UNIVERSITY OF CONNECTICUT, DEPARTMENT OF STATISTICS

# Data analysis on the CT Crash Dataset

Jianmin Chen, Jun Jin

December 1, 2019

## CONTENTS

# 1 Abstract

In this paper, the group conducts the exploratory data analysis and statistical inferences on the CT crash data, judging that whether the variables such as type of intersections, presence of lighting, presence of a left-turn lane, presence of a right-turn lane and others could post great effects on the damage accounts and exploring that how they work. The research is useful for the government as a reference when designing the intersection in Connecticut.

# 2 Introduction and Description of Data

In this paper, the group will take advantage of "CT Crash Data", it is the crash data for rural two-lane intersections and segments in Connecticut are collected in order to study highway safety. Generally, crash severity counts may be aggregated into three categories: PDO (property damage only), or B+C (possible injury or nonincapacitating injury), or K+A (fatal or incapacitating injury) crashes. Three different data sets are available on the HuskyCT course website, corresponding to the type of intersections. Dataset 3ST Intersections CT.csv consists of n = 385 cases for 3ST (three-way stop controlled) intersections. Dataset 4ST Intersections CT.csv consists of n = 61 cases for 4ST (four-way stop controlled) intersections. Dataset 4SG Intersections CT.csv consists of n = 102 cases for 4SG (four-way signalized) intersections. Each dataset includes the following variables at each intersection: Intersection ID, Annual Average Daily Traffic (AADT) for major roads and minor roads, three dummy variables to indicate presence of lighting, presence of a left-turn lane, and presence of a right-turn lane, followed by skew angle, and then the counts of PDO, KA, and BC crashes.

## 2.1 Research proposing

In this paper, we are going to explore and confirm the following questions:

- Do these different types of intersection affect the number of crashes? (Regardless of other features.)
- If we control the types of intersection, whether other features (e.g. skewed angle, lighting, approaching turns, AADT major and minor) affect the number of crashes?
- And we may want to explore what kind of relationship (negative or positive) between the features we mentioned previously with the number of crashes.

## 2.2 Qualitative an Quantitative

Qualitative: Lighting, approach left, approach right.
Quantitative: AADT major, AADT minor, skew angle, PDO crashing number, KA crashing number, BC crashing number.

# 3 Exploratory Data Analysis

We first rearranged the 3 data set and combined them into 1 single data set. One column is added to specify the intersection type, denoted as INTERTYPE. It has 3 values, "3ST","4ST" and "4SG". Then we have 12 variables and 548 observations in total. "INTERSECTION-ID" and "ID" will not be analysed here since we are not interested in specific intersection. In the rest 10 variables, "AADT-MAJOR", "AADT-MINOR", "SKEW-ANGLE" are continuous quantitative variables. "PDO","KA" and "BC" are discrete quantitative variables. "LIGHTING","APPROACH-LEFTTURN","APPROACH-RIGHTTURN", and "INTERTYPE" can be used as categorical qual-

itative variables. We are interested in the relationship between "PDO", "KA", "BC" and other variables.
A new categorical binary variable is constructed with "1" standing for there exists crash and "0" standing for no crash. This variable is denoted as "crash" and it is used as a candidate response variable when we want to analyze the data with mixed types of crash.
In this part, we will do exploratory data analysis to first find some insight of the data.

### 3.1 Overall Distribution

First, let us see overall distribution of the 3 types of crashes and the distribution across intersection types.
For the PDO(property damage only) crashes, around 38% of the intersections have 0 PDO crash overall. The overall distribution of PDO is spread out with some extreme values larger than 15. The distribution across different intersections shows some variations. The distribution is similar to the overall distribution in the 3ST type, while in 4SG and 4ST types, there are fewer zero counts and the data has larger variance with more non-zero counts had hence heavier tails. These infer that 4-way intersections tend to have more PDO crashes, especially the 4-way signalized intersection.

