Mini-Project 1: One-proportion tests

Math 247: Statistics with Applications

Warren Atkison

Outline

Come up with a research question that interests you.The research question must be answerable using a *single categorical variable*! (You could start with non-binary categorical variable, but you will need to make it into a binary categorical variable for the analysis).

Some examples are:

- Is the dominant hand faster than the non-dominant hand (measuring number of snaps in 30 seconds)?
- Do more people wear sneakers than shoes when going to the grocery store?
- Do people prefer Coke or Pepsi?
- In a typical day, do residential students eat dinner on or off campus?

I am sure you will produce more interesting and original research questions. You are given an option of collecting your own data to answer your research question or to use the results of the Math 247 survey. Be creative!

Project questions

Answer the following questions in detail. The total number of points is 25. I reserve the right to deduct points for poor document organization.

1. What is your research question? Note: Your research question and survey question are not the same thing! [1 point]

Are people who drink coffee more likely to feel stress than non-coffee drinkers?

2. What are your observational units? [1 point]

Coffee drinkers

3. What is the variable being recorded? Is it a binary variable? If not, make it into a binary categorical variable [1 point]

I will take the proportion of coffee drinkers that feel stressed often or always vs the total coffee drinkers. Our variable is is someone felt stressed often or always, or not.

4. What type of data will you collect/use? If you chose the option of collecting your own data describe the process of data collection in details. Be specific! It should be clear enough so your instructor could replicate the experiment. If you chose to use Math 247 survey results describe why you chose that variable, sample size and type of the sample.

I will use the data provided in the synthetic survey, specifically the data on cups of coffee drank a day and the how stressed everyone felt. I'm only interested in coffee drinkers, of which there where 133 and I want to know if they felt stressed often or always. This is however a convenience sample.

5. What is considered a "success" in this case? Be specific! [1 point]

If someone feels stressed often or always

6. Describe your parameter of interest in words. What is the value of π? Justify this value. [1 point]

long term proportion of coffee drinkers who feel stressed often or always. For our value of π, we will take the proportion of the whole survey of people who feel stressed often or always. In this case, $\pi = 128/200$

7. What are the null and alternative hypotheses? Use the symbolic notation! [1 point]

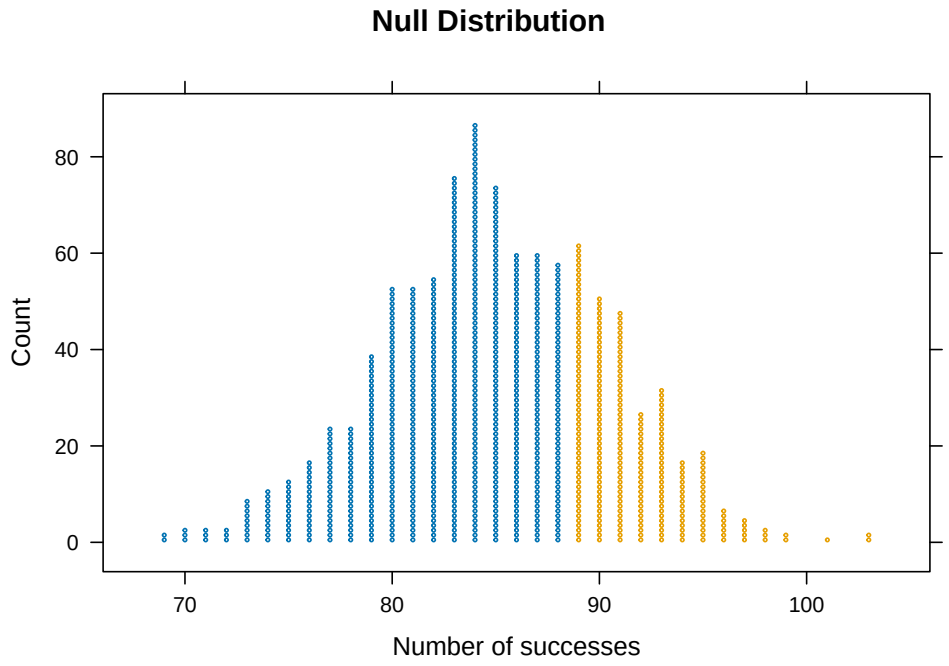
$H_0 : \pi = 0.64$
$H_A : \pi > 0.64$

8. How many successes did you find in the "n trials"? Where does the value of n come from here? [1 point]

There were a total of 133 coffee drinkers surveyed, and of them 89 felt stressed often or always

