Synthetic Biology Course

**Python Statistics Project**

compiled by Emilis Gegevicius

Dataset chosen: US Traffic Accidents (2.25 million records).

Source: <https://www.kaggle.com/sobhanmoosavi/us-accidents>

Hypothesis: Severity of traffic accidents depends on weather conditions

Out of 49 columns present in the initial dataset, 7 variables have been chosen and converted to metric scale from US:

**Severity** – ordinal variable

**Humidity(%)** – continuous variable

**Temperature(oC)** – continuous variable

**Pressure(mbar)** – continuous variable

**Wind\_Speed(kmh)** – continuous variable

**Precipitation(mm)** – continuous variable

**Visibility(km)** – continuous variable

Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).

A screenshot of a social media post

Description automatically generated

According to histogram, frequencies of severity distributed as such:

|  |  |
| --- | --- |
| Severity | Frequency |
| 0 | 17 |
| 1 | 814 |
| 2 | 1455524 |
| 3 | 715582 |
| 4 | 72002 |

Due to very low number of events of Severity = 0 (17 events, corresponding to 0.0008%), it is dropped out of dataset.

Descriptive Statistics:

A screenshot of a cell phone

Description automatically generated

**Normality Analysis**

**Temperature (oC) variable**

A screenshot of a cell phone

Description automatically generated

Lowest temperature in dataset is -61.0 °C

Highest temperature in dataset is 77.0 °C

Highest recorded temperatures in US ranges in the interval of 45-50oC, lowest -55 - -45oC. More extreme values could have occurred due to specific temperature measurement (e.g. satellite measurements are not precise) or ground temperature readings, which are usually higher than air (no additional information is available on this topic). As these outliers are possible errors in measurement, it is removed from further analysis.

30 measurements have been removed

A screenshot of a social media post

Description automatically generated

From the first look in histogram, it is obvious that temperature variable data is not normally distributed and median is shifted towards higher values. The same thing is observed in each Severity groups. Data does not represent Gauss.

Lowest Temperature(C) in dataset is -40.0

Highest Temperature(C) in dataset is 50.0

A screenshot of a cell phone

Description automatically generated

This is proved further with Anderson-Darling test , which is statistical test of whether or not a dataset comes from a certain probability distribution, e.g., the normal distribution. Shapiro-Wilk test is not suitable as there are more than 500 measurements for the variable.

Anderson-Darling test (Critical Value = p value) proves, that sample does not look Gaussian

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated

There is a possibility, that these tests could fail due to large amount of data present in the dataset. Nevertheless, from visual perspective, it could be said, that data is more or less normally distributed and we can use parametric tests.

**Humidity(%) variable**

**A screenshot of a cell phone

Description automatically generated**

Lowest humidity in dataset is 4.0%

Highest humidity in dataset is 100.0%

Humidity variable values are not normally distributed with median shifted to higher values.

A screenshot of a social media post

Description automatically generated

The same thing is seen from humidity histograms between different Severity groups.

A picture containing text, newspaper

Description automatically generated

Anderson-Darling test indicates that data does not look normal.

**Pressure(mbar) variable**

Highest ever atmospheric pressure: 1083.3 mbar

Lowest ever atmospheric preesure: 870 mbar

In this variable there are 68 extreme measurements, which has occurred due to measurement errors, which is removed from this variable. There is only 7 pressure up to 950mbars, which was removed due to skewing data. 950-1055 mbar range has been chosen

A screenshot of a cell phone

Description automatically generated

Visually, data is not distributed normally.

Lowest Pressure(mbar) in dataset is 955.6388817333427

Highest Pressure(mbar) in dataset is 1054.860069666677

A screenshot of a cell phone

Description automatically generated

A picture containing text, newspaper

Description automatically generated

**Wind Speed (kmh) variable**

Highest ever recorded wind speed in US 372km/h. In dataset there are 17 temperature values higher than this number due to measurement error, which are removed.

Only 39 values higher than 150 kmh, which is omitted

A screenshot of a cell phone

Description automatically generated

Data is obviously not normally distributed

A screenshot of a social media post

Description automatically generated

**Precipitation (mm) variable**

The highest category of rainfall, termed “Violent shower” is when precipitation per hour is greater than 50mm. There are 497 measurements of precipitation above 80mm, which can be classified as super extreme weather and must be removed from the dataset

Lowest Precipitation(mm) in dataset is 0.0

Highest Precipitation(mm) in dataset is 74.67599999999999

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

**Visibility (km) variable**

When visibility is reported to be 16km visibility (10 miles), it usually means “unlimited” visibility, thus we take 85km as a higher upper limit (dropping 404 measurements).

