Fundamental Theory of Statistical Inference

G. A. Young

November 2021

Preface

These notes cover the essential material of the LTCC course 'Fundamental Theory of Statistical Inference'. They are extracted from the key reference for the course, Young and Smith (2005), which should be consulted for further discussion and detail. The book by Cox (2006) is also highly recommended as further reading.

The course aims to provide a concise but comprehensive account of the essential elements of statistical inference and theory. It is intended to give a contemporary and accessible account of procedures used to draw formal inference from data.

The material assumes a basic knowledge of the ideas of statistical inference and distribution theory. We focus on a presentation of the main concepts and results underlying different frameworks of inference, with particular emphasis on the contrasts among frequentist, Fisherian and Bayesian approaches.

G.A. Young and R.L. Smith. 'Essentials of Statistical Inference'. Cambridge University Press, 2005.

D.R. Cox. 'Principles of Statistical Inference'. Cambridge University Press, 2006.

The following provide article-length discussions which draw together many of the ideas developed in the course:

M.J. Bayarri and J. Berger (2004). The interplay between Bayesian and frequentist analysis. Statistical Science, **19**, 58–80.

J. Berger (2003). Could Fisher, Jeffreys and Neyman have agreed on testing (with discussion)? Statistical Science, **18**, 1–32.

B. Efron. (1998). R.A. Fisher in the 21st Century (with discussion). Statistical Science, **13**, 95–122.

G. Alastair Young
Imperial College London
November 2021

1 Approaches to Statistical Inference

What is statistical inference?

In statistical inference experimental or observational data are modelled as the observed values of random variables, to provide a framework from which inductive conclusions may be drawn about the mechanism giving rise to the data.

We wish to analyse observations $y = (y_1, \ldots, y_n)$ by:

1. Regarding y as the observed value of a random variable $Y = (Y_1, \ldots, Y_n)$ having an (unknown) probability distribution, conveniently specified by a probability density, or probability mass function, $f(y)$.

2. Restricting the unknown density to a suitable family or set $\mathcal{F}$. In *parametric statistical inference*, $f(y)$ is of known analytic form, but involves a finite number of real unknown parameters $\theta = (\theta^1, \ldots, \theta^d)$. We specify the region $\Omega_\theta \subseteq \mathbb{R}^d$ of possible values of θ, the *parameter space*. To denote the dependency of $f(y)$ on θ, we write $f(y;\theta)$ and refer to this as the *model function*. Alternatively, the data could be modelled nonparametrically, a nonparametric model simply being one which does not admit a parametric representation. We will be concerned entirely with parametric statistical inference.

The objective that we then assume is that of assessing, on the basis of the observed data y, some aspect of θ, which for the purpose of the discussion in this paragraph we take to be the value of a particular component, θ^i say. In that regard, we identify three main types of inference: *point estimation*, *confidence set estimation* and *hypothesis testing*. In point estimation, a single value is computed from the data y and used as an estimate of θ^i. In confidence set estimation we provide a set of values which, it is hoped, has a predetermined high probability of including the true, but unknown, value of θ^i. Hypothesis testing sets up specific hypotheses regarding θ^i and assesses the plausibility of any such hypothesis by assessing whether or not the data y support that hypothesis.

Of course, other objectives might be considered, such as: (a) prediction of the value of some as yet unobserved random variable whose distribution depends on θ, or (b) examination of the adequacy of the model specified by $\mathcal{F}$ and Ω_θ. These are important problems, but are not the main focus here.

How do we approach statistical inference?

Following Efron (1998), we identify three main paradigms of statistical inference: the **Bayesian, Fisherian** and **frequentist**. A key objective of this course is to develop in detail the essential features of all three schools of thought and to highlight, we hope in an interesting way, the potential conflicts between them. The basic differences that emerge relate to interpretation of probability and to the objectives of statistical inference. To set the scene, it is of some value to sketch straight away the main characteristics of the three paradigms. To do so, it is instructive to look a little at the historical development of the subject.

The Bayesian paradigm goes back to Bayes and Laplace, in the late 18th century. The fundamental idea behind this approach is that the unknown parameter, θ, should itself be treated as a random variable. Key to the Bayesian viewpoint, therefore, is the specification of a *prior probability distribution* on θ, before the data analysis. We will describe in some detail in Chapter 3 the main approaches to specification of prior distributions, but this can basically be done in either some objective way, or in a subjective way which reflects the statistician's own prior state of belief. To the Bayesian, inference is the formalization of how the prior distribution changes, to the *posterior distribution*, in the light of the evidence presented by the available data y, through Bayes' formula. Central to the Bayesian perspective, therefore, is a use of probability distributions as expressing opinion.

In the early 1920's, R.A. Fisher put forward an opposing viewpoint, that statistical inference must be based entirely on probabilities with direct experimental interpretation. As Efron (1998) notes, Fisher's primary concern was the development of a logic of inductive inference which would release the statistician from the a priori assumptions of the Bayesian school. Central to the Fisherian viewpoint is the *repeated sampling principle*. This dictates that the inference we draw from y should be founded on an analysis of how the conclusions change with variations in the data samples which would be obtained through hypothetical repetitions, under exactly the same conditions, of the experiment which generated the data y in the first place. In a Fisherian approach to inference a central role is played by the concept of *likelihood*, and the associated principle of *maximum likelihood*. In essence, the likelihood measures the probability that different values of the parameter θ assign, under a hypothetical repetition of the experiment, to re-observation of the actual data y. More formally, the ratio of the likelihood at two different values of θ compares the relative plausibilities of observing the data y under the models defined by the two θ values. A further fundamental element of Fisher's viewpoint is that inference, in order to be as relevant as possible to

the data y, must be carried out *conditional* on everything that is known and uninformative about θ.

Fisher's greatest contribution was to provide for the first time an optimality yardstick for statistical estimation, a description of the optimum that it is possible to do in a given estimation problem, and the technique of maximum likelihood, which produces estimators of θ that are close to ideal in terms of that yardstick. As described by Pace and Salvan (1997), spurred on by Fisher's introduction of optimality ideas, in the 1930's and 1940's, Neyman, E.S. Pearson and, later, Wald and Lehmann, offered the third of the three paradigms, the frequentist approach. The origins of this approach lay in a detailed mathematical analysis of some of the fundamental concepts developed by Fisher, in particular likelihood and *sufficiency*. With this focus, emphasis shifted from inference as a summary of data, as favoured by Fisher, to inferential procedures viewed as decision problems. Key elements of the frequentist approach are the need for clarity in mathematical formulation, and that optimum inference procedures should be identified **before** the observations y are available, optimality being defined explicitly in terms of the repeated sampling principle.

2 Decision Theory

In this chapter we give an account of the main ideas of decision theory. Our motivation for beginning our account of statistical inference here is simple. As we have noted, decision theory requires formal specification of all elements of an inference problem, so starting with a discussion of decision theory allows us to set up notation and basic ideas that run through the remainder of the course in a formal but easy manner. In later sections, we will develop the specific techniques and ideas of statistical inference that are central to the three paradigms of inference.

Central to decision theory is the notion of a set of **decision rules** for an inference problem. Comparison of different decision rules is based on examination of the **risk functions** of the rules. The risk function describes the expected **loss** in use of the rule, under hypothetical repetition of the sampling experiment giving rise to the data y, as a function of the **parameter** of interest. Identification of an optimal rule requires introduction of fundamental principles for discrimination between rules, in particular the **minimax** and **Bayes** principles.

2.1 Formulation

A full description of a statistical decision problem involves the following formal elements.

(1) A **parameter space** Ω_θ, which will usually be a subset of $\mathbb{R}^d$ for some $d \geq 1$, so that we have a vector of d unknown parameters. This represents the set of possible unknown states of nature. The unknown parameter value $\theta \in \Omega_\theta$ is the quantity we wish to make inference about.

(2) A **sample space** $\mathcal{Y}$, the space in which the data y lie. Typically we have n observations, so the data, a generic element of the sample space, are of the form $y = (y_1, ..., y_n) \in \mathbb{R}^n$.

(3) A **family of probability distributions** on the sample space $\mathcal{Y}$, indexed by values $\theta \in \Omega_\theta$, $\{\mathbb{P}_\theta(y),\ y \in \mathcal{Y},\ \theta \in \Omega_\theta\}$. In nearly all practical cases this will consist of an assumed parametric family $f(y;\theta)$, of probability mass functions for y (in the discrete case), or probability density functions for y (in the continuous case).

(4) An **action space** $\mathcal{A}$. This represents the set of all actions or decisions available to the experimenter.

Examples of action spaces include the following.

(a) In a hypothesis testing problem where it is necessary to decide between two hypotheses H_0 and H_1, there are two possible actions corresponding to "accept H_0" and "accept H_1". So here $\mathcal{A} = \{a_0, a_1\}$, where a_0 represents accepting H_0 and a_1 represents accepting H_1.

(b) In an estimation problem where we want to estimate the unknown parameter value θ by some function of $x = (x_1, ..., x_n)$, such as $\bar{x} = \frac{1}{n}\sum x_i$ or $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ or $x_1^3 + 27\sin(\sqrt{x_2})$, etc., we should allow ourselves the possibility of estimating θ by any point in Ω_θ. So, in this context we typically have $\mathcal{A} \equiv \Omega_\theta$.

(5) A **loss function** $L : \Omega_\theta \times \mathcal{A} \to \mathbb{R}$ links the action to the unknown parameter. If we take action $a \in \mathcal{A}$ when the true state of nature is $\theta \in \Omega_\theta$, then we incur a loss $L(\theta, a)$.

Note that losses can be positive or negative, a negative loss corresponding to a gain. It is a convention that we formulate the theory in terms of trying to minimise our losses rather than trying to maximise our gains, but obviously the two come to the same thing.

(6) A set $\mathcal{D}$ of **decision rules**. An element $d : \mathcal{Y} \to \mathcal{A}$ of $\mathcal{D}$ is such that each point y in $\mathcal{Y}$ is associated with a specific action $d(y) \in \mathcal{A}$.

For example, with hypothesis testing, we might adopt the rule "Accept H_0 if $\bar{y} \leq 5.7$, otherwise accept H_1". This corresponds to a decision rule,

$$d(y) = \begin{cases} a_0 & \text{if } \bar{y} \leq 5.7, \\ a_1 & \text{if } \bar{y} > 5.7. \end{cases}$$

2.2 The risk function

For parameter value $\theta \in \Omega_\theta$, the risk associated with a decision rule d based on random data Y is defined by

$$\begin{aligned} R(\theta, d) &= \mathbb{E}_\theta L(\theta, d(Y)) \\ &= \begin{cases} \int_{\mathcal{Y}} L(\theta, d(y)) f(y;\theta) dy & \text{for continuous } Y, \\ \sum_{y \in \mathcal{Y}} L(\theta, d(y)) f(y;\theta) & \text{for discrete } Y. \end{cases} \end{aligned}$$

So, we are treating the observed data y as the realised value of a random variable Y with density or mass function $f(y;\theta)$, and defining the risk to be the expected loss, the expectation being with respect to the distribution of Y for the particular parameter value θ.

The key notion of decision theory is that different decision rules should be compared by comparing their risk functions, as functions of θ. Note that we are explicitly invoking the repeated sampling principle here, the definition of risk involving hypothetical repetitions of the sampling mechanism that generated y, through the assumed distribution of Y.

When a loss function represents the real loss in some practical problem (as opposed to some artificial loss function being set up in order to make the statistical decision problem well defined) then it should really be measured in units of "utility" rather than actual money. For example, the expected return on a UK lottery ticket is less than the £1 cost of the ticket; if everyone played so as to maximise their expected gain, nobody would ever buy a lottery ticket! The reason that people still buy lottery tickets, translated into the language of statistical decision theory, is that they subjectively evaluate the very small chance of winning, say, £1,000,000 as worth more than a fixed sum of £1, even though the chance of actually winning the £1,000,000 is appreciably less than 1/1,000,000. There is a formal theory, known as utility theory, which asserts that provided people behave rationally (a considerable assumption in its own right!), then they will always act *as if* they were maximising the expected value of a function known as the utility function. In the lottery example, this implies that we subjectively evaluate the possibility of a massive prize such as £1,000,000 to be worth more than 1,000,000 times as much as the relatively paltry sum of £1. However in situations involving monetary sums of the same order of magnitude, most people tend to be risk averse. For example, faced with a choice between

Offer 1: Receive £10,000 with probability 1,

and

Offer 2: Receive £20,000 with probability $\frac{1}{2}$, otherwise receive £0,

most of us would choose Offer 1. This means that in utility terms, we consider £20,000 as worth less than twice as much as £10,000. Either amount seems like a very large sum of money, and we may not be able to distinguish the two easily in our minds, so that the lack of risk involved in Offer 1 makes it appealing. Of course, if there was a specific reason why we really needed £20,000, for example because this was the cost of a necessary medical operation, we might be more inclined to take the gamble of Offer 2.

Detailed accounts of utility theory are given by Ferguson (1967) or Berger (1985), for example. Instead, in most of the problems we will be considering, we will use various artificial loss functions. A typical example is use of the loss function

$$L(\theta, a) = (\theta - a)^2,$$

the squared error loss function, in a point estimation problem. Then the risk $R(\theta, d)$ of a decision rule is just the mean squared error of $d(Y)$ as an estimator of θ, $\mathbb{E}_\theta\{d(Y) - \theta\}^2$. In this context, we seek a decision rule d which minimises this mean squared error.

In hypothesis testing, where we have two hypotheses H_0, H_1, identified with subsets of Ω_θ, and corresponding action space $\mathcal{A} = \{a_0, a_1\}$ in which action a_j corresponds to selecting the hypothesis H_j, $j = 0, 1$, the most familiar loss function is

$$L(\theta, a) = \begin{cases} 1 & \text{if } \theta \in H_0 \text{ and } a = a_1, \\ 1 & \text{if } \theta \in H_1 \text{ and } a = a_0, \\ 0 & \text{otherwise.} \end{cases}$$

In this case the risk is the probability of making a wrong decision:

$$R(\theta, d) = \begin{cases} \Pr_\theta\{d(Y) = a_1\} & \text{if } \theta \in H_0, \\ \Pr_\theta\{d(Y) = a_0\} & \text{if } \theta \in H_1. \end{cases}$$

In the classical language of hypothesis testing, these two risks are called, respectively, the type I error and the type II error: see Chapter 6.

2.3 Criteria for a good decision rule

In almost any case of practical interest, there will be no way to find a decision rule $d \in \mathcal{D}$ which makes the risk function $R(\theta, d)$ uniformly smallest for all values of θ. Instead, it is necessary to consider a number of criteria which help to narrow down the class of decision rules we consider. The notion is to start with a large class of decision rules d, such as the set of *all* functions from $\mathcal{Y}$ to $\mathcal{A}$, and then reduce the number of candidate decision rules by application of the various criteria, in the hope of being left with some unique best decision rule for the given inference problem.

2.3.1 Admissibility

Given two decision rules d and d', we say that d **strictly dominates** d' if $R(\theta, d) \leq R(\theta, d')$ for all values of θ, and $R(\theta, d) < R(\theta, d')$ for at least one value θ.

Given a choice between d and d', we would always prefer to use d.

Any decision rule which is strictly dominated by another decision rule (as d' is in the definition) is said to be **inadmissible**. Correspondingly, if a

decision rule d is not strictly dominated by any other decision rule, then it is **admissible**.

Admissibility looks like a very weak requirement: it seems obvious that we should always restrict ourselves to admissible decision rules. Admissibility really represents absence of a negative attribute, rather than a possession of a positive attribute. In practice, it may not be so easy to decide whether a given decision rule is admissible or not, and there are some surprising examples of natural-looking estimators which are inadmissible.

