Lab 8: Statistical Testing with Multiple Groups
KEY

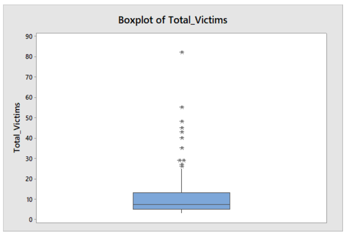
Javier E. Flores

April 19, 2019

Total Possible Points: 48

Analysis

Q1) [3 pts] 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.

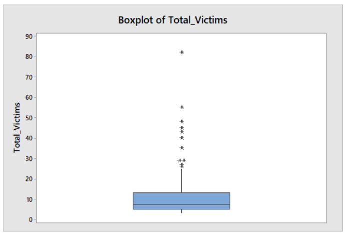


The boxplot for total victim count indicates extreme right skewness. Additionally, it appears that there may be some outliers (i.e. the data points marked with asterisks).

Q2) [3 pts] 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?

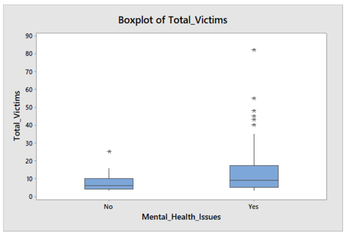
No, this outlier is a real datapoint corresponding to the 2007 Virginia Tech shooting in 2007.

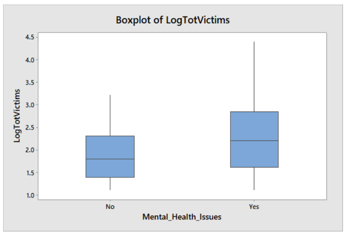
Q3) [3 pts] Apply a log-transform to the total victim count. Using the log-transformed outcome, generate the same type of plot chosen in **Q1**. Compare the log-transformed plot to the plot from **Q1**. Which plot is more 'normal'?



The boxplot for the natural log of total victim count indicates slight right skewness, but is a massive improvement towards normality relative to the untransformed count.

Q4) [3 pts] 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?





Looking at both plots, it appears that shooters with mental health issues are responsible for greater victim totals (as evident by matching the quartiles in each group). This relationship is more readily seen when looking at the log-transformed count as opposed to the untransformed count.

Q5) [4 pts] 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	95% CI for difference ($\mu_{NoMHI} - \mu_{MHI}$): (-10.11, -2.98)	We are 95% confident that the average number of victims for shooters without mental health issues is between 10 and 3 fewer people than those with mental health issues.
Relative	95% CI for difference ($\mu_{NoMHI} - \mu_{MHI}$): (-0.654, -0.169), After exponentiating: (0.520, 0.845)	We are 95% confident that the average number of victims for shooters without mental health issues is between 48% and 16% less than those with mental health issues.

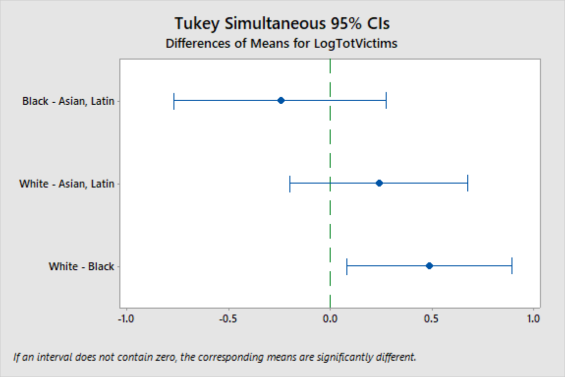
Q6) [12 pts] In addition to the mental health of a shooter, we may also be interested in determining other factors which may be associated with the total number of victims. The table on the next page lists a few of these factors. For each, state the appropriate statistical approach needed to test for 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	ANOVA	All means are the same	p-value = 0.016. We conclude that at least one mean differs. It appears that most white mass shooters are responsible for the greatest amount of victims.
Age	Correlation	$\rho = 0$	p-value = 0.465. We fail to reject the null hypothesis and conclude that we do not have sufficient evidence to suggest a relationship between the age of the shooter and the number of victims.
Cause	ANOVA	All means are the same	p-value = 0.000. We conclude that at least one mean differs. It appears that mass shooters with causes rooted in terrorism are responsible for greater amounts of victims.
Target	ANOVA	All means are the same	p-value = 0.147. We fail to reject the null hypothesis of no difference in means. There seems to be substantially more within group variability than between-group variability.

Q7) [4 pts] 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?

Including the two-sample t-test result, there were three results which were statistically significant at the $\alpha = 0.05$ level. Applying the Bonferroni correction, we obtain $\alpha^* = 0.05/5 = 0.01$. Using this as a threshold, only two results remain statistically significant. It is appropriate the apply the Bonferroni correction here since we tested associations across multiple variables with the same outcome, the log of total victim count. Without applying the correction, we have an increased risk of making a type I error.

Q8) [4 pts] 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.



If an interval does not contain zero, the corresponding means are significantly different.

Based on the above figure, the only significant difference in means is between white and black shooters. White shooters, on average, have higher total victim counts than black shooters.

Q9) [3 pts] 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?

"Cause" is a nominal categorical variable. Since this is a categorical variable, we are not concerned with skewness or outliers and cannot apply a log transformation. None of these ideas (outliers, skewness, transformations) make sense for categorical variables (e.g. how do you take the log of a category?).

Q10) [9 pts] To determine which factors are associated with ”Cause”, fill out the table below.

Variable	**Statistical Approach**	**Null Hypothesis**	**P-value and Conclusion**
Mental Health	χ^2 Test for Association	No association	p-value = 0.306. We fail to reject the null hypothesis independence (i.e. no association). There is insufficient evidence to suggest a relationship between mental health status and the proclivity for a certain cause.
Race	χ^2 Test for Association	No association	p-value = 0.536. We fail to reject the null hypothesis independence (i.e. no association). There is insufficient evidence to suggest a relationship between race and the proclivity for a certain cause. However, the reliability of this test is questionable given the low expected cell counts (i.e. ¡5). A randomization test (the better option in this case) yields a p-value of 0.565, offering no difference in conclusion.
Age	ANOVA	All means are the same	p-value = 0.078. We fail to reject the null hypothesis of no difference in means. There is, however, marginal evidence in support of a difference in means. It appears that shooters motivated by terrorism are, on average, younger than shooters aligned with other causes.