A screenshot of a cell phone

Description automatically generated

Lowest Visibility(km) in dataset is 0.0

Highest Visibility(km) in dataset is 80.4672

A screenshot of a social media post

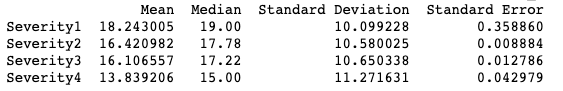
Description automatically generated

**Box plot analysis**

1. **Temperature**

A screenshot of a computer

Description automatically generated



From the first look, Severity 2 and 4 box plots are obviously skewed (median is shifted towards second half of the data) The spread of data in different severity groups are similar, nevertheless, wider whiskers and higher number of extreme outliers in 2 and 3 severity indicates, that deviations from temperature median causes more severe accidents. Severity 4 is achieved with lower temperatures (same thing can be seen in severity 1). Medians between all these measurements are different, which could indicate relationship between them.

1. **Humidity (%)**

**A screenshot of a computer

Description automatically generated**

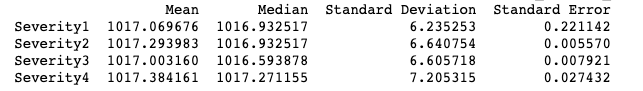
**A screenshot of a cell phone

Description automatically generated**

1. **Pressure(mbar)**

A screenshot of a computer

Description automatically generated



1. **Wind\_Speed(kmh)**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

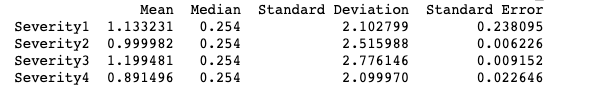
1. **Precipitation(mm)**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

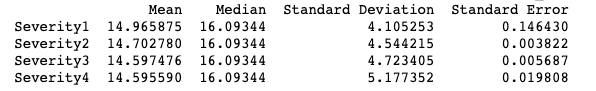
Description automatically generated**

****

1. **Visibility(km)**

A screenshot of a computer

Description automatically generated



**Significant differences analysis (ANOVA and Kruskal-Wallis)**

One of ANOVA requirements is data normality, which can only be assumed for first three variables – temperature, humidity and pressure.

In order to test, whether all these groups are statistically significantly different, several different tests can be made. ANOVA (and non parametric alternative Kruskal-Wallis test) is the most suitable test for this dataset as there are multiple groups to compare.  Multiple t-tests (difference between means of compared groups) is not recommended when more than 2 groups are compared (the more hypothesis tests are used the bigger the risk of making a type I error “rejection of true null hypothesis).

As it was seen from the dataset, there are outliers, which will affect ANOVA test values, thus, two different ANOVA tests with outliers and without outliers have been conducted.





The F ratio is the ratio of two mean square values. As it was predicted, ANOVA with outliers gave a large F ratio (value = 1353.8). Such large F ratio can be seen both when the null hypothesis is wrong (the data are not sampled from populations with the same mean) and when random sampling happened to end up with large values in some groups and small values in others. After removal of outliers, F value was significantly smaller (as predicted) and p value was <0.05, which indicates, that data are not sampled from populations with the same mean.

Kruskal-Wallis for non-parametric







For the sake of curiosity, Z-tests between different curiosity groups have been conducted and indicated, that there are significant differences between the means of compared groups as p values drop below 0.05. Z test is an alternative for t-test, when there are >30 values of specific variable.

Temperature:

A screenshot of a cell phone

Description automatically generated

HUMIDITY

A screenshot of a cell phone

Description automatically generated

PRESSURE

A screenshot of a cell phone

Description automatically generated

If the overall P value is small, then it is unlikely that the differences you observed are due to random sampling. You can reject the idea that all the populations have identical means.

MANN WHITNEY U

Wind Speed

A screenshot of a cell phone

Description automatically generated

Precipitation

A screenshot of a cell phone

Description automatically generated

Visibility

A screenshot of a cell phone

Description automatically generated

As we are dealing with ordinal variable (Severity), in order to predict severity outcome based on weather conditions, Ordinal Linear Regression model has to be fitted.

The association between *Y* and *X* is captured by *β*: a positive *βj*means that with the values for all other *X* variables fixed, an increase in *Xj* is associated with a stochastic increase in the distribution of *Y*

n. **Estimate** – These are the ordered log-odds (logit) regression coefficients. Standard interpretation of the ordered logit coefficient is that for a one unit increase in the predictor, the response variable level is expected to change by its respective regression coefficient in the ordered log-odds scale while the other variables in the model are held constant. Interpretation of the ordered logit estimates is not dependent on the ancillary parameters; the ancillary parameters are used to differentiate the adjacent levels of the response variable. However, since the ordered logit model estimates one equation over all levels of the outcome variable, a concern is whether our one-equation model is valid or if a more flexible model is required. The odds ratios of the predictors can be calculated by exponentiating the estimate.