1 Testing Hypotheses

In many statistical applications, the researcher wishes to ascertain whether a hypothesized value of a characteristic $\psi(\theta)$ of the population is consistent with the observed data, s. We write this hypothesis as $H_0 : \psi(\theta) = \psi_0$ and call this the null hypothesis.

A test of significance (or a hypothesis test) provides a measure of how unlikely the observed data s appear under the assumption that the null hypothesis is true. We can assess the evidence for the null hypothesis using a probability called the $P-$value. Small values of the $P-$value indicate that a surprising event has occured and suggest that the null hypothesis should be rejected.

In contrast, the Neyman-Pearson approach to hypothesis testing formulates the problem as a choice between two competing hypotheses concerning the population characteristic. This approach concentrates on the two error probabilities that arise when making decisions based on data.

2 Introduction to Hypothesis Testing

We will start out by considering a simple example. Suppose that we have two coins:

- coin 0 with probability of heads equal to 0.5
- coin 1 with probability of heads equal to 0.8

We randomly choose a coin and toss it eight times.

I tell you how many heads were obtained and your job is to use this information to determine which coin I tossed.

Let X denote the number of heads. The following table gives $p(x)$ for each of the coins:

x	0	1	2	3	4	5	6	7	8
coin 0	0.004	0.031	0.109	0.219	0.273	0.219	0.109	0.031	0.004
coin 1	0.000	0.000	0.001	0.009	0.046	0.147	0.294	0.336	0.168

Suppose you observe $X = 3$ coins with heads. We compare the probabilities of this outcome for the two coins using the likelihood ratio,

$$\frac{p_1(3)}{p_0(3)} = \frac{0.009}{0.219} = 0.042.$$

Thus, coin 0 is about 24 times as likely as coin 1 to produce the result $X = 3$.

On the other hand, if one observed $X = 7$ heads, the likelihood ratio would be

$$\frac{p_1(7)}{p_0(7)} = \frac{0.336}{0.031} = 10.74$$

which favors coin 1.

2.1 A Bayesian Approach

We specify two hypotheses:

- H_0 : coin 0 was used
- H_a : coin 1 was used

Suppose that we can assign prior probabilities to H_0 and H_a before observing any data. Then after observing $X = x$ heads, the posterior probabilities would be $P(H_0|x)$ and $P(H_a|x)$. For instance,

$$P(H_0|x) = \frac{P(H_0, X = x)}{P(x)} = \frac{P(x|H_0)P(H_0)}{P(x)}.$$

The ratio of the two posterior probabilities is

$$\frac{P(H_a|x)}{P(H_0|x)} = \frac{P(x|H_a)P(H_a)}{P(x|H_0)P(H_0)}.$$

Thus, the ratio of the two posterior probabilities is the product of the ratio of the prior probabilities and the likelihood ratio. Hence, the information in the data is contained in the likelihood ratio. We now examine the likelihood ratio for our example:

x	0	1	2	3	4	5	6	7	8
$\frac{P(x\|H_a)}{P(x\|H_0)}$	0.0006	0.0026	0.0104	0.042	0.168	0.671	2.684	10.74	42.95

We see that the likelihood ratio is a monotone function of x, increasing as x increases. The evidence is increasingly favorable to H_0 as x decreases and increasingly favorable to H_a as x increases. Using Bayesian reasoning, you would choose H_a if

$$\frac{P(H_a|x)}{P(H_0|x)} = \frac{P(x|H_a)P(H_a)}{P(x|H_0)P(H_0)} > 1.$$

or equivalently if,

$$\frac{P(x|H_a)}{P(x|H_0)} > c.$$

The cut-off value c is determined by the ratio of the prior probabilities. The value of c determines your decision rule. Suppose that $c = 1$ (equal prior probabilities for H_0 and H_a). Then you accept H_0 if $X \leq 5$ and accept H_a if $X > 5$. There are two possible errors:

- reject H_0 when it is true
- accept H_0 when it if false

We can evaluate the probabilities of the two types of errors:

- $P(\text{Reject } H_0|H_0) = P(X > 5|H_0) = 0.1445$
- $P(\text{Accept } H_0|H_a) = P(X \leq 5|H_a) = 0.2031$

If we use $c = 1/50$ (which corresponds to prior probability greatly favoring H_a), we accept H_0 if $X \leq 2$. The probabilities of the two types of errors are:

- $P(\text{Reject } H_0|H_0) = P(X > 2|H_0) = 0.8555$
- $P(\text{Accept } H_0|H_a) = P(X \leq 2|H_a) = 0.0012$

3 The Neyman-Pearson Paradigm for Hypothesis Testing

The Neyman-Pearson approach to hypothesis testing does not use prior probabilities for the hypotheses to make a decision, but rather concentrates on the two error probabilities. We start out with two statements concerning the distributions:

- The null hypothesis H_0
- The alternative hypothesis H_a or H_A.

We observe the data and come to one of two possible conclusions:

- Reject H_0
- Fail to reject H_0

The possible results of a hypothesis test are given in the table:

	State of Nature	
Decision	H_0 True	H_a True
Do not reject H_0	Correct	Type II error
Reject H_0	Type I error	Correct

Thus, there are two types of error:

- Type I error: Reject H_0 when H_0 is true
- Type II error: Do not reject H_0 when H_0 is false
- The probability of a type I error is called the level of significance and is denoted by α.
- The probability of a type II error is denoted by $1 - \beta$.
- The probability that H_0 is rejected when it is false is called the power of the test and equals β.

- We used the value of the likelihood ratio to determine whether to reject H_0. We also saw that this was equivalant to using the number of successes X to make our decision. The statistic used to determine whether to reject H_0 is called the test statistic.
- The subset R of the sample space S for which the test statistic leads to rejection of H_0 is called the rejection region. The subset of the sample space where the value of the test statistic leads to failure to reject H_0 (or less properly, "acceptance" of H_0) is called the acceptance region.
- The distribution of the test statistic when $H_0 : \psi(\theta) = \psi_0$ is true is called the null distribution.
- If a rejection region R satisfies $P_\theta(R) \leq \alpha$ whenever $\psi(\theta) = \psi_0$, it is called a *size α rejection region for H_0.*

If a hypothesis is completely specified (i.e., it consists of only one distribution), it is called a simple hypothesis. When a hypothesis consists of more than one distribution, it is called a composite hypothesis.

3.1 Test for the Population Mean for a Normal Distribution

We will construct a test for the population mean μ from a normal population where σ is known. In reality σ is almost never known, but this test is one of the simplest and forms the basis for the ones that follow.

We have a random sample $X_1, \ldots, X_n$ from an $N(\mu, \sigma^2)$ distribution. We want to test $H_0 : \mu = \mu_0$ against $H_a : \mu < \mu_0$

We will base our test statistic upon the estimator of μ, $\bar{X}$. Since small values of $\bar{X}$ would agree with H_a and contradict H_0, the rejection region will have the form $\bar{X} \leq x_0$. We want a rejection region that has a specified Type I error probability, say $\alpha = 0.05$.

