Basic Definition of Probability

Before going through this session, you should understand the basic definition of probability. In order to understand that, you can go through the following links -

1. Math is Fun - Probability
2. Mathopolis - Probability

1. The probability of an event is equal to the:

- Number of favourable outcomes / Number of unfavourable outcomes
- Number of unfavourable outcomes / Number of total outcomes
- Number of favourable outcomes / Number of total outcomes
    - ✓ **Correct**
    - **Feedback:**
    - Yes, this is the basic definition of probability.
- Number of total outcomes / Number of favourable outcomes

2. A regular die is thrown. The probability of getting the number 5 is:

- 5 / 6
- 1 / 6

  ✓ **Correct**

  Feedback:

  Number of favourable outcomes = 1 (i.e. the outcome that the die throws 5). Total number of outcomes = 6 (The die could throw 1, 2, 3, 4, 5 or 6). Probability = Number of favourable outcomes / Total number of outcomes = 1 / 6.
- 1 / 3
- 1 / 2

3. A regular die is thrown. The probability of getting an even number is:

- 2 / 6
- 3 / 6

  ✓ Correct

  Feedback:

  Number of favourable outcomes = 3 (the die throws 2, 4, 6). Total number of outcomes = 6 (the die could throw 1, 2, 3, 4, 5 or 6). Probability = Number of favourable outcomes / Total number of outcomes = 3 / 6
- 1 / 6
- 4 / 6

4. Out of the following values, which is/are the possible value(s) of probability?

- 0.1

  ✓ Correct

  Feedback:

  The probability of an event can only be between 0 and 1
- -0.1
- 1.0

  ✓ Correct

  Feedback:

  The probability of an event can only be between 0 and 1
- 10

5. The probability of all possible events adds up to:

- 1

  ✓ Correct

  Feedback:

  The probability of all possible events adds up to 1, as the number of favourable events becomes the total number of events.
- 0
- Infinity
- More information needed

6. A regular die is thrown. The probability of getting the number 7 is:

- 1 / 6
- 0

  ✓ **Correct**

  Feedback:

  Number of favourable outcomes = 0 (you can never get a 7 by throwing a regular die). Total number of outcomes = 6 (the die could throw 1, 2, 3, 4, 5 or 6). Probability = Number of favourable outcomes / Total number of outcomes = 0.
- 7 / 6
- 1 / 7

Advanced Concepts in Probability

1. Addition rule of probability for mutually exclusive events
2. Multiplication rule of probability for independent events
3. $^{n}C_{r}$(Combinatorics)

1. A regular die is thrown. The probability of getting the number 4 or 5 in this throw is:

- 0
- 1 / 6
- 1 / 3

  ✓ **Correct**

  Feedback:

  Define X as the number achieved after throwing the die. You know that P(X = 4) = 1 / 6 and P(X = 5) = 1 / 6. Now, using addition rule of probability, you will get that P(X = 4 or 5) = P(X = 4) + P(X = 5) = 1 / 6 + 1 / 6 = 1 / 3.
- 1 / 2

2. A regular die is thrown. The probability of getting the number 4 and 5 in this throw is:

- 0

  ✓ **Correct**

  Feedback:

  It is impossible to get two numbers in one throw of the die.
- 1 / 6
- 1 / 3
- 1 / 2

3. A regular die is thrown. The probability of getting the number 4 in the first throw and 5 in the second throw is:

- 2 / 6
- 1 / 36

  ✓ **Correct**

  Feedback:

  Again, P(getting the number 4) = 1 / 6 and P(getting the number 5) = 1 / 6. Now, using multiplication rule, you get that P(getting the number 4 in the 1st trial AND the number 5 in the 2nd trial) = P(getting the number 4 in the 1st trial)*P(getting the number 5 in the second trial) = 1 / 6 * 1 / 6 = 1 / 36.
- 20 / 36
- 9 / 36

4. A fair coin is tossed once. The probability of getting head or tail is?

- 1

  ✓ **Correct**

  Feedback:

  P(heads) = 0.5, P(tails) = 0.5, Using addition rule, you have P(heads OR tails) = 0.5 + 0.5 = 1.
- 0.25
- 0.5
- 0

5. A fair coin is tossed. The probability of getting a head and a tail is?

- 1
- 0.25
- 0.5
- 0

✓ **Correct**

Feedback:
It is impossible to get heads and tails in one toss of the coin.

6. A fair coin is tossed. The probability of getting a head in the 1st toss and a tail in the 2nd toss is?

- 1
- 0.25

✓ **Correct**

Feedback:
P(heads) = 0.5, P(tails) = 0.5, Using multiplication rule, you have P(heads in 1st toss AND tails in 2nd toss) = P(heads in 1st toss)*P(tails in 2nd toss) = 0.5*0.5 = 0.25.

- 0.5
- 0

7. MS Dhoni is selecting the Indian cricket team, and has selected 8 players. For the remaining 3 players, he has 5 options to choose from -

1. Jasprit Bumrah
2. Amit Mishra
3. Axar Patel
4. Ravindra Jadeja
5. Bhubaneshwar Kumar

He could select this combination -

- Jasprit Bumrah
- Bhubaneshwar Kumar
- Ravindra Jadeja

Or this one -

- Amit Mishra
- Axar Patel

- Ravindra Jadeja

In total, how many different combinations of 3 players can he pick for the team?

- ${}_3C_5$
- ${}_5C_3$

  ✓ Correct

  Feedback:
  Let's say we have n distinct objects, from which we are to select r of them. The number of different combinations in which this selection can be done, is given by ${}_nC_r$. For example, here, we have 5 players, out of which, 3 are to be selected. The number of combinations in which 3 players can be selected, thus, is given by ${}_5C_3$ (10).
- ${}_8C_3$
- ${}_8C_5$

The Z-Test

When you perform an analysis on a sample, you only get the statistics of the sample. You want to make claims about the entire population using the sample statistics. But remember that these are just claims; so, you cannot be sure that they are true. This kind of a claim or assumption is called a hypothesis.

