4 Tests of Hypotheses

4.1 Basic definitions

Consider again a vector $\mathbf{X}$ of observations from the family $\mathcal{F}$ of probability distributions,

$$\mathcal{F} = \{f(\mathbf{x};\theta),\ \mathbf{x} \in \mathcal{X} : \theta \in \Theta\}.$$

A hypothesis is an assertion about the true value of θ. We shall usually consider two competing hypotheses, a *null hypothesis* H_0 and an *alternative hypothesis* H_1. They will be of the following general form,

$$H_0 : \theta \in \Theta_0$$

and

$$H_1 : \theta \in \Theta_1,$$

where $\Theta_0 \subset \Theta,\ \Theta_1 \subset \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

The null hypothesis H_0 might correspond to the prediction of some scientific theory or represent some simple set of circumstances which, in the absence of evidence to the contrary, we wish to assume holds — it is the default option. The alternative hypothesis H_1 represents the family of possible departures from the null hypothesis that we wish to envisage.

If a hypothesis corresponds to a single value of the vector of parameters, so that the p.d.f. of $\mathbf{X}$ is completely specified, then the hypothesis is said to be *simple.* Otherwise it is said to be *composite.*

It is quite common to have a simple null hypothesis and a composite alternative. For example, if θ is a single parameter, so that $\Theta \subseteq \mathbb{R}$, we may wish to test the null hypothesis

$$H_0 : \theta = \theta_0$$

against the two-sided alternative hypothesis

$$H_1 : \theta \neq \theta_0,$$

for some particular value θ_0. Thus $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \bar{\Theta}_0 \equiv \Theta \setminus \Theta_0$. As another possibility, we might use the one-sided alternative hypothesis

$$H_1 : \theta > \theta_0,$$

in which case $\Theta_1 = \{\theta : \theta > \theta_0\}$.

But commoner still in practice are cases where both hypotheses are composite, usually with $\Theta_1 = \bar{\Theta}_0$. For example, in simple sampling from a normal distribution with $\theta = (\mu, \sigma^2)$, we have $\Theta = \mathbb{R} \times (0, \infty)$. If we test

$$H_0 : \mu = \mu_0$$

against

$$H_1 : \mu \neq \mu_0$$

then σ^2 is what is known as a *nuisance parameter*. We have

$$\Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0\}$$

and

$$\Theta_1 = \{(\mu, \sigma^2) : \mu \neq \mu_0\}.$$

In the general case, given the hypotheses H_0 and H_1, we base a test of H_0 against H_1 upon the sample data $\mathbf{X}$. The test procedure is specified by a *critical region* (or *rejection region*) $\mathcal{C}$, $\mathcal{C} \subset \mathcal{X}$, such that H_0 is rejected if and only if $\mathbf{X} \in \mathcal{C}$. If $\mathbf{X} \notin \mathcal{C}$ then H_0 is not rejected — there is not enough evidence in the data to reject H_0.

In a decision theoretic approach to hypothesis testing, where a definite decision has to be made between H_0 and H_1, if $\mathbf{X} \in \mathcal{C}$ then H_1 is accepted and H_0 is rejected, but if $\mathbf{X} \notin \mathcal{C}$ then H_0 is "accepted" and H_1 is "rejected". We shall tend to adopt a different approach, where the emphasis is upon weighing up the evidence against the null hypothesis H_0, upon the falsification of H_0. The alternative hypothesis H_1 has a subsidiary role. We avoid talking in terms of acceptance of H_0 or of acceptance or rejection of H_1.

- To some extent, the analogy with an English court of law may be helpful. The null hypothesis H_0 is the presumption of innocence of the accused; the alternative hypothesis H_1 is that the accused is guilty. There is an asymmetry between the role of the two hypotheses. The verdict is "guilty" if there is evidence "beyond reasonable doubt" that the accused is guilty. Otherwise, the verdict is "not guilty".