2.3.2 Minimax decision rules

The maximum risk of a decision rule $d \in \mathcal{D}$ is defined by

$$\mathrm{MR}(d) = \sup_{\theta \in \Omega_\theta} R(\theta, d).$$

A decision rule d is **minimax** if it minimises the maximum risk:

$$\mathrm{MR}(d) \leq \mathrm{MR}(d') \text{ for all decision rules } d' \in \mathcal{D}.$$

Another way of writing this is to say that d must satisfy

$$\sup_{\theta} R(\theta, d) = \inf_{d' \in \mathcal{D}} \sup_{\theta \in \Omega_\theta} R(\theta, d'). \tag{2.1}$$

In most of the problems we will encounter, the supremum and infimum are actually attained, so that we can rewrite (2.1) as

$$\max_{\theta \in \Omega_\theta} R(\theta, d) = \min_{d' \in \mathcal{D}} \max_{\theta \in \Omega_\theta} R(\theta, d').$$

[Recall that the difference between $\sup_\theta$ and $\max_\theta$ is that the maximum must actually be attained for some $\theta \in \Omega_\theta$, whereas a supremum represents a least upper bound, that may not actually be attained for any single value of θ. Similarly for infimum and minimum.]

The **minimax principle** says we should use a minimax decision rule.

2.3.3 Unbiasedness

A decision rule d is said to be **unbiased** if

$$\mathbb{E}_\theta\{L(\theta', d(Y))\} \geq \mathbb{E}_\theta\{L(\theta, d(Y))\} \text{ for all } \theta, \theta' \in \Omega_\theta.$$

Recall that in elementary statistical theory, if $d(Y)$ is an estimator for a parameter θ, then $d(Y)$ is said to be unbiased if $\mathbb{E}_\theta d(Y) = \theta$ for all θ. The connection between the two notions of unbiasedness is as follows. Suppose the loss function is the squared error loss, $L(\theta, d) = (\theta - d)^2$. Fix θ and let $\mathbb{E}_\theta d(Y) = \phi$. Then for d to be an unbiased decision rule, we require that for all θ',

$$\begin{aligned} 0 &\leq \mathbb{E}_\theta\{L(\theta', d(Y))\} - \mathbb{E}_\theta\{L(\theta, d(Y))\} \\ &= \mathbb{E}_\theta\{(\theta' - d(Y))^2\} - \mathbb{E}_\theta\{(\theta - d(Y))^2\} \\ &= (\theta')^2 - 2\theta'\phi + \mathbb{E}_\theta d^2(Y) - \theta^2 + 2\theta\phi - \mathbb{E}_\theta d^2(Y) \\ &= (\theta' - \phi)^2 - (\theta - \phi)^2. \end{aligned}$$

If $\phi = \theta$ then this statement is obviously true. If $\phi \neq \theta$, then set $\theta' = \phi$ to obtain a contradiction.

Thus we see that if $d(Y)$ is an unbiased estimator in the classical sense, then it is also an unbiased decision rule, provided the loss is squared error. However the above argument also shows that the notion of unbiased decision rule is much broader: we could have whole families of unbiased decision rules corresponding to different loss functions.

The role of unbiasedness in statistical decision theory is ambiguous. Of the various criteria being considered here, it is the only one that does not depend solely on the risk function. Often we find that biased estimators perform better than unbiased ones from the point of view of, say, minimising mean squared error. For this reason, many modern statisticians consider the whole concept of unbiasedness to be somewhere between a distraction and a total irrelevance.

2.3.4 Bayes decision rules

Bayes decision rules are based on different assumptions from the other criteria we have considered, because in addition to the loss function and the class $\mathcal{D}$ of decision rules, we must specify a **prior distribution**, which represents our prior knowledge on the value of the parameter θ, and is represented by a function $\pi(\theta)$, $\theta \in \Omega_\theta$. In cases where Ω_θ contains an open rectangle in $\mathbb{R}^d$, we would take our prior distribution to be absolutely continuous, meaning that $\pi(\theta)$ is taken to be some probability density on Ω_θ. In the case of a discrete parameter space, $\pi(\theta)$ is a probability mass function.

In the continuous case, the Bayes risk of a decision rule d is defined to be

$$r(\pi, d) = \int_{\theta \in \Omega_\theta} R(\theta, d)\pi(\theta)d\theta.$$

In the discrete case, the integral in this expression is replaced by a summation over the possible values of θ. So, the Bayes risk is just average risk, the averaging being with respect to the weight function $\pi(\theta)$ implied by our prior distribution.

A decision rule d is said to be **a Bayes rule** , with respect to a given prior $\pi(\cdot)$, if it minimises the Bayes risk, so that

$$r(\pi, d) = \inf_{d' \in \mathcal{D}} r(\pi, d') = m_\pi, \text{ say.} \tag{2.2}$$

The **Bayes principle** says we should use a Bayes decision rule.

2.3.5 Some other definitions

Sometimes the Bayes rule is not defined because the infimum in (2.2) is not attained for any decision rule d. However, in such cases, for any $\epsilon > 0$ we can find a decision rule d_ϵ for which

$$r(\pi, d_\epsilon) < m_\pi + \epsilon$$

and in this case d_ϵ is said to be ϵ-**Bayes** , with respect to the prior distribution $\pi(\cdot)$.

Finally, a decision rule d is said to be **extended Bayes** if, for , every $\epsilon > 0$, we have that d is ϵ-Bayes with respect to *some* prior, which need not be the same for different ϵ. It is often possible to derive a minimax rule through the property of being extended Bayes.

2.4 Randomised decision rules

Suppose we have a collection of I decision rules $d_1, ..., d_I$ and an associated set of probability weights $p_1, ..., p_I$, so that $p_i \geq 0$ for $1 \leq i \leq I$, and $\sum_i p_i = 1$.

Define the decision rule $d^* = \sum_i p_i d_i$ to be the rule "select d_i with probability p_i". Then d^* is a **randomised decision rule**. We can imagine that we first use some randomisation mechanism, such as tossing coins or using a computer random number generator, to select, independently of the observed data y, one of the decision rules $d_1, ..., d_I$, with respective probabilities $p_1, ..., p_I$. Then, having decided in favour of use of the particular rule d_i, under d^* we carry out the action $d_i(y)$.

For a randomised decision rule d^*, the risk function is defined by averaging across possible risks associated with the component decision rules:

$$R(\theta, d^*) = \sum_{i=1}^{I} p_i R(\theta, d_i).$$

Randomised decision rules may appear to be artificial, but minimax solutions may well be of this form. It is easy to contruct examples in which d^* is formed by randomising the rules $d_1, ..., d_I$ but

$$\sup_\theta R(\theta, d^*) < \sup_\theta R(\theta, d_i) \text{ for each } i,$$

so that d^* may be a candidate for the minimax procedure, where none of $d_1, ..., d_I$ is.

2.5 Finite decision problems

A finite decision problem is one in which the *parameter space* is a finite set: $\Omega_\theta = \{\theta_1, ..., \theta_t\}$ for some finite t, with $\theta_1, \ldots, \theta_t$ specified values. In such cases the notions of admissible, minimax and Bayes decision rules can be given a geometric interpretation: a full treatment is given by Ferguson (1967) and Young and Smith (2005).

2.6 Finding minimax rules in general

A complete classification of minimax decision rules in general problems lies outside the scope of this text, but the following two theorems give simple sufficient conditions for a decision rule to be minimax. One generalisation that is needed in passing from the finite to the infinite case is that the class of Bayes rules must be extended to include sequences of either Bayes rules, or extended Bayes rules.

Theorem 2.1 If δ_n is Bayes with respect to prior $\pi_n(\cdot)$, and $r(\pi_n, \delta_n) \to C$ as $n \to \infty$, and if $R(\theta, \delta_0) \leq C$ for all $\theta \in \Omega_\theta$, then δ_0 is minimax.

Of course this includes the case where $\delta_n = \delta_0$ for all n and the Bayes risk of δ_0 is exactly C.

To see the infinite-dimensional generalisation of the condition $R_1 = R_2$, we make the following definition.

Definition. A decision rule d is an **equaliser decision rule** if $R(\theta, d)$ is the same for every value of θ.

Theorem 2.2 An equaliser decision rule δ_0 which is extended Bayes must be minimax.

2.7 Admissibility of Bayes rules

In Chapter 3 we will present a general result that allows us to characterise the Bayes decision rule for any given inference problem. An immediate question that then arises concerns admissibility. In that regard, the rule of thumb is that Bayes rules are nearly always admissible. We complete this Chapter with some specific theorems on this point. Proofs are left as exercises.

Theorem 2.3 Assume that $\Omega_\theta = \{\theta_1, \ldots, \theta_t\}$ is finite, and that the prior $\pi(\cdot)$ gives positive probability to each θ_i. Then a Bayes rule with respect to π is admissible.

Theorem 2.4 If a Bayes rule is unique, it is admissible.

Theorem 2.5 Let Ω_θ be a subset of the real line. Assume that the risk functions $R(\theta, d)$ are continuous in θ for all decision rules d. Suppose that for any $\epsilon > 0$ and any θ the interval $(\theta - \epsilon, \theta + \epsilon)$ has positive probability under the prior $\pi(\cdot)$. Then a Bayes rule with respect to π is admissible.

3 Bayesian Methods

3.1 Fundamental elements

In non-Bayesian, or classical, statistics Y is random, with a density or probability mass function given by $f(y;\theta)$, but θ is treated as a **fixed** unknown parameter value.

Instead, in Bayesian statistics Y and θ are **both** regarded as random variables, with joint density (or probability mass function) given by $\pi(\theta)f(y;\theta)$, where $\pi(\cdot)$ represent the prior density of θ and $f(\cdot;\theta)$ is the conditional density of Y given θ.

The **posterior density** of θ, given observed value $Y = y$, is given by applying Bayes' law of conditional probabilities:

$$\pi(\theta|y) = \frac{\pi(\theta)f(y;\theta)}{\int_{\Omega_\theta} \pi(\theta')f(y;\theta')d\theta'}.$$

Commonly we write

$$\pi(\theta|y) \propto \pi(\theta)f(y;\theta)$$

where the constant of proportionality is allowed to depend on y but not on θ. This may be written in words as

$$\text{posterior} \ \propto \ \text{prior} \ \times \ \text{likelihood}$$

since $f(y;\theta)$, treated as a function of θ for fixed y, is called the **likelihood function** – for example, maximum likelihood estimation (which is not a Bayesian procedure) proceeds by maximising this expression with respect to θ.

Example 3.1 Consider a binomial experiment in which $Y \sim \text{Bin}(n,\theta)$ for known n and unknown θ. Suppose the prior density is a Beta density on (0,1),

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \quad 0 < \theta < 1,$$

where $a > 0, b > 0$ and $B(\cdot,\cdot)$ is the beta function [$B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, where Γ is the gamma function, $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$]. For the density of Y, here interpreted as a probability mass function, we have

$$f(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

Ignoring all components of π and f which do not depend on θ, we have

$$\pi(\theta|y) \propto \theta^{a+y-1}(1-\theta)^{n-y+b-1}.$$

This is also of Beta form, with the parameters a and b replaced by $a+y$ and $b+n-y$, so the full posterior density is

$$\pi(\theta|y) = \frac{\theta^{a+y-1}(1-\theta)^{n-y+b-1}}{B(a+y, b+n-y)}.$$

This example illustrates an important property of some Bayesian procedures: by adopting a prior density of Beta form, we obtained a posterior density which was also a member of the Beta family, but with different parameters. When this happens, the common parametric form of the prior and posterior are called a **conjugate prior family** for the problem. There is no universal law that says we must use a conjugate prior. However, the conjugate prior property is often a very convenient one, because it avoids having to integrate numerically to find the normalising constant in the posterior density. In non-conjugate cases where we have to do everything numerically, this is the hardest computational problem associated with Bayesian inference. Therefore, in cases where we can find a conjugate family, it is very common to use it. □

3.2 The general form of Bayes rules

We now return to our general discussion of how to solve Bayesian decision problems. For notational convenience, we shall write formulae assuming both Y and θ have continuous densities, though the concepts are exactly the same in the discrete case.

Recall that the risk function of a decision rule d is given by

$$R(\theta, d) = \int_{\mathcal{Y}} L(\theta, d(y)) f(y;\theta) dy$$

and the Bayes risk of d by

$$\begin{aligned} r(\pi, d) &= \int_{\Omega_\theta} R(\theta, d)\pi(\theta) d\theta \\ &= \int_{\Omega_\theta} \int_{\mathcal{Y}} L(\theta, d(y)) f(y;\theta)\pi(\theta) dy d\theta \\ &= \int_{\Omega_\theta} \int_{\mathcal{Y}} L(\theta, d(y)) f(y)\pi(\theta|x) dy d\theta \\ &= \int_{\mathcal{Y}} f(y) \left\{ \int_{\Omega_\theta} L(\theta, d(y))\pi(\theta|y) d\theta \right\} dy. \end{aligned}$$

In the third line here, we have written the joint density $f(y;\theta)\pi(\theta)$ in a different way as $f(y)\pi(\theta|y)$, where $f(y) = \int f(y;\theta)\pi(\theta)d\theta$ is the marginal density of Y. The change of order of integration in the fourth line is trivially justified because the integrand is non-negative.

From the final form of this expression, we can see that, to find the the action $d(y)$ specified by the Bayes rule for any y, it suffices to minimise the expression inside the brackets. In other words, for each y we choose $d(y)$ to minimise

$$\int_{\Omega_\theta} L(\theta, d(y))\pi(\theta|y)d\theta,$$

the **expected posterior loss** associated with the observed y. This greatly simplifies the calculation of the Bayes rule in a particular case. It also illustrates what many people feel is an intuitively natural property of Bayesian procedures: in order to decide what to do based on a particular observed y, it is only necessary to think about the losses that follow from one value $d(y)$. There is no need to worry (as would be the case with a minimax procedure) about all the other values of Y that might have occurred, but did not. This property illustrates just one of the features that have propelled many modern statisticians towards Bayesian methods.

Case 1. Hypothesis testing. Consider testing the hypothesis $H_0 : \theta \in \Theta_0$ against the hypothesis $H_1 : \theta \in \Theta_1 \equiv \Omega_\theta \setminus \Theta_0$, the complement of Θ_0. Now the action space $\mathcal{A} = \{a_0, a_1\}$, where a_0 denotes 'accept H_0' and a_1 denotes 'accept H_1'. Assume the following form of loss function:

$$L(\theta, a_0) = \begin{cases} 0 & \text{if } \theta \in \Theta_0, \\ 1 & \text{if } \theta \in \Theta_1, \end{cases}$$

and

$$L(\theta, a_1) = \begin{cases} 1 & \text{if } \theta \in \Theta_0, \\ 0 & \text{if } \theta \in \Theta_1. \end{cases}$$

The Bayes decision rule is: accept H_0 if

$$\Pr(\theta \in \Theta_0|y) > \Pr(\theta \in \Theta_1|y).$$

Since $\Pr(\theta \in \Theta_1|y) = 1 - \Pr(\theta \in \Theta_0|y)$, this is equivalent to accepting H_0 if $\Pr(\theta \in \Theta_0|y) > 1/2$.

Case 2. Point estimation. Suppose loss is squared error: $L(\theta, d) = (\theta - d)^2$. For observed $Y = y$, the Bayes estimator chooses $d = d(y)$ to minimise

$$\int_{\Omega_\theta} (\theta - d)^2\pi(\theta|y)d\theta.$$

Differentiating with respect to d, we find

$$\int_{\Omega_\theta} (\theta - d)\pi(\theta|y)d\theta = 0.$$

Taking into account that the posterior density integrates to 1, this becomes

$$d = \int_{\Omega_\theta} \theta\pi(\theta|y)d\theta,$$

the *posterior mean* of θ. In words, for a squared error loss function, the Bayes estimator is the mean of the posterior distribution.

