CENTRAL LIMIT THEOREM (CLT) FOR A SAMPLE PROPORTION Suppose that a random sample of size n is taken from a process or a large population (whose size is at least $20 \times n$) in which the probability of success is π. Then the sampling distribution of the sample proportion $\hat{p}$ can be approximately modeled by a normal probability curve with mean equal to π and standard deviation equal to $\sqrt{\pi(1-\pi)/n}$. This shape becomes more and more normal as the sample size n increases, and it is generally considered to be valid provided that $n\pi \geq 10$ and $n(1-\pi) \geq 10$.

INVESTIGATION 4.3.3 *COHEN V. BROWN UNIVERSITY*

In 1991, a suit was filed against Brown University after Brown terminated funding for its women's gymnastics and volleyball teams and its men's water polo and golf teams. The suit charged that Brown was violating Title IX of the Education Amendments of 1972, the federal law that prohibits sex discrimination by all educational institutions receiving federal funds. This requires men and women to have equivalent opportunities for participation. A main component of the plaintiff's case was that while 51% of the undergraduate student body was women, only 38% of the 897 students engaged in intercollegiate athletics were women.

If there is no gender discrepancy, then Title IX assumes that the proportion of women athletes should be similar to the proportion of women in the overall student body. However, we know the sample result can deviate from this population proportion "just by chance." Suppose we were randomly selecting students to be athletes, the question is whether this random process could lead to such a disparity in these proportions. Although we know the proportion of women in the population of all university students to be .51, we don't know the probability of an athlete being female. Let π refer to the probability of a Brown University athlete being female.

Test Statistic

Minitab will only report this approximate p-value to seven decimal places. Clearly, it is very small, and the z-score computed in (d) tells us that the observed sample proportion is more than 7 standard deviations from the hypothesized value!

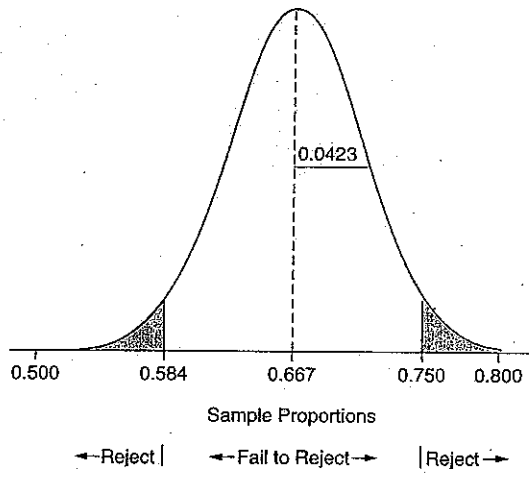
The calculation you carried out in (d) is referred to as the *test statistic*, which provides a standardized measure of the distance between our observed result and our hypothesized value. The p-value indicates how often we would obtain a test statistic at least this extreme when the null hypothesis is true.

Since we are working with a normal distribution, we know that an observation more than 2 or 3 standard deviations away from the mean is a surprising outcome.

In general, to test H_0: $\pi = \pi_0$, where π_0 indicates the hypothesized value of the population proportion, the test statistic is obtained through the formula:

$$z_0 = \frac{\text{observed} - \text{hypothesized}}{\text{standard deviation}} = \frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

DEFINITION The *rejection region* of a test of significance is the values of the sample statistic that lead us to reject the null hypothesis. The definition of the rejection region will depend on the level of significance specified. In the preceding example, for a 5% level of significance, the rejection region would be the values of $\hat{p}$ that are below .584 or above .750. These are the values that are far enough from 2/3 that we would obtain a two-sided p-value $\leq .05$ and would therefore reject the null hypothesis (that $\pi = 2/3$) at the 5% level (see figure). If the level of significance had been 1%, then the rejection region would have been $\hat{p} \leq .558$ and $\hat{p} \geq .776$. This rejection region is smaller, as we are requiring the sample result to be even more extreme in order to convince us to reject the null hypothesis.



An approximate C% confidence interval for a population proportion or process probability π can be calculated using the expression:

$$\hat{p} \pm z^* \sqrt{\hat{p}(1 - \hat{p})/n}$$

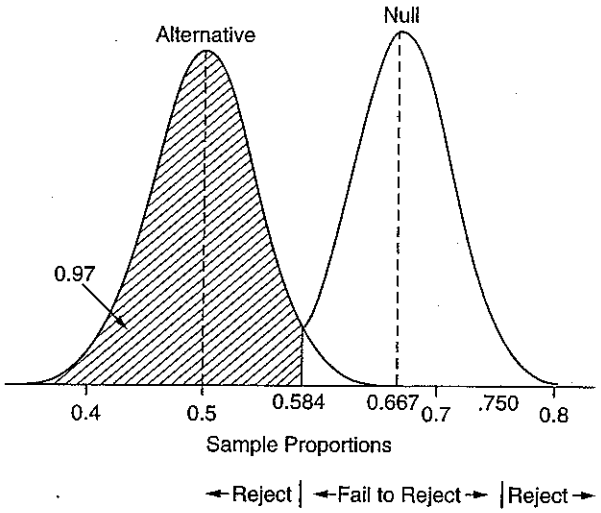
where $-z^*$ is the $(100 - C)/2$ percentile of a *standard normal* probability model. This procedure is considered valid when the data are randomly sampled from the population/process of interest and $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$, where $-z^*$ is the $(100 - C)/2$ percentile of a *standard normal*. This procedure is sometimes referred to as the *Wald interval*. It provides an alternative to the exact binomial confidence interval that you studied in Chapter 3.

