Measures of Variability

Lecture 4

Emma Ning, M.A.

Last Class

Central Tendency

Mean
Median
Mode

Distribution Shape & Modality

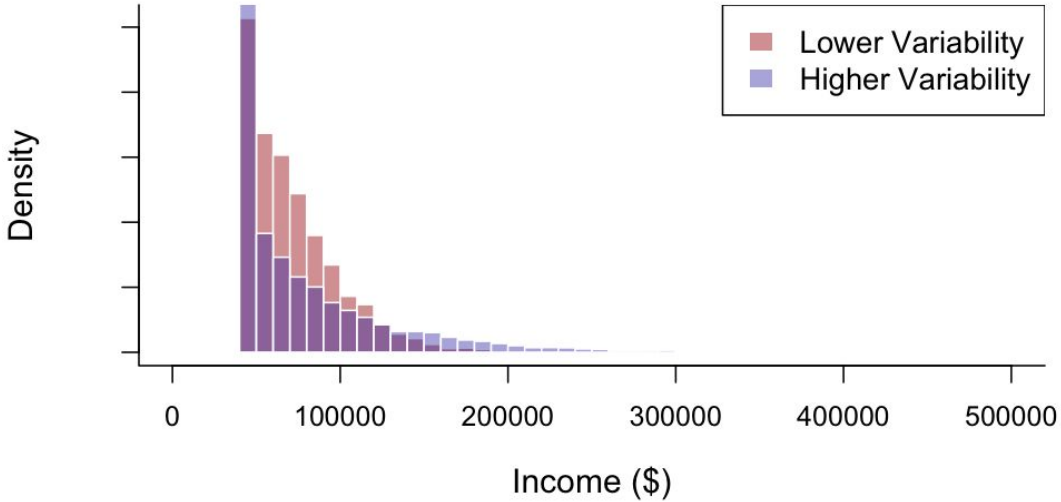
Skew
Modality

Picking the Best Summary for the Data

Apply what you learned

From our last lecture…

You are choosing between 2 companies: A  or B 



You don’t know your contract yet, which offer are you leaning toward?

TODAY'S PLAN

01

Measuring the "Spread" of Data

02

Variance & Standard Deviation

03

Interpreting Measures of Variability

04

Wrap Up

Learning objectives

- Explain why **variability** is important beyond central tendency, skewness and symmetry
- Understand and can explain the difference between **variance** and **standard deviation**
- Define and calculate **range**, **variance**, **standard deviation** for a population and/or a sample
- Explain **interquartile range**, and be able to interpret **box plots**
- Know and can explain the general idea of **degrees of freedom**

01

Measuring the “Spread” of Data

Two samples can have the same mean, but different spread

Sample 1

19 20 20 20 21

$M = \mathbf{20}$

Sample 2

5 10 15 25 45

$M = \mathbf{20}$

If we only look at the means, we might conclude that these groups are very similar, when in reality they are **not**.

Two samples can have the same mean, but different spread

Sample 1

19 20 20 20 21

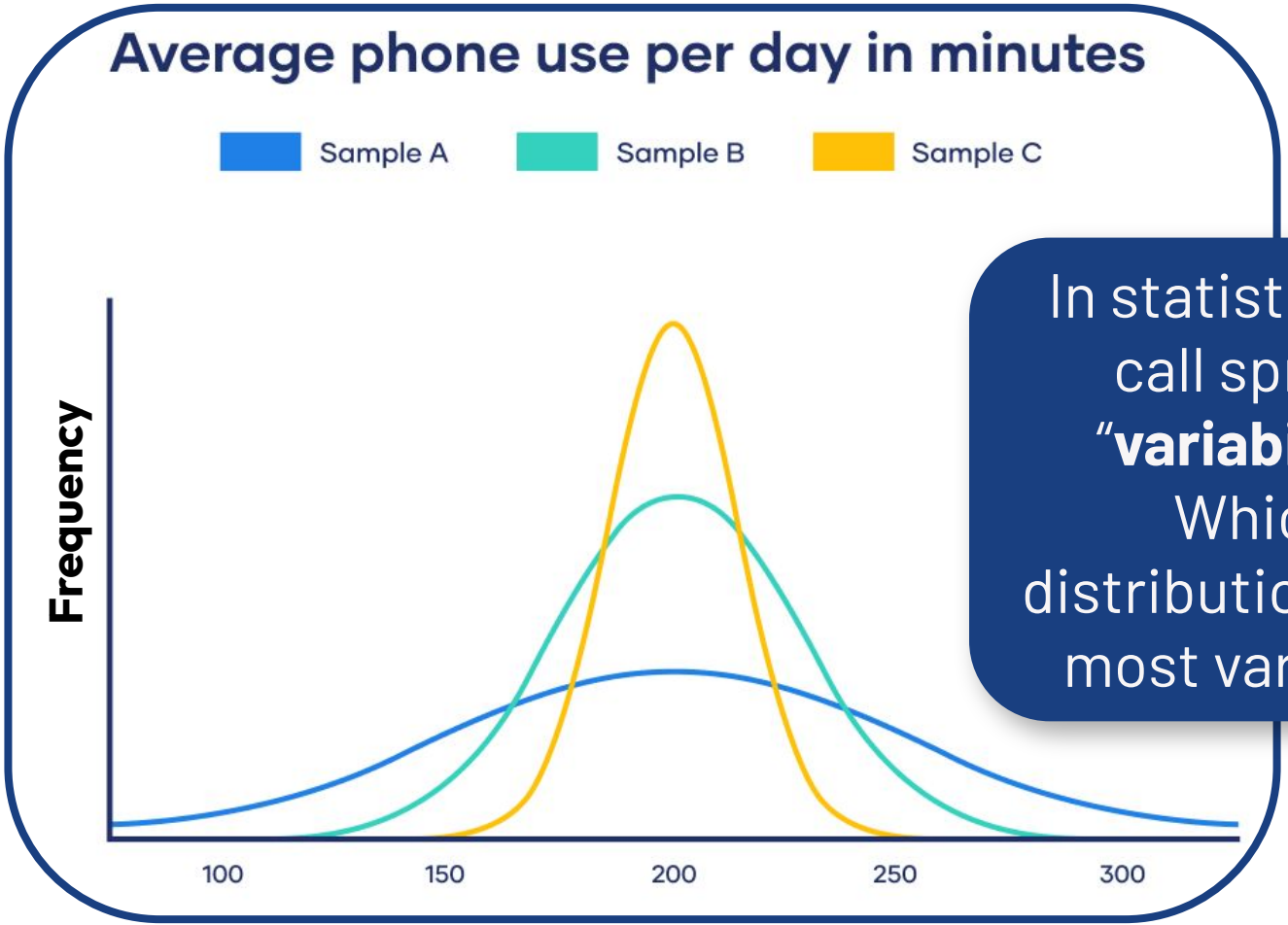
$M = \mathbf{20}$

Sample 2

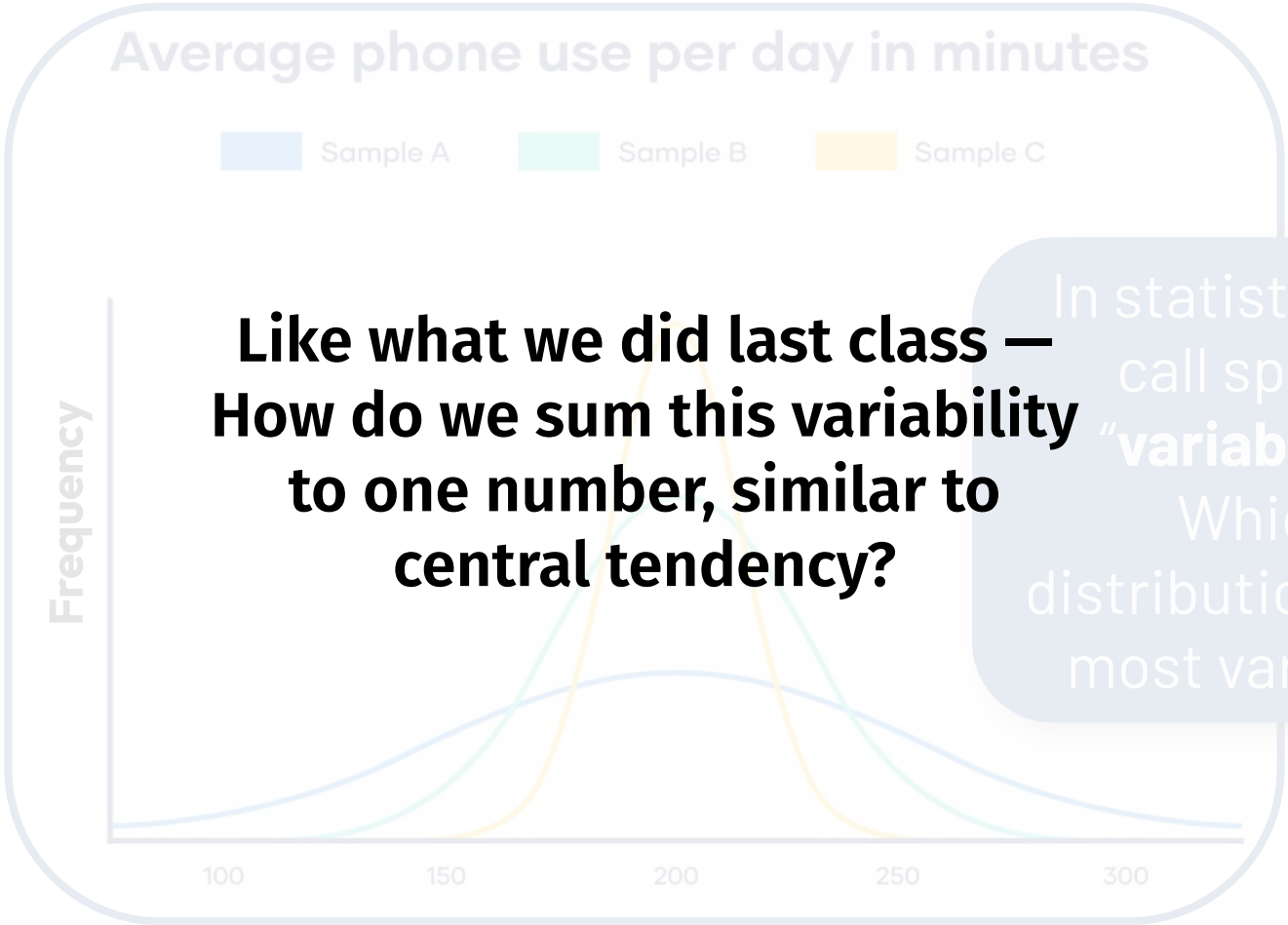
5 10 15 25 45

$M = \mathbf{20}$

The participant ages in Sample 1 are much closer together than those in Sample 2.



