EDP308: STATISTICAL LITERACY

The University of Texas at Austin, Fall 2020
RAZ: Rebecca A. Zárate, MA

Overview

- All the Z's
  - Calculating a Z-score
    - Z-score for one *person*
  - Looking up on the Unit Normal Table for Z-scores
  - 68%, 95%, 99% Rule
- Calculating a Z-test
    - Z-test for one *mean*
  - Looking up on the Unit Normal Table for Z-scores
  - Comparing the Probability of Z-score (one person) vs. Z-statistic (one mean)
- Real World Example: Flint, MI
- Unit Normal table for Z-scores Tips
- Z-scores, Z-tests, and Probability in R

All the Z’s

Different Z's

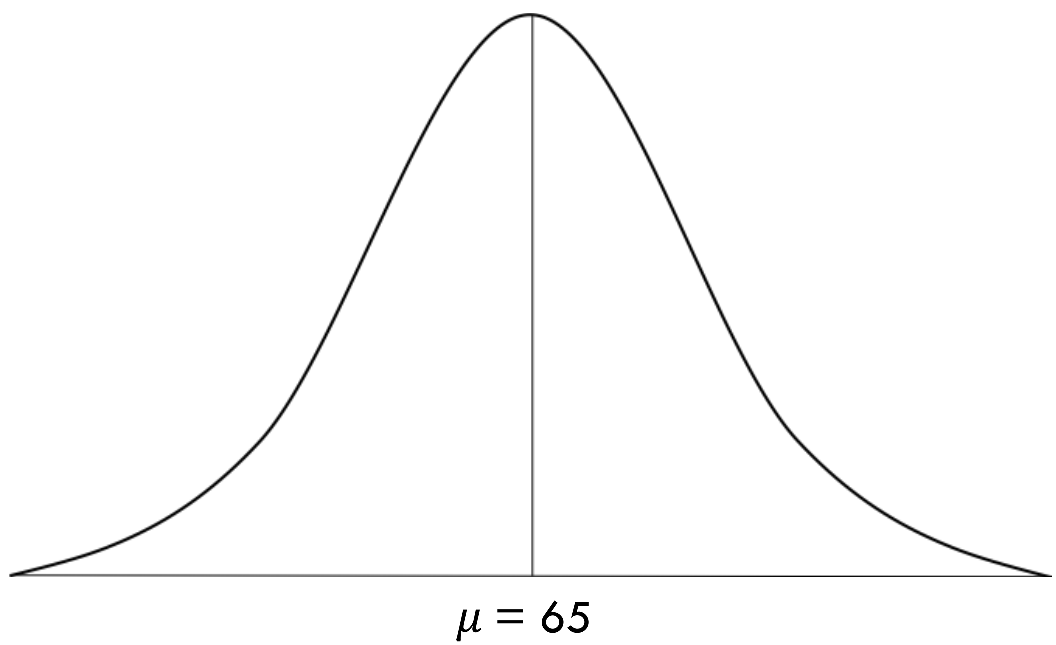
- There are different z-things… They can all mean the same thing… And some people use them interchangeably…
  - Z-score
    - The value that corresponds to how many standard deviations one person's raw score is above or below the mean
      - Ex. How is this one student (x) performing compared to the class average?
      - Ex. What is the probability of seeing a student with this score?
  - Z-test
    - A statistical test you calculate to determine a z-statistic, basically a z-score, but here we are testing to see how many standard deviations a <u>mean</u> is above or below another known mean
      - Ex. How is this one school ($\bar{x}$) performing compared to the national average?
      - Ex. What is the probability of seeing a school average with this score?
  - Z-statistic
    - The value you calculated in the z-test, again, basically a z-score but for a mean rather than one person

Back to Probability of x (Z-score)

- If the average height of women is μ = 65" (5'5), σ = 3.5, what is the probability of observing ONE woman who is 68" (5'8) or taller?
  - Which direction are we looking for?

$$P(x > 68) = ?$$

$$z - score = \frac{x - \mu}{\sigma}$$



Back to Probability of x (Z-score)

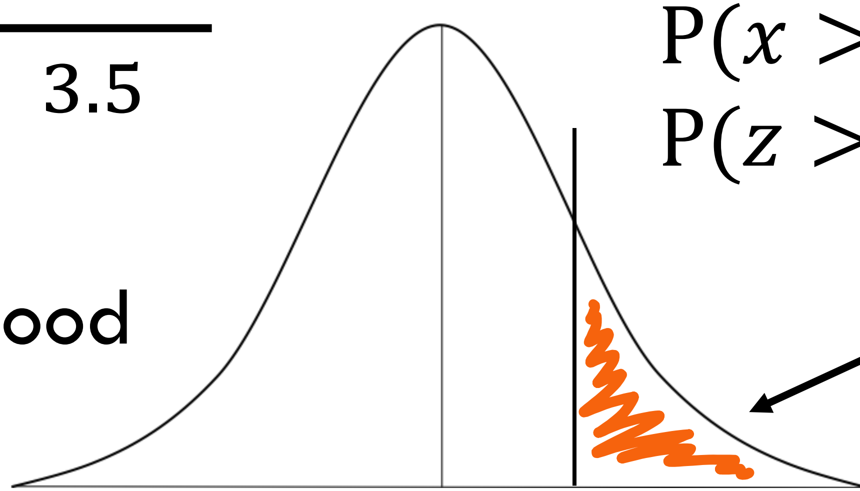
- If the average height of women is μ = 65" (5'5), σ = 3.5, what is the probability of observing ONE woman who is 68" (5'8) or taller?

$$\underset{\text{z-score}}{.86} = \frac{68-65}{3.5}$$

$$P(x > \overset{\text{raw score}}{68}) = .19$$

$$P(z > \underset{\text{z-score}}{.86}) = .19$$

19% likelihood



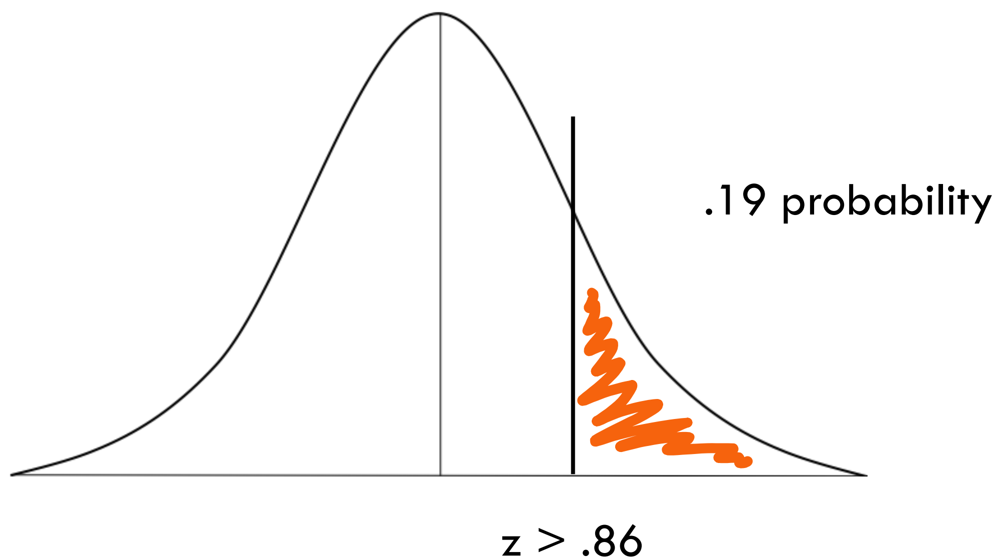
probability

z > .86

Back to Probability of x (Z-score)

- If the average height of women is μ = 65" (5'5), σ = 3.5, what is the probability of observing ONE woman who is 68" (5'8) or taller?
  - So observing one woman who is 68 inches (5'8) or taller is pretty likely, we know there are tall women out there, so it's not too uncommon or unlikely.

19% chance of seeing a woman that is 5'8 or taller

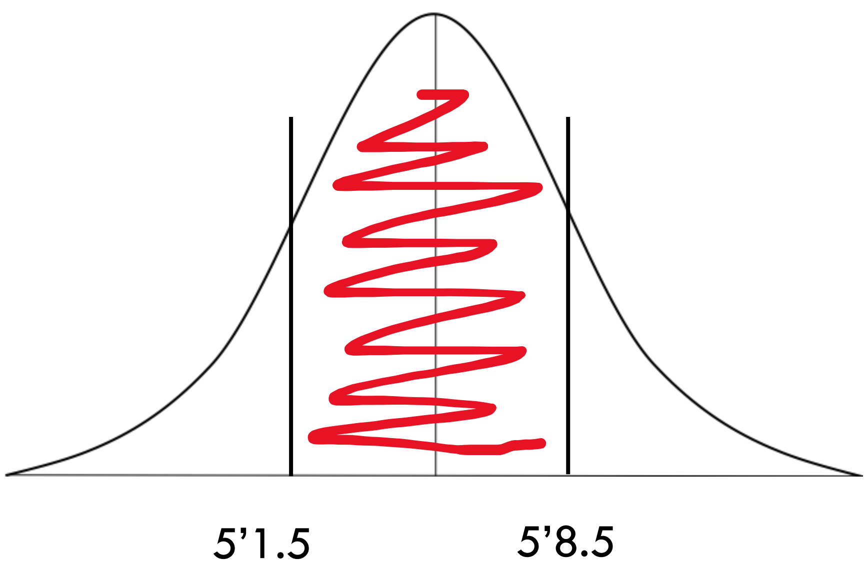


Between Two Values

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

- If the average height of women is μ = 65 inches, σ = 3.5, is the probability of seeing one woman that is between 61.5 inches (~5' 1½) and 68.5 (~5'8½) tall?

$$P(61.5 < x < 68.5) = ?$$



Before even starting the problem, how likely do ***you think*** it is to see a woman between 5'1.5 and 5'8.5?

Super likely?
Super rare?

