Forward School

Statistical Data Analysis

Instructor, Nero Chan Zhen Yu

Agenda

Topic 1: Basic Statistics

- Why Statistics Matter
- Types of Data
- Descriptive Statistics
- Probability and Conditional Probability
- Probability Distributions

Agenda

Forward School

Topic 2: Sampling and Hypothesis Testing

- Sampling
- Central Limit Theorem
- Sampling Distribution and Standard Errors of Sample Mean
- Confidence Interval
- Z and T Statistics
- Overview of Hypothesis Testing
- Types of Hypothesis Testing
- Type 1 and Type 2 Errors
- Analysis of Variance (ANOVA)

Agenda

Topic 3: Regression and Correlation Analysis

- Regression Modelling
- Residues and Mean Square Error
- Covariance and Correlation Analysis

Topic 1:
Basic Statistics

What is Statistics?

Statistics is discipline which is concerned with:

- Designing experiments and other data collection
- Summarizing information to aid understanding
- Drawing conclusion from data
- Estimating the present and predicting the future

What is Statistics?

Statistics Statements:

- “I sleep for about eight hours per night on average.”
- “You are more likely to pass the exam if you start preparing earlier.”

Why Statistics Matter?

- Environmental Study

  - Is Malaysia getting hotter over last 10 years?

- Policy Study

  - Is more people using green transport such as Bicycles, Buses, Carpool, CNG Cars, Electric Cars, Electric Scooters.

- Market Analysis

  - Is more people likely to take green transport If they’ve seen a recent TV advertisement for green transport?

- Public Transport

  - Is more people likely to commute by MRT if we have more MRT stations in the neighborhood?

- Health Care

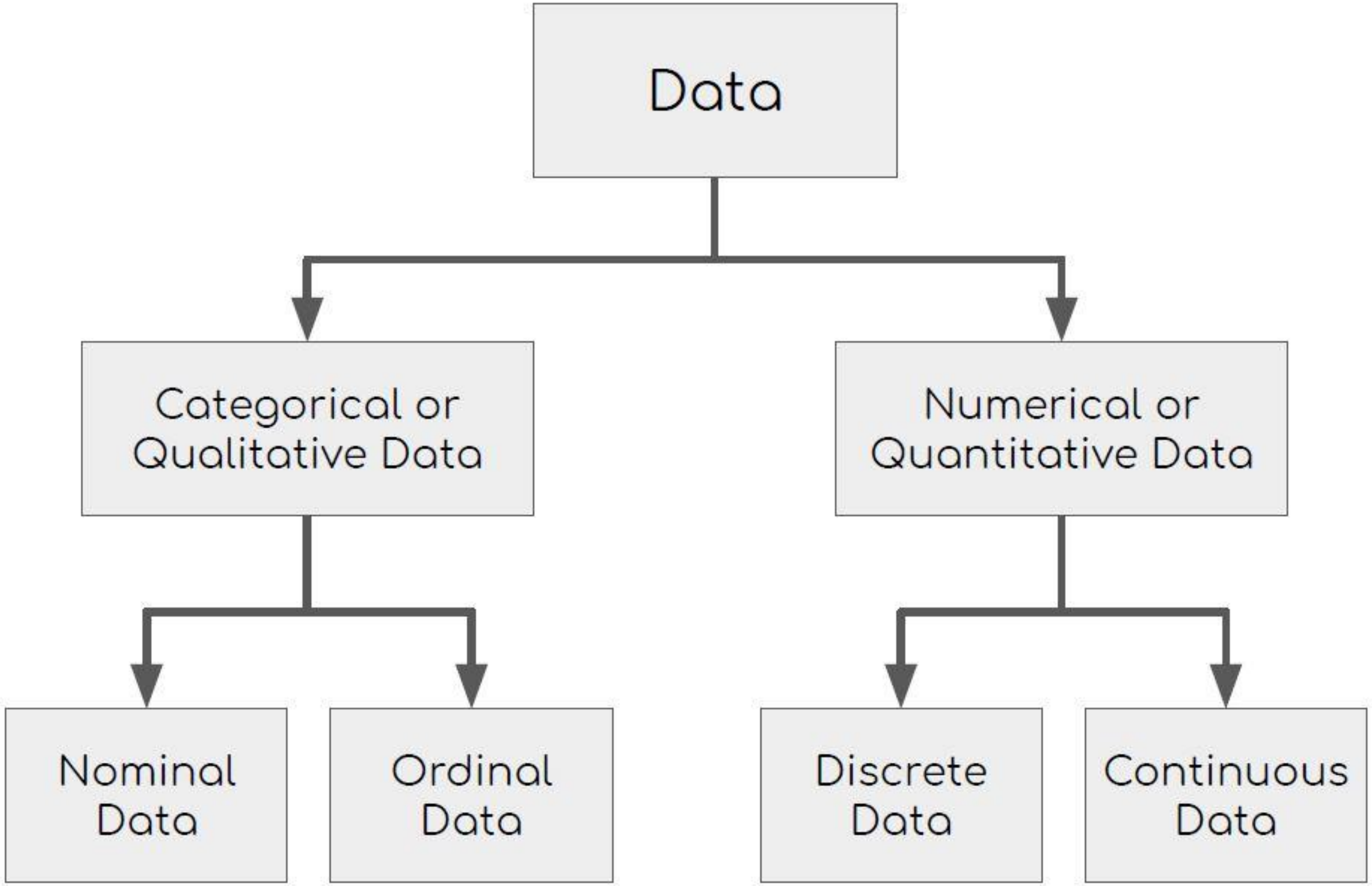
  - Does air pollution from vehicles cause any health concern?

- Data Science

  - Statistics is a fundamental for understanding Artificial Intelligence and Machine Learning.

Types of Data

Forward School



Data

Categorical or Qualitative Data

Numerical or Quantitative Data

Nominal Data

Ordinal Data

Discrete Data

Continuous Data

Categorical and Quantitative Data

- Categorical (Qualitative) Data – each observation belongs to one of a set of categories. Example:
  - Weather (Rainy/Sunny)
  - Air Pollutants (Ozone/Nitrogen Dioxide)
  - Gender (Male or Female)
  - Place of Resident (HDB, Condo)
  - Marital status (Married, Single)
- Quantitative (Numerical) Data – observations take numerical values. Examples:
  - Surface Air Temperature.
  - Weekly number of dengue cases.
  - No. of days with rainfall in a month.
  - Age.
  - Number of cars.
  - Weight.

Nonimal and Ordinal Data

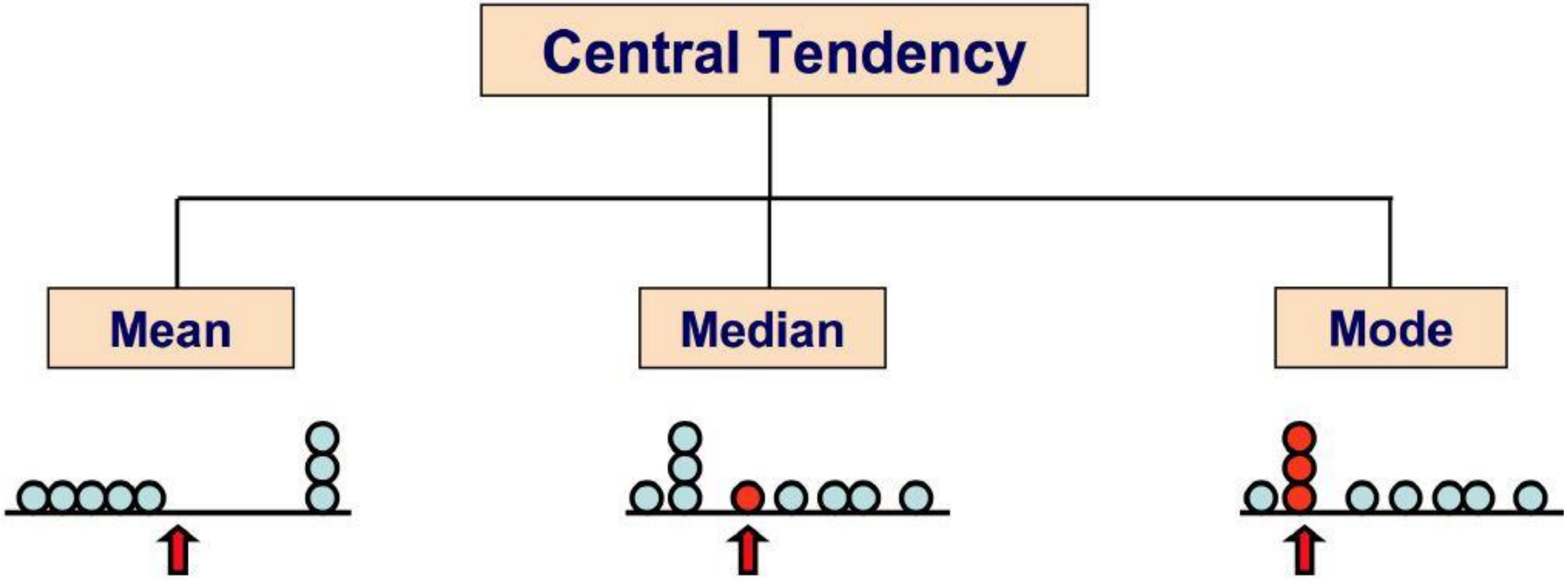
- Nominal Data is defined as data that is used for naming or labelling variables, without any quantitative value. It is sometimes called “labels” data, Eg:
  - Male/Female
  - Red/Green/Blue
- Ordinal Data is a type of categorical data with an order. The variables in ordinal data are listed in an ordered manner:
  - Disagree/Neutral/Agree/Strongly Agree
  - Very Bad/Bad/Good/Very Good

Discrete and Continuous Data

Forward School

- Discrete Data is a set of countable numbers such as 0, 1, 2, 3, … , for example:
  - No. of days with rainfall in a month.
  - Weekly no. of dengue cases.
  - Number of children in a family
  - Number of foreign languages spoken
- Continuous Data are continuous numbers from an interval. Examples:
  - Surface Air Temperature.
  - Amount of rainfall in a month.
  - Height.
  - Weight.

Measures of Central Tendency



Mean – add up all the values and divide by how many there are.

Median – Arrange all the numbers from smallest to largest:

- Odd number of points: Median = middle value
- Even number of points: Median = mean of the middle two values

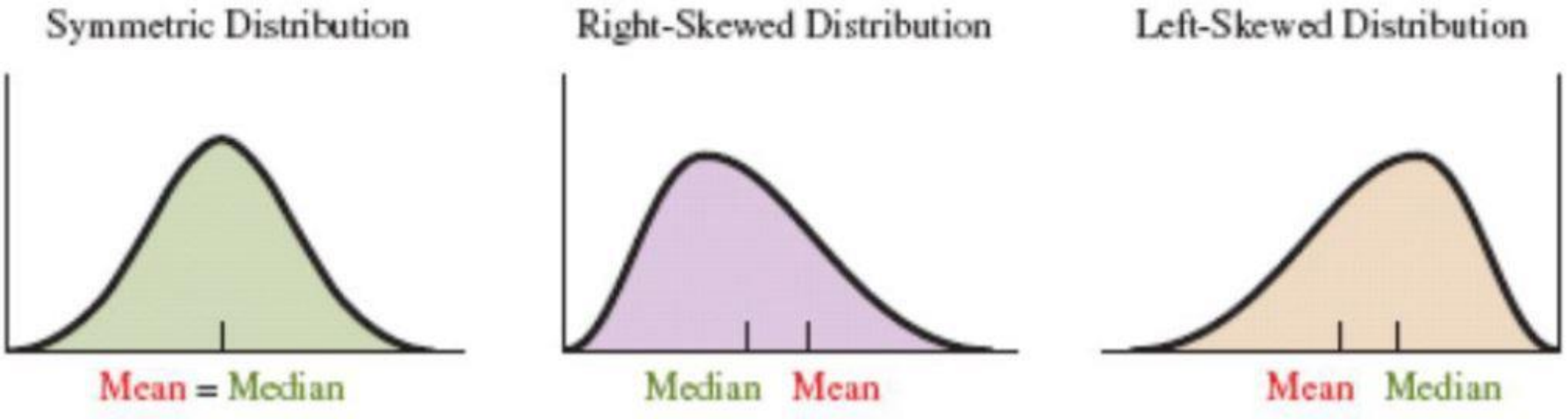
Median vs Mean

Mean

- Useful for roughly symmetric quantitative data.
- Sensitive to outliner data.

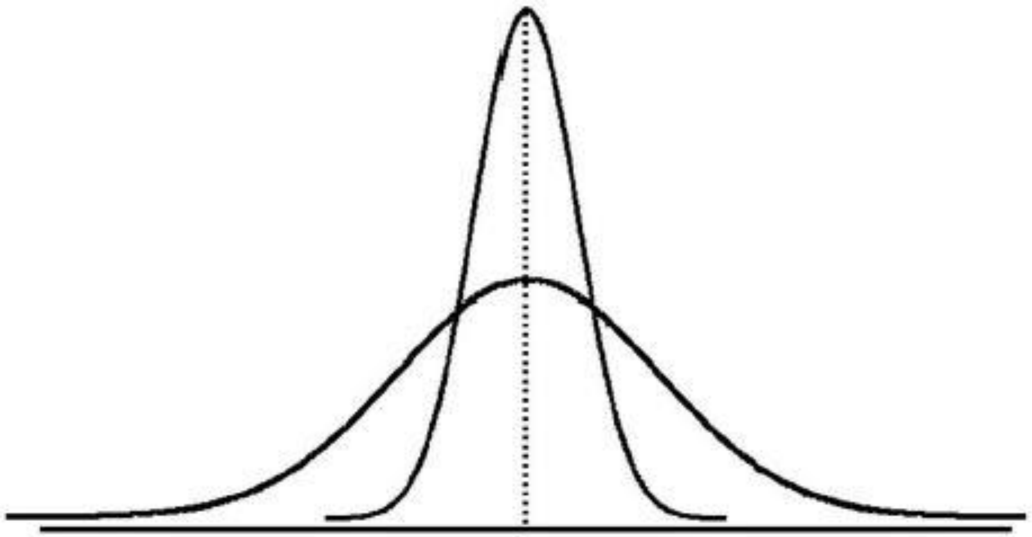
Median

- Splits the data into halves.
- Useful for highly skewed quantitative data.
- Insensitive to outliner data.



Measures of Dispersion

- The measures of dispersion measure the differences between how far “spread out” the data values are.
- Two commonly used measures for dispersion are range and standard deviation.



Same center, different variation

Standard Deviation

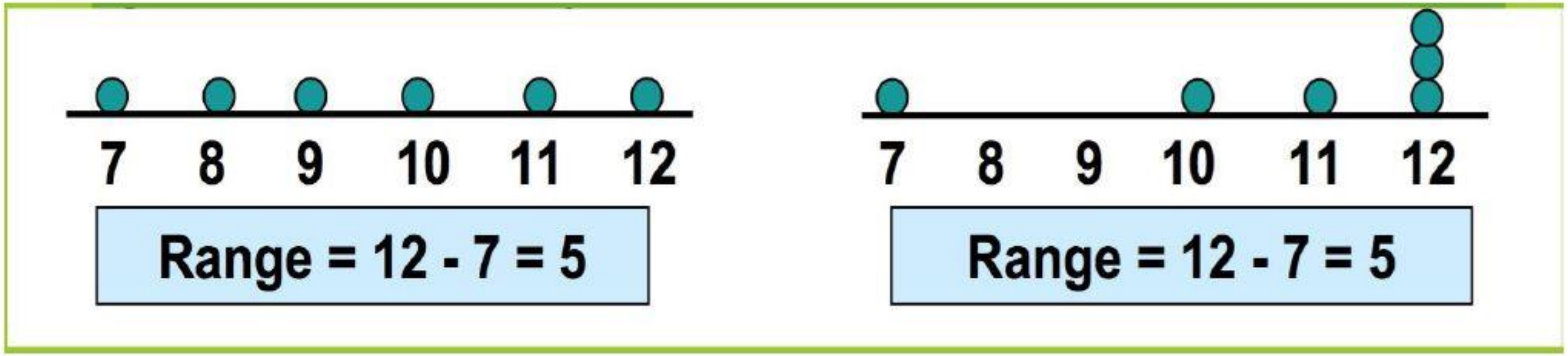
- The standard deviation measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.
- Larger standard deviation means greater variability of the data.

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

Range

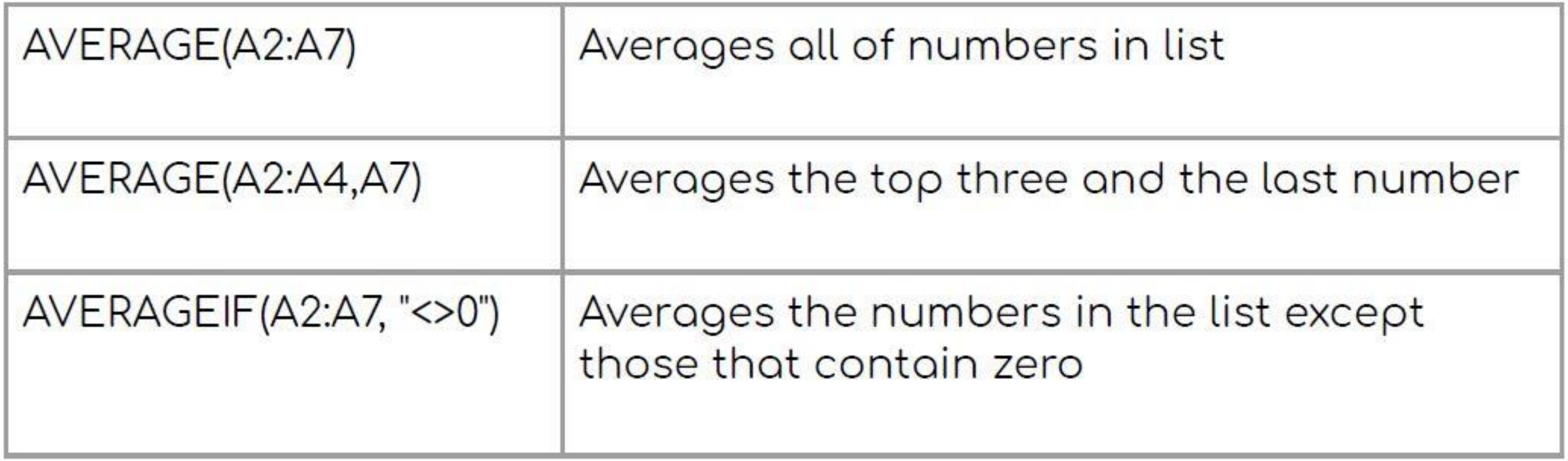
Forward School

- Range is the difference between the highest and lowest values.
- Since it uses only the extreme values, it is greatly affected by extreme values.
- Range ignores the way in which data are distributed.



Average Function in Excel

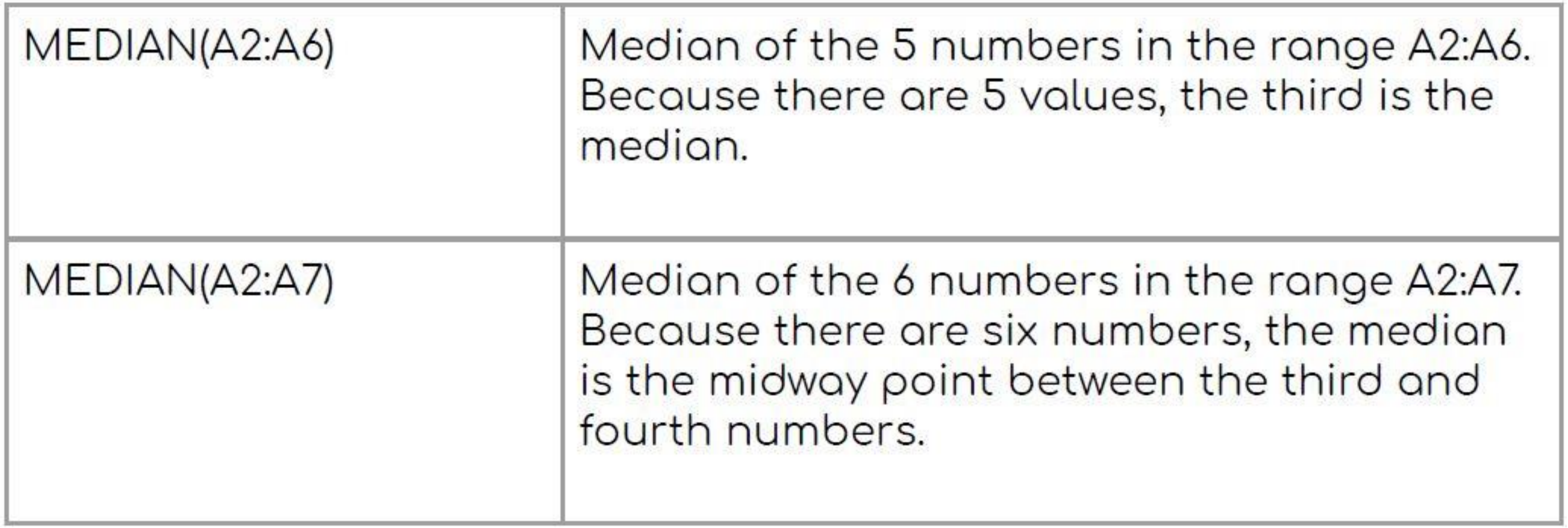
- The Average function measures central tendency, which is the location of the center of a group of numbers in a statistical distribution.



AVERAGE(A2:A7)	Averages all of numbers in list
AVERAGE(A2:A4,A7)	Averages the top three and the last number
AVERAGEIF(A2:A7, "<>0")	Averages the numbers in the list except those that contain zero

Median Function in Excel

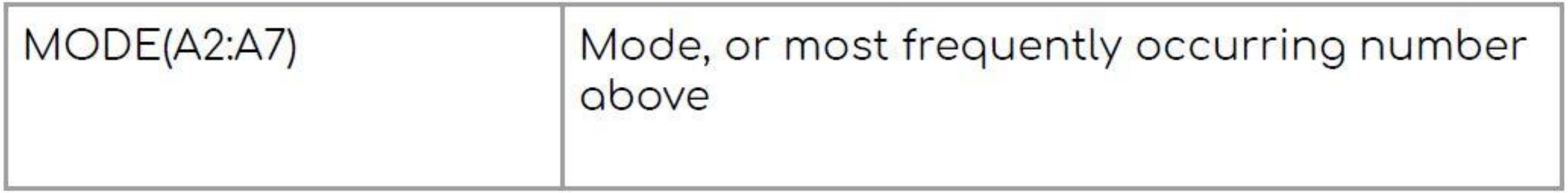
- The Median function measures central tendency, which is the location of the center of a group of numbers in a statistical distribution.



MEDIAN(A2:A6)	Median of the 5 numbers in the range A2:A6. Because there are 5 values, the third is the median.
MEDIAN(A2:A7)	Median of the 6 numbers in the range A2:A7. Because there are six numbers, the median is the midway point between the third and fourth numbers.

Mode Function in Excel

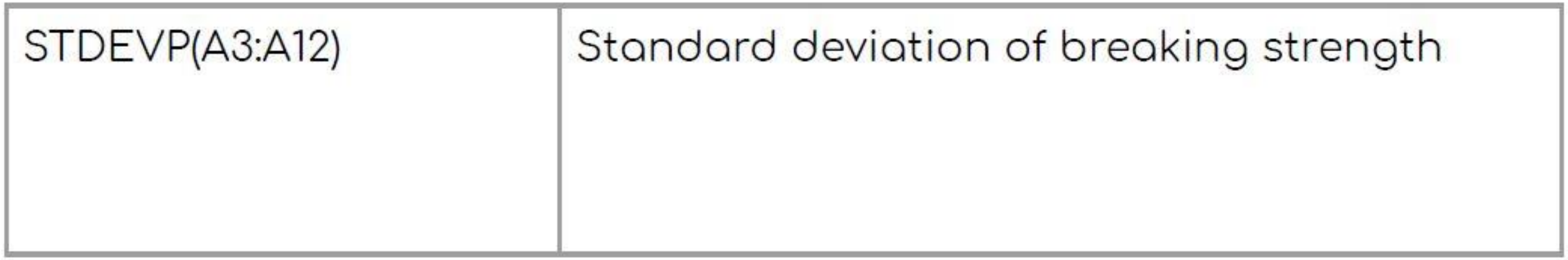
- The Mode function measures central tendency, which is the location of the center of a group of numbers in a statistical distribution.



MODE(A2:A7)	Mode, or most frequently occurring number above

STDEVP Function in Excel

- STDEVP assumes that its arguments are the entire population. If your data represents a sample of the population, then compute the standard deviation using STDEV.



STDEVP(A3:A12)	Standard deviation of breaking strength

Activity: Descriptive Statistics

Consider the following three sets of observations:

- Set 1: 8, 9, 10, 11, 12
- Set 2: 8, 9 , 10, 11, 100
- Set 3: 8, 9, 10, 11, 1000

a) Find the median and mean for each data set.

b) Find the range and standard deviation for each data set.

