Descriptive Statistics

Laxminarayen

Data Scientist, Logitech, India

Descriptive Statistics

Agenda-

In this session you will learn about

- Basics of Statistics
- Types of Variables
- Measure of Central Tendancy
- Measure of Dispersion
- Case studies of Central tendencies and Dispersion
- Percentile/Quartile & Correlation and Covariance
- Central Limit Theorem
- Data Visualization and distribution

What is Statistics?

- A branch of mathematics taking and transforming numbers into useful information for decision makers.

- The practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

Why Learn Statistics ?

Case 1 - Answer in 5 seconds !

Case 1 - Answer in 5 seconds !

A college in US has students from the following countries for a Masters degree. Which country is in majority ?

Case 1 - Answer in 5 seconds !

A college in US has students from the following countries.
Which country is in majority ?

US	China	US	Sweden	China
Canada	China	Japan	Mexico	US
China	Germany	India	India	Japan
US	US	US	China	China
India	Japan	England	India	Japan
England	India	China	Mexico	US
Mexico	US	Canada	Pakistan	India
Japan	China	US	Japan	Germany
China	India	India	China	China
Germany	Japan	China	US	Japan

Frequency Table

Country	Frequency
Canada	2
China	12
England	2
Germany	3
India	8
Japan	8
Mexico	3
Pakistan	1
Sweden	1
US	10

Case 2

Problem

A parent changes school of their Son who is studying in 11th standard since his academic results are not good in 10th Standard in his current School.

They change Student A from ABC school to XYZ school

Case 2

Problem

A parent changes school of their Son who is studying in 11th standard since his academic results are not good in 10th Standard in his current School.

They change Student A from ABC school to XYZ school

Results

1. Ranked 15th in ABC school
2. Ranked 2nd in XYZ school

What's the conclusion ?

Case 2

Problem

A parent changes school of their Son who is studying in 11th standard since his academic results are not good in 10th Standard in his current School.

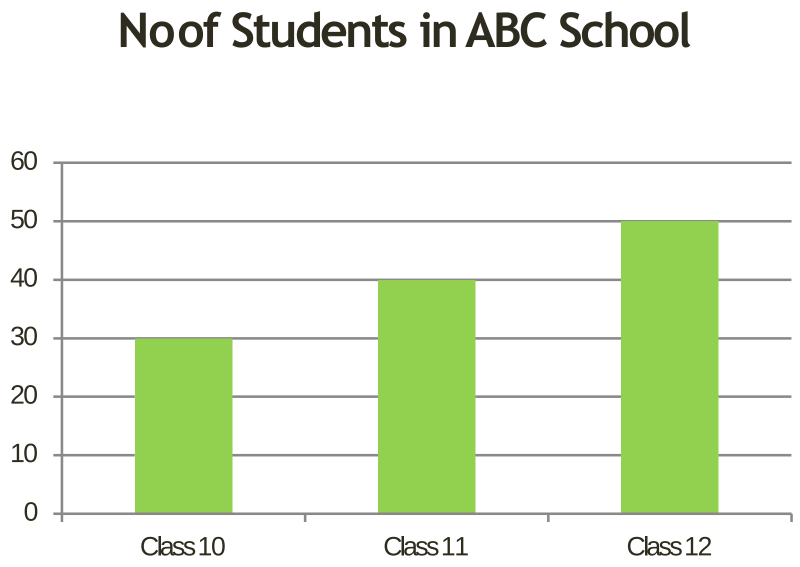
They change Student A from ABC school to XYZ school

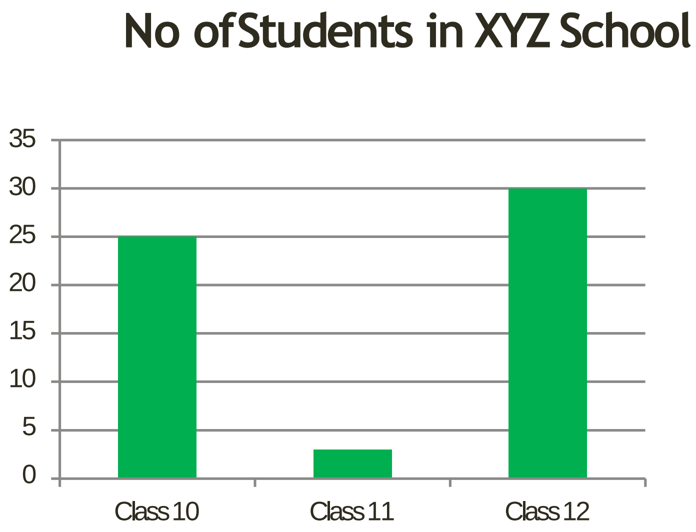
Results

1. Ranked 15th in ABC school
2. Ranked 2nd in XYZ school

What's the conclusion: Has the student improved ?

Number of Students





Why Learn Statistics ?

Why Learn Statistics ?

Knowledge of Statistics allows you to make better sense of the ubiquitous use of numbers.

Why Learn Statistics ?

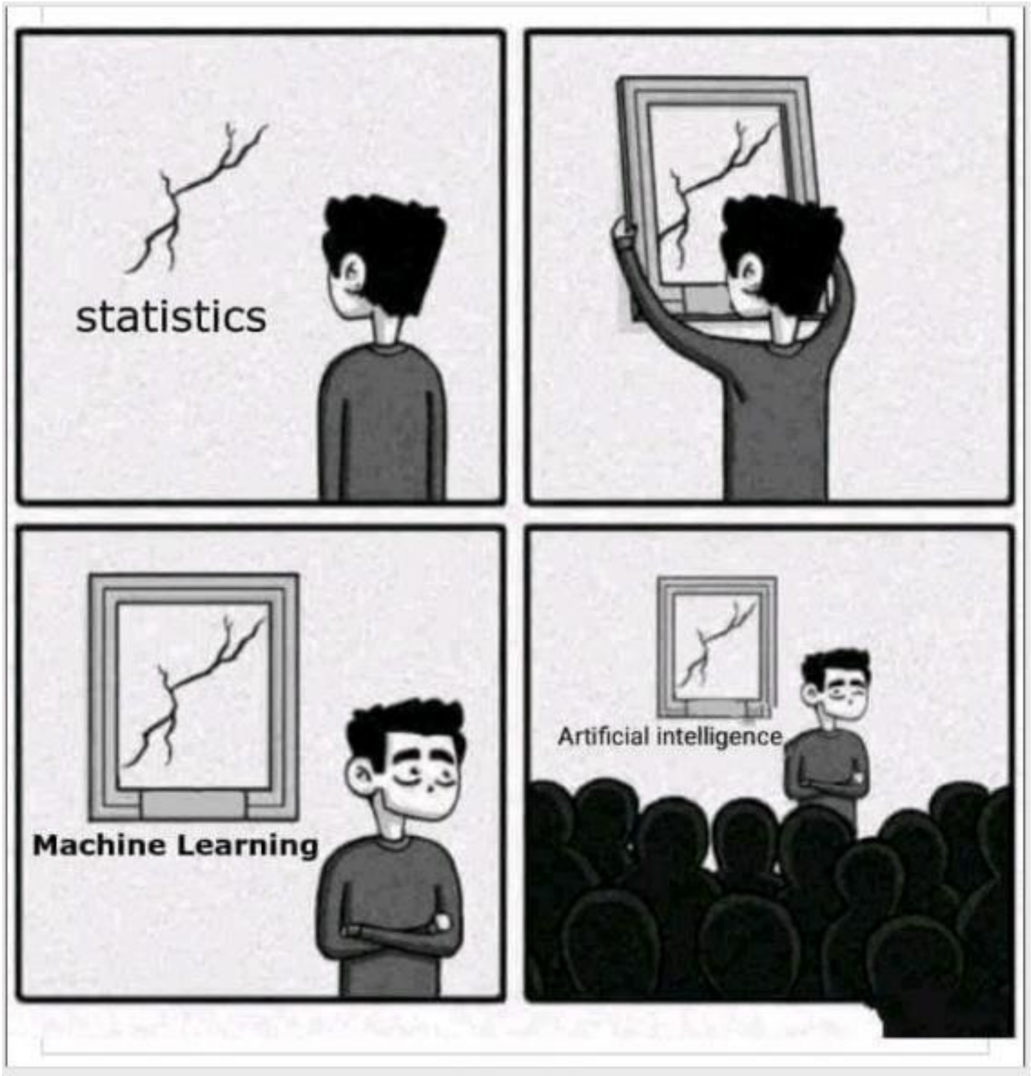
Decision Makers Use Statistics for Various Purposes:

- Present and describe business data and information properly
- Draw conclusions about large sets using information collected from subsets
- Make reliable forecasts about a business activity
- Improve business processes

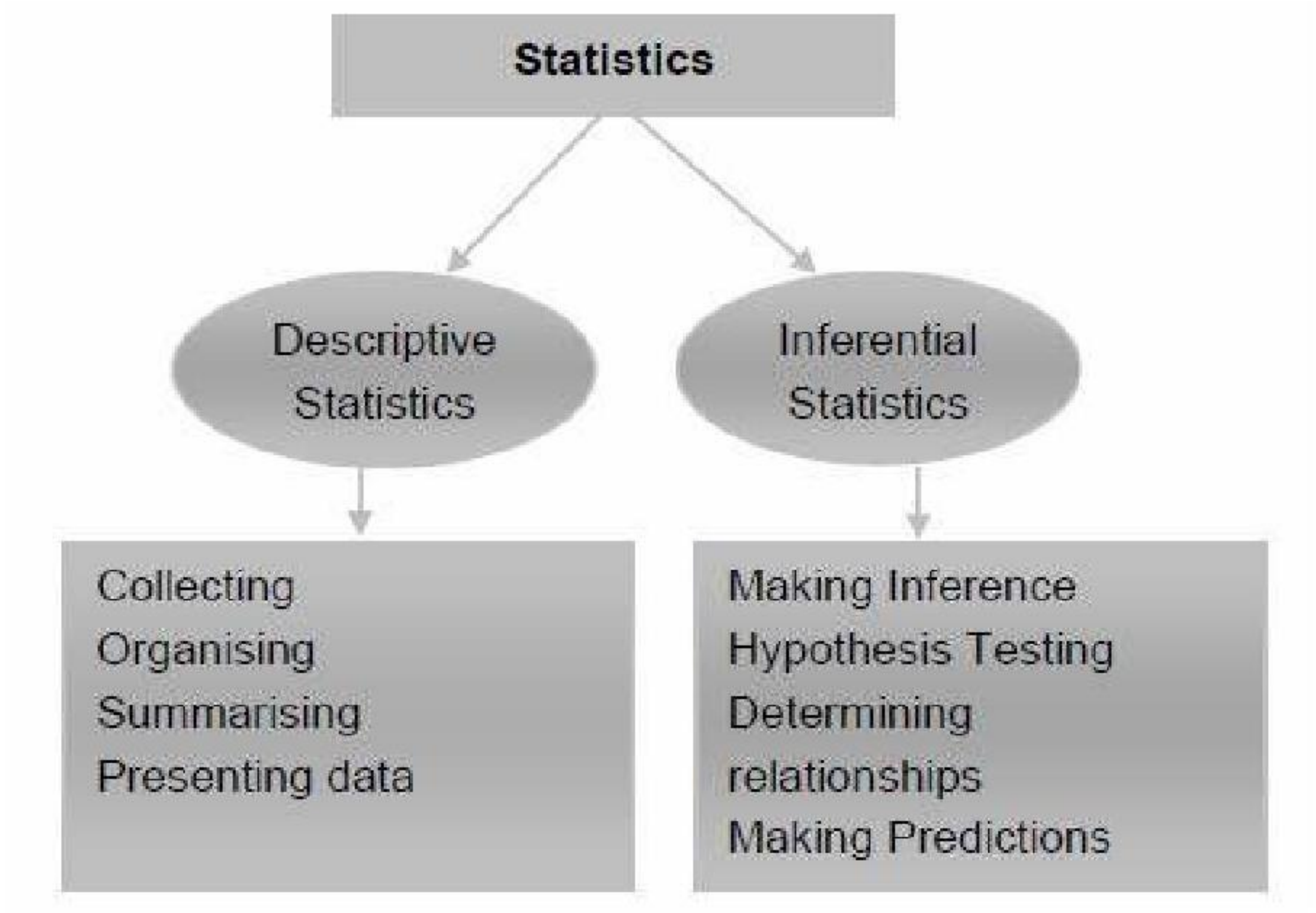
Statistics is …

1. *Collecting Data*
2. *Analyzing Data*
3. *Interpreting Data*
4. *Presenting Data*

What does it Tell?



Types of Statistics



DESCRIPTIVE STATISTICS

- Descriptive statistics are used to describe the basic features of the data in a study.

- They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

INFERENTIAL STATISTICS

- With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone.

- We use inferential statistics to make inferences from our data to more general conditions; we use descriptive statistics simply to describe what's going on in our data.

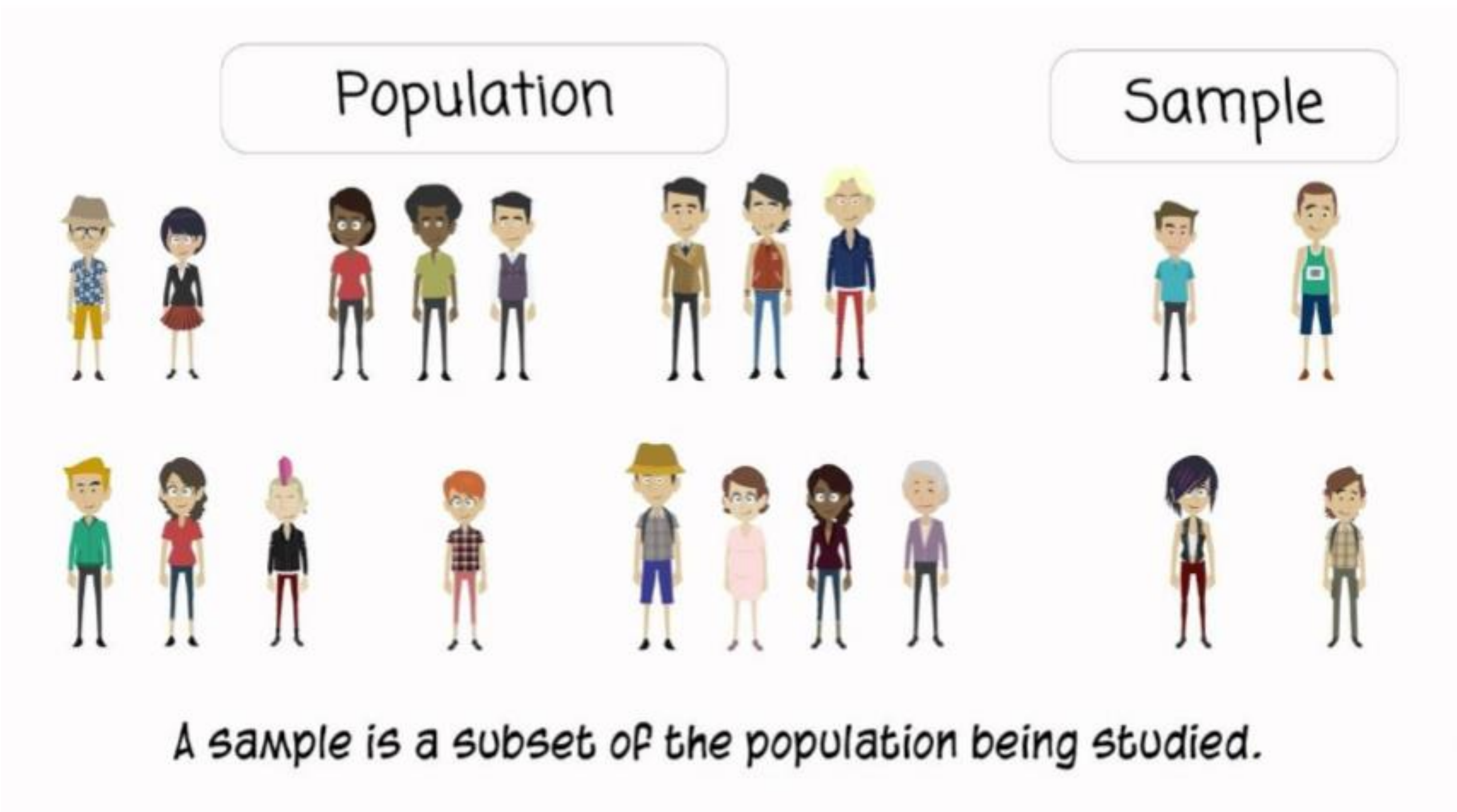
DESCRIPTIVE STATISTICS

POPULATION AND SAMPLE

- Whenever we hear the term ‘population,’ the first thing that strikes our mind is a large group of people.

- The term is often contrasted with the sample, which is nothing but a part of the population that is so selected to represent the entire group.

Population vs Sample



Census and Survey

Census: Gathering data from the whole **population** of interest.

For example, elections, 10-year census, etc.

Survey: Gathering data from the **sample** in order to make conclusions about the population.

For example, opinion polls, quality control checks in manufacturing units, etc.

PARAMETER AND STATISTIC

- Parameters are numbers that summarize data for an entire population.
- Statistics are numbers that summarize data from a sample, i.e. some subset of the entire population.