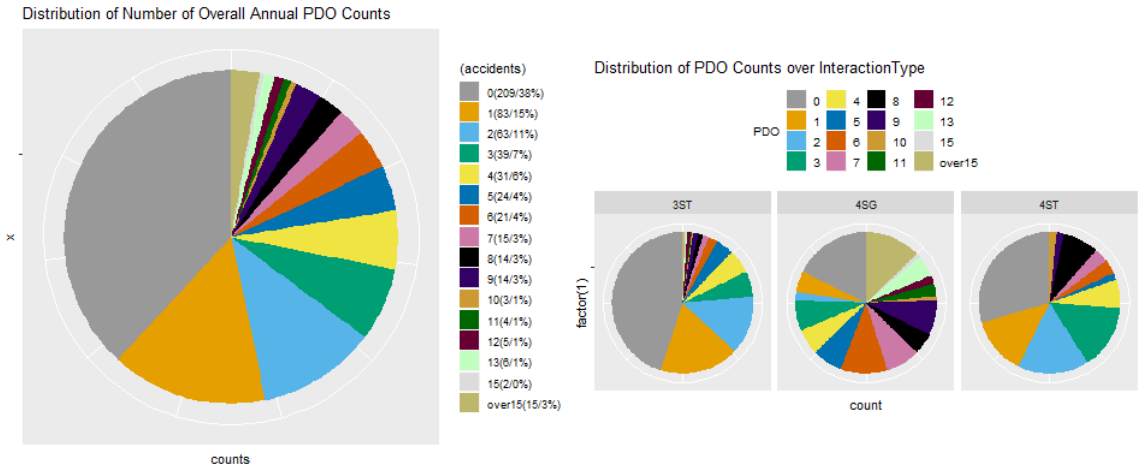


Figure 3.1: Pie chart of PDO

For the BC(possible injury or non-incapacitating injury) crashes, over one-half(59%) of the intersections have 0 crash and 76% of all intersections have less than 2 crashes. Like PDO, 4-way intersections tend to have larger number of BC crashes.

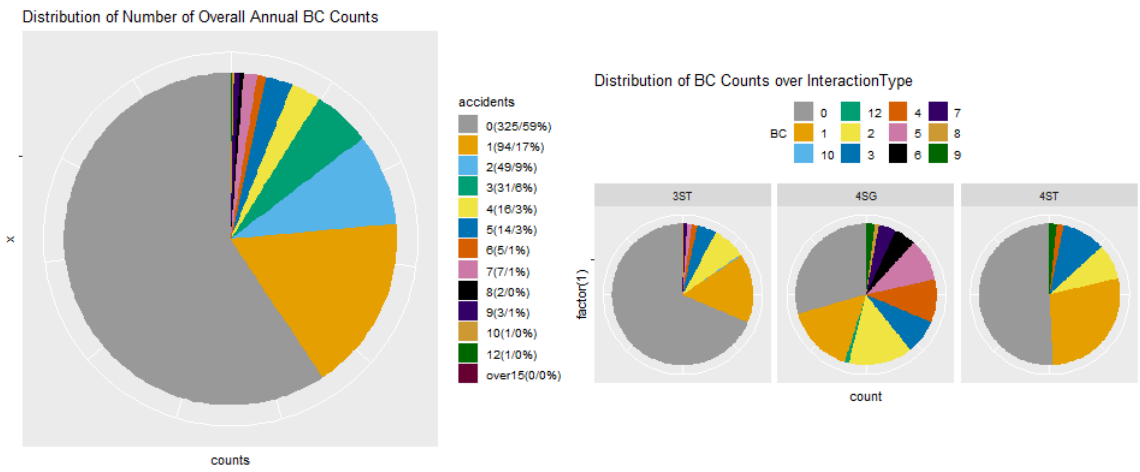


Figure 3.2: Pie chart of PDO

For the KA(fatal or incapacitating injury) crashes, only 0,1 and 2 crashes are observed. We can infer that Poisson or Negative Binomial distribution may not be suitable to fit the KA data. Quite similar to the conclusion above, 4SG and 4ST types have larger proportion of intersection with crashes.

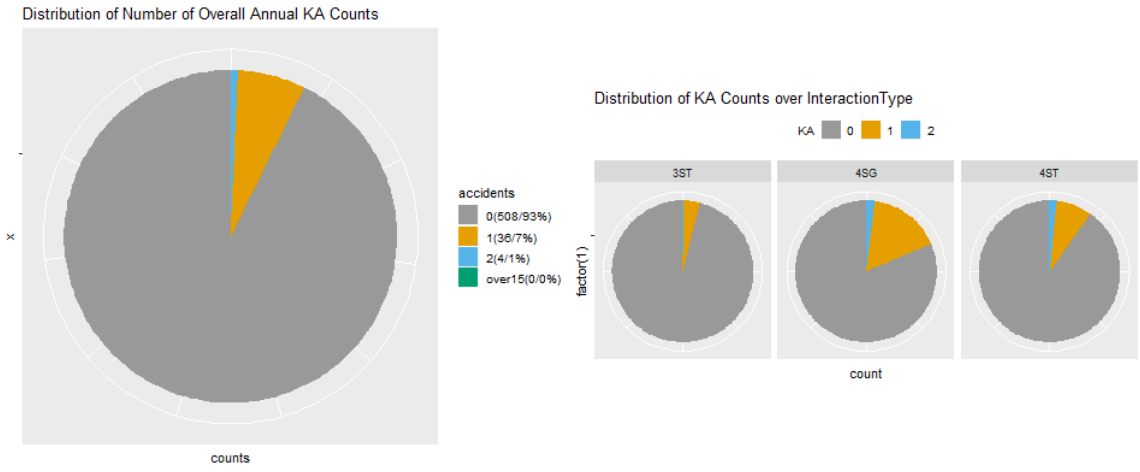


Figure 3.3: Pie chart of KA

Comparing the overall distribution of 3 types of crashes, we can see that for all 3 types, 0 crash has the maximum proportion in all 548 intersection. PDO and BC have more possible count values with larger dispersions, which is a hint that it is not plausible to mix PDO, KA and BC observations together when we want to model the distribution of crashes. Also, variance exists between types of intersections. Since intersection type is a more fixed factor that is hard to change in real traffic, in the following parts, we will analyze the data controlling intersection type to get some plausible suggestions to reduce crashes based on other variables. Since our aim is to find the vital factor of crashes, and KA crash can be considered as rare event, we will

also treat crash as binary distributed variables in some latter analysis, with 0 stands for no crash and 1 stands for at least 1 crash.

