4.3.5 Two-sided hypothesis testing with p-values

We now consider how to compute a p-value for a two-sided test. In one-sided tests, we shade the single tail in the direction of the alternative hypothesis. For example, when the alternative had the form µ > 7, then the p-value was represented by the upper tail (Figure 4.16). When the alternative was µ < 46.99, the p-value was the lower tail (Guided Practice 4.30). In a two-sided test, we shade two tails since evidence in either direction is favorable to HA.

Example 4.34

The second college randomly samples 122 students and finds a mean of ¯x = 6.83 hours and a standard deviation of s = 1.8 hours. Does this provide strong evidence against H0 in Guided Practice 4.33? Use a significance level of α = 0.05.

First, we must verify assumptions.

(1) A simple random sample of less than 10% of the student body means the observations are independent.

(2) The sample size is 122, which is greater than 30.

(3) Based on the earlier distribution and what we already know about college student sleep habits, the sample size will be acceptable.

Next we can compute the standard error (SEx¯ = s/√n = 0.16) of the estimate and create a picture to represent the p-value, shown in Figure 4.18. Both tails are shaded. An estimate of 7.17 or more provides at least as strong of evidence against the null hypothesis and in favor of the alternative as the observed estimate, ¯x = 6.83.

We can calculate the tail areas by first finding the lower tail corresponding to ¯x:

Z = 6.83-7.00/0.16 = -1.06🡪 z-table🡪 left tail = 0.1446

Because the normal model is symmetric, the right tail will have the same area as the left tail. The p-value is found as the sum of the two shaded tails:

p-value = left tail + right tail = 2 x (left tail) = 0.2892

This p-value is relatively large (larger than α = 0.05), so we should not reject H0.

That is, if H0 is true, it would not be very unusual to see a sample mean this far from 7 hours simply due to sampling variation. Thus, we do not have sufficient evidence to conclude that the mean is different than 7 hours.

4.3.6 Choosing a significance level

Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the application.

We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test. If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring HA before we would reject H0.

If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H0 when the null is actually false.

Example 4.36

A car manufacturer is considering a higher quality but more expensive supplier for window parts in its vehicles. They sample a number of parts from their current supplier and also parts from the new supplier. They decide that if the high quality parts will last more than 12% longer, it makes financial sense to switch to this more expensive supplier. Is there good reason to modify the significance level in such a hypothesis test?

The null hypothesis is that the more expensive parts last no more than 12% longer while the alternative is that they do last more than 12% longer. This decision is just one of the many regular factors that have a marginal impact on the car and company. A significance level of 0.05 seems reasonable since neither a Type 1 or Type 2 Error should be dangerous or (relatively) much more expensive.

Example 4.37

The same car manufacturer is considering a slightly more expensive supplier for parts related to safety, not windows. If the durability of these safety components is shown to be better than the current supplier, they will switch manufacturers. Is there good reason to modify the significance level in such an evaluation?

The null hypothesis would be that the suppliers’ parts are equally reliable. Because safety is involved, the car company should be eager to switch to the slightly more expensive manufacturer (reject H0) even if the evidence of increased safety is only moderately strong. A slightly larger significance level, such as α = 0.10, might be appropriate.

Guided Practice 4.38

A part inside of a machine is very expensive to replace. However, the machine usually functions properly even if this part is broken, so the part is replaced only if we are extremely certain it is broken based on a series of measurements. Identify appropriate hypotheses for this test (in plain language) and suggest an appropriate significance level.

Here the null hypothesis is that the part is not broken, and the alternative is that it is broken. If we don’t have sufficient evidence to reject H0, we would not replace the part. It sounds like failing to fix the part if it is broken (H0 false, HA true) is not very problematic, and replacing the part is expensive. Thus, we should require very strong evidence against H0 before we replace the part. Choose a small significance level, such as α= 0.01.

4.4 Examining the Central Limit Theorem

The Central Limit Theorem states that when the sample size is small, the normal approximation may not be very good. However, as the sample size becomes large, the normal approximation improves. We will investigate three cases to see roughly when the approximation is reasonable.

We consider three data sets: one from a uniform distribution, one from an exponential distribution, and the other from a log-normal distribution. These distributions are shown in the top panels of Figure 4.20.

The uniform distribution is symmetric, the exponential distribution may be considered as having moderate skew since its right tail is relatively short (few outliers), and the log-normal distribution is strongly skewed and will tend to produce more apparent outliers.

TIP: With larger n, the sampling distribution of x¯ becomes more normal

As the sample size increases, the normal model for ¯x becomes more reasonable. We can also relax our condition on skew when the sample size is very large.

Guided Practice 4.47

The proportion of students who are male in the yrbss samp sample is ˆp = 0.48. This sample meets certain conditions that ensure ˆp will be nearly normal, and the standard error of the estimate is SEpˆ = 0.05. Create a 90% confidence interval for the proportion of students in the 2013 YRBSS survey who are male.

We use z? = 1.65 (see Guided Practice 4.17 on page 180), and apply the general confidence interval formula: pˆ ± z\*SEpˆ 🡪0.48 ± 1.65\*0.05 🡪 (0.3975, 0.5625) Thus, we are 90% confident that between 40% and 56% of the YRBSS students were male

Hypothesis testing using the normal model

1. First write the hypotheses in plain language, then set them up in mathematical notation.

2. Identify an appropriate point estimate of the parameter of interest.

3. Verify conditions to ensure the standard error estimate is reasonable and the point estimate is nearly normal and unbiased.