	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Standard deviation	s	sigma
Variance	s^2	sigma^2

POPULATION

PARAMETERS

Measures used to describe the population are called **parameters**

STATISTICS

Measures computed from sample data are called **statistics**.

SAMPLE

Identify each of the following data sets as either a population or a sample:

1. The grade point averages (GPAs) of all students at a college.

2. The GPAs of a randomly selected group of students on a college campus.

3. The ages of the nine Supreme Court Justices of the United States on January1,1842January1,1842.

4. The gender of every second customer who enters a movie theater.

5. The lengths of Atlantic croakers caught on a fishing trip to the beach.

Solutions

1. Population.
2. Sample.
3. Population.
4. Sample.
5. Sample.

Data and Information

- Data is a raw and unorganized fact that required to be processed to make it meaningful. Data can be simple at the same time unorganized unless it is organized.

- Information is a set of data which is processed in a meaningful way according to the given requirement. Information is processed, structured, or presented in a given context to make it meaningful and useful.

Why DataMatters

- Helps us understand things as they are:

"What relationships if any exist between two events?"

"Do people who eat an apple a day enjoy fewer doctor's visits than those who don't?"

Why DataMatters

- Helps us predict future behavior to guide business decisions:

  "Based on a user's click history which ad is more likely to bring them to our site?"

Visualizing Data

- Compare a table:

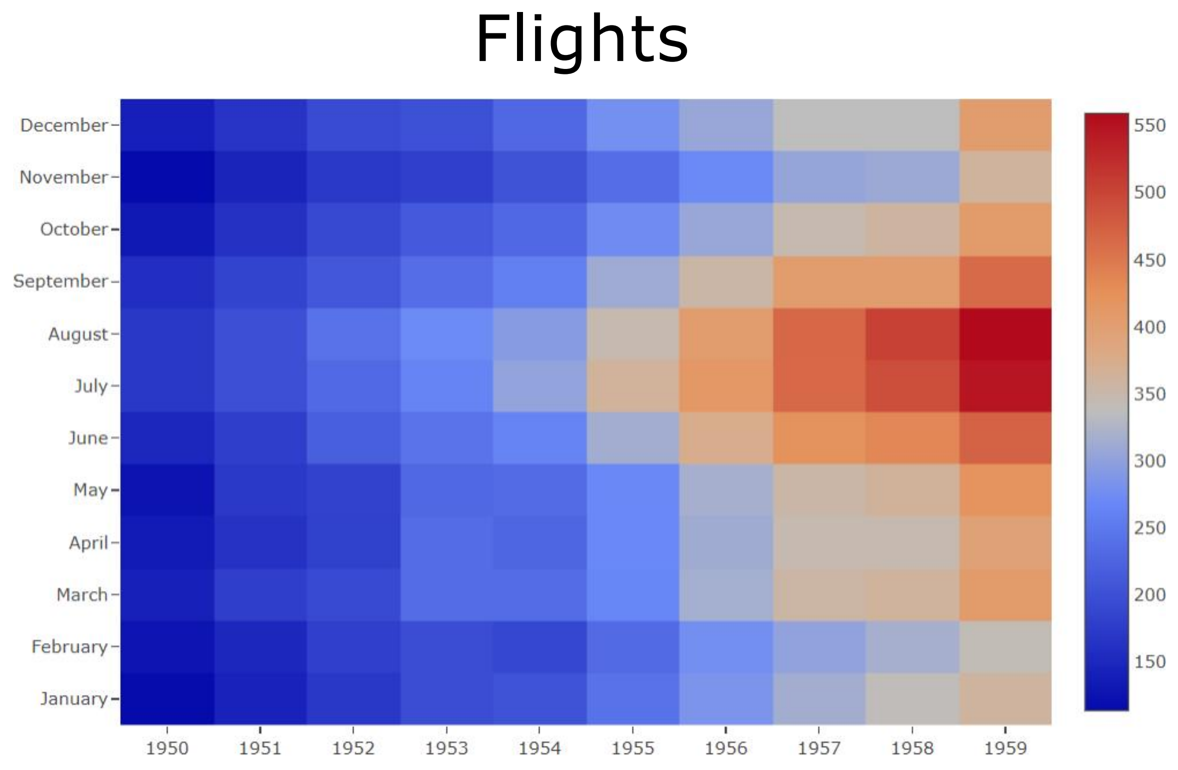
Flights

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	**year**	**month**	**passengers**		**year**	**month**	**passengers**		**year**	**month**	**passengers**		**year**	**month**	**passengers**
2	1950	January	115		1952	July	230		1955	January	242		1957	July	465
3	1950	February	126		1952	August	242		1955	February	233		1957	August	467
4	1950	March	141		1952	September	209		1955	March	267		1957	September	404
5	1950	April	135		1952	October	191		1955	April	269		1957	October	347
6	1950	May	125		1952	November	172		1955	May	270		1957	November	305
7	1950	June	149		1952	December	194		1955	June	315		1957	December	336
8	1950	July	170		1953	January	196		1955	July	364		1958	January	340
9	1950	August	170		1953	February	196		1955	August	347		1958	February	318
10	1950	September	158		1953	March	236		1955	September	312		1958	March	362
11	1950	October	133		1953	April	235		1955	October	274		1958	April	348
12	1950	November	114		1953	May	229		1955	November	237		1958	May	363
13	1950	December	140		1953	June	243		1955	December	278		1958	June	435
14	1951	January	145		1953	July	264		1956	January	284		1958	July	491
15	1951	February	150		1953	August	272		1956	February	277		1958	August	505
16	1951	March	178		1953	September	237		1956	March	317		1958	September	404
17	1951	April	163		1953	October	211		1956	April	313		1958	October	359
18	1951	May	172		1953	November	180		1956	May	318		1958	November	310

Not much can be gained by reading it.

Visualizing Data

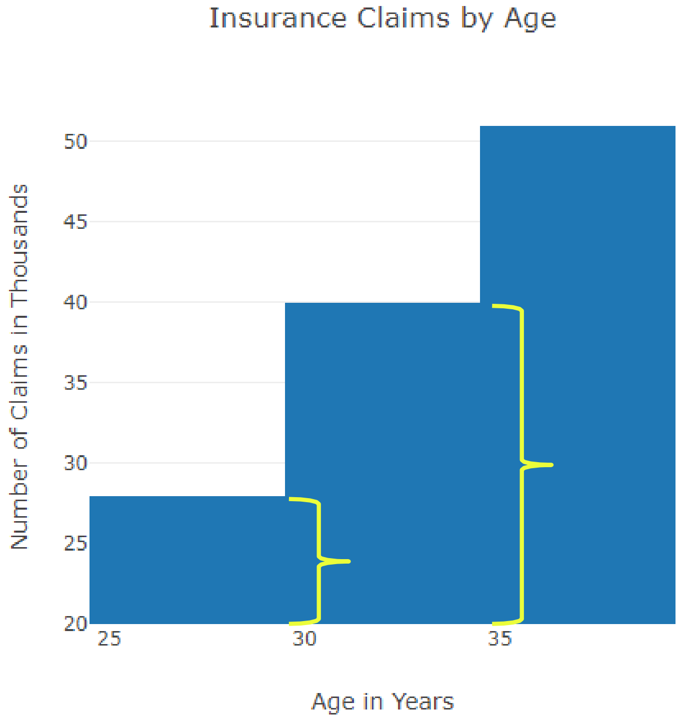
- to a **graph**:

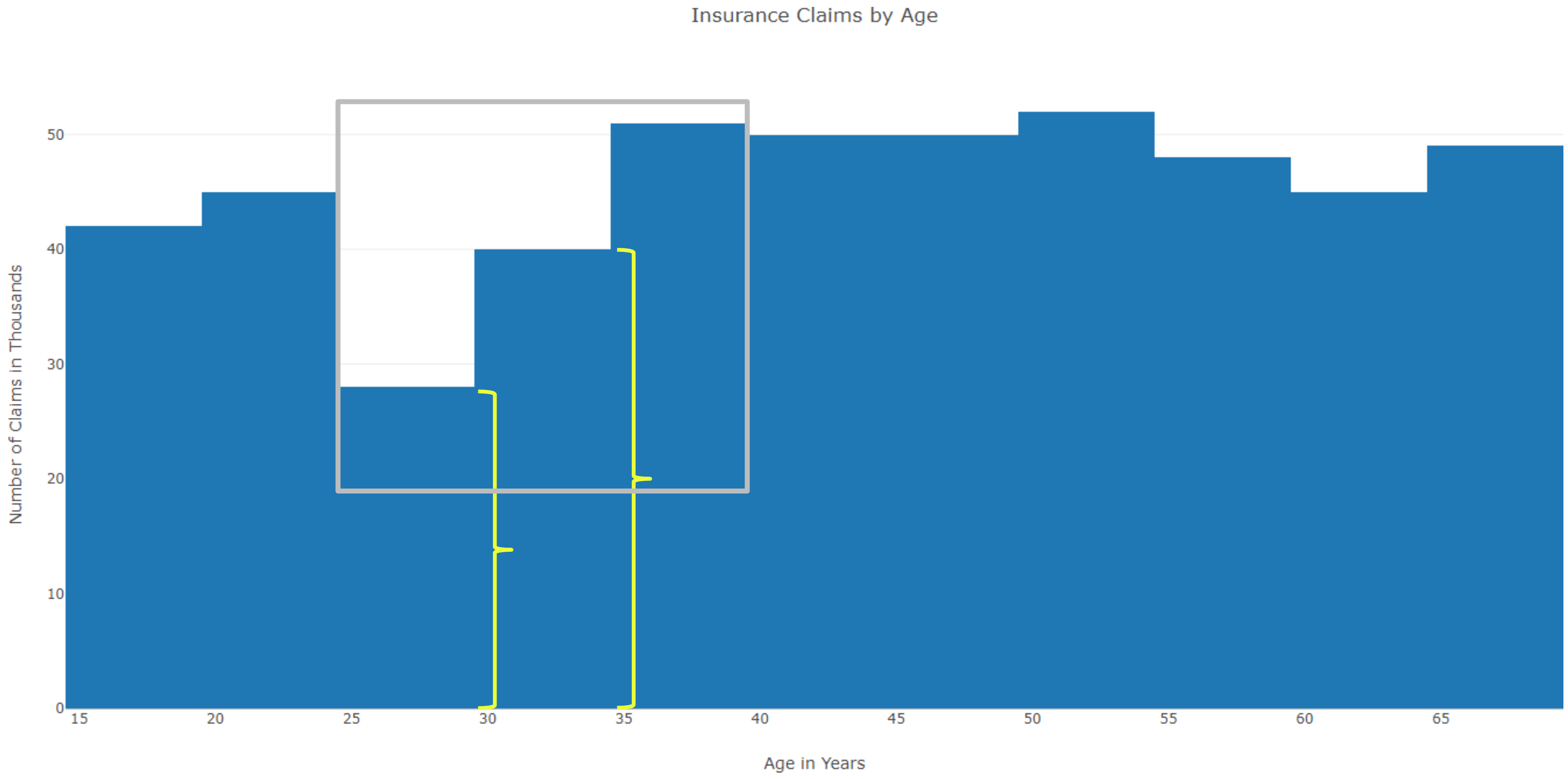


The graph uncovers two distinct trends - an increase in passengers flying over the years and a greater number of passengers flying in the summer months.

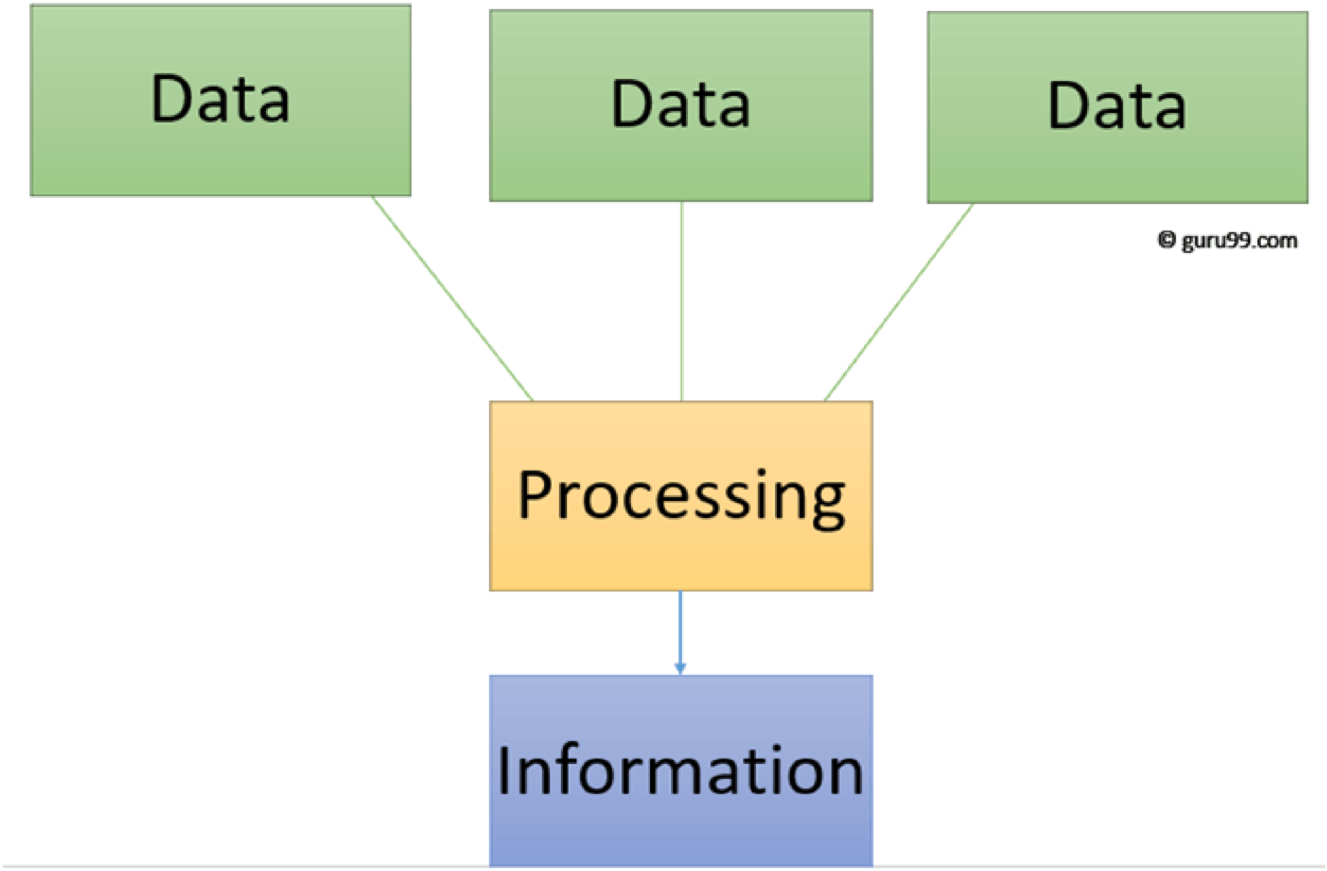
Analyze Visualizations Critically!

- Graphs can be misleading:

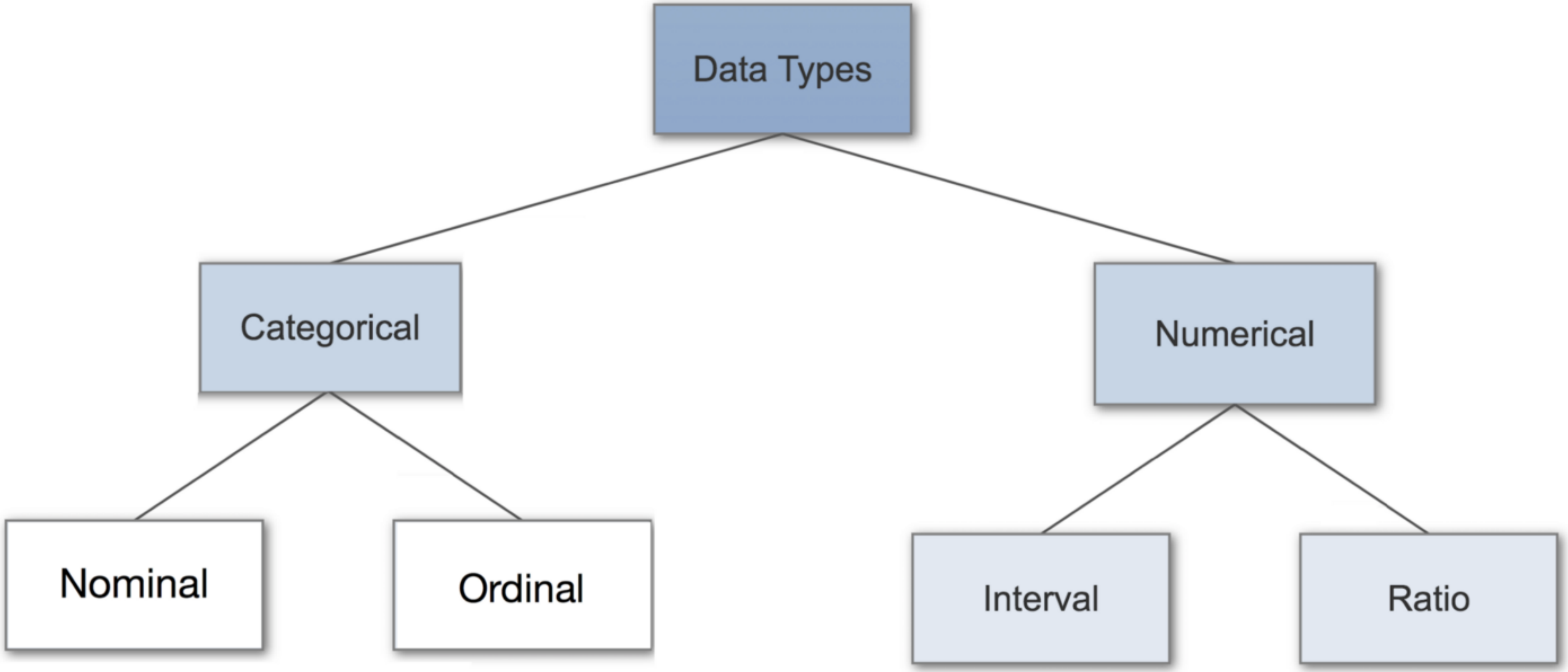




Data vs Information



TYPES OF VARIABLES



Measuring Data

Levels of Measurement

Nominal

- Predetermined categories
- Can’t be sorted

Animal classification (*mammal fish reptile*)

Political party (*republican democrat independent*)

Levels of Measurement

Ordinal

- Can be sorted
- Lacks scale

Survey responses

Levels of Measurement

Interval

- Provides scale
- Lacks a “zero” point
- Difference, Subtraction

Temperature

Levels of Measurement

Ratio

- Values have a true zero point
- Fractions, Divisions

Age, weight, salary

Numerical or Categorical?

