Module-IV

STATISTICAL INFERENCE-2

Topic Learning Objectives:

Upon Completion of this module, student will be able to:

- Solve problems on probability distribution functions two variables.
- Use statistical methodology and tools in the engineering problem-solving process
- Compute the confidence intervals for the mean of the population.

Sampling Variables

Sampling variables refers to the process of selecting data points or observations from a larger population or dataset for the purpose of analysis, experimentation, or research. It is a fundamental concept in statistics and data analysis. When you sample data, you are essentially taking a subset of the entire population to draw conclusions or make inferences about the entire population. Here are some key points related to sampling variables:

Population: The population refers to the entire set of individuals, items, or data points that you are interested in studying. Its often impractical or impossible to study an entire population, so you sample from it.

Sample: A sample is a subset of the population. It consists of a smaller number of data points or observations that are chosen in a way that they represent the larger population to some extent.

Sampling Methods: There are various methods for sampling data, including simple random sampling (each data point has an equal chance of being selected), stratified sampling (dividing the population into subgroups and then sampling from each subgroup), systematic sampling (selecting every nth data point), and more.

Sampling Error: When you take a sample from a population, there is a chance that the sample may not perfectly represent the population. This difference between the sample and the population is called sampling error.

Parameter and Statistic: In statistical analysis, a parameter is a characteristic of the population, while a statistic is a characteristic of the sample.

For example, the mean of a population is a parameter, while the mean of a sample is a statistic.

Sampling Size: The number of data points or observations you include in your sample is known as the sample size. A larger sample size generally provides more accurate estimates of population parameters.

Central limit theorem:

The theorem which explains this sort of relationship between the shape of the population distribution and the sampling distribution of the mean is known as the central limit theorem. This theorem is by far the most important theorem in statistical inference. It assures that the sampling distribution of the mean approaches normal distribution as the sample size increases. In formal terms, we may say that the central limit theorem states that the distribution of means of random samples taken from a population having mean μ and finite variance σ^2 approaches the normal distribution with mean μ and variance σ^2/n as n goes to infinity.

If $\bar{x}$ is the mean of random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of $Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$, as $n \to \infty$ is the standard normal distribution N(Z;0,1).

The significance of the central limit theorem lies in the fact that it permits us to use sample statistics to make inferences about population parameters without knowing anything about the shape of the frequency distribution of that population other than what we can get from the sample.

Confidences limit for unknown mean.

Let $\bar{x}$ be the sample mean, and n be the size of the sample. Then the interval estimate of the population mean μ is given by $\bar{x} \pm t_\alpha \frac{s}{\sqrt{n}}$

Problems:

1. A sample of size 9 from a normal population gave $\bar{x} = 15.8$ and $s^2 = 10.3$. Find a 99% interval for population mean.

Solution: Given $\bar{x} = 15.8$, $s^2 = 10.3$ and $n = 9$.

Degrees of freedom$= n - 1 = 8$

Also $t_\alpha = t_{0.01}$ for 8 d.f =3.36

99% confidence limit for the population mean $\bar{x}$ are $\bar{x} \pm t_{0.01} \frac{s}{\sqrt{n}}$

$$=15.8 \pm 3.36 \sqrt{\frac{10.3}{9}}$$

$$=12.2055, 19.3944.$$

Hence 99% confidence interval = [12.2055,19.3944].

2. A random sample of 15 observations has a mean of 20 and a standard deviation of 3.5. To estimate the population mean with 95% confidence level determine the confidence interval.

Solution: Given $\bar{x} = 20$, $s = 3.5$ and $n = 15$.

Degrees of freedom$= n - 1 = 14$

Also $t_\alpha = t_{0.05}$ for 14 d.f =2.145

95% confidence limit for the population mean $\bar{x}$ are $\bar{x} \pm t_{0.05} \frac{s}{\sqrt{n}}$

$$=20 \pm 2.145 \sqrt{\frac{3.5^2}{14}}$$

$$=18.06, 21.94.$$

Hence 95% confidence interval = [18.06,21.94].

A discussion on tests of significance for small samples

So far the problem of testing a hypothesis about a population parameter was based on the assumption that sample drawn from population is large in size (more than 30) and the probability distribution is normally distributed. However, when the size of the sample is small, (say < 30) tests considered above are not suitable because the assumptions on which they are based generally do not hold good in the case of small samples. In particular, here one cannot assume that the problem follows a normal distribution function and those values given by sample data are sufficiently close to the population values and can be used in their place for the calculation of standard error. Thus, it is a necessity to develop some alternative strategies to deal with problems having sample size relatively small. Also, we do see a number of problems involving small samples. With these in view, here, we will initiate a detailed discussion on the same.

Here, too, the problem is about testing a statement about population parameter; i.e. in ascertaining whether observed values could have arisen by sampling fluctuations from some value given in advance. For example, if a sample of 15 gives a correlation coefficient of +0.4, we shall be interested not so much in the value of the correlation in the parent population, but more generally this value could have come from an un – correlated population, i.e. whether it is significant in the parent population. It is widely accepted that when we work with small samples, estimates will vary from sample to sample.

Further, in the theory of small samples also, we begin study by assuming that parent population is normally distributed unless otherwise stated. Strictly, whatever the decision one takes in hypothesis testing problems is valid only for normal populations. Sir William Gosset and R. A. Fisher have contributed a lot to theory of small samples. Sir W. Gosset published his findings in the year 1905 under the pen name "student". He gave a test popularly known as "t – test" and Fisher gave another test known as "z – test". These tests are based on "t distribution and "z – distribution".