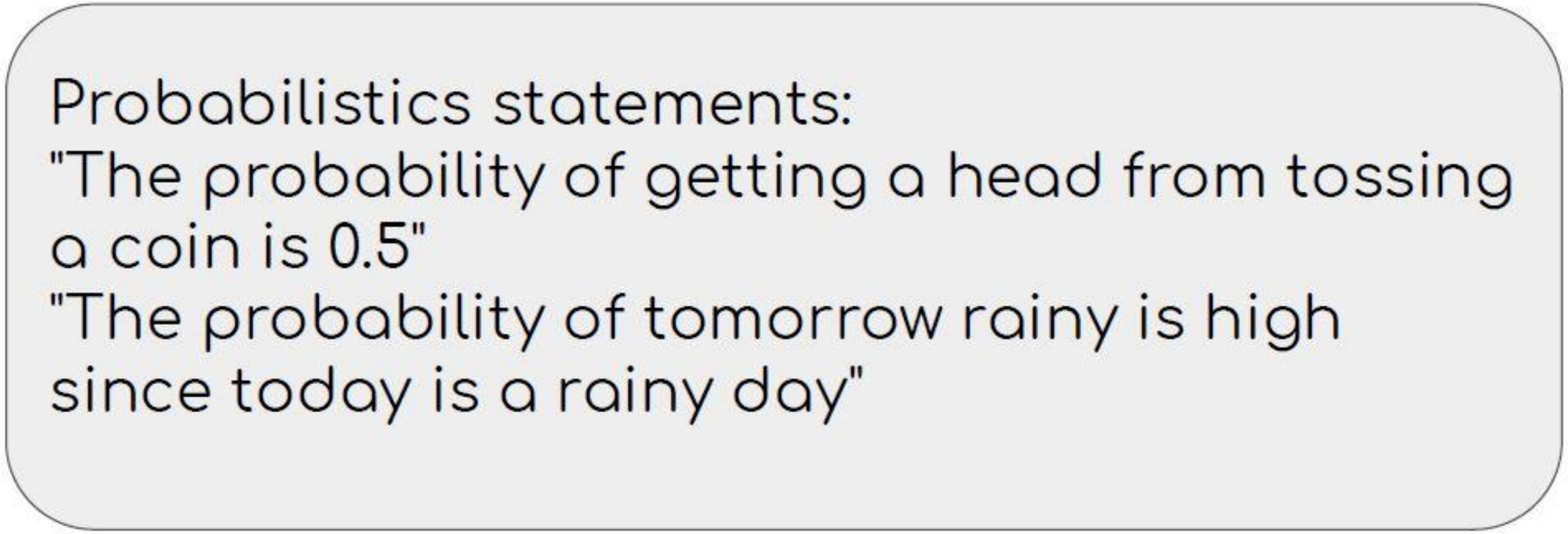
c) What do these data sets illustrate about the resistance of the median and mean?

- Verify the result with the online descriptive statistics tool to compute the answer.

  Link: https://www.calculatorsoup.com/calculators/statistics/descriptivestatistics.php

What is Probability

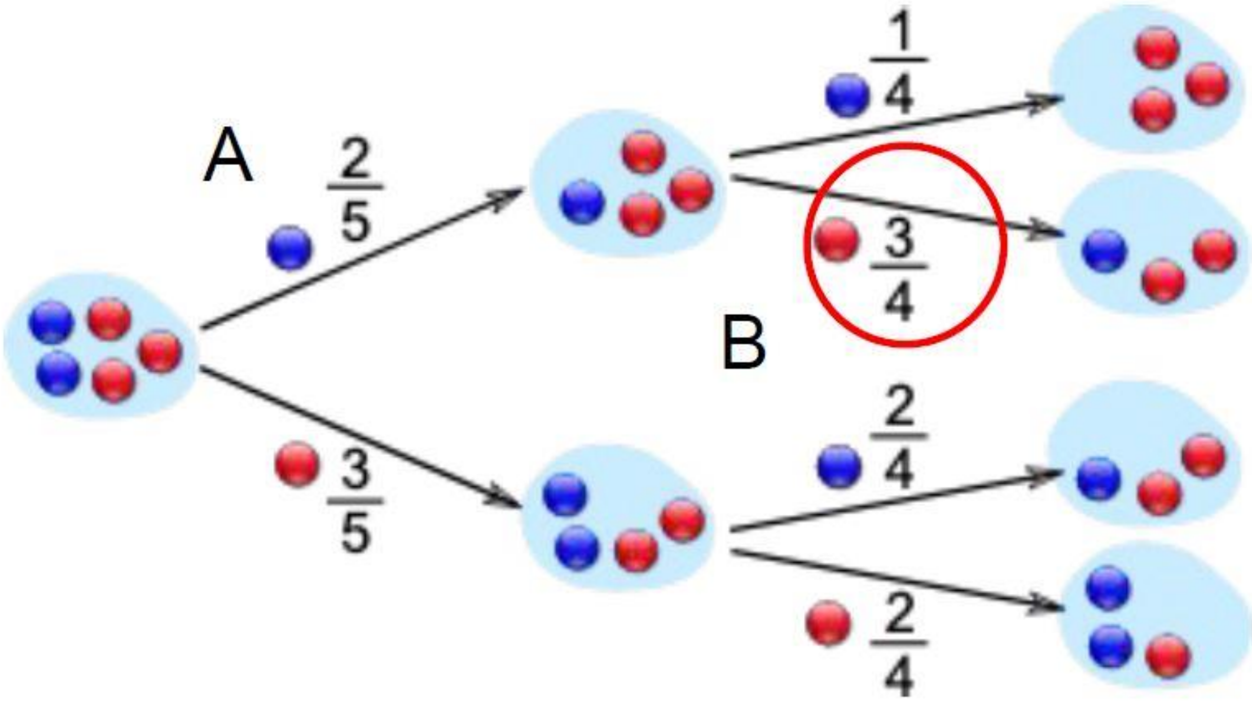
- Probability is a measure of the likelihood that an event will occur.
- Probability is quantified as a number from 0 to 1.



Probabilistics statements:
"The probability of getting a head from tossing a coin is 0.5"
"The probability of tomorrow rainy is high since today is a rainy day"

Conditional Probability

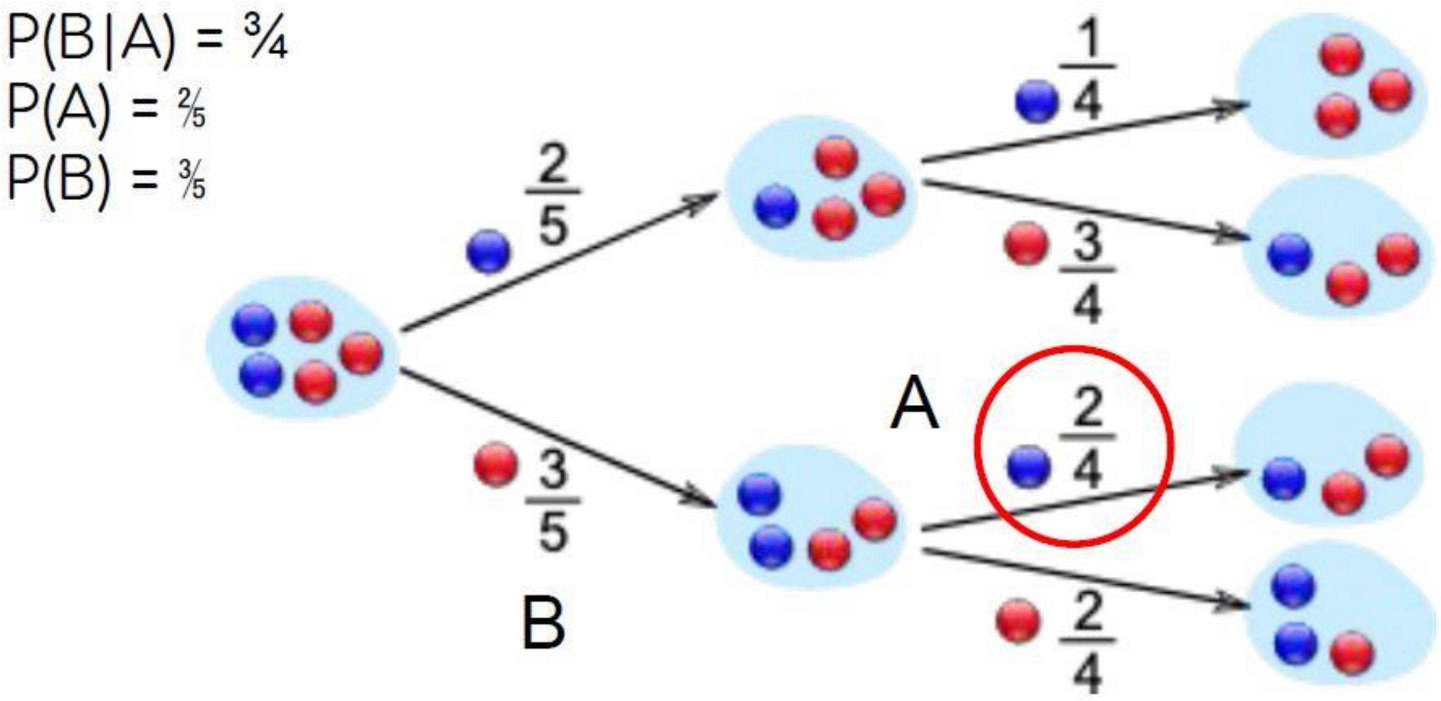
- Conditional Probability is the probability of one event(B) given another even(A) is known.
- A: get a blue marble first.
- B: get a red marble then.
- Based on the tree diagram below, P(B|A) = 3/4



Bayes’ Theorem

- Bayes’ Theorem is a way of finding a conditional probability when we know other probabilities.
- The formula is P(A|B) = [P(A)*P(B|A)]/P(B)

$$P(A|B) = P(B|A)*P(A)/P(B) = \frac{3}{4} * \frac{2}{5} / \frac{3}{5} = \frac{3}{4} * \frac{2}{3} = 2/4$$

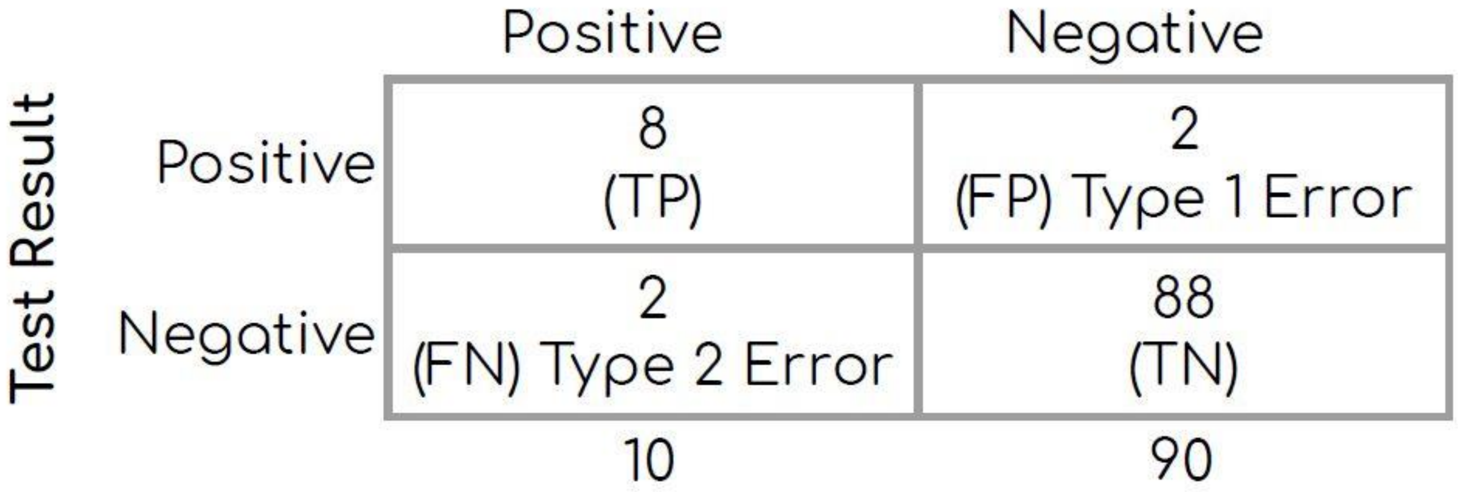


$P(B|A) = \frac{3}{4}$

$P(A) = \frac{2}{5}$

$P(B) = \frac{3}{5}$

Activity: Conditional Probability

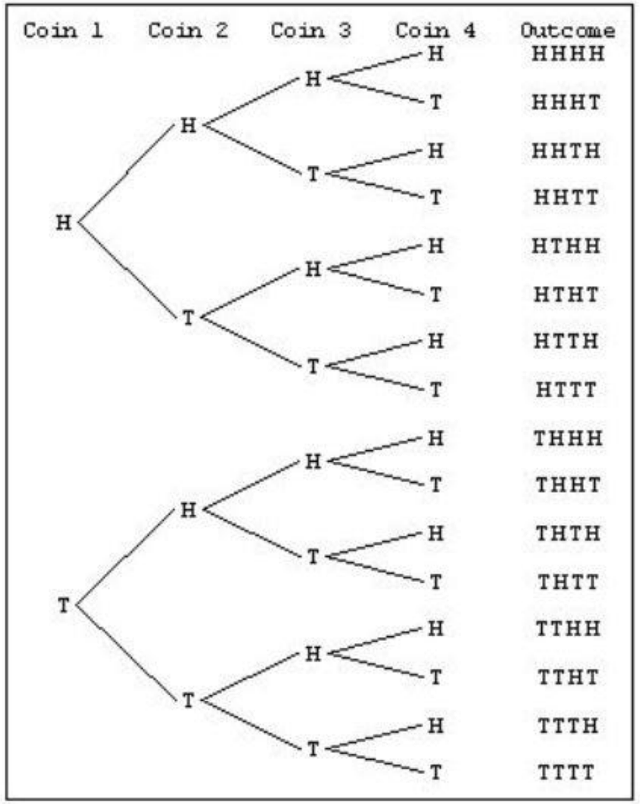


		Positive	Negative
Test Result	Positive	8 (TP)	2 (FP) Type 1 Error
	Negative	2 (FN) Type 2 Error	88 (TN)
		10	90

- If 100 people tool the COVID-19 test, the test result is shown as above.
- Compute the conditional probability that a person is positive is tested positive.
- P(test positive | actual positive)

Binomial Distribution

- The binomial distribution with parameters ‘n’ and ‘p’ is the discrete probability distribution of the number of successes.
- A binomial distribution can be thought of as simply the probability of getting # of head outcome in an experiment of tossing multiple coins.



$$P(X=k) = \binom{n}{k} p^{k} (1-p)^{n-k}$$

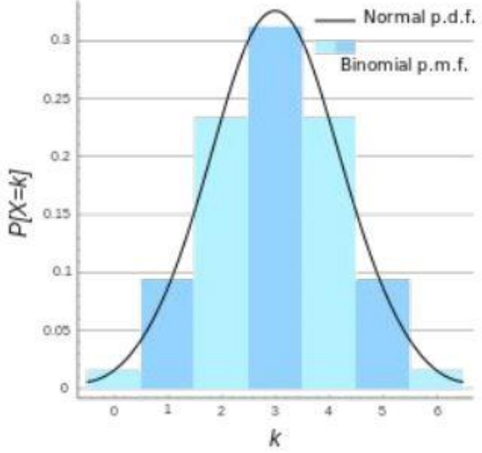
$$\mu_X = np$$

$$\sigma_X^2 = np(1-p)$$

n: No of coins/toss

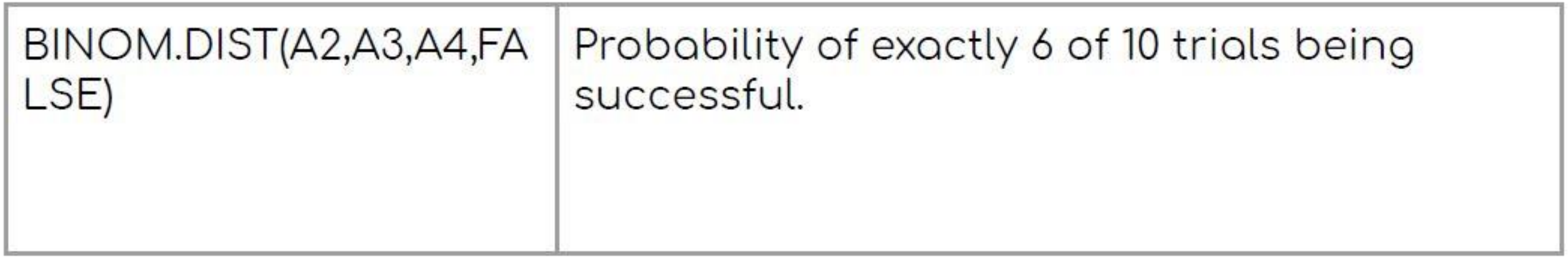
k: No of head

p = Probability for getting a head in one toss



BINOM.DIST Function in Excel

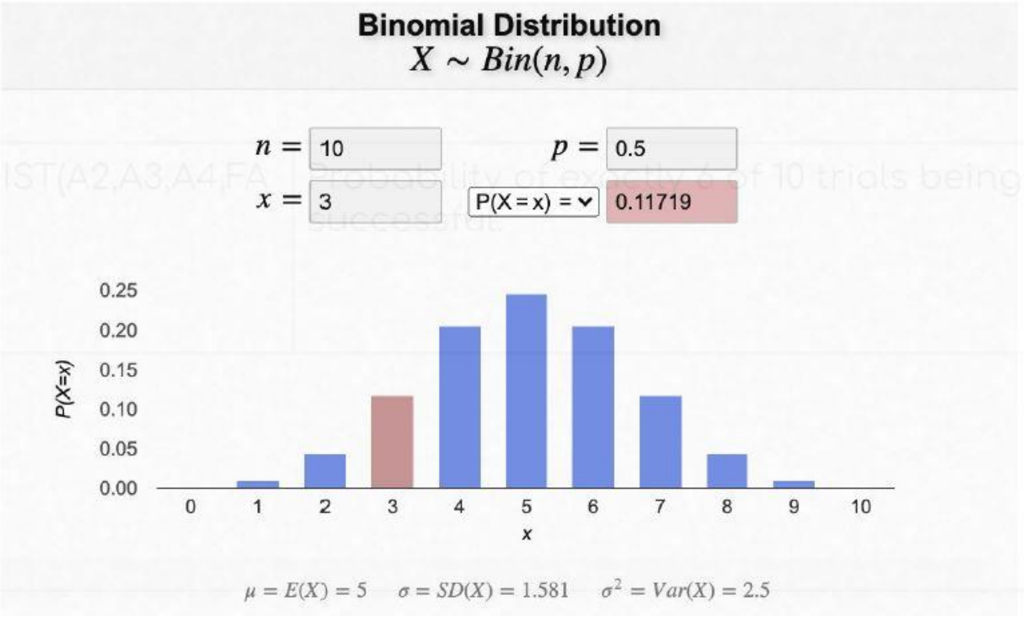
- Use BINOM.DIST in problems with a fixed number of tests or trials, when the outcomes of any trial are only success or failure, when trials are independent, and when the probability of success is constant through the experiment.
- For example, BINOM.DIST can calculate the probability that two of the next three babies born are male.



BINOM.DIST(A2,A3,A4,FALSE)	Probability of exactly 6 of 10 trials being successful.

Activity: Binomial Distribution

- Compute the Binomial probability for n=10, p=0.5, x=3, 4, 5, 6.
- Verify with the following online tool.
- Link: https://homepage.divms.uiowa.edu/~mbognar/applets/bin.html



Binomial Distribution

$X \sim Bin(n, p)$

$n =$ 10 $p =$ 0.5

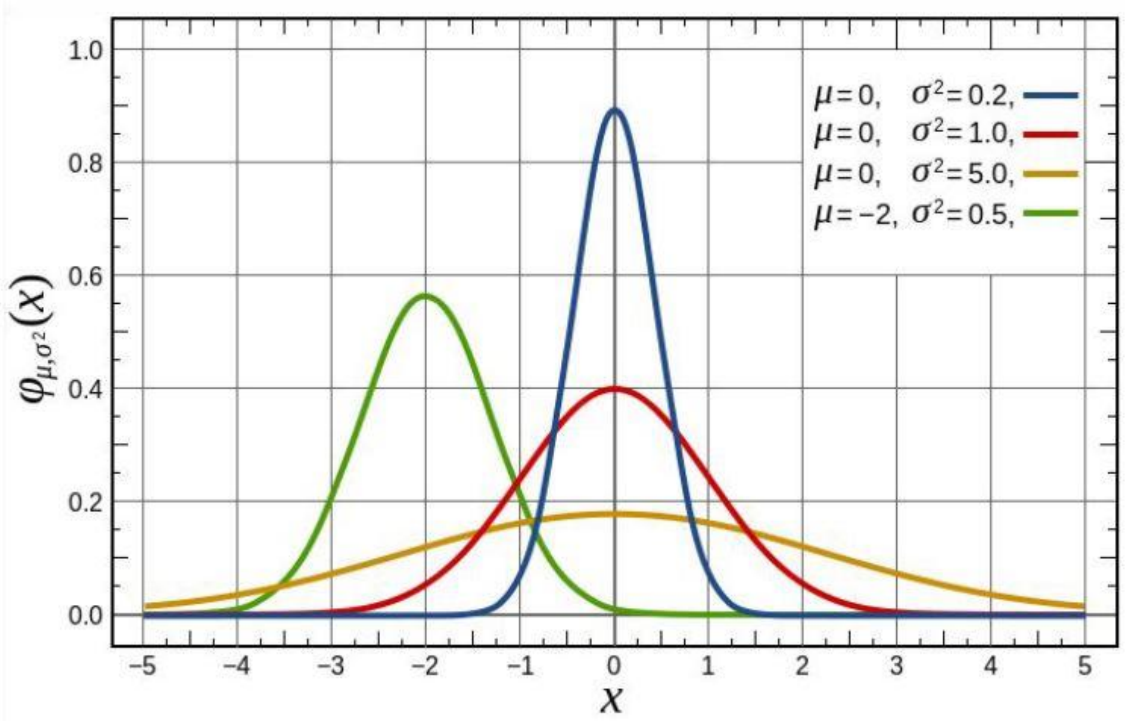
$x =$ 3 P(X = x) = 0.11719

$\mu = E(X) = 5 \quad \sigma = SD(X) = 1.581 \quad \sigma^2 = Var(X) = 2.5$

Normal Distribution

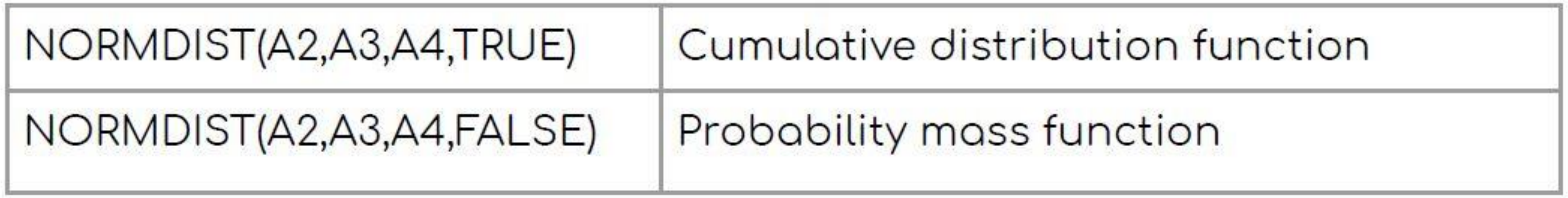
- Normal Distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.
- Normal Distribution is approximation of Binomial Distribution if n -> infinity.

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

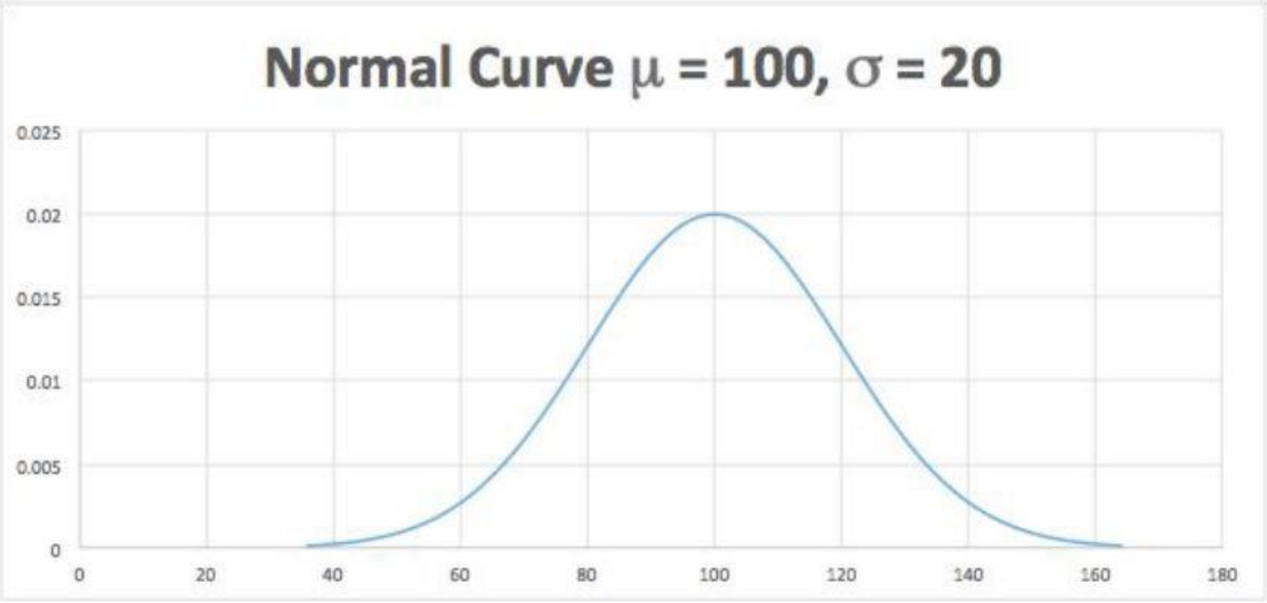


NORMDIST Function in Excel

- The NORMDIST function returns the normal distribution for the specified mean and standard deviation.



