UNIT – V

Small Sample Test

The Small Sample test can be calculated by three types of tests.

1, t-test
2, f-test
3, χ^2 (chi) test

Student-t-test:-

It is also called as t-test.

Definitions:-

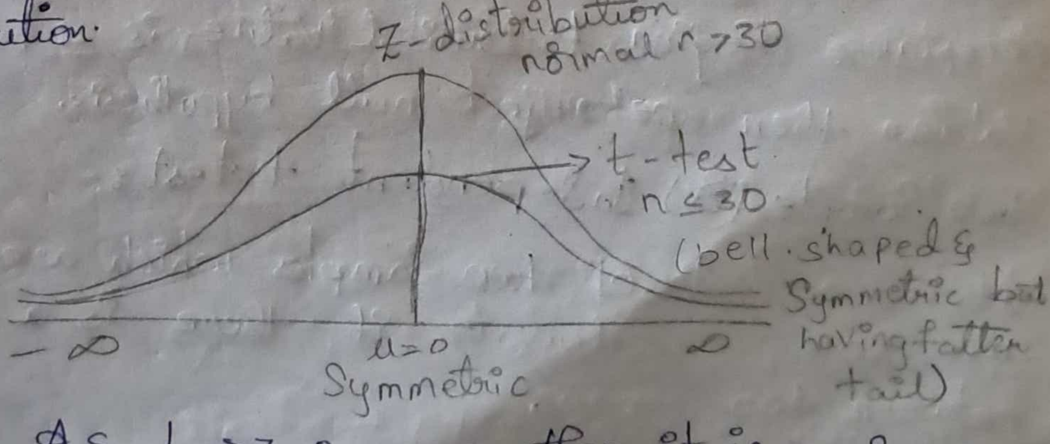
Student t distribution is a probability t-distribution that is used to calculate the Small Sample test's of Population parameters when the population variance σ is unknown.

There are two conditions for t-distribution.

1, Sample Size $n \leq 30$
2, population variance (σ) is unknown.

Student t distribution is given by W.S. Gosset but he published his studies under the name of Student that's why it is called Student t-test.

Student t-distribution is continuously probability t-distribution that generalizes the standard normal t-distribution.



Note:- As $t \to z$ increases then it is a normal when to use t-distribution.

1, Sample Size is ≤ 30.

2, population variance σ is unknown

3, population distribution is unimodel and skewed

Different types of t-tests:-

1, One Sample t-test

2, Independent Sample t-test

3, Paired t-test

One Sample t-test:-

In this case we compare the average of one group against the population mean. If population mean is greater than other then we have to perform One-tail t-test.

The formula for One Sample t-test is

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

Where $\bar{x}$ is Sample mean

μ = Population mean

n = Sample size

S = Standard deviation of Sample.

Here we take degrees of freedom as $n-1$

Acceptance Region t-critical Value is $\leq$ t-calculated Value then we reject null hypothesis.

Two Sample (or) Independent t-test:-

It is a test of two Samples which are independent. here we calculate if there is a difference between two groups.

for Example: Avg height of males is Comparing with Avg height of females.

formula for this $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2^2}}}$

where $\bar{x}_1$ and $\bar{x}_2$ are Sample means of two groups.
S_1 and S_2 are standard deviation of two Samples.
n_1 and n_2 are Sample sizes.

Acceptance Region - t-critical value $\leq$ t-calculated value then we reject null hypothesis.

Paired t-test:-

The Paired t-test Sometimes called as Dependent Sample t-test.
It is used to determine whether the mean difference between two Sets of groups at different time interval. that is each group is measured twice resulting in Pair of observations.

The formula for paired t-test is $t = \frac{\Sigma(x_1 - x_2)/n}{Sd/\sqrt{n}}$

Here x_1 and x_2 are Sample mean
$n =$ Sample size
Sd is standard error.

Degrees of freedom $n-1$

Acceptance Region - t-critical value $<$ t-calculated value then we reject null hypothesis.

Properties of t-test:-

1, It ranges from $-\infty$ to $+\infty$
2, It has bell shapped curve and symmetric.
3, Student t-distribution is different for different Sample sizes.
4, The total area under a t-curve $= 1$. but nevers touches the Peak.

4, Mean is zero.
5 Population standard deviation is unknown,
6, The data is Continuous and it has been rando
Sampled from a population.
7, An important property of test statistic is its
Sampling distribution under the null hypothesis
must be calculated either exactly or approximately.

F- Distribution:-

F-test is an statistical test which is
used to Compare the variances of two
Samples or the ratio of Variance between
multiple Samples.

A two tailed F-test is used to check whether
the Variances of two Samples are equal or
not.

F-test formula:-

the f-test formula for different hypothesis is
given as follows.

