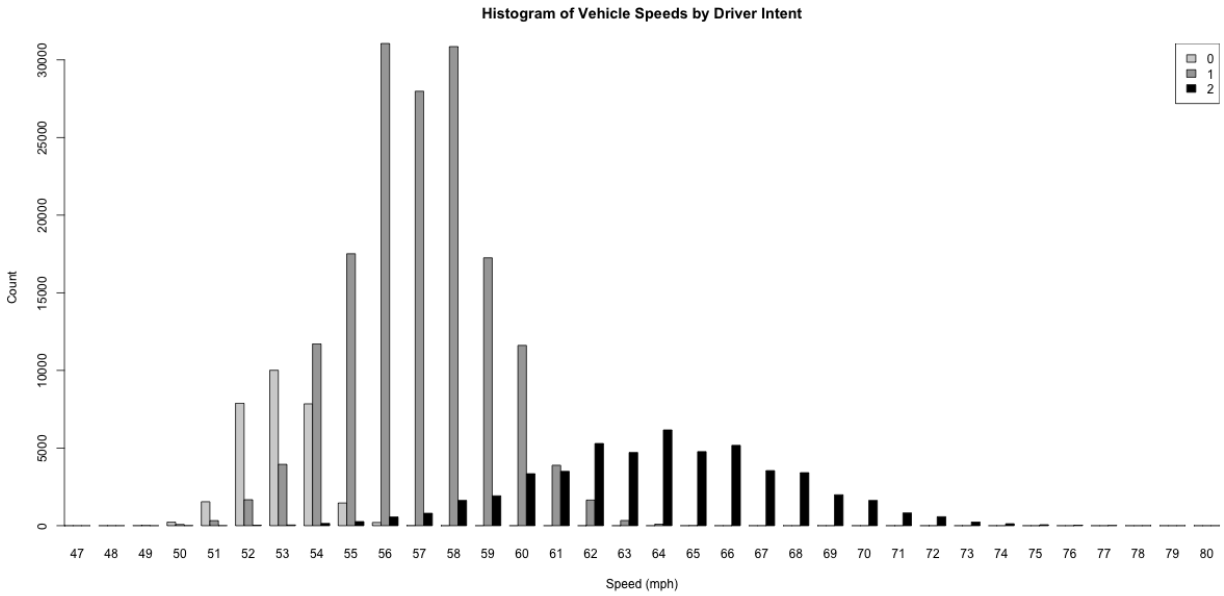
Ethan Chang
HW_01
CSCI 420

1. For Part A, what was the trend as the number of elements in each vector went up, what happened to the amount of data within one standard deviation of the origin?

As the number of dimensions increases, the fraction of data within one standard deviation of the origin decreases dramatically. In 1 dimension, about 68.7% of the data falls within one sigma. This drops to about 39% in 2 dimensions, 19% in 3 dimensions, 8.8% in 4 dimensions, 3.8% in 5 dimensions, and only 1.4% by the 6th dimension.

2. For Part B, Show a bar graph of the entire histogram of speeds, by intention.
   By the way, the legal speed limit is 55 mph.
   How might we describe this data? Is it a mixture model? What kind of mixture model?
   [It is a CRISP Gaussian Mixture model. But explain why?] What do you notice about it? Are there lumps? What is odd? Is there anything odd? Is it bi-modal? Is it tri-modal? Four-modal? Speculate about why the histogram is the way it is. (2)