Case 3. Point estimation. Suppose $L(\theta, d) = |\theta - d|$. The Bayes rule will minimise

$$\int_{-\infty}^{d} (d - \theta)\pi(\theta|y)d\theta + \int_{d}^{\infty} (\theta - d)\pi(\theta|y)d\theta.$$

Differentiating with respect to d, we must have

$$\int_{-\infty}^{d} \pi(\theta|y)d\theta - \int_{d}^{\infty} \pi(\theta|y)d\theta = 0$$

or in other words

$$\int_{-\infty}^{d} \pi(\theta|y)d\theta = \int_{d}^{\infty} \pi(\theta|y)d\theta = \frac{1}{2}.$$

The Bayes rule is the *posterior median* of θ.

Case 4. Interval estimation. Suppose

$$L(\theta, d) = \begin{cases} 0 & \text{if } |\theta - d| \leq \delta, \\ 1 & \text{if } |\theta - d| > \delta, \end{cases}$$

for prescribed $\delta > 0$. The expected posterior loss in this case is the posterior probability that $|\theta - d| > \delta$. This can be most easily motivated as a Bayesian version of interval estimation: we want to find the "best" interval of the form $(d - \delta, d + \delta)$, of predetermined length 2δ. "Best" here means the interval that maximises the posterior probability that θ is within the interval specified. The resulting interval is often called the HPD (for *highest posterior density*) interval.

3.3 Back to minimax...

We now give an example to show how some of the ideas we have developed may be applied to solve a non-trivial problem in minimax decision theory.

The problem is: find a minimax estimator of θ based on a single observation $Y \sim \text{Bin}(n, \theta)$ with n known, under squared error loss $L(\theta, d) = (\theta - d)^2$.

We know by Theorem 2.2 of Chapter 2 that if we can find a Bayes (or extended Bayes) estimator that has constant mean squared error (i.e. risk), this will also be a minimax rule.

We do not know all the possible Bayes estimators for this problem, but we do know a very large class of them, namely, all those that arise from the conjugate prior, a Beta prior with parameters $a > 0$, $b > 0$. For such a prior, the posterior distribution is Beta with the parameters a and b replaced by $a + Y$, $b + n - Y$. We also know that with squared error loss, the Bayes estimator is the mean of the posterior distribution,

$$d(Y) = \frac{a + Y}{a + b + n}.$$

The question therefore arises: is there any estimator in this class which has constant mean squared error? If there is, then it is necessarily the minimax estimator.

Recall $\mathbb{E}Y = n\theta$, $\mathbb{E}Y^2 = n\theta(1 - \theta) + n^2\theta^2$. Writing $c = a + b + n$, we have

$$\begin{aligned} \mathbb{E}\left\{\left(\frac{a+Y}{c} - \theta\right)^2\right\} &= \frac{1}{c^2}\mathbb{E}\{(Y + a - c\theta)^2\} \\ &= \frac{1}{c^2}\{n\theta(1-\theta) + n^2\theta^2 + 2n\theta(a - c\theta) + (a - c\theta)^2\}. \end{aligned}$$

This is a quadratic function of θ, and will be constant if the coefficients of θ and θ^2 are 0. This requires

$$n + 2na - 2ac = 0, \tag{3.1}$$

$$-n + n^2 - 2nc + c^2 = 0. \tag{3.2}$$

Equation (3.2) has roots $c = n \pm \sqrt{n}$, but we need $c > n$ for a proper Bayes rule, so take $c = n + \sqrt{n}$. Then (3.1) gives $a = \sqrt{n}/2$ so the final result is that

$$d(Y) = \frac{Y + \sqrt{n}/2}{n + \sqrt{n}}$$

is the minimax decision rule with respect to squared error loss. A prior with respect to which the Bayes rule is minimax is called a **least favourable prior**.

3.4 Shrinkage and the James-Stein estimator

We now move on to some broader aspects of the interplay between Bayesian methods and decision theory. Subject to certain restrictions on the prior, Bayes decision rules are admissible. However, minimax rules may not be admissible, and more generally, statistical estimators that are derived from other criteria may not be admissible. In situations like this, it may be possible to use Bayesian ideas to improve upon classical estimators, even when they are assessed by frequentist criteria. The earliest and most famous example of this is *Stein's paradox*, which we now describe.

Example 3.2. Stein's paradox. Let Y have a p-dimensional ($p \geq 3$) normal distribution with mean vector μ and known covariance matrix equal to the identity I, meaning that $Y_i \sim N(\mu_i, 1)$, independently, $i = 1, \ldots, p$.

Consider estimation of μ, with loss function $L(\mu, d) = \|\mu - d\|^2 = \sum_{i=1}^{p} (\mu_i - d_i)^2$ equal to the sum of squared errors.

If we had just one $Y \sim N(\mu, 1)$, $p = 1$, we would certainly estimate μ by Y. In the general case $p > 1$, if, as we have assumed, the Y_i are independent and we use as loss the sum of squared error losses for the individual components, it seems obvious that the Y_i have nothing to do with one another and that we should therefore use Y as the multivariate estimator of μ. Stein's paradox is so called because what seems obvious turns out not to be true.

Consider the class of "James-Stein estimators" of the form

$$d^a(Y) = \left(1 - \frac{a}{\|Y\|^2}\right) Y,$$

indexed by $a \geq 0$, which (at least if $\|Y\|^2 > a$) shrink Y towards 0.

Now $Y \equiv d^0(Y)$ has risk

$$\begin{aligned} R\big(\mu, d^0(Y)\big) = \mathbb{E}\|\mu - Y\|^2 &= \sum_{i=1}^{p} \mathbb{E}(\mu_i - Y_i)^2 = \sum_{i=1}^{p} \operatorname{var} Y_i \\ &= p, \quad \text{irrespective of } \mu. \end{aligned}$$

Integration by parts shows that, for each i, for suitably behaved real-valued functions h,

$$\mathbb{E}\{(Y_i - \mu_i) h(Y)\} = \mathbb{E}\left\{\frac{\partial h(Y)}{\partial Y_i}\right\}.$$

This result, known as *Stein's Lemma*, enables us to compute the risk of the estimator $d^a(Y)$:

$$\begin{aligned} R\big(\mu, d^a(Y)\big) &= \mathbb{E}\|\mu - d^a(Y)\|^2 \\ &= \mathbb{E}\|\mu - Y\|^2 - 2a\mathbb{E}\left[\frac{Y^T(Y-\mu)}{\|Y\|^2}\right] + a^2\mathbb{E}\left[\frac{1}{\|Y\|^2}\right]. \end{aligned}$$

We have

$$\begin{aligned} \mathbb{E}\left[\frac{Y^T(Y-\mu)}{\|Y\|^2}\right] &= \mathbb{E}\left[\sum_{i=1}^{p}\frac{Y_i(Y_i-\mu_i)}{\Sigma_j Y_j^2}\right] \\ &= \sum_{i=1}^{p}\mathbb{E}\left[\frac{\partial}{\partial Y_i}\left\{\frac{Y_i}{\Sigma_j Y_j^2}\right\}\right] \\ &= \sum_{i=1}^{p}\mathbb{E}\left[\frac{\Sigma_j Y_j^2 - 2Y_i^2}{(\Sigma_j Y_j^2)^2}\right] \\ &= \mathbb{E}\left[\frac{p-2}{\|Y\|^2}\right], \end{aligned}$$

so

$$R\big(\mu, d^a(Y)\big) = p - \big[2a(p-2) - a^2\big]\mathbb{E}\left(\frac{1}{\|Y\|^2}\right). \tag{3.3}$$

Remember that here $\mathbb{E}$ denotes expectation with respect to the distribution of Y for the given μ. We then note immediately that $R(\mu, d^a(Y)) < p \equiv R(\mu, d^0(Y))$, provided $2a(p-2) - a^2 > 0$ i.e. $0 < a < 2(p-2)$. For such a, $d^a(Y)$ strictly dominates $d^0(Y)$, so that the obvious estimator Y is inadmissible!

Note also that the risk of $d^a(Y)$ is minimised for $a = p-2$. When $\mu = 0$, $Y^TY \sim \mathcal{X}_p^2$, so that $\mathbb{E}[1/(\|Y\|^2)] = \frac{1}{p-2}$, by a straightforward direct calculation. Hence, when $\mu = 0$, $d^{p-2}(Y)$ has risk $p - [(p-2)^2/(p-2)] = 2$, which is substantially less than the risk of Y if p is large. □

This inadmissibility result was first pointed out by Charles Stein in 1956, but then proved in more detail by James and Stein (1961). Stein (1981) presented a simpler proof, on which the above analysis is essentially based. At first sight, the result seems incredible: there is no apparent "tying together" of the losses in different components, yet the obvious estimator, the sample 'mean' Y, is not admissible. It is now known that this is a very general phenomenon when comparing three or more populations — the present setting of normal means with known common variance is just the simplest case.

There are, however, many well-documented examples which give intuitive justification for not using the sample mean in practice. The most famous of these concerns an analysis of baseball batting data, by Efron and Morris (1975): see also Efron and Morris (1977).

3.4.1 Some discussion

Admissibility of $d(Y) = Y$ in dimension $p = 1$ was established by Blyth (1951). A simple, direct proof is possible: see, for example Casella and Berger (1990, Section 10.4). Admissibility is more awkward to prove in the case $p = 2$, but was established by Stein (1956). Berger (1985, Chapter 8) gives the admissibility results a Bayesian interpretation, using the notion of a *generalised Bayes rule*. Though the formal definition of a generalised Bayes rule is mathematically awkward, the rough idea is that of a rule which minimises the expected posterior loss, obtained from an improper prior. In the estimation problem at hand, any admissible estimator is a generalised Bayes rule, and results are available which determine whether or not a generalised Bayes estimator is admissible. Since Y is a generalised Bayes estimator in any dimension p, these latter results lead immediately to the conclusions that Y is admissible if $p = 1$ or 2, but not if $p \geq 3$.

A point of clarification should be noted here: although the estimator $d^a(Y)$ defined in Example 3.2 dominates $d^0(Y) = Y$ for certain values of a, this does not mean we would actually want to use the estimator $d^a(Y)$ in applications. Once the idea is presented, that we might not want to use Y as our estimator, then there are many so-called shrinkage estimators which potentially improve on Y, and the task of deciding which of them to adopt is an important focus of practical discussion. A key point to note is that the estimator $d^{p-2}(Y)$ is actually inadmissible: it is strictly dominated by the estimator $d_+^{p-2}(Y)$, which replaces the factor $(1 - \frac{p-2}{Y^T Y})$ by zero whenever it is negative.

3.5 Empirical Bayes

In a standard Bayesian analysis, there will usually be parameters in the prior distribution that have to be specified.

For example, consider the simple normal model in which $Y \mid \theta \sim N(\theta, 1)$ and θ has the prior distribution $\theta \mid \tau^2 \sim N(0, \tau^2)$. If a value is specified for the parameter τ^2 of the prior, a standard Bayesian analysis can be carried out. Noting that $f(y) = \int f(y; \theta)\pi(\theta)d\theta$, it is readily shown that the marginal

distribution of Y is $N(0, \tau^2 + 1)$, and can therefore be used to estimate τ^2, in circumstances where a value is not specified.

Empirical Bayes analysis is characterised by the estimation of prior parameter values from marginal distributions of data. Having estimated the prior parameter values, we proceed as before, as if these values had been fixed at the beginning.

3.5.1 James-Stein estimator, revisited.

In the Stein's paradox Example 3.2 above, the estimator $d^{p-2}(Y)$ may be viewed as an empirical Bayes estimator of μ, the Bayes rule with prior parameter values replaced by estimates constructed from the marginal distribution of the Y_i.

Specifically, let $Y_i \mid \mu_i$ be distributed as $N(\mu_i, 1)$, independently $i = 1, \ldots, p$, and suppose $\mu_1, \ldots, \mu_p$ are independent, identically distributed $N(0, \tau^2)$.

If τ^2 is known, the Bayes estimator $\delta^\tau(Y)$, for the given sum of squared errors loss, of $\mu = (\mu_1, \ldots, \mu_p)^T$ is the posterior mean $\delta^\tau(Y) = \frac{\tau^2}{\tau^2+1}Y$, on observing that the posterior distribution of μ_i is $N(\frac{\tau^2}{\tau^2+1}Y_i, \frac{\tau^2}{\tau^2+1})$, independently for $i = 1, \ldots, p$. Straightforward calculations then show that the Bayes risk of $\delta^\tau(y)$, $r(\tau, \delta^\tau(Y))$, say, in an obvious notation, is given by

$$r(\tau, \delta^\tau(Y)) = \sum_{i=1}^{p} \text{var}\,(\mu_i|Y_i) = \sum_{i=1}^{p} \frac{\tau^2}{\tau^2+1} = \frac{p\tau^2}{\tau^2+1}.$$

Marginally the Y_i are independent, identically distributed $N(0, \tau^2 + 1)$, so that $Y_i/\sqrt{\tau^2+1} \sim N(0,1)$ and marginally $\|Y\|^2/(\tau^2+1) \sim \chi^2_p$. Since we know that $E(1/Z) = 1/(p-2)$ if $Z \sim \chi^2_p$ and $p \geq 3$, we see that taking the expectation with respect to this marginal distribution of Y gives,

$$\mathbb{E}\left[1 - \frac{(p-2)}{\|Y\|^2}\right] = \frac{\tau^2}{\tau^2+1}, \tag{3.4}$$

if $p \geq 3$.

In the case when τ^2 is unknown, estimating $\tau^2/(\tau^2+1)$ by $1-(p-2)/(\|Y\|^2)$ yields the James-Stein estimator $d^{p-2}(Y)$.

Under our assumed model, the Bayes risk of the James-Stein estimator $d^{p-2}(Y)$ is

$$\begin{aligned} r(\tau, d^{p-2}(Y)) &= \int R(\mu, d^{p-2}(X))\pi(\mu)d\mu \\ &= \int_{\mathbb{R}^p}\int_{\mathcal{Y}} \left[p - \frac{(p-2)^2}{\|y\|^2}\right] f(y|\mu)\pi(\mu)dyd\mu \\ &= \int_{\mathcal{Y}} \left\{\int_{\mathbb{R}^p} \left[p - \frac{(p-2)^2}{\|y\|^2}\right] \pi(\mu|y)d\mu\right\} f(y)dy, \end{aligned}$$

where we have used (3.3) and then changed the order of integration. Now, the integrand in the inner integral is independent of μ, and $\int \pi(\mu|y)d\mu$ is trivially equal to 1, and therefore

$$r(\tau, d^{p-2}(Y)) = p - (p-2)^2\mathbb{E}(\frac{1}{\|Y\|^2}).$$

Now the expectation is, as in (3.4), with respect to the marginal distribution of Y, so that (3.4) immediately gives

$$r(\tau, d^{p-2}(Y)) = p - \frac{p-2}{\tau^2+1} = r(\tau, \delta^\tau(Y)) + \frac{2}{\tau^2+1}.$$

The second term represents the increase in Bayes risk associated with the need to estimate τ^2: the increase tends to 0 as $\tau^2 \to \infty$.

3.6 Choice of prior distributions

The main approaches to the selection of prior distributions may be summarised as:

(a) physical reasoning (Bayes) – too restrictive for most practical purposes;

(b) flat or uniform priors, including improper priors (Laplace, Jeffreys) – the most widely used method in practice, but the theoretical justification for this approach is still a source of argument;

(c) subjective priors (de Finetti, Savage) – used in certain specific situations such as weather forecasting (though even there it does not tend to be as part of a formal Bayesian analysis with likelihoods and posterior distributions) and for certain kinds of business applications where prior information is very important and it is worthwhile to go to the trouble of trying to establish ("elicit" is the word most commonly used for this) the client's true subjective opinions, but hardly used at all for routine statistical analysis;

(d) prior distributions for convenience, e.g. conjugate priors – in practice these are very often used just to simplify the calculations.

