Faye Bandet
ISTA 116 Fall 19
12/3/19

ISTA 116 Final Project Report

The research question I plan on investigating is "Do most California drivers in suburban neighborhoods drive over speed limits rather than under the speed limit, and does the direction driven significantly influence speeding?" To answer this question, I wrote an R program that makes precise calculations and concludes whether to support or reject the hypothesis that speeding is common in suburban neighborhoods. The R program I designed uses multiple data sets of suburban drivers that include car speed observations as well as observations of speed with the direction of traffic. My conclusions provide valuable information that will bring awareness to local road safety and show whether additional caution, as a driver or pedestrian, is of importance to the general public.

I acquired public data from The Data and Story Library (DASL), specifically two studies conducted by Stanford University Engineering scholar John Beale. The first study, titled "Car Speeds 100", was conducted over a two-month period in Stanford, California. It consists of 100 speed observations of cars that I used to visualize, analyze, and compare with data from the next study. To strengthen the external validity of my conclusions, I compared findings with a second data set that is much larger with 500 separate entries. The observations in this study are more complex and include the recorded direction of traffic for each observation. The large data set, titled "Car Speeds", will be used to calculate whether there is significant influence from the direction a car was traveling on the speed in observations. The accuracy of these findings is aided with the analysis of 600 total speed observations, 500 of which include directional observations. Information on the DASL webpages for these studies did not indicate any stated limitations in the sampling or data collection process, however I include personal observations of the data collection methods later in the reflection section. I created histograms, Q-Q plots, and contingency tables visualize the methods used for analysis in this project, shown in Figures 1-8 below.