Assuming that $\mu = \mu_0$, we want a value x_0 such that

$$P(\bar{X} \leq x_0) = 0.05.$$

When $\mu = \mu_0$,

$$\bar{X} \sim N(\mu_0, \sigma^2/n).$$

Thus,

$$P(\bar{X} \leq x_0) = P\left(Z \leq \frac{x_0 - \mu_0}{\sigma/\sqrt{n}}\right)$$

For this probability to equal α, we need

$$\frac{x_0 - \mu_0}{\sigma/\sqrt{n}} = -Z_{1-\alpha} \quad \text{or} \quad x_0 = \mu_0 - Z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

So the rejection region is

$$\bar{x} \leq \mu_0 - Z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

More commonly, the rejection region is expressed in terms of Z:

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq -Z_{1-\alpha}$$

Power of the Test: We now find the power function by computing the power of the test for a given value of the parameter specified by H_a, $\mu' < \mu_0$:

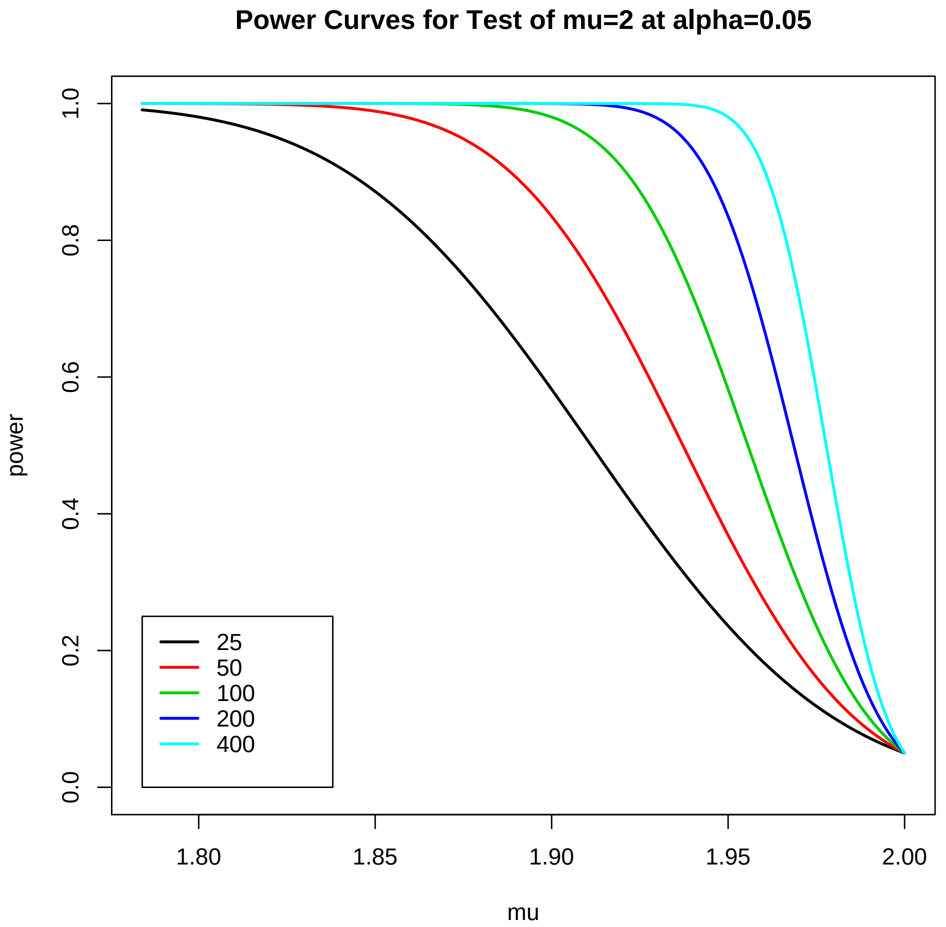
$$
\begin{aligned}
\beta(\mu') &= P(\text{Reject } H_0 \text{ when } \mu = \mu') \\
&= P\left(\bar{X} < \mu_0 - Z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \text{ when } \mu = \mu'\right) \\
&= P_{\mu'}\left(\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} < \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} - Z_{1-\alpha}\right) \\
&= \Phi\left(\frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} - Z_{1-\alpha}\right)
\end{aligned}
$$

Example: To show the effects of sample size on power, we consider testing $H_0 : \mu = 2$ versus $H_a : \mu < 2$ at level $\alpha = .05$. The variance is assumed to be $\sigma^2 = 0.27^2$ with a sample size of $n = 25$. The rejection region of this test is $\bar{X} \leq 1.911$ and the power is

$P[\bar{X} \leq 1.911 \text{ when } \mu = \mu'] = \Phi\left(\frac{1.911-\mu'}{0.27/\sqrt{25}}\right) = \Phi\left(\frac{2-\mu'}{0.27/\sqrt{25}} - 1.645\right)$.

Effect of Changing n on the Power of a Test

Often we would like to have a large power for our test detecting a particular alternative. We consider the power of the test for various sample sizes.



3.2 The Neyman-Pearson Lemma

We present a theorem that the test based on the likelihood ratio is optimal.

Neyman-Pearson Lemma: Suppose that $H_0 : \theta = \theta_0$ and $H_a : \theta = \theta_1$ are simple hypotheses. Consider the test that rejects H_0 whenever the likelihood ratio $f_{\theta_1}/f_{\theta_0}$ is greater than a constant c_0 and suppose that it has size α. Then any other test which has size less than or equal to α has power less than or equal to that of the likelihood ratio test.

Remarks:

- The rejection region of the MP level α test is comprised of values x with large LR. This says that $P[X = x|H_0]$ is small relative to $P[X = x|H_a]$. Thus, such a point x would contribute relatively little to the type I error probability, α, in contrast to its larger contribution to the power.
- The test formed by application of the Neyman-Pearson Lemma is the most powerful size α test of H_0 versus H_a.

Back to Coin Tossing Example Suppose that we have two coins:

- coin 0 with probability of heads equal to 0.5
- coin 1 with probability of heads equal to 0.8

We randomly choose a coin and toss it eight times. I tell you how many heads were obtained and your job is to use this information to determine which coin I tossed. We thus consider testing $H_0 : \theta = 0.5$ versus $H_a : \theta = 0.8$. The likelihoods and likelihood ratio are

x	0	1	2	3	4	5	6	7	8
$p_{0.5}(x)$	0.004	0.031	0.109	0.219	0.273	0.219	0.109	0.031	0.004
$p_{0.8}(x)$	0.000	0.000	0.001	0.009	0.046	0.147	0.294	0.336	0.168
LR	0.0006	0.0026	0.0104	0.042	0.168	0.671	2.684	10.74	42.95

Thus, the test that rejects for $X \geq 7$ is most powerful among all tests with size,
$$P[X \geq 7|H_0] = 0.031 + 0.004 = 0.035.$$

