LOVELY
PROFESSIONAL
UNIVERSITY

STATISTICS II

Edited By
Richa Nandra

Printed by
EXCEL BOOKS PRIVATE LIMITED
A-45, Naraina, Phase-I,
New Delhi-110028
for
Lovely Professional University
Phagwara

SYLLABUS

Statistics

Objectives:

- To understand the value of Statistics in acquiring knowledge and making decisions in today's society.
- To learn about the basic theory of Probability, random variable, moments generating function, Probability distribution, reliability theory, laws of large numbers, correlation and regression, sampling theory, theory of estimation and testing of hypotheses.

Sr. No.	Content
1	Weak Law of Large Numbers, Strong Law of Large Number, Central Limit Theorem, Confidence Intervals
2	The correlation coefficient, Conditional expectation, Regression of the mean
3	Samples, Sample Statistics, Sampling Distribution of Sample Mean and SampleVariance, t-distribution , Chi Square distribution, F- distribution
4	Estimation of Parameters: Criteria for estimates, Maximum likelihood estimates, Method of least squares
5	t-test, chi square Godness of fit, Z-test with examples

CONTENT

Unit 1: Chebyshev's Inequality

CONTENTS

Objectives

After studying this unit, you will be able to:

- Apply chebyshev's inequality
- Give example of chebyshev's inequality

Introduction

We have discussed different methods for obtaining distribution functions of random variables or random vectors. Even though it is possible to derive these distributions explicity in closed form in some special situations, in general, this is not the case. Computation of the probabilities, even when the probability distribution functions are known, is cumbersome at times. For instance, it is easy to write down the exact probabilities for a binomial distribution with parameters n = 1000 and p = $\frac{1}{50}$. However computing the individual probabilities involve factorials for integers of large order which are impossible to handle even with speed computing facilities.

In this unit, we discuss limit theorems which describe the behaviour of some distributions when the sample size n is large. The limiting distributions can be used for computation of the probabilities approximately.

Chebyshev's inequality is discussed, as an application, weak law of large numbers is derived (which describes the behaviour of the sample mean as n increases).

1.1 Chebyshev's Inequality

We prove in this section an important result known as Chebyshev's inequality. This inequality is due to the nineteenth century Russian mathematician P.L. Chebyshev.

We shall begin with a theorem.

Notes

Theorem 1: Suppose X is a random variable with mean μ and finite variance σ^2. Then for every $\varepsilon > 0$.

Proof: We shall prove the theorem for continuous r.vs. The proof in the discrete case is very similar.

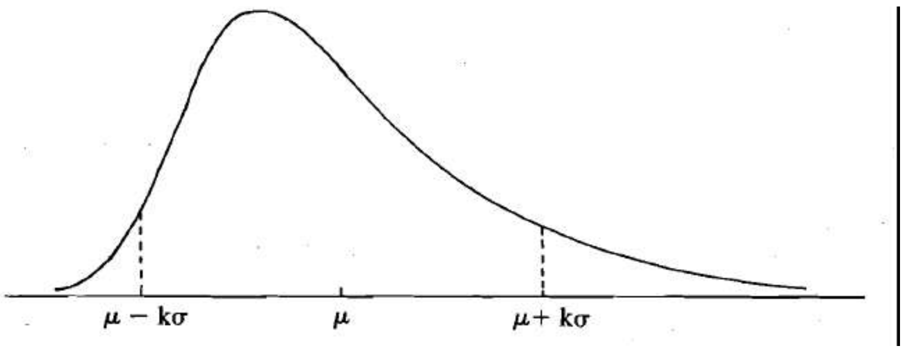
Suppose X is a random variable with probability density function f. From the definition of the variance of X, we have

$$\sigma = E[(x-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx.$$

Suppose $\varepsilon > 0$ is given. Put $\varepsilon_1 = \frac{\varepsilon}{\sigma}$. Now we divide the integral into three parts as shown in Fig. 1.

$$\sigma^2 = \int_{-\infty}^{\mu-\varepsilon_1\sigma} (x-\mu)^2 f(x)dx + \int_{\mu-\varepsilon_1\sigma}^{\mu-\varepsilon_1\sigma} (x-\mu)^2 f(x)dx + \int_{\mu+\varepsilon_1\sigma}^{\infty} (x-\mu)^2 f(x)dx \quad ...(2)$$

Figure 1.1



Since the integrand (x – m)2 f(x) is non-negative, from (2) we get the inequality

$$\sigma^2 \geq \int_{-\infty}^{\mu-\varepsilon_1\sigma} (x-\mu)^2 f(x)dx + \int_{\mu+\varepsilon_1\sigma}^{\infty} (x-\mu)^2 f(x)dx \quad ...(3)$$

Now for any $x \in]-\infty, \mu - \varepsilon_1\sigma]$, we have $x \leq \mu - \varepsilon_1\sigma$ which implies that $(x - \mu)^2 \geq \varepsilon^2\sigma^2$. Therefore we get

$$\int_{-\infty}^{\mu-\varepsilon_1\sigma} (x-\mu)^2 f(x)dx \geq \int_{-\infty}^{\mu-\varepsilon_1\sigma} \varepsilon^2\sigma^2 f(x)dx$$

$$= \varepsilon^2\sigma^2 \int_{-\infty}^{\mu-\varepsilon_1\sigma} f(x)dx.$$

Similarly for $x \in]\mu + \varepsilon_1\sigma, \infty[$ also we have $(x - \mu)^2 \geq \varepsilon_1^2\sigma^2$ and therefore

$$\int_{\mu+\varepsilon_1\sigma}^{\infty} (x-\mu)^2 f(x)dx \geq \varepsilon_1^2\sigma^2 \int_{\mu+\varepsilon_1\sigma}^{\infty} f(x)dx$$

Then by (3) we get

$$\sigma^2 \ge \varepsilon_1^2\sigma^2\left[\int_{-\infty}^{\mu-\varepsilon_1\sigma} f(x)dx + \int_{\mu+\varepsilon_1\sigma}^{\infty} f(x)dx\right]$$

i.e.,
$$\frac{1}{\varepsilon_1^2} \ge \int_{-\infty}^{\mu-\varepsilon_1\sigma} f(x)dx + \int_{\mu+\varepsilon_1\sigma}^{\infty} f(x)dx$$

whenever $\sigma^2 \neq 0$.

Now, by applying Property (iii) of the density function given in Sec. 11.3, unit 10, we get

$\frac{1}{\varepsilon_1^2} \ge P[X \le \mu - \varepsilon_1\sigma] + P[X \ge \mu + \varepsilon\sigma]$

$= P[X - \mu \le - \varepsilon_1\sigma] + P[X - \mu \ge \varepsilon 1\sigma]$

$= P[\,|X - m|\, \ge \varepsilon_1\sigma]$

That is, $P[\,|X - m|\, \ge \varepsilon_1\sigma] \le \frac{1}{\varepsilon_1^2}$ ...(4)

Substituting $\varepsilon_1 = \frac{\varepsilon}{\sigma}$ in (4), we gt the inequality

$$\left[P[\,|X - \mu| \ge \varepsilon\right] \le \frac{\sigma^2}{\varepsilon^2}$$

Chebyshev's inequality also holds when the distribution of X is neither (absolutely) continuous nor discrete. We will not discuss this general case here. Now we shall make a remark.

Remark 1: The above result is very general indeed. We need to know nothing about the probability distribution of the random variable X. It could be binomial, normal, beta or gamma or any other distribution. The only restriction is that it should have finite variance. In other words the upper bound is universal in nature. The price we pay for such generality is that the upper bound is not sharp in general. If we know more about the distribution of X, then it might be possible to get a better bound. We shall illustrate this point in the following example.

Example 1: Suppose X is $N(\mu, \sigma^2)$. Then $E(X) = \mu$ and $Var(X) = \sigma^2$. Let us compute $P[\,|X - \mu| \ge 2\sigma]$.

Here $\varepsilon = 2\sigma$. By applying Chebychev's inequality we get

$$P\left[|X - \mu| \ge 2\sigma\right] \le \frac{\sigma^2}{4\sigma^2} = \frac{1}{4} = .25$$

Since we know that the distribution of X is normal, we can directly compute the probability. Then we have

$$P\left(|X - \mu| \ge 2\sigma\right) = P\left[\left|\frac{X-\mu}{\sigma}\right| \ge 2\right]$$

Since $\frac{X-\mu}{\sigma}$ has N(0, 1) as its distribution, from the normal distribution table given in the appendix of Unit 11, we get

$$P\left(\left|\frac{X-\mu}{\sigma}\right| \geq 2\right) = 0.456$$

which is substantially small as compared to the exact value 0.25. Thus in this case we could get a better upperbound by directly using the distribution.

Let us consider another example.

Example 2: Suppose X is a random variable such that P[X = 1] = 1/2 = P[X = –1]. Let us compute an upper bound for $P[\,|X - \mu| > \sigma]$.

You can check that E(X) = 0 and Var(X) = 1. Hence, by Chebyshev's inequality, we get that

$$P\left(|X-\mu| > \sigma\right) \leq \frac{\sigma^2}{\sigma^2} = 1.$$

on the other hand, direct calculations show that

$$P\left(|X-\mu| > \sigma\right) = P\left[|X| \geq 1\right] = 1.$$

In this example, the upper bound obtained from Chebyshev's inequality as well as the one obtained from using the distribution of X are one and the same.

In the first example you can see an application of Chebyshev's inequality.

Example 3: Suppose a person makes 100 check transactions during a certain period. In balancing his or her check book transactions, suppose he or she rounds off the check entries to the nearest rupee instead of subtracting the exact amount he or she has used. Let us find an upper bound to the probability that the total error he or she has committed exceeds Rs. 5 after 100 transactions.

Let X_i denote the round off error in rupees made for the ith transaction. Then the total error is $X_1 + X_2 + ... + X_{100}$. We can assume that X_i, $1 \leq i \leq 100$ are independent and idelltically distributed random variables and that each X_i has uniform distribution on $\left[-\frac{1}{2}, \frac{1}{2}\right]$. We are interested in finding an upper bound for the $P\left[|S_{100}| > 5\right]$ where $S_{100} = X_1 + ... + X_{100}$.

In general, it is difficult and computationally complex to find the exact distribution. However, we can use Chebyshev's inequality to get an upper hound. It is clear that

$$E(S_{100}) = 100E(X_1) = 0$$

and

$$\text{var}(S_{100}) = 100 \text{ var } (X_i) = \frac{100}{12}.$$

since $E(X_1) = 0$ and $\text{Var}(X_1) = \frac{1}{12}$. Therefore by Chebyshev's inequality,

$$P\left(|S_{100}-0)|>5\right)\leq\frac{\mathrm{Var}(S_{100})}{25}$$

$$=\frac{100}{12\times 25}$$

$$=\frac{1}{3}.$$

Here are some exercises for you.

The above examples and exercises must have given you enough practise to apply Chebyshev's inequality. Now we shall use this inequality to establish an important result.

Suppose X_1, X_2, ..., X_n are independent and identically distributed random variables having mean μ and variance σ^2. We define

$$\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$$

Then $\overline{X}_n$ has mean μ and variance $\frac{\sigma^2}{n}$. Hence, by the Chebyshev's inequality, we get

$$P\left[|\overline{X}_n-\mu|\geq\varepsilon\right]\leq\frac{\sigma^2}{n\varepsilon^2}$$

for any ε > 0. If n → 0, then $\frac{\sigma^2}{n\varepsilon^2}\to 0$ and therefore

$$P\left(|\overline{X}_n-\mu|\geq\varepsilon\right)\to 0.$$

In other words, as n grows large, the probability that $\overline{X}_n$ differs from μ by more than any given positive number E, becomes small. An alternate way of stating this result is as follows :

For any ε > 0, given any positive number δ, we a n choose sufficiently large n such that

$$P\left(|\overline{X}_n-\mu|\geq\varepsilon\right)\leq\delta$$

This result is known as the weak law of large numbers. We now state it as a theorem.

Theorem 2 (Weak law of large nombers) : Suppose X_1, X_2, ..., X_n are i.i.d. random variables with mean μ and finite variance σ^2.

Let

$$\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i.$$

Then

$$P\left[|\overline{X}_n-\mu|\geq\varepsilon\right]\to 0 \text{ as } n\to\infty.$$

for any E > 0.

The above theorem is true even when the variance is infinite but the mean p is finite However this result does not follow as an application of the Chebyshev's inequality in this general set up. The proof in the general case is beyond the scope of this course.

We make a remark here.

Remark 2 : The above theorem only says that the probability that the value of the difference $|\overline{X}_n - X|$ exceeds any fixed number ε, gets smaller and smaller for successively large values of n. The theorem does not say anything about the limiting case of the actual difference. In fact there is another strong result which talks about the limiting case of the actual values of the differences. This is the reason why Theorem 2 is called 'weak law'. We hove not included the stronger result here since it is beyond the level of this course.

Let us see an example.

Example 4: Suppose a random experiment has two possihle outcomes called success (S) and Failure (F). Let p he the probability of a success. Suppose the experiment is repeated independently n times. Let X_i take the value 1 or 0 according as the outcome in the i-th trial of the experiments is success or a failure. Let us apply Theorem 2 to the set $\{X_i\}_{j=1}^{n}$.

We first note that

$$P[X_i = 1] = p \text{ and } P[X_i = 0] = 1 - p = q,$$

for $1 \leq i \leq n$. Also you can check that $E(X_i) = p$ ond var $(X_i) = p\,q$ for i = 1, ... n.

Since the mean and the variance are finite, we can apply the weak law of large numbers for the sequence $\{X_1 : 1 \leq i \leq n\}$. Then we have

$$P\left[\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right] \rightarrow 0 \text{ as } b \rightarrow \infty$$

Sn for every $\varepsilon > 0$ where $S_n = X_1 + X_2 + ... + X_n$. Now, what is $\frac{S_n}{n}$? S_n is the number of successes observed in n trials and therefore $\frac{S_n}{n}$ is the proportinn of successes in n trials. Then the above result says that as the number of trials increases, the proportion of successes tends stabilize to the probability of a success. Of course, one of the basic assumptions behind this interpretation is that the random experiment can be repeated.

In the next section we shall discuss another limit theorem which gives an approximation to the binomial distrihution.

1.2 Summary

- Suppose X is a random variable with mean μ and finite variance σ^2. Then for every $\varepsilon > 0$.
- The above theorem only says that the probability that the value of the difference $|\overline{X}_n - X|$ exceeds any fixed number ε, gets smaller and smaller for successively large values of n. The theorem does not say anything about the limiting case of the actual difference. In fact there is another strong result which talks about the limiting case of the actual values of the differences. This is the reason why Theorem 2 is called 'weak law'. We hove not included the stronger result here since it is beyond the level of this course.

- Since the mean and the variance are finite, we can apply the weak law of large numbers for the sequence $\{X_1 : 1 \leq i \leq n\}$. Then we have

$$P\left[\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right] \to 0 \text{ as } b \to \infty$$

S_n for every $\varepsilon > 0$ where $S_n = X_1 + X_2 + ... + X_n$. Now, what is $\frac{S_n}{n}$? S_n is the number of successes observed in n trials and therefore $\frac{S_n}{n}$ is the proportinn of successes in n trials. Then the above result says that as the number of trials increases, the proportion of successes tends stabilize to the probability of a success. Of course, one of the basic assumptions behind this interpretation is that the random experiment can be repeated.

1.3 Keywords

Chebyshev's inequality is discussed, as an application, weak law of large numbers is derived.

Weak law of large nombers: Suppose $X_1, X_2, ..., X_n$ are i.i.d. random variables with mean m and finite variance σ^2.

1.4 Self Assessment

1. Computation of the probabilities, even when the .... functions are known, is cumbersome at times.

   (a) Chebyshev's inequality (b) limiting distributions

   (c) (absolutely) continuous (d) probability distribution

2. The .... can be used for computation of the probabilities approximately.

   (a) Chebyshev's inequality (b) limiting distributions

   (c) (absolutely) continuous (d) probability distribution

3. .... is discussed, as an application, weak law of large numbers is derived.

   (a) Chebyshev's inequality (b) limiting distributions

   (c) (absolutely) continuous (d) probability distribution

4. Chebyshev's inequality also holds when the distribution of X is neither .... nor discrete.

   (a) Chebyshev's inequality (b) limiting distributions

   (c) (absolutely) continuous (d) probability distribution

1..5 Review Questions

1. Suppose X is $N(\mu, \sigma^2)$. Then $E(X) = \mu$ and $Var(X) = \sigma^2$. Let us compute $P[\ |X - \mu| \geq 3\sigma]$.

2. Suppose X is a random variable such that $P[X = 1] = 1/2 = P[X = -1]$. Let us compute an upper bound for $P[|X - \mu| > 1/2\sigma]$.

3. Suppose a person makes 100 check transactions during a certain period. In balancing his or her check book transactions, suppose he or she rounds off the check entries to the nearest

Notes

rupee instead of subtracting the exact amount he or she has used. Let us find an upper bound to the probability that the total error he or she has committed exceeds Rs. 5 after 100 transactions.

Answers: Self Assessment

1. (d) 2. (b) 3. (a) 4. (c)

1.6 Further Readings

Books

Introductory Probability and Statistical Applications by P.L. Meyer

Introduction to Mathematical Statistics by Hogg and Craig

Fundamentals of Mathematical Statistics by S.C. Gupta and V.K. Kapoor

Unit 2: The Weak Law

CONTENTS

Objectives

After studying this unit, you will be able to:

- Discuss the weak laws
- Describe some examples related to weak law

Introduction

James Bernoulli proved the weak law of large numbers (WLLN)around 1700 which was published posthumously in 1713 in his treatise Ars Conjectandi. Poisson generalized Bernoulli's theorem around 1800, and in 1866 Tchebychev discovered the method bearinghis name. Later on one of his students, Markov observed that Tchebychev's reasoning can be used to extend Bernoulli's theoremto dependent random variables as well.

In 1909 the French mathematician Emile Borel proved adeeper theorem known as the strong law of large numbers that furthergeneralizes Bernoulli's theorem. In 1926 Kolmogorov derived conditions that were necessary and sufficient for a set of mutually independent random variables to obey the law of large numbers.

2.1 Weak Law of Number

Let X_i be independent, identically distributed Bernoulli randomVariables such that

$$P(X_i) = p, \qquad P(X_i = 0) = 1 - p = q,$$

and let $k = X_1 + X_2 + ... + X_n$ represent the number of "successes"in n trials. Then the weak law due to Bernoulli states that [see Theorem 3-1, page 58, Text]

$$P\left\{\left|\frac{k}{h} - p\right| > \varepsilon\right\} \le \frac{pq}{n\varepsilon^2} \qquad ...(18.1)$$

i.e., the ratio "total number of successes to the total numberof trials" tends to p in probability as nincreases.

Notes

A stronger version of this result due to Borel and Cantellistates that the above ratio k/n tends to p not only in probability, but with probability 1. This is the strong law of large numbers (SLLN).

What is the difference between the weak law and the strong law? The strong law of large numbers states that if $\{\varepsilon_n\}$ is a sequence of positive numbers converging to zero, then

$$\sum_{n=1}^{\infty} P\left\{\left|\frac{k}{h} - p\right| \geq \varepsilon_n\right\} < \infty \qquad ...(18.2)$$

From Borel-Cantelli lemma [see (2-69) Text], when (13-2) is satisfied the events $An \overset{\Delta}{=} \left\{\left|\frac{k}{h} - p\right| \geq \varepsilon_n\right\}$ can occur only for a finitenumber of indices n in an infinite sequence, or equivalently, the events $\left\{\left|\frac{k}{h} - p\right| \geq \varepsilon_n\right\}$ occur infinitely often, i.e., the event k/nconverges to palmost-surely.

Proof: To prove (18.2), we proceed as follows. Since

$$\left|\frac{k}{h} - p\right| \geq \varepsilon \Rightarrow |k - np|^4 \geq \varepsilon^4 n^4$$

we have

$$\sum_{k=0}^{n}(k - np)^4 p_n(k) \geq \varepsilon^4 n^4 = \varepsilon^4 n^4 \left(P\left\{\left|\frac{k}{n} - p\right| \geq \varepsilon\right\} + P\left\{\left|\frac{k}{n} - p\right| < \varepsilon\right\}\right)$$

and hence

$$P\left\{\left|\frac{k}{n} - p\right| \geq \varepsilon\right\} \leq \frac{\sum_{k=0}^{n}(k - np)^4 p_n(k)}{\varepsilon^4 n^4} \qquad ...(13.3)$$

where

$$p_n(k) = P\left\{\sum_{i=1}^{n} X_i = k\right\} = \binom{n}{k} p^k q^{n-k}$$

By direct computation

$$\sum_{k=0}^{n}(k - np)4pn(k) = E\left\{\left(\sum_{i=1}^{n} X_i - np\right)^4\right\} = E\left\{\left(\sum_{i=1}^{n} X_i - p\right)^4\right\}$$

$$= E\left\{\left(\sum_{i=1}^{n} Y_i\right)^4\right\} = \sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n} E(Y_i Y_k Y_j Y_l)$$

$$= \sum_{i=1}^{n} E(Y_i^4) + 4n(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} E(Y_i^3)E(X_j) + 3n(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} E(Y_i^2)E(Y_j^2)$$

$$= n(p^3 + q^3)pq + 3n(n - 1)(pq)^2 \leq [n + 3n(n - 1)]pq$$

$$= 3n^2 pq, \qquad ...(18.4)$$

since

$$p^3 + q^3 = (p + q)^3 - 3p^2q - 3pq^2 < 1, pq \leq 1 / 2 < 1$$

Substituting (18.4) also (18.3) we obtain

$$P\left\{\left|\frac{k}{n} - p\right| \geq \varepsilon\right\} \leq \frac{3pq}{n^2\varepsilon^4}$$

Let $\varepsilon = \frac{1}{n^{1/8}}$ so that the above integral reads and hence

$$\sum_{n=1}^{\infty} P\left\{\left|\frac{k}{n} - p\right| \geq \frac{1}{n^{1/8}}\right\} \leq 3pq\sum_{n=1}^{\infty}\frac{1}{n^{3/2}} \leq 3pq(1 + \int_1^{\infty} x^{-3/2}dx)$$

$$= 3pq(1 + 2) = 9pq < \infty, \qquad ...(18.5)$$

thus proving the strong law by exhibiting a sequence of positive numbers $\varepsilon_n = 1/n^{1/8}$ that converges to zero and satisfies (13-2).

We return back to the same question: "What is the difference between the weak law and the strong law?.

"The weak law states that for every n that is large enough, the ratio $\left(\sum_{i=1}^{n} X_i\right)/n = k/n$ is likely to be near p with certain probability that tends to 1 as nincreases. However, it does not say that k/n is boundto stay near p if the number of trials is increased. Suppose (18.1) is satisfied for a given ε in a certain number of trials n_0. If additional trials are conducted beyond n_0, the weak law does not guarantee that the new k/n is bound to stay near p for such trials. In fact there can be events for which k/n > p + e, for $n > n_0$ in some regular manner. The probability for such an event is the sum of a large number of very small probabilities, and the weak law is unable to say anything specific about the convergence of that sum.

However, the strong law states (through (18.2)) that not onlyall such sums converge, but the total number of all such events where k/n > p + ε is in fact finite! This implies that the probability $\left\{\left|\frac{k}{n} - p\right| > \varepsilon\right\}$ of the events as n increases becomes and remains small, since with probability 1 only finitely many violations to the above inequality takes place as $n \to \infty$.

Interestingly, if it possible to arrive at the same conclusion using a powerful bound known as Bernstein's inequality that is based on the WLLN.

Bernstein's inequality : Note that

$$\left|\frac{k}{n} - p\right| > \varepsilon \Rightarrow k > n(p + \varepsilon)$$

and for any λ > 0, this gives $e^{\lambda(k - n(p+\varepsilon))} > 1$.

Thus

$$P\left\{\frac{k}{n} - p > \varepsilon\right\} = \sum_{k=[n(p+\varepsilon)]}^{n} \binom{n}{k} p^k q^{n-k}$$

$$\leq \sum_{k=[n(p+\varepsilon)]}^{n} e^{\lambda(k-n(p+\varepsilon))} \binom{n}{k} p^k q^{n-k}$$

$$\leq \sum_{k=0}^{n} e^{\lambda(k-n(p+\varepsilon))}\binom{n}{k}p^{k}q^{n-k}$$

$$P\left\{\frac{k}{n}-p>\varepsilon\right\}=e^{-\lambda n\varepsilon}\sum_{k=0}^{n}\binom{n}{k}(pe^{\lambda q})^{k}(qe^{-\lambda p})^{n-k}$$

$$= e^{-\lambda n\varepsilon}\left(pe^{\lambda q}+qe^{-\lambda p}\right)^{n} \qquad ...(18.6)$$

Since $e^x \leq x + e^{x^2}$ for any real x,

$pe^{\lambda q} + qe^{-\lambda p} \leq p(\lambda q + e^{\lambda^2 q^2}) + q(-\lambda p + e^{\lambda^2 p^2})$

$$= pe^{\lambda^2 q^2} + qe^{\lambda^2 p^2} \leq e^{\lambda^2}. \qquad ...(18.7)$$

Substituting (18.7) into (18.6), we get

$$P\left\{\frac{k}{n}-p>\varepsilon\right\}\leq e^{\lambda^2 n-\lambda n\varepsilon}.$$

But $\lambda^2 n - \lambda n\varepsilon$ is minimum for $\lambda = \varepsilon/2$ and hence

$$P\left\{\frac{k}{n}-p>\varepsilon\right\}\leq e^{-n\varepsilon^2/4},\ \varepsilon>0. \qquad ...(18.8)$$

Similarly

$$P\left\{\frac{k}{n}-p<-\varepsilon\right\}£\ e^{-n\varepsilon 2/4}$$

and hence we obtain Bernstein's inequality

$$P\left\{\left|\frac{k}{n}-p\right|>\varepsilon\right\}£\ 2e^{-n\varepsilon^2/4}. \qquad ...(18.9)$$

Bernstein's inequality is more powerful than Tchebyshev's inequalityas it states that the chances for the relative frequency k /n exceeding its probability p tends to zero exponentially fast as n → ∞.

Chebyshev's inequality gives the probability of k /nto lie between and for a specific n. We can use Bernstein's inequality to estimate the probability for k /nto lie between and for all large n

Towards this, let

$$y_n=\left\{p-\varepsilon\leq\frac{k}{n}<p+\varepsilon\right\}$$

so that

$$P(y_n^c)=P\left\{\left|\frac{n}{k}-p\right|>\varepsilon\right\}\leq 2e^{-n\varepsilon 2/4}$$

To compute the probability of the event $\bigcap_{n=m}^{\infty} y_n$, note that its complement is given by

$$\left(\bigcap_{n=m}^{\infty} y_n\right)^c = \bigcup_{n=m}^{\infty} y_n^c$$

and using Eq. (2-68) Text,

$$P\left(\bigcup_{n=m}^{\infty} y_n^c\right) \leq \sum_{n=m}^{\infty} P(y_n^c) \leq \sum_{n=m}^{\infty} 2e^{-n\varepsilon^2/4} = \frac{2e^{-m\varepsilon^2/4}}{1-e^{-\varepsilon^2/4}}.$$

This gives

$$P\left(\bigcup_{n=m}^{\infty} y_n\right) = \left\{1 - P\left(\bigcup_{n=m}^{\infty} \overline{y}_n\right)\right\} \geq 1 - \frac{2e^{-m\varepsilon^2/4}}{1-e^{-\varepsilon^2/4}} \to 1 \text{ as } m \to \infty$$

or,

$$P\left\{p - \varepsilon \leq \frac{k}{n} \leq p + \varepsilon, \text{ for all } n \geq m\right\} \to 1 \text{ as } m \to \infty.$$

Thus k/n is bound to stay near p for all large enough n, in probability,a conclusion already reached by the SLLN.

Discussion: Let Thus if we toss a fair coin 1,000 times, from the weak law

$$P\left\{\left|\frac{k}{n} - \frac{1}{2}\right| \geq 0.01\right\} \leq \frac{1}{40}.$$

Thus on the average 39 out of 40 such events each with 1000 or more trials will satisfy the inequality $\left\{\left|\frac{k}{n} - \frac{1}{2}\right| \leq 0.1\right\}$ or, it is quite possible that one out of 40 such events may not satisfy it.

As a result if we continue the coin tossing experiment for an additional 1000 moretrials, with k representing the total number of successes up to the current trial n, for n = 1000 → 2000, it is quite possible that for few such n the above inequality may be violated. This is still consistent with the weak law, but "not so often" says the strong law. According to the strong law such violations can occur only a finite number of times each with a finite probability in an infinite sequence of trials, and hence almost always the above inequality will be satisfied, i.e., the sample space of k/n coincides with that of p as n → ∞.

Next we look at an experiment to confirm the strong law:

Example: 2n red cards and 2n black cards (all distinct) are shuffled together to form a single deck, and then split into half. What is the probability that each half will contain n red and n black cards?

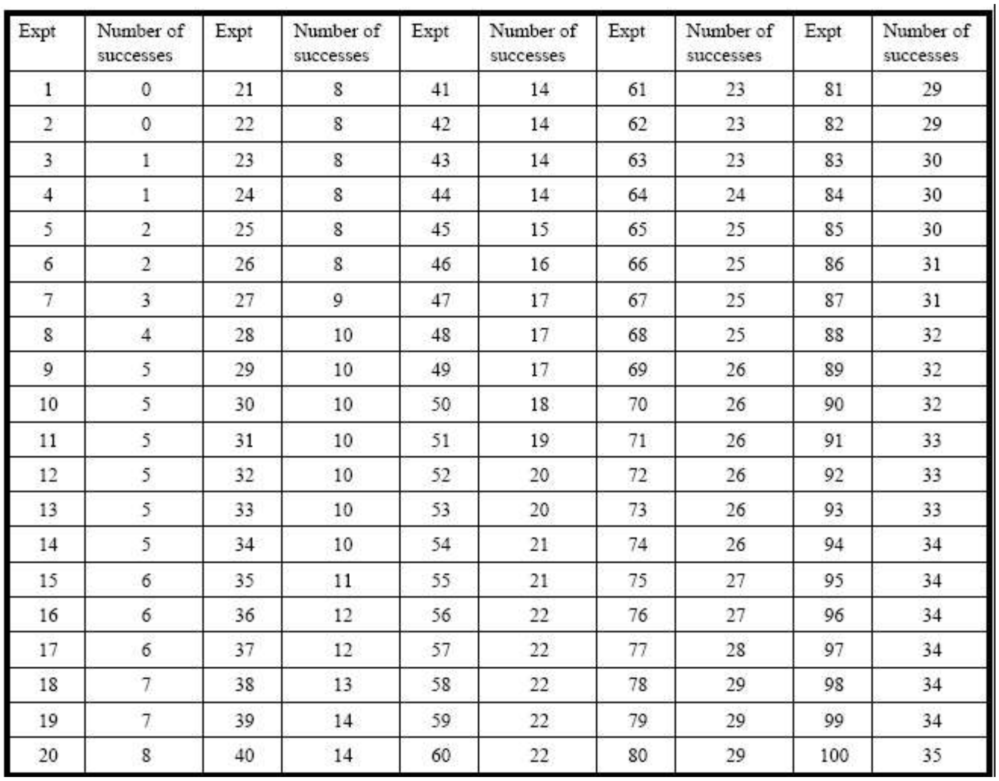
Notes

Solution: From a deck of 4n cards, 2n cards can be chosen $\binom{4n}{2n}$ in different ways. To determine the number of favorable draws of n red and n black cards in each half, consider the unique draw consisting of 2n red cards and 2n black cards in each half. Among those 2n red cards, n of them can be chosen in $\binom{2n}{n}$ different ways; similarly for each such draw there are $\binom{2n}{n}$ ways of choosing n black cards. Thus the total number of favorable draws containing n red and n black cards in each half are $\binom{2n}{n}\binom{2n}{n}$ among a total of $\binom{4n}{2n}$ draws. This gives the desired probability p_n to be

$$p_n \simeq \frac{\binom{2n}{n}\binom{2n}{n}}{\binom{4n}{2n}} = \frac{(2n!)^4}{(4n)!(n!)^4}.$$

For large n, using Stingling's formula we get

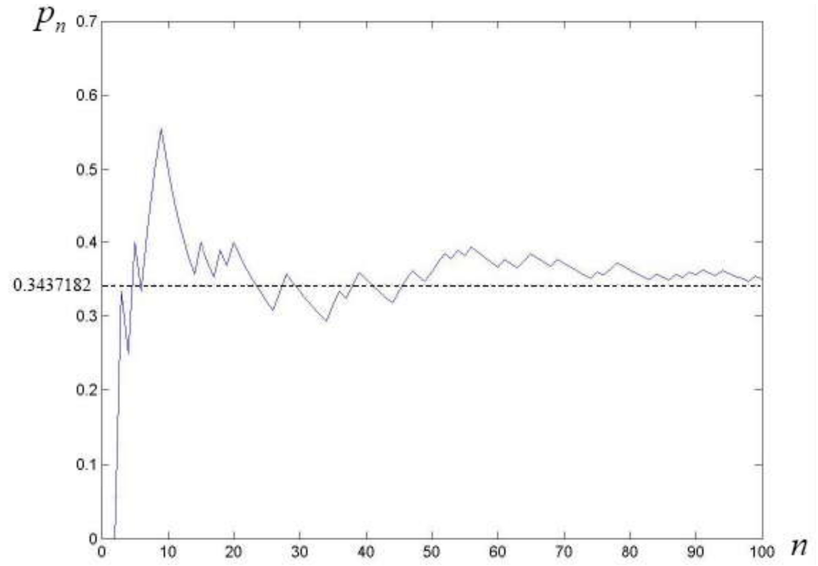
Table 2.1



Expt	Number of successes	Expt	Number of successes	Expt	Number of successes	Expt	Number of successes	Expt	Number of successes
1	0	21	8	41	14	61	23	81	29
2	0	22	8	42	14	62	23	82	29
3	1	23	8	43	14	63	23	83	30
4	1	24	8	44	14	64	24	84	30
5	2	25	8	45	15	65	25	85	30
6	2	26	8	46	16	66	25	86	31
7	3	27	9	47	17	67	25	87	31
8	4	28	10	48	17	68	25	88	32
9	5	29	10	49	17	69	26	89	32
10	5	30	10	50	18	70	26	90	32
11	5	31	10	51	19	71	26	91	33
12	5	32	10	52	20	72	26	92	33
13	5	33	10	53	20	73	26	93	33
14	5	34	10	54	21	74	26	94	34
15	6	35	11	55	21	75	27	95	34
16	6	36	12	56	22	76	27	96	34
17	6	37	12	57	22	77	28	97	34
18	7	38	13	58	22	78	29	98	34
19	7	39	14	59	22	79	29	99	34
20	8	40	14	60	22	80	29	100	35

The figure below shows results of an experiment of 100 trials.

Figure 2.1



2.1 Summary

Let X_i be independent, identically distributed Bernoulli randomVariables such that

$$P(X_i) = p, \qquad P(X_i = 0) = 1 - p = q,$$

and let $k = X_1 + X_2 + ... + X_n$ represent the number of “successes”in n trials. Then the weak law due to Bernoulli states that [see Theorem 3-1, page 58, Text]

$$P\left\{\left|\frac{k}{h} - p\right| > \varepsilon\right\} \le \frac{pq}{n\varepsilon^2} \qquad ...(18.1)$$

i.e., the ratio “total number of successes to the total numberof trials” tends to p in probability as nincreases.

A stronger version of this result due to Borel and Cantellistates that the above ratio k/n tends to p not only in probability, but with probability 1. This is the strong law of large numbers (SLLN).

2.2 Keywords

Strong law of large numbers: A stronger version of this result due to Borel and Cantellistates that the above ratio k/n tends to p not only in probability, but with probability 1. This is the strong law of large numbers (SLLN).

Bernstein’s inequality is more powerful than Tchebyshev’s inequalityas it states that the chances for the relative frequency k /n exceeding its probability p tends to zero exponentially fast as $n \to \infty$.

2.3 Self Assessment

1. .... generalized Bernoulli's theorem around 1800, and in 1866 Tchebychev discovered the method bearinghis name.

2. In .... the French mathematician Emile Borel proved adeeper theorem known as the strong law of large numbers that further generalizes Bernoulli's theorem.

3. In .... Kolmogorov derived conditions that were necessary and sufficient for a set of mutually independent random variables to obey the law of large numbers.

4. A .... of this result due to Borel and Cantellistates that the above ratio k/n tends to p not only in probability, but with probability 1. This is the strong law of large numbers (SLLN).

5. The strong law of large numbers states that if $\{e_n\}$ is a sequence of .... to zero, then

$$\sum_{n=1}^{\infty} P\left\{\left|\frac{k}{h}-p\right| \geq \varepsilon_n\right\} < \infty$$

2.4 Review Questions

1. 2n red cards and 2n black cards (all distinct) are shuffled together to form a single deck, and then split into half. What is the probability that each half will contain n red and n black cards?

2. 3n red cards and n black cards (all distinct) are shuffled together to form a single deck, and then split into half. What is the probability that each half will contain n red and n black cards?

3. 4n red cards and 4n black cards (all distinct) are shuffled together to form a single deck, and then split into half. What is the probability that each half will contain n red and n black cards?

4. n red cards and 2n black cards (all distinct) are shuffled together to form a single deck, and then split into half. What is the probability that each half will contain n red and n black cards?

Answers: Self Assessment

1. Poisson 2. 1909 3. 1926 4. stronger version

5. positive numbers converging

2.5 Further Readings

Books

Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.

Unit 3: The Laws of Large Numbers Compared

CONTENTS

Objectives

After studying this unit, you will be able to:

- Discuss the strong law of large number
- Discuss examples related to large number

Introduction

Probability Theory includes various theorems known as Laws of Large Numbers; for instance, see [Fel68, Hea71, Ros89]. Usually two major categories are distinguished: Weak Laws versus Strong Laws. Within these categories there are numerous subtle variants of differing generally. Also the Central Limit Theorems are often brought up in this context.

Many introductory probability texts treat this topic superficially, and more than once their vague formulations are misleading or plainly wrong. In this note, we consider a special case to clarify the relationship between the Weak and Strong Laws. The reason for doing so is that I have not been able to find a concise formal exposition all in one place. The material presented here is certainly not new and was gleaned from many sources.

In the following sections, X1, X2, ... is a sequence of independent and indentically distributed random variabels with finite expectation μ. We define the associated sequence $\overline{X}_i$ of partial sample means by

$$\overline{X}_i = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The Laws of Large Numbers make statements about the convergence of $\overline{X}_n$ to m. Both laws relate bounds on sample size, accuracy of approximation, and degree of confidence. The Weak Laws deal with limits of probabilities involving $\overline{X}_n$. The Strong Laws deal with probabilities involving limits of $\overline{X}_n$. Especially the mathematical underpinning of the Strong Laws requires a caretful approach ([Hea71, Ch. 5] is an accesible presentation).

3.1 Strong Law of Large Numbers

Notes

We are now ready to give Etemadi's proof of

(7.1) Strong law of large numbers. Let X1, X2, ... be pairwise independent identically distributed random variables with E | X_i | < ∞. Let $EX_i = \mu$ and $S_n = X_1 + ... + X_n$. Then $S_n/n \to \mu$ a.s. as $n \to \infty$.

Proof : As in the proof of weak law of large numbers, we begin by truncating.

(a) **Lemma.** Let $Y_k = K_k 1_{(|X_k| \le k)}$ and $T_n = Y_1 + ... + Y_n$. It is sufficient to prove that $T_n/n \to \mu$ a.s.

Proof $\sum_{k=1}^{\infty} P(|X_k| > k) \le \int_0^{\infty} P(|X_1| > t)dt = E|X_1| < \infty$ so $P(X_k \neq Yk$ i.o.) = 0. This shows that |Sn(w) – Tn(w) < ∞ a.s. for all n, from which the desired result follows.

The second step is not so intuitive but it is an important part of this proof and the one given in Section 1.8.

(b) **Lemma.** $\sum_{k=1}^{\infty} \text{var}(Y_k)/k^2 \le 4E|X_1| < \infty$.

Proof To bound the sum, we observe

$$\text{var}(Y_k) \le E(Y_k^2) = \int_0^{\infty} 2yP(|Y_k| > y)dy \le \int_0^{k} 2yP(|X_1| > y)dy$$

so using Fubini's theorem (since everything is ≤ 0 and the sum is just an integral with respect to counting measure on {1, 2, ...})

$$\sum_{k=1}^{\infty} E(Y_k^2)/k^2 \le \sum_{k=1}^{\infty} k^{-2} \int_0^{\infty} 1_{(y<k)} 2yP(|X_1| > y)dy$$

$$= \int_0^{\infty} \left\{ \sum_{k=1}^{\infty} k^{-2} 1_{(y<k)} \right\} 2yP(|X_1| > y)dy$$

Since $E|X_1| = \int_0^{\infty} P(|X_1| > y)dy$, we can complete the proof by showing

(c) Lemma. If y ≥ 0 then $2y \sum_{k>y} k^{-2} \ge 4$.

Proof We being with the observation that if m ≥ 2 then

$$\sum_{k \ge m} k^{-2} \le \int_{m-1}^{\infty} x^{-2} dx = (m-1)^{-1}$$

When y ≥ 1 the sum starts with k = [y] +1 ≥ 2 so

$$2y \sum_{k \ge m} k^{-2} \le 2y/[y] \le 4$$

since y/[y] ≤ 2 for y ≥ 1 (the worst case being y close to 2). To cover 0 ≤ y < 1 we note that in this case

$$2y \sum_{k \ge y} k^{-2} \le 2y \left(1 + \sum_{k=2}^{\infty} k^{-2} \right) \le 4$$

The first two steps, (a) and (b) above, are standard. Etemadi's inspiration was that since $X_n^+, n \ge 1$, and $X_n^-, n \ge 1$, satisfy the assumptions of the theorem of $X_n = X_n^+ - X_n^-$, we can without loss of generality suppose $X_n \ge 0$, As in proof of (6.8) we will prove the result first for a subsequence and then use monotonicity to control the values in between. This time however, we let $\alpha > 1$, and $k(n) = [a^n]$. Chebyshev's inequality implies that if $\in > 0$

$$\sum_{n=1}^{\infty} P(|T_{k(n)} - ET_{k(n)}| > \in k(n)) \leq \in - 2 \sum_{n=1}^{\infty} \mathrm{var}(T_{k(n)})/k(n)^2$$

$$= \in^{-2} \sum_{n=1}^{\infty} k(n)^{-2} \sum_{m=1}^{k(n)} \mathrm{var}(Y_m)$$

$$= \in^{-2} \sum_{m=1}^{\infty} \mathrm{var}(Y_m) \sum_{n:k(n) \geq m} k(n)^{-2}$$

where we have used Fubini's theorem to interchange the two summations (everything is ≥ 0). Now $k(n) = [\alpha^n]$ and $[\alpha^n] \geq \alpha^n/2$ for $n \geq 1$, so summing the geometric series and noting that the first term is $\leq m^{-2}$

$$\sum_{n:\alpha^n \geq m} [\alpha^n]^{-2} \geq 4 \sum_{n:\alpha^n \geq m} \alpha^{-2n} \leq 4(1-\alpha^{-2})^{-1} m^{-2}$$

Combining our computations shows

$$\sum_{n=1}^{\infty} P(|T_{k(n)} - ET_{k(n)}| > \in k(n)) \leq 4(1-\alpha^{-2})^{-1} \in^{-2} \sum_{m=1}^{\infty} E(Y_m^2) m^{-2} < \infty$$

by (b). Since $\in$ is arbitrary $(T_{k(n)} - ET_{k(n)})/k(n) \to 0$. The dominated convergence theorem implies $EY_k \to EX_1$ as $k \to \infty$, so $ET_{k(n)}/k(n) \to Ex_1$ and we have shown $T_{k(n)}/k(n) \to EX_1$ a.s. To handle the intermediate values, we observe that if $k(n) \leq m < k(n + 1)$

$$\frac{T_{k(n)}}{k(n+1)} \leq \frac{T_m}{m} \leq \frac{T_{k(n+1)}}{k(n)}$$

(here we use $Y_i \geq 0$), so recalling $k(n) = [\alpha^n]$ we have $k(n + 1)/k(n) \to \alpha$ and

$$\frac{1}{\alpha} EX_1 \leq \lim_{n \to \infty} \inf Tm/m \leq \lim_{m \to \infty} \sup T_m/m \leq \alpha EX_1$$

Since $\alpha > 1$ is arbitrary the proof is complete.

The next result shows that the strong law holds whenever EX_i exists.

(7.2) Theorem. Let $X_1, X_2, \ldots$ be i.i.d. with $EX_i^+ = \infty$ and $EX_i^- < \infty$. If $S_n = X_1 + \ldots + X_n$ then $Sn/n \to \infty$ a.s.

Proof Let $M > 0$ and $X_i^M = Xi \wedge M$. The X_i^M are i.i.d with $E|X_i^M| < \infty$ so if $S_i^M = X_i^M + \ldots + X_n^M$ then (7.1) implies $S_n^M/n \to EX_i^M$. Since $X_i \geq X_i^M$ it follows that

$$\lim_{n \to \infty} \inf S_n/n \geq \lim_{n \to \infty} S_n^M/n = EX_i^M$$

The monotone convergence theorem implies $E(X_i^M)+ \uparrow EX_i^+ = \infty$ as $M \uparrow \infty$, so $EX_i^+ = E(X_i^M)^+ - E(X_i^M)^- \uparrow \infty$ and we have $\lim \inf_{n \to \infty} S_n/n \geq \infty$ which imlies the desired result.

The rest of this section is devoted to applications of the strong law of large numbers.

Notes

Example: Renewal theory. Let X_1, X_2, ... be i.i.d. with $0 < X_i < \infty$. Let $T_n = X_1 + ... + X_n$ and think of T_n as the time of nth occurence of some event. For a concrete situation consider a diligent janitor who replaces a light bulb the instant it burns out. Suppose the first bulb is put in at time 0 and let X_i be the lifetime of the ith lightbulb. In this interpretation T_n is the time the nth light bulb burns out and $N_t = \sup\{n : T_n \leq t\}$ is the number of light bulbs that have burns out by time t.

Theorem. If $EX_1 = \mu \leq \infty$ then as $t \to \infty$, $N_t/t \to 1/\mu$ a.s. $(1/\infty = 0)$

3.2 Summary

- Many introductory probability texts treat this topic superficially, and more than once their vague formulations are misleading or plainly wrong. In this note, we consider a special case to clarify the relationship between the Weak and Strong Laws. The reason for doing so is that I have not been able to find a concise formal exposition all in one place. The material presented here is certainly not new and was gleaned from many sources.

  In the following sections, X1, X2, ... is a sequence of independent and indentically distributed random variabels with finite expectation m. We define the associated sequence $\overline{X}_i$ of partial sample means by

$$\overline{X}_i = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- Lemma. Let $Y_k = K_k 1_{(|X_k| \leq k)}$ and $T_n = Y_1 + ... + Y_n$. It is sufficient to prove that $T_n/n \to \mu$ a.s.

- Lemma. $\sum_{k=1}^{\infty} \mathrm{var}(Y_k)/k^2 \leq 4E|X_1| < \infty$.

- Lemma. If $y \geq 0$ then $2y\sum_{k>y} k^{-2} \geq 4$.

- Implies $S_n^M/n \to EX_i^M$. Since $X_i \geq X_i^M$ it follows that

$$\underline{\lim}_{n\to\infty} \inf S_n/n \geq \underline{\lim}_{n\to\infty} S_n^M/n = EX_i^M$$

  The monotone convergence theorem implies $E(X_i^M)+\uparrow EX_i^+ = \infty$ as M↑∞, so $EX_i^+ = E(X_i^M)^+ - E(X_i^M)^- \uparrow \infty$ and we have $\lim\inf_{n\to\infty} S_n/n \geq \infty$ which imlies the desired result.

3.3 Keywords

Probability Theory includes various theorems known as Laws of Large Numbers.

Strong law of large numbers. Let X1, X2, ... be pairwise independent identically distributed random variables with $E \mid X_i \mid < \infty$. Let $EX_i = \mu$ and $S_n = X_1 + ... + X_n$. Then $S_n/n \to \mu$ a.s. as $n \to \infty$.

3.4 Self Assessment

1. .... includes various theorems known as Laws of Large Numbers.
2. The Laws of Large Numbers make statements about the convergence of .... to m.
3. Lemma. $\sum_{k=1}^{\infty} \mathrm{var}(Y_k)/k^2 \leq$ ....
4. Lemma. If $y \geq 0$ then ....

5. The .... $E(X_i^M)+\uparrow EX_i^+ = \infty$ as $M\uparrow\infty$, so $EX_i^+ = E(X_i^M)^+ - E(X_i^M)^- \uparrow \infty$ and we have lim $\inf_{n\to\infty} S_n/n \geq \infty$ which imlies the desired result.

6. If $EX_1 = \mu \leq \infty$ then as $t \to \infty$, $N_t/t \to 1/\mu$ a.s. ....

3.5 Review Questions

1. Discuss the strong law of large number.

2. Discuss examples related to large number.

Answers: Self Assessment

1. Probability Theory 2. $\overline{X}_n$ 3. $4E|X_1| < \infty$ 4. $2y\sum_{k>y} k^{-2} \geq 4$

5. monotone convergence theorem implies 6. $(1/\infty = 0)$

3.6 Further Readings

Books Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.

Unit 4: Control Limit Theorem

CONTENTS

Objectives

After studying this unit, you will be able to:

- Define the central limit theorem
- Describe control limit theorem

Introduction

In Binomial distribution with parameters n and p is shown to be approximable by a Poisson distribution whenever n is large and p is such that np is a constant A z 0. An important limit theorem, known as the central limit theorem, is studied in Section 14.4. Central limit theorem essentially states that whatever the original distribution is (as long as it has finite variance), the sample mean computed from the observations following that distribution has an approximate normal distribution as long as the sample size (number of observations) is large. An important special case of this result is that binomial distribution can be approximated by an appropriate normal distribution for large samples.

4.1 Central Limit Theorem

The Central Limit Theorem (CLT) is one of the most important and useful results in probability theory. We have already seen that the sum of a finite number of independent normal random variables is normally distributed. However the sum of a finite number of independent non-normal random variables need not be normally distributed. Even then, according to the central limit theorem, the sum of a large number of independent random variables has a distribution that is approximately normal under general conditions. The CLT provides a simple method of computing the probabilities for the sum of independent random variables approximately. This theorem also suggests the reasoning behind why most of the data observed in practice leads to bell-shaped curves.

Let us now state the main theorem.

Theorem 3 (Central Limit Theorem) : Let X_1, X_2, ... be an infinite sequence of independent and identically distributed random variables with mean p and finite variance σ^2. Then, for any real x,

$$P\left[\frac{X_1+...+X_n-n\mu}{\sigma\sqrt{n}}\right] \to \phi(x) \text{ as } n \to \infty \qquad ...(5)$$

where $\phi(x)$ is the standard normal distribution function.