## 3.2 Crash vs Qualitative Variables

Next, we find some interesting relationship between crash and the 3 categorical variables, LIGHTING, APPROACH-LEFTTURN, and APPROACH-RIGHTTURNk.

### *3.2.1 Lighting*

First, let's see the relationship between LIGHTING and crashes. There are 197 intersections without lighting in all 548(around 36%) intersections. Overall, when not controlling intersection type, intersection without lighting tends to have fewer crashes than intersection with lighting in all 3 types of crashes. We can compare the proportion of 0 counts in the following bar chart of PDO. The bar chart of BC and KA has similar tendency which are contained in appendix.
When controlling intersection type, there are some interesting findings.
For PDO, in 3ST and 4SG types, partitioned data has same tendency with whole data while for 4ST group, there is lower proportion of observations with crash when there is lighting, which is opposite to the whole data. For BC and KA, partitioning with intersection type generate same result in each group as the whole set.
The plots infer that there might be a relationship between lighting and crash. We will not take 4ST PDO group out for special consideration as these group has a small number of observations which may cause larger randomness.

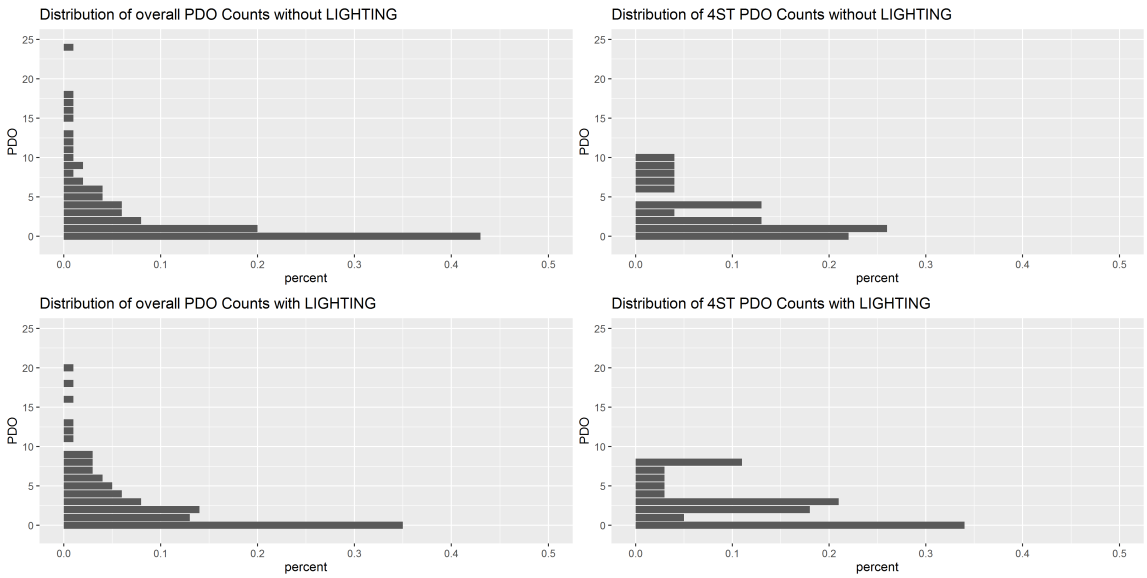


Figure 3.4: Bar Chart of Lighting

### *3.2.2 Approach Left Turn and Approach Right Turn*

There are only 65 intersections with approach left-turn. When partitioned with intersection type, we get the following contingency table. From the table we can see that the counts get even smaller when separated into 3 type, there is not much sense to analyze with data further separated by crash type. We discuss the relationship between Approach Left-Turn and crash based on the generated columns "crash".

There is similar problem with approach right-turn. And we also analyze based on crash as response variable. We observed similar association between these 2 variables and crash. Overall, when not controlling intersection type, intersection with approach left-turn or right turn has a higher crash rate, which infers that approach turn may has a positive relationship with car crashes. However, when we look into data grouped by intersection type, is it shown that left-turn or right turn does not have big effect with 3ST and 4SG intersections, even shown slightly negative relationship. The effect is only obvious in 4ST groups, where the number of observations with left or right turn are quite small. Hence, the result of overall data and 4ST data may not be reliable. We need to check the relationship with data only from 3ST and 4SG group and take intersection type into consideration.
Further, as same patterns are observed with right-turn and left-turn, we combine these 2 variables into a new binary variable, denoted as "TURN", where "1" stands for there exists an approaching turn and "0" else. The follow graph is made with combined column. (Separated result is in Appendix.)

