Project Part 3

Data Set Description

Background Information

This data set(Reference 1) contains data on Americans' preferences and usage practices regarding the Oxford comma. I understand this is a statistics course, but I appreciated the juxtaposition of statistics regarding an English subject. Anyways, for those unfamiliar with the Oxford comma, it is the comma separating a list of three or more things before "and". For example, in the sentence: "Statistics is fun, cool, and exciting." the Oxford comma is the one after "cool" and before "and." This data set examines Americans' usage of the Oxford comma and what factors determine whether an individual uses the Oxford comma or not.

Data Organization

This data is a sample of the American population and was collected through a survey ran by FiveThiryEight and SurveyMonkey Audience and polled 1,129 Americans. FiveThirtyEight is a polling aggregate website that uses polling and statistical analyses to look at topics ranging from sports to politics to pop culture. SurveyMonkey Audience is a survey tool that allows surveyors to reach a specific audience, such as a representative sample of the American population. With SurveyMonkey Audience a surveyor is able to design a survey and pick an appropriate audience. The audience is made up of "a diverse online population that voluntarily joined a program to take surveys."(Reference 2).

Each row of the dataset then represents the survey answers for one respondent, and each one of the columns represents a different question that was asked in the survey. The main survey question was "In your opinion, which sentence is more grammatically correct?" with response choices of "It's important for a person to be honest, kind and loyal." or "It's important for a person to be honest, kind, and loyal." The second question was "Prior to reading above, have you heard of the serial (or Oxford) comma before?" with reponse choices of "Yes" or "No". The third question was "How much, if at all, do you care about the use (or lack thereof) of the serial (or Oxford) comma in grammar?" with response choices of "Not at all", "Not much", "Some", or "A lot". The fourth question was "How would you write the following sentence?" with response choices of "Some experts say it's important to drink milk, but the data is inconclusive." or "Some experts say it's important to drink milk, but the data are inconclusive.” The fifth question is then “When faced with using the word”data“, have you ever spent time considering if the word was a singular or plural noun?” with response choices of “Yes” or “No”. The sixth question was “How much, if at all, do you care about the debate over the use of the word”data" as a singular or plural noun?" with response choices of “Not at all”, “Not much”, “Some”, or “A lot”. The seventh question was “In your opinion, how important is proper use of grammar?” with response choices of “Neither important nor unimportant (neutral)”, “Somewhat important”, or “Very important”. The rest of the questions were “Gender”, “Age”, “Household Income”, “Education”, and “Location(Census Region)”.

Research Question

In this dataset, the primary response variable is the answer to the question “In your opinion, which sentence is more grammatically correct?” because this answer then dictates whether the respondent typically chooses to use the Oxford comma or not. The eleven other questions are then able to function more as explanatory variables that provide insight into why that particular respondent used the Oxford comma or not. Therefore, with this many possible explanatory variables there’s a litany of possible relationships to explore through a series of different testing. However, I’m most interested in whether the respondent having a higher or lower education level makes a difference in whether the respondent uses the Oxford comma or not. Specifically, I want to know if respondents with a higher education level use the Oxford comma more often than those with lower education levels. A higher education level would qualify as Bachelor’s degree or above and a lower education level would qualify as High school degree or below. My question is then “Do those with a higher education level use Oxford commas more often that those with a lower education level?”

Relevance of Question and Data

This question certainly shares a clear link with this dataset, as the data is primarily measuring who uses the Oxford comma and what about them shows whether they use it or not. Education level is also a more defined characteristic of the respondent whereas many of the other questions rely on the respondent’s personal opinon, and I feel an analysis of a defined characteristic explaining a personal opinion would be more interesting than an analysis of a personal opinion explaining another personal opinion. Moreover, intuitively it makes sense that a variable like Education level could have an affect on grammar. Grammar is tightly tied to English and associated with being well-educated, so those with more experience in academia may prefer the Oxford comma and may have more of an appreciation of correct grammar and value its importance. This is what leads me to believe that more educated individuals may use the Oxford comma more often than lower educated individuals.

