ORIGINAL RESEARCH PAPER

Statistics

ON SOME ASPECTS OF STATISTICAL INFERENCE IN EXACT SAMPLING DISTRIBUTIONS BY USING MS-EXCEL

KEY WORDS:

K Pusphanjali — Professor, Department of statistics, S.K.University, Anantapur , Andhra Pradesh , India.

G Vijaya Lakshmi — Research Scholar , Department of statistics , S.K. University , Anantapur , Andhra Pradesh , India

ABSTRACT

We know that statistical data is nothing but a random sample of observations drawn from a Population described by a random variable whose probability distribution is unknown or partly unknown and we try to know about the properties of the population on the basis of knowledge of the properties of the sample. This inductive process of going from known sample to the unknown Population is called "Statistical Inference". The present paper gives overviews of thestatistical tools for t-test, X^2 -test and F-test and their applications in various numerical data. Statistical analysis is expected to make an important contribution to solving major Socio- economic development and activities to build for good planning and better decision – making valid Inference. Hence, it is the branch of statistics concerned with mathematical facts and data related to Business and Economic development events.

Introduction to Statistical Inference using Exact Sample Distributions and their applications

Some of the statistical Inference test based on Student's t-test, F-test and Chi-Square test and their applications.

When the sample size is ≥30, there follows normal probability law. If the sample size is <30. We cannot apply normal test because they do not follow normal probability law. i.e., there is need to know some special type of distributions which are known as Exact Sampling Distributions. Those are t, F and Chi-Square distributions.

So, their entire large sample theory, it was based on application of normal tests. If the sample size is small the distribution of various statistic's are as:

$Z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ Or $Z=\frac{\bar{x}-np}{\sqrt{npq}}$ etc., are far from normality and such normal test cannot be applied, if n is small.

Review of Literature:

The sampling tests pioneered by W.S. Gosset (1908) who wrote under the pen name of "Student". And later developed and extended by Prof. R.A. Fisher (1926). He gave a test known as t-test. In the following paper we shall discuss:

(a) t – test, (b) F – test and (c) Chi – Square test.

The exact sample tests can, however, be applied to large samples also though the converse is not true. In all the exact sample tests, the basic assumption is that, "The population from which samples are drawn is normal. i.e., Parent population is normally distributed".

1)t-Distribution:

When the sample size is smaller, the ration $z=\frac{\bar{x}-\mu}{s/\sqrt{n}}$ will follow t distribution and not the standard normal distribution. Hence, the test statistic is given as $t=\frac{\bar{x}-\mu}{s/\sqrt{n}}$ which follows normal distribution with mean 0 and 1 standard deviation. This follows a t - distribution with (n-1) degrees of freedom which can be written as t (n-1) d.f.

Applications (or) uses for student's t-test:

1. To test if the sample mean (x) differs significantly from the hypothetical value (m) of the population mean.
2. To test the significance of the difference between two sample means.
3. To test the significance of an observed sample correlation coefficient and Sample regression coefficient.
4. To test the significance of observed partial and multiple correlation coefficients.
5. To test the significance of paired samples.

Assumptions for Student's t-test:

1. The population standard deviation (σ)is unknown.
2. The sample observations are independent that the sample size is random.
3. The parent population from which the samples are drawn is normal.

(a) t- test for single Mean:

The test procedure is has follows:

1. Form the null hypothesis $H_0:\mu=\mu_0$
   i.e.,There is no significance difference between the sample mean and the population mean
2. Alternate hypothesis $H_1:\mu\neq\mu_0$ ($\mu>\mu_0$ or $\mu>\mu_0$)
3. Level of Significance: The level may be fixed at either 5% or 1%.
4. Test statistic: $t=\frac{\bar{x}-\mu}{s/\sqrt{n}}$ ~ t - distribution with (n-1) degrees of freedom. Where $\bar{x}=\frac{\sum x_i}{n}$ and $s=\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$
5. Find the table value of t corresponding to (n-1) d.f. and the specified level of significance.
6. Inference: If tcal < ttab, we accept the null hypothesis H0. We conclude that there is no significant difference sample mean and population mean (or) if tcal > ttab , we reject the null hypothesis H0. (i.e.) we accept the alternative hypothesis and conclude that there is significant difference between the sample mean and the population mean.

