e-Pathshala पाठशाला
A Gateway to All Post Graduate Courses

MHRD
Govt. of India

An MHRD Project under its National Mission on Education throught ICT (NME-ICT)

Subject: Library and Information Science

Production of Courseware

e-Content for Post Graduate Courses

Paper No : 10 Informetrics and Scientometrics

Module : 19 Basics of Testing of Hypotheses

Development Team

Principal Investigator & Subject Coordinator	Dr. Jagdish Arora, Director INFLIBNET Centre, Gandhinagar
Paper Coordinator	Dr I K Ravichandra Rao Retd Professor, Documentation Research and Training Centre
Content Writer	Dr I K Ravichandra Rao Professor, Documentation Research and Training Centre
Content Reviewer	Prof A Y Asundi Retd Professor, Bangalore University

Module 19

Basics of Testing of Hypothesis

Module Structure

OBJECTIVES

- To study an overview of testing of hypotheses
- To study procedure/steps in testing of hypotheses and related concepts

SUMMARY

Most often, the research process is incomplete without testing of hypotheses. As such there are two types of hypotheses normally referred in statistical testing viz. Null Hypothesis and Alternative Hypothesis. A research process involves:

1. Identify the general problem(s);
2. Conduct literature search;
3. Decide the design methodology;
4. Collect the data either for the population or for a sample;
5. Analyze the data;
6. Report the result; and
7. Refine the hypotheses.

It is in step 1, we generally formulate the hypotheses and in step 5, we test the hypotheses. Based on the results of testing of hypotheses, we generalize the results. The Unit 19 discusses basically the z-test, t-test and the Chi Square test.

1 INTRODUCTION

Statistical analysis aims at inferring about a population based on the information/data contained in a sample. There are methods for making inferences which are usually based on statistical tests of hypotheses. Below we discuss only those aspects concerning testing of hypothesis.

A hypothesis is a well-defined statement. However, the word hypothesis in science generally refers to a definite interpretation of a given set of facts, which is put forth as a tentative assumptions and remain partially or wholly unverified. A simple definition of hypothesis as given by Luniberg "is a tentative generalisation, the validity of which remains to be tested. In this context testing of hypothesis, with relevant statistical data becomes important either to accept or reject the tentative assumption.

1.1 Why Test Hypothesis?

Science does not accept anything as valid knowledge, until satisfactory tests confirm its validity. Therefore they need to be tested through research process for their acceptance or rejection. The hypothesis is normally tested by making use of a pre-defined assertion, rule which is applied to sample data and direct the research process in deciding to accept or reject the hypothesis. The process of testing hypothesis embodies the major part of research process. A hypothesis is tested on the basis of facts.

1.2 Types of Hypotheses and Notations

The two hypotheses in a statistical test are normally referred to as:

a) Null Hypothesis, and
b) Alternative Hypothesis.

a) The Null Hypothesis is a very useful tool in testing the significance of difference. In its simple form, the hypothesis asserts that there is no true difference in the sample and population in particular matter under consideration and that thedifference found is accidental, unimportant, arising out of fluctuations of sampling. A simple definition of hypothesis is that it is a hypothesis which is being tested.

b) The Alternative hypothesis specifies those values that the researcher considers to be true, and hopes that the sample data leads to acceptance of this whole hypothesis as true. In other words, when a null hypothesis is rejected, then alternative hypothesis is likely to be accepted. For example,

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

where μ is the population mean and μ_0 is the hypothesised value of the population mean.

1.3 Errors in Hypothesis Testing

When accepting or rejecting a null hypothesis, we may commit an error. For instance, we may reject $\mathbf{H_0}$ when it is correct; and accept $\mathbf{H_0}$ when it is not correct. These two errors are called Type I error and Type II error respectively. The probability of making a Type I error is denoted by α. The probability of making a Type II error is denoted as β. This is shown in the tabular form as below:

Conclusion of the Test	**H_0 True**	**H_0 False**
Accept H_0	Correct	Wrong (Type II Error)
Reject H_0	Wrong (Type I Error)	Correct

1.4 Empirical Test of Hypothesis

For the purpose of understanding the testing of hypotheses, let us discuss an experimental situation. Consider a condition of verifying the manufacturer's statement about its product. For example, let us take a case of investigating the container weights specified on the labels of wheat products of the manufacturer. In order to demonstrate the hypothesis testing procedure, let us show how a test on label accuracy could be made for the company's 2 kg packet of wheat flour.

The first assumption is that labels are correct. This assumption or hypothesis is subjected to a test by providing evidence regarding the truth of the claim or assumption. There are three possibilities in the case of 2 k.g. wheat flour packets. It is possible that the mean weight for the population of 2 kg packets could be;

i) $\geq$ 2 kg or
ii) $\leq$ 2 kg or
iii) = 2 kg.