Left-tailed test:-

Null hypothesis $H_0: \sigma_1^2 = \sigma_2^2$
Alternative hypothesis $H_1: \sigma_1^2 < \sigma_2^2$.
Decision Criteria : If $f_{stat} < f_{crit}$
then we reject H_0.

Right tailed test:-

Null hypothesis $H_0: \sigma_1^2 = \sigma_2^2$
Alternative hypothesis $H_1: \sigma_1^2 > \sigma_2^2$
Decision Criteria: If $f_{stat} > f_{crit}$
then we reject H_0

Two - tailed test:-

Null Hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$

Alternative hypothesis: $H_1: \sigma_1^2 \neq \sigma_2^2$

Decision Criteria: If $f_{stat} > f_{crit}$
then we reject H_0

F- Statistic formula:-

F- Statistic for small samples $F = \frac{S_1^2}{S_2^2}$

where S_1^2 is Variance of first Sample and
S_2^2 is Variance of Second Sample

Note:-

for two-tailed f test the variance with the greater value will be in the numerator.

F-test for Critical Value:-

Find the degrees of first Sample that is done by Subtracting 1 from given Sample size n_1 i.e, $x = n_1 - 1$ degrees of freedom. Similarlly for the Second Sample the degrees of freedom $y = n_2 - 1$.

Uses of f-test:-

1, whether two independent Samples have been drawn from normal population with same variance.

2, Whether two independent estimates of population Variances are homogenous or not.

Assumptions:-

1 the distribution in each group should be normally distributed.

2, Error should be independent of each observed value.

3, Variance within each group should be equal for all groups.

$$S_1^2 = \frac{1}{n-1} \Sigma (x-\bar{x}_1)^2$$

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

Properties of f- distribution:-

1, The f-distribution is a continuous probability distribution that has no negative range values.

2, It is a ratio of two independent χ^2 (chi-sque) distribution each divided by their degrees of freedom.

3, the F distribution has 2 parameters i.e, the numerator degrees of freedom df_1 and denominator degrees of freedom df_2.

4, The shape of distribution depends on degrees of freedom as the degrees of freedom increases the distribution becomes more symmetrical and approaches normal distribution

Relationship with other probability distributions.

the f distribution is related to other probability distributions such as chi square & t-distributions. the chi-square distribution is used to test for differences in variances of a single population. while the t-distribution is used for testing differences in means for a single population.

*The f-distribution can also be related to β-distribution. As β distribution is used To model the proportions or probabilities of events. while the f-distribution is used to model the ratio of two variances.

The chi-square test:-

The chi-square test is one of the most commonly used for non-parametric test. It was introduced by Carl pearson as a test of association and the Greek Letter χ^2 is used to denote this test.

Defination:-

The chi-square test is a hypothesis test that is used when you want to determine If there is a relationship between Categorial Variables.

Chi-Square distribution:-

The distribution of chi-square statistic is called chi-square distribution. chi-square distributions are a family of distributions that take only Positive values.

Condenjency table:-

	Column 1	Column 2	Total
Row 1	A	B	$R_1 = A+B$
Row 2	C	D	$R_2 = C+D$
Total	$C_1 = A+C$	$C_2 = B+D$	N

A Condenjency table is a type of table in a matrix format that displays the frequency distribution of Variables. They provide a basic picture of interrelationship between two Variables.

The chi-square statistic Compares the observed Count in each table cell to the count which would be expected under the assumption of no association between row and column Classifications

Degrees of freedom:-

In general the degrees of freedom of an estimate of a parameter is Equals to the no. of independent scores that go into the estimate(-) minus the no. of parameters used as intermediate steps in the estimation i.e, the Sample Variance has n-1 degrees of freedom.

The no. of degrees of freedom for n observations is n-k and usually denoted by V.

the degrees of freedom for chi-Square Contenjency table can be calculated as $V = (r-1, c-1)$.

where r = no. of rows

c = no. of Columns.

Chi-Square formula:-

the Chi-Square test is used to determine whether there is a Significant difference between expected frequency and observed frequencies in are more categories.

The value of chi-square is calculated as.

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots$$

where $O_1, O_2 \ldots O_i$ are observed values

$E_1, E_2 \ldots E_i$ are Expected Values

Steps to solve chi-square test.

Step-1:- Calculate the Expected frequencies

$$E = \frac{\text{Row Total} \times \text{Column total}}{\text{Grand total}}$$

step-2:- Take the difference between the observed and Expected frequencies and obtain the squares of these differences. i.e $(O-E)^2$

step-3:- Divide the values obtained in step 2 by the respective Expected frequency according to the formula. i.e $\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i}$

The chi-Square Critical value is less than or Equal to chi-Square tabulated value then we reject H_0.