(a) t- test for the difference between two sample Means:

The t- test procedure is has follows:

1. Form the null hypothesis $H_0:\mu_x=\mu_y$
   i.e., there is no significance difference between the two sample means.
2. Level of Significance: The level may be fixed at either 5% or 1%.
3. Test statistic $t=\frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ ~ t - distribution with (n1+n2-2) degrees of freedom. Where

$$\bar{x}=\frac{\sum_{i=1}^{n_1} x_i}{n_1},\ \bar{y}=\frac{\sum_{j=1}^{n_2} y_j}{n_2}\ \text{and}\ s^2=\frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(x_i-\bar{x})^2+\sum_{j=1}^{n_2}(y_j-\bar{y})^2\right]$$

4. Find the table value of t corresponding to (n1+n2-2) d.f. and the specified level of significance.
5. Inference: If tcal < ttab, we accept the null hypothesis H0. We conclude that there is no significant difference between two sample means. Otherwise, we can reject the null hypothesis H0.

Assumptions of t- test for difference of two sample means:

a) Parent population from which the samples have been drawn are normally distributed.

b) The population variances are equal and unknown.

c) The two samples are random and independent of each other.

Illustration: solve by Ms-Excel

Below are given the gain in weights (in lbs) of pigs fed on diets A and B. Gain in Weight

Diet A : 25,32,30,34,24,14,32,24,30,31,35,25.

Diet B : 44,34,22,10,47,31,40,30,32,35,18,21,35,29,22.

Test, if the two diets differ significantly as regards their effect on increase in weight.

Solution:

Here, we state the null hypothesis as:

H_0: $\mu_x = \mu_y$ i.e., there is no significant difference between the mean increase in weight due to diets A and B.

Steps to calculate t calculated value in Ms-Excel: Step 1: Enter the given data in Excel worksheet. Step 2: Select Toolbar and go to Add-Ins.

Step 3: Select Analysis toolpak and Analysis toolpak-VBA.

Step 4: Select Data Analysis and open it separate wizard "Data Analysis". The quick order for select Data Analysis Pack as:

Enter → Toolbar → Add - Ins → DataAnalysis

Step 5: Select " t-test two-sample Assuming equal variances " and press Ok button.

Step 6: Open Wizard of t-test: select data series and hypothesized mean difference is zero. Step 7: Finally select New worksheet ply and press Ok button.

The below table shows the t-test for significant difference between two sample means:

t-Test: Two-Sample Assuming Equal Variances		
	X	Y
Mean	28	30
Variance	34.5455	100.7143
Observations	12	15
Pooled Variance	71.6	
Hypothesized Mean Difference	0	
d.f.	25	
t Stat	-0.6103	
P(T<=t) one-tail	0.2736	
t Critical one-tail	1.7081	
P(T<=t) two-tail	0.5472	
t Critical two-tail	2.0595	

Inference: Now, we can compare t cal value and t cri value at required level of significance.cal Here t value = 0.6103 < t cri value with 25 d.f. (both one-tail and two-tail test) are accepted the null hypothesis H0. Hence, we can conclude that there is no significant difference between the mean increase in weight due to diets A and B.

Paired t- test for the difference between two sample Means:

The paired t - test procedure is has follows:

Let us consider the case when (i) the sample sizes are equal.i.e,n1 = n2 = n (say), and (ii) thetwo samples are not independent but the sample observations are paired together. i.e., the pair of observations (xi , yi), wherei = 1,2,3,. , n. corresponding to the same ith sample unit.The problem is to test if the sample means differ significantly or not.Consider the increments di = xi - yi ; here i = 1, 2, 3,…, n.Under the null hypothesis as:

H0: the increments are due to fluctuations of sampling. i.e., the drug is not responsible for these increments.

To test the above H0 , we can use Student's paired t-test is given by

$$t = \frac{\bar{d}}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)} \text{ d.f.}$$

where $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$

Finally, we can compare t cal value and t cri value with (n-1) d.f. at required level of significance. Then, we can draw the conclusions accordingly.

Illustration:

A certain stimulus administered to each of the 12 patients resulted in the following increase of blood pressure: 5,2,8,-1,3,0,-2,1,5,0,4,6. Can it be concluded that the stimulus will, in general, be accompanied by an increase in blood pressure?

Solution:

Here we are given the increments in blood pressure. i.e.,di = xi - yi ; here i=1,2,3,…..,n.

Now, we can state the null hypothesis as: . H_0: $\mu_x = \mu_y$ .i.e., there is no significant difference in the blood pressure readings of the patients before and after the drug. Or the given increments are just by chance and not due to the stimulus.

To test, under H0, we can use the test statistic is t = $\frac{\bar{d}}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)}$ d.f.

d	5	2	8	-1	3	0	-2	1	5	0	4	6	31
d2	25	4	64	1	9	0	4	1	25	0	16	36	185

$$\bar{d} = \frac{\sum d_i}{n} = \frac{31}{12} = 2.58 \text{ and } s^2 = \frac{1}{n-1}\sum (d_i - \bar{d})^2 = \frac{1}{n-1}\left[\sum d_i^2 - \frac{\left(\sum d_i\right)^2}{n}\right]$$

$$s^2 = \frac{1}{12-1}\left[185 - \frac{(31)^2}{12}\right] = \frac{1}{11}\left[185 - 80.08\right] = 9.5382$$

t becomes $t = \frac{\bar{d}}{\frac{s}{\sqrt{n}}} = \frac{2.58}{\sqrt{9.5382}/\sqrt{12}} = 2.89$

Therefore, t cal value = 2.89 and t cri value with 11 d.f. at 5% los = 1.80 So, if t cal value > t cri value with 11 d.f. at 5% los, then we reject H0.

