DATA ANALYSIS

Week 3: Normal distribution

logistics: problem sets

- problem set #1 revisions
  - please explain your work. the goal is not to get you to the correct answer, you already have that. I would like to see what you've learned
- problem set #2
  - you MUST show your work for variance/sd calculations
  - you can use STDEV.P and STDEV.S for checking, but I need to see all calculations
  - for Qs about normal distributions, please provide screenshots from table/website

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

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

resources on 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

today's agenda

review: z-scores

the normal distribution

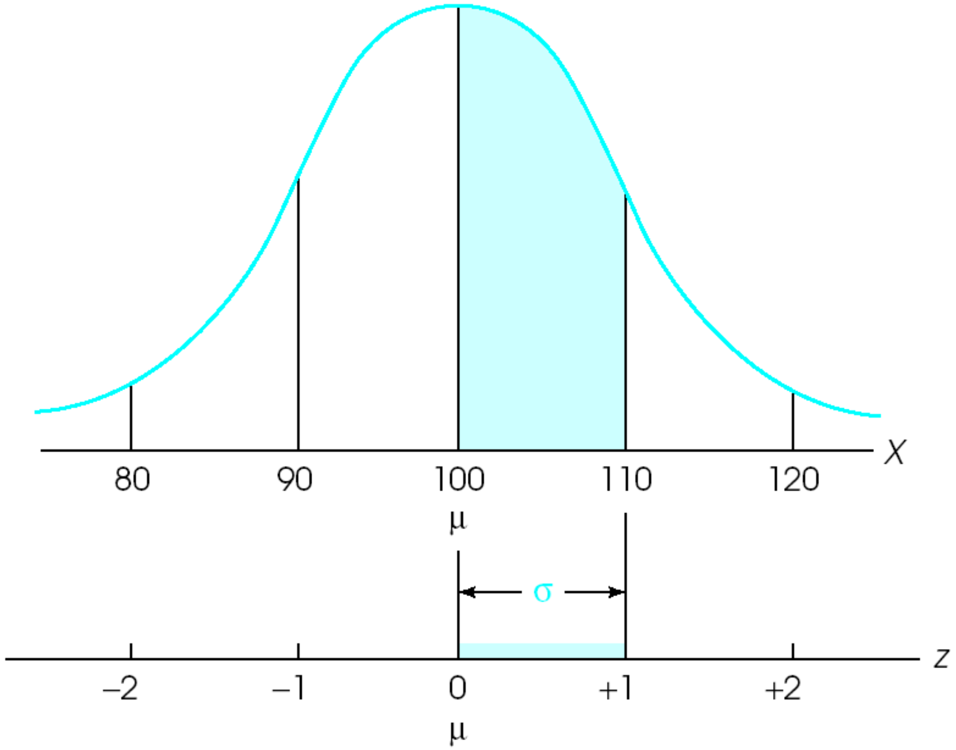
z-scores

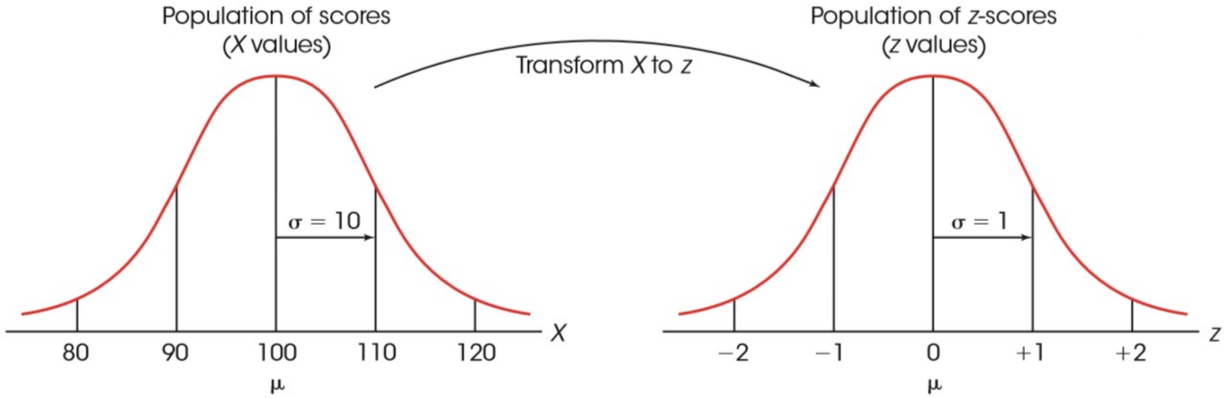
- z-scores are a way to understand how far away a score is from the mean, in standard deviation units

