

EDP308: STATISTICAL LITERACY

The University of Texas at Austin, Fall 2020

RAZ: Rebecca A. Zárate, MA

Overview

- Measures of Spread (Variation)
 - ▣ Interquartile Range (IQR)
 - ▣ Variance
 - ▣ Standard Deviation
- Calculating Measures of Variation
 - ▣ Population
 - Variance (σ^2) and Standard Deviation (σ)
 - ▣ Sample
 - Variance (s^2) and Standard Deviation (s)
- Variation in R

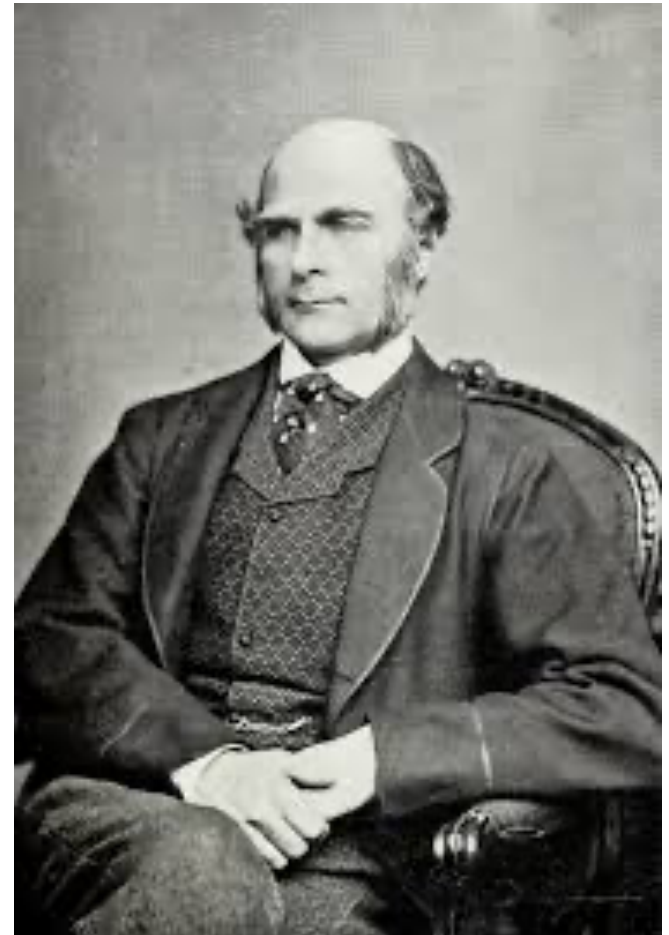
Measures of Spread (Variation)

What is variation?

- Variation is how much things differ from one another.
 - ▣ If a particular variable has a lot of variability then the observations (ex. people) are going to differ from each other quite a bit.
 - ▣ Data with low variability (i.e. observations are very similar) will cluster while data with high variability will be all over the place.
- How much observations (ex. people) differ from each other is of great interest and value in statistics.

Historical Moment: Sir Francis Galton

- **Sir Francis Galton** (Feb. 16, 1822 – Jan. 17, 1911)
 - ▣ Cousin to Charles Darwin
- Child prodigy. Adult statistician, psychology, psychometrician, anthropologist, among other things.
- Was fascinated by genetics and the idea that genius may be hereditary.
- This interest got him into measuring anything and everything in humans
 - ▣ Height, weight, IQ, birth order, race, fingerprints, and much more
- Founded Psychometrics
 - ▣ The measure of human traits like personality and intelligence



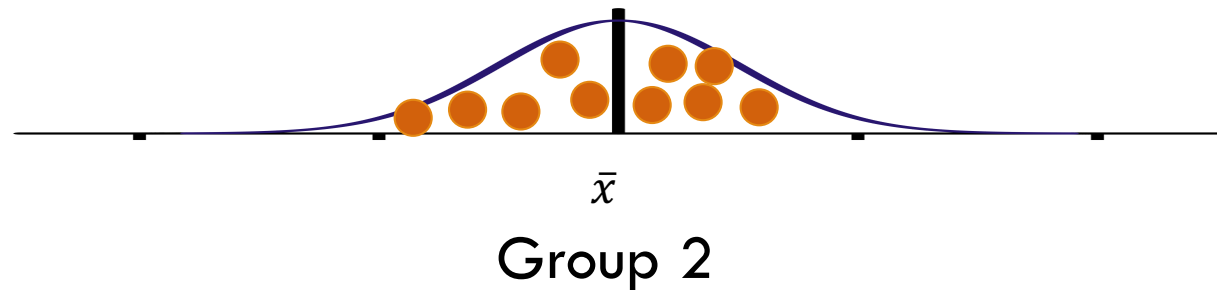
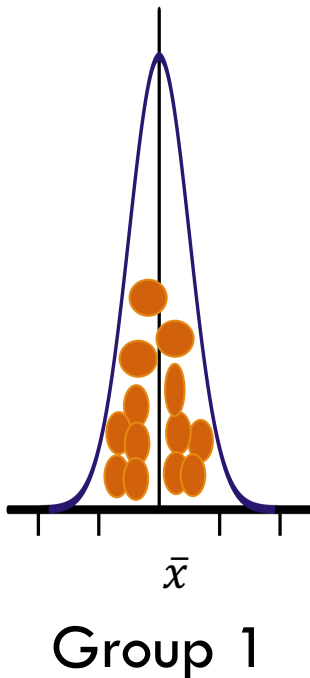
Historical Moment: Sir Francis Galton

- **Sir Francis Galton** (Feb. 16, 1822 – Jan. 17, 1911)
- Unfortunately, it was his interest with measurement and differences that lead him to a new field of study...
- Eugenics...
 - ▣ Improving the human race by encouraging “good stock” to reproduce, while discouraging undesirables from mating...
 - ▣ Eventually, the Nazi’s got their hands on this and took it to be scientific evidence for their belief.
- An example of people misusing statistics to push their own agenda.