Age	Gender	Major	Units	Housing	GPA
18	Male	Psychology	16	Dorm	3.6
21	Male	Nursing	15	Parents	3.1
20	Female	Business	16	Apartment	2.8

- Numerical
- Categorical

Numerical or Categorical?

Age	Gender	Major	Units	Housing	GPA
18	Male	Psychology	16	Dorm	3.6
21	Male	Nursing	15	Parents	3.1
20	Female	Business	16	Apartment	2.8

- Numerical
  - Age
  - Units
  - GPA

- Categorical
  - Gender
  - Major
  - Housing

Mathematical Symbols& Syntax

Symbol/Expression	Spoken as	Description
x^2	x squared	x raised to the second power $x^2 = x \times x$
x_i	x-sub-i	a subscripted variable (the subscript acts as a label)
$x!$	x factorial	$4! \; = \; 4 \times 3 \times 2 \times 1$
$\bar{x}$	x bar	symbol for the sample mean
μ	"mew"	symbol for the population mean (Greek lowercase letter mu)
Σ	sigma	syntax for writing sums (Greek capital letter sigma)

Exponents

$$\boldsymbol{x}^{\mathbf{5}} = x \times x \times x \times x \times x$$

1 2 3 4 5

EXAMPLE: $3^4 = 3 \times 3 \times 3 \times 3 = 81$

Exponents –specialcases

$$x^{-3} = \frac{1}{x \times x \times x}$$

EXAMPLE: $2^{-3} = \frac{1}{2\times2\times2} = \frac{1}{8} = 0.125$

$$x^{\left(\frac{1}{n}\right)} = \sqrt[n]{x}$$

EXAMPLE: $8^{\left(\frac{1}{3}\right)} = \sqrt[3]{8} = 2$

Factorials

$$\boldsymbol{x}! = x \times (x-1) \times (x-2) \times \cdots \times 1$$

EXAMPLE: $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$

EXAMPLE: $\frac{5!}{3!} = \frac{5\times4\times\cancel{3\times2\times1}}{\cancel{3\times2\times1}} = 5 \times 4 = 20$

Simple Sums

$$\sum_{x=1}^{n} x = 1 + 2 + 3 + \cdots + n$$

EXAMPLE: $\sum_{x=1}^{4} x = 1 + 2 + 3 + 4 = 10$

EXAMPLE: $\sum_{x=1}^{4} x^2 = 1 + 4 + 9 + 16 = 30$

Series Sums

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n$$

EXAMPLE:

$$x = \{5,3,2,8\}$$

$$n = \#\ elements\ in\ x = 4$$

$$\sum_{i=1}^{4} x_i = 5 + 3 + 2 + 8 = 18$$

Equation Example

- Formula for calculating a sample mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Read out loud:

"$\boldsymbol{x}$ bar (the symbol for the sample mean) is equal to the sum (indicated by the Greek letter sigma) of all the $\boldsymbol{x}$-sub-$\boldsymbol{i}$ values in the series as $\boldsymbol{i}$ goes from 1 to the number $\boldsymbol{n}$ items in the series divided by $\boldsymbol{n}$."

Equation Example

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

1. Start with a series of values:

   {7 8 9 10}

2. Assign placeholders to each item

   {7 8 9 10}

   1 2 3 4 n=4

3. These become x_1 x_2 etc.

   $x_1 = 7$ $x_2 = 8$ $x_3 = 9$ $x_4 = 10$

Equation Example

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

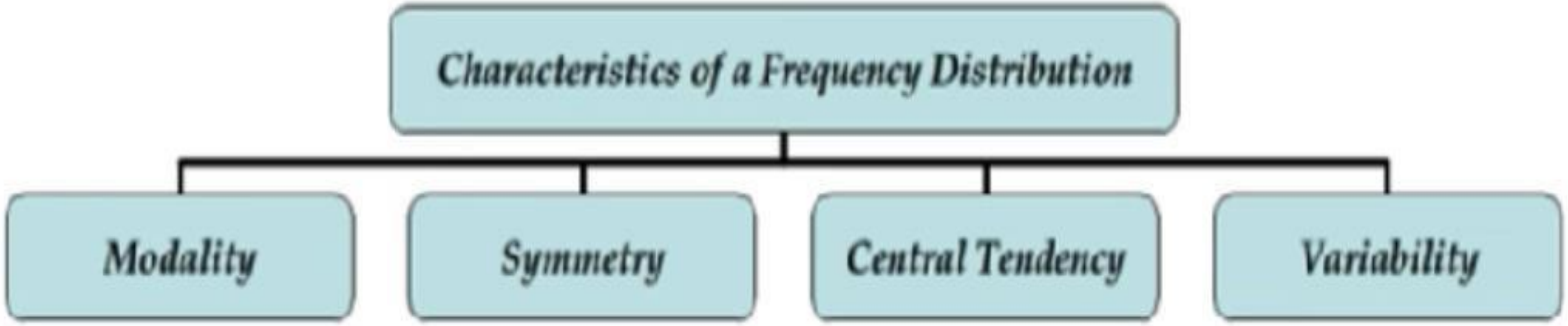
4. Plug these into the equation:

The empirical rule tells you what percentage of your data falls within a certain number of standard deviations from the mean:

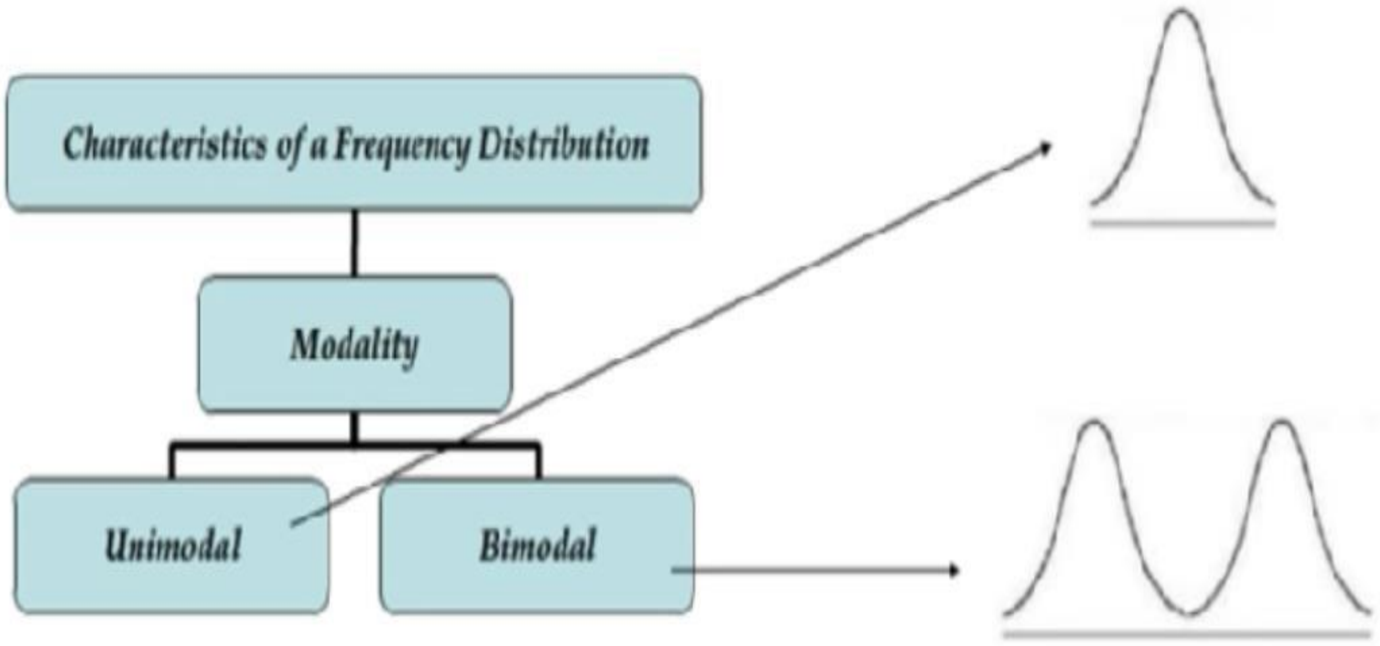
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 \ldots + x_n}{n}$$

$$= \frac{7 + 8 + 9 + 10}{4} = \frac{34}{4} = 8.5$$

Summarizing Data

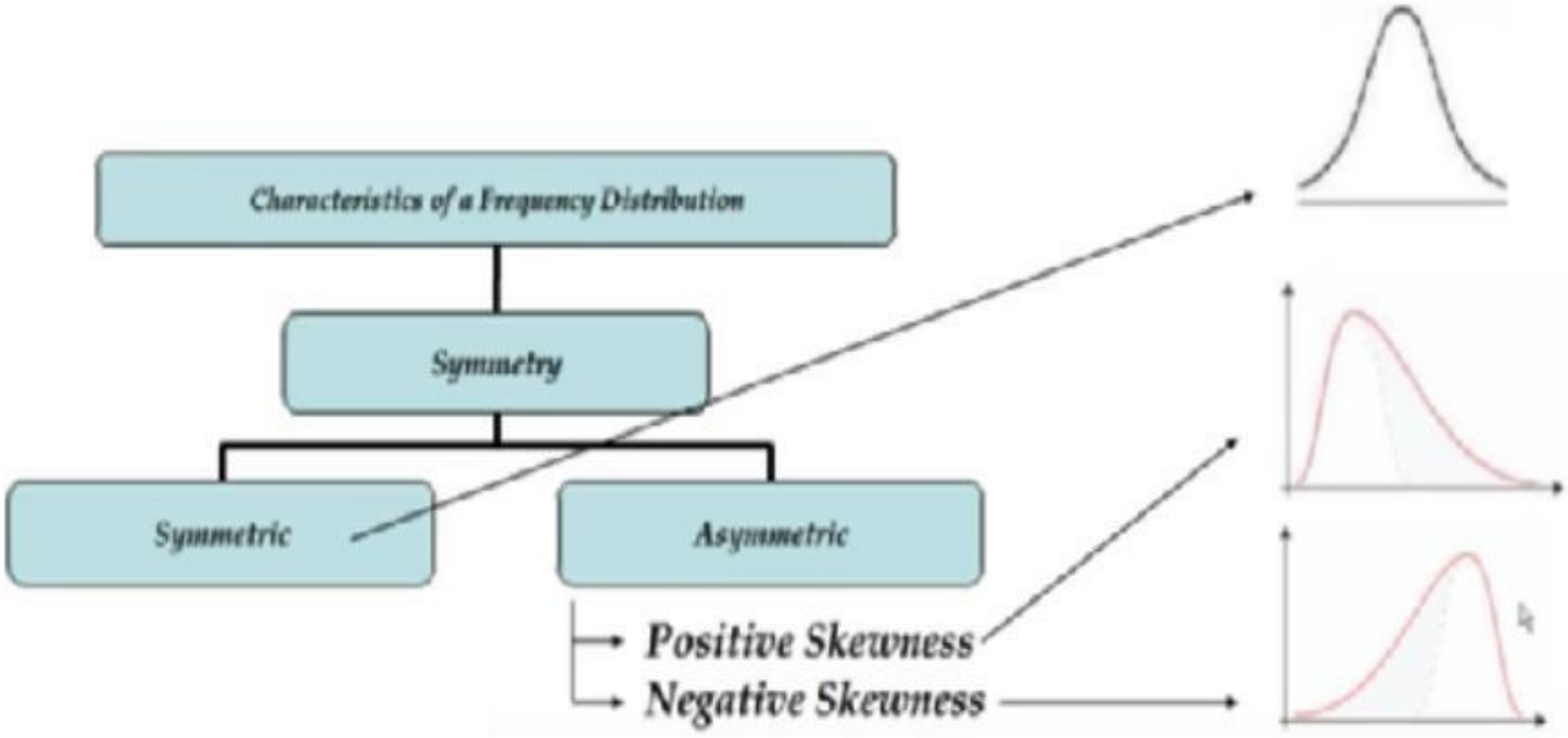


Modality



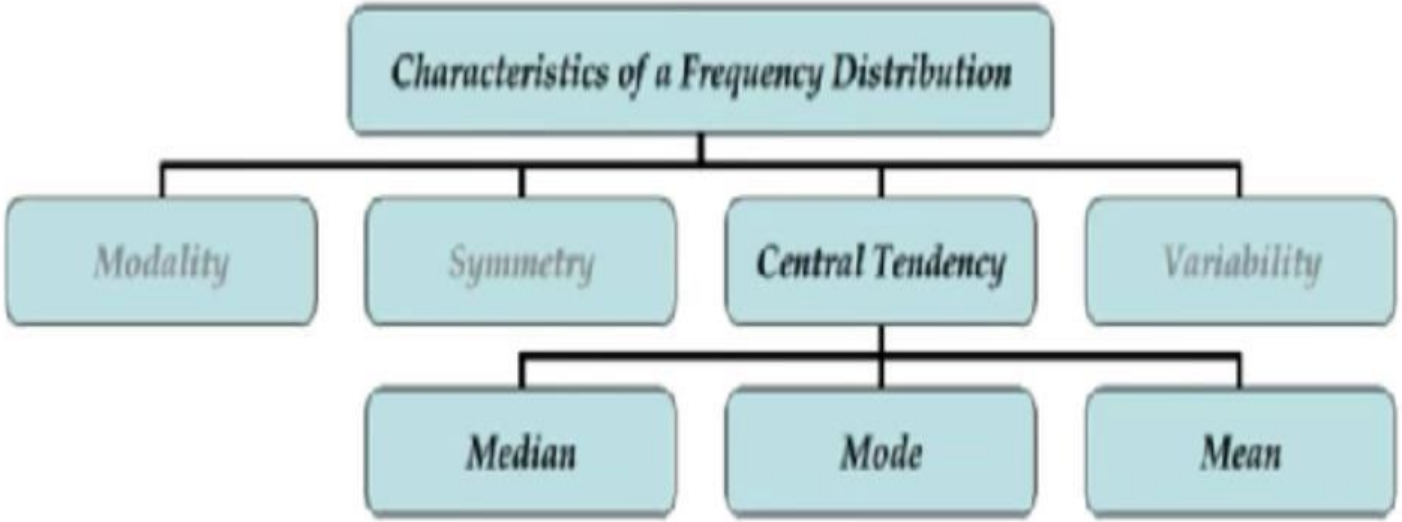
Characteristics of a Frequency Distribution
Modality
Unimodal
Bimodal

Symmetry

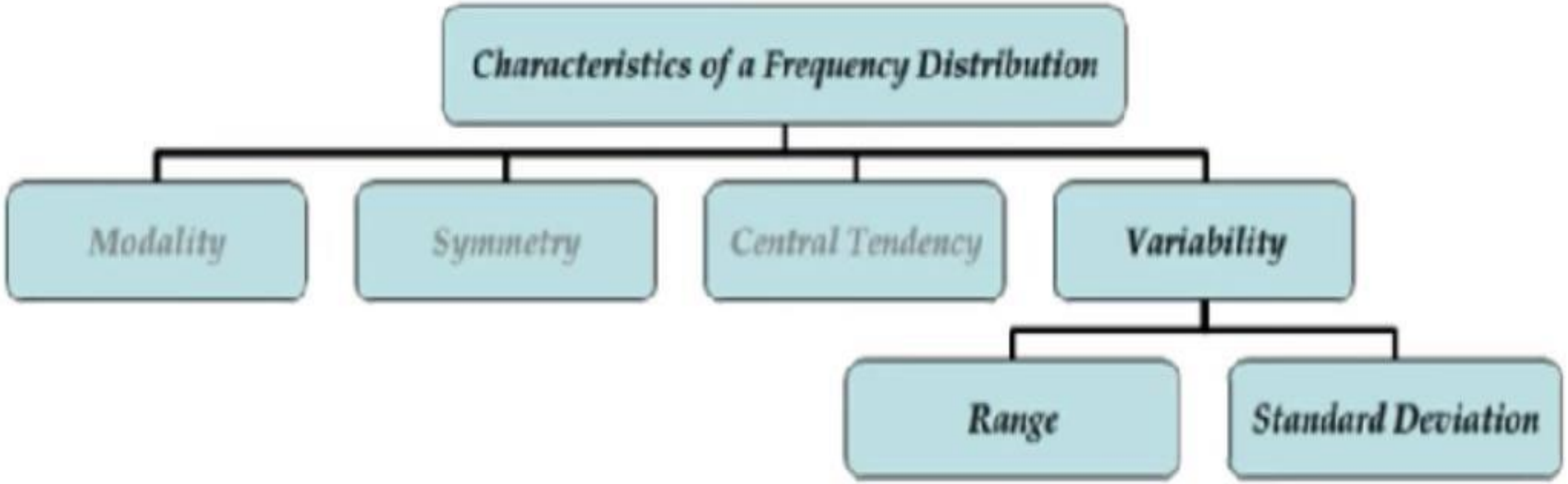


Characteristics of a Frequency Distribution
Symmetry
Symmetric
Asymmetric
Positive Skewness
Negative Skewness

