Seth Metcalf

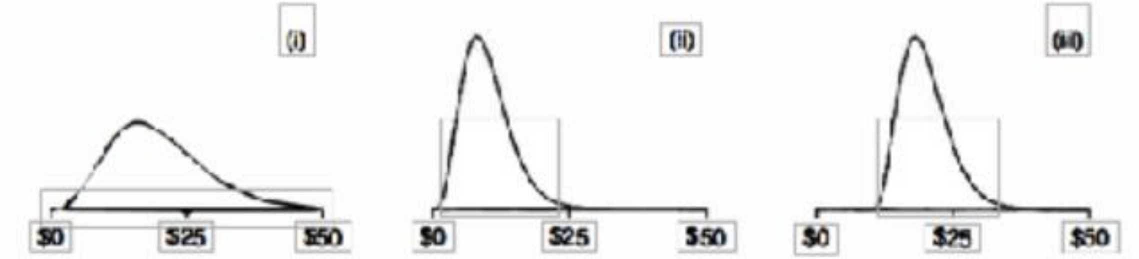
Prof. Hank Ibser

Homework 2

5 February 2021

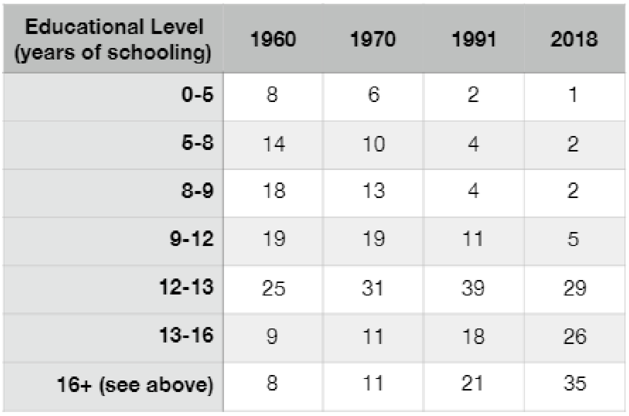
Statistics HW 2

1. Ch 3, A7:



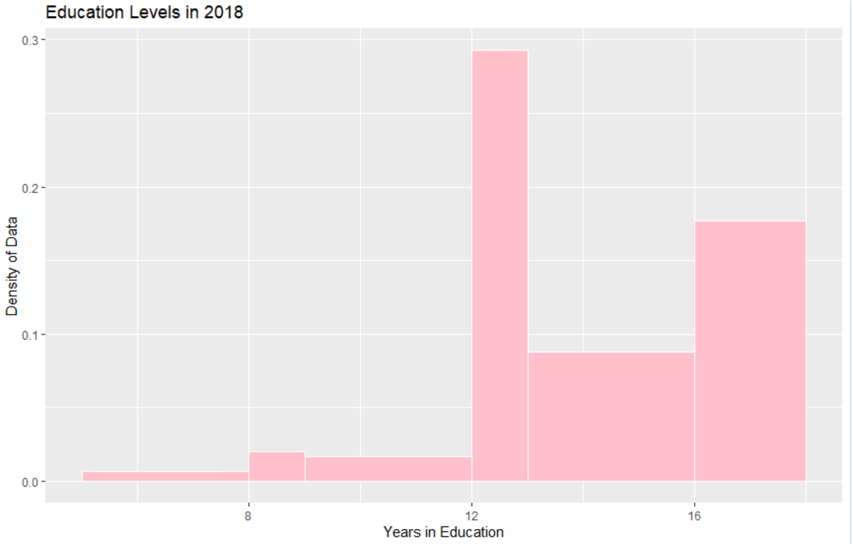
   a. I is associated with B. II is associated with A. III is associated with C. We know this since III is the same as II, just roughly $10 more average, aligning A and C with those graphs respectively. And while I is more flat, due to the spread being wider on the curve, it can be assumed that the mean is approximately double than that of II.

