Statistical Thinking

- What is statistics, anyway?

Statistics

- The science of collecting, organizing, summarizing, and analyzing data to answer questions and/or draw conclusions.

Why statistics?

- To satisfy our curiosity
  - Exploring the world around us.
  - Searching for patterns to lead to discoveries
- To make sure that we can stand on our legs
  - Evidence to show that we are right (or wrong)

Statistics Rests on Two Major Concepts

- Variation
  - Differences or changes in an item

Statistics Rests on Two Major Concepts

- Data
  - Observations gathered to draw conclusions
  - Context matters

Context matters –Always Ask:

- Who: Describe the individuals who were surveyed.
- What: Determine what is being measured.
- When: When was the research conducted?
- Where: Where was the research conducted?
- Why: What was the purpose of the survey or experiment?
- How: Describe how the survey or experiment was conducted.

Statistics: Data Types

Statistics: Data Types

How do you define data?

- Information or a set of values collected from surveys, experiments, observations, etc.

Statistics: Data Types

- In statistics, we classify data into four categories:
  - Nominal Data
  - Ordinal
  - Interval
  - Ratio

Statistics: Data Types

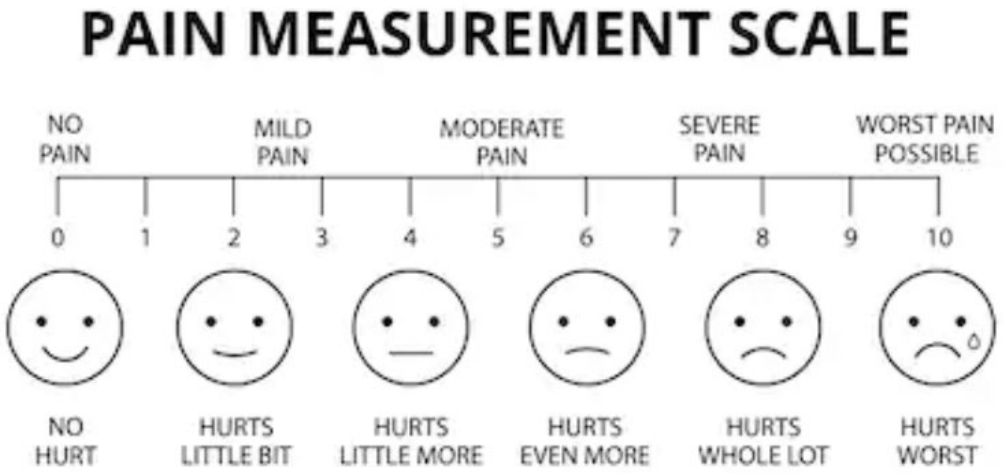
- Nominal Data
  - Labels; no quantitative value; can be grouped

Statistics: Data Types

- Nominal Data
  - Labels; no quantitative value; can be grouped

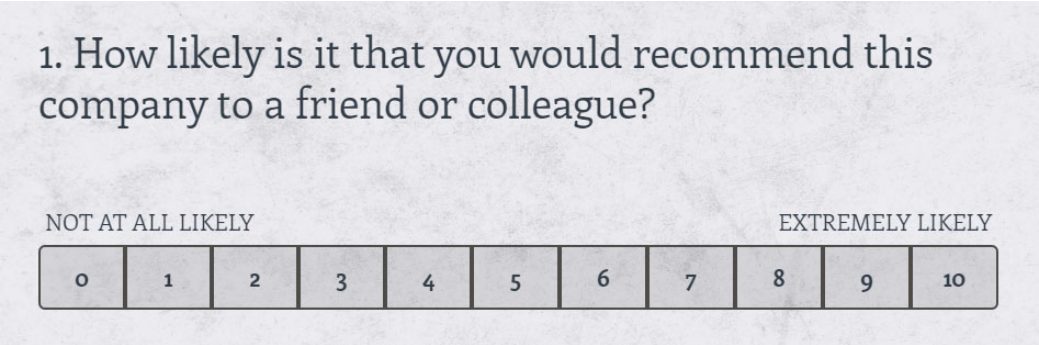
Statistics: Data Types

- Ordinal
  - Non-numerical values; can be ranked



Statistics: Data Types

- Ordinal
  - Non-numerical values; can be ranked



Statistics: Data Types

- Interval
  - Numerical values; equal distance between; known order and differences

Statistics: Data Types

- Interval
  - Numerical values; equal distance between; known order and differences

Statistics: Data Types

- Ratio
  - Can be compared

The ratio between coke cans to orange juice

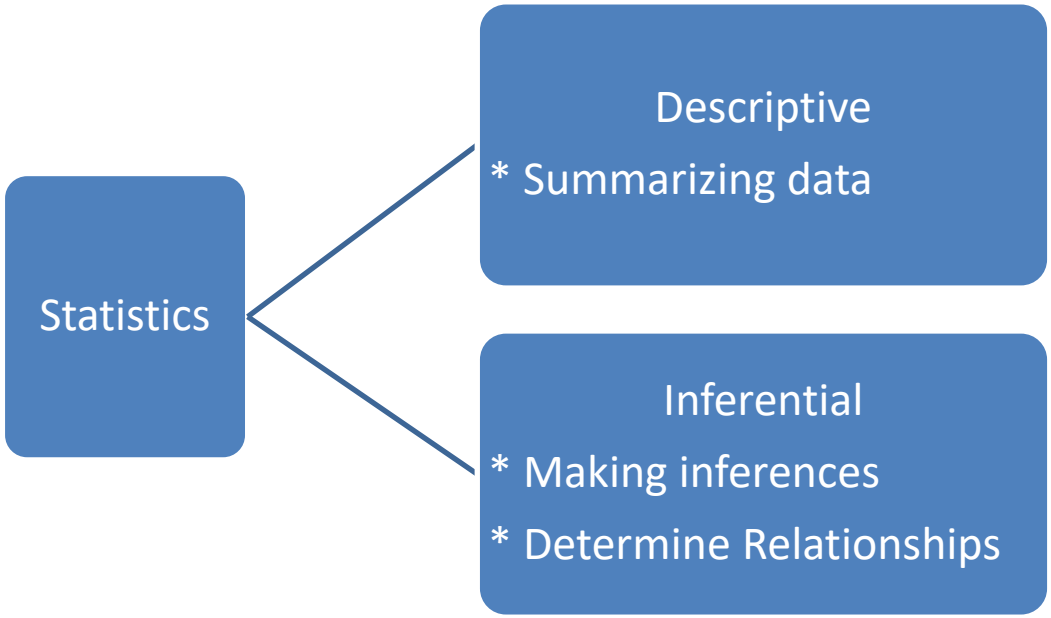
Ratio Examples

5:9	Boys:Girls

Statistics


Business Chart - Visual
Business Chart
Space Usage (750 Mb)
1,505
18,321

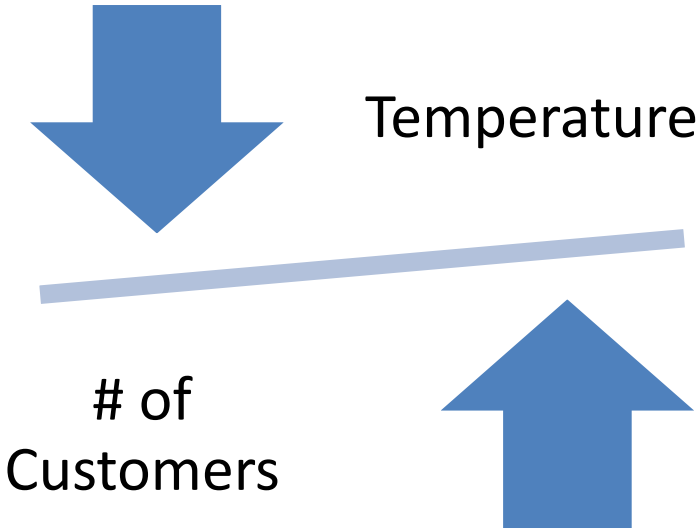
Statistics


Statistics
Descriptive
* Summarizing data
Inferential
* Making inferences
* Determine Relationships

Statistics

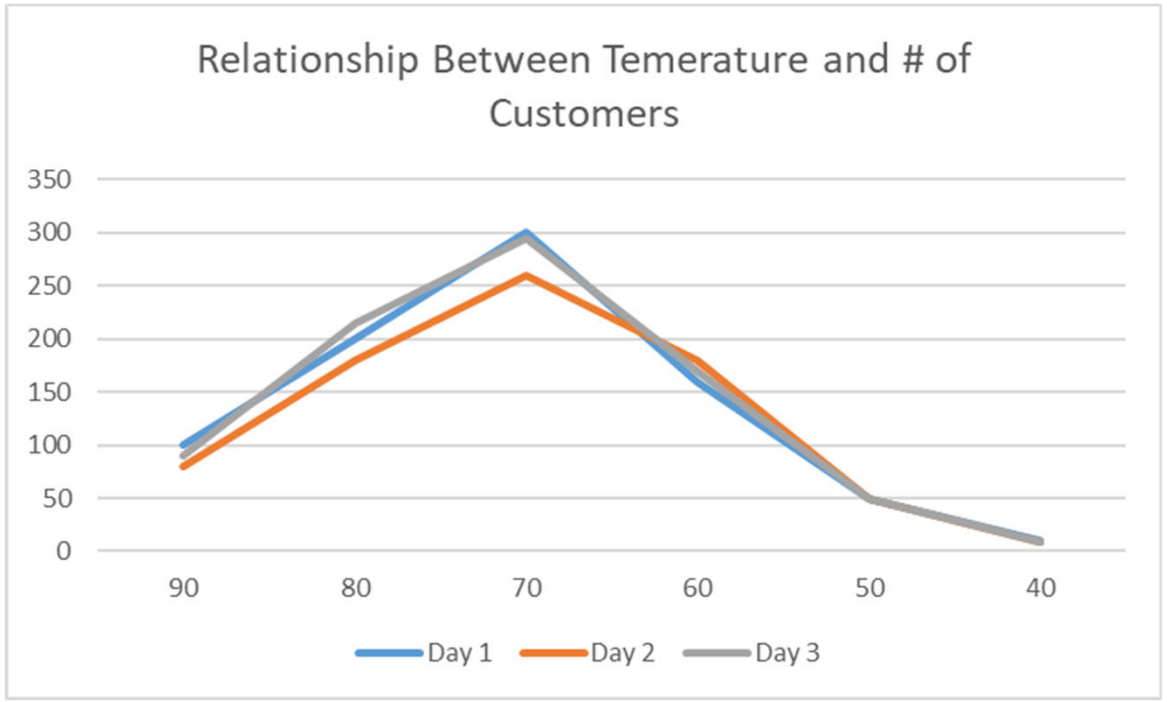
Temerature	Number of Customers		
	Day 1	Day 2	Day 3
90	100	80	90
80	200	180	215
70	300	260	295
60	160	180	170
50	50	50	49
40	10	9	8