The power of this test is $P[X \geq 7|H_a] = 0.336 + 0.168 = 0.504.$

Example: Let X be a single observation from one of the three following distributions:

x	1	2	3	4	5	6	7
$f_0(x)$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$
$f_1(x)$	$\frac{1}{28}$	$\frac{2}{28}$	$\frac{3}{28}$	$\frac{4}{28}$	$\frac{5}{28}$	$\frac{6}{28}$	$\frac{7}{28}$
$f_2(x)$	$\frac{7}{28}$	$\frac{6}{28}$	$\frac{5}{28}$	$\frac{4}{28}$	$\frac{3}{28}$	$\frac{2}{28}$	$\frac{1}{28}$
$\frac{f_1(x)}{f_0(x)}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{5}{4}$	$\frac{6}{4}$	$\frac{7}{4}$
$\frac{f_2(x)}{f_0(x)}$	$\frac{7}{4}$	$\frac{6}{4}$	$\frac{5}{4}$	1	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

The most powerful size $\alpha = \frac{2}{7}$ test of $H_0 : \theta = 0$ versus $H_a : \theta = 1$ rejects if $LR \geq \frac{6}{4}$ or for $x = 6$ or $x = 7$.

The most powerful level $\alpha = \frac{2}{7}$ test of $H_0 : \theta = 0$ versus $H_a : \theta = 2$. rejects H_0 if $LR \geq \frac{6}{4}$ or for $x = 1$ or $x = 2$.

We see that the form of the most powerful size α test depends on the alternative hypothesis.

Example: Suppose that $X_1, \ldots, X_n$ are independent exponential (λ) rvs. We wish to test $H_0 : \lambda = 5$ versus $H_a : \lambda = 1$. The likelihood ratio is

$$\text{LR} = \frac{\prod_{i=1}^n f(x_i|1)}{\prod_{i=1}^n f(x_i|5)} = \frac{1^n e^{-1\sum x_i}}{5^n e^{-5\sum x_i}} = 5^{-n} e^{4\sum x_i}$$

We see that LR $> c$ is equivalent to $\sum x_i > k$ for some k. Thus, the most powerful test rejects for $\sum x_i > k$.

To find k we need to find the null distribution of $\sum_{i=1}^n X_i$. The mgf of $\sum_{i=1}^n X_i$ is

$$M(t) = \left(\frac{\lambda}{\lambda - t}\right)^n .$$

Then the mgf of $2\lambda \sum_{i=1}^n X_i$ is

$$M_{2\lambda \sum_{i=1}^n X_i}(t) = M(2\lambda t) = \left(\frac{\lambda}{\lambda - 2\lambda t}\right)^n = \left(\frac{\frac{1}{2}}{\frac{1}{2} - t}\right)^n .$$

This is the mgf of a chi-squared rv with $2n$ d.f. Thus, the test will reject for $2\lambda_0 \sum x_i > \chi^2_{1-\alpha}(2n)$.

Remarks:

- The rejection region depends on the null distribution and not on the particular alternative hypothesis. A more realistic test for the nerve data would have the composite alternative hypothesis, $H_a : \lambda < 5$. The same reasoning as above would show that our test is the most powerful size $\alpha = 0.05$ test of $H_0 : \lambda = 5$ versus $H_a : \lambda = \lambda_1$ for any $\lambda_1 < 5$. Thus, this test is the uniformly most powerful size α test of $H_0 : \lambda = 5$ versus $H_a : \lambda < 5$.
- The exponential distribution is an example of an exponential family of distributions. If we can write the pdf or pmf of X in the form

$$f_\theta(x) = \exp[c(\theta)T(x) + d(\theta) + S(x)],$$

the distribution forms an exponential family. A sample from this exponential family has sufficient statistic $\sum_{i=1}^n T(X_i)$. We will see that tests derived by using the likelihood ratio can be expressed in terms of the sufficient statistic.
- For exponential families of distributions, we can find uniformly most powerful tests for testing $H_0 : \theta = \theta_0$ versus $H_a : \theta > \theta_0$ or versus $H_a : \theta < \theta_0$.

Application to Nerve Impulse Data:

We want to test $H_0 : \lambda = 5$ versus $H_a : \lambda < 5$ at level $\alpha = 0.05$. From the earlier example, $\sum x_i = 174.64$ and $n = 799$.

The rejection region is $10 \sum x_i > \chi^2_{0.95}(1598) = 1692.112$.

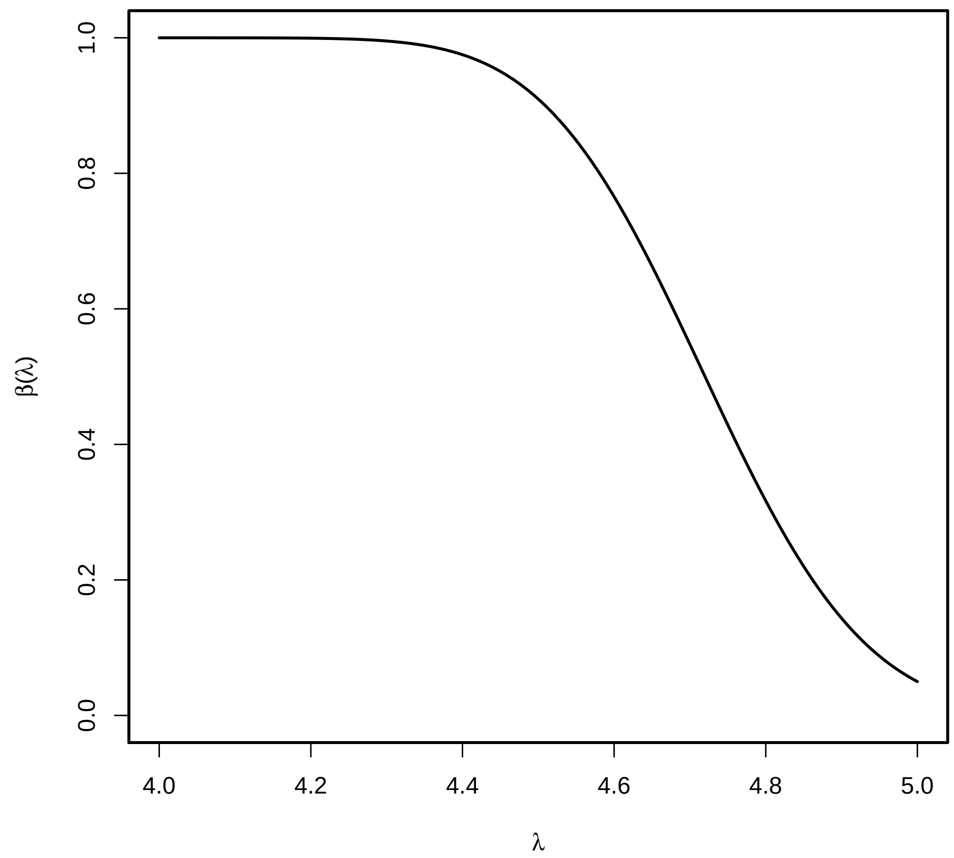
Since $10 \times 174.64 = 1746.4 > 1692.112$, we reject H_0 at level $\alpha = 0.05$.

