Statistical Methods for the Social Sciences

Fifth Edition

Alan Agresti

University of Florida

Pearson

Chapter 6

Statistical Inference: Significance Tests

CHAPTER OUTLINE

An aim of many studies is to check whether the data agree with certain predictions. The predictions, which often result from the theory that drives the research, are *hypotheses* about the study population.

Hypothesis

> In statistics, a ***hypothesis*** is a statement about a population. It takes the form of a prediction that a parameter takes a particular numerical value or falls in a certain range of values.

Examples of hypotheses are the following: "For restaurant managerial employees, the mean salary is the same for women and for men"; "There is no difference between Democrats and Republicans in the probabilities that they vote with their party leadership"; and "A majority of adult Canadians are satisfied with their national health service."

A statistical ***significance test*** uses data to summarize the evidence about a hypothesis. It does this by comparing point estimates of parameters to the values predicted by the hypothesis. The following example illustrates concepts behind significance tests.

Example 6.1

Testing for Gender Bias in Selecting Managers A large supermarket chain in Florida periodically selects employees to receive management training. A group of women employees recently claimed that the company selects males at a disproportionally high rate for such training. The company denied this claim. In past years, similar claims of gender bias have been made about promotions and pay for women who work for various companies.[1] How could the women employees statistically back up their assertion?

Suppose the employee pool for potential selection for management training is half male and half female. Then, the company's claim of a lack of gender bias is a hypothesis. It states that, other things being equal, at each choice the probability of selecting a female equals 1/2 and the probability of selecting a male equals 1/2. If the employees truly are selected for management training randomly in terms of gender, about half the employees picked should be females and about half should be male. The women's claim is an alternative hypothesis that the probability of selecting a male exceeds 1/2.

Suppose that 9 of the 10 employees chosen for management training were male. We might be inclined to believe the women's claim. However, we should analyze whether these results would be unlikely if there were *no* gender bias. Would it be highly unusual that 9/10 of the employees chosen would have the same gender if they were truly selected at random from the employee pool?

[1] For example, Wal-Mart, see `http://now.org/blog/walmart-and-sex-discrimination`.

Due to sampling variation, not exactly 1/2 of the sample need be male. How far above 1/2 must the sample proportion of males chosen be before we believe the women's claim? ■

This chapter introduces statistical methods for summarizing evidence and making decisions about hypotheses. We first present the parts that all significance tests have in common. The rest of the chapter presents significance tests about population means and population proportions. We'll also learn how to find and how to control the probability of an incorrect decision about a hypothesis.

6.1 The Five Parts of a Significance Test

Now let's take a closer look at the significance test method, also called a *hypothesis test*, or *test* for short. All tests have five parts:

Assumptions, Hypotheses, Test statistic, P-value, Conclusion.

ASSUMPTIONS

Each test makes certain assumptions or has certain conditions for the test to be valid. These pertain to

- *Type of data*: Like other statistical methods, each test applies for either quantitative data or categorical data.
- *Randomization*: Like other methods of statistical inference, a test assumes that the data gathering employed randomization, such as a random sample.
- *Population distribution*: Some tests assume that the variable has a particular probability distribution, such as the normal distribution.
- *Sample size*: Many tests employ an approximate normal or t sampling distribution. The approximation is adequate for any n when the population distribution is approximately normal, but it also holds for highly nonnormal populations when the sample size is relatively large, by the Central Limit Theorem.

HYPOTHESES

Each significance test has two hypotheses about the value of a population parameter.

Null Hypothesis, Alternative Hypothesis

> The ***null hypothesis***, denoted by the symbol H_0, is a statement that the parameter takes a particular value. The ***alternative hypothesis***, denoted by H_a, states that the parameter falls in some alternative range of values. Usually the value in H_0 corresponds, in a certain sense, to *no effect*. The values in H_a then represent an effect of some type.

In Example 6.1 about possible gender discrimination in selecting management trainees, let π denote the probability that any particular selection is a male. The company claims that $\pi = 1/2$. This is an example of a null hypothesis, *no effect* referring to a lack of gender bias. The alternative hypothesis reflects the skeptical women employees' belief that this probability actually exceeds 1/2. So, the hypotheses are H_0: $\pi = 1/2$ and H_a: $\pi > 1/2$. Note that H_0 has a *single* value whereas H_a has a range of values.

A significance test analyzes the sample evidence about H_0, by investigating whether the data contradict H_0, hence suggesting that H_a is true. The approach taken is the indirect one of *proof by contradiction*. The null hypothesis is presumed to be true. Under this presumption, if the data observed would be very unusual, the evidence supports the alternative hypothesis. In the study of potential gender discrimination, we presume that H_0: $\pi = 1/2$ is true. Then we determine whether the sample result of 9 men selected for management training in 10 choices would be unusual, under this presumption. If so, then we may be inclined to believe the women's claim. But, if the difference between the sample proportion of men chosen (9/10) and the H_0 value of 1/2 could easily be due to ordinary sampling variability, there's not enough evidence to accept the women's claim.

A researcher usually conducts a test to gauge the amount of support for the alternative hypothesis, as that typically reflects an effect that he or she predicts. Thus, H_a is sometimes called the ***research hypothesis***. The hypotheses are formulated *before* collecting or analyzing the data.

TEST STATISTIC

The parameter to which the hypotheses refer has a point estimate. The ***test statistic*** summarizes how far that estimate falls from the parameter value in H_0. Often this is expressed by the *number of standard errors* between the estimate and the H_0 value.

P-VALUE

To interpret a test statistic value, we create a probability summary of the evidence against H_0. This uses the sampling distribution of the test statistic, under the presumption that H_0 is true. The purpose is to summarize how unusual the observed test statistic value is compared to what H_0 predicts.

Specifically, if the test statistic falls well out in a tail of the sampling distribution in a direction predicted by H_a, then it is far from what H_0 predicts. We can summarize how far out in the tail the test statistic falls by the tail probability of that value and of more extreme values. These are the possible test statistic values that provide *at least as much evidence against H_0 as the observed test statistic*, in the direction predicted by H_a. This probability is called the ***P-value***.

P-value

> The ***P-value*** is the probability that the test statistic equals the observed value or a value even more extreme in the direction predicted by H_a. It is calculated by presuming that H_0 is true. The P-value is denoted by P.

A small P-value (such as $P = 0.01$) means that the data we observed would have been unusual if H_0 were true. *The smaller the P-value, the stronger the evidence is against H_0.*

For Example 6.1 on potential gender discrimination in choosing managerial trainees, π is the probability of selecting a male. We test H_0: $\pi = 1/2$ against H_a: $\pi > 1/2$. One possible test statistic is the sample proportion of males selected, which is $9/10 = 0.90$. The values for the sample proportion that provide this much or even more extreme evidence against H_0: $\pi = 1/2$ and in favor of H_a: $\pi > 1/2$ are the right-tail sample proportion values of 0.90 and higher. See Figure 6.1. A formula from Section 6.7 calculates this probability as 0.01, so the P-value equals $P = 0.01$. If the selections truly were random with respect to gender, the probability is only 0.01 of such an extreme sample result, namely, that 9 or all 10 selections would be males. Other things being equal, this small P-value provides considerable evidence against H_0: $\pi = 1/2$ and supporting the alternative H_a: $\pi > 1/2$ of discrimination against females.

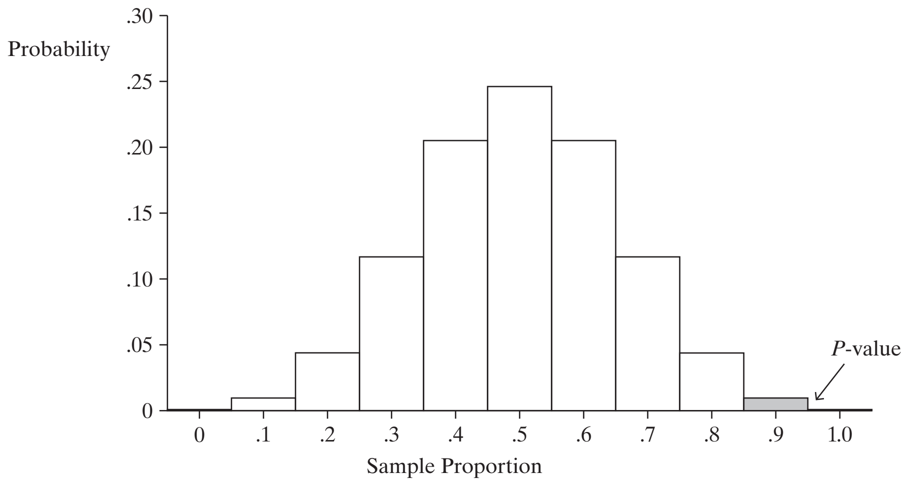


FIGURE 6.1: The P-Value Equals the Probability of the Observed Data or Even More Extreme Results. It is calculated under the presumption that H_0 is true, so a very small P-value gives strong evidence against H_0.

By contrast, a moderate to large P-value means the data are consistent with H_0. A P-value such as 0.26 or 0.83 indicates that, if H_0 were true, the observed data would not be unusual.

CONCLUSION

The P-value summarizes the evidence against H_0. Our conclusion should also *interpret* what the P-value tells us about the question motivating the test. Sometimes it is necessary to make a decision about the validity of H_0. If the P-value is sufficiently small, we reject H_0 and accept H_a.

Most studies require very small P-values, such as $P \leq 0.05$, in order to reject H_0. In such cases, results are said to be *significant at the 0.05 level*. This means that if H_0 were true, the chance of getting such extreme results as in the sample data would be no greater than 0.05.

Making a decision by rejecting or not rejecting a null hypothesis is an optional part of the significance test. We defer discussion of it until Section 6.4. Table 6.1 summarizes the parts of a significance test.

TABLE 6.1: The Five Parts of a Statistical Significance Test

1. **Assumptions**
   Type of data, randomization, population distribution, sample size condition
2. **Hypotheses**
   Null hypothesis, H_0
   (parameter value for "no effect")
   Alternative hypothesis, H_a
   (alternative parameter values)
3. **Test statistic**
   Compares point estimate to H_0 parameter value
4. ***P*-value**
   Weight of evidence against H_0; smaller P is stronger evidence
5. **Conclusion**
   Report and interpret P-value
   Formal decision (optional; see Section 6.4)

6.2 Significance Test for a Mean

For quantitative variables, significance tests usually refer to population means. The five parts of the significance test for a single mean follow:

THE FIVE PARTS OF A SIGNIFICANCE TEST FOR A MEAN

1. Assumptions

The test assumes the data are obtained using randomization, such as a random sample. The quantitative variable is assumed to have a normal population distribution. We'll see that this is mainly relevant for small sample sizes and certain types of H_a.

2. Hypotheses

The null hypothesis about a population mean μ has the form

$$H_0: \mu = \mu_0,$$

where μ_0 is a particular value for the population mean. In other words, the hypothesized value of μ in H_0 is a single value. This hypothesis usually refers to *no effect* or *no change* compared to some standard. For example, Example 5.5 in the previous chapter (page 117) estimated the population mean weight change μ for teenage girls after receiving a treatment for anorexia. The hypothesis that the treatment has *no effect* is a null hypothesis, H_0: $\mu = 0$. Here, the H_0 value μ_0 for the parameter μ is 0.

The alternative hypothesis contains alternative parameter values from the value in H_0. The most common alternative hypothesis is

$$H_a: \mu \neq \mu_0, \quad \text{such as} \quad H_a: \mu \neq 0.$$

This alternative hypothesis is called ***two-sided***, because it contains values both below and above the value listed in H_0. For the anorexia study, H_a: $\mu \neq 0$ states that the treatment has *some effect*, the population mean equaling some value other than 0.

3. Test Statistic

The sample mean $\bar{y}$ estimates the population mean μ. When the population distribution is normal, the sampling distribution of $\bar{y}$ is normal about μ. This is also approximately true when the population distribution is *not* normal but the random sample size is relatively large, by the Central Limit Theorem.

Under the presumption that H_0: $\mu = \mu_0$ is true, the center of the sampling distribution of $\bar{y}$ is the value μ_0, as Figure 6.2 shows. A value of $\bar{y}$ that falls far out in the tail provides strong evidence against H_0, because it would be unusual if truly $\mu = \mu_0$.

