Basic of Statistics

Types of Statistics

- **Descriptive statistics** deals with enumeration, organization and graphical representation of the data, e.g. Census data.

- **Inferential statistics** is concerned with making conclusions about population characteristics using information contained in a sample, that is , generalizing from the specific e.g. an opinion poll, such as the Gallop Poll.

Population and Sample

Population is the set of all measurements of interest to the researcher.

Sample is any subset of measurements selected from the population.

Population and Sample

Population

Sample

Examples of Population

1. **Average Income** of households in Pakistan.
2. **Households** in Islamabad who own at least one car.
3. **Patients** coming to PIMS Hospital

Characteristic of Population and Sample

- Parameter: The set of measurements in a population may be summarized by a descriptive characteristic, called a parameter. e.g. average income of household in Pakistan.

- Statistic: The set of measurements in a sample may be summarized by a descriptive characteristic, called a statistic. e.g. sample mean

Measurement Scales

- **Nominal Scale**

  Male-Female, Well-Sick, Child-Adult, Married-Unmarried

- **Ordinal Scale**

  Unimproved, Improved, Much Improved.
  Above average, average, below average etc.

- **Intervals Scale**

  Arbitrary Zero and unit distance e.g. temperature.

- **Ratio Scale**

  Height, weight and length etc.

Hypothesis Testing

- **Significance of Hypothesis Testing**

- Till now we have learned about the following topic:
  - Inferential statistics: Making inferences about the population using the sample data

- Inferential statistics is used to find some population parameter (mostly population mean) when you have no initial number to start with. So, you start with the sampling activity and find out the sample mean. Then, you estimate the population mean from the sample mean using the confidence interval.

- Hypothesis testing is used to confirm your conclusion (or hypothesis) about the population parameter (which you know from EDA or your intuition). Through hypothesis testing, you can determine whether there is enough evidence to conclude if the hypothesis about the population parameter is true or not.

Hypothesis Testing

- **Significance of Hypothesis Testing**

- Let’s take an example of Hypothesis Testing. We took a sample of 100 Maggi and tried to estimate the population mean lead component with 99% confidence interval. We found out this range between 2.2-2.3 PPM and the permitted range was 2.5 PPM, So we said that Maggi is safe. But how sure are we if this claim is true or not? We need to check this if the value that we estimated is a true representative of the population mean. Now here comes the concept of HT.

- Maggie Claim: The average lead content in a Maggi packet is less than 2.5PPM

- Now Hypothesis testing will help us to statically verify if the claim by Maggi holds true for population or not?

Hypothesis Testing

- **Significance of Hypothesis Testing**
- Let’s take a real world example

Hypothesis Testing

- **Null and Alternate Hypothesis**

- The two opposing Hypothesis are know as NULL HYPOTHESIS and ALTERNATE HYPOTHESIS.

Represents	No observed effect	Some observed effect
What is it?	It is what the researcher tries to disprove.	It is what the researcher tries to prove.
Acceptance	No changes in opinions or actions	Changes in opinions or actions

- Image from : https://keydifferences.com/difference-between-null-and-alternative-hypothesis.html

Hypothesis Testing Decision Chart

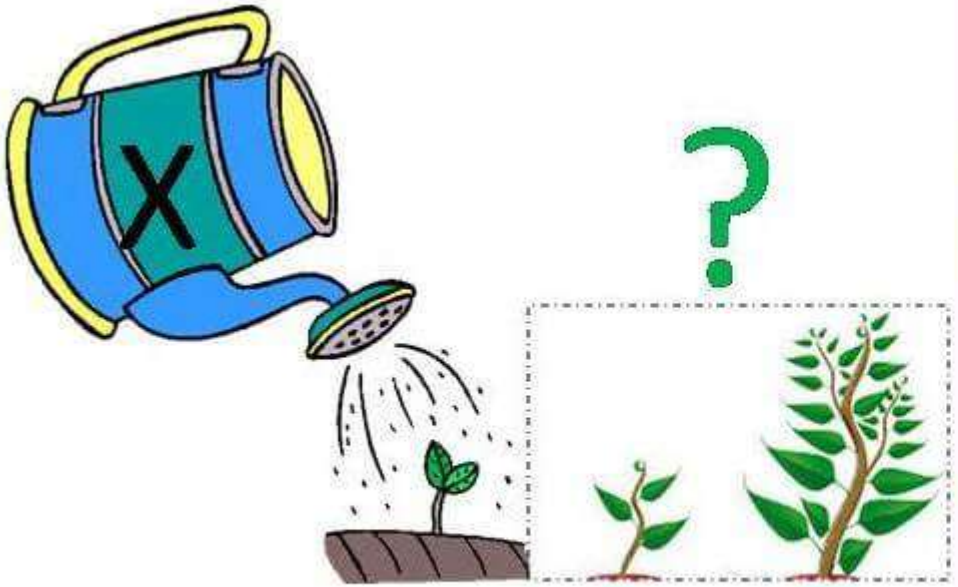
Decision \ *Reality*	Null Hypothesis (H_0) is true	Alternative Hypothesis (H1) is true
Reject (H_0)	Type I error (α) typically .05 or .01	Correct decision (Power = 1 - β) typically .80
Don't reject (H_0)	Correct decision (1 - α) typically .95 or .99	Type II error (β) typically .20

Effect of Bio-fertilizer ‘x’ on Plant growth

www.majordifferences.com

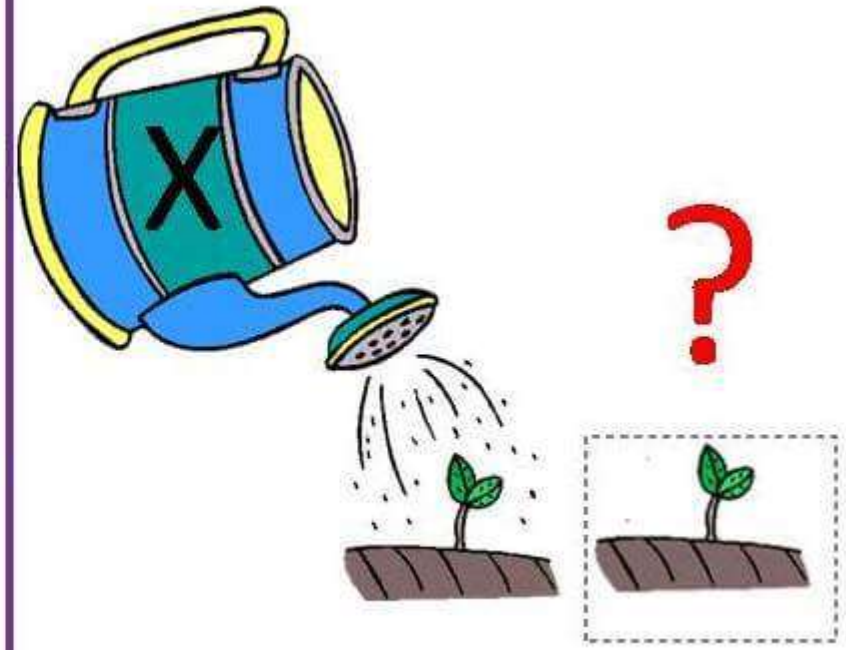
Alternative Hypothesis

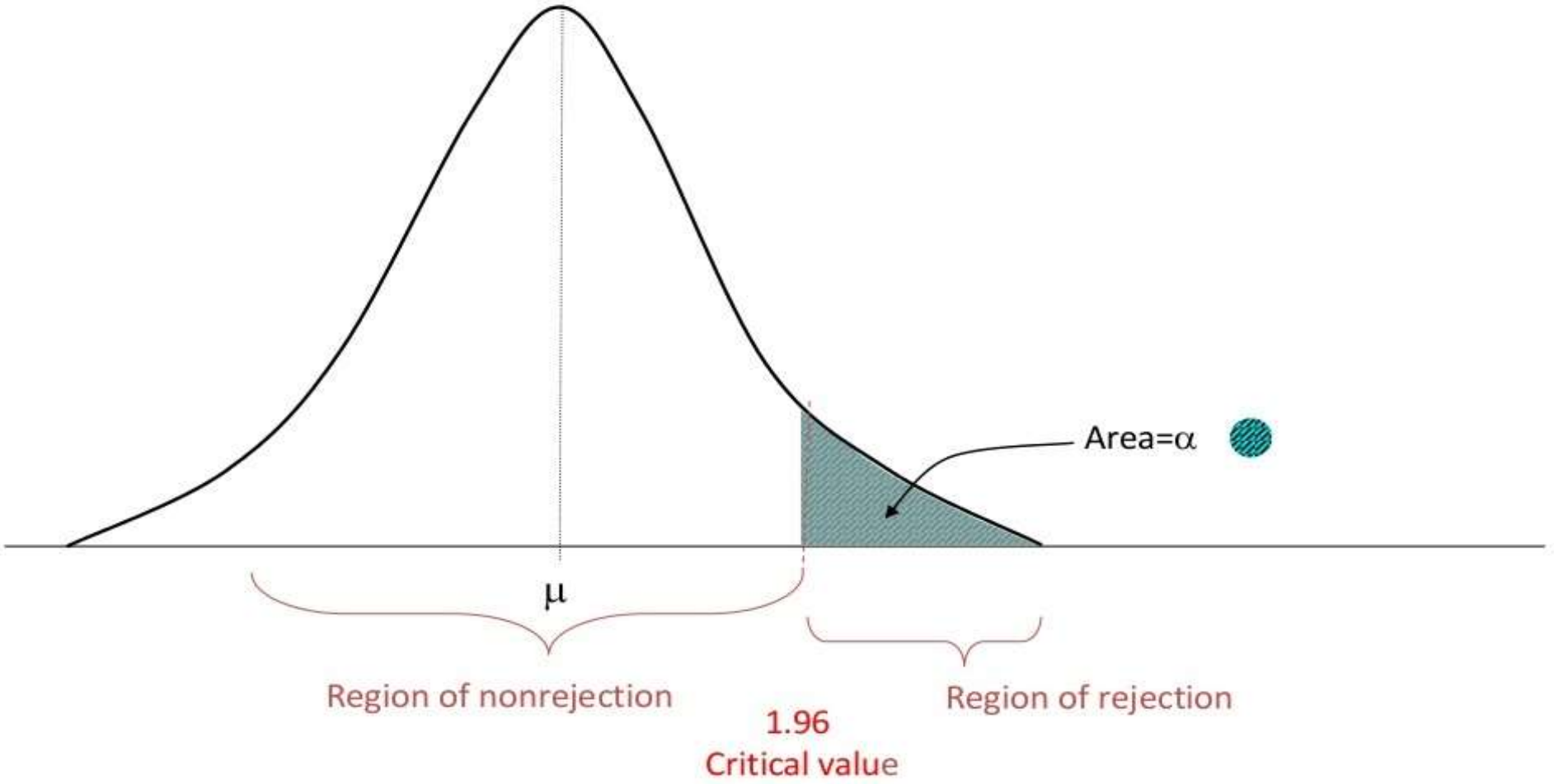
H_1: Application of bio-fertilizer ‘x’ increase plant growth.



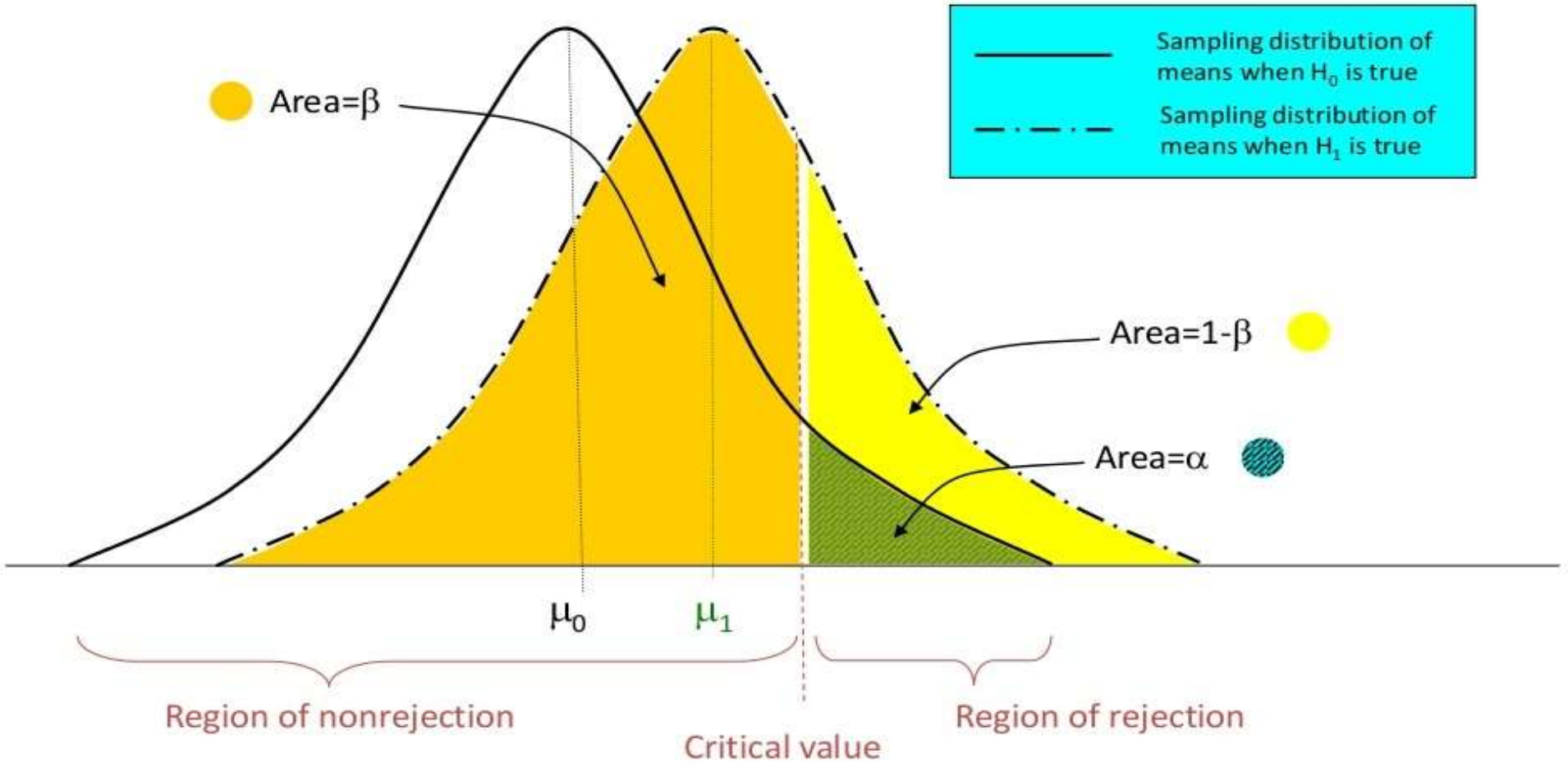
Null Hypothesis

H_0: Application of bio-fertilizer ‘x’ do not increase plant growth.



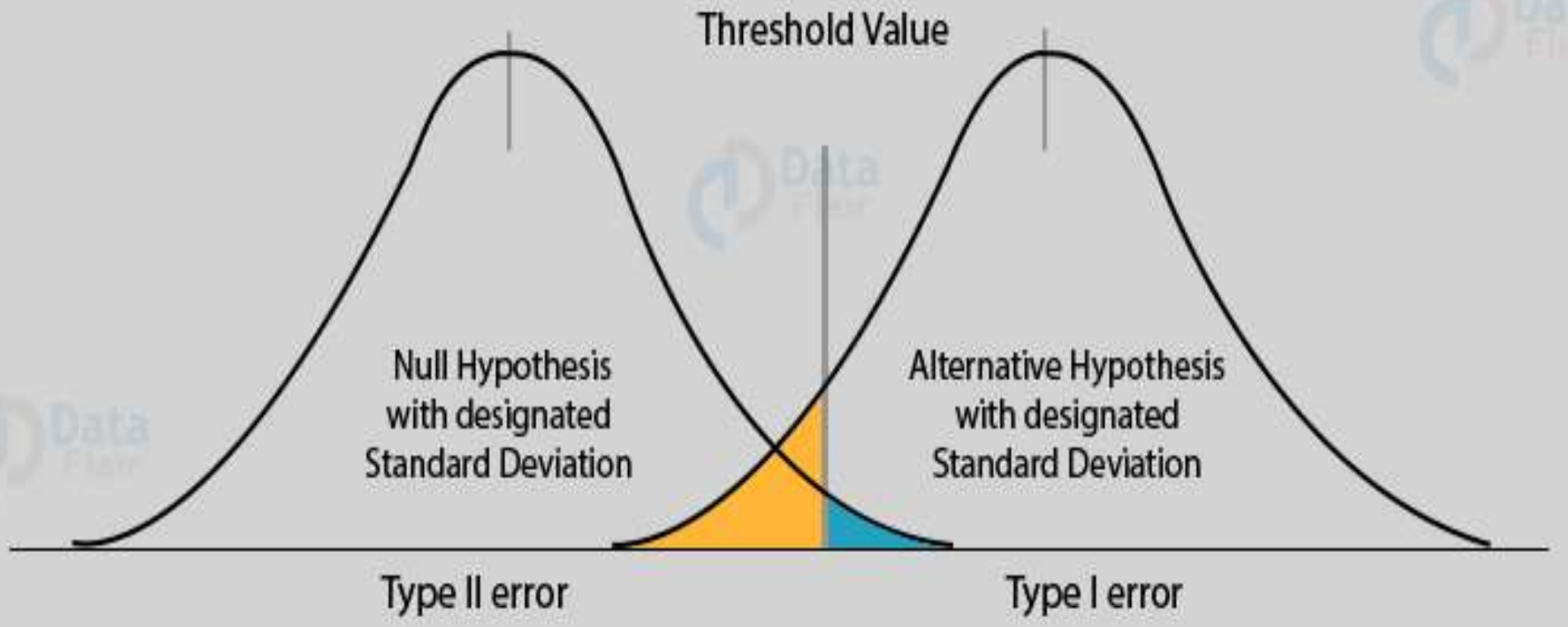

Area=α
μ
Region of nonrejection
1.96
Critical value
Region of rejection

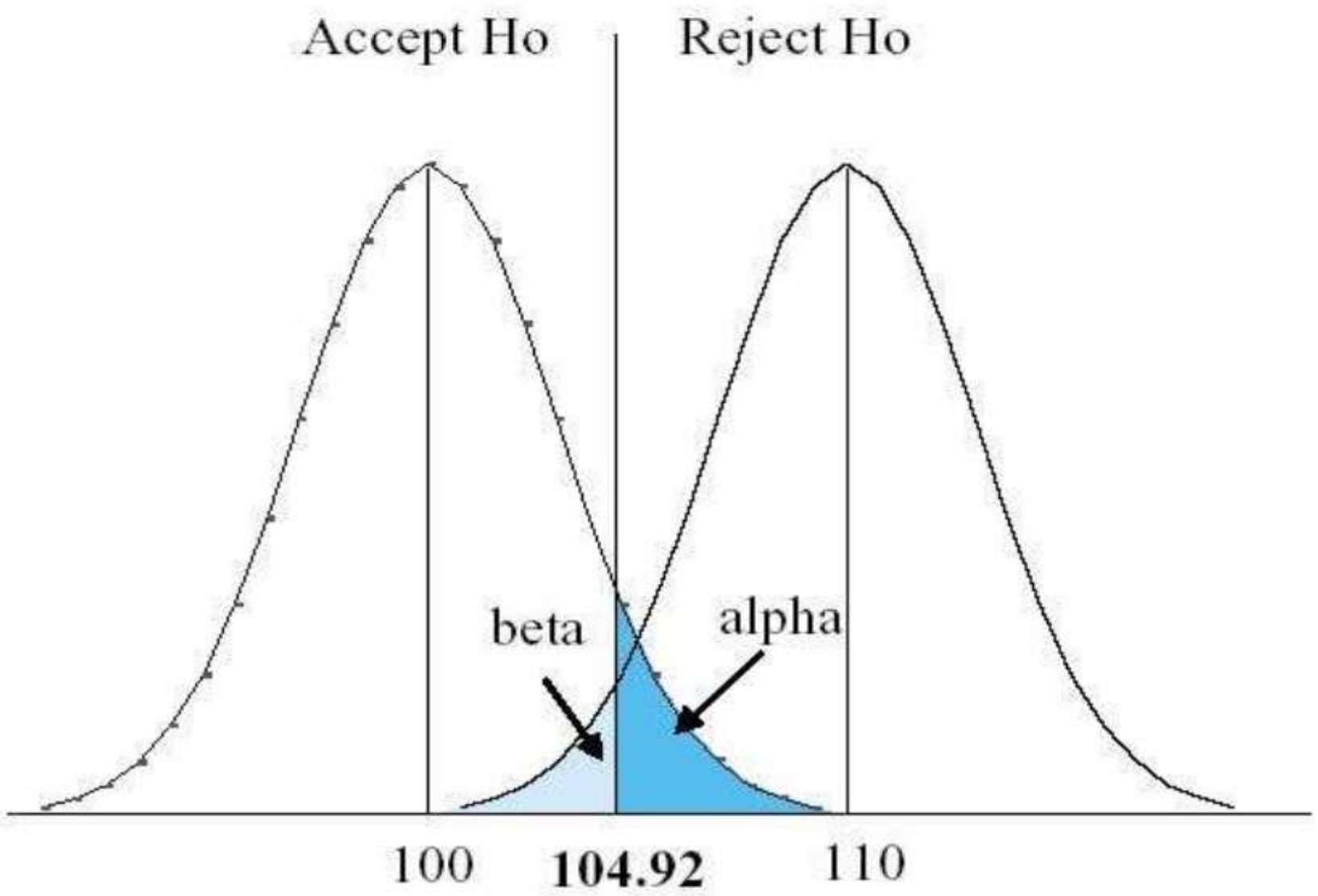
Testing a mean μ_0 against a true alternative μ_1



If the value is less than the threshold value, then we reject the alternative hypothesis and accept the null hypothesis.

If the values goes beyond the threshold value, then we accept the alternative hypothesis and reject the other one.





Hypothesis Testing

- **Formulating Hypothesis**

A restaurant owner installed a new automated drink machine. The machine is designed to dispense $530\ \text{mL}$ of liquid on the medium size setting. The owner suspects that the machine may be dispensing too much in medium drinks. They decide to take a sample of 30 medium drinks to see if the average amount is significantly greater than $530\ \text{mL}$.

What are appropriate hypotheses for their significance test?

- Questions from: https://www.khanacademy.org/

Hypothesis Testing

- **Formulating Hypothesis**

The National Sleep Foundation recommends that teenagers aged 14 to 17 years old get at least 8 hours of sleep per night for proper health and wellness.

A statistics class at a large high school suspects that students at their school are getting less than 8 hours of sleep on average. To test their theory, they randomly sample 42 of these students and ask them how many hours of sleep they get per night. The mean from the sample is $\bar{x} = 7.5$ hours.

Here's their alternative hypothesis:

H_a : The average amount of sleep students at their school get per night is...

What is an appropriate ending to their alternative hypothesis?

- Questions from: https://www.khanacademy.org/

Hypothesis Testing

- **Formulating Hypothesis**

- Now try to formulate the Null and Alternate Hypothesis for Maggi example. Remember the allowed limit by the Government is 2.5PPM

Hypothesis Testing

CONFLICTING HYPOTHESES IN CRIMINAL TRIAL EXAMPLE

Null Hypothesis: H_0-Defendent is innocent

Alternate Hypothesis : H_1-Defendent is not innocent

Rejection of Null Hypothesis - Guilty

Failure to reject null hypothesis - Not Guilty

Failure to reject null hypothesis ≠ Accept null hypothesis

General Approach to Hypothesis Testing

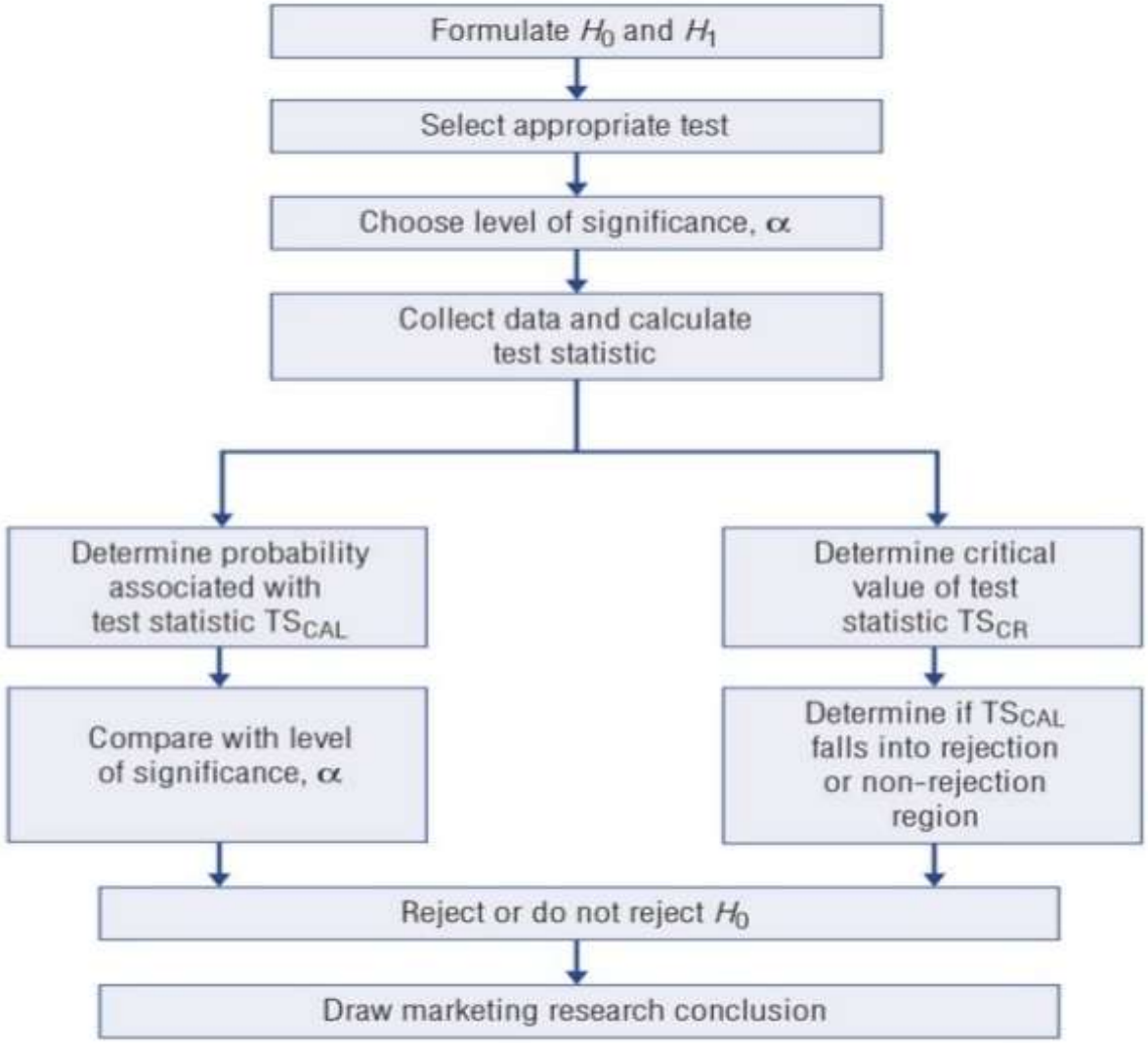


Figure 10.5 A general procedure for hypothesis testing

Malhotra et al. 2005, 293

What is a hypothesis?

- A hypothesis is an unproven proposition or supposition that tentatively explains certain facts or phenomena.
  - Empirically testable
- Null hypothesis is a statement about status quo.
  - Any change from what has been thought to be true will be due entirely to random sampling error.
- Alternative hypothesis is a statement indicating the opposite of the null.

Null and alternative hypotheses

- Example: highly dogmatic consumers will be less likely to try a new product than less dogmatic consumers.
- The null hypothesis (H_0)is where there is no difference between high dogmatics and low dogmatics in their willingness to try an innovation.
- The alternative hypothesis (H_1) is where there is a difference between high and low dogmatics.

Hypothesis testing

- The purpose of hypothesis testing is to determine which of the two hypotheses is correct.

The hypothesis–testing procedure

- The process is as follows:
  - Determine a statistical hypothesis.
  - Imagine what sampling distribution would be if this is a true statement.
  - Take an actual sample and calculate sample mean.
  - Determine if the deviation between the obtained value and expected value of sample mean could have occurred by chance alone.
  - Set a standard for determining if we reject the null or accept the alternative.

The hypothesis–testing procedure

- Significance level is the critical probability in choosing between the null and alternative hypotheses.
  - The probability level that is too low to warrant support of the null hypothesis.
- Assuming the null is true, if the probability of occurrence of the observed data is smaller than the significance level, then the data suggest that the null should be rejected.
  - Evidence supporting contradiction of null.

Type I and type II errors

- Since hypothesis testing is based on probability theory, the researcher cannot be completely certain and runs the risk of committing two types of errors.
  - Type I error is an error caused by rejecting the null hypothesis when it is true.
    - Probability of alpha (α)
  - Type II error is an error caused by failing to reject the null hypothesis when the alternative hypothesis is true.
    - Probability of beta (β)

Errors

- **Type-1 error**
- Alpha given as 5% or 1%

	The null hypothesis is true	The null hypothesis is false
We decide to reject the null hypothesis	Type I error (rejecting a true null hypothesis) α	Correct decision
We fail to reject the null hypothesis	Correct decision	

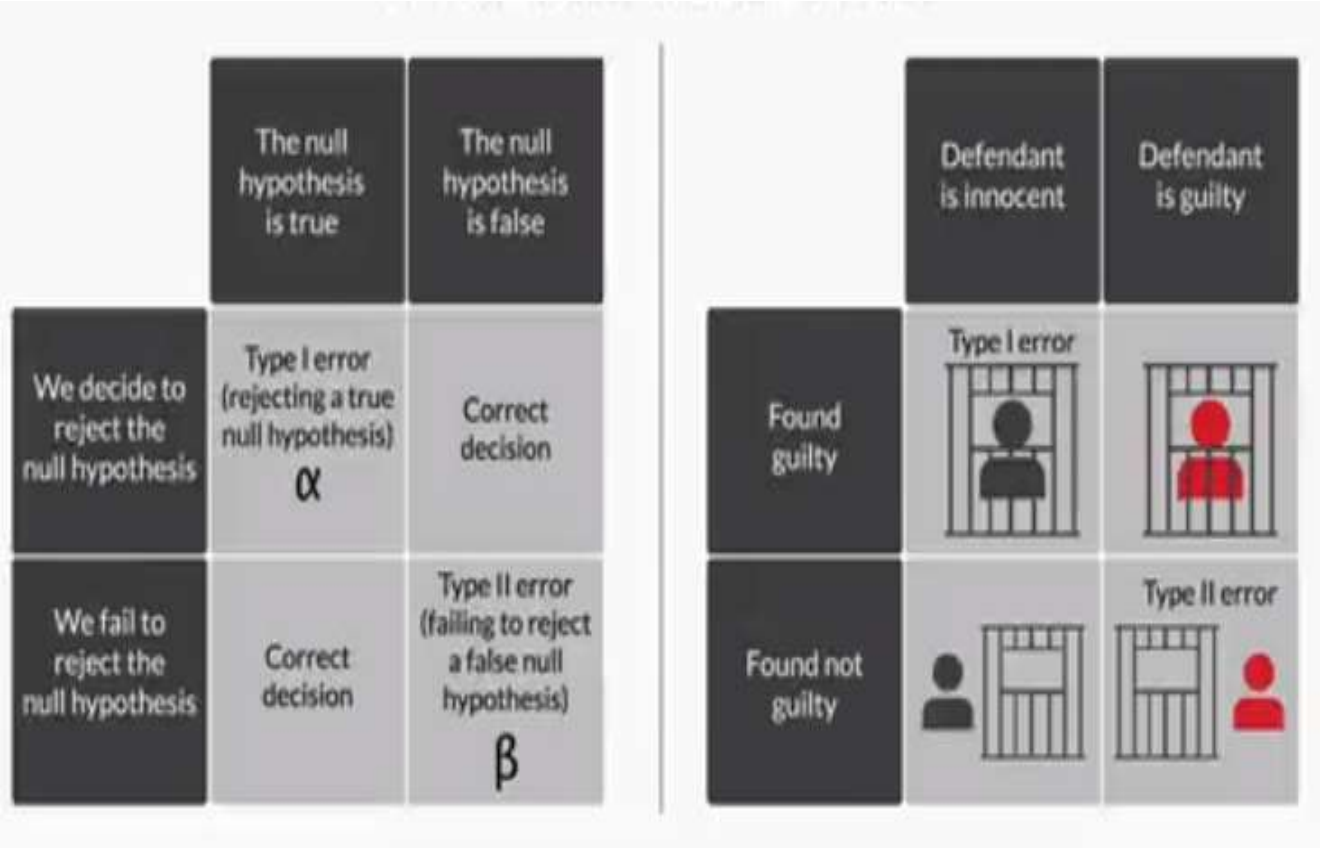
	Defendant is innocent	Defendant is guilty
Found guilty	Type I error 	
Found not guilty		

Hypothesis Testing

- **Type-II Error**

- H0: The amount of
- chemical required to
- cure a heart disease
- in a pill is equal to the
- amount require to cure
- the heart disease.

- We failed to reject the
- Null Hypothesis. The
- amount of chemical was
- very high

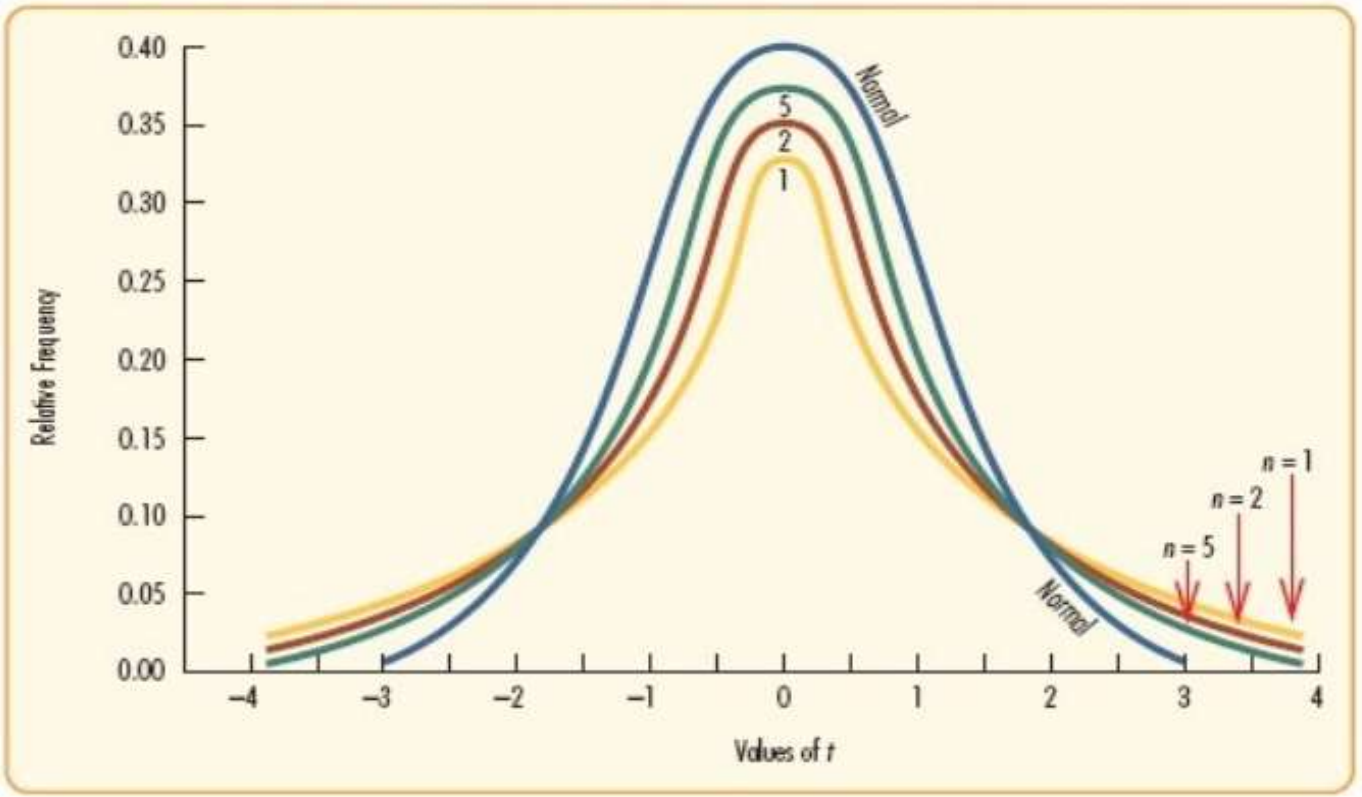


The t–distribution

- The t-distribution is a symmetrical, bell–shaped distribution that is contingent on sample size.
  - It has a mean of zero and a standard deviation equal to one.
- The shape of the t–distribution is influenced by its degrees of freedom.
  - Number of observations minus number of constraints or assumptions.
  - For sample sizes over 30, t–distribution closely approximates Z–distribution.

The t–distribution

EXHIBIT 12.10 The *t*-distribution for various degrees of freedom



Univariate hypothesis test using the t–distribution

- Store manager believes that average number of customers who return or exchange merchandise is 20.
  - **H_0: $\mu = 20$ and H_1: $\mu \neq 20$**
- Store records returns and exchanges for 25 days (n) and sample mean is 22 and standard deviation is 5.
- Confidence level of 95% or significance level of 5%.
- Referring to Table B.3 in Appendix B, we find that for 24 degrees of freedom (n–1), the t–value is 2.064. Thus:

$$\text{Lower limit} = \mu - t_{cl}\frac{S}{\sqrt{n}} = 20 - (2.064)\frac{5}{\sqrt{25}} = 17.936$$

$$\text{Upper limit} = \mu + t_{cl}\frac{S}{\sqrt{n}} = 20 + (2.064)\frac{5}{\sqrt{25}} = 22.064$$

- Since the sample mean lies within the critical limits, the null hypothesis cannot be rejected.

Univariate hypothesis test using the t–distribution

- Alternatively, we may test a hypothesis by calculating the observed t–value and comparing it to the critical t–value.
- To calculate the observed t:

$$t_{obs} = \frac{\bar{X} - \mu}{S_{\bar{x}}} = \frac{22 - 20}{\frac{5}{\sqrt{25}}} = 2$$

- Referring to Table B.3 in Appendix B, we find that for 24 degrees of freedom (n–1), the t–value is 2.064.
- Since the observed t–value is less than the critical t–value, the sample mean lies within the critical limits.
- Thus, the null hypothesis cannot be rejected.

t-test

A test which compare means two 2 groups

When should the t-distribution be used?

- When data variables are continuous
- when sample size is small
- when population SD is not known.

If you know population SD or your sample size exceeds 25, feel free to use the Standard Normal Distribution

Example: A gynecologist at PIMS feels that pregnant women coming to the hospital are mostly anemic. She takes a random sample of 25 pregnant women and test their Hb level.

Is this average Hb level less than the Hb level of normal women population? Which is 13.

H_0: $\mu = 13$ Ha: $\mu < 13$

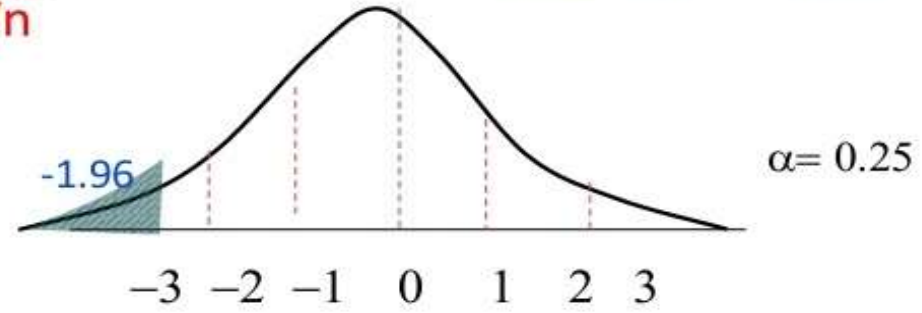
Significance level $\alpha = 0.5$

Test Statistic: $t = \frac{X - \mu_0}{S/\sqrt{n}}$

$X = 11.8$ g/dl $\mu_0 = 13$

$S = 2.6$

$t = \frac{11.8 - 13}{2.6/\sqrt{25}} = \frac{-1.2}{0.52} = -2.32$



The *t*-test

Measurements

The investigation involved **taking measurements**

Use this test when
- you wish to find out if there is a significant difference between two means
- the data are normally distributed
- the sample size is between **10-30**

t can be calculated from the formula

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where $\bar{x}_1$ = mean of first sample
$\bar{x}_2$ = mean of second sample
s_1 = standard deviation of first sample
s_2 = standard deviation of second sample
n_1 = number of measurements in first sample
n_2 = number of measurements in second sample

Standard deviation formula is also needed to solve for *t*-test

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

A table showing the critical values of *t* for different degrees of freedom.

Degrees of freedom

Degrees of freedom	Critical value	Degrees of freedom	Critical value
4	2.78		
5	2.57	15	2.13
6	2.48	16	2.12
7	2.37	18	2.10
8	2.31	20	2.09
9	2.26	22	2.07
10	2.23	24	2.06
11	2.20	26	2.06
12	2.18	28	2.05
13	2.16	30	2.04
14	2.15	40	2.02

Critical Value

The number of degrees of freedom = $(n_1 + n_2) - 2$

T-test - Is there significant difference between two means?

Set up a chart that will help you solve for **s** and **t**.

Sample Number	Sample X_1	$X_1 - \overline{X}_1$	$(X_1 - \overline{X}_1)^2$	Sample X_2	$X_2 - \overline{X}_2$	$(X_2 - \overline{X}_2)^2$
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
	$\overline{X}_1 =$		$\Sigma =$	$\overline{X}_2 =$		$\Sigma =$

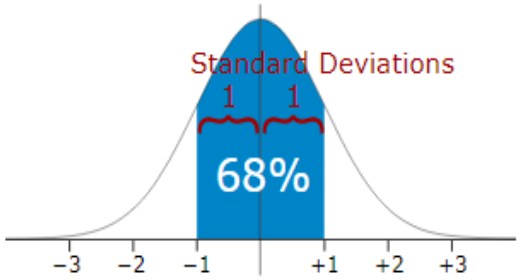
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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

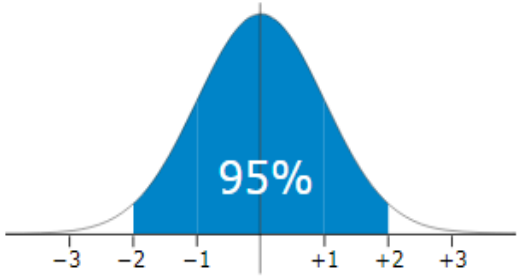
Standard Deviation

A graphical expression of the distance between numbers.

