10
Hypothesis Testing and p-values

Suppose we want to know if exposure to asbestos is associated with lung disease. We take some rats and randomly divide them into two groups. We expose one group to asbestos and leave the second group unexposed. Then we compare the disease rate in the two groups. Consider the following two hypotheses:

The Null Hypothesis: The disease rate is the same in the two groups.

The Alternative Hypothesis: The disease rate is not the same in the two groups.

If the exposed group has a much higher rate of disease than the unexposed group then we will reject the null hypothesis and conclude that the evidence favors the alternative hypothesis. This is an example of hypothesis testing.

More formally, suppose that we partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and that we wish to test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1. \tag{10.1}$$

We call H_0 the **null hypothesis** and H_1 the **alternative hypothesis**.

Let X be a random variable and let $\mathcal{X}$ be the range of X. We test a hypothesis by finding an appropriate subset of outcomes $R \subset \mathcal{X}$ called the **rejection**

L. Wasserman, *All of Statistics*

	Retain Null	Reject Null
H_0 true	$\surd$	type I error
H_1 true	type II error	$\surd$

TABLE 10.1. Summary of outcomes of hypothesis testing.

region. If $X \in R$ we reject the null hypothesis, otherwise, we do not reject the null hypothesis:

$$\begin{aligned} X \in R &\implies \text{reject } H_0 \\ X \notin R &\implies \text{retain (do not reject) } H_0 \end{aligned}$$

Usually, the rejection region R is of the form

$$R = \left\{ x : \; T(x) > c \right\} \tag{10.2}$$

where T is a **test statistic** and c is a **critical value**. The problem in hypothesis testing is to find an appropriate test statistic T and an appropriate critical value c.

Warning! There is a tendency to use hypothesis testing methods even when they are not appropriate. Often, estimation and confidence intervals are better tools. Use hypothesis testing only when you want to test a well-defined hypothesis.

Hypothesis testing is like a legal trial. We assume someone is innocent unless the evidence strongly suggests that he is guilty. Similarly, we retain H_0 unless there is strong evidence to reject H_0. There are two types of errors we can make. Rejecting H_0 when H_0 is true is called a **type I error**. Retaining H_0 when H_1 is true is called a **type II error**. The possible outcomes for hypothesis testing are summarized in Tab. 10.1.

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

$$\beta(\theta) = \mathbb{P}_\theta(X \in R). \tag{10.3}$$

The **size** *of a test is defined to be*

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta). \tag{10.4}$$

A test is said to have **level** *α if its size is less than or equal to α.*

A hypothesis of the form $\theta = \theta_0$ is called a **simple hypothesis**. A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called a **composite hypothesis**. A test of the form

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

is called a **two-sided test**. A test of the form

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

or

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

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

10.2 Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma)$ where σ is known. We want to test $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$. Hence, $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, \infty)$. Consider the test:

$$\text{reject } H_0 \text{ if } T > c$$

where $T = \overline{X}$. The rejection region is

$$R = \left\{ (x_1, \ldots, x_n) : \; T(x_1, \ldots, x_n) > c \right\}.$$

Let Z denote a standard Normal random variable. The power function is

$$\begin{aligned} \beta(\mu) &= \mathbb{P}_\mu \left(\overline{X} > c \right) \\ &= \mathbb{P}_\mu \left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} > \frac{\sqrt{n}(c - \mu)}{\sigma} \right) \\ &= \mathbb{P} \left(Z > \frac{\sqrt{n}(c - \mu)}{\sigma} \right) \\ &= 1 - \Phi \left(\frac{\sqrt{n}(c - \mu)}{\sigma} \right). \end{aligned}$$

This function is increasing in μ. See Figure 10.1. Hence

$$\text{size} = \sup_{\mu \leq 0} \beta(\mu) = \beta(0) = 1 - \Phi \left(\frac{\sqrt{n} c}{\sigma} \right).$$

For a size α test, we set this equal to α and solve for c to get

$$c = \frac{\sigma \, \Phi^{-1}(1 - \alpha)}{\sqrt{n}}.$$

We reject when $\overline{X} > \sigma \, \Phi^{-1}(1 - \alpha) / \sqrt{n}$. Equivalently, we reject when

$$\frac{\sqrt{n} \, (\overline{X} - 0)}{\sigma} > z_\alpha.$$

where $z_\alpha = \Phi^{-1}(1 - \alpha)$. ■

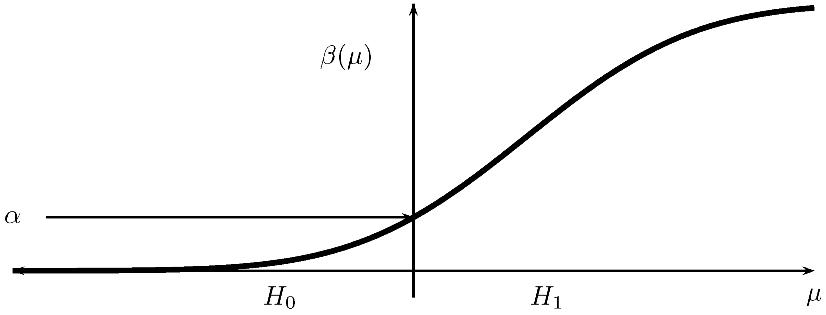

FIGURE 10.1. The power function for Example 10.2. The size of the test is the largest probability of rejecting H_0 when H_0 is true. This occurs at $\mu = 0$ hence the size is $\beta(0)$. We choose the critical value c so that $\beta(0) = \alpha$.

It would be desirable to find the test with highest power under H_1, among all size α tests. Such a test, if it exists, is called **most powerful**. Finding most powerful tests is hard and, in many cases, most powerful tests don't even exist. Instead of going into detail about when most powerful tests exist, we'll just consider four widely used tests: the Wald test,[1] the χ^2 test, the permutation test, and the likelihood ratio test.

10.1 The Wald Test

Let θ be a scalar parameter, let $\widehat{\theta}$ be an estimate of θ and let $\widehat{\mathsf{se}}$ be the estimated standard error of $\widehat{\theta}$.