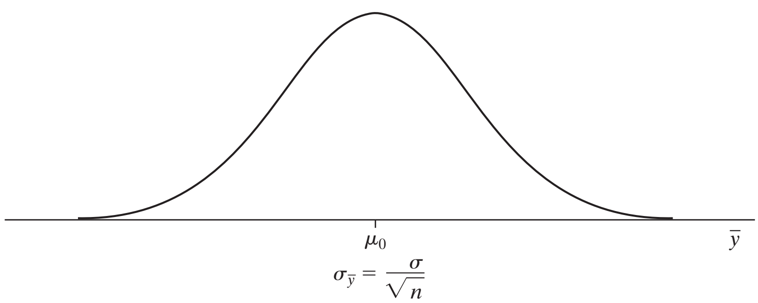


FIGURE 6.2: Sampling Distribution of $\bar{y}$ if H_0: $\mu = \mu_0$ Is True. For large random samples, it is approximately normal, centered at the null hypothesis value, μ_0.

The evidence about H_0 is summarized by the number of standard errors that $\bar{y}$ falls from the null hypothesis value μ_0.

Recall that the *true* standard error is $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. As in Chapter 5, we substitute the sample standard deviation s for the unknown population standard deviation σ to get the *estimated* standard error, $se = s/\sqrt{n}$. The test statistic is the t-score

$$t = \frac{\bar{y} - \mu_0}{se}, \quad \text{where} \quad se = \frac{s}{\sqrt{n}}.$$

The farther $\bar{y}$ falls from μ_0, the larger the absolute value of the t test statistic. Hence, the larger the value of $|t|$, the stronger the evidence against H_0.

We use the symbol t rather than z because, as in forming a confidence interval, using s to estimate σ in the standard error introduces additional error. The null sampling distribution of the t test statistic is the *t distribution* (see Section 5.3). It looks like the standard normal distribution, having mean equal to 0 but being more spread out, more so for smaller n. It is specified by its degrees of freedom, $df = n - 1$.

4. *P*-Value

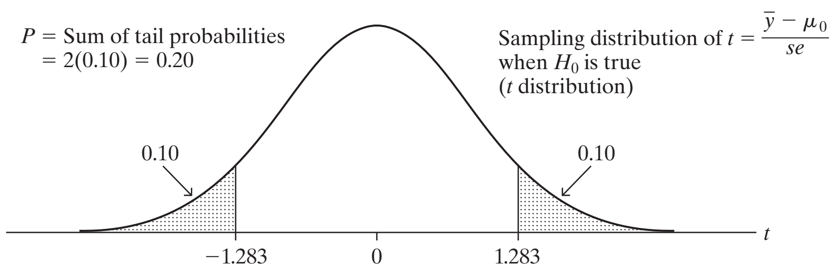
The test statistic summarizes how far the data fall from H_0. Different tests use different test statistics, though, and simpler interpretations result from transforming it to the probability scale of 0 to 1. The P-value does this.

We calculate the P-value under the presumption that H_0 is true. That is, we give the benefit of the doubt to H_0, analyzing how unusual the observed data would be if H_0 were true. The P-value is the probability that the test statistic equals the observed value or a value in the set of more extreme values that provide even stronger evidence against H_0. For H_a: $\mu \neq \mu_0$, the more extreme t-values are the ones even farther out in the tails of the t distribution. So, the P-value is the two-tail probability that the t test statistic is at least as large in absolute value as the observed test statistic. This is also the probability that $\bar{y}$ falls at least as far from μ_0 *in either direction* as the observed value of $\bar{y}$.

Figure 6.3 shows the sampling distribution of the t test statistic when H_0 is true. A test statistic value of $t = (\bar{y} - \mu_0)/se = 0$ results when $\bar{y} = \mu_0$. This is the t-value most consistent with H_0. The P-value is the probability of a t test statistic value at least as far from this consistent value as the one observed. To illustrate its calculation, suppose $t = 1.283$ for a sample size of 369. (This is the result in the example below.) This t-score means that the sample mean $\bar{y}$ falls 1.283 estimated standard errors above μ_0. The P-value is the probability that $t \geq 1.283$ or $t \leq -1.283$ (i.e., $|t| \geq 1.283$). Since $n = 369$, $df = n - 1 = 368$ is large, and the t distribution is nearly identical to the standard normal. The probability in one tail above 1.28 is 0.10, so the two-tail probability is $P = 2(0.10) = 0.20$.

Software can supply tail probabilities for the t distribution. For example, the free software R has a function `pt` that gives the cumulative probability for a particular t-score. When $df = 368$, the right-tail probability above $t = 1.283$ is 1 − the cumulative probability:

FIGURE 6.3: Calculation of P-Value when $t = 1.283$, for Testing H_0: $\mu = \mu_0$ against H_a: $\mu \neq \mu_0$. The P-value is the two-tail probability of a more extreme result than the observed one.



```
> 1 - pt(1.283, 368)
[1] 0.1001498   # right-tail probability above t=1.283, when df=368
```

With Stata software, we can find the right-tail probability with the `ttail` function:

```
. display ttail(368, 1.283)
.10014975
```

We double the single-tail probability to get the P-value, $P = 2(0.10014975) = 0.2002995$. Round such a value, say to 0.20, before reporting it. Reporting the P-value with many decimal places makes it seem as if more accuracy exists than actually does. In practice, the sampling distribution is only *approximately* the t distribution, because the population distribution is not exactly normal as is assumed with the t test.

Tail probabilities for the t distribution are also available using SPSS and SAS and Internet applets, such as Figure 5.7 showed with the *t Distribution* applet at `www.artofstat.com/webapps.html`.

5. Conclusion

Finally, the study should interpret the P-value in context. The smaller P is, the stronger the evidence against H_0 and in favor of H_a.

Example 6.2

Significance Test about Political Ideology Some political commentators have remarked that citizens of the United States are increasingly conservative, so much so that many treat "liberal" as a dirty word. We can study political ideology by analyzing responses to certain items on the General Social Survey. For instance, that survey asks where you would place yourself on a seven-point scale of political views ranging from extremely liberal, point 1, to extremely conservative, point 7. Table 6.2 shows the scale and the distribution of responses among the levels for the 2014 survey. Results are shown separately according to subjects classified as white, black, or Hispanic.

Political ideology is an ordinal scale. Often, we treat such scales in a quantitative manner by assigning scores to the categories. Then we can use quantitative summaries such as means, allowing us to detect the extent to which observations gravitate toward the conservative or the liberal end of the scale. If we assign the category

TABLE 6.2: Responses of Subjects on a Scale of Political Ideology

	Race		
Response	Black	White	Hispanic
1. Extremely liberal	16	73	5
2. Liberal	52	209	49
3. Slightly liberal	42	190	46
4. Moderate, middle of road	182	705	155
5. Slightly conservative	43	260	50
6. Conservative	25	314	50
7. Extremely conservative	11	84	14
	$n = 371$	$n = 1835$	$n = 369$

scores shown in Table 6.2, then a mean below 4 shows a propensity toward liberalism and a mean above 4 shows a propensity toward conservatism. We can test whether these data show much evidence of either of these by conducting a significance test about how the population mean compares to the moderate value of 4. We'll do this here for the Hispanic sample and in Section 6.5 for the entire sample.

1. *Assumptions*: The sample is randomly selected. We are treating political ideology as quantitative with equally spaced scores. The t test assumes a normal population distribution for political ideology, which seems inappropriate because the measurement of political ideology is discrete. We'll discuss this assumption further at the end of this section.
2. *Hypotheses*: Let μ denote the population mean ideology for Hispanic Americans, for this seven-point scale. The null hypothesis contains one specified value for μ. Since we conduct the analysis to check how, if at all, the population mean departs from the moderate response of 4, the null hypothesis is

$$H_0: \mu = 4.0.$$

   The alternative hypothesis is then

$$H_a: \mu \neq 4.0.$$

   The null hypothesis states that, on the average, the population response is politically "moderate, middle of road." The alternative states that the mean falls in the liberal direction ($\mu < 4.0$) or in the conservative direction ($\mu > 4.0$).
3. *Test statistic*: The 369 observations in Table 6.2 for Hispanics are summarized by $\bar{y} = 4.089$ and $s = 1.339$. The estimated standard error of the sampling distribution of $\bar{y}$ is

$$se = \frac{s}{\sqrt{n}} = \frac{1.339}{\sqrt{369}} = 0.0697.$$

   The value of the test statistic is

$$t = \frac{\bar{y} - \mu_0}{se} = \frac{4.089 - 4.0}{0.0697} = 1.283.$$

   The sample mean falls 1.283 estimated standard errors above the null hypothesis value of the mean. The $df = 369 - 1 = 368$.
4. *P-value*: The P-value is the two-tail probability, presuming H_0 is true, that t would exceed 1.283 in absolute value. From the t distribution with $df = 368$, this two-tail probability is $P = 0.20$. If the population mean ideology were 4.0, then the probability equals 0.20 that a sample mean for $n = 368$ subjects would fall at least as far from 4.0 as the observed $\bar{y}$ of 4.089.
5. *Conclusion*: The P-value of $P = 0.20$ is not very small, so it does not contradict H_0. If H_0 were true, the data we observed would not be unusual. It is plausible that the population mean response for Hispanic Americans in 2014 was 4.0, not leaning in the conservative or liberal direction. ■

CORRESPONDENCE BETWEEN TWO-SIDED TESTS AND CONFIDENCE INTERVALS

Conclusions using two-sided significance tests are consistent with conclusions using confidence intervals. If a test says that a particular value is believable for the parameter, then so does a confidence interval.

Example 6.3

Confidence Interval for Mean Political Ideology For the data in Example 6.2, let's construct a 95% confidence interval for the Hispanic population mean political ideology. With $df = 368$, the multiple of the standard error ($se = 0.0697$) is $t_{.025} = 1.966$. Since $\bar{y} = 4.089$, the confidence interval is

$$\bar{y} \pm 1.966(se) = 4.089 \pm 1.966(0.0697) = 4.089 \pm 0.137, \quad \text{or} \quad (3.95, 4.23).$$

At the 95% confidence level, these are the plausible values for μ. ■

This confidence interval indicates that μ may equal 4.0, since 4.0 falls inside the confidence interval. Thus, it is not surprising that the P-value ($P = 0.20$) in testing H_0: $\mu = 4.0$ against H_a: $\mu \neq 4.0$ in Example 6.2 was not small. In fact,

> Whenever the $P > 0.05$ in a two-sided test about a mean μ, a 95% confidence interval for μ *necessarily contains* the H_0 value for μ.

By contrast, suppose the P-value = 0.02 in testing H_0: $\mu = 4.0$. Then, a 95% confidence interval would tell us that 4.0 is implausible for μ, with 4.0 falling *outside* the confidence interval.

> Whenever $P \leq 0.05$ in a two-sided test about a mean μ, a 95% confidence interval for μ *does not contain* the H_0 value for μ.

ONE-SIDED SIGNIFICANCE TESTS

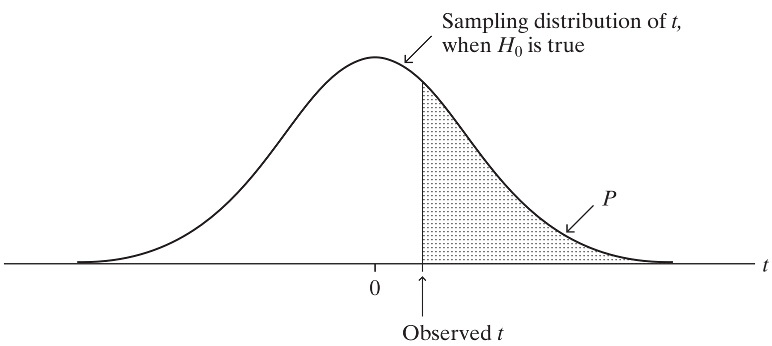
We can use a different alternative hypothesis when a researcher predicts a deviation from H_0 in a particular direction. It has one of the forms

$$H_a: \mu > \mu_0 \quad \text{or} \quad H_a: \mu < \mu_0.$$

We use the alternative H_a: $\mu > \mu_0$ to detect whether μ is *larger* than the particular value μ_0, whereas we use H_a: $\mu < \mu_0$ to detect whether μ is *smaller* than that value. These hypotheses are called *one-sided*. By contrast, we use the *two-sided* H_a to detect *any* type of deviation from H_0. This choice is made before analyzing the data.