Central Tendency



Variability



Measurements of Data

- “What was the average return?”
  Measures of Central Tendency

- “How far from the average did individual values stray?”
  Measures of Dispersion

Measurement of Central Tendency

Measures of Central Tendency (mean, median, mode)

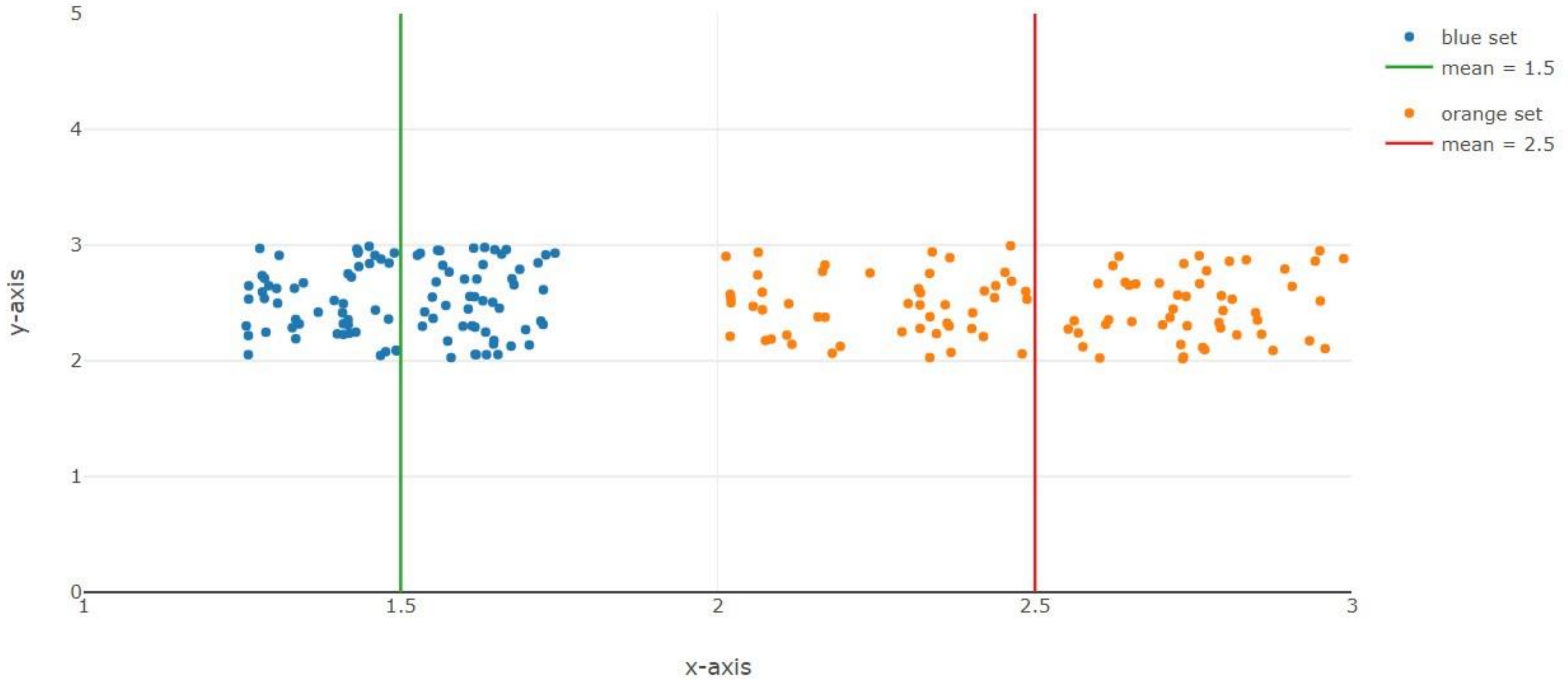
- Describe the “location” of the data
- Fail to describe the “shape” of the data

mean =“calculated average”

median =“middle value”

mode =“most occurring value”

Mean



- Shows “location” but not “how spread out”

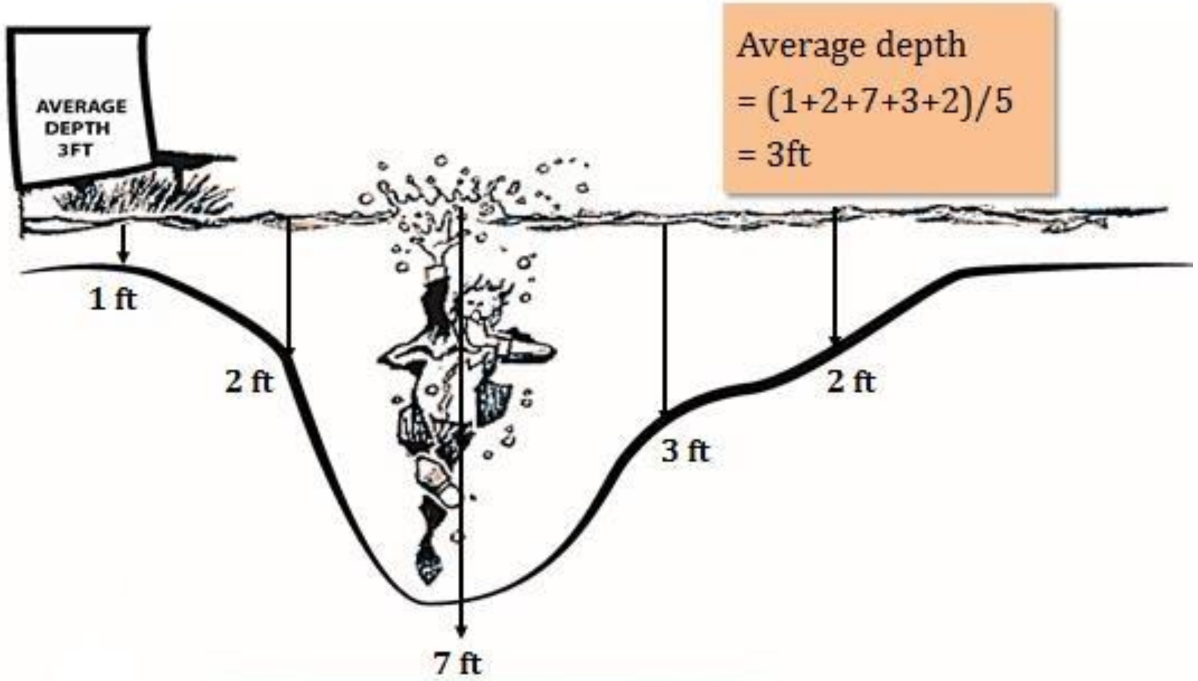
Alan went for a trek. On the way, he had to cross a stream. As Alan did not know swimming, he started exploring alternate routes to cross over.

Suddenly he saw a sign-post, which said "Average depth 3 feet". Alan was 5'7" tall and thought he could safely cross the stream.

Alan never reached the other end and drowned in the stream.

Why did Alan Drown?

Why did Alan Drown?



Beware of Averages!!!

The "Hotshot" Sales Executive

Kurt works as a sales manager at vsellhomes.com. In the monthly sales review, Kurt reports that he will achieve his quarterly target of $1M.

Kurt claims his average deal size is $100,000 and he has 10 deals in his pipeline. Kurt's boss Ross is very delighted with his numbers.

At the end of quarter, even after closing 8 deals Kurt fails to meet his target number and falls short by more than $500,000.

Discussion


Why did Kurt fail to achieve his quarterly target?
With 10 deals in pipeline and with average deal size of $100,000 and converting 7 of those deals, how did he fail?

The Reality of the "Hotshot" Salesman

- Average deal size in pipeline

  = $100,000

Deal #	Deal Value	Deal Status
1	70,000	Open
2	50,000	Closed
3	55,000	Closed
4	60,000	Closed
5	55,000	Closed
6	50,000	Closed
7	50,000	Closed
8	60,000	Closed
9	50,000	Closed
10	5,00,000	Open

The Reality of the "Hotshot" Salesman

- Average deal size in pipeline = $100,000
- Deal #10 is of significantly higher value than all the other deals and impacts the average calculation

Deal #	Deal Value	Deal Status
1	70,000	Open
2	50,000	Closed
3	55,000	Closed
4	60,000	Closed
5	55,000	Closed
6	50,000	Closed
7	50,000	Closed
8	60,000	Closed
9	50,000	Closed
10	5,00,000	Open

Median – *odd number of values*

10 10 11 13 15 16 19 19 21 23 28 30 33 34 36 44

= 19

Median - *even number of values*

10 10 11 13 15 16 19 19 21 23 28 30 33 34 36 44

$$\frac{19 + 21}{2} = 20$$

The Reality of the “Hotshot” Salesman

- Average deal size in pipeline
  = $100,000
- Deal #10 is of significantly higher value than all the other deals and impacts the average calculation
- Median = $55,000 more realistic measure

Deal #	Deal Value	Deal Status
1	70,000	Open
2	50,000	Closed
3	55,000	Closed
4	60,000	Closed
5	55,000	Closed
6	50,000	Closed
7	50,000	Closed
8	60,000	Closed
9	50,000	Closed
10	5,00,000	Open

The Reality of the “Hotshot” Salesman

- Average deal size in pipeline
  = $100,000
- Deal #10 is of significantly higher value than all the other deals and impacts the average calculation
- Median = $55,000 more realistic measure

Deal #	Deal Value	Deal Status
1	70,000	Open
2	50,000	Closed
3	55,000	Closed
4	60,000	Closed
5	55,000	Closed
6	50,000	Closed
7	50,000	Closed
8	60,000	Closed
9	50,000	Closed
10	5,00,000	Open

Median is less susceptible to the influence of Outliers.

Mean vs.Median

- The mean can be influenced by *outliers*.
- The mean of {2,3,2,3,2,12} is 4
- The median is 2.5
- The median is much closer to most of the values in the series!

Mode

10 10 11 13 15 16 16 16 21 23 28 30 33 34 36 44

= 16

Central Tendency: Example

- Timing for the Men's 500-meter Speed Skating event in Winter Olympics is tabulated.
- The Central Tendency measures are computed below:

Year	Time
1928	43.4
1932	43.4
1936	43.4
1948	43.1
1952	43.2
1956	40.2
1960	40.2
1964	40.1
1968	40.3
1972	39.44
1976	39.17
1980	38.03
1984	38.19
1988	36.4

Mean
= (43.4+…+36.4)/14
= 568.53/14
= 40.61

Year	Time
1988	36.4
1980	38.03
1984	38.19
1976	39.17
1972	39.44
1964	40.1
1956	40.2
1960	40.2
1968	40.3
1948	43.1
1952	43.2
1928	43.4
1932	43.4
1936	43.4

Median
= (7^{th} + 8^{th} Value)/2
= (40.2+40.2)/2
= 40.2

Year	Time
36.4	1
38.03	1
38.19	1
39.17	1
39.44	1
40.1	1
40.2	2
40.3	1
43.1	1
43.2	1
43.4	3

Mode
= Value with highest frequency
= 43.4

Player_A Vs Player_B – Who is Better ?

Match	Player A	Player B
1	40	40
2	40	35
3	7	45
4	40	52
5	0	30
6	90	40
7	3	29
8	11	43
9	120	37

Player_A Vs Player_B – Who is Better ?

Match	Player A	Player B
1	40	40
2	40	35
3	7	45
4	40	52
5	0	30
6	90	40
7	3	29
8	11	43
9	120	37
SUM	351	351

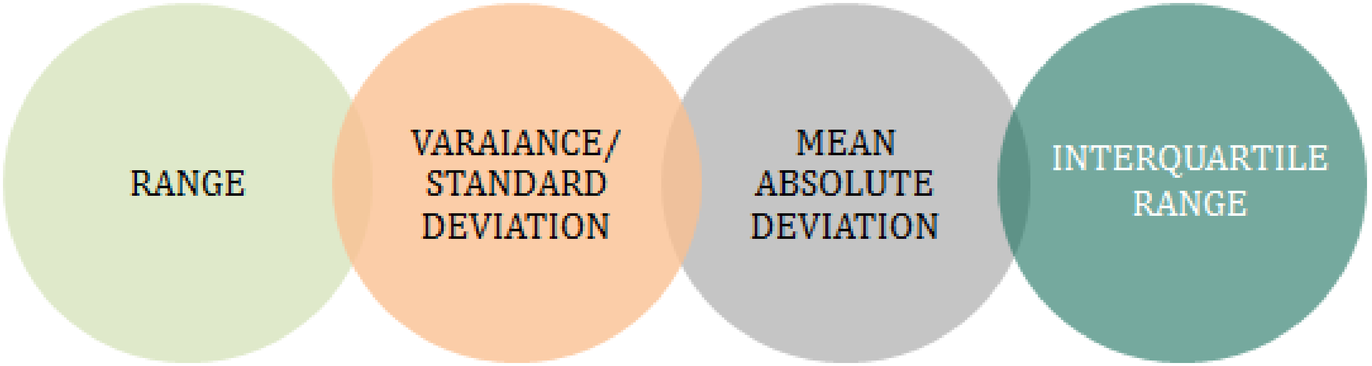
Player_A Vs Player_B – Who is Better ?

Match	Player A	Player B
1	40	40
2	40	35
3	7	45
4	40	52
5	0	30
6	90	40
7	3	29
8	11	43
9	120	37
SUM	351	351
MEAN	39	39

Player_A Vs Player_B – Who is Better ?

Match	Player A	Player B
1	40	40
2	40	35
3	7	45
4	40	52
5	0	30
6	90	40
7	3	29
8	11	43
9	120	37
SUM	351	351
MEAN	39	39
MEDIAN	40	40

Measures of Dispersion



Measures of Dispersion (range, variance, standard deviation)

9 10 11 13 15 16 19 19 21 23 28 30 33 34 36 39

- In this sample the mean is 22.25
- How do we describe how “spread out” the sample is?

Range

9 10 11 13 15 16 19 19 21 23 28 30 33 34 36 39

$$Range = max - min$$

$$= 39 - 9$$

$$= 30$$

Variance

- Calculated as the sum of square distances from each point to the mean
- There’s a difference between the SAMPLE variance and the POPULATION variance
- subject to Bessel's correction ($\boldsymbol{n-1}$)

Variance

SAMPLE VARIANCE: $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$

POPULATION VARIANCE: $\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$

Sample Variance

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

4 7 9 8 11

$$\bar{x} = \frac{4+7+9+8+11}{5} = \frac{39}{5} = 7.8$$

sample mean

$$s^2 = \frac{(4-7.8)^2+(7-7.8)^2+(9-7.8)^2+(8-7.8)^2+(11-7.8)^2}{5-1}$$

$$= 6.7$$

sample variance

Standard Deviation

- square root of the variance
- benefit: same units as the sample
- meaningful to talk about
  “*values that lie within*
  one standard deviation
  of the mean”

Sample Standard Deviation

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

Sample:

4	7	9	8	11

$$\bar{x} = \frac{4+7+9+8+11}{5} = \frac{39}{5} = 7.8$$

sample mean

$$s = \sqrt{\frac{(4-7.8)^2 + (7-7.8)^2 + (9-7.8)^2 + (8-7.8)^2 + (11-7.8)^2}{5-1}}$$

$$= \sqrt{6.7} = 2.59$$

sample standard deviation

Population Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

Population:

4	7	9	8	11

$$\mu = \frac{4+7+9+8+11}{5} = \frac{39}{5} = 7.8$$

population mean

$$\sigma = \sqrt{\frac{(4-7.8)^2+(7-7.8)^2+(9-7.8)^2+(8-7.8)^2+(11-7.8)^2}{5}}$$

$$= \sqrt{5.36} = 2.32$$

population standard deviation

Who's Best ?

Match	Player A	Player B
1	40	40
2	40	35
3	7	45
4	40	52
5	0	30
6	90	40
7	3	29
8	11	43
9	120	37
SUM	351	351
MEAN	39	39
MEDIAN	40	40
STANDARD DEVIATION	41.5180683558376	7.28010988928052

Measuring Variability and Spread

Basketball coach Statson is in a dilemma choosing between 3 players all having the same average scores.

Points scored per game	7	8	9	10	11	12	13
Frequency, f	1	1	2	2	2	1	1

Points scored per game	7	9	10	11	13
Frequency, f	1	2	4	2	1

Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1

Measuring Variability and Spread

Basketball coach Statson is in a dilemma choosing between 3 players all having the same average scores.