[1] The test is named after Abraham Wald (1902–1950), who was a very influential mathematical statistician. Wald died in a plane crash in India in 1950.

10.3 Definition. The Wald Test

Consider testing

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

Assume that $\widehat{\theta}$ is asymptotically Normal:

$$\frac{(\widehat{\theta} - \theta_0)}{\widehat{\mathsf{se}}} \rightsquigarrow N(0,1).$$

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

$$W = \frac{\widehat{\theta} - \theta_0}{\widehat{\mathsf{se}}}. \tag{10.5}$$

10.4 Theorem. *Asymptotically, the Wald test has size α, that is,*

$$\mathbb{P}_{\theta_0}\left(|W| > z_{\alpha/2}\right) \to \alpha$$

as $n \to \infty$.

Proof. Under $\theta = \theta_0$, $(\widehat{\theta} - \theta_0)/\widehat{\mathsf{se}} \rightsquigarrow N(0,1)$. Hence, the probability of rejecting when the null $\theta = \theta_0$ is true is

$$\begin{aligned} \mathbb{P}_{\theta_0}\left(|W| > z_{\alpha/2}\right) &= \mathbb{P}_{\theta_0}\left(\frac{|\widehat{\theta} - \theta_0|}{\widehat{\mathsf{se}}} > z_{\alpha/2}\right) \\ &\to \mathbb{P}\left(|Z| > z_{\alpha/2}\right) \\ &= \alpha \end{aligned}$$

where $Z \sim N(0,1)$. ∎

10.5 Remark. An alternative version of the Wald test statistic is $W = (\widehat{\theta} - \theta_0)/\mathsf{se}_0$ where se_0 is the standard error computed at $\theta = \theta_0$. Both versions of the test are valid.

Let us consider the power of the Wald test when the null hypothesis is false.

10.6 Theorem. *Suppose the true value of θ is $\theta_\star \neq \theta_0$. The power $\beta(\theta_\star)$ — the probability of correctly rejecting the null hypothesis — is given (approximately) by*

$$1 - \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} - z_{\alpha/2}\right). \tag{10.6}$$

Recall that $\widehat{\mathsf{se}}$ tends to 0 as the sample size increases. Inspecting (10.6) closely we note that: (i) the power is large if $\theta_\star$ is far from θ_0, and (ii) the power is large if the sample size is large.

10.7 Example (Comparing Two Prediction Algorithms)**.** We test a prediction algorithm on a test set of size m and we test a second prediction algorithm on a second test set of size n. Let X be the number of incorrect predictions for algorithm 1 and let Y be the number of incorrect predictions for algorithm 2. Then $X \sim \text{Binomial}(m, p_1)$ and $Y \sim \text{Binomial}(n, p_2)$. To test the null hypothesis that $p_1 = p_2$ write

$$H_0 : \delta = 0 \quad \text{versus} \quad H_1 : \delta \neq 0$$

where $\delta = p_1 - p_2$. The MLE is $\widehat{\delta} = \widehat{p}_1 - \widehat{p}_2$ with estimated standard error

$$\widehat{\mathsf{se}} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}.$$

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

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathsf{se}}} = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}}.$$

The power of this test will be largest when p_1 is far from p_2 and when the sample sizes are large.

What if we used the same test set to test both algorithms? The two samples are no longer independent. Instead we use 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 look 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

$$\delta = \mathbb{E}(D_i) = \mathbb{E}(X_i) - \mathbb{E}(Y_i) = \mathbb{P}(X_i = 1) - \mathbb{P}(Y_i = 1).$$

The nonparametric plug-in estimate of δ is $\widehat{\delta} = \overline{D} = n^{-1}\sum_{i=1}^n D_i$ and $\widehat{\mathsf{se}}(\widehat{\delta}) = S/\sqrt{n}$, where $S^2 = n^{-1}\sum_{i=1}^n (D_i - \overline{D})^2$. To test $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$

we use $W = \widehat{\delta}/\widehat{\mathsf{se}}$ and reject H_0 if $|W| > z_{\alpha/2}$. This is called a **paired comparison.** ■

10.8 Example (Comparing Two Means). Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two independent samples from populations with means μ_1 and μ_2, respectively. Let's test the null hypothesis that $\mu_1 = \mu_2$. Write this as $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ where $\delta = \mu_1 - \mu_2$. Recall that the nonparametric plug-in estimate of δ is $\widehat{\delta} = \overline{X} - \overline{Y}$ with estimated standard error

$$\widehat{\mathsf{se}} = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

where s_1^2 and s_2^2 are the sample variances. The size α Wald test rejects H_0 when $|W| > z_{\alpha/2}$ where

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathsf{se}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}. \quad ■$$

10.9 Example (Comparing Two Medians). Consider the previous example again but let us test whether the medians of the two distributions are the same. Thus, $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ where $\delta = \nu_1 - \nu_2$ where ν_1 and ν_2 are the medians. The nonparametric plug-in estimate of δ is $\widehat{\delta} = \widehat{\nu}_1 - \widehat{\nu}_2$ where $\widehat{\nu}_1$ and $\widehat{\nu}_2$ are the sample medians. The estimated standard error $\widehat{\mathsf{se}}$ of $\widehat{\delta}$ can be obtained from the bootstrap. The Wald test statistic is $W = \widehat{\delta}/\widehat{\mathsf{se}}$. ■

There is a relationship between the Wald test and the $1 - \alpha$ asymptotic confidence interval $\widehat{\theta} \pm \widehat{\mathsf{se}}\, z_{\alpha/2}$.

10.10 Theorem. *The size α Wald test rejects $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ if and only if $\theta_0 \notin C$ where*

$$C = (\widehat{\theta} - \widehat{\mathsf{se}}\, z_{\alpha/2},\ \widehat{\theta} + \widehat{\mathsf{se}}\, z_{\alpha/2}).$$

Thus, testing the hypothesis is equivalent to checking whether the null value is in the confidence interval.