For H_a: $\mu > \mu_0$, the P-value is the probability (presuming H_0 is true) of a t-score *above* the observed t-score, that is, to the right of it on the real number line. These t-scores provide more extreme evidence than the observed value in favor of H_a: $\mu > \mu_0$. So, P equals the right-tail probability under the t curve. See Figure 6.4. A t-score of 1.283 with $df = 368$ results in $P = 0.10$ for this alternative.

FIGURE 6.4: Calculation of P-Value in Testing H_0: $\mu = \mu_0$ against H_a: $\mu > \mu_0$. The P-value is the probability of values to the right of the observed test statistic.



For H_a: $\mu < \mu_0$, the P-value is the left-tail probability, *below* the observed t-score. A t-score of $t = -1.283$ with $df = 368$ results in $P = 0.10$ for this alternative. A t-score of 1.283 results in $P = 1 - 0.10 = 0.90$.

Example 6.4 **Test about Mean Weight Change in Anorexic Girls** Example 5.5 in Chapter 5 (page 117) analyzed data (available in the `Anorexia_CB` data file at the text website) from a study comparing treatments for teenage girls suffering from anorexia. For each girl, the study observed her change in weight while receiving the therapy. Let μ denote the population mean change in weight for the cognitive behavioral treatment. If this treatment has beneficial effect, as expected, then μ is positive. To test for no treatment effect versus a positive mean weight change, we test H_0: $\mu = 0$ against H_a: $\mu > 0$.

In the Chapter 5 analysis, we found that the $n = 29$ girls had a sample mean weight change of 3.007 pounds, a standard deviation of 7.309 pounds, and an estimated standard error of $se = 1.357$. The test statistic is

$$t = \frac{\bar{y} - \mu_0}{se} = \frac{3.007 - 0}{1.357} = 2.22.$$

For this one-sided H_a, the P-value is the right-tail probability above 2.22. Why do we use the right tail? Because H_a: $\mu > 0$ has values *above* (i.e., to the right of) the null hypothesis value of 0. It's the positive values of t that support this alternative hypothesis.

Now, for $n = 29$, $df = n - 1 = 28$. The P-value equals 0.02. Software can find the P-value for you. For instance, for the one-sided and two-sided alternatives with a data file with variable *change* for weight change, R reports

```
> t.test(change, mu = 0, alternative = "greater")$p.value
[1] 0.0175113
> t.test(change, mu = 0, alternative = "two.sided")$p.value
[1] 0.0350226
```

Using its `ttest` command with the data file, Stata also reports $P = 0.0175$ for the one-sided H_a: $\mu > 0$. See Table 6.3. If you have only summary statistics rather than a data file, Stata can conduct the test using them, with the `ttesti` command (or a dialog box), by entering n, $\bar{y}$, s, and μ_0 as shown in Table 6.3. Internet applets can also do this.[2]

Some software reports the P-value for a two-sided alternative as the default, unless you request otherwise. SPSS reports results for the two-sided test and confidence interval as

```
                    Test Value = 0                   95% Confidence Interval
                                             Mean      of the Difference
              t   df   Sig.(2-tailed)   Difference    Lower      Upper
change    2.216   28        .035          3.00690      .2269     5.7869
```

The one-sided P-value is $0.035/2 = 0.018$. The evidence against H_0 is relatively strong. It seems that the treatment has an effect.

The significance test concludes that the mean weight gain was not equal to 0. But the 95% confidence interval of (0.2, 5.8) is more informative. It shows just how

[2] Such as the *Inference for a Mean* applet at `www.artofstat.com/webapps.html`.

different from 0 the population mean change is likely to be. The effect could be very small. Also, keep in mind that this experimental study (like many medically oriented studies) had to use a volunteer sample. So, these results are highly tentative, another reason that it is silly for studies like this to report P-values to several decimal places. ■

TABLE 6.3: Stata Software Output (Edited) for Performing a Significance Test about a Mean

```
. ttest change == 0

One-sample t test
Variable | Obs      Mean   Std. Err.  Std. Dev.   [95% Conf. Interval]
  change |  29  3.006896   1.357155   7.308504    .2268896    5.786902

    mean = mean(change)                                      t =   2.2156
Ho: mean = 0                                     degrees of freedom = 28
    Ha: mean < 0            Ha: mean != 0                Ha: mean > 0
 Pr(T < t) = 0.9825     Pr(|T| > |t|) = 0.0350      Pr(T > t) = 0.0175

/*  Can also perform test with n, mean, std. dev., null value  */
. ttesti 29 3.007 7.309 0

    Ha: mean < 0            Ha: mean != 0                Ha: mean > 0
 Pr(T < t) = 0.9825     Pr(|T| > |t|) = 0.0350      Pr(T > t) = 0.0175
```

IMPLICIT ONE-SIDED H_0 FOR ONE-SIDED H_a

From Example 6.4, the one-sided P-value = 0.018. So, if $\mu = 0$, the probability equals 0.018 of observing a sample mean weight gain of 3.01 or greater. Now, suppose $\mu < 0$; that is, the population mean weight change is negative. Then, the probability of observing $\bar{y} \geq 3.01$ would be even smaller than 0.018. For example, a sample value of $\bar{y} = 3.01$ is even less likely when $\mu = -5$ than when $\mu = 0$, since 3.01 is farther out in the tail of the sampling distribution of $\bar{y}$ when $\mu = -5$ than when $\mu = 0$. Thus, rejection of H_0: $\mu = 0$ in favor of H_a: $\mu > 0$ also inherently rejects the broader null hypothesis of H_0: $\mu \leq 0$. In other words, one concludes that $\mu = 0$ is false *and* that $\mu < 0$ is false.

THE CHOICE OF ONE-SIDED VERSUS TWO-SIDED TESTS

In practice, two-sided tests are more common than one-sided tests. Even if a researcher predicts the direction of an effect, two-sided tests can also detect an effect that falls in the opposite direction. In most research articles, significance tests use two-sided P-values. Partly this reflects an objective approach to research that recognizes that an effect could go in either direction. In using two-sided P-values, researchers avoid the suspicion that they chose H_a when they saw the direction in which the data occurred. That is not ethical.

Two-sided tests coincide with the usual approach in estimation. Confidence intervals are two sided, obtained by adding and subtracting some quantity from the point estimate. One can form one-sided confidence intervals, for instance, having 95% confidence that a population mean weight change is *at least* equal to 0.8 pounds (i.e., between 0.8 and ∞), but in practice one-sided intervals are rarely used.

In deciding whether to use a one-sided or a two-sided H_a in a particular exercise or in practice, consider the context. An exercise that says "Test whether the mean has *changed*" suggests a two-sided alternative, to allow for increase or decrease. "Test whether the mean has *increased*" suggests the one-sided H_a: $\mu > \mu_0$.

In either the one-sided or two-sided case, hypotheses always refer to population parameters, not sample statistics. So, *never* express a hypothesis using sample statistic notation, such as H_0: $\bar{y} = 0$. There is no uncertainty or need to conduct statistical inference about sample statistics such as $\bar{y}$, because we can calculate their values exactly from the data.

THE α-LEVEL: USING THE *P*-VALUE TO MAKE A DECISION

A significance test analyzes the strength of the evidence against the null hypothesis, H_0. We start by presuming that H_0 is true. We analyze whether the data would be unusual if H_0 were true by finding the *P*-value. If the *P*-value is small, the data contradict H_0 and support H_a. Generally, researchers do not regard the evidence against H_0 as strong unless P is very small, say, $P \leq 0.05$ or $P \leq 0.01$.

Why do smaller *P*-values indicate stronger evidence against H_0? Because the data would then be more unusual if H_0 were true. When H_0 is true, the *P*-value is roughly equally likely to fall anywhere between 0 and 1. By contrast, when H_0 is false, the *P*-value is more likely to be near 0 than near 1.

Sometimes we need to decide whether the evidence against H_0 is strong enough to reject it. We base the decision on whether the *P*-value falls below a prespecified cutoff point. For example, we could reject H_0 if $P \leq 0.05$ and conclude that the evidence is not strong enough to reject H_0 if $P > 0.05$. The boundary value 0.05 is called the α*-level* of the test.

α-Level

> The ***α-level*** is a number such that we reject H_0 if the *P*-value is less than or equal to it. The α-level is also called the ***significance level***. In practice, the most common α-levels are 0.05 and 0.01.

Like the choice of a confidence level for a confidence interval, the choice of α reflects how cautious you want to be. The smaller the α-level, the stronger the evidence must be to reject H_0. To avoid bias in the decision-making process, you select α *before* analyzing the data.

Example 6.5

Examples of Decisions about H_0 Let's use $\alpha = 0.05$ to guide us in making a decision about H_0 for the examples of this section. Example 6.2 (page 145) tested H_0: $\mu = 4.0$ about mean political ideology. With sample mean $\bar{y} = 4.089$, the *P*-value was 0.20. The *P*-value is not small, so if truly $\mu = 4.0$, it would not be unusual to observe $\bar{y} = 4.089$. Since $P = 0.20 > 0.05$, there is insufficient evidence to reject H_0. It is believable that the population mean ideology was 4.0.

Example 6.4 tested H_0: $\mu = 0$ about mean weight gain for teenage girls suffering from anorexia. The *P*-value was 0.018. Since $P = 0.018 < 0.05$, there is sufficient evidence to reject H_0 in favor of H_a: $\mu > 0$. We conclude that the treatment results in an increase in mean weight. Such a conclusion is sometimes phrased as "The increase in mean weight is *statistically significant* at the 0.05 level." Since $P = 0.018$ is *not* less than 0.010, the result is *not* statistically significant at the 0.010 level. In fact, *the P-value is the smallest level for α at which the results are statistically significant*. So, with *P*-value $= 0.018$, we reject H_0 if $\alpha = 0.02$ or 0.05 or 0.10, but not if $\alpha = 0.010$ or 0.001. ■

Table 6.4 summarizes significance tests for population means.

TABLE 6.4: The Five Parts of Significance Tests for Population Means

1. **Assumptions**
   Quantitative variable
   Randomization
   Normal population (robust, especially for two-sided H_a, large n)
2. **Hypotheses**
   H_0: $\mu = \mu_0$
   H_a: $\mu \neq \mu_0$ (or H_a: $\mu > \mu_0$ or H_a: $\mu < \mu_0$)
3. **Test statistic**
   $t = \dfrac{\bar{y} - \mu_0}{se}$, where $se = \dfrac{s}{\sqrt{n}}$
4. ***P*-value**
   With the t distribution, use
   P = Two-tail probability for H_a: $\mu \neq \mu_0$
   P = Probability to right of observed t-value for H_a: $\mu > \mu_0$
   P = Probability to left of observed t-value for H_a: $\mu < \mu_0$
5. **Conclusion**
   Report P-value. Smaller P provides stronger evidence against H_0 and supporting H_a. Can reject H_0 if $P \leq \alpha$-level

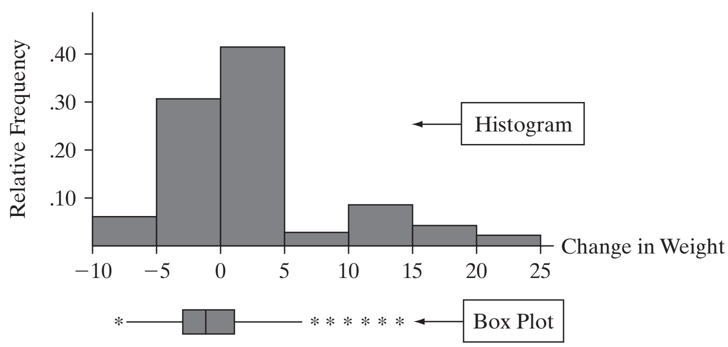
ROBUSTNESS FOR VIOLATIONS OF NORMALITY ASSUMPTION

The t test for a mean assumes that the population distribution is normal. This ensures that the sampling distribution of the sample mean $\bar{y}$ is normal (even for small n) and, after using s to estimate σ in finding the se, the t test statistic has the t distribution. As n increases, this assumption of a normal population becomes less important. We've seen that when n is roughly about 30 or higher, an approximate normal sampling distribution occurs for $\bar{y}$ regardless of the population distribution, by the Central Limit Theorem.

