Inferential Statistics **First semester, 2021**

Lecture 6: Hypothesis testing

Instructor: Erlis Ruli (ruli@stat.unipd.it), Department of Statistical Sciences

Suppose that the population of WM's has a finite average energy consumption μ_1. A manufacturer of motors for WM's claims that his next generation motors (called NGM1) have average energy consumption μ_2, with $\mu_2 < \mu_1$. There are two possibilities, either the population average energy consumptions in the two groups are **equal** or they are **different**. We call the first possibility the **Null Hypothesis** (H_0, for short) and the second the **Alternative Hypothesis** (H_1).

To asses which of the hypothesis is true we may take a group of WM's with the old motor and a group of WM's with NGM1 and measure their energy consumptions. Suppose that the average energy consumptions with the old and NGM1 are 0.9 and 0.89, respectively. Since $0.9 \neq 0.89$, we might be tempted to conclude that H_0 must be rejected, but this is wrong. In order to judge if the discrepancy $|0.9 - 0.89|$ is far enough from 0 we must assess how much surprising is this discrepancy in a repeated sampling context. We can measure this degree of suprisingness by computing a confidence interval for the difference of two population means (as in L5, Example 5.5) and that would perfectly solve our problem. However, there is an alternative: statistical hypothesis testing.

Statistical hypothesis testing aims to measure the support or credibility of an hypothesis (typically a null hypothesis) in light of the sample.

Similar problems in scientific activity are: is a new vaccine effective? Does a lot of manufactured items contain an excessive number of defectives? Is the mean lifetime of a WM at least 2 years? Ordinarily, information about such phenomena can be obtained only by performing experiments whose outcomes have some bearing on the hypotheses of interest.

Hypothesis testing thus is about assessing hypothesis on the values of the parameter θ of a statistical model. Formally, suppose that we partition the parameter space Θ into two

disjoint sets Θ_0 and Θ_1 and that we wish to know if

$$H_0 : \theta \in \Theta_0 \quad \text{or} \quad H_1 : \theta \in \Theta_1, \tag{6.1}$$

is true. Let X be r.v. with range $\mathcal{X}$. Hypothesis testing if performed by finding an appropriate subset of outcomes $R \subset \mathcal{X}$, called the *rejection region* or *critical region* and applying the following simple rule:

$$\text{if } X \in R \Longrightarrow \text{reject } H_0, \text{ otherwise accept } H_0.$$

In many practical cases, the rejection region R is of the form

$$R = \{X : T(X) \geq c\},$$

where $T(X)$ is a *test statistic* and c is a *critical value*. The problem of finding R, then translates to finding an appropriate test statistic $T(X)$ and an appropriate critical value c.

Hypothesis testing is like a legal trial. We assume someone is innocent unless the evidence strongly suggest that he is guilty. Similarly, we retain H_0 unless there is strong evidence to reject H_0. There are four possible outcomes:

(C1) accept H_0 when H_0 is true

(C2) reject H_0 when H_1 is true

(W1) reject H_0 when H_0 is true

(W2) accept H_0 when H_1 is true.

(C1) and (C2) are correct decisions, whereas (W1) and (W2) are wrong decisions. In particular, rejecting H_0 when H_0 is true is called *type I error*, whereas accepting H_0 when H_1 is true is called *type II error*.

The probability of each of the above incorrect decisions is called *size of the error*.

Definition 6.1 *The power function of a test with rejection region R is defined by*

$$\gamma(\theta) = P_\theta(X \in R).$$

The size of the rejection region or, equivalently, the size of type I error is defined by

$$\sup_{\theta \in \Theta_0} \gamma(\theta).$$

A test is said to have level α if $\sup_{\theta\in\Theta_0} \gamma(\theta) \leq \alpha$. Furthermore, if $\alpha = \sup_{\theta\in\Theta_0} \gamma(\theta)$, then we say that the test has size α.

The size of type II error is defined by

$$\beta(\theta) = 1 - \gamma(\theta), \quad \textit{for all } \theta \in \Theta_1.$$

A hypothesis of the form $H_n : \theta = \theta_0$ is called *simple hypothesis* because the underlying distribution is completely determined, $n = 0, 1$. A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called a *composite hypothesis*. A test for hypothesis of the form

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \neq \theta_0,$$

is called *two-sided* test. A test for

$$H_0 : \theta \leq \theta_0 \quad \text{against} \quad H_1 : \theta > \theta_0,$$

or

$$H_0 : \theta \geq \theta_0 \quad \text{against} \quad H_1 : \theta < \theta_0,$$

is called a *one-sided test*. The most common tests are two-sided.

The problem of hypothesis testing reduces to that of finding an *optimal* region R such that the size of the errors are minimised. In principle, an optimal R is the one for which $\alpha = \beta = 0$, that is an error-free rejection region. However, this is impossible. Intuitively, we if wanted $\alpha = 0$ then we need to take R small enough will never be able to reject H_0. This is fine if H_0 is true. But if H_0 is false, then the size of type II error will be $\beta = 1 - 0 = 1$. Thus, such an optimal rejection region cannot be achieved, some restrictions must be imposed.

The restriction typically adopted is to fix α in advance. The basic problem of testing a hypothesis H_0 is then to find a critical region of level or size α with minimum β. The value of α is often chosen taking into account practical considerations, and only critical regions of this size or less are permitted in the competition.

For a fixed θ, if there exists a critical region of size α with smallest β among all critical regions whose sizes do not exceed α, it is called **best critical region of size** α. A test based on a best critical region of size α is called a **best test of size** α.

Since the size of type II error rate is one minus the power of the test, then a best test of size α is also also a test with highest power. We would like then to construct tests with highest power under H_1, among all size α tests, for all θ. Such a test, when it exists, is called **uniformly most powerful test** (UMP). A UMP test is thus a best test, for all θ and since the latter is unknown it makes sense to focus on UMP tests.

To illustrate these concepts, consider the following example about testing a simple null hypothesis H_0 against a simple alternative hypothesis H_1.

Example 6.1 *Let X be a discrete r.v. whose p.d.f. depends upon a parameter θ and assume that we wish to test $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$, on the basis of a single observed value x. Hence, $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. Let the p.d.f. of X for $\theta = \theta_0$ and $\theta = \theta_1$ be as in the following table.*

X	*0*	*1*	*2*	*3*	*4*	*5*
$f(x;\theta_0)$	*.02*	*.03*	*.05*	*.05*	*.35*	*.50*
$f(x;\theta_1)$	*.04*	*.05*	*.08*	*.12*	*.41*	*.30*

If we choose $\alpha = .05$, the possible critical regions of this size are

(a) $R = \{0, 1\}$

(b) $R = \{2\}$

(c) $R = \{3\}$.

The value of β corresponding to these critical regions is (a) .91, (b) .92 and (c) .88. Among these critical regions, the region $R = \{3\}$ is therefore to be preferred because it leads to the lowest type II error rate, i.e it leads to highest power among the three critical regions of the same size. Before we can claim that it is the best critical region of size $\alpha = .05$, we have to assess there are no better critical region of size $\leq .05$. The non trivial critical regions are (d) $R = \{0\}$ and (e) $R = \{1\}$ for which β equals .96 and .95, respectively. It follows that the test

$$\textit{If } X = 3 \Longrightarrow \textit{ reject } H_0, \textit{ otherwise accept } H_0,$$

is a most powerful test of size .05.

A UMP test, when it exists, can be constructed through the following famous result.

Theorem 6.1 (Neyman-Pearson Lemma) *Let $x_1, \ldots, x_n$ be an observed sample from the random sample $X_i \overset{\text{iid}}{\sim} F_\theta$, $i = 1, \ldots, n$ with p.d.f. $f(x;\theta)$. If there exist a critical region C of size α and a nonnegative constant k such that*

$$\frac{\prod_{i=1}^n f(x_i;\theta_1)}{\prod_{i=1}^n f(x_i;\theta_0)} \geq k \quad \textit{for all} \quad (x_1, \ldots, x_n) \in C \tag{6.2}$$

and

$$\frac{\prod_{i=1}^n f(x_i;\theta_1)}{\prod_{i=1}^n f(x_i;\theta_0)} < k \quad \textit{for all} \ (x_1, \ldots, x_n) \notin C, \tag{6.3}$$

then C is a best critical region of size α.

Proof: To simplify notation let $x = (x_1, \ldots, x_n)$ and $dx = dx_1 \cdots dx_n$. Furthermore, we write $L_j(x) = \prod_{i=1}^n f(x_i;\theta_j)$, for $j = 0, 1$. Let C^* be any other critical region of size less then or equal to α. The two critical regions C and C^* may be represented by the sets of points labeled C and C^* in Figure 6.1. Their intersection is denoted by e and their non-intersecting parts by a and b, respectively.

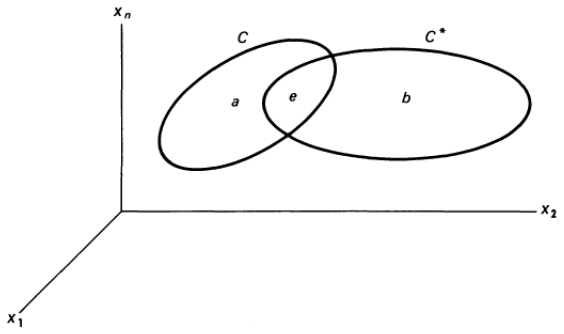

Figure 6.1: A graphical representation of the best critical region C compared to another critical region C^*.