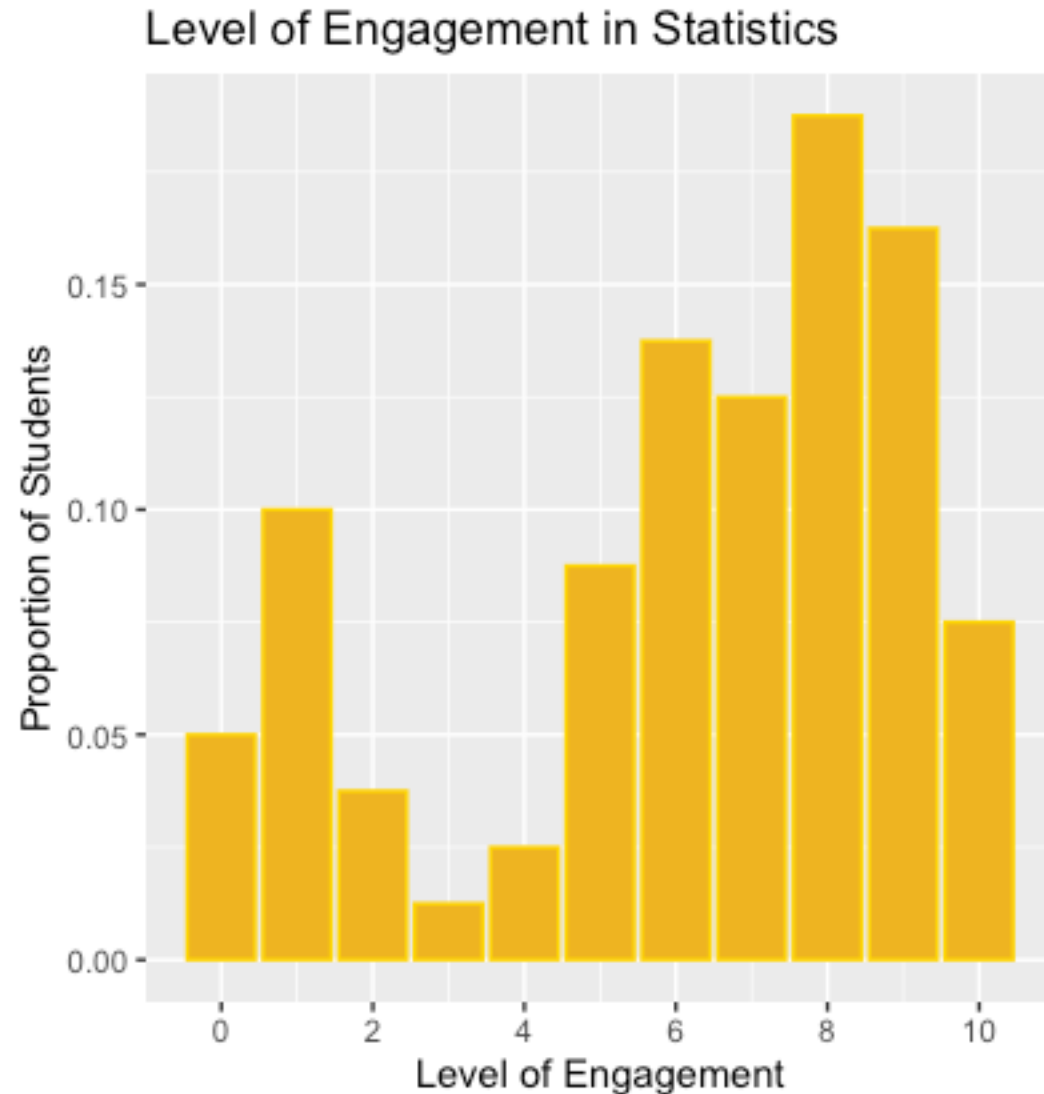
Variation

Judging purely visually, which group do you think has higher variability?



High Variability

- The average level of engagement in this class is about 6.
- But notice the difference are large. Some students are very engaged, some none at all.
- Here, students vary a lot in how much they are engaged with the class.

