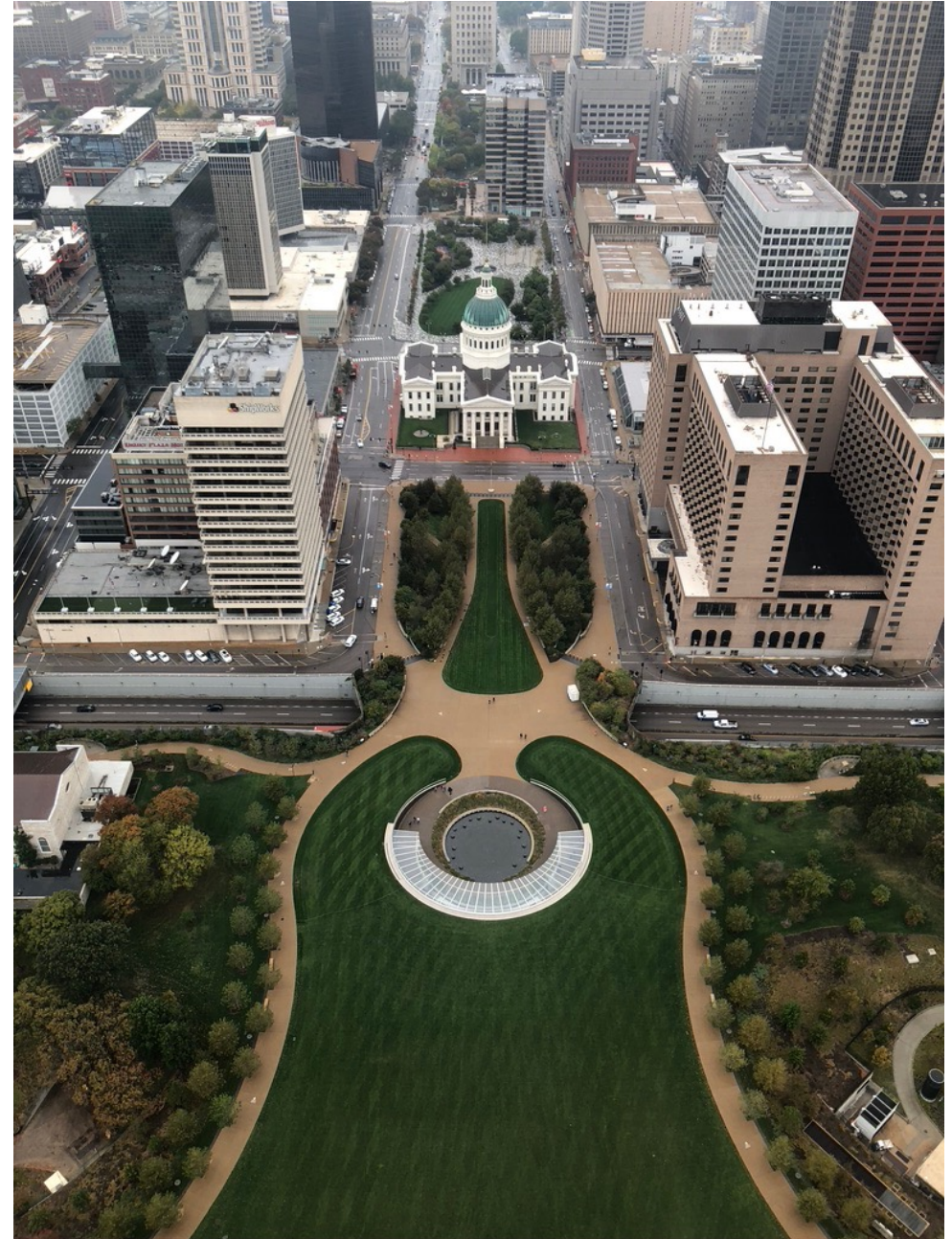


# Still working on it...

Unit 3 – Numerical Summaries, Day 2  
Non-Calculating Professor Colton



# Unit 3, Day 2 - Outline

## Unit 3 – Numerical Summaries

### Measures of Variation

- Range
- Standard Deviation
- Variance
- IQR
- Best Measure of Variation
- Standard Deviation of Grouped Data

### Using the Standard Deviation

- Normal Distribution
- Empirical Rule

### Measures of Relative Position

- Percentiles
- Quartiles
- Outliers via Boxplots
- Z-scores

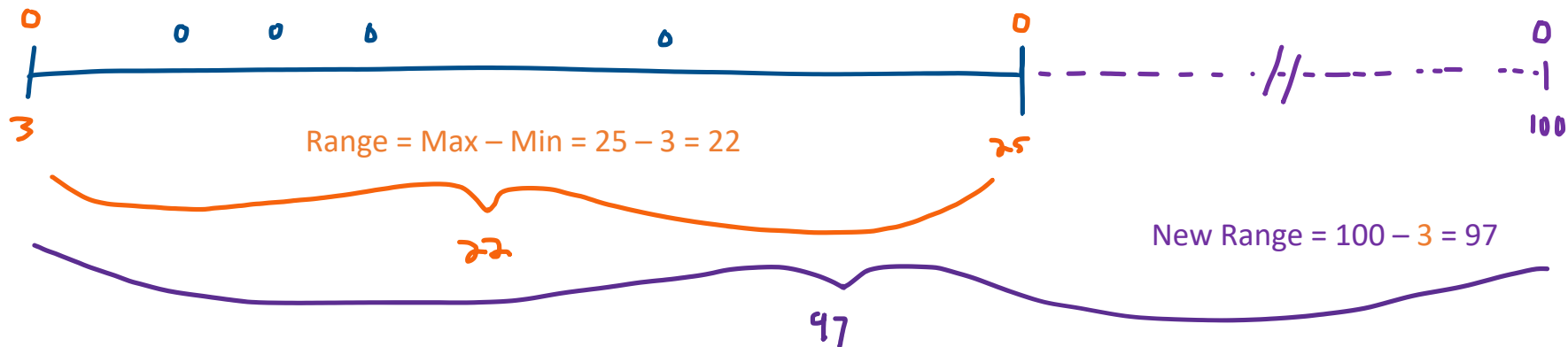
# Measures of Variation: Intro

- We talked about measures of center; these give us ideas of **location**.
- So we know where the “middle” is. Now how far do the values branch (spread) out around the middle??
- To quantify this, we use a measure of spread.
- **Measures of Variation** give an idea of the spread/variability of your data.
- Smaller values, data is closer together (i.e., *more consistent*).
- Large values, the data is spread out from one another (i.e., *less consistent*).
- To BEST **describe** a dataset, we want to give the *best, most appropriate* measure of center **paired** with the *best, most appropriate* measure of variation!

# Range

## Range

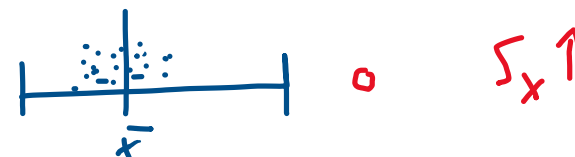
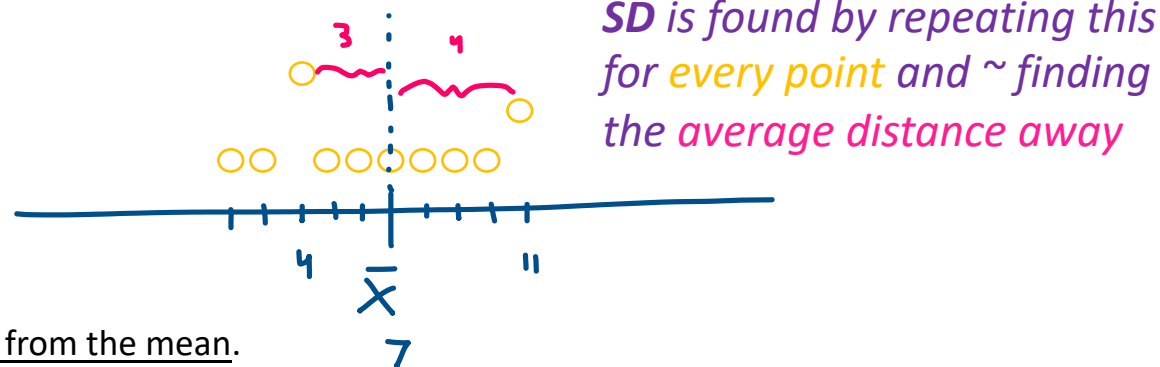
- Maybe the simplest measure of variation (spread).
- $\text{Range} = \text{Max} - \text{Min}$
- Gives idea of the entire "range" of values, how much distance do they span in total.
- Obviously range is heavily impacted by extreme observations → NOT resistant
- Very easy to find in calculator, just do a 1-Var Stats to find the info you need.



# Standard Deviation

## Standard Deviation (SD)

- Most informative measure of variation.
- Complex formula that measures the average distance that each data point is from the mean.
- Notation and formula:
  - Sample Standard Deviation =  $S = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}$
  - Population Standard Deviation =  $\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$ , (Greek letter sigma)
  - On your calculator, 1-Var Stats you will see  $Sx$  and  $\sigma x$ .
- AGAIN THERE IS A BIG DIFFERENCE BETWEEN THESE TWO!
  - $S$  is a statistic because it describes a sample! While  $\sigma$  is a parameter because it is referring to a population
- Most of the time we will be **evaluating sample data**, so look for  $Sx$  on your calculator to calculate the sample standard deviation.
- **Standard Deviation**, like the mean is **NOT resistant** to skewness and outliers.



Adding this new point causes the standard deviation to increase alot

# Variance

## Variance

- Another measure of variation.
- **Variance** is actually just equal to the square of the standard deviation.
- Notation and formula:
  - Sample variance =  $S^2$
  - Population Variance =  $\sigma^2$
  - AGAIN, THERE IS A BIG DIFFERENCE BETWEEN THESE TWO! Same distinction as with SD
- Same properties as standard deviation.
  - Measures the average **squared** distance that each data point is from the mean.
  - **NOT resistant**.
  - Generally we will work with the standard deviation, but good to know what this is as well.

# LCQ: Standard Deviation and Variance

- 1) **Determine** the population variance and population standard deviation of the following population.
  - Data: 23, 34, 12, 33, 24, 40, 33, 35, 32, 20
- 2) **Approximate** the population variance and population standard deviation of the following sample.
  - Data: 23, 34, 12, 33, 24, 40, 33, 35, 32, 20
- 3) Lets say this data is from a sample of the number of eggs hens laid in the month. **Interpret** the standard deviation in context.

# LCQ: SD and Variance

L1	L2	L3	L4	L5	1
23					
34					
12					
33					
24					
40					
33					
35					
32					
20					
-----					
L1(10)= 20					

L1	L2	L3	L4	L5	1
23					
34					
12					
33					
24					
40					
33					
35					
32					
20					
-----					
L1(10)= 20					

L1	L2	L3	L4	L5	1
23					
34					
12					
33					
24					
40					
33					
35					
32					
20					
-----					
L1(10)= 20					

1) **Determine** the population standard deviation and population variance of the following population.

- Data: 23, 34, 12, 33, 24, 40, 33, 35, 32, 20
- $\sigma = 8.077 \rightarrow$  The population SD is given with the Greek symbol sigma
- $\sigma^2 = (8.077)^2 = 65.238 \rightarrow$  Just have to square it to get the population variance

2) **Approximate** the population standard deviation and population variance of the following sample.

- Data: 23, 34, 12, 33, 24, 40, 33, 35, 32, 20
- $S = 8.514 \rightarrow$  The sample SD is given with the regular letter S
- $S^2 = (8.514)^2 = 72.488 \rightarrow$  Just have to square it to get the sample variance

3) Lets say this data is from a sample of the number of eggs hens laid in the month. **Interpret** the standard deviation in context.