For example, your hypothesis may be that the average lead content in a food product is less than 2.5 ppm, or the average time to resolve a query at a call centre is 6 minutes.

Whatever your hypothesis is, it is only a claim based on a limited amount of data and not the entire population. Hypothesis testing helps you statistically verify whether a claim is likely to be true or not for the whole population.

Thus, we can say that **hypothesis testing is a method or procedure that tests the statistical validity of a claim.**

The components involved in hypothesis testing are as follows:

- Null hypothesis: A null hypothesis is a prevailing belief about a population; it states that there is no change or no difference in the situation and assumes that the status quo is true. It is denoted by H_0.
- Alternative hypothesis: An alternative hypothesis is a claim that opposes the null hypothesis. It challenges the status quo and may or may not be proved. It is symbolised by H_1 .

Let's understand these with the help of an example.

Jeep, a well-known car maker, claims that its car 'Compass' gives a mileage of at least 17 km/litre.

The null hypothesis for this case would be:

H_0**: $\mu \geq 17$**

And the alternative hypothesis is:

H_1**: $\mu < 17$**

An important thing to note is that a null hypothesis always has the '=' sign and is a common belief about the population, while the alternative hypothesis never has the '=' sign and always challenges the status quo.

Steps in hypothesis testing

The process of hypothesis testing has a well-defined path that stems from our intuition. For the sake of simplicity, it can be broken down into easy-to-remember steps.

The first step in the process is to **define the hypothesis**. This involves stating the null and alternative hypotheses for the problem. For example, Google claims that its internet browser 'Chrome' is the best in the industry, as it has an optimum boot time of only 250 ms, with a standard deviation of 9 ms. Sam, a tech geek, wanted to test the claim of Google. So, he randomly collected boot time data of 165 devices of Chrome and got a sample mean of 247 ms.

Based on this, the hypotheses can be defined as follows:
Ho: $\mu = 250$, i.e., the mean boot time is 250 ms.
Ha: $\mu \neq 250$, i.e., the mean boot time is not 250 ms.

The next step is to **identify the associated distribution.**

Condition 1: n >30, which means that the population sample size should be greater than 30 observations.
Condition 2: $\boldsymbol{\sigma}$ is known, i.e., the population standard deviation is known.

Now, if both these conditions are satisfied, you go for a **normal distribution** or **Z-test**; otherwise, you use the **t-test**.
This problem satisfies both these conditions, as the sample size is 165 and the standard deviation is 9 ms. So, you go for a normal distribution.

The next step is to **determine the test statistic**. A test statistic, in simple terms, is a value that is to be calculated from some given data, which is then used to compare the results arrived at with the tabular values.

The test statistic for a normal distribution or a Z-test is defined as:

$$Z = \frac{x-\mu}{\sigma/\sqrt{n}}$$

Here, x is the process mean, μ is the population mean, σ is the standard deviation and n is the sample size.

In this example, the distribution is normal. So, let's calculate the test statistic using the aforementioned formula, which gives us:

Z = (247 - 250)/(9/√165)

Z = -4.3

Continuing with our demonstration on the 'Chrome' browser, we will now test our hypothesis at a 95% confidence level. For a 95% confidence interval, Z critical value = +1.96 and -1.96; these are the upper and lower critical values, respectively. The test statistic value we calculated is -4.3.

The region between +1.96 and -1.96 is called the **acceptance region**, and the region outside it is called the **critical region**.

So, the question that arises is: How do you compare the two Z-statistics? This comparison is done on the basis of whether or not the calculated Z-statistic lies in our stated confidence interval, also called the **region of acceptance**.

If the calculated Z-statistic is in the region of acceptance, you **fail to reject** the null hypothesis. If the calculated Z-statistic lies outside the region of acceptance, i.e., in the critical region, you **reject** the null hypothesis.

Going back to our previous example, the test statistic value is -4.3, which lies outside the region of acceptance of ±1.96. So, you reject the null hypothesis.

It is also important to understand that you can **never accept the null hypothesis**, you can only **fail to reject it**. This is because the whole testing takes place with an aim to **reject the present status quo**, which is what the alternative hypothesis is. So, you try and gather support for the alternative hypothesis so that you can **reject the null hypothesis**. Similarly, you can never say that you reject the alternative hypothesis; instead, you say that you **failed to reject the null hypothesis.**

For instance, suppose that your friend Diksha tells you that she has an **average score of 80** in the game of pistol shooting, based on all her past games. But you don't believe her. So, you decide to test her claim and ask for five rounds of shooting. The **null hypothesis** would be that Diksha's average score is 80, which is represented as $\mu = 80$.

Here, the **critical region lies on both sides of the population mean**. But this is not the case always.

The critical region depends upon the nature of the alternative hypothesis. An alternative hypothesis can be of two types:

1. Non-directional
2. Directional

Let's see what these terms mean. In this example, the null hypothesis, H0, is $\mu = 80$, i.e. Diksha's average score is exactly equal to 80. So, your alternative hypothesis,Ha, will be $\mu \neq 80$. This does not specifically say that it is more than 80, or less. The population mean can be more or less than 80. So, there is no indication of the direction in which it will lie, i.e., whether towards the left end or the right end of the distribution. Therefore, this kind of an alternative hypothesis is called a non-directional hypothesis.

When you test any non-directional hypothesis, you need to define the critical region on both the sides, as you need to check whether the sample mean lies to the left or the right of the assumed population mean.

This kind of a test is called a **two-tailed test** because you have to check both the tails of the sampling distribution.

But there are cases where you are only interested in finding whether the mean is lower or higher than the claimed value.

