Chapter 2

One-Sample Proportions

2.1 Introduction: Qualitative Data

The simplest types of measurements are qualitative in nature, meaning that they are non-numeric – or at least numeric manipulation of them is meaningless – and include names, labels and group membership. Examples of qualitative data are ubiquitous, but are best exemplified by dichotomous categorical data consisting of only two possible values, such as a patient's gender (male or female), diagnosis of a certain disease (positive or negative), or the result from a health intervention (success or failure).

Dichotomous categorical data are typically described in terms of the proportion (p) of some population with one of the two possible characteristics. This value is defined as the total number of subjects in some population *exhibiting* that specific characteristic *divided* by the number of total subjects (N) in that population. For instance, if 105 out of 200 physicians in a given hospital are female, then the proportion of these physicians who are female is $p_f = 105/200 = 0.525$. It should stand to reason that a proportion can only take values between 0 and 1.0, as you cannot have fewer than zero subjects with a given characteristic (reflecting $p = 0/N = 0$), just as you cannot have more than the total number subjects with a given characteristic (reflecting $p = N/N = 1.0$).

The Complement Rule: The outcomes for dichotomous data must also be mutually exclusive, in the sense that any given subject may assume only one of the two potential outcomes at one time. For instance, a subject cannot simultaneously test positive and negative for a disease. In general we will ignore instances where the outcomes are not mutually exclusive; in practice it is best to avoid these scenarios all together. One benefit of this characteristic is that we only need to know the proportion of one of the two outcomes to know the proportion for both. Returning to our previous example, if there are 105 female physicians, then there *must be* $200 - 105 = 95$ male physicians, meaning the proportion of male physicians is $p_m = 95/200 = 0.475$. Note here

R. Sabo and E. Boone, *Statistical Research Methods: A Guide for Non-Statisticians*, DOI 10.1007/978-1-4614-8708-1_2,

that there are $105 + 95 = 200$ physicians who are either male or female, meaning that the proportion of physicians who are either male or female is $p_e = 200/200 = 1.0$. Further, note that $p_f + p_m = 0.525 + 0.475 = 1.0 = p_e$. This will *always be the case* for dichotomous categorical data. So if we know that $p_f = 0.525$, then we can use what is called the *complement rule* to find $p_m = 1 - p_f = 1 - 0.525 = 0.475$.

As a final note on proportions, there is a one-to-one relationship between proportions and percentages, meaning that for every proportion between 0 and 1.0, there is a corresponding percentage between 0 and 100%. This means that we should be able to transform proportions into percentages – and percentages into proportions – with ease. Without getting into the mathematical rationale, the algorithm is simple: to turn a proportion into a percentage, move the decimal two places to the right and add a percent sign (%). For example, if we have the proportion $p = 0.525$, we turn it into the percentage 52.5%. Likewise, if we have a percentage (say 47.5%), we turn it into a proportion by moving the decimal two places to the left and removing the percent sign (0.475).

2.2 Establishing Hypotheses

The key problem here is that we generally do not know the exact value of a population proportion, and at times we might not even know the total number of subjects comprising that population. This is problematic for those who may want to base their decisions or actions on such a proportion. For instance, in deciding between two different treatments to administer to a patient, a physician might want to know the success rates – read: *proportions* – of those two treatments before choosing between them. These population values are *rarely* known, but certainly the physician – or others in a similar situation – must make a decision, so something else must be done.

Thus enters the statistical method and the formation of a hypothesis. When a population proportion is unknown, we must formulate competing and mutually exclusive hypotheses about that proportion, collect data representative of the desired population, evaluate that data, and determine which hypothesis the evidence supports. The first step in this process is to set up competing hypotheses to test. There is generally some hypothesized value (p_0 – pronounced "*p-naught*") in which we are interested; for instance, maybe it is commonly accepted that the success-rate of a given treatment is 0.5 (treatment is successful for half of all patients and unsuccessful for the other half). This value then becomes the central crux around which we form our hypotheses.

As mentioned in Chapter 1, we will create two mutually exclusive hypotheses, such that only one can be true at a time. We name these hypotheses the null and alternative hypotheses, and we have a formal process for determining which hypothesis gets which name (the naming procedure is actually important). The null hypothesis (H_0) is that hypothesis which states

our parameter is equal to some value (p_0), while the alternative hypothesis (H_A) indicates that the parameter is somehow different from p_0. Depending upon our research question, there are three possible ways in which the parameter – proportion in this case – can differ from p_0: it can be *less than* p_0 (represented by the symbol $<$), it can be *greater than* p_0 (represented by the symbol $>$), or it can be *not equal to* p_0 (represented by the symbol $\neq$). To choose between these three options we begin by translating our research question into symbolic form, which will include one of the following options: $<$, $\leq$, $>$, $\geq$, $=$ or $\neq$. As an example, suppose our research question is that the proportion of subjects with adverse toxic reactions to a particular drug is less than 0.3. In order to turn this into a symbolic statement, we must identify the operative phrase "is less than", which is stating that $p < 0.3$.

The second step is to find the functional opposite of the statement from our research question. Based on the symbolic form from our research question, we create the functional opposite by pairing the following symbols: ($<$ and $\geq$), ($>$ and $\leq$) or ($=$ and $\neq$). Note that each of these pairs comprise all possibilities for a given situation (e.g. you are either strictly less than some value or greater than or equal to some value; either greater than some value or less than or equal to some value; either equal to some value or not equal to some value). Returning to our example, the functional opposite of the symbolic form of our research question ($p < 0.3$) is $p \geq 0.3$.

The third step is to identify which of our two symbolic forms is the null hypothesis and which is the alternative hypothesis, which is easier to do than to explain. Of the two symbolic forms, the form with some equality (meaning the $=$, $\leq$ or $\geq$ signs) becomes the null hypothesis, while the symbolic form without any equality (meaning the $\neq$, $<$ or $>$ signs) becomes the alternative hypothesis. Further, regardless of the symbol in the statement that belongs to the null hypothesis, we use the $=$ sign. (We do this for practical reasons, as we're going to assume the null hypothesis is true, and doing so is much easier if H_0 contains only one value rather than a range of values. Keep in mind, however, that this practical reason is not the same as theoretical justification, which will be given elsewhere.) For our example, the statement $p \geq 0.3$ contains equality, while the statement $p < 0.3$ does not. So our alternative hypothesis becomes $H_A : p < 0.3$, while the null hypothesis becomes $H_0 : p = 0.3$. This process can be followed for most research statements concerning one population proportion, and Table 2.1 lists the possible hypotheses as well as key words to help in guiding you to the appropriate pair.

