Chapter eight

Steps in a Statistical Test: Using the Binomial Distribution to Make Decisions About Hypotheses

Statistical assumptions

What Type of Measurement is Being Used?

Are Assumptions Made About the Population Distribution?

What Sampling Method is Being Used?

What are the Hypotheses?

Sampling distribution

Which Sampling Distribution is Appropriate?

Significance level

What is the Rejection Region?

Where is It Placed?

Should a One-Tailed or a Two-Tailed Test be Used?

Test statistic and decision

What is the Test Statistic?

How is a Final Decision Made?

D. Weisburd and C. Britt, *Statistics in Criminal Justice*, DOI 10.1007/978-1-4614-9170-5_8,

In the previous chapter, you saw how probability theory is used to identify the observed significance level in a test of statistical significance. But you cannot simply rely on mathematical calculations to determine whether to reject the null hypothesis. You must make sure at the outset that the methods used are appropriate to the problem examined. You must clearly state the assumptions made. You must define the specific hypotheses to be tested and the specific significance criteria to be used. It is best to take a careful step-by-step approach to tests of statistical significance. Using this approach, you will be much less likely to make serious mistakes in developing such tests.

In this chapter, we introduce the basic elements of this step-by-step approach. To place this approach in context, we illustrate each step with a specific research problem that can be addressed using the binomial distribution. Although we use the binomial distribution as an example, you should not lose sight of the fact that our purpose here is to establish a general model for presenting tests of statistical significance, which can be used whichever sampling distribution is chosen.

The Problem: The Impact of Problem-Oriented Policing on Disorderly Activity at Violent-Crime Hot Spots

In Jersey City, New Jersey, researchers developed a problem-oriented policing program directed at violent-crime hot spots.[1] Computer mapping techniques were used to identify places in the city with a very high level of violent-crime arrests or emergency calls to the police. Jersey City police officers, in cooperation with staff of the Rutgers University Center for Crime Prevention Studies, developed strategies to solve violent-crime

[1]See Anthony Braga, "Solving Violent Crime Problems: An Evaluation of the Jersey City Police Department's Pilot Program to Control Violent Crime Places," unpublished dissertation, Rutgers University, Newark, NJ, 1996.

Table 8.1 Results at Treatment and Control Locations Derived from Observations of Disorderly Behavior Before and After Intervention

TRIAL	PLACE	OUTCOME
1	**Journal Square East** Newport Mall	+
2	**Stegman & Ocean** Clerk & Carteret	+
3	**Glenwood & JFK** Journal Square West	−
4	**Bergen & Academy** Westside & Duncan	+
5	**Westside & Clendenny** Franklin & Palisade	+
6	**Belmont & Monticello** MLK & Wade	+
7	**MLK & Atlantic** Neptune & Ocean	+
8	**MLK & Armstrong** Ocean & Eastern	+
9	**Westside & Virginia** JFK & Communipaw	+
10	**Park & Prescott** Dwight & Bergen	+
11	**Old Bergen & Danforth** Bramhall & Arlington	+

Note: Experimental or treatment hot spots are listed in boldface type.
+ = Relative improvement in experimental locations
− = Relative improvement in control locations

problems at a sample of 11 places. The strategies followed a problem-oriented policing (POP) approach, in which police collect a wide variety of information about each hot spot, analyze that information to identify the source of the problem, develop tailor-made responses to do something about the problem, and finally assess whether their approach actually had an impact.[2]

The evaluation involved a number of different components. One part of the research sought to identify whether "disorderly" activity at the hot spots had declined during the period of the study. For example, the researchers wanted to see whether the number of loiterers or homeless people had been reduced as a result of the efforts of the police. The treatment areas were compared to a matched group, or control group, of similar but untreated violent-crime places. Table 8.1 presents the overall

[2]Problem-oriented policing is an important new approach to police work formulated by Herman Goldstein of the University of Wisconsin Law School. See H. Goldstein, *Problem-Oriented Policing* (New York: McGraw-Hill, 1990).

results of pre- and posttest comparisons of outcomes for the 11 matched pairs of locations. In 10 of the 11 pairs, the experimental hot spots (those receiving POP intervention) improved relative to the control locations.

The research question asked by the evaluator was whether the POP approach has an impact on disorderly activity at violent-crime hot spots. The statistical problem faced is that the 11 comparisons are only a sample of such comparisons. What conclusions can the researcher make regarding the larger population of violent-crime hot spots? To answer this question, we use a test of statistical significance. The specific test that is appropriate for our problem is based on the binomial sampling distribution.

Assumptions: Laying the Foundations for Statistical Inference

The first step in a test of statistical significance is to establish the **assumptions** on which the test is based. These assumptions form the foundation of a test. No matter how elegant the statistics used and the approach taken, if the assumptions on which they are built are not solid, then the whole structure of the test is brought into question.

Level of Measurement

Our first assumption is related to the type of measurement used. Different types of tests of statistical significance demand different levels of measurement.

Accordingly, it is important to state at the outset the type of measurement required by a test. For the binomial test, which is based on the binomial distribution, a nominal-level binary measure is required. A binary measure has only two possible outcomes, as was the case with the coin toss example in Chapter 7. The type of outcome measure used to evaluate the impact of problem-oriented policing on disorderly activity—whether the treatment hot spot improved (or got worse) relative to the control location—fits this assumption. In stating our assumptions (as is done at the end of this section), we include a specific definition of the level of measurement required:

Level of Measurement: Nominal binary scale.

Shape of the Population Distribution

The second assumption refers to the shape of the population distribution. In statistical inference, we are generally concerned with two types of tests. In the first type—termed **parametric tests**—we make an assumption about the shape of the population distribution. For example, in a number of tests we will examine in later chapters, there is a require-

ment that for the population to which you infer, the scores on the variable be normally distributed.

The second type of test of statistical significance does not make a specific assumption regarding the population distribution. These tests are called **nonparametric tests** or **distribution-free tests.** The advantage of nonparametric tests is that we make fewer assumptions. The disadvantage is that nonparametric tests do not allow us to analyze data at higher levels of measurement. They are generally appropriate only for nominal and ordinal scales. The binomial test is a nonparametric test. Accordingly, in stating our assumptions we write:

Population Distribution: No assumption made.

