity. Information can be conceived of in many different ways so this is not necessarily straightforward.

This debate has been fuelled by several enlightening cases used as counter-examples to various inferential theories – the type of process described by Lakatos, in *Proofs*

[2] The "range of application [of exact ancillaries] is rather limited, exact ancillaries being rather a rarity in applications". Lloyd, p. 2.

and refutations[3], for mathematics generally. In this thesis we aim to push the debate further by advancing some novel counter-examples that apply to cases previously thought exempt from criticism on a conditional basis. We will consider parameter spaces containing two elements of the conventional parameter space (consistent with tests of two simple hypotheses) and find that we are often able to identify an exact ancillary statistic possessing certain optimal properties. Since these *exhaustive ancillary statistics* are part of the minimal sufficient statistic, they meet the most stringent requirements that have been proposed for valid conditioning and, since they are continuous, conditioning on them is highly informative. If we accept that the arguments of Cox and others still apply when the parameter space is reduced in size, then it follows that we can improve (frequentist) hypothesis tests by conditioning on these statistics and using the conditional error rates. Conditioning on this type of ancillary statistic produces results that are very different from those of standard inference and, although not strictly in accordance with the likelihood principle, are much more consistent with it than either unconditional or existing conditional methods. In particular it is no longer possible to get a statistically significant result against a hypothesis, H, from data where the likelihood under H is greater than the likelihood under the alternative hypothesis. (This is a common failing of traditional tests with high power.) There is a widespread perception that classical inference is more informative than likelihood inference because, for instance, standard confidence intervals are generally narrower than standard likelihood intervals. However, we will show that standard measures are *averages* of components, most of which are associated with *unobserved* values of the ancillary statistic and therefore (arguably) irrelevant. It is this undesirable feature that produces the apparently superior performance; for instance, the standard p-value is the average of a number of conditional probabilities only one of which is actually relevant, the averaging process tending to create a smaller, more ‘significant’ p-value.

[3] Lakatos.

1.2 Terminology & general approach.

Statistical terminology is often arbitrary or inconsistent and, to further complicate matters, certain terms have come to be strongly identified with particular philosophies or points of view. In this work we have tried to use those terms that will (regardless of their origins or associations) help the reader to understand what is being said by reference to familiar concepts. For example, the quantity $\mathrm{P}_H(X \leq x)$ is familiar as the left-sided 'p-value of x', and will be referred to by this name in every context involving left-sided hypothesis tests, including those where a fixed significance level has been specified. The reader may interpret the term 'p-value of x' in a purely mathematical sense, free of its Fisherian baggage.

We use the term 'evidence' quite frequently and intend it to be understood in a general, layman's sense; more particular versions (e.g. likelihood ratio) are always specified. We do assume that most investigators in the sciences and social sciences (as well as many other areas) are frequently interested in obtaining from their data something that can reasonably be described as evidence. It follows from this that any mode of inference that cannot be interpreted as providing evidence (in the broad sense) is of limited use. Although the theory of Neyman and Pearson has sometimes been described (particularly by Neyman) as pure quality control with no evidential interpretation intended, it is quite clear that this form of inference is almost invariably interpreted evidentially by those applying it to (for example) experimental data. Cox (1958) distinguished between finding out 'what the data tells us' and using rules to secure predetermined success rates; this work is about discovering the former.

We join Kendall & Stuart in making the following disclaimer (with an evidential flavour) regarding some particularly unpleasant terminology[4]:

It is necessary to make it clear at the outset that the rather peremptory terms "reject" and "accept", used of a hypothesis under test, … are now conventional usage, to which we shall adhere, and are not intended to imply that any hypothesis is ever finally accepted or rejected in science. If the reader cannot

[4] Kendall & Stuart (1964), Vol. 2, p. 163.

overcome his philosophical dislike of these admittedly inappropriate expressions, he will perhaps agree to regard them as code words, "reject" standing for "decide that the observations are unfavourable to" and "accept" for the opposite.