NORMDIST(A2,A3,A4,TRUE)	Cumulative distribution function
NORMDIST(A2,A3,A4,FALSE)	Probability mass function

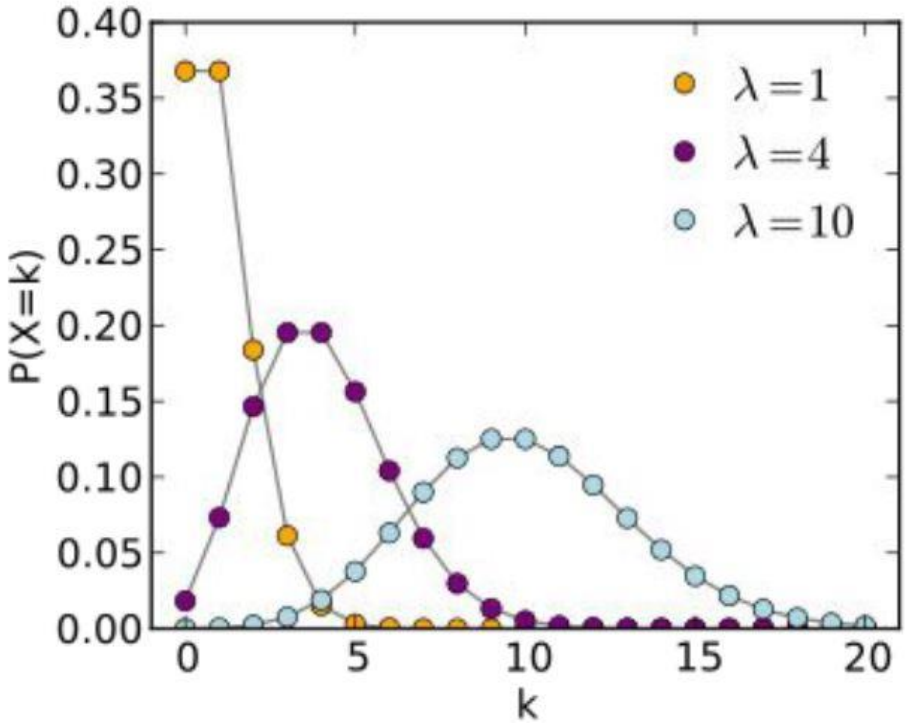


Activity: Normal Distribution

- The normal distribution for women in Malaysia has mean=160cm and standard deviation=20cm. Most major airlines have height requirements for flight attendants.
- The minimum height requirement is 170cm. What proportion of adult females in Malaysia are not tall enough to be a flight attendant?
- Compute the value above.
- Verify the answer using the following online tools.
- Link: https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html

Poisson Distribution

- The Poisson Distribution lets you estimate the number of customers who will come into a store during a given time period such as an hour or perhaps the number of seconds between times that cars arrive at a tall booth.

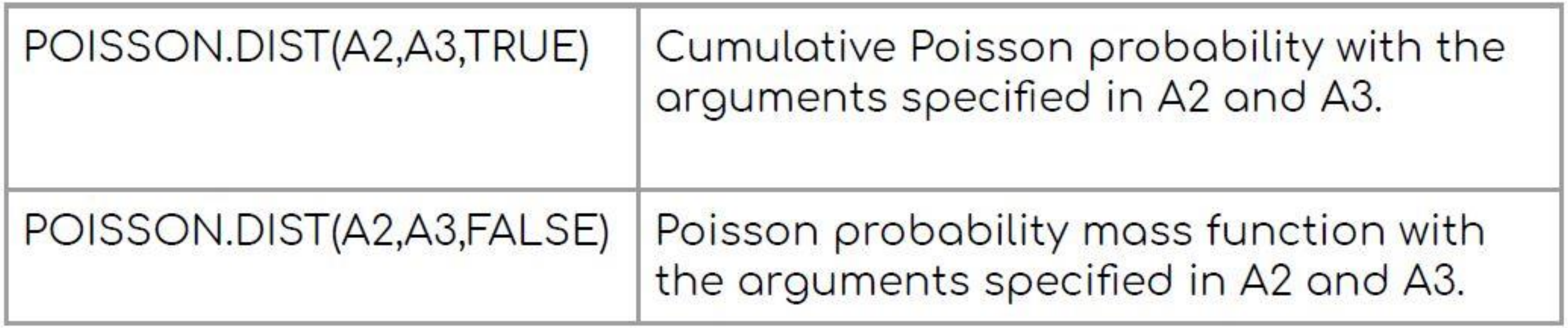


$$f(k;\lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

ere

POISSON.DIST Function in Excel

- The BINOM.DIST function returns the Poisson Distribution. A common application of the Poisson Distribution is predicting the number of events over a specific time, such as the number of cars arriving at a toll plaza in 1 minute.



POISSON.DIST(A2,A3,TRUE)	Cumulative Poisson probability with the arguments specified in A2 and A3.
POISSON.DIST(A2,A3,FALSE)	Poisson probability mass function with the arguments specified in A2 and A3.

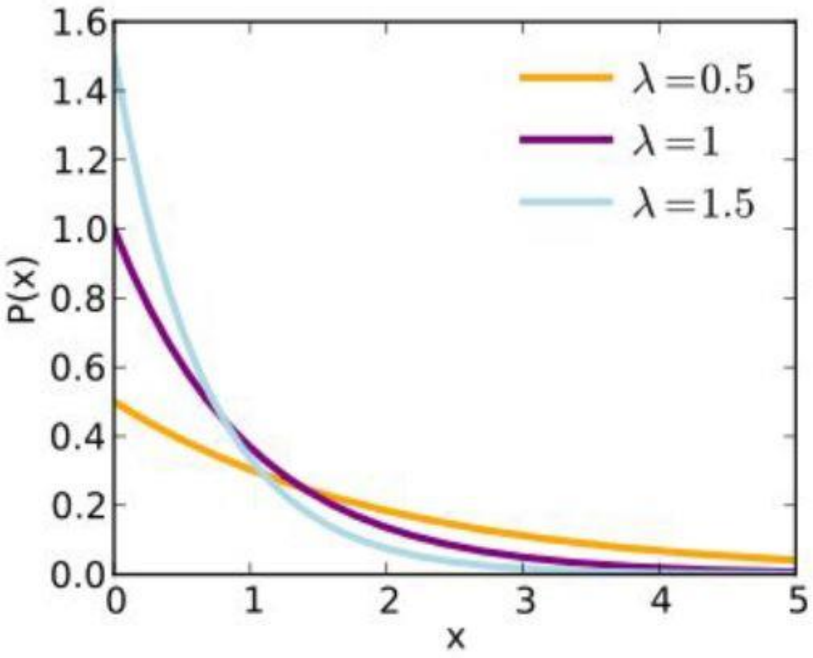
Activity: Poisson Distribution

- A bank is interest in studying the number of people who use the ATM located outside its office late at night.
- On average, 1.6 customers use the ATM during any 10 minute interval between 9pm and midnight.
- What is lambda for this problem?
- What is the probability of exactly 3 customers using the ATM during any 10 minutes interval?
- What is the probability of 3 or fewer people?

- Compute the answer and verify using the online tool.
- Link: https://www.onlinemathlearning.com/poisson-distribution.html
- Link: https://homepage.divms.uiowa.edu/~mbognar/applets/pois.html

Exponential Distribution

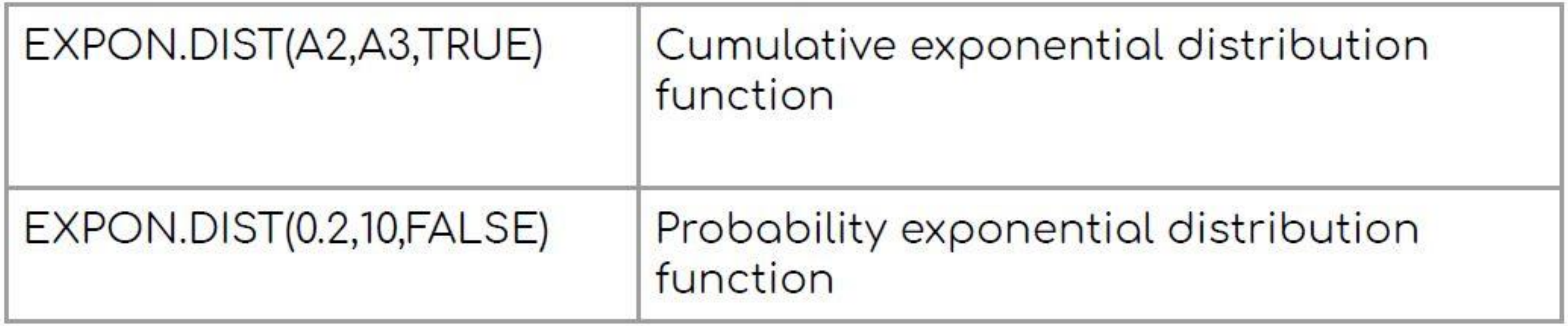
- If you sell products via your company's website, knowing the average time between orders helps you plan the number of employees you'll need to have on duty at any time.
- This type of occurrence is described by the exponential probability distribution.



$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

EXPON.DIST Function in Excel

- EXPON.DIST function returns the exponential distribution. Use EXPON.DIST to model the time between events, such as how long an automated bank teller takes to deliver cash.
- For example, you can use EXPON.DIST to determine the probability that the process takes at most 1 minute.



EXPON.DIST(A2,A3,TRUE)	Cumulative exponential distribution function
EXPON.DIST(0.2,10,FALSE)	Probability exponential distribution function

Activity: Exponential Distribution

Forward School

- The number of days ahead travelers purchase their airline tickets can be modelled by an exponential distribution with the average amount of time equal to 15 days.
- Find the probability that a traveler will purchase a ticket fewer than tendays in advance.

- Compute the answer and verify using the online tool.
- Link: https://homepage.divms.uiowa.edu/~mbognar/applets/exp-like.html

Topic 2:
Sampling and Hypothesis Testing

Sampling Theory

- Sampling Theory is the filed of statistics that is involved with the collection, analysis and interpretation of data gathered from random samples of a population under study.
- The application of sampling theory is concerned not only with the proper selection of observations from the population that will constitute the random sample.
- It also involves the use of probability theory, along with prior knowledge about the population parameters, to analyze the data from the random sample and develop conclusion from the analysis.
- The Normal Distribution is most heavily utilized in developing the theoretical background for sampling theory.

Term Definitions

- A population is the collection of all members of a group.
- A sample is a portion of the population selected for analysis.
- A parameter is a numerical measure that describes a characteristic of a population.
- A statistic is a numerical measure that describes a characteristic of a sample.

Sampling Distribution

- Parameters are usually unknown.
- Use statistics to estimate parameters.
- The sampling distribution of a statistic is the probability distribution that specifies probabilities for the possible values the statistic can take.

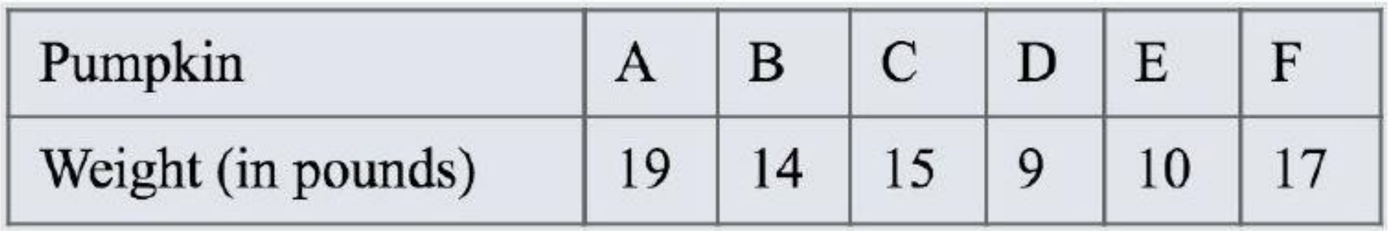
Sample Mean & Standard Deviation

- The sample mean is a statistics that varies from sample to sample. While population mean is a fixed value parameter.
- The estimate of the sample mean and standard deviation is given below.
- Note that the n-1 instead of n in the sample standard deviation is to ensure an unbiased estimate of the popular standard deviation.

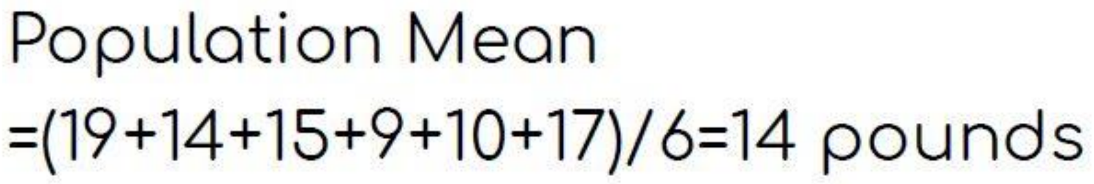
$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n} \qquad s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Example: Pumpkin Weights

- The population is the weight of six pumpkins displayed in a carnival “guess the weight” game booth.
- You are asked to guess the average weight of the six pumpkins by taking a random sample without replacement from the population.

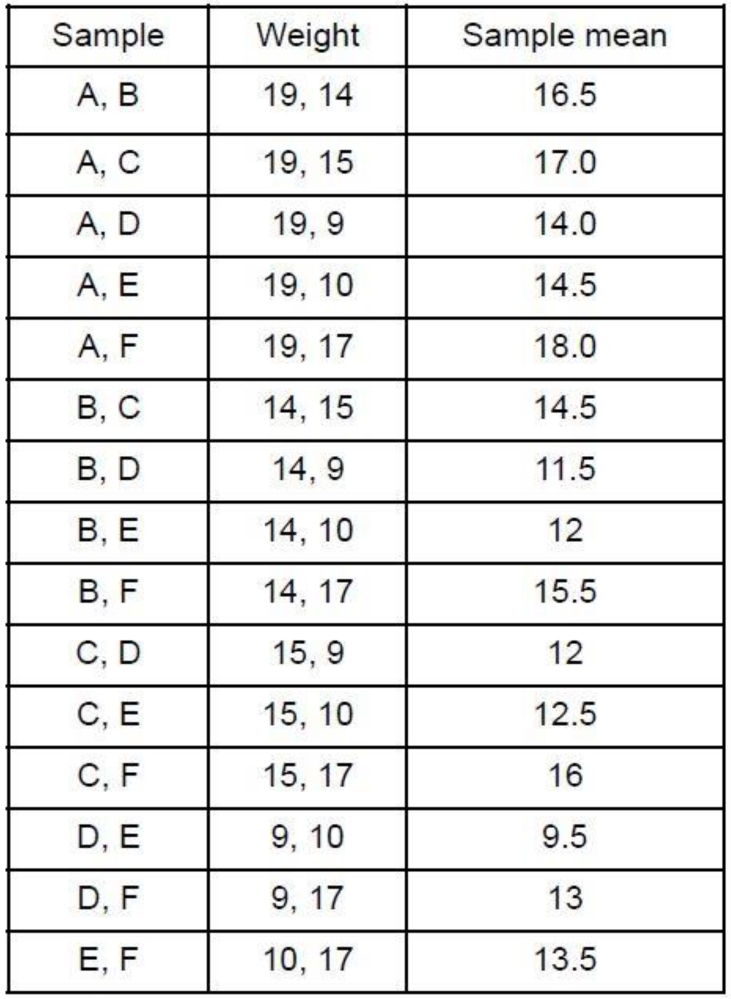


Pumpkin	A	B	C	D	E	F
Weight (in pounds)	19	14	15	9	10	17



Population Mean
=(19+14+15+9+10+17)/6=14 pounds

Example: Pumpkin Weights (sample size = 2)

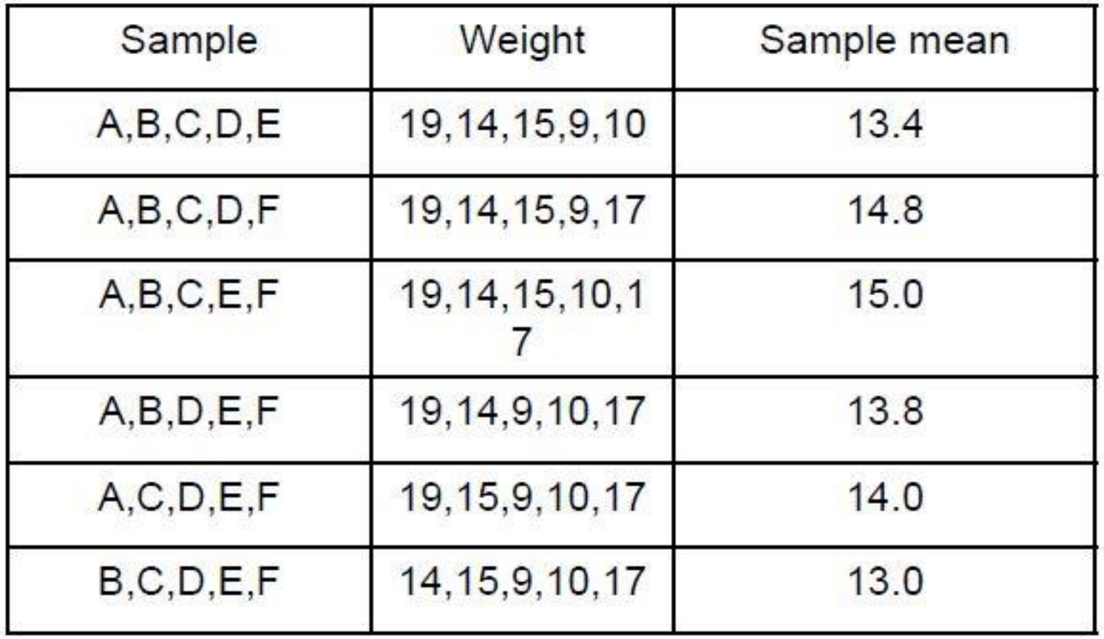


Sample	Weight	Sample mean
A, B	19, 14	16.5
A, C	19, 15	17.0
A, D	19, 9	14.0
A, E	19, 10	14.5
A, F	19, 17	18.0
B, C	14, 15	14.5
B, D	14, 9	11.5
B, E	14, 10	12
B, F	14, 17	15.5
C, D	15, 9	12
C, E	15, 10	12.5
C, F	15, 17	16
D, E	9, 10	9.5
D, F	9, 17	13
E, F	10, 17	13.5

- When using the sample mean to estimate the population mean, some possible error will be involved since sample mean is random.
- The chance that the sample mean is exactly the population mean is only 1/15.

Example: Pumpkin Weights (sample size = 5)

- The chance that the sample mean is exactly the population mean is only 1/6.
- The error with a sample of size 5 is on the average smaller than with a sample size 2.



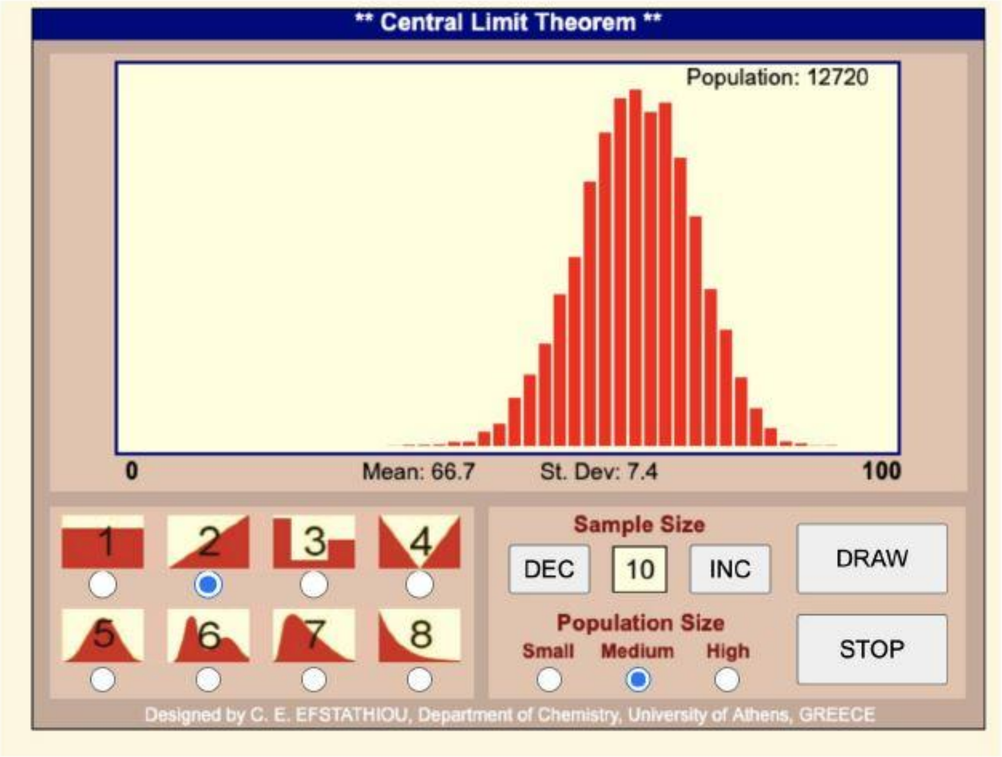
Sample	Weight	Sample mean
A,B,C,D,E	19,14,15,9,10	13.4
A,B,C,D,F	19,14,15,9,17	14.8
A,B,C,E,F	19,14,15,10,17	15.0
A,B,D,E,F	19,14,9,10,17	13.8
A,C,D,E,F	19,15,9,10,17	14.0
B,C,D,E,F	14,15,9,10,17	13.0

Central Limit Theorem

- For random sampling with a sample size “n”, the sampling distribution of the sample mean is approximately a normal distribution, no matter what the shape of the probability distribution from which the samples are taken.
- A rule of thumb for the sample size is more than 30. However, in most cases, sample size of 5 or more will work.
- The sample distribution standard deviation reduces with sample size.

Activity: Central Limit Theorem

- Try out eight different distributions, try a sample size of 5, 10, 20 and see the sample mean distribution.
- Link: http://195.134.76.37/applets/AppletCentralLimit/Appl_CentralLimit2.html



** Central Limit Theorem **

Population: 12720

Mean: 66.7 St. Dev: 7.4

○ 1 ● 2 ○ 3 ○ 4

○ 5 ○ 6 ○ 7 ○ 8

Sample Size: 10

Population Size: ○ Small ● Medium ○ High

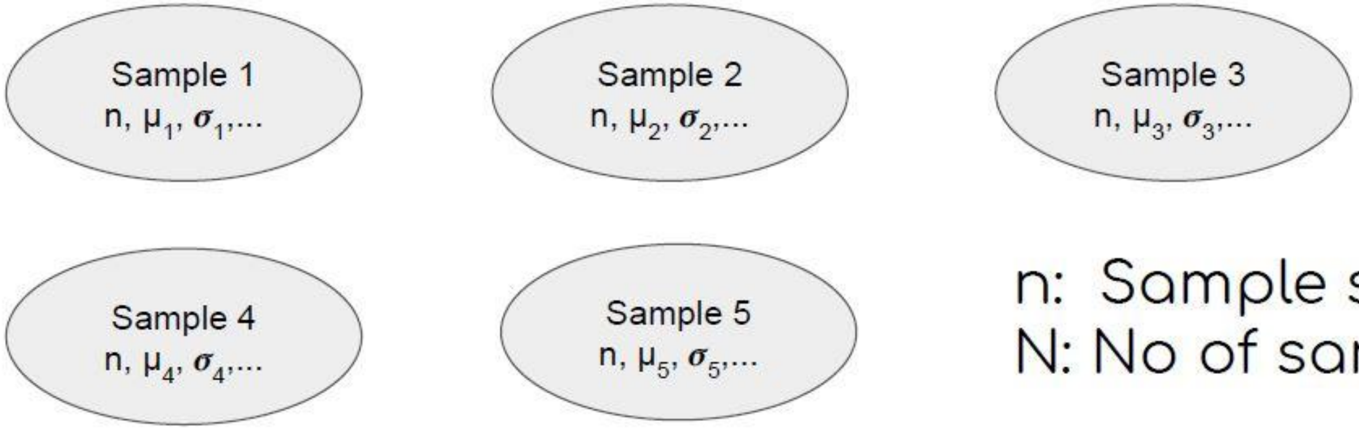
Designed by C. E. EFSTATHIOU, Department of Chemistry, University of Athens, GREECE

Applications of CLT

- Central Limit Theorem is used in a number of statistical areas such as:
    - Standard Error
    - Confidence Interval
    - Hypothesis Testing
    - ANOVA

Standard Error

- The standard error of a statistic is the standard deviation of the sampling distribution of that statistic (mean, standard deviation, mode, mediation)



Sample 1
n, μ_1, σ_1,...

Sample 2
n, μ_2, σ_2,...

Sample 3
n, μ_3, σ_3,...

Sample 4
n, μ_4, σ_4,...

Sample 5
n, μ_5, σ_5,...

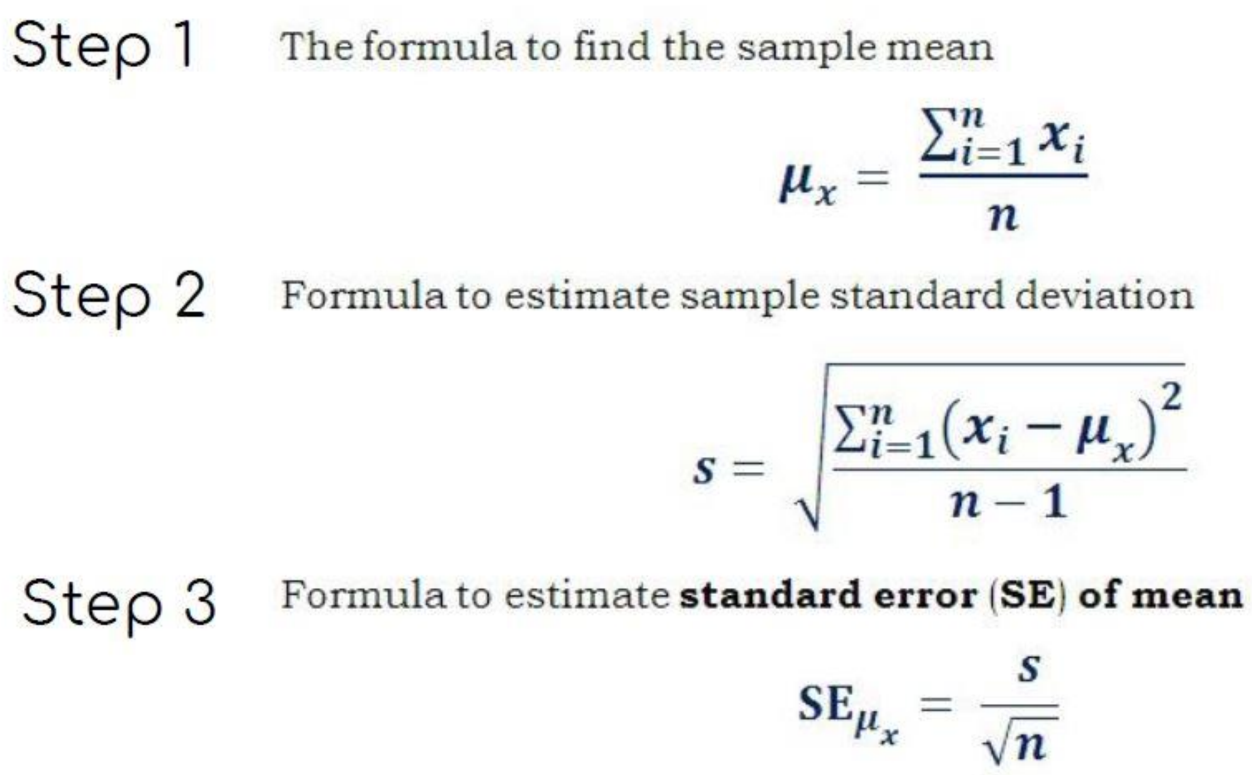
n: Sample size
N: No of samples

Standard Error (SE) of mean = $\dfrac{\sqrt{\Sigma\, \mu_i - \overline{\mu}}}{N}$

Standard Error (SE) of standard deviation = $\dfrac{\sqrt{\Sigma\, \sigma_i - \overline{\sigma}}}{N}$

Estimate Standard Error

- It is common to estimate the standard error from just one sample. One method is to use bootstrapping method, another method is to compute the standard error based on Central Limit Theorem (CLT) as follows:



Step 1 The formula to find the sample mean

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

Step 2 Formula to estimate sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu_x)^2}{n - 1}}$$

Step 3 Formula to estimate **standard error (SE) of mean**

$$\mathrm{SE}_{\mu_x} = \frac{s}{\sqrt{n}}$$

Standard Error in Excel

- Standard error in Excel can be computed as follows:
- STDEV (sampling range) / SQRT (COUNT(sampling range)).