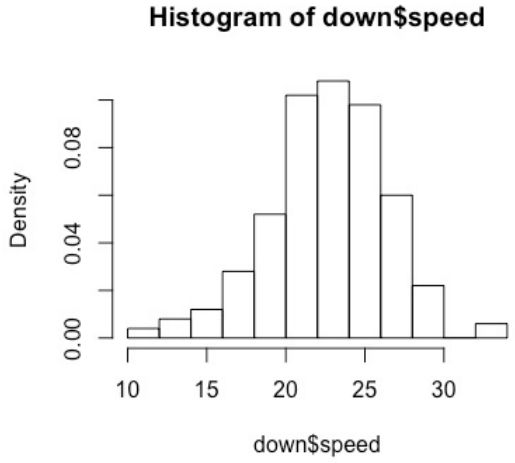

Fig. 1

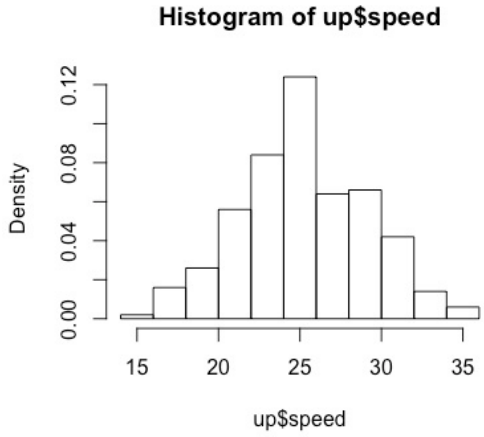

Fig. 2

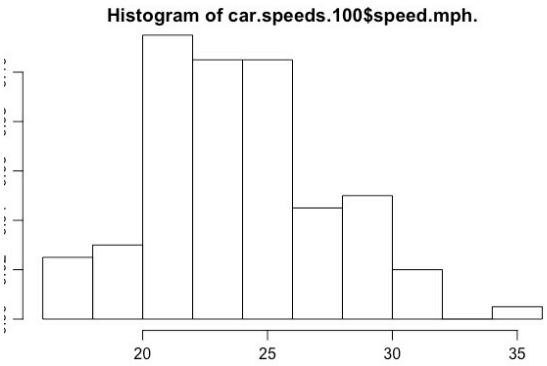

Fig. 3

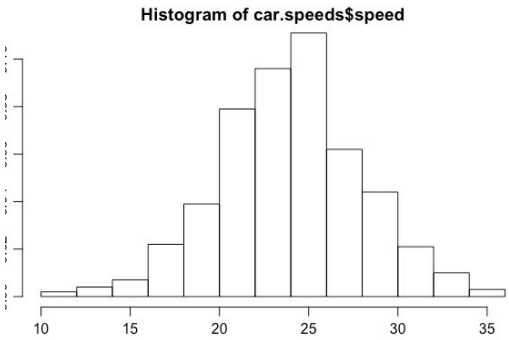

Fig. 4

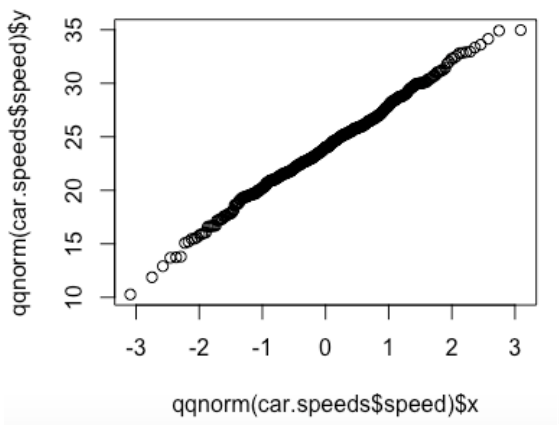

Fig. 5

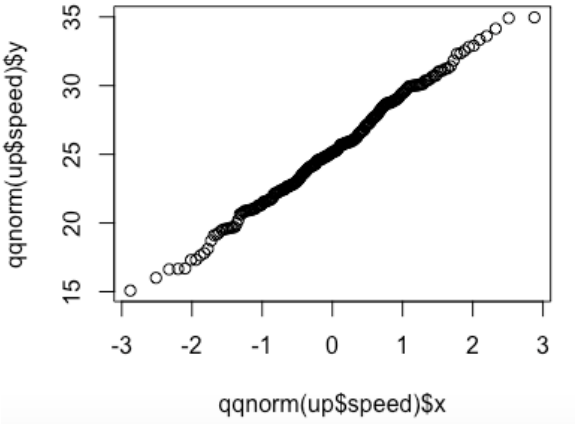

Fig. 6

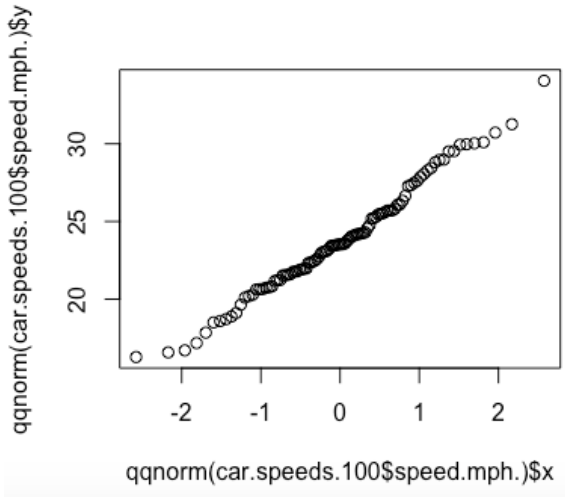

Fig. 7

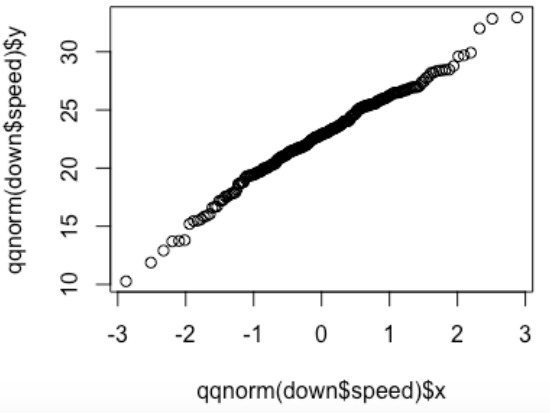

Fig. 8

Analysis of the car speed and directional data gave us insight into typical drivers of Stanford, California. For reference, the speed limit in the area is 25 miles per hour. The observations came from randomly sampled cars at the location, which satisfies the random condition for inference, and each observation was an independent observation of a car speed or speed with direction, which satisfied the independence condition. The results from my contingency tests and proportion tables found that the car.speeds.100 data set had 35%[35] of drivers speeding, and 65%[65] of drivers drove at or under the