Between Two Values

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

- If the average height of women is μ = 65 inches, σ = 3.5, is the probability of seeing one woman that is between 61.5 inches (~5' 1½) and 68.5 (~5'8½) tall?

raw score   mean

$$-1 = \frac{61.5 - 65}{3.5}$$

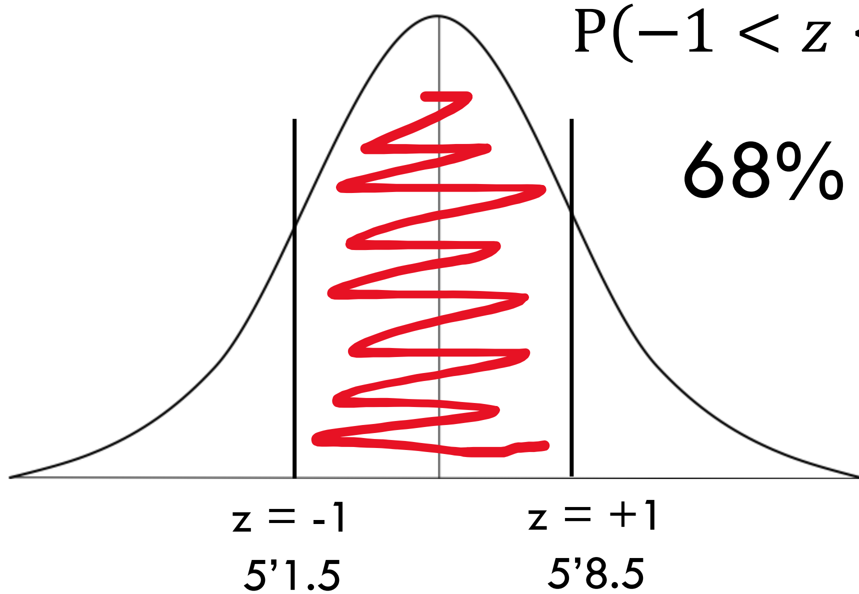
z-score   standard dev

$$+1 = \frac{68.5 - 65}{3.5}$$

$$P(61.5 < x < 68.5) = .68$$

$$P(-1 < z < +1) = .68$$

68% likelihood

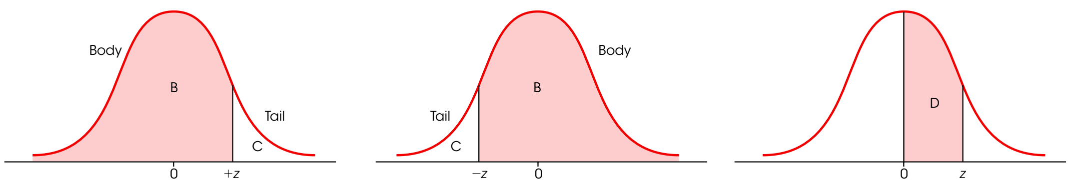


Between Two Values

- Different ways to determine the in between values
- You could determine the entire body then subtract the tail
- You could determine the proportion between the mean and given z-score for both side and add them

TABLE B.1 **The Unit Normal Table***

*Column A lists *z*-score values. A vertical line drawn through a normal distribution at a *z*-score location divides the distribution into two sections.
Column B identifies the proportion in the larger section, called the *body*.
Column C identifies the proportion in the smaller section, called the *tail*.
Column D identifies the proportion between the mean and the *z*-score.
Note: Because the normal distribution is symmetrical, the proportions for negative *z*-scores are the same as those for positive *z*-scores.

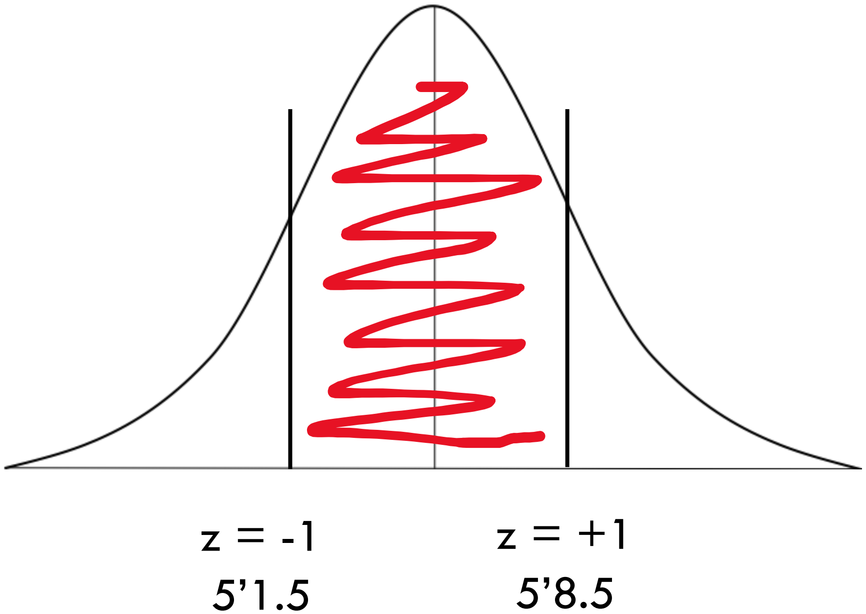


(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and z	(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and z
0.00	.5000	.5000	.0000	0.25	.5987	.4013	.0987
0.01	.5040	.4960	.0040	0.26	.6026	.3974	.1026
0.02	.5080	.4920	.0080	0.27	.6064	.3936	.1064
0.03	.5120	.4880	.0120	0.28	.6103	.3897	.1103
0.04	.5160	.4840	.0160	0.29	.6141	.3859	.1141
0.05	.5199	.4801	.0199	0.30	.6179	.3821	.1179
0.06	.5239	.4761	.0239	0.31	.6217	.3783	.1217
0.07	.5279	.4721	.0279	0.32	.6255	.3745	.1255
0.08	.5319	.4681	.0319	0.33	.6293	.3707	.1293
0.09	.5359	.4641	.0359	0.34	.6331	.3669	.1331
0.10	.5398	.4602	.0398	0.35	.6368	.3632	.1368
0.11	.5438	.4562	.0438	0.36	.6406	.3594	.1406
0.12	.5478	.4522	.0478	0.37	.6443	.3557	.1443
0.13	.5517	.4483	.0517	0.38	.6480	.3520	.1480

Between Two Values

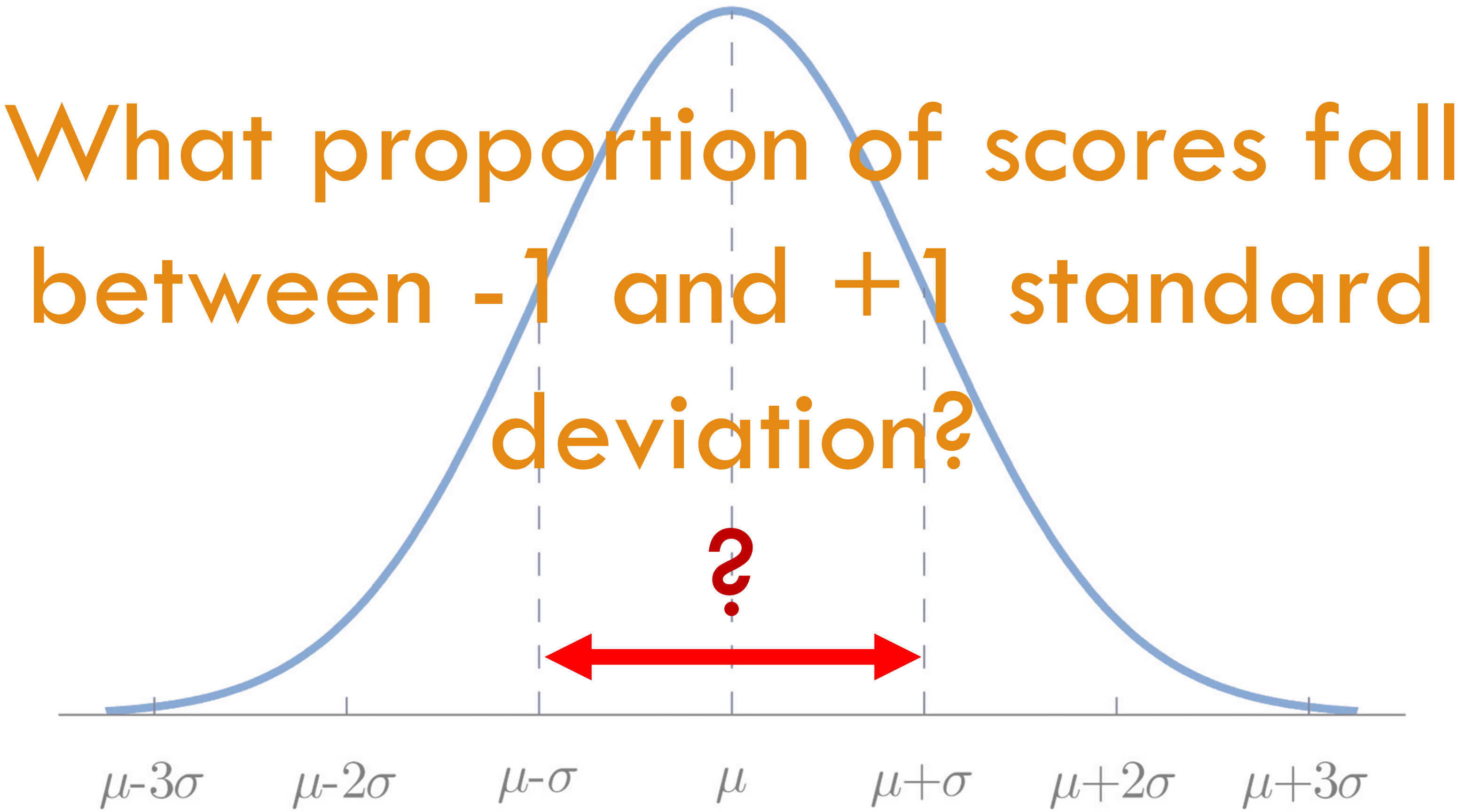
$$z = \frac{x - \mu}{\sigma}$$

- In this example, there is another quick trick. Notice that the question is asking for the proportion of people between -1 and + 1 standard deviation.
- Think back to some rules we learned about the curve…