We have omitted the proof because proof of this result involves complex analysis and other concepts which are beyond the scope of this course. Let us try to understand the above statement more clearly. Let $S_n = X_1 + X_2 +....+ X_n$. Then we know that $P\left[\frac{S_n-n\mu}{\sigma\sqrt{n}} \le x\right]$ represents the distribution of the random variable $\frac{S_n-n\mu}{\sigma\sqrt{n}}$. Then the theorem says that the distribution of $\frac{S_n-n\mu}{\sigma\sqrt{n}}$ is approximately a standard normal distribution for sufficiently large n. Therefore the distribution of S_n will be approximately normal with mean $n\pi$ and variance $n\sigma^2$. In other words the theorem asserts that if $X_1 + X_2 ... + X_n$ are i.i.d.r. v' s of any kind (discrete or continuous) with finite variances, the $\Sigma\ S_n = X_1 + X_2 + ... + X_n$ will approximately be a normal distribution for sufficiently large n. The importance of the theorem lies in this fact. This theorem has got many applications. An important application is to a sequence of Bernoulli random variables.

Normal Approximation to the Binomial Distribution

Let X_i, $i \ge 1$ be a sequence of i.i.d. random variables such that

$$P[X_i = 1] = p, P[X_i = 0] = 1 - p$$

where $0 < p < l$.

Observe that $S_n = X_1 + ... + X_n$ has the binomial distribution with parameters n qnd p. You can check that $E(X_i) = p$ and Var $(X_i) = p\ (1\text{-}p)$ for any i which is finite and positive. An application of the central limit theorem gives the following result:

For every real x,

$$P\left[\frac{Sn-np}{\sqrt{n}\sqrt{[p(1-p)}} \le x\right] \to f(x) \qquad \text{as } n \to \infty.$$

In other words, for large n

$$P[\left[S_n \le np + x\sqrt{np(1-p)}\right] \simeq \phi(x) \qquad ...(6)$$

where $\simeq$ denotes that the quantities on both sides are approximately equal to each other.

An alternate way of interrupting the above approximation that a binomial distribution tends to be close to a normal distribution for large n. Let us explain this in mote detail.

Suppose S_n has binomial distribution with parameters n and p. Then, for $1 \leq r \leq n$,

$$P[S_n \leq r] = P\left[\frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{r - np}{\sqrt{np(1-p)}}\right]$$

$$\simeq \phi\left[\frac{r - np}{\sqrt{np(1-p)}}\right]$$

for large n by (2). In general, it is computationally difficult to calculate the exact probability

$$P[S_n \leq r] = \sum_{j=0}^{r}\left[\begin{matrix} n \\ j \end{matrix}\right] p^n (1-p)^{n-j}$$

when n is large. A close approximation to this probability can be obtained by computing

$$\phi\left(\frac{r - np}{\sqrt{np(1-p)}}\right)$$

where \$ is the standard normal distribution function. It has been found from empirical studies that this approximation is good when $n \geq 30$ and a better approximation is obtained by applying a slight correction, namely,

$$\phi\left[\frac{r + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right]$$

Let us illustrate these results by an example.

Example 6: The ideal size of a first year class in a college is 150. It is known from an earlier data that on the average only 30% of those accepted for admission will actually attend. Suppose the college admits 450 students. What is the probability that more than 150 first year students attend the college?

Let us denote by S_n the number of sludents that attend the college when n are admitted. Assuming that all the students take independent decision of either attending or not attending the college, we can suppose that ***S,*** has the binomial distribution with parameters n and p = 0.3. Here n = 450 and we are interested in finding the

$$P[S_n \geq 150].$$

Note that $E(S_n) = np = (450)\,(0.3) = 135$ and

$Var(S_n) = np(1 - p) = (135)(.7)$

Further more

$$P[S_n \geq 150] = 1 - P[S_n < 150]$$

$$\simeq 1 - P[S_n \leq 149]$$

and

$$P[S_n \leq 149] = \phi\left[\frac{149 + \frac{1}{2} - 135}{\sqrt{(135)(.7)}}\right]$$

$$= \phi(1.59)$$

Hence

$$P[S_n \geq 150] = 1 - \phi(1.59)$$

$$= .0559$$

This shows that the probability that more than 150 first year students attend is less than 6%. Let us now consider a different type of application of the central limit theorem.

Example 7: Suppose X_1, X_2 .... is a sequence of i.i.d. random variables each N(0, l). Then X_1^2, X_2^2, ... is a sequence of i.i.d. random variables each with X_1^2 -distribution.

Note that E(X_1^2) = 1 and Var (X_1^2) = 2 for any i. Hence by central limit theorem we get

$$P\left[\frac{X_1^2 + ... + X_n^2 - n}{\sqrt{2n}} \leq x\right] \to \phi(x) \text{ as } n \to \infty.$$

But S_n = X_1^2 + ... + X_1^2 has X_n^2 distribution. What we have shown just now is that if Sn has X_n^2 distribution, then $\frac{S_n - n}{\sqrt{2n}}$ has an approximate standard normal distribution for large n. In other words, for every real x,

$$P\left[\frac{S_n - n}{\sqrt{2n}} \leq x\right] \simeq \phi(x)$$

for large n whenever S, has *Xz* -distribution.

We make a remark now.

Remark 3 : The central limit theorem is central to the distribution theory needed for statistical inferential techniques to he developed in Block 4. You must have noted that the distribution of individual Xi in CLT could be discrete or continuous. The only condition that is imposed is that its variance has to be finite. In general, it is not easy to specify the size of n for a good approximation as it depends on the underlying distribution of {X_i}. However, it is found in practice that, in most cases, a good approximation is obtained whenever n is greater than or equal to 30.

We will stop our discussion on limit theorem now, though we shall refer to them off and on in the next block. Let us now do quick review of what we have covered in this unit.

4.2 Summary

- Obtain Poisson approximation to binomial;
- Discussed the central limit theorem and obtained normal approximation to binomial as an application.

As usual we suggest that you go back to the beginning of the unit and see if you have achieved the objectives. We have given our solutions to the exercises in the unit in the last section. Please go through them too. With this we have come to the end of this block.

- The Central Limit Theorem (CLT) is one of the most important and useful results in probability theory. We have already seen that the sum of a finite number of independent normal random variables is normally distributed. However the sum of a finite number of independent non-normal random variables need not be normally distributed. Even then, according to the central limit theorem, the sum of a large number of independent random variables has a distribution that is approximately normal under general conditions. The CLT provides a simple method of computing the probabilities for the sum of independent random variables approximately. This theorem also suggests the reasoning behind why most of the data observed in practice leads to bell-shaped curves.

4.3 Keywords

Binomial distribution with parameters n and p is shown to be approximable by a Poisson distribution whenever n is large and p is such that np is a constant A z 0.

Central Limit Theorem (CLT): The Central Limit Theorem (CLT) is one of the most important and useful results in probability theory.

4.4 Self Assessment

1. .... with parameters n and p is shown to be approximable by a Poisson distribution whenever n is large and p is such that np is a constant A z 0.
2. An important special case of this result is that binomial distribution can be approximated by an appropriate .... for large samples.
3. The .... is one of the most important and useful results in probability theory.
4. The CLT provides a simple method of computing the probabilities for the sum of .... approximately.

4.5 Review Questions

1. If X is binomial with n = 100 and p = 1/2, find an approximation for P[X = 50].
2. Suppose X is binomial with parameters n and p = 0.55. Determine the smallest n for which

$$P\left[\frac{X}{n} > \frac{1}{2}\right] \geq 0.95$$

   approximately.
3. If 10 fair dice are rolled, find the approximate probability that the sum of the numbers observed is between 30 and 40.
4. Suppose X is binomial with n = 100 and p = 0.1. Find the approximate value of $P(12 \leq X \leq 14)$ using
   (a) the normal approximation
   (b) the poisson approximation, and
   (c) the binomial distribution.

Answers: Self Assessment

1. Binomial distribution
2. normal distribution
3. Central Limit Theorem (CLT)
4. independent random variables

4.6 Further Readings

Books

Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.

Unit 5: Confidence Intervals

CONTENTS

Objectives

After studying this unit, you will be able to:

- Discuss statistic for various testing of hypotheses problems as well as to derive power functions
- Explain confidence intervals for parameters of various distributions
- Describe large sample tests.

Introduction

You have been introduced to the problem of testing of hypothesis and also to some basic concepts of the theory of testing of hypothesis. There you have studied two important procedures fortesting statistical hypotheses,viz. using Neyman-Pearson Lemma and the likelihood ratio test. In this unit, you will be exposed to the problem of testing statistical hypotheses involving the parameters of some important distributions through some selected examples. In this unit, you will also be exposed to the problem of constructing confidence intervals for parameters of some important distributions through some selected examples. You will also learn the use of chi-square test for goodness of fit.

5.1 Some Common Tests of Hypothesis for Normal Populations

We have already described with examples two procedures for testing statistical hypotheses. In this section we will employ Neyman-Pearson Lamma and likelihood ratio test for testing of hypothesis related to a normal population.

Example 1: Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively. It is desired to obtain a test statistic for testing $H_0 : \mu_1 = \mu_2$ against $\mu_1 : \mu_1 \neq \mu_2$ when σ^2 (> 0) is unknown.

In order to obtain the test statistic, we use the likelihood ratio test. We have

$$\Omega = \{ (\mu_1, \mu_2, \sigma^2) : -\infty < \mu_1, \mu_2 < \infty, \sigma^2 > 0 \}$$

$$\Omega_0 = \{ \mu_1 = \mu_2 = \mu \text{ (say)}, \sigma^2) : - \infty < \mu < \infty, \sigma^2 > 0\}$$

We shall write $\theta = (\mu_1, \mu_2, \sigma^2)$

We have

$$\underset{\theta \in Q_0}{\{\sup} L(\theta | X, Y)$$

$$= \text{Sup} \frac{1}{(2\pi)^{\frac{m+n}{2}} (\sigma^2)^{\frac{m+n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_1^m (X_i - \mu_1)^2 + \sum_1^n (Y_i - \mu_2)^2\right]\Bigg]\right\}$$

Under H_0, $\mu_1 = \mu_2 = \mu$ and the maximum likelihood estimate of μ is

$$\hat{\mu} = \frac{m\bar{X} + n\bar{Y}}{m+n} \text{ and of } \sigma^2 \text{ is}$$

$$\hat{\sigma}^2 = \frac{1}{m+n}\left[\sum_1^m (X_i - \mu_1)^2 + \sum_1^n (X_1 - \mu_2)^2 + \frac{mn}{(m+n)}(\bar{X} - \bar{Y})^2\right]$$

= u (say)

$$\text{Thus } \underset{\theta \in \Omega_0}{\text{Sup L}} (\theta | X, Y) = \frac{1}{(2pu')^{\frac{m+n}{2}}} \exp\left[-\frac{1}{2u'}(m+n)u'\right]$$

$$= \left(\frac{1}{2\pi u'}\right)^{\frac{m+n}{2}} \exp\left(-\frac{(m+n)}{2}\right)$$

Under H_1, the maximum likelihood estimates of μ_1, μ_2 and σ^2 are respectively

$$\hat{\mu}_1 = \bar{X}, \hat{\mu}_2 = \bar{Y}, \hat{\sigma}^2 = \frac{\sum_1^m (X_i - \bar{X})^2 + \sum_1^n (Y - \bar{Y})^2}{m+n} = u(\text{say})$$

and

$$\underset{\theta \in Q_0}{\{\sup} L(\theta | X, Y)$$

$$= \left(\frac{1}{2\pi u}\right)^{\frac{m+n}{2}} \exp\left(-\frac{m+n}{2}\right)$$

The likelihood ratio test is thus

$$\lambda(X, Y) = \frac{\underset{\theta \in \Omega_0}{\{\sup} L(\underline{\theta} | X, Y)}{\underset{\theta \in \Omega}{\{\sup} L(\underline{\theta} | X, Y)}$$

$$= \left(\frac{u}{u'}\right)^{\left(\frac{m+n}{2}\right)}$$

$$= \left[\frac{\sum_{1}^{m}(Xi-\overline{X})^2+\sum_{1}^{n}(Yi-\overline{Y})^2}{\sum_{1}^{m}(Xi-\overline{X})^2+\sum_{1}^{n}(Yi-\overline{Y})^2+\frac{mn}{(m+n)}(Xi-\overline{X})^2}\right]^{\frac{m+n}{2}}$$

Now under null hypothesis, $\mu_1 = \mu_2 = \mu$, and $t = \frac{\overline{X}-\overline{Y}}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}}$ follows a Student's t distributions with m + n – 2 degrees of freedom, where $S^2 = \frac{u(m+n)}{m+n-2}$

Thus

$$t^2 = \frac{(m+n-2)mn(\overline{X}-\overline{Y})^2}{(m+n)\left\{\sum_{1}^{m}(X_i-\overline{X})^2+\sum_{1}^{n}(Y-\overline{Y})^2\right\}}$$

and

$$\lambda(X, Y) = \left[\frac{1}{1+\frac{t^2}{m+n-2}}\right]^{\frac{m+n}{2}} < c$$

The likelihood ratio critical region is given by

$$\lambda(X, Y) = \left[\frac{1}{1+\frac{t^2}{m+n-2}}\right]^{\frac{m+n}{2}} < c$$

where c is to be determined so that

$$\underset{\theta\in\Omega_0}{Sup}\, P_\theta\left[\lambda(X,Y) < c\right] = \alpha$$

Since $\lambda(X, Y)$ is a decreasing function of $t^{2/}(m + n - 2)$ we reject H_0 if

$$\frac{t^2}{(m+n-2)} > c^{2/(m+n)}$$

or

$$|t| > c_1$$

where c_1 is so chosen that

Let $c_1 = t_{m+n-2,\, \alpha/2}$ in accordance with the diskbution oft under Ho. Thus, the two sided test obtained is

$$\left| \frac{(\overline{X} - \overline{Y}}{S} \sqrt{\frac{mn}{(m+n)}} \right| > t_{m+n-2, \alpha/2}$$

Example 2: Let $X_1, \ldots, X_n$, be a random sample from N (μ, σ^2), p is known and $\sigma^2 > 0$, is unknown. We wish to obtain a test statistic for testing $H_0 : \sigma^2 - \sigma_0^2$ against an alternative $H_1 : \sigma^2 = \sigma_1^2 \ (> \sigma_0^2)$.

We have

$$P_{\theta_1}(\underline{X}) = \frac{1}{(2\pi\sigma_1)^{n/2}} \exp\left[-\frac{1}{2\sigma_1^2} \sum_1^n (Xi - \mu)^2 \right]$$

$$P_{\theta_0}(\underline{X}) = \frac{1}{(2\pi\sigma_0)^{n/2}} \exp\left[-\frac{1}{2\sigma_0^2} \sum_1^n (Xi - \mu)^2 \right]$$

Using Neyman-Pearson Lemma, the test statistic is

$$T(X) = \frac{P_{\theta_1}(\underline{X})}{P_{\theta_0}(\underline{X})} \geq k$$

$$\Rightarrow \left(\frac{\sigma_0}{\sigma_1} \right)^{n/2} \exp\left\{ 1/2 \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2} \right) \sum_1^n (X_i - \mu)^2 \right.$$

$$\Rightarrow (\sigma_1^2 - \sigma_0^2) \sum_1^n (X_i - \mu^2) \geq k \text{, taking logarithms}$$

$$\Rightarrow \sum_1^n (X_i - \mu^2) \geq k_1 \text{, since } \sigma_1^2 > \sigma_0^2, \text{under } H_1$$

Here k_1 is so determined that

$$P_{\theta_0}(T, (\underline{X}) \geq k) = \alpha$$

$$\Rightarrow P_{\theta_0}\left[\sum_1^n (X_i - \mu)^2 \geq k_1 \right] = \alpha$$

$$\Rightarrow P_{\theta_0}\left[\sum_1^n (X_i - \mu)^2 / \sigma_0^2 \geq k_1 / \sigma_0^2 \right]$$

Under the null hypothesis, since σ² = $\sigma_0^2, \sum_1^n (X_i - \mu)^2 / \sigma_0^2\}$ has a χ_n^2 distribution (chi-square distribution with n degrees of freedom). Let $\chi_{n,\alpha}^2$ be the upper -α probability point of χ_n^2. The test statistic is thus

$$\sum_1^n (X_i - \mu)^2 >= k1 \text{ and hence}$$

$$C0 = \left\{ X \mid \sum_1^n (X_i - \mu)^2 / \sigma_0^2 > c_{n,\alpha}^2 \right\}$$

On the other hand, if the alternative hypothesis is H_1 : σ² = σ_1^2 ($\sigma_1^2 < \sigma_0^2$), then the test statistic is and hence

$$\sum_1^n (X_i - \mu)^2 < k_2$$

where $\chi_{n,1-\alpha}^2$ is the lower α -probability point of the χ^2 distribution with n degrees of freedom.

Example 3: Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples from N(μ_1, σ_1^2) and N(μ_2, σ_2^2). We wish to obtain a test statistic for testing H_0 : $\sigma_1^2 = \sigma_2^2$ against H_1 : $\sigma_1^2 \neq \sigma_2^2$.

Here Ω = $\{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : -\infty < \mu_i < \infty, \sigma_1^2 > 0, i = 1, 2\}$

and Ω_0 = $\{(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : -\infty < \mu_i < \infty, i = 1, 2, \sigma_1^2 = \sigma_2^2 = \sigma^2 > 0\}$

We shall use $\underline{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

Also L (8 | X, Y)

$$= \left(\frac{1}{2\pi}\right)^{\frac{m+n}{2}} \left(\frac{1}{\sigma_1^2}\right)^{m/2} \left(\frac{1}{\sigma_1^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma_1^2}\sum_1^n (X_i - \mu_1)^2 - \frac{1}{2\sigma_2^2}\sum_1^n (Y_i - \mu_2)^2\right\}$$

The maximum likelihood estimates of μ_1, μ_2, σ_1^2, σ_2^2 are respectively

$$\hat{\mu}_1 = \frac{1}{m}\sum_1^m X_i = \overline{X}, \hat{\mu}_2 = \frac{1}{n}\sum_1^n Y_i = \overline{Y}$$

$$\hat{\sigma}_1^2 = \frac{1}{m}\sum_1^m (X_i - \overline{X})^2, \hat{\sigma}_2^2 = \frac{1}{n}\sum_1^n (Y_i - \overline{Y})^2$$

Further, if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, the maximum likelihood estimate of σ² is

$$\hat{\sigma}^2 = \frac{1}{(m+n)}\left[\sum_1^m (X_i - \overline{X})^2 + \sum_1^n (Y_i - \overline{Y})^2\right]$$

Notes

Thus

$$\underset{\theta \in Q_0}{\text{Sup}}\, L(\theta \mid X, Y)$$

$$= \frac{\exp\{-(m+n)/2\}}{[2\pi/(m+n)]^{\frac{m+n}{2}} \left\{\sum_1^m (X_i - \overline{X})^2 + \sum_1^n (Y_i - \overline{Y})^2\right\}^{\frac{m+n}{2}}}$$

and

$$\underset{\theta \in Q_0}{\text{Sup}}\, L(\theta \mid X, Y)$$

$$= \frac{\exp\{-(m+n)/2\}}{(2\pi/m)^{m/2}(2\pi/n)^{n/2} \left\{\sum_1^m (X_i - \overline{X})^2\right\}^{\frac{m}{2}} \left\{\sum_1^n (Y_i - \overline{Y})^2\right\}^{\frac{n}{2}}}$$

$$= \frac{\exp\{-(m+n)/2\}}{(2\pi/m)^{m/2}(2\pi/n)^{n/2} \left\{\sum_1^m (X_i - \overline{X})^2\right\}^{\frac{m}{2}} \left\{\sum_1^n (Y_i - \overline{Y})^2\right\}^{\frac{n}{2}}}$$

The likelihood ratio test is thus

$$\lambda\,(X, Y) = \frac{\underset{\theta \in \Omega_0}{\text{Sup}}\, L(\theta \mid X, Y)}{\underset{\theta \in \Omega_0}{\text{Sup}}\, L(\theta \mid X, Y)}$$

$$= \left(\frac{m}{m+n}\right)^{m/2} \left(\frac{n}{m+n}\right)^{n/2} \frac{\left\{\sum_1^m (X_i - \overline{X})^2\right\}^{\frac{m}{2}} \left\{\sum_1^n (Y_i - \overline{Y})^2\right\}^{\frac{n}{2}}}{\left\{\sum_1^m (X_i - \overline{X})^2 + \sum_1^n (Y_i - \overline{Y})^2\right\}^{\frac{m+n}{2}}}$$

Now

$$\frac{\left\{\sum_1^m (X_i - \overline{X})^2\right\}^{\frac{m}{2}} \left\{\sum_1^n (Y_i - \overline{Y})^2\right\}^{\frac{n}{2}}}{\left\{\sum_1^m (X_i - \overline{X})^2 + \sum_1^n (Y_i - \overline{Y})^2\right\}^{\frac{m+n}{2}}}$$

We have

$$\lambda\,(X, Y) = \frac{\left(\frac{m}{m+n}\right)^{m/2} \left(\frac{n}{m+n}\right)^{m/2}}{\left[1 + \frac{(m-1)}{(m-1)} f\right]^{n/2} \left[1 + \frac{(n-1)}{(m-1)}(1/f)\right]^{m/2}}$$

The likelihood ratio test criterion rejects Ho if λ (X, Y) < c

It is easy to see that λ (X, Y) is a monotonic function of f and h (X, Y) < c is equivalent to f < c_1 or f > c. Under H_0,

$$f = \frac{\sum_1^m (X_i - \overline{X})^2/(m-1)}{\sum_1^n (Y_i - \overline{Y})^2/(n-1)}$$

has Snedecar's F (m – 1, n – 1) distribution, so that c_1, c_2 can be selected, such that

$$\underset{\theta \in \Omega_0}{\text{Sup}}\, P_{\underline{\theta}}\, [\lambda(X, Y) < c] = \alpha$$

or

$$P(F \leq c_1) = P(F \geq c_2) = \alpha/2$$

Thus c_2 = F(m – l, n – 1, α/2) is the upper α/2 probability point of F (m – 1, n – 1) distribution and c_1 = F (m – 1, n – 1, l – α/2) is the lower α/2 probability point of F (m - 1, n - 1).

5.2 Confidence Intervals

In you have been briefly exposed to some notions of interval estimation of a parameter. In this section we discuss in detail the problem of obtaining interval estimates of parameters and describe, through examples, some methods of constructing interval extimates of parameters. We may remind you again that an interval estimate is also called a confidence interval or a confidence set. We first illustrate through small examples the need for constructing confidence intervals. Suppose X denotes the tensile strength of a copper wire. A potential user may desire to know the lower bound for the mean of X, so that he can use the wire if the average tensile strength is not less than say go. Similarly, if the random 'variable X measures the toxicity of a drug, a doctor may wish to have a knowledge t of the upper bound for the hean of X in order to prescribe this dmg. If the random variable X measures the waiting times at the emergency room of a large city hospital, one may be interested in the mean waiting time at this emergency room. In this case we wish to obtain both the lower and upper bounds for the waiting time.

In this unit we are concerned with the problem of determining confidence intervals for a parameter. A formal definition of a confidence interval has been given in Section 15.6. However, for the sake of completeness we define some terms here.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with density (or, mass) function f (x, θ), $\theta \in \Omega \subseteq R^1$. The object is to find statistics r_L ($X_1 \ldots\ldots, X_n$) and r_U ($X_1, \ldots, X_n$) such that

P_θ { (r_L ($X_1,\ldots, X_n$) ≤ θ ≤ r_U($X_1 ,\ldots, X_n$)] ≥ 1 – α for all $\theta \in \Omega \subseteq R^1$. The interval $(r_L(\underline{X}), r_U(\underline{X}))$ is called a confidence interval and the quantity

$$\inf P_\theta [r_L (X_1,\ldots,X_n) \leq \theta \leq r_U (X_1,\ldots.,X_n)]$$

will be referred to as the confidence co-efficient associated with the random interval.

We now give some examples of construction of confidence intervals.

Example 4: Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population, N (μ, σ^2). We wish to obtain a (1 – α) level confidence interval for μ.

Let $\overline{X} = n^{-1} \sum_{1}^{n} X_i$. Consider the interval $(\overline{X} - a, \overline{X} + b)$. In order for this to be a (1 – α)-level confidence interval, we must have

$$P\{\overline{X} - a < \mu < \overline{X} + b\} \geq 1 - \alpha$$

Thus

$$P\left\{-\frac{b}{\sigma}\sqrt{n} < \frac{(\overline{X} - \mu)}{\sigma}\sqrt{n} < \frac{a}{b}\sqrt{n}\right\} \geq 1 - \alpha$$

Since, $\frac{(\overline{X} - \mu)}{\sigma}\sqrt{n} / \sim N(0,1)$ we can choose a and b to satisfy

$$P\left\{-\frac{b}{\sigma}\sqrt{n} < \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} < \frac{a}{\sigma}\sqrt{n}\right\} = 1 - a$$

provided that a is known. There are infinitely many such pairs of values (a, b). In Inference particular, an intuitively reasonable choice is a = b = c , say

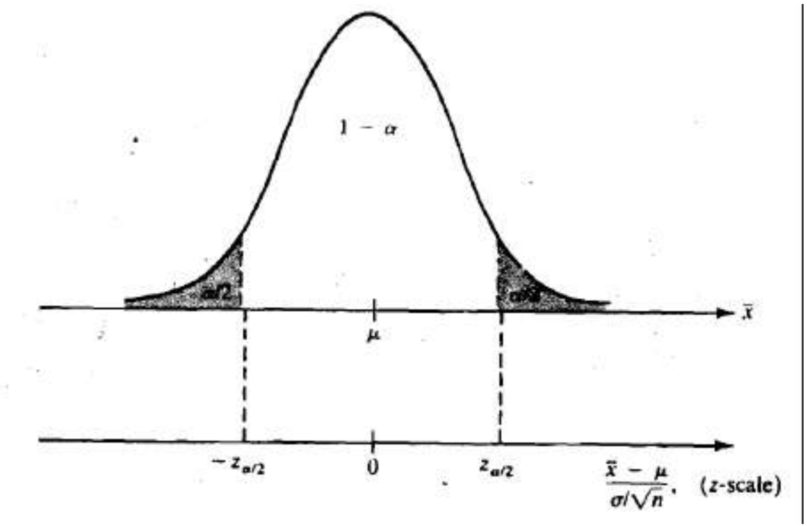
In that case

$\frac{c\sqrt{n}}{\sigma} = Z_{\alpha/2}$ where $Z_{\alpha/2}$ is the α/2 percent point of the standard normal distribution, and the confidence interval is

$$(\overline{X} - (\sigma/\sqrt{n})Z_{\alpha/2}, \overline{X} + (\sigma/\sqrt{n})Z_{\alpha/2})$$

The length of the interval is $(2\sigma/\sqrt{n})\ Z_{\alpha/2}$ Given a and a one can choose n to get a confidence interval of desired length.

Figure 5.1: Probability density curve of normal distribution with mean m and variance ●/n. Shows area ●/2 in each of two talls



If σ^2 is unknown, we have from

$$P\{-b < \overline{X} - \mu < a\} \geq 1 - \alpha$$

that

$$P\left\{-\frac{b}{S}\sqrt{n} < \frac{(\overline{X}-\mu)}{S} < \frac{a}{S}\sqrt{n}\right\} \geq 1-\alpha$$

It is known that $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$. We can choose pairs of values (a, b) using a students t-distribution with (n – 1) degrees of freedom such

$$P\left\{-\frac{b\sqrt{n}}{S} < \frac{\overline{X}-\mu}{Sb/\sqrt{n}} < \frac{a\sqrt{n}}{S}\right\} = 1-\alpha$$

In particular, an intuitively reasonable choice is a = b = c say. Then

$$\frac{c\sqrt{n}}{S} = t_{n-1,\alpha/2}$$

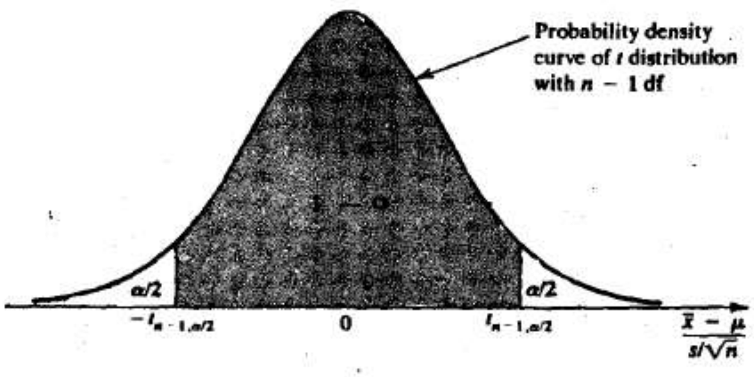
and $(\overline{X}-(S/\sqrt{n})t_{n-1,\alpha/2}, \overline{X}+(S/\sqrt{n})t_{n-1,\alpha/2})$ is 1 – α level confidence interval for μ. The length of the interval is $(2S/\sqrt{n})t_{n\text{-}1,\alpha/2}$, which is no longer constant.

Therefore, in this case one cannot choose n to get a fixed length confidence interval of level 1 – α. The expected length is, however,

$$\frac{2}{\sqrt{n}}t_{n-1,\alpha/2}Es(S) = \frac{2}{\sqrt{n}}t_{n-1,\alpha/2}\sqrt{\frac{2}{n-1}}\frac{\Gamma(n/2)}{\Gamma(n-1)/2)}\sigma$$

which can be made as small as we want by making a proper choice of n for a given σ and α.

Figure 5.2 : t Values such that there is an area ●/2 in the right tall and ●/2 in the left tall of the distribution.



Example 5: Let $X_1, X_2, \ldots, X_n$ be a random sample, from N(μ, σ²). It is desired to obtain a confidence interval for σ² when μ is unknown.

Consider the interval (aS², bS²), a, b > 0, $S^2 = (n-1)^{-1}\sum_{1}^{n}(X_i-\overline{X})^2$. We have

$$P\{ aS^2 < \sigma^2 < bS^2 \} \geq 1 - a$$

so that

$$P\left\{b^{-1} < \frac{S^2}{\sigma^2} < a^{-1}\right\} \geq 1 - \alpha$$

It is known that

$$(n - 1)S^2/\sigma^2 \sim \chi^2_{n-1}$$

We can therefore choose pairs of intervals (a, b) from the tables of the chi-square distribution. In particular we can choose a, b so that

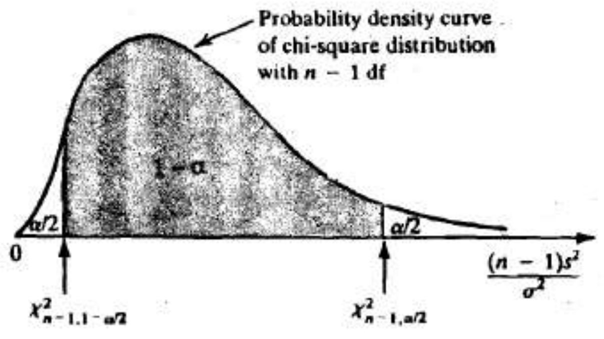
$$P\left\{\frac{S^2}{\sigma^2} \geq \frac{1}{a}\right\} = \alpha/2 = P\left\{\frac{S^2}{\sigma^2} \leq \frac{1}{b}\right\}.$$

Then $\frac{n-1}{a} = x^2_{n-1,\alpha/2}$ and $\frac{n-1}{b} = x^2_{n-1,1-\alpha/2}$ and the 1 – α level confidence interval for σ^2 when μ is unknown is

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right)$$

If however, μ is known then (n – 1) S^2 is replaced by $\sum_1^n (X_i - \mu)^2$ and the degrees of freedom of χ^2 is n instead of n – 1, for $\sum_1^n (X_i - \mu)^2/\sigma^2 \sim c^2_n$.

Figure 5.3 : Chi-square values such that area 1 – ●/2 and ●/2 are to their right.



Example 6: Let $X_1, \ldots, X_2$ and $Y_1, \ldots, Y_m$ denote respectively independent random samples from the two independent distributions having respectively the probability density functions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$. We wish to obtain a confidence interval for $\mu_1 - \mu_2$.

Consider the interval $\{(\overline{X} - \overline{Y}) - a, (\overline{X} - \overline{Y}) + b\}$. In order that this is a (1 - α) level confidence interval, we mbt have

$$P\{(\overline{X} - \overline{Y}) - a < \mu_1 - \mu_2 < (\overline{X} - \overline{Y}) + b\} \geq 1 - \alpha$$

which is the same as

$$P\{-b < (\overline{X} - \overline{Y}) - (\mu_1 - \mu_2) < a\} \geq 1 - \alpha$$

or

$$P\left\{\frac{b}{\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{a}{\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}}\right\} \geq 1 - \alpha$$

Here $\overline{X} = \frac{1}{n}\sum_1^n X_i$ and $\overline{Y} = \frac{1}{m}\sum_1^m Y_i$

Since $\dfrac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} \sim N(0,1)$.

we can choose a and b to satisfy

$$P\left\{\frac{-b}{\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{a}{\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}}\right\} \geq 1 - \alpha$$

provided that σ is known. There are infinitely many such pairs of values (a, b). In particular, an intuitively reasonable choice is a = b = c, say. In that case $c / \left\{s\left(\frac{1}{n}+\frac{1}{m}\right)^{1/2}\right\} = Z_{\alpha/2}$ and the confidence interval is

$$\left\{(\overline{X} - \overline{Y}) - \sigma\left(\frac{1}{n}+\frac{1}{m}\right)^{1/2} Z_{\alpha/2}, (\overline{X} - \overline{Y}) + \sigma\left(\frac{1}{n}+\frac{1}{m}\right)^{1/2} Z_{\alpha/2}\right\}$$

The length of the intaval is $2\sigma\left(\frac{1}{n}+\frac{1}{m}\right)^{1/2} Z_{\alpha/2}$. Given α and σ one can choose n and m to get a desired length confidence interval.

If σ^2 is unknown, we have from

$$P\{-b < (\overline{X} - \overline{Y}) - (\mu_1 - \mu_2) < a\} \geq 1 - \alpha$$

that

$$P\left\{\frac{-b}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{a}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}}\right\} \geq 1 - \alpha$$

where $\frac{\sum_{1}^{n}(X_i-\overline{X})2+\sum_{1}^{m}(Y_i-\overline{Y})^2}{(n+m-2)}=\frac{(n-1)S_x^2+(m-1)S_y^2}{n+m-2}$

It is known that

$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S2\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}}\sim t_{n+m-2}$. We can choose pairs of values (a, b) using Student's t-distribution with n + m – 2 degrees of freedom such that

$$\left\{\frac{-b}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}}<\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}}<\frac{a}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}}\right\}=1-\alpha$$

In particular, an intuitively reasonable choice is a = b = c, say. Then

$$\frac{c}{S\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)}}=t_{n+m-2,\alpha/2}$$

and $\left\{(\overline{X}-\overline{Y})-S\left(\frac{1}{n}+\frac{1}{m}\right)^{1/2}tn+m-2,a/2,(\overline{X}-\overline{Y})+S\left(\frac{1}{n}+\frac{1}{m}\right)^{1/2}t_{n+m-2,\alpha/2}\right\}$

is a 1 – α level confidence interval for μ_1 – μ_2.

Example 7: Let X_1, . . . , X_n, and Y_1, . . . , Y_m, n, m > 2, denote respectively independent random samples from the two distributions having respectively the probability density fundions $N(\mu_1,\sigma_1^2)$ and N $N(\mu_2,\sigma_2^2)$. We wish to obtain a confidence interval for the ratio σ_2^2/σ_1^2 when μ_1 and μ_2 are unknown.

Consider the interval $(a\,S_2^2/S_1^2,\,bS_2^2/S_1^2)$ a , b > 0, where

$$S_1^2=\frac{1}{(n-1)}\sum_{1}^{n}(X_i-\overline{X})^2,S_2^2=\frac{1}{(m-1)}\sum_{1}^{m}(Y_i-\overline{Y})^2,$$

$$\overline{X}=\frac{1}{n}\sum_{1}^{n}X_i,\overline{Y}=\frac{1}{m}\sum_{1}^{m}Y_i.\text{ We have}$$

$$P\left\{a\frac{S_2^2}{S_1^2}<\frac{\sigma_2^2}{\sigma_1^2}<b\frac{S_2^2}{S_1^2}\right\}\geq 1-\alpha$$

so that

$$P\left\{\frac{1}{b}<\frac{(S_2^2/S_1^2)}{(\sigma_2^2/\sigma_1^2)}<\frac{1}{a}\right\}\geq 1-\alpha$$

It is also known that if X and Y are independent χ^2 random variables with m and n degrees of freedom respectively, the random variable F = (X/m)/(Y/n) is said to have an F-distribution with (m, n) degrees of freedom. It is also known that if X has an F (m, n) distribution then l/X has an F (n, m) distribution, and $F_{m,\,n,\,1-\alpha} = 1/F_{n,\,m,\,\alpha}$. Therefore

$$\frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} = \frac{S_2^2/S_1^2}{\sigma_2^2/\sigma_1^2} \sim F_{(m-1),(n-1)}$$

We can therefore choose paris of values (a, b) from the tables of F-distribution. In particular, we can choose a and b so that

$$P\left\{\frac{(S_2^2/\sigma_2^2)}{(S_1^2/\sigma_1^2)} \geq \frac{1}{a}\right\} = \alpha/2 = P\left\{\frac{(S_2^2/\sigma_2^2)}{(S_1^2/\sigma_1^2)} \leq \frac{1}{b}\right\}$$

Then $\frac{1}{a} = F_{m,n,\alpha/2}$ and $\frac{1}{b} = F_{m,n,1-\alpha/2}$ and the 1 – α level confihnce interval for σ_2^2/σ_1^2 is

$$\left(\frac{S_2^2}{S_1^2} - F_{n,m,1-\alpha/2}, \frac{S_2^2}{S_1^2} - F_{n,m,1/\alpha/2}\right)$$

5.3 Summary

- We have already described with examples two procedures for testing statistical hypotheses. In this section we will employ Neyman-Pearson Lamma and likelihood ratio test for testing of hypothesis related to a normal population.
- In you have been briefly exposed to some notions of interval estimation of a parameter. In this section we discuss in detail the problem of obtaining interval estimates of parameters and describe, through examples, some methods of constructing interval extimates of parameters. We may remind you again that an interval estimate is also called a confidence interval or a confidence set. We first illustrate through small examples the need for constructing confidence intervals. Suppose X denotes the tensile strength of a copper wire. A potential user may desire to know the lower bound for the mean of X, so that he can use the wire if the average tensile strength is not less than say go. Similarly, if the random 'variable X measures the toxicity of a drug, a doctor may wish to have a knowledge t of the upper bound for the hean of X in order to prescribe this dmg. If the random variable X measures the waiting times at the emergency room of a large city hospital, one may be interested in the mean waiting time at this emergency room. In this case we wish to obtain both the lower and upper bounds for the waiting time.
- In this unit we are concerned with the problem of determining confidence intervals for a parameter. A formal definition of a confidence interval has been given in Section 15.6. However, for the sake of completeness we define some terms here.

5.4 Keywords

Confidence interval: Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with density (or, mass) function f (x, θ), $\theta \in \Omega \subseteq R^1$. The object is to find statistics $r_L(X_1 \ldots\ldots, X_n)$ and $r_U(X_1, \ldots, X_n)$ such that

$P_\theta\{(r_L(X_1,\ldots, X_n) \leq \theta \leq r_U(X_1,\ldots, X_n)] \geq 1 - \alpha$ for all $\theta \in \Omega \subseteq R^1$. The interval $(r_L(\underline{X}), r_U(\underline{X}))$ is called a confidence interval.

Notes

5.5 Self Assessment

1. If the random variable X measures the toxicity of a drug, a doctor may wish to have a knowledge t of the .... for the hean of X in order to prescribe this dmg.

2. If the .... X measures the waiting times at the emergency room of a large city hospital, one may be interested in the mean waiting time at this emergency room.

5.6 Review Questions

1. LetX_1, X_2, . . . , X_n be a random sample from a normal population, N (μ, σ^2). We wish to obtain a (1 – α) level confidence interval for μ.

2. Let X_1, X_2, . . . , X_n be a random sample, from N(μ, σ^2). It is desired to obtain a confidence interval for σ^2 when μ is unknown.

3. Let X_1, X_2, . . . , X_n be a random sample from a normal population, N (μ, σ^2). We wish to obtain a (1 – 2α) level confidence interval for μ.

4. Let X_1, . . . , X_2 and Y_1, . . . , Y_m denote respectively independent random samples from the two independent distributions having respectively the probability density functions N(μ_1, σ^2) and N(μ_2, σ^2). We wish to obtain a confidence interval for μ_1 – μ_2.

5. Let X_1, X_2, . . . , X_n be a random sample from a normal population, N (μ, σ^2). We wish to obtain a (1 – 3α) level confidence interval for μ.

6. Let X_1, X_2, . . . , X_n be a random sample from a normal population, N (μ, σ^2). We wish to obtain a $(1 - \alpha)^2$ level confidence interval for μ.

Answers: Self Assessment

1. upper bound
2. random variable

5.7 Further Readings

Books

Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.

Unit 6: Correlation

CONTENTS

Objectives

After studying this unit, you will be able to:

- Definition of Correlation
- Discuss Scatter Diagram
- Explain Karl Pearson's Coefficient of Linear Correlation
- Discuss Properties of Coefficient of Correlation
- Describe Probable Error of r

Introduction

So far we have considered distributions relating to a single characteristics. Such distributions are known as Univariate Distribution. When various units under consideration are observed simultaneously, with regard to two characteristics, we get a Bivariate Distribution. For example, the simultaneous study of the heights and weights of students of a college. For such data also, we can compute mean, variance, skewness etc. for each individual characteristics. In addition to this, in the study of a bivariate distribution, we are also interested in knowing whether there exists some relationship between two characteristics or in other words, how far the two variables, corresponding to two characteristics, tend to move together in same or opposite directions i.e. how far they are associated.

The knowledge of this type of relationship is useful for predicting the value of one variable given the value of the other. It also helps in understanding and analysis of various economic and business problems. It should be noted here that statistical relations are different from the exact mathematical relations. Given a statistical relation $Y = a + bX$, between two variables X and Y, we can only get a value of Y that we expect on the average for a given value of X. The study of relationship between two or more variables can be divided into two broad categories:

(i) To determine whether there exists some sort of association between the variables. If so, what is the degree of association or the magnitude of correlation between the two.

(ii) To determine the most suitable form of the relationship between the variables given that they are correlated.

The first category relates to the study of 'Correlation' which will be discussed in this chapter and the second relates to the study of 'Regression', to be discussed in next chapter.

6.1 Definition of Correlation

Various experts have defined correlation in their own words and their definitions, broadly speaking, imply that correlation is the degree of association between two or more variables. Some important definitions of correlation are given below:

(i) "*If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in other(s) then they are said to be correlated.*"

— L.R. Connor

(ii) "*Correlation is an analysis of covariation between two or more variables.*"

— A.M. Tuttle

(iii) "*When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.*"

— Croxton and Cowden

(iv) "*Correlation analysis attempts to determine the 'degree of relationship' between variables*".

— Ya Lun Chou

Correlation Coefficient: It is a numerical measure of the degree of association between two or more variables.

6.1.1 Scope of Correlation Analysis

The existence of correlation between two (or more) variables only implies that these variables (i) either tend to increase or decrease together or (ii) an increase (or decrease) in one is accompanied by the corresponding decrease (or increase) in the other. The questions of the type, whether changes in a variable are due to changes in the other, i.e., whether a cause and effect type relationship exists between them, are not answered by the study of correlation analysis. If there is a correlation between two variables, it may be due to any of the following situations:

(i) ***One of the variable may be affecting the other:*** A correlation coefficient calculated from the data on quantity demanded and corresponding price of tea would only reveal that the degree of association between them is very high. It will not give us any idea about whether price is affecting demand of tea or vice-versa. In order to know this, we need to have some additional information apart from the study of correlation. For example if, on the basis of some additional information, we say that the price of tea affects its demand, then price will be the cause and quantity will be the effect. The causal variable is also termed as independent variable while the other variable is termed as dependent variable.

(ii) ***The two variables may act upon each other:*** Cause and effect relation exists in this case also but it may be very difficult to find out which of the two variables is independent. For example, if we have data on price of wheat and its cost of production, the correlation between them may be very high because higher price of wheat may attract farmers to produce more wheat and more production of wheat may mean higher cost of production, assuming that it is an increasing cost industry. Further, the higher cost of production may in turn raise the price of wheat. For the purpose of determining a relationship between the two variables in such situations, we can take any one of them as independent variable.

(iii) ***The two variables may be acted upon by the outside influences:*** In this case we might get a high value of correlation between the two variables, however, apparently no cause and effect type relation seems to exist between them. For example, the demands of the two commodities, say X and Y, may be positively correlated because the incomes of the consumers are rising. Coefficient of correlation obtained in such a situation is called a spurious or nonsense correlation.

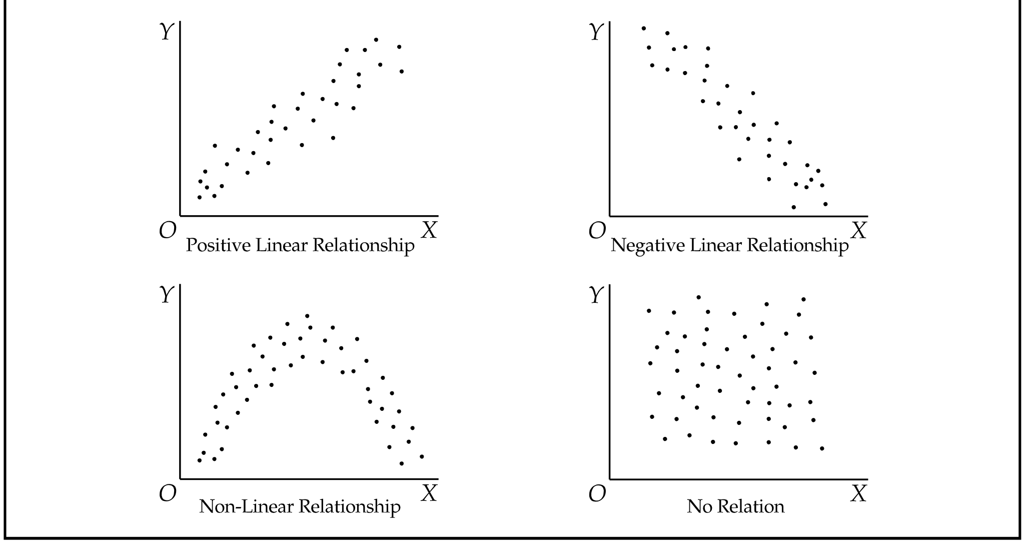
(iv) ***A high value of the correlation coefficient may be obtained due to sheer coincidence (or pure chance):*** This is another situation of spurious correlation. Given the data on any two variables, one may obtain a high value of correlation coefficient when in fact they do not have any relationship. For example, a high value of correlation coefficient may be obtained between the size of shoe and the income of persons of a locality.

6.2 Scatter Diagram

Let the bivariate data be denoted by (X_i, Y_i), where i = 1, 2 ...... n. In order to have some idea about the extent of association between variables X and Y, each pair (X_i, Y_i), i = 1, 2......n, is plotted on a graph. The diagram, thus obtained, is called a Scatter Diagram.

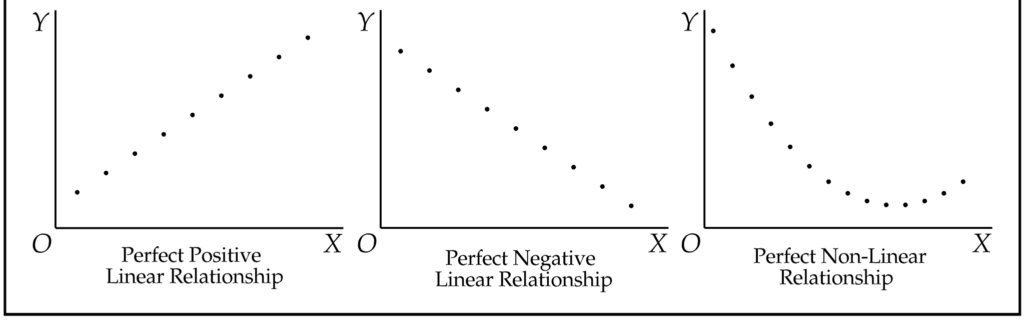
Each pair of values (X_i, Y_i) is denoted by a point on the graph. The set of such points (also known as dots of the diagram) may cluster around a straight line or a curve or may not show any tendency of association. Various possible situations are shown with the help of given diagrams:

Figure 6.1



If all the points or dots lie exactly on a straight line or a curve, the association between the variables is said to be perfect. This is shown below:

Figure 6.2



A scatter diagram of the data helps in having a visual idea about the nature of association between two variables. If the points cluster along a straight line, the association between variables is linear. Further, if the points cluster along a curve, the corresponding association is non-linear or curvilinear. Finally, if the points neither cluster along a straight line nor along a curve, there is absence of any association between the variables.

It is also obvious from the above figure that when low (high) values of X are associated with low (high) value of Y, the association between them is said to be positive. Contrary to this, when low (high) values of X are associated with high (low) values of Y, the association between them is said to be negative.

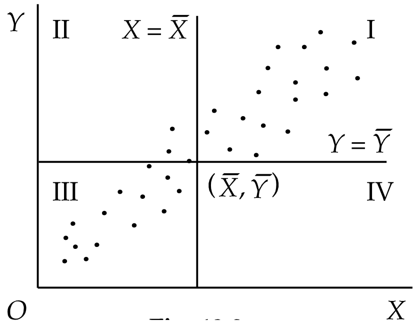
This chapter deals only with linear association between the two variables X and Y. We shall measure the degree of linear association by the Karl Pearson's formula for the coefficient of linear correlation.

6.3 Karl Pearson's Coefficient of Linear Correlation

Let us assume, again, that we have data on two variables X and Y denoted by the pairs (X_i, Y_i), i = 1,2, ... n. Further, let the scatter diagram of the data be as shown in figure 22.3.

Let $\overline{X}$ and $\overline{Y}$ be the arithmetic means of X and Y respectively. Draw two lines $X = \overline{X}$ and $Y = \overline{Y}$ on the scatter diagram. These two lines, intersect at the point ($\overline{X}, \overline{Y}$) and are mutually perpendicular, divide the whole diagram into four parts, termed as I, II, III and IV quadrants, as shown.

Figure 6.3



As mentioned earlier, the correlation between X and Y will be positive if low (high) values of X are associated with low (high) values of Y. In terms of the above figure, we can say that when values of X that are greater (less) than $\overline{X}$ are generally associated with values of Y that are greater (less) than $\overline{Y}$, the correlation between X and Y will be positive. This implies that there will be a general tendency of points to concentrate in I and III quadrants. Similarly, when correlation between X and Y is negative, the point of the scatter diagram will have a general tendency to concentrate in II and IV quadrants.

Further, if we consider deviations of values from their means, i.e., $\left(X_i - \overline{X}\right)$ and $\left(Y_i - \overline{Y}\right)$, we note that:

(i) Both $\left(X_i - \overline{X}\right)$ and $\left(Y_i - \overline{Y}\right)$ will be positive for all points in quadrant I.

(ii) $\left(X_i - \overline{X}\right)$ will be negative and $\left(Y_i - \overline{Y}\right)$ will be positive for all points in quadrant II.

(iii) Both $\left(X_i - \overline{X}\right)$ and $\left(Y_i - \overline{Y}\right)$ will be negative for all points in quadrant III.