Sampling Method

The third assumption concerns the sampling method used. When we conduct a test of statistical significance, we want our sample to be a good representation of the population from which it is drawn. Put in statistical terms, we want our study to have high **external validity.**

Let's suppose you are interested in attitudes toward the death penalty. Would a sample of your friends provide an externally valid sample of all Americans? Clearly not, because a sample of only your friends is not likely to include age or ethnic or class differences that typify the U.S. population. Even if we used your friends as a sample of U.S. college students, we could still identify threats to the external validity of the study. Colleges have differing criteria for admission, so it is not likely that one college will be representative of all colleges. Even as a sample of students at your college, your friends may not provide a valid sample. They may be drawn primarily from a specific year of college or have other characteristics that make them attractive as friends but also mean that they are a poor representation of others in the college.

How can we draw a **representative sample?** The most straightforward approach is to choose cases at random from the population. This type of sampling is called **random sampling.** Random samples are assumed to have high external validity compared with what may be termed **convenience samples.** A convenience sample consists of whatever subjects are readily available to the researcher. Your friends form a convenience sample of students at your college or of all college students.

It is important to note that convenience samples are not always bad samples. For example, if you choose to examine prisoners in one prison on the assumption that prisoners there provide a cross section of the different types of prisoners in the United States, you might argue that it is a representative sample. However, if you use a convenience sample, such as prisoners drawn from a single prison, you must always be wary of potential threats to external validity. Convenience samples are prone to systematic biases precisely because they are convenient. The characteristics that make

them easy for the researcher to define are likely as well to differentiate them in one way or another from the population the researcher seeks to study.

Statistical tests of significance generally assume that the researcher has used a type of random sampling called **independent random sampling.** Independent random sampling requires not only that cases be identified at random, but also that the selection of cases be independent. As discussed in the previous chapter, two events are statistically independent when the occurrence of one does not affect the occurrence of the other. In sampling, this means that the choice of one case or group of cases will not have any impact on the choice of another case or group of cases. This is a useful assumption in assuring the external validity of a study because it prevents biases that might be brought into the process of sampling.

For example, suppose you want to select 1,000 prisoners from the population of all prisoners in the United States. Each time you select a prisoner for your sample, you use a random method of selection. However, prison officials have told you that if you select one prisoner from a cell then you cannot select any other prisoner from that cell. Accordingly, after each selection of a prisoner, you must remove all of his cellmates from your **sampling frame,** or universe of eligible cases. The result is that there are now systematic reasons why you might suspect that your sample is not representative of the population.

In order to ensure independent random sampling, the same population of cases must be used in drawing each case for a sample. As we discussed in Chapter 7, if we want each draw from a deck of cards to be independent, we have to return the card chosen on any specific draw to the deck. If we didn't replace the card, we would influence the likelihood of a specific card being chosen on the next draw from the deck. For example, if we started with a full deck of 52 cards, the likelihood of getting the queen of spades would be 1 in 52. However, if we drew, say, a jack of hearts and didn't return it to the deck, what would be the likelihood of getting a queen of spades on our next draw? This time we would have only 51 cards to draw from, so our likelihood would change to 1 in 51. In order to gain a fully independent random sample, we must use a method of sampling called **sampling with replacement.** This means that we must use the same population each time we select a case. For every selection, the sampling frame must remain exactly the same. In this way, we can ensure that the choice of one case cannot have any impact on the choice of another.

Though this method ensures independence, it also means that a particular case may be selected more than once. For example, suppose you choose a particular prisoner as case number five in your sample. Because you must use the same sampling frame each time you select a case, that prisoner is returned to the sampling frame after selection. Later

in your study, you might choose that prisoner again. Accordingly, while sampling with replacement, or returning sampled cases to the sampling frame after each selection, makes statistical sense, it often does not make practical sense when you are carrying out research in the real world. If you are conducting an interview study, for example, independent random sampling would allow individuals to be interviewed more than once. It is likely that subjects would find it strange to be reinterviewed using the same interview schedule. Moreover, their responses would likely be influenced by their knowledge of the survey. Similarly, if a subject or place is chosen twice in a study that involves a specific treatment or intervention, then that subject or place should be given the treatment after each selection. Here there is the difficulty that it may be harmful to provide the treatment more than once.

Even when there are no specific practical barriers to sampling with replacement, it is difficult to explain to practitioners or even many researchers why an individual may appear twice in the same sample. As a result, many, if not most, criminal justice studies do not replace individuals in the sampling frame once they have been selected. Although this represents a formal violation of the assumptions of your test, in most cases its impact on your test result is negligible. This is because samples are generally very small relative to populations, and thus in practice there is little chance of selecting a case more than once even when sampling with replacement. If, however, your sample reaches one-fifth or more of the size of your population, you may want to include a correction factor in your test.[3]

For this test of statistical significance, we assume that researchers in the Jersey City POP study sampled cases randomly from a large population of hot spots during the sample selection month. Because it would not have been practical to implement treatments more than once at any site, the researchers did not sample with replacement.

---

[3]The correction factor adjusts your test to account for the fact that you have not allowed individuals to be selected from the population more than once. Not including a correction factor makes it more difficult to reject the null hypothesis. That is, the inclusion of a correction factor will make it easier for you to reject the null hypothesis. One problem criminal justice scholars face in using a correction factor is that they often want to infer to populations that are beyond their sampling frame. For example, a study of police patrol at hot spots in a particular city may sample 50 of 200 hot spots in the city during a certain month. However, researchers may be interested in making inferences to hot spots generally in the city (not just those that exist in a particular month) or even to hot spots in other places. For those inferences, it would be misleading to adjust the test statistic based on the small size of the sampling frame. For a discussion of how to correct for sampling without replacement, see Paul S. Levy and Stanley Lemeshow, *Sampling of Populations: Methods and Applications* (New York: Wiley, 1991).

The binomial test, however, like most tests of statistical significance examined in this book, assumes independent random sampling. Accordingly, in stating our assumptions, it is important to note both the requirement for this test and our failure to meet that requirement. Therefore we state our assumption:

Sampling Method: Independent random sampling (no replacement; sample is small relative to population).