Since C and C^* are critical regions of sizes α and $\leq \alpha$, respectively, it follows by the definition

of the size of a critical region that

$$\int_C \cdots \int \prod_{i=1}^{n} f(x_i;\theta_0)\,dx_1 \cdots dx_n = \int_C L_0(x)\,dx = \alpha \tag{6.4}$$

and that

$$\int_{C^*} \cdots \int \prod_{i=1}^{n} f(x_i;\theta_0)\,dx_1 \cdots dx_n = \int_{C^*} L_0(x)\,dx \leq \alpha. \tag{6.5}$$

Hence

$$\int_C L_0(x)\,dx \geq \int_{C^*} L_0(x)\,dx. \tag{6.6}$$

Writing $C = a + e$ and $C^* = b + e$, we may cancel the integral over e from both sides of (6.6) to reduce it to

$$\int_a L_0(x)\,dx \geq \int_b L_0(x)\,dx. \tag{6.7}$$

Let β and β^* denote the sizes of the type II error for the critical regions C and C^*, respectively. Since the size of a type II error is the probability that the sample point will fall outside the critical region when H_1 is true, which in turn is equal to one minus the probability that will fall inside the critical region when H_1 is true, we may write

$$\beta = 1 - \int_C L_1(x)\,dx, \quad \text{and} \quad \beta^* = 1 - \int_{C^*} L_1(x)\,dx.$$

Hence

$$\begin{aligned} \beta^* - \beta &= \int_C L_1(x)\,dx - \int_{C^*} L_1(x)\,dx \\ &= \int_a L_1(x)\,dx - \int_b L_1(x)\,dx, \end{aligned} \tag{6.8}$$

where in the last equality we have cancelled the integral over the common part of C and C^*.

From the definition of C given in (6.2), it follows that $L_1(x) \geq kL_0(x)$ for all points in C,

and hence for all points in a, and therefore that

$$\int_a L_1(x)\,dx \geq k \int_a L_0(x)\,dx.$$

Similarly, since b lies outside C, every point of b satisfies (6.3), namely $L_1(x) < kL_0(x)$; consequently

$$\int_b L_1(x)\,dx < k \int_b L_0(x)\,dx.$$

Applying these two results to (6.8) will yield the inequality

$$\beta^* - \beta \geq k \int_a L_0(x)\,dx - k \int_b L_0(x)\,dx. \tag{6.9}$$

But from (6.7) the right side must be nonnegative; therefore we arrive at the conclusion that

$$\beta^* \geq \beta.$$

Since β^* is the size of the type II error for any critical region, other than C, of size less then or equal to α, this proves that C is a best critical region of size α. ■

Here is an an illustration of how this theorem enables us to find a UMP test.

Example 6.2 *Let $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, for $i = 1, \ldots, n$ and consider the problem of testing the hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1 > \mu_0$, with σ^2 being known. Here*

$$\begin{aligned} \tfrac{L_1}{L_0} &= \frac{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu_1)^2\right]}{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu_0)^2\right]} \\ &= \exp\left[\tfrac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^n X_i + \tfrac{n(\mu_0^2-\mu_1^2)}{2\sigma^2}\right]. \end{aligned}$$

From (6.2) *it follows that the critical region C will be determined by the inequality*

$$\exp\left[\tfrac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^n X_i + \tfrac{n(\mu_1^2-\mu_0^2)}{2\sigma^2}\right] \geq k$$

for some constant $k > 0$. Taking logarithms will yield the equivalent inequality

$$\frac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^{n} X_i \geq \log k + \frac{n(\mu_1^2-\mu_0^2)}{2\sigma^2}.$$

It was assumed under H_1 that $\mu_1 > \mu_0$, so the previous inequality reduces to

$$\sum_{i=1}^{n} X_i \geq \frac{\sigma^2 \log k}{\mu_1-\mu_0} + \frac{n(\mu_1+\mu_0)}{2}.$$

Since k may be chosen to be any nonnegative number, as it ranges over values from 0 to ∞, the right side of this inequality will assume values from $-\infty$ to $+\infty$; therefore this inequality is equivalent to the inequality

$$\sum_{i=1}^{n} X_i \geq a,$$

where a may be chosen to be any real number. The equation $\sum_i X_i = a$ is that of a plane in n dimensional sample space. In Figure 6.2 the part of this plane with positive coordinates is sketched.

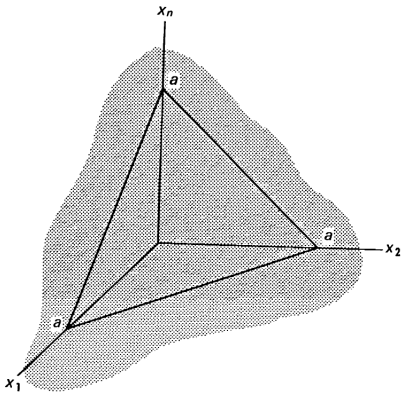

Figure 6.2: Graphical representation of the rejection region in Example 6.2.

The critical region C is therefore that part of the sample space which lies above this plane. As a assumes increasingly large numerical values, the region C includes increasingly less of the sample space, and as a goes from $-\infty$ to ∞ C shifts from including all of the sample space to none of it. Thus it is clear that by a proper choice of a, which is equivalent to a

proper choice of k we can choose C to be a region of any desired probability size α, other than 0 or 1.

The selected value of α affects only the value of the constant a and not the shape of the region C; therefore a test based on this type of critical region must be a best test of its size, whatever is the size. Furthermore, since $\sum_i X_i \geq a$ is equivalent to $\overline{X} > a/n$, the best test here is equivalent to one that is based on the statistic $\overline{X}$ and which chooses as critical region the interval $\overline{X} \geq b$, where b is a constant selected to satisfy $P_\theta(\overline{X} \geq b|H_0) = \alpha$. The striking feature of this test is its simplicity, in that its critical region can be made to depend only upon the statistic $\overline{X}$ rather than upon the n dimensional r.ve. $X_1, \ldots, X_n$.

As a numerical illustration, consider the problem of testing for the mean energy consumption of a population of WM's being either 8 or 10. Suppose that the experience has shown that energy consumption may be treated as a normal variable with $\sigma = 2$ and suppose that a random sample of size $n = 16$ yielded $\overline{x} = 9$. The problem then is to test the hypothesis $H_0 : \mu = 8$ against $H_1 : \mu = 10$ by means of the sample information. We shall choose $\alpha = .05$. Since the best test here is based on the critical region $\overline{X} \geq b$, where b is chosen to satisfy $P_\theta(\overline{X} \geq b|H_0) = .05$. Now when $\mu = \mu_0 = 8$, $\overline{X} \sim \mathrm{N}(8, 2/\sqrt{16})$, so $Z = \frac{\overline{X}-8}{.5} \sim \mathrm{N}(0,1)$, when H_0 is true.

Hence

$$\begin{aligned} P(\overline{X} \geq b|\mu = 8) &= P\left(\tfrac{\overline{X}-8}{.5} \geq \tfrac{b-8}{.5}\big|\mu = 8\right) \\ &= P(Z \geq \tfrac{b-8}{.5}). \end{aligned}$$

By the properties of the standard normal distribution we have that $P(Z \geq 1.645) = .05$; consequently b must be chosen to satisfy the equation $\frac{b-8}{.5} = 1.645$, which is equivalent to $b = 8.823$. Our critical region of size .05 therefore consists of those sample points for which $\overline{X} \geq 8.823$ and the UMP test is thus

$$\textit{Reject } H_0 \textit{ if } \overline{X} \geq 8.823.$$

The observed sample value $\overline{x} = 9$ falls in this critical region, hence H_0 is rejected in favour of H_1.

Example 6.3 (Example 6.2 cont.: Calculation of β) *Now let us evaluate β. For the example concerned with the testing a normal mean, assume again that $\mu_0 = 8$, $\mu_1 = 10$,*

$\sigma = 2$, $n = 16$ and $\alpha = .05$. The critical region for that problem was found to be $\overline{X} \geq 8.823$; therefore

$$\beta = P_\mu(\overline{X} \leq 8.823 | H_1).$$

Under H_1 $\overline{X}$ is a normal r.v. with mean 10 and standard deviation .5, hence $Z = \frac{\overline{X}-10}{.5} \sim$ $\mathrm{N}(0,1)$. Using properties of the normal distribution we obtain

$$\begin{aligned} \beta = P_\mu(\overline{X} \leq 8.823 | H_1) &= P\left(\tfrac{\overline{X}-10}{.5} \leq \tfrac{8.823-10}{.5} | \mu = 10\right) \\ &= P(Z \leq -2.36) = .009. \end{aligned}$$

The geometrical meaning of α and β for this problem are displayed in Figure 6.3. Note that the power of the test γ is given by $1 - \beta$.

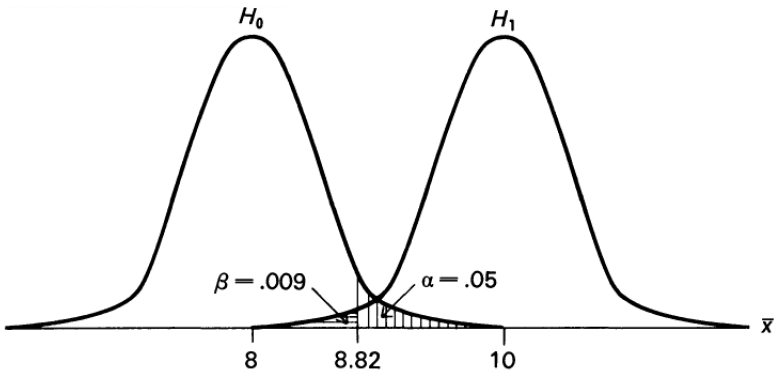

Figure 6.3: Geometrical representation of type I error and type II error when testing a normal mean.

