3

Significance tests

Summary. First a number of distinct situations are given in which significance tests may be relevant. The nature of a simple significance test is set out and its implications explored. The relation with interval estimation is emphasized. While most of the discussion is from a frequentist perspective, relations with Bayesian theory are outlined in the final section.

3.1 General remarks

So far, in our frequentist discussion we have summarized information about the unknown parameter ψ by finding procedures that would give in hypothetical repeated applications upper (or lower) bounds for ψ a specified proportion of times in a long run of repeated applications. This is close to but not the same as specifying a probability distribution for ψ; it avoids having to treat ψ as a random variable, and moreover as one with a *known* distribution in the absence of the data.

Suppose now there is specified a particular value ψ_0 of the parameter of interest and we wish to assess the relation of the data to that value. Often the hypothesis that $\psi = \psi_0$ is called the *null hypothesis* and conventionally denoted by H_0. It may, for example, assert that some effect is zero or takes on a value given by a theory or by previous studies, although ψ_0 does not have to be restricted in that way.

There are at least six different situations in which this may arise, namely the following.

- There may be some special reason for thinking that the null hypothesis may be exactly or approximately true or strong subject-matter interest may focus on establishing that it is likely to be false.

- There may be no special reason for thinking that the null hypothesis is true but it is important because it divides the parameter space into two (or more) regions with very different interpretations. We are then interested in whether the data establish reasonably clearly which region is correct, for example it may establish the value of $\mathrm{sgn}(\psi - \psi_0)$.
- Testing may be a technical device used in the process of generating confidence intervals.
- Consistency with $\psi = \psi_0$ may provide a reasoned justification for simplifying an otherwise rather complicated model into one that is more transparent and which, initially at least, may be a reasonable basis for interpretation.
- Only the model when $\psi = \psi_0$ is under consideration as a possible model for interpreting the data and it has been embedded in a richer family just to provide a qualitative basis for assessing departure from the model.
- Only a single model is defined, but there is a qualitative idea of the kinds of departure that are of potential subject-matter interest.

The last two formulations are appropriate in particular for examining model adequacy.

From time to time in the discussion it is useful to use the short-hand description of H_0 as being possibly *true*. Now in statistical terms H_0 refers to a probability model and the very word 'model' implies idealization. With a very few possible exceptions it would be absurd to think that a mathematical model is an exact representation of a real system and in that sense all H_0 are defined within a system which is untrue. We use the term to mean that in the current state of knowledge it is reasonable to proceed as if the hypothesis is true. Note that an underlying subject-matter hypothesis such as that a certain environmental exposure has absolutely no effect on a particular disease outcome might indeed be true.

3.2 Simple significance test

In the formulation of a simple significance test, we suppose available data y and a null hypothesis H_0 that specifies the distribution of the corresponding random variable Y. In the first place, no other probabilistic specification is involved, although some notion of the type of departure from H_0 that is of subject-matter concern is essential.

The first step in testing H_0 is to find a distribution for observed random variables that has a form which, under H_0, is free of nuisance parameters, i.e., is

completely known. This is trivial when there is a single unknown parameter whose value is precisely specified by the null hypothesis. Next find or determine a test statistic T, large (or extreme) values of which indicate a departure from the null hypothesis of subject-matter interest. Then if t_{obs} is the observed value of T we define

$$p_{\text{obs}} = P(T \geq t_{\text{obs}}), \tag{3.1}$$

the probability being evaluated under H_0, to be the (observed) *p-value* of the test.

It is conventional in many fields to report only very approximate values of p_{obs}, for example that the departure from H_0 is significant just past the 1 per cent level, etc.

The hypothetical frequency interpretation of such reported significance levels is as follows. If we were to accept the available data as just decisive evidence against H_0, then we would reject the hypothesis when true a long-run proportion p_{obs} of times.

Put more qualitatively, we examine consistency with H_0 by finding the consequences of H_0, in this case a random variable with a known distribution, and seeing whether the prediction about its observed value is reasonably well fulfilled.

We deal first with a very special case involving testing a null hypothesis that might be true to a close approximation.