In this situation, we have to determine whether or not the population mean (of wheat flour packets) **$\mu = \mu_0$ (say, 2).** How to determine? This is discussed below:

A statement like $\mu = \mu_0$ ***or*** $\mu \geq \mu_0$ **or** $\mu \leq \mu_0$ is called a hypothesis. As said in section 1.2, a hypothesis that is being tested is called the **Null hypothesis**. It is denoted by **H_o.** The hypothesis that we are willing to accept if we do not accept the null hypothesis is called the **Alternative hypothesis.** The two hypotheses, the Null Hypothesis *(***H_o**) and the Alternative Hypothesis (**H_1**) are so constructed that if one is correct the other is wrong. It is denoted by **H_1.** Generally, the Null Hypothesis and Alternative Hypothesis for testing the mean will be shown in the following forms:

	Null Hypothesis		Alternative Hypothesis
- Case 1:	$H_0 : \mu \geq \mu_0$	**or**	$H_1 : \mu < \mu_0$
- Case 2:	$H_0 : \mu \leq \mu_0$	**or**	$H_1 : \mu > \mu_0$
- Case 3:	$H_0 : \mu = \mu_0$	**or**	$H_1 : \mu \neq \mu_0$

In establishing the critical value for a particular hypothesis testing situation, we always assume that the **H_o holds as equality**. This allows us to control the maximum probability of Type I error. Thus, in cases 1-3 above, the null hypotheses may be treated as H_0: $\mu = \mu_0$. Now with an assumption that the null hypothesis is true, let us select a sample from the population. If the sample results do not differ significantly from the **assumed null hypothesis, we accept H_0 as being true. If the sample results differ** significantly from the hypothesis, we reject **H_0** and conclude that the alternate hypothesis **H_1** is true.

2 Z-TEST: AN EXPLANATION

In z-test, the distribution of the test statistic under the null hypothesis is approximated by a normal distribution. From the central limit theorem, we know that the sample mean $\bar{x}$ follows normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. That is, the variable $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ follows asymptotically a normal distribution with mean zero and standard deviation one. $\bar{x}$ is the sample mean; μ is the population mean; **σ** is the standard deviation of the population and n is the sample size. Hence for a large n, we have, $P = \left\{ \left| \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right| \leq 1.96 \right\} = 0.95$

This is explained further in the following Figure:

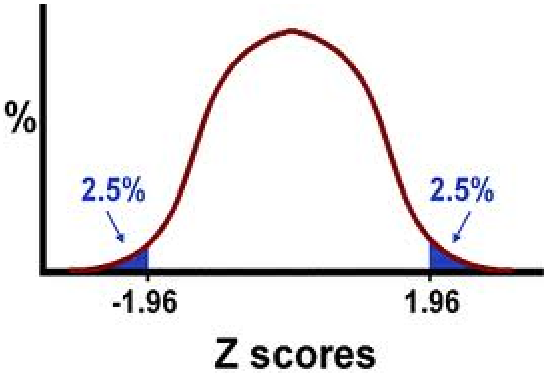


Fig.1: A Normal Distribution Curve

The shaded area in the above figure is called Rejection Area or Critical Region***;*** the critical region, in fact, in this figure is the set (z: | z | > 1.96). The values –1.96 and 1.96 are called the critical values (these values are for case 3, in section 1.3) . A general rule to define the critical region is to choose z so that

$$P = \left\{ \left| \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right| \quad \leq \quad | z_{\alpha/2} | \right\} = 1 - \alpha$$

$| z_{\alpha/2} |$ is 1.96 for α = 0.05. The α is known as the size of the Type I error (= P (rejecting $H_o|H_o$ is true)). It is normally denoted by α; also, it is called α-significance level of the test. It determines a set; if the experiment yields a value of the quantity we are using to test the null hypothesis falls in this set, and then we reject $H_{o.}$

So, while testing a hypothesis, if the computed value of z is greater than | $z_{\alpha/2}$ | (if α = 0.05, |$z_{\alpha/2}$| = 1.96), the value of z falls in the critical region. Hence we reject H_o. In other words, if the experiment is conducted a number of times and follow the test procedure mentioned above, it is likely that 5% of the times, we may commit Type I error — rejecting H_o when it should have been accepted. While computing the value of z from the sample data, we may use the sample standard deviation instead of *z* which is usually unknown.