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

  - 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

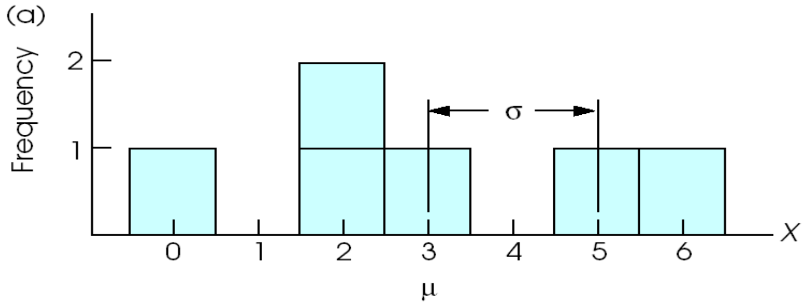


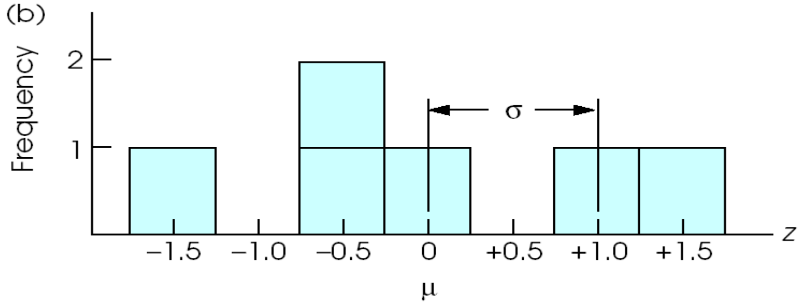


calculating z-scores

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

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





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

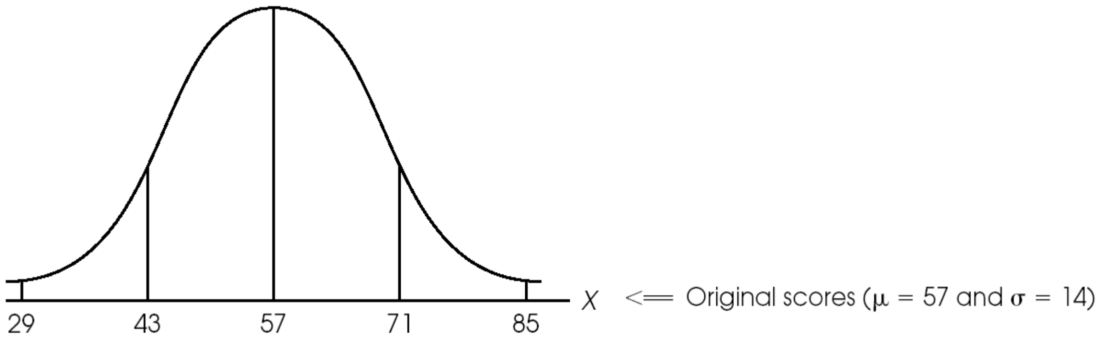
solution sheet

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

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 $\sigma = 10?$



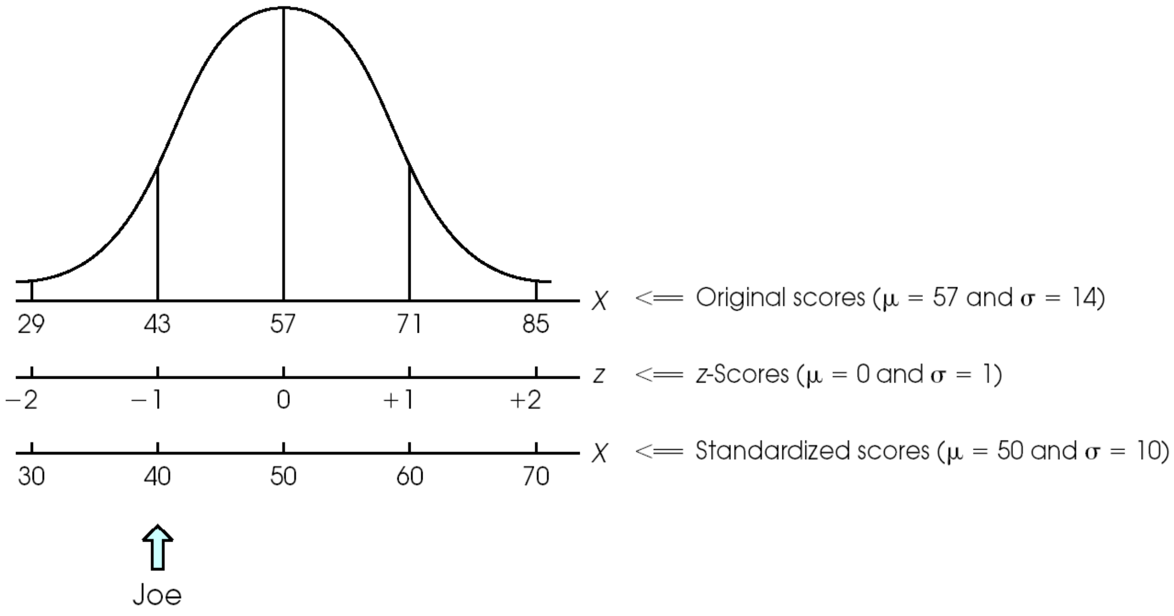
standardized scores

- 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

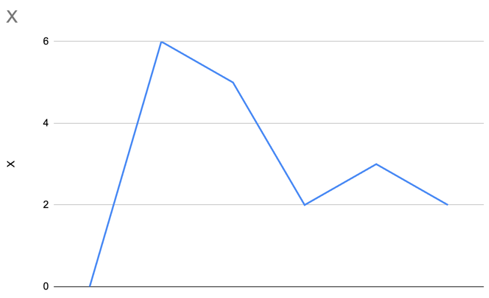
$X = \sigma_2 z + \mu_2 = 10\,(-1) + 50 = 40$

