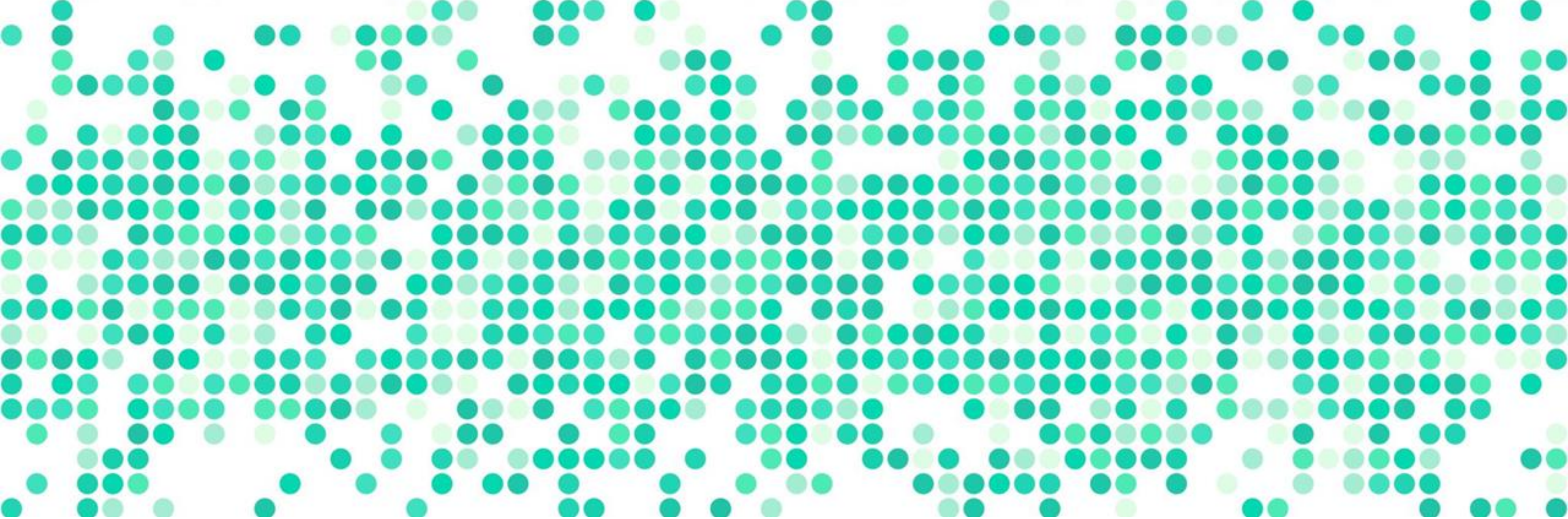
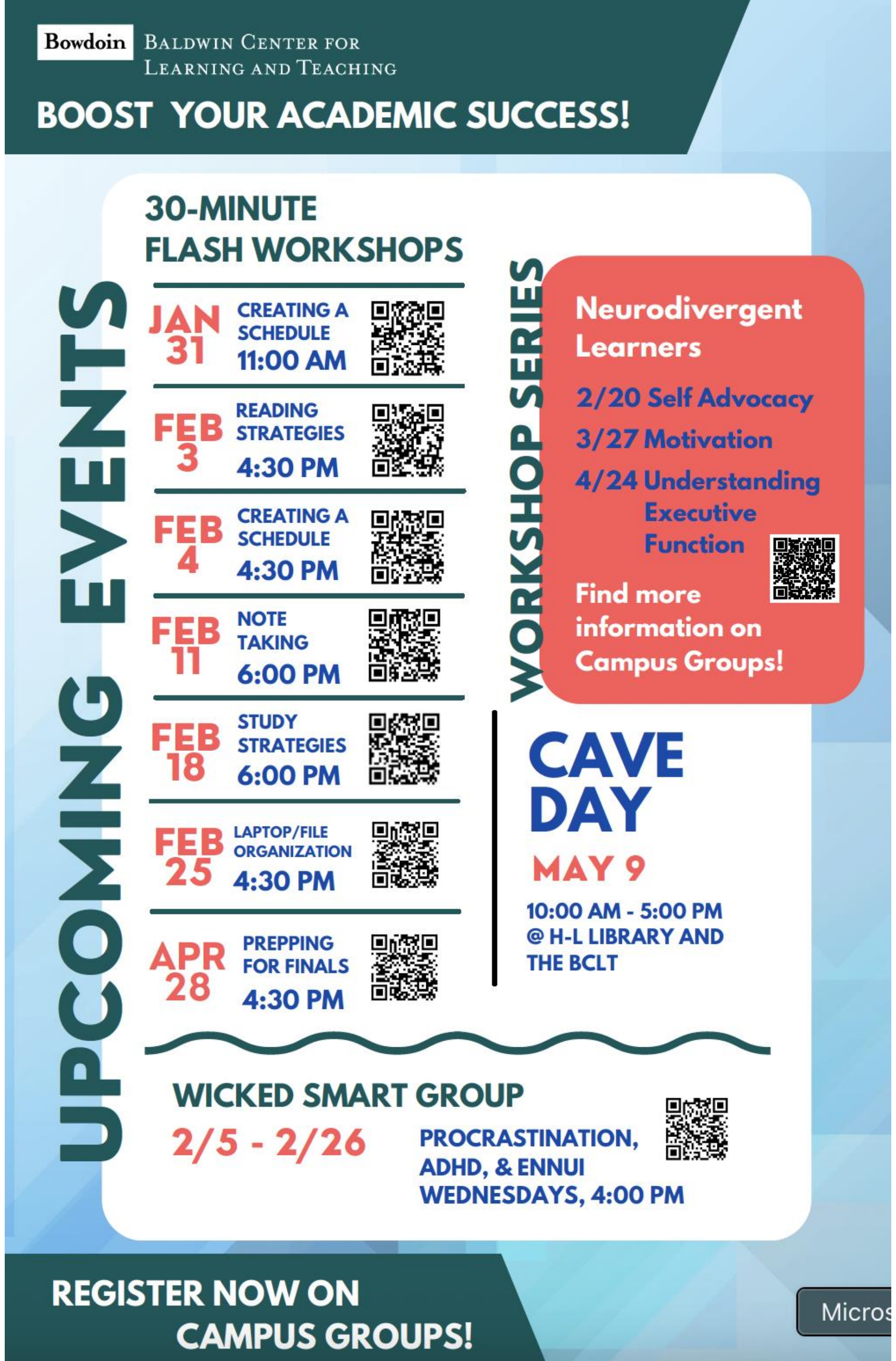


DATA ANALYSIS

Week 3: Variability



Bowdoin Baldwin Center for Learning and Teaching

BOOST YOUR ACADEMIC SUCCESS!

UPCOMING EVENTS

30-MINUTE FLASH WORKSHOPS

- JAN 31 CREATING A SCHEDULE 11:00 AM
- FEB 3 READING STRATEGIES 4:30 PM
- FEB 4 CREATING A SCHEDULE 4:30 PM
- FEB 11 NOTE TAKING 6:00 PM
- FEB 18 STUDY STRATEGIES 6:00 PM
- FEB 25 LAPTOP/FILE ORGANIZATION 4:30 PM
- APR 28 PREPPING FOR FINALS 4:30 PM

WORKSHOP SERIES

Neurodivergent Learners

2/20 Self Advocacy

3/27 Motivation

4/24 Understanding Executive Function

Find more information on Campus Groups!

