STAT 201 Chapter 2

Exploring Data

Types of Variables

- **Variable**: any characteristic that is observed for the subject. There are two types of variables, categorical variable and quantitative variable.

- **Categorical**: Observations that belong to a set of categories.

  - Examples: Hair color, gender, zip code, etc.

- **Quantitative**: Observations that take on numerical values

  - Height, Weight, Income

Types of Variables

- **Quantitative**: Observations that take on numerical values

  - **Discrete**: measured by a whole number

    Examples: Number of books, children, money, etc

  - **Continuous**: measured on an interval

    Examples: Time, weight, distances

How to Compare Discrete and Continuous

- If you think of time: going from 1 min to 2 min we have to hit all of the times, e.g. 1.5 min or 1 min 30 sec
- If you think of weight: going from 150 lbs to 140 lbs we have to be every weight between 140 and 150, e.g. 144 lbs

- If you think of the number of books and children, we jump from one number to the next, 2.5 books, 1.5 children means nothing.

- Time and weight are continuous variables. Books and children are discrete variables.

How to Compare Discrete and Continuous

- The big difference here is that we can keep coming up with smaller units for the **continuous** case and we stop at some point from the **discrete** case.

- It should be noted that when we talk about **continuous** variables, we stop somewhere so we are measuring them **discretely** for convenience. (e.g. 100 mil to Columbia)

Data Type: Example

- Let’s consider a random sample of five residents of Columbia

  - **Days:** Number of days spent on workout weekly
  - **Piercings:** Number of body piercings
  - **Gym:** Do they go to the gym or not
  - **Type:** Do they lift, run, neither or both
  - **Age:** Age of person, in years
  - **Gender:** Male or Female

Data Type: Example

Days	Piercings	Gym	Type	Age	Gender
2	0	No	Neither	46	Female
3	1	Yes	run	21	Female
1	0	Yes	run	64	Male
6	2	Yes	Both	18	Female
0	0	No	Neither	19	Female

- **Days:** Number of days spent on workout weekly
- **Piercings:** Number of body piercings
- **Gym:** Do they go to the gym or not
- **Type:** Do they lift, run, neither or both
- **Age:** Age of person, in years
- **Gender:** Male or Female

Data Type: Example

Days	Piercings	Gym	Type	Age	Gender
2	0	No	Neither	46	Female
3	1	Yes	run	21	Female
1	0	Yes	run	64	Male
6	2	Yes	Both	18	Female
0	0	No	Neither	19	Female

- Which variables are Categorical?
- Which variables are Quantitative(Discrete)?
- Which variables are Quantitative(Continuous)?

Data Type: Example

Days	Piercings	Gym	Type	Age	Gender
2	0	No	Neither	46	Female
3	1	Yes	run	21	Female
1	0	Yes	run	64	Male
6	2	Yes	Both	18	Female
0	0	No	Neither	19	Female

- **Categorical**: Gym, Type, Gender
- **Quantitative(Continuous)**: Days, Age
- **Quantitative(Discrete):** Piercings

Categorical Summary: Frequency Table

- Let’s say we had 160 people in our sample instead of the 5 in the previous example and we want to get a better look at the type of workout that a resident of Columbia has.

Type	Frequency	Relative Frequency
Lift	32	
Run	64	
Both	16	
Neither	48	
Total	**160**	

Categorical Summary: Frequency Table

Type	Frequency	Relative Frequency
Lift	32	32/160=0.2
Run	64	64/160=0.4
Both	16	16/160=0.1
Neither	48	48/160=0.3
Total	**160**	**160/160=1**

- Let's fill out the relative frequency column. The **relative frequency** is the percent of the total sample, of 160, that had the data point we're looking at.
- $\boldsymbol{Relative\ Frequency} = \frac{(\#\ of\ subjects\ in\ each\ case)}{total\ \#\ of\ subjects\ in\ total\ sample}$

Categorical Summary: Frequency Table

Type	Frequency	Relative Frequency
Lift	32	32/160=0.2 → 20%
Run	64	64/160=0.4 → 40%
Both	16	16/160=0.1 → 10%
Neither	48	48/160=0.3 → 30%
Total	**160**	**160/160=1 → 100%**

- I think all of us would rather look at percentages than decimals, right?
- **Percentage** = (Decimal*100)%

Categorical Summary: Frequency Table

Type	Frequency	Relative Frequency
Lift	32	32/160=0.2 → 20%
Run	64	64/160=0.4 → 40%
Both	16	16/160=0.1 → 10%
Neither	48	48/160=0.3 → 30%
Total	**160**	**160/160=1 → 100%**

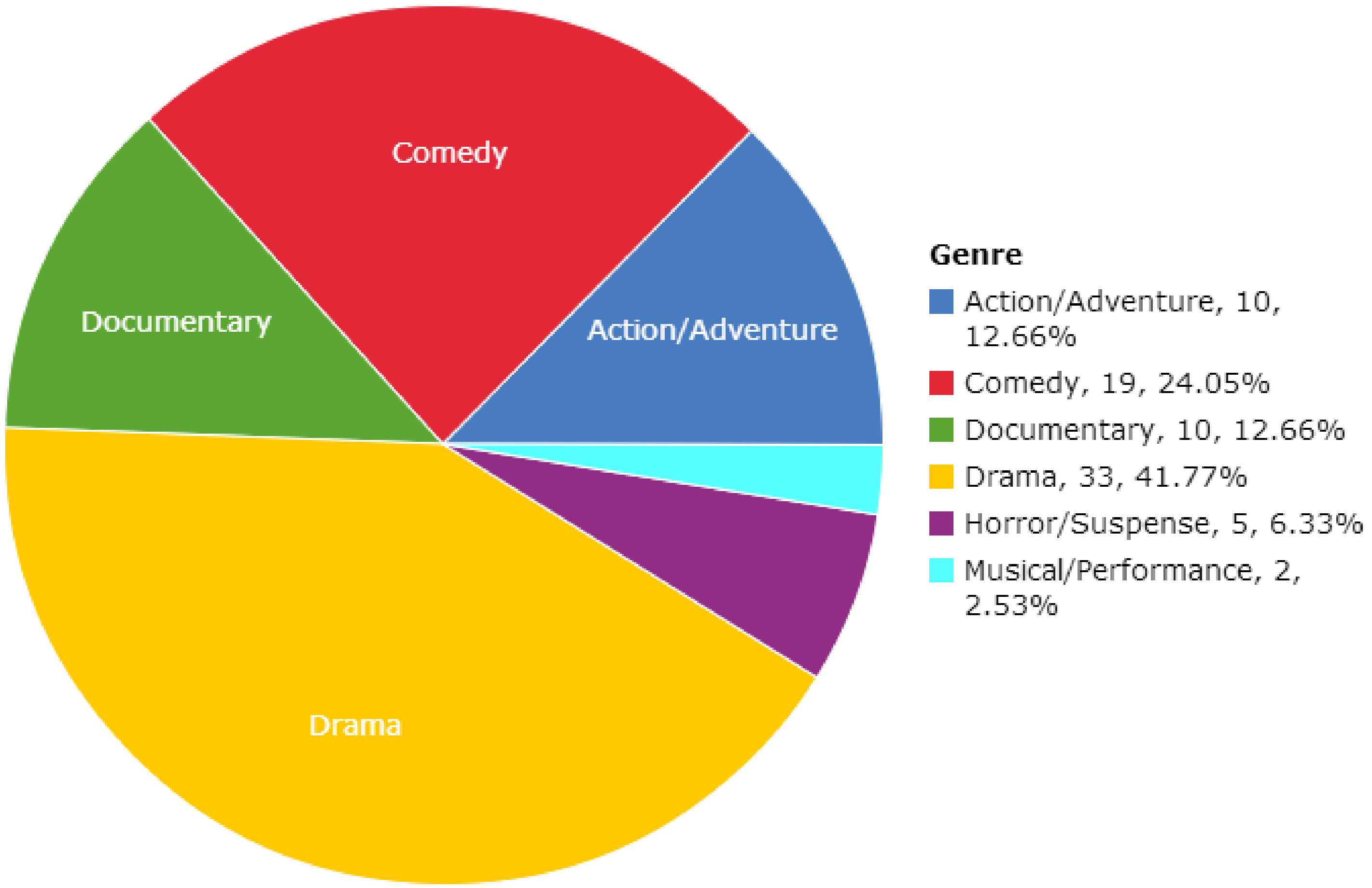
- **Q:** How many people workout with at least 1 type?
- **A:** We can just add the frequencies:
  - 32+64+16 = 112 people in our sample

English – This might be the Hardest Part!

- **At least x:** x or any number greater
  - At least 5 = 5, 6, 7, …
- **At most x:** x or any number lesser
  - At most 5 = …, 1, 2, 3 ,4, 5
- **Less than x:** any number smaller than x
  - Less than 5 = … 1, 2 , 3, 4
- **More than x:** any number larger than x
  - More than 5 = 6, 7, 8, 9, …
- **Between x and y:** we will say any number larger than x and less than y excluding x and y
  - Between 5 and 10 = 6, 7, 8, 9

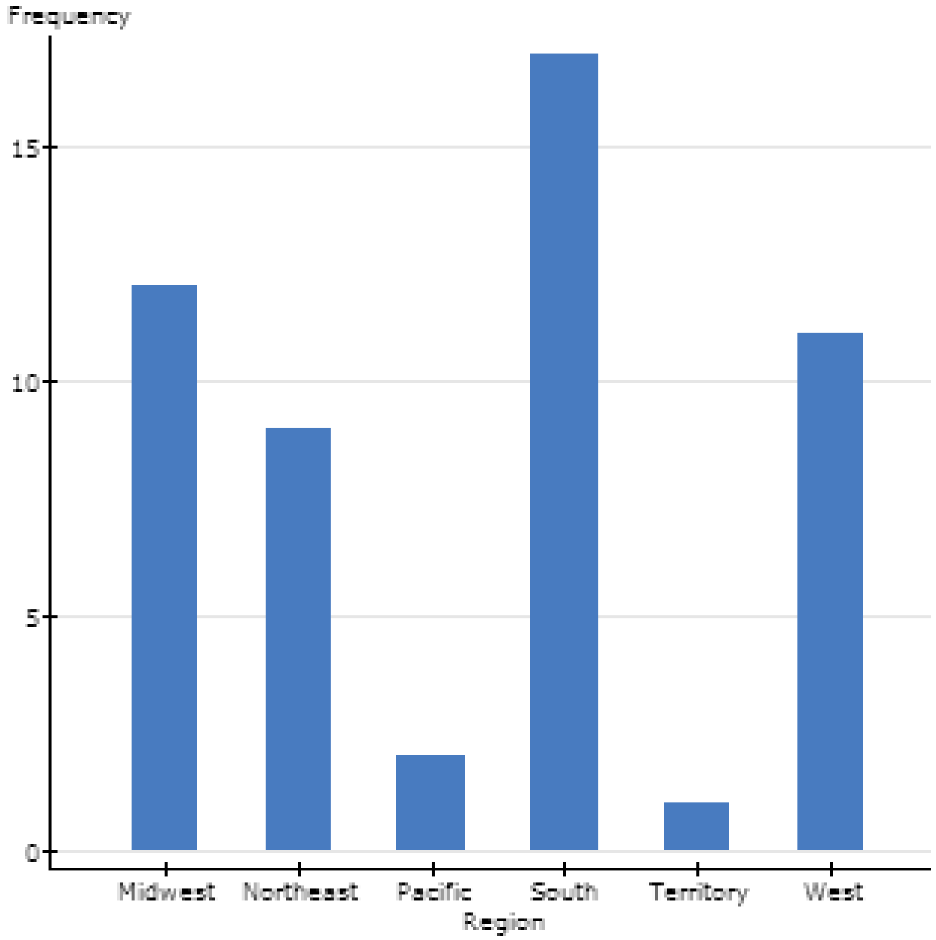
Categorical Summary: Pie Chart

- Useful when there are a small number of categories



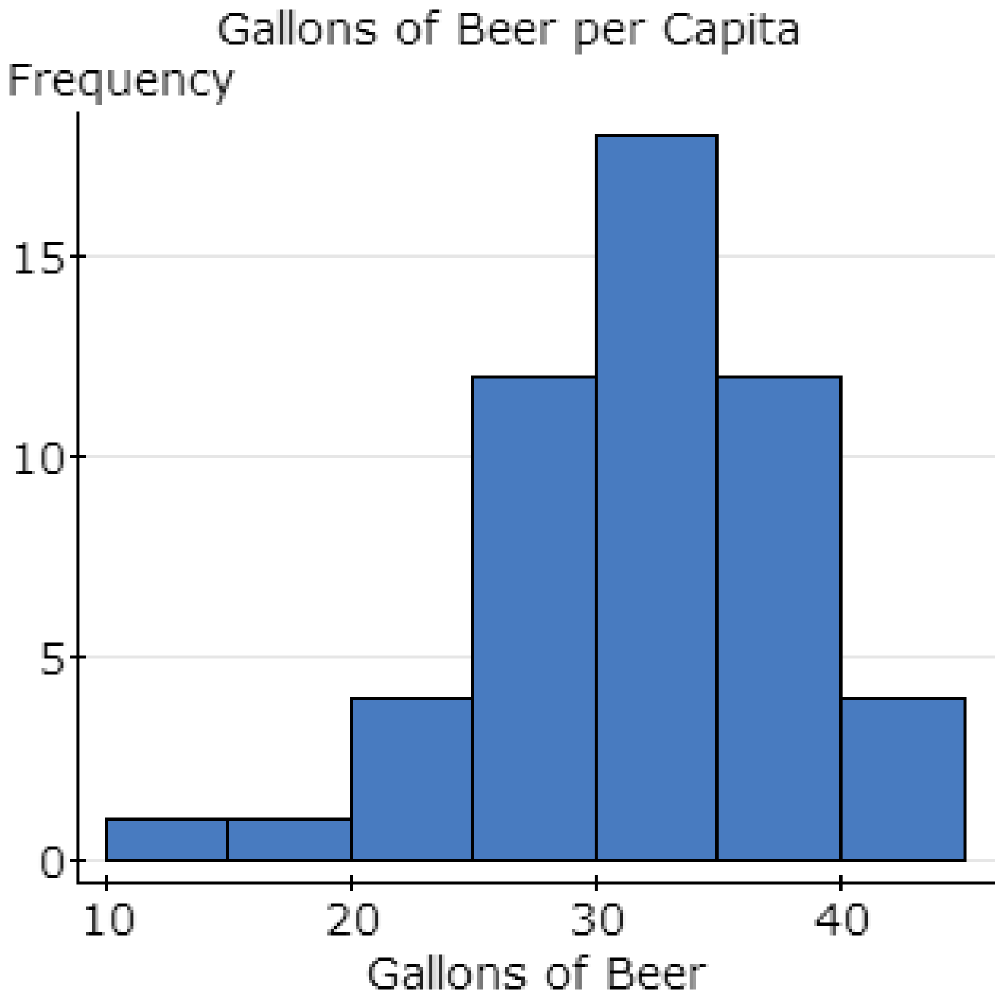
Categorical Summary: Bar Graph

- Useful when there are many categories of the variable
- Useful to compare groups



Quantitative Summary: Histograms

- Good for large data and for showing the shape of distribution
- We will use these a lot!

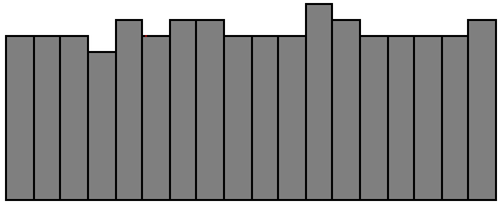


Histogram v.s. Bar Graph

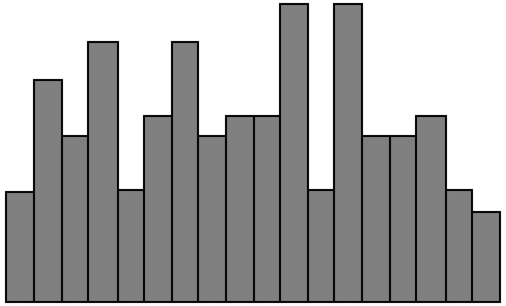
- With bar graph, each column represents a group defined by a categorical variable.

- With histograms, each column represents a group defined by a quantitative variable.