Throughout this text, we state the assumptions of a test and then place any violations of assumptions in parentheses. This is good practice, as it will alert you to the fact that in many studies there are violations of one type or another of assumptions. Some of these violations are not important. For example, not sampling with replacement in this study does not affect the test outcome because the population of hot spots is assumed to be very large relative to the sample. However, you will sometimes find more serious violations of assumptions. In those cases, you will have to take a more critical view of the results of the test.

It is good practice to define not only the sampling method used but also the sampling frame of your study. In our example, we can make inferences based on our random sample to the population of hot spots in Jersey City during the month of sample selection. Accordingly, we state in our assumptions:

Sampling Frame: Hot spots of violent crime in one month in Jersey City.

Our sampling frame reminds us of the specific population to which our sample infers. However, researchers usually want to infer beyond the specific population identified by their sampling frame. For example, the population of interest for the POP study is likely to be hot spots throughout the year, not just those in a specific month. Researchers may even want to infer to violent-crime hot spots generally, not just those in Jersey City.

We cannot assume that our sample is a representative sample for these inferences based on our sampling method, since these populations did not constitute our sampling frame. However, we can ask whether our sample is likely to provide valid inferences to those populations. In the case of hot spots in Jersey City, we would need to question whether there is any reason to suspect that hot spots chosen in the month of study were different from those that would be found in other months of the year. For inferences to the population of hot spots in other locations, we would have to assume that Jersey City hot spots are similar to those in other places and would respond similarly to POP interventions. In making any inference beyond your sampling frame, you must try to identify all possible threats to external validity.

The Hypotheses

The final assumptions we make in a test of statistical inference refer to the hypotheses of our study. As discussed in Chapter 6, hypotheses are developed from the research questions raised in a project. Hypotheses must be stated before the researcher collects outcome data for a study. If hypotheses are stated only after data have been collected and analyzed, the researcher might be tempted to make changes in the hypotheses that unfairly affect the tests of statistical significance that are conducted.

As discussed in Chapter 6, the researcher ordinarily begins by defining the research hypothesis. In the problem-oriented policing study, we might state our research hypothesis in three different ways:

Hypothesis 1. Incivilities in treatment hot spots decline relative to incivilities in control hot spots after POP intervention.

Hypothesis 2. Incivilities in treatment hot spots increase relative to incivilities in control hot spots after POP intervention.

Hypothesis 3. The level of incivilities in treatment hot spots relative to incivilities in control hot spots changes after POP intervention.

Recall from Chapter 6 that we distinguish directional from nondirectional hypotheses. The first two research hypotheses are directional hypotheses because they specify the direction, or type of relationship, that is expected. For example, hypothesis 1 is concerned only with whether the POP program is successful in *reducing* incivilities. If the researcher adopts this hypothesis, then he or she is stating that the statistical test employed will not be concerned with the second hypothesis—that the intervention makes matters worse and *increases* incivilities. The third hypothesis is a nondirectional hypothesis. In this case, the researcher is interested in testing the possibility that the intervention improves hot spots or makes them worse.

In the POP study, researchers wanted to assess both positive and negative outcomes. Although they believed that problem-oriented policing should reduce incivilities at violent-crime hot spots, they did not want to preclude at the outset a finding that the program actually made matters worse. Accordingly, they used a nondirectional research hypothesis: "The level of incivilities in treatment hot spots relative to incivilities in control hot spots changes after POP intervention." The null hypothesis is "The level of incivilities in treatment hot spots does not change relative to incivilities in control hot spots after POP intervention."

In practice, the null hypothesis may be stated in terms of probabilities, just as we could state the coin toss hypothesis in the last chapter in terms of probabilities. In this study, the researchers examined (for each matched pair of hot spots) whether the hot spot that received the problem-oriented policing intervention improved or worsened relative to

the control location. The null hypothesis suggests that the treatment and control hot spots are equally likely to improve. Put in terms of probabilities, there is a 0.50 chance of success ($P = 0.50$) for the intervention under the null hypothesis. The research hypothesis represents all other possible outcomes ($P \neq 0.50$). Remember that our hypotheses are statements about the populations examined. Accordingly, in stating the hypotheses, we use symbols appropriate for population parameters—in this case P rather than p. Stating our assumptions, we write

Hypotheses:

H_0: The level of incivilities in treatment hot spots does not change relative to incivilities in control hot spots after POP intervention, $P = 0.50$.

H_1: The level of incivilities in treatment hot spots relative to incivilities in control hot spots changes after POP intervention, $P \neq 0.50$.

Stating All of the Assumptions

Our assumptions may be stated as follows:

Assumptions:

Level of Measurement: Nominal binary scale.

Population Distribution: No assumption made.

Sampling Method: Independent random sampling (no replacement; sample is small relative to population).

Sampling Frame: Hot spots of violent crime in one month in Jersey City.

Hypotheses:

H_0: The level of incivilities in treatment hot spots does not change relative to incivilities in control hot spots after POP intervention, $P = 0.50$.

H_1: The level of incivilities in treatment hot spots relative to incivilities in control hot spots changes after POP intervention, $P \neq 0.50$.

Selecting a Sampling Distribution

In stating our hypotheses, we already noted the specific requirements of the binomial sampling distribution. Now we must state why we have chosen the binomial distribution and identify the specific characteristics of the sampling distribution that will be used to assess the risk of falsely rejecting the null hypothesis in our problem-oriented policing example. Choosing a sampling distribution is one of the most important decisions that researchers make in statistical inference. As we will show in later chapters, there are a number of different types of sampling distributions. Moreover, as with the binomial distribution, a single type of sampling distribution may have different forms depending on the problem

examined. If the sampling distribution used is inappropriate for the research problem examined, then the conclusion reached will be suspect.

Because our measure is nominal and binary (see assumptions), we selected the binomial distribution for our test. The specific distribution that we use is based on our null hypothesis and the size of our sample. As illustrated in Chapter 7, the binomial distribution provides the likelihood of gaining a particular number of successes (heads in the example of the coin toss) in a fixed number of trials. In order to assess that likelihood, we also need to know what the probability of a success or failure is on any particular trial.