Warning! 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 small. In such a case we have a result that is statistically significant but not scientifically or practically significant. The difference between statistical significance and scientific significance is easy to understand in light of Theorem 10.10. Any confidence interval that excludes θ_0 corresponds to rejecting H_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 10.2.

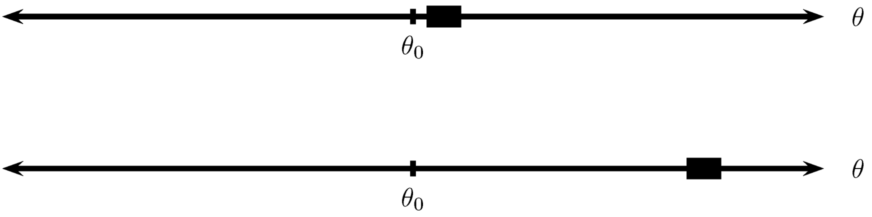

FIGURE 10.2. Scientific significance versus statistical significance. A level α test rejects $H_0 : \theta = \theta_0$ if and only if the $1 - \alpha$ confidence interval does not include θ_0. Here are two different confidence intervals. Both exclude θ_0 so in both cases the test would reject H_0. But in the first case, the estimated value of θ is close to θ_0 so the finding is probably of little scientific or practical value. In the second case, the estimated value of θ is far from θ_0 so the finding is of scientific value. This shows two things. First, statistical significance does not imply that a finding is of scientific importance. Second, confidence intervals are often more informative than tests.

10.2 p-values

Reporting "reject H_0" or "retain H_0" is not very informative. Instead, we could ask, for every α, whether the test rejects at that level. Generally, if the test rejects at level α it will also reject at level $\alpha' > \alpha$. Hence, there is a smallest α at which the test rejects and we call this number the p-value. See Figure 10.3.

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

$$\text{p-value} = \inf\left\{\alpha : \; T(X^n) \in R_\alpha\right\}.$$

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

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

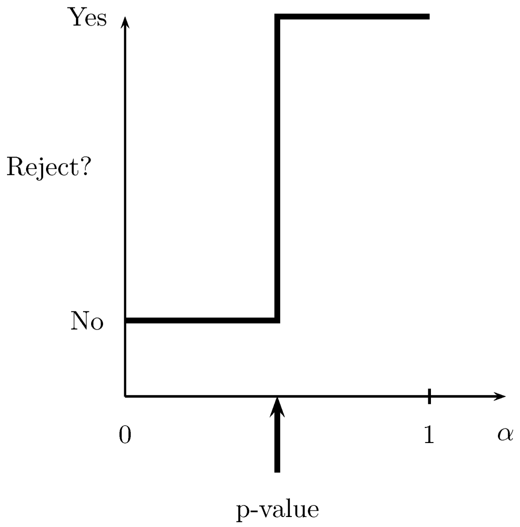

FIGURE 10.3. p-values explained. For each α we can ask: does our test reject H_0 at level α? The p-value is the smallest α at which we do reject H_0. If the evidence against H_0 is strong, the p-value will be small.

p-value	evidence
$< .01$	very strong evidence against H_0
$.01 - .05$	strong evidence against H_0
$.05 - .10$	weak evidence against H_0
$> .1$	little or no evidence against H_0

Warning! A large p-value is not strong evidence in favor of H_0. A large p-value can occur for two reasons: (i) H_0 is true or (ii) H_0 is false but the test has low power.

Warning! Do not confuse the p-value with $\mathbb{P}(H_0|\text{Data})$. [2] **The p-value is not the probability that the null hypothesis is true.**

The following result explains how to compute the p-value.

[2] We discuss quantities like $\mathbb{P}(H_0|\text{Data})$ in the chapter on Bayesian inference.

10.12 Theorem. *Suppose that the size α test is of the form*

$$\text{reject } H_0 \text{ if and only if } T(X^n) \geq c_\alpha.$$

Then,

$$\text{p-value} = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X^n) \geq T(x^n))$$

where x^n is the observed value of X^n. If $\Theta_0 = \{\theta_0\}$ then

$$\text{p-value} = \mathbb{P}_{\theta_0}(T(X^n) \geq T(x^n)).$$

We can express Theorem 10.12 as follows:

> 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 was actually observed.

10.13 Theorem. *Let $w = (\widehat{\theta} - \theta_0)/\widehat{\mathsf{se}}$ denote the observed value of the Wald statistic W. The p-value is given by*

$$\text{p} - \text{value} = \mathbb{P}_{\theta_0}(|W| > |w|) \approx \mathbb{P}(|Z| > |w|) = 2\Phi(-|w|) \qquad (10.7)$$

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

To understand this last theorem, look at Figure 10.4.

Here is an important property of p-values.

10.14 Theorem. *If the test statistic has a continuous distribution, then under $H_0 : \theta = \theta_0$, the p-value has a Uniform (0,1) distribution. Therefore, if we reject H_0 when the p-value is less than 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.

10.15 Example. Recall the cholesterol data from Example 7.15. To test if the means are different we compute

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathsf{se}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{216.2 - 195.3}{\sqrt{5^2 + 2.4^2}} = 3.78.$$

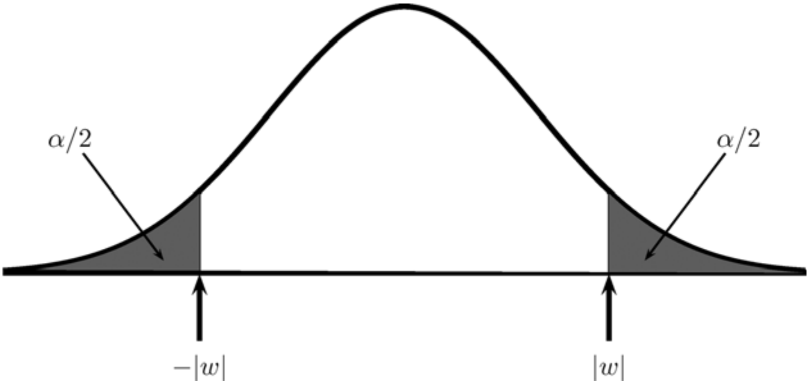

FIGURE 10.4. The p-value is the smallest α at which you would reject H_0. To find the p-value for the Wald test, we find α such that $|w|$ and $-|w|$ are just at the boundary of the rejection region. Here, w is the observed value of the Wald statistic: $w = (\widehat{\theta} - \theta_0)/\widehat{\mathsf{se}}$. This implies that the p-value is the tail area $\mathbb{P}(|Z| > |w|)$ where $Z \sim N(0,1)$.

