# Unit II

Basic Analysis Techniques : Statistical hypothesis generation and testing, Chi-Square test, t-Test, Analysis of variance, Correlation analysis, Maximum likelihood test.

# Statistical Inference

The field of statistical inference consists of those methods used to make decisions or draw conclusions about a **population.**

These methods utilize the information contained in a sample from the population in drawing conclusions.

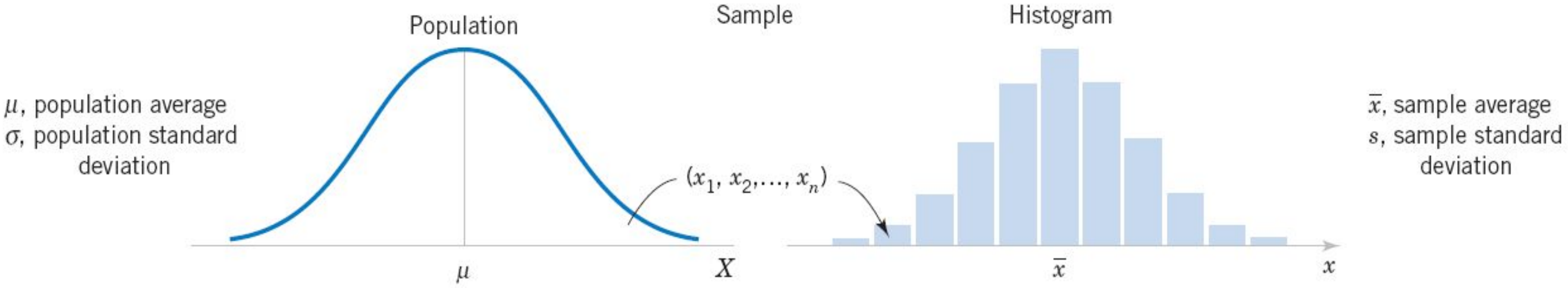

Population
Sample
Histogram
$\mu$, population average
$\sigma$, population standard deviation
$(x_1, x_2, \ldots, x_n)$
$\bar{x}$, sample average
$s$, sample standard deviation
$\mu$
$X$
$\bar{x}$
$x$

Figure 4-1 Relationship between a population and a sample.

# Statistical Inference

- In statistics, a **hypothesis** is a claim or statement about a property of a population.
- A **hypothesis test** (or **test of significance**) is a standard procedure for testing a claim about a property of a population.
- If, under a given assumption, the probability of a particular observed event is exceptionally small, one can conclude that the assumption is probably not correct.
- Hypothesis plays a crucial role in today’s data driven decision process , whether it may be making business decisions, in the health sector, academia, or in quality improvement.
- Without hypothesis & hypothesis tests, there is a risk of drawing the wrong conclusions and making bad decisions.

# Hypothesis Testing

- Hypothesis Testing is a type of statistical analysis in which assumptions are made about a population parameter to test.
- It is used to estimate the relationship between 2 statistical variables.

**Example:-**

- ✔ A teacher assumes that 60% of his college's students come from middle-class families.
- ✔ A doctor believes that 3D (Diet, Dose, and Discipline) is 90% effective for diabetic patients.
- ✔ A study says that 30% of customers are migrated to online shopping in last 3 years.

# Hypothesis Testing

- The **Null Hypothesis** is the assumption that the event will not occur. ***H0*** is the symbol for it, and it is pronounced *H-naught.*
- The **Alternate Hypothesis** is the logical opposite of the null hypothesis.
- The acceptance of the alternative hypothesis follows the rejection of the null hypothesis. ***H1*** is the symbol for it.
- Hypothesis testing is a statistical interpretation that examines a sample to determine whether the results stand true for the population.
- The test allows two explanations for the data—the null hypothesis or the alternative hypothesis. If the sample mean matches the population mean, the null hypothesis is proven true. Alternatively, if the sample mean is not equal to the population mean, the alternate hypothesis is accepted.

***A sanitizer manufacturer claims that its product kills 98 percent of germs on average.***

*To put this company's claim to the test, create a null and alternate hypothesis.*

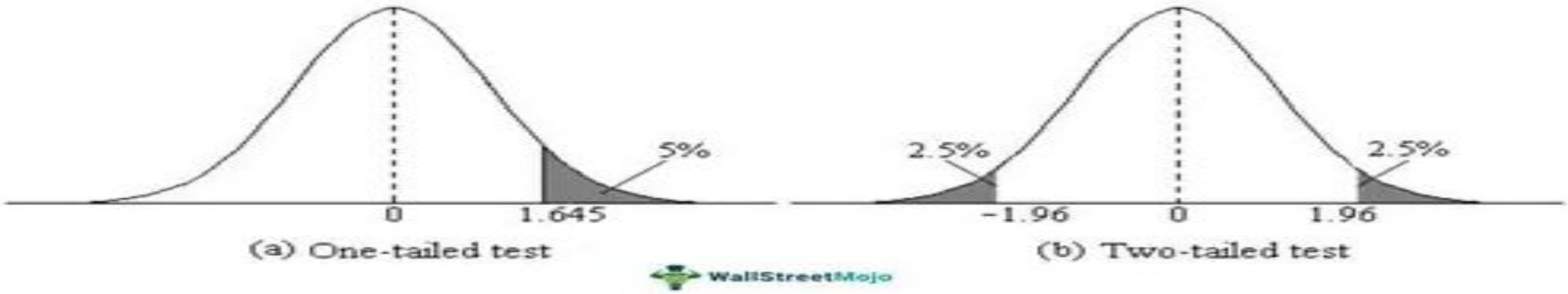
*H0 (Null Hypothesis): Average = 98%.*

*Alternative Hypothesis (H1): The average is less than 98%.*

# Hypothesis Testing Types

Based on population distribution, hypothesis testing is categorized into sub-types:

1. **Simple**: In a simple hypothesis, the population parameter is stated as a specific value, making the analysis easier. E.g.**A company is claiming that their average sales for this quarter are 1000 units.**
2. **Composite**: In a composite hypothesis, the population parameter ranges between a lower and upper value. E.g. **the company claims that the sales are in the range of 900 to 1000 units**
3. **One-tailed**: When the majority of the population is concentrated on one side, it is called a one-tailed test.
4. **Two-tailed**: The two-tailed hypothesis test works when the critical distribution of the population is two-sided.



# One-Tailed and Two-Tailed Hypothesis Testing

- The One-Tailed test, also called a **directional test**, considers a **critical region of data** that would result in the null hypothesis being rejected if the test sample falls into it, inevitably meaning the acceptance of the alternate hypothesis.
- In a one-tailed test, the critical distribution area is one-sided, meaning the test sample is either greater or lesser than a specific value.

- In two tails, the test sample is checked to be greater or less than a range of values in a Two-Tailed test, implying that the critical distribution area is two-sided.
- If the sample falls within this range, the alternate hypothesis will be accepted, and the null hypothesis will be rejected.

# Sample Hypothesis Testing Calculation

A hypothesis test for the average height of women in a particular country

Suppose null hypothesis is that the average height is 5'4".
sample size = 100 women and calculated average height is 5'5".
The standard deviation of population is 2.

Hypothesis Testing Formula

$Z = ( \bar{x} - \mu 0 ) / (\sigma / \sqrt{n})$

- Here, $\bar{x}$ is the sample mean,
- $\mu 0$ is the population mean,
- $\sigma$ is the standard deviation,
- n is the sample size.

z = (5'5" - 5'4") / (2" / √100)
z = 0.5 / (0.045)
z = 11.11

Null hypothesis will be rejected as the z-score of 11.11 is very large and conclude that there is evidence to suggest that the average height of women in the country is greater than 5'4".

# Sample Hypothesis Testing Calculation

A battery manufacturing company claims that the average life of its two-wheeler batteries is 2.1 years. The quality inspector surveyed ten customers to know the lasting period of their batteries. The following data was collected:

the **standard deviation** is 0.17 and the significance level is 0.05, conduct a hypothesis testing to prove the company’s claim.

| Customer No. | Battery Life (in years) |
|---|---|
| 1 | 1.9 |
| 2 | 2.3 |
| 3 | 2.1 |
| 4 | 2.2 |
| 5 | 1.9 |
| 6 | 2.4 |
| 7 | 2.1 |
| 8 | 2.3 |
| 9 | 2.2 |
| 10 | 2.0 |

Given:

$\boldsymbol{\mu_0}$= 2.1 years

$\boldsymbol{\sigma}$ = 0.17

**n** = 10

Level of Significance = 0.05

Assuming that the company's claim of average battery life being 2.1 years is true, We need to prove that:

**$H_0$: $\mu=\mu_0$ , or**<br>
**$H_1$: $\mu\neq\mu_0$**

Sample mean ($\bar{x}$) = (1.9 + 2.3 + 2.1 + 2.2 + 1.9 + 2.4 + 2.1 + 2.3 + 2.2 + 2.0) / 10 = 2.14 years.

Applying the Z-test formula:

**$Z = ( \bar{x} - \mu_0 ) / (\sigma /\sqrt{n})$**

$Z = (2.14 - 2.1) / (0.17 / \sqrt{10}) = 0.744$

- The Z test and the Z score are related but different concepts in statistics. **A Z test** is a statistical hypothesis test that compares the mean of a sample with a known population mean, using the standard deviation of the population. The Z test is used to determine whether the sample mean is significantly different from the population mean.

- A **Z score**, also known as a standard score, is a measure of how many standard deviations away from the mean a particular data point is. It is calculated by subtracting the population mean from an individual data point, and then dividing the result by the population standard deviation. In summary, a Z test is a statistical test used to determine if a sample mean is different from a population mean, while a Z score is a measure of how far away from the mean an individual data point is.

**If the value of z is greater than 1.96 or less than -1.96, the null hypothesis is rejected otherwewise null hypothesis accepted.**

Thus, the company's claim that the average life of its batteries is 2.1 years is proven true.

# Steps of Hypothesis Testing

**Step 1: Specify Null and Alternate Hypotheses**

**Step 2: Gather Data**

**Step 3: Conduct a Statistical Test**

**Step 4: Determine Rejection Of Your Null Hypothesis**

**Step 5: Present Your Results**

# Types of Hypothesis Testing

## Z Test:-

To determine whether a discovery or relationship is statistically significant, hypothesis testing uses a z-test. It usually checks to see if two means are the same (the null hypothesis). Only when the population standard deviation is known and the sample size is 30 data points or more, can a z-test be applied.

## T Test:-

A statistical test called a t-test is employed to compare the means of two groups. To determine whether two groups differ or if a procedure or treatment affects the population of interest, it is frequently used in hypothesis testing.

## Chi-Square :-

You utilize a Chi-square test for hypothesis testing concerning whether your data is as predicted. To determine if the expected and observed results are well-fitted, the Chi-square test analyzes the differences between categorical variables from a random sample.

# T Test

- A ***t* test** is a statistical test that is used to compare the means of two groups.
- It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.

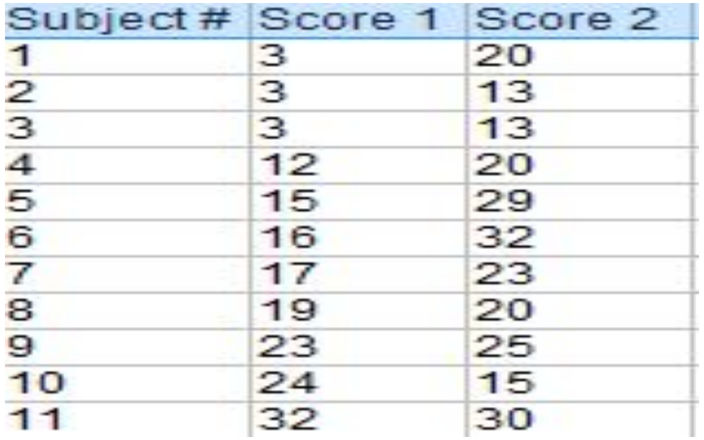
***t test example :-***

- *You want to know whether the **mean petal length** of iris flowers differs according to their species.*
- *You find two different species of irises growing in a garden and measure 25 petals of each species.*

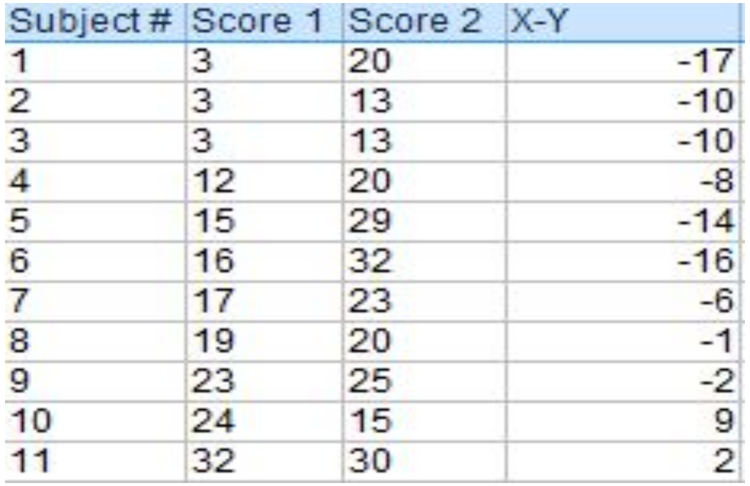
*you can test the difference between these two groups using a t test and null and alterative hypotheses.*

1. *The null hypothesis ($H_0$) is that the true difference between these group means is zero.*
2. *The alternate hypothesis ($H_a$) is that the true difference is different from zero.*

