Unit-V

t-test for single mean:

(i) If a random sample x_i $(i=1,2\ldots n)$ of size n has been drawn from a random population with specified mean say μ_0, or.

(ii) If the sample mean differs significantly from the hypothetical value μ_0 of the population mean.

Under the null hypothesis H_0:

(i) the sample has been drawn from the population with mean μ_0 or

(ii) there is no significant difference b/w the sample mean $\bar{x}$ and the population mean μ_0.

the statistic, ②

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

follows student's dist with $(n-1)$ d.f.

we now compare the calculated value of t, with the tabulated value at certain level of significance.

If calculated $|t| >$ tabulated t, null hypothesis is rejected and if calcuted $|t| <$ tabulated t, Ho may be accepted at the level of significance ~~adop~~ accept.

Remark : 1

on computation of s^2 for numerical problems:- If $\bar{x}$ comes out in integers, the formula can be convenentely used for computing s^2. However if $\bar{x}$ comes in fractions case, step deviation method given below, is quite useful.

If we take $d_i = x_i - A$ where A is any arbitrary number, then.

$$s^2 = \frac{1}{n-1}\left[\Sigma [x_i - \bar{x}]^2\right] = \frac{1}{n-1}\left[\Sigma x_i^2 - \left(\frac{\Sigma x_i}{n}\right)^2\right]$$

$$= \frac{1}{n-1}\left[\Sigma d_i^2 - \left(\frac{\Sigma d_i}{n}\right)^2\right]$$ since variance is independent change of orgin.

Also in case $\bar{x} = A + \frac{\Sigma f d_i}{N}$

2.) we know, the sample variance,

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

$$\Rightarrow n s^2 = (n-1) S^2.$$

$$\frac{S^2}{n} = \frac{s^2}{(n-1)}.$$

Hence for numerical problems the test statistic on using becomes.

$$t = \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}} = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n-1}} \sim t_{n-1}.$$

paired t-test for difference of mean:

let us know consider the case when (i.) the sample sizes are equal.

4

i.e., $n_1 = n_2 = n$ (say) and (ii) the two sample are not independent but the sample observations are paired together.

i.e the paired of observations (x_i, y_i) $(i = 1, 2 \ldots n)$ corresponds to the same (ith) sample unit. The problem is to test if the sample means differ significantly or not.

For example, suppose we want to test the efficacy of a particular drug, say the including sleep. Let x and y_i $(i = 1, 2 \ldots n)$ be the readings, in hours of sleep, on the individual before and after the drug is given respectively, Here instead of applying the difference of the means test discussed in, we apply the paired t-test given below.

Here we consider the increments $d_i = x_i - y_i$ $(i = 1, 2 \ldots n)$.

under the null hypothesis, H_0 that increments are due to fluctuations of sampling.

i.e the clag is not responsible for those increments, the statistic.

$$t = \frac{\bar{d}}{S/\sqrt{n}}.$$

where $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2.$$

follows students t-distribution with $(n-1)$ df.

T-test for diff means:

suppose we want to test if the independent samples $x_i (i=1,2..n)$ and $y_i (i=1,2...n)$ of sizes n_1 and n_2 here been drawn from two normal population with means μ_x and μ_y respectively.

Under the null hypothesis (H_0) that the samples have been drawn from the normal populations with means μ_x and μ_y and under the assumption that the population variance an equal i.e., ($\sigma_x^2 = \sigma_y^2 = \sigma^2$ (say)). the statistic

$$t = \frac{(\bar{x}-\bar{y}) - (\mu_x - \mu_y)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}.$$

where $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n} x_i$ $\bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i$ ⑥

and $S^2 = \frac{1}{n_1+n_2-2}\left[\sum_i (x_i-\bar{x})^2 + \sum_i (y_i-\bar{y})^2\right].$

is an unbiased estimate of the mean common population variance σ^2, following students distribution (n_1+n_2-2) df.

proof:-

distribution of t defined in

$$\xi_i = \frac{(\bar{x}-\bar{y}) - E(\bar{x}-\bar{y})}{\sqrt{V(\bar{x}-\bar{y})}} \sim N(0,1).$$

put $E(\bar{x}-\bar{y}) = E(\bar{x}) - E(\bar{y}) = \mu_x - \mu_y$.