Most often, as mentioned earlier, we test the null hypothesis $\mathbf{H_o: \mu = \mu_0}$ against one of the following alternative hypothesis:

1) $\mathbf{H_1 : \mu < \mu_0}$

2) $\mathbf{H_1 : \mu > \mu_0}$

3) $\mathbf{H_1 : \mu \neq \mu_0}$

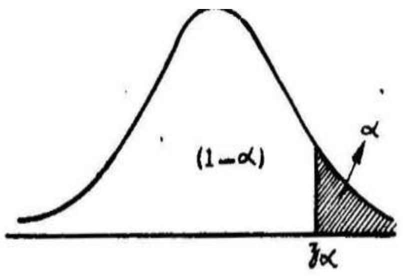
In such cases, the critical values (±zα or ±zα/2) are given by:

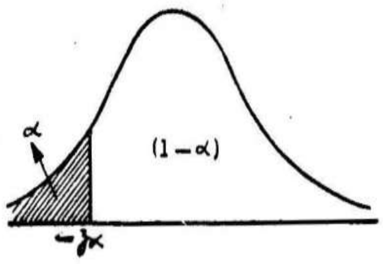
For (1): $P\left\{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \ge -z_\alpha\right\} = 1-\alpha$

For (2): $P\left\{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le z_\alpha\right\} = 1-\alpha$ and

For (3): $P = \left\{\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| \le |z_{\alpha/2}|\right\} = 1-\alpha$

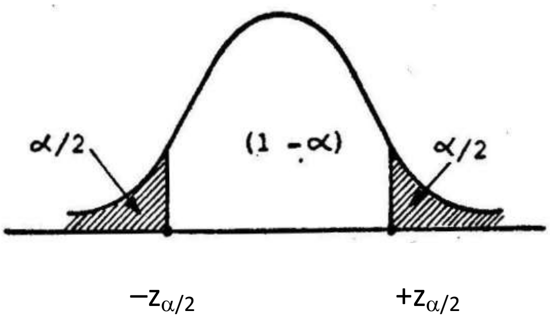
Further, if H_1: $\mu < \mu_0$ or H_1: $\mu > \mu_0$, the test is also called one-sided test; otherwise, it is called two-sided test. The critical regions for the above three alternative hypotheses are shown in the following figures.





H_1 : $\mu > \mu_0$

H_1 : $\mu < \mu_0$



H_1 : $\mu \neq \mu_0$

Case 1: H_0 : $\mu \geq \mu_0$ H_1 : $\mu < \mu_{0.}$ -- One-Tailed Hypothesis Test About Population Mean

Here, we treat the null hypothesis as $\mathbf{H_0 : \mu = \mu_0}$ instead of $\mathbf{H_0 : \mu \geq \mu_0}$, as explained in section 1.2. In this case, the decision rule is **Accept H_0 if $\bar{x} \geq c$ and Reject H_0 if $\bar{x} < c$,** c is called the critical value for the test; it is given by c = $\mu_0 - z_\alpha \sigma_{\bar{x}}$. The value of z_α can be obtained from the normal distribution table for the given α.

Case 2: H_0: $\mu \leq \mu_0$ H_1 : $\mu > \mu_{0.}$ -- One-Tailed Hypothesis Test About Population Mean

Here, we treat the null hypothesis as **H_0: $\mu = \mu_0$ instead** of **H_0: $\mu \geq \mu_0$**, as explained in section 1.2. Then the decision rule is **Accept H_0 if $\bar{x} \leq c$ and Reject H_0 if $\bar{x} > c$.** where c is called the critical value for the test; it is given by c = $\mu_0 + z_\alpha \sigma_{\bar{x}}$. The value of z_α can be obtained from the normal distribution table for the given α.

Case 3: H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_{0.}$ -- Two-Tailed Hypothesis Test About Population Mean

In this case the decision rule is **Accept H_0 if $c_1 \leq \bar{x} \leq c_2$ Reject H_0 if $\bar{x} < c_1$ or if $\bar{x} > c_2$.** where $c_1 = \mu_0 - z_{\alpha/2}\sigma_{\bar{x}}$ *and* $c_2 = \mu_0 + z_{\alpha/2}\sigma_{\bar{x}}$. The value of $z_{\alpha/2}$ can be obtained from the normal distribution table for the given α.

In all the above three cases, if σ is unknown, then use the sample standard deviation (s); in that case, n must be sufficiently large; at least n > 30.