Points scored per game	7	8	9	10	11	12	13
Frequency, f	1	1	2	2	2	1	1

Points scored per game	7	9	10	11	13
Frequency, f	1	2	4	2	1

Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1

Mean = Median = Mode = 10 for all 3.

Measuring Variability and Spread

Range = Max - Min

Points scored per game	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Frequency, f	1	1	2	2	2	1	1

Points scored per game	**7**	**9**	**10**	**11**	**13**
Frequency, f	1	2	4	2	1

Points scored per game	**3**	**6**	**7**	**10**	**11**	**13**	**30**
Frequency, f	2	1	2	3	1	1	1

Points scored per game	7	8	9	10	11	12	13
Frequency, f	1	1	2	2	2	1	1

Points scored per game	7	9	10	11	13
Frequency, f	1	2	4	2	1

Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1

MEAN = MEDIAN = MODE = 10 RANGE = 5 , 5 , 27

Points scored per game	7	8	9	10	11	12	13
Frequency, f	1	1	2	2	2	1	1

Points scored per game	7	9	10	11	13
Frequency, f	1	2	4	2	1

Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1

MEAN = MEDIAN = MODE = 10 RANGE = 5 , 5 , 27 Reject Player 3

Basketball coach Statson is in a dilemma choosing between 3 players all having the same average scores.

Points scored per game	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Frequency, f	1	1	2	2	2	1	1

Points scored per game	**7**	**9**	**10**	**11**	**13**
Frequency, f	1	2	4	2	1

STANDARD DEVIATION

Player 1 = **1.7873008824606**
Player 2 = **3.30823887354653**

What is your Decision?????????

Exercises

- Consider the following data: 3, 8, 4, 10, 6, 2.

1. Calculate its mean and variance.

2**.** If all the above data was multiplied by 3, what would the new mean and variance be?

Solution

x_i	x_i^2
2	4
3	9
4	16
6	36
8	64
10	100
33	**229**

1

$$\bar{x}_1 = \frac{33}{6} = 5.5$$

$$\sigma_1^2 = \frac{229}{6} - 5.5^2 = 7.92$$

2

$$\bar{x}_2 = 5.5 \cdot 3 = 16.5$$

$$\sigma_1^2 = 7.92 \cdot 3^2 = 71.28$$

Case Study

In an Under 19 World Cup selection squad for 2018 the BCCI needs to select 1 player based on the current performance in 2017 – 2018 Ranji Trophy. There are 2 players with similar stats and the board is not sure whom to select.

- Can you help the board members with your analysis ?

Coefficient of Variation:

- A coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. It represents the ratio of the standard deviation to the mean.
- It is also used as a measure of variability when the standard deviation is proportional to the mean, and as a means to compare variability of measurements made in different units.
- Less Coefficient of variance means less risk and more consistency.
- More coefficient of variance means more risk and less consistency.

Stats - Player X & Y

Runs scored by both players in last 14 matches

Player X	Player Y
40	35
20	40
5	7
20	23
10	20
75	26
100	12
25	30
15	27
15	102
20	18
17	17
11	14
5	7

Equation for Coefficient of Variation

CV for a population:

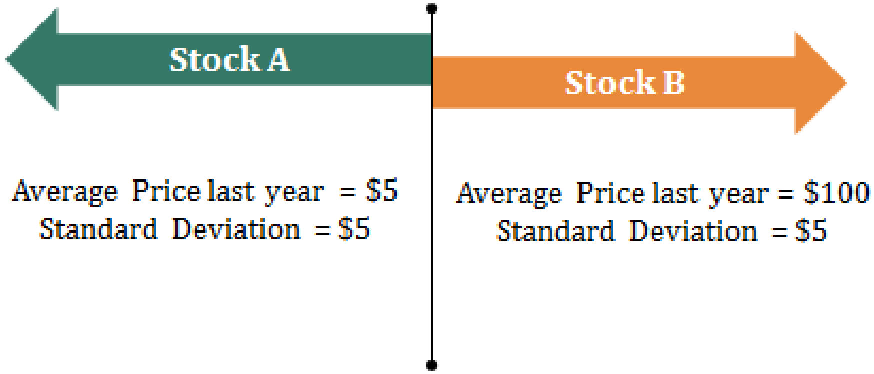
$$CV = \frac{\sigma}{\mu} * 100\%$$

CV for a sample:

$$CV = \frac{s}{\bar{x}} * 100\%$$

Coefficient of Variation

Coeff of Variation = (Standard deviation/ Mean) * 100 %



Coefficient of Variation:

Stock A: CV = 100%

(5/5*100=100%)

Stock B: CV = 5%

(5/100*100=5%)

$$CV = \left(\frac{S}{\overline{X}}\right) \cdot 100\%$$

Coefficient of Variation

Calculate the descriptive statistics of both players and if the coefficient of variation is greater than 85% then drop that player

Coeff of Variation = (Standard deviation/ Mean) * 100 %

Measurement Types Quartiles

Quartiles and IQR

- Another way to describe data is through quartiles and the interquartile range (IQR)
- Has the advantage that every data point is considered, not aggregated!

Percentile & Quartile

Nth percentile states that there are atleast N% of values less than or equal to this value and (100-N) values are greater or equal to this value

$i = (N/100)*n$

N – The percentile you are interested

n – Number of values

Key points

1. If i is decimal then round off to next value
2. If i is integer then take average of i and i+1 value

Let's calculate 85th percentile

Data:

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

Calculate 85th percentile ?

Quartile

Data:
3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730
3925

Quartile
Dividing data into ¼– 4 parts

Q1 – First Quartile – 25th percentile
Q2 – Second Quartile – 50th percentile (Median) Q3
– Third Quartile – 75th percentile

IQR (Inter Quartile Range) = Q3 – Q1

InterQuartile Range

Quartile

Dividing data into ¼– 4 parts

Q1 – First Quartile – 25th percentile
Q2 – Second Quartile – 50th percentile (Median) Q3 – Third Quartile – 75th percentile

IQR (Inter Quartile Range) = Q3 – Q1

Quartiles and IQR

- Consider the following series of 20 values:

9 10 10 11 13	15 16 19 19 21	23 28 30 33 34	36 44 45 47 60

1st quartile

2nd quartile or median

3rd quartile

1. Divide the series
2. Divide each subseries
3. These become quartiles

Quartiles and IQR

- Consider the following series of 20 values:

9 10 10 11 13	15 16 19 19 21	23 28 30 33 34	36 44 45 47 60

1st quartile — 2nd quartile or median — 3rd quartile

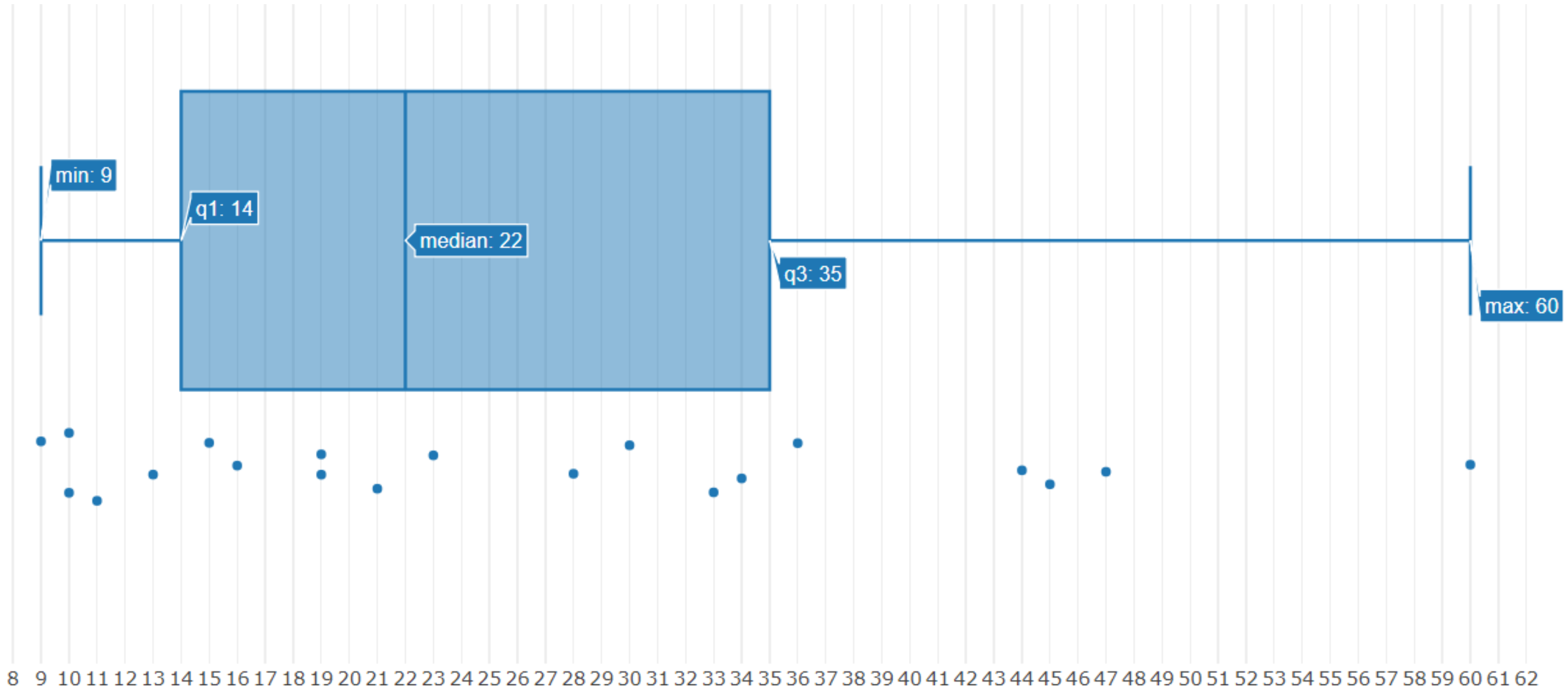
1st quartile = 14

2nd quartile = 22

3rd quartile =35

Plot theQuartiles

9 10 10 11 13	15 16 19 19 21	23 28 30 33 34	36 44 45 47 60

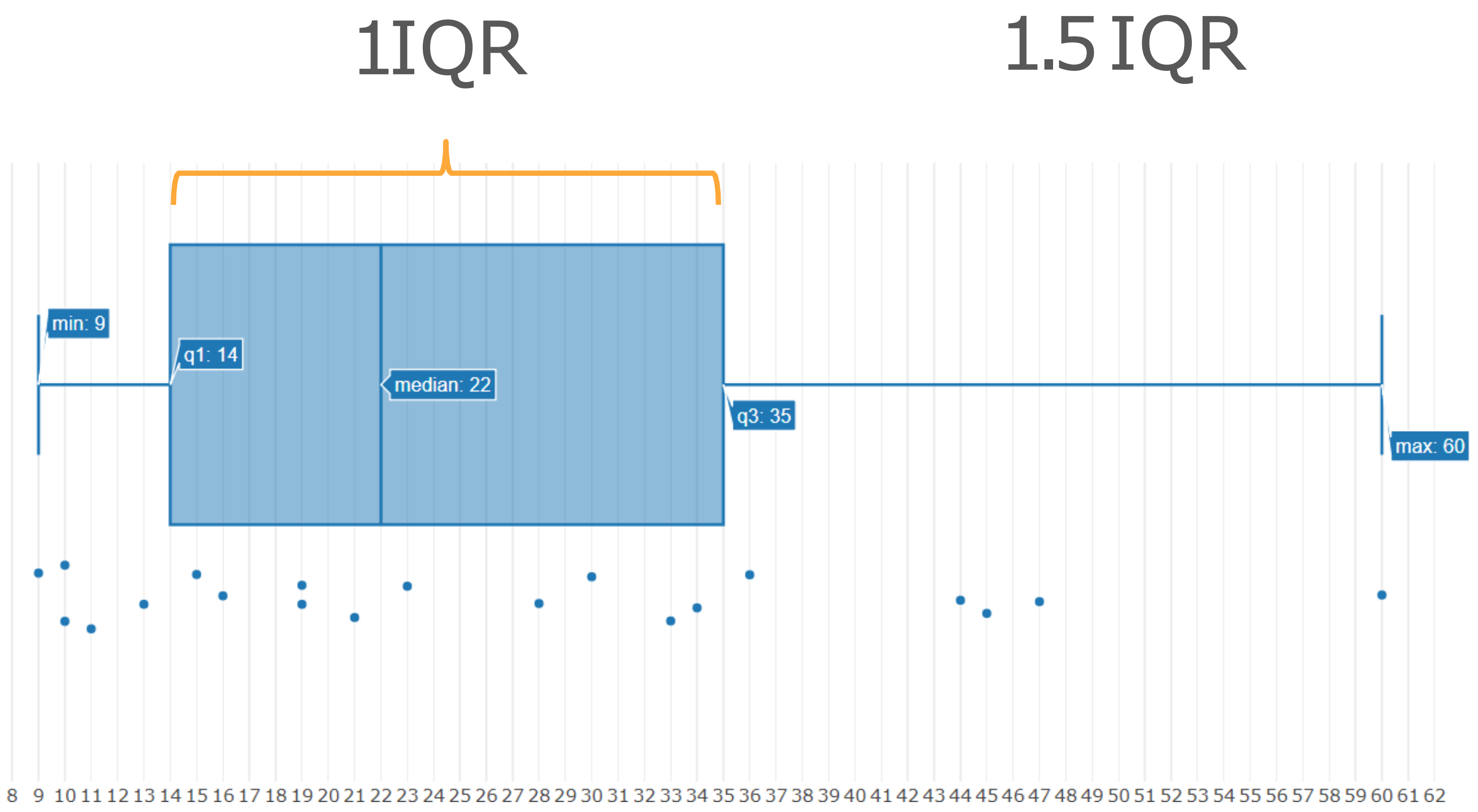


Quartile ranges are seldom the same size!

Fences & Outliers

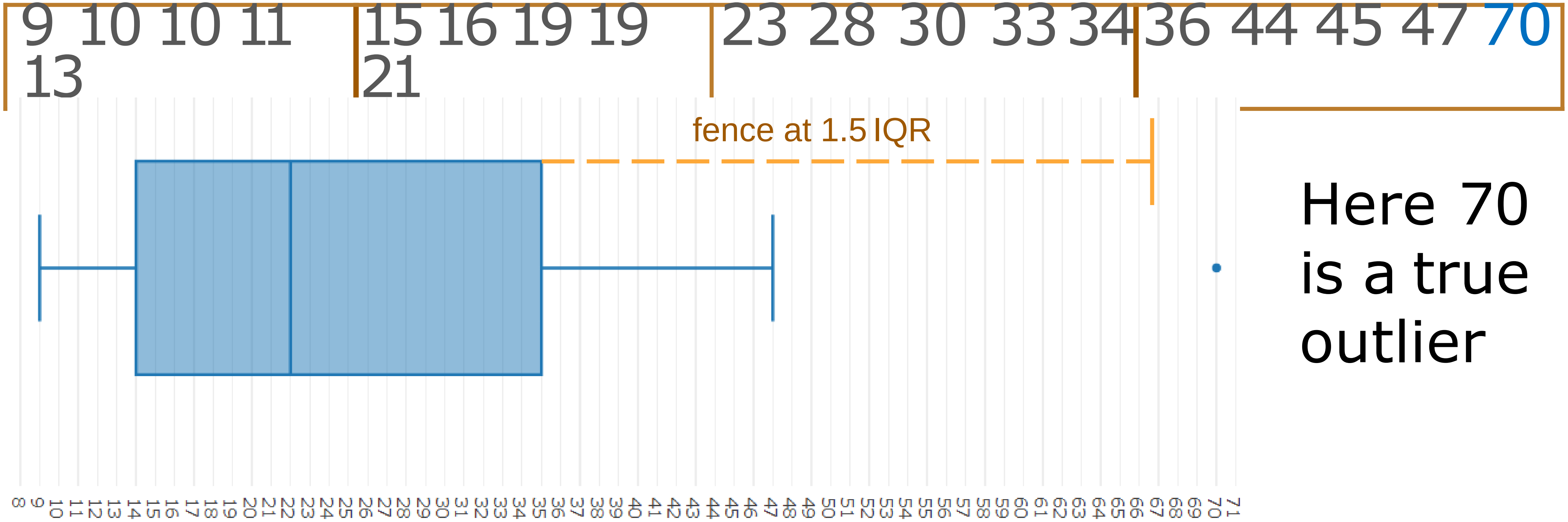
- What is considered an "outlier"?
- A common practice is to set a "fence" that is 1.5 times the width of the IQR
- Anything outside the fence is an outlier
- This is determined by the *data*, not an arbitrary percentage!

Fences &Outliers



In this set, 60 is *not* an outlier, but 70 would be

Fences &Outliers



Here 70 is a true outlier

- When drawing box plots, the whiskers are brought inward to the outermost values inside the fence.