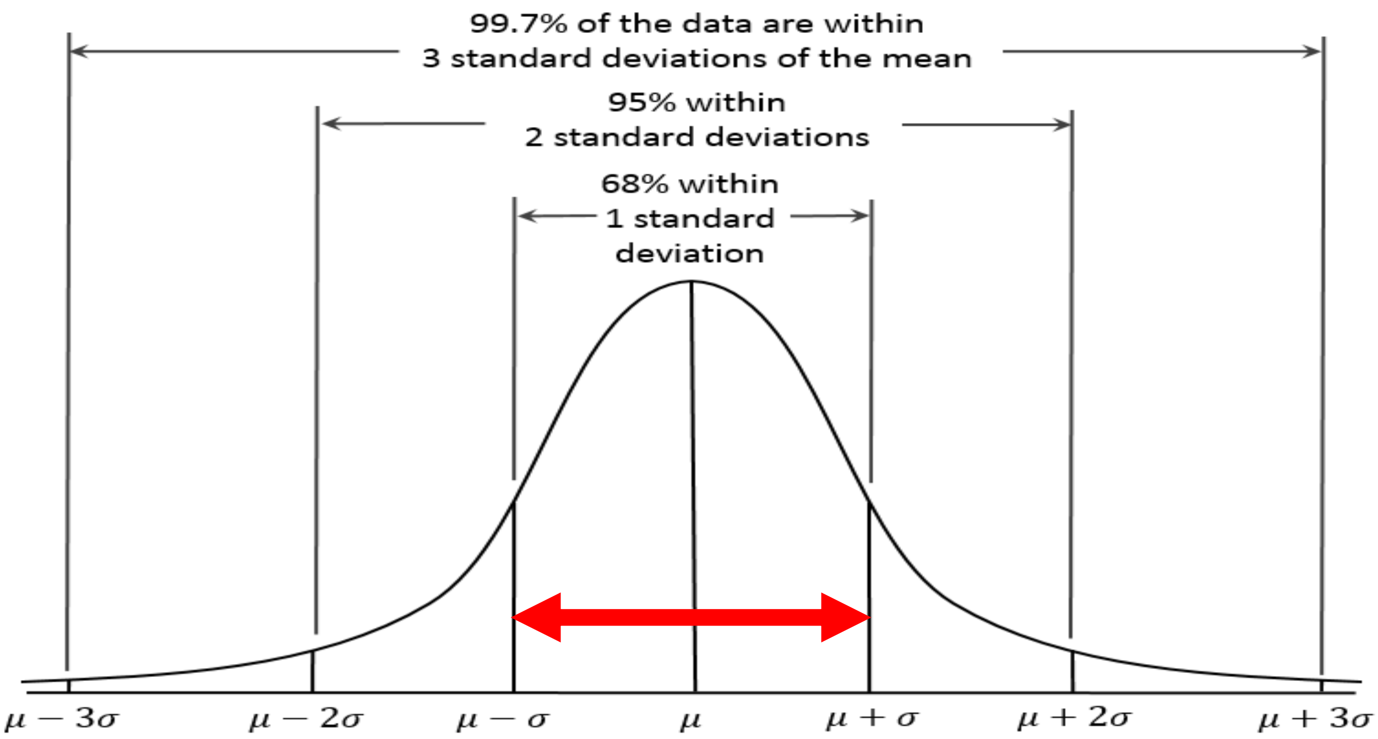


68, 95, 99 Rule

- Remember back to the normal distribution…



68, 95, 99 Rule



For normally distributed data when σ and μ are known:

68% of observations fall between $\pm$1 SDs

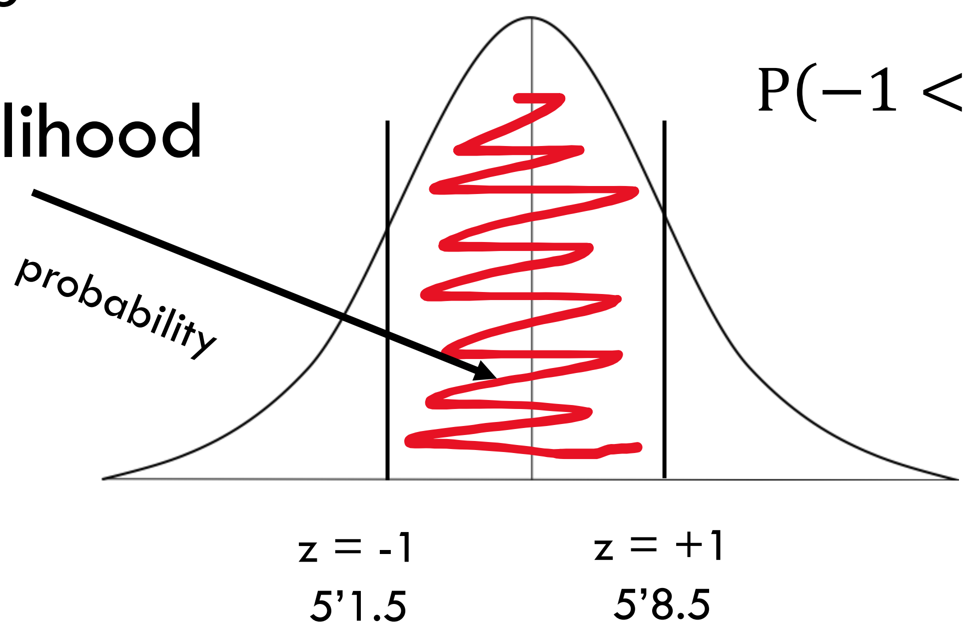
Between Two Values

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

- In this example, there is quick trick. Notice that the question is asking for the proportion of people between -1 and + 1 standard deviation.
- We know that 68% of people call between -1 and +1, so the probability of observing a woman between -1 and +1 SDs is .68