Hence, we conclude that the stimulus will, in general, be accompanied by an increase in blood pressure.

2) F-Distribution or F-Statistic:

If x_1^2 and x_2^2 are two independent x^2-variates with n1 and n2 d.f . respectively. Then, F- Statistic is defined by Or F- $= \frac{\left(\frac{\chi_1^2}{n_1}\right)}{\left(\frac{\chi_2^2}{n_2}\right)}$ statistics is defined as the ratio of two independent - x^2 variates and by their corresponding to its respective d.f.'s and it follows Snedecor's F- distribution with n1 and n2 d.f. respectively. i.e., F(n1,n2)d.f.

$$F = \frac{\left(\frac{\chi_1^2}{n_1}\right)}{\left(\frac{\chi_2^2}{n_2}\right)} \sim F_{(n1,\ n2)} \text{ d.f.}$$

Applications: 1. F-test for equality of population variances.

2. F- test for testing the significance of an observed multiple correlation coefficient.
3. F- test for equality of several means.

1) F- test for equality of population variances:

The t- test procedure is has follows:

Form the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ i.e., the population variances are equal.

2. Level of Significance: The level may be fixed at either 5% or 1%.
3. Test statistic $F = \frac{S_1^2}{S_2^2}$ F - distribution with (n1-1,n2-1) degrees of freedom. Where

$\bar{x} = \frac{\sum x_i}{n_1}$, $\bar{y} = \frac{\sum y_i}{n_2}$, $S_1^2 = \frac{1}{n_1 - 1}\left[\sum (x_i - \bar{x})^2\right]$ and $S_2^2 = \frac{1}{n_2 - 1}\left[\sum (y_i - \bar{y})^2\right]$ are

unbiased estimates of the common population variance 2.

F- statistics become F= $\frac{\left(\frac{\chi_1^2}{n_1 - 1}\right)}{\left(\frac{\chi_2^2}{n_2 - 1}\right)}$ - F (n1, n2) d.f. at required level of significance.

Here, $\chi_1^2 = \frac{n_1 S_1^2}{\sigma_1^2}$ and $\chi_2^2 = \frac{n_2 S_2^2}{\sigma_2^2}$ are independent chi-square variates with (n1-1) and (n2-1) d.f. respectively.

4.. Find the table value of F corresponding to (n1-1,n2-1) d.f. and the specified level of significance.

5. Inference: If Fcal < Ftab, we accept the null hypothesis H0. We conclude that the population variances are equal. Otherwise, we can reject the null hypothesis H_0.

Illustration: Solve by Ms- Excel:

Two random samples drawn from normal populations are:

Sample I	18	16	24	26	20	22	17	24	25	19
Sample II	28	34	42	36	33	35	38	27	42	41

Test, whether the two populations have the same variance.

Solution: Let the null hypothesis be H :σ2 = σ2 = σ2 , where σ2 and σ2 are the variances of the two populations =s(2.2)\-/2-2)

Under the H0 , the test statistic as F $= \frac{S_1^2}{S_2^2}$, $S_1 > S_2$ or $\frac{S_2^2}{S_1^2}$, $S_2 > S_1$

Where $S_1^2 = \frac{1}{n_1 - 1}\left[\sum (x_i - \bar{x})^2\right]$ and $S_2^2 = \frac{1}{n_2 - 1}\left[\sum (y_i - \bar{y})^2\right]$

Here, n1=n2=10.

Steps to calculate t calculated value in Ms-Excel:

Step 1: Enter the given data in Excel worksheet. Step 2: Select Toolbar and go to Add-Ins.

Step 3: Select Analysis toolpak and Analysis toolpak-VBA.

Step 4: Select Data Analysis and open it separate wizard "Data Analysis". The quick order for select Data Analysis Pack as:

Enter →Toolbar → Add - Ins →DataAnalysis

Step 5: Select " F-test two-sample for variances " and press Ok button. Step 6: Open Wizard of F-test: select data

Step 7: Finally select New worksheet ply and press Ok button.

F-Test Two-Sample for Variances		
	Sample I	Sample II
Mean	21.1	35.6
Variance	12.7667	28.7111
Observations	10	10
d.f.	9	9
F stat	2.2489	
F Critical two-tail	3.18	

Inference:

Now, if Fcal value < Fcri value with (9,9) d.f. at 5% los, then we accept H0. Hence, we can conclude that two population variances are equal.

