Is There a Correlation Between Manner of Collision and Severity of a Crash?

Tomas Barriga, Chris Day, and Gillian Martin

12/2/2021

Question

Is there a correlation between manner of collision and severity of a crash? For instance, are head-on crashes more likely to be severe crashes than a sideswipe crash?

Study Design

Crash data was collected from UDOT for the years 2014 to 2021 and provides information about the conditions in which the crash occurred. These conditions include roadway, weather, lighting, pavement, junction, work zone, horizontal and vertical curves, manner of collision, and first harmful event, all of which are recorded with various numerical codes. It also provides information on number of vehicles involved and the severity of the crash which is ranked on a scale from 1 to 5. 1 being property damage only, and 5 being a fatal crash. We will primarily be looking at the correlation between manner of collision and severity of a crash. We will use R to analyze this data statistically.

Statistical Methods

For our project, we used three statistical methods to analyze the data: boxplots, ANOVA, and the Tukey-Kramer method. The information each of those methods contain is further discussed below. It is important to note that the actual implementation of these methods is shown in the "Numerical Results" section. This section simply gives background information and context about each method.

Boxplot Method

A box plot, also called a box-and-whisker plot, is defined in our textbook as "a graphical display that represents the middle 50% of a group of measurements by a box and highlights various features of the upper and lower 25% by other symbols" where the middle 50% refers to the interquartile range (Ramsey/Schafer, 2013, p.18).

The graph gives an uncluttered view of the center, the spread, and the skewness of a distribution and indicates the presence of unusually small or large values, known as outliers. The spread can be seen by the tails of the data which are represented by whiskers on the box plot. The width of the box can also indicate the spread of the data. Skewness of data can be detected by a line in the box which indicates the median of the data. A data point is considered an outlier when it is more than 1.5 IQRs away from the box. An outlier is represented by a dot on a boxplot. Examples of how boxplots can visually represent different distributions of data are shown in Figures 1 and 2.

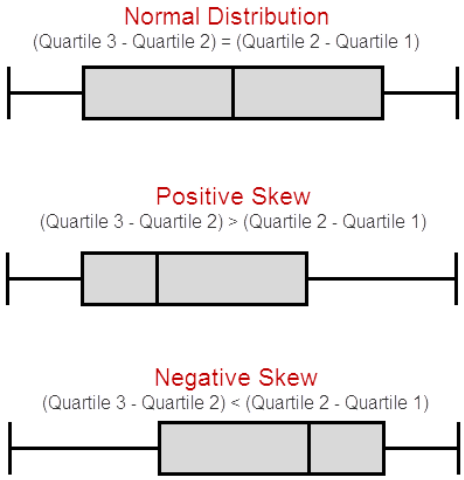


Figure 1: Boxplots visualizing skewed data.

In Figure 1, all of the spreads are pretty similar (the width of the boxes are about the same and the whiskers are all at the same values of 0 and 400). Also, no outliers are shown. Those elements of box plots are better shown in Figure 2.

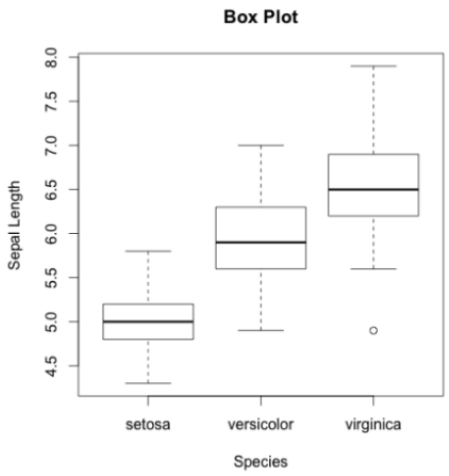


Figure 2: Boxplots visualizing different spreads and outliers within data.

We can see from Figure 2 that virginica flowers have an outlier. Furthermore, setosa has the smallest spread out of all three kinds of flowers.

Both figures also demonstrate how boxplots are useful when comparing multiple sets of data, which is why we used it when analyzing our data. We wanted an initial visualization of how the medians and spreads of crash severity varied depending on the manner of crash.

ANOVA (Analysis of Variance) Method

Another method used in our analysis is the one-way ANOVA (Analysis of Variance) method. The one-way ANOVA method is used to determine whether there are any statistically significant difference between the means of three or more independent groups. In our case, we are interested in seeing if the means of severity are different between nine independent manners of collision.

Now, how does a one-way ANOVA work? The first thing one should know is what the null and alternative hypotheses are in this method. The null hypothesis is that the difference between all of the means is 0, or that they are all equal. The model in which all means are assumed to be equal is referred to as the *reduced model* or the *equal-means model*. On the other hand, the alternative hypothesis is that there are at least two group means that are statistically different from each other. This model where the means are assumed to possible be different from each other is called the *full model* or *separate-means model*.

The next thing to know is what *residuals* are in an ANOVA test. A residual is the observation value minus its estimated mean, where the mean will be different depending on if one is referring to the null or alternative hypothesis. If the null hypothesis is incorrect, the magnitude of the residuals from the equal means model will tend to be larger. The *extra sum of squares F-statistic* is the single number that summarizes the differences in sizes of residuals from the full and reduced models.

The third piece of information to know is the *F-test*. The F-test is summarized by its corresponding p-value, the chances of finding an F-statistic as large as or larger than the observed one when all the means are indeed equal.In other words, the smaller the p-value is, the lower the chances are of finding the F-statistic that was observed. Therefore, if the *p-value* is less than 0.05 for 95% confidence, there is statistical evidence that one should reject the null hypothesis and at least two of the groups of means are different from one another.

