ABSTRACT

Within *mathematical statistics*, there are important differences of opinion about *which* features of data encapsulate the evidence for or against a hypothesis. At present the dominant school of thought in statistical inference is 'frequentist' theory, which utilises concepts such as the *p-value* of the data or the result of a test with given *error probabilities*. Unlike most non-frequentist methods, this approach is in breach of the *likelihood principle.*

In 1958 Cox published a seminal paper in which he supported Fisher's view that frequentist inferences made *conditional upon the value of an ancillary statistic* produce results that are more relevant to the question at issue (eg. hypotheses) and, therefore, evidentially superior. So compelling was his argument that, within a few years, this view had gained wide acceptance, at least in principle. Inferences carried out according to Cox's method are no more consistent with likelihood theory than other versions of frequentism, even though Birnbaum (1962) showed that certain of Cox's tenets are closely allied to the likelihood principle. In many common cases, no exact ancillary statistic exists on the conventional parameter space, and the conventional, *unconditional* inference remains unchallenged. Attempts to extend the theory of conditioning by the use of *approximately* ancillary statistics are complicated by the need to weigh the gain in relevance against the loss of information occasioned by the lack of strict ancillarity.

In this thesis, I identify a type of statistic that is exactly ancillary, in Cox's sense, on *binary* parameter spaces. Such statistics exist even in cases where there are no exact ancillaries on the conventional parameter space. Adopting Cox's approach, I argue that inferences made, conditional on the value of these statistics, are evidentially superior. Conditioning upon these statistics is *exhaustive* and takes the pursuit of relevance, via conditioning, as far as it can be taken within the frequentist framework.

Exhaustive conditioning allows us to criticise standard frequentist inference in terms of the concepts that define that theory, rather than by reference to concepts imported from alternative systems. We see, for instance, that the standard p-value is the average of a number of conditional probabilities, only one of which is actually relevant to the question at issue. Thus, the conventional p-value and error probabilities of a test do not measure anything useful from the evidential point of view, and should be replaced by their conditional counterparts, which are the measures that are *most relevant*.

Exhaustive conditional inference produces results that are more consistent with the likelihood principle than those from either unconditional or existing conditional methods. In particular, it is not possible to get a result that is statistically significant against a hypothesis, H, from data where the likelihood under H is greater than the likelihood under the alternative hypothesis. The properties of exhaustive conditional inference show that it is possible to approximate likelihood results while remaining formally frequentist.

DECLARATION

This is to certify that

i. the thesis comprises only my original work towards the PhD,
ii. due acknowledgement has been made in the text to all other material used,
iii. the thesis is less than 100,000 words in length, exclusive of tables and bibliography.

_____

Claire F. Leslie, November, 2007

ACKNOWLEDGEMENTS

I wish to take this opportunity to thank a number of people who have contributed to the development of this work at both a technical and a personal level.

My supervisor, Dr Neil Thomason (HPS Dept), has been unflagging in providing me with moral support and retaining interest and enthusiasm for the project over many years. He has provided advice on many practical matters and encouraged and assisted me in the challenging task of taking on a project that is, by its nature, somewhat controversial. Through association with Neil, I have become acquainted with a whole host of interesting questions, outside my area of specialisation, and this has hugely furthered my general education and has allowed me to see my own work in a broader context.

I would also like to thank Dr Keith Hutchison (HPS Dept) for supervision and technical feedback, particularly in the earlier stages of the work, and Dr Jason Grossman (ANU) for introducing me to Birnbaum's theorem and the Likelihood Principle and thereby having a considerable influence on the direction that this work has taken.

Thanks also to Eric Nyberg for conversations going back to the beginning of the project, and to Jeremy de Silva for going to the trouble of closely reading the first four chapters more recently.

I am very grateful for the support of fellow post-graduate student, Paul Carter, with whom I shared an office during the time I wrote this work. It is a great deal easier to keep up consistent work habits in the company of someone who is both interesting and sympathetic (and I look forward to the eventual publication of our ground-breaking research into the provincial nature of the lamington).

Finally, I want to thank my husband, Martin Kelly, and our three children for their love and support over a period that has sometimes been difficult and demoralising. No one could hope for a more sympathetic domestic environment from which to take on a challenge.

Thank you all.

November, 2007.

Glossary of technical terms and abbreviations.

Ancillary statistic: A statistic having (as a minimum requirement) a distribution that is independent of the parameter of interest in the specified parameter space.

Ancillary event: An event having a probability that is independent of the value of the parameter of interest in the specified parameter space.

Best critical region (BCR): The *rejection region* with highest *power* for a given *significance level* (α), given (in the absence of randomisation) by $\{\underset{\sim}{x} : y = LR(\underset{\sim}{x}) \leq c\}$ where $P_H(Y \leq c) = \alpha$.

Binary parameter space (BPS): A *parameter space* containing exactly two distinct values. Denoted (in general) Θ_B.

CLR: Critical likelihood ratio.