In statistics, we call spread "**variability**". Which distribution is the most variable?



Like what we did last class — How do we sum this variability to one number, similar to central tendency?

In statistics, we call spread "**variability**". Which distribution is the most variable?

Range

is simply the **largest** number subtracted from the **smallest** number

$$X_{max} - X_{min}$$

Let’s say I asked ten of you to tell me your age…

$$X_{max} - X_{min}$$

Let’s say I asked ten of you to tell me your age…

$$27 - 18 = \mathbf{9}$$

What problem do you see with range?

Let’s say I asked ten of you to tell me your age…

X_{min}

18 22 21 24 20 27 19

X_{max}

Hmm, but if we have an outlier, the range is determined entirely by that.

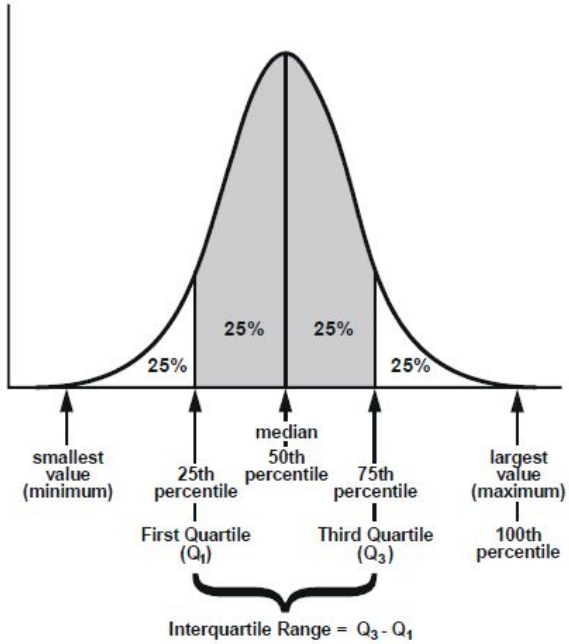
Therefore it does not give us the entire picture.

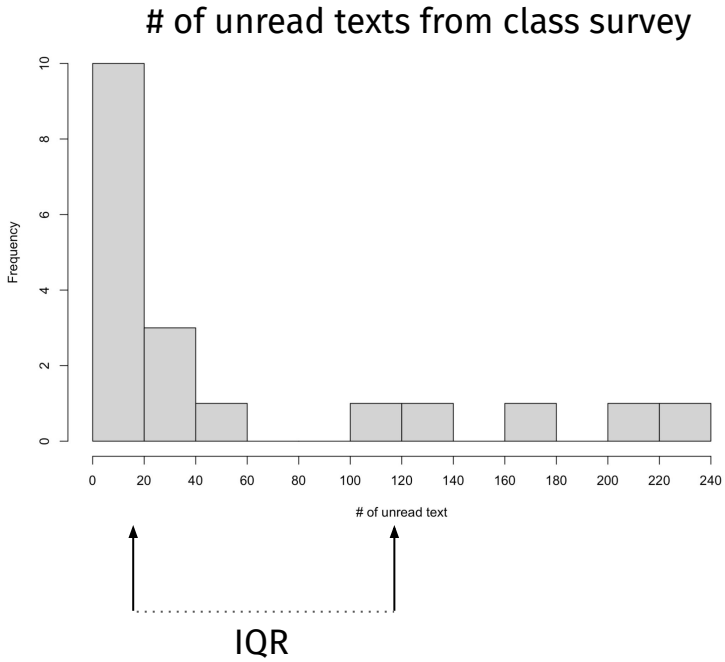
$27 - 18 = 9$

What problem do you see with range?

Interquartile Range (IQR)

The interquartile range (IQR) gives us a sense of how spread out the *middle* half of the data are.

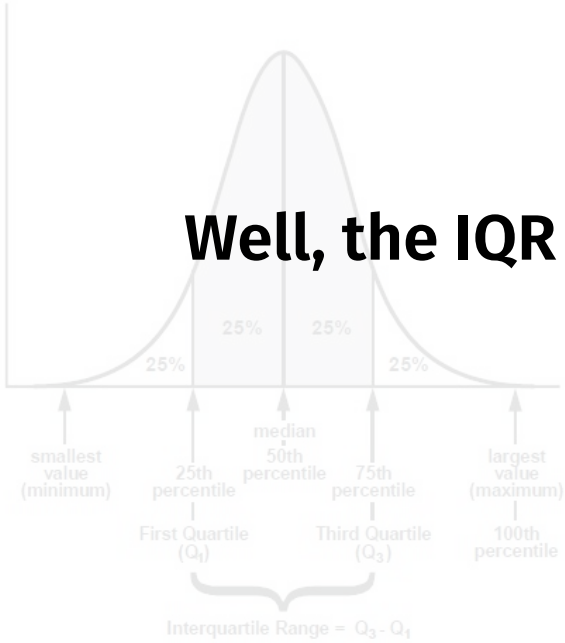




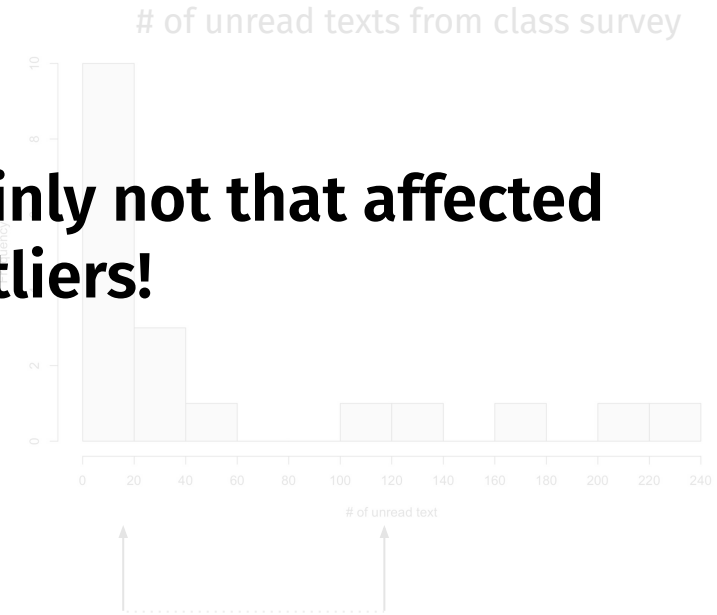
Note: Q1 = 25th percentile; Q2 = 50th percentile (median); Q3 = 75th percentile (I will not test you on this difference).

Interquartile Range (IQR)

The interquartile range (IQR) gives us a sense of how spread out the *middle* half of the data are.

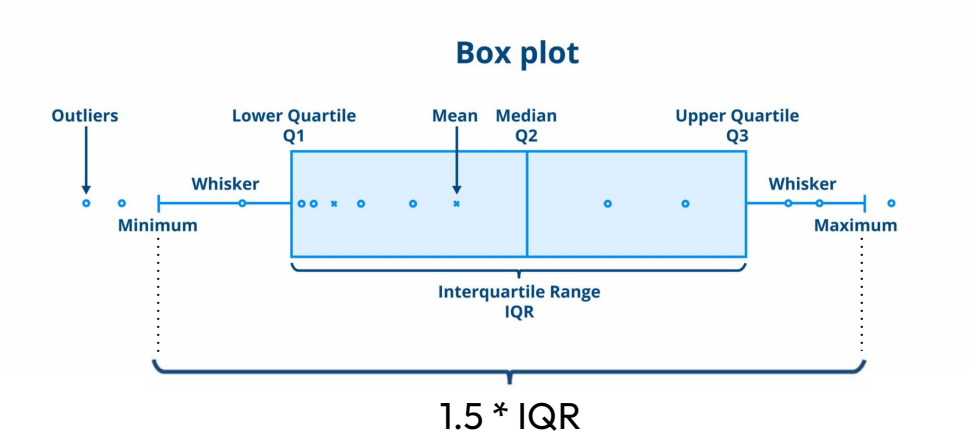


Well, the IQR is certainly not that affected by outliers!