Power of the Test: The power of the test for $\lambda_1 < 5$ is given by

$$\begin{aligned} \beta(\lambda_1) &= P[\text{Reject } H_0 \text{ when } \lambda = \lambda_1] \\ &= P_{\lambda_1}[2\lambda_0 \sum X_i > \chi^2_{1-\alpha}(2n)] \\ &= P_{\lambda_1}[2\lambda_1 \sum X_i > (\lambda_1/\lambda_0)\chi^2_{1-\alpha}(2n)] \\ &= P[V > (\lambda_1/\lambda_0)\chi^2_{1-\alpha}(2n)] \end{aligned}$$

where V has a chi-squared distribution with $2n$ degrees of freedom. For the nerve impulse example, $n = 799$, $\alpha = 0.05$, and $\lambda_0 = 5$. The plot of the power as a function of λ is below:

Power of Test for Exponential λ



3.3 Specification of Level and $P-$values

To apply the MP test, one needs to specify the level of significance α. This choice is arbitrary, and often is a small value such as 0.01 or 0.05. The power of the test can be used also as a guide in choosing the level. Typically, if one increases α, the power also increases.

Suppose in the coin example, we observed $X = 6$ heads. We would fail to reject H_0 at level $\alpha = 0.035$, but we could reject H_0 at level 0.144 since $P[X \geq 6|H_0] = 0.144$. We could use this quantity, also called a $P-$value, to summarize the evidence in the data against H_0. Thus, the form of the rejection region of the test provides a guide to the region whose probability we define as the $P-$value.

We define the $P-$value $p(x_1, \ldots, x_n)$ corresponding to the observed data $X_1 = x_1, \ldots, X_n = x_n$ as the smallest level of significance at which H_0 can be rejected using a rejection region of the given form. Another definition of $P-$value is the probability of a result at least as extreme as the observed test statistic when H_0 is true.

Nerve Impulse Example:

$P-$value

$= P[10\sum X_i \geq 1746.4|H_0] = P[\chi^2(1598) \geq 1746.4] = 0.0052$

Example of Test for Normal Mean:

Suppose a researcher claims the mean lung capacity of 50-year-old former smokers is less than two liters. We wish to test $H_0 : \mu = 2$ versus $H_a : \mu < 2$. The researcher examines a random sample of 25 fifty-year-old former smokers and measure their lung capacities. Assume that $\sigma = 0.27$ and that the data come from a normal population. For the 25 former smokers, the sample mean lung capacity was $\bar{x} = 1.88$.

$$P-\text{value} = P[\bar{X} \leq 1.88] = P\left(Z \leq \frac{1.88 - 2}{0.27/\sqrt{25}}\right) = \Phi(-2.22) = 0.013$$

Remarks on $P-$values

- We note that the $P-$value $p(x_1, \ldots, x_n)$ is a statistic that is calculated from the observed value of $X_1 = x_1, ..., X_n = x_n$.
- Computer software often provides the $P-$value for a given test. One can use the $P-$value to make a decision in a level α hypothesis test:

$$\text{Reject } H_0 \text{ at level } \alpha \text{ iff the } P - \text{value } \leq \alpha.$$

- Under fairly general conditions when the test statistic has a continuous distribution, one can prove that the distribution of the $P-$value $p(X_1, \ldots, X_n)$ when $H_0 : \theta = \theta_0$ is true is uniform $[0, 1]$.

3.4 Testing the Population Mean for a Normal Distribution

We have a random sample $X_1, \ldots, X_n$ from an $N(\mu, \sigma_0^2)$ distribution. Previously we developed a level α test of the null hypothesis $H_0 : \mu = \mu_0$ versus the one-sided alternative $H_a : \mu < \mu_0$. The test was based upon the point estimator of μ, $\bar{X}$ and rejected H_0 for small values of H_0.

Remarks

1. We could have used the Neyman-Pearson Lemma to show that the above test is the uniformly most powerful level α test of H_0 versus H_a. In a similar fashion, the UMP level α test forH_0 versus $H_a : \mu > \mu_0$ rejects for large values of $\bar{X}$.

2. Suppose that we are interesting in testing $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$. In this case, it can be shown that a UMP level α test does not exist. There are tests that are most powerful among a restricted class of tests (e.g., unbiased tests or invariant tests).

The test that rejects H_0 for $|\bar{X} - \mu_0| > C_\alpha$ can be shown to be UMP unbiased level α. We now find C_α so that this test has size α.

When $H_0 : \mu = \mu_0$,

$$\bar{X} \sim N(\mu_0, \sigma_0^2/n).$$

Thus,

$$P_{\mu_0}(|\bar{X} - \mu_0| \geq C_\alpha) = P_{\mu_0}\left(\frac{|\bar{X} - \mu_0|}{\sigma_0/\sqrt{n}} \geq \frac{C_\alpha}{\sigma_0/\sqrt{n}}\right) = P\left(|Z| \geq \frac{C_\alpha}{\sigma_0/\sqrt{n}}\right).$$

For this probability to equal α, we need

$$\frac{C_\alpha}{\sigma_0/\sqrt{n}} = Z_{1-\alpha/2} \quad \text{or} \quad C_\alpha = Z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}$$

So the rejection region is

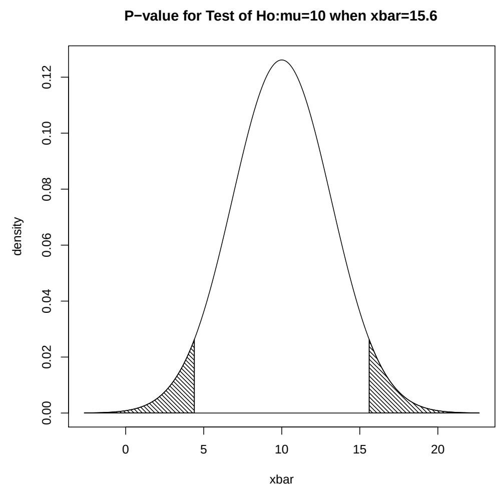
$$\frac{|\bar{x} - \mu_0|}{\sigma_0/\sqrt{n}} \geq Z_{1-\alpha/2}.$$

To obtain the $P-$value when $\bar{X} = \bar{x}$ is observed, we find the probability of a more extreme value of the test statistic $|\bar{X} - \mu_0|$ than the observed value $|\bar{x} - \mu_0|$ assuming $\mu = \mu_0$. Since

$$\bar{X} \sim N(\mu_0, \sigma^2/n),$$

the $P-$value is

$$\begin{aligned} P_{\mu_0}(|\bar{X} - \mu_0| \geq |\bar{x} - \mu_0|) &= P_{\mu_0}\left(\frac{|\bar{X} - \mu_0|}{\sigma_0/\sqrt{n}} \geq \frac{|\bar{x} - \mu_0|}{\sigma_0/\sqrt{n}}\right) \\ &= 2\left[1 - \Phi\left(\frac{|\bar{x} - \mu_0|}{\sigma_0/\sqrt{n}}\right)\right]. \end{aligned}$$