For example, if Diksha now claims that her average shooting score is greater than or equal to 80, the null hypothesis will be $\mu \geq 80$. In this case, the alternative hypothesis will be $\mu < 80$. It will hypothesise that the population mean lies in a particular direction from the assumed mean. Such an alternative hypothesis is called a directional hypothesis.

In this case, since H1: $\mu < 80$, the critical region will lie on the left tail of the sampling distribution. Thus, the hypothesis test is called a one-tailed test, and more specifically, a lower-tailed test.

Similarly, if the null hypothesis is H0: $\mu \leq 80$, then H1: $\mu > 80$.
This alternative hypothesis is also a directional hypothesis, and the critical region will lie on the right tail of the sampling distribution. So, the hypothesis test is called a one-tailed test, and more specifically, an **upper-tailed test**.

The t-Test

Now that you have learnt all the basics of hypothesis testing, you are now well equipped to frame a hypothesis, test it, and make a decision to reject or not reject the null hypothesis. (This is done considering the fact that the population standard deviation for the data is known and the sample size is greater than 30.)

But how will you test the hypothesis if these conditions are not fulfilled? Let's find out.
The t-distribution, as you studied earlier, is kind of a normal distribution; it is also symmetric and single peaked but less concentrated around its peak. In layman's terms, a t-distribution is shorter and flatter around the centre than a normal distribution. It is used to study the mean of a population that has a distribution fairly close to a normal distribution (but not an exact normal distribution).

Two simple conditions to determine when to use the t-statistic are as follows:

1. **The population standard deviation is unknown.**
2. **The sample size is less than 30.**

Even if one of them is applicable in a situation, you can comfortably go for a t-test.

The formula to determine the t-statistic is:

$$t = \frac{x - \mu}{s/\sqrt{n}}$$

Here, s is the sample standard deviation.

Let's look at a problem to get a better understanding of the t-test.

The National Highways Authority of India (NHAI) stated that the **average number of accidents per month** on national highways is **12,000**. A researcher wanted to test this claim. To that end, he collected **25 samples** for 25 months and found out that the **sample mean** was **13,105** and the **sample standard deviation** was **1638.4**.

Let's now try to solve this problem according to the steps we discussed earlier.

The hypothesis for this case will be:

H_o: $\mu = 12000$

H_a: $\mu \neq 12000$

In this case, the population standard deviation is not given. So, you will calculate the t-statistic.

$t = (x - \mu) / (s/\sqrt{(n)})$

$= (13105 - 12000)/(1638.4/\sqrt{25})$

$= 1105/327.68$

$= 3.37$

Now, as in the case of a normal test, you need to compare the value you calculated with the tabular value.

For a 90% confidence interval and a sample size of 25, the **critical t value is 1.71**. (Here is a link to the tutorial of critical t-value calculation: http://www.dummies.com/education/math/statistics/how-to-find-t-values-for-confidence-intervals/.)

Thus, our acceptance region lies between +1.71 and -1.71.

As our calculated t-value lies outside the acceptance region, you **reject the null hypothesis** and can say that you don't have sufficient evidence to support the fact that the number of accidents is equal to 12,000 per month on the highways.

With this example, you have a complete understanding of the one-sample t-test. Let's now focus on the **two sample t-test**. As the name suggests, this test is conducted on two sets of sample data in order to **compare the means of two samples.**

Note that a two-sample test can be performed for multiple statistical parameters, but you are going to focus only on the two-sample test for means, where the standard deviations of both the samples are unknown.

The formula for the two-sample t-test is:

$$t_{df} = \frac{\overline{x_1} - \overline{x_2} - \mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

df = smaller of n_1 -1 or n_2 - 1

Suppose that you want to come up with a hypothesis test regarding the mean age difference between men and women. You can use the two-sample t-test in such a case.

Chi-Squared Test

Following are the two types of chi-squared tests:

1. Chi-squared test of independence
2. Chi-squared goodness of fit (This is used to test whether the sample data correctly represents the population data.)

Chi-squared test of independence: This is used to determine whether or not there is a significant relationship between two nominal (categorical) variables.

For example, a researcher wants to examine the relationship between gender (male vs female) and the chances of developing Alzheimer's disease. The chi-squared test of independence can be used to examine this relationship. The null hypothesis (Ho) for this test is that there is no relationship between gender and life expectancy, and the

alternative hypothesis is that there is a relationship between gender and life expectancy.

Here, there are two categorical variables (nominal variables): male and female.

Let’s draw a table for both these categorical values:

	Male	**Female**
Expected Value		
Sample Value		

The expected value is calculated by assuming that the null hypothesis is correct. So, if you select a sample of, say, 100 Alzheimer’s patients, 50 should be men and 50 should be women.
Putting the expected values in the table above, you get:

	Male	**Female**
Expected Value	50	50
Sample Value		

Let’s say the sample value comes out to be a bit different, and in a sample of 100 Alzheimer’s patients, 60 are men and 40 are women.

	Male	**Female**
Expected Value	50	50
Sample Value	60	40

The test statistic for the chi-squared test is equal to $\chi 2 = \Sigma (O-E)2E$,
where O is the observed sample value and E is the expected value.

So, our test statistic will be equal to:
$\chi 2$ = (10^2)/50 + (10^2)/50 = 4

Let’s select the level of significance as 5%, or 0.05.

Degrees of freedom = (r - 1) x (c - 1), where r is the number of rows and c is the number of columns.

So, the degree of freedom, in this case, is 1.

Now, you will use the chi-squared distribution table to calculate the critical value. Select the value corresponding to the required degrees of freedom and the significance level.