From Section 5.3 (page 113), a statistical method is ***robust*** if it performs adequately even when an assumption is violated. *Two-sided* inferences for a mean using the t distribution are robust against violations of the normal population assumption. Even if the population is not normal, two-sided t tests and confidence intervals still work quite well. The test does not work so well for a one-sided test with small n when the population distribution is highly skewed.

Figure 6.5 shows a histogram and a box plot of the data from the anorexia study of Example 6.4 (page 148). They suggest skew to the right. The box plot highlights (as outliers) six girls who had considerable weight gains. As just mentioned, a two-sided

FIGURE 6.5: Histogram and Box Plot of Weight Change for Anorexia Sufferers



t test works quite well even if the population distribution is skewed. However, this plot makes us wary about using a one-sided test, since the sample size is not large ($n = 29$). Given this and the discussion in the previous subsection about one-sided versus two-sided tests, we're safest with that study to report a two-sided P-value of 0.035. Also, the median may be a more relevant summary for these data.

6.3 Significance Test for a Proportion

For a categorical variable, the parameter is the population proportion for a category. For example, a significance test could analyze whether a majority of the population support legalizing same-sex marriage by testing H_0: $\pi = 0.50$ against H_a: $\pi > 0.50$, where π is the population proportion π supporting it. The test for a proportion, like the test for a mean, finds a P-value for a test statistic that measures the number of standard errors a point estimate falls from a H_0 value.

THE FIVE PARTS OF A SIGNIFICANCE TEST FOR A PROPORTION

1. Assumptions

Like other tests, this test assumes that the data are obtained using randomization. The sample size must be sufficiently large that the sampling distribution of $\hat{\pi}$ is approximately normal. For the most common case, in which the H_0 value of π is 0.50, a sample size of at least 20 is sufficient.[3]

2. Hypotheses

The null hypothesis of a test about a population proportion has the form

$$H_0: \pi = \pi_0, \quad \text{such as} \quad H_0: \pi = 0.50.$$

Here, π_0 denotes a particular proportion value between 0 and 1, such as 0.50. The most common alternative hypothesis is

$$H_a: \pi \neq \pi_0, \quad \text{such as} \quad H_a: \pi \neq 0.50.$$

This *two-sided* alternative states that the population proportion *differs* from the value in H_0. The *one-sided* alternatives

$$H_a: \pi > \pi_0 \quad \text{and} \quad H_a: \pi < \pi_0$$

apply when the researcher predicts a deviation in a certain direction from the H_0 value.

3. Test Statistic

From Section 5.2, the sampling distribution of the sample proportion $\hat{\pi}$ has mean π and standard error $\sqrt{\pi(1-\pi)/n}$. When H_0 is true, $\pi = \pi_0$, so the standard error is $se_0 = \sqrt{\pi_0(1-\pi_0)/n}$. We use the notation se_0 to indicate that this is the standard error under the presumption that H_0 is true.

The test statistic is

$$z = \frac{\hat{\pi} - \pi_0}{se_0}, \quad \text{where} \quad se_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}.$$

[3] Section 6.7, which presents a small-sample test, gives a precise guideline.

This measures the number of standard errors that the sample proportion $\hat{\pi}$ falls from π_0. When H_0 is true, the sampling distribution of the z test statistic is approximately the standard normal distribution.

The test statistic has a similar form as in tests for a mean.

Form of Test Statistic in Test for a Proportion

$$z = \frac{\text{Estimate of parameter} - \text{Null hypothesis value of parameter}}{\text{Standard error of estimate}}$$

Here, the estimate $\hat{\pi}$ of the proportion replaces the estimate $\bar{y}$ of the mean, and the null hypothesis proportion π_0 replaces the null hypothesis mean μ_0.

Note that in the standard error formula, $\sqrt{\pi(1-\pi)/n}$, we substitute the null hypothesis value π_0 for the population proportion π. The parameter values in sampling distributions for tests presume that H_0 is true, since the P-value is based on that presumption. This is why, for tests, we use $se_0 = \sqrt{\pi_0(1-\pi_0)/n}$ rather than the estimated standard error, $se = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$. If we instead used the estimated se, the normal approximation for the sampling distribution of z would be poorer. This is especially true for proportions close to 0 or 1. By contrast, the confidence interval method does not have a hypothesized value for π, so that method uses the estimated se rather than a H_0 value.

4. *P*-Value

The P-value is a one- or two-tail probability, as in tests for a mean, except using the standard normal distribution rather than the t distribution. For H_a: $\pi \neq \pi_0$, P is the two-tail probability. See Figure 6.6. This probability is double the single-tail probability beyond the observed z-value.

For one-sided alternatives, the P-value is a one-tail probability. Since H_a: $\pi > \pi_0$ predicts that the population proportion is *larger* than the H_0 value, its P-value is the probability *above* (i.e., to the right) of the observed z-value. For H_a: $\pi < \pi_0$, the P-value is the probability *below* (i.e., to the left) of the observed z-value.

5. Conclusion

As usual, the smaller the P-value, the more strongly the data contradict H_0 and support H_a. When we need to make a decision, we reject H_0 if $P \leq \alpha$ for a prespecified α-level such as 0.05.

Example 6.6

Reduce Services, or Raise Taxes? These days, whether at the local, state, or national level, government often faces the problem of not having enough money to pay for the various services that it provides. One way to deal with this problem is to raise taxes. Another way is to reduce services. Which would you prefer? When the Florida Poll recently asked a random sample of 1200 Floridians, 52% (624 of the 1200) said raise taxes and 48% said reduce services.

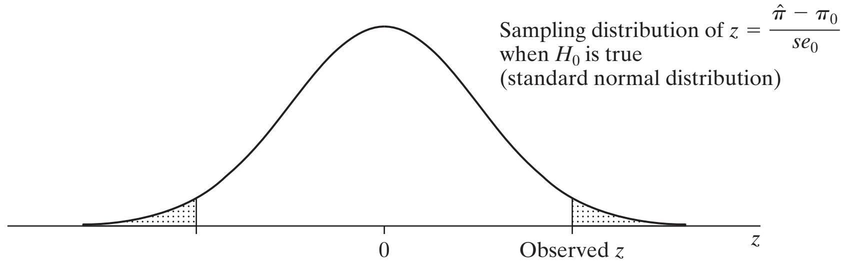


FIGURE 6.6: Calculation of P-Value in Testing H_0: $\pi = \pi_0$ against H_a: $\pi \neq \pi_0$. The two-sided alternative hypothesis uses a two-tail probability.

Let π denote the population proportion in Florida who would choose raising taxes rather than reducing services. If $\pi < 0.50$, this is a minority of the population, whereas if $\pi > 0.50$, it is a majority. To analyze whether π is in either of these ranges, we test H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$.

The estimate of π is $\hat{\pi} = 0.52$. Presuming H_0: $\pi = 0.50$ is true, the standard error of $\hat{\pi}$ is

$$se_0 = \sqrt{\frac{\pi_0(1 - \pi_0)}{n}} = \sqrt{\frac{(0.50)(0.50)}{1200}} = 0.0144.$$

The value of the test statistic is

$$z = \frac{\hat{\pi} - \pi_0}{se_0} = \frac{0.52 - 0.50}{0.0144} = 1.386.$$

The two-tail P-value is about $P = 0.17$. If H_0 is true (i.e., if $\pi = 0.50$), the probability equals 0.17 that sample results would be as extreme in one direction or the other as in this sample.

This P-value is not small, so there is not much evidence against H_0. It seems believable that $\pi = 0.50$. With an α-level such as 0.05, since $P = 0.17 > 0.05$, we would not reject H_0. We cannot determine whether those favoring raising taxes rather than reducing services are a majority or minority of the population. ■

We can conduct the test using software. Table 6.5 shows some output (edited) using the free software R applied to the number in the category, n, and the null value π_0. With Stata, you can do this for a variable in a data file, or also directly using the summary statistics as shown in Table 6.6 with the command `prtesti`. The test is also easy to conduct with an Internet applet.[4]

TABLE 6.5: R Software for Performing a Significance Test about a Proportion

```
> prop.test(624, 1200, p=0.50, alt="two.sided", correct=FALSE)

data:  624 out of 1200, null probability 0.5
p-value = 0.1659
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval: 0.4917142 0.5481581
sample estimates:  p  0.52
```

TABLE 6.6: Stata Software for Performing a Significance Test about a Proportion

```
. prtesti 1200 0.52 0.50 // provide n, sample prop., H0 prop.

One-sample test of proportion          x: Number of obs =     1200
  Variable |    Mean   Std. Err.          [95% Conf. Interval]
         x |     .52   .0144222            .491733      .548267
    p = proportion(x)                             z =   1.3856
Ho: p = 0.5
     Ha: p < 0.5           Ha: p != 0.5            Ha: p > 0.5
 Pr(Z < z) = 0.9171    Pr(|Z| > |z|) = 0.1659    Pr(Z > z) = 0.0829
```

[4] For example, with the *Inference for a Proportion* applet at www.artofstat.com/webapps.html.

NEVER "ACCEPT H_0"

In Example 6.6 about raising taxes or reducing services, the P-value of 0.17 was not small. So, H_0: $\pi = 0.50$ is plausible. In this case, the conclusion is sometimes reported as "Do not reject H_0," since the data do not contradict H_0.

It is better to say "Do not reject H_0" than "Accept H_0." The population proportion has many plausible values besides the number in H_0. For instance, the software output above reports a 95% confidence interval for the population proportion π as (0.49, 0.55). This interval shows a range of plausible values for π. Even though insufficient evidence exists to conclude that $\pi \neq 0.50$, it is improper to conclude that $\pi = 0.50$.

In summary, H_0 contains a single value for the parameter. When the P-value is larger than the α-level, saying "Do not reject H_0" instead of "Accept H_0" emphasizes that that value is merely one of *many* believable values. Because of sampling variability, there is a range of believable values, so we can never accept H_0. The reason "accept H_a" terminology is permissible for H_a is that when the P-value is sufficiently small, the entire range of believable values for the parameter falls within the range of values that H_a specifies.

EFFECT OF SAMPLE SIZE ON P-VALUES

In Example 6.6 on raising taxes or cutting services, suppose $\hat{\pi} = 0.52$ had been based on $n = 4800$ instead of $n = 1200$. The standard error then decreases to 0.0072 (half as large), and you can verify that the test statistic $z = 2.77$. This has two-sided P-value $= 0.006$. That P-value provides strong evidence against H_0: $\pi = 0.50$ and suggests that a majority support raising taxes rather than cutting services. In that case, though, the 95% confidence interval for π equals (0.506, 0.534). This indicates that π is quite close to 0.50 in practical terms.

A given difference between an estimate and the H_0 value has a smaller P-value as the sample size increases. The larger the sample size, the more certain we can be that sample deviations from H_0 are indicative of true population deviations. In particular, notice that even a small departure between $\hat{\pi}$ and π_0 (or between $\bar{y}$ and μ_0) can yield a small P-value if the sample size is very large.

6.4 Decisions and Types of Errors in Tests

When we need to decide whether the evidence against H_0 is strong enough to reject it, we reject H_0 if $P \leq \alpha$, for a prespecified α-level. Table 6.7 summarizes the two possible conclusions for α-level = 0.05. The null hypothesis is either "rejected" or "not rejected." If H_0 is rejected, then H_a is accepted. If H_0 is not rejected, then H_0 is plausible, but other parameter values are also plausible. Thus, H_0 is never "accepted." In this case, results are inconclusive, and the test does not identify either hypothesis as more valid.

TABLE 6.7: Possible Decisions in a Significance Test with α-Level = 0.05

	Conclusion	
P-Value	H_0	H_a
$P \leq 0.05$	Reject	Accept
$P > 0.05$	Do not reject	Do not accept

It is better to report the P-value than to indicate merely whether the result is "statistically significant." Reporting the P-value has the advantage that the reader can tell whether the result is significant at any level. The P-values of 0.049 and 0.001 are both "significant at the 0.05 level," but the second case provides much stronger evidence than the first case. Likewise, P-values of 0.049 and 0.051 provide, in practical terms, the same amount of evidence about H_0. It is a bit artificial to call one result "significant" and the other "nonsignificant." Some software places the symbol * next to a test statistic that is significant at the 0.05 level, ** next to a test statistic that is significant at the 0.01 level, and *** next to a test statistic that is significant at the 0.001 level.