- The number of eggs hens laid in a month was on average 8.514 eggs away from the mean of 28.6 eggs
- There's two numbers needed in the interpretation of a standard deviation, 1) the average distance away from 2) the mean. And we have to talk about these both in context

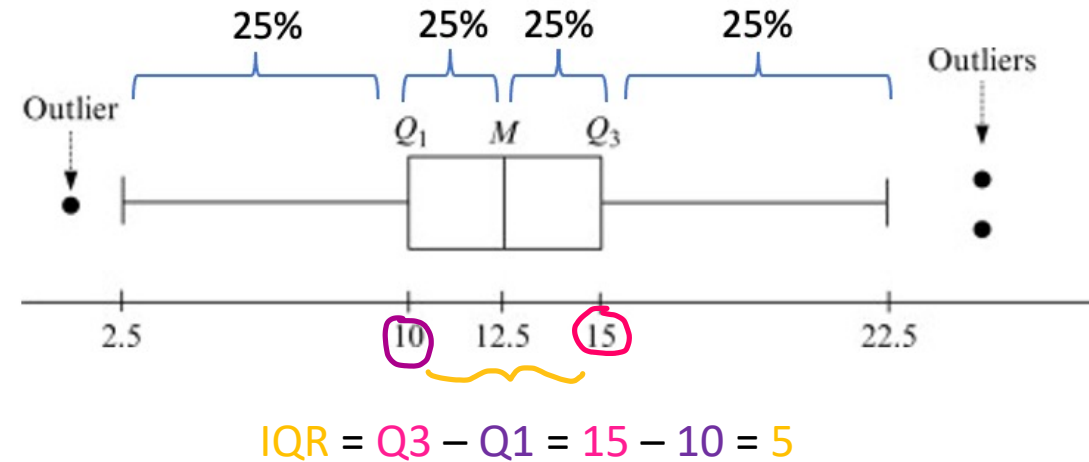
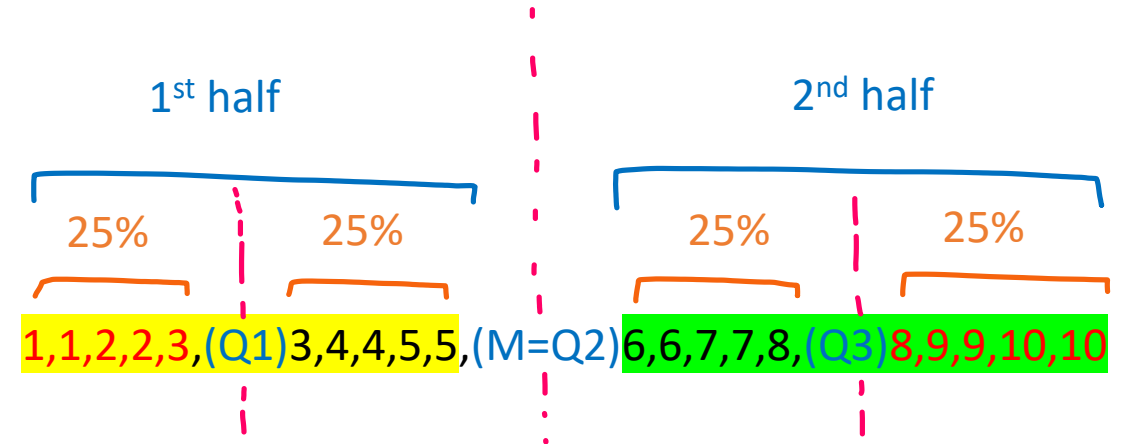
- The difference between number one and two is about perspective!
- In number one we are viewing this data set as the entire population. It includes every single member of interest.
  - So when we're finding the standard deviation and variance, they represent population values
- Whereas in number two we are viewing the same set of data, but now as a subset of our overall population
  - So it only includes a portion of the members of interest. Now it becomes a statistic because it's about a sample
- And remember the reason why we are sampling is because we cannot measure everybody in the population
  - The hope is that our sample information (our sample statistics) will be close to / estimate well the population information if we were actually able to obtain it



# IQR

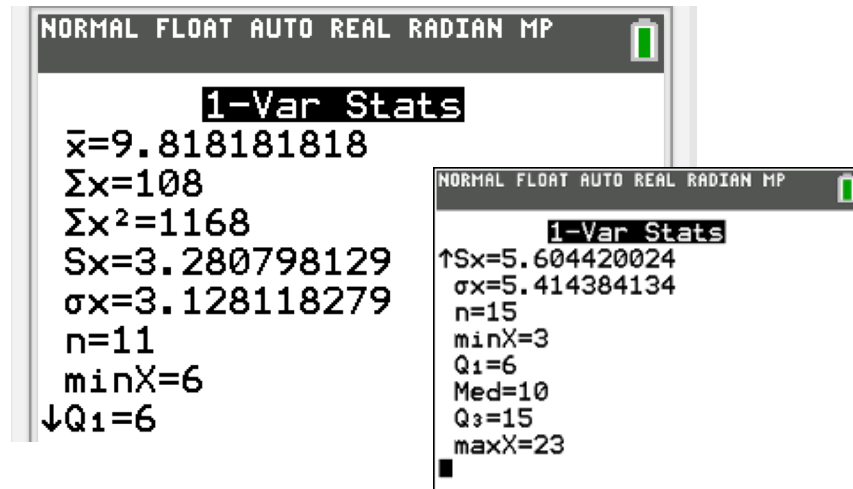
## Inter Quartile Range (IQR)

- Another measure of variation, less informative than the standard deviation.
- Uses quartiles to measure how far data is spread out around the median.
  - Specifically it measures the range of the middle 50% of the data
  - Visualized very well in boxplots! It is the length of the box!
- Notation and formula:
  - $IQR = Q3 - Q1$ 
    - $Q3$  = median (middle) of the second half of data
    - $Q1$  = median (middle) of the first half of data
- **IQR**, like the median is **resistant** to skewness and outliers.



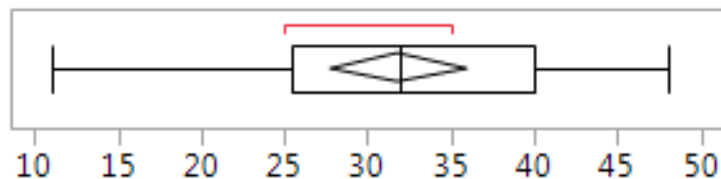
# LCQ: IQR

- 1) Using this output from a 1-Var Stat, **what** is the IQR?



IQR = ???

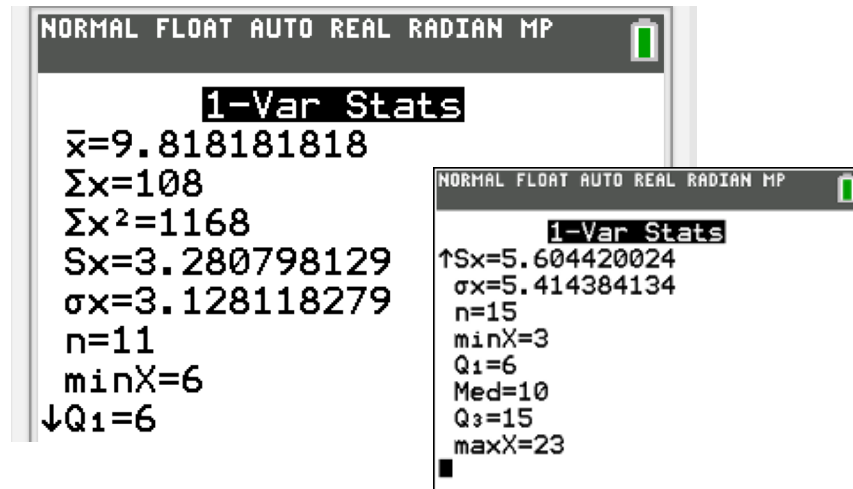
- 2) Find the IQR from this boxplot.



IQR = ???

# LCQ: IQR

- 1) Using this output from a 1-Var Stat, **calculate** the IQR? **Calculate** the range.

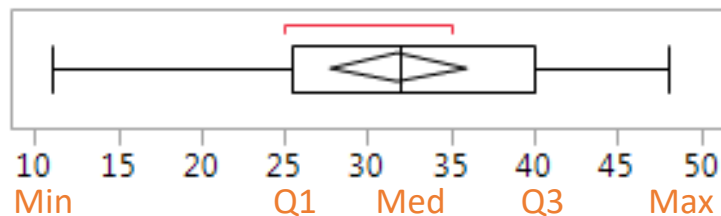


$$IQR = Q3 - Q1 = 15 - 6 = 9$$

$$Range = Max - Min = 23 - 3 = 20$$

- Another common answer  $\rightarrow$  Range: 3 to 23
- In words this is correct, but in stats context we want the actual subtraction (so the first statement)

- 2) Find the IQR from this boxplot.



$$IQR = 40 - 25 = 15$$

# Best Measure of Variation

## Comparison of SD and IQR

- BOTH measure variation, but differently.
  - Standard Deviation actually uses the VALUES of each data point → MOST INFORMATIVE
  - IQR (for the most part) ONLY uses the POSITION of each data point.
- Again, this is why SD is *NOT resistant* while IQR is *resistant*.

Mean ⇔ Standard Deviation

Median ⇔ IQR

Can *optionally* report range, but that is NOT enough information by itself to **describe** the spread.

## When to Use

- **SD** is the most informative, so we want to use this if possible. This means when...
  - When data is symmetric and there is NO outliers
  - **SD** is ALWAYS reported when reporting mean!
- **SD blows up when there are extreme observations.**
  - So, we DON'T report this when there are outliers or skewness to the data.
- When this is the case, use the **IQR**.
  - This would now be the MOST APPROPRIATE measure of variation because it is of course *unaffected by the skewness*.
  - This is ALWAYS reported with the median.

# IMPORTANT LCQ: Best Measure of Variation

Which measure of variation is needed?????

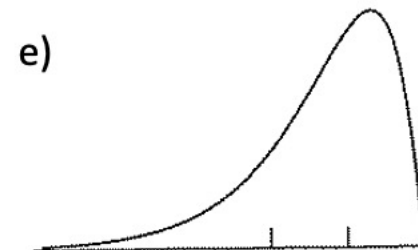
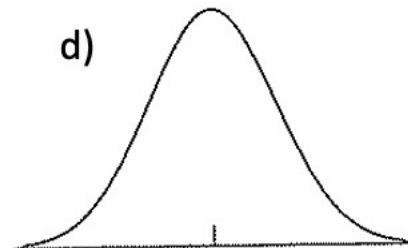
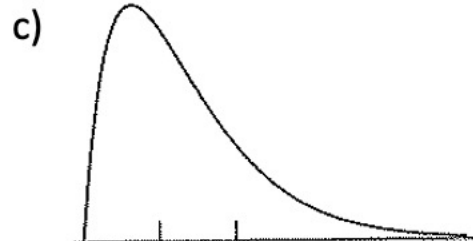
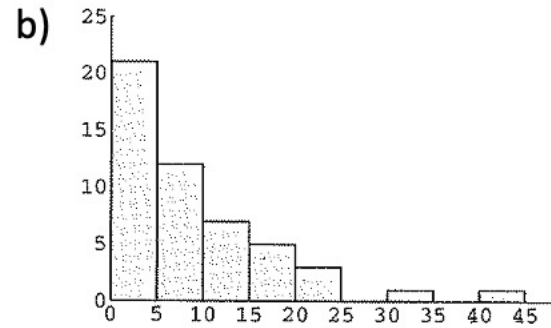
## Decisions

- Again we are looking at the shape to determine which is the best measure of variation
  - a) Start with using the SD because it is more informative than IQR
  - b) Then use the following two test (or conditions) on the data. If it fails either, then switch to using the IQR
    1. Shape: Symmetric or skewed?
    2. No outliers or outliers?
- Or we can decide based on which measure of center we reported (and pick the corresponding measure of variation)!

a) Stem-and-leaf plot for the hotel rate data

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

Stem = tens  
Leaf = ones



- a) SD (mean)
- b) IQR (median)
- c) IQR (median)
- d) SD (mean)
- e) IQR (median)

# Center and Spread: When to Report Which

When looking to describe data by the Center and Spread, ask yourself what the shape is and if there are outliers.

If the data is symmetric and has no outliers, report the Mean for the Center. If you report the Mean for the center, report the Standard Deviation for the Spread.

If the data is not symmetric or has outliers, report the Median for the Center. If you report the Median for the center, report the IQR for the Spread.

# LCQ / Example: SD of Grouped Data

**GOAL:** Find the Standard Deviation based on a Frequency Table!

1. Calculate Midpoints of each bin.
2. Enter data.
  - a) Midpoints (our estimates) go in  $L_1$ .
  - b) Frequencies (our weights) go in  $L_2$ .
3. 1-Var Stats
  - a) List is  $L_1$  (Midpoints).
  - b) FreqList is  $L_2$  (Frequencies).
  - c) Calculate!