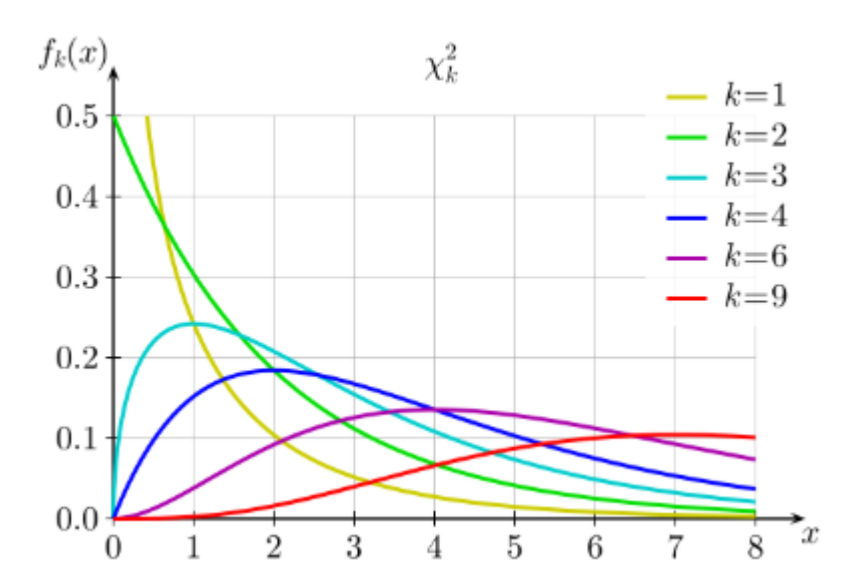
Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.90	61.16
42	22.14	23.65	26.00	28.14	30.77	54.09	58.12	61.78	66.21
46	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.20
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
55	31.73	33.57	36.40	38.96	42.06	68.80	73.31	77.38	82.29
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38

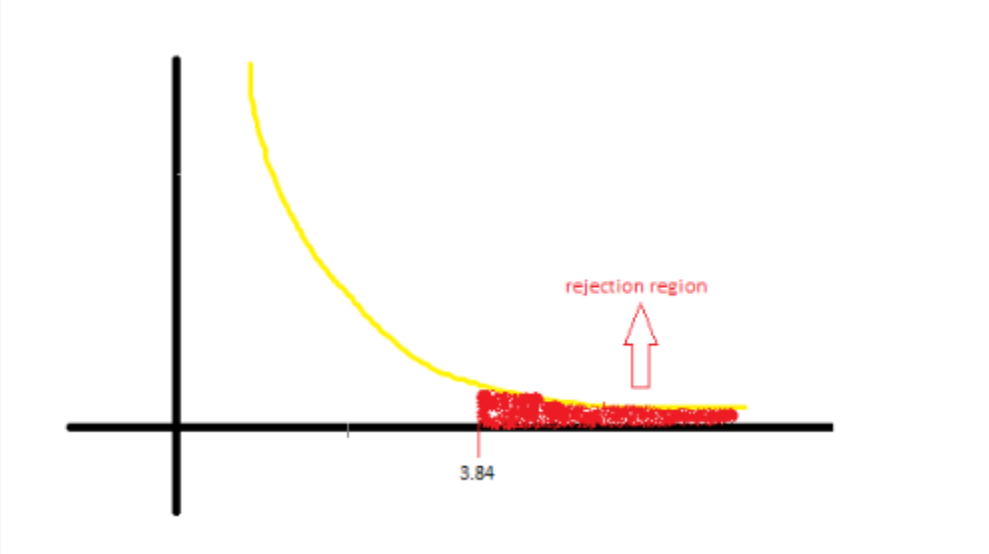
(Source: https://people.smp.uq.edu.au/YoniNazarathy/stat_models_B_course_spring_07/distributions/chisqtab.pdf)

So, the critical value is 3.84, and the test statistic value is 4.

The following is a chi-squared distribution for different values of k (degrees of freedom):



In this case, the test statistic value (4), which is greater than the critical value, lies in the rejection region.



Therefore, you reject the null hypothesis.

Types of Hypotheses

1. A doctor states to a researcher, "A malaria patient takes 10 or fewer days on average to recover."
   What will the null and alternative hypotheses be in this case? (Assume that the average number of days is represented by μ.)

- H_0: μ ≤ 10 days and H_1: μ > 10 days

  ✓ **Correct**

  Feedback:
  Recall the definition of null and alternative hypotheses.
- H_0: μ > 10 days and H_1: μ ≤ 10 days
- H_0: μ ≥ 10 days and H_1: μ < 10 days
- H_0: μ < 10 days and H_1: μ ≥ 10 days

The average growth of a certain variety of bamboo trees is less than or equal to 10.1 inches with a standard deviation of 2.1 inches in three years. A biologist claims that due to climate change, the average growth of bamboo is more than 10.1 inches over a period of three years. To prove this point, the biologist planted 35 bamboo trees and recorded that they had an average three-year growth of 10.8 inches.
Use this information to solve the following questions:

2. Select the appropriate null and alternative hypotheses to test the biologist's claim:

- H_0: μ ≤ 10.1 against H_1: μ>10.1

  ✓ **Correct**

  Feedback:
  The null hypothesis is a premise that has been existing traditionally, so here, traditionally, the average growth of the bamboo trees is less than or equal to 10.1 inches. Thus, the null hypothesis is H_0: μ ≤ 10.1, and the alternative hypothesis is that the growth was caused by a trigger factor. Thus, based on the biologist's claim, the alternative hypothesis is H_1: μ > 10.1.
- H_0: μ =10.1 against H_1: μ≠10.1
- H_0: μ < 10.1 against H_1: μ>10.1
- H_0: μ = 10.1 against H_1: μ<10.1

3. Identify the distribution associated with the example above and the test statistic to be used for the hypothesis testing.

- t-distribution; Z-statistic
- Normal distribution; Z-statistic

  ✓ **Correct**

  Feedback:

  As the sample size is more than 30 and the population standard deviation is known, we identify the distribution as normal and Z-statistic as the statistic.
- t-distribution; t-statistic
- Normal distribution; t-statistic

4. What is the Z-statistic in this case?

- 1.97

  ✓ **Correct**

  Feedback:

  The formula for Z-statistic is $Z = (\bar{x} - \boldsymbol{\mu}) / (\sigma / \sqrt{n})$.