2. Ch 3, based on B1 - 2:



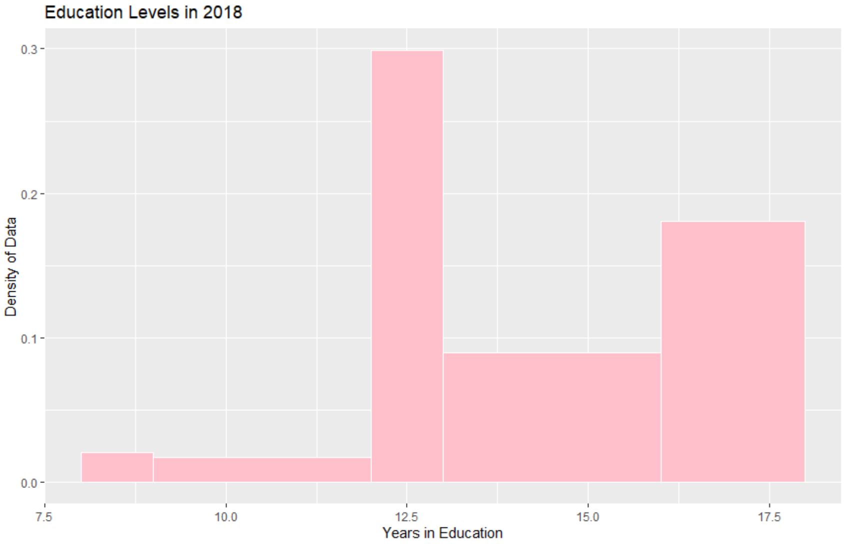
Educational Level (years of schooling)	1960	1970	1991	2018
0-5	8	6	2	1
5-8	14	10	4	2
8-9	18	13	4	2
9-12	19	19	11	5
12-13	25	31	39	29
13-16	9	11	18	26
16+ (see above)	8	11	21	35

   a. The Histogram has bumps, spikes in those areas because of the differing bin sizes. Putting concentrated data right after a large section (5-8 to 8-9, 9-12 to 12-13, etc.) creates an unevenness in the graph

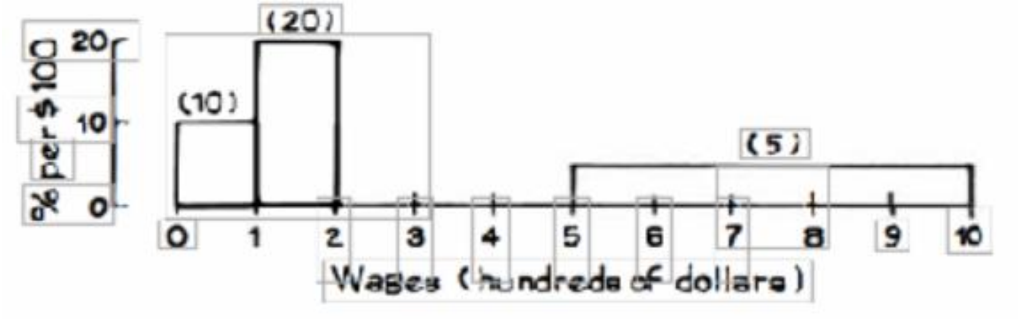


If you were to redraw the graph, creating one interval between 0-8, it would not really change the shape of the histogram. The concentration of the values happen mostly on the higher end of the years of schooling.

This is the result:



3. Ch 3, C1:



   a. It must be 15% per $100 tall. Since the areas of all of them must equal 100%; 100% - (10%*1) – (20%*1) – (5%*5) = 100% - 10% - 20% - 25% = 45% over a width of 3 = 15%.

4. Ch 3, Rev 2:

a. There is a higher chance that there are more children age 1 than adults age 71. This is because while there is 7% of the population between 0-5, there is only 6% of the population between 65-75, leaving them with a lower percent population for a wider range of age.

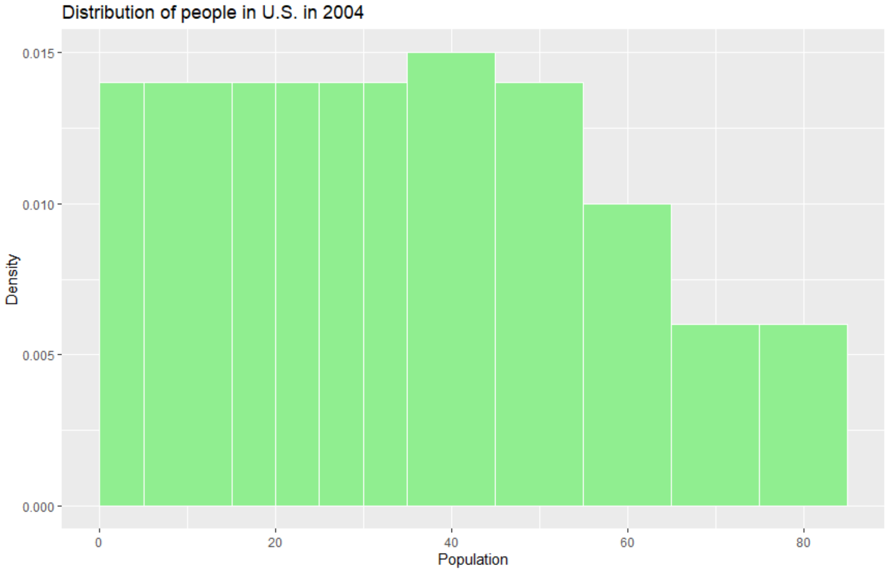
b. In this case, it is also the same logic as before. Since 7% of the population resides at 20-25 (roughly 1.4% per age), only 10% of the population resides at 55-65 (roughly 1% per age).

