APTS Introductory notes on Statistical Inference

Simon Shaw, s.shaw@bath.ac.uk
University of Bath

November 2021

1 Introduction to the course

These introductory notes are designed to help students prepare for the APTS module on Statistical Inference. The aim is to introduce the idea of a statistical model, to motivate some principles of statistical inference and to review two approaches to statistical inference: the classical approach, also called the frequentist approach, and the Bayesian approach. A recommended book which covers both introductory ideas to statistical inference and elements of this course is Casella and Berger (2002).

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.

2 Statistical endeavour

Efron and Hastie (2016, 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; 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, generalize and perhaps modify and improve the fruits of such ingenuity."

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 3.1 *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 3.1 *(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 3.2 *(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

$$f_{Y|X}(y \,|\, x, \theta) \;=\; \frac{f_{X,Y}(x, y \,|\, \theta)}{f_X(x \,|\, \theta)} \;=\; \frac{f_{X,Y}(x, y \,|\, \theta)}{\int_{\mathcal{Y}} f_{X,Y}(x, y \,|\, \theta)\, dy}. \tag{3.1}$$

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.

4.1 Likelihood

In the model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$, 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 4.1 *(Likelihood)*
The likelihood for θ given observations x is

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

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 4.1 *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,*

$$f_X(x \mid \theta) \quad = \quad \frac{\beta^{n\alpha}}{\Gamma^n(\alpha)} \left(\prod_{i=1}^{n} x_i \right)^{\alpha - 1} \exp\left(-\beta \sum_{i=1}^{n} x_i \right) \tag{4.1}$$

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

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

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]

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

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

$$f_Y(y \mid \theta) \quad = \quad f_X(x \mid \theta) \left| \frac{\partial x}{\partial y} \right|, \tag{4.2}$$

where $x = g^{-1}(y)$ and $|\cdot|$ denotes the determinant. Consequently, as Cox and Hinkley (1974; 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

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

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 (3) should lead to the same inference about θ.

4.1.1 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 4.2 *(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 (3)) but is not restricted to that.

Example 4.2 *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

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

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 4.3 *(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

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

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

4.2 Sufficient statistics

Consider the model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \,|\, \theta)\}$. 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 4.4 *(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) $f_{X|S}(x \,|\, s, \theta)$ 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 4.4 that the sufficiency of S for θ is dependent upon the choice of the family of distributions in the model.

Example 4.3 *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*

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

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

$$f_S(s \,|\, \theta) \;=\; \frac{(n\theta)^s \exp(-n\theta)}{s!}$$

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

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

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 4.4 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 4.1 *(Fisher-Neyman Factorization Theorem)*
The statistic $S = s(X)$ is sufficient for θ if and only if, for all x and θ,

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

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

Example 4.4 *We revisit Example 4.1 and the case where the X_i are independent and identically distributed Gamma(α, β) random variables. From equation (2) we have*

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

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 4.3, $S = \sum_{i=1}^n X_i$ 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 4.4, 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 Factorization Theorem, we can easily obtain the following result for models drawn from the exponential family.

Theorem 4.2 *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*

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

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 4.5 *The Poisson distribution, see Example 4.3, 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 4.4, 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, Casella and Berger (2002, p276).

4.2.1 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 4.5 *(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.

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.

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 4.1,

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

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

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

$g(\hat{\theta})$. However, there are drawbacks which come from the difficulties of finding the maximum of a function.
Hastie and Efron (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 5.1 *We revisit Examples 4.1 and 4.4 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 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

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

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 5.2 *(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

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

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

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

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 $\mathbb{E}(T^{-1} \,|\, \theta) \neq 1/\mathbb{E}(T \,|\, \theta) \;=\; \theta^{-1}$. An alternate, and better, criterion is that T has small *mean square error (MSE)*,

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

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; 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 5.3 *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 5.3 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.

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 3, see equation (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 5.4 *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.

[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.

5.2 Bayesian inference

In a Bayesian approach to statistical inference, we consider that, in addition to the parametric model $\mathcal{E} = \{\mathcal{X}, \Theta, f_X(x \mid \theta)\}$, 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* $\pi(\theta \mid x)$ of θ given $X = x$,

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

We make the following definition.

Definition 5.1 *(Bayesian statistical model)*
A Bayesian statistical model is the collection $\mathcal{E}_B = \{\mathcal{X}, \Theta, f_X(x \mid \theta), \pi(\theta)\}$.

As O'Hagan and Forster (2004; 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

$$x \quad \mapsto \quad \pi(\theta \mid x)$$

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 $f_X(x \mid \theta)$ in Section 5.1, an analogous hierarchy exists for the posterior distribution $\pi(\theta \mid x)$.

In contrast to the plug-in classical approach to prediction, the Bayesian approach can be viewed as *integrate-out*. If $\mathcal{E}_B = \{\mathcal{X} \times \mathcal{Y}, \Theta, f_{X,Y}(x, y \mid \theta), \pi(\theta)\}$ 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 $f_{Y|X}(y \mid x)$:

$$f_{Y|X}(y \mid x) \quad = \quad \int_\Theta f_{Y|X}(y \mid x, \theta)\pi(\theta \mid x)\, d\theta. \tag{5.1}$$

In terms of an algorithm, we have

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

where, as equation (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 5.5 *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.*

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

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

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.

References

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

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

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

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

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

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference.* Cambridge University Press, Cambridge, UK.

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

O'Hagan, A. and Forster, J. (2004). *Kendall's Advanced Theory of Statistics Volume 2B Bayesian Inference.* Arnold, London, UK, 2nd edition.