2.3 Summarizing Categorical Data (with R Code)

Sample Proportion: Given a set of hypotheses about a population proportion, the next step is to collect evidence that will (hopefully) support one of the two hypotheses. When we are interested in a population proportion, the

Table 2.1: Possible Sets of Hypotheses for a Population Proportion Based Upon Key Phrases in a Research Question.

	Key Phrases		
	"less than" "greater than or equal to"	"greater than" "less than or equal to"	"equal to" "not equal to"
Hypothesis	"at least"	"at most"	
Null	$H_0 : p = p_0$	$H_0 : p = p_0$	$H_0 : p = p_0$
Alternative	$H_A : p < p_0$	$H_A : p > p_0$	$H_A : p \neq p_0$

logical step would be to calculate a *sample proportion* from a representative sample drawn from the population of interest. The sample proportion $\hat{p}$ serves as an *estimate* of the population proportion, and is calculated in a similar manner as its population analogue, being the number of subjects x in a sample exhibiting a particular characteristic (often referred to as the *frequency*) divided by the total number of subjects n in the sample (referred to as the *sample size*). So if we have a random sample of 25 physicians, 13 of whom are female, then the proportion of female physicians in this sample is $\hat{p}_f = 13/25 = 0.52$. Due to the dichotomous nature of this type of measurement, we can use the complement rule to find the sample proportion of male physicians ($\hat{p}_m = 1 - 0.52 = 0.48$).

Rounding: Depending on the sample size, you will have different rules for rounding proportions. For sample sizes greater than 100, round $\hat{p}$ to at most three decimal places (e.g. $\hat{p} = 0.452$). For sample sizes less than 100 but greater than 20, round $\hat{p}$ to two decimal places (e.g. $\hat{p} = 0.45$). For small sample sizes less than $\sim$20, the ratio of frequency to sample size should be reported as a fraction (x/n) (e.g. 5/11) and the sample proportion should not be calculated (note that there is no universally agreed upon value for this last rule, and 20 was selected for presentation).

2.4 Assessing Assumptions

If our random sample is adequately representative of the parent population from which it is drawn, then our sample estimate $\hat{p}$ should be close to the population value p. Unless we have clear evidence or information to the contrary, we will assume: (i) that the sample used to calculate $\hat{p}$ is representative, and (ii) that the subjects in the sample from whom measurements were taken are independent of one another. To determine if the sample is of sufficient size, we need to check the following conditions: based on the condition set forth in the null hypothesis H_0, we need to expect there to be *at least five* subjects taking either value of the categorical variable. This expectation is determined by noting that if the proportion of subjects taking the first category is $p_0 = 0.3$ according to H_0 (meaning the proportion taking the second

category is $q_0 = 1 - p_0 = 1 - 0.3 = 0.7$) and if $n = 30$, then we would expect there to be $p_0 n = (0.3)(30) = 9$ subjects in the first category, and $(1 - p_0)n = q_0\, n = (0.7)(30) = 21$ subjects in the second category. So in this case we would have adequate sample size. However, if H_0 instead specified that $p_0 = 0.1$, then we would expect $p_0 n = (0.1)(30) = 3$ subjects in the first category and $q_0 n = (0.9)(30) = 27$ subjects in the second. Since we expect less than five subjects in the first category, we would not have adequate sample size to perform the desired hypothesis test. In cases of inadequate sample size we would report that we cannot perform the desired test (meaning we stop the entire process and either figure out how much more data we need or perform a different test). Note that we use p_0 to determine if we have adequate sample size rather than $\hat{p}$.

2.5 Hypothesis Test for Comparing a Population Proportion to a Hypothesized Value

At this point we have already translated our research question into testable hypotheses, verified our assumptions, and summarized our data. It is now time to combine the two pieces into a statistical test that will eventually support either the null hypothesis or alternative hypothesis. Since we have a sample proportion $\hat{p}$ that should resemble the population proportion p upon which we are trying to make inference, it makes sense to base our test around the sample estimate. However, before we develop a formal test, we should study further the behavior of $\hat{p}$ in order to better understand from where such a test might arise.

2.5.1 Behavior of the Sample Proportion

Consider a random and representative sample of 200 patients undergoing treatment to alleviate side-effects from a rigorous drug regimen at a particular hospital, where 33 patients experienced reduced or no side-effects. For this particular sample, we know that the sample proportion of patients who experienced little or no side-effects was $\hat{p} = 33/200 = 0.165$. So one could presume – based on the evidence from this sample – that between 16 and 17% of all patients would experience reduced side-effects when using this treatment regimen. But is this a reasonable presumption? What if we had collected a different sample of patients from this hospital (or from a different hospital, for that matter)? Would the sample proportion change? If so, how much would it change?

While we cannot answer these questions based on our particular sample, we can conduct studies that will allow us to see what we could expect to find if we could *repeatedly sample* from a population with a *known population proportion*. It is possible for us to conduct a *simulation study*, or a study in which we repeatedly simulate sets of data that reflect known population parameters (such as p_0), where summary statistics (such as frequencies or $\hat{p}$) are calculated for each of those samples and then summarized themselves. We can then determine the likelihood of the observed sample data (or sample estimate) compared to the results from the simulation study (more on this topic later).

Returning to our example of 200 hospital patients, maybe the historical rate of patients with little or no side-effects is 10.0%, and we want to determine if this new treatment increases that rate. (Imagine a bag filled with a large number – say many thousands – of chips, 10% of which are red and the rest blue, and we draw out 200 of those chips and count the number of red chips; this is not what we really do, but the idea is the same).

The results from a simulation study that generated 1,000 such samples are provided below in Table 2.2, which shows the number of samples for which we observed specific success counts (ranging from 5 to 36) out of 200. Note that there are many more success frequencies in this study other than the frequency observed in our sample (which was 33), which reflects the variability we might observe if we were to repeatedly sample from this population. Variability can mean many things, but here we are taking it to mean how our sample proportion could change in value if we were to resample.

Note that a frequency of 20 occurs most often ($\sim$12% of the time) and represents the case when $\hat{p} = 20/200 = 0.100$, which is the value (p_0) assumed in this study. Also note that most of the simulated samples yielded frequencies slightly below or slightly above 20 (e.g. 16–19, 21–24), while relatively fewer studies yielded frequencies greatly below or greatly above 20 (e.g. 5–11, 29–36), which makes sense since 20 was the value we assumed was the *true population parameter*. Based on the results from this simulation study, we would then conclude that if $p = 0.10$ was indeed true, we would expect sample proportions close to 0.10 rather than far away from it.

2.5.2 Decision Making

So how likely is our sample value of 33? Based on our simulated data, 33 occurred once out of 1,000 total simulations; specifically, 33 successes appeared at a rate of $1/1{,}000 = 0.001$, which is not often. More generally, any value *greater than or equal to* 33 (which includes 33, 34, 35 and 36) occurred only 4 out of 1,000 times, or $4/1{,}000 = 0.004$, which is still not often. In either case, both the observed event (33) or any event *at least as extreme* as our observed event (> 33) seems unlikely if the true population success rate is $p = 0.10$.

