Chapter 8:

Proportions Tests

In the previous three chapters, we learned how to perform tests of population means. In each case, we required that the measurements were interval/ratio level data. Without that assumption, mathematics on the measurements would have little meaning. However, not all interesting data is interval/ratio level. Some data is nominal, such as gender or religion. In the next chapter, we will learn how to perform some tests on multi-category data. In this chapter, we concern ourselves with binary data.

This chapter continues testing simple hypotheses. Whereas in the past we have tested population means, here we test population *proportions*. All tests herein stand upon the Binomial distribution or its Normal approximation.

❧ ❧ ❧

An opinion poll performed prior to the 2012 US Presidential election found that 370 of the 900 respondents stated they would vote for Barack Obama; 340 for Mitt Romney; and 290 would not vote. What is a symmetric 95% confidence interval on the proportion of people supporting Mitt Romney in the United States?

Not all variables are interval-level or higher. At times, you need to determine the proportion of a population that has a specific characteristic. This characteristic could be "blonde hair," or "favors Pepsi," or "is bilingual." The techniques learned in the past chapters do not directly lend themselves to such dependent variables.

However, it is rather straight-forward to test hypotheses and to estimate population proportions using our current knowledge. In this chapter, we will cover the Binomial test and the proportions test. The first is used to learn about the population proportion for a single population. The latter is used to compare population proportions of two independent populations.

8.1: The One-Population Proportion Test

The derivation of the Binomial test follows directly from our knowledge of the Binomial distribution (Appendix A.3). Recall that the Binomial distribution is the sum of n independent Bernoulli trials with a constant success probability.

The quintessential example of a Binomial experiment is flipping a coin n times and counting the number of heads. In this example, the success probability remains constant, the results of the coin flips are independent of each other, and the outcome variable is the number of successes. Thus, if we define H as the number of heads, the distribution of H is

$$H \sim Bin(n, \pi)$$

Here, n is the number of coin flips and π is the probability that the coin lands Head on any one flip. The parameter of interest is π.

8.1.1 Hypothesis Testing

To test hypotheses, one specifies the hypothesized value of π. With that information, we know all probabilities associated with the experiment.

For instance, if $n = 3$ and we hypothesize $\pi = 0.50$, then we know $\mathbb{P}[H = 0] = 0.125$, $\mathbb{P}[H = 1] = 0.375$, $\mathbb{P}[H = 2] = 0.375$, and $\mathbb{P}[H = 3] = 0.125$. We also know $\mathbb{P}[H \leq 1] = 0.500$ and $\mathbb{P}[H \leq 2] = 0.875$. If we perform this experiment and observe zero heads, then we know the corresponding p-value would be

$$\text{p-value} = \mathbb{P}[H = 0] + \mathbb{P}[H = 3] = 0.250$$

This calculation assumes that the alternative hypothesis is $\pi \neq 0.500$. If the alternative hypothesis were $\pi < 0.500$, then the p-value would just be 0.125. Were the alternative hypothesis $\pi > 0.500$, then the p-value would be 1.000.

In all cases, remember the definition of the p-value: It is the probability of observing data this extreme *or more so*, given the null hypothesis is true.

Example 8.1: At the national level, Oregon is a Democratic state. I hypothesize that 50% of all voting Oregonians will vote Democratic in the next presidential election.

To test this hypothesis, I call a random sample of $n = 1000$ Oregonians and ask them how they will vote. Of the 1000, $x = 545$ stated that they would vote for the Democratic candidate.

Do the data support my hypothesis?

Solution: Let us define the random variable D as the number of Oregonians in a sample of 1000 who state they will vote for the Democratic candidate. In this survey, I recorded $d = 615$.