An example of an ANOVA table is shown in Figure 3.

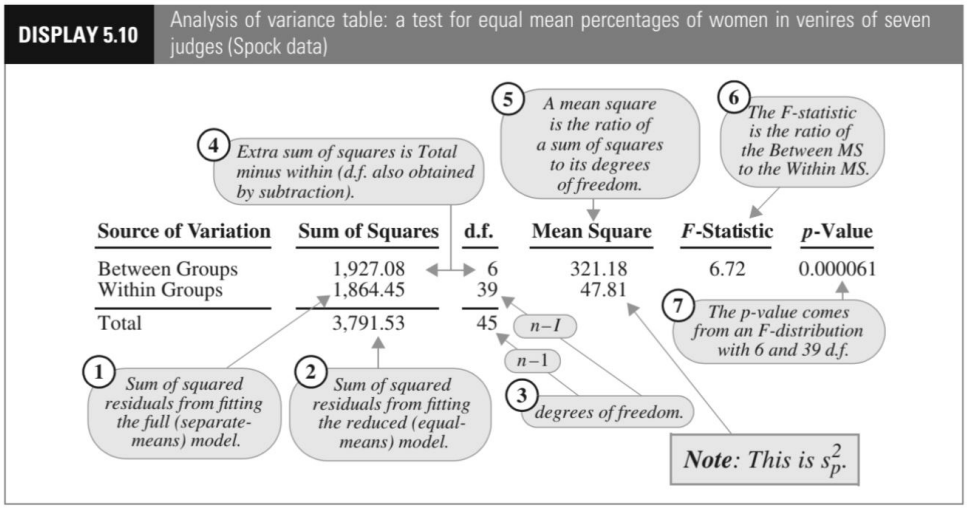


Figure 3: ANOVA table outputs and labels.

Using the one-way ANOVA method in our analysis will allow us to see if the mean severity of crash differs between any group of manner of collisions.

Tukey-Kramer Method

The Tukey-Kramer procedure, as the textbook states, utilizes "the unique structure of the multiple comparisons problem by selecting a multiplier from the studentized range distributions rather than from the t-distributions" (Ramsey/Schafer, 2013, p.161). In simpler terms, it compares the means of every treatment to the means of every other treatment and identifies any difference between two means that is greater than the expected standard error. This is useful for our question because it will tell us specifically which manners of collision have statistically different means of severity.

Logistic Regression

Logistic regression is a process of modeling the probability of a discrete outcome given input variables whereas linear regression which models the probability of continuous outcomes. Logistic regression also requires does not require a linear relationship between inputs and output variables due to the application of a nonlinear log transformation to the odds ratio. In the case of our research question, the response variable, severity of the crash, is a discrete variable where the severity is measured on a scale from 1 to 5. Therefore, logistic regression should be used in this case to see which variables have a significant impact on the severity of a crash, or which variables have a significant p-value in the logistic regression model.

Fisher's Exact Test

According to Wolfram MathWorld, "Fisher's Exact Test is a statistical test used to determine if there are non-random associations between two categorical variables." The p-value from Fisher's Exact Test corresponds to the proportion of values of the test statistic that are as extreme or more extreme than the observed value of that test statistic. The null hypothesis for Fisher's test is that the two classification variables are not different.

Chi-Squared Test

The chi-squared test determines whether there is a statistically significant difference between the expected frequencies and the observed frequencies of a table. It also tells whether two variables are independent of one another. It is similar to the Fisher's test, except it requires a larger sample size, specifically at least 5 observations or more. The chi-squared test also assumes that the observations are independent, which in our case, they are since each crash is an independent event of another.The null hypothesis and p-values are interpreted very similarly to Fisher's test, and it also gives an X^2 value which is the measure of the difference between the observed and expected frequencies.

Issues/ Limitations

There are pros and cons to each type of statistical method we are using in our analysis. The box-plot is helpful because it quickly gives the researcher an initial glance at the distribution of the data and get an idea for which groups may be significantly different from one another. They also can indicate if a transformation may be needed if the spreads of data between groups are hugely different from each other. The major issue with boxplots is that it doesn't prove anything statistically. Even if the medians or distributions look different from each other, there is no p-value to tell whether or not the means are significantly different from each other or not. Along with that, box plots do a better job at showing the medians than the means, which is not always what the researcher is interested in.

These weaknesses with boxplots is what lead my group to decide to do an ANOVA test. The severity index only ranges from 1 to 5, so even if the median severity index for each manner of collision was different, it wasn't by much. Therefore, we wanted to see if any of these small visual differences from the boxplots were statistically significant. However, the biggest issue with using one-way ANOVA is that it does not reveal which means are different from each other, only that at least two means differ from each other.

This is why our group also did a post-hoc test, in this case Tukey-Kramer. The pros of Tukey-Kramer is it highlights eactly where the differences in means do and do not occur. Some of the issues that come with Tukey-Kramer is that the confidence interval is wider than other methods such as Scheffe, and it is less powerful when testing small numbers of means. However, we still think it provides an adequate analysis of our data and we are not as concerned with the confidence interval. Furthermore, all of the assumptions of Tukey-Kramer were met: normality, homogeneity of variance, and independent observations.