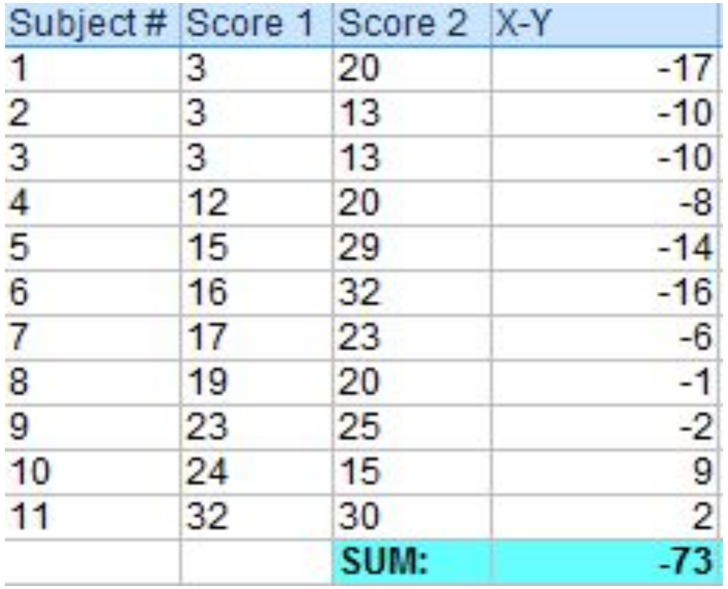
# T test



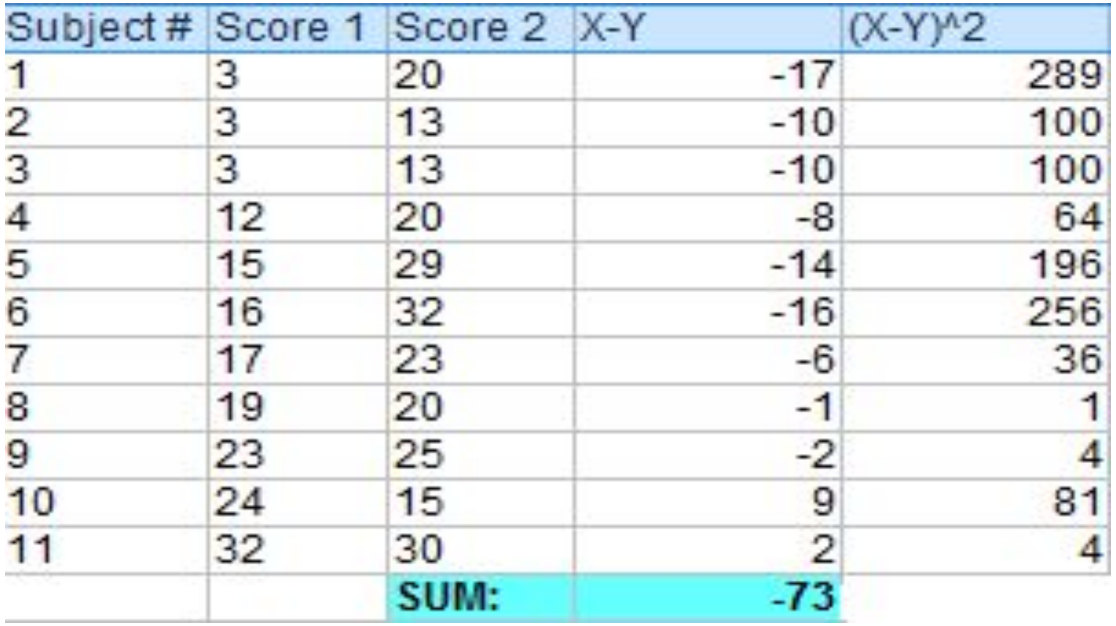
| Subject # | Score 1 | Score 2 |
|---|---|---|
| 1 | 3 | 20 |
| 2 | 3 | 13 |
| 3 | 3 | 13 |
| 4 | 12 | 20 |
| 5 | 15 | 29 |
| 6 | 16 | 32 |
| 7 | 17 | 23 |
| 8 | 19 | 20 |
| 9 | 23 | 25 |
| 10 | 24 | 15 |
| 11 | 32 | 30 |



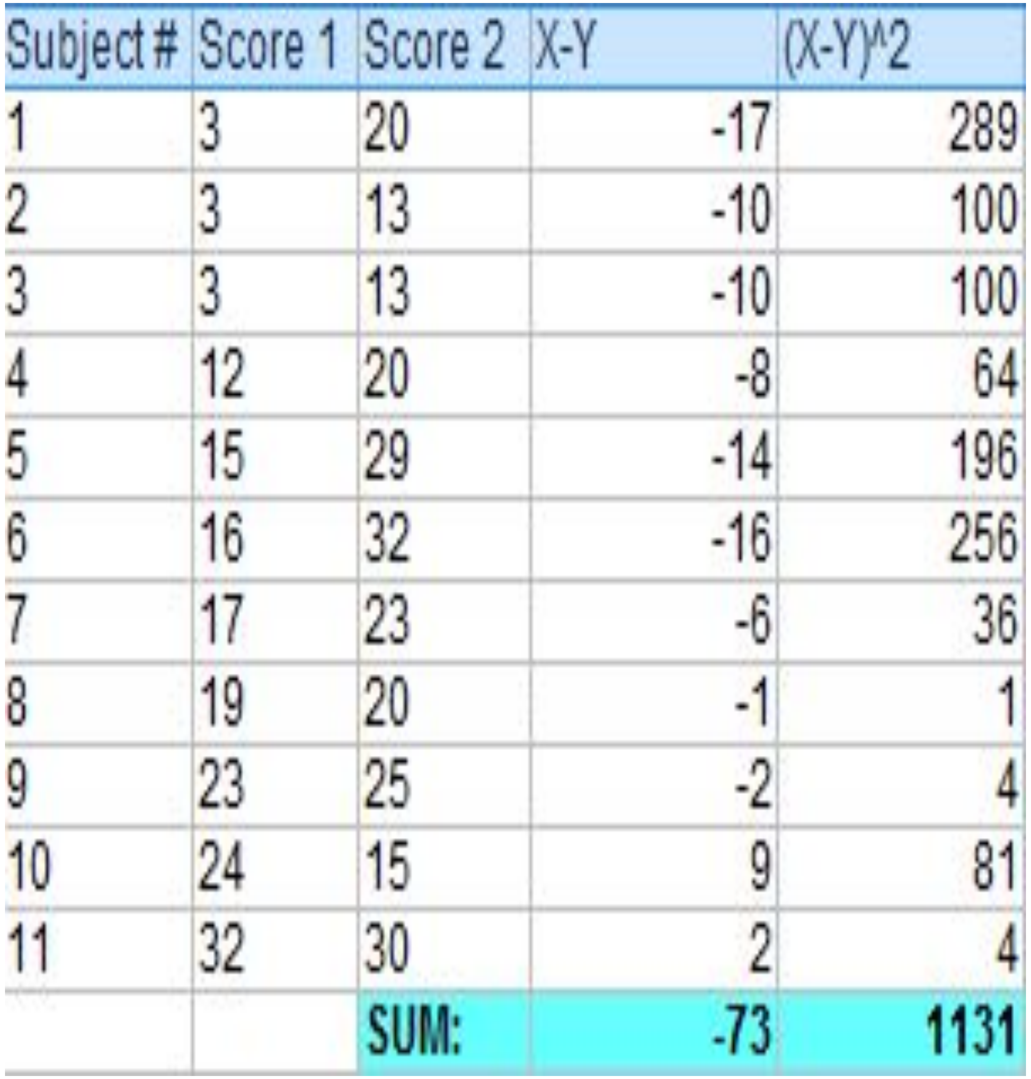
| Subject # | Score 1 | Score 2 | X-Y |
|---|---|---|---|
| 1 | 3 | 20 | -17 |
| 2 | 3 | 13 | -10 |
| 3 | 3 | 13 | -10 |
| 4 | 12 | 20 | -8 |
| 5 | 15 | 29 | -14 |
| 6 | 16 | 32 | -16 |
| 7 | 17 | 23 | -6 |
| 8 | 19 | 20 | -1 |
| 9 | 23 | 25 | -2 |
| 10 | 24 | 15 | 9 |
| 11 | 32 | 30 | 2 |



| Subject # | Score 1 | Score 2 | X-Y |
|---|---|---|---|
| 1 | 3 | 20 | -17 |
| 2 | 3 | 13 | -10 |
| 3 | 3 | 13 | -10 |
| 4 | 12 | 20 | -8 |
| 5 | 15 | 29 | -14 |
| 6 | 16 | 32 | -16 |
| 7 | 17 | 23 | -6 |
| 8 | 19 | 20 | -1 |
| 9 | 23 | 25 | -2 |
| 10 | 24 | 15 | 9 |
| 11 | 32 | 30 | 2 |
|  |  | **SUM:** | **-73** |



| Subject # | Score 1 | Score 2 | X-Y | (X-Y)^2 |
|---|---|---|---|---|
| 1 | 3 | 20 | -17 | 289 |
| 2 | 3 | 13 | -10 | 100 |
| 3 | 3 | 13 | -10 | 100 |
| 4 | 12 | 20 | -8 | 64 |
| 5 | 15 | 29 | -14 | 196 |
| 6 | 16 | 32 | -16 | 256 |
| 7 | 17 | 23 | -6 | 36 |
| 8 | 19 | 20 | -1 | 1 |
| 9 | 23 | 25 | -2 | 4 |
| 10 | 24 | 15 | 9 | 81 |
| 11 | 32 | 30 | 2 | 4 |
|  |  | **SUM:** | **-73** |  |



| Subject # | Score 1 | Score 2 | X-Y | (X-Y)^2 |
|---|---|---|---|---|
| 1 | 3 | 20 | -17 | 289 |
| 2 | 3 | 13 | -10 | 100 |
| 3 | 3 | 13 | -10 | 100 |
| 4 | 12 | 20 | -8 | 64 |
| 5 | 15 | 29 | -14 | 196 |
| 6 | 16 | 32 | -16 | 256 |
| 7 | 17 | 23 | -6 | 36 |
| 8 | 19 | 20 | -1 | 1 |
| 9 | 23 | 25 | -2 | 4 |
| 10 | 24 | 15 | 9 | 81 |
| 11 | 32 | 30 | 2 | 4 |
|  |  | SUM: | -73 | 1131 |

1. The “ΣD” is the sum of X-Y from Step 2.
2. $\Sigma D^2$: Sum of the squared differences
3. $(\Sigma D)^2$: Sum of the differences squared.

$$t = \frac{(\sum D)/N}{\sqrt{\dfrac{\sum D^2 - \left(\dfrac{(\sum D)^2}{N}\right)}{(N-1)(N)}}}$$

$$t = \frac{(\Sigma D)/N}{\sqrt{\frac{\Sigma D^2 - \left(\frac{(\Sigma D)^2}{N}\right)}{(N-1)(N)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \left(\frac{(-73)^2}{11}\right)}{(11-1)(11)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \left(\frac{5329}{11}\right)}{110}}}$$

$$t = -\ 2.74$$

Subtract 1 from the sample size to get the degrees of freedom. Total 11 items. So 11 – 1 = 10.

Find the p-value in the t-table, using the degrees of freedom in But if you don't have a specified alpha level, use 0.05 (5%).

So for this example t test problem, with df = 10, the t-value is 2.228.

**If the absolute value of the t-value is greater than the critical value, reject the null hypothesis.**

**If the absolute value of the t-value is less than the critical value, fail to reject the null hypothesis.**

- In conclusion, compare t-table value from (2.228) to calculated t-value (-2.74).
- The calculated t-value is greater than the table value at an alpha level of .05. **p value=0.0175 considering one tail**
- In addition, note that the p-value is less than the alpha level: p <.05. So we can reject the null hypothesis that there is no difference between means.
- However, note that you can ignore the minus sign when comparing the two t-values as ± indicates the direction;
- the p-value remains the same for both directions.

# Chi-Square Test

- The Chi-Square test is a statistical procedure for determining the difference between observed and expected data. The Observed values are those you gather yourselves. The expected values are the frequencies expected, based on the null hypothesis.

- It helps to find out whether a difference between two categorical variables is due to chance or a relationship between them.

- The degrees of freedom in a statistical calculation represent the number of variables that can vary in a calculation.

$$x_c^2 = \frac{\Sigma\,(O_i - E_i)^2}{E_i}$$

Where
c = Degrees of freedom
O = Observed Value
E = Expected Value

There are two main types of Chi-Square tests namely –

1. Independence :-

*For Example-*
*In a movie theatre, suppose we made a list of movie genres. Let us consider this as the first variable. The second variable is whether or not the people who came to watch those genres of movies have bought snacks at the theatre. Here the null hypothesis is that the genre of the film and whether people bought snacks or not are unrelatable. If this is true, the movie genres don't impact snack sales.*

2. Goodness-of-Fit :-

*For Example-*
*Suppose we have bags of balls with five different colours in each bag. The given condition is that the bag should contain an equal number of balls of each colour. The idea we would like to test here is that the proportions of the five colours of balls in each bag must be exact.*

# Chi square Example

We want to know if gender has anything to do with political party preference. You poll 440 voters in a simple random sample to find out which political party they prefer. The results of the survey are shown in the table .

| | Republican | Democrat | Independent | Total |
|---|---|---|---|---|
| Male | 100 | 70 | 30 | 200 |
| Female | 140 | 60 | 20 | 220 |
| Total | 240 | 130 | 50 | 420 |

$$= \frac{(240) * (200)}{440} = 109$$

**Step 1: Define the Hypothesis**

H0: There is no link between gender and political party preference.

H1: There is a link between gender and political party preference.

**Step 2: Calculate the Expected Values**

$$\text{Expected Value} = \frac{(Row\ Total) * (Column\ Total)}{Total\ Number\ Of\ Observations}$$

**For example, the expected value for Male Republicans is:**

200*240/420 = 114.28
130*200/420=61.90
50*200/420= 23.80
240*220/420=125.71
130*220/420= 68.09
50*220/420 = 26.19

## Step 3: Calculate (O-E)2 / E for Each Cell in the Table

calculate the (O - E)2 / E for each cell in the table.
Where O = Observed Value ,E = Expected Value

| | Republican | Democrat | Independent | Total |
|---|---|---|---|---|
| Male | 114.28 | 61.90 | 23.80 | 200 |
| Female | 125.71 | 68.09 | 26.19 | 220 |
| Total | 240 | 130 | 50 | 420 |

