Exercise 4: Naturalistic data analysis

Alexander Rasch
alexander.rasch@chalmers.se

Pierluigi Olleja
ollejap@chalmers.se

September 30, 2021

Introduction

In this assignment you will apply your knowledge on naturalistic data to analyze the data from the 100-car Naturalistic Driving study. Specifically, this assignment will guide you into extracting relevant safety indicators to investigate if there is a difference in average speed between near-crashes and crashes. In this assignment you will be using MATLAB and the SAFER100car toolkit.

Note: This exercise lets you do a more exploratory analysis where you are given freedom to change existing methodology and asked questions. Hence, Grader will not be used for this exercise. Be sure to document your answers in the MATLAB script.

Note: You are encouraged to discuss your solutions with others but you must write and submit your own code implementation and answers.

Deadline: October 14, 2021 (23:59)

Learning objectives

After having performed this assignment, you shall be able to:

- Explain which data (objective and subjective) are collected in the 100-car Naturalistic Driving study
- Show how you can such data to understand the causes and scenarios in which accidents happened
- Use the 100-car dataset and the SAFER100Car toolkit to test hypotheses according to the FESTA framework:

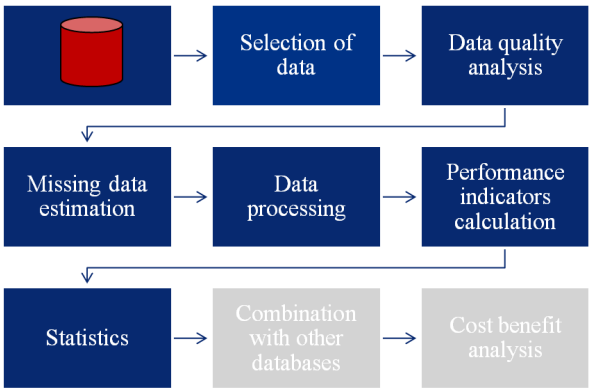


Figure 1: FESTA framework, `http://wiki.fot-net.eu/index.php?title=FESTA`

- Run queries to retrieve information on vehicle dynamics and driver state in safety critical situations from the 100-car data
- Identify quality issues related to the analysis of data collected in real traffic and implement strategies to address such issues
- Perform statistical analysis for testing null hypotheses and ensure that the data meet the assumptions of the statistical tests
- Identify contributing factors to accident related to human behavior

Preparations

1. Download the material from Canvas. `Exercise.m` is the template script to solve the exercise. The script has some guidelines to complete the exercise and the boilerplate. Complete the parts marked by `======== YOUR CODE HERE ========`. The function `isGoodQuality.m` checks the quality of the data (you are welcome to improve the criteria to discard questionable data). `Events` is the folder that contains the data for each crash and near-crash event (same data used by the SAFE100car toolkit). The folder `Dictionaries` contains the 100-car data dictionaries that are useful to understand the data structure (same dictionaries included with the SAFER100car toolkit). Note: the variable names may be slightly different between the dictionaries and the MATLAB data).
2. Open `Exercise.m`, fill in your group number in the header of the file, and start the tasks below.

Tasks

1. Within the for-loop that scans the event files, load the data for each event. Each event file holds the structure Data100Car. The structure Data100Car contains all the data collected during the 100-car study. The dictionaries are helpful to understand the data included in the structure Data100Car.

2. Analyze the event only if the incident type is a Rear-end striking. Look at the dictionary `ResearcherDictionaryVideoReductionDatav1_1.pdf` to understand how to extract this information.

3. Extract the data for the vehicle speed. Look at the dictionary `100CarTimeSeriesDataDictionary_v1_1.pdf` to understand how to extract this information.

4. Naturalistic data can be corrupted. If so, you will discard the data (but you will keep track of the amount of data that you are discarding). Use the function `isGoodQuality` to assess the overall quality of the speed data.

5. Even if the data quality is good overall, a closer look at the speed data would reveal that sometimes there are artifacts in the signal. Explore some events by plotting the vehicle speed. Can you identify artifacts in the signal?

6. If so, you could implement a simple, yet effective, strategy for dealing with this issue. For example, you could do a linear interpolation to recover the data. For example, Figure 2a shows that vehicle speed sometimes dropped suspiciously to zero. By interpolating, a more realistic speed profile can be recovered. You are encouraged to apply also better methods to recover the speed information.

7. For the analysis, we are interested only in the speed during the incident. Use the variables `start` and `end` in `Data100Car.Video` to extract the right speed segment from the event (e.g., see Figure 2b).

8. Depending on the incident category (i.e., crash or near-crash), store the event ID and the mean speed during the incident in the right variable (e.g., `crash_event_ID` vs. `near_crash_event_ID`). Be aware of some NaN values in the signal, the function `nanmean` can help.