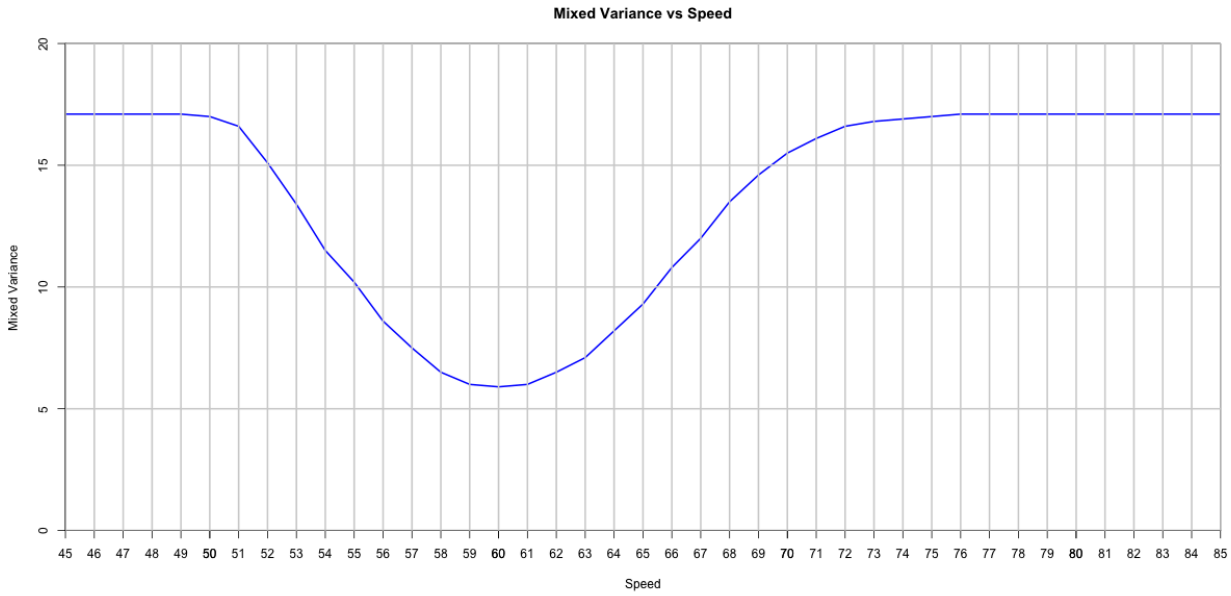


This data is split into three individual groups for each speed. For each speed, we have counts of Intent 0 (light grey), Intent 1 (dark grey), and Intent 2 (black). On the Y-axis, the chart represents the frequency of each group. From observing this bar chart, I noticed clustering from 51 to 53 that mainly focuses on Intent 0 with the highest frequency. From 55 to 58, I

noticed another clustering focused on Intent 1 with the highest frequency. Lastly, from 62 to 66, I noticed another clustering group focused on Intent 2.

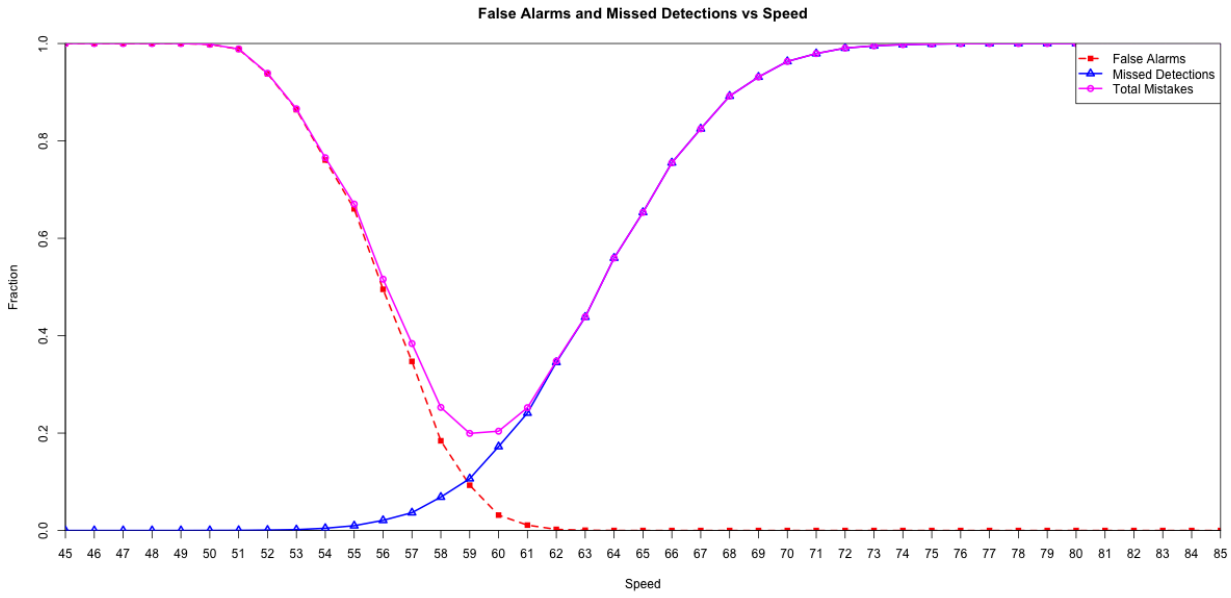
From observing the bar chart, it can be determined that this is a CRISP Gaussian Model. From researching CRISP Gaussian Models, I learned that there are separate groups by Intent 0, 1, and 2. This model would be considered tri-modal since there are three bumps or groups. The first bump contains light grey bars, which represent Intent 0; the second bump contains dark grey bars, which represent Intent 1; and the last bump contains black bars, which represent Intent 2. We know that 55 mph is considered the legal speed limit, and it centers around Intent 1. One odd thing about this data is that the Intent 0 cluster is limited to only below 55 mph, which makes me think the data may have been collected from specific kinds of roads. When driving on highways, I've noticed speed limit signs that can go up to 70 mph, so it seems unusual that the 70 mph region in the graph is categorized under Intent 2.