CAVE DAY

MAY 9

10:00 AM - 5:00 PM
@ H-L LIBRARY AND
THE BCLT

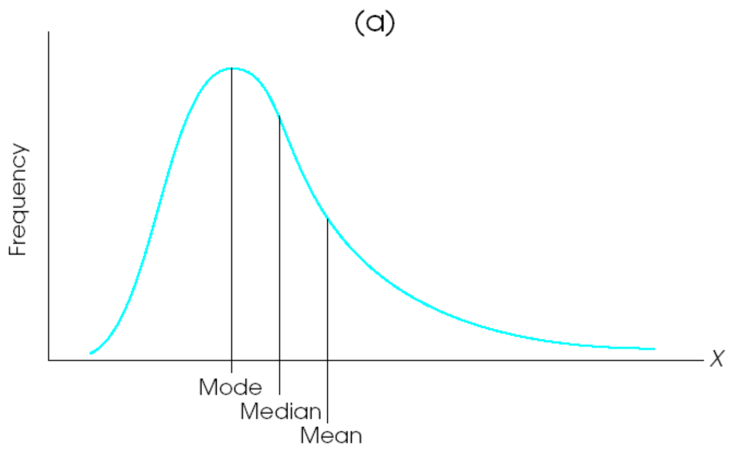
WICKED SMART GROUP

2/5 - 2/26

PROCRASTINATION, ADHD, & ENNUI
WEDNESDAYS, 4:00 PM

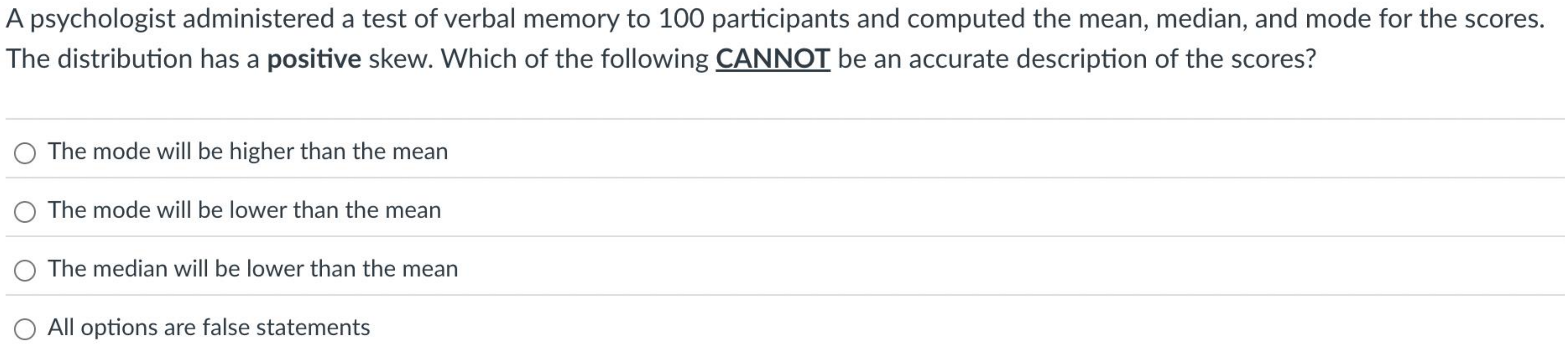
REGISTER NOW ON CAMPUS GROUPS!

lingering question



positive skew

- I did not understand the 7th question in the week 2 quiz.



A psychologist administered a test of verbal memory to 100 participants and computed the mean, median, and mode for the scores. The distribution has a **positive** skew. Which of the following **CANNOT** be an accurate description of the scores?

○ The mode will be higher than the mean

○ The mode will be lower than the mean

○ The median will be lower than the mean

○ All options are false statements

recap of fitting models

- models are fit to data: data = model + error
- we fit “central tendencies”/models to the data (mean / median / mode)
- we calculated “errors”/distances between the data and our model(s)
  - sum of squared errors (SSE or SS): $\sum_{i=1}^{N}(X_i - \mu)^2$
  - mean of squared errors (MSE): $\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N} = \frac{SS}{N}$
  - root mean squared error (RMSE): $\sqrt[2]{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}} = \sqrt{MSE}$

lingering question

- Why do we need the sum of square roots when calculating the standard deviation?

data = model + error

error = data - model

error = data - mean

$$\sum error = \sum \text{data} - \text{mean}$$

adding up negative and positive errors leads to total error canceling out

$$\sum error^2 = \sum(\text{data} - \text{mean})^2$$

we square the errors so that all errors are ≥ 0 and the sum of errors is non-zero.

$$average\ error = \sum \frac{(\text{data} - \text{mean})^2}{N}$$

we take the average so that the error is not dependent on sample size (N)

but now, the average error is not in the original units of data, it is the square of the units (e.g., if data were heights, errors are heights^2)

to bring the error back into the original units of the data, we take the square root

$$average\ error = \sqrt{\sum \frac{(\text{data} - \text{mean})^2}{N}}$$

this represents the “average” distance from the mean

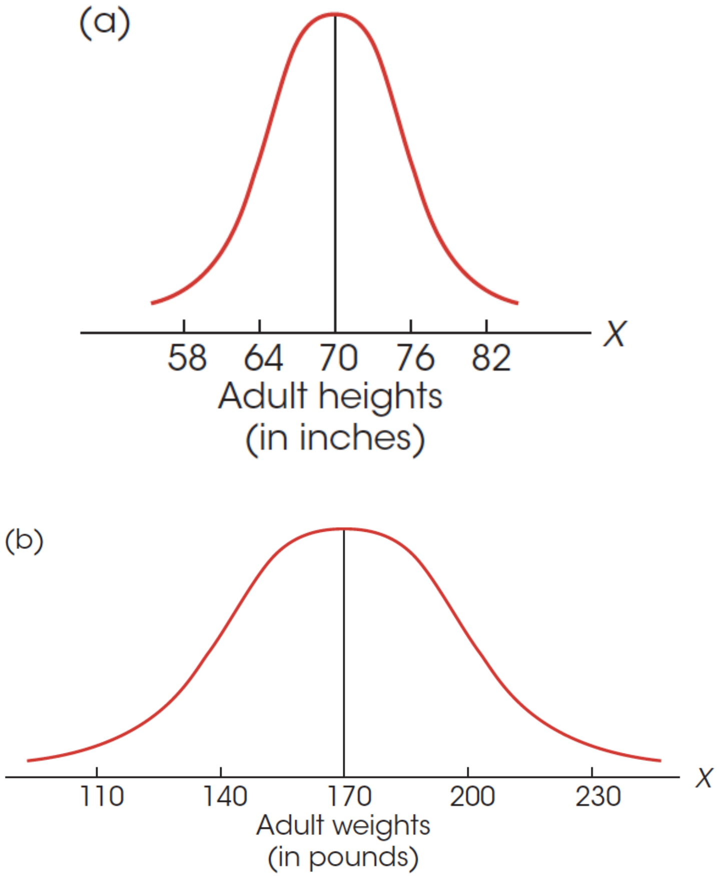
today's agenda

variability

z-scores

variability

- variability describes the spread of scores in a distribution
- measures of variability
  - range = maximum – minimum
  - variance = mean squared error from the mean (MSE or σ^2) = average of ***squared*** distances/errors from the mean
  - standard deviation = root mean squared error from the mean (RMSE or σ) = average distance/error from the mean **in original units**
- variance and standard deviation are defined relative to the mean, i.e., how well does the mean fit the data?



