

Unit 22

One-Sided and Two-Sided Hypotheses Tests

Objectives:

- To differentiate between a one-sided hypothesis test and a two-sided hypothesis test about a population proportion or a population mean
- To understand the difference between a statistically significant difference and a clinically significant difference

The first hypothesis test we considered when we introduced this topic concerned a claim by the manufacturer of a lighter. The manufacturer's claim was that the lighter would ignite on the first try 75% of the time. This led to the following hypothesis test:

$$H_0: \lambda = 0.75 \quad \text{vs.} \quad H_1: \lambda \neq 0.75 \quad (\alpha = 0.10).$$

Figure 20-1 graphically displays the rejection region for the z test statistic. Algebraically, this rejection region is defined as

$$z \geq +1.645 \quad \text{or} \quad z \leq -1.645 \quad (\text{i.e., } |z| \geq 1.645).$$

This rejection region is supposed to identify values of the z test statistic which represent a larger distance between \bar{p} and the hypothesized proportion 0.75 than we would expect from sampling error if $H_0: \lambda = 0.75$ were really true. If the value of the z test statistic is greater than +1.645, we would be inclined to believe that the population proportion is larger than the hypothesized 0.75, whereas if the value of the z test statistic is less than -1.645, we would be inclined to believe that the population proportion is smaller than the hypothesized 0.75.

However, after some reflection, you may decide that you really do not care about finding evidence that the population proportion is larger than the hypothesized 0.75. If it were in fact true that the lighter ignited on the first try more than 75% of the time, then you should be very pleased to find that the lighter actually performs better than the manufacturer claims. With this in mind, it is reasonable for you to decide that you are only concerned about finding evidence that the population proportion is smaller than the hypothesized 0.75.

A hypothesis test which is designed to identify a difference from a hypothesized value in only one direction is called a *one-sided test*. A hypothesis test which is designed to identify a difference from a hypothesized value in either direction is called a *two-sided test*. All of the hypothesis tests we have considered up to this point have been two-sided tests. We shall now consider some illustrations where a one-sided test is called for. The procedure to perform a one-sided test is exactly the same as the procedure to perform a two-sided test except for two things: first, the rejection region will consist either of positive values only or of negative values only, but not both; second, the p -value will be calculated by considering only one direction.

To illustrate a one-sided test, suppose a substance called PatchUp is designed to repair and inflate a flat bicycle tire when one tube of the substance is squeezed into the tire through the stem. The manufacturer claims that PatchUp will be completely successful 85% of the time. A bicycle club would like to see if there is any evidence that the percentage of times PatchUp is completely successful is less than the claimed 85%. (Since it would be wonderful if the percentage were actually higher than the claimed 85%, the bicycle club has no interest in looking for evidence in this direction.) A 0.10 significance level is chosen for a hypothesis test. In a simple random sample of 152 flat tires, PatchUp is found to be completely successful 123 times.

Since 0.85 is the hypothesized value, and $n = 152$ is the sample size, we can verify that the sample size is sufficiently large for the z -test about a population proportion to be appropriate by noting that $(152)(0.85) = 129.2$ and $(152)(1 - 0.85) = 22.8$ are each greater than five. We shall now perform the four steps of the hypothesis test.

The first step of our hypothesis test is to state the null and alternative hypotheses, and to choose a significance level. Remember that the null hypothesis is what we assume to be true unless there is sufficient evidence against it, and also that our null hypothesis will be a statement involving equality; in addition, the alternative hypothesis is a statement we are looking for evidence to support, and also the alternative hypothesis

is a statement involving inequality. Thinking of our hypothesis as similar to a court trial, we shall assume that the proportion of times PatchUp is completely successful is (at least) 0.85, unless we find sufficient evidence that the proportion is smaller; in other words, 0.85 is our hypothesized value for λ . If we decide on a 0.10 significance level, we can complete the first step of the hypotheses test as follows:

