Review I

- Statistic concepts: Population, sample, variables, values, cases, statistics, distributions.
- Summarizing data using statistics:
  - Centers: Mean, Median
  - Spread: Mean Square Deviation, Mean Absolute Deviation. Centers minimize the loss defined by the spread.
- Visualization of a distribution. Histogram.

Today:

- Visualization through boxplots.
- Visualizing two variables through scatter plots.
- Linear transformations of data.
- Why random sampling?

IQR, Boxplots, and Outliers I

```
quantile(myBikeCommute$Distance)
```

```
   0%   25%   50%   75%  100%
25.86 27.00 27.19 27.38 27.52
```

```
c(Q1, Q3, iqra, 1.5*iqra, Q1 - 1.5*iqra, Q3 + 1.5*iqra)
```

```
[1] 27.00 27.38  0.39  0.58 26.42 27.96
```

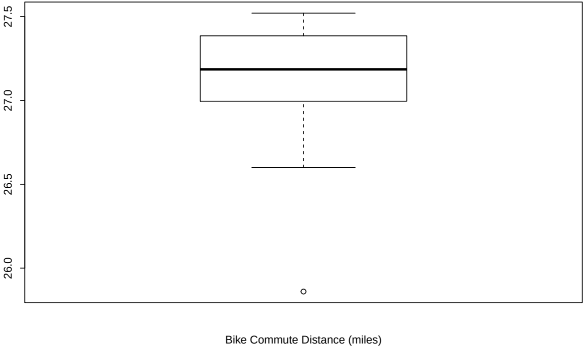
```
sort(myBikeCommute$Distance)
```

```
 [1] 25.86 26.60 26.74 26.88 26.90 26.91 26.91 26.91 26.92
[10] 26.94 26.94 26.94 26.95 26.99 27.00 27.00 27.01 27.02
[19] 27.02 27.03 27.03 27.05 27.06 27.09 27.10 27.16 27.16
[28] 27.17 27.20 27.20 27.27 27.29 27.31 27.31 27.32 27.32
[37] 27.33 27.34 27.34 27.36 27.36 27.38 27.39 27.40 27.40
[46] 27.43 27.44 27.45 27.46 27.48 27.49 27.49 27.49 27.51
[55] 27.52 27.52
```

IQR, Boxplots, and Outliers II

IQR, Boxplots, and Outliers III

```
boxplot(myBikeCommute$Distance,
      xlab="Bike Commute Distance (miles)")
```



IQR, Boxplots, and Outliers IV

Compare centers and spreads of speed distributions

```
by(myBikeCommute$AvgSpeed,myBikeCommute$Bike,mean)
```

```
myBikeCommute$Bike: Carbon
[1] 15.19
---------------------------------------------
myBikeCommute$Bike: Steel
[1] 15.04
```

```
by(myBikeCommute$AvgSpeed,myBikeCommute$Bike,summary)
```

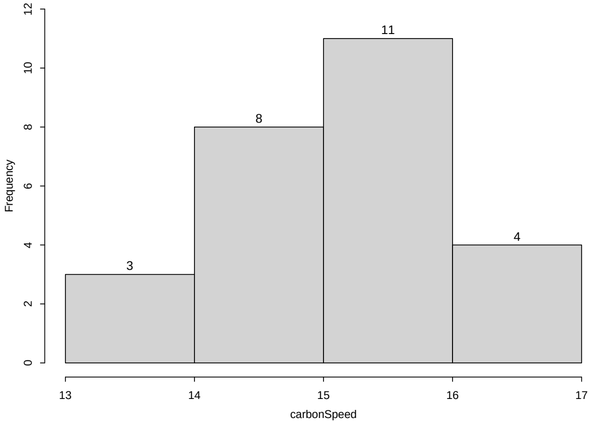
```
myBikeCommute$Bike: Carbon
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   13.4    14.6    15.2    15.2    15.9    16.3
---------------------------------------------
myBikeCommute$Bike: Steel
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   13.8    14.6    15.0    15.0    15.4    16.6
```

```
by(myBikeCommute$AvgSpeed,myBikeCommute$Bike,IQR)
```

```
myBikeCommute$Bike: Carbon
[1] 1.31
---------------------------------------------
myBikeCommute$Bike: Steel
[1] 0.8725
```

