Master of Science Mathematics

(M.Sc. Mathematics)

MMT-205

MATHEMATICAL STATISTICS

Semester-II

Author- Dr. Harsh Vardhan Harsh

SURESH GYAN VIHAR UNIVERSITY

Centre for Distance and Online Education

Mahal, Jagatpura, Jaipur-302025

Published by:

S. B. Prakashan Pvt. Ltd.

WZ-6, Lajwanti Garden, New Delhi: 110046

Tel.: (011) 28520627 | Ph.: 9205476295

Email: info@sbprakashan.com | Web.: www.sbprakashan.com

Designed & Graphic by : S. B. Prakashan Pvt. Ltd.

Printed at :

Suresh Gyanvihar University
Department of Mathematics

M.Sc., Mathematics - Syllabus – I year – II Semester (Distance Mode)

COURSE TITLE : MATHEMATICAL STATISTICS

COURSE CODE : MMT-205

COURSE CREDIT : 4

COURSE OBJECTIVES

While studying the **MATHEMATICAL STATISTICS,** the Learner shall be able to:

CO 1: Demonstrate the concept of t distribution and F distribution.

CO 2: To impart knowledge about simple hypothesis, alternative hypothesis, Type I errors, Type II errors and critical regions.

CO 3: Explain the relationship between correlation analysis and regression analysis.

CO 4: To impart knowledge about to solve the problems in analysis of variance in one way and two way classifications, completely randomized design, randomized block design and Latin Square design.

CO 5: Summarize the partitioning the covariance matrix, sample mean vector and covariance matrix.

COURSE LEARNING OUTCOMES

After completion of the **MATHEMATICAL STATISTICS**, the Learner will be able to:

CLO 1: Familiarize with sampling distribution and to find estimators for the parameters

CLO 2: Analyze and compare the tests based on normal, t distribution, Chi-square distribution and F distribution for testing of mean, variance and population.

CLO 3: Demonstrate the problems in partial correlation, multiple correlation and multiple regression.

CLO 4: Explain the difference between Completely Randomized Design, Randomized Block Design and Latin Square Design.

CLO 5: To impart the knowledge of the concept of multivariate normal distribution, multivariate normal density and its properties.

BLOCK I:SAMPLING DISTRIBUTIONS AND ESTIMATION THEORY

Sampling distributions - Characteristics of good estimators - Method of Moments - Maximum Likelihood Estimation - Interval estimates for mean, variance and proportions.

BLOCK II:TESTING OF HYPOTHESIS

Type I and Type II errors - Tests based on Normal, t, □2 and F distributions for testing of mean, variance and proportions - Tests for Independence of attributes and Goodness of fit.

BLOCK III:CORRELATION AND REGRESSION

Method of Least Squares - Linear Regression - Normal Regression Analysis - Normal Correlation

Analysis - Partial and Multiple Correlation - Multiple Linear Regression.

BLOCK IV:DESIGN OF EXPERIMENTS

Analysis of Variance - One-way and two-way Classifications - Completely Randomized Design - Randomized Block Design - Latin Square Design.

BLOCK V:MULTIVARIATE ANALYSIS

Mean Vector and Covariance Matrices - Partitioning of Covariance Matrices - Combination ofRandom Variables for Mean Vector and Covariance Matrix - Multivariate, Normal Density and its Properties - Principal Components: Population principal components - Principal components from standardized variables.

REFERENCE BOOKS :

1. Freund J.E.," Mathematical Statistics", Prentice Hall of India, Fifth Edition, 2001.
2. Johnson R.A. and Wichern D.W., "Applied Multivariate Statistical Analysis", PearsonEducation Asia, Sixth Edition, 2007.
3. Gupta S.C. and Kapoor V.K.," Fundamentals of Mathematical Statistics", Sultan Chand &Sons, Eleventh Edition, 2003.
4. Devore J.L. "Probability and Statistics for Engineers", Brooks/Cole (Cengage Learning),First India Reprint, 2008.

CONTENTS

BLOCK I: Sampling Distributions and Estimation theory

Unit – 1

Sampling Distributions

Structure

Objectives

After Studying this Unit, the student will be able to

- Explain the sampling distribution of the mean.
- Demonstrate the t distribution and F distribution.
- Elaborate the chi-square distribution with example.

Overview

In this unit, we will study the concept of sampling distribution of the mean, the chi-square distribution, t distribution and F distribution.

1.1. Introduction

Statistics concerns itself mainly with conclusions and predictions resulting from chance outcomes that occur in carefully planned experiments or investigations. Drawing such conclusions usually involves taking sample observations from a given population and using the results of the sample to make inferences about the population itself, its mean, its variance, and so forth. To do this requires that we first fin the distributions of certain functions of the random variables whose value make up the sample, called statistics. The properties of these distributions then allow us to make probability statements about the resulting inferences drawn from the sample about the population.

1.1.1. Population

A set of numbers from which a sample is drawn is referred to as a population. The distribution of the numbers constituting a population is called the population distribution.

1.1.2. Random Sample

If $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables, we say that they constitute a random sample from the infinite population given by their common distribution.

1.1.3. Sample Mean and Sample Variance

If $X_1, X_2, \ldots, X_n$ constitute a random sample, then the sample mean is given by $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and the sample variance is given by $S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$

1.1.4. Remark

Let $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ for observed sample data and refer to these statistics as the sample mean and the sample variance. Here x_i, $\bar{x} and s^2$ are values of the corresponding random variables X_i, $\bar{X} and S^2$. The formulas for $\bar{x} and s^2$ are used even when we deal with any kind of data, not necessarily sample data, in which case we refer $\bar{x} and s^2$ simple as the mean and the variance.

1.2. The Sampling Distribution of the Mean

1.2.1. Theorem

If $X_1, X_2, \ldots, X_n$ constitute a random sample from an infinite population with the mean μ and the variance σ^2, then $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \frac{\sigma^2}{n}$

Proof:

Let $Y = \bar{X}$ and hence setting $a_i = \frac{1}{n}$, we get

$E(\bar{X}) = \sum_{i=1}^{n} \frac{1}{n}.\mu = n\left(\frac{1}{n}.\mu\right) = \mu$. Since $E(X_i) = \mu$.

Then, "If the random variables $X_1, X_2, \ldots, X_n$ are independent and $Y = \sum_{i=1}^{n} a_i X_i$, then $Var(Y) = \sum_{i=1}^{n} a_i^2\ Var(X_i)$"

$$Var(\bar{X}) = \sum_{i=1}^{n} \frac{1}{n^2}.\sigma^2 = n\left(\frac{1}{n^2}.\sigma^2\right) = \frac{\sigma^2}{n}$$

1.2.2. Remark

We write $E(\bar{X})$ as $\mu_{\bar{X}}$ and $Var(\bar{X})$ as $\sigma^2_{\bar{X}}$ and refer to $\sigma_{\bar{X}}$ as the standard error of the mean. The formula for the standard error of the mean, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, the standard deviation of the distribution of $\bar{X}$ decreases when n, the sample size, is increased. This means that when n becomes larger and we actually have more information, we can expect values of $\bar{X}$ to closer to μ, the quantity that they are intended to estimate.

1.2.3. Result

For any positive constant c, the probability that $\bar{X}$ will take on a value between $\mu - c$ to $\mu + c$ is at least $1 - \frac{\sigma^2}{n\,c^2}$, When $n \to \infty$, this probability approaches 1. This result, called a law of large numbers.

1.2.4. Theorem (Central Limit Theorem)

If $X_1, X_2, \ldots, X_n$ constitute a random sample from an infinite population with the mean μ, the variance σ^2, and the moment-generating function $M_X(t)$, then the limiting distribution of $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ as $n \to \infty$ is the standard normal distribution.

Proof:

If a and b are constants, then

1. $M_{X+a}(t) = E[e^{(X+a)t}] = e^{at}.M_X(t)$

2. $M_{bX}(t) = E[e^{bXt}] = M_X(bt)$

3. $M_{\frac{X+a}{b}}(t) = E\left[e^{\left(\frac{X+a}{b}\right)t}\right] = e^{\frac{a}{b}t}.M_X\left(\frac{t}{b}\right)$ ", we get

$$M_Z(t) = M_{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}(t) = e^{-\sqrt{n}\mu t/\sigma}.M_{\bar{X}}\left(\frac{\sqrt{n}\,t}{\sigma}\right)$$

$$M_Z(t) = M_{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}(t) = e^{-\sqrt{n}\mu t/\sigma}.M_{n\bar{X}}\left(\frac{t}{\sigma\sqrt{n}}\right)$$

Since $n\bar{X} = X_1 + X_2 + \cdots + X_n$

$$M_Z(t) = e^{-\sqrt{n}\mu t/\sigma}.\left[M_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$

and hence that

$$\ln M_Z(t) = -\frac{\sqrt{n}\,\mu t}{\sigma} + n.\ln M_X\left(\frac{t}{\sigma\sqrt{n}}\right)$$

Expanding $M_X\left(\frac{t}{\sigma\sqrt{n}}\right)$ as a power series in t, we obtain

$$\ln M_Z(t) = -\frac{\sqrt{n}\,\mu t}{\sigma} + n.\ln\left[1+\mu'_1\frac{t}{\sigma\sqrt{n}}+\mu'_2\frac{t^2}{2\sigma^2 n}+\mu'_3\frac{t^3}{6\sigma^3 n\sqrt{n}}+\cdots\right]$$

Where μ'_1, μ'_2 *and* μ'_3 are the moments about the origin of the population distribution, that is, those of the original random variables X_i.

If n is sufficiently large, we can use the expansion of $\ln(1+x)$ as a power series in x, getting

$$\ln M_Z(t) = -\frac{\sqrt{n}\,\mu t}{\sigma} + n\left\{\begin{array}{l}\left[\mu'_1\frac{t}{\sigma\sqrt{n}}+\mu'_2\frac{t^2}{2\sigma^2 n}+\mu'_3\frac{t^3}{6\sigma^3 n\sqrt{n}}+\cdots\right]^1 \\ -\frac{1}{2}\left[\mu'_1\frac{t}{\sigma\sqrt{n}}+\mu'_2\frac{t^2}{2\sigma^2 n}+\mu'_3\frac{t^3}{6\sigma^3 n\sqrt{n}}+\cdots\right]^2 \\ +\frac{1}{3}\left[\left[\mu'_1\frac{t}{\sigma\sqrt{n}}+\mu'_2\frac{t^2}{2\sigma^2 n}+\mu'_3\frac{t^3}{6\sigma^3 n\sqrt{n}}+\cdots\right]\right]^3 - \ldots\end{array}\right\}$$

Then, collecting powers of t, we obtain

$$\ln M_Z(t) = \left(-\frac{\sqrt{n}\,\mu}{\sigma}+\frac{\sqrt{n}\,\mu'_1}{\sigma}\right)t + \left(\frac{\mu'_2}{2\sigma^2}-\frac{\mu'^2_1}{2\sigma^2}\right)t^2 + \left(\frac{\mu'_3}{6\sigma^3\sqrt{n}}-\frac{\mu'_1\mu'_2}{2\sigma^3\sqrt{n}}+\frac{\mu'^3_1}{3\sigma^2\sqrt{n}}\right)t^3+\cdots.$$

and since $\mu'_1 = \mu$ and $\mu'_2 - (\mu'_1)^2 = \sigma^2$, this reduces to

$$\ln M_Z(t) = \frac{1}{2}t^2 + \left(\frac{\mu'_3}{6}-\frac{\mu'_1\mu'_2}{2}+\frac{\mu'^3_1}{6}\right)\frac{t^3}{\sigma^3\sqrt{n}}+\cdots$$

Finally, observing that the coefficient of t^3 is a constant times $\frac{1}{\sqrt{n}}$ and in general, for $r \geq 2$, the coefficient of t^r is a constant times $\frac{1}{\sqrt{n^{r-2}}}$, we get

$\lim\limits_{n\to\infty} ln\, M_Z(t) = \frac{1}{2}t^2$ and hence $\lim\limits_{n\to\infty} M_Z(t) = e^{\frac{1}{2}t^2}$

Since the limit of a logarithm equals the logarithm of the limit (Provided these limit exist).

1.2.5. Example

A soft-drink vending machine is set so that the amount of drink dispensed is a random variable with a mean of 200 milliliters and a standard deviation of 15 milliliters. What is the probability that the average (mean) amount dispensed in a random sample of size 36 is at least 204 milliliters?

Solution:

The distribution of $\bar{X}$ has the mean $\mu_{\bar{X}} = 200$ and the standard deviation $\sigma_{\bar{X}} = \frac{15}{\sqrt{36}} = 2.5$ and according to the central limit theorem, this distribution is approximately normal.

Since $Z = \frac{204-200}{2.5} = 1.6$

By Statistical table, we have

$P(\bar{X} \geq 204) = P(Z \geq 1.6) = 0.5 - 0.4452 = 0.0548$

1.2.6. Theorem

If $\bar{X}$ is the mean of a random sample of size n from a normal population with the mean μ and the variance σ^2, its sampling distribution is a normal distribution with the mean μ and the variance $\frac{\sigma^2}{n}$.

Proof:

If a and b are constants, then

1. $M_{X+a}(t) = E[e^{(X+a)t}] = e^{at}.M_X(t)$

2. $M_{bX}(t) = E[e^{bXt}] = M_X(bt)$

3. $M_{\frac{X+a}{b}}(t) = E\,[e^{(\frac{X+a}{b})t}] = e^{\frac{a}{b}t}.M_X\,(\frac{t}{b})$. If $X_1, X_2, \dots, X_n$ are independent random variables and $Y = X_1 + X_2, \dots + X_n$ then $M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t)$ where $M_{X_i}(t)$ is the value of the moment-generating function of X_i at t.

We can write $M_{\bar{X}}(t) = [M_X(\frac{t}{n})]^n$ and since the moment-generating function of a normal distribution with mean μ and σ^2 is given by $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

According to the theorem $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$, we ger

$M_{\bar{X}}(t) = [e^{\mu \cdot \frac{t}{n} + \frac{1}{2}(\frac{t}{n})^2 \sigma^2}]^n = e^{\mu t + \frac{1}{2}t^2(\frac{\sigma^2}{n})}$

Ths moment-generating funciton is a normal distribution with the mean μ and the variance $\frac{\sigma^2}{n}$.

1.3. The Sampling Distribution of the Mean: Finite Populations

1.3.1. Random Sample-Finite Population

If X_1 is the first value drawn from a finite population of size N, X_2 is the second value drawn,.....,X_n is the n^{th} value drawn, and the joint probability distribution of these n random variables is given by $f(x_1, x_2, \dots, x_n) = \frac{1}{N(N-1)\dots(N-n+1)}$ for each ordered n-tuple of values of these random variables, $X_1, X_2, \dots, X_n$ are said to be constitute a random sample from the given finite population.

1.3.2. Sample Mean and Variance – Finite Population

The sample mean and the sample variance of the finite population $\{c_1, c_2, \dots, c_N\}$ are $\mu = \sum_{i=1}^{N} c_i \cdot \frac{1}{N}$ and $\sigma^2 = \sum_{i=1}^{N} (c_i - \mu)^2 \cdot \frac{1}{N}$

1.3.3. Theorem

If X_r and X_s are the r^{th} and s^{th} random variables of a random sample of size n drawn from the finite population $\{c_1, c_2, \dots, c_N\}$, then $cov(X_r, X_s) = -\frac{\sigma^2}{N-1}$

Proof:

$cov(X_r, X_s) = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{N(N-1)} (c_i - \mu)(c_j - \mu), i \neq j.$

$$cov(X_r, X_s) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (c_i - \mu) \left[\sum_{j=1}^{N} (c_j - \mu), i \neq j\right]$$

and since $i \neq j$, $\sum_{j=1}^{N} (c_j - \mu) = \sum_{j=1}^{N} (c_j - \mu) - (c_i - \mu) = -(c_i - \mu)$, we get

$$cov(X_r, X_s) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (c_i - \mu)^2 = -\frac{\sigma^2}{N-1}.$$

1.3.4. Theorem

If $\bar{X}$ is the mean of a random sample of size n taken without replacement from a finite population of size N with the mean μ and the variance σ^2, then $E(\bar{X}) = \mu$ and

$$var(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

Proof:

Substituting $a_i = \frac{1}{N}$, $var(X_i) = \sigma^2$, and $cov(X_i, X_j) = -\frac{\sigma^2}{N-1}$ into the formula

$E(Y) = \sum_{i=1}^{n} a_i E(X_i)$, we get

$E(\bar{X}) = \sum_{i=1}^{n} \frac{1}{n} . \mu = \mu$ and

$$var(\bar{X}) = \sum_{i=1}^{n} \frac{1}{n^2} . \sigma^2 + \sum\sum_{i<j} \frac{1}{n^2} (- \frac{\sigma^2}{N-1})$$

$$var(\bar{X}) = \frac{\sigma^2}{n^2} + 2. \frac{n(n-1)}{2} . \frac{1}{n^2} (- \frac{\sigma^2}{N-1})$$

$$var(\bar{X}) = \frac{\sigma^2}{n} . \frac{N-n}{N-1}$$

1.4. The Chi-Square Distribution

If X has the standard normal distribution, then X^2 has the special gamma distribution, which is known as the chi-square distribution and it is denoted by χ^2.

If a random variable X has the chi-square distribution the v degrees of freedom if its probability density is given by

$$f(x) = \begin{cases} \frac{1}{2^{\frac{v}{2}}\Gamma(v/2)} x^{\frac{v-2}{2}} e^{-\frac{x}{2}} & for x > 0 \\ 0 & elsewhere \end{cases}$$

The mean and the variance of the chi-square distribution with v degrees of freedom are $v and 2v$, respectively, and its moment-generating function is given by $M_X(t) = (1-2t)^{-v/2}$

1.4.1. Result

If X has the standard normal distribution, then X^2 has the chi-square distribution with $v = 1$ degree of freedom.

1.4.2. Theorem

If $X_1, X_2, \dots, X_n$ are independent random variables having standard normal distributions, then $Y = \sum_{i=1}^{n} X_i^2$ has the chi-square distribution with $v = n$ degrees of freedom.

Proof:

Using the moment-generating function with $v = 1$ and by above result 1.3.1., we get

$M_{X_i^2}(t) = (1-2t)^{-\frac{1}{2}}$ and from the theorem " $M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t)$ then

$$M_Y(t) = \prod_{i=1}^{n} (1-2t)^{-\frac{1}{2}} = (1-2t)^{-\frac{n}{2}}$$

This moment-generating funciton is identified as that of the chi-square distribution with $v = n$ degrees of freedom.

1.4.3. Result

If $X_1, X_2, \ldots, X_n$ are independent random variables having chi-square distribution with $v_1, v_n, \ldots v_n$ degrees of freedom, then $Y = \sum_{i=1}^{n} X_i$ has the chi-square distribution with $v_1 + v_n + \cdots + v_n$ degrees of freedom.

1.4.4. Result

If X_1 and X_2 are independent random variables, X_1 has a chi-square distribution with v_1 degrees of freedom, and $X_1 + X_2$ has a chi-sqaure distribution with $v > v_1$ defrees of freedom, then X_2 has a chi-square distribution with $v - v_1$ degrees of freedom.

1.4.5. Theorem

If $\bar{X}$ and S^2 are the mean and the variance of a random sample of size n from a normal population with the mean μ and the standard deviation σ, then The random variable $\frac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with n-1 degrees of freedom.

Proof:

Consider the identity

$$\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} (X_i - \mu)^2 + n(\bar{X} - \mu)^2$$

Now, divided each term by $\sigma^2$and substitue $(n-1)S^2$ for $\sum_{i=1}^{n} (X_i - \bar{X})^2$,

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

We know from the theorem that the one on the left-hand side of the equation is a random variable having a chi-square distribution with n degrees of freedom. Also by theorems, the second term on the right-hand side of the equation is a random variable having a chi-square distributoin with 1 degree of freedom. Now, since $\bar{X}$ and S^2 are assumed to be independent that the two terms on the right-hand side of the equation are independent, and therefore $\frac{(n-1)S^2}{\sigma^2}$ is a random variable having a chi-squre distribution with n-1 degrees of freedom.

1.4.6. Example

Suppose that the thickness of a parat used in a semiconductor is its critical dimension and that the process of manufacturing these parts is condsidered to be under control if the true variation among the thickness of the parts is given by a standard deviation not greater than $\sigma = 0.60$ thousandth of an inch. To keep a check on the process, random samples of size n = 20 are taken periodically, and it is regarded to be “out of control” if the probability that S^2 will take on a value greater than or equal to the observed sample value is 0.01 or less (even though $\sigma = 0.60$). What can one conculude about the process if the standard deviation of such a periodic random sample is $s = 0.84$ thousandth of an inch?

Solution:

The process will be declared "out of control" if $\frac{(n-1)S^2}{\sigma^2}$ with $n = 20$ and $\sigma = 0.60$ exceeds $\chi^2_{0.01,19} = 36.191$. Since $\frac{(n-1)S^2}{\sigma^2} = \frac{19(0.84)^2}{(0.60)^2} = 37.24$ exceeds 36.191, the process is declared out of control. Here we assumed that the sample regarded as a random sample from a normal population.

1.5. The t Distribution

1.5.1. Theorem

If Y and Z are independent random variables. Y has a chi-square distribution with v degrees of freedom, and Z has the standard normal distribution, then the distribution of $T = \frac{Z}{\sqrt{Y/v}}$ is given by $f(t) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi v}\Gamma(\frac{v}{2})}.(1+\frac{t^2}{v})^{-\frac{v+1}{2}}$ $for -\infty < t < \infty$ and it is called the t distribution with v degrees of freedom.

Proof:

Since Y and Z are independent, their joint probability density is given by

$$f(y,z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \frac{1}{\Gamma(\frac{v}{2}) 2^{\frac{v}{2}}} y^{\frac{v}{2}-1} e^{-\frac{y}{2}}$$

for $y > 0$ and $-\infty < z < \infty$, and $f(y,z) = 0$ elsewhere. Then, to use the change-of-variable technique, we solve $t = \frac{z}{\sqrt{y/v}}$ for z, getting $z = t\sqrt{y/v}$ and hence $\frac{\partial z}{\partial t} = \sqrt{y/v}$. Thus, the joint density of Y and T is given by

$$g(y,t) = \begin{cases} \frac{1}{\sqrt{2\pi v}\Gamma(\frac{v}{2}) 2^{\frac{v}{2}}} y^{\frac{v-1}{2}} e^{-\frac{y}{2}(1+\frac{t^2}{v})} & for\ y > 0\ and -\infty < t < \infty \\ 0 & elsewhere \end{cases}$$

and, integrating out y with the aid of the substitution $w = \frac{y}{2}(1+\frac{t^2}{v})$, we get

$$f(t) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi v}\Gamma(\frac{v}{2})}.(1+\frac{t^2}{v})^{-\frac{v+1}{2}} \quad for -\infty < t < \infty$$

1.5.2. Theorem

If $\bar{X}$ and S^2 are the mean and the variance of a random sample of size n from a normal population with the mean μ and the variance σ^2 then $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has the t distribution with n-1 degrees of freedom.

Proof:

From theorems 1.2.6. & 1.4.5. we get, the random variables $Y = \frac{(n-1)S^2}{\sigma^2}$ and $T = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ have, respectively, a chi-square distribution with n-1 degrees of freedom and the standard normal distribution. Since they are also independent, substitution into the formula for T of the above theorem we have $T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$.

1.5.3. Example

In 16 one-hour test runs, the gasoline consumption of an engine averaged 16.4 gallons with a standard deviation of 2.1 gallons. Test the claim that the average gasoline consumption of this engine is 12.0 gallons per hour.

Solution:

Substituting $n = 16, \mu = 12. \bar{x} = 16.4\ and\ s = 2.1$ into the formula for t in the above theorem

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{16.4 - 12}{2.1/\sqrt{16}} = 8.38$$

Since from statistical table we have, for $v = 15$ the probability of getting of T greater than 2.947 is 0.005, the probability of getting a value greater than 8 must be negligible. Thus, it would seen reasonable to conculud that the ture average hourly gasoline consumption of the engine exceeds 12 gallons.

1.6. The F Distribution

1.6.1. Theorem

If U and V are independent random variables having chi-squre distributions with v_1 and v_2 degrees of freedom, then $F = \frac{U/v_1}{V/v_2}$ is a random variable having an F distribution, that is, a random variable whose probability density is given by

$$g(f) = \frac{\Gamma(\frac{v_1+v_2}{2})}{\Gamma(\frac{v_1}{2})\Gamma(\frac{v_2}{2})}\left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} f^{\frac{v_1}{2}-1}\left(1+\frac{v_1}{v_2}f\right)^{-\frac{1}{2}(v_1+v_2)}$$ for $f > 0$ and $g(f) = 0$ elsewhere.

Proof:

By virute of independence, the joint density of U and V is given by

$$f(u,v) = \frac{1}{2^{v_1/2}\Gamma\left(\frac{v_1}{2}\right)} u^{\frac{v_1}{2}-1} e^{-\frac{u}{2}} \frac{1}{2^{v_2/2}\Gamma\left(\frac{v_2}{2}\right)} v^{\frac{v_2}{2}-1} e^{-\frac{v}{2}}$$

$f(u,v) = \frac{1}{2^{(v_1+v_2)/2}\Gamma(\frac{v_1}{2})\Gamma(\frac{v_2}{2})} u^{\frac{v_1}{2}-1} v^{\frac{v_2}{2}-1} e^{-\frac{(u+v)}{2}}$ for $u > 0$ and $v > 0$, and $f(u,v) = 0$ elsewhere. Then, to use the change-of-variable, we solve $f = \frac{u/v_1}{v/v_2}$ for u, getting $u = \frac{v_1}{v_2}.vf$ and hence $\frac{\partial u}{\partial f} = \frac{v_1}{v_2}v$. Thus, the joint density of F and V is given by

$g(f,v) = \frac{\left(\frac{v_1}{v_2}\right)^{v_1/2}}{2^{(v_1+v_2)/2}\Gamma(\frac{v_1}{2})\Gamma(\frac{v_2}{2})} f^{\frac{v_1}{2}-1} v^{\frac{v_1+v_2}{2}-1} e^{-\frac{v}{2}\left(\frac{v_1 f}{v_2}+1\right)}$ for $f > 0$ and $v > 0$, and $g(f,v) = 0$ elsewhere.

Now, integrating out v by making the substitutiion $w = \frac{v}{2}(\frac{v_1 f}{v_2} + 1)$, we finally get

$$g(f) = \frac{\Gamma(\frac{v_1+v_2}{2})}{\Gamma(\frac{v_1}{2})\Gamma(\frac{v_2}{2})}\left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} f^{\frac{v_1}{2}-1}\left(1+\frac{v_1}{v_2}f\right)^{-\frac{1}{2}(v_1+v_2)}$$ for $f > 0$ and $g(f) = 0$ elsewhere.

1.6.2. Result

If S_1^2 and S_2^2 are the variances of independent random samples of sizes n_1 and n_2 populations from normal populations with the variances σ_1^2 and σ_2^2 then $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$ is a random variable having an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.The F distribution is also known as the variance-ratio distribution.

Let Us Sum Up

In this unit, we explained the concept of sampling distribution of the mean, the chi-square distribution, t distribution and F distribution with illustration.

Check Your Progress

1. The stand error of $\bar{X}$ is _________.

2. The standard deviation computed from the observations of sampling distribution of a statistic is_________.

3. The standard error of $\bar{X}$ varies _______with standard deviation and ______with sample size.

Glossaries

Population: It means the whole of the information which comes under the purview of statistical investigation.

Parameter: Any statistical measure computed from population data.

Statistic: Any statistical measure computed from sample data.

Population distribution: The distribution of the numbers constituting a population.

Random sample: It is a subset of individuals chosen from a larger set in which a subset of individuals is chosen randomly, all with the same probability.

Sample mean: It is an average value found in a sample.

Suggested Readings

1. Freund. J.E.," Mathematical Statistics", Prentice Hall of India, Fifth Edition, 2001.

2. Gupta. S.C. and Kapoor. V. K., "Fundamentals of Mathematical Statistics", Sultan Chand & Sons, Eleventh Edition, 2003.

3. Devore. J. L. "Probability and Statistics for Engineers", Brooks/Cole (Cengage Learning), First India Reprint, 2008.

Answers to Check Your Progress

1. $\frac{\sigma}{\sqrt{n}}$.

2. Standard error of the statistic.

3. Directly, inversely.

Unit – 2

Point Estimation

Objectives

After Studying this Unit, the student will be able to

- Explain the unbiased estimators, efficiency, consistency and sufficiency.
- Demonstrate the concept of the method of moments and maximum likelihood.
- Illustrate the numerical problems in point estimation.

Overview

In this unit, we will study the concept of Point estimation. We will mainly focus on unbiased estimators, efficiency, consistency, sufficiency, the method of moments and the method of maximum likelihood.

2.1. Introduction

Problems of statistical inference are divided into problems of estimation and tests of hypotheses, though actually they are all decision problems and, hence, could be handled by the unified approach. The main difference between the two kinds of problems is that in problems of estimation we must determine the value of a parameter or the values of several parameters from a possible continuum of alternatives, whereas in tests of hypotheses we must decide whether to accept or reject a specific value or a set of specific values of a parameter or those of several parameters.

2.1.1 Point Estimation.

Using the value of a sample statisitc to estimate the value of a population parameters is called point estimation. We refer to the value of the statistic as a point estimate.

2.2. Unbiased Estimators

Perfect decision functions do not exist, and in connection with problems of estimation this means that there are no perfect estimators that always give the right answer. Thus, it would seem reasonable that an estimator should do so at least on the average; that is, it's expected value should equal the parameter that is supposed to estimate. If this is the case, the estimator is said to be unbiased; otherwise it is said to be biased.

2.2.1. Unbiased Estimator

A statistic $\hat{\Theta}$ is an unbaised estimataor of the parameter θ of a given distribution if and only if $E(\hat{\Theta}) = \theta$ for all possible vlaues of θ.

2.2.2. Example

Show that unless $\theta = \frac{1}{2}$ the minimax estimator of the binomial parameter θ is biased.

Solution:

Since $E(X) = n\theta$

$$E\left(\frac{X+\frac{1}{2}\sqrt{n}}{n+\sqrt{n}}\right)=\frac{E\left(X+\frac{1}{2}\sqrt{n}\right)}{n+\sqrt{n}}=\frac{n\theta+\frac{1}{2}\sqrt{n}}{n+\sqrt{n}}$$

This quantity does not equal to θ unless $\theta = \frac{1}{2}$

2.2.3. Example

If X has the binomial distribution with the parameters n and θ, show that the sample proportion, $\frac{X}{n}$, is an unbiased estimator of θ.

Solution:

$E(X) = n\theta$

$$E\left(\frac{X}{n}\right) = \frac{1}{n}.E(X) = \frac{1}{n}.n\theta = \theta$$

Hence $\frac{X}{n}$ is an unbiased estimator of θ.

2.2.4. Example

If $X_1, X_2, \ldots, X_n$ constitute a random sample from the population given by

$$f(x) = \begin{cases} e^{-(x-\delta)} & for\ x>\delta \\ 0, & otherwise \end{cases}$$

Show that $\bar{X}$ is a biased estimator of δ.

Solution:

Since the mean of the population is $\mu = \int_{\delta}^{\infty} x\,.\,e^{-(x-\delta)}dx = 1 + \delta$

From the theorem "If $\bar{X}$ is the mean of a random sample of size n taken without replacement from a finite population of size N with the mean μ and the variance σ^2, then $E(\bar{X}) = \mu$ and $var(\bar{X}) = \frac{\sigma^2}{n}.\frac{N-n}{N-1}$" that $E(\bar{X}) = 1 + \delta \neq \delta$ and hence that $\bar{X}$ is a biased estimator of δ.

2.2.5. Asymptotically unbiased Estimator

Letting $b_n(\theta) = E(\hat{\Theta}) - \theta$ express the bias of an estimator $\hat{\Theta}$ based on a random sample of size n from a given distribution, we say that $\hat{\Theta}$ is an asymptotically unbiased estimator of θ if an only if $\lim_{n\to\infty} b_n(\theta) = 0$.

2.2.6. Example

If $X_1, X_2, \ldots, X_n$ constitute a random sample from a uniform population with $\alpha = 0$. Show that the largest sample value (that is, the nth order statistic, Y_n) is a biased estimator of the parameter β. Also, modify this estimator of β to make it unbiased.

Solution:

Substituting into the formula for

$$g_n(y_n) = \begin{cases} \frac{n}{\theta} \cdot e^{-\frac{y_n}{\theta}} [1 - e^{-\frac{y_n}{\theta}}]^n & for\ y_n > 0 \\ 0, & otherwise \end{cases}$$

We find that the sampling distribution of Y_n is given by

$$g_n(y_n) = n.\frac{1}{\beta}\left(\int_0^{y_n} \frac{1}{\beta}\, dx\right)^{n-1} = \frac{n}{\beta^n} \cdot y_n^{n-1}$$

for $0 < y_n < \beta$ and $g_n(y_n) = 0$ elsewhere, and hence that

$$E(Y_n) = \frac{n}{\beta^n} \int_0^{\beta} y_n^n\, dy_n = \frac{n}{n+1} . \beta$$

Thus, $E(Y_n) \neq \beta$ and the nth order statistic is a biased estimator of the parameter β.

Since $E\left(\frac{n}{n+1} . Y_n\right) = \frac{n+1}{n} . \frac{n}{n+1} . \beta = \beta$

$\frac{n+1}{n}$ times the largest sample value is an unbiased estimator of the parameter β.

2.2.7. Theorem

If S^2 is the variance of a random sample from an infinite population with the finite variance σ^2, then $E(S^2) = \sigma^2$.

Proof:

By definition of sample mean and sample variance

$$E(S^2) = E\left[\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2\right]$$

$$E(S^2) = \frac{1}{n-1} E\left[\sum_{i=1}^{n} \{(X_i - \mu) - (\bar{X} - \mu)\}^2\right]$$

$$E(S^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n} E\{(X_i - \mu)^2\} - n.E\{(\bar{X} - \mu)^2\}\right]$$

Then, since $E\{(X_i - \mu)^2\} = \sigma^2$ and $E\{(\bar{X} - \mu)^2\} = \frac{\sigma^2}{n}$ we get

$$\boldsymbol{E(S^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n} \sigma^2 - n.\frac{\sigma^2}{n}\right] = \sigma^2}$$

2.3. Efficiency

2.3.1. Minimum Variance unbiased Estimator

The estimator for the parameter θ of a given distribution that has the samllest variance of all unbiased estimators for θ is called the minimum variance unbiased estimator, or the best unbiased estimator for θ.

2.3.2. Result

If $\hat{\Theta}$ is an unbiased estimator of θ and $var(\hat{\Theta}) = \frac{1}{n.E[(\frac{\partial \ln f(X)}{\partial \theta})^2]}$ then $\hat{\Theta}$ is a minimum variance unbiased estimator of θ.

2.3.3. Example

Show that $\bar{X}$ is a minimum variance unbiased estimator of the mean μ of a normal population.

Solution:
Since $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ $for - \infty < x < \infty$

$$\ln f(x) = -\ln \sigma\sqrt{2\pi} - \frac{1}{2}(\frac{x-\mu}{\sigma})^2$$

$\frac{\partial \ln f(x)}{\partial \mu} = \frac{1}{\sigma}(\frac{x-\mu}{\sigma})$ and hence

$$E\,[(\frac{\partial \ln f(X)}{\partial \mu})^2] = \frac{1}{\sigma^2}.E\,[(\frac{x-\mu}{\sigma})^2] = \frac{1}{\sigma^2}.1 = \frac{1}{\sigma^2}$$

Thus, $\frac{1}{n.E[(\frac{\partial \ln f(X)}{\partial \mu})^2]} = \frac{1}{n.\frac{1}{\sigma^2}} = \frac{\sigma^2}{n}$

and since $\bar{X}$ is unbiased and $Var(\bar{X}) = \frac{\sigma^2}{n}$, $\bar{X}$ is a minimum variance unbiased estimator of μ.

2.3.4. Result

Unbiased estimators of one and the same parameter are usually compared in terms of the size of their variances. If $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are wo unbiased esitmators of the parameter θ of a given population and the vairance $\hat{\Theta}_1$ is less than the variance $\hat{\Theta}_1$ is relatively more efficient than $\hat{\Theta}_2$. Also $\frac{Var(\hat{\Theta}_1)}{Var(\hat{\Theta}_2)}$ as a measure of the efficient of $\hat{\Theta}_2$ relative to $\hat{\Theta}_1$.

2.3.5. Example

If $X_1, X_2, \ldots, X_n$ constitute a random sample from a uniform population with $\alpha = 0$, then $\frac{n+1}{n}, Y_n$ is an unbiased estimator of β. (a) Show that $2\bar{X}$ is also an unbiased estimator of β. (b) Compare the efficient of these two estimators of β.

Soluiton:

(a) Since the mean of the population is $\mu = \frac{\beta}{2}$, accoring to the theorem, “The mean and the variance of the uniform distribution are given by $\mu = \frac{\alpha+\beta}{2}$ and $\sigma^2 = \frac{1}{12}(\beta - \alpha)^2$” and also from the theorem “If $X_1, X_2, \dots, X_n$ constitute a random sample from an infinite population with the mean μ and the variance σ^2, then $E(\bar{X}) = \mu$ and $var(\bar{X}) = \frac{\sigma^2}{n}$” that $E(\bar{X}) = \frac{\beta}{2}$ and hence that $E(2\bar{X}) = \beta$. Thus $2\bar{X}$ is an unbiased estimator of β.

(b) Using the sampling distribution of Y_n and the expression for $E(Y_n) = \frac{n}{\beta^n}\int_0^\beta y_n^n \, dy_n = \frac{n}{n+1}.\beta$

$E(Y_n^2) = \frac{n}{\beta^n}\int_0^\beta y_n^{n+1} \, dy_n = \frac{n}{n+2}\beta^2$ and

$$Var(Y_n) = \frac{n}{n+2}.\beta^2 - \left(\frac{n}{n+1}.\beta\right)^2$$

$$Var\left(\frac{n+1}{n}Y_n\right) = \frac{\beta^2}{n(n+2)}$$

Since the variance of the population is $\sigma^2 = \frac{\beta^2}{12}$ according to the theorem we have $Var(\bar{X}) = \frac{\beta^2}{12n}$ and hence $Var(2\bar{X}) = 4.var(\bar{X}) = \frac{\beta^2}{3n}$

Therefore, the efficiency of $2\bar{X}$ relative to $\frac{n+1}{n}.Y_n$ is given by

$\frac{Var(\frac{n+1}{n}.Y_n)}{Var(2\bar{X})} = \frac{\left(\frac{\beta^2}{n(n+2)}\right)}{\left(\frac{\beta^2}{3n}\right)} = \frac{3}{n+2}$ and for $n > 1$ the estimator based on the nth order statistic is much more efficient than the other one. For $n = 10$, foro exmpale, the relative efficiency is only 25 percent, and for $n = 25$ it is only 11 percent.

2.3.6. Example

When the mean of a normal population is estimated on the basis of a random sample of size $2n + 1$, what is the efficiency of the median relataive to the mean?

Solution:

From the theorem we know that $\bar{X}$ is unbaised and that $Var(\bar{X})$ is unbiased and that $Var(\bar{X}) = \frac{\sigma^2}{2n+1}$

For $\tilde{X}$, it is unbiased by virtue of the symmetry of the normal distribution about its mean, and for large sample $Var(\tilde{X}) = \frac{\pi\sigma^2}{4n}$

Thus for larage samples, the efficiency of the median relative to the mean is approximately

$\frac{Var(\bar{X})}{Var(\tilde{X})} = \frac{\left(\frac{\sigma^2}{2n+1}\right)}{\left(\frac{\pi\sigma^2}{4n}\right)} = \frac{4n}{\pi(2n+1)}$ and the asymptotic efficiency of the median with respect to the mean is $\lim_{n\to\infty} \frac{4n}{\pi(2n+1)} = \frac{2}{\pi}$ or about 64 percent.

2.4. Consistency

2.4.1. Consistent Estimator

The Statistic $\hat{\Theta}$ is a consistent estimator of the parameter θ of a given distribution if and only if for each $c > 0$ $\lim_{n\to\infty} P(|\hat{\Theta} - \theta| < c) = 1$

2.4.2. Result

If $\hat{\Theta}$ is an unbiased estimator of the parameter θ and $var(\hat{\Theta}) \to 0$ as $n \to \infty$, then $\hat{\Theta}$ is a consistent estimator of θ.

2.4.3. Example

Show that for a random sample from a normal population, the sample variance S^2 is a consistent estimator of σ^2.

Solution:

Since S^2 is an unbiased estimator of σ^2 by theorem.

To show that $var(S^2) \to 0$ as $n \to \infty$. From the theorem "the random variable $\frac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with $n-1$ degress of freedom"

We find that for a random sample from a normal population $var(S^2) = \frac{2\sigma^4}{n-1}$

$Var(S^2) \to 0$ as $n \to \infty$ and we have S^2 is a consistent estimator of variance of a normal population.

2.4.4. Example

If $X_1, X_2, \dots, X_n$ constitute a random sample from the population given by

$$f(x) = \begin{cases} e^{-(x-\delta)} & for\ x > \delta \\ 0 & elsewhere \end{cases}$$

Show that the smallest sample value (that is, the first order statisitic Y_1) is a consistent estimator of the parameter δ.

Solution:

Substituting into the formula for $g_1(y_1)$, we find that the sampling distribution of Y_1 is given by

$g_1(y_1) = n.e^{-(y_1-\delta)}.\left[\int_{y_1}^{\infty} e^{-(x-\delta)}\,dx\right]^{n-1} = n.e^{-n(y_1-\delta)}$ for $y_1 > \delta$ and $g_1(y_1) = 0$ elsewhere. Based on this result, we have $E(Y_1) = \delta + \frac{1}{n}$ and hence Y_1 is an asymptotically unbiased estimator of δ.

$$P(|Y_1 - \delta| < c) = P(\delta < Y_1 < \delta + c) = \int_{\delta}^{\delta+c} n.e^{-n(y_1-\delta)}dy_1 = 1 - e^{-nc}$$

Since $\lim_{n\to\infty}(1 - e^{-nc}) = 1$, from Definition we have Y_1 is a consistent estimator of δ.

2.5. Sufficiency

2.5.1. Sufficient Estimator

The statistic $\hat{\Theta}$ is a sufficient estimator of the parameter θ of a given distribution if and only if for each value of $\hat{\Theta}$ the conditional probability distribution or density of the random sample $X_1, X_2, \dots, X_n$, given $\hat{\Theta} = \theta$, is independent of θ.

2.5.2. Example

If $X_1,\ X_2,\ \dots\ ,\ X_n$ constitute a random sample of size n from a Bernoulli population, Show that $\hat{\Theta} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ is a sufficient estimator of the parameter θ.

Solution:

By the definition "BERNOULLI DISTRIBUTIONS, A random variable X has a Bernoulli distributon and it is referred to as a Bernoulli random variable if and only if its probability distribution is given by $f(x;\ \theta) = \theta^x(1-\theta)^{1-x}$ for $x = 0, 1$".

$f(x_i;\ \theta) = \theta^{x_i}(1-\theta)^{1-x_i}$ for $x_i = 0, 1$

So that $f(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$

$$f(x_1, x_2, \dots, x_n) = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}$$

$$f(x_1, x_2, \dots, x_n) = \theta^{x}(1-\theta)^{n-x}$$

$$f(x_1, x_2, \dots, x_n) = \theta^{n\delta}(1-\theta)^{n-n\delta}$$

for $x_i = 0\ or\ 1$ and $i = 1, 2, \dots, n$. Also, since $X = X_1 + X_2 + \cdots + X_n$ is a binomial random varaible with the parameters θ and n, its distribution is given by

$b(x; n, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ and the transformation-of-variable technique we have

$g(\hat{\theta}) = \binom{n}{n\hat{\theta}}\theta^{n\hat{\theta}}(1-\theta)^{n-n\hat{\theta}}$ for $\hat{\theta} = 0, \frac{1}{n}, \dots, 1$

We know that

$$f(x_1, x_2, \dots, x_n \mid \hat{\theta}) = \frac{f(x_1, x_2, \dots, x_n, \hat{\theta})}{g(\hat{\theta})}$$

$$f(x_1, x_2, \dots, x_n \mid \hat{\theta}) = \frac{f(x_1, x_2, \dots, x_n)}{g(\hat{\theta})}$$

$$f(x_1, x_2, \dots, x_n \mid \hat{\theta}) = \frac{\theta^{n\hat{\theta}}(1-\theta)^{n-n\hat{\theta}}}{\binom{n}{n\hat{\theta}}}$$

$$f(x_1, x_2, \dots, x_n \mid \hat{\theta}) = \frac{1}{\binom{n}{n\hat{\theta}}} = \frac{1}{\binom{n}{x}} = \frac{1}{\binom{n}{x_1+x_2+\cdots+x_n}}$$

for $x_i = 0$ or 1 and $i = 1, 2, 3, \dots, n$. This does not depend on θ and therefore, $\hat{\Theta} = \frac{X}{n}$ is a sufficient estimator of θ.

2.5.3. Example

Show that $Y = \frac{1}{6}(X_1 + 2X_2 + 3X_3)$ is not a sufficient estimator of the Bernoulli parameter θ.

Solution:

Since $f(x_1, x_2, x_3 \mid y) = \frac{f(x_1, x_2, x_3, y)}{g(y)}$ is not independent of θ for some values of X_1, X_2 and X_3.

Let us consider the case where $x_1 = 1, x_2 = 1$, and $x_3 = 0$.

Thus, $y = \frac{1}{6}(1 + 2.1 + 3.0) = \frac{1}{2}$ and

$$f(1, 1, 0 \mid Y = \frac{1}{2}) = \frac{P(X_1 = 1, X_2 = 1, X_3 = 0, Y = \frac{1}{2})}{P(Y = \frac{1}{2})}$$

$$f(1, 1, 0 \mid Y = \frac{1}{2}) = \frac{f(1, 1, 0)}{f(1, 1, 0) + f(0, 0, 1)}$$

Where $f(x_1, x_2, x_3) = \theta^{x_1+x_2+x_3}(1 - \theta)^{3-(x_1+x_2+x_3)}$

for $x_1 = 0$ or 1 and $i = 1, 2, 3$. Since $f(1, 1, 0) = \theta^2(1 - \theta)$ and $f(0,0,1) = \theta(1 - \theta)^2$

$$f(1, 1, 0 \mid Y = \frac{1}{2}) = \frac{\theta^2(1 - \theta)}{\theta^2(1 - \theta) + \theta(1 - \theta)^2} = \theta$$

This conditional probability depends on θ.

Thus, $Y = \frac{1}{6}(X_1 + 2X_2 + 3X_3)$ is not a sufficient estimator of the parameter θ of a Bernoulli population.