2.1 A Confidence Interval Approach to Test a Hypothesis of the Form H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$

Select a simple random sample from a population and use the value of the mean $\bar{x}$ to develop the confidence interval. Let $H_0 : \mu = \mathbf{\mu_0}$ $H_1 : \mu \neq \mathbf{\mu_0}$ **.** A sample of n observations gives a sample mean of $\bar{x}$ and gives the standard error of $\boldsymbol{\sigma}_{\bar{x}}$ .Using these results along with $z_{\alpha/2} = 1.96$, the confidence interval becomes $\bar{x} \pm z_{\alpha/2}\ \boldsymbol{\sigma}_{\bar{x}}$. If the μ_0 falls in this range, we may accept H_0 otherwise we may reject H_0

3 PROCEDURE INVOLVED IN TESTING OF HYPOTHESES

The following steps are involved in a test of significance:

Step 1: Formulate the null and alternative hypotheses. For example:

a) $H_0 : \mu \geq \mu_0 \quad H_1 : \mu < \mu_0$ or
b) $H_0 : \mu \leq \mu_0 \quad H_1 : \mu > \mu_0$ or
c) $H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_{o.}$

Step 2: Fix the value of α; that is, deciding, the level of significance. Usually, we fix $\alpha = 0.5$ or $\alpha = 0.01$.

Step 3: Select a sample of n units; compute sample mean $\bar{x}$. Then compute the following test statistic, under the assumption that the null hypothesis is true; so, replace the μ by $\mu_{0,}$ while computing. That is,

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

Further, we assume that σ is known. However, if it is unknown, *s* can be used for a sufficiently large *n*)

Step 4: Determine the critical values. For (a) and (b) the critical value c is given by μ_0 - $z_\alpha \sigma_{\bar{x}}$ and μ_0 + $z_\alpha \sigma_{\bar{x}}$ respectively. The value of z_α can be obtained from the normal distribution table for a given α. For (c), the critical values are given by c_1 and c_2; The c_1 and c_2 are given by μ_0 - $z_{\alpha/2}\sigma_{\bar{x}}$ and μ_0 + $z_{\alpha/2}\sigma_{\bar{x}}$ respectively. The value of $z_{\alpha/2}$ can be obtained from the normal distribution table for a given α.

Step 5: 1) For (a) and (b), accept the null hypothesis H_0 if $\bar{x} \geq c$, otherwise reject H_0. For (c), accept H_0, if $c_1 \leq \bar{x} \leq c_2$ is true, otherwise reject H_0.

2) Or if the computed value of $|z| \geq |z_\alpha|$ (for (a) and (b)) or $|z_{\alpha/2}|$ (for (c)) reject H_0, otherwise accept H_0.

4 EXAMPLES WITH DIFFERENT METHODS:

a) Example 1

Consider a sample of 36 Units; sample mean ($\bar{x}$) is 2.92 and **σ** is 0.18. Test whether **$H_0 : \mu \geq \mu_0 \quad H_1 : \mu < \mu_0; \mu_0 = 3$.**

Method 1:

Let $\alpha = 0.01$ and $|z_\alpha| = 2.33$; $\sigma_{\bar{x}} = 0.18/\sqrt{36} = 0.03$. Then c = μ_0 - $z_{\alpha/2}\sigma_{\bar{x}}$ = 3.0 -2.33*0.03 = 2.93.

Since $\bar{x}$ (= 2.92) < c (2.93), reject H_0 that it is $\mu \geq 3$.

Method 2

Let α = 0.01 and $|z_{\alpha/2}|$ = 2.57; $\sigma_{\bar{x}}$ = 0.18/$\sqrt{36}$ = 0.03. Then

z = $\frac{\bar{x}-\mu_0}{\sigma_{\bar{x}}} = \frac{2.92-3}{0.03}$ = -2.66

Since $|z| > |z_{\alpha/2}|$, reject H_0.

b) Example 2

Mean age ($\bar{x}$) of a sample of students is given by 20.3; n =100, $\sigma_{\bar{x}} = 0.5$

Test whether $H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$; μ_0 = 21.0.

Method 1:

Let α = 0.05 and $|z_{\alpha/2}|$ = 1.96; $\sigma_{\bar{x}}$ = 0.05. Then

c_1 = μ_0 - $z_{\alpha/2}\sigma_{\bar{x}}$ = 21.0 - (1.96)(0.50 = 20.02.

c_2 = μ_0 - $z_{\alpha/2}\sigma_{\bar{x}}$ = 21.0 - (1.96)(0.50 = 21.98

Since 20.02 ≤ $\bar{x}$ (= 20.30) ≤ 21.98, H_0 that $\mu = \mu_0$ (21.0) is accepted.

Method 2

Let α = 0.05 and $|z_{\alpha/2}|$ = 1.96; $\sigma_{\bar{x}}$ = 0.05. Then

z = $\frac{\bar{x}-\mu_0}{\sigma_{\bar{x}}} = \frac{20.3-21.0}{0.05}$ = -1.4

Since $|z| \leq |z_{\alpha/2}|$, accept H_0.

5 USE OF P-VALUE

In statistical significance testing the ***p*-value** is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true; the p-value is the smallest value of α for which the given sample outcome would lead to accepting H_0. The decision rule is to accept H_0 if the p-value ≥ α and reject H_0 if p-value < α. In the Example 1, above, the p-value is 0.0038 (i.e., P ($\bar{x}$ > 2.92) which is also equivalent to P(z ≥ 2.66)); the p-value is < α and thus it leads to rejecting H_0.