  So, here, $Z = (10.8 - 10.1)/(2.1/\sqrt{35}) = 1.97$
- 2.77
- 0.34
- 1.68

5.Test the hypothesis above at a 5% significance level and state the final result.

Fail to reject the null hypothesis

- Reject the null hypothesis.

  ✓ **Correct**

  Feedback:

  At a 5% significance level, $\boldsymbol{\alpha}$ is equal to .05. Since this is a right-tailed test, the tabular Z-value appears to be 1.64. The calculated Z-statistic is 1.97, which is more than the tabular value 1.64; therefore, you will reject the null hypothesis.

The ideal water used in concrete mixing should not have more than 200 mg/litre of solid organic content on average. So, before any construction work begins, the concerned engineers test samples of water from different water sources to select an appropriate water source, or they use filtered water.

So, the hypotheses for this test are:

H_o: Average organic content ≤ 200 (Water is fit for concrete mixing)

H_1: Average organic content > 200 (Water is unfit for concrete mixing)

Solve the following questions based on the information above:

Errors in Hypothesis Testing

1. Which of the following is a type-1 error?

- The hypothesis test declares the water unfit for construction when, in fact, it is fit.

  ✓ **Correct**

  Feedback:
  In a type-1 error, the Ho is correct, but it is rejected. Hence, if we declare the water unfit, we are rejecting the null hypothesis. But since the water is actually fit, the rejection of the null hypothesis leads to a type-1 error.
- The hypothesis test declares the water unfit for construction when it is actually unfit.
- The hypothesis test declares the water fit for construction when it is actually fit.
- The hypothesis test declares the water fit for construction when, in fact, it isn't fit.

2. Which of the following is a type-2 error?

- The hypothesis test declares the water unfit for construction, and it actually is not fit.
- The hypothesis test declares the water fit for construction when, in fact, it is not fit.

  ✓ **Correct**

  Feedback:

In a type-2 error, the H_0 is incorrect. Thus, if we fail to reject the null hypothesis, we are declaring the water fit for construction. But the water is actually unfit for construction, and hence, this is a type-2 error.

- The hypothesis test declares the water unfit for construction when, in fac t, it is fit.
- The hypothesis test declares the water fit for construction, and it actually is fit.

t-Test

1. According to a study, the daily average time spent by a user on a social media website is 50 minutes. To test the claim of this study, Ramesh, a researcher, takes a sample of 25 website users and finds out that the mean time spent by the sample users is 60 minutes and the sample standard deviation is 30 minutes.
   Based on this information, the null and the alternative hypotheses will be:
   H_0 = The average time spent by the users is 50 minutes
   H_1= The average time spent by the users is not 50 minutes
   Use a 5% significance level to test this hypothesis. Use the t-table from the link here.

   - Reject the null hypothesis
   - Fail to reject the null hypothesis

✓ Correct

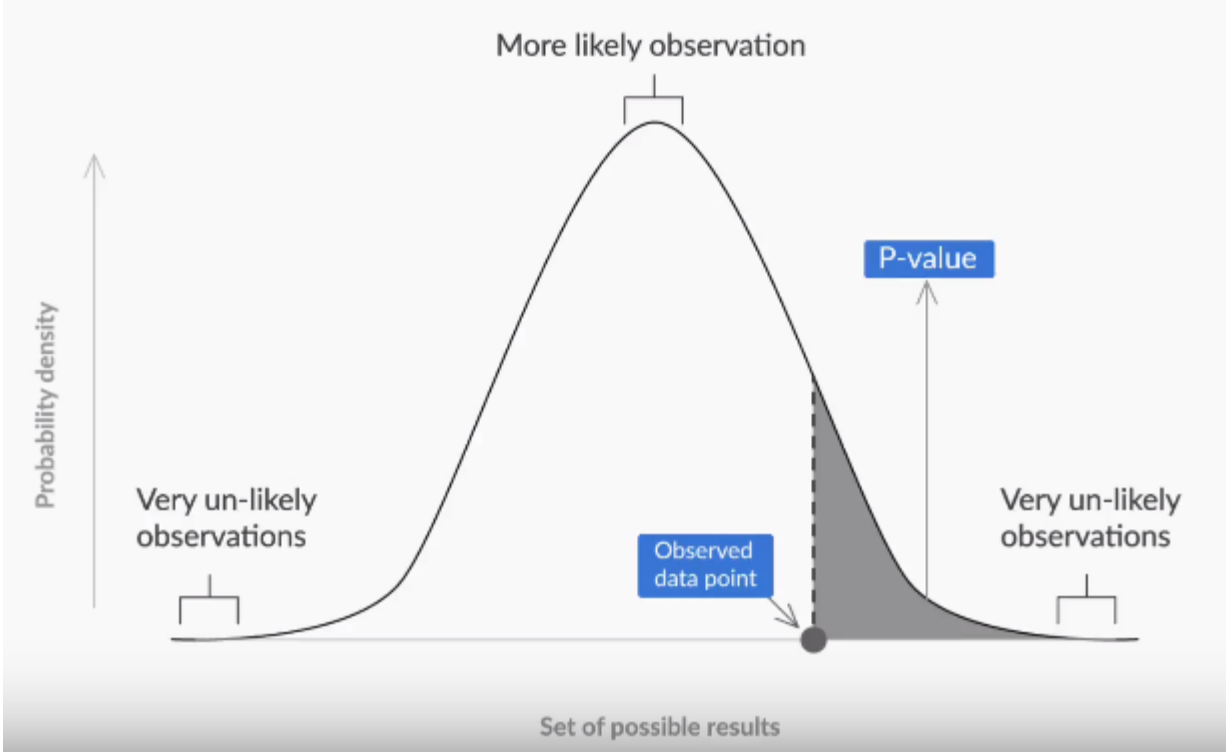
Feedback:

Here, the population standard deviation is not known and the sample size is less than 30; so, we can use the t-test.
Therefore, the t-test statistic (t) = $(x - \mu)/(s/\sqrt{(n)})$ = (60 - 50)/(30/√25) = 1.66.
The sample size is 25. So, the degree of freedom = 25 - 1 = 24.
This is a two-tailed t-test with a significance level of 5%; so, the critical t-value = t 0.05, 24 = 2.064 (using the t-table).
Since the observed t-value is smaller than the critical value, Ramesh fails to reject the null hypothesis.