In our example, there are 11 trials, or 11 matched comparisons. Our null hypothesis states that the likelihood of a success for any comparison is 0.50. To build our sampling distribution, we apply the binomial formula to each of the 12 possible outcomes that could be gained in our study, under the assumption that $P = 0.50$. This is done in Table 8.2. The resulting distribution is presented in Table 8.3.

Table 8.2 Computation of Sampling Distribution of Success or Failure in 11 Trials

	$\binom{N}{r} = \frac{N!}{r!(N-r)!}$	$\binom{N}{r}p^r(1-p)^{N-r}$
0 successes	$\frac{39,916,800}{1(11-0)!} = \frac{39,916,800}{39,916,800} = 1$	$1(0.00049) = 0.00049$
1 success	$\frac{39,916,800}{1(11-1)!} = \frac{39,916,800}{3,628,800} = 11$	$11(0.00049) = 0.00537$
2 successes	$\frac{39,916,800}{2(11-2)!} = \frac{39,916,800}{725,760} = 55$	$55(0.00049) = 0.02686$
3 successes	$\frac{39,916,800}{6(11-3)!} = \frac{39,916,800}{241,920} = 165$	$165(0.00049) = 0.08057$
4 successes	$\frac{39,916,800}{24(11-4)!} = \frac{39,916,800}{120,960} = 330$	$330(0.00049) = 0.16113$
5 successes	$\frac{39,916,800}{120(11-5)!} = \frac{39,916,800}{86,400} = 462$	$432(0.00049) = 0.22638$*
6 successes	$\frac{39,916,800}{720(11-6)!} = \frac{39,916,800}{86,400} = 462$	$432(0.00049) = 0.22638$*
7 successes	$\frac{39,916,800}{5,040(11-7)!} = \frac{39,916,800}{120,960} = 330$	$330(0.00049) = 0.16113$
8 successes	$\frac{39,916,800}{40,320(11-8)!} = \frac{39,916,800}{241,920} = 165$	$165(0.00049) = 0.08057$
9 successes	$\frac{39,916,800}{362,880(11-9)!} = \frac{39,916,800}{725,760} = 55$	$55(0.00049) = 0.02686$
10 successes	$\frac{39,916,800}{3,628,800(11-10)!} = \frac{39,916,800}{3,628,800} = 11$	$11(0.00049) = 0.00537$
11 successes	$\frac{39,916,800}{39,916,800(11-11)!} = \frac{39,916,800}{39,916,800} = 1$	$1(0.00049) = 0.00049$

*Probabilities contain rounding error.

Table 8.3 Sampling Distribution of Success or Failure in 11 Trials

OUTCOME OF TRIALS	OVERALL PROBABILITY
0 successes	0.00049
1 success	0.00537
2 successes	0.02686
3 successes	0.08057
4 successes	0.16113
5 successes	0.22559
6 successes	0.22559
7 successes	0.16113
8 successes	0.08057
9 successes	0.02686
10 successes	0.00537
11 successes	0.00049

Significance Level and Rejection Region

Having selected the distribution that will be used to assess Type I error, we are ready to define the outcomes that will lead us to reject the null hypothesis. Our first step is to choose the significance level of our test. As described in Chapter 6, the significance level of a test is the amount of Type I error we are willing to risk in rejecting the null hypothesis. By convention, criminal justice researchers use a 5% significance threshold. But, as discussed in Chapter 6, we should consider at the outset whether a more lenient or more stringent significance level is appropriate for our study.

As researchers in the problem-oriented policing study do not present any special reason for altering conventionally accepted levels of significance, we will set a 5% significance threshold for our test of statistical significance. As noted in Chapter 6, in articles and books the significance level is often expressed by the Greek letter α. For our test, $\alpha = 0.05$.

The significance level defines the Type I error we are willing to risk in our test. But it does not tell us directly what outcomes in our sample would lead us to reject the null hypothesis. For this, we need to turn to our sampling distribution and define an area within it called a **rejection region.** The rejection region of a test is the area in the sampling distribution that includes those outcomes that would lead to rejection of the null hypothesis. If the observed significance level of a test, or the p value of the test, falls within the rejection region, then the researcher rejects the null hypothesis and concludes that the outcome is statistically significant. The area covered by the rejection region is equivalent to the significance level of a test. The point at which the rejection region begins is called the **critical value** because it is the point at which the test becomes critical and leads the researcher to reject the null hypothesis.

In the problem-oriented policing example, the rejection region includes 5% of the sampling distribution. Our initial problem is to define which 5%. Should we define the rejection region to be in the middle of the distribution represented in Table 8.3—for example, at 5 or 6 successes in 11 comparisons? Or should we look only at the extreme values on the positive side of the distribution, where there are mostly successes? Or should we include the area on the negative side of the distribution, where there are no successes?

Choosing a One-Tailed or a Two-Tailed Rejection Region

The answer to our questions comes in part from common sense and in part from our assumptions. It just would not make sense to place the rejection region in the middle of the sampling distribution. We are trying to decide whether the outcomes observed in our sample are very different from the outcomes that would be expected if problem-oriented policing had no impact. Putting the rejection region in the middle of the distribution would place it among those outcomes that are most likely under the null hypothesis. Clearly, we want the rejection region to be on the edges of the distribution, or in what statisticians call the **tails of the distribution.** These are the unlikely events—those that we would not expect if the null hypothesis were true. As indicated in our sampling distribution in Table 8.3, we would expect to get 11 successes in a row in about 5 of 10,000 samples if the program had no impact on the population. This is a very unlikely event and one that would lead us to reject the null hypothesis.