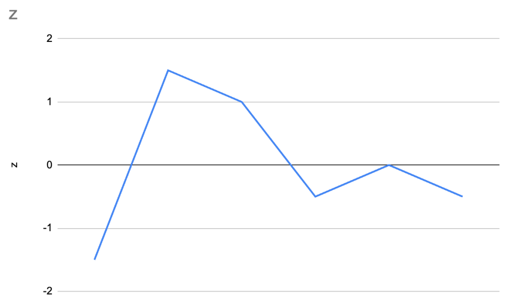


properties of z-scores

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

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

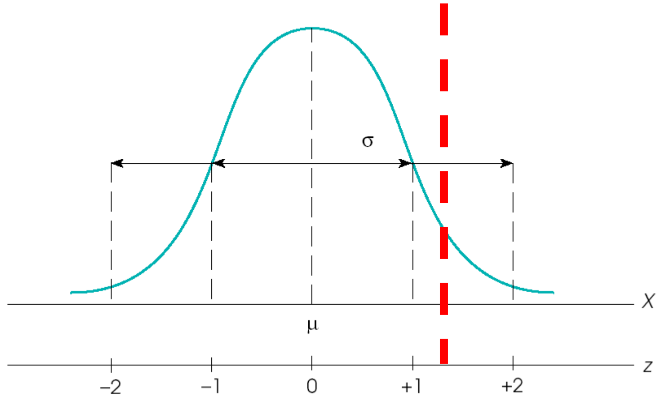
$= \frac{\frac{1}{\sigma^2}\sum(X-\mu)^2}{N} = \frac{1}{\sigma^2}(\sigma^2) = 1$

comparing apples and oranges

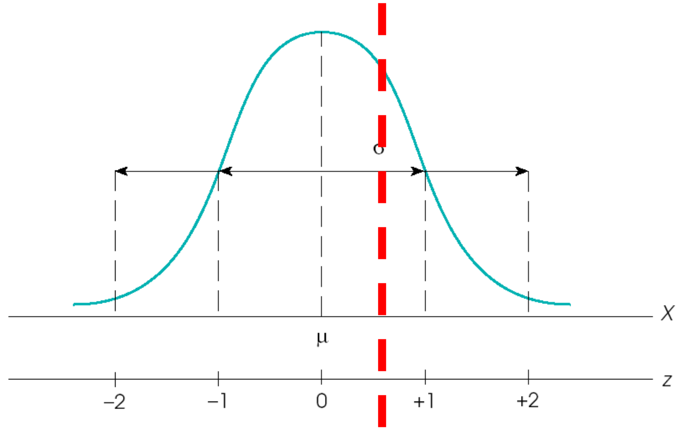
- Eric competes in two track events: standing long jump and javelin. His long jump is 49 inches, and his javelin throw was 92 ft. He then measures all the other competitors in both events and calculates the mean and standard deviation:
  - Long Jump: M = 44, s = 4
  - Javelin: M = 86ft, s = 10ft
- Which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