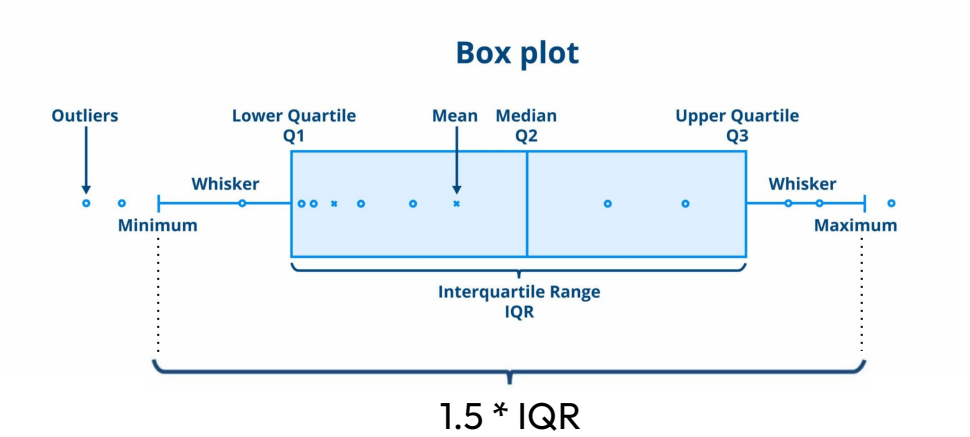
Note: Q1 = 25th percentile; Q2 = 50th percentile (median); Q3 = 75th percentile (I will not test you on this difference).

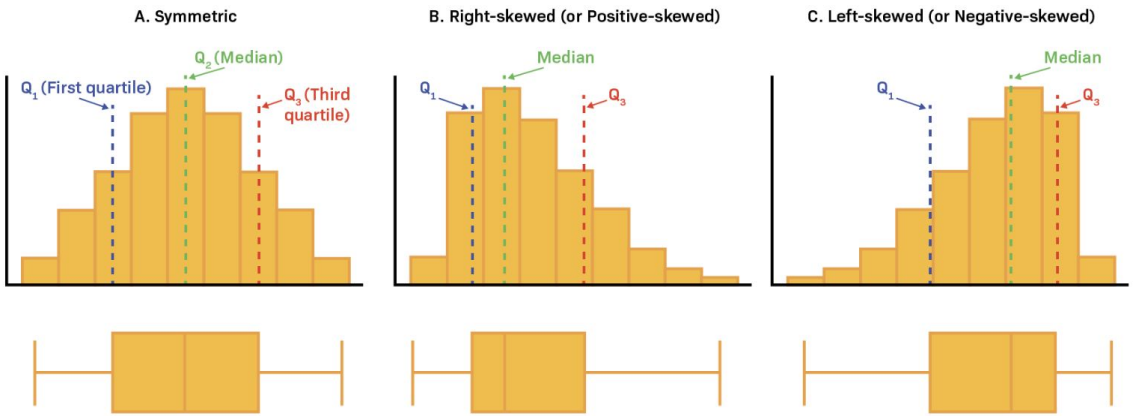
Box Plot



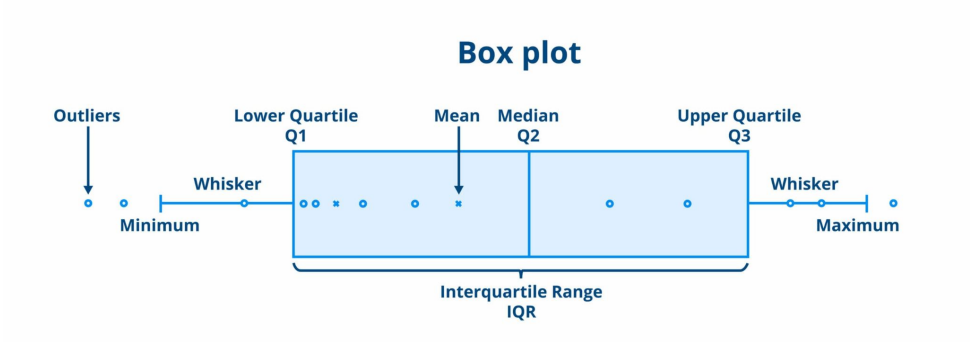
How will the box look like if our data is skewed?

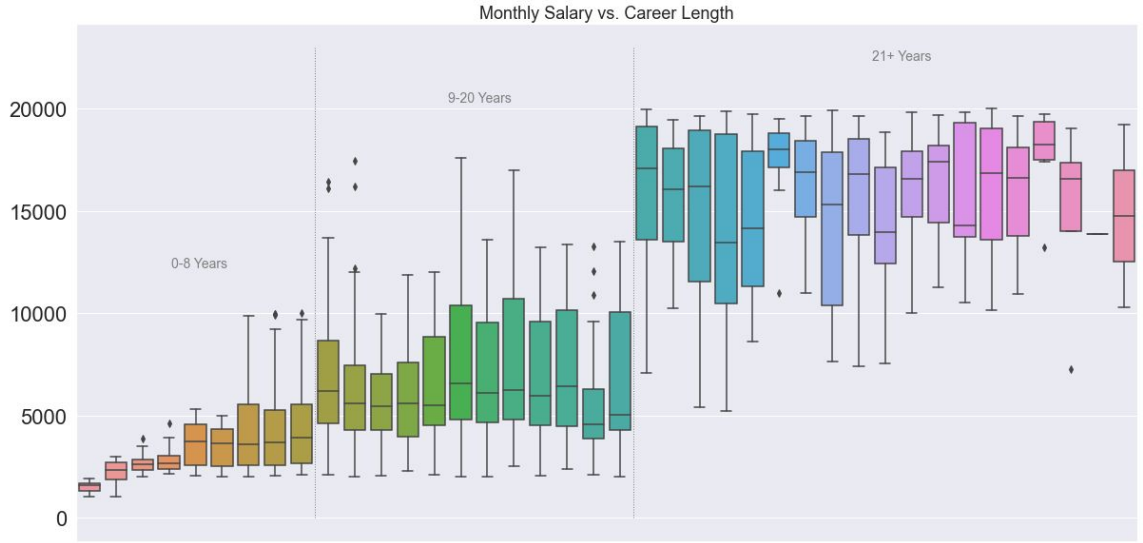
Box Plot


Box plot
Outliers
Lower Quartile
Q1
Mean
Median
Q2
Upper Quartile
Q3
Whisker
Whisker
Minimum
Maximum
Interquartile Range
IQR
1.5 * IQR

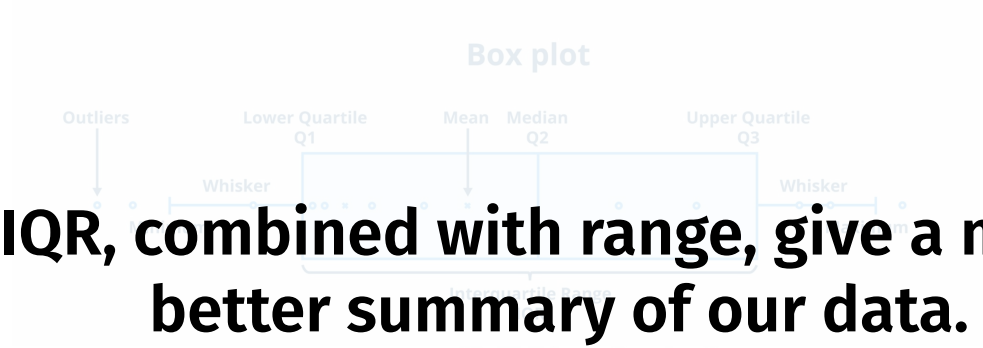

A. Symmetric
B. Right-skewed (or Positive-skewed)
C. Left-skewed (or Negative-skewed)
Q2 (Median)
Q1 (First quartile)
Q3 (Third quartile)
Median
Q1
Q3
Median
Q1
Q3

Box Plot


Box plot
Outliers
Lower Quartile
Q1
Mean
Median
Q2
Upper Quartile
Q3
Whisker
Whisker
Minimum
Maximum
Interquartile Range
IQR

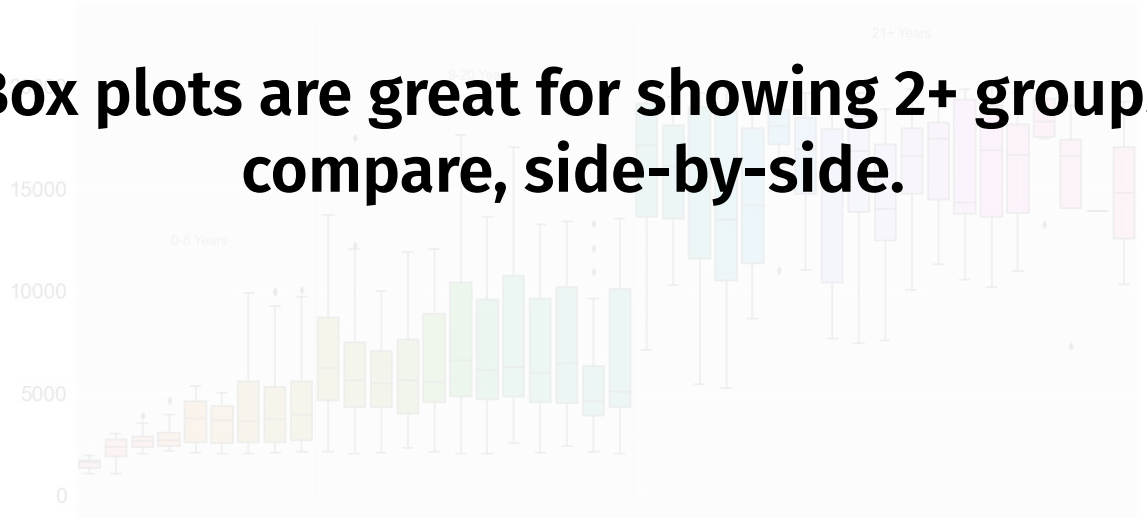

Monthly Salary vs. Career Length
21+ Years
9-20 Years
0-8 Years
20000
15000
10000
5000
0

Box Plot



IQR, combined with range, give a much better summary of our data.