Under the null hypothesis, we have

$$D \sim Bin(n = 1000, \pi = 0.50)$$

The expected value of D is $\mathbb{E}[D] = 1000 \times 0.50 = 500$. I observed $d = 545$. Thus, the p-value would be

$$\text{p-value} = \mathbb{P}[D \leq 455] + \mathbb{P}[D \geq 545]$$

The 545 came from what we observed. The 455 is the value above the expected value that is just as extreme as what we observed; that is, $455 = 500 - (545 - 500)$. Using R,

```
pbinom(455, size=1000, prob=0.50) +
    (1-pbinom(544, size=1000, prob=0.50))
```

the p-value is 0.004861736. As this is less than $\alpha = 0.05$, we conclude that there is significant evidence against the hypothesized value of $\pi = 0.50$. Since the observed proportion was $\hat{\pi} = 0.545$, we can conclude that the proportion of Oregonians voting Democratic in the next presidential election is greater than 0.50. ◇

THE P-VALUES: The calculation of the p-value is frequently not so straight forward. Its definition refers to "data as extreme or more so." Determining what is "as extreme" is not always easy.

One set of observed data is as extreme as a second set if the probability of observing them is the same. In the previous example, the probability of observing $D = 545$ is 0.0004388554. Thus, to find the corresponding value below 500 that is just as extreme as this, we need to search through all possible values and select the largest one with a probability less than this:

```
for(x in 0:500) {
  targetProbability = dbinom(545, size=1000, prob=0.500)
  thisProbability   = dbinom(  x, size=1000, prob=0.500)
  if( thisProbability > targetProbability ) break
}
print(x-1)
```

Running this snippet gives a value of 455. Thus, observing 455 is just as extreme as observing a value of 545.

> ***Note***: This value is not too surprising: 455 = 1000 – 545. However, when the underlying distribution is not symmetric, calculating the "other extreme value" requires some type of loop.

One option for estimating the p-value is to just use the doubling rule. If the underlying distribution is continuous, this will be exact. Since the underlying distribution is discrete, the resulting p-value will be slightly different than it should be.

EXAMPLE 8.2: At the national level, Oklahoma is a Republican state. I hypothesize that 66% of all voting Oklahomans will vote Republican in the next presidential election.

To test this hypothesis, I call a random sample of $n = 1000$ Oklahomans and ask them how they will vote. Of the 1000, $x = 615$ stated that they would vote for the Republican candidate.

Do the data support my hypothesis?

Solution: Again, let us define the random variable R as the number of Oklahomans in a sample of 1000 who state they will vote for the Republican candidate. In this survey, I recorded $r = 615$.

Under the null hypothesis, we now have

$$R \sim Bin(n = 1000, \pi = 0.66)$$

The expected value of R is $\mathbb{E}[R] = 1000 \times 0.66 = 660$. I observed $r = 615$. Thus, the p-value would be

$$\text{p-value} = \mathbb{P}[R \leq 615] + \mathbb{P}[R \geq ???]$$

To determine the upper extreme value, symbolized with ??? here, run the code snippet from above, with the appropriate changes:

```
for(x in 1000:500) {
  targetProbability = dbinom(615, size=1000, prob=0.66)
  thisProbability   = dbinom(  x, size=1000, prob=0.66)
  if( thisProbability > targetProbability ) break
}
print(x+1)
```

This gives a value of 705. The p-value is

$$\text{p-value} = \mathbb{P}[R \leq 615] + \mathbb{P}[R \geq 705] = 0.002956647$$

the p-value is 0.002956647. As this is less than $\alpha = 0.05$, we conclude that there is significant evidence against the hypothesized value of $\pi = 0.66$. Since the observed proportion was $\hat{\pi} = 0.615$, we can conclude that the proportion of Oklahomans voting Republican in the next presidential election is less than 0.66.

For the record, the doubling approximation is

$$\text{p-value} = 2 \times \mathbb{P}[R \leq 615] = 0.003225845$$

This is larger than the true p-value, although the same substantive conclusion holds. ◇

The P-Values, Again*: The definition of the p-value is that it is the probability of observing data this extreme, or more so, given the null hypothesis is true. A more general way of calculating the p-value is just to add the probability of all possible outcomes that are as extreme or more extreme that what was observed.

