Applied Stats: Homework Assignment 1

Example Write-up

## Instructions

This assignment asks you to present your work with the US state data we used extensively during the first week. You will be asked to phrase research questions and answer them for single variables and relationships between two variables. Although not required, **you may re-use your work from Stations 2 & 3.** Please read through the example write-up linked in the assignment description on Moodle. It should give you an idea of the level of work expected.

The goal of this assignment is for you to use your knowledge of introductory statistics tools to pose and answer three research questions. Your work should include complete sentences, presentation of statistical figures and summaries, and in-context discussions.

Each Part (1 - 3) below should include three sections:

1. **Introduction of the variable(s)**: Introduce your variable(s), and pose of a research question both in context and in terms of the relevant population parameter. Include discussion of what type of statistical test you plan to perform to answer the question.
2. **Exploratory data analysis.** Present a visual summary and a numerical summary, and discuss what you learned from them. Discuss what you think the results of your hypothesis test will be based on your EDA. Further, include comments about whether the conditions are met for performing the hypothesis test.
3. **Inference and Conclusions.** Perform the hypothesis test you posed in the introduction, interpreting the strength of the evidence and making a formal conclusion. Also build and interpret a 95% confidence interval for your parameter of interest.

# Part 1: One quantitative variable

Choose one quantitative variable from the merged states data.

**TASK 1.1–Introduction of the variable:**

The variable of interest is *Insured*, the percentage of adults (ages 18-64) that has health coverage. This is a quantitative variable that should always be between 0 and 100. It’s an interesting variable to explore because of how things may have changed due to the implementation of the ACA. Some reasonable research questions could include:

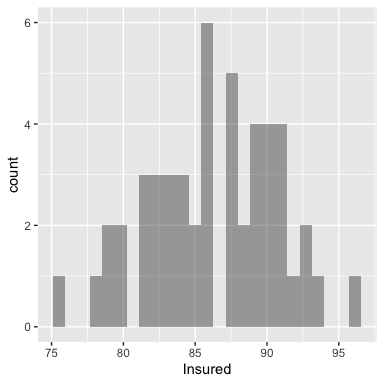
* What is the average state percentage of insured adults?
* Is the average state percentage of insured adults less than 90%?

The first question could be answered with a confidence interval, while the second should be written as a hypothesis test involving , the population mean percentage of insured adults, measured at the state level.

**TASK 1.2–Exploratory Data Analysis**

Because this is a quantitative variable, relevant plots could include histogram, dotplot, or boxplot. Here, I choose to create a histogram. I also calculate various summary statistics.

gf\_histogram( ~ Insured, data = state\_data)



favstats( ~ Insured, data = state\_data)

## min Q1 median Q3 max mean sd n missing  
## 75.2 83.425 86.2 89.675 95.8 86.244 4.478801 50 0

Based on my EDA I learned the statewide percentage of insured adults is roughly symmetric and bell-shaped, varying from 75.2 to 95.8. The mean is 86.244. Based on my EDA it seems that there will be some evidence against the null hypothesis that the population mean is 90, but I am not convinced that the evidence will be very strong. I will need to calculate the p-value. It doesn’t appear that there will be any issues with the necessary conditions for performing a one-sample t-test: the sample size is large (50) and the distribution is roughly symmetric.

**TASK 1.3–Inference and Conclusions**

t.test( ~ Insured, mu = 90, alternative = "less", data = state\_data)

##   
## One Sample t-test  
##   
## data: Insured  
## t = -5.9299, df = 49, p-value = 1.5e-07  
## alternative hypothesis: true mean is less than 90  
## 95 percent confidence interval:  
## -Inf 87.30592  
## sample estimates:  
## mean of x   
## 86.244

t.test( ~ Insured, mu = 90, alternative = "two.sided", data = state\_data)

##   
## One Sample t-test  
##   
## data: Insured  
## t = -5.9299, df = 49, p-value = 3e-07  
## alternative hypothesis: true mean is not equal to 90  
## 95 percent confidence interval:  
## 84.97114 87.51686  
## sample estimates:  
## mean of x   
## 86.244

After performing a one-sided t-test for a single mean, I see that the standardized test statistic is -5.93 and the p-value is very small. Thus we can reject the null hypothesis. There is very strong evidence that the average statewide percentage of insured adults is less than 90%. Further, based on the confidence interval I am 95% certain that the average state proportion of insured adults is between 84.97 and 87.52.

