Lab 8: Statistical Testing with Multiple Groups

Javier E. Flores

April 19, 2019

Introduction

Over the past few lectures, we've learned of methods which allow for testing comparisons across multiple groups. The first method, a χ^2 test for association, allows us to test for associations between two categorical variables - neither of which are required to be binary. Most recently, we've learned about ANOVA, which allows us to test for an association between a categorical and quantitative variable. Using ANOVA, we are able to determine whether a given categorical variable might be used to explain the variability observed in a quantitative outcome variable. The focus of this lab will be to apply these methods to real data.

As you work through the lab with your group, you will be asked to answer several questions. Please submit your responses (as a group) in a single, separate document. Include the original questions as well as your group's response in the final submission. All generated figures must also be included.

Data

The dataset we will explore in today's lab contains information on mass shootings in the United States from the years 1966 to 2017. These data were compiled by this kaggle user from several sources across the web. Certain records have been removed for the sake of providing a clean dataset. Additionally, certain variable categories have been modified for the sake of simplicity. The available variables include:

- **Location**: Location of the shooting
- **Date**: Date of the shooting
- **Area**: Exact area of the shooting
- **Open_Closed_Area**: Whether the incident took place in an open or closed area
- **Target**: Target of the shooting
- **Cause**: Cause for the shooting
- **Fatalities**: Number of fatalities
- **Injured**: Number injured
- **Total Victims**: Number of fatalities + the number injured
- **Policemen Killed**: Number of policemen kill
- **Age**: Age of shooter
- **Mental Health Issues**: Whether shooter had mental health issues
- **Race**: Race of shooter
- **Gender**: Gender of shooter

Analysis

Q1) Three quantitative variables of interest in this dataset are the total fatalities, total injured, and total victims. We will focus on total victims, which is the sum of the fatalities and injured. Construct a plot illustrating the distribution of this variable. Comment on whether you observe evidence of skew or outliers.

Q2) The vast majority of shootings recorded in this dataset have fatalities amounting to 10 individuals or less. One recorded shooting had a fatality count of more than three times this amount (32). Should this extreme observation be excluded when analyzing these data? Why or why not?

Q3) Apply a log-transform to the total victim count. Using the log-transformed outcome, generate the same type of plot chosen in **Q1**. Compare the log-transformed plot to the plot from **Q1**. Which plot is more 'normal'?

Q4) In order to visually assess the relationship between the shooter's mental health and the total number of victims, construct a boxplot comparing the distribution of the total number of victims using the mental health of the shooter as the grouping variable. Construct a similar plot using the log transformed count of total victims. Does there appear to be a relationship between shooter mental health status and the total number of victims?

Q5) Construct a 95% confidence interval to estimate the difference in total victim count between shooters with and without mental health issues. Provide both relative and absolute assessments along with a one-sentence interpretation of each.

Comparison	**95% Confidence Interval**	**Interpretation**
Absolute		
Relative		

Q6) In addition to the mental health of a shooter, we may also be interested in determining other factors which may be associated with the total number of victims. The table on the next page lists a few of these factors. For each, state the appropriate statistical approach needed to test for the association, the corresponding null hypothesis, resulting p-value, and final conclusion. Be sure to use a log-transform of your outcome if necessary to meet the assumptions of a given test.

Variable	Statistical Approach	Null Hypothesis	P-value and Conclusion
Mental Health	Two-Sample t-test	$\mu_{MHI} = \mu_{NoMHI}$	p = 0.001. We reject the null hypothesis. The average number of total victims is greater with shooters suffering from mental health issues.
Race			
Age			
Cause			
Target			

Q7) In the previous question, several hypothesis tests were performed. How many were statistically significant at the $\alpha = 0.05$ level? How many are statistically significant after applying a Bonferroni correction? Is it reasonable to apply a Bonferroni correction here? Why or why not?

Q8) In **Q6**, we observed a relationship between race and the total number of victims. Perform pairwise comparisons using Tukey's HSD to determine where there are differences across race in the total number of victims. Interpret the results of all pairwise comparisons.

Q9) One step towards the goal of ending mass shootings would be to try to understand the motivating cause behind each event and how that might relate to other factors of an individual. What type of variable (specifically) is "Cause"? Should we be concered with outliers or skew when analyzing this variable as an outcome? Can we apply a log-transform to this variable?

Q10) To determine which factors are associated with "Cause", fill out the table below.

Variable	Statistical Approach	Null Hypothesis	P-value and Conclusion
Mental Health			
Race			
Age			