Statistics



Temerature	Number of Customers		
	Day 1	Day 2	Day 3
90	100	80	90
80	200	180	215
70	300	260	295
60	160	180	170
50	50	50	49
40	10	9	8

Statistics



Temerature	Number of Customers		
	Day 1	Day 2	Day 3
90	100	80	90
80	200	180	215
70	300	260	295
60	160	180	170
50	50	50	49
40	10	9	8

Statistics

- Descriptive Statistics
  - Describe data
- Inferential Statistics
  - Describe and infer

Module 1 Video 3: Statistics

Population

- The entire set (of interest)

Sample

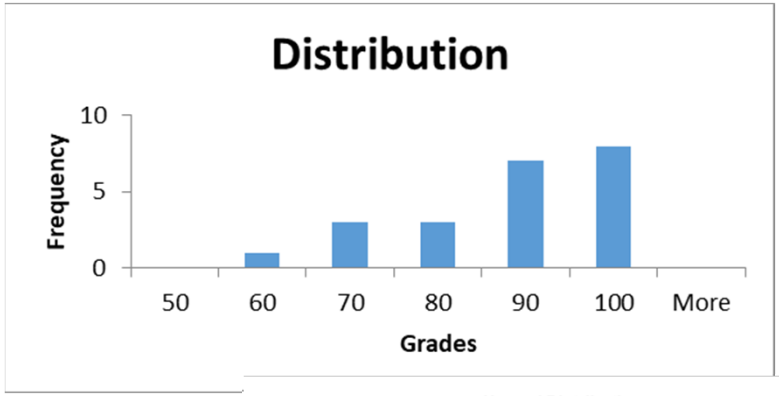
- A subset of a Population

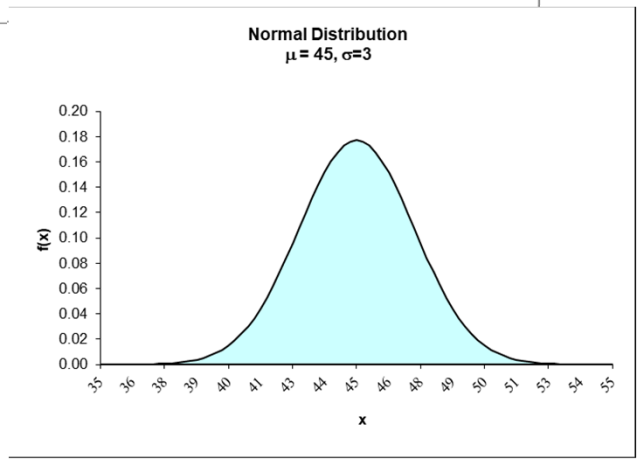
Random

- all items have equal chance to be selected

Statistics: Central Tendency

Statistics: Central Tendency

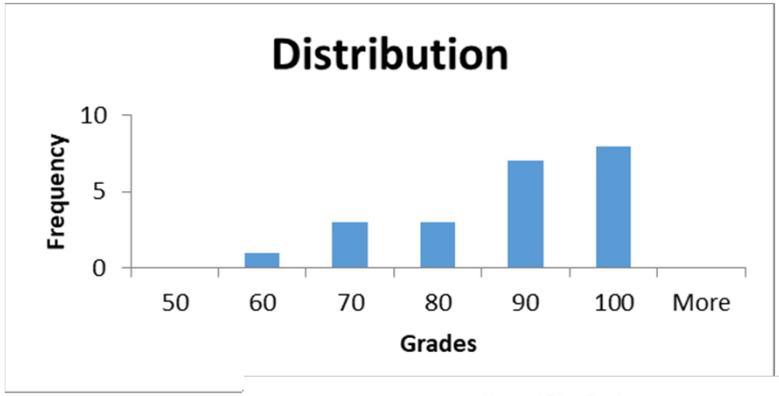


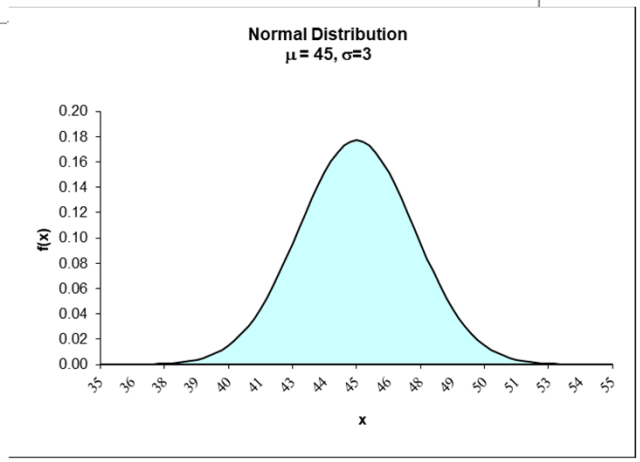


Distribution:

- Shows all values in a data set and their frequency

Statistics: Central Tendency





Central Tendency: a value describes the center or central location of a data set.

- There are three ways to describe the central tendency: mean, median, and mode.

Statistics: Central Tendency

Mean

- Numerical average of the data set

Congratulations!
Your test score is
80!

Statistics: Central Tendency

Mean (μ mu)

Studens	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
Grades	90	98	56	67	89	78	98	100	64	89	98	76	95	100	90	85	78	95	86	89	91	67

Population mean = μ = (Σ Xi) / N

Where Σ = the sum of

Xi = individual datum value

N = the number of datum in the population

(90+ 98+ 56+ 67+ 89+ 78+ 98+ 100+
64+ 89+ 98+ 76+ 95+ 100+ 90+ 85+
78+ 95+ 86+ 89+ 91+ 67) **/ 22 = 85.4**

Statistics: Central Tendency

Mean ($\bar{x}$ x bar)

Sample mean $\bar{x} = (\Sigma x_i) / n$

Where Σ = the sum of

x_i = individual datum value

n = the number of datum in the sample

Median

- Score at 50 percentile; the number in the middle

Congratulations!
Your test score is
80!

Statistics: Central Tendency

Median

Studens	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
Grades	90	98	56	67	89	78	98	100	64	89	98	76	95	100	90	85	78	95	86	89	91	67

First, we have to rank order the numbers

Studens	C	I	D	V	L	F	Q	P	S	E	J	T	A	O	U	M	R	B	G	K	H	N
Grades	56	64	67	67	76	78	78	85	86	89	89	89	90	90	91	95	95	98	98	98	100	100

Since there are two numbers in an even size data set, we will add the 2 numbers then divide the sum by 2 to obtain the median

(89+89) / 2 = 89

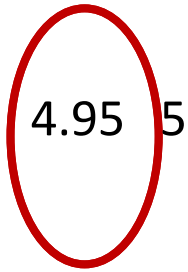
Statistics: Central Tendency

Median

Coffee 3.25 5.25 5.25 3.55 4.95

First, we have to rank order the numbers

Coffee 3.25 3.55 4.95 5.25 5.25

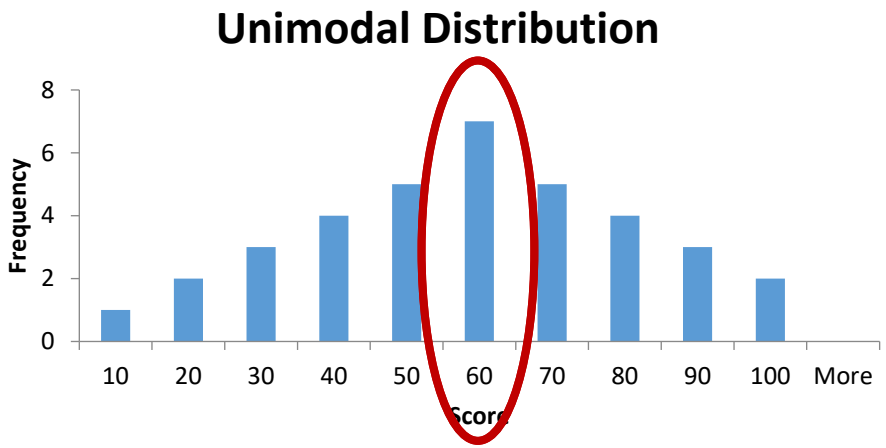


$4.95 is the median

Statistics: Central Tendency

Mode

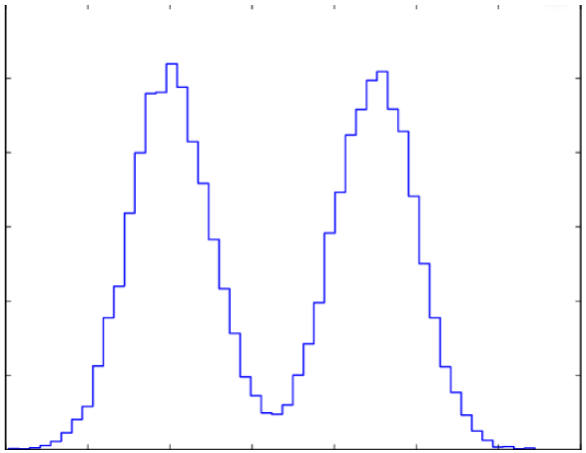
- The most frequently occurring; the most common



Statistics: Central Tendency

Mode

- The most frequently occurring; the most common



Bimodal Distribution

Mode

- The most frequently occurring; the most common

Statistics: Central Tendency

Statistics: Central Tendency 2: Mode, Mean, Median – Which One?