Example 3.1. *Test of a Poisson mean.* Suppose that Y has a Poisson distribution of unknown mean μ and that it is required to test the null hypothesis $\mu = \mu_0$, where μ_0 is a value specified either by theory or a large amount of previous experience. Suppose also that only departures in the direction of larger values of μ are of interest. Here there is no ambiguity about the choice of test statistic; it has to be Y or a monotone function of Y and given that $Y = y$ the p-value is

$$p_{\text{obs}} = \Sigma_{v=y}^{\infty} e^{-\mu_0} \mu_0^v / v!. \tag{3.2}$$

Now suppose that instead of a single observation we have n independent replicate observations so that the model is that $Y_1, \ldots, Y_n$ have independent Poisson distributions all with mean μ. With the same null hypothesis as before, there are now many possible test statistics that might be used, for example $\max(Y_k)$. A preference for sufficient statistics leads, however, to the use of ΣY_k, which under the null hypothesis has a Poisson distribution of mean $n\mu_0$. Then the p-value is again given by (3.2), now with μ_0 replaced by $n\mu_0$ and with y replaced by the observed value of ΣY_k.

We return to this illustration in Example 3.3. The testing of a null hypothesis about the mean μ of a normal distribution when the standard deviation σ_0 is

known follows the same route. The distribution of the sample mean $\bar{Y}$ under the null hypothesis $\mu = \mu_0$ is now given by an integral rather than by a sum and (3.2) is replaced by

$$p_{\text{obs}} = 1 - \Phi\left(\frac{\bar{y} - \mu_0}{\sigma_0/\sqrt{n}}\right). \tag{3.3}$$

We now turn to a complementary use of these ideas, namely to test the adequacy of a given model, what is also sometimes called model criticism. We illustrate this by testing the adequacy of the Poisson model. It is necessary if we are to parallel the previous argument to find a statistic whose distribution is exactly or very nearly independent of the unknown parameter μ. An important way of doing this is by appeal to the second property of sufficient statistics, namely that after conditioning on their observed value the remaining data have a fixed distribution.

Example 3.2. *Adequacy of Poisson model.* Let $Y_1, \ldots, Y_n$ be independent Poisson variables with unknown mean μ. The null hypothesis H_0 for testing model adequacy is that this model applies for some unknown μ. Initially no alternative is explicitly formulated. The sufficient statistic is ΣY_k, so that to assess consistency with the model we examine the conditional distribution of the data given $\Sigma Y_k = s$. This density is zero if $\Sigma y_k \neq s$ and is otherwise

$$\frac{s!}{\Pi y_k!} \frac{1}{n^s}, \tag{3.4}$$

i.e., is a multinomial distribution with s trials each giving a response equally likely to fall in one of n cells. Because this distribution is completely specified numerically, we are essentially in the same situation as if testing consistency with a null hypothesis that completely specified the distribution of the observations free of unknown parameters. There remains, except when $n = 2$, the need to choose a test statistic. This is usually taken to be either the *dispersion index* $\Sigma(Y_k - \bar{Y})^2/\bar{Y}$ or the number of zeros.

The former is equivalent to the ratio of the sample estimate of variance to the mean. In this conditional specification, because the sample total is fixed, the statistic is equivalent also to ΣY_k^2. Note that if, for example, the dispersion test is used, no explicit family of alternative models has been specified, only an indication of the kind of discrepancy that it is especially important to detect. A more formal and fully parametric procedure might have considered the negative binomial distribution as representing such departures and then used the apparatus of the Neyman–Pearson theory of testing hypotheses to develop a test especially sensitive to such departures.

A quite high proportion of the more elementary tests used in applications were developed by the relatively informal route just outlined. When a full family of

distributions is specified, covering both the null hypothesis and one or more alternatives representing important departures from H_0, it is natural to base the test on the optimal statistic for inference about ψ within that family. This typically has sensitivity properties in making the random variable P corresponding to p_{obs} stochastically small under alternative hypotheses.

For continuously distributed test statistics, p_{obs} typically may take any real value in $(0, 1)$. In the discrete case, however, only a discrete set of values of p are achievable in any particular case. Because preassigned values such as 0.05 play no special role, the only difficulty in interpretation is the theoretical one of comparing alternative procedures with different sets of achievable values.