The outcome of this question can then also be appropriately generalized to the American population, as this dataset is a sampling of the overall American population. Therefore, based on the outcome it will be possible to generalize to the American population whether an individual's higher education level makes a difference in their comma usage and makes them more likely to use the Oxford comma. Specifically, the results may be generalized to the two American populations of those with a Bachelor's degree or higher and those with a High school degree or lower.

Test

I believe the most appropriate test for answering this question is the two-sample z-test for proportions. This test fits my data well due to how all of the data is categorical and relies on counts and proportions for its categories rather than means and standard deviations. Moreover, this test was designed to test "the difference of the proportions of a certain characteristic across two independent populations"(Reference 3). In this case, the "certain characteristic" is whether the respondent preferred using the Oxford comma or not in a sentence. The "two independent populations" are those with a Bachelor's degree or higher and those with a High school degree or lower. These populations are certainly independent, as there is no obvious dependency relationship between them. The other assumptions for this test are that the samples are simple random samples, each sample includes at least 10 successes and failures, and each population is at least 20 times as big as its sample (Reference 4). All of these conditions are easily met, as the sample was random, each sample has more than 10 usages of Oxford comma and no Oxford comma, and the US population for higher educated individuals and lower educated individuals is much more than 20 times our samples.

```
## R code
# Respondents with higher education
highedu <- which(comma$Education == "Graduate degree"
                 | comma$Education == "Bachelor degree" )
highcount<-length(highedu) # Size of higher education sample (620)
# Respondents with lower education
lowedu <- which(comma$Education =="High school degree"
                | comma$Education == "Less than high school degree" )
```

```
lowcount <- length(lowedu) # Size of lower educationsample (111)
hedu <- comma[highedu, ] # Only higher education responses
OxfordHigh <- which(hedu$CommaChoice
                    == "It's important for a person to be honest, kind, and loyal.")
# How many higher educated respondents preferred the Oxford comma(371)
highcommacount <- length(OxfordHigh)
ledu <- comma[lowedu, ] # Only lower education responses
OxfordLow <- which(ledu$CommaChoice
                   == "It's important for a person to be honest, kind, and loyal.")
# How many lower educated respondents preferred the Oxford comma (60)
lowcommacount <- length(OxfordLow)
# Proportion of higher educated respondents who preferred the Oxford comma (%59.83871)
highproportion <- highcommacount / highcount
# Proportion of lower educated respondents who preferred the Oxford comma (%54.05405)
lowproportion <- lowcommacount / lowcount
phat <- c(highproportion, lowproportion)
n <- c(highcount, lowcount)
x <- phat * n
#Perform test
z2.prop <- prop.test(x, n, alternative="greater", correct=FALSE)
z2.prop
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  x out of n
## X-squared = 1.3019, df = 1, p-value = 0.1269
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.02642805  1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.5983871 0.5405405
```

```
#Verify z-value
```

```
cphat <- sum(x)/sum(n)
z <- diff(phat)/sqrt(cphat*(1-cphat)*(1/n[1]+1/n[2])) # 1.1410
## Verify p-value
p.value <- pnorm(z) # 0.1269
```

Test Conclusions

After organizing the data appropriately and finding the sample proportions for each sample, I performed the two sample z-test for proportions. This test resulted in a fairly large p-value of 0.1296, which is above the threshold of 0.05 and 0.10. Therefore, we fail to reject the null hypothesis that the proportions of the two populations for Oxford comma usage are equal. Stated non-statistically, this means that we cannot conclude from this test that individuals with a higher education use the Oxford comma more often than those with a lower education. Even though higher educated respondents used the Oxford comma 5% more often in these samples, this difference is not large enough to be generalized to the American population. Therefore, we cannot say that higher educated Americans use the Oxford comma more often than lower educated Americans.

References

1. https://github.com/fivethirtyeight/data/blob/master/comma-survey/comma-survey.csv
2. https://help.surveymonkey.com/articles/en_US/kb/SurveyMonkey-Audience
3. unit5Notes.R created by Professor Richard Ross
4. https://stattrek.com/hypothesis-test/difference-in-proportions.aspx