Test of Significance for means of two small samples by Student's t - distribution

Procedures to be followed for testing of significance for means of two small samples

1. Null Hypothesis: $H_0: \mu_1 = \mu_2$ There is no significant difference in the means.
   Alternate Hypothesis: $H_1: \mu_1 \neq \mu_2$
2. Calculation of Test Statistic.
   Estimated standard deviation:

$$S = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 1}}$$

$$\text{Test Statistic } t = \frac{\bar{x_1} - \bar{x_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

3. Level of significance: Take the level of significance $\alpha = 0.05$ if α is not known.
4. Decision: Accept H_0 if computed $t \leq$ tabled t_α
   Reject H_0 if computed $t >$ tabled t_α.

Problems:

1. The average number of articles produced by two machines per day are 200 and 250 with standard deviations 20 and 25 respectively on the basis of records of 25 days production. Can you regard both the machines equally efficient at 1% level of significance?

Solution: Given $n_1 = 25, \overline{x_1}$=200, $s_1 = 20$

$n_2 = 25, \overline{x_2}$=250, $s_2 = 25$

Assume Null Hypothesis: $H_0: \mu_1 = \mu_2$ ie., both the machines are equally efficient.

Alternate Hypothesis: $H_1: \mu_1 \neq \mu_2$

Estimated standard deviation:

$$S = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 1}} = 23.1$$

Test Statistic t= $\left| \frac{\overline{x_1} - \overline{x_2}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \right| = \frac{200-250}{23.1\sqrt{\frac{1}{25}+\frac{1}{25}}} = |-7.7| = 7.7$

The table value of t at 1% level of significance $t_{0.01,48}$ is 2.58

Calculated value > Tabulated value,

Hence reject the Null hypothesis.ie., The two machines are not equally efficient at 1% level of significance.

2. Two salesman A and B are working in a certain district. From a sample survey conducted by the Head Office, the following results were obtained. State whether there is any significant difference in the average sales between the two salesmen?

	A	B
No. of Sales	20	18
Average sales (in Rs.)	170	205
Standard Deviation (in Rs.)	20	25

Solution:

Given $n_1 = 20, \overline{x_1}$=170, $s_1 = 20$

$n_2 = 18, \overline{x_2}$=205, $s_2 = 25$

Assume Null Hypothesis: $H_0: \mu_1 = \mu_2$

i.e., There is no significant difference in the average between the two salesmen.

Alternate Hypothesis: $H_1: \mu_1 \neq \mu_2$

Estimated standard deviation:

$S = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 1}} = 23.12$

Test Statistic t= $\left|\frac{\overline{x_1} - \overline{x_2}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right| = \frac{170-205}{23.12\sqrt{\frac{1}{20}+\frac{1}{18}}} = 4.73$

The table value of t at 5% level of significance $t_{0.05,36}$ is 1.96

Calculated value > Tabulated value,

Hence reject the Null hypothesis.ie., There is a significant difference in the average between the two salesmen.

3. The mean life of a sample of 10 electric bulbs was found to be 1456 hours with a standard deviation of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1280 hours with standard deviation 398 hours. Is there significant difference between the means of the two batches?

Solution: Given $n_1 = 10, \overline{x_1}$=1456, $s_1 = 423$

$n_2 = 17, \overline{x_2}$=1280, $s_2 = 398$

Assume Null Hypothesis: $H_0: \mu_1 = \mu_2$ ie., There is no significant difference in the means of two samples.

Alternate Hypothesis: $H_1: \mu_1 \neq \mu_2$

Estimated standard deviation:

$S = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 1}} = 423.42$

Test Statistic t= $\left|\frac{\overline{x_1} - \overline{x_2}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right| = \frac{1456-1280}{423.42\sqrt{\frac{1}{10}+\frac{1}{17}}} = 1.04$

The table value of t at 5% level of significance $t_{0.05,25}$ is 2.06

Calculated value < Tabulated value,

Hence accept the Null hypothesis.ie., There is no significant difference in the means of two samples.

4. Two types of batteries are tested for their lengths of life and the following data are obtained.

	No. of Samples	Mean Life	Variance
Type A	9	600 hours	121
Type B	8	640 hours	144

Is there significance difference in the two means? Value of t for 15 degrees of freedom at 5% level is 2.131.

Solution: Given $n_1 = 9, \overline{x_1}$=600, $s_1^2 = 121$

$n_2 = 8, \overline{x_2}$=40, $s_2^2 = 144$

Assume Null Hypothesis: $H_0: \mu_1 = \mu_2$ ie., There is no significant difference in the two means.

Alternate Hypothesis: $H_1: \mu_1 \neq \mu_2$

Estimated standard deviation:

$$S = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 1}} = 12.22$$

Test Statistic t= $\left| \frac{\overline{x_1} - \overline{x_2}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \right| = \frac{600-640}{12.22\sqrt{\frac{1}{9}+\frac{1}{8}}} = 6.73$

The table value of t at 5% level of significance $t_{0.05,15}$ is 2.131

Calculated value > Tabulated value,

Hence reject the Null hypothesis.ie., There is a significant difference in the two means.

Student's t - distribution function

Gosset was employed by the Guinness and Son, Dublin bravery, Ireland which did not permit employees to publish research work under their own names. So Gosset adopted the pen name "student" and published his findings under this name. Thereafter, the t – distribution commonly called student's t – distribution or simply student's distribution.

