Data 8 Mentoring Exercise 3 Solutions

Jessica Hu and Weston Hughes

1 Visualization

1.1 Plotting

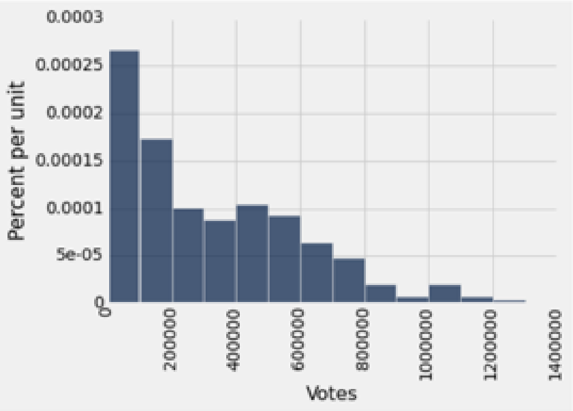
Match each picture below to a call to:

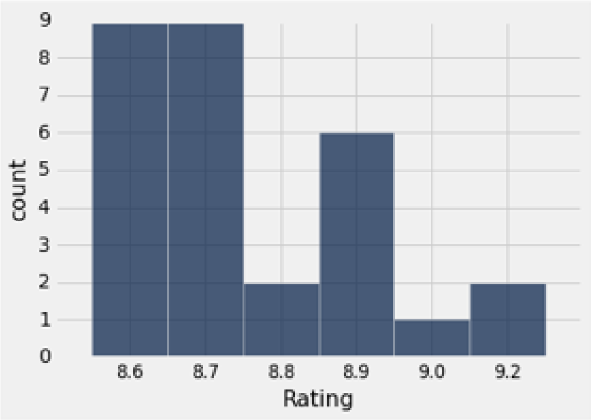
1. barh: bottom right

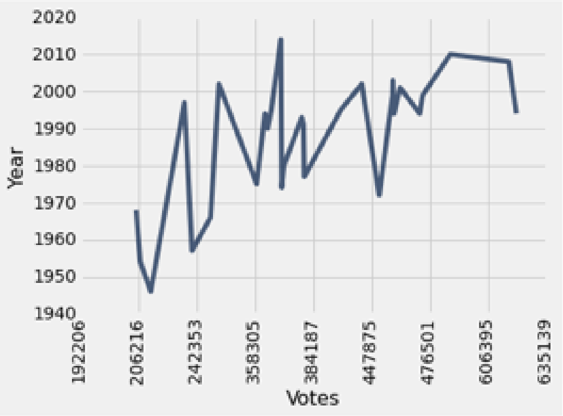
2. plot: bottom left

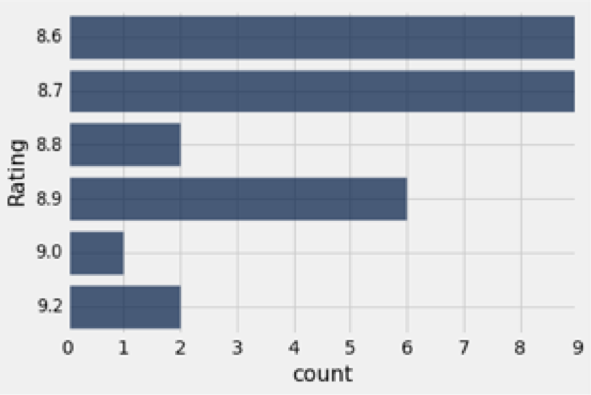
3. bar: top right

4. hist: top left







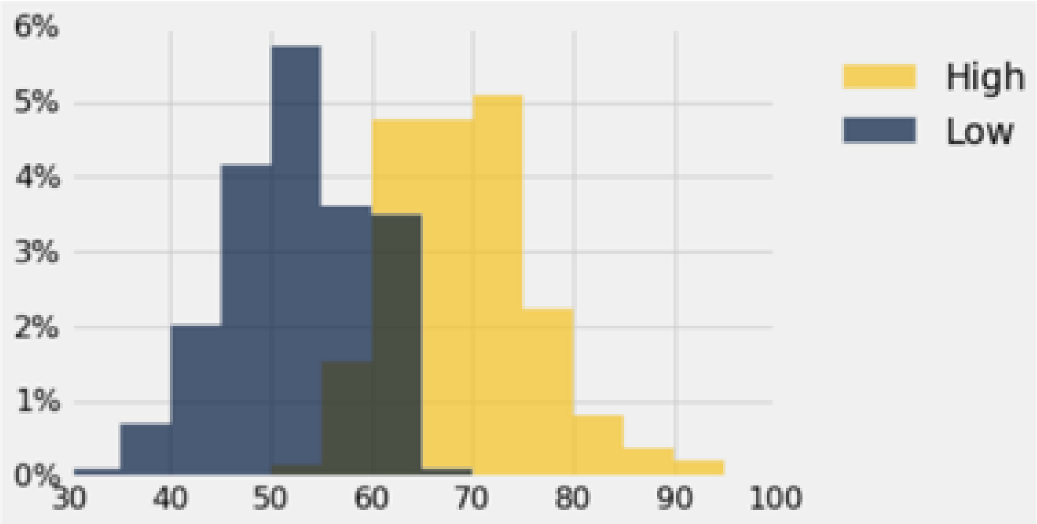


1.2 Histograms

In this chart for a year of daily high and low temperatures, answer these questions:

1. Is this chart a histogram or a bar chart? Why?
   A histogram, because it displays continuous data and has areas representing proportions.

2. If you were to pick one day out of the year, uniformly at random, what would you guess was the low temperature for that day?
   Around 50, given that the mean of the blue histogram is around there.



1.3 Probability/empirical histograms

Suppose I were to flip a fair coin 10 times and count the number of heads.

1. If I take 100 trials and plot the number of heads for each trial on a histogram, would that be a probability or an empirical histogram?
   Empirical

2. If I plotted the number of heads in all possible combinations of coin flips, would that be a probability or an empirical histogram?
   Probability

2 Sampling

2.1 iteration

Explain in words what this for loop would do:

```
lst = ['Hello', 'world', '!']
for i in np.arange(len(lst)):
    print(lst[i]*(i+1))
```

It would print Hello once, world twice, and ! three times

2.2 Bias/Variability

1. Is it better to have high or low bias? What about high or low variability?
   We want both to be low
2. What is the relationship between bias and variability?
   The higher the bias, the lower the variability.
3. If you get a test score that is below average (i.e. below the mean), then does that mean that you did worse than half of the people who took the test?
   No - this would be true of the median.

2.3 Normal Curve

If I had a normal distribution of heights for women, where the mean is 66 inches and the standard deviation is 3 inches:

1. Calculate the mean and standard deviation for this distribution in standard units
   Mean is 0, SD is 1
2. Can you write a formula to represent an arbitrary height in standard units? (e.g. if you were given height 63, how would you represent that in standard units?)
   Yes, for a an actual height h the height is standard units is $\frac{h-66}{3}$
3. What can you say about the proportion of women who are taller than 69 inches? Shorter than 63 inches?
   Given that the distribution is normal, we can say that about 16 percent of women fall into each category.