  In the law court, the verdict of "guilty" is rejection of H_0 and amounts to acceptance of H_1. The verdict of "not guilty" means that there is not enough evidence to reject H_0, although in effect it amounts to acceptance of H_0. So, in effect, in the setting of the law court, a decision is being made, although the rules for coming to a decision are weighted in favour of "not guilty". There is a greater concern to guard against finding the innocent guilty than to guard against acquitting the guilty.

There will generally be a large number of possible tests available for testing a given H_0 against a given H_1 — in the case of continuous distributions, infinitely many — since each $\mathcal{C} \subset \mathcal{X}$ may be used to specify the critical region of some test. The problem will be to find a test, or a family of tests, which is in some sense optimal.

There is never any perfect test procedure available, guaranteed always to come up with the right conclusion. There are always two types of error associated with any test, the error of rejecting H_0 when it is true (a Type I error) and the error of not rejecting H_0 when H_1 is true (a Type II error).

The performance of a test is described by its *power function* $\beta(\theta)$, where

$$\begin{aligned} \beta(\theta) &= \mathbb{P}(\mathbf{X} \in \mathcal{C}; \theta) \\ &= \mathbb{P}(H_0 \text{ is rejected}; \theta) \qquad \theta \in \Theta. \end{aligned}$$

Clearly,

$$0 \leq \beta(\theta) \leq 1 \qquad \theta \in \Theta.$$

The “power” here is the power to reject H_0. For $\theta \in \Theta_0$, $\beta(\theta)$ is an error probability, the probability of rejecting H_0 when it is true. For $\theta \in \Theta_1$, it is $1 - \beta(\theta)$ that is the error probability, the probability of not rejecting H_0 when H_0 is false.

In general terms, because we want to have small error probabilities, it is desirable that $\beta(\theta)$ should be small when H_0 is true, i.e, when $\theta \in \Theta_0$, but that $\beta(\theta)$ should be large when H_1 is true, i.e, when $\theta \in \Theta_1$. The unattainable ideal is that $\beta(\theta) = 0$ for $\theta \in \Theta_0$ and $\beta(\theta) = 1$ for $\theta \in \Theta_1$.

Definition

The *significance level* α of a test is defined by

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta).$$

Thus the significance level of a test is the supremum of the Type I error probabilities.

4.2 Simple hypotheses

We now turn to the case where the null and the alternative hypotheses are both simple, with

$$H_0 : \theta = \theta_0$$

and

$$H_1 : \theta = \theta_1,$$

for some particular values θ_0 and θ_1. Thus $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$.

This case is unrealistically simple in that neither of two simple hypotheses is likely to be exactly correct. However, it is theoretically straightforward to analyze and provides a basis for the discussion of more complex situations.

For ease of notation, we write the p.d.f.s $f(\mathbf{x}; \theta_0)$ and $f(\mathbf{x}; \theta_1)$ as $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$, respectively.

We may take the parameter space to be $\Theta = \{\theta_0, \theta_1\}$. It seems reasonable that a test procedure should be based upon the value of the *likelihood ratio*, $f_1(\mathbf{x})/f_0(\mathbf{x})$. The larger the observed value of this statistic, the greater the relative likelihood of H_1 rather than H_0 being true. Thus a plausible test procedure is one with critical region of the form

$$\mathcal{C} = \{\mathbf{x} \in \mathcal{X} : \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} \geq k\} \tag{1}$$

for some positive constant k. This is known as a *likelihood ratio test.*

In the case of simple hypotheses, the significance level and power of a test with critical region $\mathcal{C}$ may be defined particularly simply. The *significance level* α is given by

$$\alpha = \mathbb{P}(\mathbf{X} \in \mathcal{C}; \theta_0) = \mathbb{P}(H_0 \text{ is rejected}; H_0)$$

and the *power* β by

$$\beta = \mathbb{P}(\mathbf{X} \in \mathcal{C}; \theta_1) = \mathbb{P}(H_0 \text{ is rejected}; H_1).$$

Usually we choose k so as to get a specified value of α. The classical Neyman-Pearson lemma demonstrates the optimality of likelihood ratio tests according to a natural criterion: among all tests of a given significance level, or smaller, the likelihood ratio test has the greatest power.