| Observations | 3ST | 4ST | 4SG |
|---|---|---|---|
| Without Left-turn | 368 | 58 | 57 |
| With Left-turn | 17 | 44 | 4 |

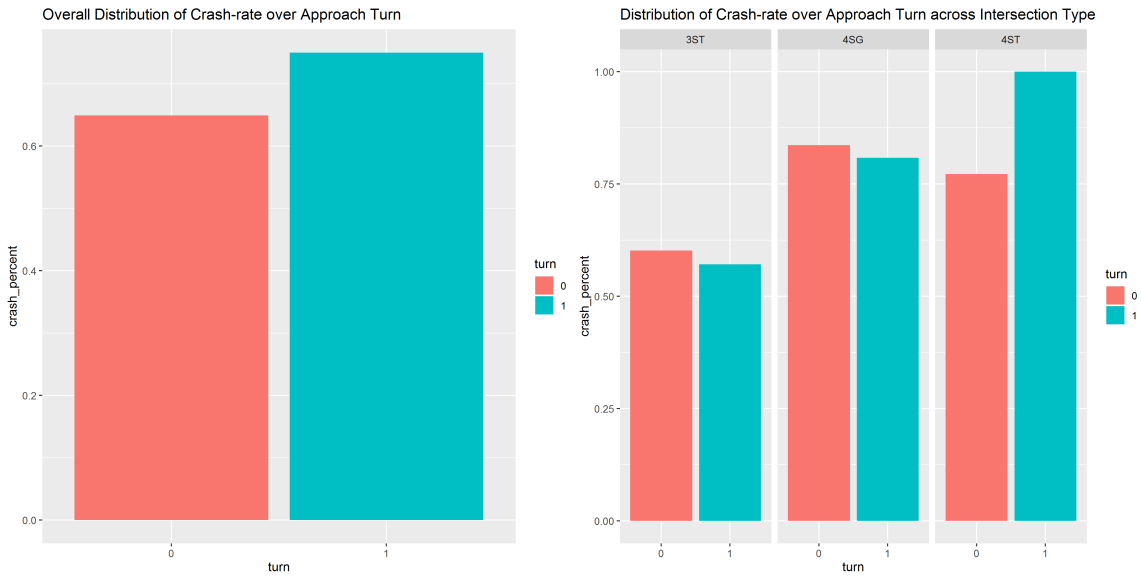


Figure 3.5: Bar Chart of Approach Left Turn

## 3.3 Crash vs Quantitative Variables

Also, there is likely to exist some association between AADT-MAJOR, AADT-MINOR, SKEW-ANGLE and crash.

### *3.3.1 AADT-MAJOR and AADT-MINOR*

Selected graphs are shown in this part.
We mainly apply the notched box-plot here to compare the distribution of AADT in intersection of crashes with intersection of no crash. The first 3 box plots below show significant difference on median of overall distribution as the notch does not overlap. This infers that generally, the intersections with crash has higher level of AADT-Major in all 3 types of crashes.

After controlling the intersection type, all partitioned data shows same pattern that higher AADT-Major appears in group with crash. However, for PDO and BC, the difference is only significant in 3ST group and for KA, no difference is significant.

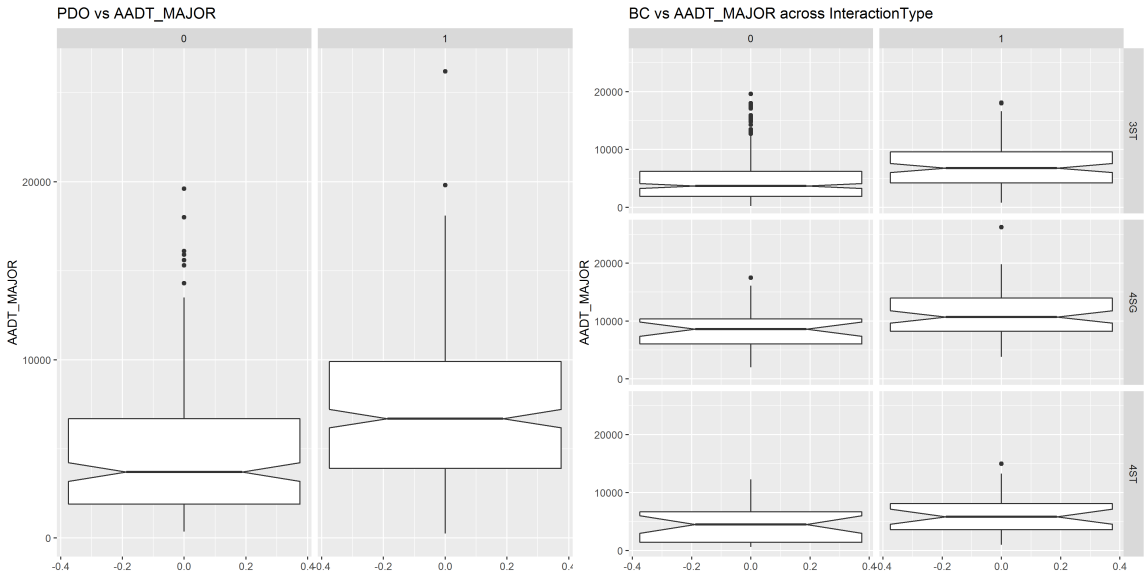


Figure 3.6: Boxplot of AADT-Major

For AADT-Minor, AADT-Minor is significantly larger in intersection with crash when exploring overall data for all of PDO, KA and BC. When taking intersection type into consideration, for 3ST group, same patters are observed as overall data and the difference are all significant. However, opposite pattern occurs in 4SG group that intersection with crash has a lower AADT-MINOR value. The difference of sample median between the with-crash and without-crash group is significant in PDO crash. For 4ST group, all difference in median is not significant.