c. There are more people age 0-5. Since 6% of the population resides between 65-75, and the age range given to us is 65-70, it can be assumed that only 3% of the population resides in that 5 year interval compared to the 7% that resides in the 5 year interval of 0-6.

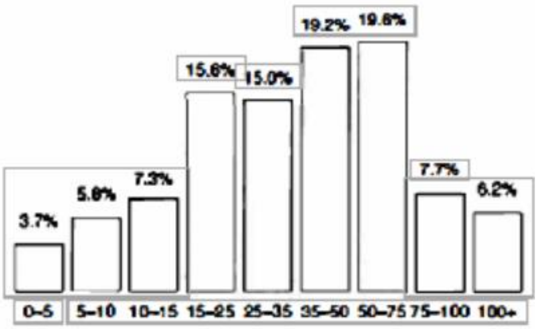
d. Closest to 50%. Since 15% of the population reside in 35-45, 14% reside in 45-55, 10% reside in 55-65, 6% reside in 65-75, and another 6% reside in 75+; adding all of these together (15+14+10+6+6) results in roughly 51% in age range 35+, which is closest to 50%.

e. Code in R

```
# Homework 2
population = rep(c(2.5, 10, 17.5, 22.5, 27.5, 32.5, 40, 50, 60, 70, 80),
c(7 ,14 ,7 ,7 ,7 ,7 ,15 ,14 ,10 ,6 ,6 ))
dist = data.frame(population)
distogram = ggplot(dist, aes(x = population, y = ..density..))
distogram + geom_histogram(col = "white", fill = "light green",
breaks=c(0,5,15,20,25,30,35,45,55,65,75,85))  + ggtitle("Distribution of people in U.S. in
2004") + xlab("Population") + ylab("Density")
```



5. Ch 3, Rev 8:



   a. True. In the range $10,000-$35,000, there are a total of 3 bins that cover the areas 10-15, 15-25, and 25-35. In the bin with a binwidth of 5 (10-15) there only resides 7.3% of the population, while those with a binwidth of 10 have closer to 15% of the population, but considering the change in binwidth, it shows that in this range, families are spread out evenly.
   b. True, in the two bins that are given for $35,000 to $75,000, they hold the same binwidth and have a comparable population percentage that they seem to be fairly spread out.
   c. False, the graph is not a histogram as the binwidths are not represented by the width of the bars on the graph and there is not a vertical axis that represents density.
6. Ch 4, B1-2:
   a. The average is around 50. The average is the same as the median.
   b. The average is around 25. The average is the same as the median.
   c. The average is around 40, the average is greater than the median.
7. Ch 4, D8:
   a. The averages should be about the same assuming that they take from the same sample. Though there is a possibility that the investigator that is taking a sample of 100 men may have a few outliers that either raise or lower his average from the average given from the 1000 men.
   b. The investigator that only sampled 100 men will get a bigger standard deviation because he did not gather enough data to make his average more precise/the spread on the graph is larger.
   c. The investigator that only sampled 100 is more likely to get the tallest of the sample men because he has less data to work with (more likely to be affected by outliers.
   d. The investigator that only sampled 100 is more likely to get the shortest of the sample men because he has less data to work with (more likely to be affected by outliers.
8. Ch 4, E4:
   a. Every class had the same average. Even though their spreads were wildly different, all of them averaged to 50 points.
   b. The biggest SD was from class C. Since they had the widest spread of numbers, it causes them to have the largest standard deviations.
   c. They all had the biggest range since they all had the same maximum and the same minimum.
9. Ch 4, Rev 6:
   a. The average for I is 60, the average for II is 50, and the average for III is 40.

b. Graph I has a median that is bigger of the average. Graph II has a median that is about equal to the average. Graph III has a median that is less than the average.
c. Based off of histogram III's spread, it seems that the standard deviation should be around 15.
d. False, SD is not based off of the "largeness" of data values, it is based off of how much they differ from the mean. Since both I and III differ from the mean in similar ways with similar structures (reversed), it can be assumed that their standard deviations are roughly similar.