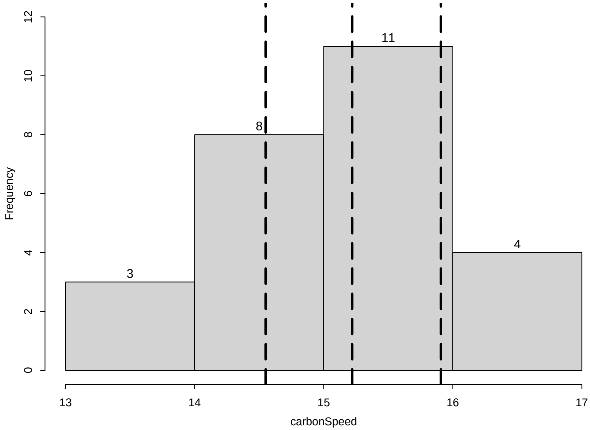
IQR, Boxplots, and Outliers V

These are NOT the quartiles of average speed for carbon bike

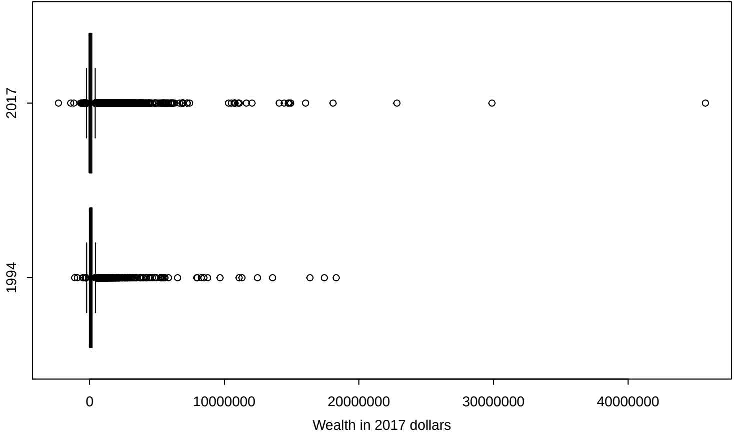


IQR, Boxplots, and Outliers VI

These ARE the quartiles of average speed for carbon bike

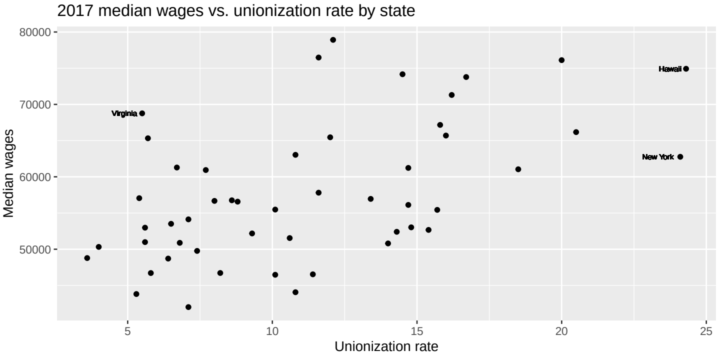


Wealth distribution I

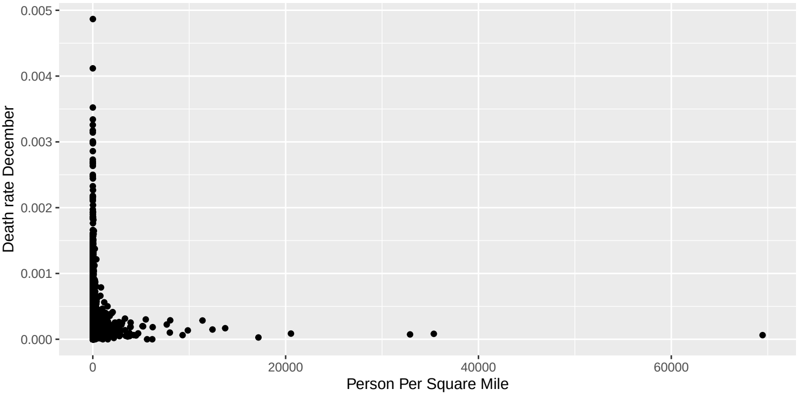


https://www.youtube.com/watch?v=QPKKQnijnsM

Looking at pairs of variables I

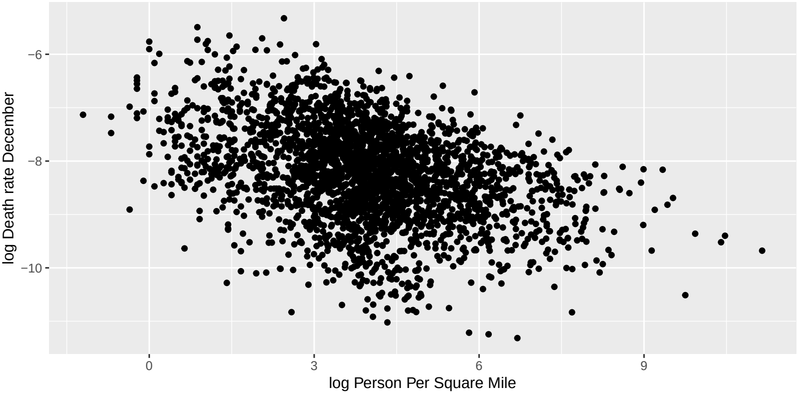


Looking at pairs of variables II



Doesn't look very informative. Lots of dots squeezed together.
Taking logs of variables helps spread things out...

Looking at pairs of variables III



https://www.huffpost.com/entry/covid-19-population-density-myth_n_5ff8c68fc5b63642b6fba9eb

Linear Transformation of Data I

Sometimes we want to analyze data in different units

- Temperature: $\text{Celsius} = \frac{5}{9}(\text{Fahrenheit} - 32)$
- Curve: $\text{exam} = \text{score} + (0.25)(100 - \text{score})$

  This curve adds back 25% of exam points missed.
- Standardized Score: $z_i = \dfrac{x_i - \overline{x}}{s}$

Claim: All 3 are examples of linear transformations: $y = a + bx$

- Temperature: $\text{Celsius} = -\left(\dfrac{160}{9}\right) + \left(\dfrac{5}{9}\right)\text{Fahrenheit}$
- Curve: $\text{exam} = 25 + (0.75)\,\text{score}$
- Standardized Score: $z_i = -\left(\dfrac{\overline{x}}{s}\right) + \left(\dfrac{1}{s}\right)x_i$