The t – distribution to be used in a situation when the sample drawn from a population is of size lower than 30 and population standard deviation is un – known. The t – statistic, t_{cal} is defined as $t_{cal} = \left(\frac{\bar{x}-\mu}{S}\right) \times \sqrt{n}$ where $S = \sqrt{\frac{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}{n-1}}$, $\bar{x}$ is the sample mean, **n** is the sample size, and x_i are the data items.

The t – distribution function has been derived mathematically under the assumption of a normally distributed population; it has the following form $f(t) = C\left(1+\frac{t^2}{\gamma}\right)^{-\left(\frac{\gamma+1}{2}\right)}$ where C is a constant term and γ = n - 1 denotes the number of degrees of freedom. As the p.d.f. of a t –

distribution is not suitable for analytical treatment. Therefore, the function is evaluated numerically for various values of t, and for particular values of γ . The t – distribution table normally given in statistics text books gives, over a range of values of γ , the probability values of exceeding by chance value of t at different levels of significance. The t – distribution function has a different value for each degree of freedom and when degrees of freedom approach a large value, t – distribution is equivalent to normal distribution function.

The application of t – distribution includes (i) testing the significance of the mean of a random sample i.e. determining whether the mean of a sample drawn from drawn from a normal population deviates significantly from a stated value (i.e. hypothetical value of the populations mean) and (ii) testing whether difference between means of two independent samples is significant or not i.e. ascertaining whether the two samples comes from the same normal population? (iii) Testing difference between means of two dependent samples is significant? (iv) Testing the significance of on observed correlation coefficient.

Procedures to be followed in testing a hypothesis made about the population parameter using student's t - distribution:

- As usual first set up null hypothesis,
- Then, set up alternate hypothesis,
- Choose a suitable level of significance,
- Note down the sample size, n and the number of degrees of freedom,
- Compute the theoretical value, t_{tab} by using t – distribution table.
- t_{tab} value is to be obtained as follows: If we set up $\alpha = 5\% = 0.05$, suppose $\gamma = 9$ then, t_{tab} is to be obtained by looking in 9th row and in the column $\alpha = 0.025$ (i.e. half of α = 0.05) .
- The test criterion is then calculated using the formula, $t_{cal} = \left(\frac{\bar{x}-\mu}{S}\right)\times\sqrt{n}$
- Later, the calculated value above is compared with tabulated value. As long as the calculated value matches with the tabulated value, we as usual accept the null hypothesis and on the other hand, when the calculated value becomes more than tabulated value, we reject the null hypothesis and accept the alternate hypothesis.

Problems:

1. The manufacturer of a certain make of electric bulbs claims that his bulbs have a mean life of 25 months with a standard deviation of 5 months. Random samples of 6 such bulbs have the following values: Life of bulbs in months: 24, 20, 30, 20, 20, and 18. Can you regard the producer's claim to valid at 1% level of significance? (Given that $t_{tab} = 4.032$ corresponding to $\gamma = 5$).

Solution: To solve the problem, we first set up the null hypothesis $H_0 : \mu = 25$ months, alternate hypothesis may be treated as $H_0 : \mu < 25$ months. To set up $\alpha = 1\%$, then tabulated value corresponding to this level of significance is $t_{tab}|_{\alpha=1\% \text{ and } \gamma=5} = 4.032$ (4.032 value has been got by looking in the 5th row) . The test criterion is given by

$t_{cal} = \left(\frac{\bar{x}-\mu}{S}\right)\times\sqrt{n}$ where $S = \sqrt{\frac{\sum_{i=1}^{i=n}\left(x_i-\bar{x}\right)^2}{n-1}}$.

Consider

x_i	$\bar{x}$	$x_i - \bar{x}$	$\left(x_i - \bar{x}\right)^2$
24		1	1
26		3	9
30		7	49
20	**23**	-3	9
20		-3	9
18		-5	25
Total = 138		-	**Total = 102**

Thus, $S = \sqrt{\frac{102}{5}} = \sqrt{20.4} = 4.517$ and $t_{cal} = \left|\frac{23-25}{4.517}\right|\sqrt{6} = 1.084$. Since the calculated value, 1.084 is lower than the tabulated value of 4.032; we accept the null hypothesis as mean life of bulbs could be about 25 hours.

2. A certain stimulus administered to each of the 13 patients resulted in the following increase of blood pressure: 5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, 6, 8. Can it be concluded that the stimulus, in general, be accompanied by an increase in the blood pressure?

Solution: We shall set up $H_0: \mu_{before} = \mu_{after}$ i.e. there is no significant difference in the blood pressure readings before and after the injection of the drug. The alternate hypothesis is $H_0: \mu_{before} > \mu_{after}$ i.e. the stimulus resulted in an increase in the blood pressure of the patients. Taking $\alpha = 1\%$ and $\alpha = 5\%$, as n = 13, $\gamma = n - 1 = 12$, respective tabulated values are $t_{tab}\big|_{\alpha=1\% \text{ and } \gamma=12} = 3.055$ and $t_{tab}\big|_{\alpha=5\% \text{ and } \gamma=12} = 2.179$. Now, we compute the value of test criterion. For this, consider

x_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
5		2	4
2		-1	1
8		5	25
-1		-4	16
3		0	0
0		-3	9
-2		-5	25
1	**3**	-2	4
5		2	4
0		-3	9
4		1	1
6		3	9
8		5	25
Total = 39		-	**Total = 132**

Consider $S=\sqrt{\frac{\sum_{i=1}^{i=n}\left(x_i-\bar{x}\right)^2}{n-1}}=\sqrt{\frac{132}{12}}=\sqrt{11}=3.317$. Therefore, $t_{cal}=\left|\frac{\bar{x}-\mu}{S}\right|\times\sqrt{n}$ may be obtained as $t_{cal}=\left|\frac{0-3}{3.317}\right|\sqrt{13}=3.2614$. As the calculated value 3.2614 is more than the tabulated values of 3.055 and 2.179, we accept the alternate hypothesis that after the drug is given to patients, there is an increase in the blood pressure level.