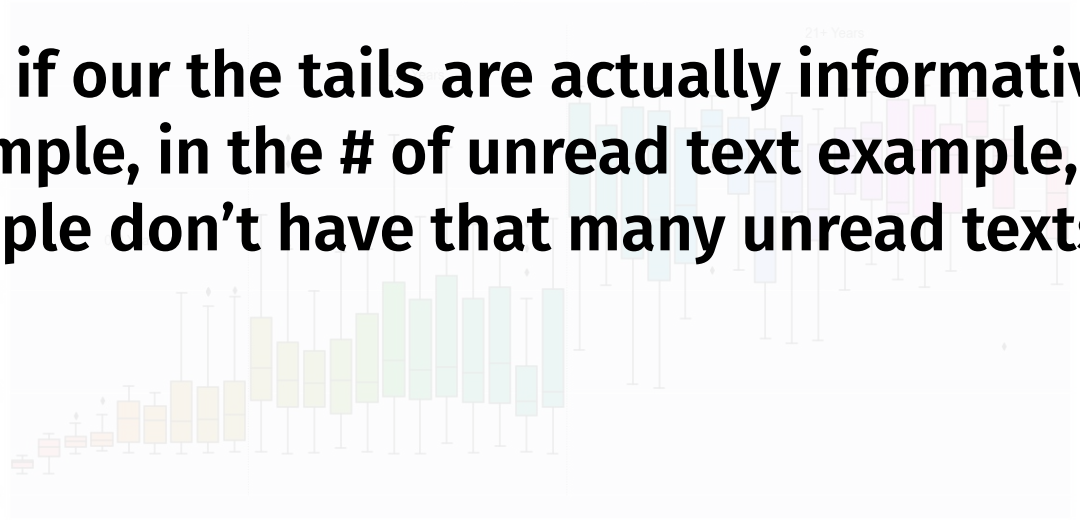
Box plots are great for showing 2+ groups compare, side-by-side.



Box Plot

However, IQR after all ignores the “tail” ends of the data.

What if our the tails are actually informative? For example, in the # of unread text example, most people don’t have that many unread texts.



02

Variance & Standard Deviation

“Deviation” Scores

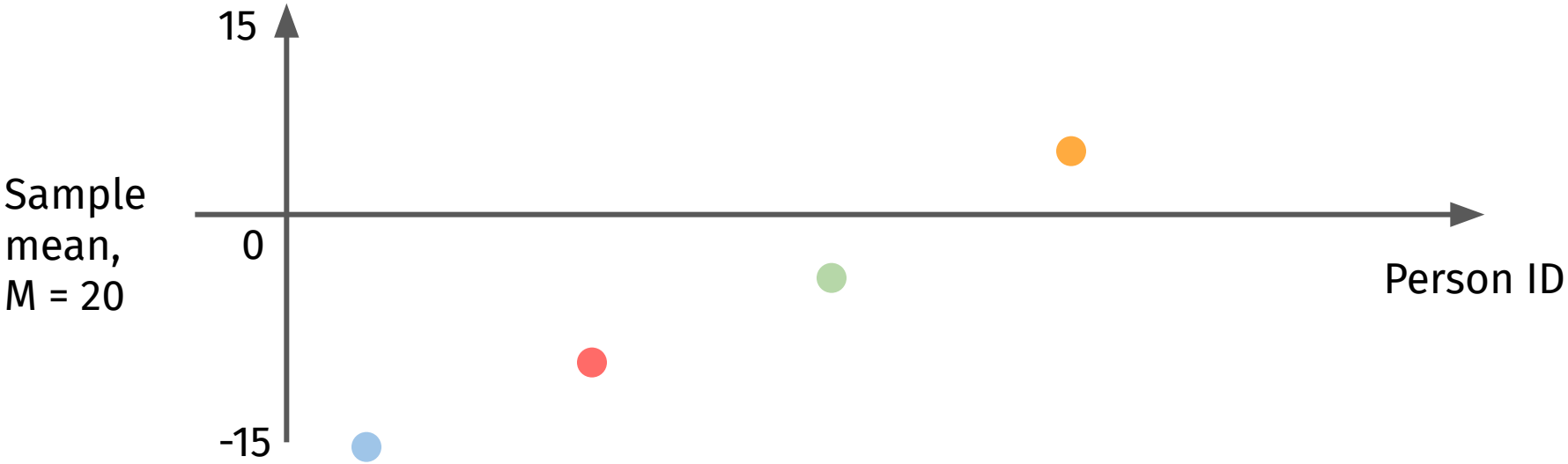
$M = 20$

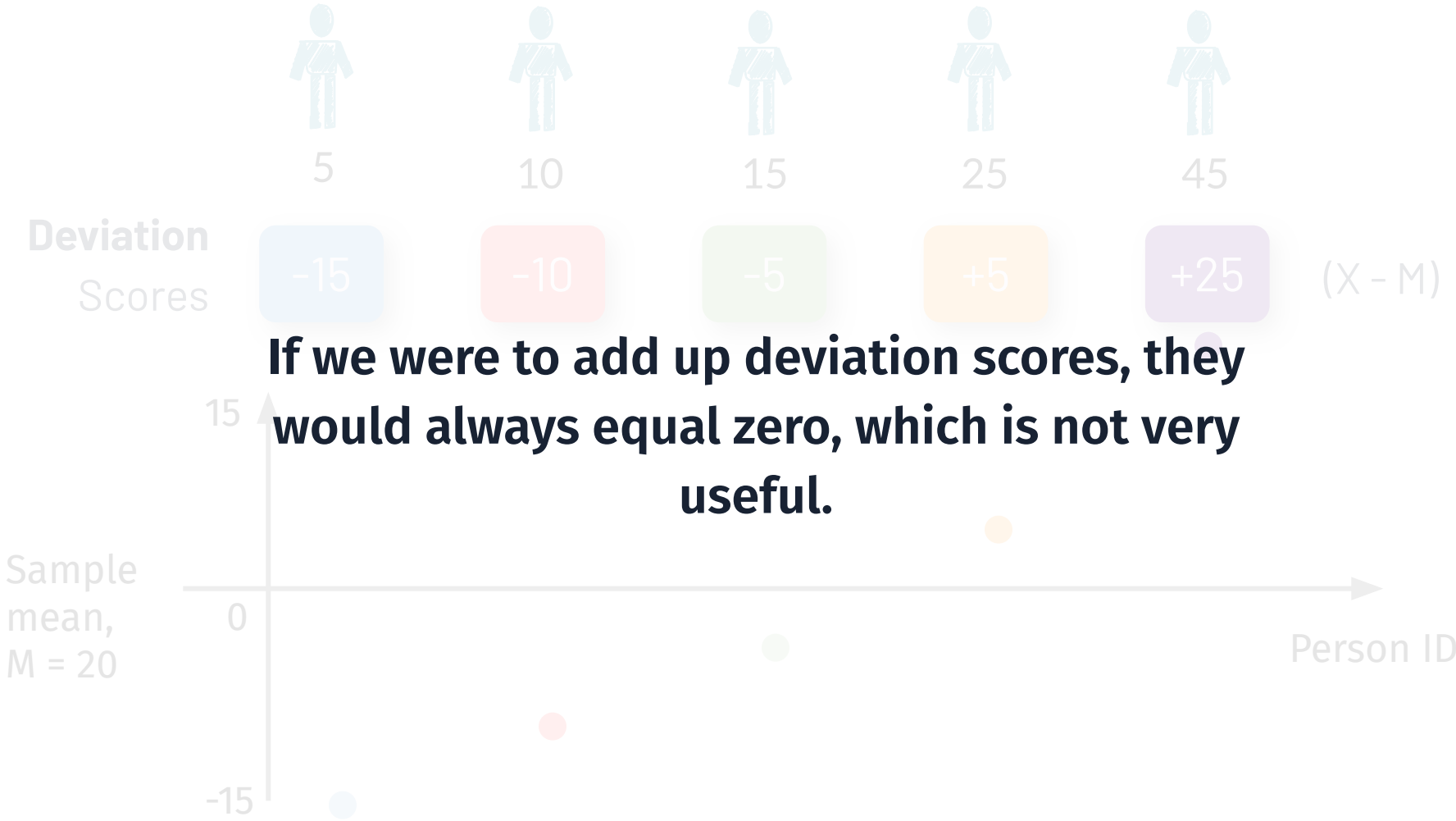
	5	10	15	25	45	
Deviation Scores	-15	-10	-5	+5	+25	(X - M)

5 10 15 25 45

Deviation Scores

-15 -10 -5 +25 (X - M)





If we were to add up deviation scores, they would always equal zero, which is not very useful.