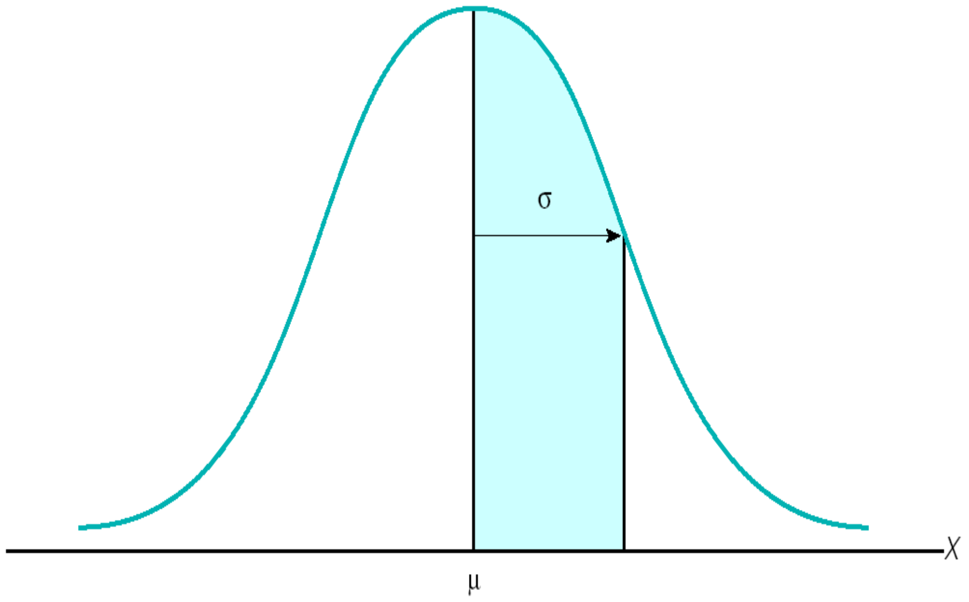


Long Jump

the normal distribution

$$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

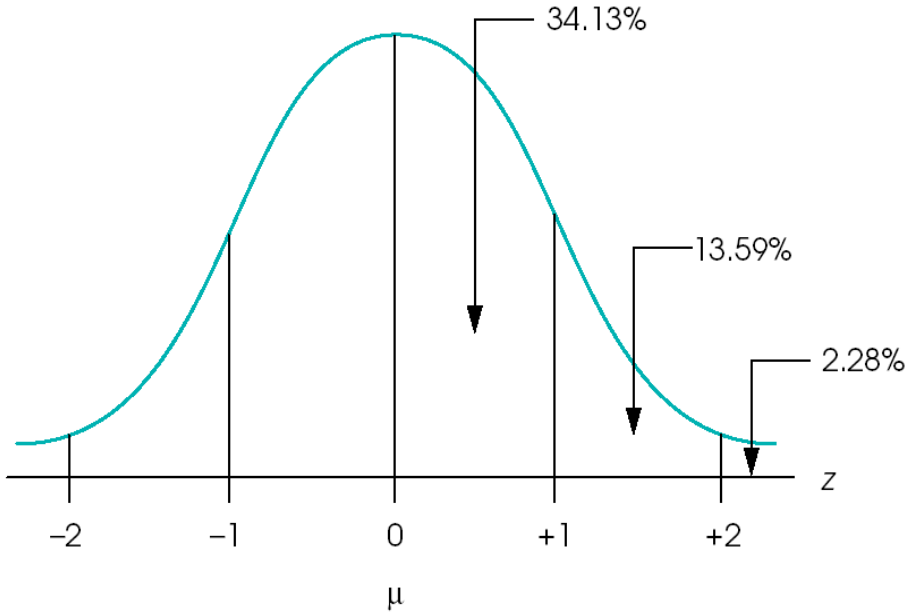
- population distributions typically take the form of a normal distribution
  - symmetric, unimodal, "bell-shaped"
- after converting the normal distribution scores to z-scores, z-scores are often used to identify parts of a normal distribution (μ= 0, σ= 1)
- proportions of areas within the normal distribution can be quantified using z-scores



the normal distribution

$$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

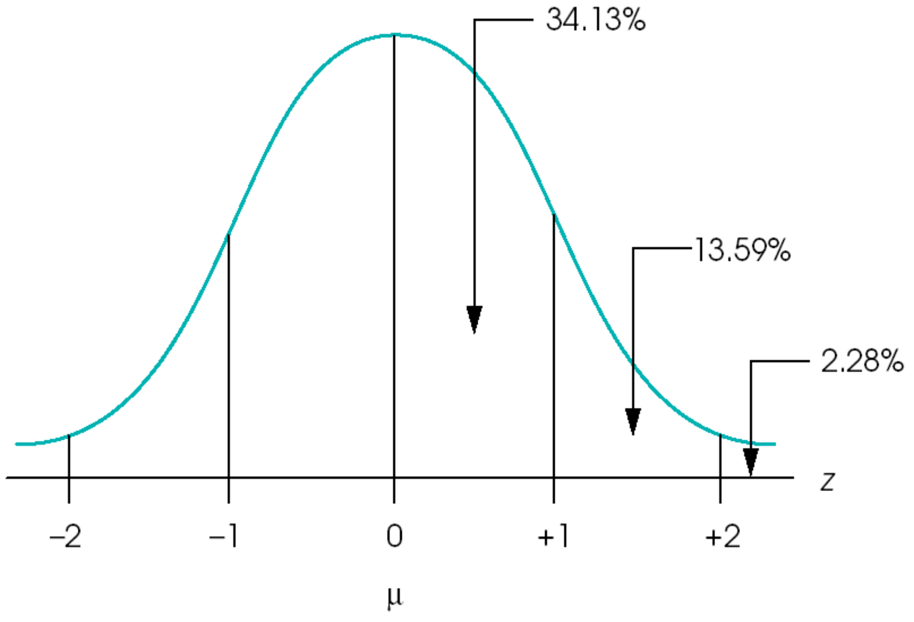
- population distributions typically take the form of a normal distribution
  - symmetric, unimodal, "bell-shaped"
- after converting the normal distribution scores to z-scores, z-scores are often used to identify parts of a normal distribution (μ= 0, σ= 1)
- proportions of areas within the normal distribution can be quantified using z-scores



area under the curve

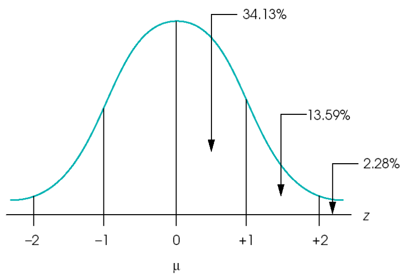
$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