9. Use R to create/simulate the null distribution and visualize it with a dotplot. [1 point]

```
set.seed(1234)
Null.sim <- do(1000) * rflip(n = 133, prob = 0.64)
dotPlot(~heads, data = Null.sim, width = 1, cex=1, groups = (prop >= 89/133),
xlab = "Number of successes",
main = "Null Distribution")
```




```
round(favstats(~prop, data = Null.sim)["mean"],2)
```

```
##  mean
##  0.64
```

```
round(favstats(~prop, data = Null.sim)["sd"],3)
```

```
##     sd
##  0.041
```

10. Where is the null distribution centered? (You should have a specific value!). Does this make sense? Why? [1 point]

It's centered at 0.64, which is the estimated mean of the population.

11. How much is this distribution varying? (You should have a specific value!) [1 point]

The standard deviation is 0.041.

12. Use Simulation methods to answer questions *a* through *d*.

- a. Find the value of the standardized statistic. Use your standardized statistic to evaluate the strength of evidence. [1 point]

$$z = \frac{89/133 - 0.64}{0.041} = 0.711$$

We are less than one standard deviation above the mean.

- b. Use R to find the p-value. Interpret it using the definition of p-value. [1 point]

```
p.value <- prop(~(prop >= 89/133), data = Null.sim)
p.value
```

```
## prop_TRUE
##     0.276
```

- c. Use the p-value to evaluate the strength of evidence. [1 point]

There is a 27.6% chance of this statistic occurring if the null hypothesis is true

- d. Using your findings from parts a and c, what is your conclusion about the null hypothesis? Explain fully. [1 point]

We cannot reject the null hypothesis, as we do not have sufficient evidence against it since our p value was above 0.10.

13. Use Theory-based method to answer questions *a* through *e*.

- a. Find the value of the standardized statistic. Use your standardized statistic to evaluate the strength of evidence. [1 point]

```
n <- 133
pi <- 0.64
p.hat <- 89/133
sd <- sqrt(pi*(1-pi)/n)
z <- (p.hat - pi)/sd
z
```

```
## [1] 0.7009139
```

We are 0.7 standard deviations above the mean.

- b. Use R to find the p-value. Interpret it! [1 point]

```
p.value <- round(pnorm(z, mean=0, sd=1, lower.tail = FALSE),10)
p.value
```

```
## [1] 0.2416784
```

- c. Use the p-value to evaluate the strength of evidence. [1 point]

We do not have any evidence against the null hypothesis

- d. Using your findings from parts a and c, what is your conclusion about the null hypothesis? [1 point]

We cannot reject the null hypothesis, as we do not have sufficient evidence since our p value is above 0.10.

- e. Are validity conditions for using theory-based approach satisfied in this case? Explain. Be specific. [1 point]

Yes, since we had more than 10 successes and more than 10 failures in our statistic. However, using the overall proportion of coffee drinkers as a parameter for the population could be unfair in some way.

Use your findings from above to answer the following questions.

14. Write a few sentences in plain English, telling a friend the results of this small research you conducted and the conclusions.[1 point]

We have not found any evidence that coffee drinkers are more stressed than non coffee drinkers.

15. Did your standardized statistic lead to the same/different conclusions than the p-value that you found? (Feel free to use either the simulated or theoretical results). [1 point]

In both the simulated and theory approach, they lead to the same conclusion.

16. With this knowledge is it necessary to always find both the standardized statistic and p-value? Explain why in two sentences or less.[1 point]

No, as they represent the same thing. In the Gaussian curve, z represents your x coordinate, and the p value is the area under the curve from that coordinate to infinity or to negative infinity, depending on what side of the distribution you're interested in. If you want both then we can add the two together. We can write the p value in terms of z as follows (two sided)

$$\text{p.value} = \int_{-\infty}^{z} e^{-x^2} + \int_{z}^{\infty} e^{-x^2}$$

17. Compare your simulated p-value to your theoretical p-value. Do they lead to different results? If so, why do you think this happened? If not, what does this tell you about the two approaches? [1 point]

They lead to the same results in my case, most likely because the validity conditions were met.