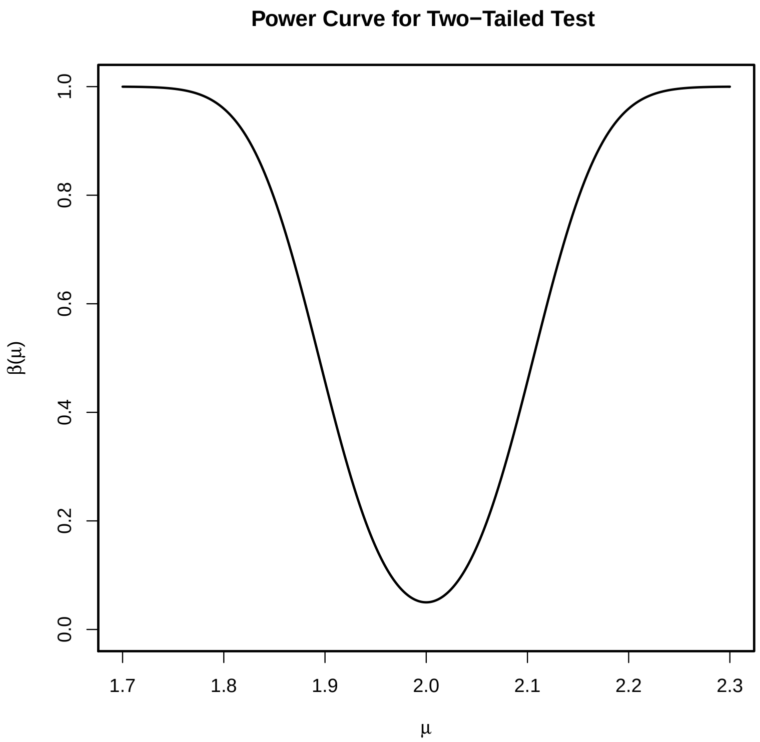


Power of the Two-Tailed Test: We now find the power function by computing the power of the test for a given value of the parameter specified by H_a, $\mu' \neq \mu_0$:

$$\begin{aligned}
\beta(\mu') &= P(\text{Reject } H_0 \text{ when } \mu = \mu') \\
&= P\left(\bar{X} < \mu_0 - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ when } \mu = \mu'\right) \\
&\quad + P\left(\bar{X} > \mu_0 + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ when } \mu = \mu'\right) \\
&= P_{\mu'}\left(\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} < \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} - Z_{1-\alpha/2}\right) \\
&\quad + P_{\mu'}\left(\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} > \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} + Z_{1-\alpha/2}\right) \\
&= \Phi\left(\frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} - Z_{1-\alpha/2}\right) + 1 - \Phi\left(\frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} + Z_{1-\alpha/2}\right)
\end{aligned}$$

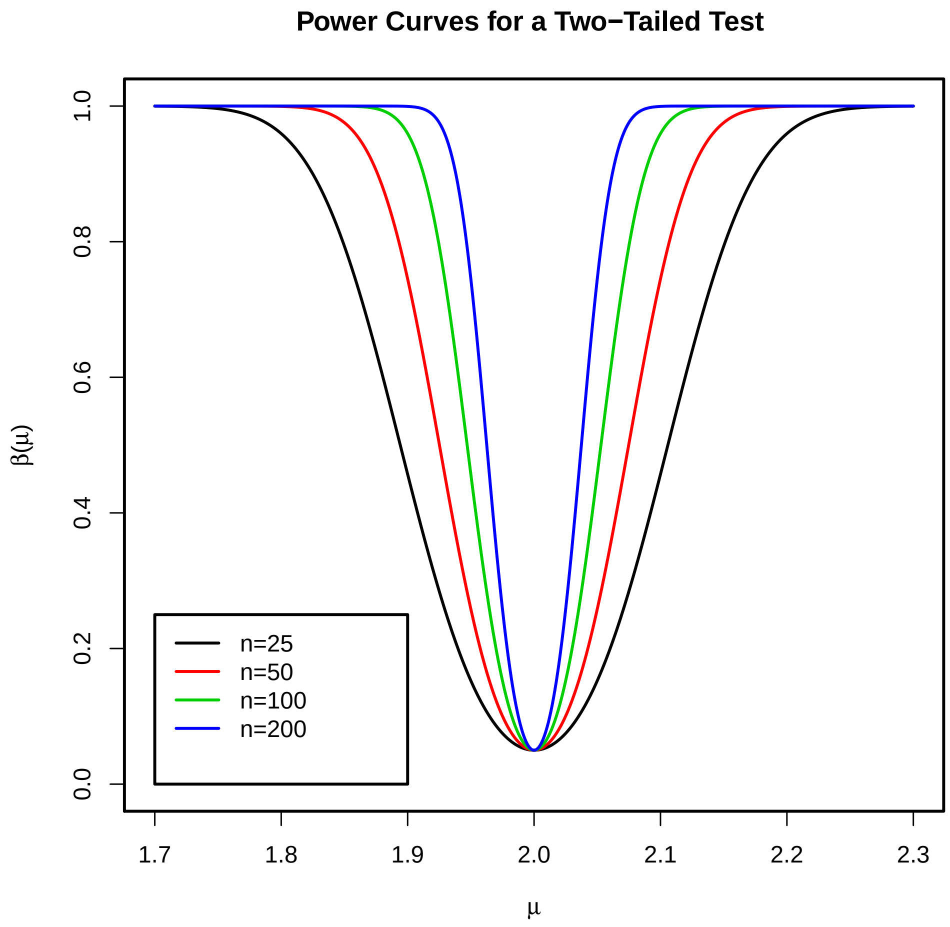
Example: Consider testing $H_0 : \mu = 2$ versus $H_a : \mu \neq 2$ at level $\alpha = .05$. The variance is assumed to be $\sigma^2 = 0.27^2$ with a sample size of $n = 25$. The rejection region of this test is $\bar{X} \leq 1.894$ or $\bar{X} \geq 2.106$ and the power function is

$$P_{\mu'}[\bar{X} \leq 1.894 \text{ or } \bar{X} \geq 2.106] = \Phi\left(\frac{1.894 - \mu'}{0.27/\sqrt{25}}\right) + 1 - \Phi\left(\frac{2.106 - \mu'}{0.27/\sqrt{25}}\right).$$



Effect of Changing n on the Power of a Two-Tailed Test

Often we would like to have a large power for our test detecting a particular alternative. We consider the power of the test for various sample sizes.



3.5 The Duality of Confidence Intervals and Hypothesis Tests

In testing of the hypotheses $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ for a random sample from a normal distribution with unknown mean μ and known variance σ_0^2, we found that the level α test rejects H_0

$$\frac{|\bar{x} - \mu_0|}{\sigma_0/\sqrt{n}} > Z_{1-\alpha/2}$$

Thus, the null hypothesis H_0 is "accepted" (actually is not rejected) when

$$\frac{|\bar{x} - \mu_0|}{\sigma_0/\sqrt{n}} < Z_{1-\alpha/2}.$$

We now carry out some algebra to related this level α two-tailed test to a level γ confidence interval for μ where $\gamma = 1 - \alpha$.