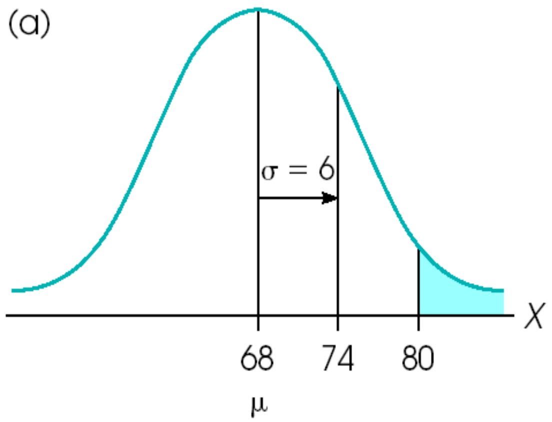
- all normal distributions have the same proportions of data/area in specific locations on the curve
- the normal distribution is symmetrical, i.e., the proportions on both sides of the mean are identical
  - what % of the scores are above the mean?
  - what % of the scores are above 2 standard deviations?
- ~ 68% of scores fall between z-scores of -1 and +1
- ~ 95% fall between z-scores of -2 and +2
- ~ 99% fall between z-scores of -3 and +3

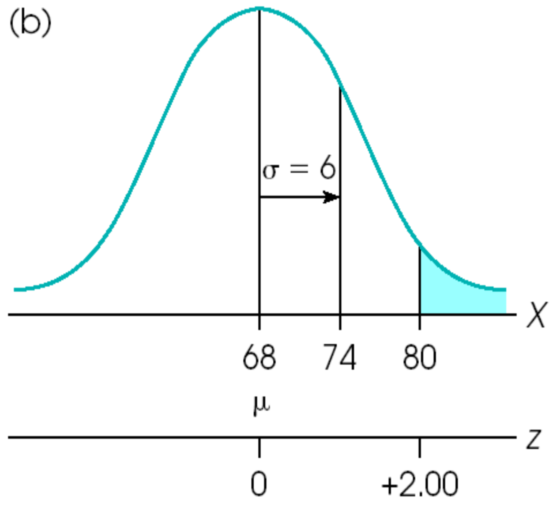


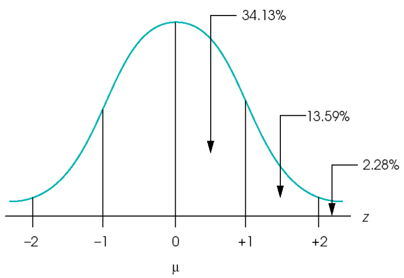
example

- body height has a normal distribution, with μ= 68, and σ= 6.
- if we select one person at random, what is the probability for selecting a person taller than 80?
  - represent the problem graphically
  - convert to z-scores, 80 is 2 σ
  - all scores **<u>above</u>** 2 σ: 2.28%



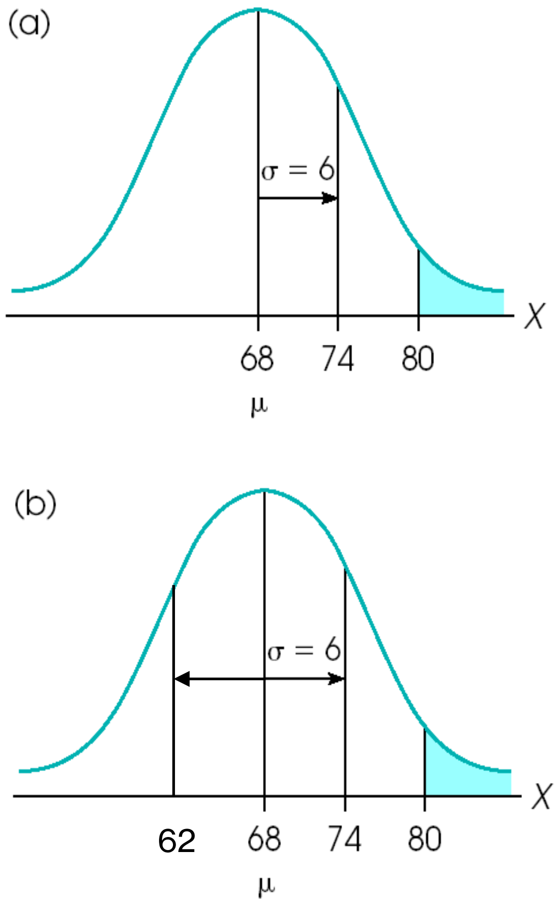






example

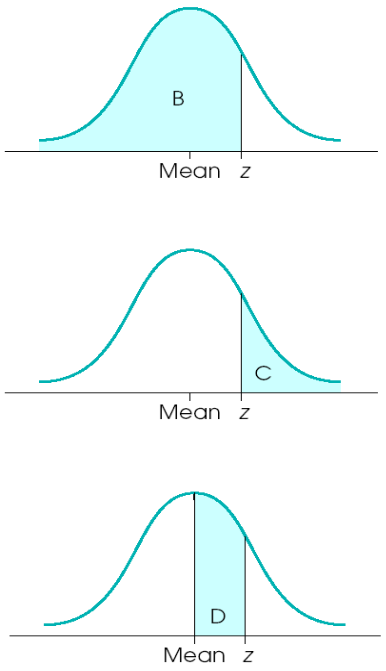
- body height has a normal distribution, with μ= 68, and σ= 6.
- what % of people have heights between 62 and 74?
  - represent the problem graphically
  - convert to z-scores, 62 is -1σ and 74 is 62 is +1σ
    - all scores **between** -σ to +σ
    - 34.13 + 34.13 = 68.26



unit normal table / calculator

- questions about whole z-scores (±1, ±2, etc.) are easily gleaned from the distribution, but estimates for fractional z-scores are trickier to obtain via eyeballing
- unit normal tables contain proportion estimates for the full scale of possible z-scores
  - column A: the z-score (vertical line)
  - column B (body): the larger section created by the z-score
  - column C (tail): smaller section created by z-score
  - column D: section between mean and z-score
- available in several places online!
  - [full table]
  - [visual calculator]