2.5.4. Result: (Factorization theorem)

The statistic $\hat{\Theta}$ is a sufficient etimator of the parameter θ if and only if the joint probability dsitrbution or density of the randam sample can be factored so that $f(x_1, x_2, \dots, x_n; \theta) = g(\hat{\theta}\theta).h(x_1, x_2, \dots, x_n)$, where $g(\hat{\theta}\theta)$ depends only on $\hat{\theta}$ and θ, and $h(x_1, x_2, \dots, x_n)$ does not depend on θ.

2.5.5. Example

Show that $\bar{X}$ is a sufficient estimator of the mean μ of a normal population iwht the known variance σ^2.

Solution:

We know that

$$f(x_1, x_2, \dots, x_n; \mu) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n . e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

and that

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}[(x_i-\bar{X}) - (\mu-\bar{X})]^2$$

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{X})^2 + \sum_{i=1}^{n}(\bar{x}-\mu)^2$$

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{X})^2 + n(\bar{x}-\mu)^2$$

We get

$$f(x_1, x_2, \dots, x_n; \mu) = \left\{\frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2}\right\} \times \left\{\frac{1}{\sqrt{n}}\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n-1} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2}\right\}$$

Where the first factor on the right-hand side depends only on the estimate $\bar{x}$ and the population mean μ, and the second factor does not involve μ. According to the theorem, $\bar{X}$ is a sufficient estimator of the mean μ of a normal population with the known variance σ^2.

2.6. The Method of Moments

2.6.1. Sampe Moments

The k^{th} sample moment of a set of observations $x_1, x_2, \dots, x_n$ is the mean of their k^{th} power and it is denoted by m'_k. Symbolically,

$$m'_k = \frac{\sum_{i=1}^{n} x_i^k}{n}$$

Thus, if a population has r parameters, the method of meoments consists of solving the system of equations $m'_k = \mu'_k, k = 1, 2, \dots, r$ for the r parameters.

2.6.2. Example

Given a random sample of size n from a unifrom population with $\beta = 1$, use the method of moments to obtain a formula for estimating the parameter α.

Solution:

The equation that we shall to solve is $m'_1 = \mu'_1$

Where $m'_1 = \bar{x}$ and $\mu'_1 = \frac{\alpha+\beta}{2} = \frac{\alpha+1}{2}$.

Thus $\bar{x} = \frac{\alpha+1}{2}$

$\hat{\alpha} = 2\bar{x} - 1$

2.6.3. Example

Given a random sample of size n from a gamma population, we use the method of meoments to obtain formulas for estimating the parameters α and β.

Solution:

The system of equations that we shall have to solve is $m'_1 = \mu'_1$ and $m'_2 = \mu'_2$

Where $\mu'_1 = \alpha\beta$ and $\mu'_2 = \alpha(\alpha+1)\beta^2$.

Thus, $m'_1 = \alpha\beta$ and $m'_2 = \alpha(\alpha+1)\beta^2$

Solving for α and β, we get the following formulas for estimating the two parameters of the gamma distribution:

$\hat{\alpha} = \frac{(m'_1)^2}{m'_2-(m'_1)^2}$ and $\hat{\beta} = \frac{m'_2-(m'_1)^2}{m'_1}$

Since $m'_1 = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$ and $m'_2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$

$\hat{\alpha} = \frac{n\,\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ and $\hat{\beta} = \frac{\sum_{i=1}(x_i-\bar{x})^2}{n\,\bar{x}}$ in terms of the original observations.

2.7. The Method of Maximum Likelihood

2.7.1. Maximum Likelihood Estimator

If $x_1, x_2, \ldots, x_n$ are the values of a random sample from a population with the parameter θ, the likelihood functin of the sample is given by $L(\theta) = f(x_1, x_2, \ldots, x_n; \theta)$ for values of θ within a given domain. Here $f(x_1, x_2, \ldots, x_n; \theta)$ is the value of the joint probability distribution or the joint probability density of the random variables $X_1, X_2, \ldots, X_n$ at $X_1 = x_1$, $X_2 = x_2, \ldots, X_n = x_n$. We refer to the value of θ that maximizes $L(\theta)$ as the maximum likelihood estimator of θ.

2.7.2. Example

Given x "successes" in n trials, find the maximum likelihood estimates of the parameter θ of the corresponding binomial distribution.

Solution:

To find the value of θ that maximizes $L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$

The value of θ that maximizes $L(\theta)$ will also maximize

$$ln\, L(\theta) = ln \binom{n}{x} + x.\, ln\, \theta + (n - x).\, ln\,(1 - \theta)$$

Thus, we get

$\frac{d[ln\, L(\theta)]}{d\theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta}$ and, equating this derivative to 0 and solving for θ, we get the likelihiood function has a maximum at $\theta = \frac{x}{n}$.

This is the maximum likelihood estimate of the binomial parameter θ, we refere to $\hat{\Theta} = \frac{X}{n}$ as the corresponding maximum likelihood estimator.

2.7.3. Example

If $x_1, x_2, \dots, x_n$ are the values of a random sample from a exponential population, find the maximum likelihood estimator of its parameter θ.

Solution:

Since the likelihood function is given by $L(\theta) = f(x_1, x_2, \dots, x_n; \theta)$

$$L(\theta) = \prod_{i=1} f(x_i;\ \theta)$$

$$L(\theta) = \left(\frac{1}{\theta}\right)^n e^{-\frac{1}{\theta}\left(\sum_{i=1}^{n} x_i\right)}$$

Differentiation of $ln\, L(\theta)$ with respect to θ, we have

$$\frac{d[ln\, L(\theta)]}{d\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i$$

Equating this derivative to zero and solving for θ, we get the maximum likelihood estimate $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$. Hence, the maximum likelihood estimator is $\hat{\Theta} = \bar{X}$

2.7.4. Example

If $x_1, x_2, \dots, x_n$ are the values of a random sample of size n form a uniform population with $\alpha = 0$, find the maximum likelihood estimator of β.

Solution: The Likelihood funciton is given by

$$L(\beta) = \prod_{i=1}^{n} f(x_i; \beta) = \left(\frac{1}{\beta}\right)^n$$

for β greater than or equal to the largest of the $x's$ and 0 otherwaise. Since the value of this likelihood function increases as β decreases, we must male β as samll as possible, and it follows that the maximum likelihood estimator of β is Y_n, the n^{th} order statistic.

2.7.5. Example

If $X_1, X, \ldots, X_n$ constitute a random sample of size n from a normal population with the mean μ and the variance σ^2, find joint maximum likelihood estimates of these two parameters.

Solution:

Since the likelihood function is given by

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} n(x_i; \mu, \sigma)$$

$$L(\mu, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n . e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

Partial differentiation of $ln\ L(\mu, \sigma^2)$ with respect to μ and σ^2, we have

$$\frac{d[ln\ L(\mu, \sigma^2)]}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$

and

$$\frac{d[ln\ L(\mu, \sigma^2)]}{d\theta^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}.\sum_{i=1}^{n}(x_i - \mu)^2$$

Equating the first of these two partial derivatives to zero and solving for μ, we get

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

and equating the second of these partial derivatives to zero and solving for σ^2 after substituting $\mu = \bar{x}$, we get

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

Let Us Sum Up

In this unit, we studied the concept of unbiased estimators, efficiency, consistency, sufficiency, the method of moments and the method of maximum likelihood.

Check Your Progress

1. A good estimator must possess________.

2. A statistic $t = t_n$ based on the sample size n is said to be consistent estimator of the parameter if________.

3. Method of moment estimators are usually less efficient than__________.

Glossaries

Point estimate: The estimate of a population parameter given by a single number.

Unbiasedness: The mean value of the sampling distribution of the statistic t is equal to the parameter of the population.

Efficiency: An estimator with less variability is said to be more efficient and consequently more reliable than the other.

Sufficiency: It contains all the information in the sample regarding the parameter.

Suggested Readings

1. Freund. J.E.," Mathematical Statistics", Prentice Hall of India, Fifth Edition, 2001.

2. Gupta. S.C. and Kapoor. V. K., "Fundamentals of Mathematical Statistics", Sultan Chand & Sons, Eleventh Edition, 2003.

3. Devore. J. L. "Probability and Statistics for Engineers", Brooks/Cole (Cengage Learning), First India Reprint, 2008.

Answers to Check Your Progress

1. Unbiasedness, Consistency, Efficiency, Sufficiency.

2. $t_n \rightarrow \theta \; as \; n \rightarrow \infty$

3. Method of Maximum likelihood.

Unit – 3

Interval Estimation

Structure

Objectives

After Studying this Unit, the student will be able to

- Distinguish between the estimation means and differences between means.
- Examine the difference between the estimation of proportions and differences between proportions.
- Explain the estimation of variances and ratio of two variances.

Overview

In this unit, we will study the concept of Interval estimation. We will mainly focus on the Estimation of Means, differences between means, proportions, and differences between proportions, variances and ratio of two variances.

3.1. Introduction

Although point estimation is a common way in which estimates are expressed. For instance, it does not tell us on how much informationthe estimate is based, nor does it tell us anything about the possible size ofthe error. Thus, we might have to supplement a point estimate $\hat{\theta}$ of θ with the size ofthe sample and the value of $Var(\hat{\Theta})$ or with some other information about the samplingdistribution of $\hat{\Theta}$ This will enable us to appraise the possible size of the error. Alternatively, we might use interval estimation.

An interval estimate of θ is an interval of the form $\hat{\theta}_1 < \theta < \hat{\theta}_2$, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are values of appropriate random variables $\hat{\Theta}_1$ and $\hat{\Theta}_2$.

3.1.1. Confidence Interval

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are values of the random variables $\hat{\Theta}_1$ and $\hat{\Theta}_2$ such that $P(\hat{\Theta}_1 < \theta < \hat{\Theta}_2) = 1 - \alpha$ for some specified probability $1 - \alpha$, we refer to the interval $\hat{\theta}_1 < \theta < \hat{\theta}_2$ as a $(1 - \alpha)100\%$ confidence interval for θ. The Probability $1 - \alpha$ is called the degree of confidence, and the endpoints of the interval are called the lower and upper confidence limits.
When $\alpha = 0.05$, the degree of confidence is 0.95 and we get a 95% confidence intrval.

3.2.The Estimation of Means

Suppose that the mean of a random sample is to be used to estimate the mean of a normal population with the known variance σ^2. By the theorem "If χ is the mean of a random sample of size n from a normal populaton with the mean μ and the variance σ^2, its sampling distribution of $\bar{X}$ for random samples of size n from a normal population with the mean μ and the variance σ^2 is a normal distribution with $\mu_{\bar{x}} = \mu$ and $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$, Then $p(|Z| < z_{\alpha/2}) = 1 - \alpha$, where $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ and $z_{\alpha/2}$ is such that the integral of the standard normal density from $z_{\alpha/2}$ $to\ \infty$ equals $\alpha/2$. Therefore, $P\left(|\bar{X} - \mu| < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$

3.2.1. Result

If $\bar{X}$ the mean of a random sample of size n from a normal population with the known variance σ^2, is to be used as an estimator of the mean of the population, the probability is $1-\alpha$ that the error will be less than $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$.

3.2.2. Example

A team of efficiency experts intends to use the mean of a random sample of size $n = 150$ to estimate the average mechanical aptitude of assembly-line workers in a large industry (as measured by a certain standardized test). If, based on experience, the efficiency experts can assume that $\sigma = 6.2$ for such data, what can they assert with probability 0.99 about the maximum error of their estimate?

Solution:

Substituting $n = 150, \sigma = 6.2,$ and $z_{0.005} = 2.575$ into the expression for the maximum error, we get

$$2.575 . \frac{6.2}{\sqrt{150}} = 1.30$$

Thus, the efficiency experts can assert with probability 0.99 that their error will be less than 1.30.

3.2.3. Result

To construct a confidence interval formula for estimating the mean of a normal population with the known variance σ^2, then $P\left(|\bar{X} - \mu| < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$, we write

$$P\left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

If $\bar{x}$ is the value of the mean of a random sample of size n from a normal population with the known variance σ^2, then $\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ is $(1-\alpha)100\%$ confidence interval for the mean of the population.

3.2.4. Example

If a random sample of size $n = 20$ from a normal population with the variance $\sigma^2 = 225$ has the mean $\bar{x} = 64.3$ construct a 95% confidence interval for population mean μ.

Solution:

Substituting $n = 20, \bar{x} = 20, \sigma = 15$ and $z_{0.025} = 1.96$ into the confidence-interval formula of above theorem, we get

$$64.3 - 1.96 \ \frac{15}{\sqrt{20}} < \mu < 64.3 + 1.96 \ \frac{15}{\sqrt{20}}$$

$$57.7 < \mu < 70.9$$

3.2.5. Remark

Confidence-interval formulas are not unique. This may be by changing the confidence-interval formula of the above result, we have

$$\bar{x} - z_{2\alpha/3}.\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/3}.\frac{\sigma}{\sqrt{n}}$$

or to the one-sided $(1-\alpha)100\%$ confidence-interval formula $\mu < \bar{x} + z_{\alpha}.\frac{\sigma}{\sqrt{n}}$

3.2.6. Example

An industrial designer wants to determine the average amount of time it takes an adult t assemble an "easy-to-assemble" toy. Use the following data (in minutes), a random sample, to construct a 95% confidence interval for the mean of the population sampled: 17, 13, 18, 19, 17, 21, 29, 22, 16, 28, 21, 15, 26, 23, 24, 20, 8, 17, 17, 21, 32, 18, 25, 22, 16, 10, 20, 22, 19, 14, 30, 22, 12, 24, 28, 11

Solution:

$$n = 36, \bar{x} = \frac{\sum x}{n} = \frac{717}{36} = 19.92$$

Let $dx = x - A = x - 20$

$$\sum dx = -3, \sum dx^2 = 1151$$

$$s = \sqrt{\frac{\sum dx^2 - \frac{(\sum dx)^2}{n}}{n-1}} = \sqrt{\frac{1151 - \frac{(-3)^2}{36}}{35}} = 5.73$$

for σ into the confidence-interval formula of the above Result, we get

$$19.92 - 1.96\ \frac{5.73}{\sqrt{36}} < \mu < 19.92 + 1.96\ \frac{5.73}{\sqrt{36}}$$

$$18.05 < \mu < 21.79$$

Thus, the 95% confidence limits are 18.05 and 21.79 minutes.

3.2.7. Result

When we are dealing with a random sample from a normal population, $n < 30$, and σ is unknown, Results 3.2.1 and 3.2.3. cannot be used. Instead, we make use of the fact that $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ is a random variable having the t distribution with $n-1$ degrees of freedom. Substituting $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ for T in $P\left(-t_{\frac{\alpha}{2},n-1} < T < t_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$ we get the following confidence interval for μ.

If $\bar{x}$ and s are the values of the mean and the standard deviation of a random sample of size n from a normal population, then $\bar{x} - t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2},n-1} \cdot \frac{s}{\sqrt{n}}$ is a $(1-\alpha)100\%$ confidence interval for the mean of the population.This confidence-interval formula is used mainly when n is samll, less than 30, we refer to it as a small-sample confidence interval for μ.

3.2.8. Example

A paint manufacturer wants to determine the average drying time of a new interior wall paint. If for 12 test areas of equal size he obtained a mean drying time of 66.3 minutues and a standard deviation of 8.4 minutes, construct a 95% confidence interval for the true mean μ.

Solution:

Substituting $\bar{x} = 66.3$, $s = 8.4$ and $t_{0.025,11} = 2.201$ (from statistical table), the 95% confidence interval for μ becomes

$$66.3 - 2.201 \times \frac{8.4}{\sqrt{12}} < \mu < 66.3 + 2.201 \times \frac{8.4}{\sqrt{12}}$$

$$61 < \mu < 71.6$$

This mean that we can assert with 95% confidence that the interval from 61 minutes to 71.6 minutes contains the true average drying time of the paint.

3.2.9. Result

When we used the random variable $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ whose value cannoe calculated without knowledge of μ, but whose distribution for random samples from normal populations, the standard normal distribution, does not involve μ. This method of condfidence interval construcation is called the pivotal method.

3.3. The Estimation of Differences Between Means

3.3.1. Result

For independent random samples from normal populations

$Z = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ has the standard normal distribution.

If we substitute this expression of Z into $P(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) = 1 - \alpha$ the pivotal method yields the following confidence interval formula for $\mu_1 - \mu_2$.

If $\bar{x}_1$ and $\bar{x}_2$ are the values of the means of independent random samples of sizes n_1 and n_2 from normal populations with the known variances σ_1^2 and σ_2^2, then

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

is a $(1-\alpha)100\%$ confidence interval for the difference between the two population means.

By the central limit theorem, this confidence-interval formula can also be used for independent random samples from nonnormal populations with known variances with n_1 and n_2 are large, that is, when $n_1 \geq 30$ and $n_2 \geq 30$.

3.3.2. Example

Construct a 94% confidence interval for the diffference between the mean lifetimes of two kinds of light bulbs, given that a random sample of 40 light bulbs of the first kind lasted on the average 418 hours of continuous use and 50 light bulbs of the second kind lasted on the average 402 hours of continuous use. The population standard deviations are known to be $\sigma_1 = 26$ and $\sigma_2 = 22$.

Solution:

For $\alpha = 0.06$, we find from the statistical table that $z_{0.03} = 1.88$. Therefore, the 94% confidence interval for $\mu_1 - \mu_2$ is

$$(418 - 402) - 1.88 \times \sqrt{\frac{26^2}{40} + \frac{22^2}{50}} < \mu_1 - \mu_2 < (418 - 402) + 1.88 \times \sqrt{\frac{26^2}{40} + \frac{22^2}{50}}$$

$$6.3 < \mu_1 - \mu_2 < 25.7$$

Hencew, we are 94% confident that the interval from 6.3 to 25.7 hours contains the actual difference betweem the mean lifetimes of the two kinds of light bulbs. The fact that both confidence limits are positive suggests that on the average the first kind of light bulb is superior to the second kind.

3.3.4. Result

To Construct a $(1 - \alpha)100\%$ confidence interval for the difference between two means when $n_1 \geq 30$ and $n_2 \geq 30$, but σ_1 and σ_2 are unknown, we simply substitute s_1 and s_2 for σ_1 and σ_2 and proceed as before. When σ_1 and σ_2 are unknown and either or both of the samples are small, the procedure for estimating the difference between the means of two normal populations is not straight forward unless it can be assumed that $\sigma_1 = \sigma_2$. If $\sigma_1 = \sigma_2 = \sigma$, then $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ is a random variable having the standard distribution, and σ^2 can be estimated by pooling the squared deivations from the means of the two samples.

The Pooled estimator $S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ is an unbiased esitmator of σ^2. Now, by two theorems, "If $\bar{X}$ and S^2 are the mean and the variance of a random sample of size n from a normal population with the mean μ and the standard deivation σ then $(i)\bar{X}$ and S^2 are independent (ii) the random variable $\frac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with $n - 1$ degrees of freedom. If $X_1, X_2, \dots X_n$ are independent random variables having chi-square distributions with $v_1, v_2, \dots v_n$ degrees of freedom, then $Y = \sum_{i=1}^{n} X_i$, has the chi-square distribution with $v_1, +v_2 + \cdots + v_n$ degrees of freedom" the independent random variables $\frac{(n_1-1)S_1^2}{\sigma^2}$ and $\frac{(n_2-1)S_2^2}{\sigma^2}$ have chi-square distributions with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, and their sum $Y = \frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2} = \frac{(n_1+n_2-2)S_p^2}{\sigma^2}$ has a chi-squre distribution with $n_1 + n_2 - 2$ degrees of freedom. Since the random variables Z and Y are independent.

$T = \dfrac{Z}{\sqrt{\dfrac{Y}{n_1+n_2-2}}} = \dfrac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ has a t distribution with $n_1 + n_2 - 2$ degrees of freedom. Substituting this expression for T into $P\left(-t_{\frac{\alpha}{2},n-1} < T < t_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$, we get the following $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$.

If $\bar{x}_1$, $\bar{x}_2$, s_1 and s_2 are the values of the means and the standard deviations of independent random samples of sizes n_1 and n_2 from normal populations with equal variances, then

$$(\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2},n_1+n_2-2} \,.\, s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2},n_1+n_2-2} \,.\, s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

is a $(1-\alpha)100\%$ confidence interval for the difference between the two population means.

This confidence-interval formula is used mainly when n_1 and/or n_2 are small, less than 30, we refer to it as a small-sample confidence interval for $\mu_1 - \mu_2$.

3.3.5. Example

A study has been made to compare the nicotine contents of two brands of cigarettes. Ten cigarettes of Brand A has an average nicotine content of 3.1 milligrams with a standard deviation of 0.5 milligram. While eight cigarettes of Brand B had an average nicotine content of 2.7 milligrams with a standard deviatoin of 0.7 miligram. Assuming that the two sets of data are independent random samples from normal populations with equal variances, construct a 95% confidence interval for the difference between the mean nicotine contents of the two brands of cigarettes.

Solution:

Substitute $n_1 = 10, n_2 = 8, s_1 = 0.5$ and $s_2 = 0.7$ into the formula for s_p, we get

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{9(0.25)+7(0.49)}{16}} = 0.596$$

Then, substituting this value together with $n_1 = 10, n_2 = 8$, $\bar{x}_1 = 3.1, \bar{x}_2 = 2.7$ and $t_{0.0.25,16} = 2.120$ (form statistical table) into the confidence-interval formula, we find that the required 95% confidence interval is

$$(3.1-2.7) - 2.120 \times 0.596\sqrt{\frac{1}{10}+\frac{1}{8}} < \mu_1 - \mu_2 < (3.1-2.7) + 2.120 \times 0.596\sqrt{\frac{1}{10}+\frac{1}{8}}$$

$$-0.20 < \mu_1 - \mu_2 < 1.00$$

Thus, the 95% confidence limits are -0.20 and 1.00 milligrams; since this includes $\mu_1 - \mu_2 = 0$, we cannot conclude that there is a real difference between the average nicotine contents of the two brands of cigarettes.

3.4.The Estimation of Proportions

3.4.1. Result

In many problems we must esitmate proportions, probabilites, percentages or rates, such as the proportion of defectives in a large shipment of transistors, the probability that a car stopped at a road block will have faulty fights, the percentage of school children with I.Q.'s over 115 or the mortality rate of a disease. In many of these it is reasonable to assume that we are sampling a binomial population and hence our problem is to estimate the binomial parameter θ. Thus we can make use of the fact that for large n the binomial distribution can br approximated with a normal distribution; that is $Z = \frac{X-n\theta}{\sqrt{n\theta(1-\theta)}}$ can be treated as a random variable having approximately the standard normal distributon.

Substituting this expectation for Z into $P\left(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}\right) = 1-\alpha$, we get, $P\left(-z_{\frac{\alpha}{2}} < \frac{X-n\theta}{\sqrt{n\theta(1-\theta)}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$ and the two inequalities $-z_{\frac{\alpha}{2}} < \frac{X-n\theta}{\sqrt{n\theta(1-\theta)}}$ and $\frac{X-n\theta}{\sqrt{n\theta(1-\theta)}} < z_{\frac{\alpha}{2}}$, whose solution will give $(1-\alpha)100\%$ confidence limit for θ.

Let us give here instead a large sample approximation by rewriting

$P\left(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}\right) = 1-\alpha$ with $\frac{X-n\theta}{\sqrt{n\theta(1-\theta)}}$ substituted for Z, as

$$P\left(\hat{\Theta} - z_{\frac{\alpha}{2}}.\sqrt{\frac{\theta(1-\theta)}{n}} < \theta < \hat{\Theta} + z_{\frac{\alpha}{2}}.\sqrt{\frac{\theta(1-\theta)}{n}}\right) = 1-\alpha$$

where $\hat{\Theta} = \frac{X}{n}$ Then, if we substitute $\hat{\theta}$ for θ inside the radicals,which is a futher approximation, we get the following

If X is a binomial random variable with the parameters n and θ, n is large, and $\hat{\theta} = \frac{x}{n}$, then $\hat{\theta} - z_{\alpha/2}.\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} < \theta < \hat{\theta} + z_{\alpha/2}.\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$ is an approximate $(1-\alpha)100\%$ confidence interval for θ.

3.4.2. Example

In a random sample, 136 of 400 persons given a flu vaccine experienced some discomfort. Construct a 95% confidence interval for the true proportion of persons who will experience some discomfort from the vaccine.

Solution:

Substituting $n = 400, \hat{\theta} = \frac{136}{400} = 0.34$ and $z_{0.025} = 1.96$ into the confidence-interval formula, we get

$$\hat{\theta} - z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} < \theta < \hat{\theta} + z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

$$0.34 - 1.96\sqrt{\frac{(0.34)(0.66)}{400}} < \theta < 0.34 + 1.96\sqrt{\frac{(0.34)(0.66)}{400}}$$

$$0.294 < \theta < 0.386$$

$$0.29 < \theta < 0.39$$

3.4.3. Result

Using the same approximations that led to above result, we can get the following

If $\hat{\theta} = \frac{x}{n}$ is used as an estimate of θ, we can assert with $(1-\alpha)100\%$ confidence that the error is less than $z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$

3.4.4. Example

A study is made to determine the proportion of voters in a sizable community who favor the construction of a nuclear power plant. If 140 of 400 voters selected at random favor the project and we use $\hat{\theta} = \frac{140}{400} = 0.35$ as an estimate of the actual proportion of all voters in the community who favor the project, what can we say with 99% confidence about the maximum error?

Solution:

Substituting $n = 400, \hat{\theta} = \frac{140}{400} = 0.35$ and $z_{0.005} = 2.575$ into the formula we get

$$z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = 2.575\sqrt{\frac{(0.35)(0.65)}{400}} = 0.061 = 0.06$$

Thus, if we use $\hat{\theta} = \frac{140}{400} = 0.35$ as an estimate of the actual proportion of voters in the community who favor the projet, we can assert with 99% confidence that the error is less than 0.06

3.5. The Estimation of Differences Between Proportions

3.5.1. Result

In many probalems we must estimate the difference between the binomial parameters θ_1 and θ_2 on the basis of independent random samples of sizes n_1 and n_2 from two binomial populations. For example, if we want to estimate the difference between the poportions of male and female voters who favor a certain candidate for governor of Illinois.

If the respective numbers of successes are X_1 and X_2 and the corresponding sample proportions are denoted by $\hat{\theta}_1 = \frac{X_1}{n_1}$ and $\hat{\theta}_2 = \frac{X_2}{n_2}$. Let us investigage the sampleing distribution of $\hat{\theta}_1 - \hat{\theta}_2$, which is an obvious estimator of $\theta_1 - \theta_2$.

Let's take $E(\hat{\Theta}_1 - \hat{\Theta}_2) = \theta_1 - \theta_2$ and $var(\hat{\Theta}_1 - \hat{\Theta}_2) = \frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}$ and since, for large samples, X_1 and X_2, and hence also their differences, can be approximated with normal distributions, we get

$Z = \frac{(\hat{\Theta}_1-\hat{\Theta}_2)-(\theta_1-\theta_2)}{\sqrt{\frac{\theta_1(1-\theta_1)}{n_1}+\frac{\theta_2(1-\theta_2)}{n_2}}}$ is a random variable having approximately the standard normal distribution. Substituting this expression for Z into $P\left(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$, we get the following

If X_1 is a binomial random variable with the parameters n_1 and θ_1, X_2 is a binomial random variable with the parameters n_2 and θ_2, n_1 and n_2 are large, and $\hat{\theta}_1 = \frac{x_1}{n_1}$ and $\hat{\theta}_2 = \frac{x_2}{n_2}$, then

$$(\hat{\theta}_1 - \hat{\theta}_2) - z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}} < \theta_1 - \theta_2$$
$$< (\hat{\theta}_1 - \hat{\theta}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}$$

is an approximate $(1-\alpha)100\%$ confidence interal for $\theta_1 - \theta_2$.

3.5.2. Example

If 132 and 200 male voters and 90 of 150 female voters favor a certain candidate running for governor of Illinois, find a 90% confidence interval for the difference between the actual proportions of male and female voters who favor the candidate.

Solution:

Substituting $\hat{\theta}_1 = \frac{132}{200} = 0.66$, $\hat{\theta}_2 = \frac{90}{150} = 0.60$ and $z_{0.005} = 2.575$ into the confidece interal formula, we get

$$(\hat{\theta}_1 - \hat{\theta}_2) - z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}} < \theta_1 - \theta_2$$
$$< (\hat{\theta}_1 - \hat{\theta}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}$$

$$(0.66 - 0.60) - 2.575 \times \sqrt{\frac{(0.66)(0.34)}{200} + \frac{(0.60)(0.40)}{150}} < \theta_1 - \theta_2 <$$
$$(0.66 - 0.60) + 2.575 \times \sqrt{\frac{(0.66)(0.34)}{200} + \frac{(0.60)(0.40)}{150}}$$

$$-0.074 < \theta_1 - \theta_2 < 0.194$$

Thus, we are 99% confident that the interval form -0.074 to 0.194 contains the difference between the actual proportions of male and female voters who favor the condidate. This includes the possibility of a zero difference between the two proportions.

3.6. The Esimation of Variances

3.6.1. Result

Given a random sample of size n from a normal population, we can obtain a $(1-\alpha)100\%$ confidence interval for σ^2 by making use of the result, $\frac{(n-1)S^2}{\sigma^2}$ is a random variable having a chi-square distribution with $n-1$ degreees of freedom. Thus,

$$P\left[\chi^2_{1-\frac{\alpha}{2},n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2},n-1}\right] = 1-\alpha$$

$$P\left[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right] = 1-\alpha$$

Thus, we get the following

If S^2 is the value of the variance of a random sample of size n from a normal population, then $\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}$ is a $(1-\alpha)100\%$ confidence interval for σ^2.

3.6.2. Example

In 16 test runs the gasoline consumption of an experimental engine had a standard of 2.2 gallons. Construct a 99% confidence interval for σ^2, which measures the true variability of the gasoline consumption of the engine.

Solution:

Assuming that the orbserved data can be looked upon as a random sample from a normal population. We substitute $n = 16$ and $s = 2.2$, along with $\chi^2_{0.005,15} = 32.801$ and $\chi^2_{0.995,15} = 4.601$, obtained from tatistical tables, into the confidence-interval formula we get,

$$\frac{15\,(2.2)^2}{32.801} < \sigma^2 < \frac{15\,(2.2)^2}{4.601}$$

$$2.21 < \sigma^2 < 15.78$$

For 99% confidence interval , $1.49 < \sigma < 3.97$

3.7. The Estimation of the Ratio of Two Variances

3.7.1. Result

If S_1^2 and S_2^2 are the variances of independent random samples of sizes n_1 and n_2 from normal populations, then, according to the theorem, "If S_1^2 and S_2^2 are the variances of independent random samples of sizes n_1 and n_2 from normal populations with the variances

σ_1^2 and σ_2^2, then $F = \frac{\left(\frac{S_1^2}{\sigma_1^2}\right)}{\left(\frac{S_2^2}{\sigma_2^2}\right)} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$ is a random variables having an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom"

$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$ is a random variable having an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. Thus, we can write $P\left(f_{1-\frac{\alpha}{2},n_1-1,n_2-1} < \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} < f_{\frac{\alpha}{2},n_1-1,n_2-1}\right) = 1 - \alpha$

Since $f_{1-\frac{\alpha}{2},n_1-1,n_2-1} = \frac{1}{f_{\frac{\alpha}{2},n_2-1,n_1-1}}$, we have the following

If s_1^2 and s_2^2 are the values of the variances of independent random samples of sizes n_1 and n_2 from normal populations, then

$\frac{s_1^2}{s_2^2} \cdot \frac{1}{f_{\frac{\alpha}{2},n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \cdot f_{\frac{\alpha}{2},n_2-1,n_1-1}$ is a $(1-\alpha)100\%$ confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$.

Corrponding $(1-\alpha)100\%$ confidence limits for $\frac{\sigma_1}{\sigma_2}$ can be obtained by taking the square roots of the confidence limit for $\frac{\sigma_1^2}{\sigma_2^2}$.

3.7.2. Example

A study has been made to compare the nicotine contents of two brands of cigarettes. Ten cigarettes of Brand A has an average nicotine content of 3.1 milligrams with a standard deviation of 0.5 milligram. While eight cigarettes of Brand B had an average nicotine content of 2.7 milligrams with a standard deviatoin of 0.7 miligram. Assuming that the two sets of data are independent random samples from normal populations with equal variances. Find a 98% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$

Solution:

Substituting $n_1 = 10, n_2 = 8, s_1 = 0.5, s_2 = 0.7$, and

$f_{0.01,9,7} = 6.72$ and $f_{0.01,7,9} = 5.61$ from the statistical table, we get

$$\frac{0.25}{0.49} \cdot \frac{1}{6.42} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{0.25}{0.49} .5.61$$

$$0.076 < \frac{\sigma_1^2}{\sigma_2^2} < 2.862$$

Since the interval obtained here includes the possibility that the ratio is 1, there is no real evidence against the assumption of equal population variances.

Let Us Sum Up

In this unit, we discussed the concept of interval estimation, in particularly the Estimation of Means, differences between means, proportions, and differences between proportions, variances and ratio of two variances.

Check Your Progress

1. The interval bounded by two limits is known as_______.

2. The end points of the confidence interval are called________.

3. A random sample of size 100 has mean 15, the population variance being 25. The interval estimates of the population mean with a confidence level of 99% is______.

Glossaries

Interval estimation: It is the range of values used in making estimation of a population parameter.

Population proportion: The population proportion P is the ratio of the number of elements possessing a characteristic to the total number of elements in the population.

Sample Proportion: The sample proportion p is the ratio of the number of elements possessing to the total number of elements n in the sample.

Degrees freedom: The degrees freedom is the number of independent random variables.

Suggested Readings

1. Freund. J.E.," Mathematical Statistics", Prentice Hall of India, Fifth Edition, 2001.

2. Gupta. S.C. and Kapoor. V. K., "Fundamentals of Mathematical Statistics", Sultan Chand & Sons, Eleventh Edition, 2003.

3. Devore. J. L. "Probability and Statistics for Engineers", Brooks/Cole (Cengage Learning), First India Reprint, 2008.

Answers to Check Your Progress

1. Confidence interval

2. Confidence limits

3. 13.71 to 16.29

BLOCK II: Testing of Hypothesis

Unit – 4

Hypothesis Testing

Objectives

After Studying this Unit, the student will be able to

- Demonstrate the simple hypothesis, alternative hypothesis, Type I and Type II errors, Critical Region.
- Explain the Neyman-Pearson lemman with example.
- Explain the Power function and the uniformly most powerful critical region test
- Summarize the Likelihood ratio test..

Overview

In this unit, we will study the concept of testing a statistical hypothesis, the Neyman-Pearson Lemma, the Power function of a test, Likelihood ratio test with examples.

4.1. Introduction

If an engineer has to decide on the basis of sample data whether the true average life time of certain kind of tire is at least 42,000 miles, if an agronomist has to decide on the basis of experiments whether one kind of fertilizer produces a higher yield of soybeans than another, and if an manufacturer of pharmaceutical products has to decide on the basis of samples whether 90 percent of all patients given a new medication will recover from a certain disease, these problems can all be translated into the language of statistical tests of hypotheses. In the first case we might say that the engineer has to test the hypothesis that θ, the parameter of an exponential population, is at least 42,000; in the second case we might say that the agronomist has to decide whether $\mu_1 > \mu_2$, where μ_1 and μ_2 are the means of two normal populations; and in the third case we might say that the manufacturer has to decide whether θ, the parameter of a binomial population, equals 0.90. In each case it must be assumed that the chosen distribution correctly describes the experimental conditions. That is, the distribution provides the correct statistical model.

4.1.1. Statistical Hypothesis

An assertion or conjecture about the distribution of one or more random variables is called a statistical hypothesis. If a statistical hypothesis completely specifies the distribution, it is called a simple hypothesis, if not; it is referred to as a composite hypothesis.

A simple hypothesis is not only the functional form of the underlying distribution, but also the values of all parameters. In the third of the above examples, the effectiveness of the new medication, the hypothesis $\theta = 0.90$ is simple, assuming that we specify the sample size and that the population is binomial. In the first of the preceding examples the hypothesis is composite since $\theta \geq 42{,}000$ does not assign a specific value to the parameter θ.

For testing statistical hypotheses, it is necessary that we formulate alternative hypotheses. In the first example dealing with the lifetimes of the tires, we might formulate the alternative hypothesis that the parameter θ of the exponential population is less than 42,000. In the second example dealing with the two kinds of fertilizer, we might formulate the alternative hypothesis$\mu_1 = \mu_2$. In the third example dealing with the new medication, we mightformulate the alternative hypothesis that the parameter θ of the given binomial population is only 0.60, which is the disease's recovery rate without the new medication.

The concept of simple simple and composite hypothesis applies also to alternative hypotheses. In the first example we can say that we testing the compositive hypothesis $\theta \geq 42000$ against the composite alternative $\theta < 42{,}000$, where $\theta < 42{,}000$, where θ is the parameter of an exponential population. In the second example we are testing the composite hypothesis $\mu_1 > \mu_2$ against the composite alternative $\mu_1 = \mu_2$ where μ_1, μ_2 are the means of two normal populations. In the third exmple we are tesing the simple hypotheis $\theta = 90$ against the simple alternative $\theta = 60$, where θ is the parameter of a binomial population for which n is given.

If we want to show that the students in one school have higher average I.Q. than those in another school, we formulate the hypothesis that there is no difference: the hypothesis $\mu_1 = \mu_2$.

In view of the assumptions of "no difference", hypotheses such as these led to the term null hypothesis, but this term is applied to any hypotheis that we may want to test.

We use the symbol H_0 for the null hypothesis that we want to test and H_1 or H_A for the alternative hypothesis.

4.2.Testing a Statistical Hypothesis

4.2.1. Type I and Type II errors

1. Rejection of a null hypothesis when it is true is called a type I error. The probability of committing a type I error is denoted by α.

2. Acceptance of the null hypothesis when it is false is called a type II error. The probability of commiting a type II error is denoted by β.

	H_0 $is\ true$	H_0 $is\ false$
$Accept\ H_0$	$No\ error$	$Type\ II\ error\ probability = \beta$
$Reject\ H_0$	$Type\ I\ error\ probability = \alpha$	$No\ error$

4.2.2. Critical Region

It is customary to refer to the rejection region for H_0 as the critical region of a test. The probability of obtaining a value of the test statistic inside the critical region when H_0 is true is called the size of the critical region. Thus, the size of the critical region is just the probability α of committing a type I error. This probability is also called the level of significance of the test.

4.2.3. Examples

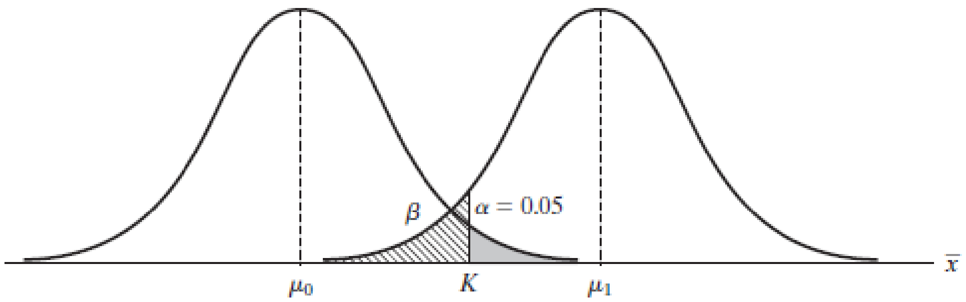
4.2.3.1. Suppose that the manufacturer of a new medication wants to test the null hypothesis $\theta = 0.90$ against the alternative hypothesis 0.60. His test statistic is X, the observed number of successes (recoveries) in 20 trials, and he will accept the null hypothesis if $x > 14$; otherwise, he will reject. Find α and β.

Solution:

The acceptance region for the hull hypothesis is $x = 15, 16, 17, 18, 19$ and 20, and correspondingly, the rejection region or critical region is $x = 0, 1, 2{,}3, \dots , 14$. Therefore, from the Binomial Probabilites table of statistical tables we have
$\alpha = P(X \leq 14;\ \theta = 0.90) = 0.0114$ and $\beta = P(X > 14;\ \theta = 0.60) = 0.1255$

4.2.3.2. Suppose that we want to test the null hypothesis that the mean of a normal population with $\sigma^2 = 1$ is μ_0 against the alternative hypothesis that it is μ_1, where $\mu_1 > \mu_0$. Find the value of K such that $\bar{x} > K$ provides a critical region of size $\alpha = 0.05$ for a random sample of size n.

Solution:



From the above figure and the standard normal distribution table of statisitcal tables, we find that $z = 1.645$ corresponds to an entry of 0.45 and hence that

$$1.645 = \frac{K - \mu_0}{1/\sqrt{n}}$$

$$K = \mu_0 + \frac{1.645}{\sqrt{n}}$$

4.2.3.3. With reference to the previous example, Determine the minimum sample size needed to test the null hypothesis $\mu_0 = 10$ against the alternative hypothesis $\mu_1 = 11$ with $\beta \leq 6$.

Solution:

Since β is given by the area of the ruled region of the above figure, we get

$$\beta = P\left(\bar{X} < 10 + \frac{1.645}{\sqrt{n}};\ \mu = 11\right)$$

$$\beta = \left[Z < \frac{\left(10 + \frac{1.645}{\sqrt{n}}\right) - 11}{1/\sqrt{n}}\right]$$

$$\beta = (Z < -\sqrt{n} + 1.645)$$

and since $z = 1.555$ corresponds to an entry of $0.5 - 0.06 = 0.44$ in the standard normal distribution table of statistical table, we get $-\sqrt{n} + 1.645$ equal to -1.555. $\sqrt{n} = 1.645 + 1.555 = 3.2$ and $n = 10.24$ or 11.

4.3. Losses and Risks

The concepts of loss functions and risk functions also play an important part in the theory of hypotesis testing. In the decision theory approach to testing the null hypothesis that a population parameter θ equals θ_0 against the alternative that it equals θ_1, the statistican either takes the action $a_0$and accepts the null hypothesis, or takes the action a_1 accepts the alternative hypothesis. Depending on the true "state of Nature" and the action that she takes, her losses are shown in the following table

		Statistician	
		a_0	a_1
Nature	θ_0	$L(a_0, \theta_0)$	$L(a_1, \theta_0)$
	θ_1	$L(a_0, \theta_1)$	$L(a_1, \theta_1)$

These losses can be positive or negative (reflecting penalties or rewards), and the only condition that we shall impose is that

$$L(a_0,\ \theta_0) < L(a_1,\ \theta_0)\ \ and\ L(a_1,\ \theta_1) < L(a_0,\ \theta_1)$$

That is, in either case the right decision is more profitable than the wrong one.

The statistician's choice will depend on the outcome of an experiment and the decision funciton d, which tell her for each possible outcome what action to take. If the null hypothesis is true and the statistician accepts the alternative hypothesis, that is, if the value of the parameter θ_0 and the statistician takean action a_1, she commits a type I error; correspondingly, if the value of the parameter is θ_1 and the statistician takes action a_0, she commits a type II error. For the decision function d, we shall let $\alpha(d)$ denote the probability of committing a type I error and $\beta(d)$ the probability of committing a type II error. The values of the risk function are that

$$R(d, \theta_0) = [1 - \alpha(d)]L(a_0,\ \theta_0) + \alpha(d)L(a_1,\ \theta_0)$$

$$R(d,\ \theta_0) = L(a_0,\ \theta_0) + \alpha(d)[L(a_1,\ \theta_0) - L(a_0,\ \theta_0)]$$

and

$$R(d, \theta_1) = \beta(d)L(a_0,\ \theta_1) + [1 - \beta(d)]\ L(a_1,\ \theta_1)$$

$$R(d,\ \theta_1) = L(a_1,\ \theta_1) + \beta(d)[L(a_0,\ \theta_1) - L(a_1,\ \theta_1)]$$

Where, by assumption, the quantities in brackets are both positive. It is apparent from this that to minimize the risks the statistician must choose a decision funciton that, keeps the probabilities of both types of errors as small as possible.

If we could assign prior probabiiities to θ_0 and θ_1 and if we know the exact values of all the losses $L(a_i,\ \theta_j)$, we could calculate the Bayes risk and look for the decision funciton that minimize this risk. Alternatively, if we looked upon nature as a malevolent oppeonent, we could use the mimimax criterion and choose the decision funciton that minimizez the maximum risk.

4.4. The Neyman-Pearson Lemma

4.4.1. The Power of a Test

When testing the null hypothesis $H_0: \theta = \theta_0$ against the alternaive hypothesis $H_1: \theta = \theta_1$, the quantitiy $1 - \beta$ is referred to as the power of the test $\theta = \theta_1$. A critical region for testing a simple null hypothesis H_0: $\theta = \theta_0$ against a simple alternative hypothesis H_1: $\theta = \theta_1$ is siad to be a best critical region or a most powerful critical region if the power of the test is maximum at $\theta = \theta_1$.

To Construct a most powerful critical region

The likelihoods of a random sample of size n from the population under consideration when $\theta = \theta_0$ and $\theta = \theta_1$. Denoting these likelihoods by L_0 and L_1, we have

$$L_0 = \prod_{i=1}^{n} f(x_i, \theta_0) \quad and\ L_1 = \prod_{i=1}^{n} f(x_i, \theta_1)$$

$\frac{L_0}{L_1}$ Should be small for sample points inside the critical region, which lead to type I erros when $\theta = \theta_0$ and to correct decisions when $\theta = \theta_1$.

$\frac{L_0}{L_1}$ Should be large for sample points inside the critical region, which lead to correct decisions when $\theta = \theta_0$ and type II erros when $\theta = \theta_1$.

4.4.2. Theorem (Neyman-Pearson Lemma)

If C is a critical region of size α and k is a constant such that $\frac{L_0}{L_1} \leq k\ inside\ C$ and $\frac{L_0}{L_1} \geq k\ \ outside\ C$ then C is a most powerful critical region of size α for testing $\theta = \theta_0$ against $\theta = \theta_1$.