(iv) $\left(X_i - \overline{X}\right)$ will be positive and $\left(Y_i - \overline{Y}\right)$ will be negative for all points in quadrant IV.

It is obvious from the above that the product of deviations, i.e., $\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$ will be positive for points in quadrants I and III and negative for points in quadrants II and IV.

Since, for positive correlation, the points will tend to concentrate more in I and III quadrants than in II and IV, the sum of positive products of deviations will outweigh the sum of negative products of deviations. Thus, $\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$ will be positive for all the n observations.

Similarly, when correlation is negative, the points will tend to concentrate more in II and IV quadrants than in I and III. Thus, the sum of negative products of deviations will outweigh the sum of positive products and hence $\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$ will be negative for all the n observations.

Further, if there is no correlation, the sum of positive products of deviations will be equal to the sum of negative products of deviations such that $\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$ will be equal to zero.

Notes

On the basis of the above, we can consider $\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$ as an absolute measure of correlation. This measure, like other absolute measures of dispersion, skewness, etc., will depend upon (i) the number of observations and (ii) the units of measurements of the variables.

In order to avoid its dependence on the number of observations, we take its average, i.e., $\frac{1}{n}\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$. This term is called covariance in statistics and is denoted as Cov(X,Y).

To eliminate the effect of units of measurement of the variables, the covariance term is divided by the product of the standard deviation of X and the standard deviation of Y. The resulting expression is known as the Karl Pearson's coefficient of linear correlation or the product moment correlation coefficient or simply the coefficient of correlation, between X and Y.

$$r_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \qquad \text{.... (1)}$$

or $$r_{XY} = \frac{\frac{1}{n}\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\frac{1}{n}\sum\left(X_i - \bar{X}\right)^2}\sqrt{\frac{1}{n}\sum\left(Y_i - \bar{Y}\right)^2}} \qquad \text{.... (2)}$$

Cancelling $\frac{1}{n}$ from the numerator and the denominator, we get

$$r_{XY} = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum\left(X_i - \bar{X}\right)^2}\sqrt{\sum\left(Y_i - \bar{Y}\right)^2}} \qquad \text{.... (3)}$$

Consider $\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = \sum\left(X_i - \bar{X}\right)Y_i - \bar{Y}\sum\left(X_i - \bar{X}\right)$

$$= \sum X_i Y_i - \bar{X}\sum Y_i \quad \text{(second term is zero)}$$

$$= \sum X_i Y_i - n\bar{X}\bar{Y} \quad \left(\sum Y_i = n\bar{Y}\right)$$

Similarly we can write $\sum\left(X_i - \bar{X}\right)^2 = \sum X_i^2 - n\bar{X}^2$

and $\sum\left(Y_i - \bar{Y}\right)^2 = \sum Y_i^2 - n\bar{Y}^2$

Substituting these values in equation (3), we have

$$r_{XY} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left[\sum X_i^2 - n\bar{X}^2\right]}\sqrt{\left[\sum Y_i^2 - n\bar{Y}^2\right]}} \qquad \text{.... (4)}$$

$$r_{XY} = \frac{\sum X_i Y_i - n \cdot \frac{\sum X_i}{n} \times \frac{\sum Y_i}{n}}{\sqrt{\sum X_i^2 - n\left(\frac{\sum X_i}{n}\right)^2}\sqrt{\sum Y_i^2 - n\left(\frac{\sum Y_i}{n}\right)^2}}$$

$$= \frac{\sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}}{\sqrt{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}\sqrt{\sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}}} \qquad \text{.... (5)}$$

On multiplication of numerator and denominator by n, we can write

$$r_{XY} = \frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{\sqrt{n\sum X_i^2 - \left(\sum X_i\right)^2}\sqrt{n\sum Y_i^2 - \left(\sum Y_i\right)^2}} \qquad \text{.... (6)}$$

Further, if we assume $x_i = X_i - \overline{X}$ and $y_i = Y_i - \overline{Y}$, equation (2), given above, can be written as

$$r_{XY} = \frac{\frac{1}{n}\sum x_i y_i}{\sqrt{\frac{1}{n}\sum x_i^2}\sqrt{\frac{1}{n}\sum y_i^2}} \qquad \text{.... (7)}$$

$$\text{or} \quad r_{XY} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \qquad \text{.... (8)}$$

$$\text{or} \quad r_{XY} = \frac{1}{n}\frac{\sum x_i y_i}{\sigma_x \sigma_y} \qquad \text{.... (9)}$$

Equations (5) or (6) are often used for the calculation of correlation from raw data, while the use of the remaining equations depends upon the forms in which the data are available. For example, if standard deviations of X and Y are given, equation (9) may be appropriate.

Example 1: Calculate the Karl Pearson's coefficient of correlation from the following pairs of values :

Values of X :	12	9	8	10	11	13	7
Values of Y :	14	8	6	9	11	12	3

Solution.

The formula for Karl Pearson's coefficient of correlation is

$$r_{XY} = \frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{\sqrt{n\sum X_i^2 - \left(\sum X_i\right)^2}\sqrt{n\sum Y_i^2 - \left(\sum Y_i\right)^2}}$$

Notes

The values of different terms, given in the formula, are calculated from the following table :

X_i	Y_i	X_iY_i	X_i^2	Y_i^2
12	14	168	144	196
9	8	72	81	64
8	6	48	64	36
10	9	90	100	81
11	11	121	121	121
13	12	156	169	144
7	3	21	49	9
70	63	676	728	651

Here n = 7 (no. of pairs of observations)

$$r_{XY} = \frac{7\times 676 - 70\times 63}{\sqrt{7\times 728-(70)^2}\sqrt{7\times 651-(63)^2}} = 0.949$$

Example 2: Calculate the Karl Pearson's coefficient of correlation between X and Y from the following data:

No. of pairs of observations n = 8, $\sum\left(X_i-\overline{X}\right)^2$ = 184, $\sum\left(Y_i-\overline{Y}\right)^2$ =148,

$\sum\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)$ = 164, $\overline{X} = 11$ and $\overline{Y} = 10$

Solution.

Using the formula, $r_{XY} = \dfrac{\sum\left(X-\overline{X}_i\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum\left(X_i-\overline{X}\right)^2}\sqrt{\sum\left(Y_i-\overline{Y}\right)^2}}$, we get

$$r_{XY} = \frac{164}{\sqrt{184}\sqrt{148}} = 0.99$$

Example 3:

(a) The covariance between the length and weight of five items is 6 and their standard deviations are 2.45 and 2.61 respectively. Find the coefficient of correlation between length and weight.

(b) The Karl Pearson's coefficient of correlation and covariance between two variables X and Y is – 0.85 and – 15 respectively. If variance of Y is 9, find the standard deviation of X.

Solution.

(a) Substituting the given values in formula (1) for correlation, we get

$$r_{XY} = \frac{6}{2.45\times 2.61} = 0.94$$

(b) Substituting the given values in the formula of correlation, we get

$$-0.85 = \frac{-15}{\sigma_X \times 3} \quad \text{or} \quad s_X = 5.88$$

6.4 Properties of Coefficient of Correlation

1. The coefficient of correlation is independent of the change of origin and scale of measurements.

   In order to prove this property, we change origin and scale of both the variables X and Y.

   Let $u_i = \frac{X_i - A}{h}$ and $v_i = \frac{Y_i - B}{k}$, where the constants A and B refer to change of origin and the constants h and k refer to change of scale. We can write

   $$X_i = A + hu_i, \quad \therefore \quad \bar{X} = A + h\bar{u}$$

   Thus, we have $X_i - \bar{X} = A + hu_i - A - h\bar{u} = h(u_i - \bar{u})$

   Similarly, $Y_i = B + kv_i, \quad \therefore \quad \bar{Y} = B + k\bar{v}$

   Thus, $Y_i - \bar{Y} = B + kv_i - B - k\bar{v} = k(v_i - \bar{v})$

   The formula for the coefficient of correlation between X and Y is

   $$r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

   Substituting the values of $(X_i - \bar{X})$ and $(Y_i - \bar{Y})$, we get

   $$r_{XY} = \frac{\sum h(u_i - \bar{u})k(v_i - \bar{v})}{\sqrt{\sum h^2(u_i - \bar{u})^2}\sqrt{\sum k^2(v_i - \bar{v})^2}} = \frac{\sum(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum(u_i - \bar{u})^2}\sqrt{\sum(v_i - \bar{v})^2}}$$

   $$\therefore \quad r_{XY} = r_{uv}$$

   This shows that correlation between X and Y is equal to correlation between u and v, where u and v are the variables obtained by change of origin and scale of the variables X and Y respectively.

   This property is very useful in the simplification of computations of correlation. On the basis of this property, we can write a short-cut formula for the computation of r_{XY} :

   $$r_{XY} = \frac{n\sum u_i v_i - (\sum u_i)(\sum v_i)}{\sqrt{n\sum u_i^2 - (\sum u_i)^2}\sqrt{n\sum v_i^2 - (\sum v_i)^2}} \qquad \text{.... (10)}$$

2. The coefficient of correlation lies between - 1 and + 1.

   To prove this property, we define

   $$x'_i = \frac{X_i - \bar{X}}{\sigma_X} \quad \text{and} \quad y'_i = \frac{Y_i - \bar{Y}}{\sigma_Y}$$

$$\therefore \qquad x'^2_i = \frac{\left(X_i - \bar{X}\right)^2}{\sigma_X^2} \text{ and } y'^2_i = \frac{\left(Y_i - \bar{Y}\right)^2}{\sigma_Y^2}$$

or
$$\sum x'^2_i = \frac{\sum\left(X_i - \bar{X}\right)^2}{\sigma_X^2} \text{ and } \sum y'^2_i = \frac{\sum\left(Y_i - \bar{Y}\right)^2}{\sigma_Y^2}$$

From these summations we can write $\sum x'^2_i = \sum y'^2_i = n$

$$\text{Also, } r = \frac{\frac{1}{n}\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sigma_X \sigma_Y} = \frac{1}{n}\cdot\sum\left(\frac{X_i - \bar{X}}{\sigma_X}\right)\left(\frac{Y_i - Y}{\sigma_Y}\right) = \frac{1}{n}\sum x'_i\, y'_i$$

Consider the sum $x_i' + y_i'$. The square of this sum is always a non-negative number, i.e., $(x_i' + y_i')^2$ ³ 0.

Taking sum over all the observations and dividing by n, we get

$$\frac{1}{n}\sum\left(x'_i + y'_i\right)^2 \geq 0 \quad \text{or} \quad \frac{1}{n}\sum\left(x'^2_i + y'^2_i + 2x'_i\, y'_i\right) \geq 0$$

or
$$\frac{1}{n}\sum x'^2_i + \frac{1}{n}\sum y'^2_i + \frac{2}{n}\sum x'_i\, y'_i \geq 0$$

or $\quad 1 + 1 + 2r \geq 0$ or $2 + 2r \geq 0$ or $r \geq -1$ .... (11)

Further, consider the difference $x_i' - y_i'$. The square of this difference is also non-negative, i.e., $(x_i' - y_i')^2$ ³ 0.

Taking sum over all the observations and dividing by n, we get

$$\frac{1}{n}\sum\left(x'_i - y'_i\right)^2 \geq 0 \quad \text{or} \quad \frac{1}{n}\sum\left(x'^2_i + y'^2_i - 2x'_i\, y'_i\right) \geq 0$$

or
$$\frac{1}{n}\sum x'^2_i + \frac{1}{n}\sum y'^2_i - \frac{2}{n}\sum x'_i\, y'_i \geq 0$$

or $\quad 1 + 1 - 2r \geq 0$ or $2 - 2r \geq 0$ or $r \leq 1$ .... (12)

Combining the inequalities (11) and (12), we get $-1 \leq r \leq 1$. Hence r lies between -1 and +1.

3. If X and Y are independent they are uncorrelated, but the converse is not true.

   If X and Y are independent, it implies that they do not reveal any tendency of simultaneous movement either in same or in opposite directions. In terms of figure 12.3, the dots of the scatter diagram will be uniformly spread in all the four quadrants. Therefore, $\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$ or Cov(X, Y) will be equal to zero and hence, $r_{XY} = 0$. Thus, if X and Y are independent, they are uncorrelated.

   The converse of this property implies that if $r_{XY} = 0$, then X and Y may not necessarily be independent. To prove this, we consider the following data :

X	1	2	3	4	5	6	7
Y	9	4	1	0	1	4	9

Here SX_i = 28, SY_i = 28 and SX_iY_i = 112.

$$\therefore \ \operatorname{Cov}(X,Y)=\frac{1}{n}\left[\sum X_iY_i-\frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}\right]=\frac{1}{7}\left[112-\frac{28\times 28}{7}\right]=0 \text{ . Thus, } r_{XY}=0$$

A close examination of the given data would reveal that although r_{XY} = 0, but X and Y are not independent. In fact they are related by the mathematical relation Y = (X - 4)^2.

Remarks: This property points our attention to the fact that r_{XY} is only a measure of the degree of linear association between X and Y. If the association is non-linear, the computed value of r_{XY} is no longer a measure of the degree of association between the two variables.

Example 4:

Calculate the Karl Pearson's coefficient of correlation from the following data:

Height of fathers (inches)	:	66	68	69	72	65	59	62	67	61	71
Height of sons (inches)	:	65	64	67	69	64	60	59	68	60	64

Solution.

Note: When there is no common factor, we can take h = k = 1 and define $u_i = X_i$ - A and $v_i = Y_i$ - B.

Calculation of r

Height of fathers (X_i)	*Height of sons* (Y_i)	$u_i = X_i - 65$	$v_i = Y_i - 64$	u_iv_i	u_i^2	v_i^2
66	65	1	1	1	1	1
68	64	3	0	0	9	0
69	67	4	3	12	16	9
72	69	7	5	35	49	25
65	64	0	0	0	0	0
59	60	- 6	- 4	24	36	16
62	59	- 3	- 5	15	9	25
67	68	2	4	8	4	16
61	60	- 4	- 4	16	16	16
71	64	6	0	0	36	0
Total		10	0	111	176	108

Here n = 10. Using formula (10) for correlation, we get

$$=\frac{10\times 111-10\times 0}{\sqrt{10\times 176-(10)^2}\sqrt{10\times 108-0^2}}=0.83$$

Example 5:

(a) Calculate the Karl Pearson's coefficient of correlation from the following data:

(i) Sum of deviations of X values = 5

(ii) Sum of deviations of Y values = 4

(iii) Sum of squares of deviations of X values = 40

(iv) Sum of squares of deviations of Y values = 50

(v) Sum of the product of deviations of X and Y values = 32

(vi) No. of pairs of observations = 10

(b) Given the following, calculate the coefficient of correlation :

(i) Sum of squares of deviations of X values from mean = 136

(ii) Sum of squares of deviations of Y values from mean = 138

(iii) Sum of products of deviations of X and Y values from their means = 122.

Solution.

(a) Let $u_i = X_i - A$ and $v_i = Y_i - B$ be the deviations of X and Y values. We are given $Su_i = 5$, $Sv_i = 4$, $Su_i^2 = 40$, $Sv_i^2 = 50$, $Su_iv_i = 32$ and n = 10.

Substituting these values in formula (10), we get

$$r_{XY} = \frac{10 \times 32 - 5 \times 4}{\sqrt{10 \times 40 - 5^2}\sqrt{10 \times 50 - 4^2}} = 0.704$$

(b) Using formula (3) for correlation, we get $r = \frac{122}{\sqrt{136}\sqrt{138}} = 0.89$

Example 6: Calculate the coefficient of correlation between age group and rate of mortality from the following data:

Age group	:	0-20	20-40	40-60	60-80	80-100
Rate of Mortality	:	350	280	540	760	900

Solution.

Since class intervals are given for age, their mid-values shall be used for the calculation of r.

Table for calculation of r

Age group	*M.V.* (X)	*Rate of Mort.* (Y)	$u_i = \frac{X_i - 50}{20}$	$v_i = \frac{Y_i - 540}{10}$	u_iv_i	u_i^2	v_i^2
0 - 20	10	350	- 2	- 19	38	4	361
20 - 40	30	280	- 1	- 26	26	1	676
40 - 60	50	540	0	0	0	0	0
60 - 80	70	760	1	22	22	1	484
80 - 100	90	900	2	36	72	4	1296
Total			0	13	158	10	2817

Here n = 5. Using the formula (10) for correlation, we get

$$r_{XY} = \frac{5 \times 158 - 0 \times 13}{\sqrt{5 \times 10 - 0^2}\sqrt{5 \times 2817 - 13^2}} = 0.95$$

Example 7:

Deviations from assumed average of the two series are given below :

Deviations, X series : - 10, - 6, - 4, - 1, 0, + 2, + 1, + 5, + 7, + 11

Deviations, Y series : - 8, - 5, + 4, - 2, - 4, 0, + 2, 0, - 2, + 4

Find out Karl Pearson's coefficient of correlation.

Solution.

Here the values of $u_i = X_i - A$ and $v_i = X_i - B$ are given.

Table for calculation of r

u_i	- 10	- 6	- 4	- 1	0	2	1	5	7	11	5
v_i	- 8	- 5	4	- 2	- 4	0	2	0	- 2	4	- 11
$u_i v_i$	80	30	- 16	2	0	0	2	0	- 14	44	128
u_i^2	100	36	16	1	0	4	1	25	49	121	353
v_i^2	64	25	16	4	16	0	4	0	4	16	149

Here n = 10.

$$r_{XY} = \frac{10 \times 128 - 5 \times (-11)}{\sqrt{10 \times 353 - 5^2}\sqrt{10 \times 149 - 11^2}} = 0.609$$

Example 8:

From the following table, find the missing values and calculate the coefficient of correlation by Karl Pearson's method :

X :	6	2	10	4	?
Y :	9	11	?	8	7

Arithmetic means of X and Y series are 6 and 8 respectively.

Solution.

The missing value in X - series = 5 × 6 – (6 + 2 + 10 + 4) = 30 – 22 = 8

The missing value in Y - series = 5 × 8 – (9 + 11 + 8 + 7) = 40 – 35 = 5

Table for calculation of r

X	Y	$X - \overline{X}$	$(Y - \overline{Y})$	$(X - \overline{X})(Y - \overline{Y})$	$(X - \overline{X})^2$	$(Y - \overline{Y})^2$
6	9	0	1	0	0	1
2	11	- 4	3	- 12	16	9
10	5	4	- 3	- 12	16	9
4	8	- 2	0	0	4	0
8	7	2	- 1	- 2	4	1
Total				- 26	40	20

Using formula (3) for correlation, we get $r = \dfrac{-26}{\sqrt{40}\sqrt{20}} = -0.92$

Example 9:

Calculate Karl Pearson's coefficient of correlation for the following series :

Price (in Rs)	:	10	11	12	13	14	15	16	17	18	19
Demand (in kgs)	:	420	410	400	310	280	260	240	210	210	200

Solution.

Table for calculation of r

Price (X)	*Demand* (Y)	$u = X - 14$	$v = \frac{Y - 310}{10}$	uv	u^2	v^2
10	420	- 4	11	- 44	16	121
11	410	- 3	10	- 30	9	100
12	400	- 2	9	- 18	4	81
13	310	- 1	0	0	1	0
14	280	0	- 3	0	0	9
15	260	1	- 5	- 5	1	25
16	240	2	- 7	- 14	4	49
17	210	3	- 10	- 30	9	100
18	210	4	- 10	- 40	16	100
19	200	5	- 11	- 55	25	121
Total		5	- 16	- 236	85	706

$$r = \frac{-10 \times 236 + 5 \times 16}{\sqrt{10 \times 85 - 25}\sqrt{10 \times 706 - 256}} = -0.96$$

Example 10:

A computer while calculating the correlation coefficient between two variables, X and Y, obtained the following results :

$n = 25, \Sigma X = 125, \Sigma X^2 = 650, \Sigma Y = 100, \Sigma Y^2 = 460, \Sigma XY = 508.$

It was, however, discovered later at the time of checking that it had copied down two pairs of observations as $\begin{array}{c|c} X & Y \\ \hline 6 & 14 \\ 8 & 6 \end{array}$ in place of the correct pairs $\begin{array}{c|c} X & Y \\ \hline 8 & 12 \\ 6 & 8 \end{array}$. Obtain the correct value of r.

Solution.

First we have to correct the values of ΣX, ΣX^2......etc.

Corrected $\Sigma X = 125 – (6 + 8) + (8 + 6) = 125$

Corrected $\Sigma X^2 = 650 – (36 + 64) + (64 + 36) = 650$

Corrected $\Sigma Y = 100 – (14 + 6) + (12 + 8) = 100$

Corrected $\Sigma Y^2 = 460 - (196 + 36) + (144 + 64) = 436$

Corrected SXY = 508 - (84 + 48) + (96 + 48) = 520

$$r = \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - (125)^2}\sqrt{25 \times 436 - (100)^2}} = 0.67$$

6.5 Probable Error of r

It is an old measure to test the significance of a particular value of r without the knowledge of test of hypothesis. Probable error of r, denoted by P.E.(r) is 0.6745 times its standard error. The value 0.6745 is obtained from the fact that in a normal distribution $\bar{r} \pm 0.6745 \times \text{S.E.}$ covers 50% of the total distribution.

According to Horace Secrist "The probable error of correlation coefficient is an amount which if added to and subtracted from the mean correlation coefficient, gives limits within which the chances are even that a coefficient of correlation from a series selected at random will fall."

Since standard error of r, i.e., $\text{S.E.}_r = \frac{1-r^2}{\sqrt{n}}, \quad \therefore \text{P.E.}(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$

6.5.1 Uses of P.E.(r)

(i) It can be used to specify the limits of population correlation coefficient ρ (rho) which are defined as r – P.E.(r) ≤ r ≤ r + P.E.(r), where ρ denotes correlation coefficient in population and r denotes correlation coefficient in sample.

(ii) It can be used to test the significance of an observed value of r without the knowledge of test of hypothesis. By convention, the rules are:

(a) If |r| < 6 P.E.(r), then correlation is not significant and this may be treated as a situation of no correlation between the two variables.

(b) If |r|> 6 P.E.(r), then correlation is significant and this implies presence of a strong correlation between the two variables.

(c) If correlation coefficient is greater than 0.3 and probable error is relatively small, the correlation coefficient should be considered as significant.

Example 11: Find out correlation between age and playing habit from the following information and also its probable error.

Age	:	15	16	17	18	19	20
No. of Students	:	250	200	150	120	100	80
Regular Players	:	200	150	90	48	30	12

Solution.

Let X denote age, p the number of regular players and q the number of students. Playing habit, denoted by Y, is measured as a percentage of regular players in an age group, i.e., Y = (p/q)×100.

Table for calculation of r

X	q	p	Y	$u = X - 17$	$v = Y - 40$	uv	u^2	v^2
15	250	200	80	- 2	40	- 80	4	1600
16	200	150	75	- 1	35	- 35	1	1225
17	150	90	60	0	20	0	0	400
18	120	48	40	1	0	0	1	0
19	100	30	30	2	- 10	- 20	4	100
20	80	12	15	3	- 25	- 75	9	625
Total				3	60	- 210	19	3950

$$r_{XY} = \frac{-6\times 210 - 3\times 60}{\sqrt{6\times 19 - 9}\sqrt{6\times 3950 - 3600}} = -0.99$$

Probable error of r, i.e., $\text{P.E.}(r) = 0.6745\times\frac{\left[1-(0.99)^2\right]}{\sqrt{6}} = 0.0055$

Example 12:

Test the significance of correlation for the values based on the number of observations (i) 10, and (ii) 100 and r = 0.4 and 0.9.

Solution.

(i) (a) Consider n = 10 and r = 0.4. Thus, $\text{P.E.}(r) = 0.6745\times\frac{1-0.4^2}{\sqrt{10}} = 0.179$ and 6 P.E. = 6 × 0.179 = 1.074. Since | r | < 6 P.E., r is not significant.

(i) (b) Take n = 10 and r = 0.9. Thus, $\text{P.E.} = 0.6745\times\frac{1-0.9^2}{\sqrt{10}} = 0.041$ and 6 P.E. = 6 × 0.041 = 0.246. Since | r | > 6 P.E., r is highly significant.

(ii) (a) Take n = 100 and r = 0.4. Thus, $6\text{P.E.} = 6\times 0.6745\frac{\left(1-0.4^2\right)}{\sqrt{100}} = 0.34$

Since | r | > 6 P.E., r is significant.

(ii) (b) Take n = 100 and r = 0.9. Thus, $6\text{P.E.} = 6\times 0.6745\frac{\left(1-0.9^2\right)}{\sqrt{100}} = 0.077$

Since | r | > 6 P.E., r is significant.

6.6 Correlation in a Bivariate Frequency Distribution

Let the two variables X and Y take respective values X_i, i = 1, 2, ... m and Y_j, j = 1, 2, ... n. These values, taken together, will make m´n pairs (X_i, Y_j). Let f_{ij} be the frequency of this pair. This frequency distribution can be presented in a tabular form as given below :

$Y\rightarrow$ $X\downarrow$	Y_1	Y_2	$\cdots$	Y_j	$\cdots$	Y_n	*Total*
X_1	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	f_{1n}	f_1
X_2	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	f_{2n}	f_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
X_i	f_{i1}	f_{i2}		f_{ij}		f_{in}	f_i
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
X_m	f_{m1}	f_{m2}	$\cdots$	f_{mj}	$\cdots$	f_{mn}	f_m
Total	f'_1	f'_2	$\cdots$	f'_j	$\cdots$	f'_n	N

Here $\sum\sum f_{ij} = \sum f_i = \sum f'_j = N$ (the total frequency).

The formula for correlation can be written on the basis of the formula discussed earlier.

$$r_{XY} = \frac{N\sum\sum f_{ij}X_iY_j - \left(\sum f_iX_i\right)\left(\sum f_j'Y_j\right)}{\sqrt{N\sum f_iX_i^2 - \left(\sum f_iX_i\right)^2}\sqrt{N\sum f_j'Y_j^2 - \left(\sum f_j'Y_j\right)^2}}$$

When we make changes of origin and scale by making the transformations $u_i = \frac{X_i - A}{h}$ and $v_j = \frac{Y_j - B}{k}$, then we can write

$$r_{XY} = \frac{N\sum\sum f_{ij}u_iv_j - \left(\sum f_iu_i\right)\left(\sum f_j'v_j\right)}{\sqrt{N\sum f_iu_i^2 - \left(\sum f_iu_i\right)^2}\sqrt{N\sum f_j'v_j^2 - \left(\sum f_j'v_j\right)^2}}$$

Example 13:

Calculate Karl Pearson's coefficient of correlation from the following data :

Age(yrs) → Marks ↓	18	19	20	21	22
20 - 25	3	2			
15 - 20		5	4		
10 - 15			7	10	
5 - 10				3	2
0 - 5					4

Solution.

Let X_i denote the mid-value of the class interval of marks. Various values of X_i can be written as 22.5, 17.5, 12.5, 7.5 and 2.5.

Further, let $u_i = (X_i - 12.5) \div 5$. Various values of u_i would be 2, 1, 0, - 1 and - 2.

Similarly, let Y_j denote age. Various values of Y_j are 18, 19, 20, 21 and 22.

Assuming $v_j = Y_j - 20$, various values of v_j would be - 2, - 1, 0, 1 and 2.

We shall use the values of u_i and v_j in the computation of r.

Table for Calculation of r

$u_i \backslash v_j$	-2	-1	0	1	2	f_i	f_iu_i	$f_iu_i^2$	$f_{ij}u_iv_j$
2	3 $\boxed{-12}$	2 $\boxed{-4}$	...	...	...	5	10	20	–16
1	...	5 $\boxed{-5}$	4 $\boxed{0}$	...	...	9	9	9	–5
0	...	...	7 $\boxed{0}$	10 $\boxed{0}$	...	17	0	0	0
–1	...	...	...	3 $\boxed{-3}$	2 $\boxed{-4}$	5	–5	5	–7
–2	...	...	...	...	4 $\boxed{-16}$	4	–8	16	–16
f_j'	3	7	11	13	6	40	6	50	–44
$f_j'v_j$	–6	–7	0	13	12	12			
$f_j'v_j^2$	12	7	0	13	24	56			

Substituting various values in the formula for r, we get

$$r = \frac{40 \times (-44) - 6 \times 12}{\sqrt{40 \times 50 - 36}\sqrt{40 \times 56 - 144}} = \frac{-1832}{\sqrt{1964}\sqrt{2096}} = -0.903$$

Examples 14:

Given the following data, compute the coefficient of correlation r, between X and Y.

Y → X ↓	30-50	50-70	70-90	Total
0-5	10	6	2	18
5-10	3	5	4	12
10-15	4	7	9	20
Total	17	18	15	50

Solution.

Note: Instead of doing the computation work in a single table, as done in example 13, it can be split into the following steps:

Taking mid-values of the class intervals, we have

Mid-values (X) : 2.5 7.5 12.5

Mid-values (Y) : 40 60 80

$$\text{Let } u_i = \frac{X_i - 7.5}{5} \text{ and } v_i = \frac{Y_i - 60}{20}$$

∴ various u values are : - 1 0 1

and various v values are : - 1 0 1

(i) Calculation of $SSf_{ij}u_iv_j$

u_i \ v_j	−1	0	1	Total
−1	$\boxed{10}$ 10	$\boxed{0}$ 6	$\boxed{-2}$ 2	8
0	$\boxed{0}$ 3	$\boxed{0}$ 5	$\boxed{0}$ 4	0
1	$\boxed{-4}$ 4	$\boxed{0}$ 7	$\boxed{9}$ 9	5
Total	6	0	7	13

\ $Sf_{ij}u_iv_j = 13$

(ii) Calculation of Sf_iu_i and $Sf_iu_i^2$

u_i	f_i	f_iu_i	$f_iu_i^2$
−1	18	−18	18
0	12	0	0
1	20	20	20
Total	50	2	38

(iii) Calculation of $Sf_j'v_j$ and $Sf_j'v_j^2$

v_j	f_j'	$f_j'v_j$	$f_j'v_j^2$
−1	17	−17	17
0	18	0	0
1	15	15	15
Total	50	−2	32

Substituting these values in the formula of r, we have

$$r = \frac{50 \times 13 - 2 \times (-2)}{\sqrt{50 \times 38 - 4}\sqrt{50 \times 32 - 4}} = \frac{654}{\sqrt{1896}\sqrt{1596}} = 0.376$$

6.7 Merits and Limitations of Coefficient of Correlation

The only merit of Karl Pearson's coefficient of correlation is that it is the most popular method for expressing the degree and direction of linear association between the two variables in terms of a pure number, independent of units of the variables. This measure, however, suffers from certain limitations, given below :

1. Coefficient of correlation r does not give any idea about the existence of cause and effect relationship between the variables. It is possible that a high value of r is obtained although none of them seem to be directly affecting the other. Hence, any interpretation of r should be done very carefully.

2. It is only a measure of the degree of linear relationship between two variables. If the relationship is not linear, the calculation of r does not have any meaning.

3. Its value is unduly affected by extreme items.

4. If the data are not uniformly spread in the relevant quadrants (see - Fig 12.3), the value of r may give a misleading interpretation of the degree of relationship between the two variables. For example, if there are some values having concentration around a point in first quadrant and there is similar type of concentration in third quadrant, the value of r will be very high although there may be no linear relation between the variables.

5. As compared with other methods, to be discussed later in this chapter, the computations of r are cumbersome and time consuming.

6.8 Spearman's Rank Correlation

This is a crude method of computing correlation between two characteristics. In this method, various items are assigned ranks according to the two characteristics and a correlation is computed between these ranks. This method is often used in the following circumstances:

(i) When the quantitative measurements of the characteristics are not possible, e.g., the results of a beauty contest where various individuals can only be ranked.

(ii) Even when the characteristics is measurable, it is desirable to avoid such measurements due to shortage of time, money, complexities of calculations due to large data, etc.

(iii) When the given data consist of some extreme observations, the value of Karl Pearson's coefficient is likely to be unduly affected. In such a situation the computation of the rank correlation is preferred because it will give less importance to the extreme observations.

(iv) It is used as a measure of the degree of association in situations where the nature of population, from which data are collected, is not known.

The coefficient of correlation obtained on the basis of ranks is called 'Spearman's Rank Correlation' or simply the 'Rank Correlation'. This correlation is denoted by ρ (rho).

Let X_i be the rank of i th individual according to the characteristics X and Y_i be its rank according to the characteristics Y. If there are n individuals, there would be n pairs of ranks (X_i, Y_i), i = 1, 2, ... n. We assume here that there are no ties, i.e., no two or more individuals are tied to a particular rank. Thus, $X_i^{'s}$ and $Y_i^{'s}$ are simply integers from 1 to n, appearing in any order.

The means of X and Y, i.e., $\bar{X} = \bar{Y} = \frac{1+2+\cdots\cdots n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$. Also,

$$\sigma_X^2 = \sigma_Y^2 = \frac{1}{n}[1^2 + 2^2 + \cdots + n^2] - \frac{(n+1)^2}{4} = \frac{1}{n}\left[\frac{n(n+1)(2n+1)}{6}\right] - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}$$

Let d_i be the difference in ranks of the i th individual, i.e.,

$$d_i = X_i - Y_i = \left(X_i - \bar{X}\right) - \left(Y_i - \bar{Y}\right) \qquad \left(\because \bar{X} = \bar{Y}\right)$$

Squaring both sides and taking sum over all the observations, we get

$$\sum d_i^2 = \sum\left[\left(X_i - \bar{X}\right) - \left(Y_i - \bar{Y}\right)\right]^2$$

$$= \sum\left(X_i - \bar{X}\right)^2 + \sum\left(Y_i - \bar{Y}\right)^2 - 2\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

Dividing both sides by n, we get

$$\frac{1}{n}\sum d_i^2 = \frac{1}{n}\sum\left(X_i - \bar{X}\right)^2 + \frac{1}{n}\sum\left(Y_i - \bar{Y}\right)^2 - \frac{2}{n}\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

$$= \sigma_X^2 + \sigma_Y^2 - 2Cov(X,Y) = 2\sigma_X^2 - 2Cov(X,Y) \qquad \left(\because \sigma_X^2 = \sigma_Y^2\right)$$

$$= 2\sigma_X^2 - 2\rho\sigma_X\sigma_Y = 2\sigma_X^2 - 2\rho\sigma_X^2 = 2\sigma_X^2(1-\rho) \qquad \left(\because \rho = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}\right)$$

From this, we can write $1-\rho=\frac{1}{n}\times\frac{\sum d_i^2}{2\sigma_X^2}$

$$or\ \ \rho=1-\frac{1}{n}\times\frac{\sum d_i^2}{2\sigma_X^2}=1-\frac{1}{n}\times\frac{\sum d_i^2}{2}\times\frac{12}{n^2-1}=1-\frac{6\sum d_i^2}{n\left(n^2-1\right)}$$

Note: This formula is not applicable in case of a bivariate frequency distribution.

Example 15:

The following table gives the marks obtained by 10 students in commerce and statistics. Calculate the rank correlation.

Marks in Statistics	:	35	90	70	40	95	45	60	85	80	50
Marks in Commerce	:	45	70	65	30	90	40	50	75	85	60

Solution.

Calculation Table

Marks in Statistics	*Marks in Commerce*	*Rank of Marks in*		$d_i=X_i-\overline{Y}_i$	d_i^2
		Statistics X	*Commerce Y*		
35	45	1	3	–2	4
90	70	9	7	2	4
70	65	6	6	0	0
40	30	2	1	1	1
95	90	10	10	0	0
45	40	3	2	1	1
60	50	5	4	1	1
85	75	8	8	0	0
80	85	7	9	–2	4
50	60	4	5	–1	1

From the above table, we have $\sum d_i^2=16$.

$\therefore$ Rank Correlation $\rho=\mathbf{1}-\frac{\mathbf{6}\sum \mathbf{d_i^2}}{\mathbf{n}\left(\mathbf{n^2}-\mathbf{1}\right)}=\mathbf{1}-\frac{\mathbf{6}\times\mathbf{16}}{\mathbf{10}\times\mathbf{99}}=\mathbf{0.903}$

6.9 Coefficient of Correlation by Concurrent Deviation Method

This is another simple method of obtaining a quick but crude idea of correlation between two variables. In this method, only direction of change in the concerned variables are noted by comparing a value from its preceding value. If the value is greater than its preceding value, it is indicated by a '+' sign; if less, it is indicated by a '-' sign and equal values are indicated by '=' sign. All the pairs having same signs, i.e., either both the deviations are positive or negative or have equal sign ('='), are known as concurrent deviations and are indicated by '+' sign in a separate column designated as 'concurrences'. The number of such concurrences is denoted by C. Similarly, the remaining pairs are marked by '-' sign in another column designated as 'disagreements'. The coefficient of correlation, denoted by r_C, is given by the formula $r_C=\pm\sqrt{\pm\left(\frac{2C-D}{D}\right)}$, where C denotes the number of concurrences and D (= number of observations - 1) is the number of pairs of deviation.

Note:

(i) The sign of r_C is taken to be equal to the sign of $\left(\frac{2C - D}{D}\right)$.

(ii) When $\left(\frac{2C-D}{D}\right)$ is negative, we make it positive for the purpose of taking its square root. However, the computed value will have a negative sign.

(iii) The sign of r_C will be positive when $\left(\frac{2C-D}{D}\right)$ is positive.

(iv) This method gives same weights to smaller as well as to the larger deviations.

(v) This method is suitable only for the study of short term fluctuations because it does not take into account the changes in magnitudes of the values.

Example 18:

The following table gives the marks obtained by 11 students of a class in micro and macro-economics papers. Calculate the coefficient of correlation by concurrent deviation method.

Roll No.	:	1	2	3	4	5	6	7	8	9	10	11
Marks in Micro - economics	:	80	45	55	56	58	60	65	68	70	75	85
Marks in Macro - economics	:	82	56	50	48	60	62	64	65	70	74	90

Solution.

Let D_1 and D_2 denote deviations from the preceding marks in micro and macro economics respectively.

Calculation Table

Roll No.	*Marks in Micro - economics*	D_1	*Marks in Macro - economics*	D_2	*Concurr - ences*	*Disagree - ments*
1	80		82			
2	45	-	56	-	+	
3	55	+	50	-		-
4	56	+	48	-		-
5	58	+	60	+	+	
6	60	+	62	+	+	
7	65	+	64	+	+	
8	68	+	65	+	+	
9	70	+	70	+	+	
10	75	+	74	+	+	
11	85	+	90	+	+	
			Total		8	2

Here C = 8 and the no. of pairs of deviation D = 10.

Now, $\frac{2C-D}{D} = \frac{16-10}{10} = 0.6$ which is positive, $\therefore\ r_C = \sqrt{0.6} = 0.77$

This value indicates the presence of a very high positive correlation between the marks obtained in two papers.

Example 19:

Find out the coefficient of correlation by concurrent deviation method from the following information:

Number of pairs of deviations = 96

Number of concurrent deviations = 32

Solution.

We are given C = 32 and D = 96

Now, $\frac{2C - D}{D} = \frac{64 - 96}{96} = -\frac{1}{3}$ which is negative, $\therefore\ r_C = -\sqrt{\frac{1}{3}} = -0.577$

6.10 Summary

- ***One of the variable may be affecting the other:*** A correlation coefficient calculated from the data on quantity demanded and corresponding price of tea would only reveal that the degree of association between them is very high. It will not give us any idea about whether price is affecting demand of tea or vice-versa. In order to know this, we need to have some additional information apart from the study of correlation. For example if, on the basis of some additional information, we say that the price of tea affects its demand, then price will be the cause and quantity will be the effect. The causal variable is also termed as independent variable while the other variable is termed as dependent variable.
- ***The two variables may act upon each other:*** Cause and effect relation exists in this case also but it may be very difficult to find out which of the two variables is independent. For example, if we have data on price of wheat and its cost of production, the correlation between them may be very high because higher price of wheat may attract farmers to produce more wheat and more production of wheat may mean higher cost of production, assuming that it is an increasing cost industry. Further, the higher cost of production may in turn raise the price of wheat. For the purpose of determining a relationship between the two variables in such situations, we can take any one of them as independent variable.
- ***The two variables may be acted upon by the outside influences:*** In this case we might get a high value of correlation between the two variables, however, apparently no cause and effect type relation seems to exist between them. For example, the demands of the two commodities, say X and Y, may be positively correlated because the incomes of the consumers are rising. Coefficient of correlation obtained in such a situation is called a spurious or nonsense correlation.
- ***A high value of the correlation coefficient may be obtained due to sheer coincidence (or pure chance):*** This is another situation of spurious correlation. Given the data on any two variables, one may obtain a high value of correlation coefficient when in fact they do not have any relationship. For example, a high value of correlation coefficient may be obtained between the size of shoe and the income of persons of a locality.
- Let the bivariate data be denoted by (X_i, Y_i), where i = 1, 2 ...... n. In order to have some idea about the extent of association between variables X and Y, each pair (X_i, Y_i), i = 1, 2......n, is plotted on a graph. The diagram, thus obtained, is called a Scatter Diagram.

- Each pair of values (X_i, Y_i) is denoted by a point on the graph. The set of such points (also known as dots of the diagram) may cluster around a straight line or a curve or may not show any tendency of association.
- It can be used to specify the limits of population correlation coefficient ρ (rho) which are defined as $r - P.E.(r) \leq r \leq r + P.E.(r)$, where ρ denotes correlation coefficient in population and r denotes correlation coefficient in sample.
- It can be used to test the significance of an observed value of r without the knowledge of test of hypothesis. By convention, the rules are:
  - If $|r| < 6$ P.E.(r), then correlation is not significant and this may be treated as a situation of no correlation between the two variables.
  - If $|r| > 6$ P.E.(r), then correlation is significant and this implies presence of a strong correlation between the two variables.
  - If correlation coefficient is greater than 0.3 and probable error is relatively small, the correlation coefficient should be considered as significant.

6.11 Keywords

Correlation: It is an analysis of covariation between two or more variables.

Correlation Coefficient: It is a numerical measure of the degree of association between two or more variables.

Scatter diagram: A scatter diagram of the data helps in having a visual idea about the nature of association between two variables. If the points cluster along a straight line, the association between variables is linear.

6.12 Self Assessment

1. Fill in the blanks :

   (i) Coefficient of correlation is a measure of the strength of the .... relationship between two variables.

   (ii) Coefficient of correlation is .... of the change of origin and scale.

   (iii) Coefficient of correlation between sale of woolen garments and the day temperature is likely to be ....

   (iv) Coefficient of correlation lies between .... and ....

   (v) Correlation between number of accidents and number of babies born in different years is termed as .... correlation.

   (vi) If two variables X and Y are such that their difference (X – Y) is always equal to 25. The correlation between X and Y is .... and positive.

2. Examine the validity of the following statements giving necessary proofs and reasons for your answer:

   (i) If $r_{XY} = 0$, then X and Y are always independent.

   (ii) If the sum of squares of the difference in ranks of 8 pairs of observations is 126, then the rank correlation coefficient is 0.5.

(iii) If u + 3x = 5, 2y - v = 7 and r_{xy} = 0.12, then r_{uv} = 0.12.

(iv) If 2x - u = 8, y - 3v = 10 and r_{xy} = 0.8, then r_{uv} = 0.8.

(v) If $\sum d_i^2$ = 33 and n = 10 then *r* = 0.8.

6.13 Review Questions

1. (a) Define correlation between two variables. Distinguish between positive and negative correlation. Illustrate by using diagrams.

   (b) Define the concept of covariance. How do you interpret it?

2. Define correlation and discuss its significance in statistical analysis. Does it signify 'cause and effect' relationship between the two variables?

3. (a) What do you understand by the coefficient of linear correlation? Explain the significance and limitations of this measure in any statistical analysis.

   (b) Write down an expression for the Karl Pearson's coefficient of linear correlation. Why is it termed as the coefficient of linear correlation? Explain.

4. (a) Describe the method of obtaining the Karl Pearson's formula of coefficient of linear correlation. What do positive and negative values of this coefficient indicate?

   (b) Does a zero value of Karl Pearson's coefficient of correlation between two variables X and Y imply that X and Y are not related? Explain.

5. Define product moment coefficient of correlation. What are the advantages of the study of correlation?

6. Show that the coefficient of correlation, r, is independent of change of origin and scale.

7. Prove that the coefficient of correlation lies between - 1 and +1.

8. "If two variables are independent the correlation between them is zero, but the converse is not always true". Explain the meaning of this statement.

9. What is Spearman's rank correlation? What are the advantages of the coefficient of rank correlation over Karl Pearson's coefficient of correlation?

10. Distinguish between the Spearman's coefficient of rank correlation and Karl Pearson's coefficient of correlation. Explain the situations under which Spearman's coefficient of rank correlation can assume a maximum and a minimum value. Under what conditions will Spearman's formula and Karl Pearson's formula give equal results?

11. Explain the method of calculating coefficient of correlation by Concurrent Deviation Method.

12. Write short notes on:

    (i) Positive and negative correlation.

    (ii) Linear and non-linear correlation.

    (iii) Probable error of correlation.

    (iv) Scatter diagram.

13. Compute Karl Pearson's coefficient of correlation from the following data :

X :	8	11	15	10	12	16
Y :	6	9	11	7	9	12

14. Calculate Karl Pearson's coefficient of correlation between the marks obtained by 10 students in economics and statistics.

Roll No. :	1	2	3	4	5	6	7	8	9	10
Marks in eco. :	23	27	28	29	30	31	33	35	36	39
Marks in stat. :	18	22	23	24	25	26	28	29	30	32

15. Find Karl Pearson's coefficient of correlation from the following data and interpret its value.

Wages (Rs) :	100	101	103	102	100	99	97	98	96	95
Cost of Living (Rs) :	98	99	99	97	95	92	95	94	90	91

16. Find the coefficient of correlation between X and Y. Assume 69 and 112 as working origins for X and Y respectively.

X :	78	89	96	69	59	79	68	61
Y :	125	137	156	112	107	136	123	108

17. The distribution of population (in thousand) and blind persons according to various age groups is given in the following table. Find out correlation between age and blindness.

Age groups :	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Population :	100	60	40	36	24	11	6	3
No. of Blind :	55	40	40	40	36	22	18	15

18. Find out the coefficient of correlation from the following data :

X :	300	350	400	450	500	550	600	650	700
Y :	1600	1500	1400	1300	1200	1100	1000	900	800

19. Calculate the coefficient of correlation of the following figures relating to the consumption of fertiliser (in metric tonnes) and the output of food grains (in metric tonnes) in a district. Comment on your result.

Chemical Fertiliser used	Output of food grains	Chemical Fertiliser used	Output of food grains
100	1000	170	1360
110	1050	180	1420
120	1080	190	1500
130	1150	200	1600
140	1200	210	1650
150	1220	220	1650
160	1300	230	1650

Answers: Self Assessment

1. (i) linear (ii) independent (iii) negative (iv) – 1, + 1 (v) spurious or nonsense (vi) perfect
2. (i) invalid (ii) invalid (iii) invalid (iv) valid (v) valid.

6.14 Further Readings

Books

Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.

Unit 7: Regression Analysis

CONTENTS

Objectives

After studying this unit, you will be able to:

- Define Two Lines of Regression
- Explain Regression Coefficient in a Bivariate Frequency Distribution
- Discuss The Coefficient of Determination
- Describe Mean of the Estimated Values
- Explain Mean and Variance of 'ei' values

Introduction

If the coefficient of correlation calculated for bivariate data (X_i, Y_i), i = 1,2, ... n, is reasonably high and a cause and effect type of relation is also believed to be existing between them, the next logical step is to obtain a functional relation between these variables. This functional relation is known as regression equation in statistics. Since the coefficient of correlation is measure of the degree of linear association of the variables, we shall discuss only linear regression equation. This does not, however, imply the non-existence of non-linear regression equations.

The regression equations are useful for predicting the value of dependent variable for given value of the independent variable. As pointed out earlier, the nature of a regression equation is different from the nature of a mathematical equation, e.g., if $Y = 10 + 2X$ is a mathematical equation then it implies that Y is exactly equal to 20 when X = 5. However, if $Y = 10 + 2X$ is a regression equation, then Y = 20 is an average value of Y when X = 5.

The term regression was first introduced by Sir Francis Galton in 1877. In his study of the relationship between heights of fathers and sons, he found that tall fathers were likely to have tall sons and vice-versa. However, the mean height of sons of tall fathers was lower than the mean height of their fathers and the mean height of sons of short fathers was higher than the mean height of their fathers. In this way, a tendency of the human race to regress or to return to a normal height was observed. Sir Francis Galton referred this tendency of returning to the mean height of all men as regression in his research paper, "Regression towards mediocrity in hereditary stature". The term 'Regression', originated in this particular context, is now used in various fields of study, even though there may be no existence of any regressive tendency.

7.1 Two Lines of Regression

For a bivariate data (X_i, Y_i), i = 1,2, ... n, we can have either X or Y as independent variable. If X is independent variable then we can estimate the average values of Y for a given value of X. The relation used for such estimation is called regression of Y on X. If on the other hand Y is used for estimating the average values of X, the relation will be called regression of X on Y. For a bivariate data, there will always be two lines of regression. It will be shown later that these two lines are different, i.e., one cannot be derived from the other by mere transfer of terms, because the derivation of each line is dependent on a different set of assumptions.

7.1.1 Line of Regression of Y on X

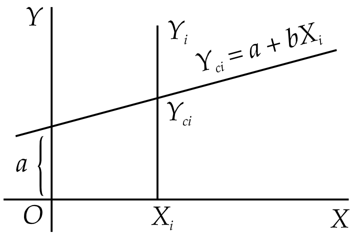
The general form of the line of regression of Y on X is $Y_{Ci} = a + bX_i$, where Y_{Ci} denotes the average or predicted or calculated value of Y for a given value of $X = X_i$. This line has two constants, a and b. The constant a is defined as the average value of Y when X = 0. Geometrically, it is the intercept of the line on Y- axis. Further, the constant b, gives the average rate of change of Y per unit change in X, is known as the regression coefficient.

The above line is known if the values of a and b are known. These values are estimated from the observed data (X_i, Y_i), i = 1,2, ... n.

Note It is important to distinguish between Y_{Ci} and Y_i. Where as Y_i is the observed value, Y_{Ci} is a value calculated from the regression equation.

Using the regression $Y_{Ci} = a + bX_i$, we can obtain Y_{C1}, Y_{C2}, ... Y_{Cn} corresponding to the X values X_1, X_2, ... X_n respectively. The difference between the observed and calculated value for a particular value of X say X_i is called error in estimation of the i th observation on the assumption of a particular line of regression. There will be similar type of errors for all the n observations. We denote by $e_i = Y_i - Y_{Ci}$ (i = 1,2,.....n), the error in estimation of the i th observation. As is obvious from figure 23.1, e_i will be positive if the observed point lies above the line and will be negative if the observed point lies below the line. Therefore, in order to obtain a figure of total error, $e_i^{'s}$ are squared and added. Let S denote the sum of squares of these errors, i.e., $S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - Y_{Ci}\right)^2$.

Figure 7.1



Note The regression line can, alternatively, be written as a deviation of Y_i from Y_{ci} i.e. $Y_i - Y_{ci} = e_i$ or $Y_i = Y_{ci} + e_i$ or $Y_i = a + bX_i + e_i$. The component $a + bX_i$ is known as the deterministic component and e_i is random component.

The value of S will be different for different lines of regression. A different line of regression means a different pair of constants a and b. Thus, S is a function of a and b. We want to find such values of a and b so that S is minimum. This method of finding the values of a and b is known as the Method of Least Squares.