6 Z-TEST: σ IS UNKNOWN

If σ^2 is unknown, it is usually estimated from the sample variance. However, the sample variance is not a reliable estimate of σ^2. To get a good approximation for σ^2, we can use the formula: $s_e^2 = \frac{1}{n-1}\Sigma(x_i - \overline{X})^2$.

In other words, $s_e^2 = \frac{n}{n-1}s^2$, where $s^2 = \frac{\Sigma(x_i - \overline{X})^2}{n}$. We can now use s_e^2 instead of σ^2 in testing the null hypothesis. To test the null hypothesis, H_o: **μ = μ₀,** compute the statistic: $\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ by replacing μ by μ_0 and **σ** by s_e; then use the *z*-test as discussed above (for large n only). But, for a small *n* (say *n* < 30) the statistic $\frac{\overline{x}-\mu}{s_e/\sqrt{n}}$ follows a Student-*t* distribution provided the sample is drawn from a normal population with mean μ; the variate is called *t*-statistic. The *t*-distribution has a single parameter known as the degrees of freedom. It is calculated by subtracting the number of restrictions placed on the calculation of the statistic (the number of values that are free to vary in the calculation of the statistic) from the sample size and . So, the variable $\frac{\overline{x}-\mu}{s_e/\sqrt{n}}$ follows a *t*-distribution with (n— 1) degree of freedom. The procedure for testing the null hypothesis using the *t*-statistic is given below. *This test is called t-test.*

6.1
Procedure for t-test

Step 1: Formulate the null and alternative hypotheses. For example,

a) $H_0 : \mu \geq \mu_0 \quad H_1 : \mu < \mu_0$
b) $H_0 : \mu \leq \mu_0 \quad H_1 : \mu > \mu_0$
c) $H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_{0.}$

Step 2: Decide the level of significance (α) and determine the critical values for a given degree of freedom from the *t*-table. For (a) and (b), the critical value is given by $|t_\alpha|$ such that $P(|t| \leq |t_\alpha|)= 1 - \alpha$. For (a), the critical values are given by at $|t_{\alpha/2}|$ such that $P(|t| \leq |t_{\alpha/2}|) = 1-\alpha$.

Step 3: Compute $t = \dfrac{\overline{x} - \mu}{s_e / \sqrt{n}}$ under the assumption that μ = μ₀.

Step 4: Accept the null hypothesis if *t* is less than the critical value; otherwise reject the null hypothesis.

7. Test of Difference between Two Population Means

Let us now deal with two populations for which standard deviations (σ_1 and σ_2) are known. However, the means (μ_1 and μ_2) are not known. Under the circumstances, can we test whether or not $\mu_1 = \mu_2$; i.e., $\mu_1 - \mu_2 = 0$? So in such cases, usually we would like to test

$H_o : \mu_1 - \mu_2 = D_o$ against a specified alternative hypothesis. It may be any one of the following:

$H_1 : \mu_1 - \mu_2 \neq D_o$
$H_1 : \mu_1 - \mu_2 > D_o$
$H_1 : \mu_1 - \mu_2 < D_o$

If D_o=0, we are actually testing whether or not $\mu_1 = \mu_2$. To test whether or not H_o is true against a specified H_1 we consider the difference between $\overline{x}_1$ and $\overline{x}_2$ and their distribution in repeated samples. It has been shown in the probability theory that for repeated independent randomly drawn samples, $\overline{x}_1 - \overline{x}_2$ follows a normal distribution with mean $\mu_1 - \mu_2$ and standard deviation $\left(\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)^{\frac{1}{2}}$.

Hence, $z = \dfrac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\left(\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)^{\frac{1}{2}}}$ is a standardized normal variate.

If H_o: $\mu_1 - \mu_2 = D_o$ is true, z becomes

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - D_o}{\left(\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)^{\frac{1}{2}}}$$

Thus, we can use the *z*-test to test the null hypothesis. The procedure is similar to the z-test explained earlier. If X_1 and X_2 follow normal distributions and if and σ_1 and σ_2 are unknown, for small samples, we can use the *t*-test. That is,

$$t = \frac{(\bar{x}_1 - \bar{x}_2 - D_o)}{s_d}\left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2}$$

$$s_d = \left[\frac{\Sigma(x_{ti} - \bar{x}_t)^2 + \Sigma(x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}\right]^{\frac{1}{2}}$$

t-statistic in this case has $n_1 + n_2 - 2$ degrees of freedom. However, for large samples, even if σ_1 and σ_2 are unknown, we can use the *z*-test. In this case, we use s_1 and s_2 instead of σ_1 and σ_2. That is,

