APTS: Statistical Inference

Simon Shaw

s.shaw@bath.ac.uk

Warwick, 13-16 December 2022

Contents

1 Introduction

1.1 Introduction to the course

Course aims: To explore a number of statistical principles, such as the likelihood principle and sufficiency principle, and their logical implications for statistical inference. To consider the nature of statistical parameters, the different viewpoints of Bayesian and Frequentist approaches and their relationship with the given statistical principles. To introduce the idea of inference as a statistical decision problem. To understand the meaning and value of ubiquitous constructs such as p-values, confidence sets, and hypothesis tests.

Course learning outcomes: An appreciation for the complexity of statistical inference, recognition of its inherent subjectivity and the role of expert judgement, the ability to critique familiar inference methods, knowledge of the key choices that must be made, and scepticism about apparently simple answers to difficult questions.

The course will cover three main topics:

1. Principles of inference: the Likelihood Principle, Birnbaum's Theorem, the Stopping Rule Principle, implications for different approaches.
2. Decision theory: Bayes Rules, admissibility, and the Complete Class Theorems. Implications for point and set estimation, and for hypothesis testing.
3. Confidence sets, hypothesis testing, and p-values. Good and not-so-good choices. Level error, and adjusting for it. Interpretation of small and large p-values.

These notes could not have been prepared without, and have been developed from, those prepared by Jonathan Rougier (University of Bristol) who lectured this course previously. I thus acknowledge his help and guidance though any errors are my own.

1.2 Statistical endeavour

Efron and Hastie (2016) on pxvi consider statistical endeavour as comprising two parts:

algorithms aimed at solving individual applications and a more formal theory of statistical inference: "very broadly speaking, algorithms are what statisticians do while inference says why they do them." Hence, it is that the algorithm comes first: "algorithmic invention is a more free-wheeling and adventurous enterprise, with inference playing catch-up as it strives to assess the accuracy, good or bad, of some hot new algorithmic methodology." This though should not underplay the value of the theory: as Cox (2006) on pxiii writes "without some systematic structure statistical methods for the analysis of data become a collection of tricks that are hard to assimilate and interrelate to one another ... the development of new problems would become entirely a matter of ad hoc ingenuity. Of course, such ingenuity is not to be undervalued and indeed one role of theory is to assimilate, generalise and perhaps modify and improve the fruits of such ingenuity."

1.3 Statistical models

A ***statistical model*** is an artefact to link our beliefs about things which we can measure, or observe, to things we would like to know. For example, we might suppose that X denotes the value of things we can observe and Y the values of the things that we would like to know. Prior to making any observations, both X and Y are unknown, they are ***random variables***. In a statistical approach, we quantify our uncertainty about them by specifying a probability distribution for (X, Y). Then, if we observe $X = x$ we can consider the conditional probability of Y given $X = x$, that is we can consider ***predictions*** about Y.

In this context, artefact denotes an object made by a human, for example, you or me. There are no statistical models that don't originate inside our minds. So there is no arbiter to determine the "true" statistical model for (X, Y): we may expect to disagree about the statistical model for (X, Y), between ourselves, and even within ourselves from one time-point to another. In common with all other scientists, statisticians do not require their models to be true: as Box (1979) writes 'it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations ... for such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?"' Statistical models exist to make prediction feasible.

Maybe it would be helpful to say a little more about this. Here is the usual procedure in

"public" Science, sanitised and compressed:

1. Given an interesting question, formulate it as a problem with a solution.
2. Using experience, imagination, and technical skill, make some simplifying assumptions to move the problem into the mathematical domain, and solve it.
3. Contemplate the simplified solution in the light of the assumptions, e.g. in terms of robustness. Maybe iterate a few times.
4. Publish your simplified solution (including, of course, all of your assumptions), and your recommendation for the original question, if you have one. Prepare for criticism.

MacKay (2009) provides a masterclass in this procedure. The statistical model represents a statistician's "simplifying assumptions".

A statistical model for a random variable X is created by ruling out many possible probability distributions. This is most clearly seen in the case when the set of possible outcomes is finite.

Example 1 *Let $\mathcal{X} = \{x^{(1)}, \ldots, x^{(k)}\}$ denote the set of possible outcomes of X so that the sample space consists of $|\mathcal{X}| = k$ elements. The set of possible probability distributions for X is*

$$\mathcal{P} \quad = \quad \left\{ p \in \mathbb{R}^k \,:\, p_i \geq 0 \,\forall i, \sum_{i=1}^{k} p_i = 1 \right\},$$

where $p_i = \mathbb{P}(X = x^{(i)})$. A statistical model may be created by considering a family of distributions $\mathcal{F}$ which is a subset of $\mathcal{P}$. We will typically consider families where the functional form of the probability mass function is specified but a finite number of parameters θ are unknown. That is

$$\mathcal{F} \quad = \quad \left\{ p \in \mathcal{P} \,:\, p_i = f_X(x^{(i)} \,|\, \theta) \text{ for some } \theta \in \Theta \right\}.$$

We shall proceed by assuming that our statistical model can be expressed as a *parametric model*.

Definition 1 *(Parametric model)*
A parametric model for a random variable X is the triple $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \,|\, \theta)\}$ where only the finite dimensional parameter $\theta \in \Theta$ is unknown.

Thus, the model specifies the sample space $\mathcal{X}$ of the quantity to be observed X, the parameter space Θ, and a family of distributions, $\mathcal{F}$ say, where $f_X(x \,|\, \theta)$ is the distribution for X when θ is the value of the parameter. In this general framework, both X and θ may be multivariate and we use f_X to represent the density function irrespective of whether X is continuous or discrete. If it is discrete then $f_X(x \,|\, \theta)$ gives the probability of an individual value x. Typically, θ is continuous-valued.

The method by which a statistician chooses the chooses the family of distributions $\mathcal{F}$ and then the parametric model $\mathcal{E}$ is hard to codify, although experience and precedent are obviously relevant; Davison (2003) offers a book-length treatment with many useful examples. However, once the model has been specified, our primary focus is to make an *inference* on the parameter θ. That is we wish to use observation $X = x$ to update our knowledge about θ so that we may, for example, estimate a function of θ or make predictions about a random variable Y whose distribution depends upon θ.

Definition 2 *(Statistic; estimator)*
Any function of a random variable X is termed a statistic. If T is a statistic then $T = t(X)$ is a random variable and $t = t(x)$ the corresponding value of the random variable when $X = x$. In general, T is a vector. A statistic designed to estimate θ is termed an estimator.

Typically, estimators can be divided into two types.

1. A ***point estimator*** which maps from the sample space $\mathcal{X}$ to a point in the parameter space Θ.
2. A ***set estimator*** which maps from $\mathcal{X}$ to a set in Θ.

For prediction, we consider a parametric model for (X, Y), $\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Theta, f_{X,Y}(x, y \,|\, \theta)\}$ from which we can calculate the ***predictive model*** $\mathcal{E}^* = \{\mathcal{Y}, \Theta, f_{Y|X}(y \,|\, x, \theta)\}$ where

$$f_{Y|X}(y \,|\, x, \theta) \;=\; \frac{f_{X,Y}(x, y \,|\, \theta)}{f_X(x \,|\, \theta)} \;=\; \frac{f_{X,Y}(x, y \,|\, \theta)}{\int_{\mathcal{Y}} f_{X,Y}(x, y \,|\, \theta)\, dy}. \tag{1.1}$$

1.4 Some principles of statistical inference

In the first half of the course we shall consider principles for statistical inference. These principles guide the way in which we learn about θ and are meant to be either self-evident, or logical implications of principles which are self-evident. In this section we aim to motivate three of these principles: the weak likelihood principle, the strong likelihood principle, and the sufficiency principle. The first two principles relate to the concept of the likelihood and the third to the idea of a sufficient statistic.

1.4.1 Likelihood

In the model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \,|\, \theta)\}$, f_X is a function of x for known θ. If we have instead observed x then we could consider viewing this as a function, termed the *likelihood*, of θ for known x. This provides a means of comparing the plausibility of different values of θ.

Definition 3 *(Likelihood)*
The likelihood for θ given observations x is

$$L_X(\theta; x) \;=\; f_X(x \,|\, \theta), \quad \theta \in \Theta$$

regarded as a function of θ for fixed x.

If $L_X(\theta_1; x) > L_X(\theta_2; x)$ then the observed data x were more likely to occur under $\theta = \theta_1$ than θ_2 so that θ_1 can be viewed as more plausible than θ_2. Note that we choose to make the dependence on X explicit as the measurement scale affects the numerical value of the likelihood.

Example 2 *Let $X = (X_1, \ldots, X_n)$ and suppose that, for given $\theta = (\alpha, \beta)$, the X_i are independent and identically distributed Gamma(α, β) random variables. Then,*

$$f_X(x \,|\, \theta) \quad = \quad \frac{\beta^{n\alpha}}{\Gamma^n(\alpha)} \left(\prod_{i=1}^{n} x_i \right)^{\alpha - 1} \exp\left(-\beta \sum_{i=1}^{n} x_i \right) \tag{1.2}$$

if $x_i > 0$ for each $i \in \{1, \ldots, n\}$ and zero otherwise. If, for each i, $Y_i = X_i^{-1}$ then the Y_i are independent and identically distributed Inverse-Gamma(α, β) random variables with

$$f_Y(y \,|\, \theta) \quad = \quad \frac{\beta^{n\alpha}}{\Gamma^n(\alpha)} \left(\prod_{i=1}^{n} \frac{1}{y_i} \right)^{\alpha + 1} \exp\left(-\beta \sum_{i=1}^{n} \frac{1}{y_i} \right)$$

if $y_i > 0$ for each $i \in \{1, \ldots, n\}$ and zero otherwise. Thus,

$$L_Y(\theta; y) \quad = \quad \left(\prod_{i=1}^{n} \frac{1}{y_i} \right)^{2} L_X(\theta; x).$$

If we are interested in inferences about $\theta = (\alpha, \beta)$ following the observation of the data, then it seems reasonable that these should be invariant to the choice of measurement scale: it should not matter whether x or y was recorded.[1]

More generally, suppose that X is a continuous vector random variable and $Y = g(X)$ a one-to-one transformation of X with non-vanishing Jacobian $\partial x / \partial y$ then the probability density function of Y is

$$f_Y(y \,|\, \theta) \quad = \quad f_X(x \,|\, \theta) \left| \frac{\partial x}{\partial y} \right|, \tag{1.3}$$

where $x = g^{-1}(y)$ and $|\cdot|$ denotes the determinant. Consequently, as Cox and Hinkley (1974) on p12 observe, if we are interested in comparing two possible values of θ, θ_1 and θ_2 say, using the likelihood then we should consider the ratio of the likelihoods rather than, for example, the difference since

$$\frac{f_Y(y \,|\, \theta = \theta_1)}{f_Y(y \,|\, \theta = \theta_2)} \quad = \quad \frac{f_X(x \,|\, \theta = \theta_1)}{f_X(x \,|\, \theta = \theta_2)}$$

so that the comparison does not depend upon whether the data was recorded as x or as $y = g(x)$. It seems reasonable that the proportionality of the likelihoods given by equation (1.3) should lead to the same inference about θ.

[1] In the course, we will see that this idea can developed into an inference principle called the Transformation Principle.

The likelihood principle

Our discussion of the likelihood function suggests that it is the ratio of the likelihoods for differing values of θ that should drive our inferences about θ. In particular, if two likelihoods are proportional for all values of θ then the corresponding likelihood ratios for any two values θ_1 and θ_2 are identical. Initially, we consider two outcomes x and y from the same model: this gives us our first possible principle of inference.

Definition 4 *(The weak likelihood principle)*
If $X = x$ and $X = y$ are two observations for the experiment $\mathcal{E}_X = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$ such that

$$L_X(\theta; y) \quad = \quad c(x, y) L_X(\theta; x)$$

for all $\theta \in \Theta$ then the inference about θ should be the same irrespective of whether $X = x$ or $X = y$ was observed.

A stronger principle can be developed if we consider two random variables X and Y corresponding to two different experiments, $\mathcal{E}_X = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$ and $\mathcal{E}_Y = \{\mathcal{Y}, \Theta, f_Y(y \mid \theta)\}$ respectively, for the *same* parameter θ. Notice that this situation includes the case where $Y = g(X)$ (see equation (1.3)) but is not restricted to that.

Example 3 *Consider, given θ, a sequence of independent Bernoulli trials with parameter θ. We wish to make inference about θ and consider two possible methods. In the first, we carry out n trials and let X denote the total number of successes in these trials. Thus, $X \mid \theta \sim Bin(n, \theta)$ with*

$$f_X(x \mid \theta) \quad = \quad \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad x = 0, 1, \ldots, n.$$

In the second method, we count the total number Y of trials up to and including the rth success so that $Y \mid \theta \sim Nbin(r, \theta)$, the negative binomial distribution, with

$$f_Y(y \mid \theta) \quad = \quad \binom{y-1}{r-1} \theta^r (1-\theta)^{y-r}, \quad y = r, r+1, \ldots.$$

Suppose that we observe $X = x = r$ and $Y = y = n$. Then in each experiment we have seen x successes in n trials and so it may be reasonable to conclude that we make the same inference about θ from each experiment. Notice that in this case

$$L_Y(\theta; y) \;=\; f_Y(y \mid \theta) \quad = \quad \frac{x}{y} f_X(x \mid \theta) \;=\; \frac{x}{y} L_X(\theta; x)$$

so that the likelihoods are proportional.

Motivated by this example, a second possible principle of inference is a strengthening of the weak likelihood principle.

Definition 5 *(The strong likelihood principle)*
Let $\mathcal{E}_X$ and $\mathcal{E}_Y$ be two experiments which have the same parameter θ. If $X = x$ and $Y = y$ are two observations such that

$$L_Y(\theta; y) \quad = \quad c(x, y) L_X(\theta; x)$$

for all $\theta \in \Theta$ then the inference about θ should be the same irrespective of whether $X = x$ or $Y = y$ was observed.

1.4.2 Sufficient statistics

Consider the model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$. If a sample $X = x$ is obtained there may be cases when, rather than knowing each individual value of the sample, certain summary statistics could be utilised as a sufficient way to capture all of the relevant information in the sample. This leads to the idea of a sufficient statistic.

Definition 6 *(Sufficient statistic)*
A statistic $S = s(X)$ is sufficient for θ if the conditional distribution of X, given the value of $s(X)$ (and θ) $f_{X|S}(x \mid s, \theta)$ does not depend upon θ.

Note that, in general, S is a vector and that if S is sufficient then so is any one-to-one function of S. It should be clear from Definition 6 that the sufficiency of S for θ is dependent upon the choice of the family of distributions in the model.

Example 4 *Let $X = (X_1, \ldots, X_n)$ and suppose that, for given θ, the X_i are independent and identically distributed $Po(\theta)$ random variables. Then*

$$f_X(x \mid \theta) \quad = \quad \prod_{i=1}^{n} \frac{\theta^{x_i} \exp(-\theta)}{x_i!} \quad = \quad \frac{\theta^{\sum_{i=1}^n x_i} \exp(-n\theta)}{\prod_{i=1}^n x_i!},$$

if $x_i \in \{0, 1, \ldots\}$ for each $i \in \{1, \ldots, n\}$ and zero otherwise. Let $S = \sum_{i=1}^n X_i$ then $S \sim Po(n\theta)$ so that

$$f_S(s \mid \theta) \quad = \quad \frac{(n\theta)^s \exp(-n\theta)}{s!}$$

for $s \in \{0, 1, \ldots\}$ and zero otherwise. Thus, if $f_S(s \mid \theta) > 0$ then, as $s = \sum_{i=1}^n x_i$,

$$f_{X|S}(x \mid s, \theta) \quad = \quad \frac{f_X(x \mid \theta)}{f_S(s \mid \theta)} \quad = \quad \frac{(\sum_{i=1}^n x_i)!}{\prod_{i=1}^n x_i!} n^{-\sum_{i=1}^n x_i}$$

which does not depend upon θ. Hence, $S = \sum_{i=1}^n X_i$ is sufficient for θ. Similarly, the sample mean $\frac{1}{n}S$ is also sufficient.

Sufficiency for a parameter θ can be viewed as the idea that S captures all of the information about θ contained in X. Having observed S, nothing further can be learnt about θ by observing X as $f_{X|S}(x \mid s, \theta)$ has no dependence on θ.

Definition 6 is confirmatory rather than constructive: in order to use it we must somehow guess a statistic S, find the distribution of it and then check that the ratio of the distribution of X to the distribution of S does not depend upon θ. However, the following theorem[2] allows us to easily find a sufficient statistic.

Theorem 1 *(Fisher-Neyman Factorisation Theorem)*
The statistic $S = s(X)$ is sufficient for θ if and only if, for all x and θ,

$$f_X(x \,|\, \theta) \quad = \quad g(s(x), \theta) h(x)$$

for some pair of functions $g(s(x), \theta)$ and $h(x)$.

Example 5 *We revisit Example 2 and the case where the X_i are independent and identically distributed Gamma(α, β) random variables. From equation (1.2) we have*

$$f_X(x \,|\, \theta) \;=\; \frac{\beta^{n\alpha}}{\Gamma^n(\alpha)} \left(\prod_{i=1}^n x_i \right)^{\alpha} \exp\left(-\beta \sum_{i=1}^n x_i \right) \left(\prod_{i=1}^n x_i \right)^{-1} \;=\; g\left(\prod_{i=1}^n x_i, \sum_{i=1}^n x_i, \theta \right) h(x)$$

so that $S = (\prod_{i=1}^n X_i, \sum_{i=1}^n X_i)$ is sufficient for θ.

Notice that S defines a data reduction. In Example 4, $S = \sum_{i=1}^n X_i$ is a scalar so that all of the information in the n-vector $x = (x_1, \ldots, x_n)$ relating to the scalar θ is contained in just one number. In Example 5, all of the information in the n-vector for the two dimensional parameter $\theta = (\alpha, \beta)$ is contained in just two numbers. Using the Fisher-Neyman Factorisation Theorem, we can easily obtain the following result for models drawn from the exponential family.

Theorem 2 *Let $X = (X_1, \ldots, X_n)$ and suppose that the X_i are independent and identically distributed from the exponential family of distributions given by*

$$f_{X_i}(x_i \,|\, \theta) \quad = \quad h(x_i) c(\theta) \exp\left(\sum_{j=1}^k a_j(\theta) b_j(x_i) \right),$$

where $\theta = (\theta_1, \ldots, \theta_d)$ for $d \leq k$. Then

$$S \quad = \quad \left(\sum_{i=1}^n b_1(X_i), \ldots, \sum_{i=1}^n b_k(X_i) \right)$$

is a sufficient statistic for θ.

Example 6 *The Poisson distribution, see Example 4, is a member of the exponential family where $d = k = 1$ and $b_1(x_i) = x_i$ giving the sufficient statistic $S = \sum_{i=1}^n X_i$. The Gamma distribution, see Example 5, is also a member of the exponential family with $d = k = 2$ and $b_1(x_i) = x_i$ and $b_2(x_i) = \log x_i$ giving the sufficient statistic $S = (\sum_{i=1}^n X_i, \sum_{i=1}^n \log X_i)$ which is equivalent to the pair $(\sum_{i=1}^n X_i, \prod_{i=1}^n X_i)$.*

[2] For a proof see, for example, p276 of Casella and Berger (2002)

The sufficiency principle

Following Section 2.2(iii) of Cox and Hinkley (1974), we may interpret sufficiency as follows. Consider two individuals who both assert the model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$. The first individual observes x directly. The second individual also observes x but in a two stage process:

1. They first observe a value $s(x)$ of a sufficient statistic S with distribution $f_S(s \mid \theta)$.
2. They then observe the value x of the random variable X with distribution $f_{X|S}(x \mid s)$ which does not depend upon θ.