Linear Transformation of Data II

Claim: If data $x_1, x_2, \ldots, x_n$
are linearly transformed to $y_i = a + bx_i$

Then, $\overline{y} = a + b\overline{x}$.

Claim: If data $x_1, x_2, \ldots, x_n$
are linearly transformed to $y_i = a + bx_i$

Then, $SD(y) = s_y = |b|s_x = |b|SD(x)$.

Proof: On your own for HW #2.

Why **random** samples I

1. Choose 10 representative words from the text that will be shown.
2. Record the length of each word â the number of letters in the word.
3. Record whether or not the word contains the letter e
4. Calculate the average word length of your 10 words.
5. Calculate the proportion of words containing an e

Text

Why **random** samples II

My personal sample of $n = 10$ words

```
mySample
```

```
 [1] "endure"  "have"    "which"   "testing" "world"
 [6] "we"      "perish"  "poor"    "never"   "detract"
```

The lengths of my $n = 10$ words:

```
 endure    have   which testing   world      we  perish
      6       4       5       7       5       2       6
   poor   never detract
      4       5       7
```

Why **random** samples III

Average length of my sample of $n = 10$ words:

```
myxbar <- mean(mySampleWordLen)
myxbar
```

```
[1] 5.1
```

$\bar{x} = 5.1$

Why **random** samples IV

How many of my words contain the letter e?

```
eWords <- mySample[grep("e", mySample)]
x=length(mySample[grep("e", mySample)])
x
```

```
[1] 7
```

