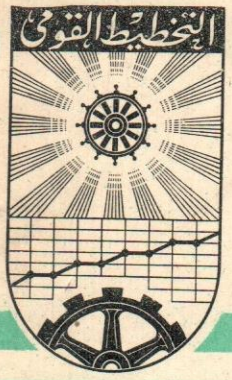
الجمهورية العربية المتحدة



معهد التخطيط القومي

نسخة واحدة فقط

Memo. No. 59o.

Note On Statistical Methods

By

Dr. Moharram W. Mahmoud

B.Sc., M.Sc., M.A., Ph.D.

Operations Research Centre.

10/8/1965.

القاهرة

٣ شارع محمد مظهر - بالزمالك

Contents

This note has been prepared by the author for a course in Statistical methods which is especially designed for those of mathematical background as engineers and scientists.

It was given as lectures for

(i) the group of operations research of the 4th long term training course 1965.

&(ii) the group who attended the short term training period (Jan. - Feb. 1965) at the Operations Research Centre.

The author takes this opportunity to acknowledge the encouragement of Dr Salah Hamid; the general director of the Operations Research Centre.

He would also like to thank Mrs. Nadia Amer and Miss Ellen Zaki for the great care they took in writing this note.

Moharram W. Mahmoud

(I) MOMENTS

(1-1) Moments around the origin:-

$$m'_r(x) = \frac{1}{n}\sum x_i^r \qquad \text{(ungrouped data)}$$

$$= \frac{1}{n}\sum f_i\, x_i^r \qquad \text{(grouped data)}$$

For grouped data x_i denote the mid points of the intervals and f_i the corresponding frequencies

(1-2) Moments around the mean:-

$$m_r(x) = \frac{1}{n}\sum (x_i-\bar{x})^r \qquad \text{(ungrouped data)}$$

$$= \frac{1}{n}\sum f_i(x_i-\bar{x})^r \qquad \text{(grouped data)}$$

Remarks (i) $m'_o(x) = m_o(x) = 1$

(ii) $m'_1(x) = \bar{x}$ = the mean

(iii) $m_1(x) = o$

(iV) $m_2(x) = s^2$ = the variance.

(1.3) Expressing $m_r(x)$ in terms of $m'_r(x)$

$$m_r(x) = \sum_{t=o}^{r} (-1)^t \binom{r}{t} (m'_1)^t\, m'_{r-t}$$

In particular

$$m_2 = m'_2 - m'^2_1$$

$$m_3 = m'_3 - 3m'_2\, m'_1 + 2m'^3_1$$

$$m_4 = m'_4 - 4m'_3\, m'_1 + 6m'_2\, m'^2_1 - 3m'^4_1$$

Remark In the case of frequency distributions with equidistant intervals we replace the mid-points x_i by d_i where

$$d_i = \frac{x_i - a}{\ell}$$

where "a" is an arbitrary origin

"ℓ" is the length of the interval

Then

$$m_r(x) = \ell^r \, m_r(d)$$

(1.4) Measures of skewness

$$\beta_1 = m_3^2 / m_2^3$$

$$\gamma_1 = \sqrt{\beta_1}$$

Measure of Kurtosis

$$\beta_2 = m_4 / m_2^2$$

$$\gamma_2 = \beta_2 - 3$$

(1.5) Numerical Example (illustrative

Intervals	f_i	d_i	d_i^2	d_i^3	d_i^4
10-20	10	-1	1	-1	1
20-30	18	0	0	0	0
30-40	14	1	1	1	1
40-50	8	2	4	8	16
	50	20	56	72	152
$m_r(d)$		$\sum f_i d_i$	$\sum f_i d_i^2$	$\sum f_i d_i^3$	$\sum f_i d_i^4$
		0.4	1.02	1.44	3.04

$m_2(d) = 1.02 - (0.4)^2 = 0.86$

$m_3(d) = 1.44 - 3(1.02)(0.4)+2(0.4)^3 = 0.344$

$m_4(d) = 3.04 - 4(1.44)(0.4)+6(1.02)(0.4)^2-3(0.4)^4 = 1.6394$

$m_1(x) = \ell\, m_1'(d)^{(0)} + a = 10(0.4) + 25 = 29$

$m_2(x) = s^2 = 10^2\,(0.86) = 86$

$m_3(x) = 10^3\,(0.344) = 344$

$m_4(x) = 10^4\,(1.6394) = 16394$

$\beta_1 = (344)^2/(86)^3 = 0.1860$

$\gamma_1 = \sqrt{0.1860} = 0.43$

& $\beta_2 = 16394/7396 = 2.22$

$\therefore \gamma_2 = \beta_2 - 3 = 2.22 - 3 = -0.78$

Exercise given the following frequency table, calculate the mean, the variance and β_1 & β_2

Intervals	f	Intervals	f
55 - 75	3	155 - 175	209
75 - 95	21	175 - 195	81
95 -115	78	195 - 215	21
115 -135	182	215 - 235	5
135 -155	305		905

(II) FUNDAMENTAL CONCEPTS OF PROBABILITY

(2-1) The theory of probability is a branch of applied mathematics dealing with the effects of chance. If we throw a die upon a board we are certain that one of the six faces will turn up, but whether a particular face will show, depends on what we call chance. Also, if equal numbers of white and black balls are put in an urn and we draw one of them blindly, we are certain that its colour will be either white or black, but whether it will be black, that depends on chance.

The word event which we are going to use frequently is used to signify an observation satisfying some specified conditions.

Two events are said to be "equally likely" if after taking into consideration all relevent evidence one of them cannot be expected in preference to the other. e.g. in the case of the urn with equal number of white and black balls, if we draw a ball, it is equally likely to be either white or black.

In the field of statistical analysis there would seem to be two definitions:

(i) Mathematical theory of arrangements which is as old as gambling & playing cards. The probability (p) of an event is the ratio of the no. of ways in which the event may happen divided by the total no. of ways in which the event may or may not happen. This is under the condition that all the events are equally likely. e.g. in the case of an unbiased coin, the probability that the head appears uppermost is $p=\frac{1}{2}$. Also in throwing a die the probability a particular face will show is $p=1/6$.

(ii) The frequency theory: If in a series of n independent trials which are absolutely identical, the event E is found to occur in m trials, then the probability of E is $\frac{m}{n}$.

This gives us a way to estimate probabilities from experimental results in a simple way.

As n increases $\frac{m}{n}$ tends to p, i.e. $p = \lim_{n \to \infty} \frac{m}{n}$.

(2-2) Definition (1) Fundamental probability set (F.P.S.). is that set of individuals or units from which the probability is calculated.
In the case of die, the F.P.S. given by the mathematical theory of arrangements would be 6. If the die is biassed in some way and it is necessary to estimate a probability from the observations, then the F.P.S. would be the total number of throws of a die.

Definition (2) Mutually exclusive: Two events E_1, E_2 are said to be mutually exclusive if no element of the P.P.S. may possess both E_1, E_2. In other words the two events do not occur together.

Remarks (i) $\Pr\{E_1+E_2\}$ means the probability of E_1 or E_2.

(ii) $\Pr\{E_1E_2\}$ means the probability of E_1 & E_2.

(2.3) Basic theorems:-

In the following theorems we are going to assume that the fundamental probability set N, consists of

n_1 elements possessing E_1
n_2 " " E_2
n_{12} " " E_1 & E_2
n_o " " $\bar{E}_1$ & $\bar{E}_2$

(where $\bar{E}_1$ means the event E_1 does not occur)

In other words $N = n_1+n_2+n_{12}+n_o$

Theorem (1) If E_1, E_2 are mutually exclusive and the only possible events, then

$$\Pr\{E_1\} + \Pr\{E_2\} = 1$$

Proof:- Since the two events are mutually exclusive then $n_{12} = o$. Also the two events are the only possible then $n_o = o$. Therefore

$$n_1 + n_2 = N$$

$$\therefore \frac{n_1}{N} + \frac{n_2}{N} = 1$$

$$\therefore \Pr\{E_1\} + \Pr\{E_2\} = 1$$

Theorem (2) $\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1E_2\}$

$$\Pr\{E_1\} = \frac{n_1+n_{12}}{N} \quad , \quad \Pr\{E_2\} = \frac{n_2 + n_{12}}{N}$$

$$\Pr\{E_1E_2\} = \frac{n_{12}}{N}$$

$$\therefore \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1E_2\} = \frac{n_1+n_{12}}{N} + \frac{n_2+n_{12}}{N} - \frac{n_{12}}{N}$$

$$= \frac{n_1 + n_2 + n_{12}}{N}$$

$$= \Pr\{E_1 + E_2\}$$

Cor. If E_1, E_2 are mutually exclusive them

$$\Pr\{E_1 + E_2\} = \Pr\{E_1\} + P_r\{E_2\}$$

In general for k mutually exclusive properties we have

$$\Pr\left\{\sum_{i=1}^{k} E_i\right\} = \sum_{i=1}^{k}\{\Pr(E_i)\}$$

Definition Conditional probability of an event E_2 given event E_1 is the probability of E_2 referred to the F.P.S. for E_1 and it is written $Pr\{E_2 \mid E_1\}$

Theorem (3) $Pr\{E_1E_2\} = Pr\{E_1\} \cdot Pr\{E_2 \mid E_1\}$

$= Pr\{E_2\} \; Pr\{E_1 \mid E_2\}$

Proof:- $Pr\{E_1\} = \frac{n_1 + n_{12}}{N}$

$Pr\{E_2\} = \frac{n_2 + n_{12}}{N}$

$Pr\{E_2 \mid E_1\} = \frac{n_{12}}{n_1 + n_{12}}$

$Pr\ E_1\ E_2 = \frac{n_{12}}{n_2 + n_{12}}$

$\therefore Pr\{E_1\} Pr\{E_2 \mid E_1\} = \frac{n_1 + n_{12}}{N} \times \frac{n_{12}}{n_1 + n_{12}} = \frac{n_{12}}{N}$

$Pr\{E_2\} Pr\{E_1 \mid E_2\} = \frac{n_2 + n_{12}}{N} \times \frac{n_{12}}{n_2 + n_{12}} = \frac{n_{12}}{N}$

$\therefore Pr\{E_1\} Pr\{E_2 \mid E_1\} = Pr\{E_2\} Pr\{E_1 \mid E_2\}$

$= \frac{n_{12}}{N} = Pr\{E_1E_2\}$

(2-4) Independence: E_1 is independent of E_2 if

$$Pr\{E_1\} = Pr\{E_1 \mid E_2\}.$$

Consequently this implies (theorem 3) that E_2 is independent of E_1

Cor. If E_1, E_2 are mutually independent then

$$\Pr\{E_1 E_2\} = \Pr\{E_1\} \Pr\{E_2\}$$

In general

$$\Pr\left\{\prod_{i=1}^{k} E_i\right\} = \prod_{i=1}^{k} \Pr\{E_i\}$$

III Random Variables

Discontinuous random

Variables (Binomial distribution)

Continuous random variables (Normal distribution

(3.1) Statistical data consist of values obtained by measurements of a certain character or characters of a number of individuals. The observed values are not usually all the same. As the number of observations is increased the proportion of observations less than a fixed number X say become more & more stable. From the viewpoint of the frequency theory of probability this defines the probability that the observed value of the character falls below the number X.

The theoritical concept which we use to correspond with "observed values" is the random variable. x is said to be a random variable, if for any number X, there is a probability F(X) that x is less than or equal to X i.e.

$$Pr\{x \leq X\} = F(X) \text{ exists.}$$

Replacing X by x in F(X) we get what is called the cumulative distribution function of x.

The properties of F(x) are

(i) $0 \leq F(x) \leq 1$

(ii) If $x_1 < x_2$ then $F(x_1) \leq F(x_2)$

i.e. F(x) is non-decreasing function.

As an example, let us take an unbiased die.