There is nothing different in this statement. The difference comes when the underlying distribution is not unimodal. The following steps to calculate the p-value will always work:

1. Calculate the likelihood of the observed data, given the null hypothesis is true.
2. Determine which values are as likely, or less likely, than what was observed.
3. Add the probabilities of each of these outcomes.

This sum will be the p-value. The next example shows how to do this general method in R.

Example 8.3: An associate hypothesized that more people preferred Pepsi to Coke. To test this hypothesis, she interviewed $n = 1000$ people. Of those people, 535 stated that they preferred Pepsi.

Do the data support her hypothesis?

Solution: Let us define X as the number of people in a sample of 1000 who prefer Pepsi to Coke. We observed $x = 535$. We hypothesize

$$X \sim Bin(1000, \pi = 0.500)$$

This is a nice distribution, because it is unimodal and symmetric. Thus, any of the above techniques will work, allowing us to check our work.

Here, however, let us use the general method for calculating p-values:

```
obs = dbinom(535, size=1000,prob=0.500)
out = 0:1000
ext = which( dbinom(out,size=1000,prob=0.500) <= obs )
sum( dbinom(out[ext],size=1000,prob=0.500) )
```

The first line calculates the likelihood of what was observed. The second line defines the sample space. The third line determines which elements of the sample space are as extreme, or more so, than what we observed. The fourth line calculates the p-value as the sum of those extreme probabilities.

Running this, we get a p-value of 0.02906112. As this is less than $\alpha = 0.05$, we reject the non-directional null hypothesis and conclude that the proportion of people preferring Pepsi over Coke is not 50%.

Note, however, that the stated null hypothesis was that more people preferred Pepsi to Coke; that is: $\pi_P > 0.500$. To determine the p-value associated with *this* hypothesis, run

```
obs = dbinom(535, size=1000,prob=0.500)
out = 500:1000
ext = which( dbinom(out,size=1000,prob=0.500) <= obs )
sum( dbinom(out[ext],size=1000,prob=0.500) )
```

This gives a p-value of 0.01453056. Here, we reject the null hypothesis and conclude that more people prefer Pepsi to Coke. ◇

8.1.2 Confidence Intervals

In general, a confidence interval is a set of values for the population parameter for which the observed data is reasonable. For this chapter, that means a confidence interval is a set of values for π. The observed data correspond to p-values greater than or equal to α (or $\alpha/2$) for each value in the confidence interval.

Thus, one way of calculating a confidence interval is to run through every possible value for the parameter, calculate the p-value for each of those values, and report those which give p-values greater than α (or $\alpha/2$).

This is the most general method.

In the Pepsi example, the code to calculate an upper bound to a 95% confidence interval is

```
ucl = seq(0.535,1, length=10000)
dd  = which(pbinom(535, size=1000, prob=ucl)>=0.025)
max(ucl[dd])
```

This gives an upper bound of 0.5662511. Similarly, a lower bound is 0.5045555:

```
lcl = seq(0,0.535, length=10000)
dd  = which(pbinom(535, size=1000, prob=lcl)<=0.975)
min(lcl[dd])
```

Thus, one 95% confidence interval for the proportion of people who prefer Pepsi to Coke is from 0.505 to 0.566.

Confidence Interval Length*: Notice that people refer to *a* confidence interval, never to *the* confidence interval. This is because there are an infinite number of confidence intervals that meet the confidence-level criterion.

Usually, confidence intervals are selected based on tradition, which means they are selected based on ease of calculation. This usually means they are "central" or "symmetric" confidence intervals. These are intervals of the form $\bar{x} \pm E$, where E is the margin of error.

Margin of Error

A different criterion to use in selecting confidence intervals is to select the one with the smallest length. The general method above will produce a minimal-length confidence interval (within rounding).

If the underlying distribution is symmetric, unimodal, and continuous, the central confidence interval will also be the minimal-length confidence interval. In the case of an asymmetric Binomial distribution, this will not be the case.

The width of the interval given above is 0.061. Using the usual method for calculating a confidence interval (given below), the width is 0.063.