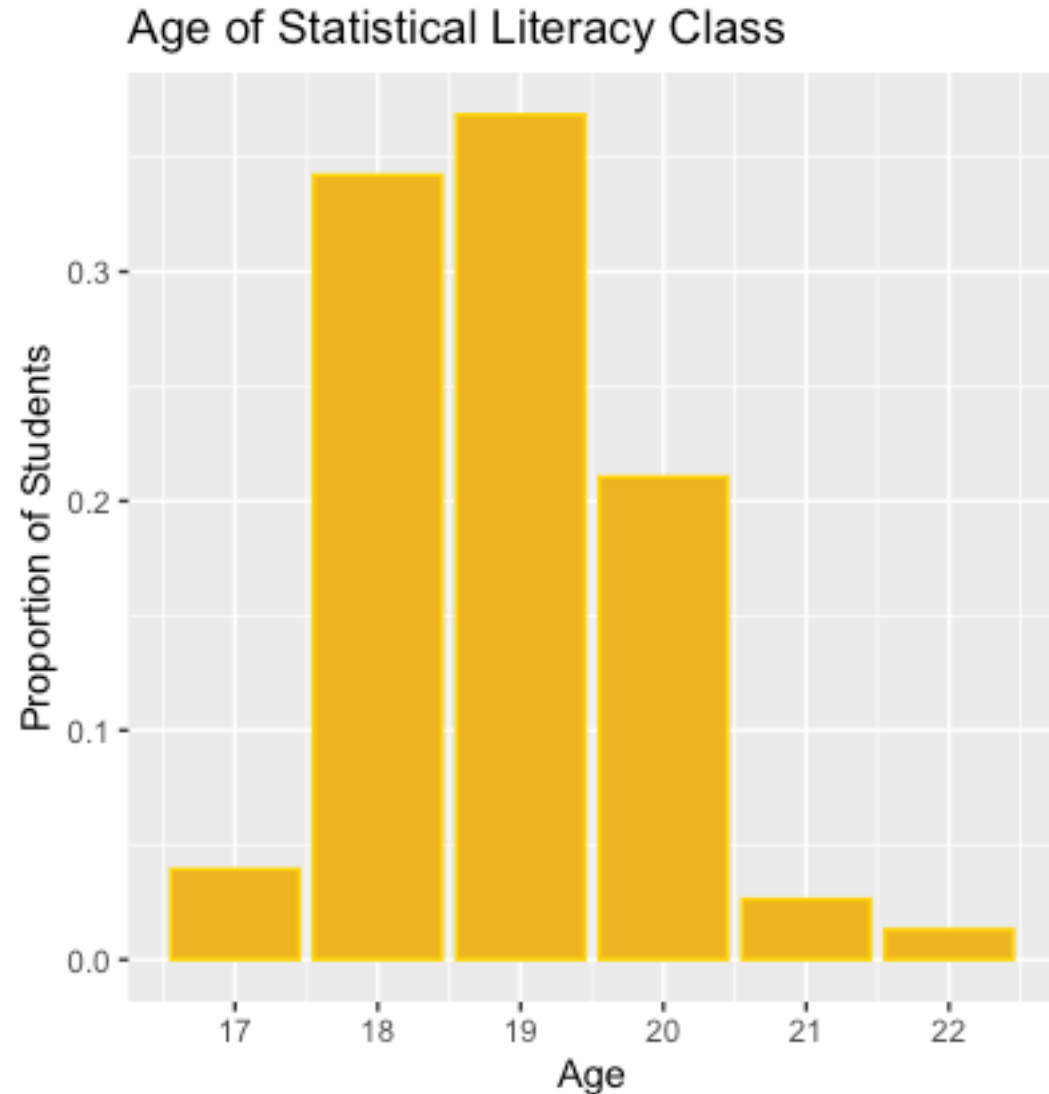


Low Variability

The average age of students in this class is about 19.

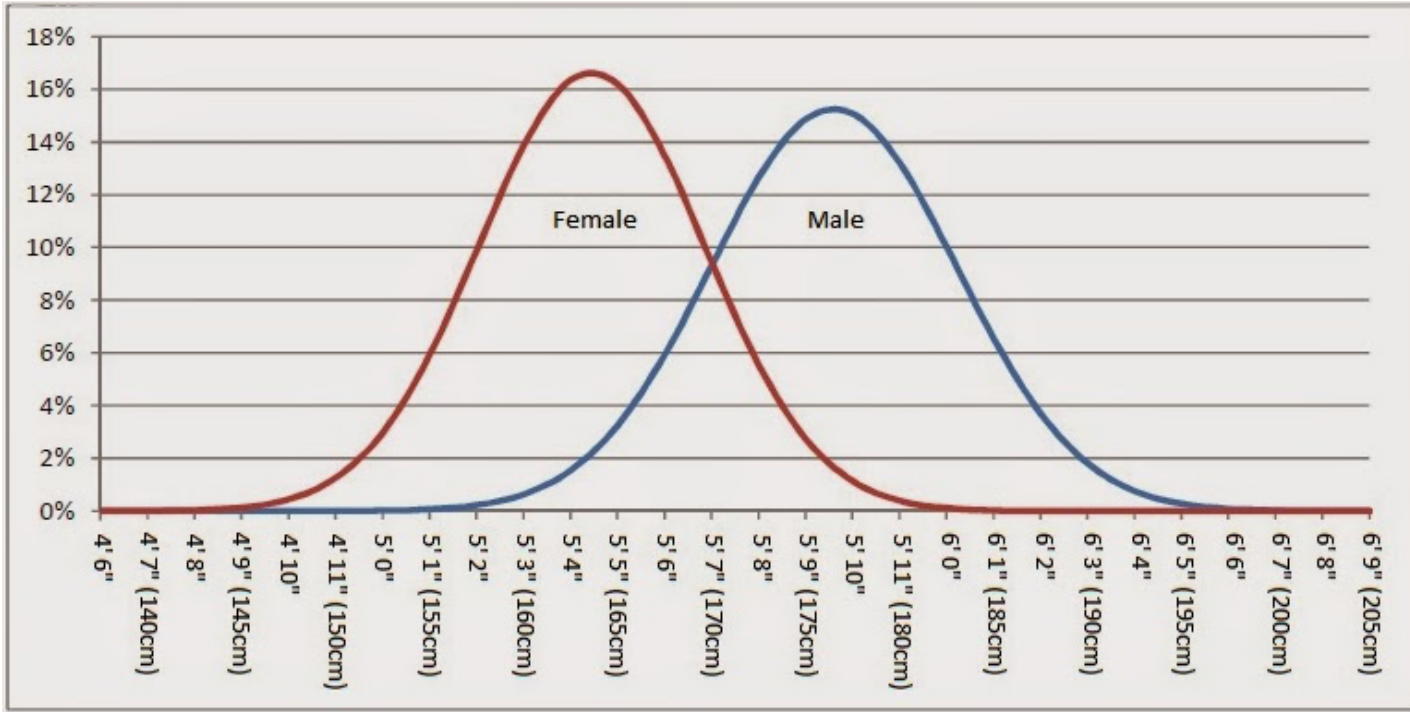
Notice how here most people are right next to that 19 years.