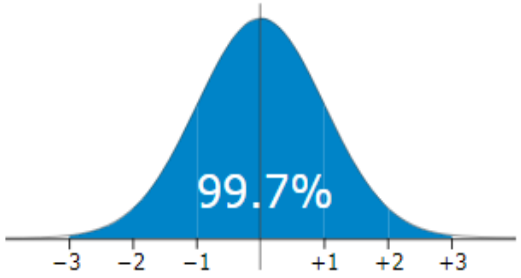
When we calculate the standard deviation we find that (generally):



68% of values are within
1 standard deviation of the mean



95% of values are within
2 standard deviations of the mean



99.7% of values are within
3 standard deviations of the mean

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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

Standard deviation will give the **+/- error value** of a measurement.

	Female	Male
mean	21	10
SD	3	1.6
2 x SD	6	3.2
Mean + (2 x SD)	27	13
Mean - (2 x SD)	15	7

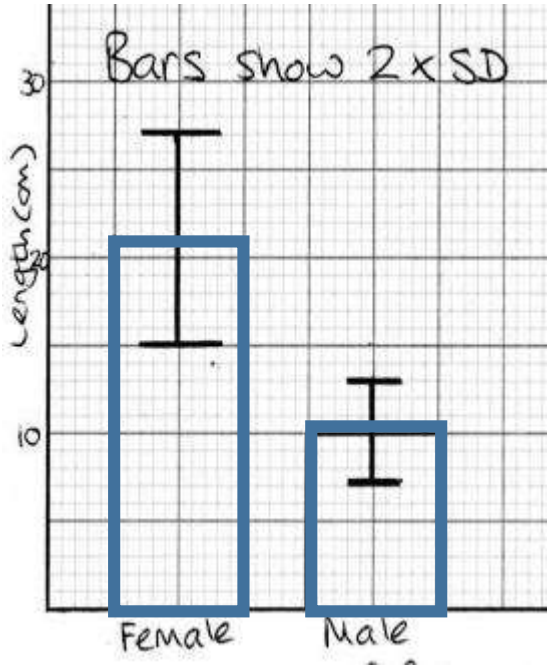
There is no overlap in the ($\pm$2 SD) bars.

This indicates that the differences in the means the size of male and females is **unlikely** to be due to chance.

Describing the results

We can draw a bar chart of the mean and plot the $\pm$ 2 Standard deviations from the mean and look at the **overlap of the bars.**

If there is NO overlap of the error bars overlap, then there is IS significant difference between the two samples.



Error Bars +/-

Note: You cannot say how 'unlikely' this is due to chance – just that it is **unlikely**!

	DATA A (mm)	DATA B (mm)
mean	32.3	48.8
SD	9.3	1.7
2 x SD	18.6	2.4
Mean + (2 x SD)	50.9	51.2
Mean - (2 x SD)	13.7	46.4

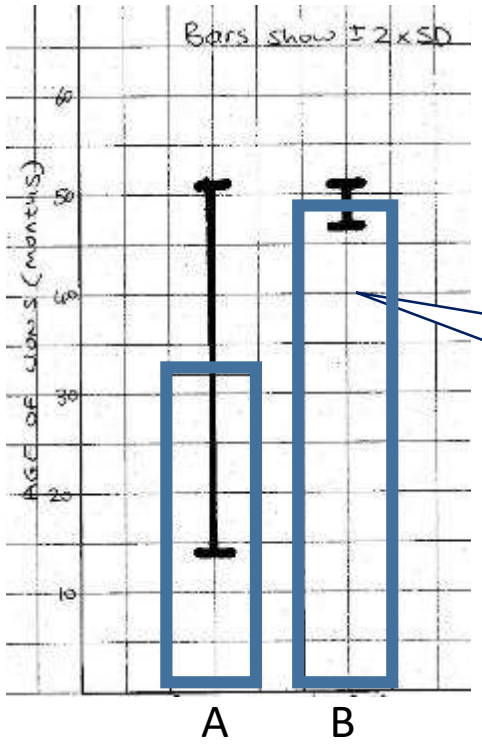
There is an overlap in the (±2 SD) bars.

This indicates that the differences in the means between A and B are **likely** to be due to chance.

Describing the results
We can draw a bar chart of the mean and plot the **± 2** Standard deviations from the mean and look at the

overlap of the bars.

If error bars overlap, then there is **no significant difference** between the two samples.



Note: You cannot say how 'likely' this is due to chance – just that it is **likely**!

Alternative Hypothesis vs Null Hypothesis

Alternative Hypothesis (a hypothesis you want to prove)	Null hypothesis
If the flowers are counted there will be only yellow flowers.	There will be no significant difference between the expected flower color and the observed flower color.
If leaves are exposed to more sunlight, then they will be larger than leaves that receive less sunlight.	There will be no significant difference in the size of the leaves.

Acceptance or **Rejection** of the of the Null Hypothesis

Our calculated value of t is less than the critical value of t.

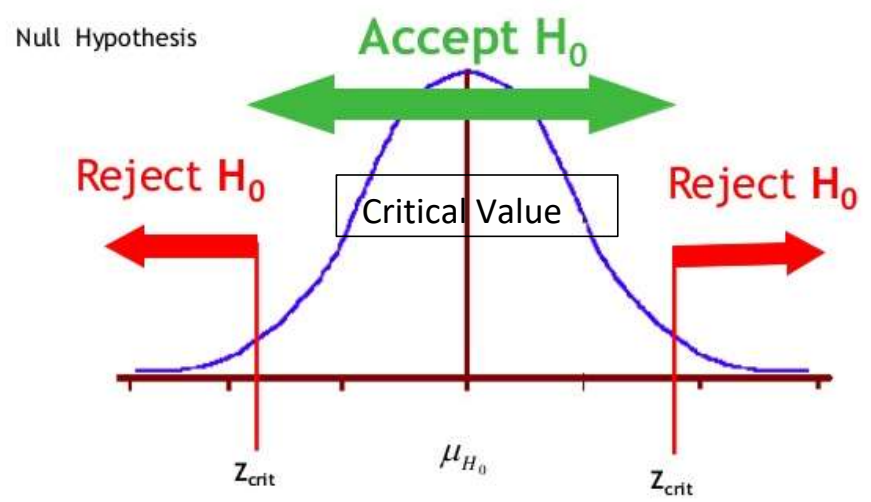
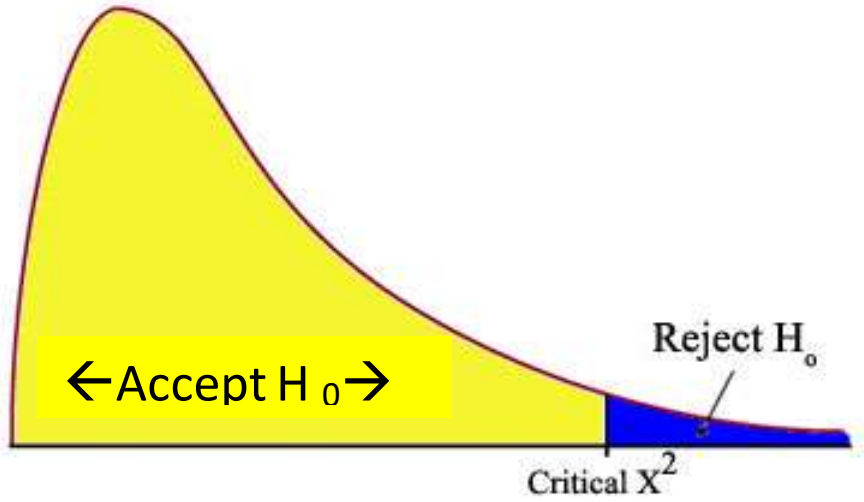
There is more than 5% probability that the differences in the means (mean of **A** and the mean of **B**) **are** due to chance.

We **accept** our null hypothesis.

Our calculated value of Chi-squared is much larger than the critical value of Chi-squared.

There is less than 5% probability that the differences (between the observed and expected data) **are** due to chance.

We **reject** our null hypothesis.

H_0 = Null Hypothesis	
t-test t > critical value reject H_0 t < critical value accept H_0	**Chi Square** χ^2 > critical value reject H_0 χ^2 < critical value accept H_0
	
Our calculated value of t is **greater than** the critical value of t. There is **more than 5% probability** that the differences in the means (mean mass of A and B) **are not** due to chance. We **reject** our null hypothesis.	Our calculated value of Chi-squared is much **larger** than the critical value of Chi-squared. There is **less than 5% probability** that the differences (between the observed and expected data) are due to chance. We **reject** our null hypothesis.
Our calculated value of t is **less than** the critical value of t. There is **more than 5% probability** that the differences in the means (mean mass A and B) **are** due to chance. We **accept** our null hypothesis.	Our calculated value of Chi-squared is **smaller** than the critical value of Chi-squared. There is **more than 5% probability** that the differences (between the observed and expected data) are due to chance. We **accept** our null hypothesis.

Statistical test	When to use it	Criteria for using the test	Examples of use	How to interpret the value you calculate
t-test	You want to know if two sets of continuous data are significantly different from one another.	• You have two sets of continuous, quantitative data (page 494). • You have more than 10 but less than 30 readings for each set of data. • Both sets of data come from populations that have normal distributions. • The standard deviations for the two sets of data are very similar.	Are the surface areas of the leaves on the north-facing side of a tree significantly different from the surface areas on the south-facing side? Are the reaction times of students who have drunk a caffeine-containing drink significantly different from students who have drunk water?	Use a *t*-test table to look up your value of *t*. If this value is greater than the *t* value for a probability of 0.05 (the critical value), then you can say that your two populations are significantly different.
χ^2 test	You want to know if your observed results differ significantly from your expected results.	• You have two or more sets of quantitative data, which belong to two or more discontinuous categories (i.e. they are nominal data – page 494)	Are the numbers of offspring of different phenotypes obtained in a genetic cross significantly different from the expected numbers?	Use a χ^2 table to look up your value of χ^2. If this value is greater than the χ^2 value for a probability of 0.05, then you can say that your observed results differ significantly from your expected results.

The formula for calculating standard deviation is:

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

where:

$\bar{x}$ is the mean

$\sum$ stands for 'sum of'

x refers to the individual values in a set of data

n is the total number of observations (individual values, readings or measurements) in one set of data

s is standard deviation

$\sqrt{}$ is the symbol for square root

Woodland vs garden petals

petal

woodland

garden

Standard deviation

A useful statistic to know about data that have an approximately normal distribution is how far they spread out on either side of the mean value. This is called the standard deviation. The larger the standard deviation, the wider the variation from the mean (Figure P2.6).

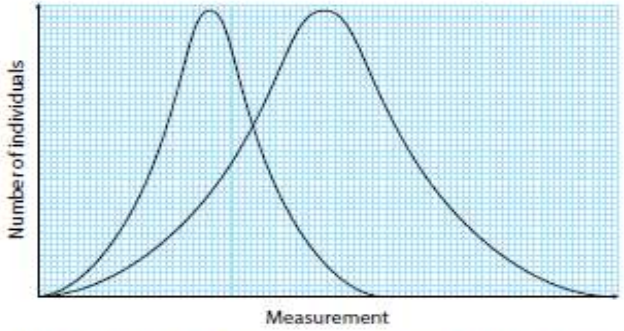


Figure P2.6 Normal distribution curves with small and large standard deviations.

A student measured the length of 21 petals from flowers of a population of a species of plant growing in woodland. These were the results:

Woodland

Petal lengths in woodland population / mm

3.1	3.2	2.7	3.1	3.0	3.2	3.3
3.1	3.1	3.3	3.3	3.2	3.2	3.3
3.2	2.9	3.4	2.9	3.0	2.9	3.2

Standard deviation formula

The formula for calculating standard deviation is:

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

where:

$\bar{x}$ is the mean
Σ stands for 'sum of'
x refers to the individual values in a set of data
n is the total number of observations (individual values, readings or measurements) in one set of data
s is standard deviation
$\sqrt{}$ is the symbol for square root

You may have a calculator that can do all the hard work for you – you just key in the individual values and it will calculate the standard deviation. However, you do need to know how to do the calculation yourself. The best way is to set your data out in a table, and work through it step by step.

1 List the measurement for each petal in the first column of a table like Table P2.1.
2 Calculate the mean for the petal length by adding all the measurements and dividing this total by the number of measurements.
3 Calculate the difference from the mean for each observation. This is $(x - \bar{x})$.
4 Calculate the squares of each of these differences from the mean. This is $(x - \bar{x})^2$.
5 Calculate the sum of the squares. This is $\Sigma(x - \bar{x})^2$.
6 Divide the sum of the squares by $n - 1$.
7 Find the square root of this. The result is the standard deviation, s, for that data set.

Table P2.1 shows the calculation of the standard deviation for petals from plants in the woodland.

garden

QUESTION

P2.5 The student measured the petal length from a second population of the same species of plant, this time growing in a garden. These are the results:

Petal lengths in garden population / mm

2.8	3.1	2.9	3.2	2.9	2.7	3.0
2.8	2.9	3.0	3.2	3.1	3.0	3.2
3.0	3.1	3.3	3.2	2.9		

Show that the standard deviation for this set of data is 0.16.

Calculating Standard Deviation

Step 1: Find the mean.

Step 2: For each data point, find the square of its distance to the mean.

Step 3: Sum the values from Step 2.

Step 4: Divide by the number of data points.

Step 5: Take the square root

3 Calculate the difference between each observation and the mean, $x - \bar{x}$.

1 List each observation, x.

4 Calculate the square of each difference, $(x - \bar{x})^2$.

5 Calculate the sum of the squares of each difference, $\Sigma(x - \bar{x})^2$.

2 Calculate the mean, $\bar{x}$.

6 Divide the sum of the squares by $n - 1$.

7 Find the square root. This is the standard deviation.

x	$(x - \bar{x})$	$(x - \bar{x})^2$
3.1	−0.02	0.001
3.2	0.08	0.006
2.7	−0.42	0.180
3.1	−0.02	0.001
3.0	−0.12	0.015
3.2	0.08	0.006
3.3	0.18	0.031
3.1	−0.02	0.001
3.1	−0.02	0.001
3.3	0.18	0.031
3.3	0.18	0.031
3.2	0.08	0.006
3.2	0.08	0.006
3.3	0.18	0.031
3.2	0.08	0.006
2.9	−0.22	0.050
3.4	0.28	0.076
2.9	−0.22	0.050
3.0	−0.12	0.015
2.9	−0.22	0.050
3.2	0.08	0.006
$\Sigma x = 65.6$		$\Sigma(x - \bar{x})^2 = 0.600$
$n = 21$		$n - 1 = 20$
$\bar{x} = 3.12$		$\frac{\Sigma(x - \bar{x})^2}{n - 1} = 0.03$
		$s = 0.17$

Table P2.1 Calculation of standard deviation for petal length in a sample of plants from woodland. All lengths are in mm.

Calculate Standard Error

Standard error

The 21 petals measured were just a sample of all the thousands of petals on the plants in the wood and in the garden. If we took another sample, would we get the same value for the mean petal length? We cannot be certain without actually doing this, but there is a calculation that we can do to give us a good idea of how close our mean value is to the true mean value for all of the petals in the wood. The calculation works out the **standard error** (S_M) for our data.

Once you have worked out the standard deviation, s, then the standard error is very easy to calculate. The formula is:

$$S_M = \frac{s}{\sqrt{n}}$$

where S_M = standard error
s = standard deviation
n = the sample size (in this case, the number of petals in the sample)

So, for the petals in woodland:

$$S_M = \frac{0.17}{\sqrt{21}} = \frac{0.17}{4.58} = 0.04$$

What does this value tell us?

The standard error tells us how certain we can be that our mean value is the true mean for the population that we have sampled.

We can be 95% certain that – if we took a second sample from the same population – the mean for that second sample would lie within 2 × our value of S_M from the mean for our first sample.

So here, we can be 95% certain that the mean petal length of a second sample would lie within 2 × 0.04 mm of our mean value for the first sample.

QUESTION

P2.6 Show that the standard error for the lengths of the sample of petals taken from the garden (Question P2.5) is also 0.04. Show each step in your working.

Error bars

The standard error can be used to draw error bars on a graph. Figure P2.7 shows the means for the two groups of petals, plotted on a bar chart.

The bars drawn through the tops of the plotted bars are called error bars.

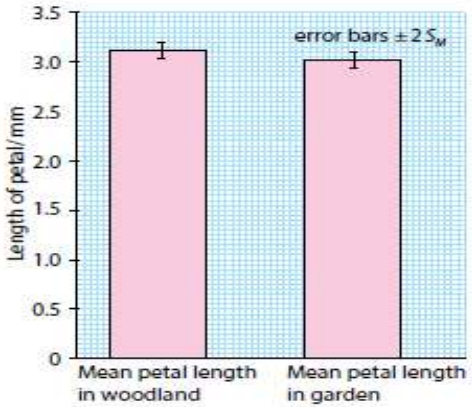


Figure P2.7 Mean petal length of plants in woodland and garden.

QUESTION

P2.7 From the data in Figure P2.7, is there strong evidence that the lengths of the petals in the woodland are significantly different from the lengths of the petals in the garden? Explain your answer.

If we draw an error bar that extends two standard errors above the mean and two standard errors below it, then we can be 95% certain that the true value of the mean lies within this range.

We can use these error bars to help us to decide whether or not there is a significant difference between the petal length in the woodland and the garden. If the error bars overlap, then the difference between the two groups is definitely not significant. If the error bars do not overlap, we still cannot be sure that the difference is significant – but at least we know it is possible that it is. You can also add error bars to line graphs, where your individual points represent mean values.

To find out whether the difference is significant, we can do a further statistical calculation, called a *t*-test.

Calculate *t*-test

The *t*-test

The t-test is used to assess whether or not the means of two sets of data with roughly normal distributions, are significantly different from one another.

For this example, we will use data from another investigation.

The corolla (petal) length of two populations of gentian were measured in mm.

Corolla lengths of population A:
13, 16, 15, 12, 18, 13, 13, 16, 19, 15, 18, 15, 15, 17, 15,

Corolla lengths of population B:
16, 14, 16, 18, 13, 17, 19, 20, 17, 15, 16, 16, 19, 21, 18,

The formula for the *t*-test is:

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

$\bar{x}_1$ is the mean of sample 1
$\bar{x}_2$ is the mean of sample 2
s_1 is the standard deviation of sample 1
s_2 is the standard deviation of sample 2
n_1 is the number of individual measurements in sample 1
n_2 is the number of individual measurements in sample 2

1 For each set of data, calculate the mean.
2 Calculate the differences from the mean of all observations in each data set. This is $x - \bar{x}$.
3 Calculate the squares of these. This is $(x - \bar{x})^2$.
4 Calculate the sum of the squares. This is $\Sigma(x - \bar{x})^2$.
5 Divide this by $n_1 - 1$ for the first set and $n_2 - 1$ for the second set.
6 Take the square root of this. The result is the standard deviation for each set of data.
For population A, $s_1 = 4.24$.
For population B, $s_2 = 4.86$.
7 Square the standard deviation and divide by the number of observations in that sample, for both samples.
8 Add these values together for the two samples and take the square root of this.
9 Divide the difference in the two sample means with the value from step 8. This is *t* and, in this case, is 1.93.
10 Calculate the total degrees of freedom for all the data (ν).
$\nu = (n_1 - 1) + (n_2 - 1) = 28$
11 Refer to the table of *t* values for 28 degrees of freedom and a value of $t = 1.93$ (Table P2.2).

Degrees of freedoms for t-test

Degrees of freedom	Value of *t*			
1	6.31	12.7	63.7	63.6
2	2.92	4.30	9.93	31.6
3	2.35	3.18	5.84	12.9
4	2.13	2.78	4.60	8.61
5	2.02	2.57	4.03	6.87
6	1.94	2.45	3.71	5.96
7	1.90	2.37	3.50	5.41
8	1.86	2.31	3.36	5.04
9	1.83	2.26	3.25	4.78
10	1.81	2.23	3.17	4.59
11	1.80	2.20	3.11	4.44
12	1.78	2.18	3.06	4.32
13	1.77	2.16	3.01	4.22
14	1.76	2.15	2.98	4.14
15	1.75	2.13	2.95	4.07
16	1.75	2.12	2.92	4.02
17	1.74	2.11	2.90	3.97
18	1.73	2.10	2.88	3.92
19	1.73	2.09	2.86	3.88
20	1.73	2.09	2.85	3.85
22	1.72	2.07	2.82	3.79
24	1.71	2.06	2.80	3.75
26	1.71	2.06	2.78	3.71
28	1.70	2.05	2.76	3.67
30	1.70	2.04	2.75	3.65
>30	1.64	1.96	2.58	3.29
Probability that chance could have produced this value of *t*	0.10	0.05	0.01	0.001
Confidence level	10%	5%	1%	0.1%

Table P2.2 Values of *t*.

Using the table of probabilities in the *t*-test

In statistical tests that compare samples, it is the convention to start off by making the assumption that there is no significant difference between the samples. You assume that they are just two samples from an identical population. This is called the null hypothesis. In this case, the null hypothesis would be:

There is no difference between the corolla length in population **A** and population **B**.

ANOVA

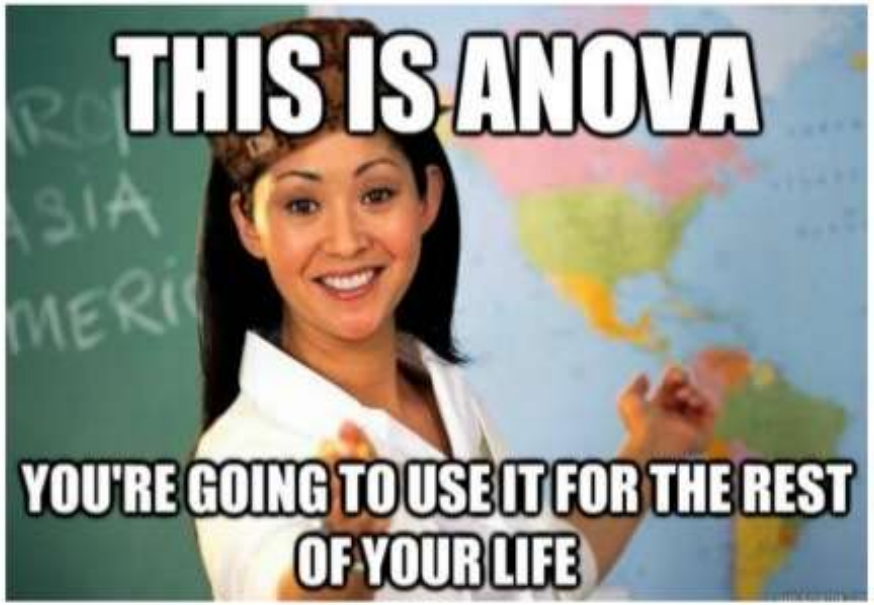
When there are more than 2 groups
then we use

ANOVA

(Analysis of Variances)

ANOVA

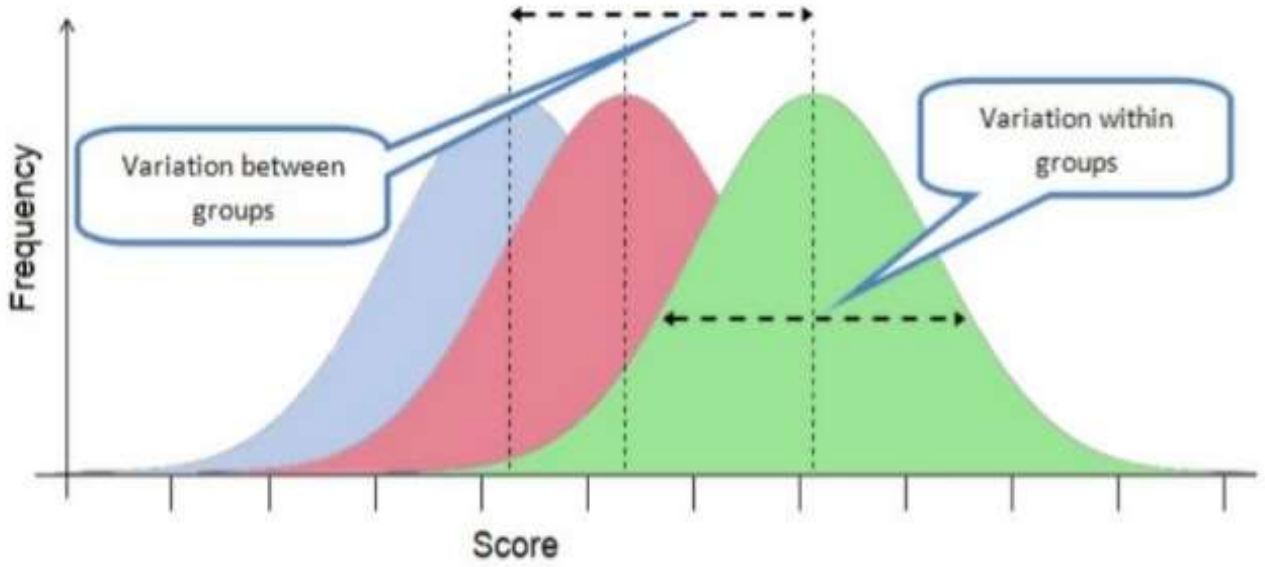
- Analysis of the variance or ANOVA is used to compare differences of means among more than 2 groups.
- It does this by looking at variation in the data and where that variation is found (hence its name).
- Specifically, ANOVA compares the amount of variation between groups with the amount of variation within groups.
- It can be used for both observational and experimental studies.



INTERNSHIPSTUDIO

ANOVA

- When we take samples from a population, we expect each sample mean to differ simply because we are taking a sample rather than measuring the whole population; this is called sampling error but is often referred to more informally as the effects of "chance".
- Thus, we always expect there to be some differences in means among different groups.
- The question is: is the difference among groups greater than that expected to be caused by chance? In other words, is there likely to be a true (real) difference in the population mean?



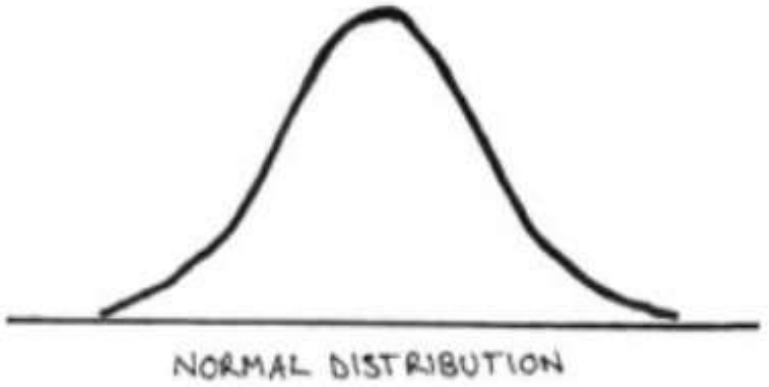
ANOVA

The ANOVA model

- Mathematically, ANOVA can be written as:

$$x_{ij} = \mu_i + \varepsilon_{ij}$$

- where x are the individual data points (i and j denote the group and the individual observation), ε is the unexplained variation and the parameters of the model (μ) are the population means of each group. Thus, each data point (xij) is its group mean plus error.



Assumptions of ANOVA

- The response is normally distributed
- Variance is similar within different groups
- The data points are independent

ANOVA

INTERNSHIPSTUDIO

Hypothesis testing

- Like other classical statistical tests, we use ANOVA to calculate a test statistic (the F-ratio) with which we can obtain the probability (the P-value) of obtaining the data assuming the null hypothesis.

- **Null hypothesis:** all population means are equal
- **Alternative hypothesis:** at least one population mean is different from the rest.

- A significant P-value (usually taken as $P<0.05$) suggests that at least one group mean is significantly different from the others. In other words, a variable with $p<0.05$ allows for us to consider including the variable within a predictive model.

- ANOVA separates the variation in the dataset into 2 parts: between-group and within-group. These variations are called the sums of squares, which can be seen in the following slides.

ANOVA

INTERNSHIPSTUDIO

Calculation of the F ratio

Step 1) Variation between groups

- The between-group variation (or between-group sums of squares, SS) is calculated by comparing the mean of each group with the overall mean of the data.

- Specifically, this is:

$$Between\ SS = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2$$

- We then divide the BSS by the number of degrees of freedom [this is like sample size, except it is n-1, because the deviations must sum to zero, and once you know n-1, the last one is also known] to get our estimate of the mean variation between groups.

ANOVA

Step 2) Variation within groups

- The within-group variation (or the within-group sums of squares) is the variation of each observation from its group mean.

$$SS_r = s^2_{\ group1}\,(n_{group1\ \_\ 1}) + s^2_{\ group2}\,(n_{group2} - 1) + s^2_{\ group3}\,(n_{group3} - 1)$$

- i.e., by adding up the variance of each group times by the degrees of freedom of each group. Note, you might also come across the total SS (sum of). Within SS is then Total SS minus Between SS.

- As before, we then divide by the total degrees of freedom to get the mean variation within groups.

ANOVA

Step 3) The F ratio

- The F ratio is then calculated as:

$$F\ Ratio = \frac{Mean\ Between\ Group\ SS}{Mean\ Within\ Group\ SS}$$

- If the average difference between groups is similar to that within groups, the F ratio is about 1. As the average difference between groups becomes greater than that within groups, the F ratio becomes larger than 1.

- Therefore, variables with higher F Ratio values provide greater explanatory power when utilized in predictive models.

- To obtain a P-value, it can be tested against the F-distribution of a random variable with the degrees of freedom associated with the numerator and denominator of the ratio. The P-value is the probably of getting that F ratio or a greater one. Larger F-ratios gives smaller P-values.

ANOVA

- Do you prefer ketchup or soy sauce?

OR

- If someone asked you this question, your answer would likely depend upon what you were eating. You probably wouldn't dunk your spicy tuna roll in ketchup. And most people (pregnant moms-to-be excluded) don't seem to fancy eating soy sauce with hot French fries.

ANOVA

A Common Error When Using ANOVA to Assess Variables

- So you collect data about your variables of interest, and now you're ready to do your analysis. This is where many people make the unfortunate mistake of looking only at each variables individually.
- In addition to considering how each variable impacts your response variable, you also need to evaluate the interaction between those variables and determine if any of those are significant as well.
- And much like your preference for ketchup versus soy sauce depends upon what you're eating, optimum settings for a given variable will depend upon the settings of another variable when an interaction is present.

ANOVA

INTERNSHIPSTUDIO

How to Evaluate and Interpret an Interaction

- Let's use a weight loss example to illustrate how we can evaluate an interaction between factors. We're evaluating 2 different diets and 2 different exercise programs: one focused on cardio and one focused on weight training. We want to determine which result in greater weight loss. We randomly assign participants to either diet A or B and either the cardio or weight training regimen, and then record the amount of weight they've lost after 1 month.

- Here is a snapshot of the data:

Exercise	Diet	WeightLoss
Cardio	A	22.6
Cardio	A	18.9
Cardio	B	5.9
Cardio	B	5.8
Weights	A	9.7
Weights	A	7.1
Weights	B	9.8
Weights	B	12.7

ANOVA

- **Example:** We are wanting to understand how to explain the WeightLoss variable from the diet variable.

Exercise	Diet	WeightLoss
Cardio	A	22.6
Cardio	A	18.9
Cardio	B	5.9
Cardio	B	5.8
Weights	A	9.7
Weights	A	7.1
Weights	B	9.8
Weights	B	12.7

Analysis of Variance Table					
	Df	**SS**	**Mean SS**	**F Value**	**P(>F)**
Between-group	1	284.6	284.62	12	0.00133
Within-group	38	901.4	23.72		

OR

Analysis of Variance Table					
	Df	**SS**	**Mean SS**	**F Value**	**P(>F)**
Diet	1	284.6	284.62	12	0.00133
Residuals	38	901.4	23.72		

Observations:

- The F Value is well over 1 indicating that this variable has some explanatory value for WeightLoss.
- The P-Value is statistically significant at the 0.05 level.

ANOVA

- Let's look at the ANOVA output for both the Diet and Exercise variables.

Analysis of Variance Table					
	Df	**SS**	**Mean SS**	**F Value**	**P(>F)**
Diet	1	284.6	284.62	13.69	0.000698
Exercise	1	132.1	132.13	6.355	0.016142
Residuals	37	769.2	20.79		

Between Group

Within Group

Observations:

- The Diet variable has a F Value of 13.69
- The Exercise variable has a F Value of 6.355
- Both variables are statistically significant at the 0.05 level

ANOVA

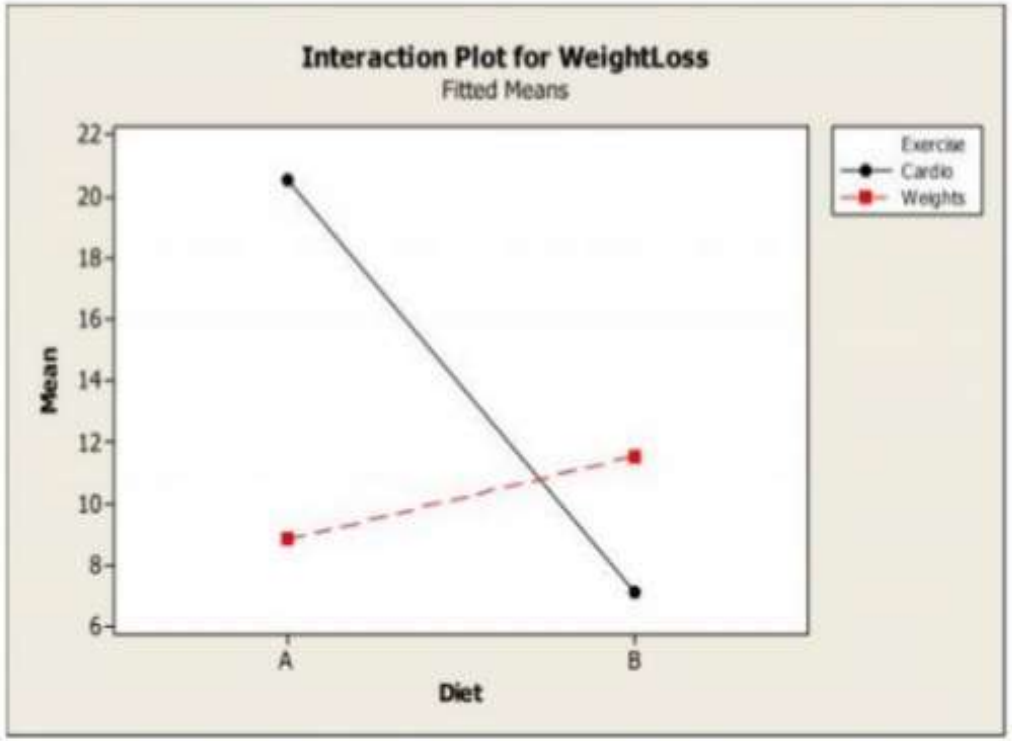
Analysis of Variance Table					
	Df	SS	Mean SS	F Value	P(>F)
Diet	1	284.6	284.62	85.09	5.13E-11
Exercise	1	132.1	132.1	39.5	2.90E-07
Diet : Exercise	1	648.8	648.8	193.97	4.56E-16
Residuals	36	120.4	3.3		

- We can see that the p-value for the Exercise*Diet interaction is 0.000. Because this p-value is so small, we can conclude that there is indeed a significant interaction between Exercise and Diet.

- So which diet is better? Our data suggest it's like asking "ketchup or soy sauce?" The answer is, "It depends."

ANOVA

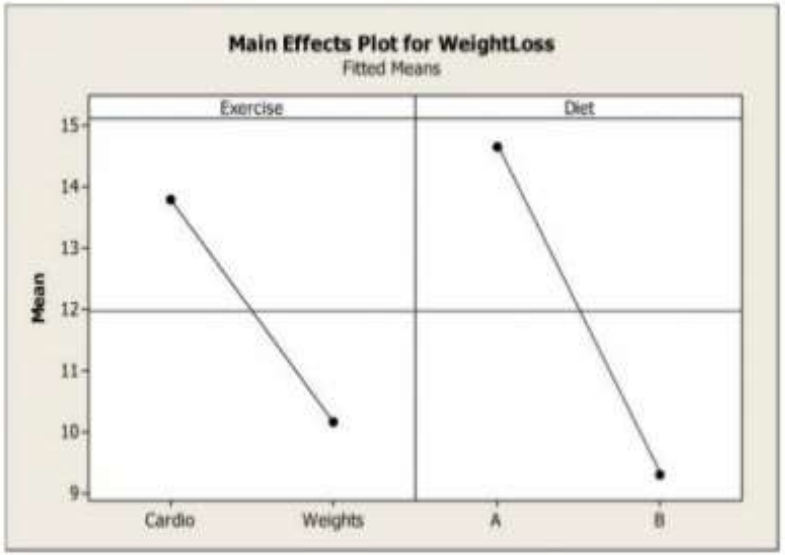
- Since the Exercise*Diet interaction is significant, let's use an interaction plot to take a closer look:



- For participants using the cardio program (shown in black), we can see that diet A is best and results in greater weight loss. However, if you're following the weight training regimen (shown in red), then diet B is results in greater weight loss than A.

ANOVA

- Suppose this interaction wasn't on our radar, and we instead focused only on the individual main effects and their impact on weight loss:



- Based on this plot, we would incorrectly conclude that diet A is better than B. As we saw from the interaction plot, that is only true IF we're looking at the cardio group.

- Clearly, we always need to evaluate interactions when analyzing multiple factors. If you don't, you run the risk of drawing incorrect conclusions...and you might just get ketchup with your sushi roll.

ANOVA

- ANOVA can also be used as a means to compare two linear regression models using the Chi-square measure.

- Here are two regression models we want to compare to each other. The order here is important so make sure you are applying the correct selection of the models.
  - Model 1: y = a
  - Model 2: y = b

Analysis of Variance Table					
	Res. Df	**RSS**	**Df**	**Sum of Sq**	**Pr(>Chi)**
Model 1	2372	2320			
Model 2	2371	2320	1	0.0489	0.82

- The p-value of the test is 0.82. It means that the fitted model "Model 1" is not significantly different from Model 2 at the level of $\alpha=0.05$. Note that this test makes sense only if Model 1 and Model 2 are nested models. (i.e. it tests whether reduction in the residual sum of squares are statistically significant or not).

ANOVA and Linear Regression

- Linear regression is used to analyze continuous relationships; however, regression is essentially the same as ANOVA.
- In ANOVA, we calculate means and deviations of our data from the means.
- In linear regression, we calculate the best line through the data and calculate the deviations of the data from this line.
- The F ratio can be calculated in both.

"I can prove it or disprove it! What do you want me to do?"

Multiple Comparison Test

After ANOVA test, to detect which group is different, many types of test can be performed.

Among these the most Powerful are

1. Duncan's Test
2. Tukey's HSD Test

MANOVA

When we are comparing more than one variable in more than two groups

Than we use

MANOVA

(Multivariate Analysis of Variances)

ANCOVA

In ANOVA table when there is a Co-Variant then we use

ANCOVA

(Analysis of Co Variants)

MANCOVA

In ANCOVA table when there are more than one

Co-Variant then we use

MANCOVA

(Multivariate Analysis of Co Variants)

Chi-square Test

The chi-square test is the most commonly used method for comparing frequencies or proportions.

It is a statistical test used to determine if observed data deviate from those expected.