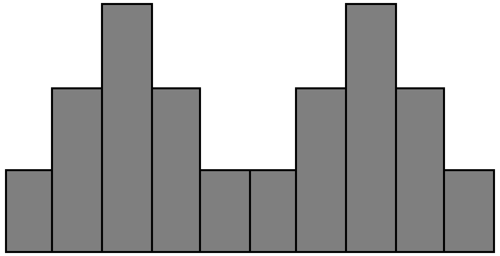
Quantitative Summary: Histogram Shapes



Uniform

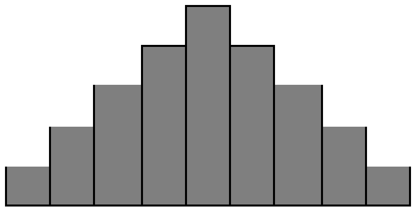


Random

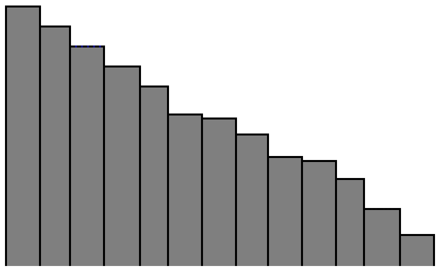


Bimodal

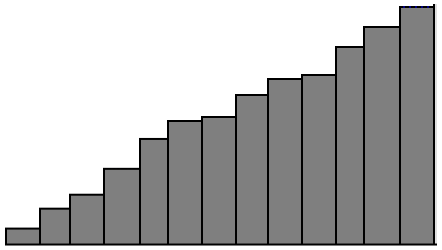
Quantitative Summary: Histogram Shapes



Bell-shaped - Unimodal



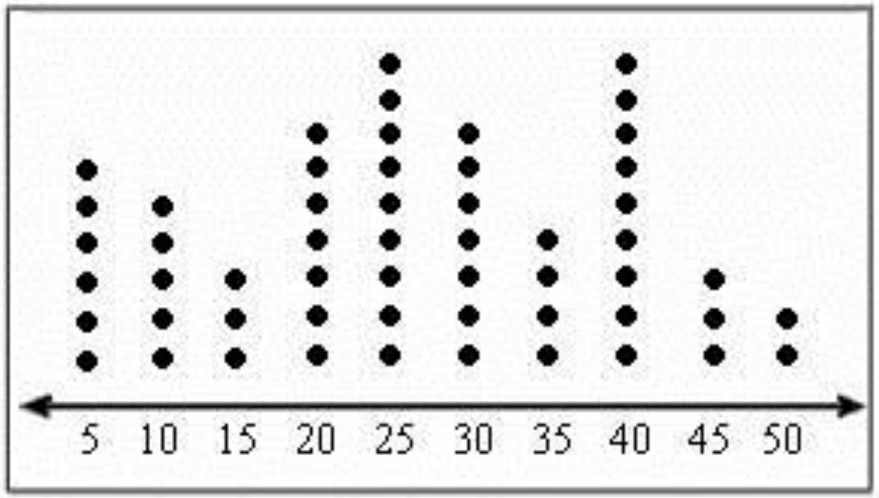
Skewed Right



Skewed Left

Quantitative Summary: Dot Plot

- Useful for smaller datasets
- Useful for finding outliers
- I don’t like these “dots” – histograms are almost always better



Quantitative Summary: Stem and Leaf

- Retain actual data values

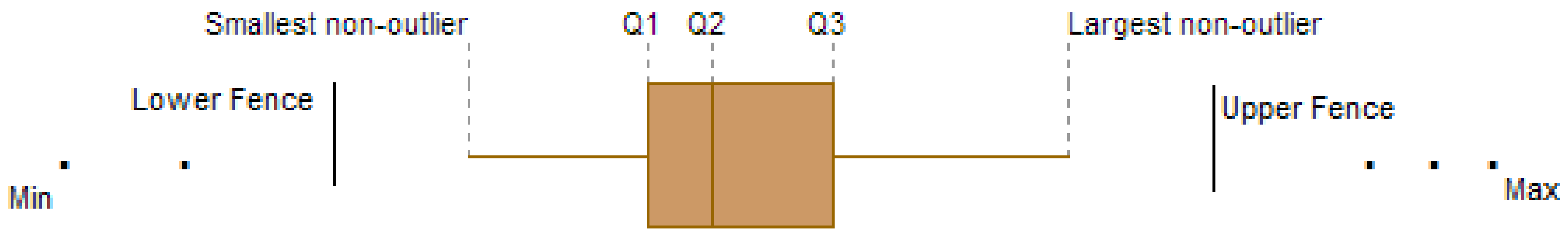
Example: Number of calories for a large serving of French Fries at Fast Food Restaurants
(source: http://www.acaloriecounter.com/fast-food.php)

570	500	500	540	566	631	610
400	400	640	550	700	280	380
480	430	370	380	490	310	620
450	730	260				

Stem Unit = hundreds, Leaf Unit = Tens
Variable: Calories

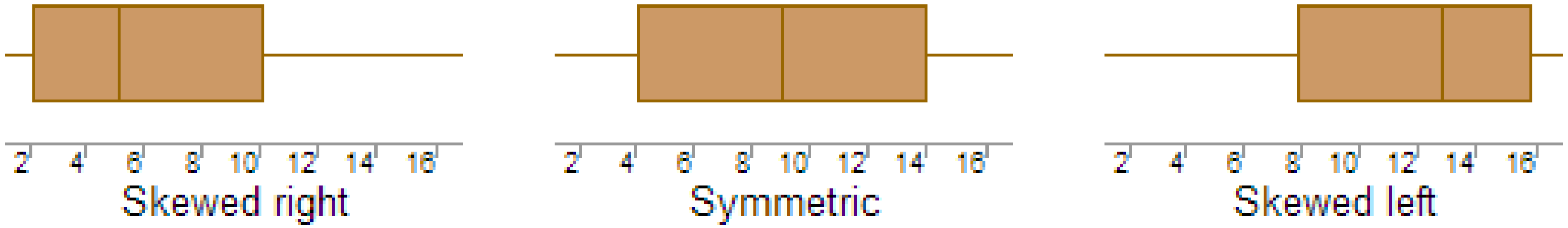
2 : 68
3 : 1788
4 : 003589
5 : 004577
6 : 1234
7 : 03

Box Plots

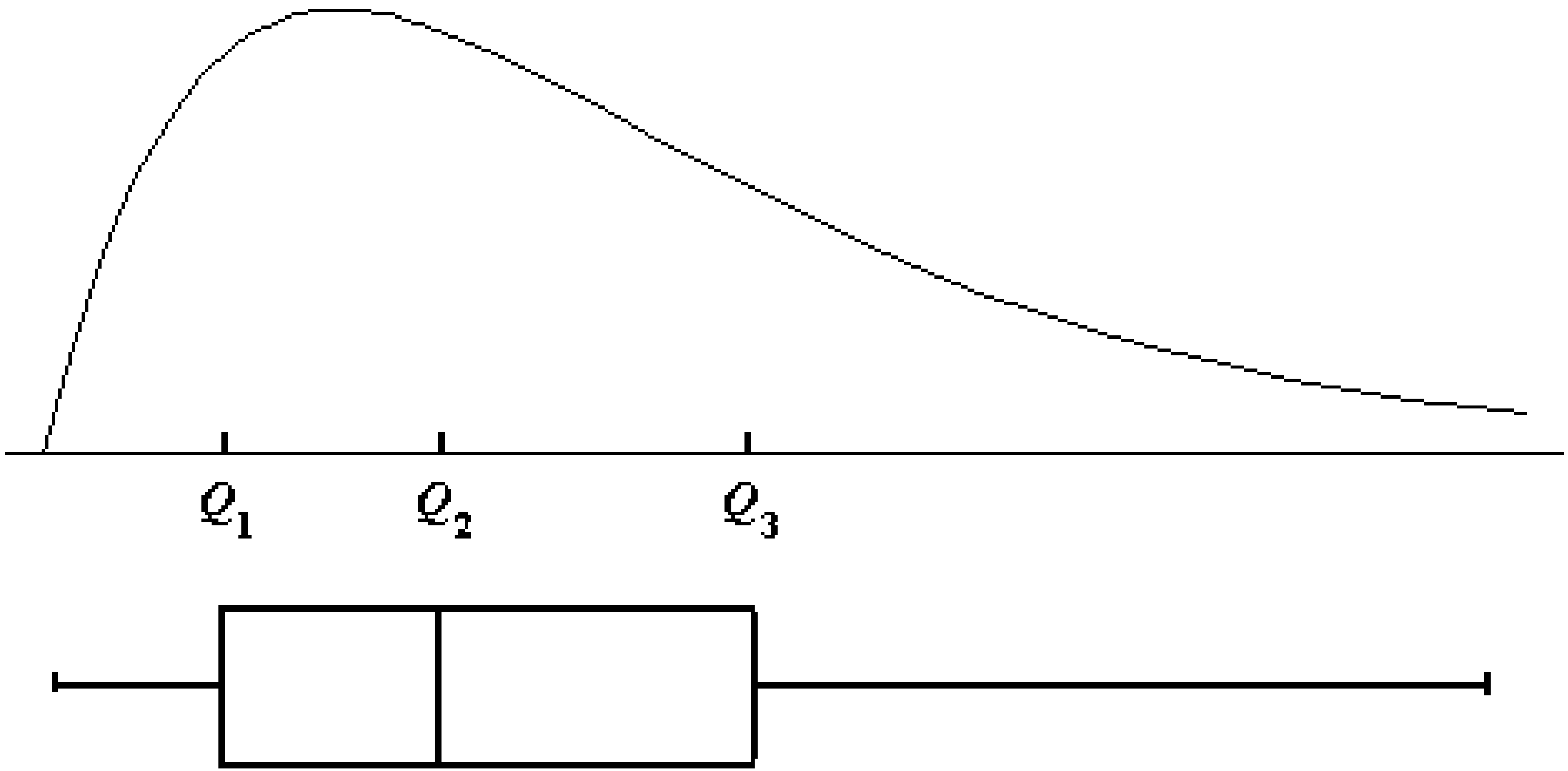


- The box is created using the quartiles
- The whiskers are created using the fences
- The points are the outlying points –if there are any