3. the life time of electric bulbs for a random sample of 10 from a large consignment gave the following data: 4.2, 4.6, 3.9, 4.1, 5.2, 3.8, 3.9, 4.3, 4.4, 5. 6 . Can we accept the hypothesis that the average life time of bulbs is 4, 000 hours?

Solution: Set up H_0:μ=4,000 hours , H_1:$\mu<4,000$ hours . Let us choose that $\alpha=5\%$. Then tabulated value is $t_{tab}|_{\alpha=5\% \text{ and } \gamma=9}=2.262$. To find the test criterion, consider

x_i	$\bar{x}$	$x_i-\bar{x}$	$\left(x_i-\bar{x}\right)^2$
4.2	**4.4**	-0.2	0.04
4.6		0.2	0.04
3.9		-0.5	0.25
4.1		-0.3	0.09
5.2		0.8	0.64
3.8		-0.6	0.36
3.9		-0.5	0.25
4.3		-0.1	0.01
4.4		0.0	0.0
5.6		1.2	1.44
Total = 44		-	**Total = 3.12**

Consider $S=\sqrt{\frac{\sum_{i=1}^{i=n}\left(x_i-\bar{x}\right)^2}{n-1}}=\sqrt{\frac{3.12}{9}}=0.589$. Therefore, $t_{cal}=\left|\frac{\bar{x}-\mu}{S}\right|\times\sqrt{n}$ is computed as $t_{cal}=\left|\frac{4.4-4.0}{0.589}\right|\cdot\sqrt{10}$ = 2.148. As the computed value is lower than the tabulated value of 2.262, we conclude that mean life of time bulbs is about 4, 000 hours.

4. Consider the sample consisting of nine numbers 45, 47, 50, 52, 48, 47, 49, 53 and 51. The sample is drawn from a population whose mean is 47.5. Find whether the sample mean differs significantly from the population mean at 5% level of significance.

Solution: for the given sample, the size is N=9. Therefore its mean is

$$\bar{X} = \frac{1}{9}(45 + 47+50+5\,2 +48+47 + 49 +53 + 51) = 49.11$$

And the variance is

$$S^2 = \frac{1}{9}\{(45-49.11)^2 + (47-49.11)^2 + (50-49.11)^2 + (52-49.11)^2 + (48-49.11)^2 + (47-49.11)^2 + (49-49.11)^2 + (53-49.11)^2 + (51-49.11)^2\}$$

$$=6.0988$$

So that the standard deviation is s = $\sqrt{6.0988}$ = 2.47.

Since N = 9, we have $\gamma = 8$ for which we find from the table that $t_{0.05}$ = 2.31

With $\mu = 47.5$, $\bar{X}$= 49.11 and s= 2.47,

we have

$$t = \left(\frac{\bar{X}-\mu}{s}\right)\sqrt{\gamma} = \frac{49.11-47.5}{2.47}\times\sqrt{8} = 1.844.$$

Thus, here the t- score is less than $t_{0.05}(\gamma)$ = 2.31. Accordingly, the difference between the sample mean and the population is not significant at 0.05 level of significance.

5. Eleven school boys were given a test in mathematics carrying a maximum of 25 marks. They were given a month's extra coaching and a second test of equal difficulty was held thereafter. The following table gives the marks in the two tests.

Boy	1	2	3	4	5	6	7	8	9	10	11
I Test Marks	23	20	19	21	18	20	18	17	23	16	19
II Test Marks	24	19	22	18	20	22	20	20	23	20	17

Do the marks given evidence that the students have benefitted by extra coaching? Use 0.05 level of significance.

Solution: We first calculate the mean and the standard deviation in the difference in marks in the two tests.

We note that the difference in marks(marks in II test – marks in I test) are

1, -1, 3, -3, 2, 2, 2, 3, 0, 4, -2.

The mean of these differences is

$$\bar{X} = \frac{1}{11}(1 - 1 + 3 - 3 + 2 + 2 + 2 + 3 + 0 + 4 - 2) = 1$$

And the variance is

$$s^2 = \frac{1}{11}\{(1-1)^2 + (-1-1)^2 + (3-1)^2 + (-3-1)^2 + (2-1)^2 + (2-1)^2 + (2-1)^2 + (3-1)^2 + (0-1)^2 + (4-1)^2 + (-2-1)^2\}$$

$$= \frac{1}{11}(0 + 4 + 4 + 16 + 1 + 1 + 1 + 4 + 1 + 9 + 9) = \frac{50}{11} = 4.545,$$

So that the standard deviation is $s = \sqrt{4.545} = 2.13$.

Since N = 11, we have $\gamma = 10$ for which we find from table that $t_{0.05} = 2.23$.

Now, we make hypothesis that the students have not been benefitted by extra coaching. That is, the difference in mean marks μ is Zero. Under this Hypothesis, the t – score is

$$t = \frac{\bar{X}-\mu}{s}\sqrt{\gamma} = \frac{1-0}{2.13}\sqrt{10} = 1.485.$$

We note that this t- score is less than $t_{0.05}(\gamma) = 2.23$. Hence, we do not reject the hypothesis at 0.05 level of significance. This means that it is likely that the students have not been benefitted by extra coaching.