We rewrite this inequality to see which values of μ_0 would be "accepted" by the level α test of $H_0 : \mu = \mu_0$.

$$\frac{|\bar{x} - \mu_0|}{\sigma_0/\sqrt{n}} \quad < \quad Z_{1-\alpha/2}$$

or

$$-Z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}} \quad < \quad \bar{x} - \mu_0 \ < \ Z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}$$

or

$$\bar{x} - Z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}} \quad < \quad \mu_0 \ < \ \bar{x} + Z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}$$

Earlier we saw that a level γ confidence interval for a normal mean μ with known variance σ_0^2 is given by

$$\bar{x} - Z_{(1+\gamma)/2}\frac{\sigma_0}{\sqrt{n}} < \mu \ < \ \bar{x} + Z_{(1+\gamma)/2}\frac{\sigma_0}{\sqrt{n}}.$$

Thus, the level γ confidence interval for μ consists of those values of μ_0 for which the hypothesis $H_0 : \mu = \mu_0$ is not rejected at level $\alpha = 1 - \gamma$.

4 Generalized Likelihood Ratio Tests

Suppose that we have a random samples $X_1, \ldots, X_n$ from a $N(\mu, \sigma_0^2)$ where σ_0^2 is known. We again consider testing the hypotheses:

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_a : \mu \neq \mu_0.$$

If we were interested in a specific alternative hypothesis, say $\mu = \mu_1$, the Neyman-Pearson Lemma implies that we should use the likelihood ratio as our test statistic:

$$\text{LR} = \frac{\prod_{i=1}^{n} f_{\mu_1}(x_i)}{\prod_{i=1}^{n} f_{\mu_0}(x_i)}$$

$$
\begin{aligned}
\text{LR} &= \frac{\prod_{i=1}^n f_{\mu_1}(x_i)}{\prod_{i=1}^n f_{\mu_0}(x_i)} = \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \exp\left[-\frac{(x_i-\mu_1)^2}{2\sigma_0^2}\right]}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \exp\left[-\frac{(x_i-\mu_0)^2}{2\sigma_0^2}\right]} \\
&= \exp\left(-\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^n (x_i-\mu_1)^2 - \sum_{i=1}^n (x_i-\mu_0)^2\right]\right) \\
&= \exp\left(\frac{(\mu_1-\mu_0)}{\sigma_0^2}\sum_{i=1}^n x_i - \frac{n(\mu_1^2-\mu_0^2)}{2\sigma_0^2}\right)
\end{aligned}
$$

- For $\mu_1 > \mu_0$, LR $> c$ is equivalent to $\sum_{i=1}^n x_i > c$.
- For $\mu_1 < \mu_0$, LR $> c$ is equivalent to $\sum_{i=1}^n x_i < c$.

Thus, we cannot form a uniformly most powerful level α test for $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$. We note that the test we found on slide 22 is uniformly most powerful for testing $H_0 : \mu = 2$ versus $H_a : \mu < 2$. We will now use the likelihood ratio to form a test that has good power characteristics, but will not be uniformly most powerful.

Since we do not have a specific value of μ_1, we choose the value of μ_1 that maximizes the likelihood under the alternative. This value will be $\hat{\mu}_1 = \bar{x}$. We substitute this into the LR statistic:

$$\begin{aligned}\text{LR} &= \exp\left(-\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^n (x_i - \bar{x})^2 - \sum_{i=1}^n (x_i - \mu_0)^2\right]\right) \\ &= \exp\left(\frac{n}{2\sigma_0^2}(\bar{x} - \mu_0)^2\right)\end{aligned}$$

The last equality holds since

$$\sum_{i=1}^n (x_i - \mu_0)^2 = \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu_0)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2.$$

Our generalized LR test would reject for large values of LR or equivalently, large values of

$$2 \log \text{LR} = \frac{n(\bar{x} - \mu_0)^2}{\sigma_0^2}$$

We next need to determine the rejection region for the test. To do this we need to obtain the null distribution for the test statistic.

When $H_0 : \mu = \mu_0$ is true, $\bar{X} \sim N(\mu_0, \sigma_0^2/n)$. Thus,

$$\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma_0} \sim N(0,1) \qquad \text{and} \qquad \frac{n(\bar{X} - \mu_0)^2}{\sigma_0^2} \sim \chi^2(1)$$

Thus, our generalized LR test rejects for

$$\frac{n(\bar{x} - \mu_0)^2}{\sigma_0^2} > \chi^2_{1-\alpha}(1)$$

Remark: We constructed a generalized LR test for a normal mean using the following steps:

- Write down the LR statistic for testing two simple hypotheses.
- Substitute in the MLE for the mean under the alternative.
- Rewrite the test statistic to obtain a new test statistic with known distribution.
- Find the rejection region of the test.

We now outline our approach to generalized likelihood ratio tests. We suppose that $X_1, \ldots, X_n$ form a random sample from a distribution with pdf or pmf $f_\theta(x)$. We wish to test the hypotheses

$$H_0 : \psi(\theta) = \psi_0 \qquad \text{versus} \qquad H_a : \psi(\theta) \neq \psi_0.$$

To determine the plausibility of the two hypotheses, we will compare the largest likelihood to the largest likelihood under the null hypothesis using the generalized LR statistic:

$$\text{LR} = \frac{L(\hat{\theta}|x_1, \ldots, x_n)}{L(\hat{\theta}_{H_o}|x_1, \ldots, x_n)}$$

We see that large values of LR discredit H_0. Thus, we need to find a threshhold c_0 so that $P[\text{LR} \geq c_0|H_0] = \alpha$. As seen in the example, we will usually rewrite the LR statistic in terms of another statistic with known distribution.

Example: Let $X_1, \ldots, X_n$ be iid Poisson (λ) rvs. Form a LR test of $H_0 : \lambda = \lambda_0$ versus $H_a : \lambda \neq \lambda_0$.

The joint likelihood is

$$p(x_1, \ldots, x_n | \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \frac{e^{-\lambda n} \lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

The mle of λ is $\hat{\lambda} = \bar{x}$. Thus, the generalized LR statistic is

$$\text{LR} = \frac{\dfrac{e^{-n\hat{\lambda}} \hat{\lambda}^{\sum x_i}}{\prod x_i!}}{\dfrac{e^{-\lambda_0 n} \lambda_0^{\sum x_i}}{\prod x_i!}} = \frac{e^{-n\hat{\lambda}} \hat{\lambda}^{\sum x_i}}{e^{-\lambda_0 n} \lambda_0^{\sum x_i}} = e^{-n(\hat{\lambda} - \lambda_0)} \left(\frac{\hat{\lambda}}{\lambda_0} \right)^{n\hat{\lambda}}$$

The null distribution of LR or of $2 \log(\text{LR})$ is a complicated discrete distribution. We will next develop an approximation to the null distribution of the LR statistic that is often useful.