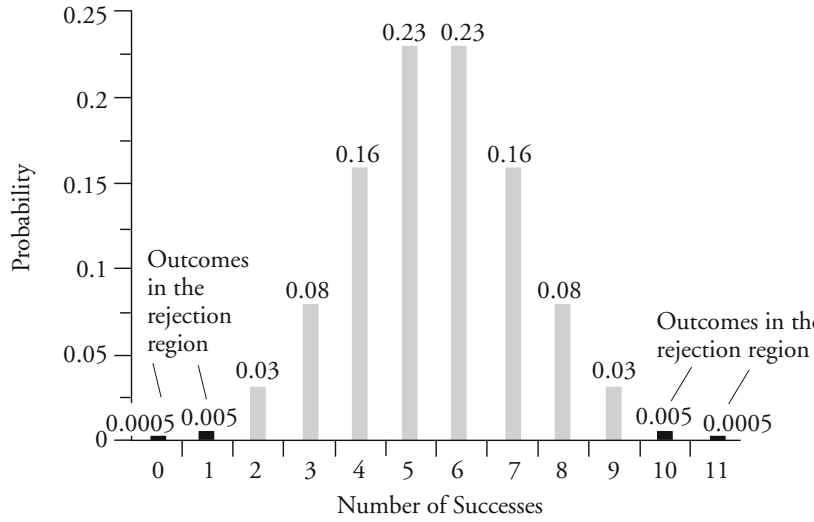
But zero successes is also an unlikely event, with the same probability of occurrence as 11 successes. Should we include only one tail of the distribution in our rejection region—the tail that assesses whether the program was a success? Or should we also include the opposite side of the distribution, which suggests that the program led to more disorder? Our answer is drawn from the research hypothesis that we stated in our assumptions. We chose a nondirectional research hypothesis, meaning that we are interested in evaluating both the possibility that the experimental sites improved relative to the control hot spots and the potential outcome that they got worse relative to the control hot spots. In terms of the sampling distribution, our research hypothesis suggests that the rejection region for our test should be split between both tails of the distribution.

This type of test is called a **two-tailed test of significance.** If we had stated a directional research hypothesis, we would be concerned with outcomes on only one side of the sampling distribution. Such a test is called a **one-tailed test of significance.** For example, if our research hypothesis were that incivilities in treatment hot spots decrease relative to incivilities in control hot spots after POP intervention, we would be concerned only with outcomes on the side of the distribution that shows program success.

The choice of a one-tailed or two-tailed test of statistical significance has important implications for the types of study outcomes that will lead to rejection of the null hypothesis. Because our test is a two-tailed test, the rejection region must be divided between both sides of the sampling distribution. This means in practice that the total significance level of 0.05 must be divided in half. Half of the rejection region, or 0.025, is found in the tail associated with success of the program, and half, or 0.025, in the tail associated with failure.

What outcomes would lead to rejection of the null hypothesis in our example? When we add 0 and 1 successes or 10 and 11 successes, we gain a probability value of 0.00586 (in each tail of the distribution, 0.00049 + 0.00537). This is less than the 0.025 value that we have defined as the rejection region for each tail of our test. Accordingly, an outcome of 0, 1, 10 or 11 would lead to an observed significance level less than the significance level of 0.05 that we have set, and thus we would reject the null hypothesis ($p < 0.05$). However, including 2 or 9 successes, each of which has a probability value of 0.027, increases the area of the distribution to 0.066. This area is larger than our rejection region. An outcome of 9 or 2 would result in an observed significance level greater than 0.05, and thus we would fail to reject the null hypothesis. Figure 8.1 presents the binomial probabilities for our example and highlights the two tails of the distribution that are used to test our nondirectional hypothesis.

Figure 8.1 *Outcomes That Would Lead to Rejecting the Null Hypothesis for a Two-Tailed Test of Significance ($\alpha = 0.05$)*

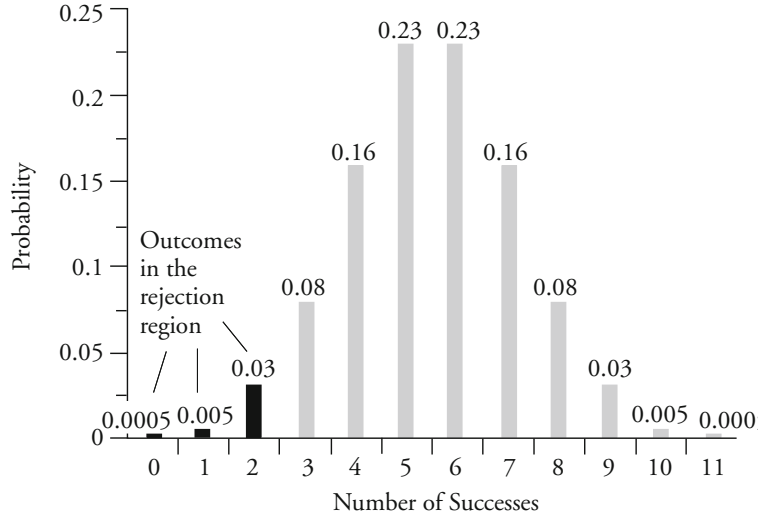


But what if we state a directional research hypothesis? How does this affect our rejection region? In this case, we calculate the area of the rejection region on only one side of the sampling distribution. Parts a and b of Figure 8.2 present the binomial probabilities for our two distinct directional hypotheses and highlight the tail of the distribution that is potentially of interest. For example, if our research hypothesis is that incivilities in treatment hot spots decline relative to incivilities in control hot spots after POP intervention, we look at outcomes only on the tail of the distribution that shows program success (Figure 8.2b). Because we are concerned only about these outcomes, all 5% of the rejection region is placed in this one tail of the distribution. We do not have to split the area of the rejection region. In this example, outcomes of 9, 10, and 11 successes are all within the rejection region, because adding their probabilities results in a value of 0.033 (0.00049 + 0.00537 + 0.02686). An outcome of 9, 10, or 11 results in an observed significance level that is less than the 5% significance threshold of our test (see Figure 8.2b). Adding the probability of 8 successes (or 0.08057) puts us above that threshold. If our research hypothesis is that incivilities increase in treatment hot spots relative to control hot spots, then we look at outcomes only on the opposite tail of the distribution (Figure 8.2a). In this case, outcomes of 0, 1, and 2 successes lead us to reject the null hypothesis.