# Part 2: Two binary variables

Choose two binary variables from the merged states data.

**TASK 2.1–Introduction of the variables:**

The two variables of interest are *TwoParents* and *Elect2016*. Both are binaary variables. *TwoParents* indicates whether a state has at least 70% of families with two parents living in the household, while *Elect2016* specifies whether a state voted for the republican or the democratic presidential candidate in 2016.

A reasonable research question would be: Is there a relationship between voting for the Republican presidential candidate in 2016 and whether or not a state is has more than 70% of households with two parents?

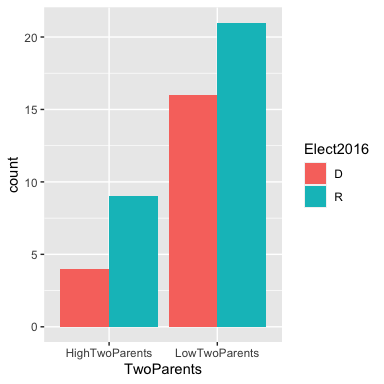
The corresponding hypotheses to test would then be:

Where is the proportion of states that voted for the Republican candidate, when considering only states with at least 70% of households having two-parents (so group 1 is states with at least 70% of households having two-parents); and is the proportion of states that voted for the Republican candidate, when considering only states with less than 70% of households having two-parents (so group 2 is states with less than 70% of households having two-parents).

**TASK 2.2–Exploratory Data Analysis**

A relevant visual summary would be a side-by-side or a stacked bar chart, and relevant numerical summaries include the difference in sample proportions.

gf\_bar( ~ TwoParents, fill = ~ Elect2016, data = state\_data, position = position\_dodge())



tally( Elect2016 ~ TwoParents, data = state\_data)

## TwoParents  
## Elect2016 HighTwoParents LowTwoParents  
## D 4 16  
## R 9 21

Based on my EDA, I see that states with more two-parent households are about twice as likely to vote for the Republican candidate in the 2016 Presidential Election. Specifically, the proportion of high two-parent states that voted Republican is 9/13 = 0.692. For low two-parent states the proportion that voted republican is less, 21/37 = 0.568. The difference between the sample proportions is thus 0.692 - 0.568 = 0.124. This may be a large enough difference to be statistically significant in a hypothesis test, but I will not know until I calculate the p-value.

The sample size is a bit small, however, as there are only 4 states that are low two-parent and voted Democratic. In an ideal world, we would a technique more apt for a small sample (such as a Fisher’s exact test), but those tests aren’t always covered in intro stats, so I will continue – *very cautiously* – with a standard test for the difference in proportions.

**TASK 2.3–Inference and Conclusions**

prop.test(Elect2016 ~ TwoParents, alternative="two.sided", data = state\_data)

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: tally(Elect2016 ~ TwoParents)  
## X-squared = 0.21223, df = 1, p-value = 0.645  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.4740834 0.2246032  
## sample estimates:  
## prop 1 prop 2   
## 0.3076923 0.4324324

The two-sided test for a difference in proportions yields a p-value of 0.645, indicating no evidence against the null and in favor of the alternative. We don’t have enough evidence to claim that there is a relationship between states having more than 70% of households with two parents and how they voted in the 2016 Presidential Election.