3) Chi-Square Distribution or Chi-Square variate:

The square of the standard normal variate is known as a χ^2 – variate with 1 d.f. Thus, if X follows N(μ, σ^2), then $Z = \frac{X-\mu}{\sigma}$ ~ N(0,1) and $Z^2 = \left(\frac{X-\mu}{\sigma}\right)^2$ ...

In general, if X_i (i=1,2,3,...,n) are 'n' independent normal variates with mean μ_i and variance σ_i^2 respectively, then $Z = \sum\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2$ or $\chi^2 = \sum\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2$ with n d.f.

Applications of Chi-Square Distribution:

Chi-Square distribution is a large number of applications in Bio-Statistics.

a) To test the independence of attributes.
b) To test the goodness of fit.
c) 0
c) To test, if the hypothetical value of the population variance is σ2 = σ2 (say).
d) To combine various probabilities obtained from independent experiments to give a single test of significance.

Conditions for the validity of Chi-Square test:

1. The sample observations should be independent.
2. Constraints on the cell frequencies, if any, should be linear.
3. N, the total frequency should be reasonably large (say,>50).
4. If any theoretical cell frequency should be less than 5, then for the application of Chi- Square test, it is pooled with the (Proceeding or Succeeding) frequency, so that, the pooled frequency is more than 5 and finally adjust for the degree of freedom lost in pooling technique.

Chi-Square test for Goodness of fit:

A very powerful test for testing the significance of the difference between theory and experiment was given by Prof. Karl Pearson (1900) and is known as "Chi-Square test for Goodness of fit". It make possible us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.If Oi (i=1,2,3,....,n) is a set of observed(experimental) frequencies and E i (i=1,2,3,....,n) is the corresponding to a set of expected (theoretical or hypothetical) frequencies , then Karl Pearson's Chi-Square statistic is given by $\chi^2 = \sum\left(\frac{(O_i - E_i)^2}{E_i}\right)$, $\sum O_i = \sum E_i$ ~ chi- Square distribution with (n-1)d.f.

Now, we can compare X^2value and X^2value with (n-1) d.f. at required los, then we can draw the conclusions accordingly.

Illustration:

Fit a Poisson distribution and test its goodness of fit for the following data.

x	0	1	2	3	4	5
f	112	63	20	3	1	1

Solution:

In order to fit Poisson distribution to the given data, we take the mean l of the Poisson distribution equal to the mean and its probability mass function is given by

$$P(X = x) = \frac{e^{\lambda}\lambda^x}{x!}, \quad x = 0,1,2,3\ldots\ldots$$

x	f	fx
0	112	0
1	63	63
2	20	40
3	3	9
4	1	4
5	1	5
Totals	200	121

$$\text{Mean} = \frac{\sum fx}{N} = \frac{121}{200} = 0.605$$

$$\lambda = \bar{x} = 0.605$$

The fitted Poisson distribution is given by

$$P(X=x) = \frac{e^{0.605} 0.605^{x}}{x}, x=0,1,2,3....$$

The expected frequencies are calculated for the following formula as

$$f(x) = N.\ P(x) = N.\left(\frac{e^{0.605} 0.605^{x}}{x!}\right), x = 0, 1, 2, 3, ...$$

Oi	Ei	(O - E)2
112	109.21	0.0713
63	66.07	0.1426
20	19.99	0
3	4.03	0.0178
1	0.61	
1	0.07	
c 2 =	0.2318	

$X_{cal}^{2} = 0.2318$ and x_{2dof} 5%los = 5.99

Inference:
so, $x_{cal}^{2}=0.2318< x_{dof}^{2}$ at 5%los=5.99, we can accept H0 Hences, we conclude that Poisson distribution is a goodness of fit to the given data.

CONCLUSIONS:

Finally, the present study was going to applications of Various Socio- economic Research problems is done for applying the Statistical Inference like t, F and Chi- Square test and also, their applications and importance.

REFERENCES:

1. Telugu Academy English Medium textbook for "Statisti cal Methods and Inference".
2. Fisher, R. A. [1956]: "Statistical Methods and Scientific Inference", Oliver & Boyd, London.
3. B.A./B.Sc. Statistics,. "Statistical Inference" by D.V.L.N. Jogiraju, Palnati Sudarsan Published with Kalyani Publishers, New Delhi, India.
4. Prof. K.V.S. Sarma, "Statistics Made Simple do it yourself on PC ", Second Edition, PHI, 2010.
5. Gupta S.C. and Kapoor V.K. , "Fundamentals of mathematical Statistics" by S. Chand Publications.