It may well then be reasonable to argue that, as the final distribution for X for the two individuals are identical, the conclusions drawn from the observation of a given x should be identical for the two individuals. That is, they should make the same inference about θ. For the second individual, when sampling from $f_{X|S}(x \mid s)$ they are sampling from a fixed distribution and so, assuming the correctness of the model, only the first stage is informative: all of the knowledge about θ is contained in $s(x)$. If one takes these two statements together then the inference to be made about θ depends only on the value $s(x)$ and not the individual values x_i contained in x. This leads us to a third possible principle of inference.

Definition 7 *(The sufficiency principle)*
If $S = s(X)$ is a sufficient statistic for θ and x and y are two observations such that $s(x) = s(y)$, then the inference about θ should be the same irrespective of whether $X = x$ or $X = y$ was observed.

1.5 Schools of thought for statistical inference

There are two broad approaches to statistical inference, generally termed the ***classical approach*** and the ***Bayesian approach***. The former approach is also called ***frequentist***. In brief the difference between the two is in their interpretation of the parameter θ. In a classical setting, the parameter is viewed as a fixed unknown constant and inferences are made utilising the distribution $f_X(x \mid \theta)$ even after the data x has been observed. Conversely, in a Bayesian approach parameters are treated as random and so may be equipped with a probability distribution. We now give a short overview of each school.

1.5.1 Classical inference

In a classical approach to statistical inference, no further probabilistic assumptions are made once the parametric model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$ is specified. In particular, θ is treated as an unknown constant and interest centres on constructing good methods of inference.

To illustrate the key ideas, we shall initially consider point estimators. The most familiar classical point estimator is the *maximum likelihood estimator (MLE)*. The MLE $\hat{\theta} = \hat{\theta}(X)$

satisfies, see Definition 3,

$$L_X(\hat{\theta}(x); x) \quad \geq \quad L_X(\theta; x)$$

for all $\theta \in \Theta$. Intuitively, the MLE is a reasonable choice for an estimator: it's the value of θ which makes the observed sample most likely. In general, the MLE can be viewed as a good point estimator with a number of desirable properties. For example, it satisfies the invariance property[3] that if $\hat{\theta}$ is the MLE of θ then for any function $g(\theta)$, the MLE of $g(\theta)$ is $g(\hat{\theta})$. However, there are drawbacks which come from the difficulties of finding the maximum of a function.

Efron and Hastie (2016) consider that there are three ages of statistical inference: the pre-computer age (essentially the period from 1763 and the publication of Bayes' rule up until the 1950s), the early-computer age (from the 1950s to the 1990s), and the current age (a period of computer-dependence with enormously ambitious algorithms and model complexity). With these developments in mind, it is clear that there exist a hierarchy of statistical models.

1. Models where $f_X(x \mid \theta)$ has a known analytic form.
2. Models where $f_X(x \mid \theta)$ can be evaluated.
3. Models where we can simulate X from $f_X(x \mid \theta)$.

Between the first case and the second case exist models where $f_X(x \mid \theta)$ can be evaluated up to an unknown constant, which may or may not depend upon θ.

In the first case, we might be able to derive an analytic expression for $\hat{\theta}$ or to prove that $f_X(x \mid \theta)$ has a unique maximum so that any numerical maximisation will converge to $\hat{\theta}(x)$.

Example 7 *We revisit Examples 2 and 5 and the case when $\theta = (\alpha, \beta)$ are the parameters of a Gamma distribution. In this case, the maximum likelihood estimators $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ satisfy the equations*

$$\begin{aligned} \hat{\beta} &= \frac{\hat{\alpha}}{\overline{X}}, \\ 0 &= n \log \hat{\alpha} - n \log \overline{X} - n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} + \sum_{i=1}^{n} \log X_i. \end{aligned}$$

Thus, numerical methods are required to find $\hat{\theta}$.

In the second case, we could still numerically maximise $f_X(x \mid \theta)$ but the maximiser may converge to a local maximum rather than the global maximum $\hat{\theta}(x)$. Consequently, any algorithm utilised for finding $\hat{\theta}(x)$ must have some additional procedures to ensure that all local maxima are ignored. This is a non-trivial task in practice. In the third case, it is extremely difficult to find the MLE and other estimators of θ may be preferable. This

[3] For a proof of this property, see Theorem 7.2.10 of Casella and Berger (2002)

example shows that the choice of algorithm is critical: the MLE is a good method of inference only if:

1. you can prove that it has good properties for your choice of $f_X(x \mid \theta)$ and
2. you can prove that the algorithm you use to find the MLE of $f_X(x \mid \theta)$ does indeed do this.

The second point arises once the choice of estimator has made. We now consider how to assess whether a chosen point estimator is a good estimator. One possible attractive feature is that the method is, on average, correct. As estimator $T = t(X)$ is said to be *unbiased* if

$$\text{bias}(T \mid \theta) \quad = \quad \mathbb{E}(T \mid \theta) - \theta$$

is zero for all $\theta \in \Theta$. This is a superficially attractive criterion but it can lead to unexpected results (which are not sensible estimators) even in simple cases.

Example 8 *(Example 8.1 of Cox and Hinkley (1974))*
Let X denote the number of independent Bernoulli(θ) trials up to and including the first success so that $X \sim Geom(\theta)$ with

$$f_X(x \mid \theta) \quad = \quad (1-\theta)^{x-1}\theta$$

for $x = 1, 2, \ldots$ and zero otherwise. If $T = t(X)$ is an unbiased estimator of θ then

$$\mathbb{E}(T \mid \theta) \;=\; \sum_{x=1}^{\infty} t(x)(1-\theta)^{x-1}\theta \;=\; \theta.$$

Letting $\phi = 1 - \theta$ we thus have

$$\sum_{x=1}^{\infty} t(x)\phi^{x-1}(1-\phi) \quad = \quad 1 - \phi.$$

Thus, equating the coefficients of powers of ϕ, we find that the unique unbiased estimate of θ is

$$t(x) \quad = \quad \begin{cases} 1 & x = 1, \\ 0 & x = 2, 3, \ldots. \end{cases}$$

This is clearly not a sensible estimator.

Another drawback with the bias is that it is not, in general, transformation invariant. For example, if T is an unbiased estimator of θ then T^{-1} is not, in general, an unbiased estimator of θ^{-1} as $\mathbb{E}(T^{-1} \mid \theta) \neq 1/\mathbb{E}(T \mid \theta) \;=\; \theta^{-1}$. An alternate, and better, criterion is that T has small *mean square error (MSE)*,

$$\begin{aligned} \text{MSE}(T \mid \theta) &= \mathbb{E}((T-\theta)^2 \mid \theta) \\ &= \mathbb{E}(\{(T - \mathbb{E}(T \mid \theta)) + (\mathbb{E}(T \mid \theta) - \theta)\}^2 \mid \theta) \\ &= Var(T \mid \theta) + \text{bias}(T \mid \theta)^2. \end{aligned}$$

Thus, estimators with a small mean square error will typically have small variance and bias and it's possible to trade unbiasedness for a smaller variance. What this discussion does make clear is that it is properties of the distribution of the estimator T, known as the ***sampling distribution***, across the range of possible values of θ that are used to determine whether or not T is a good inference rule. Moreover, this assessment is made not for the observed data x but based on the distributional properties of X. In this sense, we determine the method of inference by calibrating how they would perform were they to be used repeatedly. As Cox (2006) on p8 notes "we intend, of course, that this long-run behaviour is some assurance that with our particular data currently under analysis sound conclusions are drawn."

Example 9 *Let $X = (X_1, \ldots, X_n)$ and suppose that the X_i are independent and identically distribution normal random variables with mean θ and variance 1. Letting $\overline{X} = \frac{1}{n}\sum_{i-1}^{n} X_i$ then*

$$\mathbb{P}\left(\theta - \frac{1.96}{\sqrt{n}} \leq \overline{X} \leq \theta + \frac{1.96}{\sqrt{n}} \,\middle|\, \theta\right) \;=\; \mathbb{P}\left(\overline{X} - \frac{1.96}{\sqrt{n}} \leq \theta \leq \overline{X} + \frac{1.96}{\sqrt{n}} \,\middle|\, \theta\right) \;=\; 0.95.$$

Thus, $(\overline{X} - \frac{1.96}{\sqrt{n}}, \overline{X} + \frac{1.96}{\sqrt{n}})$ is a set estimator for θ with a coverage probability of 0.95. We can consider this as a method of inference, or algorithm. If we observe $X = x$ corresponding to $\overline{X} = \overline{x}$ then our algorithm is

$$x \quad \mapsto \quad \left(\overline{x} - \frac{1.96}{\sqrt{n}}, \overline{x} + \frac{1.96}{\sqrt{n}}\right)$$

*which produces a 95% confidence interval for θ. Notice that we report two things: the result of the algorithm (the actual interval) and the justification (the long-run property of the algorithm) or **certification** of the algorithm (95% confidence interval).*

As the example demonstrates, the certification is determined by the sampling distribution ($\overline{X}$ is a normal distribution with mean θ and variance $1/n$) whilst the choice of algorithm is determined by the certification (in this case, the coverage probability of 0.95[4]). This is an inverse problem in the sense that we work backwards from the required certificate to the choice of algorithm. Notice that we are able to compute the coverage for every $\theta \in \Theta$ as we have a ***pivot***: $\sqrt{n}(\overline{X} - \theta)$ is a normal distribution with mean 0 and variance 1 and so parameter free. For more complex models it will not be straightforward to do this.

We can generalise the idea exhibited in Example 9 into a key principle of the classical approach that

1. Every algorithm is certified by its sampling distribution, and
2. The choice of algorithm depends on this certification.

[4]For example, if we wanted a coverage of 0.90 then we would amend the algorithm by replacing 1.96 in the interval calculation with 1.645.

Thus, point estimators of θ may be certified by their mean square error function; set estimators of θ may be certified by their coverage probability; hypothesis tests may be certified by their power function. The definition of each of these certifications is not important here, though they are easy to look up. What is important to understand is that in each case an algorithm is proposed, the sampling distribution is inspected, and then a certificate is issued. Individuals and user communities develop conventions about certificates they like their algorithms to possess, and thus they choose an algorithm according to its certification. For example, in clinical trials, it is for a hypothesis test to have a type I error below 5% with large power.

We now consider prediction in a classical setting. As in Section 1.3, see equation (1.1), from a parametric model for (X, Y), $\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Theta, f_{X,Y}(x, y \,|\, \theta)\}$ we can calculate the ***predictive model***

$$\mathcal{E}^* \quad = \quad \{\mathcal{Y}, \Theta, f_{Y|X}(y \,|\, x, \theta)\}.$$

The difficulty here is that $\mathcal{E}^*$ is a family of distributions and we seek to reduce this down to a single distribution; effectively, to "get rid of" θ. If we accept, as our working hypothesis, that one of the elements in the family of distributions is true, that is that there is a $\theta^* \in \Theta$ which is the true value of θ then the corresponding predictive distribution $f_{Y|X}(y \,|\, x, \theta^*)$ is the true predictive distribution for Y. The classical solution is to replace θ^* by ***plugging-in*** an estimate based on x.

Example 10 *If we use the MLE $\hat{\theta} = \hat{\theta}(x)$ then we have an algorithm*

$$x \quad \mapsto \quad f_{Y|X}(y \,|\, x, \hat{\theta}(x)).$$

The estimator does not have to be the MLE and so we see that different estimators produce different algorithms.

1.5.2 Bayesian inference

In a Bayesian approach to statistical inference, we consider that, in addition to the parametric model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \,|\, \theta)\}$, the uncertainty about the parameter θ prior to observing X can be represented by a ***prior distribution*** π on θ. We can then utilise Bayes's theorem to obtain the ***posterior distribution*** $\pi(\theta \,|\, x)$ of θ given $X = x$,

$$\pi(\theta \,|\, x) \quad = \quad \frac{f_X(x \,|\, \theta)\pi(\theta)}{\int_\Theta f_X(x \,|\, \theta)\pi(\theta)\, d\theta}.$$

We make the following definition.

Definition 8 *(Bayesian statistical model)*
A Bayesian statistical model is the collection $\mathcal{E}_B = \{\mathcal{X}, \Theta, f_X(x \,|\, \theta), \pi(\theta)\}$.

As O'Hagan and Forster (2004) on p5 note, "the posterior distribution encapsulates all that is known about θ following the observation of the data x, and can be thought of as comprising an all-embracing inference statement about θ." In the context of algorithms, we have

$$x \quad \mapsto \quad \pi(\theta \,|\, x)$$

where each choice of prior distribution produces a different algorithm. In this course, our primary focus is upon general theory and methodology and so, at this point, we shall merely note that both specifying a prior distribution for the problem at hand and deriving the corresponding posterior distribution are decidedly non-trivial tasks. Indeed, in the same way that we discussed a hierarchy of statistical models for $f_X(x \,|\, \theta)$ in Section 1.5.1, an analogous hierarchy exists for the posterior distribution $\pi(\theta \,|\, x)$.

In contrast to the plug-in classical approach to prediction, the Bayesian approach can be viewed as ***integrate-out***. If $\mathcal{E}_B = \{\mathcal{X} \times \mathcal{Y}, \Theta, f_{X,Y}(x, y \,|\, \theta), \pi(\theta)\}$ is our Bayesian model for (X, Y) and we are interested in prediction for Y given $X = x$ then we can integrate out θ to obtain the parameter free conditional distribution $f_{Y|X}(y \,|\, x)$:

$$f_{Y|X}(y \,|\, x) \quad = \quad \int_\Theta f_{Y|X}(y \,|\, x, \theta) \pi(\theta \,|\, x) \, d\theta. \tag{1.4}$$

In terms of an algorithm, we have

$$x \quad \mapsto \quad f_{Y|X}(y \,|\, x)$$

where, as equation (1.4) involves integrating out θ according to the posterior distribution, then each choice of prior distribution produces a different algorithm.

Whilst the posterior distribution expresses all of knowledge about the parameter θ given the data x, in order to express this knowledge in clear and easily understood terms we need to derive appropriate summaries of the posterior distribution. Typical summaries include point estimates, interval estimates, probabilities of specified hypotheses.

Example 11 *Suppose that θ is a univariate parameter and we consider summarising θ by a number d. We may compute the posterior expectation of the squared distance between t and θ.*

$$\begin{aligned} \mathbb{E}((d - \theta)^2 \,|\, X) &= \mathbb{E}(d^2 - 2d\theta + \theta^2 \,|\, X) \\ &= d^2 - 2d\mathbb{E}(\theta \,|\, X) + \mathbb{E}(\theta^2 \,|\, X) \\ &= (d - \mathbb{E}(\theta \,|\, X))^2 + Var(\theta \,|\, X). \end{aligned}$$

Consequently $d = \mathbb{E}(\theta \,|\, X)$, the posterior expectation, minimises the posterior expected square error and the minimum value of this error is $Var(\theta \,|\, X)$, the posterior variance.

In this way, we have a justification for $\mathbb{E}(\theta \mid X)$ as an estimate of θ. We could view d as a decision, the result of which was to occur an error $t - \theta$. In this example we choose to measure how good or bad a particular decision was by the squared error suggesting that we were equally happy to overestimate θ as underestimate it and that large errors are more serious than they would be if an alternate measure such as $|d - \theta|$ was used.

1.5.3 Inference as a decision problem

In the second half of the course we will study inference as a decision problem. In this context we assume that we make a decision d which acts as an estimate of θ. The consequence of this decision in a given context can be represented by a specific loss function $L(\theta, d)$ which measures the quality of the choice d when θ is known. In this setting, decision theory allows us to identify a best decision. As we will see, this approach has two benefits. Firstly, we can form a link between Bayesian and classical procedures, in particular the extent to which classical estimators, confidence intervals and hypothesis tests can be interpreted within a Bayesian framework. Secondly, we can provide Bayesian solutions to the inference questions addressed in a classical approach.

2 Principles for Statistical Inference

2.1 Introduction

We wish to consider inferences about a parameter θ given a parametric model

$$\mathcal{E} \quad = \quad \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}.$$

We assume that the model is true so that only $\theta \in \Theta$ is unknown. We wish to learn about θ from observations x so that $\mathcal{E}$ represents a model for this ***experiment***. Our inferences can be described in terms of an algorithm involving both $\mathcal{E}$ and x. In this chapter, we shall assume that $\mathcal{X}$ is finite; Basu (1975) on p4 argues that "this contingent and cognitive universe of ours is in reality only finite and, therefore, discrete ... [infinite and continuous models] are to be looked upon as mere approximations to the finite realities."

Statistical principles guide the way in which we learn about θ. These principles are meant to be either self-evident, or logical implications of principles which are self-evident. What is really interesting about Statistics, for both statisticians and philosophers (and real-world decision makers) is that the logical implications of some self-evident principles are not at all self-evident, and have turned out to be inconsistent with prevailing practices. This was a discovery made in the 1960s. Just as interesting, for sociologists (and real-world decision makers) is that the then-prevailing practices have survived the discovery, and continue to be used today.

This chapter is about statistical principles, and their implications for statistical inference. It demonstrates the power of abstract reasoning to shape everyday practice.

2.2 Reasoning about inferences

Statistical inferences can be very varied, as a brief look at the 'Results' sections of the papers in an Applied Statistics journal will reveal. In each paper, the authors have decided on a different interpretation of how to represent the 'evidence' from their dataset. On the surface, it does not seem possible to construct and reason about statistical principles when the notion of 'evidence' is so plastic. It was the inspiration of Allan Birnbaum[1] Birnbaum (1962) to

[1] Allan Birnbaum (1923-1976)

see—albeit indistinctly at first—that this issue could be side-stepped. Over the next two decades, his original notion was refined; key papers in this process were Birnbaum (1972), Basu (1975), Dawid (1977), and the book by Berger and Wolpert (1988).

The model $\mathcal{E}$ is accepted as a working hypothesis. How the statistician chooses her statements about the true value θ is entirely down to her and her client: as a point or a set in Θ, as a choice among alternative sets or actions, or maybe as some more complicated, not ruling out visualisations. Dawid (1977) puts this well - his formalism is not excessive, for really understanding this crucial concept. The statistician defines, ***a priori***, a set of possible "inferences about θ", and her task is to choose an element of this set based on $\mathcal{E}$ and x. Thus the statistician should see herself as a function 'Ev': a mapping from $(\mathcal{E}, x)$ into a predefined set of 'inferences about θ', or

$$(\mathcal{E}, x) \overset{\text{statistician, Ev}}{\longmapsto} \text{Inference about } \theta.$$

Thus, $\text{Ev}(\mathcal{E}, x)$ is the inference about θ made if $\mathcal{E}$ is performed and $X = x$ is observed. For example, $\text{Ev}(\mathcal{E}, x)$ might be the maximum likelihood estimator of θ or a 95% confidence interval for θ. Birnbaum called $\mathcal{E}$ the 'experiment', x the 'outcome', and Ev the 'evidence'.

Birnbaum (1962)'s formalism, of an experiment, an outcome, and an evidence function, helps us to anticipate how we can construct statistical principles. First, there can be different experiments with the same θ. Second, under some outcomes, we would agree that it is self-evident that these different experiments provide the same evidence about θ. Thus, we can follow p3 of Basu (1975) and define the equality or equivalence of $\text{Ev}(\mathcal{E}_1, x_1)$ and $\text{Ev}(\mathcal{E}_2, x_2)$ as meaning that

1. The experiments $\mathcal{E}_1$ and $\mathcal{E}_2$ are related to the same parameter θ.
2. 'Everything else being equal', the outcome x_1 from $\mathcal{E}_1$ 'warrants the same inference' about θ as does the outcomes x_2 from $\mathcal{E}_2$.