6. Two horses A and B were tested according to the time (in seconds) to run a particular race with the following results.

Horse A: 28 30 32 33 33 29 34
Horse B: 29 30 30 24 27 29

Test whether you can discriminate between the two horses. ($t_{0.05}$=2.2 for 11 d.f.)

Solution: Let the variables x and y respectively correspond to horse A and horse B.

$$\bar{x} = \frac{\sum x}{n_1} = \frac{219}{7} = 31.3$$

$$\bar{y} = \frac{\sum y}{n_2} = \frac{169}{6} = 28.2$$

$\sum(x - \bar{x})^2 = 31.43 \quad \sum(y - \bar{y})^2 = 26.84$

$$s = \sqrt{\frac{\langle\sum(x-\bar{x})^2 + (y-\bar{y})^2\rangle}{n_1 + n_2 - 2}} = 2.30$$

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 2.42 > 2.2$$

Therefore, hypothesis rejected at 5% level of significance.

Discussion on χ^2 test and Goodness of Fit

In above section, we have discussed t – distribution function (i.e. t – test). The study was based on the assumption that the samples were drawn from normally distributed populations, or, more accurately that the sample means were normally distributed. Since test required such an assumption about population parameters. For this reason, A test of this kind is called parametric test. There are situations in which it may not be possible to make any rigid assumption about the distribution of population from which one has to draw a sample.

Thus, there is a need to develop some non – parametric tests which does not require any assumptions about the population parameters.

With this in view, now we shall consider a discussion on χ^2 distribution which does not require any assumption with regard to the population. The test criterion corresponding to this distribution may be given as $\chi^2 = \frac{\sum_i (O_i - E_i)^2}{E_i}$ where O_i : Observed values ,

E_i : Expected values.

The calculated χ^2 value (i.e. test criterion value or calculated value) is compared with the tabular value of χ^2 value for given degree of freedom at a certain prefixed level of significance. Whenever the calculated value is lower than the tabular value, we continue to accept the fact that there is not much significant difference between expected and observed results. On the other hand, if the calculated value is found to be more than the value suggested in the table, then we have to conclude that there is a significant difference between observed and expected frequencies.

As usual, degrees of freedom are $\gamma = n - k$ where k denotes the number of independent constraints. Usually, it is 1 as we will be always testing null hypothesis against only one hypothesis, namely, alternate hypothesis.

This is an approximate test for relatively a large population. For the usage of test, the following conditions must checked before employing the test. These are:

1. The sample observations should be independent.
2. Constraints on the cell frequencies, if any, must be linear.

3. i.e. the sum of all the observed values must match with the sum of all the expected values.
4. **N,** total frequency should be reasonably large
5. No theoretical frequency should be lower than 5.
6. It may be recalled this test is depends on χ^2 test: The set of observed and expected frequencies and on the degrees of freedom, it does not make any assumptions regarding the population.

Problems:

1. The following table gives the number of road accidents that occurred in a large city during the various days of a week. Test the hypothesis that the accidents are uniformly distributed over all the days of a week.

Day	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Total
No. of accidents	14	16	8	12	11	9	14	84

Solution: under the hypothesis that the accidents on each day are uniformly distributed over the week, the expected number of accidents on each day are 12. (because a total of N = 84 accidents have occurred in 7 days).

Thus, her, the expected frequencies are 12 each observed frequencies are the number of accidents shown in the given table.

Using these, we find that

$$\chi^2 = \frac{(14-12)^2}{12} + \frac{(16-12)^2}{12} + \frac{(8-12)^2}{12} + \frac{(12-12)^2}{12} + \frac{(11-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(14-12)^2}{12} = 4.17$$

We note that n=7 frequency pairs are used in the computation of χ^2. Further, N = $\sum f_i$ = 84. Is the only quantity used in the computation of e_i. Therefore, the number of degrees of freedom is v= 7-1 = 6. From the Table we find that $\chi^2_{0.05}(6) = 12.59$ and $\chi^2_{0.01}(6) = 16.81$.

Since χ^2=4.17 is much less than both of $\chi^2_{0.05}(6)$ and $\chi^2_{0.01}(6)$, we do not reject the hypothesis. This means that the accidents seem to be distributed uniformly over the week.

2. A set of five similar coins is tossed 320 times and the result is

No. of heads	0	1	2	3	4	5
Frequency	6	27	72	112	71	32

Test the hypothesis that the data follow a binomial distribution function.

Solution: We shall set up the null hypothesis that data actually follows a binomial distribution. Then alternate hypothesis is, namely, data does not follow binomial distribution. Next, to set up a suitable level of significance, $\alpha = 5\%$, with **n** = 6, degrees of freedom is $\gamma = 5$. Therefore, the tabulated value is $\chi^2|_{\alpha=0.05,\gamma=5} = 11.07$. Before proceeding to finding test criterion, first we compute the various expected frequencies. As the data is set to be following binomial distribution, clearly probability density function is $F(X) = N\binom{n}{k}p^k q^{n-k}$.

Here, $n = 320$, $p = 0.5$, $q = 0.5$, and **k** takes the values right from 0 up to 5. Hence, the expected frequencies of getting 0, 1, 2, 3, 4, 5 heads are the successive terms of the binomial expansion

Here, observed values are: O_i : 6, 27, 72, 112, 71, 32

The expected values are: E_i : 10, 50, 100, 100, 50, 10.

$$\chi^2|_{cal} = \left(\frac{(6-10)^2}{10}\right) + \left(\frac{(27-50)^2}{50}\right) + \left(\frac{(72-100)^2}{100}\right)$$

$$+ \left(\frac{(112-100)^2}{100}\right) + \left(\frac{(71-50)^2}{50}\right) + \left(\frac{(32-10)^2}{10}\right) = 78.68.$$

As the calculated value is very much higher than the tabulated value of 3.841, we reject the null hypothesis and accept the alternate hypothesis that data does not follow the binomial distribution.