Table 2.2: Results From Simulation Study of Samples with 200 Dichotomous Observations with a Known Success Rate of 0.10.

# of Successes out of 200	Frequency	Proportion	# of Successes out of 200	Frequency	Proportion
5	1	0.001	21	75	0.075
6	0	0.000	22	78	0.078
7	1	0.001	23	71	0.071
8	0	0.000	24	54	0.054
9	2	0.002	25	33	0.033
10	8	0.008	26	25	0.025
11	9	0.009	27	27	0.027
12	20	0.020	28	11	0.011
13	25	0.025	29	8	0.008
14	35	0.035	30	11	0.011
15	48	0.048	31	4	0.004
16	65	0.065	32	4	0.004
17	69	0.069	33	1	0.001
18	99	0.099	34	1	0.001
19	94	0.094	35	1	0.001
20	119	0.119	36	1	0.001

This last statement brings us to the crux of statistical decision making: based on our assumption of $p = 0.10$, the observed success rate $\hat{p} = 0.165$ does not seem likely (if we are to believe the simulation study, which we should). So what do we conclude? There are two likely outcomes: (i) our assumption of the population proportion was correct and our sample data are wrong (or at best unlikely), or (ii) our sample data is more reflective of the "truth" and our assumption was wrong.

Since our sample is the only information we have that reflects any property of the population from which it was drawn, and since it was randomly selected from and is representative of that population, *we must base our conclusions on what the data and its summaries tell us.* This is one of the most important ideas in this entire textbook: if the data do not support a given assumption, then that assumption is most likely not true. On the other hand, if our data *did* support our assumption, then we would conclude that the assumption is likely to be true (or at least more likely than some alternative).

Returning to our example, since a frequency of 33 (or greater) did not occur often in our simulation study, we would logically presume that we are not likely to observe frequencies that high (or higher) in samples drawn from a population with a success rate of 0.10. Thus, we conclude that, based on our sample proportion of 0.165, the true population proportion of patients experiencing reduced symptoms or side-effects under this treatment is probably greater than 0.10.

2.5.3 Standard Normal Distribution

While the previously conducted simulation study was helpful in discussing the behavior of a sample proportion under some hypothesized value, it is important to note that we do not usually conduct simulation studies every time we want to conduct a hypothesis test (indeed, it is often difficult or impractical to do so). Rather, statisticians from centuries past have successfully characterized the behavior of a sample proportion in such a manner that the results we would like to obtain are readily available without the need for sophisticated computing power.

Consider the histogram in Figure 2.1, which shows in graphical form the results from our simulation study. In its entirety, this histogram represents the *distribution* of sample proportions assuming $p = 0.10$. Based on this distribution, we see that it is largest in the middle (corresponding to likely values based on our assumption of $p = 0.10$), and then slowly gets smaller as we move away from 0.10 (in both directions), so that eventually we have infrequent or non-occurring outcomes. These regions are called *the tails* and represent values that are unlikely to occur if our assumption of $p = 0.10$ is true.

As mentioned earlier, we *do not* want to rely upon simulation studies or the distributions they create, though we would like a distribution that resembles that created by the simulation study. Thus, we use what is called the *standard normal distribution*, whose properties are well known and easy to use. A random variable Z has a standard normal distribution if the probability that it is between two numbers a and b is given by the following integral (given in Equation 2.1).

Figure 2.1: Histogram Summarizing Results from a Simulation Study of 1,000 Samples of 200 Dichotomous Outcomes with an Assumed Success Rate of $p = 0.10$.

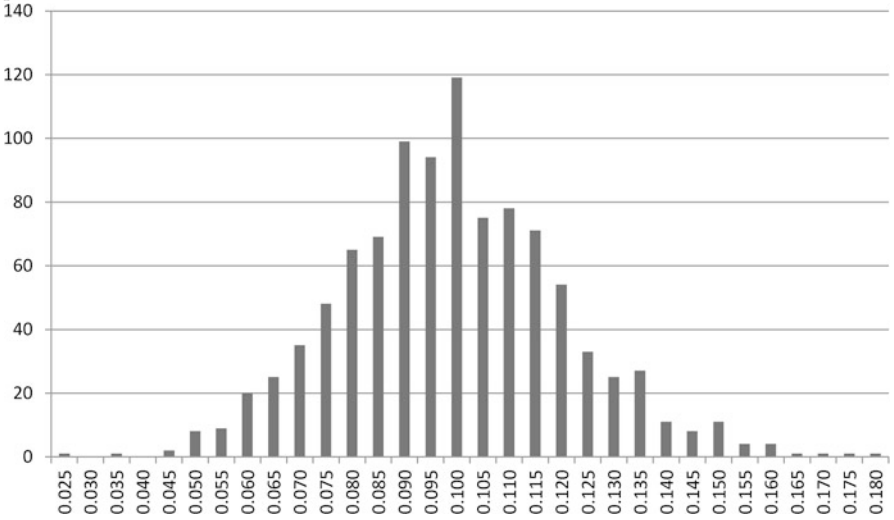
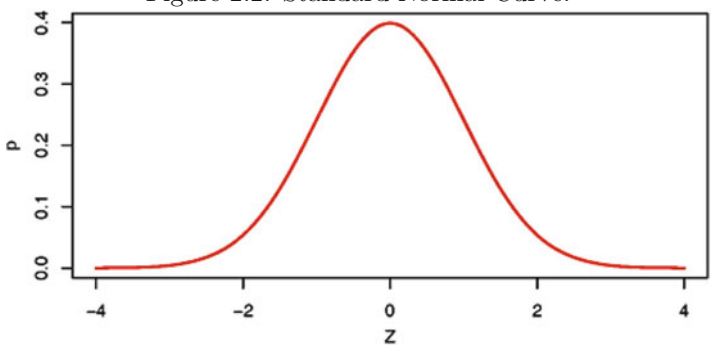


Figure 2.2: Standard Normal Curve.



$$P(a < Z < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dx \tag{2.1}$$

The standard normal distribution is centered at zero with a variance of one (we will formally define center and variance in later chapters). If the center of the distribution is any value other than zero or if the variance is not one, then we simply have a *normal distribution*. The standard normal distribution is graphically presented in Figure 2.2 below. Here we clearly see the *bulge* centered at zero, the gradual decline as we move away from zero, and the tails for unlikely large positive and large negative values far from the center.