Bivariate Data

Bivariate Data

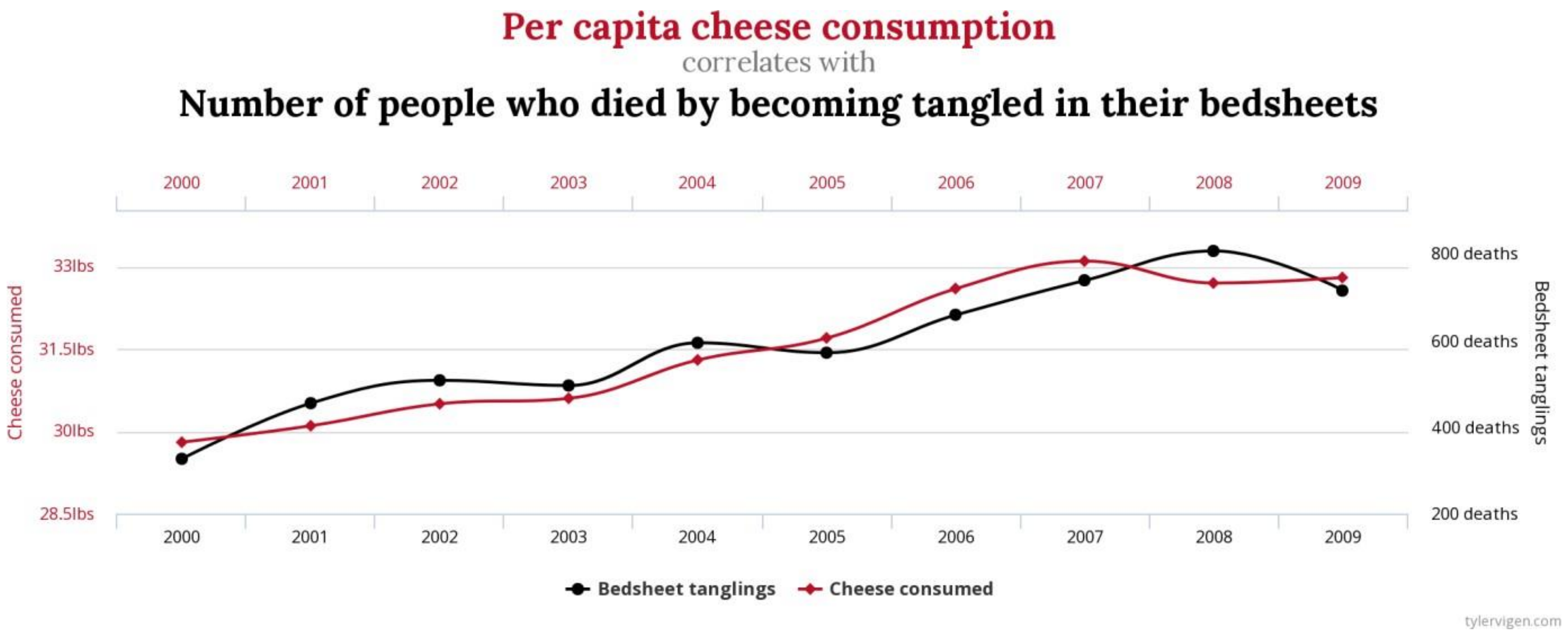
- Compares two variables
- By convention, the x-axis is set to the independent variable
- The y-axis is set to the dependent variable, or that which is being measured relative to x.

Bivariate Data

- Scatter plots may uncover a correlation between two variables
- They *can't* show causality!

Bivariate Data

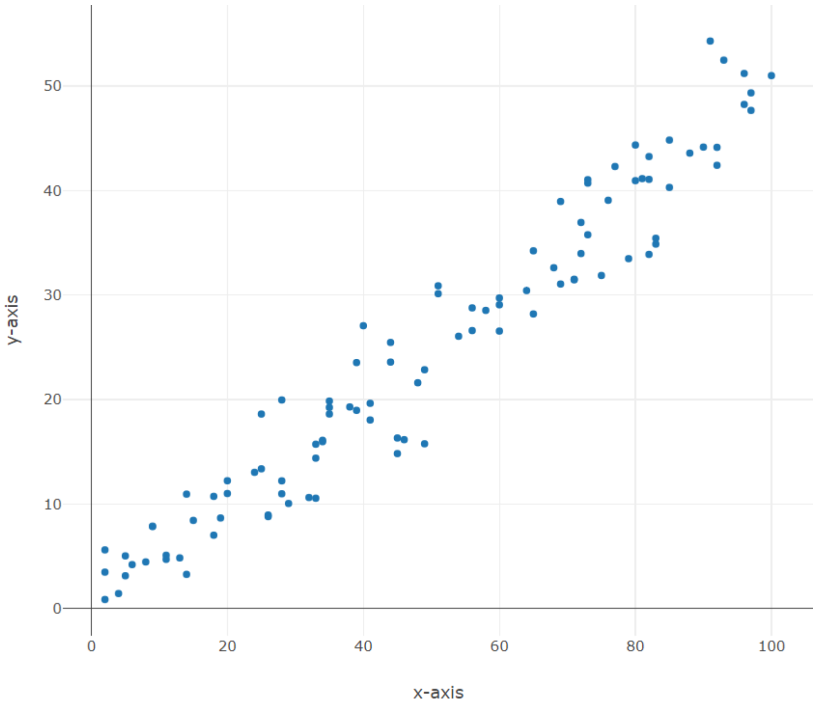
- Correlation between two variables
- Doesn't prove causality!



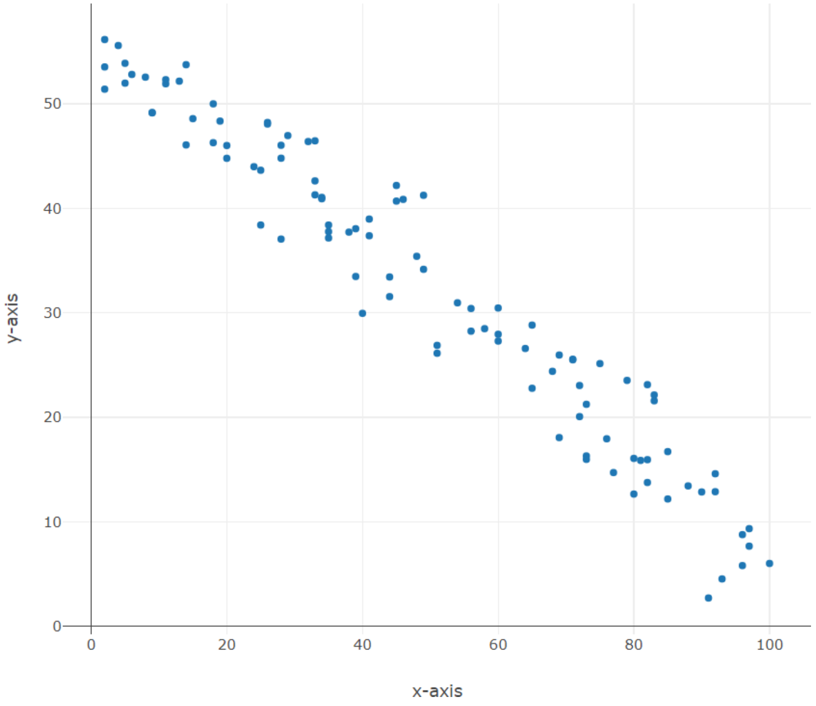
Bivariate Data

- More statistical analysis is needed to determine causality!
- For example: "Does increasing number of police officers decrease crime?"
- We would look at correlation, and do further analysis to understand causality.

Bivariate Data



Positive correlation



Negative or Inverse correlation

Covariance

- A common way to compare two variables is to compare their variances – how far from each item's mean do typical values fall?
- The first challenge is to match scale. Comparing height in inches to weight in pounds isn't meaningful unless we develop a standard score to normalize the data.

Covariance

- For simplicity, we'll consider the *population covariance:*

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Covariance Exercise

- Consider the following two tables:

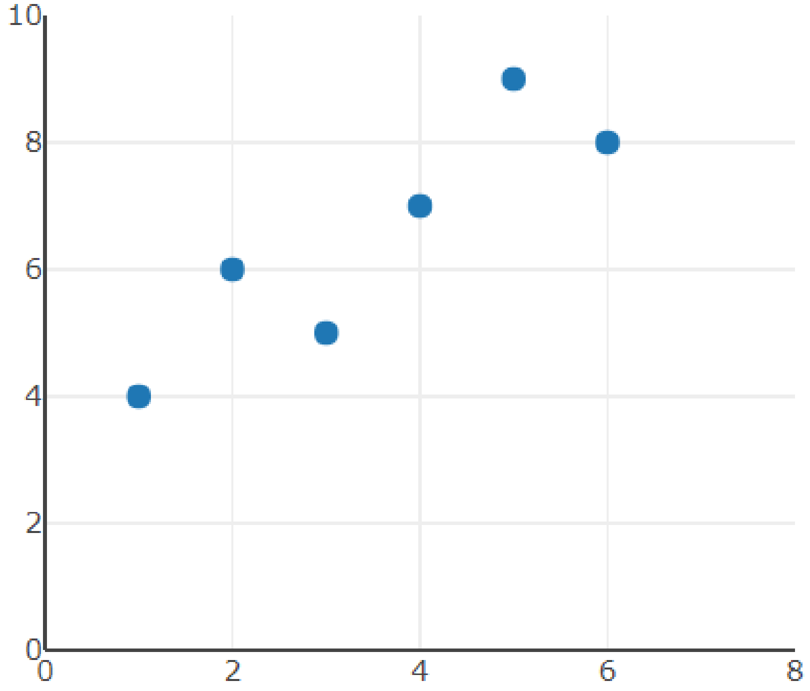
x	y
1	4
2	6
3	5
4	7
5	9
6	8

x	y
1	5
2	9
3	7
4	4
5	8
6	6

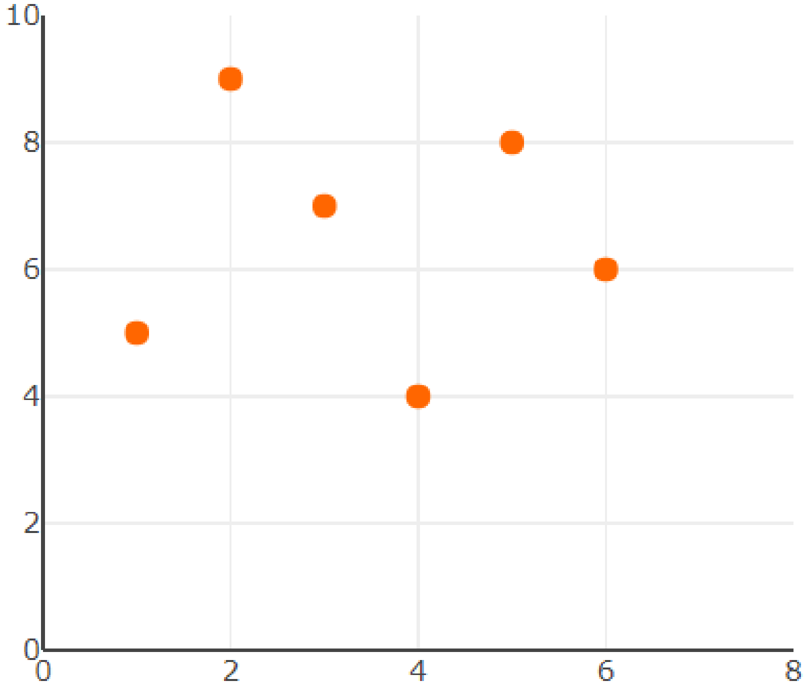
Covariance Exercise

- Plot them:

x	y
1	4
2	6
3	5
4	7
5	9
6	8



x	y
1	5
2	9
3	7
4	4
5	8
6	6



Covariance Exercise

$$\bar{x} = 3.5, \bar{y} = 6.5$$

- Calculate mean values:

x	y
1	4
2	6
3	5
4	7
5	9
6	8

$$\bar{x} = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\bar{y} = \frac{4+6+5+7+9+8}{6} = 6.5$$

x	y
1	5
2	9
3	7
4	4
5	8
6	6

$$\bar{x} = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\bar{y} = \frac{5+9+7+4+8+6}{6} = 6.5$$

Covariance Exercise

$\bar{x} = 3.5, \bar{y} = 6.5$

- Calculate $(x - \bar{x})$ and $(y - \bar{y})$:

x	y	(x - x̄)	(y - ȳ)
1	4	-2.5	-2.5
2	6	-1.5	-0.5
3	5	-0.5	-1.5
4	7	0.5	0.5
5	9	1.5	2.5
6	8	2.5	1.5

x	y	(x - x̄)	(y - ȳ)
1	5	-2.5	-1.5
2	9	-1.5	2.5
3	7	-0.5	0.5
4	4	0.5	-2.5
5	8	1.5	1.5
6	6	2.5	-0.5

Covariance Exercise

$\bar{x} = 3.5, \bar{y} = 6.5$

- Calculate $(x - \bar{x})(y - \bar{y})$:

x	y	(x - x̄)	(y - ȳ)	(x - x̄)(y - ȳ)
1	4	-2.5	-2.5	6.25
2	6	-1.5	-0.5	0.75
3	5	-0.5	-1.5	0.75
4	7	0.5	0.5	0.25
5	9	1.5	2.5	3.75
6	8	2.5	1.5	3.75

x	y	(x - x̄)	(y - ȳ)	(x - x̄)(y - ȳ)
1	5	-2.5	-1.5	3.75
2	9	-1.5	2.5	-3.75
3	7	-0.5	0.5	-0.25
4	4	0.5	-2.5	-1.25
5	8	1.5	1.5	2.25
6	6	2.5	-0.5	-1.25

Covariance Exercise

$\bar{x} = 3.5, \bar{y} = 6.5$

- Calculate sums:

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$
1	4	-2.5	-2.5	6.25
2	6	-1.5	-0.5	0.75
3	5	-0.5	-1.5	0.75
4	7	0.5	0.5	0.25
5	9	1.5	2.5	3.75
6	8	2.5	1.5	3.75
			Σ	**15.5**

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$
1	5	-2.5	-1.5	3.75
2	9	-1.5	2.5	-3.75
3	7	-0.5	0.5	-0.25
4	4	0.5	-2.5	-1.25
5	8	1.5	1.5	2.25
6	6	2.5	-0.5	-1.25
			Σ	

Covariance Exercise

$\bar{x} = 3.5, \bar{y} = 6.5$

- Calculate covariance:

x	y
1	4
2	6
3	5
4	7
5	9
6	8

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{15.5}{6} = \mathbf{2.583}$$

Σ	15.5

x	y
1	5
2	9
3	7
4	4
5	8
6	6

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

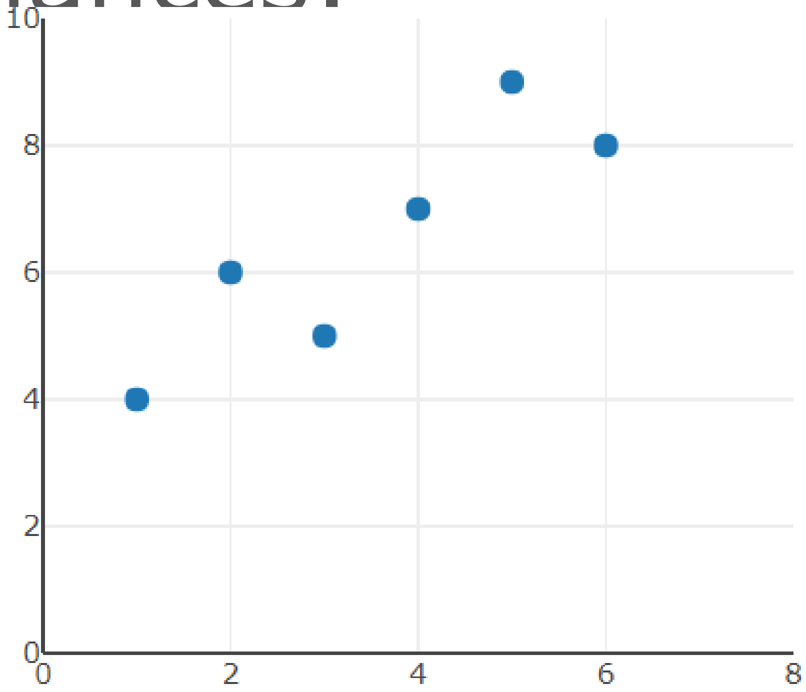
$$= \frac{-0.5}{6} = \mathbf{-0.083}$$

Σ	-0.5

Covariance Exercise

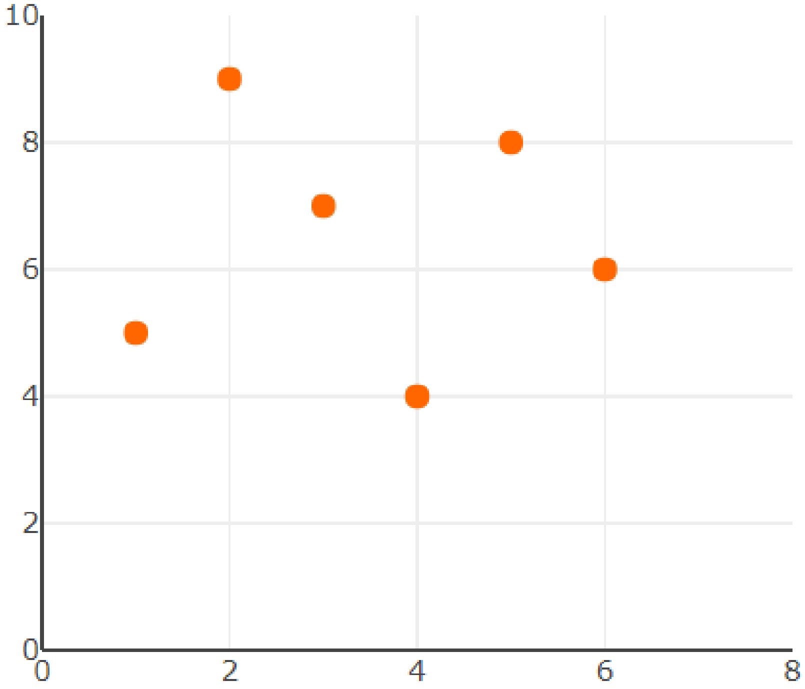
- Compare covariances:

x	y
1	4
2	6
3	5
4	7
5	9
6	8



cov(x,y) = 2.583

x	y
1	5
2	9
3	7
4	4
5	8
6	6



cov(x,y) = -0.083

Pearson Correlation Coefficient

Pearson Correlation Coefficient

- In order to normalize values coming from two different distributions, we use:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