3.7 Computational techniques

As mentioned previously, one of the main practical advantages of Bayesian methods is that they may often be applied in very complicated situations where both Y and θ are very high-dimensional. In such a situation, the main computational problem is to compute numerically the normalising constant that is required to make the posterior density a proper density function.

Direct numerical integration is usually impracticable in more than four or five dimensions. Instead, **Monte Carlo methods** – in which random numbers are drawn to simulate a sample from the posterior distribution – have become very widely used. The two key algorithms are the Gibbs sampler and the Hastings-Metropolis algorithm.

3.8 Hierarchical modelling

Another way of dealing with the specification of prior parameter values in Bayesian inference is with a hierarchical specification, in which the prior parameter values are themselves given a (second-stage) prior.

For example, in the simple normal model considered previously we might specify $Y \mid \theta \sim N(\theta, 1)$, $\theta \mid \tau^2 \sim N(0, \tau^2)$ and $\tau^2 \sim$ uniform $(0, \infty)$, an example of an improper, diffuse prior. Inference on θ is based on the marginal posterior of θ, obtained by integrating out τ^2 from the joint posterior of θ and τ^2:

$$\pi(\theta \mid y) = \int \pi(\theta, \tau^2 \mid y) d\tau^2,$$

where the joint posterior $\pi(\theta, \tau^2 \mid y) \propto f(y; \theta)\pi(\theta \mid \tau^2)\pi(\tau^2)$.

Hierarchical modelling is a very effective practical tool and usually yields answers that are reasonably robust to misspecification of the model. Often, answers from a hierarchical analysis are quite similar to those obtained from an empirical Bayes analysis. In particular, when the second-stage prior is relatively flat compared to the first-stage prior and the density of the observable Y, answers from the two approaches are close to one another.

3.9 Predictive distributions

So far, we have stressed use of the posterior distribution $\pi(\theta \mid y)$ as a means of making inference about the parameter θ. We may not be interested directly in

that parameter, but rather in some independent future observation depending on θ. It is possible to obtain the conditional distribution of the value of a future observation $Y^\dagger$, given the data y, from the posterior $\pi(\theta \mid y)$.

Suppose that $y = (y_1, \ldots, y_n)$, with the y_i independent from $f(y; \theta)$. Since, given θ, $Y^\dagger$ and y are independent and $Y^\dagger$ has density $f(y^\dagger; \theta)$, the posterior joint distribution of $Y^\dagger$ and θ is $f(y^\dagger; \theta)\pi(\theta \mid y)$. Integrating out θ gives the posterior **predictive distribution** as

$$g(Y^\dagger \mid y) = \int f(Y^\dagger; \theta)\pi(\theta \mid y) d\theta.$$

If a point prediction of $Y^\dagger$ is required, we might use the mean, median or other function of this distribution, depending on our loss function.

In the Bayesian paradigm, predictive inference is therefore, in principle, straightforward, since the logical status of the future observation $Y^\dagger$ and the parameter θ is the same, both being random. This contrasts with methods for predictive inference in frequentist approaches, which are generally more difficult, due to the observation and the parameter having *different* status, the former as a random variable, and the latter as a fixed value.

4 Special Families of Models

Two general classes of models particularly relevant in theory and practice are exponential families and transformation families.

4.1 Exponential families

Suppose that the distribution of Y depends on m unknown parameters, denoted by $\phi = (\phi^1, \ldots, \phi^m)^T$, to be called natural parameters, through a density of the form

$$f_Y(y;\phi) = h(y)\exp\{s^T\phi - K(\phi)\}, \quad y \in \mathcal{Y}, \tag{4.1}$$

where $\mathcal{Y}$ is a set not depending on ϕ. Here $s \equiv s(y) = (s_1, \ldots, s_m)^T$, are called natural statistics. The value of m may be reduced if the components of ϕ satisfy a linear constraint, or if the components of s are (with probability one) linearly dependent. So assume that the representation (4.1) is minimal, in that m is as small as possible. Provided the natural parameter space Ω_ϕ consists of all ϕ such that

$$\int h(y)\exp\{s^T\phi\}dy < \infty,$$

we refer to the family $\mathcal{F}$ as a full exponential model, or an (m, m) exponential family.

The exponential form (4.1) is preserved if we apply any fixed nonsingular linear transformation to s, provided we make the inverse transformation to ϕ, leaving $s^T\phi$ invariant. If $0 \in \Omega_\phi$, we can without loss of generality take $K(0) = 0$ and then $h(y) = f_Y(y;0)$. In other cases we can measure ϕ from some suitable origin $\phi_0 \in \Omega_\phi$, by rewriting (4.1) as

$$f_Y(y;\phi) = f_Y(y;\phi_0)\exp[s^T(\phi - \phi_0) - \{K(\phi) - K(\phi_0)\}].$$

We refer to $f_Y(y;\phi)$ as the (m, m) exponential family generated from the baseline $f_Y(y;\phi_0)$, by exponential tilting via s. We generate all the members of the family by tilting a single baseline density.

We have from (4.1) that the moment generating function of the random variable S corresponding to s is

$$\begin{aligned} M(S;t,\phi) &= E\{\exp(S^Tt)\} \\ &= \int h(y)\exp\{s^T(\phi + t)\}dy \times \exp\{-K(\phi)\} \\ &= \exp\{K(\phi + t) - K(\phi)\}, \end{aligned}$$

from which we obtain

$$E(S_i;\phi) = \frac{\partial K(\phi)}{\partial \phi^i},$$

or

$$E(S;\phi) = \nabla K(\phi),$$

where ∇ is the gradient operator $(\partial/\partial\phi^1, \ldots, \partial/\partial\phi^m)^T$. Also,

$$\mathrm{cov}(S_i, S_j;\phi) = \frac{\partial^2 K(\phi)}{\partial\phi^i \partial\phi^j}.$$

To compute $E(S_i)$ etc. it is only necessary to know the function $K(\phi)$.

Let $s(y) = (t(y), u(y))$ be a partition of the vector of natural statistics, where t has k components and u is $m-k$ dimensional. Consider the corresponding partition of the natural parameter $\phi = (\tau, \xi)$. The density of a generic element of the family can be written as

$$f_Y(y;\tau,\xi) = \exp\{\tau^T t(y) + \xi^T u(y) - K(\tau,\xi)\}h(y).$$

Two key results hold, which make exponential families particularly attractive, as they allow inference about selected components of the natural parameter, in the absence of knowledge about the other components.

First, the family of marginal distributions of $U = u(Y)$ is an $m-k$ dimensional exponential family,

$$f_U(u;\tau,\xi) = \exp\{\xi^T u - K_\tau(\xi)\}h_\tau(u),$$

say.

Secondly, the family of conditional distributions of $T = t(Y)$ given $u(Y) = u$ is a k dimensional exponential family, and the conditional densities are free of ξ, so that

$$f_{T|U=u}(t;u,\tau) = \exp\{\tau^T t - K_u(\tau)\}h_u(t),$$

say.

A proof of both of these results is given by Pace and Salvan (1997, p. 190). The key is to observe that the family of distributions of the natural statistics is an m dimensional exponential family, with density

$$f_{T,U}(t,u;\tau,\xi) = \exp\{\tau^T t + \xi^T u - K(\tau,\xi)\}p_0(t,u),$$

where $p_0(t,u)$ denotes the density of the natural statistics when $(\tau,\xi) = (0,0)$, assuming without loss of generality that $0 \in \Omega_\phi$.

In the situation described above, both the natural statistic and the natural parameter lie in m-dimensional regions. Sometimes, ϕ may be restricted to lie in a d-dimensional subspace, $d < m$. This is most conveniently expressed by writing $\phi = \phi(\theta)$ where θ is a d-dimensional parameter. We then have

$$f_Y(y;\theta) = h(y)\exp[s^T\phi(\theta) - K\{\phi(\theta)\}]$$

where $\theta \in \Omega_\theta \subset \mathbb{R}^d$. We call this system an (m, d) exponential family, noting that we required that $(\phi^1, \ldots, \phi^m)$ does not belong to a v-dimensional linear subspace of $\mathbb{R}^m$ with $v < m$: we indicate this by saying that the exponential family is curved. Think of the case $m = 2, d = 1$: $\{\phi^1(\theta), \phi^2(\theta)\}$ defines a curve in the plane, rather than a straight line, as θ varies.

Interest in curved exponential families stems from two features, related to concepts to be discussed. The maximum likelihood estimator is not a sufficient statistic, so that there is scope for conditioning on an ancillary statistic. Also, it can be shown that any sufficiently smooth parametric family can be approximated, locally to the true parameter value, to some suitable order, by a curved exponential family.

4.2 Transformation families

The basic idea behind a transformation family is that of a group of transformations acting on the sample space, generating a family of distributions all of the same form, but with different values of the parameters.

Recall that a group G is a mathematical structure having a binary operation $\circ$ such that

- if $g, g' \in G$, then $g \circ g' \in G$;
- if $g, g', g'' \in G$, then $(g \circ g') \circ g'' = g \circ (g' \circ g'')$;
- G contains an identity element e such that $e \circ g = g \circ e = g$, for each $g \in G$; and
- each $g \in G$ possesses an inverse $g^{-1} \in G$ such that $g \circ g^{-1} = g^{-1} \circ g = e$.

In the present context, we will be concerned with a group G of transformations acting on the sample space $\mathcal{Y}$ of a random variable Y, and the binary operation will simply be composition of functions: we have $e(y) = y$, $(g_1 \circ g_2)(y) = g_1(g_2(y))$.

The group elements typically correspond to elements of a parameter space Ω_θ, so that a transformation may be written as, say, g_θ. The family of densities of $g_\theta(Y)$, for $g_\theta \in G$, is called a **(group) transformation family**.

Setting $y \approx y'$ iff there is a $g \in G$ such that $y = g(y')$ defines an equivalence relation, which partitions $\mathcal{Y}$ into equivalence classes called *orbits*. These may be labelled by an index a, say. Two points y and y' on the same orbit have the same index, $a(y) = a(y')$. Each $y \in \mathcal{Y}$ belongs to precisely one orbit, and might be represented by a (which identifies the orbit) and its position on the orbit.

4.2.1 Maximal invariant

We say that the statistic t is **invariant** to the action of the group G if its value does not depend on whether y or $g(y)$ was observed, for any $g \in G$: $t(y) = t(g(y))$. An example is the index a above.

The statistic t is **maximal invariant** if every other invariant statistic is a function of it, or equivalently, $t(y) = t(y')$ implies that $y' = g(y)$ for some $g \in G$. A maximal invariant can be thought of (Davison, 2003, Section 5.3) as a reduced version of the data that represents it as closely as possible while remaining invariant to the action of G. In some sense, it is what remains of Y once minimal information about the parameter values has been extracted.

4.2.2 Equivariant statistics and a maximal invariant

As described, typically there is a one-to-one correspondence between the elements of G and the parameter space Ω_θ, and then the action of G on $\mathcal{Y}$ requires that Ω_θ itself constitutes a group, with binary operation $*$ say: we must have $g_\theta \circ g_\phi = g_{\theta * \phi}$. The group action on $\mathcal{Y}$ induces a group action on Ω_θ. If $\bar{G}$ denotes this induced group, then associated with each $g_\theta \in G$ there is a $\bar{g}_\theta \in \bar{G}$, satisfying $\bar{g}_\theta(\phi) = \theta * \phi$.

If t is an invariant statistic, the distribution of $T = t(Y)$ is the same as that of $t(g(Y))$, for all g. If, as we assume here, the elements of G are identified with parameter values, this means that the distribution of T does not depend on the parameter and is known in principle. T is said to be *distribution constant*.

A statistic $S = s(Y)$ defined on $\mathcal{Y}$ and taking values in the parameter space Ω_θ is said to be **equivariant** if $s(g_\theta(y)) = \bar{g}_\theta(s(y))$ for all $g_\theta \in G$ and $y \in \mathcal{Y}$. Often S is chosen to be an estimator of θ, and it is then called an *equivariant estimator*.

A key operational point is that an equivariant estimator can be used to construct a maximal invariant.

Consider $t(Y) = g^{-1}_{s(Y)}(Y)$. This is invariant, since

$$\begin{aligned} t(g_\theta(y)) &= g^{-1}_{s(g_\theta(y))}(g_\theta(y)) = g^{-1}_{\bar{g}_\theta(s(y))}(g_\theta(y)) = g^{-1}_{\theta * s(y)}(g_\theta(y)) \\ &= g^{-1}_{s(y)}\{g_\theta^{-1}(g_\theta(y))\} = g^{-1}_{s(y)}(y) = t(y). \end{aligned}$$

If $t(y) = t(y')$, then $g^{-1}_{s(y)}(y) = g^{-1}_{s(y')}(y')$, and it follows that $y' = g_{s(y')} \circ g^{-1}_{s(y)}(y)$, which shows that $t(Y)$ is maximal invariant.

The statistical importance of a maximal invariant will be illuminated in Chapter 5. In a transformation family, a maximal invariant plays the role of the ancillary statistic in the conditional inference on the parameter of interest indicated by a Fisherian approach. The above direct construction of a maximal invariant from an equivariant estimator facilitates identification of an appropriate ancillary statistic in the transformation family context.

4.2.3 An example

An important example is the **location-scale model**. Let $Y = \eta + \tau\epsilon$, where ϵ has a known density f, and the parameter $\theta = (\eta, \tau) \in \Omega_\theta = \mathbb{R} \times \mathbb{R}_+$. Define a group action by $g_\theta(y) = g_{(\eta,\tau)}(y) = \eta + \tau y$, so

$$g_{(\eta,\tau)} \circ g_{(\mu,\sigma)}(y) = \eta + \tau\mu + \tau\sigma y = g_{(\eta+\tau\mu,\tau\sigma)}(y).$$

The set of such transformations is closed with identity $g_{(0,1)}$. It is easy to check that $g_{(\eta,\tau)}$ has inverse $g_{(-\eta/\tau,\tau^{-1})}$. Hence, $G = \{g_{(\eta,\tau)} : (\eta, \tau) \in \mathbb{R} \times \mathbb{R}_+\}$ constitutes a group under the composition of functions operation $\circ$ defined above.

The action of $g_{(\eta,\tau)}$ on a random sample $Y = (Y_1, \ldots, Y_n)$ is $g_{(\eta,\tau)}(Y) = \eta + \tau Y$, with $\eta \equiv \eta 1_n$, where 1_n denotes the $n \times 1$ vector of 1's, and Y is written as an $n \times 1$ vector.

The induced group action on Ω_θ is given by $\bar{g}_{(\eta,\tau)}((\mu, \sigma)) \equiv (\eta, \tau) * (\mu, \sigma) = (\eta + \tau\mu, \tau\sigma)$.

The sample mean and standard deviation are equivariant, because with

$s(Y) = (\bar{Y}, V^{1/2})$, where $V = (n-1)^{-1}\sum(Y_j - \bar{Y})^2$, we have

$$\begin{aligned} s(g_{(\eta,\tau)}(Y)) &= \left(\overline{\eta + \tau Y}, \left\{(n-1)^{-1}\sum(\eta + \tau Y_j - \overline{(\eta + \tau Y)})^2\right\}^{1/2}\right) \\ &= \left(\eta + \tau\bar{Y}, \left\{(n-1)^{-1}\sum(\eta + \tau Y_j - \eta - \tau\bar{Y})^2\right\}^{1/2}\right) \\ &= \left(\eta + \tau\bar{Y}, \tau V^{1/2}\right) \\ &= \bar{g}_{(\eta,\tau)}(s(Y)). \end{aligned}$$

A maximal invariant is $A = g^{-1}_{s(Y)}(Y)$, and the parameter corresponding to $g^{-1}_{s(Y)}$ is $(-\bar{Y}/V^{1/2}, V^{-1/2})$. Hence a maximal invariant is the vector of residuals

$$A = (Y - \bar{Y})/V^{1/2} = \left(\frac{Y_1 - \bar{Y}}{V^{1/2}}, \ldots, \frac{Y_n - \bar{Y}}{V^{1/2}}\right)^T,$$

called the *configuration*. It is easily checked directly that the distribution of A does not depend on θ. Any function of A is also invariant. The orbits are determined by different values a of the statistic A, and Y has a unique representation as $Y = g_{s(Y)}(A) = \bar{Y} + V^{1/2}A$.