Proof:

Suppose that C is a critial region satisfying the conditions of the theorem and that D is some other critical region of size α. Thus,

$$\int_C \dots \int L_0\ dx = \int_D \dots \int L_0\, dx = \alpha$$

where dx stads for $dx_1, dx_2, \dots, dx_n$ and the two multiple integrals are taken over the respective n-diemansional regions $C\ and\ D$. Now, making use of the fact that C is the union of disjoint sets $C \cap D$ and $C \cap D'$, while D is the union of the disjoint sets $C \cap D$ and $C' \cap D$, we can write

$$\int_{C \cap D} \dots \int L_0\ dx + \int_{C \cap D'} \dots \int L_0\ dx = \int_{C \cap D} \dots \int L_0\ dx + \int_{C' \cap D} \dots \int L_0\ dx = \alpha$$

and hence

$$\underset{C \cap D'}{\int \dots \int} L_0 \ dx = \underset{C' \cap D}{\int \dots \int} L_0 \ dx$$

Then, since $L_1 \geq \frac{L_0}{k}$ inside C and $L_1 \leq \frac{L_0}{k}$ outside C,

$$\underset{C \cap D'}{\int \dots \int} L_1 \ dx \geq \underset{C \cap D'}{\int \dots \int} \frac{L_0}{k} \ dx = \underset{C' \cap D}{\int \dots \int} \frac{L_0}{k} dx \geq \underset{C' \cap D}{\int \dots \int} L_1 \ dx$$

and hence

$$\underset{C \cap D'}{\int \dots \int} L_1 \ dx \geq \underset{C' \cap D}{\int \dots \int} L_1 \ dx$$

Finally,

$$\underset{C}{\int \dots \int} L_1 \ dx = \underset{C \cap D}{\int \dots \int} L_1 \ dx + \underset{C \cap D'}{\int \dots \int} L_1 \ dx$$

$$\underset{C}{\int \dots \int} L_1 \ dx = \underset{C \cap D}{\int \dots \int} L_1 \ dx + \underset{C' \cap D}{\int \dots \int} L_1 \ dx = \underset{D}{\int \dots \int} L_1 \ dx$$

So that

$$\underset{C}{\int \dots \int} L_1 \, dx \geq \underset{D}{\int \dots \int} L_1 \, dx = \alpha$$

The final inequality states that for the critical region C the probability of not committing a type II error is greater than or equal to the corresponding probability for any other region of size α.

4.4.3. Example

A random sample of size n from a normal population with $\sigma^2 = 1$ is to be used to test the null hypothesis $\mu = \mu_0$ against the alternatiave hypothesis $\mu = \mu_1$, where $\mu_1 > \mu_0$. Use the Neyman-Pearson lemma to find the most powerful critical region of size α.
Solution:

The two likelihoods are

$$L_0 = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\Sigma(x_i - \mu_0)^2} \quad and \ L_1 = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\Sigma(x_i - \mu_1)^2}$$

Where the summations extend from $i = 1 to \ n$, and after some simplification their ratio becomes

$$\frac{L_0}{L_1} = e^{\frac{n}{2}(\mu_1^2-\mu_0^2)+(\mu_0-\mu_1)\sum x_i}$$

Thus, we find a constant k and a region C of the sample sapce such that

$$e^{\frac{n}{2}(\mu_1^2-\mu_0^2)+(\mu_0-\mu_1)\sum x_i} \leq k \quad inside\ C$$

$$e^{\frac{n}{2}(\mu_1^2-\mu_0^2)+(\mu_0-\mu_1)\sum x_i} \geq k \quad Outside\ C$$

and after taking logarithms, subtractng $\frac{n}{2}(\mu_1^2 - \mu_0^2)$, and dividing by the negative quantity $n(\mu_0 - \mu_1)$, these two inequalities become

$$\bar{x} \leq K \quad inside\ C$$

$$\bar{x} \geq K \quad Outside\ C$$

where K is an expresssion in $k, n, \mu_0\ and\ \mu_1$.

4.5. The Power Funciton of a Test

4.5.1. Power Function

The Power function of a test of a statistical hypothesis H_0 against an alternative hypothesis H_1 is given by

$$\pi(\theta) = \begin{cases} \alpha(\theta) & for\ values\ of\ \theta\ assumed\ under\ H_0 \\ 1-\beta(\theta) & for\ values\ of\ \theta\ assumed\ under\ H_1 \end{cases}$$

4.5.2. Uniformly Most Powerful Critical Region (Test)

If, for a given problem, a critical region of size α is uniformly more powerful than any other critical region of size α, it is said to be uniformly most powerful critical region, or a uniformly most powerful test.

4.6. Likelihood Ratio Tests

The Neyman-Pearson lemma provides a means of constructing most powerful critical regions for testing a simple null hypothesis against a simple alternative hypothesisi, but it does not always apply to compositie hypotheses. We shall now present a general method for constructing critical regions for tests of composite hypotheses that in most caes have very satisfactory properties. The resulting tests, called Likelihood ratio tests, are based on a generalization of the method of Neyman-Pearson lemma, but they are not necessarily uniformly most powerful.

To Illustrate the likelihood ratio technique, Let us suppose that $X_1, X_2, \ldots X_n$ constitute a random sample of size n form population whose density at x is $f(x;\ \theta)$ and that Ω is the set of values that can be taken on by the parameter θ. We refer Ω as the parameter space for θ. To test the null hypothesis is $H_0: \theta \in \omega$ and the alternative hypothesis is $H_1: \theta \in \omega'$, where ω is the subset of Ω and ω' is the complement of ω with respect to Ω. Thus, the parameter space for θ is partitioned into the disjoint sets ω and ω'. The null hypothesis is θ is an element of the first set and the alternative hypothesis θ is an element of the second set. Ω is either the set of all real numbers, the set of all positive real numbers, some interval of real numbers or a discrete set of real numbers.

When H_0 and H_1 are both simple hypotheses, c each have one element, and in 4.4. we constructed tests by comparing the likelihood L_0 and L_1. In the general case, where at least one of the two hypotheses is composite, we comare instead the two quantitties max $L_0$and max L, where max L_0 is the maximum value of the likelihood funciton for all values of θ in ω, and max L is the maximum value of the likelihood function for all values of θ in Ω. In other words, if we have a random sample of size n from a population whose density at x is $f(x;\ \theta)$, $\hat{\theta}$ is the maximum likelihood estimate of θ subject to the restriction that θ must be an element of ω, and $\hat{\hat{\theta}}$ is the maximum likelihood estimate of θ for all value of θ in Ω, then

$max\ L_0 = \prod_{i=1}^{n} f(x;\hat{\theta})$ and $max\ L = \prod_{i=1}^{n} f(x;\hat{\hat{\theta}})$

These quantities are both values of random varaibles, since they depend on the observed values $x_1, x_2, \ldots, x_n$, and their ratio $\lambda = \frac{max\ L_0}{max\ L}$ is known as a value of the likelihood ratio statistic.

Since $max\ L_0$ and $max\ L$ are both values of a likelihood function and therefore are never negative. Therefore $\lambda \geq 0$; also, since ω is a subset of the parameter space Ω, therefore $\lambda \leq 1$. When the null hypothesis is false, would expect $max\ L_0$ to be small compared to $max\ L$, in which case λ would close to zero. On the other hand, when the null hypothesis is true and $\theta \in \omega$, we would expect $max\ L_0$ to be close to $max\ L$, in which case λ would be close to 1. A likelihood ratio test states that the null hypothesis H_0 is rejected if and only if λ falls in a critical region of the form $\lambda \leq k$, where $0 < k < 1$.

4.6.1. Likelihood Ratio Test

If ω and ω' are complementary subsets of the parameter space Ω and if the likelihood ratio statistic $\lambda = \frac{max\ L_0}{max\ L}$ where $max\ L_0$ and $max\ L$ are the maximum values of the likelihood function for all values of θ in ω and Ω, respectively, then the critical region $\lambda \leq k$, where $0 < k < 1$, defines a likelihood ratio test of the null hypothesis $\theta \in \omega$ against the alternative hypothesis $\theta \in \omega'$.

If H_0 is a simple hypothesis, k is chosen so that the size of the critical region equals α; if H_0 is composite, k is chosen so that the probability of a type I error is less than or equal to α for all θ in ω, and equal to α, if possible, for at least one value of θ in ω. Thus, if H_0 is a simple hypothesis and $g(\lambda)$ is density of Λ at λ when H_0 is true, then k must be such that $P(\Lambda \leq k) = \int_0^k g(\lambda)d\lambda = \alpha$

In the discrete case, the integral is replaced by a summation, and k is taken to be the large values for which the sume is less than or equal to α.

4.6.2. Example

Find the critical region of the likelihood ratio test for testing the null hypothesis H_0: $\mu = \mu_0$ against the composite alternative H_1: $\mu \neq \mu_0$ on the basis of a random sample of size n from a normal population with the known variance σ^2.

Solution:

Since ω contains only μ_0, $\hat{\mu} = \mu_0$, and since Ω is the set of all real number, $\hat{\mu} = \bar{x}$. Thus,

$max\, L_0 = (\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\mu_0)^2}$ and $max\, L = (\frac{1}{\sigma\sqrt{2\pi}})^n e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\bar{x})^2}$

where the summation from $i = 1\ to\ n$, and the value of the likelihood ratio statistic becomes

$$\lambda = \frac{e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\mu_0)^2}}{e^{-\frac{1}{2\sigma^2}\Sigma(x_i-\bar{x})^2}}$$

$$\lambda = e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu_0)^2}$$

Hence , the critical region of the likelihood ratio test is $e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu_0)^2} \leq k$ and taking logarithms and dividing by $-\frac{n}{2\sigma^2}$, we have

$$(\bar{x} - \mu_0)^2 \geq -\frac{2\sigma^2}{n}.\ln k$$

$$|\bar{x} - \mu_0| \geq K$$

Where K will have to be determined so that the size of the critical region is α.

Since $\bar{X}$ has a normal distribution with the mean μ_0 and the variance $\frac{\sigma^2}{n}$, the critical region of this likelihood ratio test is

$$|\bar{x} - \mu_0| \geq z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}$$

$|z| \geq z_{\frac{\alpha}{2}}$ where $z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$

In other words, the null hypothesis must be rejected when Z takes on a value greater than or equal to $z_{\frac{\alpha}{2}}$ or a value less than or equal to $-z_{\frac{\alpha}{2}}$

4.6.3. Example

On the basis of a single observation, we want to test the simple null hypothesis that the probability distribution of X is

x	1	2	3	4	5	6	7
$f(x)$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

against the composite alternative that the probability distribution is

x	1	2	3	4	5	6	7
$g(x)$	$\frac{a}{3}$	$\frac{b}{3}$	$\frac{c}{3}$	$\frac{2}{3}$	0	0	0

where $a+b+c=1$. Show that the critical region obtained by means of the likelihood ratio technique is inadmissible.

Soluiton:

The compositie alternative bypothesis includes all the probability distributions that we get by assigning different values from $0\ to\ 1\ to\ 1, a, b, and\ c,$ subject only to the restriction that $a+b+c=1$.
For each value of x, let $x=1$, for this value we get $max\ L_0 = \frac{1}{12}, max\ L = \frac{1}{3}$ (corresponding to $a=1$) and hence $\lambda = \frac{1}{4}$.

Determining λ for the other values of x in the same way, we get

x	1	2	3	4	5	6	7
λ	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	1	1	1

If the size of the critical region is to be $\alpha = 0.25$, we find that the likelihood ratiod technique yields, the critical region for which the null hypothesis is rejected when $\lambda = \frac{1}{4}$. This is, when $x=1, x=2, x=3$, we have $f(1)+f(2)+f(3) = \frac{1}{12}+\frac{1}{12}+\frac{1}{12} = 0.25$. The corresponding probability of a type II error is given by $g(4)+g(5)+g(6)+g(7) = \frac{2}{3}$

Now, let us consider the critical region for which the null hypothesis is rejected only when $x=4$. Its size is also $\alpha = 0.25$ since $f(4) = \frac{1}{4}$, but the corresponding probability of a type II error is $g(1)+g(2)+g(3)+g(5)+g(6)+g(7) = \frac{a}{3}+\frac{b}{3}+\frac{c}{3}+0+0+0 = \frac{1}{3}$.

Since theis is less than $\frac{2}{3}$, the critical region obtained by means of the likelihood ratio technique is inadmissible.

Let Us Sum Up

In this unit, we discussed the concept of testing a statistical hypothesis, the Neyman-Pearson Lemma, the Power function of a test, Likelihood ratio test with examples.

Check Your Progress

1. The Probabilities of committing the type I and type II errors are called______.

2. The Power of a test is maximum, when the probability of type II error is______.

3. If both null hypothesis and alternative hypothesis are simple hypotheses, then Likelihood ratio test is__________.

4. The Likelihood ratio test is a generalization of____________.

Glossaries

Hypothesis: A hypothesis is a statement about the population parameter.

Type I error: It is the error of rejecting null hypothesis when it is true.

Type II error: It is the error of accepting the null hypothesis when it is false.

Critical Region: It is the region of the standard normal curve corresponding to a predetermined level of significance.

Suggested Readings

1. Freund. J.E.," Mathematical Statistics", Prentice Hall of India, Fifth Edition, 2001.

2. Gupta. S.C. and Kapoor. V. K., "Fundamentals of Mathematical Statistics", Sultan Chand & Sons, Eleventh Edition, 2003.

3. Devore. J. L. "Probability and Statistics for Engineers", Brooks/Cole (Cengage Learning), First India Reprint, 2008.

Answers to Check Your Progress

1. Sizes of errors

2. Minimum.

3. Neyman-Pearson Lemma

4. Neyman-Pearson Lemma

Unit – 5

Testing of Hypothesis involving Means, Variances and Proportions

Objectives

After Studying this Unit, the student will be able to

- Analyse and compare the tests based on normal, t, $\chi^2$and F distributions for testing of mean, variance and proportions.
- Explain the tests for Independence of attributes and Goodness of fit.
- Illustrate with the numerical examples in normal, t, $\chi^2$and F distributions.

Overview

In this unit, we will study the tests based on normal, t, $\chi^2$and F distributions for testing of means, variance and proportions and tests for Independence of attributes and Goodness of fit.

5.1. Introduction

We shall preseent some of the standard tests that are most widely used in applications. Most of these tests, at least those based on known population distributions, can be obtained by the likelihood ratio technique.

5.1.1. Test of Significance

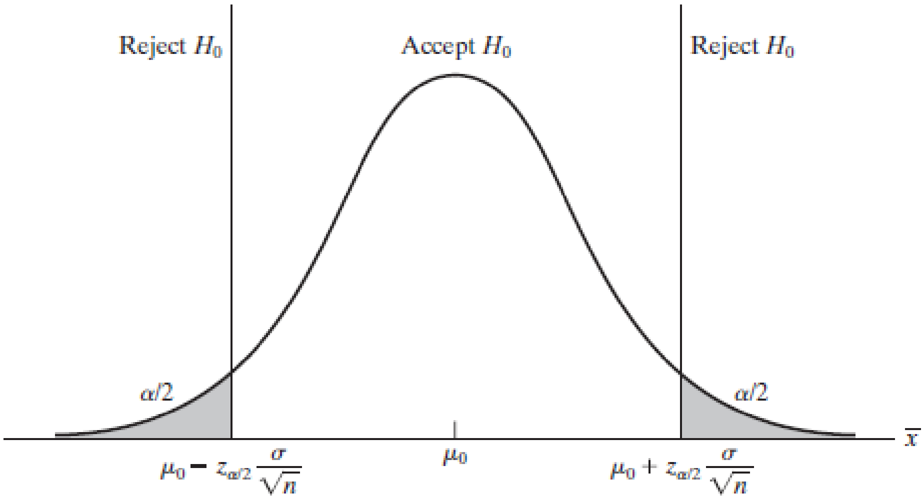
A statistical test which specifies a simple null hypothesis, the size of the critical region, α, and a composite alternative hypothesis is called a tet of significance. In such a test, α is referred to as the level of signficance.

5.1.2. Two Tailed Test

When thes test of hypothesis is made on the basis of rejection region represented by both sides of the standard normal curve, it is called a two tailed test. A test of statistical hypothesis where the alternative hypothesis is two tailed such as

Null Hypothesis $H_0 : \mu = \mu_0$

Alterntaive Hypothesis $H_1 : \mu \neq \mu_0$



Critical region for two-tailed test.

Or

$$\bar{x} \le \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ and } \bar{x} \ge \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

5.1.3. One tailed test

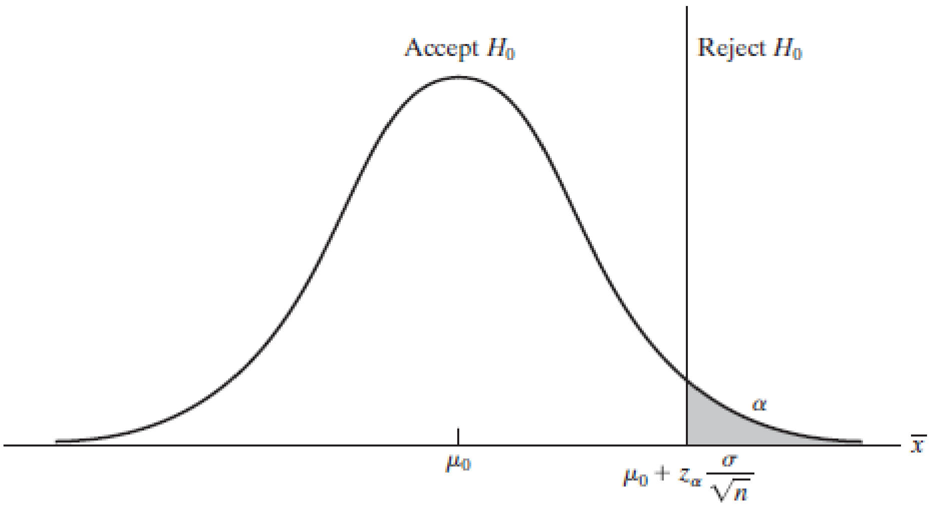
A test of statistical hypothesis, where the alternative hypothesis is one side is called as one tailed test.

There are two types of one tailed test.

1. Right tailed test: In the right tailed test the rejection region or critical region lies entirely on the right tail of the normal curve.

Null Hypothesis $H_0 : \mu = \mu_0$

Alterntaive Hypothesis $H_1 : \mu > \mu_0$ (Right tailed)

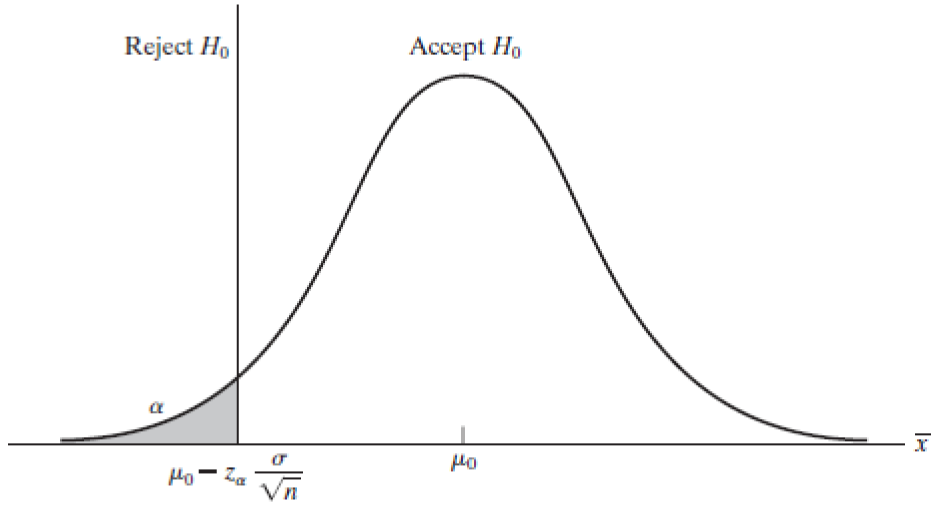


Critical region for one-tailed test ($H_1 : \mu > \mu_0$).

2. Left tailed test: In the left tailed test the rejection region or critical region lies entirely on the left tail of the normal curve.

Null Hypothesis $H_0 : \mu = \mu_0$

Alterntaive Hypothesis H_1: $\mu < \mu_0$ (Left tailed)



Critical region for one-tailed test ($H_1: \mu < \mu_0$).

5.1.4. The following are the steps for testing of hypothesis by means

1. Formulate H_0 and H_1, and specify α.

2. Usng the sampling distribution of an appropriate test statistic, determine a critical region of size α.

3. Determine the value of the test statistic from the sample data.

4. Check whether the value of the test statistic falls into the critical region and accordingly, reject the null hypothesis, or reserve judgement. (Note that we do not accept the null hypothesis because β, the probability of false acceptance, is not specified in a test of significance)

Definition: (P- Value) Corresponding to an observed value of a test statistic, the P-value is the lowest level of significance at which the null hypothesis could have been rejected.

5.1.5. Alternative approach to testing hypotheses

1. Formulate H_0 and H_1, and specify α.

2. Specify the tst statistic.

3. Determine the value of the test statistic and the corresponding P-value from the sample data.

4. Check whether the P-value is less than or equal to α and, accordingly, reject the null hypothesis, or reserve judgement.

5.2. Test Concerning Means

Suppose that we want to test the null hypothesis $\mu = \mu_0$ against one of the alternatives $\mu \neq \mu_0$, $\mu > \mu_0$ or $\mu < \mu_0$ on the basis of a random sample of size n from a normal population with the known variance σ^2. Thre critical regions for the respective alternatives are $|z| \geq z_{\alpha/2}$, $z \geq z_\alpha$ and $z \leq -z_\alpha$, where $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

The most commonly used levels of significance are 0.05 and 0.01, and the corresponing values of z_α and $z_{\alpha/2}$ are $z_{0.05} = 1.645, z_{0.01} = 2.33, z_{0.025} = 1.96$ and $z_{0.005} = 2.575$.

5.2.1. Example

Suppoe that it is known from experience that the standard deviation of the weight of 8-ounce package of cookies made by a certain bakery is 0.16 ounce. To chcek whether its production is under control on a given day, that is, to check whether the true average weight of the packages is 8 ounces, employees select a random sample of 25 packages and find that their mean weight is $\bar{x} = 8.091$ ounces. Since the bakery stands to lose money when $\mu > 8$ and the custormer loses out when $\mu < 8$, test the null hypothesis $\mu = 8$ against the alternative hypothesis $\mu \neq 8$ at the 0.01 level of significance.

Solution:

1. $H_0: \mu = 8$
$H_1: \mu \neq 8$
$\alpha = 0.01$

2. Reject the null hypothesis if $z \leq -2.575$ or $z \geq 2.575$

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

3. Substituting $\bar{x} = 8.091, \mu_0 = 8, \sigma = 0.16, and\ n = 25$, we get

$$z = \frac{8.091 - 8}{0.16/\sqrt{25}} = 2.84$$

4. Since $z = 2.84$ exceeds 2.575, the null hypothesis must be rejected and suitable adjustments should be made in the production process.

5.2.2. Larage-sample test.

When we dealing with a large sample of size $n \geq 30$ from a population that need not be normal but has a finite variance, when σ^2 is unknown we can approximate its value with s^2 in the computation of the test statistic. The following example is a larage-sample test.

5.2.3. Example

Suppose that 100 high-performance tires made by a certain manufacturer lasted on the average 21,819 miles with a standard deviation of 1,295 miles. Test the null hypothesis $\mu = 22{,}000$ miles against the alternative hypothesis $\mu < 22{,}000$ miles at the 0.05 level of significance.

Solution:

1. $H_0: \mu = 22{,}000$
$H_1: \mu < 22{,}000$
$\alpha = 0.05$

2. Reject the null hypothesis if $z \leq -1.645$

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

3. Substituting $\bar{x} = 21{,}819, \mu_0 = 22{,}000, s = 1.295\ for\ \sigma, and\ n = 100$, we get

$$z = \frac{21{,}819 - 22{,}000}{1{,}295/\sqrt{100}} = -1.40$$

4. Since $z = -1.40$ is greater than -1.645, the null hypothesis cannot be rejected; there is no convincing evidence that the tires are not as good as assumed under the null hypothesis.

5.2.4. One-Sample t test

When $n < 30$ and σ^2 is unknown, for random samples from normal populations, the likelihood ratio techniques yidles a corresponding test based on $t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ which is a value of a random variable having the t distribution with $n - 1$ degrees of freedom. Thus, critifcal regions of size α for testing the null hypothesis $\mu = \mu_0$ against the alternatives $\mu \neq \mu_0, \mu > \mu_0\ or\ \mu < \mu_0$ are, respectively, $|t| \geq t_{\frac{\alpha}{2}, n-1}$, $t \geq t_{\alpha, n-1}$ and $t \leq -t_{\alpha, n-1}$.

5.2.5. Example

The specifications for a certain kind of ribbon call for a mean breaking strength of 185 pounds. If five pieces randomly selected from different rolls have breaking strength of 171.6, 191.8, 178.3, 184.9, and 189.1 pounds, test the null hypothesis $\mu = 185$ pounds against the alternative hypothesis $\mu < 185$ pounds at the 0.05 level of significance.

Solution:

1. $H_0: \mu = 185$
$H_1: \mu < 185$
$\alpha = 0.05$

2. Reject the null hypothesis if $t \leq -2.132$, where 2.132 is the value of $t_{0.05,4}$

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

3.

x	$dx = x - A$ $dx = x - 183$	dx^2
171.6	-11.4	129.96
191.8	8.8	77.44
178.3	-4.7	22.09
184.9	1.9	3.61
189.1	6.1	37.21
$\sum x = 915.7$	$\sum dx = 0.7$	$\sum dx^2 = 270.31$

$\bar{x} = \frac{\sum x}{n} = \frac{915.7}{5} = 183.1$

Standard deviation $s = \sqrt{\frac{\sum dx^2 - \frac{(\sum dx)^2}{n}}{n-1}} = \sqrt{\frac{270.31 - \frac{(0.7)^2}{5}}{4}} = 8.2$

Substituting $\bar{x} = 183.1, \mu_0 = 185, s = 8.2\ for\ \sigma, and\ n = 5$, we get

$$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{183.1 - 185}{8.2/\sqrt{5}} = -0.51$$

4. Since $t = -0.49$ is greater than $-\ 2.132$, the null hypothesis cannot be rejected. If we went beyond this and concluded that the rolls of ribbon from which the sample was selected meet spectifications.

5.3. Tests Concerning Differences Between Means

Let us suppose that we are dealing with independent random samples of sizes n_1 and n_2 from two normal populations having the means μ_1 and μ_2 and the known variances σ_1^2 and σ_2^2 and that we want to test the hypothesis $\mu_1 - \mu_2 = \delta$, where δ is a given constant, against one of the alternatives $\mu_1 - \mu_2 \neq \delta$, $\mu_1 - \mu_2 > \delta\ or\ \mu_1 - \mu_2 < \delta$. Applying the likelihood ratio technique, we will arrive at a test based on $\bar{x}_1 - \bar{x}_2$ and the respective critical regions can be written as $|z| \geq z_{\alpha/2}$, $z \geq z_\alpha$ and $z \leq -z_\alpha$, where

$$z = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

When we deal with independent random samples from populations with unknown variances that may not even be normal, we can still use the test that we have just descirbed with s_1 substituted for σ_1 and s_2 substituted for σ_2 as long as both samples are large enough to invoke the central limit theorem.

5.3.1. Example

An experiment is performed to determine whether the average nicotine content of one kind of cigarette exceeds that of another kind by 0.20 miligram. If $n_1 = 50$ cigarettes of the first kind had an average nicotine content of $\bar{x}_1 = 2.61$ miligrams with a standard deviation of $s_1 = 0.12$ miligram, whereas $n_2 = 40$ cigarettes of the other kind had an average nicotine content of $\bar{x}_2 = 2.38$ miligrams with a standard deviation of $s_2 = 0.14$ miligram, test the null hypothesis $\mu_1 - \mu_2 = 0.20$ against the alternative hypothesis $\mu_1 - \mu_2 \neq 0.20$ at the 0.05 level fo significance. Based the decision on the P-Value corresponding to the value of the appropriate test statistic.

Solution:

1. $H_0 : \mu_1 - \mu_2 = 0.20$
$H_1 : \mu_1 - \mu_2 \neq 0.20$
$\alpha = 0.05$

2. Use the test statistic Z, where $z = \frac{\bar{x}_1-\bar{x}_2-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$

3 Substituting$\bar{x}_1 = 2.61, \bar{x}_1 = 2.61, \delta = 0.20, s_1 = 0.12\ for\ \sigma_1, s_2 = 0.14\ for\ \sigma_2, n_1 = 50\ and$ $n_2 = 40$ into this formula, we get

$$z = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} = \frac{2.61 - 2.38 - 0.20}{\sqrt{\frac{(0.12)^2}{50}+\frac{(0.14)^2}{40}}} = 1.08$$

This corrsponding P-value is $2\ (0.5 - 0.3599)$, where 0.3599 is the entry in the statisical table for z = 1.08.

4. Since 0.2802 exceeds 0.05, the null hypothesis cannot be rejected; we say that the difference between $2.61 - 2.38 = 0.23$ and 0.20 is not significant. This means that the difference may well be atributed to chance.

5.3.2. Two-Sample t test

When n_1 and n_2 are samll and σ_1 and σ_2 are unknown. For independent random samples from two normal populations having the same unknown variance σ^2, the likelihood ratio technique yields a test based on

$t = \frac{\bar{x}_1-\bar{x}_2-\delta}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$Where $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$

Under the given assumptions and the null hypothesis $\mu_1 - \mu_2 = \delta$, this expression for t is a vlaue of a random variable having the t distribution with $n_1 + n_2 - 2$ degrees of freedom. Thus, the appropriate critical regions of size α for testing the null hypothesis $\mu_1 - \mu_2 = \delta$ against the alternatives $\mu_1 - \mu_2 \neq \delta, \mu_1 - \mu_2 > \delta\ or\ \mu_1 - \mu_2 < \delta$ under the given assumptions are, respectively, $|t| \geq t_{\frac{\alpha}{2},n_1+n_2-2}$, $t \geq t_{\alpha,n_1+n_2-2}, and\ t \leq -t_{\alpha,n_1+n_2-2}$.

5.3.3. Example

In the comparison of two kinds of paint, a consumer testing service finds that four 1-gallon cans of the one brand over on the average 546 square feet with a standard deviation of 31 square feet, whereas four 1-gallon cans of another brand cover on the average 492 square feet with a standard deviation of 26 square feet. Assuming that the two populations sampled are normal and have equal variances, test the null hypothesis $\mu_1 - \mu_2 = 0$ at the 0.05 level of significance.

Solution:

1. $H_0 : \mu_1 - \mu_2 = 0$
   $H_1 : \mu_1 - \mu_2 > 0$
   $\alpha = 0.05$

2. Reject the null hypothesis if $t \geq 1.943$, and 1.943 is the value of $t_{0.05,6}$.

3. $s_p = \sqrt{\frac{3\ (31)^2+3(26)^2}{4+4-2}} = 28.609$ and then substituting its value together with $x_1 = 546$, $\bar{x}_2 = 492, \delta = 0, n_1 = n_2 = 4$, we get

$$t = \frac{546 - 492}{28.609\sqrt{\frac{1}{4} + \frac{1}{4}}} = 2.67$$

4. Since $t = 2.67$ exceeds 1.943 the null hypothesis must be rejected; we conclude that on the average the first kind of paint covers a greater area than the second.

5.4. The Concerning Variances

Given a random sample of size n from a normal population, we shall want to the null hypothesis $\sigma^2 = \sigma_0^2$ against one the alternatives $\sigma^2 \neq \sigma_0^2, \sigma^2 > \sigma_0^2$, or $\sigma^2 < \sigma_0^2$, and the likelihood ratio technique leads to a test based on s^2, the value of the sample variance. Based on theorem " If X_1 and X_2 are independent random variables, X_1 has a chi-square distribution with v_1 degrees of freedom and $X_1 + X_2$ has a chi-square distribution with $v > v_1$ degrees of freedom, then X_2 has a chi-square distributon with $v - v_1$ degrees of freedom". Thus, the critical regions for testing the null hypothesis against the two one-sided alternatives as $\chi^2 \geq \chi^2_{\alpha,n-1}$ and $\chi^2 \leq \chi^2_{1-\alpha,n-1}$, where $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$

For the two-sided alterntaive, we rejct the null hypothesis if $\chi^2 \geq \chi^2_{\alpha/2,n-1}$ or $\chi^2 \leq \chi^2_{1-\alpha/2,n-1}$, and the size of all these critical regions is equal to α.

5.4.1. Example

Suppose that the uniformity of the thickness of a part used in a semiconductor is critical and that measurements of the thickness of a random sample of 18 such parats have the variance $s^2 = 0.68$, where the measurements are in thousandths of an inch. The process is considered to be under control if the variation of the thickness is given by a variance not greater than 0.36. Assuming that the measurements constitute a random sample from a normal population, test the null hypothesis $\sigma^2 = 0.36$ against the alternative hypothesis $\sigma^2 > 0.36$ at the 0.05 level of significance.

Solution:

1. $H_0: \sigma^2 = 0.36$
   $H_1: \sigma^2 > 0.36$
   $\alpha = 0.05$

2. Reject the null hypothesis if $\chi^2 \geq 27.587$ and 27.587 is the value of $\chi^2_{0.05,17}$

3. Substituting $s^2 = 0.68, \sigma_0^2 = 0.36$ and $n = 18$ we get

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{17(0.68)}{0.36} = 32.11$$

4. Since $\chi^2 = 32.11$ exceeds 27.587, the null hypothesis must be rejected and the process used in the manufacture of the parts must be adjusted.

5.4.2. Note

In the above example, if $\alpha = 0.01$, the null hypothesis could not have been rejected, since $\chi^2 = 32.11$ does not exceed $\chi^2_{0.01,17} = 33.409$.

5.4.3. Remark

The likelihood ratio statistic for testing the equality of the variances of two normal populations can be expressed in terms of the ratio of the two sample variances. Given indepdendent random samples of sizes n_1 and n_2 form two normal populations with the variances σ_1^2 and σ_2^2, from the theorem " S_1^2 and S_2^2 are the variances of independent random sampels of sizes n_1 and n_2 from normal populations with the variances σ_1^2 and σ_2^2, then $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$ is a random variable having an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom" that corresponding critical regions of size α for testing the null hypothesis $\sigma_1^2 = \sigma_2^2$ against the one-sided alternaive $\sigma_1^2 > \sigma_2^2$ or $\sigma_1^2 < \sigma_2^2$ are respectively.

$$\frac{s_1^2}{s_2^2} \geq f_{\alpha, n_1-1, n_2-1} \quad and \quad \frac{s_2^2}{s_1^2} \geq f_{\alpha, n_1-1, n_1-1}$$

The appropriate critical region for testing the null hypothesis against the two-sided alternative $\sigma_1^2 \neq \sigma_2^2$ is $\frac{s_1^2}{s_2^2} \geq f_{\alpha/2, n_1-1, n_2-1}$ if $s_1^2 \geq s_2^2$ and $\frac{s_2^2}{s_1^2} \geq f_{\alpha/2, n_2-1, n_1-1}$ if $s_1^2 < s_2^2$

5.4.5.. Example

In comparing the variabliity of the tensile strength of two kinds of structural steel, an experiment yielded the following results: $n_1 = 13$, $s_1^2 = 19.2$, $n_2 = 16$ and $s_2^2 = 3.5$, where the units of measurement are 1,000 pounds per square inch. Assuming that the measurements constitute independent random samples from two normal populations, tet the hypothesis $\sigma_1^2 = \sigma_2^2$ against the alternaive $\sigma_1^2 \neq \sigma_2^2$ at the 0.02 level significance.

Solution:

1. $H_0: \sigma_1^2 = \sigma_2^2$
   $H_1: \sigma_1^2 \neq \sigma_2^2$
   $\alpha = 0.02$

2. Since $s_1^2 \geq s_2^2$, reject the null hypothesis if $\frac{s_1^2}{s_2^2} \geq 3.67$, where 3.67 is the value of $f_{0.01,12,15}$

3. Substituting $s_1^2 = 19.2$ and $s_2^2 = 3.5$, we get

$$\frac{s_1^2}{s_2^2} = \frac{19.2}{3.5} = 5.49$$

4. Since $f = 5.49$ exceeds 3.67, the null hypothesis must be rejected; we conculde that the variability of the tensile strength of the two kinds of steel is not the same.

5.5. Test Concerning Proportions

Let's take the most powerful critical region for testing the null hypothesis $\theta = \theta_0$ against the alternative hypothesis $\theta = \theta_1 < \theta_0$, where θ is the parameter of a binomial population, is based on the value of X, the number of "successes" obtained in n trials. When it comes to compositve alternaives, the likelihood ratio technique also yields test based on the observed number of successes. If we want to test the null hypothesis $\theta = \theta_0$ against the one-sided alternative $\theta > \theta_0$, the critical region of size α of the likelihood ratio criterion is $x \geq k_\alpha$ wher k_α is the smallest integer for which $\sum_{y=k_\alpha}^{n} b(y; n, \theta_0) \leq \alpha$ and $b(y; n, \theta_0)$ is the probability of getting y successes in n binomial trials when $\theta = \theta_0$. The size of this critical region is thus as close as possible to α without exceeding it.

The corresponding critical region for testing the null hypothesis $\theta = \theta_0$ against the one-sided alternative $\theta < \theta_0$ is $x \leq k'_{\alpha}$. Where k'_{α} is the largest integer for which $\sum_{y=k_\alpha}^{k_\alpha^F} b(y; n, \theta_0)$ and finally, the critical region for testing the null hypothesis $\theta = \theta_0$ against two-sided alternative $\theta \neq \theta_0$ is $x \geq k_{\alpha/2}$ or $x \leq k'_{\alpha/2}$.

5.5.1. Example

If $x = 4$ of $n = 20$ patients suffered serious side effects from a new medication, test the null hypothesis $\theta = 0.50$ against the alternative hypothesis $\theta \neq 0.50$ at the 0.05 level of significance. Here θ is the true poportion of patients suffering serious side effects from the new medication.

Solution:

1. $H_0: \theta = 0.50$
   $H_1: \theta \neq 0.50$
   $\alpha = 0.05$

2. Use the test statistic X, observed number of successes.

3. x =4, and since $P(X \leq 4) = 0.0059$, the P-value is $2(0.0059) = 0.0118$

4. Since the P-value, 0.0118 is less than 0.05, the null hypothesis is must be rejected; we conclude that $\theta \neq 0.50$.

5.5.2. Remark

For large values of n we can use the normal approximation to the binomial distribution and treat $z = \frac{x-n\theta}{\sqrt{n\theta(1-\theta)}}$ as a value of a random variable haing the standard normal distribution. For large n, we can thus test the null hypothesis $\theta = \theta_0$ against the alternatives $\theta \neq \theta_0$, $\theta > \theta_0$ or $\theta < \theta_0$using, respectively, the critical regions $|z| \geq z_{\alpha/2}$, $z \geq z_{\alpha}$ and $z \leq -z_{\alpha}$, where $z = \frac{x-n\theta 0}{\sqrt{n\theta_0(1-\theta_0)}}$ or $z = \frac{(x \pm \frac{1}{2}) - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}$

If we use the continuity correction. We use the minus sign when x exceeds $n\theta_0$ and the plus sign when x is less than $n\theta_0$.

5.5.3. Example

An oil claims that less than 20 percent of all car owners have not tried its gasoline. Test this claim at the 0.01 level of significance if a random check reveals that 22 of 200 car owners have not tried the oil company’s gasoline.

Solution:

1. $H_0: \theta = 0.20$
   $H_1: \theta < 0.20$
   $\alpha = 0.01$

2. Reject the null hypothesis of $z \leq -2.33$, where (without the continuity correction)

$$z = \frac{x - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}$$

3. Substituting $x = 22, n = 200, and\ \theta_0 = 0.20$ we get

$$z = \frac{22 - 200(0.20)}{\sqrt{200(0.20)(0.80)}} = -3.18$$

4. Since $z = -3.18$ is less than -2.33, the null hypothesis must be rejected; we conclude that, as claimed, less than 20 percent of all car owners have not tried the oil company's gasoline.

5.5.4. Note

If we had used the continuity correction in the above problem, we get

$$z = \frac{(x \pm \frac{1}{2}) - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} = \frac{(22 + 0.5) - 200(0.20)}{\sqrt{200(0.20)(0.80)}} = -3.09$$

Since $z = -3.09$ is less than -2.33, the null hypothesis must be rejected; we conclude that, as claimed, less than 20 percent of all car owners have not tried the oil company's gasoline.

5.6. Tests Concerning Differences among k proportions

Suppose that $x_1, x_2, \dots, x_k$ are observed values of k independent random variables $X_1, X_2, \dots, X_k$ having binomial distributions with the parameters n_1 and θ_1, n_2 and $\theta_2 \dots . n_k$ and θ_k. If n's are sufficiently large, we can approximate the distributions of the independent random variables

$$Z_i = \frac{X_i - n_i\theta_i}{\sqrt{n_i\theta_i(1-\theta_i)}} \text{ for } i = 1, 2, \dots k$$

With standard normal distributions, and , according to the theorem : If $X_1, X_2, \dots, X_n$ are independnet random variables having standard normal distributions, then $Y = \sum_{i=1}^{n} X_i^2$ has the chi-square distribution with $v = n$ degrees of freedom, we have

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - n_i\theta_i)^2}{n_i\theta_i(1-\theta_i)}$$

as a value of a random variable having the chi-square distribution with k degrees of freedom. To test the null hypothesis $\theta_1 = \theta_2 = \cdots . = \theta_k = \theta_0$ (against the alternative that the least one of the $\theta's$ does not equal θ_0), we can thus use the critica region $\chi^2 \geq \chi^2_{\alpha,k}$ where

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - n_i\theta_0)^2}{n_i\theta_0(1-\theta_0)}$$

When θ_0 is not specified, that is, when we are interested only in the null hypothesis $\theta_1 = \theta_2 = \cdots . = \theta_k$, we substitute for θ the pooled estimate

$$\hat{\theta} = \frac{x_1 + x_2 + \cdots . + x_k}{n_1 + n_2 + \cdots . + n_k}$$

and the critical region becomes $\chi^2 \geq \chi^2_{\alpha,k-1}$, where

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - n_i\hat{\theta})^2}{n_i\hat{\theta}(1-\hat{\theta})}$$

The loss of 1 degree of freedom, thati is, the change in the critical region from $\chi^2_{\alpha,k}$ to $\chi^2_{\alpha,k-1}$, is due to the fact that an estimate is substituted for the unkown parameter θ.

Let us now present an alternative formula for the chi-square statistic. If we arrange the data as in the following table, let us refer to its entries as the observed cell frequencies f_{ij}, where the first subscript indicates the row and the second subscript indicates the column of this $k \times 2$ tables

	Successes	Failures
Sample 1	x_1	$n_1 - x_1$
Sample 2	x_2	$n_2 - x_2$
⋮	⋮	⋮
Sample k	x_k	$n_k - x_k$

Under the null hypothesis $\theta_1 = \theta_2 = \cdots . = \theta_k = \theta_0$ the expected cell frequencies for the first column are $n_i\theta_0$ for $i = 1,2, \ldots, k$, and those for the second column are $n_i(1 - \theta_0)$. when θ_0 is not known, we substitute for it, the pooled estimate $\hat{\theta}$, and estimate the expected cell frequencies as $e_{i1} = n_i\hat{\theta}$ and $e_{i2} = n_i(1-\hat{\theta})$ for $i = 1,2,\ldots,k$. The chi-square statistic $\chi^2 = \sum_{i=1}^{k} \frac{(x_i - n_i\hat{\theta})^2}{n_i\hat{\theta}(1-\hat{\theta})}$ can also b written as $\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{2} \frac{(f_{ij}-e_{ij})^2}{e_{ij}}$.

5.6.1. Example

Determine on the basis of the sample data shown in the following table, whether the true proportion of shoppers favoring detergent A over detergent B is the same in all three cities:

	Number favoring detergent A	Number favoring detergent B	
Mumbai	232	168	400
Chennai	260	240	500
Kerala	197	203	400

Use the 0.05 level of significance.

Solution:

1. $H_0: \theta_1 = \theta_2 = \theta_3$
   $H_0: \theta_1, \theta_2, and\ \theta_3$ are not all equal.
   $\alpha = 0.05$

2. Reject the null hypothesis if $\chi^2 \geq 5.991$, where
$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{2} \frac{(f_{ij}-e_{ij})^2}{e_{ij}}$, and 5.991 is the value of $\chi^2_{0.05,2}$.

3. Since the pooled estimate of θ is

$$\hat{\theta} = \frac{232 + 260 + 197}{400 + 500 + 400} = 0.53$$

The expected cell frequencies are

$e_{11} = 400(0.53) = 212,\quad e_{12} = 400(0.47) = 188,\quad e_{21} = 500(0.53) = 265$

$e_{22} = 500(0.47) = 235,\quad e_{31} = 400(0.53) = 212,\quad e_{32} = 400(0.47) = 188$

and substituted into the formula we get

$$\chi^2 = \frac{(232-212)^2}{212} + \frac{(260-265)^2}{265} + \frac{(197-212)^2}{212} + \frac{(168-188)^2}{188} + \frac{(240-235)^2}{235} + \frac{(203-188)^2}{188} = 6.48$$

4. Since $\chi^2 = 6.48$ exceeds 5.991, the null hypothesis must be rejected; That is, the true proportions of shoppers favoring detergent A over detergent B in the three cities are not the same.

5.7. The Analysis of an $r \times c$ Table

5.7.1. Contingency Table

A table having r rows and c columns where each row represents c values of a non-numerical variable and each column represents r values of a different nonnumerical variable is called a contingency table. In such a table, the entries are count data (Positive integers) and both the row and the column total are left to chance. Such a table is assembled for the purpose of testing whether the row variable and the column variable are independent.

We denote the observed frequency for the cell in the i^{th} rwo and the j^{th}column by f_{ij}, the row totals by $f_{i.}$, the column totals by $f_{.j}$, and the grand total, the sum all the cell frequencies, by f, With this notation, we estimate the probabilities $\theta_{i.}$and $\theta_{.j}$ as

$$\hat{\theta}_{i.} = \frac{f_{i.}}{f} \text{ and } \hat{\theta}_{.j} = \frac{f_{.j}}{f}$$

and under the null hypothesis of independenfce we get

$$e_{ij} = \hat{\theta}_{i.} \cdot \hat{\theta}_{.j} \cdot f = \frac{f_{i.}}{f} \cdot \frac{f_{.j}}{f} \cdot f = \frac{f_{i.} \cdot f_{.j}}{f}$$

for the expected frequency for the cell in the i^{th} rwo and the j^{th}column. e_{ij} is obtained by multiplying the total of the row to wich the cell belongs by the total of the coulmn to which it belongs and then dividing by the grand total.

The value of $\chi^2 = \sum_{i=1} \sum_{j=1}^{c} \frac{(f_{ij}-e_{ij})^2}{e_{ij}}$

Reject the null hypothesis if χ^2 exceeds $\chi^2_{\alpha,(r-1)(c-1)}$.

5.7.2. Example

Use the data shown in the following table to test at the 0.01 level of significance whether a person's ability in mathematics is independent of his or her interest in statistics.

		Ability in Mathematics		
		Low	Average	High
Interest in statistics	Low	63	42	15
	Average	58	61	31
	High	14	47	29

Solution:

1. H_0: Ability in mathemtics and interest in statistics are independent.

H_1: Ability in mathematics and interest in statistics are not independent.

$\alpha = 0.01$

2. Reject the null hypothesis if $\chi^2 \geq 13.277$, where $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}$ and 13.277 is the value of $\chi^2_{\alpha,(r-1)(c-1)} = \chi^2_{0.01,(3-1)(3-1)} = \chi^2_{0.01,4}$

3. The expected frequencies for the first row are $\frac{120\times135}{360} = 45$, $\frac{120\times150}{360} = 50$, $\frac{120\times75}{360} = 25$.