Bin	<i>Midpoint</i>	Frequency
55-75	65	80
75-95	85	45
95-115	105	60
115-135	125	72

Show work:

## LCQ / Example: SD of Grouped Data

Bin	<i>Midpoint</i>	Frequency
55-75	<i>65</i>	80
75-95	<i>85</i>	45
95-115	<i>105</i>	60
115-135	<i>125</i>	72

**GOAL:** Find the Standard Deviation based on a Frequency Table!

1. Calculate Midpoints of each bin.
2. Enter data.
  - a) Midpoints (our estimates) go in  $L_1$ .
  - b) Frequencies (our weights) go in  $L_2$ .
3. 1-Var Stats
  - a) List is  $L_1$  (Midpoints).
  - b) FreqList is  $L_2$  (Frequencies).
  - c) Calculate!

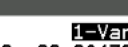
[illegible]

NORMAL FLOAT AUTO REAL RADIAN MP

**1-Var Stats**

List:L1  
FreqList:L2  
Calculate

**1-Var Stats**  
 $\bar{x}$ =94.64980545  
 $\Sigma x$ =24325  
 $\Sigma x^2$ =2449625  
 **$Sx=23.98473493$**   
 $\sigma x$ =23.93802634  
n=257  
minX=65  
↓Q1=65



NORMAL FLOAT AUTO REAL RADIAN MP

**1-Var Stats**

$\bar{x}$  = 23.98473493  
 $\sigma_x$  = 23.93802654  
 $n$  = 257  
 $\min X$  = 65  
 $Q_1$  = 65  
 $Med$  = 105  
 $Q_3$  = 125  
 $\max X$  = 125

***\*\* This is the SAME exact process as finding the mean of grouped data!  
Just looking for different information in the results***

Show work:

Midpoints calculated in Frequency Table  
 $S = 23.98$ , 1-Var Stats(List =  $L_1$ , FreqList =  $L_2$ )

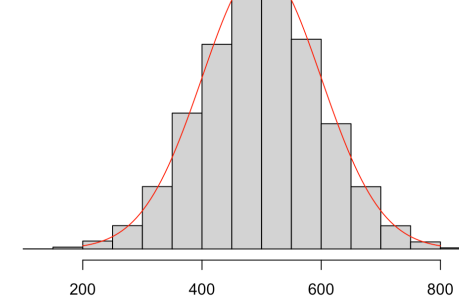
- *Technically this is a “weighted” or estimated standard deviation because it’s from grouped data and we don’t have the real numbers*
- *But we can just refer to it as the standard deviation like usual*



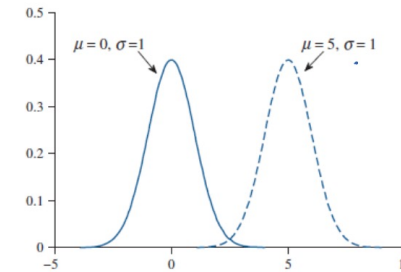
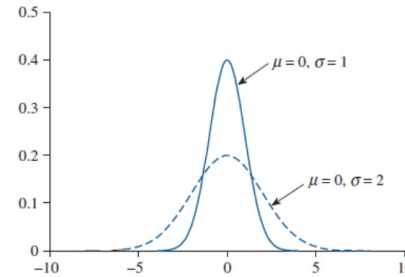
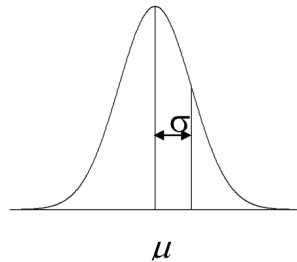
# Normal Distribution

## Normal Distribution

- This is a special distribution to describe data that is **bell-shaped**
- The smooth curve we can overlay is called a normal (“bell”) curve which describe normal distributions.
- A normal distribution has the following properties:
  - Describes a continuous random variable.
  - It’s a symmetric, unimodal, bell-shaped distribution.
  - Completely described by its **mean  $\mu$**  (which determines the location of the peak) and **standard deviation  $\sigma$**  (which measures the spread of the population)
- If a random variable X follows a Normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , we can summarize this in fancy stats notation with:



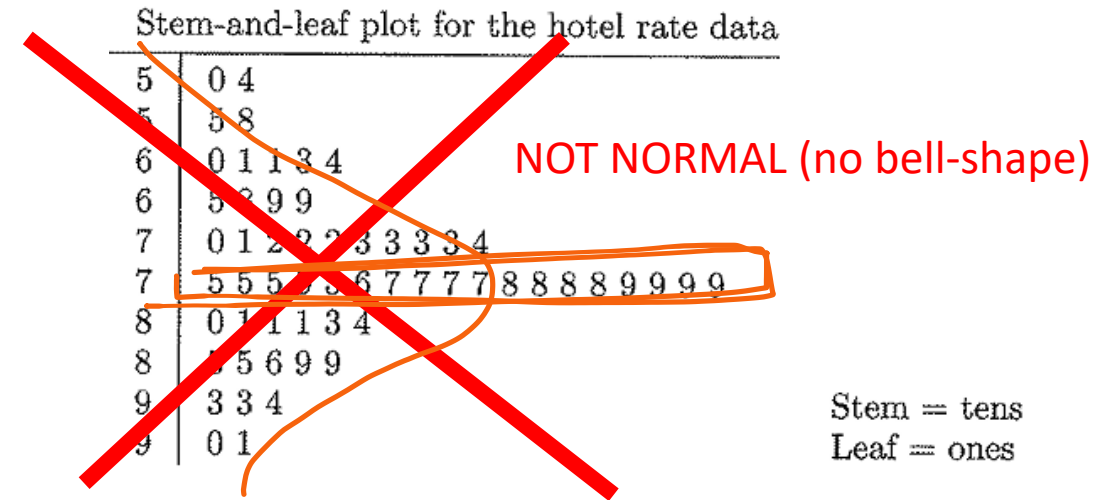
$$X \sim \text{Normal}(\mu, \sigma)$$



- If we know this information (Normal, mean and SD), we know everything about this distribution!
  - So, we can do lots of things, like find probabilities, make generalizations, etc.

# Bell-Shaped

- **Bell-shaped *implies* Normal distribution**
  - Can NOT say that a differently shaped symmetric, unimodal distribution is Bell-shaped.
- IF you say a boxplot is shape is bell → **WRONG!!!!**
  - DON'T KNOW THE MODALITY FROM A BOXPLOT

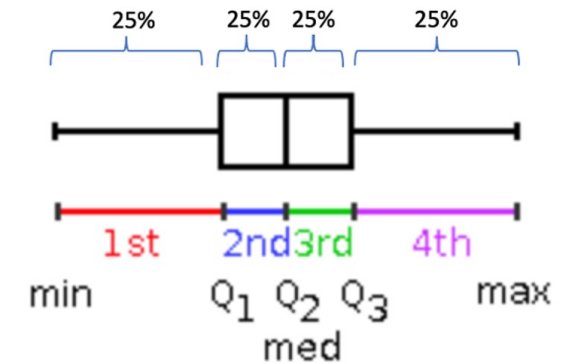


# Empirical Rule Motivation

## Before

- With boxplots, we could find ranges with certain probabilities
  - Ex) Middle 50% in between Q1 and Q3, lower 25% between Min and Q1, upper 50% between Med and Max, etc.
  - BUT, in order to calculate a 5 number summary (and draw the boxplot), we have to have every single data point...

1,1,2,2,3,(Q1)3,4,4,5,5,(M=Q2)6,6,7,7,8,(Q3)8,9,9,10,10



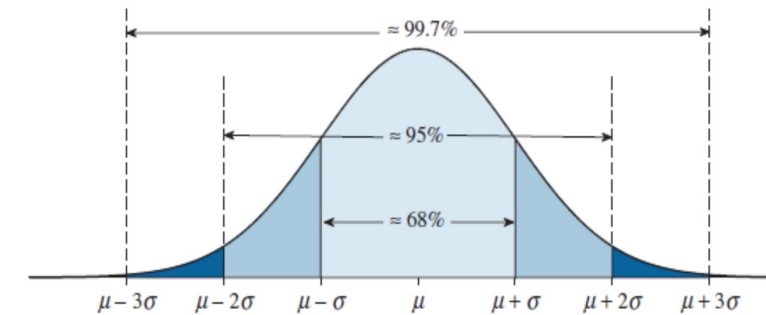
## Now

- The advantage of the Empirical Rule is that if we know these three things:
  1. Normal distribution / Bell-shaped
  2. Mean
  3. Standard deviation
- This can be used to make generalizations about a dataset / population without actually having the observations!
- We can estimate where the middle 68%, 95% and 99.7% (nearly all) are!

### Example

- Lets say ACT Scores follow a normal distribution with mean 22 and SD 4
- Based on this new rule, we know between which two scores almost all students are between:

$$22 \pm 3(4) = (10, 34)$$

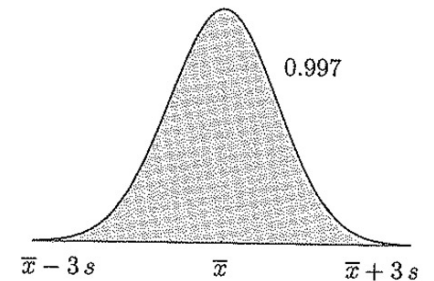
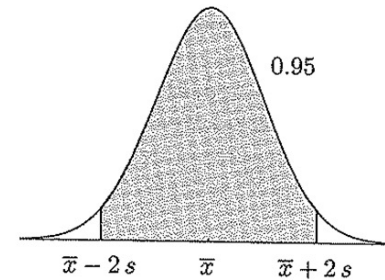
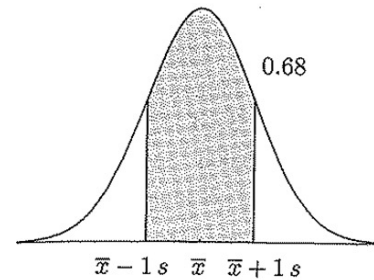
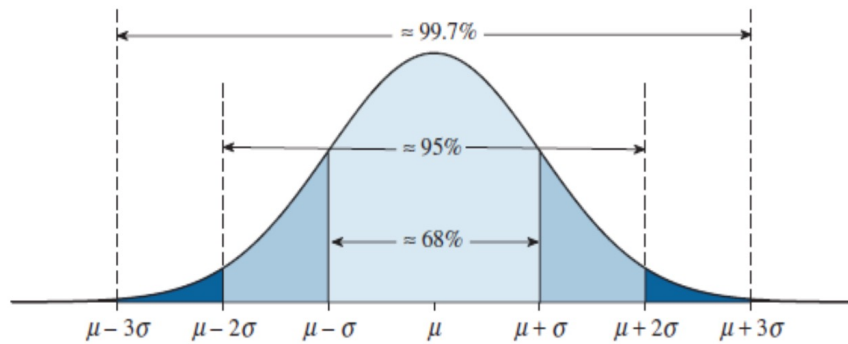


# Empirical Rule

## Empirical Rule (68, 95, 99.7 Rule)

If we have a normal distribution, we can generalize how much data lies within a certain distance from the mean. Specifically...