4. Compute the standard error. Draw a picture depicting the distribution of the estimate under the idea that H0 is true. Shade areas representing the p-value.

5. Using the picture and normal model, compute the test statistic (Z-score) and identify the p-value to evaluate the hypotheses. Write a conclusion in plain language.

Guided Practice 4.48

A drug called sulphinpyrazone was under consideration for use in reducing the death rate in heart attack patients. To determine whether the drug was effective, a set of 1,475 patients were recruited into an experiment and randomly split into two groups: a control group that received a placebo and a treatment group that received the new drug. What would be an appropriate null hypothesis? And the alternative?

The skeptic’s perspective is that the drug does not work at reducing deaths in heart attack patients (H0), while the alternative is that the drug does work (HA).

We can formalize the hypotheses from Guided Practice 4.48 by letting pcontrol and ptreatment represent the proportion of patients who died in the control and treatment groups, respectively.

Then the hypotheses can be written as

H0 : pcontrol = ptreatment (the drug doesn’t work)

HA : pcontrol > ptreatment (the drug works) or

equivalently,

H0 : pcontrol-ptreatment = 0 (the drug doesn’t work)

HA : pcontrol-ptreatment > 0 (the drug works)

Strong evidence against the null hypothesis and in favor of the alternative would correspond to an observed difference in death rates,

point estimate = ˆpcontrol-pˆtreatment

being larger than we would expect from chance alone. This difference in sample proportions represents a point estimate that is useful in evaluating the hypotheses.

Example 4.49

We want to evaluate the hypothesis setup from Guided Practice 4.48 using data from the actual study.36 In the control group, 60 of 742 patients died. In the treatment group, 41 of 733 patients died. The sample difference in death rates can be summarized as

point estimate = ˆpcontrol-pˆtreatment = (60/742)-(41/733) = 0.025

This point estimate is nearly normal and is an unbiased estimate of the actual difference in death rates. The standard error of this sample difference is SE = 0.013. Evaluate the hypothesis test at a 5% significance level: α = 0.05.

We would like to identify the p-value to evaluate the hypotheses. If the null hypothesis is true, then the point estimate would have come from a nearly normal distribution, like the one shown in Figure 4.22. The distribution is centered at zero since

pcontrol-ptreatment = 0

under the null hypothesis. Because a large positive difference provides evidence against the null hypothesis and in favor of the alternative, the upper tail has been shaded to represent the p-value. We need not shade the lower tail since this is a one-sided test: an observation in the lower tail does not support the alternative hypothesis. The p-value can be computed by using the Z-score of the point estimate and the normal probability table.

Z = (point estimate-null value)/SEpoint estimate = 0.025-0/0.013 = 1.92

Examining Z in the normal probability table, we find that the lower unshaded tail is about 0.973. Thus, the upper shaded tail representing the p-value is

p-value = 1-0.973 = 0.027

Because the p-value is less than the significance level (α = 0.05), we say the null hypothesis is implausible. That is, we reject the null hypothesis in favor of the alternative and conclude that the drug is effective at reducing deaths in heart attack patients.

Test Statistic

The Z-score in Equation (4.50) is called a test statistic. In most hypothesis tests, a test statistic is a particular data summary that is especially useful for computing the p-value and evaluating the hypothesis test. In the case of point estimates that are nearly normal, the test statistic is the Z-score.

A test statistic is a summary statistic that is particularly useful for evaluating a hypothesis test or identifying the p-value. When a point estimate is nearly normal, we use the Z-score of the point estimate as the test statistic.

4.5.3 Non-normal point estimates

We may apply the ideas of confidence intervals and hypothesis testing to cases where the point estimate or test statistic is not necessarily normal. There are many reasons why such a situation may arise:

• the sample size is too small for the normal approximation to be valid;

• the standard error estimate may be poor; or

• the point estimate tends towards some distribution that is not the normal distribution.

For each case where the normal approximation is not valid, our first task is always to understand and characterize the sampling distribution of the point estimate or test statistic. Next, we can apply the general frameworks for confidence intervals and hypothesis testing to these alternative distributions.

4.5.4 When to retreat

Statistical tools rely on conditions. When the conditions are not met, these tools are unreliable and drawing conclusions from them is treacherous. The conditions for these tools typically come in two forms.

• The individual observations must be independent. A random sample from less than 10% of the population ensures the observations are independent. In experiments, we generally require that subjects are randomized into groups. If independence fails, then advanced techniques must be used, and in some such cases, inference may not be possible.

• Other conditions focus on sample size and skew. For example, if the sample size is too small, the skew too strong, or extreme outliers are present, then the normal model for the sample mean will fail. Verification of conditions for statistical tools is always necessary. Whenever conditions are not satisfied for a statistical technique, there are three options. The first is to learn new methods that are appropriate for the data. The second route is to consult a statistician.37 The third route is to ignore the failure of conditions. This last option effectively invalidates any analysis and may discredit novel and interesting findings. Finally, we caution that there may be no inference tools helpful when considering data that include unknown biases, such as convenience samples. For this reason, there are books, courses, and researchers devoted to the techniques of sampling and experimental design. See Sections 1.3-1.5 for basic principles of data collection.