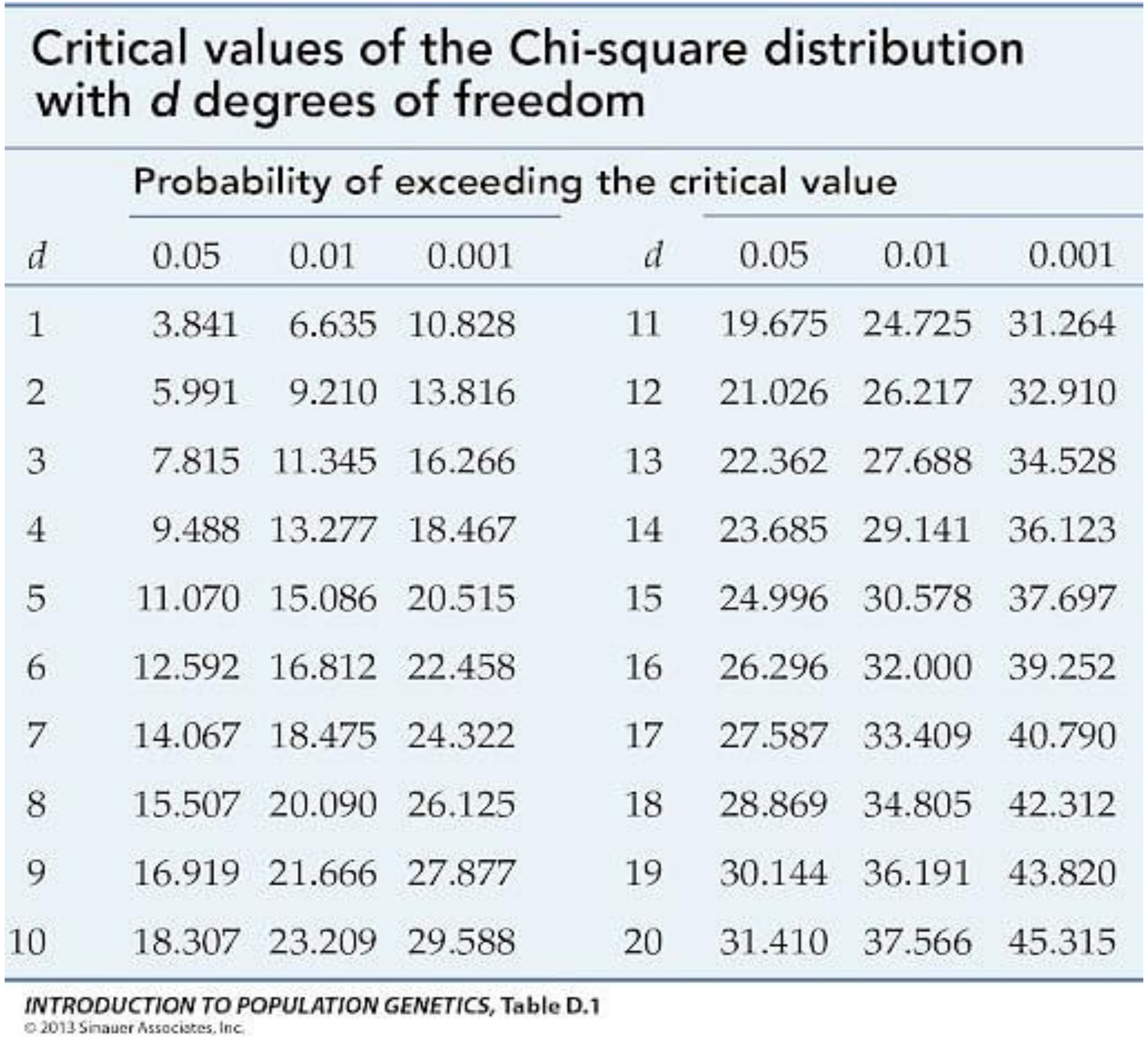
⍰Expected Value

| | Republican | Democrat | Independent | Total |
|---|---|---|---|---|
| Male | 1.783 | 1.059 | 1.281 | 200 |
| Female | 1.458 | 1.09 | 1.91 | 220 |
| Total | 240 | 130 | 50 | 420 |

## Step 4: Calculate the Test Statistic X2

$X^2$ is the sum of all the values in the last table
= 1.783+1.059+1.281+1.458+1.09+1.91
= **8.581**

*The degrees of freedom for this example is equal to the table's number of columns minus one multiplied by the table's number of rows minus one, or (r-1) (c-1). We have (3-1)(2-1) = 2.*



Critical values of the Chi-square distribution with *d* degrees of freedom

Probability of exceeding the critical value

| *d* | 0.05 | 0.01 | 0.001 | *d* | 0.05 | 0.01 | 0.001 |
|---|---|---|---|---|---|---|---|
| 1 | 3.841 | 6.635 | 10.828 | 11 | 19.675 | 24.725 | 31.264 |
| 2 | 5.991 | 9.210 | 13.816 | 12 | 21.026 | 26.217 | 32.910 |
| 3 | 7.815 | 11.345 | 16.266 | 13 | 22.362 | 27.688 | 34.528 |
| 4 | 9.488 | 13.277 | 18.467 | 14 | 23.685 | 29.141 | 36.123 |
| 5 | 11.070 | 15.086 | 20.515 | 15 | 24.996 | 30.578 | 37.697 |
| 6 | 12.592 | 16.812 | 22.458 | 16 | 26.296 | 32.000 | 39.252 |
| 7 | 14.067 | 18.475 | 24.322 | 17 | 27.587 | 33.409 | 40.790 |
| 8 | 15.507 | 20.090 | 26.125 | 18 | 28.869 | 34.805 | 42.312 |
| 9 | 16.919 | 21.666 | 27.877 | 19 | 30.144 | 36.191 | 43.820 |
| 10 | 18.307 | 23.209 | 29.588 | 20 | 31.410 | 37.566 | 45.315 |

*INTRODUCTION TO POPULATION GENETICS*, Table D.1

© 2013 Sinauer Associates, Inc.

Finally, compare obtained statistic to the critical statistic found in the chi-square table.

As can be seen, for an alpha level of 0.05 and two degrees of freedom, the critical statistic is 5.991, which is less than obtained statistic of 8.581.

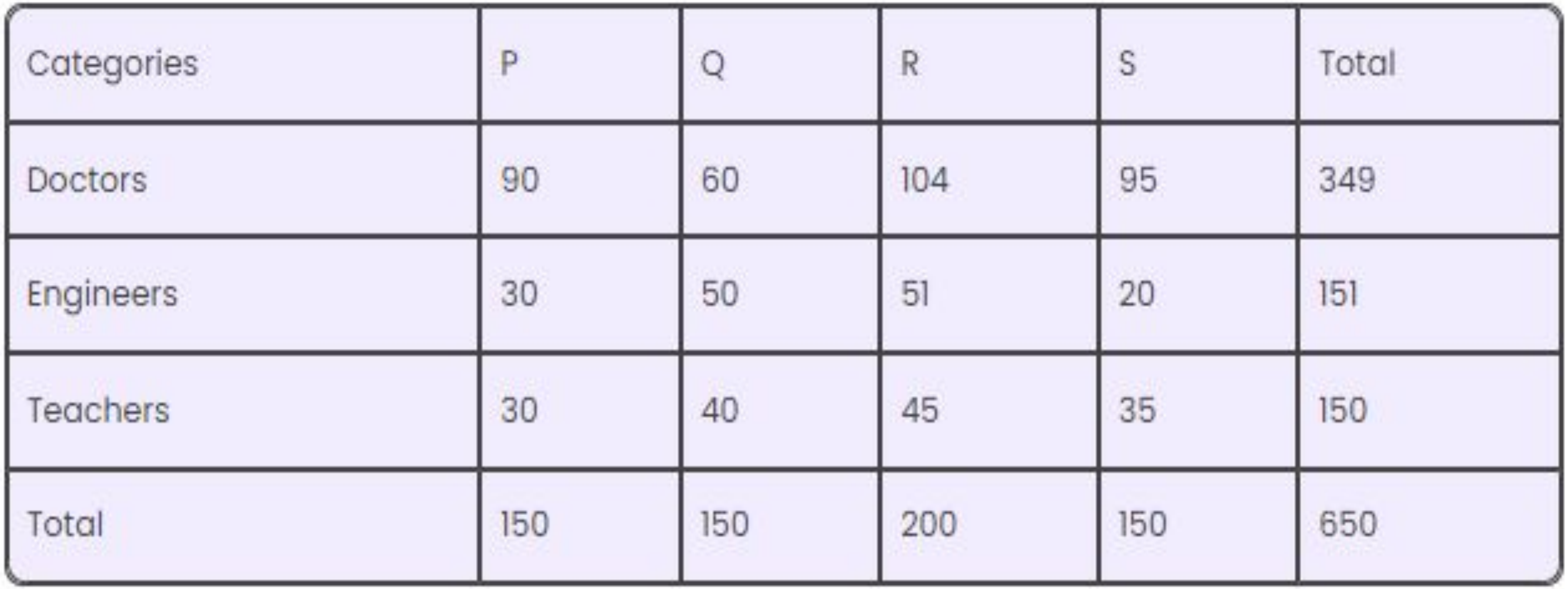
**If your chi-square calculated value is greater than the chi-square critical value, then you reject your null hypothesis.**

**If your chi-square calculated value is less than the chi-square critical value, then you "fail to reject" your null hypothesis.**

This means you have sufficient evidence to say that there is an association between gender and political party preference.

Take an example of a categorical data where there is a society of 1000 residents with four neighbourhoods, P, Q, R and S. A random sample of 650 residents of the society is taken whose occupations are doctors, engineers and teachers. The null hypothesis is that

**each person’s neighbourhood of residency is independent of the person’s professional division.** The data are categorised as:



| Categories | P | Q | R | S | Total |
|---|---|---|---|---|---|
| Doctors | 90 | 60 | 104 | 95 | 349 |
| Engineers | 30 | 50 | 51 | 20 | 151 |
| Teachers | 30 | 40 | 45 | 35 | 150 |
| Total | 150 | 150 | 200 | 150 | 650 |

# What is Analysis of Variance (ANOVA)

- ANOVA is to test for differences among the means of the population by examining the amount of variation within each sample, relative to the amount of variation between the samples.

- ***Analyzing variance tests the hypothesis*** that the means of two or more populations are equal.

- In a regression study, analysts use the ANOVA test to determine the impact of independent variables on the dependent variable.

# The steps to perform the ANOVA test :

**Step 1:** Calculate the mean for each group.

**Step 2:** Calculate the total mean. This is done by adding all the means and dividing it by the total number of means.

**Step 3:** Calculate the SSB. **SSB= Sum of squares between groups**

**Step 4:** Calculate the between groups degrees of freedom.

Degrees of freedom between groups, $df_1 = k - 1$. Here, k denotes the number of groups.

Degrees of freedom of errors, $df_2 = N - k$, where N denotes the total number of observations across k groups.

Total degrees of freedom, $df_3 = N - 1$.

**Step 5:** Calculate the SSE. **Sum of squares of errors**

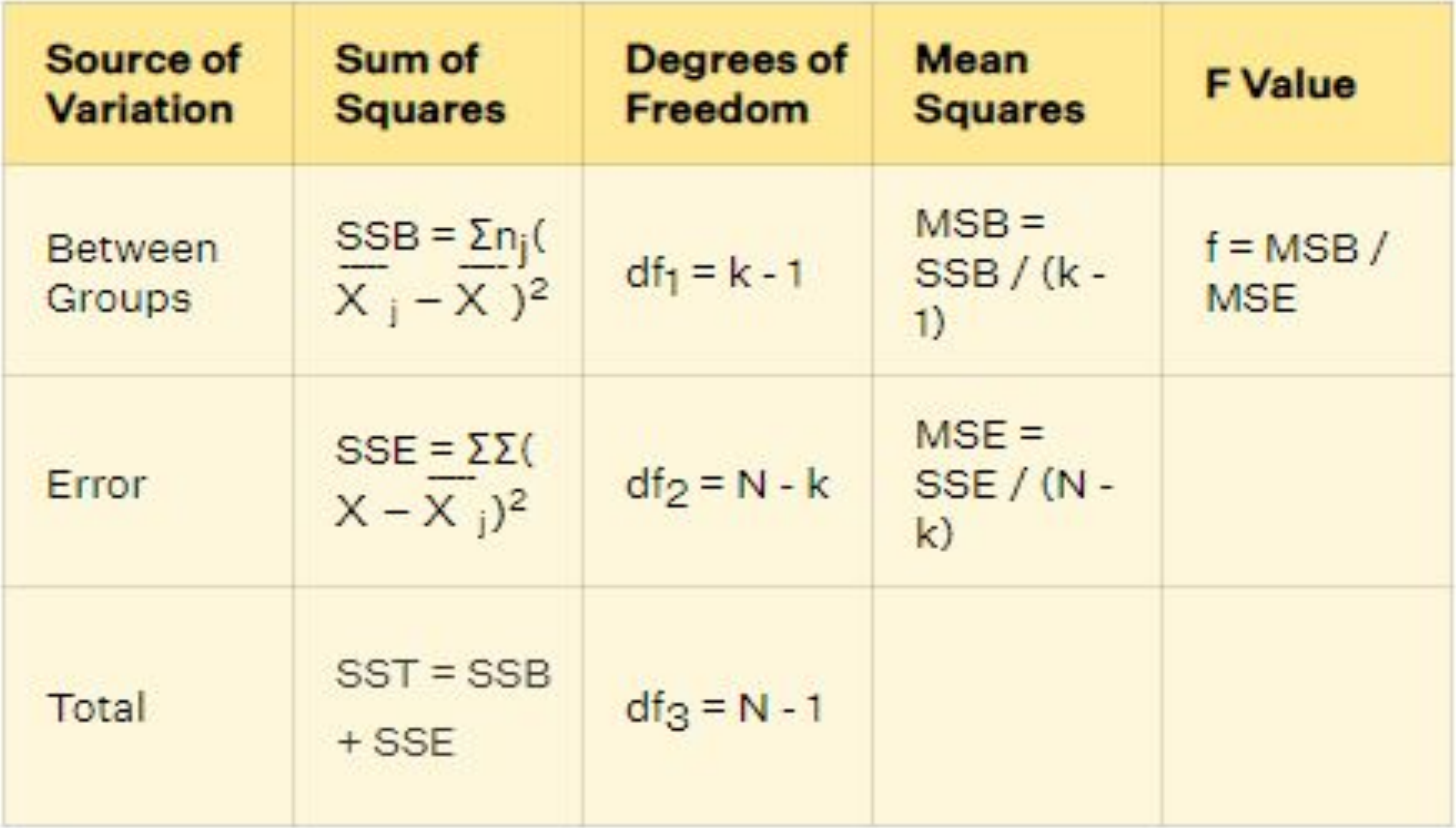
**Step 6:** Calculate the degrees of freedom of errors. **Total sum of squares, SST = SSB + SSE**

**Step 7:** Determine the MSB and the MSE.

**Step 8:** Find the f test statistic.

**Step 9:** Using the f table for the specified level of significance, α, find the critical value. This is given by $F(\alpha, df_1. df_2)$.