The expected frequencies for the second row are $\frac{150\times135}{360} = 56.25$, $\frac{150\times150}{360} = 62.5$, $\frac{150\times75}{360} = 31.25$.

The expected frequencies for the fourth row are $\frac{90\times135}{360} = 33.75$, $\frac{90\times150}{360} = 37.5$, $\frac{90\times75}{360} = 18.75$.

$\chi^2 = \frac{(63-45)^2}{45} + \frac{(42-50)^2}{50} + \frac{(15-25)^2}{25} + \frac{(58-56.25)^2}{56.25} + \frac{(61-62.5)^2}{62.5} + \frac{(31-31.25)^2}{31.25} + \frac{(14-33.75)^2}{33.75} + \frac{(47-37.5)^2}{37.5} + \frac{(29-18.75)^2}{18.75} = 32.14.$

4. Since $\chi^2 = 32.14$ excedds 13.277, the null hypothesis must be rejected; we conclude that there is a relationship between a person's ability in mathematics and his or her interest in statistics.

5.7.3. Goodness of Fit

The goodness-of-fit test considered here applies to situtations in which we want to determine whether a set of data may be looked upon as a random sample from a population having a given distribution.

5.7.4. Example

From the following table, test at the 0.05 level of significance whether the number of errors the compositor makes in setting a galley of type is a random variable having a Poisson distribution.

Number of errors	0	1	2	3	4	5	6	7	8	9
Observed frequencies	18	53	103	107	82	46	18	10	2	1

Solution:

Since the expected frquencies corresponding to eight and nine errors are less than 5, the two classes are combined.

1. H_0: Number of errors is a Poisson random variable.

H_1: Number of errors is not a Poisson random variable.

$\alpha = 0.05$

Number of errors	Observed frequencies f_i	Poisson Probabilites with $\lambda = 3$	Expected frequencies e_i
0	18	0.0498	21.9
1	53	0.1494	65.7
2	103	0.2240	98.6
3	107	0.2240	98.6
4	82	0.1680	73.9
5	46	0.1008	44.4
6	18	0.0504	22.2
7	10	0.0216	9.5
8	2	0.0081	3.6
9	1	0.0038	1.7

2. Reject the null hypothesis if $\chi^2 \geq 14.067$, where $\chi^2 = \sum_{i=1}^{m} \frac{(f_i - e_i)^2}{e_i}$ and 14.067 is the value of $\chi^2_{0.05,7}$.

3.
$$\chi^2 = \frac{(18-21.9)^2}{21.9} + \frac{(53-65.7)^2}{65.7} + \frac{(103-98.6)^2}{98.6} + \frac{(107-98.6)^2}{98.6} + \frac{(82-73.9)^2}{73.9} + \frac{(46-44.4)^2}{44.4} + \frac{(18-22.2)^2}{22.2} + \frac{(10-9.6)^2}{9.5} + \frac{(3-5.3)^2}{5.3} = 6.83.$$

4. Since $\chi^2 = 6.83$ is less than 14.067, the null hypothesis cannot be rejected, the close agreement between the observed and expected frequencies suggest that the Poisson distribution provides a "good fit"

Let Us Sum Up

In this unit, we studied the tests based on normal, t, $\chi^2$and F distributions for testing of mean, variance and proportions and tests for Independence of attributes and Goodness of fit.

Check Your Progress

1. The χ^2 test is one of the simplest and most widely used________test.

2. The range of F-distribution is_______.

3. The range of t-distribution is_______.

4. In a $r \times c$ contingency table, the degrees of freedom is _______.

Glossaries

Level of Significance: The level of significance is the maximum probability of making a type I error.

Two tailed test: When the test of hypothesis is made on the basis of critical region represented by both sides of the standard normal curve.

One tailed test: A test of statistical hypothesis, where the alternative hypothesis is one sided.

Critical value: The value of the sample statistic that defines the region of acceptance and rejection.

Suggested Readings

1. Freund. J.E.," Mathematical Statistics", Prentice Hall of India, Fifth Edition, 2001.

2. Gupta. S.C. and Kapoor. V. K., "Fundamentals of Mathematical Statistics", Sultan Chand & Sons, Eleventh Edition, 2003.

3. Devore. J. L. "Probability and Statistics for Engineers", Brooks/Cole (Cengage Learning), First India Reprint, 2008.

Answers to Check Your Progress

1. Non-parametric test

2. $1 \ to \ \infty$

3. $-\infty \ to \ \infty$

4. $(r-1) \times (c-1)$

BLOCK III: Correlation and Regression

Unit 6 Correlation and Regression Analysis

Unit 7 Partial and Multiple correlation and regression Analysis

Unit – 6

Correlation and Regression Analysis

Structure

Objectives

After Studying this Unit, the student will be able to

- Explain the relationship between two variables and the relationship between the average values of two variables.
- Relationship between correlation analysis and regression analysis.
- Solving problems in correlation and regression analysis.

Overview

In this unit, we will study the concept of correlation and Regression analysis. That is, correlation is the relationship between two variables and regression means relationship between the average values of two variables. Regression is very useful in estimating and predicting the average value of one variable for a given value of the other variable.

6.1. Introduction

The main objective of many statistical investigations is to establish relationships that make it possible to predict one or more variables in terms of others. Thus, studies are made to predict the potential sales of a new product in terms of its price, a patient's weight in terms of the number of weeks he or she has been on a diet, family expenditures on entertainment in terms of family income etc.

If we are given the joint distribution of two random variables X and Y, and X is known to take on the value x, the main objective of bivariate regression is that of determining the conditional mean $\mu_{Y \mid x}$, that is, " the average value of Y for the given value of X. In Problems involving more than two random variables, that is, in multiple regression, we are concerned with quantities such as $\mu_{Z \mid x,y}$, the mean of Z for given values of X and Y, $\mu_{W \mid x,y,z}$, the mean of W for given values of X, Y, Z and so on.

6.1.1. Bivariate Regression (Regression equation)

If $f(x,\ y)$ is the value of the joint density of two random variables X and Y, bivariate regression consists of determining the conditional density of Y, given X = x and then evaluating the integral

$$\mu_{Y \mid x} = E(Y \mid x) = \int_{-\infty}^{\infty} y.\, w(y|x)dy$$

The resulting equation is called the regression equation of Y on X. Alternately, the regression equation of X on Y is given by

$$\mu_{X \mid y} = E(X \mid y) = \int_{-\infty}^{\infty} x.\, f(x|y)dx$$

6.2. Linear Regression

The Linear regression equation is of the form $\mu_{Y \mid x} = \alpha + \beta x$, where α and β are constant, called the regression coefficients.

Let us express the regression coefficients α and β in terms of some of the lower moments of the joint distribution of X and Y, that is, in terms of $E(X) = \mu_1, E(Y) = \mu_2$, $Var(X) = \sigma_1^2, Var(Y) = \sigma_2^2$ and $cov(X, Y) = \sigma_{12}$. Then, also using the correlation coefficient $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$.

6.2.1. Theorem

If the regression of Y on X is linear, then $\mu_{Y\,|\,x} = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)$ and if the regression of X on Y is linear, then $\mu_{X\,|\,y} = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2)$

Proof:

Since $\mu_{Y\,|\,x} = \alpha + \beta x$

$\int y.w(y \mid x)dy = \alpha + \beta x$

and if we multiply the expression on both sides of this equation by $g(x)$, the corresponding value of the marginal density of X, and integrate on x, we obtain

$\int\int y.w(y\,|x)g(x)dy\,dx = \alpha\int g(x)dx + \beta\int x.g(x)dx$

$\mu_2 = \alpha + \beta\mu_1$

Since $w(y\,|x)g(x) = f(x, y)$. If we had multiplied the equation for $\mu_{Y\,|\,x}$ on both sides by x. g(x) before integrating on x, we obtain

$\int\int xy\,.f(x, y)dydx = \alpha\int x.g(x)dx + \beta\int x^2.g(x)dx$

$E(XY) = \alpha\mu_1 + \beta E(X^2)$

Solving $\mu_2 = \alpha + \beta\mu_1$ and $E(XY) = \alpha\mu_1 + \beta E(X^2)$ for α and β and using

$E(XY) = \sigma_{12} + \mu_1\mu_2$ and $E(X^2) = \sigma_1^2 + \mu_1^2$, we get

$\alpha = \mu_2 - \frac{\sigma_{12}}{\sigma_1^2}.\mu_1 = \mu_2 - \rho\frac{\sigma_2}{\sigma_1}.\mu_1$ and $\beta = \frac{\sigma_{12}}{\sigma_1^2} = \rho\frac{\sigma_2}{\sigma_1}$

The linear regression equation of Y on X as $\mu_{Y\,|\,x} = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)$

Similarly we prove the regression equation of X on Y is linear, $\mu_{X\,|\,y} = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2)$

6.2.2. Remark

If the regression equation is linear and $\rho = 0$ then $\mu_{Y|x}$ does not depend on x or $\mu_{X|y}$ does not depend on y. When $\rho = 0$ and hence $\sigma_{12} = 0$, the two random variables X and Y are uncorrelated and we can say that if two random variables are independent, they are also uncorrelated, but if two random variables are uncorrelated, they are not necessarily independent.

6.3. The Method of Least Squares

6.3.1. Least Squares Estimate

If we are given a set of paired data $\{(x_i, y_i); i = 1, 2, \dots, n\}$. The least squares estimates of the regression coefficients in bivariate linear regression are those that make the quantity $q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\hat{\alpha} + \hat{\beta} x_i)]^2$ a minimum with respect to $\hat{\alpha}$ and $\hat{\beta}$.

6.3.2. Theorem

Given the sample data $\{(x_i, y_i); i = 1,2, \dots, n\}$, the coefficients of the least squares line $\hat{y} = \hat{\alpha} + \hat{\beta} x$ are $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ and $\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$.

Proof:

$q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\hat{\alpha} + \hat{\beta} x_i)]^2$ a minimum with respect to $\hat{\alpha}$ and $\hat{\beta}$.

Differentiating partially with respect to $\hat{\alpha}$ and $\hat{\beta}$ we have

$\frac{\partial q}{\partial \hat{\alpha}} = \sum_{i=1}^{n} (-2)[y_i - (\hat{\alpha} + \hat{\beta} x_i)]$ and

$$\frac{\partial q}{\partial \hat{\beta}} = \sum_{i=1}^{n} (-2) x_i [y_i - (\hat{\alpha} + \hat{\beta} x_i)]$$

For the finding the minimum value, $\frac{\partial q}{\partial \hat{\alpha}} = \sum_{i=1}^{n} (-2)[y_i - (\hat{\alpha} + \hat{\beta} x_i)] = 0$ and

$$\frac{\partial q}{\partial \hat{\beta}} = \sum_{i=1}^{n} (-2) x_i [y_i - (\hat{\alpha} + \hat{\beta} x_i)] = 0$$

Therefrore we have the system of normal equations

$$\sum_{i=1}^{n} y_i = \hat{\alpha} n + \hat{\beta} \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\alpha} \sum_{i=1}^{n} x_i + \hat{\beta} \sum_{i=1}^{n} x_i^2$$

Solving this system of equations, we have, the least squares estimate of β is

$$\hat{\beta} = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

Then the least sqaures estimate of α is

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} y_i - \hat{\beta} \sum_{i=1}^{n} x_i}{n}$$

By solving the first of the two normal equations for $\hat{\alpha}$

Therefore $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

Let us consider

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

$$\hat{\beta} = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

6.4. Normal Regression Analysis

When we analyse a set of paired data $\{(x_i, y_i): 1,2, \dots, n\}$ by regression analysis, we look upon the x_i as constants and the y_i as values of corresponding independent random variables. For example, If we want to analyze data on the ages and prices of used cars, treating the ages as known constants and the price as values of random variables, this is a problem of regression analysis.

Assume that the for each fixed x_i the conditional density of the corresponding random variable y_i is the normal density

$$W(y_i|x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{y_i-(\alpha+\beta x_i)}{\sigma}\right]^2} \quad ; -\infty < y_i < \infty$$

Where $\alpha, \beta and \sigma$ are the same for each i. Given a random sample of such paired data, normal regression analysis concerns itself mainly with the estimation of σ and the regression coefficients $\alpha$and β, with tests of hypothesis concerning these three parameters, and the predictions based on the estimated regression equation $\hat{y} = \hat{\alpha} + \hat{\beta}x$, where $\hat{\alpha}$ and $\hat{\beta}$ are estimates of α and β.

6.4.1. To Obtain maximum likelihood estimates of the parameters α, Q$\boldsymbol{and\sigma}$.

Differentiate partially the likelihood function (or its logarithm, which is easier) with respect to $\alpha, \beta and \sigma$, equate the expressions to zero, we get

$$ln\, L = -n.\, ln\, \sigma - \frac{n}{2}.\, ln\, 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)]^2$$

$$\frac{\partial \ln L}{\partial \alpha} = \frac{1}{\sigma^2}\sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)] = 0$$

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i[y_i - (\alpha + \beta x_i)] = 0$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)]^2 = 0$$

Since the first two equations are equivalent to the two normal equations. The maximum likelihood estimates of α and β are identical with the least squares estimate of the above theorem.

If we substitute these estimates of α and β into the equation obtained by $\frac{\partial \ln L}{\partial \sigma}$ to zero, we get the maximum likelihood estimate of σ is

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i - (\hat{\alpha} + \beta x_i)]^2}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}(S_{yy} - \hat{\beta}.S_{xy})}$$

Let us now investigate their use in testing hypotheses concerning α and β and in constructing confidence intervals for these two parameters.

To study the sampling distribution of $\hat{B}$ let us write

$$\hat{B} = \frac{S_{xY}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{S_{xx}} = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_{xx}}\right)Y_i$$

which is a linear combination of the n independent normal random variables Y_i. $\hat{B}$ itself has a normal distribution with the mean

$$E(\hat{B}) = \sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{S_{xx}}\right] E(Y_i|x_i) = \sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{S_{xx}}\right](\alpha + \beta x_i) = \beta$$

and the variance

$$Var(\hat{B}) = \sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{S_{xx}}\right] Var(Y_i|x_i) = \sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{S_{xx}}\right]\sigma^2 = \frac{\sigma^2}{S_{xx}}$$

6.4.2. Result

Under the assumptions of normal regression analysis, $\frac{n\hat{\sigma}^2}{\sigma^2}$ is a value of a random variable having the chi-square distribution with n-2 degree of freedom. Furthermore, this random variable and $\hat{B}$ are independent.

6.4.3. Result

Under the assumptions of normal regression analysis,

$$t = \frac{\left(\frac{\hat{\beta}-\beta}{\alpha/\sqrt{S_{xx}}}\right)}{\sqrt{\frac{n\hat{\sigma}^2}{\sigma^2}/(n-2)}} = \frac{\hat{\beta}-\beta}{\hat{\alpha}}\sqrt{\frac{(n-2)S_{xx}}{n}}$$

is a value of a random variable having the t distribution with $n-2$ of freedom.

6.4.4. Result

Let $\hat{\Sigma}$ be the random variable whose variable are $\hat{\sigma}$ then

$$P\left(-t_{\frac{\alpha}{2},n-2} < \frac{\hat{B}-\beta}{\hat{\Sigma}}\sqrt{\frac{(n-2)S_{xx}}{n}} < t_{\frac{\alpha}{2},n-2}\right) = 1-\alpha$$

By Result 2, we write this as

$$P\left[\hat{B} - t_{\frac{\alpha}{2},n-2}.\Sigma\sqrt{\frac{n}{(n-2)S_{xx}}} < \beta < \hat{B} + t_{\frac{\alpha}{2},n-2}.\Sigma\sqrt{\frac{n}{(n-2)S_{xx}}}\right] = 1-\alpha$$

6.4.5. Result

Under the assumptions of normal regression analysis,

$$\hat{\beta} - t_{\frac{\alpha}{2},n-2}.\hat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}} < \beta < \beta + t_{\frac{\alpha}{2},n-2}.\hat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}}$$

Is a $(1-\alpha)100\%$ confidence interval for the parameter β.

6.5. Normal Correlation Analysis

Assume that the x_i are fixed constants analyzing the set of paired data $\{(x_i, y_i): 1,2, \dots, n\}$, where $x_i's$ and $y_i's$ are values of a random sample from a bivariagte normal population with the parameters $\mu_1, \mu_2, \sigma_1, \sigma_2 and \rho$.

6.5.1. To estimate the parameters $\mu_1, \mu_2, \sigma_1, \sigma_2 and \rho$ by the method of maximum likelihood

we shall have to maximize the likelihood $L = \prod_{i=1}^{n} f(x_i, y_i)$

$\frac{\partial \ln L}{\partial \mu_1}$ and $\frac{\partial \ln L}{\partial \mu_2}$ are equated to zero, we get

$$-\frac{\sum_{i=1}^{n}(x_i-\mu_1)}{\sigma_1^2} + \frac{\rho\sum_{i=1}^{n}(y_i-\mu_2)}{\sigma_1\sigma_2} = 0 \text{ and } -\frac{\rho\sum_{i=1}^{n}(x_i-\mu_1)}{\sigma_1\sigma_2} + \frac{\sum_{i=1}^{n}(y_i-\mu_2)}{\sigma_2^2} = 0$$

Solve these two equations for μ_1 and μ_2, we get the maximum likelihood estimates of these two parameters are $\hat{\mu}_1 = \bar{x}$ and $\hat{\mu}_2 = \bar{y}$ are the respective sample means.

$\frac{\partial \ln L}{\partial \sigma_1}$, $\frac{\partial \ln L}{\partial \sigma_2}$ and $\frac{\partial \ln L}{\partial \rho}$ are equated to zero and substituting $\mu_1 = \bar{x}$ and $\mu_2 = \bar{y}$, we get

$\hat{\sigma}_1 = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$, $\hat{\sigma}_2 = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}}$ and $\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

The estimate $\hat{\rho}$ is called the sample correlation coefficient, is usually denoted by r.

6.5.2. Result

If $\{(x_i, y_i): 1,2, \dots, n\}$ are the values of a random sample from a bivariate population then $r = \frac{S_{xy}}{\sqrt{S_{xx}.S_{yy}}}$

6.6. Examples

6.6.1. Given the two random variables X and Y that have the joint density

$$f(x, y) = \begin{cases} x.e^{-x(1+y)}, & for\ x > 0\ and\ y > 0 \\ 0, & elseshere \end{cases}$$

Find the regression equation of Y on X and sketch the regression curve.

Solution:

Integrating out y, we find that the marginal density of X is given by

$$g(x) = \begin{cases} e^{-x} & for\ x > 0 \\ 0 & elsewhere \end{cases}$$

and hence the conditional density of Y given X = x is given by

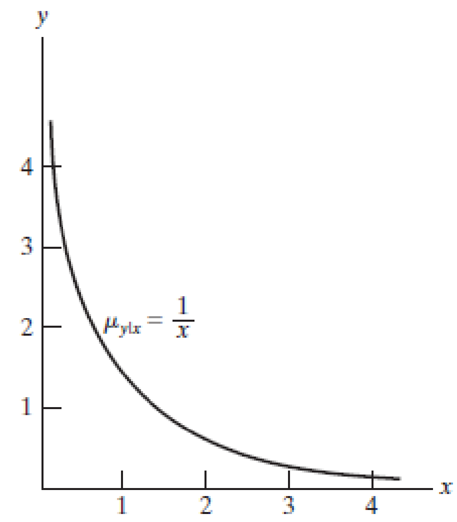
$$w(y\,|x) = \frac{f(x, y)}{g(x)} = \frac{x.\,e^{-x(1+y)}}{e^{-x}} = x.\,e^{-xy}$$

for $y > 0$ and $w(y\,|x) = 0$ elsewhere, which we recognize as an exponential density with $\theta = \frac{1}{x}$. Hence, by

$$\mu_{Y\,|\,x} = \int_0^{\infty} y\,.x.\,e^{-xy}\,dy$$

The mean and the variance of the exponential distribution are given by $\mu = \theta$ and $\sigma^2 = \theta^2$, so that the regression equation Y on X is $\mu_{Y\,|\,x} = \frac{1}{x}$

The corresponding regression curve is shown the following figure



6.6.2. If X and Y have the multinomial distribution

$$f(x, y) = \binom{n}{x, y, n-x-y} . \theta_1^x . \theta_2^y (1 - \theta_1 - \theta_2)^{n-x-y}$$

for $x = 0, 1, 2, \ldots n$, and $y = 0, 1, 2, \ldots n$, with $x + y \leq n$, find the regression equation of Y on X.

Solution:

The Marginal distribution of X is given by

$$g(x) = \sum_{y=0}^{n-x} \binom{n}{x, y, n-x-y} \theta_1^x . \theta_2^y (1 - \theta_1 - \theta_2)^{n-x-y} = \binom{n}{x} \theta_1^x . (1 - \theta_1)^{n-x}$$

for $x = 0, 1, 2, \ldots n$, which we recognize as a binomial distribution with the parameters n and θ_1. Hence,

$$w(y\,|x) = \frac{f(x, y)}{g(x)} = \frac{\binom{n-x}{y} \theta_2^y (1 - \theta_1 - \theta_2)^{n-x-y}}{(1 - \theta_1)^{n-x}}$$

for $y = 0, 1, 2, \ldots n$,

$$w(y\,|x) = \binom{n-x}{y} \left(\frac{\theta_2}{1 - \theta_1}\right)^y \left(\frac{1 - \theta_1 - \theta_2}{1 - \theta_1}\right)^{n-x-y}$$

The conditional distribution of Y given $X = x$ is binomial distribution with parameters $n - x$ and $\frac{\theta_2}{1-\theta_1}$, so that the regression equation of Y on X is $\mu_{Y\,|\,x} = \frac{(n-x)\theta_2}{1-\theta_1}$

Note: In the Previous example, if we let X be the number of times that an even number comes cup in 30 rools of a balanced die and Y be the number of times that the reulst is a 5, then the regression equation becomes

$$\mu_{Y\,|\,x} = \frac{(n - x)\theta_2}{1 - \theta_1} = \frac{(30 - x)\frac{1}{6}}{1 - \frac{1}{2}} = \frac{1}{3}(30 - x)$$

Because there are equally likely possibilities 1, 3 or 5, for each of the $30 - x$ outcomes that are not even.

6.6.3. If the joint density of X_1, X_2 and X_3 is given by

$$f(x_1, x_2, x_3) = \begin{cases} (x_1 + x_2)e^{-x_3}, & for\ 0 < x_1 < 1, \quad 0 < x_2 < 1 \\ 0 & elsewhere \end{cases}$$

Find the regression equtaion of X_2 on X_1 and X_3.

Solution:

The Joint marginal density of X_1 and X_3 is given by

$$m(x_1, x_3) = \begin{cases} (x_1 + \frac{1}{2})\, e^{-x_3}\ for\ 0 < x_1 < 1, x_3 > 0 \\ 0 \qquad elsewhere \end{cases}$$

$$\mu_{X_2|x_1,x_3} = \int_{-\infty}^{\infty} x_2 \cdot \frac{f(x_1, x_2, x_3)}{m(x_1, x_3)}\, dx_2 = \int_0^1 \frac{x_2(x_1 + x_2)}{x_1 + \frac{1}{2}}\, dx_2 = \frac{x_1 + \frac{2}{3}}{2x_1 + 1}$$

6.6.4. Consider the following data on the number of hours that 10 persons studies for a French test and their scores on the test:

Hours studied x	4	9	10	14	4	7	12	22	1	17
Test score y	31	58	65	73	37	44	60	91	21	84

(a) Find the equation of the least squares line that approximates the regression of the test scores on the number of hours studied.

(b) Predict the average test score of persons who studied 14 hours for the test.

Solution:

(a) $n = 10, \sum x = 100, \sum x^2 = 1{,}376, \sum y = 564, \sum xy = 6{,}945, \bar{y} = \frac{\sum y}{n} = \frac{564}{10} = 56.4,$

$$\bar{x} = \frac{\sum x}{n} = \frac{100}{10} = 10$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 1{,}376 - \frac{1}{10}(100)^2 = 376$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = 6{,}945 - \frac{1}{10}(100)(564) = 1{,}305$$

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{1{,}305}{376} = 3.471$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 56.4 - 3.471(10) = 56.4 - 34.71 = 21.69$

Therefore, the equtaion of the least squares is $\hat{y} = 21.69 + 3.471x$

(b) Substituting $x = 14$ into the equation obtained in part (a), we get

$\hat{y} = 21.69 + 3.471(14) = 70.284$

6.6.5. Consider the following data on the number of hours that 10 persons studies for a French test and their scores on the test:

Hours studied x	4	9	10	14	4	7	12	22	1	17
Test score y	31	58	65	73	37	44	60	91	21	84

Test the null hypothesis $\beta = 3$ against the alternative hypothesis $\beta > 3$ at the 0.01 level of significance.

Solution:

1. $H_0: \beta = 3$

$H_1: \beta > 3$

$\alpha = 0.01$

2. Reject the null hypothesis if $t \geq 2.896$, where 2.896 is the value of $t_{0.01,8}$ from the statistical table.

3. Caculate $n = 10, \sum x = 100, \sum x^2 = 1{,}376, \sum y = 564, \sum xy = 6{,}945, \bar{y} = \frac{\sum y}{n} = \frac{564}{10} = 56.4$, $\sum y^2 = 36{,}562$.

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 36{,}562 - \frac{1}{10}(564)^2 = 4{,}752.4$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = 6{,}945 - \frac{1}{10}(100)(564) = 1{,}305$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{1{,}305}{376} = 3.471$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}[S_{yy} - (\hat{\beta})(S_{xy})]} = \sqrt{\frac{1}{10}[4{,}752 - (3{,}471)(1{,}305)]} = 4.720$$

$$t = \frac{\hat{\beta} - \beta}{\hat{\alpha}}\sqrt{\frac{(n-2)S_{xx}}{n}} = \frac{3.471 - 3}{4{,}720}\sqrt{\frac{8.376}{10}} = 1.73$$

since $t = 1.73$ is less than 2.896, the null hypothesis cannot be rejected; we cannot conclude that one the average an extra hour of study will increase the score by more than 3 points.

6.6.6. Consider the following data on the number of hours that 10 persons studies for a French test and their scores on the test:

Hours studied x	4	9	10	14	4	7	12	22	1	17
Test score y	31	58	65	73	37	44	60	91	21	84

Construct a 95% confidence interval for β.

Solution:

$n = 10, \sum x = 100, \sum x^2 = 1{,}376, \sum y = 564, \sum xy = 6{,}945, \bar{y} = \frac{\sum y}{n} = \frac{564}{10} = 56.4,$

$$\bar{x} = \frac{\sum x}{n} = \frac{100}{10} = 10$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 1{,}376 - \frac{1}{10}(100)^2 = 376$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 36{,}562 - \frac{1}{10}(564)^2 = 4{,}752.4$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = 6{,}945 - \frac{1}{10}(100)(564) = 1{,}305$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{1{,}305}{376} = 3.471$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}[S_{yy} - (\hat{\beta})(S_{xy})]} = \sqrt{\frac{1}{10}[4{,}752 - (3{,}471)(1{,}305)]} = 4.720$$

$$t_{0.025,8} = 2.306$$

$$\hat{\beta} - t_{\frac{\alpha}{2},n-2}\,\hat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}} < \beta < \beta + t_{\frac{\alpha}{2},n-2}\,\hat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}}$$

$$3.471 - (2.306)(4.720)\sqrt{\frac{10}{8(376)}} < \beta < 3.471 + (2.306)(4.720)\sqrt{\frac{10}{8(376)}}$$

$$2.84 < \beta < 4.10$$

6.6.7. Suppose that we want to determine on the basis of the following data whether there is a relationship between the time, in minutes, it takes a secretary to compute certain form in the morning and in the late in the late afternoon:

Morning x	8.2	9.6	7	9.4	10.9	7.1	9	6.6	8.4	10.5
Afternoon y	8.7	9.6	6.9	8.5	11.3	7.6	9.2	6.3	8.4	12.3

Compute and interpret the sample correlation coefficient.

Solution:

From the data we get

$n = 10, \sum x = 86.7, \sum x^2 = 771.35, \quad \sum y = 88.8, \sum xy = 792.92, \sum y^2 = 819.34$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 771.35 - \frac{1}{10}(86.7)^2 = 19.661$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 891.34 - \frac{1}{10}(88.8)^2 = 30.796$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = 792.92 - \frac{1}{10}(86.7)(88.8)$$

$S_{xy} = 23.024$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{23.024}{\sqrt{(19.661)(30.796)}} = 0.936$$

Result: The confidence intervals for ρ and tests concerning ρ on the statistic $\frac{1}{2} . ln \frac{1+R}{1-R}$ whose distribution can be approximately normal with mean $\frac{1}{2} ln \frac{1+\rho}{1-\rho}$ and the variance $\frac{1}{n-3}$. Thus,

$$z = \frac{\frac{1}{2} ln \frac{1+r}{1-r} - \frac{1}{2} ln \frac{1+\rho}{1-\rho}}{\frac{1}{\sqrt{n-3}}} = \frac{\sqrt{n-3}}{2} . ln \frac{(1+r)(1-\rho)}{(1-r)(1+\rho)}$$

Using this approximation, we can test the null hypothesis $\rho = \rho_0$ against the alternative hypothesis.

6.6.8. Suppose that we want to determine on the basis of the following data whether there is a relationship between the time, in minutes, it takes a secretary to compute certain form in the morning and in the late in the late afternoon:

Morning x	8.2	9.6	7	9.4	10.9	7.1	9	6.6	8.4	10.5
Afternoon y	8.7	9.6	6.9	8.5	11.3	7.6	9.2	6.3	8.4	12.3

Test the null hypothesis $\rho = 0$ against the alternative hypothesis $\rho \neq 0$ at the 0.01 level of significance.

Solution:

1. $H_0: \rho = 0$
   $H_1: \rho \neq 0$
   $\alpha = 0.01$

2. Reject the null hypothesis if $z \leq -2.575$ or $z \geq 2.575$, where $z = \frac{\sqrt{n-3}}{2} . ln \frac{1+r}{1-r}$

3. Substituting $n = 10$ and $r = 0.936$, we get, $z = \frac{\sqrt{7}}{2} . ln \frac{1.936}{0.064} = 4.5$

4. Since $z = 4.5$ $exceeds$ 2.575, the null hypothesis must be rejected.
We conculde that there is a linear relationship between the time it takes a secretary to complete the form in the morning and in the late afternoon.

Let Us Sum Up

In this unit we discussed the Normal Correlation analysis, linear regression, Method of least squares and normal regression analysis are illustrated with numerical examples.

Check your Progress

1. The Coefficient of correlation lies between_________.

2. The two-regression linear always intersect at their__________.

3. The regression lines become identical if________.

Glossaries

Correlation: The relationship between two variables such that a change in one variable results in corresponding greater or smaller change in the other variable.

Regression: It shows a relationship between the average values of two variables. It is very helpful in estimating and predicting the average value of one variable for a given value of the other variable.

Linear Regression: The relationship between two variables x and y is linear.

Method of Least squares: It is a mathematical device. It is used for obtaining the equation of a curve which fits best to a given set of observations.

Suggested Readings

1. Freund. J.E.," Mathematical Statistics", Prentice Hall of India, Fifth Edition, 2001.

2. Gupta. S.C. and Kapoor. V. K., "Fundamentals of Mathematical Statistics", Sultan Chand & Sons, Eleventh Edition, 2003.

3. Devore. J. L. "Probability and Statistics for Engineers", Brooks/Cole (Cengage Learning), First India Reprint, 2008.

4. Veerarajan. T, "Fundamentals of Mathematical Statistics", Yee Dee Publishing Pvt. Ltd, 2017.

Answers to Check Your Progress

1. $-1 \; and +1$

2. Means

3. The correlation coefficient ± 1

Unit – 7

Partial and Multiple Correlation and Regression Analysis

Structure

Objectives

After Studying this Unit, the student will be able to

- Understand the Yule's notation
- Explain the concept of plane of regression, properties of residuals, coefficient of partial and multiple correlation.
- Demonstrate the problems in partial correlation, multiple correlation and multiple regression.

Overview

In this unit, we will study the concept of Partial and Multiple Correlation and Regression Analysis. That is, in partial correlation, the relationship between dependent variables and one of the independent variables by excluding the effect of other variables and in multiple correlation the effect of all the independent factors on a dependent factor.

7.1. Introduction

Simple correlation that deals with the degree of relationship between two variables, such as heights and weights of individuals, supply and demand of a commodity, ages of husbands and wives and so on. But there are situations when there is interrelation between many variables and the value of one variable may be influenced by other variables. For example, the yield of crop in a year depends upon fertility of soil, amount of rainfall, type of manure used, average temperature and so on. When we are interested in knowing the combined effect group of variables upon a variable not included in that group, we resort to the study of multiple correlation and multiple regression.

The simple correlation between two variables in a group when the influence of other variables in the group has been eliminated from both is called partial correlation. For example, the correlation between the heights and weights of boys of the same age and from families of the same income group is partial correlation. Here the influence of the age factor and the income factor of the family have been eliminated as they are kept constant and so the heights and weights are the variable factors. Even if it is not possible to eliminate the entire influence of variables other than the variables whose partial correlation is measured, we can reduce the influence by easily eliminating the linear effect of those variables. Thus, the simple correlation and regression between two variables in a group, when the linear effect of other variables in the group eliminated, are called partial correlation and partial regression.

7.2. Yule's Subscript notation

We shall study the group of three variables only, through the meanings and arguments will apply to the case of n variables also.

To find the equation of the regression plane x on y and z, we shall assume that
$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \quad (1)$

assuming that the variables have been measured from their respective means. The $b's$ are called partial regression coefficients. In $b_{12.3}$, the first suffix 1 preceding the dot indicates the dependent variable, the second suffix 2 preceding the dot indicates the variables to which the coefficient $b_{12.3}$ is attached and the third suffix 3 succeeding the dot indicates the remaining variable. Similar meanings are attached to $b_{13.2}$. The suffixes preceding the dot are called primary subscripts and those succeeding the dot are called secondary subscripts.

If we consider n variables $x_1, x_2, \dots, x_n$ then the equation of the regression plane $x_1\ on\ x_2, \dots, x_n$ will be assumed as $x_1 = b_{12.24\dots n}x_2 + b_{13.24\dots n}x_3 + \cdots + b_{1n.23\dots(n-1)}x_n$. The order of the primary subscripts cannot be altered, but the secondary subscripts can be written in any order. Note that $b_{12.34} = b_{12.43}; but\ b_{12.34} \neq b_{21.43}$.

The order of any regression coefficient is the number of secondary subscripts in its representation. Thus b_{12} is the regression coefficient of order zero is called simple or total regression coefficient.$b_{12.3}$ is of order 1 and $b_{12.34\dots n}$ is of order $(n-2)$. The quantity $x_{12.3}$ is defined as $x_{12.3} = x_1 - b_{12.3}x_2 - b_{13.2}x_3$ is called the residual of x_1, given by the plane of regression (1) and is said to be of order 2. Residual of x_1 is also called the error of estimate of x_1. The quantity $x_1 - x_{12.3} = b_{12.3}x_2 + b_{13.2}x_3$ is called the estimate of x_1 and it is denoted by $e_{1.23}\ or\ x_{1(23)}$.

7.3. Plane of Regression

Consider a trivariate distribution consisting of three random variablesx_1, x_2, x_3.

Let the equation of the plane of regression of $x_1\ on\ x_2\ and\ x_3\ be$

$$x_1 = a + b_{12.3}x_2 + b_{13.2}x_3 \quad (1)$$

where the variables are assumed to have been, measured from their respective means namely $E[x_1] = 0, E[x_2] = 0, E[x_3] = 0$ (2)

Taking expectations of both sides of (1) and (2) and using (2), we get $a = 0$.

(1) becomes $x_1 = b_{12.3}x_2 + b_{13.2}x_3 \quad (3)$

The constants $b_{12.3}$ and $b_{13.2}$ are determined by the principle of least squares which states that if the (3) is to be the equation of the best fitting regression plane for a given data consisting of N sets of corresponding values x_1, x_2, x_3, the sum of the squares of the residuals should be a minimum.

The best estimates of $b_{12.3}$ and $b_{13.2}$ are obtained by minimizing

$$S = \sum x_{1.23}^2 = \sum (x_1 - b_{12.3}x_2 - b_{13.2}x_3)^2$$

The normal equations for getting the best estimates of $b_{12.3}$ and $b_{13.2}$ are

$\frac{\partial S}{\partial b_{12.3}} = 0$ and $\frac{\partial S}{\partial b_{13.2}} = 0$

$-2\sum x_2(x_1 - b_{12.3}x_2 - b_{13.2}x_3) = 0$ and $-2\sum x_3(x_1 - b_{12.3}x_2 - b_{13.2}x_3) = 0$

$\sum x_1x_2 - b_{12.3}\sum x_2^2 - b_{13.2}\sum x_2x_3 = 0 \quad (4)$ and $\sum x_1x_3 - b_{12.3}\sum x_2x_3 - b_{13.2}\sum x_3^2 = 0 \quad (5)$

Since the variables are measured from their respective means,

$\sum x_1 x_2 = Cov(x_1, x_2) = N r_{12} \sigma_1 \sigma_2$ and $r_{12} = \frac{Cov(x_1, x_2)}{\sigma_1 \sigma_2} = \frac{\left(\frac{\sum x_1 x_2}{N}\right)}{\sigma_1 \sigma_2}$

$\sum x_1 x_3 = Cov(x_1, x_3) = N r_{13} \sigma_1 \sigma_{23}$ and $r_{13} = \frac{Cov(x_1, x_3)}{\sigma_1 \sigma_3} = \frac{\left(\frac{\sum x_1 x_3}{N}\right)}{\sigma_1 \sigma_3}$

Now, σ_i is the standard deviation of x_i.

From (4) we have

$$N r_{12} \sigma_1 \sigma_2 = N b_{12.3} \sigma_2^2 + N b_{13.2} r_{23} \sigma_2 \sigma_3$$

$$r_{12} \sigma_1 = b_{12.3} \sigma_2 + b_{13.2} r_{23} \sigma_3 \quad (6)$$

From (5) we have

$$N r_{13} \sigma_1 \sigma_3 = N b_{12.3} r_{23} \sigma_2 \sigma_3 + N b_{13.2} \sigma_3^2$$

$$r_{13} \sigma_1 = b_{12.3} r_{23} \sigma_2 + b_{13.2} \sigma_3 \quad (7)$$

Solving equations (6) and (7) by Cramer's rule

$$b_{12.3} = \begin{vmatrix} r_{12}\sigma_1 & r_{23}\sigma_3 \\ r_{13}\sigma_1 & \sigma_3 \end{vmatrix} \div \begin{vmatrix} \sigma_2 & r_{23}\sigma_3 \\ r_{23}\sigma_2 & \sigma_3 \end{vmatrix}$$

$$b_{12.3} = \frac{\sigma_1}{\sigma_2} \left\{ \begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix} \div \begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix} \right\} \quad (8)$$

Similarly

$$b_{13.2} = \frac{\sigma_1}{\sigma_3} \left\{ \begin{vmatrix} 1 & r_{12} \\ r_{23} & r_{13} \end{vmatrix} \div \begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix} \right\} \quad (9)$$

Consider the determinant $\Delta = \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix}$

where r_{ii} is the simple correlation between x_i and x_i; $i = 1, 2, 3$

Let the cofactor of r_{ij} in Δ be denoted by R_{ij}.

Using these notations and definitions in (8) and (9) we have

$b_{12.3} = -\frac{\sigma_1}{\sigma_2} \cdot \frac{R_{12}}{R_{11}}$ and $b_{13.2} = -\frac{\sigma_1}{\sigma_3} \cdot \frac{R_{13}}{R_{11}}$ since $r_{ij} = r_{ji}$

Using these values of $b_{12.3}$ and $b_{13.2}$ in (3), the required equation of the regression plane of x_1 on x_2 and x_3 becomes

$$x_1 = \left(-\frac{\sigma_1}{\sigma_2} \cdot \frac{R_{12}}{R_{11}}\right) x_2 + \left(-\frac{\sigma_1}{\sigma_3} \cdot \frac{R_{13}}{R_{11}}\right) x_3$$

$$\frac{x_1}{\sigma_1} R_{11} + \frac{x_2}{\sigma_2} R_{12} + \frac{x_3}{\sigma_3} R_{13} = 0$$

$$\begin{vmatrix} \frac{x_1}{\sigma_1} & \frac{x_2}{\sigma_2} & \frac{x_3}{\sigma_3} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix} = 0$$

7.3.1. Note

1. The equation of the regression plane of x_2 on x_3 and x_1 is $\frac{x_1}{\sigma_1} R_{21} + \frac{x_2}{\sigma_2} R_{22} + \frac{x_3}{\sigma_3} R_{23} = 0$

2. The equation of the regression plane of x_3 on x_1 and x_2 is $\frac{x_1}{\sigma_1} R_{31} + \frac{x_2}{\sigma_2} R_{32} + \frac{x_3}{\sigma_3} R_{33} = 0$

3. If the variables x_1, x_2, x_3 are not measured from their respective means, the equation of the regression plane of x_1 on x_2 and x_3 is $\frac{x_1 - \bar{x}_1}{\sigma_1} R_{11} + \frac{x_2 - \bar{x}_2}{\sigma_2} R_{12} + \frac{x_3 - \bar{x}_3}{\sigma_3} R_{13} = 0$

That is

$$\begin{vmatrix} \frac{x_1 - \bar{x}_1}{\sigma_1} & \frac{x_2 - \bar{x}_2}{\sigma_2} & \frac{x_3 - \bar{x}_3}{\sigma_3} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix} = 0$$

4. If we consider a multivariate distribution consisting of n random variables $x_1, x_2, \ldots, x_n$, the equation of the regression plane of x_1 on $x_2, \ldots, x_n$ will be assumed as $x_1 = b_{12.24\ldots n} x_2 + b_{13.24\ldots n} x_3 + \cdots + b_{1n.23\ldots.(n-1)} x_n$ and it is denoted in determinant notation as

$$\begin{vmatrix} \frac{x_1}{\sigma_1} & \frac{x_2}{\sigma_2} & \frac{x_3}{\sigma_3} & \ldots & \frac{x_n}{\sigma_n} \\ r_{21} & r_{22} & r_{23} & \ldots & r_{2n} \\ r_{31} & r_{32} & r_{33} & \cdots & r_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{n1} & r_{n2} & r_{n3} & \cdots & r_{nn} \end{vmatrix} = 0$$

7.4. Properties of Residuals

7.4.1. The sum of the products of any variable with every residual is zero, provided the subscript of the variable occurs among the secondary subscripts of the residual.

Proof: In the derivation of the equation of the regression plane of x_1 on x_2 and x_3 in a univariate distribution, the normal equations for getting $b_{12.3}$ and $b_{13.2}$ are

$$\sum x_2(x_1 - b_{12.3}x_2 - b_{13.2}x_3) = 0$$

$\sum x_2 . x_{1.23} = 0 \quad (1)$ and

$$\sum x_3(x_1 - b_{12.3}x_2 - b_{13.2}x_3) = 0$$

$$\sum x_3 . x_{1.23} = 0 \quad (2)$$

From (1) and (2), the sum of the products of any variable with every residual is zero.

Similarly, From derivation of the equation of the regression plane of x_2 on x_3 and x_1 , we get $\sum x_1 . x_{2.31} = 0$ and $\sum x_3 . x_{2.31} = 0$

From derivation of the equation of the regression plane of x_3 on x_1 and x_2 , we get $\sum x_1 . x_{3.21} = 0$ and $\sum x_1 . x_{3.21} = 0$

7.4.2. The sum of the products of two residuals is unaltered, if we omit from one of the residuals, any or all the secondary subscripts which are common to those of the other.

Proof: In a trivariate distribution, the residual $x_{1.2}$ is given by $x_{1.2} = x_1 - b_{12}x_2$, where $x_{1.2}$ is got from $x_{1.23}$ by omitting 3 and the residual $x_{1.23} = x_1 - b_{12.3}x_2 - b_{13.2}x_3$

Now,

$\sum x_{1.23} . x_{1.2} = \sum x_{1.23}(x_1 - b_{12}x_2) = \sum x_{1.23}x_1$ since $\sum x_2 . x_{1.23} = 0$ (by above property)

Also

$\sum x_{1.23} . x_{1.23} = \sum x_{1.23}(x_1 - b_{12.3}x_2 - b_{13.2}x_3) = \sum x_{1.23}x_1$ (by above property)

7.4.3. The sum of the products of two residuals is zero, if all the subscripts (primary and secondary) of one residual occur among the secondary subscripts of the order.

Proof:

Consider

$\sum x_{1.2}x_{3.12} = \sum(x_1 - b_{12}x_2)x_{3.12} = \sum x_1 x_{3.12} - b_{12} \sum x_2 x_{3.12} = 0 - b_{12} \times 0 = 0$ (by property 1)

7.4.4. The variance of the residual of a variable given by the plane of regression can be expressed in terms of the variance of the variable itself. That is, $\sigma^2_{1.23} = \frac{\Delta}{R_{11}} \sigma^2_1$, in a trivariate distribution.

Proof:

$$\sigma^2_{1.23} = \frac{1}{N} \sum x^2_{1.23}$$

$$N \sigma^2_{1.23} = \sum x^2_{1.23}$$

$N \sigma^2_{1.23} = \sum x_{1.23} . x_{1.23} = \sum x_1 x_{1.23}$, by property 2

$$N \sigma^2_{1.23} = \sum x_1(x_1 - b_{12.3}x_2 - b_{13.2}x_3)$$

$$N \sigma^2_{1.23} = \sum x^2_1 - b_{12.3} \sum x_1 x_2 - b_{13.2} \sum x_1 x_3$$

$$N \sigma^2_{1.23} = N\sigma^2_1 - b_{12.3}Nr_{12}\sigma_1\sigma_2 - b_{13.2}Nr_{13}\sigma_1\sigma_3$$

$$\sigma^2_{1.23} = \sigma^2_1 - b_{12.3}r_{12}\sigma_1\sigma_2 - b_{13.2}r_{13}\sigma_3$$

$$\sigma^2_{1.23} = \sigma^2_1 + \frac{R_{12}\sigma_1}{R_{11}\sigma_2} r_{12}\sigma_1\sigma_2 + \frac{R_{13}\sigma_1}{R_{11}\sigma_3} r_{13}\sigma_3$$

$$\sigma^2_{1.23} = \frac{\sigma_1^2}{R_{11}}(r_{11}R_{11} + r_{12}R_{12} + r_{13}R_{13}) = \frac{\Delta}{R_{11}}\sigma_1^2$$

7.4.5. Note

1. $\sigma^2_{i,jk} = \frac{\Delta}{R_{ii}}\sigma^2_i$

2. $\sigma^2_{1.23\ldots n} = \frac{\Delta}{R_{11}}\sigma^2_1$ where $\Delta = \begin{vmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & r_{22} & r_{23} & \cdots & r_{2n} \\ r_{31} & r_{32} & r_{33} & \cdots & r_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & r_{n3} & \cdots & r_{nn} \end{vmatrix}$

7.5. Coefficient of multiple correlation

The simple correlation coefficient between x_1 and the totality of all the other variables $x_2, x_3, \ldots, x_n$ is called the coefficient of multiple correlation between x_1 and $(x_2, x_3, \ldots, x_n)$ and it is denoted by $R_{1.23\ldots n}$ or $R_{1(23\ldots n)}$.