3. A set of five identical coins is tossed 320 times and the result is shown in the following table.

No. of heads	*0*	*1*	*2*	*3*	*4*	*5*
Frequency	*6*	*27*	*72*	*112*	*71*	*32*

Test the hypothesis that the data follows a binomial distribution associated with a fair coin.

Solution: The Probability that x number of fair coins out of 5 shows a head in a single toss is given by the binomial function

b (5, ½, x) = $5_{C_x}(1/2)^x(1/2)^{5-x} = \frac{1}{2^5}(5_{C_x}) = \frac{1}{32}(5_{C_x})$ = b(x), say,

accordingly, in 320 tosses the expected number of tosses in which x number of coins show a head is 320 × b(x). Hence the expected frequencies (i,e. the number of tosses in which 0,1,2,3,4,5 coins show a head) are, respectively,

$e_1 = 320 \times b(0) = 320 \times \frac{1}{32} \times 5_{C_0} = 10,$

$e_2 = 320 \times b(1) = 320 \times \frac{1}{32} \times 5_{C_1} = 50,$

$e_3 = 320 \times b(2) = 320 \times \frac{1}{32} \times 5_{C_2} = 100,$

$e_4 = 320 \times b(4) = 320 \times \frac{1}{32} \times 5_{C_4} = 100,$

$e_5 = 320 \times b(5) = 320 \times \frac{1}{32} \times 5_{C_5} = 50,$

$e_6 = 320 \times b(6) = 320 \times \frac{1}{32} \times 5_{C_6} = 10,$

The corresponding observed frequencies are

$f_1 = 6,\ f_2 = 27,\ f_3 = 72,\ f_4 = 112,\ f_5 = 71,\ f_6 = 32$

We find that

$$\chi^2 = \frac{(6-10)^2}{10} + \frac{(27-50)^2}{50} + \frac{(72-100)^2}{100} + \frac{(112-100)^2}{100} + \frac{(71-50)^2}{50} + \frac{(32-10)^2}{10}$$

$$= \frac{16}{10} + \frac{529}{50} + \frac{784}{100} + \frac{144}{100} + \frac{441}{50} + \frac{484}{10} = 78.68$$

We note that the number of degrees of freedom is 6-1 = 5. From the table we find that $\chi^2_{0.05}(5) = 11.07$ and $\chi^2_{0.01}(5) = 15.09$. We observe that $\chi^2 = 78.68$, is very much greater than both of $\chi^2_{0.05}(5)$ and $\chi^2_{0.01}(5)$. Therefore, we reject the hypothesis that the observed data follows a binomial distribution associated with a fair coin.

4. Five dice were thrown 96 times and the numbers 1, 2 or 3 appearing on the dice follows the frequency distribution as below.

No. of dice showing 1, 2 or 3	5	4	3	2	1	0
Frequency	7	19	35	24	8	3

Test the hypothesis that the data follows a binomial distribution. ($\chi^2_{0.05} = 11.07$ for 5 d.f).

Solution:

$p = q = 0.5$

$F(x) = N\left({}^nC_x\right) p^x q^{n-x}$

By fitting of Binomial distribution, we get

0i	7	19	35	24	8	3
Ei	3	15	30	30	15	3

$$\chi^2 = \sum \frac{(E_i - O_i)^2}{E_i} = 11.7 > 11.07$$

Therefore, hypothesis rejected at 5% level of significance.

5. Fit a Poisson distribution to the following data and test for its goodness of fit at a level of significance 0.05. ($\chi^2_{0.05}$ with 3 d.f = 9.48)

X	0	1	2	3	4
f	419	352	154	56	19

Solution:

$\bar{x} = \frac{\sum fx}{N} = \frac{904}{1000} = 0.904 = m$, the mean of Poisson distribution.

Hence $P(x) = \frac{m^x e^{-m}}{x!} = \frac{(0.904)^x e^{0.904}}{x!}$, $x = 0,\ 1,\ 2,\ 3,\ 4$

Hence the expected frequency for 'x' successes is

$E_x = N \times P(x) = \frac{1000 \times (0.904)^x e^{-0.904}}{x!}$, where x = 0, 1, 2, 3, 4.

Putting x = 0, 1, 2, 3 , 4 we get

$$E_0 = N \times P(0) = \frac{1000 \times (0.904)^0 e^{-0.904}}{0!} = 405,$$

$$E_1 = N \times P(1) = \frac{1000 \times (0.904)^1 e^{-0.904}}{1!} = 366,$$

$$E_2 = N \times P(2) = \frac{1000 \times (0.904)^2 e^{-0.904}}{2!} = 165.4,$$

$$E_3 = N \times P(3) = \frac{1000 \times (0.904)^3 e^{-0.904}}{3!} = 49.8,$$

$$E_4 = N \times P(4) = \frac{1000 \times (0.904)^4 e^{-0.904}}{4!} = 11.2,$$

Hence the theoretical frequencies are

x:	0	1	2	3	4
f:	405	366	165.4	49.8	11.2

$\chi^2 = \sum \frac{(E_i - O_i)^2}{E_i} = \frac{(419-405)^2}{405} + \frac{(352-366)^2}{366} + \frac{(154-164.5)^2}{164.5} + \frac{(56-49.8)^2}{49.8} + \frac{(19-11.2)^2}{11.2}$

= 7.87

Here calculated $\chi^2 < 9.48$. So we accept H_0.