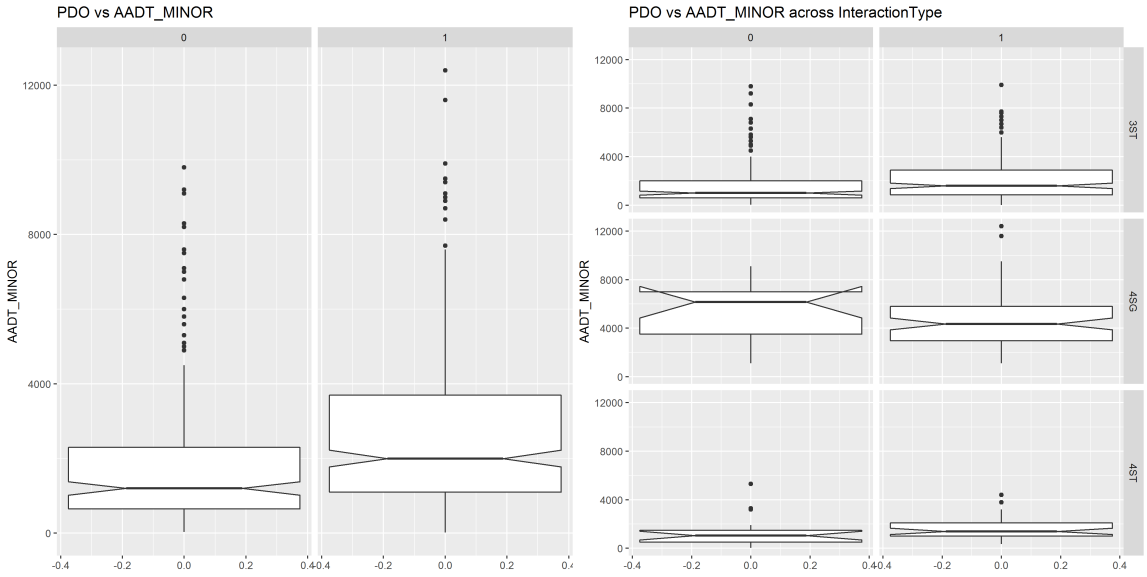


Figure 3.7: Boxplot of AADT-Minor

### *3.3.2 Skew Angle*

For Skew Angle, we apply notched box plot also. In all 3 types of crash, groups with crash have smaller median for skew angle while the difference is only significant in KA crash type. When controlling intersection type, only difference in 3ST intersection and PDO crash shows significant larger pattern in no crash group. All other are insignificant. 4SG group get insignificant opposite pattern in PDO and BC crash. Skew angle is a variable that is considered less important.