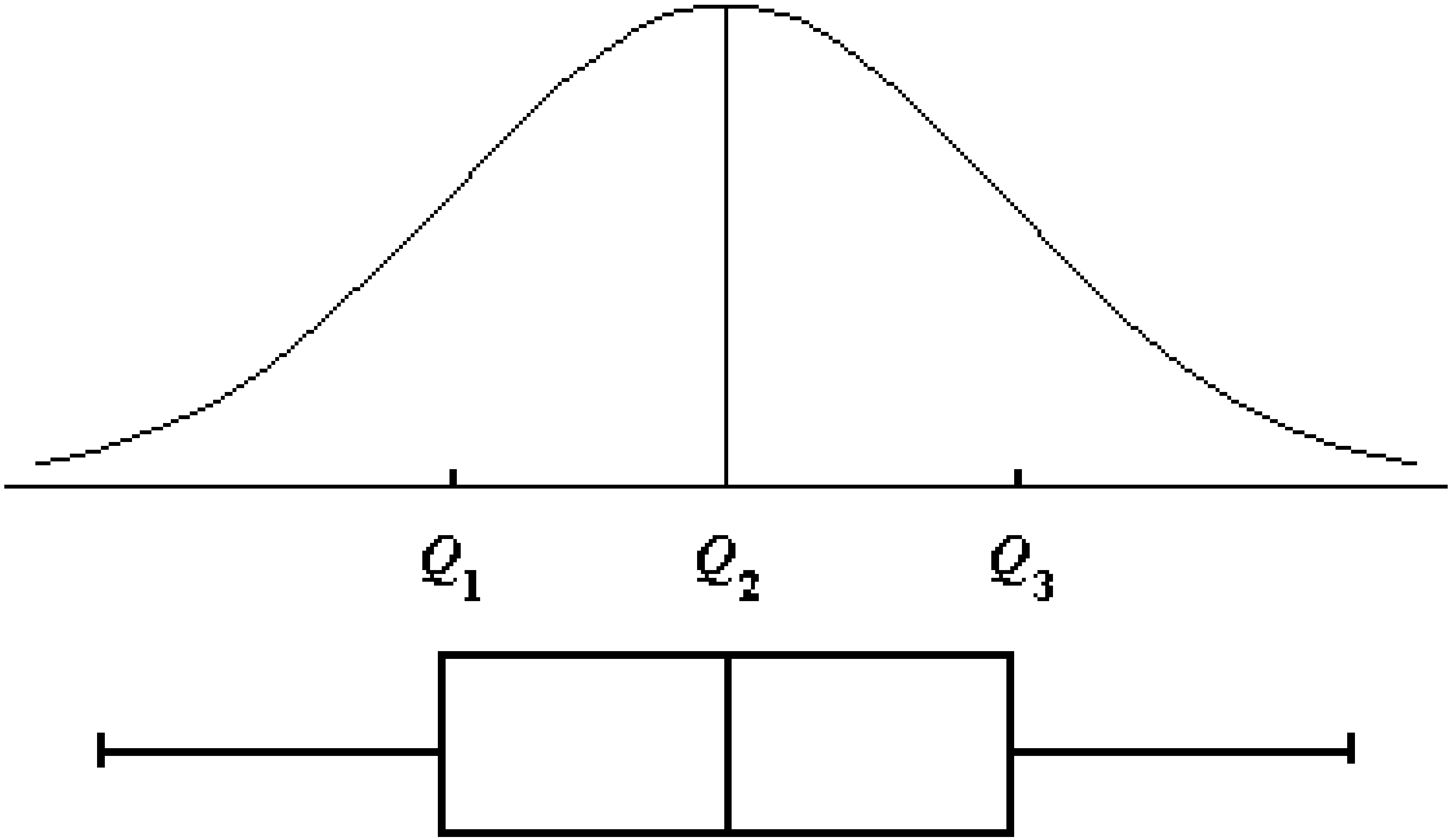
Skewness in Boxplots



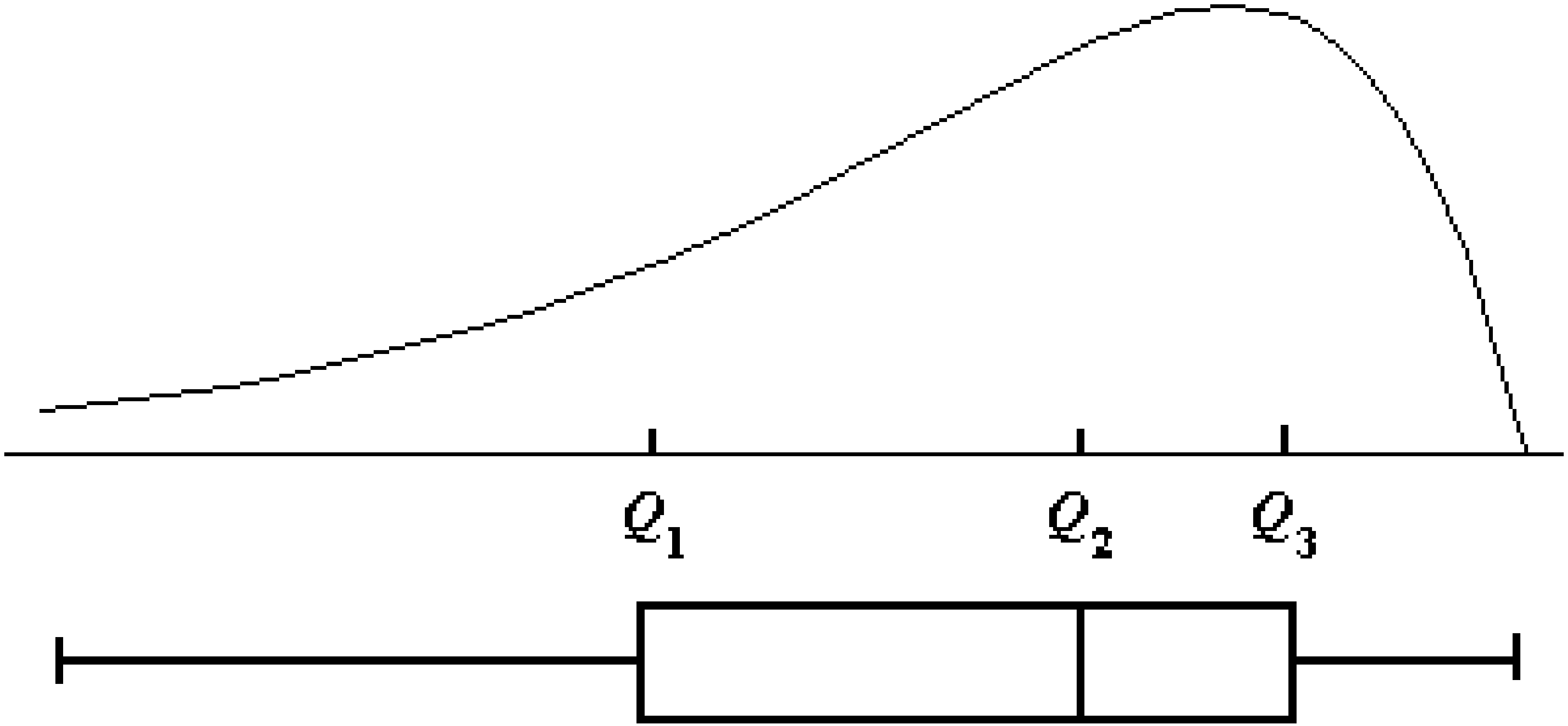
Right Skewed w/ Boxplots



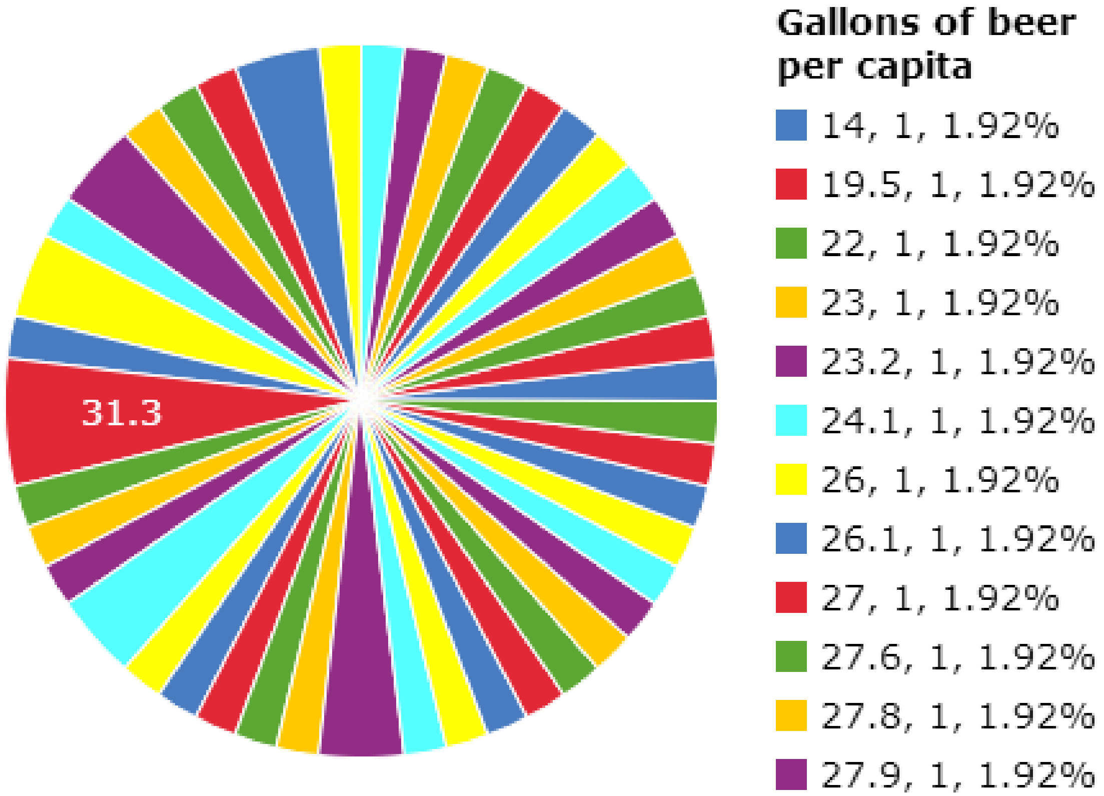
Bell Shaped w/ Boxplots



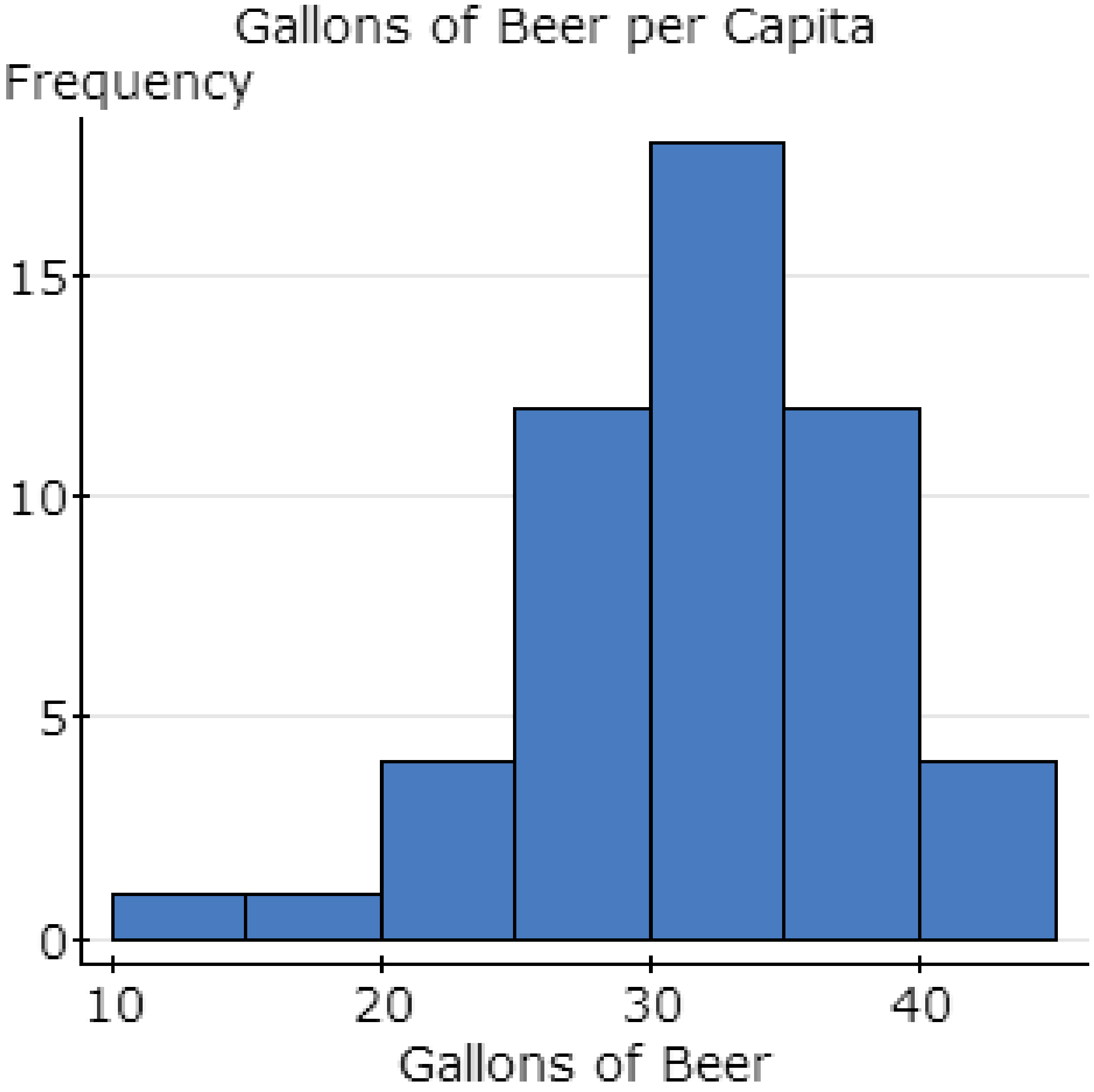
Left Skewed w/ Boxplots



Remember: **With graphs, if it's ugly it's probably not right**



Much Better!


Gallons of Beer per Capita
Frequency
15
10
5
0
10
20
30
40
Gallons of Beer

The Greek Letter Sigma in Math

- Before the Sigma was famous for representing organizations on campus it was used in mathematics
- This is a mathematical operator just like “+” .
- This weird looking E, uses for summation, tells you to add everything up

$$\Sigma$$

The Greek Letter Sigma in Math

- $\sum\{1,2,3,4,5,6,7,8,9\}$
- $= 1+2+3+4+5+6+7+8+9$

  $= 45$

- This is easy, you could have learned this in first grade – don’t make it harder than it actually is
- You can add, I have faith in you

Σ

Quantitative Summary: Mean

- **Mean (Average) –** The mean is the sum of observations divided by the number of observations
  - **Properties:** Sensitive to outliers

$$\bar{x} = \frac{\sum x}{n}$$

- X are the **variable** values for our sample
- n is the size of the sample

Quantitative Summary: Median

- **Median** – the median is the midpoint of the observations when they are ordered from the smallest to largest
  - Properties: Resistant to outliers
  - In position .5(n+1) when the data is in ascending order

Example: Median

X Value	1	1	2	3	4	5	5	5	5	6	10
Position	1	2	3	4	5	6	7	8	9	10	11

- Position = .5*(n+1) = .5(11+1) = 6^{th} position
- Median = 5

Example: Median

X Value	0.2	0.7	1.1	1.2	1.8	2.3	9.8	19.7
Position	1	2	3	4	5	6	7	8

- Position = .5*(n+1) = .5*(8+1) = 4.5th position
- Median = (1.2 + 1.8)/2 = 1.5

Quantitative Summary: Mode

- **Mode**– the mode is the observation that shows up the most in the data set.
  - Mode doesn’t necessary exist when we meet tie

Example: Mode

- X = {.2, .7, 1.1, 1.2, 1.8, 2.3, 9.8, 19.7}
  - There is no mode; all observations are tied with one occurrence

- X ={1, 1, 2, 3, 4, 5, 5, 5, 5, 6, 10}
  - Mode = 5 because 5 is the observation that occurred most.

Quantitative Summary: Range

- **Range –** The range is the difference between the maximum and minimum observations
  - **Properties:** easy to calculate but relies on only two values, which may be outliers

Range = Maximum - Minimum

Quantitative Summary: Variance

- **Variance –** the average, squared deviation of each observation from the mean
  - The idea is that it measures the spread of the data about the mean
  - **Properties:** difficult to interpret because it's in squared units, cannot be negative and is only zero when all data points are equal

Variance = $s^2 = \frac{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\mathbf{2}}}{\boldsymbol{n}-\mathbf{1}}$

Quantitative Summary: Standard Deviation

- **Standard Deviation** – the standard deviation is an adjusted average deviation of each observations' distance from the mean
  - The idea is that it measures the spread of the data about the mean
  - We prefer this to the variance because it isn't in squared units.
  - **Properties:** The larger the value the more spread or variability in the data, influenced by outliers and it's always positive.

Standard Deviation = $s = \sqrt{Variance} = \sqrt{\frac{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\mathbf{2}}}{\boldsymbol{n}-\mathbf{1}}}$

Let’s do an example!

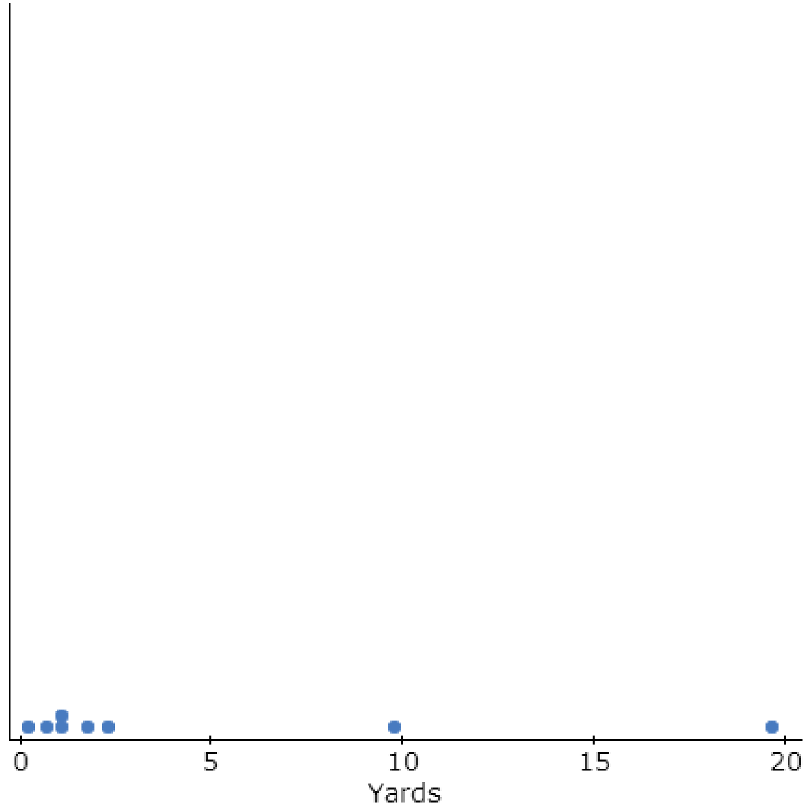
Quantitative Summary: Example

- X = distance (yards per carry for Marcus Lattimore) = {.2, .7, 1.1, 1.2, 1.8, 2.3, 9.8, 19.7}
- What kind of **data type** is this?
  - We know that distance or length is a **Continuous Quantitative** variable but we measure it discretely here by tenths of a yard
  - What type of graphs would be appropriate?
  - **Dot plot**, **Box plot**, **steam and leaf plot**, or a **histogram**

Quantitative Summary: Example

- Let’s try a **dot plot**!

- Our outlier is clear because it is highlighted and far away but the graph is awkward and hard to read



Quantitative Summary: Example

- Let’s try a **Stem and Leaf Plot**!

- Our outlier is clear because it is highlighted and far away but the graph is awkward and hard, or at least annoying to read

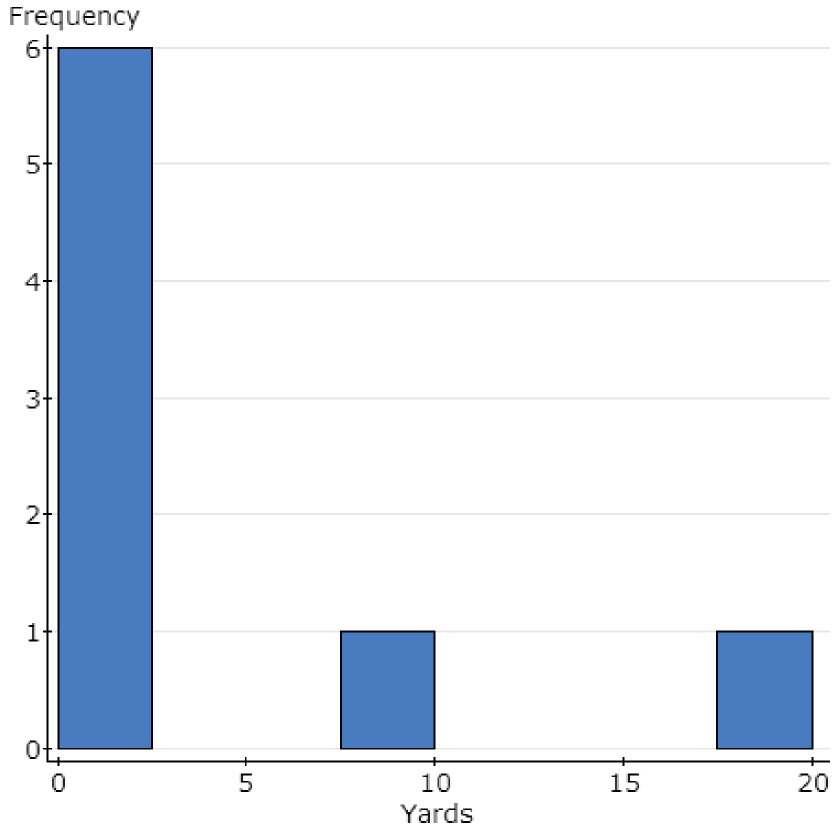
```
Decimal point is at the colon.
Leaf unit = 0.1

 0 : 27
 1 : 128
 2 : 3
 3 :
 4 :
 5 :
 6 :
 7 :
 8 :
 9 : 8
10 :
11 :
12 :
13 :
14 :
15 :
16 :
17 :
18 :
19 : 7
```

Quantitative Summary: Example

- Let's try a **histogram**

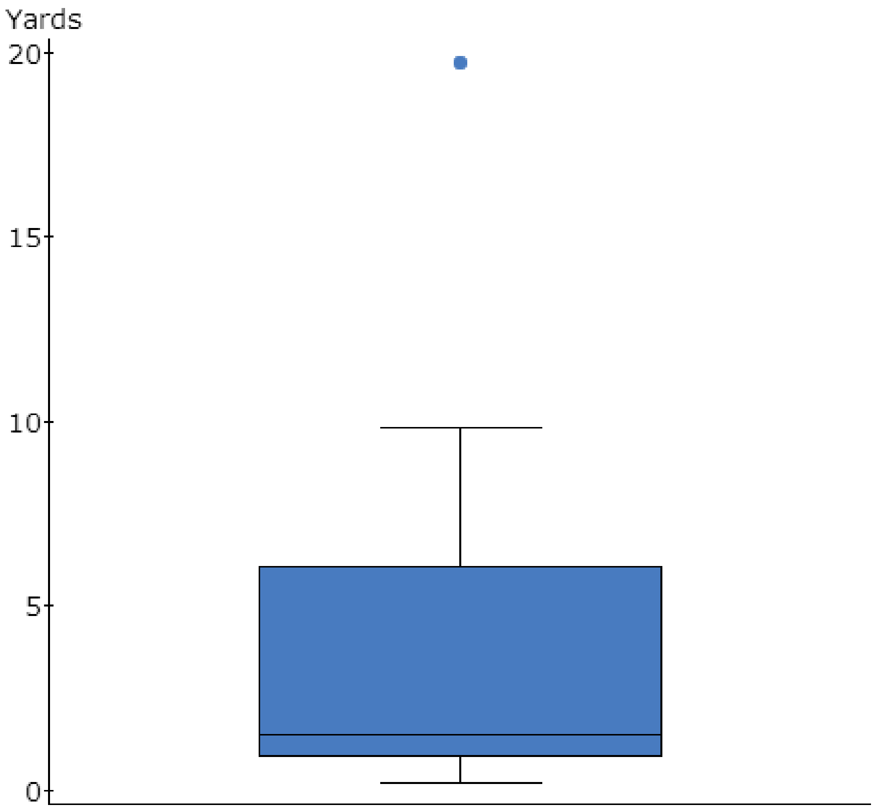
- This is better, but we still have some awkwardness with the gaps and the outlier isn't as obvious here



Quantitative Summary: Example

- Let's try a **box plot**

- This is really the best choice
  - Our outlier is clearly shown
  - The rest of the graph is readable and not as awkward



Back to the example!

Quantitative Summary: Example

- **Mean:** $\bar{x} = \frac{\sum x}{n}$

  = (.2 + .7 + 1.1 + 1.2 + 1.8 + 2.3 + 9.8 + 19.7) **/** 8

  = 4.6

- **Median:**
  - **Position** = .5(8+1) = 4.5th position

  = (1.2 + 1.8) / 2   We take the average of the two

  = 1.5

- **Mode:** there is no mode

Quantitative Summary: Example

After removing the outlier,

- **Mean:** $\bar{x} = \frac{\sum x}{n}$

  = **(**.2 + .7 + 1.1 + 1.2 + 1.8 + 2.3 + 9.8) **/** 7

  = 2.442857

- **Median:**
  - **Position =** .5(7+1) = 4

  = 1.2

Quantitative Summary: Example

- Before Removing Outlier: **Mean =** 4.6
  Median = 1.5
- After Removing Outlier:**Mean =** 2.442857
  Median = 1.2
- Notice that the mean changes much more than the median. Remember that the median is resistant to outliers and the mean is not.
- Notice the **mean > median** so it is **right skewed** in both cases!

Quantitative Summary: Example

- X = yards per carry for Marcus Lattimore = {.2, .7, 1.1, 1.2, 1.8, 2.3, 9.8, 19.7}

- **Range** = Maximum – Minimum

  = 19.7 - .2

  = 19.5

Quantitative Summary: Example

- X={.2, .7, 1.1, 1.2, 1.8, 2.3, 9.8, 19.7}
- **Variance =** $\frac{\sum(x-\bar{x})^2}{n-1} = \frac{326.56}{8-1} = 46.6514 \text{ yds}^2$

x	(x - mean)		(x - mean) ^2	
0.2	.2 - 4.6 =	-4.4	(-4.4)^2 =	19.36
0.7	.7 - 4.6 =	-3.9	(-3.9)^2 =	15.21
1.1	1.1 - 4.6 =	-3.5	(-3.5)^2 =	12.25
1.2	1.2 - 4.6 =	-3.4	(-3.4)^2 =	11.56
1.8	1.8 - 4.6 =	-2.8	(-2.8)^2 =	7.84
2.3	2.3 - 4.6 =	-2.3	(-2.3)^2 =	5.29
9.8	9.8 - 4.6 =	5.2	(5.2)^2 =	27.04
19.7	19.7 - 4.6 =	15.1	(15.1)^2 =	228.01
			Total	326.56

Quantitative Summary: Example

- X = {.2, .7, 1.1, 1.2, 1.8, 2.3, 9.8, 19.7}

- Standard Deviation =$\sqrt{Variance}$

$$= \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$
$$= \sqrt{46.6514}$$
$$= 6.8302 \text{ yds}$$

Let’s do a tricky example!