5 Principles of Inference and Data Reduction

5.1 Likelihood

We have a parametric model, involving a model function $f_Y(y;\theta)$ for a random variable Y and parameter $\theta \in \Omega_\theta$. The likelihood function is

$$L(\theta; y) = f_Y(y;\theta).$$

Usually we work with the log-likelihood

$$l(\theta; y) = \log f_Y(y;\theta).$$

Quite generally, even for dependent random variables, if $Y_{(j)} = (Y_1, \ldots, Y_j)$, we may write

$$l(\theta; y) = \sum_{j=1}^{n} l_{Y_j|Y_{(j-1)}}(\theta; y_j \mid y_{(j-1)}),$$

each term being computed from the conditional density given all the previous values in the sequence.

A simple example concerning Bernoulli trials illustrates this. The log likelihood function corresponding to r successes in n trials is essentially the same whether (i) only the number of successes in a prespecified number of trials is recorded or (ii) only the number of trials necessary to achieve a prespecified number of successes is recorded, or (iii) whether the detailed results of individual trials are recorded, with an arbitrary data-dependent stopping rule.

5.2 Sufficiency

5.2.1 Definitions

Let the data y correspond to a random variable Y with density $f_Y(y;\theta), \theta \in \Omega_\theta$. Let $s(y)$ be a statistic such that if $S \equiv s(Y)$ denotes the corresponding random variable, then the conditional density of Y given $S = s$ does not depend on θ, for all s, so that

$$f_{Y|S}(y \mid s;\theta) = g(y, s) \tag{5.1}$$

for all $\theta \in \Omega_\theta$. Then S is said to be sufficient for θ.

The definition (5.1) does not define S uniquely. We usually take the minimal S for which (5.1) holds, the minimal sufficient statistic. S is minimal sufficent if it is sufficient and is a function of every other sufficient statistic.

The determination of S from the definition (5.1) is often difficult. Instead we use the factorisation theorem: a necessary and sufficient condition that S is sufficient for θ is that for all y, θ

$$f_Y(y;\theta) = g(s,\theta)h(y),$$

for some functions g and h. Without loss of generality, $g(s,\theta)$ may be taken as the unconditional density of S for given θ.

The following result is easily proved (see Young and Smith, 2005, pp.92–93) and useful for identifying minimal sufficient statistics. A statistic T is minimal sufficient iff

$$T(x) = T(y) \Leftrightarrow \frac{L(\theta_1;x)}{L(\theta_2;x)} = \frac{L(\theta_1;y)}{L(\theta_2;y)}, \quad \forall \theta_1, \theta_2 \in \Omega_\theta.$$

5.2.2 Examples

Exponential models Here the natural statistic S is a (minimal) sufficient statistic. In a curved (m,d) exponential model the dimension m of the sufficient statistic exceeds that of the parameter.

Transformation models Except in special cases, such as the normal distribution, where the model is also an exponential family model, there is no reduction of dimensionality by sufficiency: the minimal sufficient statistic has the same dimension as the data vector $Y = (Y_1, \ldots, Y_n)$.

5.3 Completeness

A sufficient statistic $T(Y)$ is **complete** if for any real function g,

$$\mathbb{E}_\theta\{g(T)\} = 0 \text{ for all } \theta$$

implies

$$\Pr_\theta\{g(T) = 0\} = 1 \text{ for all } \theta.$$

This definition has a number of consequences. For instance, if there exists an unbiased estimator of a scalar parameter θ which is a function of a complete sufficient statistic T, then it is the *unique* such estimator (except possibly on

a set of measure 0). This follows because if, for instance, $g_1(T)$ and $g_2(T)$ are two such estimators, then $\mathbb{E}_\theta\{g_1(T) - g_2(T)\} = \theta - \theta = 0$, so $g_1(T) = g_2(T)$ with probability 1.

A key example of a complete sufficient statistic is the following: if $S \equiv s(Y) = (s_1(Y), ..., s_m(Y))$ is the natural statistic for a full exponential family in its natural parametrisation, as given by (4.1), and if Ω_ϕ contains an open rectangle in $\mathbb{R}^m$, then S is complete.

5.4 Conditioning

In connection with methods of statistical inference, probability is used in two quite distinct ways. The first is to define the stochastic model assumed to have generated the data. The second is to assess uncertainty in conclusions, via significance levels, confidence regions, posterior distributions etc. We enquire how a given method would perform if, hypothetically, it were used repeatedly on data derived from the model under study. The probabilities used for the basis of inference are long-run frequencies under hypothetical repetition. The issue arises of how these long-run frequencies are to be made relevant to the data under study. The answer lies in conditioning the calculations so that the long run matches the particular set of data in important respects.

5.4.1 The Bayesian stance

In a Bayesian approach the issue of conditioning is dealt with automatically. It is supposed that the particular value of θ is the realised value of a random variable Θ, generated by a random mechanism giving a known density $\pi_\Theta(\theta)$ for Θ, the prior density. Then Bayes' Theorem gives the posterior density

$$\pi_{\Theta|Y}(\theta \mid Y = y) \propto \pi_\Theta(\theta) f_{Y|\Theta}(y \mid \Theta = \theta),$$

where now the model function $f_Y(y; \theta)$ is written as a conditional density $f_{Y|\Theta}(y \mid \Theta = \theta)$. The insertion of a random element in the generation of θ allows us to condition on the whole of the data y: relevance to the data is certainly accomplished. This approach is uncontroversial if a meaningful prior can be agreed. In many applications, there may be major obstacles to specification of a meaningful prior and we are forced to adopt a less direct route to conditioning.

5.4.2 The Fisherian stance

Suppose first that the whole parameter vector θ is of interest. Reduce the problem by sufficiency. If, with parameter dimension $d = 1$, there is a one-dimensional sufficient statistic, we have reduced the problem to that of one observation from a distribution with one unknown parameter and there is little choice but to use probabilities calculated from that distribution. The same notion occurs if there is a d-dimensional θ of interest and a d-dimensional sufficient statistic. If the dimension of the (minimal) sufficient statistic exceeds that of the parameter, there is scope and need for ensuring relevance to the data under analysis by conditioning.

We therefore aim to

1. partition the minimal sufficient statistic s in the form $s = (t, a)$, so that $\dim(t) = \dim(\theta)$ and A has a distribution not involving θ;
2. use for inference the conditional distribution of T given $A = a$.

Conditioning on $A = a$ makes the distribution used for inference involve (hypothetical) repetitions like the data in some respects.

In the next section we extend this discussion to the case where there are nuisance parameters.

5.4.3 An example

Suppose that $Y_1, \ldots, Y_n$ are independent and identically uniformly distributed on $(\theta-1, \theta+1)$. The (minimal) sufficient statistic is the pair of order statistics $(Y_{(1)}, Y_{(n)})$, where $Y_{(1)} = \min\{Y_1, \ldots, Y_n\}$ and $Y_{(n)} = \max\{Y_1, \ldots, Y_n\}$. Suppose we make a (one-to-one) transformation to the mid-range $\bar{Y} = \frac{1}{2}(Y_{(1)} + Y_{(n)})$ and the range $R = Y_{(n)} - Y_{(1)}$. The sufficient statistic may equivalently be expressed as $(\bar{Y}, R)$. A direct calculation shows that R has a distribution not depending on θ, so we have the situation where the dimension of the sufficient statistic exceeds the dimension of θ and the statistic R, being distribution constant, plays the role of A. Inference should be based on the conditional distribution of $\bar{Y}$, given $R = r$, which it is easily checked to be uniform over $(\theta - 1 + \frac{1}{2}r, \theta + 1 - \frac{1}{2}r)$.

5.5 Ancillarity and the Conditionality Principle

A component a of the minimal sufficient statistic such that the random variable A is distribution constant is said to be ancillary, or sometimes ancillary in the simple sense.

The Conditionality Principle says that inference about a parameter of interest θ is to be made conditional on $A = a$ i.e. on the basis of the conditional distribution of Y given $A = a$, rather than from the model function $f_Y(y; \theta)$.

The Conditionality Principle is discussed most frequently in the context of transformation models, where the maximal invariant is ancillary.

5.5.1 Nuisance parameter case

In our previous discussion, the argument for conditioning on $A = a$ rests not so much on the distribution of A being known as on its being totally uninformative about the parameter of interest.

Suppose, more generally, that we can write $\theta = (\psi, \chi)$, where ψ is of interest. Suppose that

1. $\Omega_\theta = \Omega_\psi \times \Omega_\chi$, so that ψ and χ are variation independent;
2. the minimal sufficient statistic $s = (t, a)$;
3. the distribution of T given $A = a$ depends only on ψ;
4. one or more of the following conditions holds:
   (a) the distribution of A depends only on χ and not on ψ;
   (b) the distribution of A depends on (ψ, χ) in such a way that from observation of A alone no information is available about ψ;

Then the extension of the Fisherian stance of Section 5.4.2 argues that inference about ψ should be based upon the conditional distribution of T given $A = a$, and we would still speak of A as being ancillary and refer to this stance as the Conditionality Principle. The most straightforward extension corresponds to (a). In this case A is said to be a cut and to be S-ancillary for ψ and S-sufficient for χ. The arguments for conditioning on $A = a$ when ψ is the parameter of interest are as compelling as in the case where A has a fixed distribution. Condition (b) is more problematical to qualify. See the discussion in Barndorff-Nielsen and Cox (1994, pp.38–41) for detail and examples. The same authors discuss problems associated with existence and non-uniqueness of ancillary statistics.

6 Key Elements of Frequentist Theory

In this section we describe key elements of frequentist theory, starting with the Neyman-Pearson framework for hypothesis testing. The fundamental notion is that of seeking a test which maximises **power**, the probability under repeated sampling of correctly rejecting an incorrect hypothesis, subject to some pre-specified fixed **size**, the probability of incorrectly rejecting a true hypothesis. We end with a brief discussion of optimal point estimation.

6.1 Formulation of the hypothesis testing problem

Throughout we have a parameter space Ω_θ, and consider hypotheses of the form

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1$$

where Θ_0 and Θ_1 are two disjoint subsets of Ω_θ, possibly, but not necessarily, satisfying $\Theta_0 \cup \Theta_1 = \Omega_\theta$.

If a hypothesis consists of a single member of Ω_θ, for example if $\Theta_0 = \{\theta_0\}$ for some $\theta_0 \in \Omega_\theta$, then we say that it is a **simple** hypothesis. Otherwise it is called **composite**.

Sometimes hypotheses which at first sight appear to be simple hypotheses are really composite. This is especially common when we have **nuisance parameters**. For example, suppose $Y_1, ..., Y_n$ are independent, identically distributed $N(\mu, \sigma^2)$, with μ and σ^2 both unknown, and we want to test $H_0 : \ \mu = 0$. This is a composite hypothesis because $\Omega_\theta = \{(\mu, \sigma^2) : \ -\infty < \mu < \infty, \ 0 < \sigma^2 < \infty\}$ while $\Theta_0 = \{(\mu, \sigma^2) : \ \mu = 0, \ 0 < \sigma^2 < \infty\}$. Here σ^2 is a nuisance parameter: it does not enter into the hypothesis we want to test, but nevertheless we have to take it into account in constructing a test.

For most problems we adopt the following criterion: fix a small number α (often fixed to be 0.05, but any value in (0,1) is allowable), and seek a test of **size** α, so that

$$\Pr_\theta\{\text{Reject } H_0\} \leq \alpha \quad \text{for all } \theta \in \Theta_0.$$

Thus H_0 and H_1 are treated *asymetrically*. Usually H_0 is called the **null hypothesis** and H_1 the **alternative hypothesis**.

6.1.1 Test functions

The usual way hypothesis testing is formulated in elementary statistics texts is as follows: choose a test statistic $t(Y)$ (some function of the observed data Y) and a critical region C_α, then reject H_0 based on $Y = y$ if and only if $t(y) \in C_\alpha$. The critical region must be chosen to satisfy

$$\Pr_\theta\{t(Y) \in C_\alpha\} \leq \alpha \text{ for all } \theta \in \Theta_0.$$

We consider here a slight reformulation of this. Define a **test function** $\phi(y)$ by

$$\phi(y) = \begin{cases} 1 & \text{if } t(y) \in C_\alpha, \\ 0 & \text{otherwise.} \end{cases}$$

So whenever we observe $\phi(Y) = 1$, we reject H_0, while if $\phi(Y) = 0$, we accept.

In decision theory it was necessary sometimes to adopt a randomised decision rule. The same concept arises in hypothesis testing as well: sometimes we want to use a randomised test. This may be done by generalising the concept of a test function to allow $\phi(y)$ to take on any value in the interval $[0, 1]$. Thus having observed data y and evaluated $\phi(y)$, we use some independent randomisation device to draw a Bernoulli random number W which takes value 1 with probability $\phi(y)$, and 0 otherwise. We then reject H_0 if and only if $W = 1$. Thus we may interpret $\phi(y)$ to be "the probability that H_0 is rejected when $Y = y$".

If we want to construct a theory of hypothesis tests of a *given* size, we have to allow the possibility of randomised tests, regardless of whether we would actually want to use a test in a practical problem.

6.1.2 Power

We now need some criterion for deciding whether one test is better than another. We do this by introducing the concept of **power**.

The power function of a test ϕ is defined to be

$$w(\theta) = \Pr_\theta\{\text{Reject } H_0\} = \mathbb{E}_\theta\{\phi(Y)\}$$

which is defined for all $\theta \in \Omega_\theta$. When testing a simple null hypothesis against a simple alternative hypothesis, the term 'power' is often used to signify the

probability of rejecting the null hypothesis when the alternative hypothesis is true.

The idea is this: a good test is one which makes $w(\theta)$ as large as possible on Θ_1 while satisfying the constraint $w(\theta) \leq \alpha$ for all $\theta \in \Theta_0$.

Within this framework, we can consider various classes of problems:

(i) Simple H_0 vs. simple H_1: here there is an elegant and complete theory which tells us exactly how to construct the best test, given by the Neyman-Pearson Theorem.

(ii) Simple H_0 vs. composite H_1: in this case the obvious approach is to pick out a representative value of Θ_1, say θ_1, and construct the Neyman-Pearson test of H_0 against θ_1. In some cases the test so constructed is the same for every $\theta_1 \in \Theta_1$. When this happens, the test is called "uniformly most powerful" or UMP. We would obviously like to use a UMP test if we can find one, but there are many problems for which UMP tests do not exist, and then the whole problem is harder.

(iii) Composite H_0 vs. composite H_1: in this case the problem is harder again. It may not be so easy even to find a test which satisfies the size constraint, because of the requirement that $\mathbb{E}_\theta\{\phi(X)\} \leq \alpha$ for all $\theta \in \Theta_0$; if Θ_0 contains a nuisance parameter such as σ^2 in the above $N(\mu, \sigma^2)$ example, we must find a test which satisfies this constraint regardless of the value of σ^2.