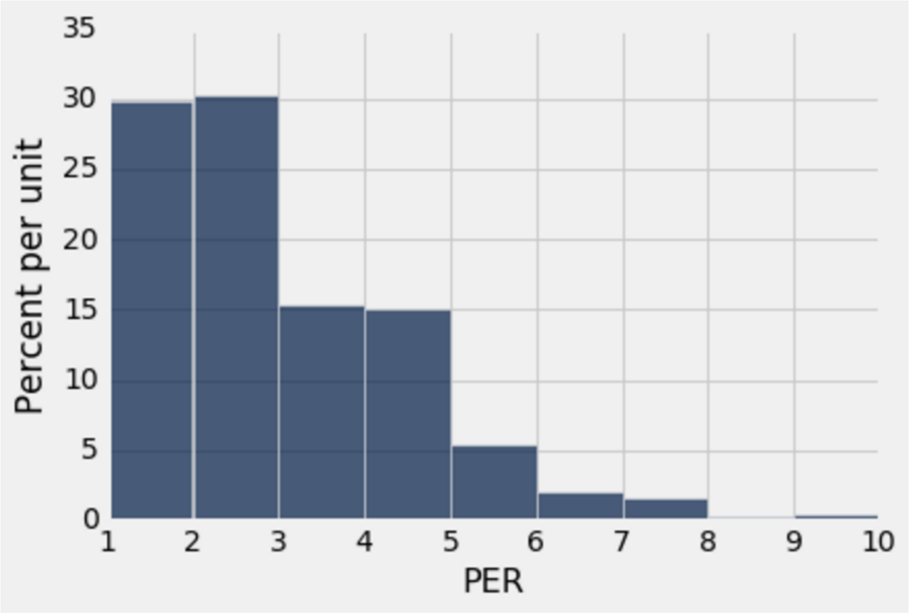
3 Prediction

3.1 Central Limit Theorem

Consider the following normed histogram of the number of people in each household in the Bay Area, generated in homework 4. Recall that the sample size that this histogram was generated from is huge. *Hint: When does the Central Limit Theorem apply?*

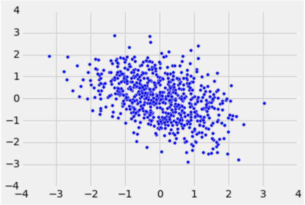
1. Say we took a sample of 20 households, and counted the number of people living in each one, and plotted the results on a histogram. What, if anything, could we predict about this histogram?
   Very little - the values would mostly fall towards the lower end, between 1-5. The sample size is insufficient to say anything more.
2. Say we took a sample of 1000 households, and counted the number of people living in each one, and plotted the results on a histogram. What, if anything, could we predict about this histogram?
   The histogram would be very similar to the given one.

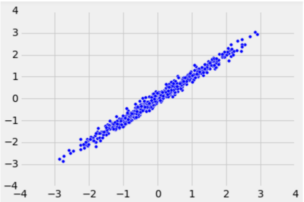
3. Say we took 500 samples of 20 households each, and for each sample we took the average number of people living in a household, and then plotted the result of each sample on a histogram. What, if anything, could we predict about this histogram?
By the central limit theorem, it would be a bell curve around the mean of the given histogram.

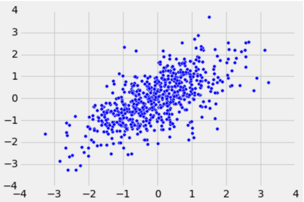


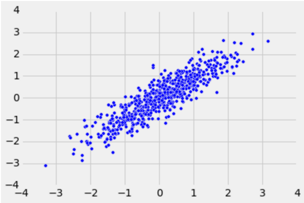
3.2 r

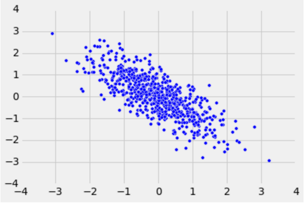
Order the following 7 scatter plots in ascending order by their correlation coefficient (r).

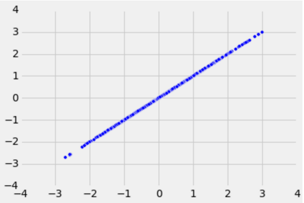
a)


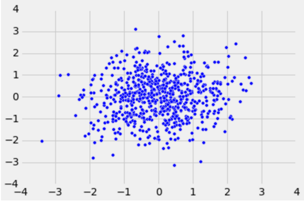
b)


c)


d)


e)


f)


g)


$e < a < g < c < d < b < f$

3.3 Predicting Scores

Suppose we want to predict a Sarah's performance on an exam based on her score on a prior midterm. We are given the following information: The average score on the midterm was a 63, with a standard deviation of 11. The average score on the final was a 70, with a standard deviation of 16. The regression coefficient between scores on the midterm and scores on the final was .48. Sarah scored a 68 on the midterm.

1. Convert Sarah's score on the midterm to Standard Units.
   .4545

2. Without doing any more calculations, should we expect Sarah's score on the final, in standard units, to be greater than or less than her score on the midterm?
   Less than, by regression to the mean.

3. Calculate Sarah's expected score, in Standard Units, on the final. Was your prediction in the last part correct?
   .218

4. Convert Sarah's expected score to a real percentage.
   73.5