The Neyman-Pearson Lemma is designed for testing simple null hypothesis versus a simple alternatives, i.e. $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. However, it can also be used to test one-sided hypothesis $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$, provided the model at hand satisfies the *monotone likelihood ratio* property.

Definition 6.2 *A family of distributions indexed by the real parameter θ is said to have a monotone likelihood ratio if there is a statistic T_n such that for each pair (θ, θ'), where $\theta > \theta'$, the likelihood ratio $L(\theta)/L(\theta')$ is a non decreasing function of T_n.*

Example 6.4 *Let $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$ where σ^2 is known. We wish to test $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$. Thus we have $\Theta_0 = (-\infty, \mu_0]$ and $\Theta_1 = (\mu_0, \infty)$, a composite null versus a composite alternative.*

To see if a UMP test exists we have to check if the model satisfies the monotone likelihood ratio property. For consider the pair (μ, μ'), where $\mu > \mu'$, then by the previous example we have that

$$\frac{L(\mu)}{L(\mu')} = \exp\left[\frac{\mu-\mu'}{\sigma^2}\sum_{i=1}^{n} X_i + \frac{n((\mu')^2-\mu^2)}{2\sigma^2}\right].$$

With $T_n = \overline{X}$ we see that the monotone likelihood ratio property is satisfied since, $\mu - \mu' > 0$, thus the NP Lemma tells us that the test:

$$\textit{Reject } \ H_0 \ \textit{ if } \ \overline{X} \geq c,$$

is UMP for $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$. To define the rejection region, take any θ_0, θ_1, such that $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$ and get the likelihood ratio as in Example 6.2. The rejection region is still of the form

$$R = \{(X_1 \ldots, X_n) : \overline{X} \geq b\},$$

regardless of the value of θ_0 and θ_1. We now define b such that R has size α. For we have to solve in b the equation

$$\sup_{\mu \leq \mu_0} \gamma(\mu) = \alpha \iff \sup_{\mu \leq \mu_0} P_\mu(\overline{X} \geq b) = \alpha.$$

Since $\sqrt{n}(\overline{X} - \mu)/\sigma$, when μ is the true value of the parameter, has the standard normal distribution, we see that

$$P_\mu\left(\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \geq \frac{\sqrt{n}(b-\mu)}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}(b-\mu)}{\sigma}\right).$$

Therefore

$$\begin{aligned}
\alpha &= \sup_{\mu \le \mu_0} P_\mu(\overline{X} \ge b) \\
&= \sup_{\mu \le \mu_0} P_\mu\left(\tfrac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \ge \tfrac{\sqrt{n}(b-\mu)}{\sigma}\right) \\
&= \sup_{\mu \le \mu_0} \left(1 - \Phi\left(\tfrac{\sqrt{n}(b-\mu)}{\sigma}\right)\right) \\
&= 1 - \inf_{\mu \le \mu_0} \Phi\left(\tfrac{\sqrt{n}(b-\mu)}{\sigma}\right) \\
&= 1 - \Phi\left(\tfrac{\sqrt{n}(b-\mu_0)}{\sigma}\right),
\end{aligned}$$

since $\Phi(-x)$ is a decreasing function in x. From the above equation we find that

$$\Phi\left(\tfrac{\sqrt{n}(b-\mu_0)}{\sigma}\right) = 1 - \alpha,$$

so $\frac{\sqrt{n}(b-\mu_0)}{\sigma} = z_\alpha$, and thus, $b = \mu_0 + z_\alpha\sigma/\sqrt{n}$, where z_α is the upper αth quantile of the standard normal distribution.

On the other hand, the size of type II error depends on μ and is

$$\beta(\mu) = \Phi\left(\tfrac{\sqrt{n}(\mu_0-\mu)}{\sigma} + z_\alpha\right), \text{ for all } \mu > \mu_0.$$

From this we deduce that the size of type II error is lower when:

- *μ is further apart from μ_0*
- *$\sqrt{n}$ is large*
- *α is large.*

The test is thus to reject H_0 when $\overline{X} \ge \mu_0 + z_\alpha\sigma/\sqrt{n}$ or, equivalently, we reject when $\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma} \ge z_\alpha$.

Most powerful tests do not always exist. For instance, in the above problem for the hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \ne \mu_0$ there is no UMP test. Thus instead of going deeper into UPM tests we'll just consider three widely used near optimal tests: the Wald test, the χ^2 test, and the likelihood ratio test.

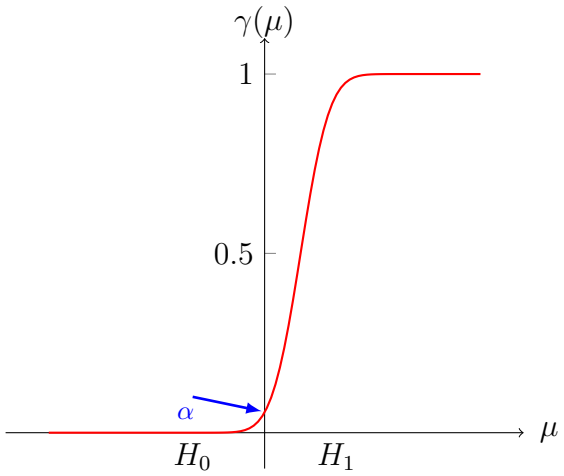

Figure 6.4: The power function for Example 6.4. The size of the test is the largest probability of rejecting H_0 when H_0 is true. This occurs at μ_0, hence the size is $\gamma(\mu_0)$. We choose the critical value b so that $\gamma(\mu_0) = \alpha$.

6.1 Wald test

Let θ be a scalar parameter, let $\widehat{\theta}$ be an estimator of θ and let $\widehat{\text{se}} = \widehat{\text{se}}(\widehat{\theta})$ (see Lecture 4, p. 18), be the estimated standard error of $\widehat{\theta}$. Consider testing

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \neq \theta_0,$$

and assume that $\widehat{\theta}$ is asymptotically normal, i.e. for large n,

$$\widehat{\theta} \;\dot{\sim}\; \text{N}(\theta_0, \widehat{\text{se}}^2).$$

The Wald test of approximate size α is to reject H_0 when $|W_n| > z_{\alpha/2}$, where

$$W_n = \frac{\widehat{\theta} - \theta_0}{\widehat{\text{se}}}$$

is called the *Wald statistic*.

The Wald test has size α asymptotically, as $n \to \infty$. Indeed, under $H_0 : \theta = \theta_0$ we have that

$$\begin{aligned} P_{\theta_0}(|W_n| > z_{\alpha/2}) &= P_{\theta_0}\left(\frac{|\widehat{\theta}-\theta_0|}{\widehat{\text{se}}} > z_{\alpha/2}\right) \\ &\to P(|Z| > z_{\alpha/2}) \\ &= \alpha, \end{aligned}$$

where $Z \sim \text{N}(0,1)$.

Remark 6.1 *An alternative version of the Wald test statistic is* $W_{n,0} = (\widehat{\theta} - \theta_0)/\text{se}_0$*, where* se_0 *is the standard error, i.e. the standard deviation of* $\widehat{\theta}$*, computed at* θ_0*. Both versions of the test are valid and are asymptotically equivalent.*

Let us consider the power of the Wald test under the alternative hypothesis. Since we are under the alternative hypothesis, the true value that generates the data is say $\theta_\star$, with $\theta_\star \neq \theta_0$. The power function at $\theta_\star$ is the probability of correctly rejecting the null hypothesis, is approximately (for large n) equal to

$$1 - \Phi\left(\frac{\theta_0-\theta_\star}{\widehat{\text{se}}} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0-\theta_\star}{\widehat{\text{se}}} - z_{\alpha/2}\right).$$

With all other things held equal, as $n \to \infty$ we see that the power goes to 1; recall that $\widehat{\text{se}} \to 0$ as $n \to \infty$ since $\widehat{\theta}$ is consistent. Furthermore, the power is large if $\theta_\star$ is far from θ_0.

Example 6.5 *(Comparing Two Prediction Algorithms). We run prediction algorithm 1 on a test set of size* n_1 *and prediction algorithm 2 on a second test set of size* n_2*. Let* X *be the number of incorrect predictions for algorithm 1 and let* Y *be the number of incorrect predictions of algorithm 2. Then* $X \sim \text{Bin}(n_1, \theta_1)$ *and* $Y \sim \text{Bin}(n_2, \theta_2)$*. We wish to verify if the two algorithms give the same number of incorrect predictions, thus the null hypothesis is* $H_0 : \theta_1 = \theta_2$ *against the alternative* $H_1 : \theta_1 \neq \theta_2$*. Letting* $\delta = \theta_1 - \theta_2$*, these two hypotheses can also be stated as*

$$H_0 : \delta = 0 \quad \textit{against} \quad H_1 : \delta \neq 0.$$

Let $\widehat{\theta}_1$ *and* $\widehat{\theta}_2$ *be the MLE of* θ_1 *and* θ_2*, respectively. By the equivariance principle of the MLE, we have that the MLE of* δ *is* $\widehat{\delta} = \widehat{\theta}_1 - \widehat{\theta}_2$*. The standard error of* $\widehat{\delta}$ *can be found by*

noting that $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are asymptotically independent normals (by invoking CLT or the large sample property of the MLE), thus

$$\widehat{\text{se}}(\widehat{\delta}) = \sqrt{\frac{\widehat{\theta}_1(1-\widehat{\theta}_1)}{n_1} + \frac{\widehat{\theta}_2(1-\widehat{\theta}_2)}{n_2}}.$$

The size α Wald test is to reject H_0 when $|W_n| > z_{\alpha/2}$ where

$$W_n = \frac{\widehat{\delta}-0}{\widehat{\text{se}}(\widehat{\delta})} = \frac{\widehat{\theta}_1-\widehat{\theta}_2}{\sqrt{\frac{\widehat{\theta}_1(1-\widehat{\theta}_1)}{n_1}+\frac{\widehat{\theta}_2(1-\widehat{\theta}_2)}{n_2}}}.$$

The power function of this test will be largest when θ_1 is far from θ_2 and when n_1 and n_2 are large.