Statistics: Central Tendency

- Mode
  - Nominal data – outliers are fine
    - Which brand do you prefer?
- Mean
  - Interval and ratio data not excessively skewed
    - What is the average salary?
- Median
  - Ordinal Data – skewed data is fine
    - How satisfy are you?

Statistics: Central Tendency

Statistics: Standard Deviation and Variance

Spread of data

Statistics: Standard Deviation and Variance

- Standard Deviation
  - Average distance from the mean

Statistics: Standard Deviation and Variance

- Standard Deviation (Population)

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Where σ = lowercase sigma = standard deviation

Σ = the sum of

Xi = individual datum value

μ = mean of population

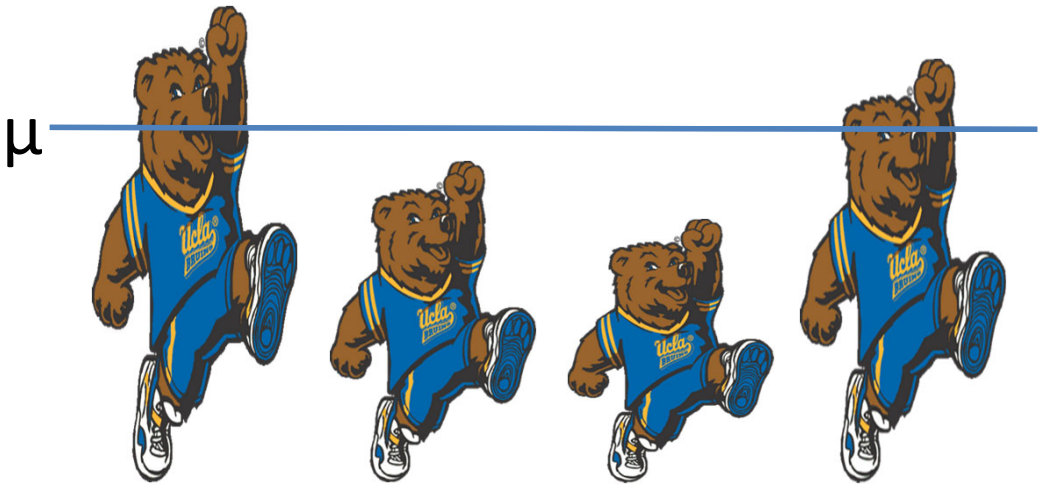
N = the number of datum in the population

Statistics: Standard Deviation and Variance

- Standard Deviation (Population)
  - Average distance from the mean

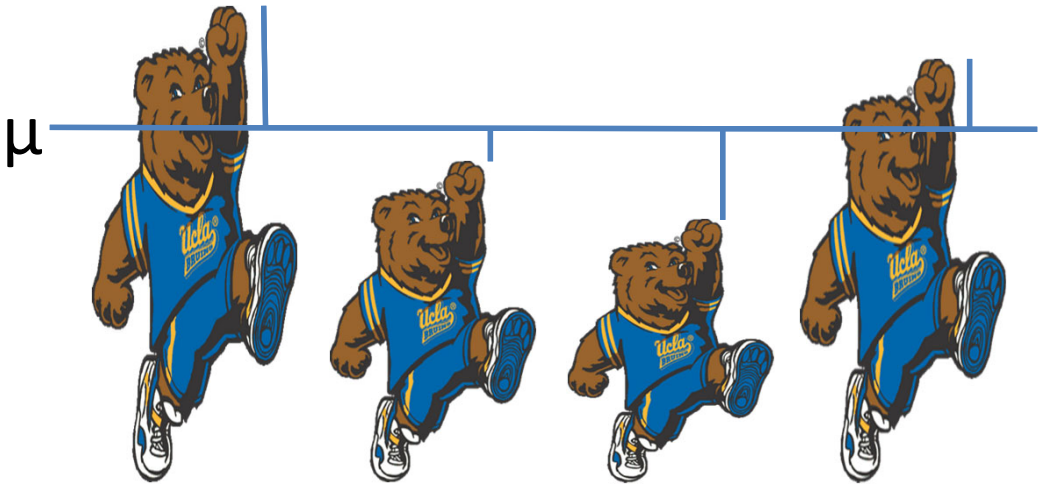
Statistics: Standard Deviation and Variance

- Standard Deviation (Population)
  - Average distance from the mean



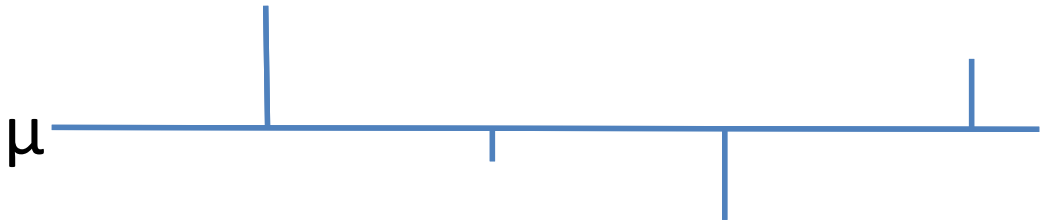
Statistics: Standard Deviation and Variance

- Standard Deviation (Population)
  - Average distance from the mean



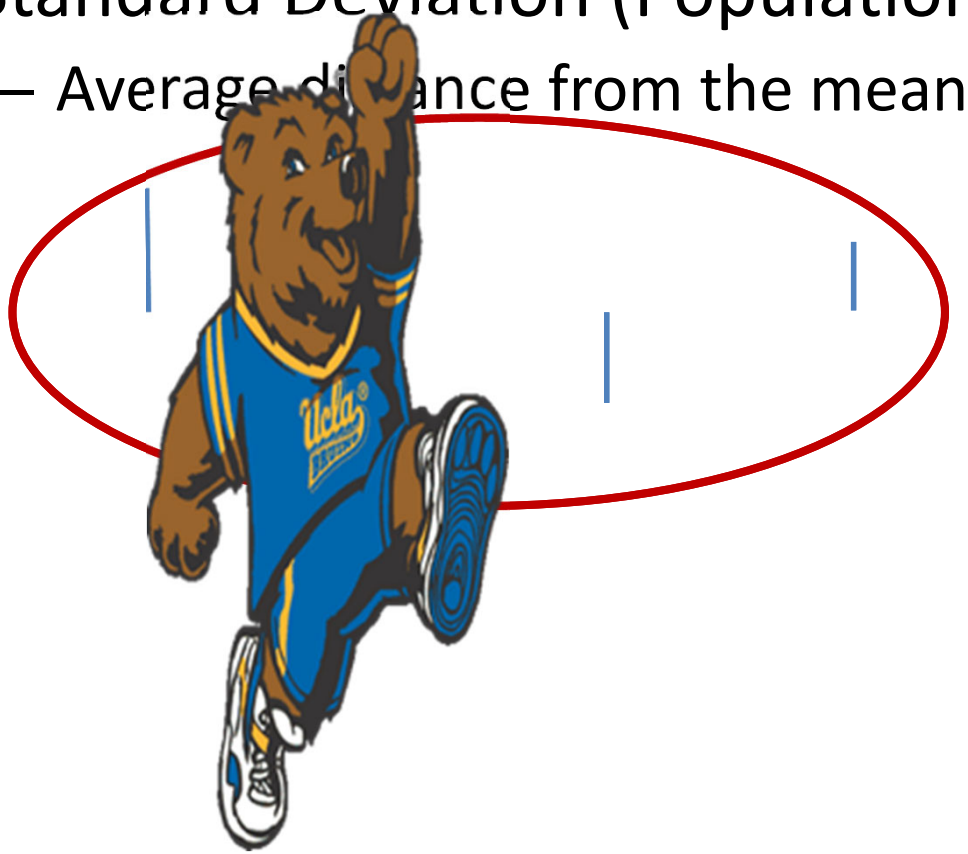
Statistics: Standard Deviation and Variance

- Standard Deviation (Population)
  - Average distance from the mean



Statistics: Standard Deviation and Variance

- Standard Deviation (Population)
  - Average distance from the mean



Statistics: Standard Deviation and Variance

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

- Standard Deviation
  - Price of Ice Tea **(1, 3, 6, 8, 7)**
    - N = 5
    - μ = (1+3+6+8+7)/ 5 = 5

$$\sigma = \sqrt{\frac{(1-5)^2 + (3-5)^2 + (6-5)^2 + (8-5)^2 + (7-5)^2}{5}}$$

$$\sigma = \sqrt{\frac{16 + 4 + 1 + 9 + 4}{5}} = \sqrt{\frac{34}{5}}$$

$$\sigma = \sqrt{\frac{34}{5}} = \sqrt{6.8} = \boxed{2.61}$$

Statistics: Standard Deviation and Variance

- Standard Deviation (Sample)

