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** – categorical 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

Temperature(oC) variable analysis

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

There are 11 temperature values between -70 and -55 °C

There are 7 temperature values between 55 and 80 °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.

23 measurements have been removed

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.

A screenshot of a social media post

Description automatically generated

Temperature data between Severity groups are not normally distributed as median is shifted towards the second half of the plot and do not represent Gauss.

This is proved further with Shapiro-Wilk test. The [Shapiro-Wilk test](https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test) evaluates a data sample and quantifies how likely it is that the data was drawn from a Gaussian distribution.

A screenshot of a cell phone

Description automatically generated

Anderson-Darling test is another statistical test of whether or not a dataset comes from a certain probability distribution, e.g., the normal distribution.

A screenshot of a social media post

Description automatically generated

A screenshot of a social media post

Description automatically generated

These test could fail due to large amount of data present in the dataset. Visually, it could be said, that data is more or less normally distributed and we can use parametric tests.

T-tests between different severity groups indicates, that there are significant differences between the means of compared groups as p values drop below 0.05.

A screenshot of a cell phone

Description automatically generated

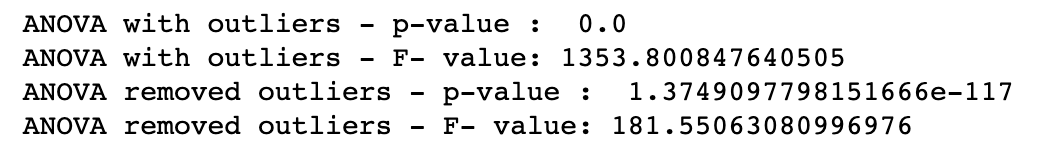
For such dataset, Z test is better, which is an alternative for t-test, when there are >30 values of variable. This test also indicates, that there are significant differences between the means of compared groups.

A screenshot of a cell phone

Description automatically generated

A large F ratio means that the variation among group means is more than you'd expect to see by chance. You'll see a large F ratio 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.

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.



Kruskal-Wallis test

A screenshot of a cell phone

Description automatically generated