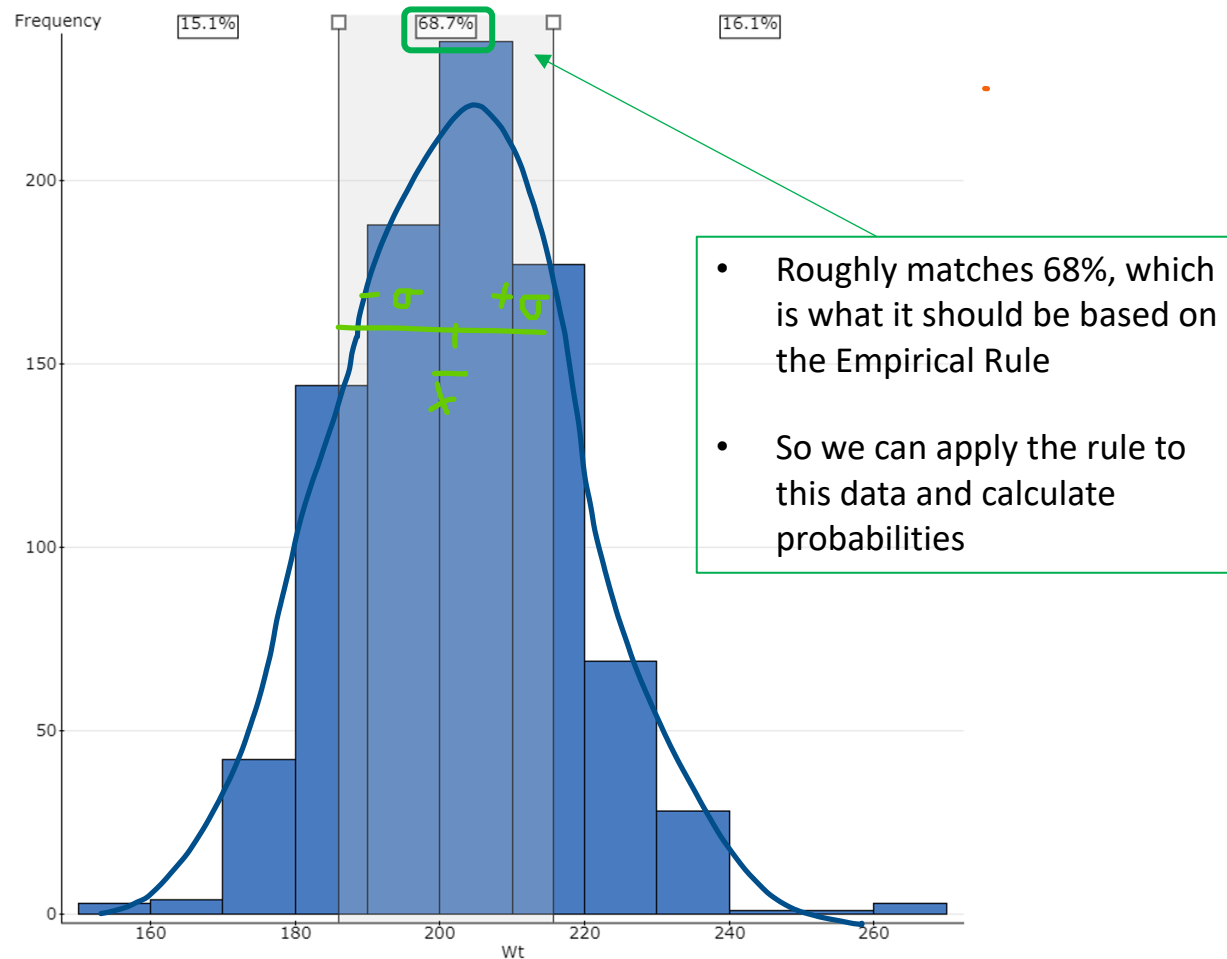
- Roughly 68% of our data falls within 1 Standard Deviation of the mean
- Roughly 95% of our data falls within 2 Standard Deviations of the mean
- Roughly 99.7% (nearly all) of our data falls within 3 Standard Deviations of the mean



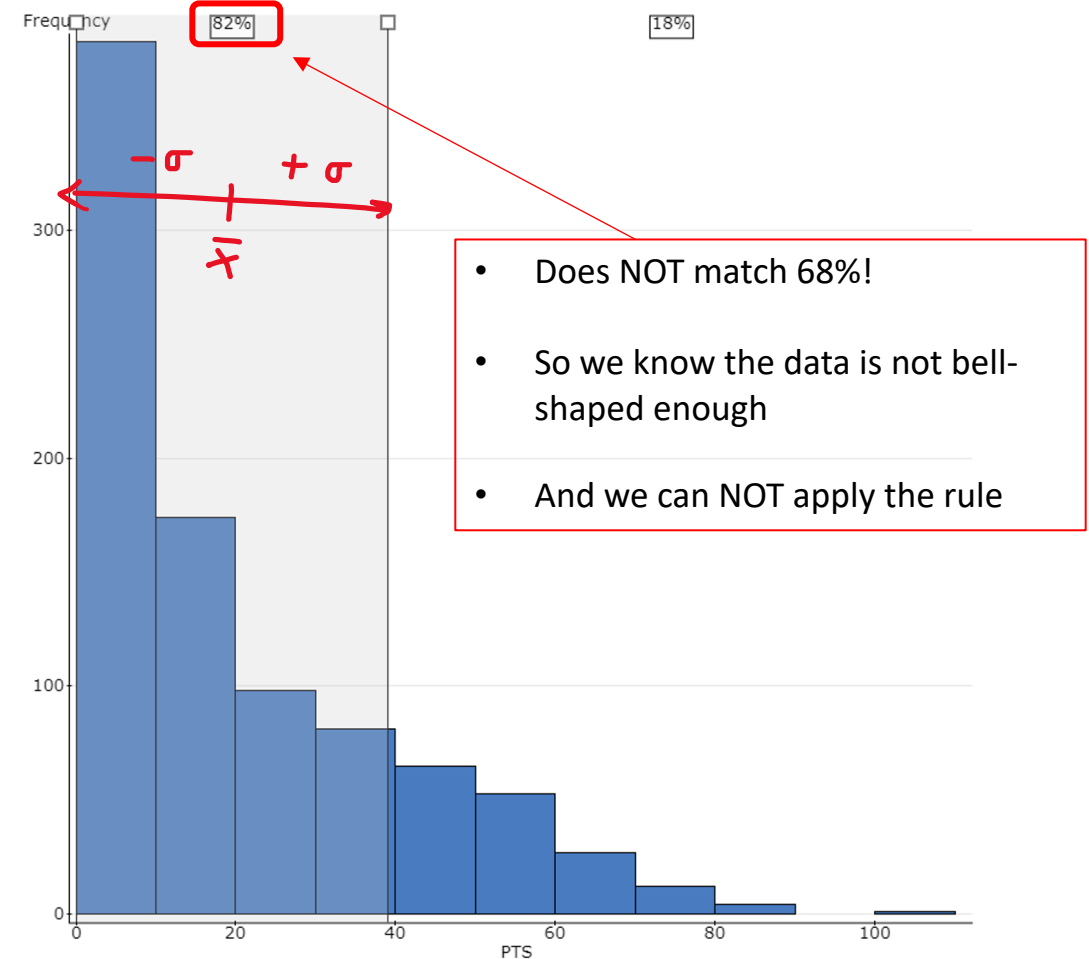
- In problems, it will have to be stated that you have a **Normal distribution**.
- We are only going to use the Empirical Rule with Normal Distributions.

# Empirical Rule: Examples for within 1 SD

## Where it works



## Where it doesn't work



# Example 1 – the Empirical Rule

**Assume** the distribution of each data set is approximately normal, with  $\mu$  and  $\sigma$  given.

**Find the intervals** (referred to by the Empirical Rule) that are one, two, and three standard deviations from the mean.

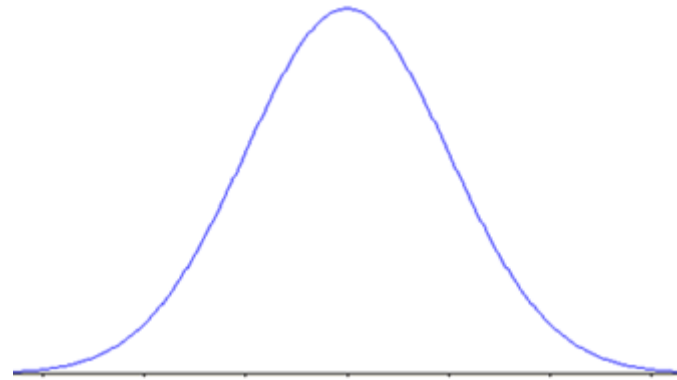
Carefully **sketch** the corresponding normal curve for each data set, indicating the endpoints of each interval.

a)  $\mu = 20, \sigma = 5$

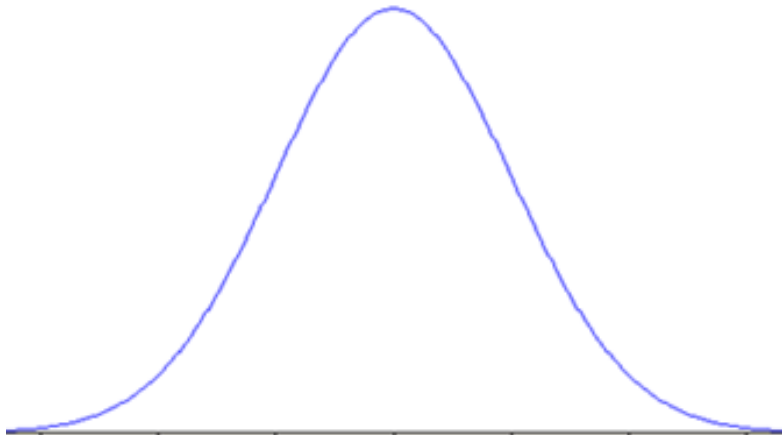
b)  $\mu = -5.5, \sigma = 12$

# Example 1 Solution

a)  $\mu = 20, \sigma = 5$



b)  $\mu = -5.5, \sigma = 12$



## How to Build Curve

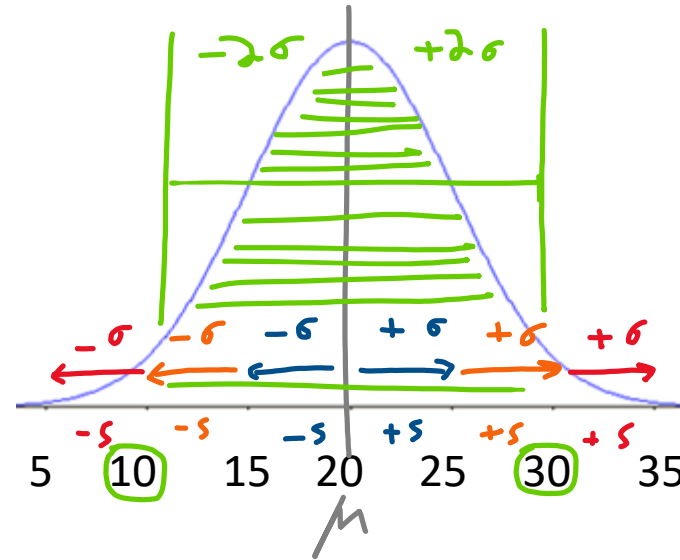
1. The mean is 20, so the curve is centered at 20
2. Then each step outwards is of size 5 because each step represents one standard deviation ( $\sigma$ )
  - And we go three steps in each direction from the mean

## Further example:

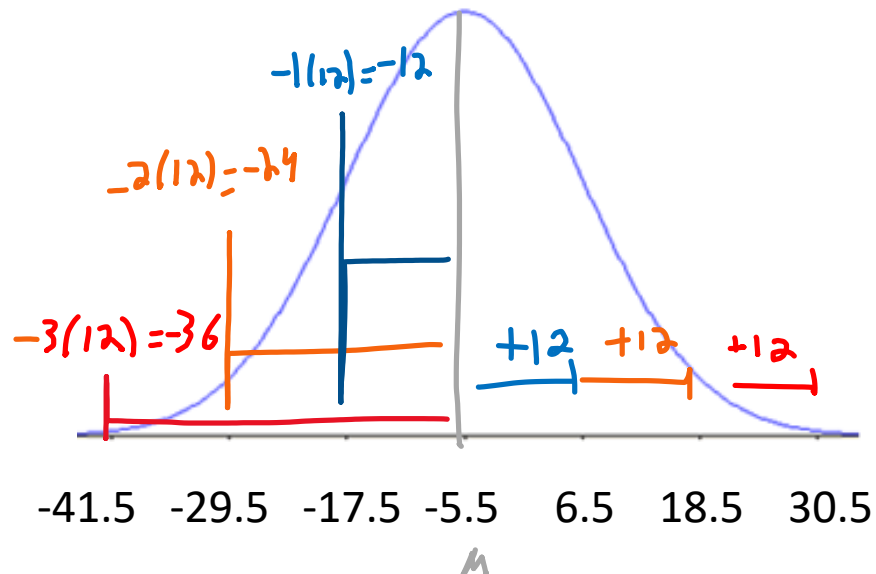
- Lets say this data for (a) was about the hours students spent at work per week.
- **Question:** What percent of the data is within 2 standard deviations of the mean?
- How can we **interpret** this in context?