What if we tested both algorithms on the same test set? The two samples are no longer independent. Instead we have to resort to the following strategy. Let $X_i = 1$ if algorithm 1 is correct on test case i and $X_i = 0$ otherwise. Let $Y_i = 1$ if algorithm 2 is correct on test case i, and $Y_i = 0$ otherwise. Define $D_i = X_i - Y_i$. A typical dataset will be something like this:

Test case	x_i	y_i	$d_i = x_i - y_i$
1	*1*	*0*	*1*
2	*1*	*1*	*0*
3	*1*	*1*	*0*
4	*0*	*1*	*-1*
5	*0*	*0*	*0*
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	*0*	*1*	*-1*

Let $\delta = E(D_i) = E(X_i) - E(Y_i) = P(X_i = 1) - P(Y_i = 1)$. We can estimate δ by the sample average of $d_1, \ldots, d_n$, thus let $\widehat{\delta} = \overline{d} = n^{-1}\sum_{i=1}^n d_i$. Furthermore, we can estimate the sampling variance of $\widehat{\delta}$ by s_d^2/n where $s_d^2 = (n-1)^{-1}\sum(d_i - \overline{d})^2$ and thus set $\widehat{\text{se}}(\widehat{\delta}) = \sqrt{s_d^2/n}$. To test $H_0 : \delta = 0$ against $H_1 : \delta \neq 0$ we use the test statistic $W_n = \widehat{\delta}/\widehat{\text{se}}(\widehat{\delta})$ and reject H_0 if $|W_n| > z_{\alpha/2}$. This last is called a test for paired samples.

Example 6.6 *(Nonparametric Comparison of Two Means). Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two independent random samples from populations with means μ_1 and μ_2, respectively,*

with both populations having finite variance. We are interested in testing the null hypothesis that $\mu_1 = \mu_2$. *Write this as* $H_0 : \delta = 0$ *against* $H_1 : \delta \neq 0$, *where* $\delta = \mu_1 - \mu_2$. *Let* $\widehat{\delta} = \widehat{\mu}_1 - \widehat{\mu}_2$, *where* $\widehat{\mu}_1 = \overline{X}$ *and* $\widehat{\mu}_2 = \overline{Y}$. *For large values of* m *and* n *the standard error of* $\widehat{\delta}$ *is*

$$\widehat{\text{se}} = \sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}.$$

The size α *Wald test rejects* H_0 *when* $|W_n| > z_{\alpha/2}$, *where*

$$W_n = \frac{\widehat{\delta} - 0}{\widehat{\text{se}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}.$$

There is close connection between Wald tests and Wald confidence intervals, which permits us to perform hypothesis testing through confidence intervals. Indeed, the size α Wald test rejects $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ if and only if $\theta_0 \notin \text{IC}_{1-\alpha}$, where $\text{IC}_{1-\alpha}$ is the $1 - \alpha$ Wald-type confidence interval for θ

$$\text{IC}_{1-\alpha} = [\widehat{\theta} \pm z_{\alpha/2}\, \widehat{\text{se}}].$$

Remark 6.2 *When we reject* H_0 *we often say that the result is statistically significant. A result might be statistically significant and yet the size of the effect might be practically or scientifically negligible. Thus a result could be statistically significant but not scientifically significant. The difference between statistical significance and scientific significance can be better understood in light the above connection between hypothesis testing and confidence intervals. Any interval that excludes* θ_0 *corresponds to a test which rejects* $H_0 : \theta = \theta_0$. *But the values in the interval could be close to* θ_0 *(not scientifically significant) or far from* θ_0 *(scientifically significant); see Figure 6.5. The message from this figure is that statistical significance does not imply that the finding is of scientific importance. Furthermore, confidence intervals are often more informative than tests.*

6.2 p-values

Reporting "reject H_0" or "accept H_0" is not very informative. Instead, we could try to see, for every α, whether the test rejects at that level. Generally, if the test rejects at level α it

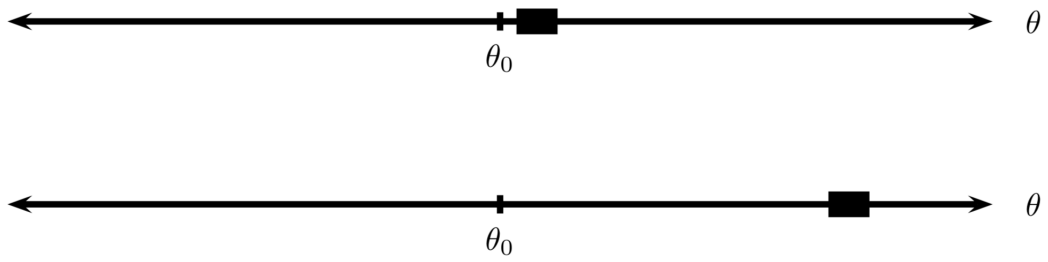

Figure 6.5: Scientific significance vs statistical significance. A level α test rejects $H_0 : \theta = \theta_0$ if and only if the $1 - \alpha$ confidence interval does not include θ_0. Here are shown two different confidence intervals, both excluding θ_0, so in both cases the test would both reject H_0. But in case on top, the estimated value of θ is close to θ_0 so the finding is probably of little scientific or practical value. In the bottom case, the estimated value of θ is far from θ_0 so the finding is of scientific value.

will also reject at level $\alpha' > \alpha$. Thus, there is smallest α at which the test rejects and we call this number the p-value.

Definition 6.3 *Suppose that for every $\alpha \in (0, 1)$ we have a size α test with rejection region R_α. Then*

$$p\text{-value} = \inf\{\alpha : T(X_1, \ldots, X_n) \in R_\alpha\}.$$

That is, the p-value is the smallest size at which we can reject H_0.

Informally, the p-value is a measure of evidence against H_0: the smallest the p-value, the stronger is the evidence against H_0. Typically, researchers use the following evidence scale:

- p-value $< .01 \Rightarrow$ very strong evidence against H_0.
- p-value $\in [.01, .05) \Rightarrow$ strong evidence against H_0.
- p-value $\in [.05, .1) \Rightarrow$ weak evidence against H_0.
- p-value $> .1 \Rightarrow$ little or no evidence against H_0.

Be aware that a large p-value is not strong evidence in favour of H_0. A large p-value can occur because (i) H_0 is true or (ii) H_0 is false but the test has low power. Another common confusion about the p-value is that it is sometimes interpreted as the probability of H_0 being true given the data. This is clearly false.

The following result shows how the p-value is computed.

Definition 6.4 *Let $T_n = T(X_1, \ldots, X_n)$ be a test statistic with observed value $t_n = t(x_1, \ldots, x_n)$. Then*

(i) if a size α test has rejection region of the form

$$\{X_1, \ldots, X_n : T(X_1, \ldots, X_n) \geq c\}$$

*then the p-*value *is defined by*

$$\sup_{\theta \in \Theta_0} P_\theta \left(T(X_1, \ldots, X_n) \geq t_n \right);$$

(ii) if a size α test has rejection region of the form

$$\{X_1, \ldots, X_n : T(X_1, \ldots, X_n) \leq c\}$$

*then the p-*value *is defined by*

$$\sup_{\theta \in \Theta_0} P_\theta \left(T(X_1, \ldots, X_n) \leq t_n \right);$$

(iii) if a size α test has rejection region of the form

$$\{X_1, \ldots, X_n : T(X_1, \ldots, X_n) \geq c_1\} \cup \{X_1, \ldots, X_n : T(X_1, \ldots, X_n) \leq c_2\},$$

then

$$p\text{-value} = 2 \min \left(\sup_{\theta \in \Theta_0} P_\theta(T(X_1, \ldots, X_n) \leq t_n), \sup_{\theta \in \Theta_0} P_\theta(T(X_1, \ldots, X_n) \geq t_n) \right).$$

If $\Theta_0 = \{\theta_0\}$ then replace $\sup_{\theta \in \Theta_0} P_\theta$ by P_{θ_0}.

In words, Definition 6.4 (i) says that the p-value is the probability under H_0 of observing a value of the test statistic the same as or more extreme than what is actually observed.

In the case of the Wald test, if we let $w_n = (\widehat{\theta} - \theta_0)/\widehat{\text{se}}$ denote the observed value of the Wald statistic W_n, the p-value is given by

$$P_{\theta_0}(|W_n| \geq |w_n|) \doteq P(|Z| \geq |w_n|) = 2\Phi(-|w_n|).$$

This is further illustrated in Figure 6.6.