There is low variability in age in the statistical literacy course.



Height by Gender

Probability of Seeing Something X Tall



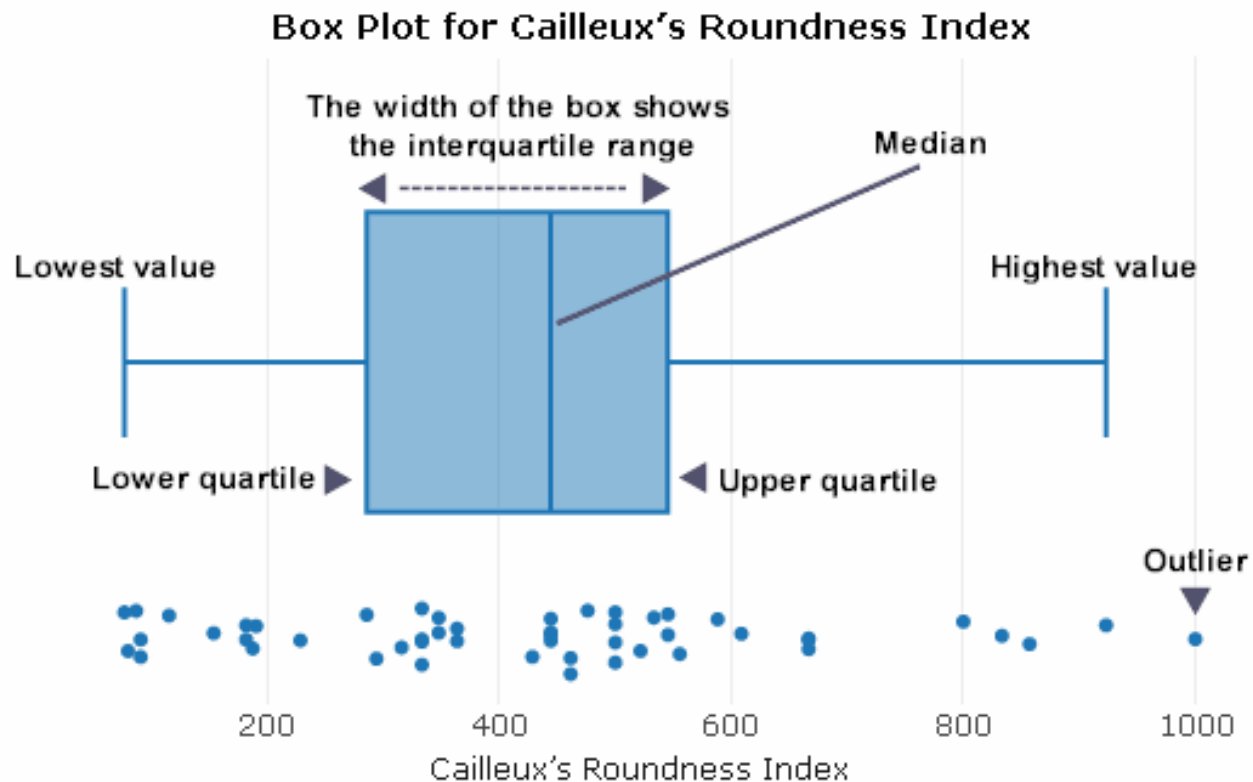
Which sex has more variability?