The simple correlation between x_1 and the estimate of x_1 in terms of x_2 and x_3 namely, $e_{1.23}$ is $R_{1.23}$ for a trivariate distribution.

7.5.1. Multiple correlation coefficient in terms of simple correlation coefficients

In a trivariate distribution

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{23}r_{31}}{1 - r^2_{23}}$$

$$Cov(x_1, e_{1.23}) = Cov(x_1,\ x_1 - x_{1.23})$$

$$Cov(x_1, e_{1.23}) = Cov(x_1, x_1) - Cov(x_1, x_{1.23})$$

$Cov(x_1, e_{1.23}) = \sigma_1^2 - E(x_1.x_{1.23})$ since $E(x_1) = E(x_{1.23}) = 0$

$Cov(x_1, e_{1.23}) = \sigma_1^2 - E(x_{1.23}.x_{1.23})$ by property 2 of residuals

$$Cov(x_1, e_{1.23}) = \sigma^2_1 - \sigma^2_{1.23} \quad (1)$$

$$Var(e_{1.23}) = Var(x_1 - x_{1.23})$$

$$Var(e_{1.23}) = Var(x_1) + Var(x_{1.23}) - Cov(x_1, x_{1.23})$$

$Var(e_{1.23}) = \sigma^2_1 + \sigma^2_{1.23} - 2\sigma^2_{1.23}$ by (1)

$$Var(e_{1.23}) = \sigma^2_1 - \sigma^2_{1.23} \quad (2)$$

Now

$$R_{1.23} = \frac{Cov(x_1, e_{1.23})}{\sqrt{Var(x_1).Var(e_{1.23})}}$$

$$R_{1.23} = \frac{\sigma_1^2 - \sigma_{1.23}^2}{\sigma_1\sqrt{\sigma_1^2 - \sigma_{1.23}}}$$ by (1) & (2)

$$R_{1.23} = \frac{\sqrt{\sigma_1^2 - \sigma_{1.23}}}{\sigma_1}$$

$$R_{1.23} = \sqrt{1 - (\frac{\sigma_{1.23}}{\sigma_1})^2}$$

$$R_{1.23}^2 = 1 - (\frac{\sigma_{1.23}}{\sigma_1})^2$$

$$1 - R_{1.23}^2 = (\frac{\sigma_{1.23}}{\sigma_1})^2 = \frac{\Delta}{R_{11}}$$

Now,

$$\Delta = \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix}$$

$$\Delta = 1 - r_{23}^2 - r_{12}(r_{12} - r_{23}r_{31}) + r_{13}(r_{12}.r_{23} - r_{31})$$

$$\Delta = 1 - r_{12}^2 - r_{23}^2 - r_{31}^2 + 2r_{12}r_{23}r_{31}$$

and $R_{11} = 1 - r_{23}^2$

Therefore,

$$1 - R_{1.23}^2 = \frac{\Delta}{R_{11}} = \frac{1 - r_{12}^2 - r_{23}^2 - r_{31}^2 + 2r_{12}r_{23}r_{31}}{1 - r_{23}^2}$$

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}^2}$$

7.5.2. Note

1. $R_{i.jk}^2 = \frac{r_{ij}^2 + r_{ik}^2 - 2r_{ij}r_{jk}r_{ik}}{1 - r_{jk}^2}$

2. $\sigma_{1.23}^2 = \sigma_1^2(1 - R_{1.23}^2)$

3. For a n-variate distribution,

$R_{1.23...n}^2 = 1 - (\frac{\sigma_{1.23...n}}{\sigma_1})^2 = 1 - \frac{\sigma_{1.23...n}^2}{\sigma_1^2}$ and $\sigma_{1.23...n}^2 = \sigma_1^2(1 - R_{1.23...n}^2)$

4. Since $Var(e_{1.23}) = Cov(x_1, e_{1.23})$ From (1) & (2), $Cov(x_1, e_{1.23}) \geq 0$ and hence $R_{1.23} \geq 0$.

$$0 \leq R_{1.23} \leq 1.$$

5. If $R_{1.23} = 1$, then $\sigma^2_{1.23} = \frac{1}{N}\sum x^2_{1.23} = 0$

That is, all the regression residuals are zero and hence $x_1 = b_{12.3}x_2 + b_{13.2}x_3$.
The equation of the regression plane may be treated as a perfect prediction formula for $\boldsymbol{x_1}$.

7.6. Partial correlation coefficient in terms of simple correlation coefficients

In the case of trivariate distribution, the correlation coefficient between x_1 and x_2 after the linear effect of x_3 on them has been eliminated is the partial correlation coefficient between x_1 and x_2 and it is denoted by $r_{12.3}$.

$x_{1.3} = x_1 - b_{13}x_3$ and $x_{2.3} = x_2 - b_{23}x_3$ are the residuals that may be regarded as the parts of the variables x_1 and x_2 that remain after the linear effects of x_3 on them have been eliminated.

Therefore,

$$r_{12.3} = \frac{cov(x_{1.3}, x_{2.3})}{\sqrt{Var(x_{1.3}).\,Var(x_{2.3})}} \quad (1)$$

Now,

$$cov(x_{1.3}, x_{2.3}) = Cov\{(x_1 - b_{13}x_3), (x_2 - b_{23}x_3)\}$$

$$cov(x_{1.3}, x_{2.3}) = Cov(x_1, x_2) - b_{23}Cov(x_1, x_3) - b_{13}Cov(x_2, x_3) + b_{13}b_{23}Cov(x_1, x_2)$$

$$cov(x_{1.3}, x_{2.3}) = r_{12}\sigma_1\sigma_2 - r_{23}\frac{\sigma_2}{\sigma_3}r_{13}\sigma_1\sigma_3 - r_{13}\frac{\sigma_1}{\sigma_3}r_{23}\sigma_2\sigma_3 + r_{13}\frac{\sigma_1}{\sigma_3}r_{23}\frac{\sigma_2}{\sigma_3}\sigma_3^2$$

$$cov(x_{1.3}, x_{2.3}) = \sigma_1\sigma_2(r_{12} - r_{13}r_{23}) \quad (2) \text{ since } cov(x_3, x_3) = Var(x_3)$$

$$Var(x_{1.3}) = Var(x_1 - b_{13}x_3)$$

$$Var(x_{1.3}) = Var(x_1) + b_{13}^2Var(x_3) - 2b_{13}Cov(x_1, x_3)$$

$$Var(x_{1.3}) = \sigma_1^2 + r_{13}^2\frac{\sigma_1^2}{\sigma_3^2}.\sigma_3^2 - 2r_{13}\frac{\sigma_1}{\sigma_3}r_{13}\sigma_1\sigma_3$$

$$Var(x_{1.3}) = \sigma_1^2(1 - r_{13}^2) \quad (3)$$

$$Var(x_{2.3}) = \sigma_2^2(1 - r_{23}^2) \quad (4)$$

Using (2), (3) and (4) in (1), we get

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Also $r_{12.3} = \frac{-R_{12}}{\sqrt{R_{11}R_{22}}}$

7.6.1. Note

$$r_{ij.k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}, where\ i, j, k = 1, 2, 3\ and\ \ i \neq j \neq k$$

7.7. Examples

7.7.1. A teacher wished to find the relationship marks in the final examination to those in the two class tests during the semester. Denoting the marks of a student in the first two test and the final examination by x_1, x_2, x_3 respectively, he obtained the following information

$\bar{x}_1 = 6.8, \bar{x}_2 = 7, \bar{x}_3 = 7.3, \sigma_1 = 1, \sigma_2 = 0.8, \sigma_3 = 0.9, r_{12} = 0.6, r_{13} = 0.7, r_{23} = 0.65$

(i) Find the least square regression equation x_3 on x_1 and x_2
(ii) Estimate the marks in the final examination of two students who secured respectively 9 and 7, 4 and 8 in the two tests.
(iii) Compute $R_{3.12}$
(iv) Compute $r_{12.3}$

Solution:

(i) Equation of the regression plane of x_3 on x_1 and x_2 is given by

$$\begin{vmatrix} \frac{x_1 - \bar{x}_1}{\sigma_1} & \frac{x_2 - \bar{x}_2}{\sigma_2} & \frac{x_3 - \bar{x}_3}{\sigma_3} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix} = 0$$

$$\begin{vmatrix} \frac{x_1 - 6.8}{1} & \frac{x_2 - 7}{0.8} & \frac{x_3 - 7.3}{0.9} \\ 1 & 0.6 & 0.7 \\ 0.6 & 1 & 0.65 \end{vmatrix} = 0$$

$$\frac{(x_1 - 6.8)}{1}(-0.31) - \frac{(x_2 - 7)}{0.8}(0.23) + \frac{(x_3 - 7.3)}{0.9}(0.64) = 0$$

$0.711x_3 - 0.288x_2 - 0.310x_1 - 1.071 = 0$

$x_3 = 0.436x_1 = 0.402x_2 + 1.506$

(ii) When $x_1 = 9, x_2 = 7$ then $x_3 = 8.244$

When $x_1 = 4, x_2 = 8$ then $x_3 = 6.466$

(iii)

$$R^2_{3.12} = \frac{r_{31}^2 + r_{32}^2 + 2r_{12}r_{23}r_{31}}{1 - r_{12}^2} = \frac{(0.7)^2 + (0.65)^2 - 2 \times 0.6 \times 0.7 \times 0.65}{1 - (0.6)^2} = 0.5727$$

$R_{3.12} = 0.757$

(iv)

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

$$r_{12.3} = \frac{0.6 - (0.7 \times 0.65)}{\sqrt{(1 - 0.7^2)(1 - 0.65^2)}} = 0.263$$

7.7.2. Prove that the necessary and sufficient condition for the three regression planes to coincide is $r_{12}^2 + r_{23}^2 + r_{31}^2 - 2r_{12}r_{23}r_{31} = 1$

Solution:

The equations of the three regression planes are

$$\frac{x_1}{\sigma_1}R_{11} + \frac{x_2}{\sigma_2}R_{12} + \frac{x_3}{\sigma_3}R_{13} = 0 \quad (1)$$

$$\frac{x_1}{\sigma_1}R_{21} + \frac{x_2}{\sigma_2}R_{22} + \frac{x_3}{\sigma_3}R_{23} = 0 \quad (2)$$

$$\frac{x_1}{\sigma_1}R_{31} + \frac{x_2}{\sigma_2}R_{32} + \frac{x_3}{\sigma_3}R_{33} = 0 \quad (3)$$

Planes (1) and (2) coincide, if and only if the corresponding coefficients are proportional. Namely, if

$$\frac{R_{11}}{R_{21}} = \frac{R_{12}}{R_{22}} = \frac{R_{13}}{R_{23}} \quad (4)$$

Taking the first two rations, the required condition is

$$R_{11}R_{22} - R_{12}R_{21} = 0$$

$$(1 - r_{23}^2)(1 - r_{31}^2) - (r_{12} - r_{23}r_{13})^2 = 0$$

$$(1 - r_{23}^2 - r_{31}^2 + r_{23}^2 r_{31}^2) - (r_{12}^2 + r_{23}^2 r_{31}^2 - 2r_{12}r_{23}r_{31}) = 0$$

$$r_{12}^2 + r_{23}^2 + r_{31}^2 - 2r_{12}r_{23}r_{31} = 1 \quad (5)$$

Taking the second and third ratios in (4), we will get the same condition (5) as the required condition.

Now the planes (2) and (3) will coincide. If

$$\frac{R_{21}}{R_{31}} = \frac{R_{22}}{R_{32}} = \frac{R_{23}}{R_{33}} \quad (6)$$

Proceeding as before, (6) will also reduce to the same condition as (5)
Therefore, the necessary and sufficient condition required is given by (5).

7.7.3. For a trivariate distribution, express the multiple correlation coefficient in terms of simple and partial correlation coefficients.

or

Prove that $1 - R^2_{1.23} = (1 - r^2_{12})(1 - r^2_{13.2})$. Hence deduce that the multiple correlation coefficient is not less than any simple correlation or any partial correlation coefficient.

Solution:

If we express $R_{1.23}$ and $r_{13.2}$is terms of simple correlation coefficients, we have

$1 - R^2_{1.23} = \frac{\Delta}{R_{11}}$ (1) and $r_{13.2} = \frac{-R_{13}}{\sqrt{R_{11} R_{33}}}$

$$1 - r^2_{13.2} = 1 - \frac{R^2_{13}}{R_{11}R_{33}}$$

$$1 - r^2_{13.2} = \frac{R_{11}R_{33} - R^2_{13}}{R_{11}R_{33}} \quad (2)$$

Dividing (1) by (2) we get

$\frac{1-R^2_{1.23}}{1-r^2_{13.2}} = \frac{\Delta R_{33}}{R_{11}R_{33}-R^2_{13}}$ where $\Delta = \begin{vmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{vmatrix}$ and R's are the cofactors of the corresponding r's

$$\frac{1 - R^2_{1.23}}{1 - r^2_{13.2}} = \frac{(1 - r^2_{12} - r^2_{23} - r^2_{31} + 2r_{12}r_{23}r_{31})(1 - r^2_{12})}{(1 - r^2_{23})(1 - r^2_{12}) - (r_{12}r_{23} - r_{31})^2}$$

$$\frac{1 - R^2_{1.23}}{1 - r^2_{13.2}} = \frac{(1 - r^2_{12} - r^2_{23} - r^2_{31} + 2r_{12}r_{23}r_{31})(1 - r^2_{12})}{(1 - r^2_{12} - r^2_{23} - r^2_{31} + 2r_{12}r_{23}r_{31})}$$

$$\frac{1 - R^2_{1.23}}{1 - r^2_{13.2}} = 1 - r^2_{12}$$

$$1 - R^2_{1.23} = (1 - r^2_{12})(1 - r^2_{13.2}) \quad (3)$$

Since

$$0 \le r^2_{12} \le 1, \quad 0 \le 1 - r^2_{12} \le 1 \; and \; 0 \le r^2_{13.2} \le 1, 0 \le 1 - r^2_{13.2} \le 1 \quad (4)$$

From (3) & (4) we get

$1 - R^2_{1.23} \le 1 - r^2_{12} \Rightarrow R^2_{1.23} \ge r^2_{12}$ (5) and $1 - R^2_{1.23} \le 1 - r^2_{13.2} \Rightarrow R^2_{1.23} \ge r^2_{13.2}$ (6)

From (5) & (6) the required result holds.

Note: $1 - R^2_{1.23} = (1 - r^2_{13})(1 - r^2_{12.3})$

7.7.4. If $r_{23} = 0$ then prove that $R^2_{1.23} = r^2_{12} + r^2_{13}$ and deduce that if $R_{1.23} = 0$ then prove that $r_{12} = r_{13} = 0$

Solution:

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} + 2r_{12}r_{23}r_{31}}{1 - r^2_{23}}$$

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13}}{1}$$

$$R^2_{1.23} = r^2_{12} + r^2_{13}$$

$$0 = r^2_{12} + r^2_{13}$$

$$r^2_{12} + r^2_{13} = 0$$

Therefore, $r_{12} = 0 \; and \; r_{13} = 0$

7.7.5. If $r_{12} = r_{23} = r_{31} = p$ then prove that $r_{12.3} = r_{23.1} = r_{31.2} = \frac{p}{p+1}$ and also prove that $1 - R^2_{1.23} = \dfrac{(1-p)(1+2p)}{1+p}$

Solution:

$$r_{ij.k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{(1 - r^2_{ik})(1 - r^2_{jk})}}, where \; i,j,k = 1,2,3 \; and \;\; i \neq j \neq k$$

$$r_{ij.k} = \frac{p - p^2}{\sqrt{(1-p^2)^2}} = \frac{p(1-p)}{(1-p)(1+p)} = \frac{p}{1+p}$$

$$1 - R^2_{1.23} = (1 - r^2_{12})(1 - r^2_{13.2})$$

$$1 - R^2_{1.23} = (1 - p^2)\{1 - \frac{p^2}{(1+P)^2}\}$$

$$1 - R^2_{1.23} = \frac{(1-p^2)(1+2p)}{(1+P)^2}$$

$$1 - R^2_{1.23} = \frac{(1-p)(1+2p)}{1+p}$$

7.7.6. If $r_{23} = 1$ then show that (i) $R^2_{1.23} = r^2_{12} = r^2_{13}$ and (ii) $\sigma^2_{1.23} = \sigma^2_1(1 - r^2_{12})$

Solution:

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{23}r_{31}}{1 - r^2_{23}} \quad (1)$$

$$R^2_{1.23}(1 - r^2_{23}) = r^2_{12} + r^2_{13} - 2r_{12}r_{23}r_{31}$$

When $r_{23} = 1$

$$r^2_{12} + r^2_{13} - 2r_{12}r_{31} = 0$$

$(r_{12} - r_{13})^2 = 0$

$r_{12} = r_{13} \quad (2)$

Using (2) In (1) we get

$$R^2_{1.23} = \frac{2r^2_{12} - 2r^2_{12} r_{23}}{1 - r^2_{23}}$$

$$R^2_{1.23} = \frac{2r^2_{12}(1 - r_{23})}{1 - r^2_{23}}$$

$$R^2_{1.23} = \frac{r^2_{12}}{(1 + r_{23})} = r^2_{12}$$

$R^2_{1.23} = r^2_{12} = r^2_{13}$ By (2)

Now

$$\sigma^2_{1.23} = \sigma^2_1(1 - R^2_{1.23}) = \sigma^2_1(1 - r^2_{12}) = \sigma^2_1(1 - r^2_{13})$$

7.7.7. If r_{12} and r_{13} are given, then show that r_{23} must lie in the range $r_{12}r_{13} \pm (1 - r^2_{12} - r^2_{13} + r^2_{12} r^2_{13})^{1/2}$. Hence Prove that r_{23} lies between -1 and $1 - 2k^2$, if $r_{12} = k$ and $r_{13} = -k$.

Solution:

Since $r_{12.3}$ is a partial correlation coefficient

$0 \le r^2_{12.3} \le 1$

$r_{12.3} \le 1$

$$\frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r^2_{13})(1 - r^2_{23})}} \le 1$$

$$r_{12} - r_{13}r_{23} \le \sqrt{(1 - r^2_{13})(1 - r^2_{23})}$$

$(r_{12} - r_{13}r_{23})^2 \le (1 - r^2_{13})(1 - r^2_{23})$

$r^2_{12} - 2r_{12}r_{23}r_{31} + r^2_{13} r^2_{23} \le (1 - r^2_{13})(1 - r^2_{23})$

$r^2_{12} - 2r_{12}r_{23}r_{31} + r^2_{13} r^2_{23} \le 1 - r^2_{23} - r^2_{13} + r^2_{13} r^2_{23}$

$r^2_{23} - 2r_{12}r_{23}r_{31} + r^2_{12} r^2_{13} \le 1 - r^2_{12} - r^2_{13} + r^2_{12} r^2_{13}$

$(r_{23} - r_{12}r_{13})^2 \le 1 - r^2_{12} - r^2_{13} + r^2_{12} r^2_{13}$

$|r_{23} - r_{12}r_{13}| \le \sqrt{1 - r^2_{12} - r^2_{13} + r^2_{12} r^2_{13}}$

Therefore, r_{23} lie in the range $r_{12}r_{13} \pm (1 - r^2_{12} - r^2_{13} + r^2_{12} r^2_{13})^{1/2}$ (1)

Put $r_{12} = k$ and $r_{13} = -k$ in (1), we get that r_{23} lies in the range $-k^2 \pm \sqrt{1 - 2k^2 + k^4}$

$-k^2 \pm (1 - k^2)$

$-k^2 - (1 - k^2) \leq r_{23} \leq -k^2 + (1 - k^2)$

$-1 \leq r_{23} \leq 1 - 2k^2$

7.7.8. If the variables x_1, x_2, x_3 are measured from the respective means and have the same variance. Prove that (i) $r_{12} + r_{23} + r_{31} \geq -\frac{3}{2}$ (ii) $r^2_{12} + r^2_{23} + r^2_{31} \leq 1 + 2r_{12}r_{23}r_{31}$

Solution:

$E[x_1 + x_2 + x_3]^2 = E[x^2_1 + x^2_2 + x^2_3 + 2(x_1x_2 + x_2x_3 + x_3x_1)]$ (1)

Since the variables are measured from the respective means and have variance.

$E[x^2_i] = \sigma^2$ for $i = 1, 2, 3$

$$r_{ij} = \frac{Cov(x_i, x_j)}{\sigma_{x_i}\sigma_{x_j}}$$

$$r_{ij} = \frac{E[x_ix_j]}{\sigma^2} for\ i, j = 1, 2, 3 and\ i \neq j$$

Consider

$E[x_1 + x_2 + x_3]^2 \geq 0$

$3\sigma^2 + 2\sigma^2(r_{12} + r_{23} + r_{31}) \geq 0$

$$r_{12} + r_{23} + r_{31} \geq -\frac{3}{2}$$

Since $R^2_{1.23} \leq 1$

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{23}r_{31}}{1 - r^2_{23}} \leq 1$$

$r^2_{12} + r^2_{13} - 2r_{12}r_{23}r_{31} \leq 1 - r^2_{23}$

$r^2_{12} + r^2_{23} + r^2_{31} \leq 1 + 2r_{12}r_{23}r_{31}$

7.7.9. If $x_1 = ax_2 + bx_3$ then prove that the three partial correlations are numerically equal to unity. Also show that $r_{1.23}$ has got the same sign of a, $r_{13.2}$ has got the same sign as b and $r_{23.1}$ has the opposite sign of $(\frac{a}{b})$.

Solution:

Since given $x_1 = ax_2 + bx_3$, assume that x_2 and x_3 are independent variables and x_1 is the dependent variable, depending on x_2 and x_3. Therefore $r_{23} = 0$ and hence

$$\frac{Cov(x_2, x_3)}{\sigma_2\sigma_3} = 0$$

$$Cov(x_2, x_3) = 0$$

$$Var(x_1) = Var(ax_2 + bx_3)$$

$$Var(x_1) = a^2Var(x_2) + b^2Var(x_3)$$

$$Var(x_1) = a^2\sigma_2^2 + b^2\sigma_3^2$$

$$Cov(x_1, x_2) = Cov(ax_2 + bx_3, x_2)$$

Therefore

$$r_{12} = \frac{[a\sigma_2^2 + bCov(x_2, x_3)]}{\sqrt{Var(ax_2 + bx_3)Var(x_2)}}$$

$$r_{12} = \frac{a\sigma_2}{\sqrt{a^2\sigma_2^2 + b^2\sigma_3^2}}$$

Similarly

$$r_{13} = \frac{b\sigma_3}{\sqrt{a^2\sigma_2^2 + b^2\sigma_3^2}}$$

Now,

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = \frac{\left(\frac{a\sigma_2}{k}\right)}{\sqrt{\frac{k^2 - b^2\sigma_3^2}{k^2}}} = \frac{a\sigma_2}{a\sigma_2} = \pm 1$$

Where $k = \sqrt{a^2\sigma_2^2 + b^2\sigma_3^2}$

Since $\sqrt{a^2\sigma_2^2} = \pm\, a\ \sigma_2$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} = \frac{\left(\frac{b\sigma_3}{k}\right)}{\sqrt{\frac{k^2 - a^2\sigma_2^2}{k^2}}} = \frac{b\sigma_3}{\sqrt{b^2\sigma_3^2}} = \pm 1$$

Since $\sqrt{b^2\sigma_3^2} = \pm\, b\ \sigma_3$

Therefore, $r_{12.3}$ has the same sign as a and $r_{13.2}$ has the same sign as b and they are numerically equal to unity.

Now,

$$r_{23.1} = \frac{r_{23} - r_{12}r_{31}}{\sqrt{(1 - r_{12}^2)(1 - r_{31}^2)}}$$

$$r_{23.1} = \frac{\frac{-a\sigma_2}{k}\frac{b\sigma_3}{k}k^2}{\sqrt{(k^2 - a^2\sigma_2^2)(k^2 - b^2\sigma_3^2)}}$$

$$r_{23.1} = \frac{-ab\sigma_2\sigma_3}{\sqrt{(a^2\sigma_2^2)(b^2\sigma_3^2)}}$$

$$r_{23.1} = \frac{-ab}{\sqrt{a^2b^2}}$$

$$r_{23.1} = \frac{-(\frac{a}{b})}{\sqrt{\frac{a^2}{b^2}}}$$

$$r_{23.1} = \frac{-(\frac{a}{b})}{\pm(\frac{a}{b})} = \mp 1$$

$r_{23.1}$ has opposite sign of $(\frac{a}{b})$ and its numerical value is 1.

Let Us Sum Up

In this unit, we explained the concept and the differences between simple, partial and multiple correlation analysis with examples and also discussed plane of regression and Properties of residuals.

Check your Progress

1. The partial correlation coefficient lies between________.

2. Multiple correlation coefficient is a_______coefficient.

3. If $R_{12.3} = 0$ then $r_{12} =$____________and $r_{13} =$____________.

4. In Multiple regression analysis, the independent variable is a random variable whereas the independent variables___________random variables.

Glossaries

Partial Correlation: It is the measure of association between two variables, while controlling or adjusting the effect of one or more additional variables.

Multiple Correlation: It is a statistical technique that predicts values of one variable on the basis of two or more other variables.

Multiple Regression: It's statistical technique that can be to analyse the relationship between a single dependent variable and several independent variables.

Suggested Readings

1. Freund. J.E.," Mathematical Statistics", Prentice Hall of India, Fifth Edition, 2001.

2. Gupta. S.C. and Kapoor. V. K., "Fundamentals of Mathematical Statistics", Sultan Chand & Sons, Eleventh Edition, 2003.

3. Devore. J. L. "Probability and Statistics for Engineers", Brooks/Cole (Cengage Learning), First India Reprint, 2008.

4. Veerarajan. T, "Fundamentals of Mathematical Statistics", Yee Dee Publishing Pvt. Ltd, 2017.

Answers to Check Your Progress

1. $-1\ and + 1$

2. Non-negative

3. $r_{12} = 0$ and $r_{13} = 0$

4. Need not be a

BLOCK IV: Design of Experiments	
Unit 8	**Analysis of Variance one-way, two-way classification and Design of Experiments**

Unit – 8

Analysis of Variance one-way, two-way classification and Design of experiments

Objectives

After Studying this Unit, the student will be able to

- Explain the design of experiments, analysis of variance one way and two-way classifications.
- Distinguish between Completely Randomized Design, Randomized Block Design and Latin Square Design.
- Solve the problems in analysis of variance one way and two-way classifications, Completely Randomized Design, Randomized Block Design and Latin Square Design

Overview

In this unit, we will study the concept of Design of Experiments. We will only focus on the ANOVA one-way classification, ANOVA two-way classification and the most commonly used design of experiments such as Completely Randomized Design, Randomized Block Design and Latin Square Design.

8.1. Introduction

Experiment, what is meant is collection of data (which usually consist of a series of measurement of some feature of an object) for a scientific investigation, according to certain specified procedures. Statistics provides not only the principles and the basic for the proper planning of the experiments but also the methods for proper interpretation of the results of the experiment.

In the beginning, the study of the design of experiments was restricted only to agricultural experimentation. The need to save time and money has led to the study of methods to obtain maximum information with minimum cost and labour. Such considerations resulted in the subsequent acceptance and wide use of the design of experiments and related analysis of variance techniques in many fields of scientific experimentation.

A statistical experiment in any field is performed to verify a particular hypothesis. For example, an agricultural experiment may be performed to verify the claim that a particular manure has got the effect of increasing the yield of paddy. Here the quantity of the manure used and the amount of yield of paddy are the two variables involved directly. They are called experimental variables. Apart from these two, there are other variables such as fertility of the soil, the quality of the seed used and the amount of rain fall which also affect the yield of paddy. Such variables are called extraneous variables. The main aim of the design of experiments is to control the contribution of extraneous variables and hence to minimize the experimental error so that the results of the experiments could be attributed only to the experimental variables.

8.2. Basic Principles of Experimental Design

In order to achieve the objectives, usually the following three principles are adopted while designing experiments.

1. Randomisation

It is not possible to eliminate completely the contribution of extraneous variables to the value of the response variable (namely; the amount of yield of paddy). So, we try to minimize it by randomization technique. As per this technique, plots of the same size are taken and divided into two groups. In one group called the experimental group the manure is used in all the plots (units). In the other group of plots in which manure is not used but will provide a basis for comparison is called the control group.

If any information regarding the extraneous variables and the nature and magnitude of their effect on the response variable in question is not available, we resort t randomization which means selection of plots for the experimental and control group in a random manner. This technique provides the most effective way of eliminating any unknown bias in the experiment.

2. Replication

If the effects of different manures on the yield of paddy are studied, each manure is used in more than one plot. In other words, we resort to replication which means repetition. In order to estimate the amount of experimental error and hence to get some idea of the precision of the estimate of the manure effects, it is essential to carry out more than one test on each manure.

3. Local Control

In order to achieve adequate control of extraneous variables, another important principle used in the experimental design is the local control. This includes techniques such as grouping, blocking and balancing of the experimental Plots (units) used in the experimental design. By grouping, we mean combining sets homogeneous plots into groups, so that different manures may be used in different groups. The number of plots in different groups need not necessarily be the same.

By blocking we mean assigning the same number of plots in different blocks. The plots in the same block may be assumed to be relatively homogeneous. We use as many manures as the number of plots in a block in a random manner.

By balancing, we mean making minor changes in the procedures of grouping and blocking and then assigning the manures in such a manner that a balanced configuration is obtained.

The following are the commonly used design of experiments

1. Completely Randomized Design (C.R.D.)

C.R.D. is a design in which N values of a given random variable X (the yield of Paddy) contained in a sample are sub-divided into h classes according to one factor of classification (different manures)

Let us assume that we wish to compare h treatments (namely; h different manures) and there are n plots available for the experiment.

Let i^{th} treatment be replicated (repeated) n, times, so that $n_1 + n_2 + \cdots + n_k = n$. The plots to which the different treatments are to be given are found by the following randomization principle. The plots are numbered from 1 to n serially, n identical cards are taken which are also numbered from 1 to n and shuffled thoroughly. The numbers on the first n_1 card drawn randomly give the numbers of the plots to which the first treatment is to be given. The numbers on the next n_2 card drawn at random give the numbers of the plots to which the second treatment is to be given and so on. This design, known as completely randomized design is used when the plots are homogeneous or pattern of heterogeneity of the plots is not known.

2. Randomized Block Design (R.B.D.)

R.B.D. is a design in which the N variate values (yield of paddy) are classified according to two factors.

Assuming that there are N plots and they are divided into h blocks (rows) representing one factor of classification (say, soil fertility) each block containing k plots (columns) representing the other factor of classification (say, treatments). The plots in each block will be of homogeneous fertility as far as possible and k treatments are given to the k plots in each block in perfectly random manner such that each treatment occurs only once in any block. But the same k treatments are repeated from block to block.

3. Latin Square Design (L.S.D.)

L.S.D. is a design in which $N = n^2$ plots are taken and arranged in the form of an $n \times n$ square, such that the plots in each row will be homogeneous as far as possible with respect to one factor of classification, say, soil fertility. Plots in each column will be homogeneous as far as possible with respect to another factor of classification, say, seed quality. Then n treatments (third factor of classification) represented by letters are given to these plots such that each treatment occurs only once in each row and only once in each column. The various possible arrangement obtained in this manner are known as Latin squares of order n.

Analysis of Variance (ANOVA)

After planning and conducting experiments, the results obtained must be analysed and interpreted. The technique for making statistical inferences is known as the analysis of variance, which is widely used technique developed by R.A. Fisher. In general, there are several factors involved, in an experiment each one of which may cause a certain amount of variability in the observed values of the response variable.

In analysis of variance technique, we divide the total variation (represented by variance) in a group into parts which might have been caused by different factors and a residual random variation which could not be accounted for by any of these factors. The variation due to any specific factor is compared with the residual variation for significance by applying the F-test and thus test the homogeneity of the observed data, namely, test if all the observations have been drawn from the same normal population.

8.3. Analysis of Variance (ANOVA) for one factor classification

We assume that the N values of a given random variable (yield of paddy) contained in a sample are subdivided into h classes according to a factor of classification (manure)

We proceed with the assumption that the factor of classification has no effect on the variable and test if this assumption (null hypothesis) can be accepted.

Let x_{ij} be the value of the j^{th} member of the i^{th} class, which contains n, members. Let the general mean of all the N values be $\bar{x}$ and the mean of the n_i values in the i^{th} class be$\bar{x}_i$.

Now,

$$\sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i \sum_j \{(x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x})\}^2$$

$$\sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 + \sum_i \sum_j (\bar{x}_i - \bar{x})^2 + 2 \sum_i \sum_j (x_{ij} - \bar{x}_i)(\bar{x}_i - \bar{x})$$

$$\sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 + \sum_i \sum_j (\bar{x}_i - \bar{x})^2 + 2 \sum_i (\bar{x}_i - \bar{x}) \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) \qquad (1)$$

$$\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) = \text{Sum of the deviation of the } \mathrm{n_i} \text{ values of } \mathrm{x_{ij}} \text{ in the } \mathrm{i^{th}} \text{ class from their mean } \bar{x}_i$$

$$= 0 \qquad (2)$$

Using (2) in (1) we get

$$\sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 + \sum_i n_i (\bar{x}_i - \bar{x})^2$$

$P = P_2 + P_1$

Where $P = \text{Total variation}$

$$P_1 = \sum_i n_i (\bar{x}_i - \bar{x})^2$$

$= \text{Sum of the squared deviations of class means from the general mean}$
$\text{(namely, the variation between calsses)}$

$$P_2 = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$$

$= \text{sum of the squared deviation of variates from the corresponding class means}$
$\text{(variation within classes)}$

Since

$P_2 = \text{vartion within classes} = \mathrm{P} - P_1$ can be considered to have been obtained after removing the variation P_1 between classes from the total variation P.

Hence P_2 is regarded as the residual variation.

Now the items in the i^{th}class with variance $s_i^2 = \frac{1}{n_i} \sum_{j=1}^{ni} (x_{ij} - \bar{x}_i)^2$may be considered as a sample of size n_i drawn from a population with variance σ^2, hence $E\left(\frac{n_i s_i^2}{n_i - 1}\right) = \sigma^2$

By theory of estimation,

$$E\left[\sum_j (x_{ij} - \bar{x}_i)^2\right] = (n_i - 1)\sigma^2$$

$$E\left[\sum_i \sum_j (x_{ij} - \bar{x}_i)^2\right] = \sum_{i=1}^{h} (n_i - 1)\sigma^2$$

$$E[P_2] = (N - h)\sigma^2$$

$$E\left[\frac{P_2}{(N - h)}\right] = \sigma^2$$

$\frac{P_2}{(N-h)}$ is an unbiased estimate of σ^2 with $(N - h)$ degrees of freedom.

Now, if we consider the entire group of N items with variance

$$s^2 = \frac{1}{N} \sum_i \sum_j (x_{ij} - \bar{x})^2$$

As a sample of size N drawn from the same population

$$E\left[\sum_i \sum_j (x_{ij} - \bar{x})^2\right] = (N - 1)\sigma^2$$

$$E\left[\frac{P}{(N - 1)}\right] = \sigma^2$$

$\frac{P}{(N-1)}$ is an unbiased estimate of σ^2 with $(N - 1)$ degrees of freedom.

Now $P_1 = P - P_2$

$$E[P_1] = E[P] - E[P_2]$$

$$E[P_1] = (N - 1)\sigma^2 - (N - h)\sigma^2$$

$$E[P_1] = (h - 1)\sigma^2$$

$$E\left[\frac{P_1}{(h - 1)}\right] = \sigma^2$$

$\frac{P_1}{(h-1)}$ is also an unbiased estimate of σ^2 with $(h - 1)$ degrees of freedom.

If we assume that the sample population is normal, then the estimate $\frac{P1}{(h-1)}$ and $\frac{P2}{(N-h)}$ are independent and hence the ratio

$$\frac{\left[\frac{P_1}{(h-1)}\right]}{\left[\frac{P_2}{(N-h)}\right]}$$

Follows a F-distribution with $(h-1, N-h)$ degree of freedom or the ratio

$$\frac{\left[\frac{P_2}{(N-h)}\right]}{\left[\frac{P_1}{(h-1)}\right]}$$

Follows a F-distribution with $(N-h, h-1)$degree of freedom.

Choosing the ratio which is greater than 1, we employ the F-test.

If Calculated value of F is less than the table value of F at 5% , our hypothesis holds good, that is , different treatments do not contribute significantly by different yields.

ANOVE table for one factor of classification

Source of variation (S.V.)	Sum of square (S.S.)	Degrees of freedom (d.f.)	Mean square (M.S.)	Variance ratio (F)
Between Columns	P_1	$h-1$	$\frac{P_1}{(h-1)}$	$\left\{\frac{\frac{P_1}{(h-1)}}{\frac{P_2}{(N-h)}}\right\}^{\pm 1}$
Within classes	P_2	$N-h$	$\frac{P_2}{(N-h)}$	
Total	P	$N-1$		

10.3.1. Note

For calculating P, P_1, P_2 the following computational formulae may be used

$$P = N\{\frac{1}{N}\sum\sum x_{ij}^2 - x^2\}$$

$$P = N\{\frac{1}{N}\sum\sum x_{ij}^2 - (\frac{1}{N}\sum\sum x_{ij})^2\}$$

$$P = \sum\sum x_{ij}^2 - \frac{T^2}{N}, \text{where } \mathrm{T} = \sum\sum x_{ij}$$

Similarly, for the i^{th} class

$$\sum_j (x_{ij} - \bar{x}_i)^2 = \sum_j x_{ij}^2 - \frac{T_i^2}{n_i},\ where\ T_i = \sum_j x_{ij}$$

Therefore,

$$P_2 = \sum_i\sum_j (x_{ij} - \bar{x}_i)^2 = \sum_i\sum_j x_{ij}^2 - \sum_i \frac{T_i^2}{n_i}$$

Therefore, $P_1 = P - P_2 = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N}$

8.4. Analysis of Variance (ANOVA) for two factors of classification

We assume that the N values of the random variable (yield of paddy) contained in a sample are classified according to two factors-one factor classification (soil fertility) represented by h rows and the other factor (treatment) represented by k columns. So that $N = hk$.

We assume that the rows and columns are homogeneous, namely; there no difference in the variance values (yields of paddy) between the various rows and between the various columns and test if this assumption (null hypothesis) can be accepted.

Let x_{ij} be the value of the variable in the i^{th} row and j^{th} column.

Let $\bar{x}$ be the general mean of all the N values, $\bar{x}_i$, be the mean of the k values in the i^{th} row and $\bar{x}_{*j}$ be the mean of the h values in the j^{th} column.

Now,

$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_{i*} - \bar{x}_{*j} + \bar{x}) + (\bar{x}_{i*} - \bar{x}) + (\bar{x}_{*j} - \bar{x})$$

Therefore,

$$\sum\sum(x_{ij} - \bar{x})^2$$
$$= \sum\sum(x_{ij} - \bar{x}_{i*} - \bar{x}_{*j} + \bar{x})^2$$
$$+ \sum\sum(\bar{x}_{i*} - \bar{x})^2 + \sum\sum(x_{*j} - \bar{x})^2 + 2\sum\sum(x_{ij} - x_{i*} - \bar{x}_{j*} + \bar{x})(x_{i*} - \bar{x})$$
$$+ 2\sum\sum(x_{ij} - \bar{x}_{i*} - \bar{x}_{j*} + \bar{x})(\bar{x}_{*j} - \bar{x}) + 2\sum\sum(\bar{x}_{i*} - \bar{x})(\bar{x}_{*j} - \bar{x}) \quad (1)$$

Now the fourth member in the R.H.S. of (1)

$$= 2\sum(\bar{x}_{i*} - \bar{x})\sum_{j=1}^{k}(x_{ij} - \bar{x}_{i*} - \bar{x}_{*j} + \bar{x})$$
$$= 2\sum_i(\bar{x}_{i*} - \bar{x})(k\bar{x}_{i*} - k\bar{x}_{i*} - k\bar{x} + k\bar{x}) = 0$$

Similarly, the last two members in the R.H.S. of (1) also become each.

Omitting these zero valued terms, (1) becomes

$P = P_3 + P_1 + P_2$, say where

$$P_1 = \sum_i\sum_j(\bar{x}_{i*} - \bar{x})^2 = k\sum_i(\bar{x}_{i*} - \bar{x})^2$$

$$P_2 = \sum_i \sum_j (\bar{x}_{*j} - \bar{x})^2 = h \sum_i (\bar{x}_{*j} - \bar{x})^2$$

$$P_3 = \sum \sum (\bar{x}_{ij} - \bar{x}_{i*} - \bar{x}_{*j} + \bar{x})^2$$

$P =$ Total Variation

$P_1 =$ Sum of the squares due to the variations in the rows

$P_2 =$ Sum of the squares due to the variations in the columns

$P_3 =$ Sum of the squares due to the residual variations.

Using one factor of classification, we can prove that

$\frac{P_1}{(h-1)}, \frac{P_2}{(k-1)}, \frac{P_3}{(h-1)(k-1)}, \frac{P}{(hk-1)}$ are unbiased estimates of the population variance σ^2 with degrees of freedom $(h-1), (k-1), (h-1)(k-1)$ and $(hk-1)$ respectively.

If the sample population is assumed normal, all these estimates are independent.

Therefore,

$$\frac{\left[\frac{P_1}{(h-1)}\right]}{\left[\frac{P_3}{(h-1)(k-1)}\right]}$$

Or

It's reciprocal follows a F-distribution with $\{(h-1), (h-1)(k-1)\}$ degrees of freedom or with $\{(h-1)(k-1), (h-1)\}$ degrees of freedom, depending on the value of F. Similarly

$$\frac{\left[\frac{P_2}{(k-1)}\right]}{\left[\frac{P_3}{(h-1)(k-1)}\right]}$$

Or

It's reciprocal follows a F-distribution with $\{(k-1), (h-1)(k-1)\}$ degrees of freedom or with $\{(h-1)(k-1), (k-1)\}$ degrees of freedom, depending on the value of F then the F-test is applied as usual and the significance of the difference between rows and between columns analysed.

ANOVA table for two factors of classification

Source of variation (S.V.)	Sum of square (S.S.)	Degrees of freedom (d.f.)	Mean square (M.S.)	Variance ratio (F)
Between Rows	P_1	$h-1$	$\frac{P_1}{(h-1)}$	$\left\{\frac{\frac{P_1}{(h-1)}}{\frac{P_3}{(h-1)(k-1)}}\right\}^{\pm 1}$
Between Columns	P_2	$k-1$	$\frac{P_2}{(k-1)}$	$\left\{\frac{\frac{P_2}{(k-1)}}{\frac{P_3}{(h-1)(k-1)}}\right\}^{\pm 1}$
Residual	P_3	$(h-1)(k-1)$	$\frac{P_3}{(h-1)(k-1)}$	
Total	P	$hk-1$		

8.4.1. Note

For computing P, P_1, P_2 *and* P_3, the following working formulae may used

1. $P = \sum_i \sum_j x_{ij}^2 - \frac{T^2}{N}, \quad where\ T = \sum_i \sum_j x_{ij}$

2. $P_1 = \frac{1}{k} \sum T_i^2 - \frac{T^2}{N}, where\ T_i = \sum_{j=1}^{k} x_{ij}$

3. $P_2 = \frac{1}{h} \sum T_j^2 - \frac{T^2}{N}, where\ T_j = \sum_{i=1}^{h} x_{ij}$

4. $P_3 = P - P_1 - P_2, \quad Also\ \sum_i T_i = \sum_j T_j = T$

8.5. Analysis of Variance (ANOVA) for three factors of classification

We assume that $N\ (= n^2)$ variate values (yield of paddy) contained in a sample are classified to three factors, namely soil fertility, seed quality and treatment represented by the rows, columns and letters respectively.

We assume that the rows, columns and letters are homogeneous, namely, there is no difference in the variate values between rows, between the columns and between the letters and test if this assumption (null hypothesis) can be accepted.

Let x_{ij} be the value of the variate corresponding to the i^{th} row, j^{th} column and k^{th} letter.

Let

$$\bar{x} = \frac{1}{n^2} \sum \sum x_{ij}$$

$$\bar{x}_{i*} = \frac{1}{n} \sum_j x_{ij}$$

$x_{*j} = \frac{1}{n} \sum_i x_{ij}$ and x_k be the mean of the values of x_{ij} corresponding to the k^{th} treatment.

Now

$$x_{ij} - \bar{x} = (\bar{x}_{i*} - \bar{x}) + (\bar{x}_{*j} - \bar{x}) + (\bar{x}_k - \bar{x}) + (x_{ij} - \bar{x}_{i*} - \bar{x}_{*j} - \bar{x}_k + 2\bar{x})$$

Therefore,

$$\sum \sum (x_{ij} - \bar{x})^2$$
$$= n \sum_j (\bar{x}_{i*} - \bar{x})^2$$
$$+ n \sum_j (\bar{x}_{*j} - \bar{x})^2 + n \sum_k (\bar{x}_k - \bar{x})^2 + \sum \sum (x_{ij} - \bar{x}_{i*} - \bar{x}_{*j} - \bar{x}_k + 2\bar{x})^2$$

Since all the product terms vanish, we have

$$P = P_1 + P_2 + P_3 + P_4$$

We can prove that $\frac{P1}{(n-1)}, \frac{P_2}{(n-1)}, \frac{P_3}{(n-1)}, \frac{P_4}{(n-1)(n-2)}, \frac{P}{(n^2-1)}$ are unbiased estimates of the population variance σ^2 with degrees of freedom$(n-1), (n-1), (n-1), (n-1)(n-2)$, (n^2-1) respectively.

If the sample population is assumed normal, all these estimates are independent. Therefore, each of

$$\frac{[\frac{P_1}{(n-1)}]}{[\frac{P_4}{(n-1)(n-2)}]}$$

$$\frac{[\frac{P_2}{(n-1)}]}{[\frac{P_4}{(n-1)(n-2)}]}$$

$$\frac{[\frac{P_3}{(n-1)}]}{[\frac{P_4}{(n-1)(n-2)}]}$$

Or their reciprocal follows a F-distribution with $\{(n-1), (n-1)(n-2)\}$ degrees of freedom or $\{(n-1)(n-2), (n-1)\}$ degrees of freedom, depending on the value of F then the F-test is applied as usual and the significance of the differences between rows, columns and letters is analysed.