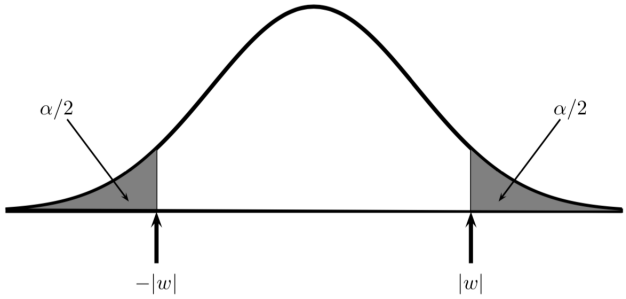

Figure 6.6: The p-value is the smallest α at which you would reject H_0. To find the p-value for the Wald test, we find α such that $|w_n|$ and $-|w_n|$ are just at the boundary of the rejection region. Here w_n is the observed value of the Wald statistic. Thus the p-value is the tail area $P(|Z| \geq |w_n|)$.

Note that the p-value depends on the observed data through w_n, thus it is a r.v. We state this more formally by the following theorem.

Theorem 6.2 *If the test statistic has a continuous distribution, then under $H_0 : \theta = \theta_0$, the* p-value *has distribution* $\text{Unif}(0, 1)$*. Therefore, if we reject H_0 when the* p-value *is less then α, the probability of a type I error is α.*

In other words, if H_0 is true, the p-value is like a random draw from a $\text{Unif}(0, 1)$ distribution. If H_1 is true, the distribution of the p-value will tend to concentrate closer to 0.

Example 6.7 *Regarding our motivating application about energy consumption of WM's, suppose the average energy consumption measured from $m = 10$ WM's with the old motor be $\overline{x} = 216$ and the let the sample variance be $s_1^2 = 5$. Furthermore, suppose that from a sample of $n = 15$ WM's with NGM1 we got $\overline{y} = 213$ and $s_2^2 = 2.5$. To verify that the average energy consumption of the population of WM's with the old motor μ_1 is equal to μ_2, the average energy consumption of the population of WM's with NGM1, consider $H_0 : \delta = 0$ against $H_1 : \delta \neq 0$, where $\delta = \mu_1 - \mu_2$. The observed Wald test statistic is*

$$w_n = \frac{\widehat{\delta} - 0}{\widehat{\text{se}}} = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{216 - 195}{\sqrt{0.5 + 0.167}} = 3.43.$$

The p-value is given by

$$P(|Z| \geq 3.43) = 2P(Z \leq -3.43) = .0006$$

which is very strong evidence against the null hypothesis.

6.3 Pearson's χ^2 test for multinomial data

Recall that if the r.ve. $(X_1, \ldots, X_k) \sim \text{Mn}(k, \theta_1, \ldots, \theta_k)$, then the MLE of $\theta = (\theta_1, \ldots, \theta_k)$ is $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$, where $\widehat{\theta}_i = X_i/n$, with $n = \sum_{i=1}^k X_i$.

Let $\theta = (\theta_1, \ldots, \theta_k)$ and let $\theta_0 = (\theta_{01}, \ldots, \theta_{0k})$ be some fixed vector and suppose we want to test

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \neq \theta_0.$$

Consider the statistic

$$T_n = \sum_{i=1}^k \frac{(X_j - n\theta_{0j})^2}{n\theta_{0j}} = \sum_{i=1}^k \frac{(X_j - E_j)^2}{E_j},$$

where $E_j = E(X_j) = n\theta_{0j}$ is the expected value of X_j under H_0. It can be shown that under H_0,

$$T_n \xrightarrow{d} \chi^2_{k-1} \quad \text{as } n \to \infty.$$

Hence the test: reject H_0 if $T_n \geq \chi^2_{k-1,\alpha}$, has asymptotic level α; here $\chi^2_{k-1,\alpha}$. The p-value is $P(\chi^2_{k-1} \geq t_n)$, where t_n is the observed value of test statistic T_n.

Example 6.8 *Consider again Mendel's experiment on peas, where round yellow seeds are breed with wrinkled green seeds. There are four type of progeny: round yellow, wrinkled yellow, round green, wrinkled green. The number of each type is a multinomial with probability $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$. His theory of inheritance predicts that θ is equal to*

$$\theta_0 = \left(\tfrac{9}{16}, \tfrac{3}{16}, \tfrac{3}{16}, \tfrac{1}{16}\right).$$

In $n = 556$ trials he observed $x = (315, 101, 108, 32)$. We will test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Since $n\theta_{01} = 312.75$, $n\theta_{02} = n\theta_{02} = 104.25$ and $n\theta_{04} = 34.75$, the observed test statistic is

$$t_n = \frac{(315-312.75)^2}{312.75} + \frac{(101-104.25)^2}{104.25} + \frac{(108-104.25)^2}{104.25} + \frac{(32-34.75)^2}{34.75} = 0.47.$$

With $\alpha = .05$ the threshold is $\chi^2_{3,.05} = 7.815$. Since 0.47 is not larger than 7.815 we do not reject H_0. In addition, the p-value is equal to $P(\chi^2_3 \geq 0.47) = .93$, which is not evidence against H_0. Hence the data do not contradict Mendel's theory.

6.4 The likelihood ratio test

The Wald test is mostly useful for testing a scalar parameter, although vector parameters can also be tested. The likelihood ratio test is more general and can be used for testing scalar and vector-valued parameters.

Definition 6.5 *Consider testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. The log-likelihood ratio statistic is defined as*

$$\lambda_n = 2 \log \left(\frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)} \right) = 2 \log \left(\frac{L(\widehat{\theta})}{L(\widehat{\theta}_0)} \right), \tag{6.10}$$

where $\widehat{\theta}$ is the MLE and $\widehat{\theta}_0$ is the restricted MLE in the space Θ_0. The likelihood ratio test (LRT) is then

Reject H_0 if the observed λ_n is greater or equal to c.

where c is a positive critical value such that $\sup_{\theta \in \Theta_0} P_\theta(\lambda_n \geq c) = \alpha$.

The log-likelihood ratio statistic λ_n is a r.v., although we have suppressed the dependence on the sample to ease notation. The LRT thus has always rejection region of the form

$$\{X_1, \ldots, X_n : \lambda_n \geq c\},$$

irrespective of the specific form of H_0 and H_1.

The likelihood ratio $\Lambda_n = \sup_{\theta \in \Theta} L(\theta) / \sup_{\theta \in \Theta_0} L(\theta)$ is always greater or equal to 1 and if H_0 is false, the ratio Λ_n takes on large values since the most likely values of θ will be in Θ_1. Thus large values of the likelihood ratio statistic $\lambda_n = 2 \log \Lambda_n$ indicate that the hypothesis H_1 is more likely to be true than H_0.

To determine the critical value c we need to know the distribution of λ_n. In many cases this is impossible, but under mild regularity conditions on the model F_θ, it can be shown that λ_n has limiting distribution distribution as $n \to \infty$.

To see this formally, let $\theta \in \mathbb{R}^r$ and suppose that Θ_0 consists of all θ such that some components are fixed at some particular values and the rest are left free.

Theorem 6.3 *Suppose that* $\theta = (\theta_1, \ldots, \theta_q, \theta_{q+1}, \ldots, \theta_r)$ *and let*

$$\Theta_0 = \{\theta : \theta_{q+1} = \theta_{0,q+1}, \theta_{q+2} = \theta_{0,q+2}, \ldots, \theta_r = \theta_{0,r}\}.$$

Under $H_0 : \theta \in \Theta_0$

$$\lambda_n \xrightarrow{d} \chi^2_{r-q} \quad \text{as } n \to \infty.$$

The degrees of freedom in the limiting distribution are $r - q = dim(\Theta) - dim(\Theta_0)$*;* $dim(S)$ *denotes the dimension of the space* S*.*

For example, if $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$ and we want to test the null hypothesis that $\theta_4 = \theta_5 = 0$, then the limiting distribution of the likelihood ratio test has $5 - 3 = 2$ degrees of freedom. Or equivalently, the degrees of freedom in the χ^2 distribution of the LRT are equal to the number of fixed parameters under H_0.

The critical value for a size α LRT is $c = \chi^2_{r-q,\alpha}$ and the p-value for this test is $P(\chi^2_{r-q} \geq \lambda_n^{obs})$, where λ_n^{obs} is the observed value of log-likelihood ratio statistic.

Example 6.9 *(Example 6.8 revisited) The observed value of the likelihood ratio test in the case of Example 6.8 is*

$$\begin{aligned}
\lambda_n &= 2\log\left(\tfrac{L(\widehat{\theta})}{L(\theta_0)}\right) \\
&= 2\sum_{j=1}^{k} x_j \log\left(\tfrac{\widehat{\theta}_j}{\theta_{0j}}\right) \\
&= 2\left(315\log\left(\tfrac{315/556}{9/16}\right) + 101\log\left(\tfrac{101/556}{3/16}\right) + 108\log\left(\tfrac{108/556}{3/16}\right) + 32\log\left(\tfrac{32/556}{1/16}\right)\right) \\
&= 0.48.
\end{aligned}$$

Under H_1 *there are four parameters. However, the parameters must sum to one, so the dimension of the parameter space is three. Under* H_0 *there are no free parameters so the dimension of the restricted parameter space is zero. The difference of these two dimension is three, so the limiting distribution of* Λ_n *under* H_0 *is* χ^2_3 *and the p-value is* $P(\chi^2_3 \geq 0.48) = .92$*. The conclusion is the same as with the Pearson's* χ^2 *test.*

6.4.1 Likelihood-based confidence sets

From the previous section we see that under H_0 the log-likelihood ratio statistic λ_n is an asymptotic pivot. In Lecture 5, we learned that a confidence set with the desired confidence level can be obtained by inverting a (asymptotic) pivot. If the pivot is λ_n, the implied confidence sets are called likelihood-based confidence sets.