Example 3.3. *More on the Poisson distribution.* For continuous observations the random variable P can, as already noted, in principle take any value in $(0, 1)$. Suppose, however, we return to the special case that Y has a Poisson distribution with mean μ and that the null hypothesis $\mu = \mu_0$ is to be tested checking for departures in which $\mu > \mu_0$, or more generally in which the observed random variable Y is stochastically larger than a Poisson-distributed random variable of mean μ_0. Then for a given observation y the p-value is

$$p^+_{\text{obs}} = \Sigma_{v=y}^{\infty} e^{-\mu_0}\mu_0^v/v!, \tag{3.5}$$

whereas for detecting departures in the direction of small values the corresponding p-value is

$$p^-_{\text{obs}} = \Sigma_{v=0}^{y} e^{-\mu_0}\mu_0^v/v!. \tag{3.6}$$

Table 3.1 shows some values for the special case $\mu_0 = 2$. So far as use for a one-sided significance test is concerned, the restriction to a particular set of values is unimportant unless that set is in some sense embarrassingly small. Thus the conclusion from Table 3.1(b) that in testing $\mu = 2$ looking for departures $\mu < 2$ even the most extreme observation possible, namely zero, does not have a particularly small p-value is hardly surprising. We return to the implications for two-sided testing in the next section.

In forming upper confidence limits for μ based on an observed y there is no difficulty in finding critical values of μ such that the relevant lower tail area is some assigned value. Thus with $y = 0$ the upper 0.95 point for μ is such that $e^{-\mu^*} = 0.05$, i.e., $\mu^* = \log 20 \simeq 3$. A similar calculation for a lower confidence limit is not possible for $y = 0$, but is possible for all other y. The discreteness of the set of achievable p-values is in this context largely unimportant.

Table 3.1. *Achievable significance levels for testing that a Poisson-distributed random variable with observed value y has mean 2: (a) test against alternatives larger than 2; (b) test against alternatives less than 2*

(a)	y	2	3	4	5	6
	p	0.594	0.323	0.143	0.053	0.017
(b)	y	0	1	2		
	p	0.135	0.406	0.677		

3.3 One- and two-sided tests

In many situations observed values of the test statistic in either tail of its distribution represent interpretable, although typically different, departures from H_0. The simplest procedure is then often to contemplate two tests, one for each tail, in effect taking the more significant, i.e., the smaller tail, as the basis for possible interpretation. Operational interpretation of the result as a hypothetical error rate is achieved by doubling the corresponding p, with a slightly more complicated argument in the discrete case.

More explicitly we argue as follows. With test statistic T, consider two p-values, namely

$$p^+_{\text{obs}} = P(T \geq t; H_0), \quad p^-_{\text{obs}} = P(T \leq t; H_0). \tag{3.7}$$

In general the sum of these values is $1 + P(T = t)$. In the two-sided case it is then reasonable to define a new test statistic

$$Q = \min(P^+_{\text{obs}}, P^-_{\text{obs}}). \tag{3.8}$$

The level of significance is

$$P(Q \leq q_{\text{obs}}; H_0). \tag{3.9}$$

In the continuous case this is $2q_{\text{obs}}$ because two disjoint events are involved. In a discrete problem it is q_{obs} plus the achievable p-value from the other tail of the distribution nearest to but not exceeding q_{obs}. As has been stressed the precise calculation of levels of significance is rarely if ever critical, so that the careful definition is more one of principle than of pressing applied importance. A more important point is that the definition is unaffected by a monotone transformation of T.

In one sense very many applications of tests are essentially two-sided in that, even though initial interest may be in departures in one direction, it will

rarely be wise to disregard totally departures in the other direction, even if initially they are unexpected. The interpretation of differences in the two directions may well be very different. Thus in the broad class of procedures associated with the linear model of Example 1.4 tests are sometimes based on the ratio of an estimated variance, expected to be large if real systematic effects are present, to an estimate essentially of error. A large ratio indicates the presence of systematic effects whereas a suspiciously small ratio suggests an inadequately specified model structure.

3.4 Relation with acceptance and rejection

There is a conceptual difference, but essentially no mathematical difference, between the discussion here and the treatment of testing as a two-decision problem, with control over the formal error probabilities. In this we fix in principle the probability of rejecting H_0 when it is true, usually denoted by α, aiming to maximize the probability of rejecting H_0 when false. This approach demands the explicit formulation of alternative possibilities. Essentially it amounts to setting in advance a threshold for p_{obs}. It is, of course, potentially appropriate when clear decisions are to be made, as for example in some classification problems. The previous discussion seems to match more closely scientific practice in these matters, at least for those situations where analysis and interpretation rather than decision-making are the focus.