speed limit. In the car.speeds data set, 40.8%[204] of drivers were speeding and 59.2%[296] of drivers drove at or under the speed limit. In the car.speeds data set, specifically analyzing the drivers going "Up", 52%[130] of drivers were speeding and 48%[120] of drivers drove at or under the speed limit. In the car.speeds data set, specifically analyzing the drivers going "Down", 29.6%[74] of drivers were speeding and 70.4%[176] of drivers drove at or under the speed limit. These tests proved that the condition of 10 failures and 10 successes, in this case speeding, was satisfied for inference. We know that with 500 and 100 observations each, these studies pass the condition of having a large sample size for inference. Overall the full data sets showed that most drivers drove at or under the speed limit [car.speeds = 59.2%, car.speeds.100 = 65%]. Histograms stated in figures 1-7 above show that there are even distributions, that are relatively normal and uni-modal with little to no skew, which satisfies approximate normalcy for inference. Q-Q plots supported these findings and showed that observations had little skew with few residual speed observations that had close variance to the majority of observations. According to summary statistics, on average, the car.speeds.100 data set had lower median, mean, and mode speeds than car.speeds, respectively [23.52, 23.84, 22 and 24 (both had equal 15 obs.)] versus [24.02, 23.98, 25]. The minimum and maximum speeds of car.speeds.100 were 16.27 mph and 34.06 mph, compared to car.speeds with a minimum of 10.27, and a maximum of 34.97. I analyzed the 500 reported observations of speed connected to categorical observations of directions to see if there is a significant difference in speed connected to the direction driven. I checked the variance of speeds between the directions with a proportion test that had one degree of freedom, and concluded that there is a significant difference in speeds driven in each direction, because the p-value is $< 2.2e-16$, using a confidence interval of .95. This means there is an association between speed and direction driven. Additionally, using two sample t-tests with a .95 confidence interval on the speeds of cars going the directions up and down revealed a p-level of 1.771e-13. This low p-level indicated that there was a significant difference in speed going each direction, which supports the association between direction and speed. This is supported by the proportion tests for the directions, since cars driving "Up" had 52% of observations speeding, compared to only 29.6% of cars speeding while driving "Down". I found that standard deviation for the car.speeds data set was 3.946692 mph, car.speeds.100 was 3.563338 mph, and the up and down subsets of car.speeds were 0.6023039 and 0.6068254 respectively. I used the standard error, which for the car.speeds data set was 0.1765014 mph, car.speeds.100 was 0.3563338 mph, and the up and down subsets of car.speeds were 0.2438958 and 0.2290758 respectively, to calculate the critical values. The critical values observed with .95 confidence intervals for the car.speeds data set was -0.9459427 mph, car.speeds.100 was -0.3922807 mph, and the up and down subsets of car.speeds were -0.7134025 and -0.7609318, which would be expected according to previous tests. The extent to which these results may be generalized beyond the sample could be strong for the immediate surrounding area such as the sample city, Stanford, California. It would be less strong or valid of a finding to generalize to other suburban neighborhoods because there was only one location used for data collection.

Looking back critically at my original statistical question, "Do most California drivers in suburban neighborhoods drive over speed limits rather than under the speed limit, and does the direction driven significantly influence speeding?", we can now

clearly see the answer: my analysis found that in both studies, most drivers drove at or under the speed limit, supported by proportion tables. Also, the direction of observations was associated with and significantly influenced speed according to proportion tests. I found that the recordings of drivers going up the street drove significantly faster, with a maximum of almost 10 mph over the speed limit and were significantly more likely to speed than people driving down the street. I can conclude that because most people were not speeding, my hypothesis was rejected, and the null is true that drivers generally follow the speed limit. On the other hand, I did prove that the direction driven significantly influenced speed within those subsets of data. These claims are consistent with and supported by the data found in my analysis with plots, proportion tables, proportion tests, t-tests, as well as multiple other data tests. I used publicly available data, which upon reflection had some limitations and weaknesses, primarily the limitation of data sample collection locations. All observations were collected in the same city, which makes the data heavily represent the local driving speeds rather than sampling from multiple suburban neighborhoods in California. This made it hard to generalize across the state without data recordings of car speeds across multiple locations/cities. Generally, the data reported is not missing any important variables, however additional data would always be beneficial in strengthening the external validity of the analysis. Beale's data collection was recorded in a less- than-optimal manner because it did not describe in depth how the data collection process was completed. Finally, I propose the idea that this study could be followed up with additional data collection and studies to append to this data and compile new observations of speed with the driving direction in multiple locations that would be a stronger reflection across the state. Furthermore, this data could be more detailed and complex if there were additional variables such as the state the vehicle is registered in, because we can eliminate out of state drivers from the data. It would be ideal to have data from both northern and southern California to create an accurate profile of driving patterns.

Works Cited

Beale, John. “Car Speeds 100.” *The Data and Story Library*, 2019, dasl.datadescription.com/datafile/car-speeds-100/?_sfm_cases=4+59943&sf_paged=6.

Beale, John. “Car Speeds 100.” *The Data and Story Library*, 2019, Beale, John. “Car Speeds 100.” *The Data and Story Library*, 2019, https://dasl.datadescription.com/datafile/car-speeds/?_sfm_cases=4+59943&sf_paged=6