Quantitative Summary: A Tricky One

- Scores for Class A: 30, 65, 70, 76, 93, 99
- Scores for Class B: 68, 72, 73, 73, 74, 77

Class	n	Mean	Median
Class A	6	72.1667	73
Class B	6	72.8333	73

- Now, these are very similar. Would you say the students in each class performed the same?
  - Yes, the **mean** and **median** are almost identical

Quantitative Summary: A Tricky One

- A more complete summary will include a measure of spread

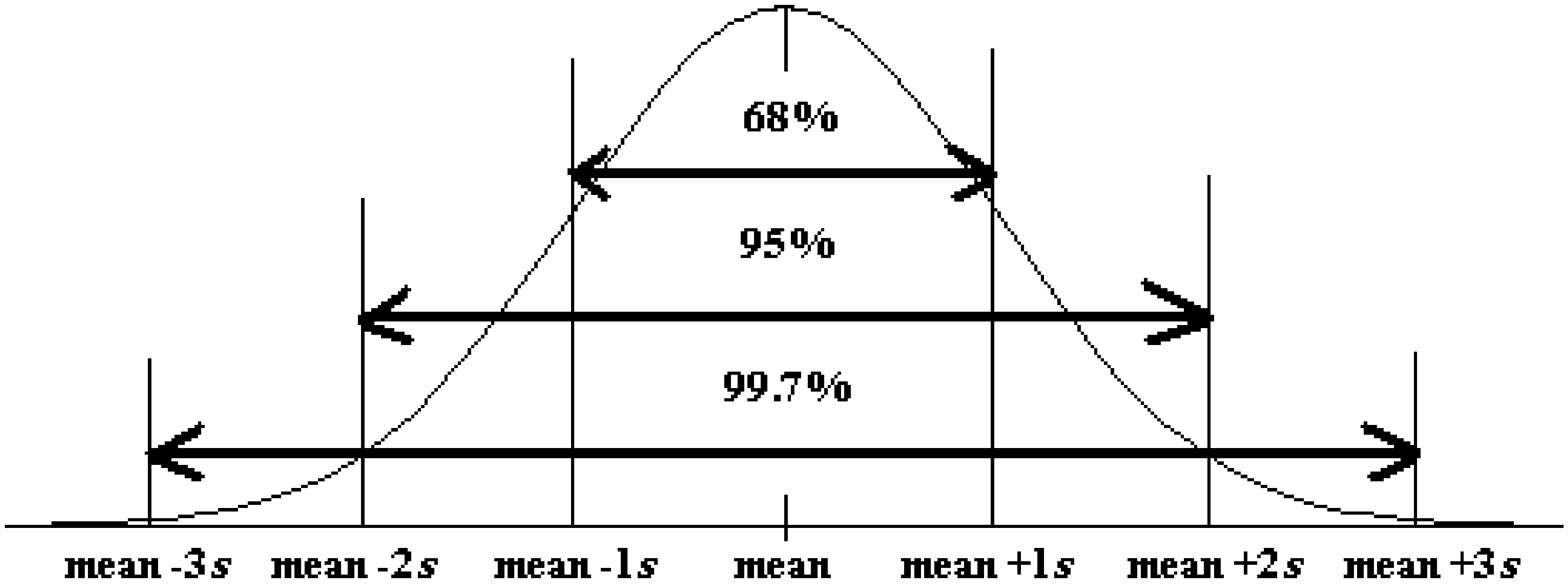
Class	n	Mean	Median	Variance	St. Dev
Class A	6	72.1667	73	600.5667	24.5065
Class B	6	72.8333	73	8.5667	2.9269

- Note, now we can say that although the mean and median were almost identical, the scores of Class A were more varied.

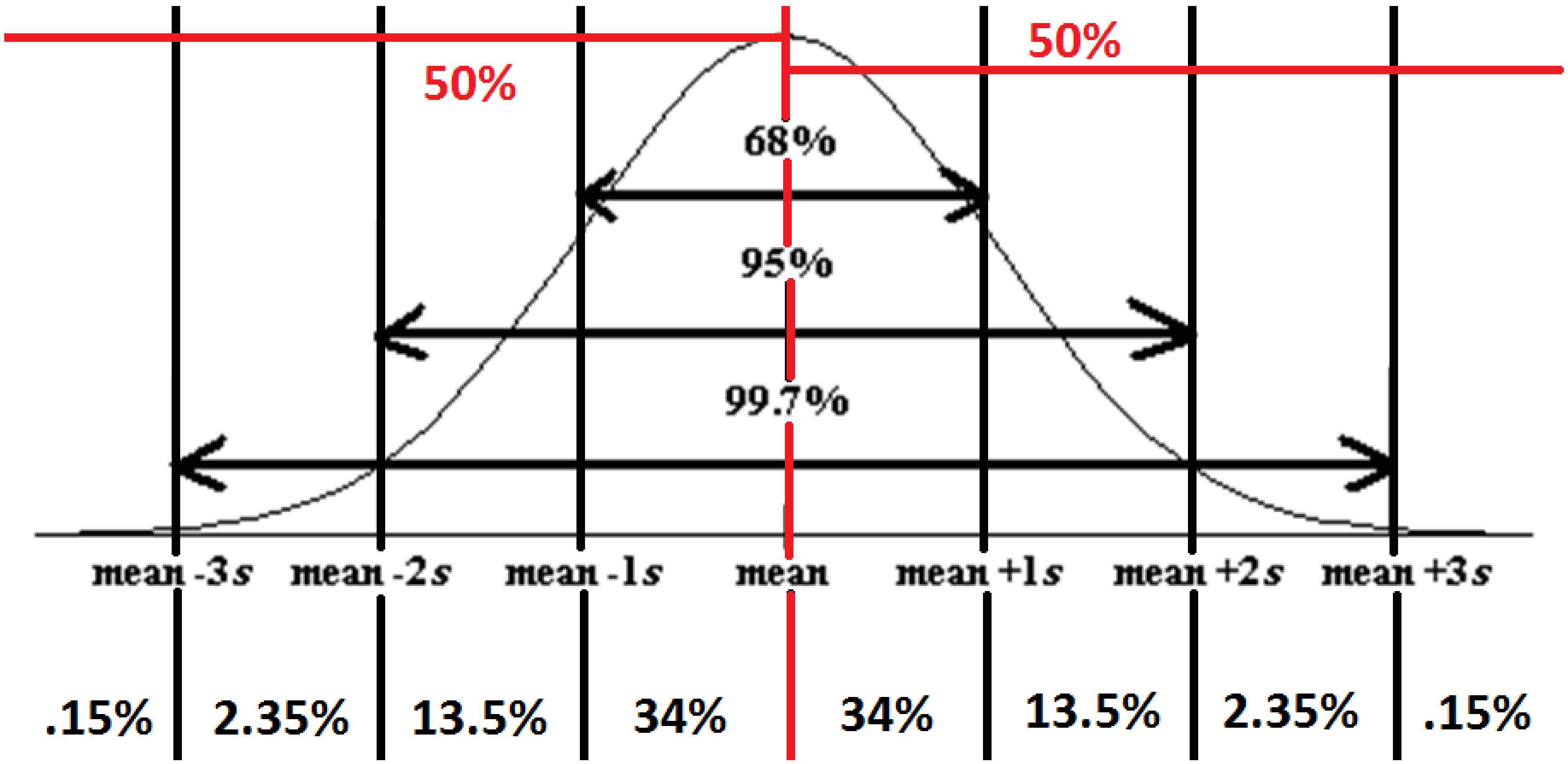
The Empirical Rule

- About 68% of data fall within 1 standard deviation of the mean
- About 95% of data fall within 2 standard deviation of the mean
- About 99.7% of data fall within 3 standard deviation of the mean
- **The distribution must be symmetric and bell shaped to use this Rule**

The Empirical Rule



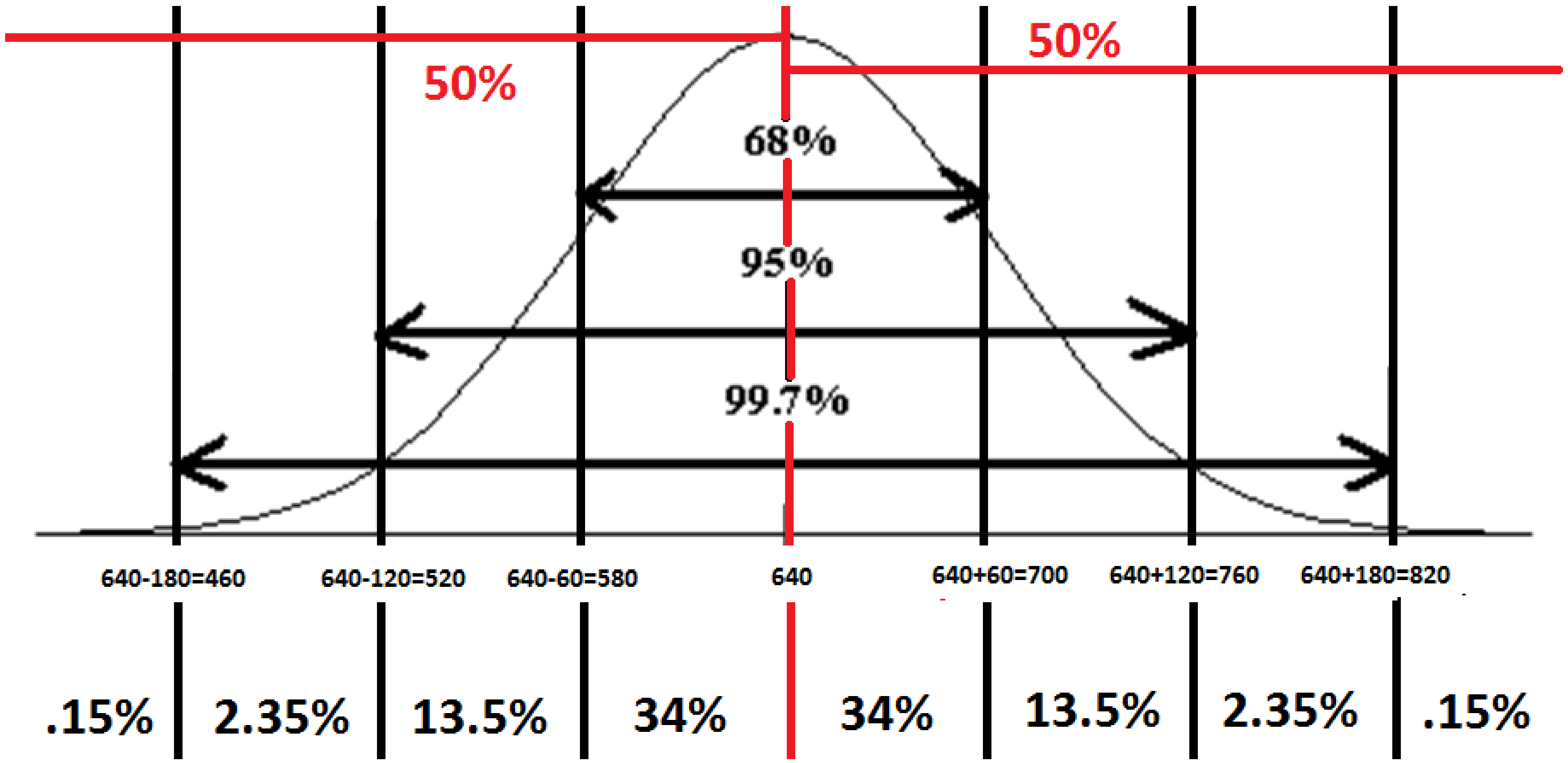
The Empirical Rule



The Empirical Rule: Example

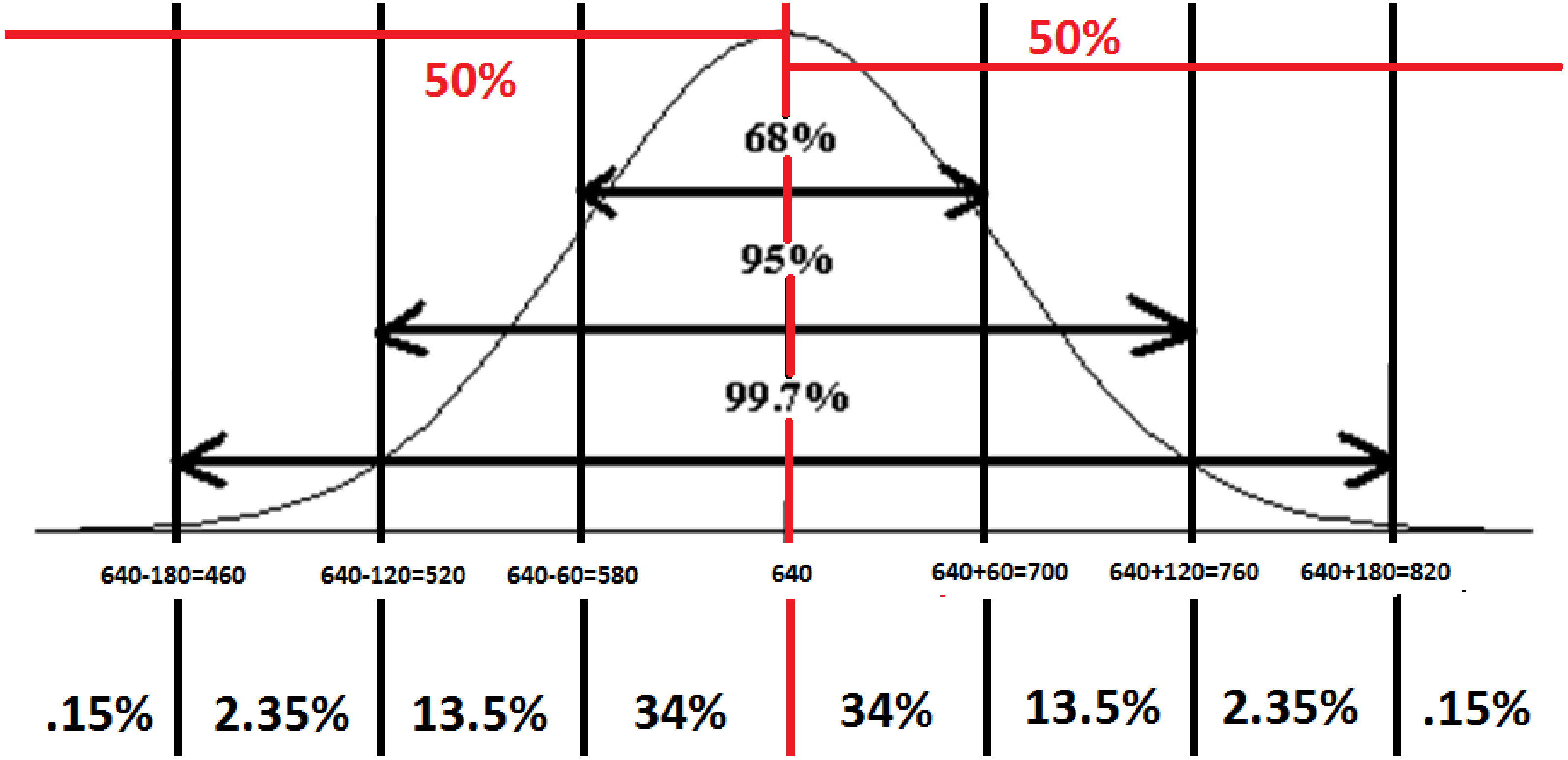
- The average college student consumes 640 cans of beer each year.
- Assume the distribution of cans of beers consumed per college student is **bell-shaped** with a **mean of 640 cans** and a **standard deviation of 60 cans**.

The Empirical Rule: Example



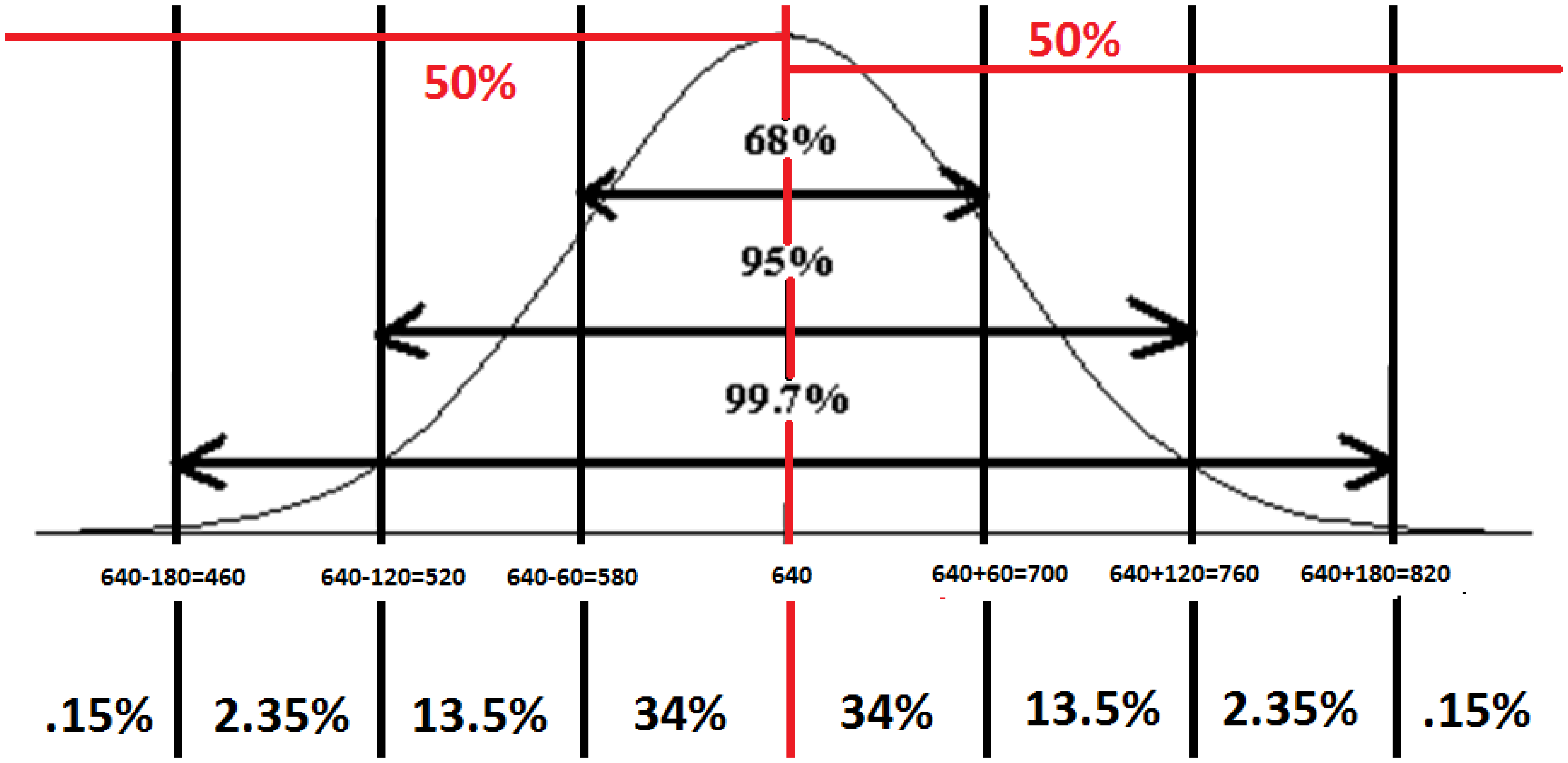
The Empirical Rule: Example

- About 68% of college students consume between ??? And ??? cans of beer per year