To compute the p-value, let $Z \sim N(0,1)$ denote a standard Normal random variable. Then,

$$\text{p-value} = \mathbb{P}(|Z| > 3.78) = 2\mathbb{P}(Z < -3.78) = .0002$$

which is very strong evidence against the null hypothesis. To test if the medians are different, let $\widehat{\nu}_1$ and $\widehat{\nu}_2$ denote the sample medians. Then,

$$W = \frac{\widehat{\nu}_1 - \widehat{\nu}_2}{\widehat{\mathsf{se}}} = \frac{212.5 - 194}{7.7} = 2.4$$

where the standard error 7.7 was found using the bootstrap. The p-value is

$$\text{p-value} = \mathbb{P}(|Z| > 2.4) = 2\mathbb{P}(Z < -2.4) = .02$$

which is strong evidence against the null hypothesis. ■

10.3 The χ^2 Distribution

Before proceeding we need to discuss the χ^2 distribution. Let $Z_1, \ldots, Z_k$ be independent, standard Normals. Let $V = \sum_{i=1}^k Z_i^2$. Then we say that V has a χ^2 distribution with k degrees of freedom, written $V \sim \chi^2_k$. The probability density of V is

$$f(v) = \frac{v^{(k/2)-1}e^{-v/2}}{2^{k/2}\Gamma(k/2)}$$

for $v > 0$. It can be shown that $\mathbb{E}(V) = k$ and $\mathbb{V}(V) = 2k$. We define the upper α quantile $\chi^2_{k,\alpha} = F^{-1}(1-\alpha)$ where F is the CDF. That is, $\mathbb{P}(\chi^2_k > \chi^2_{k,\alpha}) = \alpha$.

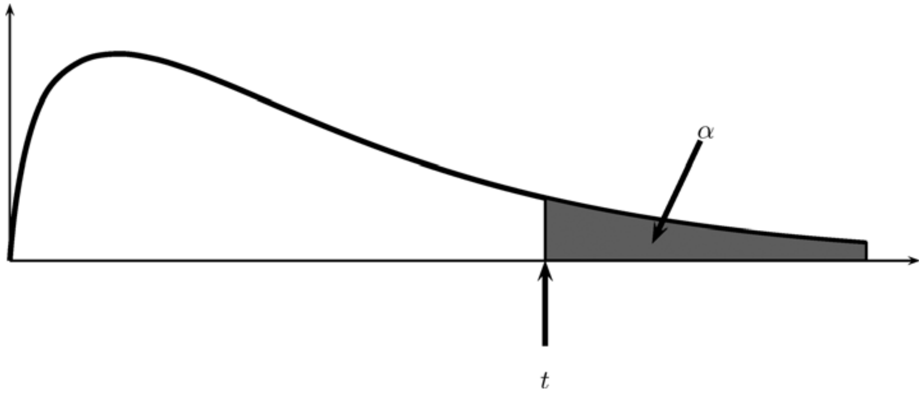

FIGURE 10.5. The p-value is the smallest α at which we would reject H_0. To find the p-value for the χ^2_{k-1} test, we find α such that the observed value t of the test statistic is just at the boundary of the rejection region. This implies that the p-value is the tail area $\mathbb{P}(\chi^2_{k-1} > t)$.

10.4 Pearson's χ^2 Test For Multinomial Data

Pearson's χ^2 test is used for multinomial data. Recall that if $X = (X_1, \ldots, X_k)$ has a multinomial (n, p) distribution, then the MLE of p is $\widehat{p} = (\widehat{p}_1, \ldots, \widehat{p}_k) = (X_1/n, \ldots, X_k/n)$.

Let $p_0 = (p_{01}, \ldots, p_{0k})$ be some fixed vector and suppose we want to test

$$H_0 : p = p_0 \text{ versus } H_1 : p \neq p_0.$$

10.16 Definition. Pearson's χ^2 statistic *is*

$$T = \sum_{j=1}^{k} \frac{(X_j - np_{0j})^2}{np_{0j}} = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j}$$

where $E_j = \mathbb{E}(X_j) = np_{0j}$ is the expected value of X_j under H_0.

10.17 Theorem. *Under H_0, $T \rightsquigarrow \chi^2_{k-1}$. Hence the test: reject H_0 if $T > \chi^2_{k-1,\alpha}$ has asymptotic level α. The p-value is $\mathbb{P}(\chi^2_{k-1} > t)$ where t is the observed value of the test statistic.*

Theorem 10.17 is illustrated in Figure 10.5.

10.18 Example (Mendel's peas)**.** Mendel bred peas with round yellow seeds and wrinkled green seeds. There are four types of progeny: round yellow, wrinkled yellow, round green, and wrinkled green. The number of each type is multinomial with probability $p = (p_1, p_2, p_3, p_4)$. His theory of inheritance predicts that p is equal to

$$p_0 \equiv \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right).$$

In $n = 556$ trials he observed $X = (315, 101, 108, 32)$. We will test $H_0 : p = p_0$ versus $H_1 : p \neq p_0$. Since, $np_{01} = 312.75, np_{02} = np_{03} = 104.25$, and $np_{04} = 34.75$, the test statistic is

$$\begin{aligned}\chi^2 &= \frac{(315 - 312.75)^2}{312.75} + \frac{(101 - 104.25)^2}{104.25} \\ &\quad + \frac{(108 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} = 0.47.\end{aligned}$$

The $\alpha = .05$ value for a χ^2_3 is 7.815. Since 0.47 is not larger than 7.815 we do not reject the null. The p-value is

$$\text{p-value} = \mathbb{P}(\chi^2_3 > .47) = .93$$

which is not evidence against H_0. Hence, the data do not contradict Mendel's theory.[3] ■

In the previous example, one could argue that hypothesis testing is not the right tool. Hypothesis testing is useful to see if there is evidence to reject H_0. This is appropriate when H_0 corresponds to the status quo. It is not useful for proving that H_0 is true. Failure to reject H_0 might occur because H_0 is true, but it might occur just because the test has low power. Perhaps a confidence set for the distance between p and p_0 might be more useful in this example.

