DATA ANALYSIS

Week 3: Variability

logistics: quiz 2 / problem set #1

- quiz 2
  - bar graph / histogram question was regraded
- problem set #1
  - going forward, please submit a PDF of your document with link to sheet as before

logistics: midterm 1

- Feb 23rd 2024 (Friday)
- content: Week 1 through Week 5
- two parts
  - in-person conceptual portion (quiz-like)
    - closed book
  - take-home computational portion (problem set-like: due Monday)
    - open book but NOT open person
- practice questions will be made available before Week 5

today's agenda

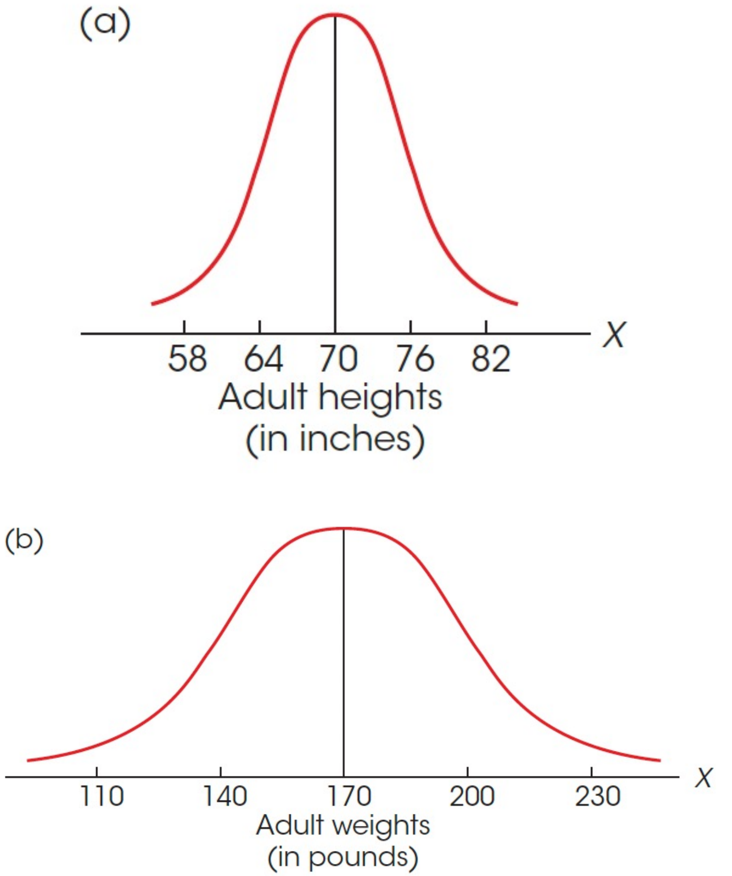
- variability
- z-scores

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

variability

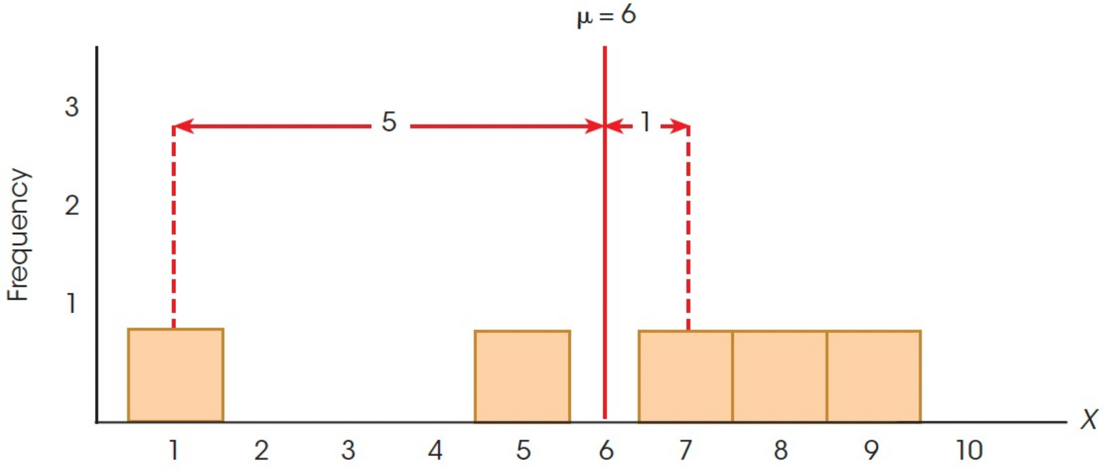
- variability describes the spread of scores in a distribution
- measures of variability
  - range = maximum – minimum
  - variance = mean squared error from the mean (MSE or σ^2) = average of ***squared*** distances/errors from the mean
  - standard deviation = root mean squared error from the mean (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