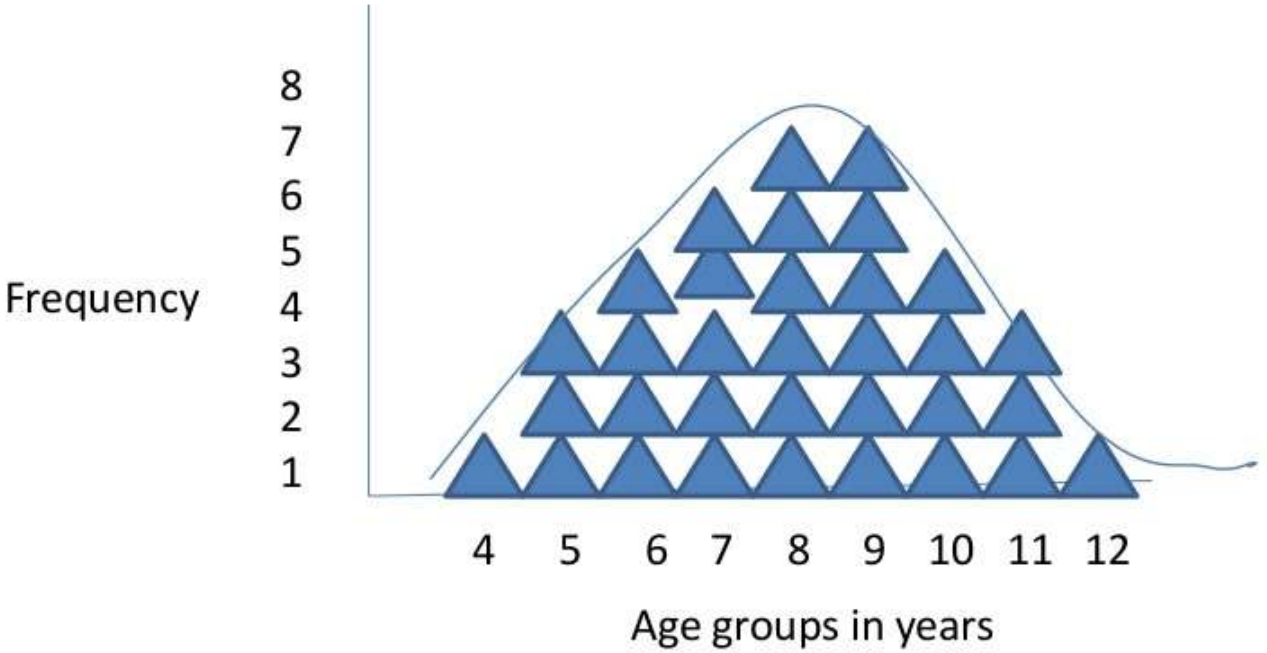
$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

	Observed	Expected	Total
Heads	108	100	208
Tails	92	100	192
Total	200	200	400

Types of Chi-Square Test

- Chi Square Test Goodness-of-fit
- Chi Square Test of independence
- Chi Square Test of Homogeneity

Normal Distribution



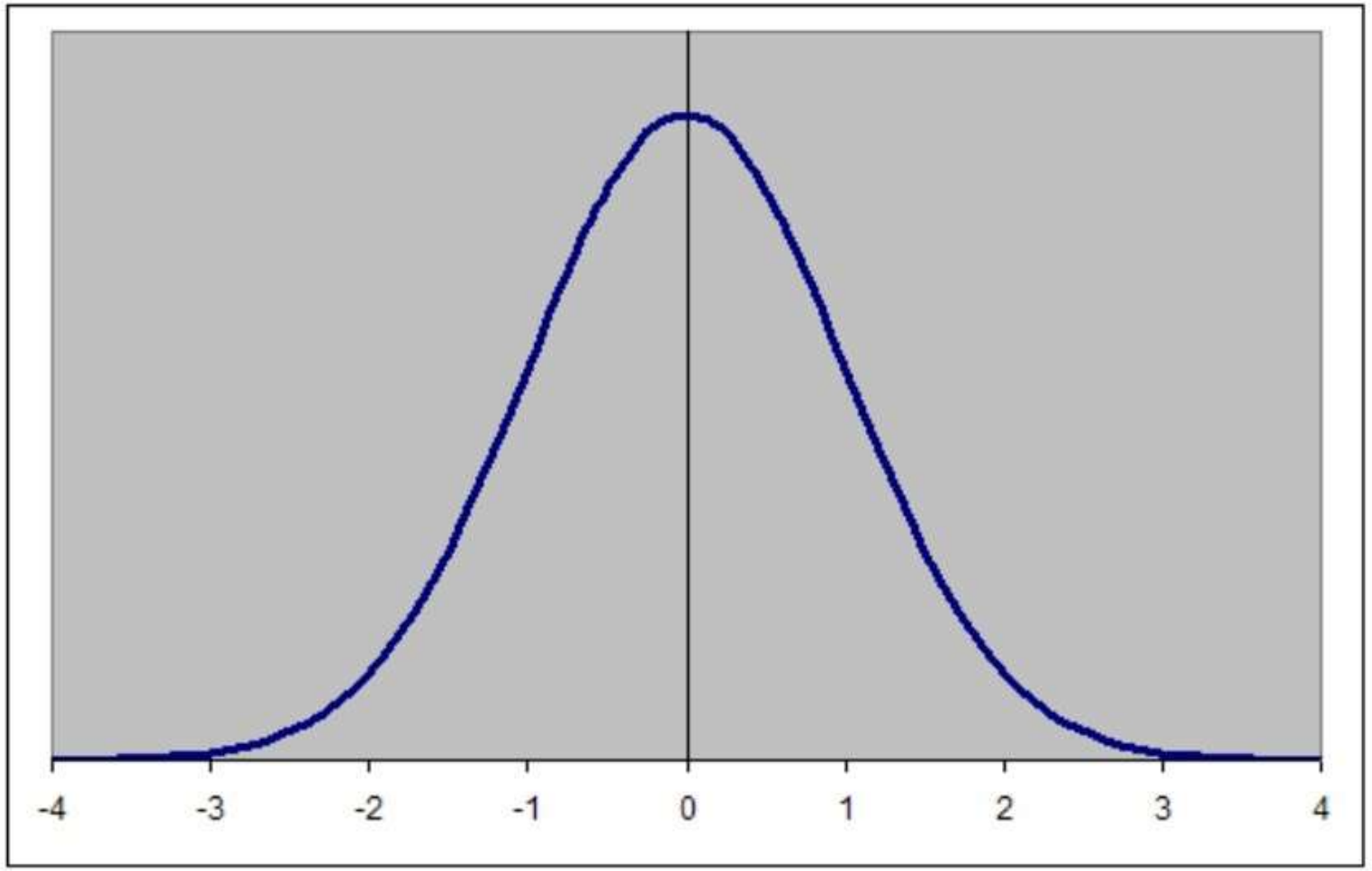
Standard Deviation

It is the mean deviation of the data points from the mean

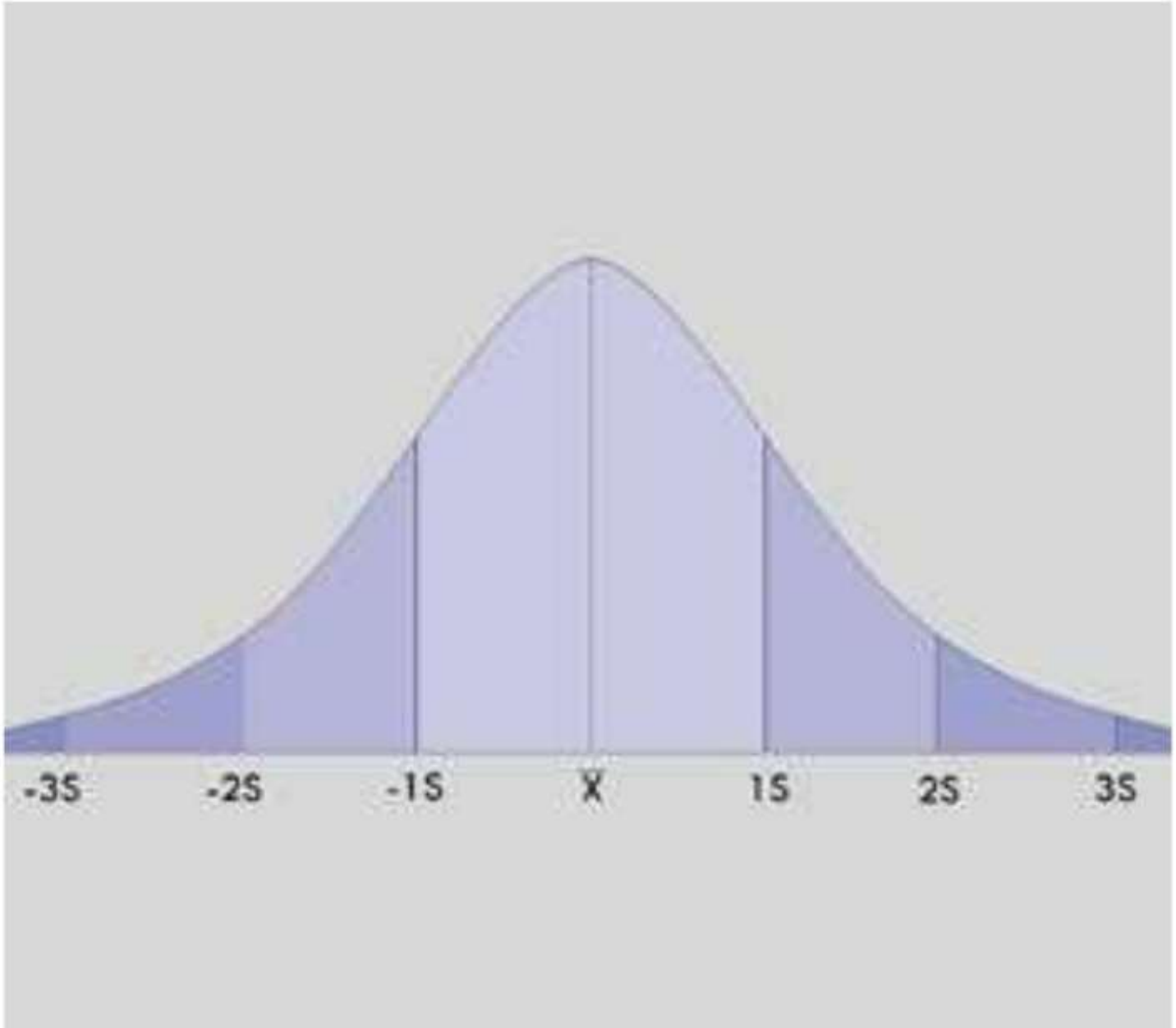
and

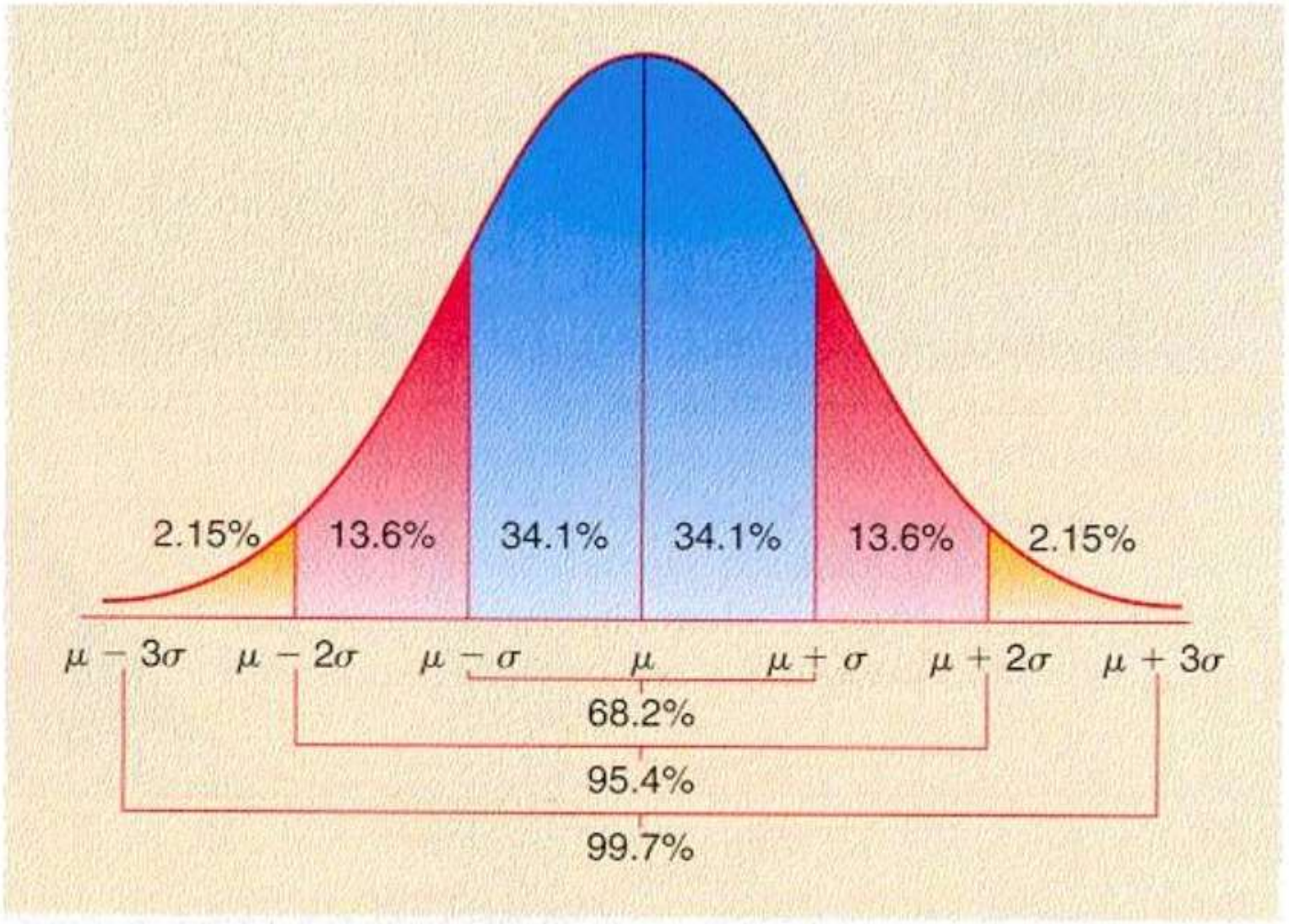
by far the most useful measure of variation

Normal Distribution Curve



Normal Distribution Curve





If the average height for adult men in Pakistan is 70” with a SD of 3”.

— 68% have a height within 67–73 inches

—95% have a height within 64–76 inches

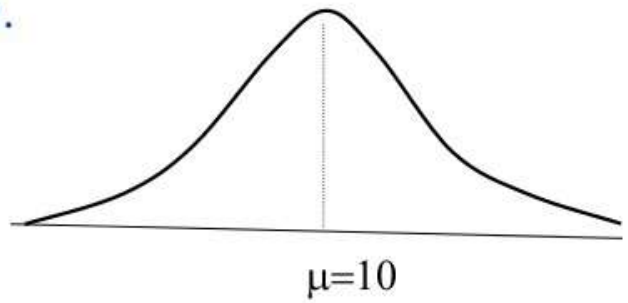
—99.9% have a height within 61-79 inches

If the SD were 0, then all men would be exactly 70”high.

Example: Mean reaction time of a particular drug X is 10 minutes with SD of 2 minutes .

1) Find 50th percentile reaction time.

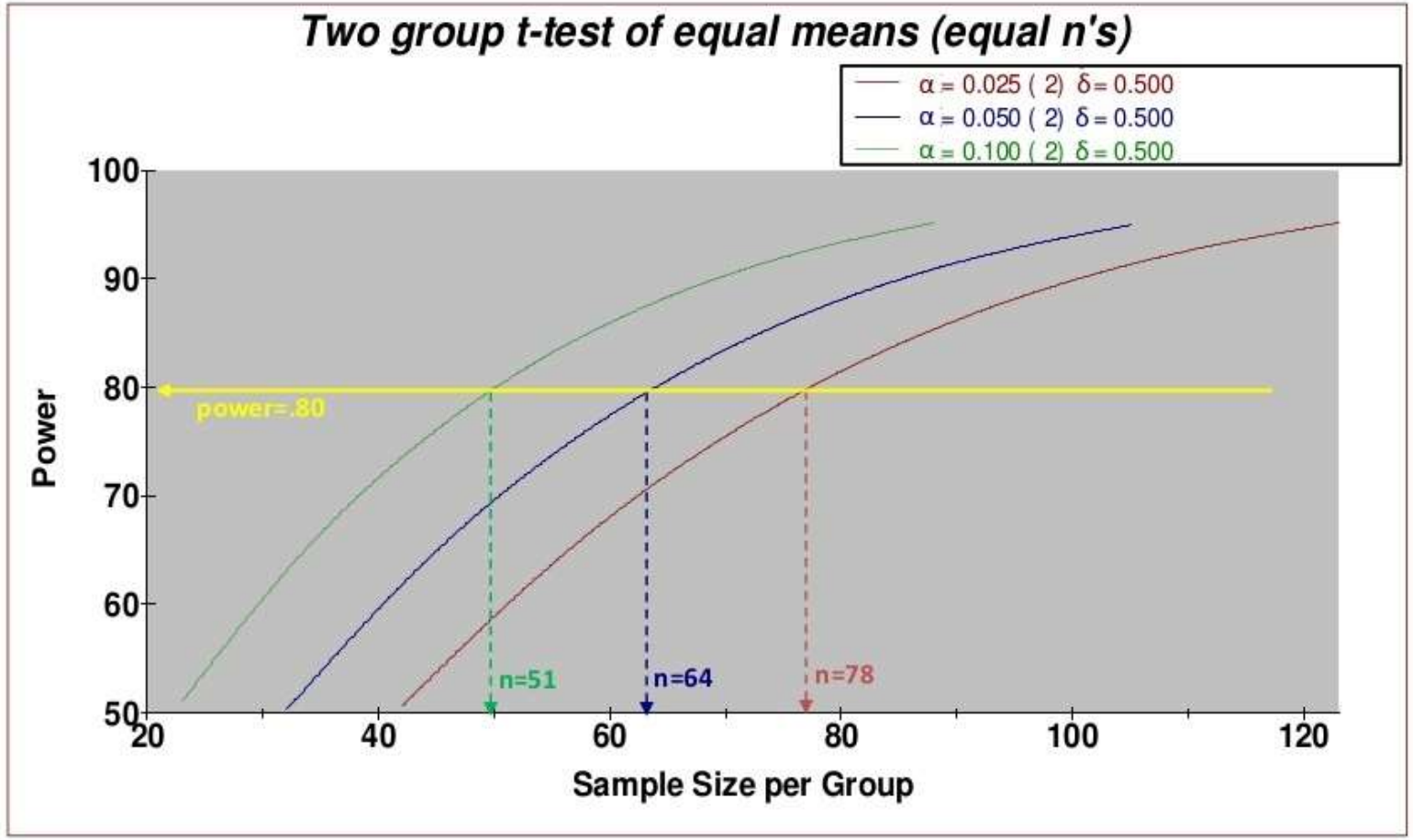
Ans: 10 minutes



2) Find 99.9 Percentile reaction time.

X= 3.09 x 2 + 10 = 16.18 minutes

Relationship Between Alpha(α), Sample Size (n), and Power (1-β)



The Chi-square test for goodness of fit

- The Chi-square (c2) test allows for investigation of statistical significance in the analysis of a frequency distribution.
- Categorical data on variables like sex, education, etc.,
- Allows us to compare the observed frequencies with the expected frequencies based on theoretical ideas about the population distribution.
- Tests whether the data came from a certain probability distribution.

The Chi-square test for goodness of fit

- The process is as follows:
- Formulate the null hypothesis and determine the expected frequency of each answer.
- Determine the appropriate significance level.
- Calculate the c2 value, using the observed frequencies from the sample and the expected frequencies.
- Make the statistical decision by comparing the calculated c2 value with the critical c2 value.

The Chi-square test for goodness of fit

TABLE 12.14 One-way frequency table for brand awareness

Awareness of Tyre Manufacturer's Brand	Frequency
Aware	60
Unaware	40
	100

TABLE 12.15 Calculating the Chi-square statistic

Brand Awareness	Observed Frequency (O_i)	Expected Probability	Expected Frequency (E_i)	$(O_i - E_i)$	$(O_i - E_i)E_i 2$
Aware	60	0.5	50	10	100 50 = 2.0
Unaware	40	0.5	50	–10	100 50 = 2.0
Total	100	1.0	100	0	$\chi^2 = 4.0$

The Chi-square test for goodness of fit

- To calculate the Chi-square statistic:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

  - where c2 is the Chi–square statistic, Oi is the observed frequency in the ith cell, and Ei is the expected frequency in the ith cell.
- Table 12.15 shows that calculated Chi–square value is 4.
- The degrees of freedom is the number of cells associated with column or row data minus one (k–1).
  - k is 2 since there are only two categorical responses.
- Referring to Table B.4 in Appendix B, we find that for 1 degree of freedom (k–1), the Chi-square value is 3.84.
- Since the calculated value is larger than the critical value, the null hypothesis is rejected.

Choosing the appropriate statistical technique

- The choice of statistical analysis depends on:
  - The type of question to be answered
    - Example: researcher concerned with comparing average monthly sales central value would use t–test.
  - The number of variables
    - Example: researcher concerned with one variable at a time would use univariate statistical analysis.
  - The scale of measurement
    - Example: testing a hypothesis about a mean requires interval scaled or ratio scaled data.
    - Parametric versus nonparametric statistics.

Parametric versus hypothesis tests

- Parametric statistics are used when the data are interval or ratio scaled, when the sample size is large, and when the data are drawn from a population with a normal distribution.
- Nonparametric statistics are used when data are either nominal or ordinal.

Table 2 Area under the normal curve

Using the Normal tables

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.49865	.49869	.49874	.49878	.49882	.49886	.49889	.49893	.49897	.49900

Using the p-value

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
How much have you spent, in total, on Internet shopping over the past 12 months?	448	1150.5960	2705.08330	127.80317

One-Sample Test

	Test Value = 800					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
How much have you spent, in total, on Internet shopping over the past 12 months?	2.743	447	.006	350.5960	99.4263	601.7657

$H_0 : \mu \leq 800$

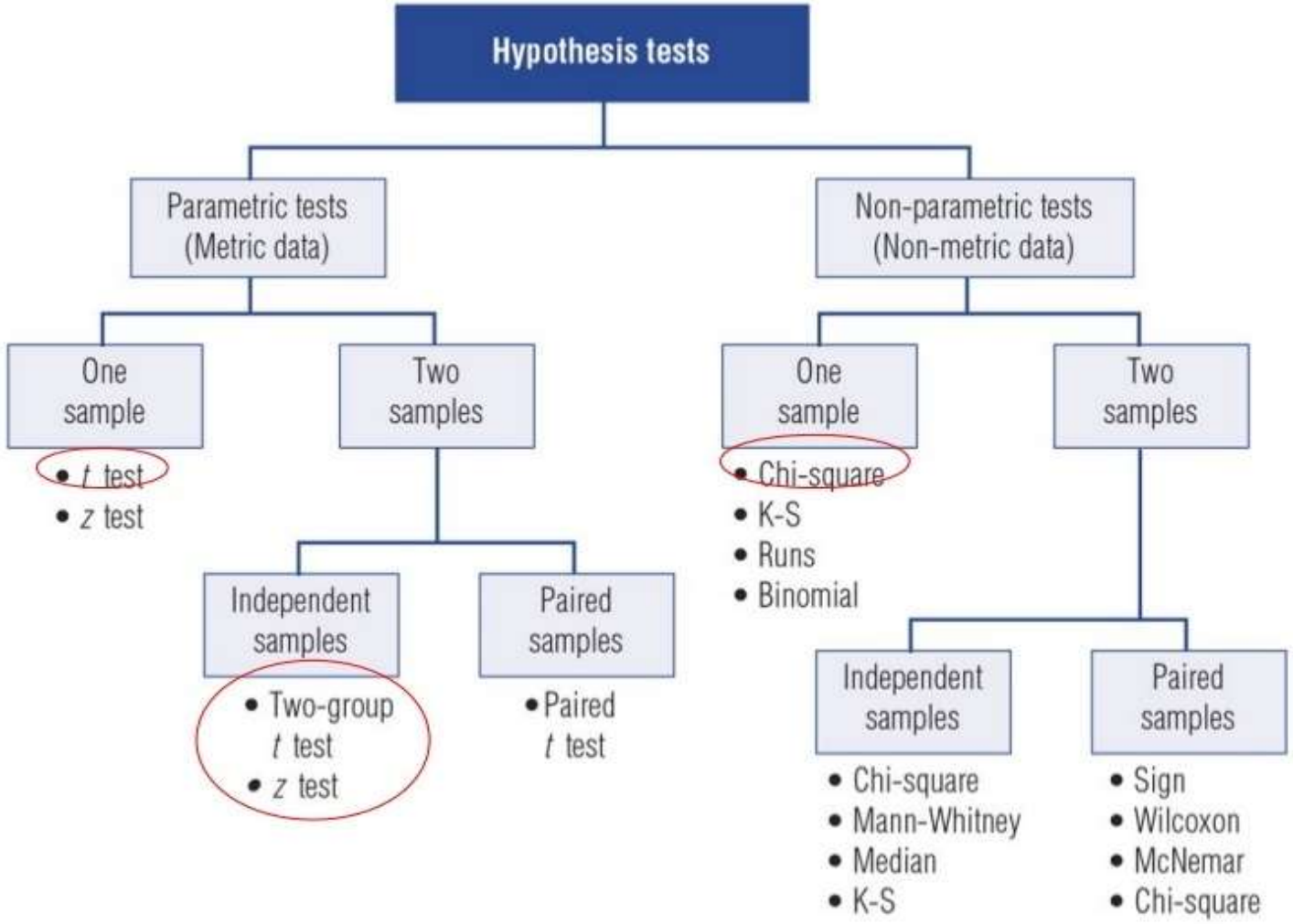
$H_1 : \mu > 800$

If p-value $\leq$ 0.05, Reject H_0

Conclude that the average amount spent on the internet is more than $800 per year

Types of Hypothesis Tests

Figure 10.6
Hypothesis tests related to differences



Parametric Tests

One sample t test

We are testing the hypothesis that the mean satisfaction rating exceeds 4.0, the neutral value on a 7-point scale.

$$H_0 : \mu \leq 4$$

$$H_1 : \mu > 4$$

Parametric Tests cont.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Shopping at this website is usually a satisfying experience	443	5.19	1.079	.051

One-Sample Test

	Test Value = 4					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Shopping at this website is usually a satisfying experience	23.112	442	.000	1.19	1.08	1.29

The p-value < 0.05, hence reject H_0 and conclude that the satisfaction rating for the website is greater than 4 (generally agree)

Parametric Tests cont.

Two Independent samples (Means)

We are testing the hypothesis that mean amount spent on shopping on the Internet is different for males and females

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
How much have you spent, in total, on Internet shopping over the past 12 months?	Male	236	1283.4237	3502.02542	227.96244
	Female	212	1002.7311	1342.04673	92.17215

Parametric Tests cont.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
How much have you spent, in total, on Internet shopping over the past 12 months?	Equal variances assumed	2.166	.142	1.097	446	.273	280.6926	255.91581	-222.258	783.64323
	Equal variances not assumed			1.142	308.922	.255	280.6926	245.89139	-203.141	764.52641

Since p-value > 0.05, t test assuming equal variances should be used

Since p-value > 0.05, we do not reject H_0 and conclude that there is no difference between men and women on the amount they spend on internet shopping

Two Independent samples (Proportions)

We are testing the hypothesis that the proportion of heavy internet users is the same for male and females.

$H_0: \pi_1 = \pi_2$

$H_1: \pi_1 \neq \pi_2$

Sample data

Internet usage * Gender Crosstabulation

Count

		Gender		Total
		Male	Female	
Internet usage	Light	39	58	97
	Heavy	199	154	353
Total		238	212	450

Parametric Tests cont.

$$Z_{calc.} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}$$

$$= \frac{(.84 - .73) - 0}{\sqrt{\frac{(.84)(.16)}{238} + \frac{(.73)(.27)}{212}}}$$

$$= 2.75$$

Parametric Tests cont.

If Z_{crit} = 1.645 (using the normal tables where α =0.05)

We reject H_0 and conclude that there is a difference in the percentage (proportion) of heavy user of the internet between males and females.

Non-Parametric Tests

Chi-square

H_0: There is no association between Internet usage and age of respondents

H_1: There is an association between Internet usage and age of respondents

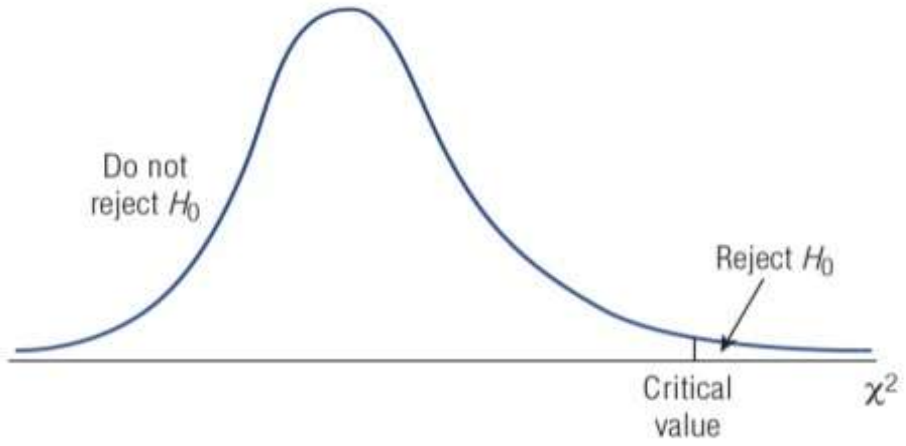


Figure 10.7
Chi-square test of association

Non-Parametric Tests cont.

Internet usage * Age of respondents Crosstabulation

Count

		Age of respondents				
		18 - 24	25 - 39	40-59	60 years or over	Total
Internet usage	Light	22	17	44	14	97
	Heavy	164	107	71	11	353
Total		186	124	115	25	450

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	51.444[a]	3	.000
Likelihood Ratio	47.450	3	.000
Linear-by-Linear Association	43.858	1	.000
N of Valid Cases	450		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.39.

P-value < 0.05 hence reject H_0 and conclude that there is an association between internet usage and age of respondents

Non-Parametric Tests cont.

Chi-square (Another one!)

VU's Open Day organisers are investigating whether visitors' overall rating of Open Day is independent of the age of the visitor. Test at .05 level of significance.

H_0: Overall rating of Open Day and age are independent [no association]

H_1: Overall rating of Open Day and age are not independent [association]

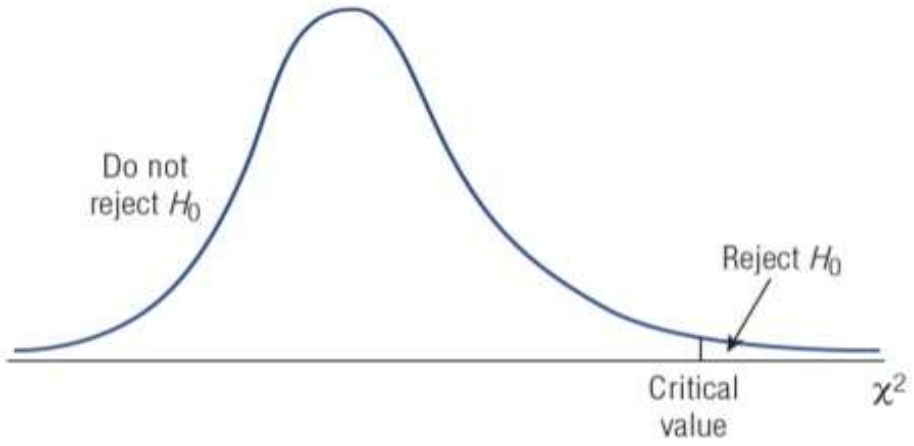


Figure 10.7
Chi-square test of association

Non-Parametric Tests cont.

Overall rating of Open Day * Age of respondent Crosstabulation

Count

		Age of respondent			Total
		18 or under	19 - 29	Over 29	
Overall rating of Open Day	Poor	1		2	3
	Fair	6	2	1	9
	Good	24	15	4	43
	Very Good	100	25	12	137
	Excellent	66	29	18	113
Total		197	71	37	305

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	18.369[a]	8	.019
Likelihood Ratio	15.161	8	.056
Linear-by-Linear Association	.023	1	.879
N of Valid Cases	305		

a. 5 cells (33.3%) have expected count less than 5. The minimum expected count is .36.

P-value < 0.05 hence reject H_0 and conclude that there is an association between ratings of open day and age of respondents

Parametric or Non-parametric?

- Parametric tests are restricted to data that:
  1) show a normal distribution
  2) * are independent of one another
  3) * are on the same *continuous* scale of measurement

- Non-parametric tests are used on data that:
  1) show an other-than normal distribution
  2) are dependent or conditional on one another
  3) in general, do not have a continuous scale of measurement

e.g., the length and weight of something –> parametric
vs.
did the bacteria grow or not grow –> non-parametric

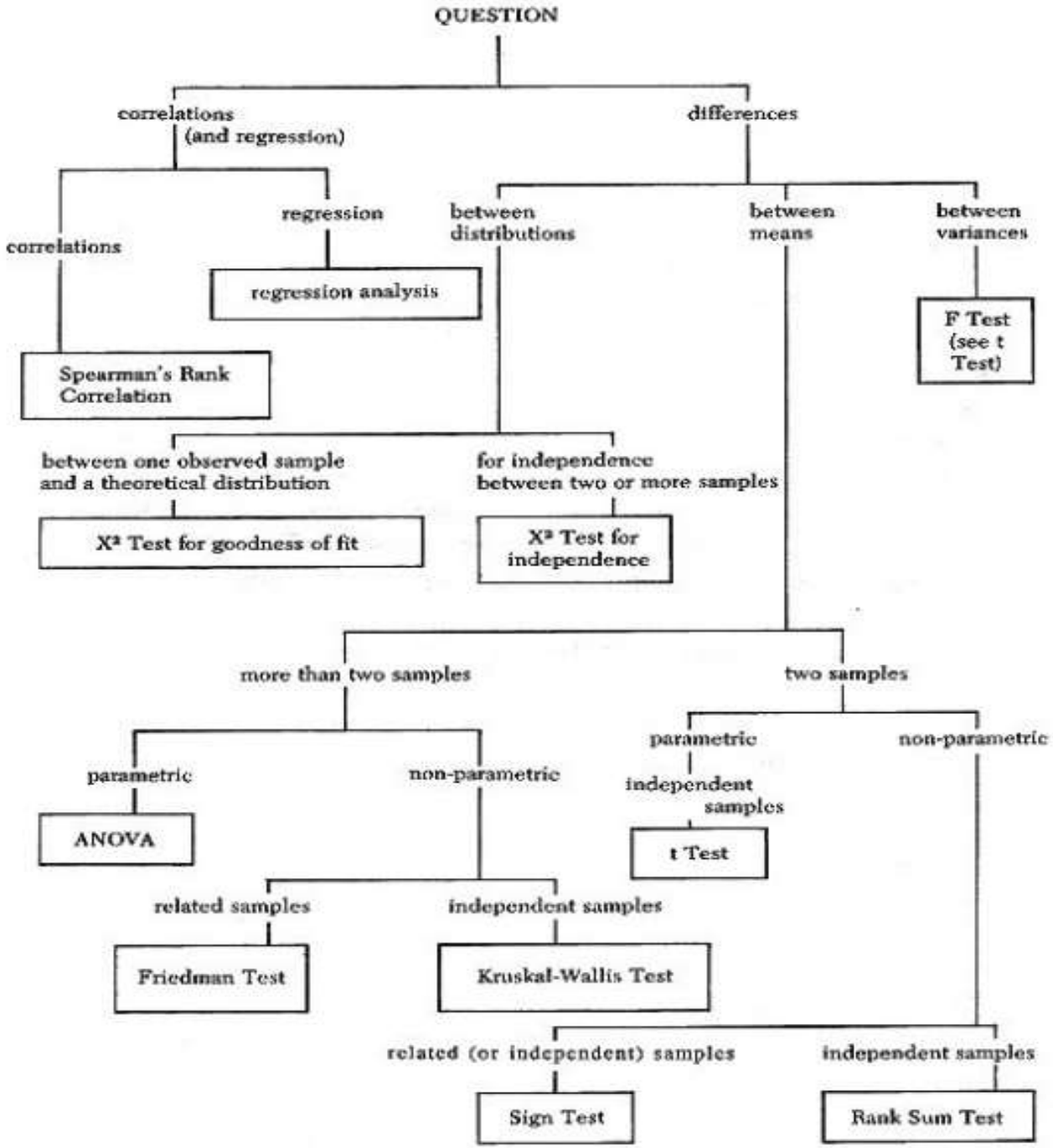
The First Question

After examining your data, ask: does what you're testing seem to be a question of **relatedness** or a question of **difference**?

If **relatedness** (between your control and your experimental samples or between you dependent and independent variable), you will be using tests for correlation (positive or negative) or regression.

If **difference** (your control differs from your experimental), you will be testing for independence between distributions, means or variances. Different tests will be employed if your data show parametric or non-parametric properties.

See Flow Chart on page 50 of HBI.



Tests for Differences

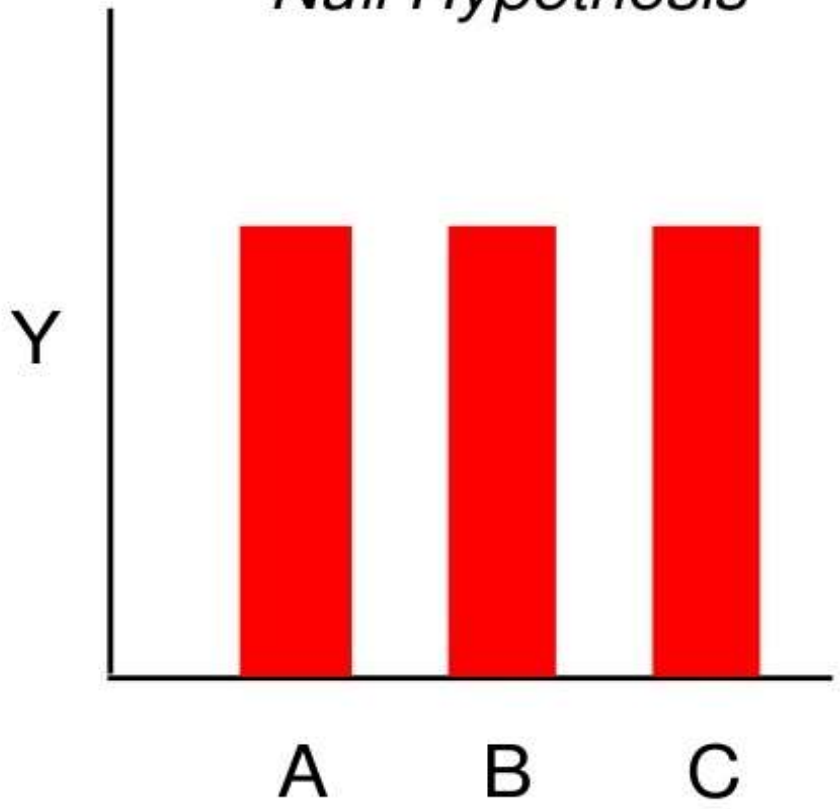
- Between **Means**
  - t-Test - **P**
  - ANOVA - **P**
  - Friedman Test
  - Kruskal-Wallis Test
  - Sign Test
  - Rank Sum Test
- Between **Distributions**
  - Chi-square for goodness of fit
  - Chi-square for independence
- Between **Variances**
  - F-Test – **P**

P – parametric tests

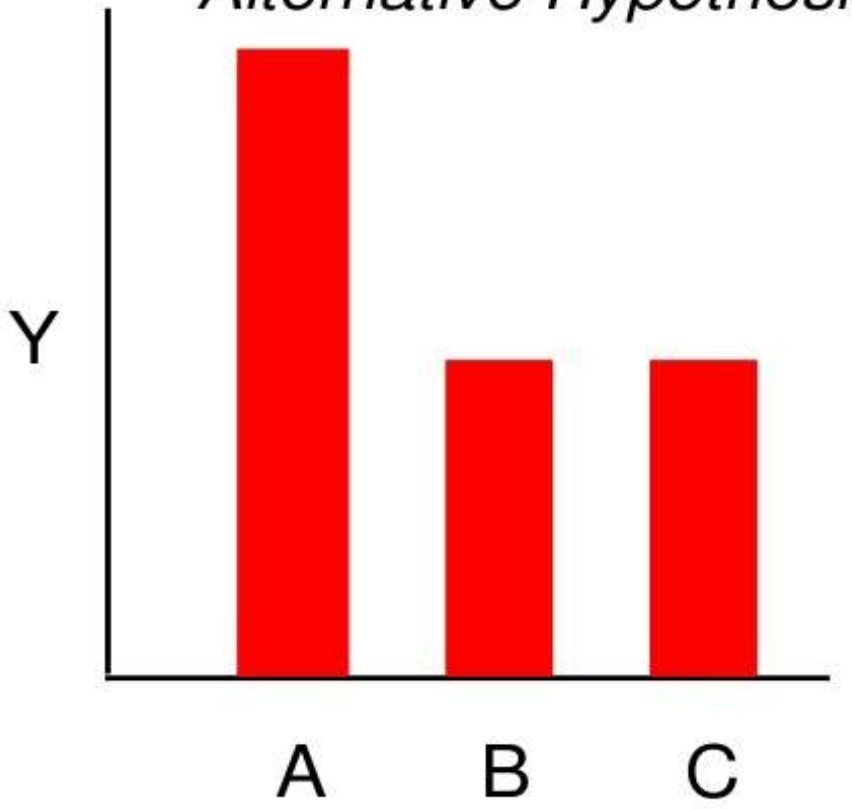
Differences Between Means

Asks whether samples come from populations with different means

Null Hypothesis



Alternative Hypothesis



There are different tests if you have 2 vs more than 2 samples

Differences Between Means – Parametric Data

t-Tests compare the means of **two *parametric*** samples

E.g. Is there a difference in the mean height of men and women?

HBI: t-Test

Excel: t-Test (paired and unpaired) – in Tools – Data Analysis

A researcher compared the height of plants grown in high and low light levels. Her results are shown below. Use a T-test to determine whether there is a statistically significant difference in the heights of the two groups

Low Light	High Light
49	45
31	40
43	59
31	58
40	55
44	50
49	46
48	53
33	43

Differences Between Means – Parametric Data

ANOVA (Analysis of Variance) compares the means of **two or more *parametric*** samples.

E.g. Is there a difference in the mean height of plants grown under red, green and blue light?