10.5 The Permutation Test

The permutation test is a nonparametric method for testing whether two distributions are the same. This test is "exact," meaning that it is not based on large sample theory approximations. Suppose that $X_1, \ldots, X_m \sim F_X$ and $Y_1, \ldots, Y_n \sim F_Y$ are two independent samples and H_0 is the hypothesis that

[3] There is some controversy about whether Mendel's results are "too good."

the two samples are identically distributed. This is the type of hypothesis we would consider when testing whether a treatment differs from a placebo. More precisely we are testing

$$H_0 : F_X = F_Y \quad \text{versus} \quad H_1 : F_X \neq F_Y.$$

Let $T(x_1, \ldots, x_m, y_1, \ldots, y_n)$ be some test statistic, for example,

$$T(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = |\overline{X}_m - \overline{Y}_n|.$$

Let $N = m + n$ and consider forming all $N!$ permutations of the data X_1, ..., X_m, Y_1, ..., Y_n. For each permutation, compute the test statistic T. Denote these values by $T_1, \ldots, T_{N!}$. Under the null hypothesis, each of these values is equally likely. [4] The distribution $\mathbb{P}_0$ that puts mass $1/N!$ on each T_j is called the **permutation distribution** of T. Let $t_{\rm obs}$ be the observed value of the test statistic. Assuming we reject when T is large, the p-value is

$$\text{p-value} = \mathbb{P}_0(T > t_{obs}) = \frac{1}{N!}\sum_{j=1}^{N!} I(T_j > t_{obs}).$$

10.19 Example. Here is a toy example to make the idea clear. Suppose the data are: $(X_1, X_2, Y_1) = (1, 9, 3)$. Let $T(X_1, X_2, Y_1) = |\overline{X} - \overline{Y}| = 2$. The permutations are:

permutation	value of T	probability
(1,9,3)	2	1/6
(9,1,3)	2	1/6
(1,3,9)	7	1/6
(3,1,9)	7	1/6
(3,9,1)	5	1/6
(9,3,1)	5	1/6

The p-value is $\mathbb{P}(T > 2) = 4/6$. ∎

Usually, it is not practical to evaluate all $N!$ permutations. We can approximate the p-value by sampling randomly from the set of permutations. The fraction of times $T_j > t_{obs}$ among these samples approximates the p-value.

[4] More precisely, under the null hypothesis, given the ordered data values, $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ is uniformly distributed over the $N!$ permutations of the data.

Algorithm for Permutation Test

1. Compute the observed value of the test statistic
   $t_{\text{obs}} = T(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$.
2. Randomly permute the data. Compute the statistic again using the permuted data.
3. Repeat the previous step B times and let $T_1, \ldots, T_B$ denote the resulting values.
4. The approximate p-value is

$$\frac{1}{B}\sum_{j=1}^{B} I(T_j > t_{\text{obs}}).$$

10.20 Example. DNA microarrays allow researchers to measure the expression levels of thousands of genes. The data are the levels of messenger RNA (mRNA) of each gene, which is thought to provide a measure of how much protein that gene produces. Roughly, the larger the number, the more active the gene. The table below, reproduced from Efron et al. (2001) shows the expression levels for genes from ten patients with two types of liver cancer cells. There are 2,638 genes in this experiment but here we show just the first two. The data are log-ratios of the intensity levels of two different color dyes used on the arrays.

	Type I					Type II				
Patient	1	2	3	4	5	6	7	8	9	10
Gene 1	230	-1,350	-1,580	-400	-760	970	110	-50	-190	-200
Gene 2	470	-850	-.8	-280	120	390	-1730	-1360	-1	-330
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Let's test whether the median level of gene 1 is different between the two groups. Let ν_1 denote the median level of gene 1 of Type I and let ν_2 denote the median level of gene 1 of Type II. The absolute difference of sample medians is $T = |\widehat{\nu}_1 - \widehat{\nu}_2| = 710$. Now we estimate the permutation distribution by simulation and we find that the estimated p-value is .045. Thus, if we use a $\alpha = .05$ level of significance, we would say that there is evidence to reject the null hypothesis of no difference. ■

In large samples, the permutation test usually gives similar results to a test that is based on large sample theory. The permutation test is thus most useful for small samples.

10.6 The Likelihood Ratio Test

The Wald test is useful for testing a scalar parameter. The likelihood ratio test is more general and can be used for testing a vector-valued parameter.

10.21 Definition. *Consider testing*

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \notin \Theta_0.$$

The **likelihood ratio statistic** *is*

$$\lambda = 2 \log \left(\frac{\sup_{\theta \in \Theta} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)} \right) = 2 \log \left(\frac{\mathcal{L}(\widehat{\theta})}{\mathcal{L}(\widehat{\theta}_0)} \right)$$

where $\widehat{\theta}$ is the MLE *and $\widehat{\theta}_0$ is the* MLE *when θ is restricted to lie in Θ_0.*

You might have expected to see the maximum of the likelihood over Θ_0^c instead of Θ in the numerator. In practice, replacing Θ_0^c with Θ has little effect on the test statistic. Moreover, the theoretical properties of λ are much simpler if the test statistic is defined this way.

The likelihood ratio test is most useful when Θ_0 consists of all parameter values θ such that some coordinates of θ are fixed at particular values.