8.1.3 The R Function

The previous sections are important for better understanding p-values and confidence intervals. However, there is a function in R that performs these calculations rather quickly.

The function is `binom.test`. It takes five parameters: the number of successes, the number of trials, the hypothesized population proportion, the direction of the alternative hypothesis, and the confidence level.

Example 8.4: Recently, a researcher sought to determine if people could tell the difference between the different colors of Skittles. To determine this, she fed Skittles to blindfolded people, recording both the color and flavor stated.

Of the $n = 253$ Skittle tastings by the 11 people, $x = 163$ were called correctly.

Do the data suggest the Skittle colors are not related to the flavors?

Solution: There are five Skittle colors (flavors). If there were no relationship between the colors and the flavors, one would expect to get the color correctly 20% of the time. Thus, if we define X as the number of Skittles guessed correctly, the null hypothesis is

$$X \sim Bin(n = 253, \pi = 0.20)$$

According to the Binomial test, a 95% confidence interval for the proportion of Skittle colors correctly chosen is between 0.5819 and 0.7032. The p-value associated with the null hypothesis that $\pi = 0.20$ is less than one in 10,000. Because the p-value is so small, we reject the null hypothesis and conclude that the colors and flavors are not independent:

```
binom.test(x=163, n=253, p=0.20)
```

That line of code produces the following output:

```
        Exact binomial test

data:  163 and 253
number of successes = 163, number of trials = 253, p-value
    < 2.2e-16
alternative hypothesis: true probability of success is not
   equal to 0.2
95 percent confidence interval:
 0.5818797 0.7032329
sample estimates:
probability of success
             0.6442688
```

Instead of the default confidence interval given by R, we could use the script above to calculate a minimal-length confidence interval. According to that script, the minimal-length confidence interval is from 0.5859 to 0.7032.

```
ucl = seq(163/253,1, length=1e6)
dd  = which(pbinom(163, size=253, prob=ucl)>=0.025)
max(ucl[dd])

lcl = seq(0,163/253, length=1e6)
dd  = which(pbinom(163, size=253, prob=lcl)<=0.975)
min(lcl[dd])
```

The width of the usual confidence interval is 0.1213; of the minimal-length interval, 0.1173. ◇

8.2: The Two-Population Proportion Test

Recall Chapter 6, where we generated a two-sample test based on our knowledge of the one-sample test. In that case, the hypothesized distribution generating the data was the Normal distribution. We relied on the fact that the difference between two Normal distributions is also a Normal distribution.

This method will not work with the Binomial distribution. The difference between two Binomial distribution is *not* another Binomial distribution. To easily see this, note that a Binomially-distributed random variable must be non-negative but that the difference between two such random variables may be negative.

This means we have two options when exploring the difference between two Binomial populations. The first is to use the Normal approximation to the Binomial distribution. The second is to use Monte Carlo simulation to estimate the distribution of the difference. In this section, we do the former. In the next section, we do the latter.

8.2.1 The Proportions Test

Appendix C discusses the Central Limit Theorem. In short, for distributions with a finite variance, sums and averages of independent random variables tend to the Normal distribution. This is what makes the Normal distribution so important in statistics.

The Binomial distribution is the sum of independent Bernoulli random variables. It has a finite variance. As such, the Binomial distribution converges to the Normal distribution with expected value $n\pi$ and variance $n\pi(1-\pi)$. In other words, if X is a Binomially-distributed random variable, then

$$X \sim Bin\Big(n, \pi\Big) \dot{\sim} \mathcal{N}\Big(n\pi,\ n\pi(1-\pi)\Big)$$

As n gets larger, this approximation is better.[1]

Thus, to compare the population proportions of two Binomially-distributed populations, we have the following approximations:

$$P_x := \frac{X}{n_x} \dot{\sim} \mathcal{N}\left(\pi_x,\ \frac{\pi_x(1-\pi_x)}{n_x}\right)$$