Table 1: Comparions of Means and Standard Deviations

collisionType	mean	sd
1 Angle	1.592196	0.8472743
2 FrontToRear	1.458208	0.6983760
3 HeadOn	1.943038	1.0434443
4 SideSwipeSame	1.256635	0.6223845
5 SideSwipeOpp	1.535861	0.8558139
6 ParkedVeh	1.423800	0.7630393
7 RearToSide	1.188235	0.4740389
8 RearToRear	1.184615	0.4640955
96 SingleVeh	1.496187	0.8682057
97 Other	1.222222	0.5595814
99 Unknown	1.443558	0.7421246

The chi-squared test applies an approximation assuming that the sample is large whereas Fisher's exact test runs an exact procedure for small-sized samples. Therefore, in our analysis, after figuring out how large the sample size was in the Fisher's test, we also decided to perform a chi-squared test to confirm that the overall results were consistent and/or accurate.

One general limitation that we ran into during our analysis was when we appended the annual average daily travel dataset onto the crash data. The problem was that it gave over 30,000,000 observations and that vector was too large to work in within R Studio. Therefore, for our logistic regression analysis, Fisher's exact test, and chi-squared test, we filtered the data set to be within only Provo and Orem counties only (county ID = 49). This can be seen in the appendix of this report where we include the scrips that we wrote to import and clean the crash data we used for our analyses.

Numerical Results

We conducted an array of statistical tests to help us determine if the severity of a crash can be directly associated with the manner of the collision that occurred during a crash. One cannot simply assume that certain types of collisions will automatically result in certain severity. As stated in the Statistical Method section, the tests we conducted include a comparison of means and standard deviations, a boxplot, an ANOVA test, and a Tukey-Kramer analysis. The numerical results of each of these tests are displayed in the following subsections.

Comparison of Means and Standard Deviations

The first result that we computed is the means and the standard deviation of the crash severity of every manner of collision. The results this computation is displayed in Table 1. Although a simple table, it shows the numerical results of the first analysis that was conducted.

An interesting result found in the table is the mean of 3 HeadOn. The mean of this collision type is 1.92, which is much higher than the rest of the collision types. Unfortunately, the standard deviation of this result is also the highest. 7 RearToSide, 8 RearToRear, 4 SideSwipeSame, and 6 ParkedVeh all have values between 1.1 and 1.3, which forms the lowest severity value means.

Boxplot Analysis

The second numerical results is a box plot displaying the mean, 25th percentile, 75th percentile, and outliers of each of the manner of collisions in respect to crash severity. This result shows a more detailed version of Table 1 except it is displayed in a more visual manner. This time, it is clearly seen that HeadOn collisions, for example, have a higher mean crash severity than all other crash types. The boxplot is seen in Figure 4.

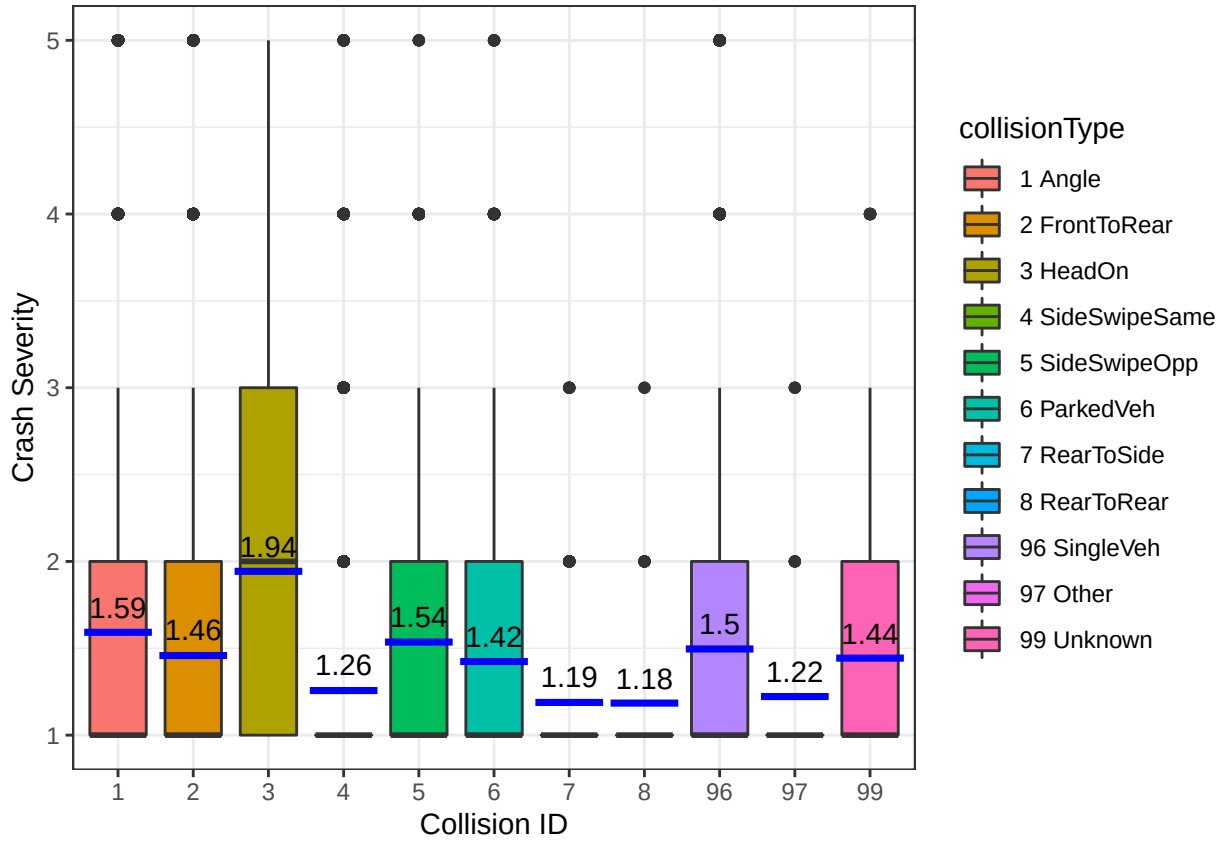


Figure 4: A boxplot analysis of Crash Severity by Collision ID.