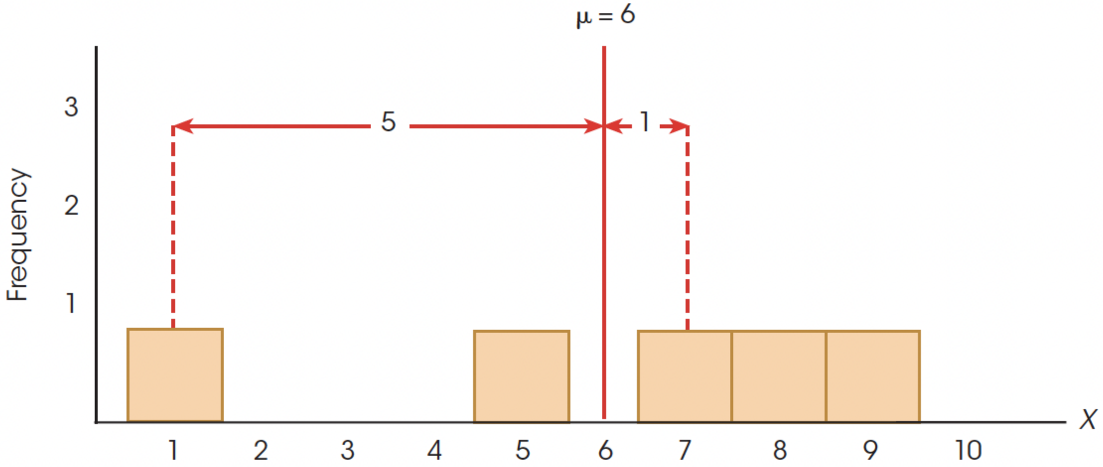
visual inspection

- we can estimate/calculate the mean
- the farthest score is 5 points away
- the closest score is 1 point away
- on average, scores are likely $\frac{5+1}{2}$ away = 3 points away
- what is our actual estimate of standard deviation for these scores?

$$\sqrt[2]{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

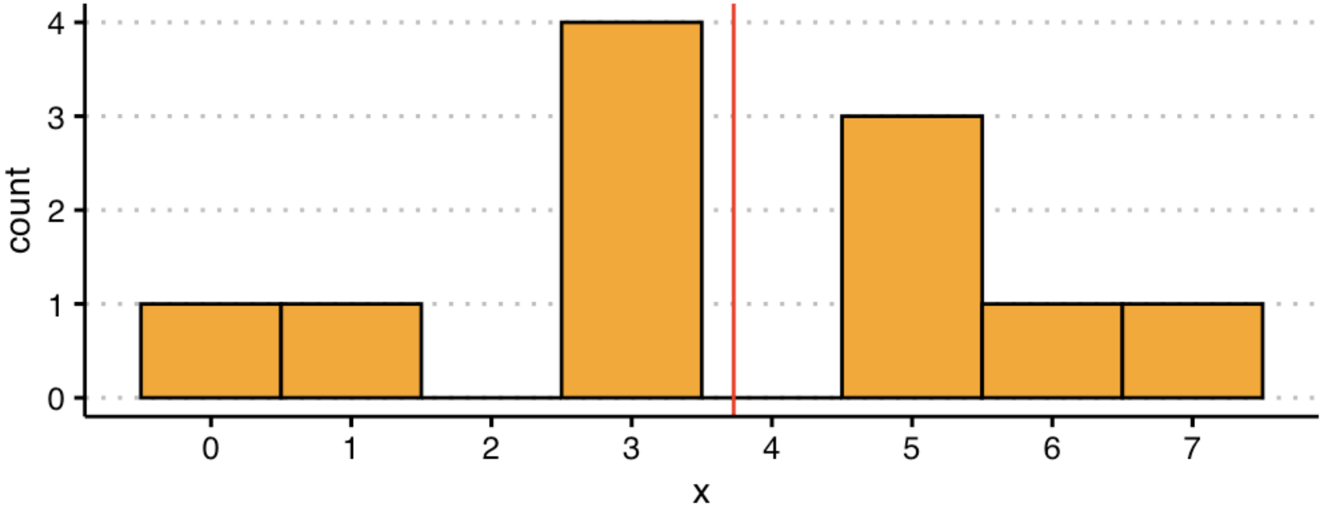
$$= \sqrt{\frac{(1-6)^2 + (5-6)^2 + (7-6)^2 + (8-6)^2 + (9-6)^2}{5}}$$

$$= \sqrt{\frac{25 + 1 + 1 + 4 + 9}{5}} = 2.83$$



activity

- 5,5,5, 3,3,3,3,6, 7, 1, 0
- calculate the mean
- visually estimate the standard deviation

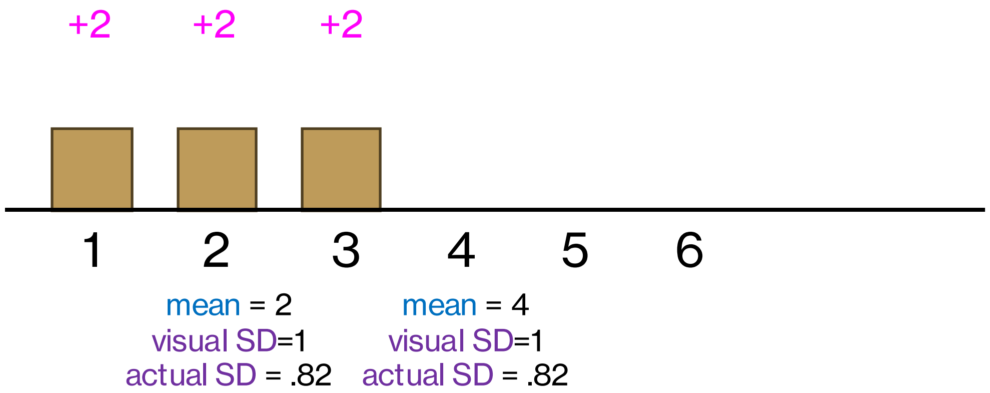


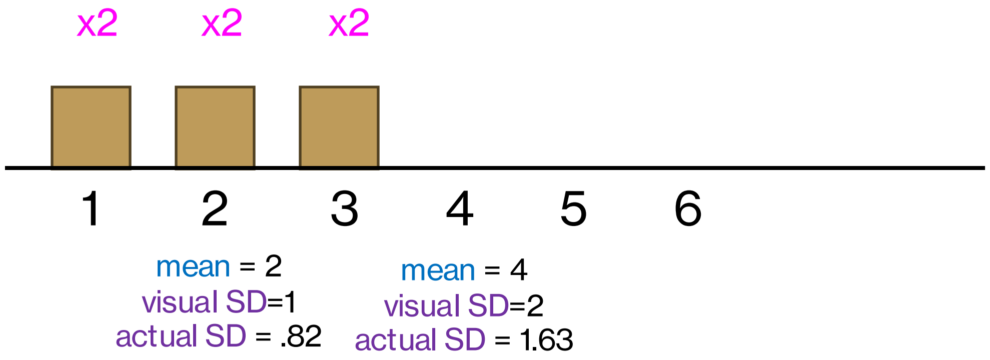
W3 Activity 1

- complete the activity on your own
- discuss the logic behind your answers with a peer
- re-attempt the activity

properties of standard deviation

- adding/subtracting a constant to all scores will have no impact on standard deviation
- multiplying/dividing a constant to all scores will change the standard deviation by the same constant





questions?