[1] This approximation can also be made better by applying certain "continuity corrections" to the random variable. For this discussion, we will ignore these corrections.

$$P_y := \frac{Y}{n_y} \dot{\sim} \mathcal{N}\left(\pi_y,\ \frac{\pi_y(1-\pi_y)}{n_y}\right)$$

This means that the difference between the two population proportions is distributed

$$P_x - P_y \dot{\sim} \mathcal{N}\left(\pi_x - \pi_y,\ \frac{\pi_x(1-\pi_x)}{n_x} + \frac{\pi_y(1-\pi_y)}{n_y}\right)$$

If, as is frequently the case, we wish to test if the two population proportions are equal, this reduces to

$$P_x - P_y \dot{\sim} \mathcal{N}\left(0,\ \pi(1-\pi)\frac{n_x + n_y}{n_x n_y}\right)$$

From this, we can get the p-values and confidence intervals as we did back in Chapter 6.

Example 8.5: In Oklahoma, the speed limit in school zones when children are *not* present is 30mph. In an effort to reduce the proportion of people speeding through the zones, the Stillwater City Council estimated the proportion of people speeding through a specific school zone, installed an electronic sign that flashed when the car traveled faster than 30 mph, and then estimated the proportion of cars speeding through the same zone.

Before installing the sign, 50 of 250 cars traveled faster than 35 mph in the school zone. After installing the flashing warning sign, 75 of 315 cars traveled faster than 35 mph in the school zone.

Did installing the sign result in a significant change in speeding proportions?

Solution: The null hypothesis, as the question is written, is $H_0 : \pi_b = \pi_a$. Here, π_b is the proportion of people speeding **b**efore the flashing sign was installed, and π_a is the proportion speeding **a**fter the sign was installed. From the discussion above, we have

$$P_x - P_y \dot{\sim} \mathcal{N}\left(0,\ \pi(1-\pi)\frac{n_x + n_y}{n_x n_y}\right)$$

From this, we get a p-value of 0.2785786 and a 95% confidence interval from -0.1070 to 0.0308. Both indicate that we are unable to detect a difference in

the proportion of those speeding before and after the signs were installed. There is no significant evidence that the signs helped curb speeding in the school zone. ◇

The above was done without using a continuity correction. As such, its accuracy relies on a large sample size. Unfortunately, without a continuity correction, the needed sample sizes for a good estimate are rather large. It is better to use a continuity correction, which the R function does by default:

```
prop.test( x=c(50,75), n=c(250,315) )
```

Note that, since we are dealing with two populations, we need to give the function at least four values: the numbers of successes for sample 1 and for sample 2, and the numbers of trials for sample 1 and for sample 2. The function above gives an output of

```
        2-sample test for equality of proportions with
    continuity correction

data:  c(50, 75) out of c(250, 315)
X-squared = 0.9633, df = 1, p-value = 0.3263
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1100258  0.0338353
sample estimates:
   prop 1    prop 2
0.2000000 0.2380952
```

The Chi-Square Test*: If you look closely at the output from the `prop.test` function, you will notice that the test statistic is `X2`, which is the symbol for the standard Chi-Square distribution (Appendix B.4). The sum of ν independent squared standard Normal random variables has a Chi-Square distribution with ν degrees of freedom.

Because of this relationship between the Normal distribution and the Chi-Square distribution, the Chi-Square distribution is typically used to compare population proportions. While R could have used the Normal distribution above, the Chi-Square distribution is necessary when there are more than two possible outcomes per population. This, we will cover in Chapter 10.

8.2.2 Monte Carlo Simulation

The reason that the proportions test is only approximate is that we do not know the exact distribution of the test statistic. In lieu of using the Normal approximation, which requires moderately large sample sizes, we can simulate the distribution of the test statistic. This will give us a distribution against which to compare our observed data.

to illustrate the method, let us return to the speed zone example from above.

Example 8.6: Before installing the sign, 50 of 250 cars traveled faster than 35 mph in the school zone. After installing the flashing warning sign, 75 of 315 cars traveled faster than 35 mph in the school zone.