HBI: ANOVA

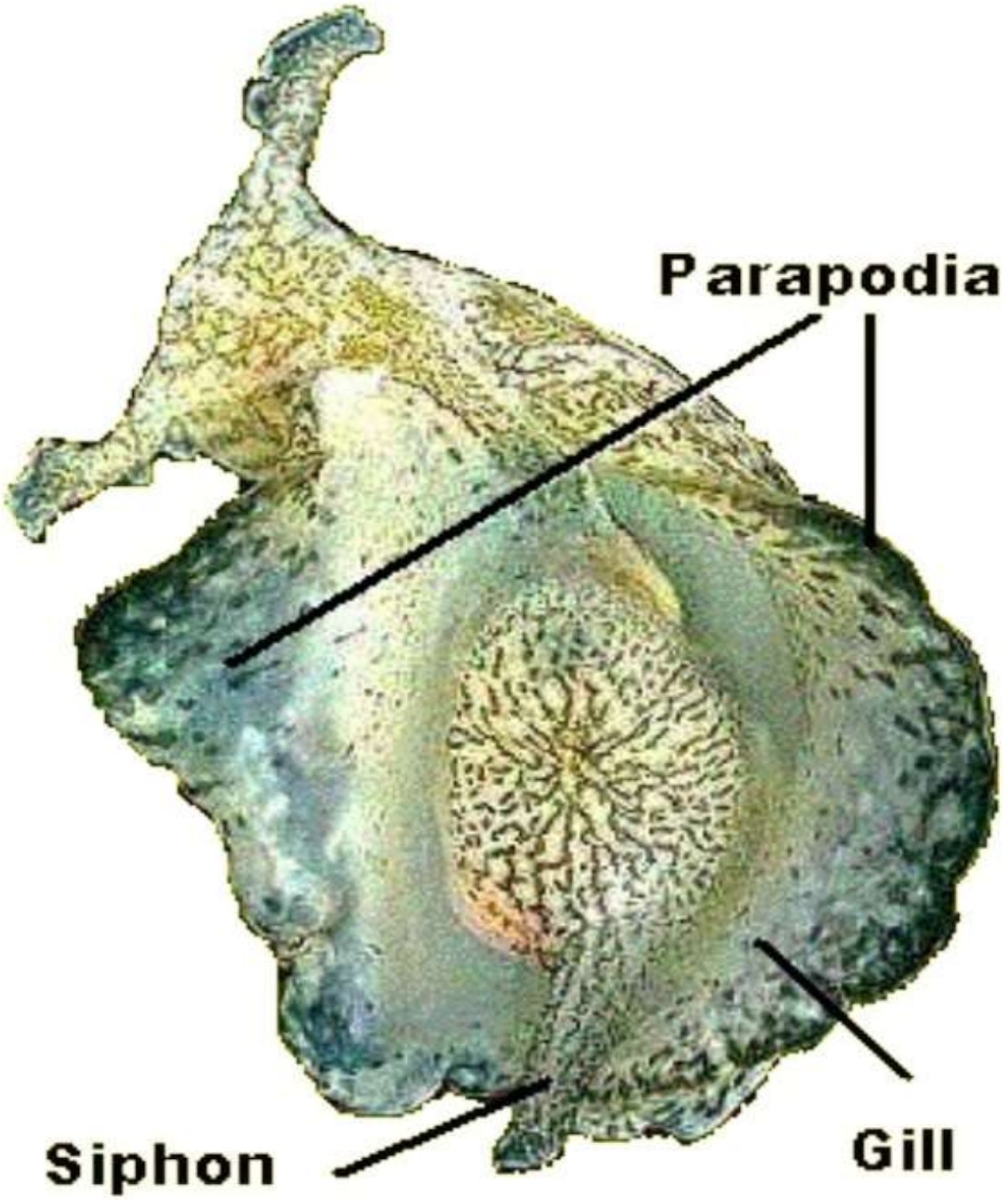
Excel: ANOVA – check type under Tools – Data Analysis

A researcher fed pigs on four different foods. At the end of a month feeding, he weighed the pigs. Use an ANOVA test to determine if the different foods resulted in differences in growth of the pigs.

weight of pigs fed different foods

food 1	food 2	food 3	food 4
60.8	68.7	102.6	87.9
57.0	67.7	102.1	84.2
65.0	74.0	100.2	83.1
58.6	66.3	96.5	85.7
61.7	69.8		90.3

Aplysia punctata – the sea hare



Aplysia parts

Differences Between Means – Non-Parametric Data

The Sign Test compares the means of **two "paired"**, ***non-parametric*** samples

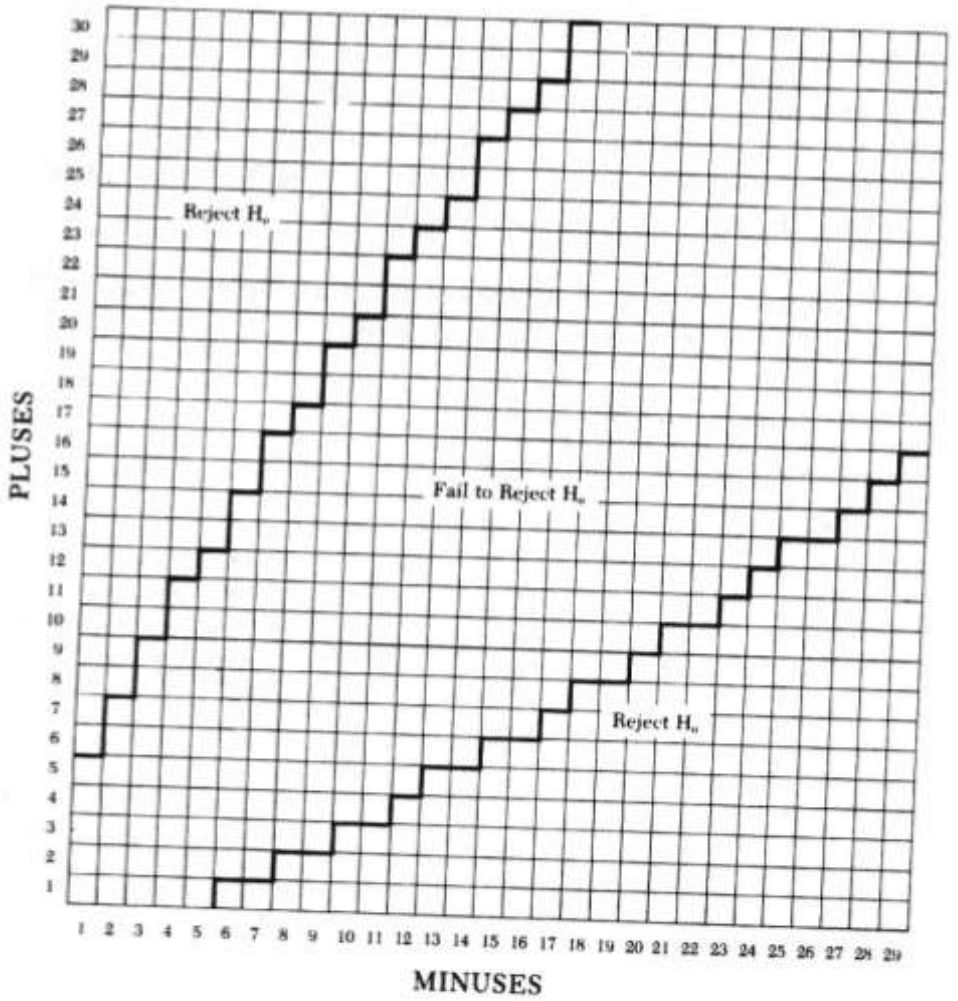
E.g. Is there a difference in the gill withdrawal response of *Aplysia* in night versus day? Each subject has been tested once at night and once during the day –> paired data.

HBI: Sign Test

Excel: N/A

Subject	Night Response	Day Response
1	2	5
2	1	3
3	2	2

Table for the Sign Test at the alpha level of .05.[1]



1. Constructed from values found in The Chemical Rubber Co., *Handbook of Probability and Statistics*, 2nd ed. (W.H. Beyer, ed.) The Chemical Rubber Company, Cleveland, Ohio. 1968.

Differences Between Means – Non-Parametric Data

The Friedman Test is like the Sign test, (compares the means of "paired", non-parametric samples) for **more than two samples**.

E.g. Is there a difference in the gill withdrawal response of *Aplysia* between morning, afternoon and evening? Each subject has been tested once during each time period –> paired data

HBI: Friedman Test

Excel: N/A

Subject	Morning Response	Afternoon Response	Evening. Response
1	4	3	2
2	5	2	1
3	3	4	3

Table of critical values for the Friedman Test at the .05 alpha level.[1]

k	b	Critical Value
3	3	6.000
3	4	6.500
3	5	6.400
3	6	7.000
3	7	7.143
3	8	6.250
3	9	6.222
4	2	6.000
4	3	7.400
4	4	7.800

(for larger sample sizes, use the χ^2 table on page 95, with k-1 d.f.)

1. M. Friedman, 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* 32:675–701.

Differences Between Means – Non-Parametric Data

The Rank Sum test compares the means of **two *non-parametric*** samples

E.g. Is there a difference in the gill withdrawal response of *Aplysia* in night versus day? Each subject has been tested once, either during the night ***or*** during the day –> unpaired data.

HBI: Rank Sum

Excel: N/A

Subject	Night Response	Day Response
1		5
2	1	
3		2
4	3	
5		4
6	1	
7		5

Table of critical values for the Rank Sum Test at the .05 alpha level.[1]

sample sizes for n_2	sample sizes for n_1													
	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4			10											
5		6	11	17										
6		7	12	18	26									
7		7	13	20	27	36								
8	3	8	14	21	29	38	49							
9	3	8	15	22	31	40	51	63						
10	3	9	15	23	32	42	53	65	78					
11	4	9	16	24	34	44	55	68	81	96				
12	4	10	17	26	35	46	58	71	85	99	115			
13	4	10	18	27	37	48	60	73	88	103	119	137		
14	4	11	19	28	38	50	63	76	91	106	123	141	160	
15	4	11	20	29	40	52	65	79	94	110	127	145	164	185
16	4	12	21	31	42	54	67	82	97	114	131	150	169	
17	5	12	21	32	43	56	70	84	100	117	135	154		
18	5	13	22	33	45	58	72	87	103	121	139			
19	5	13	23	34	46	60	74	90	107	124				
20	5	14	24	35	48	62	77	93	110					
21	6	14	25	37	50	64	79	95						
22	6	15	26	38	51	66	82							
23	6	15	27	39	53	68								
24	6	16	28	40	55									
25	6	16	28	42										
26	7	17	29											
27	7	17												
28	7													

1. Adapted from R.G.D. Steel and J.H. Torrie. *Principles and Procedures of Statistics*. McGraw-Hill Book Co., N.Y. 1960.

Differences Between Means – Non-Parametric Data

The Kruskal-Wallis Test compares the means of **more than two *non-parametric*, non-paired** samples

E.g. Is there a difference in the gill withdrawal response of *Aplysia* in night versus day? Each subject has been tested once, either during the morning, afternoon or evening –> unpaired data.

HBI: Kruskal-Wallis Test

Excel: N/A

Subject	Morning Response	Afternoon Response	Evening. Response
1	4		
2	5		
3		4	
4		3	
5			2
6			3

Table of critical values at the .05 level for the Kruskal-Wallis Test[1]

Sample Sizes			
n_1	n_2	n_3	Critical Values
3	2	2	4.714
3	3	1	5.143
3	3	2	5.361
3	3	3	5.600
4	2	1	4.8214*
4	2	2	5.333
4	3	1	5.208
4	3	2	5.444
4	3	3	5.727
4	4	1	4.967
4	4	2	5.455
4	4	3	5.598
4	4	4	5.692
5	2	1	5.000
5	2	2	5.160
5	3	1	4.960
5	3	2	5.251
5	3	3	5.648
5	4	1	4.986
5	4	2	5.273
5	4	3	5.656
5	4	4	5.657
5	5	1	5.127
5	5	2	5.338
5	5	3	5.705
5	5	4	5.666
5	5	5	5.780

(*this is actually at .057, no alpha level of .05 was available at this sample size)

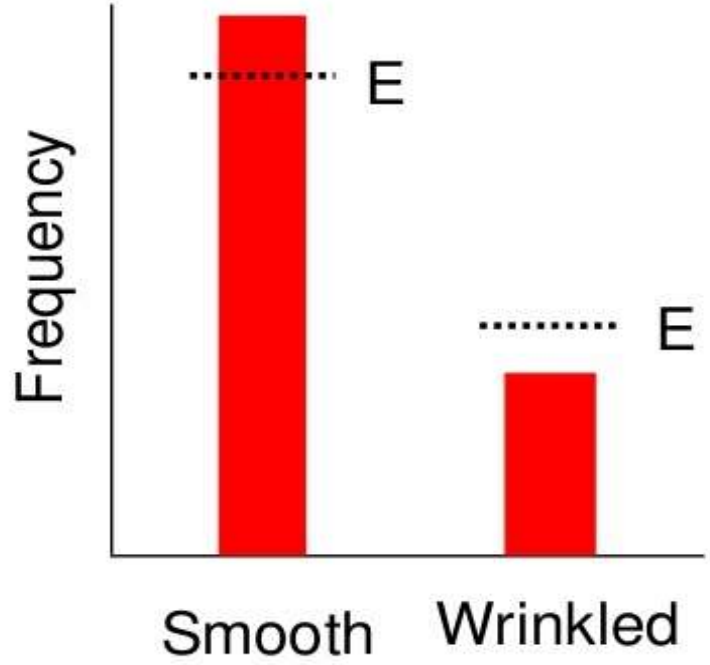
1. Adapted from J.H. Zar, *Biostatistical Analysis*. Prentice-Hall, Inc. Englewood Cliffs, N.J. 1974.

Differences Between Distributions

Chi square tests compare observed frequency distributions, either to theoretical expectations or to other observed frequency distributions.

Differences Between Distributions

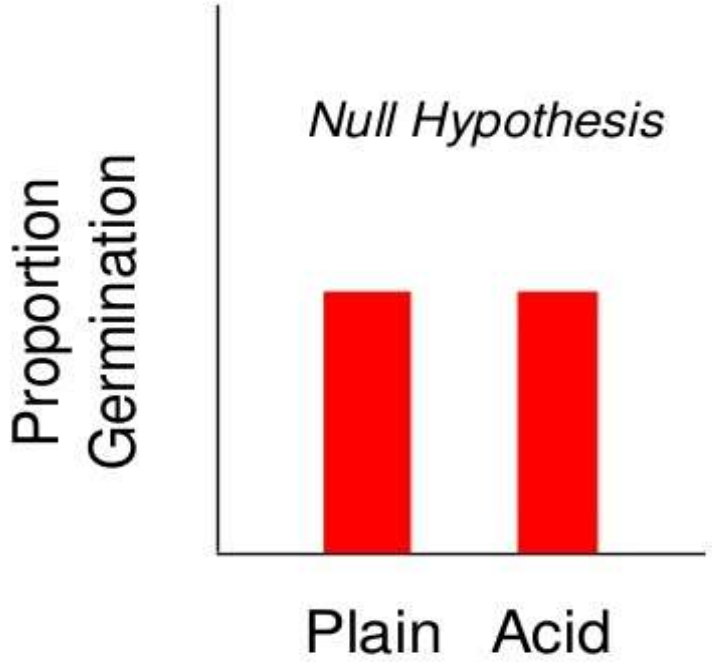
E.g. The F2 generation of a cross between a round pea and a wrinkled pea produced 72 round individuals and 20 wrinkled individuals. Does this differ from the expected 3:1 round:wrinkled ratio of a simple dominant trait?

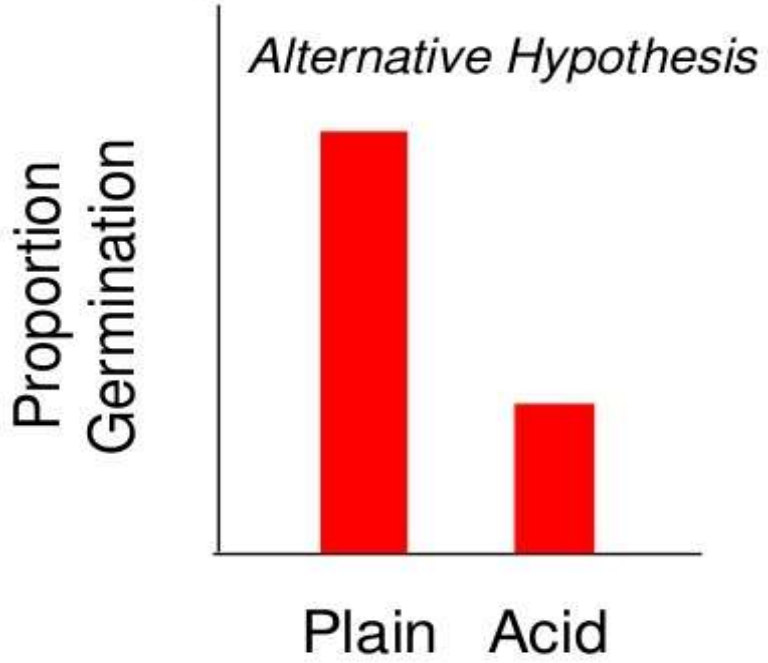


HBI: Chi-Square One Sample Test (goodness of fit)
Excel: Chitest – under Function Key – Statistical

Differences Between Distributions

E.g. 67 out of 100 seeds placed in plain water germinated while 36 out of 100 seeds placed in “acid rain” water germinated. Is there a difference in the germination rate?





HBI: Chi-Square Two or More Sample Test (independence)

Excel: Chitest – under Function key - Statistical

Primary Scales of Measurement

Scale					
Nominal	Numbers Assigned to Runners				**Finish**
Ordinal	Rank Order of Winners	 **Third place**	 **Second place**	 **First place**	**Finish**
Interval	Performance Rating on a 0 to 10 Scale	**8.2**	**9.1**	**9.6**	
Ratio	Time to Finish, in	**15.2**	**14.1**	**13.4**	

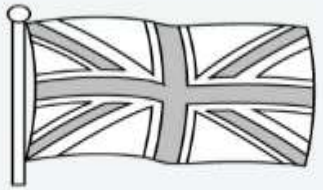
Exhibit 4.3 Levels of Measurement

Qualitative

Nominal or categorical level of measurement: Nationality

American

Canadian

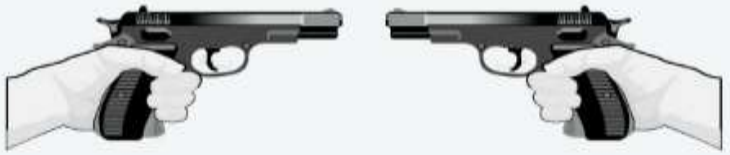


British

Quantitative

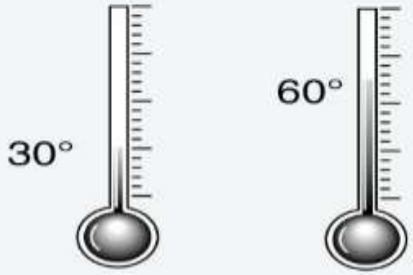
Ordinal level of measurement: Level of conflict

Low



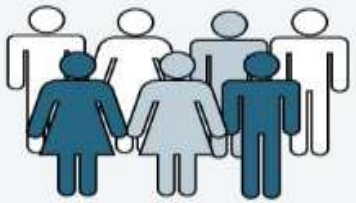
High

Interval level of measurement: Temperature in degrees Fahrenheit

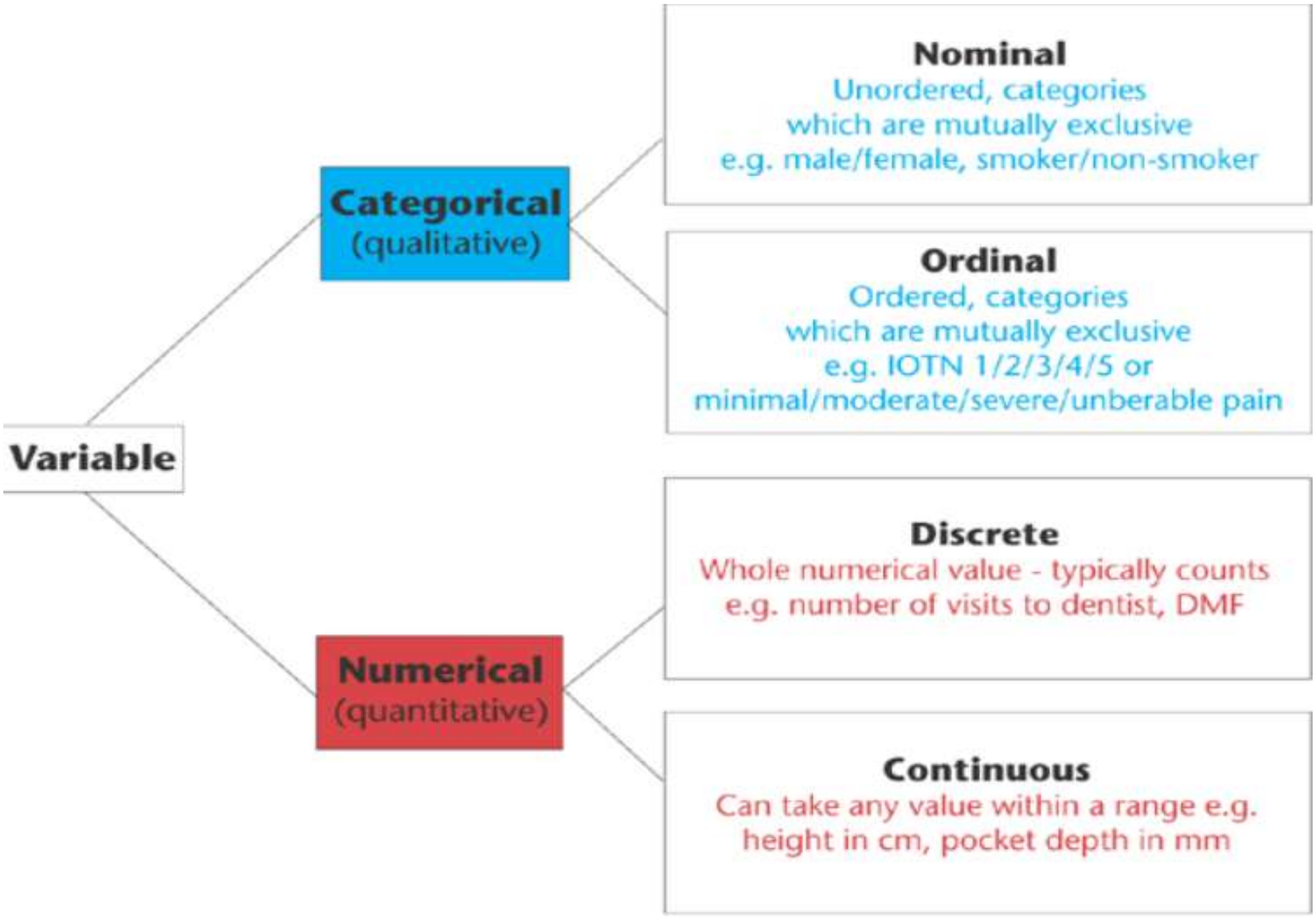


Ratio level of measurement: Group size

5



7



Types of data

Level	Properties	Observations reflect	Example	Type of data
Nominal	Classification	Differences in kind	Favorite food	Qualitative
Ordinal	Order classification	Differences in degree	Letter grade	Ranked
Interval/Ratio*	• Equal intervals • Order classification	Differences in total amount	Height	Quantitative

Correlation

Correlations look for relationships between two variables which may not be functionally related. The variables may be ordinal, interval, or ratio scale data. Remember, correlation does not prove causation; thus there may not be a cause and effect relationship between the variables.

E.g. Do species of birds with longer wings also have longer necks?

HBI: Spearman's Rank Correlation (NP)

Excel: Correlation (P)

Question – is there a relationship between students aptitude for mathemathics and for biology?

Student	Math score	Math Rank	Biol. score	Biology rank
1	57	3	83	7
2	45	1	37	1
3	72	7	41	2
4	78	8	84	8
5	53	2	56	3
6	63	5	85	9
7	86	9	77	6
8	98	10	87	10
9	59	4	70	5
10	71	6	59	4

Table of critical values for different sample sizes at the .05 alpha level to be used with the Spearman's Rank Correlation test.[1] (n = sample size)

n	critical value	n	critical value	n	critical value	n	critical value
5	1.00	27	0.382	49	0.282	92	0.205
6	0.886	28	0.375	50	0.279	94	0.203
7	0.786	29	0.368	52	0.274	96	0.201
8	0.738	30	0.362	54	0.268	98	0.199
9	0.700	31	0.356	56	0.264	100	0.197
10	0.648	32	0.350	58	0.259		
11	0.618	33	0.345	60	0.255		
12	0.587	34	0.340	62	0.250		
13	0.560	35	0.335	64	0.246		
14	0.538	36	0.330	66	0.243		
15	0.521	37	0.325	68	0.239		
16	0.503	38	0.321	70	0.235		
17	0.485	39	0.317	72	0.232		
18	0.472	40	0.313	74	0.229		
19	0.460	41	0.309	76	0.226		
20	0.447	42	0.305	78	0.221		
21	0.435	43	0.301	80	0.220		
22	0.425	44	0.298	82	0.217		
23	0.415	45	0.294	84	0.215		
24	0.406	46	0.291	86	0.212		
25	0.398	47	0.288	88	0.210		
26	0.390	48	0.285	90	0.207		

1. Adapted from J.H. Zar. *Biostatistical Analysis*. Prentice-Hall, Englewood Cliffs, N.J. 1974.

Regression

Regressions look for functional relationships between two continuous variables. A regression assumes that a change in X causes a change in Y.

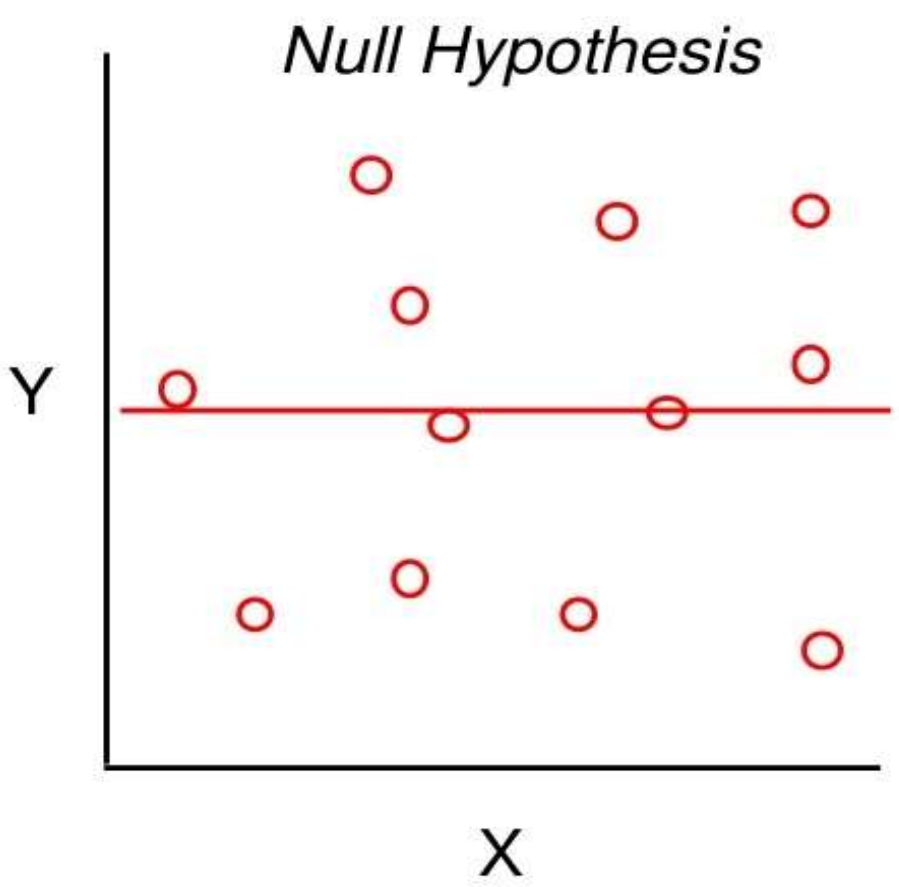
E.g. Does an increase in light intensity cause an increase in plant growth?

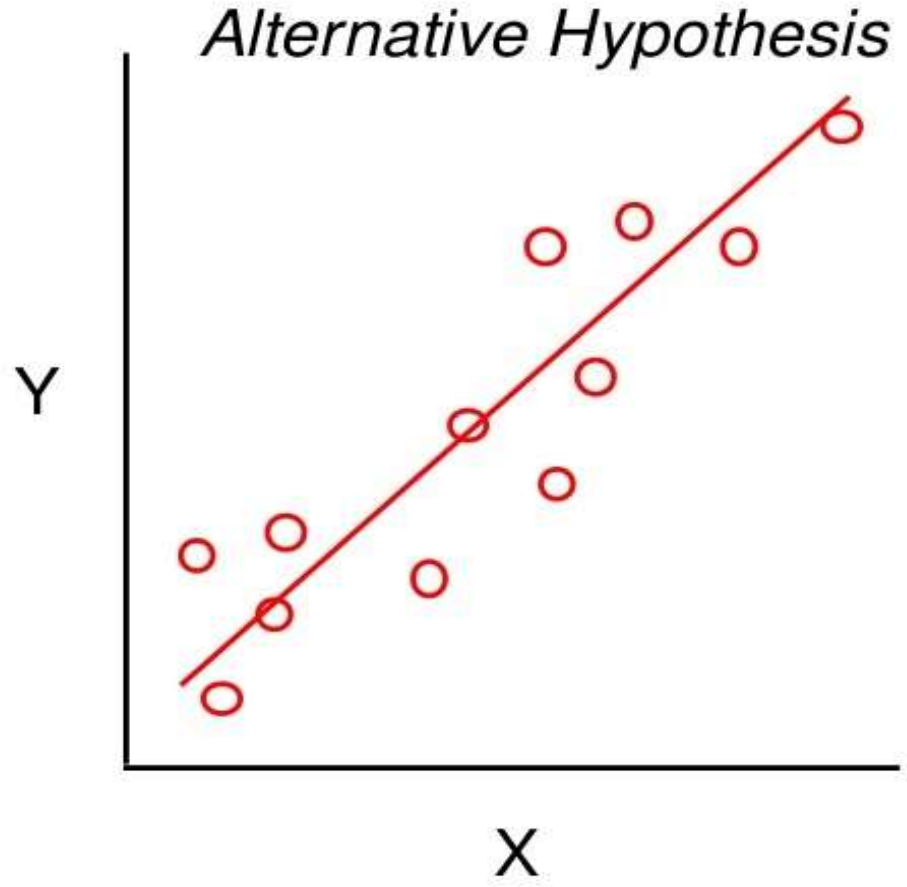
HBI: Regression Analysis (P)

Excel: Regression (P)

Correlation & Regression

Looks for relationships between two continuous variables





Is there a relationship between wing length and tail length in songbirds?

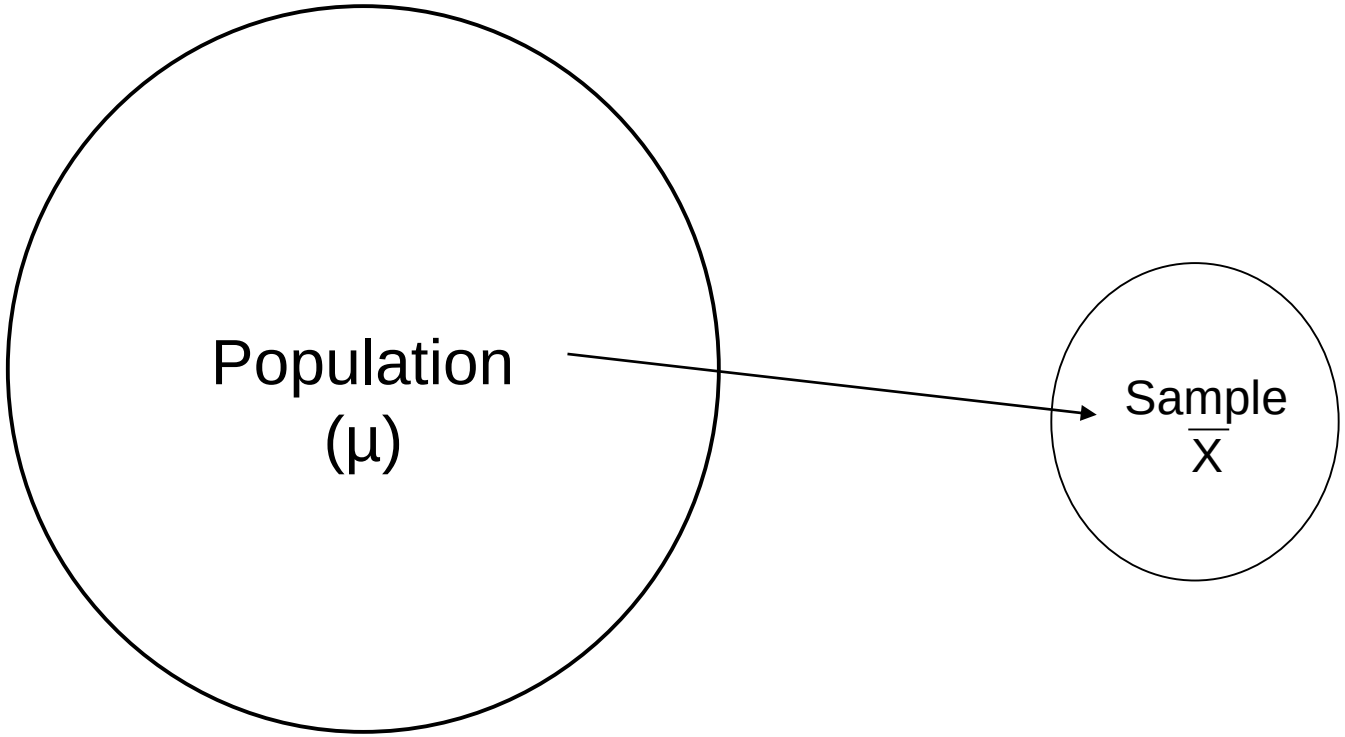
wing length cm	tail length cm
10.4	7.4
10.8	7.6
11.1	7.9
10.2	7.2
10.3	7.4
10.2	7.1
10.7	7.4
10.5	7.2
10.8	7.8
11.2	7.7
10.6	7.8
11.4	8.3

Is there a relationship between age and systolic blood pressure?

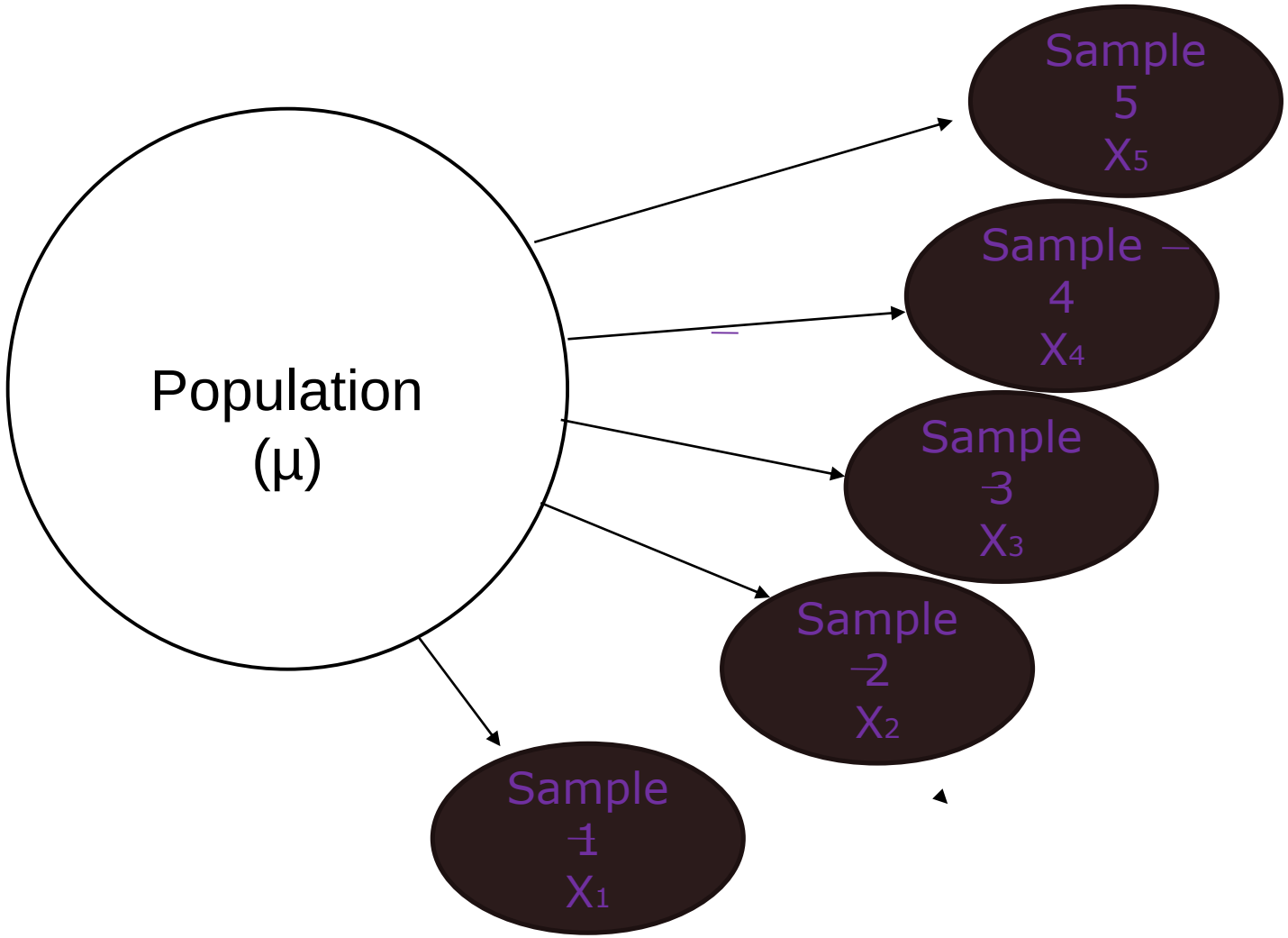
Age (yr)	systolic blood pressure mm hg
30	108
30	110
30	106
40	125
40	120
40	118
40	119
50	132
50	137
50	134
60	148
60	151
60	146
60	147
60	144
70	162
70	156
70	164
70	158
70	159

Some Commonly Used Statistical Tests

Number of samples	Level of Measurement		
	Nominal	Ordinal	Interval/Ratio
1 sample	χ^2 test	Kolmogorov-Smirnoff 1 sample test	1 sample t-test
2 independent samples	χ^2 test	Mann-Whitney U test	2 samples t-test
2 dependent samples	McNemar test	Wilcoxon test	Paired t-test
>2 independent samples	χ^2 test	Kruskal-Wallis test	ANOVA
>2 dependent samples	Cochran Q test	Friedman ranks test	Repeated measures ANOVA



- The sampling distribution of a sample statistic (such as X) is the probability distribution of that statistic.



•The sampling distribution of the mean consists of *all possible sample means* – for all possible samples of size *n* – that you could take from the population

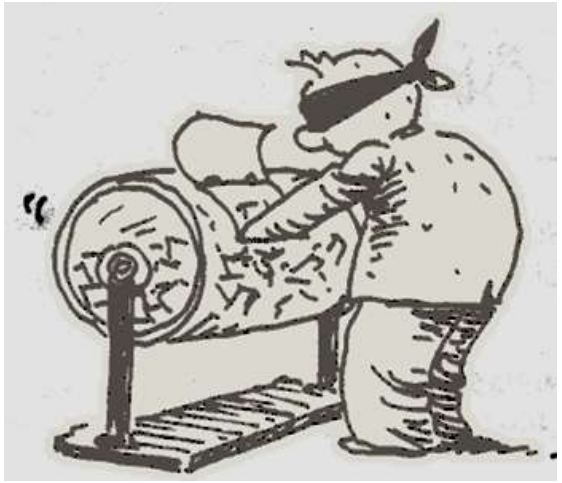
Objective of inferential statistics is to determine characteristics of a population based on a sample

Why sample?

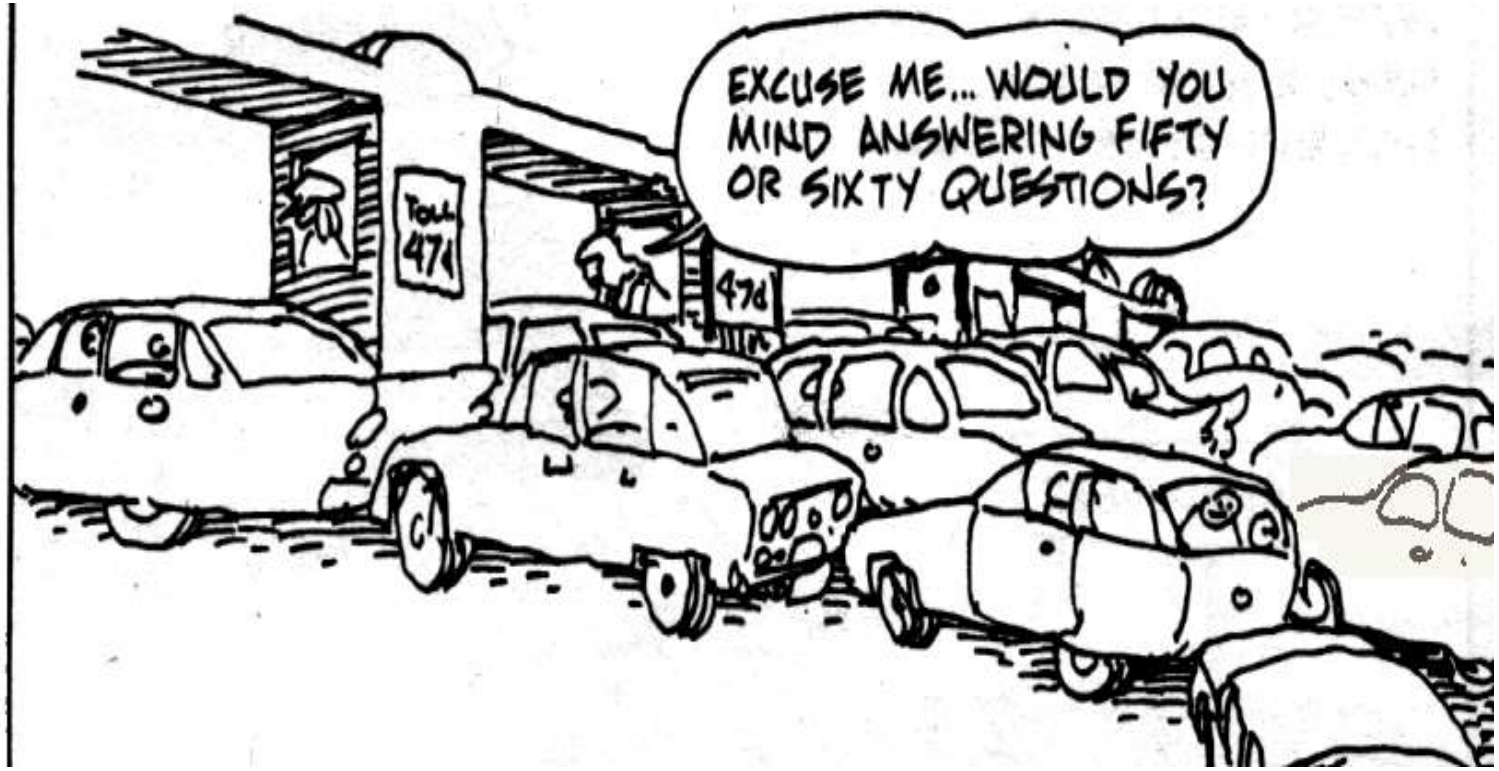
- The physical impossibility of checking all items in the population.
- The cost of studying all the items in a population.
- The time-consuming aspect of contacting the whole population.
- The destructive nature of certain tests.
- The adequacy of sample results in most cases.

Sampling Methods

Simple Random Sample: **A sample selected so that each item or person in the population has the same chance of being included.**



Systematic Random Sampling: Every k^{th} member of the population is selected for the sample.