SSE: definitional vs. computational formulas

$$\sum(X-\mu)^2 = \sum X^2 - \frac{(\sum X)^2}{N}$$

definitional formula

computational formula

$$\sum(X-\mu)^2 = \sum(X^2+\mu^2-2X\mu) = \sum X^2 + \sum \mu^2 - 2\sum X\mu = \sum X^2 + N\mu^2 - 2\mu\sum X =$$

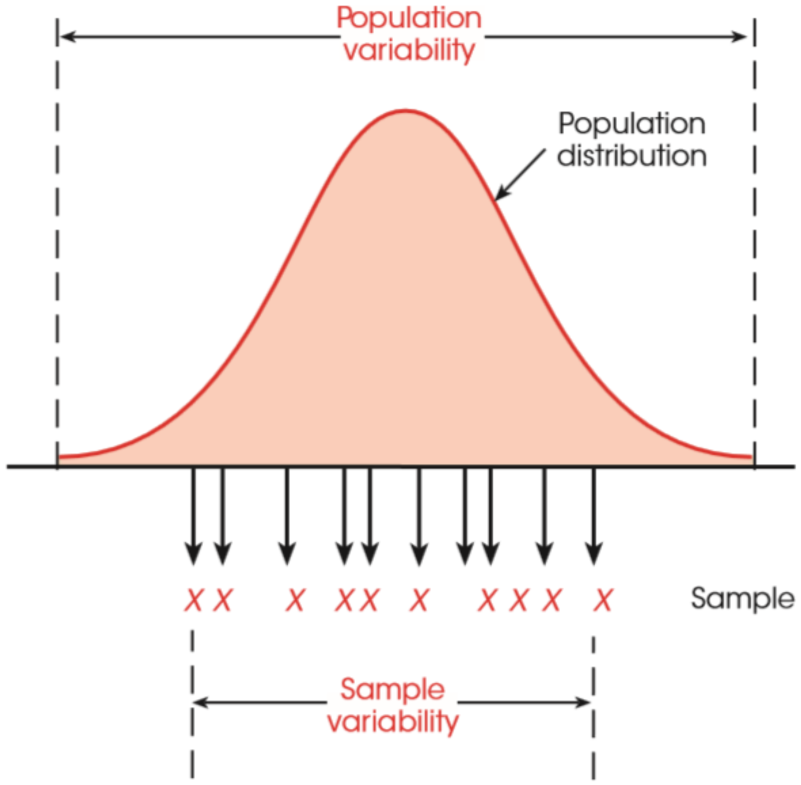
$$= \sum X^2 + N\mu^2 - 2\,\frac{\sum X}{N}\sum X = \sum X^2 + N\mu^2 - 2\,\frac{(\sum X)^2}{N} = \sum X^2 + N\frac{\sum X}{N}\frac{\sum X}{N} - 2\,\frac{(\sum X)^2}{N} = \sum X^2 + \frac{(\sum X)^2}{N} - 2\,\frac{(\sum X)^2}{N}$$

$$= \sum X^2 - \frac{(\sum X)^2}{N}$$

only for your curiosity, stick to definitional formula for this class: easier to remember and understand

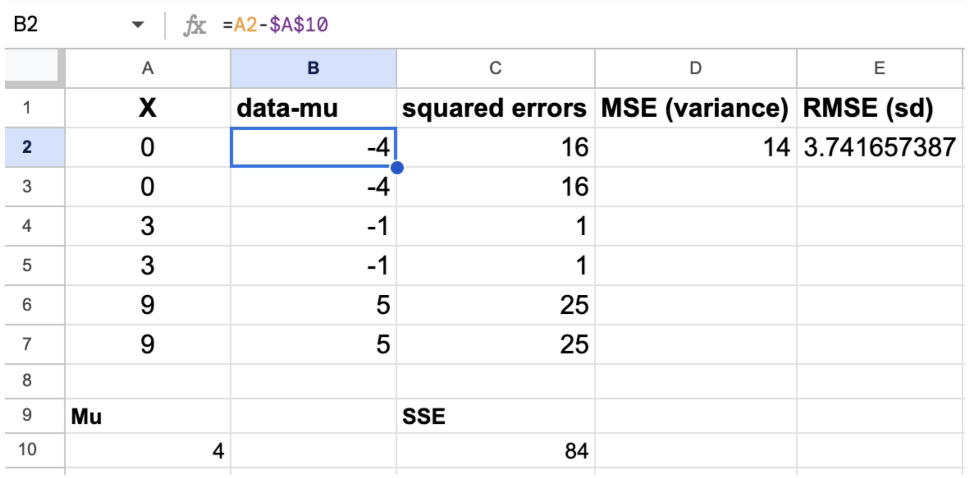
from populations to samples

- we have been talking about central tendencies and spread for populations, but we hardly ever have access to the populations!
- sample means (*M*) contribute to sample-based estimates of variance (s^2) and standard deviation (*s*)
- sampling tends to focus more on “typical” scores, so we tend to miss out on extreme scores from the population
- as a result, samples tend to underestimate population variability



a demonstration: small population

- consider an island population (N = 6) where people were asked to report how many trees they own on the island
- 2 people owned no trees, 2 people owned 3 trees each, and 2 people owned 9 trees each!
- we calculate the mean and standard deviation of trees owned for this population

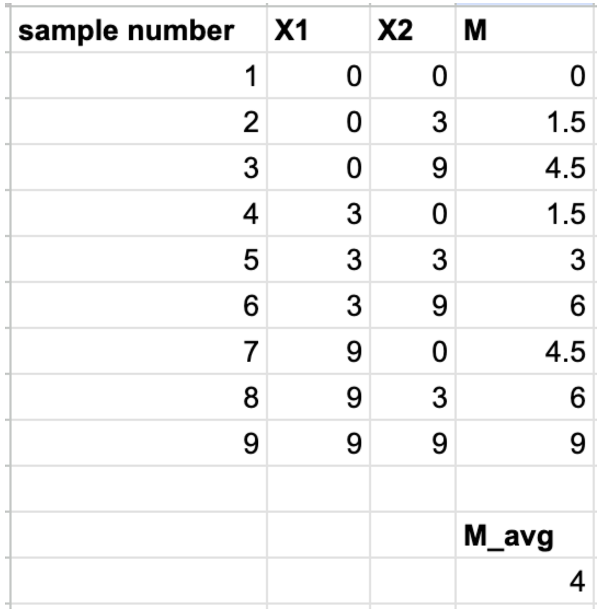


B2 fx =A2-A10

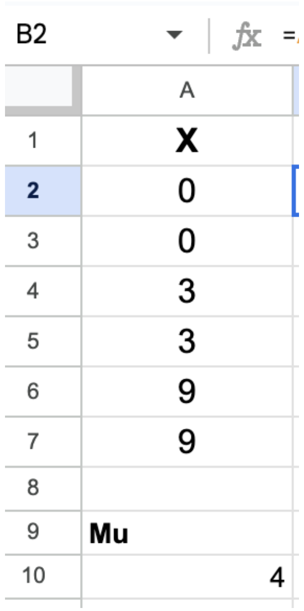
	A	B	C	D	E
1	X	data-mu	squared errors	MSE (variance)	RMSE (sd)
2	0	-4	16	14	3.741657387
3	0	-4	16		
4	3	-1	1		
5	3	-1	1		
6	9	5	25		
7	9	5	25		
8					
9	Mu		SSE		
10	4		84		

a demonstration: small samples

- now we take all possible samples of size 2 from this population
- calculate the mean M for each sample
- average M from all possible samples is equal to the population M: mean is an unbiased statistic!



sample number	X1	X2	M
1	0	0	0
2	0	3	1.5
3	0	9	4.5
4	3	0	1.5
5	3	3	3
6	3	9	6
7	9	0	4.5
8	9	3	6
9	9	9	9
			M_avg
			4