Let $\theta \in \mathbb{R}^r$, and suppose $H_0 : \theta_1 = \theta_{0,1}, \ldots, \theta_{0,r}$. Then by the previous result, under H_0

$$\lambda_n \xrightarrow{d} \chi^2_r,$$

and the rejection region of the LRT of size α is

$$B_\alpha = \{X_1, \ldots, X_n : \lambda_n \geq \chi^2_{r,\alpha}\}.$$

A $1-\alpha$ likelihood-based confidence set for θ is defined by the set

$$\{\theta : \lambda_n(\theta) \notin B_\alpha\} = \{\theta : 2(\ell(\widehat{\theta}) - \ell(\theta)) \leq \chi^2_{r,\alpha}\}.$$

This is the set of all parameter values that under H_0 and for a given sample, lead to log-likelihood ratio statistics outside the rejection region of size α. That is, the set of all values θ_0 for which we would end up not rejecting $H_0 : \theta = \theta_0$ at the level α. Confidence set of subset of components of the parameter can also be defined in a meaningful way.

Although both Wald and likelihood-based confidence sets have coverage probability that converges to $1-\alpha$ as $n \to \infty$, likelihood-based confidence sets tend to have coverages closer to the prescribed confidence level. However, the likelihood-based confidence sets are more difficult to compute since λ_n is typically not invertible analytically and numerical methods must be used.

6.5 Further examples of tests

6.5.1 Tests with exact null distributions

Example 6.10 (Example 5.4 revised) *As in Example 5.4, let* $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ *be a random*

sample of size n from a normal distribution with both parameters being unknown; thus $\theta = (\mu, \sigma^2)$. *We wish to test the hypothesis* $H_0 : \mu = \mu_0$ *against* $H_1 : \mu \neq \mu_0$. *Under* H_0, *the test statistic*

$$T_n = \frac{\sqrt{n}(\overline{X}-\mu_0)}{\sqrt{S^2}} \sim t_{n-1},$$

and the test which rejects H_0 *if* $|T_n| \geq t_{n-1,\alpha/2}$ *is called the t-test or Student t-test and has size* α. *The p-value is computed as* $P(|T_n| \geq |t_n^{obs}|) = 2P(t_{n-1} \geq |t_n^{obs}|)$; *here we use* t_n^{obs} *to denote the observed value of* T_n *in order to not confuse it with* t_{n-1}, *the t-Student distribution with* $n-1$ *degrees of freedom.*

Equivalently, the test accepts H_0 *if*

$$\begin{aligned} |T_n| \leq t_{n-1,\alpha/2} &\iff |\overline{X} - \mu_0| \leq t_{n-1,\alpha/2}\sqrt{S^2/n} \\ &\iff -t_{n-1,\alpha/2}\sqrt{S^2/n} \leq \overline{X} - \mu_0 \leq t_{n-1,\alpha/2}\sqrt{S^2/n} \\ &\iff \overline{X} - t_{n-1,\alpha/2}\sqrt{S^2/n} \leq \mu_0 \leq \overline{X} + t_{n-1,\alpha/2}\sqrt{S^2/n}, \end{aligned}$$

i.e. if $\mu_0 \in [\overline{X} \pm t_{n-1,\alpha/2}\sqrt{S^2/n}]$, *the* $1-\alpha$ *confidence interval for* μ *includes* μ_0.

It can be shown that the Student t-test is also a likelihood ratio test for the same null and alternative hypothesis as above. Let's work this out.

Under H_0 *we have that*

$$\sup_{\theta\in\Theta_0} L(\theta) = \frac{\exp\left[-\frac{1}{2\widehat{\sigma^2_{\mu_0}}}\sum_{i=1}^n (X_i-\mu_0)^2\right]}{(2\pi)^{n/2}\widehat{\sigma^2_{\mu_0}}^{n/2}}, \tag{6.11}$$

where $\widehat{\sigma^2_{\mu_0}} = \sum_{i=1}^n (X_i - \mu_0)^2/n$. *Under* H_1 *we have*

$$\sup_{\theta\in\Theta} L(\theta) = \frac{e^{-n/2}}{(2\pi)^{n/2}}\left[\frac{\sum_i (X_i-\overline{X})^2}{n}\right]^{-n/2}. \tag{6.12}$$

Dividing (6.12) *by* (6.11) *gives*

$$\Lambda_n = \left[\frac{\sum_i (X_i-\overline{X})^2}{\sum_i (X_i-\mu_0)^2}\right]^{-n/2},$$

which has critical region $2\log\Lambda_n \geq a$. *But this critical region is equivalent to the critical region* $\Lambda_n \geq \log(a/2)$, *which is equivalent to* $(\Lambda_n)^{n/2} \geq (\log(a/2))^{n/2} = b$. *But then*

$$\begin{aligned}\frac{\sum_i (X_i-\mu_0)^2}{\sum_i (X_i-\overline{X})^2} \geq b &\iff \frac{n(\overline{X}-\mu_0)^2}{\sum_i (X_i-\overline{X})^2} \geq b-1 \\ &\iff \frac{n(\overline{X}-\mu_0)^2(n-1)}{\sum_i (X_i-\overline{X})^2} \geq (b-1)(n-1) \\ &\iff \frac{n(\overline{X}-\mu_0)^2}{S^2} \geq (b-1)(n-1) \\ &\iff T_n^2 \geq (b-1)(n-1).\end{aligned}$$

We thus see that the critical region for the LRT is of the type $T_n^2 \geq d$ *or* $|T_n| \geq \sqrt{d} = c$*. In order to define a size* α *test it is sufficient then to find* c *such that* $P(|T_n| \geq c) = \alpha$*. This is given by* $c = t_{n-1,\alpha/2}$*, since* T_n *follows a t-Student distribution with* $n-1$ *degrees of freedom.*

As a numerical example suppose $x_1, \ldots, x_n$ *is a sample of energy consumption of* $n = 10$ *WM's for which* $\overline{x} = 201$ *and* $s^2 = 5^2$ *and suppose we wish to test* $H_0 : \mu = 200$ *against* $H_1 : \mu \neq 200$ *at the level* $\alpha = .05$*. Then*

$$t_n^{obs} = \frac{\sqrt{10}(201-200)}{5} = .632.$$

Since $t_{9,0.025} = 2.26$ *and .632 is not greater than 2.26 we do not reject* H_0 *at level* $\alpha = .05$*. Furthermore, the p-value for this observed statistic is* $2P(t_{n-1} > .632) = .543$*, which is no evidence against* H_0*.*

Example 6.11 (Example 5.4 revised II) *As in Example 5.4, let again* $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ *be a random sample of size n from a normal distribution with both parameters being unknown. We wish to test the hypothesis* $H_0 : \sigma^2 = \sigma_0^2$ *against* $H_1 : \sigma^2 \neq \sigma_0^2$*. There are two exact test that can be applied here, one is a log-likelihood ratio test and the other is an equi-tailed test. For the LRT we have*

$$\sup_{\theta \in \Theta_0} L(\theta) = \sup_{\mu} L(\mu, \sigma_0^2) = (2\pi\sigma_0^2)^{-n/2} \exp\left(-\tfrac{n\widehat{\sigma}^2}{2\sigma_0^2}\right), \quad \textit{and}$$

$$\sup_{\theta \in \Theta} L(\theta) = \sup_{\mu, \sigma^2} L(\mu, \sigma^2) = (2\pi\widehat{\sigma}^2)^{-n/2} \exp\left(-\tfrac{n}{2}\right),$$

thus the log-likelihood ratio statistic is

$$\lambda_n = 2\log\left(\frac{(2\pi\widehat{\sigma}^2)^{-n/2}\exp\left(-\frac{n}{2}\right)}{(2\pi\sigma_0^2)^{-n/2}\exp\left(-\frac{n\widehat{\sigma}^2}{2\sigma_0^2}\right)}\right).$$

To perform a LRT we need to find $c > 0$ such that $\sup_{\theta\in\Theta_0} P_\theta(X_1,\ldots,X_n : \lambda_n \geq c) = \alpha$. Now $\lambda_n \geq c$ is equivalent to $\Lambda_n \geq e^{c/2} = b$, thus the rejection region must be of the type

$$\left(\frac{\widehat{\sigma}^2}{\sigma_0^2}\right)^{-n/2}\exp\left(\frac{n\widehat{\sigma}^2}{2\sigma_0^2}\right) \geq be^{n/2} = a \tag{6.13}$$