We have defined the likelihood ratio statistic as is standard in statements of the Neyman-Pearson theorem. That is, for a test of two simple hypotheses H (the null hypothesis) and K (the alternative hypothesis) with data, x, the *likelihood ratio of* x, written $LR(x)$, is given by

$$LR(x) = \frac{f_H(x)}{f_K(x)},$$

where $f_H(\cdot)$ and $f_K(\cdot)$ are the density functions of X under H and K respectively. In some works, for instance Royall, likelihood ratio is defined as the reciprocal of this expression.

1.3 Who will this interest?

The bulk of this work concerns various forms of conditional inference. This places it firmly within the realm of frequentist inference. A likelihoodist need not consider this issue since, as Birnbaum showed, the likelihood principle implies the strongest conditional principle. Also, methods consistent with the likelihood principle do not make inferences dependent on any values in the sample space other than the one observed. For these reasons, this work will be of most interest to practitioners of frequentist inference, particularly those who have been impressed by the arguments in favour of conditioning, but it also throws some light on the differences and similarities between the conflicting schools of thought (see §11.7).

1.4 Why look at tests of two simple hypotheses.

In this work we concentrate on comparisons of simple hypotheses. We do this partly because, in such cases, it is easier to judge whether the result given by an inference procedure accurately describes the available evidence, and because it is not clear to us that those who test composite hypotheses have a clear idea of what question(s) they are seeking to answer. Also we have no reason to believe that any method that delivers flawed results for simple hypotheses is likely to be any more successful in dealing with complex situations, although the complexity of the scenario may hide the flaws. In the case of Neyman-Pearson inference, the theorem on which it is based[5] applies to tests of two simple hypotheses in the first instance, and the theory for testing composite hypotheses is an extension of the method used in the simple case. It is only possible to calculate the significance level *and* power of a test when both hypotheses are simple. Looking at simple hypotheses is not as restrictive as it may appear since interval estimation (for example, confidence intervals) can be equated to the results of tests on pairs of simple hypotheses as long as we can test all such pairs defined by a natural parameter space.

The examples we consider are free of nuisance parameters. This reduces the generality, but follows the long tradition of avoiding discussion of a situation that creates problems for almost all statistical theories.

1.5 Overview

This work falls into two halves. Chapters 2 to 6 contain discussions of some evidential flaws in the analyses of data produced by orthodox frequentist inference, and detail the conditioning debate from 1939 to the present, including the most illuminating examples and the main points of theoretical contention. Although there is reason to believe that conditioning might hold the key to improving frequentist inferences, to date it has not done so. In Chapter 7 we consider, as a contrast, the non-frequentist likelihood approach advocated by Royall. In Chapters 8 to 10, we show how to find exact ancillary statistics for binary parameter spaces in a reasonably wide range of cases. This extends the impact of (frequentist) conditional theory to cases where it was not previously applicable and produces very different results. We consider the results of such conditioning, and discuss the implications – theoretical

[5] The “Neyman-Pearson theorem”, see, for example, Stuart et al (1999)

and practical – of accepting this approach, placing it in the context of the wider debate. A brief outline of the contents of each chapter is given below.

Chapter 2.

This chapter is largely descriptive, laying the groundwork of concepts and definitions central to the development of this work. In it we describe frequentist inference, particularly the ‘optimal’ theory of Neyman and Pearson, who developed a test procedure satisfying their own criteria of good performance. We contrast the formal properties of this approach with the informal intuitive properties that are frequently attributed to it. We describe the concepts of sufficiency and ancillarity and the two related inferential principles of *likelihood* and *sufficiency*.

Chapter 3.

In Chapter 3 we establish that optimal Neyman-Pearson inferences have severe defects when it comes to assessing and describing the evidential content of data, even in straightforward cases. These flaws are so intuitively apparent that they can be discussed without recourse to any specific, alternative notion of ‘evidence’.

In particular, we discuss:

a) Bias between the hypotheses and the implications of trying to control it;
b) Lack of sensitivity of the test procedure to the alternative hypothesis;
c) The problem of weak evidence;
d) The way in which tests of simple hypotheses are generalised to tests of composite hypotheses – why this is possible and whether it is reasonable.

Chapter 4.