Squared “Deviation” Scores

$M = 20$

5

10

15

25

45

	-15	-10	-5	+5	+25	
Squared Deviation Scores	225	100	25	25	625	$(X - M)^2$

Sum of Squares

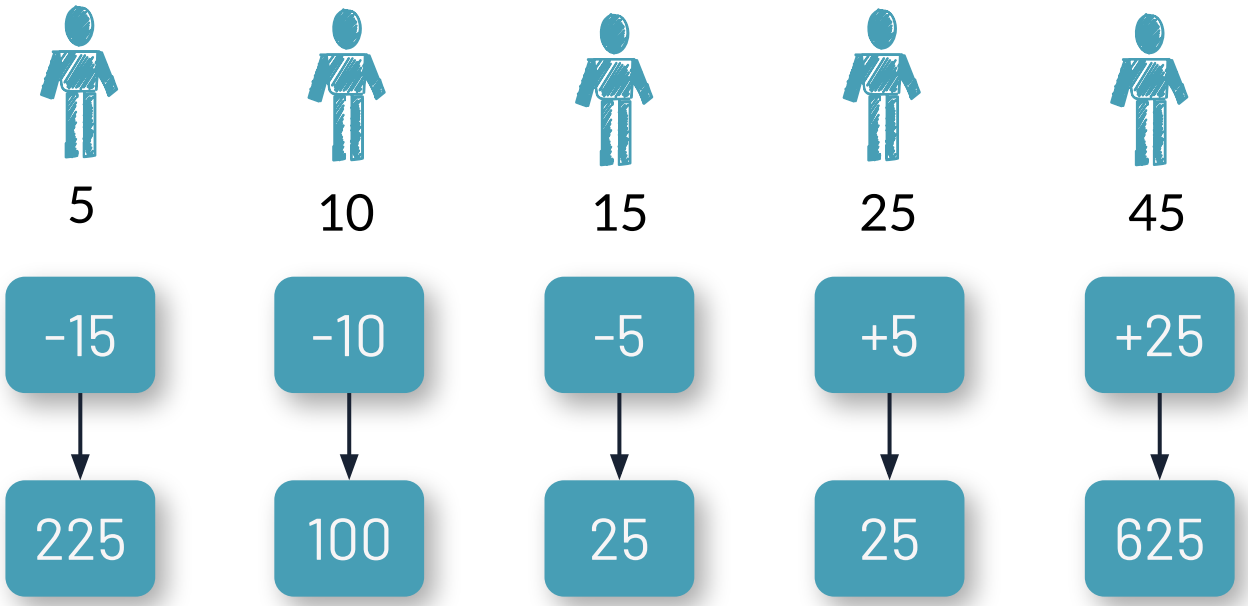
(SS)

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

SS is the number we get if we **add all the squared deviation scores**.

Sum of Squares

$M = 20$



SS = 225 + 100 + 25 + 25 + 625 = **1000**

Population Variance (σ^2)

Sometimes we need to calculate this.

score

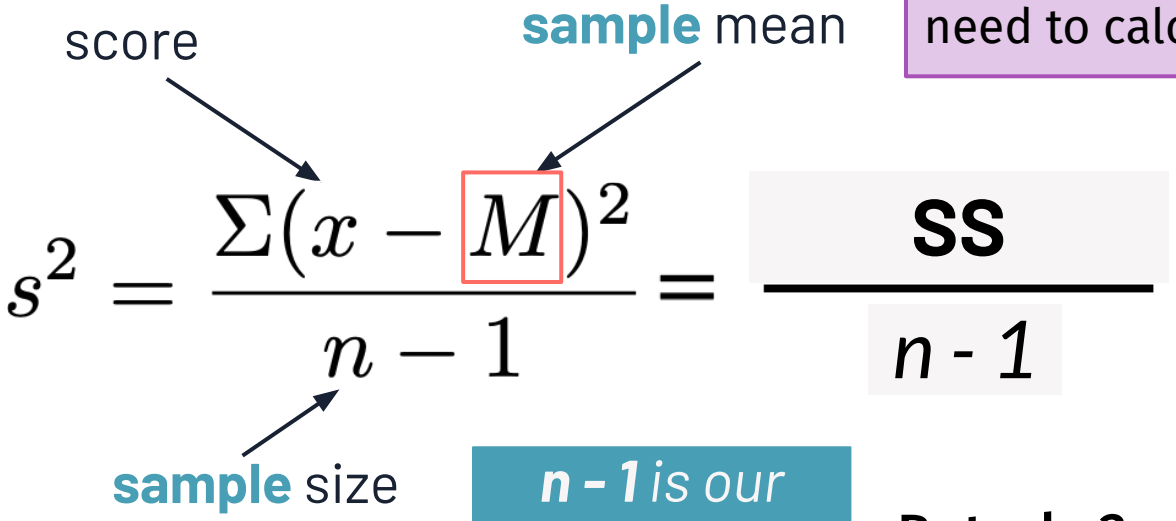
population mean

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N} = \frac{\mathbf{SS}}{N}$$

population size

Sample Variance (s^2)

Most of the time we need to calculate this.



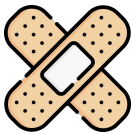
But why?

Remember our mean calculation from last lecture?

4 1 6 5

$$M = \frac{\Sigma x}{n} = \frac{4 + 1 + 6 + 5}{4} = \mathbf{4}$$

If I cover up one number like this:

4 1  5

And tell you: “The sample mean is M = 4, tell me what the covered number is”.
There is only one right answer, 6. It’s like sudoku.

Remember our mean calculation from last lecture?

This “sudoku”-like idea, is called <u>degrees of freedom</u>, an important idea in many calculations you will see throughout the semester.

In our case, the degrees of freedom is 3, which is calculated using n - 1.

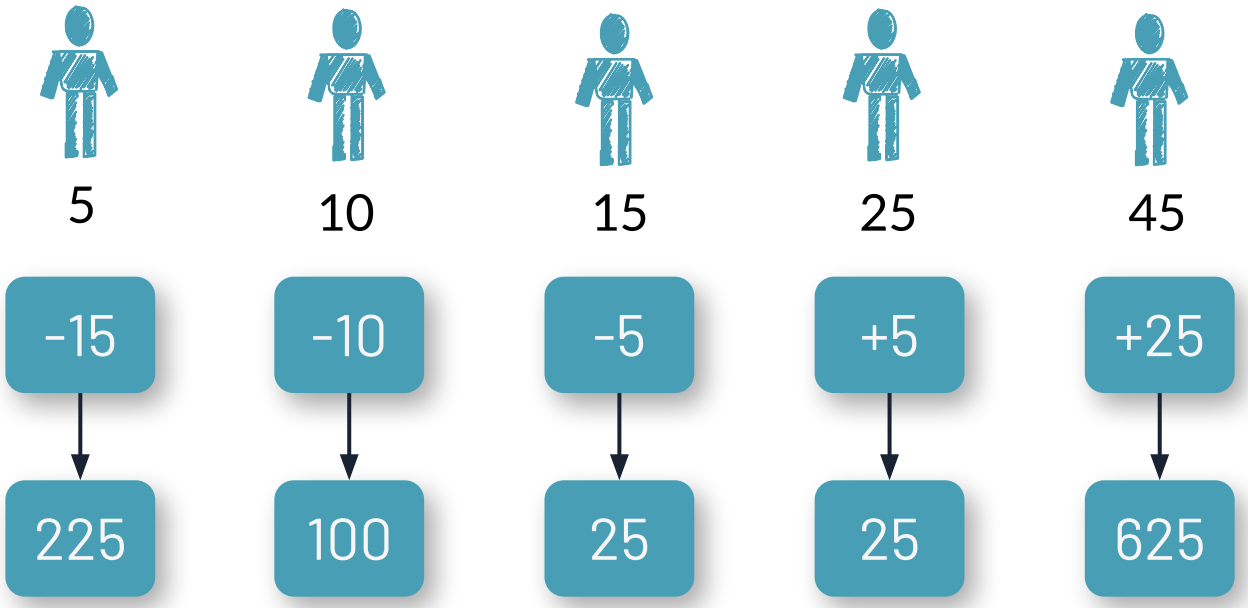
If I cover up one number like this:

This is how the sample variance is corrected.

And tell you: “The sample mean is M = 4, tell me what the covered number is”. There is only one right answer, 6. It’s like sudoku.

Sample Variance

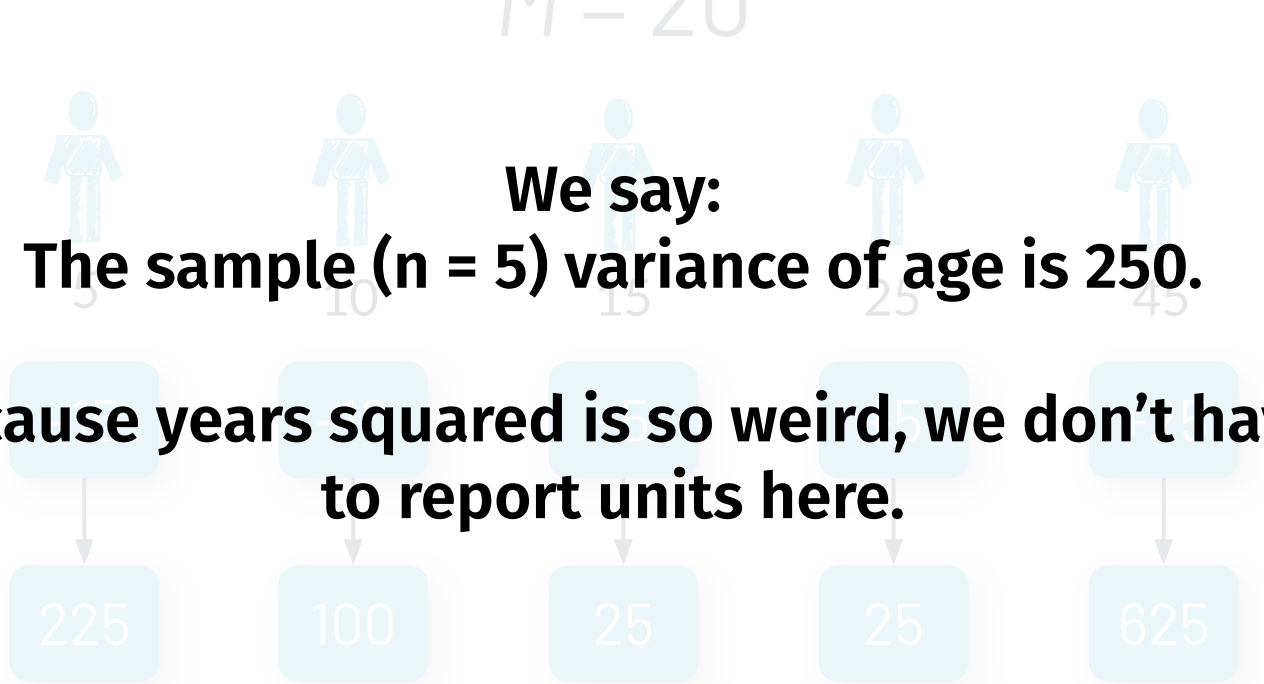
$M = 20$



SS = 1000

$s^2 = 1000/(5-1) = 1000/4 = \mathbf{250}$ years^2

Sample Variance



We say:
The sample (n = 5) variance of age is 250.