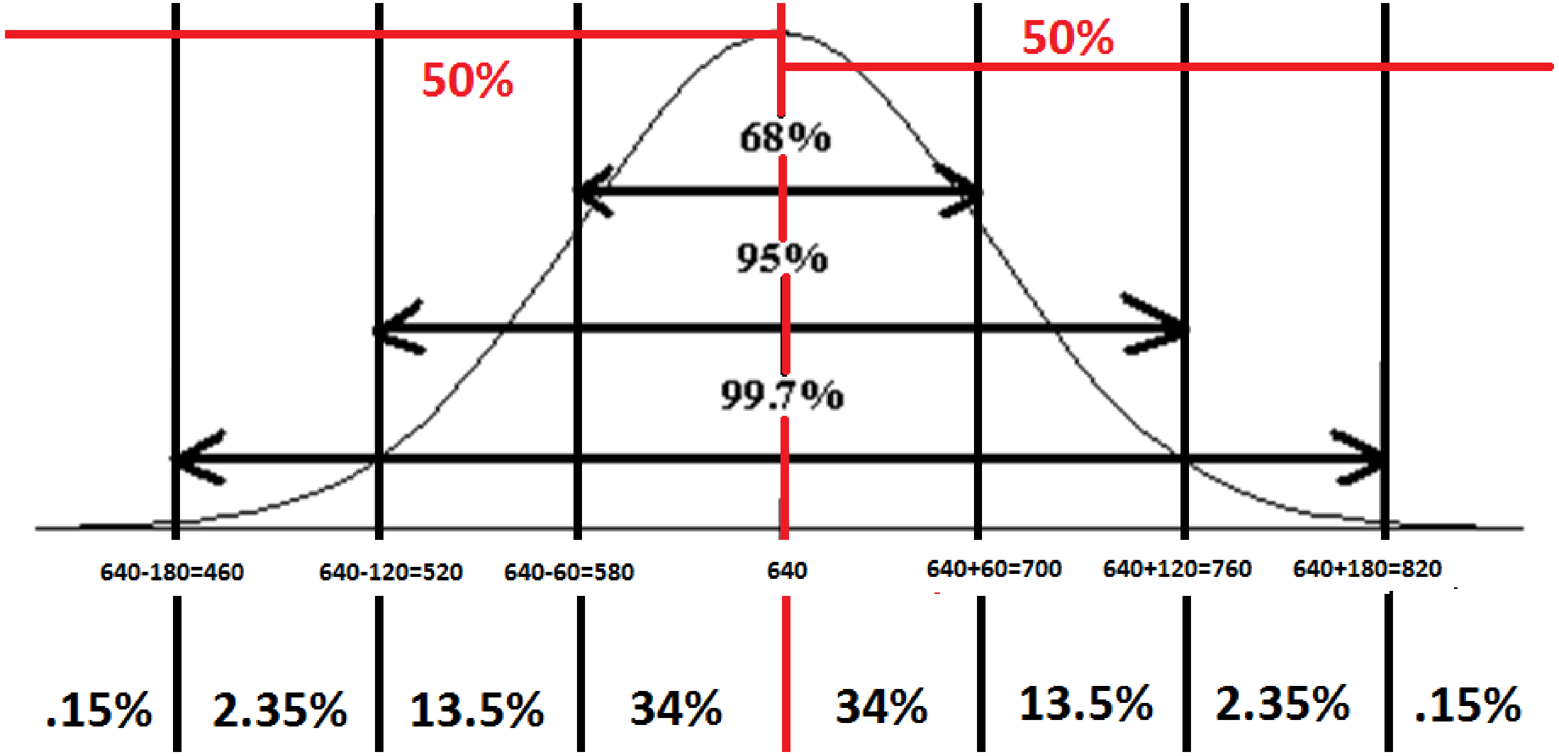
The Empirical Rule: Example

- About 68% of college students consume between 580 and 700 cans of beer per year



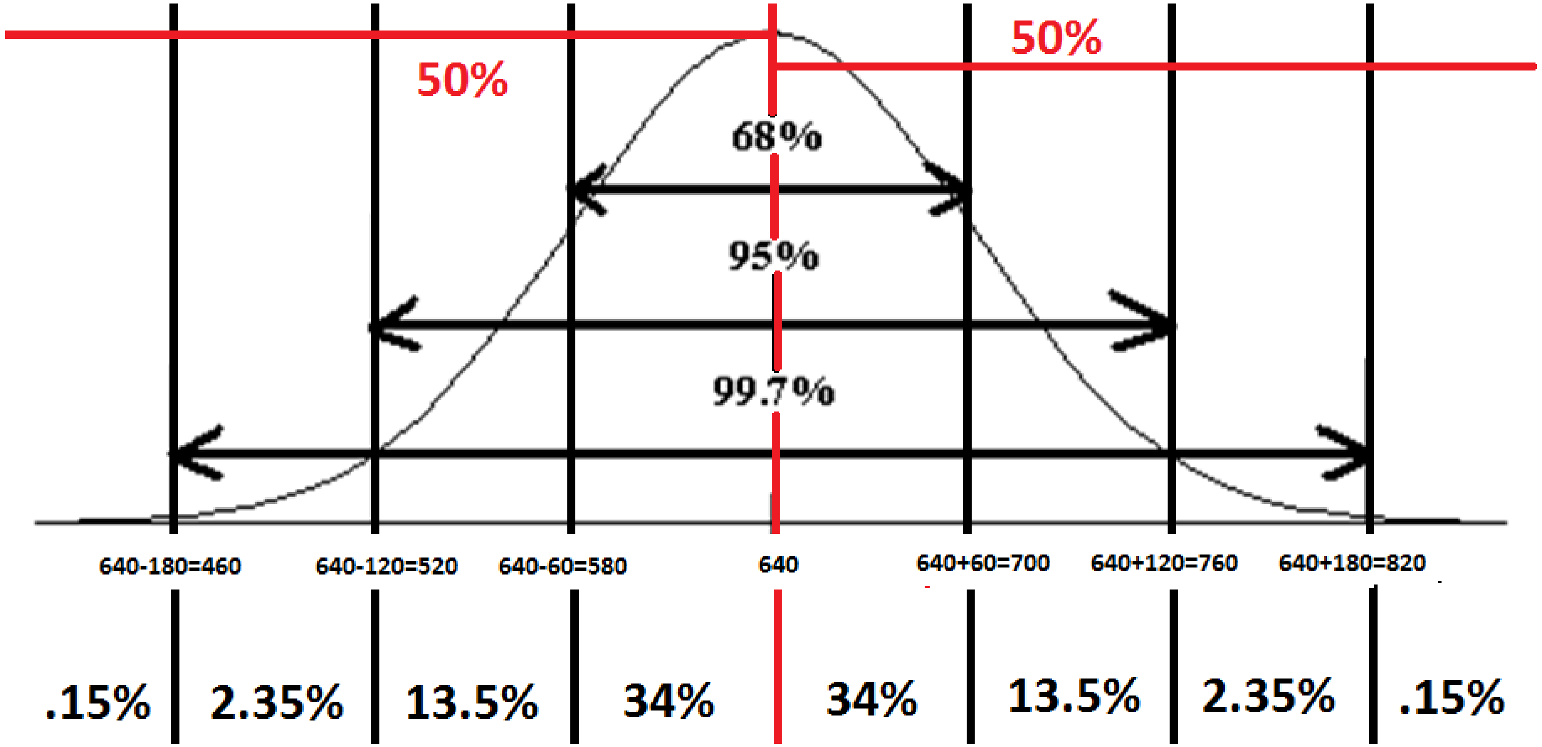
The Empirical Rule: Example

- About 95% of college students consume between ??? and ??? cans of beer per year



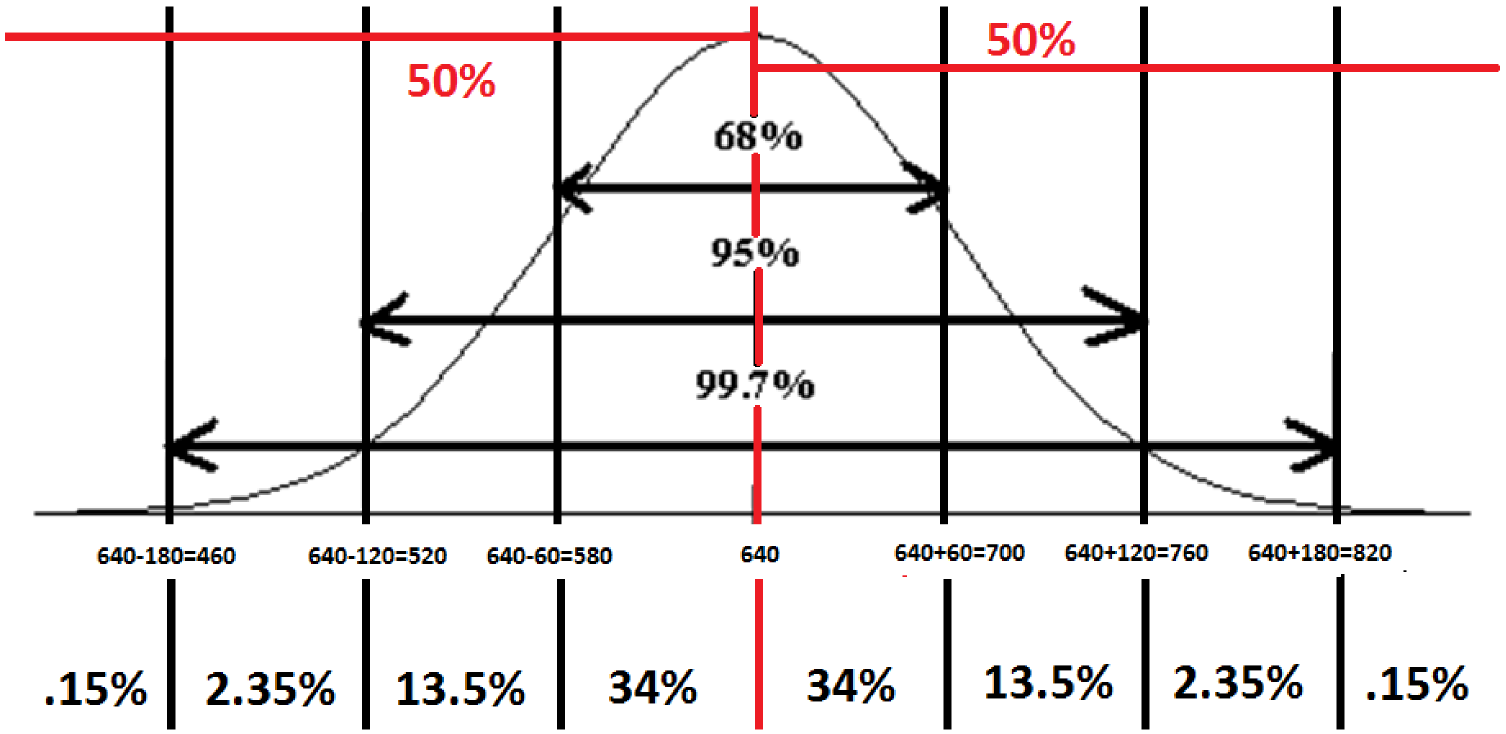
The Empirical Rule: Example

- About 95% of college students consume between 520 and 760 cans of beer per year



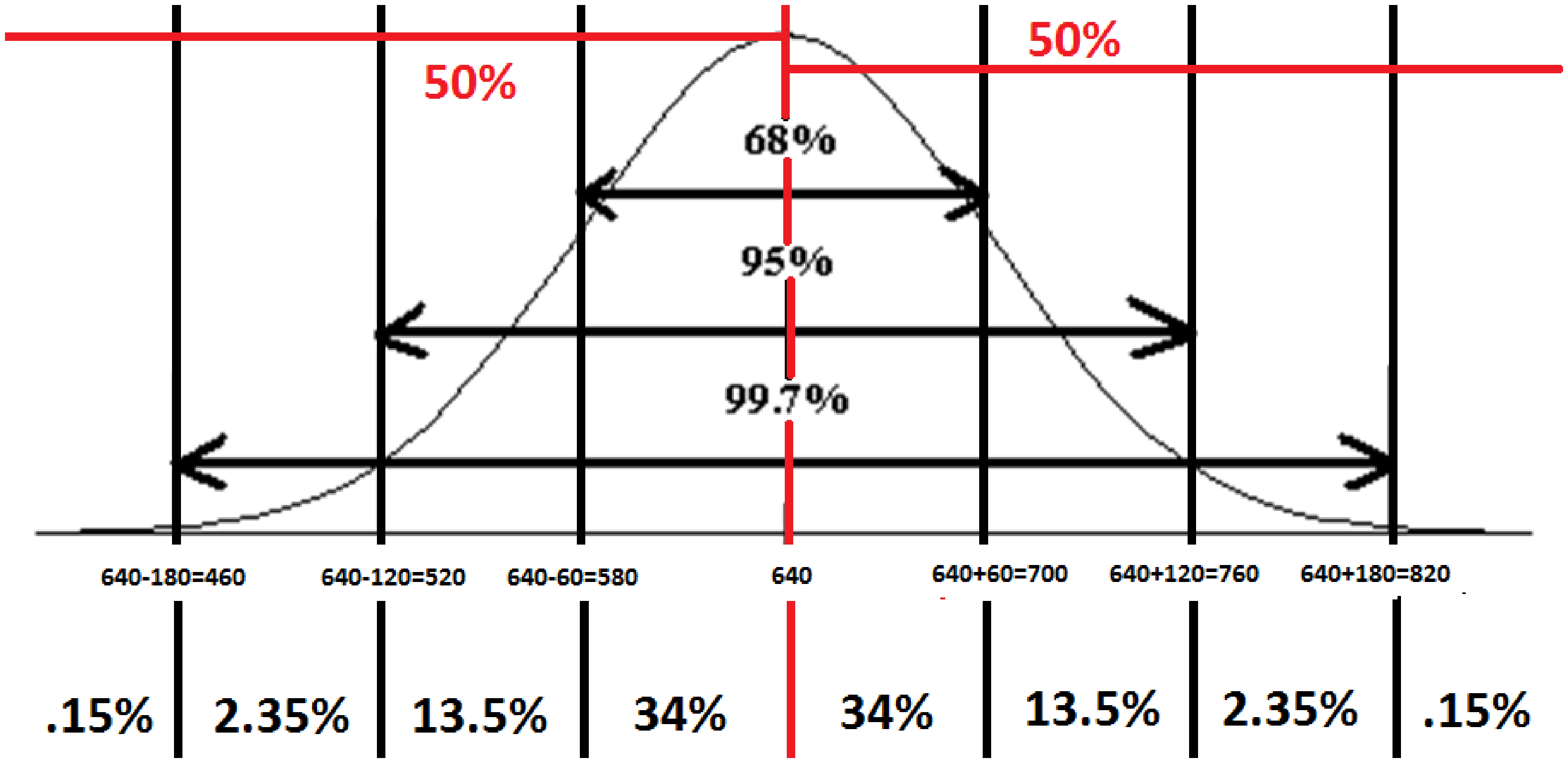
The Empirical Rule: Example

- About 99.7% of college students consume between ??? and ??? cans of beer per year



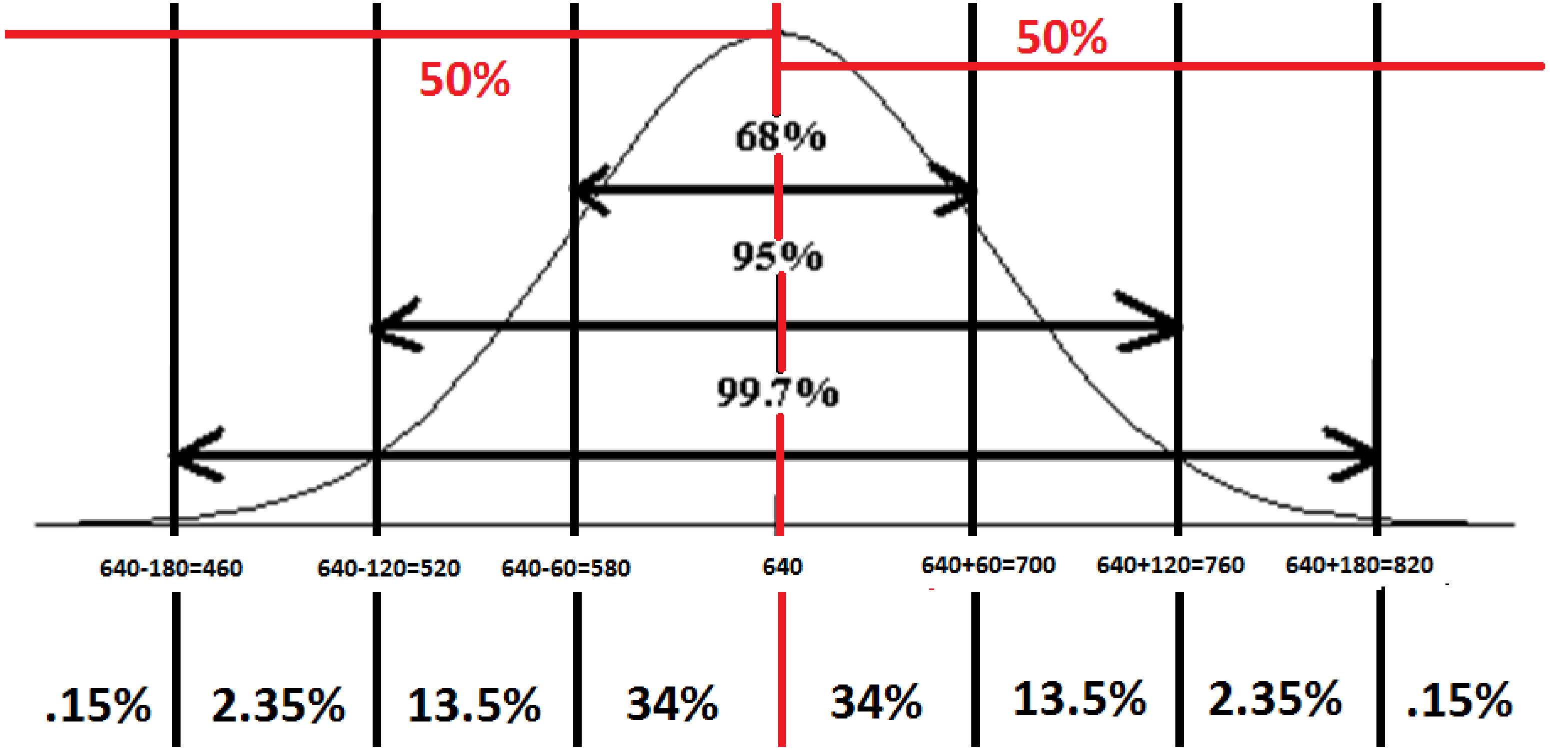
The Empirical Rule: Example

- About 99.7% of college students consume between 460 and 820 cans of beer per year