In Chapter 4, we describe the historical evolution of the conditional debate via three seminal works on the topic:

a) Welch (1939), where the ‘optimal’ theory of Neyman & Pearson and the conditional theory of Fisher where critically compared;

b) Cox (1958), where the ‘relevance argument’ for conditioning was developed;
c) Birnbaum (1962), where a theorem linking the conditional principle to the likelihood principle was proved.

We highlight the responses of a number of eminent statisticians to these papers, but critical analyses of these and other arguments are postponed until Chapter 5. We describe the examples used by Welch and Cox in great detail in order to facilitate further development of them in later chapters.

Chapter 5.

Chapter 5 introduces the ‘gambling scenario’ test for the reasonableness of an inference (Buehler (1959)), which can be used as an argument for conditioning, distinct from the ‘relevance’ argument of Cox. The examples raised in Chapter 4 are subjected to this test via simulations.

We illuminate Birnbaum’s theorem by analysing the proof and thereby show that the unrestricted conditional principle does indeed rule out the use of frequentist methods. By contrast the restricted conditional principle can be adhered to within a frequentist framework but the restriction itself causes some coherence problems. We show that the unrestricted conditional principle is *not* inherently inconsistent with the sufficiency principle – a claim often made to justify preference for the restricted CP.

We note that the wider community of statistics users is not aware of these important controversies, even though they are of great practical significance, are not difficult to understand, and have been current for more than thirty years.

Chapter 6.

We establish that Welch’s example, which has played a critical role on both sides of the debate and is still regarded as the paradigmatic case for conditioning[6], has been misunderstood, and that neither of the standard competing methods produces sensible inferences. We show that the example will still justify conditional inference if the usual ancillary statistic is replaced with another, defined on a binary parameter space.

[6] Keifer, p. 105.

Chapter 7.

The concept of ‘evidence’ is formalised through Royall’s Quantitative Law of Likelihood[7]. We compare the results obtained by this method with those from conventional frequentism for a number of simple cases, noting that the likelihood interpretations are more consistent with our intuitions about the meaning of the data. Using this interpretation, we show formally that having a conventional p-value that is small is a necessary but not sufficient condition for the data to strongly favour the alternative hypothesis; this result is consistent with the cases we have observed in Chapter 3 where the p-value is low and yet the data does not appear to justify rejecting H.

Chapter 8.

In Chapter 8, we consider cases with certain symmetry properties and identify a type of statistic that is ancillary on binary parameter spaces in all such cases, including the Normal location model. We describe the effect of conditioning upon the ancillary statistic and the practical and theoretical differences between these results and those derived from conventional inference. These include the fact that the conditional p-value is always greater than the conventional p-value, and that the conditional p-value is consistent with the likelihood interpretation described in Chapter 7. We show why it is reasonable to take the view that the conventional measures are not relevant to evidential questions. Birnbaum’s Binomial example and Welch’s Uniform example are instances of the cases discussed here.

Chapter 9.

In Chapter 9 we prove that, in many cases that lack the symmetry of those discussed in Chapter 8, it is nevertheless possible to find a similar ancillary statistic. Some general implications of conditioning on such *exhaustive ancillary statistics* are derived and the statistics are shown to possess a number of optimal properties. We show that, with exhaustive conditional inference, the model-dependent interpretations of a given likelihood ratio tend to converge, in contrast to conventional inference.

[7] Royall, p. 3.

This approach constitutes a major extension of conditional inference within the frequentist framework.

Chapter 10.

We use the approach outlined in Chapter 9 to find conditional inferences for a range of different models and these are compared with conventional and likelihood results. We find that the conditional results are frequently superior to the conventional results when it comes to describing the balance of evidence between the hypotheses, and that they are usually in reasonable agreement with the likelihood results. We show under what circumstances this technique gives intuitively unsatisfactory results, and conjecture that larger samples may overcome this problem. Among the examples considered is a new model (the ‘gradient model’) devised to highlight the difference between conditional and unconditional results more effectively than Welch’s model.

Chapter 11.

In Chapter 11, we summarise the main points and analyse their relevance and implications for the wider debate. We note that, since our approach tends to produce likelihood-type results while still being formally frequentist, it shows that the frequentist/non-frequentist divide does not constitute the most important *practical* distinction between competing approaches.