To show how the normal distribution works, we have reproduced the histogram in Figure 2.3 and now overlaid a normal curve (like the one in Figure 2.2, but with a mean and variance matching those from the distribution in Figure 2.1). Notice how well the simulated data and the theoretical normal curve align. This generally happens if our simulation study is conducted adequately enough, and this result is actually supported by a statistical law (known as the *central limit theorem*, which we will discuss later). Thus, if our assumptions are met, we should feel comfortable using the normal distribution to represent the distribution of our sample estimate.

2.6 Performing the Test and Decision Making (with R Code)

2.6.1 Test Statistic

So while we can use the standard normal distribution to answer probabilistic statements about our data and hypotheses about the population parameter, a quick glance at Figure 2.1 will show you that the distribution of $\hat{p}$ *is not*

Figure 2.3: Histogram from Simulation Study with Overlaid Normal Curve.

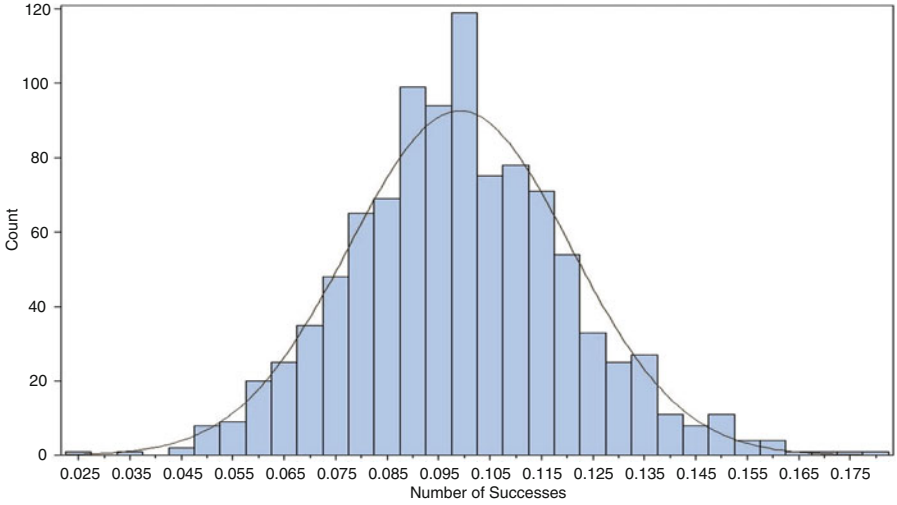
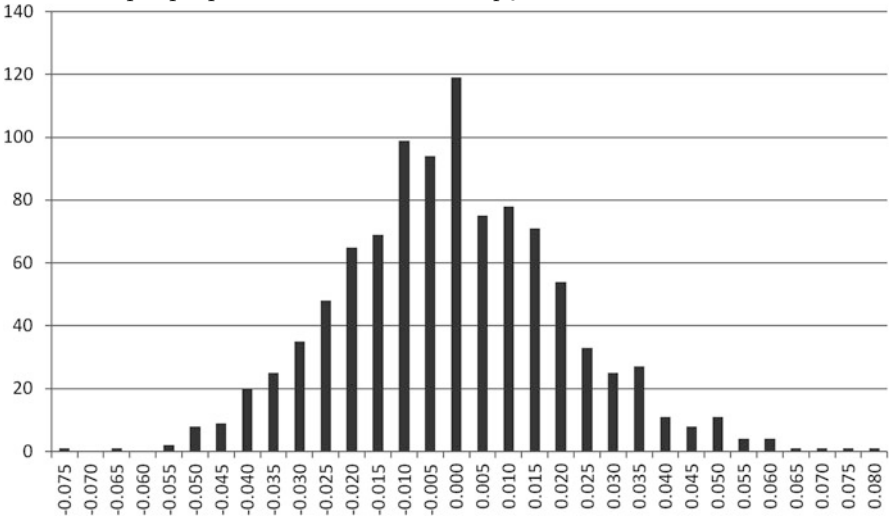


Figure 2.4: Histogram Summarizing Results from Simulation Study of 1,000 Samples of 200 Dichotomous Outcomes with Assumed Success Rate $p = 0.10$, where sample proportions are centered at $p_0 = 0.10$.



centered at zero. Since there is no other standard normal distribution for us to use, we will need to manipulate our sample estimate $\hat{p}$ so that *it will* have a standard normal distribution.

To do this, note that the distribution of $\hat{p}$ in Figure 2.1 is centered near the hypothesized value $p_0 = 0.10$. Thus, to get this distribution centered at zero, we can subtract the hypothesized value from our sample proportion: $\hat{p} - p_0$. Figure 2.4 shows the (nonsensical) adjusted distribution of $\hat{p}$ (nonsensical since it contains negative proportions), which is centered at zero.

While the distribution is now centered correctly, it still requires a variance of one. For reasons that are easy to mathematically justify, but are not easy to explain, the variability of a sample proportion that was drawn from a population with known proportion p is given by the standard error: $\sqrt{\frac{p(1-p)}{n}}$ (note that we use the population proportion and not the sample proportion). Thus, to transform our sample proportion into a random variable that has a standard normal distribution, we center at zero by subtracting p_0 and scale to a variance of one by dividing by the standard error to get Equation 2.2 below.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \tag{2.2}$$

Since the statistic z – known as a *test statistic* – has a standard normal distribution, we can use it to calculate probabilistic statements regarding our hypotheses and ultimately answer our question as to which hypothesis (the null or alternative) the data supports.

Returning to our example, recall that $x = 33$, $n = 200$, $\hat{p} = 0.165$, and $p_0 = 0.10$. Thus, our test statistics is

$$z = \frac{0.165 - 0.10}{\sqrt{\frac{0.10(1-0.10)}{200}}} = \frac{0.065}{0.0212} = 3.064$$

which means that the sample proportion 0.165 is slightly more than three standard deviations above the hypothesized population proportion 0.10 (standard deviation is another measure of variability, which we will define later). Three standard deviations is a lot, and means that in view of our sample data, the hypothesized value is unlikely. We can now use the test statistic z to make a formal decision. Note that we report test statistics – of the form z – to two decimal places, meaning we would report $z = 3.06$.

Unfortunately, R does not compute the test statistic just provided. However, the R function `prop.test()` does provide an equivalent (though cosmetically different) test for proportions, the syntax for which is as follows:

```
prop.test( x, n, p=p_0} )
```

The R code for our example is given in Program 2 below.

Notice that R provides a considerable amount of output (some of which is not needed or has not yet been defined). The `data:` line states what R is testing, which corresponds to the x, n and p_0. The next line gives the value of the test statistic `X-squared = 9.3889` and states that `p-value =`

Program 2 Program to conduct a hypothesis test on a single proportion.

