7 Further aspects of hypothesis testing

Statisticians do it with only a 5% chance of being rejected.

7.1 The p-value of an observation

The **significance level of a test** is another name for α, the size of the test. For a composite hypothesis $H_0 : \theta \in \Theta_0$ and rejection region C this is

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}(X \in C \mid \theta)\,.$$

For a likelihood ratio test the **p-value** of an observation x is defined to be

$$p^* = \sup_{\theta \in \Theta_0} P_\theta\big(L_X(H_0, H_1) \geq L_x(H_0, H_1)\big)\,.$$

The p-value of x is the probability under H_0 of seeing x or something at least as 'extreme', in the sense of containing at least as mush evidence against H_0. E.g., in Example 6.4, the p-value is p^* where

$$p^* = \mathbb{P}(Z > z \mid \mu = \mu_0) = 1 - \Phi(z)\,.$$

A test of size α rejects H_0 if and only if $\alpha \geq p^*$. Thus *the p-value of x is the smallest value of α for which H_0 would be rejected on the basis of seeing x.* It is often more informative to report the p-value than merely to report whether or not the null hypothesis has been rejected.

The p-value is also sometimes called the **significance level** of x. Historically, the term arises from the practice of 'significance testing'. Here, we begin with an H_0 and a test statistic T, which need not be the likelihood ratio statistic, for which, say, large positive values suggest falsity of H_0. We observe value t_0 for the statistic: the significance level is $\mathbb{P}(T \geq t_0 \mid H_0)$. If this probability is small, H_0 is rejected.

7.2 The power of a test

For a parametric hypothesis about $\theta \in \Theta$, we define the **power function** of the test specified by the critical region C as

$$W(\theta) = \mathbb{P}(X \in C \mid \theta).$$

Notice that $\alpha = \sup_{\theta \in \Theta_0} W(\theta)$, and $1 - W(\theta) = \mathbb{P}(X \in \bar{C} \mid \theta) = \mathbb{P}(\text{type II error} \mid \theta)$ for $\theta \in \Theta_1$.

7.3 Uniformly most powerful tests

For the test of $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$, $\mu_1 > \mu_0$ described in section 6.4 the critical region turned out to be

$$C(\mu_0) = \left\{x : \sqrt{n}(\bar{x} - \mu_0)/\sigma > \alpha_z\right\}.$$

This depends on μ_0 but not on the specific value of μ_1. The test with this critical region would be optimal for any alternative $H_1 : \mu = \mu_1$, provided $\mu_1 > \mu_0$. This is the idea of a uniformly most powerful (UMP) test.

We can find the power function of this test.

$$\begin{aligned} W(\mu) &= \mathbb{P}(Z > \alpha_z \mid \mu) \\ &= \mathbb{P}\big(\sqrt{n}(\bar{X} - \mu_0)/\sigma > \alpha_z \mid \mu\big) \\ &= \mathbb{P}\big(\sqrt{n}(\bar{X} - \mu)/\sigma + \sqrt{n}(\mu - \mu_0)/\sigma > \alpha_z \mid \mu\big) \\ &= 1 - \Phi\big(\alpha_z - \sqrt{n}(\mu - \mu_0)/\sigma\big) \end{aligned}$$

Note that $W(\mu)$ increases from 0 to 1 as μ goes from $-\infty$ to ∞ and $W(\mu_0) = \alpha$.

More generally, suppose $H_0 : \theta \in \Theta_0$ is to be tested against $H_1 : \theta \in \Theta_1$, where $\Theta_0 \cap \Theta_1 = \emptyset$. Suppose H_1 is composite. H_0 can be simple or composite. We want a test of size α. So we require $W(\theta) \leq \alpha$ for all $\theta \in \Theta_0$, $W(\theta_0) = \alpha$ for some $\theta_0 \in \Theta_0$.