Measures of Variability

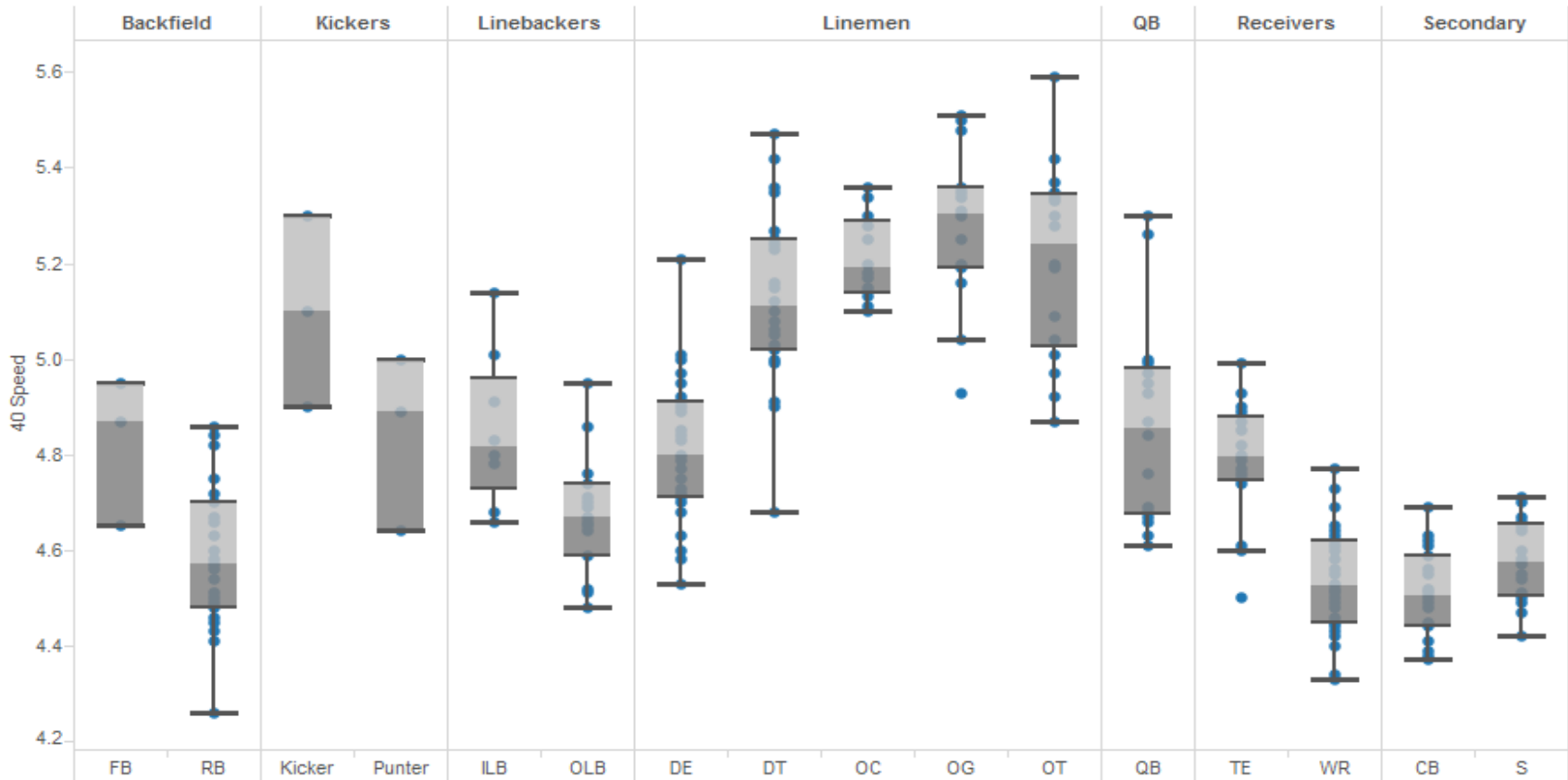
- A few different methods:
 - Range
 - The scores were between 75 to 100
 - $100 - 75 = 25$, a range of 25 points
 - Interquartile Range
 - The scores from the 25th% to the 75th%
 - The middle 50%
 - Variance
 - Standard Deviation

Interquartile Range (IQR)

- The IQR is the box portion in Box-and-Whisker plots. It tells you where the middle 50% of the data is.



Running the 40 Yard Dash



Which position have some of the highest variability? Lowest?

Measures of Variability

- In the last few examples, we saw how the level of engagement in this class varies a lot more than the average age in this class. Men vary more than women in their heights. And linemen have the most variability in the running the 40-yard dash. We could see that visually in the graphs but we can also quantify it.

How do we quantify/figure out the “average amount of variation” in our data?

Deep breath... Here comes the math...

Variance and SD for POPULATIONS

Deviation Example for Population

| X (unit) | - μ (mean) = | Difference (Deviation) |
|----------|------------------|------------------------|
| 8 - | | |
| 7 - | | |
| 5 - | | |
| 6 - | | |
| 10 - | | |
| 9 - | | |
| 7 - | | |
| 9 - | | |
| 8 - | | |
| 11 - | | |
| n = 10 | | |

Here are some quiz grades out of an 11 point scale.