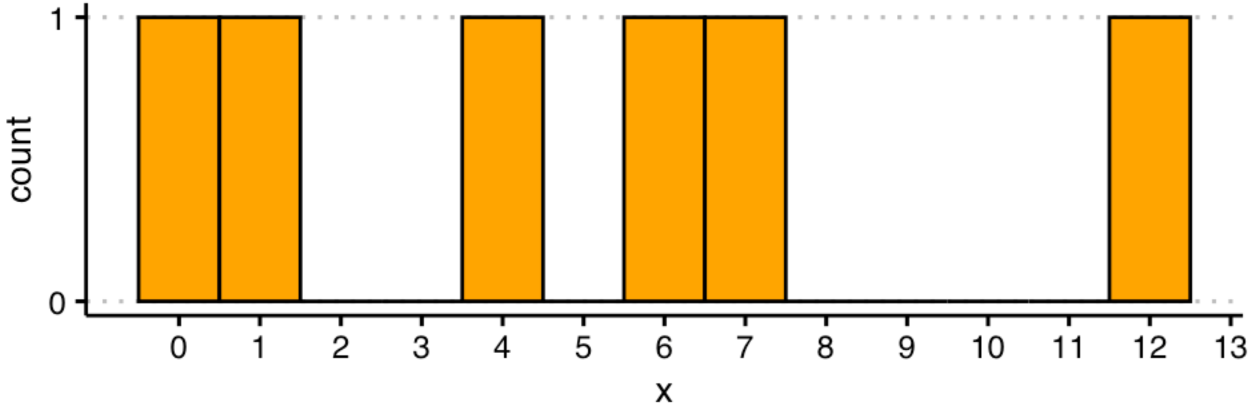
- we can estimate/calculate the mean
- the largest score is 5 points away
- the smallest score is 1 point away
- on average, scores are likely $\frac{5+1}{2}$ away = 3 points away
- what is our actual estimate of standard deviation for these scores?

$$\sqrt[2]{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}} = \sqrt{\frac{25+1+1+4+9}{5}} = 2.83$$



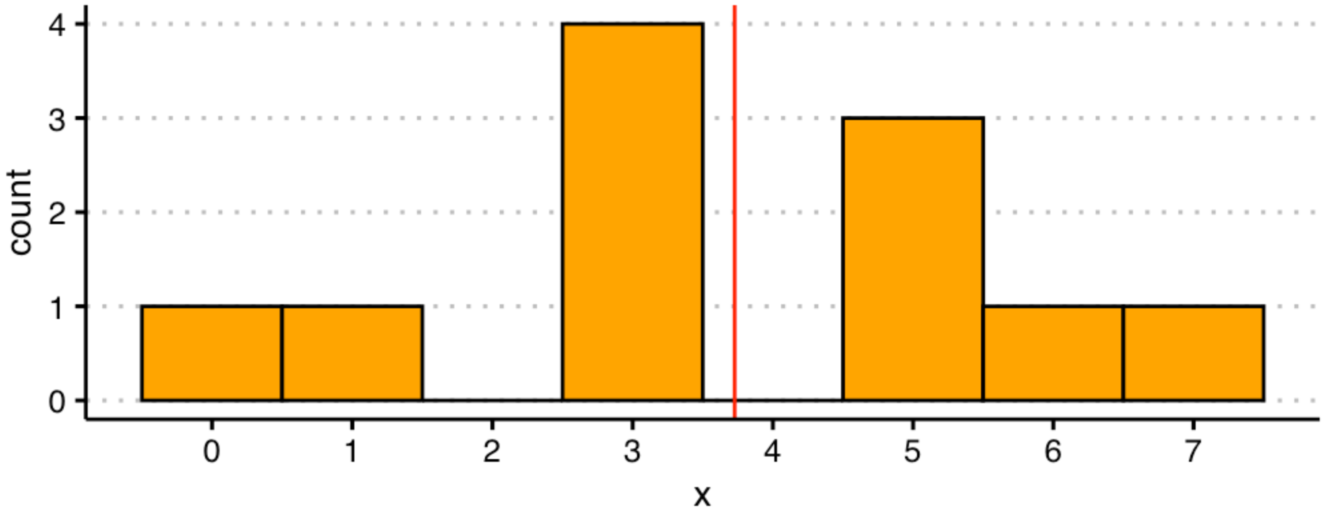
activity

- 6 scores: 12, 0, 1, 7, 4, and 6
- calculate the mean
- visually estimate the standard deviation