As we will show, these self-evident principles imply other principles. These principles all have the same form: under such and such conditions, the evidence about θ should be the same. Thus they serve only to rule out inferences that satisfy the conditions but have different evidences. They do not tell us how to do an inference, only what to avoid.

2.3 The principle of indifference

We now give our first example of a statistical principle, using the name conferred by Basu (1975).

Principle 1 *(Weak Indifference Principle, WIP)*
Let $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$. *If* $f_X(x \mid \theta) = f_X(x' \mid \theta)$ *for all* $\theta \in \Theta$ *then* $Ev(\mathcal{E}, x) = Ev(\mathcal{E}, x')$.

As Birnbaum (1972) notes, this principle, which he termed mathematical equivalence, asserts that we are indifferent between two models of evidence if they differ only in the manner of

the labelling of sample points. For example, if $X = (X_1, \ldots, X_n)$ where the X_is are a series of independent Bernoulli trials with parameter θ then $f_X(x \mid \theta) = f_X(x' \mid \theta)$ if x and x' contain the same number of successes. We will show that the WIP logically follows from the following two principles, which I would argue are self-evident, for which we use the names conferred by Dawid (1977).

Principle 2 *(Distribution Principle, DP)*
If $\mathcal{E} = \mathcal{E}'$, then $Ev(\mathcal{E}, x) = Ev(\mathcal{E}', x)$.

As Dawid (1977) on p247 writes "informally, this says that the only aspects of an experiment which are relevant to inference are the sample space and the family of distributions over it."

Principle 3 *(Transformation Principle, TP)*
Let $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$. For the bijective $g : \mathcal{X} \to \mathcal{Y}$, let $\mathcal{E}^g = \{\mathcal{Y}, \Theta, f_Y(y \mid \theta)\}$, the same experiment as $\mathcal{E}$ but expressed in terms of $Y = g(X)$, rather than X. Then $Ev(\mathcal{E}, x) = Ev(\mathcal{E}^g, g(x))$.

This principle states that inferences should not depend on the way in which the sample space is labelled.

Example 12 *Recall Example 2. Under TP, inferences about θ are the same if we observe $x = (x_1, \ldots, x_n)$ where each independent $X_i \sim Gamma(\alpha, \beta)$ or $X^{-1} = (1/x_1, \ldots, 1/x_n)$ where each independent $X_i^{-1} \sim$ Inverse-Gamma(α, β).*

We have the following result, see Basu (1975), Dawid (1977).

Theorem 3 *$(DP \wedge TP) \to WIP$.*

Proof: Fix $\mathcal{E}$, and suppose that $x, x' \in \mathcal{X}$ satisfy $f_X(x \mid \theta) = f_X(x' \mid \theta)$ for all $\theta \in \Theta$, as in the condition of the WIP. Now consider the transformation $g : \mathcal{X} \to \mathcal{X}$ which switches x for x', but leaves all of the other elements of $\mathcal{X}$ unchanged. In this case $\mathcal{E} = \mathcal{E}^g$. Then

$$\begin{aligned} \mathrm{Ev}(\mathcal{E}, x') &= \mathrm{Ev}(\mathcal{E}^g, x') && (2.1) \\ &= \mathrm{Ev}(\mathcal{E}^g, g(x)) && (2.2) \\ &= \mathrm{Ev}(\mathcal{E}, x), && (2.3) \end{aligned}$$

where equation (2.1) follows by the DP and (2.3) follows from (2.2) by the TP. We thus have the WIP. □

Therefore, if I accept the principles DP and TP then I must also accept the WIP. Conversely, if I do not want to accept the WIP then I must reject at least one of the DP and TP. This is the pattern of the next few sections, where either I must accept a principle, or, as a matter of logic, I must reject one of the principles that implies it.

2.4 The Likelihood Principle

Suppose we have experiments $\mathcal{E}_i = \{\mathcal{X}_i, \Theta, f_{X_i}(x_i \,|\, \theta)\}$, $i = 1, 2, \ldots$, where the parameter space Θ is the same for each experiment. Let $p_1, p_2, \ldots$ be a set of known probabilities so that $p_i \geq 0$ and $\sum_i p_i = 1$. The ***mixture*** $\mathcal{E}^*$ of the experiments $\mathcal{E}_1, \mathcal{E}_2, \ldots$ according to mixture probabilities $p_1, p_2, \ldots$ is the two-stage experiment

1. A random selection of one of the experiments: $\mathcal{E}_i$ is selected with probability p_i.
2. The experiment selected in stage 1. is performed.

Thus, each outcome of the experiment $\mathcal{E}^*$ is a pair (i, x_i), where $i = 1, 2, \ldots$ and $x_i \in \mathcal{X}_i$, and family of distributions

$$f^*((i, x_i) \,|\, \theta) \quad = \quad p_i f_{X_i}(x_i \,|\, \theta). \tag{2.4}$$

The famous example of a mixture experiment is the 'two instruments' (see Section 2.3 of Cox and Hinkley (1974). There are two instruments in a laboratory, and one is accurate, the other less so. The accurate one is more in demand, and typically it is busy 80% of the time. The inaccurate one is usually free. So, *a priori*, there is a probability of $p_1 = 0.2$ of getting the accurate instrument, and $p_2 = 0.8$ of getting the inaccurate one. Once a measurement is made, of course, there is no doubt about which of the two instruments was used. The following principle asserts what must be self-evident to everybody, that inferences should be made according to which instrument was used and not according to the *a priori* uncertainty.

Principle 4 *(Weak Conditionality Principle, WCP)*
Let $\mathcal{E}^*$ be the mixture of the experiments $\mathcal{E}_1$, $\mathcal{E}_2$ according to mixture probabilities p_1, $p_2 = 1 - p_1$. Then $Ev(\mathcal{E}^*, (i, x_i)) = Ev(\mathcal{E}_i, x_i)$.

Thus, the WCP states that inferences for θ depend only on the experiment performed. As Casella and Berger (2002) on p293 state "the fact that this experiment was performed rather than some other, has not increased, decreased, or changed knowledge of θ."

In Section 1.4.1, we motivated the strong likelihood principle, see Definition 5. We now reassert this principle.[2]

Principle 5 *(Strong Likelihood Principle, SLP)*
Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two experiments which have the same parameter θ. If $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ satisfy $f_{X_1}(x_1 \,|\, \theta) = c(x_1, x_2) f_{X_2}(x_2 \,|\, \theta)$, that is

$$L_{X_1}(\theta; x_1) \quad = \quad c(x_1, x_2) L_{X_2}(\theta; x_2)$$

for some function $c > 0$ for all $\theta \in \Theta$ then $Ev(\mathcal{E}_1, x_1) = Ev(\mathcal{E}_2, x_2)$.

[2] The SLP is self-attributed to G. Barnard, see his comment to Birnbaum (1962), p. 308. But it is alluded to in the statistical writings of R.A. Fisher, almost appearing in its modern form in Fisher (1956).

The SLP thus states that if two likelihood functions for the same parameter have the same shape, then the evidence is the same. As we shall discuss in Section 2.8, many classical statistical procedures violate the SLP and the following result was something of the bombshell, when it first emerged in the 1960s. The following form is due to Birnbaum (1972) and Basu (1975).[3]

Theorem 4 *(Birnbaum's Theorem)*
$(WIP \wedge WCP) \leftrightarrow SLP$.

Proof: Both SLP $\rightarrow$ WIP and SLP $\rightarrow$ WCP are straightforward. The trick is to prove (WIP $\wedge$ WCP) $\rightarrow$ SLP. So let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two experiments which have the same parameter, and suppose that $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ satisfy $f_{X_1}(x_1 \,|\, \theta) = c(x_1, x_2) f_{X_2}(x_2 \,|\, \theta)$ where the function $c > 0$. As the value c is known (as the data has been observed) then consider the mixture experiment with $p_1 = 1/(1+c)$ and $p_2 = c/(1+c)$. Then, using equation (2.4),

$$\begin{aligned} f^*((1, x_1) \,|\, \theta) &= \frac{1}{1+c} f_{X_1}(x_1 \,|\, \theta) \\ &= \frac{c}{1+c} f_{X_2}(x_2 \,|\, \theta) \qquad (2.5) \\ &= f^*((2, x_2) \,|\, \theta) \qquad (2.6) \end{aligned}$$

where equation (2.6) follows from (2.5) by (2.4). Then the WIP implies that

$$\mathrm{Ev}\left(\mathcal{E}^*, (1, x_1)\right) = \mathrm{Ev}\left(\mathcal{E}^*, (2, x_2)\right).$$

Finally, applying the WCP to each side we infer that

$$\mathrm{Ev}(\mathcal{E}_1, x_1) = \mathrm{Ev}(\mathcal{E}_2, x_2),$$

as required. □

Thus, either I accept the SLP, or I explain which of the two principles, WIP and WCP, I refute. Methods which violate the SLP face exactly this challenge.

2.5 The Sufficiency Principle

In Section 1.4.2 we considered the idea of sufficiency. From Definition 6, if $S = s(X)$ is sufficient for θ then

$$f_X(x \,|\, \theta) = f_{X|S}(x \,|\, s, \theta) f_S(s \,|\, \theta) \qquad (2.7)$$

where $f_{X|S}(x \,|\, s, \theta)$ does not depend upon θ. Consequently, we consider the experiment $\mathcal{E}^S = \{s(\mathcal{X}), \Theta, f_S(s \,|\, \theta)\}$.

[3] Birnbaum's original result, Birnbaum (1962), used a stronger condition than WIP and a slightly weaker condition than WCP. Theorem 4 is clearer.

Principle 6 *(Strong Sufficiency Principle, SSP)*
If $S = s(X)$ is a sufficient statistic for the experiment $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$ then $Ev(\mathcal{E}, x) = Ev(\mathcal{E}^S, s(x))$.

A weaker, Basu (1975) terms it 'perhaps a trifle less severe', but more familiar version which is in keeping with Definition 7 is as follows.

Principle 7 *(Weak Sufficiency Principle, WSP)*
If $S = s(X)$ is a sufficient statistic for the experiment $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$ and $s(x) = s(x')$ then $Ev(\mathcal{E}, x) = Ev(\mathcal{E}, x')$.

Theorem 5 *$SLP \rightarrow SSP \rightarrow WSP \rightarrow WIP$.*

Proof: From equation (2.7), $f_X(x \mid \theta) = c f_S(s \mid \theta)$ where $c = f_{X|S}(x \mid s, \theta)$ does not depend upon θ. Thus, from the SLP, Principle 5, $\mathrm{Ev}(\mathcal{E}, x) = \mathrm{Ev}(\mathcal{E}^S, s(x))$ which is the SSP, Principle 6. Note, that from the SSP,

$$\begin{aligned} \mathrm{Ev}(\mathcal{E}, x) &= \mathrm{Ev}(\mathcal{E}^S, s(x)) && (2.8)\\ &= \mathrm{Ev}(\mathcal{E}^S, s(x')) && (2.9)\\ &= \mathrm{Ev}(\mathcal{E}, x') && (2.10) \end{aligned}$$

where (2.9) follows from (2.8) as $s(x) = s(x')$ and (2.10) from (2.9) from the SSP. We thus have the WSP, Principle 7. Finally, notice that if $f_X(x \mid \theta) = f_X(x' \mid \theta)$ as in the statement of WIP, Principle 1, then $s(x) = x'$ is sufficient for x and so from the WSP, $\mathrm{Ev}(\mathcal{E}, x) = \mathrm{Ev}(\mathcal{E}, x')$ giving the WIP. □

Finally, we note that if we put together Theorem 4 and Theorem 5 we get the following corollary.

Corollary 1 *$(WIP \wedge WCP) \rightarrow SSP$.*

2.6 Stopping rules

Suppose that we consider observing a sequence of random variables $X_1, X_2, \ldots$ where the number of observations is not fixed in advanced but depends on the values seen so far. That is, at time j, the decision to observe X_{j+1} can be modelled by a probability $p_j(x_1, \ldots, x_j)$. We can assume, resources being finite, that the experiment must stop at specified time m, if it has not stopped already, hence $p_m(x_1, \ldots, x_m) = 0$. The *stopping rule* may then be denoted as $\tau = (p_1, \ldots, p_m)$. This gives an experiment $\mathcal{E}^\tau$ with, for $n = 1, 2, \ldots$, $f_n(x_1, \ldots, x_n \mid \theta)$ where consistency requires that

$$f_n(x_1, \ldots, x_n \mid \theta) = \sum_{x_{n+1}} \cdots \sum_{x_m} f_m(x_1, \ldots, x_n, x_{n+1}, \ldots x_m \mid \theta).$$

We utilise the following example from p42 of Basu (1975) to motivate the stopping rule principle. Consider four different coin-tossing experiments (with some finite limit on the number of tosses).

$\mathcal{E}_1$ Toss the coin exactly 10 times;

$\mathcal{E}_2$ Continue tossing until 6 heads appear;

$\mathcal{E}_3$ Continue tossing until 3 consecutive heads appear;

$\mathcal{E}_4$ Continue tossing until the accumulated number of heads exceeds that of tails by exactly 2.

One could easily adduce more sequential experiments which gave the same outcome. Notice that $\mathcal{E}_1$ corresponds to a binomial model and $\mathcal{E}_2$ to a negative binomial. Suppose that all four experiments have the same outcome $x =$ (T,H,T,T,H,H,T,H,H,H).

In line with Example 3, we may feel that the evidence for θ, the probability of heads, is the same in every case. Once the sequence of heads and tails is known, the intentions of the original experimenter (i.e. the experiment she was doing) are immaterial to inference about the probability of heads, and the simplest experiment $\mathcal{E}_1$ can be used for inference. We can consider the following principle which Basu (1975) claims is due to George Barnard.[4]

Principle 8 *(Stopping Rule Principle, SRP)*
In a sequential experiment $\mathcal{E}^\tau$, $Ev(\mathcal{E}^\tau, (x_1, \ldots, x_n))$ does not depend on the stopping rule τ.

The SRP is nothing short of revolutionary, if it is accepted. It implies that that the intentions of the experimenter, represented by τ, are irrelevant for making inferences about θ, once the observations $(x_1, \ldots, x_n)$ are available. Once the data is observed, we can ignore the sampling plan. Thus the statistician could proceed as though the simplest possible stopping rule were in effect, which is $p_1 = \cdots = p_{n-1} = 1$ and $p_n = 0$, an experiment with n fixed in advance. Obviously it would be liberating for the statistician to put aside the experimenter's intentions (since they may not be known and could be highly subjective), but can the SRP possibly be justified? Indeed it can.

Theorem 6 *SLP $\rightarrow$ SRP.*

Proof: Let τ be an arbitrary stopping rule, and consider the outcome $(x_1, \ldots, x_n)$, which we will denote as $x_{1:n}$. We take the first observation with probability one and, for $j = 1, \ldots, n-1$, the $(j+1)$th observation is taken with probability $p_j(x_{1:j})$, and we stop after the nth observation with probability $1 - p_n(x_{1:n})$. Consequently, the probability of this

[4] George Barnard (1915-2002)

outcome under τ is

$$\begin{aligned} f_\tau(x_{1:n} \,|\, \theta) &= f_1(x_1 \,|\, \theta) \left\{ \prod_{j=1}^{n-1} p_j(x_{1:j})\, f_{j+1}(x_{j+1} \,|\, x_{1:j}, \theta) \right\} (1 - p_n(x_{1:n})) \\ &= \left\{ \prod_{j=1}^{n-1} p_j(x_{1:j}) \right\} (1 - p_n(x_{1:n}))\, f_1(x_1 \,|\, \theta) \prod_{j=2}^{n} f_j(x_j \,|\, x_{1:(j-1)}, \theta) \\ &= \left\{ \prod_{j=1}^{n-1} p_j(x_{1:j}) \right\} (1 - p_n(x_{1:n}))\, f_n(x_{1:n} \,|\, \theta). \end{aligned}$$

Now observe that this equation has the form

$$f_\tau(x_{1:n} \,|\, \theta) = c(x_{1:n}) f_n(x_{1:n} \,|\, \theta) \tag{2.11}$$

where $c(x_{1:n}) > 0$. Thus the SLP implies that $\mathrm{Ev}(\mathcal{E}^\tau, x_{1:n}) = \mathrm{Ev}(\mathcal{E}^n, x_{1:n})$ where $\mathcal{E}^n = \{\mathcal{X}^n, \Theta, f_n(x_{1:n} \,|\, \theta)\}$. Since the choice of stopping rule was arbitrary, equation (2.11) holds for all stopping rules, showing that the choice of stopping rule is irrelevant. □

The Stopping Rule Principle has become enshrined in our profession's collective memory due to this iconic comment from L.J. Savage[5], one of the great statisticians of the Twentieth Century:

> May I digress to say publicly that I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resist an idea so patently right. (Savage et al. (1962), p76).

This comment captures the revolutionary and transformative nature of the SRP.

2.7 A stronger form of the WCP

The new concept in this section is 'ancillarity'. This has several different definitions in the Statistics literature; the one we use is close to that of Section 2.2 of Cox and Hinkley (1974).

Definition 9 *(Ancillarity)*
Y is ancillary in the experiment $\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Theta, f_{X,Y}(x, y \,|\, \theta)\}$ exactly when $f_{X,Y}$ factorises as

$$f_{X,Y}(x, y \,|\, \theta) = f_Y(y) f_{X|Y}(x \,|\, y, \theta).$$

[5] Leonard Jimmie Savage (1917-1971)

In other words, the marginal distribution of Y is completely specified. Not all families of distributions will factorise in this way, but when they do, there are new possibilities for inference, based around stronger forms of the WCP.

Here is an example, which will be familiar to all statisticians. We have been given a sample $x = (x_1, \ldots, x_n)$ to evaluate. In fact n itself is likely to be the outcome of a random variable N, because the process of sampling itself is rather uncertain. However, we seldom concern ourselves with the distribution of N when we evaluate x; instead we treat N as known. Equivalently, we treat N as ancillary and condition on $N = n$. In this case, we might think that inferences drawn from observing (n, x) should be the same as those for x conditioned on $N = n$.

When Y is ancillary, we can consider the conditional experiment

$$\mathcal{E}^{X\,|\,y} = \{\mathcal{X}, \Theta, f_{X\,|\,Y}(x\,|\,y, \theta)\},$$

This is an experiment where we condition on $Y = y$, i.e. treat Y as known, and treat X as the only random variable. This is an attractive idea, captured in the following principle.

Principle 9 *(Strong Conditionality Principle, SCP)*
If Y is ancillary in $\mathcal{E}$, then $Ev(\mathcal{E}, (x, y)) = Ev(\mathcal{E}^{X|y}, x)$.

As a second example, consider a regression of Y on X appears to make a distinction between Y, which is random, and X, which is not. This distinction is insupportable, given that the roles of Y and X are often interchangeable, and determined by the *hypothèse du jour*. What is really happening is that (X, Y) is random, but X is being treated as ancillary for the parameters in $f_{Y|X}$, so that its parameters are auxiliary in the analysis. Then the SCP is invoked (implicitly), which justifies modelling Y conditionally on X, treating X as known.

Clearly the SCP implies the WCP, with the experiment indicator $I \in \{1, 2\}$ being ancillary, since p is known. It is almost obvious that the SCP comes for free with the SLP. Another way to put this is that the WIP allows us to 'upgrade' the WCP to the SCP.