That is, there is a distinction between the Neyman–Pearson formulation of testing regarded as clarifying the meaning of statistical significance via hypothetical repetitions and that same theory regarded as in effect an instruction on how to implement the ideas by choosing a suitable α in advance and reaching different decisions accordingly. The interpretation to be attached to accepting or rejecting a hypothesis is strongly context-dependent; the point at stake here, however, is more a question of the distinction between assessing evidence, as contrasted with deciding by a formal rule which of two directions to take.

3.5 Formulation of alternatives and test statistics

As set out above, the simplest version of a significance test involves formulation of a null hypothesis H_0 and a test statistic T, large, or possibly extreme, values of which point against H_0. Choice of T is crucial in specifying the kinds of departure from H_0 of concern. In this first formulation no alternative probability models are explicitly formulated; an implicit family of possibilities is specified via T. In fact many quite widely used statistical tests were developed in this way.

A second possibility is that the null hypothesis corresponds to a particular parameter value, say $\psi = \psi_0$, in a family of models and the departures of main interest correspond either, in the one-dimensional case, to one-sided alternatives $\psi > \psi_0$ or, more generally, to alternatives $\psi \neq \psi_0$. This formulation will suggest the most sensitive test statistic, essentially equivalent to the best estimate of ψ, and in the Neyman–Pearson formulation such an explicit formulation of alternatives is essential.

The approaches are, however, not quite as disparate as they may seem. Let $f_0(y)$ denote the density of the observations under H_0. Then we may associate with a proposed test statistic T the exponential family

$$f_0(y)\exp\{t\theta - k(\theta)\}, \tag{3.10}$$

where $k(\theta)$ is a normalizing constant. Then the test of $\theta = 0$ most sensitive to these departures is based on T. Not all useful tests appear natural when viewed in this way, however; see, for instance, Example 3.5.

Many of the test procedures for examining model adequacy that are provided in standard software are best regarded as defined directly by the test statistic used rather than by a family of alternatives. In principle, as emphasized above, the null hypothesis is the conditional distribution of the data given the sufficient statistic for the parameters in the model. Then, within that null distribution, interesting directions of departure are identified.

The important distinction is between situations in which a whole family of distributions arises naturally as a base for analysis versus those where analysis is at a stage where only the null hypothesis is of explicit interest.

Tests where the null hypotheses itself is formulated in terms of arbitrary distributions, so-called *nonparametric* or *distribution-free* tests, illustrate the use of test statistics that are formulated largely or wholly informally, without specific probabilistically formulated alternatives in mind. To illustrate the arguments involved, consider initially a single homogenous set of observations.

That is, let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables with arbitrary cumulative distribution function $F(y) = P(Y_k \leq y)$. To avoid minor complications we suppose throughout the following discussion that the distribution is continuous so that, in particular, the possibility of ties, i.e., exactly equal pairs of observations, can be ignored. A formal likelihood can be obtained by dividing the real line into a very large number of very small intervals each having an arbitrary probability attached to it. The likelihood for

data y can then be specified in a number of ways, namely by

- a list of data values actually observed,
- the *sample cumulative distribution function*, defined as

$$F_n(y) = n^{-1}\Sigma I(y_k \leq y), \tag{3.11}$$

  where the indicator function $I(y_k \leq y)$ is one if $y_k \leq y$ and zero otherwise,
- the set of order statistics $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$, i.e., the observed values arranged in increasing order.

The second and third of these are reductions of the first, suppressing the information about the order in which the observations are obtained. In general no further reduction is possible and so either of the last two forms the sufficient statistic. Thus if we apply the general prescription, conclusions about the function $F(y)$ are to be based on one of the above, for example on the sample distribution function, whereas consistency with the model is examined via the conditional distribution given the order statistics. This conditional distribution specifies that starting from the original data, all $n!$ permutations of the data are equally likely. This leads to tests in general termed *permutation tests*. This idea is now applied to slightly more complicated situations.