1. Calculate the mean.

Deviation Example for Population

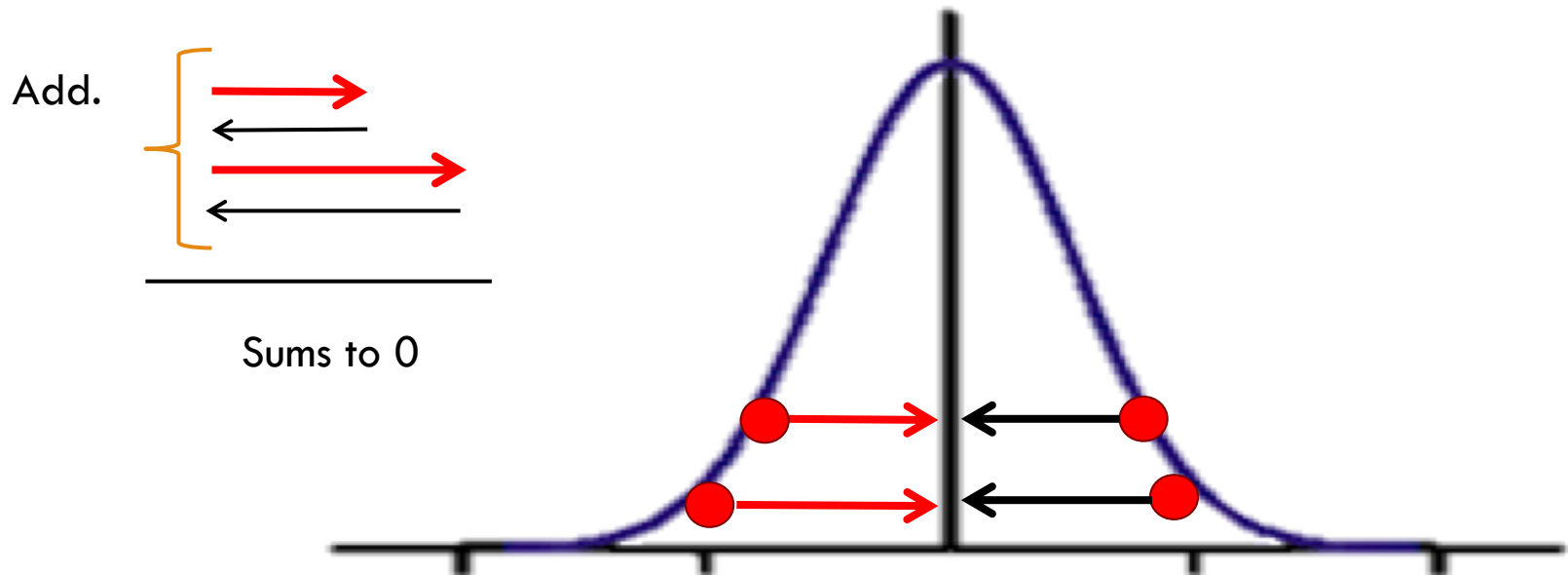
| X (unit) | - μ (mean) = | Deviation |
|---------------|------------------|--------------|
| 8 - | 8 = | 0 |
| 7 - | 8 = | -1 |
| 5 - | 8 = | -3 |
| 6 - | 8 = | -2 |
| 10 - | 8 = | 2 |
| 9 - | 8 = | 1 |
| 7 - | 8 = | -1 |
| 9 - | 8 = | 1 |
| 8 - | 8 = | 0 |
| 11 - | 8 = | 3 |
| $\Sigma = 80$ | $\mu = 8$ | $\Sigma = ?$ |

What is the sum of the deviations?

Sums to Zero

- The sum of the deviations will always be 0.

What do we do to make use out of the deviations?



Sum of Deviations = ZERO

- And we encounter our first road block...
- If we add all the deviations, they will sum to zero.
 - ▣ Why?
- Because there are an equal number of scores below the mean and above, so they cancel out and sum to zero, which does not help us...

What do we need to do to our
precious deviations?

Making Use of Our Deviations

- Think back to your math classes... We want to keep the magnitude of these numbers. These differences are important, but as you see we can't add them.
- So, we must... SQUARE them.
 - ▣ Because any negative number squared turns positive
 - Why not absolute value? It doesn't play as nice when it comes to math-ing things... It's hard to keep track and undo.

Deviation Example for Population

| X (unit) | - μ (mean) = | Deviation | Deviation Squared |
|---------------|------------------|-----------|-------------------|
| 8 - | 8 = | 0^2 | 0 |
| 7 - | 8 = | -1^2 | 1 |
| 5 - | 8 = | -3^2 | 9 |
| 6 - | 8 = | -2^2 | 4 |
| 10 - | 8 = | 2^2 | 4 |
| 9 - | 8 = | 1^2 | 1 |
| 7 - | 8 = | -1^2 | 1 |
| 9 - | 8 = | 1^2 | 1 |
| 8 - | 8 = | 0^2 | 0 |
| 11 - | 8 = | 3^2 | 9 |
| $\Sigma = 80$ | $\mu = 8$ | | $\Sigma = 30$ |

“Squares”

Sum the “Squares”

Sum of Squares (SS)

- We have arrived at our “Sum of Squares”!
- This is our first “number translation” we do to figure out the average about of deviation

$$SS = \sum (x - \mu)^2$$

Sum of Squares is equal to the sum of each score minus the mean squared*

*You do the squaring to each (score – mean)²,
NOT just one time at the end.
Remember PEMDAS!

From SS to Variance

- We have the sum of all our squared deviations...
now what?

What do we usually do to get the
average of something?

Variance for Population

- Just like when calculating the mean, we divide by the total number of people (observations, scores, etc.) to get an average difference value

$$\frac{SS}{N} = \text{VARIANCE} \\ (\sigma^2)$$

$$\frac{30}{10} = 3.00 \\ (\sigma^2)$$

Which Number Scale?

But... What “number language”
is this in?

Is it the scale of the original
scores?

Which Number Scale?