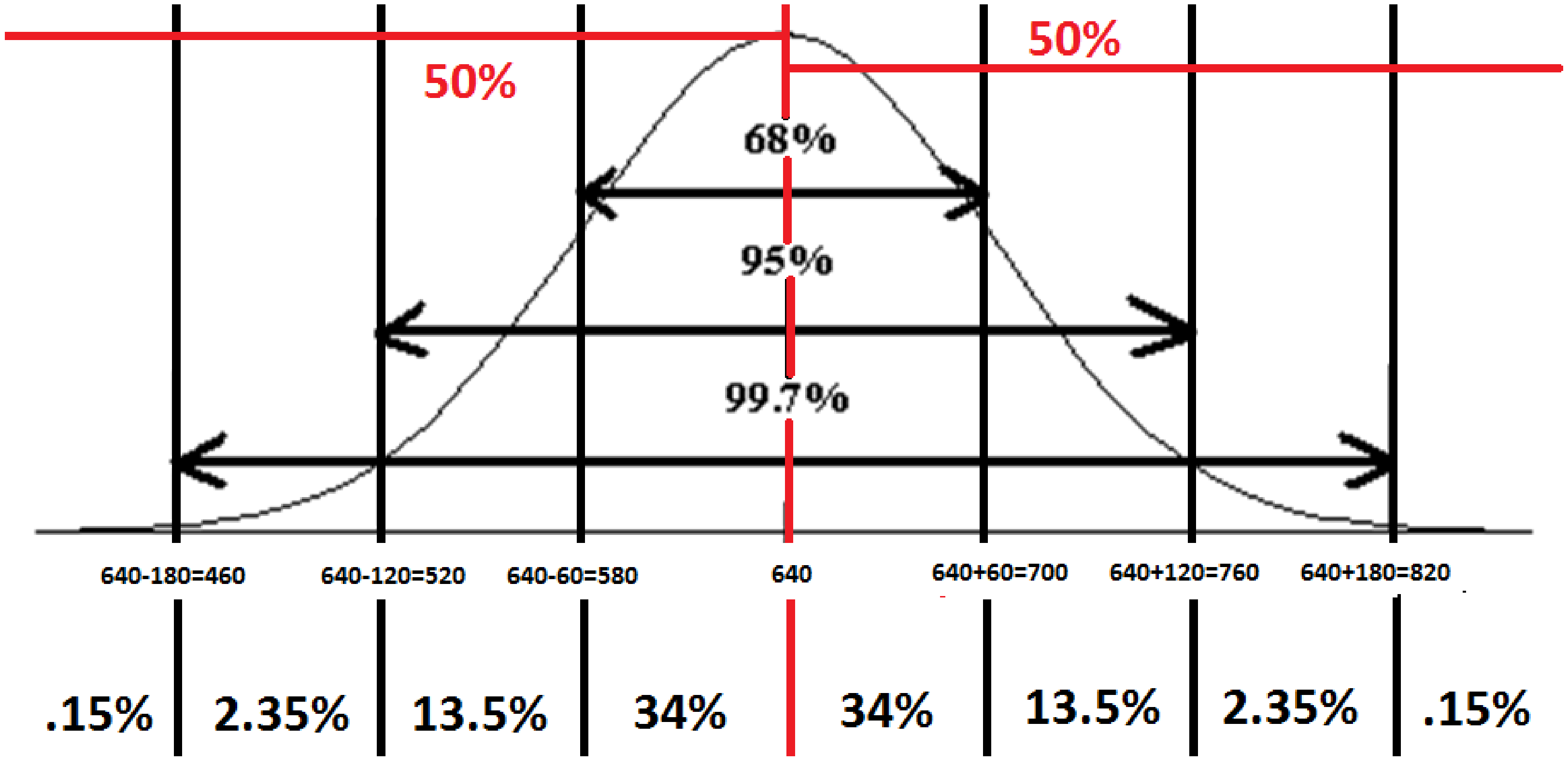
The Empirical Rule: Example

- About ??% of college students consume between 580 and 700 cans of beer per year



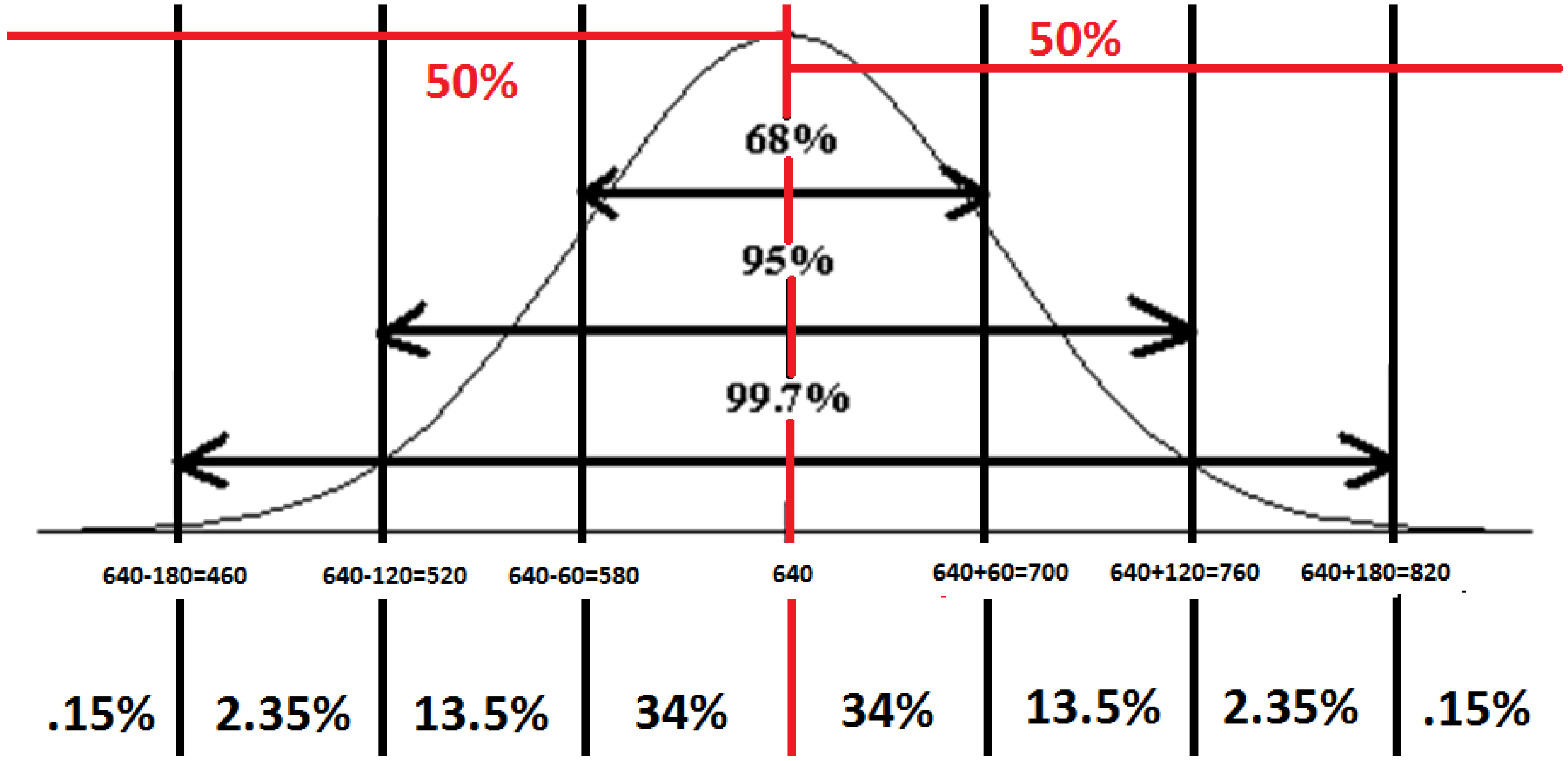
The Empirical Rule: Example

- About 68% of college students consume between 580 and 700 cans of beer per year



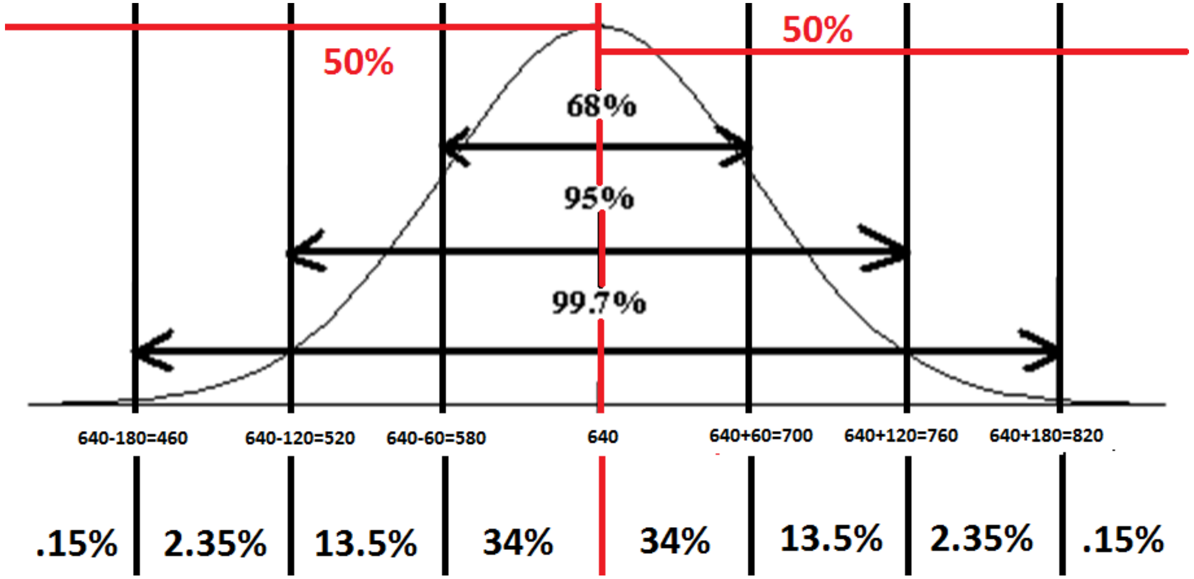
The Empirical Rule: Example

- What if I'm tricky and ask what percent of students consume less than 700 cans of beer per year?



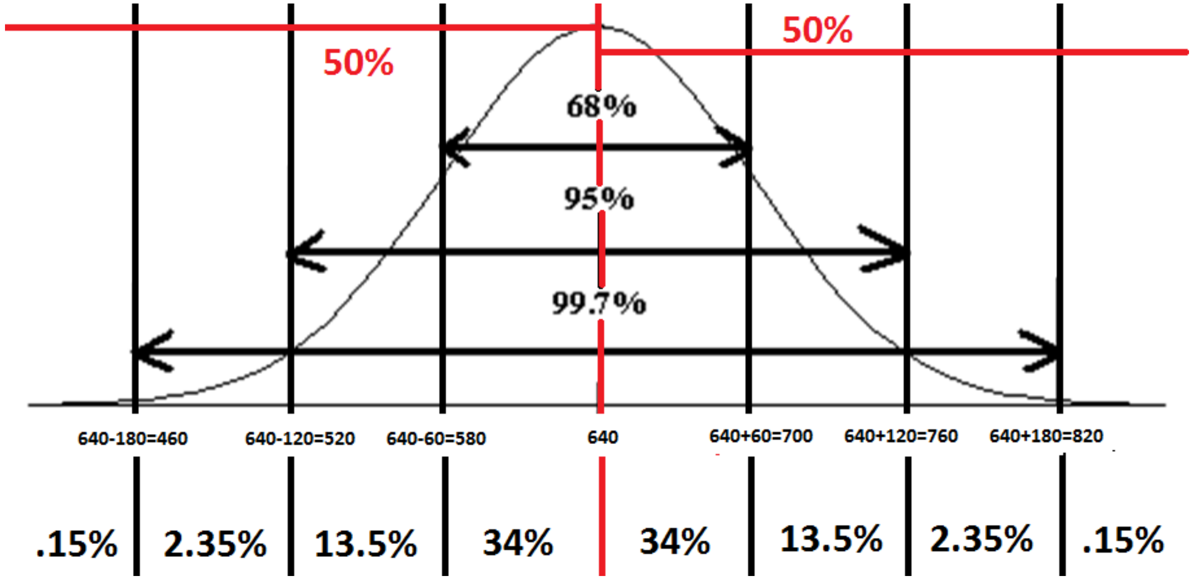
The Empirical Rule: Example

- What if I'm tricky and ask what percent of students consume less than 700 cans of beer per year?
- We can add up the area under the curve as we go left 50%+34%= 84%



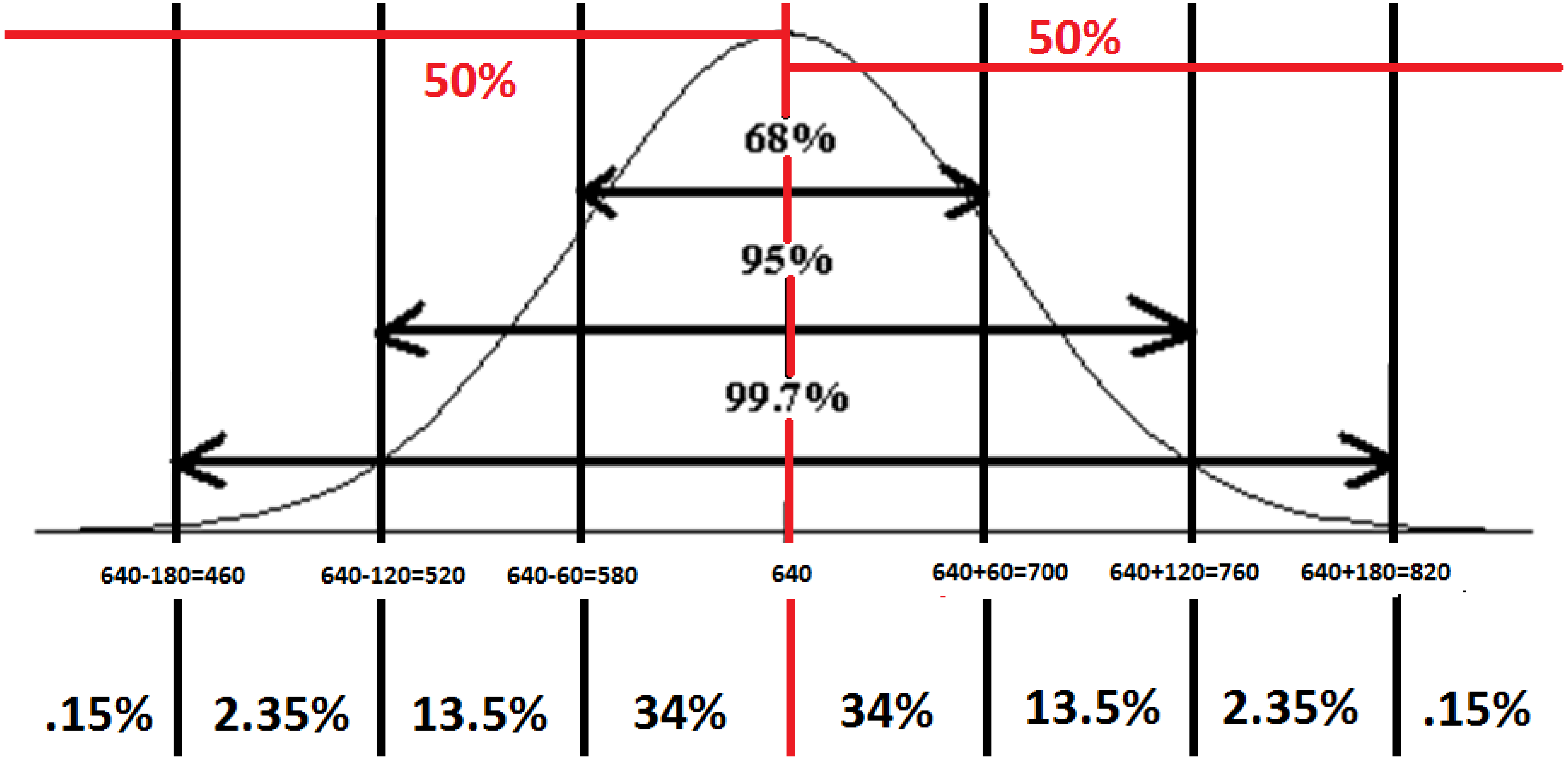
The Empirical Rule: Example

- What if I'm tricky and ask what percent of students consume less than 700 cans of beer per year?
- We can also subtract the area from 100% as we go right
  100%-13.5%-2.35%-.15%
  = 84%



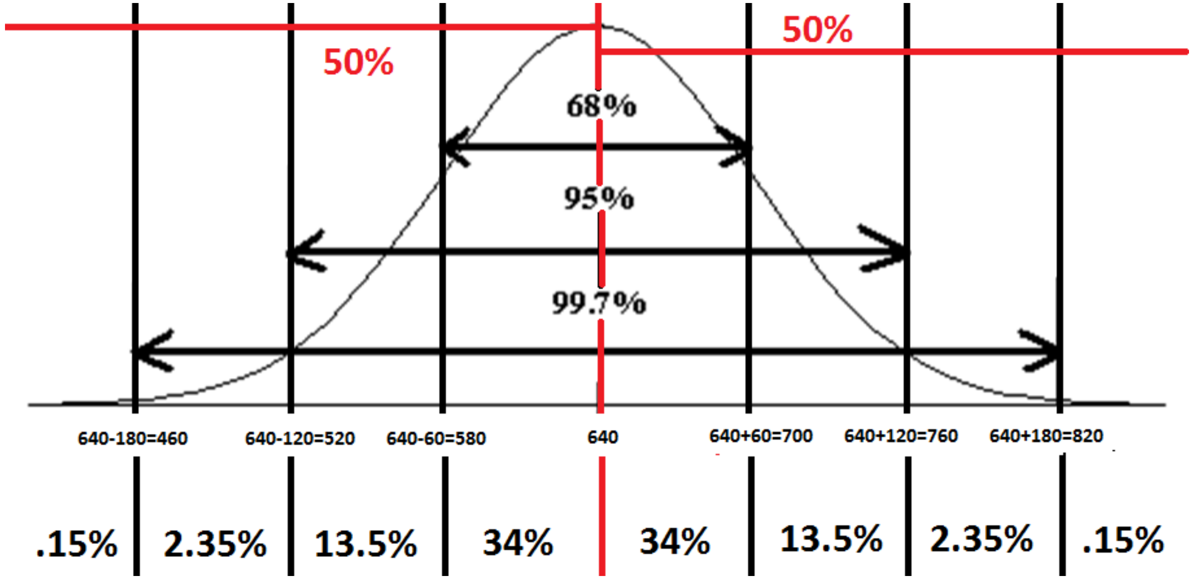
The Empirical Rule: Example

- What if I'm tricky and ask what percent of students consume more than 700 cans of beer per year?