Theorem 7 *SLP $\rightarrow$ SCP.*

Proof: Suppose that Y is ancillary in $\mathcal{E} = \{\mathcal{X} \times \mathcal{Y}, \Theta, f_{X,Y}(x, y\,|\,\theta)\}$. Thus, for all $\theta \in \Theta$,

$$\begin{aligned} f_{X,Y}(x, y\,|\,\theta) &= f_Y(y) f_{X|Y}(x\,|\,y, \theta) \\ &= c(y) f_{X|Y}(x\,|\,y, \theta) \end{aligned}$$

Then the SLP implies that

$$\mathrm{Ev}\,(\mathcal{E}, (x, y)) = \mathrm{Ev}(\mathcal{E}^{X|y}, x),$$

as required. □

2.8 The Likelihood Principle in practice

Now we should pause for breath, and ask the obvious questions: is the SLP vacuous? Or trivial? In other words, Is there any inferential approach which respects it? Or do all inferential approaches respect it? We shall focus on the classical and Bayesian approaches, as outlined in Section 1.5.1 and Section 1.5.2 respectively.

Recall from Definition 8 that a Bayesian statistical model is the collection

$$\mathcal{E}_B \quad = \quad \{\mathcal{X}, \Theta, f_X(x \,|\, \theta), \pi(\theta)\}.$$

The posterior distribution

$$\pi(\theta \,|\, x) \quad = \quad c(x) f_X(x \,|\, \theta) \pi(\theta) \tag{2.12}$$

where $c(x)$ is the normalising constant,

$$c(x) \quad = \quad \left\{ \int_\Theta f_X(x \,|\, \theta) \pi(\theta) \, d\theta \right\}^{-1}.$$

From a Bayesian perspective, all knowledge about the parameter θ given the data x are represented by $\pi(\theta \,|\, x)$ and any inferences made about θ are derived from this distribution. If we have two Bayesian models with the *same* prior distribution, $\mathcal{E}_{B,1} = \{\mathcal{X}_1, \Theta, f_{X_1}(x_1 \,|\, \theta), \pi(\theta)\}$ and $\mathcal{E}_{B,2} = \{\mathcal{X}_2, \Theta, f_{X_2}(x_2 \,|\, \theta), \pi(\theta)\}$ and $f_{X_1}(x_1 \,|\, \theta) = c(x_1, x_2) f_{X_2}(x_2 \,|\, \theta)$ then

$$\begin{aligned} \pi(\theta \,|\, x_1) \quad &= \quad c(x_1) f_{X_1}(x_1 \,|\, \theta) \pi(\theta) \\ &= \quad c(x_1) c(x_1, x_2) f_{X_2}(x_2 \,|\, \theta) \pi(\theta) \\ &= \quad \pi(\theta \,|\, x_2) \end{aligned} \tag{2.13}$$

so that the posterior distributions are the same. Consequently, the same inferences are drawn from either model and so the Bayesian approach satisfies the SLP. Notice that this assumes that the prior distribution exists independently of the outcome, that is the prior does not depend upon the form of the data. In practice, though, is hard to do. Some methods for making default choices for π depend on f_X, notably Jeffreys priors and reference priors, see for example, Section 5.4 of Bernardo and Smith (2000). These methods violate the SLP.

The classical approach however violates the SLP. As we noted in Section 1.5.1, algorithms are certified in terms of their sampling distributions, and selected on the basis of their certification. For example, the mean square error of an estimator T, $MSE(T \,|\, \theta) = Var(T \,|\, \theta) + \text{bias}(T \,|\, \theta)^2$ depends upon the first and second moments of the distribution of $T \,|\, \theta$. Consequently, they depend on the whole sample space $\mathcal{X}$ and not just the observed $x \in \mathcal{X}$.

Example 13 *(Example 1.3.5 of Robert (2007))*
Suppose that X_1, X_2 are iid $N(\theta, 1)$ so that

$$f(x_1, x_2 \,|\, \theta) \quad \propto \quad \exp\left\{-(\overline{x} - \theta)^2\right\}.$$

Now, consider the alternate model for the same parameter θ

$$g(x_1, x_2 \,|\, \theta) \;=\; \pi^{-\frac{3}{2}} \frac{\exp\left\{-(\overline{x} - \theta)^2\right\}}{1 + (x_1 - x_2)^2}$$

We thus observe that $f(x_1, x_2 \,|\, \theta) \propto g(x_1, x_2 \,|\, \theta)$ as a function of θ. If the SLP is applied, then inference about θ should be the same in both models. However, the distribution of g is quite different from that of f and so estimators of θ will have different classical properties if they do not depend only on $\overline{x}$. For example, g has heavier tails than f and so respective confidence intervals may differ between the two.

We can extend the idea of this example by showing that if $\mathrm{Ev}(\mathcal{E}, x)$ depends on the value of $f_X(x' \,|\, \theta)$ for some $x' \neq x$ then we can create an alternate experiment $\mathcal{E}_1 = \{\mathcal{X}, \Theta, f_1(x \,|\, \theta)\}$ where $f_1(x \,|\, \theta) = f_X(x \,|\, \theta)$ for the observed x but $f_1(x \,|\, \theta) \neq f_X(x \,|\, \theta)$ for all $x \in \mathcal{X}$. In particular, we can ensure that $f_1(x' \,|\, \theta) \neq f_X(x' \,|\, \theta)$. Then, typically, Ev does not respect the SLP.

To do this, let $\tilde{x} \neq x, x'$ and set

$$\begin{aligned} f_1(x' \,|\, \theta) &= \alpha f_X(x' \,|\, \theta) + \beta f_X(\tilde{x} \,|\, \theta) \\ f_1(\tilde{x} \,|\, \theta) &= (1 - \alpha) f_X(x' \,|\, \theta) + (1 - \beta) f_X(\tilde{x} \,|\, \theta) \end{aligned}$$

with $f_1 = f_X$ elsewhere. Clearly $f_1(x' \,|\, \theta) + f_1(\tilde{x} \,|\, \theta) = f_X(x' \,|\, \theta) + f_X(\tilde{x} \,|\, \theta)$ and so f_1 is a probability distribution. By suitable choice of α, β we can redistribute the mass to ensure $f_1(x' \,|\, \theta) \neq f_X(x' \,|\, \theta)$. Consequently, whilst $f_1(x \,|\, \theta) = f_X(x \,|\, \theta)$ for the observed x we will not have that $\mathrm{Ev}(\mathcal{E}, x) = \mathrm{Ev}(\mathcal{E}_1, x)$ and so will violate the SLP.

This illustrates that classical inference typically does not respect the SLP because the sampling distribution of the algorithm depends on values of f_X other than $L(\theta; x) = f_X(x \,|\, \theta)$. The two main difficulties with violating the SLP are:

1. To reject the SLP is to reject at least one of the WIP and the WCP. Yet both of these principles seem self-evident. Therefore violating the SLP is either illogical or obtuse.

2. In their everyday practice, statisticians use the SCP (treating some variables as ancillary) and the SRP (ignoring the intentions of the experimenter). Neither of these is self-evident, but both are implied by the SLP. If the SLP is violated, then they both need an alternative justification.

Alternative formal justifications for the SCP and the SRP have not been forthcoming.

2.9 Reflections

The statistician takes delivery of an outcome x. Her standard practice is to assume the truth of a statistical model $\mathcal{E}$, and then turn $(\mathcal{E}, x)$ into an inference about the true value of the parameter θ. As remarked several times already (see Chapter 1), this is not the end of

her involvement, but it is a key step, which may be repeated several times, under different notions of the outcome and different statistical models. This chapter concerns this key step: how she turns $(\mathcal{E}, x)$ into an inference about θ.

Whatever inference is required, we assume that the statistician applies an algorithm to $(\mathcal{E}, x)$. In other words, her inference about θ is not arbitrary, but transparent and reproducible - this is hardly controversial, because anything else would be non-scientific. Following Birnbaum, the algorithm is denoted Ev. The question now becomes: how does she choose her Ev?

As discussed in Chapter 1 of Smith (2010), there are three players in an inference problem, although two roles may be taken by the same person. There is the client, who has the problem, the statistician whom the client hires to help solve the problem, and the auditor whom the client hires to check the statistician's work. The statistician needs to be able to satisfy an auditor who asks about the logic of their approach. This chapter does not explain how to choose Ev; instead it describes some properties that 'Ev' might have. Some of these properties are self-evident, and to violate them would be hard to justify to an auditor. These properties are the DP (Principle 2), the TP (Principle 2), and the WCP (Principle 4). Other properties are not at all self-evident; the most important of these are the SLP (Principle 5), the SRP (Principle 8) and the SCP (Principle 9). These not self-evident properties would be extremely attractive, were it possible to justify them. And as we have seen, they can all be justified as logical deductions from the properties that are self-evident. This is the essence of Birnbaum's Theorem (Theorem 4).

For over a century, statisticians have been proposing methods for selecting algorithms for Ev, independently of this strand of research concerning the properties that such algorithms ought to have (remember that Birbaum's Theorem was published in 1962). Bayesian inference, which turns out to respect the SLP, is compatible with all of the properties given above, but classical inference, which turns out to violate the SLP, is not. The two main consequences of this violation are described in Section 2.8.

Now it is important to be clear about one thing. Ultimately, an inference is a single element in the space of 'possible inferences about θ'. An inference cannot be evaluated according to whether or not it satisfies the SLP. What is being evaluated in this chapter is the algorithm, the mechanism by which $\mathcal{E}$ and x are turned into an inference. It is quite possible that statisticians of quite different persuasions will produce effectively identical inferences from different algorithms. For example, if asked for a set estimate of θ, a Bayesian statistician might produce a 95% High Density Region, and a classical statistician a 95% confidence set, but they might be effectively the same set. But it is not the inference that is the primary concern of the auditor: it is the justification for the inference, among the uncountable other inferences that might have been made but weren't. The auditor checks the 'why', before passing the 'what' on to the client.

So the auditor will ask: why do you choose algorithm Ev? The classical statistician will reply, "Because it is a 95% confidence procedure for θ, and, among the uncountable

number of such procedures, this is a good choice [for some reasons that are then given]." The Bayesian statistician will reply "Because it is a 95% High Posterior Density region for θ for prior distribution $\pi(\theta)$, and among the uncountable number of prior distributions, $\pi(\theta)$ is a good choice [for some reasons that are then given]." Let's assume that the reasons are compelling, in both cases. The auditor has a follow-up question for the classicist but not for the Bayesian: "Why are you not concerned about violating the Likelihood Principle?" A well-informed auditor will know the theory of the previous sections, and the consequences of violating the SLP that are given in Section 2.8. For example, violating the SLP is either illogical or obtuse - neither of these properties are desirable in an applied statistician.

This is not an easy question to answer. The classicist may reply "Because it is important to me that I control my error rate over the course of my career", which is incompatible with the SLP. In other words, the statistician ensures that, by always using a 95% confidence procedure, the true value of θ will be inside at least 95% of her confidence sets, over her career. Of course, this answer means that the statistician puts her career error rate before the needs of her current client. I can just about imagine a client demanding "I want a statistician who is right at least 95% of the time." Personally, though, I would advise a client against this, and favour instead a statistician who is concerned not with her career error rate, but rather with the client's particular problem.

3 Statistical Decision Theory

3.1 Introduction

The basic premise of Statistical Decision Theory is that we want to make inferences about the parameter of a family of distributions in the statistical model

$$\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \,|\, \theta)\},$$

typically following observation of sample data, or information, x. We would like to understand how to construct the 'Ev' function from Chapter 2, in such a way that it reflects our needs, which will vary from application to application, and which assesses the consequences of making a good or bad inference.

The set of possible inferences, or decisions, is termed the ***decision space***, denoted $\mathcal{D}$. For each $d \in \mathcal{D}$, we want a way to assess the consequence of how good or bad the choice of decision d was under the event θ.

Definition 10 *(Loss function)*
A loss function is any function L from $\Theta \times \mathcal{D}$ to $[0, \infty)$.

The loss function is measures the penalty or error, $L(\theta, d)$ of the decision d when the parameter takes the value θ. Thus, larger values indicate worse consequences.

The three main types of inference about θ are (i) point estimation, (ii) set estimation, and (iii) hypothesis testing. It is a great conceptual and practical simplification that Statistical Decision Theory distinguishes between these three types simply according to their decision spaces, which are:

Type of inference	Decision space $\mathcal{D}$
Point estimation	The parameter space, Θ. See Section 3.4.
Set estimation	A set of subsets of Θ. See Section 3.5.
Hypothesis testing	A specified partition of Θ, denoted $\mathcal{H}$. See Section 3.6.

3.2 Bayesian statistical decision theory

In a Bayesian approach, a statistical decision problem $[\Theta, \mathcal{D}, \pi(\theta), L(\theta, d)]$ has the following ingredients.

1. The possible values of the parameter: Θ, the parameter space.
2. The set of possible decisions: $\mathcal{D}$, the decision space.
3. The probability distribution on Θ, $\pi(\theta)$. For example,
   (a) this could be a prior distribution, $\pi(\theta) = f(\theta)$.
   (b) this could be a posterior distribution, $\pi(\theta) = f(\theta \,|\, x)$ following the receipt of some data x.
   (c) this could be a posterior distribution $\pi(\theta) = f(\theta \,|\, x, y)$ following the receipt of some data x, y.
4. The loss function $L(\theta, d)$.

In this setting, only θ is random and we can calculate the expected loss, or risk.

Definition 11 *(Risk)*
The risk of decision $d \in \mathcal{D}$ under the distribution $\pi(\theta)$ is

$$\rho(\pi(\theta), d) \quad = \quad \int_{\theta} L(\theta, d)\pi(\theta)\, d\theta. \tag{3.1}$$

We choose d to minimise this risk.

Definition 12 *(Bayes rule and Bayes risk)*
The Bayes risk $\rho^*(\pi)$ minimises the expected loss,

$$\rho^*(\pi) \quad = \quad \inf_{d \in \mathcal{D}} \rho(\pi, d)$$

with respect to $\pi(\theta)$. A decision $d^* \in \mathcal{D}$ for which $\rho(\pi, d^*) = \rho^*(\pi)$ is a Bayes rule against $\pi(\theta)$.

The Bayes rule may not be unique, and in weird cases it might not exist. Typically, we solve $[\Theta, \mathcal{D}, \pi(\theta), L(\theta, d)]$ by finding $\rho^*(\pi)$ and (at least one) d^*.

Example 14 *Quadratic Loss. Suppose that $\Theta \subset \mathbb{R}$. We consider the loss function*

$$L(\theta, d) \quad = \quad (\theta - d)^2.$$

From (3.1), the risk of decision d is

$$\begin{aligned} \rho(\pi, d) &= \mathbb{E}\{L(\theta, d) \,|\, \theta \sim \pi(\theta)\} \\ &= \mathbb{E}_{(\pi)}\{(\theta - d)^2\} \\ &= \mathbb{E}_{(\pi)}(\theta^2) - 2d\mathbb{E}_{(\pi)}(\theta) + d^2, \end{aligned}$$

where $\mathbb{E}_{(\pi)}(\cdot)$ is a notational device to define the expectation computed using the distribution $\pi(\theta)$. Differentiating with respect to d we have

$$\frac{\partial}{\partial d}\rho(\pi, d) \quad = \quad -2\mathbb{E}_{(\pi)}(\theta) + 2d.$$

So, the Bayes rule $d^* = \mathbb{E}_{(\pi)}(\theta)$. The corresponding Bayes risk is

$$\begin{aligned} \rho^*(\pi) \;=\; \rho(\pi, d^*) \quad &= \quad \mathbb{E}_{(\pi)}(\theta^2) - 2d^*\mathbb{E}_{(\pi)}(\theta) + (d^*)^2 \\ &= \quad \mathbb{E}_{(\pi)}(\theta^2) - 2\mathbb{E}^2_{(\pi)}(\theta) + \mathbb{E}^2_{(\pi)}(\theta) \\ &= \quad \mathbb{E}_{(\pi)}(\theta^2) - \mathbb{E}^2_{(\pi)}(\theta) \\ &= \quad Var_{(\pi)}(\theta) \end{aligned}$$

where $Var_{(\pi)}(\theta)$ is the variance of θ computed using the distribution $\pi(\theta)$.

1. *If $\pi(\theta) = f(\theta)$, a prior for θ, then the Bayes rule of an immediate decision is $d^* = \mathbb{E}(\theta)$ with corresponding Bayes risk $\rho^* = Var(\theta)$.*

2. *If we observe sample data x then the Bayes rule given this sample information is $d^* = \mathbb{E}(\theta \,|\, X)$ with corresponding Bayes risk $\rho^* = Var(\theta \,|\, X)$ as $\pi(\theta) = f(\theta \,|\, x)$.*

Typically we can solve $[\Theta, \mathcal{D}, f(\theta), L(\theta, d)]$, the immediate decision problem, and solve $[\Theta, \mathcal{D}, f(\theta \,|\, x), L(\theta, d)]$, the decision problem after sample information. Often, we may be interested in the ***risk of the sampling procedure***, before observing the sample, to decide whether or not to sample. For each possible sample, we need to specify which decision to make. This gives us the idea of a ***decision rule***.

Definition 13 *(Decision rule)*
A decision rule $\delta(x)$ is a function from $\mathcal{X}$ into $\mathcal{D}$,

$$\delta : \mathcal{X} \to \mathcal{D}.$$

If $X = x$ is the observed value of the sample information then $\delta(x)$ is the decision that will be taken. The collection of all decision rules is denoted by Δ so that $\delta \in \Delta \Rightarrow \delta(x) \in \mathcal{D}\ \forall x \in X$.

In this case, we wish to solve the problem $[\Theta, \Delta, f(\theta, x), L(\theta, \delta(x))]$. In analogy to Definition 12, we make the following definition.

Definition 14 *(Bayes (decision) rule and risk of the sampling procedure)*
The decision rule δ^* is a Bayes (decision) rule exactly when

$$\mathbb{E}\{L(\theta, \delta^*(X))\} \quad \leq \quad \mathbb{E}\{L(\theta, \delta(X))\} \tag{3.2}$$

for all $\delta(x) \in \mathcal{D}$. The corresponding risk $\rho^* = \mathbb{E}\{L(\theta, \delta^*(X))\}$ is termed the risk of the sampling procedure.

If the sample information consists of $X = (X_1, \ldots, X_n)$ then ρ^* will be a function of n and so can be used to help determine sample size choice.

Theorem 8 *(Bayes rule theorem, BRT)*
Suppose that a Bayes rule exists[1] *for* $[\Theta, \mathcal{D}, f(\theta \mid x), L(\theta, d)]$. *Then*

$$\delta^*(x) = \arg\min_{d \in \mathcal{D}} \mathbb{E}(L(\theta, d) \mid X = x). \tag{3.3}$$

Proof: Let δ be arbitrary. Then

$$\begin{aligned}
\mathbb{E}\{L(\theta, \delta(X))\} &= \int_x \int_\theta L(\theta, \delta(x)) f(\theta, x)\, d\theta dx \\
&= \int_x \int_\theta L(\theta, \delta(x)) f(\theta \mid x) f(x)\, d\theta dx \\
&= \int_x \left\{ \int_\theta L(\theta, \delta(x)) f(\theta \mid x)\, d\theta \right\} f(x)\, dx \\
&= \int_x \mathbb{E}\{L(\theta, \delta(x)) \mid X\} f(x)\, dx
\end{aligned} \tag{3.4}$$

where, from (3.1), $\mathbb{E}\{L(\theta, \delta(x)) \mid X\} = \rho(f(\theta \mid x), \delta(x))$, the posterior risk. We want to find the Bayes decision function δ^* for which

$$\mathbb{E}\{L(\theta, \delta^*(X))\} = \inf_{\delta \in \Delta} \mathbb{E}\{L(\theta, \delta(X))\}.$$

From (3.4), as $f(x) \geq 0$, δ^* may equivalently be found as

$$\rho(f(\theta), \delta^*) = \inf_{\delta(x) \in \mathcal{D}} \mathbb{E}\{L(\theta, \delta(x)) \mid X\}, \tag{3.5}$$

giving equation (3.3). □

This astounding result indicates that the minimisation of expected loss over the space of all functions from $\mathcal{X}$ to $\mathcal{D}$ can be achieved by the pointwise minimisation over $\mathcal{D}$ of the expected loss conditional on $X = x$. It converts an apparently intractable problem into a simple one. We could consider Δ, the set of decision rules, to be our possible set of inferences about θ when the sample is observed so that $\mathrm{Ev}(\mathcal{E}, x)$ is $\delta^*(x)$. We thus have the following result.