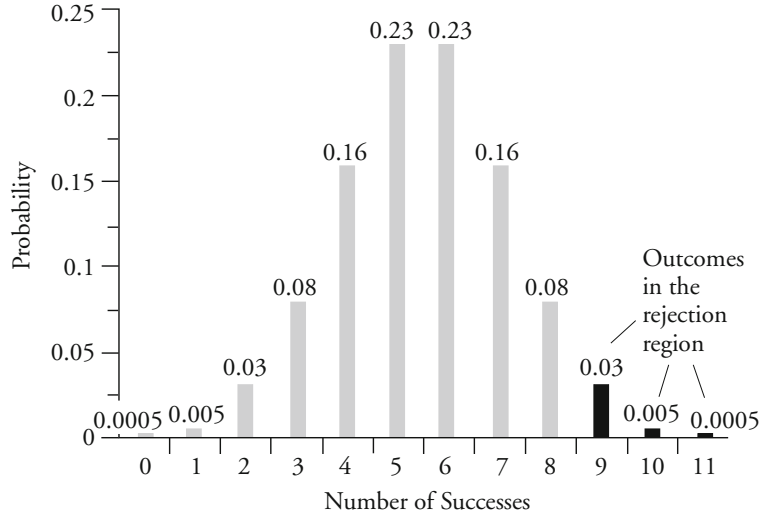
This example reinforces a rule described earlier: It is important to state the research hypothesis before you gain study outcomes. What if the problem-oriented policing hot spots improved relative to control locations in nine comparisons? With a one-tailed test, the result would fall within our rejection region and lead to rejection of the null hypothesis. With a two-tailed test, the result would be outside our rejection region. The choice of a directional or nondirectional research hypothesis can have an important impact on our conclusions. Merely by stating the research hypothesis a bit differently, we can change the outcome of the test.

A one-tailed test makes it easier to reject the null hypothesis based on outcomes on one side of a sampling distribution because it precludes rejection of the null hypothesis based on outcomes on the opposite side. The price of a larger rejection region in one-tail of the sampling distribution is no rejection region in the other tail. Similarly, the price of being able to examine outcomes on both sides of the distribution, as is the case with a two-tailed test, is that the rejection region will be smaller on each side. The benefit is that you can assess results in both directions. If you already know the outcomes of a test, you might be tempted to adjust the direction of the test according to the observed outcomes of a study. Taking such an approach unfairly adjusts the rejection region to your advantage.

Figure 8.2 *Outcomes That Would Lead to Rejecting the Null Hypothesis for a One-Tailed Test of Significance ($\alpha = 0.05$)*



(a) *Focus on Program Failures*



(b) *Focus on Program Successes*

The Test Statistic

In most tests of statistical significance, it is necessary to convert the specific outcome of a study to a **test statistic.** A test statistic expresses the value of your outcome in units of the sampling distribution employed in your test. For the binomial distribution, the units are simply the number of successes in the total number of trials. The test statistic for our POP intervention example is 10.

Making a Decision

The final step in a test of statistical significance is making a decision. If you have laid out all of the steps discussed above, then your choice should be easy. If your test statistic falls within the rejection region, then you reject the null hypothesis. This means in practice that the observed significance level of your test is less than the criterion significance level that you set when you defined the significance level and rejection region for your test. If the test statistic does not fall in the rejection region, you cannot reject the null hypothesis. In our example, the test statistic (10) does fall in the rejection region, which includes 0, 1, 10, and 11 successes. In this case, our observed significance level is less than the 0.05 threshold we set earlier. Our decision, then, is to reject the null hypothesis that incivilities in treatment hot spots do not change relative to incivilities in control hot spots after POP intervention. We conclude that the differences observed are **statistically significant.**

But what does this mean? When we say that a result is statistically significant, we are not claiming that it is substantively important. The importance of a result depends on such issues as whether the research affects real-life criminal justice decision making or whether it contributes new knowledge to a specific area of criminology or criminal justice. We also are not stating that we are certain that the null hypothesis is untrue for the population. Without knowledge of the population parameter, we cannot answer this question with certainty. Statistical significance has a very specific interpretation. The fact that an outcome is statistically significant means that it falls within the rejection region of your test. This happens when the observed significance level for a test is smaller than the significance criterion, or significance level, set at the outset of the test. A statistically significant result is one that is unlikely if the null hypothesis is true for the population. Whenever we make a statement that a result is statistically significant, we do it with the recognition that we are risking a certain level of Type I error. In this test, as

in most tests of statistical significance in criminal justice, we were willing to take a 5% risk of falsely rejecting the null hypothesis.

Chapter Summary

The first stage in a test of statistical significance is to state one's **assumptions.** The first assumption is about the type of measurement used. The second assumption concerns the shape of the population distribution. A **parametric test** is one that makes assumptions about the shape of the population distribution. A **nonparametric test** makes no such assumptions. Although nonparametric tests have the advantage of making fewer assumptions, they are generally used only for nominal and ordinal scales. The third assumption relates to the sampling method. A **random sample** is generally considered to be more representative, or to have greater **external validity,** than a **convenience sample. Independent random sampling** is the most accepted form of sampling. To ensure the independence of the sampling, it is in theory necessary to return the subject to the **sampling frame** after selection. **Sampling with replacement** creates practical problems, however, and is generally not required if the sample is small relative to the population. The fourth assumption states the null and research hypotheses. Care should be taken in framing them and in deciding whether the research hypothesis should be directional.

The second stage is to select an appropriate sampling distribution. The third stage is to select a significance level. The significance level determines the size of the **rejection region** and the location of the **critical values** of the test. If a test result falls within the rejection region, the researcher is prepared to reject the null hypothesis. This means that the observed significance level of the test is less than the significance level the researcher set at the outset of the test. If the hypotheses are directional, then the researcher will be concerned only with one **tail of the distribution,** and the entire rejection region will be placed on one side of the distribution (a **one-tailed test of significance**). If the hypotheses are nondirectional, then the researcher is concerned with results in both tails, and the rejection region will be divided equally between both sides of the distribution (a **two-tailed test of significance**).

