6 Hypothesis testing

Statistics is the only profession which demands the right to make mistakes 5 per cent of the time – Thomas Huxley.

6.1 The Neyman–Pearson framework

The second major area of statistical inference is **hypothesis testing**. A statistical **hypothesis** is an assertion or conjecture about the distribution of one or more random variables, and a test of a statistical hypothesis is a rule or procedure for deciding whether to reject that assertion.

Example 6.1 It has been suggested that dying people may be able to postpone their death until after an important occasion. In a study of 1919 people with Jewish surnames it was found that 922 occurred in the week before Passover and 997 in the week after. Is there any evidence in this data to reject the hypothesis that a person is as likely to die in the week before as in the week after Passover?

Example 6.2 In one of his experiments, Mendel crossed 556 smooth, yellow male peas with wrinkled, green female peas. Here is what he obtained and its comparison with predictions based on genetic theory.

type	observed count	predicted frequency	expected count
smooth yellow	315	9/16	312.75
smooth green	108	3/16	104.25
wrinkled yellow	102	3/16	104.25
wrinkled green	31	1/16	34.75

Is there any evidence in this data to reject the hypothesis that theory is correct?

We follow here an approach developed by Neyman and Pearson. Suppose we have data $x = (x_1, x_2, \dots, x_n)$ from a density f. We have two hypotheses about f. On the basis of the data one is **accepted**, the other **rejected**. The two hypotheses have different philosophical status. The first, called the **null hypothesis**, and denoted by H_0, is a conservative hypothesis, not to be rejected unless evidence is clear. The second, the **alternative hypothesis**, denoted by H_1, specifies the kind of departure from the null hypothesis of interest to us.

It is often assumed that f belongs to a specified parametric family $f(\cdot \mid \theta)$ indexed by a parameter $\theta \in \Theta$ (e.g. $N(\theta, 1)$, $B(n, \theta)$). We might then want to test a parametric hypothesis

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

with $\Theta_0 \cap \Theta_1 = \emptyset$. We may, or may not, have $\Theta_0 \cup \Theta_1 = \Theta$.

We will usually be concerned with testing a parametric hypothesis of this kind, but alternatively, we may wish to test

$$H_0 : f = f_0 \quad \text{against} \quad H_1 : f \neq f_0$$

where f_0 is a specified density. This is a '**goodness-of-fit**' test.

A third alternative is that we wish to test

$$H_0 : f = f_0 \quad \text{against} \quad H_1 : f = f_1$$

where f_0 and f_1 are specified, but do not necessarily belong to the same family.

6.2 Terminology

A hypothesis which specifies f completely is called **simple**, e.g., $\theta = \theta_0$. Otherwise, a hypothesis is **composite**, e.g., $\theta > \theta_0$.

Suppose we wish to test H_0 against H_1. A test is defined by a **critical region** C. We write $\bar{C}$ for the complement of C.

$$\text{If } x = (x_1, x_2, \dots, x_n) \in \begin{cases} C \text{ then } H_0 \text{ is rejected, and} \\ \bar{C} \text{ then } H_0 \text{ is accepted (not rejected).} \end{cases}$$

Note that when $x \in \bar{C}$ we might sometimes prefer to say 'not rejected', rather than 'accepted'. This is a minor point which need not worry us, except to note that sometimes 'not rejected' does more accurately express what we are doing: i.e., looking to see if the data provides any evidence to reject the null hypothesis. If it does not, then we might want to consider other things before finally 'accepting H_0'.

There are two possible types of error we might make:

H_0 might be rejected when it is true (**a type I error**), or

H_0 might be accepted when it is false (**a type II error**).

Since H_0 is conservative, a type I error is generally considered to be 'more serious' that a type II error. For example, the jury in a murder trial should take as its null hypothesis that the accused is innocent, since the type I error (that an innocent person is convicted and the true murderer is never caught) is more serious than the type II error (that a murderer is acquitted).

Hence, we fix (an upper bound on) the probability of type I error, e.g., 0.05 or 0.01, and define the critical region C by minimizing the type II error subject to this.

If H_0 is simple, $\Theta_0 = \{\theta_0\}$, the probability of a type I error is called the **size**, or **significance level**, of the test. If H_0 is composite, the size of the test is $\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}(X \in C \mid \theta)$. .

The **likelihood** of a simple hypothesis $H : \theta = \theta^*$ given data x is

$$L_x(H) = f_X(x \mid \theta = \theta^*).$$

If H is composite, $H : \theta \in \Theta$, we define

$$L_x(H) = \sup_{\theta \in \Theta} f_X(x \mid \theta).$$

The **likelihood ratio** for two hypotheses H_0, H_1 is

$$L_x(H_0, H_1) = \frac{L_x(H_1)}{L_x(H_0)}.$$

Notice that if $T(x)$ is a sufficient statistic for θ then by the factorization criterion $L_x(H_0, H_1)$ is simply a function of $T(x)$.

6.3 Likelihood ratio tests

A test given by a critical region C of the form $C = \{x : L_x(H_0, H_1) > k\}$, for some constant k, is called a **likelihood ratio test**. The value of k is determined by fixing the size α of the test, so that $\mathbb{P}(X \in C \mid H_0) = \alpha$.

Likelihood ratio tests are optimal for simple hypotheses. Most standard tests are likelihood ratio tests, though tests can be built from others statistics.