It is important to note that Crash Severity only exists as an integer between 1 and 5. For this reason, no values exists in between integer values. Although, the mean and standard devations do include decimal values.

ANOVA Table

The third numerical analysis is the ANOVA analysis. This was conducted as a way to statistically determine if a difference in means existed.

```
anovaModel <- aov(crash_severity_id ~ collisionType, data = severity)
summary(anovaModel)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## collisionType     10   1552  155.15   263.7 <2e-16 ***
## Residuals     110368  64943    0.59
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The numerical results of the ANOVA Table includes 10 degrees of freedom, an F-statistic of 2363, and a p-value of less than 2e-16. This means that at least one of the manner of collisions mean values differs from at least one other manner of collision mean value.

Tukey-Kramer

The last numerical analysis is the results of the Tukey-Kramer test. This table displays the result of every combination of manner of collision, and if their means differ or not. The results of this test can be seen below.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = crash_severity_id ~ collisionType, data = severity)
##
## $collisionType
##                                       diff          lwr          upr     p adj
## 2 FrontToRear-1 Angle          -0.13398789 -0.152409596 -0.115566190 0.0000000
## 3 HeadOn-1 Angle                0.35084197  0.298084637  0.403599310 0.0000000
## 4 SideSwipeSame-1 Angle        -0.33556097 -0.362046834 -0.309075102 0.0000000
## 5 SideSwipeOpp-1 Angle         -0.05633450 -0.128699442  0.016030441 0.3009380
## 6 ParkedVeh-1 Angle            -0.16839562 -0.218906063 -0.117885171 0.0000000
## 7 RearToSide-1 Angle           -0.40396071 -0.593879752 -0.214041662 0.0000000
## 8 RearToRear-1 Angle           -0.40758062 -0.714164545 -0.100996688 0.0009537
## 96 SingleVeh-1 Angle           -0.09600902 -0.121340455 -0.070677589 0.0000000
## 97 Other-1 Angle               -0.36997378 -0.738314381 -0.001633177 0.0478635
## 99 Unknown-1 Angle             -0.14863842 -0.233267333 -0.064009503 0.0000009
## 3 HeadOn-2 FrontToRear          0.48482987  0.432866338  0.536793394 0.0000000
## 4 SideSwipeSame-2 FrontToRear  -0.20157308 -0.226440178 -0.176705973 0.0000000
## 5 SideSwipeOpp-2 FrontToRear    0.07765339  0.005865117  0.149441667 0.0213830
## 6 ParkedVeh-2 FrontToRear      -0.03440772 -0.084088474  0.015273025 0.4837584
## 7 RearToSide-2 FrontToRear     -0.26997281 -0.459672881 -0.080272747 0.0002427
## 8 RearToRear-2 FrontToRear     -0.27359272 -0.580041050  0.032855603 0.1314273
## 96 SingleVeh-2 FrontToRear      0.03797887  0.014345141  0.061612601 0.0000125
## 97 Other-2 FrontToRear         -0.23598589 -0.604213629  0.132241857 0.6042436
## 99 Unknown-2 FrontToRear       -0.01465053 -0.098786873  0.069485822 0.9999756
## 4 SideSwipeSame-3 HeadOn       -0.68640294 -0.741741447 -0.631064436 0.0000000
## 5 SideSwipeOpp-3 HeadOn        -0.40717647 -0.494340308 -0.320012641 0.0000000
## 6 ParkedVeh-3 HeadOn           -0.51923759 -0.589324353 -0.449150829 0.0000000
## 7 RearToSide-3 HeadOn          -0.75480268 -0.950838628 -0.558766733 0.0000000
## 8 RearToRear-3 HeadOn          -0.75842259 -1.068832890 -0.448012290 0.0000000
## 96 SingleVeh-3 HeadOn          -0.44685100 -0.501646348 -0.392055643 0.0000000
## 97 Other-3 HeadOn              -0.72081575 -1.092347243 -0.349284262 0.0000000
## 99 Unknown-3 HeadOn            -0.49948039 -0.597065751 -0.401895033 0.0000000
## 5 SideSwipeOpp-4 SideSwipeSame  0.27922647  0.204958719  0.353494216 0.0000000
## 6 ParkedVeh-4 SideSwipeSame     0.16716535  0.113964611  0.220366091 0.0000000
## 7 RearToSide-4 SideSwipeSame   -0.06839974 -0.259051928  0.122252450 0.9870114
## 8 RearToRear-4 SideSwipeSame   -0.07201965 -0.379058276  0.235018979 0.9996242
## 96 SingleVeh-4 SideSwipeSame    0.23955195  0.209208987  0.269894905 0.0000000
## 97 Other-4 SideSwipeSame       -0.03441281 -0.403131962  0.334306341 0.9999999
```

```
## 99 Unknown-4 SideSwipeSame      0.18692255  0.100660931  0.273184169 0.0000000
## 6 ParkedVeh-5 SideSwipeOpp     -0.11206112 -0.197883622 -0.026238611 0.0013265
## 7 RearToSide-5 SideSwipeOpp    -0.34762621 -0.549822747 -0.145429666 0.0000017
## 8 RearToRear-5 SideSwipeOpp    -0.35124612 -0.665583359 -0.036908873 0.0143304
## 96 SingleVeh-5 SideSwipeOpp    -0.03967452 -0.113538442  0.034189400 0.8208955
## 97 Other-5 SideSwipeOpp        -0.31363928 -0.688457896  0.061179340 0.2026731
## 99 Unknown-5 SideSwipeOpp      -0.09230392 -0.201739975  0.017132139 0.1929974
## 7 RearToSide-6 ParkedVeh       -0.23556509 -0.431008333 -0.040121846 0.0049871
## 8 RearToRear-6 ParkedVeh       -0.23918500 -0.549221324  0.070851326 0.3143388
## 96 SingleVeh-6 ParkedVeh        0.07238660  0.019751064  0.125022127 0.0004947
## 97 Other-6 ParkedVeh           -0.20157816 -0.572797257  0.169640934 0.8105526
## 99 Unknown-6 ParkedVeh          0.01975720 -0.076631965  0.116146363 0.9998891
## 8 RearToRear-7 RearToSide      -0.00361991 -0.363676101  0.356436282 1.0000000
## 96 SingleVeh-7 RearToSide       0.30795169  0.117456442  0.498446928 0.0000106
## 97 Other-7 RearToSide           0.03398693 -0.379923153  0.447897009 1.0000000
## 99 Unknown-7 RearToSide         0.25532229  0.048419390  0.462225187 0.0034575
## 96 SingleVeh-8 RearToRear       0.31157159  0.004630396  0.618512793 0.0430467
## 97 Other-8 RearToRear           0.03760684 -0.441189587  0.516403263 1.0000000
## 99 Unknown-8 RearToRear         0.25894220 -0.058442853  0.576327250 0.2351103
## 97 Other-96 SingleVeh          -0.27396476 -0.642602781  0.094673267 0.3715717
## 99 Unknown-96 SingleVeh        -0.05262940 -0.138543583  0.033284789 0.6688827
## 99 Unknown-97 Other             0.22133536 -0.156042918  0.598713639 0.7256155
```