- But... What “number language” is this in? Is it the scale of the original scores?
 - ▣ No, we squared everything, remember? So we have an average deviation score that is in the “squared scale”, the “squared number language”

What do we need to do to put it back into the original scale we started with?

Standard Deviation for Population

- Square root that Variance to get us back into the original scale and to give us
- Standard Deviation,
 - ▣ How much scores, on average, differ (deviate) from the mean

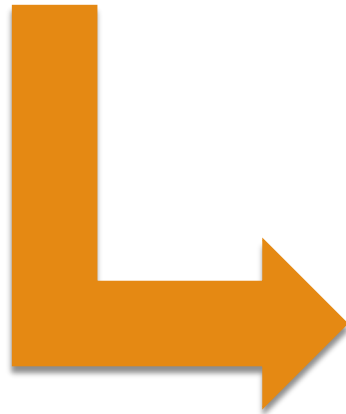
$$\sqrt{\text{VARIANCE } (\sigma^2)} = \text{STANDARD DEVIATION } (\sigma)$$

Standard Deviation for Population

VARIANCE

$$\frac{30}{10} = 3.00$$

(σ^2)



STANDARD
DEVIATION
 (σ)

“On average, the scores differ 1.73 points from the mean.”

$$\sqrt{3.00} = 1.73$$

Variance and SD for SAMPLES

Remember this?

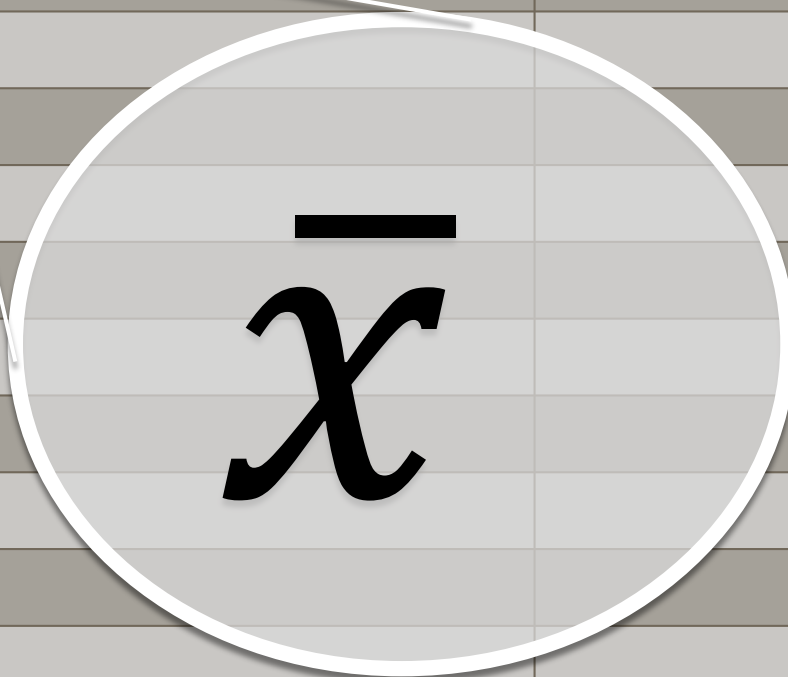
This is calculated a little different.

| Attribute | Population | Sample |
|---|---|---|
| <input type="checkbox"/> Includes | <input type="checkbox"/> Complete set | <input type="checkbox"/> Subset of population |
| <input type="checkbox"/> Mean | <input type="checkbox"/> μ ("mu") | <input type="checkbox"/> \bar{x} ("x bar") |
| <input type="checkbox"/> Sum of Squares | <input type="checkbox"/> SS ("Sum of Squares") | <input type="checkbox"/> SS ("Sum of Squares") |
| <input type="checkbox"/> Variance | <input type="checkbox"/> σ^2 ("sigma squared") | <input type="checkbox"/> s^2 ("variance") |
| <input type="checkbox"/> Standard Deviation | <input type="checkbox"/> σ ("sigma") | <input type="checkbox"/> s ("standard deviation") |
| <input type="checkbox"/> Size | <input type="checkbox"/> N | <input type="checkbox"/> n |
| <input type="checkbox"/> Numerical Descriptor | <input type="checkbox"/> "Parameter" | <input type="checkbox"/> "Statistic" |

Most things are very similar with some tweaking. We'll go through the same example as before but this time, we are treating it as a **SAMPLE** rather than the complete **POPULATION**. Keep an eye out for notational differences.

Deviation Example for Samples

| X (sample unit) | - \bar{x} (sample mean) = | Difference (Deviation) |
|-----------------|-----------------------------|------------------------|
| 8 - | | |
| 7 - | | |
| 5 - | | |
| 6 - | | |
| 10 - | | |
| 9 - | | |
| 7 - | | |
| 9 - | | |
| 8 - | | |
| 11 - | | |
| n = 10 | | |



1. Calculate the mean.

Deviation Example for Samples

| X (sample unit) | - \bar{x} (sample mean) = | Deviation | Deviation Squared |
|-----------------|-----------------------------|-----------------|-------------------|
| 8 - | 8 = | 0 ² | 0 |
| 7 - | 8 = | -1 ² | 1 |
| 5 - | 8 = | -3 ² | 9 |
| 6 - | 8 = | -2 ² | 4 |
| 10 - | 8 = | 2 ² | 4 |
| 9 - | 8 = | 1 ² | 1 |
| 7 - | 8 = | -1 ² | 1 |
| 9 - | 8 = | 1 ² | 1 |
| 8 - | 8 = | 0 ² | 0 |
| 11 - | 8 = | 3 ² | 9 |
| $\Sigma = 80$ | $\bar{x} = 8$ | $\Sigma =$ | $\Sigma = 30$ |