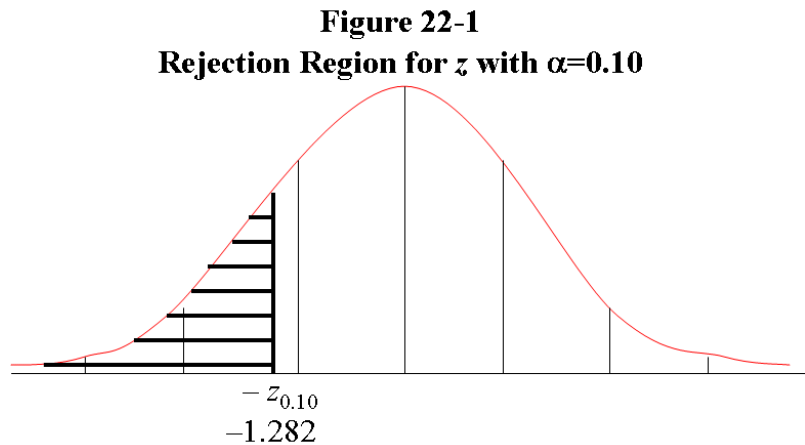
$$H_0: \lambda = 0.85 \quad \text{vs.} \quad H_1: \lambda \leq 0.85 \quad (\alpha = 0.10, \text{ one-sided test}) .$$

It would be perfectly correct to state the null hypothesis as $H_0: \lambda \geq 0.85$, because by not rejecting H_0 we are not only accepting the manufacturer's claim that PatchUp is completely successful 85% of the time, but we are also accepting the possibility that the percentage may be higher than the manufacturer claims. However, since we understand this to be implicit, we have decided to simply state the null hypothesis as $H_0: \lambda = 0.85$. Together with the significance level stated in parentheses, we have included a reminder that the test is one-sided.

The second step of the hypothesis test is to collect data and calculate the value of the test statistic. Since PatchUp was found to be completely successful 123 times out of 152 tries, we find that the sample proportion is $\bar{p} = 123/152 = 0.8092$; we then calculate the value of our test statistic z as follows:

$$z = \frac{\bar{p} - \lambda_0}{\sqrt{\frac{\lambda_0(1 - \lambda_0)}{n}}} = \frac{0.8092 - 0.85}{\sqrt{\frac{0.85(1 - 0.85)}{152}}} = -1.408 .$$

The third step of the hypothesis test is to define the rejection region, decide whether or not to reject the null hypothesis, and obtain the p -value of the test. Figure 22-1 displays the rejection region graphically. The shaded area corresponds to values of the test statistic z where \bar{p} is less than the hypothesized 0.85 and which are unlikely to occur if $H_0: \lambda = 0.85$ is true. In order that this shaded area be equal to $\alpha = 0.10$, the rejection region is defined by the z -score below which 0.10 of the area lies.

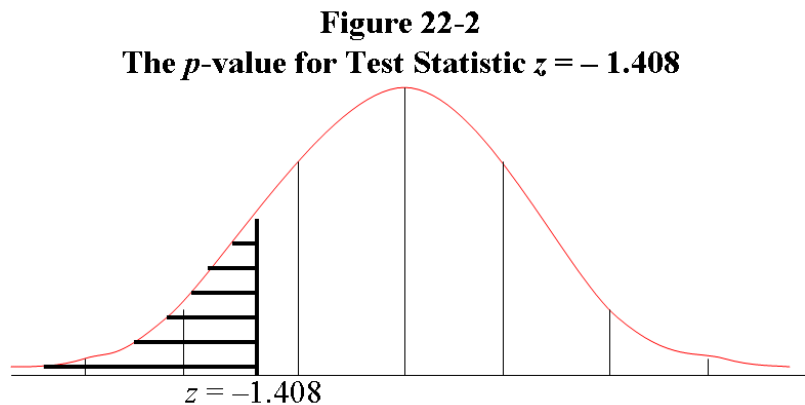


From either the bottom of Table A.2 or from the last row in Table A.3, we find that the z -score above which 0.10 of the area lies is $z_{0.10} = 1.282$, which implies that the z -score below which 0.10 of the area lies is -1.282 . We can then define our rejection region algebraically as

$$z \leq -1.282 .$$

Since our test statistic, calculated in the second step, was found to be $z = -1.408$ which is in the rejection region, our decision is to reject $H_0: \lambda = 0.85$; in other words, our data provides sufficient evidence to suggest that $H_1: \lambda \leq 0.85$ is true.