Because years squared is so weird, we don't have to report units here.

$s^2 = 1000/(5-1) = 1000/4 = \mathbf{250}$ years^2

Variance and Standard Deviation

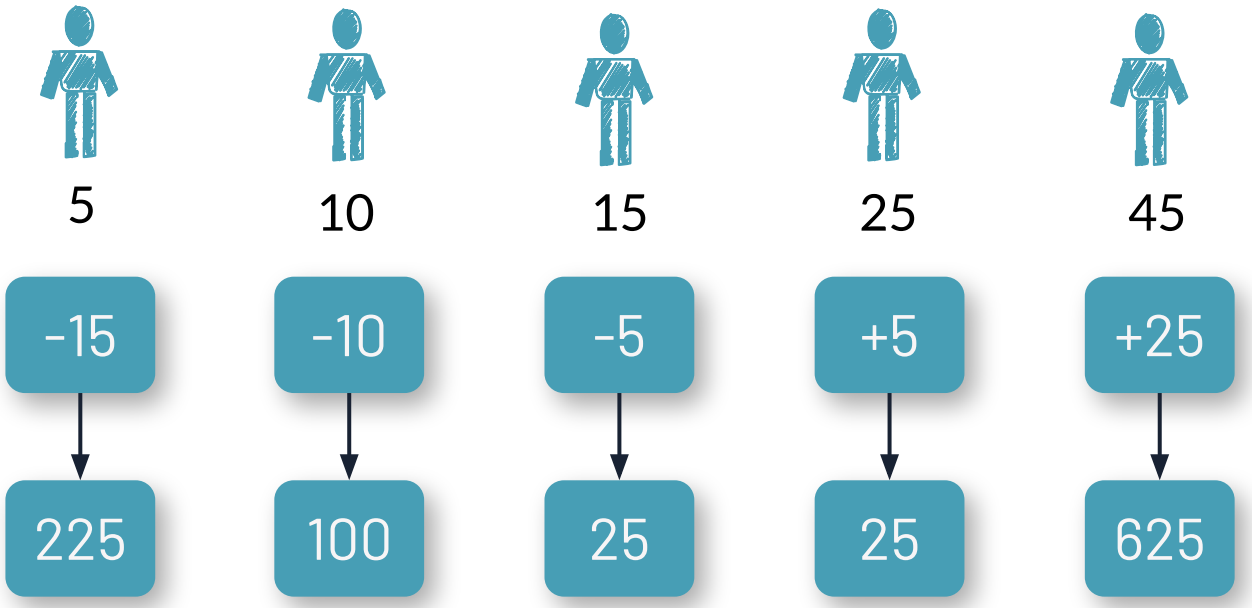
$$s^2 = \frac{SS}{n - 1}$$

Squared units? That does not make much sense, so let's take the square root of the variance.

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

Sample Standard Deviation

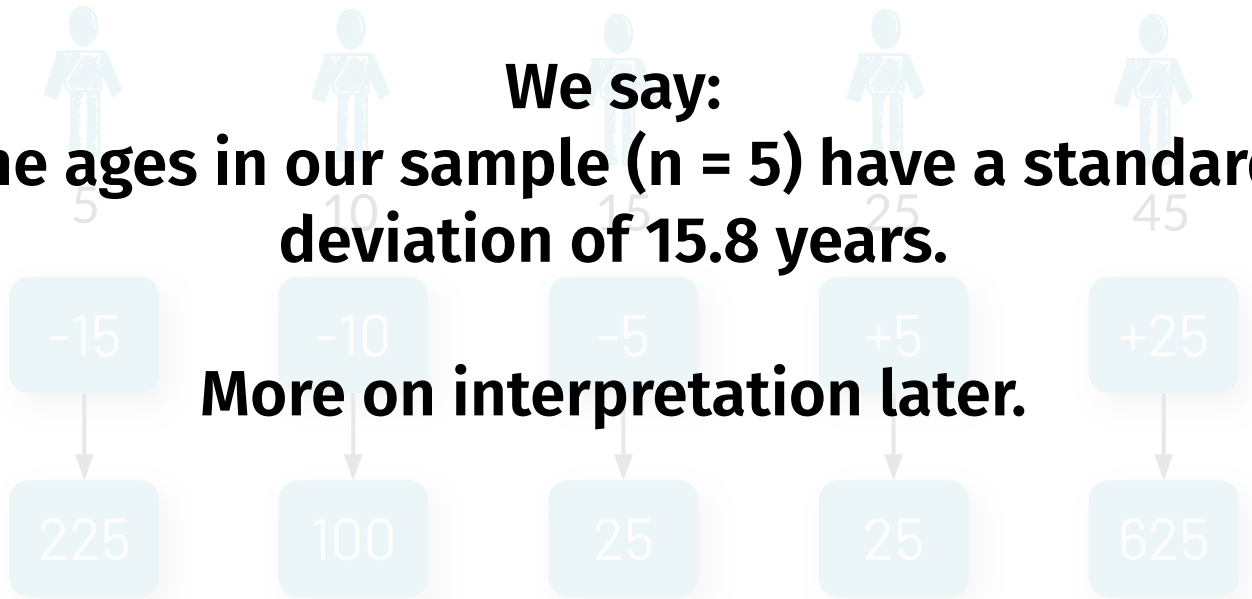
$M = 20$



$s = \sqrt{250} =$ **15.8** years

Sample Standard Deviation

$M = 20$



$s = \sqrt{250} = \mathbf{15.8}$ years

We say:
The ages in our sample (n = 5) have a standard deviation of 15.8 years.

More on interpretation later.

Calculating Sample Variance & SD

$$s^2 = \frac{\Sigma(x - M)^2}{n - 1}$$

STEP 1: Calculate the **mean**

x		
22		
18		
19		
21		
17		
23		
20		

140 / 7 = **20**

Calculating Sample Variance & SD

$$s^2 = \frac{\Sigma(x - M)^2}{n - 1}$$

STEP 2: Find the **deviation** for each score.

x	x - M	
22	22 - 20 = +2	
18	18 - 20 = -2	
19	19 - 20 = -1	
21	21 - 20 = +1	
17	17 - 20 = -3	
23	23 - 20 = +3	
20	20 - 20 = 0	

M = 20

Remember this means we just **subtract each score from the mean** (*M*).

Calculating Sample Variance & SD

$$s^2 = \frac{\Sigma(x - M)^2}{n - 1}$$

STEP 3: **Square** the deviation scores.

x	x - M	$(x - M)^2$
22	22 - 20 = +2	+4
18	18 - 20 = -2	+4
19	19 - 20 = -1	+1
21	21 - 20 = +1	+1
17	17 - 20 = -3	+9
23	23 - 20 = +3	+9
20	20 - 20 = 0	0

M = 20

Calculating Sample Variance & SD

$$s^2 = \frac{\Sigma(x - M)^2}{n - 1}$$

STEP 4: **Add up** the deviation scores to get **SS**

x	x - M	$(x - M)^2$
22	22 - 20 = +2	+4
18	18 - 20 = -2	+4
19	19 - 20 = -1	+1
21	21 - 20 = +1	+1
17	17 - 20 = -3	+9
23	23 - 20 = +3	+9
20	20 - 20 = 0	0

M = 20

Σ = 28

These should all be **positive** numbers!

Remember **"SS"** from earlier? This is what we just calculated!

Calculating Sample Variance & SD

STEP 5: Plug the **SS** into the formula and calculate

x	x - M	$(x - M)^2$
22	22 - 20 = +2	+4
18	18 - 20 = -2	+4
19	19 - 20 = -1	+1
21	21 - 20 = +1	+1
17	17 - 20 = -3	+9
23	23 - 20 = +3	+9
20	20 - 20 = 0	0

M = 20

SS = 28

$$s^2 = \frac{\Sigma(x - M)^2}{n - 1}$$

s^2 = SS / n-1

s^2 = 28 / (7-1)

s^2 = **4.67**

Calculating Sample Variance & SD

STEP 5: Plug the **SS** into the formula and calculate

x	x - M	$(x - M)^2$
22	22 - 20 = +2	+4
18	18 - 20 = -2	+4
19	19 - 20 = -1	+1
21	21 - 20 = +1	+1
17	17 - 20 = -3	+9
23	23 - 20 = +3	+9
20	20 - 20 = 0	0

M = 20

SS = 28

$$s^2 = \frac{\Sigma(x - M)^2}{n - 1}$$

s^2 = SS / n-1

s^2 = 28 / (7-1)

s^2 = **4.67**

s = **2.16**

If Asked to Calculate Population Variance & SD

1) swap out sample mean for population mean; 2) denominator is N

x	x - μ	$(X - \mu)^2$
22	22 - 20 = +2	+4
18	18 - 20 = -2	+4
19	19 - 20 = -1	+1
21	21 - 20 = +1	+1
17	17 - 20 = -3	+9
23	23 - 20 = +3	+9
20	20 - 20 = 0	0

$\mu = 20$ $SS = 28$

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

$\sigma^2 = SS / N$

$\sigma^2 = 28 / 7$

$\sigma^2 = 4$

$\sigma = 2$

In-Class Activity (ICA 4)

1. Record data for your table on: ***How many pets do you have?***
2. Calculate your group's **range**, **variance**, and **standard deviation** on the white board (use the table below to help guide you; show all work).

x (# of pets)	**x - M**	**$(x - M)^2$**
?		
?		
?		
?		
?		