If x denotes the number of dots on the upper-most face, then

$F(x) = 0$ for $x < 1$

$F(x) = \frac{1}{6}$ for $1 \leq x < 2$

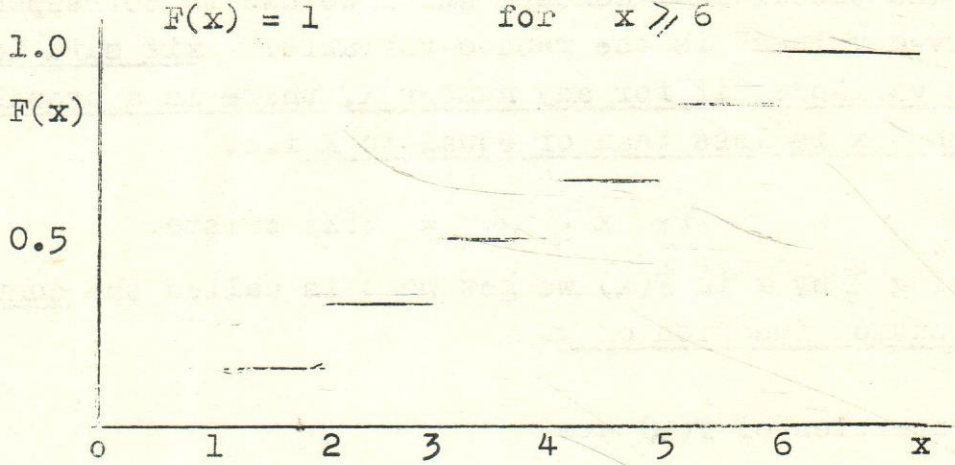
$F(x) = \frac{1}{3}$ for $2 \leq x < 3$

$F(x) = \frac{1}{2}$ for $3 \leq x < 4$

$F(x) = \frac{2}{3}$ for $4 \leq x < 5$

$F(x) = \frac{5}{6}$ for $5 \leq x < 6$

$F(x) = 1$ for $x \geq 6$



(3.2) Discontinuous random variables

Random variables that take isolated values as number of children in a family are called discontinuous. Their cumulative distribution functions are step-functions and the heights of the steps being equal to the corresponding probabilities. Suppose in general we have k definite values $x_1 < x_2 < x_3 \ldots < x_k$ and the probabilities of x taking these values are p_1, p_2, p_3, ..., p_k respectively, i.e. $\Pr\{x = x_i\} = p_i$ $i=1,2,..,k$

If $x_{i-1} \leq x < x_i$ then

$$F(x) = \Pr\{x \leq x_{i-1}\} = \sum_{j=1}^{i-1} p_j$$

Since the events $x = x_i$ are mutually exclusive then

$$\sum_{i=1}^{k} p_i = 1$$

A discontinuous random variable may have an infinity of possible discrete values

$$\ldots\ x_{-2}\ ,\ x_{-1},\ x_0,\ x_1,\ x_2,\ \ldots$$

If p_i be the probability of such a variable taking x_i then the conditions

$$p_i \geq 0$$

$$\sum_{-\infty}^{\infty} p_i = 1$$

must be satisfied.

Example Consider that we have 3 coins and we throw them on a board. The possible outcomes

(F.P.S.) are :-		No. of cases	Probability
No head :	T T T	1	1/8
One head :	H T T T H T T T H	3	3/8
Two heads:	H H T H T H T H H	3	3/8
Three heads:	H H H	1	1/8

In this case we have n=3 independent trials. If x denotes the number of heads in tossing the three coins (or tossing one coin three times) then

$$\Pr\{x = 0 \mid n=3\} = 1/8$$

$$\Pr\{x = 1 \mid n=3\} = 3/8$$

$$\Pr\{x = 2 \mid n=3\} = 3/8$$

$$\Pr\{x = 3 \mid n=3\} = 1/8$$

It is obvious that $\sum p_i = 1/8 + 3/8 + 3/8 + 1/8$

$$= 1$$

Now since the probability that the head occurs in tossing a single coin is ½, then we find that the probabilities that x = 0,1,2,3, are theterms in the expansion $(\frac{1}{2} + \frac{1}{2})^3$ respectively This of course is under the assumption that the two faces are equally likely. On the other hand if p denotes the probability that a head (H) occurs in tossing a single coin and q = 1-p the probability that a tail(T) occurs i.e. p will associated with (H) and q with (T), then

x	P(x)
0	q^3
1	$3\,q^2\,p$
2	$3\,q\,p^2$
3	p^3

$$\sum_{x=0}^{3} p(x) = q^3 + 3q^2 + 3qp^2 + p^3 = (q+p)^3 = 1$$

In general, suppose the probability of a success in a trial is "p" and the probability of a failure is q = 1-p. We can represent these probabilities in a functional form $f(\alpha)$ where $f(\alpha) = p$ for $\alpha = 1$ i.e. a success & $f(\alpha) = q$ for $\alpha = o$ i.e. a failure. The probabilities associated with n trials which are mutually independent in the probability sense is

$$f(\alpha_1)\, f(\alpha_2) \ldots\ f(\alpha_n)$$

The probability of x successes & n-x failures in a specified order is

$$\left\{f(1)\right\}^{x} \left\{f(o)\right\}^{n-x} = p^{x} q^{n-x}$$

The no. of orders in which x successes & n-x failures can occur is ${}^{n}C_{x}$. These orders are mutually exclusive events, then

$$p(x) = {}^{n}C_{x}\ p^{x}\ q^{n-x} \qquad (1)$$

This is the general term in the expansion $(q+p)^{n}$. $p(x)$ is called the <u>binomial probability distribution</u>

The mean of this distribution is given by

$$\mu_1'(x) = \sum_{x=o}^{n} x\ {}^{n}C_{x}\ p^{x}\ q^{n-x}$$

$$= np. \qquad (2)$$

Also

$$\mu_2'(x) = \sum_{x=o}^{n} x^2\ {}^{n}C_{x}\ p^{x}\ q^{n-x}$$

$$= n(n-1)p^2 + np.$$

Hence $\mu_2(x) = \mu_2'(x) - \left\{\mu_1'(x)\right\}^2$

$$= npq.$$

i.e. the standard deviation = $\sqrt{npq}$.

Exercise

prove that $\beta_1 = \frac{1}{npq} - \frac{4}{n}$

$\beta_2 = 3 + \frac{1-6pq}{npq}$

Relationship between the binomial distribution and the incomplete beta function.

The complete beta function

$$B(k,n-k+1) = \int_0^1 x^{k-1} (1-x)^{n-k} \, dx = \frac{\Gamma(k)\Gamma(n-k+1)}{\Gamma(n+1)}$$

$$= \frac{(k-1)!\,(n-k)!}{n!}$$

The incomplete beta function

$$B_p(k,n-k+1) = \int_0^p x^{k-1} (1-x)^{n-k} \, dx$$

$$\therefore \quad I_p(k,n-k+1) = B_p(k,n-k+1) \,/\, B(k,n-k+1)$$

$$= \frac{n!}{(k-1)!(n-k)!} \int_0^p x^{k-1} (1-x)^{n-k} \, dx$$

$$= \frac{n!}{(k-1)!(n-k)!} \left[\frac{p^k q^{n-k}}{k} + \frac{p^{k+1} q^{n-k-1} (n-k)}{k(k+1)} + \ldots \right.$$

$$\left. + \frac{p^n (n-k)!}{k(k+1) \ldots (n-1)n} \right]$$

$$= \frac{n!}{k!(n-k)!} p^k q^{n-k} + \frac{n!}{(k+1)!(n-k-1)!} p^{k+1} q^{n-k-1} +$$

$$\ldots\ldots\ldots + p^n$$

$$= P_{k,n} + P_{k+1,n} + \ldots + P_{n,n}$$

where

$$P_{k,n} = Pr\left\{x = k \mid n\right\}$$

$$\therefore \quad I_p(k,n-k+1) = \sum_{x=k}^{n} P_{x,n} \qquad (1)$$

Similarly it can be shown that

$$I_p(k+1,n-k) = P_{k+1,n} + P_{k+2,n} + \ldots + P_{n,n}$$

$$= \sum_{x=k+1}^{n} P_{x,n} \qquad (2)$$

From (1) & (2) we get

$$P_{k,n} = I_p(k,n-k+1) - I_p(k+1,n-k) \qquad (3)$$

Fitting a binomial distribution

Throwing 5 coins 100 times, the following table gives the frequency distribution of the number of coins (x) on which the head shows uppermost.

x:	0	1	2	3	4	5
f:	2	14	20	34	22	8

To fit a binomial distribution we 1st calculate the mean of the given frequency distribution

i.e. $\bar{x} = 2.84$

Then equating $\bar{x}$ to the mean of the binomial

i.e. $\bar{x} = np$ where $n = 5$ we get

$$5p = 2.84.$$
$$\therefore \quad p = 0.568$$

$$\therefore \; p(x) = \binom{5}{x} 0.568^{x} \quad (0.432)^{5-x}$$

where $x = 0,1,2,3,4,5.$

Values of $p(x)$ are calculated for the corresponding values of x as given in the table below:

x	p(x)	Expected frequencies	observed frequencies
0	0.015	1.5	2
1	0.099	9.9	14
2	0.260	26.0	20
3	0.342	34.2	34
4	0.225	22.5	22
5	0.059	5.9	8
		100.0	100.0

Note: The recurrent relation is

$$p(x+1) = \frac{n-x}{x+1} \cdot \frac{p}{q} \; p(x) \qquad x=0,1,2,\ldots,n-1$$

e.g. $p(1) = n \cdot \frac{p}{q} \; p(0)$

$p(2) = \frac{n-1}{2} \cdot \frac{p}{q} \; p(1)$

.....

$p(n) = \frac{1}{n} \cdot \frac{p}{q} \; p(n-1)$

(3-3) Continuous random variables

A continuous random variable (x) is a variable which may take any real value between certain limits such as age, height, weight etc. Its cumulative distribution function F(x) is continuous and differentiable for all values of x. F'(x) must exist and is called the probability density function of x and usually denoted by p(x),

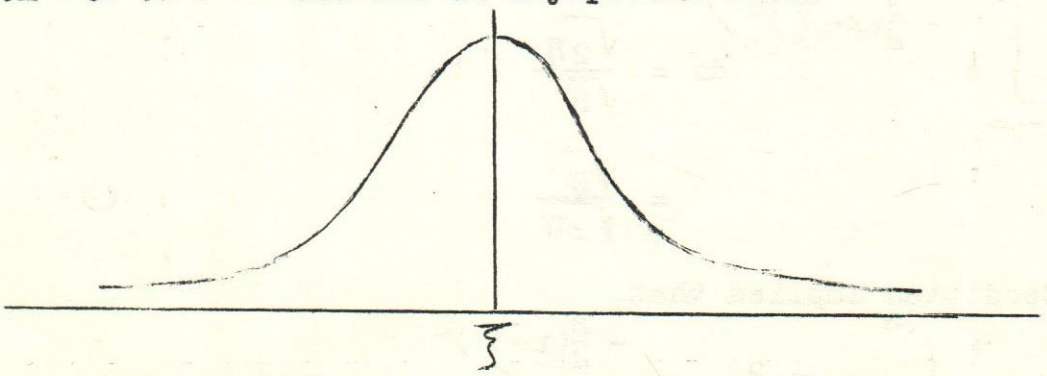
i.e. $p(x) = F'(x)$. Also

$$p(x) \geq 0$$

& $$\int_{-\infty}^{\infty} p(x)\,dx = 1$$

As an example, let us consider the normal distribution. It was discovered by De Moivre in 1733, and at the same time but independently, Laplace & gauss derived the formula for this distribution.

The normal curve is a belt shaped with its ends extending from $-\infty$ to $+\infty$ and has an asymptotic base.



To construct a mathematical function for this curve,

$$\frac{dy}{dx} = 0 \quad \text{at} \quad x = \xi$$

& $$\frac{dy}{dx} = 0 \quad \text{at} \quad y = 0$$

$\therefore$ $$\frac{dy}{dx} = -By(x- \quad) \qquad (1)$$ B is + ve.

It is clear that for $x < \xi$, $\frac{dy}{dx}$ is + ve and for $x > \xi$ $\frac{dy}{dx}$ is - ve. This means that the function attains its maximum at $x=\xi$. Solving the differential equation (1) we get

$$\int \frac{dy}{y} = -B \int (x-\xi)\, dx$$

$$\therefore \quad \log y = -\frac{B}{2}(x-\xi)^2 + \log C$$

$$\therefore \qquad y = p(x) = c\, e^{-\frac{B}{2}(x-\xi)^2} \qquad (2)$$

where C & B are Constants to be determined. We are going to impose two conditions:-

(i) The area below the curve should be equal to unity.