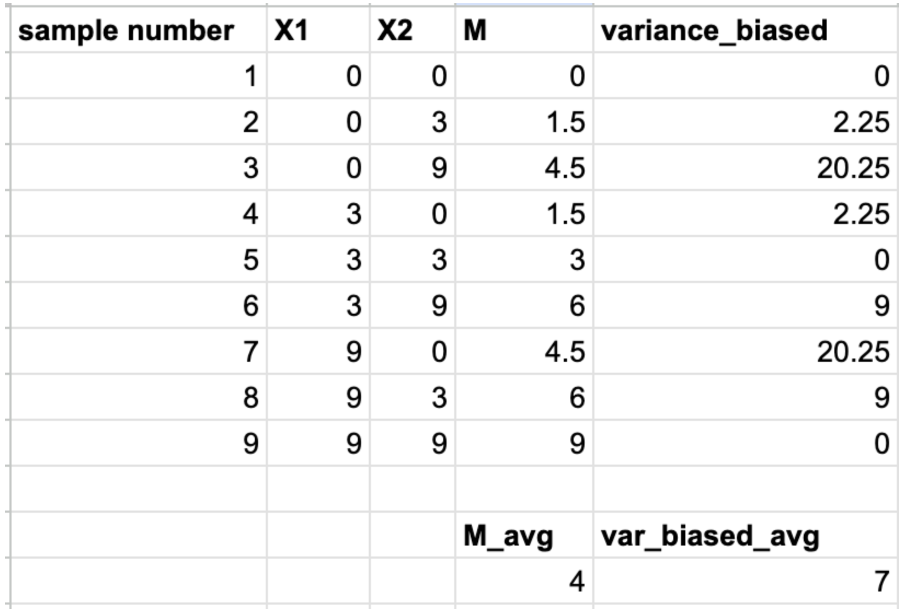
	A
1	X
2	0
3	0
4	3
5	3
6	9
7	9
8	
9	Mu
10	4

a demonstration: small samples

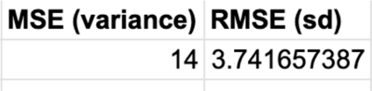
- calculate the variance (MSE) of each sample

$$= \frac{(X_1 - M_{sample})^2 + (X_2 - M_{sample})^2}{2}$$

- average variance of all samples is LOWER than the population variance: variance is a biased statistic!



sample number	X1	X2	M	variance_biased
1	0	0	0	0
2	0	3	1.5	2.25
3	0	9	4.5	20.25
4	3	0	1.5	2.25
5	3	3	3	0
6	3	9	6	9
7	9	0	4.5	20.25
8	9	3	6	9
9	9	9	9	0
			M_avg	var_biased_avg
			4	7



MSE (variance)	RMSE (sd)
14	3.741657387

a demonstration: small samples

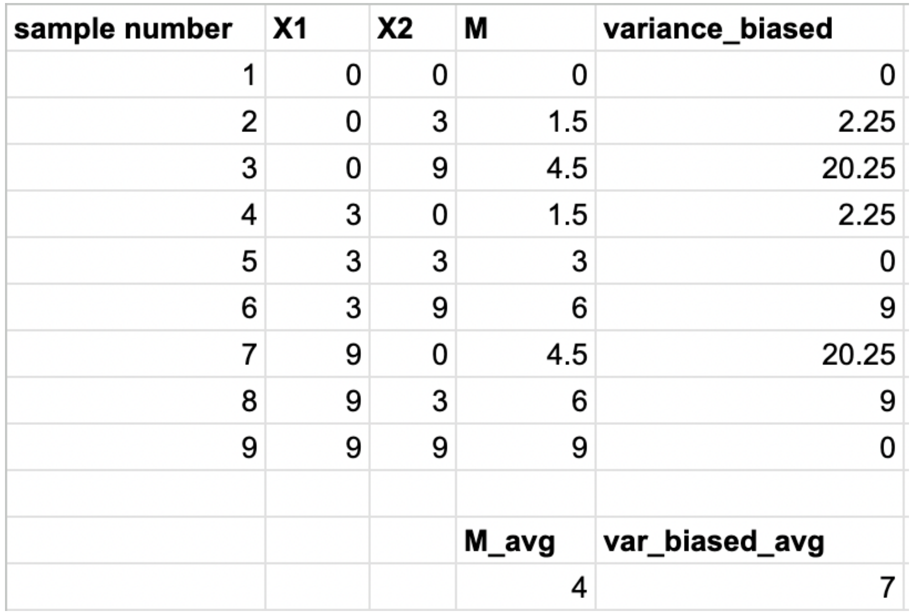
- we need to penalize the sample variance so that it accurately estimates the population variance
- we need to make variance (MSE) a larger number

$$\frac{\sum_{i=1}^{N}(X_i - M_{sample})^2}{n}$$

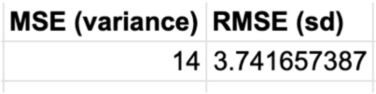
- we can decrease the the denominator: divide by (n – 1) instead

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - M_{sample})^2}{n-1}$$

- also called the Bessel's correction

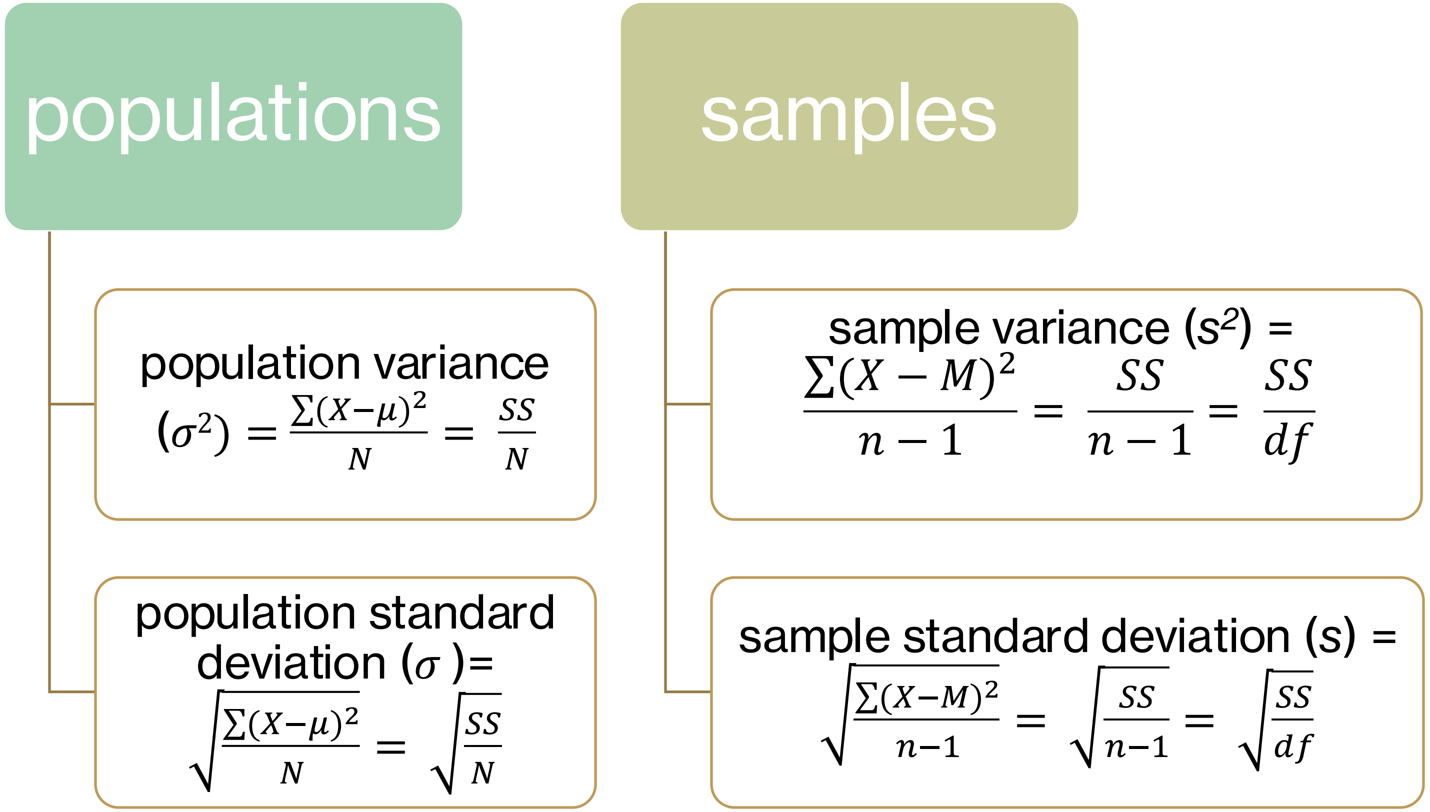


sample number	X1	X2	M	variance_biased
1	0	0	0	0
2	0	3	1.5	2.25
3	0	9	4.5	20.25
4	3	0	1.5	2.25
5	3	3	3	0
6	3	9	6	9
7	9	0	4.5	20.25
8	9	3	6	9
9	9	9	9	0
			M_avg	var_biased_avg
			4	7