What proportion of my words contain the letter e?

```
myphat <- x / n
myphat
```

```
[1] 0.7
```

$\hat{p} = 0.7$

Why **random** samples V

Student sample averages ($\overline{x}$'s):

How many students? That is, how many sample averages ($\overline{x}$'s) ?

```
[1] 70
```

Why **random** samples VI

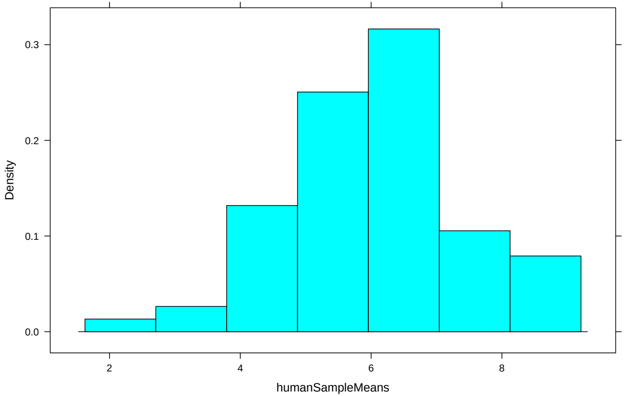
```
stem(humanSampleMeans, scale=2)


  The decimal point is at the |

  2 | 2
  2 | 9
  3 | 2
  3 | 9
  4 | 1444
  4 | 668889
  5 | 11112344
  5 | 5556678899
  6 | 011122333444
  6 | 667778899
  7 | 000112234
  7 | 66
  8 | 22344
  8 | 7
```

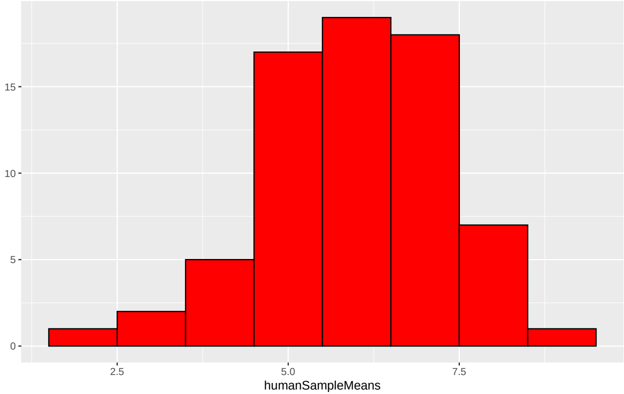
Why **random** samples VII

```
histogram(humanSampleMeans)
```



Why **random** samples VIII

What is the mean length ($\overline{x}$) "on average" for students' 70 samples?



Why **random** samples IX

```
mean(humanSampleMeans)
```

```
[1] 6.024
```

```
worddata <- read.csv("./address.csv")
```

What is stored in the data we just read in?

How many words are in the Gettysburg Address?

```
glimpse(worddata)
```

```
Rows: 268
Columns: 4
$ id       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,...
$ word     <fct> Four, score, and, seven, years, ago, o...
$ wordlen  <int> 4, 5, 3, 5, 5, 3, 3, 7, 7, 5, 4, 4, 9,...
$ containE <fct> No, Yes, No, Yes, Yes, No, No, Yes, No...
```

Why **random** samples X

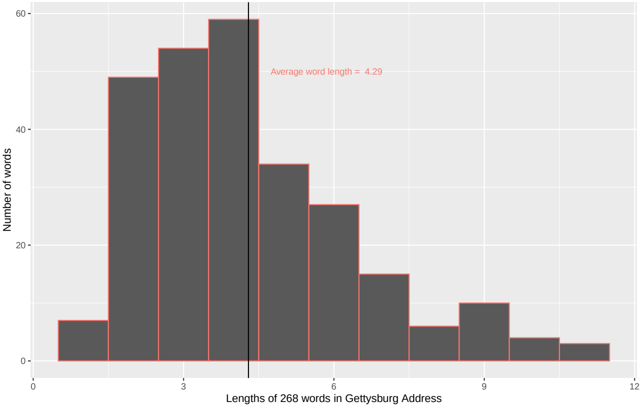
What is actual average length of all 268 words in the Address?
The **Population** mean (μ)

```
mean(worddata$wordlen)
```

```
[1] 4.295
```

Why **random** samples XI

A histogram of the lengths of all 268 words



Why **random** samples XII

Let's let R randomly sample $n_2 = 5$ words from the Gettysburg Address and record their average length ($\overline{x}$).