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Where s = standard deviation (sample)

Σ = the sum of

xi = individual datum value

$\bar{x}$ = mean of sample

N = the number of datum in the population

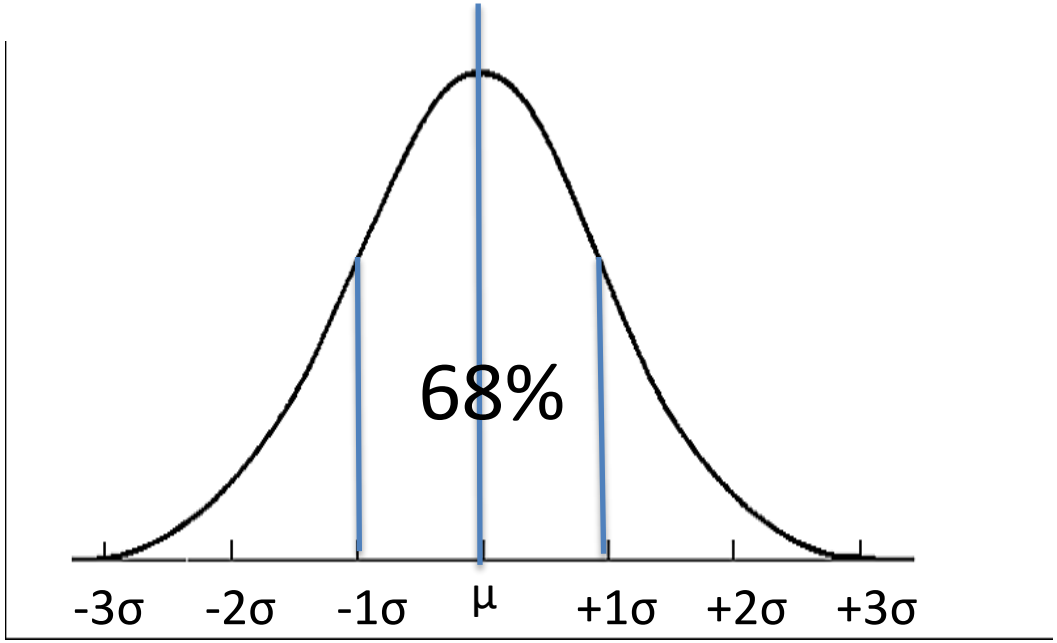
Statistics: Standard Deviation and Variance

- Variance (Population) = σ^2
- Variance (Sample) = s^2

Statistics: Standard Deviation and Variance

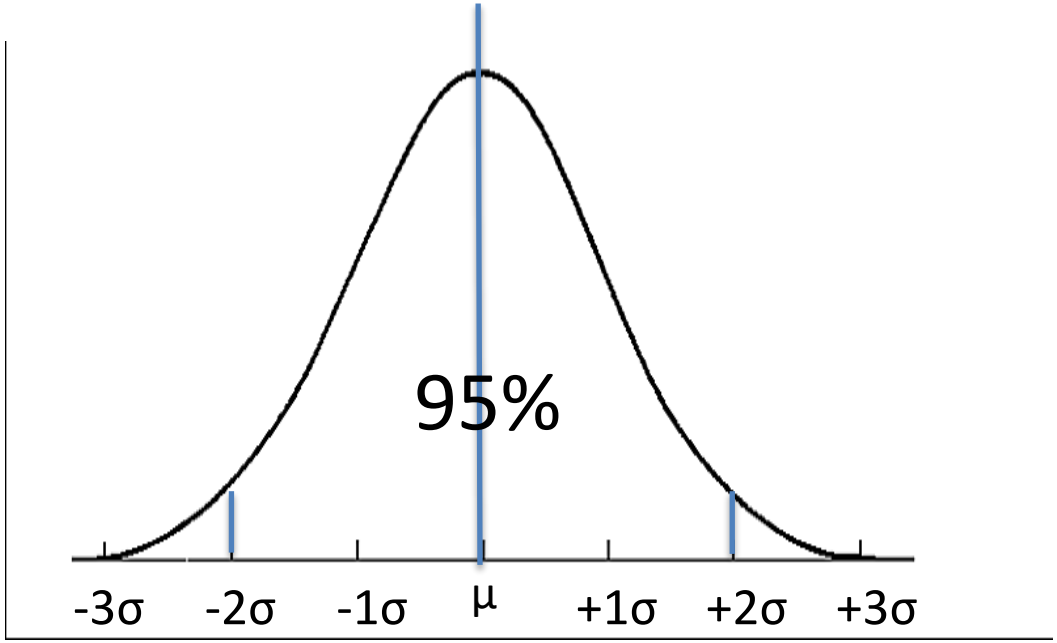
Statistics: Standard Deviation and Variance Empirical Rule

Statistics: Standard Deviation and Variance Empirical Rule



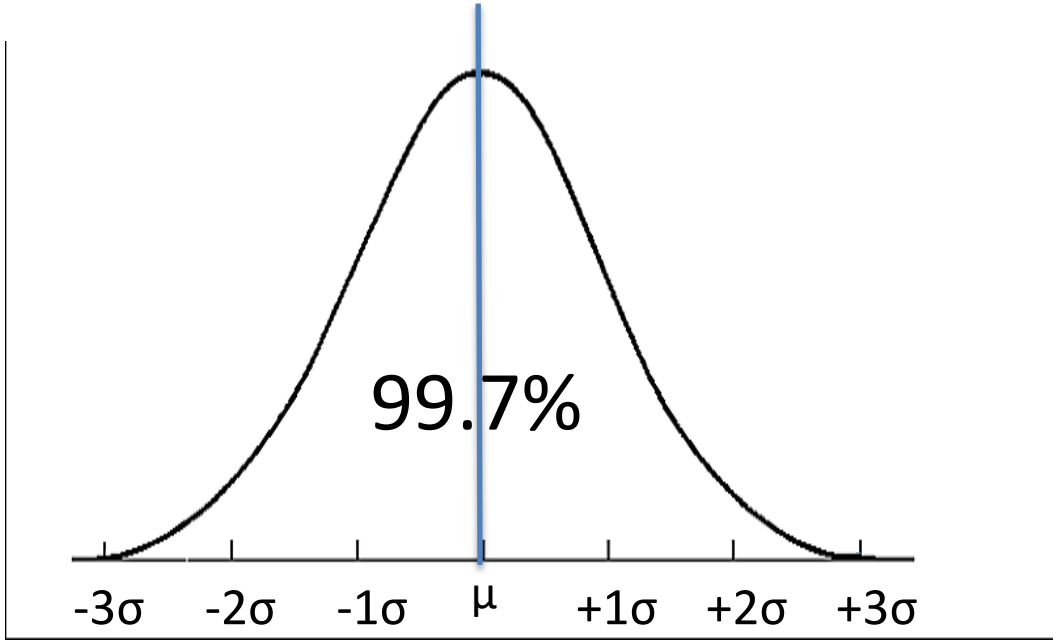
- Standard Deviation
  - 68–95–99.7 rule
  - Empirical Rule