Code:

```
prop.test(x=33, n=200, p=0.1, correct=FALSE)
```

Output:

```
        1-sample proportions test without continuity correction

data:  33 out of 200, null probability 0.1
X-squared = 9.3889, df = 1, p-value = 0.0021
alternative hypothesis: true p is not equal to 0.1
95 percent confidence interval:
 0.1199686 0.2226578
sample estimates:
    p
0.165
```

0.003216; this will be defined below. Hopefully, you are aware that the stated test statistic (9.3889) is different from the 3.064 we calculated by hand. However, note $\sqrt{9.3889} = 3.064$, which is the value we obtained. This relationship will be explained further in the next Chapter.

2.7 Formal Decision Making

2.7.1 Critical Value Method

The most traditional method of making a decision in a hypothesis test is to use critical values. A *critical value* is literally the boundary – in this case from the standard normal distribution – between values of our test statistic that seem likely and values that seem unlikely *if we were to assume that the null hypothesis was true*. Identification of such a critical value (or values) is helpful in the sense that we would only have to calculate our test statistic z and compare it to the critical value to make our decision.

Finding the critical value depends upon our alternative hypothesis and what is called the significance level. The *significance level* – denoted by the Greek letter α – is formally defined as the probability that we reject the null hypothesis (i.e. we don't believe it is true) when the situation it describes is actually true (i.e. rejecting H_0 was a mistake). Informally, α represents the lack of evidence we would need to observe in order for us doubt the veracity of the null hypothesis. Generally, we set the significance level at $\alpha = 0.05$, meaning that we would need to observe a statistic (or a statistic of a *more extreme value*) that we would expect to occur less than 5% of the time in order for us to reject the null hypothesis in favor of the alternative.

Given a specified significance level (and $\alpha = 0.05$ is generally used), the critical value then depends upon our alternative hypothesis. If the alternative specified that the proportion is less than some specified value ($H_A : p < p_0$) then we would expect small sample proportions (or negative values of our tests statistic z) to be rare if we assume $H_0 : p = p_0$ is true, and thus our critical value should be negative. For a similar reason, we have a positive critical value if our alternative hypothesis is $H_A : p > p_0$. In cases of a *two-sided alternative hypothesis* (or $H_A : p \neq p_0$), we need two critical values, since sample values much greater or much lower than our hypothesized value would lead us to reject the null hypothesis. All possible cases and decisions are presented in Table 2.3 for $\alpha = 0.05$ and $\alpha = 0.01$, which are the most commonly used significance levels. Thus, rather than having to determine critical values for each hypothesis test we wish to perform, we can consult Table 2.3 to obtain: (i) the critical value specific to the desired significance level and alternative hypothesis, and (ii) the criterion under which we would select the null or alternative hypothesis.

Table 2.3: Critical Values and Rejection (Acceptance) Regions for Hypothesis Test of a Proportion for Given Significance Levels (α) and Alternative Hypotheses.

	$\alpha = 0.05$		$\alpha = 0.01$	
Alternative Hypothesis	Critical Value	Select Hypothesis	Critical Value	Select Hypothesis
$H_A : p < p_0$ (Left-Tailed Test)	-1.645	H_0 if $z \geq -1.645$	-2.33	H_0 if $z \geq -2.33$
		H_A if $z < -1.645$		H_A if $z < -2.33$
$H_A : p > p_0$ (Right-Tailed Test)	1.645	H_0 if $z \leq 1.645$	2.33	H_0 if $z \leq 2.33$
		H_A if $z > 1.645$		H_A if $z > 2.33$
$H_A : p \neq p_0$ (Left-Tailed Test)	-1.96, 1.96	H_0 if $-1.96 \leq z \geq 1.96$	-2.575, 2.575	H_0 if $-2.575 \leq z \geq 2.575$
		H_A if $z < -1.96$ or $z > 1.96$		H_A if $z < -2.575$ or $z > 2.575$

In our example, say our original research statement was: *the proportion of subjects who experience reduced side-effects from this treatment is greater than 0.10.* This means our null hypothesis is $H_0 : p = 0.10$ and our alternative hypothesis is $H_A : p > 0.10$, and thus our critical value is 1.645 (if we have $\alpha = 0.05$), meaning that we will reject H_0 in favor of H_A if the test statistic is greater than 1.645, and we will not reject H_0 if the test statistic is less than or equal to 1.645. Earlier, we calculated our test statistic as $z = 3.06$,

which falls in the rejection region, so we reject H_0 in favor of H_A. We will discuss what this means and how we react later.

2.7.2 p-value Method

As an alternative to the critical value method, we can calculate what is called a *p-value*, which is defined as the probability of observing a test statistic *at least as extreme* as the one we actually observed, given that the null hypothesis is true. This is a tricky definition that has three distinct pieces. First, we must assume that the null hypothesis is true, otherwise we have no bearing to gauge how likely or unlikely the observed data are. Second, the meaning of *at least as extreme* depends upon the alternative hypothesis. If we have a left-tailed test (i.e. $H_A : p < p_0$), then at least as extreme means less than or equal to our observed test statistic. If we have a right-tailed test (i.e. $H_A : p > p_0$), then at least as extreme means greater than or equal to our observed test statistic. If we have a two-tailed test (i.e. $H_A : p \neq p_0$), then at least as extreme means both greater than or equal to the absolute value of our observed test statistic ($|z|$) and less than or equal to the negative of the absolute value of our observed test statistic ($-|z|$).

The third part is calculating the desired probability, which of course depends upon our observed test statistic (z, which itself depends upon the null hypothesis) and the meaning of "at least as extreme" (which is particularly dependent upon the alternative hypothesis). The standard normal distribution is used to calculate p-values, and we generally rely upon statistical software for their computation. Z-tables are used in many elementary Statistics courses, but we will not consult them. P-values can be calculated in Microsoft Excel, and are routinely provided by most statistical software packages (including R; see Program 2 above).

Regardless of the method of computation, the probability being calculated will be the same. If we have a left-tailed test, we calculate the probability that a standard normal random variable Z is less than our observed test statistic z given that the null hypothesis is true (or $P(Z < z|H_0)$). If we have a right-tailed test, we calculate the probability that Z is greater than z (or $P(Z > z|H_0)$). If we have a two-tailed test, then we calculate *two-times* the probability that Z is greater than $|z|$ (or $2P(Z > |z|\,|H_0)$), or we calculate *two-times* the probability that Z is less than $-|z|$ (or $2P(Z < -|z|\,|H_0)$). Admittedly, these definitions are complicated, but the good news is that you will not have to calculate them by hand.

To put the p-value into practice, we must compare it to the stated significance level α. In order to reject the null hypothesis, we would need an outcome (or something more extreme) that is less likely than our significance level. Thus, we reject the null hypothesis if our p-value is less than the significance level (p-value $< \alpha$), and we fail to reject the null hypothesis when our p-value is greater than or equal to the significance level (p-value $\geq \alpha$).