Repeat this 500 times.

Will all of the 500 sample averages be the same?

Why **random** samples XIII

To get started, look at a couple of samples and their means

```
sample1 <- sample(1:268, 5)
sample1
```

```
[1] 214 202 105  91  96
```

```
as.character(word[sample1])
```

```
[1] "cause" "us"    "a"     "live"  "and"
```

```
wordlen[sample1]
```

```
[1] 5 2 1 4 3
```

```
mean(wordlen[sample1])
```

```
[1] 3
```

Why **random** samples XIV

```
sample2 <- sample(1:268, 5);   sample2

[1]  54 143  26 262  45

as.character(word[sample2])

[1] "endure" "little" "all"    "people" "any"

wordlen[sample2]

[1] 6 6 3 6 3

mean(wordlen[sample2])

[1] 4.8
```

Why **random** samples XV

```
mean(wordlen[sample1])
```

```
[1] 3
```

```
mean(wordlen[sample2])
```

```
[1] 4.8
```

```
mu
```

```
[1] 4.295
```

Why **random** samples XVI

Now, let's repeat the random sampling a few times

```
replicate(10, wordlen[sample(1:268, 5)] )
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    4    3    5    4    6    1    4    6    9     4
[2,]    3    4    7    2    4    2    5    4    6     9
[3,]    4    2    2    2    2    3    2    6    3     4
[4,]    4    2    4    9    2   11    4    2    5     3
[5,]    3    6    2    4    4    4    4    1    8     6
```

```
replicate(10, mean(wordlen[sample(1:268, 5)]) )
```

```
 [1] 3.4 4.2 4.2 3.8 5.2 5.0 4.2 4.2 4.8 4.4
```

Let's repeat the random sampling 500 times

Why **random** samples XVII

```
randomSampleMeans = replicate(500, mean(wordlen[sample(1:268,5)]
sort(randomSampleMeans)[1:20]
```

```
 [1] 2.2 2.2 2.4 2.4 2.4 2.6 2.6 2.6 2.6 2.6 2.6 2.6 2.6 2.8
[15] 2.8 2.8 2.8 2.8 2.8 2.8
```

```
mu
```

```
[1] 4.295
```

```
sort(randomSampleMeans)[1:20] - mu
```

```
 [1] -2.095 -2.095 -1.895 -1.895 -1.895 -1.695 -1.695 -1.695
 [9] -1.695 -1.695 -1.695 -1.695 -1.695 -1.495 -1.495 -1.495
[17] -1.495 -1.495 -1.495 -1.495
```

Why **random** samples XVIII

What is the average length ($\overline{x}$) "on average" for many, many ($M = 500$) samples each with $n_2 = 5$ randomly chosen words?

```
mean(randomSampleMeans)
```

```
[1] 4.289
```

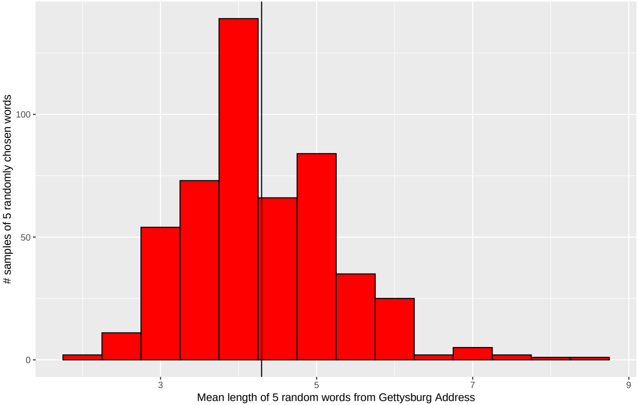
If this "mean of the averages" (or "mean of the means") is close to the true mean we say that the statistic ($\overline{x}$) is an **unbiased** statistic (estimator) for the parameter (μ).

```
mu
```

```
[1] 4.295
```