TYPE I AND TYPE II ERRORS FOR DECISIONS

Because of sampling variability, decisions in tests always have some uncertainty. The decision could be erroneous. The two types of potential errors are conventionally called *Type I* and *Type II* errors.

Type I and Type II Errors

When H_0 is true, a ***Type I error*** occurs if H_0 is rejected.
When H_0 is false, a ***Type II error*** occurs if H_0 is not rejected.

The two possible decisions cross-classified with the two possibilities for whether H_0 is true generate four possible results. See Table 6.8.

TABLE 6.8: The Four Possible Results of Making a Decision in a Significance Test. Type I and Type II errors are the incorrect decisions.

		Decision	
		Reject H_0	Do Not Reject H_0
Condition of H_0	H_0 true	Type I error	Correct decision
	H_0 false	Correct decision	Type II error

REJECTION REGIONS: STATISTICALLY SIGNIFICANT TEST STATISTIC VALUES

The collection of test statistic values for which the test rejects H_0 is called the ***rejection region***. For example, the rejection region for a test of level $\alpha = 0.05$ is the set of test statistic values for which $P \leq 0.05$.

For two-sided tests about a proportion, the two-tail P-value is ≤ 0.05 whenever the test statistic $|z| \geq 1.96$. In other words, the rejection region consists of values of z resulting from the estimate falling at least 1.96 standard errors from the H_0 value.

THE α-LEVEL IS THE PROBABILITY OF TYPE I ERROR

When H_0 is true, let's find the probability of Type I error. Suppose $\alpha = 0.05$. We've just seen that for the two-sided test about a proportion, the rejection region is $|z| \geq 1.96$. So, the probability of rejecting H_0 is exactly 0.05, because the probability of the values in this rejection region under the standard normal curve is 0.05. But this is precisely the α-level.

> The probability of a Type I error is the α-level for the test.

With $\alpha = 0.05$, if H_0 is true, the probability equals 0.05 of making a Type I error and rejecting H_0. We control P(Type I error) by the choice of α. The more serious the consequences of a Type I error, the smaller α should be. In practice, $\alpha = 0.05$ is most common, just as an error probability of 0.05 is most common with confidence intervals (i.e, 95% confidence). However, this may be too high when a decision has serious implications.

For example, consider a criminal legal trial of a defendant. Let H_0 represent innocence and H_a represent guilt. The jury rejects H_0 and judges the defendant to be guilty if it decides the evidence is sufficient to convict. A Type I error, rejecting a true H_0, occurs in convicting a defendant who is actually innocent. In a murder trial, suppose a convicted defendant may receive the death penalty. Then, if a defendant is actually innocent, we would hope that the probability of conviction is much smaller than 0.05.

When we make a decision, we do not know whether we have made a Type I or Type II error, just as we do not know whether a particular confidence interval truly contains the parameter value. However, we can control the probability of an incorrect decision for either type of inference.

AS P(TYPE I ERROR) GOES DOWN, P(TYPE II ERROR) GOES UP

In an ideal world, Type I or Type II errors would not occur. However, errors do happen. We've all read about defendants who were convicted but later determined to be innocent. When we make a decision, why don't we use an extremely small P(Type I error), such as $\alpha = 0.000001$? For instance, why don't we make it almost impossible to convict someone who is really innocent?

When we make α smaller in a significance test, we need a smaller P-value to reject H_0. It then becomes harder to reject H_0. But this means that it will also be harder even if H_0 is false. The stronger the evidence required to convict someone, the more likely we will fail to convict defendants who are actually guilty. In other words, the smaller we make P(Type I error), the larger P(Type II error) becomes, that is, failing to reject H_0 even though it is false.

If we tolerate only an extremely small P(Type I error), such as $\alpha = 0.000001$, the test may be unlikely to reject H_0 even if it is false—for instance, unlikely to convict someone even if they are guilty. This reasoning reflects the fundamental relation:

- The smaller P(Type I error) is, the larger P(Type II error) is.

For instance, in an example in Section 6.6, when P(Type I error) = 0.05 we'll find that P(Type II error) = 0.02, but when P(Type I error) decreases to 0.01, P(Type II error) increases to 0.08. Except in Section 6.6, we shall not find P(Type II error), as it is beyond our scope. In practice, making a decision requires setting only α, the P(Type I error).

Section 6.6 shows that P(Type II error) depends on just how far the true parameter value falls from H_0. If the parameter is nearly equal to the value in H_0, P(Type II error) is relatively high. If it falls far from H_0, P(Type II error) is relatively low. The farther the parameter falls from the H_0 value, the less likely the sample is to result in a Type II error.

For a fixed P(Type I error), P(Type II error) depends also on the sample size n. The larger the sample size, the more likely we are to reject a false H_0. To keep both P(Type I error) and P(Type II error) at low levels, it may be necessary to use a very large sample size. The P(Type II error) may be quite large when the sample size is small, unless the parameter falls quite far from the H_0 value.

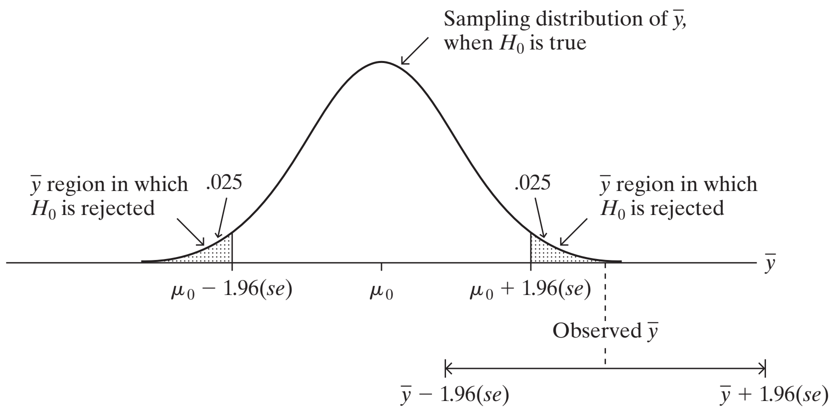
EQUIVALENCE BETWEEN CONFIDENCE INTERVALS AND TEST DECISIONS

We now elaborate on the equivalence for means[5] between decisions from two-sided tests and conclusions from confidence intervals, first alluded to in Example 6.3 (page 147). Consider the significance test of

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

When $P < 0.05$, H_0 is rejected at the $\alpha = 0.05$ level. When n is large (so the t distribution is essentially the same as the standard normal), this happens when the test statistic $t = (\bar{y} - \mu_0)/se$ is greater in absolute value than 1.96, that is, when $\bar{y}$ falls more than $1.96(se)$ from μ_0. But if this happens, then the 95% confidence interval for μ, namely, $\bar{y} \pm 1.96(se)$, does not contain the null hypothesis value μ_0. See Figure 6.7. These two inference procedures are consistent.

FIGURE 6.7: Relationship between Confidence Interval and Significance Test. For large n, the 95% confidence interval does not contain the H_0 value μ_0 when the sample mean falls more than 1.96 standard errors from μ_0, in which case the test statistic $|t| > 1.96$ and the P-value < 0.05.



Significance Test Decisions and Confidence Intervals

> In testing H_0: $\mu = \mu_0$ against H_a: $\mu \neq \mu_0$, when we reject H_0 at the 0.05 α-level, the 95% confidence interval for μ does not contain μ_0. The 95% confidence interval consists of those μ_0 values for which we do not reject H_0 at the 0.05 α-level.

In Example 6.2 about mean political ideology (page 145), the P-value for testing H_0: $\mu = 4.0$ against H_a: $\mu \neq 4.0$ was $P = 0.20$. At the $\alpha = 0.05$ level, we do not reject H_0: $\mu = 4.0$. It is believable that $\mu = 4.0$. Example 6.3 (page 147) showed that a 95% confidence interval for μ is (3.95, 4.23), which contains $\mu_0 = 4.0$.

Rejecting H_0 at a particular α-level is equivalent to the confidence interval for μ with the same error probability not containing μ_0. For example, if a 99% confidence interval does not contain 0, then we would reject H_0: $\mu = 0$ in favor of H_a: $\mu \neq 0$ at the $\alpha = 0.01$ level with the test. The α-level is P(Type I error) for the test and the probability that the confidence interval method does not contain the parameter.

---

[5] This equivalence also holds for proportions when we use the two-sided test of Section 6.3 and the confidence interval method presented in Exercise 5.77.

MAKING DECISIONS VERSUS REPORTING THE *P*-VALUE

The approach to hypothesis testing that incorporates a formal decision with a fixed *P*(Type I error) was developed by the statisticians Jerzy Neyman and Egon Pearson in the late 1920s and early 1930s. In summary, this approach formulates null and alternative hypotheses, selects an α-level for the *P*(Type I error), determines the rejection region of test statistic values that provide enough evidence to reject H_0, and then makes a decision about whether to reject H_0 according to what is actually observed for the test statistic value. With this approach, it's not even necessary to find a *P*-value. The choice of α-level determines the rejection region, which together with the test statistic determines the decision.

The alternative approach of finding a *P*-value and using it to summarize evidence against a hypothesis is due to the great British statistician R. A. Fisher. He advocated merely reporting the *P*-value rather than using it to make a formal decision about H_0. Over time, this approach has gained favor, especially since software can now report precise *P*-values for a wide variety of significance tests.

This chapter has presented an amalgamation of the two approaches (the decision-based approach using an α-level and the *P*-value approach), so you can interpret a *P*-value yet also know how to use it to make a decision when that is needed. These days, most research articles merely report the *P*-value rather than a decision about whether to reject H_0. From the *P*-value, readers can view the strength of evidence against H_0 and make their own decision, if they want to.

6.5 Limitations of Significance Tests

A significance test makes an inference about whether a parameter differs from the H_0 value and about its direction from that value. In practice, we also want to know whether the parameter is sufficiently different from the H_0 value to be practically important. In this section, we'll learn that a test does not tell us as much as a confidence interval about practical importance.

STATISTICAL SIGNIFICANCE VERSUS PRACTICAL SIGNIFICANCE

It is important to distinguish between *statistical significance* and *practical significance*. A small *P*-value, such as $P = 0.001$, is highly statistically significant. It provides strong evidence against H_0. It does not, however, imply an *important* finding in any practical sense. The small *P*-value merely means that if H_0 were true, the observed data would be very unusual. It does not mean that the true parameter value is far from H_0 in practical terms.

Example 6.7

Mean Political Ideology for All Americans The political ideology $\bar{y} = 4.089$ reported in Example 6.2 (page 145) refers to a sample of Hispanic Americans. We now consider the entire 2014 GSS sample who responded to the political ideology question. For a scoring of 1.0 through 7.0 for the ideology categories with 4.0 = moderate, the $n = 2575$ observations have $\bar{y} = 4.108$ and standard deviation $s = 4.125$. On the average, political ideology was the same for the entire sample as it was for Hispanics alone.[6]

As in Example 6.2, we test H_0: $\mu = 4.0$ against H_a: $\mu \neq 4.0$ to analyze whether the population mean differs from the moderate ideology score of 4.0. Now,

[6] And it seems stable over time, equaling 4.13 in 1980, 4.16 in 1990, and 4.10 in 2000.

$se = s/\sqrt{n} = 1.425/\sqrt{2575} = 0.028$, and

$$t = \frac{\bar{y} - \mu_0}{se} = \frac{4.108 - 4.0}{0.028} = 3.85.$$

The two-sided P-value is $P = 0.0001$. There is *very* strong evidence that the true mean exceeds 4.0, that is, that the true mean falls on the conservative side of moderate. But, on a scale of 1.0 to 7.0, 4.108 is close to the moderate score of 4.0. Although the difference of 0.108 between the sample mean of 4.108 and the H_0 mean of 4.0 is highly significant statistically, the magnitude of this difference is very small in practical terms. The mean response on political ideology for all Americans is essentially a moderate one. ■

In Example 6.2, the sample mean of $\bar{y} = 4.1$ for $n = 369$ Hispanic Americans had a P-value of $P = 0.20$, not much evidence against H_0. But now with $\bar{y} = 4.1$ based on $n = 2575$, we have instead found $P = 0.0001$. This is highly *statistically significant*, but not *practically significant*. For practical purposes, a mean of 4.1 on a scale of 1.0 to 7.0 for political ideology does not differ from 4.00.