Rewrite the above equation as S = S(Y_i - a - bX_i)^2 ($\because$ Y_{Ci} = a + bX_i).

The necessary conditions for minima of S are

(i) $\frac{\partial S}{\partial a} = 0$ and (ii) $\frac{\partial S}{\partial b} = 0$, where $\frac{\partial S}{\partial a}$ and $\frac{\partial S}{\partial b}$ are the partial derivatives of S w.r.t. a and b respectively.

$$\text{Now } \frac{\partial S}{\partial a} = -2\sum_{i=1}^{n}\left(Y_i - a - bX_i\right) = 0$$

$$or \quad \sum_{i=1}^{n}\left(Y_i - a - bX_i\right) = \sum_{i=1}^{n} Y_i - na - b\sum_{i=1}^{n} X_i = 0$$

$$or \quad \sum_{i=1}^{n} Y_i = na + b\sum_{i=1}^{n} X_i \qquad .... (1)$$

Also, $\frac{\partial S}{\partial b} = 2\sum_{i=1}^{n}\left(Y_i - a - bX_i\right)\left(-X_i\right) = 0$

$$or \quad -2\sum_{i=1}^{n}\left(X_iY_i - aX_i - bX_i^2\right) = \sum_{i=1}^{n}\left(X_iY_i - aX_i - bX_i^2\right) = 0$$

$$or \quad \sum_{i=1}^{n} X_iY_i - a\sum_{i=1}^{n} X_i - b\sum_{i=1}^{n} X_i^2 = 0$$

$$or \quad \sum_{i=1}^{n} X_iY_i = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 \qquad .... (2)$$

Equations (1) and (2) are a system of two simultaneous equations in two unknowns a and b, which can be solved for the values of these unknowns. These equations are also known as normal equations for the estimation of a and b. Substituting these values of a and b in the regression equation Y_{Ci} = a + bX_i, we get the estimated line of regression of Y on X.

Expressions for the Estimation of a and b.

Dividing both sides of the equation (1) by n, we have

$$\frac{\sum Y_i}{n} = \frac{na}{n} + \frac{b\sum X_i}{n} \quad \text{or} \quad \overline{Y} = a + b\overline{X} \qquad .... (3)$$

This shows that the line of regression Y_{Ci} = a + bX_i passes through the point $\left(\overline{X}, \overline{Y}\right)$.

From equation (3), we have $a = \overline{Y} - b\overline{X}$ .... (4)

Substituting this value of a in equation (2), we have

$$\sum X_iY_i = \left(\overline{Y} - b\overline{X}\right)\sum X_i + b\sum X_i^2$$

$$= \overline{Y}\sum X_i - b\overline{X}\sum X_i + b\sum X_i^2 = n\overline{X}\overline{Y} - b.n\overline{X}^2 + b\sum X_i^2$$

$$\text{or} \quad \sum X_iY_i - n\overline{X}\overline{Y} = b\left(\sum X_i^2 - n\overline{X}^2\right)$$

$$\text{or} \quad b = \frac{\sum X_iY_i - n\overline{X}\overline{Y}}{\sum X_i^2 - n\overline{X}^2} \qquad .... (5)$$

Also, $\sum X_iY_i - n\overline{X}\overline{Y} = \sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$ (See Chapter 12)

and $\sum X_i^2 - n\overline{X}^2 = \sum\left(X_i - \overline{X}\right)^2$

$$\therefore \quad b = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum\left(X_i - \overline{X}\right)^2} \qquad .... (6)$$

$$\text{or} \quad b = \frac{\sum x_iy_i}{\sum x_i^2} \qquad .... (7)$$

where x_i and y_i are deviations of values from their arithmetic mean.

Dividing numerator and denominator of equation (6) by n we have

$$b = \frac{\frac{1}{n}\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\frac{1}{n}\sum\left(X_i - \overline{X}\right)^2} = \frac{Cov(X,Y)}{\sigma_X^2} \qquad \text{.... (8)}$$

The expression for b, which is convenient for use in computational work, can be written from equation (5) is given below:

$$b = \frac{\sum X_i Y_i - n\frac{\sum X_i}{n}\cdot\frac{\sum Y_i}{n}}{\sum X_i^2 - n\left(\frac{\sum X_i}{n}\right)^2} = \frac{\sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}$$

Multiplying numerator and denominator by n, we have

$$b = \frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{n\sum X_i^2 - \left(\sum X_i\right)^2} \qquad \text{.... (9)}$$

To write the shortcut formula for b, we shall show that it is independent of change of origin but not of change of scale.

As in case of coefficient of correlation we define

$$u_i = \frac{X_i - A}{h} \quad \text{and} \quad v_i = \frac{Y_i - B}{k}$$

or $\quad X_i = A + hu_i \quad$ and $\quad Y_i = B + kv_i$

$\therefore \quad \overline{X} = A + h\overline{u} \quad$ and $\quad \overline{Y} = B + k\overline{v}$

also $\quad \left(X_i - \overline{X}\right) = h\left(u_i - \overline{u}\right) \quad$ and $\quad Y_i - \overline{Y} = k\left(v_i - \overline{v}\right)$

Substituting these values in equation (6), we have

$$b = \frac{hk\sum\left(u_i - \overline{u}\right)\left(v_i - \overline{v}\right)}{h^2\sum\left(u_i - \overline{u}\right)^2} = \frac{k\sum\left(u_i - \overline{u}\right)\left(v_i - \overline{v}\right)}{h\sum\left(u_i - \overline{u}\right)^2}$$

$$= \frac{k}{h}\left[\frac{n\sum u_i v_i - \left(\sum u_i\right)\left(\sum v_i\right)}{n\sum u_i^2 - \left(\sum u_i\right)^2}\right] \qquad \text{.... (10)}$$

(Note: if h = k they will cancel each other)

Consider equation (8), $b = \frac{Cov(X,Y)}{s_X^2}$

Writing $Cov(X,Y) = r.\sigma_X\sigma_Y$, we have $b = \frac{r.\sigma_X\sigma_Y}{\sigma_X^2} = r\cdot\frac{\sigma_Y}{\sigma_X}$

The line of regression of Y on X, i.e $Y_{Ci} = a + bX_i$ can also be written as

$$Y_{Ci} = \overline{Y} - b\overline{X} + bX_i \text{ or } Y_{Ci} - \overline{Y} = b\left(X_i - \overline{X}\right) \quad .... (11)$$

or $$\left(Y_{Ci} - \overline{Y}\right) = r \cdot \frac{\sigma_Y}{\sigma_X}\left(X_i - \overline{X}\right) \quad .... (12)$$

7.1.2 Line of Regression of X on Y

The general form of the line of regression of X on Y is $X_{Ci} = c + dY_i$, where X_{Ci} denotes the predicted or calculated or estimated value of X for a given value of $Y = Y_i$ and c and d are constants. d is known as the regression coefficient of regression of X on Y.

In this case, we have to calculate the value of c and d so that

$S' = \Sigma(X_i - X_{Ci})^2$ is minimised.

Figure 7.2

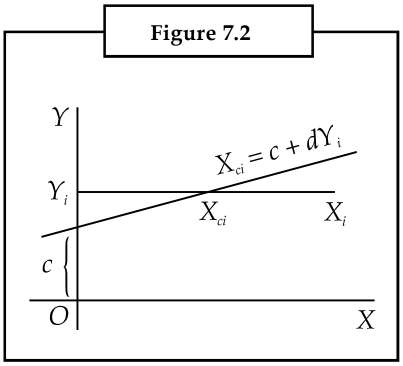
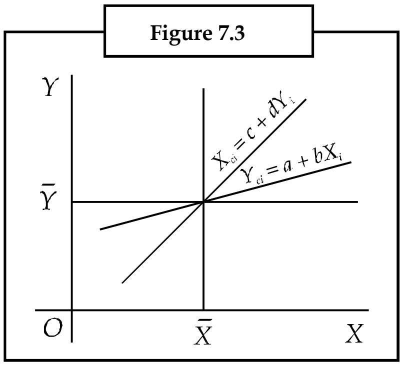


Figure 7.3



As in the previous section, the normal equations for the estimation of c and d are

$$\Sigma X_i = nc + d\Sigma Y_i \quad .... (13)$$

and $$\Sigma X_i Y_i = cSY_i + d\Sigma Y_i^2 \quad .... (14)$$

Dividing both sides of equation (13) by n, we have $\overline{X} = c + d\overline{Y}$.

This shows that the line of regression also passes through the point $\left(\overline{X}, \overline{Y}\right)$. Since both the lines of regression passes through the point $\left(\overline{X}, \overline{Y}\right)$, therefore $\left(\overline{X}, \overline{Y}\right)$ is their point of intersection as shown in Figure 23.3.

We can write $c = \overline{X} - d\overline{Y}$ .... (15)

As before, the various expressions for d can be directly written, as given below.

$$d = \frac{\sum X_i Y_i - n\overline{X}\overline{Y}}{\sum Y_i^2 - n\overline{Y}^2} \quad .... (16)$$

or $$d = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum\left(Y_i - \overline{Y}\right)^2} \quad .... (17)$$

or $$d = \frac{\sum x_i y_i}{\sum y_i^2} \quad .... (18)$$

$$= \frac{\frac{1}{n}\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\frac{1}{n}\sum\left(Y_i - \bar{Y}\right)^2} = \frac{Cov(X,Y)}{\sigma_Y^2} \quad \text{.... (19)}$$

Also $$d = \frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{n\sum Y_i^2 - \left(\sum Y_i\right)^2} \quad \text{.... (20)}$$

This expression is useful for calculating the value of d. Another short-cut formula for the calculation of d is given by

$$d = \frac{h}{k}\left[\frac{n\sum u_i v_i - \left(\sum u_i\right)\left(\sum v_i\right)}{n\sum v_i^2 - \left(\sum v_i\right)^2}\right] \quad \text{.... (21)}$$

where $u_i = \frac{X_i - A}{h}$ and $v_i = \frac{Y_i - B}{k}$

Consider equation (19)

$$d = \frac{Cov(X,Y)}{\sigma_Y^2} = \frac{r\sigma_X\sigma_Y}{\sigma_Y^2} = r\cdot\frac{\sigma_X}{\sigma_Y} \quad \text{.... (22)}$$

Substituting the value of c from equation (15) into line of regression of X on Y we have

$$X_{Ci} = \bar{X} - d\bar{Y} + dY_i \quad or \quad \left(X_{Ci} - \bar{X}\right) = d\left(Y_i - \bar{Y}\right) \quad \text{.... (23)}$$

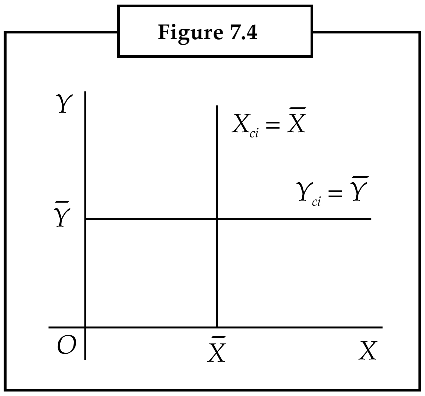
$$or \quad \left(X_{Ci} - \bar{X}\right) = r\cdot\frac{\sigma_X}{\sigma_Y}\left(Y_i - \bar{Y}\right) \quad \text{.... (24)}$$

Remarks: It should be noted here that the two lines of regression are different because these have been obtained in entirely two different ways. In case of regression of Y on X, it is assumed that the values of X are given and the values of Y are estimated by minimising $\Sigma(Y_i - Y_{Ci})^2$ while in case of regression of X on Y, the values of Y are assumed to be given and the values of X are estimated by minimising $\Sigma(X_i - X_{Ci})^2$. Since these two lines have been estimated on the basis of different assumptions, they are not reversible, i.e., it is not possible to obtain one line from the other by mere transfer of terms. There is, however, one situation when these two lines will coincide. From the study of correlation we may recall that when r = ±1, there is perfect correlation between the variables and all the points lie on a straight line. Therefore, both the lines of regression coincide and hence they are also reversible in this case. By substituting r = ±1 in equation (12) or (24) it can be shown that the lines of regression in both the cases become

$$\left(\frac{Y_i - \bar{Y}}{\sigma_Y}\right) = \pm\left(\frac{X_i - \bar{X}}{\sigma_X}\right)$$

Further when r = 0, equation (12) becomes $Y_{Ci} = \overline{Y}$ and equation (24) becomes $X_{Ci} = \overline{X}$. These are the equations of lines parallel to X-axis and Y-axis respectively. These lines also intersect at the point $\left(\overline{X}, \overline{Y}\right)$ and are mutually perpendicular at this point, as shown in figure 23.4.

Figure 7.4



7.1.3 Correlation Coefficient and the two Regression Coefficients

Since $b = r \cdot \dfrac{\sigma_Y}{\sigma_X}$ and $d = r \cdot \dfrac{\sigma_X}{\sigma_Y}$, we have

$b.d = r\dfrac{\sigma_Y}{\sigma_X} \cdot r\dfrac{\sigma_X}{\sigma_Y} = r^2$ or $r = \sqrt{b.d}$. This shows that correlation coefficient is the geometric mean of the two regression coefficients.

Remarks:

The following points should be kept in mind about the coefficient of correlation and the regression coefficients :

(i) Since $r = \dfrac{Cov(X,Y)}{\sigma_X \sigma_Y}$, $b = \dfrac{Cov(X,Y)}{\sigma_X^2}$ and $d = \dfrac{Cov(X,Y)}{\sigma_Y^2}$, therefore the sign of r, b and d will always be same and this will depend upon the sign of Cov (X, Y).

(ii) Since bd = r^2 and 0 £ r^2 £ 1, therefore either both b and d are less than unity or if one of them is greater than unity, the other must be less than unity such that 0 £ b.d £ 1 is always true.

Example 1:

Obtain the two regression equations and find correlation coefficient between X and Y from the following data :

X	:	10	9	7	8	11
Y	:	6	3	2	4	5

Solution.

Calculation table

X	Y	XY	X^2	Y^2
10	6	60	100	36
9	3	27	81	9
7	2	14	49	4
8	4	32	64	16
11	5	55	121	25
45	20	188	415	90

(a) Regression of Y on X

$$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{5 \times 188 - 45 \times 20}{5 \times 415 - (45)^2} = 0.8$$

Also, $\overline{X} = \frac{45}{5} = 9$ and $\overline{Y} = \frac{20}{5} = 4$

Now $a = \overline{Y} - b\overline{X} = 4 - 0.8 \times 9 = -3.2$

$\therefore$ Regression of Y on X is $Y_C = -3.2 + 0.8X$

(b) Regression of X on Y

$$d = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2} = \frac{5 \times 188 - 45 \times 20}{5 \times 90 - (20)^2} = 0.8$$

Also, $c = \overline{X} - d\overline{Y} = 9 - 0.8 \times 4 = 5.8$

$\therefore$ The regression of X on Y is $X_C = 5.8 + 0.8Y$

(c) Coefficient of correlation $r = \sqrt{b.d} = \sqrt{0.8 \times 0.8} = 0.8$

Example 2:

From the data given below, find :

(a) The two regression equations.

(b) The coefficient of correlation between marks in economics and statistics.

(c) The most likely marks in statistics when marks in economics are 30.

Marks in Eco. :	25	28	35	32	31	36	29	38	34	32
Marks in Stat. :	43	46	49	41	36	32	31	30	33	39

Solution.

Calculation table

Marks in Eco. *(X)*	*Marks in Stat.* *(Y)*	$u = X - 31$	$v = Y - 41$	uv	u^2	v^2
25	43	– 6	2	– 12	36	4
28	46	– 3	5	– 15	9	25
35	49	4	8	32	16	64
32	41	1	0	0	1	0
31	36	0	– 5	0	0	25
36	32	5	– 9	– 45	25	81
29	31	– 2	– 10	20	4	100
38	30	7	– 11	– 77	49	121
34	33	3	– 8	– 24	9	64
32	39	1	– 2	– 2	1	4
Total		10	– 30	– 123	150	488

From the table, we have

$$\overline{X} = 31 + \frac{10}{10} = 32 \textit{ and } \overline{Y} = 41 - \frac{30}{10} = 38.$$

(a) The lines of regression

(i) Regression of Y on X

$$b = \frac{n\sum uv - \left(\sum u\right)\left(\sum v\right)}{n\sum u^2 - \left(\sum u\right)^2} = \frac{-1230 + 300}{1500 - 100} = -0.66$$

$$a = \overline{Y} - b\overline{X} = 38 + 0.66 \times 32 = 59.26$$

∴ Regression equation is

$$Y_C = 59.26 - 0.66X$$

(ii) Regression of X on Y

$$d = \frac{n\sum uv - \left(\sum u\right)\left(\sum v\right)}{n\sum v^2 - \left(\sum v\right)^2} = \frac{-1230 + 300}{4880 - 900} = -0.23$$

$$c = \overline{X} - d\overline{Y} = 32 + 0.23 \times 38 = 40.88$$

∴ Regression equation is

$$X_C = 40.88 - 0.23Y$$

(b) Coefficient of correlation

$$r = \sqrt{b \cdot d} = -\sqrt{-0.66 \times -0.23} = -0.39$$

Note that r, b and d are of same sign.

(c) Since we have to estimate marks in statistics denoted by Y, therefore, regression of Y on X will be used. The most likely marks in statistics when marks in economics are 30, is given by

$$Y_C = 59.26 - 0.66 \times 30 = 39.33$$

Example 3:

Obtain the two lines of regression from the following data and estimate the blood pressure when age is 50 years. Can we also estimate the blood pressure of a person aged 20 years on the basis of this regression equation? Discuss.

Age (X) (in years) :	56	42	72	39	63	47	52	49	40	42	68	60
Blood Pressure (Y) :	127	112	140	118	129	116	130	125	115	120	135	133

Solution.

Calculation table

X	Y	$u = X - 52$	$v = Y - 125$	uv	u^2	v^2
56	127	4	2	8	16	4
42	112	- 10	- 13	130	100	169
72	140	20	15	300	400	225
39	118	- 13	- 7	91	169	49
63	129	11	4	44	121	16
47	116	- 5	- 9	45	25	81
52	130	0	5	0	0	25
49	125	- 3	0	0	9	0
40	115	- 12	- 10	120	144	100
42	120	- 10	- 5	50	100	25
68	135	16	10	160	256	100
60	133	8	8	64	64	64
	Total	6	0	1012	1404	858

From the table, we have

$$\overline{X} = 52 + \frac{6}{12} = 52.5 \quad \text{and} \quad \overline{Y} = 125$$

(a) Regression of Y on X

$$b = \frac{n\sum uv - (\sum u)(\sum v)}{n\sum u^2 - (\sum u)^2} = \frac{12 \times 1012 - 6 \times 0}{12 \times 1404 - (6)^2} = 0.72$$

Also $a = \overline{Y} - b\overline{X} = 125 - 0.72 \times 52.5 = 87.2$

∴ The line of regression of Blood pressure (Y) on Age (X) is

$Y_C = 87.2 + 0.72X$

(b) Regression of X on Y

$$d = \frac{n\sum uv - (\sum u)(\sum v)}{n\sum v^2 - (\sum v)^2} = \frac{12 \times 1012 - 6 \times 0}{12 \times 858 - 0} = 1.18$$

Also $c = \overline{X} - d\overline{Y} = 52.5 - 1.18 \times 125 = -95$

∴ Line of regression of Age (X) on Blood pressure (Y) is

$X_C = - 95 + 1.18Y$

(c) (i) To estimate blood pressure (Y) for a given age, X = 50 years, we shall use regression of Y on X

$\therefore Y_C = 87.2 + 0.72 \times . 50 = 123.2$

(ii) The estimate of blood pressure when age is 20 years

$Y_C = 87.2 + 0.72 \times . 20 = 101.6$

It should be noted here that this estimate is wrong because the blood pressure of a normal person cannot be less than 110.

This result reflects the limitations of regression analysis with regard to estimation or prediction. It is important to note that the prediction, based on regression line, should be done only for those values of the variable that are not very far from the range of the observed data, used to derive the line of regression. The prediction from a regression line for a value of the variable that is far away from the observed data is likely to give inconsistent results like the one obtained above.

Example 4:

A panel of judges P and Q graded seven dramatic performances by independently awarding marks as follows :

Performance	:	1	2	3	4	5	6	7
Marks by P	:	46	42	44	40	43	41	45
Marks by Q	:	40	38	36	35	39	37	41

The eighth performance which Judge Q could not attend, was awarded 37 marks by Judge P. If Judge Q had also been present, how many marks would be expected to have been awarded by him to eighth performance?

Solution.

Let us denote marks awarded by the Judge P as X and marks awarded by the Judge Q as Y. Since we have to estimate marks that would have been awarded by Judge Q, we shall fit a line of regression of Y on X to the given data.

Calculation table

X	Y	$u = X - 43$	$v = X - 37$	uv	u^2	v^2
46	40	3	3	9	9	9
42	38	-1	1	-1	1	1
44	36	1	-1	-1	1	1
40	35	-3	-2	6	9	4
43	39	0	2	0	0	4
41	37	-2	0	0	4	0
45	41	2	4	8	4	16
Total		0	7	21	28	35

From the table, we have

$$\overline{X} = 43 \quad \text{and} \quad \overline{Y} = 37 + \frac{7}{7} = 38$$

Further, $b = \dfrac{n\sum uv - (\sum u)(\sum v)}{n\sum u^2 - (\sum u)^2} = \dfrac{7 \times 21 - 0}{7 \times 28 - 0} = 0.75$

Also $a = \overline{Y} - b\overline{X} = 38 - 0.75 \times 43 = 5.75$

∴ $Y_C = 5.75 + 0.75X$ is the fitted line of regression.

Estimate of Y when X = 37

$Y_C = 5.75 + 0.75 \times 37 = 33.5$ marks

∴ It is expected that the Judge Q would have awarded 33.5 marks to the eighth performance.

Example 5:

Find out the regression coefficients of Y on X, X on Y and correlation coefficient between X and Y on the basis of the following data :

ΣXY = 350, ΣX = 50, ΣY = 60, n = 10, Variance of X = 4 and Variance of Y = 9.

Solution.

Regression coefficient of Y on X is given by

$$b = \frac{\frac{\sum XY}{n} - \left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right)}{\sigma_X^2} = \frac{\frac{350}{10} - \left(\frac{50}{10}\right)\left(\frac{60}{10}\right)}{4} = 1.25$$

Regression coefficient of X on Y is given by

$$d = \frac{\frac{\sum XY}{n} - \left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right)}{\sigma_Y^2} = \frac{35 - 30}{9} = 0.55$$

Coefficient of correlation between X and Y is given by

$$r = \sqrt{1.25 \times 0.55} = 0.83$$

Example 6:

The following results were worked out from scores in statistics and mathematics in a certain examination :

	Scores in Statistics (X)	*Scores in Mathematics (Y)*
Mean	39.5	47.5
Standard Deviation	10.8	17.8

Karl Pearson's correlation coefficient between X and Y = 0.42. Find both the regression lines. Use these lines to estimate the value of Y when X = 50 and the value of X when Y = 30.

Solution.

(a) Regression of Y on X

Regression coefficient $b = r \cdot \frac{\sigma_Y}{\sigma_X} = 0.42 \times \frac{17.8}{10.8} = 0.69$

and $a = \overline{Y} - b\overline{X} = 47.5 - 0.69 \times 39.5 = 20.24$

∴ The line of regression of Y on X is $Y_C = 20.24 + 0.69X$,

Notes

and the predicted value of Y when X = 50, is given by

$Y_C = 20.24 + 0.69 \times 50 = 54.74.$

(b) Regression of X on Y

Regression coefficient $d = r \cdot \frac{\sigma_X}{\sigma_Y} = 0.42 \times \frac{10.8}{17.8} = 0.25$

and $c = \overline{X} - d\overline{Y} = 39.5 - 0.25 \times 47.5 = 27.62$

∴ The line of regression of X on Y is $X_C = 27.62 + 0.25Y$

and the predicted value of X when Y = 30 is given by

$X_C = 27.62 + 0.25 \times 30 = 35.12$

Example 7:

For a bivariate data, you are given the following information :

$\Sigma(X - 58) = 46$ $\qquad \Sigma(X - 58)^2 = 3086$

$\Sigma(Y - 58) = 9$ $\qquad \Sigma(Y - 58)^2 = 483$

$\Sigma(X - 58)(Y - 58) = 1095.$

Number of pairs of observations = 7. You are required to determine (i) the two regression equations and (ii) the coefficient of correlation between X and Y.

Solution.

Let u = X - 58 and v = Y - 58. In terms of our notations, we are given $\Sigma u = 46$, $\Sigma u^2 = 3086$, $\Sigma v = 9$, $\Sigma v^2 = 483$, $\Sigma uv = 1095$ and $n = 7$.

Now $\overline{X} = 58 + \frac{46}{7} = 64.7$ and $\overline{Y} = 58 + \frac{9}{7} = 59.29$

(a) For regression equation of Y on X, we have

$$b = \frac{7 \times 1095 - 46 \times 9}{7 \times 3086 - (46)^2} = 0.37$$

and $a = \overline{Y} - b\overline{X} = 59.29 - 0.37 \times 64.57 = 35.40$

∴ The line of regression of Y on X is given by

$Y_C = 35.40 + 0.37X$

(b) For regression equation of X on Y, we have

$$d = \frac{7 \times 1095 - 46 \times 9}{7 \times 483 - (9)^2} = 2.20$$

and $c = \overline{X} - d\overline{Y} = 64.57 - 2.2 \times 59.29 = -65.87$

∴ The line of regression of X on Y is given by

$X_C = -65.87 + 2.2Y$

(c) The coefficient of correlation

$$r = \sqrt{b \cdot d} = \sqrt{0.37 \times 2.2} = 0.90$$

Example 8:

Find the means of X and Y variables and the coefficient of correlation between them from the following two regression equations :

$$3Y - 2X - 10 = 0$$

$$2Y - X - 50 = 0$$

Solution.

(a) The means of X and Y

We know that both the lines of regression intersect at the point $\left(\overline{X}, \overline{Y}\right)$. The simultaneous solution of the given equations will give the mean values of X and Y as

$\overline{X} = 130$ and $\overline{Y} = 90$ respectively.

(b) Correlation Coefficient

Let us assume that the first equation be regression of Y on X. Rewriting this equation as 3Y = 2X + 10 or $Y = \frac{2}{3}X + \frac{10}{3}$.

$\therefore$ The corresponding regression coefficient, $b = \frac{2}{3}$

Further, assuming the second equation as regression of X on Y, we can rewrite this equation as X = 2Y - 50.

$\therefore$ The regression coefficient, d = 2

Since b.d $= \frac{2}{3} \cdot 2 = \frac{4}{3}$ > 1, therefore, our assumptions regarding the two regression lines are wrong.

Now we reverse these assumptions and assume that the first equation is regression of X on Y and second the regression of Y on X.

$\therefore$ The first equation can be written as 2X = 3Y - 10 or $X = \frac{3}{2}Y - 5$, so that the corresponding regression coefficient is $d = \frac{3}{2}$. Further, the second equation can be written as 2Y = X + 50 or $Y = \frac{1}{2}X + 25$, so that the corresponding regression coefficient is $b = \frac{1}{2}$. Since b.d $= \frac{3}{2} \times \frac{1}{2} = \frac{3}{4} < 1$, our assumption is correct.

Also r^2 = b.d = $\frac{3}{4}$ $\quad \therefore \; r = \sqrt{\frac{3}{4}} = 0.87$

7.2 Regression Coefficient in a Bivariate Frequency Distribution

Notes

As in case of calculation of correlation coefficient (see § 12.6), we can directly write the formula for the two regression coefficients for a bivariate frequency distribution as given below :

$$b = \frac{N\sum\sum f_{ij}X_iY_j - \left(\sum f_iX_i\right)\left(\sum f_j'Y_j\right)}{N\sum f_iX_i^2 - \left(\sum f_iX_i\right)^2}$$

or, if we define $u_i = \frac{X_i - A}{h}$ and $v_j = \frac{Y_j - B}{k}$,

$$b = \frac{k}{h}\left[\frac{N\sum\sum f_{ij}u_iv_j - \left(\sum f_iu_i\right)\left(\sum f_j'v_j\right)}{N\sum f_iu_i^2 - \left(\sum f_iu_i\right)^2}\right]$$

Similarly, $d = \frac{N\sum\sum f_{ij}X_iY_j - \left(\sum f_iX_i\right)\left(\sum f_j'Y_j\right)}{N\sum f_j'Y_j^2 - \left(\sum f_j'Y_j\right)^2}$

or $d = \frac{h}{k}\left[\frac{N\sum f_{ij}u_iv_j - \left(\sum f_iu_i\right)\left(\sum f_j'v_j\right)}{N\sum f_j'v_j^2 - \left(\sum f_j'v_j\right)^2}\right]$

Example 12:

By calculating the two regression coefficients obtain the two regression lines from the following data:

$Y \rightarrow$ $X \downarrow$	0 - 5	5 - 10	10 - 15
0 - 10	2	5	7
10 - 20	1	3	2
20 - 30	8	4	0

Solution.

The mid points of X-values are 5, 15, 25.

Let $u = \frac{X - 15}{10}$, $\therefore$ Corresponding u-values become - 1, 0, 1

Similarly, the mid-points of Y-values are 2.5, 7.5, 12.5

Let $v = \frac{Y - 7.5}{5}$, $\therefore$ Corresponding v-values become - 1, 0, 1

Notes

Calculation Table

$u \backslash v$	−1	0	1	f_i	f_iu_i	$f_iu_i^2$	$f_{ij}u_iv_j$
−1	$2\ \boxed{2}$	$5\ \boxed{0}$	$7\ \boxed{-7}$	14	−14	14	−5
0	$1\ \boxed{0}$	$3\ \boxed{0}$	$2\ \boxed{0}$	6	0	0	0
1	$8\ \boxed{-8}$	$4\ \boxed{0}$	$0\ \boxed{0}$	12	12	12	−8
f'_j	11	12	9	32	−2	26	−13
$f'_jv'_j$	−11	0	9	−2			
$f'_jv'^2_j$	11	0	9	20			

From the table N = 32 (total frequency)

(a) Regression of Y on X

Regression Coefficient (here h = 10 and k = 5)

$$b=\left[\frac{-32\times 13-2\times 2}{32\times 26-4}\right]\times\frac{5}{10}=\frac{-416-4}{832-4}\times\frac{1}{2}=-\ 0.25$$

Also, $\overline{X}=15+\dfrac{10(-2)}{32}=14.73$ and $\overline{Y}=7.5+\dfrac{5(-2)}{32}=7.19$

$\therefore\ a=\overline{Y}-b\overline{X}$ = 7.19 + 0.25 × 14.73 = 10.87

Hence, the regression of Y on X becomes Y_C = 10.87 - 0.25X

(b) Regression of X on Y

Regression coefficient $d=\left[\dfrac{-420}{32\times 20-4}\right]\times\dfrac{10}{5}=-1.32$

Also, $c=\overline{X}-d\overline{Y}$ =14.73 + 1.32 × 7.19 = 24.22

Hence, the regression of X on Y becomes X_C = 24.22 – 1.32Y

7.3 The Coefficient of Determination

We recall that in the line of regression Y_C = a + bX, X is used to estimate the value of Y. Further, the estimate of Y, independently of X, is given by a constant. Let this constant be A. Thus, we can write Y_C = A.

Given the observations Y_1, Y_2, ... Y_n, A will be the best estimate of Y if $S=\sum_{i=1}^{n}\left(Y_i-A\right)^2$ is minimum.

The necessary condition for minimum of S is $\dfrac{\partial S}{\partial A}=0$.

$i.e.,\qquad 2\sum(Y_i-A)=0\qquad or\qquad \sum Y_i-nA=0\qquad or\qquad A=\overline{Y}$.

$\therefore$ The best estimate (an estimate having minimum sum of squares of errors) of Y, independently of X, is given by $Y_C = \overline{Y}$.

Remarks: If X and Y are independent variables, the two lines of regression are $Y_C = \overline{Y}$ and $X_C = \overline{X}$.

Very often, when we use X for the estimation of Y, we are interested in knowing how far the use of X enables us to explain the variations in Y values from $\overline{Y}$ or, in other words, how much of the variations in Y, from $\overline{Y}$, are being explained by the regression equation $Y_{Ci} = a + bX_i$? To answer this question, we write

$$Y_i - \overline{Y} = Y_i - Y_{Ci} + Y_{Ci} - \overline{Y} \quad \text{(Subtracting and adding } Y_{Ci}\text{)}$$

$$\text{or} \quad Y_i - \overline{Y} = \left(Y_i - Y_{Ci}\right) + \left(Y_{Ci} - \overline{Y}\right)$$

Squaring both sides and taking sum over all the observations, we have

$$\sum\left(Y_i - \overline{Y}\right)^2 = \sum\left(Y_i - Y_{Ci}\right)^2 + \sum\left(Y_{Ci} - \overline{Y}\right)^2 + 2\sum\left(Y_i - Y_{Ci}\right)\left(Y_{Ci} - \overline{Y}\right) \quad ....(1)$$

Consider the product term

$$2\sum\left(Y_i - Y_{Ci}\right)\left(Y_{Ci} - \overline{Y}\right) = 2\sum\left[\left\{Y_i - \overline{Y} - b\left(X_i - \overline{X}\right)\right\}\left\{b\left(X_i - \overline{X}\right)\right\}\right]$$

$$= 2b\sum\left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right) - 2b^2\sum\left(X_i - \overline{X}\right)^2$$

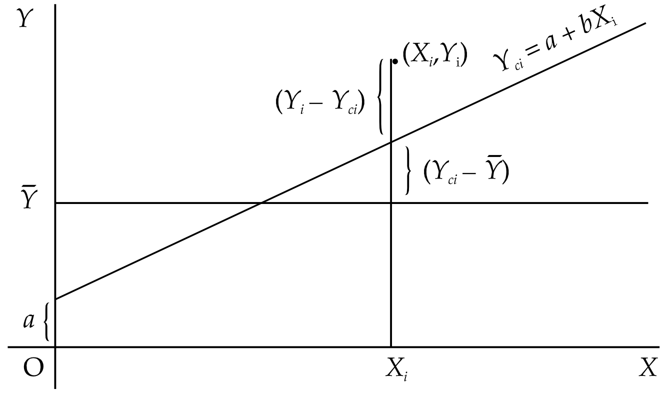
$$= 2b^2\sum\left(X_i - \overline{X}\right)^2 - 2b^2\sum\left(X_i - \overline{X}\right)^2 = 0$$

Thus, equation (1) becomes

$$\sum\left(Y_i - \overline{Y}\right)^2 = \sum\left(Y_i - Y_{Ci}\right)^2 + \sum\left(Y_{Ci} - \overline{Y}\right)^2 \quad .... (2)$$

From the above figure, we note that $Y_{Ci} - \overline{Y}$ is the deviation of the estimated value from $\overline{Y}$. This deviation has occurred because X and Y are related by the regression equation $Y_{Ci} = a + bX_i$, so that the estimate of Y is Y_{Ci} when $X = X_i$. Similar type of deviations would occur for other values of X. Thus, the magnitude of the term $\sum\left(Y_{Ci} - \overline{Y}\right)^2$ gives the strength of the relationship, $Y_{Ci} = a + bX_i$, between X and Y or, equivalently, the variations in Y that are explained by the regression equation.

Figure 23.5



Notes

The other term Y_i - Y_{Ci} gives the deviation of i th observed value from the regression line and thus the magnitude of the term $\sum\left(Y_i - Y_{Ci}\right)^2$ gives the variations in Y about the line of regression. These variations are also known as unexplained variations in Y.

Adding the two types of variations, we get the magnitude of total variations in Y. Thus, equation (2) can also be written as

Total variations in Y = Unexplained variations in Y + Explained variations in Y.

Dividing both sides of equation (2) by $\sum\left(Y_i - \overline{Y}\right)^2$, we have

$$1 = \frac{\sum\left(Y_i - Y_{Ci}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2} + \frac{\sum\left(Y_{Ci} - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2} \quad \text{.... (3)}$$

or 1 = Proportion of unexplained variations + Proportion of variations explained by the regression equation.

The proportion of variation explained by regression equation is called the coefficient of determination.

Thus, the coefficient of determination $= \dfrac{\sum\left(Y_{Ci} - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2}$

$$= \frac{b^2\sum\left(X_i - \overline{X}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2} = \frac{\left[\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right]^2}{\sum\left(X_i - \overline{X}\right)^2\sum\left(Y_i - \overline{Y}\right)^2} = r^2$$

This result shows that the coefficient of determination is equal to the square of the coefficient of correlation, i.e., r^2 gives the proportion of variations explained by each regression equation.

Remarks:

(i) It should be obvious from the above that it is desirable to calculate the coefficient of correlation prior to the fitting of a regression line. If r^2 is high enough, the fitted line will explain a greater proportion of the variations in the dependent variable. A low value of r^2 would, however, indicate that the proposed fitting of regression would not be of much use.

(ii) The expression for the coefficient of determination for regression of X on Y can be written in a similar way. Here we can write $r^2 = \dfrac{\sum\left(X_{Ci} - \overline{X}\right)^2}{\sum\left(X_i - \overline{X}\right)^2}$.

7.3.1 The Coefficient of Non-Determination

The proportion of unexplained variations is also termed as the coefficient of non-determination. It is denoted by k^2, where k^2 = (1 - r^2). The square root of k^2 is termed as the coefficient of alienation, i.e., $k = \sqrt{\left(1 - r^2\right)}$.

Example 13:

Comment on the following statements :

(i) The two regression coefficients of bivariate data are 0.7 and 1.4.

(ii) A correlation coefficient r = 0.8, between the two variables X and Y, implies a relationship twice as close as r = 0.4.

Solution.

(i) This statement implies that $r^2 = 0.7 \times 1.4 = 0.98$, i.e., a linear regression fitted to the data would explain 98% of the variations in the dependent variable.

(ii) The given statement is wrong. Since r = 0.8 implies that a regression fitted to the data would explain 64% of the variations in the dependent variable while r = 0.4 implies that the proportion of such variations is only 16%. Thus, r = 0.8 implies a relation that is four times as close as r = 0.4.

Example 14:

The correlation coefficient between two variables is found to be 0.8. Explain the meaning of this statement.

Solution.

The given statement implies that :

(i) Two variables are highly correlated.

(ii) There is positive association between them, i.e., an increase in value of one is accompanied by the increase in value of the other and vice-versa.

(iii) A linear regression fitted to the data would explain 64% of the variations in the dependent variable.

7.4 Mean of the Estimated Values

We may recall that Y_C and X_C are the estimated values from the regressions of Y on X and X on Y respectively.

Consider the regression equation $Y_{Ci} - \overline{Y} = b\left(X_i - \overline{X}\right)$.

Taking sum over all the observations, we get

$$\sum\left(Y_{Ci} - \overline{Y}\right) = b\sum\left(X_i - \overline{X}\right) = 0$$

$$\Rightarrow \quad \sum Y_{Ci} - n\overline{Y} = 0 \quad \text{or} \quad \frac{\sum Y_{Ci}}{n} = \overline{Y}_C = \overline{Y} \qquad .... (1)$$

Similarly, it can be shown that $\overline{X}_C = \overline{X}$.

This implies that the mean of the estimated values is also equal to the mean of the observed values.

7.5 Mean and Variance of 'e_i' values

(i) Mean of e_i values

We know that $e_i = Y_i - Y_{Ci}$.

Taking sum over all the observations, we have

$$\sum e_i = \sum\left(Y_i - Y_{Ci}\right) = \sum Y_i - \sum Y_{Ci} = 0 \quad \text{[from equation (1)]}$$

$\therefore$ Mean of e_i values is equal to zero.

(ii) Variance of e_i values

The variance of e_i values, in case of regression of Y on X, is given by

$$S_{Y.X}^2 = \frac{1}{n}\sum\left(e_i - 0\right)^2 = \frac{1}{n}\sum\left(Y_i - Y_{Ci}\right)^2 \qquad \text{.... (2)}$$

[Note that $\sum\left(Y_i - Y_{Ci}\right)^2$ is the magnitude of unexplained variation in Y]

$$S_{Y.X}^2 = \frac{1}{n}\sum\left[\left(Y_i - \bar{Y}\right) - b\left(X_i - \bar{X}\right)\right]^2$$

$$= \frac{\sum\left(Y_i - \bar{Y}\right)^2}{n} + \frac{b^2\sum\left(X_i - \bar{X}\right)^2}{n} - \frac{2b\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n}$$

$$= \sigma_Y^2 + b^2\sigma_X^2 - 2b\cdot b\sigma_X^2 = \sigma_Y^2 - b^2\sigma_X^2$$

$$= \sigma_Y^2 - r^2\sigma_Y^2 = \sigma_Y^2\left(1 - r^2\right)$$

Similarly, it can be shown that the mean of e'_i (= $X_i - X_{Ci}$) values, in case of regression of X on Y, is also equal to zero. Further, their variance, i.e.,

$$S_{X.Y}^2 = \sigma_X^2\left(1 - r^2\right)$$

Alternatively equation (2) can be written as

$$S_{Y.X}^2 = \frac{1}{n}\sum\left(Y_i - Y_{ci}\right)Y_i = \frac{1}{n}\left[\sum Y_i^2 - a\sum Y_i - b\sum X_iY_i\right]$$

Similarly, we can write

$$S_{X.Y}^2 = \frac{1}{n}\left[\sum X_i^2 - c\sum X_i - d\sum X_iY_i\right]$$

Remarks:

The above expressions for the variance are based on the following:

$$\Sigma(Y_i - Y_{ci})^2 = \Sigma(Y_i - Y_{ci})(Y_i - Y_{ci})$$

$$= \Sigma(Y_i - Y_{ci})Y_i - \Sigma(Y_i - Y_{ci})Y_{ci}$$

It can be shown that the last term is zero.

$$\Sigma(Y_i - Y_{ci})Y_{ci} = \Sigma[(Y_i - \bar{Y}) - b(X_i - \bar{X})][\bar{Y} + b(X_i - \bar{X})]$$

$$= \bar{Y}\,\Sigma(Y_i - \bar{Y}) - b\bar{Y}\,\Sigma(X_i - \bar{X}) + b\Sigma(X_i - \bar{X})(Y_i - \bar{Y}) - b^2\Sigma(X_i - \bar{X})^2$$

$$= 0 - 0 + b^2\Sigma(X_i - \bar{X})^2 - b^2\Sigma(X_i - \bar{X})^2 = 0$$

7.5.1 Standard Error of the Estimate

The standard error of the estimate of regression is given by the positive square root of the variance of e_i values.

The standard error of the estimate of regression of Y on X or simply the standard error of the estimate of Y is given as, $S_{Y.X} = \sigma_Y\sqrt{1-r^2}$.

Similarly, $S_{X.Y} = \sigma_X\sqrt{1-r^2}$ is the standard error of the estimate X.

According to the theory of estimation, to be discussed in Chapter 21, an unbiased estimate of the variance of e_i values is given by

$$s^2_{Y.X} = \frac{\sum e_i^2}{n-2} = \frac{n}{n-2}\cdot\frac{\sum e_i^2}{n} = \frac{n}{n-2}\cdot\sigma_Y^2\left(1-r^2\right)$$

∴ The standard errors of the estimate of Y and that of X are written as

$$s_{Y.X} = \sigma_Y\sqrt{\frac{n}{(n-2)}\left(1-r^2\right)} \text{ and } s_{X.Y} = \sigma_X\sqrt{\frac{n}{(n-2)}\left(1-r^2\right)} \text{ respectively.}$$

Example 15:

From the following data, compute (i) the coefficient of correlation between X and Y, (ii) the standard error of the estimate of Y :

$$\sum x^2 = 24 \quad \sum y^2 = 42 \quad \sum xy = 30 \quad N = 10\text{, where } x = X - \bar{X} \text{ and } y = Y - \bar{Y}\text{ .}$$

Solution.

The coefficient of correlation between X and Y is given by

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = \frac{30}{\sqrt{24}\sqrt{42}} = 0.94$$

The standard error of the estimate of Y is given by (n < 30)

$$s_{Y.X} = \sqrt{\frac{\left(1-r^2\right)\sum y^2}{n-2}} = \sqrt{\frac{\left(1-0.94^2\right)\times 42}{8}} = 0.79$$

Example 16: For 100 items, it is given that the regression equations of Y on X and X on Y are 8X – 10Y + 66 = 0 and 40X – 18Y = 214 respectively. Compute the arithmetic means of X and Y and the coefficient of determination. If the standard deviation of X is given to be 3, compute the standard error of the estimate of Y.

Solution.

(a) The means of X and Y

Since the lines of regression pass through the point $\left(\overline{X}, \overline{Y}\right)$, the simultaneous solution of the given regression equations would give the mean values of X and Y as $\overline{X} = 13, \overline{Y} = 17$

(b) The coefficient of determination

We assume that 8X - 10Y + 66 = 0 is the regression of Y on X and 40X - 18Y = 214 is the regression of X on Y. Thus, the respective regression coefficients b and d are given by $\frac{8}{10}$ and $\frac{18}{40}$.

$\therefore$ The coefficient of determination r^2 = b.d $= \frac{8}{10} \times \frac{18}{40} = 0.36$

(c) The standard error of the estimate of Y

We know that $s_{Y.X} = \sigma_Y \sqrt{1-r^2}$. To find s_Y we use the relation $b = r \cdot \frac{\sigma_Y}{\sigma_X}$.

Also $r^2 = \frac{9}{25} \ \backslash \ r = \frac{3}{5}$ Thus, $\sigma_Y = \frac{b.\sigma_X}{r} = \frac{8}{10} \times \frac{5}{3} \times 3 = 4$

Hence, $s_{Y.X} = 4\sqrt{1-0.36} = 3.2$

7.6 Summary of Formulae

I. Regression of Y on X

1. Regression coefficient $b = \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sum X^2 - n\overline{X}^2} = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum X^2 - \left(\sum X\right)^2}$

   Also b $= \frac{Cov(X,Y)}{\sigma_X^2} = r \cdot \frac{\sigma_Y}{\sigma_X}$

2. Change of scale and origin

   If $u = \frac{X-A}{h}$ and $v = \frac{Y-B}{h}$, then $b = \frac{k}{h}\left[\frac{n\sum uv - \left(\sum u\right)\left(\sum v\right)}{n\sum u^2 - \left(\sum u\right)^2}\right]$.

3. Constant term $a = \overline{Y} - b\overline{X}$

4. Alternative form of regression equation

   $Y_C - \overline{Y} = \left(X - \overline{X}\right) \quad or \quad Y_C - \overline{Y} = r \cdot \frac{\sigma_Y}{\sigma_X}\left(X - \overline{X}\right)$

5. Regression coefficient in bivariate frequency distribution

$$b = \frac{k}{h}\left[\frac{N\sum\sum f_{ij}u_i v_j - \left(\sum f_i u_i\right)\left(\sum f_j' v_j\right)}{N\sum f_i u_i^2 - \left(\sum f_i u_i\right)^2}\right]$$

6. Standard Error of the estimate

$s_{Y.X} = \sigma_Y\sqrt{1-r^2}$ for large n (i.e., n > 30)

$= \sqrt{\dfrac{\sum\left(Y_i - \bar{Y}\right)^2\left(1-r^2\right)}{n-2}}$ for small n

II. Regression of X on Y

1. Regression Coefficient $d = \dfrac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n\bar{Y}^2} = \dfrac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2}$

$= \dfrac{Cov(X,Y)}{\sigma_Y^2} = r\cdot\dfrac{\sigma_X}{\sigma_Y}$

2. Change of scale and origin

If $u = \dfrac{X-A}{h}$ and $v = \dfrac{Y-B}{h}$, then $d = \dfrac{h}{k}\left[\dfrac{n\sum uv - (\sum u)(\sum v)}{n\sum v^2 - (\sum v)^2}\right]$.

3. Constant term $c = \bar{X} - d\bar{Y}$

4. Alternative form of regression equation
$X_C - \bar{X} = d\left(Y - \bar{Y}\right)$ *or* $X_C - \bar{X} = r.\dfrac{\sigma_X}{\sigma_Y}\left(Y - \bar{Y}\right)$

5. Regression coefficient in a bivariate frequency distribution

$$d = \frac{h}{k}\left[\frac{N\sum\sum f_{ij}u_i v_j - \left(\sum f_i u_i\right)\left(\sum f_j' v_j\right)}{N\sum f_j' v_j^2 - \left(\sum f_j' v_j\right)^2}\right]$$

6. Standard error of the estimate

$s_{X.Y} = \sigma_X\sqrt{1-r^2}$ for large n (i.e., n > 30)

$= \sqrt{\dfrac{\sum\left(X_i - \bar{X}\right)^2\left(1-r^2\right)}{n-2}}$ for small n

Notes

III. Relation of r with b and d

$$b \times d = r \cdot \frac{\sigma_Y}{\sigma_X} \cdot r \cdot \frac{\sigma_X}{\sigma_Y} = r^2$$

or $r = \sqrt{b \times d}$

7.7 Keywords

Coefficient of correlation: If the coefficient of correlation calculated for bivariate data (X_i, Y_i), i = 1,2, ... n, is reasonably high and a cause and effect type of relation is also believed to be existing between them, the next logical step is to obtain a functional relation between these variables.

Term regression: The term regression was first introduced by Sir Francis Galton in 1877.

Independent variable: For a bivariate data (X_i, Y_i), i = 1,2, ... n, we can have either X or Y as independent variable.

7.8 Self Assessment

1. Fill in the blanks :

   (i) The two regression coefficients are of .... sign.

   (ii) If a regression coefficient is negative then the correlation between the variables would also be ....

   (iii) The coefficient of determination is a real number lying between .... and ....

   (iv) Regression analysis is used to study .... between the variables.

   (v) If correlation between two variables is zero, the two regression lines are .... to each other and if it is equal to ± 1, the two lines are the ....

   (vi) The smaller is the angle between the two lines of regression, the .... is correlation between the variables.

   (vii) If r ≠ ± 1, the two regression lines are ....

7.9 Review Questions

1. Distinguish between correlation and regression. Discuss least square method of fitting regression.

2. What do you understand by linear regression ? Why there are two lines of regression? Under what condition(s) can there be only one line ?

3. Define the regression of Y on X and of X on Y for a bivariate data (X_i, Y_i), i = 1, 2, ... n. What would be the values of the coefficient of correlation if the two regression lines (a) intersect at right angle and (b) coincide?

4. (a) Show that the proportion of variations explained by a regression equation is r^2

   (b) What is the relation between Total Sum of Squares (TSS), Explained Sum of Squares (ESS) and Residual Sum of squares (RSS). Use this relationship to prove that the coefficient of correlation has a value between –1 and +1.

   Hint: See § 23.3

5. Write a note on the standard error of the estimate.

6. " The regression line gives only a 'best estimate' of the quantity in question. We may assess the degree of uncertainty in this estimate by calculating its standard error ". Explain.

7. Given a scatter diagram of a bivariate data involving two variables X and Y. Find the conditions of minimisation of $\sum\left(Y - Y_C\right)^2$ and hence derive the normal equations for the linear regression of Y on X. What sum is to be minimised when X is regressed on Y? Write down the normal equation in this case.

8. Explain, fully, the meaning of regression of one variable Y on another variable X. Discuss the method of least squares for fitting a linear regression of the form Y = a + bX. Write down the normal equations and show that $b = r \cdot \dfrac{\sigma_Y}{\sigma_X}$, where the symbols have their usual meaning.