The numerical results of this test include the confidence intervals as well as the p-value for each combination of manner of collision. More specifically, all combinations that result in a p adj value of 0, or less than 0.05 are to be considered statistically significant. In other words, the null-hypothesis is rejected and there is substantial evidence that the severity level means differ between the two manner of collisions. An exampe of this is HeadOn collision types. The mean severity level of HeadOn collisions is significantly different than all other collision type mean severity levels. Combinations of values greater than 0.05 show that there is no difference between mean severity levels.

Clearly, multiple manner of collision combinations do in fact have sufficient evidence showing that crash severity differs between the two. In addition, many manner of collision combinations do not differ between each other, showing no evidence that crash severity differs between the two. Fortunately, a more inferential conclusions is interpreted from the numerical analysis in the following section.

Logistic Regression

By combining the crash data with AADT data we can perform a logistic regression and test whether different factor such a AADT, manner of collision, weather condtion, and light condition have an effect on the severity of a crash.

```
lr <- glm(crash_severity_id ~
            AADT +
            as.factor(manner_collision_id) +
            as.factor(weather_condition_id) +
            as.factor(light_condition_id),
          data = crash)
summary(lr)
```

```
##
## Call:
## glm(formula = crash_severity_id ~ AADT + as.factor(manner_collision_id) +
```

```
##     as.factor(weather_condition_id) + as.factor(light_condition_id),
##     data = crash)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0816  -0.4942  -0.4311   0.4594   3.9135
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1.585e+00  4.932e-03 321.460  < 2e-16 ***
## AADT                              7.190e-07  4.128e-08  17.417  < 2e-16 ***
## as.factor(manner_collision_id)2  -1.579e-01  5.871e-03 -26.890  < 2e-16 ***
## as.factor(manner_collision_id)3   3.540e-01  1.637e-02  21.623  < 2e-16 ***
## as.factor(manner_collision_id)4  -3.702e-01  8.507e-03 -43.514  < 2e-16 ***
## as.factor(manner_collision_id)5  -4.475e-02  2.244e-02  -1.994  0.04616 *
## as.factor(manner_collision_id)6  -1.624e-01  1.572e-02 -10.332  < 2e-16 ***
## as.factor(manner_collision_id)7  -3.932e-01  5.887e-02  -6.679 2.42e-11 ***
## as.factor(manner_collision_id)8  -3.915e-01  9.504e-02  -4.119 3.80e-05 ***
## as.factor(manner_collision_id)96 -1.120e-01  8.270e-03 -13.541  < 2e-16 ***
## as.factor(manner_collision_id)97 -3.701e-01  1.142e-01  -3.242  0.00119 **
## as.factor(manner_collision_id)99 -1.567e-01  2.624e-02  -5.972 2.35e-09 ***
## as.factor(weather_condition_id)2 -1.892e-03  6.239e-03  -0.303  0.76167
## as.factor(weather_condition_id)3 -3.229e-02  1.060e-02  -3.045  0.00232 **
## as.factor(weather_condition_id)4 -1.456e-01  9.972e-03 -14.598  < 2e-16 ***
## as.factor(weather_condition_id)5 -1.643e-01  3.869e-02  -4.246 2.17e-05 ***
## as.factor(weather_condition_id)6 -1.512e-01  9.089e-02  -1.664  0.09616 .
## as.factor(weather_condition_id)7  4.166e-02  4.920e-02   0.847  0.39716
## as.factor(weather_condition_id)8 -3.072e-01  1.006e-01  -3.055  0.00225 **
## as.factor(weather_condition_id)9  2.715e-02  1.447e-01   0.188  0.85113
## as.factor(weather_condition_id)89  1.210e-01  3.870e-01   0.313  0.75458
## as.factor(weather_condition_id)99 -5.684e-02  4.303e-02  -1.321  0.18647
## as.factor(light_condition_id)2    4.101e-03  6.770e-03   0.606  0.54466
## as.factor(light_condition_id)3    1.009e-03  9.364e-03   0.108  0.91415
## as.factor(light_condition_id)4    1.390e-02  2.804e-02   0.496  0.62019
## as.factor(light_condition_id)5   -3.170e-02  2.147e-02  -1.477  0.13974
## as.factor(light_condition_id)6    2.107e-02  1.637e-02   1.287  0.19812
## as.factor(light_condition_id)89  -5.276e-01  3.826e-01  -1.379  0.16793
## as.factor(light_condition_id)99   8.001e-03  5.677e-02   0.141  0.88792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.5855322)
##
##     Null deviance: 66494  on 110378  degrees of freedom
## Residual deviance: 64613  on 110350  degrees of freedom
## AIC: 254194
##
## Number of Fisher Scoring iterations: 2
```