For our example, our observed test statistic ($z = 3.06$) and the right-tailed hypothesis test means that our p-value is 0.001092. This is less than the significance level $\alpha = 0.05$, so we reject the null hypothesis in favor of the alternative hypothesis. Note that you will make the same decision with the p-value method as you will using the critical value method (meaning, if you come to different conclusions, at least one of them is wrong). We also round p-values to at most four decimal places, so we should report p-value = 0.0011.

2.7.3 Conclusion

Whether we used the critical value or p-value method, we report our results in the same manner. First, we firmly declare whether we rejected or failed to reject the null hypothesis, the former case in favor of the alternative. We then state *in words* what this statistical decision means; as mentioned earlier, statistical methods – such as hypothesis testing – are only useful if we can phrase the results in ways that clinical or non-statistical researchers can understand and interpret.

In our example, our test statistic fell in the rejection region (the p-value was also smaller than the significance level), so we rejected the null hypothesis ($H_0 : p = 0.10$) in favor of the alternative ($H_A : p > 0.10$). So we would declare that the evidence suggests the success rate of this treatment at reducing side-effects is *likely* greater than 0.10. Notice that we did not claim that the success rate is greater than 0.10. This is because we only have statistical evidence, which is not the same as definitive proof.

The R software conducts the two-sided test ($H_A : p \neq p_0$) by default, though we can easily modify the code to conduct either of the one-sided tests. By adding the `alternative` statement to the R function `prop.test()`, R performs the test corresponding to the specified hypothesis. The specific syntax of the `alternative` statement for each type of hypothesis test is given below. Note that if you do not specify the `alternative` statement, R will default to the `"two.sided"` case and will perform the two-sided test.

$H_A : p \neq p_0$:

```
prop.test( x, n, p=p0, alternative="two.sided", correct=FALSE)
```

$H_A : p > p_0$:

```
prop.test( x, n, p=p0, alternative="greater", correct=FALSE)
```

$H_A : p < p_0$:

```
prop.test( x, n, p=p0, alternative="less", correct=FALSE)
```

For the right-tailed hypothesis in our example ($H_A : p > 0.10$) we would use the following R code (`prop.test`(x=33, n=200 p=0.10, alternative = "greater", correct=FALSE),) to produce the correct right tailed test; note that the p-value you get with this code (0.001092; try it yourself) matches what we reported for the z-test results.

2.7.4 Confidence Intervals

The sample proportion $\hat{p}$ is dependent upon the sample we collect and the particular subjects observed within that sample. In other words, $\hat{p}$ may change if we collect a different sample consisting of different subjects. This is a source of variability that is not expressed if we focus solely upon the current sample estimate. Thus, we often accompany each sample estimate with a *confidence interval* that takes into account sampling variability.

A *confidence interval* is straight forward to calculate, though somewhat tricky to define. What is definitive is what a confidence interval is not. A confidence interval has a stated level of confidence (defined as the complement of the stated significance level, or $1 - \alpha$). For instance, if our significance level is 0.05, then our confidence level is $1 - 0.05 = 0.95$, and we would then construct a 95% confidence interval. This level of confidence is often taken as the quantification of our belief that the true population parameter resides within the estimated confidence interval; this is false. Once calculated, a population parameter is either in a confidence interval or it is not. Rather, the confidence level reflects our belief *in the process of constructing confidence intervals*, so that we believe that 95% of our estimated confidence intervals would contain the true population parameter, if we could repeatedly sample from the same population. This is an important distinction that underlies what classical statistical methods and inference can and cannot state (i.e. we don't know anything about the population parameter, only our sample data).

To calculate a $(1 - \alpha)$% confidence interval (or CI) we need three things: a point estimate, a measure of variability of that point estimate, and a probabilistic measure that distinguishes between likely and unlikely values of our point estimate. With these three pieces, our CI would take the form:

$$(\text{Point Estimate} \pm \text{Measure of Variability} \times \text{Probabilistic Measure})$$

The $\pm$ sign indicates that by adding and subtracting the second part from the first we will obtain the upper and lower bounds, respectively, of our confidence interval. For a point estimate we use $\hat{p}$; in our example, this values is 0.165. As a measure of variability, we will use the square root of $\hat{p}(1-\hat{p})/n$, which is similar to the standard error used in hypothesis testing, except here we use $\hat{p}$ instead of p_0 since we don't necessarily want to use a null hypothesis to summarize our data (e.g. sometimes we may only want the CI and not the hypothesis test). Based on our sample data, this value would be $SE = \sqrt{0.165(1-0.165)/200} = 0.026$. As a probabilistic measure, we use the positive critical value from a two-tailed test for the given confidence level. For instance, if we want 95% confidence, then we would have $\alpha = 0.05$, and a two-tailed test would yield critical values ± 1.96, of which we take the positive value 1.96. Putting these values from our example together, our 95% confidence interval is

$$(0.165 - 1.96 \times 0.026, 0.165 + 1.96 \times 0.026) = (0.114, 0.216)$$

To interpret this interval, we would say "a 95% confidence interval of the population proportion of subjects who experienced reduced side-effects with this treatment is $(0.114, 0.216)$". In general, we round the confidence interval to the same degree of precision as our point estimate, in this case the sample proportion. Note that some researchers use confidence intervals to conduct hypothesis tests, where they estimate a confidence interval and determine whether some hypothesized value is within the interval (if not, reject H_0; if so, fail to reject H_0). While the confidence interval approach is similar in many ways to hypothesis testing, they are not the same and may not produce the same inference. For this and other reasons, we will use confidence intervals only as a form of data summarization, and will not use them for inference. For the record, we do not recommend or condone the use of confidence intervals for making statistical decisions or inference, and strongly encourage you to refrain from this practice.

Note that we could have used R to produce this confidence interval, but it will not *immediately* be the same, since R calculates confidence intervals using what is called the "continuity correction". This adjustment and the resulting type of interval is an equally valid but all together different type of confidence interval than the method described above; note that what we learned is by far the most commonly accepted form of calculating confidence intervals for dichotomous data. Moving forward, you can choose to calculate 95% CIs on a proportion using the method outlined in this chapter (which requires you to calculate the interval by hand), or you may use the two methods provided by the R software. To get the 95% CI in R, we make use of the `prop.test()` function with the following specifications (`x=33`, $n = 200$, $p = 0.1$, alternative="two.sided"), which produces a 95% CI of $(0.118, 0.225)$. Note that this method uses what's called the continuity correction, which we can turn off by specifying "*correct=False*" in the `prop.test()` function, which gives a 95% CI of $(0.120, 0.223)$. Both of these intervals are similar to but not equal to the interval provided above $(0.114, 0.216)$; ultimately, we would have to create our own code in R (which is not too difficult) to obtain the confidence interval we obtained by hand.