9. Show that the coefficient of correlation is the geometric mean of the two regression coefficients.

10. What is the method of least squares ? Show that the two lines of regression obtained by this method are irreversible except when r = ± 1. Explain.

11. Show that, in principle, there are always two lines of regression for a bivariate data. Prove that the coefficient of correlation between two variables is either + 1 or - 1 when the two lines are identical and is zero when they are perpendicular.

12. Fit a linear regression of Y on X to the following data :

X	:	1	2	3	4	5	6	7	8
Y	:	65	80	45	86	178	205	200	250

13. Obtain the two lines of regression from the following data and show them on a graph. Also construct a scatter diagram of the data.

Age of husband (X) (*in years*)	:	23	27	28	28	28	30	30	33	35	38
Age of wife (Y) (*in years*)	:	18	20	22	27	21	29	27	29	28	29

14. The following table gives the information on the years of education (X) of nine farmers and annual yields per acre (Y) on their farms :

X	:	0	2	4	6	8	10	12	14	16
Y	:	4	4	6	10	10	8	12	8	6

(a) Find the regression equation of yield per acre on education and give an economic interpretation to it.

(b) What is the magnitude of the 'explained variation' in the dependent variable? Find the coefficient of correlation from it.

15. The following table gives the data relating to purchases and sales. Obtain the two regression equations by the method of least squares and estimate the likely sales when purchases equal 100.

Purchases	:	62	72	98	76	81	56	76	92	88	49
Sales	:	112	124	131	117	132	96	120	136	97	85

Answers: Self Assessment

1. (i) same (ii) negative (iii) 0 and 1 (iv) dependence (v) perpendicular, coincident (vi) more (vii) irreversible.

7.10 Further Readings

Books

Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.

Unit 8: Sampling Distributions

CONTENTS

Objectives

After studying this unit, you will be able to:

- Distinction Between Parameter and Statistic
- Sampling Distribution of Sample Mean
- Sampling Distribution of the Number of Successes

Introduction

A theoretical probability distribution is constructed on the basis of the specification of the conditions of a random experiment. In contrast to this, if the construction of the probability distribution is based upon the random experiment of obtaining a sample from a population, the resulting distribution is termed as a sampling distribution.

As we know that the main aim of obtaining a sample from a population is to draw certain conclusions about it. The process of drawing such conclusions, known as 'Statistical Inference', is based upon the rules or the framework provided by various sampling distributions.

It may be recalled here that simple random sampling is a procedure of obtaining a sample of size n from a population of size N such that each combination of n units has an equal chance of being selected as a sample. This definition also implies that every unit of the population has an equal chance of being selected in the sample.

The above definition of random sampling holds in both the situations, i.e., in simple random sampling with replacement *(srswr)* and in simple random sampling with out replacement (srswor).

8.1 Distinction between Parameter and Statistic

Let P_1, P_2 ...... P_N denote the observations on N units of a population and X_1, X_2 ...... X_n be a simple random sample of size n from it.

A parameter is a measure computed from the observation of the population. For example :

Population Mean $(\mu) = \dfrac{P_1 + P_2 + \cdots\cdots + P_N}{N}$,

Population Variance $\left(\sigma^2\right) = \dfrac{1}{N}\sum\left(P_i - \mu\right)^2$, etc. are parameters.

In a similar way, a statistics is a measure computed from the observations of a sample. For example:

Sample Mean $\left(\overline{X}\right) = \dfrac{X_1 + X_2 + \cdots\cdots + X_n}{n}$,

Sample Variance $\left(S^2\right) = \dfrac{1}{n}\sum\left(X_i - \overline{X}\right)^2$, etc. are statistic.

Formally, a parameter is any function of population values while a statistic is a function of sample values.

Very often, the values of various parameters are unknown and these are estimated by the corresponding statistic. For example, sample mean $\overline{X}$ is used as an estimator of population mean *m,* sample standard deviation S is used as an estimator of population standard deviation *s*, etc. The difference between a statistic and the corresponding parameter is known as sampling error. For example, the sampling error in estimation of *m* is $\overline{X}$ - *m*. It may be noted that the sampling error is an error caused by pure chance factors.

When we take a random sample X_1, X_2 ...... X_n from a population P_1, P_2 ...... P_N, the first sample observation X_1 could be any one of the N population observations P_1, P_2 ...... P_N. We know that the probability of selection of any one of the population observation is $\dfrac{1}{N}$ and therefore, we can regard X_1 as a random variable which can take values P_1, P_2 ...... P_N each with probability $\dfrac{1}{N}$.

Further, $E\left(X_1\right) = \dfrac{1}{N}\cdot P_1 + \dfrac{1}{N}\cdot P_2 + \cdots + \dfrac{1}{N}\cdot P_N = \dfrac{1}{N}\sum P_i = \mu$ and

Variance of X_1 = E(X_1 - *m*)^2

$$= \frac{1}{N}\left[\left(P_1 - \mu\right)^2 + \left(P_2 - \mu\right)^2 + \cdots + \left(P_N - \mu\right)^2\right] = \frac{1}{N}\sum\left(P_i - \mu\right)^2 = \sigma^2$$

In a similar way, X_2, X_3 ...... X_n are all random variables, each with mean *m* and variance *s*^{2}. The magnitude of covariance between any two of these variables, say X_i and X_j, will depend upon whether the sampling is with or without replacement.

In the case of sampling with replacement, X_1, X_2 ...... X_n would be statistically independent and the Cov(X_i, X_j) = 0 for i $\neq$ j.

In the case of sampling without replacement, we can write

Cov(X_i, X_j) = E(X_i - m)(X_j - m) $= \sum_{r=1}^{N} \sum_{\substack{s=1, \\ s \neq r}}^{N} \left(P_r - \mu\right)\left(P_s - \mu\right) \cdot p_{rs}$,

where p_{rs} is the joint probability that the rth unit of population is drawn at the ith draw and the sth unit of population is drawn at the jth draw. We note that $p_{rs} = \dfrac{1}{N(N-1)}$. Thus, we have

$$Cov\left(X_i, X_j\right) = \sum_{r=1}^{N} \sum_{\substack{s=1, \\ s \neq r}}^{N} \left(P_r - \mu\right)\left(P_s - \mu\right) \cdot \frac{1}{N(N-1)}$$

$$= \frac{1}{N(N-1)} \sum_{r=1}^{N} \left(P_r - \mu\right) \sum_{\substack{s=1, \\ s \neq r}}^{N} \left(P_s - \mu\right)$$

$$= \frac{1}{N(N-1)} \sum_{r=1}^{N} \left(P_r - \mu\right) \left[\sum_{s=1}^{N} \left(P_s - \mu\right) - \left(P_r - \mu\right)\right]$$

$$= \frac{1}{N(N-1)} \sum_{r=1}^{N} \left(P_r - \mu\right) \left[0 - \left(P_r - \mu\right)\right]$$

$$= -\frac{1}{N(N-1)} \sum_{r=1}^{N} \left(P_r - \mu\right)^2 = -\frac{1}{N(N-1)} \cdot N\sigma^2 = -\frac{\sigma^2}{(N-1)}$$

8.2 Sampling Distribution of Sample Mean

We know that $\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$. In the previous section we have shown that if the sample is random, then each of the X_i's are random variable with mean m and variance s^2. Since $\bar{X}$ is a linear combination of these random variables, therefore, it is also a random variable with mean equal to $E\left(\bar{X}\right) = \dfrac{1}{n}\left[E\left(X_1\right) + E\left(X_2\right) + \cdots + E\left(X_n\right)\right] = \dfrac{1}{n} \cdot n\mu = \mu$ and variance equal to

$$Var\left(\bar{X}\right) = E\left(\bar{X} - \mu\right)^2 = E\left[\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right]^2$$

$$= E\left[\frac{\left(X_1 + X_2 + \cdots + X_n\right) - n\mu}{n}\right]^2 = \frac{1}{n^2} E\left[\sum\left(X_i - \mu\right)\right]^2$$

$$= \frac{1}{n^2} E\left[\sum\left(X_i - \mu\right)^2 + \sum\sum_{i \neq j}\left(X_i - \mu\right)\left(X_j - \mu\right)\right]$$

$$= \frac{1}{n^2}\left[\sum E\left(X_i - \mu\right)^2 + \sum\sum_{i \neq j} E\left(X_i - \mu\right)\left(X_j - \mu\right)\right]$$

$$= \frac{1}{n^2}\left[n\sigma^2 + \sum\sum_{i \neq j} Cov\left(X_i, X_j\right) \right]$$

Case I. If the sample is drawn with replacement, then X_1, X_2 ...... X_n are independent random variates and hence, $Cov(X_i, X_j) = 0$. Thus, we have

$$Var\left(\overline{X}\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Case II. If the sample is drawn without replacement, then

$$Cov\left(X_i, X_j\right) = -\frac{\sigma^2}{N-1}, \text{ therefore,}$$

$$Var\left(\overline{X}\right) = \frac{1}{n^2}\left[n\sigma^2 - n(n-1)\frac{\sigma^2}{N-1} \right] = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$$

We note that if N → ∞ (i.e., population becomes large), $\frac{N-n}{N-1} \to 1$ and therefore, in this case also, $Var\left(\overline{X}\right) = \frac{\sigma^2}{n}$.

Remarks:

1. The standard deviation of a statistic is termed as standard error. The standard error of $\overline{X}$, to be written in abbreviated form as $S.E.\left(\overline{X}\right)$, is equal to $\frac{\sigma}{\sqrt{n}}$, when sampling is with replacement and it is equal to $\frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$, when sampling is without replacement.
2. $S.E.\left(\overline{X}\right)$ is inversely related to the sample size.
3. The term $\sqrt{\frac{N-n}{N-1}}$ is termed as finite population correction (fpc). We note that fpc tends to become closer and closer to unity as population size becomes larger and larger.
4. As a general rule, fpc may be taken to be equal to unity when sample size is less than 5% of population size, i.e., n < 0.05N.

Example 1: Construct a sampling distribution of the sample mean for the following population when random samples of size 2 are taken from it (a) with replacement and (b) without replacement. Also find the mean and standard error of the distribution in each case.

Population Unit	:	1	2	3	4
Observation	:	22	24	26	28

Solution.

The mean and standard deviation of population are

$$\mu = \frac{22+24+26+28}{4} = 25 \text{ and}$$

$$\sigma = \sqrt{\frac{(22)^2+(24)^2+(26)^2+(28)^2}{4}-(25)^2} = \sqrt{5} = 2.236 \text{ respectively.}$$

(a) When random samples of size 2 are drawn, we have $4^2 = 16$ samples, shown below :

Sample No.	*Sample Values*	$\overline{X}$
1	22,22	22
2	22,24	23
3	22,26	24
4	22,28	25
5	24,22	23
6	24,24	24
7	24,26	25
8	24,28	26
9	26,22	24
10	26,24	25
11	26,26	26
12	26,28	27
13	28,22	25
14	28,24	26
15	28,26	27
16	28,28	28

Since all of the above samples are equally likely, therefore, the probability of each value of $\overline{X}$ is $\frac{1}{16}$. Thus, we can write the sampling distribution of $\overline{X}$ as given below:

$\overline{X}$	22	23	24	25	26	27	28	*Total*
$p(\overline{X})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	1

The mean of $\overline{X}$, i.e.,

$$\mu_{\overline{X}} = E(X) = 22\times\frac{1}{16}+23\times\frac{2}{16}+24\times\frac{3}{16}+25\times\frac{4}{16}+26\times\frac{3}{16}+$$
$$27\times\frac{2}{16}+28\times\frac{1}{16} = 25$$

Further, $S.E.(\overline{X}) = \sigma_{\overline{X}} = \sqrt{E(\overline{X}^2)-\left[E(\overline{X})\right]^2}$, where

$$E(\overline{X}^2) = \frac{1}{16}\left(22^2+23^2\times 2+24^2\times 3+25^2\times 4+26^2\times 3+27^2\times 2+28^2\right)$$
$$= 627.5$$

Thus, $\sigma_{\overline{X}} = \sqrt{627.5-25^2} = \sqrt{2.5}$ which is equal to $\frac{\sigma}{\sqrt{n}}$.

(b) When random samples of size 2 are drawn without replacement, we have ${}^4C_2$ samples,

Notes

shown below:

Sample No.	*Sample Values*	$\overline{X}$
1	22, 24	23
2	22, 26	24
3	22, 28	25
4	24, 26	25
5	24, 28	26
6	26, 28	27

Since all the samples are equally likely, the probability of each value of $\overline{X}$ is $\frac{1}{6}$. Thus, we can write the sampling distribution of $\overline{X}$ as

$\overline{X}$	23	24	25	26	27	*Total*
$p(\overline{X})$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

Further, $\mu_{\overline{X}} = E(\overline{X}) = \frac{1}{6}[23+24+25\times 2+26+27] = 25.$

To find $S.E.(\overline{X})$, we first find $E(\overline{X}^2)$ given by

$$E(\overline{X}^2) = \frac{1}{6}\left[23^2+24^2+2\times 25^2+26^2+27^2\right] = \frac{3760}{6} = 626.67.$$

Thus, $\sigma_{\overline{X}} = \sqrt{626.67-25^2} = \sqrt{1.67} = 1.292.$

Alternatively, $\sigma_{\overline{X}} = \sqrt{\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}} = \sqrt{\frac{4-2}{3}\times\frac{5}{2}} = \sqrt{1.67} = 1.292.$

8.2.1 Nature of the Sampling Distribution of Mean

It can be deduced that when a random sample X_1, X_2 ...... X_n is obtained from a normal population with mean *m* and standard deviation *s*, then each of the X_i's are also distributed normally with mean *m* and standard deviation *s*.

By the use of additive (or reproductive) property of normal distribution, it follows that the distribution of $\overline{X}$, a linear combination of X_1, X_2 ...... X_n, will also be normal. As shown in the previous section, the mean and standard error of the distribution would be *m* and $\frac{\sigma}{\sqrt{n}}$ respectively.

Remarks: Since normal population is often a large population, the fpc is always taken equal to unity.

The nature of the sampling distribution of $\overline{X}$, when parent population is not normal, is provided by Central Limit Theorem. This theorem states that:

If X_1, X_2 ...... X_n is a random sample of size n from a non-normal population of size N with mean *m* and standard deviation *s*, then the sampling distribution of $\overline{X}$ will approach normal distribution with mean m and standard error $\frac{\sigma}{\sqrt{n}}\left(or\sqrt{\frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}}\right)$ as n becomes larger and larger.

Remarks: As a general rule, when n ³ 30, the sampling distribution of $\overline{X}$ is taken to be normal for practical purposes.

Application of the Sampling Distribution

Decisions by various government and non-government agencies are made on the basis of sample results. For example, a sales manager may take a sample of quantities purchased of its product to predict sales. A government agency may take a sample of residents to assess the effect of a certain welfare program etc. Thus, in order to draw reliable conclusions, we must have a sound knowledge regarding the sample. An extremely common and quite useful knowledge about the sample is given by the sampling distribution of the relevant statistic.

An important application of sampling distribution is to determine the probability of the statistic lying in a given interval.

8.2.2 Sampling Distribution of the Difference Between two Sample Means

Let there be two populations of sizes N_1 and N_2, means m_1 and m_2 and standard deviations s_1 and s_2 respectively. Let $\overline{X}_1$ be the mean of the random sample of size n_1 obtained from the first population and $\overline{X}_2$ be the mean of the random sample of size n_2 obtained from the second population. Thus, we can regard $\overline{X}_1$ and $\overline{X}_2$ as two independent random variables with means m_1 and m_2 and standard errors as

$$\frac{\sigma_1}{\sqrt{n_1}}\left(or\sqrt{\frac{N_1-n_1}{N_1-1}\cdot\frac{\sigma_1^2}{n_1}}\right) \text{ and } \frac{\sigma_2}{\sqrt{n_2}}\left(or\sqrt{\frac{N_2-n_2}{N_2-1}\cdot\frac{\sigma_2^2}{n_2}}\right) \text{ respectively.}$$

Further, their difference, $\overline{X}_1 - \overline{X}_2$, will also be a random variable with mean $= E\left(\overline{X}_1 - \overline{X}_2\right) = E\left(\overline{X}_1\right) - E\left(\overline{X}_2\right) = m_1 - m_2$ and standard error

$$= \sqrt{Variance\left(\overline{X}_1 - \overline{X}_2\right)} = \sqrt{Var\left(\overline{X}_1\right) + Var\left(\overline{X}_2\right)}$$

$$= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \text{ when both the samples are drawn using srswr) or}$$

$$= \sqrt{\frac{N_1-n_1}{N_1-1}\cdot\frac{\sigma_1^2}{n_1} + \frac{N_2-n_2}{N_2-1}\cdot\frac{\sigma_2^2}{n_2}} \text{ (when both the samples are drawn using srswor).}$$

Remarks:

1. When both the populations are normal, then $\overline{X}_1 - \overline{X}_2$ will be distributed normally with mean $m_1 - m_2$ and standard error $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

2. Using Central Limit Theorem, the above result will also hold for a non-normal population when both n_1 and $n_2 > 30$ and fpc is approximately equal to unity, i.e., $n_i < 0.05\ N_i$ (for i = 1, 2).

8.3 Sampling Distribution of the Number of Successes

Let p denote the proportion of successes in population, i.e.,

$$\pi = \frac{Number\ of\ successes\ in\ population}{Total\ number\ of\ units\ in\ population}$$

Let us take a random sample of n units from this population and let X denote the number of successes in the sample. Thus, X is a random variable with mean np and standard error

$$\sqrt{n\pi(1-\pi)}\ \left(or\sqrt{\frac{N-n}{N-1}\cdot n\pi(1-\pi)}\right)$$

If sampling is done with replacement, then X is a binomial variate with mean np and standard error $\sqrt{n\pi(1-\pi)}$. Using central limit theorem, we can say that the distribution of the number of successes will approach a normal variate with mean np and standard error $\sqrt{n\pi(1-\pi)}$ or $\sqrt{\frac{N-n}{N-1}\cdot n\pi(1-\pi)}$ for sufficiently large sample. The sample size is said to be sufficiently large if both n π and n(1 - π) are greater than 5.

8.3.1 Sampling Distribution of Proportion of Successes

Let $p = \frac{X}{n}$ be the proportion of successes in sample. Since X is a random variable, therefore, p is also a random variable with mean

$$E(p) = \frac{E(X)}{n} = \frac{n\pi}{n} = \pi \text{ and standard error}$$

$$= \sqrt{\frac{1}{n^2}Var(X)} = \sqrt{\frac{n\pi(1-\pi)}{n^2}} = \sqrt{\frac{\pi(1-\pi)}{n}} \text{ (when srswr)}$$

or
$$= \sqrt{\frac{N-n}{N-1}\cdot\frac{\pi(1-\pi)}{n}} \text{ (when srswor)}$$

As in the previous section, the sampling distribution of p will also be normal if both n π and n(1 - π) are greater than 5.

Example 2: There are 500 mangoes in a basket out of which 80 are defective. If obtaining a defective mango is termed as a success, determine the mean and standard error of the proportion of successes in a random sample of 10 mangoes, drawn (a) with replacement and (b) without replacement.

Solution.

Notes

It is given that $\pi = \frac{80}{500} = \frac{4}{25}$. Therefore, $E(p) = \pi = \frac{4}{25}$ and

(a) S.E.(p) $= \sqrt{n\pi(1-\pi)} = \sqrt{10 \times \frac{4}{25} \times \frac{21}{25}} = 1.159$ (srswr)

(b) S.E.(p) $= \sqrt{\frac{500-80}{499} \times 10 \times \frac{4}{25} \times \frac{21}{25}} = 1.063$ (srswor)

Example 3: 20% under graduates of a large university are found to be smokers. A sample of 100 students is selected at random. Construct the sampling distribution of the number of smokers. Also find the probability that the number of smokers in the sample is greater than 25.

Solution.

It is given that $\pi = \frac{20}{100} = \frac{1}{5}$. Since sample size, n = 100, is large, the number of successes X will be distributed normally with mean $100 \times \frac{1}{5} = 20$ and standard error $\sqrt{100 \times \frac{1}{5} \times \frac{4}{5}} = 4$.

Further, $P(X > 25) = P\left(z > \frac{25-20}{4}\right) = P(z > 1.25) = 0.1056.$

8.3.2 Sampling Distribution of the Difference of two Proportions

Let p_1 be proportion of successes in a random sample of size n_1 from a population with proportion of successes = p_1 and p_2 be the proportion of successes in a random sample of size n_2 from second population with proportion of successes = p_2. Assuming that the sample sizes are large, we can write

$$p_1 \sim N\left(\pi_1, \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}}\right) \text{ and } p_2 \sim N\left(\pi_2, \sqrt{\frac{\pi_2(1-\pi_2)}{n_2}}\right)$$

Thus, their difference (p_1 - p_2) will be distributed normally with mean = π_1 - π_2 and standard error

$$\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}.$$

Note: The above result will hold when we ignore fpc and the sample size, n_1 and n_2, is greater than 5 divided by the minimum of π_1, $(1 - \pi_1)$, π_2 and $(1 - \pi_2)$.

Some other Sampling Distributions

We have seen that the sampling distributions of mean and proportion of successes are normal.

Apart from normal distribution, there are certain other probability distributions that are useful in sampling theory. These distributions are:

1. Chi - square $\left(\chi^2\right)$ distribution.
2. Student's t - distribution.
3. Snedecor's F - distribution.

8.4 Summary

- Let P_1, P_2 ...... P_N denote the observations on N units of a population and X_1, X_2 ...... X_n be a simple random sample of size n from it.

  A parameter is a measure computed from the observation of the population. For example:

  Population Mean $(\mu) = \dfrac{P_1 + P_2 + \cdots\cdots + P_N}{N}$,

  Population Variance $\left(\sigma^2\right) = \dfrac{1}{N}\sum\left(P_i - \mu\right)^2$, etc. are parameters.

  In a similar way, a statistics is a measure computed from the observations of a sample. For example:

  Sample Mean $\left(\overline{X}\right) = \dfrac{X_1 + X_2 + \cdots\cdots + X_n}{n}$,

  Sample Variance $\left(S^2\right) = \dfrac{1}{n}\sum\left(X_i - \overline{X}\right)^2$, etc. are statistic.

- The standard deviation of a statistic is termed as standard error. The standard error of $\overline{X}$, to be written in abbreviated form as $S.E.\left(\overline{X}\right)$, is equal to $\dfrac{\sigma}{\sqrt{n}}$, when sampling is with replacement and it is equal to $\dfrac{\sigma}{\sqrt{n}} \cdot \sqrt{\dfrac{N-n}{N-1}}$, when sampling is without replacement.

- $S.E.\left(\overline{X}\right)$ is inversely related to the sample size.

- The term $\sqrt{\dfrac{N-n}{N-1}}$ is termed as finite population correction (fpc). We note that fpc tends to become closer and closer to unity as population size becomes larger and larger.

8.5 Keywords

Theoretical probability: A theoretical probability distribution is constructed on the basis of the specification of the conditions of a random experiment.

Parameter: A parameter is any function of population values while a statistic is a function of sample values.

8.6 Self Assessment

1. State whether the following statements are True or False:

   (i) Mean of the sample means is equal to population mean.

   (ii) Random variable of a sampling distribution is called a statistic.

   (iii) The sampling distribution of $\overline{X}$ is normal if the drawn samples are of size 20.

   (iv) When population is large, the finite population correction (fpc) is negligible, i.e., approximately equal to zero.

   (v) In order that a statistic t follows a t - distribution, the sample should have been obtained from a normal population.

8.7 Review Questions

1. Explain the concept of sampling distribution of a statistics.

2. Find the mean and standard error of sample mean in (a) Simple random sampling with replacement, (b) Simple random sampling without replacement.

3. Distinguish between:

   (a) Parameter and Statistic.

   (b) Sampling distribution and Probability distribution.

   (c) Standard deviation and Standard error.

4. (a) Distinguish between sampling with replacement and sampling without replacement. How many random samples of size n can be drawn from a population consisting N items if the sampling is done (i) with replacement, (ii) without replacement?

   (b) What is the variance of the sample mean if sampling is done (i) with replacement (ii) without replacement?

   (c) Under what conditions do the answers in (b) approach each other?

5. If X_i (i = 1, 2, ... n) are n independent normal variates with respective mean μ_i and standard deviation σ_i, then show that the variate $u = \sum X_i$ is normally distributed with mean $\sum_{mi}$ and variance $\sum \sigma_i^2$.

Answers: Self Assessment

1. (i) T (ii) T (iii) F (iv) F (v) T

8.8 Further Readings

Books

Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.

Richa Nandra, Lovely Professional University

Unit 9: Chi - Sqaure (χ^2) Distribution

CONTENTS

Objectives

After studying this unit, you will be able to:

- Discuss Chi - Square (χ^2) Distribution
- Describe some examples related to Chi - Square

Introduction

When sampling is done with replacement, each unit of the population has a probability of its selection equal to $\frac{1}{N}$. Further, there are N^n possible samples that are equally likely, and therefore, the probability of selection of each sample is $\frac{1}{N^n}$.

When sampling is done without replacement, the units are either drawn one by one, without replacement, or all the n units are selected in one attempt. We know that the permutations of N objects taking n at a time is ${}^{N}P_{n}$ and this becomes the number of ordered samples. Corresponding to this, the number of unordered samples will be ${}^{N}C_{n}$, each with probability $\frac{1}{{}^{N}C_{n}}$. In this case also, the probability of selection of a unit at any draw is $\frac{1}{N}$. For example, the probability of selection of a unit at the first draw = $\frac{1}{N}$, the probability of its selection at the second draw is $\frac{N-1}{N} \times \frac{1}{N-1} = \frac{1}{N}$ and so on, the probability of its selection at the rth draw is

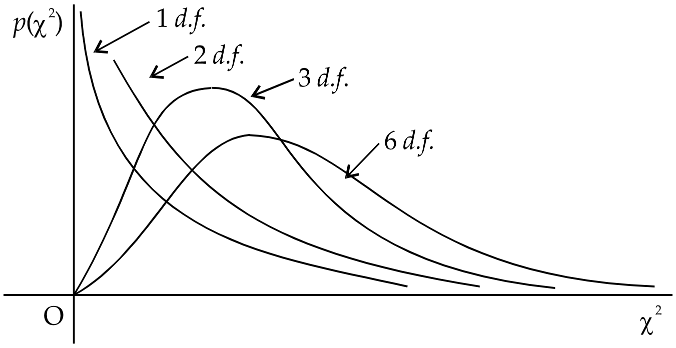
$$\frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdot \;\ldots\ldots\; \frac{N-r+1}{N-r+2} \cdot \frac{1}{N-r+1} = \frac{1}{N}$$

9.1 Chi - Square Distribution

We know that if X is a random variate distributed normally with mean *m* and standard deviation *s*, then $z = \frac{X - \mu}{\sigma}$ is a standard normal variate. Square of z, i.e., $z^2 = \frac{(X - \mu)^2}{\sigma^2}$ if distributed as χ^2 - variate with one degree of freedom and is written as χ_1^2. Further, the value of χ_1^2, a squared value, will lie between 0 to ∞, for z lying between - ∞ to ∞. Since most of the z-values are close to zero, the probability density of χ^2 will be highest near zero. The χ^2 distribution with one degree of freedom is shown in Figure 25.1.

Generalising the above result, we can say that if X_1, X_2 ...... X_n are n independent normal variates each with mean m_i and standard deviations σ_i, i = 1, 2, ... n, respectively, then the sum of squares $\sum z_i^2 = \sum \frac{(X_i - \mu_i)^2}{\sigma_i^2}$ is a χ^2 variate with n degrees of freedom, i.e., χ_n^2. Thus, we can say that χ_n^2 is sum of squares of n independent standard normal variate.

Figure 9.1



Features of χ^2 Distribution

1. The distribution has only one parameter, i.e., number of degrees of freedom or d.f. (in abbreviated form) which is a positive integer.
2. We may note that as the d.f. increases, the height of the probability density function decreases. The distribution is positively skewed and the skewness decreases asd.f. increases. For large values of d.f., the distribution approaches normal distribution. The curves for various d.f. are shown in figure 20.1.
3. The mean of χ_n^2, i.e., $E\left(\chi_n^2\right) = n$ and its variance = 2n, where n = d.f.

4. Additive property

The sum of two independent χ^2 variates is also a χ^2 variate with degrees of freedom equal to the sum of their individual degrees of freedom.

If χ_n^2 and χ_m^2 are two independent random variates with n and m degrees of freedom respectively, then $\chi_n^2 + \chi_m^2$ is also a χ^2 variate with n + m degrees of freedom.

Remarks:

1. The degrees of freedom is defined as the number of independent random variables. If n is the number of variables and k is the number of restrictions on them, the degrees of freedom are said to be n - k.
2. On the basis of the definition of degrees of freedom, given above, we can say that $\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2$ is a χ^2 variate with (n - 1) degrees of freedom. It may be pointed out here that one degree of freedom is reduced because for a given value of $\overline{X}$, the number of independent variables is (n - 1).

9.1.1 Sampling Distribution of Variance

Using χ^2-distribution, we can construct the sampling distribution of $S^2 = \frac{1}{n}\sum\left(X_i - \overline{X}\right)^2$.

Let X_1, X_2 ...... X_n be a random sample of size n from a normal population with mean *m* and variance s^2. We can write

$$X_i - \mu = \left(X_i - \overline{X}\right) + \left(\overline{X} - \mu\right)$$

Squaring both sides and taking sum over all the n observations, we get

$$\sum_{i=1}^{n}\left(X_i - \mu\right)^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + \sum_{i=1}^{n}\left(\overline{X} - \mu\right)^2 + 2\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(\overline{X} - \mu\right)$$

$$= \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + n\left(\overline{X} - \mu\right)^2 + 2\left(\overline{X} - \mu\right)\sum_{i=1}^{n}\left(X_i - \overline{X}\right)$$

We note that the last term is zero. Therefore, we have

$$\sum_{i=1}^{n}\left(X_i - \mu\right)^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + n\left(\overline{X} - \mu\right)^2$$

Dividing both sides by s^2, we get

$$\frac{\sum_{i=1}^{n}\left(X_i - \mu\right)^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sigma^2} + \frac{n\left(\overline{X} - \mu\right)^2}{\sigma^2}$$

or $$\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\sigma^2}=\frac{\sum_{i=1}^{n}\left(X_i-\mu\right)^2}{\sigma^2}-\frac{\left(\bar{X}-\mu\right)^2}{\sigma^2/n}=\chi_n^2-\chi_1^2=\chi_{n-1}^2$$

Thus, $\dfrac{\sum\left(X_i-\bar{X}\right)^2}{\sigma^2}$ *or* $\dfrac{nS^2}{\sigma^2}$ is a χ^2-variate with (n - 1) d.f.

Mean and Standard Error of S^2

Since the random variable $\dfrac{nS^2}{\sigma^2}$ is a χ^2-variate with (n - 1) d.f.,

therefore $E\left[\dfrac{nS^2}{\sigma^2}\right]=n-1$ *or* $\dfrac{n}{\sigma^2}E\left(S^2\right)=n-1$.

Thus, we have $E\left(S^2\right)=\dfrac{n-1}{n}\cdot\sigma^2$

Further, if we define $s^2=\dfrac{1}{n-1}\sum\left(X_i-\bar{X}\right)^2$ so that $s^2=\dfrac{n}{n-1}\cdot S^2$, we have

$E\left(s^2\right)=\dfrac{n}{n-1}\cdot E\left(S^2\right)=\dfrac{n}{n-1}\cdot\dfrac{n-1}{n}\cdot\sigma^2=\sigma^2$ (See Remarks 2 below).

To find variance of S^2, we make use of the fact that variance of $\dfrac{nS^2}{\sigma^2}$ is 2(n - 1). This implies that

$$E\left[\frac{nS^2}{\sigma^2}-(n-1)\right]^2=2(n-1)\ \textit{or}\ \frac{n^2}{\sigma^4}E\left(S^2-\frac{n-1}{n}\cdot\sigma^2\right)^2=2(n-1)$$

$$\therefore\ E\left[S^2-E\left(S^2\right)\right]^2=\frac{2(n-1)}{n^2}\cdot\sigma^4\ \textit{or}\ Var\left(S^2\right)=\frac{2(n-1)}{n^2}\cdot\sigma^4$$

Further, variance of s^2 = variance of $\left(\dfrac{n}{n-1}\cdot S^2\right)$. This gives

$$Var\left(s^2\right)=\frac{n^2}{(n-1)^2}\cdot Var\left(S^2\right)=\frac{n^2}{(n-1)^2}\times\frac{2(n-1)}{n^2}\cdot\sigma^4=\frac{2}{n-1}\cdot\sigma^4$$

Remarks:

1. The distributions of c^2 and S^2 are based upon the assumption that the parent population is normal. If the parent population is not normal, it is not possible to comment upon the nature of the distribution of the above statistics.

2. It will be discussed in the following chapter that when expected value of a statistic equals the value of parameter, it is said to be an unbiased estimate of the parameter.

Problem 1

The Acme Battery Company has developed a new cell phone battery. On average, the battery lasts 60 minutes on a single charge. The standard deviation is 4 minutes.

Suppose the manufacturing department runs a quality control test. They randomly select 7 batteries. The standard deviation of the selected batteries is 6 minutes. What would be the chi-square statistic represented by this test?

Solution

We know the following:

- The standard deviation of the population is 4 minutes.
- The standard deviation of the sample is 6 minutes.
- The number of sample observations is 7.

To compute the chi-square statistic, we plug these data in the chi-square equation, as shown below.

$x^2 = [(n - 1) * s^2] / \sigma^2$

$x^2 = [(7 - 1) * 6^2] / 4^2 = 13.5$

where x^2 is the chi-square statistic, n is the sample size, s is the standard deviation of the sample, and σ is the standard deviation of the population.

Problem 2

Let's revisit the problem presented above. The manufacturing department ran a quality control test, using 7 randomly selected batteries. In their test, the standard deviation was 6 minutes, which equated to a chi-square statistic of 13.5.

Suppose they repeated the test with a new random sample of 7 batteries. What is the probability that the standard deviation in the new test would be greater than 6 minutes?

Solution

We know the following:

- The sample size *n* is equal to 7.
- The degrees of freedom are equal to *n* - 1 = 7 - 1 = 6.
- The chi-square statistic is equal to 13.5 (see Example 1 above).

Given the degrees of freedom, we can determine the cumulative probability that the chi-square statistic will fall between 0 and any positive value. To find the cumulative probability that a chi-square statistic falls between 0 and 13.5, insert the values in formula then result is the cumulative probability: 0.96.

This tells us that the probability that a standard deviation would be less than or equal to 6 minutes is 0.96. This means (by the subtraction rule) that the probability that the standard deviation would be *greater than* 6 minutes is 1 – 0.96 or .04.

9.2 Summary

- We know that if X is a random variate distributed normally with mean *m* and standard deviation *s*, then $z = \frac{X-\mu}{\sigma}$ is a standard normal variate. Square of z, i.e., $z^2 = \frac{(X-\mu)^2}{\sigma^2}$ if distributed as χ^2 - variate with one degree of freedom and is written as χ_1^2. Further, the value of χ_1^2, a squared value, will lie between 0 to ∞, for z lying between - ∞ to ∞. Since most of the z-values are close to zero, the probability density of χ^2 will be highest near zero. The χ^2 distribution with one degree of freedom is shown in Figure 25.1.

  Generalising the above result, we can say that if X_1, X_2 ...... X_n are n independent normal variates each with mean m_i and standard deviations σ_i, i = 1, 2, ... n, respectively, then the sum of squares $\sum z_i^2 = \sum \frac{(X_i - \mu_i)^2}{\sigma_i^2}$ is a χ^2 variate with n degrees of freedom, i.e., χ_n^2.

  Thus, we can say that χ_n^2 is sum of squares of n independent standard normal variate.

- Since the random variable $\frac{nS^2}{\sigma^2}$ is a χ^2 -variate with (n - 1) d.f.,

  therefore $E\left[\frac{nS^2}{\sigma^2}\right] = n-1 \; or \; \frac{n}{\sigma^2}E\left(S^2\right) = n-1$.

  Thus, we have $E\left(S^2\right) = \frac{n-1}{n}\cdot\sigma^2$

  Further, if we define $s^2 = \frac{1}{n-1}\sum\left(X_i - \overline{X}\right)^2$ so that $s^2 = \frac{n}{n-1}\cdot S^2$, we have

  $E\left(s^2\right) = \frac{n}{n-1}\cdot E\left(S^2\right) = \frac{n}{n-1}\cdot\frac{n-1}{n}\cdot\sigma^2 = \sigma^2$ (See Remarks 2 below).

9.3 Keywords

Standard normal variate: if X is a random variate distributed normally with mean *m* and standard deviation *s*, then $z = \frac{X-\mu}{\sigma}$ is a standard normal variate.

Distribution: The distribution has only one parameter, i.e., number of degrees of freedom or d.f. (in abbreviated form) which is a positive integer.

9.4 Self Assessment

1. Fill in the blanks:

   (i) The positive square root of variance of a sampling distribution is known as ....

   (ii) The standard error of $\overline{X}$ varies ...... with standard deviation and ...... with sample size.

   (iii) The sampling distribution of proportion would be approximately normal when n is greater than or equal to ....

   (iv) The mean and variance of a χ^2-variate depend upon its ....

9.5 Review Questions

1. We are given the fact that 30% of all patients admitted to a medical clinic fail to pay their bills and the bills are eventually forgiven. If the clinic treats 2000 different patients over a period of one year, what is the expected number of bills that would have to be forgiven. If X is the number of forgiven bills in the group of 2000 patients, find the variance and standard deviation of X. What can you say about the probability that X will exceed 700?

2. A random sample of 10 observations is to be taken from a normal population with variance equal to 16. What is the probability of obtaining a sample with variance greater than 20?

   Hint : Use χ^2 - distribution.

3. Two independent random samples of sizes 15 and 12 are taken from a normal population. Find the probability that the ratio of their variances is greater than 3. Assume that the variance of the sample of size 15 is greater than the variance of the other.

Answers: Self Assessment

1. (i) standard error (ii) directly, inversely (iii) 50 (iv) parameter

9.6 Further Readings

Books

Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.

Unit 10 : T - Distributions

CONTENTS

Objectives

After studying this unit, you will be able to:

- Discuss T - Distribution
- Explain example of T - Distribution

Introduction

In last unit you have studied about chi-square. This unit will provide you information related to T - Distribution.

10.1 The Student's T-Distribution

Let X_1, X_2 ...... X_n be n independent random variables from a normal population with mean *m* and standard deviation *s* (unknown).

When *s* is not known, it is estimated by s, the sample standard deviation $\left(s = \sqrt{\frac{1}{n-1} \sum \left(X_i - \bar{X} \right)^2} \right)$.

In such a case we would like to know the exact distribution of the statistic $\frac{\bar{X} - \mu}{s / \sqrt{n}}$ and the answer to this is provided by t - distribution.

W.S. Gosset defined t statistic as $t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$ which follows t - distribution with (n - 1) degrees of freedom.

Features of t- distribution

1. Like χ^2 - distribution, t - distribution also has one parameter n = n - 1, where n denotes sample size. Hence, this distribution is known if n is known.
2. Mean of the random variable t is zero and standard deviation is $\sqrt{\dfrac{\nu}{\nu-2}}$, for n > 2.
3. The probability curve of t - distribution is symmetrical about the ordinate at t = 0. Like a normal variable, the t variable can take any value from - ∞ to ∞.
4. The distribution approaches normal distribution as the number of degrees of freedom become large.
5. The random variate t is defined as the ratio of a standard normal variate to the square root of χ^2 - variate divided by its degrees of freedom.

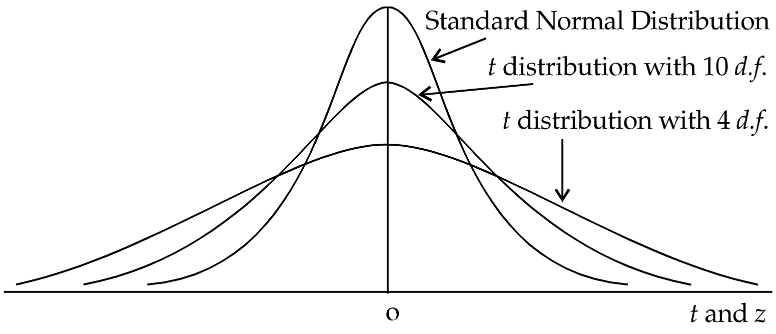
To show this we can write $t = \dfrac{\bar{X}-\mu}{s/\sqrt{n}} = \dfrac{\left(\bar{X}-\mu\right)\sqrt{n}}{s}$

Dividing numerator and denominator by s, we get

$$t = \frac{\dfrac{\left(\bar{X}-\mu\right)\sqrt{n}}{\sigma}}{\dfrac{s}{\sigma}} = \frac{\dfrac{\left(\bar{X}-\mu\right)}{\sigma/\sqrt{n}}}{\sqrt{s^2/\sigma^2}} = \frac{\dfrac{\left(\bar{X}-\mu\right)}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{1}{n-1}\cdot\dfrac{\sum\left(X_i-\bar{X}\right)^2}{\sigma^2}}}$$

$$= \frac{\dfrac{\left(\bar{X}-\mu\right)}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{\chi^2_{n-1}}{n-1}}} = \frac{\textit{Standard Normal Variate}}{\sqrt{\chi^2\textit{-variate}}}$$

Figure 10.1



10.2 T test

Acme Corporation manufactures light bulbs. The CEO claims that an average Acme light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days. If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?

> *Note* There are two ways to solve this problem, using the T Distribution Calculator. Both approaches are presented below. Solution A is the traditional approach. It requires you to compute the t score, based on data presented in the problem description. Then, you use the T Distribution Calculator to find the probability. Solution B is easier. You simply enter the problem data into the T Distribution Calculator. The calculator computes a t score "behind the scenes", and displays the probability. Both approaches come up with exactly the same answer.

Solution A

The first thing we need to do is compute the t score, based on the following equation:

$$t = [\ x - \mu\ ] \ / \ [\ s \ / \ \text{sqrt}(\ n\ )\ ]$$

$$t = (\ 290 - 300\ ) \ / \ [\ 50 \ / \ \text{sqrt}(\ 15)\ ] = -10 \ / \ 12.909945 = -\ 0.7745966$$

where x is the sample mean, ì is the population mean, s is the standard deviation of the sample, and n is the sample size.

Now, we are ready to use the T Distribution Calculator. Since we know the t score, we select "T score" from the Random Variable dropdown box. Then, we enter the following data:

- The degrees of freedom are equal to 15 – 1 = 14.
- The t score is equal to – 0.7745966.

The calculator displays the cumulative probability: 0.226. Hence, if the true bulb life were 300 days, there is a 22.6% chance that the average bulb life for 15 randomly selected bulbs would be less than or equal to 290 days.

Solution B:

This time, we will work directly with the raw data from the problem. We will not compute the t score; the T Distribution Calculator will do that work for us. Since we will work with the raw data, we select "Sample mean" from the Random Variable dropdown box. Then, we enter the following data:

- The degrees of freedom are equal to 15 - 1 = 14.
- Assuming the CEO's claim is true, the population mean equals 300.
- The sample mean equals 290.
- The standard deviation of the sample is 50.

The calculator displays the cumulative probability: 0.226. Hence, there is a 22.6% chance that the average sampled light bulb will burn out within 290 days.

Problem 2

Suppose scores on an IQ test are normally distributed, with a mean of 100. Suppose 20 people are randomly selected and tested. The standard deviation in the sample group is 15. What is the probability that the average test score in the sample group will be at most 110?

Solution:

To solve this problem, we will work directly with the raw data from the problem. We will not compute the t score; the T Distribution Calculator will do that work for us. Since we will work with the raw data, we select "Sample mean" from the Random Variable dropdown box. Then, we enter the following data:

- The degrees of freedom are equal to 20 - 1 = 19.
- The population mean equals 100.
- The sample mean equals 110.
- The standard deviation of the sample is 15.

We enter these values into the T Distribution Calculator. The calculator displays the cumulative probability: 0.996. Hence, there is a 99.6% chance that the sample average will be no greater than 110.

10.3 Summary

- Let X_1, X_2 ...... X_n be n independent random variables from a normal population with mean *m* and standard deviation *s* (unknown).

  When *s* is not known, it is estimated by s, the sample standard deviation $\left(s = \sqrt{\frac{1}{n-1}\sum\left(X_i - \overline{X}\right)^2}\right)$. In such a case we would like to know the exact distribution of

  the statistic $\frac{\overline{X} - \mu}{s / \sqrt{n}}$ and the answer to this is provided by t - distribution.

  W.S. Gosset defined t statistic as $t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$ which follows t - distribution with (n - 1) degrees of freedom.

- Like χ^2 - distribution, t - distribution also has one parameter *n* = n - 1, where n denotes sample size. Hence, this distribution is known if *n* is known.
- Mean of the random variable t is zero and standard deviation is $\sqrt{\frac{\nu}{\nu - 2}}$, for *n* > 2.
- The probability curve of t - distribution is symmetrical about the ordinate at t = 0. Like a normal variable, the t variable can take any value from - ∞ to ∞.
- The distribution approaches normal distribution as the number of degrees of freedom become large.
- The random variate t is defined as the ratio of a standard normal variate to the square root of χ^2 - variate divided by its degrees of freedom.

10.4 Keywords

Mean: Mean of the random variable t is zero and standard deviation is $\sqrt{\frac{\nu}{\nu-2}}$, for $n > 2$.

T - distribution: The probability curve of t - distribution is symmetrical about the ordinate at t = 0. Like a normal variable, the t variable can take any value from - ∞ to ∞.

10.4 Self Assessment

1. State whether the following statements are true or false:

   (i) Both, t and χ^2 distributions depend only one parameter.

   (ii) Total number of samples of size 4, with replacement, from a population of 15 units is 1365.

   (iii) F - statistic is equal the ratio of two χ^2 variates.

   (iv) In sampling with replacement if N = n, the standard error of $\overline{X}$ is equal to zero.

   (v) χ^2 - distribution depends upon two parameters.

10.5 Review Questions

1. A population consists of 4 families consisting of 2, 3, 4 and 5 children. By considering all possible random samples of size two, with replacement, find mean and standard error of $\overline{X}$. Show that S.E. of $\overline{X}$ depends upon the sample size.

2. If X_1, X_2, X_3 is a simple random sample of size three from a large population with mean 5 and variance 4, evaluate the expected value and standard error of the statistics T = ($2X_1 + X_2 - 3X_3$).

3. If X_1, X_2 and X_3 constitute a random sample of size 3 from a normal population with mean μ and the variance σ^2, find the efficiency of $\frac{X_1 + 2X_2 + X_3}{4}$ relative to $\frac{X_1 + X_2 + X_3}{3}$.

4. The diameter of a component produced on a semi-automatic machine is known to be distributed normally with mean of 10 mm. and a standard deviation of 0.1 mm. If we pick up a random sample of size 25, what is the probability that the sample mean will be between 9.95 and 10.05 mm?

5. It is known that 10% of the bolts manufactured by a factory are defective. If a random sample of 100 bolts is chosen at random from a day's production, construct the sampling distribution of (i) the number of defective bolts, (ii) the proportion of defective bolts.

6. The mean and standard deviation of per capita consumption of wheat in rural and urban areas of Delhi are estimated to be 450 gms, 75 gms and 410 gms, 100 gms respectively. Assuming that the per capita consumption of wheat is distributed normally, construct the sampling distribution of the difference between two sample means obtained from random samples of sizes 80 and 60 from rural and urban populations respectively.

Answers: Self Assessment

Notes

1. (i) T (ii) F (iii) T (iv) F (v) T

10.6 Further Readings

Books

Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.

Notes

Unit 11: F-distribution

CONTENTS

Objectives

After studying this unit, you will be able to:

- Define F - distribution
- Discuss F - distribution examples

Introduction

In the last unit you studied about samples distribution and T - Distribution. This unit provides you information related to F - distribution.

11.1 Snedecor's F- Distribution

Let there be two independent random samples of sizes n_1 and n_2 from two normal populations with variances s_1^2 and s_2^2 respectively. Further, let $s_1^2 = \frac{1}{n_1 - 1}\sum\left(X_{1i} - \bar{X}_1\right)^2$ and $s_2^2 = \frac{1}{n_2 - 1}\sum\left(X_{2i} - \bar{X}_2\right)^2$ be the variances of the first sample and the second samples respectively.

Then F - statistic is defined as the ratio of two χ^2 - variates. Thus, we can write

$$F = \frac{\frac{\chi^2_{n_1-1}}{n_1-1}}{\frac{\chi^2_{n_2-1}}{n_2-1}} = \frac{\frac{(n_1-1)s_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)s_2^2}{\sigma_2^2}/(n_2-1)} = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$$

Features of F- distribution

1. This distribution has two parameters n_1 (= n_1 - 1) and n_2 (= n_2 - 1).

2. The mean of F - variate with n_1 and n_2 degrees of freedom is $\frac{\nu_2}{\nu_2 - 2}$ and standard error is

$$\left(\frac{\nu_2}{\nu_2 - 2}\right)\sqrt{\frac{2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 4)}}.$$

   We note that the mean will exist if $\nu_2 > 2$ and standard error will exist if $\nu_2 > 4$. Further, the mean > 1.

3. The random variate F can take only positive values from 0 to ∞. The curve is positively skewed, as shown in Fig. 20.3

4. For large values of ν_1 and ν_2, the distribution approaches normal distribution. This behaviour is shown in the following figure.

5. If a random variate follows t-distribution with ν degrees of freedom, then its square follows F-distribution with 1 and n d.f. i.e. $t^2_\nu = F_{1,\nu}$

6. F and c^2 are also related as $F_{\nu_1, \nu_2} = \frac{(\chi^2_{\nu_1})}{\nu_1}$ as $\nu_2 \rightarrow \infty$

Example 1: Suppose you randomly select 7 women from a population of women, and 12 men from a population of men. The table below shows the standard deviation in each sample and in each population.

Population	Population standard deviation	Sample standard deviation
Women	30	35
Men	50	45

Compute the f statistic.

Solution A: The f statistic can be computed from the population and sample standard deviations, using the following equation:

$$f = [\, s_1^2/\sigma_1^2 \,] \,/\, [\, s_2^2/\sigma_2^2 \,]$$

where σ_1 is the standard deviation of population 1, s_1 is the standard deviation of the sample drawn from population 1, σ_2 is the standard deviation of population 2, and s_1 is the standard deviation of the sample drawn from population 2.

As you can see from the equation, there are actually two ways to compute an f statistic from these data. If the women's data appears in the numerator, we can calculate an f statistic as follows:

$$f = (\, 35^2 \,/\, 30^2 \,) \,/\, (\, 45^2 \,/\, 50^2 \,) = (1225 \,/\, 900) \,/\, (2025 \,/\, 2500) = 1.361 \,/\, 0.81 = 1.68$$

For this calculation, the numerator degrees of freedom ν_1 are 7 - 1 or 6; and the denominator degrees of freedom ν_2 are 12 - 1 or 11.