(ii) The variance is σ^2 i.e. $\mu_2 = \sigma^2$

The 1st condition implies that

$$C \int_{-\infty}^{\infty} e^{-\frac{B}{2}(x-\xi)^2} dx = 1$$

But $$C \int_{-\infty}^{\infty} e^{-\frac{B}{2}(x-\xi)^2} dx = \frac{\sqrt{2\pi}}{\sqrt{B}} C$$

$$\therefore \quad C = \frac{\sqrt{B}}{\sqrt{2\pi}} \qquad (3)$$

The 2nd Condition implies that

$$\mu_2 = C \int_{-\infty}^{\infty} (x-\xi)^2 \, e^{-\frac{B}{2}(x-\xi)^2} dx = \sigma^2$$

But

$$C \int_{-\infty}^{\infty} (x-\xi)^2 \, e^{-\frac{B}{2}(x-\xi)^2} dx = \frac{\sqrt{2\pi}}{B^{3/2}} C = \frac{1}{B}$$

i.e. $B = \frac{1}{\sigma^2}$ (4)

$\therefore \quad C = \frac{1}{\sqrt{2\pi}\,\sigma}$ (5)

$\therefore \quad p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \; e^{-\frac{1}{2\sigma^2}(x-\xi)^2}$ (6)

ξ & σ are the two parameters of the normal distribution.

If $Z = \frac{x-\xi}{\sigma}$ i.e. the standardized normal variate

then

$$p(z) = \frac{1}{\sqrt{2\pi}} \; e^{-\frac{1}{2}z^2} \qquad (7)$$

which is the unit normal distribution

Remarks

(1) Now, if ξ is known, formula (6) represents a family of normal curves having the same mean but differs in their variances. If σ is known, the same formula represents a family of normal curves having the same variance but differs in their means

(2) $$Pr\{x \leq x_1\} = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{x_1} e^{-\frac{1}{2}\left(\frac{x-\xi}{\sigma}\right)^2} dx$$

Write $Z = \frac{x-\xi}{\sigma}$

$\therefore dZ = dx/\sigma$

$$\therefore P_r\{x \leq x_1\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_1} e^{-\frac{1}{2}z^2} dz$$

$$= Pr\{Z \leq Z_1\} \qquad (8)$$

$\Pr\left\{x \leq x_1\right\}$ is called the cumulative distribution function denoted by $F(x)$ at $x = x_1$.

Also

$$\Pr\left\{Z \leq -Z_1\right\} = \int_{-\infty}^{-Z_1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2} dZ$$

$$= \frac{1}{\sqrt{2\pi}} \int_{Z_1}^{\infty} e^{-\frac{1}{2}Z^2} dZ = 1 - \Pr\left\{Z \leq Z_1\right\}$$

The values of the cumulative distribution function at different values of Z are given in "Tables of the normal probability function"

Example: The mean weight of 500 students at a certain college is 151 lb. and the standard deviation is 15 lb. Assuming that the weights are normally distributed, find how many students weight (i) more than 185 lb.

(ii) between 120 & 155 lb.

(i) standardized value of 185 is

$$Z = \frac{185 - 151}{15} = \frac{34}{15} = 2.30$$

$$\Pr\left\{Z > 2.3\right\} = 1 - \Pr\left\{Z \leq 2.3\right\} = 1 - .9893$$

$$= .0107$$

$\therefore$ The number of students is = 500 X 0.0107 = 5

(ii) standardized values

of 120 & 155 are

$$Z_1 = \frac{120 - 151}{15} = -\frac{31}{15} = -2.0$$

$$Z_2 = \frac{155 - 151}{15} = \frac{4}{15} = 0.3$$

$$\Pr\left\{-2.0 \leqslant Z \leqslant 0.3\right\} = \Pr\left\{Z \leqslant .3\right\} - \Pr\left\{Z \leq -2.0\right\}$$

$$= 0.6179 - .0228 = 0.5951$$

The number of students is = 500 X 0.5951 = 298

Fitting a normal curve

given a frequency distribution, it is required to fit a normal curve having the same mean and variance.

		(1)	(2)	(3)	(4)
Intervals	Observed frequencies	Standardized values	Areas $\Pr\{Z \leqslant Z\}$	Differences (Δ)	Expected freq.
60		-3.66	0.0001		
60- 65	3	-2.74	0.0031	0.0030	3.
65- 70	21	-1.83	0.0336	0.0305	30.5
70- 75	150	-0.91	0.1814	0.1478	147.8
75- 80	335	0.01	0.5040	0.3226	322.6
80- 85	326	0.93	0.8238	0.3198	319.8
85- 90	135	1.85	0.9678	0.1440	144.0
90- 95	26	2.77	0.9972	0.0294	29.4
95-100	4	3.68	0.9999	0.0027	2.7
	1000				

The following steps are followed:-

(i) Calculate the mean and the variance for the observed frequency distribution i.e.,

$\bar{x} = 79.945$

$s = 5.445$

(ii) Calculate the standardized values for the upper lemits of the intervals, e.g. the upper limit of the 1st interval becomes

$$\frac{65-79.945}{5.445} = -2.74$$

and so on as in column (1)

(iii) Calculate the areas below these standardized values e.g.

$$\Pr\{x \leqslant -2.74\} = 0.0031$$

$$\Pr\{x \leqslant -1.83\} = 0.0336$$

and so on as in column (2)

(iv) Calculate the areas corresponding to the intervals ( ) by successive subtraction e.g. the area corresponding to the interval 60-65

$= 0.0031 - 0.0001 = 0.0030.$

and the area corresponding to the interval
65-70 = 0.0336 - 0.0031 = 0.0305
and so on as in column (3)

(v) Calculate the expected frequencies N where N is the total number of frequencies.
e.g. the expected frequency corresponding to the interval

60-65 = 1000 X 0.003 = 3
the expected frequency corresponding to the interval

65-70 = 1000 X 0.0305 = 30.5 and so on as in column (4)

(IV) Expected Values and Characteristic Functions

IV Expected Values and Characteristic functions

(4.1) Expected Values

Suppose N observations each of which may have any one of k finite number of values $x_1, x_2, \ldots, x_k$. Suppose further that f_1 of these observations have the value x_1, f_2 have the value x_2, ... and in general f_i have the value x_i where $\sum f_i = N$.

$$\bar{x} = \frac{1}{N} \sum f_i x_i = \sum_i x_i \left(\frac{f_i}{N} \right)$$

as $n \to \infty$, $f_i / N \to p_i$

$\therefore$ The quantity $\sum_{i=1}^{k} x_i p_i$ is called the expected value of the discontinuous random variable x which takes the possible values $x_1, x_2, \ldots, x_k$ with probabilities $p_1, p_2, \ldots, p_k$ respectively. The expected value is denoted by $\mathcal{E}(x)$ i.e.

$$\mathcal{E}(x) = \sum_{i=1}^{k} x_i p_i$$

The expected value is a theoritical parameter analogous to the observed sample mean.

The definition may be extended to discontinous random values with an infinite number of possible values, i.e.

$$\mathcal{E}(x) = \sum_{-\infty}^{\infty} x_i p_i$$

For continuous random variables

$$\mathcal{E}(x) = \int_{-\infty}^{\infty} x \; p(x) \; dx$$

Any function of a random variable x, say f(x), is also a random variable and has an expected value

$$\mathcal{E}\{f(x)\} = \sum f(x_i)\, p_i \quad \text{for discontinuous random variables and}$$

$$\mathcal{E}\{f(x)\} = \int f(x)\, p(x)\, dx \quad \text{for continuous random variables.}$$

Note The joint probability density function $p(x,y)$ has the following properties

(i) $p(x,y) \geq 0$

(ii) $Pr\{\alpha < x < \beta \; ; \quad \alpha_1 < y < \beta_1\} = \int_{\alpha_1}^{\beta_1}\int_{\alpha}^{\beta} p(x,y)\,dx\,dy$

(iii) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\, dx\, dy = 1$

(iv) $p_1(x) = \int_{-\infty}^{\infty} p(x,y)\, dy$

$$p_2(y) = \int_{-\infty}^{\infty} p(x,y)\, dx$$

(v) If x,y are independent

$$p(x,y) = p_1(x) \cdot p_2(y).$$

Theorems on expected values

(i) $\mathcal{E}(A) = A$

(ii) $\mathcal{E}(Ax) = A\,\mathcal{E}(x)$.

(iii) $\mathcal{E}(x+y) = \mathcal{E}(x) + \mathcal{E}(y)$

Consider the case x,y are continuous random variables then

$$\mathcal{E}(x+y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)\, p(x,y)\, dx\,dy$$

$$= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} p(x,y)dy\ dx + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} p(x,y)dx\ dy$$

$$= \int_{-\infty}^{\infty} x\ \ p_1(x)\ dx + \int_{-\infty}^{\infty} y\ p_2(y)\ dy$$

$$= E(x) + E(y)$$

(iv) If x & y are independent then

$$E(xy) = E(x) \cdot E(y)$$

$$E(xy) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\ p(x,y)\ dx\ dy$$

$$= \int\int_{-\infty}^{\infty} xy\ p_1(x) \cdot p_2(y)\ dx\ dy$$

$$= \int_{-\infty}^{\infty} x\ p_1(x)\ dx \int_{-\infty}^{\infty} y p_2(y)\ dy$$

$$= E(x) \cdot E(y)$$

Remarks

(i) The moments of a distribution are the expected values of the powers of the random variable which has the given distribution. i.e.

$$\mu'_r(x) = E(x^r) = \int_{-\infty}^{\infty} x^r\ p(x)\ dx$$

$$\text{or} \quad = \sum_{-\infty}^{\infty} x_i^r\ p_i$$

$$\mu_r(x) = E\{(x-\mu'_1)^r\} = \int_{-\infty}^{\infty} (x-\mu'_1)^r\ p(x)\ dx$$

$$\text{or} \quad = \sum_i (x_i-\mu'_1)^r\ p_i$$

(ii) The variance

$$\sigma_x^2 = \mathcal{E}\left\{x - \mathcal{E}(x)\right\}^2$$

(iii) The coefficient of correlation

$$\rho = \frac{\mathcal{E}\left\{x - \mathcal{E}(x)\right\}\left\{y - \mathcal{E}(y)\right\}}{\sqrt{\mathcal{E}\left\{x - \mathcal{E}(x)\right\}^2 . \mathcal{E}\left\{y - \mathcal{E}(y)\right\}^2}}$$

Linear functions of random variables

Given $\mathcal{E}(x_i) = \zeta_i$, $\mathcal{E}(x_i - \mathcal{E}(x_i))^2 = \sigma_i^2$, then the standard error of any linear function

$$y = \sum_{i=1}^{n} \alpha_i \; x_i$$

of n random variables $x_1, x_2, \ldots, x_n$ is

$$\sigma_y^2 = \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2 + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \alpha_i \alpha_j \sigma_i \sigma_j \rho_{ij}$$

α are constants, σ_i, σ_j are standard errors of x_i, x_j & ρ_{ij} their correlation coefficient.

$$\mathcal{E}(y) = \mathcal{E}\left\{\sum_{i=1}^{n} \alpha_i \; x_i\right\} = \sum_{i=1}^{n} \alpha_i \zeta_i$$

$$\sigma_y^2 = \mathcal{E}\left\{y - \mathcal{E}(y)\right\}^2 = \mathcal{E}\left\{\sum_{i=1}^{n} \alpha_i \; x_i - \sum_{i=1}^{n} \alpha_i \zeta_i\right\}^2$$

$$= \mathcal{E}\left\{\sum_{i=1}^{n} \alpha_i (x_i - \zeta_i)\right\}^2$$

$$= \mathcal{E}\left\{\sum_{i=1}^{n} \alpha_i^2 (x_i - \zeta_i)^2 + 2 \sum_i \sum_{j=i+1} \alpha_i \alpha_j (x_i - \zeta_i)(x_j - \zeta_j)\right\}$$

$$= \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2 + 2 \sum_i \sum_{j=i+1} \alpha_i \alpha_j \sigma_i \sigma_j \rho_{ij}$$

Examples

(1) $\bar{x} = \frac{1}{n} \Sigma x_i$

Here $\alpha = \frac{1}{n}$ $\therefore \mathcal{E}(\bar{x}) = \frac{1}{n} \sum_{i=1}^{n} \zeta_i$

& $\sigma_{\bar{x}}^2 = \frac{1}{n^2} \Sigma \sigma_i^2 + \frac{2}{n^2} \sum_i \sum_{j=i+1} \sigma_i \sigma_j \rho_{ij}$

(2) If $y = x_1 \pm x_2$

$$\therefore \mathcal{E}(y) = \sum_{i=1}^{2} \zeta_i = \zeta_1 \pm \zeta_2$$

$$\sigma_y^2 = \sigma_1^2 + \sigma_2^2 \pm 2 \sigma_1 \sigma_2 \rho_{12}$$

(3) If $y = x_1 \pm x_2$ and x_1, x_2 are independent then

$$\sigma_y^2 = \sigma_1^2 + \sigma_2^2$$

(4) If the x's are independent and have the same standard error then

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

(5) If $\bar{x}_1$ is the mean of n_1 independent random variables, each with standard deviation σ_1 and $\bar{x}_2$ the mean of n_2 independent random variables each with standard deviation σ_2, then