6.2 The Neyman-Pearson Theorem

Consider the test of a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative hypothesis $H_1 : \theta = \theta_1$, where θ_0 and θ_1 are specified. Let the probability density function or probability mass function of Y be $f(y;\theta)$, specialised to $f_0(y) = f(y;\theta_0)$ and $f_1(y) = f(y;\theta_1)$. Define the likelihood ratio $\Lambda(y)$ by

$$\Lambda(y) = \frac{f_1(y)}{f_0(y)}.$$

According to the Neyman-Pearson Theorem, the best test of size α is of the form: reject H_0 when $\Lambda(Y) > k_\alpha$ where k_α is chosen so as to guarantee that the test has size α. However, we have seen above that this method of constructing the test can fail when Y has a discrete distribution (or more precisely, when $\Lambda(Y)$ has a discrete distribution under H_0). In the following generalised form of Neyman-Pearson Theorem, we remove this difficulty by allowing for the possibility of randomised tests.

The (randomised) test with test function ϕ_0 is said to be a **likelihood ratio test** (LRT for short) if it is of the form

$$\phi_0(y) = \begin{cases} 1 & \text{if } f_1(y) > Kf_0(y), \\ \gamma(y) & \text{if } f_1(y) = Kf_0(y), \\ 0 & \text{if } f_1(y) < Kf_0(y), \end{cases}$$

where $K \geq 0$ is a constant and $\gamma(y)$ an arbitrary function satisfying $0 \leq \gamma(y) \leq 1$ for all y.

Theorem 6.1 (Neyman-Pearson)

(a) (Optimality). For any K and $\gamma(y)$, the test ϕ_0 has maximum power among all tests whose size is no greater than the size of ϕ_0.

(b) (Existence). Given $\alpha \in (0, 1)$, there exist constants K and γ_0 such that the LRT defined by this K and $\gamma(y) = \gamma_0$ for all y has size exactly α.

(c) (Uniqueness). If the test ϕ has size α, and is of maximum power amongst all possible tests of size α, then ϕ is necessarily a likelihood ratio test, except possibly on a set of values of y which has probability 0 under both H_0 and H_1.

Proof of the Theorem is straightforward: see, for example, Young and Smith (2005, pp.68–69).

6.3 Uniformly most powerful tests

A **uniformly most powerful** or UMP test of size α is a test $\phi_0(\cdot)$ for which

(i) $\mathbb{E}_\theta \phi_0(Y) \leq \alpha$ for all $\theta \in \Theta_0$;

(ii) Given any other test $\phi(\cdot)$ for which $\mathbb{E}_\theta \phi(Y) \leq \alpha$ for all $\theta \in \Theta_0$, we have $\mathbb{E}_\theta \phi_0(Y) \geq \mathbb{E}_\theta \phi(Y)$ for all $\theta \in \Theta_1$.

In general, it is asking a very great deal to expect that UMP tests exist – in effect, it is asking that the Neyman-Pearson test for simple vs. simple hypotheses should be the same for *every* pair of simple hypotheses contained within H_0 and H_1. Nevertheless, for one-sided testing problems involving just a single parameter, for which $\Omega_\theta \subseteq \mathbb{R}$, there is a wide class of parametric families for which just such a property holds. Such families are said to have **monotone likelihood ratio** or MLR.

6.3.1 Monotone Likelihood Ratio

Definition. The family of densities $\{f(y;\theta),\ \theta \in \Omega_\theta \subseteq \mathbb{R}\}$ with real scalar parameter θ is said to be of **monotone likelihood ratio** (MLR for short) if there exists a function $t(y)$ such that the likelihood ratio

$$\frac{f(y;\theta_2)}{f(y;\theta_1)}$$

is a non-decreasing function of $t(y)$ whenever $\theta_1 \leq \theta_2$.

Note that any family for which the likelihood ratio turned out to be non-increasing (rather than non-decreasing) as a function of $t(y)$ is still MLR: simply replace $t(y)$ by $-t(y)$.

The main result of this section is that for a one-sided test in a MLR family a UMP test exists. For simplicity, we restrict ourselves to absolutely continuous distributions so as to avoid the complications of randomised tests.

Theorem 6.2 Suppose Y has a distribution from a family which is MLR with respect to a statistic $t(Y)$, and that we wish to test $H_0 :\ \theta \leq \theta_0$ against $H_1 :\ \theta > \theta_0$. Suppose the distribution function of $t(Y)$ is continuous.

(a) The test

$$\phi_0(y) = \begin{cases} 1 & \text{if } t(y) > t_0, \\ 0 & \text{if } t(y) \leq t_0, \end{cases}$$

is UMP among all tests of size $\leq \mathbb{E}_{\theta_0}\{\phi_0(Y)\}$.

(b) Given some α, where $0 < \alpha \leq 1$, there exists some t_0 such that the test in (a) has size exactly α.

A proof is provided by Young and Smith (2005, pp.72–73).

6.4 Two-sided tests and conditional inference

Our discussion here is concerned with two separate but interrelated themes. The first has to do with extending the discussion above to more complicated hypothesis testing problems, and the second is concerned with conditional inference.

We will consider first testing two-sided hypotheses of the form $H_0 :\ \theta \in [\theta_1, \theta_2]$ (with $\theta_1 < \theta_2$) or $H_0 :\ \theta = \theta_0$ where, in each case, the alternative H_1 includes all θ not part of H_0. For such problems we cannot expect to find a uniformly most powerful test However, by introducing an additional

concept of **unbiasedness** (Section 6.4.2), we are able to define a family of **uniformly most powerful unbiased**, or UMPU, tests. In general, characterising UMPU tests for two-sided problems is a much harder task than characterising UMP tests for one-sided hypotheses, but for one specific but important example, that of a one-parameter exponential family, we are able to find UMPU tests. The details of this are the subject of Section 6.4.3.

The extension to multiparameter exponential families involves the notion of **conditional tests**, discussed in Section 6.4.5. In some situations, a statistical problem may be greatly simplified by working not with the unconditional distribution of a test statistic, but the conditional distribution given some other statistic. We discuss two situations where conditional tests naturally arise, one when there are **ancillary statistics**, and the other where conditional procedures are used to construct **similar tests**. Recall that the basic idea behind an ancillary statistic is that of a quantity with distribution not depending on the parameter of interest. The Fisherian paradigm then argues that **relevance** to the data at hand demands conditioning on the observed value of this statistic. The notion behind similarity is that of eliminating dependence on nuisance parameters. We specialise to the case of a multiparameter exponential family in which one particular parameter is of interest while the remaining $m-1$ are regarded as nuisance parameters.

6.4.1 Two-sided hypotheses and two-sided tests

We consider a general situation with a one-dimensional parameter $\theta \in \Omega_\theta \subseteq \mathbb{R}$. We are particularly interested in the case when the null hypothesis is $H_0 : \ \theta \in \Theta_0$ where Θ_0 is either the interval $[\theta_1, \theta_2]$ for some $\theta_1 < \theta_2$, or else the single point $\Theta_0 = \{\theta_0\}$, and $\Theta_1 = \mathbb{R} \setminus \Theta_0$.

If we have an exponential family with natural statistic $S = s(Y)$, or a family with MLR with respect to $s(Y)$, we might still expect tests of the form

$$\phi(y) = \begin{cases} 1 & \text{if} \quad s(y) > t_2 \text{ or } s(y) < t_1, \\ \gamma(y) & \text{if} \quad s(y)) = t_2 \text{ or } s(y) = t_1, \\ 0 & \text{if} \quad t_1 < s(y) < t_2, \end{cases}$$

where $t_1 < t_2$ and $0 \leq \gamma(y) \leq 1$, to have good properties. Such tests are called **two-sided tests**.

6.4.2 Unbiased tests

Definition A test ϕ of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is called **unbiased of size** α if

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta\{\phi(Y)\} \leq \alpha$$

and

$$\mathbb{E}_\theta\{\phi(Y)\} \geq \alpha \text{ for all } \theta \in \Theta_1.$$

An unbiased test captures the natural idea that the probability of rejecting H_0 should be higher when H_0 is false than when it is true.

Definition A test which is uniformly most powerful amongst the class of all unbiased tests is called **uniformly most powerful unbiased**, abbreviated UMPU.

The requirement that a test be unbiased is one way of resolving the obvious conflict between the two sides of a two-sided alternative hypothesis. We use it as a criterion by which to assess two-sided tests. Nevertheless the objections to unbiasedness that we have noted previously are still present — unbiasedness is not by itself an optimality criterion and, for any particular decision problem, there is no reason why the optimal decision procedure should turn out to be unbiased. The principal role of unbiasedness is to restrict the class of possible decision procedures and hence to make the problem of determining an optimal procedure more manageable than would otherwise be the case.

6.4.3 UMPU tests for one-parameter exponential families

Consider an exponential family for a random variable Y, which may be a vector of independent, identically distributed observations, with real-valued parameter $\theta \in \mathbb{R}$ and density of form

$$f(y;\theta) = c(\theta)h(y)e^{\theta s(y)},$$

where $S = s(Y)$ is a real-valued natural statistic.

Then S also has an exponential family distribution, with density of form

$$f_S(s;\theta) = c(\theta)h_S(s)e^{\theta s}.$$

We shall assume that S is a continuous random variable with $h_S(s) > 0$ on the open set which defines the range of S. By restricting ourselves to families of this form we avoid the need for randomised tests and make it

easy to prove the existence and uniqueness of two-sided tests, though in a more general version of the theory such assumptions are not required: see, for example Ferguson (1967).

We consider initially the case

$$\Theta_0 = [\theta_1, \theta_2], \quad \Theta_1 = (-\infty, \theta_1) \cup (\theta_2, \infty),$$

where $\theta_1 < \theta_2$.

Theorem 6.3 Let ϕ be any test function. Then there exists a unique two-sided test ϕ' which is a function of S such that

$$\mathbb{E}_{\theta_j}\phi'(Y) = \mathbb{E}_{\theta_j}\phi(Y), \quad j = 1, 2.$$

Moreover,

$$\mathbb{E}_\theta\phi'(Y) - \mathbb{E}_\theta\phi(Y) \begin{cases} \leq 0 & \text{for } \theta_1 < \theta < \theta_2, \\ \geq 0 & \text{for } \theta < \theta_1 \text{ or } \theta > \theta_2. \end{cases}$$

Corollary For any $\alpha > 0$, there exists a UMPU test of size α, which is of two-sided form in S.

6.4.4 Testing a point null hypothesis

Now consider the case $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ for a given value of θ_0. By analogy with the case just discussed, letting $\theta_2 - \theta_1 \to 0$, there exists a two-sided test ϕ' for which

$$\mathbb{E}_{\theta_0}\{\phi'(Y)\} = \alpha, \quad \frac{d}{d\theta}\mathbb{E}_\theta\{\phi'(Y)\}\bigg|_{\theta=\theta_0} = 0.$$

Such a test is in fact UMPU, but we shall not prove this directly. We note in passing that differentiability, as a function of θ, of the power function (for any test function) is a consequence of our assumption of an exponential family distribution.

6.4.5 Conditional inference, ancillarity and similar tests

Consider the following hypothetical situation. An experiment is conducted to measure the carbon monoxide level in the exhaust of a car. A sample of exhaust gas is collected, and is taken along to the laboratory for analysis. Inside the laboratory are two machines, one of which is expensive and very

accurate, the other an older model which is much less accurate. We will use the accurate machine if we can, but this may be out of service or already in use for another analysis. We do not have time to wait for this machine to become available, so if we cannot use the more accurate machine we use the other one instead (which is always available). Before arriving at the laboratory we have no idea whether the accurate machine will be available, but we do know that the probability that it is available is $\frac{1}{2}$ (independently from one visit to the next).

This situation may be formalised as follows: we observe (δ, Y), where δ (=1 or 2) represents the machine used and Y the subsequent observation. The distributions are $\Pr\{\delta = 1\} = \Pr\{\delta = 2\} = \frac{1}{2}$ and, given δ, $Y \sim N(\theta, \sigma_\delta^2)$ where θ is unknown and σ_1, σ_2 are known, with $\sigma_1 < \sigma_2$. We want to test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. Consider the following tests:

Procedure 1. Reject H_0 if $Y > c$, where c is chosen so that the test has prescribed size α,

$$\Pr(Y > c) = \Pr(Y > c \mid \delta = 1)\Pr(\delta = 1) + \Pr(Y > c \mid \delta = 2)\Pr(\delta = 2) = \alpha,$$

which requires

$$\frac{1}{2}\left\{1 - \Phi\left(\frac{c - \theta_0}{\sigma_1}\right)\right\} + \frac{1}{2}\left\{1 - \Phi\left(\frac{c - \theta_0}{\sigma_2}\right)\right\} = \alpha.$$

Procedure 2. Reject H_0 if $Y > z_\alpha \sigma_\delta + \theta_0$, where z_α is the upper α-quantile of $N(0, 1)$.

Thus Procedure 1 sets a single critical level c, regardless of which machine is used, while Procedure 2 determines its critical level solely on the standard deviation for the machine that was actually used, without taking the other machine into account at all. Procedure 2 is called a *conditional* test because it conditions on the observed value of δ. Note that the distribution of δ itself does not depend in any way on the unknown parameter θ, so we are not losing any information by doing this.

Intuitively, one might expect Procedure 2 to be more reasonable, because it makes sense to use all the information available and one part of that information is which machine was used. However, if we compare the two in terms of power, our main criterion for comparing tests up until now, it is not so clear-cut. Figure 6.1 shows the power curves of the two tests in the case $\sigma_1 = 1,\ \sigma_2 = 3,\ \alpha = 0.05$, for which $z_\alpha = 1.6449$ and it is determined numerically that $c = 3.8457 + \theta_0$. When the difference in means, $\theta_1 - \theta_0$,

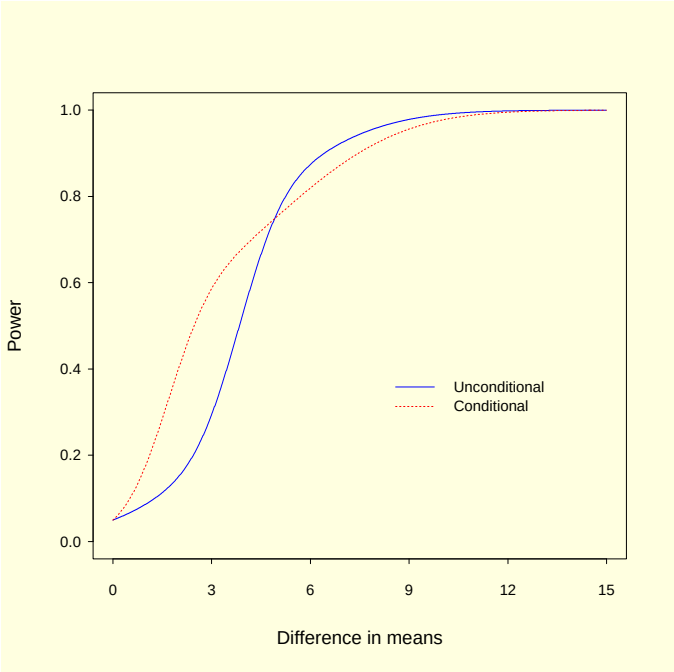

Figure 6.1: Power functions of tests for normal mixture problem.

is small, Procedure 2 is much more powerful, but for larger values when $\theta_1 > \theta_0 + 4.9$, Procedure 1 is better.

At first sight this might seem counterintuitive, but closer thought shows what is going on. Let us compute $\alpha_j = \Pr_{\theta_0}\{Y > c|\delta = j\}$ — we find $\alpha_1 = 0.00006$, $\alpha_2 = 0.09994$ (so that the overall size is $(\alpha_1 + \alpha_2)/2 = 0.05$). For large $\theta_1 - \theta_0$, this extra power when $\delta = 2$ is decisive in allowing procedure 1 to perform better than procedure 2. But is this really sensible? Consider the following scenario.