Example 3.4. *Test of symmetry.* Suppose that the null hypothesis is that the distribution is symmetric about a known point, which we may take to be zero. That is, under the null hypothesis $F(-y) + F(y) = 1$. Under this hypothesis, all points y and $-y$ have equal probability, so that the sufficient statistic is determined by the order statistics or sample distribution function of the $|y_k|$. Further, conditionally on the sufficient statistic, all 2^n sample points $\pm y_k$ have equal probability 2^{-n}. Thus the distribution under the null hypothesis of any test statistic is in principle exactly known.

Simple one-sided test statistics for symmetry can be based on the number of positive observations, leading to the *sign test*, whose null distribution is binomial with parameter $\frac{1}{2}$ and index n, or on the mean of all observations. The distribution of the latter can be found by enumeration or approximated by finding its first few moments and using a normal or other approximation to the distribution.

This formulation is relevant, for example, when the observations analyzed are differences between primary observations after and before some intervention, or are differences obtained from the same individual under two different regimes.

A strength of such procedures is that they involve no specification of the functional form of $F(y)$. They do, however, involve strong independence assumptions, which may often be more critical. Moreover, they do not extend easily to relatively complicated models.

Example 3.5. *Nonparametric two-sample test.* Let $(Y_{11}, \ldots, Y_{1n_1}; Y_{21}, \ldots, Y_{2n_2})$ be two sets of mutually independent random variables with cumulative distribution functions respectively $F_1(y)$ and $F_2(y)$. Consider the null hypothesis $F_1(y) = F_2(y)$ for all y.

When this hypothesis is true the sufficient statistic is the set of order statistics of the combined set of observations and under this hypothesis all $(n_1 + n_2)!$ permutations of the data are equally likely and, in particular, the first set of n_1 observations is in effect a random sample drawn without replacement from the full set, allowing the null distribution of any test statistic to be found.

Sometimes it may be considered that while the ordering of possible observational points is meaningful the labelling by numerical values is not. Then we look for procedures invariant under arbitrary strictly monotone increasing transformations of the measurement scale. This is achieved by replacing the individual observations y by their rank order in the full set of order statistics of the combined sample. If the test statistic, T_W, is the sum of the ranks of, say, the first sample, the resulting test is called the *Wilcoxon rank sum test* and the parallel test for the single-sample symmetry problem is the *Wilcoxon signed rank test*.

The distribution of the test statistic under the null hypothesis, and hence the level of significance, is in principle found by enumeration. The moments of T_W under the null hypothesis can be found by the arguments to be developed in a rather different context in Section 9.2 and in fact the mean and variance are respectively $n_1(n_1 + n_2 + 1)/2$ and $n_1 n_2(n_1 + n_2 + 1)/12$. A normal approximation, with continuity correction, based on this mean and variance will often be adequate.

Throughout this discussion the full set of values of y is regarded as fixed.

The choice of test statistic in these arguments is based on informal considerations or broad analogy. Sometimes, however, the choice can be sharpened by requiring good sensitivity of the test were the data produced by some unknown monotonic transformation of data following a specific parametric model.

For the two-sample problem, the most obvious possibilities are that the data are transformed from underlying normal or underlying exponential distributions, a test of equality of the relevant means being required in each case. The exponential model is potentially relevant for the analysis of survival data. Up to a scale and location change in the normal case and up to a scale change in the exponential case, the originating data can then be reconstructed approximately under the null hypothesis by replacing the rth largest order statistic out of n in the data by the expected value of that order statistic in samples

from the standard normal or unit exponential distribution respectively. Then the standard parametric test statistics are used, relying in principle on their permutation distribution to preserve the nonparametric propriety of the test.

It can be shown that, purely as a test procedure, the loss of sensitivity of the resulting nonparametric analysis as compared with the fully parametric analysis, were it available, is usually small. In the normal case the expected order statistics, called Fisher and Yates scores, are tabulated, or can be approximated by $\Phi^{-1}\{(r-3/8)/(n+1/4)\}$. For the exponential distribution the scores can be given in explicit form or approximated by $\log\{(n+1)/(n+1-r-1/2)\}$.