Theorem 1 (The Neyman-Pearson Lemma) *Given the simple hypotheses H_0 and H_1, let $\mathcal{C}$ be the critical region as specified in Equation (1) of a likelihood ratio test for some given $k > 0$. Let α be the significance level and β the power of this test. Any other test with significance level less than or equal to α has power less than or equal to β.*

Proof. We give the proof for the case of continuous distributions. The proof for the case of discrete distributions is identical, except that integrals are replaced by summations.

Let $\mathcal{C}^*$ be the critical region of any other test with significance level less than or equal to α. Thus

$$\int_{\mathcal{C}} f_0(\mathbf{x})d\mathbf{x} = \alpha \tag{2}$$

and

$$\int_{\mathcal{C}^*} f_0(\mathbf{x})d\mathbf{x} \leq \alpha. \tag{3}$$

We require to prove that

$$\int_{\mathcal{C}} f_1(\mathbf{x})d\mathbf{x} \geq \int_{\mathcal{C}^*} f_1(\mathbf{x})d\mathbf{x}. \tag{4}$$

Note that

$$\mathcal{C} = (\mathcal{C} \setminus \mathcal{C}^*) \cup (\mathcal{C} \cap \mathcal{C}^*)$$

and

$$\mathcal{C}^* = (\mathcal{C}^* \setminus \mathcal{C}) \cup (\mathcal{C} \cap \mathcal{C}^*),$$

both the above unions being disjoint.

It follows from Equations (2) and (3) that

$$\int_{\mathcal{C}} f_0(\mathbf{x})d\mathbf{x} \geq \int_{\mathcal{C}^*} f_0(\mathbf{x})d\mathbf{x}$$

and hence, subtracting $\int_{\mathcal{C}\cap\mathcal{C}^*} f_0(\mathbf{x})d\mathbf{x}$ from both sides, that

$$\int_{\mathcal{C}\setminus\mathcal{C}^*} f_0(\mathbf{x})d\mathbf{x} \geq \int_{\mathcal{C}^*\setminus\mathcal{C}} f_0(\mathbf{x})d\mathbf{x}. \tag{5}$$

But from Equation (1),

$$f_1(\mathbf{x}) \geq kf_0(\mathbf{x}) \qquad \mathbf{x} \in \mathcal{C}$$

and

$$f_1(\mathbf{x}) < kf_0(\mathbf{x}) \qquad \mathbf{x} \in \bar{\mathcal{C}}.$$

Hence from Equation (5)

$$\int_{\mathcal{C}\setminus\mathcal{C}^*} f_1(\mathbf{x})d\mathbf{x} \geq \int_{\mathcal{C}^*\setminus\mathcal{C}} f_1(\mathbf{x})d\mathbf{x}.$$

Adding $\int_{\mathcal{C}\cap\mathcal{C}^*} f_1(\mathbf{x})d\mathbf{x}$ to both sides, we obtain Equation (4).

4.3 Composite hypotheses

Returning to the consideration of composite hypotheses, a common strategy in searching for a "best" test is to attempt to find one that has maximum power $\beta(\theta)$ for all $\theta \in \Theta_1$, subject to having a significance level less than or equal to some specified value α. A test is said to be a *uniformly most powerful (UMP)* test of significance level α if it has significance level less than or equal to α and $\beta(\theta) \geq \beta^*(\theta)$ for all $\theta \in \Theta_1$ for the power function $\beta^*(\theta)$ of every other test with significance level less than or equal to α.

Only in certain special situations, usually involving one-sided alternative hypotheses, is it possible to find a UMP test. Much mathematical work has been done on investigating this and rather less stringent criteria for finding tests that are in some sense optimal.