Statistics: Standard Deviation and Variance Empirical Rule



- Standard Deviation
  - 68–95–99.7 rule
  - Empirical Rule

Statistics: Standard Deviation and Variance Empirical Rule

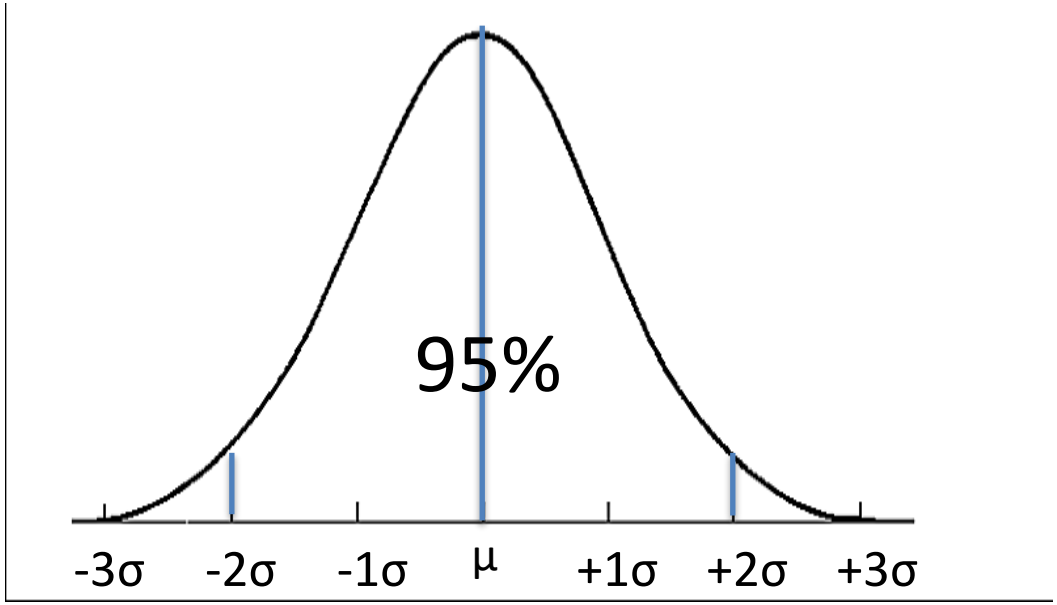


- Standard Deviation
  - 68–95–99.7 rule
  - Empirical Rule

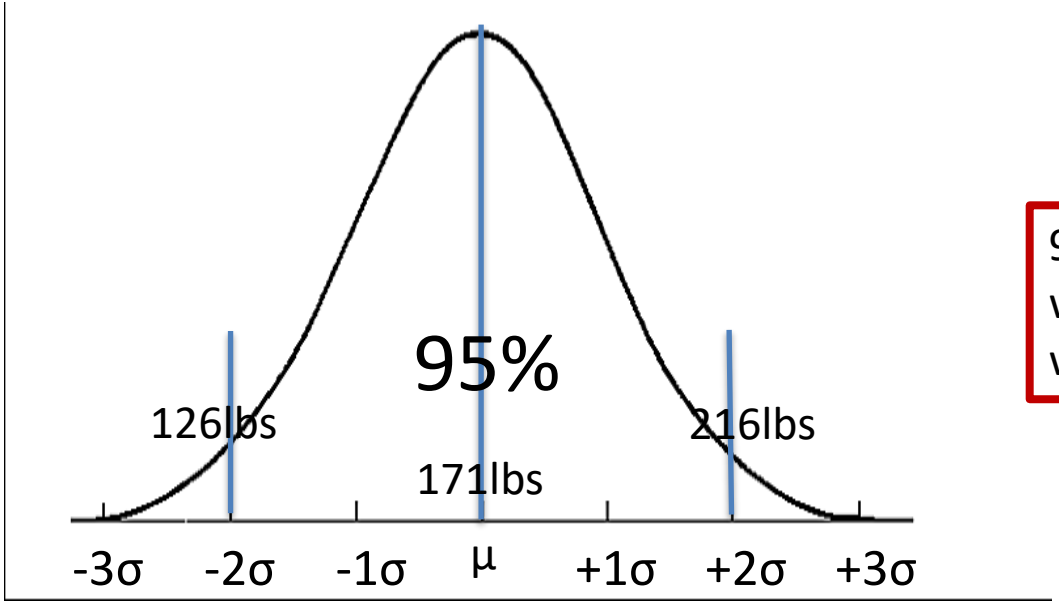
Statistics: Standard Deviation and Variance Empirical Rule

- Standard Deviation
  - U.S. males average weight is 171 pounds. The value at 2 σ is 216 pounds
  - What is the standard deviation σ?

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$



Statistics: Standard Deviation and Variance Empirical Rule



$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

95 Percent males in the United States weight between 126lbs and 216lbs with a standard deviation of 22.5lbs

σ = ?

μ = 171 lbs

Value at 2 standard deviation from mean is 216 lbs

σ = (216 -171)/2; **σ = 22.5**

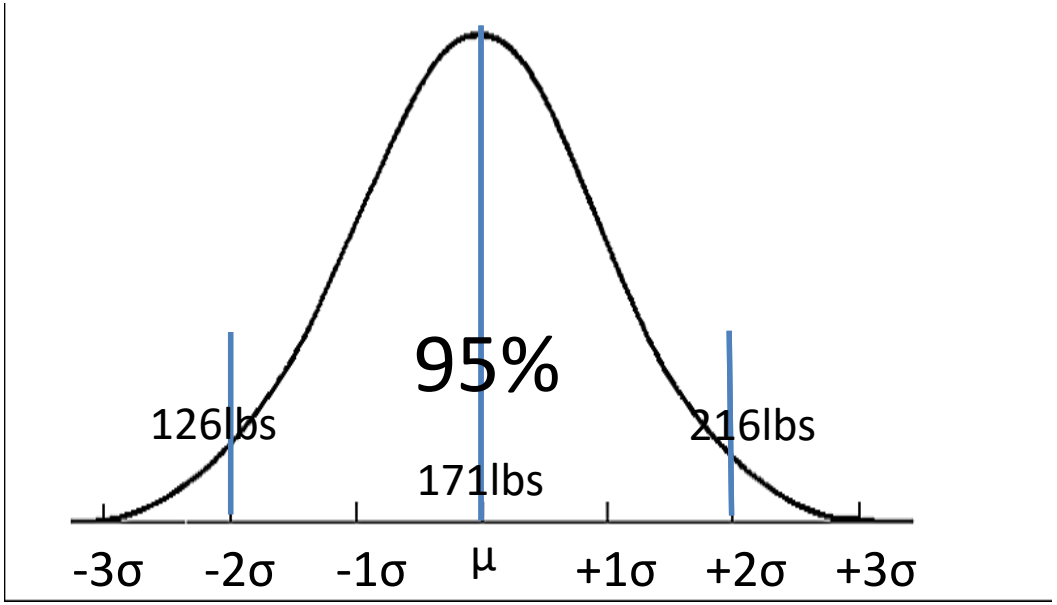
Statistics: Standard Deviation and Variance Empirical Rule

Statistics: Z-Score

Z-score

- Describes the location of a raw value in relations to the mean and standard deviation

Statistics: Z-Score for Population



$$Z_X = (X - \mu) / \sigma$$

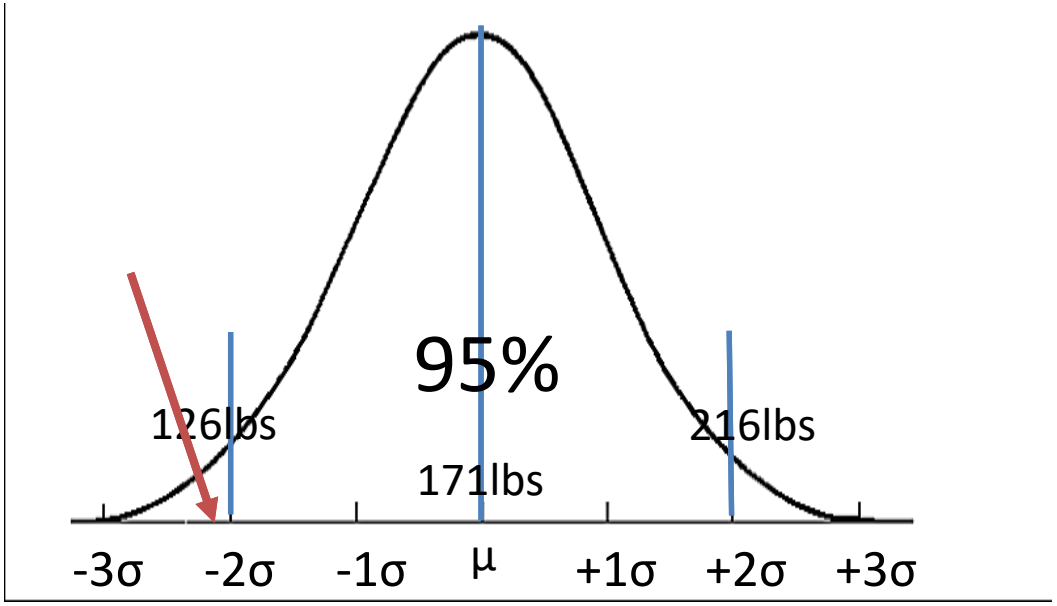
σ = ?

μ = 171 lbs

A 95 percentile male is 216 lbs (95 percentile is 2 standard deviation from mean)

σ = (216 -171)/2; **σ = 22.5**

Statistics: Z-Score for Population



σ = ?

μ = 171 lbs

A 95 percentile male is 216 lbs (95 percentile is 2 standard deviation from mean)

σ = (216 -171)/2; **σ = 22.5**

What is the Z-Score for 120lbs?

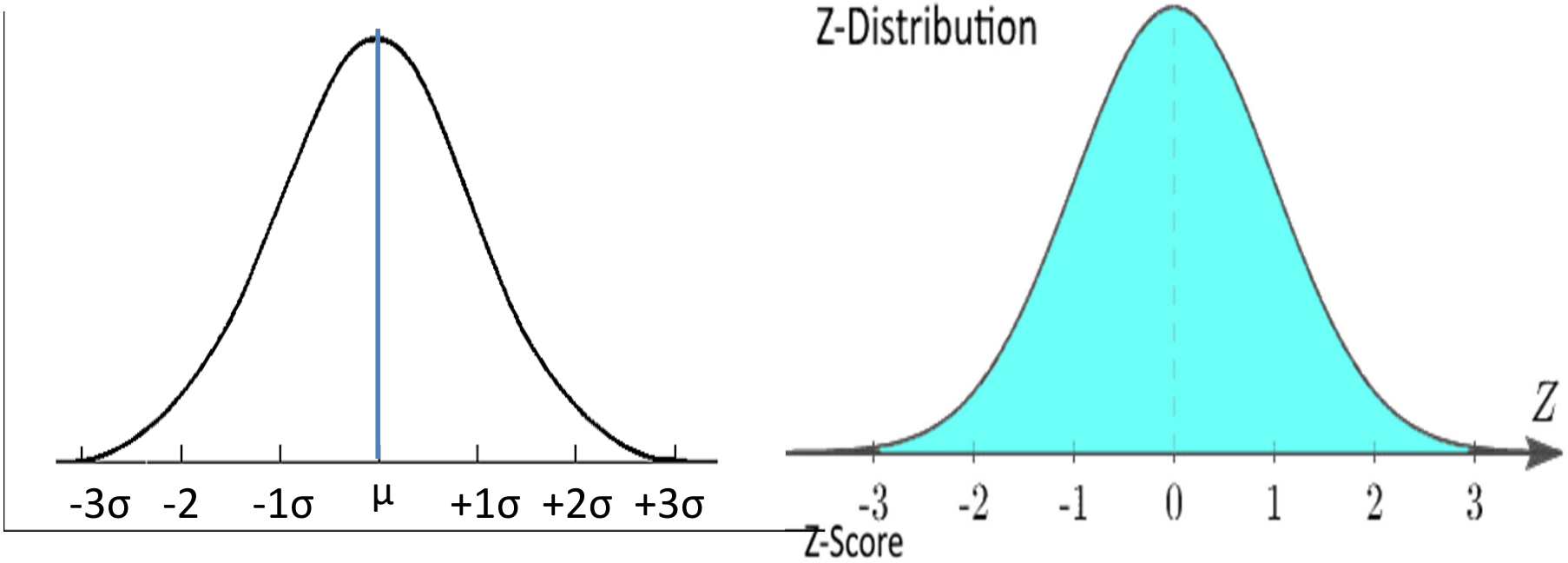
$Z_X = (X - \mu) / \sigma$

$Z_{120} = (120 - 171) / 22.5$

$Z_{120} = -2.27$

- a 120lbs male is 2.27 standard deviation from the mean

Statistics: Z-Score for Population



Normal Distribution → Standard Normal Distribution

$\sigma = 1$

$\mu = 0$

Statistics: Z-Score for Sample

$$z_x = (x - \bar{x}) / s$$

Statistics: Z-Score

Statistics: z-Distribution and t-Distribution

In the context of sample size

Statistics: z-Distribution and t-Distribution

- Standard Deviation (Population)