Solution: The first step is to determine an appropriate test statistic. Many will work, as long as the test statistic is directly related to the parameters of interest. Here, because we are comparing the difference in population proportions, let us use the difference in the sample proportions as our test statistic.

Next, we need to generate random draws from the appropriate distributions. From above, we know

$$X \sim Bin(n_x,\ \pi_x) = Bin(250, \hat{\pi})$$
$$Y \sim Bin\Big(n_y,\ \pi_y\Big) = Bin(315, \hat{\pi})$$

Notice that these are the exact distributions. Here, $\hat{\pi}$ is the observed speeding proportion, $\hat{\pi} = (50+75)/(250+315) \approx 0.2212$.

Next, we repeatedly draw from these two distributions and calculate the test statistic for each.

Finally, we compare our observed test statistic ($\frac{50}{250} - \frac{75}{315} \approx -0.0381$) to the distribution above. This gives our p-value and a confidence interval.

The actual code for the p-value is

```
obs = 50/250-75/315

X = rbinom(1e6, size=250, prob=125/565 )
Y = rbinom(1e6, size=315, prob=125/565 )

Px = X/250
Py = Y/315

TS = Px-Py

mean( TS<=obs ) * 2
```

According to this, the p-value is 0.280. Thus, we lack sufficient evidence that the new signage reduced the speeding proportion.

The code for the confidence interval is

```
X = rbinom(1e6, size=250, prob=50/250 )
Y = rbinom(1e6, size=315, prob=75/315 )

Px = X/250
Py = Y/315

TS = Px-Py

quantile( TS, c(0.025,0.975) )
```

Thus, a 95% confidence interval for the difference in speeding proportions is from -0.106 to 0.030. ◇

Note: Remember that p-values are based on the null hypothesis and the data. The confidence interval is based solely on the data. This is why the distribution of X and Y differ between the two calculations. For the p-value calculation, the X and Y values are generated from the null hypothesis. For the confidence interval calculation, they are generated from the observed data.

8.3: Conclusion

This chapter introduced methods for learning about population proportions. In previous chapters, we examined population means.

Because we dealt with population proportions, the Binomial distribution became very important. It served as the underlying distribution for observed cell counts. The one-sample test for proportions was the Binomial distribution.

Unfortunately, when exploring differences between two population proportions, we could not justify the distribution of the test statistic being Binomial. We had two options: use the Normal approximation or use Monte Carlo simulation. The first led to the proportions test with its test statistic distributed Normal or Chi-Square. The second led to a simulated distribution, from which we could obtain the p-value and a confidence interval.

One assumption that we made throughout this chapter is that the counts were distributed according to a Binomial distribution. This is usually an easy assumption to make as the counts really are the sum of *independent* Bernoulli events. However, this assumption does not always hold. When dealing with geographic groups, it is likely that the events are not independent; like people tend to live near each other. When the Bernoulli events are not independent, their sums will not be Binomial. If their sums are not Binomial, this chapter is of little use.

8.4: End of Chapter Materials

8.4.1 R Functions

In this chapter, we were introduced to several R functions that will be useful in the future. These are listed here.

Statistics:

binom.test(x,n,p) Returns the p-value and a confidence interval (95% is the default) for `x` successes in `n` trials against the hypothesis that $\pi =$ `p`. By default, `p` = 0.500.

prop.test(x=c(x1,x2),n=c(n1,n2),p) Returns the p-value and a confidence interval (95% is the default) for `x` successes in `n` trials against the hypothesis that $\pi =$ `p`. Here, however, `x` is a vector of two values, as is `n`. By default, `p` = 0.500.

Probability:

rbinom(n, size, prob) This returns `n` random values drawn from a Binomial distribution with `size` trials and success probability `prob`.

rnorm(n, m, s) This returns `n` random values drawn from a Normal distribution with mean `m` and standard deviation `s`. By default, `m` = 0 and `s` = 1.