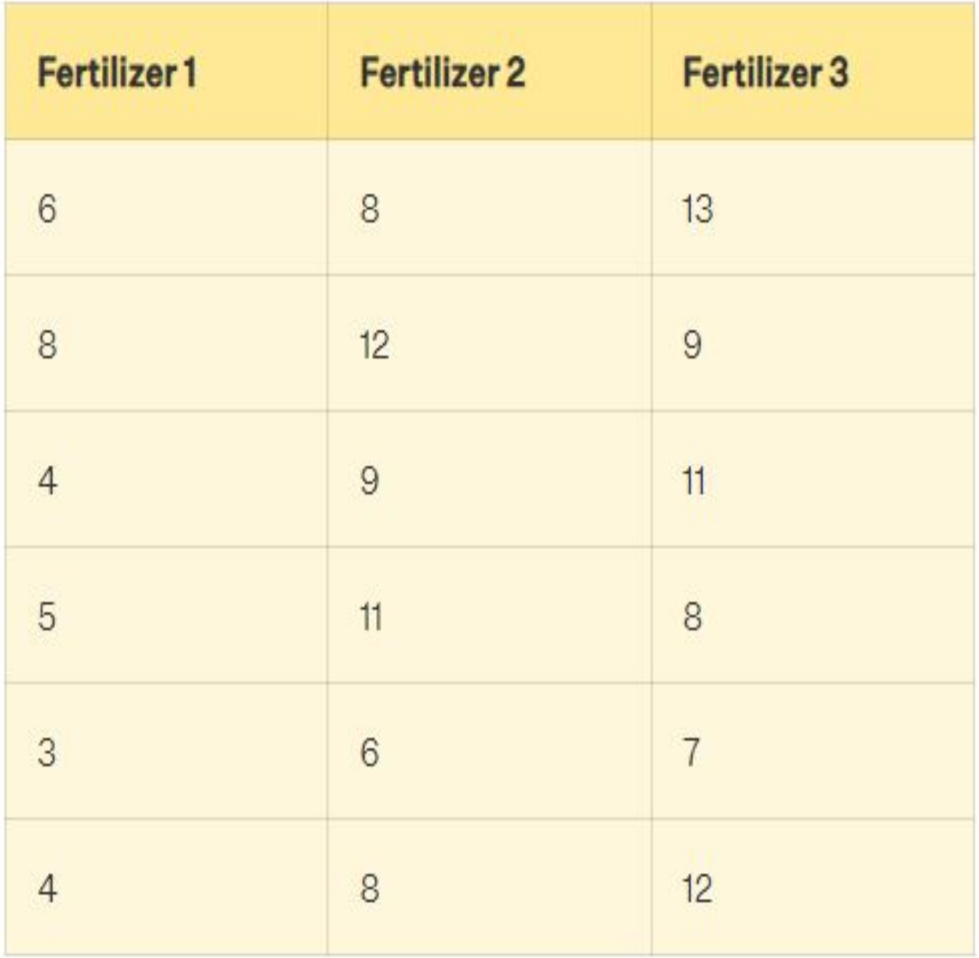
**Step 10:** If $f > F$ then reject the null hypothesis.



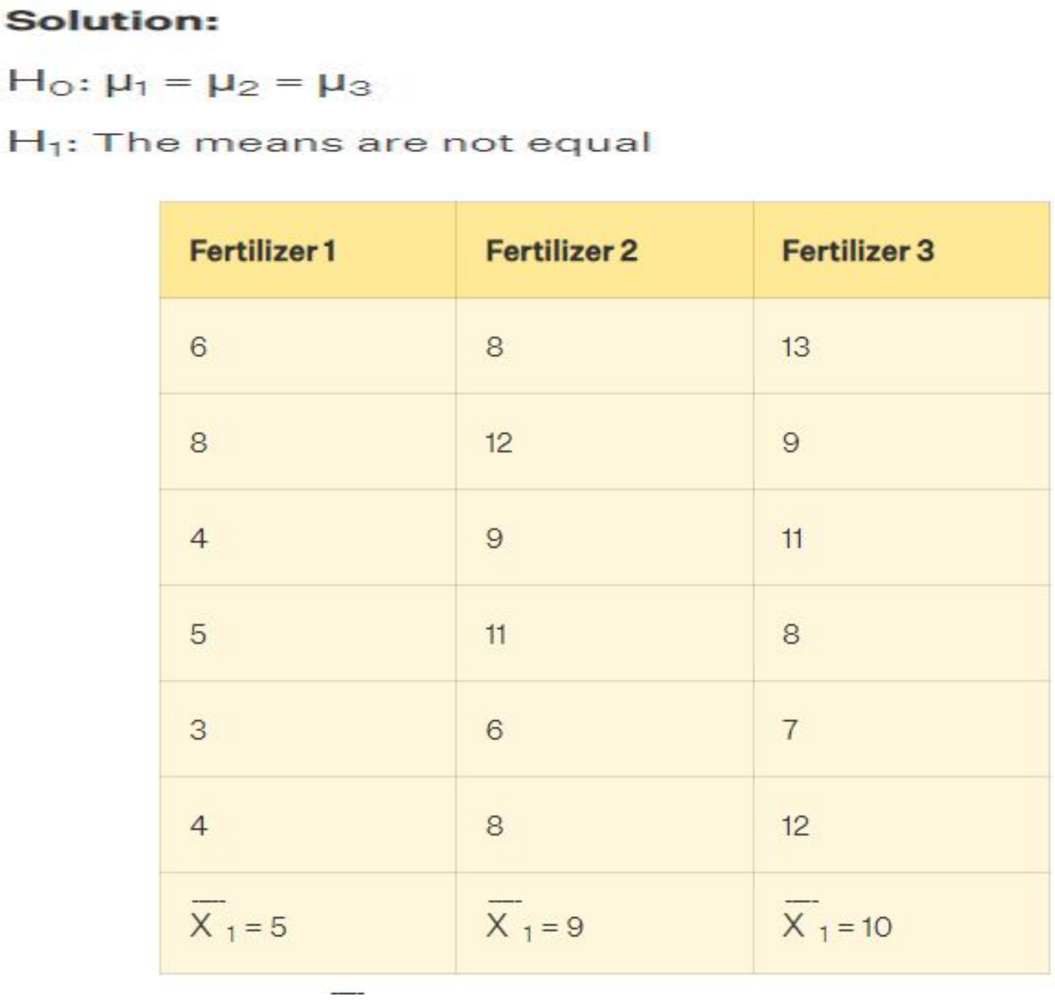
| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Squares | F Value |
|---|---|---|---|---|
| Between Groups | $SSB = \Sigma n_j(\overline{X}_j - \overline{X})^2$ | $df_1 = k - 1$ | $MSB = SSB / (k - 1)$ | $f = MSB / MSE$ |
| Error | $SSE = \Sigma\Sigma(X - \overline{X}_j)^2$ | $df_2 = N - k$ | $MSE = SSE / (N - k)$ | |
| Total | $SST = SSB + SSE$ | $df_3 = N - 1$ | | |

# Examples on ANOVA Test

Three types of fertilizers are used on three groups of plants for 5 weeks. We want to check if there is a difference in the mean growth of each group. Using the data given below apply a one way ANOVA test at 0.05 significant level.



| Fertilizer 1 | Fertilizer 2 | Fertilizer 3 |
|---|---|---|
| 6 | 8 | 13 |
| 8 | 12 | 9 |
| 4 | 9 | 11 |
| 5 | 11 | 8 |
| 3 | 6 | 7 |
| 4 | 8 | 12 |



**Solution:**

$H_O: \mu_1 = \mu_2 = \mu_3$

$H_1$: The means are not equal

| Fertilizer 1 | Fertilizer 2 | Fertilizer 3 |
|---|---|---|
| 6 | 8 | 13 |
| 8 | 12 | 9 |
| 4 | 9 | 11 |
| 5 | 11 | 8 |
| 3 | 6 | 7 |
| 4 | 8 | 12 |
| $\overline{X}_1 = 5$ | $\overline{X}_1 = 9$ | $\overline{X}_1 = 10$ |

Total mean, $\overline{X} = 8$

$n_1 = n_2 = n_3 = 6, k = 3$

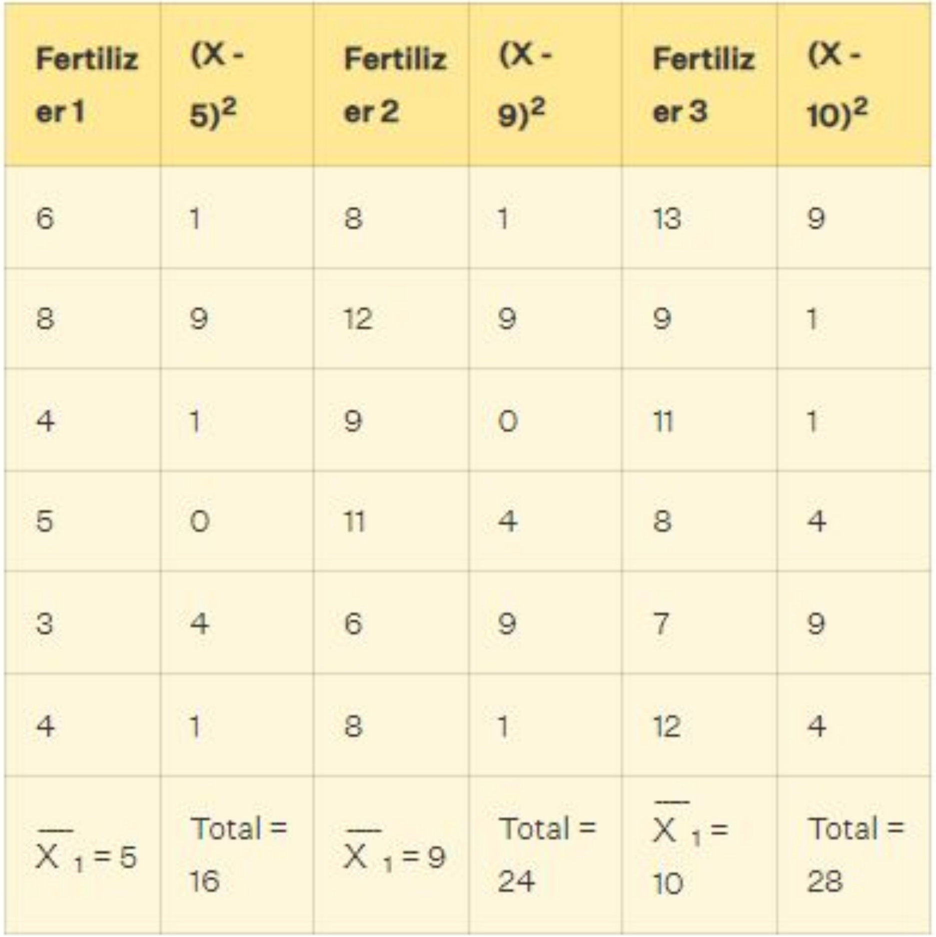
$SSB = 6(5-8)^2 + 6(9-8)^2 + 6(10-8)^2$

$= 84$

$df_1 = k - 1 = 2$

**SSB= Sum of squares between groups**

**SSE= Sum of squares of errors**



| Fertilizer 1 | $(X - 5)^2$ | Fertilizer 2 | $(X - 9)^2$ | Fertilizer 3 | $(X - 10)^2$ |
|---|---|---|---|---|---|
| 6 | 1 | 8 | 1 | 13 | 9 |
| 8 | 9 | 12 | 9 | 9 | 1 |
| 4 | 1 | 9 | 0 | 11 | 1 |
| 5 | 0 | 11 | 4 | 8 | 4 |
| 3 | 4 | 6 | 9 | 7 | 9 |
| 4 | 1 | 8 | 1 | 12 | 4 |
| $\overline{X}_1 = 5$ | Total = 16 | $\overline{X}_1 = 9$ | Total = 24 | $\overline{X}_1 = 10$ | Total = 28 |

df1= k-1=3-1=2
$df_2$ = N - k = 18 - 3 = 15
$df_3$ = N - 1 = 18 - 1 = 17

SSE = 16 + 24 + 28 = 68

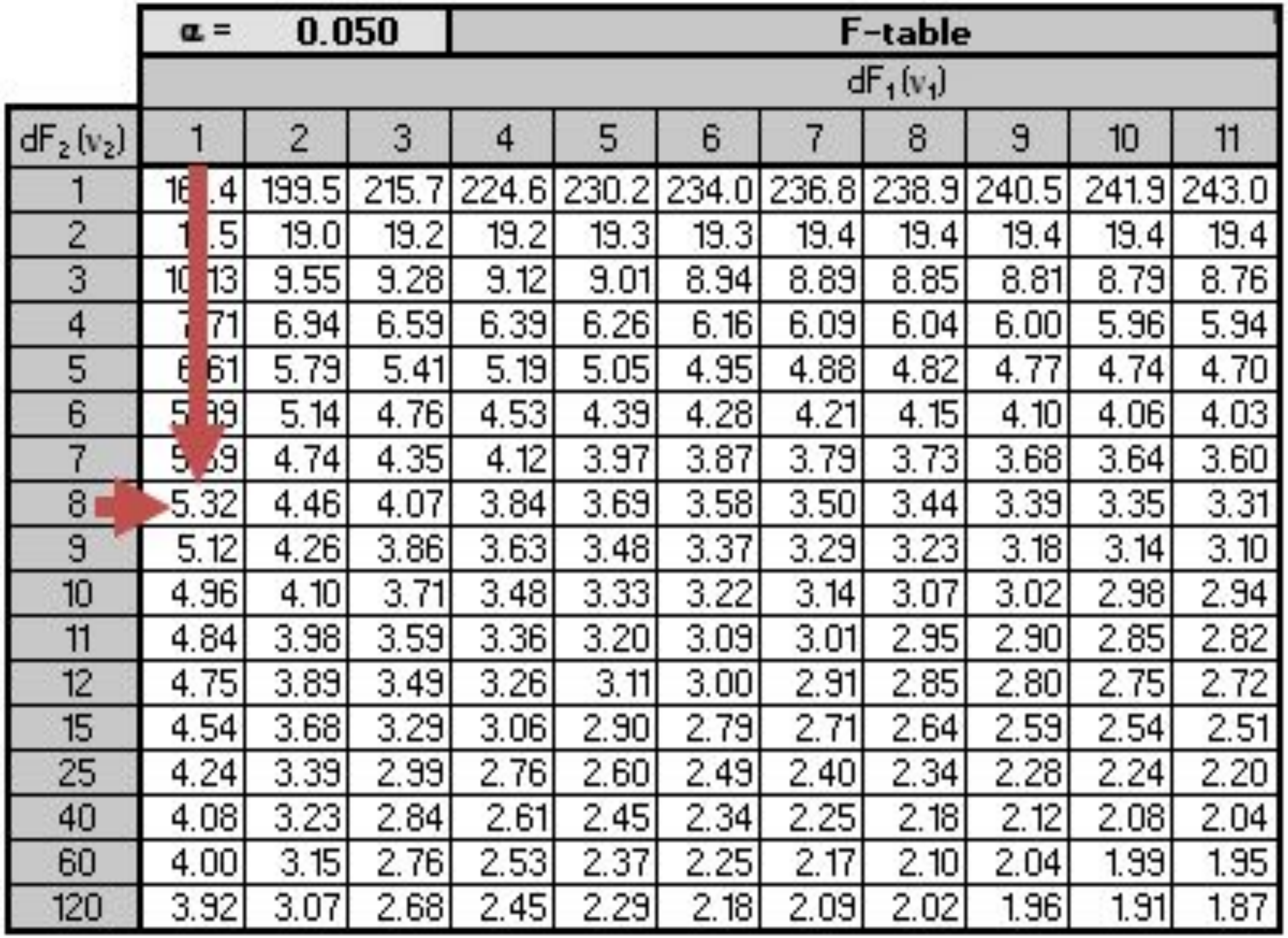
SST= SSB + SSE= 84 + 68= 152

MSB = SSB / $df_1$ = 84 / 2 = 42
MSE = SSE / $df_2$ = 68 / 15 = 4.53

ANOVA test statistic, f = MSB / MSE = 42 / 4.53 = 9.33

Using the f table at α = 0.05 the critical value is given as F(0.05, 2, 15) = 3.68
As f > F, thus, the null hypothesis is rejected and it can be concluded that there is a difference in the mean growth of the plants.