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_o}{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^{\frac{1}{2}}}$$

where $s_1^2 = \dfrac{\Sigma(x_{ti} - \bar{x}_1)^2}{n_1 - 1}$ and $s_2^2 = \dfrac{\Sigma(x_{2i} - \bar{x}_2)^2}{n_2 - 1}$

8 TESTS ABOUT PROPORTIONS

A hypothesis that is tested with respect to the theoretical proportion of successes is that $\mathbf{P} = \mathbf{P_o}$ (i.e. H_o: $P = P_o$). An alternative hypothesis is that $\mathbf{P} \neq \mathbf{P_o}$ ($\mathbf{H_t}$: $P \neq P_o$). From the probability theory, we know that for a large *n*, the binomial distribution (with mean *np* and standard deviation $\sqrt{npq}$) tends to a normal distribution. So, when we perform a binomial experiment, *n* times, if the null hypothesis H_o.: $P = P_o$ is true, then the following statistic:

$$z = \frac{p - P_o}{\sqrt{\frac{P_o Q_o}{n}}}$$

is a standardized normal variate, wherein *p* is the proportion of successes in a sample. Hence, to test the null hypothesis, we can use z-test as discussed above. The procedure involved in testing H_o is given below.

Step 1: Formulate the null and alternative hypothesis; for instance

H_o: $P = P_o$ H_1: $P \neq P_o$
H_o: $P \geq P_o$ H_1: $P < P_o$
H_o: $P \leq P_o$ H_1: $P > P_o$

Step 2: Fix the α value and then determine the critical value from the normal distribution table. | $Z_{\alpha/2}$, | = 1.96 and 2.58 for α = 0.05 and 0.01 respectively. | Z_α | = 1.64 and 2.33 for α = 0.05 and 0.01 respectively.

Step 3: Compute p and q (= 1-p) for the sample data.

Step 4: Compute the *z*-statistic, that is, $z = \frac{p - P_o}{\sqrt{\frac{P_o Q_o}{n}}}$, where Q_o=l--P_o.

Step 5: For two-sided test, case (a). That is, reject H_o if $|z| > |z_{\alpha/2}|$.

For one-sided test, case (b): Reject H_o if $z > z_\alpha$ and

For one-sided test, case (c): Reject H_o if $z < - z_\alpha$

8.1 Difference between Two Proportions

A general hypothesis that is tested regarding the theoretical proportion of successes (in two sample cases) is that H_o:P_1- P_2=P_o against a specified alternative hypothesis. The alternative hypothesis may be any one of the following:

$H_1 : P_1 - P_2 \neq P_0$
$H_1 : P_1 - P_2 > P_0$
$H_1 : P_1 - P_2 < P_0$

To test whether or not H_o is true against a specified H_1, we consider the difference P_1 and P_2 and their distribution in repeated samples. It has been shown in the probability theory for repeated independent randomly drawn samples, that P_1 - P_2 follows a normal distribution with mean P_1 - P_2 and variance

$$\frac{P_1(1-P_1)}{n_1}+\frac{P_2(1-P_2)}{n_2}$$

Hence,

$$z=\frac{(p_1-p_2)-(P_1-P_2)}{\left[\frac{P_1(1-P_1)}{n_1}+\frac{P_2(1-P_2)}{n_2}\right]^{1/2}}$$

is a standardized normal variate. If H_o: P_1-P_2=P_o is true, z becomes

$$\frac{(p_1-p_2)-P_o}{\left[\frac{P_1(1-P_1)}{n_1}+\frac{P_2(1-P_2)}{n_2}\right]^{1/2}}$$

However, if P_o*=0,* we are actually testing H_0: P_1=P_2 in which case *z* becomes

$$\frac{(p_1-p_2)}{pq\left[\frac{1}{n_1}+\frac{1}{n_2}\right]^{1/2}}$$

9 TESTS ABOUT CORRELATION COEFFICIENT

The correlation coefficient is defined as:

$$r=\frac{\Sigma x_i y_i-n\overline{xy}}{ns_x s_y}$$

n is the size of the sample, $\bar{x}$ and $\bar{y}$ are the sample means of *X* and *Y* respectively. s_x and s_y are the sample standard deviations of *X* and *Y* respectively. We may like to test the null hypothesis that ρ = 0 (where ρ is the population correlation coefficient) against a specified alternative hypothesis. We are actually using *r* to test the hypothesis about ρ since *r* is an estimate of ρ*.* The testing is usually done by determining the calculated value of *r* as significantly different from zero. This can be done by using a *t*-test. For the purpose of testing the null hypothesis, the following *t*-statistic is computed

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

where the *t*-statistic is with (*n*—2) degrees of freedom. If the calculated value lies in the critical region, we have to reject the null hypothesis. If we accept the null hypothesis it means that there is no correlation between the two variables other than that due to chance.