The p-Value Approach

The concept of p-value is very important in the field of statistics because of one solid advantage it has over the critical value method; you don't have to state the significance level before conducting the hypothesis test in the case of the p-value method. It is easier to understand intuitively whether or not you are going to reject the null hypothesis. In this segment, we will be looking at a very typical problem of testing whether a coin is fair or not using the concept of p-value.

Recall the definition of p-value: It states the probability of observing a similar or more extreme observation, given that the null hypothesis is true.



Interpretation of P-value

Let's try to understand the definition a little better here because you may not have noticed, but this definition allows us to conduct hypothesis testing on distributions that are **not normal** in nature. (In fact, hypothesis testing can be done on non-normal distributions. However, given the concepts that you learnt in the previous sessions, only the p-value method is within the scope of what we can discuss here.)

This method is best explained using an example. This is a very common type of question asked in interviews.

Demonstration

Suppose you toss a coin, the nature of which (whether it is biased or unbiased) you are not aware of. After tossing for 10 times, you observed 8 heads and 2 tails. Now you are asked to test the hypothesis of whether the coin is biased or unbiased. You are also asked to measure the p-value at a 0.05 significance level and make a decision.

Now, the solution methodology for this case may not seem straightforward at first glance, but as a matter of fact, it is quite neat and intuitive.

First, as we always do while conducting a hypothesis test, let's define the null and the alternative hypotheses.

So, what would the null hypothesis be in this case?

Well, according to the question, the null hypothesis of this test is that the coin is unbiased, i.e., $P(H) = P(T) = 0.5$.

And the alternative hypothesis would be $P(H) \neq 0.5$, or $P(T) \neq 0.5$. Observe that both these cases are similar in nature, as both the hypotheses denote the same situation.

Now, let's formally create the null and alternative hypotheses.

$H0 : P(H) = 0.5$
$H1 : P(H) \neq 0.5$
(You can also use P(T) to denote the null and alternative hypotheses in the case above.)

Now, as stated in the problem, we have observed 8 heads.
Recall what the p-value definition states: It is the probability of observing a similar or more extreme observation, given that the null hypothesis is true.
Let's use this definition in our solution methodology to get the answer.
The solution methodology using the definition of p-value would look somewhat like this:

Solution methodology

1. Assume the null hypothesis to be true, i.e., P(H)=0.5.
2. Here, a similar or more extreme observation would be denoted by (Heads ≥ 8) and its probability would be given by P(Heads ≥ 8).
3. Calculate the probability of P(Heads ≥ 8), given that P(H) = 0.5.
4. Observe that the hypothesis-test is two-tailed. Hence, multiply the previous probability by 2. This would be the p-value of this test.

Explanation

First, we assumed that the null hypothesis is true. Then we checked the current observation and tried to deduce what the extreme version of this observation might be from the given null hypothesis.

In the ideal case, we would have got 5 heads. But here, we got 8 heads. Thus, the more extreme versions lie towards 8 or more heads, rather than 8 or fewer heads. This is analogous to the way critical regions are found.

(But you can also say that observing 1 or 2 heads can also be an extreme observation. How do we take that into consideration? You will see how in a short while.)

Step 3 is the most crucial step. Here, we leverage the definition to calculate the p-value. Given that the null hypothesis is true, i.e., $P(H) = 0.5$, we are about to calculate the probability of getting similar or extreme observations, which is the probability as given by $P(\text{Heads} \geq 8)$.

If you observe carefully, you will see that it is equivalent to calculating the probability of observing 8 or more heads in a coin toss experiment where the unbiased coin is flipped 10 times.

Or, the aforementioned problem can be reduced to that of calculating the cumulative probability of a binomial distribution, with $p = 0.5$, $n = 10$ and $r = 8$.

Thus, $P(\text{Heads} \geq 8) = P(X \geq r) = P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10)$

$= {}^{10}C_8 0.5^8 0.5^2 + {}^{10}C_9 0.5^9 0.5^1 + {}^{10}C_{10} 0.5^{10} = 0.055.$

Thus, the probability of $P(\text{Heads} \geq 8)$ is now calculated. Now, note that this would be analogous to a two-tailed test because from the null hypothesis, we can infer that the

extreme observations can occur at both ends, i.e., it can be biased towards the tails or heads. (Take a look at the image above to understand the position of the extreme observations.)

So, we can have observations of 2 or 3 heads as another extreme. What do we do now?

Well, since the binomial distribution is symmetric, we need not do much here; simply multiply the previous value by 2 in a manner similar to how you calculated in the case of a normal distribution.

And voila! We have the p-value as $2 * P(\text{Heads} \geq 8) = 2 * 0.055 = 0.11$.

Given the significance level of 0.05 and the calculated p-value, we can safely say that we fail to reject the null hypothesis.

Now, try to answer the following question to understand an alternative approach to solving this problem. You can learn more about this method here.

The p-Value Approach

Let's say that you want to calculate the p-value using P(T) = 0.5 as the null hypothesis. How would your approach the change here? Does the solution, i.e., the p-value, change? Please write the answer below. Use the step-wise methodology mentioned above.