68% likelihood

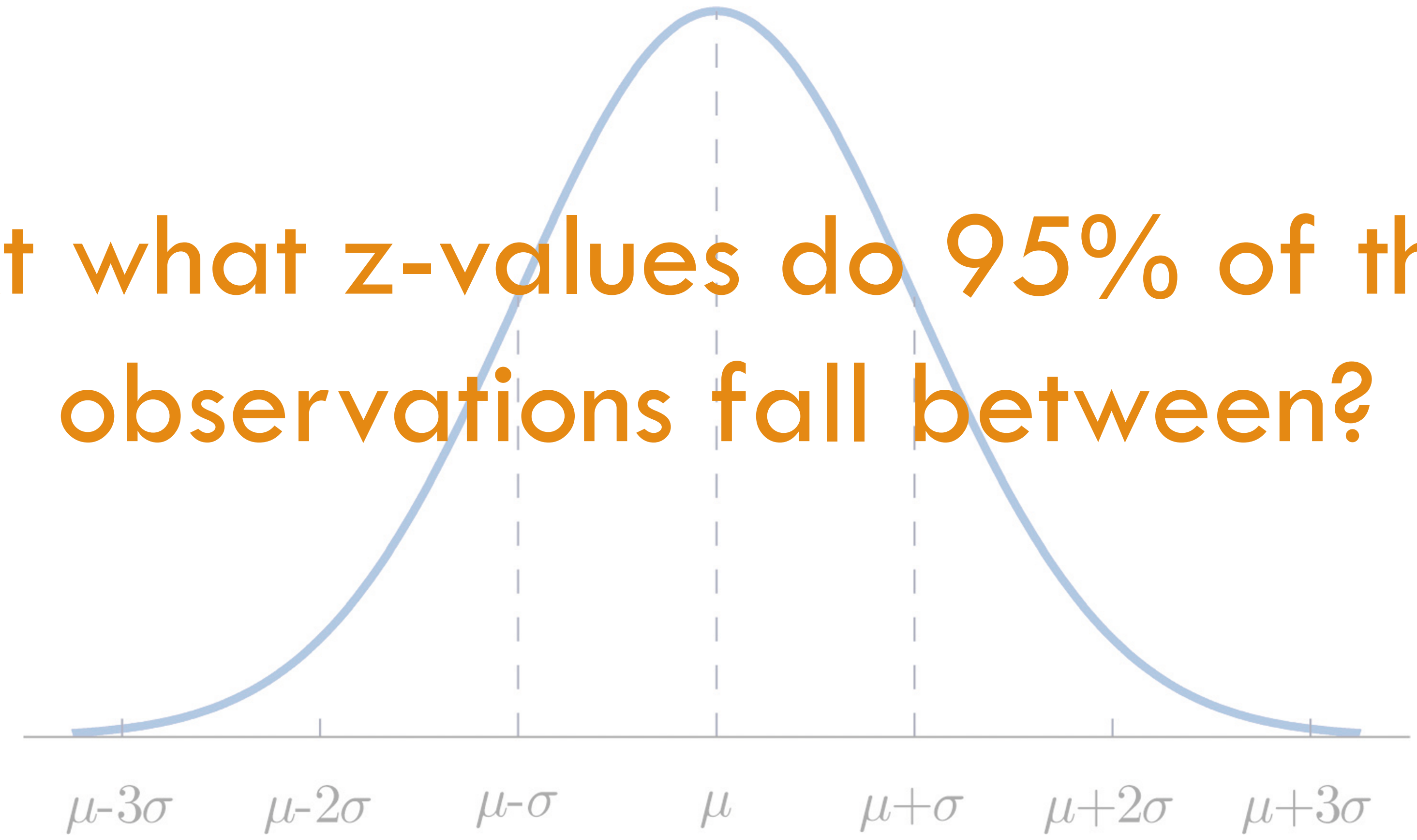
$$P(-1 < z < +1) = .68$$



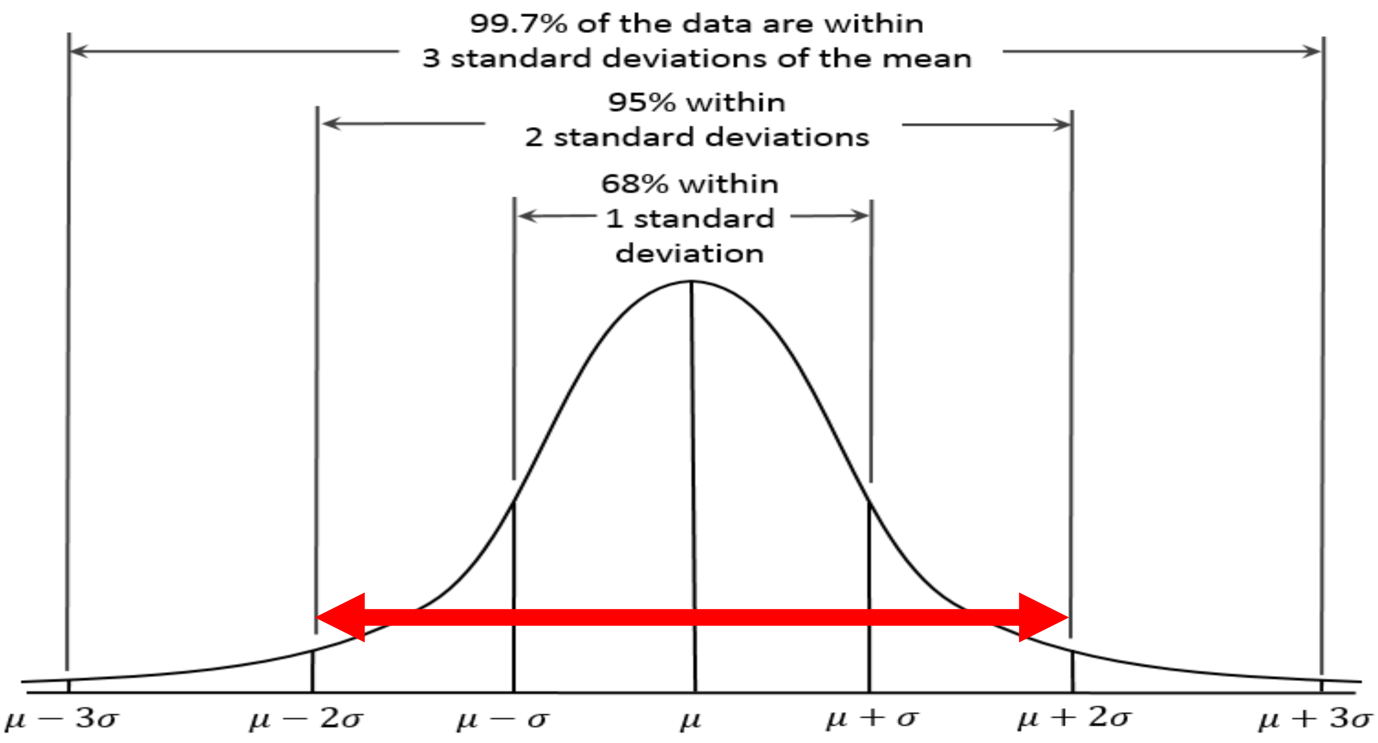
68, 95, 99 Rule

- Remember back to the normal distribution…

At what z-values do 95% of the observations fall between?



68, 95, 99 Rule



For normally distributed data: 68% of observations fall between ± 1 SDs
95% of observations fall between ± 2 SDs
99.7% of observations fall between ± 3 SDs

Z-test and Z-statistic

Moving from a Person to a Mean

- We just calculated the z-score and the associated probability of observing *one person* that is a certain height.
  - But do we usually care about just one person's score?
- Now, I want to know the probability of seeing an entire group of people, *a sample mean* that is a certain height.
  - Now we are going to run a z-test and calculate basically a z-score for a *mean* instead of just one person. This is now called a z-statistic instead of z-score.

Let's Ztandardize

- We were using population data to calculate a person's z-scores to compare and determine probabilities
- But now we are working with a sample, and all samples have some uncertainty associated with them.
  - We have to take this into account…

<u>Regular Z-score</u>

<u>Using Population Data</u>

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

z-score

<u>Sampling Z-statistic</u>

<u>Using Population and Sample Data</u>

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

z-statistic

Let's Ztandardize

- Because we are dealing with a sample mean, now we have to take into account the number of people in the sample.
- We do this with Standard Error (SE) $= {}^{\sigma}/_{\sqrt{n}}$.

<u>Regular Z-score</u>
<u>Using Population Data</u>

$$\underset{\text{z-score}}{z} = \frac{x-\mu}{\sigma}$$

<u>Sampling Z-statistic</u>
<u>Using Population and Sample Data</u>

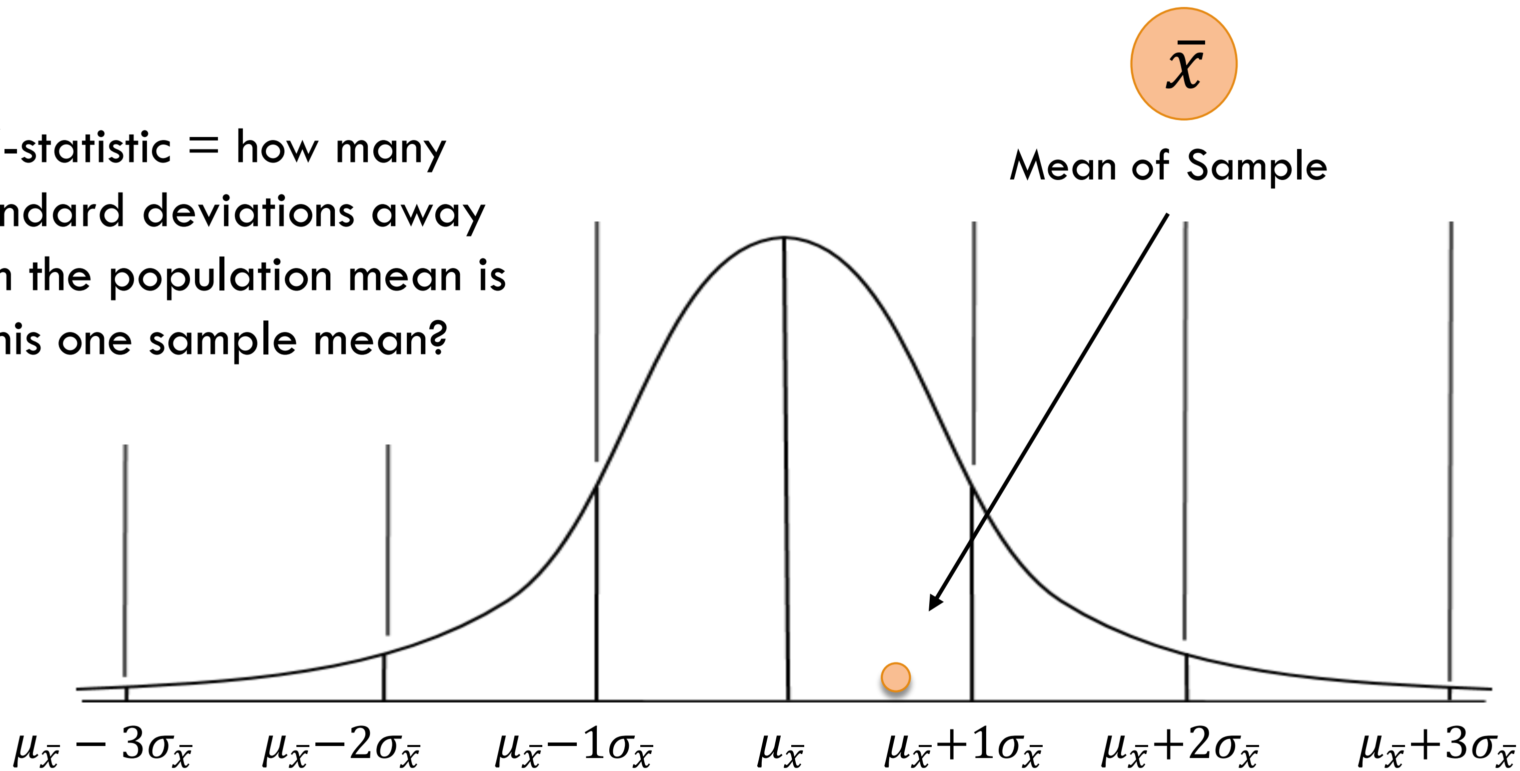
$$\underset{\text{z-statistic}}{z} = \frac{\bar{x}-\mu}{\underbrace{{}^{\sigma}/_{\sqrt{n}}}_{\text{SE}}}$$

Distribution of Statistics

There are MEANS, not PEOPLE!

- Now we have a new distribution… This one is called the "Sampling Distribution". Rather than a distribution of people, this is now a distribution of samples MEANS.

Z-statistic = how many standard deviations away from the population mean is this one sample mean?



Z-statistic (Z-test)

$$z - statistic = \frac{sample\ mean - population\ mean}{standard\ error}$$

<u>Sampling Z-statistic</u>

<u>Using Population and Sample Data</u>

$$\underset{\text{z-statistic}}{z} = \frac{\bar{x} - \mu}{\underbrace{\sigma / \sqrt{n}}_{\text{SE}}}$$

Probability $\bar{x}$

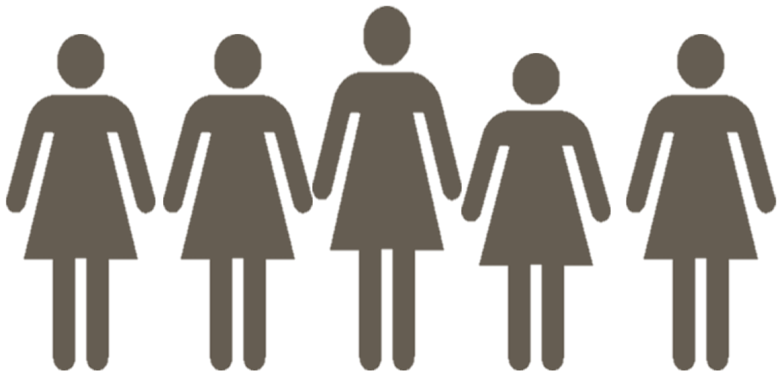
- What is the probability of observing ONE woman who is 68 inches or taller?
  - .19 probability or a 19% chance
- But now, if the average height of women is $\mu = 65$ inches, σ = 3.5, how likely is it that I collect a <u>SAMPLE</u> of 5 people (n = 5) and calculate a <u>**MEAN**</u> of 68" or taller?

How different will these probabilities be?

Probability of one person 68" or taller vs.
probability of a group of 5 people with an average of 68" or taller…

Think About It…

- Which has a higher probability… Seeing one woman that is 68'' (5'8) or seeing a group of five women with an average height of 68''?



$x = 68''$ vs $\bar{x} = 68''$

Let's Ztandardize

- If we standardize the value, then we can refer back to the z-table and see the probability of getting a certain mean just like we did for one person's score

<u>Regular Z-score</u>

<u>Using Population Data</u>

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

z-score

<u>Sampling Z-statistic</u>

<u>Using Population and Sample Data</u>

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

z-statistic

What is dividing by $\sqrt{n}$ going to do to the z value?

Increase or decrease z-statistic value?