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Where σ = lowercase sigma = standard deviation

Σ = the sum of

Xi = individual datum value

μ = mean of population

N = the number of datum in the population

Statistics: Standard Deviation and Variance

- Standard Deviation (Sample)

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Where s = standard deviation (sample)

Σ = the sum of

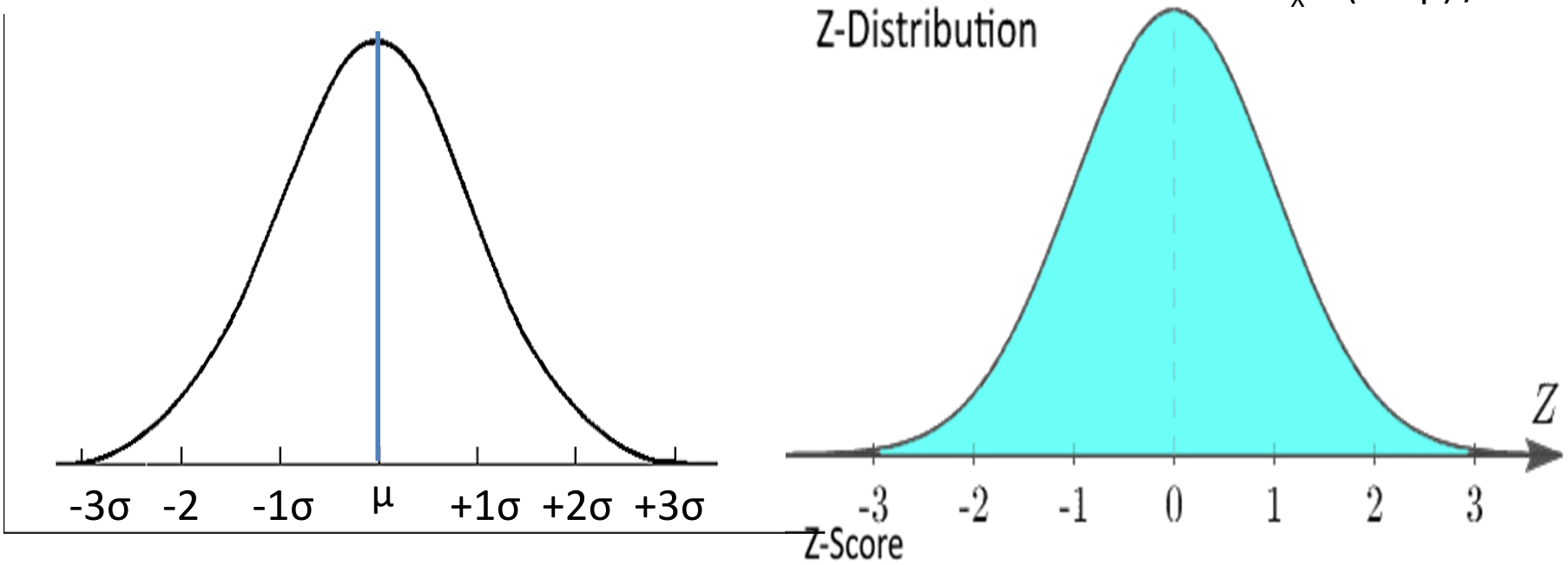
xi = individual datum value

$\bar{x}$ = mean of population

N = the number of datum in the population

Statistics: z-Distribution and t-Distribution

$Z_X = (X - \mu) / \sigma$

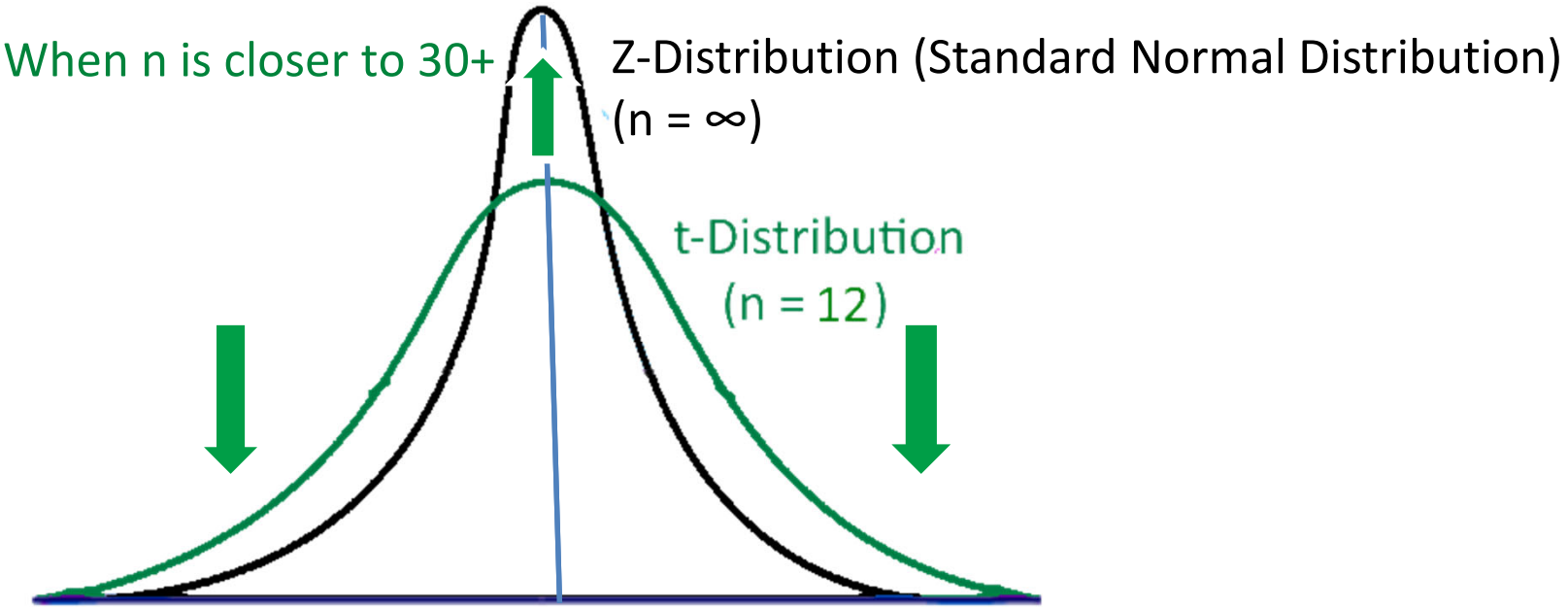


Normal Distribution → Standard Normal Distribution

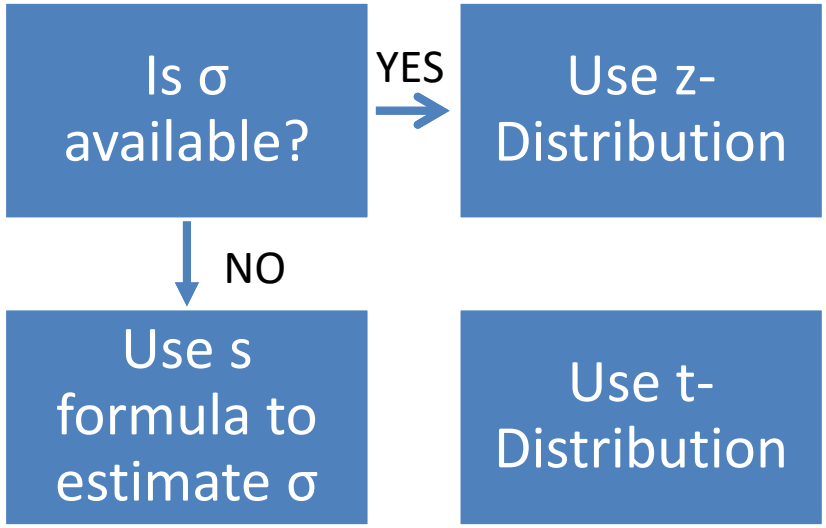
$\sigma = 1$

$\mu = 0$

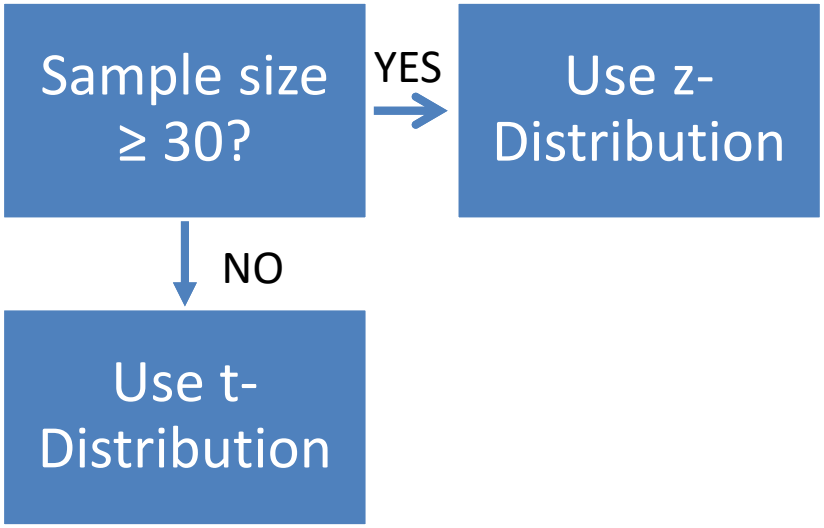
Statistics: z-Distribution and t-Distribution