Figure 22-2 graphically illustrates the p -value. Our test statistic was found to be $z = -1.408$, and the p -value of this hypothesis test is the probability of



obtaining a test statistic value z which represents a larger distance below the hypothesized proportion than the observed $z = -1.408$. In Figure 22-2, the observed test statistic value $z = -1.408$ has been located on the horizontal axis. The shaded area corresponds to test statistic values which represent a larger distance below the

It is important to realize the distinction between a *statistically significant difference* and a *clinically significant difference*. A statistically significant difference between the actual value of a parameter and the hypothesized value of the parameter is a difference which is detected by the rejection of H_0 , as we have already seen. A clinically significant difference between the actual value of a parameter and the hypothesized value of the parameter is a difference which is large enough to have some practical impact. A statistically significant difference occurs as the result of a hypothesis test, whereas a clinically significant difference occurs as a matter of judgment.

To illustrate, suppose the manager of a plant would like to see if there is any evidence that the mean amount of cereal per box on the plant assembly line is different from the advertised amount of 12 ounces. If the plant manager concluded that the mean amount of cereal per box was 11.1 ounces, he might feel strongly inclined to make some adjustment in the assembly line in order to increase the mean amount of cereal per box, because producing boxes of cereal which contained an average almost one ounce less than the advertised amount may very well be considered unacceptable. On the other hand, if the plant manager concluded that the mean amount of cereal per box was 11.97 ounces, he might feel no need to make any adjustment in the assembly line, because producing boxes of cereal which contained an average considerably less than one tenth of an ounce smaller than the advertised amount may not be considered of any practical importance.

A statistically significant difference is not necessarily always clinically significant, and a clinically significant difference may not necessarily be statistically significant. The difference of almost one ounce between the actual mean amount of cereal per box and the advertised 12 ounces per box may very well be considered clinically significant by the manager; consequently, he would be motivated to instigate some change, if he were aware of this difference. Whether or not this difference of almost an ounce is statistically significant will depend on whether or not the sample size is sufficiently large to make the probability of detecting a difference of this size highly likely. The difference of less than one tenth of an ounce between the actual mean amount of cereal per box and the advertised 12 ounces per box may very well be considered insignificant by the manager; consequently, he would have no motivation for making any changes as a result of this difference. Whether or not this of difference less than one tenth of an ounce is statistically significant will depend on whether or not the sample size is sufficiently large to make the probability of detecting a difference of this size highly likely.

In general, the decision as to whether or not a difference should be treated as clinically significant involves some subject judgment. For instance, there does not exist any precise rule to tell us exactly how far below the advertised 12 ounces should be considered clinically significant. The hope in any hypothesis test is always that the sample size will be large enough to detect differences that would be considered clinically significant. On the other hand, whenever a difference is statistically significant, some judgment must be made to decide whether or not the difference should be treated as clinically significant.

In Hypothesis tests are not really designed to identify a clinically significant difference. In the earlier hypothesis test concerning the substance called PatchUp, we concluded that the proportion of times PatchUp is completely successful at repairing and inflating a bicycle tire is less than 0.85. Although we found the sample proportion of complete successes to be statistically significantly less than the claimed 0.85, the results of the hypothesis test do not tell us how much less than the claimed 0.85, the population proportion actually is. If we were to find out that the actual proportion of complete successes is 0.848, we would most likely not really think that the manufacturer had made an exaggerated claim; however, if we were to discover that the actual proportion of complete successes is 0.805, we might be very inclined to think of the manufacturer's claim as exaggerated. The results of the hypothesis test do not suggest to us how much less than the claimed 0.85 the population proportion actually is, giving us no basis from which to make a judgment about the clinical significance of the difference. One possible approach though is to look at the difference between the sample proportion $\bar{p} = 123/152 = 0.8092$ and the hypothesized proportion 0.85.