Theorem 9 *The Bayes rule for the posterior decision respects the strong likelihood principle.*

Proof: If we have two Bayesian models with the *same* prior distribution, $\mathcal{E}_{B,1} = \{\mathcal{X}_1, \Theta, f_{X_1}(x_1 \mid \theta), \pi(\theta)\}$ and $\mathcal{E}_{B,2} = \{\mathcal{X}_2, \Theta, f_{X_2}(x_2 \mid \theta), \pi(\theta)\}$ then, as in (2.13), if $f_{X_1}(x_1 \mid \theta) = c(x_1, x_2) f_{X_2}(x_2 \mid \theta)$ then the corresponding posterior distributions $\pi(\theta \mid x_1)$ and $\pi(\theta \mid x_2)$ are the same and so the corresponding Bayes rule (and risk) is the same. □

3.3 Admissible rules

Bayes rules rely upon a prior distribution for θ: the risk, see Definition 11, is a function of d only. In classical statistics, there is no distribution for θ and so another approach is needed. This involves the classical risk.

[1] Finiteness of $\mathcal{D}$ ensures existence. Similar but more general results are possible, but they require more topological conditions to ensure a minimum occurs within $\mathcal{D}$.

Definition 15 *(The classical risk)*
For a decision rule $\delta(x)$, the classical risk for the model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$ is

$$R(\theta, \delta) \quad = \quad \int_X L(\theta, \delta(x)) f_X(x \mid \theta)\, dx.$$

The classical risk is thus, for each δ, a function of θ.

Example 15 *Let $X = (X_1, \ldots, X_n)$ where $X_i \sim N(\theta, \sigma^2)$ and σ^2 is known. Suppose that $L(\theta, d) = (\theta - d)^2$ and consider a conjugate prior $\theta \sim N(\mu_0, \sigma_0^2)$. Possible decision functions include:*

1. *$\delta_1(x) = \overline{x}$, the sample mean.*
2. *$\delta_2(x) = med\{x_1, \ldots, x_n\} = \tilde{x}$, the sample median.*
3. *$\delta_3(x) = \mu_0$, the prior mean.*
4. *$\delta_4(x) = \mu_n$, the posterior mean where*

$$\mu_n \quad = \quad \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2}\right),$$

   the weighted average of the prior and sample mean accorded to their respective precisions.

The respective classical risks are

1. *$R(\theta, \delta_1) = \frac{\sigma^2}{n}$, a constant for θ, since $\overline{X} \sim N(\theta, \sigma^2/n)$.*
2. *$R(\theta, \delta_2) = \frac{\pi\sigma^2}{2n}$, a constant for θ, since $\tilde{X} \sim N(\theta, \pi\sigma^2/2n)$ (approximately).*
3. *$R(\theta, \delta_3) = (\theta - \mu_0)^2 = \sigma_0^2 \left(\frac{\theta - \mu_0}{\sigma_0}\right)^2$.*
4. *$R(\theta, \delta_4) = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-2} \left\{\frac{1}{\sigma_0^2}\left(\frac{\theta - \mu_0}{\sigma_0}\right)^2 + \frac{n}{\sigma^2}\right\}$.*

Which decision do we choose? We observe that $R(\theta, \delta_1) < R(\theta, \delta_2)$ for all $\theta \in \Theta$ but other comparisons depend upon θ.

The accepted approach for classical statisticians is to narrow the set of possible decision rules by ruling out those that are obviously bad.

Definition 16 *(Admissible decision rule)*
A decision rule δ_0 is inadmissible if there exists a decision rule δ_1 which dominates it, that is

$$R(\theta, \delta_1) \quad \leq \quad R(\theta, \delta_0)$$

for all $\theta \in \Theta$ with $R(\theta, \delta_1) < R(\theta, \delta_0)$ for at least one value $\theta_0 \in \Theta$. If no such δ_1 exists then δ_0 is admissible.

If δ_0 is dominated by δ_1 then the classical risk of δ_0 is never smaller than that of δ_1 and δ_1 has a smaller risk for θ_0. Thus, you would never want to use δ_0.[2] Hence, the accepted approach is to reduce the set of possible decision rules under consideration by only using admissible rules. It is hard to disagree with this approach, although one wonders how big the set of admissible rules will be, and how easy it is to enumerate the set of admissible rules in order to choose between them. It turns out that admissible rules can be related to a Bayes rule δ^* for a prior distribution $\pi(\theta)$ (as given by Definition 13).

Theorem 10 *If a prior distribution $\pi(\theta)$ is strictly positive for all Θ with finite Bayes risk and the classical risk, $R(\theta, \delta)$, is a continuous function of θ for all δ, then the Bayes rule δ^* is admissible.*

Proof: We follow p75 of Robert (2007). Suppose that δ^* is inadmissible and dominated by δ_1 so that in an open set C of θ, $R(\theta, \delta_1) < R(\theta, \delta^*)$ with $R(\theta, \delta_1) \leq R(\theta, \delta^*)$ elsewhere. Then, in an analogous way to the proof of Theorem 8 but now writing $f(\theta, x) = f_X(x \mid \theta)\pi(\theta)$, for any decision rule δ,

$$\mathbb{E}\{L(\theta, \delta(X))\} \;=\; \int_\Theta R(\theta, \delta)\pi(\theta)\, d\theta.$$

Thus, if δ_1 dominates δ^* then $\mathbb{E}\{L(\theta, \delta_1(X))\} < \mathbb{E}\{L(\theta, \delta^*(X))\}$ which is a contradiction to δ^* being the Bayes rule. □

The relationship between a Bayes rule with prior $\pi(\theta)$ and an admissible decision rule is even stronger and described in the following very beautiful result, originally due to an iconic figure in Statistics, Abraham Wald.[3]

Theorem 11 *(Wald's Complete Class Theorem, CCT)*
In the case where the parameter space Θ and sample space $\mathcal{X}$ are finite, a decision rule δ is admissible if and only if it is a Bayes rule for some prior distribution $\pi(\theta)$ with strictly positive values.

An illuminating blackboard proof of this result can be found in Section 11.6 of Cox and Hinkley (1974). There are generalisations of this theorem to non-finite decision sets, parameter spaces, and sample spaces but the results are highly technical. See Chapter 3 of Schervish (1995), Chapters 4 and 8 of Berger (1985), and Chapter 2 of Ghosh and Meeden (1997) for more details and references to the original literature. In the rest of this section, we will assume the more general result, which is that a decision rule is admissible if and only if it is a Bayes rule for some prior distribution $\pi(\theta)$, which holds for practical purposes.

So what does the CCT say? First of all, admissible decision rules respect the SLP. This follows from the fact that admissible rules are Bayes rules which respect the SLP: see

[2] Here I am assuming that all other considerations are the same in the two cases: e.g. for all $x \in \mathcal{X}$, $\delta_1(x)$ and $\delta_0(x)$ take about the same amount of resource to compute.

[3] Abraham Wald (1902-1950)

Theorem 9. Insofar as we think respecting the SLP is a good thing, this provides support for using admissible decision rules, because we cannot be certain that inadmissible rules respect the SLP. Second, if you select a Bayes rule according to some positive prior distribution $\pi(\theta)$ then you cannot ever choose an inadmissible decision rule. So the CCT states that there is a very simple way to protect yourself from choosing an inadmissible decision rule.

But here is where you must pay close attention to logic. Suppose that δ' is inadmissible and δ is admissible. It does not follow that δ dominates δ'. So just knowing of an admissible rule does not mean that you should abandon your inadmissible rule δ'. You can argue that although you know that δ' is inadmissible, you do not know of a rule which dominates it. All you know, from the CCT, is the family of rules within which the dominating rule must live: it will be a Bayes rule for some positive $\pi(\theta)$. Statisticians sometimes use inadmissible rules. They can argue that yes, their rule δ' is or may be inadmissible, which is unfortunate, but since the identity of the dominating rule is not known, it is not wrong to go on using δ'. Do not attempt to explore this rather arcane line of reasoning with your client!

3.4 Point estimation

For point estimation the decision space is $\mathcal{D} = \Theta$, and the loss function $L(\theta, d)$ represents the (negative) consequence of choosing d as a point estimate of θ. There will be situations where an obvious loss function $L : \Theta \times \Theta \to \mathbb{R}$ presents itself. But not very often. Hence the need for a generic loss function which is acceptable over a wide range of situations. A natural choice in the very common case where Θ is a convex subset of $\mathbb{R}^p$ is a *convex loss function*,

$$L(\theta, d) \quad = \quad h(d - \theta)$$

where $h : \mathbb{R}^p \to \mathbb{R}$ is a smooth non-negative convex function with $h(0) = 0$. This type of loss function asserts that small errors are much more tolerable than large ones. One possible further restriction would be that h is an even function, $h(d - \theta) = h(\theta - d)$ so that $L(\theta, \theta + \epsilon) = L(\theta, \theta - \epsilon)$ so that under-estimation incurs the same loss as over-estimation.

As we saw in Example 14, the (univariate) quadratic loss function $L(\theta, d) = (\theta - d)^2$ has attractive features and is also, in terms of the classical risk, related to the MSE. As we will see, this result generalises to $\mathbb{R}^p$ in a similar way.

There are many situations where this is *not* appropriate and the loss function should be asymmetric and a generic loss function should be replaced by a more specific one.

Example 16 *(Bilinear loss)*
The bilinear loss function for $\Theta \subset \mathbb{R}$ is, for $\alpha, \beta > 0$,

$$L(\theta, d) \quad = \quad \begin{cases} \alpha(\theta - d) & \text{if } d \leq \theta, \\ \beta(d - \theta) & \text{if } d \geq \theta. \end{cases}$$

The Bayes rule is a $\frac{\alpha}{\alpha+\beta}$-fractile of $\pi(\theta)$.

Note that if $\alpha = \beta = 1$ then $L(\theta, d) = |\theta - d|$, the **absolute loss** which gives a Bayes rule of the median of $\pi(\theta)$. $|\theta - d|$ is smaller that $(\theta - d)^2$ for $|\theta - d| > 1$ and so absolute loss is smaller than quadratic loss for large deviations. Thus, it takes less account of the tails of $\pi(\theta)$ leading to the choice of the median. The choice of α and β can account for asymmetry. If $\alpha > \beta$, so $\frac{\alpha}{\alpha+\beta} > 0.5$, then under-estimation is penalised more than over-estimation and so that Bayes rule is more likely to be an over-estimate.

Example 17 *(Example 2.1.2 of Robert (2007))*
Suppose X is distributed as the p-dimensional normal distribution with mean θ and known variance matrix Σ which is diagonal with diagonal elements σ_i^2 for each $i = 1, \ldots, p$. Then $\mathcal{D} = \mathbb{R}^p$. We might consider a loss function of the form

$$L(\theta, d) \quad = \quad \sum_{i=1}^{p} \left(\frac{d_i - \theta_i}{\sigma_i} \right)^2$$

so that the total loss is the sum of the squared component-wise errors.

In this case, we observe that if $Q = \Sigma^{-1}$ then the loss function is a form of quadratic loss which we generalise in the following example.

Example 18 *If $\Theta \in \mathbb{R}^p$, the Bayes rule δ^* associated with the prior distribution $\pi(\theta)$ and the quadratic loss*

$$L(\theta, d) \quad = \quad (d - \theta)^T Q \, (d - \theta)$$

is the posterior expectation $\mathbb{E}(\theta \,|\, X)$ for every positive-definite symmetric $p \times p$ matrix Q.

Thus, as the Bayes rule does not depend upon Q, it is the same for an uncountably large class of loss functions. If we apply the Complete Class Theorem, Theorem 11, to this result we see that for quadratic loss, a point estimator for θ is admissible if and only if it is the conditional expectation with respect to some positive prior distribution $\pi(\theta)$. The value, and interpretability, of the quadratic loss can be further observed by noting that, from a Taylor series expansion, an even, differentiable and strictly convex loss function can be approximated by a quadratic loss function.

3.4.1 Stein's example

Let $X = (X_1, \ldots, X_p)^T$ and suppose that $X \,|\, \theta \sim N_p(\theta, I_p)$ where I_p is the $p \times p$ identity matrix and $\theta = (\theta_1, \ldots, \theta_p)^T$ is a vector of parameters. Thus, given θ, the X_is are independent $N(\theta_i, 1)$. Suppose we consider a single observation, $X = x$. The likelihood for θ is

$$L_X(\theta; x) \quad = \quad \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left\{ -(x_i - \theta_i)^2 \right\}.$$

so that the maximum likelihood estimate is x. The corresponding maximum likelihood estimator is X which is unbiased.

Estimation under quadratic loss

We consider point estimation of θ using quadratic loss. Following Examples 17 and 18, $Q = \Sigma^{-1} = I_p$, we have

$$L(\theta, d) \;=\; (d-\theta)^T(d-\theta) \;=\; \sum_{i=1}^{p}(d_i-\theta_i)^2$$

for decision $d = (d_1, \ldots, d_p)^T \in \mathbb{R}^p$. We consider the decision rule $\delta_0(X) = X$, the maximum likelihood estimator. From Definition 15, the classical risk of δ_0 is

$$\begin{aligned} R(\theta, \delta_0) &= \mathbb{E}[L(\theta, \delta_0(X)) \,|\, \theta] \\ &= \sum_{i=1}^{p} \mathbb{E}[(\theta_i - X_i)^2 \,|\, \theta] \\ &= \sum_{i=1}^{p} Var(X_i \,|\, \theta) \;=\; p. \end{aligned}$$

We'll show that, for $p \geq 3$, δ_0 is inadmissible by finding a decision rule which dominates it. Consider the set of James-Stein estimators

$$\delta_a(X) \;=\; \left(1 - \frac{a}{X^T X}\right) X$$

for $a \geq 0$. Notice that $a = 0$ gives $\delta_0(X) = X$ and that for $a > 0$, $\delta_a(X)$ is biased. If $X^T X > 0$ then $\delta_a(X)$ shrinks X towards 0.

Lemma 1 *(Stein's Lemma)*
If $X \,|\, \theta \sim N_p(\theta, I_p)$ and $g(X)$ a suitably behaved real valued function then

$$\mathbb{E}(g(X)(X_i - \theta_i) \,|\, \theta) \;=\; \mathbb{E}\left[\frac{\partial g(X)}{\partial X_i} \,\middle|\, \theta\right].$$

Proof: First note that,

$$\mathbb{E}(g(X)(X_i - \theta_i) \,|\, \theta) \;=\; \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x)(x_i - \theta_i) \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left\{-(x_j - \theta_j)^2\right\} dx_1 \ldots dx_p$$

Now, for $j = i$, using integration by parts,

$$\begin{aligned} \int_{-\infty}^{\infty} g(x)(x_i - \theta_i) \exp\left\{-(x_i - \theta_i)^2\right\} dx_i &= \int_{-\infty}^{\infty} \frac{\partial g(x)}{\partial x_i} \exp\left\{-(x_i - \theta_i)^2\right\} dx_i + \\ &\quad \left[-g(x) \exp\left\{-(x_i - \theta_i)^2\right\}\right]_{-\infty}^{\infty} \\ &= \int_{-\infty}^{\infty} \frac{\partial g(x)}{\partial x_i} \exp\left\{-(x_i - \theta_i)^2\right\} dx_i \end{aligned}$$

for suitable $g(x)$. Consequently,

$$\begin{aligned} \mathbb{E}(g(X)(X_i - \theta_i) \,|\, \theta) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial g(x)}{\partial x_i} \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left\{-(x_j - \theta_j)^2\right\} dx_1 \ldots dx_p \\ &= \mathbb{E}\left[\frac{\partial g(X)}{\partial X_i} \,\middle|\, \theta\right] \end{aligned}$$

which completes the proof. □

We now calculate the classical risk of δ_a under quadratic loss. We have

$$\begin{aligned} R(\theta,\delta_a) &= \mathbb{E}[(\theta-\delta_a(X))^T(\theta-\delta_a(X))\,|\,\theta] \\ &= \mathbb{E}\left[\left((\theta-X)+\frac{aX}{X^TX}\right)^T\left((\theta-X)+\frac{aX}{X^TX}\right)\,\middle|\,\theta\right] \\ &= \mathbb{E}[(\theta-X)^T(\theta-X)\,|\,\theta]+a^2\mathbb{E}\left[\frac{1}{X^TX}\,\middle|\,\theta\right] \\ &\quad -2a\mathbb{E}\left[\frac{X^T(X-\theta)}{X^TX}\,\middle|\,\theta\right] \\ &= R(\theta,\delta_0)+a^2\mathbb{E}\left[\frac{1}{X^TX}\,\middle|\,\theta\right]-2a\sum_{i=1}^{p}\mathbb{E}\left[\frac{X_i(X_i-\theta_i)}{X^TX}\,\middle|\,\theta\right] \end{aligned}$$

Now, using Stein's Lemma with $g(X)=X_i/X^TX$ we have

$$\begin{aligned} \sum_{i=1}^{p}\mathbb{E}\left[\frac{X_i}{X^TX}(X_i-\theta_i)\,\middle|\,\theta\right] &= \sum_{i=1}^{p}\mathbb{E}\left[\frac{\partial}{\partial X_i}\frac{X_i}{X^TX}\,\middle|\,\theta\right] \\ &= \sum_{i=1}^{p}\mathbb{E}\left[\frac{X^TX-2X_i^2}{(X^TX)^2}\,\middle|\,\theta\right] \\ &= \mathbb{E}\left[\frac{pX^TX-2\sum_{i=1}^{p}X_i^2}{(X^TX)^2}\,\middle|\,\theta\right] \\ &= (p-2)\mathbb{E}\left[\frac{1}{X^TX}\,\middle|\,\theta\right] \end{aligned}$$

since $\sum_{i=1}^{p}X_i^2=X^TX$. Thus,

$$R(\theta,\delta_a) = R(\theta,\delta_0)+(a^2-2a(p-2))\mathbb{E}\left[\frac{1}{X^TX}\,\middle|\,\theta\right].$$

Now, $\mathbb{E}[1/X^TX\,|\,\theta]>0$ and thus if $a^2-2a(p-2)<0$ then

$$R(\theta,\delta_a) < R(\theta,\delta_0).$$

Hence, if $0<a<2(p-2)$ δ_0 is inadmissible as, for all θ, $R(\theta,\delta_a)<R(\theta,\delta_0)$. Consequently, for $p\geq 3$ the maximum likelihood estimate δ_0 is inadmissible.

Notice that $a=p-2$ minimises $R(\theta,\delta_a)$ and that if $\theta=0$ then $X^TX\sim\chi^2_p$, the chi-squared distribution with p degrees of freedom. Thus, $\mathbb{E}[1/X^TX\,|\,\theta=0]=1/(p-2)$. Recalling that $R(\theta,\delta_0)=p$ then $R(0,\delta_{p-2})=2$ and so when p is large, this will be much smaller than the corresponding risk of δ_0.