The fourth stage involves calculating a **test statistic.** The study result is now converted into the units of the sampling distribution. Finally, a decision is made: The null hypothesis will be rejected if the test statistic falls within the rejection region. When such a decision can be made, the results are said to be **statistically significant.**

Key Terms

assumptions Statements that identify the requirements and characteristics of a test of statistical significance. These are the foundations on which the rest of the test is built.

convenience sample A sample chosen not at random, but according to criteria of expedience or accessibility to the researcher.

critical value The point at which the rejection region begins.

distribution-free tests Another name for nonparametric tests.

external validity The extent to which a study sample is reflective of the population from which it is drawn. A study is said to have high external validity when the sample used is representative of the population to which inferences are made.

independent random sampling A form of random sampling in which the fact that one subject is drawn from a population in no way affects the probability of drawing any other subject from that population.

nonparametric tests Tests of statistical significance that make no assumptions as to the shape of the population distribution.

one-tailed test of significance A test of statistical significance in which the region for rejecting the null hypothesis falls on only one side of the sampling distribution. One-tailed tests are based on directional research hypotheses.

parametric tests Tests of statistical significance that make assumptions as to the shape of the population distribution.

random sampling Drawing samples from the population in a manner that ensures every individual in that population an equal chance of being selected.

rejection region The area of a sampling distribution containing the test statistic values that will cause the researcher to reject the null hypothesis.

representative sample A sample that reflects the population from which it is drawn.

sampling frame The universe of eligible cases from which a sample is drawn.

sampling with replacement A sampling method in which individuals in a sample are returned to the sampling frame after they have been selected. This raises the possibility that certain individuals in a population may appear in a sample more than once.

statistically significant Describing a test statistic that falls within the rejection region defined by the researcher. When this occurs, the researcher is prepared to reject the null hypothesis and state that the outcome or relationship is statistically significant.

tails of the distribution The extremes on the sides of a sampling distribution. The events represented by the tails of a sampling distribution are those deemed least likely to occur if the null hypothesis is true for the population.

test statistic The outcome of the study, expressed in units of the sampling distribution. A test statistic that falls within the rejection region will lead the researcher to reject the null hypothesis.

two-tailed test of significance A test of statistical significance in which the region for rejecting the null hypothesis falls on both sides of the sampling distribution. Two-tailed tests are based on nondirectional research hypotheses.

Exercises

8.1 Answer the following conceptual questions:

a. Is it better to have more or fewer assumptions at the beginning of a test of statistical significance? Explain your answer.

b. Why is it important to state all of the assumptions at the outset of the test?

c. In what sense can stating the null and research hypotheses be seen as making assumptions?

8.2 Gatley University is an elite university of 1,000 students. Nadia, a student studying Chinese at the university, wishes to determine the average IQ of students at Gatley. She has decided that her sample size will be 50, and she is considering several different sampling methods. For each method, state the sampling frame and discuss whether the sampling method is random and whether it is independent.

a. Nadia chooses 50 names at random from the list of language students at the university.

b. Nadia asks 50 of her acquaintances at the university if they would mind taking an IQ test.

c. Nadia chooses the first two students from the alphabetical list of each of the 25 university departments.

d. Nadia takes all 1,000 names and puts them into a hat. She draws out a name, writes it down, and then puts it back in the hat and draws again. This procedure is repeated 50 times.

8.3 Hale Prison is renowned for its poor internal discipline. The new prison governor wants to tackle this problem and decides to investigate whether removing prisoners' visiting privileges will act as a deterrent against future misbehaving. From 100 prisoners who recently took part in a violent prison riot, he selects the 25 inmates with the worst disciplinary records, removes their visiting privileges, and begins to monitor their progress relative to the others.

a. Does this method meet the criteria of independent random sampling?

b. Is independent sampling possible in this case?

c. Describe a more appropriate sampling method.

8.4 For each of the following hypotheses, state whether a one-tailed or a two-tailed test of statistical significance would be appropriate. In each case, explain your choice.

a. H_1: Citizens over the age of 50 are more likely to be the victims of assault than citizens under the age of 50.

b. H_1: Children raised by adopted parents have rates of delinquency different from those of children raised by their biological parents.

c. H_1: The experience of imprisonment has an impact on the chances of an ex-convict reoffending.

d. H_1: Women are more likely than men to support increased sentences for rapists.

e. H_1: Persons who are not victims of assault have lower levels of anger than persons who have been victims of assault.

f. H_1: White offenders are less likely to be sentenced to prison than Hispanic offenders.

g. H_1: Teenagers have rates of crime that are different from adult rates of crime.

h. H_1: Defendants charged with property crimes have different rates of pretrial misconduct than defendants charged with violent crimes.

i. H_1: Male defendants are more likely to be held on bail than female defendants.

j. H_1: Women are more supportive of capital punishment than men.

k. H_1: States with higher unemployment rates have higher rates of property crime.

l. H_1: The level of poverty in a neighborhood affects the neighborhood's crime rate.

m. H_1: Democrats are less supportive of cutting taxes than Republicans.

n. H_1: Graduates from private law schools are more likely to become federal judges than graduates from state law schools.

8.5 In Chapter 7, we constructed a binomial distribution showing the chances of success and failure for ten tosses of a fair coin. The distribution was as follows:

0 heads	0.001
1 head	0.010
2 heads	0.044
3 heads	0.118
4 heads	0.206
5 heads	0.247
6 heads	0.206
7 heads	0.118
8 heads	0.044
9 heads	0.010
10 heads	0.001

Consider the following alternative hypotheses:

Alternative 1: H_0: The coin is fair.
H_1: The coin is biased.

Alternative 2: H_0: The coin is fair.
H_1: The coin is biased in favor of heads.

a. Would a one-tailed or a two-tailed test be more appropriate for a researcher who chose alternative 1? Explain why.

b. For a sequence of ten throws, what results would cause a researcher operating under the hypotheses listed under alternative 1 to reject the null hypothesis at a significance level of 5%?

c. Would a one-tailed or a two-tailed test be more appropriate for a researcher who chose alternative 2? Explain why.

d. For a sequence of ten throws, what results would cause a researcher operating under the hypotheses listed under alternative 2 to reject the null hypothesis at a significance level of 5%?