MSE (variance)	RMSE (sd)
14	3.741657387

populations vs. samples



populations

population variance $(\sigma^2) = \frac{\sum(X-\mu)^2}{N} = \frac{SS}{N}$

population standard deviation (σ)= $\sqrt{\frac{\sum(X-\mu)^2}{N}} = \sqrt{\frac{SS}{N}}$

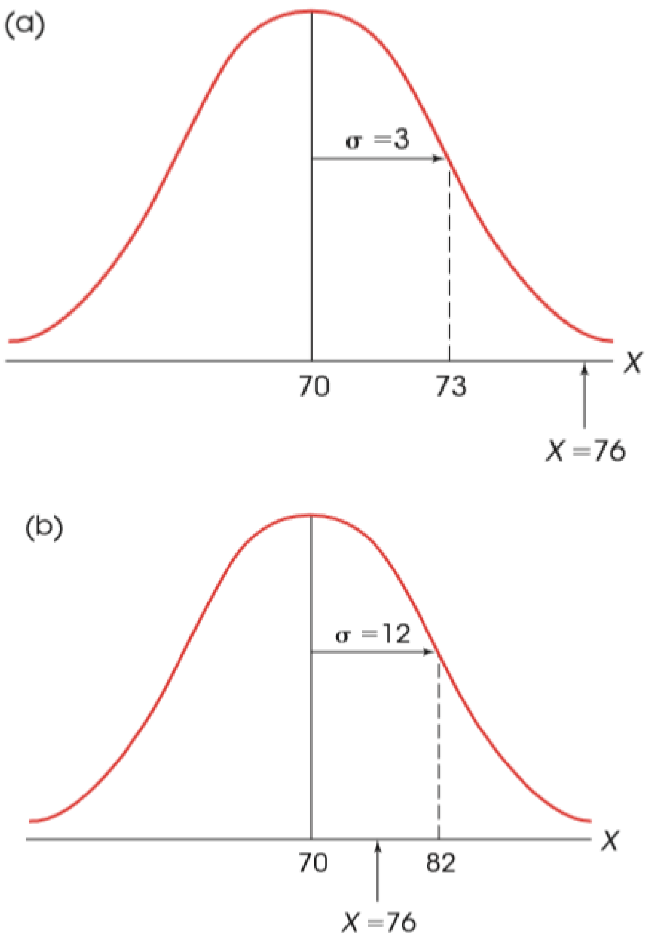
samples

sample variance (s^2) = $\frac{\sum(X-M)^2}{n-1} = \frac{SS}{n-1} = \frac{SS}{df}$

sample standard deviation (s) = $\sqrt{\frac{\sum(X-M)^2}{n-1}} = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{SS}{df}}$

locating scores within distributions

- we have used means and standard deviations as ways to summarize distributions
- but, if you wanted to know how well you performed on a test, how would you apply this knowledge of the distribution to know how well you did?
- means and standard deviations together can be informative in describing a data point's relationship to the distribution



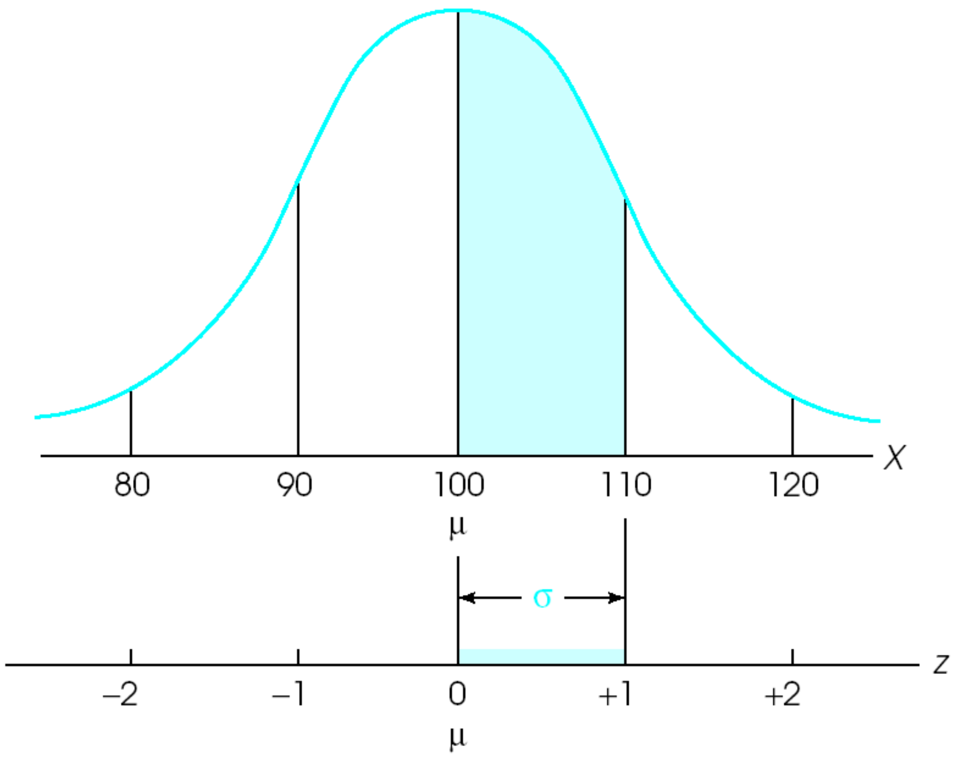
z-scores

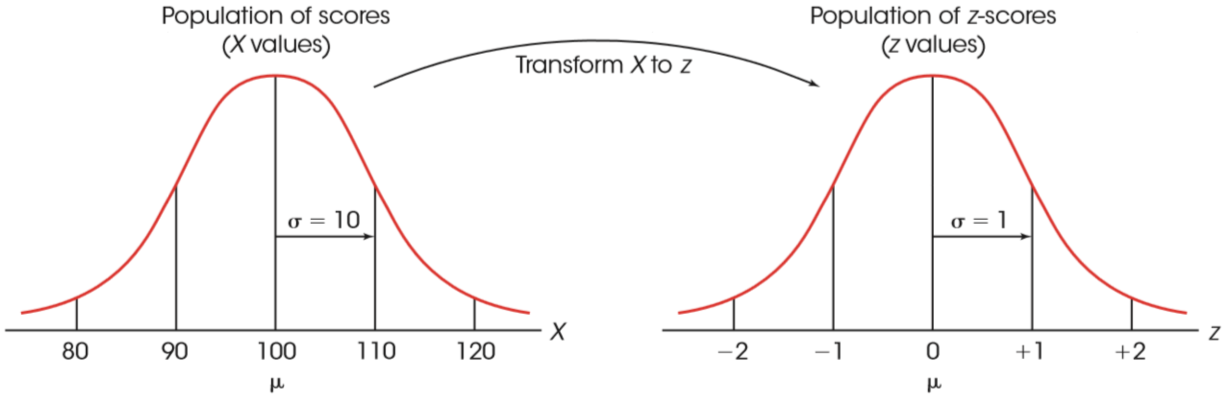
- z-scores are a way to understand how far away a score is from the mean, in standard deviation units

$$z = \frac{X - \mu}{\sigma}$$

  - calculate “distances” or deviation scores and divide by the standard deviation
  - z-score is essentially a ratio that is asking: how extreme is my score relative to the average distance I can expect based on this distribution?