A way of summarizing practical significance is to measure the ***effect size*** by the number of standard deviations (*not* standard errors) that $\bar{y}$ falls from μ_0. In this example, the estimated effect size is $(4.108 - 4.0)/1.425 = 0.08$. This is a tiny effect. Whether a particular effect size is small, medium, or large depends on the substantive context, but an effect size of about 0.2 or less in absolute value is usually not practically important.

SIGNIFICANCE TESTS ARE LESS USEFUL THAN CONFIDENCE INTERVALS

We've seen that, with large sample sizes, P-values can be small even when the point estimate falls near the H_0 value. The size of P merely summarizes the extent of evidence about H_0, not how far the parameter falls from H_0. Always inspect the difference between the estimate and the H_0 value to gauge the practical implications of a test result.

Null hypotheses containing single values are rarely true. That is, rarely is the parameter *exactly* equal to the value listed in H_0. With sufficiently large samples, so that a Type II error is unlikely, these hypotheses will normally be rejected. What is more relevant is whether the parameter is sufficiently different from the H_0 value to be of practical importance.

Although significance tests can be useful, most statisticians believe they are overemphasized in social science research. It is preferable to construct confidence intervals for parameters instead of performing only significance tests. A test merely indicates whether the particular value in H_0 is plausible. It does not tell us which other potential values are plausible. The confidence interval, by contrast, displays the entire set of believable values. It shows the extent to which reality may differ from the parameter value in H_0 by showing whether the values in the interval are far from the H_0 value. Thus, it helps us to determine whether rejection of H_0 has practical importance.

To illustrate, for the complete political ideology data in Example 6.7, a 95% confidence interval for μ is

$$\bar{y} \pm 1.96(se) = 4.108 \pm 1.96(0.028), \text{ or } (4.05, 4.16).$$

This indicates that the difference between the population mean and the moderate score of 4.0 is very small. Although the P-value of $P = 0.0001$ provides very strong evidence against H_0: $\mu = 4.0$, in practical terms the confidence interval shows that

H_0 is not wrong by much. By contrast, if $\bar{y}$ had been 6.108 (instead of 4.108), the 95% confidence interval would equal (6.05, 6.16). This indicates a substantial practical difference from 4.0, the mean response being near the conservative score rather than the moderate score.

When a P-value is not small but the confidence interval is quite wide, this forces us to realize that the parameter might well fall far from H_0 even though we cannot reject it. This also supports why it does not make sense to "accept H_0," as we discussed on page 155.

The remainder of the text presents significance tests for a variety of situations. It is important to become familiar with these tests, if for no other reason than their frequent use in social science research. However, we'll also introduce confidence intervals that describe how far reality is from the H_0 value.

SIGNIFICANCE TESTS AND *P*-VALUES CAN BE MISLEADING

We've seen it is improper to "accept H_0." We've also seen that statistical significance does not imply practical significance. Here are other ways that results of significance tests can be misleading:

- **It is misleading to report results only if they are statistically significant.** Some research journals have the policy of publishing results of a study only if the P-value ≤ 0.05. Here's a danger of this policy: Suppose there truly is no effect, but 20 researchers independently conduct studies. We would expect about $20(0.05) = 1$ of them to obtain significance at the 0.05 level merely by chance. (When H_0 is true, about 5% of the time we get a P-value below 0.05 anyway.) If that researcher then submits results to a journal but the other 19 researchers do not, the article published will be a Type I error. It will report an effect when there really is not one.
- **Some tests may be statistically significant just by chance.** You should never scan software output for results that are statistically significant and report only those. If you run 100 tests, even if all the null hypotheses are correct, you would expect to get P-values ≤ 0.05 about $100(0.05) = 5$ times. Be skeptical of reports of significance that might merely reflect ordinary random variability.
- **It is incorrect to interpret the *P*-value as the probability that H_0 is true.** The P-value is P(test statistic takes value like observed or even more extreme), presuming that H_0 is true. It is not $P(H_0 \text{ true})$. Classical statistical methods calculate probabilities about variables and statistics (such as test statistics) that vary randomly from sample to sample, not about parameters. Statistics have sampling distributions, parameters do not. In reality, H_0 is not a matter of probability. It is either true or not true. We just don't know which is the case.
- **True effects are often smaller than reported estimates.** Even if a statistically significant result is a real effect, the true effect may be smaller than reported. For example, often several researchers perform similar studies, but the results that receive attention are the most extreme ones. The researcher who decides to publicize the result may be the one who got the most impressive sample result, perhaps way out in the tail of the sampling distribution of all the possible results. See Figure 6.8.

Example 6.8

Are Many Medical "Discoveries" Actually Type I Errors? In medical research studies, suppose that an actual population effect exists only 10% of the time. Suppose also that when an effect truly exists, there is a 50% chance of making a Type II error

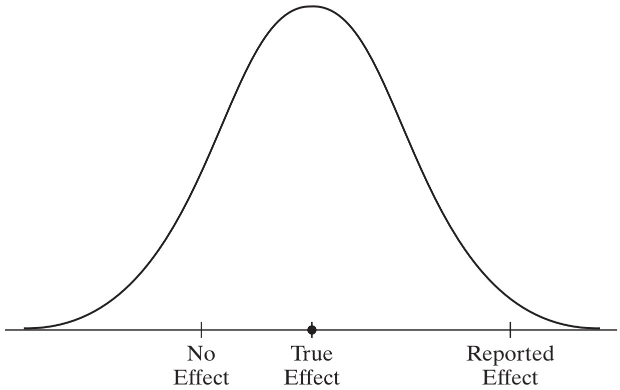


FIGURE 6.8: When Many Researchers Conduct Studies about a Hypothesis, the Statistically Significant Result Published in a Journal and Reported by Popular Media Often Overestimates the True Effect

and failing to detect it. These were the hypothetical percentages used in an article in a medical journal.[7] The authors noted that many medical studies have a high Type II error rate because they are not able to use a large sample size. Assuming these rates, could a substantial percentage of medical "discoveries" actually be Type I errors?

Figure 6.9 is a ***tree diagram*** showing what we would expect with 1000 medical studies that test various hypotheses. If a population effect truly exists only 10% of the time, this would be the case for 100 of the 1000 studies. We do not obtain a small enough P-value to detect this true effect 50% of the time, that is, in 50 of these 100 studies. An effect will be reported for the other 50 of the 100 that do truly have an effect. For the 900 cases in which there truly is no effect, with the usual significance level of 0.05 we expect 5% of the 900 studies to incorrectly reject H_0. This happens for $(0.05)900 = 45$ studies. In summary, of the 1000 studies, we expect 50 to report an effect that is truly there, but we also expect 45 to report an effect that does not actually exist. So, a proportion of $45/(45 + 50) = 0.47$ of medical studies that report effects are actually reporting Type I errors.

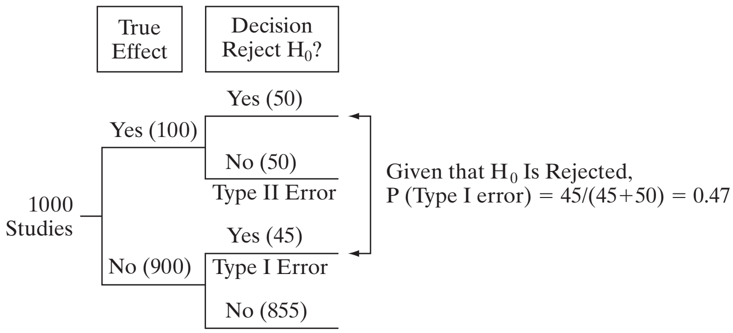


FIGURE 6.9: Tree Diagram of 1000 Hypothetical Medical Studies. This assumes a population effect truly exists 10% of the time and a 50% chance of a Type II error when an effect truly exists.

The moral is to be skeptical when you hear reports of new medical advances. The true effect may be weaker than reported, or there may actually be no effect at all. ■

Related to this is the *publication bias* that occurs when results of some studies never appear in print because they did not obtain a small enough P-value to seem important. One investigation[8] of this reported that 94% of medical studies that had positive results found their way into print whereas only 14% of those with disappointing or uncertain results did.

[7] By J. Sterne, G. Smith, and D. R. Cox, *BMJ*, vol. 322 (2001), pp. 226–231.
[8] Reported in *The New York Times*, January 17, 2008.

6.6 Finding P(Type II Error)*

We've seen that decisions in significance tests have two potential types of error. A Type I error results from rejecting H_0 when it is actually true. Given that H_0 is true, the probability of a Type I error is the α-level of the test; when $\alpha = 0.05$, the probability of rejecting H_0 equals 0.05.

When H_0 is false, a Type II error results from *not* rejecting it. This probability has more than one value, because H_a contains a range of possible values. Each value in H_a has its own P(Type II error). This section shows how to calculate P(Type II error) at a particular value.

Example 6.9

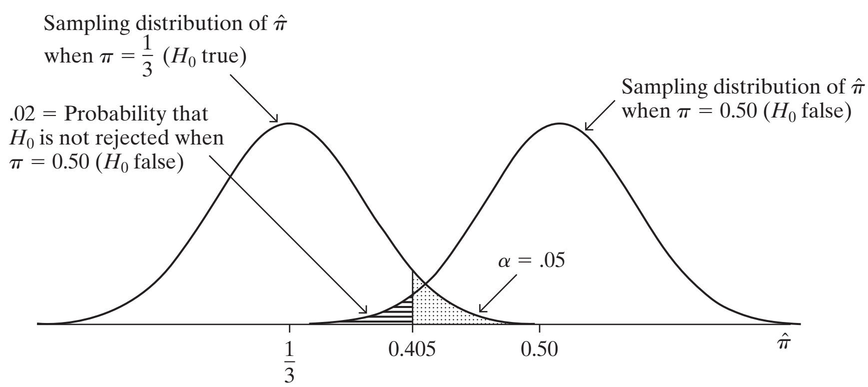
Testing whether Astrology Really Works One scientific test of the pseudoscience astrology used the following experiment[9]: For each of 116 adult subjects, an astrologer prepared a horoscope based on the positions of the planets and the moon at the moment of the person's birth. Each subject also filled out a California Personality Index survey. For each adult, his or her birth data and horoscope were shown to an astrologer with the results of the personality survey for that adult and for two other adults randomly selected from the experimental group. The astrologer was asked which personality chart of the three subjects was the correct one for that adult, based on their horoscope.

Let π denote the probability of a correct prediction by an astrologer. If the astrologers' predictions are like random guessing, then $\pi = 1/3$. To test this against the alternative that the guesses are better than random guessing, we can test H_0: $\pi = 1/3$ against H_a: $\pi > 1/3$. The alternative hypothesis reflects the astrologers' belief that they can predict better than random guessing. In fact, the National Council for Geocosmic Research, which supplied the astrologers for the experiment, claimed π would be 0.50 or higher. So, let's find P(Type II error) if actually $\pi = 0.50$, for an $\alpha = 0.05$-level test. That is, if actually $\pi = 0.50$, we'll find the probability that we'd fail to reject H_0: $\pi = 1/3$.

To determine this, we first find the sample proportion values for which we would not reject H_0. For the test of H_0: $\pi = 1/3$, the sampling distribution of $\hat{\pi}$ is the curve shown on the left in Figure 6.10. With $n = 116$, this curve has standard error

$$se_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{n}} = \sqrt{\frac{(1/3)(2/3)}{116}} = 0.0438.$$

FIGURE 6.10: Calculation of P(Type II Error) for Testing H_0: $\pi = 1/3$ against H_a: $\pi > 1/3$ at $\alpha = 0.05$ Level, when True Proportion Is $\pi = 0.50$ and $n = 116$. A Type II error occurs if $\hat{\pi} < 0.405$, since then the P-value > 0.05 even though H_0 is false.



[9] S. Carlson, *Nature*, vol. 318 (1985), pp. 419–425.