activity

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



SSE: definitional vs. computational formulas

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

definitional formula

computational formula

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

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

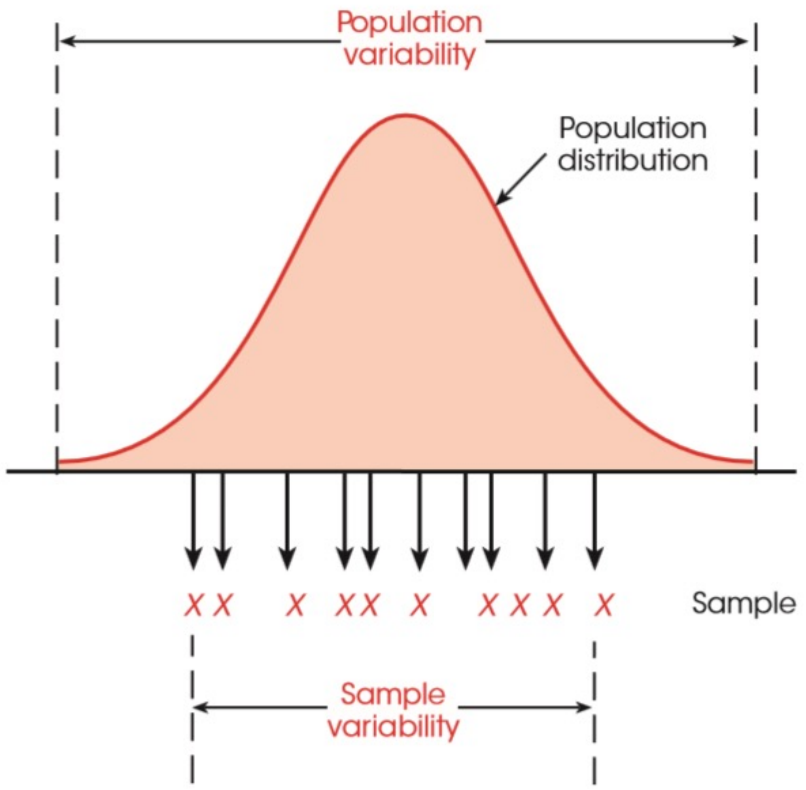
$$= \sum X^2 - \frac{(\sum X)^2}{N}$$

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

questions?

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
1	0	0
2	0	3
3	0	9
4	3	0
5	3	3
6	3	9
7	9	0
8	9	3
9	9	9

B2 fx =

	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

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

- average variance is LOWER than the population variance: variance is a biased 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

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

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

an example

- if the mean of quiz scores for 5 students is 9 points and four students' scores are 8, 10, 8, and 9, what is the score of the fifth student?

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) =

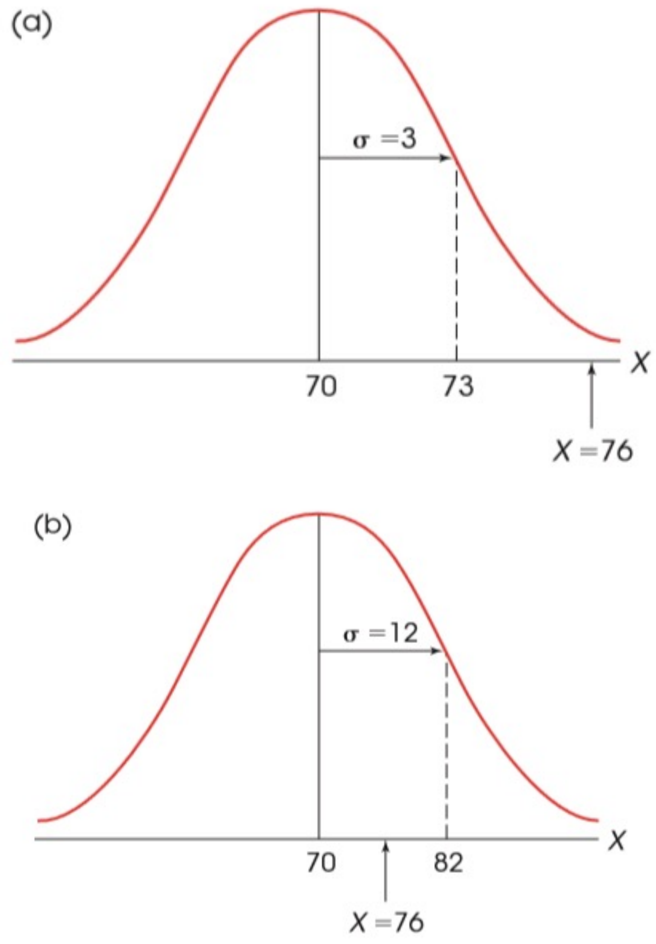
$$\sqrt{\frac{\sum(X-M)^2}{n-1}} = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{SS}{df}}$$