$$V(\bar{x}-\bar{y}) = V(\bar{x}) + V(\bar{y}) = \frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}.$$

$$= \sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right).$$

$$\xi_i = \frac{(\bar{x}-\bar{y}) - (\mu_x-\mu_y)}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N(0,1) \rightarrow ①$$

let $\chi^2 = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} (x_i-\bar{x})^2 + \sum_{i=}^{n} (y_i-\bar{y})^2\right].$

$$= \left[\sum_i (x_i-\bar{x})^2/\sigma^2\right] + \left[\sum_i (y_i-\bar{y})^2/\sigma^2\right].$$

$$= \frac{n_1 s_x^2}{\sigma^2} + \frac{n_2 s_y^2}{\sigma^2} \rightarrow ②$$

Since $n_1 s_x^2/\sigma^2$ and $n_2 s_y^2/\sigma^2$ are independent χ^2 variates with (n_1-1) and (n_1-1) d.f respectively, by the additive property of chi-square distribution χ^2 defined ① ② is a χ^2 variate with $(n_1-1)+(n_2-1)$.

i.e n_1+n_2-2 d.f further, since sample variance are independently distributed s and χ^2 are independent distribution variables.

Hence fisher's t-statistic is given by

$$t = \frac{\xi}{\sqrt{\frac{\chi^2}{n_1+n_2-2}}}$$

$$= \frac{(\bar{x}-\bar{y}) - (\mu_x-\mu_y)}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \times \frac{1}{n_1+n_2-2\left\{\sum_i (x_i-\bar{x})^2 + \sum_i (y_i-\bar{y})^2\right\}}$$

$$= \frac{(\bar{x}-\bar{y}) - (\mu_x-\mu_y)}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \text{ where } s^2 = \frac{1}{n_1+n_2-2}\left[\sum_i (x_i-\bar{x})^2 + \sum_i (y_i-\bar{y})^2\right].$$

and it follows students t-distribution with (n_1+n_2-2) d.f.

Remark:

S^2 defined in is an unbiased estimate of the common population curve σ^2, since

$$E(S^2) = \frac{1}{n_1+n_2-2} E\left[\sum_i (x_i-\bar{x})^2 + \sum_i (y_i-\bar{y})^2\right].$$

$$= \frac{1}{n_1+n_2-2} E[(n_1-1) s_x^2 + (n_2-1) s_y^2].$$

$$= \frac{1}{n_1+n_2-2} [(n_1-1) E(s_x^2) + (n_2-1) E(s_y^2)].$$

$$= \frac{1}{n_1+n_2-2} [(n_1-1)\sigma^2 + (n_2-1)\sigma^2] = \sigma^2.$$

An important dedution which is of much practical utility is discussed below.

Suppose we want to test if: (a) two independent samples x_i ($i=1,2..n_1$) and x_i ($1,2..n_2$) have been drawn from the populations with same means or (b) the two samples mean $\bar{x}$ and $\bar{y}$ differ significantly or not.

t-test for testing the significance of an observed sample correlation:

If r is the observed correlation co-efficient in a sample of n point of observations from a bivariate normal population, then proof, Fisher proved that under the null hypothesis Ho: ρ=0 i.e., population correlation co-eff is zero,

the statistic,

$$t = \frac{r}{\sqrt{(1-r^2)}}\sqrt{n-2}.$$

follows student's t-dist with (n-2) d.f. If the value of t comes out to be significant, we reject Ho at the level of significance adopted and conclude that ρ≠0 i.e 'r' is significant or correlation in the population.

If t comes out be non-significant then Ho may be accepted and we conclude that variables may be regarded as un-correlated in the population.

Application of the chi-square Test:

(10)

χ^2 dist has a large number of applications in statistics, some of when enumerated below.

(i) to test if the hypothetical value of population variance is $\sigma^2 = \sigma_0^2$.

(ii) To test the "goodness of fit".

(iii) To test the independence of attributes.

(iv) to test the homogenaity of independent estimates of the population variance.

(v) To combine various probablities obtained from independent experience give a sample test of significance.

(vi) to test the homogeneity of independent estimates of the population correlation co-efficient.

Goodness of fit test:

A very powerful test for testing the significance of the discrepancy b/w theory and experiment was given by proof. Karl pearson 1900 and in known as " chisquare test of goodness of fit ". It enables us to find it the deviation of the experiment from theory is just by chance (or) is it really due to for independency of the theory to fit the observed data.

If f_i $(1,2,\ldots n)$ is a set of observed (experimental) frequencies and e_i $(i=1,2\ldots n)$ is the corresponding set of expected (theoretical or hypothetical) frequency then Kar pearson's chisquare given by

$$\psi^2 = \sum_{i=1}^{n} \left[\frac{(f_i-e_i)^2}{e_i}\right] \left[\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} e_i\right].$$

follows chi-square dist with $(n-1)$ df.

Remark:

This is an approximate test for large values of n.