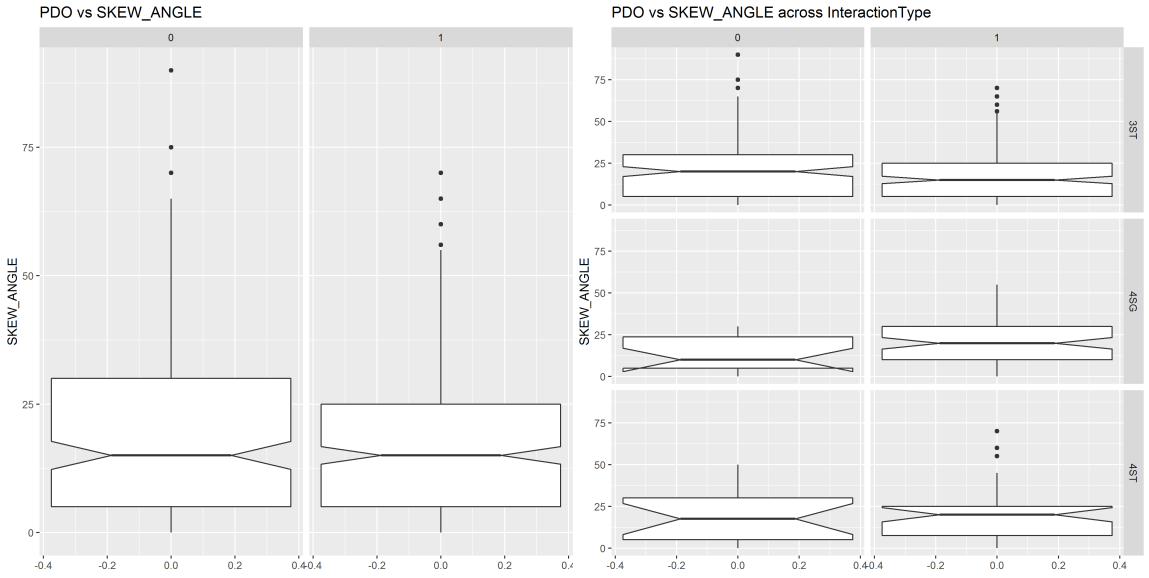


Figure 3.8: Boxplot of Skew Angle

## 3.4 EDA Conclusion

We get the following guess from the EDA part and the verification will be shown in the next inference part.
1. PDO, BC and KA has different distribution. They are not coming from different distribution family.
2. Intersection type is a vital factor to crash. Four-way intersections are at a higher probability for all 3 types of crash.
3. Intersection without lighting has a lower rate for all 3 types of crash. Controlling intersection type, same results are yielded except the PDO, 4ST group which has a lower crash rate when there is lighting. These results are kind of opposite to common sense.
4. Approach (left or right) turn has negative effect on overall crash rate, which infers that when there is an approach turn, the crash rate goes down. The result is only for 3ST and 4ST group individually. Data for 4SG group will not be analyzed with too few observations.
5. AADT-Major and AADT-Minor is at a higher level in intersections with crash for whole data set. When controlling intersection, However, for AADT-Major, the tendency is only significant in 3ST group with PDO and BC. For AADT-Minor, opposite significant pattern occurs in 4SG group of PDO crash.
6. Skew angle may not act as a vital factor to crash.

# 4 Statistical Inference

## 4.1 One v.s. PDO crashing independence test

Firstly, we can see obviously that all PDO crashing number for three kinds of intersections are not normal data, so our test is based on Mann-Whitney test.
For PDO crashing between 3st intersections and 4st intersections,

$$\begin{aligned} &H_0 : q_A = q_B \\ &H_1 : q_A > q_B \quad or \quad q_B > q_A \end{aligned}$$

where $q_A$ represents the median PDO crashing for 3st intersections, $q_B$ represents the median PDO crashing for 4st intersections. Then, the statistic can be constructed by:

$$\frac{W - E(W)}{\sqrt{Var(W)}}$$

and

$$\begin{aligned} &W = \sum_{j=1}^{n} R_j \\ &E(W) = \tfrac{n(n+m+1)}{2} \\ &Var(W) = \tfrac{mn}{12}\left\{(n+m+1) - \tfrac{1}{(m+n)(m+n-1)} \sum_{j=1}^{g} t_j\left(t_j^2 - 1\right)\right\} \end{aligned} \tag{4.1}$$

where $n$ is PDO crashing sample size for 4st intersection; $m$ is PDO crashing sample size for 3st intersection; $R_j$ is the $j-$th sample rank for PDO crashing rank of 4st intersection samples; $g$ is the tie group numbers and $t_j$ is the tie group size for $j-$th tie group. According to the code in Appendix 2, our conclusion is $q_B > q_A$, i.e. the 4st intersection is more likely to happen PDO crash than 3st intersection.
For PDO crashing between 4st intersections and 4sg intersections,

$$\begin{aligned} &H_0 : q_A = q_B \\ &H_1 : q_A > q_B \quad or \quad q_B > q_A \end{aligned}$$

where $q_A$ represents the median PDO crashing for 4st intersections, $q_B$ represents the median PDO crashing for 4sg intersections. Then, the statistic can be constructed by:

$$\frac{W - E(W)}{\sqrt{Var(W)}}$$

and

$$\begin{aligned} &W = \sum_{j=1}^{n} R_j \\ &E(W) = \tfrac{n(n+m+1)}{2} \\ &Var(W) = \tfrac{mn}{12}\left\{(n+m+1) - \tfrac{1}{(m+n)(m+n-1)} \sum_{j=1}^{g} t_j\left(t_j^2 - 1\right)\right\} \end{aligned}$$

where $n$ is PDO crashing sample size for 4sg intersection; $m$ is PDO crashing sample size for 4st intersection; $R_j$ is the $j-$th sample rank for PDO crashing rank of 4sg intersection samples; $g$ is the tie group numbers and $t_j$ is the tie group size for $j-$th tie group. According to the code in Appendix 2, our conclusion is $q_B > q_A$, i.e. the 4sg intersection is more likely to happen PDO crash than 4st intersection.
For PDO crashing between 3st intersections and 4sg intersections,

$$H_0 : q_A = q_B$$
$$H_1 : q_A > q_B \quad or \quad q_B > q_A$$

where $q_A$ represents the median PDO crashing for 3st intersections, $q_B$ represents the median PDO crashing for 4sg intersections. Then, the statistic can be constructed by:

$$\frac{W - E(W)}{\sqrt{Var(W)}}$$

and

$$W = \sum_{j=1}^{n} R_j$$
$$E(W) = \frac{n(n+m+1)}{2}$$
$$Var(W) = \frac{mn}{12}\left\{(n+m+1) - \frac{1}{(m+n)(m+n-1)}\sum_{j=1}^{g} t_j\left(t_j^2 - 1\right)\right\}$$