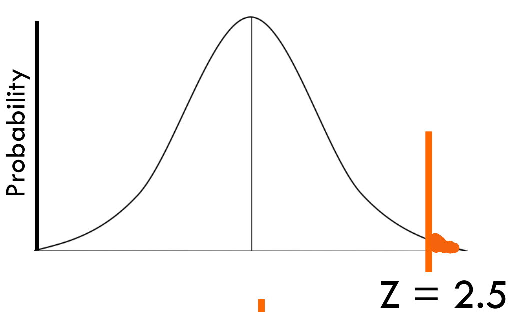
Let's Ztandardize

- Dividing by the $\sqrt{n}$ is going to *increase* the (absolute) value of the z-statistic…
- A higher z-statistic means the probability is *lower.*

Regular Z-score Using Population Data		Sampling Z-statistic Using Population and Sample Data
$z = \frac{x - \mu}{\sigma}$	$<$ Z-value	$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$
$z = \frac{x - \mu}{\sigma}$	$>$ Probability	$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

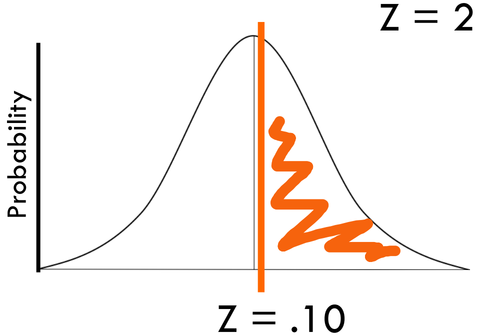
Relationship Between Z and Probability

- As a z-statistic value moves further away from the midpoint zero, the probability of seeing this sample mean goes down (lower probability), as it moves closer to the the midpoint, the probability goes up.



$\uparrow Z_{statistic}$ $\downarrow Probability$

Ex. z-statistic = 2.5 Ex. probability = .005

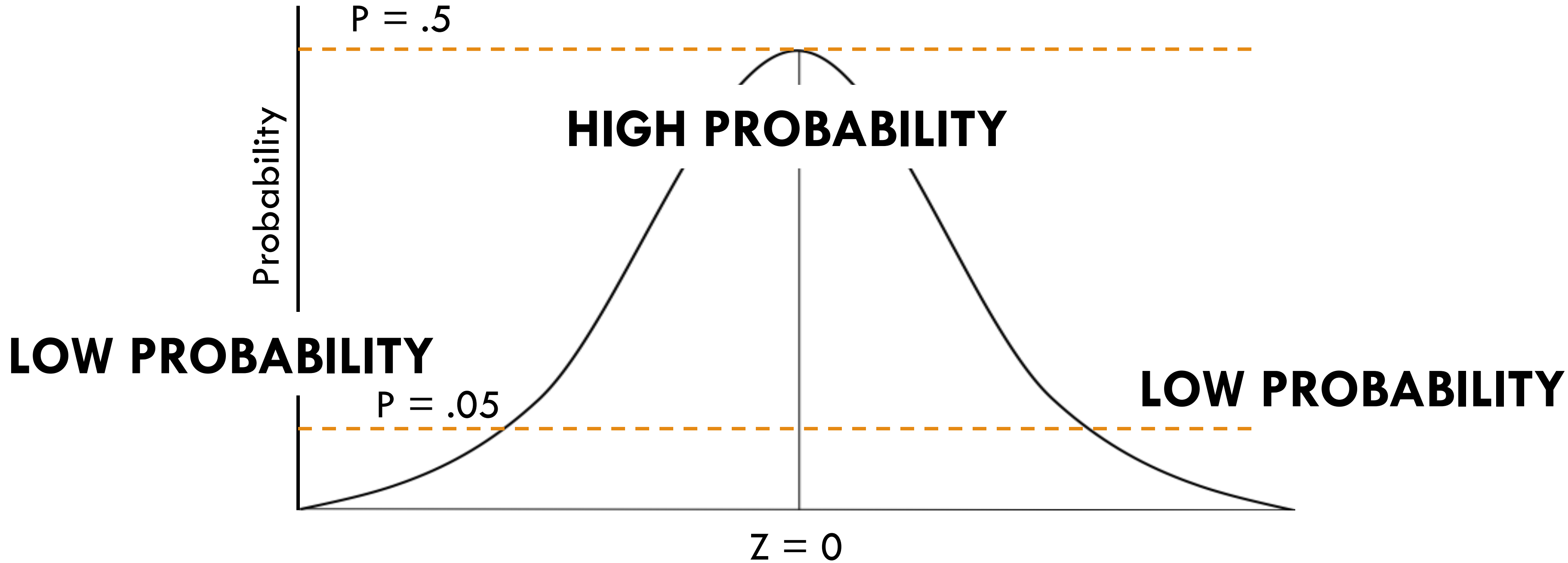


$\downarrow Z_{statistic}$ $\uparrow Probability$

Ex. z-statistic = 0.10 Ex. probability = .460

Relationship Between Z and Probability

- As a z-statistic value moves further away from the midpoint zero, the probability of seeing this sample mean goes down (lower probability), as it moves closer to the the midpoint, the probability goes up.

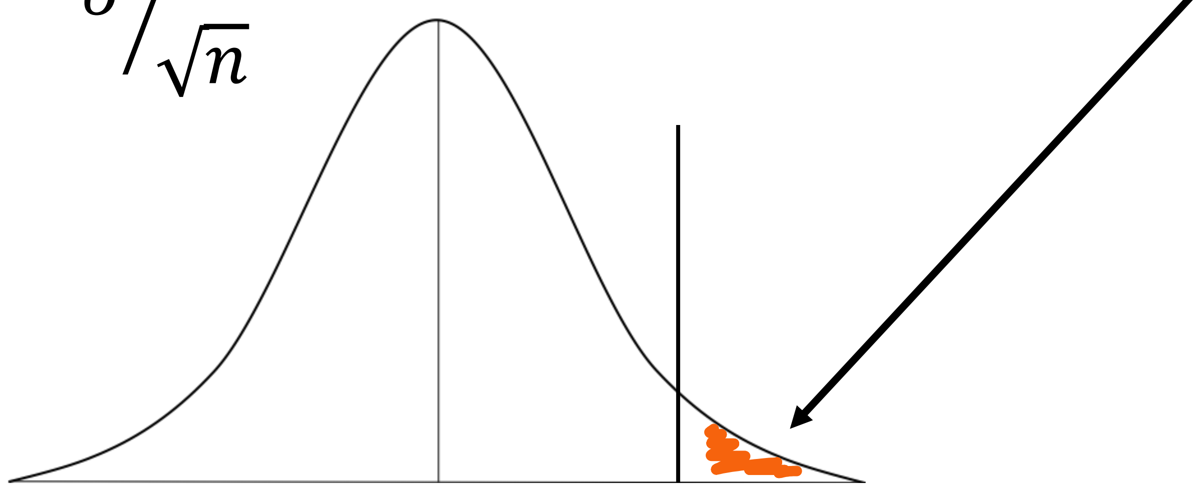


Probability $\bar{x}$, Z-statistic

- What is probability you see a SAMPLE MEAN of 68"or greater when we know the population is distributed as $x \sim N(65, 3.5)$?

$$z - statistic = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$\mathrm{P}(\bar{x} > 68) = ?$$



Probability $\bar{x}$, Z-statistic

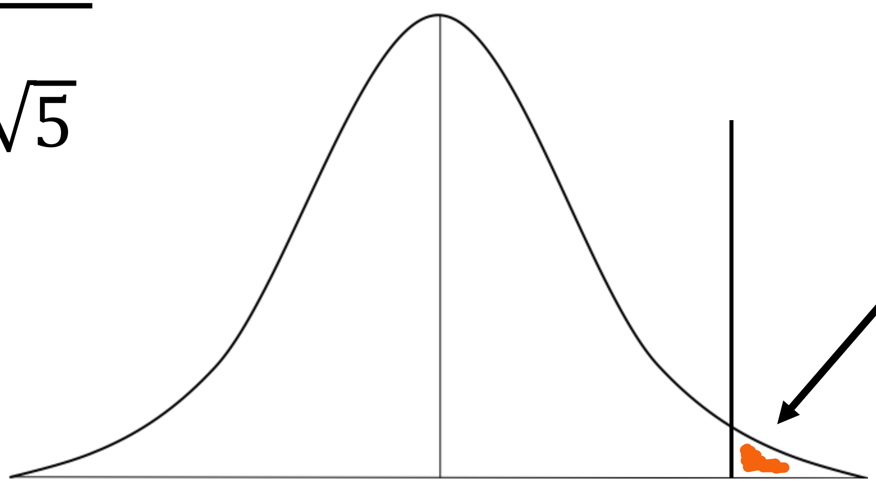
$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- What is probability you see a SAMPLE MEAN of 68”or greater when we know the population is distributed as $x \sim N(65, 3.5)$?

$$P(\bar{x} > 68) = ?$$

$$P(z > 1.92) = ?$$

$$1.92 = \frac{68 - 65}{3.5 / \sqrt{5}}$$



Z-table

- Looking up the z-statistic on the z-table.

$$1.92 = \frac{68 - 65}{3.5/\sqrt{5}}$$

We want the probability of 1.92 or higher… so we look at…

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and z
1.50	.9332	.0668	.4332
1.51	.9345	.0655	.4345
1.52	.9357	.0643	.4357
1.53	.9370	.0630	.4370
1.54	.9382	.0618	.4382
1.55	.9394	.0606	.4394
1.56	.9406	.0594	.4406
1.57	.9418	.0582	.4418
1.58	.9429	.0571	.4429
1.59	.9441	.0559	.4441
1.60	.9452	.0548	.4452
1.61	.9463	.0537	.4463
1.62	.9474	.0526	.4474
1.63	.9484	.0516	.4484
1.64	.9495	.0505	.4495
1.65			
1.66			
1.67			
1.68			
1.69			
1.70			
1.71			
1.72			
1.73			
1.74			
1.75	[illegible]	[illegible]	[illegible]
1.76	.9608	.0392	.4608
1.77	.9616	.0384	.4616
1.78	.9625	.0375	.4625
1.79	.9633	.0367	[illegible]
1.80	.9641	.0359	
1.81	.9649	.0351	
1.82	.9656	.0344	
1.83	.9664	.0336	
1.84	.9671	.0329	
1.85	.9678	.0322	
1.86	.9686	.0314	
1.87	.9693	.0307	
1.88	.9699	.0301	
1.89	.9706	.0294	
1.90	.9713	.0287	
1.91	.9719	.0281	
1.92	.9726	.0274	
[illegible]	.9732	.0268	
1.94	.9738	.0262	
1.95	.9744	.0256	.4744
1.96	.9750	.0250	.4750
1.97	.9756	.0244	.4756
1.98	.9761	.0239	.4761
1.99	.9767	.0233	.4767