I am 95% confident that the difference in proportion of states voting Republican is between -0.474 and 0.225 when comparing high two-parent states to low two-parent states. Notice that this interval *does* include zero, which corresponds to our failure to find statistical significance in the corresponding hypothesis test. However, I should take my results with a grain of salt since my one of my four sample counts was low.

Both of the variables we used for this research question were categorical (binary) versions of quantitative variables. It would be interesting to look at an analysis of the numerical versions of these variables.

# Part 3: One binary and one quantitative variable

Choose one binary variable and one quantitative variable from the merged states data.

**TASK 3.1–Introduction of the variables:**

I choose to examine the quantitative variable *Highschool* with the categorical variable *Elect2016*. *HighSchool* gives the percentage of state residents aged 25-34 that are high school graduates, while *Elect2016* specifies whether a state voted for the republican or the democratic presidential candidate in 2016.

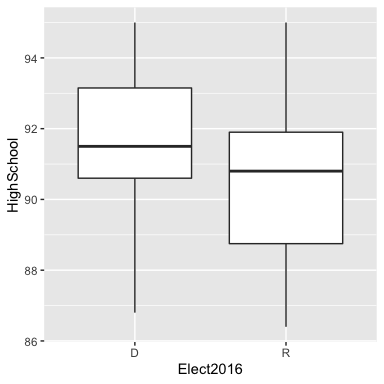
Some reasonable research questions would be: ‘Is the statewide percentage of high school graduates different, on average, when comparing states that voted Republican in 2016 versus states that voted Democratic?’ ‘How different?’ The former can be answered with a hypothesis test, and the latter with a confidence interval. The relevant hypotheses would be:

where and are the mean statewide percentage of high school graduates for states that Trump won and that Clinton won, respectively.

**TASK 3.2–Exploratory Data Analysis**

The best option here is side-by-side boxplots, but other plots that could work are side-by-side dotplots or side-by-side histograms. It is also important to calculate the summary stats for both groups, and the difference in those stats (the difference in means is of particular interest).

gf\_boxplot(HighSchool ~ Elect2016, data = state\_data)



favstats(HighSchool ~ Elect2016, data = state\_data )

## Elect2016 min Q1 median Q3 max mean sd n missing  
## 1 D 86.8 90.60 91.5 93.15 95 91.40500 2.376967 20 0  
## 2 R 86.4 88.75 90.8 91.90 95 90.33333 2.349370 30 0

diff(mean(HighSchool ~ Elect2016, data = state\_data ))

## R   
## -1.071667

Based on the difference of means from our data, states that voted republican have a slightly lower percentage of high school graduates compared to states that voted Democratic. Both groups have distributions that were only slightly skewed. I doubt that we will find a statistically significant difference between the two groups, based on this EDA. Further, it appears that all conditions are met to perform a two-sample t-test: both sample distributions are ‘well-behaved’ and have ‘large enough’ sample sizes.

**TASK 3.3–Inference and Conclusions**

t.test(HighSchool ~ Elect2016, alternative = "two.sided",data = state\_data)

##   
## Welch Two Sample t-test  
##   
## data: HighSchool by Elect2016  
## t = 1.5691, df = 40.541, p-value = 0.1244  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.3081457 2.4514790  
## sample estimates:  
## mean in group D mean in group R   
## 91.40500 90.33333

After performing a two-sided t test for the difference in means, I see that the standardized test statistic is 1.5691 and the p-value is 0.1244 This indicates that If there is truly no difference in statewide percentage of high school graduates when comparing states that Trump won to states that Clinton won, then the chance of seeing a sample difference of 1.07 or any value more extreme (in either direction) is 0.1244.

We cannot reject the null hypothesis. There is no evidence of a difference in statewide percentage of high school graduates when comparing states that Trump won to states that Clinton won. Correspondingly, the 95% confidence interval includes the null value of 0. I am 95% confident that the true different in statewide percentage of high school graduates is between -0.308 and 2.45 when comparing states that Trump won versus states that Clinton won.