The ith term of $\delta_a(X)=\left(1-\frac{a}{X^TX}\right)X$ is $\left(1-\frac{a}{X^TX}\right)X_i$ and so depends on all $X_1,\ldots,X_p$ even though the X_is are independent. As Young and Smith (2005) on p35 comment: "at first sight, the result seems incredible: there is no apparent 'tying together' of the losses, yet the obvious estimator, the sample 'mean' X, is not admissible." Indeed, the result caused

such consternation when first published, see Stein (1956) and James and Stein (1961), that it might be termed 'Stein's bombshell' and, it can be shown, to occur in many situations beyond normal means with known common variance when comparing three or more populations. See Efron and Morris (1977) for more details, and Samworth (2012) for an accessible overview. Note that whilst its admissibility under quadratic loss is questionable, the MLE remains the dominant point estimator in applied statistics.

3.5 Set estimation

For set estimation the decision space is a set of subsets of Θ so that each $d \subset \Theta$. There are two contradictory requirements for set estimators of Θ. We want the sets to be small, but we also want them to contain θ. There is a simple way to represent these two requirements as a loss function, which is to use

$$L(\theta, d) \quad = \quad |d| + \kappa(1 - \mathbb{1}_{\theta \in d}) \tag{3.6}$$

for some $\kappa > 0$ where $|d|$ is the volume of d. The value of κ controls the trade-off between the two requirements. If $\kappa \downarrow 0$ then minimising the expected loss will always produce the empty set. If $\kappa \uparrow \infty$ then minimising the expected loss will always produce Θ. For κ in-between, the Bayes rule will depend on beliefs about X and the value x. For loss functions of the form (3.6) there is a a simple necessary condition for a rule to be a Bayes rule. A set $d \subset \Theta$ is a *level set* of the posterior distribution exactly when $d = \{\theta : \pi(\theta \,|\, x) \geq k\}$ for some k.

Theorem 12 *(Level set property, LSP)*
If δ^* is a Bayes rule for the loss function in (3.6) then it is a level set of the posterior distribution.

Proof: For fixed x, we show that if d is not a level set of the posterior distribution then there is a $d' \neq d$ which has a smaller expected loss so that $\delta^*(x) \neq d$. Note that

$$\mathbb{E}\{L(\theta, d) \,|\, X\} \quad = \quad |d| + \kappa \mathbb{P}(\theta \notin d \,|\, X). \tag{3.7}$$

Suppose that d is not a level set of $\pi(\theta \,|\, x)$. Then there is a $\theta \in d$ and $\theta' \notin d$ for which $\pi(\theta' \,|\, x) > \pi(\theta \,|\, x)$. Let $d' = d \cup d\theta' \setminus d\theta$ where $d\theta$ is the tiny region of Θ around θ and $d\theta'$ is the tiny region of Θ around θ' for which $|d\theta| = |d\theta'|$. Then $|d'| = |d|$ but

$$\mathbb{P}(\theta \notin d' \,|\, X) < \mathbb{P}(\theta \notin d \,|\, X)$$

Thus, from equation (3.7), $\mathbb{E}\{L(\theta, d') \,|\, X\} < \mathbb{E}\{L(\theta, d) \,|\, X\}$ showing that $\delta^*(x) \neq d$. □

Now relate this result to the CCT (Theorem 11). First, Theorem 12 asserts that δ having the LSP is necessary (but not sufficient) for δ to be a Bayes rule for loss functions of the form (3.6). Second, the CCT asserts that being a Bayes rule is a necessary (but not sufficient)

condition for δ to be admissible. So, being a level set of a posterior distribution for some prior distribution $\pi(\theta)$ is a necessary condition for being admissible for loss functions of the form (3.6). Bayesian HPD regions satisfy the necessary condition for being a set estimator whilst classical set estimators achieve a similar outcome if they are level sets of the likelihood function, because the posterior is proportional to the likelihood under a uniform prior distribution.[4]

3.6 Hypothesis tests

For hypothesis tests, the decision space is a partition of Θ, denoted

$$\mathcal{H} \quad := \quad \{H_0, H_1, \ldots, H_d\}.$$

Each element of $\mathcal{H}$ is termed a ***hypothesis***; it is traditional to number the hypotheses from zero. The loss function $L(\theta, H_i)$ represents the (negative) consequences of choosing element H_i, when the true value of Θ is θ. It would be usual for the loss function to satisfy

$$\theta \in H_i \implies L(\theta, H_i) \quad = \quad \min_j L(\theta, H_j)$$

on the grounds that an incorrect choice of element should never incur a smaller loss than the correct choice. There is a generic loss function for hypothesis tests: the ***0-1 ('zero-one') loss function***

$$L(\theta, H_i) = 1 - \mathbb{1}_{\{\theta \in H_i\}},$$

i.e., zero if θ is in H_i, and one if it is not. The corresponding Bayes rule is to select the hypothesis with the largest posterior probability.

Its arguable about why the 0-1 loss function would approximate a wide range of actual loss functions and an alternative approach has proved more popular. This is to co-opt the theory of set estimators, for which there *is* a defensible generic loss function, which has strong implications for the selection of decision rules (see Section 3.5). The statistician can use her set estimator δ to make at least some distinctions between the members of $\mathcal{H}$:

- 'Accept' H_i exactly when $\delta(x) \subset H_i$,
- 'Reject' H_i exactly when $\delta(x) \cap H_i = \emptyset$,
- 'Undecided' about H_i otherwise.

Note that these three terms are given in quotes, to indicate that they acquire a technical meaning in this context. We do not use the quotes in practice, but we always bear in mind that we are not "accepting H_i" in the vernacular sense, but simply asserting that $\delta(x) \subset H_i$ for our particular choice of δ.

[4] In the case where Θ is unbounded, this prior distribution may have to be truncated to be proper.

4 Confidence sets and p-values

4.1 Confidence procedures and confidence sets

We consider interval estimation, or more generally set estimation. Consider the model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \,|\, \theta)\}$. For given data $X = x$, we wish to construct a set $C = C(x) \subset \Theta$ and the inference is the statement that $\theta \in C$. If $\theta \in \mathbb{R}$ then the set estimate is typically an interval. As Casella and Berger (2002) note in Section 9.1, the goal of a set estimator is to have some guarantee of capturing the parameter of interest. With this in mind, we make the following definition.

Definition 17 *(Confidence procedure)*
A random set $C(X)$ is a level-$(1-\alpha)$ confidence procedure exactly when

$$\mathbb{P}(\theta \in C(X) \,|\, \theta) \quad \geq \quad 1 - \alpha$$

for all $\theta \in \Theta$. C is an exact *level-$(1-\alpha)$ confidence procedure if the probability equals $(1-\alpha)$ for all θ.*

Thus, exact is a special case and typically $\mathbb{P}(\theta \in C(X) \,|\, \theta)$ will depend upon θ. The value $\mathbb{P}(\theta \in C(X) \,|\, \theta)$ is termed the ***coverage*** of C at θ. Thus a 95% confidence procedure has coverage of at least 95% for all θ, and an exact 95% confidence procedure has coverage of exactly 95% for all θ. If it is necessary to emphasise that C is not exact, then the term ***conservative*** is used.

Example 19 *Let $X_1, \ldots, X_n$ be independent and identically distributed Unif$(0, \theta)$ random variables where $\theta > 0$. Let $Y = \max\{X_1, \ldots, X_n\}$. For observed $x_1, \ldots, x_n$, we have that $\theta > y$. Noting that $X_i/\theta \sim$ Unif$(0,1)$ then if $T = Y/\theta$ we have that $\mathbb{P}(T \leq t) = t^n$ for $0 \leq t \leq 1$. We consider two possible sets: (aY, bY) where $1 \leq a < b$ and $(Y + c, Y + d)$ where $0 \leq c < d$. Notice that*

$$\begin{aligned} \mathbb{P}(\theta \in (aY, bY) \,|\, \theta) &= \mathbb{P}(aY < \theta < bY \,|\, \theta) \\ &= \mathbb{P}(b^{-1} < T < a^{-1} \,|\, \theta) \\ &= \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n. \end{aligned}$$

Thus, the coverage probability of the interval does not depend upon θ. However,

$$\begin{aligned}\mathbb{P}(\theta \in (Y+c, Y+d) \,|\, \theta) &= \mathbb{P}(Y+c < \theta < Y+d \,|\, \theta) \\ &= \mathbb{P}\left(1-\frac{d}{\theta} < T < 1-\frac{c}{\theta} \,|\, \theta\right) \\ &= \left(1-\frac{c}{\theta}\right)^n - \left(1-\frac{d}{\theta}\right)^n.\end{aligned}$$

In this case, the coverage probability of the interval does depend upon θ.

It is helpful to distinguish between the confidence procedure C, which is a random interval and so a function for each possible x, and the result when C is evaluated at the observation x, which is a set in Θ. We follow the terms used in Morey et al. (2016), which we will later adapt to p-values, see for example Definition 24.

Definition 18 *(Confidence set)*
The observed $C(x)$ is a level-$(1-\alpha)$ confidence set exactly when the random $C(X)$ is a level-$(1-\alpha)$ confidence procedure.

If $\Theta \subset \mathbb{R}$ and $C(x)$ is convex, i.e. an interval, then a confidence set (interval) is represented by a lower and upper value. We should write, for example, "using procedure C, the 95% confidence interval for θ is $(0.55, 0.74)$", inserting "exact" if the confidence procedure C is exact.

The challenge with confidence procedures is to construct one with a specified level. One could propose an arbitrary C, and then laboriously compute the coverage for every $\theta \in \Theta$. At that point we would know the level of C as a confidence procedure, but it is unlikely to be 95%; adjusting C and iterating this procedure many times until the minimum coverage was equal to 95% would be exceedingly tedious. So we need to go backwards: start with the level, e.g. 95%, then construct a C guaranteed to have this level. With this in mind, we can generalise Definition 17.

Definition 19 *(Family of confidence procedures)*
$C(X;\alpha)$ is a family of confidence procedures exactly when $C(X;\alpha)$ is a level-$(1-\alpha)$ confidence procedure for every $\alpha \in [0,1]$. C is a nesting family exactly when $\alpha < \alpha'$ implies that $C(x;\alpha') \subset C(x;\alpha)$.

If we start with a family of confidence procedures for a specified model, then we can compute a confidence set for any level we choose.

4.2 Constructing confidence procedures

The general approach to construct a confidence procedure is to invert a test statistic. In Example 19, the coverage of the procedure (aY, bY) does not depend upon θ because the coverage probability could be expressed in terms of $T = Y/\theta$ where the distribution of T

did not depend upon θ. T is an example of a ***pivot***. As Example 19 shows, confidence procedures are straightforward to compute from a pivot. However, a drawback to this approach in general is that there is no hard and fast method for finding a pivot.

An alternate method which does work generally is to exploit the property that every confidence procedure corresponds to a hypothesis test and vice versa. Consider a hypothesis test where we have to decide either to accept that an hypothesis H_0 is true or to reject H_0 in favour of an alternative hypothesis H_1 based on a sample $x \in \mathcal{X}$. The set of x for which H_0 is rejected is called the ***rejection region*** with its complement, where H_0 is accepted, the ***acceptance region***. A hypothesis test can be constructed from any statistic $T = T(X)$, one popular method which is optimal in some cases is the likelihood ratio test.

Definition 20 *(Likelihood Ratio Test, LRT)*
The likelihood ratio test (LRT) statistic for testing $H_0 : \theta \in \Theta_0$ *versus* $H_1 : \theta \in \Theta_0^c$, *where* $\Theta_0 \cup \Theta_0^c = \Theta$, *is*

$$\lambda(x) \quad = \quad \frac{\sup_{\theta \in \Theta_0} L_X(\theta; x)}{\sup_{\theta \in \Theta} L_X(\theta; x)}. \tag{4.1}$$

A LRT at significance level α *has a rejection region of the form* $\{x : \lambda(x) \leq c\}$ *where* $0 \leq c \leq 1$ *is chosen so that* $\mathbb{P}(\text{Reject } H_0 \,|\, \theta) \leq \alpha$ *for all* $\theta \in \Theta_0$.

Example 20 *Let* $X = (X_1, \ldots, X_n)$ *and suppose that the* X_i *are independent and identically distributed* $N(\theta, \sigma^2)$ *random variables where* σ^2 *is known and consider the likelihood ratio test for* $H_0 : \theta = \theta_0$ *versus* $H_1 : \theta \neq \theta_0$. *Then, as the maximum likelihood estimate of* θ *is* $\overline{x}$,

$$\begin{aligned} \lambda(x) &= \frac{L_X(\theta_0; x)}{L_X(\overline{x}; x)} \\ &= \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left((x_i - \theta_0)^2 - (x_i - \overline{x})^2 \right) \right\} \\ &= \exp\left\{ -\frac{1}{2\sigma^2} n(\overline{x} - \theta_0)^2 \right\}. \end{aligned}$$

Notice that, under H_0, $\frac{\sqrt{n}(\overline{X} - \theta_0)}{\sigma} \sim N(0, 1)$ *so that*

$$-2 \log \lambda(X) \quad = \quad \frac{n(\overline{X} - \theta_0)^2}{\sigma^2} \sim \chi^2_1, \tag{4.2}$$

the chi-squared distribution with one degree of freedom. Letting $\chi^2_{1,\alpha}$ *be such that* $\mathbb{P}(\chi^2_1 \geq \chi^2_{1,\alpha}) = \alpha$ *then, as the rejection region* $\{x : \lambda(x) \leq c\}$ *corresponds to* $\{x : -2 \log \lambda(x) \geq k\}$ *where* $k = -2 \log c$, *setting* $k = \chi^2_{1,\alpha}$ *gives a test at the (exact) significance level* α. *The corresponding acceptance region of this test is* $\{x : -2 \log \lambda(x) < \chi^2_{1,\alpha}\}$ *where*

$$\mathbb{P}\left(\frac{n(\overline{X} - \theta_0)^2}{\sigma^2} < \chi^2_{1,\alpha} \,\middle|\, \theta = \theta_0 \right) \quad = \quad 1 - \alpha. \tag{4.3}$$

This holds for all θ_0 *and so, additionally rearranging,*

$$\mathbb{P}\left(\overline{X} - \sqrt{\chi^2_{1,\alpha}} \frac{\sigma}{\sqrt{n}} < \theta < \overline{X} + \sqrt{\chi^2_{1,\alpha}} \frac{\sigma}{\sqrt{n}} \,\middle|\, \theta \right) \quad = \quad 1 - \alpha. \tag{4.4}$$

Thus, $C(X) = (\overline{X} - \sqrt{\chi^2_{1,\alpha}}\frac{\sigma}{\sqrt{n}}, \overline{X} + \sqrt{\chi^2_{1,\alpha}}\frac{\sigma}{\sqrt{n}})$ *is an exact level-*$(1-\alpha)$ *confidence procedure with* $C(x)$ *the corresponding confidence set. Noting that* $\sqrt{\chi^2_{1,\alpha}} = z_{\alpha/2}$*, where* $z_{\alpha/2}$ *is such that* $\mathbb{P}(Z \geq z_{\alpha/2}) = \alpha/2$ *for* $Z \sim N(0,1)$*, this confidence set is more familiarly written as* $C(x) = (\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$.

The level-$(1-\alpha)$ confidence procedure defined by equation (4.4) is obtained by inverting the acceptance region, see equation (4.3), of the level α significance test. This correspondence between acceptance regions of tests and confidence sets is a general property.

Theorem 13 *(Duality of Acceptance Regions and Confidence Sets)*

1. For each $\theta_0 \in \Theta$*, let* $A(\theta_0)$ *be the acceptance region of a test of* $H_0 : \theta = \theta_0$ *at significance level* α*. For each* $x \in \mathcal{X}$*, define* $C(x) = \{\theta_0 : x \in A(\theta_0)\}$*. Then* $C(X)$ *is a level-*$(1-\alpha)$ *confidence procedure.*

2. Let $C(X)$ *be a level-*$(1-\alpha)$ *confidence procedure and, for any* $\theta_0 \in \Theta$*, define* $A(\theta_0) = \{x : \theta_0 \in C(x)\}$*. Then* $A(\theta_0)$ *is the acceptance region of a test of* $H_0 : \theta = \theta_0$ *at significance level* α.

Proof: 1. As we have a level α test for each $\theta_0 \in \Theta$ then $\mathbb{P}(X \in A(\theta_0) \,|\, \theta = \theta_0) \geq 1 - \alpha$. Since θ_0 is arbitrary we may write θ instead of θ_0 and so, for all $\theta \in \Theta$,

$$\mathbb{P}(\theta \in C(X) \,|\, \theta) \;=\; \mathbb{P}(X \in A(\theta) \,|\, \theta) \;\geq 1 - \alpha.$$

Hence, from Definition 17, $C(X)$ is a level-$(1-\alpha)$ confidence procedure.
2. For a test of $H_0 : \theta = \theta_0$, the probability of a Type I error (rejecting H_0 when it is true) is

$$\mathbb{P}(X \notin A(\theta_0) \,|\, \theta = \theta_0) \;=\; \mathbb{P}(\theta_0 \notin C(X), \,|\, \theta = \theta_0) \;\leq\; \alpha$$

since $C(X)$ is a level-$(1-\alpha)$ confidence procedure. Hence, we have a test at significance level α. □

A possibly easier way to understand the relationship between significance tests and confidence sets is by defining the set $\{(x,\theta) : (x,\theta) \in \tilde{C}\}$ in the space $\mathcal{X} \times \Theta$ where $\tilde{C}$ is also a set in $\mathcal{X} \times \Theta$.

- For fixed x, we may define the confidence set as $C(x) = \{\theta : (x,\theta) \in \tilde{C}\}$.
- For fixed θ, we may define the acceptance region as $A(\theta) = \{x : (x,\theta) \in \tilde{C}\}$.

Example 21 *We revisit Example 20 and, recalling that* $x = (x_1, \ldots, x_n)$*, define the set*

$$\{(x,\theta) : (x,\theta) \in \tilde{C}\} \quad = \quad \left\{(x,\theta) : -z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{x} - \theta < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}.$$

The confidence set is then

$$C(x) = \left\{\theta : -z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{x} - \theta < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = \left\{\theta : \overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta < \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}$$

and acceptance region

$$A(\theta) = \left\{ x \,:\, -z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{x} - \theta < z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right\} = \left\{ x \,:\, \theta - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{x} < \theta + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right\}.$$

4.3 Good choices of confidence procedures

Section 3.5 made a recommendation about set estimators for θ, which was that they should be based on level sets of $f_X(x \,|\, \theta)$. This was to satisfy a necessary condition to be admissible under the loss function (3.6). With this in mind, a good choice of confidence procedure would be one that satisfied a level set property.

Definition 21 *(Level set property, LSP)*
A confidence procedure C has the level set property exactly when

$$C(x) \quad = \quad \{\theta \,:\, f_X(x \,|\, \theta) > g(x)\}$$

for some $g : \mathcal{X} \to \mathbb{R}$.

We now show that we can construct a family of confidence procedures with the LSP. The result has pedagogic value, because it can be used to generate an uncountable number of families of confidence procedures, each with the LSP.

Theorem 14 *Let h be any probability density function for X. Then*

$$C_h(x; \alpha) \quad := \quad \{\theta \in \Theta \,:\, f_X(x \,|\, \theta) > \alpha h(x)\} \tag{4.5}$$

is a family of confidence procedures, with the LSP.

