4.5 Confidence Intervals, Hypothesis Tests, and Significance Tests

The use of the symbol α in the development of both confidence intervals and hypothesis testing is not a coincidence.

In confidence intervals, the value of α determines the value of the stated coverage of the confidence interval, which is $1-\alpha$. If the confidence interval $L < \theta < U$ is an exact confidence interval, for instance, then $P(L < \theta < U) = 1-\alpha$.

In hypothesis testing, the value of the significance level α is the probability of committing a Type I error, that is, $\alpha = P(\text{Type I error}) = P(\text{rejecting } H_0 \,|\, H_0 \text{ true})$.

This subsection links the α that is used in constructing a confidence interval with the α used in hypothesis testing. Assume that the goal associated with the analysis of a data set is to conduct a test. As seen in Table 4.5, such a test associated with the null hypothesis $H_0 : \theta = \theta_0$ can be conducted in three different ways: using confidence intervals, using traditional hypothesis testing, and using significance testing. Each column in Table 4.5 is associated with one of the three methods of conducting the test. The rows in Table 4.5 show the sequential steps associated with the three methods. The first three steps in each of the three methods are identical. The methods differ in the last two steps. Regardless of the methods taken, the conclusion that is drawn, either reject H_0 or fail to reject H_0, will be identical for the same data set which is based on the same pivotal value/test statistic.

Confidence interval method	**Hypothesis testing method**	**Significance testing method**
Formulate H_0 and H_1	Formulate H_0 and H_1	Formulate H_0 and H_1
Determine significance level α	Determine significance level α	Determine significance level α
Determine sample size n	Determine sample size n	Determine sample size n
Construct confidence interval	Compute test statistic	Compute test statistic
Reject H_0 if θ_0 does not fall in the confidence interval	Reject H_0 if the test statistic is in the critical region	Reject H_0 if the p-value is less than α

Table 4.5: Steps associated with confidence intervals, hypothesis tests, and significance tests.

The next example illustrates these three methods for conducting a test based on a single data set. The three methods all result in the same conclusion being drawn.

Example 4.21 IQ scores are traditionally centered around a population mean of 100. An experimenter would like to test the simple null hypothesis

$$H_0 : \mu = 100$$

versus the two-sided alternative hypothesis

$$H_1 : \mu \neq 100$$

based on a random sample of IQ scores that is assumed to come from a normally distributed population. The significance level is set to $\alpha = 0.01$ prior to collecting the data. It is also decided that a sample of $n = 78$ IQ scores will be collected in order to conduct the test. More specifically, the $n = 78$ IQ scores of seventh-grade students at a rural school in the midwestern United States from Example 3.9, and also listed below, will be used to conduct the test. This completes the first three common steps for the three methods given in Table 4.5.

111	107	100	107	114	115	111	97	100	112	104	89	104
102	91	114	114	103	106	105	113	109	108	113	130	128
128	118	113	120	132	111	124	127	128	136	106	118	119
123	124	126	116	127	119	97	86	102	110	120	103	115
93	72	111	103	123	79	119	110	110	107	74	105	112
105	110	107	103	77	98	90	96	112	112	114	93	106

The test will now be completed using each of the three methods. To illustrate the use of a confidence interval to complete the test, begin by computing the sample mean and sample standard deviation of the data values, which are approximately

$$\bar{x} = 109 \qquad \text{and} \qquad s = 13.$$

Using Table 3.1, an exact two-sided 99% confidence interval for the population mean IQ score μ is

$$\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

or

$$109 - 2.6412\frac{13}{\sqrt{78}} < \mu < 109 + 2.6412\frac{13}{\sqrt{78}}$$

or

$$105 < \mu < 113.$$

Since this 99% confidence interval does not contain $\mu_0 = 100$, we reject H_0 and conclude that these IQ scores are not drawn from a population centered around 100.

Shifting to traditional hypothesis testing to complete the test, the appropriate test statistic from Table 4.4 is

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - 100}{S/\sqrt{n}},$$

which has the $t(n-1) = t(77)$ distribution under H_0. For the $n = 78$ IQ scores, the value of this test statistic is

$$\frac{109 - 100}{13/\sqrt{78}} = 5.98.$$

The critical values for the test are the 0.5th and 99.5th percentile of the t distribution with 77 degrees of freedom, which are

$$c_l = -t_{77,0.005} = -2.6412 \qquad \text{and} \qquad c_u = t_{77,0.005} = 2.6412,$$

which are values that appear in the confidence interval formula given earlier. So H_0 is rejected if the test statistic is less than $c_l = -2.6412$ or greater than $c_u = 2.6412$. Since the test statistic 5.98 is much greater than the upper critical value $c_u = 2.6412$, H_0 is overwhelmingly rejected. We again conclude that these IQ scores are not drawn from a population centered around population mean 100.

Finally, using a significance test to conduct the test, the choice of the appropriate test statistic from Table 4.4 remains the same:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - 100}{S/\sqrt{n}}.$$

For the $n = 78$ IQ scores, the value of this test statistic is again

$$\frac{109-100}{13/\sqrt{78}} = 5.98.$$

The p-value for the two-sided significance test is twice the area under the probability density function of a $t(77)$ random variable to the right of 5.98, that is,

$$p = 2P(t(77) > 5.98) = 0.00000006.$$

Since this p-value is much smaller than the significance level $\alpha = 0.01$, we again decisively reject H_0 and conclude that the IQ scores are not drawn from a population that is centered around 100.