questions?

- explore the variability sheet

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

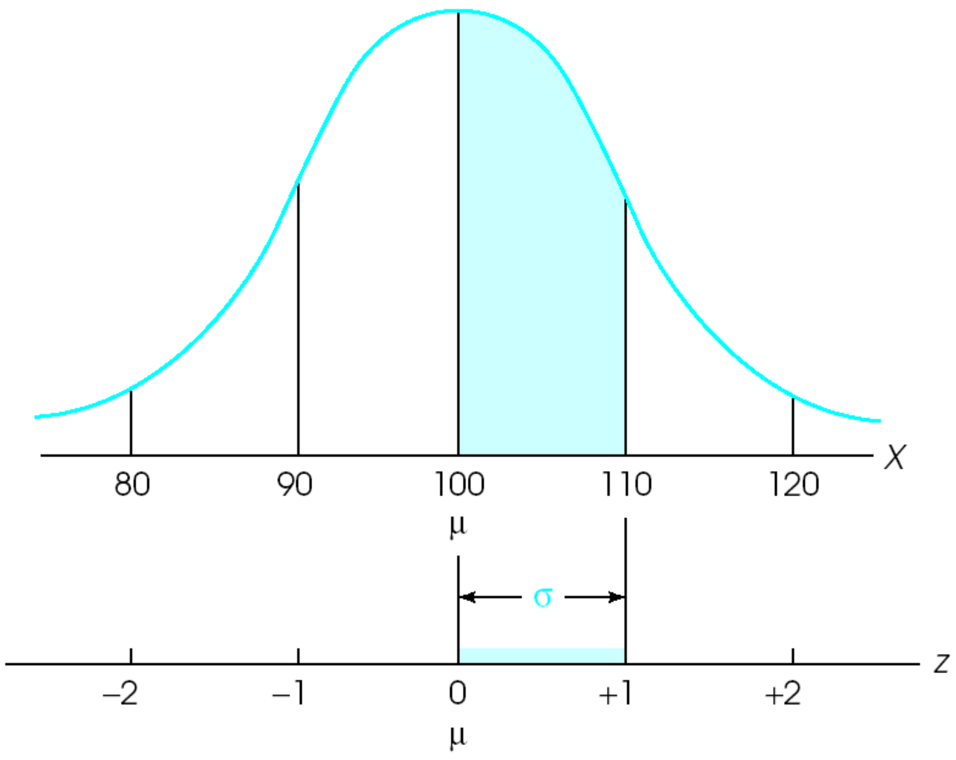


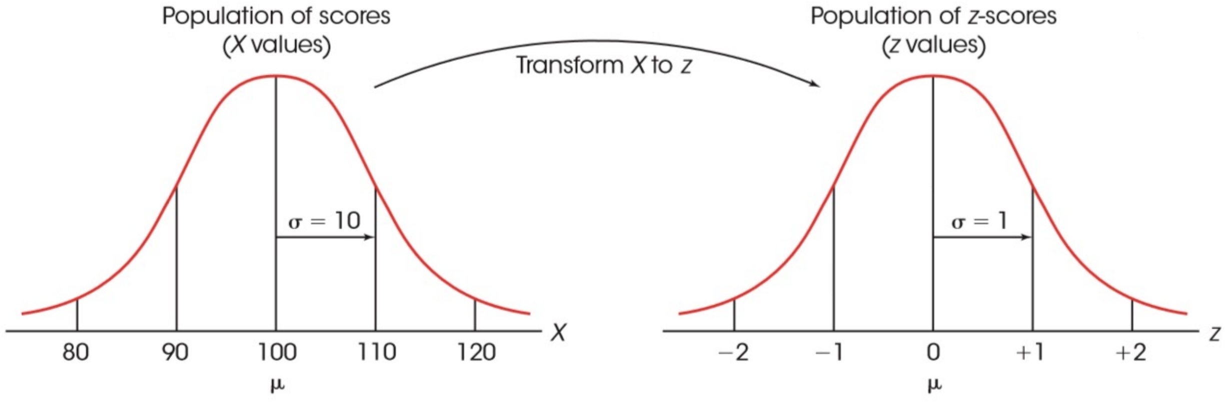
z-scores

- 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

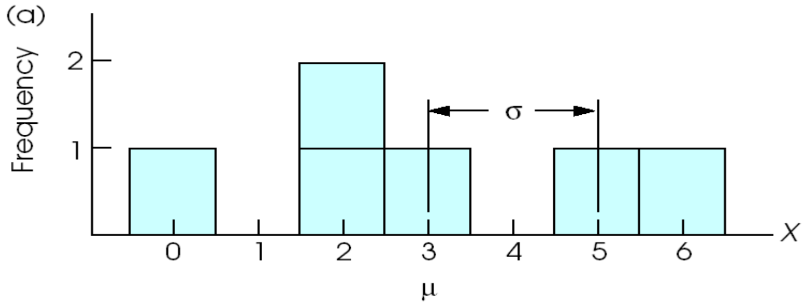




calculating z-scores

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