The goodness of fit test uses the chi-square test to determine if a hypothesis prob distribution for a population provides a good fit.

Decision rule:

Accept H_0 if $\psi^2 \le \psi^2_{\alpha}(n-1)$ and reject H_0 if $\psi^2 > \psi^2_{\alpha}(n-1)$ where is the calculated value of chi-square obtain on using and $\psi^2_{\alpha}(n-1)$ is to tabulated value of chisquare of $(n-1)$ df and level of significance α.

Test of independence of attributes

Contingency tables:

Let us consider two attributes A and B, A divided to s classes $B_1, B_2 \ldots B_s$ such a classification in which attributes are divided into where than two classes in known as manifold classification. The various cell frequencies can be expressed in the following table known as $r \times s$ manifold contingency table where is the number of persons possessing the attribute A_i $(i=1,2..r)$ (B_j) is the no. of persons possessing attribute B_j $(j=1,2\ldots s)$. and (A_iB_j) is the no. of persons possessing both the attributes A_i and B_j $(i=1,2\ldots r,$ $j=1,2\ldots s)$.

Also $\sum_{i=1}^{r} A_i = \sum_{j=1}^{s} B_j = N$

($\because$ where N is the total frequency!

A / B	A_1	$A_2 \cdots$	$A_i \cdots$	A_r	total
B_1	$A_1 B_1$	$A_2 B_1 \cdots$	$A_i B_i$	$A_r B_1$	B_1
B_2	$A_1 B_2$	$A_2 B_1 \cdots$	$A_i B_2$	$A_r B_2$	B_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_j	$A_1 B_j$	$A_2 B_j \cdots$	$A_i B_j$	$A_r B_j$	B_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_s	$A_1 B_s$	$A_2 B_s \cdots$	$A_i B_s$	$A_r B_s$	B_s
total	A_1	A_2	A_i	A_r	N.

The problem is to test if the two attributes A and B under consider to independent or not.

under the null hypothesis that the attributes are independent theory of frequency are calculated follows.

$P[A_i]$ = prob that a person possesses that attribute $A_i = \frac{(A_i)}{N}, i=1,2..n$.

$P[B_j]$ = prob that a person possesses the attribute $B_j = \frac{(B_j)}{N}\ j=1,2..n$.

$P[A_iB_j]$ = prob that a person possesses the attributes A_i and $B_j = P(A_i)P(B_j)$.

(By Compound prob theorem, since the attributes A_i and B_j are independent under the null hypothesis.") ∴

$$P[A_iB_j] = \left(\frac{A_i}{N}\right)\left(\frac{B_j}{N}\right) : i=1,2\ldots r.$$

$(A_iB_j)_0$ = Expected no. of persons possesing both the attributes A_i and B_j

$$= N \cdot P[A_iB_j] = \frac{A_iB_j}{N}.$$

$$\Rightarrow (A_iB_j)_0 = \frac{(A_i)(B_j)}{N}, \ (i=1,2\ldots s).$$

By using this formula, we want find out expected frequencies for each of the frequencies (A_i, B_j) $(i=1,2\ldots r, j=1,2\ldots s)$ under the null hypothesis at independence of attributes.

The exact test for the independence of attributes is very complicated but a fair degree of approximation is given, for large samples; (large N), by the χ^2 test of goodness of fit, viz...

$$\psi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \left[\frac{\left[A_i B_j - A_i B_j \right]_0^2}{A_i B_{j0}} \right]$$ (16)

$$= \sum_{i} \sum_{j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}.$$

where, f_{ij} = observed frequency for contingency table category in column i and row j.

e_{ij} = ex. frequency for contingency table category in column i and row j, which is distributed as a ψ^2 variate with $(r-1)(s-1)$ d.f [c.f Note b/w on d.f].

Remark:

$\phi^2 = \psi^2/N$ is known as mean square contingency.

Since the limits for ψ^2 and ϕ^2 very in different cases, they cannot be used for establishing the closeness of the relationship b/w quantitative characters under study.

proof: Karl pearson suggested another measure, known as co-eff of mean square contingency which is denoted by C and is given by:

$$C = \sqrt{\frac{\psi^2}{\psi^2 + N}} = \sqrt{\frac{\phi^2}{1+\phi^2}}.$$

Obviously c is always less than unity. (17)

The max value of c depends on r and s. the no. of. classes into which A and B divided. In a r×r Contingency table, the max value $c = \sqrt{\left(\frac{r-1}{r}\right)}$.

Since the max value of c differs for different classification, viz.. r × r (r = 2, 3, 4..) strictly, speaking, the values of c obtained from different types of classifications. are not Comparable.