For H_a: $\pi > 1/3$, the P-value equals 0.05 if the test statistic $z = 1.645$. That is, 1.645 is the z-score that has a right-tail probability of 0.05. So, we *fail to reject* H_0, getting a P-value *above* 0.05, if $z < 1.645$. In other words, we fail to reject H_0: $\pi = 1/3$ if the sample proportion $\hat{\pi}$ falls less than 1.645 standard errors above 1/3, that is, if

$$\hat{\pi} < 1/3 + 1.645(se_0) = 1/3 + 1.645(0.0438) = 0.405.$$

So, the right-tail probability above 0.405 is $\alpha = 0.05$ for the curve on the left in Figure 6.10.

To find P(Type II error) if π actually equals 0.50, we must find $P(\hat{\pi} < 0.405)$ when $\pi = 0.50$. This is the left-tail probability *below* 0.405 for the curve on the right in Figure 6.10, which is the curve that applies when $\pi = 0.50$. When $\pi = 0.50$, the standard error for a sample size of 116 is $\sqrt{[(0.50)(0.50)]/116} = 0.0464$. (This differs a bit from se_0 for the test statistic, which uses 1/3 instead of 0.50 for π.) For the normal distribution with a mean of 0.50 and standard error of 0.0464, the $\hat{\pi}$ value of 0.405 has a z-score of

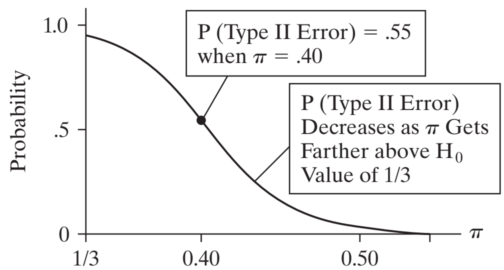
$$z = \frac{0.405 - 0.50}{0.0464} = -2.04.$$

The probability that $\hat{\pi} < 0.405$ is the probability that a standard normal variable falls below −2.04, which equals 0.02. So, for a sample of size 116, the probability of not rejecting H_0: $\pi = 1/3$ is 0.02, if in fact $\pi = 0.50$. In other words, if astrologers truly had the predictive power they claimed, the chance of failing to detect this with this experiment would have only been about 0.02. To see what actually happened in the experiment, see Exercise 6.17. ■

This probability calculation of P(Type II error) was rather involved. Such calculations can be performed easily with an Internet applet.[10]

The probability of Type II error increases when the parameter value moves closer to H_0. To verify this, you can check that P(Type II error) = 0.55 at $\pi = 0.40$. So, if the parameter falls near the H_0 value, there may be a substantial chance of failing to reject H_0. Likewise, the farther the parameter falls from H_0, the less likely a Type II error. Figure 6.11 plots P(Type II error) for the various π values in H_a.

FIGURE 6.11: Probability of Type II Error for Testing H_0: $\pi = 1/3$ against H_a: $\pi > 1/3$ at $\alpha = 0.05$ Level, Plotted for the Potential π Values in H_a



For a fixed α-level and alternative parameter value, P(Type II error) decreases when the sample size increases. If you can obtain more data, you will be less likely to make this sort of error.

TESTS WITH SMALLER α HAVE GREATER P(TYPE II ERROR)

As explained on page 157, the smaller $\alpha = P$(Type I error) is in a test, the larger P(Type II error) is. To illustrate, suppose the astrology study in Example 6.9 used

[10] See, for example, the *Errors and Power* applet at www.artofstat.com/webapps.html.

$\alpha = 0.01$. Then, when $\pi = 0.50$ you can verify that $P(\text{Type II error}) = 0.08$, compared to $P(\text{Type II error}) = 0.02$ when $\alpha = 0.05$.

The reason that extremely small values are not normally used for α, such as $\alpha = 0.0001$, is that P(Type II error) is too high. We may be unlikely to reject H_0 even if the parameter falls far from the null hypothesis. In summary, for fixed values of other factors,

- P(Type II error) decreases as
  - the parameter value is farther from H_0.
  - the sample size increases.
  - P(Type I error) increases.

THE POWER OF A TEST

When H_0 is false, you want the probability of rejecting H_0 to be high. The probability of rejecting H_0 is called the ***power*** of the test. For a particular value of the parameter from within the H_a range,

$$\text{Power} = 1 - P(\text{Type II error}).$$

In Example 6.9, for instance, the test of H_0: $\pi = 1/3$ has $P(\text{Type II error}) = 0.02$ at $\pi = 0.50$. Therefore, the power of the test at $\pi = 0.50$ is $1 - 0.02 = 0.98$.

The power increases for values of the parameter falling farther from the H_0 value. Just as the curve for P(Type II error) in Figure 6.11 decreases as π gets farther above $\pi_0 = 1/3$, the curve for the power increases.

In practice, studies should ideally have high power. Before granting financial support for a planned study, research agencies often expect principal investigators to show that reasonable power (usually, at least 0.80) exists at values of the parameter that are practically significant.

When you read that results of a study are not statistically significant, be skeptical if no information is given about the power. The power may be low, especially if n is small or the effect is not large.

6.7 Small-Sample Test for a Proportion—The Binomial Distribution*

For a population proportion π, Section 6.3 presented a significance test that is valid for relatively large samples. The sampling distribution of the sample proportion $\hat{\pi}$ is then approximately normal, which justifies using a z test statistic.

For small n, the sampling distribution of $\hat{\pi}$ occurs at only a few points. If $n = 5$, for example, the only possible values for the sample proportion $\hat{\pi}$ are $0, 1/5, 2/5, 3/5, 4/5$, and 1. A continuous approximation such as the normal distribution is inappropriate. In addition, the closer π is to 0 or 1 for a given sample size, the more skewed the actual sampling distribution becomes.

This section introduces a small-sample test for proportions. It uses the most important probability distribution for discrete variables, the *binomial distribution*.

THE BINOMIAL DISTRIBUTION

For categorical data, often the following three conditions hold:

- Each observation falls into one of two categories.
- The probabilities for the two categories are the same for each observation. We denote the probabilities by π for category 1 and $(1 - \pi)$ for category 2.

- The outcomes of successive observations are independent. That is, the outcome for one observation does not depend on the outcomes of other observations.

Flipping a coin repeatedly is a prototype for these conditions. For each flip, we observe whether the outcome is head (category 1) or tail (category 2). The probabilities of the outcomes are the same for each flip (0.50 for each if the coin is balanced). The outcome of a particular flip does not depend on the outcome of other flips.

Now, for n observations, let x denote the number of them that occur in category 1. For example, for $n = 5$ coin flips, $x =$ number of heads could equal 0, 1, 2, 3, 4, or 5. When the observations satisfy the above three conditions, the probability distribution of x is the ***binomial distribution***.

The binomial variable x is discrete, taking one of the integer values $0, 1, 2, \ldots, n$. The formula for the binomial probabilities follows:

Probabilities for a Binomial Distribution

Denote the probability of category 1, for each observation, by π. For n independent observations, the probability that x of the n observations occur in category 1 is

$$P(x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

The symbol $n!$ is called ***n factorial***. It represents $n! = 1 \times 2 \times 3 \times \cdots \times n$. For example, $1! = 1,\ 2! = 1 \times 2 = 2,\ 3! = 1 \times 2 \times 3 = 6$, and so forth. Also, $0!$ is defined to be 1.

For particular values for π and n, substituting the possible values for x into the formula for $P(x)$ provides the probabilities of the possible outcomes. The sum of the probabilities equals 1.0.

Example 6.10

Gender and Selection of Managerial Trainees Example 6.1 (page 139) discussed a case involving potential bias against females in selection of management trainees for a large supermarket chain. The pool of employees is half female and half male. The company claims to have selected 10 trainees at random from this pool. If they are truly selected at random, how many females would we expect to be chosen?

The probability that any one person selected is a female is $\pi = 0.50$, the proportion of available trainees who are female. Similarly, the probability that any one person selected is male is $(1 - \pi) = 0.50$. Let $x =$ number of females selected. This has the binomial distribution with $n = 10$ and $\pi = 0.50$. For each x between 0 and 10, the probability that x of the 10 people selected are female equals

$$P(x) = \frac{10!}{x!(10-x)!}(0.50)^x(0.50)^{10-x}, \quad x = 0, 1, 2, \ldots, 10.$$

For example, the probability that no females are chosen ($x = 0$) is

$$P(0) = \frac{10!}{0!10!}(0.50)^0(0.50)^{10} = (0.50)^{10} = 0.001.$$

(Recall that any number raised to the power of 0 equals 1.) The probability that exactly one female is chosen is

$$P(1) = \frac{10!}{1!9!}(0.50)^1(0.50)^9 = 10(0.50)(0.50)^9 = 0.010.$$

Table 6.9 lists the entire binomial distribution for $n = 10$, $\pi = 0.50$. Binomial probabilities for any n, π, and x value can be found with Internet applets.[11]

TABLE 6.9: The Binomial Distribution for $n = 10$, $\pi = 0.50$. The binomial variable x can take any value between 0 and 10.

x	$P(x)$	x	$P(x)$
0	0.001	6	0.205
1	0.010	7	0.117
2	0.044	8	0.044
3	0.117	9	0.010
4	0.205	10	0.001
5	0.246		

In Table 6.9, the probability is about 0.98 that x falls between 2 and 8, inclusive. The least likely values for x are 0, 1, 9, and 10, which have a combined probability of only 0.022. If the sample were randomly selected, somewhere between about two and eight females would probably be selected. It is especially unlikely that none or 10 would be selected.

The probabilities for females determine those for males. For instance, the probability that 9 of the 10 people selected are male equals the probability that 1 of the 10 selected is female. ■

PROPERTIES OF THE BINOMIAL DISTRIBUTION

The binomial distribution is perfectly symmetric only when $\pi = 0.50$. In Example 6.10, for instance, since the population proportion of females equals 0.50, $x = 10$ has the same probability as $x = 0$.

The sample proportion $\hat{\pi}$ relates to the binomial variable x by

$$\hat{\pi} = x/n.$$

For example, for $x = 1$ female chosen out of $n = 10$, $\hat{\pi} = 1/10 = 0.10$. The sampling distribution of $\hat{\pi}$ is also symmetric when $\pi = 0.50$. When $\pi \neq 0.50$, the distribution is skewed, the degree of skew increasing as π gets closer to 0 or 1. Figure 6.12 illustrates this. When $\pi = 0.10$, for instance, the sample proportion $\hat{\pi}$ can't fall much below 0.10 since it must be positive, but it could fall considerably above 0.10.

Like the normal distribution, the binomial can be characterized by its mean and standard deviation.

Binomial Mean and Standard Deviation

The binomial distribution for x = how many of n observations fall in a category having probability π has mean and standard deviation

$$\mu = n\pi \quad \text{and} \quad \sigma = \sqrt{n\pi(1-\pi)}.$$

For example, suppose the probability of a female in any one selection for management training is 0.50, as the supermarket chain claims. Then, out of 10 trainees, we expect $\mu = n\pi = 10(0.50) = 5.0$ females.

We've seen (in Sections 5.2 and 6.3) that the sampling distribution of the sample proportion $\hat{\pi}$ has mean π and standard error $\sqrt{\pi(1-\pi)/n}$. To obtain these formulas,

[11] For example, with the *Binomial Distribution* applet at `www.artofstat.com/webapps.html`.