Returning to the hypothesis test of Table 22-1, even though you concluded that the mean disintegration time of the magnesium tablets is greater than 100 seconds, the results of this hypothesis test do not suggest to us how much more than the hypothesized 100 seconds, the population mean actually is. Consequently, we have no basis from which to make a judgment about the clinical significance of the difference. One possible approach is to look at the difference between the sample mean $\bar{x} = 104.5$ seconds and the hypothesized mean 100 seconds.

Answers to Self-Test Problems

- 22-1** (a) Step 1: $H_0: \lambda = 0.7$ vs. $H_1: \lambda < 0.7$ ($\alpha = 0.10$, one-sided)
Step 2: $\bar{p} = 133/200 = 0.665$ and $z = -1.080$
Step 3: The rejection is $z \leq -1.282$. H_0 is not rejected; $0.10 < p$ -value.
Step 4: Since $z = -1.080$ and $z_{0.10} = 1.282$, we do not have sufficient evidence to reject H_0 . We conclude that the proportion of female patrons of the HairCare shop is not less than 0.7 ($0.10 < p$ -value).
- (b) $200(0.7) = 140$ and $200(1 - 0.7) = 60$ are both larger than 5 implying n is sufficiently large.
- (c) Since H_0 is not rejected, a Type II error is possible, which is concluding that $\lambda = 0.7$ when really $\lambda < 0.7$.
- (d) H_0 would not have been rejected with both $\alpha = 0.01$ and $\alpha = 0.05$.
- (e) a bar chart or pie chart.
- 22-2** (a) $n = 30$, $\bar{x} = 45.4$ thousand dollars, and $s = 15.895$ thousand dollars are all correct.
- (b) Step 1: $H_0: \mu = 42$ vs. $H_1: \mu > 42$ ($\alpha = 0.01$, one-sided)
Step 2: $n = 30$, $\bar{x} = 45.4$ thousand dollars, $s = 15.895$ thousand dollars, and $t_{29} = +1.172$
Step 3: The rejection is $t_{29} \geq +2.462$. H_0 is not rejected; $0.10 < p$ -value.
Step 4: Since $t_{29} = +1.172$ and $t_{29;0.01} = 2.462$, we do not have sufficient evidence to reject H_0 . We conclude that the mean yearly income in the state is not more than 42 thousand dollars ($0.10 < p$ -value).
- (c) Since yearly income is a quantitative variable, a box plot is an appropriate graphical display.
- (d) The five-number summary is 25, 33, 40.5, 60, 78. Since there are no potential outliers, and the box plot does not look extremely skewed, there is no reason to think that the t statistic is not appropriate.
- (e) Since H_0 is not rejected, a Type II error is possible, which is concluding that $\mu = 42$ when really $\mu \neq 42$.
- (f) H_0 would not have been rejected with both $\alpha = 0.05$ and $\alpha = 0.10$.

Summary

A hypothesis test which is designed to identify a difference from a hypothesized value in only one direction is called a *one-sided test*. A hypothesis test which is designed to identify a difference from a hypothesized value in either direction is called a *two-sided test*. A *clinically significant difference* between the actual value of a parameter and the hypothesized value of the parameter is a difference which is large enough to have some practical impact. A *statistically significant difference* between the actual value of a parameter and the hypothesized value of the parameter is a difference is one detected by hypothesis test. A statistically significant difference occurs as the result of a hypothesis test, whereas a clinically significant difference occurs as a matter of judgment. A clinically significant difference may or may not be statistically significant, and a statistically significant difference may or may not be clinically significant.