Proof: First notice that if we let $\mathcal{X}(\theta) := \{x \in \mathcal{X} \,:\, f_X(x \,|\, \theta) > 0\}$ then

$$\begin{aligned} \mathbb{E}(h(X)/f_X(X \,|\, \theta) \,|\, \theta) \quad &= \quad \int_{x \in \mathcal{X}(\theta)} \frac{h(x)}{f_X(x \,|\, \theta)} f_X(x \,|\, \theta)\, dx \\ &= \quad \int_{x \in \mathcal{X}(\theta)} h(x)\, dx \\ &\leq \quad 1 \end{aligned} \tag{4.6}$$

because h is a probability density function. Now,

$$\mathbb{P}(f_X(X \,|\, \theta)/h(X) \leq u \,|\, \theta) \quad = \quad \mathbb{P}(h(X)/f_X(X \,|\, \theta) \geq 1/u \,|\, \theta) \tag{4.7}$$

$$\leq \quad \frac{\mathbb{E}(h(X)/f_X(X \,|\, \theta) \,|\, \theta)}{1/u} \tag{4.8}$$

$$\leq \quad \frac{1}{1/u} = u \tag{4.9}$$

where (4.8) follows from (4.7) by Markov's inequality[1] and (4.9) from (4.8) by (4.6). □

[1] If X is a nonnegative random variable and $a > 0$ then $\mathbb{P}(X \geq a) \leq \mathbb{E}(X)/a$.

Notice that if we define $g(x, \theta) := f_X(x \,|\, \theta)/h(x)$, which may be ∞ then the proof shows that $\mathbb{P}(g(X, \theta) \leq u \,|\, \theta) \leq u$. As we will see in Definition 23 this means that $g(X, \theta)$ is ***super-uniform*** for each θ.

Among the interesting choices for h, one possibility is $h(x) = f_X(x \,|\, \theta_0)$, for some $\theta_0 \in \Theta$. Note that with this choice, the confidence set of equation (4.5) always contains θ_0. So we know that we can construct a level-$(1 - \alpha)$ confidence procedure whose confidence sets will always contain θ_0. Two statisticians can both construct 95% confidence sets for θ which satisfy the LSP, using different families of confidence procedures. Yet the first statistician may reject the null hypothesis that $H_0 : \theta = \theta_0$ (see Section 3.6), and the second statistician may fail to reject it, for any $\theta_0 \in \Theta$. This does not fill one with confidence about using confidence procedures for hypothesis tests.

Actually, the situation is not as grim as it seems. Markov's inequality is very slack, and so the coverage of the family of confidence procedures defined in Theorem 14 is likely to be much larger than $(1 - \alpha)$, e.g. much larger than 95%.

For any confidence procedure, the diameter[2] of $C(x)$ can grow rapidly with its coverage. In fact, the relation must be extremely convex when coverage is nearly one, because, in the case where $\Theta = \mathbb{R}$, the diameter at 100% coverage is unbounded. So an increase in the coverage from, say 95% to 99%, could easily correspond to a doubling or more of the diameter of the confidence procedure.

A more likely outcome in the two statisticians situation is that $C_h(x; 0.05)$ is large for many different choices of h, in which case no one rejects the null hypothesis, which is not a useful outcome for a hypothesis test. But perhaps it is a useful antidote to the current 'crisis of reproducibility', in which far too many null hypotheses are being rejected in published papers.

All in all, it would be much better to use an exact family of confidence procedures which satisfy the LSP, if one existed. And, for perhaps the most popular model in the whole of Statistics, this is the case. This is the linear model with a normal error.

4.3.1 The linear model

We briefly discuss the linear model and, in what can be viewed as an extension of Example 20, consider constructing a confidence procedure using the likelihood ratio. Wood (2017) is a recommended textbook discussion of the whole (generalised) theory.

Let $Y = (Y_1, \ldots, Y_n)$ be an n-vector of observables with $Y = X\theta + \epsilon$, where $\mu = X\theta$ is an $(n \times p)$ matrix X[3] of ***regressors***, θ is a p-vector of ***regression coefficients***, and ϵ is an

[2] The diameter of a set in a metric space such as Euclidean space is the maximum of the distance between two points in the set.

[3] We typically use X to denote a generic random variable and so it is not ideal to use it here for a specified

n-vector of ***residuals***. Assume that $\epsilon \sim N_n(0, \sigma^2 I_n)$, the n-dimensional multivariate normal distribution, where σ^2 is known.

We will utilise the following two properties of the multivariate normal distribution.

Theorem 15 *(Properties of the multivariate normal distribution)*
Let $W = (W_1, \ldots, W_k)$ with $W \sim N_k(\mu, \Sigma)$, the k-dimensional multivariate normal distribution with mean vector μ and variance matrix Σ.

1. If $Y = AW + c$, where A is any $(l \times k)$ matrix and c any l-dimensional vector, then $Y \sim N_l(A\mu + c, A\Sigma A^T)$.

2. If $\Sigma > 0$ then $Y = \Sigma^{-\frac{1}{2}}(W - \mu) \sim N_k(0, I_k)$, where I_k is the $(k \times k)$ identity matrix, and $(W-\mu)^T\Sigma^{-1}(W-\mu) = \sum_{i=1}^k y_i^2 \sim \chi^2_k$.

Proof: See for example, Theorem 3.2.1 and Corollary 3.2.1.1 of Mardia et al. (1979). □

It is thus immediate from the first property that for the linear model, $Y \sim N_n(\mu, \sigma^2 I_n)$ where $\mu = X\theta$. Now,

$$L_Y(\theta; y) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta)\right\}. \tag{4.10}$$

Let $\hat{\theta} = \hat{\theta}(y) = \left(X^TX\right)^{-1}X^Ty$ then

$$\begin{aligned}(y - X\theta)^T(y - X\theta) &= (y - X\hat{\theta} + X\hat{\theta} - X\theta)^T(y - X\hat{\theta} + X\hat{\theta} - X\theta) \\ &= (y - X\hat{\theta})^T(y - X\hat{\theta}) + (X\hat{\theta} - X\theta)^T(X\hat{\theta} - X\theta) \\ &= (y - X\hat{\theta})^T(y - X\hat{\theta}) + (\hat{\theta} - \theta)^TX^TX(\hat{\theta} - \theta). \end{aligned} \tag{4.11}$$

Thus, $(y - X\theta)^T(y - X\theta)$ is minimised when $\theta = \hat{\theta}$ and so, from equation (4.10), $\hat{\theta} = \left(X^TX\right)^{-1}X^Ty$ is the maximum likelihood estimator of θ. From equation (4.1), we can calculate the likelihood ratio

$$\begin{aligned}\lambda(y) &= \frac{L_Y(\theta; y)}{L_Y(\hat{\theta}; y)} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left[(y - X\theta)^T(y - X\theta) - (y - X\hat{\theta})^T(y - X\hat{\theta})\right]\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}(\hat{\theta} - \theta)^TX^TX(\hat{\theta} - \theta)\right\}, \end{aligned} \tag{4.12}$$

where equation (4.12) follows from equation (4.11). Thus,

$$-2\log\lambda(y) = \frac{1}{\sigma^2}(\hat{\theta} - \theta)^TX^TX(\hat{\theta} - \theta).$$

Now, as $\hat{\theta}(Y) = \left(X^TX\right)^{-1}X^TY$ then, from Property 1. of Theorem 15,

$$\hat{\theta}(Y) \sim N_p\left(\theta, \sigma^2\left(X^TX\right)^{-1}\right)$$

matrix but this is the standard notation for linear models.

so that, from Property 2. of Theorem 15, $-2\log\lambda(Y) \sim \chi^2_p$. Hence, with $\mathbb{P}(\chi^2_p \geq \chi^2_{p,\alpha}) = \alpha$,

$$\begin{aligned} C(y;\alpha) &= \left\{\theta \in \mathbb{R}^p \,:\, -2\log\lambda(y) = -2\log\frac{f_Y(y\,|\,\theta,\sigma^2)}{f_Y(y\,|\,\hat{\theta},\sigma^2)} < \chi^2_{p,\alpha}\right\} \\ &= \left\{\theta \in \mathbb{R}^p \,:\, f_Y(y\,|\,\theta,\sigma^2) > \exp\left(-\frac{\chi^2_{p,\alpha}}{2}\right) f_Y(y\,|\,\hat{\theta},\sigma^2)\right\} \end{aligned} \tag{4.13}$$

is a family of exact confidence procedures for θ which has the LSP.

4.3.2 Wilks confidence procedures

This outcome where we can find a family of exact confidence procedures with the LSP is more-or-less unique to the regression parameters of the linear model but it is found, approximately, in the large n behaviour of a much wider class of models. The result can be traced back to Wilks (1938) and, as such, the resultant confidence procedures are often termed Wilks confidence procedures.

Theorem 16 *(Wilks Theorem)*
Let $X = (X_1, \ldots, X_n)$ where each X_i is independent and identically distributed, $X_i \sim f(x_i\,|\,\theta)$,where f is a regular model and the parameter space Θ is an open convex subset of $\mathbb{R}^p$ (and invariant to n). The distribution of the statistic $-2\log\lambda(X)$ converges to a chi-squared distribution with p degrees of freedom as $n \to \infty$.

The definition of 'regular model' is quite technical, but a working guideline is that f must be smooth and differentiable in θ; in particular, the support must not depend on θ. Chapter 6 of Cox (2006) provides a summary of this result and others like it, and more details can be found in Chapter 10 of Casella and Berger (2002) or, for the full story, in van der Vaart (1998). Analogous to equation (4.13), we thus have that if the conditions of Theorem 16 are met,

$$C(x;\alpha) = \left\{\theta \in \mathbb{R}^p \,:\, f_X(x\,|\,\theta) > \exp\left(-\frac{\chi^2_{p,\alpha}}{2}\right) f_X(x\,|\,\hat{\theta})\right\} \tag{4.14}$$

is a family of approximately exact confidence procedures which satisfy the LSP. The pertinent question, as always with methods based on asymptotic properties for particular types of model, is whether the approximation is a good one. The crucial concept here is ***level error***. The coverage that we want is at least $(1-\alpha)$ everywhere, which is termed the 'nominal level'. But were we to evaluate a confidence procedure such as (4.14) for a general model (not a linear model) we would find that, over all $\theta \in \Theta$, that the minimum coverage was not $(1-\alpha)$ but something else; usually something less than $(1-\alpha)$. This is the 'actual level'. The difference is

$$\text{level error} = \text{nominal level} - \text{actual level}.$$

Level error exists because the conditions under which (4.14) provides an exact confidence procedure are not met in practice, outside the linear model. Although it is tempting to ignore level error, experience suggests that it can be large, and that we should attempt to correct for level error if we can. One method for making this correction is ***bootstrap calibration***, described in DiCiccio and Efron (1996).

4.4 Significance procedures and duality

A hypothesis test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$, where $\Theta_0 \cup \Theta_0^c = \Theta$, with a significance level of 5% (or any other specified value) returns one bit of information, either we 'accept H_0' or 'reject H_0'. We do not know whether the decision was borderline or nearly conclusive; i.e. whether, for rejection, H_0 and $C(x; 0.05)$ were close, or well-separated. Of more interest is to consider what is the smallest value of α for which $C(x; \alpha)$ does not intersect H_0. This value is termed the p-value.

Definition 22 *(p-value)*
A p-value $p(X)$ is a statistic satisfying $p(x) \in [0, 1]$ for every $x \in \mathcal{X}$. Small values of $p(x)$ support the hypothesis that H_1 is true. A p-value is valid if, for every $\theta \in \Theta_0$ and every $\alpha \in [0, 1]$,

$$\mathbb{P}(p(X) \leq \alpha \,|\, \theta) \quad \leq \quad \alpha. \tag{4.15}$$

If $p(X)$ is a valid p-value then a ***significance test*** that rejects H_0 if and only if $p(X) \leq \alpha$ is, from (4.15), a test with ***significance level*** α. In this section we introduce the idea of significance procedures and derive a duality between a significance procedure at level α and a confidence procedure at level $1 - \alpha$. We first need some additional concepts. Let X and Y be two scalar random variables. Then X ***stochastically dominates*** Y exactly when

$$\mathbb{P}(X \leq v) \quad \leq \quad \mathbb{P}(Y \leq v)$$

for all $v \in \mathbb{R}$. Visually, the distribution function for X is never to the left of the distribution function for Y.[4] Recall that if $U \sim \text{Unif}(0, 1)$, the standard uniform distribution, then $\mathbb{P}(U \leq u) = u$ for $u \in [0, 1]$. With this in mind, we make the following definition.

Definition 23 *(Super-uniform)*
The random variable X is super-uniform *exactly when it stochastically dominates a standard uniform random variable. That is*

$$\mathbb{P}(X \leq u) \quad \leq \quad u \tag{4.16}$$

for all $u \in [0, 1]$.

Example 22 *From Definition 22, we see that for $\theta \in \Theta_0$, the p-value $p(X)$ is super-uniform.*

[4] Recollect that the distribution function of X has the form $F(x) := \mathbb{P}(X \leq x)$ for $x \in \mathbb{R}$.

We now define a significance procedure which can be viewed as an extension of Definition 22. Note the similarities with the definitions of a confidence procedure which are not coincidental.

Definition 24 *(Significance procedure)*

1. $p : \mathcal{X} \to \mathbb{R}$ *is a significance procedure for* $\theta_0 \in \Theta$ *exactly when* $p(X)$ *is super-uniform under* θ_0*. If* $p(X)$ *is uniform under* θ_0*, then* p *is an exact significance procedure for* θ_0*.*

2. For $X = x$*,* $p(x)$ *is a significance level or (observed) p-value for* θ_0 *exactly when* p *is a significance procedure for* θ_0*.*

3. $p : \mathcal{X} \times \Theta \to \mathbb{R}$ *is a family of significance procedures exactly when* $p(x; \theta_0)$ *is a significance procedure for* θ_0 *for every* $\theta_0 \in \Theta$*.*

We now show that there is a duality between significance procedures and confidence procedures.

Theorem 17 *(Duality theorem)*

1. Let p *be a family of significance procedures. Then*

$$C(x; \alpha) \quad := \quad \{\theta \in \Theta \, : \, p(x; \theta) > \alpha\}$$

is a nesting family of confidence procedures.

2. Conversely, let C *be a nesting family of confidence procedures. Then*

$$p(x; \theta_0) \quad := \quad \inf\{\alpha \, : \, \theta_0 \notin C(x; \alpha)\}$$

is a family of significance procedures.
If either is exact, then the other is exact as well.

Proof: If p is a family of significance procedures then for any $\theta \in \Theta$,

$$\begin{aligned} \mathbb{P}(\theta \in C(X; \alpha) \,|\, \theta) &= \mathbb{P}(p(X; \theta) > \alpha \,|\, \theta) \\ &= 1 - \mathbb{P}(p(X; \theta) \leq \alpha \,|\, \theta). \end{aligned} \tag{4.17}$$

Now, as p is super-uniform for θ then $\mathbb{P}(p(X; \theta) \leq \alpha \,|\, \theta) \leq \alpha$. Thus, from equation (4.17),

$$\mathbb{P}(\theta \in C(X; \alpha) \,|\, \theta) \quad \geq \quad 1 - \alpha \tag{4.18}$$

so that, from Definition 17, $C(X; \alpha)$ is a level-$(1-\alpha)$ confidence procedure. From Definition 19 it is clear that C is nesting. If p is exact then the inequality in (4.18) can be replaced by an equality and so C is also exact. We thus have 1. Now, if C is a nesting family of confidence procedures then[5]

$$\inf\{\alpha \, : \, \theta_0 \notin C(x; \alpha)\} \leq u \quad \iff \quad \theta_0 \notin C(x; u).$$

[5]Here we're finessing the issue of the boundary of C by assuming that if $\alpha^* := \inf\{\alpha \, : \, \theta_0 \notin C(x; \alpha)\}$ then $\theta_0 \notin C(x; \alpha^*)$.

Let θ_0 and $u \in [0,1]$ be arbitrary. Then,

$$\mathbb{P}(p(X;\theta_0) \leq u \,|\, \theta_0) \;=\; \mathbb{P}(\theta_0 \notin C(X;u) \,|\, \theta_0) \leq u$$

as $C(X;u)$ is a level-$(1-u)$ confidence procedure. Thus, p is super-uniform. If C is exact, then the inequality is replaced by an equality, and hence p is exact as well. □

Theorem 17 shows that confidence procedures and significance procedures are two sides of the same coin. If we have a way of constructing families of confidence procedures then we have a way of constructing families significance procedures, and vice versa. If we have a good way of constructing confidence procedures then (presumably, and in principle) we have a good way of constructing significance procedures. This is helpful because, as Section 4.5 will show, there are an uncountable number of families of significance procedures, and so there are an uncountable number of families of confidence procedures. Naturally, in both these cases, almost all of the possible procedures are useless for our inference. So just being a confidence procedure, or just being a significance procedure, is never enough. We need to know how to make good choices.

4.5 Families of significance procedures

We now consider a very general way to construct a family of significance procedures. We will then show how to use simulation to compute the family.

Theorem 18 *Let $t : \mathcal{X} \to \mathbb{R}$ be a statistic. For each $x \in \mathcal{X}$ and $\theta_0 \in \Theta$ define*

$$p_t(x;\theta_0) \quad := \quad \mathbb{P}(t(X) \geq t(x) \,|\, \theta_0).$$

Then p_t is a family of significance procedures. If the distribution function of $t(X)$ is continuous, then p_t is exact.

Proof: We follow Theorem 8.3.27 of Casella and Berger (2002). Now,

$$\begin{aligned} p_t(x;\theta_0) &= \mathbb{P}(t(X) \geq t(x) \,|\, \theta_0) \\ &= \mathbb{P}(-t(X) \leq -t(x) \,|\, \theta_0). \end{aligned}$$

Let F denote the distribution function of $Y(X) = -t(X)$ then

$$p_t(x;\theta_0) \;=\; F(-t(x) \,|\, \theta_0).$$

If $t(X)$ is continuous then $Y(X) = -t(X)$ is continuous and, using the Probability Integral Transform, see Theorem 23,

$$\begin{aligned} \mathbb{P}(p_t(X;\theta_0) \leq \alpha \,|\, \theta_0) &= \mathbb{P}(F(Y) \leq \alpha \,|\, \theta_0) \\ &= \mathbb{P}(Y \leq F^{-1}(\alpha) \,|\, \theta_0) = F(F^{-1}(\alpha)) = \alpha. \end{aligned}$$

Hence, p_t is uniform under θ_0. If $t(X)$ is not continuous then, via the Probability Integral Transform, $\mathbb{P}(F(Y) \leq \alpha \,|\, \theta_0) \leq \alpha$ and so $p_t(X;\theta_0)$ is super-uniform under θ_0. □

So there is a family of significance procedures for each possible function $t : \mathcal{X} \to \mathbb{R}$. Clearly only a tiny fraction of these can be useful functions, and the rest must be useless. Some, like $t(x) = c$ for some constant c, are always useless. Others, like $t(x) = \sin(x)$ might sometimes be a little bit useful, while others, like $t(x) = \sum_i x_i$ might be quite useful - but it all depends on the circumstances. Some additional criteria are required to separate out good from poor choices of the test statistic t, when using the construction in Theorem 18. The most pertinent criterion is:

- Select a test statistic for which $t(X)$ which will tend to be larger for decision-relevant departures from θ_0.

Example 23 *For the likelihood ratio, $\lambda(x)$, given by equation (4.1), small observed values of $\lambda(x)$ support departures from θ_0. Thus, $t(X) = -2\log\lambda(X)$, is a test statistic for which large values support departures from θ_0.*

In the context of Definition 22, large values of $t(X)$ will correspond to small values of the p-value, supporting the hypothesis that H_1 is true. Thus, this criterion ensures that $p_t(X;\theta_0)$ will tend to be smaller under decision-relevant departures from θ_0; small p-values are more interesting, precisely because significance procedures are super-uniform under θ_0.