4.1 The Asymptotic Distribution of the LR Statistic

In Section 3.6 of Chapter 6 we saw the mle under certain regularity conditions had an asymptotically normal distribution. We can approximate 2 times the LR statistic using a quadratic function of the mle. Using this approximation, one can show that 2 times the log of the LR statistic has approximately a chi-squared distribution.

Theorem Under the conditions for the asymptotic normality of the mle (see slide 60 of Chapter 6), the null distribution of $2\log$ LR converges to a chi-squared distribution with $df = \dim(\Omega) - \dim(H_0)$ as n tends to infinity.

Remark: This theorem is very general and can be applied to many models useful in applications. We will look at several examples where the null hypothesis is completely specified ($\dim(H_0) = 0$).

Example: Consider the test statistic for $H_0 : \lambda = \lambda_0$ versus $H_a : \lambda \neq \lambda_0$ for a sample from the Poisson (λ) distribution. We found that the LR statistic is

$$\mathsf{LR} = e^{-n(\hat{\lambda}-\lambda_0)} \left(\frac{\hat{\lambda}}{\lambda_0} \right)^{n\hat{\lambda}}$$

The theorem implies that the following statistic has approximately a $\chi^2(1)$ distribution:

$$2 \log \mathsf{LR} = -2n(\hat{\lambda} - \lambda_0) + 2n\hat{\lambda} \log \left(\frac{\hat{\lambda}}{\lambda_0} \right).$$

The level α LR test has rejection region

$$2 \log \mathsf{LR} > \chi^2_{1-\alpha}(1).$$

4.2 Application to One-Parameter Problems

Suppose that $X_1, \ldots, X_n$ is a random sample from a distribution with pdf or pmf $f_\theta(x)$ where $\theta \in \Omega$ is a single parameter.

Note: More generally, we could consider $X_1, \ldots, X_n$ having certain joint pdfs or pmfs of the form $f_\theta(x_1, \ldots, x_n)$.

We consider testing $H_0 : \theta = \theta_0$.

There are three likelihood-based approaches to hypothesis testing:

- Generalized likelihood ratio test
- Wald test
- Score test

1. **Generalized Likelihood Ratio Test**

We wish to compare the likelihood under H_0, $L(\theta_0|x_1,...,x_n)$ to the largest likelihood, $L(\hat{\theta}|x_1,...,x_n)$, using the *likelihood ratio statistic*:

$$G^2 = 2\log \text{LR} = 2\ \log\left[\frac{L(\hat{\theta}|x_1,...,x_n)}{L(\theta_0|x_1,...,x_n)}\right] \xrightarrow{D} \chi^2(1) \text{ as } n \longrightarrow \infty.$$

We can also write
$$G^2 = 2\left[\log[L(\hat{\theta}|x_1,...,x_n)] - \log[L(\theta_0|x_1,...,x_n)]\right].$$

- Now $L(\theta) \le L(\hat{\theta})$ for all $\theta \in \Omega$, so $G^2 > 0$.
- When H_0 is true, we would expect $\hat{\theta}$ to be close to θ_0 and the ratio inside G^2 to be close to 1.
- When H_0 is false, the value of $\hat{\theta}$ would differ from θ_0 and $L(\theta_0) < L(\hat{\theta})$. We reject H_0 for large values of G^2.

Example: $Y \sim$ Binomial (n, θ)

Consider testing $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$.

Then

$$L(\theta|y) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

We earlier derived the mle, $\hat{\theta} = \frac{Y}{n}$.

We base the test on the statistic

$$\begin{aligned} G^2 &= 2\left[\log[L(\hat{\theta}|y)] - \log[L(\theta_0)]\right] \\ &= 2[y\log(\hat{\theta}) + (n-y)\log(1-\hat{\theta}) \\ &\quad -y\log(\theta_0) - (n-y)\log(1-\theta_0)] \\ &= 2\left[y\log\left(\frac{\hat{\theta}}{\theta_0}\right) + (n-y)\log\left(\frac{1-\hat{\theta}}{1-\theta_0}\right)\right] \end{aligned}$$

We reject H_0 for $G^2 > \chi^2_{1-\alpha}(1)$.

2. **Wald Test**

The Wald test is based on the asymptotic normality of the mle, $\hat{\theta}$:

$$\frac{\hat{\theta} - \theta_0}{\sqrt{I_n(\theta_0)^{-1}}} \xrightarrow{D} N(0,1) \quad \text{as} \quad n \longrightarrow \infty$$

We define the *Wald statistic* by substituting $\hat{\theta}$ into $I_n(\theta)$:

$$Z = \frac{\hat{\theta} - \theta_0}{\sqrt{I_n(\hat{\theta})^{-1}}} \sim N(0,1) \qquad \text{or} \qquad W = Z^2 = \frac{(\hat{\theta} - \theta_0)^2}{I_n(\hat{\theta})^{-1}} \sim \chi^2(1)$$

Example: Binomial (n, θ)

$$Z = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{\hat{\theta}(1 - \hat{\theta})}} \quad \text{or} \quad W = \frac{n(\hat{\theta} - \theta_0)^2}{\hat{\theta}(1 - \hat{\theta})}$$

3. **Score Test**

The score function is defined as

$$U(\theta) = \frac{\partial \log(L(\theta|x_1, \ldots, x_n))}{\partial \theta}.$$

Recall that the mle is the solution to

$$U(\theta) = \frac{\partial \log(L(\theta|x_1, \ldots, x_n))}{\partial \theta} = 0.$$

We evaluate the score function at the hypothesized value θ_0 and see how close it is to zero.

When H_0 is true, the mean and variance of the score are

$$E_{\theta_0}(U(\theta_0)) = 0 \qquad \text{and} \qquad \mathrm{Var}_{\theta_0}(U(\theta_0)) = I_n(\theta_0).$$

The score statistic found by standardizing the score function is asymptotically normal:

$$Z = \frac{U(\theta_0)}{\sqrt{I_n(\theta_0)}} \sim N(0,1) \qquad \text{or} \qquad S = Z^2 = \frac{U(\theta_0)^2}{I_n(\theta_0)} \sim \chi^2(1).$$

Example: Bernoulli random sample

$$U(\theta) = \frac{\partial \log(L(\theta|x_1,\ldots,x_n))}{\partial\theta} = \frac{y}{\theta} - \frac{n-y}{1-\theta}$$

$$S = \frac{\left(\frac{y}{\theta_0} - \frac{n-y}{1-\theta_0}\right)^2}{\frac{n}{\theta_0(1-\theta_0)}} = \frac{n(\hat\theta - \theta_0)^2}{\theta_0(1-\theta_0)}$$

Remark: We note that the score statistic is equivalent to the Z^2 statistic obtain by substituting θ_0 into $I_n(\theta)$.