2.8 Contingency Methods (with R Code)

Occasionally we will experience the situation where we wish to compare the proportion to some hypothesized value, but (at least) one of our expected frequencies is less than 5, meaning we do not have a large enough sample size to perform the z-test. In that case, we must instead use the *Binomial Test*, which is a test that compares the proportion to a hypothesized standard and *is valid for any sample size*. This test works for any sample size because it is based on the concept of *enumeration*, or counting all possible outcomes that *could be observed* within one group of categorical data. In this instance enumerating all possible outcomes is not difficult, and can even be done by hand when the sample sizes are small enough.

For instance, imagine the case where someone gives you a cup and tells you it is either filled with Pepsi or Coke (let's say it is actually Pepsi). If you were asked to taste the soda and guess which soda was in which cup, there are only two possible outcomes: you guess correctly or incorrectly. This scenario is numerically represented in Table 2.4. If we had no way to discern between the unique tastes of Pepsi and Coke (i.e we were simply guessing), then we would assume that either outcome (we guess correctly or incorrectly) would have the same probability (0.5). Given this assumption (which is our null hypothesis), we can calculate the p-value of having as many or more correct guesses than what we observed. Based on the one-cup experiment (Table 2.4), if we were to guess 0 correct, then the p-value = 1.0, because we are certain of getting an outcome at least as extreme as the one we got (i.e. 0 or more correct) the next time we do this experiment. If we guessed correctly, then the p-value = 0.5, meaning there is an equal likelihood of getting a 1 or 0 the next time we do this experiment (the "1" being at least as extreme). Both p-values are much larger than 0.05, so even if we selected correctly, this is not enough evidence for us to reject the null hypothesis.

Table 2.4: Enumeration of Outcomes from One-Cup Experiment (Y: Correct Guess, N: Incorrect Guess).

Actual Soda in Cup Pepsi	# Correct	Frequency	Proportion	p-value
N	0	1	0.5	0.5+0.5 = 1.0
Y	1	1	0.5	0.5

Now let's assume that we have two cups, where the first is filled with Pepsi and the other with Coke. Of course, we do not know which sodas are actually in each cup, so we could guess that they are both filled with Pepsi, they are both filled with Coke, or they are filled with one soda each (and there are two ways in which this can happen: Pepsi in the first and Coke in the second, or Coke in the first and Pepsi in the second). Thus there are four ways in which we can guess, one resulting in no correct guesses, two resulting in one correct guess (and one incorrect guess), and one resulting in two correct guesses. These outcomes are summarized in Table 2.5. Since the four outcomes are equally probable if we are only guessing (assuming the null hypothesis is true), then each particular outcome has a 0.25 chance of occurring. So in this case, even if we guess the contents of both cups correctly, our p-value (0.25) would still not lead us to reject the null hypothesis.

If we have three cups (filled with Pepsi, Coke and Coke, respectively), there are now eight possible ways in which we can guess, which lead to 0, 1, 2 or 3 correct guesses. The possibilities are listed in Table 2.6. Here we see that even if we were to guess correctly the contents of each cup, the evidence that we actually know what we are doing is still low (p-value = 0.125). So in

Table 2.5: Enumeration of Outcomes from Two-Cup Experiment (Y: Correct, N: Incorrect).

Actual Soda in Cups					
Pepsi	Coke	# Correct	Frequency	Proportion	p-value
N	N	0	1	0.25	$0.25 + 0.50 + 0.25 = 1.0$
N	Y	1	2	$2 \times 0.25 = 0.50$	$0.50 + 0.25 = 0.75$
Y	N	1			
Y	Y	2	1	0.25	0.25

Table 2.6: Enumeration of Outcomes from Three-Cup Experiment (Y: Correct, N: Incorrect).

Actual Soda in Cups						
Pepsi	Coke	Coke	# Correct	Frequency	Proportion	p-value
N	N	N	0	1	0.125	1.0
N	N	Y	1	3	$3 \times 0.125 = 0.375$	$0.375 + 0.375 + 0.125$
N	Y	N	1			$= 0.875$
Y	N	N	1			
N	Y	Y	2	3	$3 \times 0.125 = 0.375$	$0.375 + 0.125 = 0.500$
Y	N	Y	2			
Y	Y	N	2			
Y	Y	Y	3	1	0.125	0.125

this case guessing all of the cups correctly would still lead us to not reject the null hypothesis.

While we will not enumerate the outcomes, Table 2.7 presents the outcomes from both four-cup and five-cup experiments. Here there is still only one way of getting them all correct, but the number of ways in which we can get 0, 1, 2, 3 (or 4) correct answers is quite larger than previously seen. Note that if we select all of the cups correctly in the five-cup experiment, we get a p-value of 0.03125 (in the four-cup case, we still get a high p-value for guessing all cups correctly: p-value = 0.0625). So in this case, the only way we could convince someone that we know how to discern between the tastes of Pepsi and Coke is if we guessed the contents of 5 cups correctly, since there is a small likelihood that we could guess our way to 5 correct cups if we didn't know what we were doing.

In each of these cases, the number of correct guesses follows a *binomial distribution*, where the probability of a given number of correct guesses depends upon both the probability of any given event and the number of different ways in which that outcome can occur. Using this method, we can also calculate the probability (in the form of a p-value) of observing 33 or more successes out of 200 trials assuming the actual rate is 0.10. This value comes out to 0.002916, which is sufficiently small compared to our significance level of $\alpha = 0.05$, and thus we reject the null hypothesis that the population success rate is 0.10 in favor of the alternative that the population success rate is larger.

Table 2.7: Outcomes from Four- and Five-Cup Experiments.

Four-Cups				Five-Cups			
# Correct	Frequency	Proportion	p-value	# Correct	Frequency	Proportion	p-value
0	1	0.0625	1.0000	0	1	0.03125	1.00000
1	4	0.2500	0.9375	1	5	0.15625	0.96875
2	6	0.3750	0.6875	2	10	0.31250	0.81250
3	4	0.2500	0.3125	3	10	0.31250	0.50000
4	1	0.0625	0.0625	4	5	0.15625	0.18750
				5	1	0.03125	0.03125

We use the `binom.test()` function to calculate the *exact binomial test* in R. The general syntax is similar to the test on proportions using the `prop.test()` function and is given by:

```
binom.test(x, n, p=p0, alternative = c("two.sided", "less",
          "greater"), conf.level = 0.95)
```

where `x` is the number of successes, `n` is the number of trials, `p0` is the hypothesized value, `alternative` corresponds to which type of alternative hypothesis you have (with options: `"two.sided"` `"less"` and `"greater"`) and `"conf.level"` is the desired confidence level (thus, the significance level is *one minus* the confidence level). The statements `x`, `n`, and `p0` are required for a hypothesis test. The default values are 0.5 for `p0`, `"two.sided"` for `alternative`, and 0.95 for `conf.level`. For the example where we have 33 successes in 200 trials we can calculate the exact p-value as given in Program 3.