# Example 1 Solution

a)  $\mu = 20, \sigma = 5$



b)  $\mu = -5.5, \sigma = 12$



## How to Build Curve

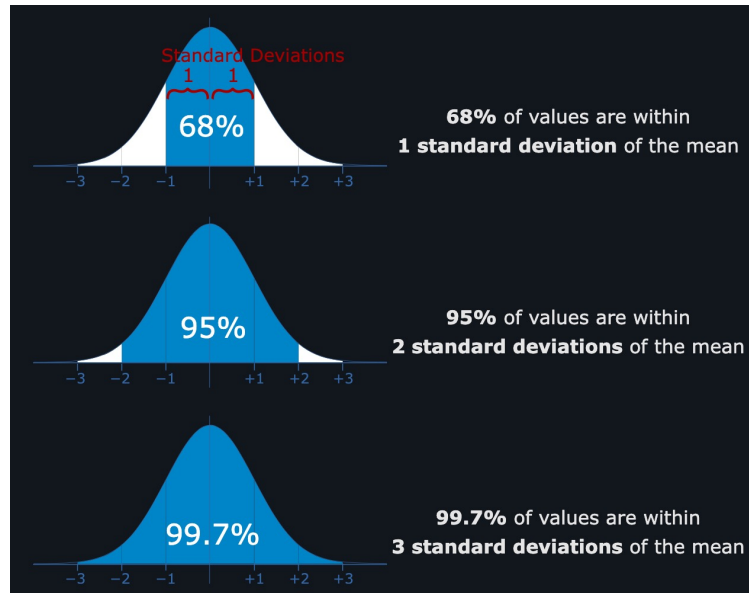
1. The mean is 20, so the curve is centered at 20
2. Then each step outwards is of size 5 because each step represents one standard deviation ( $\sigma$ )
  - And we go three steps in each direction from the mean

### Further example:

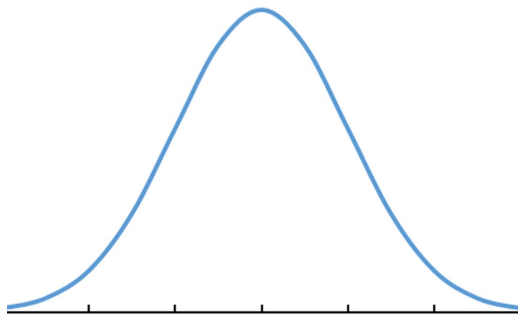
- Lets say this data for (a) was about the hours students spent at work per week.
- **Question:** What percent of the data is within 2 standard deviations of the mean?
- *95% based on rule*
- How can we **interpret** this in context?
- *2 SDs away makes the interval (10, 30)*
- *So according to the Empirical, approximately 95% of students work between 10 and 30 hours each week.*



# LCQ: Empirical Rule



<https://www.mathsisfun.com/data/standard-normal-distribution.html>



**Setup:** Oak trees heights follow a normal distribution with mean 75 m and standard deviation 7 m.

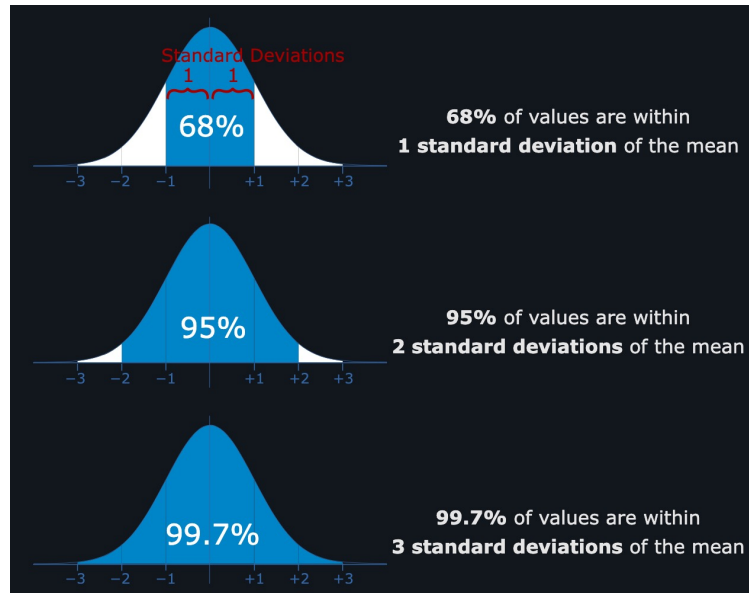
a) What heights do the middle 99.7% of trees lie between?

b) What percent of Oak trees are between 68 m and 82 m?

c) Which heights are two standard deviations from the mean?

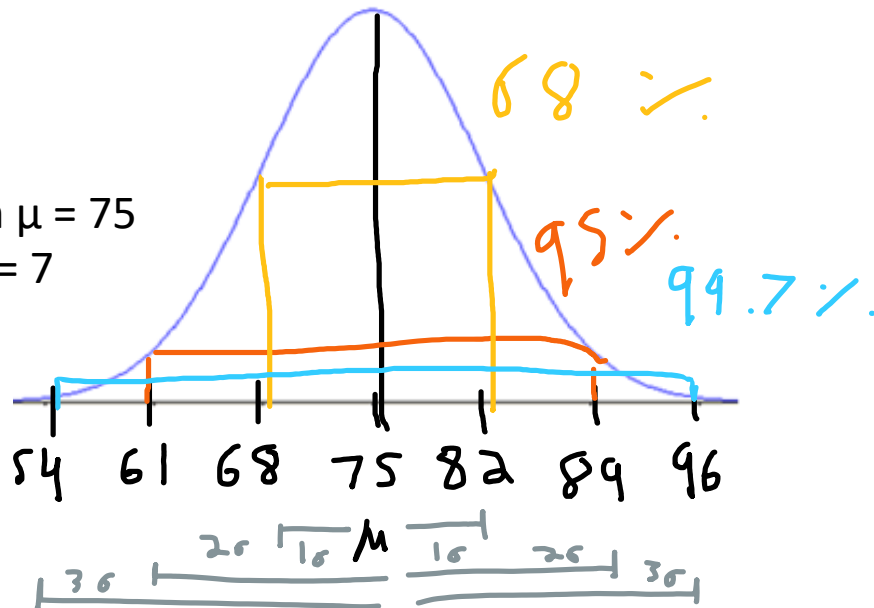
d) What percentage of trees are in between the heights from (c)?

# LCQ: Empirical Rule



<https://www.mathsisfun.com/data/standard-normal-distribution.html>

Mean  $\mu = 75$   
SD  $\sigma = 7$



**Setup:** Oak trees heights follow a normal distribution with mean 75 m and standard deviation 7 m.

a) What heights do the middle 99.7% of trees lie between?

*First need to set up the curve*

- Start with the mean and then step out the standard deviations
- Then find the interval we want based on the question!

*According to rule, 99.7% of values are within 3 SD of the mean*

$$\mu \pm 3\sigma = 75 \pm 3(7) = (54 \text{ m}, 96 \text{ m})$$

b) What percent of Oak trees are between 68 m and 82 m?

*68 m and 82 m correspond to 1 SD from the mean*

*According to rule, 68% of the data is between these heights*

c) Which heights are two standard deviations from the mean?

*Two SD is two steps out from the center, then look at the actual values*

$$\mu \pm 2\sigma = 75 \pm 2(7) = (61 \text{ m}, 89 \text{ m})$$

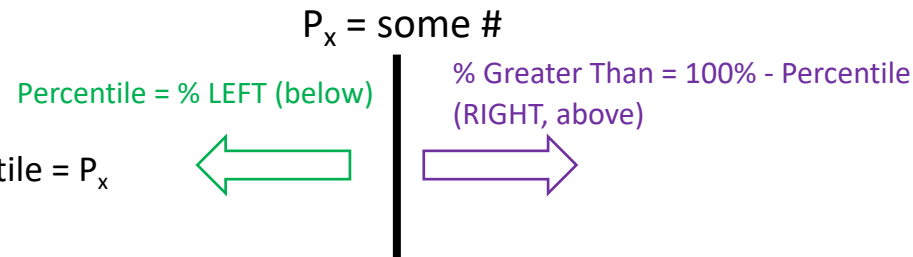
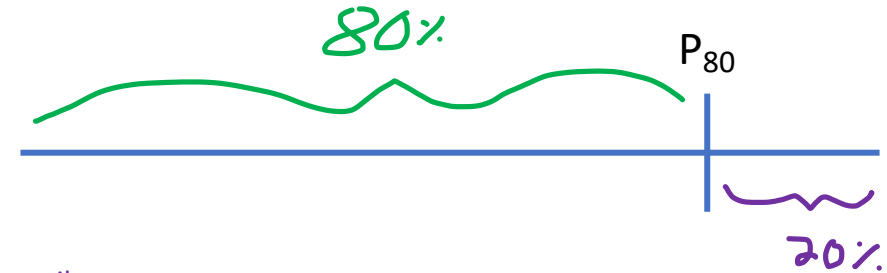
d) What percentage of trees are in between the heights from (c)?

*According to rule, 95% of values are within 2 SD of the mean*

# Percentiles

## Percentiles

- A **Percentile** tells you the percent of observations/individuals you are higher than.
  - Ex: You are told you scored in the 90th percentile on GRE. This means you have a score that is higher than 90% of all others that took the test.
- Range from greater than zero to 100<sup>th</sup> percentile!
- There is complement aspect to **percentiles** as well:
  - Example) If you are the 80<sup>th</sup> Percentile, there is 20% greater than you!
- **Best way to remember!**



- Notation: X<sup>th</sup> Percentile =  $P_x$

## How to Calculate

- To determine the percentile rank of a piece of data:
  1. Order data from smallest to largest (can use calc to do this)
  2. Count the number of pieces of data that are numerically below or equal to that value
  3. Divide by the sample size
- Example: 1, 4, 5, 5, 7
  - 4 is the second largest value out of 5:  $2/5 = 0.4 \rightarrow 40^{\text{th}}$  Percentile =  $P_{40}$
  - 5 is the 4th largest value out of 5:  $4/5 = 0.8 \rightarrow 80^{\text{th}}$  Percentile =  $P_{80} \rightarrow$  We want to use the second 5 because including less than OR equal to

This is the same idea cumulative relative frequency on frequency tables!

NOTE: You might see this defined differently elsewhere, but we will stick to this for our class

# Quartiles and IQR

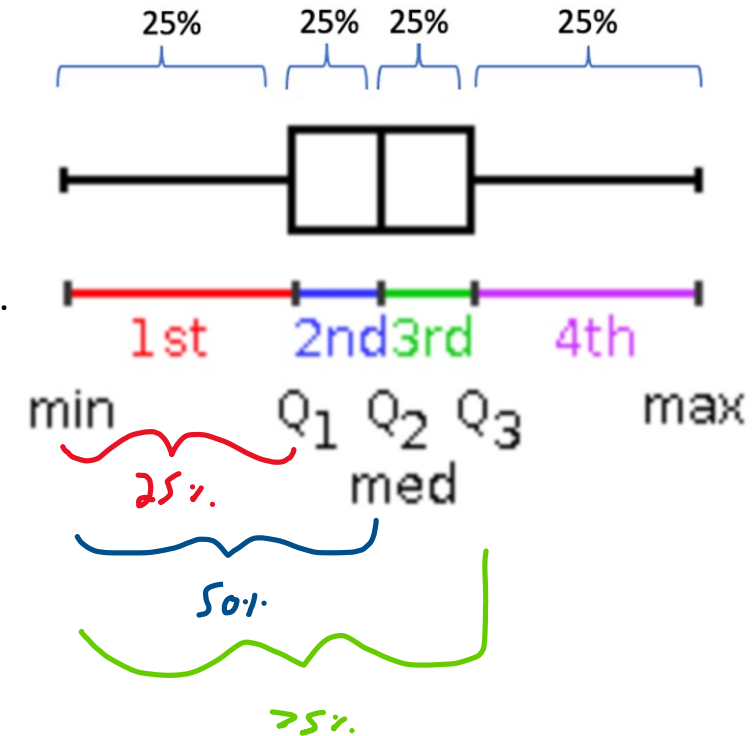
## Quartiles

- **Quartiles** are specific percentiles.
  - Q1 is the 25th Percentile.
  - Q3 is the 75th Percentile.
  - Q2 is the 50th Percentile (you will hear this called the Median more often).

## Inner Quartile Range (IQR)

- Remember this is another measure of spread.
- It looks at the middle 50% of the data
- Notation and formula:
  - $\text{IQR} = Q3 - Q1 = P_{75} - P_{25}$

1,1,2,2,3,(Q1)3,4,4,5,5,(M=Q2)6,6,7,7,8,(Q3)8,9,9,10,10



# LCQ: Percentiles and Quartiles

1) We have a sample of 14 cat weights in pounds. **Find** is the percentile for my cat Bubbles that weighs 32 lbs.