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

Quadratic functions of random variables

Given $x_1, x_2, \ldots, x_n$ are independent random variables, $\mathcal{E}(x_i) = \zeta_i$, $\mathcal{E}(x_i-\zeta_i)^2 = \sigma_i^2$ for $i=1,2,\ldots,n$ then

$$\mathcal{E}\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2\right\} = \frac{n-1}{n}\sum_{i=1}^{n}\sigma_i^2 + \sum_{i=1}^{n}(\zeta_i-\bar{\zeta})^2$$

Let $y = \sum(x_i-\bar{x})^2$

$$\therefore\ y = \sum\left[(x_i-\zeta_i) + (\zeta_i-\bar{\zeta}) - (\bar{x}-\bar{\zeta})\right]^2$$

$$= \sum_i(x_i-\zeta_i)^2 + \sum_i(\zeta_i-\bar{\zeta})^2 + n(\bar{x}-\bar{\zeta})^2$$

$$+ 2\sum(x_i-\zeta_i)(\zeta_i-\bar{\zeta}) - 2n(\bar{x}-\bar{\zeta})^2$$

$$= \sum(x_i-\zeta_i)^2 + \sum_i(\zeta_i-\bar{\zeta})^2 - n(\bar{x}-\bar{\zeta})^2 + 2\sum(x_i-\zeta_i)(\zeta_i-\bar{\zeta})$$

$$\mathcal{E}(y) = \sum_{i=1}^{n}\sigma_i^2 + \sum_{i=1}^{n}(\zeta_i-\bar{\zeta})^2 - n\,\mathcal{E}(\bar{x}-\bar{\zeta})^2 + 0$$

$$= \sum_{i=1}^{n}\sigma_i^2 + \sum_{i=1}^{n}(\zeta_i-\bar{\zeta})^2 - \frac{1}{n}\mathcal{E}\left\{\sum(x_i-\zeta_i)^2\right\}$$

$$= \sum_{i=1}^{n} \sigma_i^2 + \sum_{i=1}^{n} (\zeta_i - \bar{\zeta})^2 - \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2$$

$$= \frac{n-1}{n} \sum_{i=1}^{n} \sigma_i^2 + \sum_{i=1}^{n} (\zeta_i - \bar{\zeta})^2$$

Lor:- (1) Suppose the x's have the same mean and standard deviation i.e.

$$\zeta_i = \zeta \ , \ \sigma_i = \sigma$$

$$\therefore \bar{\zeta} = \frac{1}{n} \sum_i \zeta_i = \zeta$$

$$\mathcal{E}(y) = \mathcal{E} \sum (x_i - \bar{x})^2 = (n-1)\sigma^2$$

$$\frac{1}{n} \left\{ \sum_i (x_i - \bar{x})^2 \right\} = \frac{n-1}{n} \sigma^2$$

$$\& \ \mathcal{E} \left\{ \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 \right\} = \sigma^2$$

Exercise:

Show that for two discontinuous random variables x&y

(i) $\mathcal{E}(x+y) = \mathcal{E}(x) + \mathcal{E}(y)$

(ii) $\mathcal{E}(xy) = \mathcal{E}(x)\,\mathcal{E}(y)$ if x & y are independent

x_i \ y_j	1 ..... j	
1	p_{11} ..... p_{1j}	$p_{1.}$
.	. ...	.
.	. ...	.
.	. ...	.
i	p_{i1} ..... p_{ij}	$p_{i.}$
.	. .	
.	. .	
	$p_{.1}$ ..... $p_{.j}$	1

$$\mathcal{E}(x_i + y_j) = \sum_j \sum_i p_{ij}(x_i + y_j)$$

$$= \sum_j \sum_i p_{ij} x_i + \sum_j \sum_i p_{ij} y_j$$

$$= \sum_i p_{i.} x_i + \sum_j p_{.j} y_j$$

$$= \mathcal{E}(x) + \mathcal{E}(y)$$

Remember that for x&y independent $p_{ij} = p_{i.} p_{.j}$

(4.2) Characteristic functions

$\varphi_x(t)$ is defined as the characteristic function of the random variable x if

$$\varphi_x(t) = \mathcal{E}(e^{itx})$$

This function will always exist since

$$|e^{itx}| = |\cos^2 tx + \sin^2 tx| = 1$$

$$\therefore \varphi_x(t) = \sum e^{itx} p(x) \text{ for discontinuous variables}$$

$$= \int e^{itx} p(x)\, dx \text{ for continuous variables}$$

Properties of characteristic functions

1- If "a" is constant then

$$\varphi_x(at) = \varphi_{ax}(t)$$

2- If $x_1, x_2, \ldots, x_n$ are independent

random variables then,

$$\varphi_{\Sigma x_j}(t) = \prod_{j=1}^{n} \varphi_{x_j}(t)$$

This is obvious because

$$\varphi_{\Sigma x}(t) = \mathcal{E}(e^{it\Sigma x_j}) = \mathcal{E}\left\{e^{itx_1} e^{itx_2} \ldots e^{itx_n}\right\}$$

$$= \mathcal{E}(e^{itx_1})\, \mathcal{E}(e^{itx_2}) \ldots \mathcal{E}(e^{itx_n})$$

$$= \varphi_{x_1}(t) \cdot \varphi_{x_2}(t) \ldots \varphi_{x_n}(t)$$

$$= \prod_{j=1}^{n} \varphi_{x_j}(t)$$

Lor (1) If the random variables have the same distribution then

$$\varphi_{\Sigma x}(t) = \left\{\varphi_x(t)\right\}^n$$

Consequently it follows that

$$\varphi_{\bar{x}}(t) = \left\{\varphi_x\left(\frac{t}{n}\right)\right\}^n$$

Characteristic functions as moment generating functions

$$\begin{aligned}\varphi_x(t) &= \int_{-\infty}^{\infty} e^{itx}\, p(x)\, dx \\ &= \int_{-\infty}^{\infty} p(x) \sum_{r=0}^{\infty} \frac{(itx)^r}{r!}\, dx \\ &= \sum_{r=0}^{\infty} \frac{(it)^r}{r!} \int_{-\infty}^{\infty} x^r\, p(x)\, dx \\ &= \sum_{r=0}^{\infty} \frac{(it)^r}{r!}\, \mu'_r(x).\end{aligned}$$

∴ Provided we can expand $\varphi_x(t)$ we can pick the moments around the origin. They are the coefficients of $\frac{(it)^r}{r!}$ in the expansion of $\varphi_x(t)$.

Also it can be shown that

$$i\mu'_1 = \varphi'_x(o)$$

$$i^2\mu'_2 = \varphi''_x(o)$$

.....

$$i^r\mu'_r = \varphi^{(r)}_x(o).$$

Characteristic functions used for calculating cumulants

The cumulative function is defined as

$$\psi_x(t) = \log\left\{\varphi_x(t)\right\}$$

$$\therefore \psi_x(t) = \log\left[1 + \frac{it}{1!}\mu'_1 + \frac{(it)^2}{2!}\mu'_2 + \frac{(it)^3}{3!}\mu'_3 + \ldots\right]$$

$$= \frac{it}{1!}\mu'_1 + \frac{(it)^2}{2!}\mu'_2 + \frac{(it)^3}{3!}\mu'_3 + \ldots$$

$$- \tfrac{1}{2}\left[\frac{it}{1!}\mu'_1 + \frac{(it)^2}{2!}\mu'_2 + \frac{(it)^3}{3!}\mu'_3 + \ldots\right]^2$$

$$+ \frac{1}{3}\left[\frac{it}{1!}\mu'_1 + \frac{(it)^2}{2!}\mu'_2 + \frac{(it)^3}{3!}\mu'_3 + \ldots\right]^3$$

$$- \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$= itk_1 + \frac{(it)^2}{2!}k_2 + \frac{(it)^3}{3!}k_3 + \frac{(it)^4}{4!}k_4 + \ldots$$

Equating the coefficients of the same power of t on both sides we get

$$k_1 = \mu'_1 \quad , \quad k_2 = \mu_2$$

$$k_3 = \mu_3 \quad , \quad k_4 = \mu_4 - 3\mu_2^2$$

$$k_5 = \mu_5 - 10\mu_3\mu_2$$

Examples

(i) For the normal distribution we have

$$\varphi_x(t) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\xi}{\sigma}\right)^2} e^{itx}\, dx$$

$$= e^{i\xi t - \frac{1}{2}t^2\sigma^2}$$

$$\therefore \psi_x(t) = \log \varphi_x(t) = i\xi t + \frac{(it)^2}{2}\sigma^2$$

$$\therefore \quad k_1 = \xi$$

$$k_2 = \sigma^2$$

$$k_r = 0 \quad \text{for} \quad r > 2$$

(ii) For the binomial distribution

$$\varphi_x(t) = \sum_{k=0}^{n} e^{itx} \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{k=0}^{n} \binom{n}{x} (p e^{it})^x q^{n-x}$$

$$= (q + p e^{it})^n$$

$$\psi_x(t) = \log \varphi_x(t) = n \log (q + p e^{it})$$

$$= n \log \left[1 - p(1-e^{it})\right]$$

$$= n \log \left[1 + p \left\{\frac{it}{1!} + \frac{(it)^2}{2!} + \cdots\right\}\right]$$

$$= \frac{(it)}{1!} np + \frac{(it)^2}{2!} n(p-p^2)$$

$$+ \frac{(it)^3}{3!} n(p - 3p^2 + 2p^3)$$

$$+ \frac{(it)^4}{4!} n(p-7p^2 + 12p^3 - 6p^4)$$

$$+ \cdots\cdots\cdots\cdots\cdots$$

$$\therefore \quad k_1 = np$$

$$k_2 = n(p-p^2) = npq$$

$$k_3 = n(p-3p^2+2p^3) = np(1-p)(1-2p)$$

$$= npq(q-p)$$

$$k_4 = np(1-7p+12p^2-6p^3)$$

$$= np\left[(1-p)(1-6p)+6p^2(1-p)\right]$$

$$= np(1-p)(1-6p+6p^2)$$

$$= npq(1-6pq)$$

Exercises (1) Given $p(x) = \frac{e^{-x}\, x^{\ell-1}}{(1)}$, the gamma distribution with parameter ℓ, using the cumulant function prove that $2\beta_2 - 3\beta_1 - 6 = 0$

(2) Given $p_k = \frac{\lambda^k e^{-\lambda}}{k!}$ prove that

$$\psi_x(t) = -\lambda(1 - e^{it})$$

The Inverse theorem

This theorem states that the characteristic function uniquely determines the distribution function.

(i) If x is an absolutely continuous random variable, $p(x)$ is the p.d.f. and $\varphi_x(t)$ is its characteristic function, then

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi_x(t)\, dt$$

(ii) If p_k is the p.d.f. for a discontinuous random variable, then

$$p_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itk} \varphi_k(t)\, dt$$

Example : Given

$$\varphi_x(t) = e^{-\frac{1}{2} t^2 \sigma^2}$$

then

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx - \frac{1}{2} t^2 \sigma^2}\, dt$$

$$p(x) = \frac{1}{2\pi}\, e^{-\frac{x^2}{2\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(t\sigma + \frac{ix}{\sigma})^2}\, dt$$

Taking $u = t + \frac{ix}{\sigma}$ then

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{1}{2}\frac{x^2}{\sigma^2}}$$

One-to-one transformation

Given $p(x)$ the probability density function of x & $y = f(x)$, we want $p(y)$ the probability density function of y.