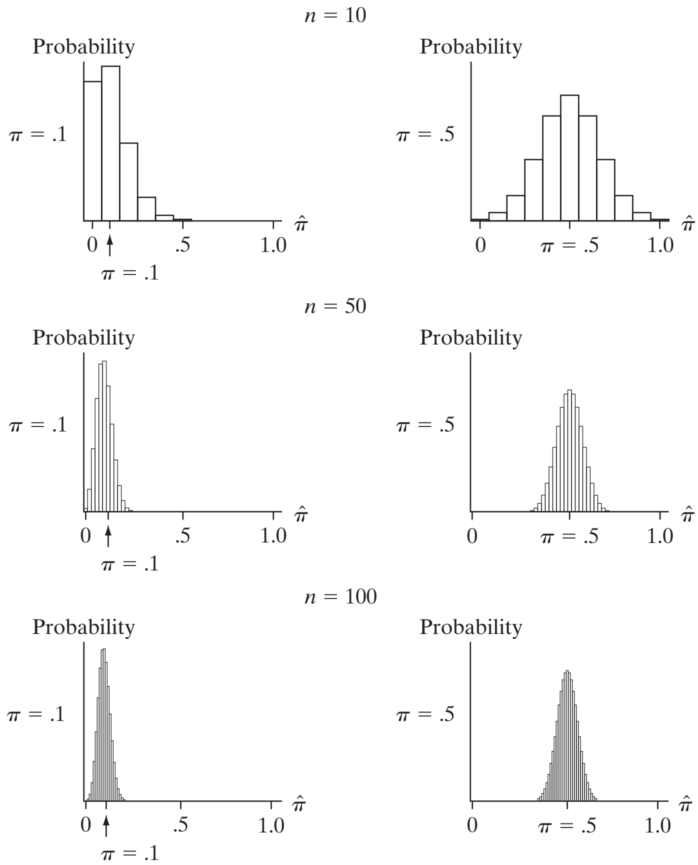


FIGURE 6.12: Sampling Distribution of $\hat{\pi}$ when $\pi = 0.10$ or 0.50, for $n = 10, 50, 100$

we divide the binomial mean $\mu = n\pi$ and standard deviation $\sigma = \sqrt{n\pi(1-\pi)}$ by n, since $\hat{\pi}$ divides x by n.

Example 6.11 **How Much Variability Can an Exit Poll Show?** Example 4.6 (page 78) discussed an exit poll of 1824 voters for the 2014 California gubernatorial election. Let x denote the number in the exit poll who voted for Jerry Brown. In the population of more than 7 million voters, 60.0% voted for him. If the exit poll was randomly selected, then the binomial distribution for x has $n = 1824$ and $\pi = 0.600$. The distribution is described by

$$\mu = 1824(0.600) = 1094, \quad \sigma = \sqrt{1824(0.600)(0.400)} = 21.$$

Almost certainly, x would fall within three standard deviations of the mean. This is the interval from 1031 to 1157. In fact, in that exit poll, 1104 people of the 1824 sampled reported voting for Brown. ■

THE BINOMIAL TEST

The binomial distribution and the sampling distribution of $\hat{\pi}$ are approximately normal for large n. This approximation is the basis of the large-sample test of Section 6.3. How large is "large"? A guideline is that the expected number of observations should be at least 10 for both categories. For example, if $\pi = 0.50$, we need at least about $n = 20$, because then we expect $20(0.50) = 10$ observations in one category and $20(1-0.50) = 10$ in the other category. For testing H_0: $\pi = 0.90$ or H_0: $\pi = 0.10$, we need $n \geq 100$. The sample size requirement reflects the fact that a symmetric bell shape for the sampling distribution of $\hat{\pi}$ requires larger sample sizes when π is near 0 or 1 than when π is near 0.50.

If the sample size is not large enough to use the normal test, we can use the binomial distribution directly. Refer to Example 6.10 (page 166) about potential gender discrimination. For random sampling, the probability π that a person selected for management training is female equals 0.50. If there is bias against females, then $\pi < 0.50$. Thus, we can test the company's claim of random sampling by testing

$$H_0: \pi = 0.50 \qquad \text{versus} \qquad H_a: \pi < 0.50.$$

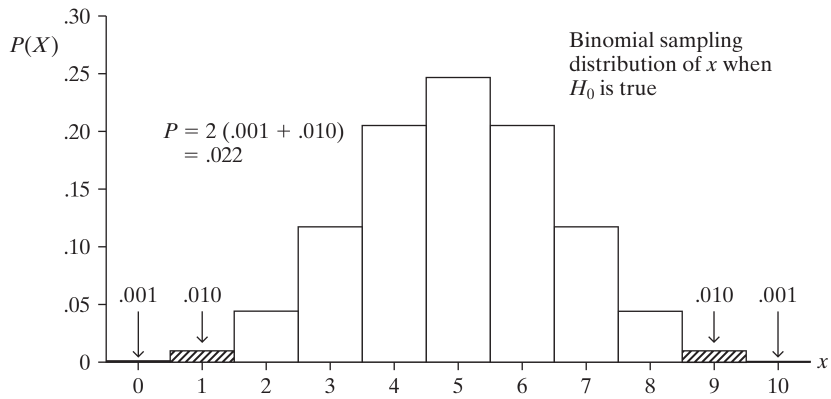
Of the 10 employees chosen for management training, let x denote the number of women. Under H_0, the sampling distribution of x is the binomial distribution with $n = 10$ and $\pi = 0.50$. Table 6.9 tabulated it. As in Example 6.1 (page 139), suppose $x = 1$. The P-value is then the left-tail probability of an outcome at least this extreme; that is, $x = 1$ or 0. From Table 6.9, the P-value is

$$P = P(0) + P(1) = 0.001 + 0.010 = 0.011.$$

If the company selected trainees randomly, the probability of choosing one or fewer females is only 0.011. This result provides evidence against the null hypothesis of a random selection process. We can reject H_0 for $\alpha = 0.05$, though not for $\alpha = 0.010$.

Even if we suspect bias in a particular direction, the most even-handed way to perform a test uses a two-sided alternative. For H_a: $\pi \neq 0.50$, the P-value is $2(0.011) = 0.022$. This is a two-tail probability of the outcome that one or fewer *of either sex* is selected. Figure 6.13 shows the formation of this P-value.

FIGURE 6.13: Calculation of P-Value in Testing H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$, when $n = 10$ and $x = 1$



The assumptions for the binomial test are the three conditions for the binomial distribution. Here, the conditions are satisfied. Each observation has only two possible outcomes, female or male. The probability of each outcome is the same for each selection, 0.50 for selecting a female and 0.50 for selecting a male (under H_0). For random sampling, the outcome of any one selection does not depend on any other one.

6.8 Chapter Summary

Chapter 5 and this chapter have introduced two methods for using sample data to make inferences about populations—***confidence intervals*** and ***significance tests***.

A confidence interval provides a range of plausible values for a parameter. A significance test judges whether a particular value for the parameter is plausible. Both methods utilize the sampling distribution of the estimator of the parameter.

Significance tests have five parts:

1. ***Assumptions:***
   - Tests for *means* apply with quantitative variables whereas tests for *proportions* apply with categorical variables.
   - Tests assume *randomization*, such as a random sample.
   - Large-sample tests about proportions require no assumption about the population distribution, because the Central Limit Theorem implies approximate normality of the sampling distribution of the sample proportion.
   - Tests for means use the t distribution, which assumes the population distribution is normal. In practice, two-sided tests (like confidence intervals) are *robust* to violations of the normality assumption.
2. ***Null and alternative hypotheses*** about the parameter: The null hypothesis has the form H_0: $\mu = \mu_0$ for a mean and H_0: $\pi = \pi_0$ for a proportion. Here, μ_0 and π_0 denote values hypothesized for the parameters, such as 0.50 in H_0: $\pi = 0.50$. The most common alternative hypothesis is *two sided*, such as H_a: $\pi \neq 0.50$. Hypotheses such as H_a: $\pi > 0.50$ and H_a: $\pi < 0.50$ are *one sided*, designed to detect departures from H_0 in a particular direction.
3. A ***test statistic*** describes how far the point estimate falls from the H_0 value. The z statistic for proportions and t statistic for means measure the number of standard errors that the point estimate ($\hat{\pi}$ or $\bar{y}$) falls from the H_0 value.
4. The ***P-value*** describes the evidence about H_0 in probability form.
   - We calculate the P-value by presuming that H_0 is true. It equals the probability that the test statistic equals the observed value or a value even more extreme.
   - The "more extreme" results are determined by the alternative hypothesis. For two-sided H_a, the P-value is a two-tail probability.
   - Small P-values result when the point estimate falls far from the H_0 value, so that the test statistic is large. When the P-value is small, it would be unusual to observe such data if H_0 were true. The smaller the P-value, the stronger the evidence against H_0.
5. A ***conclusion*** based on the sample evidence about H_0: We report and interpret the P-value. When we need to make a decision, we reject H_0 when the P-value is less than or equal to a fixed α-level (such as $\alpha = 0.05$). Otherwise, we cannot reject H_0.

When we make a decision, two types of errors can occur.

- When H_0 is true, a Type I error results if we reject it.
- When H_0 is false, a Type II error results if we fail to reject it.

The choice of α, the cutoff point for the P-value in making a decision, equals P(Type I error). Normally, we choose small values such as $\alpha = 0.05$ or 0.01. For fixed α, P(Type II error) decreases as the distance increases between the parameter and the H_0 value or as the sample size increases.

Table 6.10 summarizes the five parts of the tests this chapter presented.

Sample size is a critical factor in both estimation and significance tests. With small sample sizes, confidence intervals are wide, making estimation imprecise.

TABLE 6.10: Summary of Significance Tests for Means and Proportions

Parameter	Mean	Proportion
1. Assumptions	Random sample, quantitative variable, normal population	Random sample, categorical variable, null expected counts at least 10
2. Hypotheses	H_0: $\mu = \mu_0$ H_a: $\mu \neq \mu_0$ H_a: $\mu > \mu_0$ H_a: $\mu < \mu_0$	H_0: $\pi = \pi_0$ H_a: $\pi \neq \pi_0$ H_a: $\pi > \pi_0$ H_a: $\pi < \pi_0$
3. Test statistic	$t = \frac{\bar{y} - \mu_0}{se}$ with $se = \frac{s}{\sqrt{n}}$, $df = n - 1$	$z = \frac{\hat{\pi} - \pi_0}{se_0}$ with $se_0 = \sqrt{\pi_0(1 - \pi_0)/n}$
4. P-value	Two-tail probability in sampling distribution for two-sided test (H_0: $\mu \neq \mu_0$ or H_a: $\pi \neq \pi_0$); one-tail probability for one-sided test	
5. Conclusion	Reject H_0 if P-value $\leq \alpha$-level such as 0.05	

Small sample sizes also make it difficult to reject false null hypotheses unless the true parameter value is far from the null hypothesis value. P(Type II error) may be high for parameter values of interest.

This chapter presented significance tests about a single parameter for a single variable. In practice, it is usually artificial to have a particular fixed number for the H_0 value of a parameter. One of the few times this happens is when the response score results from taking a difference of two values, such as the change in weight in Example 6.4 (page 148). In that case, $\mu_0 = 0$ is a natural baseline. Significance tests much more commonly refer to comparisons of means for two samples than to a fixed value of a parameter for a single sample. The next chapter shows how to compare means or proportions for two groups.

Exercises

Practicing the Basics

6.1. For (a)–(c), is it a null hypothesis, or an alternative hypothesis?

(a) In Canada, the proportion of adults who favor legalized gambling equals 0.50.

(b) The proportion of all Canadian college students who are regular smokers now is less than 0.20 (the value it was 10 years ago).

(c) The mean IQ of all students at Lake Wobegon High School is larger than 100.

(d) Introducing notation for a parameter, state the hypotheses in (a)–(c) in terms of the parameter values.

6.2. You want to know whether adults in your country think the ideal number of children is equal to 2, or higher or lower than that.

(a) Define notation and state the null and alternative hypotheses for studying this.

(b) For responses in a recent GSS to the question "What do you think is the ideal number of children to have?" software shows results:

```
----------------------------------------------------
Test of mu = 2.0 vs mu not = 2.0

Variable    n    Mean   StDev  SE Mean    T   P-value
Children  1302  2.490  0.850  0.0236   20.80  0.0000
----------------------------------------------------
```

Report the test statistic value, and show how it was obtained from other values reported in the table.

(c) Explain what the P-value represents, and interpret its value.

6.3. For a test of H_0: $\mu = 0$ against H_a: $\mu \neq 0$ with $n = 1000$, the t test statistic equals 1.04.

(a) Find the P-value, and interpret it.

(b) Suppose $t = -2.50$ rather than 1.04. Find the P-value. Does this provide stronger, or weaker, evidence against the null hypothesis? Explain.

(c) When $t = 1.04$, find the P-value for (i) H_a: $\mu > 0$, (ii) H_a: $\mu < 0$.

6.4. The P-value for a test about a mean with $n = 25$ is $P = 0.05$.