$\chi^2_{cri}(\alpha, k-1) \leq \chi^2_{tab}$.

α = level of significance.

k = Degrees of freedom.

Limitations of chi Square test:-

The chi-Square test doesnot gives us much information about the strength of the relationship. The chi-square test is sensitive to sample size. The chi-Square should be used together with measures of association like Cramer's method (or) γ (Gama) method to guide in deciding whether a relationship is important and worth perceving.

~~The~~ It can be used only when not more than 20% of cells have an expected frequency of less than 5.

Types of chi-Square test.

There are two commonly used chi-Square test

1, the chi-Square goodness of fit and

2, chi-square test of independence.

Both tests involve ~~the~~ variables that divide your data into categories.

<u>chi-square test properties</u>:-

The chi-square distribution is a continuous probability distribution.

The values ranging from '0' to infinity in the +ve direction (never assumes -ve value).

The sum of independent chi-square is itself a chi-square variant

Chi-square distribution depends on degrees of freedom as its shape changes when the change in v.

As v becomes greater then chi-square gets approximation of Normal distribution.

<u>Sign test</u>:-

The sign test is a Rank test in which the test statistic is calculated by forming differences in paired sample of dependent groups.

One sample test (median)

It is the simplest form of entire non parametric test as the name suggests it is based on sign (- & +) of deviation rather than exact magnitude of variable value

It is used to test the hypothesis concerning the median for one population To test the hypothesis that median (η) of population has a specified value η_0

The null hypothesis $H_0 : \eta = \eta_0$

the Alternative hypothesis $H_1 : \eta \neq \eta_0$

Procedure:-

Let $x_1, x_2 ---- x_n$ be a random Sample of size n. from a given population with median $\eta = \eta_0$. Subtract the median Value η_0 from Each and Every Variable of x_i. and then Write

1, '+' sign Variables If the deviation is '+ve'.

2, '-' sign of Variable If the deviation is '-ve'.

3, '0' If the deviation is 'zero' (0).

From the defination of median we have

$P(x > median) = P(x < median)$

$= \frac{1}{2}$ (or) 0.5

$\therefore P(x > \eta_0) = P(x < \eta_0) = \frac{1}{2}$ (or) 0.5

Hence If H_0 is true then the '+ve' signs should be approximately equals to 'negative' signs.

Sign test for Small Samples:-

Here in sign test has ≤ 25 Samples is Considered as Small Samples. and the procedure is as follows.

1, Set the hypothesis

Null hypothesis $H_0 : \eta = \eta_0$

Alternative hypothesis $H_1 : \eta \neq \eta_0$

2, Consider a level of significance i.e α.

3, Compute $T^+ - T^-$

where T^+ = Total no. of positive signs

T^- = Total no. of '−' signs.

Critical regions:-

If define critical region as $T_t \le T_c$

where T_c is critical region

T_t is test Statistic at given level of Significance.

* If $T_t \le T_c$ we reject H_0. otherwise accept H_0.

Paired Samples:-

this Sample is used to test the difference between two population medians when the populations are not normally distributed. For Paired Sample test we must have two conditions

1, A Sample must be randomly selected from each population.

2, the Samples must be dependent.

The difference between Corresponding data entries is found and sign of difference is recorded.

Procedure:-

1, Identify Null and Alternative hypothesis.

2, Specify the level of significance 'α'.

3, Determine the Sample size n by finding the difference for Each data Pair. i.e, Assign +ve Sign for +ve & −ve sign for −ve & '0' for zero difference

$\therefore n$ = Total no. of +ve & −ve signs.

4, Determine Critical Value.

5, Find test statistic i.e X = lesser no. of +ve & −ve sign.

6, Make a decision If test statistic is lessthan or Equal to test critical then we reject H_0.

Limitation.

It often has lower efficiency and lower power than test that require stronger assumptions, when those assumptions are valid.

1. ***Mann-Whitney U Test***:
   - The Mann-Whitney U test, also known as the Wilcoxon rank-sum test, is a non-parametric test used to compare two independent groups to determine whether their distributions differ significantly from each other.
   - It's suitable for ordinal or continuous data when the assumptions of parametric tests like the t-test are not met.
   - Here's how it works:
     1. Combine the data from both groups and rank all the observations from smallest to largest.
     2. Assign ranks to tied values by averaging the ranks they would occupy.
     3. Calculate the sum of ranks for each group.
     4. Compute the U statistic, which is the smaller of the two sums of ranks. If the sample sizes are equal, U can be calculated directly. Otherwise, a correction is applied to account for the different sample sizes.
     5. Compare the calculated U value to a critical value from the Mann-Whitney U distribution table or use statistical software to determine statistical significance.