Exercises:

1. A random sample of size 2 is drawn from the population 3, 4, 5. Find the sampling distribution of the sample mean. (a) with replacement (b) without replacement. Find the sample mean and sample variance in these two case.
2. 500 ball bearings have a mean weight of 142.30 gms. and S.D of 8.5 gms. Find the probability that a random sample of 100 balls bearings chosen from this group will have a combined weight (a) between 14,061 and 14,175 gms. (b) more than 14,460 gms.
3. The weights of packages received by a department store have a mean of 136 kgs. and a S.D of 22.5 kgs. What is the probability that 25 packages received at random and loaded on an elevator will exceed the safety limit of the elevator quoted as 3720 kgs.
4. A 'die' was thrown 9000 times and a throw of 5 or 6 was obtained 3240 times. On the assumption of random throwing, do the data indicate an unbiased die?
5. A sample of 900 days is taken from meteorological records of a certain district and 100 of them are found to be foggy. Find the 99.73% confidence level probable limits of the percentage of foggy days in the district.
6. The mean and S.D marks of a sample of 100 students are 67.45 and 2.92 respectively. Find (a) 95% (b) 99% confidence intervals for estimating the mean marks of the population.
7. The mean of sample of size 1000 and 2000 are 67.5 cms and 68 cms respectively. Can the samples be regarded as drawn from the same population of S.D 2.5 cms?
8. A machine produced 20 defective units in a sample of 400. After over oiling the machine it produced 10 defective in a batch of 300. Has the machine improved due to over oiling?
9. Ten individuals are chosen at random from a population and their heights in inches are found to be 63, 63, 64, 65, 66, 69,69, 70, 70, 71. Discuss the suggestion that the mean height of the population is 65 inches given that $t_{.05} = 2.262$ for 9 d.f.

   ot the percentage ot toggy days in the district.
10. From a random sample of 10 pigs fed on diet A, the increase in weight in a certain period were 10, 6, 16, 17, 13, 12, 8, 14, 15, 9 lbs. For another random sample of 12 pigs fed on diet B, the increase in weight in the same period were 7, 13, 22, 15,1 2,1 4,18, 8 , 21, 23, 10, 7 lbs. Test whether diets A and B differ significantly regarding their effect on increase in weight. ($t_{.05}$ fir 20 d.f is equal to 2.09)
11. 4 coins were tossed 160 times and the following results were obtained.

No. of heads	0	1	2	3	4
Frequency	17	52	54	31	6

Test the goodness of fit of the binomial distribution ($\chi^2_{0.05} = 9.49$ for 4 d.f)

12. Fit a Poisson distribution for the following data and test the goodness of fit given that $\chi^2_{0.05} = 9.49$ for 4 d.f

x	0	1	2	3	4
f	419	352	154	56	19

F – test or Fisher's F-test

The F-test was first originated by the statistician R.A. Fisher. This test is also known as Fisher's F-test or simply F-test. It is based on the F-distribution, which is defined as the ratio of two independent chi-square variates which is derived by dividing each variable by its corresponding degree of freedom $\boldsymbol{F = \frac{\psi^2/\nu_1}{\psi^2/\nu_2}}$

To test if the two samples have come from same population we use F test (OR) To test there is any significant difference between two estimates of population variance.

F= greater variance/smaller variance

$$F = \frac{S_1^2}{S_2^2}$$

Where

$S_1^2 = \frac{\sum(x-\bar{x})^2}{n_1-1}$

$S_2^2 = \frac{\sum(y-\bar{y})^2}{n_2-1}$

Where n_1 is the first sample size and n_2 is the second sample size.

If the sample variance S^2 is not given we can obtain the population variance byusing the relation $S_1^2 = \frac{n_1 s_1^2}{n_1-1}$ and $S_2^2 = \frac{n_2 s_2^2}{n_2-1}$

Assumptions in F-test.

The F-Test is based on the following assumptions:

1. Normality: The values in each group should be normally distributed.
2. Independence of Error: The variation of each value around its own group mean.
3. Homogeneity: The variances within each group should be equal for all groups.

If, however, the sample sizes are large enough, we do not need the assumption of normality.

Test of hypothesis about the variance of two populations

We have the following steps:

1. Null Hypothesis: $H_0: S_1^2 = S_2^2$

   Alternate Hypothesis: $H_1: S_1^2 \neq S_2^2$
2. Calculation of Test Statistic.

$F = \frac{S_1^2}{S_2^2}$ if $S_1^2 > S_2^2$ so that $F \geq 1$ or $F = \frac{S_2^2}{S_1^2}$ if $S_2^2 > S_1^2$ so that $F \geq 1$

3. Level of significance: Take the level of significance $\alpha = 0.05$ if α is not known.
4. Decision: Accept H_0 if computed $F \leq$ tabled F_α

   Reject H_0 if computed $F >$ tabled F_α.

Problems

1. In one sample of 8 observations the sum of the squares of deviations of the sample values from the sample mean was 84.4 and in the other sample of 10 observation it was 102. 6. Test whether this difference is significant at 5 % level.