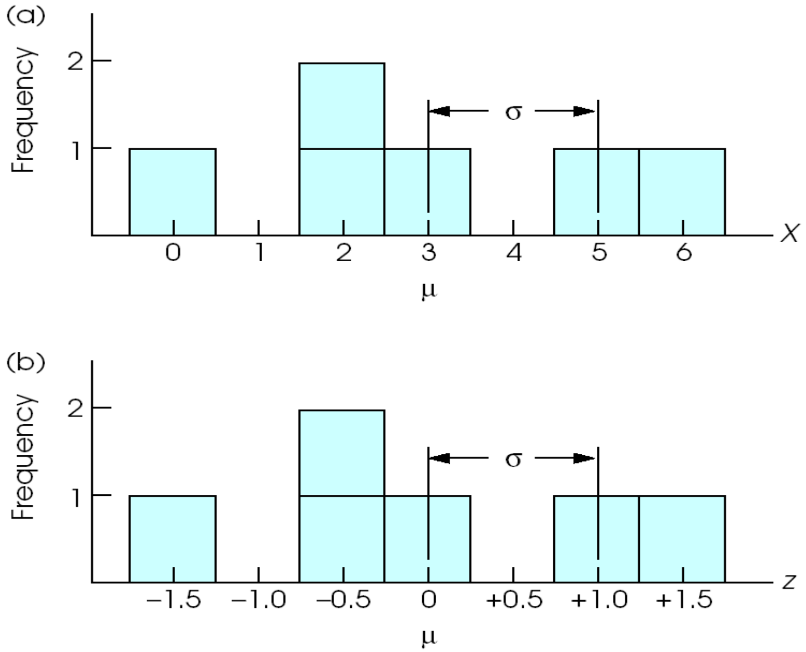
- six scores, calculate μ, $\boldsymbol{\sigma}$, and z



calculating z-scores

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

- six scores, calculate μ, σ, and z

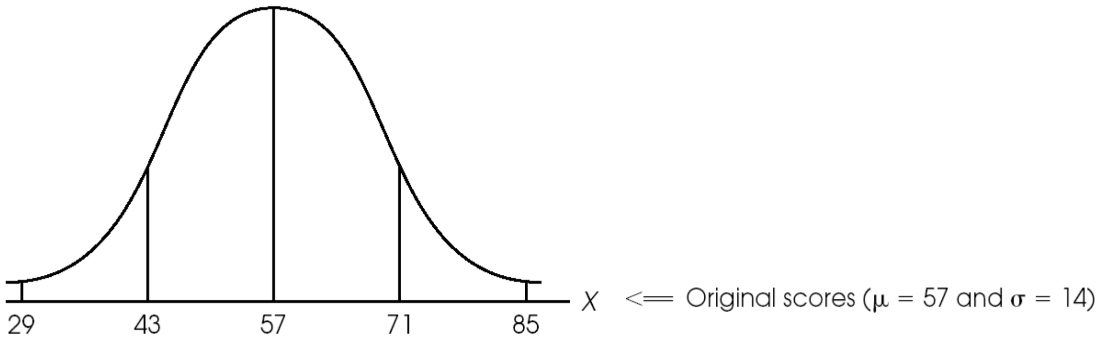


X	mu	X-mu	squared_errors	MSE	RMSE	z
0	3	-3	9	4	2	-1.5
6		3	9			1.5
5		2	4			1
2		-1	1			-0.5
3		0	0			0
2		-1	1			-0.5