Example 1

Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli r.v.s with parameter θ, $0 < \theta < 1$ and consider first the testing of the simple hypotheses,

$$H_0 : \theta = \theta_0$$

against

$$H_1 : \theta = \theta_1,$$

where $\theta_1 > \theta_0$. In this case,

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = \frac{\theta_1^t(1-\theta_1)^{n-t}}{\theta_0^t(1-\theta_0)^{n-t}},$$

where $t \equiv T(\mathbf{x}) = \sum x_i$. The form of a likelihood ratio test is thus to reject H_0 if and only if

$$\frac{\theta_1^t(1-\theta_1)^{n-t}}{\theta_0^t(1-\theta_0)^{n-t}} \geq k^*,$$

for some constant k^*. Because the LHS of the above inequality is monotonic increasing in t, the critical region $\mathcal{C}$ of the test reduces to the form

$$\mathcal{C} = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) \geq k\}, \tag{6}$$

for some constant k.

As $T(\mathbf{x})$ can only take the integer values $0, 1, \ldots, n$, the constant k in the critical region (6) may without loss of generality be restricted to the same values $0, 1, \ldots, n$. Thus there are only $n+1$ distinct critical regions, one for each of the above values of k, and hence also only $n+1$ possible values of the significance level of the test. Since $T(\mathbf{X})$ has the binomial $\mathrm{Bin}(n, \theta)$ distribution, the significance level of the test for any particular value of k is given by

$$\begin{aligned} \alpha &= \mathbb{P}(T(\mathbf{X}) \geq k; \theta_0) \\ &= \sum_{t=k}^{n} \binom{n}{t} \theta_0^t (1-\theta_0)^{n-t}. \end{aligned} \tag{7}$$

The form of the likelihood ratio test is the same for all pairs of parameter values θ_0 and θ_1 with $\theta_1 > \theta_0$. It follows that a test with critical region of the form (6) is UMP for testing

$$H_0 : \theta = \theta_0 \qquad \text{against} \qquad H_1 : \theta > \theta_0,$$

or, indeed, for testing

$$H_0 : \theta \leq \theta_0 \qquad \text{against} \qquad H_1 : \theta > \theta_0.$$

The corresponding power function $\beta(\theta) \equiv \mathbb{P}(T(\mathbf{X}) \geq k; \theta)$ is given by

$$\beta(\theta) = \sum_{t=k}^{n} \binom{n}{t} \theta^t (1-\theta)^{n-t} \qquad \text{for } 0 \leq \theta \leq 1. \tag{8}$$

It may be checked by differentiating Equation (8) that $\beta(\theta)$ is an increasing function of θ. The significance level α is equal to $\beta(\theta_0)$, as given by Equation (7).

Suppose in particular that we have reason to believe that somebody has tampered with a coin so that the probability θ of obtaining heads has been increased from its "fair" value of $\frac{1}{2}$. We therefore wish to test

$$H_0 : \theta = \frac{1}{2} \qquad \text{against} \qquad H_1 : \theta > \frac{1}{2}.$$

We toss the coin 10 times and carry out the test procedure of rejecting H_0 if and only if the number of heads observed is either 9 or 10. This corresponds to taking $n = 10$ and using the critical region (6) with $k = 9$. (We take n to be so small only for the sake of illustration, to keep the calculations simple.) The expression (8) reduces to

$$\beta(\theta) = \theta^9(10 - 9\theta) \qquad \text{for } 0 \leq \theta \leq 1.$$

In particular, the significance level $\alpha = \beta(\frac{1}{2})$ is given by

$$\alpha = 11 \left(\frac{1}{2}\right)^{10} = 0.0107,$$

almost exactly equal to a conventional 1% significance level.

The test is not very powerful unless $\theta \approx 1$. Even for $\theta = 0.9$, $\beta(\theta) = 0.736$. A better test procedure, one that had a significance level less than or equal to the above α and a higher power function for values of $\theta > \frac{1}{2}$, could only be obtained by increasing the value of the sample size n.

Note that there is no UMP test for testing $H_0 : \theta = \theta_0$ against the two-sided alternative, $H_1 : \theta \neq \theta_0$.

4.4 Test statistics

Many commonly used tests have the following structure.

1. There exists a test statistic T such that the critical region is of the form

   $$\mathcal{C} = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) \geq k\},$$

   for some constant k.