9. Compute the proportion of data that you discarded because of low quality. Is the proportion of missing data usual for a naturalistic dataset? **Please, write your answer in the MATLAB script within a comment.**

10. Now you have two groups: the average speed for (1) crashes and for (2) near-crashes. You want to test the following hypothesis: "The average speed is higher in rear-end striking near-crashes than in rear-end striking crashes". To test this hypothesis, you could run a t-test.

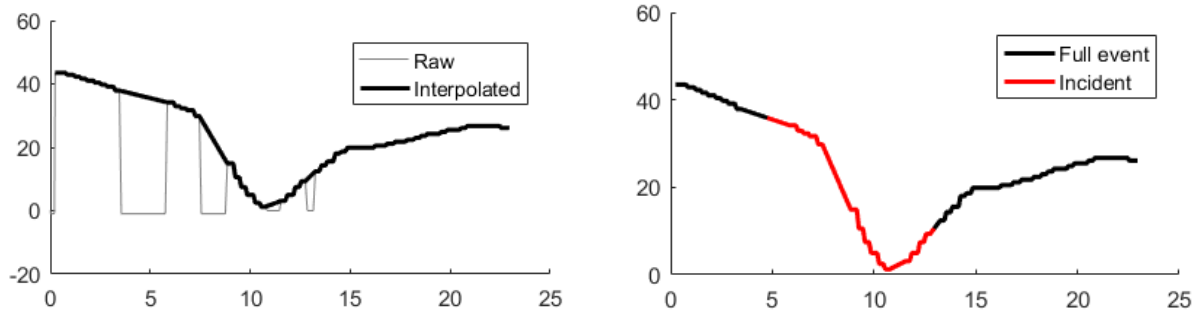
11. However, before running the t-test, you should verify that the data meet the assumption of normality. If the data are not normally distributed, the result from the t-test may be wrong or misleading. To do so, plot the distribution of average speed values for the crash and near-crash events. You should obtain something like Figure 3a.

12. If the distribution for the average speed does not quite follow a normal distribution, you could try a data transformation. For example, you could try to apply the square root. You should obtain something like Figure 3b. It seems like the transformation made the data fit the normality assumption better.

13. Run the t-test. You can run a one-tailed t-test with unequal variance. Read the documentation for `ttest2` to understand how to use the command and interpret the results of the test.

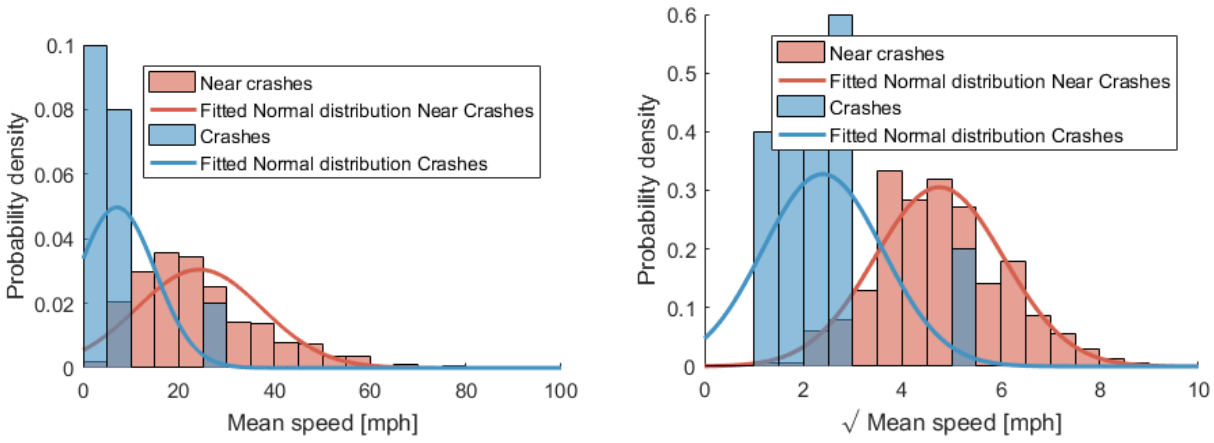
14. Does the t-test confirm that the "Average velocity is higher in rear-end striking near-crashes than in rear-end striking crashes"? **Please, write your answer in the MATLAB script within a comment.**

15. Finally, inspect the narratives of the crash events in `Data100car.Narratives`. The list of event IDs of interest are stored in the variable `crash_event_ID`. By reading the narrative of the events, what do you think is the most common scenario and critical reason that lead to a crash? **Please, write your answer in the MATLAB script within a comment.**



(a) Whole event (b) Marked incident within event

Figure 2: Vehicle speed profile in the event (raw and interpolated).



(a) Original scale (b) Square-root scale

Figure 3: Probability density of the average speed for crashes and near-crashes

Submission

Submit your solutions (`Exercise.m`) in the assignment in Canvas. It is sufficient if at least one group member submits your solution in Canvas. See deadline above.