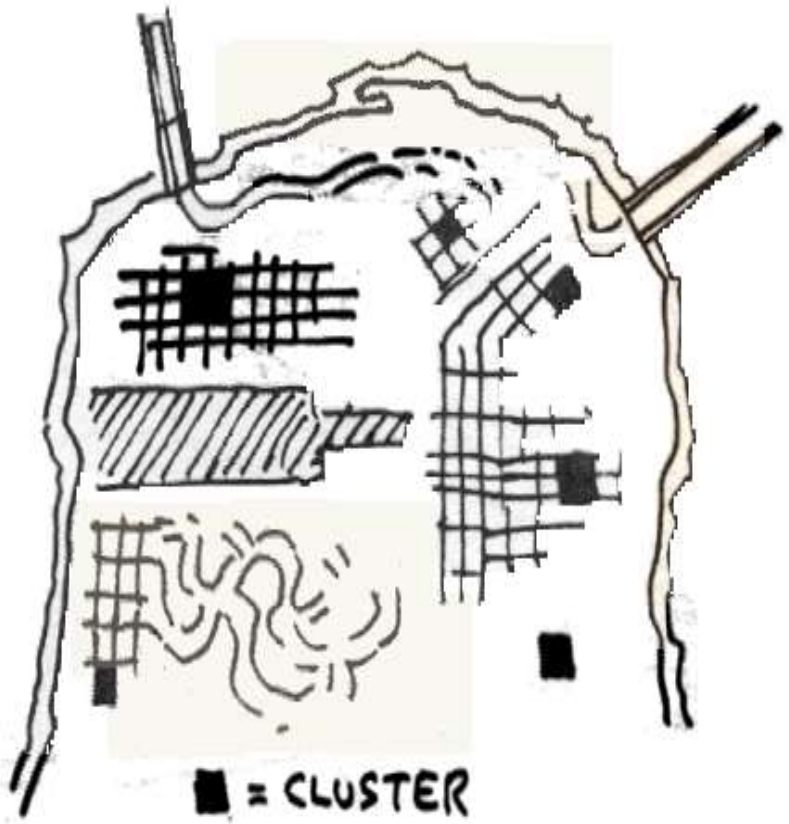


Stratified Random Sampling:

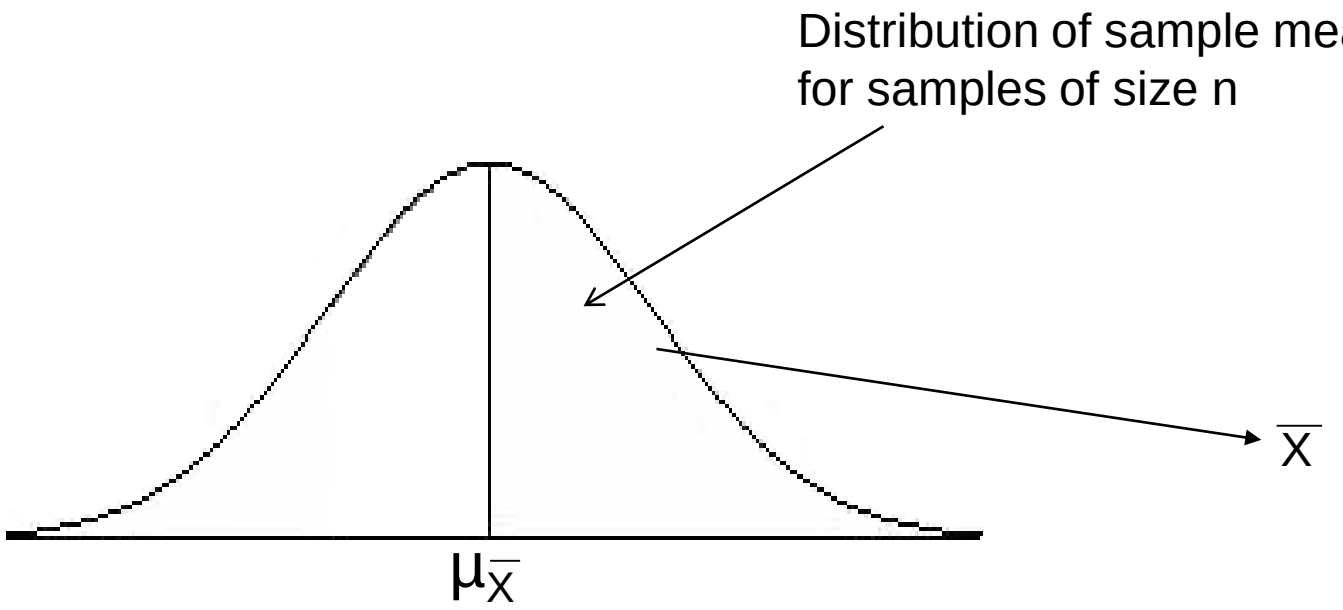
A population is first divided into subgroups, called strata, and a sample is selected from each stratum.

Eg. College students may be stratified into freshmen, sophomore, etc. or simply male and female

Cluster Sampling: A population is first divided into primary units then samples are selected from the primary units.



AN EXAMPLE IS A CITY HOUSING SURVEY WHICH DIVIDES A CITY INTO BLOCKS, RANDOMLY SAMPLES THE BLOCKS, AND LOOKS AT EVERY HOUSING UNIT IN EACH SAMPLED BLOCK.



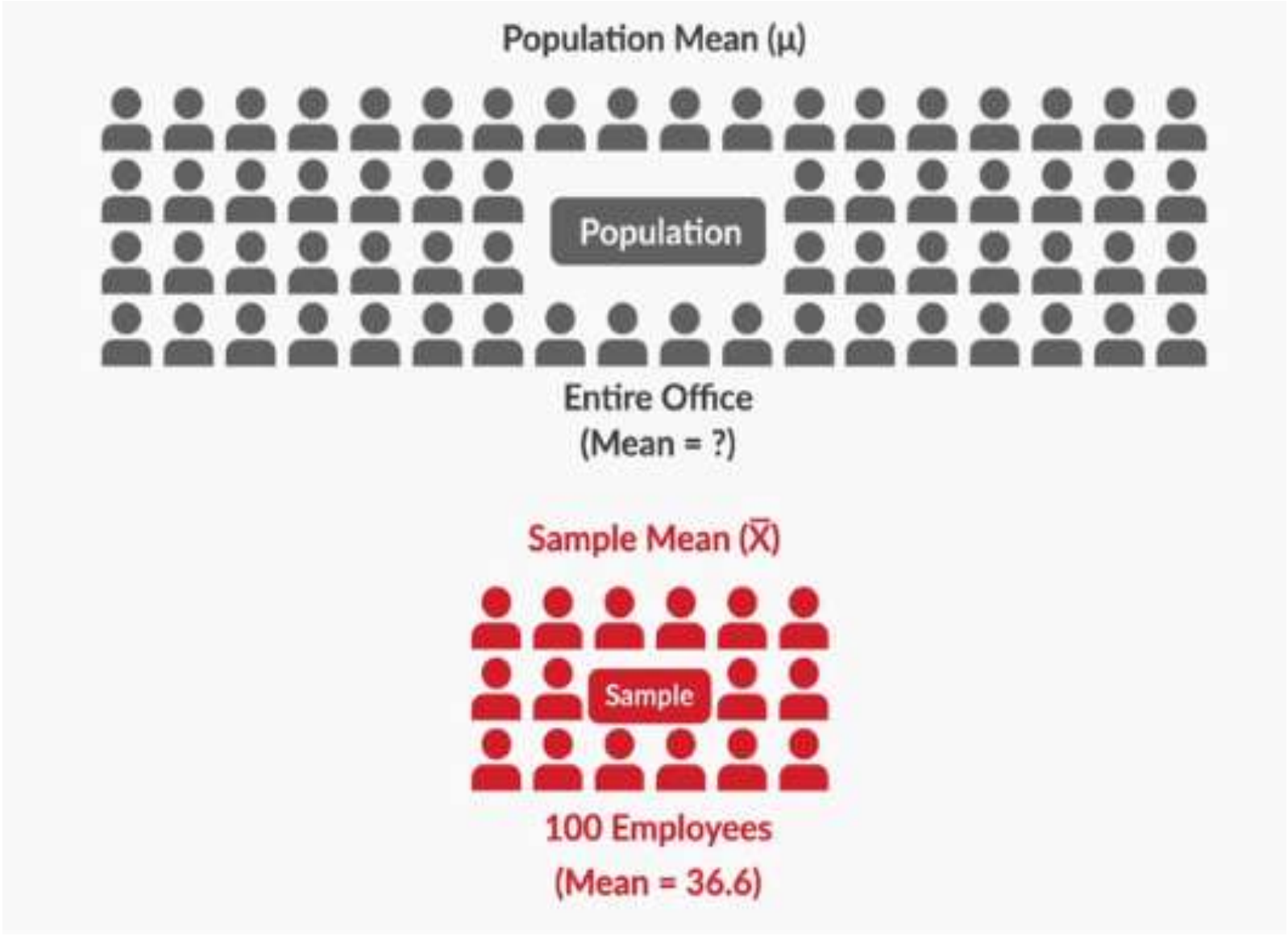
• When we draw a sample from a population, we are at the same time drawing a sample mean from the distribution of sample means for samples of size *n*

Samples

Let’s say that, for a business application, you want to find out the average number of times people in urban India visited malls last year. That’s 400 million (40 crore) people! You can't possibly go and ask every single person how many times they visited the mall. That’s a costly and time-consuming process. How can you reduce the time and money spent on finding this number?

Another example: If there are 30k employees in a company and you want to know the average commute time for each of the individual, what you will do? Will you go to each and every one and ask this information?

Samples



SAMPLING TERMINOLOGY

1. Sample Mean ($\bar{X}$) = 36.6
2. Population Mean (μ) = ?
3. Sample Size (n) = 100
4. Population Size (N) = 30,000

INTERNSHIPSTUDIO

Sample

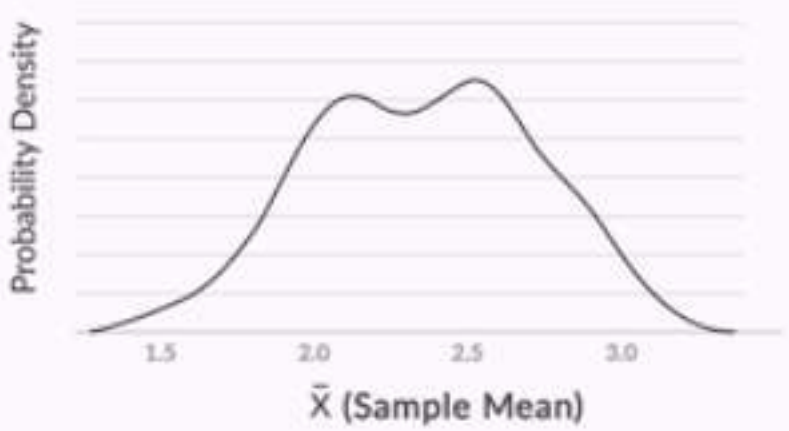
- **Samples**

Population/Sample	Term	Notation	Formula
Population $(X_1, X_2, X_3, \ldots\ldots, X_N)$	Population Size	N	Number of items/elements in the population
	Population Mean	μ	$\frac{\sum_{i=1}^{i=N} X_i}{N}$
	Population Variance	σ^2	$\frac{\sum_{i=1}^{i=N}(X_i - \mu)^2}{N}$
Sample $(X_1, X_2, X_3, \ldots\ldots, X_n)$ (Sample of Population)	Sample Size	n	Number of items/elements in the sample
	Sample Mean	$\bar{X}$	$\frac{\sum_{i=1}^{i=n} X_i}{n}$
	Sample Variance	S^2	$\frac{\sum_{i=1}^{i=n}(X_i - \bar{X})^2}{n-1}$

Sampling Distributions

Serial No.	Name	X (No. of red balls)
1	Manish	3
2	Rohit	2
3	Pearl	4
4	Prakhar	3
.	.	.
.	.	.
39	Rajiv	2
40	Ajay	3
41	Romil	3
42	Nikhil	2
43	Himanshu	2
44	Parth	3
45	Raman	2
.	.	.
.	.	.
70	Sachin	1
71	Salma	2
72	Sakshi	3
73	Sandeep	3
74	Kushal	4
75	Suri	1

~~$\mu = 2.385$~~



$\bar{x}_1 = 2.8$
$\bar{x}_2 = 2.0$
$\bar{x}_3 = 2.6$
$\bar{x}_4 = 3.2$
.
.
$\bar{x}_{39} = 2.0$
$\bar{x}_{40} = 2.8$
$\bar{x}_{41} = 2.2$
$\bar{x}_{42} = 2.6$
$\bar{x}_{43} = 2.0$
$\bar{x}_{44} = 2.6$
$\bar{x}_{45} = 3.0$
.
.
$\bar{x}_{95} = 2.8$
$\bar{x}_{96} = 2.6$
$\bar{x}_{97} = 1.8$
$\bar{x}_{98} = 2.4$
$\bar{x}_{99} = 2.4$
$\bar{x}_{100} = 2.6$

INTERNSHIPSTUDIO

Sampling distributions

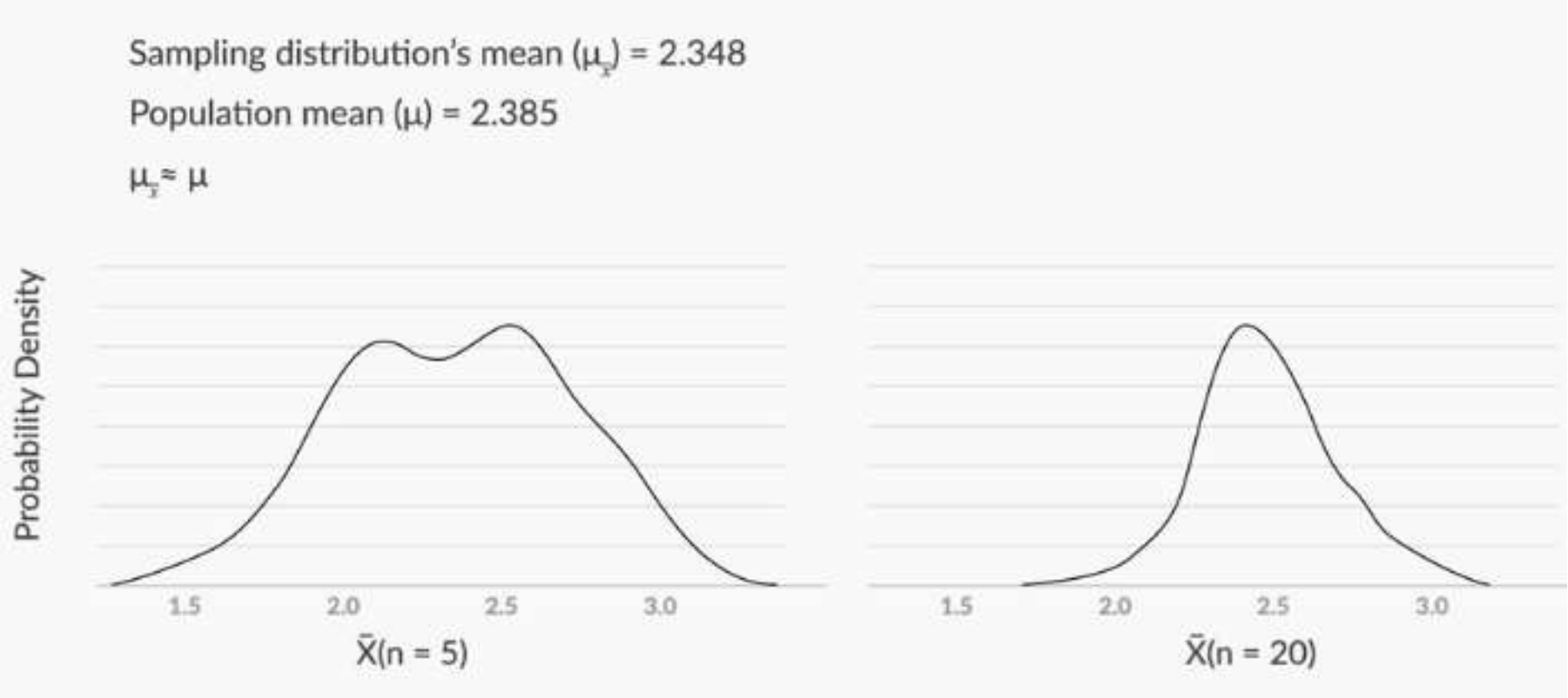
- The sampling distribution of a sample statistic is the probability distribution of that statistic.

- We can have sampling distributions of any sample statistic

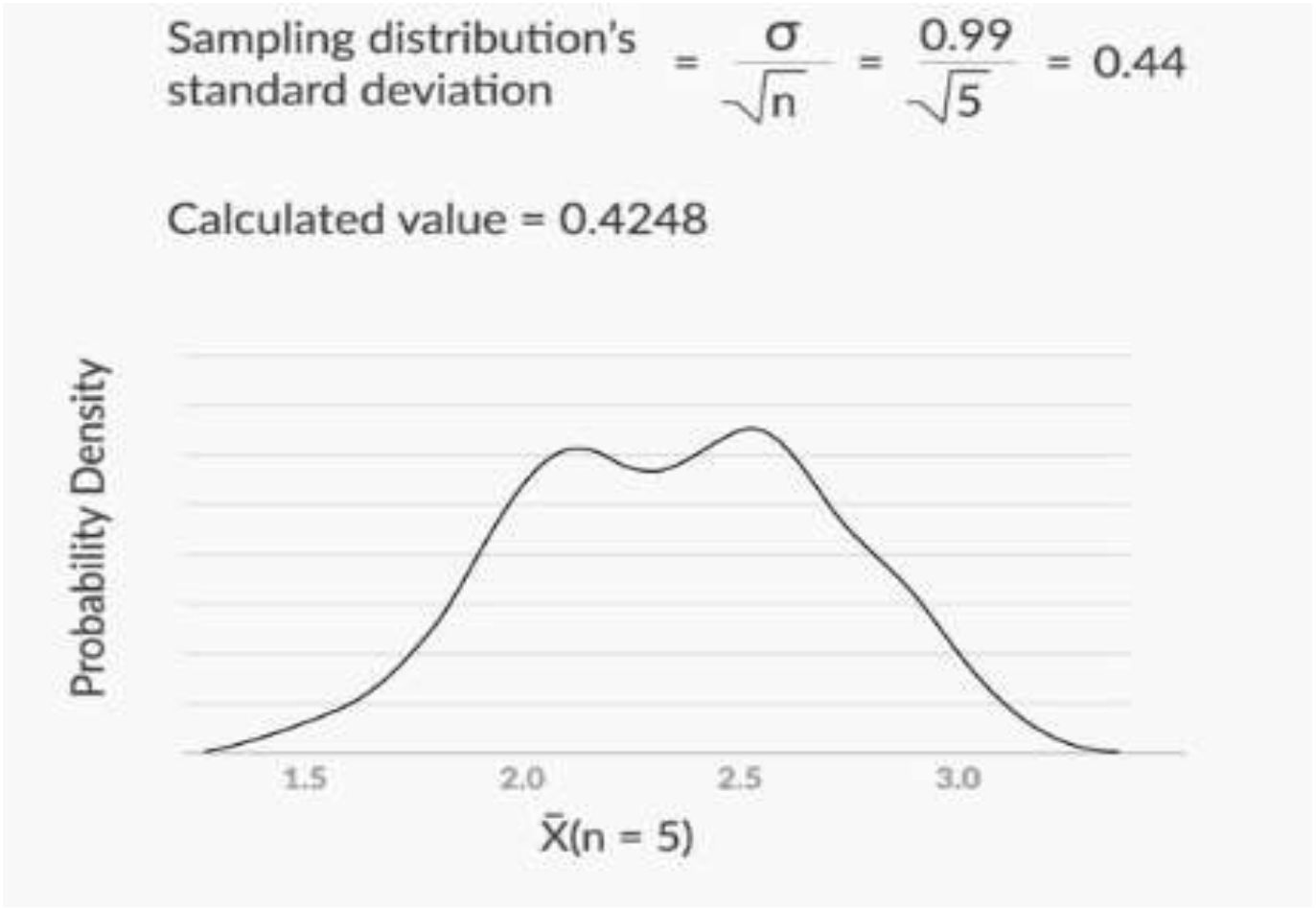
- Mean $\overline{X}$
- Median M
- Variance s^2
- Std devn s

Properties of Sampling Distributions

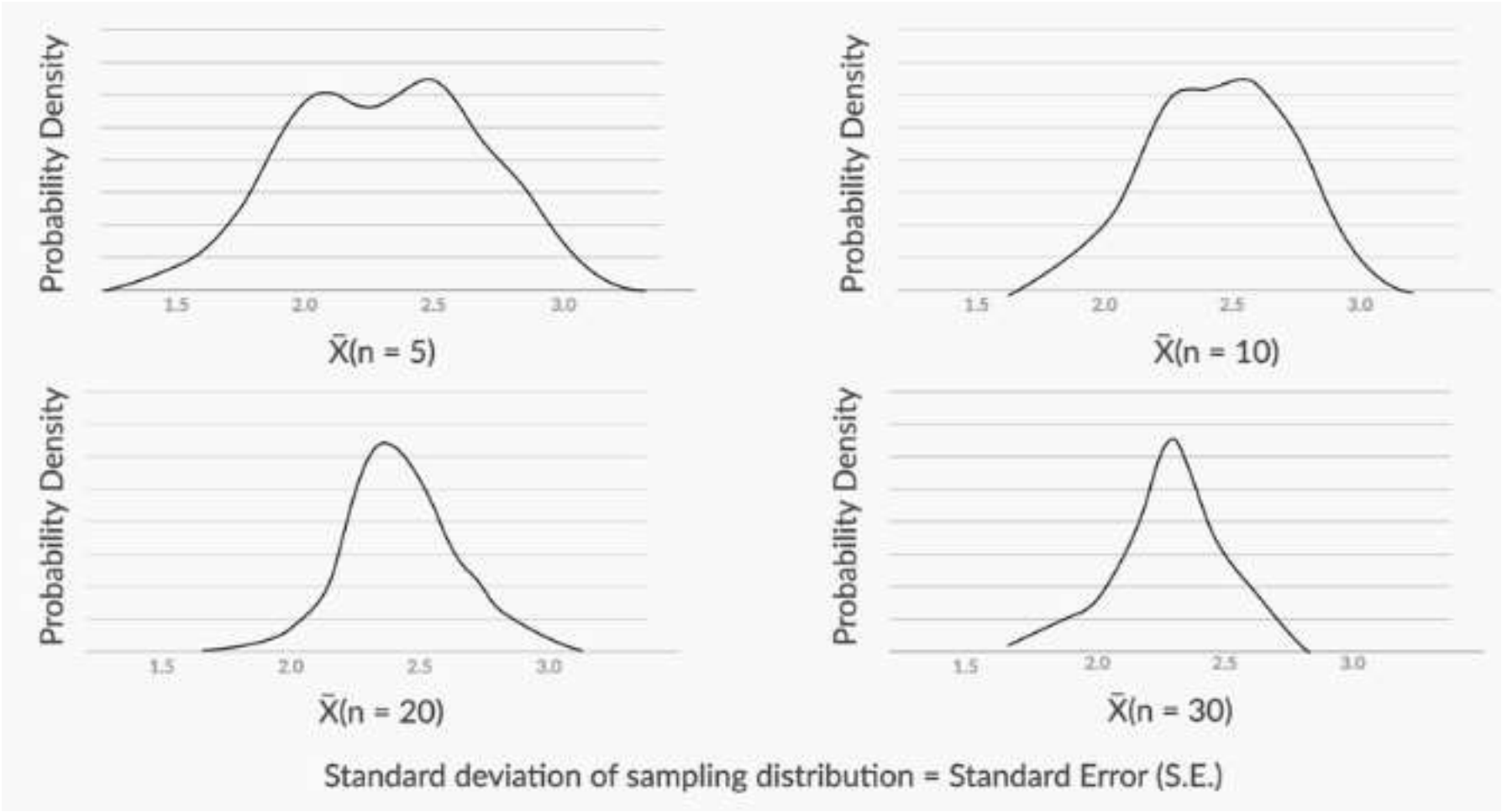
1. The mean of sampling distribution is very close to the actual mean.
2. As the number of samples increases, the mean of the sampling distribution becomes more and more closer to the population mean.



Properties of Sampling Distributions



Properties of Sampling Distributions



The sampling distribution of the mean

- The sampling distribution of the sample mean $\bar{X}$.

$$E(\bar{X}) = \mu_{\bar{X}} = \mu$$

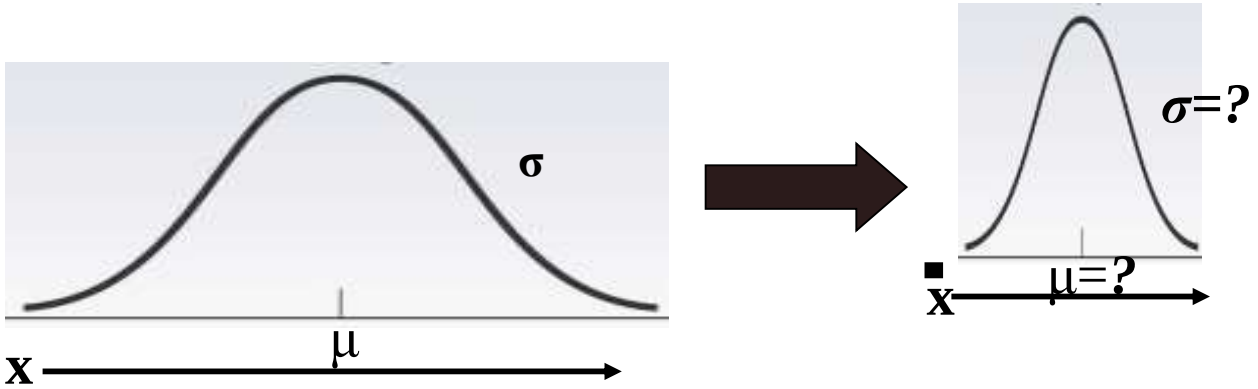
- Variability of this distribution is given by the *standard error of the mean*:

$$\sigma_{\bar{X}} = \sigma_{\bar{X}} \cong s$$

Question ?

If you repeatedly take samples from a population and calculate the sample mean for each sample,

what would the distribution of the sample means look like?



Generalizing the result

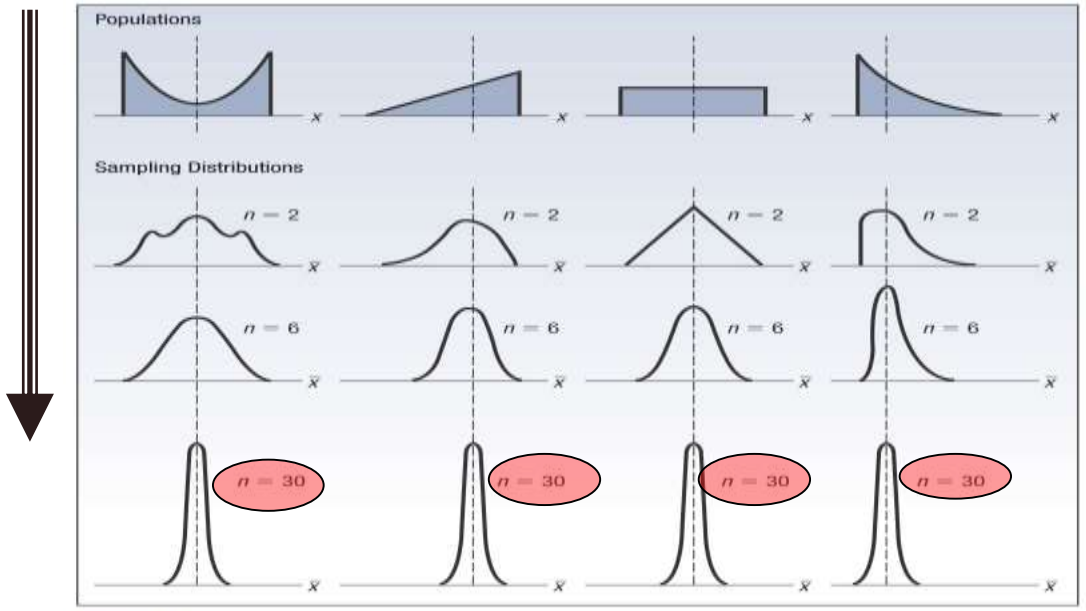


CHART 8–2 Results of the Central Limit Theorem for Several Populations

Irrespective of the shape of distribution of data in the original population, as you increase the sample size *(minimum recommended is n=30),* the distribution of the sample mean will become a normal distribution.

Note: *If the population distribution is known to be normal, then sample means is guaranteed to be normally distributed (even if n<30).*

The Central Limit Theorem

- Consider a random sample of n observations from a population with mean μ and standard deviation σ.

- When n is sufficiently large, the *sampling distribution of* $\bar{x}$ will be **approximately normal** with mean $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

- Note: this is true regardless of the shape of the underlying distribution of raw scores

The Central Limit Theorem

- The larger the sample size, the better the approximation to the normal distribution.
- For most populations, $n \geq 30$ will be "sufficiently large."

The Central Limit Theorem

- When we draw a sample and measure its mean, by the CLT, **we may assume the sampling distribution of the sample mean is normal**.

- That means we can use the *standard normal distribution* (SND) to work out the probability of finding a sample mean in a given range relative to the population mean.

Central Limit Theorem

If a sufficiently large random sample is drawn from a population with mean u and variance $s^{2,}$ the distribution of samples means will have the following characteristics:

- An approximately normal distribution regardless of the distribution of the underlying population
- The mean of the sample means will be equal to the population mean
- The variance of the sample means will be equal to the population variance divided by the sample size.

Central Limit Theorem

CLT is a statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.

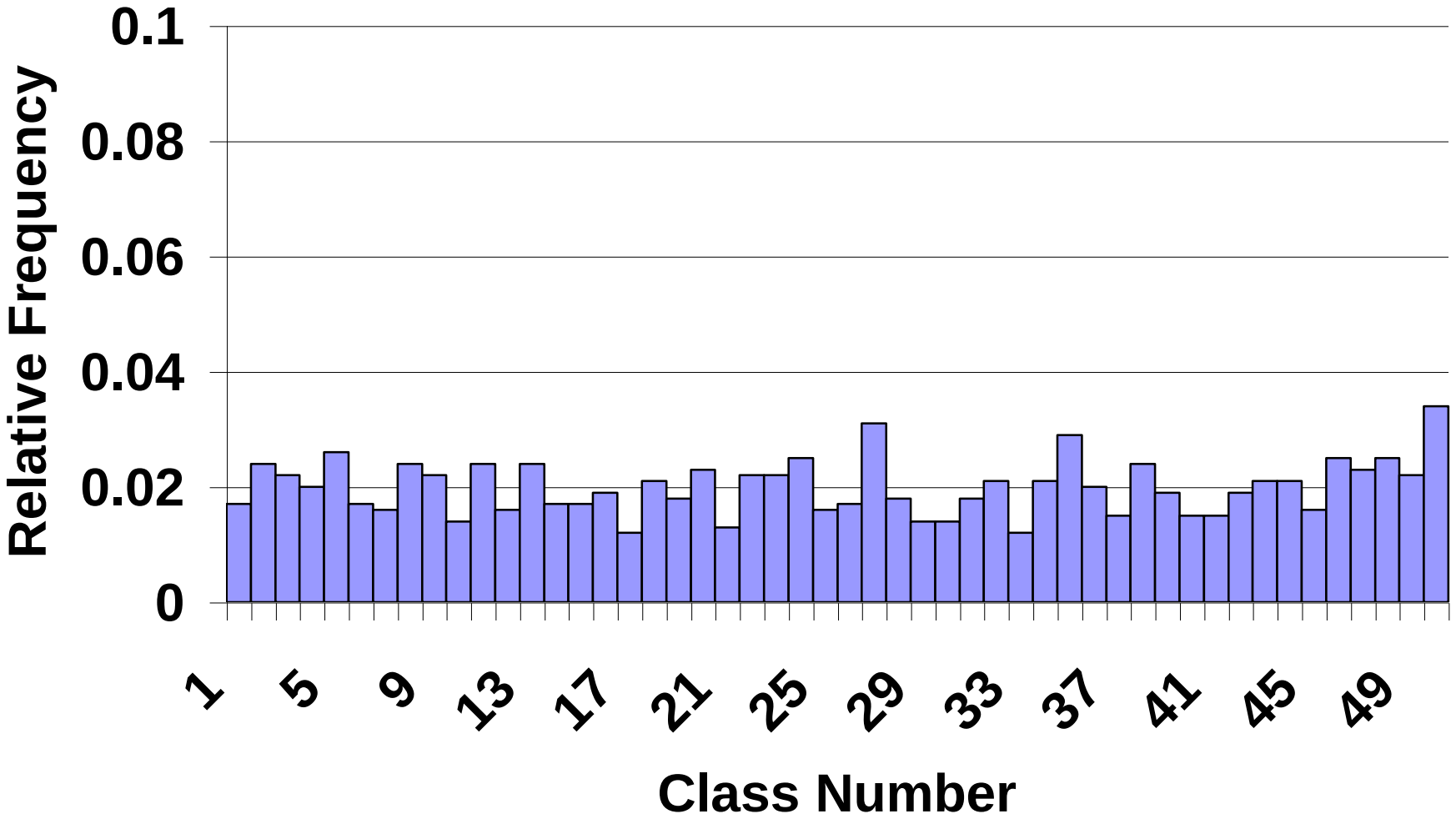
1. Sampling distribution's mean ($\mu_{\bar{x}}$) = Population mean (μ)
2. Sampling distribution's standard deviation (Standard Error) = $\frac{\sigma}{\sqrt{n}}$
3. For n > 30, the sampling distribution becomes normally distributed

This is called the "Central Limit Theorem"

http://onlinestatbook.com/stat_sim/sampling_dist/index.html

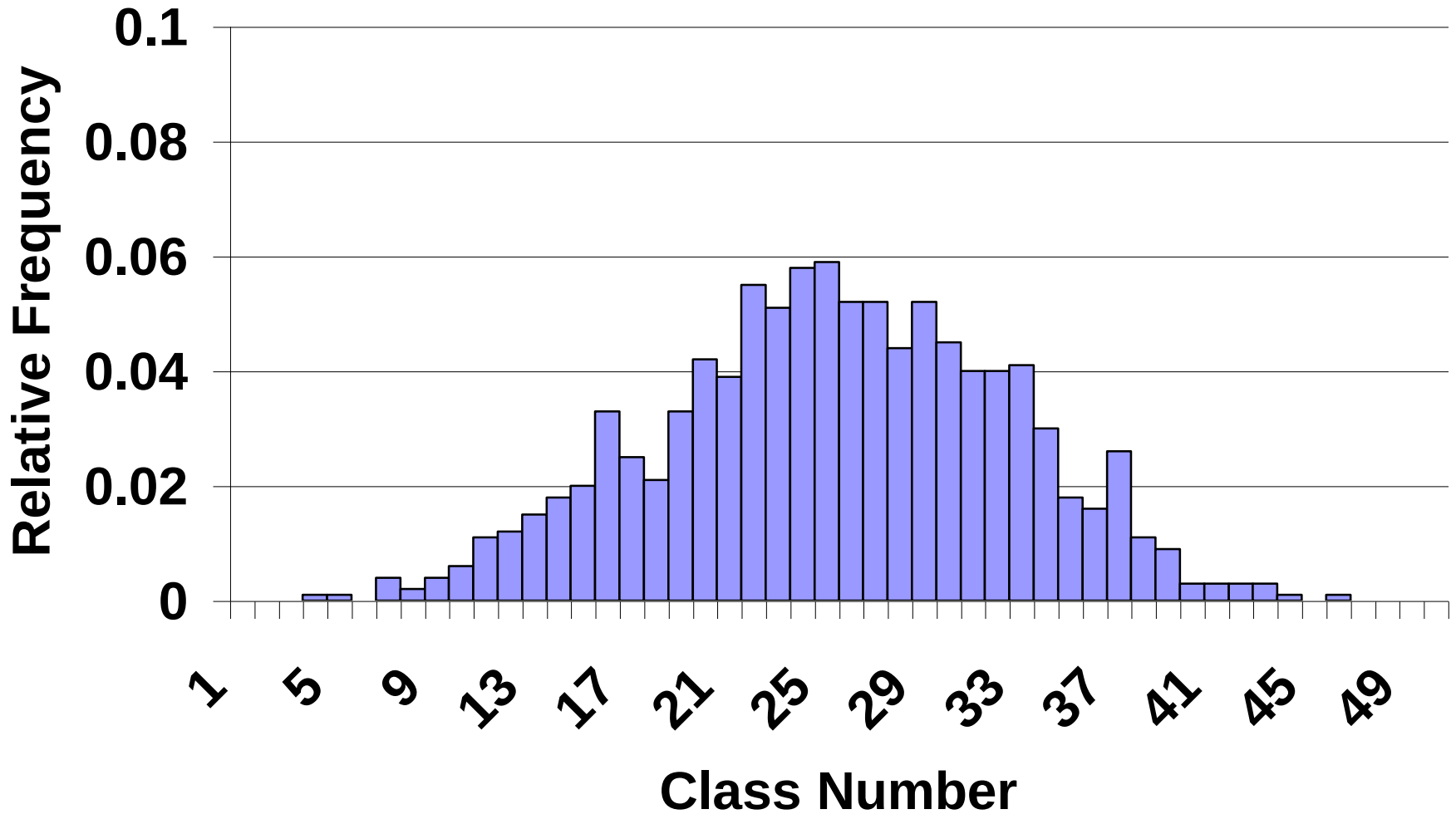
INTERNSHIPSTUDIO

Sample Size and Mean

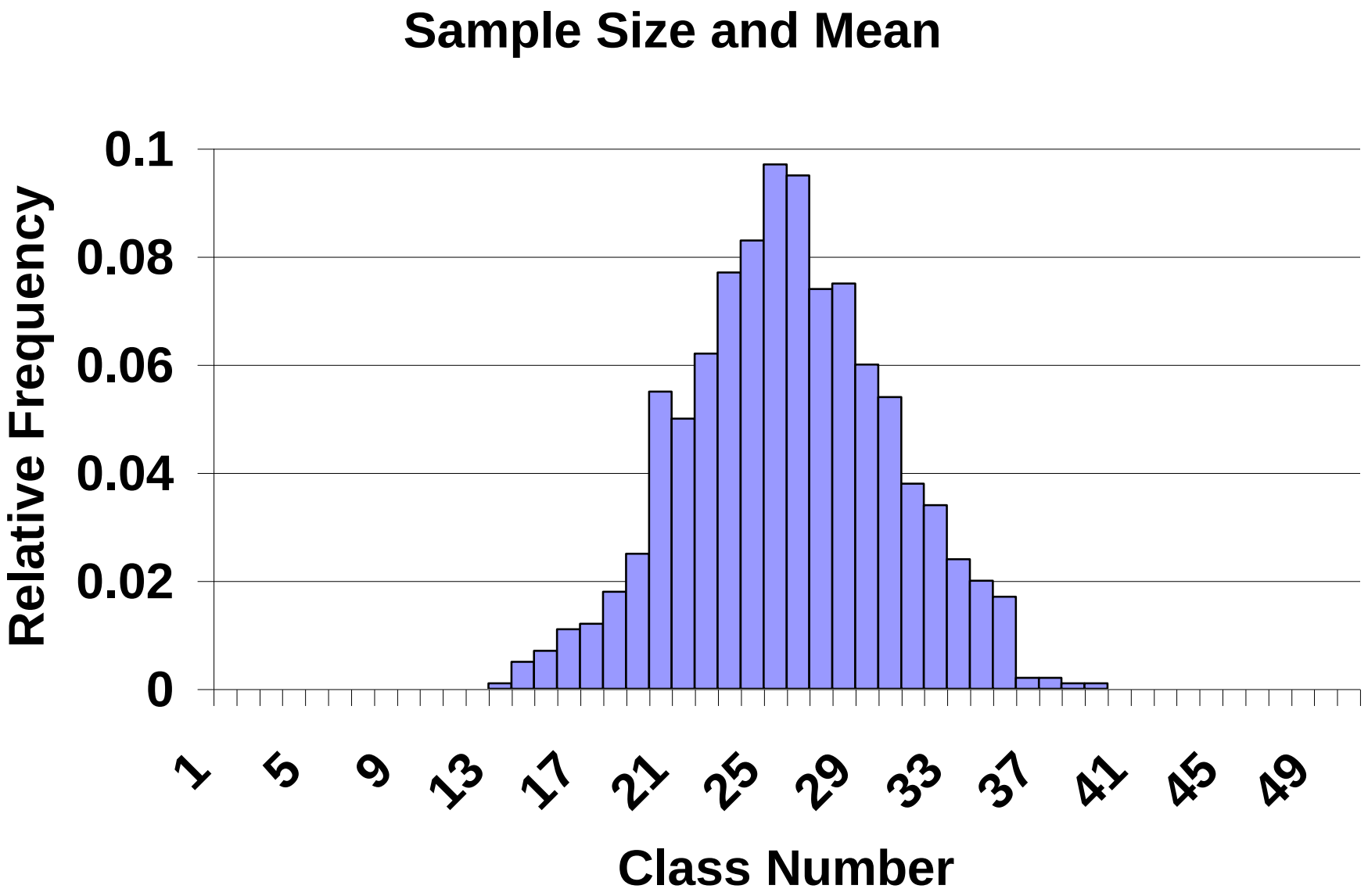


Distribution of random numbers

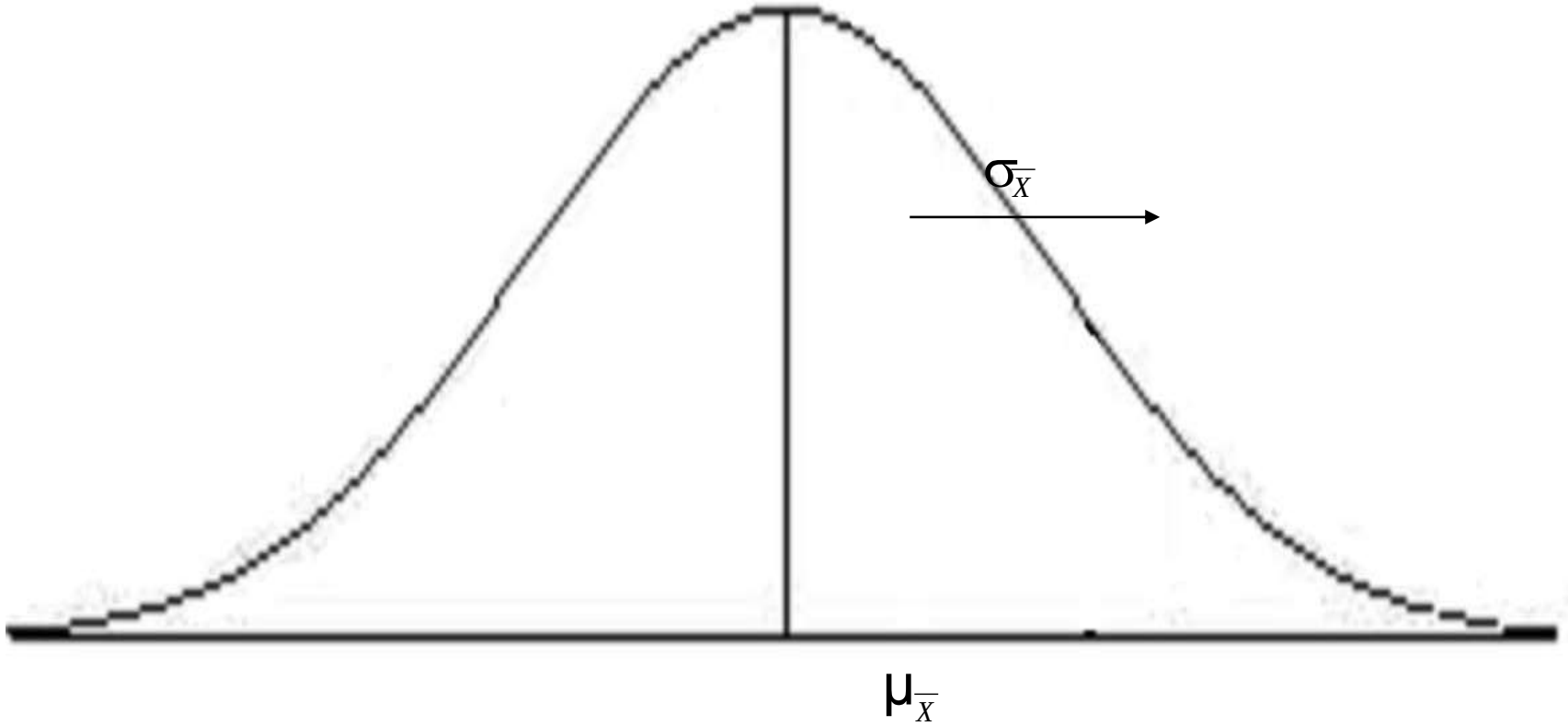
Sample Size and Mean



Distribution of means of *n* random numbers, *n*=4



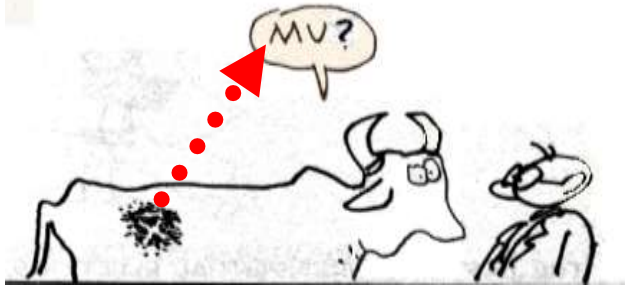
Distribution of means of *n* random numbers, *n*=10



The sampling distribution of the sample mean

As n increases $\bar{\mu}_x$ will approach μ.