2. The distribution of T is (at least approximately) the same for all $\theta \in \Theta_0$, i.e., under H_0. Hence $\mathbb{P}(T(\mathbf{X}) \geq k; \theta)$ takes the same value for all $\theta \in \Theta_0$, and we can write

   $$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}(\mathbf{X} \in \mathcal{C}; \theta)$$

   as

   $$\alpha = \mathbb{P}(T(\mathbf{X}) \geq k; H_0).$$

3. The distribution of T under H_0 is known, at least approximately.

4. The statistic $T(\mathbf{X})$ has a continuous distribution, so that, given any $\alpha > 0$, we can find a constant k_α such that

   $$\mathbb{P}(T(\mathbf{X}) \geq k_\alpha; H_0) = \alpha.$$

   To be specific, $k_\alpha = F^{-1}(1 - \alpha)$, where $F(\cdot)$ is the distribution function of T under H_0. (The values of k_α can in many cases be found from within a statistical package or from tables.) Let $\mathcal{C}_\alpha$ denote the corresponding critical region,

   $$\mathcal{C}_\alpha = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) \geq k_\alpha\}.$$

   We then have that $k_\alpha \uparrow$ and $\mathcal{C}_\alpha \downarrow$ as $\alpha \downarrow 0$.

5. In such situations, if we observe a value t of the test statistic T then we may refer to

   $$p \equiv \mathbb{P}(T \geq t; H_0) \equiv 1 - F(t)$$

   as the *p-value* (or *significance level*) of the observed value t. This is the probability under H_0 of obtaining a value of the test statistic T that is greater than or equal to t.

   The p-value is the value of α such that $t = k_\alpha$, i.e., the smallest value of α such that H_0 is rejected at the α significance level.

The p-value may be regarded as a measure of the weight of evidence against H_0. The larger the observed value of T and the smaller the corresponding p-value, the greater the evidence against H_0. If the p-value is small then **either** H_0 is true and an extreme outcome of the experiment has occurred **or** H_0 is false.

- If T has a discrete rather than a continuous distribution, we may still calculate p-values, but the p-values are restricted to a discrete set of values rather than having a continuous range.

4.5 Generalized likelihood ratio tests

There is a general method of constructing tests of composite hypotheses, which is a generalization of the particular likelihood ratio test whose optimality for testing simple hypotheses was proved in the Neyman-Pearson lemma. We assume that we are to test

$$H_0 : \theta \in \Theta_0$$

against

$$H_1 : \theta \in \bar{\Theta}_0.$$

A natural way of generalizing the earlier definition of the likelihood ratio is to adopt the form

$$\frac{\sup_{\theta \in \bar{\Theta}_0} L(\theta; \mathbf{x})}{\sup_{\theta \in \Theta_0} L(\theta; \mathbf{x})},$$

in which, in the numerator, the likelihood is maximized over all parameter values consistent with the alternative hypothesis and in the denominator over all parameter values consistent with the null hypothesis. However, it turns out to be more convenient to write the likelihood ratio in the following form.

Define the *generalized likelihood ratio* $\lambda(\mathbf{x})$ by

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta} L(\theta; \mathbf{x})}{\sup_{\theta \in \Theta_0} L(\theta; \mathbf{x})}, \tag{9}$$

in which, in the numerator, the likelihood is maximized over all possible parameter values and in the denominator over all parameter values consistent with the null hypothesis.

With this definition of the likelihood ratio, it is necessarily the case that $\lambda(\mathbf{x}) \geq 1$. A *generalized likelihood ratio test* (or *maximum likelihood ratio test*) is a test that has a critical region $\mathcal{C}$ of the form

$$\mathcal{C} = \{\mathbf{x} \in \mathcal{X} : \lambda(\mathbf{x}) \geq k\} \tag{10}$$

for some constant $k \geq 1$, where $\lambda(\mathbf{x})$ is as defined in Equation (9).

A generalized likelihood ratio test usually provides us with a sensible test procedure, but it is not necessarily optimal in any sense, except in the special case of testing simple hypotheses.