Solution: Assume Null Hypothesis: $H_0: S_1^2 = S_2^2$ (There is no significant difference)

Alternate Hypothesis: $H_1: S_1^2 \neq S_2^2$

Given $\sum(x - \bar{x})^2 = 84.4, n_1 = 8, \sum(y - \bar{y})^2 = 102.6, n_2 = 10$

$$S_1^2 = \frac{\sum(x - \bar{x})^2}{n_1 - 1} = \frac{84.4}{8 - 1} = 12.057$$

$$S_2^2 = \frac{\sum(y - \bar{y})^2}{n_2 - 1} = \frac{102.6}{10 - 1} = 11.4$$

$$F = \frac{S_1^2}{S_2^2} = 1.057$$

Calculated F value = 1.057

Tabulated Value = 3.29 (at 5% level of significance with (7,9) degrees of freedom)

Calculated value < Tabulated value,

Hence accept Ho (Null hypothesis)

2. Two random samples gave the following results.

Sample	Size	Sample mean	Sum of squares of deviations from the mean
1	10	15	90
2	12	14	108

Test whether the samples come from the same normal population.

Solution: Assume Null Hypothesis: $H_0: S_1^2 = S_2^2$ (the samples come from the same normal population)

Alternate Hypothesis: $H_1: S_1^2 \neq S_2^2$

Given $\sum(x - \bar{x})^2 = 90, n_1 = 10, \sum(y - \bar{y})^2 = 108, n_2 = 12$

$$S_1^2 = \frac{\sum(x-\bar{x})^2}{n_1 - 1} = \frac{90}{9} = 10$$

$$S_2^2 = \frac{\sum(y-\bar{y})^2}{n_2 - 1} = \frac{108}{11} = 9.82$$

$$F = \frac{S_1^2}{S_2^2} = 1.018$$

Calculated F value = 1.018

Tabulated Value at 5% level of significance with (9,11) degrees of freedom= 2.90

Calculated value < Tabulated value,

Hence accept Ho (Null hypothesis)

3. The time taken by workers in performing a job by method I and method II isgiven below.

Method I	20	16	26	27	23	22	
Method II	27	33	42	35	32	34	38

Do the data show that the variances of time distribution from population fromwhich these samples are drawn do not differ significantly?

Solution: Assume Null Hypothesis: $\boldsymbol{H_0: S_1^2 = S_2^2}$ (The two samples have the same variance)

Alternate Hypothesis: $H_1: S_1^2 \neq S_2^2$

x	$x-\bar{x}$	$(x-\bar{x})^2$	y	$y-\bar{y}$	$(y-\bar{y})^2$
20	-2	4	27	-8	64
16	-6	36	33	-2	4
26	4	16	42	7	49
27	5	25	35	0	0
23	1	1	32	-3	9
22	0		34	-1	1
			38	3	9
$\sum$X =134		$\sum(x-\bar{x})^2 = 82$	$\sum y = 241$		$\sum(y-\bar{y})^2 = 136$

Given $\bar{x} = \frac{134}{6} = 22, \bar{y} = \frac{241}{7} = 34.428 = 35$

$$S_1^2 = \frac{\sum(x-\bar{x})^2}{n_1 - 1} = \frac{82}{5} = 16.4$$

$$S_2^2 = \frac{\sum(y-\bar{y})^2}{n_2 - 1} = \frac{136}{6} = 22.66$$

$$F = \frac{S_1^2}{S_2^2} = 1.38$$

Calculated F value = 1.37

Tabulated Value = 4.95 (at 5% level of significance with (6,5) degrees of freedom)

Calculated value < Tabulated value, Accept Ho (Null hypothesis)

4. In a test given to two groups of students drawn from two normal populations, the marks obtained were as follows:

Group A	18	20	36	50	49	36	34	49	41
Group B	29	28	26	35	30	44	46		

Examine at 5% level, Whether the two populations have the same variance.

Solution: Assume Null Hypothesis: $\boldsymbol{H_0: S_1^2 = S_2^2}$ (The two samples have the same variance)

Alternate Hypothesis: $H_1: S_1^2 \neq S_2^2$

x	$x-\bar{x}$	$(x-\bar{x})^2$	y	$y-\bar{y}$	$(y-\bar{y})^2$
18	-19	361	29	5	25
20	-7	289	28	6	36
36	-1	1	26	8	64
50	13	169	35	1	1
49	12	144	30	4	16
36	-1	1	44	10	100
34	-3	9	46	12	144
49	12	144			
41	4	16			
$\sum x$ =333		$\sum(x-\bar{x})^2$ = 1134	$\sum y$ = 238		$\sum(y-\bar{y})^2$ = 386

Given $\bar{x} = \frac{333}{9} = 37, \bar{y} = \frac{238}{7} = 34$

$$S_1^2 = \frac{\sum(x-\bar{x})^2}{n_1 - 1} = \frac{1134}{8} = 141.75$$

$$S_2^2 = \frac{\sum(y-\bar{y})^2}{n_2 - 1} = \frac{386}{6} = 64.33$$

$$F = \frac{S_1^2}{S_2^2} = 2.203$$

Calculated F value = 2.203

The table value of F at 5% level for 8 and 6 degrees of freedom is 4.15

Calculated value < Tabulated value,

Hence accept the Null hypothesis.

Exercises

1. The nicotine content in milligrams of two samples of tobacco were found to beas follows:

Sample A	24	27	26	21	25	
Sample B	27	30	28	31	22	36

Can it be said that two samples come from normal populations having the samevariances.

2. The standard deviations calculated from two random samples of size 9 and 13 are 2 and 1.9 respectively May the sample be regarded as drawn from the normal population with the same standard deviation.

Video links:

1. Hypothesis Testing - Statistics - YouTube
2. Student's t-test - YouTube
3. Chi-square distribution introduction | Probability and Statistics | Khan Academy - YouTube
4. F-test - YouTube