Smith and Jones are two statisticians. Smith works for the environmental health department of Cambridge City Council and Jones is retained as a consultant by a large haulage firm which operates in the Cambridge area. Smith carries out a test of the exhaust fumes emitted by one of the lorries belonging to the haulage firm. On this particular day he has to use machine

2 and the observation is $Y = \theta_0 + 4.0$, where θ_0 is the permitted standard. It has been agreed in advance that all statistical tests will be carried out at the 5% level and therefore, following Procedure 1 above, he reports that the company is in violation of the standard.

The company is naturally not satisfied with this conclusion and therefore sends the results to Jones for comment. The information available to Jones is that a test was conducted on a machine for which the standard deviation of all measurements is 3 units, that the observed measurement exceeded the standard by 4 units, and that therefore the null hypothesis (that the lorry is meeting the standard) is rejected at the 5% level. Jones calculates that the critical level should be $\theta_0 + 3z_{0.05} = \theta_0 + 3 \times 1.645 = \theta_0 + 4.935$ and therefore queries why the null hypothesis was rejected.

The query is referred back to Smith who now describes the details of the test, including the existence of the other machine and Smith's preference for Procedure 1 over Procedure 2 on the grounds that Procedure 1 is of higher power when $|\theta_1 - \theta_0|$ is large. This however is all news to Jones, who was not previously aware that the other machine even existed.

The question facing Jones now is: should she revise her opinion on the basis of the new information provided by Smith? She does not see why she should. There is no new information about either the sample that was collected or the way that it was analysed. All that is new is that there was another machine which might have been used for the test, but which in the event was unavailable. Jones cannot see why this is relevant. Indeed, given the knowledge that there are two machines and that the probability of a false positive test (when the company is complying with the standard) is much higher using machine 2 than machine 1, she might be inclined to query the circumstances under which machine 2 was chosen to test her company's sample. She therefore advises the company to challenge the test in court.

In terms of our discussion in Section 5.5, the minimal sufficient statistic for θ is (Y, δ), with δ having a distribution which does not depend on θ. The Conditionality Principle then argues that inference should be made conditional on the observed value of δ.

The conclusion we can draw from this discussion is that while maximising power is a well-established principle for choosing among statistical tests, there are occasions when it can lead to conclusions that appear to contradict common sense, and where the Conditionality Principle is compelling.

6.4.6 Discussion

The need for a conditionality principle highlights a weakness in the emphasis on the power of tests which is characteristic of the Neyman-Pearson theory, and more generally, in the emphasis on the risk function which is central to non-Bayesian decision theory. If one uses power as the sole criterion for deciding between two tests, then in our example concerning laboratory testing there are at least some circumstances where one would prefer to use Procedure 1, but this may not be sensible for other reasons. The historical disagreement between Fisher and Neyman centred on Fisher's opinion that the Neyman-Pearson theory did not take adequate account of the need for conditional tests in this kind of situation. Another point of view might be to adopt a Bayesian approach. Bayesian procedures always try to minimise the expected loss based on the observed data and do not take account of other experiments that might have been conducted but were not. Thus in the situation with two machines discussed above, a Bayesian procedure would always act conditionally on which machine was actually used so the kind of conflict that we saw between the two statisticians would not arise. However, Fisher did not accept Bayesian methods, because of the seeming arbitrariness of choosing the prior distribution, and so this would not have resolved the difficulty for him!

6.4.7 Similar tests

Suppose we have $\theta = (\psi, \lambda)$, with ψ the parameter of interest and λ a nuisance parameter. Suppose the minimal sufficient statistic T can be partitioned as $T = (S, C)$, where the conditional distribution of S given $C = c$ depends on ψ, but not on λ, for each c. [We don't necessarily require the distribution of C to depend only on λ, as in our discussion of ancillarity]. We may construct a test on the interest parameter based on the conditional distribution of S given C. The reason is that such a test will then be **similar**.

Definition Suppose $\theta = (\psi, \lambda)$ and the parameter space is of the form $\Omega_\theta = \Omega_\psi \times \Omega_\lambda$. Suppose we wish to test the null hypothesis $H_0 : \psi = \psi_0$ against the alternative $H_1 : \psi \neq \psi_0$, with λ treated as a nuisance parameter. Suppose $\phi(y),\ y \in \mathcal{Y}$ is a test of size α for which

$$\mathbb{E}_{\psi_0,\lambda}\{\phi(Y)\} = \alpha \text{ for all } \lambda \in \Omega_\lambda.$$

Then ϕ is called a **similar test of size** α.

More generally, if the parameter space is $\theta \in \Omega_\theta$ and the null hypothesis is of the form $\theta \in \Theta_0$, where Θ_0 is a subset of Ω_θ, then a similar test is one for which $\mathbb{E}_\theta\{\phi(Y)\} = \alpha$ on the boundary of Θ_0.

By analogy with UMPU tests, if a test is uniformly most powerful among the class of all similar tests, we call it **UMP similar.**

The concept of similar tests has something in common with that of unbiased tests. In particular, if the power function is continuous in θ (a property which actually holds automatically for exponential families), then any unbiased test of size α must have power exactly α on the boundary between Θ_0 and Θ_1, i.e. such a test must be similar. In such cases, if we can find a UMP similar test, and if this test turns out also to be unbiased, then it is necessarily UMPU.

Moreover, in many cases we can demonstrate, under our assumptions, that a test which is UMP among all tests based on the conditional distribution of S given C, is UMP amongst all similar tests. In particular, this statement will be valid when C is a complete sufficient statistic for λ.

The upshot of this discussion is that there are many cases when a test which is UMP (one-sided) or UMPU (two-sided), based on the conditional distribution of S given C, is in fact UMP similar or UMPU among the class of all tests.

Note that we have now seen two quite distinct arguments for conditioning. In the first, when the conditioning statistic is ancillary, we have seen that the failure to condition may lead to paradoxical situations in which two analysts may form completely different viewpoints of the same data, though we also saw that the application of this principle may run counter to the strict Neyman-Pearson viewpoint of maximising power. The second point of view is based on power, and shows that under certain circumstances a conditional test may satisfy the conditions needed to be UMP similar or UMPU.

6.4.8 Multiparameter exponential families

Consider a full exponential family model in its natural parametrisation,

$$f(y;\theta) = c(\theta)h(y)\exp\left(\sum_{i=1}^{m} t_i(y)\theta^i\right),$$

where y represents the value of a data vector Y and $t_i(Y),\ i = 1, ..., m$ are the natural statistics. We also write T_i in place of $t_i(Y)$.

Suppose our main interest is in one particular parameter, which we may without loss of generality take to be θ^1. Consider the test $H_0 : \theta^1 \leq \theta^{1*}$ against

$H_1 : \theta^1 > \theta^{1*}$, where θ^{1*} is prescribed. Take $S = T_1$ and $C = (T_2, ..., T_m)$. Then the conditional distribution of S given C is also of exponential family form and does not depend on $\theta^2, ..., \theta^m$. Therefore, C is sufficient for $\lambda = (\theta^2, ..., \theta^m)$ and since it is also complete (from the general property that the natural statistics are complete sufficient statistics for exponential families) the arguments concerning similar tests suggest that we ought to construct tests for θ^1 based on the conditional distribution of S given C.

In fact such tests do turn out to be UMPU, though we shall not attempt to fill in the details of this: the somewhat intricate argument is given by Ferguson (1967). Finally, it sometimes (though not always) turns out that C is an ancillary statistic for θ^1, so has a distribution not depending on θ^1. When this happens, there is a far stronger argument based on the conditionality principle that says we ought to condition on C.

In cases where the distribution of T_1 is continuous, the optimal one-sided test will then be of the following form. Suppose we observe $T_1 = t_1, ..., T_m = t_m$. Then we reject H_0 if and only if $t_1 > t_1^*$, where t_1^* is calculated from

$$\Pr_{\theta^{1*}}\{T_1 > t_1^* | T_2 = t_2, ..., T_m = t_m\} = \alpha.$$

It can be shown that this test is UMPU of size α.

The following result is often useful. Suppose, as above, $C = (T_2, \ldots, T_m)$, and suppose that $V \equiv V(T_1, C)$ is a statistic independent of C, with $V(t_1, c)$ increasing in t_1 for each c. Then the UMPU test above is equivalent to that based on the marginal distribution of V. The conditional test is the same as that obtained by testing H_0 against H_1 using V as test statistic. An example is provided by the normal distribution $N(\mu, \sigma^2)$: given an independent sample $X_1, \ldots, X_n$, to test a hypothesis about σ^2, the conditional test is based on the conditional distribution of $T_1 \equiv \sum_{i=1}^n X_i^2$, given the observed value of $C \equiv \bar{X}$. Let $V = T_1 - nC^2 \equiv \sum_{i=1}^n (X_i - \bar{X})^2$. We know that V is independent of C (from general properties of the normal distribution), so the optimal conditional test is equivalent to that based on the marginal distribution of V: we have that V/σ^2 is chi-squared, χ^2_{n-1}.

In similar fashion, if we want to construct a two-sided test of $H_0 : \theta^{1*} \leq \theta^1 \leq \theta^{1**}$ against the alternative, $H_1 : \theta^1 < \theta^{1*}$ or $\theta^1 > \theta^{1**}$, where $\theta^{1*} < \theta^{1**}$ are given, we can proceed by defining the conditional power function of a test ϕ based on T_1 as

$$w_{\theta^1}(\phi; t_2, ..., t_m) = \mathbb{E}_{\theta^1}\{\phi(T_1) | T_2 = t_2, ..., T_m = t_m\}.$$

Note that it is a consequence of our previous discussion that this quantity depends only on θ^1 and not on $\theta^2, .., \theta^m$.

We can then consider a two-sided conditional test of the form

$$\phi'(t_1) = \begin{cases} 1 & \text{if } t_1 < t_1^* \text{ or } t_1 > t_1^{**}, \\ 0 & \text{if } t_1^* \leq t_1 \leq t_1^{**}, \end{cases}$$

where t_1^* and t_1^{**} are chosen such that

$$w_{\theta^1}(\phi'; t_2, ..., t_m) = \alpha \quad \text{when } \theta^1 = \theta^{1*} \text{ or } \theta^1 = \theta^{1**}.$$

If the hypotheses are of the form $H_0 : \ \theta^1 = \theta^{1*}$ against $H_1 : \ \theta^1 \neq \theta^{1*}$, then the test is of the same form but with (6.4.8) replaced by

$$w_{\theta^{1*}}(\phi'; t_2, ..., t_m) = \alpha, \quad \frac{d}{d\theta^1}\left\{ w_{\theta^1}(\phi'; t_2, ..., t_m) \right\}\Big|_{\theta^1 = \theta^{1*}} = 0.$$

It can be shown that these tests are also UMPU of size α.

6.5 Optimal point estimation

We finish by discussing optimal point estimation of a parameter θ.

Jensen's inequality is a well-known result that is proved in elementary analysis texts. It states that if $g : \mathbb{R} \to \mathbb{R}$ is a convex function [so that $g(\lambda x_1 + (1-\lambda)x_2) \leq \lambda g(x_1) + (1-\lambda)g(x_2)$ for all x_1, x_2 and $0 < \lambda < 1$] and X is a real-valued random variable, then $\mathbb{E}\{g(X)\} \geq g\{\mathbb{E}(X)\}$.

Theorem 6.4 Suppose we want to estimate a real-valued parameter θ with an estimator $d(Y)$ say. Suppose the loss function $L(\theta, d)$ is a convex function of d for each θ. Let $d_1(Y)$ be an (arbitrary) unbiased estimator for θ and suppose T is a sufficient statistic. Then the estimator

$$\chi(T) = \mathbb{E}\{d_1(Y)|T\}$$

is also unbiased and is at least as good as $d_1 : R(\theta, d_1) \geq R(\theta, \chi)$, where R is the risk.

Note that the definition of $\chi(T)$ does not depend on θ, because T is sufficient.

For a proof, see, for example, Young and Smith (2005, p.96).

Remark 1. The inequality will be strict unless L is a linear function of d, or the conditional distribution of $d_1(Y)$ given T is degenerate. In all other cases, $\chi(T)$ strictly dominates $d_1(Y)$.

Remark 2. If T is also complete, then $\chi(T)$ is the **unique** unbiased estimator minimising the risk.

Remark 3. If $L(\theta, d) = (\theta - d)^2$ then this is the Rao-Blackwell Theorem. In this case the risk of an unbiased estimator is just its variance, so the theorem asserts that there is a unique minimum variance unbiased estimator which is a function of the complete sufficient statistic. However it is still possible that there are biased estimators which achieve a smaller mean squared error: the example of a minimax estimator of the parameter of a binomial distribution given in Chapter 3 is one such, and Stein's paradox example is another.

References

Barndorff-Nielsen, O. E. and Cox, D. R. (1994) *Inference and Asymptotics*. London: Chapman & Hall.

Bayes, T. (1763) An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.* **53**, 370–418.

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, Second Edition. New York: Springer.

Berger, J. O. and Sellke, T. (1987) Testing a point null hypothesis: The irreconcilability of P values and evidence. *J. Amer. Statist. Assoc.* **82**, 112–122.

Blyth, C. R. (1951) On minimax statistical decision procedures and their admissibility. *Annals of Mathematical Statistics* **22**, 22–42.

Casella, G. and Berger, R. L. (1990) *Statistical Inference*. Pacific Grove, California: Wadsworth & Brooks/Cole.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman & Hall. [Classic text.]

Efron, B. (1998) R.A. Fisher in the 21st Century. *Statistical Science*, **13**, 95–122.

Efron, B. and Morris, C. M. (1975) Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70**, 311–319.

Efron, B. and Morris, C. M. (1977) Stein's paradox in statistics. *Scientific American* **236**, 119–127.

Ferguson, T. S. (1967) *Mathematical Statistics: A Decision-Theoretic Approach*. New York; Academic Press. [A classical reference work.]

Fisher, R. A. (1922) On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. A* **222**, 309–368. [This and the next reference are often debated as the most important single paper in statistical theory.]

Fisher, R. A. (1925) Theory of statistical estimation. *Proc. Camb. Phil. Soc.* **22**, 700-725.

Fisher, R. A. (1934) Two new properties of mathematical likelihood. *Proc. R. Soc. Lond. A* **144**, 285–307. [The origin of the conditional inference procedure for location-scale families, and therefore of much of the Fisherian viewpoint.]

Fisher, R. A. (1990) *Statistical Methods, Experimental Design and Scientific Inference*. Oxford: Clarendon Press. [A relatively recent reprint of three of Fisher's best known works on statistics.]

James, W. and Stein, C. (1961) Estimation with quadratic loss. *Proc. Fourth Berk. Symp. Math. Statist. Probab.* **1**, 361–379. Berkeley: University of California Press.

Jeffreys, H. (1939) *Theory of Probability*. Oxford: Oxford University Press. (Third edition: 1961.)

Neyman, J. and Pearson, E. S. (1933) On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. A*, **231**, 289–337.

Stein, C. (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berk. Symp. Math. Statist. Probab.* **1**, 197–206. Berkeley: University of California Press.

Stein, C. (1981) Estimation of the mean of a multivariate normal distribution. *Annals of Statistics* **9**, 1135–1151.

Problems

Problems are numbered as in Young & Smith (2005), but listed here in an order which is closer to that in which we will cover the material in the course.

2.1 Let X be uniformly distributed on $[0, \theta]$ where $\theta \in (0, \infty)$ is an unknown parameter. Let the action space be $[0, \infty)$ and the loss function $L(\theta, d) = (\theta - d)^2$ where d is the action chosen. Consider the decision rules $d_\mu(x) = \mu x, \mu \geq 0$. For what value of μ is d_μ unbiased? Show that $\mu = 3/2$ is a necessary condition for d_μ to be admissible.