(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and z
2.00	.9772	.0228	.4772
2.01	.9778	.0222	.4778
2.02	.9783	.0217	.4783
2.03	.9788	.0212	.4788
2.04	.9793	.0207	.4793
2.05	.9798	.0202	.4798
2.06	.9803	.0197	.4803
2.07	.9808	.0192	.4808
2.08	.9812	.0188	.4812
2.09	.9817	.0183	.4817
2.10	.9821	.0179	.4821
2.11	.9826	.0174	.4826
2.12	.9830	.0170	.4830
2.13	.9834	.0166	.4834
2.14	.9838	.0162	.4838
2.26	.9881	.0119	.4881
2.27	.9884	.0116	.4884
2.28	.9887	.0113	.4887
2.45	.9929	.0071	.4929
2.46	.9931	.0069	.4931
2.47	.9932	.0068	.4932
2.48	.9934	.0066	.4934
2.49	.9936	.0064	.4936

1.90	.9713	.0287	.4713
1.91	.9719	.0281	.4719
1.92	.9726	$\boxed{.0274}$	.4726
1.93	.9732	.0268	.4732
1.94	.9738	.0262	.4738

Round to 2 decimal places with probabilities, .0274 → .03

Probability $\bar{x}$, Z-statistic

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

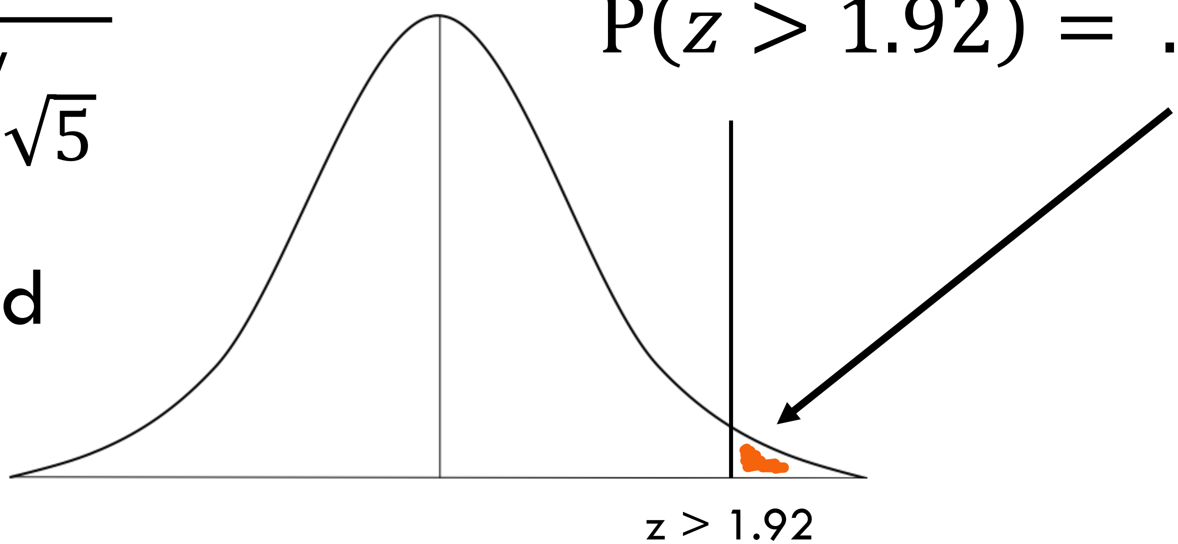
- What is probability you see a SAMPLE MEAN of 68”or greater when we know the population is distributed as $x \sim N(65, 3.5)$?

$$P(\bar{x} > 68) = .03$$

$$1.92 = \frac{68 - 65}{3.5/\sqrt{5}}$$

$$P(z > 1.92) = .03$$

3% likelihood



Side Note - Short Hand Distribution

- A common way to write out the important attributes of a distribution…
  - "The variable x is normally distributed with a mean of ___ and standard deviation of ____."
  - "Female height is normally distributed with a mean of 65 inches and a standard deviation of 3.5 inches."

$$x \sim N\,(\text{mean, standard deviation})$$

certain variable

mean

standard deviation

N is for "Normally Distributed"

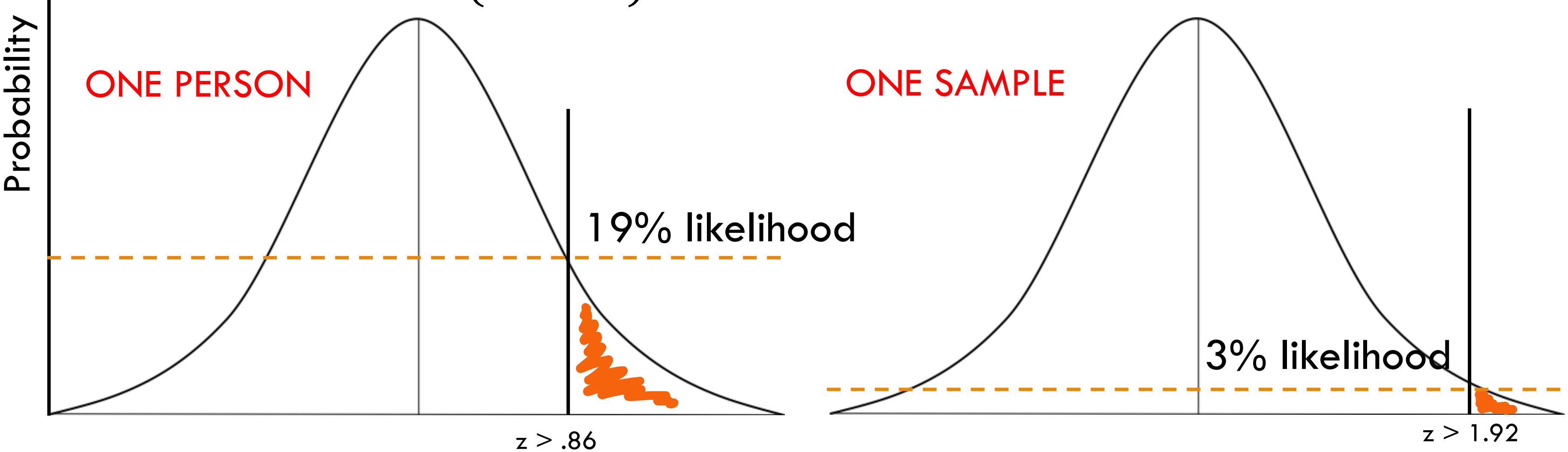
One Person vs One Mean

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- Notice how at the individual (x) level, there is more variability and a greater likelihood (19%) you'll get ONE tall person, but to get a SAMPLE MEAN that is greater than 68", perhaps you sampled a basketball team or some Scandinavians because these is less likely.

$P(x > 68) = .19$
$P(z > .86) = .19$

$P(\bar{x} > 68) = .03$
$P(z > 1.92) = .03$



Salary Person Example

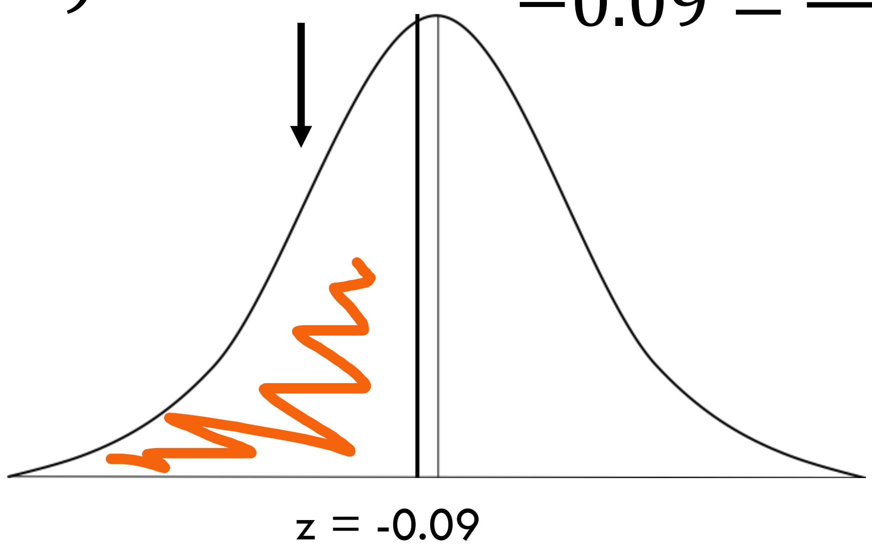
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- A large company has a salary mean of $\mu = \$62{,}000$ and a standard deviation of $\sigma = \$32{,}000$. If I select ONE employee at random, what is the probability their salary x is less than $\$59{,}000$?

$$P(x < \$59{,}000) = ?$$

$$P(z < -0.09) = ?$$

$$-0.09 = \frac{59{,}000 - 62{,}000}{32{,}000}$$



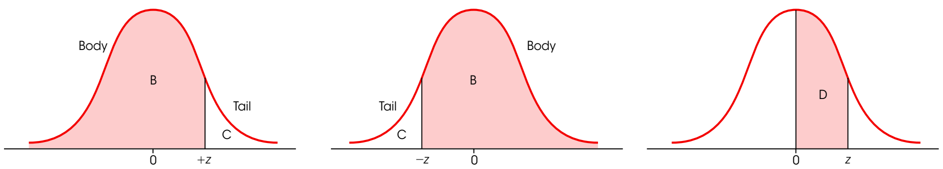
Salary Person Example

- Looking for:
  - z-score of -0.09
- The question asked for the probability of one person's salary being $59,000 or less… Which direction do we want?

$$P(z < -0.09) = ?$$

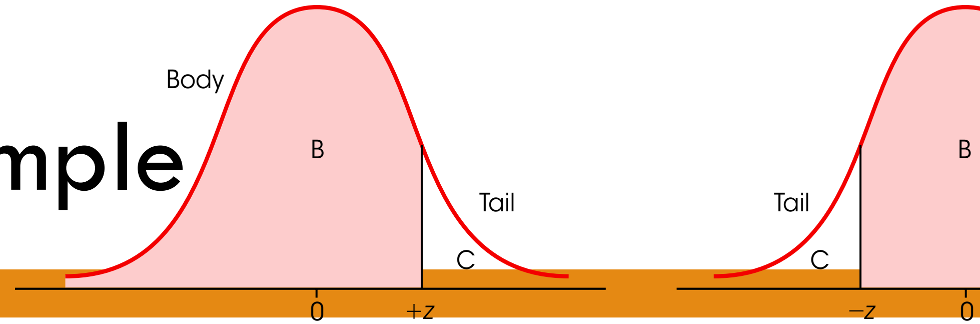
TABLE B.1 The Unit Normal Table*

*Column A lists z-score values. A vertical line drawn through a normal distribution at a z-score location divides the distribution into two sections.
Column B identifies the proportion in the larger section, called the *body.*
Column C identifies the proportion in the smaller section, called the *tail.*
Column D identifies the proportion between the mean and the z-score.
Note: Because the normal distribution is symmetrical, the proportions for negative z-scores are the same as those for positive z-scores.