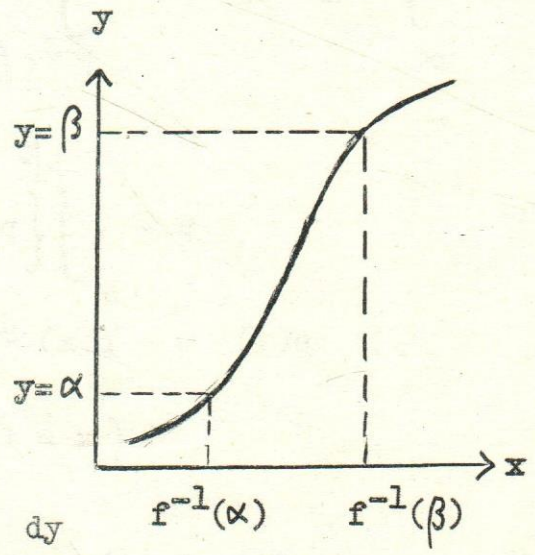
In general, given $p(x_1, x_2, \ldots, x_n)$, we want $p(y_1, y_2, \ldots, y_n)$ we are going to deal with simple types of transformation for all of which $f(x)$ will be continuous function of x. In most cases the transformation will be also, one-to-one, i.e., to any value of x there corresponds one and only one value of, while to any value of y there corresponds one and only one value of x. Since $y = f(x)$ is continuous and differentiable, this implies that $f(x)$ must be either a decreasing or an increasing function of x. This also implies that $x = f^{-1}(y)$

(i) $f(x)$ is an increasing function of x

$$\Pr\{\alpha < y < \beta\} = \Pr\{f^{-1}(\alpha) < x < f^{-1}(\beta)\}$$

i.e. $$\int_{\alpha}^{\beta} p(y)dy = \int_{f^{-1}(\alpha)}^{f^{-1}(\beta)} p(x)\, d_x$$

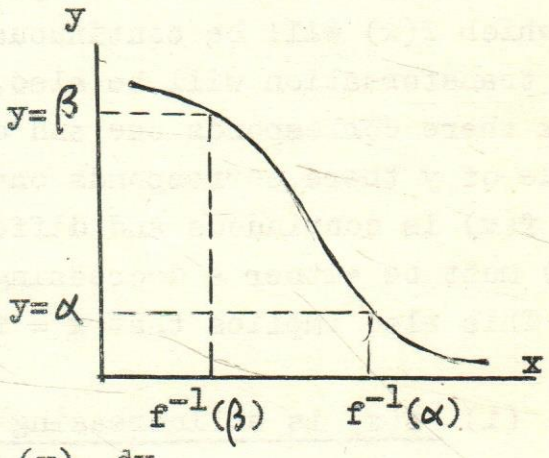
$$= \int_{\alpha}^{\beta} \left\{ p(x) \quad \frac{dx}{dy} \right\} dy$$



$$\therefore p(y) = p(x) \frac{dx}{dy} \qquad \left(\frac{dx}{dy} \geq 0\right) \qquad (1)$$

$$\left(x = f^{-1}(y)\right)$$

(ii) f(x) is a decreasing function of x



$$\Pr\{\alpha < y < \beta\}$$

$$= \Pr\left\{f^{-1}(\beta) < x < f^{-1}(\alpha)\right\}$$

i.e. $$\int_{\alpha}^{\beta} p(y)\, dy = \int_{f^{-1}(\beta)}^{f^{-1}(\alpha)} p(x)\, dx$$

$$= -\int_{\alpha}^{\beta} \left\{p(x) \frac{dx}{dy}\right\} dy$$

$$\therefore \quad p(y) = -p(x) \frac{dx}{dy} \qquad \left(\frac{dx}{dy} \leq 0\right) \qquad (2)$$

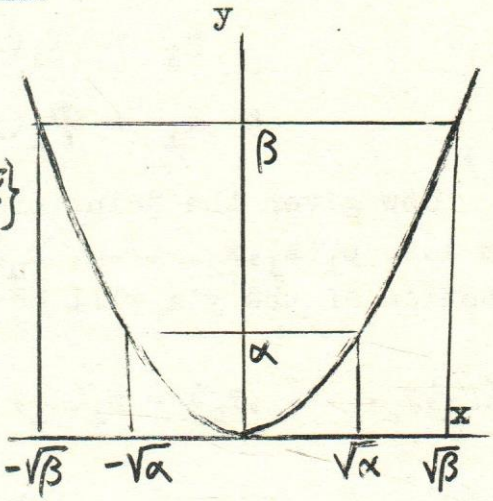
$$\left(x = f^{-1}(y)\right)$$

from (1) & (2) we have

$$p(y) = p(x) \left|\frac{dx}{dy}\right|$$

$$x = f^{-1}(y)$$

The transformation $y = x^2$



$$\Pr\{y \text{ in } (\alpha, \beta)\} = \Pr\{\sqrt{\alpha} < x < \sqrt{\beta}\} + \Pr\{-\sqrt{\beta} < x < -\sqrt{\alpha}\}$$

$$\therefore \int_{\alpha}^{\beta} p(y)\,dy = \int_{\sqrt{\alpha}}^{\sqrt{\beta}} p(x)\,dx + \int_{-\sqrt{\beta}}^{-\sqrt{\alpha}} p(x)\,dx$$

$$= \int_{\sqrt{\alpha}}^{\sqrt{\beta}} p(x)\,dx - \int_{-\sqrt{\beta}}^{-\sqrt{\alpha}} p(x)\,dx$$

$$= \int_{\alpha}^{\beta} p(x)\,\frac{dx}{dy}\,dy + \int_{\alpha}^{\beta} p(x)\,\frac{dx}{dy}\,dy$$

$$x = \sqrt{y} \qquad\qquad x = -\sqrt{y}$$

$$= \int_{\alpha}^{\beta} \frac{1}{2\sqrt{y}} \left\{ p(x)_{x=\sqrt{y}} + p(x)_{x=-\sqrt{y}} \right\} dy$$

$$\therefore \quad p(y) = \frac{1}{2\sqrt{y}} \left\{ p(x)_{x=\sqrt{y}} + p(x)_{x=-\sqrt{y}} \right\} \qquad (3)$$

One to one transformations of sets of n random variables

A set of n variables $x_1, x_2, \ldots, x_n$ and another set of the same number of variables $y_1, y_2, \ldots, y_n$. A one-to-one transformation between these two sets means that each set of values of the x's corresponds to a unique set of values of the y's and conversely. Such a transformation may be written

$$y_i = f_i(x_1, x_2, \dots, x_n)$$

$$\& \quad x_i = \phi_i(y_1, y_2, \dots, y_n)$$

Now given the joint probability density function of the x's i.e. $p_1(x_1, x_2, \dots, x_n)$, the joint probability density function of the y's will be

$$p_2(y_1, y_2, \dots, y_n) = p_1(\ _1, \ _2, \dots, \ _n) \left| \frac{\partial(x_1, x_2, \dots, x_n)}{\partial(y_1, y_2, \dots, y_n)} \right|$$

where $p_1(\ _1, \ _2, \dots, \ _n)$ is the joint probability density function of the x's expressed in terms of the y's.

Example If x_1, x_2 are two independent normally distributed variables with expected value zero and standard deviation , find the distribution of $\frac{1}{2}(x_1 + x_2)$

$$p(x_1, x_2) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}(x_1^2 + x_2^2)}$$

Let $T = \frac{1}{2}(x_1 + x_2)$

$U = \frac{1}{2}(x_1 - x_2)$

$x_1 = T + u$

$x_2 = T - U$

$$\frac{\partial(x_1, x_2)}{\partial(T, U)} = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix} = -2$$

$$\therefore\ p(T,U) = p(x_1,x_2)\left|\frac{\partial(x_1,x_2)}{\partial(T,U)}\right|$$

$$= \frac{1}{\pi\sigma^2}\ e^{-\frac{1}{\sigma^2}(T^2+U^2)}$$

$$\therefore\ p(T) = \int_{-\infty}^{\infty} p(T,U)\,du = \frac{1}{\pi\sigma^2}\int_{-\infty}^{\infty} e^{-(T^2+U^2)/\sigma^2}\,dU$$

$$= \frac{1}{\sqrt{\pi}\,\sigma}\ e^{-T^2/\sigma^2}$$

χ^2 - distribution

Let u_j be independent unit normal variates. χ^2 is defined as

$$\chi^2_\nu = u_1^2 + u_2^2 + \ldots + u_\nu^2 = \sum_{j=1}^{\nu} u_j^2 \qquad (1)$$

But $p(u_j) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}u_j^2} \qquad -\infty \leq u_j \leq \infty \qquad (2)$

write $v_j = u_j^2$

$$\therefore\ p(v_j) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}v}\ v^{-\frac{1}{2}} \qquad 0 \leq v \leq \infty \qquad (3)$$

(i) To get $p(\chi_1^2)$ Let $v_1 = \chi_1^2$

$$\therefore\ p(x_1^2) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}x_1^2}\ (x_1^2)^{-\frac{1}{2}} \qquad 0 \leq x_1^2 \leq \infty$$

(ii) To get $p(\chi_2^2)$

Let $\chi_2^2 = v_1 + v_2 = \chi_1^2 + v_2$, $y = \chi_1^2$

$$\therefore \quad p(x_1^2, v_2) = p(x_1^2) \cdot p(v_2)$$

$$= \frac{1}{2\pi} (x_1^2)^{-\frac{1}{2}} v_2^{-\frac{1}{2}} e^{-\frac{1}{2}(x_1^2 + v_2)}$$

$$\therefore \quad p(x_2^2, y) = p(x_1^2, v_2) \left| \frac{\partial(x_1^2, v_2)}{\partial(x_2^2, y)} \right|$$

$$= \frac{1}{2\pi} y^{-\frac{1}{2}} (x_2^2 - y)^{-\frac{1}{2}} e^{-\frac{1}{2}x_2^2} \qquad 0 < y < x_2^2$$

$$p(x_2^2) = \int_0^{x_2^2} p(x_2^2, y)\, dy$$

$$= \frac{1}{2} e^{-\frac{1}{2} x_2^2}$$

(iii) To get $p(x_3^2)$

$$x_3^2 = x_2^2 + v_3 \quad , \quad y = x_2^2$$

$$\therefore \quad p(x_3^2, v_3) = p(x_3^2)\, p(v_3)$$

$$= \frac{1}{2\sqrt{2\pi}} (x_3^2 - y)^{-\frac{1}{2}} e^{-\frac{1}{2} x_3^2}$$

$$p(x_3^2) = \int_0^{x_3^2} p(x_3^2, y)\, dy$$

$$= \frac{1}{\sqrt{2\pi}} (x_3^2)^{\frac{1}{2}} e^{-\frac{1}{2} x_3^2}$$

$\therefore$ In general

$$\therefore \quad p(x^2) = c\, (x^2)^{\frac{1}{2}(\nu-2)} e^{-\frac{1}{2} x_\nu^2}$$

where $c = \dfrac{1}{2^{\nu/2}\,\Gamma(\frac{\nu}{2})}$

By mathematical induction, we proceed to find the p.d.f. of

$$\chi^2_{\nu+1} = \chi^2_{\nu} + v_{\nu+1}$$

write $y = \chi^2$

$$\therefore\ p(\chi^2_{\nu+1},\ y) = \frac{c}{\sqrt{2\pi}}\ e^{-\frac{1}{2}\chi^2_{\nu+1}}\ y^{(\nu-2)/2}\ (\chi^2_{\nu+1}-y)^{-\frac{1}{2}}$$

$$\therefore\ p(\chi^2_{\nu+1}) = \int_0^{\chi^2_{\nu+1}} p(\chi^2_{\nu+1},\ y)\ dy$$

$$p(\chi^2_{\nu+1}) = c'\,(\chi^2_{\nu+1})^{\frac{1}{2}(\nu+1)-1}\ e^{-\frac{1}{2}\chi^2_{\nu+1}}$$

This is the same as $p(\chi^2)$ but with $\nu+1$ instead of ν which completes the proof.

Another proof

Let $v_j = u_j^2$

We know that $p(v_j) = \dfrac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}v}\ v^{-\frac{1}{2}}$

and $\varphi_v(t) = \dfrac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{itv-\frac{1}{2}v}\ v^{-\frac{1}{2}}\ dv$

$$= (1-2it)^{-\frac{1}{2}}$$

But v_j are independent where $j=1,2,\ldots,\nu$

$$\therefore\ \varphi_{\Sigma v_j}(t) = \left\{\varphi_v(t)\right\}^{\nu} = (1-2it)^{-\frac{\nu}{2}}$$

i.e. $\varphi_{\chi^2}(t) = (1 - 2it)^{-\nu/2}$

This is the characteristic function for Γ-function with parameter $\frac{1}{2}\nu$

$$\therefore\ p(\chi^2_\nu) = c\, e^{-\frac{1}{2}\chi^2_\nu}\, (\chi^2_\nu)^{\frac{1}{2}\nu - 1}$$

Note : Given that u_j is a normal variate $N(0,1)$ & $v_j = u_j^2$ then

$$p(v_j) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}v_j}\, v_j^{-\frac{1}{2}}$$

By definition

$$\chi^2_\nu = \sum_{j=1}^{\nu} u_j^2 = \sum_{j=1}^{\nu} v_j$$

where v_j are independent variates

Let $y_j = \frac{1}{2} v_j$

$$\therefore\ p(y_j) = \frac{1}{\Gamma(\frac{1}{2})}\, e^{-y_j}\, y_j^{-\frac{1}{2}}$$

Hence $y_j = \frac{1}{2} v_j$ is distributed as Γ-variate with parameter $\frac{1}{2}$.