On the other hand, if the men's data appears in the numerator, we can calculate an f statistic as follows:

$$f = (\, 45^2 \,/\, 50^2 \,) \,/\, (\, 35^2 \,/\, 30^2 \,) = (2025 \,/\, 2500) \,/\, (1225 \,/\, 900) = 0.81 \,/\, 1.361 = 0.595$$

For this calculation, the numerator degrees of freedom v_1 are 12 – 1 or 11; and the denominator degrees of freedom v_2 are 7 – 1 or 6.

When you are trying to find the cumulative probability associated with an f statistic, you need to know v_1 and v_2. This point is illustrated in the next example.

Example 2: Find the cumulative probability associated with each of the f statistics from Example 1, above.

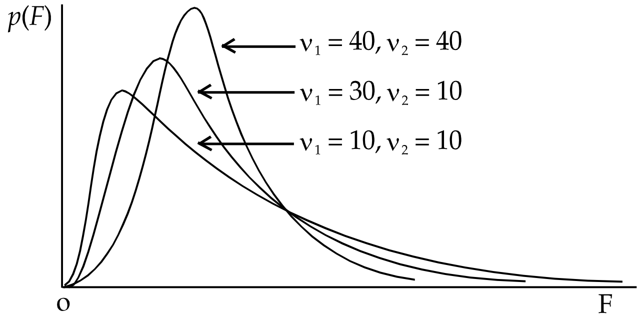
Solution: To solve this problem, we need to find the degrees of freedom for each sample. Then, we will use the F Distribution Calculator to find the probabilities.

- The degrees of freedom for the sample of women is equal to n – 1 = 7 – 1 = 6.
- The degrees of freedom for the sample of men is equal to n – 1 = 12 – 1 = 11.

Therefore, when the women's data appear in the numerator, the numerator degrees of freedom v_1 is equal to 6; and the denominator degrees of freedom v_2 is equal to 11. And, based on the computations shown in the previous example, the f statistic is equal to 1.68. We plug these values into the F Distribution Calculator and find that the cumulative probability is 0.78.

On the other hand, when the men's data appear in the numerator, the numerator degrees of freedom v_1 is equal to 11; and the denominator degrees of freedom v_2 is equal to 6. And, based on the computations shown in the previous example, the f statistic is equal to 0.595. We plug these values into the F Distribution Calculator and find that the cumulative probability is 0.22.

Figure 28.3



11.2 Summary

- Let there be two independent random samples of sizes n_1 and n_2 from two normal populations with variances s_1^2 and s_2^2 respectively. Further, let $s_1^2 = \frac{1}{n_1 - 1}\sum\left(X_{1i} - \bar{X}_1\right)^2$ and $s_2^2 = \frac{1}{n_2 - 1}\sum\left(X_{2i} - \bar{X}_2\right)^2$ be the variances of the first sample and the second samples respectively. Then F - statistic is defined as the ratio of two χ^2 - variates. Thus, we can write

$$F = \frac{\dfrac{\chi^2_{n_1-1}}{n_1 - 1}}{\dfrac{\chi^2_{n_2-1}}{n_2 - 1}} = \frac{\dfrac{(n_1 - 1)s_1^2}{\sigma_1^2}\Big/(n_1 - 1)}{\dfrac{(n_2 - 1)s_2^2}{\sigma_2^2}\Big/(n_2 - 1)} = \frac{\dfrac{s_1^2}{\sigma_1^2}}{\dfrac{s_2^2}{\sigma_2^2}}$$

- This distribution has two parameters n_1 (= n_1 - 1) and n_2 (= n_2 - 1).

- The mean of F - variate with n_1 and n_2 degrees of freedom is $\frac{\nu_2}{\nu_2 - 2}$ and standard error is

  $$\left(\frac{\nu_2}{\nu_2-2}\right)\sqrt{\frac{2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)}}.$$

  We note that the mean will exist if $\nu_2 > 2$ and standard error will exist if $\nu_2 > 4$. Further, the mean > 1.

- The random variate F can take only positive values from 0 to ∞. The curve is positively skewed, as shown in Fig. 20.3

- For large values of ν_1 and ν_2, the distribution approaches normal distribution. This behaviour is shown in the following figure.

- If a random variate follows t-distribution with ν degrees of freedom, then its square follows F-distribution with 1 and n d.f. i.e. $t^2_\nu = F_{1,\nu}$

11.3 Keywords

The random variate: The random variate F can take only positive values from 0 to ∞. The curve is positively skewed.

F-distribution: If a random variate follows t-distribution with ν degrees of freedom, then its square follows F-distribution with 1 and n d.f. i.e. $t^2_\nu = F_{1,\nu}$

11.4 Self Assessment

1. Fill in the Blanks:

   (1) The mean and standard error of a F-variate depend upon its ...... parameters.

   (ii) The sum of squares of standard normal variates is a ...... variate.

   (iii) If N = 8 and n = 3, the number of samples without replacement is equal to ....

   (iv) The ratio of two sample variances follows F - distribution when the variances of their parent population are ....

   (v) In sampling without replacement if N = n, the standard error of $\overline{X}$ is equal to ....

   (vi) Both χ^2 and F-distributions are ...... skewed distributions.

11.5 Review Questions

1. Define F distribution and discuss feature of F distribution.

2. Two random samples of sizes 100 and 150 are drawn from two different normal populations. Find mean and standard error of the statistic F.

3. Suppose you randomly select 8 women from a population of women, and 10 men from a population of men. The table below shows the standard deviation in each sample and in each population.

Population	Population standard deviation	Sample standard deviation
Women	40	45
Men	60	35

Compute the f statistic.

4. Find the cumulative probability associated with each of the f statistics. Suppose you randomly select 6 women from a population of women, and 10 men from a population of men. The table below shows the standard deviation in each sample and in each population.

Population	Population standard deviation	Sample standard deviation
Women	30	35
Men	50	45

5. Suppose you randomly select 8 women from a population of women, and 12 men from a population of men. The table below shows the standard deviation in each sample and in each population.

Population	Population standard deviation	Sample standard deviation
Women	30	35
Men	60	35

Compute the f statistic.

Answers: Self Assessment

1. (i) two (ii) χ^2 (iii) 56 (iv) equal (v) zero (vi) positively.

11.6 Further Readings

Books

Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.

Unit 12: Estimation of Parameters: Criteria for Estimates

Notes

CONTENTS

Objectives

After studying this unit, you will be able to:

- Discuss Theory of Estimation
- Explain Point Estimation (Properties of Good Estimators)
- Describe Interval Estimation

Introduction

Estimation: It is a procedure by which sample information is used to estimate the numerical magnitude of one or more parameters of the population. A function of sample values is called an estimator (or statistic) while its numerical value is called an estimate. For example is an estimator of population mean m. On the other hand if for a sample, the estimate of population mean is said to be 50.

12.1 Theory of Estimation

Let X be a random variable with probability density function (or probability mass function) f(X ; θ_1, θ_2, ... θ_k), where θ_1, θ_2, ... θ_k are k parameters of the population.

Given a random sample X_1, X_2, ... X_n from this population, we may be interested in estimating one or more of the k parameters θ_1, θ_2, ... θ_k. In order to be specific, let X be a normal variate so that its probability density function can be written as N(X : μ, σ). We may be interested in estimating m or s or both on the basis of random sample obtained from this population.

It should be noted here that there can be several estimators of a parameter, e.g., we can have any of the sample mean, median, mode, geometric mean, harmonic mean, etc., as an estimator of population mean μ. Similarly, we can use either $S = \sqrt{\frac{1}{n}\sum\left(X_i - \bar{X}\right)^2}$ or $s = \sqrt{\frac{1}{n-1}\sum\left(X_i - \bar{X}\right)^2}$ as an estimator of population standard deviation s. This method of estimation, where single statistic like Mean, Median, Standard deviation, etc. is used as an estimator of population parameter, is known as Point Estimation. Contrary to this it is possible to estimate an interval in which the value of parameter is expected to lie. Such a procedure is known as Interval Estimation. The estimated interval is often termed as Confidence Interval.

12.2 Point Estimation

As mentioned above, there can be more than one estimators of a population parameter. Therefore, it becomes necessary to determine a good estimator out of a number of available estimators. We may recall that an estimator, a function of random variables X_1, X_2, ... X_n, is a random variable. Therefore, we can say that a good estimator is one whose distribution is more concentrated around the population parameter. R. A. Fisher has given the following properties of a good estimators. These are:

(i) Unbiasedness (ii) Consistency (iii) Efficiency (iv) Sufficiency.

12.2.1 Unbiasedness

An estimator t (X_1, X_2, ... X_n) is said to be an unbiased estimator of a parameter q if E(t) = θ.

If E(t) ≠ q, then t is said to be a biased estimator of θ. The magnitude of bias = E(t) – θ. We have seen in § 20.2 that $E\left(\bar{X}\right) = \mu$, therefore, $\bar{X}$ is said to be an unbiased estimator of population mean m. Further, refer to § 20.4.1, we note that $E\left(S^2\right) = \frac{n-1}{n} \cdot \sigma^2$, where $S^2 = \frac{1}{n}\sum\left(X_i - \bar{X}\right)^2$. Therefore, S^2 is a biased estimator of s^2. The magnitude of bias $= \left(\frac{n-1}{n} - 1\right)\sigma^2 = -\frac{1}{n}\sigma^2$.

Contrary to this, if we define $s^2 = \frac{1}{n-1}\sum\left(X_i - \bar{X}\right)^2$, we have seen in § 20.4.1 that E(s^2) = σ^2. Thus, s^2 is an unbiased estimator of s^2. Also from § 20.3.1 we note that E(p) = π, therefore, p is an unbiased estimator of π.

12.2.2 Consistency

It is desirable to have an estimator, with a probability distribution, that comes closer and closer to the population parameter as the sample size is increased. An estimator possessing this property is called a consistent estimator. An estimator $t_n(X_1, X_2, ... X_n)$ is said to be consistent if its probability distribution converges to θ as n → ∞.

Symbolically, we can write $P(t_n \to \theta) = 1$ as $n \to \infty$. Alternatively, t_n is said to be a consistent estimator of q if $E(t_n) \to q$ and $Var(t_n) \to 0$, as $n \to \infty$.

We may note that $\overline{X}$ is a consistent estimator of population mean m because $E(\overline{X}) = \mu$ and

$Var(\overline{X}) = \frac{\sigma^2}{n} \to 0$ as $n \to \infty$.

Note: An unbiased estimator is necessarily a consistent estimator.

12.2.3 Efficiency

Let t_1 and t_2 be two estimators of a population parameter *q* such that both are either unbiased or consistent. To select a good estimator, from t_1 and t_2, we consider another property that is based upon its variance.

If t_1 and t_2 are two estimators of a parameter *q* such that both of them are either unbiased or consistent, then t_1 is said to be more efficient than t_2 if $Var(t_1) < Var(t_2)$. The efficiency of an estimator is measured by its variance.

For a normal population, we know that both the sample mean and median are unbiased estimator of population mean. However, their respective variances are $\frac{\sigma^2}{n}$ and $\frac{\pi}{2} \cdot \frac{\sigma^2}{n}$, where σ^2 is population variance. Since $\frac{\sigma^2}{n} < \frac{\pi}{2} \cdot \frac{\sigma^2}{n}$, therefore, sample mean is said to be efficient estimator of population mean.

Remarks: The precision of an estimator = 1/ S. E. of estimator.

An estimator having minimum variance among all the estimators of a population parameter is termed as Most Efficient Estimator or Best Estimator. If an estimator is unbiased and best, then it is termed as Best Unbiased Estimator. Further, if the best unbiased estimator is a linear function of the sample observations, it is termed as Best Linear Unbiased Estimator (BLUE). It may be pointed out here that sample mean is best linear unbiased estimator of population mean.

Cramer Rao Inequality:

This inequality gives the minimum possible value of the variance of an unbiased estimator. If t is an unbiased estimator of parameter *q* of a continuous population with probability density function f(X, *q*), then

$$Var(t) \geq \frac{1}{nE\left(\frac{\partial \log f(X,\theta)}{\partial \theta}\right)^2}$$

12.2.4 Sufficiency

An estimator t is said to be a sufficient estimator of parameter θ if it utilises all the information given in the sample about θ. For example, the sample mean $\overline{X}$ is a sufficient estimator of μ because no other estimator of μ can add any further information about μ.

Let X$_1$, X$_2$, ... X$_n$ be a random sample of n independent observations from a population with p.d.f. (or p.m.f.) given by f(X; θ$_1$, θ$_2$), where q$_1$ and q$_2$ are two parameters. The joint probability distribution of X$_1$, X$_2$, ... X$_n$, denoted by L(X; θ$_1$, θ$_2$) is given by :

$$L(X; \theta_1, \theta_2) = f(X_1; \theta_1, \theta_2) \times f(X_2; \theta_1, \theta_2) \times \ldots\ldots \times f(X_n; \theta_1, \theta_2)$$

An estimator t is said to be sufficient for q_1 if the conditional p.d.f. (or p.m.f.) of X$_1$, X$_2$, ... X$_n$ given t is independent of q_1, i.e.,

$$\frac{f(X_1;\theta_1,\theta_2)\times f(X_2;\theta_1,\theta_2)\times \;....\; \times f(X_n;\theta_1,\theta_2)}{g(t,\theta_1)} = h(X_1, X_2, \;....\; X_n)$$, where g(t, q_1) is p.d.f.

(or p.m.f.) of t and h is a function of sample values that is independent of θ$_1$. We may note that each of the functions g(t, θ$_1$) and h(X$_1$, X$_2$, ... X$_n$) may or may not be function of θ$_2$.

Alternatively, we can write the sufficiency condition as

f(X$_1$; θ$_1$, θ$_2$) × f(X$_2$; θ$_1$, θ$_2$) × ...... × f(X$_n$; θ$_1$, θ$_2$) = g(t, q$_1$) × h(X$_1$, X$_2$, ... X$_n$), which implies that if the joint p.d.f. (or p.m.f.) of X$_1$, X$_2$, ... X$_n$ can be written as a function of t and θ$_1$ multiplied by a function independent of θ$_1$, then t is sufficient estimator of θ$_1$.

Sufficient estimators are the most desirable but are not very commonly available. The following points must be noted about sufficient estimators:

1. A sufficient estimator is always consistent.
2. A sufficient estimator is most efficient if an efficient estimator exists.
3. A sufficient estimator may or may not be unbiased.

Example 1: If X$_1$, X$_2$, ... X$_n$ is a sample of n independent observations from a normal population with mean *m* and variance *s*^{2}, show that $\overline{X}$ is a sufficient estimator of *m* but $S^2 = \frac{1}{n}\sum\left(X_i - \overline{X}\right)^2$ is not sufficient estimator of *s*^{2}.

Solution.

The probability density function of a normal variate is given by

$$f(X;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(X-\mu)^2}$$

Thus, the joint probability density function of X$_1$, X$_2$, ... X$_n$ is given by

$$f(X_1;\mu,\sigma)\times f(X_2;\mu,\sigma)\times \;....\; \times f(X_n;\mu,\sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(X_i-\mu)^2}$$

We can write $X_i - \mu = \left(X_i - \overline{X}\right) + \left(\overline{X} - \mu\right)$.

Squaring both sides and taking sum over n observations, we get

$$\sum(X_i-\mu)^2 = \sum\left(X_i-\overline{X}\right)^2 + \sum\left(\overline{X}-\mu\right)^2 + 2\sum\left(X_i-\overline{X}\right)\left(\overline{X}-\mu\right)$$

$$= \sum\left(X_i-\overline{X}\right)^2 + n\left(\overline{X}-\mu\right)^2 + 2\left(\overline{X}-\mu\right)\sum\left(X_i-\overline{X}\right)$$

$$= \sum\left(X_i - \bar{X}\right)^2 + n\left(\bar{X} - \mu\right)^2 \qquad \text{(last term is zero)}$$

$$= nS^2 + n\left(\bar{X} - \mu\right)^2$$

Therefore, we can write $-\frac{1}{2\sigma^2}\sum\left(X_i - \mu\right)^2 = -\frac{n}{2\sigma^2}S^2 - \frac{n}{2\sigma^2}\left(\bar{X} - \mu\right)^2$.

Hence $f\left(X_1; \mu, \sigma\right) \times f\left(X_2; \mu, \sigma\right) \times \; .... \; \times f\left(X_n; \mu, \sigma\right)$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{n}{2\sigma^2}S^2 - \frac{n}{2\sigma^2}\left(\bar{X}-\mu\right)^2} = e^{-\frac{n}{2\sigma^2}\left(\bar{X}-\mu\right)^2} \times \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{n}{2\sigma^2}S^2}$$

$$= g\left(\bar{X}, \mu, \sigma\right) \times h\left(S^2, \sigma\right)$$

Since h is independent of m, therefore $\bar{X}$ is a sufficient estimator of μ. However, S^2 is not sufficient estimator of σ^2 because g is not independent of σ.

Further, if we define $S^2 = \frac{1}{n}\sum\left(X_i - \mu\right)^2$, then

$$f\left(X_1; \mu, \sigma\right) \times f\left(X_2; \mu, \sigma\right) \times \; .... \; \times f\left(X_n; \mu, \sigma\right) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{n}{2\sigma^2}S^2}$$

Thus, the newly defined S^2 becomes a sufficient estimator of σ^2. We note that h(X$_1$, X$_2$, ... X$_n$) = 1 in this case.

The above result suggests that if *m* is known, then we should use $S^2 = \frac{1}{n}\sum\left(X_i - \mu\right)^2$ rather than $S^2 = \frac{1}{n}\sum\left(X_i - \bar{X}\right)^2$ because former is better estimator of *s*^{2}.

12.2.5 Methods of Point Estimation

Given various criteria of a good estimator, the next logical step is to obtain an estimator possessing some or all of the above properties.

There are several methods of obtaining a point estimator of the population parameter. For example, we can use the method of maximum likelihood, method of least squares, method of minimum variance, method of minimum χ^2, method of moments, etc. We shall, however, use the most popular method of maximum likelihood.

Method of Maximum Likelihood

Let X$_1$, X$_2$, ... X$_n$ be a random sample of n independent observations from a population with probability density function (or p.m.f.) f(X; θ), where θ is unknown parameter for which we desire to find an estimator.

Since X$_1$, X$_2$, ... X$_n$ are independent random variables, their joint probability function or the probability of obtaining the given sample, termed as likelihood function, is given by

$$L = f(X_1 ; \theta) . f(X_2 ; \theta) . ... f(X_n ; \theta) = \prod_{i=1}^{n} f\left(X_i ; \theta\right).$$

We have to find that value of *q* for which L is maximum. The conditions for maxima of L are : $\frac{dL}{d\theta} = 0$ and $\frac{d^2L}{d\theta^2} < 0.$ The value of *q* satisfying these conditions is known as Maximum Likelihood Estimator (MLE).

Generalising the above, if L is a function of k parameters $\theta_1, \theta_2, ... \theta_k$, the first order conditions for maxima of L are: $\frac{\partial L}{\partial \theta_1} = \frac{\partial L}{\partial \theta_1} = \;......\; \frac{\partial L}{\partial \theta_k} = 0$.

This gives k simultaneous equations in k unknowns $\theta_1, \theta_2, ... \theta_k$, and can be solved to get k maximum likelihood estimators.

Sometimes it is convenient to work using logarithm of L. Since log L is a monotonic transformation of L, the maxima of L and maxima of log L occur at the same value.

Properties of Maximum Likelihood Estimators

1. The maximum likelihood estimators are consistent.
2. The maximum likelihood estimators are not necessarily unbiased. If a maximum likelihood estimator is biased, then by slight modifications it can be converted into an unbiased estimator.
3. If a maximum likelihood estimator is unbiased, then it will also be most efficient.
4. A maximum likelihood estimator is sufficient provided sufficient estimator exists.
5. The maximum likelihood estimators are invariant under functional transformation, i.e., if t is a maximum likelihood estimator of θ, then f(t) would be maximum likelihood estimator of f(θ).

Example 2: Obtain a maximum likelihood estimator of *p* (the proportion of successes) in a population with p.m.f. given by $f(X;\pi) = {}^{n}C_X \pi^X (1-\pi)^{n-X}$, where X denotes the number of successes in a sample of n trials.

Solution.

Since ${}^{n}C_X \pi^X (1-\pi)^{n-X}$ is the probability of X successes out of n trials, therefore, this is also the likelihood function. Thus, we can write $L = {}^{n}C_X \pi^X (1-\pi)^{n-X}$.

Taking logarithm of both sides, we get

$$\log L = \log{}^{n} C_X + X \log \pi + (n - X) \log (1 - \pi)$$

Differentiating w.r.t. *p*, we get

$$\frac{d \log L}{d\pi} = 0 + \frac{X}{\pi} - \frac{n-X}{1-\pi} = 0 \quad \text{for maxima of L.}$$

or $\quad X(1- p) - (n - X)p = 0$

This gives $\hat{\pi} = \frac{X}{n}$, where $\hat{\pi}$ denotes an estimator of *p*.

It can also be shown that $\frac{d^2 \log L}{d\pi^2} < 0$ when $\hat{\pi} = \frac{X}{n}$.

Example 3: Obtain the maximum likelihood estimator of the parameter m of the Poisson distribution.

Solution.

Let $X_1, X_2, \ldots\ldots X_n$ be a random sample of n independent observations from the given population. Therefore, we can write

$$L = \frac{e^{-m}.m^{X_1}}{X_1!} \times \frac{e^{-m}.m^{X_2}}{X_2!} \times \ \ldots\ldots \ \times \frac{e^{-m}.m^{X_n}}{X_n!} = \frac{e^{-nm}.m^{\sum X_i}}{\prod (X_i!)}$$

Taking logarithm of both sides, we get

$$\log L = -nm + \sum X_i \log m - \sum \log(X_i!)$$

Differentiating w.r.t. m, we get

$$\frac{d \log L}{dm} = -n + \frac{\sum X_i}{m} = 0 \ \Rightarrow \ \hat{m} = \frac{\sum X_i}{n} = \bar{X}$$

Thus, sample mean is MLE of parameter m.

Example 4: For a normal population with parameter μ and σ^2, obtain the maximum likelihood estimators of the parameters.

Solution.

The probability density function of normal distribution is

$$f(X;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(X-\mu)^2}{\sigma^2}}$$

Given a random sample of n independent observations, the likelihood function L is given by

$$L = \prod_{i=1}^{n} f(X_i;\mu,\sigma).$$

Taking logarithm of both sides, we get

$$\log L = \sum \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(X-\mu)^2}{\sigma^2}}\right) = \sum \log\frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2}\sum (X_i - \mu)^2$$

$$= \sum\left(-\log\sigma - \frac{1}{2}\log 2\pi\right) - \frac{1}{2\sigma^2}\sum\left(X_i - \mu\right)^2$$

$$= -\,n\log\sigma - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum\left(X_i - \mu\right)^2 \qquad .... (1)$$

(i) MLE of *m*

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{2\sigma^2}\cdot 2\sum\left(X_i - \mu\right) = 0 \text{ or } \sum\left(X_i - \mu\right) = 0 \;\Rightarrow\; \hat{\mu} = \frac{\sum X_i}{n} = \bar{X}$$

(ii) MLE of s^2

Rewriting equation (1) as a function of s^2, we get

$$\log L = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum\left(X_i - \mu\right)^2$$

$$\therefore\; \frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum\left(X_i - \mu\right)^2}{2\sigma^4} = 0 \text{ or } -n\sigma^2 + \sum\left(X_i - \mu\right)^2 = 0$$

$$\Rightarrow\; \hat{\sigma}^2 = \frac{\sum\left(X_i - \mu\right)^2}{n}$$

12.3 Interval Estimation

Using point estimation, it is possible to provide a single quantity as an estimator of a parameter. Any point estimator, even if it satisfies all the characteristics of a good estimator, has a limitation that it provides no information about the magnitude of errors due to sampling. This problem is taken care of by the method of interval estimation, that gives a range of the estimator of the parameter.

The method of interval estimation is based upon the sampling distribution of an estimator. The standard error of the estimator is used in the construction of an interval so that the probability of the parameter lying within the interval can be specified.

Given a random sample of n observations X_1, X_2, ... X_n, we can find two values l_1 and l_2 such that the probability of population parameter *q* lying between l_1 and l_2 is (say) *h*. Using symbols, we can write P(l_1 £ *q* £ l_2) = *h*.

Such an interval is termed as a Confidence Interval for *q* and the two limits l_1 and l_2 are termed as Confidential or Fiducial Limits. The percentage probability or confidence is termed as the Level of Confidence or Confidence Coefficient of the interval. For example, the level of confidence of the above interval is 100*h*%. The level of confidence implies that if a large number of random samples are taken from a population and confidence intervals are constructed for each, then 100*h*% of these intervals are expected to contain the population parameter *q*. Alternatively, a 100 *h*% confidence interval implies that we are 100 *h*% confident that the population parameter *q* lies between l_1 and l_2.

As compared to point estimation, the interval estimation is better because it takes into account the variability of the estimator in addition to its single value and thus, provides a range of values. Unlike point estimation, interval estimation indicates that estimation is an uncertain process.

The methods of construction of confidence intervals in various situations are explained through the following examples.

Confidence Interval for Population Mean

Example 5: Construct 95% and 99% confidence intervals for mean of a normal population.

Solution.

Let X_1, X_2, ... X_n be a random sample of size n from a normal population with mean m and standard deviation s.

We know that sampling distribution of $\overline{X}$ is normal with mean *m* and standard error $\frac{\sigma}{\sqrt{n}}$.

Therefore, $z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ will be a standard normal variate.

From the tables of areas under standard normal curve, we can write

$$P[- 1.96 \le z \le 1.96] = 0.95 \text{ or } P[- 1.96 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 1.96] = 0.95 \qquad .... (1)$$

The inequality $- 1.96 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ can be written as

$$-1.96\frac{\sigma}{\sqrt{n}} \le \bar{X} - \mu \text{ or } \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}} \qquad .... (2)$$

Similarly, from the inequality $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 1.96$, we can write

$$\mu \ge \bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \qquad .... (3)$$

Combining (2) and (3), we get

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

Thus, we can write equation (1) as

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

This gives us a 95% confidence interval for the parameter m. The lower limit of μ is $\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}$ and the upper limit is $\bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$. The probability of m lying between these limits is 0.95 and therefore, this interval is also termed as 95% confidence interval for μ.

In a similar way, we can construct a 99% confidence interval for *m* as

$$P\left(\bar{X} - 2.58\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 2.58\frac{\sigma}{\sqrt{n}}\right) = 0.99$$

Thus, the 99% confidence limits for *m* are $\bar{X} \pm 2.58\frac{\sigma}{\sqrt{n}}$.

Remarks: When *s* is unknown and n < 30, we use t value instead of 1.96 or 2.58 and use S in place of *s*.

Confidence Interval for Population Proportion

Example 6: Obtain the 95% confidence limits for the proportion of successes in a binomial population.

Solution.

Let the parameter *p* denote the proportion of successes in population. Further, p denotes the proportion of successes in n (≥ 50) trials. We know that the sampling distribution of p will be approximately normal with mean *p* and standard error $\sqrt{\frac{\pi(1-\pi)}{n}}$.

Since *p* is not known, therefore, its estimator p is used in the estimation of standard error of p, i.e., $S.E.(p) = \sqrt{\frac{p(1-p)}{n}}$

Thus, the 95% confidence interval for p is given by

$$P\left(p - 1.96\sqrt{\frac{p(1-p)}{n}} \le \pi \le p + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

This gives the 95% fiducial limits as $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$.

Example 7: In a newspaper article of 1600 words in Hindi, 64% of the words were found to be of Sanskrit origin. Assuming that the simple sampling conditions hold good, estimate the confidence limits of the proportion of Sanskrit words in the writer's vocabulary.

Solution.

Let *p* be the proportion of Sanskrit words in the writer's vocabulary. The corresponding proportion in the sample is given as p = 0.64.

$$\therefore \quad S.E.(p) = \sqrt{\frac{0.64 \times 0.36}{1600}} = \frac{0.48}{40} = 0.012$$

We know that almost whole of the distribution lies between 3*s* limits. Therefore, the confidence interval is given by

$$P[p - 3S.E.(p) \le p \le p + 3\,S.E.(p)] = 0.9973$$

Thus, the 99.73% confidence limits of *p* are 0.604 (= 0.64 - 3 × 0.012) and 0.676 (= 0.64 + 3 × 0.012) respectively.

Hence, the proportion of Sanskrit words in the writer's vocabulary are between 60.4% to 67.6%.

Example 8: A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Estimate the proportion of bad pineapples in the consignment and obtain the standard error of the estimator. Deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8.5 and 17.5.

Solution.

Let p be the proportion of bad pine apples in the large consignment. Its estimate based on the sample is $\hat{p} = \frac{65}{500} = 0.13$ with $S.E.(\hat{p}) = \sqrt{\frac{0.13 \times 0.87}{500}} = 0.015$

Thus, the 99.73% confidence limits of p are $0.13 \pm 3 \times 0.015$, i.e., 0.085 and 0.175. Hence, the proportion of bad pineapples in the given consignment almost certainly lies between 8.5% and 17.5%.

Remarks: The width of a confidence interval can be controlled in two ways:

(i) By adjusting the sample size: More is the sample size the narrower will be the interval.

(ii) By adjusting the level of confidence: Lower the level of confidence the narrower will be the interval.

12.3.1 Determination of an Approximate Sample Size for a Given Degree of Accuracy

Let us assume that we want to find the size of a sample to be taken from the population such that the difference between sample mean and the population mean would not exceed a given value, say Î, with a given level of confidence. In other words, we want to find n such that

$$P\left(\left|\bar{X} - \mu\right| \leq \in\right) = 0.95 \text{ (say)} \qquad .... (1)$$

Assuming that the sampling distribution of $\bar{X}$ is normal with mean *m* and $S.E._{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, we can write

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq 1.96\right) = 0.95 \text{ or } P\left(\left|\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}\right| \leq 1.96\right) = 0.95$$

$$\text{or } P\left(\left|\bar{X} - \mu\right| \leq 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95 \qquad .... (2)$$

Comparing (1) and (2), we get

$$\in = 1.96 \cdot \frac{\sigma}{\sqrt{n}} \text{ or } n = \left(\frac{1.96\sigma}{\in}\right)^2 = \frac{3.84\sigma^2}{\in^2}$$

Remarks:

1. The sample size required with a maximum error of estimation, Î and with a given level of confidence is $n = \frac{z^2\sigma^2}{\in^2}$, where z is the value of standard normal variate for a given level of confidence and σ^2 is the variance of population.

Notes

2. For a given level of confidence and σ², n is inversely related to ∈², the square of the maximum error of estimation. This implies that to reduce $\in$ to $\frac{\in}{k}$, the size of the sample must be k² times the original sample size.

3. The lesser the magnitude of Î, the more precise will be the interval estimate.

Example 9: What should be the sample size for estimating mean of a normal population if the probability that sample mean differs from population mean by not more than 30% of standard deviation is 0.99.

Solution.

Let n be the size of the sample. It is given that

$$P\left(\left|\bar{X}-\mu\right| \leq 0.30\sigma\right)=0.99 \qquad \text{.... (1)}$$

Assuming that the sampling distribution of $\overline{X}$ is normal with mean *m* and $S.E._{\bar{X}}=\frac{\sigma}{\sqrt{n}}$, we can write

$$P\left(\left|\bar{X}-\mu\right| \leq 2.58\frac{\sigma}{\sqrt{n}}\right)=0.99 \text{ (from table of areas)} \qquad \text{.... (2)}$$

Comparing (1) and (2), we get

$$0.30\sigma=2.58\frac{\sigma}{\sqrt{n}} \Rightarrow n=\left(\frac{2.58}{0.30}\right)^2=73.96 \text{ or } 74 \text{ (approx.)}$$

Example 10: A survey of middle class families of Delhi is proposed to be conducted for the estimation of average monthly consumption (in Rs) per family. What should be the size of the sample so that the average consumption is estimated within a range of Rs 300 with 95% level of confidence. It is known that the standard deviation of the consumption in population is Rs 1,600.

Solution.

Let n denote the size of the sample to be drawn. With usual notations, we want to find n such that

$$P\left(\left|\bar{X}-\mu\right| \leq 300\right)=0.95 \qquad \text{.... (1)}$$

Assuming that the sampling distribution of $\overline{X}$ is normal with mean *m* and $S.E._{\bar{X}}=\frac{\sigma}{\sqrt{n}}$, we can write $P\left(\left|\bar{X}-\mu\right| \leq 1.96\frac{\sigma}{\sqrt{n}}\right)=0.95$

or $$P\left(\left|\bar{X}-\mu\right| \leq \frac{1.96\times 1600}{\sqrt{n}}\right)=0.95 \qquad \text{.... (2)}$$

Comparing (1) and (2), we get

$$300 = \frac{1.96 \times 1600}{\sqrt{n}} \quad \text{or} \quad n = \left(\frac{1.96 \times 1600}{300}\right)^2 = 109.3$$

Since this value is greater than 109, therefore, the size of the sample should be 110.

12.3.2 Confidence Interval for Population Standard Deviation

Let $S = \sqrt{\frac{1}{n}\sum\left(X_i - \bar{X}\right)^2}$ be the sample standard deviation of a random sample of size n drawn from a normal population with standard deviation s. It can be shown that the sampling distribution of S is approximately normal, for large values of n, with mean s and standard error $\frac{\sigma}{\sqrt{2n}}$. Thus,

$z = \frac{S - \sigma}{\sigma / \sqrt{2n}}$ can be taken as a standard normal variate.

Example 11: A random sample of 50 observations gave a value of its standard deviation equal to 24.5. Construct a 95% confidence interval for population standard deviation σ.

Solution.

It is given that S = 24.5 and n = 50 (large). We know that $S.E.(S) = \frac{\sigma}{\sqrt{2n}}$. Since s is not known, we use its estimate based on sample. Thus, we can write $S.E.(S) = \frac{S}{\sqrt{2n}} = \frac{24.5}{\sqrt{100}} = 2.45$.

Hence 95% confidence interval for s is given by

$24.5 - 1.96 \times 2.45 \le \sigma \le 24.5 + 1.96 \times 2.45$ or $19.7 \le \sigma \le 29.3$

Note More examples on confidence intervals are given later with the questions on test of significance.

12.4 Summary

- Let X be a random variable with probability density function (or probability mass function) f(X ; θ_1, θ_2, ... θ_k), where θ_1, θ_2, ... θ_k are k parameters of the population.

  Given a random sample X_1, X_2, ... X_n from this population, we may be interested in estimating one or more of the k parameters θ_1, θ_2, ... θ_k. In order to be specific, let X be a normal variate so that its probability density function can be written as N(X : μ, σ). We may be interested in estimating m or s or both on the basis of random sample obtained from this population.

  It should be noted here that there can be several estimators of a parameter, e.g., we can have any of the sample mean, median, mode, geometric mean, harmonic mean, etc., as an estimator of population mean μ. Similarly, we can use either $S = \sqrt{\frac{1}{n}\sum\left(X_i - \bar{X}\right)^2}$ or $s = \sqrt{\frac{1}{n-1}\sum\left(X_i - \bar{X}\right)^2}$ as an estimator of population standard deviation s. This method of estimation, where single statistic like Mean, Median, Standard deviation, etc. is used as an estimator of population parameter, is known as Point Estimation. Contrary to this it is possible to estimate an interval in which the value of parameter is expected to lie. Such a procedure is known as Interval Estimation. The estimated interval is often termed as Confidence Interval.

- The maximum likelihood estimators are consistent.
- The maximum likelihood estimators are not necessarily unbiased. If a maximum likelihood estimator is biased, then by slight modifications it can be converted into an unbiased estimator.
- If a maximum likelihood estimator is unbiased, then it will also be most efficient.
- A maximum likelihood estimator is sufficient provided sufficient estimator exists.
- The maximum likelihood estimators are invariant under functional transformation, i.e., if t is a maximum likelihood estimator of θ, then f(t) would be maximum likelihood estimator of f(θ).

12.5 Keywords

Estimation: It is a procedure by which sample information is used to estimate the numerical magnitude of one or more parameters of the population.

Cramer Rao Inequality: This inequality gives the minimum possible value of the variance of an unbiased estimator.

Estimator: An estimator t is said to be a sufficient estimator of parameter θ if it utilises all the information given in the sample about θ.

12.6 Self Assessment

1. State whether the following statements are True or False:

   (i) Sample mean is an unbiased estimator of population mean.

   (ii) Sample standard deviation is an unbiased estimator of population standard deviation.

   (iii) An estimator whose variance tends to zero as sample size tends to infinity is called a consistent estimator.

   (iv) An efficient estimator may or may not be unbiased.

   (v) A sufficient estimator is always consistent.

   (vi) The width of the confidence interval depends upon the level of significance as well as on the sample size.

12.7 Review Questions

1. A random sample of 400 farms in certain year revealed that the average yield per acre of sugarcane was 925 kgs with a standard deviation of 88 kgs.

   (a) Determine the 95% confidence interval for the population mean.

   (b) What should be the size of the sample if the width of 95% confidence interval estimate of *m* is not more than 15?

   Hint : (a) See example 5, (b) $\in = 15/2$.

2. A random sample of 100 sale receipts of a firm showed that its average sales per customer are Rs 250 with a standard deviation of Rs 50 (assume that there is one receipt for each customer).

   (a) Determine the 99% confidence interval for the mean sales.

   (b) How does the width of the confidence interval change if sample size is 400 instead?

   (c) How many sale receipts should be included in the sample in order that a 98% confidence interval has a maximum error of estimation equal to Rs 10.

   Hint: (a) z = 2.58 and since s is not known, use S as its estimate. (b) Sample size is inversely related to the width of Confidence interval. (c) z = 2.33.

3. A survey revealed that 30% of the persons of a state are suffering from a particular disease. How many persons should be included in the sample so that the maximum width of the 95% confidence interval of proportion of persons suffering from the disease is 0.15 units?

   Hint : $n = \frac{z^2 pq}{\in^2}$.

4. A random sample of size 64 has been drawn from a population with standard deviation 20. The mean of the sample is 80. (i) Calculate 95% confidence limits for the population mean. (ii) How does the width of the confidence interval changes if the sample size is 256 instead?

   Hint : σ is given to be 20.

5. In a random sample of 100 articles taken from a large batch of articles, 10 are found to be defective. Obtain a 95% confidence interval for the true proportion of defectives in the batch.

   Hint : See example 6.

6. A random sample of size 10 from a normal population gives the values 64, 72, 65, 70, 68, 71, 65, 62, 66, 67. If it is known that the standard error of the sample mean is $\sqrt{0.7}$, find 95% confidence limits for the population mean. Also find the population variance.

   Hint : $\frac{\sigma}{\sqrt{n}} = \sqrt{0.7}$.

Answers: Self Assessment

1. (i) T (ii) F (iii) F (iv) T (v) T (vi) T

12.8 Further Readings

Books

Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.

Unit 13: Method of Least Square

Notes

CONTENTS

Objectives

After studying this unit, you will be able to:

- Discuss Method of Least Squares
- Describe Method of Selected Points and Method of Semi-Averages
- Explain Seasonal Variations

Introduction

A series of observations, on a variable, recorded after successive intervals of time is called a time series. The successive intervals are usually equal time intervals, e.g., it can be 10 years, a year, a quarter, a month, a week, a day, an hour, etc. The data on the population of India is a time series data where time interval between two successive figures is 10 years. Similarly figures of national income, agricultural and industrial production, etc., are available on yearly basis.

It should be noted here that the time series data are bivariate data in which one of the variables is time. This variable will be denoted by t. The symbol Yt will be used to denote the observed value, at point of time t, of the other variable. If the data pertains to n periods, it can be written as (t, Yt), t = 1, 2, ... n.

13.1 Method of Least Squares

This is one of the most popular methods of fitting a mathematical trend. The fitted trend is termed as the best in the sense that the sum of squares of deviations of observations, from it, are minimised. We shall use this method in the fitting of following trends:

1. Linear Trend
2. Parabolic Trend
3. Exponential Trend

13.1.1 Fitting of Linear Trend

Given the data (Y_t, t) for n periods, where t denotes time period such as year, month, day, etc., we have to find the values of the two constants, a and b, of the linear trend equation $Y_t = a + bt$.

Using the least square method, the normal equation for obtaining the values of a and b are :

$$\Sigma Y_t = na + b\Sigma t \text{ and}$$

$$\Sigma tY_t = a\Sigma t + b\Sigma t^2$$

Let X = t - A, such that ΣX = 0, where A denotes the year of origin.

The above equations can also be written as

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

(Dropping the subscript t for convenience).

Since SX = 0, we can write $a = \frac{\sum Y}{n}$ and $b = \frac{\sum XY}{\sum X^2}$

Note: The procedure for calculation of the two constants is slightly different for even and odd number of observations. This distinction will become obvious from the following two examples.

Example:

Fit a straight line trend to the following data and estimate the likely profit for the year 1986. Also calculate various trend values.

Year	:	1977	1978	1979	1980	1981	1982	1983
Profit (*in lacs of Rs*)	:	60	72	75	65	80	85	95

Solution.

Calculation Table

Years (t)	Y	$X = t - 1980$	XY	X^2	*Trend Values*
1977	60	-3	-180	9	61.42
1978	72	-2	-144	4	66.28
1979	75	-1	-75	1	71.14
1980	65	0	0	0	76.00
1981	80	1	80	1	80.86
1982	85	2	170	4	85.72
1983	95	3	285	9	90.58
Total	532	0	136	28	

From the table we can write $a = \frac{532}{7} = 76$ (n = 7, the no. of observations)

and $b = \frac{136}{28} = 4.86$

Thus, the fitted line of trend is Y = 76 + 4.86X

Note: It is very important to provide the following details for any trend equations:

(i) The year of origin, (ii) unit of X and (iii) the nature of Y values such as annual figures, monthly figures or monthly averages, quarterly figures or quarterly averages, etc. Thus, the appropriate way of writing the trend equation would be : Y = 76 + 4.86X, where (i) year of origin = 1st July 1980 (the year in which X = 0), (ii) unit of X = 1 year and (iii) Y'^s are annual figures of profits.

Calculation of trend values

Trend value of a particular year is obtained by substituting the associated value of X in the trend equation. For example, X = - 3 for 1977, therefore, trend for 1977 is Y = 76 + 4.86 × (- 3) = 61.42

Alternatively, trend values can be calculated as follows:

We know that a is the trend value in the year of origin and b gives the rate of change per unit of time. Thus, the trend for 1980 = 76, for 1979 = 76 - 4.86 = 71. 14, for 1978 = 71.14 - 4.86 = 66.28 and for 1977 = 66.28 - 4.86 = 61.42, etc. Similarly, trend for 1981 = 76 + 4.86 = 80.86, for 1982 = 80.86 + 4.86 = 85.72, etc.

Prediction of trend for a year

Using the trend equation we can predict a trend value for a year which doesn't belong to the observed data. To predict the value for 1986, the associated value of X = 6. Substituting this in the trend equation we get Y = 76 + 6 × 4.86 = Rs 105.16 lacs.

Remarks: The prediction of trend is only valid for periods that are not too far from the observed data.

Example 8: Fit a straight line trend, by the method of least squares, to the following data. Assuming that the same rate of change continues, what would be the predicted sales for 1993?

Year	:	1987	1988	1989	1990	1991	1992
Sales(*in* '000 *Rs*)	:	15	17	20	21	23	24

Solution.

We note that n is even in the given example.

Calculation Table

$Year\ (t)$	$Sales\ (Y)$	$d = t - 1989.5$	$X = 2d$	XY	X^2	$Trend\ Values$
1987	15	-2.5	-5	-75	25	15.45
1988	17	-1.5	-3	-51	9	17.27
1989	20	-0.5	-1	-20	1	19.09
1990	21	0.5	1	21	1	20.91
1991	23	1.5	3	69	3	22.73
1992	24	2.5	5	120	5	24.55
Total	120		0	64	70	

From the above table, we can write

$$a = \frac{120}{6} = 20 \text{ and } b = \frac{64}{70} = 0.91$$

∴ The fitted trend line is Y = 20 + 0.91 X

Year of origin : Middle of 1989 and 1990 or 1st Jan. 1990

Unit of X : $\frac{1}{2}$ year (Since X changes by 2 units in one year)

Nature of Y values : Annual figures of sales.

Calculation of trend values

Trend for 1989 = 20 - 0.91 = 19.09

Trend for 1988 = 19.09 - 2 × 0.91 = 17.27

Trend for 1990 = 20 + 0.91 = 20.91

Trend for 1991 = 20.91 + 2 × 0.91 = 22.73, etc.

To predict the sales for 1993, we note that X = 7

Thus, the predicted sales = 20 + 7 × 0.91 = Rs 26.37 (thousand).

Shifting of Origin of a Trend Equation

Let Y = a + bX be the equation of linear trend, with 1985 as the year of origin and unit of X equal to 1 year.

To shift origin of the above equation, say to 1990, we proceed as follows : The associated value of X for 1990 is 5. Thus, the trend for 1990 = a + 5b. We know that a linear trend equation is given by Y = trend value in the year of origin + bX. Thus, we can write the trend equation, with origin at 1990, as Y = a + 5b + bX = a + b (X + 5). This implies that the required equation can be obtained by replacing X by X + 5 in the original trend equation.

Similarly, the trend equation with 1984 as origin can be written as Y = a + b (X - 1) = (a - b) + bX.

Further, if the unit of X is given to be half year, the trend equation with 1990 as the year of origin can be written as Y = a + b (X + 10) = (a + 10b) + bX.

Example 9:

Given the following trend equations:

(a) $Y = 50 + 3X$, with 1985 as the year of origin and unit of $X = 1$ year. Shift the origin to 1991.

(b) $Y = 100 + 2.5\ X$, with origin at the middle of 1987 and 1988 and unit of $X = \frac{1}{2}$ year. Shift the origin to (i) 1988 and (ii) 1992.

Solution.

(a) Replacing X by X + 6, in the trend equation, we get

$Y = 50 + 3(X + 6) = 68 + 3X$, the required trend equation.

(b) (i) For shifting origin to 1988 (i.e., middle of 1988), we have to replace X by X + 1. (note that $X = 1$ for $\frac{1}{2}$ year)

$\therefore\ Y = 100 + 2.5(X + 1) = 102.5 + 2.5X$

(ii) Replace X by X + 9, to get the required equation

$\therefore\ Y = 100 + 2.5(X + 9) = 122.5 + 2.5X$

Conversion of Annual Trend Equation into Monthly trend Equation

Usually a trend is fitted to the annual figures because the fitting of a monthly trend is time consuming. However, monthly trend equations are often obtained from annual trend equations.

Let the annual trend equation be $Y = a + bX$, where Y denotes annual figures and the unit of X = 1 year.

To obtain the monthly trend equation, we have to convert the constants a and b into monthly values.

Thus, when a denotes an annual value, $\frac{a}{12}$ would give the value of the corresponding constant for the monthly equation.

Further, the value of b denotes the annual change in Y per unit of X, i.e., per year. Therefore $\frac{b}{12}$ would be the monthly (average) change in Y per year. Thus, the equation $Y = \frac{a}{12} + \frac{b}{12}X$, denotes a monthly average equation, where Y denotes monthly average for the year and unit of X = 1 year.

In a similar way, the value $\frac{b}{12 \times 12} = \frac{b}{144}$ would denote the monthly change in Y per month. Thus, $Y = \frac{a}{12} + \frac{b}{144}X,$ is the monthly trend equation, where Y denotes monthly figures and the unit of X = 1 month.

A quarterly trend equation can also be obtained in a similar way. We can write $Y = \frac{a}{4} + \frac{b}{4}X,$ as the quarterly average equation and $Y = \frac{a}{4} + \frac{b}{16}X,$ as the quarterly trend equation.

Example 10: The equation for yearly sales (in '000 Rs) of a commodity with 1st July, 1971, as origin is Y = 91.6 + 28.8X.

(i) Determine the trend equation to give monthly trend values with 15th January, 1972, as origin.

(ii) Calculate the trend values for March, 1972 to August, 1972.

Solution.

(i) The monthly trend equation with 1st July, 1971, as origin is given by $Y = \frac{91.6}{12} + \frac{28.8}{144}X$ = 7.63 + 0.2X, where unit of X is one month.

To shift the origin to 15th January, 1972, we replace X by X + 6.5 in the above equation. Note that the associated value of X for 15th January, 1972, is 6.5. Thus, the required equation is Y = 7.63 + 0.2(X + 6.5) = 8.93 + 0.2X

(ii) Calculation of trend values

Trend value for March, 1972 = 8.93 + 0.2 × 2 = Rs 9.33

Trend value for April, 1972 = 9.33 + 0.2 = Rs 9.53

Trend value for May, 1972 = 9.53 + 0.2 = Rs 9.73

Trend value for June, 1972 = 9.73 + 0.2 = Rs 9.93

Trend value for July, 1972 = 9.93 + 0.2 = Rs 10.13

Trend value for August, 1972 = 10.13 + 0.2 = Rs 10.33.

Example 11:

Convert the following into annual trend equation :

Y = 350 + 3X with origin = I - II Quarter, 1986, unit of X = one quarter and Y denotes quarterly production.

Solution.

Important Note : To convert a quarterly (or monthly) equation into an annual equation, it is necessary to first shift the origin to the middle of the year.

In the given example, since the middle of the year lies a quarter ahead, we shall replace X by X + 1 in the above equation. Thus, the quarterly equation with middle of the year as origin is Y = 350 + 3(X +1) = 353 + 3X.

Then, the annual trend equation can be written as

$$Y = 353 \times 4 + 3 \times 16X = 1412 + 48X$$

Notes

Example 12: Convert the following annual trend equation, for the production of cloth in a factory, into monthly average equation and predict the monthly averages for 1988 and 1989.

Y = 96 + 7.2X, with origin = 1986, unit of X = 1 year and Y denotes annual cloth production in '000 metres.

Solution.

The average monthly equation is given by

$Y = \frac{96}{12} + \frac{7.2}{12}X$ = 8 + 0.6 X, where origin = 1986, unit of X = 1 year and Y denotes monthly average production in the year.

The predicted values of Y are 8 + 0.6 × 2 = 9.2 thousand metres for 1988 and 9.2 + 0.6 = 9.8 thousand metres for 1989.

13.1.2 Fitting of Parabolic Trend

The mathematical form of a parabolic trend is given by $Y_t = a + bt + ct^2$ or $Y = a + bt + ct^2$ (dropping the subscript for convenience). Here a, b and c are constants to be determined from the given data.

Using the method of least squares, the normal equations for the simultaneous solution of a, b, and c are :

$$\Sigma Y = na + b\Sigma t + c\Sigma t^2$$

$$\Sigma tY = a\Sigma t + b\Sigma t^2 + c\Sigma t^3$$

$$\Sigma t^2 Y = a\Sigma t^2 + b\Sigma t^3 + c\Sigma t^4$$

By selecting a suitable year of origin, i.e., define X = t - origin such that ΣX = 0, the computation work can be considerably simplified. Also note that if ΣX = 0, then ΣX^3 will also be equal to zero. Thus, the above equations can be rewritten as:

$$SY = na + cSX^2 \qquad .... \text{(i)}$$

$$SXY = bSX^2 \qquad .... \text{(ii)}$$

$$SX^2Y = aSX^2 + cSX^4 \qquad .... \text{(iii)}$$

From equation (ii), we get $b = \frac{\sum XY}{\sum X^2}$ .... (iv)

Further, from equation (i), we get $a = \frac{\sum Y - c\sum X^2}{n}$ .... (v)

And from equation (iii), we get $c = \frac{n\sum X^2 Y - \left(\sum X^2\right)\left(\sum Y\right)}{n\sum X^4 - \left(\sum X^2\right)^2}$ .... (vi)

Thus, equations (iv), (v) and (vi) can be used to determine the values of the constants a, b and c.