Note in Program 3 that the output is similar to that of `prop.test()`. Here we see the *exact* p-value is given by 0.002916. Using the $\alpha = 0.05$ significance level we would conclude that it is *likely* that the population success rate is greater than 0.10.

2.9 Communicating the Results (IMRaD Write-Up)

The following write-up is an example of the material that we need to communicate if we had actually conducted the example study used for the majority of concepts included in this section. This will form a template of the material that you should include in the write-ups for actual data analyses, though you must note that the specific material that we include in any given IMRaD write-up will depend upon the types of statistical methods we use as well as the specific research question at hand (which will itself call for additional material than what is provided here).

Introduction: Treatments designed to treat certain diseases or conditions often have adverse side-effects that can complicate a patient's reaction to

Program 3 Program to conduct an *exact* hypothesis test on a single proportion.

Code:

```
binom.test(x=33, n=200, p=0.1, alternative="greater")
```

Output:

```
    Exact binomial test

data:  33 and 200
number of successes = 33, number of trials = 200, p-value =
0.002916
alternative hypothesis: true probability of success is
greater than 0.1
95 percent confidence interval:
 0.1232791 1.0000000
sample estimates:
probability of success
                 0.165
```

the treatment, and can ultimately result in a worse disease or condition prognosis. Clinicians and practitioners are then interested in treatments that have minimal to no side-effects. It was of interest to determine whether the proportion of patients undergoing a particular treatment who experienced little to no adverse side-effects was greater than 0.10.

Methods: The frequency of subjects reporting reduced side-effects from treatment out of 200 subjects is reported, and the proportion of subjects reporting reduced side-effects is summarized with a sample proportion and a 95% confidence interval. We test the null hypothesis of a 0.10 success rate ($H_0 : p = 0.10$) against a one-sided alternative hypothesis that the success rate is greater than 0.10 ($H_A : p > 0.10$) by using a chi-square-test with significance level $\alpha = 0.05$. We will reject the null hypothesis in favor of the alternative hypothesis if the p-value is less than α; otherwise we will not reject the null hypothesis. The R statistical software was used for all analyses.

Results: Assuming that the sample was representative and subjects were independent, the sample was large enough to conduct the statistical analysis. Out of a sample of 200 total patients, 33 reported reduced symptoms ($\hat{p} = 0.165, 95\%CI : 0.120, 0.223$). Using this data, a chi-square test yielded a p-value of 0.0011, which is less than the stated significance level. Thus, we reject the null hypothesis in favor of the alternative hypothesis.

Discussion: The sample data suggest that the proportion of patients who reported reduced side-effects using this treatment is greater than 0.10. Thus, clinicians and practitioners interested in treating patients with reduced treatment-related side-effects may wish to consider this treatment.

2.10 Process

1. State research question in form of testable hypothesis.
2. Determine whether assumptions are met.
   (a) Representative sample.
   (b) Independent measurements.
   (c) Sample size: calculate expected frequencies
      i. If $np_0 > 5$ and $n(1 - p_0) > 5$, then use z-test or chi-square test (in R).
      ii. If $np_0 < 5$ or $n(1 - p_0) < 5$, then use binomial test.
3. Summarize data.
   (a) If $np_0 > 5$ and $n(1 - p_0) > 5$, then report: frequency, sample size, sample proportion and CI.
   (b) If $np_0 < 5$ or $n(1 - p_0) < 5$, then report: frequency and sample size.
4. Calculate test statistic.
5. Compare test statistic to critical value or calculate p-value.
6. Make decision (reject H_0 or fail to reject H_0).
7. Summarize with IMRaD write-up.

2.11 Exercises

1. A researcher is interested in the proportion of active duty police officers who pass the standard end of training fitness test. They take a random sample of 607 officers from a major metropolitan police force and administer the fitness test to the officers. They find that 476 of the officers were able to successfully pass the fitness test. Create 96% confidence interval for the proportion of all active duty police officers on the police force that can pass the fitness test.
2. Occupational health researchers are interested in the health effects of sedentary occupations such as call center workers. Specifically they are interested in lower back pain. They conduct a survey of 439 call center workers and record whether or not the worker has back pain at the end of their shift. The surveys show that 219 workers reported back pain. In the general population there are reports that 25% of workers have back pain. Conduct a hypothesis test to determine if call center workers have a higher rate of back pain than the general population of workers.

3. In December 2012 Gallup Poll conducted a survey of 1,015 Americans to determine if they had delayed seeking healthcare treatment due to the associate costs. Of the participants, 325 reported delaying seeking treatment due to costs. Create a 95% confidence interval for the proportion of all Americans who have delayed seeking healthcare treatment due to costs.

4. Norman et al. (2013) Consider the preference for walkability of neighborhoods for obese men. They studied 240 obese men and asked them their preference for walking behavior in neighborhoods. They found that 63 responded as they walked for transportation. Create a 99% confidence interval for the proportion of obese men who walk for transportation.

5. Barrison et al. (2001) are interested in the reasons that proton pump inhibitors were prescribed. Of the 182 gastroenterologists 122 of them prescribed proton pump inhibitors to patients. Create a 98% confidence interval for the proportion of all gastroenterologists who prescribe proton pump inhibitors.

6. Barrison et al. (2001) were interested in the proportion of physicians who deemed that proton pump inhibitors (PPI) should be sold over the counter. Of the 391 physicians surveyed 59 responded that PPIs should be sold over the counter. Create a 92% confidence interval for the proportion of physicians who think that PPIs should be sold over the counter.

7. Keightley et al. (2011) are concerned with obese peoples self perceptions. They hypothesize that a majority of obese people can identify their own body type. They conduct a study with 87 obese people and find that 7 can correctly identify their body type. Conduct a hypothesis test to determine whether or not their hypothesis is warranted.

8. Salerno et al. (2013) is interested in determining the current infection rate of Chlamydia and Gonorrhea infections. They obtained a sample of 508 high school students who consented to a urine test to screen for the two diseases. Of the participants 46 tested positive for at least one of the diseases. Create a 99% confidence interval for the proportion of all high school students who have one of the two diseases.

http://www.springer.com/978-1-4614-8707-4

Statistical Research Methods
A Guide for Non-Statisticians
Sabo, R.; Boone, E.
2013, IX, 214 p. 33 illus., 13 illus. in color., Hardcover
ISBN: 978-1-4614-8707-4