From these results we can see that AADT, all manner of collisions, some weather conditions and no light conditions have an effect on the severity of crashes. The specific factor which influence severity can be seen from the results which show each of the factors that have a p-value below the significant level. Because light condition and not all weather conditions had an effect on severity a more in-depth analysis will be conducted regarding head on crashes and its effect on severity.

Fisher's Exact Test

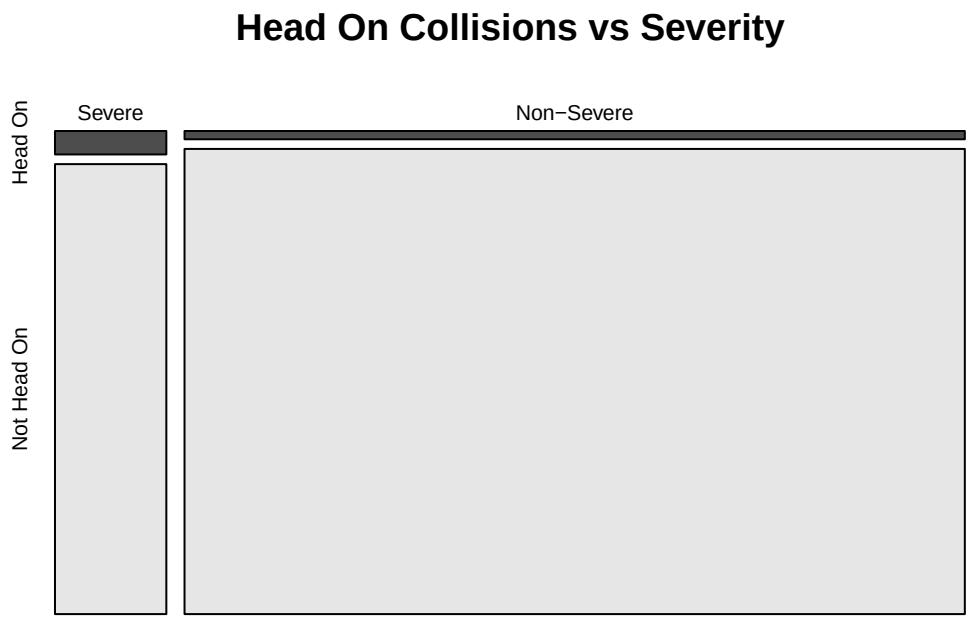
```
fishcrash <- crash %>%
  mutate(IsHeadOn = ifelse(manner_collision_id == 3,1,0),
         Severe = ifelse(crash_severity_id %in% c(3,4,5),1,0)) %>%
  group_by(IsHeadOn, Severe) %>% summarize(Sum = n())
fish <- data.frame(
  "Head On" = c(683,1687),
  "Not Head On" = c(13103, 94906),
  row.names = c("Severe", "Non-Severe"),
  stringsAsFactors = F
)
colnames(fish) <- c ("Head On", "Not Head On")
fish
```

```
##            Head On Not Head On
## Severe         683       13103
## Non-Severe    1687       94906
```

```
fisher.test(fish, alternative = "two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  fish
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  2.673983 3.212926
## sample estimates:
## odds ratio
##     2.9324
```

```
mosaicplot(fish, main = "Head On Collisions vs Severity", color=TRUE)
```



By summarizing the data we can see the total number of severe, non-severe, severe head on, and non-severe head on crashes. With the new data set up we can conduct Fisher's exact test to see the odds of a head-on manner of collision resulting in a severe crash versus any other manner of collision. A mosaic plot was also created to show if this inference might have any basis and from the plot we can see that it does appear there is a greater proportion of severe crashes resulting from head on collisions. The Fisher's exact test will show us if this hypothesis is valid. From these results we can see that a head-on crashes is 2.93 times more likely to be a severe crash than any other manner of collision with a 95% confidence interval of 2.67 to 3.21. Following these results another analysis will be conducted to further provide evidence to support the conclusion.

Chi-Squared Test

```
chisq <- chisq.test(fish, correct = F)
chisq$expected
```

```
##              Head On Not Head On
## Severe      296.0058    13489.99
## Non-Severe 2073.9942    94519.01
```

```
chisq
```

```
##
##  Pearson's Chi-squared test
##
```

```
## data:  fish
## X-squared = 590.85, df = 1, p-value < 2.2e-16
```

The Chi-squared test was conducted on the summarized data to see if the results from previous analysis were significant. From the p-value we can see that the results are significant. We can conclude that the results from Fisher's test were also significant and therefore odds of a head on collision being severe is 3 times that of any other manner of collision.

Inferential Conclusions

The inferential conclusions from our analysis can be interpreted through the numerical results that were presented. Specifically, the mean and standard deviation table, box plot, ANOVA table, Tukey-Kramer, Logisitic Regression, and Fisher's Test results provide sufficient evidence to formulate inferential conclusions.