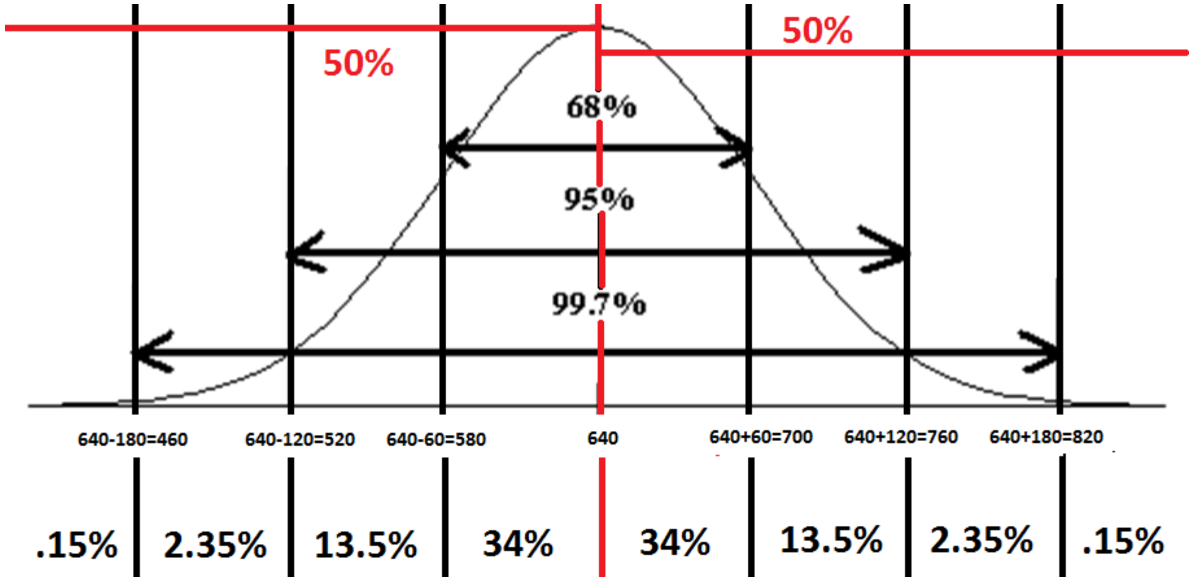
The Empirical Rule: Example

- What if I'm tricky and ask what percent of students consume more than 700 cans of beer per year?
- We can add up the area under the curve as we go right
  13.5% + 2.35% + .15% = 16%



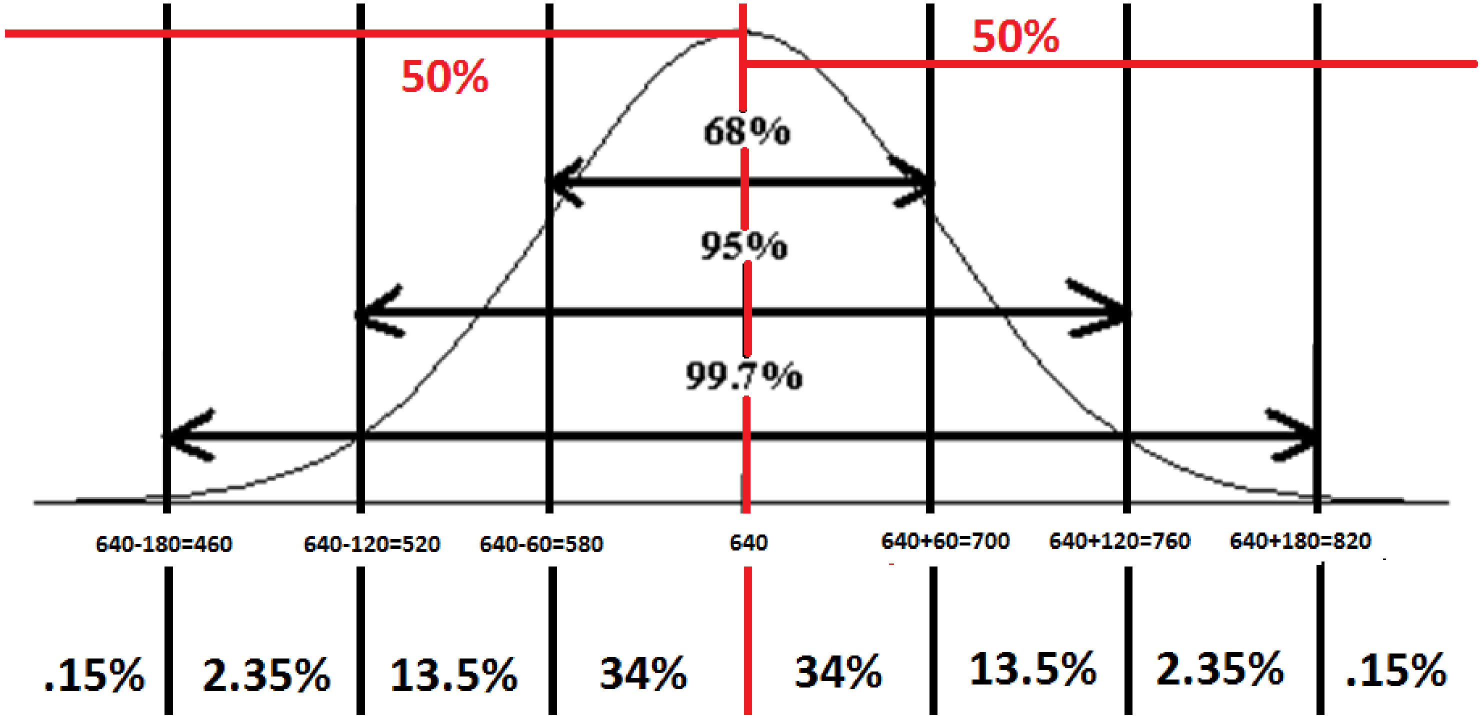
The Empirical Rule: Example

- What if I'm tricky and ask what percent of students consume more than 700 cans of beer per year?
- We can also subtract the area from 100% as we go left
  100%-84% (we know 84% from the last question)
  = 16%



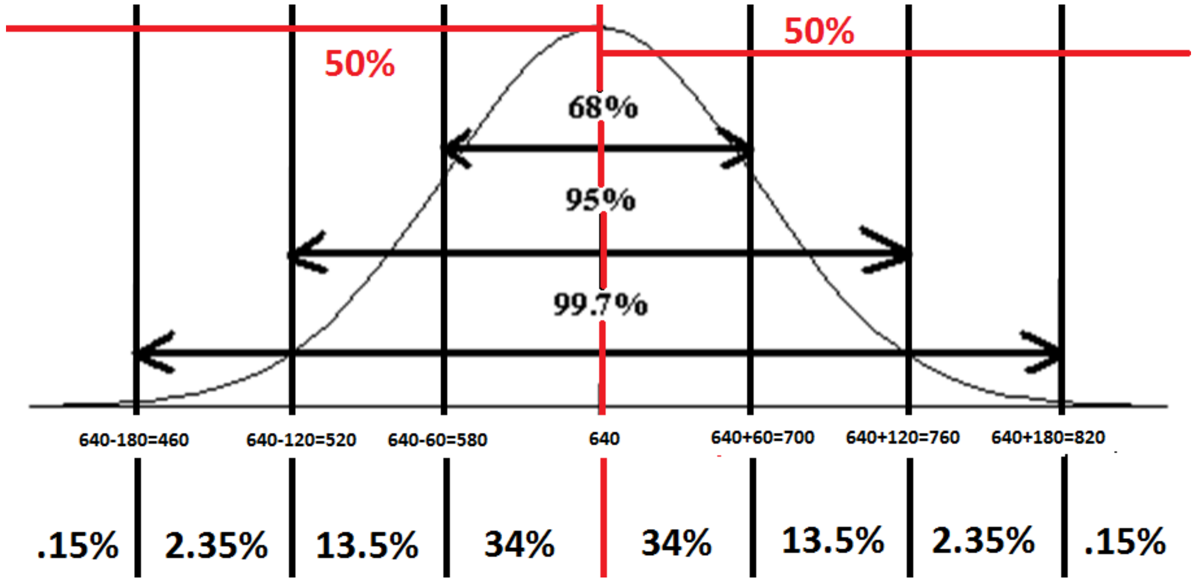
The Empirical Rule: Example

- What if I'm tricky and ask what percent of students consume between 460 and 700 cans of beer per year?



The Empirical Rule: Example

- What if I’m tricky and ask what percent of students consume between 460 and 700 pounds of beer each year?
- We can add up the area under the curve as we go from 460 to 700
  2.35%+13.5%+34%+34% = 83.85%



Z Score

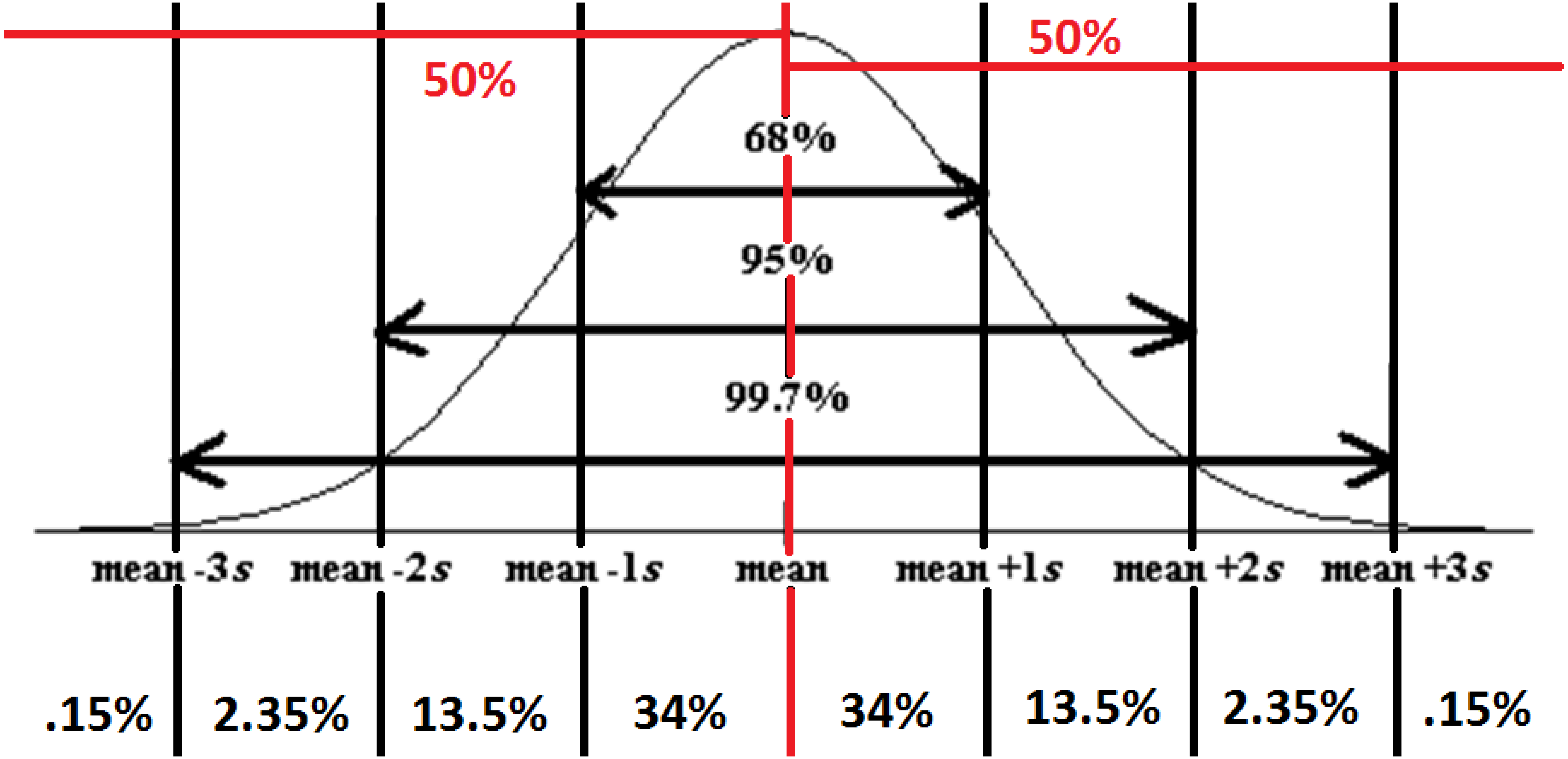
- We have learnt **Mean**, **Median**, and **Mode** to describe the center
- We have learnt **Range**, **Variance**, and **Standard Deviation** to describe the variability
- **Z Score** is what we use to describe the position

Z Score: What are We Doing Here?

- What did we do with the empirical rule?
  - We looked at how many standard deviations away the data values were

- The idea here is to be able to find out how many standard deviations the data values we're looking at are from the mean but we allow fractional answers

The Empirical Rule



Z Score: How Do We Calculate It?

- $z = \frac{observation - mean}{standard\ deviation}$

- This gives us the number of standard deviations from the mean the observation is, and the direction

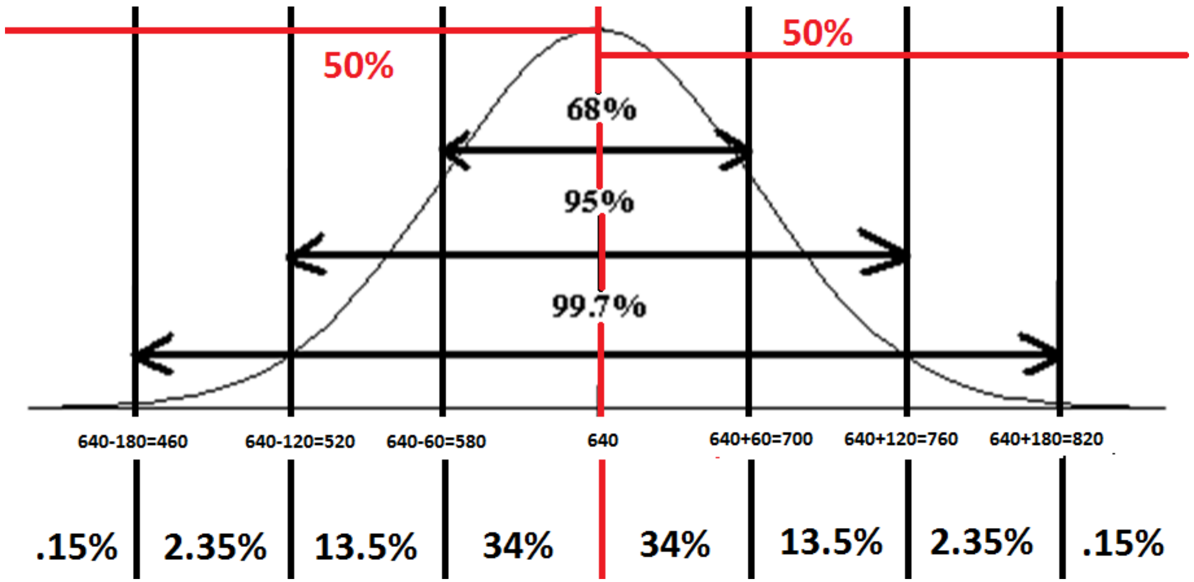
- **Note: we consider any observation with a Z score above 3 or below -3 an outlier**

Z Score: Example

- The average college student consumes 640 cans of beer per year. Assume the distribution of beers consumed per year per college student is **bell-shaped** with a **mean of 640 cans** and a **standard deviation of 60 cans**.