A **uniformly most powerful** (UMP) test of size α satisfies (i) it is of size α, (ii) $W(\theta)$ is as large as possible for every $\theta \in \Theta_1$.

UMP tests may not exist. However, likelihood ratio tests are often UMP.

Example 7.1 *Let $X_1, \ldots, X_n$ be IID $N(\mu, \sigma^2)$, where μ is known. Suppose $H_0 : \sigma^2 \leq 1$ is to be tested against $H_1 : \sigma^2 > 1$.*

We begin by finding the most powerful test for testing $H_0' : \sigma^2 = \sigma_0^2$ against $H_1' : \sigma^2 = \sigma_1^2$, where $\sigma_0^2 \leq 1 < \sigma_1^2$. The Neyman–Pearson test rejects H_0' for large values of the likelihood ratio:

$$\begin{aligned} \frac{f(x \mid \mu, \sigma_1^2)}{f(x \mid \mu, \sigma_0^2)} &= \frac{(2\pi\sigma_1^2)^{-n/2} \exp\left[-\sum_{i=1}^n (x_i - \mu)^2 / 2\sigma_1^2\right]}{(2\pi\sigma_0^2)^{-n/2} \exp\left[-\sum_{i=1}^n (x_i - \mu)^2 / 2\sigma_0^2\right]} \\ &= (\sigma_0/\sigma_1)^n \exp\left[\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right) \sum_{i=1}^n (x_i - \mu)^2\right] \end{aligned}$$

which is large when $\sum_i (x_i - \mu)^2$ is large. If $\sigma^2 = 1$ then

$$\sum_{i=1}^n (x_i - \mu)^2 \sim \chi_n^2\,.$$

So a test of the form 'reject H_0 if $T := \sum_i (x_i - \mu)^2 > F_\alpha^{(n)}$', has size α where $F_\alpha^{(n)}$ is the upper α point of χ_n^2. That is, $\mathbb{P}\left(T > F_\alpha^{(n)} \mid \sigma^2 \leq 1\right) \leq \alpha$, for all $\sigma^2 \leq 1$, with equality for $\sigma^2 = 1$. But this test doesn't depend on the value of σ_1^2, and hence is the UMP test of H_0 against H_1.

Example 7.2 Consider Example 6.1. Let p be the probability that if a death occurs in one of the two weeks either side of Passover it actually occurs in the week after the Passover. Let us test $H_0 : p = 0.5$ vs. $H_1 : p > 0.5$.

The distribution of the number of deaths in the week after Passover, say X, is $B(n, p)$, which we can approximate by $N\big(np, np(1-p)\big)$; under H_0 this is $N(0.5n, 0.25n)$. So a size 0.05 test is to reject H_0 if $z = \sqrt{n}\big(\bar{x} - \mu_0\big)/\sigma > 1.645$, where here $z = \sqrt{1919}\big(997/1919 - 0.5\big)/0.5 = 1.712$. So the data is just significant at the 5% level. We reject the hypothesis that death $p = 1/2$.

It is important to realise that this does not say anything about *why* this might be. It might be because people really are able to postpone their deaths to enjoy the holiday. But it might also be that deaths increase after Passover because of over-eating or stress during the holiday.

7.4 Confidence intervals and hypothesis tests

There is an interesting duality between confidence intervals and hypothesis tests. In the following, we speak of the **acceptance region**, i.e., the complement of the critical (or rejection) region C.

Theorem 7.3

(i) Suppose that for every θ_0 there is a size α test of $H_0 : \theta = \theta_0$ against some alternative. Denote the acceptance region by $A(\theta_0)$. Then $I(X) = \{\theta : X \in A(\theta)\}$ is a $100(1-\alpha)\%$ confidence interval for θ.

(ii) Conversely, if $I(X)$ is a $100(1-\alpha)\%$ confidence interval for θ then an acceptance region for a size α test of $H_0 : \theta = \theta_0$ is $A(\theta_0) = \{X : \theta_0 \in I(X)\}$.