10.22 Theorem. *Suppose that $\theta = (\theta_1, \ldots, \theta_q, \theta_{q+1}, \ldots, \theta_r)$. Let*

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

Let λ be the likelihood ratio test statistic. Under $H_0 : \theta \in \Theta_0$,

$$\lambda(x^n) \rightsquigarrow \chi^2_{r-q,\alpha}$$

where $r - q$ is the dimension of Θ minus the dimension of Θ_0. The p-value for the test is $\mathbb{P}(\chi^2_{r-q} > \lambda)$.

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 has $5 - 3 = 2$ degrees of freedom.

10.23 Example (Mendel's Peas Revisited). Consider example 10.18 again. The likelihood ratio test statistic for $H_0 : p = p_0$ versus $H_1 : p \neq p_0$ is

$$\begin{aligned}
\lambda &= 2\log\left(\frac{\mathcal{L}(\widehat{p})}{\mathcal{L}(p_0)}\right) \\
&= 2\sum_{j=1}^{4} X_j \log\left(\frac{\widehat{p}_j}{p_{0j}}\right) \\
&= 2\left(315\log\left(\frac{\frac{315}{556}}{\frac{9}{16}}\right) + 101\log\left(\frac{\frac{101}{556}}{\frac{3}{16}}\right)\right. \\
&\quad \left. +108\log\left(\frac{\frac{108}{556}}{\frac{3}{16}}\right) + 32\log\left(\frac{\frac{32}{556}}{\frac{1}{16}}\right)\right) \\
&= 0.48.
\end{aligned}$$

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 dimensions is three. Therefore, the limiting distribution of λ under H_0 is χ^2_3 and the p-value is

$$\text{p-value} = \mathbb{P}(\chi^2_3 > .48) = .92.$$

The conclusion is the same as with the χ^2 test. ■

When the likelihood ratio test and the χ^2 test are both applicable, as in the last example, they usually lead to similar results as long as the sample size is large.

10.7 Multiple Testing

In some situations we may conduct many hypothesis tests. In example 10.20, there were actually 2,638 genes. If we tested for a difference for each gene, we would be conducting 2,638 separate hypothesis tests. Suppose each test is conducted at level α. For any one test, the chance of a false rejection of the null is α. But the chance of at least one false rejection is much higher. This is the **multiple testing problem.** The problem comes up in many data mining situations where one may end up testing thousands or even millions of hypotheses. There are many ways to deal with this problem. Here we discuss two methods.

Consider m hypothesis tests:

$$H_{0i} \quad \text{versus} \quad H_{1i}, \quad i = 1, \ldots, m$$

and let $P_1, \ldots, P_m$ denote the m p-values for these tests.

The Bonferroni Method

Given p-values $P_1, \ldots, P_m$, reject null hypothesis H_{0i} if

$$P_i < \frac{\alpha}{m}.$$

10.24 Theorem. *Using the Bonferroni method, the probability of falsely rejecting any null hypotheses is less than or equal to α.*

Proof. Let R be the event that at least one null hypothesis is falsely rejected. Let R_i be the event that the i^{th} null hypothesis is falsely rejected. Recall that if $A_1, \ldots, A_k$ are events then $\mathbb{P}(\bigcup_{i=1}^k A_i) \leq \sum_{i=1}^k \mathbb{P}(A_i)$. Hence,

$$\mathbb{P}(R) = \mathbb{P}\left(\bigcup_{i=1}^m R_i\right) \leq \sum_{i=1}^m \mathbb{P}(R_i) = \sum_{i=1}^m \frac{\alpha}{m} = \alpha$$

from Theorem 10.14. ■

10.25 Example. In the gene example, using $\alpha = .05$, we have that $.05/2,638 = .00001895375$. Hence, for any gene with p-value less than .00001895375, we declare that there is a significant difference. ■

The Bonferroni method is very conservative because it is trying to make it unlikely that you would make even one false rejection. Sometimes, a more reasonable idea is to control the **false discovery rate** (FDR) which is defined as the mean of the number of false rejections divided by the number of rejections.

Suppose we reject all null hypotheses whose p-values fall below some threshold. Let m_0 be the number of null hypotheses that are true and let $m_1 = m - m_0$. The tests can be categorized in a 2×2 as in Table 10.2.

Define the **False Discovery Proportion** (FDP)

$$\text{FDP} = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{if } R = 0. \end{cases}$$

The FDP is the proportion of rejections that are incorrect. Next define $\text{FDR} = \mathbb{E}(\text{FDP})$.

	H_0 Not Rejected	H_0 Rejected	Total
H_0 True	U	V	m_0
H_0 False	T	S	m_1
Total	$m - R$	R	m

TABLE 10.2. Types of outcomes in multiple testing.

The Benjamini-Hochberg (BH) Method

1. Let $P_{(1)} < \cdots < P_{(m)}$ denote the ordered p-values.
2. Define

$$\ell_i = \frac{i\alpha}{C_m m}, \qquad \text{and} \qquad R = \max\left\{ i : P_{(i)} < \ell_i \right\} \tag{10.8}$$

   where C_m is defined to be 1 if the p-values are independent and $C_m = \sum_{i=1}^{m} (1/i)$ otherwise.

3. Let $T = P_{(R)}$; we call T the **BH rejection threshold**.
4. Reject all null hypotheses H_{0i} for which $P_i \le T$.