- Data: 23, 23, 25, 26, 27, 30, 31, 32, 33, 35, 36, 36, 36, 37

2) **Interpret** this percentile in context.

3) If my friend's cat Mr. Tibbles is in the upper 20% of cat weights, what is his percentile?

4) For any given dataset, what is the percentile for the maximum value?

5) **Find** the 5-number summary for the dataset in Question 1.

6) What weight is 75% of the cat weights greater than?

NORMAL FLOAT AUTO REAL RADIAN MP	
1-Var Stats	
↑Sx=	5.090391725
σx=	4.905224204
n=	14
minX=	23
Q1=	26
Med=	31.5
Q3=	36
maxX=	37

# LCQ: Percentiles and Quartiles

1) We have a sample of 14 cat weights in pounds. **Find** is the percentile for my cat Bubbles that weighs 32 lbs.

- Data: 23, 23, 25, 26, 27, 30, 31, 32, 33, 35, 36, 36, 36, 37

*Percentile = number of observations less or equal to / sample size =  $8 / 14 = 0.57 \rightarrow 57^{\text{th}}$  percentile*

2) **Interpret** this percentile in context.

*Multiple ways we could phrase it, both are correct and imply the same thing*

*57% of other cats weigh less than Bubbles*

*Bubbles is heavier than 57% of the other cats*

3) If my friend's cat Mr. Tibbles is in the upper 30% of cat weights, what is his percentile?

*70<sup>th</sup> percentile  $\rightarrow$  Remember the complement aspect to Percentiles*

*If there is 30% ABOVE (this is what you should think when read "upper 30%", there must be 100% - 30% = 70% BELOW. And Percentiles ALWAYS refer to the % BELOW*

4) For any given dataset, what is the percentile for the maximum value?

*100<sup>th</sup> Percentile  $\rightarrow$  All of the data is obviously equal to or less than the max, so 100%*

5) **Find** the 5-number summary for the dataset in Question 1.

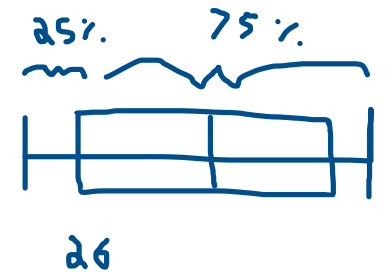
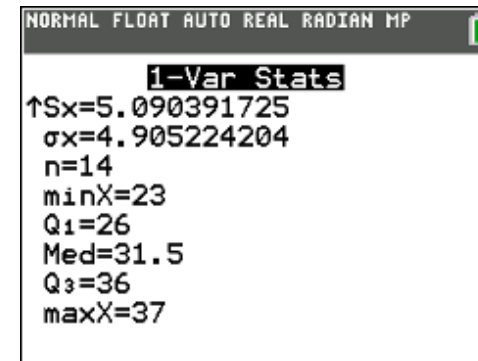
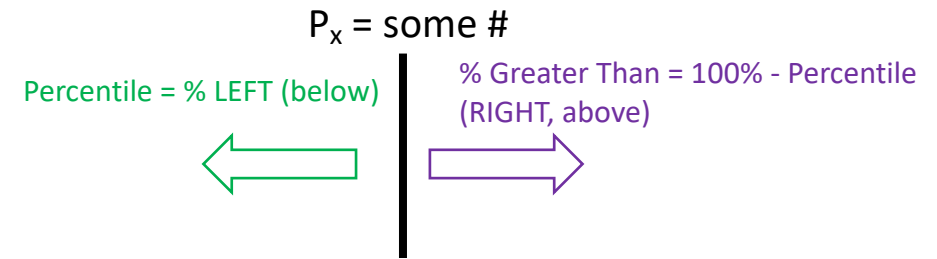
*1-Var Stat(List = L<sub>1</sub>)*

*Min = 23, Q1 = 26, Med = 31.5, Q3 = 36, Max = 37*

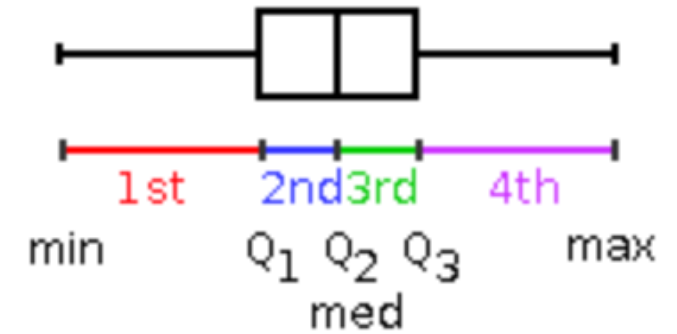
6) What weight is 75% of the cat weights greater than?

*75% greater == 25% less  $\rightarrow$  25<sup>th</sup> Percentile = Q1 = 26 lbs*

*This is just using the complement aspect and recognizing that Q1 corresponds to the 25<sup>th</sup> Percentile!*

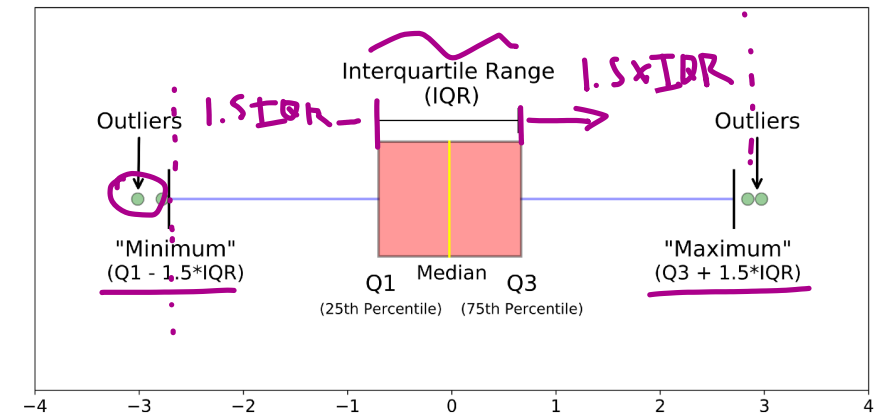


# Outliers via Boxplots



## Two Types of Boxplots

- Regular boxplot
  - Left fence is at the minimum and right fence is at the maximum.
- Modified (“Advanced”) boxplot
  - **Outliers** are drawn separately.
  - **Fences** are drawn at the next values that are NOT considered outliers!
  - How do we classify a point as an **outlier**?????



<https://towardsdatascience.com/understanding-boxplots-5e2df7bcd51>

## Outliers

- An **outlier** is an observation that is different/far away from the other observations.
- We define ‘far away’ as any data that are more than 1.5 times the box length away from the box.<sup>3</sup>
- This can be figured out by hand, but we are just going to draw modified boxplots to determine if / where the outliers are.

# LCQ: Outliers via Boxplots

## Problem

- Come back to the cat data, but now with two additional kittens:
  - Data: 5, 10, 23, 23, 25, 26, 27, 30, 31, 32, 33, 35, 36, 36, 36, 37
- **Determine** which (if any) of our new kittens are outliers by graphing the modified boxplot



# LCQ: Outliers via Boxplots

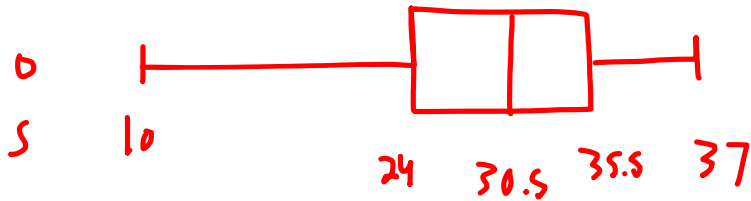
## Problem

- Come back to the cat data, but now with two additional kittens:
  - Data: 5, 10, 23, 23, 25, 26, 27, 30, 31, 32, 33, 35, 36, 36, 36, 37
- **Determine** which (if any) of our new kittens are outliers by graphing the modified boxplot

Show work (on Quiz or Exam)

*Entered data in L1, graphed modified boxplot*

*This shows data point @ 5 is an outlier, while 10 is NOT an outlier*

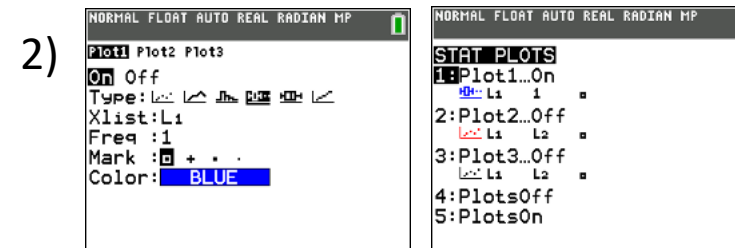
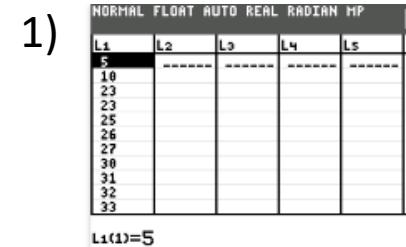


*Explain what you did in the calc, interpret the resulting image and give a sketch as well!*

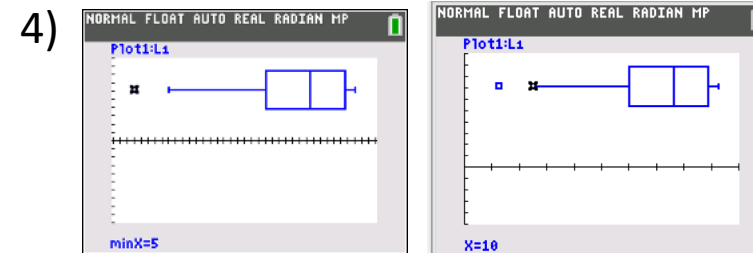
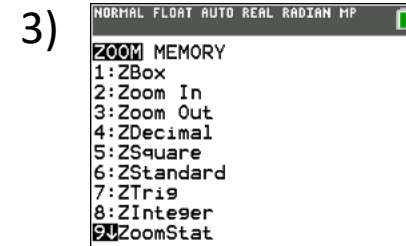
*ALL of this is what would be expected for FULL CREDIT!!*

### Calc Steps Again

- 1) Enter data in  $L_1$
- 2) Setting the STAT PLOT for the boxplot with outliers should already be done
  - We can leave this set so it is like our “default” option now
- 3) Now we will reset the Zoom
  - Remember we have to do this each time we make a new boxplot
- 4) Graph and Trace!



### Modified boxplot



# LCQ: Review

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Acres	36	46.50	47.76	6	18.50	33.50	55	250

Here are summary statistics for the sizes (in acres) of upstate New York vineyards.

- 1) If my house vineyard is in the 17<sup>th</sup> percentile, **explain** what this means?
- 2) From the summary statistics, would you **describe** this distribution as symmetric or skewed? **Explain**.
- 3) Using these summary statistics, **sketch** a boxplot.

# LCQ: Review

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Acres	36	46.50	47.76	6	18.50	33.50	55	250

Here are summary statistics for the sizes (in acres) of upstate New York vineyards.

1) If my house vineyard is in the 17<sup>th</sup> percentile, **explain** what this means?

*My vineyard is larger than 17% of the other upstate New York vineyards.*

2) From the summary statistics, would you **describe** this distribution as symmetric or skewed? **Explain.**

*The distribution is skewed right since the mean is greater than the median.*

3) Using these summary statistics, **sketch** a boxplot.



# How to compare values on different scales??

## Scenario

- You and your friend are very competitive and want to know who did better on their college entrance exam!
- But you took the ACT and your friend took the SAT
  - Lets say you scored a 30 on the ACT
  - And your friend scored a 1400 on the SAT

	Min	Max	Mean	St Dev
ACT	1	36	23	4
SAT	400	1600	1000	180

- Here is some info about the two exams overall:
- How can we figure out who did better???
  - Maybe we can think about who's score is further above the mean, but the scales are still wayyy different.....

# Z-Scores Example

The Mean Corporation has two major franchise operations each with many outlets across the country: a coffee house franchise and a fast food franchise. For both franchises, a statistician has collected and analyzed data for the annual profits of the outlets. Figures for the means and standard deviations are presented in the table below.

John Q owns both a coffee and fast food franchise. His annual profit for the coffee house is \$377,800 and his annual profit for the fast food franchise is \$838,600.