ANOVA table for three factors of classification

Source of variation (S.V.)	Sum of square (S.S.)	Degrees of freedom (d.f.)	Mean square (M.S.)	Variance ratio (F)
Between Rows	P_1	$n-1$	$\frac{P_1}{(n-1)}$	$\left\{\frac{\frac{P_1}{(n-1)}}{\frac{P_4}{(n-1)(n-2)}}\right\}^{\pm 1}$
Between Columns	P_2	$n-1$	$\frac{P_2}{(n-1)}$	$\left\{\frac{\frac{P_2}{(n-1)}}{\frac{P_4}{(n-1)(n-2)}}\right\}^{\pm 1}$
Between letters	P_3	$n-1$	$\frac{P_3}{(n-1)}$	$\left\{\frac{\frac{P_3}{(n-1)}}{\frac{P_4}{(n-1)(n-2)}}\right\}^{\pm 1}$
Residual	P_4	$(n-1)(n-2)$	$\frac{P_4}{(n-1)(n-2)}$	
Total	P	n^2-1		

8.5.1. Note

For computing P, P_1, P_2, P_3 and P_4 the following working formulae may used

1. $P = \sum_i \sum_j x_{ij}^2 - \frac{T^2}{N}$, $where\ T = \sum_i \sum_j x_{ij}$ and $N = n^2$

2. $P_1 = \frac{1}{n} \sum T_i^2 - \frac{T^2}{N}$, $where\ T_i = \sum_{j=1}^{n} x_{ij}$ and $N = n^2$

3. $P_2 = \frac{1}{n} \sum T_j^2 - \frac{T^2}{N}$, $where\ T_j = \sum_{i=1}^{n} x_{ij}$ and $N = n^2$

4. $P_3 = \frac{1}{n} \sum T_k^2 - \frac{T^2}{N}$, $where\ T_K$ is the sum of all $x_{ij}'s$ receiving the k^{th} treatment and $N = n^2$

5. $P_4 = P - P_1 - P_2 - P_3$ $Also$ $\sum_i T_i = \sum_j T_j = \sum_k T_k = T$

8.5.2. Note

Simplification of Computational work: The Variance of a set of values is independent of the origin and so a shift of origin does not affect the variance calculations. Hence in analysis of variance problems, we can subtract a convenient number from the original values and work out the problem with the new values obtained. Also, since we are concerned with variance ratios change of scale also may be introduced without affecting the value of F.

8.6. Examples

8.6.1. A Car rental agency, which uses 5 different brands of tyres in the process of deciding the brand of tyre to purchase as standard equipment for its fleet, finds that each of 5 tyres of each brand last the following number of kilometers (in thousands)

Tyre brands				
A	B	C	D	E
36	46	35	45	41
37	39	42	36	39
42	35	37	39	37
38	37	43	35	35
47	43	38	32	38

Test the hypothesis is that the five tyre brands have almost the same average life.

Solution:

Null hypothesis H_0: There is no significant difference between in the average life of the five tyre brands.

Alternative hypothesis H_1: There is a significant difference between in the average life of the five tyre brands.

Let $X_{ij} = x_{ij} - 40$

	Tyre brands					
	A	B	C	D	E	Total
X_{ij} Values	-4	6	-5	5	1	
	-3	-1	2	-4	-1	
	2	-5	-3	-1	-3	
	-2	-3	3	-5	-5	
	7	3	-2	-8	-2	
T_i	0	0	-5	-13	-10	-28
n_i	5	5	5	5	5	25
$\frac{T_i^2}{n_i}$	0	0	5	33.8	20	58.8
$\sum_{j=1}^{5} x_{ij}^2$	82	80	51	131	40	$\sum\sum x_{ij}^2 = 384$

$$T = \sum_i T_i = -28$$

$$N = \sum n_i = 25$$

$$\sum\sum x^2{}_{ij} = 384$$

$$P = \sum\sum x_{ij}^2 - \frac{T^2}{N} = 384 - \frac{(-28)^2}{25} = 352.64$$

$P_1 = \sum \frac{T_i^2}{n_i} - \frac{T^2}{N} = 58.8 - 31.36 = 27.44$

$P_2 = P - P_1 = 352.64 - 27.44 = 325.20$

ANOVA Table

Source of variation (S.V.)	Sum of square (S.S.)	Degrees of freedom (d.f.)	Mean square (M.S.)	Variance ratio (F)
Between tyre brands	$P_1 = 27.44$	$h - 1 =$ $5 - 1 = 4$	$\frac{P_1}{(h-1)} = 6.86$	$\frac{16.26}{6.86} = 2.37$
Within tyre brands	$P_2 = 325.20$	$N - h =$ $25 - 5 = 20$	$\frac{P_2}{(N-h)} = 16.26$	
Total	$P = 352.64$	$N - 1 =$ $25 - 1 = 24$		

Table value of F at 5% level of significance for (20, 4) degrees of freedom is 5.80

Calculated value of F is less than table value of F.

Therefore, Null Hypothesis H_0 is accepted.

Hence, the five tyre brands have almost the same average. That is, they do not differ significantly in their lives.

8.6.2. The following data represent the number of units of production per day turned out by different workers using 4 different types of machines.

		Machine Type			
		A	B	C	D
Workers	1	44	38	47	36
	2	46	40	52	43
	3	34	36	44	32
	4	43	38	46	33
	5	38	42	49	39

(a) Test whether the five workers differ with respect to mean productivity (b) Test whether the mean productivity is the same for the four different machine types.

Solution:

Null Hypothesis H_0: (a) There is no significant difference between in the mean productivity of the 5 workers and (b) There is no significant difference between in the mean productivity of the 4 machine types.

Alternative Hypothesis H_1: (a) There is a significant difference between in the mean productivity of the 5 workers and (b) There is a significant difference between in the mean productivity of the 4 machine types.

Let $X_{ij} = x_{ij} - 40$

Workers	Machine Type				T_i	$\frac{T_i^2}{k}$	$\sum_j X_{ij}^2$
	A	B	C	D			
1	4	-2	7	-4	5	6.25	85
2	6	0	12	3	21	110.25	189
3	-6	-4	4	-8	-14	49	132
4	3	-2	6	-7	0	0	0
5	-2	2	9	-1	8	16	90
T_j	5	-6	38	-17	T=20	$\sum \frac{T_i^2}{k} = 181.5$	$\sum_i \sum_j X_{ij}^2 = 594$
$\frac{T_j^2}{h}$	5	7.2	288.8	57.8	$\sum \frac{T_j^2}{h} = 358.8$		
$\sum_i X_{ij}^2$	110	28	326	139	$\sum_i \sum_j X_{ij}^2 = 594$		

$$P = \sum \sum x_{ij}^2 - \frac{T^2}{N} = 594 - \frac{(20)^2}{20} = 574$$

$$P_1 = \sum \frac{T_i^2}{k} - \frac{T^2}{N} = 181.5 - 20 = 161.5$$

$$P_2 = \sum \frac{T_i^2}{h} - \frac{T^2}{N} = 358.8 - 20 = 338.8$$

$$P_3 = P - P_1 - P_2 = 574 - 161.5 - 338.8 = 73.7$$

ANOVA table

Source of variation (S.V.)	Sum of square (S.S.)	Degrees of freedom (d.f.)	Mean square (M.S.)	Variance ratio (F)
Between Rows (Workers)	$P_1 = 161.5$	$h-1 = 5-1 = 4$	$\frac{P_1}{(h-1)} = 40.375$	$F_R = \frac{40.375}{6.142} = 6.57$
Between Columns (machine types)	$P_2 = 338.8$	$k-1 = 4-1 = 3$	$\frac{P_2}{(k-1)} = 112.933$	$F_C = \frac{112.933}{6.142} = 18.39$
Residual	$P_3 = 73.7$	$(h-1)(k-1) = 12$	$\frac{P_3}{(h-1)(k-1)} = 6.142$	
Total	$P = 574$	$hk - 1 = 19$		

Table value of F_R at 5% level of significance of (4,12) degrees of freedom is 3.26

Calculated value of F_R is greater than table value of F_R.

Null Hypothesis H_0 is rejected. (For Rows)

Therefore, there is significant difference between the mean productivity of the workers.

Table value of F_C at 5% level of significance of (3,12) degrees of freedom is 3.49

Calculated value of F_C is greater than table value of F_C.

Null Hypothesis H_0 is rejected. (For Columns)

Therefore, there is significant difference between the mean productivity for the four different machine types.

8.6.3. A completely randomised design (CRD) experiment with 10 plots and 3 treatments gave the following results:

Plot. No.	1	2	3	4	5	6	7	8	9	10
Treatment	A	B	C	A	C	C	A	B	A	B
Yield	5	4	3	7	5	1	3	4	1	7

Analyse the results for treatment effects.

Solution:

Rearranging the data (yields) according to the treatments, the following table is obtained.

	Treatment		
	A	B	C
Yield from plots(x_{ij})	5	4	3
	7	4	5
	3	7	1
	1	-	-

Null hypothesis H_0: Treatments do not differ significantly.

Alternative hypothesis H_1: Treatments differ significantly.

	A	B	C	Total
x_{ij} Values	5	4	3	
	7	4	5	
	3	7	1	
	1	-	-	
T_i	16	15	9	40
T_i^2	256	225	81	-
n_i	4	3	3	N=10
$\frac{T_i^2}{n_i}$	64	75	27	166
$\sum_{j=1}^{5} x_{ij}^2$	84	81	35	$\sum\sum x_{ij}^2 = 200$

$$T = \sum_i T_i = 40$$

$$N = \sum n_i = 10$$

$$\sum \sum x^2_{ij} = 200$$

$$P = \sum \sum x^2_{ij} - \frac{T^2}{N} = 200 - \frac{(40)^2}{10} = 40$$

$$P_1 = \sum \frac{T_i^2}{n_i} - \frac{T^2}{N} = 166 - 160 = 6$$

$$P_2 = P - P_1 = 40 - 6 = 34$$

ANOVA Table

Source of variation (S.V.)	Sum of square (S.S.)	Degrees of freedom (d.f.)	Mean square (M.S.)	Variance ratio (F)
Between Classes (Treatments) Within Classes (Treatments)	$P_1 = 6$ $P_2 = 34$	$h - 1 = 3 - 1 = 2$ $N - h = 10 - 3 = 7$	$\frac{P_1}{(h-1)} = 3$ $\frac{P_2}{(N-h)} = 4.86$	$\frac{4.86}{3} = 1.62$
Total	$P = 40$	$N - 1 = 10 - 1 = 9$		

Table value of F at 5% level of significance for (7, 2) degrees of freedom is 19.35

Calculated value of F is less than table value of F.

Therefore, Null Hypothesis H_0 is accepted.

Hence, the treatments do not give significantly different yields.

8.6.4. Three varieties A, B, C of a crop are tested in randomised block design with four replications, the layout being as given below. The yields are given kilograms. Analyse for significance

C48	A51	B52	A49
A47	B49	C52	C51
B49	C53	A49	B50

Solution:

Rewriting the given data such that the rows represent the blocks and columns represent the varieties of crop, we have the following table

Blocks	Crops		
	A	B	C
1	47	49	48
2	51	49	53
3	49	52	52
4	49	50	51

Null Hypothesis H_0: (a)There is no significant difference between rows (Blocks) and (b) There is no significant difference between columns (Crops)

Alternative Hypothesis H_1: (a) There is a significant difference between rows (Blocks) and (b) There is a significant difference between columns (Crops)

Let $X_{ij} = x_{ij} - 50$

Blocks	Crops			T_i	$\frac{T_i^2}{k}$	$\sum_j X_{ij}^2$
	A	B	C			
1	-3	-1	-2	-6	12	14
2	1	-1	3	3	3	11
3	-1	2	2	3	3	9
4	-1	0	1	0	0	2
T_j	-4	0	4	T=0	$\sum \frac{T_i^2}{k} = 18$	$\sum_i \sum_j X_{ij}^2 = 36$
$\frac{T_j^2}{h}$	4	0	4	$\sum \frac{T_j^2}{h} = 8$		
$\sum_i X_{ij}^2$	12	6	18	$\sum_i \sum_j X_{ij}^2 = 36$		

$$P = \sum \sum x_{ij}^2 - \frac{T^2}{N} = 36 - \frac{(0)^2}{12} = 574$$

$$P_1 = \sum \frac{T_i^2}{k} - \frac{T^2}{N} = 18 - 0 = 18$$

$$P_2 = \sum \frac{T_i^2}{h} - \frac{T^2}{N} = 8 - 0 = 8$$

$$P_3 = P - P_1 - P_2 = 36 - 18 - 8 = 10$$

ANOVA table

Source of variation (S.V.)	Sum of square (S.S.)	Degrees of freedom (d.f.)	Mean square (M.S.)	Variance ratio (F)
Between Rows (Blocks)	$P_1 = 18$	$h-1 = 4-1 = 3$	$\frac{P_1}{(h-1)} = 6$	$F_R = \frac{6}{1.67} = 3.6$
Between Columns (Crops)	$P_2 = 8$	$k-1 = 3-1 = 2$	$\frac{P_2}{(k-1)} = 4$	$F_C = \frac{4}{1.67} = 2.4$
Residual	$P_3 = 10$	$(h-1)(k-1) = 6$	$\frac{P_3}{(h-1)(k-1)} = 1.67$	
Total	$P = 10$	$hk - 1 = 11$		

Table value of F_R at 5% level of significance of (3, 6) degrees of freedom is 4.76

Calculated value of F_R is Less than table value of F_R.

Null Hypothesis H_0 is accepted. (For Rows)

Therefore, there is no significant difference between Rows (Blocks)

Table value of F_C at 5% level of significance of (2, 6) degrees of freedom is 5.14

Calculated value of F_C is Less than table value of F_C.

Null Hypothesis H_0 is accepted. (For Columns)

Therefore, there is no significant difference between Columns (Crops)

Hence the blocks do not differ significantly and the varieties of crop do not differ significantly with respect to the yield.

8.6.5. Analyse the variance in the following Latin square of yields (in kgs) of paddy, where A, B, C, D denote the different methods of cultivation:

D122	A121	C123	B122
B124	C123	A122	D125
A120	B119	D120	C121
C122	D123	B121	A122

Examine whether the different methods of cultivation have given significantly different yields.

Solution:

Null hypothesis H_0: (a) There is no significant difference between rows (b) There is no significant difference between columns and (c) There is no significant difference between letters (method of cultivation)

Alternative hypothesis H_1: (a) There is a significant difference between rows (b) There is a significant difference between columns and (c) There is a significant difference between letters (method of cultivation)

Let $X_{ij} = x_{ij} - 120$

Rows	Columns				T_i	$\frac{T_i^2}{n}$	$\sum_j X_{ij}^2$
	I	II	III	IV			
1	D2	A1	C3	B2	8	16	18
2	B4	C3	A2	D5	14	49	54
3	A0	B-1	D0	C1	0	0	2
4	C2	D3	B1	A2	8	16	18
T_j	8	6	6	10	T=20	$\sum \frac{T_i^2}{n} = 81$	$\sum_i \sum_j X_{ij}^2 = 92$
$\frac{T_j^2}{n}$	16	9	9	25	$\sum \frac{T_j^2}{n} = 59$		
$\sum_i X_{ij}^2$	24	20	14	34	$\sum_i \sum_j X_{ij}^2 = 92$		

Rearranging the $X_{ij}'s$ values according to the letters (method of cultivation), we get the following table

Letter	Value of X_k				T_k	$\frac{T_k^2}{n}$
A	1	2	0	2	5	6.25
B	2	4	-1	1	6	9
C	3	3	1	2	9	20.25
D	2	5	0	3	10	25
Total					T=30	$\sum \frac{T_k^2}{n} = 60.50$

$$P = \sum_i \sum_j x_{ij}^2 - \frac{T^2}{N} = 92 - \frac{(30)^2}{16} = 35.75$$

$$P_1 = \frac{1}{n} \sum T_i^2 - \frac{T^2}{N} = 81 - 56.25 = 24.75$$

$$P_2 = \frac{1}{n} \sum T_j^2 - \frac{T^2}{N} = 59 - 56.25 = 2.75$$

$$P_3 = \frac{1}{n}\sum T^2_k - \frac{T^2}{N} = 60.50 - 56.25 = 4.25$$

$$P_4 = P - P_1 - P_2 - P_3 = 35.75 - 24.75 - 2.75 - 4.25 = 4$$

ANOVA table

Source of variation (S.V.)	Sum of square (S.S.)	Degrees of freedom (d.f.)	Mean square (M.S.)	Variance ratio (F)
Between Rows	$P_1 = 24.75$	$n-1 = 4-1-3$	$\frac{P_1}{(n-1)} = 8.25$	$F_R = \frac{8.25}{0.67} = 12.31$
Between Columns	$P_2 = 2.75$	$n-1 = 4-1=3$	$\frac{P_2}{(n-1)} = 0.92$	$F_c = \frac{0.92}{0.67} = 1.37$
Between letters	$P_3 = 4.25$	$n-1 = 4-1=3$	$\frac{P_3}{(n-1)} = 1.42$	$F_T = \frac{1.42}{0.67} = 2.12$
Residual	$P_4 = 4$	$(n-1)(n-2) = 6$	$\frac{P_4}{(n-1)(n-2)} = 0.67$	
Total	$P = 35.75$	$n^2 - 1$		

Table value of F_R at 5% level of significance of (3, 6) degrees of freedom is 4.76

Calculated value of F_R is greater than table value of F_R.

Null Hypothesis H_0 is rejected. (For Rows)

Therefore, there is a significant difference between Rows.

Table value of F_C at 5% level of significance of (3, 6) degrees of freedom is 4.76

Calculated value of F_C is Less than table value of F_C.

Null Hypothesis H_0 is accepted. (For Columns)

Therefore, there is no significant difference between Columns

Table value of F_T at 5% level of significance of (3, 6) degrees of freedom is 4.76

Calculated value of F_T is Less than table value of F_T.

Null Hypothesis H_0 is accepted. (For Letters)

Therefore, there is no significant difference between letters (method of cultivation)

Hence the difference between the methods of cultivation is not significant.

Let Us Sum Up

In this unit we studied the design of experiments. We focused only on analysis of variance one-way classification, two-way classification, Completely Randomized Design, Randomized Block Design and Latin Square Design.

Check Your Progress

1. The term Analysis of variance was introduced by________.

2. The Analysis of variance originated in____________.

3. ANOVA table stands for________.

4. The stimulus to the development of theory and practice of experimental design came from ________.

5. The most widely used all experimental design is______.

6. The science of experimental designs is associated with the name________.

7. The Latin square model assumes that interactions between treatments and rows and columns groupings are_______.

8. The randomised block design is available for a wide range of treatments_________.

9. _______Latin square design is not possible.

10. The assumptions in analysis of variance are the same as________.

Glossaries

Analysis of variance (ANOVA): It is the separation of variance ascribable to one group of causes from the variance ascribable to other groups.

One-way classification: In one-way classification the data are classified according to only one criterion or factor.

Two-way classification: In two-way classification the data are classified according to the two different criteria or factors.

Design of experiment: The logical construction of the experiment in which the degree of uncertainty with which the inference is drawn may be will defined.

Completely Randomized Design: In this Design, treatments are allocated at random to the experimental units over the entire experimental material.

Randomized block Design: It is an experimental design where the experimental units are in groups called block. The treatments are randomly allocated to the experimental units inside each block. When all treatments appear at least once in each block, we have a completely randomized block.

Latin Square Design: It is the arrangement of t treatments, each one repeated t times, in such a way that each treatment appears exactly one time in each row and each column in the design. This kind of design is used to reduce systematic error due to rows (treatments) and columns.

Suggested Readings

1. Freund. J.E.," Mathematical Statistics", Prentice Hall of India, Fifth Edition, 2001.

2. Gupta. S.C. and Kapoor. V. K., "Fundamentals of Mathematical Statistics", Sultan Chand & Sons, Eleventh Edition, 2003.

3. Devore. J. L. "Probability and Statistics for Engineers", Brooks/Cole (Cengage Learning), First India Reprint, 2008.

Answers to Check Your Progress

1. R. A. Fisher

2. Agrarian research

3. Analysis of Variance table

4. Agricultural research

5. Randomised block design

6. Latin square

7. non-existent

8. 2 to 24

9. 2×2

10. F-test

BLOCK V: Multivariate Analysis

Unit – 9

Matrix Algebra and Random variables

Structure

Objectives

After Studying this Unit, the student will be able to

- Explain the random vectors and matrices
- Demonstrate the concept of mean vectors and covariance matrices
- Summarize the partitioning the covariance matrix, sample mean vector and covariance matrix.

Overview

In this unit, we will study the concept of random variables, random matrices, mean vectors, covariance matrices, partitioning the covariance matrix, partitioning the sample mean vector and covariance matrix.

9.1. Introduction

Scientific inquiry is an iterative learning process. Objectives pertaining to the explanation of a social or physical phenomenon must be specified and then tested by gathering and analysing data. In turn, an analysis of the data gathered by experimentation or observation will usually suggest a modified explanation of the phenomenon. Throughout this iterative learning process, variables are often added or deleted from the study. Thus, the complexities of most phenomena require an investigator to collect observations on many different variables. This block concerned with statistical methods designed to elicit information from these kinds of data sets. Because the data include simultaneous measurements on many variables, this body of methodology is called multivariate analysis.

9.1.1. Arrays

Multivariate data arise whenever an investigator, seeking to understand a social or physical phenomenon, selects a number *p* ~ 1 of *variables* or *characters* to record. The values of these variables are all recorded for each distinct *item,* individual, or experimental unit.

We will use the notation $x_{jk}$to indicate the particular value of the k^{th}variable that is observed on the j^{th}item, or trial. That is,
x_{jk}= measurement of the k^{th}variable on the j^{th}item

Consequently, *n* measurements on *p* variables can be displayed as follows:

	Variable 1	Variable 2	…	Variable k	…	Variable p
Item 1	x_{11}	x_{12}	…	x_{1k}	…	x_{1p}
Item 2	x_{21}	x_{22}	…	x_{2k}	…	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
Item j	x_{j1}	x_{j2}	…	x_{jk}	…	x_{jp}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
Item n	x_{n1}	x_{n2}	…	x_{nk}	…	x_{mp}

Or we can display these data as a rectangular array, called X, of *n* rows and *p* columns:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

The array X, then, contains the data consisting of all of the observations on all of the variables.

9.1.2. Example (A data array)

A selection of four receipts from a university bookstore was obtained in order to investigate the nature of book sales. Each receipt provided, among other things, the number of books sold and the total amount of each sale. Let the first variable be total dollar sales and the second variable be number of books sold. Then we can regard the corresponding numbers on the receipts as four measurements on two variables. Suppose the data, in tabular form, are

Variable 1 (dollar sales)	45	52	48	58
Variable 2 (number of books)	4	5	4	3

Solution:

Using the notation just introduced, we have

$x_{11} = 42, x_{21} = 52, \quad x_{31} = 48, \ x_{41} = 58$

$x_{12} = 4, x_{22} = 5, \quad x_{32} = 4, \ x_{42} = 3$

and the data array X is $X = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$ with four rows and two columns.

9.1.3. Vectors

An array x of *n* real numbers $x_1, x_2, \dots, x_n$ is called a *vector,* and it is written as

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \text{ or } X' = [x_1, x_2, \dots, x_n]$$

where the prime denotes the operation of *transposing* a column to a row.

9.2. Random Vectors and Matrices

A *random vector* is a vector whose elements are random variables. Similarly, a *random matrix* is a matrix whose elements are random variables. The expected value of a random matrix (or vector) is the matrix (vector) consisting of the expected values of each of its elements. Specifically, let $X = \{X_{ij}\}$ be an *n* ×*P* random matrix. Then the expected value of X, denoted by *E(X),* is the *n* ×*P* matrix of numbers (if they exist)

$$E(X) = \begin{bmatrix} E(X_{11}) & E(X_{12}) & \dots & E(X_{1p}) \\ E(X_{21}) & E(X_{22}) & \dots & E(X_{2p}) \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ E(X_{n1}) & E(X_{n2}) & \dots & E(X_{np}) \end{bmatrix}$$

where, for each element of the matrix

$$E(X_{ij}) = \begin{cases} \int_{-\infty}^{\infty} x_{ij} f_{ij}(x_{ij})\, dx_{ij} & \text{if } X_{ij} \text{ is a continuous random variable with} \\ & \text{probability density function } f_{ij}(x_{ij}) \\ \sum_{all\, x_{ij}} x_{ij} p_{ij}(x_{ij}) & \text{if } X_{ij} \text{ is a discrete random variable with} \\ & \text{probability function } p_{ij}(x_{ij}) \end{cases}$$

9.2.1. Example (Computing expected values for discrete random variables)

Suppose p= 2 and n = 1, and consider the random vector $X = \{X_1, X_2\}$. Let the discrete random variable X_1 have the following probability function:

x_1	-1	0	1
$p_1(x_1)$	0.3	0.3	0.4

Solution:

Then $E(X_1) = \sum_{all\, x_1} x_1 p_1(x_1) = (-1)(0.3) + (0)(0.3) + (1)(0.4) = 0.1$

Similarly, let the discrete random variable X_2 have the probability function

x_2	0	1
$p_2(x_2)$	0.8	0.2

Then $E(X_2) = \sum_{all\, x_2} x_2 p_2(x_2) = (0)(0.8) + (1)(0.2) = 0.2$

Thus, $E[X] = \begin{bmatrix} E(X_1) \\ E(X_2) \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix}$

9.3. Mean Vectors and Covariance Matrices

Suppose $X' = [X_1, X_2, \dots, X_P]$ is a p x 1 random vector. Then each element of X is a random variable with its own marginal probability distribution. The marginal means μ_i and variances σ_i^2 are defined as $\mu_i = E(X_i)$ and $\sigma_i^2 = E(X_i - \mu_i)^2$, $i = 1, 2, \dots, p$, respectively. Specifically,

$$\mu_i = \begin{cases} \displaystyle\int_{-\infty}^{\infty} x_i f_i(x_i)\, dx_i & \text{if } X_i \text{ is a continuous random variable with} \\ & \text{probability density function } f_i(x_i) \\ \displaystyle\sum_{all\, x_i} x_i p_i(x_i) & \text{if } X_i \text{ is a discrete random variable with} \\ & \text{probability function } p_i(x_i) \end{cases}$$

$$\sigma_i^2 = \begin{cases} \displaystyle\int_{-\infty}^{\infty} (x_i - \mu_i)^2 f_i(x_i)\, dx_i & \text{if } X_i \text{ is a continuous random variable with} \\ & \text{Probability density function } f_i(x_i) \\ \displaystyle\sum_{all\, x_i} (x_i - \mu_i)^2 p_i(x_i) & \text{if } X_i \text{ is a discrete random variable with} \\ & \text{prob ability function } p_i(x_i) \end{cases}$$

It will be convenient to denote the marginal variances by σ_{ii} rather than the more traditional σ_i^2, consequently, we shall adopt this notation.

The behaviour of any pair of random variables, such as X_i and X_k is described by their joint probability function, and a measure of the linear association between them is provided by the covariance.

$\sigma_{ik} = E(X_i - \mu_i)(X_k - \mu_k)$

$$\sigma_{ik} = \begin{cases} \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_i - \mu_i)(x_k - \mu_k)\, f_{ik}(x_i, x_k)\, dx_i\, dx_k & \text{if } X_i, X_k \text{ are continuous random variables with the joint densinty function } f_{ik}(x_i, x_k) \\ \displaystyle\sum_{all\, x_i}\sum_{all\, x_k} (x_i - \mu_i)(x_k - \mu_k)\, p_{ik}(x_i, x_k) & \text{if } X_i, X_k \text{ are discrete random variables with joint probability function } p_{ik}(x_i, x_k) \end{cases}$$

and μ_i and μ_k, $i, k = 1,2,\ldots,P$, are the marginal means. When $i = k$, the covariance becomes the marginal variance.

The collective behaviour of the *P* random variables $X_1, X_2, \ldots, X_P$ or, equivalently, the random vector $X' = [X_1, X_2, \ldots, X_P]$ is described by a joint probability density function $f(x_1, x_2, \ldots, x_p) = f(x)$. $f(x)$ will often be the multivariate normal density function.

If the joint probability $P[X_i \leq x_i \text{ and } X_K \leq x_k]$ can be written as the product of the corresponding marginal probabilities, so that
$P[X_i \leq x_i \text{ and } X_K \leq x_k] = P[X_i \leq x_i]P[X_K \leq x_k]$ for all pairs of values x_i and x_k then X_i and X_k are said to be *statistically independent.*

When X_i and X_k are continuous random variables with joint density $f_{ik}(x_i, x_k)$ and marginal densities $f_i(x_i)$ and $f_k(x_k)$ the independence condition becomes
$f_{ik}(x_i, x_k) = f_i(x_i) f_k(x_k)$ for all pairs (x_i, x_k).

The P continuous random variables $X_1, X_2, \ldots, X_p$ are mutually statistically independent if their joint density can be factored as

$f_{1.2......p}(x_1.x_2,\ldots,x_p) = f_1(x_1)f_2(x_2)\ldots f_p(x_p)$ for all p-tuples $(x_1,x_2,\ldots,x_p)$

Statistical independence has an important implication for covariance. The factorization in $f_{1.2......p}(x_1.x_2,\ldots,x_p) = f_1(x_1)f_2(x_2)\ldots f_p(x_p)$ implies that $Cov(X_i,X_k) = 0$.

Thus, $Cov(X_i,X_k) = 0$ If X_i and X_k are independent.

The converse of the above statement is not true in general; there are situations where $Cov(X_i,X_k) = 0$but X_i and $X_k$are not independent.

The means and covariances of the $P \times 1$ random vector X can be set out as matrices. The expected value of each element is contained in the vector of means $\mu = E(X)$and the P variances $\sigma_{ii}$and the *p(p-1)/2* distinct covariances σ_{ik} $(i < k)$are contained in the symmetric variance-covariance matrix $\sum = E(X-\mu)(X-\mu)'$.Specifically

$$E(X) = \begin{bmatrix} E(x_1) \\ E(x_2) \\ \cdot \\ \cdot \\ E(x_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \mu_n \end{bmatrix} = \mu \text{ and}$$

$$\Sigma = E(X-\mu)(X-\mu)'$$

$$\Sigma = E\left(\begin{bmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ \cdot \\ \cdot \\ X_p-\mu_p \end{bmatrix} [X_1-\mu_1 \quad X_1-\mu_1 \quad \ldots \quad X_k-\mu_k]\right)$$

$$\Sigma = E\begin{bmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \ldots & (X_1-\mu_1)(X_p-\mu_p) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \ldots & (X_2-\mu_2)(X_p-\mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (X_p-\mu_p)(X_1-\mu_1) & (X_p-\mu_p)(X_2-\mu_2) & \ldots & (X_p-\mu_p)^2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} E(X_1-\mu_1)^2 & E(X_1-\mu_1)(X_2-\mu_2) & \ldots & E(X_1-\mu_1)(X_p-\mu_p) \\ E(X_2-\mu_2)(X_1-\mu_1) & E(X_2-\mu_2)^2 & \ldots & E(X_2-\mu_2)(X_p-\mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_p-\mu_p)(X_1-\mu_1) & E(X_p-\mu_p)(X_2-\mu_2) & \ldots & E(X_p-\mu_p)^2 \end{bmatrix}$$

$$\Sigma = Cov(X) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1P} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{P1} & \sigma_{P2} & \ldots & \sigma_{PP} \end{bmatrix}$$

Because of $\sigma_{ik} = E(X_i-\mu_i)(X_k-\mu_k) = \sigma_{ki}$, it is convenient to write the above matrix as

$$\Sigma = E(X-\mu)(X-\mu)' = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1P} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{P} & \ldots & \sigma_{PP} \end{bmatrix}$$

9.3.1. Example (Computing the covariance matrix)

Find the covariance matrix for the two random variables $X_1$and X_2introduced in Example 9.2.1. When their joint probability function $p_{1,2}(x_1,\ x_2)$is represented by the entries in the body of the following table:

	x_2		
x_1	0	1	$p_1(x_1)$
-1	0.24	0. 06	0.3
0	0.16	0.14	0.3
1	0.40	0. 00	0.4
$p_1(x_1)$	0.8	0.2	1

Solution:

We have already shown that $\mu_1 = E(X_1) = 0.1$ and $\mu_2 = E(X_2) = 0.2$ (See Example 9.2.1.) In addition,

$$\sigma_{11} = E(X_1 - \mu_1)^2 = \sum_{all\, x_1} (x_1 - 0.1)^2\, p_1(x_1)$$

$$\sigma_{11} = (-1 - 0.1)^2(0.3) + (0 - 0.1)^2(0.3) + (1 - 0.1)^2(0.4) = 0.69$$

$$\sigma_{22} = E(X_2 - \mu_2)^2 = \sum_{all\, x_2} (x_1 - 0.2)^2\, p_2(x_2)$$

$$\sigma_{22} = (0 - 0.2)^2(0.8) + (1 - 0.2)^2(0.2) = 0.16$$

$$\sigma_{12} = E(X_1 - \mu_1)(X_2 - \mu_2) = \sum_{all\ pairs\ (x_1,x_2)} (x_1 - 0.1)(x_2 - 0.2)\, p_{12}\,(x_1, x_2)$$

$$\sigma_{12} = (-1 - 0.1)(0 - 0.2)(0.24) + \ (-1 - 0.1)(1 - 0.2)(0.06)$$
$$+ \cdots + (1 - 0.1)(1 - 0.2)(0.00) = -0.08$$

$$\sigma_{21} = E(X_2 - \mu_2)(X_1 - \mu_1) = E(X_1 - \mu_1)(X_2 - \mu_2) = \sigma_{12} = -0.08$$

Consequently, with $X' = [X_1, X_2]$

$$\mu = E(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix}$$

$$\Sigma = E(X - \mu)(X - \mu)'$$

$$\Sigma = E\begin{bmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} E(X_1 - \mu_1)^2 & E(X_1 - \mu_1)(X_2 - \mu_2) \\ E(X_2 - \mu_2)(X_1 - \mu_1) & E(X_2 - \mu_2)^2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 0.69 & -0.08 \\ -0.08 & 0.16 \end{bmatrix}$$

9.3.2. Note

The computation of means, variances, and covariances for *discrete* random variables involves summation (as in Examples 9.2.1. and 9.3.1.), while analogous computations for *continuous* random variables involve integration.

We shall refer to μ and Σ as the *population mean* (vector) and *population variance-covariance* (matrix), respectively.

The multivariate normal distribution is completely specified once the mean vector $\mu$and variance-covariance matrix Σ are given, so it is not surprising that these quantities play an important role in many multivariate procedures.

It is frequently informative to separate the information contained in variances σ_{ii} from that contained in measures of association and, in particular, the measure of association known as the *population correlation coefficient* ρ_{ik}.

The correlation coefficient $\rho_{ik}$is defined in terms of the covariance $\sigma_{ik}$and variances $\sigma_{ii}$and σ_{kk} as $\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}}$

The correlation coefficient measures the amount of *linear* association between the random variables X_i and X_k.

Let the Population correlation matrix be the $p \times p$ symmetric matrix

$$\rho = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{bmatrix}$$

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1P} \\ \rho_{12} & 1 & \cdots & \rho_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2P} & \cdots & 1 \end{bmatrix}$$

and let the $p \times p$ standard matrix be

$$V^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

Then

$$V^{\frac{1}{2}} \rho V^{\frac{1}{2}} = \Sigma \quad \text{and} \quad \rho = \left(V^{\frac{1}{2}}\right)^{-1} \Sigma \left(V^{\frac{1}{2}}\right)^{-1}$$

That is Σ can be obtained from $V^{1/2}$ and ρ, whereas ρcan be obtained from Σ. Moreover, the expression of these relationships in terms of matrix operations allows the calculations to be conveniently implemented on a computer.

9.3.3 Example (Computing the correlation matrix from the covariance matrix)

Suppose

$$\Sigma = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix}$$ Obtain $V^{1/2}$ and ρ.

Solution:

$$V^{\frac{1}{2}} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & 0 \\ 0 & \sqrt{\sigma_{22}} & 0 \\ 0 & 0 & \sqrt{\sigma_{33}} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

$$\left(V^{\frac{1}{2}}\right)^{-1} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{5} \end{bmatrix}$$

The correlation matrix p is given by

$$\left(V^{\frac{1}{2}}\right)^{-1} \Sigma \left(V^{\frac{1}{2}}\right)^{-1} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{5} \end{bmatrix} \begin{bmatrix} 4 & 1 & 2 \\ 1 & 9 & -1 \\ 2 & -3 & 25 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{5} \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{6} & \frac{1}{5} \\ \frac{1}{6} & 1 & -\frac{1}{5} \\ \frac{1}{5} & -\frac{1}{5} & 1 \end{bmatrix}$$

9.4. Partitioning the Covariance Matrix

The characteristics measured on individual trials will fall naturally into two or more groups. As examples, consider measurements of variables representing consumption and income or variables representing personality traits and physical characteristics. One approach to handling these situations is to let the characteristics defining the distinct groups be subsets of the *total* collection of characteristics. If the total collection is represented by a *(p* × 1)-dimensional random vector X, the subsets can be regarded as components of X and can be sorted by partitioning X.

In general, we can partition the *p* characteristics contained in the *p* × 1 random vector X into, for instance, two groups of size *q* and *p −q,* respectively. For example, we can write

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_q \\ \cdots\cdots\cdots \\ X_{q+1} \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} X^{(1)} \\ \cdots\cdots\cdots \\ X^{(2)} \end{bmatrix} \text{ and } \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_q \\ \cdots\cdots\cdots \\ \mu_{q+1} \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \mu^{(1)} \\ \cdots\cdots\cdots \\ \mu^{(2)} \end{bmatrix}$$

From the definitions of their transpose and matrix multiplication

$$(X^{(1)}-\mu^{(1)})(X^{(2)}-\mu^{(2)})' = \begin{bmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ \vdots \\ X_q-\mu_q \end{bmatrix} [X_{q+1}-\mu_{q+1} \quad X_{q+2}-\mu_{q+2} \quad \dots\dots \quad X_p-\mu_p]$$

$$(X^{(1)}-\mu^{(1)})(X^{(2)}-\mu^{(2)}) = \begin{bmatrix} (X_1-\mu_1)(X_{q+1}-\mu_{q+1}) & (X_1-\mu_1)(X_{q+2}-\mu_{q+2}) & \dots & (X_1-\mu_1)(X_p-\mu_p) \\ (X_2-\mu_2)(X_{q+1}-\mu_{q+1}) & (X_2-\mu_2)(X_{q+2}-\mu_{q+2}) & \dots & (X_2-\mu_2)(X_p-\mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (X_q-\mu_q)(X_{q+1}-\mu_{q+1}) & (X_q-\mu_q)(X_{q+2}-\mu_{q+2}) & \dots & (X_q-\mu_q)(X_p-\mu_p) \end{bmatrix}$$

Upon taking the expectation of the matrix $(X^{(1)}-\mu^{(1)})(X^{(2)}-\mu^{(2)})'$, we get

$$E(X^{(1)}-\mu^{(1)})(X^{(2)}-\mu^{(2)})' = \begin{bmatrix} \sigma_{1,q+1} & \sigma_{1,q+1} & \cdots & \sigma_{1,p} \\ \sigma_{2,q+1} & \sigma_{2,q+2} & \cdots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q,q+1} & \sigma_{q,q+2} & \dots & \sigma_{q,p} \end{bmatrix} = \Sigma_{12}$$

Which gives all the covariances $\sigma_{ij}, i = 1,2,\dots,q, j = q = 1, q+2, \dots, p,$ between a component of $X^{(1)}$ and a component of $X^{(2)}$.

The matrix Σ_{12} is not necessarily symmetric or even square.

With help of Partitioning, we can get

$$= \begin{bmatrix} \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1q} \\ \vdots & \ddots & \vdots \\ \sigma_{q1} & \cdots & \sigma_{qq} \end{pmatrix} & \begin{pmatrix} \sigma_{1,q+1} & \cdots & \sigma_{1,p} \\ \vdots & \ddots & \vdots \\ \sigma_{q,q+1} & \cdots & \sigma_{q,p} \end{pmatrix} \\ \begin{pmatrix} \sigma_{q+1,1} & \cdots & \sigma_{q+1,q} \\ \vdots & \ddots & \vdots \\ \sigma_{p,1} & \cdots & \sigma_{p,q} \end{pmatrix} & \begin{pmatrix} \sigma_{q+1,p+1} & \cdots & \sigma_{q+1,p} \\ \vdots & \ddots & \vdots \\ \sigma_{p,q+1} & \cdots & \sigma_{pp} \end{pmatrix} \end{bmatrix}$$

$$(X-\mu)(X-\mu)' = \begin{bmatrix} \underset{(q\times 1)}{(X^{(1)}-\mu^{(1)})}\underset{(1\times q)}{(X^{(1)}-\mu^{(1)})'} & \underset{(q\times 1)}{(X^{(1)}-\mu^{(1)})}\underset{(1\times(p-q))}{(X^{(2)}-\mu^{(2)})'} \\ \underset{((p-q)\times 1)}{(X^{(2)}-\mu^{(2)})}\underset{(1\times q)}{(X^{(1)}-\mu^{(1)})'} & \underset{((p-q)\times 1)}{(X^{(2)}-\mu^{(2)})}\underset{(1\times(p-q))}{(X^{(2)}-\mu^{(2)})'} \end{bmatrix}$$

$$\underset{(p\times p)}{\Sigma} = E(X-\mu)(X-\mu)' = \begin{array}{c} \\ \mathrm{q} \\ \mathrm{p-q} \end{array} \begin{array}{c} \mathrm{p} \qquad \mathrm{p-q} \\ \begin{bmatrix} (\Sigma_{11}) & (\Sigma_{12}) \\ (\Sigma_{21}) & (\Sigma_{22}) \end{bmatrix} \\ p\times p \end{array}$$

$$= \begin{bmatrix} \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1q} \\ \vdots & \ddots & \vdots \\ \sigma_{q1} & \cdots & \sigma_{qq} \end{pmatrix} & \begin{pmatrix} \sigma_{1,q+1} & \cdots & \sigma_{1,p} \\ \vdots & \ddots & \vdots \\ \sigma_{q,q+1} & \cdots & \sigma_{q,p} \end{pmatrix} \\ \begin{pmatrix} \sigma_{q+1,1} & \cdots & \sigma_{q+1,q} \\ \vdots & \ddots & \vdots \\ \sigma_{p,1} & \cdots & \sigma_{p,q} \end{pmatrix} & \begin{pmatrix} \sigma_{q+1,p+1} & \cdots & \sigma_{q+1,p} \\ \vdots & \ddots & \vdots \\ \sigma_{p,q+1} & \cdots & \sigma_{pp} \end{pmatrix} \end{bmatrix}$$

Note that $\Sigma_{12} = \Sigma'_{21}$. The covariance matrix of $X^{(1)}$ is Σ_{11}, that of $X^{(2)}$ is Σ_{22}, and that of elements from $X^{(1)}$ and $X^{(2)}$ is Σ_{12} or Σ_{21}.

It is convenient to use the $Cov(X^{(1)}, X^{(2)})$ notation where $Cov(X^{(1)}, X^{(2)}) = \Sigma_{12}$.is a matrix containing all the covariances between a component of $X^{(1)}$ and a component of $X^{(2)}$.

The Mean Vector and Covariance Matrix for linear Combinations of Random Variables

Recall that if a single random variable, such as X_1, is multiplied by a constant c, then $E(cX_1) = cE(X_1) = c\mu_1$ and $Var(cX_1) = E(cX_1 - c\mu_1) = c^2 var(X_1) = c^2\sigma_{11}$.

If X_2 is a second random variable and a and b are constants, then, using additional properties of expectation, we get

$$Cov(aX_1, bX_2) = E(aX_1 - a\mu_1)(bX_2 - b\mu_2) = abE(X_1 - \mu_1)(X_2 - \mu_2) = abCov(aX_1, X_2)$$
$$= ab\sigma_{12}$$

Finally, for the linear combination $aX_1 + bX_2$, we have

$$E(aX_1 + bX_2) = aE(X_1) + bE(X_2) = a\mu_1 + b\mu_2$$

$$Var(aX_1 + bX_2) = E[(aX_1 + bX_2) - (a\mu_1 + b\mu_2)]^2$$

$$Var(aX_1 + bX_2) = E[a(X_1 - \mu_1) + b(X_2 - \mu_2)]^2$$

$$Var(aX_1 + bX_2) = E[a^2(X_1 - \mu_1)^2 + b^2(X_2 - \mu_2)^2 + 2ab(X_1 - \mu_1)(X_2 - \mu_2)]$$

$$Var(aX_1 + bX_2) = a^2Var(X_1) + b^2Var(X_2) + 2abCov(X_1, X_2)$$

$$Var(aX_1 + bX_2) = a^2\sigma_{11} + b^2\sigma_{22} + 2ab\sigma_{12}$$

With $c' = [a, b]$, $aX_1 + bX_2$ can be written as

$$[a \quad b]\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = c'X$$

Similarly, $E(aX_1 + bX_2) = aE(X_1) + bE(X_2) = a\mu_1 + b\mu_2$ can be expressed as

$$[a \quad b]\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = c'\mu$$

If we let $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$ be the variance-covariance matrix of X, then we have

$$Var(aX_1 + bX_2) = Var(c'X) = c' \Sigma c$$

Since $c' \Sigma c = [a \quad b]\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = a^2\sigma_{11} + 2ab\sigma_{12} + b^2\sigma_{22}$

The preceding results can be extended to a linear combination of p random variables:

The linear combination $c'X = c_1X_1 + \cdots + c_pX_p$ has

Mean = $E(c'X) = c'\mu$

Variance = $Var(c'X) = c'\Sigma c$

Where $\mu = E(X)$ and $\Sigma = Cov(X)$

In general, consider the q linear combinations of the p random variables $X_1, \dots, X_p$:

$Z_1 = c_{11}X_1 + c_{12}X_2 + \cdots + c_{1p}X_p$

$Z_2 = c_{21}X_1 + c_{22}X_2 + \cdots + c_{2p}X_p$

$\vdots$

$Z_q = c_{q1}X_1 + c_{q2}X_2 + \cdots + c_{qp}X_p$

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_q \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1P} \\ c_{21} & c_{22} & \cdots & c_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ c_{q1} & c_{q2} & \cdots & c_{qp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = CX$$

The linear combinations $Z = CX$ have

$\mu_z = E(Z) = E(CX) = C_{\mu x}$

$\Sigma_z = Cov(Z) = Cov(CX) = C\Sigma_X C'$

Where μ_x and Σ_x are the mean vector and variance-covariance matrix of X, respectively.