Example 13: Fit a parabolic trend Y = a + bt + ct^2 to the following data, where t denotes years and Y denotes output (in thousand units).

t	:	1981	1982	1983	1984	1985	1986	1987	1988	1989
Y	:	2	6	7	8	10	11	11	10	9

Also compute the trend values. Predict the value for 1990.

Solution.

Calculation Table

t	Y	$X = t - 1985$	XY	X^2Y	X^2	X^3	X^4	*Trend Values*
1981	2	−4	−8	32	16	−64	256	2.28
1982	6	−3	−18	54	9	−27	81	5.02
1983	7	−2	−14	28	4	−8	16	7.22
1984	8	−1	−8	8	1	−1	1	8.88
1985	10	0	0	0	0	0	0	10.00
1986	11	1	11	11	1	1	1	10.58
1987	11	2	22	44	4	8	16	10.62
1988	10	3	30	90	9	27	81	10.12
1989	9	4	36	144	16	64	256	9.08
Total	74	0	51	411	60	0	708	

From the above table, we can write

$$b = \frac{51}{60} = 0.85$$

$$c = \frac{9 \times 411 - 60 \times 74}{9 \times 708 - (60)^2} = -0.27$$

$$a = \frac{74 - (-0.27) \times 60}{9} = 10.0$$

∴ The fitted trend equation is Y = 10.0 + 0.85X – 0.27X^2,

with origin = 1985 and unit of X = 1 year.

Various trend values are calculated by substituting appropriate values of X in the above equation. These values are shown in the last column of the above table.

The predicted value for 1990 is given by

$$Y = 10.0 + 0.85 \times 5 - 0.27 \times 25 = 7.5$$

Example 14: The prices of a commodity during 1981-86 are given below. Fit a second degree parabola to the following data. Calculate the trend values and estimate the price of the commodity in 1986.

Year	:	1981	1982	1983	1984	1985	1986
Price	:	110	114	120	138	152	218

Solution.

Calculation Table

Year (t)	*Price* (Y)	$X = 2(t - 1983.5)$	XY	X^2Y	X^2	X^4	*Trend Values*
1981	110	-5	-550	2750	25	625	114.40
1982	114	-3	-342	1026	9	81	109.12
1983	120	-1	-120	120	1	1	116.08
1984	138	1	138	138	1	1	135.28
1985	152	3	456	1368	9	81	166.72
1986	218	5	1090	5450	25	625	210.40
	852	0	672	10852	70	1414	

From the above table, we get

$$b = \frac{672}{70} = 9.6\,,\; c = \frac{6 \times 10852 - 70 \times 852}{6 \times 1414 - (70)^2} = 1.53 \text{ and } a = \frac{852 - 1.53 \times 70}{6} = 124.15$$

$\therefore$ The equation of parabolic trend is $Y = 124.15 + 9.6X + 1.53X^2$, with year of origin = 1983.5 or 1st January, 1984 and the unit of X = $\frac{1}{2}$ year.

The calculated trend values are shown in the last column of the above table.

The price of the commodity in 1986 is obtained by substituting X = 5, in the above equation.

Thus, $Y = 124.15 + 9.5 \times 5 + 1.53 \times 25 = 210.4$

13.1.3 Fitting of Exponential Trend

The general form of an exponential trend is $Y = a.b^t$, where a and b are constants to be determined from the observed data.

Taking logarithms of both sides, we have $\log Y = \log a + t \log b$.

This is a linear equation in log Y and t and can be fitted in a similar way as done in case of linear trend. Let A = log a and B = log b, then the above equation can be written as log Y = A + Bt.

The normal equations, based on the principle of least squares are

$\Sigma \log Y = nA + B\,\Sigma t$

and $\Sigma t \log Y = A\Sigma t + B\,\Sigma t^2$.

By selecting a suitable origin, i.e., defining X = t - origin, such that SX = 0, the computation work can be simplified. The values of A and B are given by $A = \frac{\sum \log Y}{n}$ and $B = \frac{\sum X \log Y}{\sum X^2}$ respectively.

Thus, the fitted trend equation can be written as log Y = A + BX

or $Y = \text{Antilog}\,[A + BX] = \text{Antilog}\,[\log a + X \log b]$

$= \text{Antilog}\,[\log a.b^X] = a.b^X$.

Example 15:

Fit a simple exponential trend to the following data and calculate the trend values. Also estimate the trend for 1992.

Year	:	1985	1986	1987	1988	1989
Sales (*Rs Crores*)	:	100	105	112	120	130

Solution.

Calculation Table

Year (*t*)	*Sales* (*Y*)	$X = t - 1987$	*logY*	*XlogY*	X^2	*log of Trend Values*	*Trend Values*
1985	100	−2	2.0000	−4.0000	4	1.9955	98.97
1986	105	−1	2.0212	−2.0212	1	2.0241	105.71
1987	112	0	2.0492	0.0000	0	2.0527	112.90
1988	120	1	2.0792	2.0792	1	2.0813	120.59
1989	130	2	2.1139	4.2278	4	2.1099	128.80
Total		0	10.2635	0.2858	10		

From the above table, we get

$$A = \frac{10.2635}{5} = 2.0527 \quad \text{and} \quad B = \frac{0.2858}{10} = 0.0286$$

Further, a = antilog 2.0527 = 112.90 and b = antilog 0.0286 = 1.07

Thus, the fitted trend equation is $Y = 112.90(1.07)^X$

Origin : 1st July, 1987, unit of X = 1 year.

The trend values, computed by the equation Y = antilog [2.0527 + 0.0286X], are written in the last column of the above table. Further, the trend for 1992 is obtained by substituting X = 5, in the above equation.

$\therefore$ Y = antilog[2.0527 + 0.0286 × 5] = antilog[2.1957] = 156.93.

Remarks: The exponential trend equation plotted on a semilogarithmic graph is a straight line.

Example 16:

Fit an exponential trend $Y = a.b^t$ to the following data :

Census Year (*t*)	:	1941	1951	1961	1971	1981	1991
Population of India (*in Crores*)	:	31.9	36.1	43.9	54.8	68.3	84.4

Predict the population for 2001.

Solution.

Calculation Table

Census Year t	Population Y	$X = \frac{(t-1966)}{5}$	$\log Y$	$X\log Y$	X^2
1941	31.9	−5	1.5038	−7.5190	25
1951	36.1	−3	1.5575	−4.6725	9
1961	43.9	−1	1.6425	−1.6425	1
1971	54.8	1	1.7388	1.7388	1
1981	68.3	3	1.8344	5.5032	9
1991	84.4	5	1.9263	9.6315	25
Total		0	10.2033	3.0395	70

From the above table, we get $A = \frac{10.2033}{6} = 1.70$ and $B = \frac{3.0395}{70} = 0.043$

Further, a = antilog 1.70 = 50.12 and b = antilog 0.043 = 1.10

Thus, the fitted trend equation is $Y = 50.12(1.10)^X$,

Origin : 1st July, 1966 and unit of X = 5 years.

The trend values can be computed by the equation Y = antilog [1.70 + 0.043X]. Further, the prediction of population for 2001 is obtained by substituting X = 7, in the above equation.

$\therefore$ Y = antilog[1.70 + 0.043 × 7] = antilog[2.001] = 100.2 crores

13.1.4 Merits and Demerits of Least Squares Method

Merits

1. Given the mathematical form of the trend to be fitted, the least squares method is an objective method.

2. Unlike the moving average method, it is possible to compute trend values for all the periods and predict the value for a period lying outside the observed data.

3. The results of the method of least squares are most satisfactory because the fitted trend satisfies the two important properties, i.e., (i) $S(Y_o - Y_t) = 0$ and (ii) $S(Y_o - Y_t)^2$ is minimum. Here Y_o denotes the observed value and Y_t denotes the calculated trend value.

   The first property implies that the position of fitted trend equation is such that the sum of deviations of observations above and below this is equal to zero. The second property implies that the sum of squares of deviations of observations, about the trend equation, are minimum.

Demerits

1. As compared with the moving average method, it is a cumbersome method.

2. It is not flexible like the moving average method. If some observations are added, then the entire calculations are to be done once again.

3. It can predict or estimate values only in the immediate future or past.

4. The computation of trend values, on the basis of this method, doesn't take into account the other components of a time series and hence not reliable.

5. Since the choice of a particular trend is arbitrary, the method is not, strictly, objective.

6. This method cannot be used to fit growth curves, the pattern followed by the most of the economic and business time series.

13.2 Summary

- Given the data (Y_t, t) for n periods, where t denotes time period such as year, month, day, etc., we have to find the values of the two constants, a and b, of the linear trend equation $Y_t = a + bt$.

  Using the least square method, the normal equation for obtaining the values of a and b are:

  $$\Sigma Y_t = na + b\Sigma t \text{ and}$$

  $$\Sigma tY_t = a\Sigma t + b\Sigma t^2$$

  Let X = t - A, such that ΣX = 0, where A denotes the year of origin.

  The above equations can also be written as

  $$\Sigma Y = na + b\Sigma X$$

  $$\Sigma XY = a\Sigma X + b\Sigma X^2$$

  (Dropping the subscript t for convenience).

  Since SX = 0, we can write $a = \frac{\sum Y}{n}$ and $b = \frac{\sum XY}{\sum X^2}$

- Unlike the moving average method, it is possible to compute trend values for all the periods and predict the value for a period lying outside the observed data.

- The results of the method of least squares are most satisfactory because the fitted trend satisfies the two important properties, i.e., (i) $S(Y_o - Y_t) = 0$ and (ii) $S(Y_o - Y_t)^2$ is minimum. Here Y_o denotes the observed value and Y_t denotes the calculated trend value.

  The first property implies that the position of fitted trend equation is such that the sum of deviations of observations above and below this is equal to zero. The second property implies that the sum of squares of deviations of observations, about the trend equation, are minimum.

- It is not flexible like the moving average method. If some observations are added, then the entire calculations are to be done once again.

- It can predict or estimate values only in the immediate future or past.

- The computation of trend values, on the basis of this method, doesn't take into account the other components of a time series and hence not reliable.

13.3 Keywords

The fitted trend is termed as the best in the sense that the sum of squares of deviations of observations, from it, are minimised.

Parabolic trend: The mathematical form of a parabolic trend is given by $Y_t = a + bt + ct^2$ or $Y = a + bt + ct^2$ (dropping the subscript for convenience). Here a, b and c are constants to be determined from the given data.

13.4 Self Assessment

1. Fill in the blanks:

    (i) Series of figures arranged in chronological order is known as ....

    (ii) .... is that irreversible movement which continues in the same direction for a considerable period of time.

    (iii) The trend equation fitted by the method of least squares is known as the equation of .... fit.

    (iv) In case of .... trend, the successive observations differ by a constant number.

    (v) In the case of an exponential trend, the successive observations differ by a constant ....

    (vi) In the case of linear trend Y = a + bX, a is termed as the .... value in the year of ....

13.5 Review Questions

1. Determine the trend and short-term fluctuations, assuming additive model, from the following data by calculating 3 yearly moving averages. The figures of profit are in Rs '000.

Years :	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Profits :	34	46	52	55	58	61	58	61	64	55

2. Calculate the long-term trend and short-term oscillations, assuming multiplicative model, with a three-year period from the following data on output (in tonnes) of tea.

Year	*Output*	*Year*	*Output*
1969	1632	1973	2620
1970	1557	1974	3120
1971	1652	1975	3236
1972	2100	1976	3562

3. Construct a four-year moving average from the following data on the consumption (in '000 bales) of imported cotton in India.

Year :	1920	1930	1940	1950	1960	1970	1980
Consumption :	129	131	106	91	95	84	93

4. Determine trend values by method of moving average if the observations, given below, are known to have a business cycle of 4 years.

Year :	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992
Values :	41	61	55	48	53	67	62	60	67	73	78	76	84

5. Assuming five-yearly cycle, determine trend of bank clearings (in Rs crores) by moving average method:

Years :	1	2	3	4	5	6	7	8	9	10	11	12
Bank Clearings :	53	79	76	66	69	94	105	87	79	104	97	92

6. Find trend of the following series using a three-year weighted moving average with weights 1, 2, 1.

Year :	1	2	3	4	5	6	7
Value :	2	4	5	7	8	10	13

Unit 14: χ^2 - Test Hypothesis

CONTENTS

Objectives

After studying this unit, you will be able to:

- Discuss Test of Hypothesis Concerning Significance of Correlation Coefficient
- Describe Test of Hypothesis concerning Correlation Coefficient using Fisher's Z test
- Explain Test Concerning Equality of Correlations in two Populations

Introduction

In last unit you have studied about hypothesis concerning standard deviation. In this unit you will go through χ^2 - test hypothesis.

14.1 Test of Hypothesis concerning Correlation Coefficient

Let ρ be coefficient of linear correlation in a bivariate normal population and r be its estimator based on a sample of n observations (X_i, Y_i).

14.1.1 Test of Hypothesis Concerning Significance of Correlation Coefficient

Here we have to test whether ρ is different from zero. Accordingly, H_0 and H_a are $\rho = 0$ and $\rho \neq 0$ respectively.

For small samples, the test statistic can be obtained from the sampling distribution of b. We note that if *r* = 0, then *b* would also be zero.

Therefore, $\frac{b}{S.E.(b)} = r\cdot\frac{S_Y}{S_X}\cdot\frac{1}{S.E.(b)} = r\cdot\frac{S_Y}{S_X}\cdot\frac{S_X}{S_Y}\sqrt{\frac{n-2}{1-r^2}} = r\sqrt{\frac{n-2}{1-r^2}}$ will follow t - distribution with (n - 2) d.f. Hence, $r\sqrt{\frac{n-2}{1-r^2}}$ can be taken as the test statistic. We note that $S.E.(r) = \sqrt{\frac{1-r^2}{n-2}}$. Therefore, 100(1 - *a*)% confidence limits of *r* can be written as r ± $t_{a/2}$ S.E.(r).

Example 1: A random sample of 11 pairs of observations from a bivariate normal population gave r = 0.29. Test the significance of correlation in population.

Solution.

We have to test $H_0 : \rho = 0$ against $H_a : \rho \neq 0$.

$$t_{cal} = |0.29|\sqrt{\frac{9}{1-0.29^2}} = 0.91.$$

The value of t from tables at 5% level of significance and 9 d.f. is 2.26. Thus, there is no evidence against H_0.

14.1.2 Test of Hypothesis concerning Correlation Coefficient using Fisher's Z test

This test is applicable whether n is small or large. If r is correlation in sample, then its Fisher's Z transformation is given by $Z = \frac{1}{2}\log_e\frac{1+r}{1-r}$.

Further, if *r* is correlation in population, its Fisher's Z transformation, denoted by *x*, is given by

$$\xi = \frac{1}{2}\log_e\frac{1+\rho}{1-\rho}$$

Fisher has shown that the sampling distribution of Z is approximately normal with mean *x* and standard error $\frac{1}{\sqrt{n-3}}$. Thus, $(Z-\xi)\sqrt{n-3} \sim N(0,1)$.

Note: Since the values of Z and *x* are defined using e as the base of the logarithms, it is necessary to convert them into logarithms with base 10 for calculation purposes. Accordingly, we write

$$Z = \frac{1}{2}\log_e\frac{1+r}{1-r} = \frac{1}{2}\log_{10}\frac{1+r}{1-r}\times\log_e 10 = \frac{1}{2}\log_{10}\frac{1+r}{1-r}\times\frac{1}{\log_{10}e}$$

$$= \frac{1}{2}\times 2.3026\times\log_{10}\frac{1+r}{1-r} = 1.1513\log_{10}\frac{1+r}{1-r}$$

Similarly, we have $\xi = 1.1513\log_{10}\frac{1+\rho}{1-\rho}$

Example 2: If the correlation of 10 pairs of observations (X, Y) is 0.96, test the hypothesis that correlation in population is 0.99.

Solution.

We have to test $H_0 : \rho = 0.99$ against $H_a : \rho \neq 0.99$

Further,

$$Z = 1.1513 \log_{10} \frac{1+0.96}{1-0.96} = 1.1513 \log_{10} 49 = 1.1513 \times 1.6902 = 1.9459 \text{ and}$$

$$\xi = 1.1513 \log_{10} \frac{1+0.99}{1-0.99} = 1.1513 \log_{10} 199 = 1.1513 \times 2.2989 = 2.6464$$

The test statistic is $z = |Z - \xi|\sqrt{n-3} = |1.9459 - 2.6464|\sqrt{7} = 1.8563.$

Since this value is less than 1.96, there is no evidence against H_0 at 5% level of significance.

14.1.3 Test Concerning Equality of Correlations in two Populations

Let there be two independent random samples of sizes n_1 and n_2 from two normal populations with correlations ρ_1 and ρ_2 respectively. Let r_1 and r_2 be the correlations computed from the respective samples.

If Z_1, Z_2, ξ_1 and ξ_2 denote Fisher's transformation of r_1, r_2, r_1 and ρ_2 respectively, then $Z_1 \sim N\left(\xi_1, \frac{1}{\sqrt{n_1 - 3}}\right)$ and $Z_2 \sim N\left(\xi_2, \frac{1}{\sqrt{n_2 - 3}}\right)$

$$\therefore \; Z_1 - Z_2 \sim N\left(\xi_1 - \xi_2, \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}\right)$$

$$\text{or } \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \sim N(0,1) \text{ under } H_0 : \rho_1 = \rho_2$$

Example 3: The correlation coefficients 0.89 and 0.85 were computed from two independent samples of sizes 12 and 16 respectively. Test whether they can be regarded to have come from two bivariate populations with different correlation coefficients?

Solution.

We shall test $H_0 : \rho_1 = \rho_2$ against $H_a : \rho_1 \neq \rho_2$.

$$\text{Now } Z_1 = 1.1513 \log_{10} \frac{1.89}{0.11} = 1.1513 \log_{10} 17.18 = 1.1513 \times 1.2350 = 1.42$$

$$\text{and } Z_2 = 1.1513 \log_{10} \frac{1.85}{0.15} = 1.1513 \log_{10} 12.33 = 1.1513 \times 1.0911 = 1.26$$

$\therefore$ The test statistic is $z = \dfrac{|1.42 - 1.26|}{\sqrt{\dfrac{1}{9} + \dfrac{1}{13}}} = 0.16 \times \sqrt{\dfrac{9 \times 13}{22}} = 0.369.$

Since this value is less than 1.96, there is no evidence against H_0 at 5% level of significance. Thus, the given samples provide no evidence of different correlations in two populations.

14.2 Uses of χ^2 test

In addition to the use of χ^2 in tests of hypothesis concerning the standard deviation, it is used as a test of goodness of fit and as a test of independence of two attributes. These tests are explained in the following sections.

14.2.1 χ^2 - test as a Goodness of Fit

χ^2 - test can be used to test, how far the fitted or the expected frequencies are in agreement with the observed frequencies. We know that for large values of n, the sampling distribution of X, the number of successes, is normal with mean np and variance np(1 - π). Thus, $z = \dfrac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \sim N(0,1).$

Further, square of z is a c^2 -variate with one degree of freedom. We can write

$$z^2 = \frac{(X - n\pi)^2}{n\pi(1-\pi)} = (X - n\pi)^2 \left[\frac{(1-\pi)+\pi}{n\pi(1-\pi)}\right] \qquad (\because 1 = 1 - \pi + \pi)$$

$$= (X - n\pi)^2 \left[\frac{1}{n\pi} + \frac{1}{n(1-\pi)}\right] = \frac{(X - n\pi)^2}{n\pi} + \frac{(X - n\pi)^2}{n(1-\pi)} \qquad \text{.... (1)}$$

We can write $\dfrac{(X - n\pi)^2}{n(1-\pi)} = \dfrac{(X - n + n - n\pi)^2}{n(1-\pi)} = \dfrac{[(X-n) + n(1-\pi)]^2}{n(1-\pi)}$

$$= \frac{[(n-X) - n(1-\pi)]^2}{n(1-\pi)} = \frac{[(n-X) - E(n-X)]^2}{E(n-X)}$$

Similarly $\dfrac{(X - n\pi)^2}{n\pi} = \dfrac{[X - E(X)]^2}{E(X)}$.

Thus, equation (1) can be written as $z^2 = \dfrac{[X - E(X)]^2}{E(X)} + \dfrac{[(n-X) - E(n-X)]^2}{E(n-X)}$

Here X denotes the observed number of successes and (n - X) the observed number of failures.

Let O_1, E_1 denote the observed and expected number of successes respectively and O_2, E_2 denote the observed and expected number of failures respectively.

$\therefore \qquad z^2 = \dfrac{(O_1 - E_1)^2}{E_1} + \dfrac{(O_2 - E_2)^2}{E_2}$ is a χ^2 - variate with 1 d.f.

Also we note that $O_1 + O_2 = E_1 + E_2 = n$.

The above result can be generalised for a manifold classification. If a population is divided into k mutually exclusive classes with observed and expected frequencies as $O_1, O_2, \ldots\ldots O_k$ and $E_1, E_2, \ldots\ldots E_k$ respectively, then $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ is a c^2 - variate with (k - 1) d.f. Here again we have

$$\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i = N \text{ (total frequency).}$$

Example 40: 300 digits were chosen from a table of numbers and the following frequency distribution was obtained :

Digit	:	0	1	2	3	4	5	6	7	8	9
Frequency	:	26	28	33	32	28	37	33	30	30	23

Test the hypothesis that the digits are uniformly distributed over the table.

Solution.

When H_0 is true, the expected frequency of each digit would be 30.

$$\therefore \quad \chi^2 = \frac{1}{30} \sum O_i^2 - N$$

$$= \frac{1}{30}\left(26^2 + 28^2 + 33^2 + 32^2 + 28^2 + 37^2 + 33^2 + 30^2 + 30^2 + 23^2\right) - 300 = 4.8$$

The value of χ^2 from table for 5% level of significance and 9 d.f. is 16.92. Since the calculated value is less than tabulated, there is no evidence against H_0. Thus, the distribution of numbers over the table may be treated as uniform.

Example 41: A sample analysis of examination results of 200 M.B.A.'s was made. It was found that 46 students had failed, 68 secured a third division, 62 secured a second division and the rest were placed in the first division. Are these figures commensurate with the general examination result which is in the ratio of 2 : 3 : 3 : 2 for the various categories, respectively? (Given : Table value of chi-square for 3 d.f. at 5% level of significance is 7.81.)

Solution.

H_0 : The students in various categories are distributed in the ratio 2 : 3 : 3 : 2.

The expected number of students, under the assumption that H_0 is true, are :

expected number of failures $= \frac{2}{(2+3+3+2)} \times 200 = 40$,

expected number of third divisioners $= \frac{3}{10} \times 200 = 60$,

expected number of second divisioners $= \frac{3}{10} \times 200 = 60$ and

expected number of first divisioners $= \frac{2}{10} \times 200 = 40$.

Thus, we have $\chi^2 = \frac{(46-40)^2}{40} + \frac{(68-60)^2}{60} + \frac{(62-60)^2}{60} + \frac{(24-40)^2}{40} = 8.44.$

Since this value is greater than the tabulated value, 7.81, for 3 d.f. and 5% level of significance, H_0 is rejected.

Example 42: A survey of 320 families with 5 children each revealed the following distribution:

No. of boys	:	5	4	3	2	1	0
No. of girls	:	0	1	2	3	4	5
No. of families	:	14	56	110	88	40	12

Is the result consistent with the hypothesis that male and female births are equally probable?

Solution.

Assuming that H_0 (i.e., male and female births are equally probable) is true, the expected number of families having r boys (or equivalently 5 - r girls) is given by $E_r = 320 \times {}^5C_r \left(\frac{1}{2}\right)^5 = 10 \times {}^5C_r$. On substituting r = 5, 4, 3, 2, 1, 0, the respective expected frequencies are 10, 50, 100, 100, 50 and 10.

$$\therefore\ \chi^2 = \frac{(14-10)^2}{10} + \frac{(56-50)^2}{50} + \frac{(110-100)^2}{100} + \frac{(88-100)^2}{100} + \frac{(40-50)^2}{50} + \frac{(12-10)^2}{10} = 7.16.$$

The value from table for 5 d.f. at 5% level of significance is 11.07, which is greater than the calculated value. Thus, there is no evidence against H_0.

Example 43:

The record for a period of 180 days, showing the number of electricity failures per day in Delhi are shown in the following table :

No.of failures	:	0	1	2	3	4	5	6	7
No.of days	:	12	39	47	40	20	17	3	2

Determine, by using c^2 - test, whether the number of failures can be regarded as a Poisson variate?

Solution.

We have to test H_0 : No. of failures is a Poisson variate against H_a : No. of failures is not a Poisson variate.

The mean of the Poisson distribution is

$$m = \frac{0 \times 12 + 1 \times 39 + 2 \times 47 + 3 \times 40 + 4 \times 20 + 5 \times 17 + 6 \times 3 + 7 \times 2}{180} = 2.5$$

The computations of χ^2 are done in the following table :

No.of families	*Expected freq.*(E_i)	*Observed freq.*(O_i)	$\frac{(O_i - E_i)^2}{E_i}$
0	$180 \times e^{-2.5} = 14.76$	12	0.52
1	$E_0 \times 2.5 = 36.94$	39	0.11
2	$E_1 \times 2.5/2 = 46.17$	47	0.01
3	$E_2 \times 2.5/3 = 38.48$	40	0.06
4	$E_3 \times 2.5/4 = 24.05$	20	0.68
5	$E_4 \times 2.5/5 = 12.02$	17	2.06
6 *or more*	*by difference* = 7.58	5	0.88
Total	180		$\chi^2 = 4.32$

The value of χ^2 from table at 5% level of significance and 5 d.f. is 11.07. Since the calculated value is less than the tabulated value, there is no evidence against H_0.

14.2.2 χ^2 - test as a Test for Independence of Two Attributes

Let us assume that a population is classified into m mutually exclusive classes, A_1, A_2, ... A_m, according to an attribute A and each of these m classes are further classified into n mutually exclusive classes, like A_iB_1, A_iB_2, ... A_iB_n , etc., according to another attribute B.

If O_{ij} is the observed frequency of A_iB_j , i.e., $(A_iB_j) = O_{ij}$, the above classification can be expressed in form of following table, popularly known as contingency table.

$B \rightarrow$ $A \downarrow$	B_1	B_2	$\cdots\cdots$	B_n	*Total*
A_1	O_{11}	O_{12}	$\cdots\cdots$	O_{1n}	(A_1)
A_2	O_{21}	O_{22}	$\cdots\cdots$	O_{2n}	(A_2)
$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$	
A_m	O_{m1}	O_{m2}	$\cdots\cdots$	O_{mn}	(A_m)
Total	(B_1)	(B_2)	$\cdots\cdots$	(B_n)	N

Assuming that A and B are independent, we can compute the expected frequencies of each cell,

i.e., $E_{ij} = \frac{(A_i)(B_j)}{N}$. Thus, $\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ would be a c^2 - variate with (m - 1)(n - 1) d.f.

Remarks : The expected frequencies of some cells may be obtained by the application of the above formula while the remaining cell frequencies can be obtained by subtraction. The minimum number of cell frequencies, that must be computed by the use of the formula, is equal to the degrees of freedom of the c^2 statistic.

Example 44: The employees in 4 different firms are distributed in three skill categories shown in the following table. Test the hypothesis that there is no relationship between the firm and the type of labour. Let the level of significance be 5%.

Firm → Type of labour ↓	A	B	C	D
Skilled	24	24	23	49
Semi-skilled	32	60	37	51
Manual	24	56	40	80

Solution.

H_0 : There is no relation between the firm and the nature of labour.

Calculation of Expected Frequencies

Firm → labour ↓	A	B	C	D	Total
Skilled	$\frac{80\times 120}{500}$ $= 19.2$	$\frac{140\times 120}{500}$ $= 33.6$	$\frac{100\times 120}{500}$ $= 24.0$	$\frac{180\times 120}{500}$ $= 43.2$	120
Semi-skilled	$\frac{80\times 180}{500}$ $= 28.8$	$\frac{140\times 180}{500}$ $= 50.4$	$\frac{100\times 180}{500}$ $= 36.0$	$\frac{180\times 180}{500}$ $= 64.8$	180
Manual	$\frac{80\times 200}{500}$ $= 32.0$	$\frac{140\times 200}{500}$ $= 56.0$	$\frac{100\times 200}{500}$ $= 40.0$	$\frac{180\times 200}{500}$ $= 72.0$	200
Total	80	140	100	180	500

We note that the totals of corresponding rows or columns are same for the observed as well as the expected frequencies.

From the observed and the expected frequencies, we get $c^2 = 12.81$. Further, the value of c^2 from the table for (4 - 1)(3 - 1) = 6 d.f. and 5% level of significance is 12.59. Since the calculated value is greater than the tabulated value H_0 is rejected.

Example 44: Samples of household income were taken from four cities. Test whether the cities are homogeneous with regard to the distribution of income?

Cities → Income(Rs) ↓	A	B	C	D	Total
Under 3000	10	15	15	10	50
3000-5000	5	10	15	10	40
Over 5000	15	15	10	20	60
Total	30	40	40	40	150

Solution.

H_0 : Various cities are homogeneous with regard to the distribution of income.

Notes

Computation of Expected Frequencies

Cities → *Income*(*Rs*) ↓	*A*	*B*	*C*	*D*	*Total*
Under 3000	10.00	13.33	13.33	13.33	50
3000-5000	8.00	10.67	10.67	10.67	40
Over 5000	12.00	16.00	16.00	16.00	60
Total	30	40	40	40	150

Note that the expected frequencies for city A, under various income groups, are computed as $\frac{30\times 50}{150}=10.00, \frac{30\times 40}{150}=8.00$ and $\frac{30\times 60}{150}=12.00.$ Other frequencies have also been computed in a similar manner.

Using the observed and expected frequencies, the value of c^2 = 8.28.

Further, the value of X^2 from tables for 6 d.f. at 5% level of significance is

Since the calculated value is less than the tabulated value, there is no evidence against H_0.

The value of χ^2 for a 2 × 2 Contingency table

For a 2 × 2 contingency table,

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	$a+b+c+d=N$

, the value of χ^2 can be directly computed with the use of the following formula :

$$\chi^2=\frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Yate's correction for continuity

We know that c^2 is a continuous random variate but the frequencies of various cells of a contingency table are integers. When N is large, the distribution of $\sum\frac{(O-E)^2}{E}$ is approximately c^2. However, the corrections for continuity are required when N is small. Yates has suggested the following corrections for continuity in a 2 × 2 contingency table :

If ad > bc, reduce a and d by $\frac{1}{2}$ and increase b and c by $\frac{1}{2}$. Similarly, If ad < bc, increase a and d by $\frac{1}{2}$ and decrease b and c by $\frac{1}{2}$.Thus, the contingency tables in the two situations become

$$\begin{array}{c|c} a-\frac{1}{2} & b+\frac{1}{2} \\ \hline c+\frac{1}{2} & d-\frac{1}{2} \end{array} \text{ and } \begin{array}{c|c} a+\frac{1}{2} & b-\frac{1}{2} \\ \hline c-\frac{1}{2} & d+\frac{1}{2} \end{array} \text{ respectively.}$$

The value of *c*² can now be obtained as $\chi^2 = \dfrac{N\left(|ad-bc|-\dfrac{N}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}$.

Brand and Snedecor formula for a 2 ´ r Contingency table

For a 2 ´×r contingency table,

$A \rightarrow$ $B \downarrow$	A_1	A_2	$\cdots$	A_r	*Total*
B_1	a_1	a_2	$\cdots$	a_r	a
B_2	b_1	b_2	$\cdots$	b_r	b
Total	n_1	n_2	$\cdots$	n_r	N

, the value of *c*² can be directly computed by the use of the following formula :

$$\chi^2 = \frac{N^2}{ab}\left(\sum_{i=1}^{r}\frac{a_i^2}{n_i} - \frac{a^2}{N}\right) \text{ with (r - 1) d.f.}$$

Example 46: In a recent diet survey, the following results were obtained in an Indian city:

No.of families	*Hindus*	*Muslims*	*Total*
Tea takers	1236	164	1400
Non-tea takers	564	36	600
Total	1800	200	2000

Discuss whether there is any significant difference between the two communities in the matter of taking tea? Use 5% level of significance.

Solution.

The null hypothesis to be tested can be written as H_0 : There is no difference between the two communities in the matter of taking tea.

Using the direct formula, we have $\chi^2 = \dfrac{2000(1236\times 36 - 164\times 564)^2}{1400\times 1800\times 200\times 600} = 15.24.$

The value of *c*² from table for 1 d.f. and 5% level of significance is 3.84. Since the calculated value is greater than the tabulated value, H_0 is rejected.

Example 47: A certain drug is claimed to be effective in curing cold. In an experiment on 160 persons with cold, half of them were given the drug and the remaining half were given sugar pills. The patients' reactions to the treatment are recorded in the following table :

	Helped	*Harmed*	*No effect*	*Total*
Drug	52	10	18	80
Sugar pills	44	10	26	80
Total	96	20	44	160

Test the hypothesis that the drug is no better than the sugar pills for curing cold.

Solution.

Notes

H_0: The drug is not effective in curing cold

Using the Brandt and Snedecor formula, we have

$$\chi^2 = \frac{160 \times 160}{80 \times 80}\left(\frac{52^2}{96} + \frac{10^2}{20} + \frac{18^2}{44} - \frac{80^2}{160}\right) = 2.12.$$

This value is less than the tabulated value (= 5.99) for 2 d.f. and 5% level of significance. Thus, there is no evidence against H_0.

14.3 Summary

- Here we have to test whether ρ is different from zero. Accordingly, H_0 and H_a are ρ = 0 and ρ ≠ 0 respectively.

  For small samples, the test statistic can be obtained from the sampling distribution of b. We note that if *r* = 0, then *b* would also be zero.

  Therefore, $\frac{b}{S.E.(b)} = r \cdot \frac{S_Y}{S_X} \cdot \frac{1}{S.E.(b)} = r \cdot \frac{S_Y}{S_X} \cdot \frac{S_X}{S_Y}\sqrt{\frac{n-2}{1-r^2}} = r\sqrt{\frac{n-2}{1-r^2}}$ will follow t - distribution with (n - 2) d.f. Hence, $r\sqrt{\frac{n-2}{1-r^2}}$ can be taken as the test statistic. We note that $S.E.(r) = \sqrt{\frac{1-r^2}{n-2}}$. Therefore, 100(1 - *a*)% confidence limits of *r* can be written as r ± $t_{a/2}$ S.E.(r).

- Let there be two independent random samples of sizes n_1 and n_2 from two normal populations with correlations ρ_1 and ρ_2 respectively. Let r_1 and r_2 be the correlations computed from the respective samples.

  If Z_1, Z_2, ξ_1 and ξ_2 denote Fisher's transformation of r_1, r_2, r_1 and ρ_2 respectively, then

  $$Z_1 \sim N\left(\xi_1, \frac{1}{\sqrt{n_1 - 3}}\right) \text{ and } Z_2 \sim N\left(\xi_2, \frac{1}{\sqrt{n_2 - 3}}\right)$$

  $$\therefore\ Z_1 - Z_2 \sim N\left(\xi_1 - \xi_2, \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}\right)$$

  $$\text{or } \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \sim N(0,1) \text{ under } H_0 : \rho_1 = \rho_2$$

14.4 Keywords

Fisher's test: It is applicable whether n is small or large. If r is correlation in sample, then its Fisher's Z transformation is given by $Z = \frac{1}{2}\log_e \frac{1+r}{1-r}$.

χ^2 - test: It can be used to test, how far the fitted or the expected frequencies are in agreement with the observed frequencies.

Notes

14.5 Self Assessment

1. Let ρ be coefficient of .... in a bivariate normal population and r be its estimator based on a sample of n observations (X_i, Y_i).
2. Let there be two .... samples of sizes n_1 and n_2 from two normal populations with correlations ρ_1 and ρ_2 respectively. Let r_1 and r_2 be the correlations computed from the respective samples.
3. In addition to the use of χ^2 in tests of hypothesis concerning the standard deviation, it is used as a test of .... and as a test of independence of two attributes. These tests are explained in the following sections.
4. .... can be used to test, how far the fitted or the expected frequencies are in agreement with the observed frequencies.

14.6 Review Questions

1. We want to decide whether a cubic die is balanced or not. For this purpose the die is thrown 300 times and various outcomes are recorded. If the observed frequencies of the six faces, namely 1, 2, 3, 4, 5 and 6 are 35, 40, 32, 60, 68 and 65 respectively, can we conclude that the die is unbiased?

   Hint : The expected frequency of each face under H_0 is 50.

2. Four coins are tossed 320 times and the number of heads obtained were recorded as follows :

No. of heads	:	0	1	2	3	4
Frequency	:	15	102	108	68	27

   Can we regard all the coins as unbiased?

   Hint : Find expected frequencies of the number of heads on the assumption that the coins are unbiased.

3. Three dice were thrown 80 times and the number of times 2, 4 or 6 was obtained, were recorded as given below:

No. of dice showing 2,4 or 6	:	0	1	2	3
Frequency	:	8	28	32	12

   Test the hypothesis that all the three dice are fair.

   Hint : Under H_0, the probability of success, i.e., getting 2 or 4 or 6 on a die is 0.5. If r denotes the number of dice giving successes in a throw, we have $p(r) = {}^4C_r 0.5^4$.

4. The health department of municipal corporation of a city believes that 14% persons of the city are smokers as well as drinkers, 30% are drinkers while 40% are smokers. In a random sample of 150 persons, it was found that 24 persons were smokers as well as drinkers, 21 were only drinkers and 36 were only smokers. Do the above data support the belief of the department. Use 5% level of significance.

   Hint : Use the figures of belief, given in percentages, to find the expected frequencies.

5. A normal distribution was fitted to the distribution of new business brought by 100 insurance agents with the following results:

New business ('000Rs)	:	10-20	20-30	30-40	40-50	50-60
Observed Frequency	:	10	20	33	22	15
Expected Frequency	:	9	22	32	25	12

Test the goodness of fit of the distribution.

Hint : The degrees of freedom of the χ^2 statistic would be 4.

Answers: Self Assessment

1. linear correlation
2. independent random
3. goodness of fit
4. χ^2 - test

14.7 Further Readings

Books

Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.

7. From the following data calculate the 4-yearly moving average and determine the trend values. Find short-term fluctuations, assuming multiplicative model, and indicate their composition. Plot the original data and the trend values on a graph.

Year	:	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987
Value	:	50.0	36.5	43.0	44.5	38.9	38.1	32.6	41.7	41.1	33.8

Answers: Self Assessment

1. (i) time series (ii) trend (iii) best (iv) linear (v) proportion (vi) trend, origin.

14.6 Further Readings

Books

Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.

Unit 15: Hypothesis Testing

Notes

CONTENTS

Objectives

After studying this unit, you will be able to:

- Discuss Hypothesis Testing
- Explain Hypothesis Concerning Mean

Introduction

A hypothesis is a preconceived idea about the nature of a population or about the value of its parameters. The statements like the distribution of heights of students of a university is normally distributed, the number of road accidents per day in Delhi is 10, etc., are some examples of a hypothesis.

The test of a hypothesis is a procedure by which we test the validity of a given statement about a population. This is done on the basis of a random sample drawn from it.

The hypothesis to be tested is termed as Null Hypothesis, denoted by H_0. This hypothesis asserts that there is no difference between population and sample in the matter under consideration. For example, if H_0 is that population mean $\mu = \mu_0$, then we regard the random sample to have been obtained from a population with mean m_0.

Corresponding to any H_0, we always define an Alternative Hypothesis. This hypothesis, denoted by H_a, is alternate to H_0 , i.e., if H_0 is false then H_a is true and vice-versa.

15.1 Test of Hypothesis

In order to illustrate the procedure of testing a null hypothesis, let us assume that the life of electric bulbs of a company is distributed normally with standard deviation of 150 hours and we want to test the null hypothesis that the mean life of bulbs is 1600 hours against the alternative hypothesis that the mean life is not 1600 hours.

Assuming that H_0 is true, we can construct a sampling distribution of $\overline{X}$, the mean life of bulbs in the sample. If a random sample of 100 bulbs is taken from this population, we know that the distribution of $\overline{X}$ will be normal with mean *m* = 1600 hours and standard error, $S.E._{\bar{X}} = \frac{150}{10} = 15$ hours. Further, we know that for a normal distribution

$$P\left(-2 \le \frac{\bar{X}-1600}{15} \le 2\right) = 0.9544$$

or $$P\left(1600 - 2\times 15 \le \bar{X} \le 1600 + 2\times 15\right) = 0.9544$$

or $$P\left(1570 \le \bar{X} \le 1630\right) = 0.9544$$

This result shows that the likelihood of getting a random sample, from the given population, with mean lying between 1570 and 1630 hours is 95.44% or equivalently, the likelihood of getting a random sample having its mean either less than 1570 or more than 1630 hours is only 4.56%. Thus, a random sample with its mean lying outside these limits is highly unlikely under the assumption that null hypothesis is true.

However, if the mean computed from the drawn sample is found to lie outside these limits, it may imply that either null hypothesis is false or the rare event, with probability = 4.56%, has occurred.

Thus, if we decide to reject the null hypothesis whenever the computed sample mean falls outside the above limits, the probability of our decision being wrong is only 4.56% or 0.0456.

Two Types of Errors

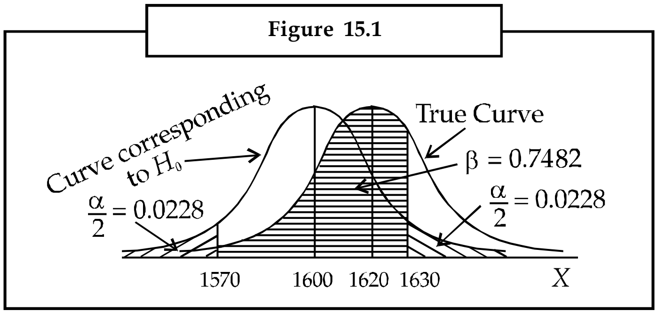
The decision of acceptance or rejection of a null hypothesis is made on the basis of a sample from a population and hence, an element of uncertainty is always involved in making such decisions. Two types of errors are likely to be committed in the procedure of testing a hypothesis. These are Type I and Type II errors. Type I error is committed when a true null hypothesis is rejected. The probability of this error is termed as the Level of Significance of the test and will be denoted by *a*. The probability of committing an error is also termed as its size. Note that size of type I error, i.e., *a* = 0.0456, in the above example.

Contrary to this, type II error is committed when a false null hypothesis is accepted. The probability of type II error is denoted by *b*. To understand the meaning of type II error, we assume that the true value of *m* is 1620 instead of the hypothesised value of 1600 hours. If the standard deviation is same, the value of *b* is given by P (1570 £ $\overline{X}$ £ 1630) when *m* = 1620 or P

$\left(\frac{1570-1620}{15} \le z \le \frac{1630-1620}{15}\right)$ = P(–3.33 £ Z £ 0.67) = 0.4996 + 0.2486 = 0.7482

The two types of errors are shown by the following figure.

Figure 15.1



It is obvious, from the above figure, that it is not possible to simultaneously control both types of errors because a decrease in probability of committing one type of error is accompanied by the increase in probability of committing the other type of error. Further, we may note that farther the true value of parameter from the hypothesised value, smaller would be the size of type II error, *b*. The graph of various values of *m* against *b* is known as the Operating Characteristic Surve.

In the procedure of testing a hypothesis, the probability or size of type I error, i.e., *a* is specified in advance. Usually we take *a* = 0.05 (i.e., 5%) or 0.01 (i.e., 1%). Also see remarks (1) given at the end of this section.

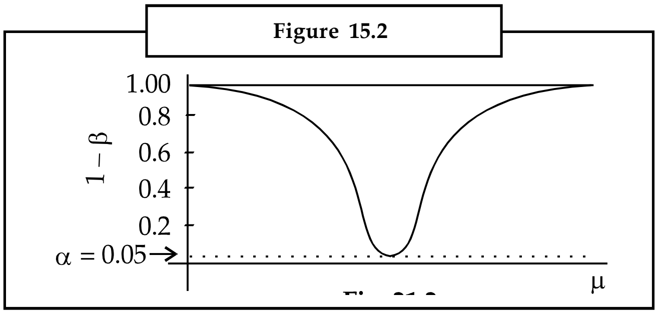
Power of a Test

The power of a test is defined as the probability of rejecting a false null hypothesis. Since *b* is the probability of accepting a false hypothesis, the power of test is given by 1 - *b*. More precisely, we can write

Power of a test = P [Rejecting H_0/H_0 is false] = 1 – *b*

Since the value of β depends upon the true value of population parameter (μ in the above example), the relationship between various values of m and 1 - β is termed as power function, as shown in Figure 31.2.

Figure 15.2



Critical Region and One Tailed versus Two Tailed Tests

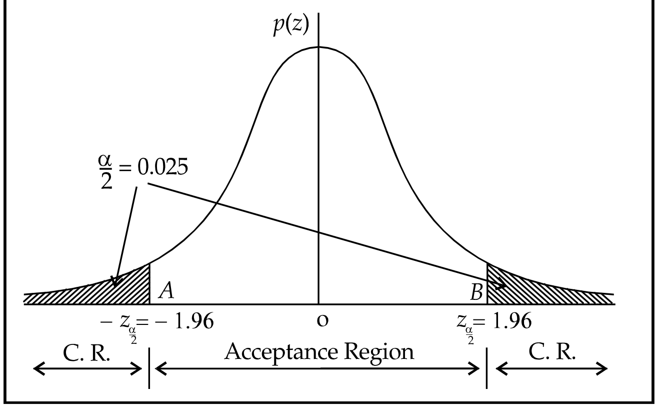
Let $H_0 : \mu = \mu_0$ against $H_a : \mu \neq \mu_0$, where μ_0 denotes some specified value of population mean m. For example, $\mu_0 = 1600$, in the example considered above.

If we decide to have a = 0.05, we know that for a standard normal variate P[- 1.96 ≤ z ≤ 1.96] = 1 - 0.05 = 0.95, the procedure of testing of hypothesis can be outlined as:

Reject H_0 if the computed value of z from the sample $\left(i.e., \; z_{cal} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \right)$ lies outside the interval (- 1.96, 1.96) and accept it otherwise.

In terms of figure, the portion of z axis covering the interval (- 1.96, 1.96), i.e, A to B is termed as the Acceptance Region and its remaining portions, which lie to the left of point A and to the right of point B, are termed as the Region of Rejection or Critical Region (C.R.).

Figure 15.3

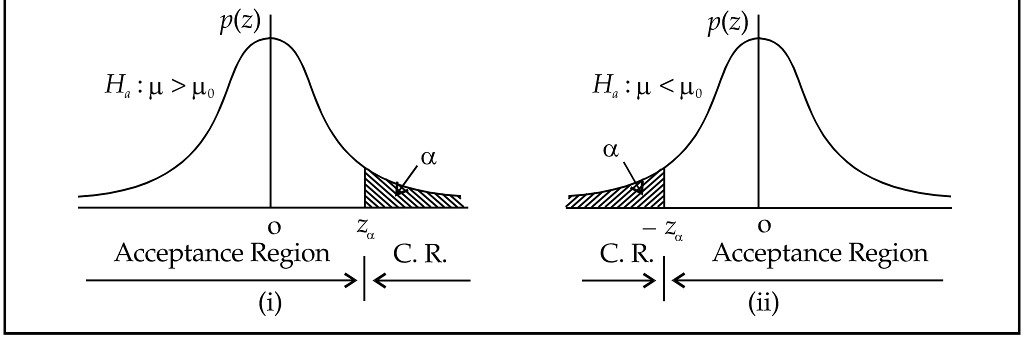

$p(z)$
$\frac{\alpha}{2} = 0.025$
A
B
$-z_{\frac{\alpha}{2}} = -1.96$
o
$z_{\frac{\alpha}{2}} = 1.96$
C. R.
Acceptance Region
C. R.

The specification of the critical region for a test depends upon the nature of the alternative hypothesis and the value of α. For example, $H_a : \mu \neq \mu_0$, this implies that m may be less or greater than μ_0. Thus, the critical region is to be specified on both tails of the curve with each part corresponding to half of the value of a. A test having critical region at both the tails of the probability curve is termed as a two tailed test.

Further, if H_a : $\mu > \mu_0$ or $\mu < \mu_0$, the critical region is to be specified only at one tail of the probability curve and the corresponding test is termed as a one tailed test. These situations are shown in the following figures.

The values of the random variable separating the acceptance region from critical region are termed as critical value(s). For example, $z_{a/2}$ and z_a, shown above, are critical values. Similarly, for a normal distribution the critical values for a two tailed test are - 1.96 and 1.96 for α = 0.05 or - 2.58 and 2.58 for α = 0.01 and the corresponding value for a one tailed test is ± 1.645 or ± 2.33 depending upon whether a = 0.05 or 0.01.

Figure 15.4



Remarks:

1. Out of the two types of errors, the type I error is considered to be more serious. Consequently, the probability of type I error is fixed at a low value (often 0.05 or lower). Thus, when the computed value of a statistic falls in the critical region, implying thereby that the probability of H_0 being true is low or equivalently the probability of H_0 being false is high, we reject H_0. However, if the computed value of statistics lies in the acceptance region, it would not be appropriate to say that the probability of H_0 being true is very high because the probability of accepting a false H_0 (the value of *b*) may also be high. Thus, accepting H_0 only implies that the sample information does not provide any evidence of H_0 being false. Because of this nature of the tests of hypothesis, the conclusion "accept H_0" is often replaced by "do not reject H_0" or "there is no evidence against H_0 on the basis of available sample information", etc.

2. The tests of hypothesis are also known as the Tests of Significance. We know that if the sample result is highly unlikely, H_0 is rejected because the sample result is significantly different from the hypothesised value. Alternatively, it implies that the observed difference between the computed and the hypothesised value is not attributable due to chance or fluctuations of sampling.

15.2 Tests of Hypothesis Concerning Mean

These tests can be divided into two broad categories depending upon whether *s*, the population standard deviation, is known or not.

15.2.1 Test of Hypothesis Concerning Population Mean (s being known)

This test is applicable when the random sample $X_1, X_2, \ldots\ldots X_n$ is drawn from a normal population. We can write

$H_0 : \mu = \mu_0$ (specified) against $H_a : \mu \neq \mu_0$ (two tailed test)

The test statistic $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Let the value of this statistic calculated from sample be denoted as $z_{cal} = \left|\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right|$. The decision rule would be :

Reject H_0 at 5%(say) level of significance if $z_{cal} > 1.96$. Otherwise, there is no evidence against H_0 at 5% level of significance.

Example 12: A manufacturer claims that the average mileage of scooters of his company is 40 kms/litre. A random sample of 20 scooters of the company showed an average mileage of 42 kms/litre. Test the claim of the manufacturer on the assumption that the mileage of scooter is normally distributed with a standard deviation of 2 kms/litre.

Solution.

Here, we have to test $H_0 : \mu = 40$ against $H_a : \mu \neq 40$.

$$z_{cal} = \left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| = \left|\frac{42-40}{2/\sqrt{20}}\right| = 4.47.$$

Since $z_{cal} > 1.96$, is rejected at 5% level of significance.