Statistics: z-Distribution and t-Distribution



Statistics: z-Distribution and t-Distribution



Statistics: z-Distribution and t-Distribution

Statistics: Covariance

Covariance

- Measures linear relationship between two variables

Statistics: Covariance

- Covariance (Population)

$$\mathrm{Cov}\,(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

Where

Σ = the sum of

Xi = individual datum value

$\bar{X}$ = population mean

n = the number of datum in the population

Statistics: Covariance

- Covariance (Sample)

$$\mathrm{Cov}\,(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n - 1}$$

Where

Σ = the sum of

Xi = individual datum value

$\bar{X}$ = sample mean

n = the number of datum in the sample

Statistics: Covariance

$$\mathrm{Cov}\,(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n - 1}$$

X	Y	$\bar{X}$	$\bar{y}$	(X-$\bar{X}$)	(Y-$\bar{Y}$)	(X-$\bar{X}$)(Y-$\bar{Y}$)
Temerature	Customer					
100	60	87.5	180	12.5	-120	-1495.83
95	98	87.5	180	7.5	-81.7	-612.5
90	100	87.5	180	2.5	-79.7	-199.167
85	200	87.5	180	-2.5	20.3	-50.8333
80	300	87.5	180	-7.5	120	-902.5
75	320	87.5	180	-12.5	140	-1754.17

Σ (X-$\bar{X}$)(Y-$\bar{Y}$) = -5015

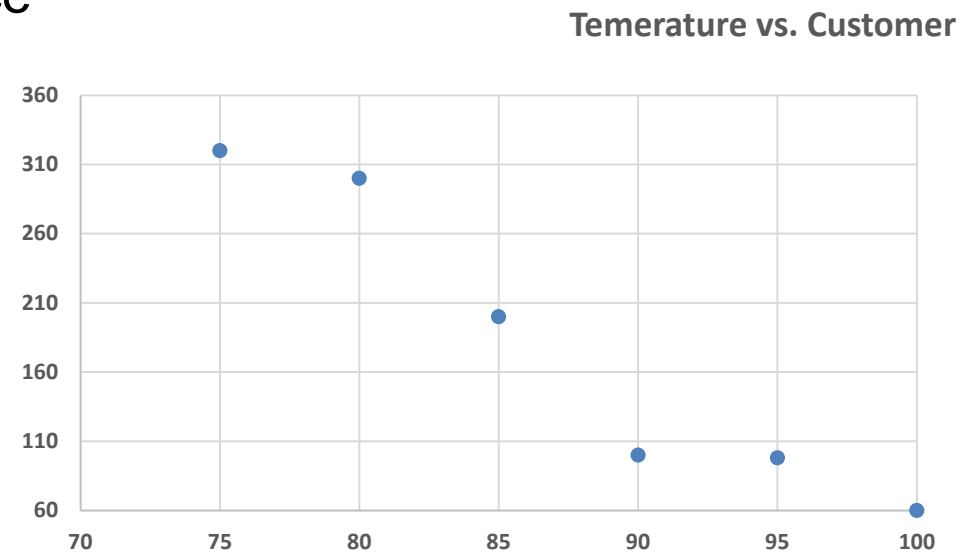
Σ (X-$\bar{X}$)(Y-$\bar{Y}$) / n-1 = -5015 / 5

Cov(X,Y) = -1003

Statistics: Covariance

X	Y
Temerature	**Customer**
100	60
95	98
90	100
85	200
80	300
75	320

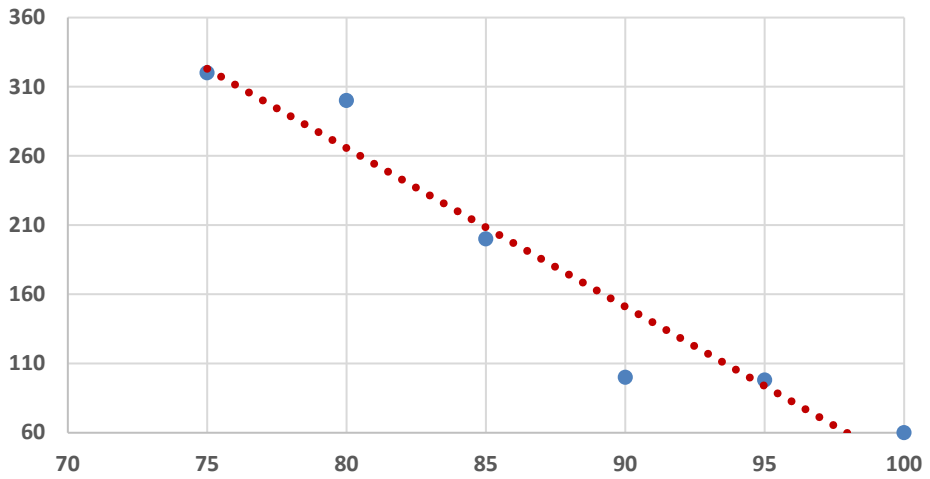
Cov(X,Y) = -1003



Statistics: Covariance

X	Y
Temerature	**Customer**
100	60
95	98
90	100
85	200
80	300
75	320

Cov(X,Y) = -1003



Statistics: Covariance

X	Y	$\bar{X}$	$\bar{y}$	$(X-\bar{X})$	$(y-\bar{y})$	$(X-\bar{X})(y-\bar{y})$
Chips	**Weight**					
2	120	4.83	185	-2.83	-65.3	185.111
5	207	4.83	185	0.17	21.7	3.61111
3	122	4.83	185	-1.83	-63.3	116.111
8	247	4.83	185	3.17	61.7	195.278
5	227	4.83	185	0.17	41.7	6.94444
6	189	4.83	185	1.17	3.67	4.27778

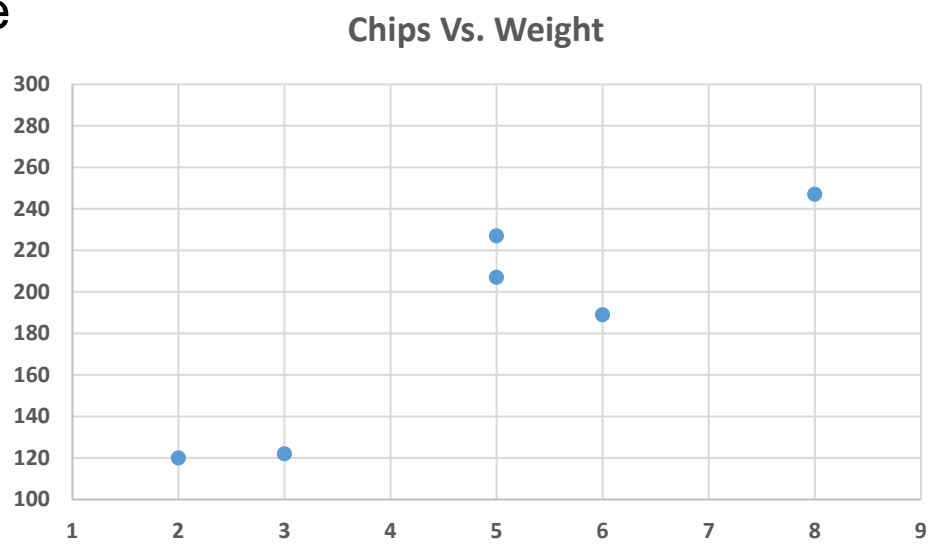
$\Sigma (X-\bar{X})(Y-\bar{Y}) = 511.33$