10 NON-PARAMETRIC TESTS

The use of t-test or z-test requires an assumption that the sample data come from a normal or binomial population (or at least, the sample distribution must tend to a normal distribution for a large sample). Both the z-test and t-test are used to test the null hypothesis concerned with population means, variances and proportions. Hypotheses related to the independence of two criteria of classification, goodness-of-fit test, median of the population, etc. can be tested using the statistical tests called non-parametric tests. The non-parametric tests do not require many assumptions (like the *z*-test and *t*-test). A non-parametric tests is discussed below.

10.1 **Chi-Square Test**

The Chi Square test is normally applicable in situations in which determination of population parameters such as the mean and standard deviation are not an issue. The data in question falling into discrete categories and are presented in a contingency table. The entries in a contingency table are known cells. Let us consider the result of a survey of 100 adults (say, 50 females and 50 males). Let us say that it has been observed that among the 100 adults, 34 of them are library users and the rest are non-users. So, in this hypothetical example, we have two nominal variables sex and library use. On categorizing the 100 adults, using these two nominal variables, we get a contingency table like the one shown in Table below:

A Typical 2 x 2 Table

	User	Non-User l	Tota
Male	10(**17**)	40(**3**3)	50
Female	24(**17**)	26(**33**)	50
Total	34	66	**100**

Let us now try to find out whether or not the two categories—gender (Male or Female) and library use—are independent; let us assume that there is no relationship between the gender and library use. Under this assumption, compute the frequencies in each of the cells. Such frequencies are called *theoretical frequencies.* They are usually referred to as the expected numbers or expected frequencies. The logic for computation of the theoretical frequencies for the data given in the above Table is as follows:

We have 50 each of males and females; the ratio is 1:1. So we would expect that half of the library users are males and also half of the non-users are males, that is, out of 100 adults (*N*) we have 50 males (row total). Out of 34 users (column total), how many of them are males?

Using the cross multiplication technique, we have:

The number of male users = $\frac{50 \times 34}{100} = 17$.

Similarly, we can compute the number of male non-users, female users and female non-users. The results are shown in bold face in the corresponding cells in Table 8.1. On generalizing the above logic, we can easily prove that the following formula can be used to obtain theoretical frequencies *(E_{ij})* in each of the cells:

$$E_{ij} = \frac{r_i . \times c._j}{N}$$

$r_i.$ is the total number of observations in the ith row

c_{ij} is the total number of observations in the jth column

E_{ij} is the expected/theoretical frequencies in the ijth cell (ith row and jth column).

Degree of Freedom

Degrees of freedom are commonly discussed in relation to chi-square and other forms of hypothesis testing statistics. It is important to calculate the degree(s) of freedom when determining the significance of a chi square statistic and the validity of the null hypothesis. It is obvious that the theoretical frequencies need not necessarily be equal to the observed frequencies. If they are equal, one could perhaps conclude that there is no relationship between the two variables. If they are not equal, the question is, "Is the difference between the observed and expected frequencies statistically significant?" To answer this question, we use a statistic called χ^2 (chi-square). It has a parameter called *degrees of freedom*. The values of χ^2 for n degrees of freedom can be obtained from chi-square table. It has been shown in statistics and probability theory that the random variable $\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ has χ^2 distribution with (*n*- 1) degrees of freedom. The O_{ij} and E_{ij} are the observed and theoretical frequencies in the ij^th^ cell; $\sum_i \sum_j O_{ij} = N$. In the case of ***r*** × c contingency table, the degrees of freedom is given by (r — 1) × (c— 1) where r and c are the numbers of rows and columns respectively. ***Thus the χ^2*** is equal to the sum over all cells of the squared differences between the observed and expected frequencies divided by the expected frequencies. We will reject the null hypothesis (in an analysis of 2×2 contingency table) at 0.05 level if the computed value of $\chi^2$is greater than the critical value of the chi-square (that is, 3.841); The critical values can be obtained from the Chi square table. For the data given in the above Table, the χ^2 is given by:

$$\chi^2 = \frac{(17-10)^2}{17} + \frac{(33-40)^2}{33} + \frac{(17-24)^2}{17} + \frac{(33-26)^2}{33}$$

χ^2 = 2.8824 + 1.4848 + 2.8824 + 1.4848

χ^2 = 8.7344

Since χ^2 (8.7344) is greater than the critical value) of the chi-square for one degree of freedom, 3.814, we will reject the null hypothesis that the variables are independent. This implies that there may be reasons to believe that men and women differ in library use.