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and z
0.00	.5000	.5000	.0000
0.01	.5040	.4960	.0040
0.02	.5080	.4920	.0080
0.03	.5120	.4880	.0120
0.21	.5832	.4168	.0832
0.22	.5871	.4129	.0871
0.23	.5910	.4090	.0910
0.24	.5948	.4052	.0948
0.25	.5987	.4013	.0987
0.26	.6026	.3974	.1026
0.27	.6064	.3936	.1064
0.28	.6103	.3897	.1103
0.29	.6141	.3859	.1141
0.30	.6179	.3821	.1179
0.31	.6217	.3783	.1217
0.32	.6255	.3745	.1255
0.33	.6293	.3707	.1293
0.34	.6331	.3669	.1331

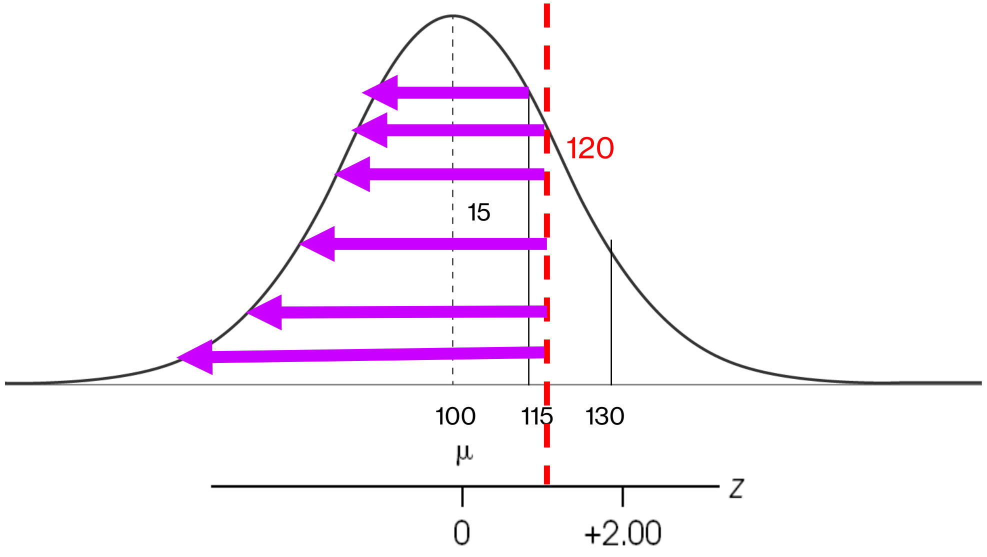


probabilities from scores: example 1

- for an IQ test, the known population parameters are μ= 100, and σ= 15. What is the probability of randomly selecting an individual with an IQ score of **<u>less than 120</u>**?
- represent the problem graphically
- transform X into z
- look up full table (column B) or visual calculator

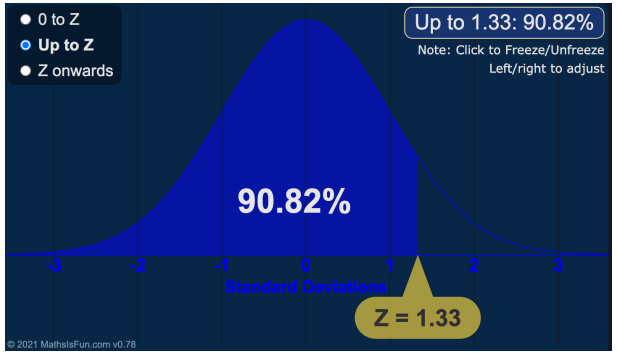
probabilities from scores: example 1

- for an IQ test, the known population parameters are μ= 100, and σ= 15. What is the probability of randomly selecting an individual with an IQ score of **<u>less than 120</u>**?
- represent the problem graphically
- transform X into z = $\frac{X-\mu}{\sigma} = \frac{120-100}{15} = 1.33$
- full table (column B) or visual calculator
- proportion or probability = .9082
- percentage = 90.82%



Unit Normal Table

(A) z	(B) Body	(C) Tail	(D) Mean & z
1.25	0.8943	0.1057	0.3943
1.26	0.8961	0.1039	0.3961
1.27	0.8979	0.1021	0.3979
1.28	0.8997	0.1003	0.3997
1.29	0.9015	0.0985	0.4015
1.30	0.9032	0.0968	0.4032
1.31	0.9049	0.0951	0.4049
1.32	0.9066	0.0934	0.4066
1.33	0.9082	0.0918	0.4082
1.34	0.9099	0.0901	0.4099