Activity: Standard Error

Consider the following three sets of observations:

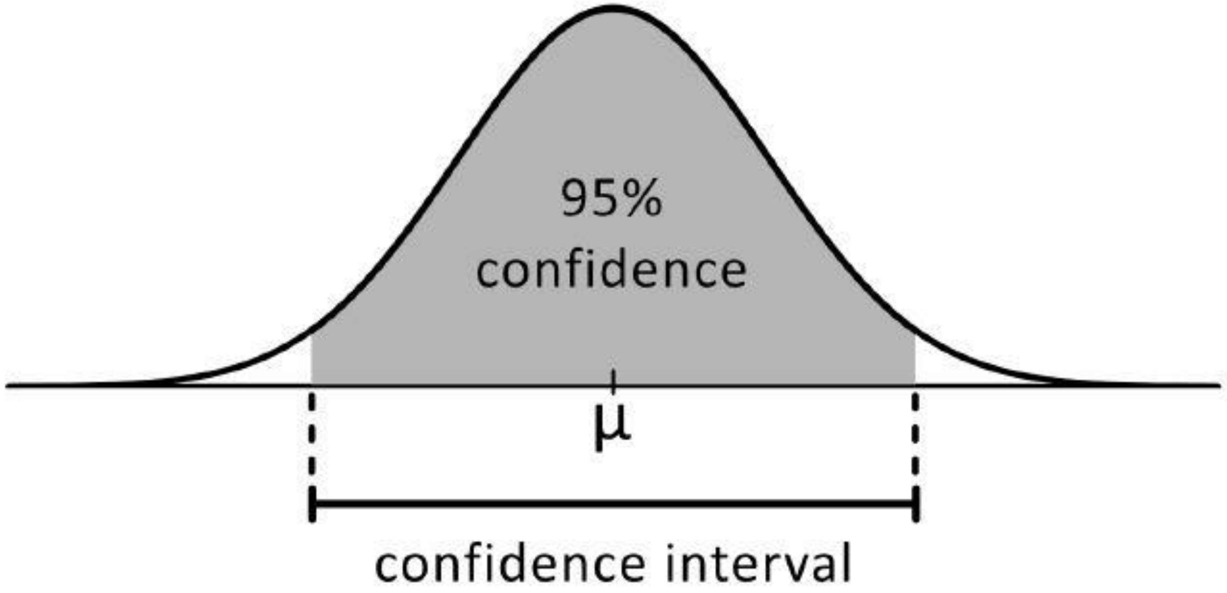
- Set 1: 8, 9, 10, 11, 12
- Set 2: 8, 9 , 10, 11, 100
- Set 3: 8, 9, 10, 11, 1000

- Compute the above standard errors.
- Verify the standard errors using the following online tool.

Link: https://ncalculators.com/statistics/standard-error-calculator.htm

Confidence Interval

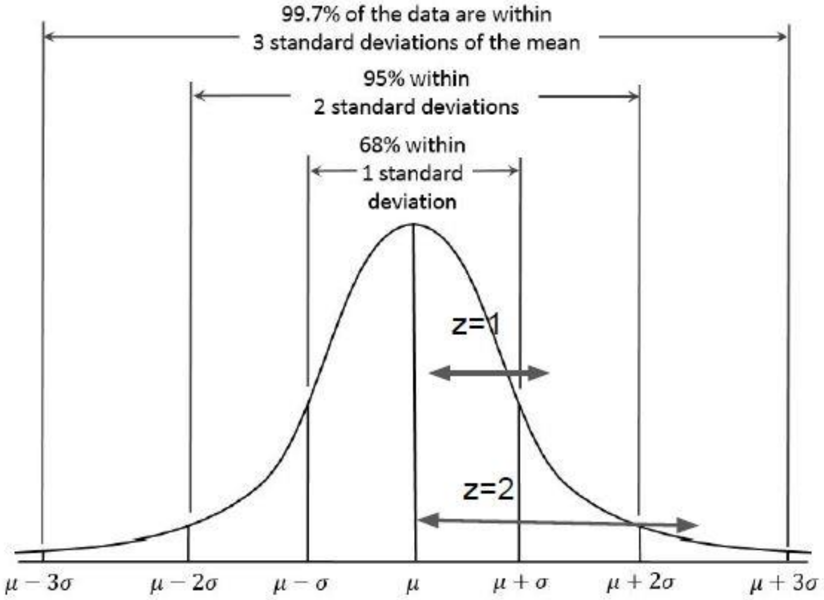
- A confidence Interval is a range of values we are fairly sure our true value lies in.
- This is a number chosen to be close to 1, most commonly 0.95.
- When the sampling distribution is approximately normally, a 95% confidence interval has margin of error equal to 1.96 standard errors.



Z-Score

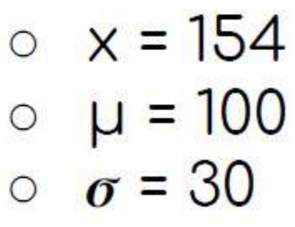
- The z-score for an observation is the number of standard deviations that it falss from the mean.
- The z-score is given by formula as below.

$$z = \frac{(x - \mu)}{\sigma} \quad \text{(population)} \quad z = \frac{(\overline{x} - \mu)}{\sigma / \sqrt{N}} \quad \text{(sample)}$$



Activity: Z-Score

Compute the z-score (population) given:

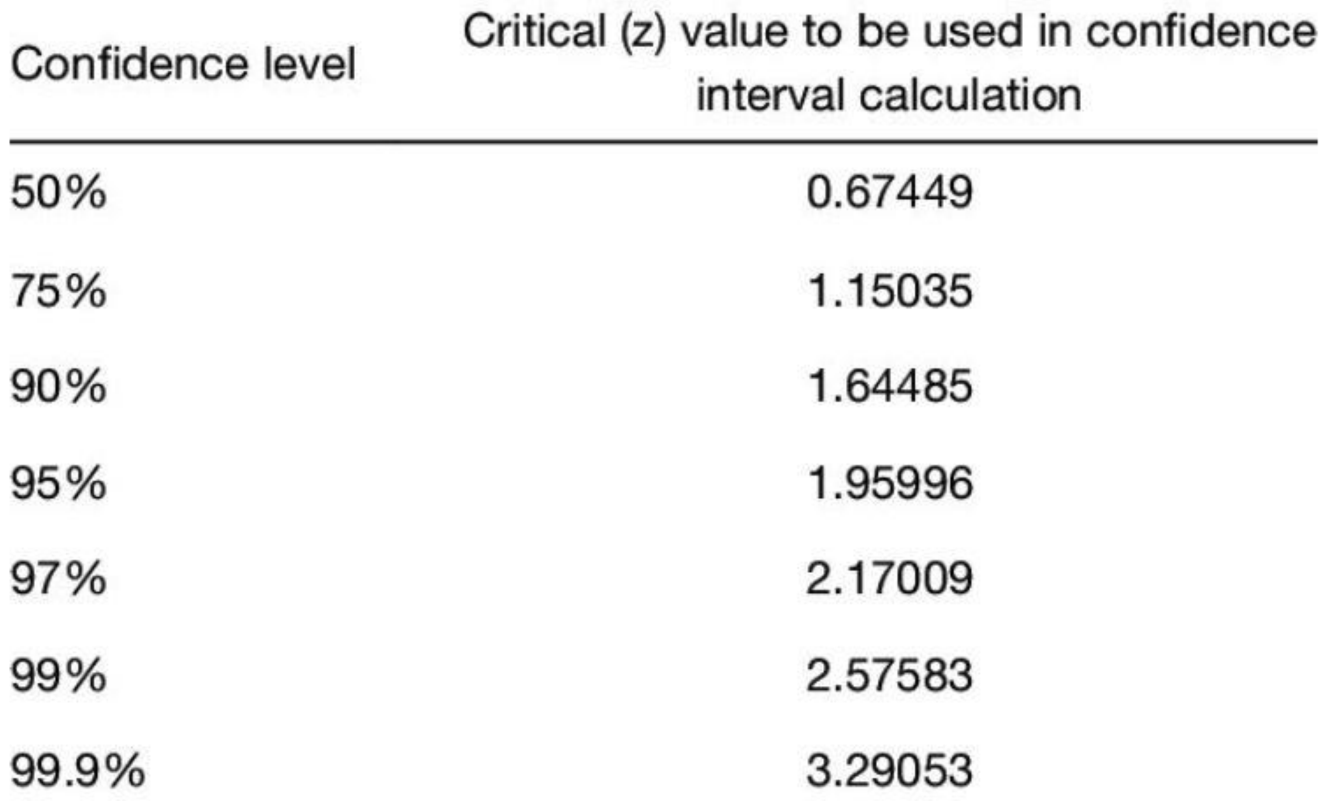


- ○ $x = 154$
- ○ $\mu = 100$
- ○ $\sigma = 30$

- Compute the z-score.
- Verify your answer with the online z-score calculator below.
- Link 1: http://www.learningaboutelectronics.com/Articles/Z-score-calculator.php
- Link 2: https://www.calculator.net/z-score-calculator.html

Confidence Interval with Z-Score

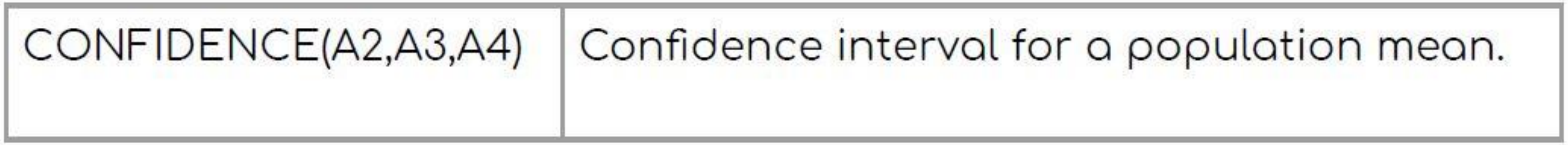
- The confidence interval for the population mean based on z-score, estimated from a sample from size “n” is:



Confidence level	Critical (z) value to be used in confidence interval calculation
50%	0.67449
75%	1.15035
90%	1.64485
95%	1.95996
97%	2.17009
99%	2.57583
99.9%	3.29053

Confidence Interval (Z-Score) in Excel

- The CONFIDENCE (alpha, sigma, n) function returns a value that you can use to construct a confidence interval for a population mean.
- The confidence interval is a range of values that are centered at a known sample mean.
- Observations in the sample are assumed to come form a normal distribution with known standard deviation, sigma and the number of observations in the sample is “n”.
- The syntax is CONFIDENCE(alpha, sigma, n).



CONFIDENCE(A2,A3,A4)	Confidence interval for a population mean.

Activity: Confidence Interval (Z-Score)

Suppose we know that the IQ scores of all incoming college freshman are normally distributed with standard deviation of 15. We have a simple random sample of 100 freshmen and the mean IQ score for this sample is 120. Find a 90% confidence interval for the mean IQ score for the entire population of incoming college freshmen.

- Compute the confidence interval.
- Verify your answer with confidence interval tools.
- Link 1: https://www.mathsisfun.com/data/confidence-interval-calculator.html
- Link 2: https://www.socscistatistics.com/confidenceinterval/default3.aspx

Confidence Interval with T-Score

- You use the t-score for small sample size (N<30) or unknow population standard deviation.
- The t-score is computed by:

$$t = \frac{(\overline{x} - \mu)}{s/\sqrt{N}}$$

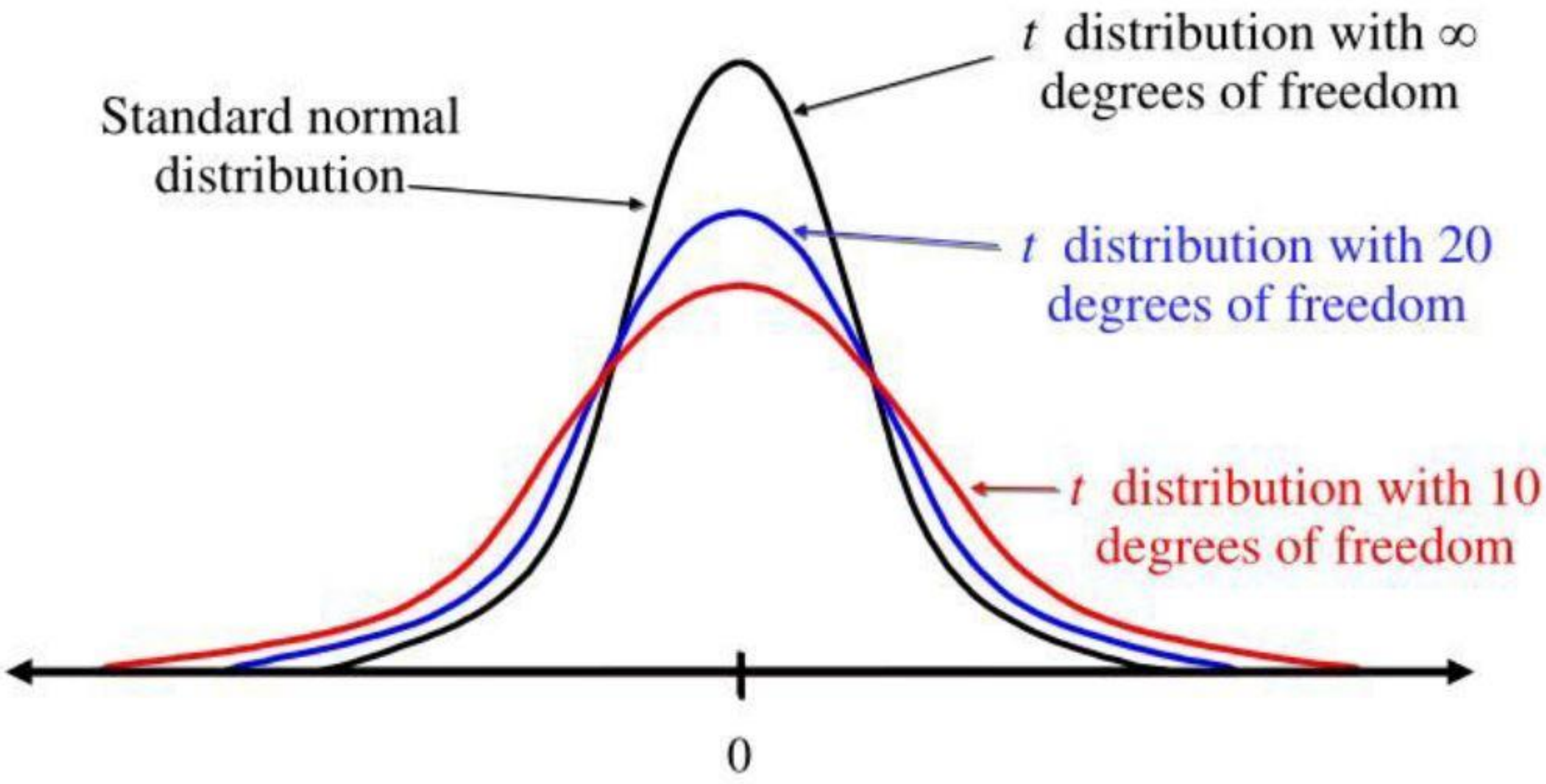
- Traditionally we look up a t-values in a t-table. The number of items in your sample, minus one, is your degrees of freedom. For example, if you have 20 items in your sample, then df=19.
- The confidence interval for the population mean based on t-score is:

$$\overline{x} \pm t \frac{s}{\sqrt{N}}$$

Student’s t-Distribution

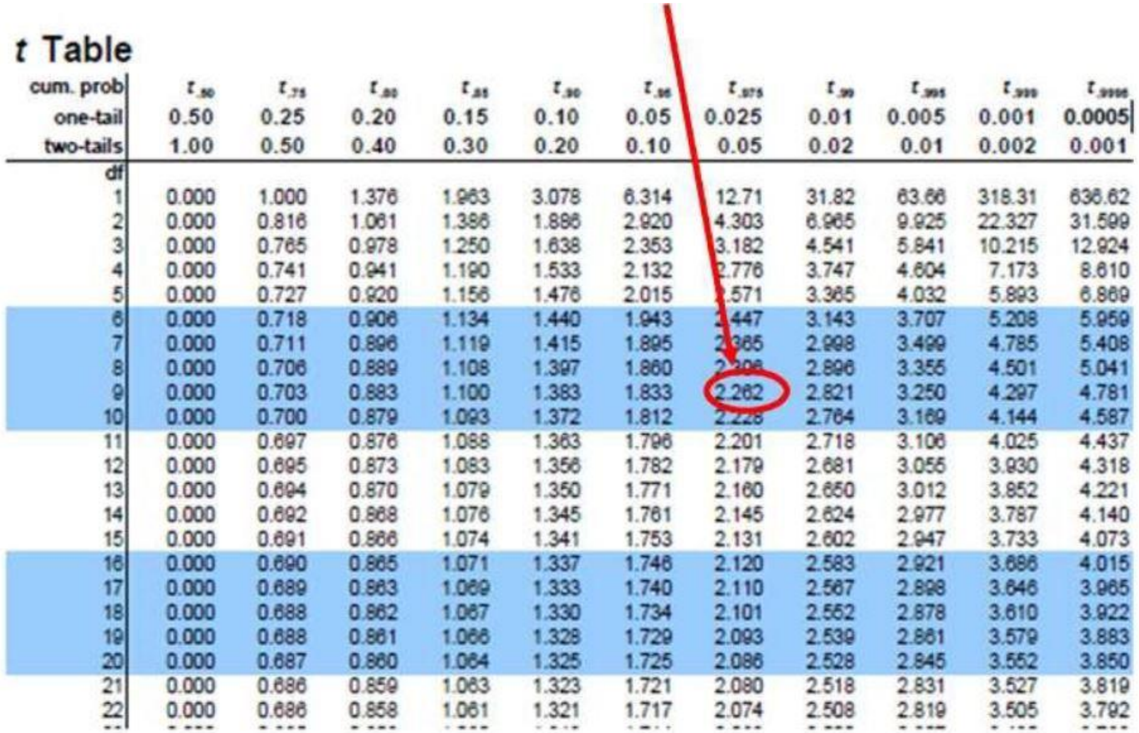
Forward School

The t-distribution is used when n is **small** and σ is **unknown**.



Confidence Interval using T-score

- Because the sample size is small, we need to use the t-distribution. For 95% confidence and df=n-1 is 9, t = 2.262

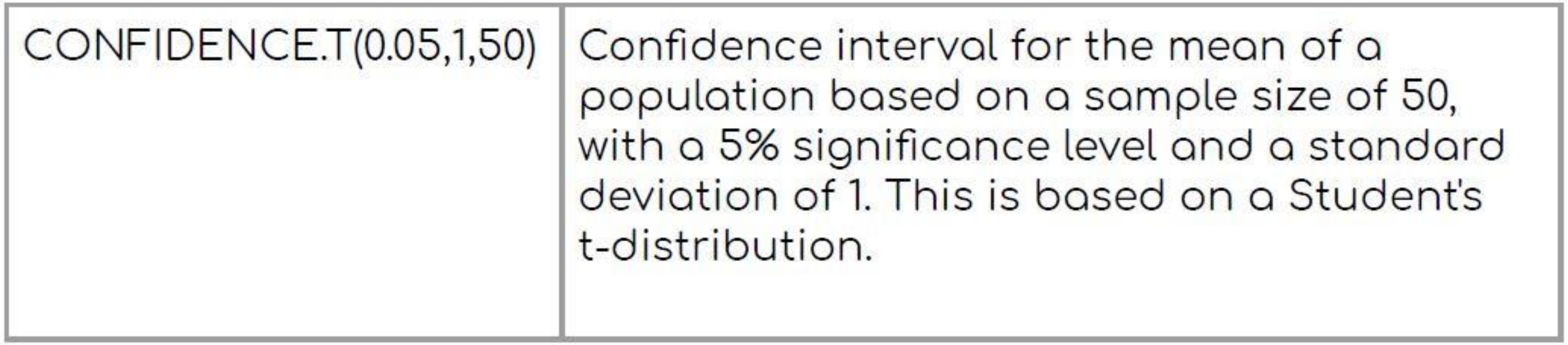


***t* Table**

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792

Confidence Interval (T-Score) in Excel

- The CONFIDENCE.T function returns the confidence interval for a population mean, using a student’s t-Distribution.



CONFIDENCE.T(0.05,1,50)	Confidence interval for the mean of a population based on a sample size of 50, with a 5% significance level and a standard deviation of 1. This is based on a Student's t-distribution.

Activity: T-Value

Forward School

- Verify the T-value for 95% confidence for a sample size of 10 is consistent with the T-value table using the following online t-value tool.
- Link 1: http://www.learningaboutelectronics.com/Articles/T-value-calculator.php
- Link 2: https://goodcalculators.com/student-t-value-calculator/

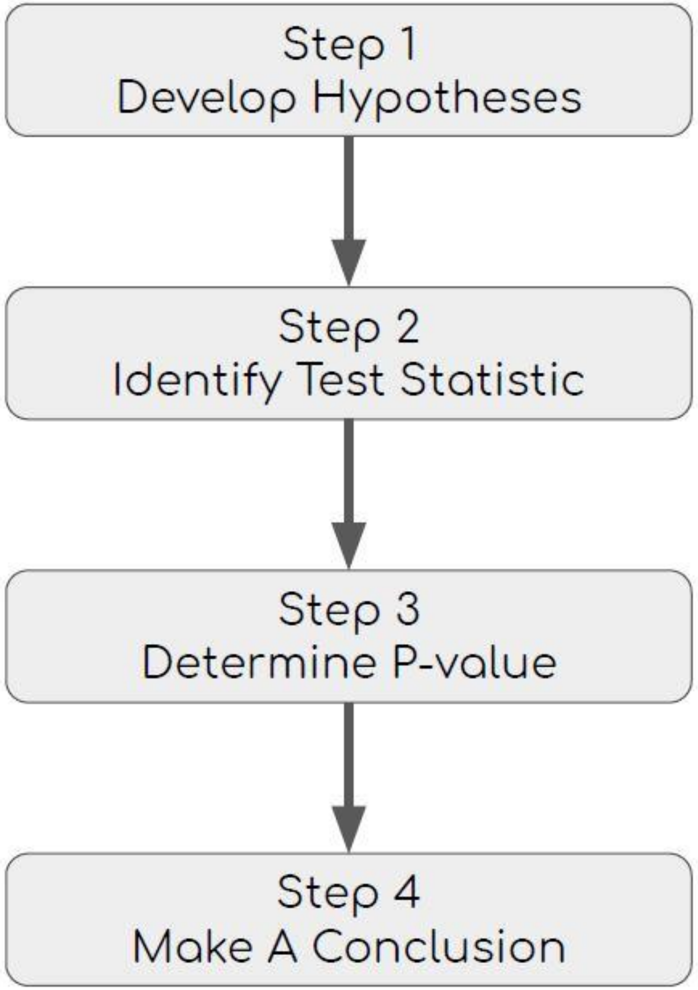
Activity: Confidence Interval (T-Score)

- We have a small random sample of 10 students from the IQ scores of all PSLE students. The mean IQ score for this sample is 120 and sample standard deviation is 15.
- Find a 90% confidence interval for the mean IQ score for the entire population of PSLE students.
- Compute the confidence interval using T-Score.
- Verify using the following confidence interval tools based on T-Score.
- Link: https://www.socscistatistics.com/confidenceinterval/default2.aspx

What is Hypothesis Testing

- A hypothesis is a statement about a population, usually of the form that a certain parameter takes a particular numerical value or falls in a certain range of values.
- The main goal is many research studies is to check whether the data support certain hypotheses.
- A hypothesis testing (significance test) is a method of using data to summarize the evidence about a hypothesis.