2.3 Each winter evening between Sunday and Thursday, the superintendent of the Chapel Hill School District has to decide whether to call off the next day's school because of snow conditions. If he fails to call off school and there is snow, there are various possible consequences, including children and teachers failing to show up for school, the possibility of traffic accidents etc. If he calls off school, then regardless of whether there actually is snow that day, there will have to be a make-up day later in the year. After weighing up all the possible outcomes he decides that the costs of failing to close school when there is snow are twice the costs incurred by closing school, so he assigns two units of loss to the first outcome and one to the second. If he does not call off school and there is no snow, then of course there is no loss.

Two local radio stations give independent and identically distributed weather forecasts. If there is to be snow, each station will forecast this with probability 3/4, but predict no snow with probability 1/4. If there is to be no snow, each station predicts snow with probability 1/2.

The superintendent will listen to the two forecasts this evening, and then make his decision on the basis of the data x, the number of stations forecasting snow.

Write down an exhaustive set of non-randomised decision rules based on x.

Find the superintendent's admissible decision rules, and his minimax rule. Before listening to the forecasts, he believes there will be snow with probability 1/2; find the Bayes rule with respect to this prior.

[Include randomised rules in your analysis when determining admissible, minimax and Bayes rules].

2.5 Bacteria are distributed at random in a fluid, with mean density θ per unit volume, for some $\theta \in H \subseteq [0, \infty)$. This means that

$$\Pr_\theta(\text{no bacteria in volume } v) = e^{-\theta v}.$$

We remove a sample of volume v from the fluid and test it for the presence or absence of bacteria. On the basis of this information we have to decide whether there are any bacteria in the fluid at all. An incorrect decision will result in a loss of 1, a correct decision in no loss.

(i) Suppose $H = [0, \infty)$. Describe all the non-randomised decision rules for this problem and calculate their risk functions. Which of these rules are admissible?

(ii) Suppose $H = \{0, 1\}$. Identify the risk set

$$S = \{(R(0, d), R(1, d)) : d \text{ a randomised rule}\} \subseteq \mathbb{R}^2,$$

where $R(\theta, d)$ is the expected loss in applying d under $\Pr_\theta$. Determine the minimax rule.

(iii) Suppose again that $H = [0, \infty)$.

Determine the Bayes decision rules and Bayes risk for prior

$$\begin{aligned} \pi(\{0\}) &= 1/3, \\ \pi(A) &= 2/3 \int_A e^{-\theta} d\theta, \quad A \subseteq (0, \infty). \end{aligned}$$

[So the prior probability that $\theta = 0$ is $1/3$, while the prior probability that $\theta \in A \subseteq (0, \infty)$ is $2/3 \int_A e^{-\theta} d\theta$.]

(iv) If it costs $v/24$ to test a sample of volume v, what is the optimal volume to test? What if the cost is $1/6$ per unit volume?

2.6 Prove Theorems 2.3, 2.4 and 2.5, concerning admissibility of Bayes rules.

2.8 In a Bayes decision problem, a prior distribution π is said to be **least favourable** if $r_\pi \geq r_{\pi'}$, for all prior distributions π', where r_π denotes the Bayes risk of the Bayes rule d_π with respect to π.

Suppose that π is a prior distribution, such that

$$\int R(\theta, d_\pi)\pi(\theta)d\theta = \sup_\theta R(\theta, d_\pi).$$

Show that (i) d_π is minimax, (ii) π is least favourable.

3.3 Find the form of the Bayes rule in an estimation problem with loss function

$$L(\theta, d) = \begin{cases} a(\theta - d) & \text{if } d \leq \theta \\ b(d - \theta) & \text{if } d > \theta, \end{cases}$$

where a and b are given positive constants.

3.4 Suppose that X is distributed as a binomial random variable with index n and parameter θ. Calculate the Bayes rule (based on the single observation X) for estimating θ when the prior distribution is the uniform distribution on $[0, 1]$ and the loss function is

$$L(\theta, d) = (\theta - d)^2 / \{\theta(1 - \theta)\}.$$

Is the rule you obtain minimax?

3.5 At a critical stage in the development of a new aeroplane, a decision must be taken to continue or to abandon the project. The financial viability of the project can be measured by a parameter θ, $0 < \theta < 1$, the project being profitable if $\theta > \frac{1}{2}$. Data x provide information about θ.

If $\theta < \frac{1}{2}$, the cost to the taxpayer of continuing the project is $(\frac{1}{2} - \theta)$ [in units of \$billion], whereas if $\theta > \frac{1}{2}$ it is zero (since the project will be privatised if profitable). If $\theta > \frac{1}{2}$ the cost of abandoning the project is $(\theta - \frac{1}{2})$ (due to contractual arrangements for purchasing the aeroplane from the French), whereas if $\theta < \frac{1}{2}$ it is zero. Derive the Bayes decision rule in terms of the posterior mean of θ given x.

The Minister of Aviation has prior density $6\theta(1-\theta)$ for θ. The Prime Minister has prior density $4\theta^3$. The prototype aeroplane is subjected to trials, each independently having probability θ of success, and the data x consist of the total number of trials required for the first successful result to be obtained. For what values of x will there be serious ministerial disagreement?

3.8 Suppose $X_1, \ldots, X_n$ are independent, identically distributed random variables which, given μ, have the normal distribution $N(\mu, \sigma_0^2)$, with σ_0^2 known. Suppose also that the prior distribution of μ is normal with known mean ξ_0 and known variance ν_0.

Let X_{n+1} be a single future observation from the same distribution which is, given μ, independent of $X_1, \ldots, X_n$. Show that, given $(X_1, \ldots, X_n)$, X_{n+1} is normally distributed with mean

$$\left\{ \frac{1}{\sigma_0^2/n} + \frac{1}{\nu_0} \right\}^{-1} \left\{ \frac{\overline{X}}{\sigma_0^2/n} + \frac{\xi_0}{\nu_0} \right\}$$

and variance

$$\sigma_0^2 + \left\{ \frac{1}{\sigma_0^2/n} + \frac{1}{\nu_0} \right\}^{-1}.$$

3.9 Let $X_1, \ldots, X_n$ be independent, identically distributed $N(\mu, \sigma^2)$, with both μ and σ^2 unknown. Let $\bar{X} = n^{-1} \sum_{i=1}^n X_i$, and $s^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$.

Assume the (improper) prior $\pi(\mu, \sigma)$ with

$$\pi(\mu, \sigma) \propto \sigma^{-1}, (\mu, \sigma) \in \mathbb{R} \times (0, \infty).$$

Show that the marginal posterior distribution of $n^{1/2}(\mu - \bar{X})/s$ is the t distribution with $n-1$ degrees of freedom, and find the marginal posterior distribution of σ.

3.10 Consider a Bayes decision problem with scalar parameter θ. An estimate is required for $\phi \equiv \phi(\theta)$, with loss function

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

Find the form of the Bayes estimator of ϕ.

Let $X_1, \ldots, X_n$ be independent, identically distributed random variables from the density $\theta e^{-\theta x}, x > 0$, where θ is an unknown parameter. Let Z denote some hypothetical future value derived from the same distribution, and suppose we wish to estimate $\phi(\theta) = \Pr(Z > z)$, for given z.

Suppose we assume a gamma prior, $\pi(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}$ for θ. Find the posterior distribution for θ, and show that the Bayes estimator of ϕ is

$$\widehat{\phi}_B = \left(\frac{\beta + S_n}{\beta + S_n + z} \right)^{\alpha+n},$$

where $S_n = X_1 + \ldots X_n$.

3.11 Let the distribution of X, given θ, be normal with mean θ and variance 1. Consider estimation of θ with squared error loss $L(\theta, a) = (\theta - a)^2$ and action space $A \equiv \Omega_\theta \equiv \mathbb{R}$.

Show that the usual estimate of $\theta, d(X) = X$, is not a Bayes rule.

[Show that if $d(X)$ were Bayes with respect to a prior distribution π, we should have $r(\pi, d) = 0$.]

Show that X is extended Bayes and minimax.

5.1 Prove that random samples from the following distributions form (m, m) exponential families with either $m = 1$ or $m = 2$: Poisson, binomial, geometric, gamma (index known), gamma (index unknown). Identify the natural statistics and the natural parameters in each case. What are the distributions of the natural statistics?

The negative binomial distribution with both parameters unknown provides an example of a model that is not of exponential family form. Why?

[If Y has a gamma distribution of known index k, its density function is of the form

$$f_Y(y;\lambda) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{\Gamma(k)}.$$

The gamma distribution with index unknown has both k and λ unknown.]

5.2 Let $Y_1, \ldots, Y_n$ be independent, identically distributed $N(\mu, \mu^2)$.

Show that this model is an example of a curved exponential family.

5.3 Find the general form of a conjugate prior density for θ in a Bayesian analysis of the one-parameter exponential family density

$$f(x;\theta) = c(\theta)h(x)\exp\{\theta t(x)\},\ x \in \mathbb{R}.$$

5.4 Verify that the family of gamma distributions of known index constitutes a transformation model under the action of the group of scale transformations.

[This provides an example of a family of distributions which constitutes *both* an exponential family, *and* a transformation family. Are there any others?]

5.5 The **maximum likelihood estimator** $\widehat{\theta}(x)$ of a parameter θ maximises the **likelihood function** $L(\theta) = f(x;\theta)$ with respect to θ. Verify that maximum likelihood estimators are equivariant with respect to the group of one-to-one transformations.

5.6 Verify directly that in the location-scale model the configuration has a distribution which does not depend on the parameters.

6.1 Let $X_1, \ldots, X_n$ be independent, identically distributed $N(\mu, \mu^2)$ random variables.

Find a minimal sufficient statistic for μ and show that it is not complete.

6.2 Find a minimal sufficient statistic for θ based on an independent sample of size n from each of the following distributions:

(i) the gamma distribution with density

$$f(x;\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)},\ x > 0,$$

with $\theta = (\alpha, \beta)$;

(ii) the uniform distribution on $(\theta - 1, \theta + 1)$;

(iii) the Cauchy distribution with density

$$f(x; \alpha, b) = \frac{b}{\pi\{(x-a)^2 + b^2\}}, \ x \in \mathbb{R},$$

with $\theta = (a, b)$.

6.3 Independent factory-produced items are packed in boxes each containing k items. The probability that an item is in working order is $\theta, 0 < \theta < 1$. A sample of n boxes are chosen for testing, and X_i, the number of working items in the ith box, is noted. Thus $X_1, \ldots, X_n$ are a sample from a binomial distribution, $\text{Bin}(k, \theta)$, with index k and parameter θ. It is required to estimate the probability, θ^k, that all items in a box are in working order. Find the minimum-variance unbiased estimator, justifying your answer.

6.4 A married man who frequently talks on his mobile is well known to have conversations whose lengths are independent, identically distributed random variables, distributed as exponential with mean $1/\lambda$. His wife has long been irritated by his behaviour and knows, from infinitely many observations, the exact value of λ. In an argument with her husband, the woman produces $t_1, \ldots, t_n$, the times of n telephone conversations, to prove how excessive her husband is. He suspects that she has randomly chosen the observations, conditional on their all being longer than the expected length of conversation. Assuming he is right in his suspicion, the husband wants to use the data he has been given to infer the value of λ. What is the minimal sufficient statistic he should use? Is it complete? Find the maximum likelihood estimator for λ.

4.1 A random variable X has one of two possible densities:

$$f(x; \theta) = \theta e^{-\theta x}, \quad x \in (0, \infty), \quad \theta \in \{1, 2\}.$$

Consider the family of decision rules

$$d_\mu(x) = \begin{cases} 1 & \text{if } x \geq \mu \\ 2 & \text{if } x < \mu, \end{cases}$$

where $\mu \in [0, \infty]$. Calculate the risk function $R(\theta, d_\mu)$ for loss function $L(\theta, d) = |\theta - d|$, and sketch the parametrised curve $\mathcal{C} = \{(R(1, d_\mu), R(2, d_\mu)) : \mu \in [0, \infty]\}$ in $\mathbb{R}^2$.

Use the Neyman–Pearson Theorem to show that $\mathcal{C}$ corresponds precisely to the set of admissible decision rules.

For what prior mass function for θ does the minimax rule coincide with the Bayes rule?

4.3 Let $X_1, \ldots, X_n$ be independent random variables with a common density function

$$f(x;\theta) = \theta e^{-\theta x}, x \geq 0,$$

where $\theta \in (0, \infty)$ is an unknown parameter. Consider testing the null hypothesis $H_0 : \theta \leq 1$ against the alternative $H_1 : \theta > 1$. Show how to obtain a uniformly most powerful test of size α.

4.5 Let $X_1, \ldots, X_n$ be an independent sample of size n from the uniform distribution on $(0, \theta)$.

Show that there exists a uniformly most powerful size α test of $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$, and find its form.

Let $T = \max(X_1, \ldots, X_n)$.

Show that the test

$$\phi(x) = \begin{cases} 1, & \text{if } t > \theta_0 \text{ or } t \leq b \\ 0, & \text{if } b < t \leq \theta_0, \end{cases}$$

where $b = \theta_0 \alpha^{1/n}$, is a uniformly most powerful test of size α for testing H_0 against $H_1' : \theta \neq \theta_0$.

[Note that in a 'more regular' situation, a UMP test of H_0 against H_1' doesn't exist.]

7.1 Let $X_1, \ldots, X_n$ be an independent sample from a normal distribution with mean 0 and variance σ^2. Explain in as much detail as you can how to construct a UMPU test of $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma \neq \sigma_0$.

7.2 Let $X_1, \ldots, X_n$ be an independent sample from $N(\mu, \mu^2)$. Let $T_1 = \bar{X}$ and $T_2 = \sqrt{(1/n)\sum X_i^2}$. Show that $Z = T_1/T_2$ is ancillary. Explain why the Conditionality Principle would lead to inference about μ being drawn from the conditional distribution of T_2 given Z. Find the form of this conditional distribution.

7.4 Suppose X is normally distributed as $N(\theta, 1)$ or $N(\theta, 4)$, depending on whether the outcome, Y, of tossing a fair coin is heads ($y = 1$) or tails ($y = 0$). It is desired to test $H_0 : \theta = -1$ against $H_1 : \theta = 1$. Show that the most powerful (unconditional) size $\alpha = 0.05$ test is the test with rejection region given by $x \geq 0.598$ if $y = 1$ and $x \geq 2.392$ if $y = 0$.

Suppose instead that we condition on the outcome of the coin toss in construction of the tests. Verify that, given $y = 1$, the resulting most powerful

size $\alpha = 0.05$ test would reject if $x \geq 0.645$ while, given $y = 0$ the rejection region would be $x \geq 2.290$.

7.7 Let $X \sim \text{Bin}(m, p)$ and $Y \sim \text{Bin}(n, q)$, with X and Y independent. Show that, as p and q range over $[0, 1]$, the joint distributions of X and Y form an exponential family. Show further that if $p = q$ then

$$\Pr(X = x \mid X + Y = x + y) = \binom{m}{x}\binom{n}{y} \Big/ \binom{m+n}{x+y}.$$

Hence find the form of a UMPU test of the null hypothesis $H_0 : p \leq q$ against $H_1 : p > q$.

In an experiment to test the efficacy of a new drug for treatment of stomach ulcers, 5 patients are given the new drug and 6 patients are given a control drug. Of the patients given the new drug, 4 report an improvement in their condition, while only 1 of the patients given the control drug reports improvement. Do these data suggest, at level $\alpha = 0.1$, that patients receiving the new drug are more likely to report improvement than patients receiving the control drug?

[This is the *hypergeometric distribution* and the test presented here is conventionally referred to as *Fisher's exact test* for a 2×2 table.]