$\Sigma (X-\bar{X})(Y-\bar{Y}) / n-1 = 511.33 / 5$

Cov(X,Y) = 102.26

Statistics: Covariance

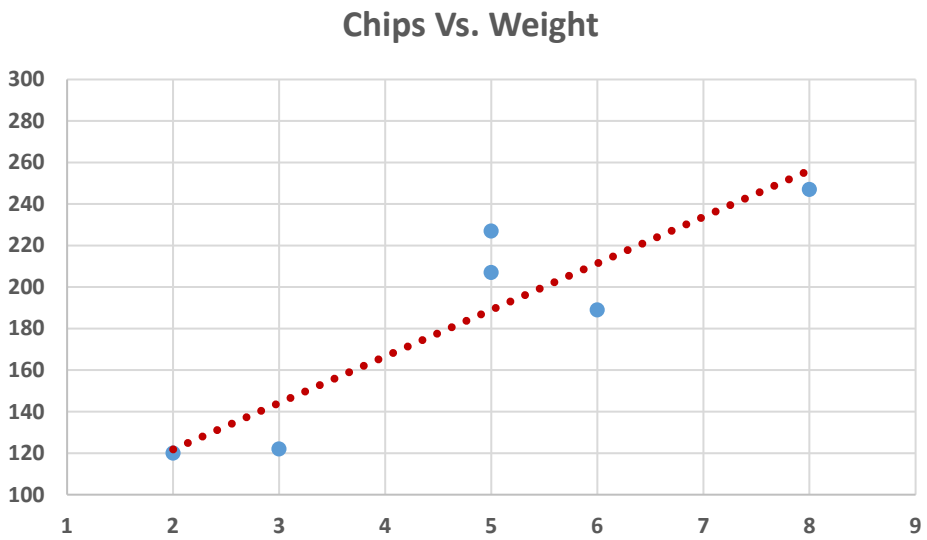
X	Y
Chips	Weight
2	120
5	207
3	122
8	247
5	227
6	189



Cov(X,Y) = 102.26

Statistics: Covariance

X	Y
Chips	Weight
2	120
5	207
3	122
8	247
5	227
6	189



Cov(X,Y) = 102.26

Statistics: : Covariance

- **Cov(X,Y) = 0 ?**

Statistics: : Covariance

- + means positive linear relationship
- - means negative linear relationship
- 0 means no linear relationship

- It only gives the direction of the relationship
- No indication of the strength of the relationship

Statistics: Correlation

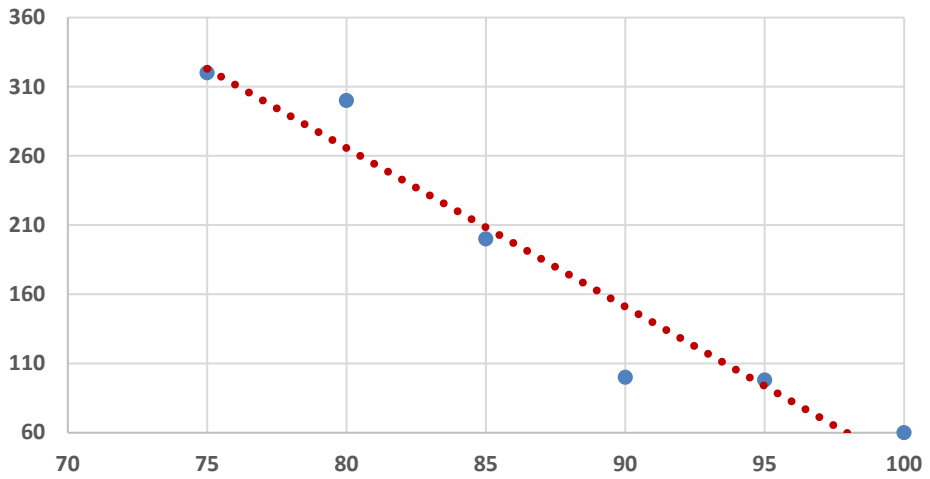
Correlation

- Measures relationship and strength of the relationship between two variables

Statistics: Covariance

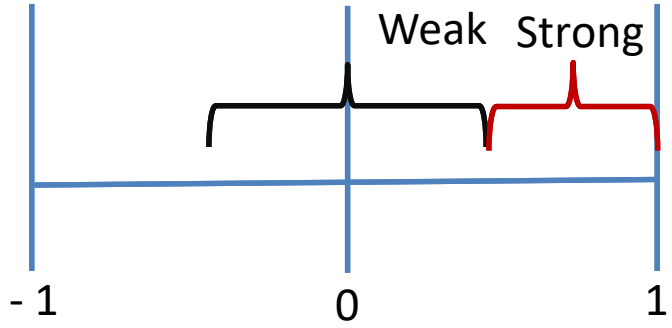
X	Y
Temerature	**Customer**
100	60
95	98
90	100
85	200
80	300
75	320

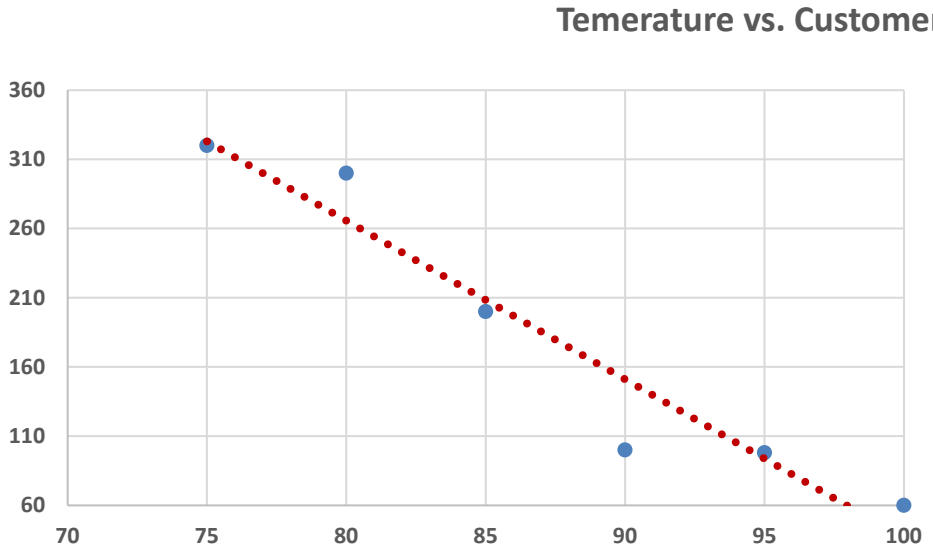
Cov(X,Y) = -1003



Statistics: Correlation

- Pearson Correlation Coefficient
  - Pearson r

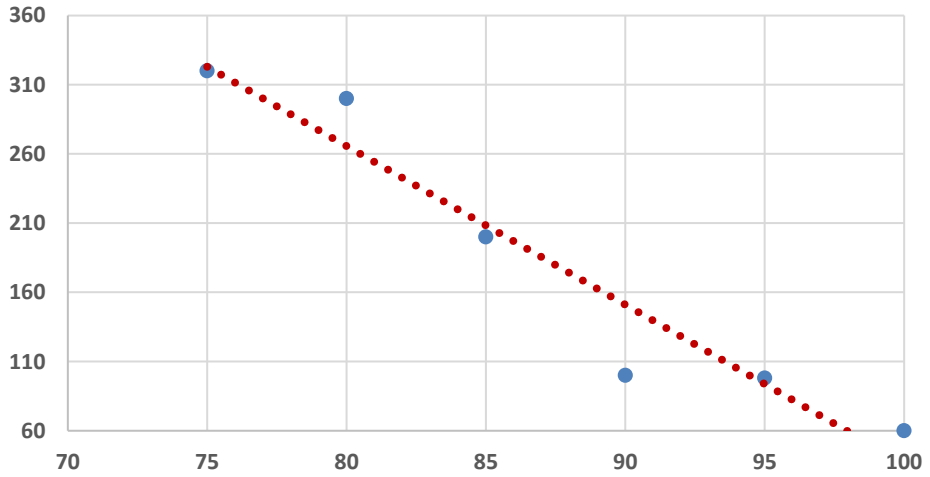




$$r = Cov(X,Y) / s_x s_y$$

Statistics: Correlation

X	Y
Temerature	Customer
100	60
95	98
90	100
85	200
80	300
75	320



Cov(X,Y) = -1003

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

$S_x = 9.35$

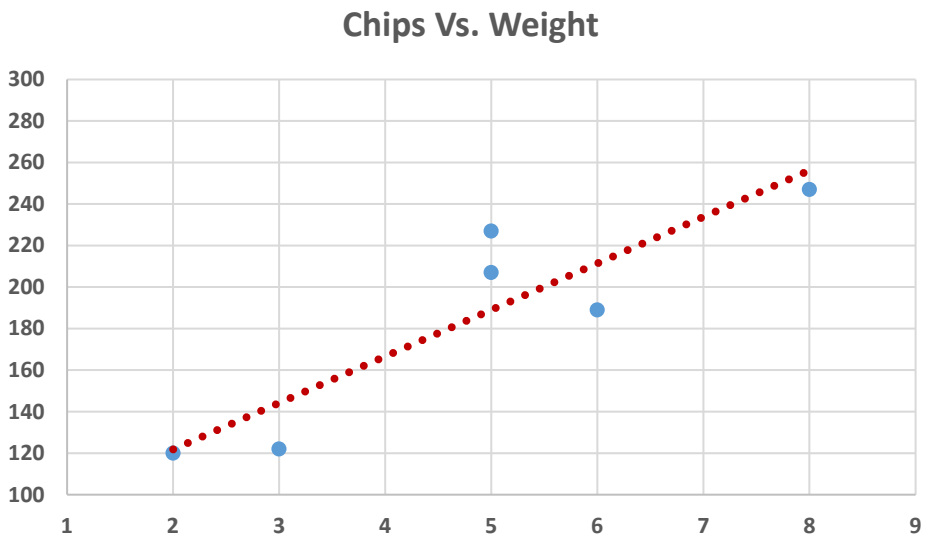
$S_y = 111.29$

$r = Cov(X,Y) / s_x s_y = -1003/(9.35 * 111.29)$

$r = -0.9635$

Statistics: Correlation

X	Y
Chips	Weight
2	120
5	207
3	122
8	247
5	227
6	189



Cov(X,Y) = 102.26

r= Cov(X,Y) / s_x s_y = 0.8948

Statistics: Correlation

- Correlation does not equal to causation

Statistics: Correlation