**Answer: Reject the null hypothesis**



| α = | 0.050 | F-table | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | dF$_1$ ($v_1$) | | | | | | | | | | |
| dF$_2$ ($v_2$) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 1 | 161.4 | 199.5 | 215.7 | 224.6 | 230.2 | 234.0 | 236.8 | 238.9 | 240.5 | 241.9 | 243.0 |
| 2 | 18.5 | 19.0 | 19.2 | 19.2 | 19.3 | 19.3 | 19.4 | 19.4 | 19.4 | 19.4 | 19.4 |
| 3 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 | 8.76 |
| 4 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 | 5.94 |
| 5 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 | 4.70 |
| 6 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 | 4.03 |
| 7 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 | 3.60 |
| 8 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 | 3.31 |
| 9 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 | 3.10 |
| 10 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 | 2.94 |
| 11 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 | 2.82 |
| 12 | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 | 2.72 |
| 15 | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 | 2.51 |
| 25 | 4.24 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.28 | 2.24 | 2.20 |
| 40 | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 | 2.08 | 2.04 |
| 60 | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 | 1.99 | 1.95 |
| 120 | 3.92 | 3.07 | 2.68 | 2.45 | 2.29 | 2.18 | 2.09 | 2.02 | 1.96 | 1.91 | 1.87 |

# Correlation analysis

- The most commonly used techniques for investigating the relationship between two quantitative variables are **correlation** and **linear regression.**

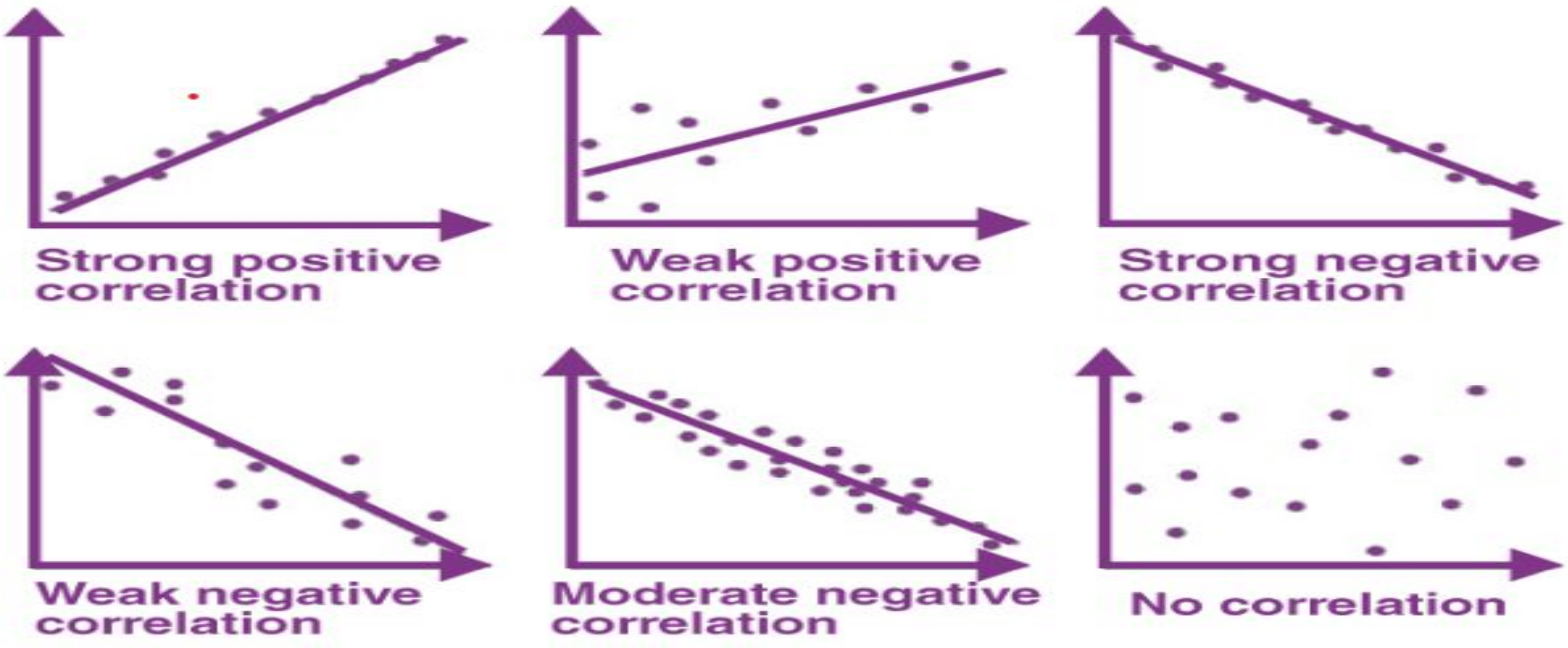
- **Correlation** quantifies the strength of the linear relationship between a pair of variables, whereas **regression** expresses the relationship in the form of an equation.

- Correlation analysis, also known as bivariate, is primarily concerned with finding out whether a relationship exists between variables and then determining the magnitude and action of that relationship.

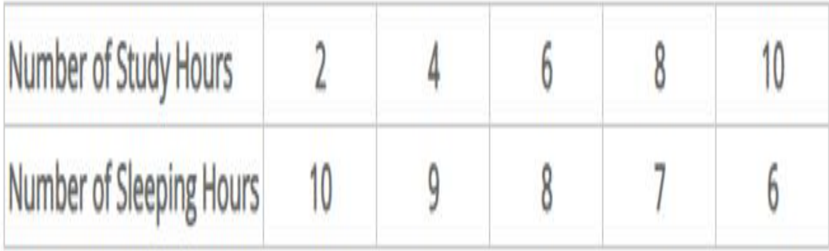
# Correlation analysis

There are three basic types of correlation:

1. **Positive correlation**: the two variables change in the same direction.
2. **Negative correlation**: the two variables change in opposite directions.
3. **No correlation:** there is no association or relevant relationship between the two variables.

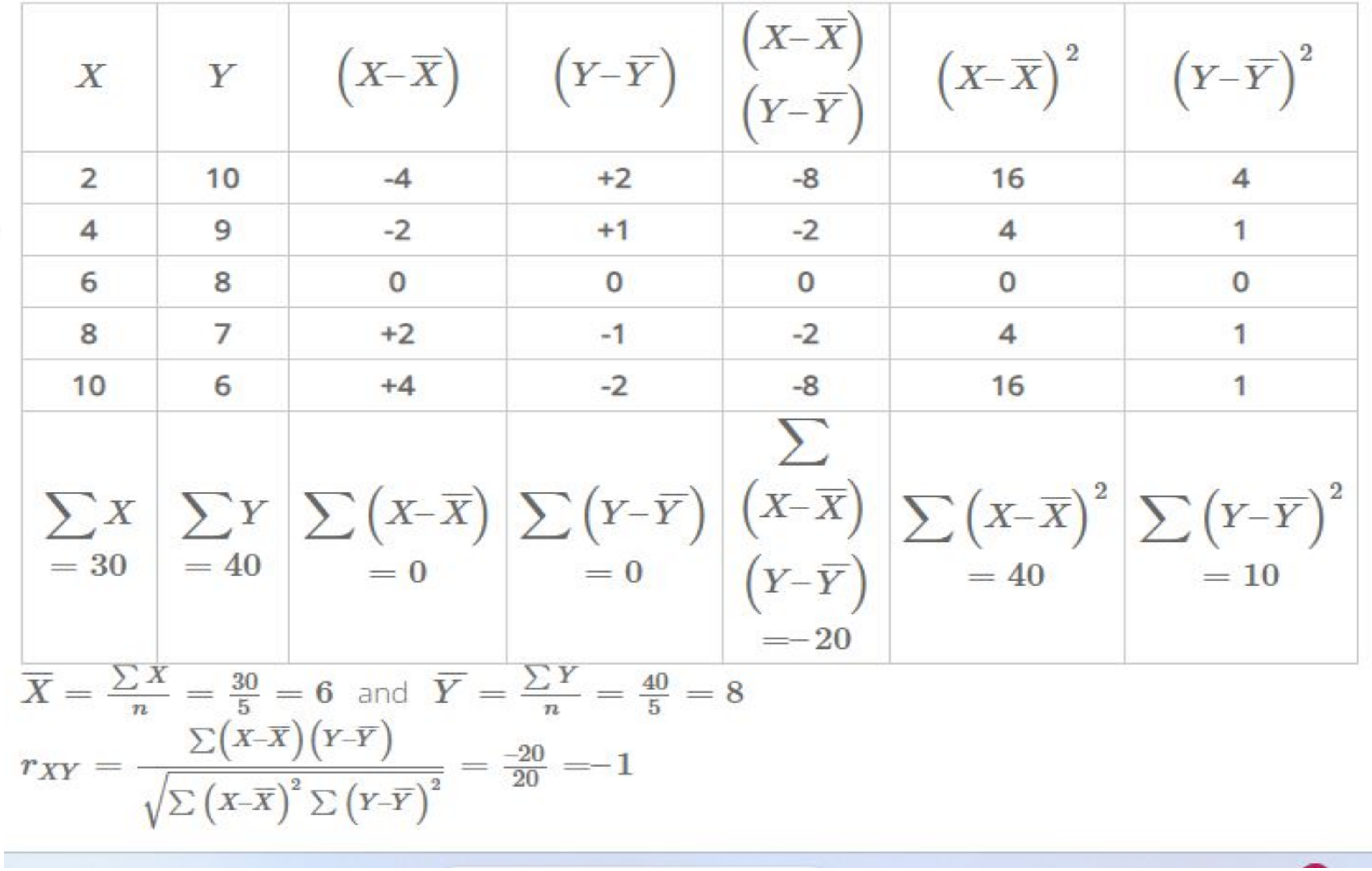


Calculate and analyze the correlation coefficient between the number of study hours and the number of sleeping hours of different students



| Number of Study Hours | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| Number of Sleeping Hours | 10 | 9 | 8 | 7 | 6 |

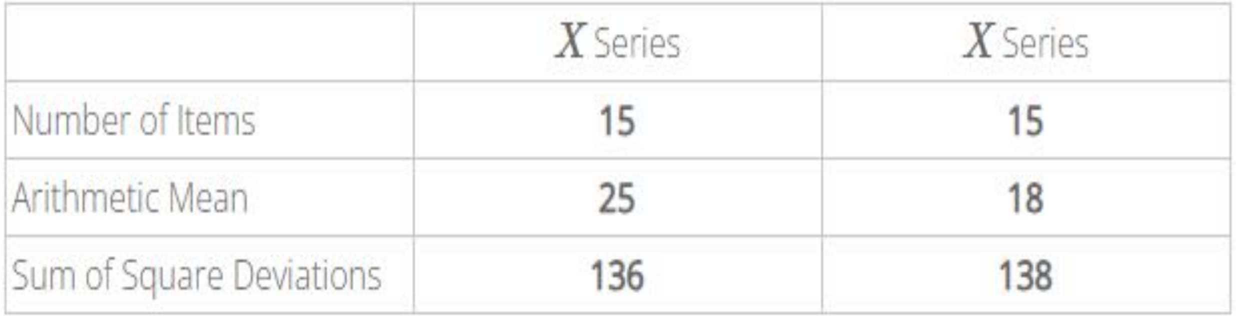
There is a perfect negative correlation between the number of study hours and the number of sleeping hours.



| $X$ | $Y$ | $(X-\overline{X})$ | $(Y-\overline{Y})$ | $(X-\overline{X})(Y-\overline{Y})$ | $(X-\overline{X})^2$ | $(Y-\overline{Y})^2$ |
|---|---|---|---|---|---|---|
| 2 | 10 | -4 | +2 | -8 | 16 | 4 |
| 4 | 9 | -2 | +1 | -2 | 4 | 1 |
| 6 | 8 | 0 | 0 | 0 | 0 | 0 |
| 8 | 7 | +2 | -1 | -2 | 4 | 1 |
| 10 | 6 | +4 | -2 | -8 | 16 | 1 |
| $\sum X = 30$ | $\sum Y = 40$ | $\sum(X-\overline{X}) = 0$ | $\sum(Y-\overline{Y}) = 0$ | $\sum (X-\overline{X})(Y-\overline{Y}) =- 20$ | $\sum(X-\overline{X})^2 = 40$ | $\sum(Y-\overline{Y})^2 = 10$ |

$\overline{X} = \frac{\sum X}{n} = \frac{30}{5} = 6$ and $\overline{Y} = \frac{\sum Y}{n} = \frac{40}{5} = 8$

$$r_{XY} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2\sum(Y-\overline{Y})^2}} = \frac{-20}{20} =-1$$