8.6 Use the following binomial distribution showing the chances of success and failure for 12 trials.

Number of Successes	Probability
0 successes	0.00118
1 success	0.01065
2 successes	0.04418
3 successes	0.11110
4 successes	0.18857
5 successes	0.22761
6 successes	0.20032
7 successes	0.12953
8 successes	0.06107
9 successes	0.02048
10 successes	0.00463
11 successes	0.00064
12 successes	0.00004

Using a significance level of 0.05, what outcomes would lead you to reject the null hypothesis for each of the following pairs of hypotheses?

a. H_0: $P = 0.50$
H_1: $P \neq 0.50$

b. H_0: $P = 0.50$
H_1: $P < 0.50$

c. H_0: $P = 0.50$
   H_1: $P > 0.50$

d. If you changed the significance level to 0.01, how would your answers to parts a, b, and c change?

8.7 Use the following binomial distribution showing the chances of success and failure for 15 trials.

Number of Successes	Probability
0 successes	0.00000
1 success	0.00000
2 successes	0.00001
3 successes	0.00006
4 successes	0.00042
5 successes	0.00228
6 successes	0.00930
7 successes	0.02928
8 successes	0.07168
9 successes	0.13650
10 successes	0.20051
11 successes	0.22313
12 successes	0.18210
13 successes	0.10288
14 successes	0.03598
15 successes	0.00587

Using a significance level of 0.05, what outcomes would lead you to reject the null hypothesis for each of the following pairs of hypotheses?

a. H_0: $P = 0.50$
   H_1: $P \neq 0.50$

b. H_0: $P = 0.50$
   H_1: $P < 0.50$

c. H_0: $P = 0.50$
   H_1: $P > 0.50$

d. If you changed the significance level to 0.01, how would your answers to parts a, b, and c change?

8.8 Locate a research article in a recent issue of a criminology or criminal justice journal.

a. State the research hypotheses tested by the researcher(s).

b. Describe the sampling method, the sample, and the sampling frame used by the researcher(s).

Computer Exercises

SPSS and Stata both have the capability to test hypotheses using the binomial distribution. As discussed in each subsection below, there are sample syntax files in both SPSS (Chapter_8.sps) and Stata (Chapter_8.do) that illustrate the commands for testing against the binomial distribution.

SPSS

The NPTESTS command will use a two-tailed test to calculate an observed significance level for a binary variable (e.g., success vs. failure) and compare the probabilities observed to a binomial distribution (among many other options in the NPTESTS command). The default probability of a single success in the BINOMIAL command is $p = 0.50$, meaning that this command tests the following hypotheses:

$$H_0: P = 0.50$$
$$H_1: P \neq 0.50$$

As we illustrate below, the probability can be changed easily in the NPTESTS command line.

To try out the NPTESTS command and apply it to a binomial distribution, open the SPSS syntax file Chapter_8.sps. A small data file (ex_8_1.sav) will be read into SPSS when you execute the first two lines of command syntax. This small data file contains the data from Table 8.1 in the text. Relative decreases in post-intervention crime are indicated by a value of 1, and relative increases in post-intervention crime are indicated by a value of 0.

The structure to the NPTESTS command for a comparison to a binomial distribution is:

```
NPTESTS /ONESAMPLE TEST (crime) BINOMIAL(TESTVALUE=0.5
SUCCESSCATEGORICAL=LIST(1)).
```

Where /ONESAMPLE indicates that we have data from only one sample that we are going to compare with a hypothesized population. The TEST(crime) statement indicates that the variable "crime" (the only one in the data file) is to be the focus of the test. Within the BINOMIAL option—which tells SPSS to compare "crime" against a binomial distribution—the TESTVALUE is the hypothesized value of the probability of a success (i.e., P in the hypotheses above) and SUCCESSCATEGORICAL = LIST(1) tells SPSS that we have coded our measure so that a value of 1 is a success. All other values for a variable would be interpreted as failures.

After executing this command, the output window presents a table of results that indicates the null hypothesis being tested and the observed significance level (labeled "Exact Significance" in the table). You will see from this output window that the observed significance level is 0.012, which is identical to the value calculated on p. 184 in the text.

Stata

The binomial test in Stata is remarkably simple in form:

bitest variable_name **== hypothesized_probability**

The output from running the **bitest** command will be a table indicating the observed number of success and then both one-tail and the two-tail tests of the hypothesized probability. Should you be interested in testing a different hypothesized probability, just alter the value from a presumed 0.5 to the value you want to test.

Open the file with the Stata do file Chapter_8.do to reproduce the results from Table 8.1 in the text. The first command line opens the data file, which is identical to that referred to in the discussion of SPSS above.

The binomial test of the crime variable is then simply:

bitest crime **== 0.5**

The results for the two-tail test show a significance level of 0.012 (rounded), exactly the same as reported above.

Problems

1. The director of a special drug treatment program claims to have found a cure to drug addiction. As supporting evidence, the director produces information on a random sample of 13 former clients who were followed for 12 months after completing the program. Here is how the director classified each former client:

   Success, Failure, Success, Success, Success, Success, Success, Failure, Success, Success, Failure, Success, Success

   Enter these data into SPSS.

   a. State all the assumptions of the hypothesis test.

   b. What is the test statistic?

   c. What decision can be made about the null hypothesis? (Assume that the significance level is 0.05.)

   d. Can the director conclude that the program is effective? Explain why.

2. A group of researchers wanted to replicate previous research on hot spot interventions in another city, using a sample of 25 hot spots. When comparing post-intervention crime levels, they classified the 25 locations as follows:

   Decrease, Decrease, Increase, Decrease, Decrease, Increase, Increase, Decrease, Decrease, Decrease, Decrease, Decrease, Decrease, Decrease, Decrease, Increase, Increase, Decrease, Decrease, Decrease, Decrease, Decrease, Increase, Decrease, Decrease

Enter these data into SPSS.

a. State all the assumptions of the hypothesis test.

b. What is the test statistic?

c. What decision can be made about the null hypothesis? (Assume that the significance level is 0.05.)

d. Did this study show a post-intervention change in crime?

e. If the significance level had been set at 0.01, would the researchers have come to the same conclusion? Explain why.