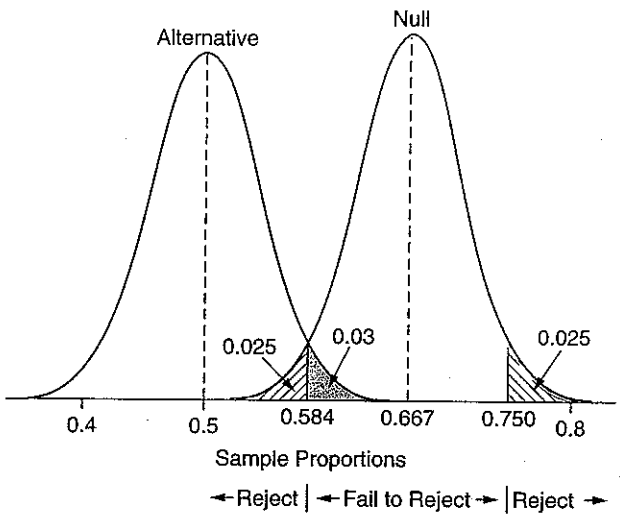
The *half-width* of the interval $z^* \sqrt{\hat{p}(1 - \hat{p})/n}$ is also referred to as the *margin of error*. So we can write the interval as *estimate* $\pm$ *margin of error* or *estimate* $\pm$ *critical value* $\times$ *SE(estimate)*, both of which will be common forms for confidence intervals that estimate other parameters, as you will study later.

We will interpret this interval as the plausible value for the parameter based on the observed sample statistic. There is a *duality* between the confidence interval and *two-sided* tests of significance: If a hypothesized value is rejected at the α level of significance, then it will not be included in the $100(1 - \alpha)\%$ confidence interval.

Discussion In (j), you found the *power* of the test against the alternative value of .5 by finding $P(\hat{p} \leq .584$ or $\hat{p} \geq .750$ assuming $\pi = .5) = .9693$. This means that if π actually equals .5 and we repeatedly sample from this population (or process), we will reject H_0: $\pi = 2/3$ in about 97% of samples.

As you saw in Chapter 3, the probability of a Type I error is equal to the level of significance. The probability of a Type II error is 1 minus the power, so in this case $P(\text{Type II error}) = 1 - .9693 = .0307$. We could have found this directly by finding $P(.584 \leq \hat{p} \leq .750$ when $\pi = .5)$. In order to do these calculations, you must first determine the *rejection region* based on the null hypothesis and the level of significance, and then see how often you are in (or not in) this rejection region for a particular alternative value of the parameter.





I. Discuss whether the power of the test will be larger, the same, or smaller for the following scenarios (*Hint:* Sketches should help! Note that the change in each scenario appears in bold.):

1. $n = 124$, $\boldsymbol{\alpha = .10}$, $\pi_a = .5$

2. $\boldsymbol{n = 250}$, $\alpha = .05$, $\pi_a = .5$

3. $n = 124$, $\alpha = .05$, $\boldsymbol{\pi_a = .6}$

Discussion You have now seen three ways to calculate a p-value in the binomial setting:

1. Approximate the p-value through simulation. This method has the advantage of emphasizing what a p-value represents: the fraction of times that such a result at least this far from the hypothesized value would occur by chance alone if the process were repeated over and over. Another advantage is that simulation can be used as a very flexible tool for approximating a p-value in almost any setting. One obvious disadvantage is that the p-value produced is only approximate. Another potential disadvantage is that simulation requires computer software to produce enough repetitions for the approximation to be reasonably accurate.
2. Calculate the exact p-value using the binomial probability model. This method has the virtue of producing an exact p-value. A disadvantage is that calculating the p-value can be cumbersome without technology, particularly when n is large. Another potential disadvantage, as compared to simulation, is that this method does not directly relate to the long-run interpretation of a p-value.
3. Approximate the p-value using the normal approximation to the binomial distribution. An advantage of this method is that calculating a z-score (test statistic) is not cumbersome, and this test statistic provides a nice measure of how far (how many standard deviations) the observed value is from the hypothesized value. This is a flexible idea that applies in many settings. However, to then convert the test statistic to a p-value requires a normal probability table or technology. Other disadvantages are that this method produces only an approximate p-value, and the approximation produced is only reasonable when certain conditions are satisfied (the validity of the normal approximation to the binomial).

TESTS OF SIGNIFICANCE STRUCTURE The preceding investigation outlined steps that you will follow in general in carrying out a test of significance.

1. ***Define the population parameter in words.*** What is the unknown quantity that you are trying to make a decision about?
2. ***State the null and alternative hypotheses about this parameter.*** We will always write the null hypothesis as "parameter = value" and the alternative hypothesis will be the same parameter symbol and the same value, but then you will choose among $<$, $>$, and $\neq$, depending on the research question. This choice should be made before you collect the data.
3. ***Check the "technical conditions" and sketch the sampling distribution.*** In the last chapter, you verified whether the sampling process was binomial, and to use the methods in this chapter you also have to first verify whether the normal approximation to the binomial is valid (the conditions of the Central Limit Theorem).
4. ***Calculate the test statistic (normal-based inference).*** This provides a measure of the distance between the sample result and the hypothesized value of the parameter, often in terms of "how many standard deviations" apart.
5. ***Calculate the p-value.*** This can be done directly from the sample result or from the test statistic and provides a measure of how likely you are to find a sample result at least this extreme (at least this many standard deviations away) when the null hypothesis is true, by chance alone.
6. ***Make a decision about the null hypothesis.*** You will either "reject H_0" or "fail to reject H_0," depending on the size of the p-value and the level of significance.
7. ***State your conclusion in context.*** You should never stop at step 6, but need to go back and answer the researcher's question (e.g., there is significant evidence that the proportion of women among athletes is smaller than the proportion of women among the general population at this university).