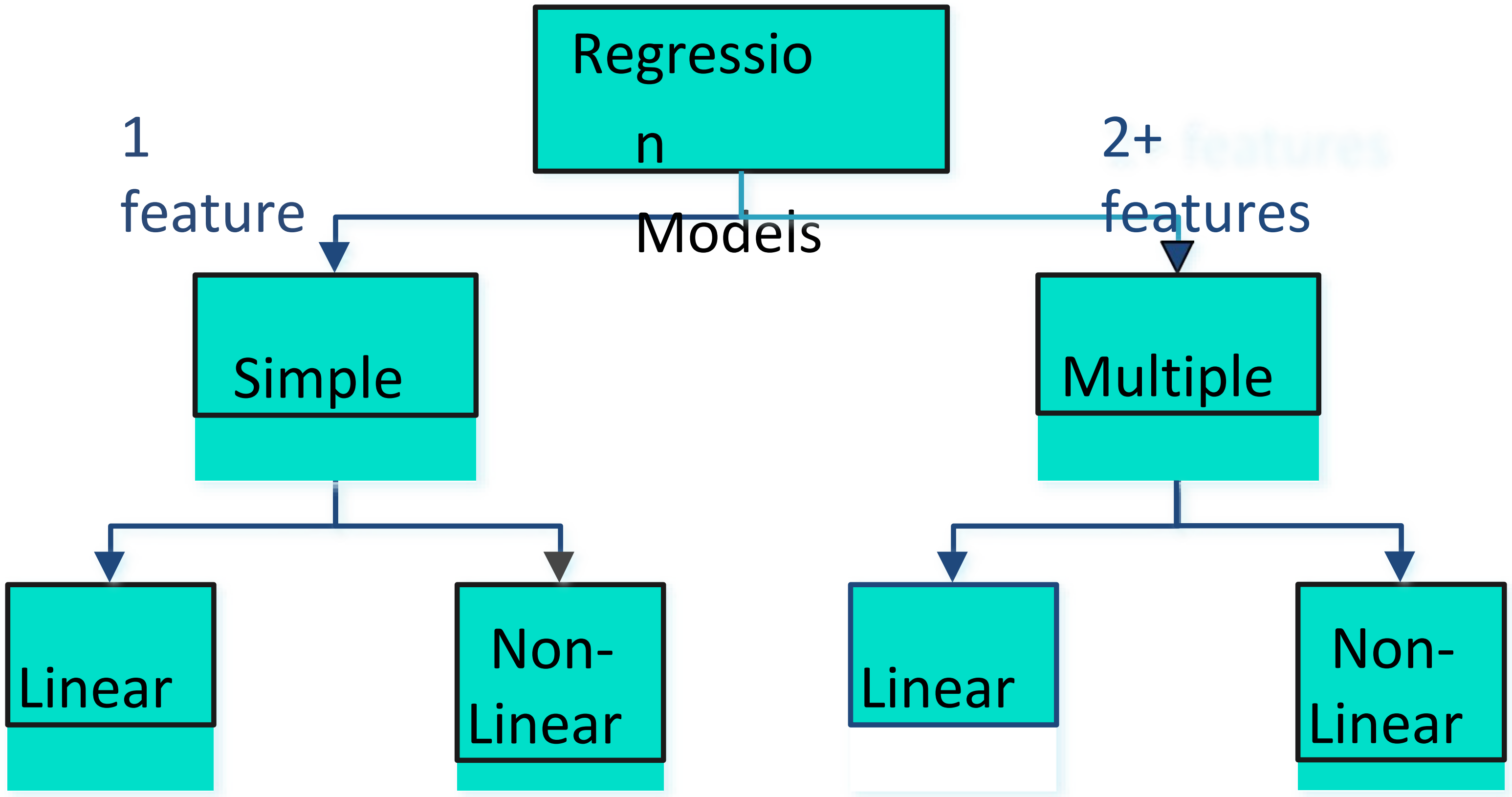
| | $X$ Series | $X$ Series |
|---|---|---|
| Number of Items | 15 | 15 |
| Arithmetic Mean | 25 | 18 |
| Sum of Square Deviations | 136 | 138 |

Here $n = 15, \bar{X} = 25, \bar{Y} = 18, \sum\left(X-\bar{X}\right)^2 = \sum\left(Y-\bar{Y}\right)^2 = 138$

$$\sum\left(X-\bar{X}\right)\left(Y-\bar{Y}\right) = 122$$

$$r = \frac{\sum\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)}{\sqrt{\sum\left(X-\bar{X}\right)^2 \sum\left(Y-\bar{Y}\right)^2}} = \frac{122}{\sqrt{(136)(138)}} = \frac{122}{137} = 0.89$$

# Types of Regression Models



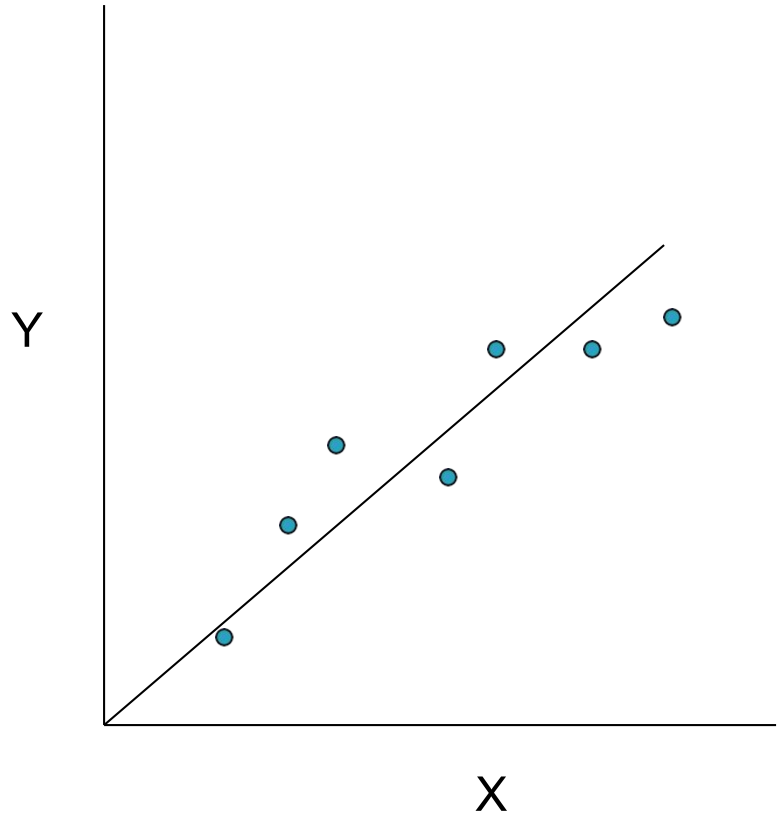
# Linear regression

- Given an input x compute an output y
- For example:

  - Predict height from age

  -Predict house price from house area

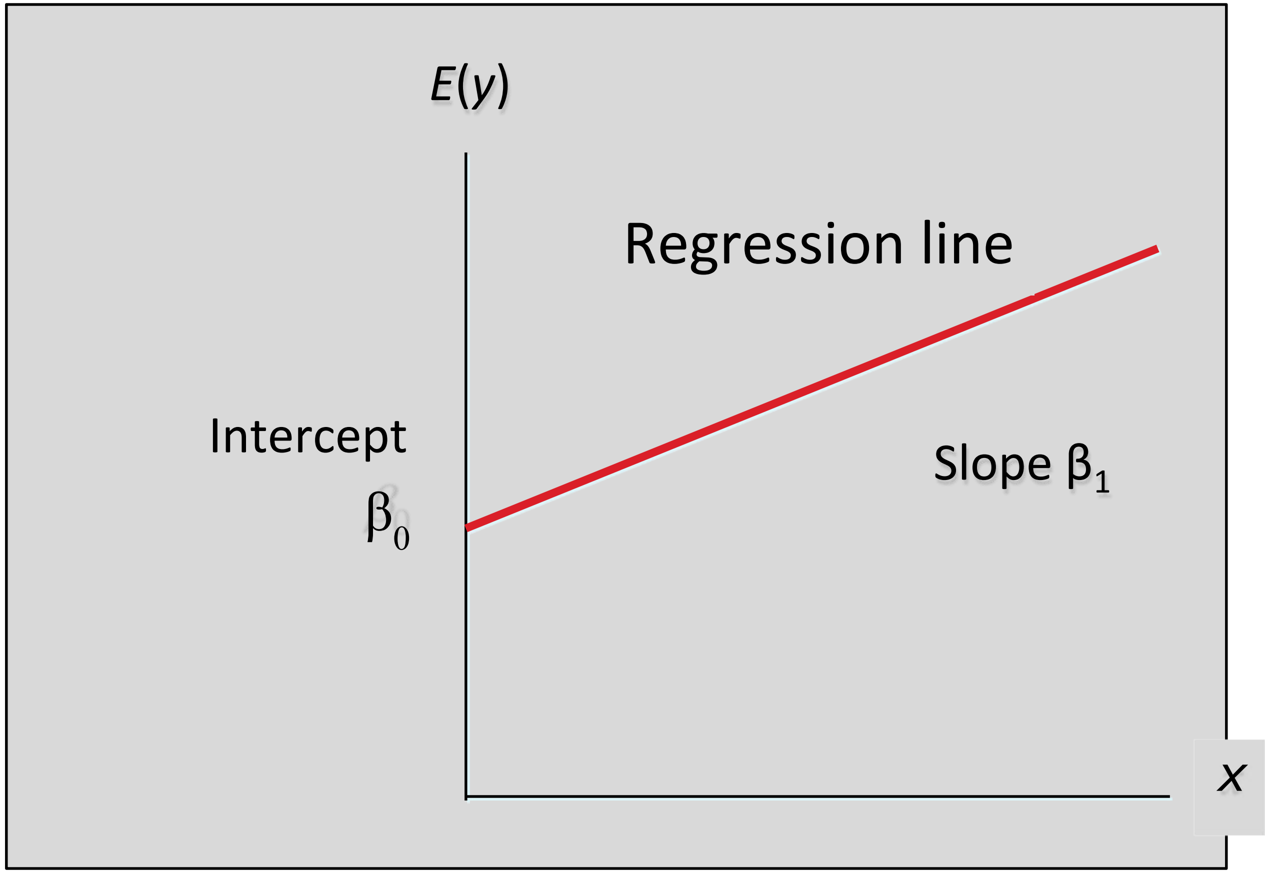
  -Predict distance from wall from sensors



# Simple Linear Regression Equation

- Relationship Between Variables Is a Linear Function

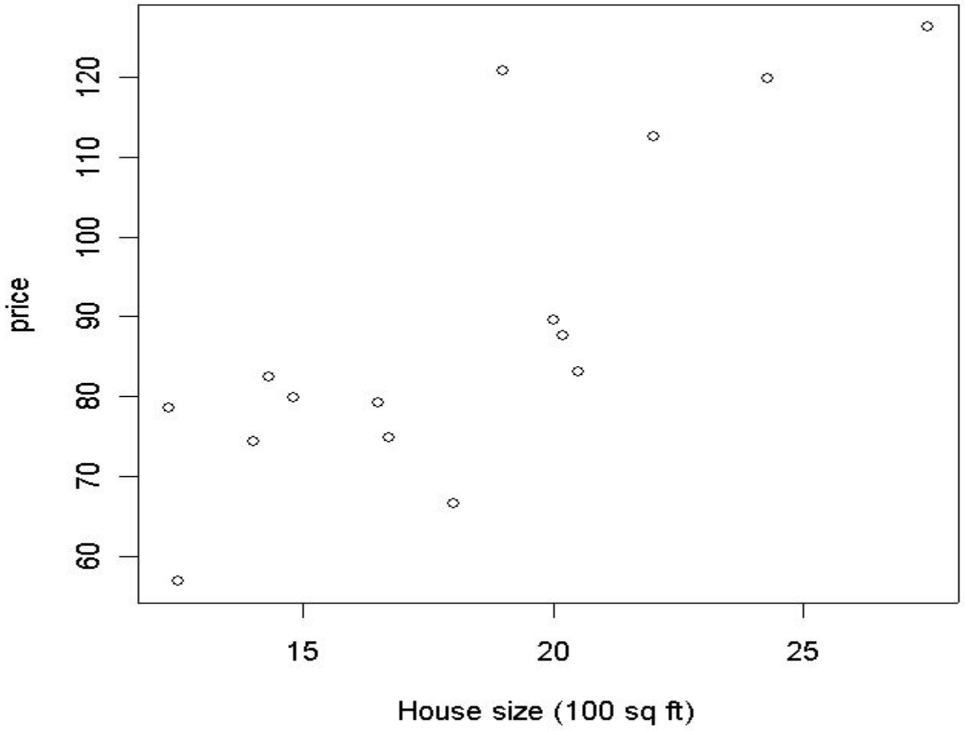
$$Y=\overset{\text{Intercept}}{\beta_0}+\overset{\text{Slope}}{\beta_1}x+\overset{\text{Random Error}}{\epsilon}$$



| House Number | *Y:* Actual Selling Price | *X*: House Size (100s ft2) |
|---|---|---|
| 1 | 89.5 | 20.0 |
| 2 | 79.9 | 14.8 |
| 3 | 83.1 | 20.5 |
| 4 | 56.9 | 12.5 |
| 5 | 66.6 | 18.0 |
| 6 | 82.5 | 14.3 |
| 7 | 126.3 | 27.5 |
| 8 | 79.3 | 16.5 |
| 9 | 119.9 | 24.3 |
| 10 | 87.6 | 20.2 |
| 11 | 112.6 | 22.0 |
| 12 | 120.8 | .019 |
| 13 | 78.5 | 12.3 |
| 14 | 74.3 | 14.0 |
| 15 | 74.8 | 16.7 |
| Averages | 88.84 | 18.17 |

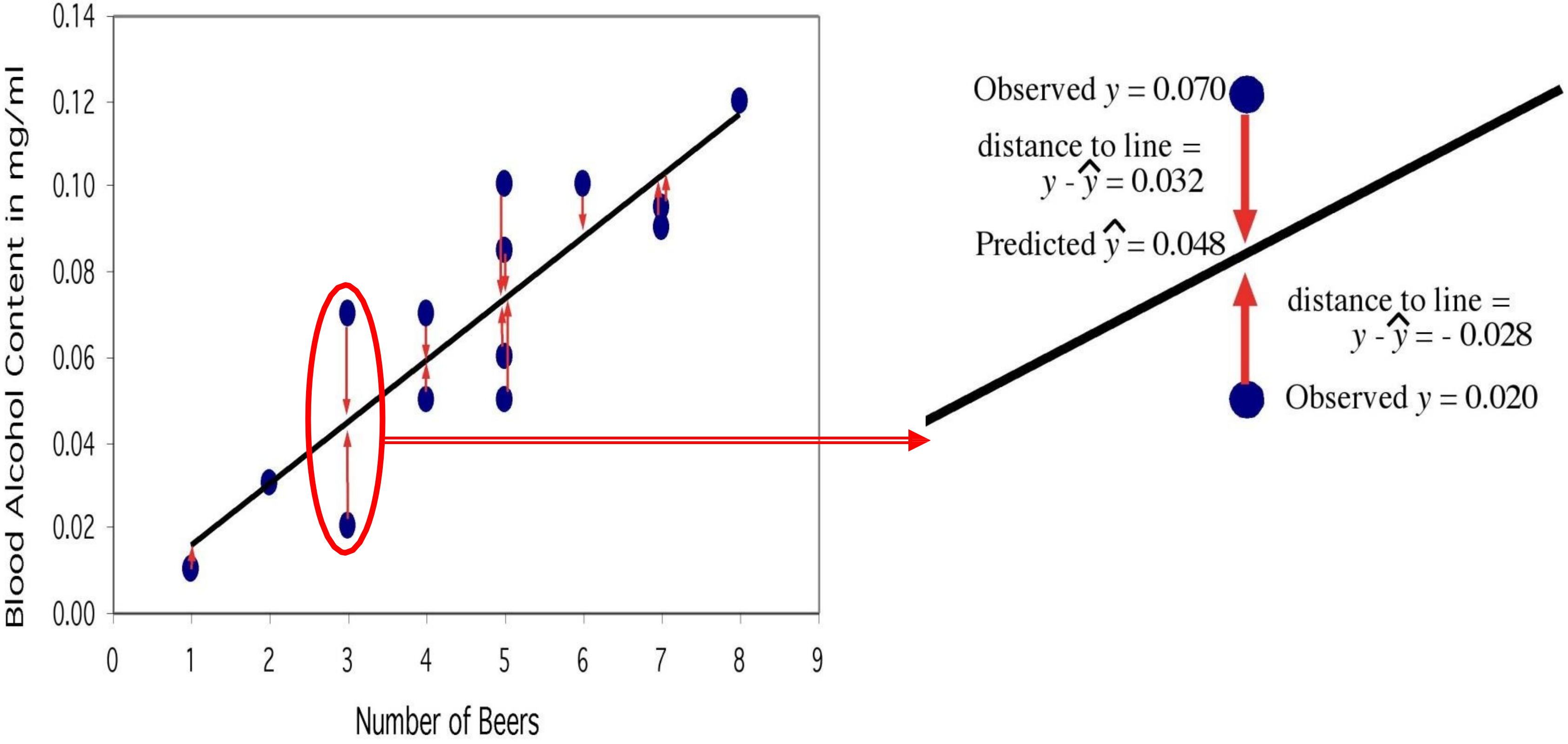
Sample 15 houses
from the region.