Below, four specific inferential conclusions drawn from the data are stated and then explained. In addition, an overarching conclusion is drawn.

1. On average, some collision types produce more severe crashes.

The first conclusion that we draw is that on average, specific collision types produce more severe crashes than other collision types. By looking at Table 1 and the Figure 4, we see that Head On, Angle, and Single Vehicle collisions seem to have a higher crash severity mean than all the other collision types. Head On collisions in particular seem to have the highest average crash severity level. This means that in the data that was collected, crash intensity was on average highest amoung Head On data points.

This inference does not conclude that Head On collisions are always the most severe. By analyzing the Box Plot data, we see that all collision types except Rear To Side collisions can produce crash severity of level 5. It does conclude, however, that the average severity of Head On, Angle, and Single Vehicle collisions are higher than all other collision types.

This conclusion is backed up by the results in the ANOVA table and the Tukey-Kramer test. The ANOVA table proved that a difference in means existed. The Tukey-Kramer tests displayed overwhelming evidence that the mean of Head On collisions and Angle collisions is different than all other collision types. For the most part, Single Vehicle crashes also showed a difference in means from other collision types. The p-value of less than 0.05 provides this evidence.

2. Certain collision types can be more likely to produce a more severe crash.

Another inference that is made is that Head On, Angle, and Single Vehicle crashes will more likely produce a more severe crash than a collision involving a Parked Vehicle, Rear to Side, or Rear to Rear type. The mean severitys between all of these combinations is significanly different. This is shown in the p-values displayed in the Tukey-Kramer test. In addition, Head On, Angle, and Single Vehicle crashes have the highest average crash severity whereas Parked Vehicle, Rear to Side, and Rear to Rear collisions have the lowest average crash severity. This is shown in the Box Plots and in Table 1.

This inference is also backed up by analyzing the results of the Fisher's Exact Test. Looking at the odds ratio of 2.9324, with a 95% confidence interval from 2.67 to 3.21, there is a much higher odds of getting into a severe car crash if your collision was head-on than if it was not. This is backed up by analyzing the mosaic plot. Although there is a greater total number of severe crashes that are not Head on compared to those that are, percentage wise there are more severe crashes when it is a Head On collision.

3. On average, some collision types produce less severe crashes.

Although very similar to the first conclusional inference that was made, here we infer that on average, specific collision types produce less severe crashes than other collision types. Again, by looking at Table 1 and Figure 4, we see that Parked Vehicle, Rear to Rear, and Rear to Side collisions seem to have lower crash severity means than all the other collision types. Rear to Side collisions in particular seem to have the lowest average crash severity level. This means taht in the data that was collected, crash intensity was on average lowest amoung Rear to Side data points.

This inference does not conclude that Rear to Side collisions are always the least severe. It does conclude, however, that the average severity of Parked Vehicle, Rear to Rear, and Rear to Side collisions are lower than all other collision types. Similar to the first conclusion, this inference is backed up by the results in the ANOVA table and the Tukey-Kramer test.

4. AADT, manner of collision, and some weather conditions play a role in predicting crash severity.

In order to predict the crash severity, a logistic regression model was developed using AADT, manner of collision, weather conditions, and light conditions are predictors. The results of the logistic regression show that indeed AADT and manner of collision do play a significan role in predicting severity. Most of these values had p-values less than 2e-16. Only some weather conditions were significant predictors. The weather conditions of Rain, Snowing, Blowing Snow, and Severe Crosswinds all were significant predictors in predicting crash severity. This makes sense as all these conditions make driving significantly more difficult. Lastly, lighting conditions was determined to be an insignificant predictor, meaning it does not play a role in determining a crash severity prediction.

A connection exists between manner of collision and crash severity.

The overarching conclusion we make, however, is that *a connection exists between manner of collision and crash severity.* Drawing from specific conclusions, we see that on average, some collision types produce more sever crashes and some collision types produce less severe crashes. We also see that certain collision types are more likely to produce a more severe crash. We also see that AADT volumes and some weather conditions play a role in predicting crash severity. Overall though, it is clear to see that some sort of valid connection exists between collision type and crash severity. This conclusion is backed up specifically by the results of the ANOVA table. There is a difference in crash severity means, and no debate exists about that.

Additional Discussion

The application of this study would be helping UDOT achieve their goal of Zero Fatalities on Utah Roadways and creating safer roadways to prevent manner of collisions that may result in more severe or fatal crashes. If roadway data was added into this statistical analysis, you could potentially analyze specific segments or intersections and the crashes in them.

Alternative analysis methods may include looking at other roadway conditions that may effect severity of a crash just as roadway surface condition, weather conditions, lighting conditions. To do this, we could use two-way ANOVA or MANOVA tables to simultaneously analyze the effects of each of these conditions.

Limitations of the study include that a causal relationship cannot be introduced between the data because of the stochastic nature of crash data and its unpredictability. Another limitation includes human error in data collection because data is not self reported; it is recorded by first responders. The low number of fatal crashes compared to non fatal crashes also makes it difficult to properly analyze the data at hand.

Division of Tasks

Tommy- Data Collection, Problem Statement, Additional Discussion, Tukey Code, Overall Writing and Code Check

Gillian- Statistical Methods, Issues, ANOVA Code, Regression Code, Chi-Squared code

Chris- Numerical Results, Inferential Conclusions, BoxPlot Code, Fisher's Code

Appendix

For more information on the design of this project, please see this github page: https://github.com/gillianriches/Stats512Project.