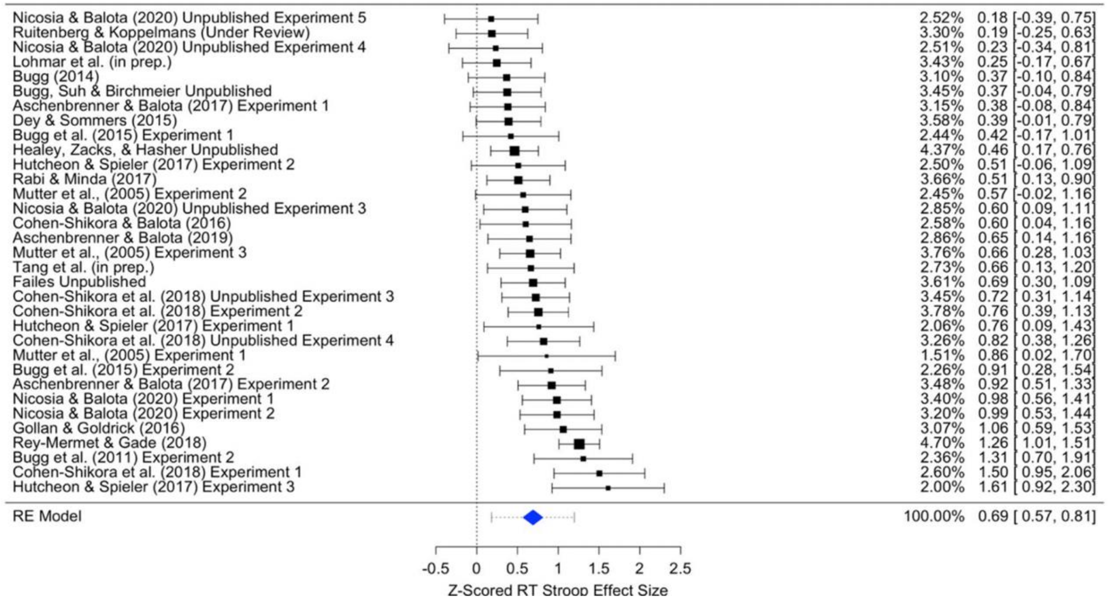
- any distribution can be transformed into a distribution of z-scores





why z-score?

- to understand the position of a score relative to all other scores
- to compare scores on one scale to scores on another scale
- examples from actual research:
  - comparing exam scores from one subject to another
  - younger and older adults performing reaction time-based tasks



Nicosia & Balota (2020) Unpublished Experiment 5	2.52%	0.18	[-0.39, 0.75]
Ruitenberg & Koppelmans (Under Review)	3.30%	0.19	[-0.25, 0.63]
Nicosia & Balota (2020) Unpublished Experiment 4	2.51%	0.23	[-0.34, 0.81]
Lohmar et al. (in prep.)	3.43%	0.25	[-0.17, 0.67]
Bugg (2014)	3.10%	0.37	[-0.10, 0.84]
Bugg, Suh & Birchmeier Unpublished	3.45%	0.37	[-0.04, 0.79]
Aschenbrenner & Balota (2017) Experiment 1	3.15%	0.38	[-0.08, 0.84]
Dey & Sommers (2015)	3.58%	0.39	[-0.01, 0.79]
Bugg et al. (2015) Experiment 1	2.44%	0.42	[-0.17, 1.01]
Healey, Zacks, & Hasher Unpublished	4.37%	0.46	[0.17, 0.76]
Hutcheon & Spieler (2017) Experiment 2	2.50%	0.51	[-0.06, 1.09]
Rabi & Minda (2017)	3.66%	0.51	[0.13, 0.90]
Mutter et al., (2005) Experiment 2	2.45%	0.57	[-0.02, 1.16]
Nicosia & Balota (2020) Unpublished Experiment 3	2.85%	0.60	[0.09, 1.11]
Cohen-Shikora & Balota (2016)	2.58%	0.60	[0.04, 1.16]
Aschenbrenner & Balota (2019)	2.86%	0.65	[0.14, 1.16]
Mutter et al., (2005) Experiment 3	3.76%	0.66	[0.28, 1.03]
Tang et al. (in prep.)	2.73%	0.66	[0.13, 1.20]
Failes Unpublished	3.61%	0.69	[0.30, 1.09]
Cohen-Shikora et al. (2018) Unpublished Experiment 3	3.45%	0.72	[0.31, 1.14]
Cohen-Shikora et al. (2018) Experiment 2	3.78%	0.76	[0.39, 1.13]
Hutcheon & Spieler (2017) Experiment 1	2.06%	0.76	[0.09, 1.43]
Cohen-Shikora et al. (2018) Unpublished Experiment 4	3.26%	0.82	[0.38, 1.26]
Mutter et al., (2005) Experiment 1	1.51%	0.86	[0.02, 1.70]
Bugg et al. (2015) Experiment 2	2.26%	0.91	[0.28, 1.54]
Aschenbrenner & Balota (2017) Experiment 2	3.48%	0.92	[0.51, 1.33]
Nicosia & Balota (2020) Experiment 1	3.40%	0.98	[0.56, 1.41]
Nicosia & Balota (2020) Experiment 2	3.20%	0.99	[0.53, 1.44]
Gollan & Goldrick (2016)	3.07%	1.06	[0.59, 1.53]
Rey-Mermet & Gade (2018)	4.70%	1.26	[1.01, 1.51]
Bugg et al. (2011) Experiment 2	2.36%	1.31	[0.70, 1.91]
Cohen-Shikora et al. (2018) Experiment 1	2.60%	1.50	[0.95, 2.06]
Hutcheon & Spieler (2017) Experiment 3	2.00%	1.61	[0.92, 2.30]
RE Model	100.00%	0.69	[0.57, 0.81]

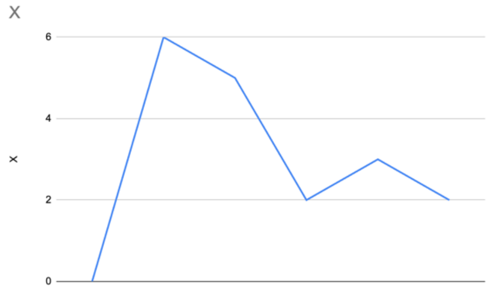
W3 Activity 2

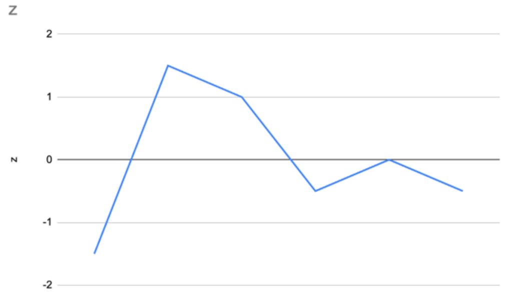
- complete the activity on your own
- discuss the logic behind your answers with a peer
- re-attempt the activity

properties of z-scores

$$z = \frac{X - \mu}{\sigma}$$

- shape of the distribution remains the same before and after z-scoring
- sum of z-scores?
  - always zero! why?
- mean of z-scores?
  - always zero! why?





$$\sum z = \sum \frac{X - \mu}{\sigma} = \frac{1}{\sigma} \sum (X - \mu) = \frac{1}{\sigma}(0) = 0$$

$$M_z = \frac{\sum z}{N} = \frac{0}{N} = 0$$

properties of z-scores

- variance of z-scores?
- always 1! why?