Now the function $p(x) = x^{-m}e^{mx}$, for $m > 0$, is convex and with unique minim at 1. Therefore, $p(x)$ is monotone increasing, for $x > 1$ and monotone decreasing for $x < 1$ for all $x > 0$. In our case $x = \frac{\widehat{\sigma}^2}{\sigma_0^2}$, thus in order to define the rejection region we need to solve

$$\begin{aligned}
\alpha &= P_{\sigma_0^2}(X_1,\ldots,X_n : \lambda_n \geq c)\\
&= P_{\sigma_0^2}\left(X_1,\ldots,X_n : \left(\frac{\widehat{\sigma}^2}{\sigma_0^2}\right)^{-n/2}\exp\left(\frac{n\widehat{\sigma}^2}{2\sigma_0^2}\right) \geq a\right)\\
&= P_{\sigma_0^2}\left(X_1,\ldots,X_n : p(\widehat{\sigma}^2/\sigma_0^2) \geq a\right)\\
&= P_{\sigma_0^2}\left(X_1,\ldots,X_n : \widehat{\sigma}^2/\sigma_0^2 \leq a_2 \ \text{ or } \ \widehat{\sigma}^2/\sigma_0^2 \geq a_1\right)\\
&= P_{\sigma_0^2}\left(X_1,\ldots,X_n : \frac{\widehat{\sigma}^2}{\sigma_0^2} \leq a_2\right) + P_{\sigma_0^2}\left(X_1,\ldots,X_n : \frac{\widehat{\sigma}^2}{\sigma_0^2} \geq a_1\right)\\
&= P_{\sigma_0^2}\left(X_1,\ldots,X_n : \frac{n\widehat{\sigma}^2}{\sigma_0^2} \leq na_2\right) + P_{\sigma_0^2}\left(X_1,\ldots,X_n : \frac{n\widehat{\sigma}^2}{\sigma_0^2} \geq na_1\right),
\end{aligned}$$

where $a_1 < 1 < a_2$ are such that $p(a_1) = p(a_2) = a$. Setting $b_1 = na_1$ and $b_2 = na_2$ and recalling that under H_0, $T_n = \frac{n\widehat{\sigma}^2}{\sigma_0^2} \sim \chi^2_{n-1}$, the problem then becomes to find that value of a for which b_1 and b_2 satisfy

$$P\left(\chi^2_{n-1} \leq b_1\right) + P\left(\chi^2_{n-1} \geq b_2\right) = \alpha.$$

This problem must be solved numerically. The α level LRT would then be to reject H_0 if $T_n \leq b_1$ or if $T_n \geq b_2$. The complement of this rejection region corresponds the highest probability density $1-\alpha$ confidence interval for σ^2. The p-value also must be computed numerically by applying Definition 6.3.

The second approach would be to take $b_1 = \chi^2_{n-1,1-\alpha/2}, b_2 = \chi^2_{n-1,\alpha/2}$. In this case, the test

rejects H_0 at the level α if $T_n > \chi^2_{n-1,\alpha/2}$ or if $T_n < \chi^2_{n-1,1-\alpha/2}$; The p-value for this approach $2\min\{P(\chi^2_{n-1} > t_n^{obs}), 1 - P(\chi^2_{n-1} > t_n^{obs})\}$ since we may reject H_0 either because the sample variance is greater than σ_0^2, in this case $P(\chi^2_{n-1} > t_n^{obs})$ is low or because the sample variance is much lower than σ_0^2. Alternatively, we reject H_0 if σ_0^2 is outside the equi-tailed $1-\alpha$ confidence interval obtained in Example 5.4.

Let's consider a numerical example again. Suppose we wish to test $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 \neq 1$ at the level $\alpha = .05$. From the observed with $n = 10$ suppose we got $s^2 = 2^2$. Then the observed value of the statistic T_n is $t_n^{obs} = \frac{9\times 4}{1} = 36$. Since $\chi^2_{9,.025} = 19.02277$, and $t_n^{obs} > 19.02277$ we reject H_0. On the other hand the p-value is

$$p\text{-}value = 2\min[P(\chi^2_9 > t_n^{obs}), 1-P(\chi^2_9 > t_n^{obs})] = 2\min(1-0.99996, 0.99996) = 2(1-0.99996).$$

Note that although both test have exact level α none of them is optimal. An interesting question would be: which is better?

Example 6.12 (Example 5.5. revised) *As in Example 5.5, let $Y_i \overset{\text{iid}}{\sim} N(\mu_1, \sigma_1^2)$ be a random sample of size n and $X_j \overset{\text{iid}}{\sim} N(\mu_2, \sigma_2^2)$, be a random sample of size m where we assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and μ_1, μ_2, σ^2 are unknown. Now we wish to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$. Considering the pivot given in Example 5.5, under H_0 we have that*

$$T_n = \frac{\overline{Y} - \overline{X}}{\sqrt{S^2_{pool}\left(\frac{1}{m}+\frac{1}{n}\right)}},$$

where S^2_{pool} is the pooled variance (see L5, p.8). Thus the test rejects H_0 for values $|T_n| \geq t_{n+m-2,\alpha/2}$. This test is called the two-sample t-test and has level α. The p-value is computed by $P(|t_{n+m-2}| \geq |t_n^{obs})|$, where t_n^{obs} is the observed value of T_n.

By a reasoning similar to that applied in Example 6.10 it is easy to see that H_0 will be rejected if the $1-\alpha$ confidence interval

$$\left[\overline{Y} - \overline{X} \pm t_{n+m-2,\alpha/2}\sqrt{S^2_{pool}\left(\tfrac{1}{m}+\tfrac{1}{n}\right)}\right]$$

does not contain 0.

It is also possible to show that this test is also a likelihood ratio test. This sheds some light on where did the pivot introduced in L5 come from.

In applied work it often happens that one has measurements under k different experimental conditions and interest is on testing if the means of the k experimental conditions are equal. This problem is known as the Analysis of Variance (ANOVA).

Example 6.13 (The ANOVA test) *Following Example 2.6, and changing slightly notation in order deal with a more general situation, let y_{ij} be the observed measurements across the $j = 1, \ldots, k$ experimental conditions and the $i = 1, \ldots, n_j$ replications or sample units at each experimental condition. A typical dataset for this problem looks like the table below.*

Conditions				
1	*2*	*3*	$\cdots$	*k*
y_{11}	y_{12}	y_{13}	$\cdots$	y_{1k}
y_{21}	y_{22}	y_{23}	$\cdots$	y_{2k}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$y_{n_1 1}$	$y_{n_2 2}$	$y_{n_3 3}$	$\cdots$	$y_{n_k k}$

In Example 2.6 we assumed each variable to have normal distribution with mean μ_j and variance σ^2. In this case we have k variables (or treatments, in the ANOVA jargon), thus we have

$$Y_{ij} = \mu_j + \epsilon_{ij}\,, \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, k$$
$$\epsilon_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2),$$

where μ_j are often referred to as treatment means.

Let $\mu = \frac{1}{n}\sum_{j=1}^{k} n_j \mu_j$ be the overall mean, that is, the common mean of all treatments pulled in a single variable, where $n = \sum_{j=1}^{n} n_j$.

The ANOVA test is a test for the null hypothesis that all treatment means are equal against the alternative that at least two treatments have different means, i.e.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \textit{against} \quad H_1 : \mu_j \neq \mu_l, \textit{ for some } j, l = 1, \ldots, k.$$

To find a suitable test statistic for these hypothesis let's first estimate the μ_j by the usual sample average estimator $\overline{Y}_{\bullet j} = n_j^{-1}(Y_{1j} + \ldots, Y_{n_j j})$; the filled dot is to remind us that

we are summing over i's. Similarly, we can estimate μ *by its sample average estimator* $\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} Y_{ij} = \frac{1}{n}\sum_{j=1}^{k} n_j \overline{Y}_{\bullet j} = \overline{Y}$. *Now consider the overall variability of the data around the estimator of the overall mean, which we call* <u>total sum of squares</u> *(SST), given by*

$$\begin{aligned} SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y})^2 &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}[(\overline{Y}_{\bullet j} - \overline{Y}) + (Y_{ij} - \overline{Y}_{\bullet j})]^2 \\ &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\overline{Y}_{\bullet j} - \overline{Y})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{\bullet j})^2 \\ &= SSR + SSE, \end{aligned}$$

Thus the overall variability around the estimator of the overall mean μ *is equal to the variability of treatments' means estimators around the estimator of the overall mean (SSR) plus the variability of the data around the estimators of the treatment means (SSE). Now, if* H_0 *is true, then treatments' means estimators* $\overline{Y}_{\bullet j}$ *would all be approximately equal to the overall mean estimate* $\overline{Y}$*, thus SSR would be approximately zero. On the other hand if* $Y_{\bullet j}$ *are all different, SSR would be high, dominating over SSE. Hence a test statistic based on the ratio SSR/SSE seems a reasonable choice.*

Let S_j^2 *be the sample variance of the variable for the* j *treatment, thus*

$$S_j^2 = \tfrac{1}{n_j - 1}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{\bullet j})^2,$$

and note that

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_{\bullet j})^2 = \sum_{j=1}^{k}(n_j - 1)S_j^2.$$

But since $(n_j - 1)S_j^2/\sigma^2 \sim \chi^2_{n_j - 1}$ *and* S_j^2*'s are independent, then by the closure with respect to addition property of the gamma distribution it follows that*

$$SSE/\sigma^2 \sim \chi^2_{n-k}.$$

Furthermore, it can be shown that under H_0, $SSR/\sigma^2 \sim \chi^2_{k-1}$ *and that SSE and SSR are*

independent. Therefore the statistic

$$F = \frac{\frac{SSR}{\sigma^2(k-1)}}{\frac{SSE}{\sigma^2(n-k)}} = \frac{SSR/(k-1)}{SSE/(n-k)},$$

has distribution $F_{k-1,n-k}$ *and an* α *level test for* $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ *rejects the null hypothesis for values of* $F \geq F_{k-1,n-k,\alpha}$. *The p-value is computed as* $P(F_{k-1,n-k} \geq F^{obs})$, *where* F^{obs} *is the value of the F statistic at the observed data.*

The quantities involved in the ANOVA test are often reported in a table such as the following

	d.f.	*SS*	*Mean Square*	*F*	*p-value*
Treatment	$k-1$	*SSR*	*SSR/(k-1)*	$\frac{SSR/(k-1)}{SSE/(n-k)}$	$P(F_{k-1,n-k} \geq F^{obs})$
Residual	$n-k$	*SSE*	*SSE/(n-k)*		
Total	$n-1$	*SST*			

As a numerical illustration consider the following problem. Gas mileages are recorded during a series of road tests with four new models of Japanese luxury sedans. We wish to test the null hypothesis that all four models, on the average, give the same mileage.

	model A	*model B*	*Model C*	*Model D*
	22	*28*	*29*	*23*
	26	*24*	*32*	*24*
		29	*28*	
$Y_{\bullet j}$	*24*	*27*	*29.67*	*23.5*
$(n_j - 1)S_j^2$	*8*	*14*	*8.66*	*.5*

The overall sample average is $\overline{Y} = 26.5$. *The ANOVA for this data is thus*

	d.f.	*SS*	*Mean Square*	*F*	*p-value*
Treatment	3	*61.34*	*20.45*	3.94	$P(F_{k-1,n-k} \geq 3.94) = 0.072$
Residual	6	*31.16*	*5.19*		
Total	9	*92.5*			

Since the p-value is higher than $\alpha = .05$ *we do not reject* H_0. *On the other hand,* $F_{3,6,.05} = 4.757$ *and since the observed value of the F statistic is 3.94, again we do not reject* H_0.