So sample mean is a good estimator of population mean.



This s.d. is called the standard error

(ie., of the mean distribution).

$$\bar{\sigma}_x = \frac{\sigma}{\sqrt{n}}$$

Note that the Std Error is smaller

The sampling distribution of the mean

- We use the sampling distribution of the mean the way we used the SND. We obtain probabilities of finding sample means in a given range relative to the population mean, for samples of size n.

- Don’t forget to use the standard error, $\sigma_{\bar{X}}$, rather than the standard deviation, σ!

Variance of the sample mean distribution

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{x1 + x2 + \ldots + xn}{n}\right)$$

Where x1 is mean of sample 1,x2 is …)

$$= \frac{1}{n^2}[\mathrm{Var}(x1) + \mathrm{Var}(x2) + \ldots + \mathrm{Var}(xn)]$$

$$= \frac{1}{n^2}[\sigma^2 + \sigma^2 + \ldots + \sigma^2] = \frac{1}{n^2}[n.\,\sigma^2] = \frac{n\,\sigma^2}{n^2}$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

therefore, **Standard Deviation = $\sigma/\sqrt{n}$**

(Remember this formula!)

Confidence Intervals

- There are two ways to estimate population parameters such as the mean:

1. Point estimates, such as X
2. Interval estimates, which tell us a range of values that will contain the parameter with known probability.

The Z Score Formula: One Sample

The basic z score formula for a sample is:

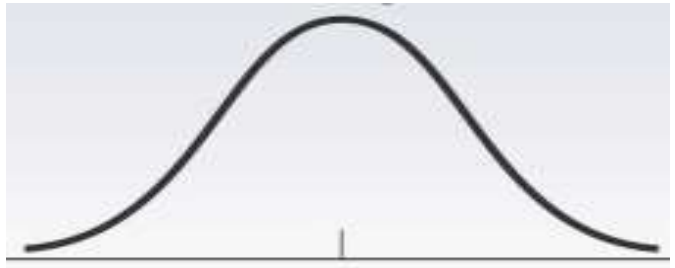
$$z = (x – \mu) / \sigma$$

For example, let’s say you have a test score of 190. The test has a mean (μ) of 150 and a standard deviation (σ) of 25. Assuming a normal distribution, your z score would be:

$$z = (x – \mu) / \sigma$$
$$= (190 – 150) / 25 = 1.6.$$

The z score tells you how many standard deviations from the mean your score is. In this example, your score is 1.6 standard deviations above the mean.

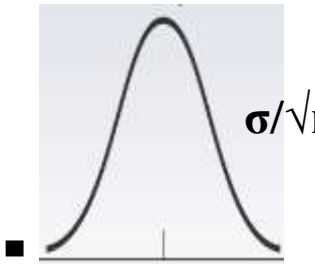
Distribution of population



σ

x̄ μ

Std.Error



σ/√n

x̄ μ

Distribution of sample

The Z score formula for the distribution of sample means is:

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

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

Use s in place of σ if the population standard deviation is unknown, so long as $n \geq 30$.

Z score formula is:

$$z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

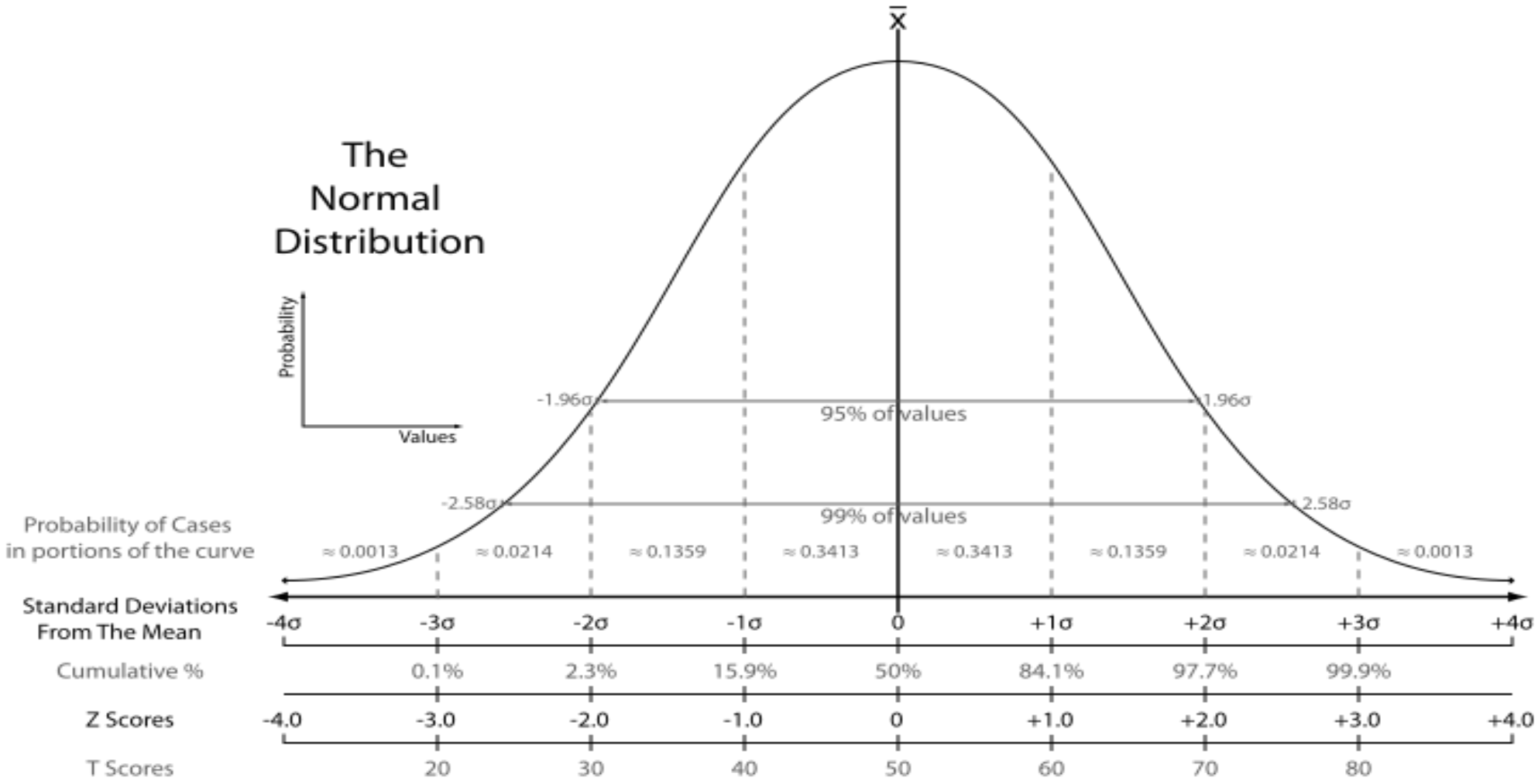
How to understand the formula?

With comparing each one’s score with the mean: x - μ, we will get a kind of deviation. But at this point we still don’t know whether each one’s deviation is big or small. We need a "standard" to compare each deviation.

So we want to compare each deviation with the Standard deviation: deviation ÷ σ

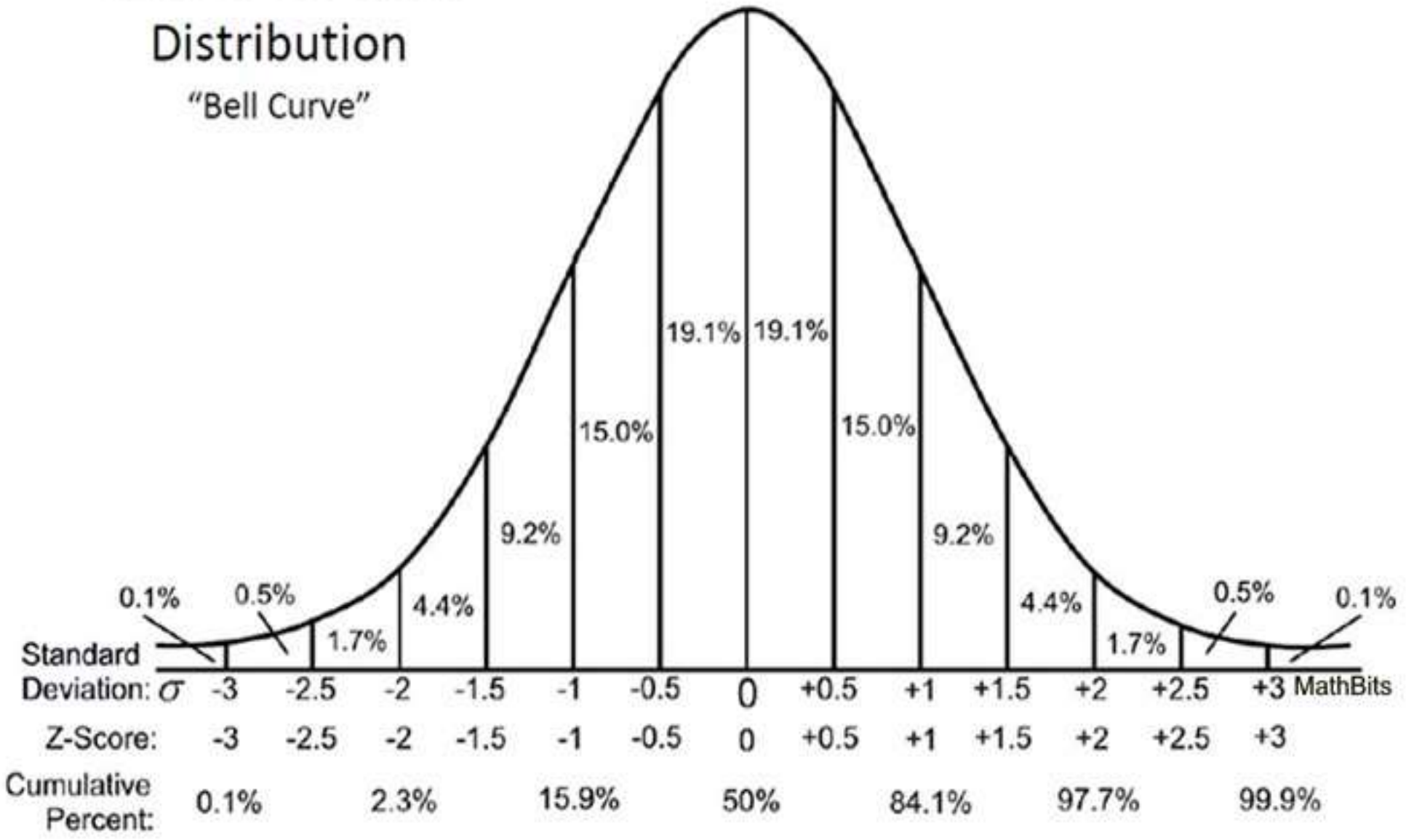
Standard Score = (x - μ) / σ

$$Z = \frac{data\ point - mean\ value}{standard\ deviation}$$


The Normal Distribution
Probability
Values
$\bar{x}$
-1.96σ
1.96σ
95% of values
-2.58σ
2.58σ
99% of values
Probability of Cases in portions of the curve
≈ 0.0013
≈ 0.0214
≈ 0.1359
≈ 0.3413
≈ 0.3413
≈ 0.1359
≈ 0.0214
≈ 0.0013
Standard Deviations From The Mean
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative %
0.1%
2.3%
15.9%
50%
84.1%
97.7%
99.9%
Z Scores
-4.0
-3.0
-2.0
-1.0
0
+1.0
+2.0
+3.0
+4.0
T Scores
20
30
40
50
60
70
80

Standard Normal Distribution

"Bell Curve"

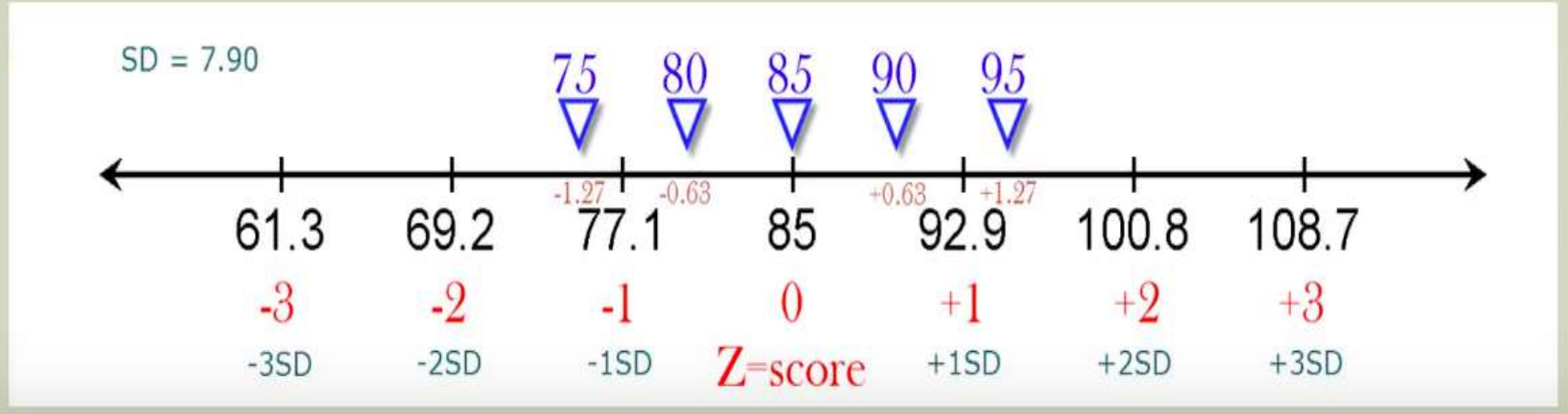


There's some exam data of a class:

Z-SCORE VISUALIZATION

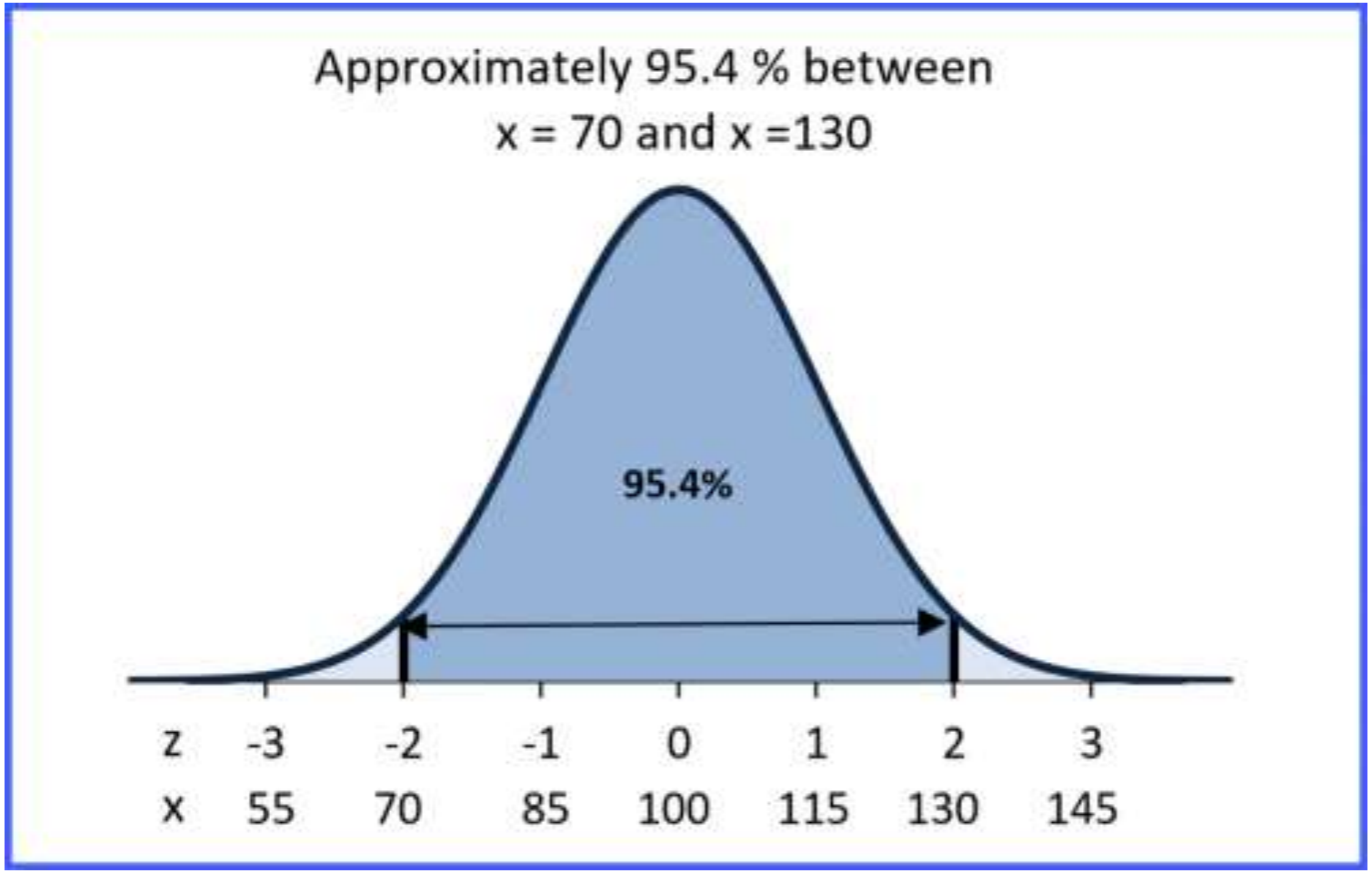
61.3	69.2	77.1	85	92.9	100.8	108.7
$\dot{x} - 3\sigma$	$\dot{x} - 2\sigma$	$\dot{x} - 1\sigma$	$\dot{x}$	$\dot{x} + 1\sigma$	$\dot{x} + 2\sigma$	$\dot{x} + 3\sigma$
-3	-2	-1	$z = 0$	$+1$	$+2$	$+3$

$$\sigma = 7.90$$



IQ Scores have a bell-shaped distribution with a mean of 100 and a standard deviation of 15. What percentage of IQ scores are between 70 and 130?

Solution: 130 minus the mean of 100 = 30, which is 2 times 15, the standard deviation. Thus, 130 is 2 standard deviations to the right of the mean. 100 – 70 = 30 which is 2 times 15. Thus, 70 is 2 standard deviations to the left of the mean. Since 70 to 130 is within 2 standard deviations of the mean, we know that about 95.4% of the IQ scores would be between 70 and 130.


Approximately 95.4 % between
x = 70 and x =130
95.4%
z -3 -2 -1 0 1 2 3
x 55 70 85 100 115 130 145

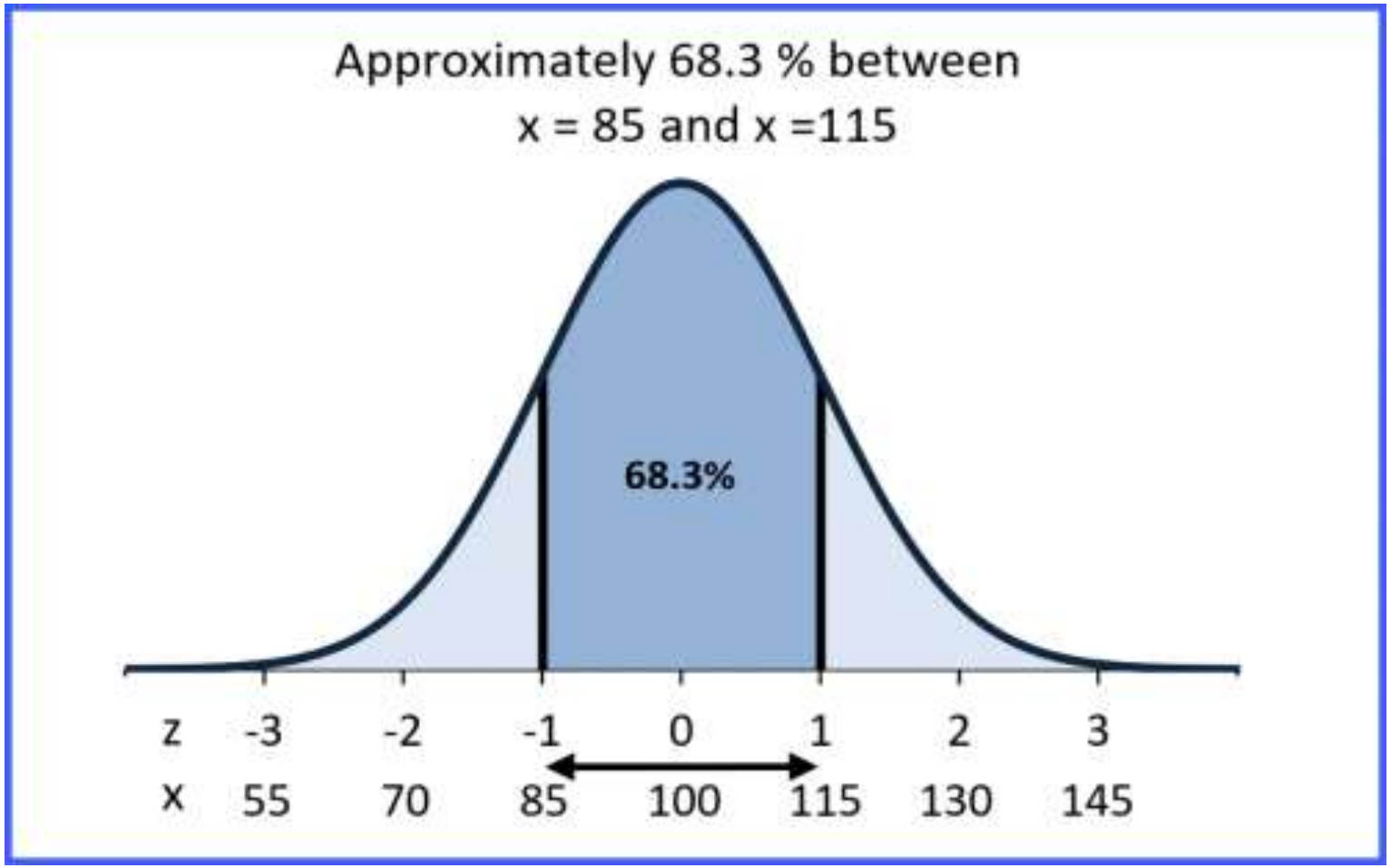
What is the probability a randomly selected individual will have an IQ between 85 and 115?

Solution: Because the area under the bell curve represents 100% of the possible data points, it also represents probability. Recall that the probability of an event is the number of ways the event can happen divided by the total number of outcomes. Thus, an area under the bell curve equal to 10% is also a 10% probability of “x” falling in that area.

First, find the z-scores of the two IQs of interest:

So, we need the area under the curve between $z = +1$ and $z = -1$

Thus, there is about 68.3% probability a randomly selected individual will have an IQ between 85 and 115.



Estimating Mean Using CLT

Now suppose we want to estimate the population mean and we have taken a sample of 30 people and the sample mean we obtained as 36.6.

Sample mean ($\bar{X}$)

Sample standard deviation (S)

Sample size (n)

Confidence interval (y% confidence level) = $\left(\bar{X} - \frac{Z^*S}{\sqrt{n}} , \bar{X} + \frac{Z^*S}{\sqrt{n}} \right)$

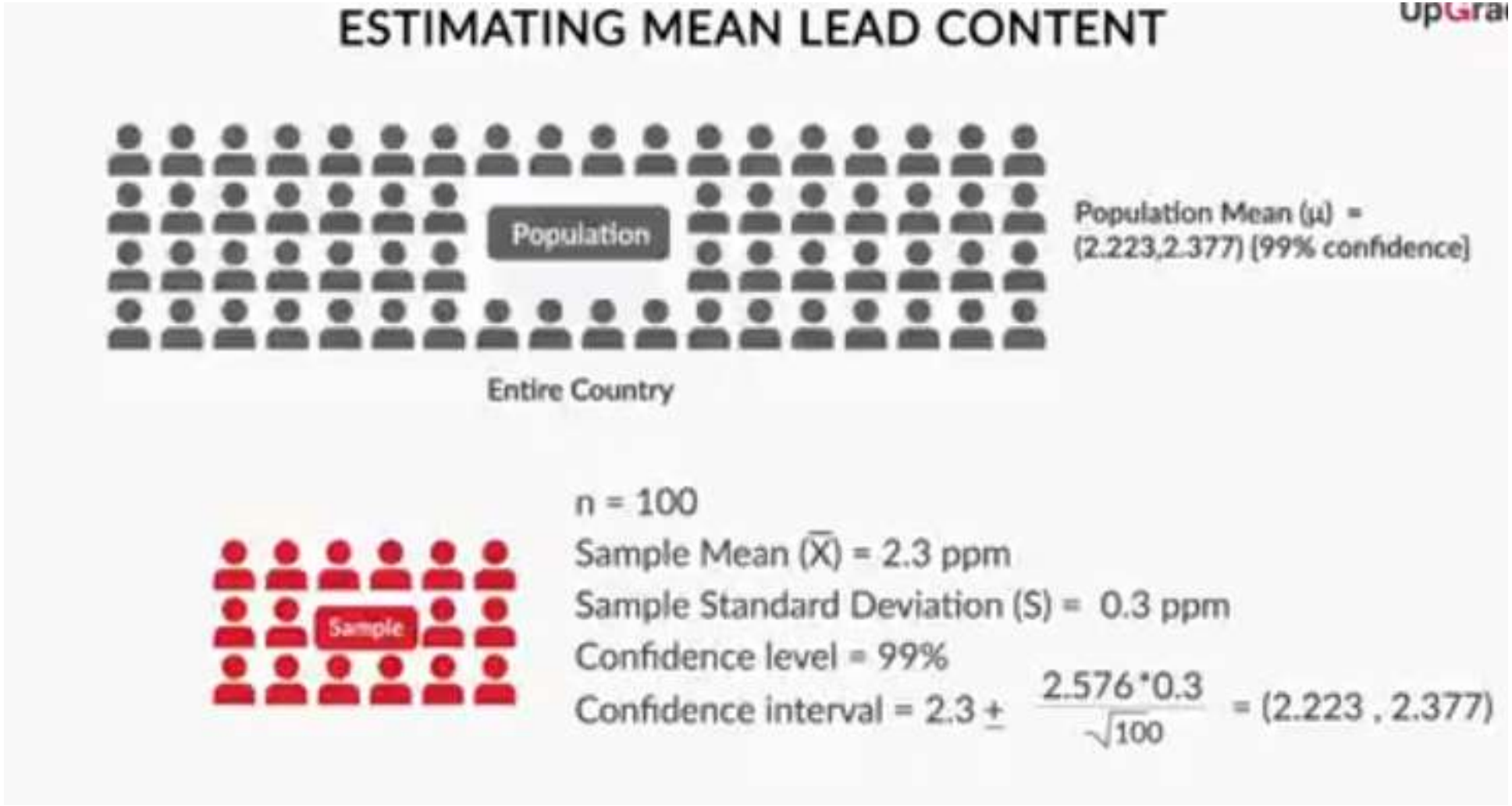
where Z^* is the Z score associated with y% confidence level

Estimating Mean Using CLT

Confidence Level	Z*
90%	±1.65
95%	±1.96
99%	±2.58

Confidence Interval - Example

Let suppose that we want to test maggi samples for lead component. The allowed lead is 2.5PPM.



Agenda

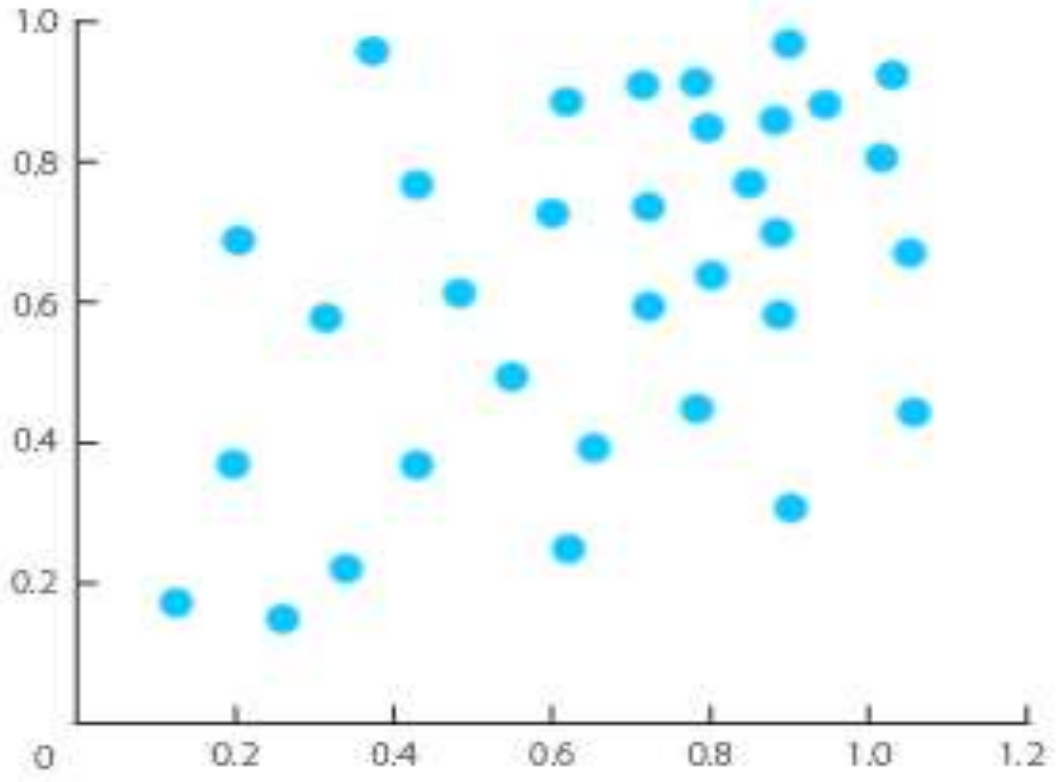
1. Introduction to Data Visualization
2. Scatter plot
3. Line plot
4. Bar chart
5. Pie chart
6. Histogram plot
7. Box plot

Introduction

- Let's begin this session by understanding how data visualisation can add value to the information you want to convey.
- Graphics and visuals, if used intelligently and innovatively, can convey a lot more than what raw data alone can convey.
- You will learn about the pre-attentive and attentive attributes, which are the differentiating factors in making your visuals interesting for the reader.

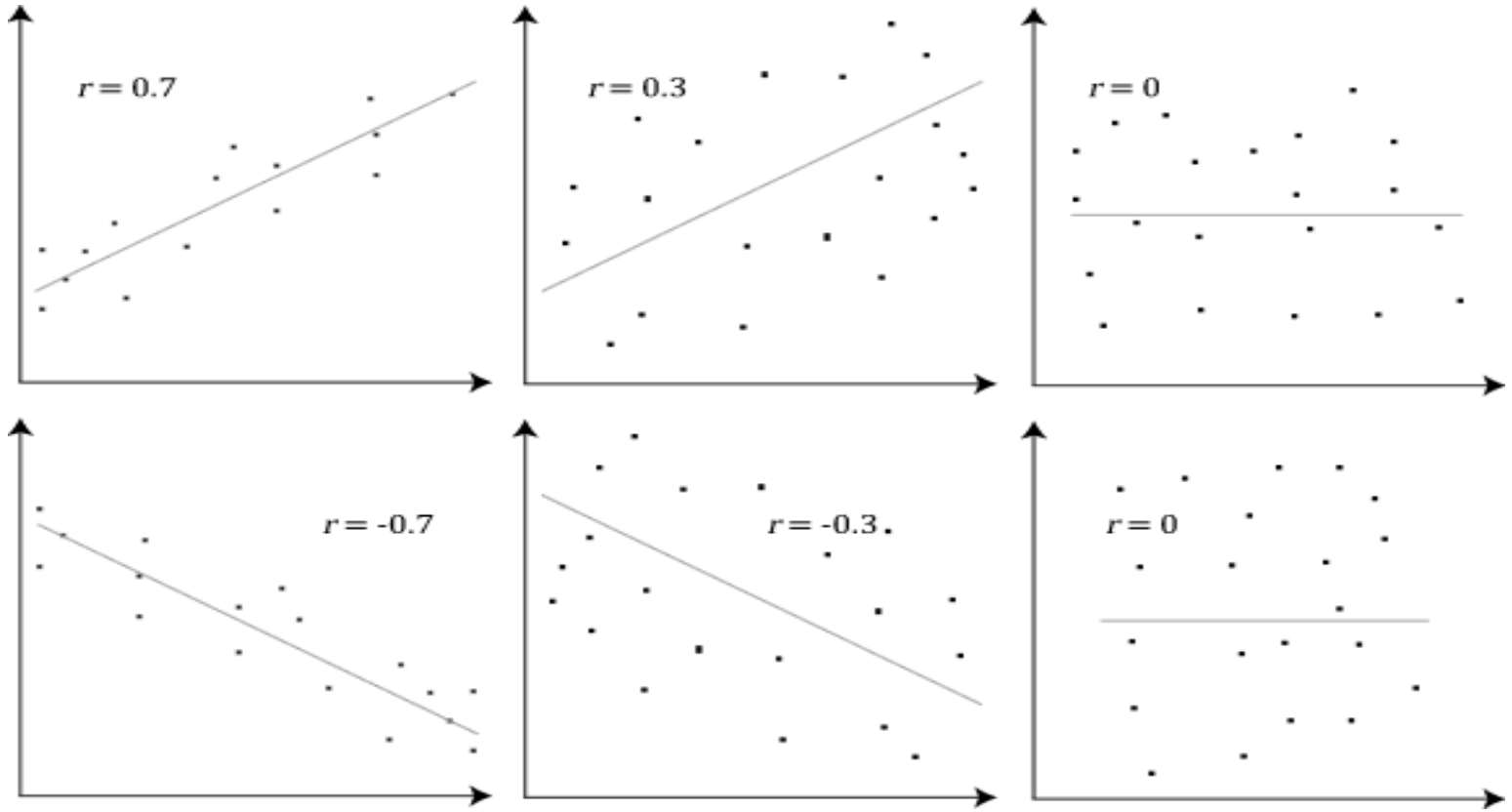
Scatter Plot

- Scatter plot explains the relationship between the two different variables. Most important it shows the correlation between two variables. Which is very important for understanding the dependency between the variables.



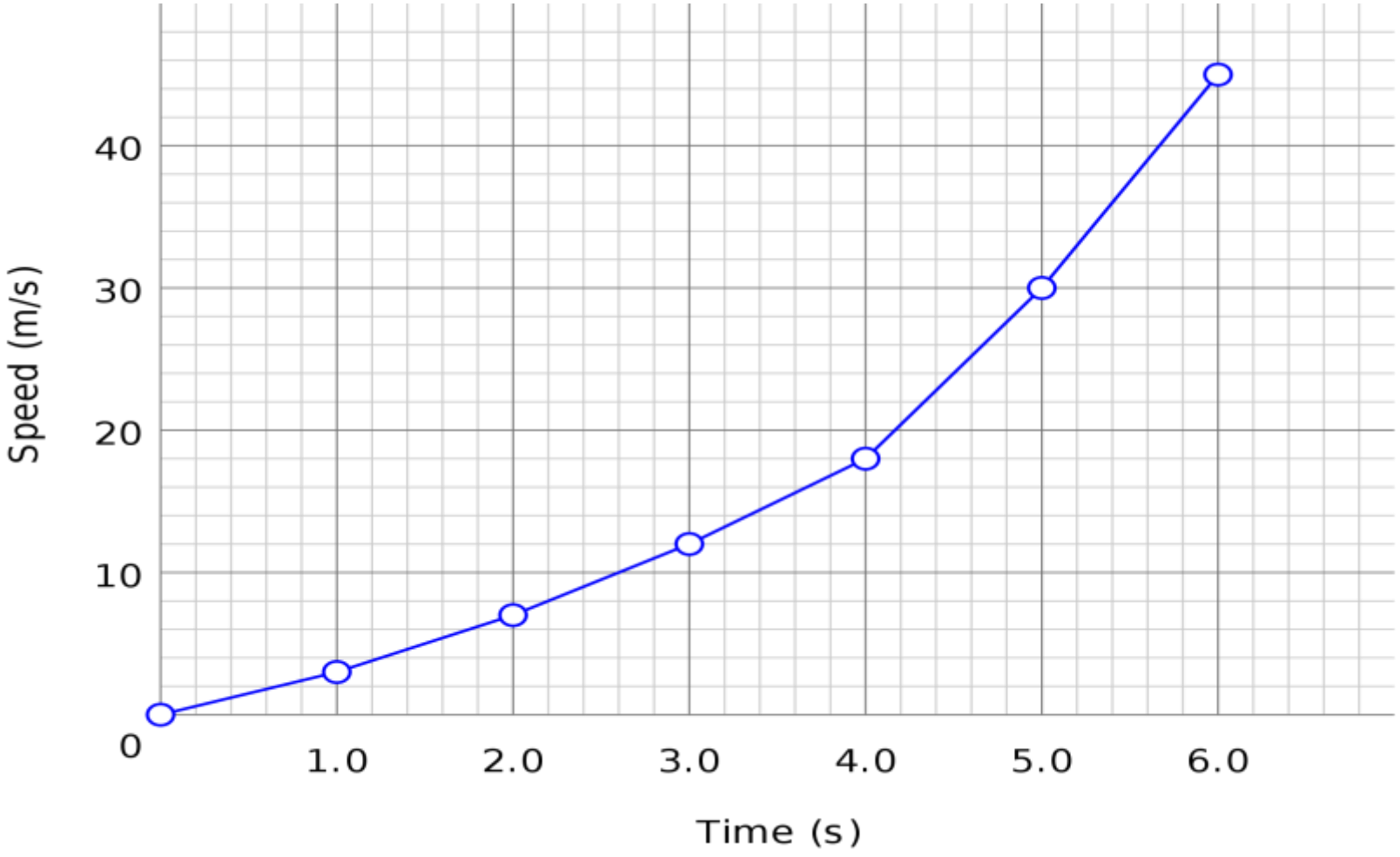
Scatter Plot

- Below figure shows the Pearson Correlation Coefficient and associated scatterplots. Correlation can be positive, zero and negative.



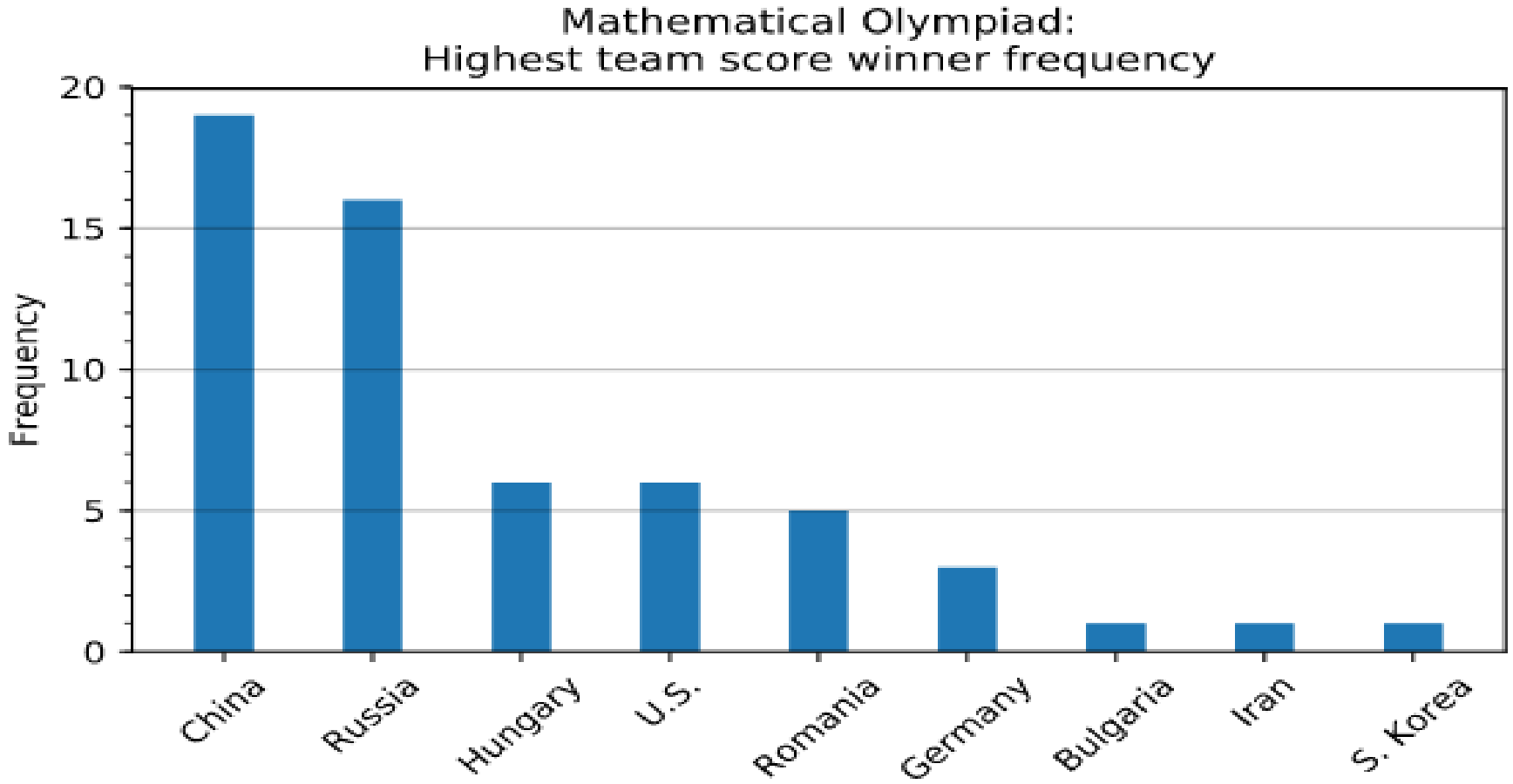
Line Plot

Generally used for plotting the time series data. Most of the time X-axis will be time. Line plot Consists of the markers and the connecting line segments. It is Similar to Scatter plot but the markers are joined through the line segments.