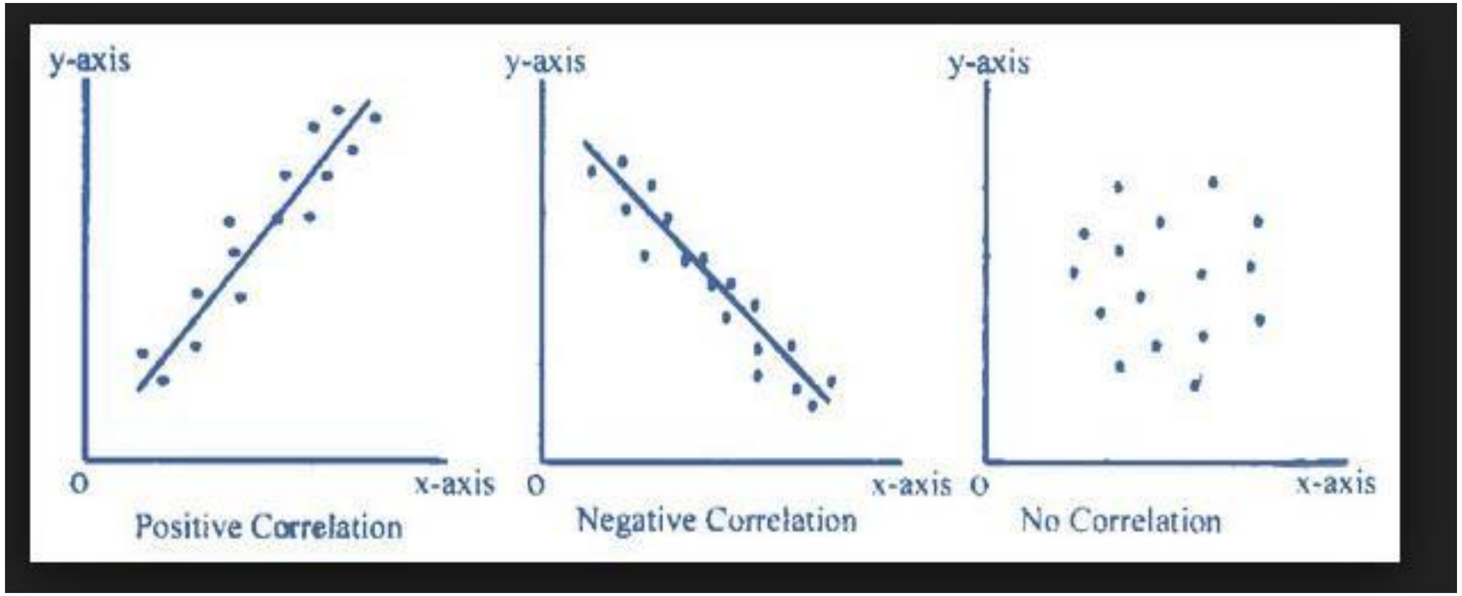
$\rho =$ Greek letter "rho"

$\sigma =$ standard deviation

$cov =$ covariance

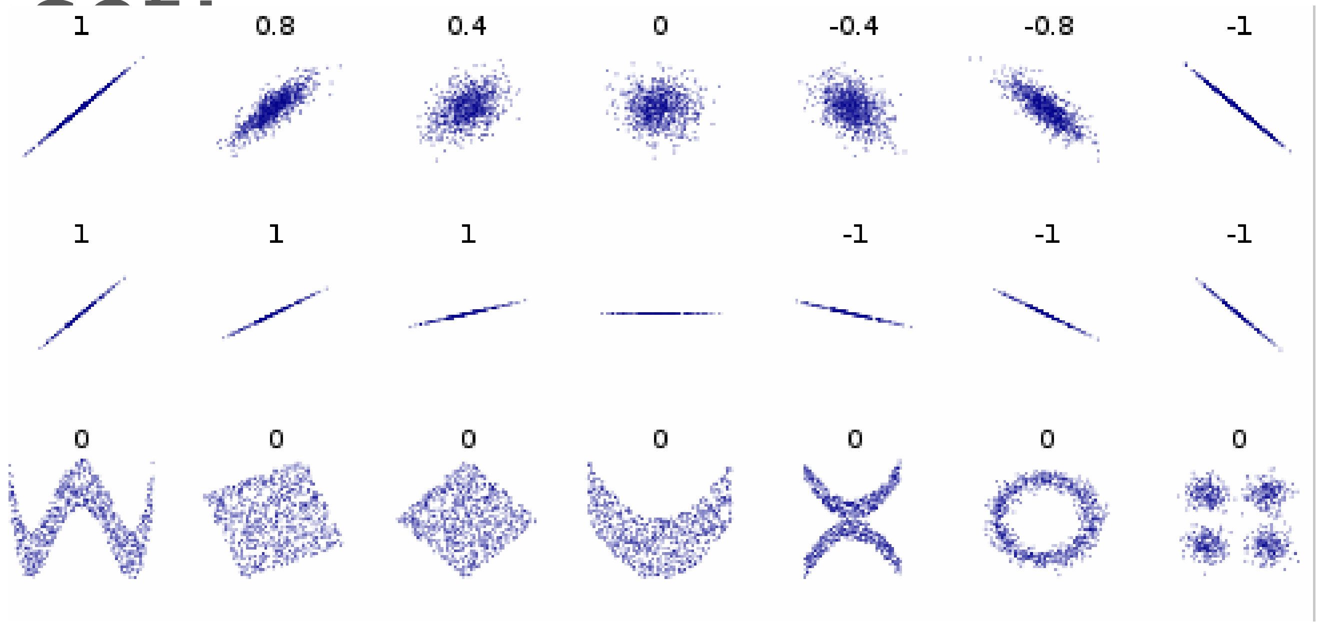
$\bar{x} =$ mean of X

Types of Correlation



Pearson CorrelationCoefficient

- Several sets of (x, y) points, with the correlation coefficient for each set:



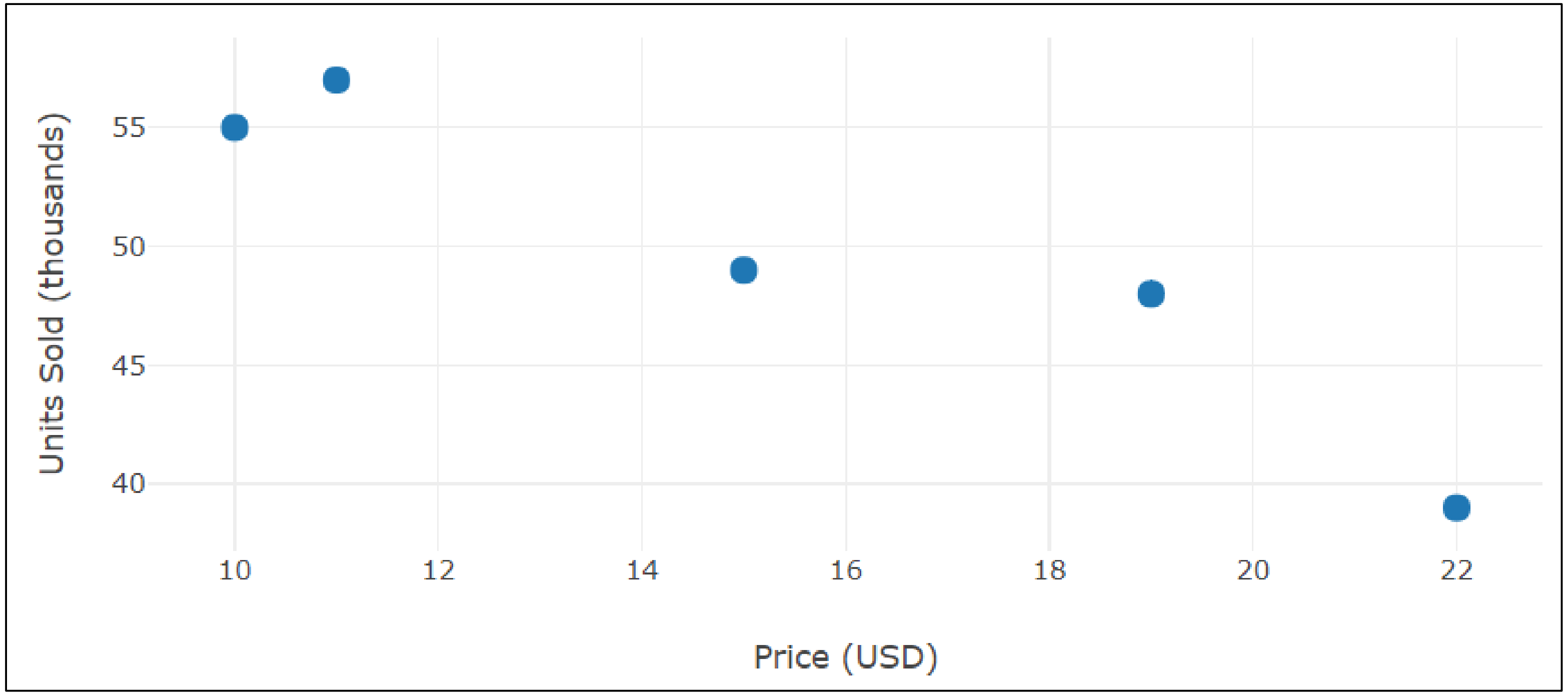
Correlation Exercise

- A company decides to test sales of a new product in five separate markets, to determine the best price point.
- They set a different price in each market and record sales volume over the same 30 day period.

Correlation Exercise

- These are the results
- Plot the results

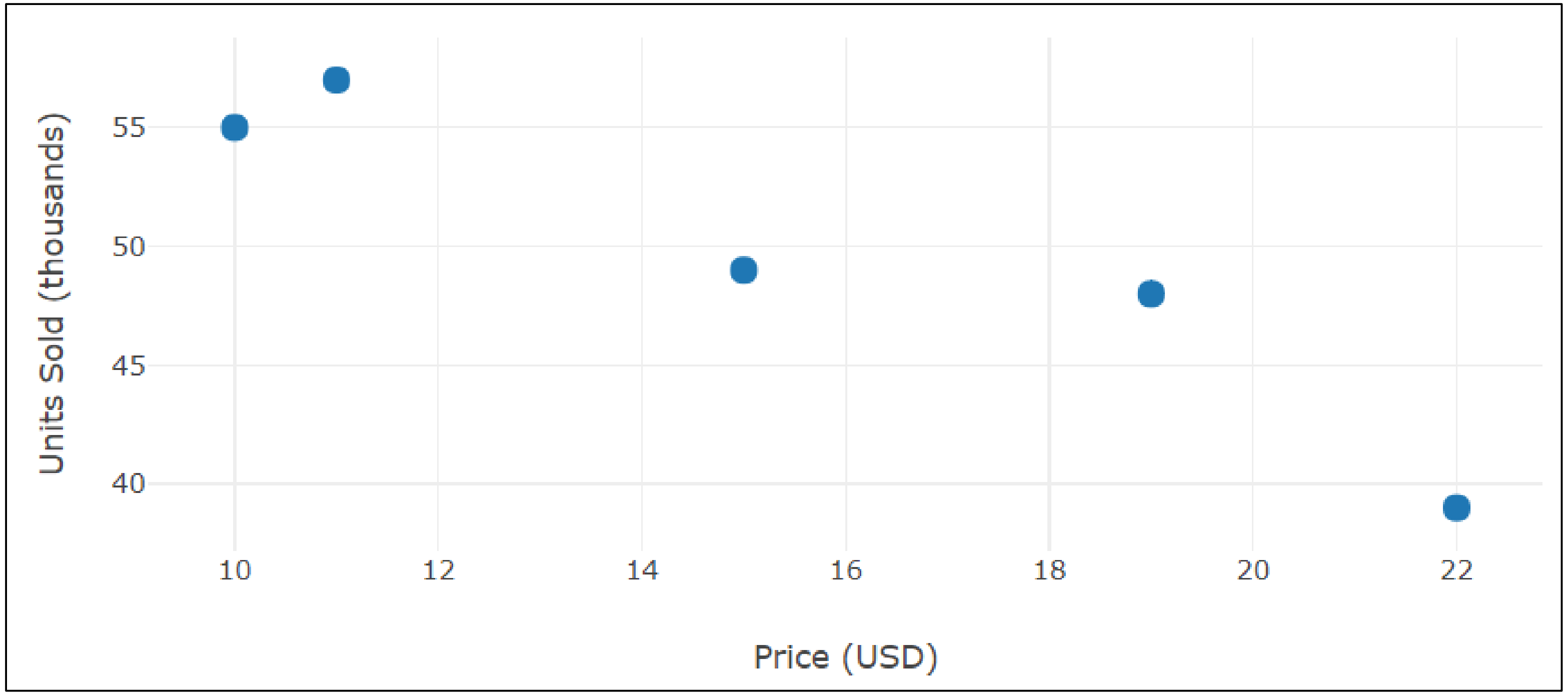
Price (USD)	Units Sold (thousands)
10	55
11	57
15	49
19	48
22	39



Correlation Exercise

- There appears to be a strong correlation, but how strong?

Price (USD)	Units Sold (thousands)
10	55
11	57
15	49
19	48
22	39



Correlation Exercise

1. Recall the simplified correlation formula:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

Price (USD)	Units Sold (thousands)
10	55
11	57
15	49
19	48
22	39

2. Find the mean of x and y:

$$\bar{x} = \frac{10+11+15+19+22}{5} = 15.4$$

$$\bar{y} = \frac{55+57+49+48+39}{5} = 49.6$$

Correlation Exercise

$$\rho_{X,Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

$\bar{x} = 15.4 \quad \bar{y} = 49.6$

3. Calculate $(x-\bar{x})$ and $(y-\bar{y})$:

Price (USD)	Units Sold (thousands)	$(x-\bar{x})$	$(y-\bar{y})$
10	55	-5.4	5.4
11	57	-4.4	7.4
15	49	-0.4	-0.6
19	48	3.6	-1.6
22	39	6.6	-10.6

Correlation Exercise

$$\rho_{X,Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

$\bar{x} = 15.4 \quad \bar{y} = 49.6$

4. Calculate $(x-\bar{x})(y-\bar{y})$:

Price (USD)	Units Sold (thousands)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
10	55	-5.4	5.4	-29.16
11	57	-4.4	7.4	-32.56
15	49	-0.4	-0.6	0.24
19	48	3.6	-1.6	-5.76
22	39	6.6	-10.6	-69.96

Correlation Exercise

$$\rho_{X,Y} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

$\bar{x} = 15.4 \quad \bar{y} = 49.6$

5. Calculate $(x-\bar{x})^2$ and $(y-\bar{y})^2$:

Price (USD)	Units Sold (thousands)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
10	55	-5.4	5.4	-29.16	29.16	29.16
11	57	-4.4	7.4	-32.56	19.36	54.76
15	49	-0.4	-0.6	0.24	0.16	0.36
19	48	3.6	-1.6	-5.76	12.96	2.56
22	39	6.6	-10.6	-69.96	43.56	112.36

Correlation Exercise

$$\rho_{X,Y} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

$\bar{x} = 15.4 \quad \bar{y} = 49.6$

6. Compute the sums:

Price (USD)	Units Sold (thousands)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
10	55	-5.4	5.4	-29.16	29.16	29.16
11	57	-4.4	7.4	-32.56	19.36	54.76
15	49	-0.4	-0.6	0.24	0.16	0.36
19	48	3.6	-1.6	-5.76	12.96	2.56
22	39	6.6	-10.6	-69.96	43.56	112.36
			Σ	**-137.2**	**105.2**	**199.2**

Correlation Exercise

$$\rho_{X,Y} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2}\sqrt{\Sigma(y - \bar{y})^2}}$$

$\bar{x} = 15.4 \quad \bar{y} = 49.6$

7. Plug these into the original formula:

Price (USD)	Units Sold (thousands)	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
10	55	-5.4	5.4	-29.16	29.16	29.16
11	57	-4.4	7.4	-32.56	19.36	54.76
15	49	-0.4	-0.6	0.24	0.16	0.36
19	48	3.6	-1.6	-5.76	12.96	2.56
22	39	6.6	-10.6	-69.96	43.56	112.36
			Σ	**-137.2**	**105.2**	**199.2**