Remark 6.3 *If the ANOVA test rejects H_0, there are many different ways in which the means of the k variables could be different, i.e $\mu_1 \neq \mu_2$ or $\mu_1 \neq \mu_3$ or both, etc. Typically a practitioner is interested in these differences and not on H_0 per se, thus after H_0 has been rejected, it is often of interest to learn which of the pair of means is significantly different from zero. This can be done either through t-tests or confidence intervals for the difference of means $\mu_i - \mu_j$, for $i \neq j = 1, \ldots, k$. In ANOVA jargon this is known as post-hoc analysis.*

Given that there are k means, we could compute at most $k(k-1)/2$ confidence intervals (or tests). But while the level of each confidence intervals is the nominal value $1-\alpha$, when we conduct simultaneously $k(k-1)/2$ confidence intervals, the joint confidence level is much lower. To see this let $\mathrm{IC}_{i,j}^{\alpha}$ denote the $1-\alpha$ confidence interval for $\delta_{i,j} = \mu_i - \mu_j$ for all $i \neq j = 1, \ldots, k$. Then, the probability that all confidence intervals contain their true parameter value is

$$
\begin{aligned}
P\left(\bigcap_{\substack{i,j=1\\ i<j}}^{k} \left\{\delta_{i,j} \in \mathrm{IC}_{i,j}^{\alpha}\right\}\right) &= P\left(\left\{\delta_{1,2} \in \mathrm{IC}_{1,2}^{\alpha}\right\} \cap \cdots \cap \left\{\delta_{k-1,k} \in \mathrm{IC}_{k-1,k}^{\alpha}\right\}\right) \\
&= 1 - P\left(\bigcup_{\substack{i,j=1\\ i<j}}^{k} \left\{\delta_{i,j} \notin \mathrm{IC}_{i,j}^{\alpha}\right\}\right) \\
&\geq 1 - \sum_{\substack{i,j=1\\ i<j}}^{k} \alpha \\
&= 1 - k(k-1)\alpha/2.
\end{aligned}
$$

Thus the probability that all the confidence intervals contain their true parameter is far from the nominal value. What's even worse is that, for large k the joint confidence level decreases to zero. The message from this is that when conducting a large number of confidence intervals (or test) simultaneously we could end up by finding a significant results just by luck. This is known as the simultaneous inference or the multiple comparison problem.

There are many solutions to the simultaneous inference problem in the statistical literature, but Bonferroni is the easiest. In post-hoc analysis and when the desired confidence level is, say 95%, the Bonferroni's method is to fix $\alpha < 5\%$, depending on the number of confidence intervals (or hypothesis tests) performed. For instance if there are k variables, and after

rejecting H_0 one is interested in the confidence intervals for all $\delta_{i,j}$, then Bonferroni's methods is to use as α the value $\alpha_B = 2\alpha/k(k-1)$. Indeed, with α_B in place of α we see that the joint confidence level of the intervals is no less than $1-\alpha$. The price paid for correcting for simultaneous inference is: wider confidence intervals for $\delta_{i,j}$. For instance, if $k = 4$ and we wanted simultaneous inference with confidence 95%, then the confidence intervals for $\delta_{i,j}$ should be built using $\alpha = \alpha_B \doteq 0.0083$.

6.5.2 Approximate α level tests based on the likelihood ratio

It can be shown that in the above situations the LRT does have size α and this happens because of of the simple structure of the implied model. Unfortunately, in many other occasions it is not possible to find a threshold c for which $P_\theta(\lambda_n \geq c) = \alpha$, let alone to satisfy the more stringent condition $\sup_{\theta\in\Theta_0} P_\theta(\lambda_n \geq c) = \alpha$. However, for a fixed $\theta \in \Theta_0$ Theorem 6.3 provides a good approximation for finite n and it can be used to build a LRT which has level close to α.

Example 6.14 (Poisson population). *Consider again Example 5.8. Let $Y_i \overset{\text{iid}}{\sim} \text{Poi}(\theta)$, $i = 1, \ldots, n$, and suppose we wish to test $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$. Let $\widehat{\theta} = \overline{Y}$ and consider the likelihood ratio*

$$\Lambda_n = \frac{\sup_{\theta\in\Theta} L(\theta)}{\sup_{\theta\in\Theta_0} L(\theta)} = \frac{e^{-n\widehat{\theta}}\widehat{\theta}^{\sum_{i=1}^n Y_i}}{e^{-n\theta_0}\theta_0^{\sum_{i=1}^n Y_i}} = e^{n(\theta_0-\overline{Y})}\left(\frac{\overline{Y}}{\theta_0}\right)^{n\overline{Y}}.$$

By Theorem 6.3 we know that under H_0,

$$\lambda_n = 2\log\Lambda_n = 2n(\theta_0 - \overline{Y}) + 2n\overline{Y}[\log\overline{Y} - \log\theta_0] \overset{\cdot}{\sim} \chi^2_1, \quad \text{as } n \to \infty.$$

Thus an asymptotic α level likelihood ratio test is to reject H_0 if the statistic $2n(\theta_0 - \overline{Y}) + 2n\overline{Y}[\log\overline{Y} - \log\theta_0]$ computed at the observed sample is larger than $\chi^2_{1,\alpha}$. The p-value for this test is given by $P(\chi^2_1 \geq 2\log\Lambda_n^{obs})$, where Λ_n^{obs} is the observed value of Λ_n.

As a numerical example consider the following number of bugs observed during the operation of a certain software installed on a server:

(5, 4, 1, 0, 0, 1, 1, 2, 1, 1)

Suppose we want to test $H_0 : \theta = 3$ vs $H_1 : \theta \neq 3$. Since the value of the statistic at this observed sample (7.884) is greater than $\chi^2_{1,.05}$ (3.841) we reject H_0. In addition, the p-value is .005, much lower than $\alpha = .05$ thus again we reject the null hypothesis.

We can also compute a likelihood confidence set for θ using Section 6.4.1. We show this set pictorially in Figure 6.14 in which a 95% confidence set is illustrated. In this example, the confidence set is an interval.

To obtain this confidence set we consider λ_n as a function of θ. We cut this function horizontally at the level $\chi^2_{1,\alpha}$, with $\alpha = .05$ and find the two points of intersection. The 95% likelihood confidence interval is then given by all values of θ which are within the two intersection points. The 95% confidence interval found is thus [0.93,2.52] and since 3, the value under H_0, is not included in this interval we reject H_0 and conclude again that the population mean is statistically different from 3.

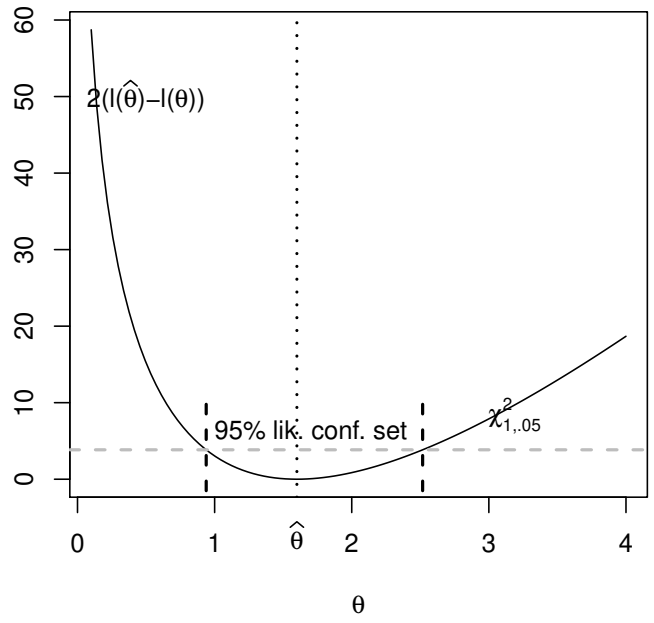

Figure 6.7: Graphical representation of a 95% likelihood confidence set for the mean θ of a Poisson population.

References

[HMC20] Hogg, R. V., McKeen, J. W. and Craig, A. T. (2018) *Introduction to Mathematical Statistics*, 8th edition, global ed., Pearson Education, Chapp. 4, 8, 9.