2. *Run Test*:
- The Run test is a non-parametric test used to analyze the randomness of a sequence of observations. It's particularly useful for detecting patterns or trends in time series data.
- The test involves counting the number of runs in the data sequence, where a run is defined as a sequence of consecutive observations with the same characteristic (e.g., all increasing or all decreasing values).
- Here's how it works:
  1. Arrange the data sequence in chronological order.
  2. Count the number of runs (R) in the sequence.
  3. Calculate the expected number of runs (ER) under the assumption of randomness. For a sequence of n observations, ER is given by: $ER = \frac{2n_1n_2}{n} + 1$, where n_1 and n_2 are the number of positive and negative deviations from the median, respectively.
  4. Compare the observed number of runs (R) to the expected number of runs (ER) using a suitable test statistic, such as the z-score or chi-squared statistic.
- The Run test helps determine whether a sequence of data exhibits randomness or if there's a systematic pattern present.

Both the Mann-Whitney U test and the Run test are valuable tools in statistics for analyzing data in situations where parametric assumptions are not met or when assessing patterns in data sequences.

Kolmogorov Smirnov One Sample Test

Introduction

The Kolmogorov Smirnov (K-S) test may be used to evaluate whether two sets of data are significantly different from one another.

The test compares empirical distribution (observed sample data) with a hypothetical distribution (expected distribution).

Like the chi-square goodness of fit test, the purpose of the K-S test is to examine the extent of agreement between the two distributions (observed and unknown).

Advantages

The K-S test for goodness-of fit compares the cumulative theoretical frequency distribution with the cumulative known (sample) frequency distribution.

The K-S test is an exact test even for small sample sizes, as it is not limited by minimum expected values as in the chi-square test.

K-S test is useful on ordinal data, whereas chi-square is appropriate for nominal data.

Assumptions

The sample was drawn from a specified theoretical distribution, and

Every observed value is close to the hypothesized value from the theoretical distribution ordinal data

Advantages

The K-S test for goodness-of fit compares the cumulative theoretical frequency distribution with the cumulative known (sample) frequency distribution.

The K-S test is an exact test even for small sample sizes, as it is not limited by minimum expected values as in the chi-square test.

K-S test is useful on ordinal data, whereas chi-square is appropriate for nominal data.

Assumptions

The sample was drawn from a specified theoretical distribution, and

Every observed value is close to the hypothesized value from the theoretical distribution

Hypothesis

H0 = The data are normally distributed

Ha= The data are not normally distributed

The Kolmogorov-Smirnov Test (K-S Test) :- [one - Sample]

- To test, $H_0 : F(x) = F_0(x)$
  $H_1 : F(x) \neq F_0(x)$

N-P Test-
Dist^n Free
Tests

- Test statistic

$$D_n = \sup_x |F_n(x) - F_0(x)|$$

or

$$= \text{Max} |S_n(x) - F_0(x)|$$

here $S_n(x)$ = empirical df (observed).
$F_0(x)$ = theoritical df (expected)

- Test statistic

$$D_n = \sup_x |F_n(x) - F_0(x)|$$

or

$$= \text{Max} |S_n(x) - F_0(x)|$$

here $S_n(x)$ = empirical df (observed)

$F_0(x)$ = theoritical df (expected)

- conclusion

Reject H_0 if

$$D_n \geq D_{n,\alpha}$$

where $D_{n,\alpha}$ is critical value obtained from table

Steps of Kruskal-Walis Test

- All observations from k samples (k groups) are combined into a single series and arranged in order magnitude from smallest to largest.
- The observations are then replaced by ranks. The smallest observation is replaced by rank 1, the next to smallest by rank 2 and the largest by rank N.
- The sum of the ranks in each sample (column) is taken.
- The Kruskal-Walis Test determines whether these sums of ranks are so disparate that they are not likely to come from same population or not.
- H value is compared to a table of critical values for U based on the sample size of each group. If H exceeds the critical value for H at some significance level (usually 0.05) it means that there is evidence to reject the null hypothesis in favor of the alternative hypothesis.

Definition:

The Kruskal–Wallis one-way analysis of variance by ranks is a non-parametric method for testing whether samples originate from the same distribution. It is also called Kruskal-Wallis H test.

Kruskal-Wallis was presented by :
William Kruskal and W. Allen Wallis

Kruskal-Wallis test
(three or more separate groups)

•The Kruskal-Wallis test is used to compare the medians of more than two groups, just like the one-way analysis of variance

The Kruskal-Wallis H Test

H_0: the k distributions are identical versus

H_a: at least one distribution is different

Test statistic: ***Kruskal-Wallis H***

When H_0 is true, the test statistic H has an approximate chi-square distribution with $df = k-1$.

[illegible]e a right-tailed rejection region or p-value based on the Chi-square distribution.