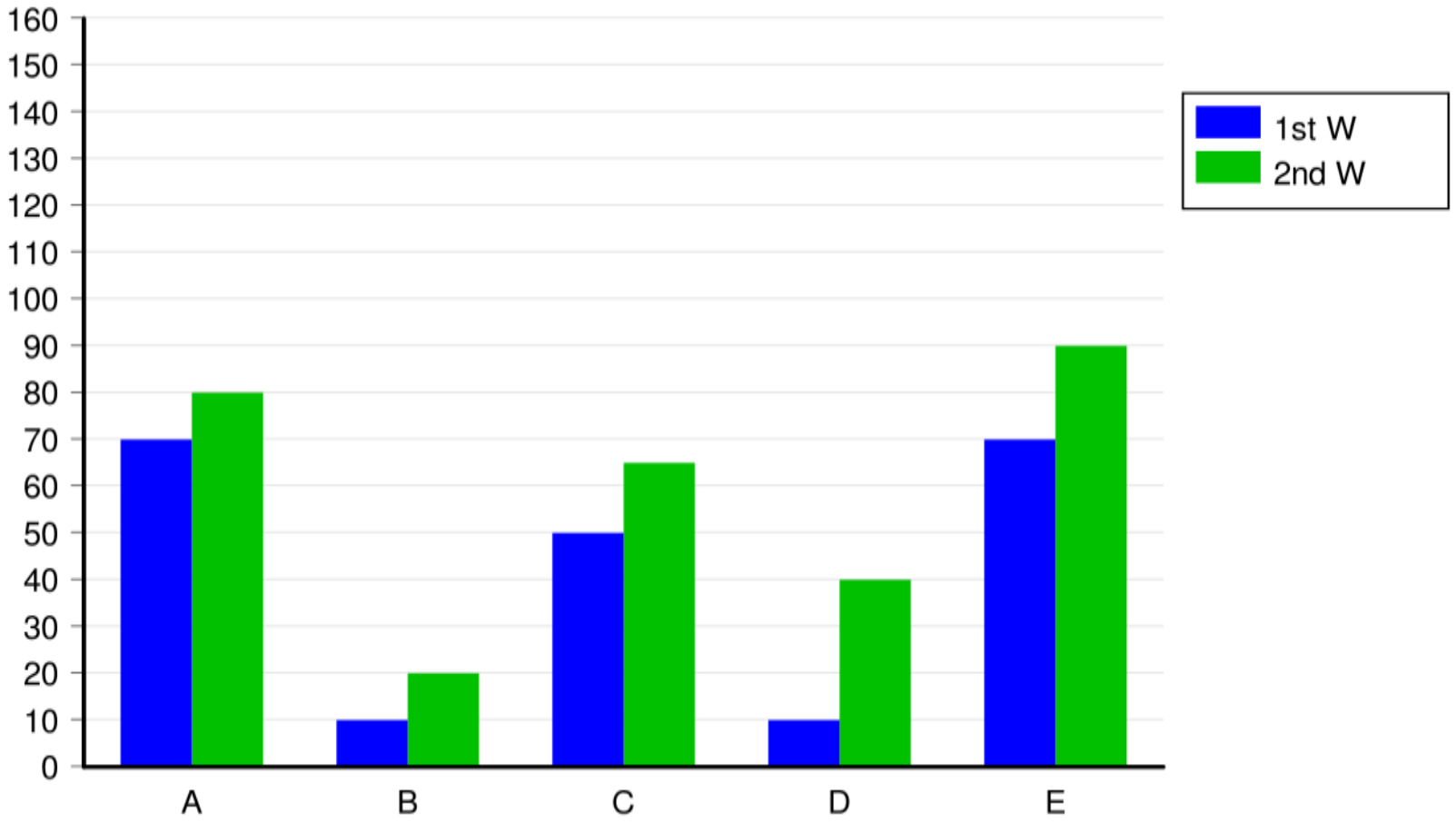
Bar chart

Bar chart represents the categorical data using the rectangular bars which represent each category. X - axis consists of categories and Y - axis consists of respective values. It is can be presented both horizontally and vertically.



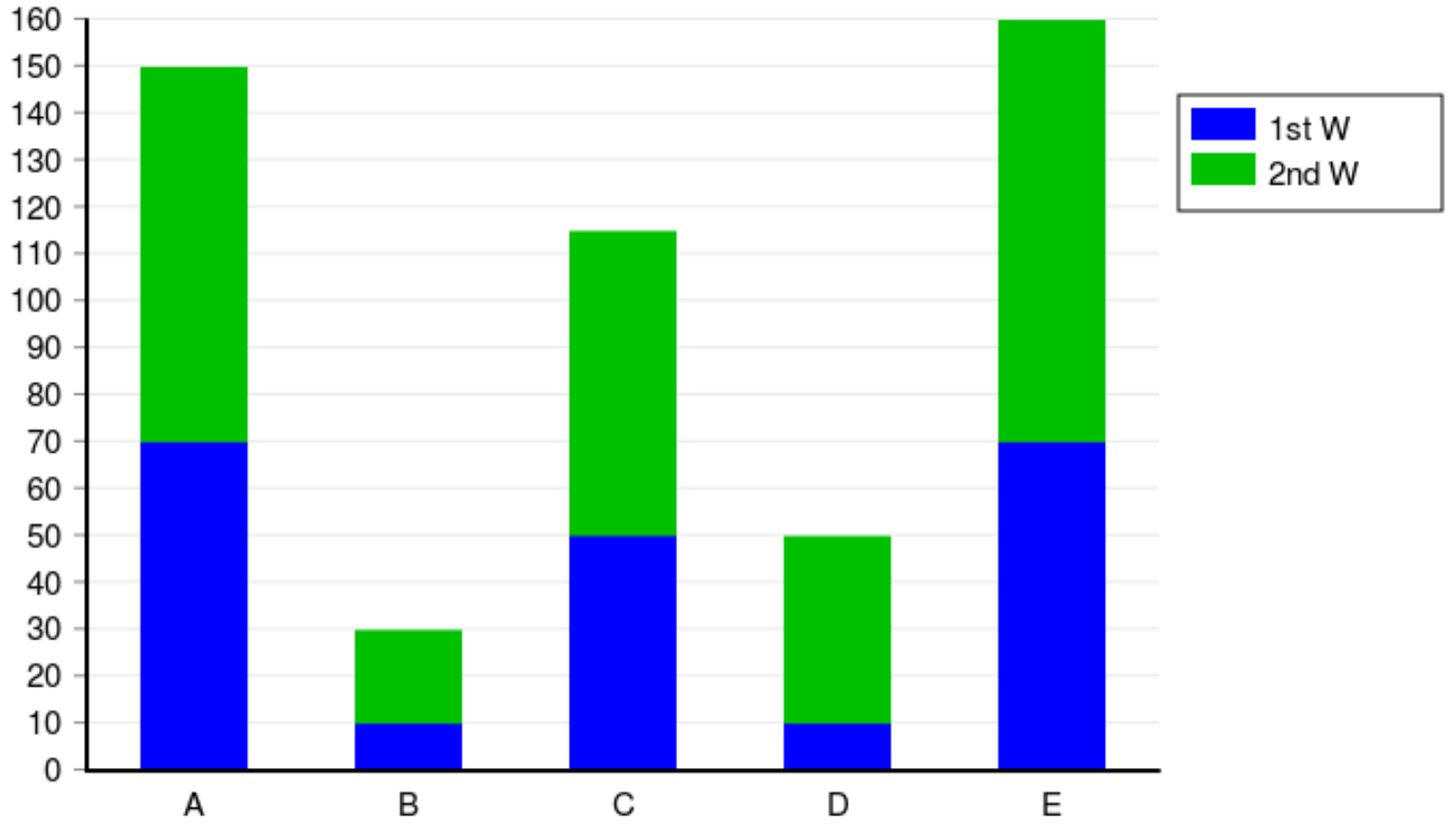
Grouped Bar chart

Is a type of barchart that represent the groups of categories or different datasets presented as bars side by side for comparison.



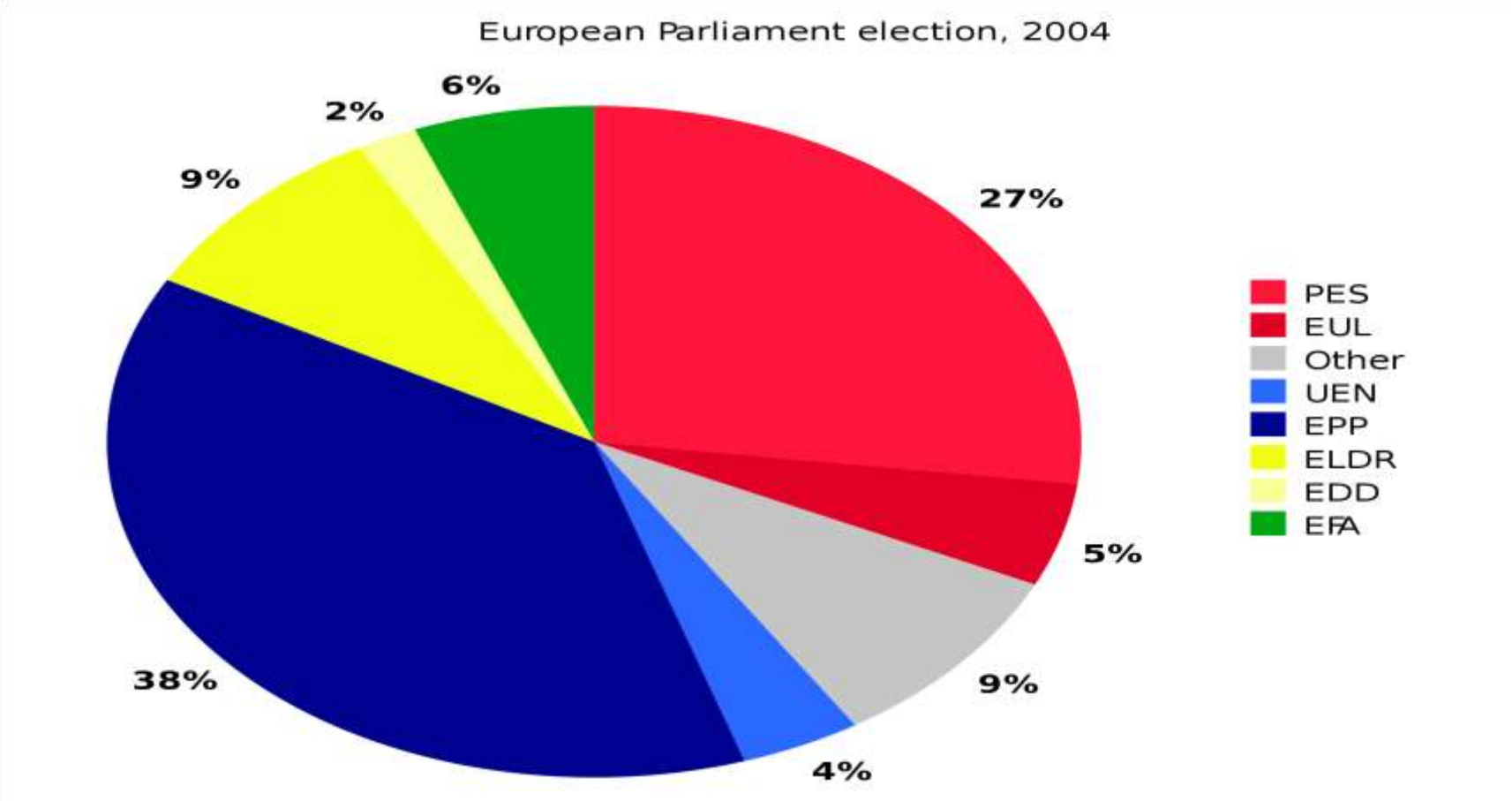
Stacked Bar chart

Is a type of barchart that represent the groups of categories or different datasets presented as bars stacked vertically. Each standard bar is divided in to sub bars that are stacked together



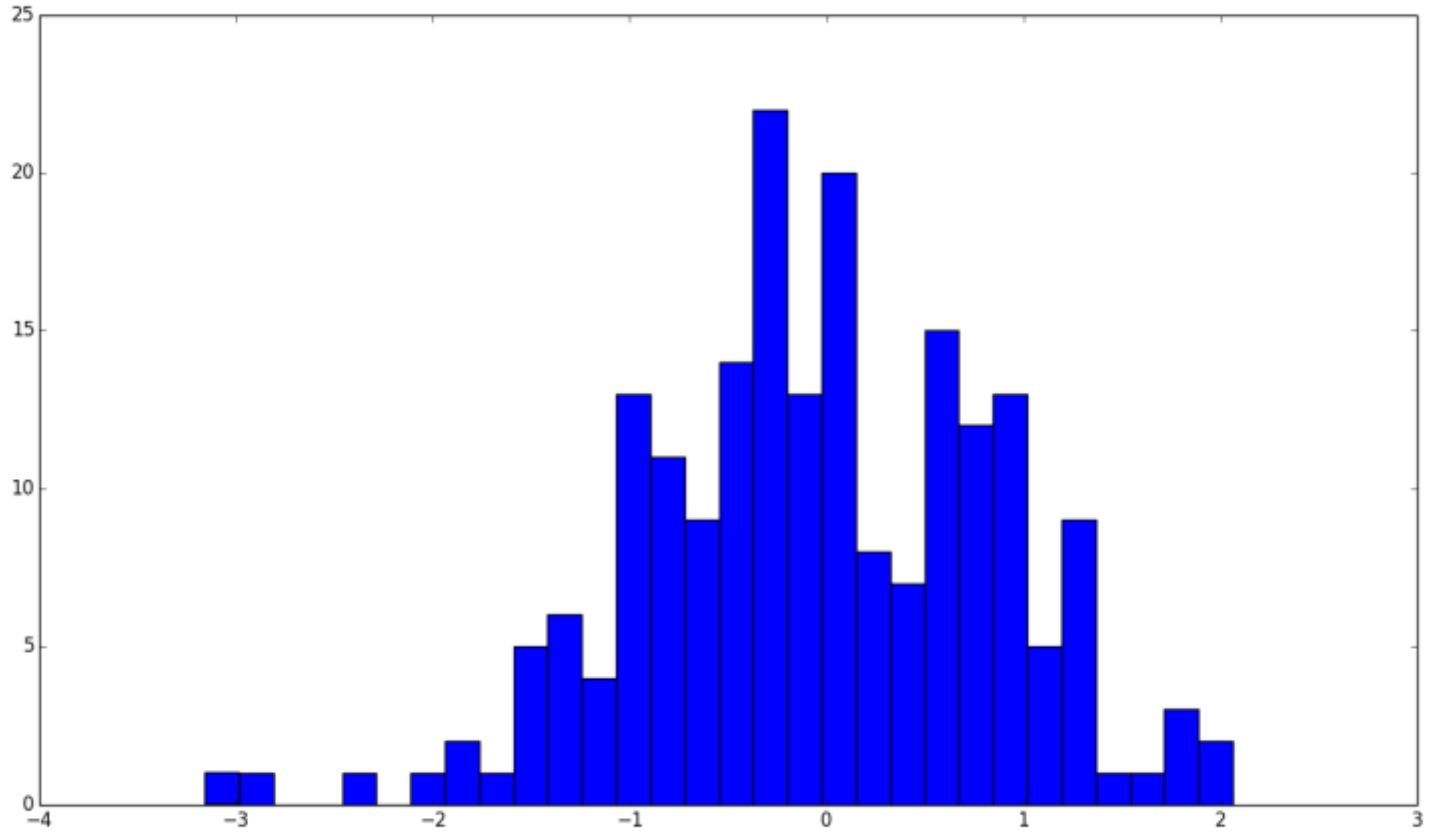
Pie chart

Pie chart is a type of chart that shows the size or percentage of the data by using the pie slices. Basically it is a circle but as it is sliced across like a pie got that name.



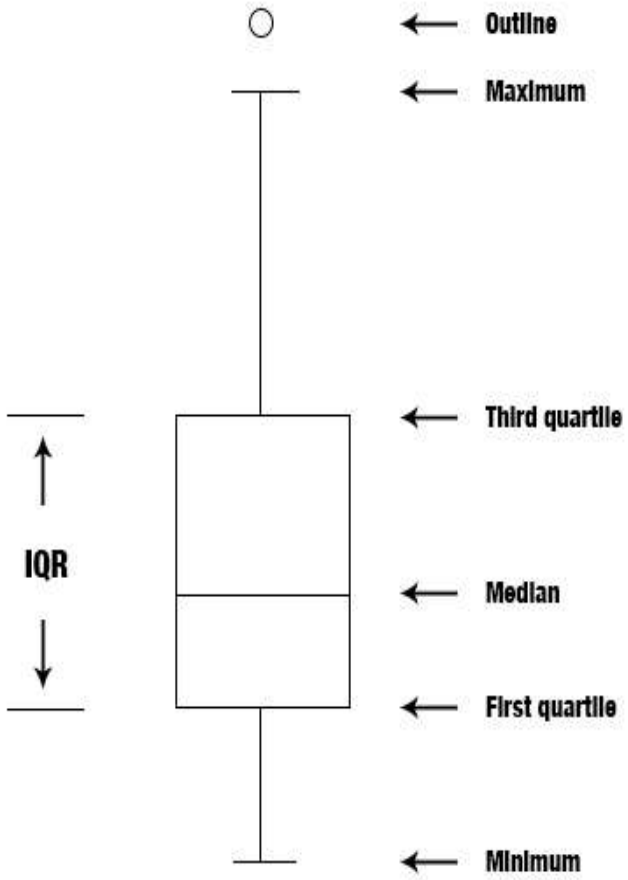
Histogram plot

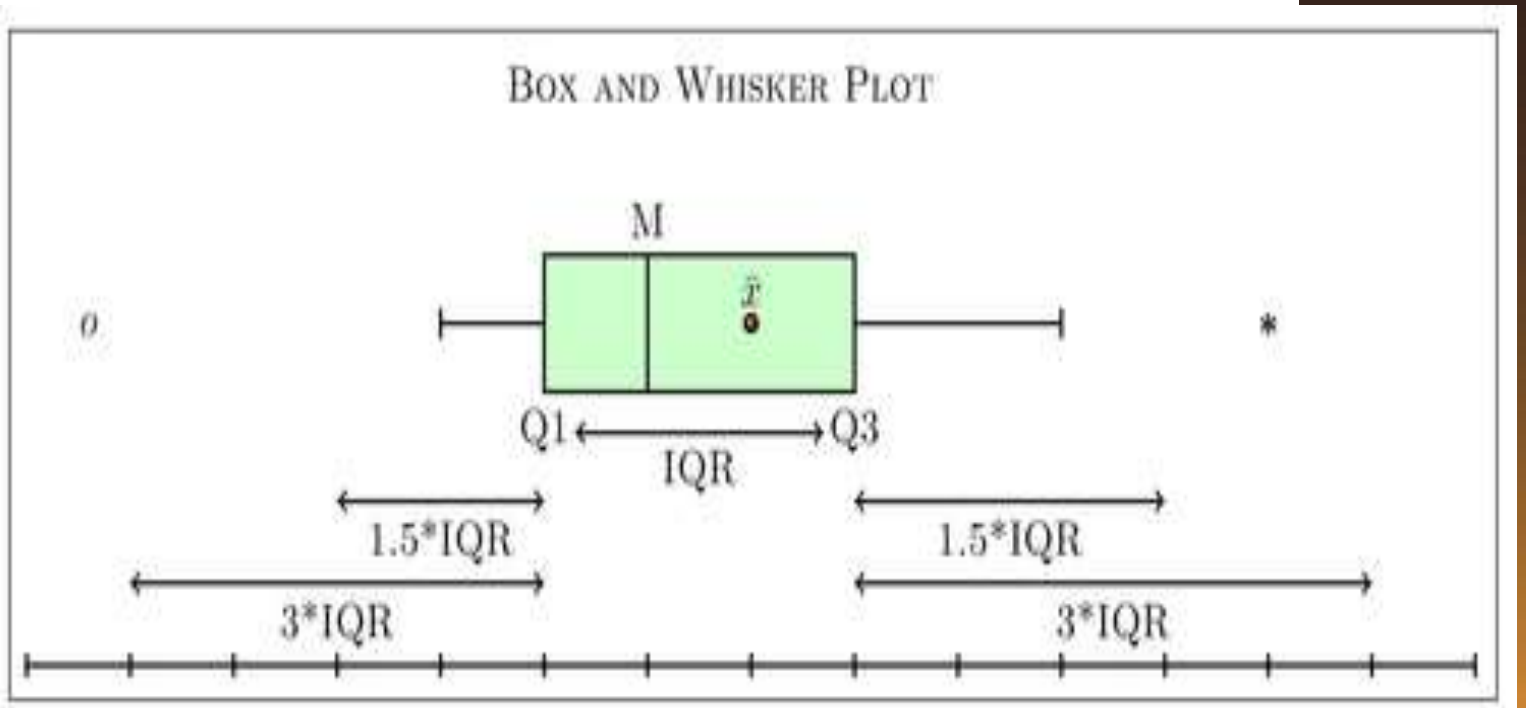
Histogram plot represents the frequency distribution of the numerical continuous data. presents the data grouped into different rectangular bins as shown



Box plot

Represents the data distribution by with respect to median as the reference. It is mainly used for outlier identification in the univariate analysis of data.





Export only template files

- **Random Variable**

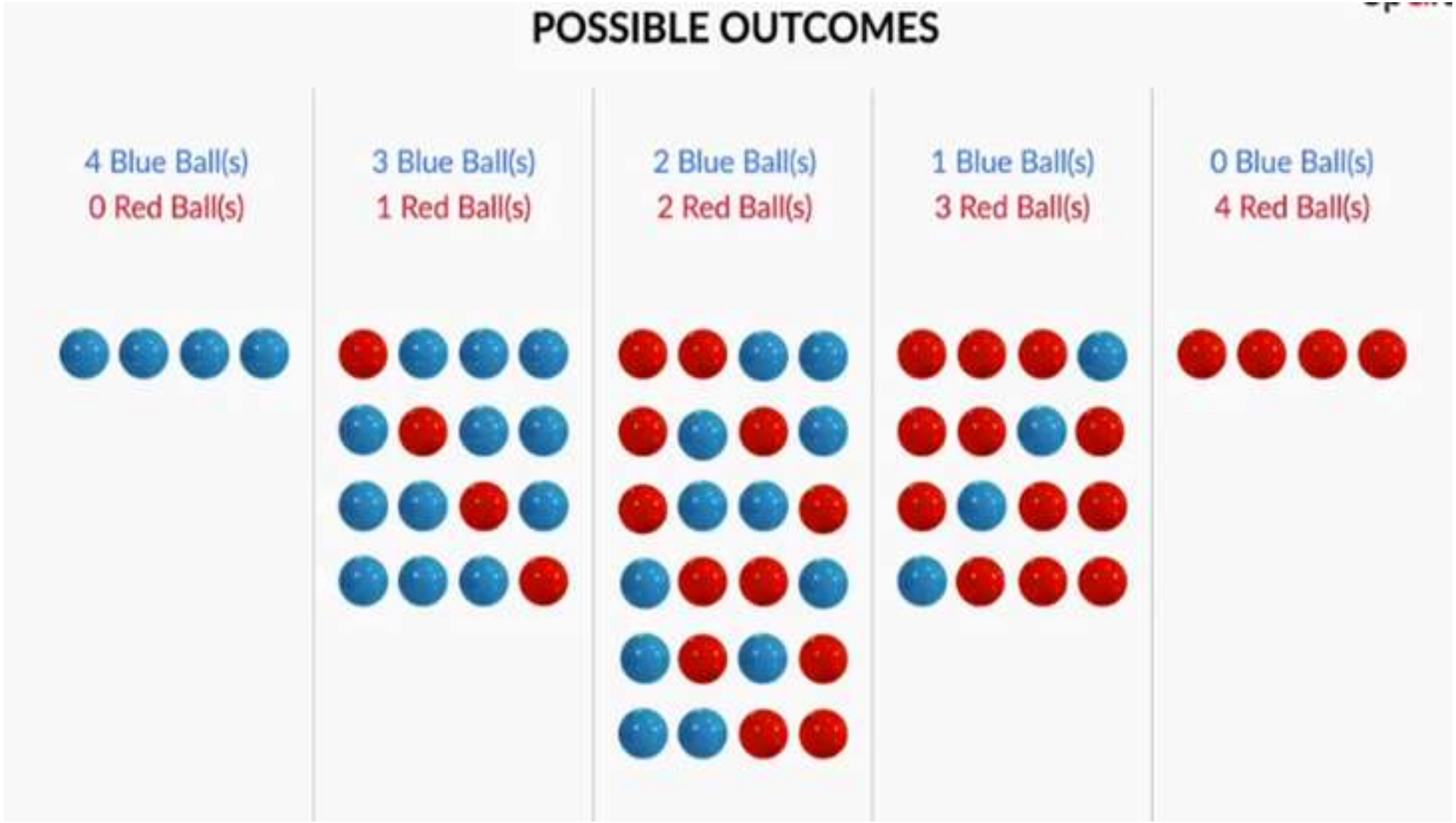
APPROACH FOR SOLVING THE UPGRAD GAME PROBLEM

1. Find possible combinations
2. Find probability of each combination
3. Use the probabilities to estimate profit/loss per player

- Take ball from the bag 4 times and not the colour-With Replacement.

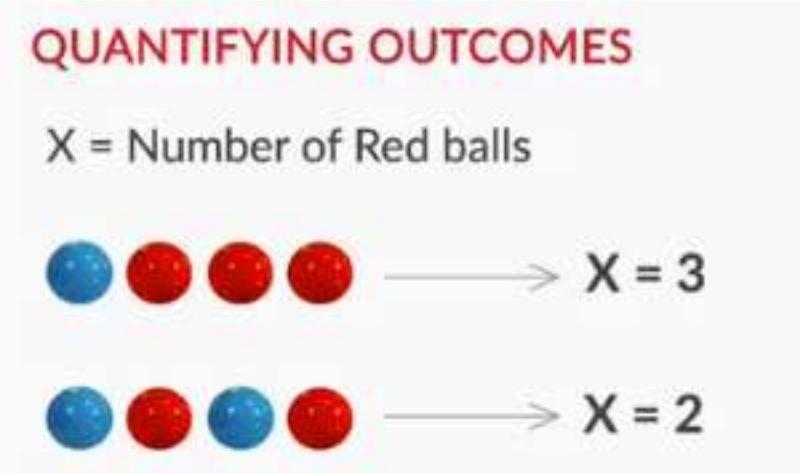
Random Variable

- **Random Variable**



Random Variable

- **Random Variable**

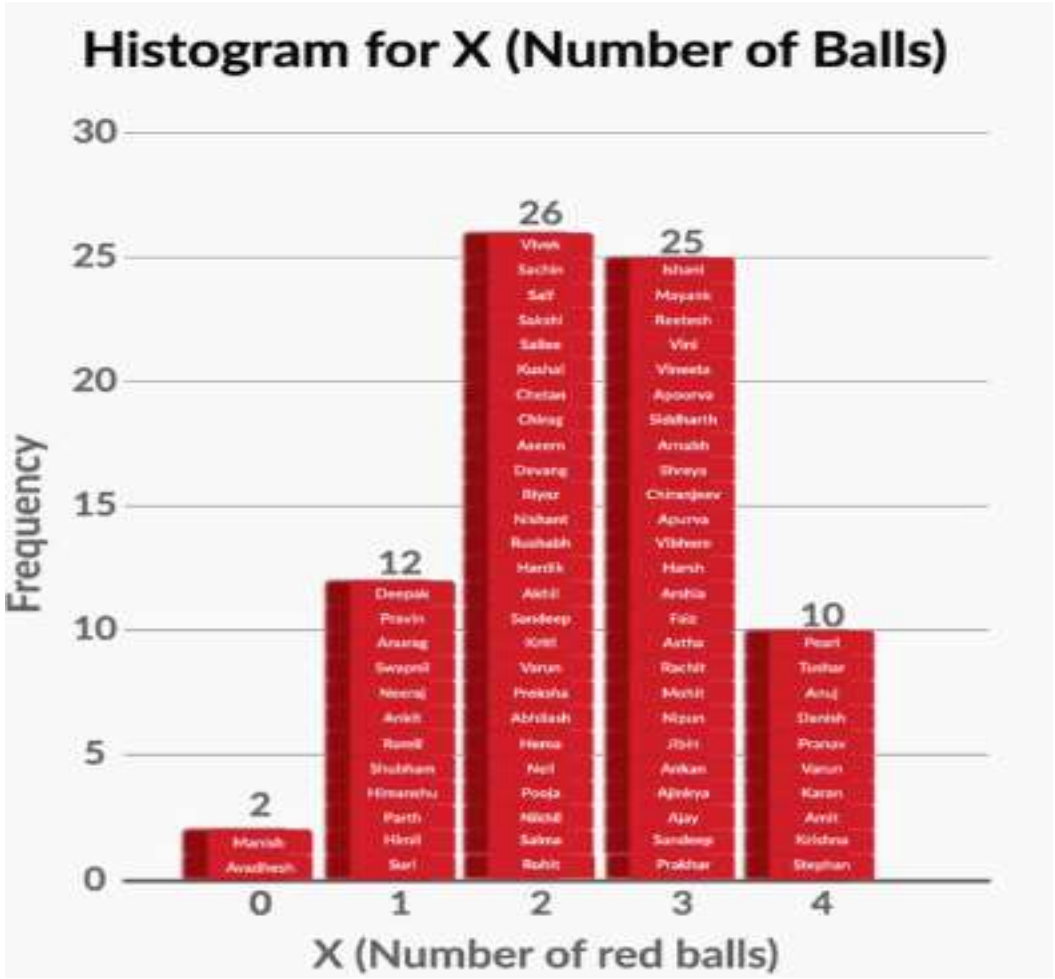


RANDOM VARIABLE

1. Denoted by X
2. Any variable that associates an outcome to a number

Random Variable

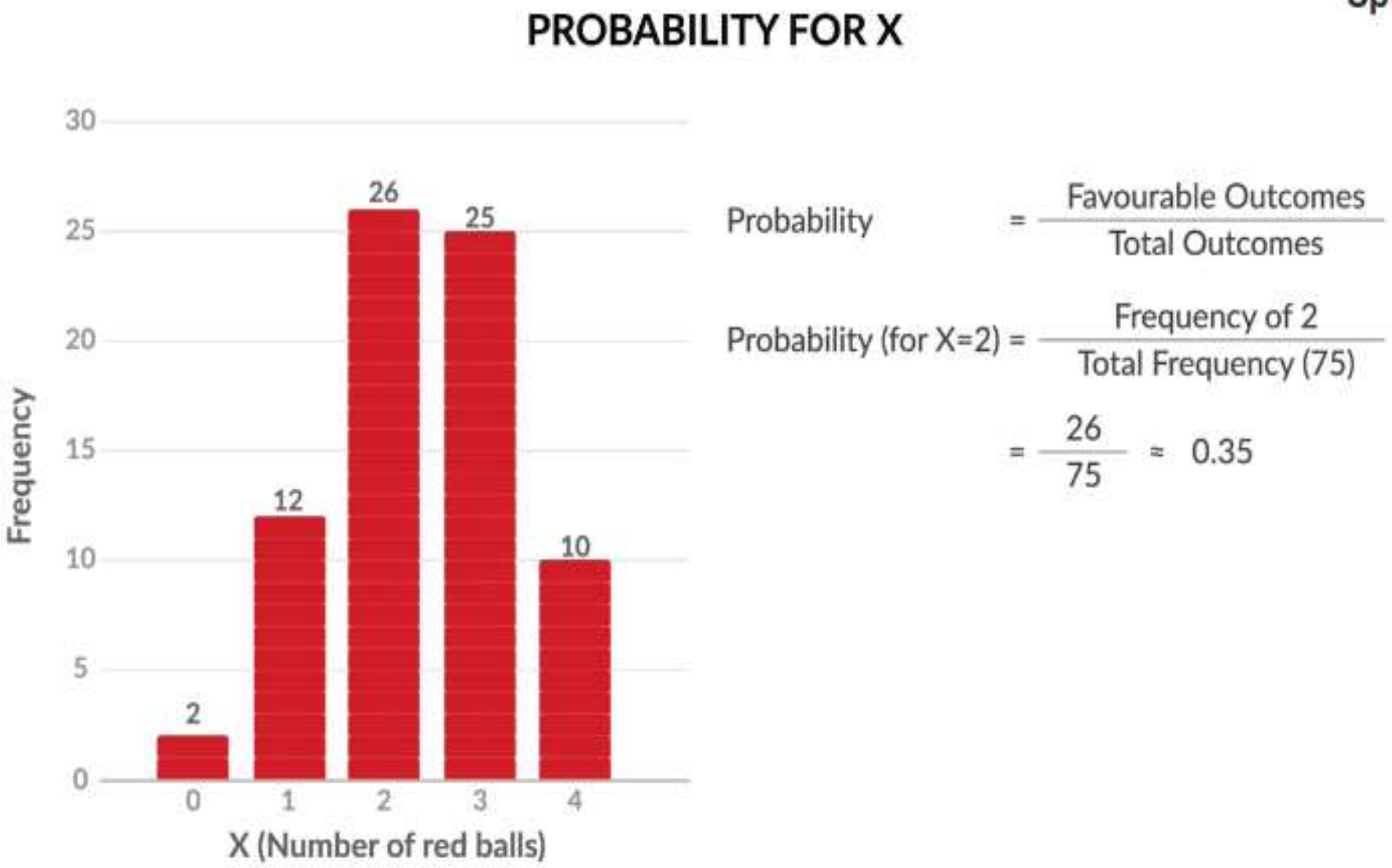
- **Random Variable**



Random Variable

- **Random Variable**

UpGrad



$$\text{Probability} = \frac{\text{Favourable Outcomes}}{\text{Total Outcomes}}$$

$$\text{Probability (for X=2)} = \frac{\text{Frequency of 2}}{\text{Total Frequency (75)}}$$

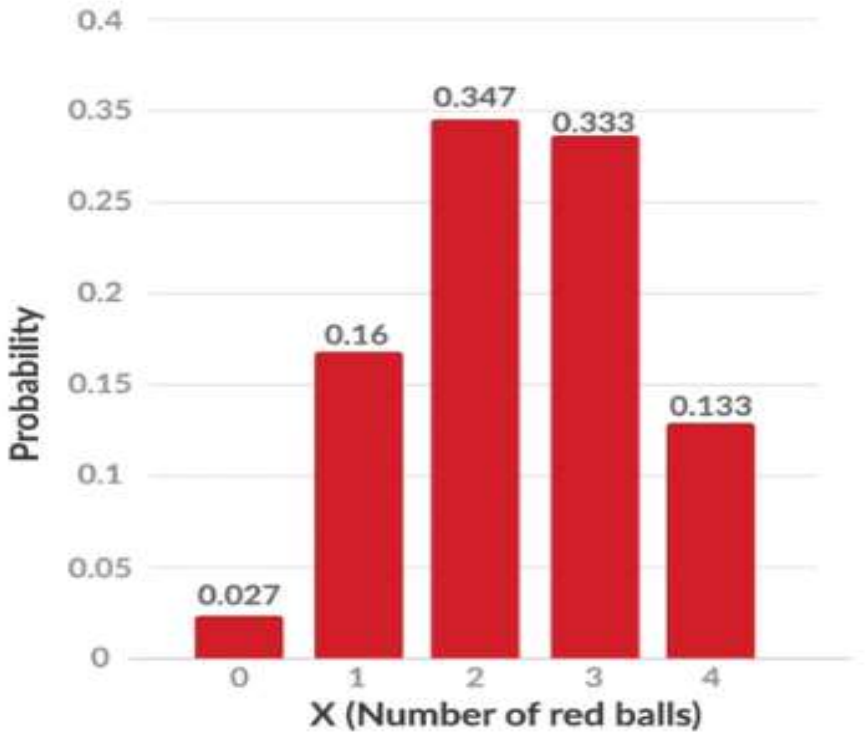
$$= \frac{26}{75} \approx 0.35$$

Random Variable

- **Random Variable**

PROBABILITY DISTRIBUTION – TABLE FORM vs BAR CHART FORM

x	P(x)
0	0.027
1	0.160
2	0.347
3	0.333
4	0.133



Random Variable

- **Random Variable**

UpGrad

EXPECTED VALUE CALCULATION

X can take $x_1, x_2, x_3, \ldots\ldots, x_n$

$EV = x_1 * P(X = x_1) + x_2 * P(X = x_2) + x_3 * P(X = x_3) + \ldots\ldots + x_n * P(X = x_n)$

Here, $P(X = x_i)$ denotes the probability that X is equal to x_i

So, $EV = 0*(0.027) + 1*(0.160) + 2*(0.347) + 3*(0.333) + 4*(0.133)$

Random Variable

- **Random Variable**

RANDOM VARIABLE DEFINITION

X = Money won after playing the game once

X can take two values: +150 and -10

P(X = +150) = P(4 red balls) = 0.133

P(X = -10) = P(0, 1, 2 or 3 red balls) = 0.027 + 0.160 + 0.347 + 0.333 = 0.867

So, EV = (150 x 0.133) + (-10 x 0.867) = +11.28

- https://learn.upgrad.com/v/course/376/session/44930/segment/245422

Probability Without Experiment

- **Random Variable**

UpGrad

MULITIPLICATION RULE OF PROBABILITY

$$P(\text{1 red ball in 1 trial}) = \frac{\text{Number of red balls}}{\text{Number of total balls}} = \frac{3}{5} = 0.6$$

$$P(\text{Event 1 AND Event 2}) = P(\text{Event 1}) \times P(\text{Event 2})$$

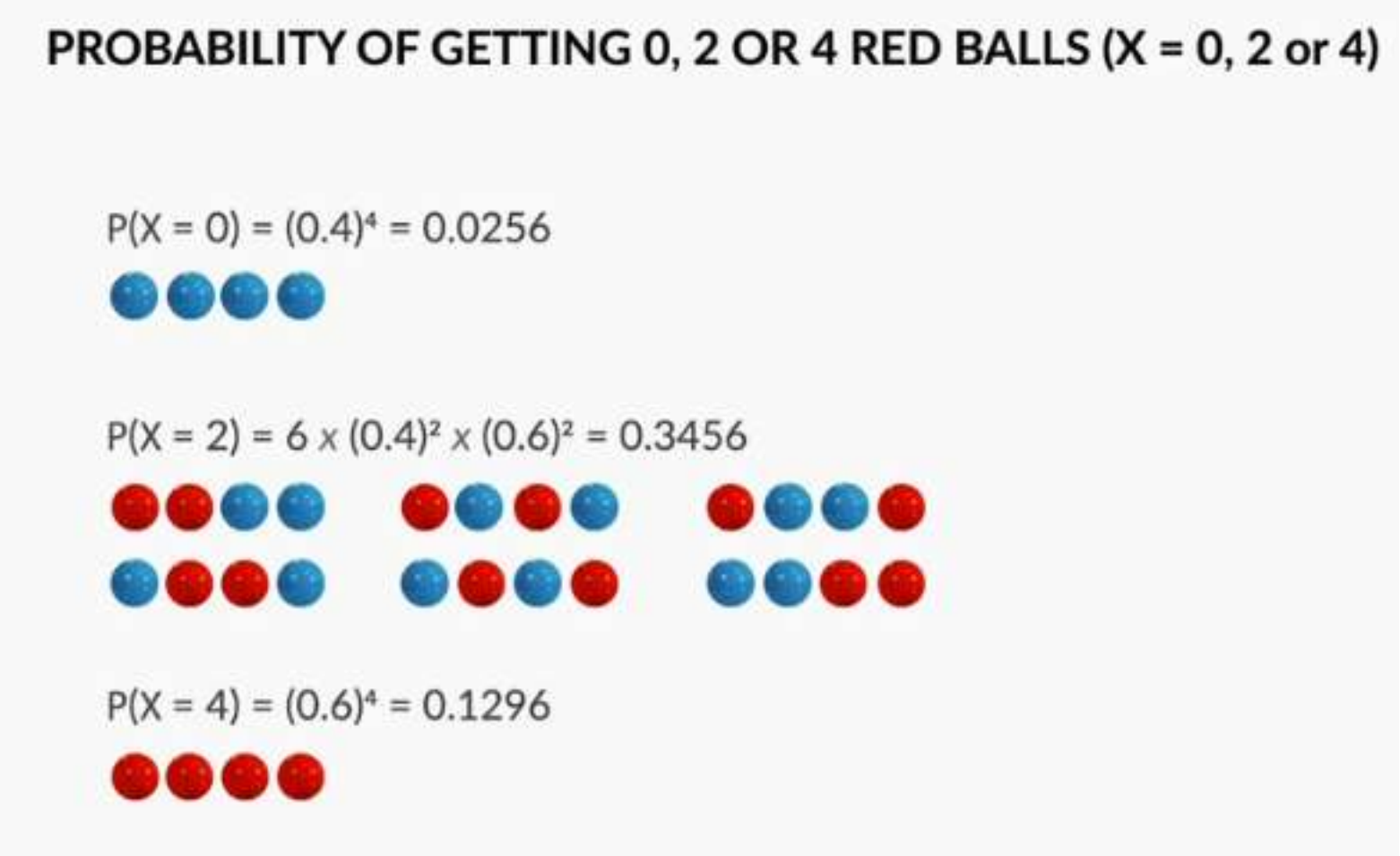
$$P(\text{2 red balls in 2 trials}) = P(\text{red ball in 1st trial}) \times P(\text{red ball in 2nd trial})$$

$$= 0.6 \times 0.6$$

$$= 0.36$$

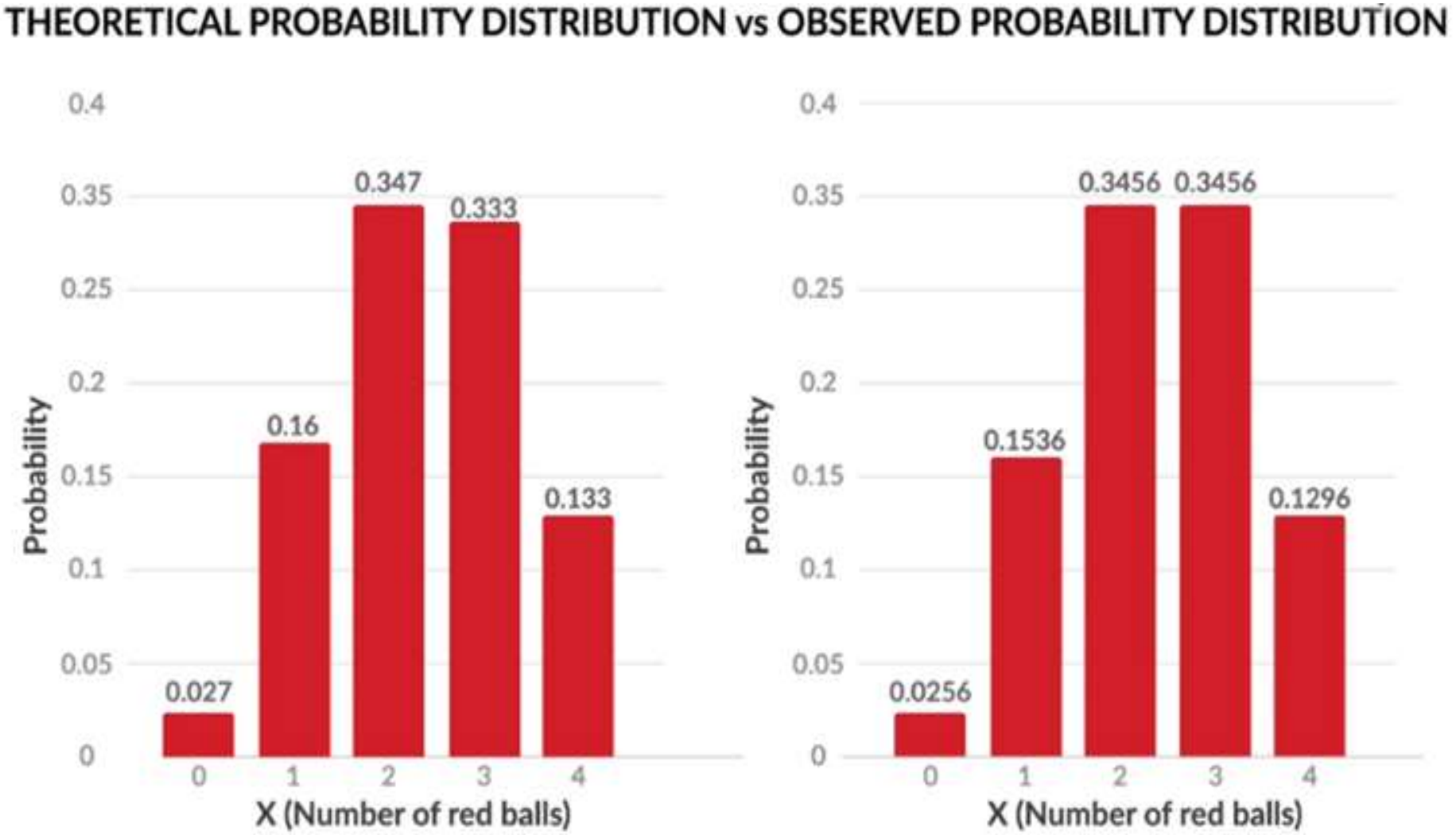
Probability Without Experiment

- **Random Variable**



Probability Without Experiment

- **Random Variable**



Probability Without Experiment

•Random Variable

So, the formula for finding **binomial probability** is given by:

$$P(X = r) = {}^{n}C_{r}(p)^{r}(1-p)^{n-r}$$

Where **n** is **the number of trials**, **p** is the **probability of success**, and **r** is the **number of successes after n trials**.