Usually the vector of parameters, $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$, is such that the parameter space Θ is a q-dimensional space, a subset of $\mathbb{R}^q$, and the null hypothesis corresponds to r, say, of the parameters being set to zero, or, more generally, to there being r restrictions on the parameters, so that Θ_0 is a $q-r$ dimensional subset of Θ. In such a situation, under some regularity conditions, it may be shown that, for all $\theta \in \Theta_0$, as the sample size $n \to \infty$, the distribution of the statistic

$$2 \ln \lambda(\mathbf{X})$$

converges to the χ^2_r distribution.

We may use $2 \ln \lambda(\mathbf{X})$ instead of $\lambda(\mathbf{X})$ as the test statistic for the generalised likelihood ratio test. The critical region of Equation (10) may equivalently be written in the form

$$\mathcal{C} = \{\mathbf{x} \in \mathcal{X} : 2 \ln \lambda(\mathbf{x}) \geq c\} \tag{11}$$

for some constant $c \geq 0$. Use of the large sample approximation that under H_0 the test statistic $2 \ln \lambda(\mathbf{X})$ has the χ^2_r distribution enables us to make the identification that, for a test of significance level α, the constant c has to be the $100\alpha\%$ point of the χ^2_r distribution, i.e. $\chi^2_{r,\alpha}$.

4.6 Confidence sets

We may consider hypothesis testing and estimation as two complementary approaches to classical statistical inference. In this section we do no more than state the way in which estimation is extended to the construction of confidence sets.

When we construct a point estimator $\hat{\theta}(\mathbf{X})$ of a parameter θ, as discussed in Chapters 2 and 3, the probability that the estimate $\hat{\theta}(\mathbf{x})$ gives exactly the true value of θ is generally zero. If we state only the value of an estimate, without any further comment, this does not indicate how accurate we may expect the estimate to be. We may construct a subset of the parameter space, a *confidence set*, which we can be reasonably confident contains the true parameter value. In the special case of a single, real-valued parameter θ, the subset of the parameter space will usually be an interval of the real line, a *confidence interval.* Correspondingly, we sometimes refer to the task of constructing such a set or interval as the problem of *set estimation* or *interval estimation*, respectively, as opposed to point estimation as dealt with previously.

Consider again a vector $\mathbf{X}$ of observations from the family $\mathcal{F}$ of probability distributions,

$$\mathcal{F} = \{f(\mathbf{x};\theta),\ \mathbf{x} \in \mathcal{X} : \theta \in \Theta\}.$$

Definition

Given any α with $0 < \alpha < 1$, let $\{\mathcal{S}(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ be a family of subsets of the parameter space Θ having the property that

$$\mathbb{P}(\theta \in \mathcal{S}(\mathbf{X});\theta) = 1 - \alpha \tag{12}$$

for all $\theta \in \Theta$. Given any observed value $\mathbf{x}$ of $\mathbf{X}$, the set $\mathcal{S}(\mathbf{x})$ is said to be a $1-\alpha$ (or a $100(1-\alpha)\%$) *confidence set* for θ.

The intuitive interpretation of a confidence set $\mathcal{S}(\mathbf{x})$ is that we are $100(1-\alpha)\%$ "confident" that $\theta \in \mathcal{S}(\mathbf{x})$.

The subset $\mathcal{S}(\mathbf{X})$ is a random subset of the parameter space Θ. It varies from repetition to repetition of the statistical experiment of observing a value of $\mathbf{X}$. The vector of parameters θ is fixed, although unknown. The probability statement of Equation (12) is thus **not** a probability statement about θ but an expression for the probability that the random subset $\mathcal{S}(\mathbf{X})$ contains the true value θ. A correct interpretation is that, whatever the true value of θ, in a long sequence of repetitions of the statistical experiment, the confidence set contains this true value in a proportion $1-\alpha$ of cases.

- If we wish to examine critically this frequentist approach to the construction of confidence sets, we might consider how relevant is the long-run behaviour of the sets $\mathcal{S}(\mathbf{X})$ to the specific instance of wishing to make a confidence statement about θ on the basis of the value of $\mathbf{x}$ observed on a particular occasion.