Steps of Hypothesis Testing



Step 1
Develop Hypotheses

Step 2
Identify Test Statistic

Step 3
Determine P-value

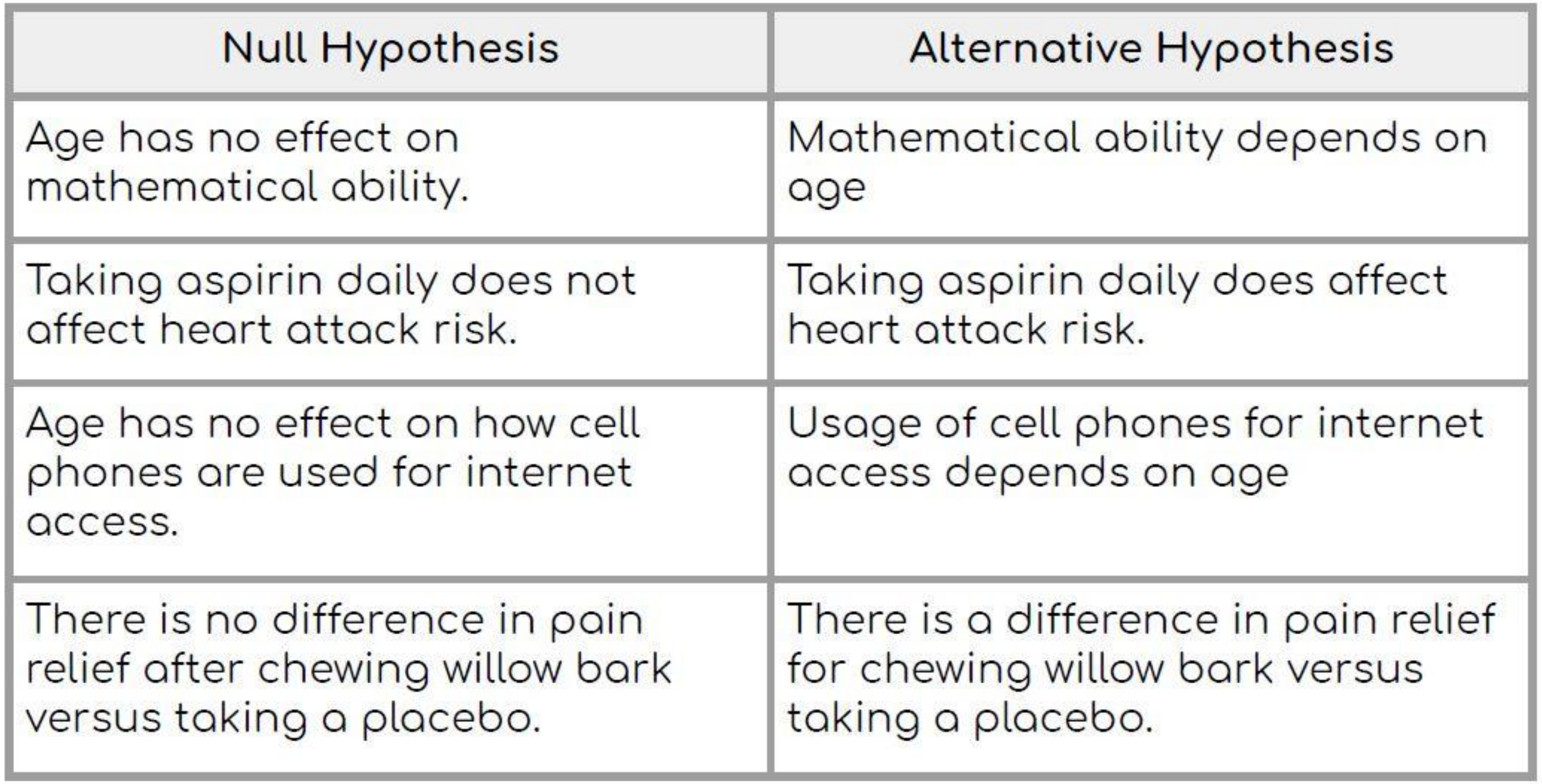
Step 4
Make A Conclusion

Step 1: Develop Hypotheses

Each significance test has two hypotheses:

- The null hypothesis states that:
  - A parameter takes a particular value.
  - A method has null effect (no effect).
- The alternative hypothesis states that:
  - A parameter falls in some alternative range of values.
  - A method has better or worst effect.
  - We usually set the hypothesis that one wants to conclude as the alternative hypothesis.

Examples of Hypothesis



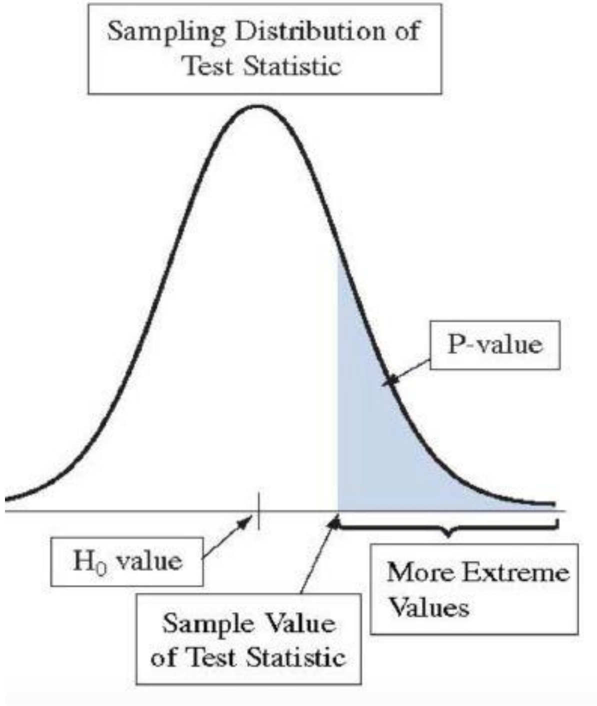
Null Hypothesis	Alternative Hypothesis
Age has no effect on mathematical ability.	Mathematical ability depends on age
Taking aspirin daily does not affect heart attack risk.	Taking aspirin daily does affect heart attack risk.
Age has no effect on how cell phones are used for internet access.	Usage of cell phones for internet access depends on age
There is no difference in pain relief after chewing willow bark versus taking a placebo.	There is a difference in pain relief for chewing willow bark versus taking a placebo.

Step 2: Identify Test Statistic

- A test statistic describes how far that estimate (the sample statistic) falls from the parameter value given in the null hypothesis.
- A test statistic is either z-score or t-score.
- In most cases, t-score is used as the sample size is small and the population variance is unknown.

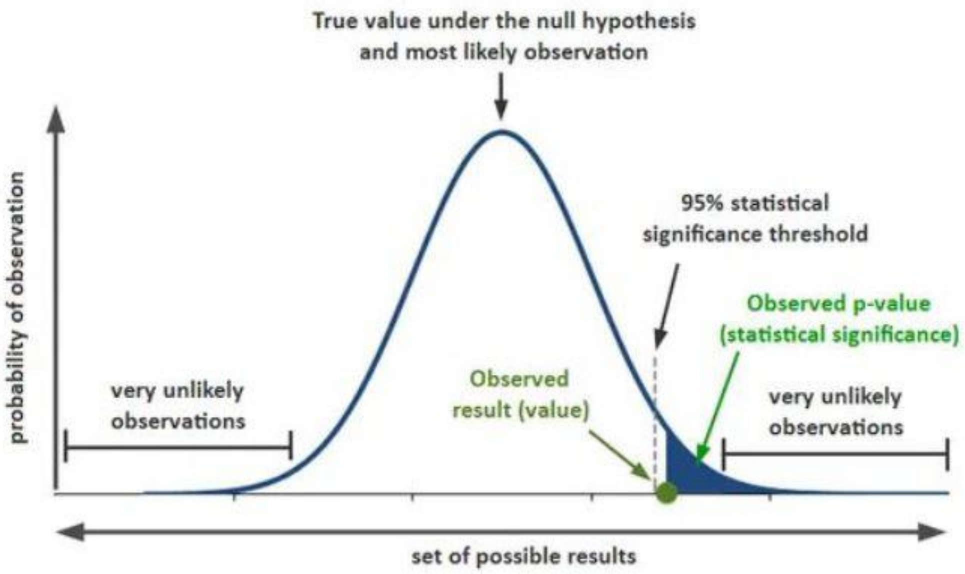
Step 3: P-value for T-Statistic

- The P-value is the probability that the test statistic equals the observed value or a value even more extreme.
- The smaller the P-value, the stronger the evidence is against null hypothesis.


Sampling Distribution of Test Statistic
P-value
H_0 value
Sample Value of Test Statistic
More Extreme Values

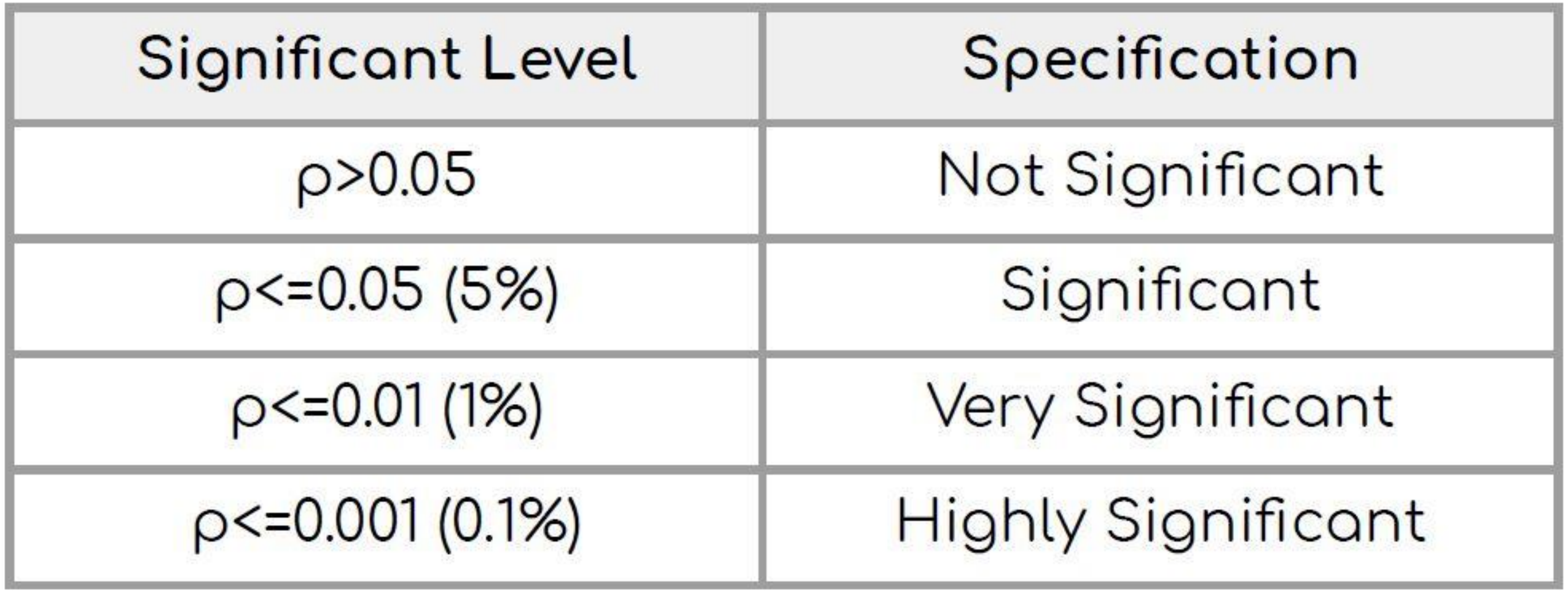
Significance Level

- The level of statistical significance is often expressed as a p-value between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.
- A p-value less than 0.05 (typically <= 0.05) is statistically significant.



Significance Level

- In practice, the most common significance level is 0.05



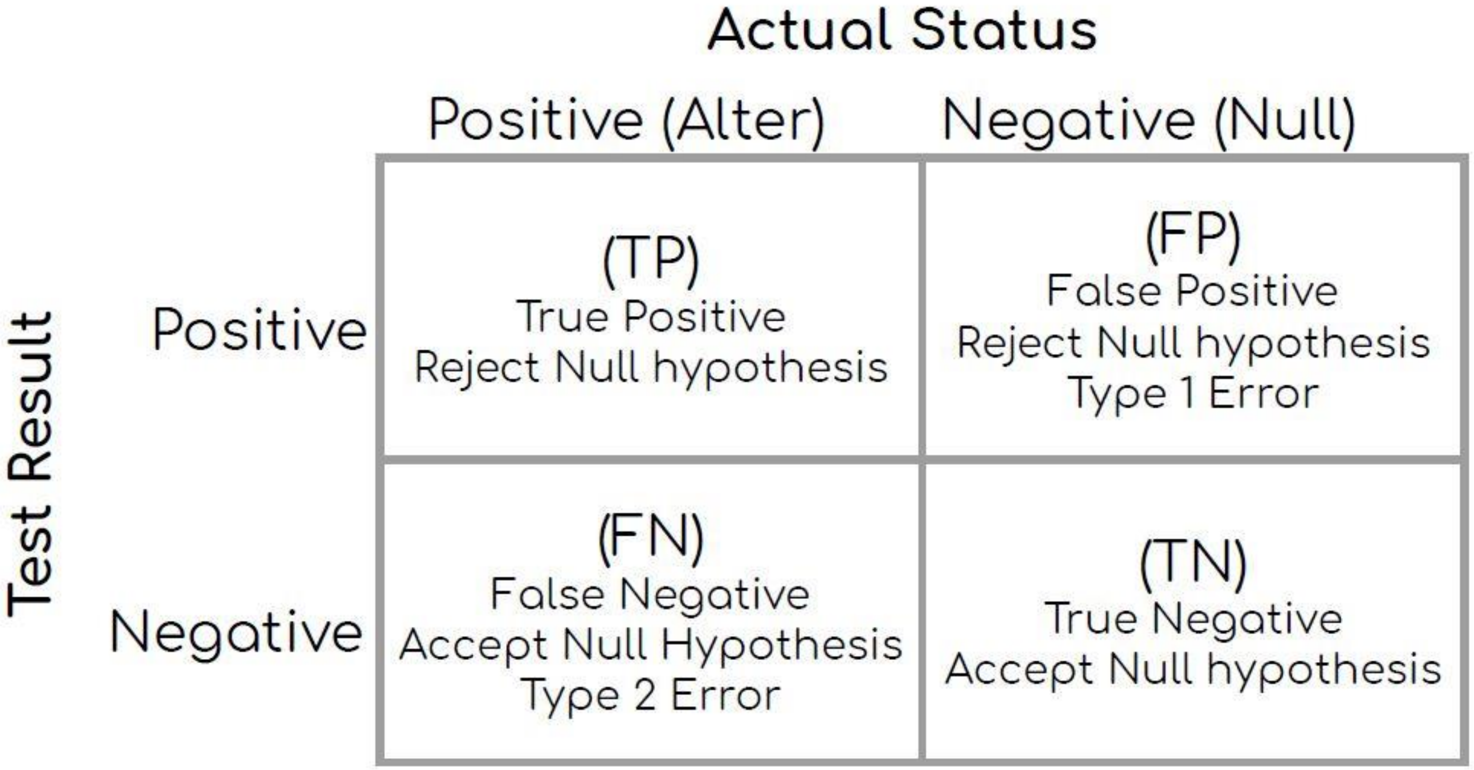
Significant Level	Specification
p>0.05	Not Significant
p<=0.05 (5%)	Significant
p<=0.01 (1%)	Very Significant
p<=0.001 (0.1%)	Highly Significant

Step 4: Make Conclusion

- Compare P-value with significance level.
- If the P-value < significance level, then reject the null hypothesis and accept the alternative hypothesis.

Type I and Type II Errors

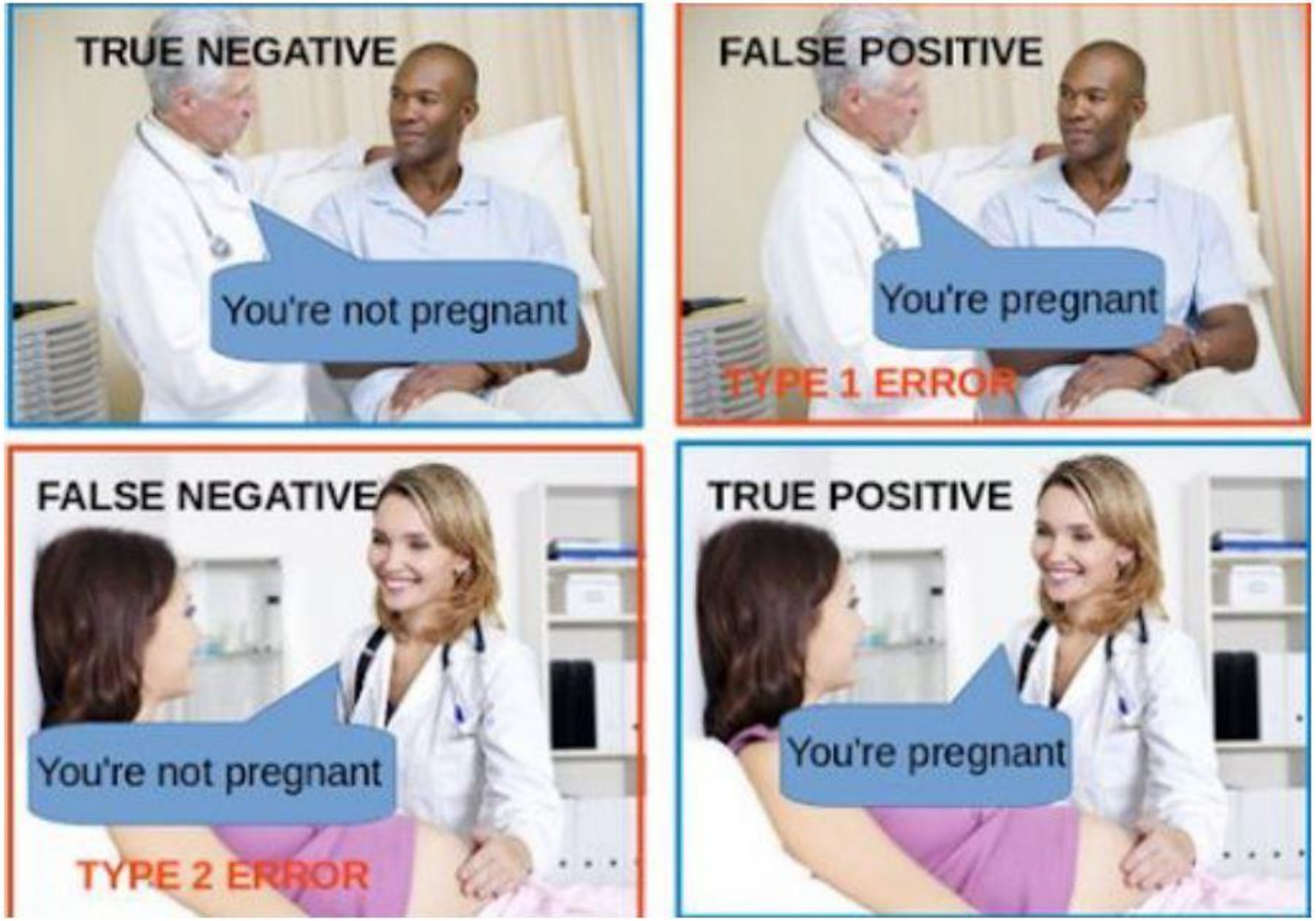
- Null Hypothesis => Negative
- Alternative Hypothesis => Positive



		Actual Status	
		Positive (Alter)	Negative (Null)
Test Result	Positive	(TP) True Positive Reject Null hypothesis	(FP) False Positive Reject Null hypothesis Type 1 Error
	Negative	(FN) False Negative Accept Null Hypothesis Type 2 Error	(TN) True Negative Accept Null hypothesis

Example of Type I and Type II Errors

Forward School

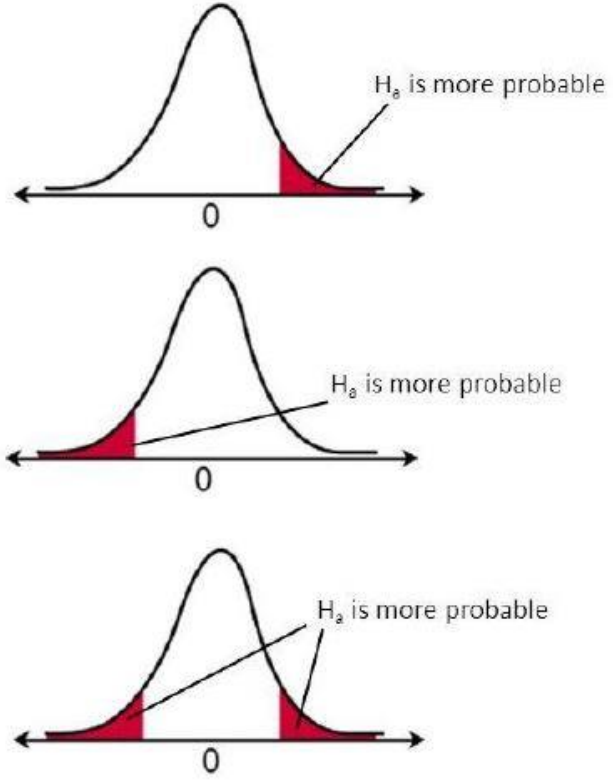


Types of T-Tests

- One sample t-test: Compare a sample mean to a hypothetical mean.
- One sample paired t-test: Compare the difference from the same sample before and after treatment.
- Two sample t-test: Compare two sample means from two population with equal variances.
- Two sample pooled variance t-test: Compare two sample means from two population with unequal variance.
- It is recommended to use a two sample t-test with unequal variance as we don’t need to make assumption on the data.

One Tails vs Two Tail T-Tests

- One tail t-test assumes the mean is higher or lower than a value.
- It is recommended to use a two tail test as we don’t need to make assumption on the data.



Right-tail test

H_a: $\mu >$ value

Left-tail test

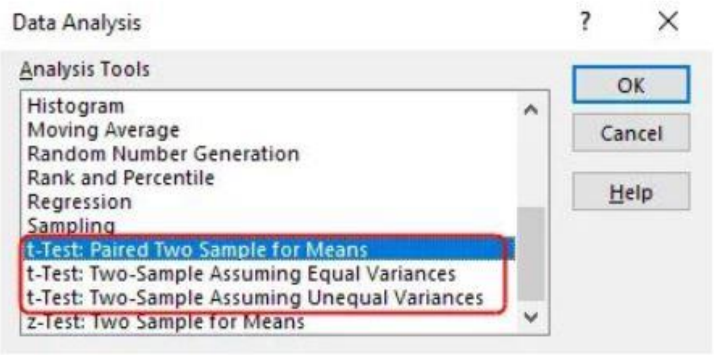
H_a: $\mu <$ value

Two-tail test

H_a: $\mu \neq$ value

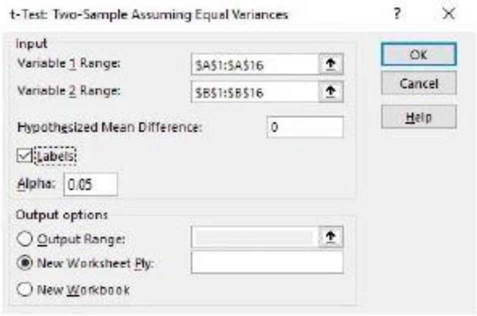
Hypothesis Testing in Excel

- In Excel, click Data Analysis on the Data tab.
- From the Data Analysis pop up, choose t-Test: Two-Sample Assuming Equal Variances.
- Under Input, select the ranges for both Variable 1 and Variable 2.
- In the Hypothesized Mean Difference, you'll typically enter zero. This value is the null hypothesis value, which represents no effect. In this case, a mean difference of zero represents no difference between the two methods which is no effect.
- Check the Labels checkbox if you have meaningful variables labels in row 1. This option helps make the output easier to interpret. Ensure that you include the label row in step #3.
- Excel uses a default Alpha value of 0,05 which is usually a good value. Alpha is the significance level. Change this value only when you have a specific reason for doing so.
- Click OK.