Relative to the rest of the outlets in each group, which of John Q's stores is performing better?

Franchise	Mean (\$)	Standard deviation (\$)
Coffee	245,500	49,000
Fast food	681,000	98,500

# Z-Scores Example

Relative to the rest of the outlets in each group, which of John Q's stores is performing better?

Start by calculating the z-score each. Why? **A z-score will let us see how each of John's stores is doing in their respective groups.**

Coffee:

$$Z_{Coffee} = \frac{(John's Store) - (Mean Coffee)}{(Standard Deviation Coffee)} = \frac{377,800 - 254,000}{49,000} = 2.7$$

John's Coffee Shop is 2.7 standard deviations away from the mean performance of all coffee shops

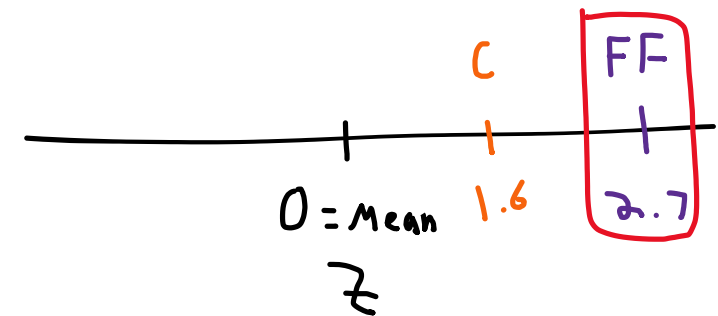
Franchise	Mean (\$)	Standard deviation (\$)
Coffee	245,500	49,000
Fast food	681,000	98,500

Fast Food:

$$Z_{FF} = \frac{(John's Store) - (Mean FF)}{(Standard of FF)} = \frac{838,600 - 681,000}{98,500} = 1.6$$

John's Fast Food store is 1.6 standard deviations above the mean of all fast food.

*Because John's Coffee shop has a higher Z-score, it is performing better than John's fast food within each store type.*



# Z-Scores

## Z-Scores

- A **z-score** standardizes observations based on the mean (center) and standard deviation (spread) of the distribution.
- Allows for comparisons between observations from different distributions (contexts).
  - Can think of it as a process to cancel out the units, puts everything on a common scale

## Formula

- $$Z = \frac{x - \mu}{\sigma} = \frac{\text{obs} - \text{mean}}{\text{st dev}}$$

## Interpretations (IMPORTANT!!)

- Technically, a **z-score** tells us how many standard deviations an observation is away from the mean.
- The unit of a **z-score** is standard deviations.
- Examples:
  - If you have z-score of 1.5, you are 1.5 standard deviations above the mean → Of course that distance depends on the value of sigma for the context.
  - Think about a Z-score as the number of 'Steps' from the mean, and the size of the step depends on the context!
  - $Z = 0$  is our reference point! When  $Z = 0$  you are equal to the mean.
  - If my z-score is negative, I am below the mean; Positive  $z$  = above mean
  - If my z-score is 2 and yours is 2.3, you have a relatively greater value than me.

## Outliers

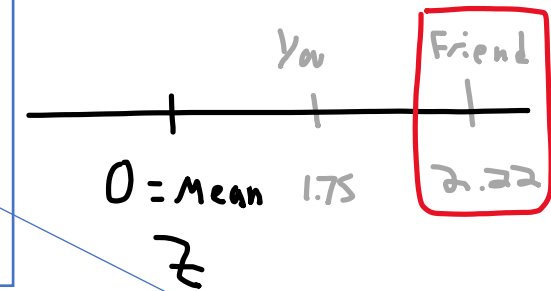
- Another way to detect outliers! → If a **z-score** is above 2 or below -2, we can say that the observation is unusual

### Return to Example

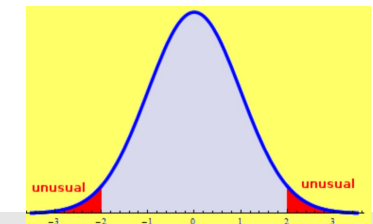
- Using Z-scores, we get the following results:
  - Try the calculations on your own!
- So, your friend performed better on their exam, relative to all others who took the same exam!

	Min	Max	Mean	St Dev
ACT	1	36	23	4
SAT	400	1600	1000	180

	Score	Z-Score
You (ACT)	30	1.75
Friend (SAT)	1400	2.22



Here, you are 1.75 'steps' above the mean  $\mu$  and each step is size 4, the original St Dev  $\sigma$



# LCQ: Z-scores

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

**Problem 1:** For each data set with the stated  $\mu$  and  $\sigma$ , **find** the z-score corresponding to the given observation,  $x$ .

a)  $\mu = 8, \sigma = 3, x = 17$

b)  $\mu = 100, \sigma = 16, x = 80$

c) Which observation is further from the mean, relatively? Are either of these considered unusual observations?

**Problem 2:** For each data set with the stated  $\mu$  and  $\sigma$ , **find** the original observation  $x$  corresponding to the given z-score.

d)  $\mu = 33, \sigma = 6, z = -1.2$

e)  $\mu = -40, \sigma = 1.5, z = 3$



# LCQ: Z-scores

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

**Problem 1:** For each data set with the stated  $\mu$  and  $\sigma$ , **find** the z-score corresponding to the given observation,  $x$ .

a)  $\mu = 8, \sigma = 3, x = 17$

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 8}{3} = 3 \rightarrow \text{just have to plug and chug}$$

b)  $\mu = 100, \sigma = 16, x = 80$

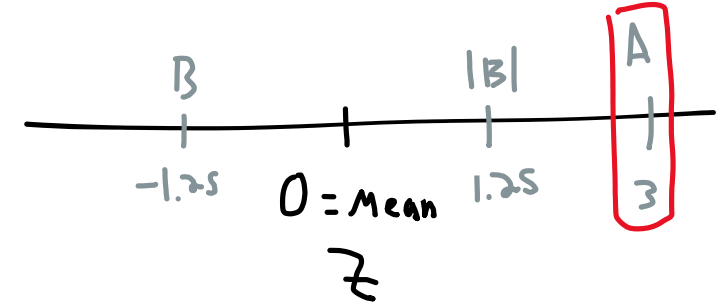
$$z = \frac{x - \mu}{\sigma} = \frac{80 - 100}{16} = -1.25$$

c) Which observation is further from the mean, relatively? Are either of these considered unusual observations?

$z = 3$  (a) is further from the mean (larger absolute value)  $\rightarrow$  Recall absolute value is just the positive #,  $|z| = \text{positive } z$

- So this just says which is further, regardless of direction

(a) would be considered unusual because more than 2 SD away from mean ( $|z| > 2$ )



**Problem 2:** For each data set with the stated  $\mu$  and  $\sigma$ , **find** the original observation  $x$  corresponding to the given z-score.

d)  $\mu = 33, \sigma = 6, z = -1.2$

Start with the formula that we know, and plug in the values we have:

$$z = \frac{x - \mu}{\sigma} \rightarrow -1.2 = \frac{x - 33}{6} \rightarrow \text{now solve for } x \text{ with algebra} \rightarrow x = 25.8$$

- 25.8 is below the mean which makes sense because there was a negative Z score

e)  $\mu = -40, \sigma = 1.5, z = 3$

$$z = \frac{x - \mu}{\sigma} \rightarrow 3 = \frac{x - (-40)}{1.5} \rightarrow \text{now solve for } x \text{ with algebra} \rightarrow x = -35.5$$

- Now -35.5 is above the mean, positive z-score

# Z-Scores and Standard Normal Curve

## Z-Score - Conceptual

We can standardize ANY distribution, i.e. turn it into z-scores.

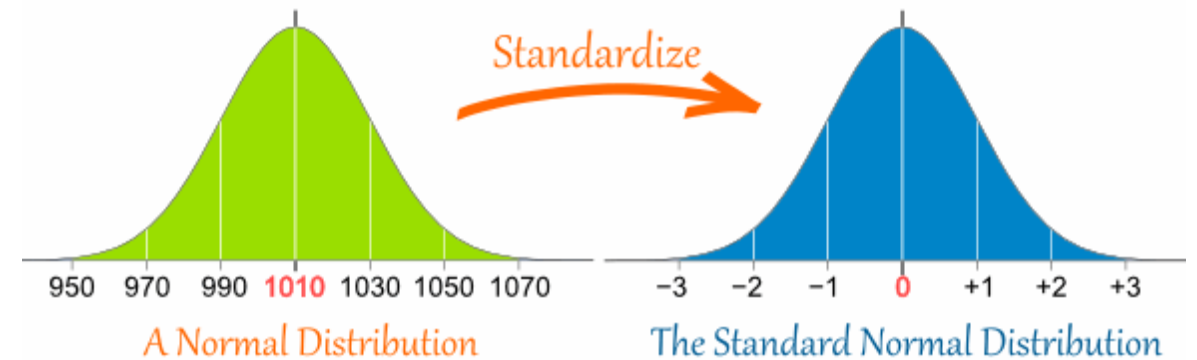
\*\* After standardizing (i.e. converting to Z), our new distribution has mean  $\mu_{new} = 0$

$$\text{if } x = \mu \rightarrow z = \frac{x - \mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0$$

and SD  $\sigma_{new} = 1$

(This is why the new Z-scores scale has “steps” equal to 1, which represent 1 SD, 1 Z-score)

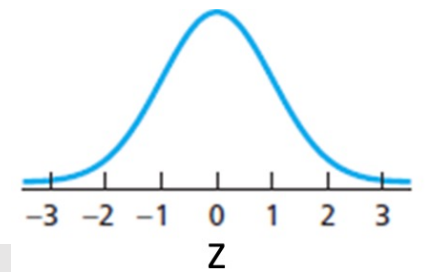
But when starting with a Normal distribution, we get a result that is very common!



<https://www.mathsisfun.com/data/standard-normal-distribution.html>

## Standard Normal Curve

- Just a specific Normal distribution that has mean  $\mu = 0$  and standard deviation  $\sigma = 1$
- If a random variable follows a Standard Normal distribution then,  
 $Z \sim \text{Normal}(\mu = 0, \sigma = 1)$



# Empirical Rule Again

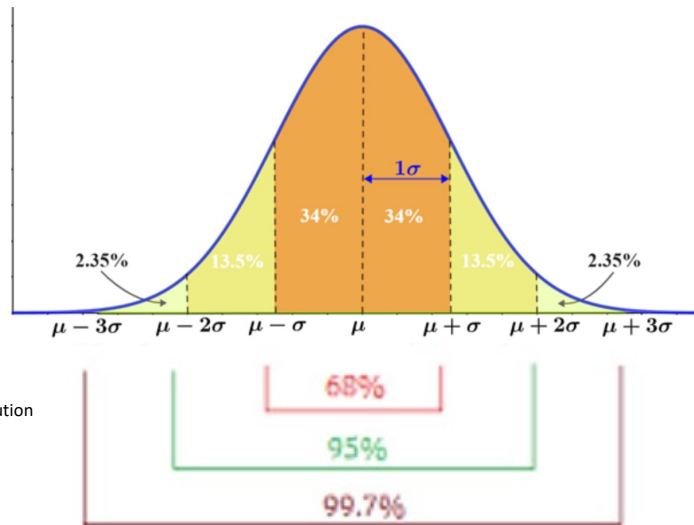
- Return to Empirical!

## Empirical Rule (68, 95, 99.7 Rule)

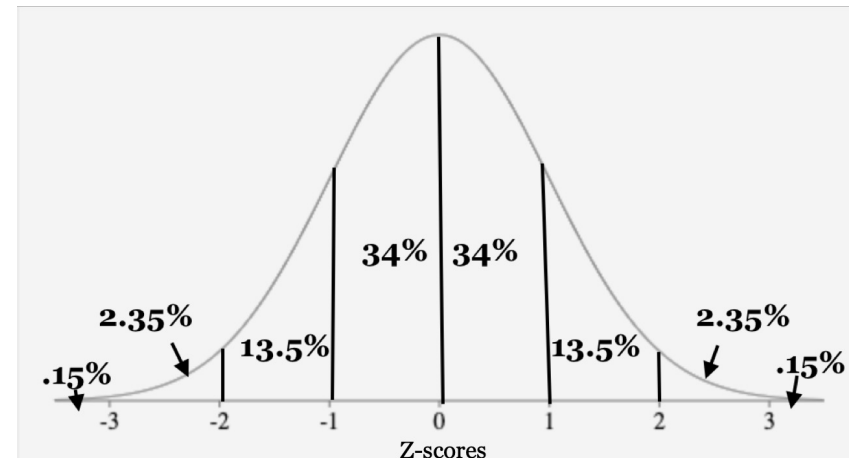
If we have a Normal distribution:

- Roughly 68% of the data is within 1 SD of the mean, **1 Z-SCORE!**
- Roughly 95% of the data is within 2 SDs of the mean, **2 Z-SCORES!**
- Roughly 99.7% (nearly all) of the data is within 3 SDs of the mean, **3 Z-SCORES!**

We can also break the curve down into smaller sections and find those probabilities!