3.6 Relation with interval estimation

While conceptually it may seem simplest to regard estimation with uncertainty as a simpler and more direct mode of analysis than significance testing there are some important advantages, especially in dealing with relatively complicated problems, in arguing in the other direction. Essentially confidence intervals, or more generally confidence sets, can be produced by testing consistency with every possible value in Ω_ψ and taking all those values not 'rejected' at level c, say, to produce a $1-c$ level interval or region. This procedure has the property that in repeated applications any true value of ψ will be included in the region except in a proportion $1-c$ of cases. This can be done at various levels c, using the same form of test throughout.

Example 3.6. *Ratio of normal means.* Given two independent sets of random variables from normal distributions of unknown means μ_0, μ_1 and with known variance σ_0^2, we first reduce by sufficiency to the sample means $\bar{y}_0, \bar{y}_1$. Suppose that the parameter of interest is $\psi = \mu_1/\mu_0$. Consider the null hypothesis $\psi = \psi_0$. Then we look for a statistic with a distribution under the null hypothesis that does not depend on the nuisance parameter. Such a statistic is

$$\frac{\bar{Y}_1 - \psi_0 \bar{Y}_0}{\sigma_0\sqrt{(1/n_1 + \psi_0^2/n_0)}}; \tag{3.12}$$

this has a standard normal distribution under the null hypothesis. This with ψ_0 replaced by ψ could be treated as a pivot provided that we can treat $\bar{Y}_0$ as positive.

Note that provided the two distributions have the same variance a similar result with the Student t distribution replacing the standard normal would apply if the variance were unknown and had to be estimated. To treat the probably more realistic situation where the two distributions have different and unknown variances requires the approximate techniques of Chapter 6.

We now form a $1-c$ level confidence region by taking all those values of ψ_0 that would not be 'rejected' at level c in this test. That is, we take the set

$$\left\{\psi : \frac{(\bar{Y}_1 - \psi\bar{Y}_0)^2}{\sigma_0^2(1/n_1 + \psi^2/n_0)} \le k^*_{1;\,c}\right\}, \tag{3.13}$$

where $k^*_{1;\,c}$ is the upper c point of the chi-squared distribution with one degree of freedom.

Thus we find the limits for ψ as the roots of a quadratic equation. If there are no real roots, *all* values of ψ are consistent with the data at the level in question. If the numerator and especially the denominator are poorly determined, a confidence interval consisting of the whole line may be the only rational conclusion to be drawn and is entirely reasonable from a testing point of view, even though regarded from a confidence interval perspective it may, wrongly, seem like a vacuous statement.

Depending on the context, emphasis may lie on the possible explanations of the data that are reasonably consistent with the data or on those possible explanations that have been reasonably firmly refuted.

Example 3.7. *Poisson-distributed signal with additive noise.* Suppose that Y has a Poisson distribution with mean $\mu + a$, where $a > 0$ is a known constant representing a background process of noise whose rate of occurrence has been estimated with high precision in a separate study. The parameter μ corresponds to a signal of interest. Now if, for example, $y = 0$ and a is appreciable, for example, $a \ge 4$, when we test consistency with each possible value of μ *all* values of the parameter are inconsistent with the data until we use very small values of c. For example, the 95 per cent confidence interval will be empty. Now in terms of the initial formulation of confidence intervals, in which, in particular, the model is taken as a firm basis for analysis, this amounts to making a statement that is certainly wrong; there is by supposition some value of the parameter that generated the data. On the other hand, regarded as a statement of which values of μ are consistent with the data at such-and-such a level the statement is perfectly reasonable and indeed is arguably the only sensible frequentist conclusion possible at that level of c.

3.7 Interpretation of significance tests

There is a large and ever-increasing literature on the use and misuse of significance tests. This centres on such points as:

1. Often the null hypothesis is almost certainly false, inviting the question why is it worth testing it?

2. Estimation of ψ is usually more enlightening than testing hypotheses about ψ.
3. Failure to 'reject' H_0 does not mean that we necessarily consider H_0 to be exactly or even nearly true.
4. If tests show that data are consistent with H_0 and inconsistent with the minimal departures from H_0 considered as of subject-matter importance, then this may be taken as positive support for H_0, i.e., as more than mere consistency with H_0.
5. With large amounts of data small departures from H_0 of no subject-matter importance may be highly significant.
6. When there are several or many somewhat similar sets of data from different sources bearing on the same issue, separate significance tests for each data source on its own are usually best avoided. They address individual questions in isolation and this is often inappropriate.
7. p_{obs} is not the probability that H_0 is true.