9.4.1. Example (Means and covariances of linear combinations)

Let $X' = [X_1, X_2]$ be a random vector with mean vector $\mu'_x = [\mu_1, \mu_2]$ and variance-covariance matrix $\Sigma_x = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$

Solution:

Find the mean vector and covariance matrix for the linear combinations

$Z_1 = X_1 - X_2$

$Z_2 = X_1 + X_2$

$$z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = CX$$

in terms of μ_X and Σ_X

$$\mu_Z = E(Z) = C_{\mu x} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \mu_1 - \mu_2 \\ \mu_1 + \mu_2 \end{bmatrix}$$

$$\Sigma_z = Cov(Z) = C\Sigma_X C' = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

$$\Sigma_z = \begin{bmatrix} \sigma_{11} - 2\sigma_{12} + \sigma_{22} & \sigma_{11} - \sigma_{22} \\ \sigma_{11} - \sigma_{22} & \sigma_{11} + 2\sigma_{12} + \sigma_{22} \end{bmatrix}$$

9.4.2. Note

If $\sigma_{11} = \sigma_{22}$, that is, if X_1 and X_2 have equal variances, the off-diagonal terms in Σ_z vanish. This demonstrates the well-known result that the sum and difference of two random variables with identical variances are uncorrelated.

9.5. Partitioning the sample mean vector and Covariance matrix

Let $\bar{x}' = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p]$ be the vector of sample averages constructed from n observations on p variables $X_1, X_2, \ldots, X_p$, and let

$$S_n = \begin{pmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{1p} & \cdots & s_{pp} \end{pmatrix}$$

$$S_n = \begin{bmatrix} \frac{1}{n}\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)^2 & \cdots & \frac{1}{n}\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)(x_{jp} - \bar{x}_p) \\ \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)(x_{jp} - \bar{x}_p) & \cdots & \frac{1}{n}\sum_{j=1}^{n}(x_{jp} - \bar{x}_p)^2 \end{bmatrix}$$

be the corresponding sample variance-covariance matrix.

The sample mean vector and the covariance matrix can be partitioned in order to distinguish quantities corresponding to groups of variables. Thus,

$$\underset{(p\times 1)}{\bar{x}} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_q \\ \cdots \\ \bar{x}_{q+1} \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} \bar{x}^{(1)} \\ \cdots \\ \bar{x}^{(2)} \end{bmatrix} \text{ and}$$

$$\underset{(p\times p)}{S_n} = \begin{bmatrix} \begin{pmatrix} s_{11} & \cdots & s_{1q} \\ \vdots & \ddots & \vdots \\ s_{q1} & \cdots & s_{qq} \end{pmatrix} & \begin{pmatrix} s_{1,q+1} & \cdots & s_{1,p} \\ \vdots & \ddots & \vdots \\ s_{q,q+1} & \cdots & s_{q,p} \end{pmatrix} \\ \begin{pmatrix} s_{q+1,1} & \cdots & s_{q+1,q} \\ \vdots & \ddots & \vdots \\ s_{p,1} & \cdots & s_{p,q} \end{pmatrix} & \begin{pmatrix} s_{q+1,p+1} & \cdots & s_{q+1,p} \\ \vdots & \ddots & \vdots \\ s_{p,q+1} & \cdots & s_{pp} \end{pmatrix} \end{bmatrix} = \begin{bmatrix} (S_{11}) & (S_{12}) \\ (S_{21}) & (S_{22}) \end{bmatrix}$$

Where $\bar{x}^{(1)}$ and $\bar{x}^{(2)}$ are the sample mean vectors constructed from observations $\bar{x}^{(1)} = [x_1, \ldots, x_q]$ and $\bar{x}^{(2)} = [x_{q+1}, \ldots, x_p]'$, respectively; S_{11} is the sample covariance matrix computed from observations $\bar{x}^{(1)}$; S_{22} is the sample covariance matrix computed from observations $\bar{x}^{(2)}$; and $S_{12} = S_{21}$ is the sample covariance matrix for elements of $\bar{x}^{(1)}$ and elements of $\bar{x}^{(2)}$.

Let Us Sum Up

In this unit we studied the random variables, random matrices, mean vectors, covariance matrices, partitioning the covariance matrix, partitioning the sample mean vector and covariance matrix.

Check Your Progress

1. $Cov(x_1, x_2) =$______if x_1 and x_2 are independent.

2. Let X be a random variable and let A and B be conformable matrices of constants. Then $E(AXB) =$_____.

3. The P continuous random variables $X_1, X_2, \dots, X_p$ are mutually statistically independent if their joint density can be factored as__________.

Glossaries

Random vector: It is a vector whose elements are random variables.

Random matrix: It is a matrix whose elements are random variables.

Correlation coefficient: It measures the amount of linear association between the random variables.

Suggested Readings

1. Johnson. R. A. and Wichern. D. W., "Applied Multivariate Statistical Analysis", Pearson Education Asia, Sixth Edition, 2007.

Answers to Check Your Progress

1. Zero

2. $AE(X)B$

3.$f_{1.2......p}(x_1 . x_2, \dots, x_p) = f_1(x_1) f_2(x_2) \dots f_p(x_p)$ for all p-tuples $(x_1, x_2, \dots, x_p)$

Unit – 10

The Multivariate Normal Distribution

Objectives

After Studying this Unit, the student will be able to

- Explain the multivariate normal distribution
- Demonstrate the concept of the Multivariate Normal Density and its properties

Overview

In this unit, we will study the concept of multivariate normal distribution and multivariate normal density and its properties.

10.1. Introduction

A generalization of the bell-shaped normal density to several dimensions plays a fundamental role in multivariate analysis. Most of the techniques encountered in this unit are based on the assumption that the data were generated from a multivariate normal distribution. While real data are never exactly multivariate normal, the normal density is a useful approximation to the true population distribution.

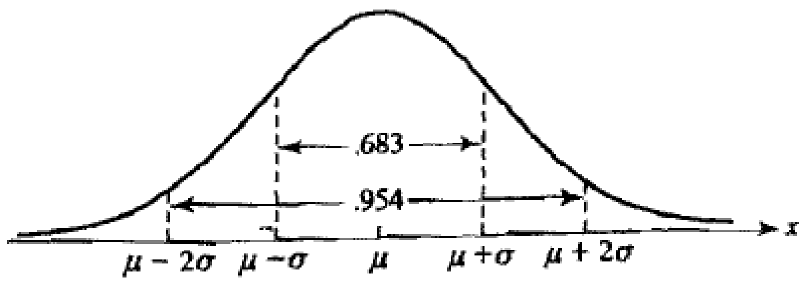
One advantage of the multivariate normal distribution is mathematically tractable and nice results can be obtained. The normal distributions are useful for two reasons: First, the normal distribution serves as a bona fide population model in some instances; Second, the sampling distributions of many multivariate statistics are approximately normal, regardless of the form of the parent population, because of central limit effect.

Many real-world problems fall naturally within the framework of normal theory. The importance of the normal distribution rests on its dual role as both population model for certain natural phenomena and approximate sampling distribution for many statistics.

10.2. Multivariate Normal Density and its properties

The multivariate normal density is a generalization of the univariate normal density to $p \geq 2$ dimensions. Recall that the univariate normal distribution, with mean $\mu$and variance σ^2,has the probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-[(x-\mu)/\sigma]^2/2}, -\infty < x < \infty$$



The above figure is a Normal density with mean μ and variance σ^2 and selected areas under the curve

A plot of this function yields the familiar bell-shaped curve shown in the above figure. Also shown in the figure are approximate areas under the curve within ± 1 standard deviations and ± 2 standard deviations of the mean. These areas represent probabilities, and thus, for the normal random variable X.

$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68$

$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95$

It is convenient to denote the normal density function with mean $\mu$and variance σ^2 by $N(\mu, \sigma^2)$. Therefore, N(10, 4) refers to the function $f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-[(x-\mu)/\sigma]^2/2}, -\infty < x < \infty$ with $\mu = 2$ and $\sigma = 2$.

The term $\left(\frac{x-\mu}{\sigma}\right)^2 = (x - \mu)(\sigma^2)^{-1}(x - \mu)$ is the exponent of the univariate normal density function This can be generalized for a $p \times 1$ vector x of observations on several variables as $(x - \mu)'\Sigma^{-1}(x - \mu)$.

The $p \times 1$ vector μ represents the expected value of the random vector X, and the $p \times p$ matrixΣ is the variance-covariance matrix of X. We shall assume that the symmetric matrix Σ is positive definite, so the expression $(x - \mu)'\Sigma^{-1}(x - \mu)$ is the square of the generalized distance from x to μ.

The multivariate normal density is obtained by replacing the univariate distance in the function $\left(\frac{x-\mu}{)\sigma}\right)^2 = (x - \mu)(\sigma^2)^{-1}(x - \mu)$ by the multivariate generalized distance of $(x - \mu)'\Sigma^{-1}(x - \mu)$ in the density function of $(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-[(x-\mu)/\sigma]^2/2}, -\infty < x < \infty$.

When this replacement is made, the univariate normalizing constant $(2\pi)^{-1/2}(\sigma^2)^{-1/2}$ must be changed to a more general constant that makes the volume under the surface of the multivariate density function unity for any p. This is necessary because, in the multivariate case, the probabilities are represented by volumes under the surface over regions defined by intervals of the x_i values. Consequently, a p $-$dimensional normal density for the random vector $X' = [X_1, X_2, \dots, X_p]$ has the form
$f(x) = \frac{1}{(2\pi)^{p/2}[\Sigma]^{1/2}} e^{-(x-\mu)\Sigma^{-1}(x-\mu)/2}$, where $-\infty < x_i < \infty, i = 1,2, \dots p$.

We shall denote this p-dimensional normal density by $N_p(\mu, \Sigma)$ which is analogous to the normal density in the univariate case.

10.2.1. Example (Bivariate normal density)

Evaluate the $p = 2$ $-$variate normal density in terms of the individual parameters $\mu_1 = E(X_1), \mu_2 = E(X_2)$. $\sigma_{11} = Var(X_1), \sigma_{22} = Var(X_2)$, and $\rho_{12} = \frac{\sigma_{12}}{(\sqrt{\sigma_{11}}\sqrt{\sigma_{22}})} = Corr(X_1, X_2)$.

Solution

The inverse of the covariance matrix $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$ is $\Sigma^{-1} = \frac{1}{\sigma_{11}\sigma_{22}} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}$

Introducing the correlation coefficient ρ_{12} by writing $\sigma_{12} = \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}$, we obtain $\sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1-\rho_{12}^2)$, and the squared distance becomes

$$(x-\mu)'\Sigma^{-1}(x-\mu) = [x_1-\mu_1, x_2-\mu_2]\begin{bmatrix} \sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} \\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11} \end{bmatrix}\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}$$

$$(x-\mu)'\Sigma^{-1}(x-\mu) = \frac{\sigma_{22}(x_1-\mu_1)^2 + \sigma_{11}(x_2-\mu_2)^2 - 2\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}(x_1-\mu_1)(x_2-\mu_2)}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}$$

$$(x-\mu)'\Sigma^{-1}(x-\mu) = \frac{1}{1-\rho_{12}^2}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]$$

The last expression is written in terms of the standardized values $\frac{x1-\mu1}{\sqrt{\sigma_{11}}}$ and $\frac{x2-\mu2}{\sqrt{\sigma_{22}}}$

Next, since $|\Sigma| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1-\rho_{12}^2)$. We can substitute for Σ^{-1} and $|\Sigma|$ in $f(x) = \frac{1}{(2\pi)^{p/2}[\Sigma]^{1/2}} e^{-(x-\mu)\Sigma^{-1}(x-\mu)/2}$ to get the expression for the bivariate ($p = 2$) normal density involving the individual parameters $\mu_1, \mu_2, \sigma_{11}, \sigma_{22}$ *and* ρ_{12}

$$f(x_1,x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}} exp\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right\}$$

The above expression is somewhat unwidely, and the compact general form $f(x) = \frac{1}{(2\pi)^{p/2}[\Sigma]^{1/2}} e^{-(x-\mu)\Sigma^{-1}(x-\mu)/2}$ is more inofmative in many ways. On the other hand, the above expression is useful for discussing certain properties of the normal distribution.

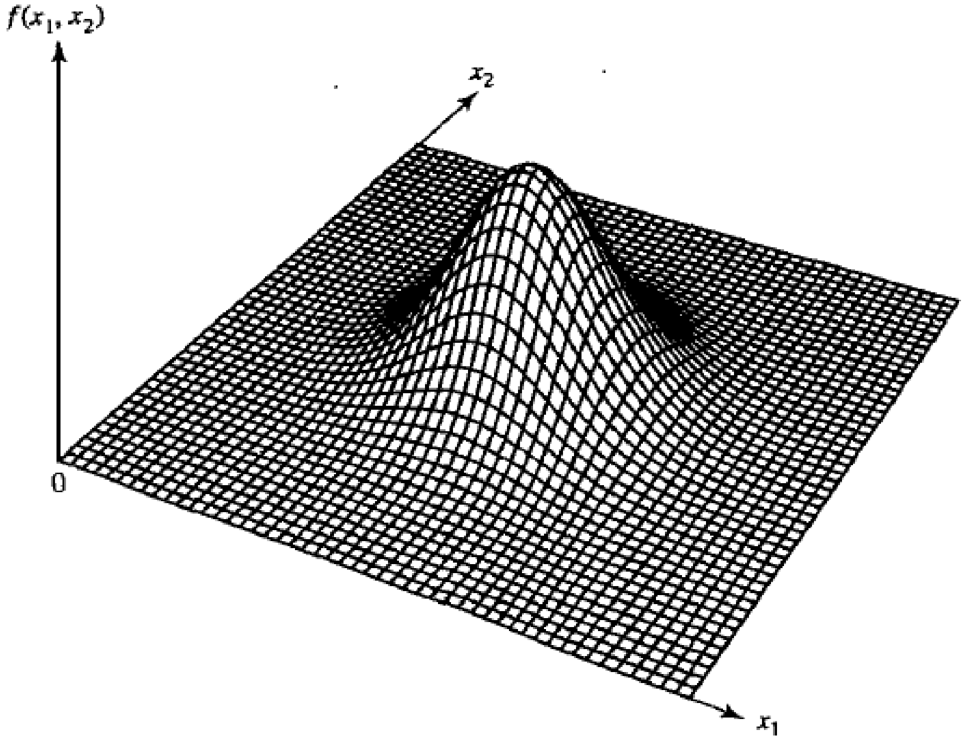
For example, if the randam variables X_1 *and* X_2 are uncorrelated, so that $\rho_{12} = 0$, the joint density can be written as the product of two univariate normal densities each of the form of $(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-[(x-\mu)/\sigma]^2/2}, -\infty < x < \infty$.
That is, $f(x_1,x_2) = f(x_1)f(x_2)$ and X_1 *and* X_2 are independent.

Two bivariate distributions with $\sigma_{11} = \sigma_{22}$ in the following figures.

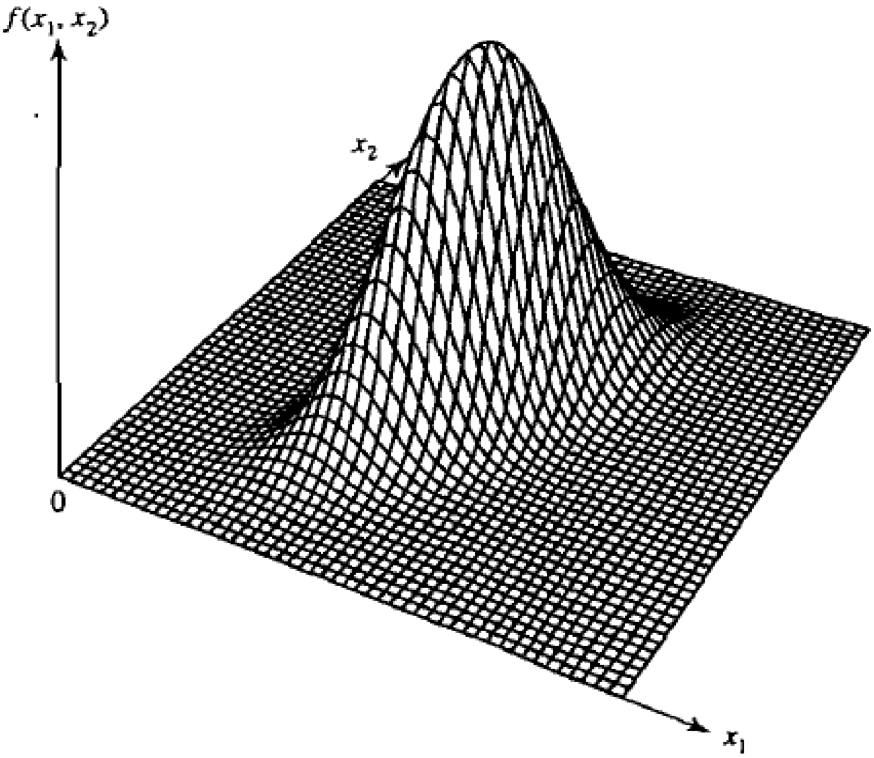
In Figure (a), X_1 *and* X_2 are independent $\rho_{12} = 0$.

In Figure (b) $\rho_{12} = 0.75$.

Notice how the presence of correlation causes the probability to concntrate along a line.



(a)



(b)

In the above two figures, Two bivariate normal distributions *(a)* $\sigma_{11} = \sigma_{22}$ *and* $\rho_{12} = 0$ *(b)* $\sigma_{11} = \sigma_{22}$ *and* $\rho_{12} = 0.75$

From the expression $f(x) = \frac{1}{(2\pi)^{p/2}[\Sigma]^{1/2}} e^{-(x-\mu)\Sigma^{-1}(x-\mu)/2}$ for the density of a p-dimensional normal variable, it should be clear that the paths of x values yielding a constant height for the density are ellipsoids. That is, the multivariate normal density is constant on surfaces where the square of the distance$(x - \mu)'\Sigma^{-1}(x - \mu)$ is constant. These paths are called *contours.*

Constant probability density contour = $\{all\ x\ such\ that\ (x-\mu)'\Sigma^{-1}(x-\mu) = c^2\}$

Constant probability density contour = surface of an ellipsoid centred at μ.

The axes of each ellipsoid of constant density are in the direction of the eigenvectors of $\Sigma^{-1}$and their lengths are proportional to the reciprocals of the square roots of the Eigen values of Σ^{-1}. Fortunately, we can avoid the calculation of Σ^{-1} when determining the axes, since these ellipsoids are also determined by the eigenvalues and eigenvectors of Σ.

10.2.2. Result

If Σ is positive definite, so that Σ^{-1} exists, then $\Sigma e = \lambda e$ imples $\Sigma^{-1}e = (\frac{1}{\lambda})\, e$ so (λ, e) is an eigen value - eigen vector pair for Σ corresponding to the pair $(\frac{1}{\lambda}, e)$ for c. Also, Σ^{-1} is positive definite.

Proof:

For Σ is positive definite and $e \neq 0$ an eigen vector, we have

$0 < e'\Sigma e = e'(\Sigma e) = e'(\lambda e)$

$e = \lambda\Sigma^{-1}e$, and division by $\lambda > 0$, we have

$$\Sigma^{-1}e = (\frac{1}{\lambda})\, e$$

Thus, $(\frac{1}{\lambda}, e)$ is an eigen value - eigen vector pair for Σ^{-1}. Also, for any $p \times 1x$

We know that $A^{-1} = PA^{-1}P' = \sum_{i=1}^{k} (\frac{1}{\lambda_i})\, e_i\, e_i'$

$$x\Sigma^{-1}x' = x' (\sum_{i=1}^{k} (\frac{1}{\lambda_i})\, e_i e_i')\, x$$

$$x\Sigma^{-1}x' = \sum_{i=1}^{k} (\frac{1}{\lambda_i})\, (x'e_i)^2 \geq 0$$

Since each term $\lambda_i^{-1}(x'e_i)^2$ is nonnegative. In addition, $x'e_i = 0$ for all I only if $x = 0$. So $x \neq 0$ implies that $\sum_{i=1}^{k} (\frac{1}{\lambda_i})\, (x'e_i)^2 > 0$ and therefore Σ^{-1} is positive definite.

The following summarizes these concepts:

Contours of constant density for the p −dimensional normal distribution are ellipsoids defined by x such that $(x-\mu)'\Sigma^{-1}(x-\mu)=c^2$.

These ellipsoids are centred at μ and have axes $\pm c\sqrt{\lambda_i}e_i$ where $\Sigma e_i=\lambda_i e_i$ for $i=1,2,\dots,p$.

A contour of constant density for a bivariate normal distribution with $\sigma_{11}=\sigma_{22}$ is obtained in the following example.

10.2.3. Example(contours of the bivariate normal density)

Obtain the axes of constant probability density contours for a bivariate normal distribution when $\sigma_{11}=\sigma_{22}$. From $(x-\mu)'\Sigma^{-1}(x-\mu)$ these axes given by the eigen values and eigenvectors of Σ.

Here $|\Sigma-\lambda I|=0$ becomes

$$0=\begin{vmatrix}\sigma_{11}-\lambda & \sigma_{12}\\ \sigma_{12} & \sigma_{11}-\lambda\end{vmatrix}=(\sigma_{11}-\lambda)^2-\sigma_{12}^2=(\lambda-\sigma_{11}-\sigma_{12})(\lambda-\sigma_{11}+\sigma_{12})$$

Consequently, the Eigen values are $\lambda_1=\sigma_{11}+\sigma_{12}$ and $\lambda_2=\sigma_{11}-\sigma_{12}$. The eigen vector e_1 is determined from

$$\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{bmatrix}\begin{bmatrix}e_1\\ e_2\end{bmatrix}=(\sigma_{11}+\sigma_{12})\begin{bmatrix}e_1\\ e_2\end{bmatrix}$$

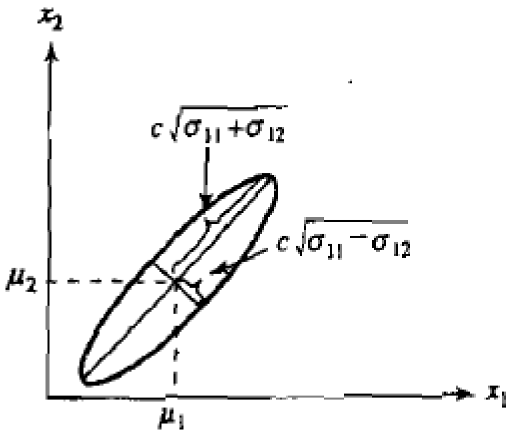
$$\sigma_{11}e_1+\sigma_{12}e_2=(\sigma_{11}+\sigma_{12})e_1$$

$$\sigma_{12}e_1+\sigma_{11}e_2=(\sigma_{11}+\sigma_{12})e_2$$

These equations imply that $e_1=e_2$ and after normalization, the firstr eign value – eigen vector pair is $\lambda_1=\sigma_{11}+\sigma_{12}$, $e_1=[\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}]$; the second eigen value – eigen vector pair is $\lambda_2=\sigma_{11}-\sigma_{12}$; $e_1=[\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}}]$

When the covariance σ_{12}or correlation$\rho_{12}$is positive, $\lambda_1=\sigma_{11}+\sigma_{12}$is the *largest* eigenvalue, and its associated eigenvector $e_1'=[\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}]$lies along the 45° line through the point $\mu'=[\mu_1,\mu_2]$. This is true for any positive value of the covariance (correlation). Since the axes of the constant-density ellipses are given by $\pm c\sqrt{\lambda_1}e_1$ and $\pm c\sqrt{\lambda_2}e_2$,and the eigenvectors each have length unity, the major axis will be associated with the largest eigenvalue. For positively correlated normal random variable, then, the *major* axis of the constant-density ellipses will be along the 45° line through μ.

The following figure is a constant-density contour for a bivariate normal distribution with $\sigma_{11}=\sigma_{22}$ and $\sigma_{12}>0$ or $\rho_{12}>0$.



When the covariance or correlation is negative, $\lambda_2 = \sigma_{11} + \sigma_{12}$will be the largest eigenvalue, and the major axes of the constant-density ellipses will lie along a line at right angles to the 45^o line through μ. These results are ture only for $\sigma_{11} = \sigma_{22}$.

To summarize the axes of the ellipses of constant denstiy for a bivariate normal distribution with $\sigma_{11} = \sigma_{22}$ are determined by

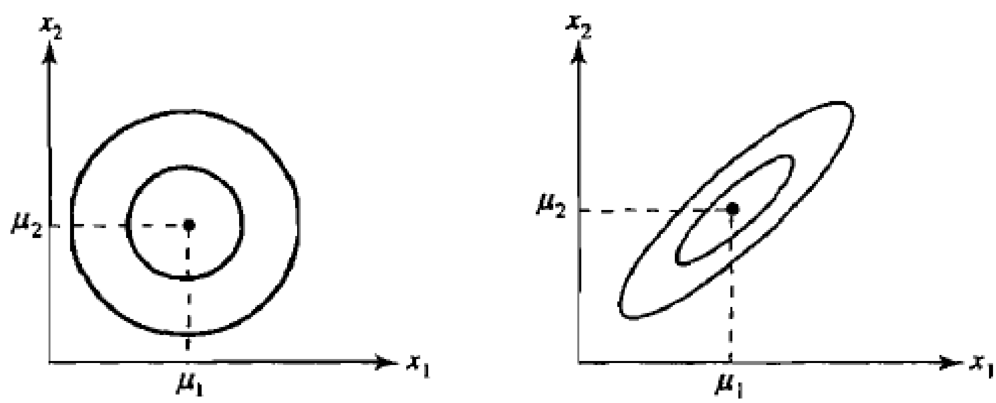
$$\pm c\sqrt{\sigma_{11} + \sigma_{12}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \text{ and } \pm c\sqrt{\sigma_{11} - \sigma_{12}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$$

From the result ,$(x-\mu)'\Sigma^{-1}(x-\mu) = c^2$ that the choice $c^2 = \chi^2_p(\alpha)$, $\chi^2_p(\alpha)$ is the upper $(100\alpha)th$ percentile of a chi-square distibution with p degress of freedom, leads to contours that contian $(1-\alpha) \times 100\%$ of the probability, specifically, the following is true for a p-dimensional normal distibution.

The solid ellipsoid of x values satisfying $(x-\mu)'\Sigma^{-1}(x-\mu) \le \chi^2_p(\alpha)$ has probability $1-\alpha$.

The constant-density contours contianing 50% and 90% of the probaility under the bivariate normal surfaces.

The following figure is the 50% and 90% contours for the bivariate normal distibutions.



The p-variate normal density in $f(x) = \frac{1}{(2)^{p/2}[\Sigma]^{1/2}} e^{-(x-\mu)}\Sigma^{-1}(x - \mu)/2$ has a maximum value when the squared distance in$(x-\mu)'\Sigma^{-1}(x-\mu)$. is zero - that is, when x = μ. Thus, $\mu$is the point of maximum density, or *mode,* as well as the expected value of X, or *mean.* The fact that $\mu$is the mean of the multivariate normal distribution follows from the symmetry exhibited by the constant-density contours. These contours are cantered, or balanced, at μ.

Let Us Sum Up

In this unit we studied the concept of multivariate normal distribution and multivariate normal density and its properties.

Check Your Progress

1. The multivariate normal density is a generalization of the univariate normal density to ______dimensions.

2. The normal density function with mean $\mu$and variance σ^2 is denoted by_____.

3. The $p \times 1$ vector μ represents________.

4. The $p \times p$ matrix Σ represents_________.

Glossaries

Univariate normal distribution: It is defined by two parameters mean, which is expected value of the distribution and standard deviation, which corresponds to the expected square deviation from the mean.

Bivariate normal distribution: It is made up of two independent random variables. The two variables in a bivariate normal are both normally distributed and they have normal distribution when both are added together.

Multivariate normal distribution: It is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions.

Suggested Readings

1. Johnson. R. A. and Wichern. D. W., "Applied Multivariate Statistical Analysis", Pearson Education Asia, Sixth Edition, 2007.

Answers to Check Your Progress

1. $p \geq 2$
2. $N(\mu, \sigma^2)$
3. The expected value of the random vector X
4. The variance-covariance matrix of X.

Unit – 11

Principal Components

Objectives

After Studying this Unit, the student will be able to

- Explain the principal components
- Summarize the uses of population principal components

Overview

In this unit, we will study the concept of the principal components and the population principal components.

11.1. Introduction

A principal component analysis is concerned with explaining the variance-covariance structure of a set of variables through a few *linear* combinations of these variables. Its general objectives are (1) data reduction and (2) interpretation.

Although *p* components are required to reproduce the total system variability, often much of this variability can be accounted for by a small number *k* of the principal components. If so, there is (almost) as much information in the *k* components as there is in the original *p* variables. The *k* principal components can then replace the initial *p* variables, and the original data set, consisting of *n* measurements on *p* variables, is reduced to a data set consisting of *n* measurements on *k* principal components.

An analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result.

Analyses of principal components are more of a means to an end rather than an end in themselves, because they frequently serve as intermediate steps in much larger investigations.

11.2. Population Principal Components

Algebraically, principal components are particular linear combinations of the *p* random variables$X_1,\ X_2,\ \dots\ ,\ X_P$. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $X_1,\ X_2,\ \dots\ ,\ X_P$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariance structure.

Principal components depend solely on the covariance matrix Σ or the correlation matrix ρ of$X_1,\ X_2,\ \dots\ ,\ X_P$. Their development does not require a multivariate normal assumption. On the other hand, principal components derived for multivariate normal populations have useful interpretations in terms of the constant density ellipsoids. Further, inferences can be made from the sample components when the population is multivariate normal.

Let the random vector $X' = [X_1, X_2, \dots, X_P]$ have the covariance matrix Σ with Eigen values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Consider the linear combinations

$Y_1 = a_1' X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$

$Y_2 = a_2' X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$

$\vdots$

$Y_p = a_p' X = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p$

Then by using

The linear combinations $Z = CX$ we have

$\mu_z = E(Z) = E(CX) = C_{\mu_X}$

$\sum_Z = Cov(Z) = Cov(CX) = C\sum_X C'$

We obtain

$Var(Y_i) = a_i' \sum a_i \qquad i = 1,2, \dots, p$

$Cov(Y_i, Y_k) = a_i' \sum a_k \qquad i, k = 1,2, \dots, p$

The Principal components are those uncorrelated linear combinations $Y_1, Y_2, \dots, Y_P$ whose variances in $Var(Y_i) = a_i' \sum a_i, i = 1,2 \dots, p$ are as large possible.

The first principal component is the linear combination with maximum variance. That is, it maximizes $Var(Y_1) = a_1' \sum a_1$. It is clear that $Var(Y_1) = a_1' \sum a_1$ can be increased by multiplying any a_1 by some constant. To eliminate this indeterminacy, it is convenient to restrict attention to coefficient vectors of unit length.

We define

First principal component = linear combination $a_1' X$ that maximizes $Var(a_1' X)$ subject to $a_1' a_1 = 1$

Second principal component = linear combination $a_2' X$ that maximizes $Var(a_2' X)$ subject to $a_2' a_2 = 1$ and $Cov(a_1' X, a_2' X) = 0$

At the i^{th} step

i^{th} principal component = linear combination $a_i' X$ that maximizes $Var(a_i' X)$ subject to $a_i' a_i = 1$ and $Cov(a_i' X, a_k' X) = 0$ for $k < i$

11.2.1. Result

Let Σ be the covariance matrix associated with the random vector $X' = [X_1, X_2, \dots, X_P]$. Let Σ have the eigenvalue-eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$ where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$. Then ith principal component is given by

$$Y_i = e_i' = e_{i1}X_1 + e_{i2}X_2 + \dots + e_{ip}X_p, \qquad i = 1,2,\dots,p$$

With these choices,

$$Var(Y_i) = e_i' \sum e_i = \lambda_i, \; i = 1,2,\dots,p$$

$$Cov(Y_i, Y_k) = e_i' \sum e_k \,, i \neq k$$

If some λ_i are equal, the choices of the corresponding coefficient vectorse_i, and hence Y_i are not unique.

Proof:

We know that, with $B = \Sigma$, that

$$\max_{a \neq 0} \frac{a' \Sigma a}{a'a} = \lambda_1 \text{ (Attained when } a = e_1\text{)}$$

But $e_1'e_1 = 1$ since the eigen vectors are normalized. Thus,

$$\max_{a \neq 0} \frac{a'\Sigma a}{a'a} = \lambda_1 = \frac{e_1' \Sigma e_1}{e_1' e_1} = e_1' \sum e_1 = Var(Y_1)$$

Similarly, we get

$$\max_{a \perp e_1, e_2, \dots e_k} \frac{a'\Sigma a}{a'a} = \lambda_{k+1}, \; k = 1,2,\dots,p-1$$

For the choice $a = e_{k+1}$ with $e_{k+1}'e_i = 0$, for $i = 1,2,\dots,k$ and $k = 1,2,\dots,p-1$

$$\frac{e_{k+1}' \sum e_{k+1}}{e_{k+1}' e_{k+1}} = e_{k+1}' \sum e_{K=1} = Var(Y_{k+1})$$

But $e_{k+1}' (\sum e_{k+1}) = \lambda_{k+1} e_{k+1}' e_{k+1} = \lambda_{k+1}$, So $Var(Y_{k+1}) = \lambda_{k+1}$.

It remains to show that e_i perpendicular to e_k. That is $e'e_{k_i} = 0, \; i \neq k$ gives $Cov(Y_i, \, Y_k) = 0$. Now, the eigen vectors of $\sum$ are orthogonal if all the eigen values $\lambda_1, \, \lambda_2, \, \dots \, , \, \lambda_P$ are distinct. If the eigen values are not all distinct, the eigen vectors corresponding to common eigen values may be chosen to be orthogonal. Therefore, for any two eigen vectors e_i and e_k, $e_i' e_k = 0, i \neq k$. Since $\sum e_k = \lambda_k e_k$, premultiplication by e_i', gives

$$Cov(Y_i, Y_k) = e_i' \sum e_k = e_i' \lambda_k e_k = \lambda_k e_i' e_k = 0 \text{ for any } i \neq k$$

From the above result, the principal components are uncorrelated and have variances equal to the Eigen values of Σ.

11.2.2. Result

Let $X' = [X_1, X_2, \dots, X_P]$ have covariance matrix Σ, with eigenvalue-eigenvecor pairs $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Let $Y_1 = e_1'X, Y_2 = e_2'X, \dots, Y_p = e_p'X$ be the principal components. Then

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} Var(Y_i)$$

Proof:

We know that$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = tr(\Sigma)$.

Also from

$$\underset{(k \times k)}{A} = \sum_{i=1}^{k} \lambda_i \underset{(k \times 1)}{e_i} \underset{(1 \times k)}{e_i'} = \underset{(k \times k)}{P} \underset{(k \times k)}{\Lambda} \underset{(k \times k)}{P'}$$

With $A = \Sigma$, we can write $\Sigma = P\Lambda P'$ where Λ ia the diagonal matrix of eigen values and $P = [e_1, e_2, \dots, e_p]$ so othat $PP' = P'P = I$.

$$tr(\Sigma) = tr(P\Lambda P') = tr(\Lambda P'P) = tr(A) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$$

Thus,

$$\sum_{i=1}^{p} Var(X_i) = tr(\Sigma) = tr(A) = \sum_{i=1}^{p} Var(Y_i)$$

Result. 11.2.2. Says that

Total population variance = $\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \lambda_1 + \lambda_2 + \cdots + \lambda_p$ and consequently, the proportion of total variance due to the k^{th} principal component is

$$\begin{pmatrix} Proportion\ of\ total \\ population\ variance \\ due\ to\ k^{th} \\ principal \\ component \end{pmatrix} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p},\ k = 1,2, \dots, p$$

If most (for instance 80 to 90) of the total population variance, for large p, can be attributed to the first one, two, or three components, then these components can “replace” the original p variables without much loss of information.

Each component of the coefficient vector $e_i' = [e_{i1}, e_{i2}, \dots, e_{ip}]$ also merits inspection. The magnitude of e_{ik} measures the importance of the k^{th} variable to the i^{th} principal component, irrespective of the other variables. In particular, e_{ik} is proportional to the correlation coefficient between Y_i and X_k.

11.2.3. Result

If $Y_1 = e_1'X, Y_2 = e_2'X, \dots, Y_p = e_p'X$ are the principal components obtained from the covariance matrix Σ then $\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$, $i, k = 1,2,\dots,p$ are the correlation coefficients between the components Y_i and the variables X_k. Here $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$ are the eigenvalue-eigenvector pairs for Σ.

Proof:

Set $a_k' = [0, \dots, 0, 1, 0, \dots 0]$ so that $X_k = a_k'X$ and $Cov(X_k, Y_i) = Cov(a_k'X, e_i'X) = a_k' \sum e_i$. Since $\sum e_i = \lambda_i e_i$, $Cov(X_k, Y_i) = a_k'\lambda_i e_i = \lambda_i e_{ik}$ Then $Var(Y_i) = \lambda_i$ and $Var(X_k) = \sigma_{kk}$ yield

$$\rho Y_i, X_k = \frac{Cov(Y_i, X_k)}{\sqrt{Var(Y_i)}\sqrt{Var(X_k)}} = \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \quad i, k = 1,2,\dots,p$$

11.2.4. Remark

Although the correlations of the varibales with the principal components often help to interpret the components, they measure only the univariate contribution of an individual X to a component Y. That is, they do not indicate the importance of an X to a component Y in the presence of the other X's. For this reason, some statisticians recommend that only the coefficients$e_{ik}$and not the correlations, be used to interpret the components. Although the coefficients and the correlations can lead to different rankings as measures of the importance of the variables to a given component, it is our experience that these rankings are often not *appreciably* different. In practice, variables with relatively large coefficients (in absolute value) tend to have relatively large correlations, so the two measures of importance, the first multivariate and the second univariate, frequently give similar results. We recommend that both the coefficients and the correlations be examined to help interpret the principal components.

The following hypothetical example illustrates the contents of Results 11.2.1, 11.2.2 and 11.2.3.

11.2.5. Example (Calculating the population principal components)

Suppose the random variables X_1, X_2 and X_3 have the covariance matrix

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

The Eigen value - Eigen vector pairs are

$\lambda_1 = 5.83, \qquad e_1' = [0.383, -0.924, 0]$

$\lambda_2 = 2.00, \qquad e_2' = [0, 0, 1]$

$\lambda_3 = 0.17, \qquad e_3' = [0.924, 0.383, 0]$

Therefore, the principal components become

$Y_1 = e_1'X = 0.383X_1 - 0.924X_2$

$Y_2 = e_2'X = X_3$

$Y_3 = e_3'X = 0.924X_1 + 0.383X_2$

,
The variable X_3 is one of the principal components, because it is uncorrelated with the other two variables.

We know that

$Var(Y_i) = e_i' \sum e_i = \lambda_i,\ i = 1,2, \ldots, p$

$Cov(Y_i, Y_k) = e_i' \sum e_k\ , i \neq k$ can be demonstrated from first principles.

For example,

$Var(Y_1) = Var(0.383X_1 - 0.924X_2)$

$Var(Y_1) = (0.383)^2 Var(X_1) + (-0.924)^2 Var(X_2) + 2(0.383)(-0.924)Cov(X_1, X_2)$

$Var(Y_1) = 0.147(1) + 0.854(5) - 0.708(-2) = 5.83 = \lambda_1$

$Cov(Y_1, Y_2) = Coc(0.383X_1 - 0.924X_2, X_3) = 0.383\ Cov(X_1, X_3) - 0.924\ Cov(X_2, X_3)$

$Cov(Y_1, Y_2) = 0.383(0) - 0.924(0) = 0$

$\sigma_{11} + \sigma_{22} + \sigma_{33} = 1 + 5 + 2 = \lambda_1 + \lambda_2 + \lambda_3 = 5.83 + 2.00 + 0.17 = 8$

The proportion of total variance accounted for by the first principal component is $\frac{\lambda_1}{(\lambda_1+\lambda_2+\lambda_3)} = \frac{5.83}{8} = 0.73$.

Further, the first two components account for a proportion $\frac{5.83+2}{8} = 0.98$ of the population variance. In this case, the components Y_1 and Y_2 could replace the original three variables with little loss of information

Using $\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}},\quad i,k = 1,2, \ldots, p$ we obtain

$$\rho Y_1, X_1 = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{0.383\sqrt{5.83}}{\sqrt{1}} = 0.925$$

$$\rho Y_1, X_2 = \frac{e_{12}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{-0.924\sqrt{5.83}}{\sqrt{5}} = -0.998$$

The variableX_2, with coefficient - 0.924 receives the greatest weight in the componentY_1. It is also having the largest correlation (in absolute value) with Y_1. The correlation of X_1, with Y_1, 0.925, is almost as large as that for X_2, indicating that the variables are about equally important to the first principal component. The relative sizes of the coefficients of X_1 and X_2 suggest, however, that X_2contributes more to the determination of Y_1 than doesX_1. Since, in this case, both coefficients are reasonably large and they have opposite signs. We would argue that both variables aid in the independent of Y_1.

Finally

$\rho Y_2, X_1 = \rho Y_2, X_2 = 0$ and $\rho Y_2, X_3 = \frac{\sqrt{\lambda_2}}{\sqrt{\sigma_{33}}} = \frac{\sqrt{2}}{\sqrt{2}} = 1$

The remaining correlations can be neglected, since the third component is unimportant.

11.2.6. Reamark

Consider principal components derived from multivariate normal random variables. Supoose X is distsributed as $N_p(\mu, \Sigma)$.

We know that the density of X is constant on the μ centered ellipsoids $(x-\mu)'\Sigma^{-1}(x-\mu) = c^2$ which have axes $\pm c\sqrt{\lambda_i}e_i$, $i = 1,2, \dots, p$ where the (λ_i, e_i) are the eigenvalue-eigenvector pairs of Σ. A point lying on the i^{th} axis of the ellipsoid will have coordinates proportional to $e' = [e_{i1}, e_{i2}, \dots_i, e_{ip}]$ in the coordinate system that has origin μ and axes that are parallel to the original axes $x_1, x_2, \dots, x_p$.

Set $\mu = \theta$ and $A = \Sigma^{-1}$, we can write

$$c^2 = x'\Sigma^{-1}x = \frac{1}{\lambda_1}(e'_1 x)^2 + \frac{1}{\lambda_2}(e'_2 x)^2 + \cdots + \frac{1}{\lambda_p}(e'_p x)^2$$

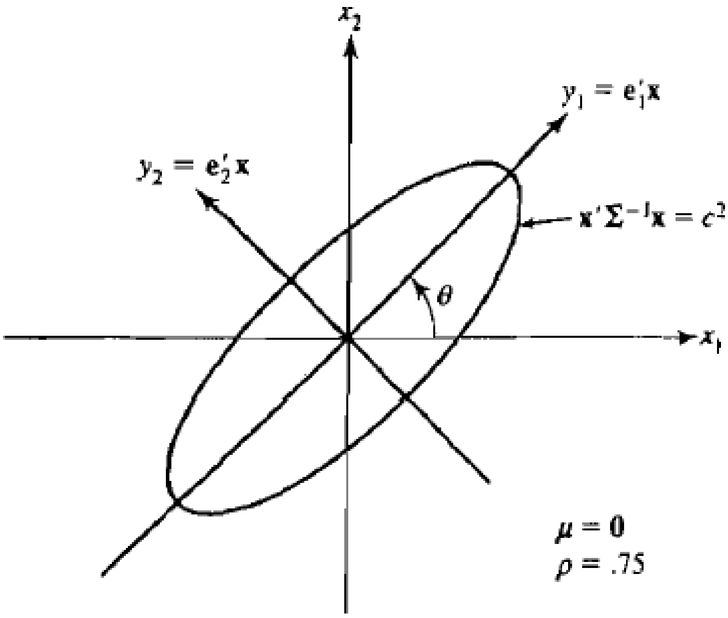
Where $e'_1 x, e'_2 x, \dots, e'_p x$ are recognized as the principal components of x. Setting $y_1 = e'_1 x, y_2 = e'_2 x, \dots, y_p = e'_p x$ we have

$c^2 = \frac{1}{\lambda_1}y_1^2 + \frac{1}{\lambda_2}y_2^2 + \cdots + \frac{1}{\lambda_p}y_p^2$ and this equation defines an ellipsoid (since $\lambda_1, \lambda_2, \dots, \lambda_p$ are positive) in a coordinate system with axes$y_1, y_2, \dots, y_p$ lying in the directions $e_1, e_2, \dots, e_p$respectively. If λ_1 is the largest eigenvalues, then the major axis lies in the directions e_1. The remaining minor axes line in the directions defined by $e_2, \dots, e_p$.

To summarize, the principal components $y_1 = e'_1 x, y_2 = e'_2 x, \dots, y_p = e'_p x$ lie in the directions of the axes of a constant density ellipsoid. Therefore, any point on the i^{th} ellipsoid axis has x cooridinates proportional to $e'_i = [e_{i1}, e_{i2}, \dots, e_{ip}]$ and necessarily, pincipal compoent coordinates of the form $[0, \dots, 0, y_i, 0, \dots, 0]$.

When $\mu \neq 0$, it is the mean-centreed principal component $y_i = e'_i(x-\mu)$ that mean 0 and lies in the direction e_i.

A constant density ellipse and the pincipal components for a bivariate normal random vector with $\mu = 0$ and $\rho = 0.75$ are shown in the following figur. We see that the principal components are obtained by rotating the orginal coordinate axes through an angle θ until they coincide with the axes of the constant density ellipse. This result holds for $\rho > 2$ diemensions as well



The above figure is the constant density ellispe $x'\Sigma^{-1} x = c^2$ and the principal components y_1, y_2 for a bivariate normal random vector X having mean 0.

11.2.7. Principal Components Obtained from Standardized Variables

Principal components may also be obtained for the standardized variables

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}}$$

$$Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}}$$

$$Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}}$$

In matrix notation

$Z = (V^{1/2})^{-1} (X - \mu)$ where $V^{1/2}$ is the diagonal standard deviation matrix, E(Z) = 0 and $Cov(Z) = (V^{1/2})^{-1}\Sigma(V^{1/2})^{-1} = \rho$. The principal components of Z may be obtained from the eigenvectos of the correlation matrix ρ of X. Since the variance of each Z_i is unity. We shall continue to use the notation Y_i to refer to the i^{th} principal component and (λ_i, e_i) for the eigenvalue-eigenvector pair from either ρ or Σ. The (λ_i, e_i) derived from Σ are not the same as the ones derived from ρ.

11.2.8. Result

The i^{th} principal component of the standardized variables $Z' = [Z_1, Z_2, \dots, Z_p]$ with $Cov(Z) = \rho$ is given by$Y_i = e_i' Z = e_i' (V^{\frac{1}{2}})^{-1} (X - \mu), i = 1,2, \dots, p$

Moreover, $\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p$ and $\rho Y_i, Z_k = e_{ik}\sqrt{\lambda_i}, i, k = 1,2, \dots, p$

In this case, $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$ are the eigen value-eigenvector pairs for ρ, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

Proof:

Result. 11.2.8. Follows from Results 11.2.1., 11.2.2. and 11.2.3. with $Z_1, Z_2, , , , , Z_p$in place of $X_1, X_2, , , , , X_p$ and ρ in place of Σ.