10.26 Theorem (Benjamini and Hochberg). *If the procedure above is applied, then regardless of how many nulls are true and regardless of the distribution of the p-values when the null hypothesis is false,*

$$\text{FDR} = \mathbb{E}(\text{FDP}) \le \frac{m_0}{m}\alpha \le \alpha.$$

10.27 Example. Figure 10.6 shows six ordered p-values plotted as vertical lines. If we tested at level α without doing any correction for multiple testing, we would reject all hypotheses whose p-values are less than α. In this case, the four hypotheses corresponding to the four smallest p-values are rejected. The Bonferroni method rejects all hypotheses whose p-values are less than α/m. In this case, this leads to no rejections. The BH threshold corresponds to the last p-value that falls under the line with slope α. This leads to two hypotheses being rejected in this case. ■

10.28 Example. Suppose that 10 independent hypothesis tests are carried leading to the following ordered p-values:

```
0.00017 0.00448 0.00671 0.00907 0.01220
0.33626 0.39341 0.53882 0.58125 0.98617
```

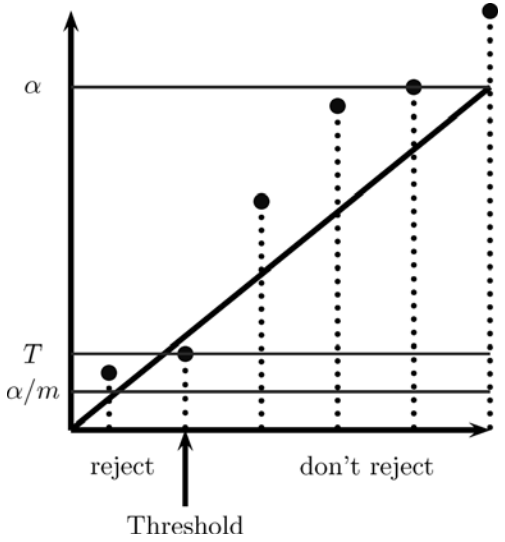

FIGURE 10.6. The Benjamini-Hochberg (BH) procedure. For uncorrected testing we reject when $P_i < \alpha$. For Bonferroni testing we reject when $P_i < \alpha/m$. The BH procedure rejects when $P_i \leq T$. The BH threshold T corresponds to the rightmost undercrossing of the upward sloping line.

With $\alpha = 0.05$, the Bonferroni test rejects any hypothesis whose p-value is less than $\alpha/10 = 0.005$. Thus, only the first two hypotheses are rejected. For the BH test, we find the largest i such that $P_{(i)} < i\alpha/m$, which in this case is $i = 5$. Thus we reject the first five hypotheses. ■

10.8 Goodness-of-fit Tests

There is another situation where testing arises, namely, when we want to check whether the data come from an assumed parametric model. There are many such tests; here is one.

Let $\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\}$ be a parametric model. Suppose the data take values on the real line. Divide the line into k disjoint intervals $I_1, \ldots, I_k$. For $j = 1, \ldots, k$, let

$$p_j(\theta) = \int_{I_j} f(x;\theta)\, dx$$

be the probability that an observation falls into interval I_j under the assumed model. Here, $\theta = (\theta_1, \ldots, \theta_s)$ are the parameters in the assumed model. Let N_j be the number of observations that fall into I_j. The likelihood for θ based on the counts $N_1, \ldots, N_k$ is the multinomial likelihood

$$Q(\theta) = \prod_{j=1}^{k} p_i(\theta)^{N_j}.$$

Maximizing $Q(\theta)$ yields estimates $\widetilde{\theta} = (\widetilde{\theta}_1, \ldots, \widetilde{\theta}_s)$ of θ. Now define the test statistic

$$Q = \sum_{j=1}^{k} \frac{(N_j - np_j(\widetilde{\theta}))^2}{np_j(\widetilde{\theta})}. \tag{10.9}$$

10.29 Theorem. *Let H_0 be the null hypothesis that the data are* IID*draws from the model $\mathfrak{F} = \{f(x;\theta) : \ \theta \in \Theta\}$. Under $H - 0$, the statistic Q defined in equation (10.9) converges in distribution to a χ^2_{k-1-s} random variable. Thus, the (approximate) p-value for the test is $\mathbb{P}(\chi^2_{k-1-s} > q)$ where q denotes the observed value of Q.*

It is tempting to replace $\widetilde{\theta}$ in (10.9) with the MLE $\widehat{\theta}$. However, this will not result in a statistic whose limiting distribution is a χ^2_{k-1-s}. However, it can be shown — due to a theorem of Herman Chernoff and Erich Lehmann from 1954 — that the p-value is bounded approximately by the p-values obtained using a χ^2_{k-1-s} and a χ^2_{k-1}.

Goodness-of-fit testing has some serious limitations. If reject H_0 then we conclude we should not use the model. But if we do not reject H_0 we cannot conclude that the model is correct. We may have failed to reject simply because the test did not have enough power. This is why it is better to use nonparametric methods whenever possible rather than relying on parametric assumptions.

10.9 Bibliographic Remarks

The most complete book on testing is Lehmann (1986). See also Chapter 8 of Casella and Berger (2002) and Chapter 9 of Rice (1995). The FDR method is due to Benjamini and Hochberg (1995). Some of the exercises are from Rice (1995).

10.10 Appendix

10.10.1 *The Neyman-Pearson Lemma*

In the special case of a simple null $H_0 : \theta = \theta_0$ and a simple alternative $H_1 : \theta = \theta_1$ we can say precisely what the most powerful test is.

10.30 Theorem (Neyman-Pearson). *Suppose we test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. Let*

$$T = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = \frac{\prod_{i=1}^n f(x_i; \theta_1)}{\prod_{i=1}^n f(x_i; \theta_0)}.$$

Suppose we reject H_0 when $T > k$. If we choose k so that $\mathbb{P}_{\theta_0}(T > k) = \alpha$ then this test is the most powerful, size α test. That is, among all tests with size α, this test maximizes the power $\beta(\theta_1)$.

10.10.2 *The t-test*

To test $H_0 : \mu = \mu_0$ where $\mu = \mathbb{E}(X_i)$ is the mean, we can use the Wald test. When the data are assumed to be Normal and the sample size is small, it is common instead to use the **t-test**. A random variable T has a *t-distribution with k degrees of freedom* if it has density

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)\left(1 + \frac{t^2}{k}\right)^{(k+1)/2}}.$$

When the degrees of freedom $k \to \infty$, this tends to a Normal distribution. When $k = 1$ it reduces to a Cauchy.