# House price vs size



# The regression line

The least-squares regression line is the unique line such that the sum of the squared vertical (*y*) distances between the data points and the line is the smallest possible.



Observed $y = 0.070$

distance to line = $y - \hat{y} = 0.032$

Predicted $\hat{y} = 0.048$

distance to line = $y - \hat{y} = - 0.028$

Observed $y = 0.020$

Calculate the regression coefficient and obtain the lines of regression for the following data

| X | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Y | 9 | 8 | 10 | 12 | 11 | 13 | 14 |

| $X$ | $Y$ | $X^2$ | $Y^2$ | $X^Y$ |
|---|---|---|---|---|
| 1 | 9 | 1 | 81 | 9 |
| 2 | 8 | 4 | 64 | 16 |
| 3 | 10 | 9 | 100 | 30 |
| 4 | 12 | 16 | 144 | 48 |
| 5 | 11 | 25 | 121 | 55 |
| 6 | 13 | 36 | 169 | 78 |
| 7 | 14 | 49 | 196 | 98 |
| $\sum X = 28$ | $\sum Y = 77$ | $\sum X^2 = 140$ | $\sum Y^2 = 875$ | $\sum XY = 334$ |

$$\overline{X} = \frac{\Sigma X}{N} = \frac{28}{7} = 4,$$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{77}{7} = 11$$

**Regression coefficient of $X$ on $Y$**

$$b_{xy} = \frac{N\Sigma XY-(\Sigma X)(\Sigma Y)}{N\ \Sigma Y^2-(\Sigma Y)^2}$$

$$= \frac{7(334)-(28)(77)}{7(875)-(77)^2}$$

$$= \frac{2338-2156}{6125-5929}$$

$$= \frac{182}{196}$$

$$b_{xy} = 0.929$$

**(ii) Regression coefficient *of* $Y$ on $X$**

$$b_{yx} = \frac{N\Sigma XY-(\Sigma X)(\Sigma Y)}{N\ \Sigma X^2-(\Sigma X)^2}$$

$$= \frac{7(334)-(28)(77)}{7(140)-(28)^2}$$

$$= \frac{2338-2156}{980-784}$$

$$= \frac{182}{196}$$

$$\therefore \quad b_{yx} = 0.929$$

**(i) Regression equation of $X$ on $Y$**

$$X-\overline{X} = b_{xy}(\mathrm{Y}-\overline{Y})$$

$$X-4 = 0.929(Y-11)$$

$$X-4 = 0.929Y-10.219$$

$\therefore$ The regression equation $X$ on $Y$ is $X= 0.929Y-6.219$

**(iii) Regression equation of $Y$ on $X$**

$$Y-\overline{Y} = b_{yx}(\mathrm{X}-\overline{X})$$

$$\mathrm{Y}-11 = 0.929\ (X-4)$$

$$\mathrm{Y} = 0.929X-3.716+11$$

$$= 0.929X+7.284$$

The regression equation of $Y$ on $X$ is $Y= 0.929X + 7.284$

# Example-1

| X [midterm exam] | Y [Final Exam] |
|---|---|
| 72 | 84 |
| 50 | 63 |
| 81 | 77 |
| 74 | 78 |
| 94 | 90 |
| 86 | 75 |
| 59 | 49 |
| 83 | 79 |
| 65 | 77 |
| 33 | 52 |
| 88 | 74 |
| 81 | 90 |

The given table shows the midterm and final exam grades obtained for students in a database course.

a) Use the method of least squares to find an equation for the prediction of a student's final exam grade based on the student's midterm grade in the course.

b) Predict the final exam grade of student who received an 86 in the midterm exam based on the equation of least squares.

# Example-2

## Consider a set of points x and the predicted values y as given below:

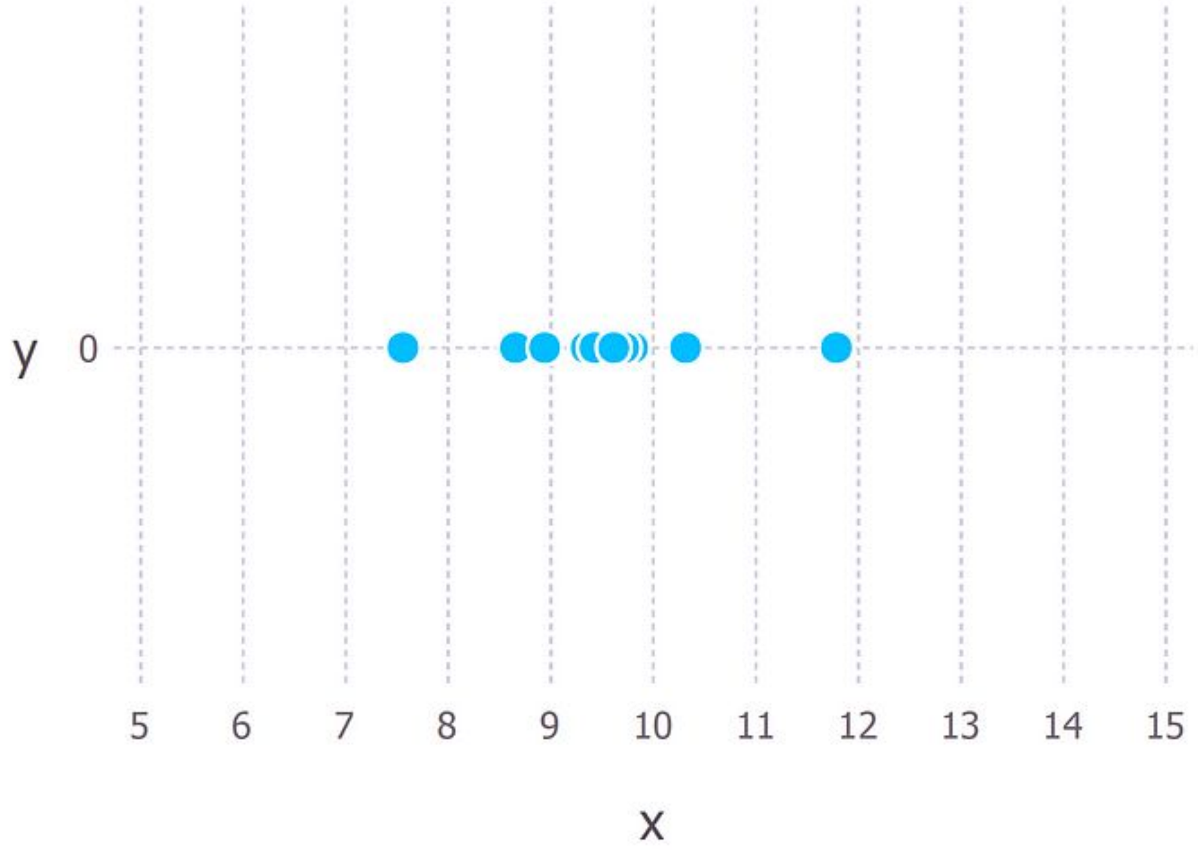
| x | y |
|---|---|
| 0 | 2 |
| 2 | 4 |
| 3 | 5 |
| 5 | 6 |
| 6 | 7.5 |
| 7 | 8.5 |
| 8 | 9 |
| 8 | 10 |
| 9 | 11 |
| 10 | 12 |

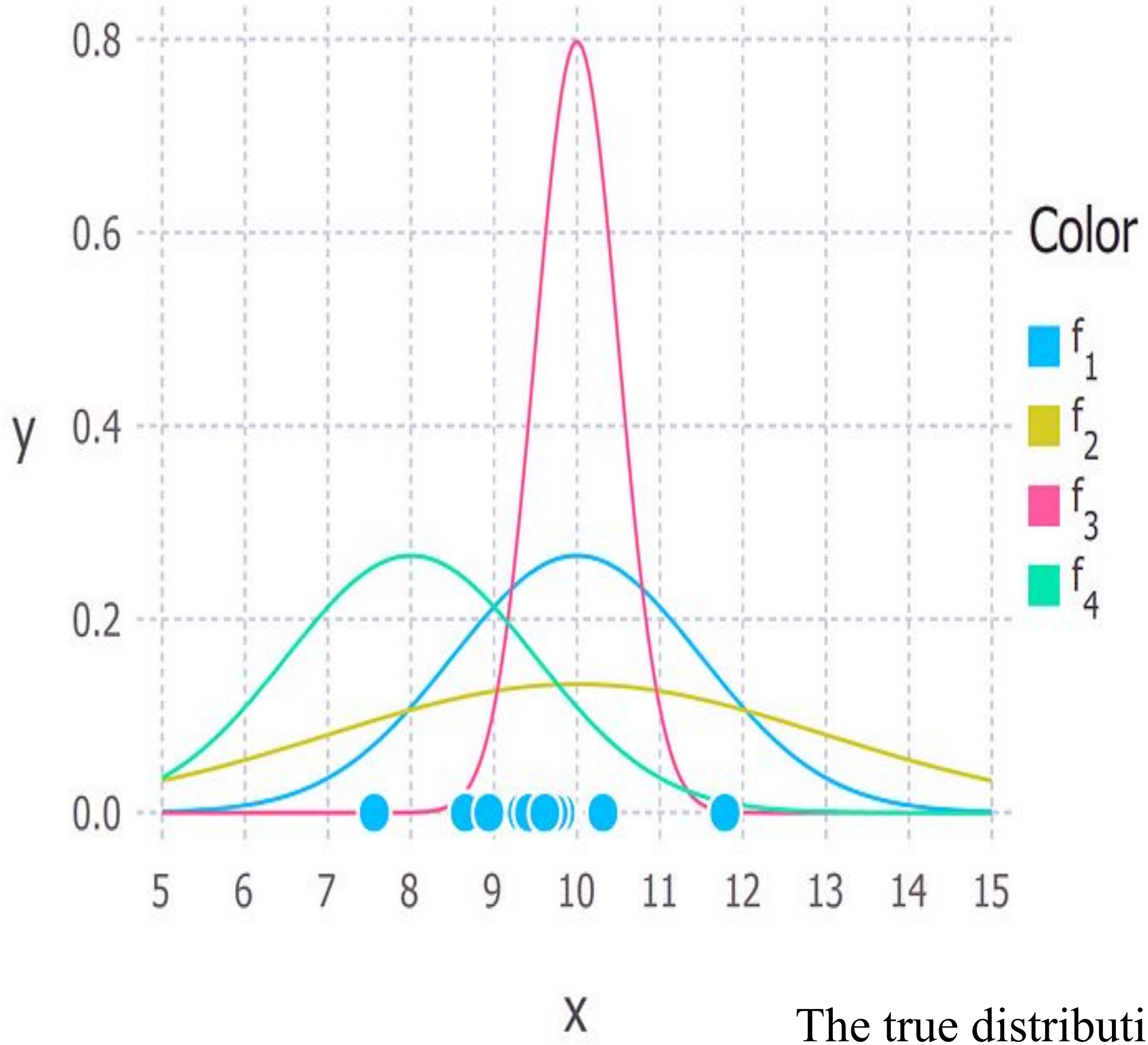
a) Plot the data.

b) Use Linear Regression model and consider a data value for x and predict the value of y using the above model.

# Maximum likelihood test

- Maximum likelihood estimation is a method that determines values for the parameters of a model. The parameter values are found such that they maximize the likelihood that the process described by the model produced the data that were actually observed.





The 10 data points and possible Gaussian distributions from which the data were drawn.
f1 is normally distributed with mean 10 and variance 2.25 (variance is equal to the square of the standard deviation),
this is also denoted
f1 ~ N (10, 2.25).
f2 ~ N (10, 9),
f3 ~ N (10, 0.25) and
f4 ~ N (8, 2.25).
The goal of maximum likelihood is to find the parameter values that give the distribution that maximize the probability of observing the data.

The true distribution from which the data were generated was f1 ~ N(10, 2.25), which is the blue curve in the figure

Suppose there are three data points and assume that they have been generated from a process that is adequately described by a Gaussian distribution.

**These points are 9, 9.5 and 11.**

To calculate the total probability of observing all of the data, i.e. the joint probability distribution of all observed data points., you need to calculate some conditional probabilities.

So *the assumption is that each data point is generated independently of the others*.

If the events (i.e. the process that generates the data) are independent, then the total probability of observing all of data is the product of observing each data point individually (i.e. the product of the marginal probabilities).