where $n$ is PDO crashing sample size for 4sg intersection; $m$ is PDO crashing sample size for 3st intersection; $R_j$ is the $j-$th sample rank for PDO crashing rank of 4sg intersection samples; $g$ is the tie group numbers and $t_j$ is the tie group size for $j-$th tie group. According to the code in Appendix 2, our conclusion is $q_B > q_A$, i.e. the 4sg intersection is more likely to happen PDO crash than 3st intersection.
According to the conclusions above, we have:

$$q_A < q_B < q_C$$

where $q_A$ represents the median number of PDO crashes of 3st intersection, $q_B$ represents the median number of PDO crashes of 4st intersection, $q_C$ represents the median number of PDO crashes of 4sg intersection.
However, we can use another method - Pearson test in categorical analysis - to analyze this dependence. We fistly, transform the PDO crashing number into dummy variable. If the crashing number equals to 0, then we denote it as "noncrash", while if the crashing number is greater than 0, then we denote it as "crash". Thus, we can form the following table:

| | 3ST | 4ST | 4SG | Total |
|---|---|---|---|---|
| Nocrash | 173 | 18 | 18 | 209 |
| Crash | 212 | 43 | 84 | 339 |
| Total | 385 | 61 | 102 | 548 |

Then our hypothesis will be

$$H_0 : p_{ij} = p_i p_j$$
$$H_1 : p_{ij} \neq p_i p_j$$

and our statistic is:

$$Q_P = \frac{\sum_{i=1}^{2}\sum_{j=1}^{3}\left(n_{ij} - n_{i\cdot}n_{\cdot j}/n\right)^2}{n_{i\cdot}n_{\cdot j}/n}$$

and it approximately obeys a $\chi^2(1)$ distribution. Thus, according to the code in Appendix 2, our conclusion is: they are dependence.

## 4.2 One v.s. BC and KA crashing independence test

We next conduct the independence test for BC crashes and KA crashes similar with the way we conduct for PDO test, so the testing procedures are omitted (codes are in Appendix 2). The conclusion is

- $q_A < q_B < q_C$, where $q_A$ represents the median number of BC crashes of 3st intersection, $q_B$ represents the median number of BC crashes of 4st intersection, $q_C$ represents the median number of BC crashes of 4sg intersection.
- $q_A < q_B = q_C$, where $q_A$ represents the median number of KA crashes of 3st intersection, $q_B$ represents the median number of KA crashes of 4st intersection, $q_C$ represents the median number of KA crashes of 4sg intersection.

and two categorical tables are:

Table 4.1: Categorical table for BC crash

| | 3ST | 4ST | 4SG | Total |
|---|---|---|---|---|
| Nocrash | 264 | 31 | 30 | 325 |
| Crash | 121 | 30 | 72 | 223 |
| Total | 385 | 61 | 102 | 548 |

Table 4.2: Categorical table for KA crash

| | 3ST | 4ST | 4SG | Total |
|---|---|---|---|---|
| Nocrash | 370 | 55 | 83 | 508 |
| Crash | 15 | 6 | 19 | 50 |
| Total | 385 | 61 | 102 | 548 |

And finally both of these two Pearson tests show dependences.

## 4.3 Two v.s. crashing independence test

### *4.3.1 Crossing type + skewed angle v.s. PDO crashing number*

As we can see that the skewed angle holds little choices of values, so we decide to transform skewed angle into categorical data and also transform PDO crashing number as our previous method, forming the categorical table below. We settle a standard: let $s$ denote the skewed angle, if $0 \leq s \leq 30$, then we record it as "1"; If $31 \leq s \leq 60$, then we record it as "2"; If $61 \leq s \leq 90$, then we record it as "3".

Table 4.3: Categorical table for PDO crash

| | 3ST_1 | 3ST_2 | 3ST_3 | 4ST_1 | 4ST_2 | 4ST_3 | 4SG_1 | 4SG_2 | Total |
|---|---|---|---|---|---|---|---|---|---|
| Nocrash | 131 | 31 | 11 | 14 | 4 | 0 | 18 | 0 | 209 |
| Crash | 168 | 39 | 5 | 36 | 6 | 1 | 71 | 13 | 339 |
| Total | 299 | 70 | 16 | 50 | 10 | 1 | 89 | 13 | 548 |

Then, the Pearson test shows that there exists a dependence (codes are in Appendix 2.)

### 4.3.2 Crossing type + lighting for PDO crash

As the lighting itself is a dummy variable, we can also form the following categorical table:

Table 4.4: Categorical table for PDO crash

| | 3ST_0 | 3ST_1 | 4ST_0 | 4ST_1 | 4SG_0 | 4SG_1 | Total |
|---|---|---|---|---|---|---|---|
| Nocrash | 74 | 99 | 5 | 13 | 6 | 12 | 209 |
| Crash | 76 | 136 | 18 | 25 | 28 | 66 | 339 |
| Total | 150 | 235 | 23 | 38 | 24 | 78 | 548 |

Then, the Pearson test shows that there exists a dependence (codes are in Appendix 2.)