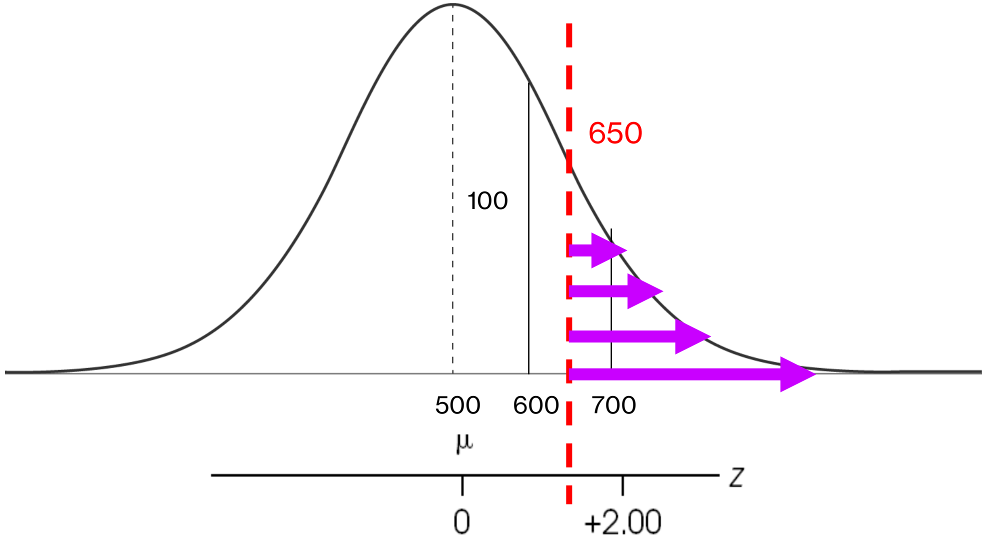


probabilities from scores: example 2

- for a normal distribution with μ= 500, and σ= 100, find the probability of selecting an individual whose score is **above 650**.
- represent the problem graphically
- transform X into z
- full table (column C) or visual calculator

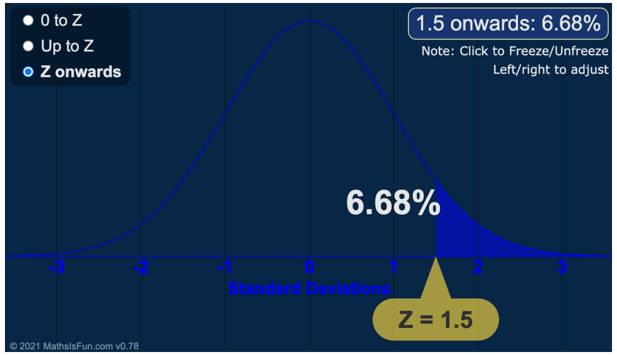
probabilities from scores: example 2

- for a normal distribution with μ= 500, and σ= 100, find the probability of selecting an individual whose score is **above 650**.
- represent the problem graphically
- transform X into z = 1.5
- full table (column C) or visual calculator
- proportion or probability = .0668
- percentage = 6.68%



Unit Normal Table

(A) z	(B) Body	(C) Tail	(D) Mean & z
1.50	0.9332	0.0668	0.4332
1.51	0.9345	0.0655	0.4345

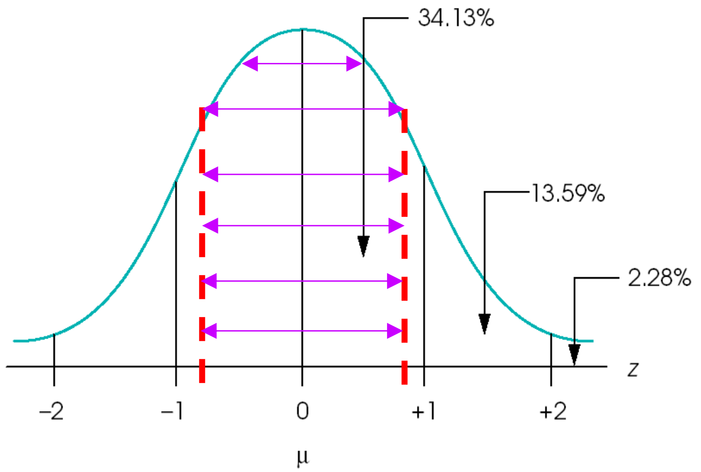


z-scores from proportions: example 3

- what z-score values form the boundaries that separate the middle 60% of the distribution?
- represent the problem graphically
- transform X into z
- full table (column D) or visual calculator

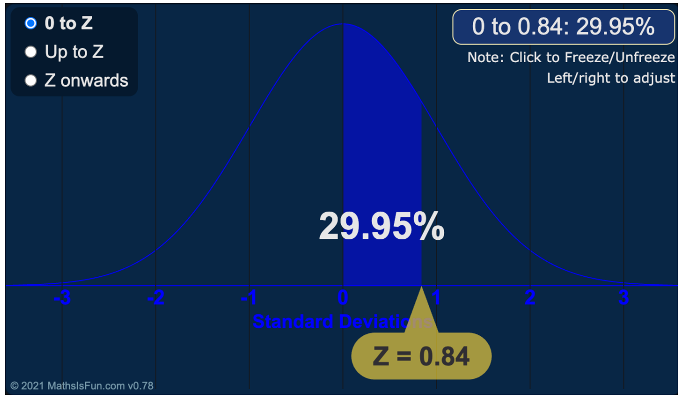
z-scores from proportions: example 3

- what z-score values form the boundaries that separate the middle 60% of the distribution?
- represent the problem graphically
  - somewhere less than 34% on both sides
- transform X into z
  - find z-value for p close to ±0.30
- full table (column D) or visual calculator
- z = -0.84 to +0.84