8.4.2 Exercises and Extensions

This section offers suggestions on things you can practice from this chapter.

Summary:

1. What are the two parameters for a Binomial distribution?
2. What are the two parameters for a Normal distribution?
3. What is the ratio of the variance to the mean for a Binomial distribution?
4. Can the ratio of the variance to the mean be greater than 1? Either give an example where it is greater than one or explain why it is impossible.
5. What is the relationship between a Normal distribution and a Chi-Squared distribution?

Theory:

6. Let $Y \sim \mathcal{N}(0,1)$. What is the distribution of Y^2?
7. Let $Y \sim \mathcal{N}(1,1)$. What is the mean of Y^2? What is the variance of Y^2?
8. Let $Y \sim \mathcal{N}(2,1)$. Calculate $\mathbb{P}[Y \leq 2]$.
9. Let $Y \sim \mathcal{N}(3,1)$. Define $Z = Y - 3$. What is the distribution of Z^2? Calculate $\mathbb{P}\left[Y^2 \leq 9\right]$. Calculate $\mathbb{P}\left[Y^2 \leq 1\right]$.
10. Let $X \sim Bin(100, 0.500)$. What is Normal approximation for X? Calculate $\mathbb{P}[X \leq 50]$ for both the exact distribution and for the approximation. Comment on the closeness of the estimates to each other.
11. Let $W \sim Bin(3, 0.001)$. What is the Normal approximation for W? Calculate $\mathbb{P}[W \leq 1]$ for both the exact distribution and for the approximation. Comment on the closeness of the estimates to each other.

Data:

12. At a different school zone, a researcher randomly sampled from all people passing through. On Monday morning, 15 of 290 traveled faster than 35 mph. On Friday afternoon, 75 of 182 traveled faster than 35 mph. Is there significant evidence that Friday afternoon drivers speed more often through the school zone than do Monday morning drivers?

    In addition to the above hypothesis test, provide 95% confidence intervals for the proportion of cars speeding through the zone on Monday mornings and on Friday afternoons.

13. The `someCollege` datafile contains seven variables measured on an allegedly random sample of the students at a small liberal arts college. Assuming this is indeed a random sample, what proportion of the students are from a public high school? Are males more or less likely to be home schooled?

    In both cases, provide 95% confidence intervals for the population proportion.

Monte Carlo:

14. A certain manufacturer claims that 99.99% of their USB drives are free of defects. To test this hypothesis, I test 100 of their USB drives. Of those 100, only 1 was defective. Do the data provide sufficient evidence against the manufacturer's claim? Provide both a p-value and a confidence interval.

    Use the Binomial test, the proportions test, and Monte Carlo simulation to answer this question. Compare and contrast the three methods.

15. A researcher stated that only 1% of all students at a specific university could locate Sri Lanka on a world map. To test this, he asked 15 people to locate Sri Lanka; one could. Do the data support his hypothesis?

    Use the Binomial test, the proportions test, and Monte Carlo simulation to answer this question. Compare and contrast the three methods.

8.4.3 Applied Research

This section offers some readings that are connected with the topics in this chapter.

-

8.4.4 References and Additional Readings

- Alan Agresti. (2002) *Categorical Data Analysis*, Second Edn. New York: Wiley-Interscience.
- Alan Agresti. (2007) *An Introduction to Categorical Data Analysis*. New York: Wiley Series in Probability and Statistics.
- Alan Agresti. (2010) *Analysis of Ordinal Categorical Data*, Second Edition New York: Wiley Series in Probability and Statistics.
- Alan Agresti and Brent A. Coull. (1998) "Approximate Is Better than 'Exact' for Interval Estimation of Binomial Proportions." *The American Statistician*, **52**(2): 119-126.
- Lee J. Bain and Max Engelhardt. (1992) *Introduction to Probability and Mathematical Statistics*, 2nd edn. Brooks/Cole: Belmont, CA.
- William C. Navidi. (2006) *Statistics for Engineering and Scientists*, 2nd edn. McGraw-Hill: New York.