Data Analysis

Analysis Tools

Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means


t-Test: Two-Sample Assuming Equal Variances

Input

Variable 1 Range: $A$1:$A$16

Variable 2 Range: $B$1:$B$16

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.05

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Activity: One Sample Hypothesis Test

- We know that the national average (population) on the PSLE scoring system is 554 with standard deviation of 99. Our sample of 90 students from ABC school had an average score of 568.
- Is the 14 points difference in averages enough to say that students in ABC school perform better than the national average at significance level 0.05?
- What is the ABC school average score is 590?
- Perform hypothesis testing for one sample.
- Verify using the one sample hypothesis testing.
- Link: http://www.learningaboutelectronics.com/Articles/Hypothesis-testing-calculator.php#answer

One Sample Paired T-test

- If you are studying the same group of students (one sample) before and after taking a special GRE preparation session, you can use one sample paired t-test.
- You can use the following online for paired T-test.
- Link: http://www.learningaboutelectronics.com/Articles/Paired-t-test-calculator.php

Two Samples T-test using Pooled Variance

- For two samples from two populations with different variance, the null hypothesis is given by:

$$H_o = \mu_1 - \mu_2$$

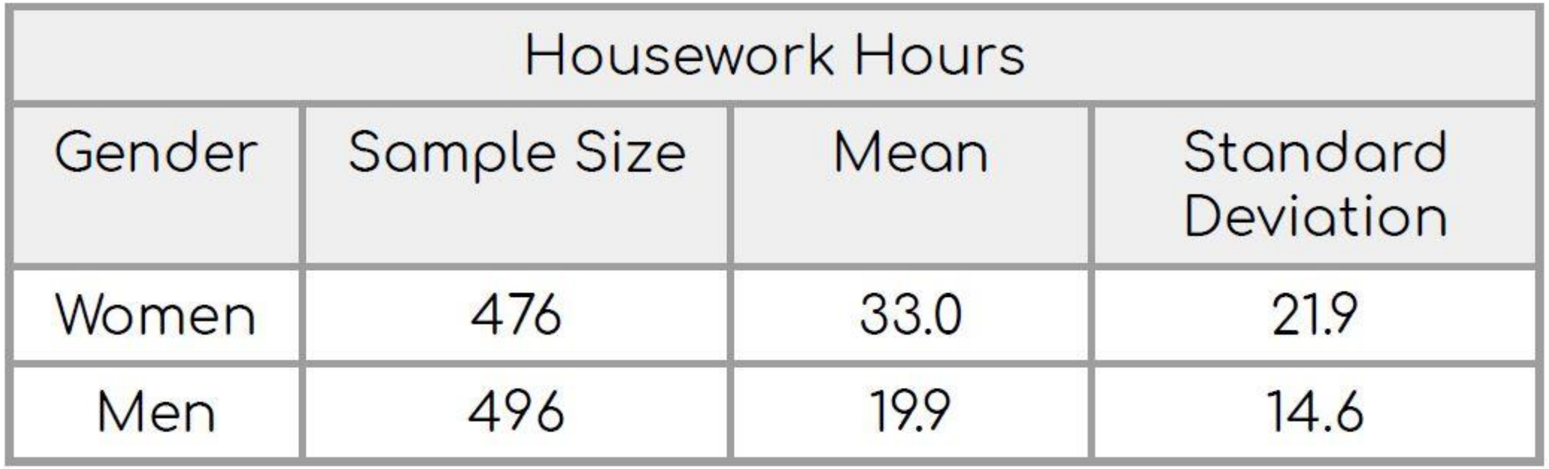
- The test statistics is given by:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- With degrees of freedom equal to n1+n2-2

Activity: Two Samples Hypothesis Test

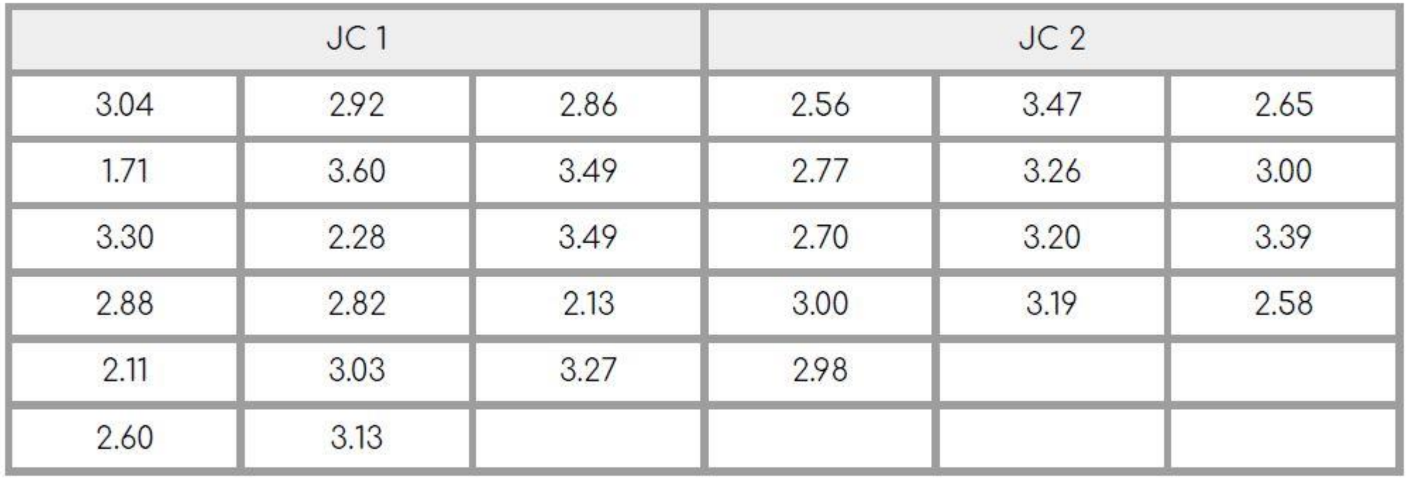
- Do women tend to spend more time on housework than men? If so, how much more?
- Perform hypothesis testing for one sample.
- You can use the following online tool for two samples hypothesis testing to analyze the above data.
- Link: https://www.statskingdom.com/140MeanT2eq.html



Housework Hours			
Gender	Sample Size	Mean	Standard Deviation
Women	476	33.0	21.9
Men	496	19.9	14.6

Activity: Two Samples Hypothesis Test

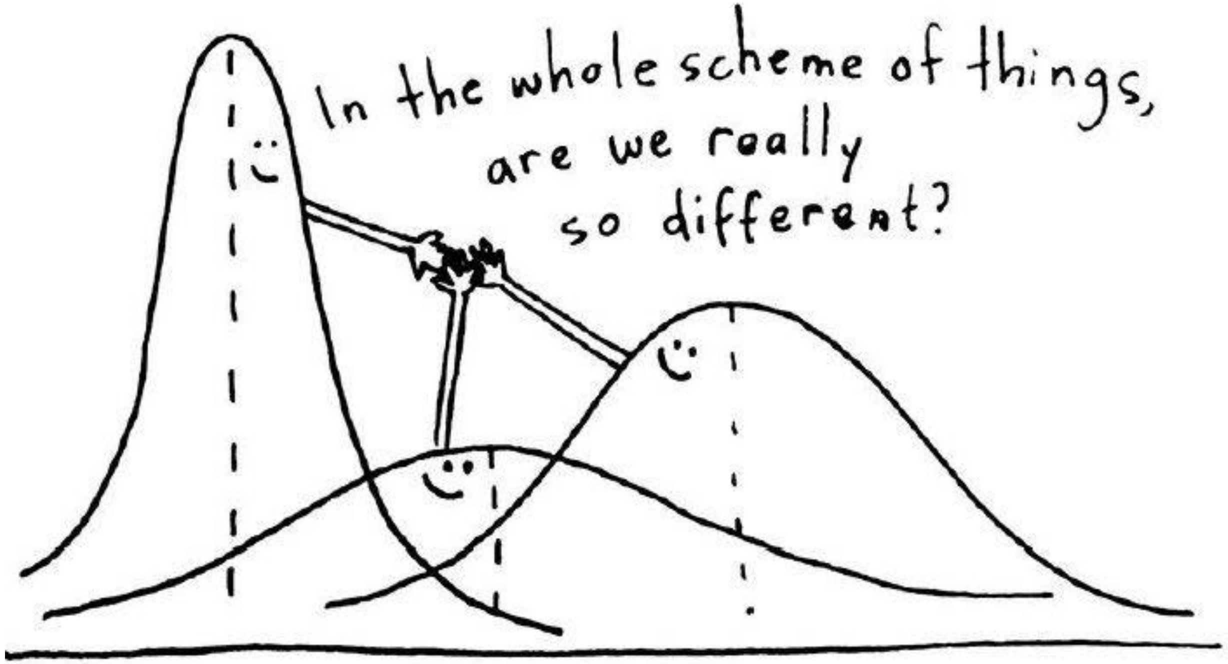
- Independent random samples of 17 students from JC1 and 13 students from JC2 yield the following grade points.
- Is there any difference in grade points between JC1 and JC2 students?
- Perform hypothesis testing for one sample.
- Verify with two samples hypothesis testing tool below:
- Link 1: https://www.socscistatistics.com/tests/studentttest/default2.aspx
- Link 2: https://ncalculators.com/statistics/t-test-calculator.htm
- Link 3: http://www.learningaboutelectronics.com/Articles/Unpaired-t-test-calculator.php



JC 1			JC 2		
3.04	2.92	2.86	2.56	3.47	2.65
1.71	3.60	3.49	2.77	3.26	3.00
3.30	2.28	3.49	2.70	3.20	3.39
2.88	2.82	2.13	3.00	3.19	2.58
2.11	3.03	3.27	2.98		
2.60	3.13				

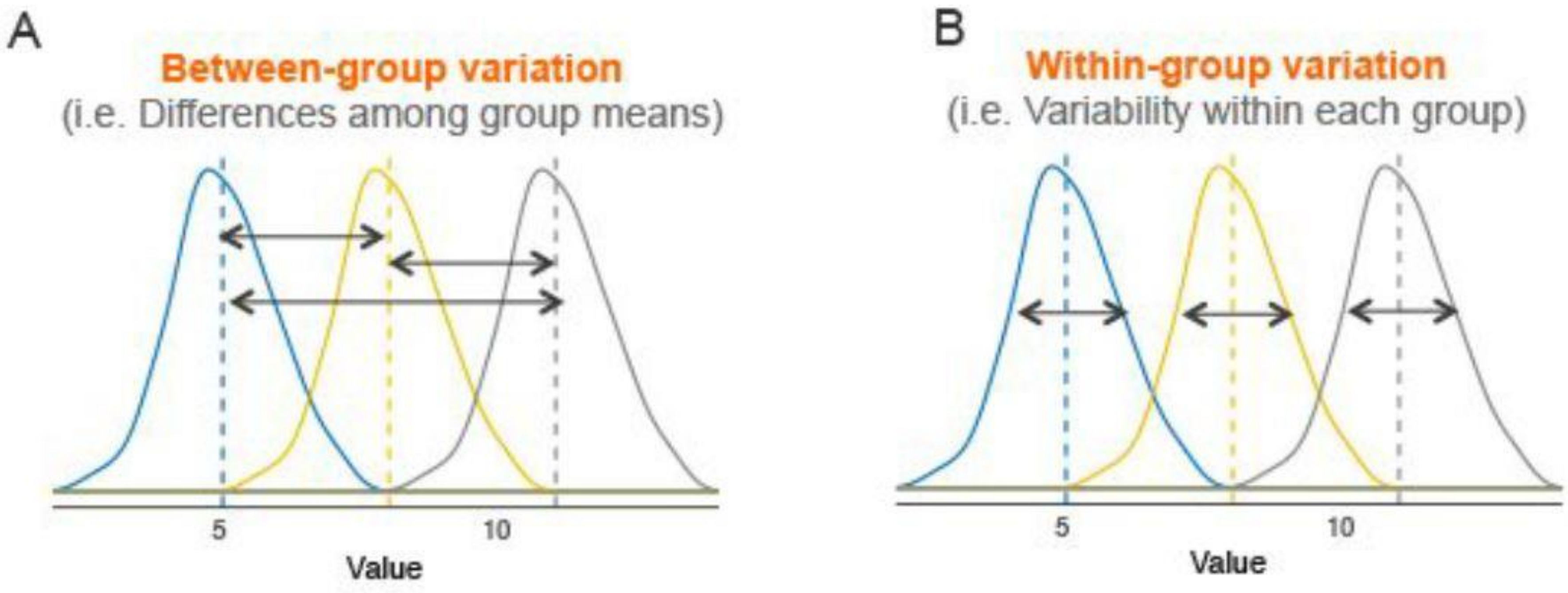
Analysis of Variance (ANOVA)

- Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.
- ANOVA check the impact of one or more factors by comparing the means of different samples.



Variability

- ANOVA takes into account between group variation and within group variation.
- Total variation = between-group variation + within-group variation


A
Between-group variation
(i.e. Differences among group means)
5
10
Value
B
Within-group variation
(i.e. Variability within each group)
5
10
Value

ANOVA Hypothesis

- Similar to t-test and chi-square test, ANOVA also uses a Null hypothesis and an Alternate hypothesis.
- The null hypothesis in ANOVA is valid when all the sample means are equal, or they don’t have any significant difference. Thus, they can be considered as a part of a larger set of the population.
- On the other hand, the alternate hypothesis is valid when at least one of the sample means is different from the rest of the sample means. In mathematical form, they can be represented as:

$$H_o : \quad \mu_1 = \mu_2 = \cdots = \mu_L \qquad \textit{Null hypothesis}$$

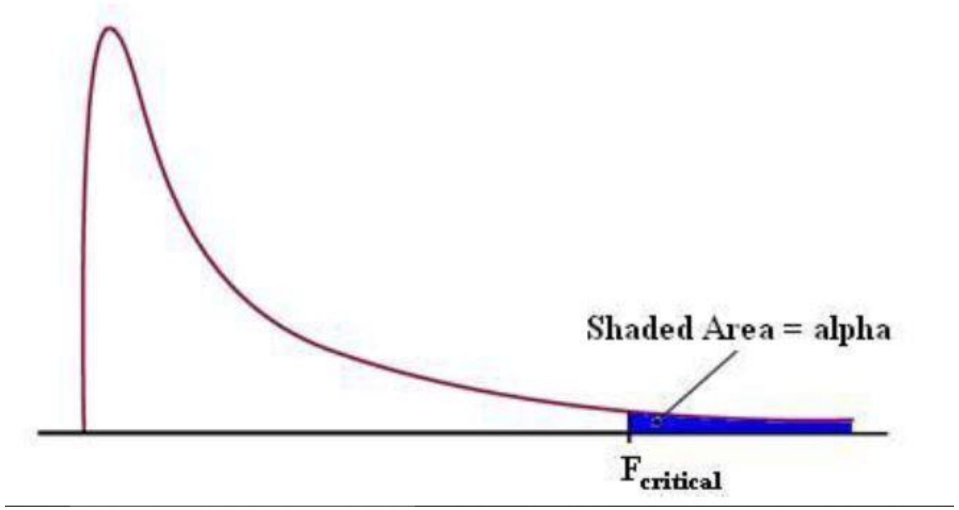
$$H_1 : \quad \mu_l \neq \mu_m \qquad \textit{Alternate hypothesis}$$

F Statistics

- The statistic which measures if the means of different samples are significantly different or not is called F-Ratio.
- Lower the F-Ratio, more similar are the sample means. In that case, we cannot reject the null hypothesis.
- F = Between group variability / Within group variability
- The numerator term in the F-statistic calculation defines the between-group variability. As we read earlier, as between group variability increases, sample means grow further apart from each other. In other words, the samples are more probable to be belonging to totally different populations.

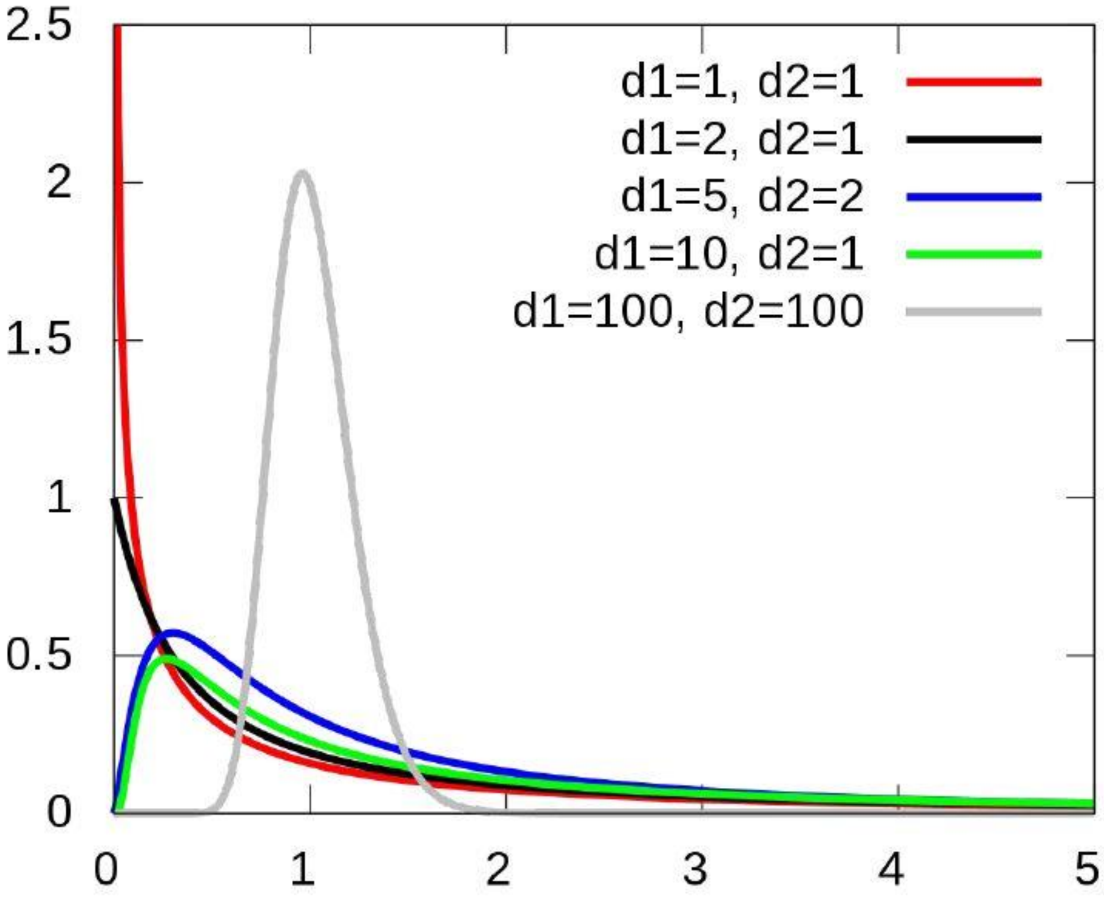
P-Value for F-Statistics

- This F-statistic calculated here is compared with the F-critical value for making a conclusion.
- If the value of the calculated F-statistic is more than the F-critical value, then we reject the null hypothesis.
- Alternatively, one can compute the p-value of F-statistics. If the p-value is less than the significance level, then we reject the null hypothesis.
- You can compute the critical F-Value from this online tool.
- Link 1: https://www.danielsoper.com/statcalc/calculator.aspx?id=4
- Link 2: https://www.danielsoper.com/statcalc/calculator.aspx?id=7



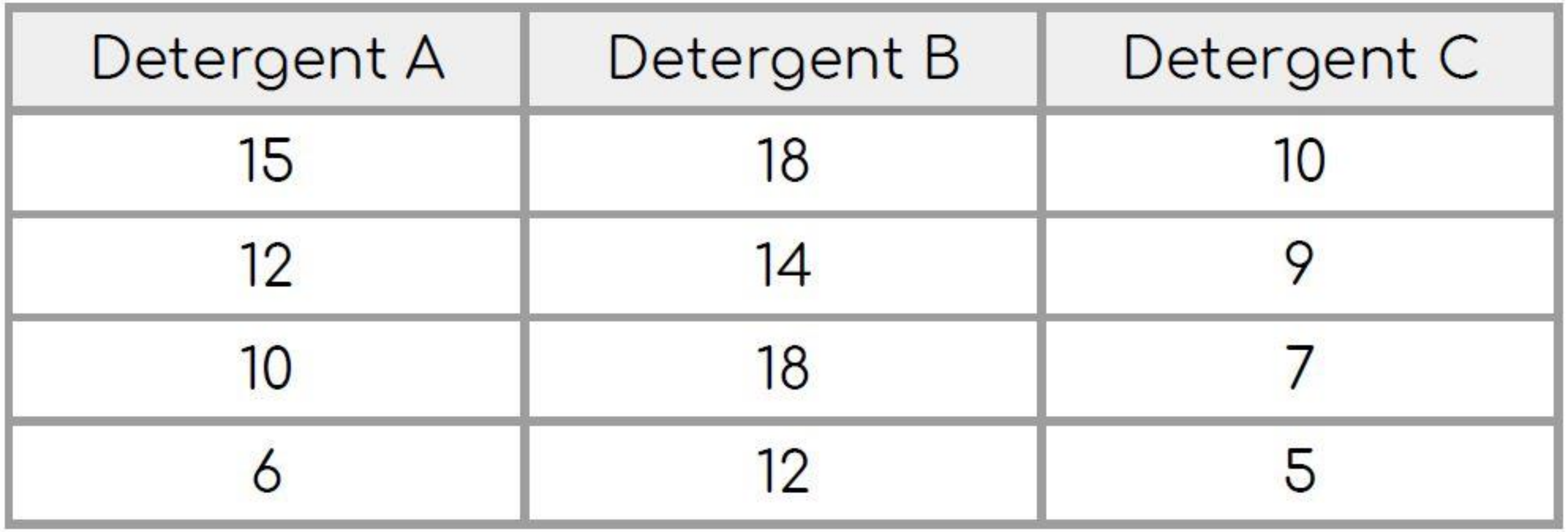
F-Distribution

- The F-Distribution is the probability distribution associated with the f-statistic.



One Way ANOVA

- The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups whilst considering only one independent variable or factor.
- Example of data for one-way ANOVA could be:

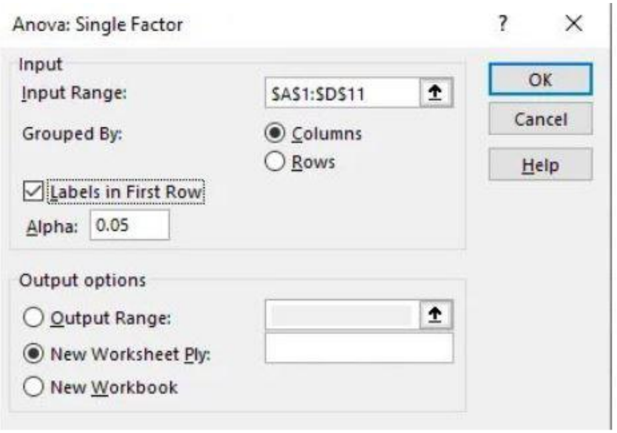


Detergent A	Detergent B	Detergent C
15	18	10
12	14	9
10	18	7
6	12	5

One Way ANOVA in Excel

Forward School

- Click Data Analysis on the Data tab.
- From the Data Analysis pop up. Choose ANOVA: Single Factor.
- Under Input, select the ranges for all columns of data.
- In Grouped By, choose Columns.
- Check the Labels checkbox if you have meaningful variables labels in row 1. This option helps make the out put easier to interpret. Ensure that you include the label row in step #3.
- Excel uses a default Alpha value of 0.05, which is usually a good value. Alpha is the significance level. Change this value only when you have a specific reason for doing so.
- Click OK.



Anova: Single Factor

Input

Input Range: $A$1:$D$11

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

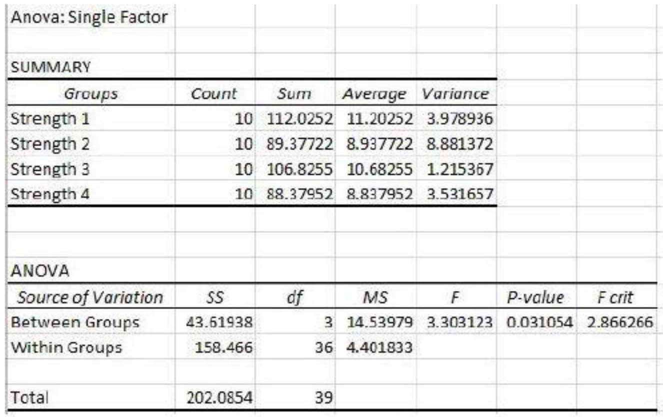
Alpha: 0.05

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook



Anova: Single Factor

SUMMARY

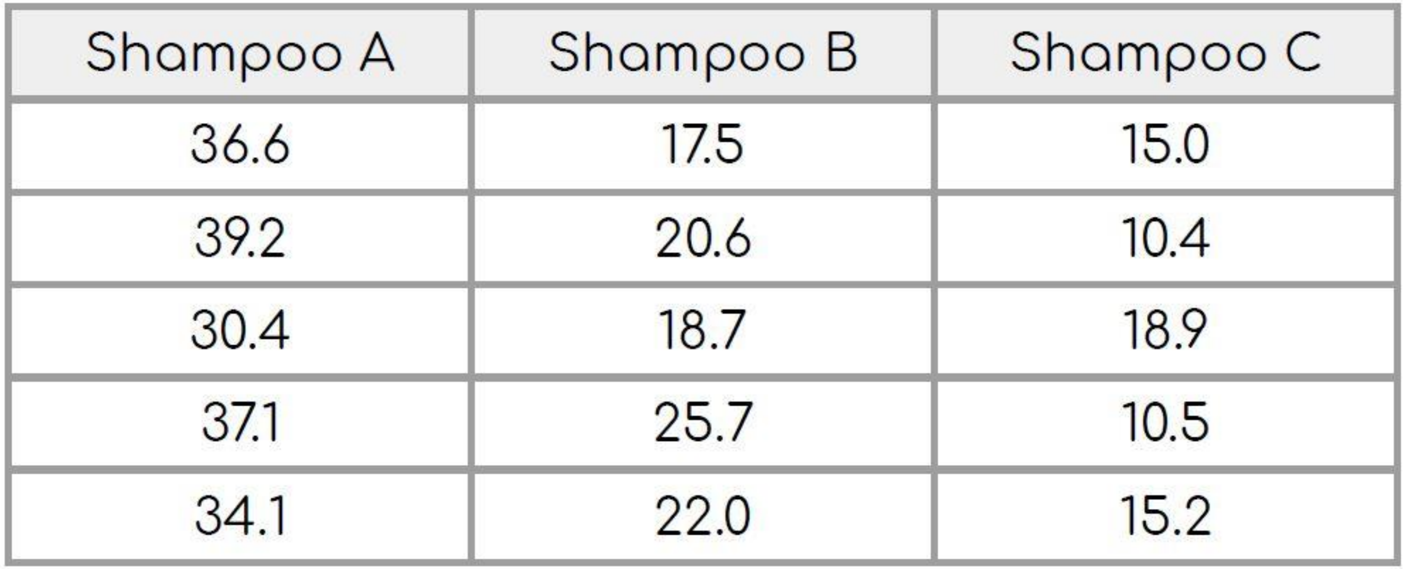
Groups	Count	Sum	Average	Variance
Strength 1	10	112.0252	11.20252	3.978936
Strength 2	10	89.37722	8.937722	8.881372
Strength 3	10	106.8255	10.68255	1.215367
Strength 4	10	88.37952	8.837952	3.531657

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	43.61938	3	14.53979	3.303123	0.031054	2.866266
Within Groups	158.466	36	4.401833			
Total	202.0854	39				

Activity: One Way ANOVA

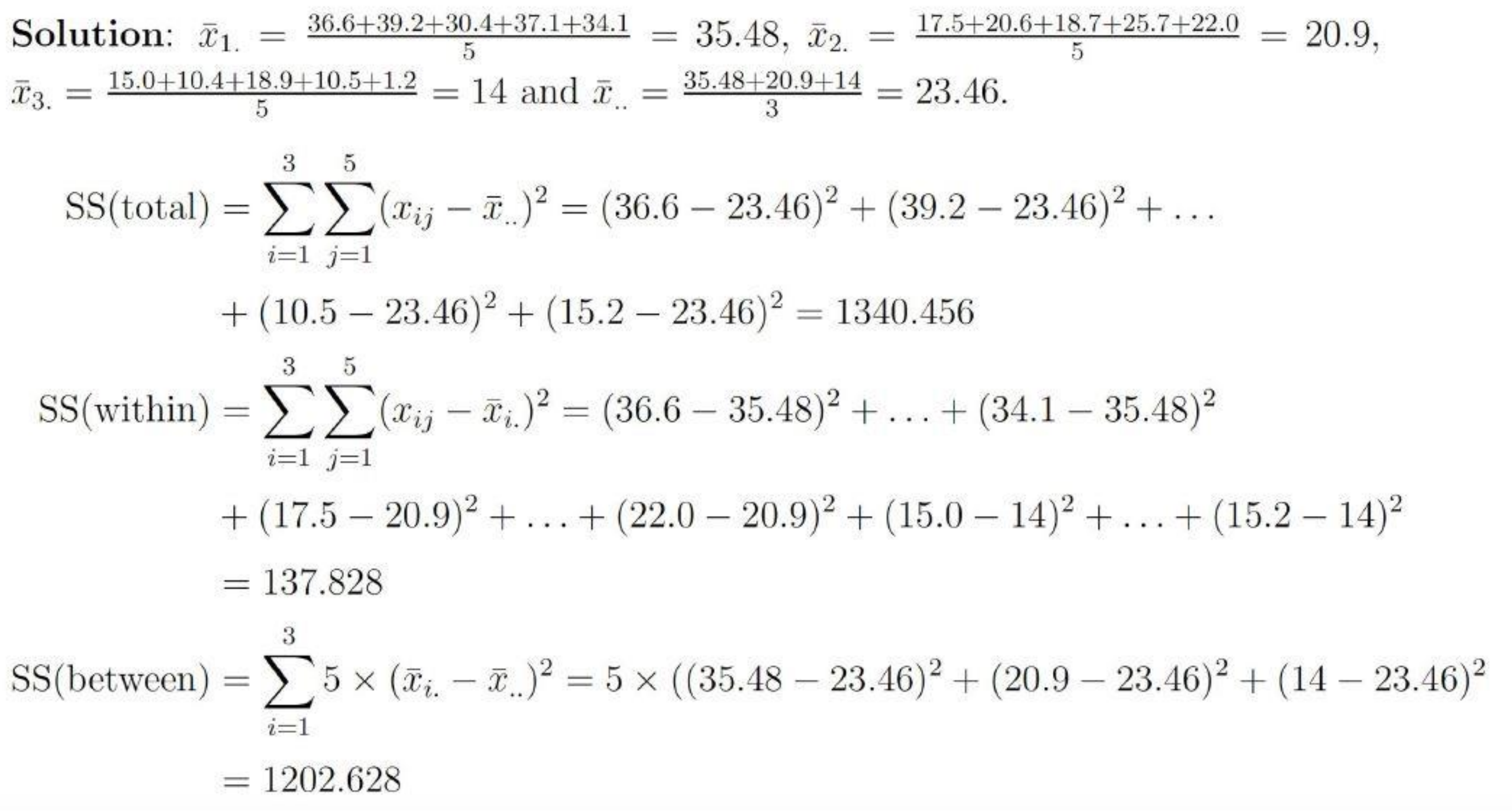
- Test the hypothesis of whether there is any significant difference among the 3 shampoos below using one-way ANOVA test.
- Perform the one-way ANOVA and verify using the following online tool.
- Link: https://www.socscistatistics.com/tests/anova/default2.aspx