Comments

- The above tests all reject for large values of the test statistic based on chi-squared critical values.
- The three tests are asymptotically equivalent. That is, in large samples they will tend to have similar values and lead to the same decision.
- For moderate sample sizes, the LR test is usually more reliable than the Wald test.
- A large difference in the values of the three statistics may indicate that the distribution of $\hat{\theta}$ may not be normal.
- The Wald test is based on the behavior of the log-likelihood at the mle $\hat{\theta}$. The ASE of $\hat{\theta}$ depends on the curvature of the log-likelihood function at $\hat{\theta}$.
- The score test is based on the behavior of the log-likelihood function at θ_0. It uses the derivative (or slope) of the log-likelihood at the null value, θ_0. Recall that the slope at $\hat{\theta}$ equals zero.

- The LR statistic combines information about the log-likelihood function both at $\hat{\theta}$ and at θ_0. Thus, the LR statistic uses more information than the other two statistics and is usually the most reliable among the three.

- These statistics can be used for multiparameter models. Often we have a parameter vector $(\theta, \beta_1, \ldots, \beta_p)$. We wish to test $H_0 : \theta = \theta_0$. The following are the differences that hold for this model:
  - The score function is now a vector of $p + 1$ partial derivatives of the log-likelihood function.
  - The MLE is determined by solving the resulting set of $p + 1$ equations in $p + 1$ unknowns.
  - Fisher's information is now a $(p + 1) \times (p + 1)$ matrix.
  - All three statistics are asymptotically equivalent and asymptotically have a chi-squared distribution with 1 d.f.

4.3 Forming Confidence Intervals from LR Tests

Let's return to the random sample from a normal distribution with unknown mean μ and known variance σ_0^2. Consider testing the hypotheses:

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_a : \mu \neq \mu_0$$

We found that the generalized likelihood ratio test rejects $H_0 : \mu = \mu_0$ when

$$\frac{n(\bar{x} - \mu_0)^2}{\sigma_0^2} > \chi^2_{1-\alpha}(1)$$

Thus, the null hypothesis H_0 is "accepted" when

$$\frac{n(\bar{x} - \mu_0)^2}{\sigma_0^2} \quad < \quad \chi^2_{1-\alpha}(1)$$

We now rewrite this inequality to see which values of μ_0 would be "accepted" by the level α test.

$$\frac{n(\bar{x}-\mu_0)^2}{\sigma_0^2} \quad < \quad \chi^2_{1-\alpha}(1)$$

or

$$(\bar{x}-\mu_0)^2 \quad < \quad \frac{\chi^2_{1-\alpha}(1)\sigma_0^2}{n}$$

or

$$-z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}} \quad < \quad \bar{x}-\mu_0 \; < \; z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}$$

or

$$\bar{x}-z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}} \quad < \quad \mu_0 \; < \; \bar{x}+z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}$$

A level $1-\alpha$ confidence interval for a normal mean μ with known variance σ_0 is given by

$$\bar{x}-z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}} < \mu \; < \; \bar{x}+z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}.$$

Thus, the level $1 - \alpha$ confidence interval for μ consists of those values of μ_0 for which the hypothesis $H_0 : \mu = \mu_0$ is accepted.

This duality holds more generally and provides a method for forming confidence intervals from hypothesis tests. Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ come from a distribution with parameter θ with parameter space Θ. We present two theorems that summarize the duality between tests and confidence intervals:

Theorem A Let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. Then the set

$$C(\boldsymbol{X}) = \{\theta : \boldsymbol{X} \in A(\theta)\}.$$

forms a level $1 - \alpha$ confidence interval for θ.

Remark: This method of forming a confidence interval is called *inverting a test.*

Theorem B Suppose that $C(\boldsymbol{X})$ is a level $1-\alpha$ confidence interval for θ; that is, for every θ_0,

$$P[\theta_0 \in C(\boldsymbol{X})|\theta = \theta_0] = 1-\alpha.$$

Then an acceptance region for a level α test of $H_0 : \theta = \theta_0$ is given by

$$A(\theta_0) = \{\boldsymbol{X} : \theta_0 \in C(\boldsymbol{X})\}.$$

Remark: We can use the first theorem to form approximate confidence intervals for θ based on the large sample distribution of the likelihood ratio statistic.

Example: Form approximate level $1-\alpha$ confidence intervals for the binomial probability of success θ based on LR, Wald, and score tests.

- LR interval: The acceptance region of the approximately level α LR test for $H_0 : \theta = \theta_0$ is given by

$$A(\theta_0) = \left\{ y : 2\left[y \log\left(\frac{\hat{\theta}}{\theta_0}\right) + (n-y)\log\left(\frac{1-\hat{\theta}}{1-\theta_0}\right)\right] \le \chi^2_{1-\alpha}(1) \right\}$$

Hence, the approximately level $1-\alpha$ confidence interval for θ is given by

$$C(y) = \left\{ \theta_0 : 2\left[y \log\left(\frac{\hat{\theta}}{\theta_0}\right) + (n-y)\log\left(\frac{1-\hat{\theta}}{1-\theta_0}\right)\right] \le \chi^2_{1-\alpha}(1) \right\}$$

For the Bill of Rights example, $n = 50$ and $y = 14$. The approximate level 0.95 LR confidence interval for θ is $(0.169, 0.413)$.

- Wald interval

The Wald test for $H_0 : \theta = \theta_0$ has acceptance region

$$|Z| = \frac{\sqrt{n}\,|\hat{\theta} - \theta_0|}{\sqrt{\hat{\theta}(1 - \hat{\theta})}} < Z_{1-\alpha/2}$$

We invert the test to obtain the approximate level $1 - \alpha$ confidence interval for θ:

$$\left(\hat{\theta} - Z_{1-\alpha/2}\sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}, \hat{\theta} + Z_{1-\alpha/2}\sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}\right)$$

For the Bill of Rights example, $n = 50$ and $y = 14$. The approximate level 0.95 Wald confidence interval for θ is $(0.156, 0.404)$.

- Score interval

  The approximate level α score test for $H_0 : \theta = \theta_0$ has acceptance region

$$S = \frac{\left(\frac{y}{\theta_0} - \frac{n-y}{1-\theta_0}\right)^2}{\frac{n}{\theta_0(1-\theta_0)}} = \frac{n(\hat{\theta} - \theta_0)^2}{\theta_0(1-\theta_0)} < Z^2_{1-\alpha/2}.$$

  Thus, the approximate level $1 - \alpha$ confidence interval for θ is given by

$$C(y) = \left\{\theta_0 : \frac{n(\hat{\theta} - \theta_0)^2}{\theta_0(1-\theta_0)} < Z^2_{1-\alpha/2}\right\}.$$

  This confidence interval has endpoints

$$\frac{\hat{\theta} + \frac{Z^2_{1-\alpha/2}}{2n} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n} + \frac{Z^2_{1-\alpha/2}}{4n^2}}}{1 + Z^2_{1-\alpha/2}/n}$$

  For the Bill of Rights data, the score confidence interval is $(0.175, 0.417)$.