Lemma 6.3 (Neyman–Pearson Lemma) $H_0 : f = f_0$ *is to be tested against* $H_1 : f = f_1$. *Assume that* f_0 *and* f_1 *are* > 0 *on the same regions and continuous.*

Then, among all tests of size $\leq \alpha$*, the test with smallest probability of type II error is given by* $C = \{x : f_1(x)/f_0(x) > k\}$*, where* k *is determined by*

$$\alpha = \mathbb{P}(X \in C \mid H_0) = \int_C f_0(x)\, dx.$$

Proof. Consider any test with size $\leq \alpha$, i.e., with a critical region D such that $\mathbb{P}(X \in D \mid H_0) \leq \alpha$. Define

$$\phi_D(x) = \begin{cases} 1 \\ 0 \end{cases} \text{as } x \begin{matrix} \in \\ \notin \end{matrix} D$$

and let C and k be defined as above. Note that

$$0 \leq \big(\phi_C(x) - \phi_D(x)\big)\big(f_1(x) - kf_0(x)\big), \text{ for all } x.$$

since this is always the product of two terms with the same sign. Hence

$$\begin{aligned}
0 &\leq \int_x \big(\phi_C(x) - \phi_D(x)\big)\big(f_1(x) - kf_0(x)\big)\, dx \\
&= \mathbb{P}(X \in C \mid H_1) - \mathbb{P}(X \in D \mid H_1) - k\big[\mathbb{P}(X \in C \mid H_0) - \mathbb{P}(X \in D \mid H_0)\big] \\
&= \mathbb{P}(X \in C \mid H_1) - \mathbb{P}(X \in D \mid H_1) - k\big[\alpha - \mathbb{P}(X \in D \mid H_0)\big] \\
&\leq \mathbb{P}(X \in C \mid H_1) - \mathbb{P}(X \in D \mid H_1)
\end{aligned}$$

This implies $\mathbb{P}(X \notin C \mid H_1) \leq \mathbb{P}(X \notin D \mid H_1)$ as required. ∎

6.4 Single sample: testing a given mean, simple alternative, known variance (z-test)

Let $X_1, \ldots, X_n$ be IID $N(\mu, \sigma^2)$, where σ^2 is known. We wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$, where $\mu_1 > \mu_0$.

The Neyman–Pearson test is to reject H_0 if the likelihood ratio is large (i.e., greater than some k). The likelihood ratio is

$$\begin{aligned}
\frac{f(x \mid \mu_1, \sigma^2)}{f(x \mid \mu_0, \sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2} \exp\left[-\sum_{i=1}^n (x_i - \mu_1)^2/2\sigma^2\right]}{(2\pi\sigma^2)^{-n/2} \exp\left[-\sum_{i=1}^n (x_i - \mu_0)^2/2\sigma^2\right]} \\
&= \exp\left[\sum_{i=1}^n \left\{(x_i - \mu_0)^2 - (x_i - \mu_1)^2\right\}/2\sigma^2\right] \\
&= \exp\left[n\left\{2\bar{x}(\mu_1 - \mu_0) + (\mu_0^2 - \mu_1^2)\right\}/2\sigma^2\right]
\end{aligned}$$

It often turns out, as here, that the likelihood ratio is a monotone function of a sufficient statistic and we can immediately rephrase the critical region in more convenient terms. We notice that the likelihood ratio above is increasing in the sufficient statistic $\bar{x}$ (since $\mu_1 - \mu_0 > 0$). So the Neyman–Pearson test is equivalent to 'reject H_0 if $\bar{x} > c$', where we choose c so that $\mathbb{P}(\bar{X} > c \mid H_0) = \alpha$. There is no need to try to write c in terms of k.

However, under H_0 the distribution of $\bar{X}$ is $N(\mu_0, \sigma^2/n)$. This means that

$$Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma \sim N(0, 1)\ .$$

It is now convenient to rephrase the test in terms of Z, so that a test of size α is to reject H_0 if $z > z_\alpha$, where $z_\alpha = \Phi^{-1}(1 - \alpha)$ is the 'upper α point of $N(0, 1)$' i.e., the point such that $\mathbb{P}(N(0, 1) > z_\alpha) = \alpha$. E.g., for $\alpha = 0.05$ we would take $z_\alpha = 1.645$, since 5% of the standard normal distribution lies to the right of 1.645.

Because we reject H_0 only if z is in the upper tail of the normal distribution we call this a **one-tailed test**. We shall see other tests in which H_0 is rejected if the test statistic lies in either of two tails. Such a test is called a **two-tailed test**.

Example 6.4 *Suppose* $X_1, \ldots, X_n$ *are IID* $N(\mu, \sigma^2)$ *as above, and we want a test of size* 0.05 *of* $H_0 : \mu = 5$ *against* $H_1 : \mu = 6$*, with* $\sigma^2 = 1$*. Suppose the data is* $x = (5.1, 5.5, 4.9, 5.3)$. Then $\bar{x} = 5.2$ and $z = 2(5.2 - 5)/1 = 0.4$. Since this is less than 1.645 we do not reject $\mu = 5$.

Suppose the hypotheses are reversed, so that we test $H_0 : \mu = 6$ against $H_1 : \mu = 5$. The test statistic is now $z = 2(5.2 - 6)/1 = -1.6$ and we should reject H_0 for values of Z less than -1.645. Since z is more than -1.645, it is not significant and we do not reject $\mu = 6$.

This example demonstrates the preferential position given to H_0 and therefore that it is important to choose H_0 in a way that makes sense in the context of the decision problem with which the statistical analysis is concerned.