11.2.9. Ramark

From $\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p$, we have the total (standard varaibles) population variances is simply p, the sum of the diagonal elements of the matrix ρ. Using

$\begin{pmatrix} Proportion\ of\ total \\ population\ variance \\ due\ to\ k^{th} \\ principal \\ component \end{pmatrix} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$, $k = 1,2, \dots, p$ with Z in place of Z, we find the proportion of total variances explained by the k^{th} principal component of Z is

$$\begin{pmatrix} Proportion\ of\ (Standard) \\ population \\ variance\ due\ to\ k^{th} \\ principal \\ component \end{pmatrix} = \frac{\lambda_k}{p}, k = 1,2, \dots, p$$

Where the $\lambda_k's$ are the eigenvalues of ρ.

11.2.10. Example (Principal components obtained from covariance and correlation matrices are different)

Consider the covariance matrix $\Sigma = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$ and the derived correlation matrix $\rho = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 0 \end{bmatrix}$

The Eigen value - Eigen vector pairs from $\sum$ are

$\lambda_1 = 100.16, \qquad e_1' = [0.040, 0.999]$

$\lambda_2 = 0.84, \qquad e_2' = [0.999, -0.040]$

Similarly, the Eigen value - Eigen vector pairs from ρ are

$\lambda_1 = 1 + \rho = 1.4, \qquad e_1' = [0.707, 0.707]$

$\lambda_2 = 1 - \rho = 0.6, \qquad e_2' = [0.707, -0.707]$

The respective principal components become

Σ: $\begin{matrix} Y_1 = 0.040X_1 + 0.999X_2 \\ Y_1 = 0.999X_1 - 0.040X_2 \end{matrix}$ and

ρ:

$Y_1 = 0.707Z_1 + 0.707Z_2$

$$Y_1 = 0.707\left(\frac{X_1 - \mu_1}{1}\right) + 0.707\left(\frac{X_2 - \mu_2}{10}\right)$$

$Y_1 = 0.707(X_1 - \mu_1) + 0.0707(X_2 - \mu_2)$

$Y_2 = 0.707Z_1 - 0.707Z_2$

$$Y_1 = 0.707\left(\frac{X_1 - \mu_1}{1}\right) - 0.707\left(\frac{X_2 - \mu_2}{10}\right)$$

$Y_1 = 0.707(X_1 - \mu_1) - 0.0707(X_2 - \mu_2)$

Because of its large variance, X_2 completely dominates the first principal component determined from Σ. This first principal component explains a proportion $\frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{100.16}{101} = 0.992$ of the total population variance.

When the variables X_1 and X_2 are standardized, the resulting variables contribute equally to the principal components determined from ρ. Using Result 4 we obtain

$\rho Y_1, Z_1 = e_{11}\sqrt{\lambda_1} = 0.707\sqrt{1.4} = 0.837$

$\rho Y_1, Z_2 = e_{21}\sqrt{\lambda_1} = 0.707\sqrt{1.4} = 0.837$

In this case, the first principal compoent explains a proportion $\frac{\lambda_1}{p} = \frac{1.4}{2} = 0.7$ of the total (standardized) population variance.

The relative importance of the variables to the first pincipal compoenent is greatly affected by the standardization.

When the first principal component obatined from ρ is expressed in terms of X_1 and X_2, the relative magnitudes of the weights 0.707 and 0.0707 are in direct opposition to those of the weights 0.040 and 0.999 attached to these variables in the pincipal component obtained from Σ.

11.2.11. Note

The above example demonstrates that the principal components derived from Σ are different from those derived from ρ. One set of principal components is not a simple function of the other. This suggests that the standardization is not inconsequential.

Variables should probably be standardized if they are measured on scales with widely differing ranges or if the units of measurement are not commensurate. For example, if X_1represents annual sales in \$10,000 to \$35,000 range and X_2 is the ratio (net annual income)/(total assets) that falls in the 0.01 to 0.06 range, then the total variation will be due almost exclusively to dollar sales. In this case, we would expect a single (important) principal component with a heavy weighting of X_1. Alterntively, if both variables are standardized, their subsequent magnitudes will be of the same order, and X_2 or (Z_2) will play a larager role in the construction of the principal components. This behavior was observed in Exmple 11.2.10.

Let Us Sum Up

In this unit we studied the principal components and the population principal components.

Check Your Progress

1. Let $X_1, X_2, \ldots, X_P$ be the p random variables. Then the principal components depend on the ______.

2. The first principal component is the linear combination with__________.

Glossaries

Principal component analysis: It is concerned with explaining the variance-covariance structure of a set of variables through a few *linear* combinations of these variables.

i^{th} principal component: It is a linear combination $a_i'X$ that maximizes $Var(a_i'X)$ subject to $a_i'a_i = 1$ and $Cov(a_i'X, a_k'X) = 0$ for $k < i$

Suggested Readings

1. Johnson. R. A. and Wichern. D. W., "Applied Multivariate Statistical Analysis", Pearson Education Asia, Sixth Edition, 2007.

Answers to Check Your Progress

1. Maximum variance

2. Covariance matrix Σ or the correlation matrix ρ

STATISTICAL TABLES

I. Binomial Probabilities

II. Poisson Probabilities

III. Standard Normal Distribution

IV. Values of $t_{\frac{\alpha}{2},v}$

V. Values of $\chi^2{}_{\alpha,v}$

VI. Values of $f_{0.05,v_1,v_2}$ and $f_{0.01,v_1,v_2}$

VII. Factorials and Binomial Coefficients

VIII. Values of e^x and e^{-x}

Table I: Binomial Probabilities[†]

		θ									
n	x	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
1	0	.9500	.9000	.8500	.8000	.7500	.7000	.6500	.6000	.5500	.5000
	1	.0500	.1000	.1500	.2000	.2500	.3000	.3500	.4000	.4500	.5000
2	0	.9025	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.0950	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0025	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.1354	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0071	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0001	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.1715	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0135	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0005	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4	.0000	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0312
	1	.2036	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1562
	2	.0214	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0011	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4	.0000	.0004	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1562
	5	.0000	.0000	.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0312
6	0	.7351	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.2321	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2	.0305	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3	.0021	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125
	4	.0001	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5	.0000	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0938
	6	.0000	.0000	.0000	.0001	.0002	.0007	.0018	.0041	.0083	.0156
7	0	.6983	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1	.2573	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547
	2	.0406	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641
	3	.0036	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734
	4	.0002	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734
	5	.0000	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641
	6	.0000	.0000	.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547
	7	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0016	.0037	.0078
8	0	.6634	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039
	1	.2793	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0312
	2	.0515	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094
	3	.0054	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188
	4	.0004	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734
	5	.0000	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188
	6	.0000	.0000	.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094
	7	.0000	.0000	.0000	.0001	.0004	.0012	.0033	.0079	.0164	.0312
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0017	.0039

[†] Based on *Tables of the Binomial Probability Distribution*, National Bureau of Standards Applied Mathematics Series No. 6. Washington, D.C.: U.S. Government Printing Office, 1950.

Table I: *(continued)*

		θ									
n	*x*	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
9	0	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020
	1	.2985	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176
	2	.0629	.1722	.2597	.3020	.3003	.2668	.2162	.1612	.1110	.0703
	3	.0077	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641
	4	.0006	.0074	.0283	.0061	.1168	.1715	.2194	.2508	.2600	.2461
	5	.0000	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461
	6	.0000	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641
	7	.0000	.0000	.0000	.0003	.0012	.0039	.0098	.0212	.0407	.0703
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0035	.0083	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0020
10	0	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1	.3151	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098
	2	.0746	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439
	3	.0105	.0574	.1298	.2013	.2503	.2668	.2522	.2150	.1665	.1172
	4	.0010	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051
	5	.0001	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461
	6	.0000	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051
	7	.0000	.0000	.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172
	8	.0000	.0000	.0000	.0001	.0004	.0014	.0043	.0106	.0229	.0439
	9	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016	.0042	.0098
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0016
11	0	.5688	.3138	.1673	.0859	.0422	.0198	.0088	.0036	.0014	.0005
	1	.3293	.3835	.3248	.2362	.1549	.0932	.0518	.0266	.0125	.0054
	2	.0867	.2131	.2866	.2953	.2581	.1998	.1395	.0887	.0513	.0269
	3	.0137	.0710	.1517	.2215	.2581	.2568	.2254	.1774	.1259	.0806
	4	.0014	.0158	.0536	.1107	.1721	.2201	.2428	.2365	.2060	.1611
	5	.0001	.0025	.0132	.0388	.0803	.1321	.1830	.2207	.2360	.2256
	6	.0000	.0003	.0023	.0097	.0268	.0566	.0985	.1471	.1931	.2256
	7	.0000	.0000	.0003	.0017	.0064	.0173	.0379	.0701	.1128	.1611
	8	.0000	.0000	.0000	.0002	.0011	.0037	.0102	.0234	.0462	.0806
	9	.0000	.0000	.0000	.0000	.0001	.0005	.0018	.0052	.0126	.0269
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0007	.0021	.0054
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0005
12	0	.5404	.2824	.1422	.0687	.0317	.0138	.0057	.0022	.0008	.0002
	1	.3413	.3766	.3012	.2062	.1267	.0712	.0368	.0174	.0075	.0029
	2	.0988	.2301	.2924	.2835	.2323	.1678	.1088	.0639	.0339	.0161
	3	.0173	.0852	.1720	.2362	.2581	.2397	.1954	.1419	.0923	.0537
	4	.0021	.0213	.0683	.1329	.1936	.2311	.2367	.2128	.1700	.1208
	5	.0002	.0038	.0193	.0532	.1032	.1585	.2039	.2270	.2225	.1934
	6	.0000	.0005	.0040	.0155	.0401	.0792	.1281	.1766	.2124	.2256
	7	.0000	.0000	.0006	.0033	.0115	.0291	.0591	.1009	.1489	.1934
	8	.0000	.0000	.0001	.0005	.0024	.0078	.0199	.0420	.0762	.1208
	9	.0000	.0000	.0000	.0001	.0004	.0015	.0048	.0125	.0277	.0537
	10	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0025	.0068	.0161
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0029
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
13	0	.5133	.2542	.1209	.0550	.0238	.0097	.0037	.0013	.0004	.0001
	1	.3512	.3672	.2774	.1787	.1029	.0540	.0259	.0113	.0045	.0016
	2	.1109	.2448	.2937	.2680	.2059	.1388	.0836	.0453	.0220	.0095

Table I: *(continued)*

n	x	θ = .05	.10	.15	.20	.25	.30	.35	.40	.45	.50
13	3	.0214	.0997	.1900	.2457	.2517	.2181	.1651	.1107	.0660	.0349
	4	.0028	.0277	.0838	.1535	.2097	.2337	.2222	.1845	.1350	.0873
	5	.0003	.0055	.0266	.0691	.1258	.1803	.2154	.2214	.1989	.1571
	6	.0000	.0008	.0063	.0230	.0559	.1030	.1546	.1968	.2169	.2095
	7	.0000	.0001	.0011	.0058	.0186	.0442	.0833	.1312	.1775	.2095
	8	.0000	.0000	.0001	.0011	.0047	.0142	.0336	.0656	.1089	.1571
	9	.0000	.0000	.0000	.0001	.0009	.0034	.0101	.0243	.0495	.0873
	10	.0000	.0000	.0000	.0000	.0001	.0006	.0022	.0065	.0162	.0349
	11	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0012	.0036	.0095
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
14	0	.4877	.2288	.1028	.0440	.0178	.0068	.0024	.0008	.0002	.0001
	1	.3593	.3559	.2539	.1539	.0832	.0407	.0181	.0073	.0027	.0009
	2	.1229	.2570	.2912	.2501	.1802	.1134	.0634	.0317	.0141	.0056
	3	.0259	.1142	.2056	.2501	.2402	.1943	.1366	.0845	.0462	.0222
	4	.0037	.0349	.0998	.1720	.2202	.2290	.2022	.1549	.1040	.0611
	5	.0004	.0078	.0352	.0860	.1468	.1963	.2178	.2066	.1701	.1222
	6	.0000	.0013	.0093	.0322	.0734	.1262	.1759	.2066	.2088	.1833
	7	.0000	.0002	.0019	.0092	.0280	.0618	.1082	.1574	.1952	.2095
	8	.0000	.0000	.0003	.0020	.0082	.0232	.0510	.0918	.1398	.1833
	9	.0000	.0000	.0000	.0003	.0018	.0066	.0183	.0408	.0762	.1222
	10	.0000	.0000	.0000	.0000	.0003	.0014	.0049	.0136	.0312	.0611
	11	.0000	.0000	.0000	.0000	.0000	.0002	.0010	.0033	.0093	.0222
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0019	.0056
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0009
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
15	0	.4633	.2059	.0874	.0352	.0134	.0047	.0016	.0005	.0001	.0000
	1	.3658	.3432	.2312	.1319	.0668	.0305	.0126	.0047	.0016	.0005
	2	.1348	.2669	.2856	.2309	.1559	.0916	.0476	.0219	.0090	.0032
	3	.0307	.1285	.2184	.2501	.2252	.1700	.1110	.0634	.0318	.0139
	4	.0049	.0428	.1156	.1876	.2252	.2186	.1792	.1268	.0780	.0417
	5	.0006	.0105	.0449	.1032	.1651	.2061	.2123	.1859	.1404	.0916
	6	.0000	.0019	.0132	.0430	.0917	.1472	.1906	.2066	.1914	.1527
	7	.0000	.0003	.0030	.0138	.0393	.0811	.1319	.1771	.2013	.1964
	8	.0000	.0000	.0005	.0035	.0131	.0348	.0710	.1181	.1647	.1964
	9	.0000	.0000	.0001	.0007	.0034	.0116	.0298	.0612	.1048	.1527
	10	.0000	.0000	.0000	.0001	.0007	.0030	.0096	.0245	.0515	.0916
	11	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0074	.0191	.0417
	12	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052	.0139
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0032
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
16	0	.4401	.1853	.0743	.0281	.0100	.0033	.0010	.0003	.0001	.0000
	1	.3706	.3294	.2097	.1126	.0535	.0228	.0087	.0030	.0009	.0002
	2	.1463	.2745	.2775	.2111	.1336	.0732	.0353	.0150	.0056	.0018
	3	.0359	.1423	.2285	.2463	.2079	.1465	.0888	.0468	.0215	.0085
	4	.0061	.0514	.1311	.2001	.2252	.2040	.1553	.1014	.0572	.0278
	5	.0008	.0137	.0555	.1201	.1802	.2099	.2008	.1623	.1123	.0667
	6	.0001	.0028	.0180	.0550	.1101	.1649	.1982	.1983	.1684	.1222

Table I: *(continued)*

n	x	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
						θ					
16	7	.0000	.0004	.0045	.0197	.0524	.1010	.1524	.1889	.1969	.1746
	8	.0000	.0001	.0009	.0055	.0197	.0487	.0923	.1417	.1812	.1964
	9	.0000	.0000	.0001	.0012	.0058	.0185	.0442	.0840	.1318	.1746
	10	.0000	.0000	.0000	.0002	.0014	.0056	.0167	.0392	.0755	.1222
	11	.0000	.0000	.0000	.0000	.0002	.0013	.0049	.0142	.0337	.0667
	12	.0000	.0000	.0000	.0000	.0000	.0002	.0011	.0040	.0115	.0278
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0029	.0085
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0018
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
17	0	.4181	.1668	.0631	.0225	.0075	.0023	.0007	.0002	.0000	.0000
	1	.3741	.3150	.1893	.0957	.0426	.0169	.0060	.0019	.0005	.0001
	2	.1575	.2800	.2673	.1914	.1136	.0581	.0260	.0102	.0035	.0010
	3	.0415	.1556	.2359	.2393	.1893	.1245	.0701	.0341	.0144	.0052
	4	.0076	.0605	.1457	.2093	.2209	.1868	.1320	.0796	.0411	.0182
	5	.0010	.0175	.0668	.1361	.1914	.2081	.1849	.1379	.0875	.0472
	6	.0001	.0039	.0236	.0680	.1276	.1784	.1991	.1839	.1432	.0944
	7	.0000	.0007	.0065	.0267	.0668	.1201	.1685	.1927	.1841	.1484
	8	.0000	.0001	.0014	.0084	.0279	.0644	.1134	.1606	.1883	.1855
	9	.0000	.0000	.0003	.0021	.0093	.0276	.0611	.1070	.1540	.1855
	10	.0000	.0000	.0000	.0004	.0025	.0095	.0263	.0571	.1008	.1484
	11	.0000	.0000	.0000	.0001	.0005	.0026	.0090	.0242	.0525	.0944
	12	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0081	.0215	.0472
	13	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0021	.0068	.0182
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
18	0	.3972	.1501	.0536	.0180	.0056	.0016	.0004	.0001	.0000	.0000
	1	.3763	.3002	.1704	.0811	.0338	.0126	.0042	.0012	.0003	.0001
	2	.1683	.2835	.2556	.1723	.0958	.0458	.0190	.0069	.0022	.0006
	3	.0473	.1680	.2406	.2297	.1704	.1046	.0547	.0246	.0095	.0031
	4	.0093	.9700	.1592	.2153	.2130	.1681	.1104	.0614	.0291	.0117
	5	.0014	.0218	.0787	.1507	.1988	.2017	.1664	.1146	.0666	.0327
	6	.0002	.0052	.0301	.0816	.1436	.1873	.1941	.1655	.1181	.0708
	7	.0000	.0010	.0091	.0350	.0820	.1376	.1792	.1892	.1657	.1214
	8	.0000	.0002	.0022	.0120	.0376	.0811	.1327	.1734	.1864	.1669
	9	.0000	.0000	.0004	.0033	.0139	.0386	.0794	.1284	.1694	.1855
	10	.0000	.0000	.0001	.0008	.0042	.0149	.0385	.0771	.1248	.1669
	11	.0000	.0000	.0000	.0001	.0010	.0046	.0151	.0374	.0742	.1214
	12	.0000	.0000	.0000	.0000	.0002	.0012	.0047	.0145	.0354	.0708
	13	.0000	.0000	.0000	.0000	.0000	.0002	.0012	.0045	.0134	.0327
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0011	.0039	.0117
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0009	.0031
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0006
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
19	0	.3774	.1351	.0456	.0144	.0042	.0011	.0003	.0001	.0000	.0000
	1	.3774	.2852	.1529	.0685	.0268	.0093	.0029	.0008	.0002	.0000

Table I: *(continued)*

		θ									
n	x	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
19	2	.1787	.2852	.2428	.1540	.0803	.0358	.0138	.0046	.0013	.0003
	3	.0533	.1796	.2428	.2182	.1517	.0869	.0422	.0175	.0062	.0018
	4	.0112	.0798	.1714	.2182	.2023	.1491	.0909	.0467	.0203	.0074
	5	.0018	.0266	.0907	.1636	.2023	.1916	.1468	.0933	.0497	.0222
	6	.0002	.0069	.0374	.0955	.1574	.1916	.1844	.1451	.0949	.0518
	7	.0000	.0014	.0122	.0443	.0974	.1525	.1844	.1797	.1443	.0961
	8	.0000	.0002	.0032	.0166	.0487	.0981	.1489	.1797	.1771	.1442
	9	.0000	.0000	.0007	.0051	.0198	.0514	.0980	.1464	.1771	.1762
	10	.0000	.0000	.0001	.0013	.0066	.0220	.0528	.0976	.1449	.1762
	11	.0000	.0000	.0000	.0003	.0018	.0077	.0233	.0532	.0970	.1442
	12	.0000	.0000	.0000	.0000	.0004	.0022	.0083	.0237	.0529	.0961
	13	.0000	.0000	.0000	.0000	.0001	.0005	.0024	.0085	.0233	.0518
	14	.0000	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0082	.0222
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0022	.0074
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0018
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
20	0	.3585	.1216	.0388	.0115	.0032	.0008	.0002	.0000	.0000	.0000
	1	.3774	.2702	.1368	.0576	.0211	.0068	.0020	.0005	.0001	.0000
	2	.1887	.2852	.2293	.1369	.0669	.0278	.0100	.0031	.0008	.0002
	3	.0596	.1901	.2428	.2054	.1339	.0716	.0323	.0123	.0040	.0011
	4	.0133	.0898	.1821	.2182	.1897	.1304	.0738	.0350	.0139	.0046
	5	.0022	.0319	.1028	.1746	.2023	.1789	.1272	.0746	.0365	.0148
	6	.0003	.0089	.0454	.1091	.1686	.1916	.1712	.1244	.0746	.0370
	7	.0000	.0020	.0160	.0545	.1124	.1643	.1844	.1659	.1221	.0739
	8	.0000	.0004	.0046	.0222	.0609	.1144	.1614	.1797	.1623	.1201
	9	.0000	.0001	.0011	.0074	.0271	.0654	.1158	.1597	.1771	.1602
	10	.0000	.0000	.0002	.0020	.0099	.0308	.0686	.1171	.1593	.1762
	11	.0000	.0000	.0000	.0005	.0030	.0120	.0336	.0710	.1185	.1602
	12	.0000	.0000	.0000	.0001	.0008	.0039	.0136	.0355	.0727	.1201
	13	.0000	.0000	.0000	.0000	.0002	.0010	.0045	.0146	.0366	.0739
	14	.0000	.0000	.0000	.0000	.0000	.0002	.0012	.0049	.0150	.0370
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0013	.0049	.0148
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0013	.0046
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0011
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	20	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Table II: Poisson Probabilities†

x	λ = 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4	.0000	.0001	.0002	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005
7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

x	λ = 1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353
1	.3662	.3614	.3543	.3452	.3347	.3230	.3106	.2975	.2842	.2707
2	.2014	.2169	.2303	.2417	.2510	.2584	.2640	.2678	.2700	.2707
3	.0738	.0867	.0998	.1128	.1255	.1378	.1496	.1607	.1710	.1804
4	.0203	.0260	.0324	.0395	.0471	.0551	.0636	.0723	.0812	.0902
5	.0045	.0062	.0084	.0111	.0141	.0176	.0216	.0260	.0309	.0361
6	.0008	.0012	.0018	.0026	.0035	.0047	.0061	.0078	.0098	.0120
7	.0001	.0002	.0003	.0005	.0008	.0011	.0015	.0020	.0027	.0034
8	.0000	.0000	.0001	.0001	.0001	.0002	.0003	.0005	.0006	.0009
9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002

x	λ = 2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	.1225	.1108	.1003	.0907	.0821	.0743	.0672	.0608	.0550	.0498
1	.2572	.2438	.2306	.2177	.2052	.1931	.1815	.1703	.1596	.1494
2	.2700	.2681	.2652	.2613	.2565	.2510	.2450	.2384	.2314	.2240
3	.1890	.1966	.2033	.2090	.2138	.2176	.2205	.2225	.2237	.2240
4	.0992	.1082	.1169	.1254	.1336	.1414	.1488	.1557	.1622	.1680
5	.0417	.0476	.0538	.0602	.0668	.0735	.0804	.0872	.0940	.1008
6	.0146	.0174	.0206	.0241	.0278	.0319	.0362	.0407	.0455	.0504
7	.0044	.0055	.0068	.0083	.0099	.0118	.0139	.0163	.0188	.0216
8	.0011	.0015	.0019	.0025	.0031	.0038	.0047	.0057	.0068	.0081
9	.0003	.0004	.0005	.0007	.0009	.0011	.0014	.0018	.0022	.0027
10	.0001	.0001	.0001	.0002	.0002	.0003	.0004	.0005	.0006	.0008
11	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002	.0002
12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

x	λ = 3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	.0450	.0408	.0369	.0334	.0302	.0273	.0247	.0224	.0202	.0183
1	.1397	.1304	.1217	.1135	.1057	.0984	.0915	.0850	.0789	.0733
2	.2165	.2087	.2008	.1929	.1850	.1771	.1692	.1615	.1539	.1465
3	.2237	.2226	.2209	.2186	.2158	.2125	.2087	.2046	.2001	.1954
4	.1734	.1781	.1823	.1858	.1888	.1912	.1931	.1944	.1951	.1954
5	.1075	.1140	.1203	.1264	.1322	.1377	.1429	.1477	.1522	.1563
6	.0555	.0608	.0662	.0716	.0771	.0826	.0881	.0936	.0989	.1042
7	.0246	.0278	.0312	.0348	.0385	.0425	.0466	.0508	.0551	.0595
8	.0095	.0111	.0129	.0148	.0169	.0191	.0215	.0241	.0269	.0298
9	.0033	.0040	.0047	.0056	.0066	.0076	.0089	.0102	.0116	.0132

†Based on E. C. Molina, *Poisson's Exponential Binomial Limit*, 1973 Reprint, Robert E. Krieger Publishing Company, Melbourne, Fla., by permission of the publisher.

Table II: *(continued)*

x	λ 3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
10	.0010	.0013	.0016	.0019	.0023	.0028	.0033	.0039	.0045	.0053
11	.0003	.0004	.0005	.0006	.0007	.0009	.0011	.0013	.0016	.0019
12	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005	.0006
13	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0002	.0002
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

x	λ 4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	.0166	.0150	.0136	.0123	.0111	.0101	.0091	.0082	.0074	.0067
1	.0679	.0630	.0583	.0540	.0500	.0462	.0427	.0395	.0365	.0337
2	.1393	.1323	.1254	.1188	.1125	.1063	.1005	.0948	.0894	.0842
3	.1904	.1852	.1798	.1743	.1687	.1631	.1574	.1517	.1460	.1404
4	.1951	.1944	.1933	.1917	.1898	.1875	.1849	.1820	.1789	.1755
5	.1600	.1633	.1662	.1687	.1708	.1725	.1738	.1747	.1753	.1755
6	.1093	.1143	.1191	.1237	.1281	.1323	.1362	.1398	.1432	.1462
7	.0640	.0686	.0732	.0778	.0824	.0869	.0914	.0959	.1002	.1044
8	.0328	.0360	.0393	.0428	.0463	.0500	.0537	.0575	.0614	.0653
9	.0150	.0168	.0188	.0209	.0232	.0255	.0280	.0307	.0334	.0363
10	.0061	.0071	.0081	.0092	.0104	.0118	.0132	.0147	.0164	.0181
11	.0023	.0027	.0032	.0037	.0043	.0049	.0056	.0064	.0073	.0082
12	.0008	.0009	.0011	.0014	.0016	.0019	.0022	.0026	.0030	.0034
13	.0002	.0003	.0004	.0005	.0006	.0007	.0008	.0009	.0011	.0013
14	.0001	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005
15	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0002

x	λ 5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	6.0
0	.0061	.0055	.0050	.0045	.0041	.0037	.0033	.0030	.0027	.0025
1	.0311	.0287	.0265	.0244	.0225	.0207	.0191	.0176	.0162	.0149
2	.0793	.0746	.0701	.0659	.0618	.0580	.0544	.0509	.0477	.0446
3	.1348	.1293	.1239	.1185	.1133	.1082	.1033	.0985	.0938	.0892
4	.1719	.1681	.1641	.1600	.1558	.1515	.1472	.1428	.1383	.1339
5	.1753	.1748	.1740	.1728	.1714	.1697	.1678	.1656	.1632	.1606
6	.1490	.1515	.1537	.1555	.1571	.1584	.1594	.1601	.1505	.1606
7	.1086	.1125	.1163	.1200	.1234	.1267	.1298	.1326	.1353	.1377
8	.0692	.0731	.0771	.0810	.0849	.0887	.0925	.0962	.0998	.1033
9	.0392	.0423	.0454	.0486	.0519	.0552	.0586	.0620	.0654	.0688
10	.0200	.0220	.0241	.0262	.0285	.0309	.0334	.0359	.0386	.0413
11	.0093	.0104	.0116	.0129	.0143	.0157	.0173	.0190	.0207	.0225
12	.0039	.0045	.0051	.0058	.0065	.0073	.0082	.0092	.0102	.0113
13	.0015	.0018	.0021	.0024	.0028	.0032	.0036	.0041	.0046	.0052
14	.0006	.0007	.0008	.0009	.0011	.0013	.0015	.0017	.0019	.0022
15	.0002	.0002	.0003	.0003	.0004	.0005	.0006	.0007	.0008	.0009
16	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003
17	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001

x	λ 6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
0	.0022	.0020	.0018	.0017	.0015	.0014	.0012	.0011	.0010	.0009
1	.0137	.0126	.0116	.0106	.0098	.0090	.0082	.0076	.0070	.0064
2	.0417	.0390	.0364	.0340	.0318	.0296	.0276	.0258	.0240	.0223

Table II: *(continued)*

x	λ 6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8	6.9	7.0
3	.0848	.0806	.0765	.0726	.0688	.0652	.0617	.0584	.0552	.0521
4	.1294	.1249	.1205	.1162	.1118	.1076	.1034	.0992	.0952	.0912
5	.1579	.1549	.1519	.1487	.1454	.1420	.1385	.1349	.1314	.1277
6	.1605	.1601	.1595	.1586	.1575	.1562	.1546	.1529	.1511	.1490
7	.1399	.1418	.1435	.1450	.1462	.1472	.1480	.1486	.1489	.1490
8	.1066	.1099	.1130	.1160	.1188	.1215	.1240	.1263	.1284	.1304
9	.0723	.0757	.0791	.0825	.0858	.0891	.0923	.0954	.0985	.1014
10	.0441	.0469	.0498	.0528	.0558	.0588	.0618	.0649	.0679	.0710
11	.0245	.0265	.0285	.0307	.0330	.0353	.0377	.0401	.0426	.0452
12	.0124	.0137	.0150	.0164	.0179	.0194	.0210	.0227	.0245	.0264
13	.0058	.0065	.0073	.0081	.0089	.0098	.0108	.0119	.0130	.0142
14	.0025	.0029	.0033	.0037	.0041	.0046	.0052	.0058	.0064	.0071
15	.0010	.0012	.0014	.0016	.0018	.0020	.0023	.0026	.0029	.0033
16	.0004	.0005	.0005	.0006	.0007	.0008	.0010	.0011	.0013	.0014
17	.0001	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0006
18	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002
19	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001

x	λ 7.1	7.2	7.3	7.4	7.5	7.6	7.7	7.8	7.9	8.0
0	.0008	.0007	.0007	.0006	.0006	.0005	.0005	.0004	.0004	.0003
1	.0059	.0054	.0049	.0045	.0041	.0038	.0035	.0032	.0029	.0027
2	.0208	.0194	.0180	.0167	.0156	.0145	.0134	.0125	.0116	.0107
3	.0492	.0464	.0438	.0413	.0389	.0366	.0345	.0324	.0305	.0286
4	.0874	.0836	.0799	.0764	.0729	.0696	.0663	.0632	.0602	.0573
5	.1241	.1204	.1167	.1130	.1094	.1057	.1021	.0986	.0951	.0916
6	.1468	.1445	.1420	.1394	.1367	.1339	.1311	.1282	.1252	.1221
7	.1489	.1486	.1481	.1474	.1465	.1454	.1442	.1428	.1413	.1396
8	.1321	.1337	.1351	.1363	.1373	.1382	.1388	.1392	.1395	.1396
9	.1042	.1070	.1096	.1121	.1144	.1167	.1187	.1207	.1224	.1241
10	.0740	.0770	.0800	.0829	.0858	.0887	.0914	.0941	.0967	.0993
11	.0478	.0504	.0531	.0558	.0585	.0613	.0640	.0667	.0695	.0722
12	.0283	.0303	.0323	.0344	.0366	.0388	.0411	.0434	.0457	.0481
13	.0154	.0168	.0181	.0196	.0211	.0227	.0243	.0260	.0278	.0296
14	.0078	.0086	.0095	.0104	.0113	.0123	.0134	.0145	.0157	.0169
15	.0037	.0041	.0046	.0051	.0057	.0062	.0069	.0075	.0083	.0090
16	.0016	.0019	.0021	.0024	.0026	.0030	.0033	.0037	.0041	.0045
17	.0007	.0008	.0009	.0010	.0012	.0013	.0015	.0017	.0019	.0021
18	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
19	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003	.0003	.0004
20	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002
21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001

x	λ 8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
0	.0003	.0003	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001
1	.0025	.0023	.0021	.0019	.0017	.0016	.0014	.0013	.0012	.0011
2	.0100	.0092	.0086	.0079	.0074	.0068	.0063	.0058	.0054	.0050
3	.0269	.0252	.0237	.0222	.0208	.0195	.0183	.0171	.0160	.0150
4	.0544	.0517	.0491	.0466	.0443	.0420	.0398	.0377	.0357	.0337

Table II: *(continued)*

x	λ 8.1	8.2	8.3	8.4	8.5	8.6	8.7	8.8	8.9	9.0
5	.0882	.0849	.0816	.0784	.0752	.0722	.0692	.0663	.0635	.0607
6	.1191	.1160	.1128	.1097	.1066	.1034	.1003	.0972	.0941	.0911
7	.1378	.1358	.1338	.1317	.1294	.1271	.1247	.1222	.1197	.1171
8	.1395	.1392	.1388	.1382	.1375	.1366	.1356	.1344	.1332	.1318
9	.1256	.1269	.1280	.1290	.1299	.1306	.1311	.1315	.1317	.1318
10	.1017	.1040	.1063	.1084	.1104	.1123	.1140	.1157	.1172	.1186
11	.0749	.0776	.0802	.0828	.0853	.0878	.0902	.0925	.0948	.0970
12	.0505	.0530	.0555	.0579	.0604	.0629	.0654	.0679	.0703	.0728
13	.0315	.0334	.0354	.0374	.0395	.0416	.0438	.0459	.0481	.0504
14	.0182	.0196	.0210	.0225	.0240	.0256	.0272	.0289	.0306	.0324
15	.0098	.0107	.0116	.0126	.0136	.0147	.0158	.0169	.0182	.0194
16	.0050	.0055	.0060	.0066	.0072	.0079	.0086	.0093	.0101	.0109
17	.0024	.0026	.0029	.0033	.0036	.0040	.0044	.0048	.0053	.0058
18	.0011	.0012	.0014	.0015	.0017	.0019	.0021	.0024	.0026	.0029
19	.0005	.0005	.0006	.0007	.0008	.0009	.0010	.0011	.0012	.0014
20	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0005	.0006
21	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0002	.0003
22	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001

x	λ 9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10
0	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0000
1	.0010	.0009	.0009	.0008	.0007	.0007	.0006	.0005	.0005	.0005
2	.0046	.0043	.0040	.0037	.0034	.0031	.0029	.0027	.0025	.0023
3	.0140	.0131	.0123	.0115	.0107	.0100	.0093	.0087	.0081	.0076
4	.0319	.0302	.0285	.0269	.0254	.0240	.0226	.0213	.0201	.0189
5	.0581	.0555	.0530	.0506	.0483	.0460	.0439	.0418	.0398	.0378
6	.0881	.0851	.0822	.0793	.0764	.0736	.0709	.0682	.0656	.0631
7	.1145	.1118	.1091	.1064	.1037	.1010	.0982	.0955	.0928	.0901
8	.1302	.1286	.1269	.1251	.1232	.1212	.1191	.1170	.1148	.1126
9	.1317	.1315	.1311	.1306	.1300	.1293	.1284	.1274	.1263	.1251
10	.1198	.1210	.1219	.1228	.1235	.1241	.1245	.1249	.1250	.1251
11	.0991	.1012	.1031	.1049	.1067	.1083	.1098	.1112	.1125	.1137
12	.0752	.0776	.0799	.0822	.0844	.0866	.0888	.0908	.0928	.0948
13	.0526	.0549	.0572	.0594	.0617	.0640	.0662	.0685	.0707	.0729
14	.0342	.0361	.0380	.0399	.0419	.0439	.0459	.0479	.0500	.0521
15	.0208	.0221	.0235	.0250	.0265	.0281	.0297	.0313	.0330	.0347
16	.0118	.0127	.0137	.0147	.0157	.0168	.0180	.0192	.0204	.0217
17	.0063	.0069	.0075	.0081	.0088	.0095	.0103	.0111	.0119	.0128
18	.0032	.0035	.0039	.0042	.0046	.0051	.0055	.0060	.0065	.0071
19	.0015	.0017	.0019	.0021	.0023	.0026	.0028	.0031	.0034	.0037
20	.0007	.0008	.0009	.0010	.0011	.0012	.0014	.0015	.0017	.0019
21	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
22	.0001	.0001	.0002	.0002	.0002	.0002	.0003	.0003	.0004	.0004
23	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002
24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001

x	λ 11	12	13	14	15	16	17	18	19	20
0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Table II: *(continued)*

	λ									
x	11	12	13	14	15	16	17	18	19	20
2	.0010	.0004	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000
3	.0037	.0018	.0008	.0004	.0002	.0001	.0000	.0000	.0000	.0000
4	.0102	.0053	.0027	.0013	.0006	.0003	.0001	.0001	.0000	.0000
5	.0224	.0127	.0070	.0037	.0019	.0010	.0005	.0002	.0001	.0001
6	.0411	.0255	.0152	.0087	.0048	.0026	.0014	.0007	.0004	.0002
7	.0646	.0437	.0281	.0174	.0104	.0060	.0034	.0018	.0010	.0005
8	.0888	.0655	.0457	.0304	.0194	.0120	.0072	.0042	.0024	.0013
9	.1085	.0874	.0661	.0473	.0324	.0213	.0135	.0083	.0050	.0029
10	.1194	.1048	.0859	.0663	.0486	.0341	.0230	.0150	.0095	.0058
11	.1194	.1144	.1015	.0844	.0663	.0496	.0355	.0245	.0164	.0106
12	.1094	.1144	.1099	.0984	.0829	.0661	.0504	.0368	.0259	.0176
13	.0926	.1056	.1099	.1060	.0956	.0814	.0658	.0509	.0378	.0271
14	.0728	.0905	.1021	.1060	.1024	.0930	.0800	.0655	.0514	.0387
15	.0534	.0724	.0885	.0989	.1024	.0992	.0906	.0786	.0650	.0516
16	.0367	.0543	.0719	.0866	.0960	.0992	.0963	.0884	.0772	.0646
17	.0237	.0383	.0550	.0713	.0847	.0934	.0963	.0936	.0863	.0760
18	.0145	.0256	.0397	.0554	.0706	.0830	.0909	.0936	.0911	.0844
19	.0084	.0161	.0272	.0409	.0557	.0699	.0814	.0887	.0911	.0888
20	.0046	.0097	.0177	.0286	.0418	.0559	.0692	.0798	.0866	.0888
21	.0024	.0055	.0109	.0191	.0299	.0426	.0560	.0684	.0783	.0846
22	.0012	.0030	.0065	.0121	.0204	.0310	.0433	.0560	.0676	.0769
23	.0006	.0016	.0037	.0074	.0133	.0216	.0320	.0438	.0559	.0669
24	.0003	.0008	.0020	.0043	.0083	.0144	.0226	.0328	.0442	.0557
25	.0001	.0004	.0010	.0024	.0050	.0092	.0154	.0237	.0336	.0446
26	.0000	.0002	.0005	.0013	.0029	.0057	.0101	.0164	.0246	.0343
27	.0000	.0001	.0002	.0007	.0016	.0034	.0063	.0109	.0173	.0254
28	.0000	.0000	.0001	.0003	.0009	.0019	.0038	.0070	.0117	.0181
29	.0000	.0000	.0001	.0002	.0004	.0011	.0023	.0044	.0077	.0125
30	.0000	.0000	.0000	.0001	.0002	.0006	.0013	.0026	.0049	.0083
31	.0000	.0000	.0000	.0000	.0001	.0003	.0007	.0015	.0030	.0054
32	.0000	.0000	.0000	.0000	.0001	.0001	.0004	.0009	.0018	.0034
33	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0005	.0010	.0020
34	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0012
35	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0007
36	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0004
37	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
38	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
39	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

Table III: Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4988
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Also, for $z = 4.0$, 5.0, and 6.0, the probabilities are 0.49997, 0.4999997, and 0.499999999.

Table IV: Values of $t_{\alpha,\nu}$ †

ν	$\alpha = .10$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .005$	ν
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
inf.	1.282	1.645	1.960	2.326	2.576	inf.

†Based on Richard A. Johnson and Dean W. Wichern, *Applied Multivariate Statistical Analysis*, 2nd ed., © 1988, Table 2, p. 592. By permission of Prentice Hall, Upper Saddle River, N.J.

Table V: Values of $\chi^2_{\alpha,\nu}$ †

ν	$\alpha = .995$	$\alpha = .99$	$\alpha = .975$	$\alpha = .95$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .005$	ν
1	.0000393	.000157	.000982	.00393	3.841	5.024	6.635	7.879	1
2	.0100	.0201	.0506	.103	5.991	7.378	9.210	10.597	2
3	.0717	.115	.216	.352	7.815	9.348	11.345	12.838	3
4	.207	.297	.484	.711	9.488	11.143	13.277	14.860	4
5	.412	.554	.831	1.145	11.070	12.832	15.086	16.750	5
6	.676	.872	1.237	1.635	12.592	14.449	16.812	18.548	6
7	.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188	10
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757	11
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300	12
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819	13
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319	14
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801	15
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267	16
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718	17
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156	18
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582	19
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997	20
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401	21
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796	22
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181	23
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558	24
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928	25
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290	26
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645	27
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993	28
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336	29
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672	30

† Based on Table 8 of *Biometrika Tables for Statisticians*, Vol. 1, Cambridge University Press, 1954, by permission of the *Biometrika* trustees.

Table VI: Values of $f_{0.05,\nu_1,\nu_2}$ †

ν_2 = Degrees of freedom for denominator	ν_1 = Degrees of freedom for numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07

†Reproduced from M. Merrington and C. M. Thompson, "Tables of percentage points of the inverted beta (*F*) distribution," *Biometrika*, Vol. 33 (1943), by permission of the *Biometrika* trustees.

Table VI: (*continued*) Values of $f_{0.05,\nu_1,\nu_2}$

ν_2 = Degrees of freedom for denominator	ν_1 = Degrees of freedom for numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Table VI: *(continued)* Values of $f_{0.01,\nu_1,\nu_2}$

ν_2 = Degrees of freedom for denominator	ν_1 = Degrees of freedom for numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Table VII: Factorials and Binomial Coefficients

Factorials

n	$n!$	$\log n!$
0	1	0.0000
1	1	0.0000
2	2	0.3010
3	6	0.7782
4	24	1.3802
5	120	2.0792
6	720	2.8573
7	5,040	3.7024
8	40,320	4.6055
9	362,880	5.5598
10	3,628,800	6.5598
11	39,916,800	7.6012
12	479,001,600	8.6803
13	6,227,020,800	9.7943
14	87,178,291,200	10.9404
15	1,307,674,368,000	12.1165

Binomial Coefficients

n	$\binom{n}{0}$	$\binom{n}{1}$	$\binom{n}{2}$	$\binom{n}{3}$	$\binom{n}{4}$	$\binom{n}{5}$	$\binom{n}{6}$	$\binom{n}{7}$	$\binom{n}{8}$	$\binom{n}{9}$	$\binom{n}{10}$
0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1
11	1	11	55	165	330	462	462	330	165	55	11
12	1	12	66	220	495	792	924	792	495	220	66
13	1	13	78	286	715	1287	1716	1716	1287	715	286
14	1	14	91	364	1001	2002	3003	3432	3003	2002	1001
15	1	15	105	455	1365	3003	5005	6435	6435	5005	3003
16	1	16	120	560	1820	4368	8008	11440	12870	11440	8008
17	1	17	136	680	2380	6188	12376	19448	24310	24310	19448
18	1	18	153	816	3060	8568	18564	31824	43758	48620	43758
19	1	19	171	969	3876	11628	27132	50388	75582	92378	92378
20	1	20	190	1140	4845	15504	38760	77520	125970	167960	184756

Table VIII: Values of e^x and e^{-x}

x	e^x	e^{-x}	x	e^x	e^{-x}
0.0	1.000	1.000	**5.0**	148.4	0.0067
0.1	1.105	0.905	**5.1**	164.0	0.0061
0.2	1.221	0.819	**5.2**	181.3	0.0055
0.3	1.350	0.741	**5.3**	200.3	0.0050
0.4	1.492	0.670	**5.4**	221.4	0.0045
0.5	1.649	0.607	**5.5**	244.7	0.0041
0.6	1.822	0.549	**5.6**	270.4	0.0037
0.7	2.014	0.497	**5.7**	298.9	0.0033
0.8	2.226	0.449	**5.8**	330.3	0.0030
0.9	2.460	0.407	**5.9**	365.0	0.0027
1.0	2.718	0.368	**6.0**	403.4	0.0025
1.1	3.004	0.333	**6.1**	445.9	0.0022
1.2	3.320	0.301	**6.2**	492.8	0.0020
1.3	3.669	0.273	**6.3**	544.6	0.0018
1.4	4.055	0.247	**6.4**	601.8	0.0017
1.5	4.482	0.223	**6.5**	665.1	0.0015
1.6	4.953	0.202	**6.6**	735.1	0.0014
1.7	5.474	0.183	**6.7**	812.4	0.0012
1.8	6.050	0.165	**6.8**	897.8	0.0011
1.9	6.686	0.150	**6.9**	992.3	0.0010
2.0	7.389	0.135	**7.0**	1,096.6	0.0009
2.1	8.166	0.122	**7.1**	1,212.0	0.0008
2.2	9.025	0.111	**7.2**	1,339.4	0.0007
2.3	9.974	0.100	**7.3**	1,480.3	0.0007
2.4	11.023	0.091	**7.4**	1,636.0	0.0006
2.5	12.18	0.082	**7.5**	1,808.0	0.00055
2.6	13.46	0.074	**7.6**	1,998.2	0.00050
2.7	14.88	0.067	**7.7**	2,208.3	0.00045
2.8	16.44	0.061	**7.8**	2,440.6	0.00041
2.9	18.17	0.055	**7.9**	2,697.3	0.00037
3.0	20.09	0.050	**8.0**	2,981.0	0.00034
3.1	22.20	0.045	**8.1**	3,294.5	0.00030
3.2	24.53	0.041	**8.2**	3,641.0	0.00027
3.3	27.11	0.037	**8.3**	4,023.9	0.00025
3.4	29.96	0.033	**8.4**	4,447.1	0.00022
3.5	33.12	0.030	**8.5**	4,914.8	0.00020
3.6	36.60	0.027	**8.6**	5,431.7	0.00018
3.7	40.45	0.025	**8.7**	6,002.9	0.00017
3.8	44.70	0.022	**8.8**	6,634.2	0.00015
3.9	49.40	0.020	**8.9**	7,332.0	0.00014
4.0	54.60	0.018	**9.0**	8,103.1	0.00012
4.1	60.34	0.017	**9.1**	8,955.3	0.00011
4.2	66.69	0.015	**9.2**	9,897.1	0.00010
4.3	73.70	0.014	**9.3**	10,938	0.00009
4.4	81.45	0.012	**9.4**	12,088	0.00008
4.5	90.02	0.011	**9.5**	13,360	0.00007
4.6	99.48	0.010	**9.6**	14,765	0.00007
4.7	109.95	0.009	**9.7**	16,318	0.00006
4.8	121.51	0.008	**9.8**	18,034	0.00006
4.9	134.29	0.007	**9.9**	19,930	0.00005