Notes

Remarks:

1. If the manufacturer claims that the average mileage is more than 40 kms/litre rather than equal to 40 kms/litre, we have to use a one tailed test. Now we shall test $H_0 : \mu = 40$ against $H_a : \mu > 40$ and z_{cal} would be defined as $z_{cal} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$. Since this value is also equal to 4.47 and lies in the critical region, we reject at 5% level of significance. This implies that the claim of the manufacturer may be taken as correct.

2. In one tailed tests the alternative hypothesis is expressed as a strict inequality and the null hypothesis as a weak inequality or simply equality.

3. The decision rule can also be specified in terms of prob or p-value of the observed sample result. The p-value is the smallest level of significance at which the null hypothesis can be rejected. We define p-value

$$= 2P\left(z \geq \left| \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \right| \right), \text{ for a two tailed test,}$$

$$= P\left(z \geq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \right), \text{ when } H_a : m > m_0 \text{ and}$$

$$= P\left(z \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \right), \text{ when } H_a : m < m_0$$

   The decision rule is : If p-value < α, reject H_0.

   In the above example p -value is approximately equal to zero when H_a is either $\mu \neq 40$ or $\mu > 40$, therefore H_0 is rejected. However, if H_a is taken as $\mu < 40$, the p -value is almost equal to unity and consequently H_0 would be accepted.

4. As per the central limit theorem, even if the parent population is not normal, the sampling distribution of z will be approximately normal when n > 30.

Example 13: A filling machine at a soft drink factory is designed to fill bottles of 200 ml with a standard deviation of 10 ml. A sample of 50 bottles was selected at random from the filled bottles and the volume of soft drink was computed to be 198 ml per bottle. Test the hypothesis that the mean volume of soft drink per bottle is not less than 200 ml.

Solution.

Here n > 30, therefore, the sampling distribution of mean volume of soft drink per bottle will be normal.

We have to test $H_0 : \mu \geq 200$ against $H_a : \mu < 200$.

It is given that $\bar{X} = 198$ and $\sigma = 10$.

Thus, the test static is $z_{cal} = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} = \frac{198 - 200}{10 / \sqrt{50}} = -1.41$

Since this value is greater than - 1.645, z_{cal} lies in the acceptance region. Hence, there is no evidence against H_0 at 5% level of significance.

Remarks:

Alternatively, a null hypothesis can be tested by computing critical sample mean $\overline{X}_C$ for a given standard error and the level of significance.

(i) Let H_0: $\mu = \mu_0$ against H_a: $\mu \neq \mu_0$

If a = 0.05, then $\overline{X}_C = \mu_0 \pm 1.96 \dfrac{\sigma}{\sqrt{n}}$

If $\mu_0 - 1.96 \dfrac{\sigma}{\sqrt{n}} < \overline{X} < \mu + 1.96 \dfrac{\sigma}{\sqrt{n}}$, we accept H_0.

(ii) Let H_0: $\mu \leq \mu_0$ against H_a: $\mu > \mu_0$ (Right tailed test)

If a = 0.05 then $\overline{X}_C = \mu_0 + 1.645 \dfrac{\sigma}{\sqrt{n}}$

If $\overline{X} > \overline{X}_C$, we reject H_0.

In the above example,

H_0: $\mu \geq 200$ against $\mu < 200$ (Left tailed test)

$$\therefore \ \overline{X}_C = 200 - 1.645 \times \frac{10}{\sqrt{50}} = 197.67$$

It is given that $\overline{X} = 198$. Since $\overline{X} > \overline{X}_C$, we accept H_0 at 5% level of significance.

15.2.2 Test of Hypothesis Concerning Population Mean (σ being unknown)

When *s* is not known, we use its estimate computed from the given sample. Here, the nature of the sampling distribution of $\overline{X}$ would depend upon sample size n. There are the following two possibilities:

(i) If parent population is normal and n < 30 (popularly known as small sample case), use t - test. The unbiased estimate of *s* in this case is given by $s = \sqrt{\dfrac{\sum\left(X_i - \overline{X}\right)^2}{n-1}}$.

Also, like normal test, the hypothesis may be one or two tailed.

(ii) If n ³ 30 (large sample case), use standard normal test. The unbiased estimate of *s* in this case can be taken as $S = \sqrt{\dfrac{\sum\left(X_i - \overline{X}\right)^2}{n}}$, since the difference between n and n - 1 is negligible for large values of n. Note that the parent population may or may not be normal in this case.

Example 14: The yield of alfalfa from six test plots is 2.75, 5.25, 4.50, 2.50, 4.25 and 3.25 tonnes per hectare. Test at 5% level of significance whether this supports the contention that true average yield for this kind of alfalfa is 3.50 tonnes per hectare.

Notes

Solution.

We note that *s* is not given and n = 6 (< 30), $\therefore$ t - test is applicable.

Using sample information we have

$$\overline{X} = \frac{2.75 + 5.25 + 4.50 + 2.50 + 4.25 + 3.25}{6} = 3.75.$$

To calculate s, we define $u_i = \frac{X_i - 3.75}{0.25} = \left(X_i - 3.75\right) \times 4$

X_i	2.75	5.25	4.50	2.50	4.25	3.25
u_i	– 4	6	3	– 5	2	– 2
u_i^2	16	36	9	25	4	4

From the above table $\sum u_i^2 = 94$. Therefore, $s = 0.25\sqrt{\frac{94}{6-1}} = 1.085$

We have to test H_0 : *m* = 3.50 against H_a : *m* ¹ 3.50.

The test statistic $\frac{\overline{X} - \mu_0}{s / \sqrt{n}}$ ~ t - distribution with (n - 1) d.f.

Thus, $t_{cal} = \left|\frac{3.75 - 3.50}{1.085 / \sqrt{6}}\right| = 0.564$

Further, the critical value of t, from table at 5% level of significance and with 5 d.f. is 2.571. Since t_{cal} is less than this value, there is no evidence against at 5% level of significance.

Example 15: Daily sales figures of 40 shopkeepers showed that their average sales and standard deviation were Rs 528 and Rs 600 respectively. Is the assertion that daily sales on the average is Rs 400, contradicted at 5% level of significance by the sample?

Solution.

Since n > 30, standard normal test is applicable. It is given that n = 40, $\overline{X}$ = 528 and S = 600.

We have to test H_0 : μ = 400 against H_a : μ ≠ 400.

$$z_{cal} = \left|\frac{528 - 400}{600 / \sqrt{40}}\right| = 1.35.$$

Since this value is less than 1.96, there is no evidence against H_0 at 5% level of significance. Hence, the given assertion is not contradicted by the sample.

15.2.3 Test of Hypothesis Concerning Equality of two Population Means

If random samples are obtained from each of the two normal populations, refer to § 20.2.2, the sampling distribution of the difference of their means is given by

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

Case I. If σ_1 and σ_2 are known, use standard normal test.

(a) To test $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$ (two tailed test), the test statistic is

$$z_{cal} = \frac{\left|\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ under } H_0.$$

This value is compared with 1.96 (2.58) for 5% (1%) level of significance.

(b) To test $H_0 : \mu_1 \leq \mu_2$ against $H_a : \mu_1 > \mu_2$ (one tailed test), the test statistic is $z_{cal} = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$,

and the critical value for 5% (1%) level of significance is 1.645 (2.33).

(c) To test $H_0 : \mu_1 \geq \mu_2$ against $H_a : \mu_1 < \mu_2$ (one tailed test), the test statistic, i.e., z_{cal} is same as in (b) above, however, the critical value for 5% (or 1%) level of significance is - 1.645 (or - 2.33).

Case II. If σ_1 and σ_2 are not known, their estimates based on samples are used. This category of tests can be further divided into two sub-groups.

1. Small Sample Tests (when either n_1 or n_2 or both are less than or equal to 30). To test $H_0 : \mu_1 = \mu_2$, we use t - test. The respective estimates of σ_1 and σ_2 are given by

$$s_1 = \sqrt{\frac{\sum\left(X_{1i} - \bar{X}_1\right)^2}{n_1 - 1}} = S_1\sqrt{\frac{n_1}{n_1 - 1}} \text{ and } s_2 = \sqrt{\frac{\sum\left(X_{2i} - \bar{X}_2\right)^2}{n_2 - 1}} = S_2\sqrt{\frac{n_2}{n_2 - 1}}$$

This test is more restrictive because it is based on the assumption that the two samples are drawn from independent normal populations with equal standard deviations, i.e., $\sigma_1 = \sigma_2 = \sigma$ (say). The pooled estimate of σ, denotes by s, is defined as

$$s = \sqrt{\frac{\sum\left(X_{1i} - \bar{X}_1\right)^2 + \sum\left(X_{2i} - \bar{X}_2\right)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2}{n_1 + n_2 - 2}}$$

(a) To test $H_0 : m_1 = m_2$ against $H_a : m_1 {}^1 m_2$ (two tailed test), the test statistic is

$$t_{cal} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{s} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}},$$ which follows t - distribution with $(n_1 + n_2 - 2)$ d.f.

This value is compared with the value of t from tables, to be denoted as $t_{a/2}(n_1 + n_2 - 2)$, at 100*a*% level of significance with $(n_1 + n_2 - 2)$ d.f.

(b) To test $H_0 : \mu_1 \le \mu_2$ against $H_a : \mu_1 > \mu_2$ (one tailed test), the test statistic is $t_{cal} = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{s} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$. This value is compared with $t_a(n_1 + n_2 - 2)$ from tables.

(c) To test $H_0 : \mu_1 \ge \mu_2$ against $H_a : \mu_1 < \mu_2$ (one tailed test), the test statistic, i.e., t_{cal} is same as in (b) above. This value is compared with $- t_a(n_1 + n_2 - 2)$.

2. Large Sample Tests (when both n_1 and n_2 is greater than 30)

In this case s_1 and s_2 are estimated by their respective sample standard deviations S_1 and S_2.

The test statistics for two and one tailed tests are $z_{cal} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ and $z_{cal} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ respectively. The remaining procedure is same as in case I above.

Remarks:

1. 100(1 - *a*)% confidence limits for $m_1 - m_1$ are given by $\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2} S.E._{\left(\bar{X}_1 - \bar{X}_2\right)}$.

   If $\bar{X}_1 - \bar{X}_2$ ~ t - distribution, $z_{a/2}$ is replaced by $t_{a/2}(n_1 + n_2 - 2)$.

2. If the two sample are drawn from populations with same standard deviations, i.e., $s_1 = s_2 = s$ (say), then $S.E._{\left(\bar{X}_1 - \bar{X}_2\right)} = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ for problems covered under case I and $S.E._{\left(\bar{X}_1 - \bar{X}_2\right)} = S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ for problems covered under case II, large sample tests. S is a pooled estimate of *s*, is given by

$$S = \sqrt{\frac{\sum\left(X_{1i} - \bar{X}_1\right)^2 + \sum\left(X_{2i} - \bar{X}_2\right)^2}{n_1 + n_2}} = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}}$$

Example 16: An investigation of the relative merits of two kinds of flashlight batteries showed that a random sample of 100 batteries of brand X lasted on the average 36.5 hours with a standard deviation of 1.8 hours, while a random sample of 80 batteries of brand Y lasted on the average 36.8 hours with a standard deviation of 1.5 hours. Use a level of significance of 1% to test whether the observed difference between average life times is significant.

Solution.

Let X and Y denote the life time of flashlight batteries of type X and type Y respectively and let μ_X and μ_Y be their respective population means.

It is given that $\bar{X} = 36.5$, $S_X = 1.8$, $n_X = 100$, $\bar{Y} = 36.8$, $S_Y = 1.5$, $n_Y = 80$.

We have to test H_0 : $\mu_X = \mu_Y$ against H_a : $\mu_X \neq \mu_Y$.

Since sample sizes are large (> 30), it is a large sample case.

The test statistic is $z_{cal} = \dfrac{|36.5 - 36.8|}{\sqrt{\dfrac{1.8^2}{100} + \dfrac{1.5^2}{80}}} = \dfrac{0.3}{0.246} = 1.219$

Since this value is less than 2.58, there is no evidence against H_0 at 1% level of significance and thus, the observed difference between average life times cannot be regarded as significant.

Example 17: Measurements performed on random samples of two kinds of cigarettes yielded the following results on their nicotine content (in mgs)

Brand A : 21.4, 23.6, 24.8, 22.4, 26.3

Brand B : 22.4, 27.7, 23.5, 29.1, 25.8

Assuming that the nicotine content is distributed normally, test the hypothesis that brand B has a higher nicotine content than brand A.

Solution.

We have to test $H_0 : \mu_A \geq \mu_B$ against $H_a : \mu_A < \mu_B$.

Note that the rejection of H_0 would imply that brand B has a higher nicotine content than brand A.

The means of the two samples are

$$\bar{X}_A = \frac{21.4 + 23.6 + 24.8 + 22.4 + 26.3}{5} = 23.7$$

and $$\bar{X}_B = \frac{22.4 + 27.7 + 23.5 + 29.1 + 25.8}{5} = 25.7.$$

Also $\sum\left(X_{Ai} - \bar{X}_A\right)^2 = 14.96 \; and \; \sum\left(X_{Bi} - \bar{X}_B\right)^2 = 31.30$

The pooled estimate of s is $s = \sqrt{\dfrac{14.96 + 31.30}{5 + 5 - 2}} = 2.40$

Thus, the test statistic is $t_{cal} = \dfrac{(23.7 - 25.7)}{2.40} \times \sqrt{\dfrac{5 \times 5}{5 + 5}} = -1.318.$

The critical value of t at 5% level of significance and 8 d.f. is - 1.86. Since t_{cal} is greater than this value, it lies in the region of acceptance and hence, there is no evidence against at 5% level of significance. Thus, the nicotine content in brand B is not higher than in brand A.

Example 18: Two salesmen 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.

Since n_1, n_2 < 30, it is a small sample case.

We have to test $H_0 : \mu_A = \mu_B$ against $H_a : \mu_A \neq \mu_B$.

Assuming that the two samples have come from the same population with S.D. *s*, we find its pooled estimate as

$$s = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{20 \times 20^2 + 18 \times 25^2}{36}} = 23.12$$

Also $t_{cal} = \frac{|170 - 205|}{23.12}\sqrt{\frac{20 \times 18}{20 + 18}} = 4.66.$ This value is highly significant, therefore, H_0 is rejected at 5% level of significance.

Example 19: The mean life of a random sample of 10 light bulbs was found to be 1456 hours with a S.D. of 423 hours. A second sample of 17 bulbs chosen at random from a different batch showed a mean life of 1280 hours with S.D. of 398 hours. Is there a significant difference between the mean life of the two batches?

Solution.

Note that the two samples have been obtained from the same population with unknown *s*.

We have to test $H_0 : m_1 = m_2$ against $H_a : m_1$ ¹ m_2.

It is given that $\overline{X}_1 = 1456$, $S_1 = 423$, $n_1 = 10$, $\overline{X}_2 = 1280$, $S_2 = 398$, $n_2 = 17$.

The pooled estimate of *s* is $s = \sqrt{\frac{10 \times 423^2 + 17 \times 398^2}{10 + 17 - 2}} = 423.42$

Therefore $t_{cal} = \frac{|1456 - 1280|}{423.42} \times \sqrt{\frac{10 \times 17}{10 + 17}} = 1.04$

The value of t from table at 5% level of significance and with 25 d.f. is 2.06. Since t_{cal} is less than this value, there is no evidence against H_0. Hence, the observed difference in mean life of bulbs of the two batches can be regarded as due to fluctuations of sampling.

When the Hypothesized Difference is not Zero

Let H_0: m_1 £ m_2 + k against H_a: $m_1 > m_2 + k$, where k is constant. The above can also be written as.

H_0: $m_1 - m_2$ £ k against H_a: $m_1 - m_2 > k$

Thus we can write

$$\overline{X}_1 - \overline{X}_2 \sim \mathrm{N}\left(k, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

or $$Z_{cal} = \frac{\left|\overline{X}_1 - \overline{X}_2 - k\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ under H_0.

In a similar way, we can write the expressions for t_{cal} under different situations.

Example 20: A sample of 100 electric bulbs of 'Philips' gave a mean life of 1500 hours with a standard deviation of 60 hours. Another sample of 100 electric bulbs of "HMT" gave a mean life of 1615 hours with a standard deviation of 80 hours. Can we conclude that the mean life of 'HMT' bulbs is greater then that of 'Philips' bulbs by 100 hours?

Let $\overline{X}_1 = 1615, S_1 = 80, n_1 = 100, \overline{X}_2 = 1500, S_2 = 60, n_2 = 100$.

We can write

H_0: $\mu_1 \le \mu_2 + 100$ against H_a: $\mu_1 > \mu_2 + 100$

$$Z_{cal} = \frac{\left|1615 - 1500 - 100\right|}{\sqrt{\frac{80^2}{100} + \frac{60^2}{100}}} = 1.5$$

Since $Z_{cal} < 1.645$, we accept H_0 at 5% and say that the difference in mean life of 'HMT' bulbs and that of 'Philips' bulbs is less than or equal to 100 hours.

15.2.4 Paired t - Test

This test is used in situations where there is a pairing of observations (X_{1i}, X_{2i}), like marks obtained by students of a class in two subjects, performance of the patients before and after the administration of a drug, etc. We define $d_i = X_{1i} - X_{2i}$, the difference in the observations for the i th item.

Then, we compute $\overline{d} = \frac{\sum d_i}{n}$ and $s_d = \sqrt{\frac{\Sigma\left(d_i - \overline{d}\right)^2}{n-1}} = \sqrt{\frac{\Sigma d_i^2 - n\overline{d}^2}{n-1}}$

As before, we can test $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \ne \mu_2$ (two tailed test) or $H_0 : \mu_1 \le$ (or $\ge$) μ_2 against $H_a : \mu_1 >$ (or $<$) μ_2 (one tailed test).

The test statistic $t = \frac{\left|\overline{d}\right|}{s_d / \sqrt{n}} = \frac{\left|\overline{d}\right|\sqrt{n}}{s_d} \sim t$-distribution with (n - 1) d.f.

Example 21: Eleven students of B.Com. (Hons) were given a test in economic analysis. They were imparted a month's special coaching and a second test was held at the end of it. The result were as follows :

Student No.	:	1	2	3	4	5	6	7	8	9	10	11
Marks in 1*st Test*	:	36	40	36	34	46	32	38	46	40	38	42
Marks in 2*nd Test*	:	40	44	40	40	46	40	34	48	38	44	36

Do the marks give an evidence that the students have benefited by extra coaching?

Solution.

We have to test $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$.

Note that H_0 implies that students have not benefited by the extra coaching.

Let X_1 and X_2 denote the marks in 1st and 2nd tests respectively.

Calculation of $\overline{d}$ and s_d

Student No.	:	1	2	3	4	5	6	7	8	9	10	11
d_i	:	-4	-4	-4	-6	0	-8	4	-2	2	-6	6
d_i^2	:	40	44	40	40	46	40	34	48	38	44	36

From the above table, we can write $\sum d_i = -22$ and $\sum d_i^2 = 244$

Thus, $\overline{d} = \frac{-22}{11} = -2$ and $s_d = \sqrt{\frac{244 - 11 \times 4}{10}} = 4.47$

Further, $t_{cal} = \frac{\left|-2 \times 11\right|}{4.47} = 1.48$. The value of t at 5% level of significance and 10 d.f. is 2.228. Therefore, the sample information provides no evidence that students have benefited by extra coaching.

Example 22: A random sample of heights of 20 students gave a mean of 68 inches with S.D. of 3 inches. Test the hypothesis that mean height in population is 70 inches under the assumption that the heights are normally distributed. Also construct a 95% confidence interval for the population mean.

Solution.

We have to test $H_0 : \mu_1 = 70$ against $H_a : \mu_1 \neq 70$.

It is given that n = 20, $\overline{X} = 68$ and S = 3.

The unbiased estimate of s.d. is $s = S\sqrt{\frac{n}{n-1}} = 3\sqrt{\frac{20}{19}} = 3.08$.

$$\therefore\ S.E._{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{3.08}{\sqrt{20}} = 0.688.$$

Alternatively, we can directly write

$$S.E._{\overline{X}} = \frac{s}{\sqrt{n}} = S\sqrt{\frac{n}{n-1}} \times \frac{1}{\sqrt{n}} = \frac{S}{\sqrt{n-1}} = \frac{3}{\sqrt{19}} = 0.688.$$

Thus, $t_{cal} = \frac{\left|68 - 70\right| \times \sqrt{19}}{3} = 2.906$

This value is greater than 2.093, the value of t from tables at 5% level of significance and 19 d.f. Thus, H_0 is rejected.

The 100(1 - a)% confidence limits for m are $\overline{X} \pm t_{\alpha/2} S.E._{\overline{X}}$.

Thus, the 95% confidence limits for m are given by $68 \pm 2.093 \times \frac{3}{\sqrt{19}} = 68 \pm 1.44$, i.e., 66.56 and 69.44 inches.

Example 23: Ten individuals are chosen at random from a normal population and their weights (in kgs) are found to be 63, 63, 66, 67, 68, 69, 70, 70, 71, 71. In the light of this data, discuss the suggestion that the mean height in the population is 66 inches.

Solution.

We have to test $H_0 : \mu = 66$ against $H_a : \mu \neq 66$.

From the given data, we can compute $\overline{X} = 67.8$ and $s = 3.01$.

$$\therefore \quad t_{cal} = \frac{\left|(67.8 - 66.0)\sqrt{10}\right|}{3.01} = 1.89.$$

This value is less than 2.262, the value of t from tables for 9 d.f. at 5% level of significance. Thus, there is no evidence against H_0.

15.3 Tests of Hypothesis concerning Proportion

Like the tests concerning sample mean, the null hypothesis to be tested would be either $\pi = \pi_0$, i.e., the proportion of successes in population is π_0 or $\pi_1 = \pi_2$, i.e., two populations have the same proportion of successes. These tests are based upon the sampling distribution of p, the proportion of successes in sample and the sampling distribution of $p_1 - p_2$, the difference between two sample proportions.

15.3.1 Test of Hypothesis that Population Proportion is π_0

The null hypothesis to be tested is $H_0 : \pi = \pi_0$ against $H_a : \pi \neq \pi_0$ for a two tailed test and $\pi >$ or $< \pi_0$ for a one tailed test. The test statistic is

$$z_{cal} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} = (p - \pi_0)\sqrt{\frac{n}{\pi_0(1 - \pi_0)}}$$

Remarks: The 100(1 - *a*)% confidence limits for *p* are $p \pm z_{\alpha/2}$S.E.(p).

Example 24: A wholesaler in apples claims that only 4% of the apples supplied by him are defective. A random sample of 600 apples contained 36 defective apples. Test the claim of the wholesaler.

Solution.

We have to test $H_0 : \pi \leq 0.04$ against $H_a : \pi > 0.04$.

It is given that $p = \frac{36}{600} = 0.06$ and n = 600.

Notes

$$\therefore \qquad z_{cal} = (0.06 - 0.04)\sqrt{\frac{600}{0.04 \times 0.96}} = 2.5$$

This value is highly significant in comparison to 1.645, therefore, H_0 is rejected at 5% level of significance.

Example 25: The manufacturer of a spot remover claims that his product removes at least 90% of all spots. What can be concluded about his claim at the level of significance *a* = 0.05, if the spot remover removed only 174 of the 200 spots chosen at random from the spots on clothes brought to a dry cleaning establishment?

Solution.

We have to test $H_0 : \pi \geq 0.9$ against $H_a : p < 0.9$.

It is given that $p = \frac{174}{200} = 0.82$ and n = 200.

$$\therefore \qquad z_{cal} = (0.82 - 0.90)\sqrt{\frac{200}{0.9 \times 0.1}} = -\,3.77$$

Since this value is less than - 1.645, H_0 is rejected at 5% level of significance. Thus, the sample evidence does not support the claim of the manufacturer.

Example 26: 470 heads were obtained in 1,000 throws of an unbiased coin. Can the difference between the proportion of heads in sample and their proportion in population be regarded as due to fluctuations of sampling?

Solution.

We have to test $H_0 : \pi = 0.5$ against $H_a : \pi \neq 0.5$.

It is given that $p = \frac{470}{1000} = 0.47$ and n = 1000.

$$\therefore \qquad z_{cal} = |0.47 - 0.50|\sqrt{\frac{1000}{0.5 \times 0.5}} = 1.897.$$

Since this value is less than 1.96, the coin can be regarded as fair and thus, the difference between sample and population proportion of heads are only due to fluctuations of sampling.

15.3.2 Test of Hypothesis Concerning Equality of Proportions

The null hypothesis to be tested is $H_0 : \pi_1 = \pi_2$ against $H_a : \pi_1 \neq \pi_2$ for a two tailed test and $\pi_1 >$ or $< \pi_2$ for a one tailed test.

The test statistic is $z_{cal} = (p_1 - p_2)\sqrt{\frac{n_1 n_2}{\pi(1-\pi)(n_1+n_2)}}$ under the assumption that $\pi_1 = \pi_2 = \pi$, where π is known. Often population proportion p is unknown and it is estimated on the basis of samples. The pooled estimate of π, denoted by p, is given by $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$.

Thus, the test statistic becomes $z_{cal} = (p_1 - p_2)\sqrt{\frac{n_1 n_2}{p(1-p)(n_1+n_2)}}$.

Remarks: 100(1 - α)% confidence limits of $(\pi_1 - \pi_2)$ can be written as

$$(p_1 - p_2) \pm z_{a/2} S.E.(p_1 - p_2)$$

Example 27: In a random sample of 600 persons from a large city, 450 are found to be smokers. In another sample of 900 persons from another large city, 450 are smokers. Do the data indicate that the cities are significantly different with respect to the prevalence of smoking? Let the level of significance be 5%.

Solution.

We have to test $H_0 : \pi_1 = \pi_2$ against $H_a : \pi_1 \neq \pi_2$.

It is given that $n_1 = 600$, $n_2 = 900$, $X_1 = X_2 = 450$.

$$\therefore \quad p_1 = \frac{X_1}{n_1} = \frac{450}{600} = 0.75 \text{ and } p_2 = \frac{X_2}{n_2} = \frac{450}{900} = 0.50$$

The pooled estimate of p, i.e., $p = \frac{450+450}{600+900} = 0.6$

Thus, $z_{cal} = |0.75 - 0.50|\sqrt{\frac{600 \times 900}{0.6 \times 0.4 \times 1500}} = 9.682$

This value is highly significant, therefore, H_0 is rejected. Thus, the given samples indicate that the two cities are significantly different with regard to the prevalence of smoking.

Example 28: A company is considering two different television advertisements for the promotion of a new product. Management believes that advertisement A is more effective than advertisement B. Two test market areas with virtually identical consumer characteristics are selected ; advertisement A is used in one area and advertisement B is used in the other area. In a random sample of 60 customers who saw the advertisement A, 18 tried the product. In a random sample of 100 customers who saw advertisement B, 22 tried the product. Does this indicate that advertisement A is more effective than advertisement B, if a 5% level of significance is used?

Solution.

We have to test $H_0 : \pi_A \leq \pi_B$ against $H_a : \pi_A > \pi_B$.

It is given that $n_A = 60$, $X_A = 18$, $n_B = 100$ and $X_B = 22$.

Thus, $p_A = \frac{18}{60} = 0.30$ and $p_B = \frac{22}{100} = 0.22$.

Also, the pooled estimate of p, i.e., $p = \frac{18+22}{160} = 0.25$.

$$\therefore \quad z_{cal} = (0.30 - 0.22)\sqrt{\frac{60 \times 100}{0.25 \times 0.75 \times 160}} = 1.131$$

Notes

Since this value is less than 1.645, there is no evidence against H_0 at 5% level of significance. Thus, the sample information provides no indication that advertisement A is more effective than advertisement B.

Remarks:

As in the variable case, we can also test the hypothesis $\pi_1 = \pi_2 + k$. Since $\pi_1 \neq \pi_2$, pooling of proportions is not allowed for the computations of standard error of $p_1 - p_2$. The standard error in this case is

$$\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

15.5 Summary

- A hypothesis is a preconceived idea about the nature of a population or about the value of its parameters. The statements like the distribution of heights of students of a university is normally distributed, the number of road accidents per day in Delhi is 10, etc., are some examples of a hypothesis.
- The test of a hypothesis is a procedure by which we test the validity of a given statement about a population. This is done on the basis of a random sample drawn from it.
- The hypothesis to be tested is termed as Null Hypothesis, denoted by H_0. This hypothesis asserts that there is no difference between population and sample in the matter under consideration. For example, if H_0 is that population mean $\mu = \mu_0$, then we regard the random sample to have been obtained from a population with mean m_0.
- Corresponding to any H_0, we always define an Alternative Hypothesis. This hypothesis, denoted by H_a, is alternate to H_0 , i.e., if H_0 is false then H_a is true and vice-versa.
- If the manufacturer claims that the average mileage is more than 40 kms/litre rather than equal to 40 kms/litre, we have to use a one tailed test. Now we shall test $H_0 : \mu = 40$ against $H_a : \mu > 40$ and z_{cal} would be defined as $z_{cal} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$. Since this value is also equal to 4.47 and lies in the critical region, we reject at 5% level of significance. This implies that the claim of the manufacturer may be taken as correct.
- In one tailed tests the alternative hypothesis is expressed as a strict inequality and the null hypothesis as a weak inequality or simply equality.
- The decision rule can also be specified in terms of prob or p-value of the observed sample result. The p-value is the smallest level of significance at which the null hypothesis can be rejected. We define p-value

15.6 Keywords

Hypothesis: A hypothesis is a preconceived idea about the nature of a population or about the value of its parameters.

Power of a test: The power of a test is defined as the probability of rejecting a false null hypothesis. Since *b* is the probability of accepting a false hypothesis, the power of test is given by 1 - *b*. More precisely, we can write

Power of a test = P [Rejecting H_0/H_0 is false] = 1 – *b*

15.7 Self Assessment

1. Fill in the blanks:

   (i) The reciprocal of standard error of an estimator is ....

   (ii) ...... tailed test is used when H_0 is $\theta \geq$ or $\leq \theta_0$.

   (iii) For testing $H_0 : \mu = \mu_0$ or $\mu_1 = \mu_2$ (σ known), we always use ...... normal test.

   (iv) When $\sigma_1 = \sigma_2 = \sigma$ is not known, we compute its ...... estimate from sample.

   (v) The χ^2 - test is used to test $H_0 : \sigma = \sigma_0$ only in case of a ...... sample.

   (vi) The test of hypothesis regarding equality of standard deviations makes use of ...... statistics.

   (vii) The test of goodness of fit or of independence is always a ...... tailed test.

   (viii) When sample (from normal population) sizes are small and s_1 and s_2 are not known, the sampling distribution of the difference of sample means follows t - distribution under the assumption that ....

   (ix) The existence of a strong linear relationship between two variables implies that the regression coefficient is ....

   (x) Yate's correction for continuity is needed when sample size is ....

15.8 Review Questions

1. Certain motor oil is packed in tins holding 5 litres each. The filling machine can maintain this but with a S.D. of 0.15 litre. Two samples of 36 tins each are taken from the production line. If the sample means are 5.20 and 4.95 litres respectively, can we be 99% sure that the sample have come from a population of 5 liters?

   Hint : Check whether the two sample means lie in the interval $5 \pm \frac{2.58 \times 0.15}{6}$ or not.

2. The Industrial Placement Unit of Unisex Polytechnic believes that the average salary paid to the students during their industrial year is Rs 2,800. A sample of 17 of its own students reveals that their average salary is Rs 2,860 with a S.D. of Rs 105. Does this evidence suggest that the countrywide average salary is higher than Rs 2,800? Let the level of significance be 5%.

   Hint : Use one tailed test.

4. In a survey of buying habits, 400 women shoppers are chosen at random in super market A located in a certain section of the city. Their average weekly food expenditure is Rs 250 with a S.D. of Rs 40. For 400 women shoppers chosen at random in super market B in another section of the city, the average weekly food expenditure is Rs 220 with a S.D. of Rs 55. Test at 1% level of significance whether the average weekly food expenditure of the population of shoppers are equal?

   Hint : Apply two tailed test to test the hypothesis regarding equality of means. Also note that both the samples are large.

5. Samples of two types of electric bulbs were tested for length of life (in hours) and the following data were obtained :

	Type I	*Type II*
Sample size	9	8
Mean of the sample	1235	1125
S.D. of the sample	30	35

Test, at 5% level of significance, whether the difference in sample means is significant?

Hint : Use small sample test, i.e., t-test.

6. A company selects 9 salesmen at random and their sales figures for the previous month are recorded. These salesmen then undergo a course devised by a business consultant and their sales figures for the following month are compared as shown in the following table. Has the training course caused an improvement in the salesmen's ability? Let the level of significance be 5%.

Previous Month	75	90	94	85	100	90	69	70	64
Following Month	77	101	93	92	105	88	73	76	68

Hint : Use paired t-test to test $H_0 : \mu_1 \geq \mu_2$ against $H_a : \mu_1 < \mu_2$.

7. A trader wants to compare the delivery times for two suppliers A and B. The trader wishes to continue with his current supplier A if his mean delivery time is less than or equal to that of supplier B, otherwise will switch over to B. He has obtained the following two independent samples for the above purpose :

$$\text{Supplier A} : n_1 = 40, \ \overline{X}_1 = 10 \text{ days}, \ S_1 = 3 \text{ days}$$

$$\text{Supplier B} : n_2 = 30, \ \overline{X}_2 = 8 \text{ days}, \ S_2 = 4 \text{ days}.$$

Answers: Self Assessment

1. (i) precision (ii) one (iii) standard (iv) pooled (v) small (vi) F (vii) one (viii) $\sigma_1 = \sigma_2$ (ix) significant (x) small.

15.9 Further Readings

Books

Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.

Unit 16: Hypothesis Concerning Standard Deviation

Notes

CONTENTS

Objectives

After studying this unit, you will be able to:

- Discuss Hypothesis concerning population standard deviation (n ≤ 30)
- Describe Hypothesis concerning population for large sample.

Introduction

In last unit you have studied about hypothesis testing. In this unit you will be studying about hypothesis concerning standard deviation.

These tests can be divided into two broad categories depending upon whether the size of the sample is large or small.

16.1 Test of Hypothesis Concerning Population Standard Deviation (n $\leq$ 30)

Refer to § 20.4.1, the statistic $\frac{\sum\left(X_i - \bar{X}\right)^2}{\sigma^2}$ *or* $\frac{nS^2}{\sigma^2}$ is a χ^2 - variate with (n - 1) degrees of freedom.

Under $H_0 : \sigma = \sigma_0$ (or $\sigma^2 = \sigma_0^2$), $\frac{nS^2}{\sigma_0^2}$ would be a χ^2 - variate with (n - 1) degrees of freedom.

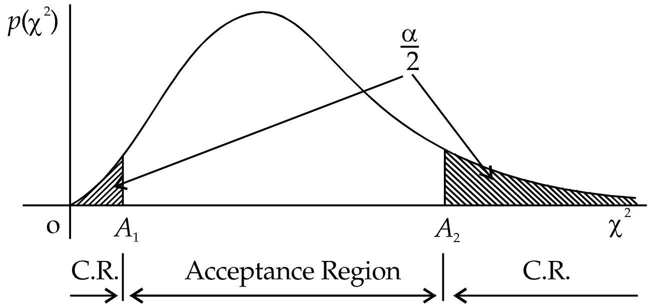
Example 1: A random sample of 20 bulbs from a large lot revealed a standard deviation of 150 hours. Assuming that the life of bulbs follow normal distribution, test the hypothesis that the standard deviation of the population is 130 hours.

Solution.

We have to test $H_0 : \sigma = 130$ against $H_a : \sigma \neq 130$ (two tailed test).

The test statistics, under H_0 is $\chi^2_{cal} = \dfrac{20 \times 150^2}{130^2} = 26.63$.

Figure 16.1



From the table of χ^2 at 5% level of significance and 19 degrees of freedom, the critical values are $A_1 = 8.91$ and $A_2 = 32.9$. Since χ^2_{cal} lies in the acceptance region, there is no evidence against H_0.

Remarks: To write (1 - *a*)% confidence interval for s^2, we write

$$P(A_1 \le c^2 \le A_2) = 1 - a \quad \text{or} \quad P\left(A_1 \le \frac{nS^2}{\sigma^2} \le A_2\right) = 1 - \alpha$$

The inequality $A_1 \le \dfrac{nS^2}{\sigma^2}$ can be written as $\sigma^2 \le \dfrac{nS^2}{A_1}$. Similarly, we can write $\dfrac{nS^2}{A_2} \le \sigma^2$. Thus, the (1 - *a*)% confidence interval for s^2 is given by

$$P\left(\frac{nS^2}{A_2} \le \sigma^2 \le \frac{nS^2}{A_1}\right) = 1 - \alpha.$$

Example 2: The standard deviation of a random sample of 25 units, taken from a normal population with $s = 8.5$, was calculated to be 10.8. Test the hypothesis that the observed value of standard deviation is significantly higher than the population standard deviation.

Solution.

We have to test $H_0 : \sigma = 8.5$ against $H_a : \sigma > 8.5$. (one tailed test)

The test statistic is $\chi^2_{cal} = \dfrac{25 \times 10.8^2}{8.5^2} = 40.36.$

Notes

χ^2 from tables at 5% level of significance and 24 d.f. is 36.4. Since this value is less than the calculated value, H_0 is rejected. Thus, the observed value of standard deviation is significantly higher than the population standard deviation.

16.2 Test of Hypothesis Concerning Population Standard Deviation (Large Sample)

It can be shown that for large samples (n > 30), the sampling distribution of S is approximately normal with mean *s* and standard error $\frac{\sigma}{\sqrt{2n}}$. Thus,

$$z = \frac{(S-\sigma)\sqrt{2n}}{\sigma} \sim N(0,1).$$

Alternatively, using Fisher's approximation, we can say that when n > 30, the statistic $\sqrt{2\chi^2}$ follows a normal distribution with mean $\sqrt{2n}$ and standard error unity. Thus $z = \sqrt{2\chi^2} - \sqrt{2n}$ can be taken as standard normal variate for sufficiently large values of n.

Example 3: In a random sample of 300 units, the standard deviation was found to be 8.5. Can it reasonably be regarded as to have come from a population with standard deviation equal to 9.0?

Solution.

We have to test $H_0 : \sigma = 9.0$ against $H_a : \sigma \neq 9.0$ (two tailed test).

It is given that S = 8.5 and n = 300 (large).

Thus, the test statistic is $z_{cal} = \frac{|8.5 - 9.0|\sqrt{600}}{9.0} = 1.36.$

Since this value is less than 1.96, there is no evidence against H_0 at 5% level of significance.

Note: The same value of z is obtained by the use of the statistic $z = \sqrt{2\chi^2} - \sqrt{2n}$.

We can write

$$z_{cal} = \left|\sqrt{\frac{2nS^2}{\sigma^2}} - \sqrt{2n}\right| = \left|\sqrt{\frac{2 \times 300 \times 8.5^2}{9.0^2}} - \sqrt{600}\right| = 1.36$$

If *s* is unknown it is estimated by S. The 95% confidence limits for *s* are $S \pm 1.96\frac{S}{\sqrt{2n}}$ *or* $S\left(1 \pm \frac{1.96}{\sqrt{2n}}\right)$.

Notes

Example 4: The standard deviation of a random sample of size 81 was found to be 12. Test the hypothesis that population standard deviation is greater than 10.

Solution.

We have to test H_0 : *s* £ 10 against H_a : *s* > 10.

$$z = \frac{(12-10)}{10}\sqrt{2\times 81} = 2.55.$$

Since this value is greater than 1.645, H_0 is rejected. Hence, the sample information supports the contention that *s* is greater than 10.

16.3 Test of Hypothesis Concerning the Equality of Standard Deviations (Small Samples)

We have to test $H_0 : \sigma_1 = \sigma_2$ against $\sigma_1 > \sigma_2$. Refer to § 20.6, the statistic $F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$, would become $\frac{s_1^2}{s_2^2}$ under H_0, follows F - distribution with n_1 (= n_1 - 1) and n_2 (= n_2 - 1) degrees of freedom.

Remarks:

1. We can write $s_1^2 = \frac{1}{n_1-1}\sum\left(X_{1i}-\bar{X}_1\right)^2 = \frac{n_1}{n_1-1}S_1^2 = \frac{1}{n_1-1}\left(\sum X_{1i}^2 - \frac{\sum X_{1i}^2}{n_1}\right)$ and

$$s_2^2 = \frac{1}{n_2-1}\sum\left(X_{2i}-\bar{X}_2\right)^2 = \frac{n_2}{n_2-1}S_2^2 = \frac{1}{n_2-1}\left(\sum X_{2i}^2 - \frac{\sum X_{2i}^2}{n_2}\right).$$

2. In the variance ratio $F = \frac{s_1^2}{s_2^2}$, we take, by convention the largest of the two sample variance as σ_1^2. Thus, this test is always a one tailed test with critical region at the right hand tail of the F - curve.

3. The 100(1 - *a*)% confidence limits for the variance ratio $\frac{\sigma_1^2}{\sigma_2^2}$, are given by

$$P\left[\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{1-\alpha/2}}\right] = 1-\alpha.$$

Example 5: Two independent samples of sizes 10 and 12 from two normal populations have their mean square deviations about their respective means equal to 12.8 and 15.2 respectively. Test the equality of variances of the two populations.

Solution.

We have to test $H_0 : \sigma_1 = \sigma_2$ against $\sigma_1 > \sigma_2$.

It is given that $S_1^2 = 15.2$, $S_2^2 = 12.8$, $n_1 = 12$ and $n_2 = 10$.

The unbiased estimates of respective population variances are

$$s_1^2 = \frac{12}{11} \times 15.2 = 16.58 \ \ and \ \ s_2^2 = \frac{10}{9} \times 12.8 = 14.22.$$

Thus, $F_{cal} = \dfrac{16.58}{14.22} = 1.166.$

The value of F from tables at 5% level of significance with 11 and 9 d.f. is 3.10. Since this value is greater than F_{cal}, there is no evidence against H_0.

Example 6: The increase in weight (in 100 gms) due to food A and food B given to two independent samples of children was recorded as follows. Test whether (i) mean weights and (ii) standard deviations of the two samples are equal.

Sample I : 6, 12, 10, 14, 12, 12, 10, 7, 5, 7.

Sample II : 9, 11, 8, 5, 6, 12, 7, 13, 10.

Solution.

We shall first test $H_0 : \sigma_1 = \sigma_2$ against $\sigma_1 > \sigma_2$.

The means of the samples are $\bar{X}_1 = \dfrac{95}{10} = 9.5 \ and \ \bar{X}_2 = \dfrac{81}{9} = 9.0$, respectively.

We can write $s_k^2 = \dfrac{n_k}{n_k - 1}\left(\dfrac{\sum X_{ki}^2}{n_k} - \bar{X}_k^2\right) = \dfrac{\sum X_{ki}^2}{n_k - 1} - \dfrac{n_k}{n_k - 1}\bar{X}_k^2$ (k = 1, 2)

Thus, we have $s_1^2 = \dfrac{987}{9} - \dfrac{10}{9} \times 9.5^2 = 9.39$ and $s_2^2 = \dfrac{789}{8} - \dfrac{9}{8} \times 9^2 = 7.50.$

Further, the test statistic is $F = \dfrac{9.39}{7.50} = 1.25.$

The critical value of F at 5% level of significance and (9,8) d.f. is 3.39, therefore, there is no evidence against H_0. Hence, s_1 and s_2 may be treated as equal.

To test $H_0 : \mu_1 = \mu_2$ against Ha ; $\mu_1 \neq \mu_2$, we note that samples are small, t-test is to be used. Since $\sigma_1 = \sigma_2 = s$ (say), its unbiased estimate is

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{9 \times 9.39 + 8 \times 7.50}{10 + 9 - 2}} = 2.92.$$

The test statistic is $t_{cal} = \dfrac{\left|\bar{X}_1 - \bar{X}_2\right|}{s}\sqrt{\dfrac{n_1 n_2}{n_1 + n_2}} = \dfrac{|9.5 - 9.0|}{2.92}\sqrt{\dfrac{10 \times 9}{10 + 9}} = 0.37.$

The critical value of t at 5% level of significance and 17 d.f. is 2.11. Since this value is greater than the calculated, there is no evidence against H_0. Thus, we conclude that the two samples may be regarded to have drawn from a population with same means and same standard deviations.

16.4 Test of Hypothesis Concerning Equality of Standard Deviations (Large Samples)

It can be shown that when sample sizes are large, i.e., n_1, n_2 > 30, the sampling distribution of the statistic S_1 - S_2 is approximately normal with mean s_1 - s_2 and standard error $\sqrt{\frac{\sigma_1^2}{2n_1}+\frac{\sigma_2^2}{2n_2}}$.

Therefore $z=\frac{\left(S_1-S_2\right)-\left(\sigma_1-\sigma_2\right)}{\sqrt{\frac{\sigma_1^2}{2n_1}+\frac{\sigma_2^2}{2n_2}}}\sim N\left(0,1\right)$

or $\quad z=\frac{S_1-S_2}{\sigma\sqrt{\frac{1}{2n_1}+\frac{1}{2n_2}}}$ under $H_0 : \sigma_1 = \sigma_2 = \sigma$.

Very often σ is not known and is estimated on the basis of sample. The pooled estimate of σ is $S=\frac{n_1S_1^2+n_2S_2^2}{n_1+n_2}$. Thus, the test statistic becomes

$$z_{cal}=\frac{S_1-S_2}{S\sqrt{\frac{1}{2n_1}+\frac{1}{2n_2}}}=\frac{S_1-S_2}{S}\times\sqrt{\frac{2n_1n_2}{n_1+n_2}}.$$

Example 7: The standard deviation of a random sample of the heights of 500 individuals from country A was found to be 2.58 inches and that of 600 individuals from country B was found to be 2.35 inches. Do the data indicate that the standard deviation of heights in country A is greater than that in country B?

Solution.

We have to test $H_0 : \sigma_1 = \sigma_2$ against $H_a : \sigma_1 > \sigma_2$.

It is given that S_1 = 2.58, n_1 = 500, S_2 = 2.35 and n_2 = 600.

The pooled estimate of s is $S=\sqrt{\frac{500\times 2.58^2+600\times 2.35^2}{1100}}=2.46$

The test statistic is $z_{cal}=\frac{2.58-2.35}{2.46}\times\sqrt{\frac{600000}{1100}}=2.17$

Since this value is greater than 1.645, H_0 is rejected at 5% level of significance. Thus, the sample evidence indicates that the standard deviation of heights in country A is greater.

16.5 Summary

- We can write $s_1^2 = \frac{1}{n_1-1}\sum\left(X_{1i}-\bar{X}_1\right)^2 = \frac{n_1}{n_1-1}S_1^2 = \frac{1}{n_1-1}\left(\sum X_{1i}^2 - \frac{\sum X_{1i}^2}{n_1}\right)$ and

  $$s_2^2 = \frac{1}{n_2-1}\sum\left(X_{2i}-\bar{X}_2\right)^2 = \frac{n_2}{n_2-1}S_2^2 = \frac{1}{n_2-1}\left(\sum X_{2i}^2 - \frac{\sum X_{2i}^2}{n_2}\right).$$

- In the variance ratio $F = \frac{s_1^2}{s_2^2}$, we take, by convention the largest of the two sample variance as σ_1^2. Thus, this test is always a one tailed test with critical region at the right hand tail of the F - curve.

- The 100(1 - *a*)% confidence limits for the variance ratio $\frac{\sigma_1^2}{\sigma_2^2}$, are given by

  $$P\left[\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{1-\alpha/2}}\right] = 1-\alpha.$$

16.6 Keywords

F - distribution: If $H_0 : \sigma_1 = \sigma_2$ against $\sigma_1 > \sigma_2$. Refer to § 20.6, the statistic $F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$, would become $\frac{s_1^2}{s_2^2}$ under H_0, follows F - distribution with n_1 (= n_1 - 1) and n_2 (= n_2 - 1) degrees of freedom.

16.7 Self Assessment

Fill in the blanks:

1. These tests can be divided into two broad categories depending upon whether the .... of the sample is large or small.

2. If $H_0 : \sigma_1 = \sigma_2$ against $\sigma_1 > \sigma_2$. Refer to § 20.6, the statistic $F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$, would become $\frac{s_1^2}{s_2^2}$ under H_0, follows .... with n_1 (= n_1 - 1) and n_2 (= n_2 - 1) degrees of freedom.

3. In the .... $F = \frac{s_1^2}{s_2^2}$, we take, by convention the largest of the two sample variance as σ_1^2. Thus, this test is always a one tailed test with critical region at the right hand tail of the F - curve.

4. The 100(1 - *a*)% .... for the variance ratio $\frac{\sigma_1^2}{\sigma_2^2}$, are given by

$$P\left[\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2}}\leq\frac{\sigma_1^2}{\sigma_2^2}\leq\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{1-\alpha/2}}\right]=1-\alpha.$$

16.8 Review Questions

1. Test the hypothesis that $\sigma = 8$, given that S = 10 for a random sample of size 51. Also construct 95% confidence interval for σ.

   Hint : Use a two tailed normal test.

2. A random sample of size 10 from a normal population gave the following observations : 169, 173, 171, 177, 161, 163, 174, 168, 172, 165.

   Test the hypothesis that population variance is 25.

   Hint : Use a two tailed χ^2 test.

3. The following two samples are drawn from two normal populations. Test at 5% level of significance whether their variance can be regarded as equal?

   Sample I : 60, 65, 71, 74, 76, 82, 85, 57.

   Sample II : 61, 66, 67, 85, 78, 63, 85, 86, 88, 91.

   Hint : Use F - test.

4. Can the following two samples obtained from two normal populations, be regarded to have same variances?

Sample No.	Sample Size	Sample Variance
1	15	20
2	25	35

   Test at 10% level of significance.

   Hint : Use F - test.

5. Two independent random samples, one of 12 observations with mean 15 and sum of squares of deviations from mean equal to 135 and another of 16 observations with mean 22 and sum of squares of deviations from mean equal to 250, were obtained from two normal populations. Test at 5% level of significance whether the two samples can be regarded to have come form the same population?

   Hint : Test $\sigma_1 = \sigma_2$ and $\mu_1 = \mu_2$ an in example 34.

6. The following figures relate to the number of units produced per shift by two workers A and B for a number of days:

   A : 19, 22, 24, 27, 24, 18, 20, 19 and 25.

   B : 26, 37, 40, 35, 30, 30, 40, 26, 30, 35 and 45.

   Can it be inferred that A is more stable worker compared to B? Answer using 5% level of significance.

   Hint : Use F - test.

Notes

7. In one sample of 10 observations from a normal population, the sum of squares of the deviations of sample values from their mean is 100.4 and in another sample of 12 observations from another normal population, the sum of squares of the deviations of sample values from their mean is 115.5. Test at 5% level whether the two normal populations have the same variance?

   Hint : Use F - test.

8. In a test given to two groups of students, 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.

   Assuming that the marks obtained follows normal distribution, examine at 5% level of significance whether the two groups of students can be regarded to have come from populations with same standard deviation?

   Hint : Use F - test.

Answers: Self Assessment

1. size 2. F - distribution 3. variance ratio 4. confidence limits

16.9 Further Readings

Books

Sheldon M. Ross, Introduction to Probability Models, Ninth Edition, Elsevier Inc., 2007.

Jan Pukite, Paul Pukite, Modeling for Reliability Analysis, IEEE Press on Engineering of Complex Computing Systems, 1998.