### 4.3.3 Crossing type + turns for PDO crash

As the lighting itself is a dummy variable, we can also form the following categorical table:

Table 4.5: Categorical table for PDO crash

| | 3ST_00 | 3ST_01 | 3ST_10 | 3ST_11 | 4ST_00 | 4ST_10 | 4ST_11 | 4SG_00 | 4SG_01 | 4SG_10 | 4SG_11 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Nocrash | 164 | 2 | 3 | 4 | 18 | 0 | 0 | 9 | 1 | 4 | 4 | 209 |
| Crash | 200 | 2 | 7 | 3 | 39 | 3 | 1 | 46 | 2 | 18 | 18 | 339 |
| Total | 364 | 4 | 10 | 7 | 57 | 3 | 1 | 55 | 3 | 22 | 22 | 548 |

where “00” represents no left approach or right approach; “01” represents no left approach but a right approach; “10” represents no right approach but left approach; “11” represents a left approach and right approach. Then, the Pearson test shows that there exists a dependence (codes are in Appendix 2.)

### 4.3.4 Crossing type + AADT major for PDO crash

We firstly partition the full samples into 6 parts: 3st intersection + crash; 3st intersection + noncrash; 4st intersection + crash; 4st intersection + noncrash; 4sg intersection + crash; 4sg intersection + noncrash. And we carry out a Shapiro normality test on these 6 sample pieces, however, few of them pass the test. Hence, once again, we need to turn to the Mann-Whitney test.
Comparing the AADT major and controlling the intersection type in 3st, we have the following hypotheses:

$$H_0 : q_A = q_B$$
$$H_1 : q_A < q_B \quad or \quad q_A > q_B$$

where $q_A$ represents the median of AADT major in PDO noncrash situation in 3st intersection, while $q_B$ represents the median of AADT major with PDO crash situation in 3st intersection. And the statistic is similar with (4.1). And our final conclusion (codes are in Appendix 2) is:

$$q_A < q_B.$$

Comparing the AADT major and controlling the intersection type in 4st, we have the following hypotheses:

$$H_0 : q_A = q_B$$
$$H_1 : q_A < q_B \quad or \quad q_A > q_B$$

where $q_A$ represents the median of AADT major in PDO noncrash situation in 4st intersection, while $q_B$ represents the median of AADT major with PDO crash situation in 4st intersection. And the statistic is similar with (4.1). And our final conclusion (codes are in Appendix 2) is:

$$q_A = q_B.$$

Comparing the AADT major and controlling the intersection type in 4sg, we have the following hypotheses:

$$H_0 : q_A = q_B$$
$$H_1 : q_A < q_B \quad or \quad q_A > q_B$$

where $q_A$ represents the median of AADT major in PDO noncrash situation in 4sg intersection , while $q_B$ represents the median of AADT major with PDO crash situation in 4sg intersection. And the statistic is similar with (4.1). And our final conclusion (codes are in Appendix 2) is:

$$q_A = q_B.$$

*4.3.5 Crossing type + AADT minor for PDO crash*

With the same partition made for AADT major,we also carry out a Shapiro normality test on these 6 sample pieces for AADT minor, however, few of them pass the test. Hence, once again, we need to turn to the Mann-Whitney test.
Comparing the AADT minor and controlling the intersection type in 3st, we have the following hypotheses:

$$H_0 : q_A = q_B$$
$$H_1 : q_A < q_B \quad or \quad q_A > q_B$$

where $q_A$ represents the median of AADT minor in PDO noncrash situation in 3st intersection, while $q_B$ represents the median of AADT minor with PDO crash situation in 3st intersection. And the statistic is similar with (4.1). And our final conclusion (codes are in Appendix 2) is:

$$q_A < q_B.$$

Comparing the AADT minor and controlling the intersection type in 4st, we have the following hypotheses:

$$H_0 : q_A = q_B$$
$$H_1 : q_A < q_B \quad or \quad q_A > q_B$$

where $q_A$ represents the median of AADT minor in PDO noncrash situation in 4st intersection, while $q_B$ represents the median of AADT minor with PDO crash situation in 4st intersection. And the statistic is similar with (4.1). And our final conclusion (codes are in Appendix 2) is:

$$q_A = q_B.$$

Comparing the AADT minor and controlling the intersection type in 4sg, we have the following hypotheses:

$$H_0 : q_A = q_B$$
$$H_1 : q_A < q_B \quad or \quad q_A > q_B$$

where $q_A$ represents the median of AADT minor in PDO noncrash situation in 4sg intersection , while $q_B$ represents the median of AADT minor with PDO crash situation in 4sg intersection. And the statistic is similar with (4.1). And our final conclusion (codes are in Appendix 2) is:

$$q_A = q_B.$$

## 5 Discussion of Results and Concluding Remarks

## 6 Appendix

## 7 References