Z Score: Example

- Recall from the Empirical Rule that about 99.7% of college students consume between 460 and 820 cans of beer per year (+,- 3 standard deviations)



Z Score: Example

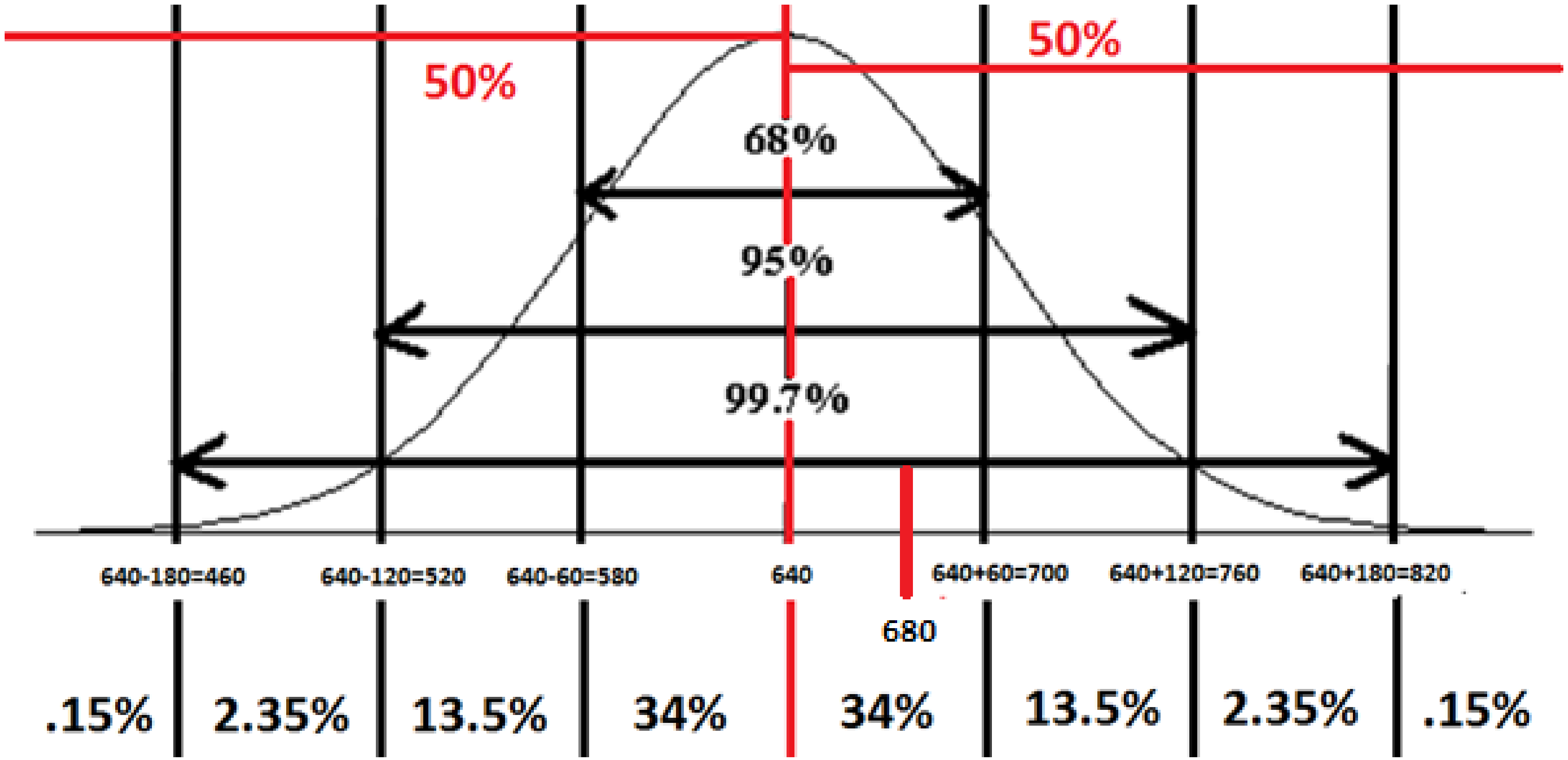
- $z_{460} = \frac{460 - 640}{60} = \frac{-180}{60} = -3$
- $z_{820} = \frac{820 - 640}{60} = \frac{180}{60} = 3$

- Note the Z score has given us the correct number of standard deviations from the mean for each case!

Z Score: Example

- Let’s consider an observation of 680 cans of beer.
  - 680 is not 1, 2, or 3 standard deviations away
  - $z = \frac{680-640}{60} = .6667$
    - X=680 is .6667 standard deviations above the mean
    - .6667 indicates this observation is not an outlier because .6667<3 and .6667>-3

Z Score: Example



Z Score: Example

- Let's consider an observation of 1080 cans of beer.
  - 1080 is not 1, 2, or 3 standard deviations away
  - $z = \frac{1080-640}{60} = 7.3333$
    - X=1080 is 7.3333 standard deviations above the mean
    - .67 indicates this observation is an outlier because 7.3333>3

Z Score: Example

- Let’s consider an observation of 500 cans of beer.
  - 500 is not 1, 2, or 3 standard deviations away
  - $z = \frac{500-640}{60} = -2.3333$
    - X=500 is 2.3333 standard deviations below the mean
    - -2.3333 indicates this observation isn’t an outlier because -2.3333<3 and -2.3333>-3

Percentiles

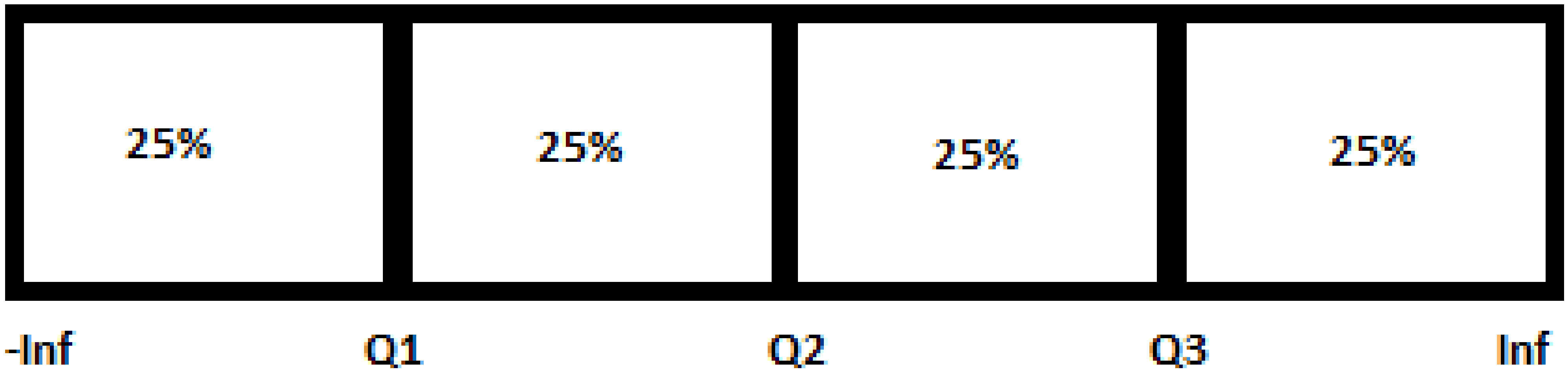
- **Percentile**: the p-th percentile is a value such that p percentage of the observations fall below or at the value.
- Consider an ordered population of 10 data values {3,6,7,8,8,10,13,15,16,20}
- What are the 70^{th} and 15^{th} percentile?
- 70^{th} percentile = $(0.7 * 10)^{th}$ position = 7^{th} position = 13
- 15^{th} percentile = $(0.15 * 10)^{th}$ position = 1.5^{th} position < 2^{nd} position = 6

Percentile and Quartile

- **Quartiles** because they split the data into quarters
- Q1: the observation at the 25th percentile
- Q2: the observation at the 50th percentile (Median)
- Q3: the observation at the 75th percentile

- **IQR (Interquartile range)**=Q3-Q1: another measure of spread used in place of standard deviation

Percentile and Quartile



Five Number Summary

- The five number summary includes:

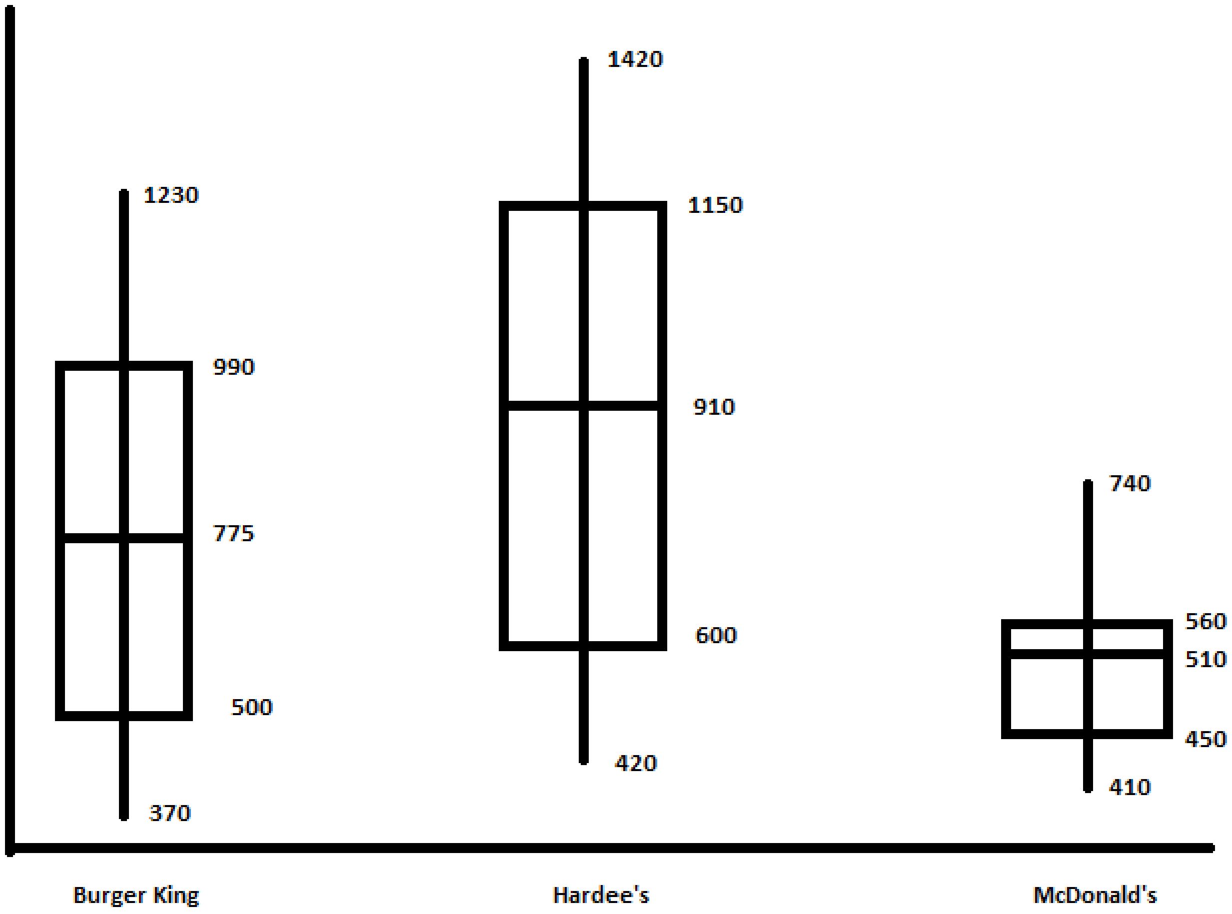
Maximum
3rd Quartile
2nd Quartile (Median)
1th Quartile
Minimum

Five Number Summary: Example

- Let’s consider this summarized data of calories per item

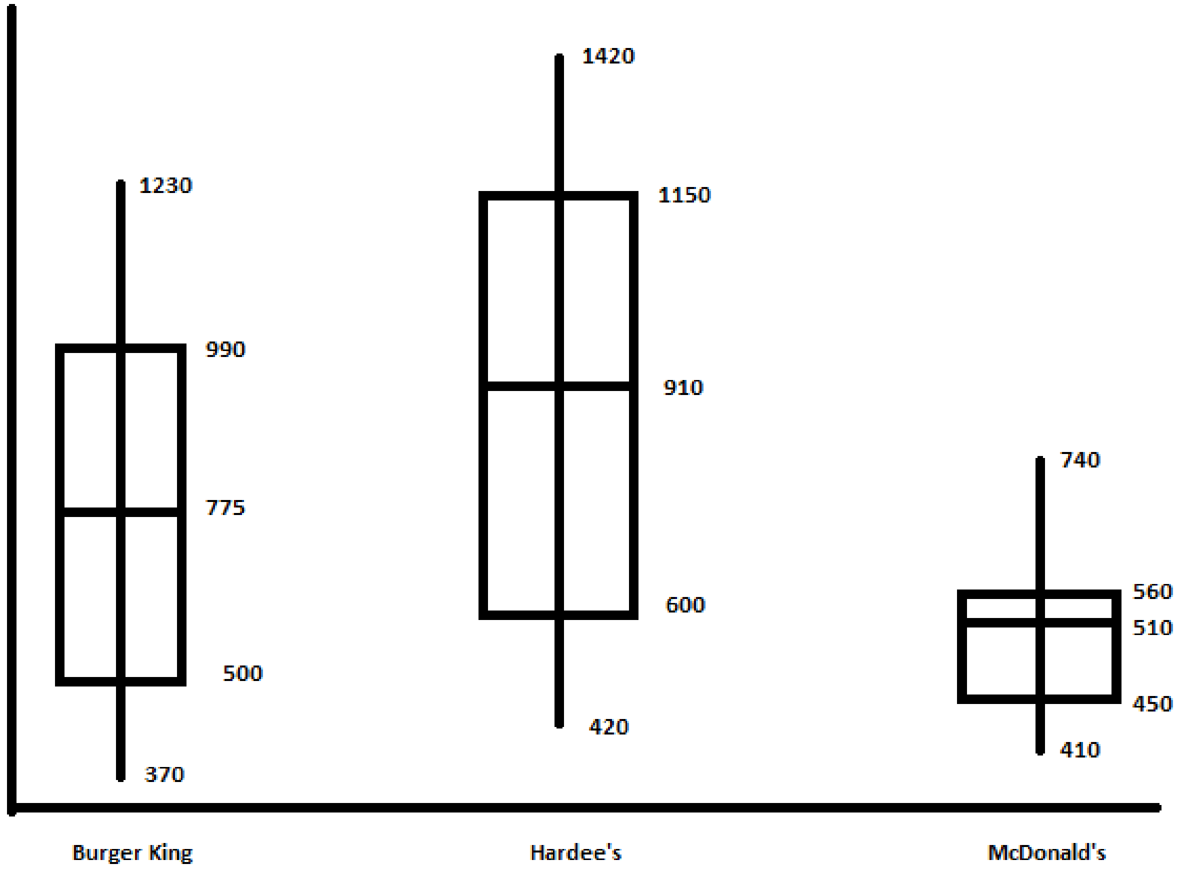
Restaurant	Min	Max	Q1	Q2	Q3	IQR
Burger King	370	1230	500	775	990	990-500=490
Hardee’s	420	1420	600	910	1150	1150-600=550
McDonalds	410	740	450	510	560	560-450=110

Five Number Summary: Example



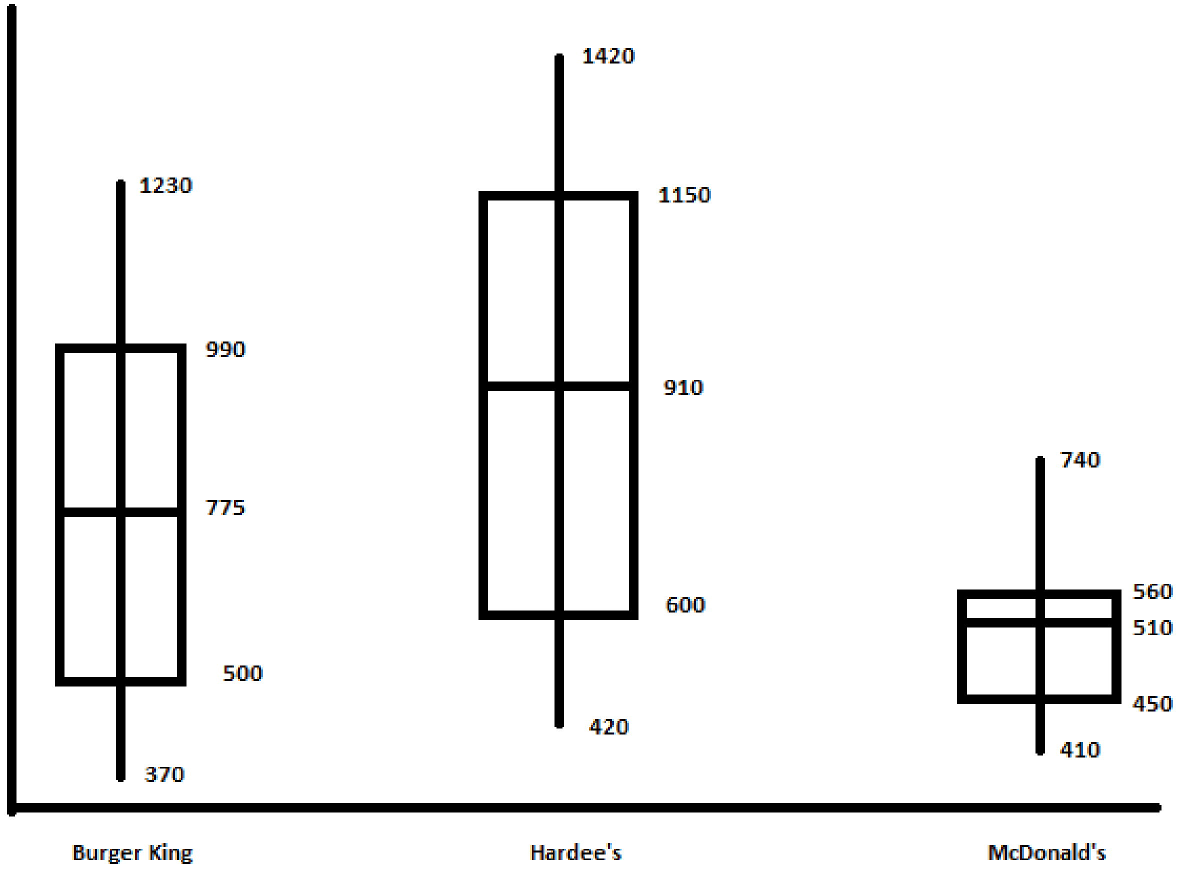
Five Number Summary: Example

- What can we say to compare these three restaurants?



Five Number Summary: Example

- Hardee's has the highest calorie item
- BK has the lowest calorie item
- McDonald's has the least spread (range)
- Hardee's has the most spread



Measures of Central Tendency

Measure	Computation	Interpretation	When to Use
Mean Statistic: $\bar{x}$ Parameter: μ	$\bar{x} = \frac{\sum x}{n}$	Center	Use for quantitative data when the distribution is roughly symmetric
Median	The point halfway through the data when it is arranged in ascending order.	The point which splits the data in half.	Use for quantitative data when the distribution is skewed
Mode	We report the observation with the highest frequency	Most frequent observation	When the most frequent observation is the desired measure or when data is qualitative.

Measure of Dispersion

Measure	Computation	Interpretation
Range	Maximum – Minimum	The difference between the largest and smallest data point
Standard Deviation Statistic: s Parameter: σ	$\sqrt{Variance}$	The square root of the mean of squared deviations from the mean in the original units – this usually makes the standard deviation easier to interpret
Variance Statistic: s^2 Parameter: σ^2	$\frac{\sum(x-\overline{x})^2}{n-1}$	The square root of the mean of squared deviations from the mean in units squared

Graphical Displays

Variable Type	Graphical Display	Numerical Summary
Categorical	Pie chart or bar graph	Frequency table
Quantitative	Histogram or box plot – can also try dotplot or stem & leaf	Quantitative Summary
1-Categorical and 1-Quantitative	Side by Side boxplots	Quantitative Summary for groups
2-Categorical	Side by side pie charts or bar graphs best: stacked bar chart	Contingency Table or side by side frequency tables
2-Quantitative	Scatter plot	Side by side Quantitative Summaries