Shampoo A	Shampoo B	Shampoo C
36.6	17.5	15.0
39.2	20.6	10.4
30.4	18.7	18.9
37.1	25.7	10.5
34.1	22.0	15.2

How Variances is Calculated

- SS: Sum of Square = Variance
- SS(total) = SS (within) + SS (between)



Solution: $\bar{x}_{1.} = \frac{36.6+39.2+30.4+37.1+34.1}{5} = 35.48,\ \bar{x}_{2.} = \frac{17.5+20.6+18.7+25.7+22.0}{5} = 20.9,$
$\bar{x}_{3.} = \frac{15.0+10.4+18.9+10.5+1.2}{5} = 14$ and $\bar{x}_{..} = \frac{35.48+20.9+14}{3} = 23.46.$

$$\mathrm{SS(total)} = \sum_{i=1}^{3}\sum_{j=1}^{5}(x_{ij} - \bar{x}_{..})^2 = (36.6 - 23.46)^2 + (39.2 - 23.46)^2 + \ldots$$

$$+ (10.5 - 23.46)^2 + (15.2 - 23.46)^2 = 1340.456$$

$$\mathrm{SS(within)} = \sum_{i=1}^{3}\sum_{j=1}^{5}(x_{ij} - \bar{x}_{i.})^2 = (36.6 - 35.48)^2 + \ldots + (34.1 - 35.48)^2$$

$$+ (17.5 - 20.9)^2 + \ldots + (22.0 - 20.9)^2 + (15.0 - 14)^2 + \ldots + (15.2 - 14)^2$$

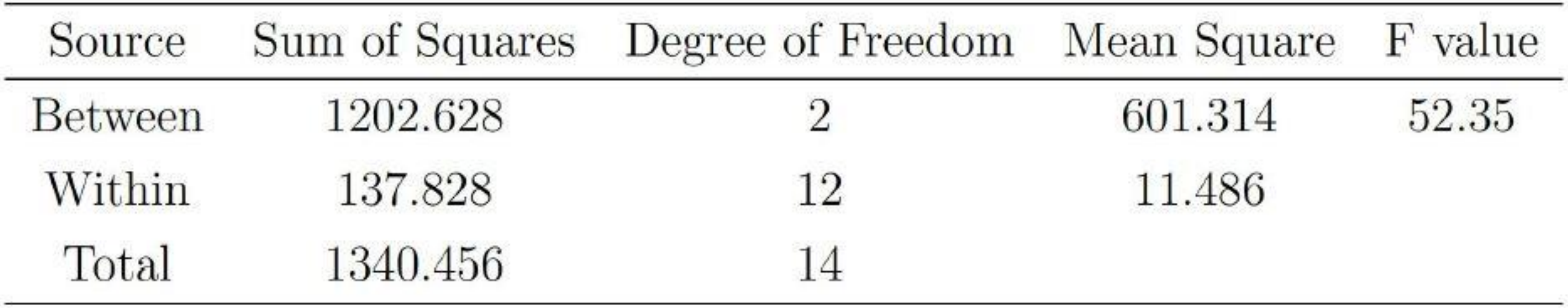
$$= 137.828$$

$$\mathrm{SS(between)} = \sum_{i=1}^{3} 5 \times (\bar{x}_{i.} - \bar{x}_{..})^2 = 5 \times ((35.48 - 23.46)^2 + (20.9 - 23.46)^2 + (14 - 23.46)^2$$

$$= 1202.628$$

ANOVA Table

- MS: Mean Squares = Sum of Square/DF
- F = MS (Between) / MS (Within)
- ANOVA table:

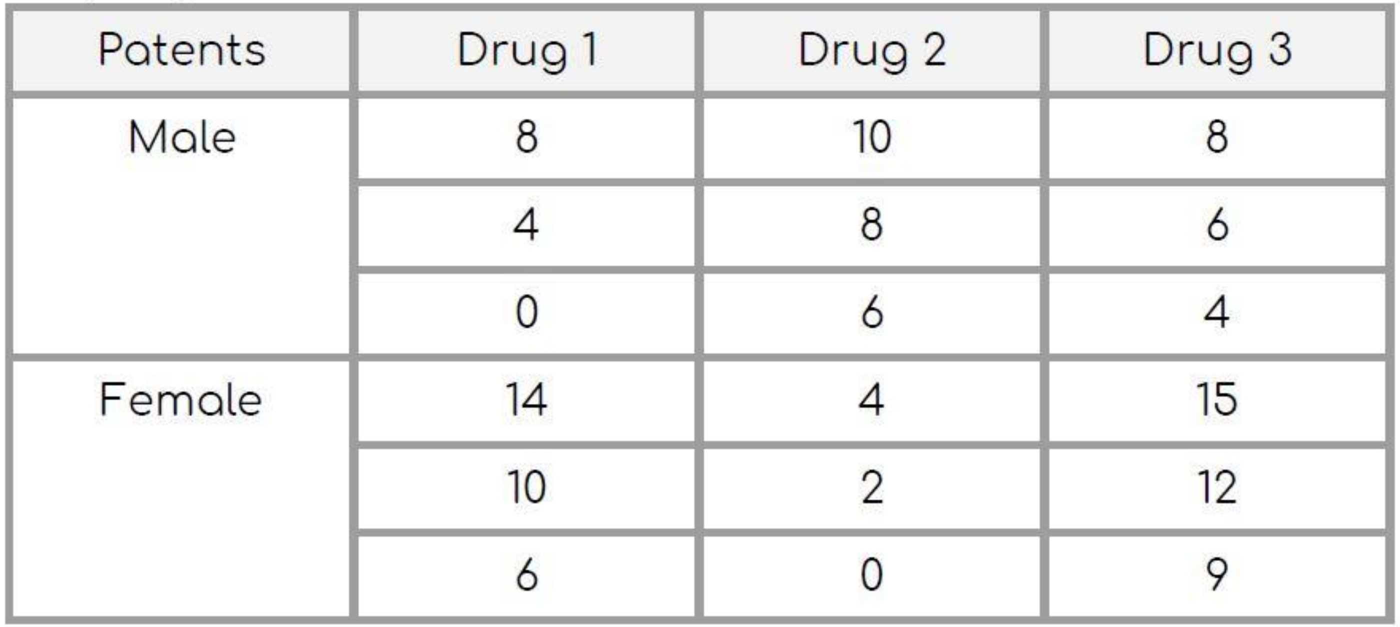


Source	Sum of Squares	Degree of Freedom	Mean Square	F value
Between	1202.628	2	601.314	52.35
Within	137.828	12	11.486	
Total	1340.456	14		

Two Way ANOVA

Forward School

- The two-way ANOVA compares the mean differences between groups that have been split on two independent variables (called factors).
- The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable.
- For example, you could use a two-way ANOVA to understand whether there is an interaction between gender and drug level on anxiety amongst patients where gender (males/females) and drug level (1, 2, 3).



Patents	Drug 1	Drug 2	Drug 3
Male	8	10	8
	4	8	6
	0	6	4
Female	14	4	15
	10	2	12
	6	0	9

Hypothesis for Two Way ANOVA

- Because the two-way ANOVA consider the effect of two categorical factors and the effect of the categorical factors on each other, there are three pairs of null or alternative hypothesis for the two-ANOVA . For example:

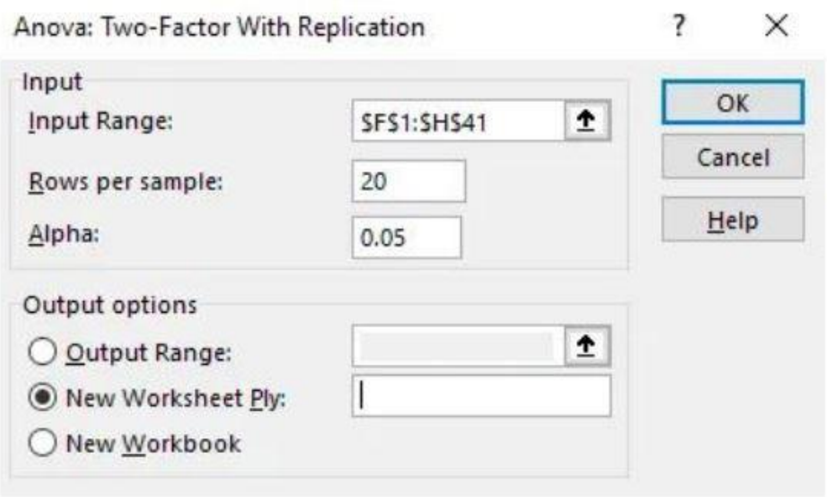
- H0: The means of all drug levels are equal.
- H1: The means of at least drug level is different.

- H0: The means of the gender groups are equal.
- H1: The means of the gender groups are different.

- H0: There is no interaction between the drug level and gender.
- H1: There is interaction between the drug level and gender.

Two Way ANOVA in Excel

- Click Data Analysis on the Data tab.
- From the Data Analysis pop up, choose ANOVA: Two-Factors with Replication.
- Under Input, select the ranges for all columns of data.
- In rows per sample, enter 20. This represents the number of observations per group.
- Excel uses a default Alpha value of 0.05 , which is usually a good value. Alpha is the significance level. Change this value only when you have a specific reason for doing so.
- Click OK.



Anova: Two-Factor With Replication

Input

Input Range: $F$1:$H$41

Rows per sample: 20

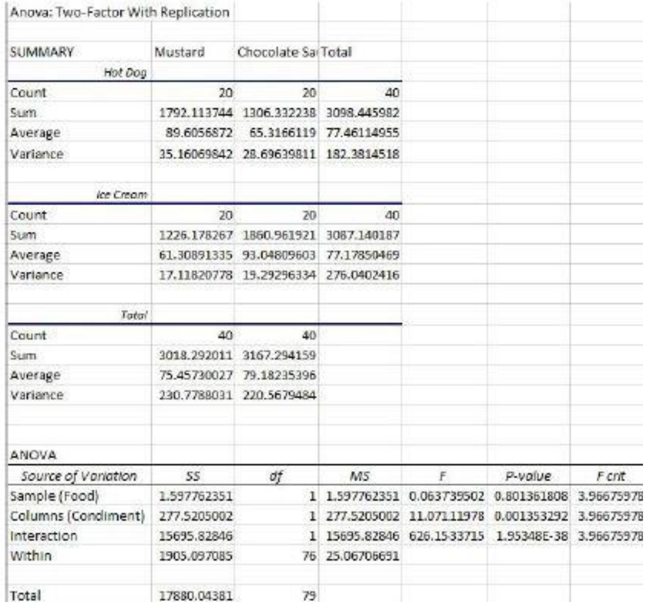
Alpha: 0.05

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook



Anova: Two-Factor With Replication

SUMMARY	Mustard	Chocolate Sa	Total
Hot Dog			
Count	20	20	40
Sum	1792.113744	1306.332238	3098.445982
Average	89.6056872	65.3166119	77.46114955
Variance	35.16069842	28.69639811	182.3814518
Ice Cream			
Count	20	20	40
Sum	1226.178267	1860.961921	3087.140187
Average	61.30891335	93.04809603	77.17850469
Variance	17.11820778	19.29296334	276.0402416
Total			
Count	40	40	
Sum	3018.292011	3167.294159	
Average	75.45730027	79.18235396	
Variance	230.7788031	220.5679484	

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample (Food)	1.597762351	1	1.597762351	0.063739502	0.801361808	3.96675978
Columns (Condiment)	277.5205002	1	277.5205002	11.07111978	0.001353292	3.96675978
Interaction	15695.82846	1	15695.82846	626.1533715	1.95348E-38	3.96675978
Within	1905.097085	76	25.06706691			
Total	17880.04381	79				

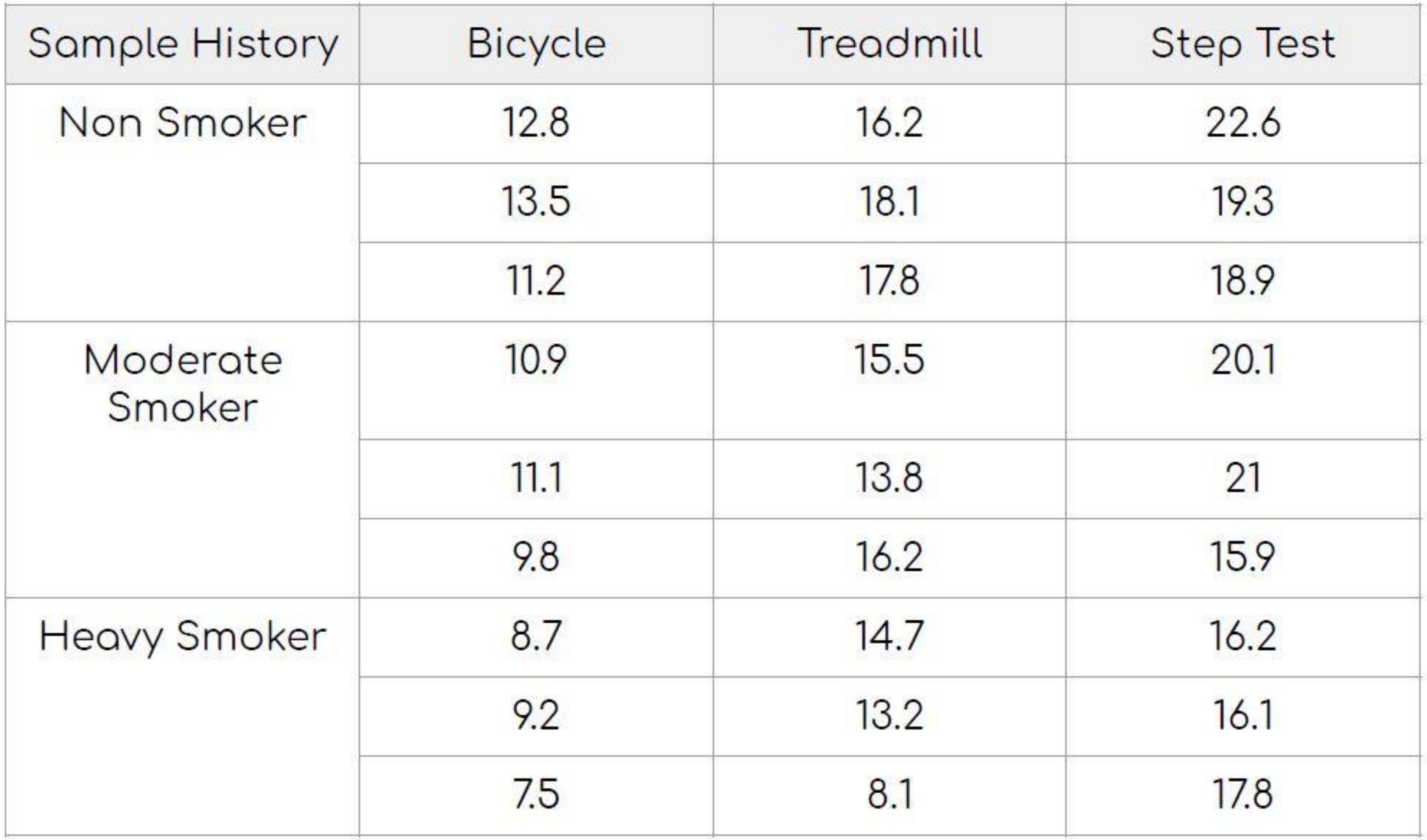
Activity: Two Way ANOVA

- A physiologist was interested in learning whether smoking history and different types of stress tests influence the timing of subject’s maximum oxygen uptake as measured in minutes.
- The researcher classified a subject’s smoking history as either heavy smoking, moderate smoking, or non-smoking. He was interested in seeing the effects of three different types of stress tests a test performed on a bicycle, a test on treadmill and test on steps.
- The physiologist recruited 9 non-smokers, 9 moderate smokers and 9 heavy smokers to participate in his experiment, for a total of n=27 subjects.
- He then randomly assigned each of his recruited subjects to undergo one of the three types of stress test.

Activity: Two Way ANOVA

Forward School

- Perform the two-way ANOVA and verify using the following online tools:
- Link: http://vassarstats.net/anova2u.html
- Here are his resulting data:



Sample History	Bicycle	Treadmill	Step Test
Non Smoker	12.8	16.2	22.6
	13.5	18.1	19.3
	11.2	17.8	18.9
Moderate Smoker	10.9	15.5	20.1
	11.1	13.8	21
	9.8	16.2	15.9
Heavy Smoker	8.7	14.7	16.2
	9.2	13.2	16.1
	7.5	8.1	17.8

One Way vs Two Way ANOVA

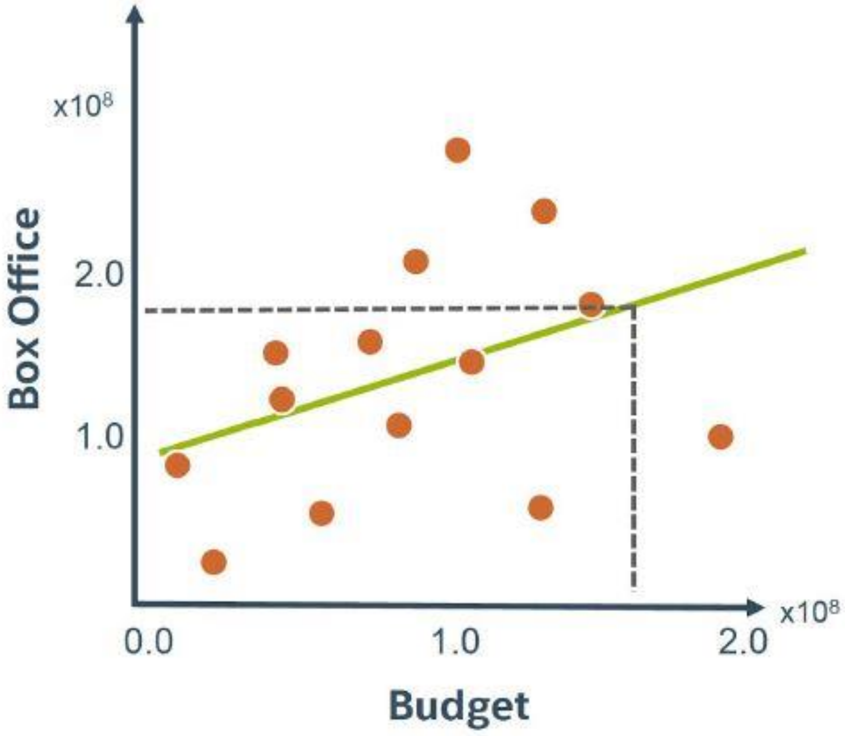
- A one-way ANOVA is primarily designed to enable the equality testing between three or more means. A two-way ANOVA is designed to assess the interrelationship of two independent variables on dependent variable.
- A one-way ANOVA only involves one factor or independent variable, whereas there are two independent variables in a two-way ANOVA.
- In a one-way ANOVA, the one factor or independent variable analyzed has three or more categorical groups. A two-way ANOVA instead compares multiple groups of two factors.
- One-way ANOVA need to satisfy only two principles of design of experiments, etc. replication and randomization. As opposed to Two-way ANOVA, which meets all three principles of designs of experiments which are replication , randomization and local control.

Forward School

Topic 3:
Regression and Correlation Analysis

Linear Regression

- Linear Regression is the most common regression model. Many predictive models use linear regression models.
- You can use a linear regression model to predict the box office from the budget.



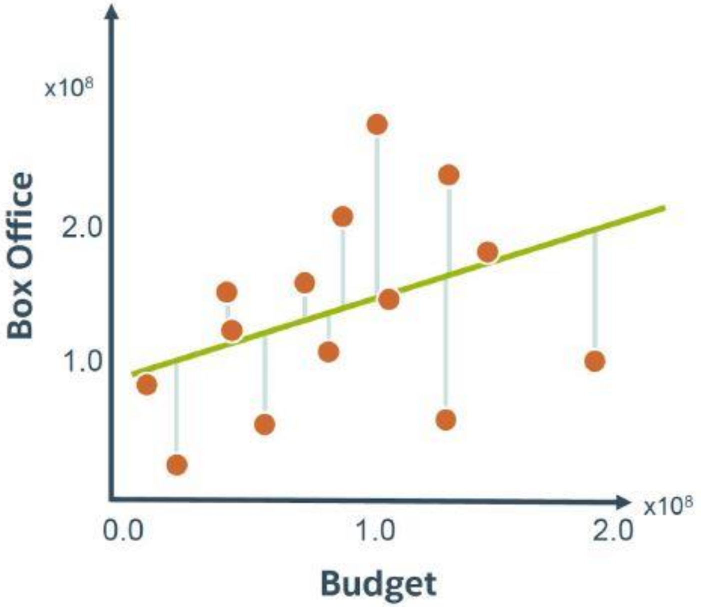
$$y_\beta(x) = \beta_0 + \beta_1 x$$

β_0= 80 million, β_1= 0.6

Predict 175 Million Gross for 160 Million Budget

Residues

- Residue is the difference between the predicted value and actual value.



$$y_\beta\left(x_{obs}^{(i)}\right) - y_{obs}^{(i)}$$

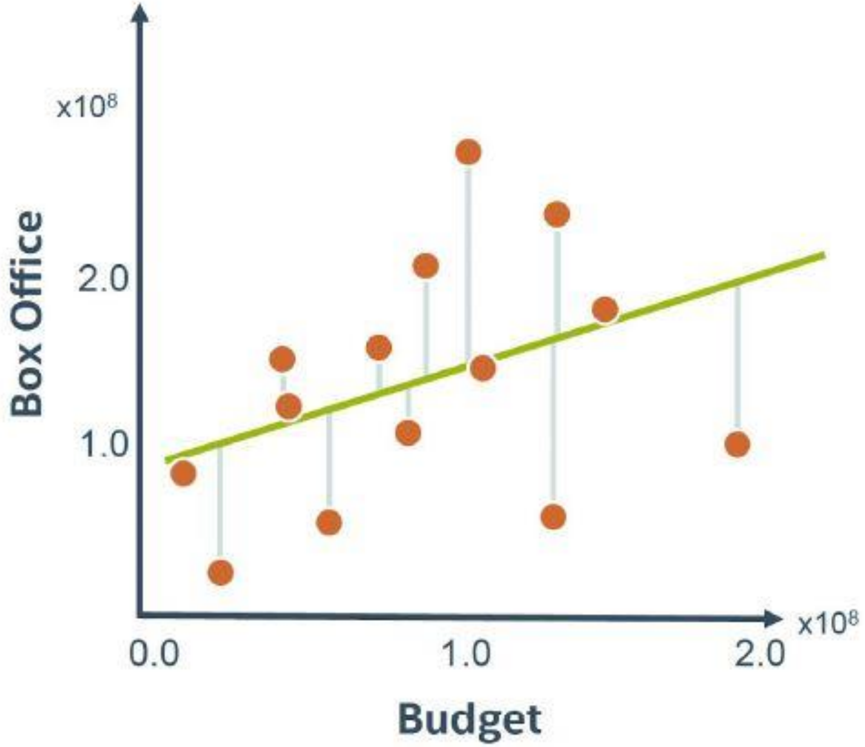
↑ Predicted value

↑ Observe d value

$$\left(\beta_0 + \beta_1 x_{obs}^{(i)}\right) - y_{obs}^{(i)}$$

Mean Square Error

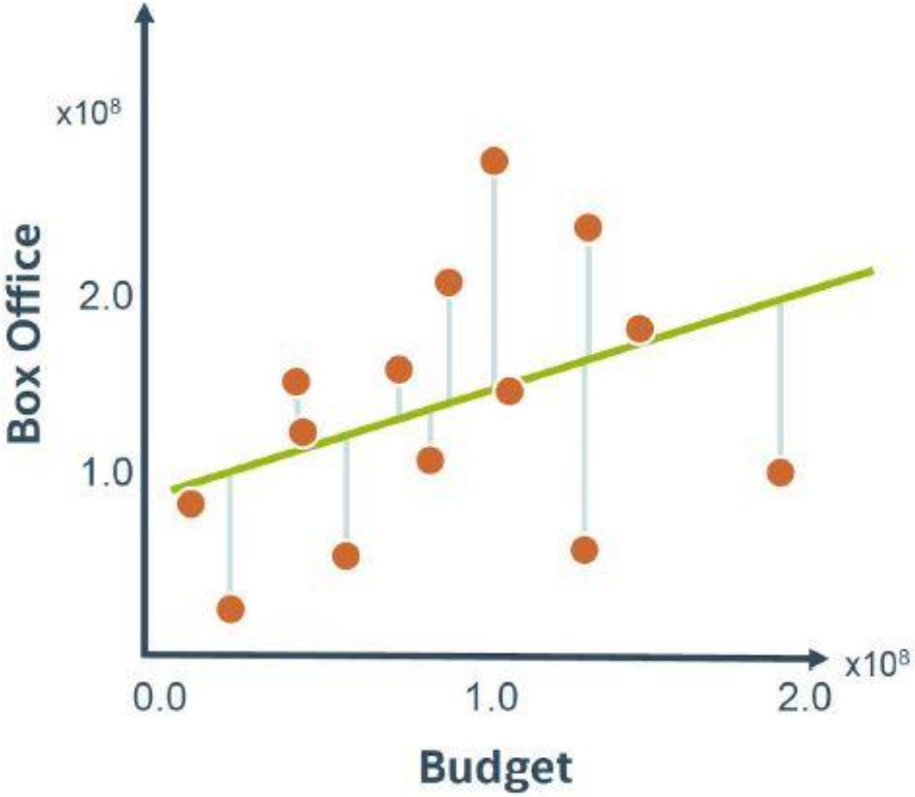
- Mean Square Error (MSE) is the common loss function to measure how good is the linear regression model.



$$\frac{1}{m}\sum_{i=1}^{m}\left(\left(\beta_0 + \beta_1 x_{obs}^{(i)}\right) - y_{obs}^{(i)}\right)^2$$

Minimum Mean Square Error

- Regression aims to minimize the MSE to find the best linear regression model.