Discussion of these points would take us too far afield. Point 7 addresses a clear misconception. The other points are largely concerned with how in applications such tests are most fruitfully applied and with the close connection between tests and interval estimation. The latter theme is emphasized below. The essential point is that significance tests in the first place address the question of whether the data are reasonably consistent with a null hypothesis in the respect tested. This is in many contexts an interesting but limited question. The much fuller specification needed to develop confidence limits by this route leads to much more informative summaries of what the data plus model assumptions imply.

3.8 Bayesian testing

A Bayesian discussion of significance testing is available only when a full family of models is available. We work with the posterior distribution of ψ. When the null hypothesis is quite possibly exactly or nearly correct we specify a prior probability π_0 that H_0 is true; we need also to specify the conditional prior distribution of ψ when H_0 is false, as well as aspects of the prior distribution concerning nuisance parameters λ. Some care is needed here because the issue of testing is not likely to arise when massive easily detected differences are present. Thus when, say, ψ can be estimated with a standard error of $\sigma_0/\sqrt{n}$ the conditional prior should have standard deviation $b\sigma_0/\sqrt{n}$, for some not too large value of b.

When the role of H_0 is to divide the parameter space into qualitatively different parts the discussion essentially is equivalent to checking whether the posterior interval at some suitable level overlaps the null hypothesis value of ψ. If and only if there is no overlap the region containing ψ is reasonably firmly established. In simple situations, such as that of Example 1.1, posterior and confidence intervals are in exact or approximate agreement when flat priors are used, providing in such problems some formal justification for the use of flat priors or, from a different perspective, for confidence intervals.

We defer to Section 5.12 the general principles that apply to the choice of prior distributions, in particular as they affect both types of testing problems mentioned here.

Notes 3

Section 3.1. The explicit classification of types of null hypothesis is developed from Cox (1977).

Section 3.2. The use of the conditional distribution to test conformity with a Poisson distribution follows Fisher (1950). Another route for dealing with discrete distributions is to define (Stone, 1969) the p-value for test statistic T by $P(T > t_{\text{obs}}) + P(T = t_{\text{obs}})/2$. This produces a statistic having more nearly a uniform distribution under the null hypothesis but the motivating operational meaning has been sacrificed.

Section 3.3. There are a number of ways of defining two-sided p-values for discrete distributions; see, for example, Cox and Hinkley (1974, p.79).

Section 3.4. The contrast made here between the calculation of p-values as measures of evidence of consistency and the more decision-focused emphasis on accepting and rejecting hypotheses might be taken as one characteristic difference between the Fisherian and the Neyman–Pearson formulations of statistical theory. While this is in some respects the case, the actual practice in specific applications as between Fisher and Neyman was almost the reverse. Neyman often in effect reported p-values whereas some of Fisher's use of tests in applications was much more dichotomous. For a discussion of the notion

of severity of tests, and the circumstances when consistency with H_0 might be taken as positive support for H_0, see Mayo (1996).

Section 3.5. For a thorough account of nonparametric tests, see Lehmann (1998).

Section 3.6. The argument for the ratio of normal means is due to E. C. Fieller, after whom it is commonly named. The result applies immediately to the ratio of least squares regression coefficients and hence in particular to estimating the intercept of a regression line on the z-coordinate axis. A substantial dispute developed over how this problem should be handled from the point of view of Fisher's fiducial theory, which mathematically but not conceptually amounts to putting flat priors on the two means (Creasy, 1954). There are important extensions of the situation, for example to inverse regression or (controlled) calibration, where on the basis of a fitted regression equation it is desired to estimate the value of an explanatory variable that generated a new value of the response variable.

Section 3.8. For more on Bayesian tests, see Jeffreys (1961) and also Section 6.2.6 and Notes 6.2.

4

More complicated situations

Summary. This chapter continues the comparative discussion of frequentist and Bayesian arguments by examining rather more complicated situations. In particular several versions of the two-by-two contingency table are compared and further developments indicated. More complicated Bayesian problems are discussed.