Below are the scripts that correspond to the datamaker.R script

```
#build the crash dataset
crashprep <- function() {
  location <- read_csv("data/Crash_Data_14-20.csv")
  vehicle <- read_csv("data/Vehicle_Data_14-20.csv")
  rollups <- read_csv("data/Rollups_14-20.csv")
  aadt <- read_csv("data/AADT_Unrounded.csv")

  location <- check_location(location)
  vehicle <- check_vehicle(vehicle)
  rollups <- check_rollups(rollups)
  aadt <- check_aadt(aadt)

  crash <- left_join(location,rollups,by='crash_id')
  crash <- left_join(crash,vehicle,by='crash_id') %>%
    filter(county_id == 49)
  crash <- left_join(crash, aadt, by=c('route')) %>%
    filter(milepoint >= START_ACCU, milepoint < END_ACCUM) %>%
    mutate(AADT = case_when(
      grepl("2014",crash_datetime) == TRUE ~ AADT2014,
      grepl("2015",crash_datetime) == TRUE ~ AADT2015,
      grepl("2016",crash_datetime) == TRUE ~ AADT2016,
      grepl("2017",crash_datetime) == TRUE ~ AADT2017,
      grepl("2018",crash_datetime) == TRUE ~ AADT2018,
      grepl("2019",crash_datetime) == TRUE ~ AADT2019,
      grepl("2020",crash_datetime) == TRUE ~ AADT2020
    ))
  rm("location","vehicle","rollups", "aadt")

  crash
}
```

```
#build severity dataset
buildSeverity <- function(crashData) {
  crashData %>%
    select(3,6) %>%
    mutate(collisionType = case_when(
      manner_collision_id == 1 ~ "1 Angle",
```

```
      manner_collision_id == 2 ~ "2 FrontToRear",
      manner_collision_id == 3 ~ "3 HeadOn",
      manner_collision_id == 4 ~ "4 SideSwipeSame",
      manner_collision_id == 5 ~ "5 SideSwipeOpp",
      manner_collision_id == 6 ~ "6 ParkedVeh",
      manner_collision_id == 7 ~ "7 RearToSide",
      manner_collision_id == 8 ~ "8 RearToRear",
      manner_collision_id == 96 ~ "96 SingleVeh",
      manner_collision_id == 97 ~ "97 Other",
      manner_collision_id %in% c(99,89) ~ "99 Unknown"
    ))
}
```

```
buildSeveritySummary <- function(severityData) {
  severityData %>%
    group_by(collisionType) %>%
    summarize(
      mean = mean(crash_severity_id),
      sd = sd(crash_severity_id)
    )
}
```

```
fun_mean <- function(x){return(round(data.frame(y=mean(x),label=mean(x,na.rm=T)),digit=2))}

makeBoxPlot <- function(severity){
  ggplot(severity) +
    aes(x = as.factor(manner_collision_id), y = as.numeric(crash_severity_id)) +
    geom_boxplot(aes(fill = collisionType)) +
    stat_summary(fun.y = mean, geom="crossbar",colour="blue", size=.5) +
    stat_summary(fun.data = fun_mean, geom="text", vjust=-0.7) +
    labs(x = "Collision ID", y = "Crash Severity") +
    theme_bw()
}
```

Below are the functions relating to the crashprep.R script.

```
# The following is the functions that prepare the crash data to be joined

# Keeps the following columns from the location file
check_location <- function(df){
  df %>%
    select(crash_id = crash_id,
           crash_datetime = crash_datetime,
           crash_severity_id = crash_severity_id,
           light_condition_id = light_condition_id,
           weather_condition_id = weather_condition_id,
           manner_collision_id = manner_collision_id,
           roadway_surf_condition_id = roadway_surf_condition_id,
           route = route,
           milepoint = milepoint,
           county_id = county_id)
}
```

```r
# Keeps the following columns from the vehicle file
check_vehicle <- function(df){
  df %>%
    select(crash_id = crash_id,
           vehicle_num = vehicle_num,
           vehicle_year = vehicle_year,
           travel_direction_id = travel_direction_id,
           most_harmful_event_id = most_harmful_event_id,
           vehicle_maneuver_id = vehicle_maneuver_id)
}
```

```r
# Keeps the following columns from the rollups file
check_rollups <- function(df){
  df %>%
    select(crash_id = crash_id,
           number_fatalities = number_fatalities,
           number_four_injuries = number_four_injuries,
           number_three_injuries = number_three_injuries,
           number_two_injuries = number_two_injuries,
           number_one_injuries = number_one_injuries,
           pedalcycle_involved = pedalcycle_involved,
           pedestrian_involved = pedestrian_involved,
           motorcycle_involved = motorcycle_involved,
           distracted_driving = distracted_driving,
           speed_related = speed_related,
           adverse_weather = adverse_weather,
           adverse_roadway_surf_condition = adverse_roadway_surf_condition,
           roadway_geometry_related = roadway_geometry_related,
           roadway_departure = roadway_departure,
           overturn_rollover = overturn_rollover,
           route_type = route_type,
           night_dark_condition = night_dark_condition,
           single_vehicle = single_vehicle)
}
```

```r
check_aadt <- function(df){
  df %>%
    mutate(route = RT_NUM) %>%
    select(START_ACCU = START_ACCU,
           END_ACCUM = END_ACCUM,
           route,
           AADT2020 = AADT2020,
           AADT2019 = AADT2019,
           AADT2018 = AADT2018,
           AADT2017 = AADT2017,
           AADT2016 = AADT2016,
           AADT2015 = AADT2015,
           AADT2014 = AADT2014)
}
```