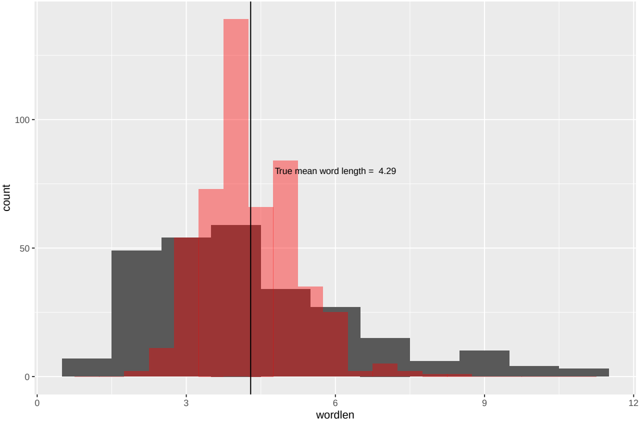
Why **random** samples XIX

Histogram of the average lengths ($n_2 = 5$)



Why **random** samples XX

How does the original population of word lengths compare with the $M = 500$ average lengths ($\overline{x}$'s) of $n_2 = 5$ randomly chosen words?



Why **random** samples XXI

Let's randomly sample $n = 15$ words instead of 5.

Let's repeat the random sampling 500 times.

```
randomSampleMeans.15 = replicate(500, mean(wordlen[sample(1:268,
sort(randomSampleMeans.15[1:20])
```

```
 [1] 3.600 3.800 3.800 3.867 3.867 3.933 3.933 4.067 4.067
[10] 4.133 4.267 4.267 4.333 4.533 4.533 4.533 4.533 4.600
[19] 5.133 5.400
```

```
mu
```

```
[1] 4.295
```

```
sort(randomSampleMeans.15[1:20] - mu)
```

```
 [1] -0.69478 -0.49478 -0.49478 -0.42811 -0.42811 -0.36144
 [7] -0.36144 -0.22811 -0.22811 -0.16144 -0.02811 -0.02811
[13]  0.03856  0.23856  0.23856  0.23856  0.23856  0.30522
[19]  0.83856  1.10522
```

Why **random** samples XXII

What is the mean length "on average" for many, many ($N = 500$) samples of $n = 15$ randomly chosen words?

```
mean(randomSampleMeans.15)
```

```
[1] 4.308
```

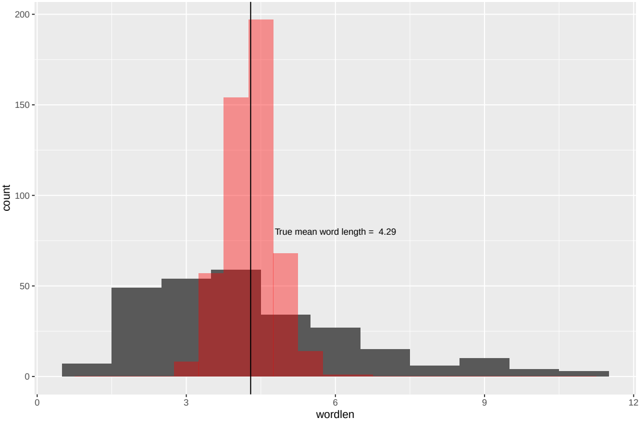
If this "mean of the averages" (or "mean of the means") is close to the true mean we say that the statistic ($\overline{x}$) is an **unbiased** statistic (estimator) for the parameter (μ).

```
mu
```

```
[1] 4.295
```

Why **random** samples XXIII

How does the original population of word lengths compare with the $M = 500$ mean lengths of $n = 15$ randomly chosen words?



Why **random** samples XXIV

Conclusion:

Word samples obtained by students were biased - did not provide a good estimator.

Word samples obtained from random sampling were unbiased - the average of the sample averages was very close to the population mean μ and the spread of sample averages of size 15 was quite small.

To understand the behavior of these random sample averages and how their ‘spread’ depends on the sample size we need to study Probability...