Conditional confidence interval (CCI): Unless otherwise specified, this interval is based on *ECI*, i.e. on tests of simple hypotheses (covering all values of the conventional parameter space) conditional on the observed value of the *DDF statistic*.

Confidence interval (CI): Two-sided optimal (NP) confidence interval, i.e. 'uniformly most accurate unbiased' of specified coverage.

Critical region: See *rejection region*.

cp-value: p-value conditional on the observed value of the *DDF statistic*, $D(Y)$. The function $cp(\cdot)$ operates on the *likelihood ratio* value.

CP: Conditional Principle (assumed 'unrestricted' unless specified as 'restricted').

DDF statistic: The 'difference of the distribution functions' statistic. This statistic is an *EAS* whenever Y is continuous. Denoted $D(\cdot)$ as a function of Y.

Exhaustive ancillary statistic (EAS): An *ancillary statistic* that partitions the *support* of the *likelihood ratio statistic* into pairs of values, except for LR=1, which comprises an entire sub-set.

H: (*Simple*) null hypothesis (often given in this work as $\theta = \theta_1$).

K: (*Simple*) alternative hypothesis (often given as $\theta = \theta_2$).

LI: Likelihood interval.

Likelihood: Denoted $L(x;\theta)$.
(1) As a function of x: the probability or density of random x for a given (fixed) value of θ.
(2) As a function of θ: the probability or density of (fixed) x at (variable) θ.
[NB: The use of the term 'likelihood' as distinct from 'density' is, by some authors, confined to (2), however we use it as both a function of x and a function of θ, made explicit in the discussion].

Likelihood ratio (LR): The **likelihood ratio statistic** is generally denoted 'Y' in this work. $Y = LR(\underset{\sim}{X}) = LR(\underset{\sim}{X};\theta_1,\theta_2) = \frac{g_H(\underset{\sim}{X})}{g_K(\underset{\sim}{X})} = \frac{f_H(Y)}{f_K(Y)}$, where g is the probability or density of $\underset{\sim}{X}$ calculated under the two *simple hypotheses* H $(\theta = \theta_1)$ and K $(\theta = \theta_2)$, and f is the probability or density of Y. For the likelihood ratio **value**, substitute $\underset{\sim}{x}$ for $\underset{\sim}{X}$.

LP: Likelihood Principle.

LL: Law of Likelihood.

QLL: Quantitative Law of Likelihood.

Minimum sufficient statistic (MSS): A sufficient statistic that is a function of each and every other sufficient statistic.

MLE: Maximum likelihood estimator (random variable) or estimate (observed value).

Neyman-Pearson optimal test (Best test): A test based on the *BCR*.

Parameter space: Set containing all values under consideration for the parameter of interest. Denoted (in general) Θ.

Power: $P_K(\text{Reject H})$. Denoted κ.

p-value: For the purposes of this work, best thought of as $P_H(Y \leq y_0)$ where Y is the *likelihood ratio statistic* and y_0 is its observed value. The p-value is the smallest *significance level* value at which H is rejected (by a *NP optimal test*) given this data.

Rejection region: Set of values of a test statistic (often a sufficient statistic) for which H is rejected at the α level.

Scenario: Let $\mathcal{M}$, be a model connecting a natural statistic, X, with a parameter of interest, θ, via some probability density $f_{\mathcal{M}}(x;\theta)$. A *scenario*, $\mathcal{S} \equiv (\mathcal{M}, H, K)$, is a combination of the model with two distinct, ordered hypotheses each specifying the value of θ.

Significance level: $P_H(\text{Reject H})$. Denoted α.

Simple hypothesis: A hypothesis that completely defines the distribution of the test statistic, for example (in the absence of nuisance parameters), by assigning a single specific value to the parameter of interest, θ.

SP: Sufficiency Principle.

Sufficient statistic: A statistic, $s(\underset{\sim}{X})$, is sufficient for $\theta \in \Theta$ if and only if $L(\underset{\sim}{x};\theta) = g(s(\underset{\sim}{x});\theta) \cdot h(\underset{\sim}{x}),\ \forall \underset{\sim}{x},\ \forall \theta \in \Theta.$

Support: The support of the random variable, X, (having a specified distribution) is the set of all values, x, such that the probability or density of X at x is non-*zero*.

Type I error: Rejecting H when H is true. The probability of this event is the *significance level* of the test.

Type II error: Accepting H when K is true. The probability of this event (denoted β) equals one minus the *power* of the test (i.e. $\beta = 1 - \kappa$).

Y: Denotes the *likelihood ratio statistic* throughout most of this work.

ϕ and Φ: Pdf and cdf of a *standard Normal* (i.e. *Gaussian*) variable:

$$\phi(z) = \tfrac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2} z^2\},\ z \in \mathbb{R}$$

$$\Phi(z) = \int_{-\infty}^{z} \tfrac{1}{\sqrt{2\pi}} \exp\{-\tfrac{1}{2} t^2\} dt,\ z \in \mathbb{R}.$$