Consequently - applying the additive property of Γ-we have

$\chi^2_\nu/2 = \sum_{j=1}^{\nu} y_j$ = a Γ-variate with parameter $\nu/2$

$$\therefore\ p(\chi^2_\nu/2) = \frac{1}{\Gamma(\nu/2)}\, e^{-\frac{1}{2}\chi^2_\nu}\, (\chi^2/2)^{\frac{1}{2}\nu - 1}$$

$$\therefore\ p(\chi^2_\nu) = \frac{1}{2^{\frac{1}{2}\nu}\Gamma(\nu/2)}\, e^{-\frac{1}{2}\chi^2}\, (\chi^2)^{\frac{1}{2}\nu - 1}$$

The Moments of χ^2 :

The cumulant function is $\psi_{\chi^2}(t) = -\frac{1}{2}\nu \log(1-2it)$

$$= \frac{it}{1!}\nu + \frac{(it)^2}{2!}(2\nu) + \frac{(it)^3}{3!}(8\nu) + \frac{(it)^4}{4!}(48\nu) + \ldots\ldots$$

$$\therefore\ K_r = 2^{r-1}(r-1)!\,\nu$$

Hence we have the following relationship $2\beta_2 - 3\beta_1 - 6 = 0$

Remark : It can be shown that

$$p(\chi_\nu) = \frac{1}{2^{\frac{1}{2}\nu - 1}\Gamma(\frac{1}{2}\nu)}\, \chi_\nu^{\nu-1}\, e^{-\frac{1}{2}\chi^2_\nu}$$

Also $\mu'_{2k}(\chi_\nu) = \mu'_k(\chi^2_\nu)$

(V) Some Exact Distributions

A Discontinuous Distributions

1- Poisson Distribution

It is in a sense a particular limiting form of the binomial distribution where $p \to 0$, $n \to \infty$ and np remains constant $= \lambda$ say.

$$P_{n.x} = \frac{n!}{x!\,(n-x)!}\; p^x\; q^{n-x}$$

$$= \frac{n(n-1)\ldots(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= 1\left(1-\frac{1}{n}\right)\ldots\left(1-\frac{x-1}{n}\right)\left(\frac{\lambda}{n}\right)^x \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^x}$$

when $n \to \infty$

$$P_{n.x} \to \frac{\lambda^x}{x!}\; e^{-\lambda}$$

The Poisson distribution is applicable to problems dealing with occurrence of events in a time interval of a given length such as:

- emission of rays from radio active substances
- certain traffic problems
- demands for telephone service
- bacteria count in cells
- distribution of bomb fragments in space.

The characteristic function

$$\varphi_x(t) = \sum_{x=0}^{\infty} \frac{\lambda}{x!}\; e^{-\lambda}\; e^{itx}$$

$$= e^{-\lambda(1-e^{it})}$$

The cumulant function is

$$\psi_x(t) = \log \varphi_x(t) = -\lambda(1-e^{it})$$

all cumulants are constant and each $= \lambda$

$$\therefore \quad \beta_1 = 1/\lambda$$
$$\beta_2 = 3 + \frac{1}{\lambda}$$

The relationship to Γ-function

$$\Gamma_x(n) = \int_0^x v^{n-1} e^{-v}\, dv = \int_0^\infty v^{n-1} e^{-v}\, dv - \int_x^\infty v^{n-1} e^{-v}\, dv$$

$$= \Gamma(n) + \int_x^\infty v^{n-1}\, d(e^{-x})$$

$$= \Gamma(n) - x^{n-1} e^{-x} - (n-1)\, x^{n-2} e^{-x} - (n-1)(n-2)\int_x^\infty e^{-v} v^{n-3} dv$$

$$\Gamma_x(n)/\Gamma(n) = 1 - e^{-x}\left[1 + x + x^2/2! + \ldots + x^{n-2}/(n-2)! + x^{n-1}/(n-1)!\right]$$

$$I_x(n) = 1 - \sum_{t=0}^{n} e^{-x} x^t/t! = 1 - Pr\left\{ r \le n \right\}$$

$$\therefore Pr\left\{ r \le n \right\} = 1 - I_x(n)$$

2-The Negative Binomial

It is closely related to the Bernoulli binomial distribution If we expand, according to the binomial theorem, $(q-p)^{-k}$ where $q = 1+p$, $k > 0$, $p > 0$, we get the general term in

$$q^{-k}(1 - p/q)^{-x} = q^{-k}\left[1 + k(p/q) + \frac{k(k+1)}{2!}\left(\frac{p}{q}\right)^2 + \ldots\right]$$

i.e. $$q^{-k}\, \frac{(k+x)}{x!\ (k)} \left(\frac{p}{q}\right)^x$$

Then $$p(x) = \frac{(k+x)}{x!\ (k)} \left(\frac{p}{q}\right)^x \qquad x=0 \text{ to } x=\infty$$

is called the negative binomial distribution.

(2.1) The moment generating function

$$\varphi_x(t) = \sum_{x=0}^{\infty} q^{-k} e^{tx} \frac{(k+x)}{x!\ (k)} \left(\frac{p}{q}\right)^x = (q - pe^t)^{-k}$$

The mean is

$$\mathcal{E}(x) = \left.\frac{\partial \varphi}{\partial t}\right|_{t=0} = kp$$

& $$Var(x) = kpq$$

3- The Sequential Binomial

In this case we are going to fix the number of successes and vary the number of trials .We want k successes and we shall stop when they are achieved. The trials are independent and p is constant from trial to trial. The result will be the probability if k-1 achieved successes in the preceeding n-1 trials multiplied by the probability of the n^{th} trial ,i.e.

$$P_{n.k} = \binom{n-1}{k-1} p^{k-1} q^{n-k} p = \binom{n-1}{k-1} p^{k} q^{n-k}$$

(3.1) To prove that $\sum_{n=k}^{\infty} P_{n.k} = 1$

$$\sum_{n=k}^{\infty} \binom{n-1}{k-1} p^k q^{n-k} = p^k + kp^k q + \frac{(k+1)k}{2!} p^k q^2 + \cdots\cdots$$

$$= p^k (1 - q)^{-k} = 1$$

(3.2) Moments

$$\mu_1' = \sum_{n=k}^{\infty} n \binom{n-1}{k-1} p^k q^{n-k} = k/p$$

$$\mu_2' = \sum_{n=k}^{\infty} n^2 \binom{n-1}{k-1} p^k q^{n-k} = k(k+1)/p^2 - k/p$$

$$\mu_2 = kq/p^2$$

Exercise Find μ_3 & μ_4

4- Hypergeometric Distribution

Finite population of balls of size N; where there is Np black balls and Nq white balls. A sample of size n is drawn without replacement.What is the probability of k black balls in the sample of n?

$$\Pr\{k=o\} = \frac{Nq}{N} \frac{Nq-1}{N-1} \frac{Nq-2}{N-2} \cdots\cdots\cdots \frac{Nq-n+1}{N-n+1}$$

$$\Pr\{k=1\} = \frac{NP}{N} \frac{Nq}{N-1} \frac{Nq-1}{N-2} \cdots\cdots\cdots\cdots \frac{Nq-n+2}{N-n+2} \binom{n}{1}$$

$$\Pr\{k=2\} = \frac{Np}{N}\,\frac{Np-1}{N-1}\,\frac{Nq}{N-2} \cdots\cdots\cdots \frac{Nq-n+3}{N-n+3}\,\binom{n}{2}$$

.....

$$\Pr\{k=r\} = \frac{Np}{N-1}\,\frac{Np-1}{N-1} \cdots\cdots \frac{Np-r+1}{N-r+1}\,\frac{Nq}{N-r+2}\,\frac{Nq-1}{N-r+3} \cdots\cdots \frac{Nq-n+r+1}{N-n+1}\,\binom{n}{r}$$

$$= \frac{Np!}{(Np-r)!}\,\frac{Nq!}{(Nq-n+r)!}\,\frac{(N-n)!}{N!}\,\frac{n!}{r!(n-r)!}$$

$$= \frac{Np!}{r!(Np-r)!}\,\frac{Nq!}{(n-r)!(Nq-n+r)!}\,\frac{(N-n)!\;n!}{N!}$$

$$= \binom{Np}{r}\binom{Nq}{n-r} \Big/ \binom{N}{n}$$

(4.1) To prove that the sum of the probabilities =1

$$\sum_{r=o}^{n} \frac{\binom{Np}{r}\binom{Nq}{n-r}}{\binom{N}{n}} = \frac{1}{\binom{N}{n}} \sum_{r=o}^{n} \binom{Np}{r}\binom{Nq}{n-r}$$

Now since $(1+x)^{Np}(1+x)^{Nq} = (1+x)^{N}$

Equating the coefficients of x^r on both sides we get

$$\binom{Nq}{r} + \binom{Np}{1}\binom{Nq}{r-1} + \cdots\cdots\cdots + \binom{Np}{r} = \binom{N}{r}$$

For r=n, we have

$$\binom{Nq}{n} + \binom{Np}{1}\binom{Nq}{n-1} + \cdots\cdots\cdots + \binom{Np}{n} = \binom{N}{n}$$

i.e. $$\sum_{r=o}^{n} \binom{Np}{r}\binom{Nq}{n-r} = \binom{N}{n}$$

Hence

$$\sum_{r=o}^{n} \frac{\binom{Np}{r}\binom{Nq}{n-r}}{\binom{N}{n}} = \frac{1}{\binom{N}{n}} \sum_{r=o}^{n} \binom{Np}{r}\binom{Nq}{n-r} = 1$$

(4.2) The moments

$$\mu_1' = \sum_{k=0}^{n} k \binom{Np}{k}\binom{Nq}{n-k} / \binom{N}{n}$$

$$= \sum_{k=0}^{n} \frac{Np!}{(k-1)!(N-p-k)!} \cdot \frac{Nq!}{(n-k)!(Nq-n+k)!} \cdot \frac{n!(N-n)!}{N!}$$

$$= np \sum_{k=0}^{n} \binom{Np-1}{k-1}\binom{Nq}{n-k} / \binom{N-1}{n-1}$$

$$= np$$

$$2 = \sum_{k=0}^{n} k^2 \binom{Np}{k}\binom{Nq}{n-k} / \binom{N}{n} = \sum_{k=0}^{n} \left\{ k(k+1)+k \right\} \binom{Np}{k}\binom{Nq}{n-k} / \binom{N}{n}$$

$$= np\left[1 + \frac{(n-1)(Np-1)}{N-1}\right]$$

$$\therefore \mu_2 = npq\left[1 - \frac{n-1}{N-1}\right]$$

As $N \longrightarrow \infty$, $\mu_2 \longrightarrow npq$

(4.3) Remark The problem can be represented in the following table

	A	$\bar{A}$	
Sample	r	n-r	n
Remainder	Np-r	Nq-n+r	N-n
Population	Np	Nq	N

As we mentioned before

$$Pr\{k=r\} = \frac{Np!}{r!\,(Np-r)!} \cdot \frac{Nq!}{(n-r)!(Nq-n+r)!} \cdot \frac{(N-n)!n!}{N!}$$

In a 2×2 corresponding table, we have

a =r	c=n-r	n
b=Np-r	d=Nq-n+r	m=N-n
r=Np	s=Nq	N

Then $Pr\{k=r\}$ becomes $Pr\{k=r\} = \dfrac{r!\ s!\ m!\ n!}{a!\ b!\ c!\ d!\ N!}$

Also

$$\text{the mean} = np = n\frac{Np}{N} = \frac{nr}{N}$$

& the variance $= npq\left(\dfrac{N-n}{N-1}\right)$

$$= np(1-p)\left(\frac{N-n}{N-1}\right)$$

$$= \frac{nNp}{N}\ \frac{N-Np}{N}\ \frac{N-n}{N-1}$$

$$= \frac{nrsm}{N^2(N-1)}$$

(4.4) To test the significance of the difference between two treatments after being randomly assigned to a group of N=n+m individuals

We are observing the abscence or presence of reaction X .The 1st treatment T_1 is applied to n individuals & the 2nd treatment T_2 to m individuals . As a result a/n,b/m show reaction X.The random process has been applied within the group of N individuals & its repetition would simply involve other random reassignments of the two treatments among the N. No assumption is made as to how the N individuals were selected from some large universe .On the null hypothesis there are r=a+b individuals who will react & s=c+d who do not react,whatever the arrangement of the treatments.