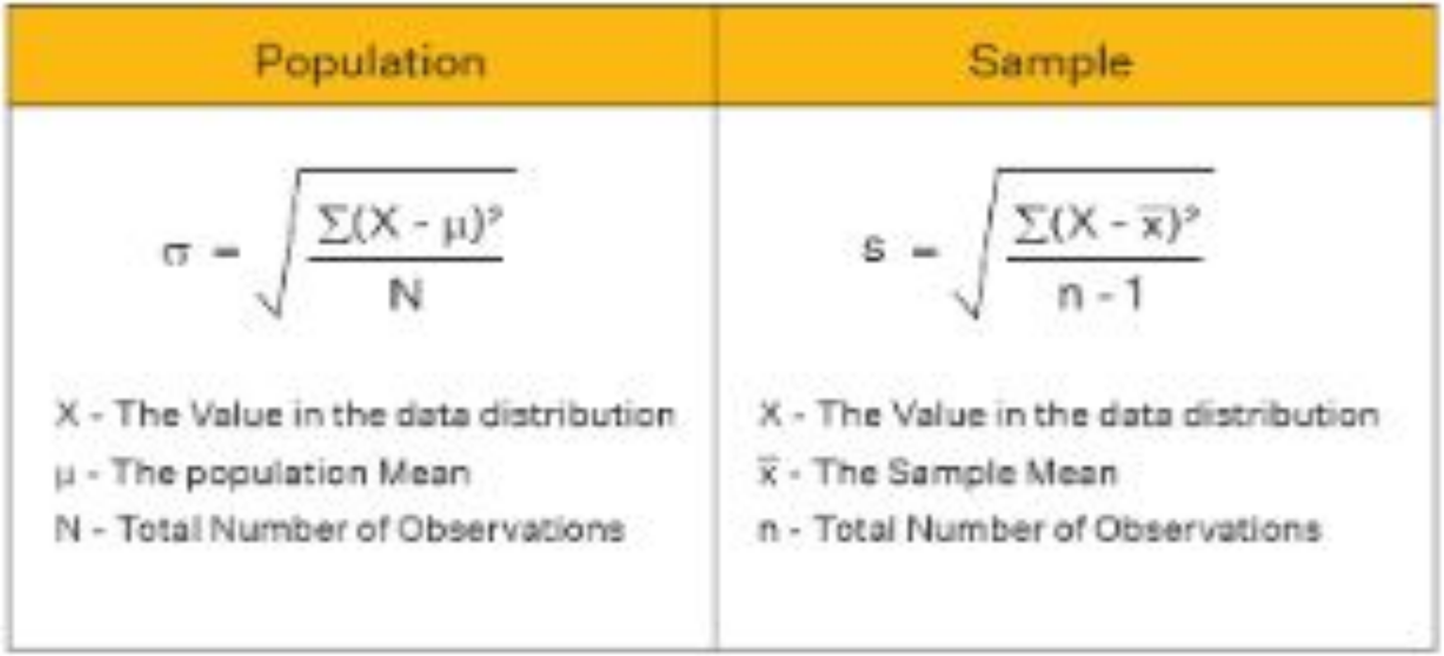
The probability density of observing a single data point *x*, that is generated from a Gaussian distribution is given by:

$$P(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The total (joint) probability density of observing the three data points is given by:

$$P(9, 9.5, 11; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9.5-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(11-\mu)^2}{2\sigma^2}\right)$$

Sample Mean =
(Sum of terms) ÷ (Number of Terms)

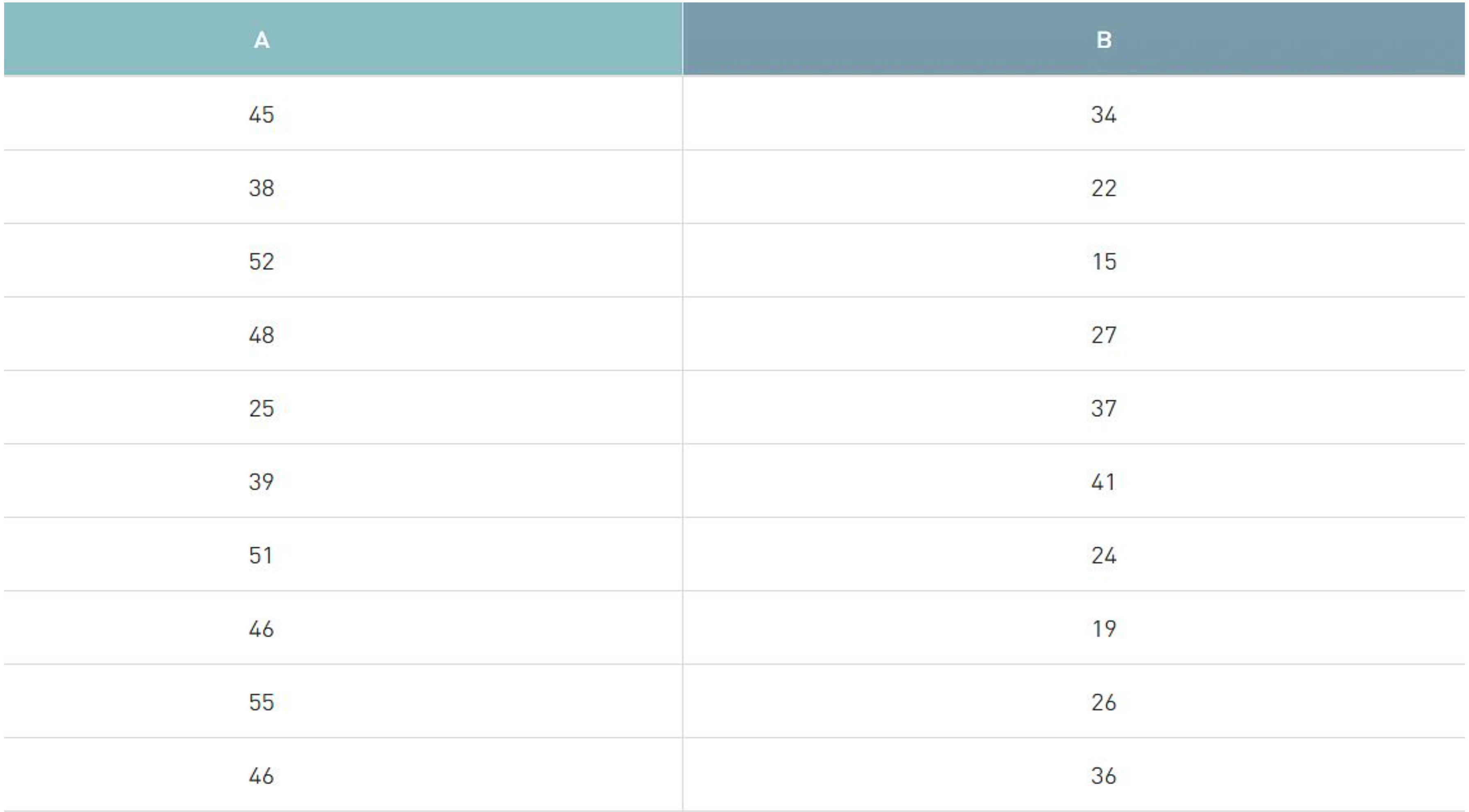


| Population | Sample |
|---|---|
| $\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$ | $s = \sqrt{\frac{\Sigma(X-\bar{x})^2}{n-1}}$ |
| X - The Value in the data distribution<br>µ - The population Mean<br>N - Total Number of Observations | X - The Value in the data distribution<br>x̄ - The Sample Mean<br>n - Total Number of Observations |

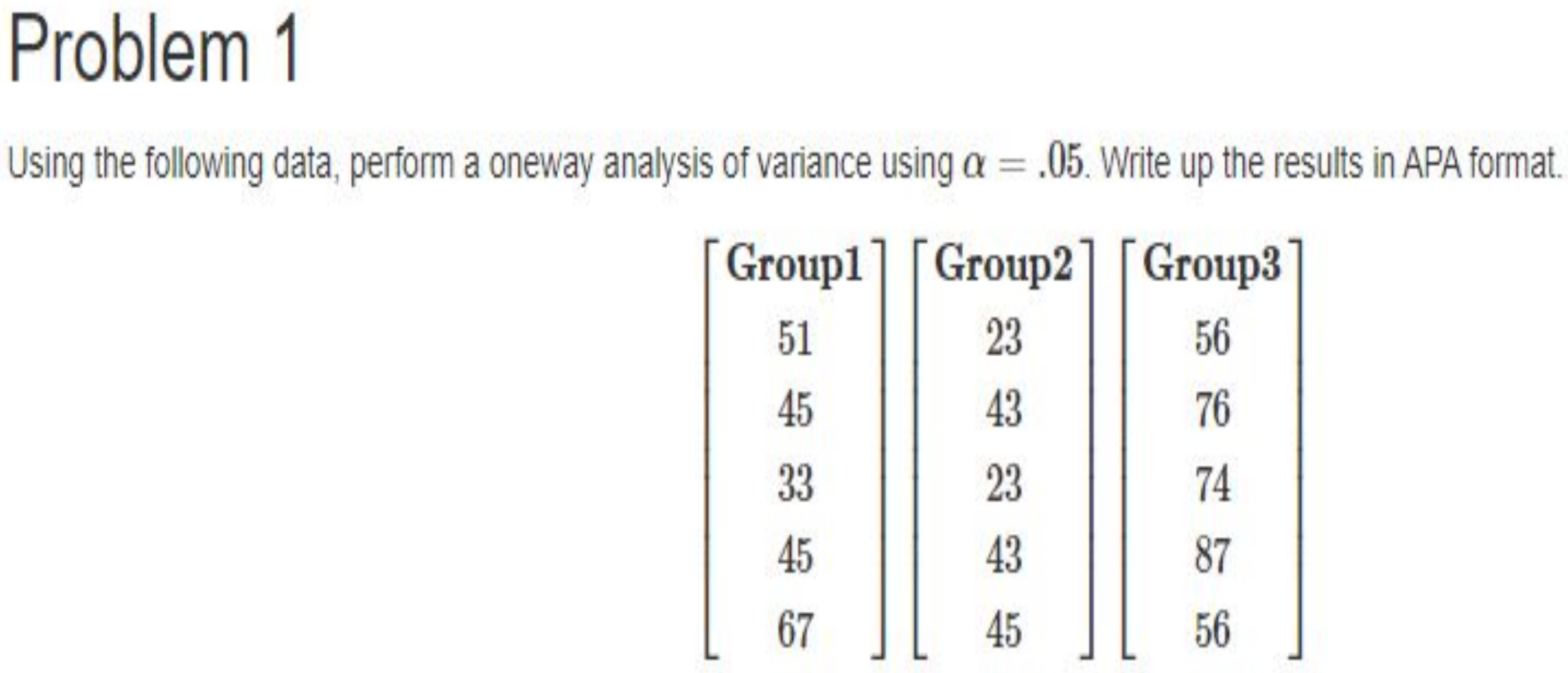
A doctor claims that a particular hospital contains more than 100 diabetes patients with a sugar level of 234 or more. To verify the claim, a random test was conducted on 90 diabetes patients. The test resulted in a mean blood sugar level of 279. In addition, the test resulted in a standard deviation of 18. Here, we set the significance level at **22.50.**

A professor wants to know if her introductory statistics class has a good grasp of basic math. Six students are chosen at random from the class and given a math proficiency test. The professor wants the class to be able to score above 70 on the test. The six students get scores of 62, 92, 75, 68, 83, and 95. Can the professor have 90 percent confidence that the mean score for the class on the test would be above **70**?

# Calculate t-test for the following data.



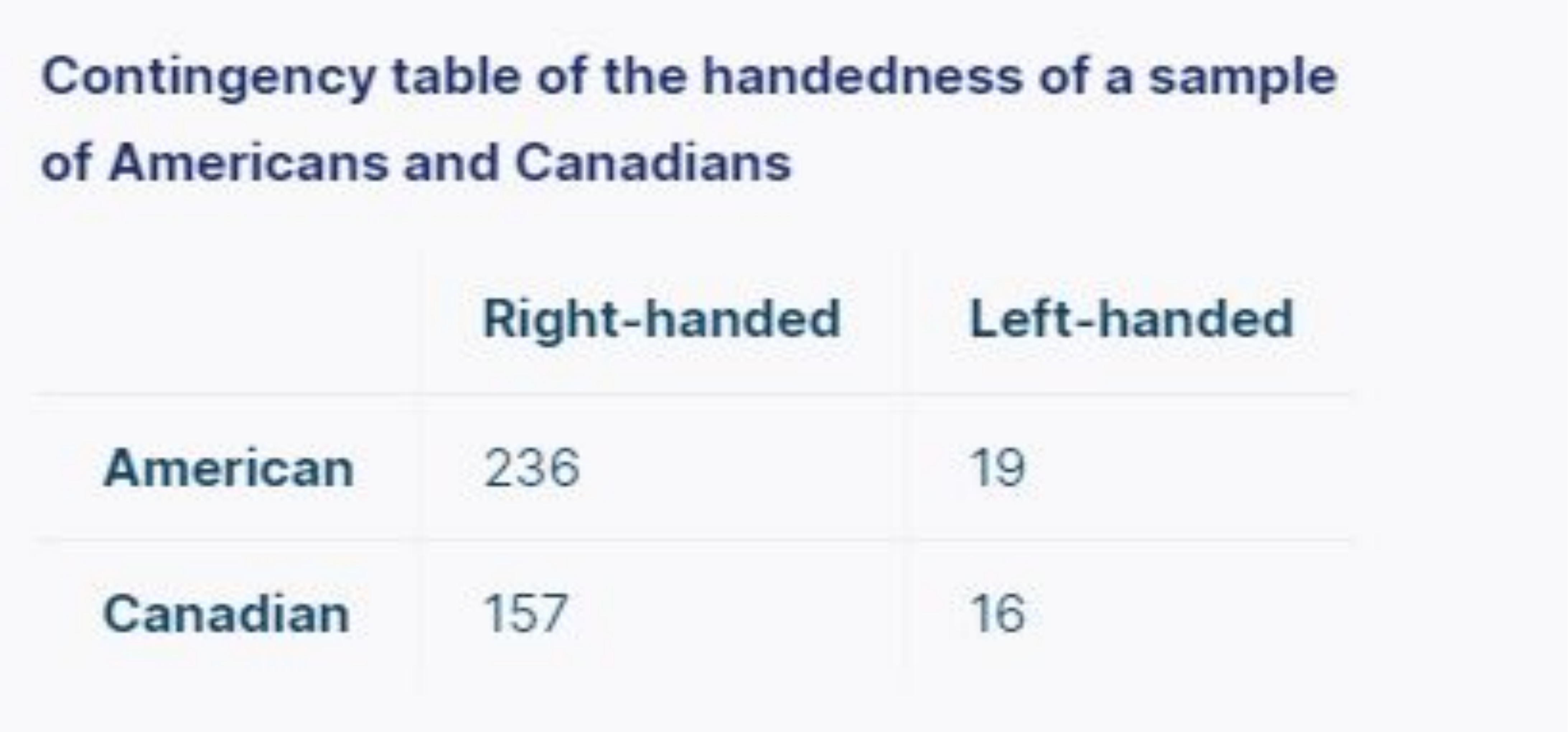
| A | B |
|---|---|
| 45 | 34 |
| 38 | 22 |
| 52 | 15 |
| 48 | 27 |
| 25 | 37 |
| 39 | 41 |
| 51 | 24 |
| 46 | 19 |
| 55 | 26 |
| 46 | 36 |



# Problem 1

Using the following data, perform a oneway analysis of variance using $\alpha = .05$. Write up the results in APA format.

$$\begin{bmatrix} \textbf{Group1} \\ 51 \\ 45 \\ 33 \\ 45 \\ 67 \end{bmatrix} \begin{bmatrix} \textbf{Group2} \\ 23 \\ 43 \\ 23 \\ 43 \\ 45 \end{bmatrix} \begin{bmatrix} \textbf{Group3} \\ 56 \\ 76 \\ 74 \\ 87 \\ 56 \end{bmatrix}$$



**Contingency table of the handedness of a sample of Americans and Canadians**

| | Right-handed | Left-handed |
|---|---|---|
| **American** | 236 | 19 |
| **Canadian** | 157 | 16 |