SS = ?

Interpreting Measures of Variability

Interpret the SD in plain English

$M = 20$

$s = \sqrt{250} =$ **15.8** years

A standard deviation of 15.8 means that, on average, participants' ages differ from the sample mean age by about 15.8 years.

Reviewing **Notation**

Meaning	Population	Sample
Mean	μ	M (or $\bar{x}$)
Variance	σ^2	s^2
Standard Deviation	σ	s (or SD)

Usually, you can tell which type of variance or SD you can calculate based on the information given to you (i.e., is it population or sample mean? What does the context imply?)

Interpret the SD in plain English

$M = 20$

$s = \sqrt{250} = 15.8$ years

A standard deviation of 15.8 means that, on average, participants' ages differ from the sample mean age by about 15.8 years.

The plain English version is what you want to say when you explain what you are learning about/what your research is about to your <u>friends and family</u>.

A Note on APA Style

The most common descriptive statistics are the **mean** (central tendency) and the **standard deviation** (variability), which are usually reported together in research.

Note: In many journals, especially those following APA style, the symbol ***SD*** is used for the **sample** standard deviation.

"Participants who played violent video games displayed more aggressive responses **(*M* = 8.45, *SD* = 1.72)** than those who viewed the a cartoon **(*M* = 4.21, *SD* = 1.01)**."

Italics!

A Note on APA Style

The most common descriptive statistics are the **mean** (central tendency) and the **standard deviation** (variability), which are usually reported together in research.

Note: In many journals, especially those following APA style, the symbol SD is used for the sample standard deviation.

The APA version is what you want to use when you write to fellow scientists & psychologists.

For your homework & exams, please include <u>both</u> interpretations.

"Participants who played violent video games displayed more aggressive responses (*M* = 8.45, *SD* = 1.72) than those who viewed the a cartoon (*M* = 4.21, *SD* = 1.01)."

Italics!

Descriptive Statistics Table Example

Table 2

Descriptive Statistics

Variable	*N*	*M*	*SD*	*Mdn*	Range	Skew	Kurtosis
ACE	211	2.63	2.51	2	0–10	0.92	0.19
BAI Raw Score	165	13.59	10.91	12	0–51	0.91	0.61
BDI-II Raw Score	159	18.24	13.12	16	0–56	0.67	−0.17
TOPF Word Reading SS	169	99.88	16.52	101	50–131	−0.32	−0.89
Digit Span ss	178	8.73	3.66	9	1–19	0.20	0.19
Trail Making Test-Part A T	178	44.20	13.04	45.4	9–74	−0.36	0.01
Trail Making Test-Part B T	150	44.81	11.36	44.5	15–81	−0.04	0.46
RAVLT Learning T	105	43.52	14.45	43	4–71	−0.32	−0.11
RAVLT Long Delay Recall T	105	46.29	13.74	47	5–68	−0.50	−0.37

Bing-Canar et al. (2024)

Wrap Up

A Note on APA Style

The most common descriptive statistics are the **mean** (central tendency) and the **standard deviation** (variability), which are usually reported together.

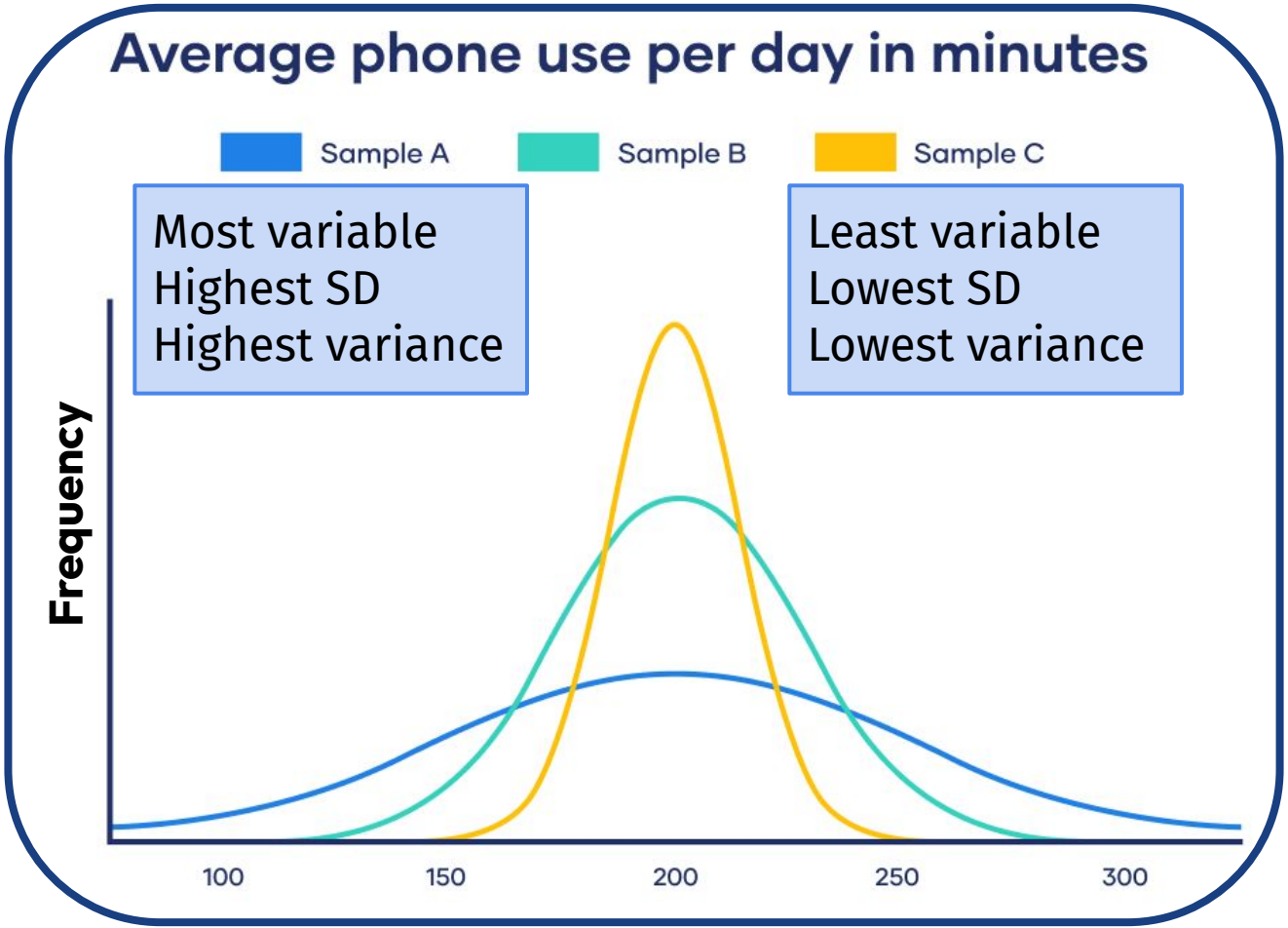
Note: In many journals, especially those following APA style, the symbol *SD* is used for the sample standard deviation.

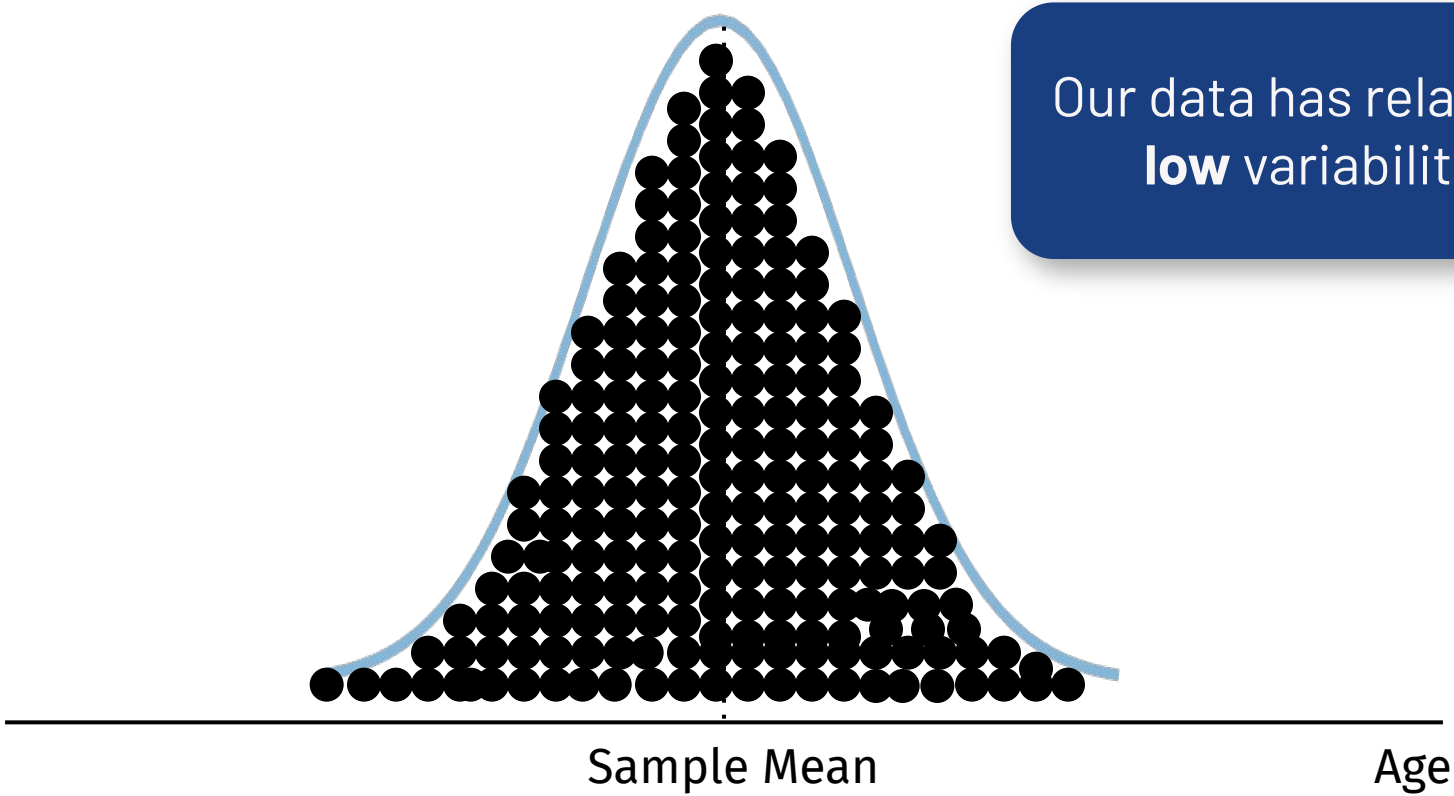
We have explained how to interpret measures of variability through numbers.