The chance that 'a' will react in n & 'b' in m , if the null hypothesis is true , will be

$$Pr\left\{a \mid N,r,n\right\} = \frac{r!n!m!s!}{a!b!c!d!N!}$$

	X	$\bar{X}$	
T_1	a	c	n
T_2	b	d	m
	r	s	N

The mean of 'a' = rn/N and

the variance $= \dfrac{nrsm}{N^2(N-1)}$

Example :

Mean = rn/N = 6

Variance = = 1.5

s.d. = 1.224

	Survive	Die	
T_1	2	8	10
T_2	13	2	15
	15	10	25

Using the approximation of the normal for the chance that 'a' = 2 or less we get u =(2+0.5-6)/1.224 = -2.857

From the table of the normal probability integral we get the probability of obtaining as large as a -ve deviation by chance is 0.00213. It is significantly small & it would be reasonable to suppose that T_2 is more effective than T_1.
We can also use the hypergeometric

$$Pr\{a=2\} = \frac{10!15!15!10!}{2!8!13!\ 2!\ 25!} = 0.001445$$

$$Pr\{a=1\} = \frac{10!15!15!10!}{1!\ 9!\ 14!\ 1!\ 25!} = 0.0000458$$

$$Pr\{a=0\} = \frac{10!15!15!10!}{0!\ 10!\ 15!\ 0!\ 25!} = 0.000003$$

$$Pr\{a \leqslant 2\} = 0.001445 + 0.0000458 + 0.000003 = 0.0015$$

which is significant.

B Continuous Distributions

1- Gamma distribution

A continuous variable x is called a Gamma variate, if distributed with probability density function

$$p(x) = \frac{e^{-x}\, x^{l-1}}{\Gamma(l)} \qquad (1)$$

throughout the range of values of x from $0 \longrightarrow \infty$
The Characteristic function is given as

$$\varphi_x(t) = \int_0^\infty \frac{e^{-x}\, x^{l-1}}{\Gamma(l)}\, e^{itx}\, dx = (1 - it)^{-l} \qquad (2)$$

$$\therefore \psi_x(t) = -l \log(1-it) = l\left[it + \frac{(it)^2}{2} + \frac{(it)^3}{3} + \ldots\right]$$

$$\therefore \mu_1' = l\ ,\ \mu_2 = l\ ,\ \mu_3 = 2l\ ,\ \mu_4 = 3l(l+2)$$

Theorem(1) If x is normally distributed with mean 'a' & s.d. $= \sigma$, then $(x-a)^2/2\sigma^2$ is a gamma variate with parameter ½

$$p(x) = \frac{1}{\sqrt{2\pi}\ \sigma} \exp -\tfrac{1}{2}\left(\frac{x-a}{\sigma}\right)^2$$

Let $y = \frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2$, $\therefore \frac{dy}{dx} = \frac{x-a}{\sigma^2}$

$$\therefore p(y) = \left[p(x)_{x-a=+\sigma\sqrt{2y}} + p(x)_{x-a=-\sigma\sqrt{2y}} \right] \left| \frac{dx}{dy} \right|$$

$$= \frac{1}{\sqrt{\pi}} e^{-y} y^{-\frac{1}{2}}$$

Hence y is a gamma variate with parameter $\frac{1}{2}$.

Theorem (2) The sum of two independent gamma variates with parameters ℓ & m is a gamma variate with parameter $\ell + m$.

Let $p(x) = e^{-x} x^{\ell-1} / \Gamma(\ell)$, $p(y) = e^{-y} y^{m-1} / \Gamma(m)$

$$\therefore p(x,y) = \frac{e^{-(x+y)} x^{\ell-1} y^{m-1}}{\Gamma(\ell)\,\Gamma(m)}$$

Write $w = x+y$, $v = x$; then $x = v$, $y = w-x$ & $|\partial(x,y)/\partial(w,v)| = 1$

$$p(w,v) = \frac{e^{-w} v^{\ell-1} (w-v)^{m-1}}{\Gamma(\ell)\,\Gamma(m)}$$

$$\therefore p(w) = \int_0^w p(w,v)\, dv = e^{-w} w^{\ell+m-1} / \Gamma(\ell+m)$$

Hence w is a gamma variate with parameter $\ell + m$

We can proceed in another way:-

$$\varphi_x(t) = (1-it)^{-\ell} , \quad \varphi_y(t) = (1-it)^{-m}$$

Since x,y are independent then

$$\varphi_{x+y}(t) = (1-it)^{-(\ell+m)}$$

which is the characteristic function of a gamma variate with parameter $\ell + m$

2- Beta Distribution

If x is a continuous variate distributed with probability density function

$$p(x) = \frac{x^{\ell-1} (1-x)^{m-1}}{B(\ell, m)}$$

throughout the range of values of x=o to 1; then x is a Beta variate of the 1st kind with parameters ℓ,m
The characteristic function is given as

$$\varphi_x(t) = \int_0^1 x^{\ell-1}(1-x)^{m-1} e^{itx}/B(\ell,m)\, dx$$

$$= \frac{1}{B(\ell,m)}\left[B(\ell,m) + \frac{it}{1!}B(\ell+1,m) + \frac{(it)^2}{2!}B(\ell+2,m) + \cdots\cdots\right]$$

The r^{th} moment is

$$\mu'_r = \frac{B(\ell+r,m)}{B(\ell,m)}$$

In particular

$$\mu'_1 = \frac{\ell}{\ell+m} \quad \& \quad \mu_2 = \frac{\ell m}{(\ell+m)^2(\ell+m+1)}$$

Theorem If x,y are independent gamma variates with parameters ℓ,m respectively, the quotient x/(x+y) is a Beta variate with parameters ℓ,m of the 1st kind

Since x,y are independent ,then

$$p(x,y) = \frac{e^{-(x+y)}x^{\ell-1}y^{m-1}}{\Gamma(\ell)\Gamma(m)}$$

Write $w = x/(x+y)$, $v=x$;then $x=v$, $y = v(1-w)/w$ and $|\partial(x,y)/\partial(w,v)| = v/w^2$

$$\therefore p(w,v) = \frac{e^{-v/w}v^{\ell+m-1}(1-w)^{m-1}w^{-m-1}}{\Gamma(\ell)\Gamma(m)}$$

$$\therefore p(w) = \int_0^\infty p(w,v)\, dv = \frac{(1-w)^{m-1}w^{\ell-1}}{B(\ell,m)} = B_1(\ell, m)$$

Definition : A Beta variate of the 2nd type with +ve parameters ℓ,m is defined as a continuous variate x which is distributed with probability density function

$$p(x) = \frac{x^{\ell-1}(1+x)^{-(\ell+m)}}{B(\ell,m)}$$

throughout the range of values of x=o to x= ∞ and denoted by $B_2(\ell,m)$.

$$\mathcal{E}(x) = \int_0^1 x\,p(x)\,dx = \frac{1}{B(\ell,m)} \int_0^1 (1+x)^{-(\ell+m)}\,dx$$

Write $y = x/(x+1)$

$$\therefore \quad (x) = \frac{1}{B(\ell,m)} \int_0^1 y^{\ell}(1-y)^{m-2}\,dy = \frac{}{m-1}$$

Also $\mu_2' = \frac{\ell(\ell+1)}{(m-1)(m-2)}$

$$\therefore \text{Var}(x) = \frac{\ell(\ell+m-1)}{(m-1)^2(m-2)}$$

Theorem : If x is a Beta variate of 2nd kind , its reciprocal is a variate of the same kind with parameters interchanged.

$$p(x) = \frac{x^{\ell-1}}{B(\ell,m)(1+x)^{\ell+m}} = B_2(\ell,m)$$

Write $y = 1/x$

$$\therefore \quad p(y) = p(x)\left|(\partial x/\partial y)\right| = \frac{y^{m-1}}{B(\ell,m)(1+y)^{\ell+m}} = B_2(m,\ell)$$

EXERCISES

1-If x_i $(i=1,2,\ldots,n)$ are n independent variates normally distributed about a common mean o, with s.d. σ_i and $\chi^2 = \sum x_i^2/\sigma_i^2$ then ½ χ^2 is a gamma variate with parameter n/2

2--Prove that a product of a $B_1(\ell,m)$ variate and an independent $\gamma(\ell+m)$ variate is a $\gamma(\ell)$ variate.

3- Prove that the quotient of two independent gamma variates with parameters ℓ, m is a $B_2(\ell,m)$ variate.

(VI) Bivariate Normal & Binomial Distributions

1- Bivariate Normal Distribution

(1-1) To build a mathematical equation of the surface representing the joint distribution of x,y. Take the origin at (x,y) and assume that:-

(i) The marginal distribution of x is normal with s.d.= σ_x

(ii) The regression of y on x is linear

(iii) For given x, y varies normally

(IV) Common array with s.d. = $\sigma_y \sqrt{1-\rho^2}$

From (i); if n_x is the frequency in the small interval δx

$$\therefore n_x = \frac{N}{\sqrt{2\pi}\,\sigma} \; e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2}} \qquad (1)$$

From (ii); $$Y_x = \rho \frac{\sigma_y}{\sigma_x} \; x \qquad (2)$$

From (iii),(iv)

$$n_{xy} = \frac{n_x}{\sqrt{2\pi}\,\sigma_y \sqrt{1-\rho^2}} \; e^{-\frac{1}{2(1-\rho^2)}\left(\frac{y - Y_x}{\sigma_y}\right)^2} \delta y$$

$$= \frac{N}{2\pi \sigma_x \sigma_y \sqrt{1-\rho^2}} \; e^{-\frac{1}{2}\left[\frac{x^2}{\sigma_x^2} + \frac{y - Y_x}{\sigma_y^2(1-\rho^2)}\right]} \delta x \, \delta y \qquad (3)$$

Substituting from (2) in (3) we get

$$n_{xy} = \frac{N}{2\pi \sigma_x \sigma_y \sqrt{1-\rho^2}} \; e^{-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_x^2} - 2\rho\frac{xy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right]} \delta x \, \delta y$$

If $\delta x, \delta y$ each $\to 0$, we can say

$$n_{xy} = Z\ \delta x\ \delta y$$

and if the volume under the surface is 1, then

$$Z = f(x,y) = \frac{1}{2\pi\sigma_x\ \sigma_y\sqrt{1-\rho^2}}\ e^{-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_x^2} - 2\rho\frac{xy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right]} \quad (4)$$

(1-2) Contours of the surface

$$Z = f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\ e^{-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_1^2} - 2\rho\frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)}$$

Z is constant when

$$\frac{x^2}{\sigma_1^2} - 2\rho\frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} = k^2$$

which is the equation of an ellipse. If we change k, we get a family of coaxal similar ellipses. The regression line y on x is the diameter of the ellipse conjugate to y- axis is also the regression line of x on y is conjugate to x- axis.

(i) To get the slope of major axis:-

The equation of the ellipse is

$$\frac{x^2}{\sigma_1^2} - 2\rho\frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} = k^2$$

This can be written in the form

$$Ax^2 - 2Hxy + By^2 = k^2$$

If the angle of rotation of axis is θ, then θ is given by

$$\tan 2\theta = \frac{-2H}{A-B} = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}$$

(ii) The area of the ellipse:-
Again the equation of the ellipse is

$$\frac{x^2}{\sigma_1^2} - 2\rho\frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} = k^2 \qquad (1)$$

Consider the concentric cricle

$$x^2 + y^2 = r^2 \qquad (2)$$

$$\therefore\ x^2\left(\frac{1}{\sigma_1^2} - \frac{k^2}{r^2}\right) - 2\rho\frac{xy}{\sigma_1\sigma_2} + y^2\left(\frac{1}{\sigma_2^2} - \frac{k^2}{r^2}\right) = 0 \qquad (3)$$

This equation has equal roots if

$$\rho^2\frac{1}{\sigma_1^2\sigma_2^2} = \left(\frac{1}{\sigma_1^2} - \frac{k^2}{r^2}\right)\left(\frac{1}{\sigma_2^2} - \frac{k^2}{r^2}\right)$$

i.e. $$\frac{1-\rho^2}{\sigma_1^2\sigma_2^2}r^4 - \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)k^2r^2 + k^4 = 0$$

$\therefore$ The area is $\pi r_1 r_2$ where

$$r_1^2 r_2^2 = \frac{k^4\sigma_1^2\sigma_2^2}{1-\rho^2}$$

$\therefore$ The area is $$\frac{\pi k^2\sigma_1\sigma_2}{\sqrt{1-\rho^2}}$$

Also, for the ellipse

$$\frac{x^2}{\sigma_1^2} - 2\rho\frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} = (k+\delta k)^2 ,$$

The area is

$$\frac{\pi(k+\delta k)^2\sigma_1\sigma_2}{\sqrt{1-\rho^2}}$$

Hence, the area between the two ellipses is

$$dx\,dy = \frac{2k(\delta k)\sigma_1\sigma_2}{\sqrt{1-\rho^2}}$$

$\therefore$ The probability of a point to fall between the two ellipses is

$$\frac{dx\ dy}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\ e^{-\frac{1}{2(1-\rho^2)}k^2}$$

$$= \frac{k\,(\delta k)}{1-\rho^2}\ e^{-\frac{1}{2(1-\rho^2)}k^2}$$

$\therefore$ The probability that the point will fall inside the ellipse k is

$$\frac{1}{1-\rho^2}\int_{o}^{k} k\ e^{-\frac{1}{2(1-\rho^2)}}\ dk$$

$$= 1 - e^{-\frac{k^2}{2(1-\rho^2)}}$$

e.g. if the probability is 0.95, then

$$= 1 - e^{-\frac{1}{2(1-\rho^2)}k^2} = 0.95$$

i.e. $e^{-\frac{1}{2(1-\rho^2)}k^2} = 0.05$ from which we get k.