.....

F-Test

Two sample t-tests can validate a hypothesis containing only two groups at a time. For samples involving three or more groups, the t-test becomes tedious, as you have to perform the tests for each combination of the groups. Also, the possibility of a **type-1 error increases in this process**. You use ANOVA in such cases.

Analysis of variance (ANOVA) can determine whether the means of three or more groups are different. ANOVA uses F-tests to statistically test the equality of means.
To understand how ANOVA is applied, let's go over a simple case:
A test was conducted at a workplace, and the feedback on the three e-commerce platforms was recorded in a data set, which is as follows:

Amazon	**Flipkart**	**Snapdeal**
7.5	7	5
8.5	9.5	7.5
6	10	8.5
10	6	3
8.5	7.5	6
8	8.5	5
8	10	7
6	6.5	
9.5	6.5	
10	9	
6.5	10	

To begin with, create a null hypothesis (H_0) for your ANOVA test. In this case, your null hypothesis will be: “All the platforms are equally popular”. The alternate hypothesis (), thus, becomes “At least one of the platforms has different popularity from the rest”. Represent this information as:

H0: $\boldsymbol{\mu}1 = \boldsymbol{\mu}2 = ... = \boldsymbol{\mu}k$ (where k is the number of different **populations** or **groups** or **treatment levels**, in your case, it’s 3). By writing this, you suggest that the ‘mean’ of the different **populations** will be the same, which is your null hypothesis. If the statement above is proved at the end of your test, it will imply that all the platforms are equally popular. If not, then you accept your alternative hypothesis (HA).

There are a couple of things you should keep a note of while using ANOVA:

- You must be thinking that it is a fairly simple problem for us. You calculate the means of the three groups and compare them to reject or not reject the null hypothesis. Unfortunately, it is not that simple after all because your hypothesis considers the mean of a particular ‘population’, while your data set only has a ‘sample’ of that ‘population’. So, the mean that you calculate will be of the sample and not of the population. For instance, in your case, the people who have given feedback for, say, Amazon, are not the only ones who have used Amazon. There are many others too. But they form your sample, which is why the mean that you will calculate here will be of this ‘sample’ and not of the ‘population’.
- Another question that might crop up in your mind is: Why is the process called analysis of ‘variance’ when you are comparing ‘means’? This is because the

math that you will use later in the process will require the concept of variance to study the means of the groups. It will tell you how the means vary or differ.

Now, as variance is the central idea behind ANOVA, let's briefly revisit the topic:

1. Variance is the average squared deviation of a datapoint from the distribution mean. The distance between the sample mean and each datapoint is measured and squared. Then, you add them together and take the average. The formula is:

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

Here, s2 represents the variance, x represents the sample datapoints, x- bar represents the sample mean, and n represents the number of sample points.

2. If you momentarily ignore the average part, what you are left with is the 'sum of squares'. So, the sum of squares is the variance without finding the average of the sum of squared deviations.

The sum of squares is given by:

$$SS = \sum(x - \bar{x})^2$$

Let’s have a look at the dataset again:

Amazon	Flipkart	Snapdeal
7.5	7	5
8.5	9.5	7.5
6	10	8.5
10	6	3
8.5	7.5	6
8	8.5	5
8	10	7
6	6.5	
9.5	6.5	
10	9	
6.5	10	

We have talked about two kinds of calculations that you have to make in accordance with the variance. ‘Sum of squares between’ accounts for the variation between the groups, and ‘sum of squares within’ accounts for the variation within a group. The total sum of squares is the sum of all the variations that there are, and it gives us the deviation of each observation from the grand mean of the dataset. To understand this more clearly, let’s look at your case:

- **SSB** represents the variation of the mean feedback of a company, say Flipkart, from the grand mean of all the feedback.
- **SSW** represents the variation of all the feedback in a company from the mean of its feedback.

- **TSS** represents the variation of all the feedback in your dataset from the grand mean.

Let's look at the basic formula you will be using:

Total sum of squares = Sum of squares between + sum of squares within the group

(TSS) (SSB) (SSW)

$$SS_{total} = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \overline{x})^2$$

$$SS_{between} = \sum_{j=1}^{p} n_j(\overline{x}_j - \overline{x})^2$$

$$SS_{within} = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \overline{x}_j)^2$$

(Image source: https://www.easycalculation.com/formulas/eta-squared-formula.html)

Here, 'i' represents the **observations in a group** or a **treatment level**, and '**j**' refers to a **particular group** or a **treatment level**. In your case, 'i' will represent all the feedback of, say, Amazon, and 'j' refers to a particular group and can be Amazon, Flipkart, or Snapdeal.

n_j : It represents the number of observations in a group. In your case, it will be the number of times feedback is received for, say, Snapdeal.

x_{ij} : It represents all the observations that have been recorded in the dataset.

$\overline{x}_j$: It represents the mean of a particular group or treatment.

$\overline{x}$: It represents the grand mean of all the observations. In your case, this will be the mean of all the feedback that has been collected.

Let's now calculate the aforementioned measures for your data:

	Amazon	Flipkart	Snapdeal	
	7.5	7	5	
	8.5	9.5	7.5	
	6	10	8.5	
	10	6	3	
	8.5	7.5	6	
	8	8.5	5	
	8	10	7	
	6	6.5		
	9.5	6.5		
	10	9		
	6.5	10		
n(j)	11	11	7	
x(j)-bar	8.045454545	8.227272727	6	
x-bar				7.620689655
n(j)*((x(j)-bar)-(x-bar))	1.984677332	4.047373257	18.38644471	
SSB	24.4184953			
SSW	66.40909091			

After you have calculated this data, the next step is to analyse the ANOVA table:

ANOVA

	Sum of Squares	df	Mean Square	F
Between Groups				
Within Groups				
Total				

You have already calculated the sum of squares. Now, 'df' here represents the degrees of freedom.