Unit Normal Table

(A) z	(B) Body	(C) Tail	(D) Mean & z
0.75	0.7734	0.2266	0.2734
0.76	0.7764	0.2236	0.2764
0.77	0.7793	0.2207	0.2793
0.78	0.7823	0.2177	0.2823
0.79	0.7852	0.2148	0.2852
0.80	0.7881	0.2119	0.2881
0.81	0.791	0.209	0.291
0.82	0.7939	0.2061	0.2939
0.83	0.7967	0.2033	0.2967
0.84	0.7995	0.2005	0.2995
0.85	0.8023	0.1977	0.3023

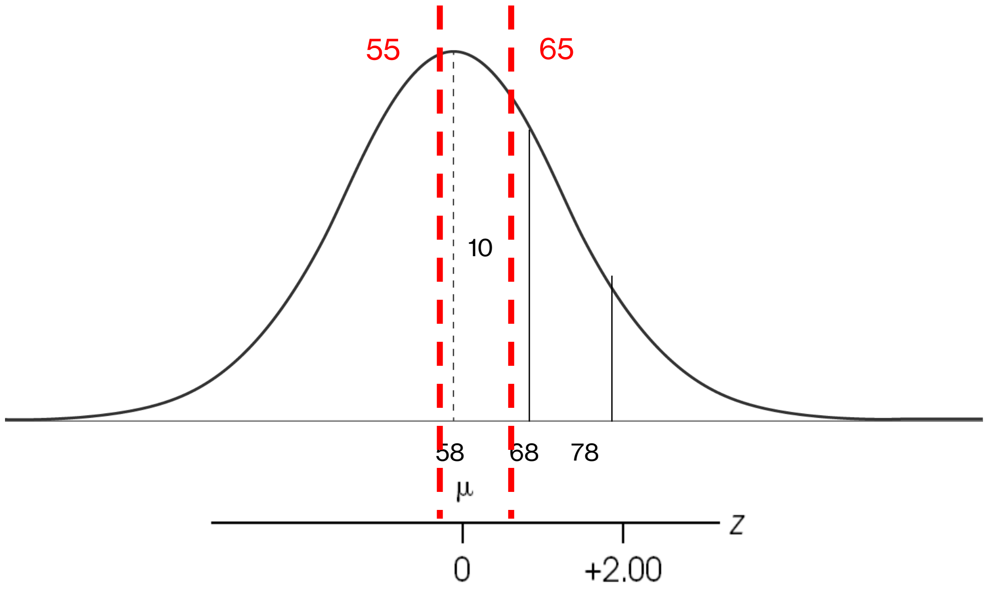


proportions between two scores

- highway department survey, average speed μ= 58 mph, and σ= 10. What proportion of cars travel between 55 and 65 miles per hour?
- represent the problem graphically
- transform into z scores
- full table (column D) or visual calculator

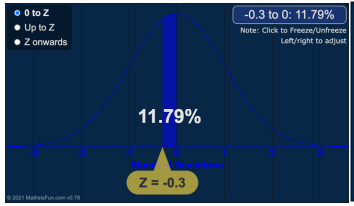
proportions between two scores

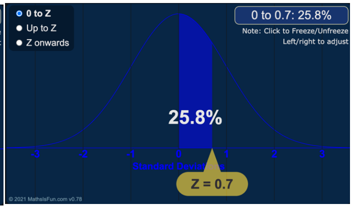
- highway department survey, average speed μ= 58 mph, and σ= 10. What proportion of cars travel between 55 and 65 miles per hour?
- represent the problem graphically
- transform into z scores
    - z for X = 55 = (55-58) / 10 = -0.3
    - z for X = 65 = (65-58)/10 = 0.7
- full table (column D) or visual calculator
    - find -0.3 to mean : 0.1179
    - find mean to 0.7: 0.258
- total = .3759 or 35.79% cars



Unit Normal Table

(A) z	(B) Body	(C) Tail	(D) Mean & z
0.30	0.6179	0.3821	0.1179
0.70	0.758	0.242	0.258





review activity

- a researcher is interested in the relationship between performance in a science course and performance in a history course. The researcher believes that intelligence is a general ability. Thus, one would expect that those students who do well in history would also do well in science. The researcher randomly selects a group of 10 students from all the students who are taking both a history and a science course, and records the number of errors the students made on their final exams (both exams have 100 total points and errors do not have any partial scoring).
- answer the questions

next time

- **before** class
  - *prep*: Ch 15 and Ch 16 (specific parts – see website)
  - *try*: week 3 quiz
  - *submit*: problem set #2
- **during** class
  - building models again