(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and z	(A) z	(B) Proportion in Body	(C) Proportion in Tail	(D) Proportion Between Mean and z
0.00	.5000	.5000	.0000	0.25	.5987	.4013	.0987
0.01	.5040	.4960	.0040	0.26	.6026	.3974	.1026
0.02	.5080	.4920	.0080	0.27	.6064	.3936	.1064
0.03	.5120	.4880	.0120	0.28	.6103	.3897	.1103
0.04	.5160	.4840	.0160	0.29	.6141	.3859	.1141
0.05	.5199	.4801	.0199	0.30	.6179	.3821	.1179
0.06	.5239	.4761	.0239	0.31	.6217	.3783	.1217
0.07	.5279	.4721	.0279	0.32	.6255	.3745	.1255
0.08	.5319	.4681	.0319	0.33	.6293	.3707	.1293
0.09	.5359	.4641	.0359	0.34	.6331	.3669	.1331
0.10	.5398	.4602	.0398	0.35	.6368	.3632	.1368
0.11	.5438	.4562	.0438	0.36	.6406	.3594	.1406
0.12	.5478	.4522	.0478	0.37	.6443	.3557	.1443
0.13	.5517	.4483	.0517	0.38	.6480	.3520	.1480
0.14	.5557	.4443	.0557	0.39	.6517	.3483	.1517
0.15	.5596	.4404	.0596	0.40	.6554	.3446	.1554
0.16	.5636	.4364	.0636	0.41	.6591	.3409	.1591
0.17	.5675	.4325	.0675	0.42	.6628	.3372	.1628
0.18	.5714	.4286	.0714	0.43	.6664	.3336	.1664
0.19	.5753	.4247	.0753	0.44	.6700	.3300	.1700
0.20	.5793	.4207	.0793	0.45	.6736	.3264	.1736
0.21	.5832	.4168	.0832	0.46	.6772	.3228	.1772
0.22	.5871	.4129	.0871	0.47	.6808	.3192	.1808
0.23	.5910	.4090	.0910	0.48	.6844	.3156	.1844
0.24	.5948	.4052	.0948	0.49	.6879	.3121	.1879

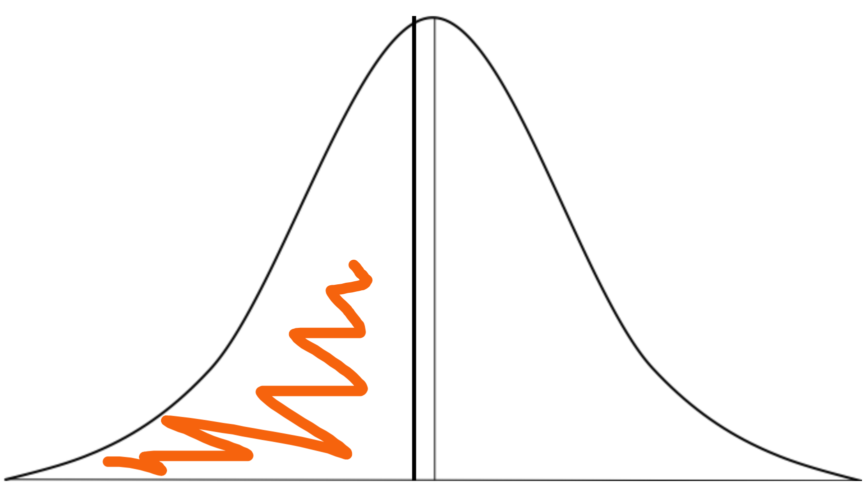
Salary Person Example



$$P(z < -0.09) = ?$$

(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
0.04	.5160	.4840	.0160
0.05	.5199	.4801	.0199
0.06	.5239	.4761	.0239
0.07	.5279	.4721	.0279
0.08	.5319	.4681	.0319
0.09	.5359	$\boxed{.4641}$	.0359

Notice that the proportion we want does NOT cross the mean (z=0), so we know that the probability has to be less than .500, we want the "tail"



z = -0.09

$$P(z < -0.09) = .46$$

46% of people have a salary less than $59,000
The probability of one person having a salary less than $59,000 is .46.

Salary Sample Example

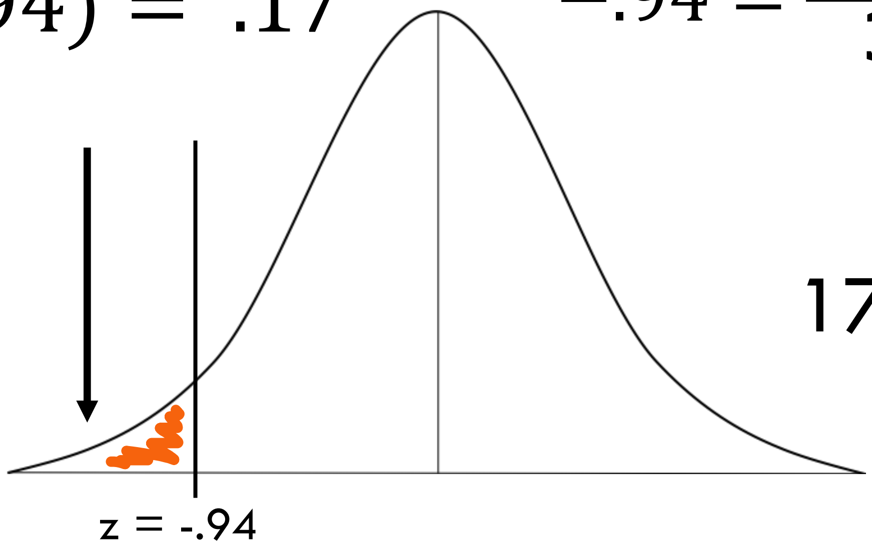
$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- A large company has a salary mean of $\mu = \$62{,}000$ and a standard deviation of $\sigma = \$32{,}000$. If I select 100 employees at random, what is the probability their **average** salary $\bar{x}$ is less than $\$59{,}000$?

$$P(\bar{x} < \$59{,}000) = .17$$

$$P(z < -.94) = .17$$

$$-.94 = \frac{59{,}000 - 62{,}000}{32{,}000 / \sqrt{100}}$$



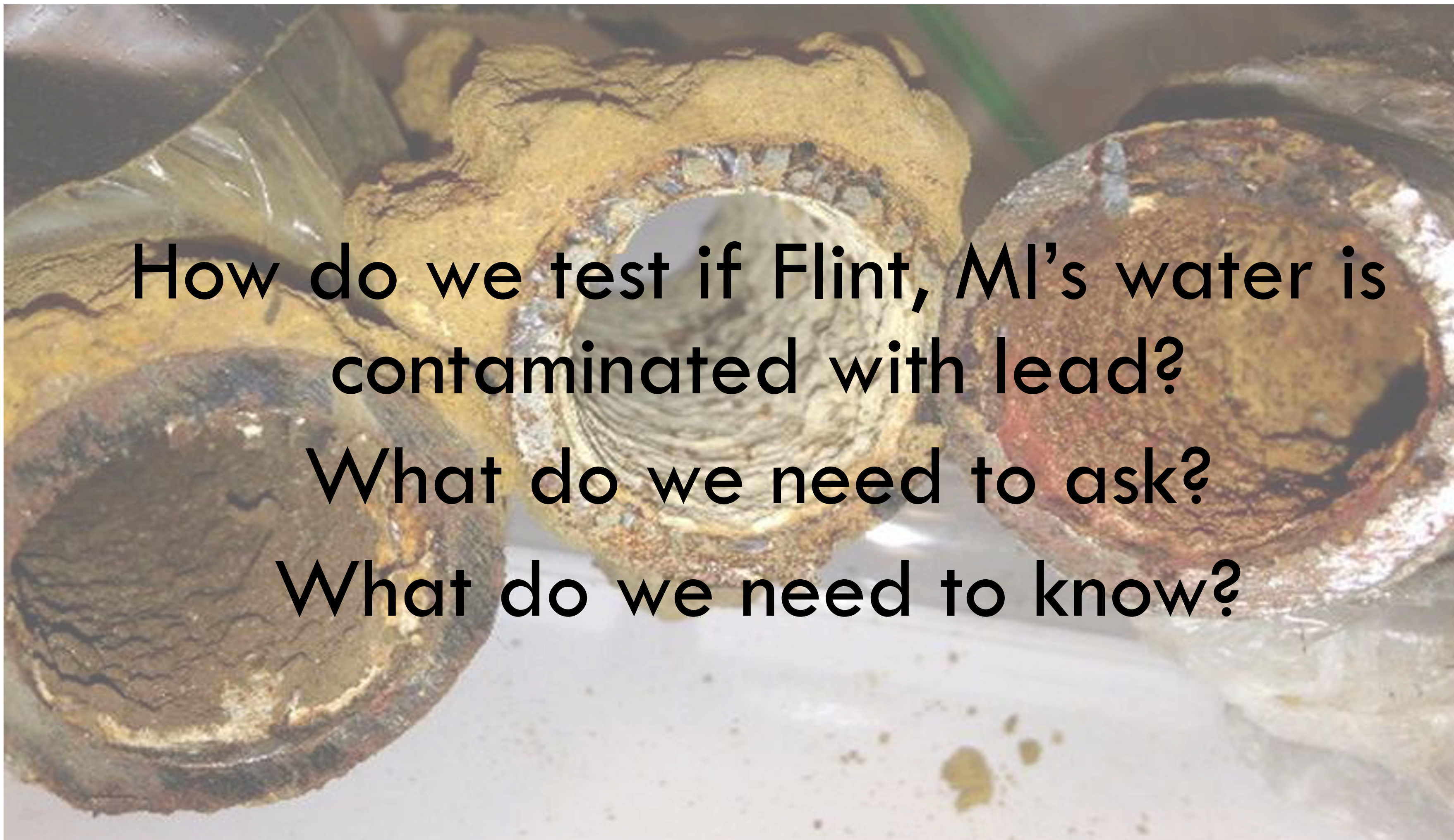
17% likelihood

Z-test

- When we test things in statistics, we are looking for a PROBABILITY, not really a simple "Yes" or "No"
- We usually ask, "Assuming everyone is similar to the average, what is the probability I see a sample mean this high (or low)?"
  - Ex. "Assuming the average height of women is 65", what is the probability I take a sample of 5 women and their average height is 68" or greater?"