solution sheet

standardized scores

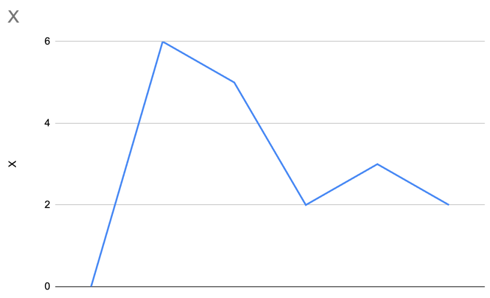
- z-scoring on original distribution and then obtaining scores on a predetermined μ and σ
- Joe got 43 on original test. Where $\mu = 57$ and $\sigma = 14.$ What should his score be on a new distribution with $\mu = 50$ and $\sigma = 10?$

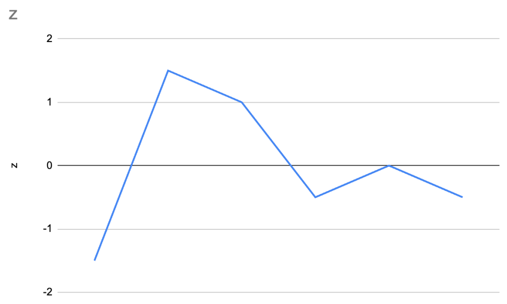


properties of z-scores

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

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





properties of z-scores

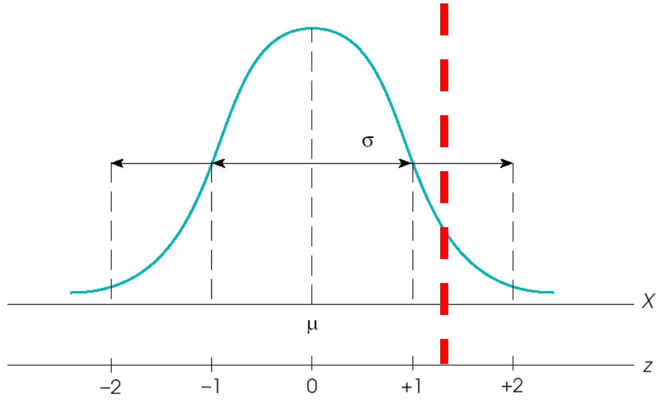
- variance of z-scores?
- always 1! why?

comparing apples and oranges

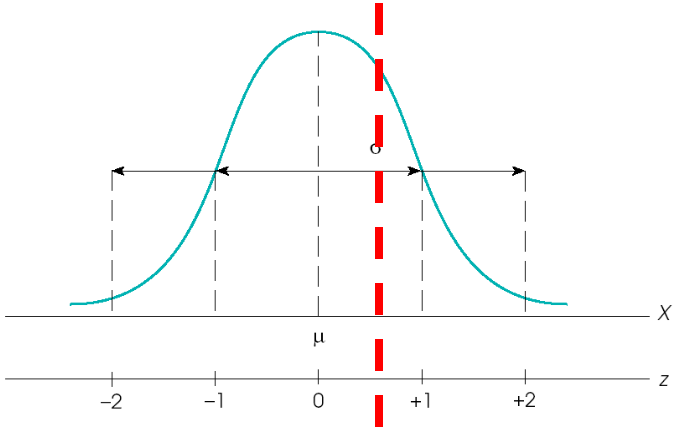
- Eric competes in two track events: standing long jump and javelin. His long jump is 49 inches, and his javelin throw was 92 ft. He then measures all the other competitors in both events and calculates the mean and standard deviation:
  - Long Jump: $M = 44$, $s = 4$
  - Javelin: $M = 86\text{ft}$,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

- **before** class
  - *watch*: Variability and z-scores
  - *prep*: chapter 6 (specific sections – see course website)
  - *try*: problem set #2 (chapter 4 and 5 problems)
- **during** class
  - deep dive into the normal distribution