10.2 Measures of Association

The statistical significance of the null hypothesis depends both on the strength of the observed relationship and the size of the sample. Tests of statistical significance indicate only the likelihood that an observed relationship actually exists in the universe; but they do not reveal the fact as to how strong the relationship is. Further, a relationship may be statistically significant being substantially important.

There are a few measures that will describe the strength of the association between two nominal variables. They are:

1. Contingency coefficient -- $C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$

2. Phi-square measure -- $\Phi^2 = \frac{\chi^2}{n}$

3. Cramer's-*V* measure -- $\mathrm{V} = \frac{\Phi^2}{\sqrt{\min(\mathrm{r}-1),(\mathrm{c}-1)}}$, r and c are the number of rows and columns respectively.

All these measures are the functions of the chi-square. The values of these measures are zero when no relationship between the two variables exists which implies that the variables are independent; the value is one when the variables are perfectly related, which means that they are dependent. The maximum value of the contingency coefficient depends on the size of the contingency table; e.g. for 2x2 table, the maximum value is 0:707; for 3 X3 table, it is 0.816. In general, if the number of columns and the number of rows are equal to each other, the maximum value of *c* is given by

$$\sqrt{\frac{\text{number of columns - 1}}{\text{number of columns}}}.$$

10.3 **Goodness-of-Fit Test**

The chi-square statistic is also used to test the hypothesis that whether or not the probability distribution (of the population) is similar to that of the sample distribution. This type of test is often referred to as a goodness-of-fit test. The goodness-of-fit test is illustrated with following example. Examine whether or not the distribution of transactions follows a negative binomial distribution for the data shown below.

x	f(x)	x	f(x)	x	f(x)
0	324	3	16	6	2
1	108	4	7	7	1
2	43	5	4		

We will use the goodness-of-fit test for this purpose. The procedure is:

Step 1: Formulate the null and alternative hypothesis.

H_o: The sample data belongs to a population which follows a negative binomial distribution.

H_1: The sample data belongs to a population which does not follow a negative binomial distribution.

Step 2: Compute the parameters (such as mean, variance, etc.); estimate the parameters of the theoretical probability distribution (which is assumed in the null hypothesis). Use, as far as possible, the maximum likelihood estimators.

Step 3: Compute the probabilities under the assumption that the H_o is true.

Step 4: Compute the theoretical or expected frequencies (use the formula, that expected frequency is equal to *n·P(x),* where *n* is the sample size, and *P(x)* is the theoretical probability distribution function). In this case, *P(x)* is the mass function of the negative binomial distribution.

Step 5: Decide α and determine the critical region (for α = 0.05). Find out X^2 for *(k* — 1) degrees of freedom; *k* is the number of frequency classes.

Step 6: Compute: $\chi^2 = \sum_{i=1} \frac{(O_i - E_t)^2}{E_t}$, O_i is the observed frequency in the *i*th class; E_i is the expected frequency in the *i*th class, and *k* is the number of frequency classes.

Step 7: Reject H_o if $\chi^2 \geq \chi^2_\alpha$

The result of the goodness-of-fit test is shown in the Table below.

Thus, the χ^2= 1.1472; the χ^2 (α = 0.05, the degrees of freedom is 5) = 11.070. Since $\chi^2 < \chi^2_\alpha$ we accept the null hypothesis that the distribution of transactions in the population follows a negative binomial distribution.

Table 1: A goodness-of-fit test: Chi-square test

X	Observed frequencies	$p(x) = \frac{(x+k-1)!}{(k-1)!\,x!} p^k (1-p)^x$	Expected frequencies		$\frac{(O_i - E_i)^2}{E_i}$
0	324	0.6412	324		0
1	108	0.2135	108		0
2	43	0.0841	43		0
3	16	0.0349	18		0.2222
4	7	0.0148	8		0.125
5	4	0.0064	3	**	0.8
6	2	0.0028	1		
7	1	0.0022*	1		
	505				1.1472

Note: $\bar{x} = 0.6118811, \sigma^2 = 1.1246114,$ p = 0.5441, q=0.4559. k 0.7302052

* Computed such that $\sum_x p(x) = 1$.

** Combined together so that the expected frequency is at least 5 in the last class; after combining these frequencies, the number of classes

(Source: reference number 2)

11. Conclusion

In this Unit, we have discussed the basics of z-test, t-test and chi square test.

REFERENCES

1 Anderson, David R.; Sweepney, Dennis J; and Williams, Thomas. A. (1981) Statistics for Bussiness and Economics. Edition 2. International Edition. West Publishing Company. SanFrancisco.

2 Ravichandra Rao, I.K. (1983) Quantitative Methods for Library and Information Science. Wiley Eastern. New Delhi.

3 Yule, G.H. and Kendall, M.G. (1950) An introduction to theory of statistics. London, Charles Griffin and Company.