However, there are some **conditions** that need to be met in order for us to be able to apply the formula.

1. The **total number** of trials is **fixed** at **n**.
2. Each trial is **binary**, i.e., it has **only two possible outcomes:** success or failure.
3. **Probability of success** is the **same** in all trials, denoted by **p**.

Probability Without Experiment

- **Random Variable**

CUMULATIVE PROBABILITY DISTRIBUTION

$F(x) = P(X \leq x)$

E.g. $F(1) = P(X \leq 1) = P(X=0) + P(X=1)$

Similarly, $F(2) = P(X \leq 2) = P(X=0) + P(X=1) + P(X=2)$

Probability Without Experiment

- **Random Variable**

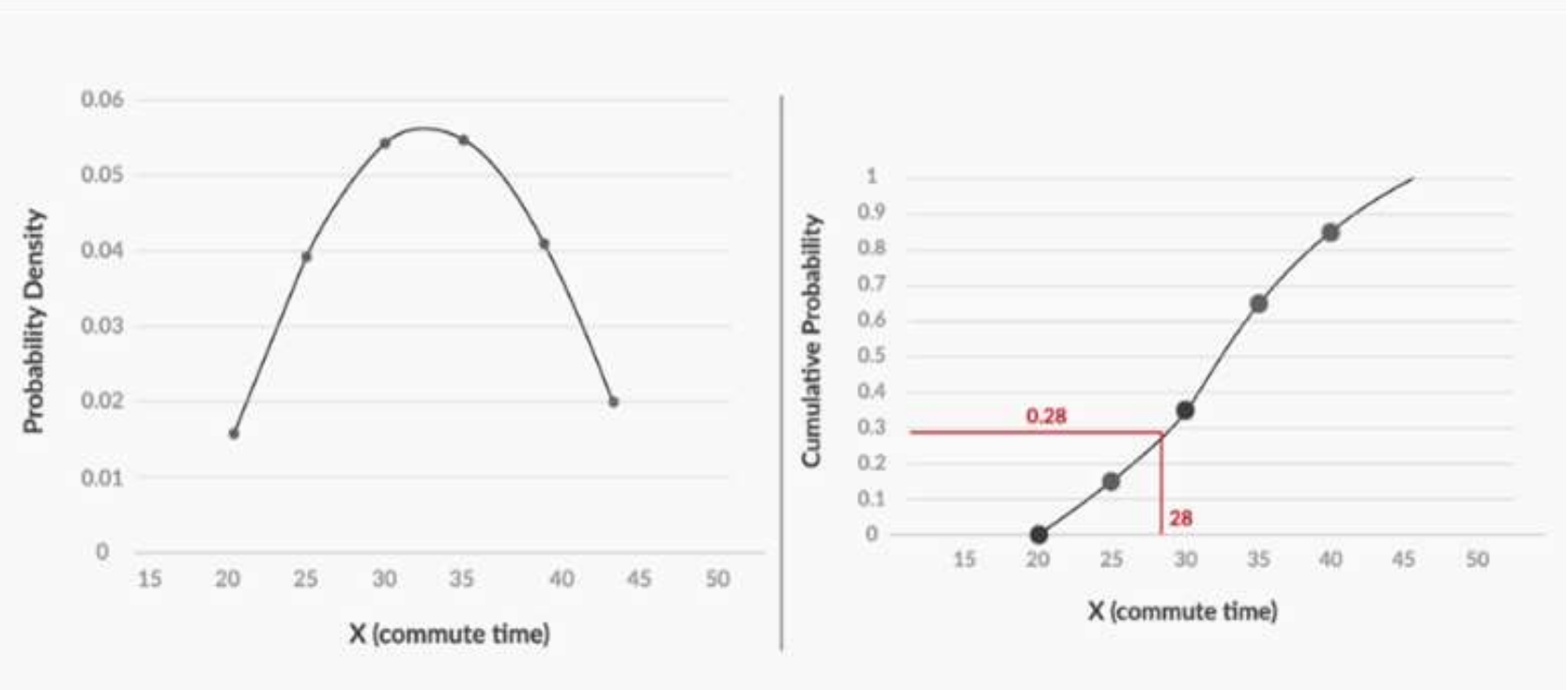
Daily Commute Time of Company X's Employees

x(Commute Time)	Probability
20-25	0.15
25-30	0.20
30-35	0.30
35-40	0.20
40-45	0.15

x(Commute Time)	P(X≤x)
25	0.15
30	0.35
35	0.65
40	0.85
45	1.00

Probability Without Experiment

- **Random Variable**



- https://learn.upgrad.com/v/course/376/session/44932/segment/245438

Probability Without Experiment

- **Random Variable**

Normally distributed data follows the **1-2-3 rule**. This rule states that there is a:

1. **68%** probability of the variable lying **within 1 standard deviation** of the mean,
2. **95%** probability of the variable lying **within 2 standard deviations** of the mean, and
3. **99.7%** probability of the variable lying **within 3 standard deviations** of the mean.

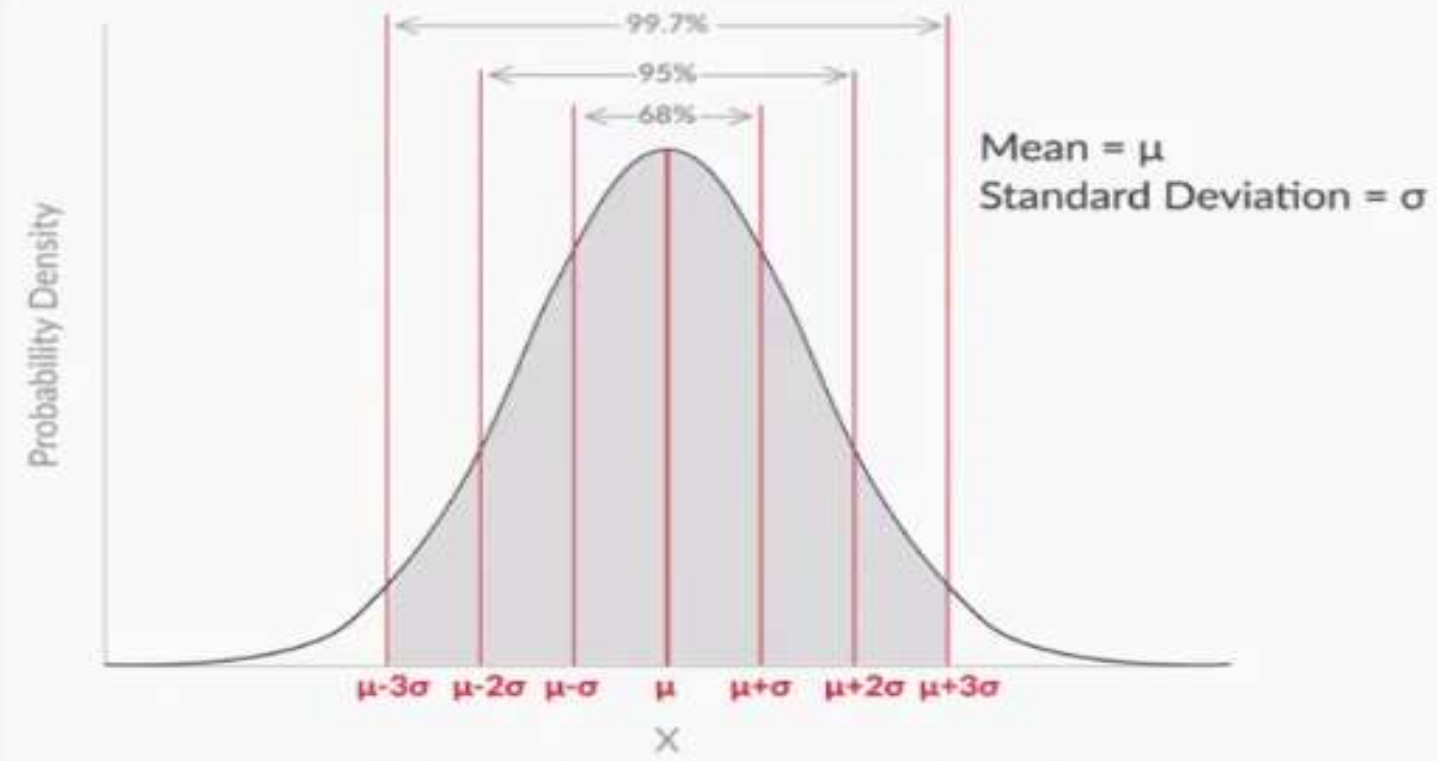
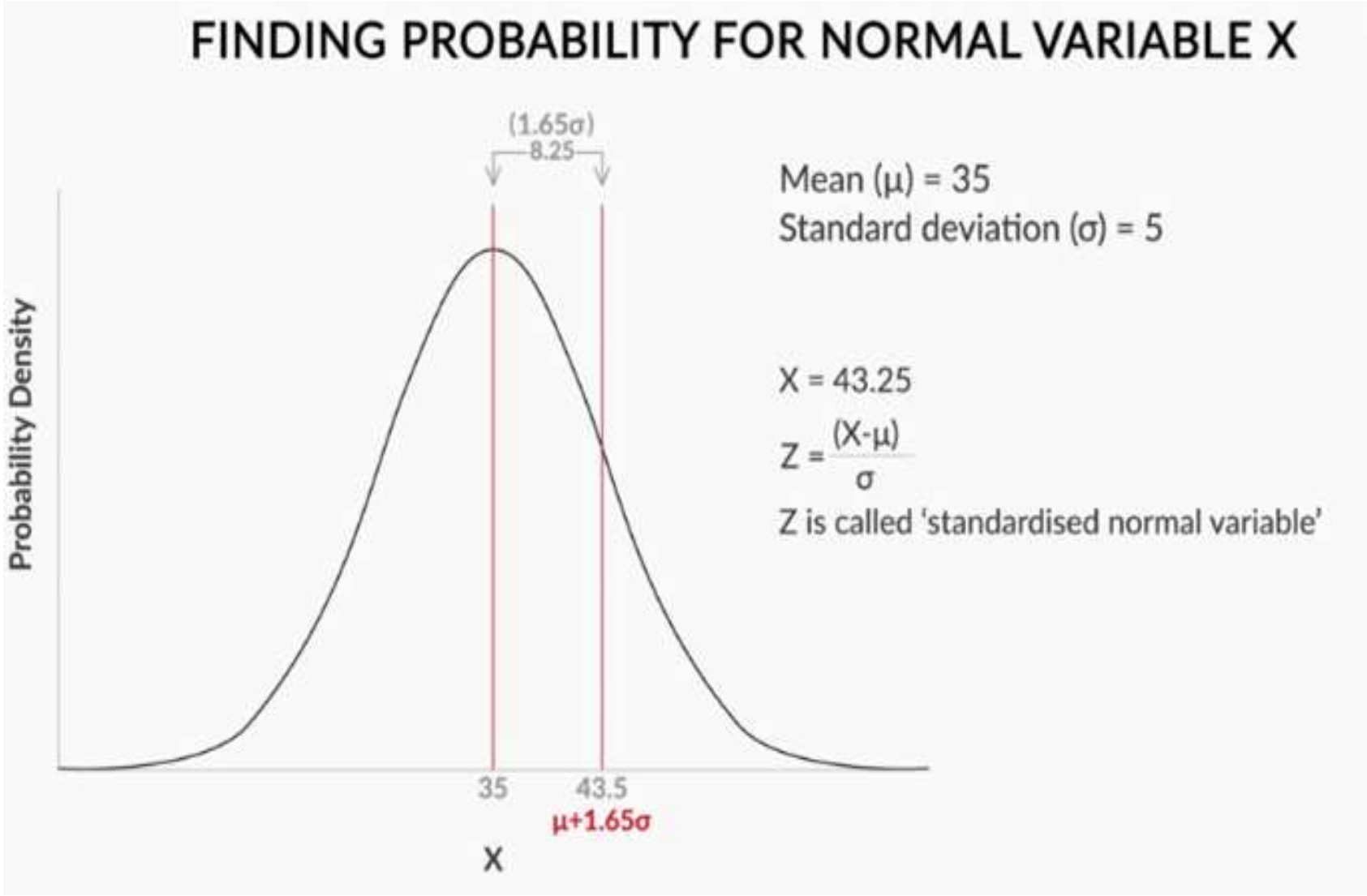


Figure 9 - 1-2-3 Rule for a Normal Distribution

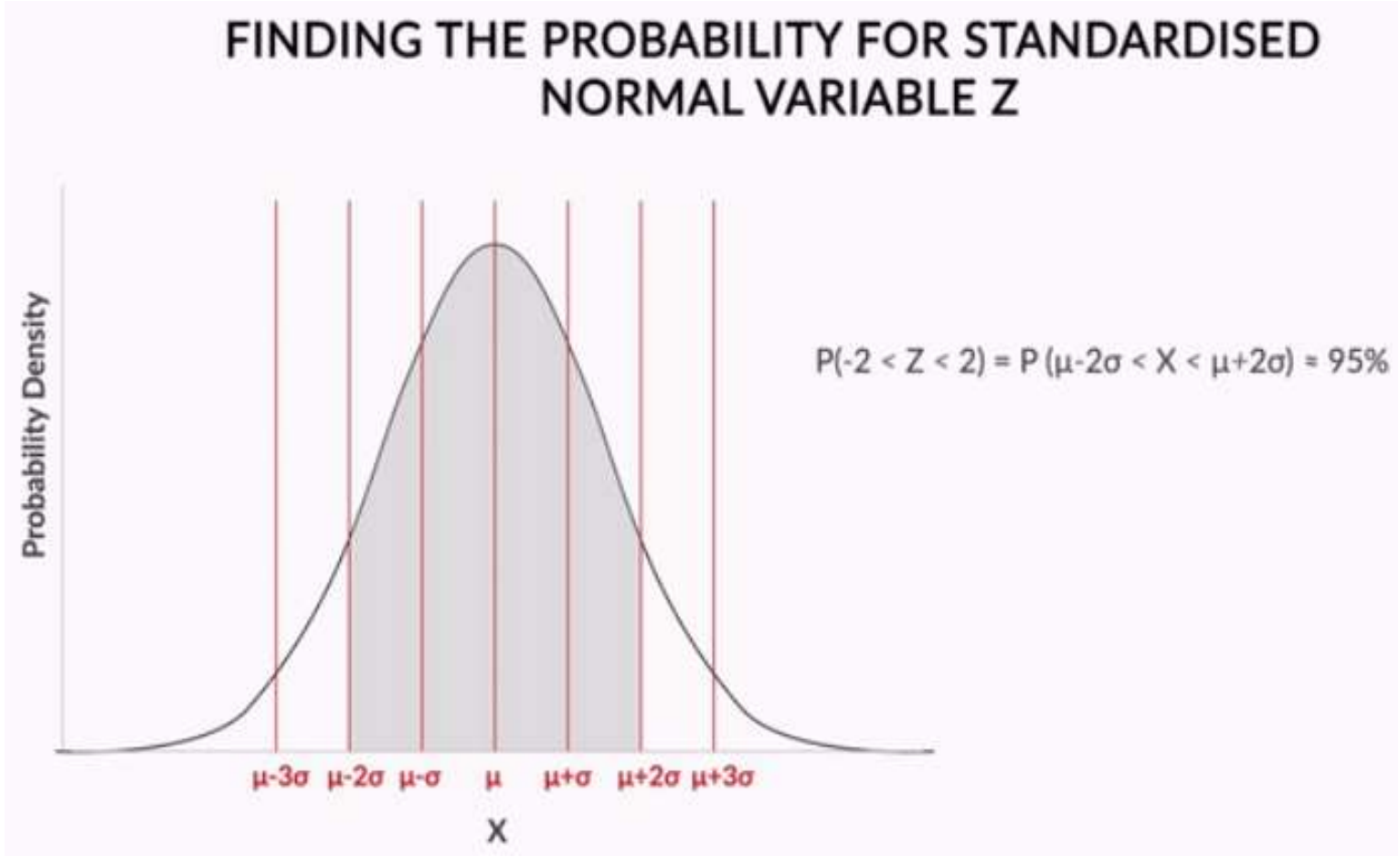
Probability Without Experiment

- **Random Variable**



Probability Without Experiment

- **Random Variable**



Continuous Variables

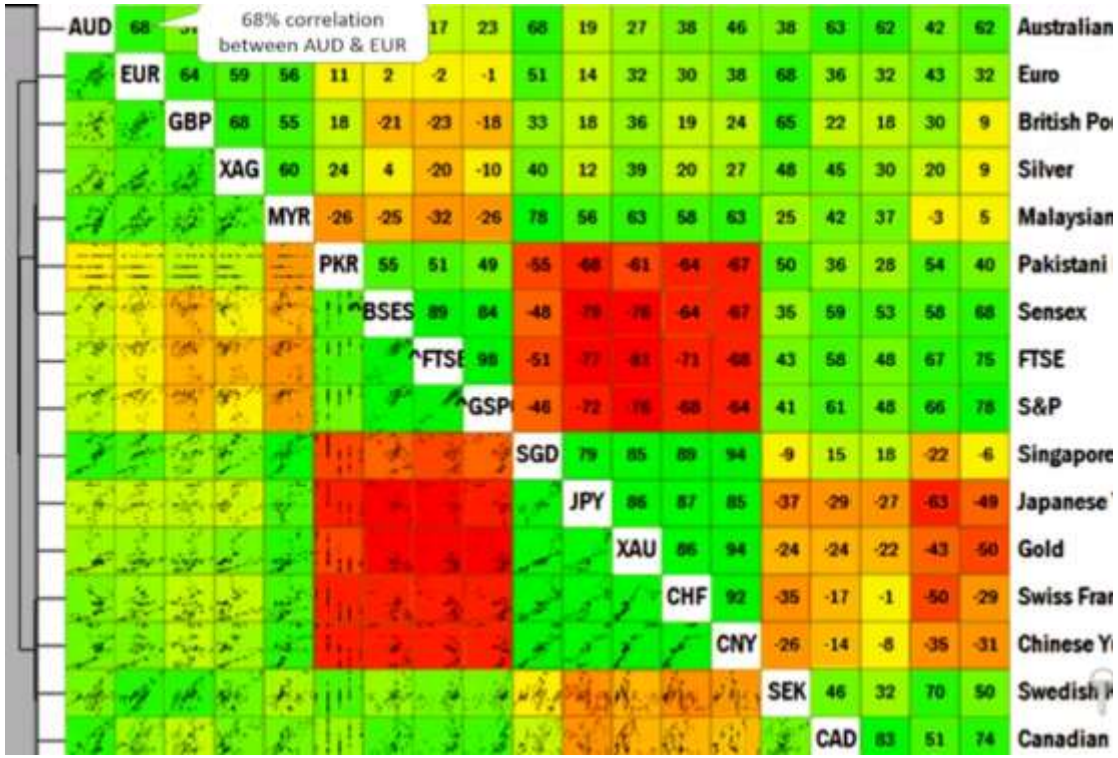
Bivariate Analysis on Continuous Variables

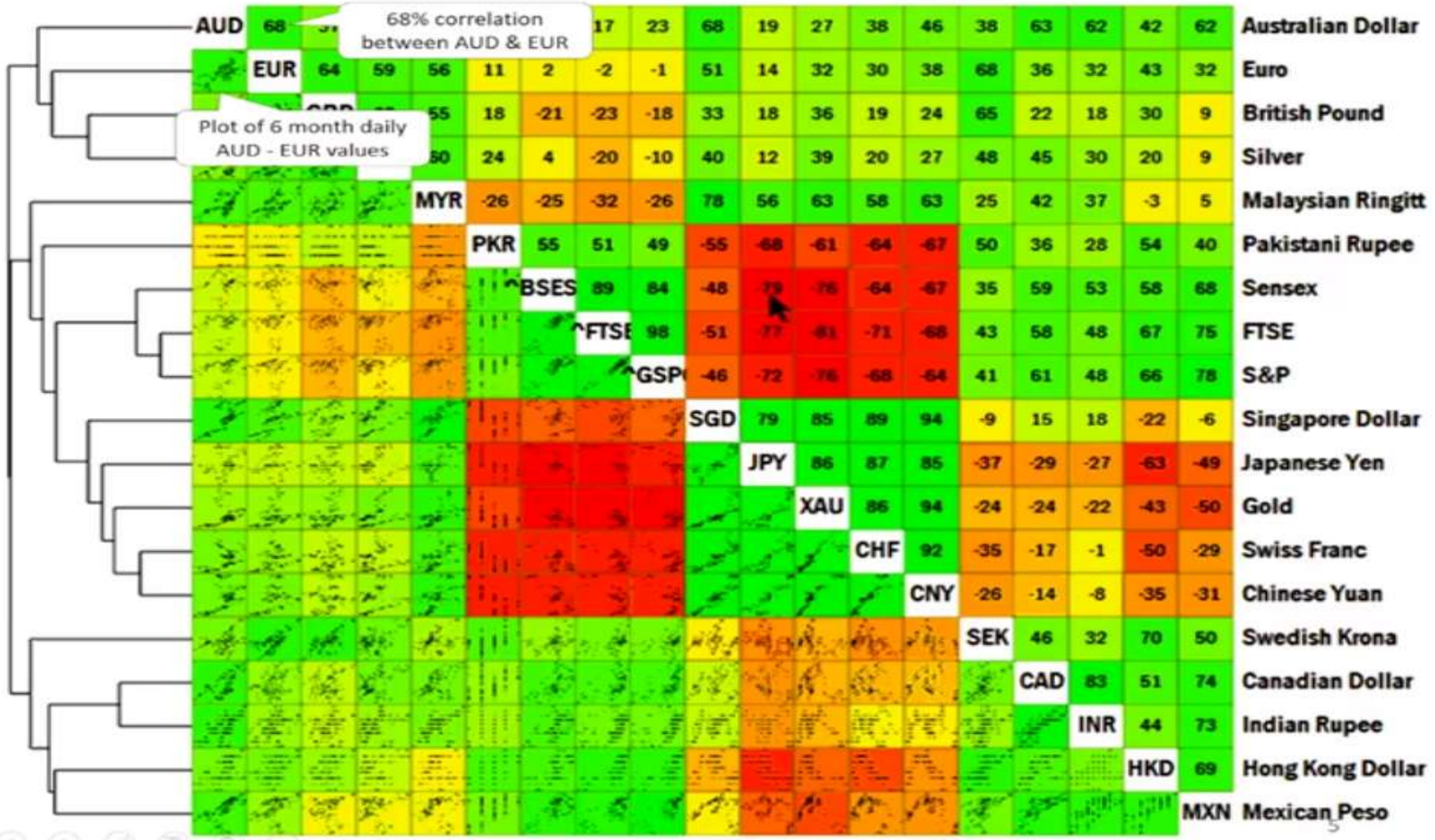
- The type of analysis that we do for continuous variable is to check that when one variable raises does another variable gets affected by it or not? This is calculated by correlation.
- correlation is a number between -1 and 1 which quantifies the extent to which two variables 'correlate' with each other.
  - If one increases as the other increases, the correlation is positive
  - If one decreases as the other increases, the correlation is negative
  - If one stays constant as the other varies, the correlation is zero

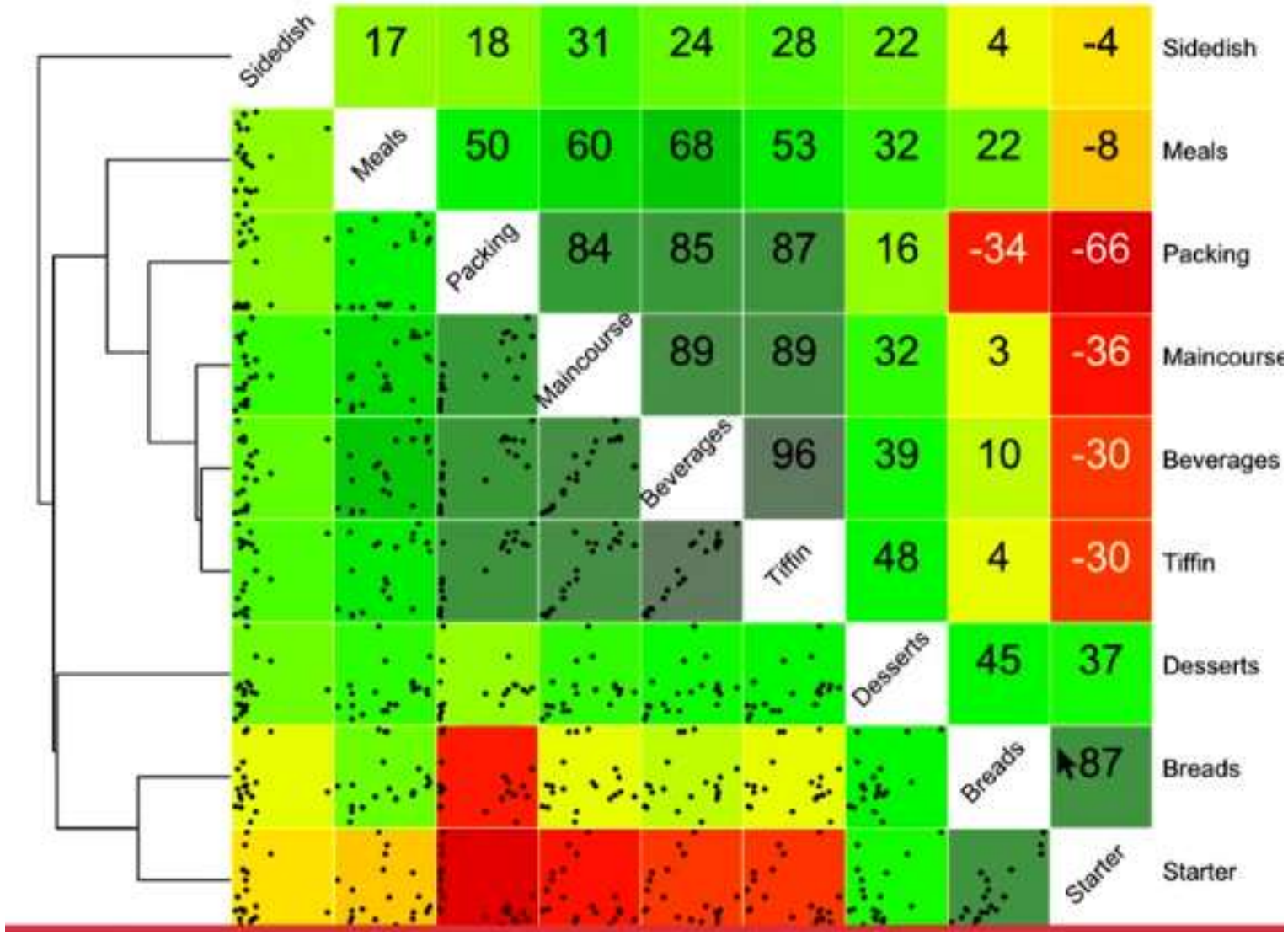
- In general, a positive correlation means that two variables will increase together and decrease together, e.g. an increase in rain is accompanied by an increase in humidity. A negative correlation means that if one variable increases the other decreases, e.g. in some cases, as the price of a commodity decreases its demand increases.

- A perfect positive correlation means that the correlation coefficient is exactly 1. This implies that as one variable moves, either up or down, the other one moves in the same direction. A perfect negative correlation means that two variables move in opposite directions, while a zero correlation implies no relationship at all.

Business Problems Involving Correlation




68% correlation between AUD & EUR
Plot of 6 month daily AUD - EUR values
AUD 68 17 23 68 19 27 38 46 38 63 62 42 62 Australian Dollar
EUR 64 59 56 11 2 -2 -1 51 14 32 30 38 68 36 32 43 32 Euro
55 18 -21 -23 -18 33 18 36 19 24 65 22 18 30 9 British Pound
50 24 4 -20 -10 40 12 39 20 27 48 45 30 20 9 Silver
MYR -26 -25 -32 -26 78 56 63 58 63 25 42 37 -3 5 Malaysian Ringitt
PKR 55 51 49 -55 -68 -61 -64 -67 50 36 28 54 40 Pakistani Rupee
^BSES 89 84 -48 -79 -76 -64 -67 35 59 53 58 68 Sensex
^FTSE 98 -51 -77 -81 -71 -68 43 58 48 67 75 FTSE
^GSP -46 -72 -76 -68 -64 41 61 48 66 78 S&P
SGD 79 85 89 94 -9 15 18 -22 -6 Singapore Dollar
JPY 86 87 85 -37 -29 -27 -63 -49 Japanese Yen
XAU 86 94 -24 -24 -22 -43 -50 Gold
CHF 92 -35 -17 -1 -50 -29 Swiss Franc
CNY -26 -14 -8 -35 -31 Chinese Yuan
SEK 46 32 70 50 Swedish Krona
CAD 83 51 74 Canadian Dollar
INR 44 73 Indian Rupee
HKD 69 Hong Kong Dollar
MXN Mexican Peso



INTERNSHIPSTUDIO

Bivariate Analysis on categorical variables

For categorical there isn't much that we can do. For correlation we can check if there is a increase of 10% in one category how much % the other category is getting changed. But category variables does not have any concept know as “Increase” or “Decrease”, It just have buch of category and it’s count.

Let’s take an example. What is the impact of class 8th science marks on father occupation?

Bivariate Analysis on categorical variables

Is the pattern same even if the parent is below poverty?

Below poverty	No	Yes
Father occupation		
Unemployed	34.3%	33.4%
Labourer	36.5%	34.9%
Farmer	37.5%	36.1%
Clerk	40.0%	36.2%
Skilled Worker	38.2%	36.0%
Business	39.7%	36.3%
Teacher/Lecturer	44.9%	36.6%
Professional	46.7%	37.6%

Bivariate Analysis on categorical variables

Does the use of calculator influence the passing %age in maths?

Does this same for both boys and girls?

Gender	Boy	Girl
Use calculator		
No	32.0%	31.8%
Yes	32.5%	33.1%

Bivariate Analysis on categorical variables

Demo in Excel- Maths Calculator

To Summaries:

1. To see the distribution of two categorical variables. For example, if you want to compare the number of boys and girls who play games, you can make a ‘cross table’ as given below:

	Everyday	Never	Once a month	Once a week	Total
Boy	3474	154	150	780	4558
Girl	2776	175	200	1046	4197
Total	6250	329	350	1826	8755

Bivariate Analysis on categorical variables

Demo in Excel- Maths Calculator

To Summaries:

1. To see the distribution of two categorical variables. For example, if you want to compare the number of boys and girls who play games, you can make a ‘cross table’ as given below:

	Everyday	Never	Once a month	Once a week	Total
Boy	3474	154	150	780	4558
Girl	2776	175	200	1046	4197
Total	6250	329	350	1826	8755

From this table, firstly, you can compare boys and girls across a fixed level of ‘play games’, e.g. a higher number of boys play games every day than girls, a higher number of girls never play games than boys, etc. And secondly, you can compare the levels of ‘play games’ across a fixed value of gender, e.g. most boys play every day and very few play once a month or never.

1. Demo using the image of restaurant sales.

What was given as the data was not the day wise sales but the date-time wise sales, we extracted day from date-time column and then the data makes more sense. This is why Derived Variables can make more sense then the normal variable.

2. In the second example, by plotting the marks against the ‘month of birth’ (derived variable), it was observed that the children who were born after June would have missed the cutoff by a few days and gotten admission at the age of 5. The ones born after June (e.g. July, August, etc) were intellectually and emotionally more mature than their peers because of their higher age, resulting in better performance.

Types of Derived Metrics: Type Driven Metrics

- Type-driven metrics
- Business-driven metrics
- Data-driven metrics

Derived Metrics

INTERNSHIPSTUDIO

Type-Driven Metrics

These metrics can be derived by understanding the variable’s typology. You have already learnt one simple way of classifying variables/attributes — categorical (ordered, unordered) and quantitative or numeric. Similarly, there are various other ways of classification, one of which is Steven's typology.

Steven’s typology classifies variables into four types — nominal, ordinal, interval and ratio:

- Nominal variables: Categorical variables, where the categories differ only by their names; there is no order among categories, e.g. colour (red, blue, green), gender (male, female), department (HR, analytics, sales)
  - These are the most basic form of categorical variables
- Ordinal variables: Categories follow a certain order, but the mathematical difference between categories is not meaningful, e.g. education level (primary school, high school, college), height (high, medium, low), performance (bad, good, excellent), etc.
  - Ordinal variables are nominal as well

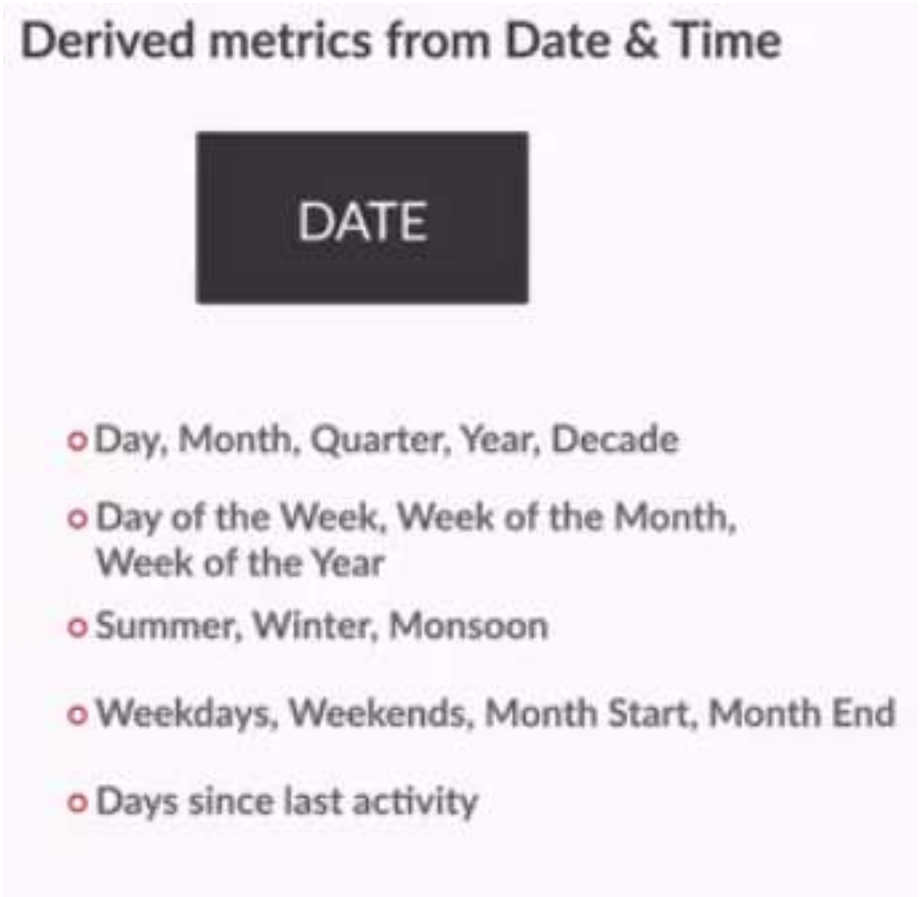
Derived Metrics

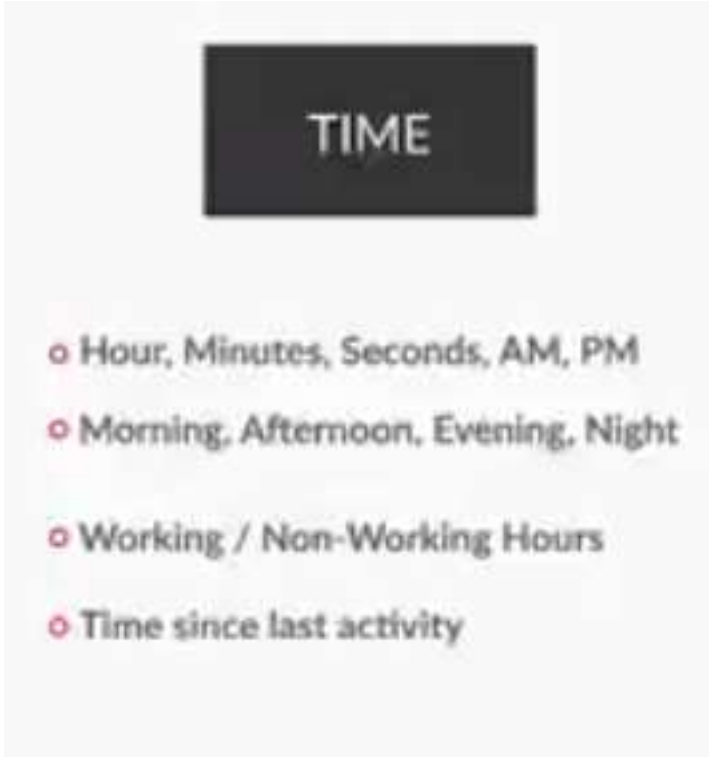
- **Interval variables:**
  - Categories follow a certain order, and the mathematical difference between categories is meaningful but division or multiplication is not, e.g. temperature in degrees celsius (the difference between 40 and 30 degrees C is meaningful, but 30 degrees x 40 degrees is not), dates (the difference between two dates is the number of days between them, but 25th May / 5th June is meaningless), etc.
  - Interval variables are both nominal and ordinal
- Ratio variables:
  - Apart from the mathematical difference, the ratio (division/multiplication) is possible, e.g. sales in dollars ($100 is twice $50), marks of students (50 is half of 100), etc.
  - Ratio variables are nominal, ordinal and interval type

For example, age in years is a ratio attribute, but you can convert it into an ordinal type by binning it into categories such as children (< 13 years), teenagers (13-19 years), young adults (20-25 years), etc. This enables you to ask questions, e.g. do teenagers do X better than children, are young adults more likely to do X than the other two types, etc. Here, X is an action you are interested in measuring.

Derived Metrics

- **Email Address**
- **URL**





Derived Metrics

- **Type Driven**

Customer feedback: Length of feedback

Derived Metrics

- **Business Driven**

Telecom example: Usage variable

INTERNSHIPSTUDIO

Derived Metrics

- **Data Driven**

Data-driven metrics can be created based on the variables present in the existing data set. For example, if you have two variables in your data set such as "weight" and "height" which shows a high correlation. So, instead of analysing "weight" and "height" variables separately, you can think of deriving a new metric "Body Mass Index (BMI)". Once you get the BMI, you can easily categorise people based on their fitness, e.g. a BMI below 18.5 should be considered as an underweight category, while BMI above 30.0 is considered as obese, by standard norms. This is how data-driven metrics can help you discover hidden patterns out of the data.