Real World Example: Flint, MI

Is Flinch, MI Water Contaminated?



How do we test if Flint, MI's water is contaminated with lead?

What do we need to ask?

What do we need to know?

What do we need to ask?

- There may be trace amounts of lead in all water, but, on average, how much lead is in America's drinking water?
- Some cities may have more or less lead in their water, but how much does that average amount of lead vary from city to city?
- How many Flint, MI water samples were taken?
- What was the average lead level of those Flint, MI samples?

What do we need to know?

- There may be trace amounts of lead in all water, but, on average, how much lead is in America's drinking water?
  - EPA says average is 2.8 ppb
  - Anything over 15 ppb is "action level"
- Some cities may have more or less lead in their water, but how much does that average amount of lead vary from city to city?
  - This is hard to find… For the sake of this example, let's say it is 1.5* ppb
- How many Flint, MI water samples were taken? What was the average lead level of those Flint, MI samples?
  - 269, no overall average found, but we do know that 40% of those samples (n = 101) were over 5ppb
  - The top 10 percent, i.e. the 90^{th} percentile, (n = 27) was 25ppb
  - A few outlier samples were over 100 ppb, one exceeded 1000 ppb

*This is a made up value for the example, the rest of the values are from water reports.

Flint, MI One Water Sample

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

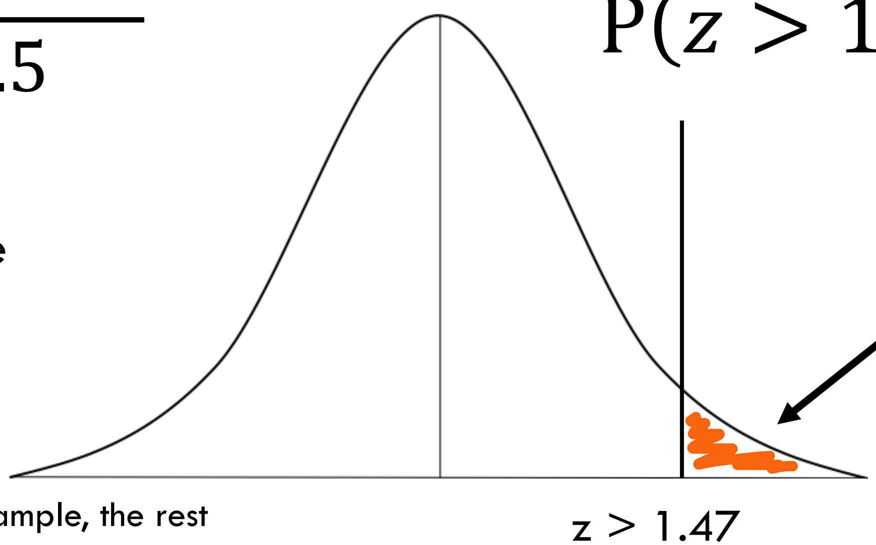
- *Assuming the water is not contaminated*, what is probability you see ONE water sample (not an average) with a lead concentration of 5 ppb or greater, considering the mean for the US is 2.8 with a "known" standard deviation of 1.5*.

$$1.47 = \frac{5.0 - 2.8}{1.5}$$

$$P(x > 5.0) = .07$$

$$P(z > 1.47) = .07$$

7% chance of seeing one sample greater than 5 ppb of lead



*This is a made up value for the example, the rest of the values are from water reports.

Flint, MI Water

- *Assuming all the water is fine*, there is still a 7% chance of seeing one water sample greater than 5 ppb. It's not impossible or completely unlikely to see some water samples with high lead levels because samples vary.
  - Maybe you sampled a particularly old lead pipe.
- But… *Assuming all the water is fine,* what is the probability of seeing an average from multiple water samples of 5 ppb or greater?

Flint, MI Average Water Sample

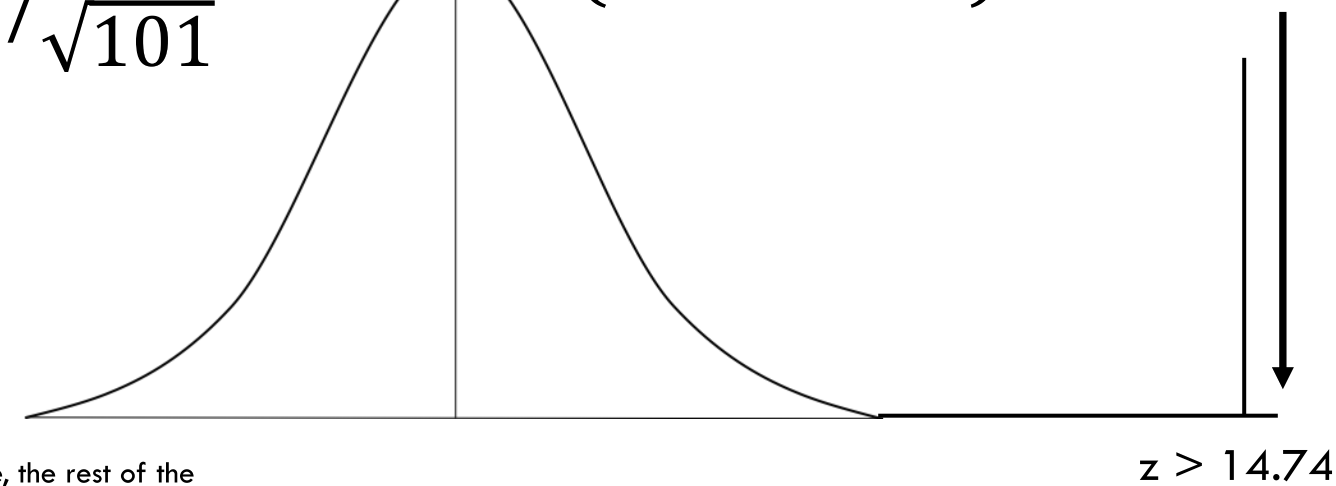
- *Assuming the water is not contaminated*, what is probability you see an average lead concentration of 5 ppb or greater from a sample of n = 101, considering the mean for the US is 2.8 with a “known” standard deviation of 1.5*.

$$14.74 = \frac{5.0 - 2.8}{1.5/\sqrt{101}}$$

$$P(\bar{x} > 5.0) = .000$$

$$P(z > 14.74) = .000$$

0.000% chance of seeing an average of samples this high if the water was fine



*This is a made up value for the example, the rest of the values are from water reports.

Flint, MI Water Example

- *Assuming all the water is fine*, there a 0.0000% chance of seeing an average lead concentration of 5 ppb from an average of 101 water samples.
  - But, there is still technically a really, really small chance (somewhere at the end of all those zeros) that this could just be a fluke…
- But what is more likely? That this was a fluke of 101 water samples and everything is fine, as some politicians might claim, or is it more likely that Flint, MI water is different from the average of 2.8 ppb, and the water is contaminated…?

This is the essence of statistical hypothesis testing.

Z-table Tips

- When you calculate a z-score, round the z-score to two decimal places
- When you find a probability in a z-table, round the probability to two decimal places
- When you find the z-score that corresponds to a probability, go with the z-score that is *smaller*
- There are no negative z-scores on the table, but you can still look up negative z-scores values
  - Ex. z = -1.5, look for 1.5 and based on the question, determine if you want proportion of people above of below the mean

Next up…

- For the rest of this course, we will focus mainly on means we derive from samples (i.e. one sample's mean), rather than one single observation (i.e. one person's score)
- Next we shift our attention to the process of taking a good sample.

Z-scores, Z-tests, and Probability in R

Probability $\bar{x}$, Z-statistic

```
## 3.2-2 Probability, z-scores, and z-test

```{r}
################################
##### z-test & Probability  ####
################################

sample_mean <- 68
pop_mean <- 65
sd <- 3.5
n <- 5

z_stat <- (sample_mean - pop_mean)/(sd/sqrt(n))

# When asking for the probability of seeing a sample_mean GREATER than a certain value,
# P(sample_mean > 68), P(z > 1.92)
pnorm(z_stat, lower.tail = F) #0.03

# When asking for the probability of seeing a sample_mean LESS than a certain value,
# P(sample_mean < 68), P(z < 1.92)
pnorm(z_stat, lower.tail = T) #0.97

# When asking for the probability of seeing a sample_mean BETWEEN two symmetric values,
# P(-1.92 < z < 1.92)
1 - pnorm(z_stat, lower.tail = F)*2 #0.94
```
```

Flint, MI One Sample vs. Average Water Sample

- The first part of the code calculates a z-score for just one sample value.
- The second part of the code calculates a z-test for a sample mean.

```r
```{r}
###############################
##### z-score, z-test, #########
#####  & Probability  ##########
###############################
########## Flint, MI ###########
###############################

#####  One water sample: z-score #####

water_mean <- 2.8
water_sd <- 1.5
water_sample <- 5

water_Z_score <- (water_sample - water_mean)/water_sd

# When asking for the probability of seeing a sample GREATER than a certain value,
# P( one sample > 5.0), P(z > 1.47)
pnorm(water_Z_score, lower.tail = F) #0.07

##### A sample average: z-test #####

sample_water_mean <- 5
water_mean <- 2.8
water_sd <- 1.5
n <- 101

z_stat_water_sample <- (sample_water_mean - water_mean)/(water_sd/sqrt(n))

# When asking for the probability of seeing a sample_mean GREATER than a certain value,
# P(sample_mean > 5.0), P(z > 14.74)
pnorm(z_stat_water_sample, lower.tail = F) #0.000000
```
```