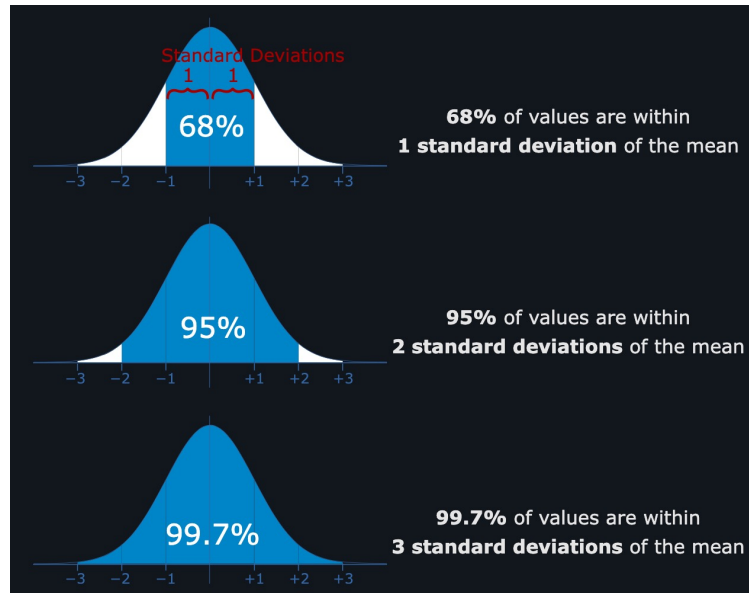


[math.net/normal-distribution](http://math.net/normal-distribution)



<https://spot.pcc.edu/~evega/>

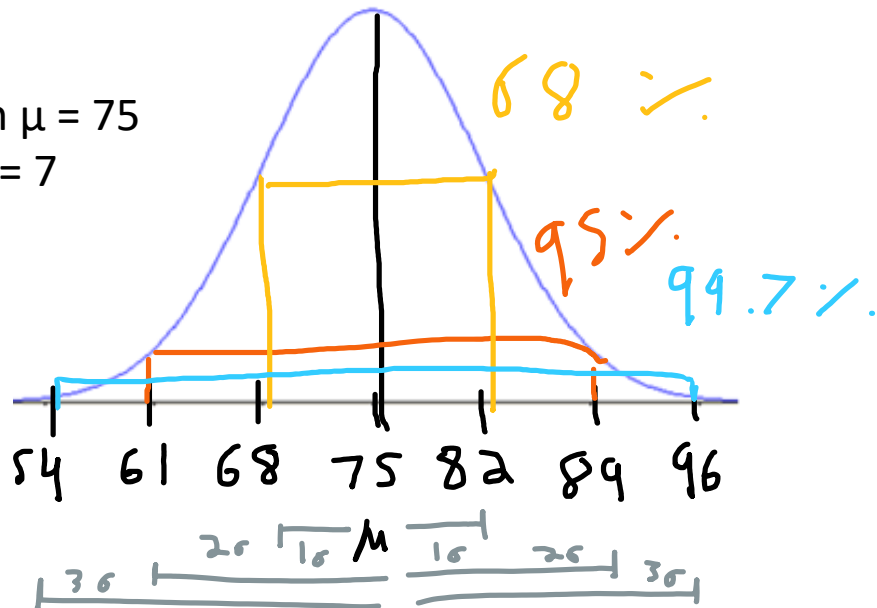
# LCQ: Summary



<https://www.mathsisfun.com/data/standard-normal-distribution.html>

Mean  $\mu = 75$

SD  $\sigma = 7$



**Setup:** Oak trees heights follow a normal distribution with mean 75 m and standard deviation 7 m.

a) The 50<sup>th</sup> Percentile of tree heights corresponds to what height?

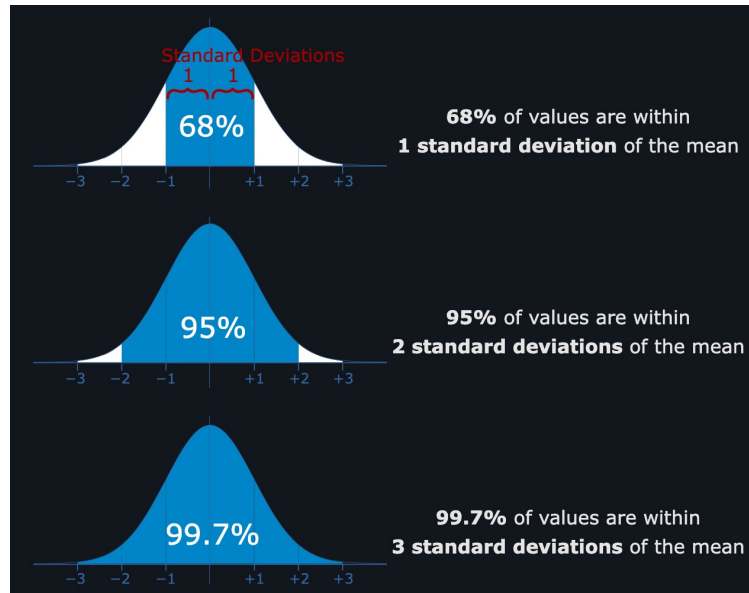
b) What percentile is a height of 68m?

c) What percent of trees are taller than 89 m?

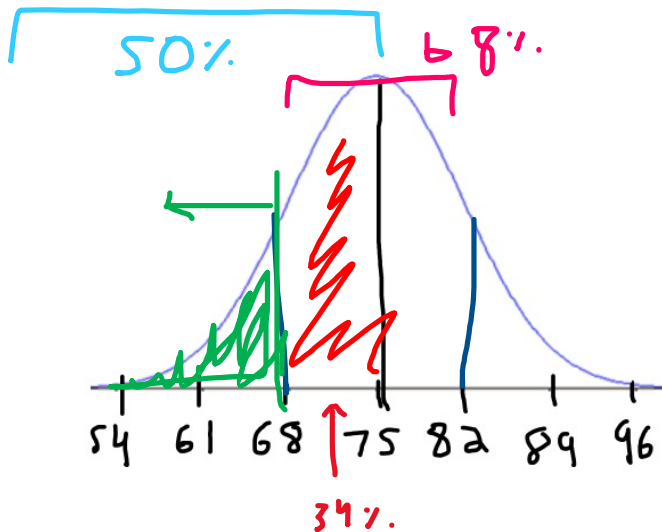
d) What heights would be considered outliers?

e) Is a tree height of 60 m more extreme than a height of 92 m? Why or why not?

# LCQ: Summary



<https://www.mathsisfun.com/data/standard-normal-distribution.html>



**Setup:** Oak trees heights follow a normal distribution with mean 75 m and standard deviation 7 m.

a) The 50<sup>th</sup> Percentile of tree heights corresponds to what height?

- *Percentile = % to the left*
- *This is a symmetric distribution, so the mean (which is equal to the median) splits distribution in half*

*75m*

b) What percentile is a height of 68m?

*This one is trickier, we just need to think about what we know based on the rule and break the curve down into parts*

- *We want to the left of 68*
- *We know to the left of 75 is 50%*
- *If we find what % is in between 68 and 75, we can just subtract from 50% to leave us with what we want!*

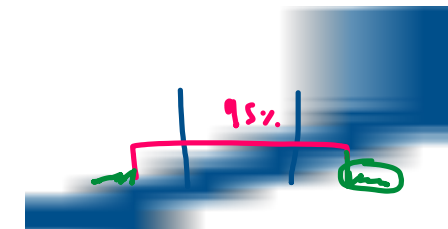
*In between 68 and 82 is 68% based on rule, and this is a symmetric interval. Which means  $0.68/2 = 0.34$  must be on the left side!  
 $0.5 - 0.34 = 0.16 \rightarrow 16^{\text{th}}$  Percentile*

c) What percent of trees are taller than 89 m?

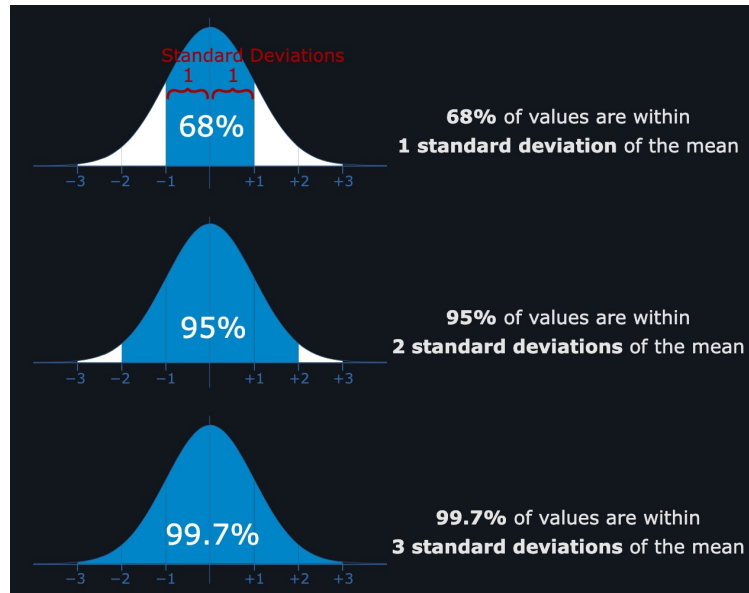
*Again, same process of breaking the curve into pieces and finding the areas we want*

- *We want to the right of 89 (the two SD mark)*
- *We know the middle interval at 2 SD is 95% based on the rule*
- *Based on the complement, this means  $100\% - 95\%$  is on the outside of that*
- *And we just want the right side*

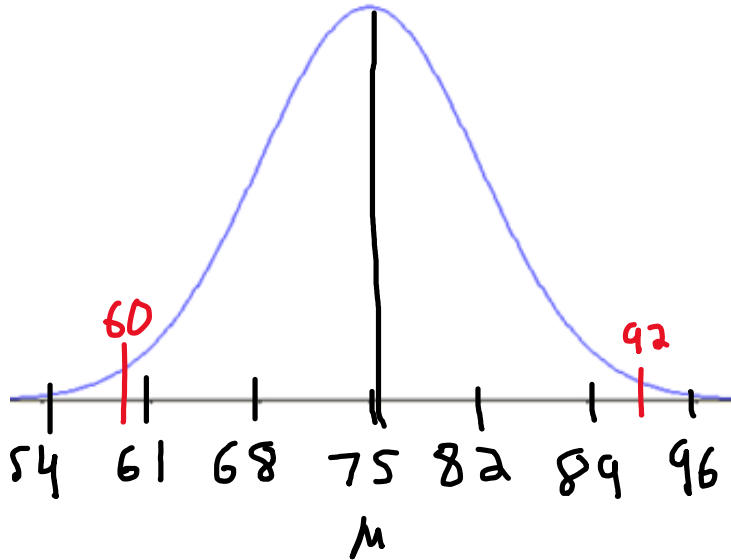
*$1 - 0.95 = 0.05 \rightarrow 0.05/2 = 0.025$*



# LCQ: Summary



<https://www.mathsisfun.com/data/standard-normal-distribution.html>



**Setup:** Oak trees heights follow a normal distribution with mean 75 m and standard deviation 7 m.

d) What heights would be considered outliers?

*More than 2 SD away from the mean is one way to classify points as outliers*  
*Heights below 61 m, above 89 m*

e) Is a tree height of 60 m more extreme than a height of 92 m? Why or why not?

- *Could probably tell visually by plotting, but to be sure we need to see which is further from the mean!*
- *The best way to do this is with Z-scores*

- $z = \frac{x - \mu}{\sigma} = \frac{60 - 75}{7} = -2.14$

- $z = \frac{x - \mu}{\sigma} = \frac{92 - 75}{7} = 2.42 \rightarrow$  Larger  $|z|$  score, so height of 90 is more extreme