- Between groups, df = Number of groups - 1
- Within a group, df = Total number of observations - Number of groups
- Degrees of freedom for the complete dataset = Total number of observations - 1

Let's calculate the degrees of freedom for your observations:

Source of Variation	Sum of Squares	DOF	Mean Square	F Ratio
Between	24.4184953	2		
Within	66.40909091	26		
Total	90.82758621	28		

Mean square = Sum of squares/df

Using this formula, you can find the mean square between the groups as well as within the group.

Let’s calculate the mean squares for your calculations:

Source of Variation	Sum of Squares	DOF	Mean Square	F Ratio
Between	24.4184953	2	12.20924765	
Within	66.40909091	26	2.554195804	
Total	90.82758621	28		

Before moving any further, let’s first see what the F-test is.

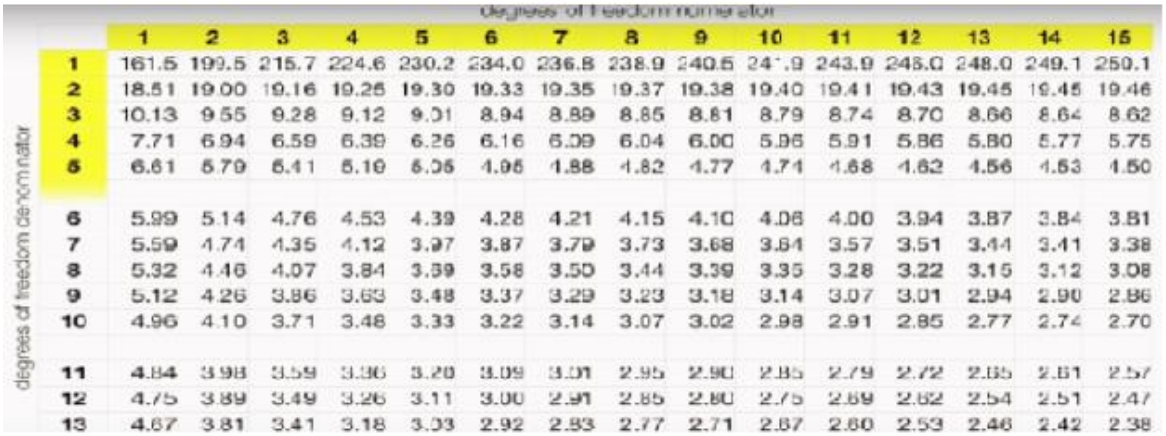
F-tests are named after the test statistic F, which was named in honour of Sir Ronald Fisher. The F-statistic is simply a ratio of two variances.

To use the F-test to determine whether group means are equal, all you need to do is include the correct variances in the ratio. In one-way ANOVA, the F-statistic is given by this ratio:

F = Variation between the sample means / variation within the samples

= (MSB/MSW)

Now, you have to calculate the critical F-value using the F-distribution table for a given significance level and compare it with your calculated F-value. In your case, $p < 0.05$. The table looks like this:



degrees of freedom denominator	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	161.5	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.0	248.0	249.1	250.1
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38

Degrees of freedom of the numerator will be that of the df between the groups. Degrees of freedom of the denominator will be that of the df within the group.

The intersection will give us the critical F-value. Now, you compare your calculated F-value with the critical F-value.

- **If the calculated F < the critical F, you will fail to reject the null hypothesis.**
- **If the calculated F > the critical F, you will reject the null hypothesis.**

Let's now do the final calculations in your case to see whether Amazon, Flipkart and Snapdeal are equally popular:

Source of Variation	Sum of Squares	DOF	Mean Square	F Ratio
Between	24.4184953	2	12.20924765	4.780075055
Within	66.40909091	26	2.554195804	
Total	90.82758621	28		

Your F critical value is 3.3690, and your calculated value comes out to be 4.78.

Therefore, you will reject the null hypothesis and accept the alternative hypothesis.

Therefore, Amazon, Flipkart and Snapdeal are not equally popular.

ANOVA

1. In an Analysis of Variance (ANOVA) problem, the total sum of squares is 150 and the sum of squares error is 50. What will the value in the sum of squares column be?
   - 200
   - 75
   - 100

     ✓ **Correct**

     Feedback:
     Total sum of squares = Sum of squares column + Sum of squares error
   - 125

2. In an Analysis of Variance (ANOVA) problem, there are four groups with 10 observations in each group. If the total sum of squares is 278.4 and the sum of square error is 56.2, what will the value in the mean square column be?

   - 111.1
   - 74.06

     ✓ **Correct**

     Feedback:
     Calculate the sum of squares column using the following equation: Total sum of squares = Sum of squares column + Sum of squares error. The degrees of freedom column = Number of columns - 1. Now, the mean squares column is = Sum of squares column / Degrees of freedom column.
   - 6.17
   - 5.69

3. From a dataset, we have extracted some information:

The total sum of squares (TSS) = 2784;

The sum of squares error (SSE) = 970;

The total number of observations = 30; and

H0: μ1 = μ2 = μ3 = μ4.

What is the total mean square value?

- 604.667
- 37.30
- 96

  ✓ Correct

  Feedback:
  Find the total degrees of freedom to calculate the mean square value.
- 928

F-Ratio

1. In the previous question, what is the value of the F-ratio?

- 16.21

✓ Correct

Feedback:
Calculate the sum of squares column using the formula 'Total sum of squares = Sum of squares column + Sum of square error'. Now, calculate the degrees of freedom. The degrees of freedom column is represented by 'The number of columns - 1'. Moreover, the degree of freedom error = total number of observations - number of columns. Now, find the mean squares from the sum of squares and the degrees of freedom. Calculate the F-ratio by dividing the mean squares column by the mean square error.

- 6.29
- 18.45
- 5.80