4.5.1 Computing p-values

Only in very special cases will it be possible to find a closed-form expression for p_t from which we can compute the p-value $p_t(x;\theta_0)$. Instead,we can use simulation, according to the following result (adapted from Besag and Clifford (1989)).

Theorem 19 *For any finite sequence of scalar random variables $X_0, X_1, \ldots, X_m$, define the rank of X_0 in the sequence as*

$$R \quad := \quad \sum_{i=1}^{m} \mathbb{1}_{\{X_i \leq X_0\}}.$$

If $X_0, X_1, \ldots, X_m$ are exchangeable[6] then R has a discrete uniform distribution on the integers $\{0, 1, \ldots, m\}$, and $(R+1)/(m+1)$ has a super-uniform distribution.

Proof: By exchangeability, X_0 has the same probability of having rank r as any of the other X_i's, for any r, and therefore

$$\mathbb{P}(R = r) \quad = \quad \frac{1}{m+1} \tag{4.19}$$

[6]If $X_0, X_1, \ldots, X_m$ are exchangeable then their joint density function satisfies $f(x_0, \ldots, x_m) = f(x_{\pi(0)}, \ldots, x_{\pi(m)})$ for all permutations π defined on the set $\{0, \ldots, m\}$.

for $r \in \{0, 1, \ldots, m\}$ and zero otherwise, proving the first claim. For the second claim,

$$\begin{aligned} \mathbb{P}\left(\frac{R+1}{m+1} \leq u\right) &= \mathbb{P}(R+1 \leq u(m+1)) \\ &= \mathbb{P}(R+1 \leq \lfloor u(m+1) \rfloor) \end{aligned}$$

since R is an integer and $\lfloor x \rfloor$ denotes the largest integer no larger than x. Hence,

$$\begin{aligned} \mathbb{P}\left(\frac{R+1}{m+1} \leq u\right) &= \sum_{r=0}^{\lfloor u(m+1) \rfloor - 1} \mathbb{P}(R = r) && (4.20) \\ &= \sum_{r=0}^{\lfloor u(m+1) \rfloor - 1} \frac{1}{m+1} && (4.21) \\ &= \frac{\lfloor u(m+1) \rfloor}{m+1} \leq u, \end{aligned}$$

as required where equation (4.21) follows from (4.20) by (4.19). □

To use this result, fix the test statistic $t(x)$ and define $T_i = t(X_i)$ where $X_1, \ldots, X_m$ are independent and identically distributed random variables with density $f(\cdot \,|\, \theta_0)$. Define

$$R_t(x; \theta_0) := \sum_{i=1}^{m} \mathbb{1}_{\{-T_i \leq -t(x)\}} = \sum_{i=1}^{m} \mathbb{1}_{\{T_i \geq t(x)\}},$$

where θ_0 is an argument to R because θ_0 needs to be specified in order to simulate $T_1, \ldots, T_m$. Then Theorem 19 implies that

$$P_t(x; \theta_0) := \frac{R_t(x; \theta_0) + 1}{m+1}$$

has a super-uniform distribution under $X \sim f(\cdot \,|\, \theta_0)$, because in this case $t(X), T_1, \ldots, T_m$ are exchangeable. Furthermore, the Weak Law of Large Numbers (WLLN) implies that

$$\begin{aligned} \lim_{m \to \infty} P_t(x; \theta_0) &= \lim_{m \to \infty} \frac{R_t(x; \theta_0) + 1}{m+1} \\ &= \lim_{m \to \infty} \frac{R_t(x; \theta_0)}{m} \\ &= \mathbb{P}(T \geq t(x) \,|\, \theta_0) = p_t(x; \theta_0). \end{aligned}$$

Therefore, not only is $P_t(x; \theta_0)$ super-uniform under θ_0, so that P_t is a family of significance procedures for every m, but the limiting value of $P_t(x; \theta_0)$ as m becomes large is $p_t(x; \theta_0)$.

In summary, if you can simulate from your model under θ_0 then you can produce a p-value for any test statistic t, namely $P_t(x; \theta_0)$, and if you can simulate cheaply, so that the number of simulations m is large, then $P_t(x; \theta_0) \approx p_t(x; \theta_0)$.

The less-encouraging news is that this simulation-based approach is not well-adapted to constructing confidence sets. Let C_t be the family of confidence procedures induced by p_t

using duality, see Theorem 17. We can answer the question 'Is $\theta_0 \in C_t(x;\alpha)$?' with one set of m simulations. These simulations give a value $P_t(x;\theta_0)$ which is either larger or not larger than α. If $P_t(x;\theta_0) > \alpha$ then $\theta_0 \in C_t(x;\alpha)$, and otherwise it is not. Clearly, though, this is not an effective way to enumerate all of the points in $C_t(x;\alpha)$, because we would need to do m simulations for each point in Θ.

4.6 Generalisations

So far, confidence procedures and significance procedures have been defined with respect to a point $\theta_0 \in \Theta$. Often, though, we require a more general treatment, where a confidence procedure is defined for some $g : \theta \mapsto \phi$, where g may not be bijective; or where a significance procedure is defined for some $\Theta_0 \subset \Theta$, where Θ_0 may not be a single point. These general treatments are always possible, but the result is often very conservative. As discussed at the end of Section 4.3, conservative procedures are formally correct but they can be practically useless.

4.6.1 Marginalisation of confidence procedures

Suppose that $g : \theta \mapsto \phi$ is some specified function, and we would like a confidence procedure for ϕ. If C is a level-$(1-\alpha)$ confidence procedure for ϕ then it must have ϕ-coverage of at least $(1-\alpha)$ for all $\theta \in \Theta$. The most common situation is where $\Theta \subset \mathbb{R}^p$, and g extracts a single component of θ: for example, $\theta = (\mu, \sigma^2)$ and $g(\theta) = \mu$.

Theorem 20 *(Confidence Procedure Marginalisation, CPM)*
Suppose that $g : \theta \mapsto \phi$, and that C is a level-$(1-\alpha)$ procedure for θ. Then

$$gC \quad := \quad \{\phi \,:\, \phi = g(\theta) \text{ for some } \theta \in C\}$$

is a level-$(1-\alpha)$ confidence procedure for ϕ.

Proof: The result follows immediately by noting that $\theta \in C(x)$ implies that $\phi \in gC(x)$ for all x, and hence

$$\mathbb{P}(\theta \in C(X) \,|\, \theta) \quad \leq \quad \mathbb{P}(\phi \in gC(X) \,|\, \theta)$$

for all $\theta \in \Theta$. So if C has θ-coverage of at least $(1-\alpha)$, then gC has ϕ-coverage of at least $(1-\alpha)$ as well. □

This result shows that we can derive level-$(1-\alpha)$ confidence procedures for functions of θ directly from level-$(1-\alpha)$ confidence procedures for θ. Furthermore, if the confidence procedure for θ is easy to enumerate, then the confidence procedure for ϕ is easy to enumerate too - just by transforming each element. But it also shows that the coverage of such derived procedures will typically be more than $(1-\alpha)$, even if the original confidence procedure

is exact: thus gC is a conservative confidence procedure. As already noted, conservative confidence procedures can often be far larger than they need to be: sometimes too large to be useful.

4.6.2 Generalisation of significance procedures

We now give a simple result which extends a family of significance procedures over a set in Θ.

Theorem 21 *Let $\Theta_0 \subset \Theta$. If p is a family of significance procedures, then*

$$P(x;\Theta_0) \quad := \quad \sup_{\theta_0 \in \Theta_0} p(x;\theta_0)$$

is super-uniform for all $\theta \in \Theta_0$.

Proof: $P(x;\Theta_0) \leq u$ implies that $p(x;\theta_0) \leq u$ for all $\theta_0 \in \Theta$. Let $\theta \in \Theta_0$ be arbitrary; then, for any $u \geq 0$,

$$\mathbb{P}(P(X;\Theta_0) \leq u \,|\, \theta) \;\leq\; \mathbb{P}(p(X;\Theta_0) \leq u \,|\, \theta) \;\leq\; u$$

for $\theta \in \Theta_0$, showing that $P(x;\Theta_0)$ is super-uniform for all $\theta \in \Theta_0$. □

As with the marginalisation of confidence procedures, this result shows that we can derive a significance procedure for an arbitrary $\Theta_0 \subset \Theta$. The difference, though, is that this is rather impractical, because of the need, in general, to maximise over a possibly unbounded set Θ_0. As a result, this type of p-value is not much used in practice. It is sometimes replaced by simple approximations. For example, if the parameter is (v, θ) then a p-value for v_0 could be approximated by plugging-in a specific value for θ, such as the maximum likelihood value, and treating the model as though it were parameterised by v alone. But this does not give rise to a well defined significance procedure for v_0 on the basis of the original model. Adopting this type of approach is something of an act of desperation, for when Theorem 21 is intractable. The difficulty is that you get a number, but you do not know what it signifies.

4.7 Reflections

4.7.1 On the definitions

The first thing to note is the abundance of families of confidence procedures and significance procedures, most of which are useless. For example, let U be a uniform random quantity. Based on the definition alone,

$$C(x;\alpha) \quad = \quad \begin{cases} \{0\} & U < \alpha \\ \Theta & U \geq \alpha \end{cases}$$

is a perfectly acceptable family of exact confidence procedures, and

$$p(x;\theta_0) \quad = \quad U$$

is a perfectly acceptable family of exact significance procedures. They are both useless. You cannot object that these examples are pathological because they contain the auxiliary random quantity U, because the most accessible method for computing p-values also contains auxiliary random quantities (see Section 4.5.1). You could object that the family of significance procedures does not have the LSP property (Definition 21), which is a valid objection if you intend to apply the LSP rigorously. But would you then have to insist that every significance procedure's dual confidence procedure (see Theorem 17) should also have the LSP?

The second thing to note is how often confidence procedures and significance procedures will be conservative. This means that there is some region of the parameter space where the actual coverage of the confidence procedure is more than the nominal coverage of $(1-\alpha)$. Or where the significance procedure has a super-uniform but not uniform distribution under θ_0. As shown in this chapter:

- A generic method for constructing families of confidence procedures with the LSP (see Theorem 14) is always conservative.
- Confidence procedures for non-bijective functions of the parameters are always conservative (see Theorem 20).
- Significance procedures based on test statistics where $t(X)$ is discrete are always conservative (see Theorem 18).
- Significance procedures for composite hypotheses are always conservative (see Theorem 21).

4.7.2 On the interpretations

It is a very common observation, made repeatedly over the last 50 years see, for example, Rubin (1984), that clients think more like Bayesians than classicists. Classical statisticians have to wrestle with the issue that their clients will likely misinterpret their results. This is bad enough for confidence sets (see, e.g., Morey et al. (2016)), but potentially disastrous for p-values.

A p-value $p(x;\theta_0)$ refers only to θ_0, making no reference at all to other hypotheses about θ. But a posterior probability $\pi(\theta_0 \mid x)$ contrasts θ_0 with other values in Θ which θ might have taken. The two outcomes can be radically different, as first captured in Lindley's paradox, Lindley (1957), see also Bartlett (1957). A p-value can be viewed as measuring the fit of a model, that under H_0, to the observed data. A large p-value indicates only that the data is not unusual under the model but it does not imply that the model is correct. For example,

there may be many other models defined by other hypotheses which may be exhibit greater consistency with the observed data. Greenland et al. (2016) discuss 25 misinterpretations of p-values, confidence intervals, and power. Wasserstein and Lazar (2016) is a statement from the American Statistical Association (ASA) on statistical significance and p-values. The statement gives six principles for the correct use and interpretation of p-values, some of which, in particular Principles 3 and 4, are applicable more generally and dovetail to Smith (2010)'s view of there being three players in an inference problem: the client, the statistician, and the auditor. We state them here as values that should be at the heart of any work that we do.

Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

- Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making.
- Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis.

Proper inference requires full reporting and transparency.

- Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis.
- Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p-values computed.

4.8 Appendix: The Probability Integral Transform

Here is a very elegant and useful piece of probability theory. Let X be a scalar random variable with sample space $\mathcal{X}$ and distribution function $F(x) := \mathbb{P}(X \leq x)$. By convention, F is defined for all $x \in \mathbb{R}$. By construction, $\lim_{x \downarrow -\infty} F(x) = 0$, $\lim_{x \uparrow \infty} F(x) = 1$, F is non-decreasing, and F is continuous from the right, i.e.

$$\lim_{x' \downarrow x} F(x') \quad = \quad F(x).$$

Define the ***quantile function***

$$F^{-}(u) \quad := \quad \inf\{x \in \mathbb{R} : F(x) \geq u\}. \tag{4.22}$$

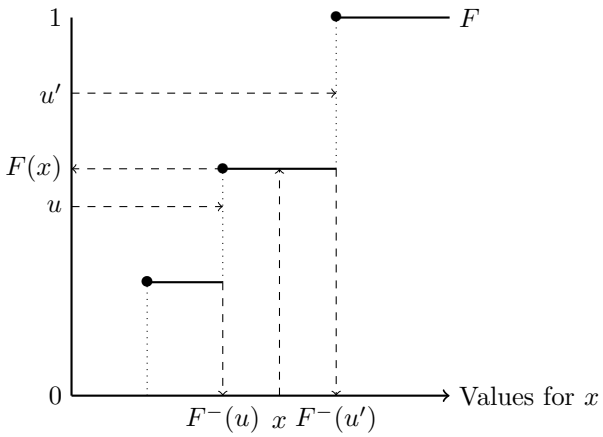

Figure 4.1: Figure for the proof of Theorem 22. The distribution function F is non-decreasing and continuous from the right. The quantile function F^- is defined in equation (4.22).

The following result is a cornerstone of generating random variables with easy-to-evaluate quantile functions.

Theorem 22 *(Probability Integral Transform, PIT)*
Let U have a standard uniform distribution. If F^- is the quantile function of X, then $F^-(U)$ and X have the same distribution.

Proof: Let F be the distribution function of X. We must show that

$$F^-(u) \leq x \quad \Longleftrightarrow \quad u \leq F(x) \tag{4.23}$$

because then

$$\mathbb{P}(F^-(U) \leq x) \;=\; \mathbb{P}(U \leq F(x)) \;=\; F(x)$$

as required. Now, see Figure 4.1, it is easy to check that

$$u \leq F(x) \quad \Longrightarrow \quad F^-(u) \leq x,$$

which is one half of equation (4.23). It is also easy to check that

$$u' > F(x) \quad \Longrightarrow \quad F^-(u') > x.$$

Taking the contrapositive of this second implication gives

$$F^-(u') \leq x \quad \Longrightarrow \quad u' \leq F(x),$$

which is the other half of equation (4.23). □

Theorem 22 is the basis for the following result; recollect the definition of a super-uniform random quantity from Definition 23. This result is used in the proof of Theorem 18.

Theorem 23 *If F is the distribution function of X, then $F(X)$ has a super-uniform distribution. If F is continuous then $F(X)$ has a uniform distribution.*

Proof: As we can see from Figure 4.1, $F(F^-(u)) \geq u$. Then, from Theorem 22,

$$\begin{aligned} \mathbb{P}(F(X) \leq u) &= \mathbb{P}(F(F^-(U)) \leq u) \\ &\leq \mathbb{P}(U \leq u) \\ &= u. \end{aligned}$$

In the case where F is continuous, it is strictly increasing except on sets which have probability zero. Then

$$\mathbb{P}(F(X) \leq u) \;=\; \mathbb{P}(F(F^-(U)) \leq u) \;=\; \mathbb{P}(U \leq u) \;=\; u,$$

as required. □

Bibliography

Bartlett, M. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika 44*, 533–534.

Basu, D. (1975). Statistical information and likelihood. *Sankhyā 37*(1), 1–71. With discussion.

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis* (second ed.). NY, USA: Springer-Verlag New York, Inc.

Berger, J. and R. Wolpert (1988). *The Likelihood Principle* (second ed.). Hayward CA, USA: Institute of Mathematical Statistics. Available online, http://projecteuclid.org/euclid.lnms/1215466210.

Bernardo, J. and A. Smith (2000). *Bayesian Theory*. Chichester, UK: John Wiley & Sons Ltd. (paperback edition, first published 1994).

Besag, J. and P. Clifford (1989). Generalized Monte Carlo significance tests. *Biometrika 76*(4), 633–642.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association 57*, 269–306.

Birnbaum, A. (1972). More concepts of statistical evidence. *Journal of the American Statistical Association 67*, 858–861.

Box, G. (1979). Robustness in the strategy of scientific model building. In R. Launer and G. Wilkinson (Eds.), *Robustness in Statistics*, pp. 201–236. Academic Press, New York, USA.

Casella, G. and R. Berger (2002). *Statistical Inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Cox, D. (2006). *Principles of Statistical Inference*. Cambridge, UK: Cambridge University Press.

Cox, D. and D. Hinkley (1974). *Theoretical Statistics*. London, UK: Chapman and Hall.

Davison, A. (2003). *Statistical Models*. Cambridge, UK: Cambridge University Press.

Dawid, A. (1977). Conformity of inference patterns. In J. Barra et al. (Eds.), *Recent Developments in Statistcs*. Amsterdam: North-Holland Publishing Company.

DiCiccio, T. and B. Efron (1996). Bootstrap confidence intervals. *Statistical Science 11*(3), 189–212. with discussion and rejoinder, 212–228.

Efron, B. and T. Hastie (2016). *Computer Age Statistical Inference*. New York NY, USA: Cambridge University Press.

Efron, B. and C. Morris (1977). Stein's paradox in statistics. *Scientific American 236*(5), 119–127. Available at http://statweb.stanford.edu/~ckirby/brad/other/Article1977.pdf.

Fisher, R. (1956). *Statistical Methods and Scientific Inference*. Edinburgh and London: Oliver and Boyd.

Ghosh, M. and G. Meeden (1997). *Bayesian Methods for Finite Population Sampling*. London, UK: Chapman & Hall.

Greenland, S., S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology 31*, 337–350.

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 361–380. University of California Press.

Lindley, D. (1957). A statistical paradox. *Biometrika 44*, 187–192. See also Bartlett (1957).

MacKay, D. (2009). *Sustainable Energy – Without the Hot Air*. Cambridge, UK: UIT Cambridge Ltd. available online, at http://www.withouthotair.com/.

Mardia, K., J. Kent, and J. Bibby (1979). *Multivariate Analysis*. London, UK: Harcourt Brace & Co.

Morey, R., R. Hoekstra, J. Rouder, M. Lee, and E.-J. Wagenmakers (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bullentin & Review 23*(1), 103–123.

O'Hagan, A. and J. Forster (2004). *Bayesian Inference* (2nd ed.), Volume 2b of *Kendall's Advanced Theory of Statistics*. London: Edward Arnold.

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York, USA: Springer.

Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics 12*(4), 1151–1172.

Samworth, R. (2012). Stein's paradox. *Eureka 62*, 38–41. Available online at http://www.statslab.cam.ac.uk/~rjs57/SteinParadox.pdf. Careful readers will spot a typo in the maths.

Savage, L. et al. (1962). *The Foundations of Statistical Inference*. London, UK: Methuen.

Schervish, M. (1995). *Theory of Statistics*. New York NY, USA: Springer. Corrected 2nd printing, 1997.

Smith, J. (2010). *Bayesian Decision Analysis: Principle and Practice*. Cambridge, UK: Cambridge University Press.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 197–206. University of California Press.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

Wasserstein, R. and N. Lazar (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician 70*(2), 129–133.

Wilks, S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics 9*(1), 60–62.

Wood, S. (2017). *Generalized Linear Models: An Introduction with R* (2nd ed.). Boca Raton FL, USA: CRC Press.

Young, G. and R. Smith (2005). *Essentials of Statistical Inference*. Cambridge UK: Cambridge University Press.