Correlation Exercise

$$\rho_{X,Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

$\bar{x} = 15.4 \quad \bar{y} = 49.6$

7. Plug these into the original formula:

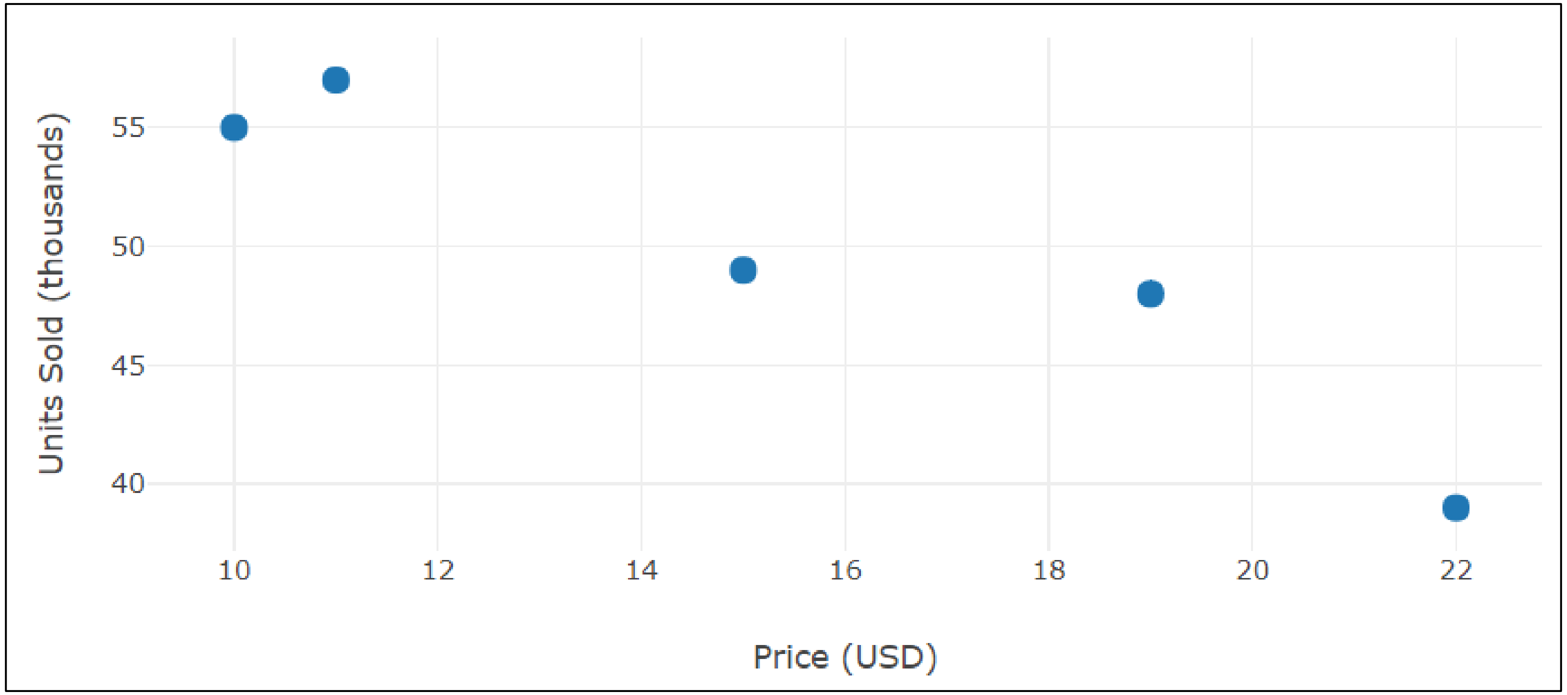
$$\rho_{X,Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} = \frac{-137.2}{\sqrt{105.2}\sqrt{199.2}}$$

$$= \frac{-137.2}{10.26 \times 14.11} = \frac{-137.2}{144.8} = \mathbf{-0.948}$$

Σ	-137.2	105.2	199.2

Correlation Exercise

- $\rho_{X,Y} = -0.948$ shows a *very* strong negative correlation!



Central Limit Theorem

When samples of size **n>=30** are drawn from a population and distributed with individual samples mean then any distribution changes to normal distribution

Standard Deviation of sample mean = $\frac{\sigma}{\sqrt{n}}$

Key Points

1. Also called as Standard Error (SE)

Standard deviation of sample mean = **(population standard deviation/square root(n))**

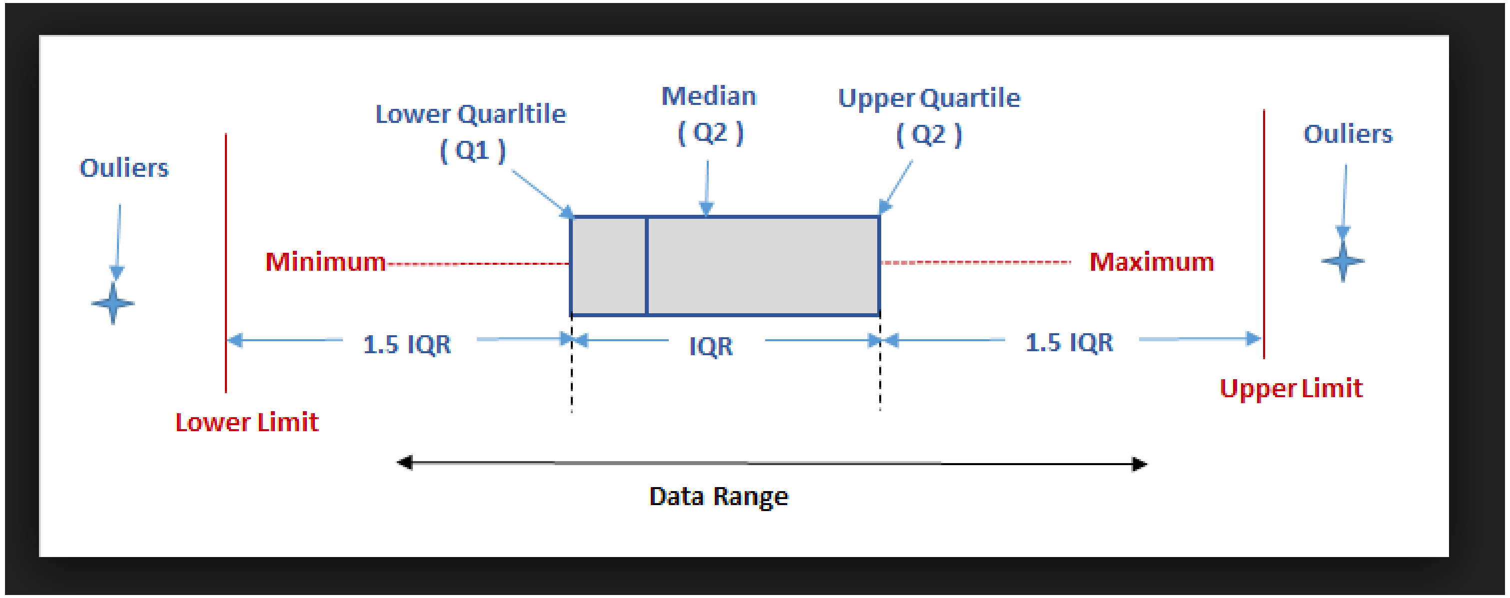
2. Mean of sample means distribution = **Population mean**

NOTE: As n increases SE decreases - SE is inversely proportional to n

Data Visualization - Plots

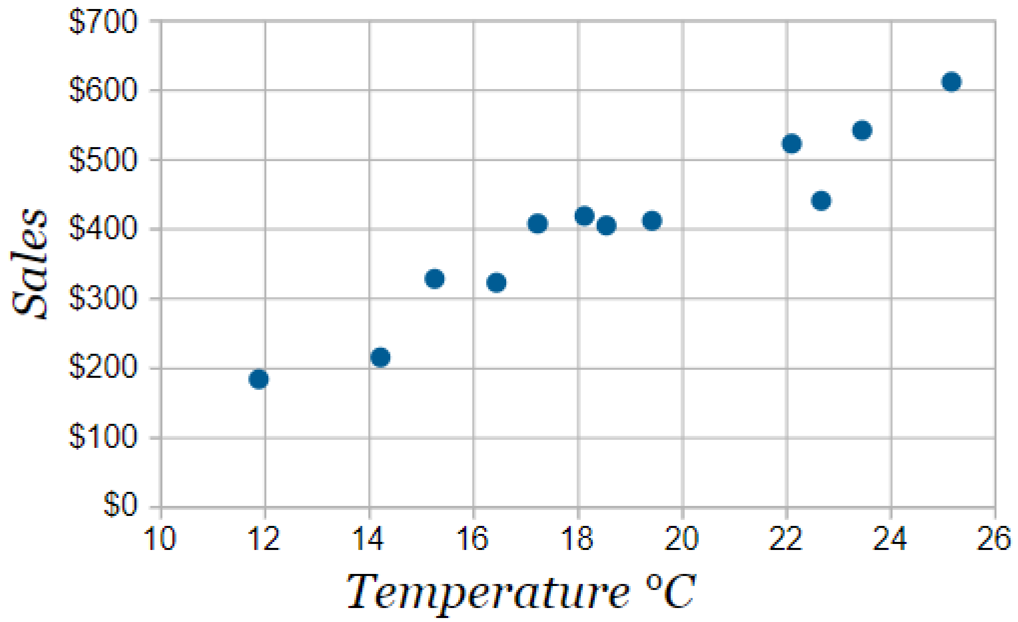
1. *Box Plot*
2. *Scatter plot*
3. *Histogram*
4. *Density Plot*

Box Plot - Shows the data spread for individual columns



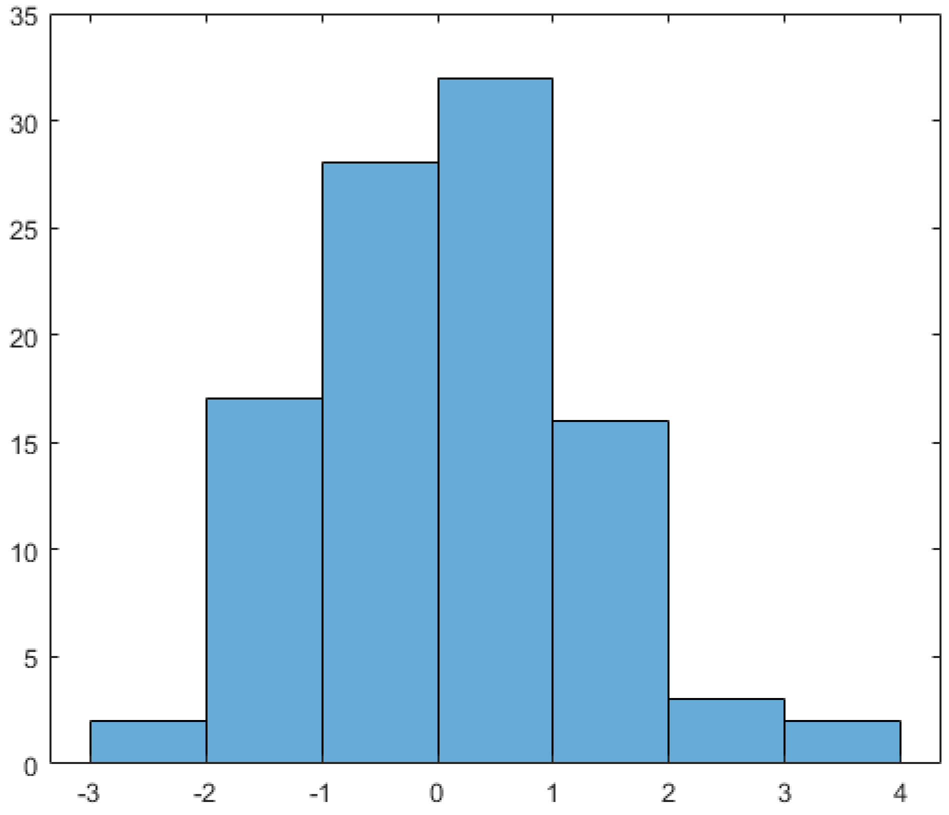
Scatter Plot - Shows relationship between 2 columns

Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

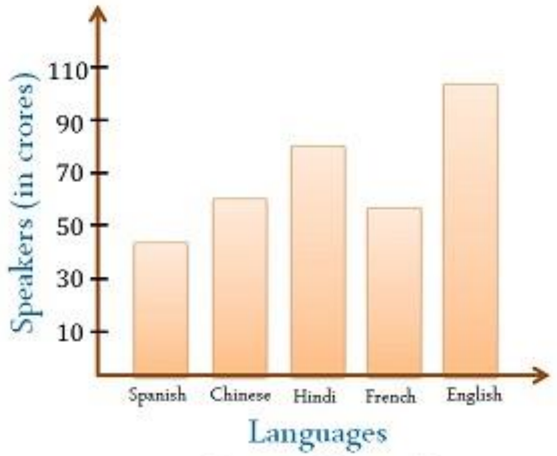


Histograms

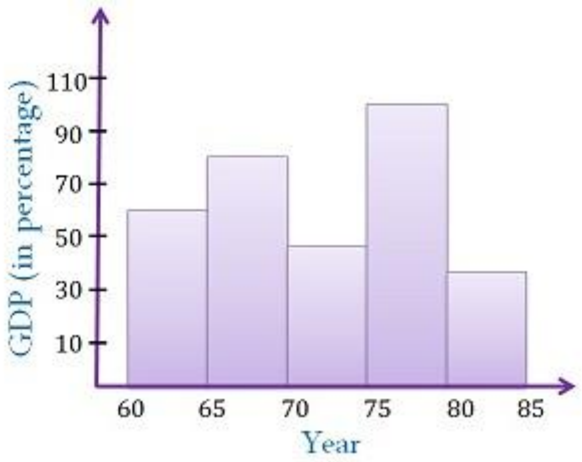
A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable



Difference between Histogram and Bar graph?

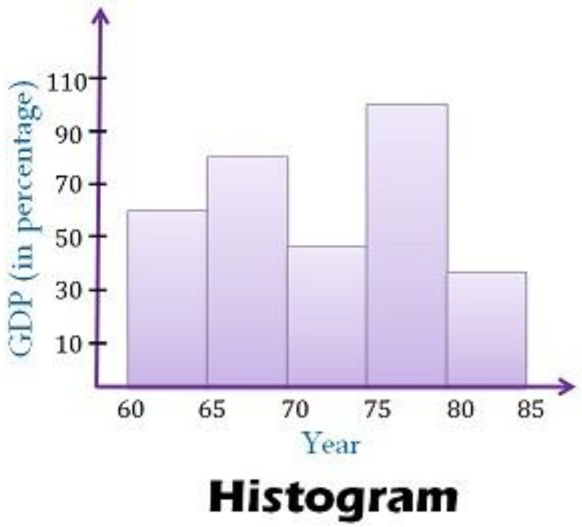


Bar Graph



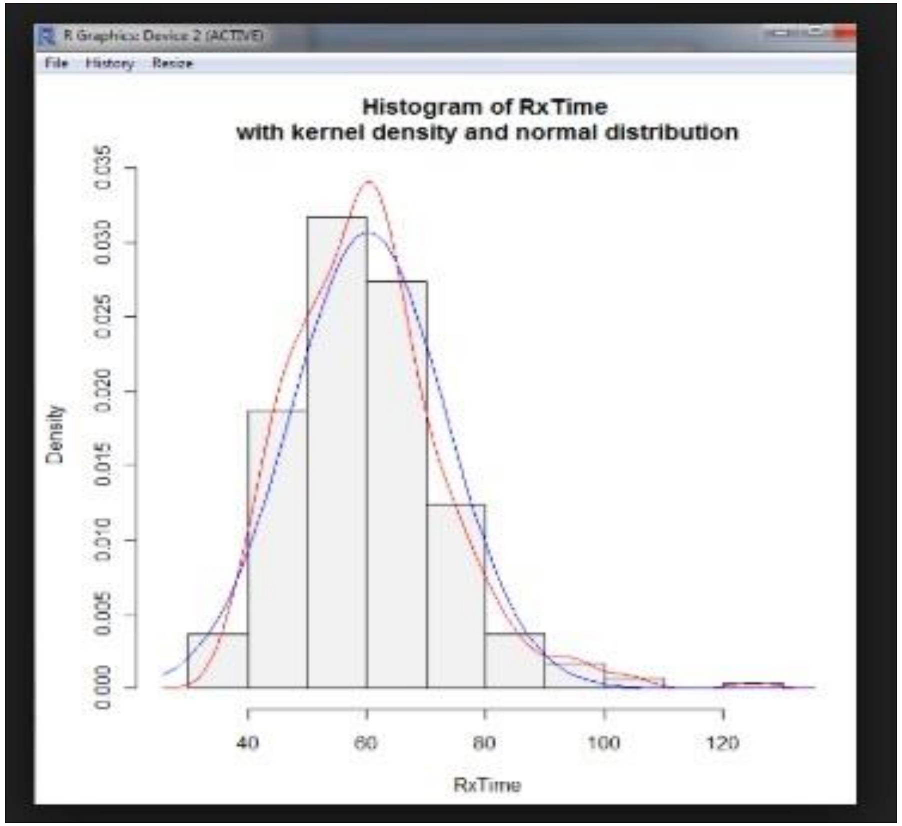
Histogram

Difference between Histogram and Bar graph?



A **histogram** represents the frequency distribution of continuous variables. Conversely, a **bar graph** is a diagrammatic comparison of discrete variables. Histogram presents numerical data whereas **bar graph** shows categorical data.

Density Plot - Shows the distribution of data



Statistical simulation link

http://www.shodor.org/interactivate/activities/