But what about visually?

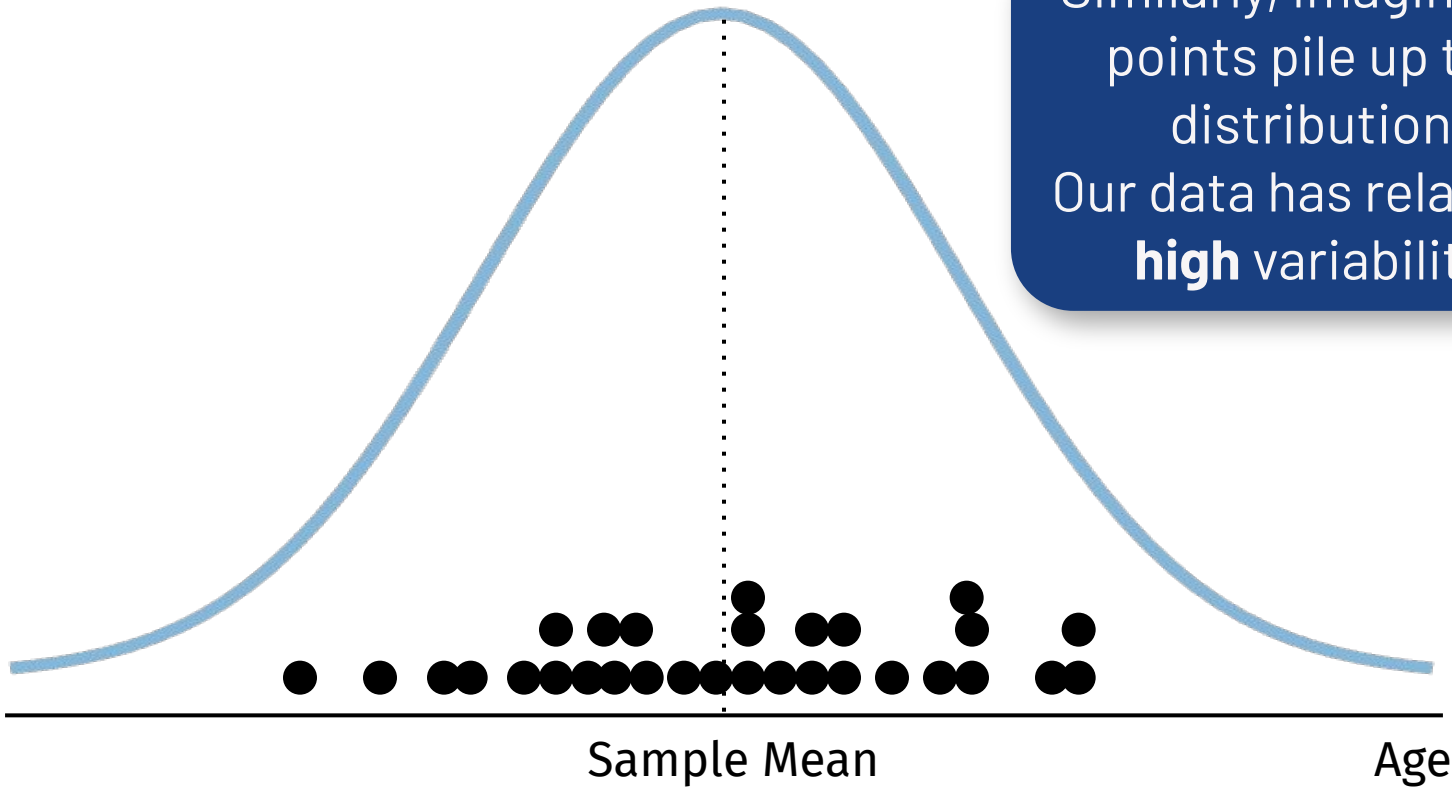
"Participants who played violent video games displayed more aggressive responses (*M* = 8.45, *SD* = 1.72) than those who viewed the a cartoon (*M* = 4.21, *SD* = 1.01)."

Italics!

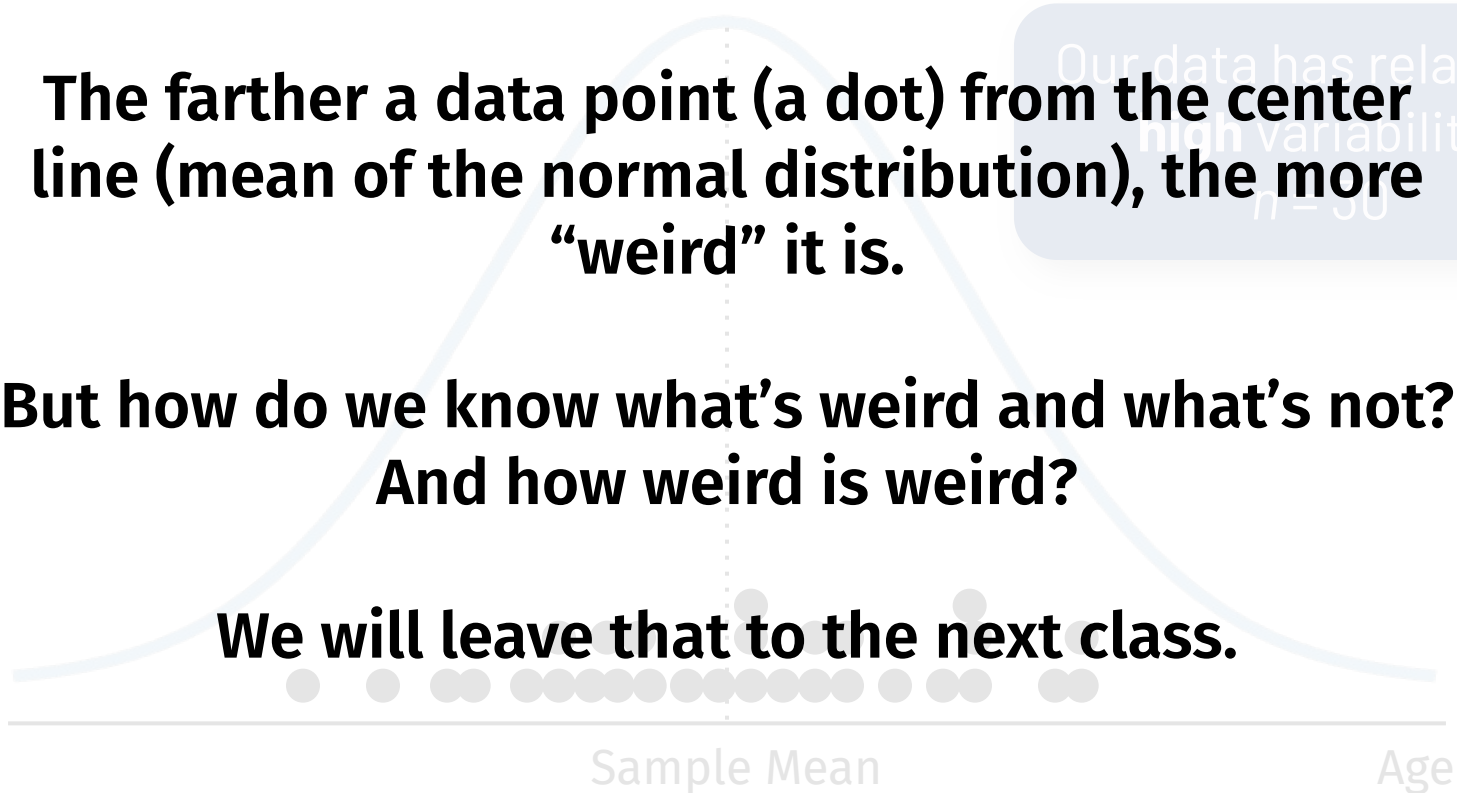

Average phone use per day in minutes
Sample A
Sample B
Sample C
Most variable
Highest SD
Highest variance
Least variable
Lowest SD
Lowest variance
Frequency
100
150
200
250
300
Minutes



Our data has relatively **low** variability.



Similarly, imagine the points pile up to a distribution.
Our data has relatively **high** variability.



The farther a data point (a dot) from the center line (mean of the normal distribution), the more “weird” it is.

But how do we know what’s weird and what’s not? And how weird is weird?

We will leave that to the next class.