Proof. The definitions in the theorem statement give

$$\mathbb{P}\big(X \in A(\theta_0) \mid \theta = \theta_0\big) = \mathbb{P}\big(\theta \in I(X) \mid \theta = \theta_0\big).$$

By assumption the l.h.s. is $1-\alpha$ in case (i) and the r.h.s. is $1-\alpha$ in case (ii). ∎

This duality can be useful. In some circumstances it can be easier to see what is the form of a hypothesis test and then work out a confidence interval. In other circumstances it may be easier to see the form of the confidence interval.

In Example 4.1 we saw that a 95% confidence interval for μ based upon $X_1, \dots, X_n$ being IID samples from $N(\mu, \sigma^2)$, σ^2 known, is

$$\left[\bar{X} - \frac{1.96\sigma}{\sqrt{n}}\,,\ \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right].$$

Thus $H_0 : \mu = \mu_0$ is rejected in a 5% level test against $H_1 : \mu \neq \mu_0$ if and only if μ_0 is not in this interval; i.e., if and only if

$$\sqrt{n}\big|\bar{X} - \mu_0\big|/\sigma > 1.96\,.$$

7.5 The Bayesian perspective on hypothesis testing

Suppose a statistician wishes to test the hypothesis that a coin is fair ($H_0 : p = 1/2$) against the alternative that it biased towards heads with probability p, ($H_1 : p = p_1 > 1/2$). He decides to toss it five times. It is easy to see that the best test is to reject H_0 if the total number of heads, say T, is large. Here $T \sim B(5, p)$. Suppose he observes `H,H,H,H,T`, so $T = 4$. The p-value (or significance level) of this result is the probability that under the null hypothesis he should see a result which is equally or more extreme, i.e.,

$$\mathbb{P}(T = 4 \text{ or } 5) = 5(0.5^4)(0.5) + 0.5^5 = 6/32 = 0.1875\,.$$

Another statistician wishes also wishes to test H_0 against H_1, but with a different experiment. He plans to toss the coin until he gets a tail. Now the best test is to reject H_0 if the number of tosses, say N, is large. Here $N \sim \text{geometric}(1-p)$. Suppose this statistician also observes `H,H,H,H,T`, so $N = 5$. He figures the p-value as the probability of seeing this result or one that is even more extreme and obtains

$$\mathbb{P}(N \geq 5) = 0.5^4 = 1/16 = 0.0625\,.$$

Thus the two statisticians come to different conclusions about the significance of what they have seen. This is disturbing! The coin knew nothing about the experimental procedure. Maybe there were four tosses simply because that was the point at which the experimenter spilt his coffee and so decided to stop tossing the coin. What then is the significance level?

The Bayesian perspective on hypothesis testing avoids this type of problem. Suppose the experimenter places prior probabilities of $\mathbb{P}(H_0)$ and $\mathbb{P}(H_1)$ on the truth of two mutually exclusive hypotheses. Having observed the data x, these are modified into posterior probabilities in the usual way, i.e.,

$$\frac{\mathbb{P}(H_1 \mid x)}{\mathbb{P}(H_0 \mid x)} = \frac{\mathbb{P}(x \mid H_1)}{\mathbb{P}(x \mid H_0)}\frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} = L_x(H_0, H_1)\frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)}\,.$$

Thus the ratio of the posterior probabilities is just the ratio of the prior probabilities multiplied by the likelihood ratio.

Under both procedures above $x = \{\texttt{H,H,H,H,T}\}$, and the likelihood ratios is

$$L_x(H_0, H_1) = \frac{p_1}{0.5} \times \frac{p_1}{0.5} \times \frac{p_1}{0.5} \times \frac{p_1}{0.5} \times \frac{(1-p_1)}{0.5} = \frac{p_1^4(1-p_1)}{0.5^5}\,.$$

In general, the Bayesian analysis does not depend on how the data was obtained.