As indicated in Table 4.5, all three methods for conducting the test, namely confidence intervals, traditional hypothesis testing, and significance testing, draw the same conclusion.

Both confidence intervals and traditional hypothesis testing address the question of consistency between a population parameter and a sample statistic. In the previous example, the confidence interval gives a range of parameter values, in this case $105 < \mu < 113$, that are consistent with the sample statistic $\bar{x} = 109$. On the other hand, the traditional hypothesis test fixes the parameter according to the null hypothesis, in this case $H_0 : \mu = 100$, and asks which values of the sample statistic are consistent with H_0.

The geometry associated with these two points of view for the $n = 78$ IQ data values and $\alpha = 0.01$ is illustrated in Figure 4.40, which contains a plot of the sample mean $\bar{x}$ on the horizontal axis and the population mean μ on the vertical axis. The equation of the top solid diagonal line is

$$\mu = \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

for $n = 78$, $\alpha = 0.01$, and $s = 13$. The equation of the bottom solid diagonal line is

$$\mu = \bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}.$$

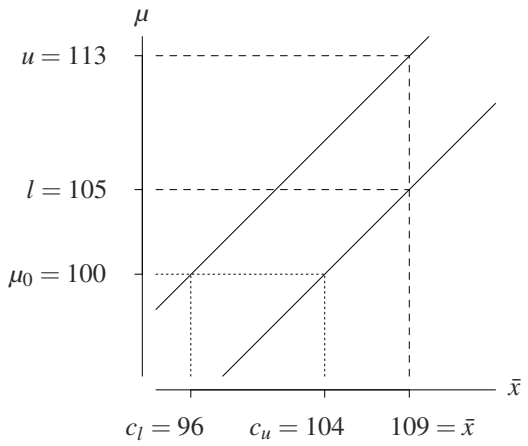


Figure 4.40: Equivalence between confidence intervals and traditional hypothesis testing.

The dashed lines illustrate the construction of a confidence interval. Beginning on the horizontal axis at $\bar{x} = 109$, the vertical dashed line intersects the diagonal solid lines at heights that correspond to the lower confidence limit $l = 105$ and upper confidence limit $u = 113$. The resulting exact two-sided 99% confidence interval for μ is $105 < \mu < 113$. The dotted lines illustrate a traditional hypothesis test. Beginning on the vertical axis at $\mu_0 = 100$ (associated with the null hypothesis $H_0 : \mu = 100$), the horizontal dotted line intersects the diagonal solid lines at the lower critical value $c_l = 96$ and the upper critical value $c_u = 104$. These two critical values are associated with using the test statistic $\bar{x}$. (In the previous example, $(\bar{x} - 100)/(s/\sqrt{n})$ was used as a test statistic, but the geometric interpretation given here uses $\bar{x}$ as the test statistic.) Since the slopes of the two solid lines both equal 1, the confidence interval width $u - l$ equals the width of the region associated with failing to reject H_0, which is $c_u - c_l$, thus establishing the equivalence between the confidence interval and hypothesis testing methods for conducting the test for this particular data set.

The three methods for conducting a test, namely confidence intervals, traditional hypothesis testing, and significance testing, have been shown to be equivalent for one particular test, one particular data set, and one particular choice of α. This is promising, but it does not establish the equivalence of the three methods in general.

The equivalence between traditional hypothesis testing and significance testing should be clear from the development of the two methods given earlier in this section. But the equivalence between using confidence intervals and traditional hypothesis testing to conduct a test is not so obvious.

We illustrate the equivalence for the confidence interval from the previous example for a general value of α and any random sample $X_1, X_2, \ldots, X_n$ assumed to be drawn from a normal population with unknown population mean and variance. The derivation below will show that the confidence interval method for conducting a test is equivalent to traditional hypothesis testing in this slightly more general context. Consider testing

$$H_0 : \mu = \mu_0$$

versus

$$H_1 : \mu \neq \mu_0.$$

Using the confidence interval method to conduct this test, Table 4.5 indicates that we fail to reject H_0 when μ_0 falls in the exact two-sided confidence interval for μ (from Table 3.1), that is, when

$$\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \mu_0 < \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}.$$

In order to show the equivalence of this method with traditional hypothesis testing, some algebra is performed on the inequality given above resulting in

$$-t_{n-1,\alpha/2} < \frac{\bar{X} - \mu_0}{S/\sqrt{n}} < t_{n-1,\alpha/2}.$$

The random variable in the center of this inequality happens to be the appropriate test statistic for the hypothesis test in this particular setting (from Table 4.4), and the values on the left-hand side and right-hand side of the inequality are the associated critical values c_l and c_u. Thus, failing to reject H_0 using the confidence interval method in this setting is equivalent to failing to reject H_0 using traditional hypothesis testing.

We have now shown the duality between hypothesis testing and interval estimation in one particular statistical setting. The good news is that duality generalizes, as established next.