4.1 General remarks

The previous frequentist discussion in especially Chapter 3 yields a theoretical approach which is limited in two senses. It is restricted to problems with no nuisance parameters or ones in which elimination of nuisance parameters is straightforward. An important step in generalizing the discussion is to extend the notion of a Fisherian reduction. Then we turn to a more systematic discussion of the role of nuisance parameters.

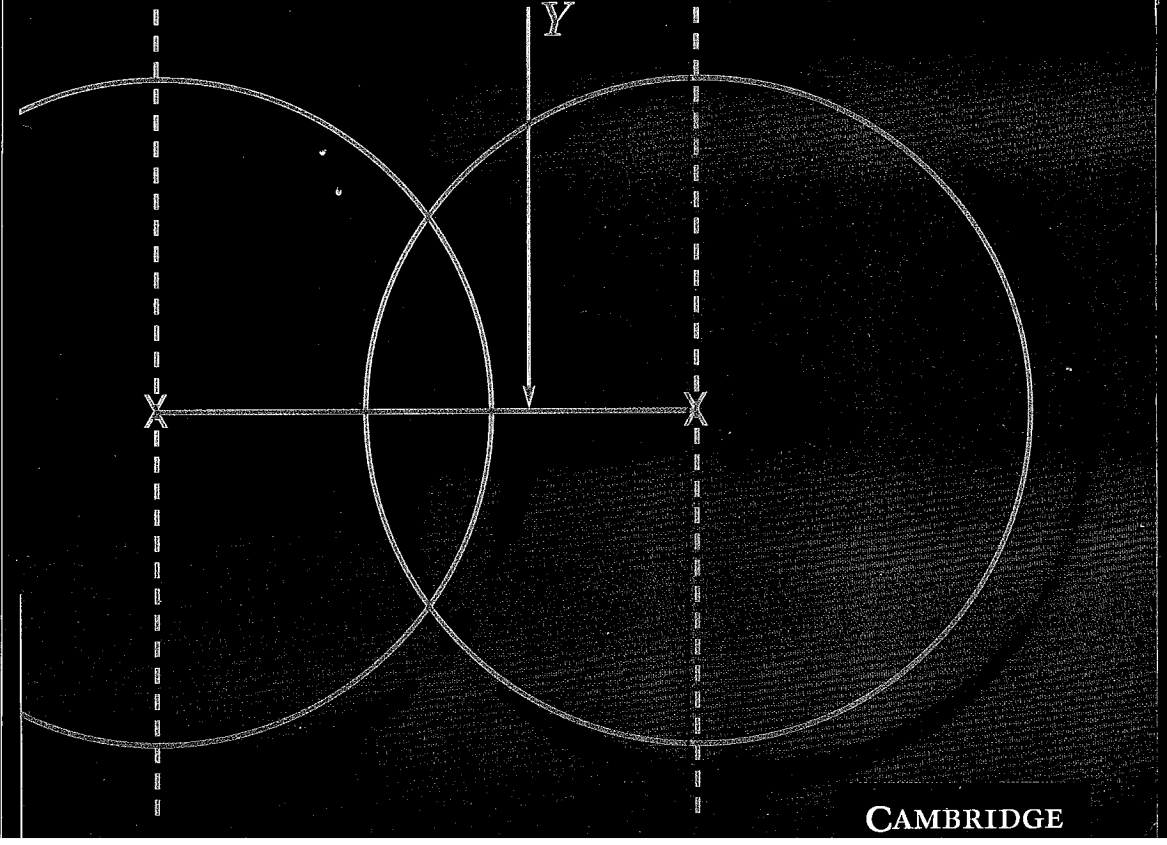
By comparison, as noted previously in Section 1.5, a great formal advantage of the Bayesian formulation is that, once the formulation is accepted, all subsequent problems are computational and the simplifications consequent on sufficiency serve only to ease calculations.

4.2 General Bayesian formulation

The argument outlined in Section 1.5 for inference about the mean of a normal distribution can be generalized as follows. Consider the model $f_{Y|\Theta}(y \mid \theta)$, where, because we are going to treat the unknown parameter as a random variable, we now regard the model for the data-generating process as a conditional density. Suppose that Θ has the prior density $f_\Theta(\theta)$, specifying the marginal

D.R. COX

Principles of Statistical Inference



CAMBRIDGE