Mean, Deviation, and Sum of Squares are calculated the SAME for Populations and Samples

“Squares”

Sum the “Squares”

Sum of Squares (SS)

- We have
- This is our
- out the av

Sum of Squares are calculated the SAME for Populations and Samples, but the NOTATION is different!

$$SS = \sum (x - \bar{x})^2$$

Sum of Squares is equal to the sum of each score minus the mean squared*

*You do the squaring to each (score – mean), not just one time at the end.
Remember PEMDAS!

Variance for Samples

- Here is the major difference for calculating POPULATION vs. SAMPLE variance!

$$\frac{SS}{n-1} = \text{VARIANCE} \quad (s^2)$$

When calculating variance for a SAMPLE, you divide the SS by (n-1) rather than N.

$$\frac{30}{10-1} = 3.33\dots \quad (s^2)$$

Variance for Samples

- Here is the major difference for calculating
POPULATION VARIANCE vs. SAMPLE VARIANCE

Which will have more
(larger) variance?
A population or a
sample?

$$\frac{1}{n-1} (s^2)$$

Why $(n-1)$?

- “Statistics” (which come from SAMPLES, it would be a “parameter” if it was a POPULATION) inherently have **BIAS** because we do not sample the entire population
- Because you are dividing SS by $(n-1)$ rather than N , SAMPLE variance will always be GREATER than a population variance of the same data.

$(n-1)$

Tip: Pay extra attention to whether you are working with a population or a sample so you know which formula to use.

Standard Deviation for Sample

- Square root that Variance to get us back into the original scale and
- Standard Deviation, average, differ/

Variance to Standard Deviation is the same but different notation!

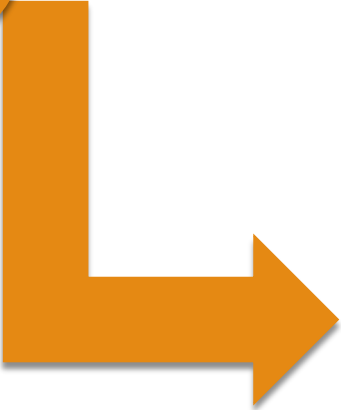
$$\sqrt{\text{VARIANCE } (s^2)} = \text{STANDARD DEVIATION } (s)$$

Standard Deviation for Sample

VARIANCE

$$\frac{30}{10-1} = 3.33\dots$$

(s^2)



STANDARD
DEVIATION

(s)

“On average, the scores differ 1.83 points from the mean.”

$$\sqrt{3.33\dots} = 1.83$$

Population vs. Sample

- Notice how in our first example, where we are working with a population, the variation ($\sigma = 1.73$) is **LESS** than the variation in the second example where we treat the data as if they were from a sample ($s = 1.83$).
- All else being equal, the variance of a Sample will always be greater than the variance of a Population

Population vs. Sample

- The main difference is when you divide the Sum of Squares by $(n-1)$ rather than N
- Other differences are notational
 - ▣ Population vs Sample
 - Size: N vs n
 - Mean: μ vs \bar{x}
 - Variance: σ^2 vs s^2
 - Standard Deviation: σ vs s

So how big is the average deviation?

How big is a standard deviation of 2?

Or 34.5? Or 4,927?

Or how would you know if a person's score of 47 was high, low, or average?

Up Next...

- To evaluate a standard deviation we need to know the original scale and the mean... Surely there is a better way...? What could we do to the data to make it is interpretable with out being tied to the original scale?
- Next we will learn about the wonders of **Standardization...** But first a little visual detour.

Variation in R

Variation in R

- It would be incredibly time consuming, not to mention error prone, to calculate these things by hand... No one calculates things by hand anymore because we have powerful tools like R, which is basically a giant calculator.

Variation in R

I used the same quiz grade data from our previous example. It is MUCH faster to do it in R than by hand on paper...

```
#####  
##### VARIATION #####  
#####  
  
##### POPULATION DATA #####  
quiz_grades <- c(8, 7, 5, 6, 10, 9, 7, 9, 8, 11)  
  
mean(quiz_grades) # mean = 8  
  
# to get the variance of a population, we need to do some tweeking  
# R's default is to calculate variance for a SAMPLE  
  
# first we need to know how many observations there are, we can use the "length()" function  
# which tells us how long an object is, i.e. how many row (how many observations)  
n <- length(quiz_grades) # 10 observations saved as object "n"  
  
# we can use the "var()" function to get the variance for a sample,  
# but to get variance for a population, we need to multiple by ((n-1)/n)  
var(quiz_grades) * ((n-1)/n) # variance = 3  
  
# to get the standard deviation, we can just take the squareroot of 3  
sqrt(3) # standard deviation = 1.73  
  
##### SAMPLE DATA #####  
  
# R makes this much easier...  
  
var(quiz_grades) # variance = 3.33...  
sd(quiz_grades) #standard deviation = 1.83
```