3. **Mixed Variance versus speed:** In your PDF, show your graph of Mixed Variance versus speed. (2) This could be used to break the drivers into two groups. Which speed would you use to split the drivers into two groups?



From observing the graph, the curve dips to the lowest point around 60 mph. This dip in the graph represents where the mixed variance is the lowest. So the two groups I would split up are drivers below 60 mph and drivers above 60 mph. The first group, who drive below 60 mph, most likely contains cautious drivers, while drivers above 60 mph are intentional speeders going beyond the speed limit. 60 mph is a well-balanced threshold, and sometimes drivers can go an extra 5 mph within the speed limit.

4. In your PDF, show your graph of the numbers of false alarms, misses, and mistakes as a function of speed. Eventually, we will want to minimize the number of mistakes. What one-rule would you use to decide if a driver was trying to speed? (2)



The graph showcases the false alarms (red) decreasing as speed increases, while total misses (blue) increase at the same time. The overall mistake curve (magenta) reaches its lowest point around 60 mph, which balances between the two errors of false alarms and total misses.

One-Rule: If the driver's speed is greater than 60 mph, then they are intentionally attempting to speed.

Conclusion:
Write up what you learned here using at least three paragraphs. (2)
What did you discover? Was anything unusual? What was surprising?
Was there anything particularly challenging? Did anything go wrong?
Provide strong evidence of learning.

From this homework, the first thing I did before starting the assignment was decide to select R as my computing language, not only because the instructions mentioned it was easy to use, but also for me to learn the R language. Throughout this assignment, I found it challenging to use R, especially as a beginner. Starting from Part A to Part C, learning the R language gradually became harder and harder. I believe the most difficult part was figuring out the plots and how to handle files such as datasets including speed and intents. For example, in Part C, I not only had to learn specific variable names in plots like xlab and col, but I also had to learn other functions such as creating new lines. R is also known for statistical analysis and modeling and has functions that create vectors, which actually reminded me of the C programming language but with memory location. However, overall I developed problem-solving and R language skills, which was the positive side of this project.

From Part A, B, and C, I looked closely and analyzed datasets and mathematical models in written formulas and interpreted them into coding. For example, in Part B, I had no idea there was more than one Gaussian model, such as the CRISP model. Studying the histogram I programmed for Part B helped me understand the situation a lot better. It made me realize that histograms are one of the essential parts of data visualization since they show different groups of intents with each speed and frequency count. By visualizing the data, I was able to understand how to read it more clearly and identify the clusters. I learned that normally Gaussians have one cluster, but for CRISP it actually has multiple clusters, and I ended up with a tri-modal distribution with three different peaks. Another thing this histogram showed me is that the data was likely taken from a specific street or road, since it emphasizes 55 mph as the legal speed.

This homework assignment helped me gain insights and a better understanding of why data science is very important and can be an essential field in the future. I learned how to apply math models with data to solve real-world problems and figure out driver behaviors. Another thing I realized is that data is very complex, and numbers alone cannot fully confirm driver behavior. We would also have to hear the driver's side of the story or consider their conditions to understand what caused them to speed up. It is possible they could have

different intentions, for example a driver might speed up because they are late for work, or simply for fun. Additionally, we do not even know the specific time they are driving. I find it fascinating and fun to apply math equations to figure out real-world problems by using data.