$$z_i = \frac{X_i \;-\; \mu}{\sigma}$$

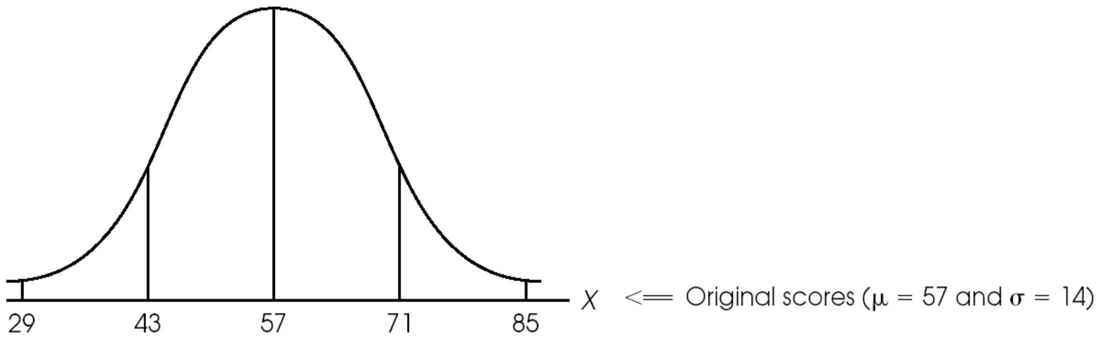
variance of z-scores = $\sigma_z{}^2 = \frac{\sum(z_i - M_z)^2}{N}$

$M_z = 0$, thus $\sigma_z{}^2 = \frac{\sum(z_i)^2}{N} = \frac{\sum(\frac{X-\mu}{\sigma})^2}{N}$

$= \frac{\frac{1}{\sigma^2}\sum(X-\mu)^2}{N} = \frac{1}{\sigma^2}\,(\sigma^2) = 1$

standardized scores

- z-scoring on original distribution and then obtaining scores on a predetermined μ and σ
- Joe got 43 on original test, where $\mu = 57$ and $\sigma = 14$. What should his score be on a new distribution with $\mu = 50$ and $\sigma = 10$?

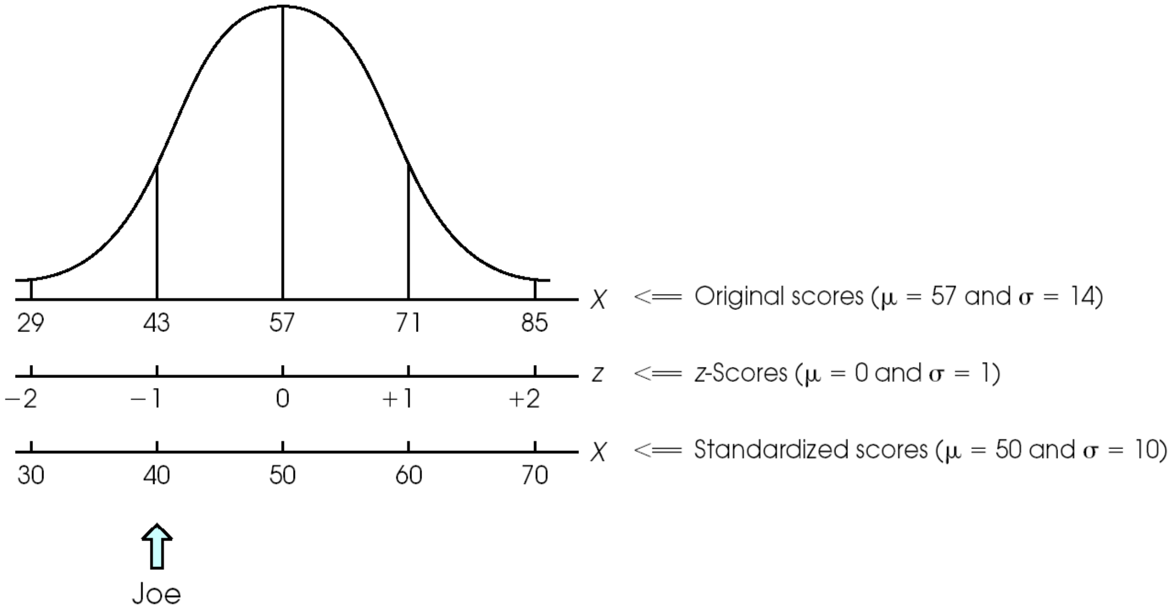


standardized scores

- compute Joe’s z-score on original distribution

$$z = \frac{X - \mu_1}{\sigma_1} = \frac{43 - 57}{14} = -1$$

- compute Joe’s score on new distribution $X = \sigma_2 z + \mu_2 = 10\,(-1) + 50 = 40$

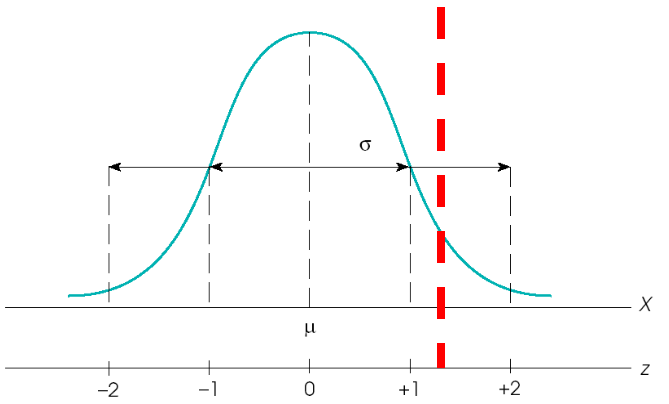


comparing apples and oranges

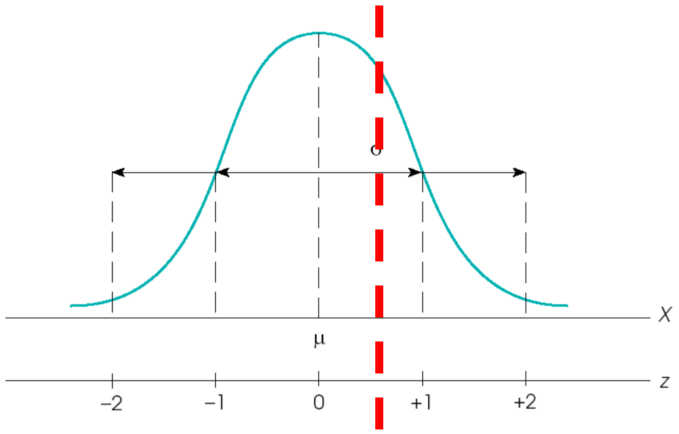
- Eric competes in two track events: standing long jump and javelin. His long jump is 49 inches, and his javelin throw was 92 ft. He then measures all the other competitors in both events and calculates the mean and standard deviation:
    - Long Jump: M = 44, s = 4
    - Javelin: M = 86ft, s = 10ft
- Which event did Eric do best in?

comparing apples and oranges

- we calculate Eric’s z-score on both events
- $z_{javelin} = (49 – 44\ )/4\ \ = 1.25$
- $z_{long\text{-}jump} = (92 – 86\ )/10\ \ = 0.6$



Javelin



Long Jump

next time

- deep dive into the normal distribution

optional: why (n-1)? degrees of freedom

- df = number of values that are *free to vary* in the calculation of a statistic
- for populations, we use the population mean (μ) to compute deviation scores (X - μ)
- however, for samples, μ is unknown and we estimate it using our sample mean *M*
- computing M restricts the scores that went into the calculation
  - why? because changing even a single score would change M
  - if M is known, you only need to know n-1 scores to find the last score
  - only n-1 scores are free to vary once M is known

optional: understanding df

- Bessel’s correction (n-1 instead of n in sample variance formula)
  - sample variance is an estimate of population variance (off by a factor of $\frac{n-1}{n}$)
  - mathematical proof
  - video (+ proof)
- degrees of freedom
  - the term’s origin is based in physical systems (e.g., a pulley)
  - the video describes how to visualize data in a visual space