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)$ are both unknown. Suppose we want to test $\mu = \mu_0$ versus $\mu \neq \mu_0$. Let

$$T = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{S_n}$$

where S_n^2 is the sample variance. For large samples $T \approx N(0, 1)$ under H_0. The exact distribution of T under H_0 is t_{n-1}. Hence if we reject when $|T| > t_{n-1,\alpha/2}$ then we get a size α test. However, when n is moderately large, the t-test is essentially identical to the Wald test.

10.11 Exercises

1. Prove Theorem 10.6.

2. Prove Theorem 10.14.

3. Prove Theorem 10.10.

4. Prove Theorem 10.12.

5. Let $X_1, ..., X_n \sim \text{Uniform}(0, \theta)$ and let $Y = \max\{X_1, ..., X_n\}$. We want to test

   $H_0 : \theta = 1/2$ versus $H_1 : \theta > 1/2$.

   The Wald test is not appropriate since Y does not converge to a Normal. Suppose we decide to test this hypothesis by rejecting H_0 when $Y > c$.

   (a) Find the power function.

   (b) What choice of c will make the size of the test .05?

   (c) In a sample of size $n = 20$ with Y=0.48 what is the p-value? What conclusion about H_0 would you make?

   (d) In a sample of size $n = 20$ with Y=0.52 what is the p-value? What conclusion about H_0 would you make?

6. There is a theory that people can postpone their death until after an important event. To test the theory, Phillips and King (1988) collected data on deaths around the Jewish holiday Passover. Of 1919 deaths, 922 died the week before the holiday and 997 died the week after. Think of this as a binomial and test the null hypothesis that $\theta = 1/2$. Report and interpret the p-value. Also construct a confidence interval for θ.

7. In 1861, 10 essays appeared in the *New Orleans Daily Crescent.* They were signed "Quintus Curtius Snodgrass" and some people suspected they were actually written by Mark Twain. To investigate this, we will consider the proportion of three letter words found in an author's work. From eight Twain essays we have:

   .225 .262 .217 .240 .230 .229 .235 .217

   From 10 Snodgrass essays we have:

   .209 .205 .196 .210 .202 .207 .224 .223 .220 .201

   (a) Perform a Wald test for equality of the means. Use the nonparametric plug-in estimator. Report the p-value and a 95 per cent confidence interval for the difference of means. What do you conclude?

   (b) Now use a permutation test to avoid the use of large sample methods. What is your conclusion? (Brinegar (1963)).

8. Let $X_1, \ldots, X_n \sim N(\theta, 1)$. Consider testing

$$H_0 : \theta = 0 \text{ versus } \theta = 1.$$

Let the rejection region be $R = \{x^n : T(x^n) > c\}$ where $T(x^n) = n^{-1}\sum_{i=1}^n X_i$.

(a) Find c so that the test has size α.

(b) Find the power under H_1, that is, find $\beta(1)$.

(c) Show that $\beta(1) \to 1$ as $n \to \infty$.

9. Let $\widehat{\theta}$ be the MLE of a parameter θ and let $\widehat{\mathsf{se}} = \{nI(\widehat{\theta})\}^{-1/2}$ where $I(\theta)$ is the Fisher information. Consider testing

$$H_0 : \theta = \theta_0 \text{ versus } \theta \neq \theta_0.$$

Consider the Wald test with rejection region $R = \{x^n : |Z| > z_{\alpha/2}\}$ where $Z = (\widehat{\theta} - \theta_0)/\widehat{\mathsf{se}}$. Let $\theta_1 > \theta_0$ be some alternative. Show that $\beta(\theta_1) \to 1$.

10. Here are the number of elderly Jewish and Chinese women who died just before and after the Chinese Harvest Moon Festival.

Week	Chinese	Jewish
-2	55	141
-1	33	145
1	70	139
2	49	161

Compare the two mortality patterns. (Phillips and Smith (1990)).

11. A randomized, double-blind experiment was conducted to assess the effectiveness of several drugs for reducing postoperative nausea. The data are as follows.

	Number of Patients	Incidence of Nausea
Placebo	80	45
Chlorpromazine	75	26
Dimenhydrinate	85	52
Pentobarbital (100 mg)	67	35
Pentobarbital (150 mg)	85	37

(a) Test each drug versus the placebo at the 5 per cent level. Also, report the estimated odds–ratios. Summarize your findings.

(b) Use the Bonferroni and the FDR method to adjust for multiple testing. (Beecher (1959)).

12. Let $X_1, ..., X_n \sim \text{Poisson}(\lambda)$.

    (a) Let $\lambda_0 > 0$. Find the size α Wald test for

    $$H_0 : \lambda = \lambda_0 \quad \text{versus} \quad H_1 : \lambda \neq \lambda_0.$$

    (b) (Computer Experiment.) Let $\lambda_0 = 1$, $n = 20$ and $\alpha = .05$. Simulate $X_1, \ldots, X_n \sim \text{Poisson}(\lambda_0)$ and perform the Wald test. Repeat many times and count how often you reject the null. How close is the type I error rate to .05?

13. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Construct the likelihood ratio test for

    $$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

    Compare to the Wald test.

14. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Construct the likelihood ratio test for

    $$H_0 : \sigma = \sigma_0 \quad \text{versus} \quad H_1 : \sigma \neq \sigma_0.$$

    Compare to the Wald test.

15. Let $X \sim \text{Binomial}(n, p)$. Construct the likelihood ratio test for

    $$H_0 : p = p_0 \quad \text{versus} \quad H_1 : p \neq p_0.$$

    Compare to the Wald test.

16. Let θ be a scalar parameter and suppose we test

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

    Let W be the Wald test statistic and let λ be the likelihood ratio test statistic. Show that these tests are equivalent in the sense that

    $$\frac{W^2}{\lambda} \xrightarrow{\text{P}} 1$$

    as $n \to \infty$. Hint: Use a Taylor expansion of the log-likelihood $\ell(\theta)$ to show that

    $$\lambda \approx \left(\sqrt{n}(\widehat{\theta} - \theta_0)\right)^2 \left(-\frac{1}{n}\ell''(\widehat{\theta})\right).$$