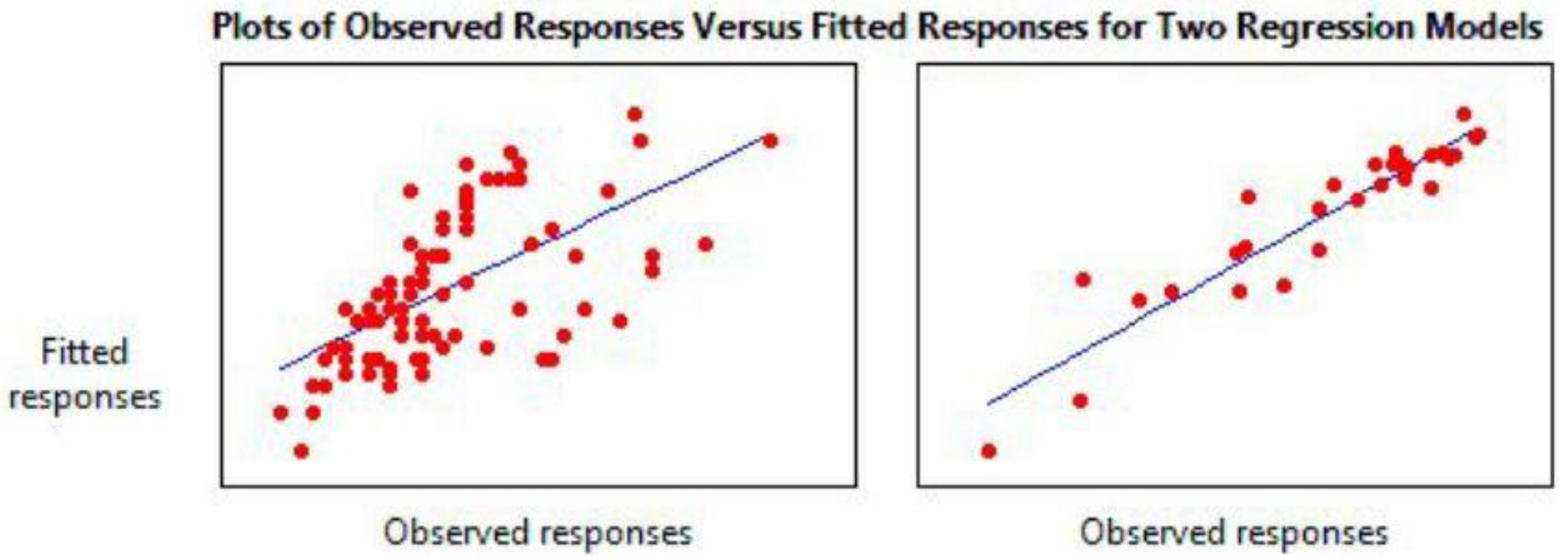


$$\min_{\beta_0,\beta_1} \frac{1}{m} \sum_{i=1}^{m} \left(\left(\beta_0 + \beta_1 x_{obs}^{(i)} \right) - y_{obs}^{(i)} \right)^2$$

R Square (Goodness of Fit)

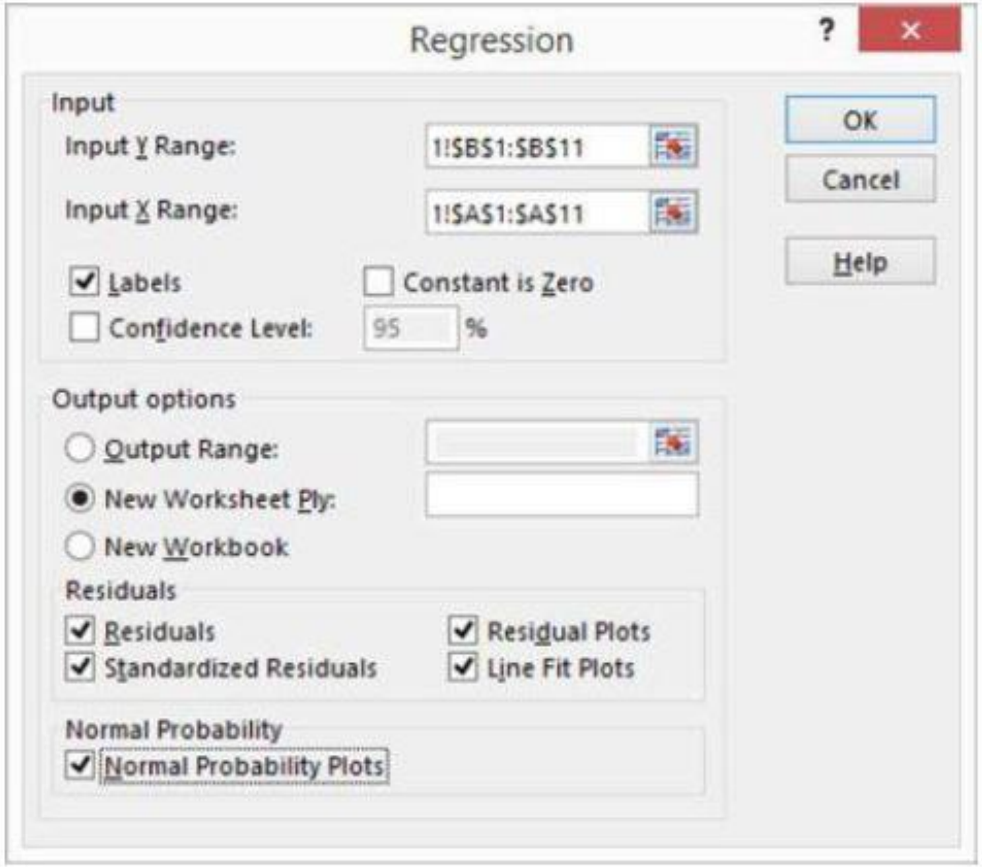
Forward School

- R-squared is a statistical measure of how close the data are to the fitted regression line.
- R-squared = Explained variation / Total variation
- R-squared is always between 0 and 1.
- 0 indicates that the model explains none of the variability of the response data around its mean.
- 1 indicates that the model explains all the variability of the response data around its mean.



Regression in Excel

- When Excel displays the Data Analysis dialog box, select the Regression tool from the Analysis Tools list and then click OK.
- Identify you Y and X values.
- Select a location for the regression analysis results.
- Identify what data you want returned.
- Click OK.



Regression

Input

Input Y Range: 1!$B$1:$B$11

Input X Range: 1!$A$1:$A$11

☑ Labels　□ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☑ Residuals　☑ Residual Plots

☑ Standardized Residuals　☑ Line Fit Plots

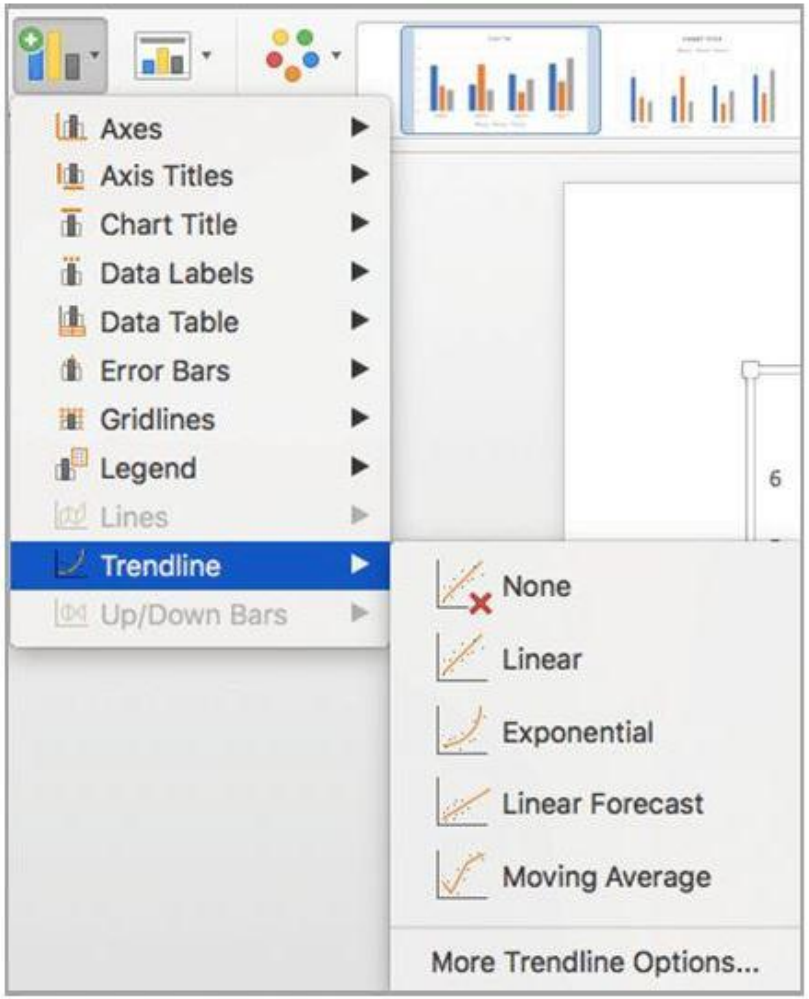
Normal Probability

☑ Normal Probability Plots

Add Trend Line in Excel

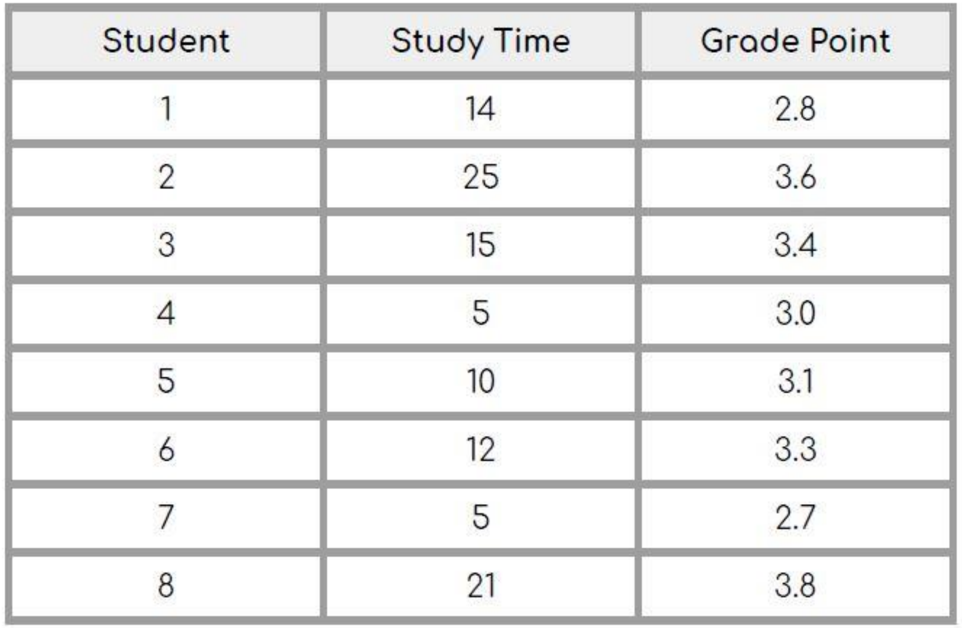
Forward School

- On the View menu, click Print Layout.
- In the chart, select the data series that you want to add a trend line to, and then click the Chart Design tab.
- On the Chart Design tab, click Add Chart Element, and then Click Trendline.
- Choose a trendline option or click more Trendline options.



Activity: Regression

- The data are collected at the end of an introductory statistics course. The table shows the data for the eight males in the class on these variables and on the number of class lectures for the course that the student reported skipping during the term.
- Investigate the relationship between x=study time and y=GPA. Find the prediction equation and interpret the slope.
- Perform a regression and verify by using the following one tool.
- Link: https://www.graphpad.com/quickcalcs/linear1/



Student	Study Time	Grade Point
1	14	2.8
2	25	3.6
3	15	3.4
4	5	3.0
5	10	3.1
6	12	3.3
7	5	2.7
8	21	3.8

What is Covariance

- Variance is a measure of the variability or spread in a set of data.
- We use the following formula to compute variance for population and sample respectively.

$$Var(x) = \frac{\sum(x - \overline{x})^2}{N} \qquad Var(x) = \frac{\sum(x - \overline{x})^2}{N - 1}$$

- Covariance is a measure of the extent to which corresponding elements from two sets of ordered data move in the same direction.
- We use the following formula to compute covariance for population and sample respectively.

$$Cov(x, y) = \frac{\sum(x - \overline{x})(y - \overline{y})}{N} \qquad Cov(x, y) = \frac{\sum(x - \overline{x})(y - \overline{y})}{N - 1}$$

Covariance Matrix

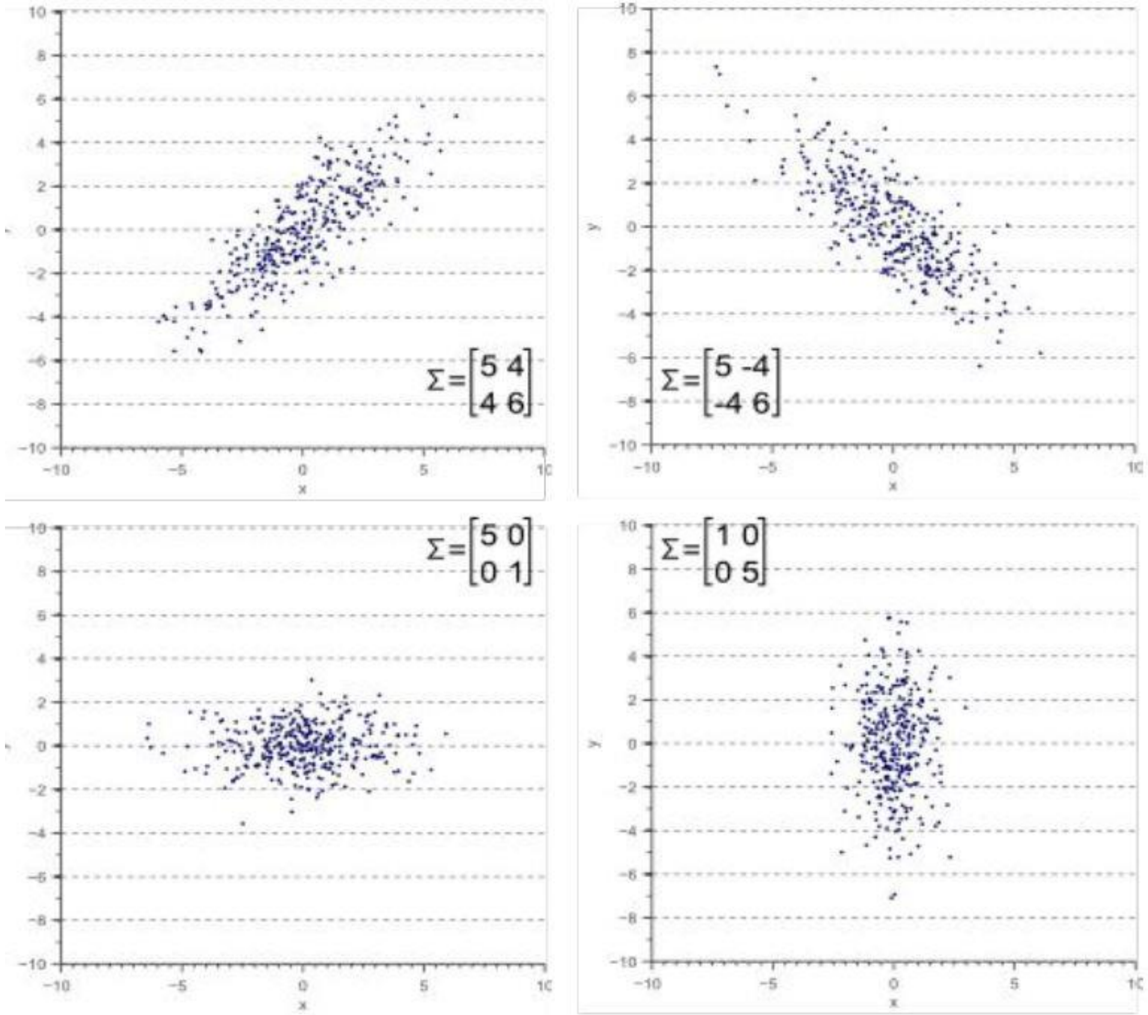
- Variance and covariance are often displayed together in a covariance matrix given as follows:

$$\mathrm{Cov}(A)=\begin{bmatrix} \dfrac{\sum(x_i-\bar{X})(x_i-\bar{X})}{N} & \dfrac{\sum(x_i-\bar{X})(y_i-\bar{Y})}{N} \\ \dfrac{\sum(x_i-\bar{X})(y_i-\bar{Y})}{N} & \dfrac{\sum(y_i-\bar{Y})(y_i-\bar{Y})}{N} \end{bmatrix}$$

$$=\begin{bmatrix} \mathrm{Cov}(X,X) & \mathrm{Cov}(Y,X) \\ \mathrm{Cov}(X,Y) & \mathrm{Cov}(Y,Y) \end{bmatrix}$$

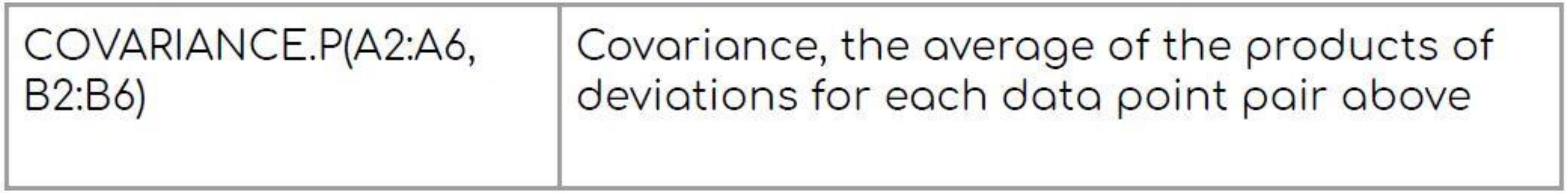
Covariance Matrix Visualization

- Variance and covariance are often displayed together in a covariance matrix given as follows:



COVARIANCE.P function in Excel

- The COVARIANCE.P function returns population covariance, the average of the products of deviations for each data point pair in two data sets. Use covariance to determine the relationship between two data sets.
- For example, you can examine whether greater income accompanies greater levels of educations.



COVARIANCE.P(A2:A6, B2:B6)	Covariance, the average of the products of deviations for each data point pair above

Activity: Covariance

Compute the covariance for the following data.

X: 90, 90, 60, 60, 30

Y: 60, 90, 60, 60, 30

- Verify the answer using the online covariance calculator.
- Link: https://www.calculatored.com/math/algebra/covariance-calculator

What is Correlation

- The correlation coefficient is also known as the Pearson product-moment correlation coefficient, or Pearson correlation coefficient.
- It is obtained by dividing the covariance of the two variables by the product of their standard deviations.

$$Corr(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

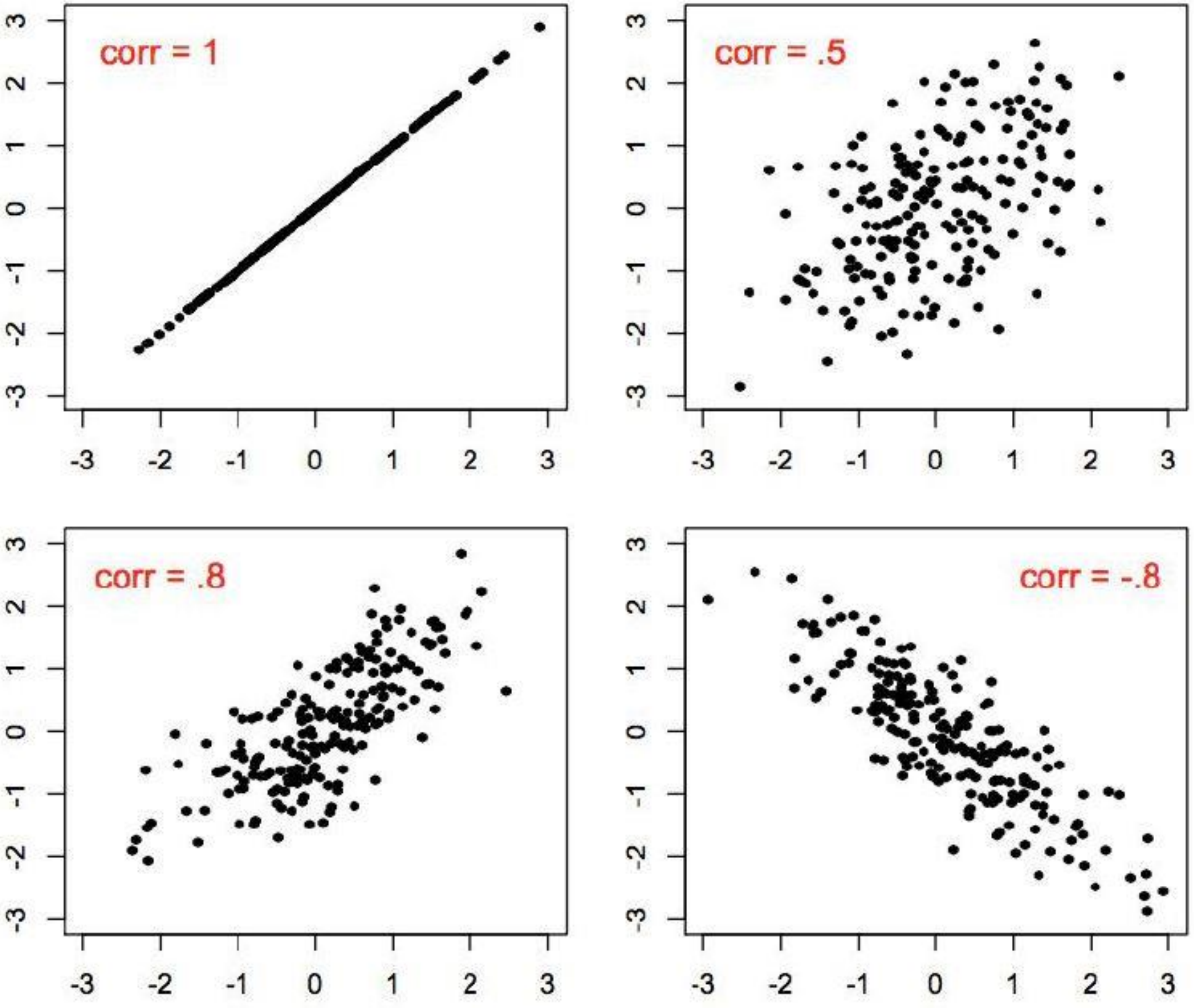
- The values of the correlation coefficient can range from -1 to +1. The closer it is to +1 or -1, the more closely are the two variables are related.
- The positive sign signifies the direction of the correlation i.e. if one of the variables increases, the other variable is also supposed to increase.

Correlation Matrix

- For multiple variables, we can display all the correlation coefficients in the matrix form as below:

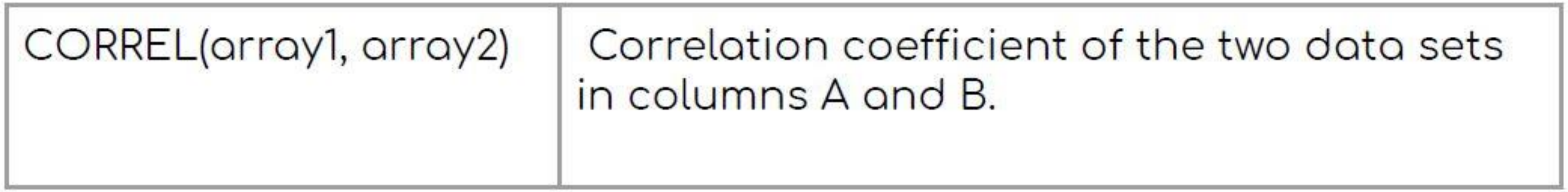
$$\begin{bmatrix} 1 & \mathrm{Corr}(X,Y) & \mathrm{Corr}(X,Z) \\ \mathrm{Corr}(X,Y) & 1 & \mathrm{Corr}(Y,Z) \\ \mathrm{Corr}(X,Z) & \mathrm{Corr}(Y,Z) & 1 \end{bmatrix}$$

Correlation Coefficient



CORREL function in Excel

- The CORREL function returns the correlation coefficient of two cell ranges. Use the correlation coefficient to determine the relationship between 2 properties.
- For example, you can examine the relationship between a location’s average temperature and the use of air conditioners.



CORREL(array1, array2)	Correlation coefficient of the two data sets in columns A and B.

Activity: Correlation

Compute the Person correlation coefficient for the following data.

X: 90, 90, 60, 60, 30

Y: 60, 90, 60, 60, 30

- Verify the answer using the online correlation coefficient calculator.
- Link: https://www.socscistatistics.com/tests/pearson/default2.aspx