(1-3) The Marginal & Conditional distributions

If the mean of the x's is "a" and that of the y's is b and if we write

$$x = \frac{x-a}{\sigma_1}\ ,\qquad y = \frac{y-b}{\sigma_2}$$

then the equation for the bivariate normal can be written in the form

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\ e^{-\frac{1}{2(1-\rho^2)}\left[x^2-2\rho xy+y^2\right]}$$

Integrating out y we get

$$f_1(x) = e \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_1^2} - \rho\frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right]} dy$$

$$= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{x^2}{2\sigma_1^2}} = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-a)^2}{2\sigma_1^2}}$$

This is the marginal distribution of x.

Similarly we can get the marginal distribution of y

$$f_2(y) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(y-b)^2}{2\sigma_2^2}}$$

Now the conditional probability function for y given x is

$$f(y|x) = f(x,y) / f_1(x)$$

$$\therefore f(y|x) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2\sigma_2^2(1-\rho^2)}\left[y-b-\rho\frac{\sigma_2}{\sigma_1}(x-a)\right]^2}$$

This means that for fixed x, y is distributed normally with

(i) $\mathcal{E}(y|x) = b + \rho\frac{\sigma_2}{\sigma_1}(x-a)$

So the regression function of y on x is linear

(ii) $\text{Var}(y|x) = \sigma_2^2(1-\rho^2)$

Then the nearer ρ^2 is to 1, the smaller is the variance. If $\rho=0$, y does not depend on x and the two variates are independent. Similarly

$$f(x|y) = \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}} e^{-\frac{1}{2\sigma_1^2(1-\rho^2)}\left[x-a-\rho\frac{\sigma_1}{\sigma_2}(y-b)\right]^2}$$

and

(i) $\mathcal{E}(x|y) = a + \rho \frac{\sigma_1}{\sigma_2}(y-b)$

(ii) $\mathrm{Var}(x|y) = \sigma_1^2 (1-\rho^2)$

The Characteristic function

$$\varphi_{x,y}(t_1,t) = \int_{-\infty}^{\infty}\int e^{it_1x_1+it_2x_2} f(x,y)\,dx\,dy$$

$$= e^{\frac{(it_1)^2}{2!}\sigma_1^2 + \frac{it_1}{1!}\frac{it_2}{2!}\sigma_1\sigma_2 + \frac{(it_2)^2}{2!}\sigma_2^2}$$

$$\therefore \Psi_{x,y}(t_1,t_2) = \frac{(it_1)^2}{2!}\sigma_1^2 + \frac{it_1}{1!}\frac{it_2}{1!}\rho\sigma_1\sigma_2 + \frac{(it_2)^2}{2!}\sigma_2^2$$

$$\therefore K_{20} = \sigma_1^2 \,,\ K_{11} = \sigma_1\sigma_2\rho \,,\ K_{02} = \sigma_2^2$$

and all other cummulants 2 are = 0. In other words

$$\mu_{10} = 0 \,,\ \mu_{20} = \sigma_1^2 \,,\ \mu_{40} = 3\sigma_1^4$$

$$\mu_{01} = 0 \,,\ \mu_{02} = \sigma_2^2 \,,\ \mu_{04} = 3\sigma_2^4$$

$$\mu_{11} = \sigma_1\sigma_2\rho \,,\ \mu_{22} = (1+2\rho^2)\sigma_1^2\sigma_2^2$$

$$\mu_{13} = 3\rho\sigma_1\sigma_2^3$$

$$\mu_{31} = 3\rho\sigma_1^3\sigma_2$$

Note: The bivariate normal distribution function contains five parameters. Four of them $(a,\sigma_1),(b,\sigma_2)$ characterize the two marginal distributions, while the fifth, ρ, characterizes the relationship between the two variables. It can be shown that the parameter ρ is equal to the mean value of the product of the standardized variables. As an estimate, r, of the parameter ρ we compute $r = s_{12}/s_1 s_2$ where s_{12} denotes the estimate of the covariance. The parameter ρ is the correlation coefficient and r is its best estimate.

2- Bivariate Binomial Distribution

(2.1) In generalization to the binomial, consider the drawing of a sample of n from a population where the individuals may or may not have two attributes A & $\bar{A}$, B & $\bar{B}$. Suppose the proportions of the individuals with attributes AB, $\bar{A}B$, $A\bar{B}$, $\bar{A}\bar{B}$ are p_1, p_2, p_3, p_4 respectively where $p_1+p_2+p_3+p_4=1$

	A	$\bar{A}$	
B	p_1	p_3	p_1+p_3
$\bar{B}$	p_2	p_4	p_2+p_4
	p_1+p_2	p_3+p_4	1

	A	$\bar{A}$	
B	h	k	x_2
$\bar{B}$	j	ℓ	$n-x_2$
	x_1	$n-x_1$	n

In exactly the same way as for the binomial, it is seen that the proportion of samples with h of the attribute (A B), j of the attribute ($A\bar{B}$), k of the attribute ($\bar{A}B$) & ℓ of the attribute ($\bar{A}\bar{B}$), is

$$\frac{n!}{h!\,j!\,k!\,l!}\; p_1^h\, p_2^j\, p_3^k\, p_4^{\ell}$$

and the distribution of samples is given by the expansion of the multinomial form

$$(p_1+p_2+p_3+p_4)^n$$

The distribution given by this form is bivariate, one variate being the number of the A's i.e. (x_1) and the other being the number of the B's i.e. (x_2)

(2.2) The characteristic function

$$\varphi_{x_1,x_2} = \mathcal{E}(e^{it_1x_1+it_2x_2})$$

$$= \sum e^{it_1x_1+it_2x_2}\; \frac{n!}{h!j!k!l!}\; p_1^h\, p_2^j\, p_3^k\, p_4^{\ell}$$

$$= \sum e^{it_1(h+j)+it_2(h+k)} \frac{n!}{h!j!k!l!} p_1^h p_2^j p_3^k p_4^l$$

$$= \sum \frac{n!}{h!j!k!l!} \left\{p_1 e^{i(t_1+t_2)}\right\}^h \left\{p_2 e^{it_1}\right\}^j \left\{p_3 e^{it_2}\right\}^k p_4^l$$

$$= \left\{ p_1 e^{it_1+it_2} + p_2 e^{it_1} + p_3 e^{it_2} + p_4 \right\}^n$$

$$\therefore \psi_{x_1,x_2}(t_1,t_2) = n \log \left\{ p_1 e^{it_1+it_2} + p_2 e^{it_1} + p_3 e^{it_2} + p_4 \right\}$$

$$= n \log \left[(p_1+p_2+p_3+p_4) + \frac{it_1}{1!}(p_1+p_2) + \frac{it_2}{1!}(p_1+p_3) + \frac{(it_1)^2}{2!}(p_1+p_2) + \frac{(it_2)^2}{2!}(p_1+p_3) + \frac{it_1}{1!}\frac{it_2}{1!} p_1 + \ldots\ldots \right]$$

$$= n \left[\frac{it_1}{1!}(p_1+p_2) + \frac{it_2}{1!}(p_1+p_3) + \frac{(it_1)^2}{2!}\left\{p_1+p_2-(p_1+p_2)^2\right\} + \frac{(it_2)^2}{2!}\left\{p_1+p_3-(p_1+p_3)^2\right\} + \frac{it_1}{1!}\frac{it_2}{1!}\left\{p_1-(p_1+p_2)(p_1+p_3)\right\} + \ldots\ldots\ldots\ldots \right]$$

$$\therefore K_{10} = n(p_1+p_2) \quad , \quad K_{01} = n(p_1+p_3)$$

$$K_{20} = n(p_1+p_2)\left\{1-(p_1+p_2)\right\}$$

$$K_{02} = n(p_1+p_3)\left\{1-(p_1+p_3)\right\}$$

$$K_{11} = n\left\{p_1-(p_1+p_2)(p_1+p_3)\right\}$$

If we transfer the origin to the mean of the variates we have

$$\psi = \log \varphi$$

$$= -\frac{n}{2}\left[t_1^2(p_1+p_2)\left\{1-(p_1+p_2)\right\} + t_2^2(p_1+p_3)\left\{1-(p_1+p_3)\right\} + 2t_1t_2\left\{p_1-(p_1+p_2)(p_1+p_3)\right\}\right] + O(n)$$

(2-3) The approximation of Biv.Binomial to the bivariate normal

When the distribution is expressed in standard measure we have

$$\varphi_{x_1-k_{10},\,x_2-K_{01}}(t_1,t_2) = e^{it_1(x_1-K_{10})+it_2(x_2-K_{01})}$$

$$= e^{-(it_1K_{10}+it_2K_{01})}\,\varphi_{x_1,x_2}(t_1,t_2)$$

$$\therefore \varphi_{\frac{x_1-K_{10}}{\sigma_{20}},\,\frac{x_2-K_{01}}{\sigma_{02}}}(t_1,t_2) = e^{-\left(\frac{it_1K_{10}}{\sigma_{20}}+\frac{it_2K_{01}}{\sigma_{02}}\right)}\varphi_{x_1,x_2}(t_1,t_2)$$

$$\log \varphi_{\frac{x_1-K_{10}}{\sigma_{20}},\,\frac{x_2-K_{01}}{\sigma_{02}}}(t_1,t_2) \quad .$$

$$= \frac{(it_1)^2}{2!} + \frac{(it_2)^2}{2!} + \frac{it_1}{1!}\,\frac{it_2}{1!}\,\frac{K_{11}}{\sigma_{20}\sigma_{02}} + O\left(\frac{1}{\sqrt{n}}\right)$$

$$\longrightarrow -\frac{1}{2}\left(t_1^2 + t_2^2 + 2\,t_1t_2\right) \text{ as } n \longrightarrow \infty \qquad (1)$$

where

$$\rho = \frac{K_{11}}{\sigma_{20}\sigma_{02}} = \frac{p_1-(p_1+p_2)(p_1+p_3)}{\sqrt{\left[(p_1+p_2)(1-\overline{p_1+p_2})\right]\left[(p_1+p_3)(1-\overline{p_1+p_3})\right]}}$$

$$\therefore \varphi_{\frac{x_1-K_{10}}{\sigma_{20}},\ \frac{x_2-K_{01}}{\sigma_{02}}}(t_1,t_2) \to e^{-\frac{1}{2}(t_1^2+2\rho t_1 t_2+t_2^2)}$$

which is the characteristic of the bivariate normal surface with zero mean, unit standard deviation and coefficient of correlation ρ i.e. of the form

$$p(x_1,x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}(x_1^2-2\rho x_1 x_2+x_2^2)} \qquad (2)$$

In other words the standardized bivariate binomial distribution tends to the bivariate normal distribution (2) as $n \to \infty$.

(VII) Sampling and Sampling Distributions

Sampling and Sampling distributions

5-1 Basic definitions

Population is a totality of objects under Consideration It refers not alone to persons but to physical objects and operations as well.

A real population is one which actually exists in the form of specific population objects.

A hypothetical population is one which the statistician immagines to exist.

A finite population is one in which the number of population objects is countable and definitely limited in number.

An infinite population is one in which the number of objects or operations is unlimited or cannot be counted.

Sample is a part selected or drawn from a statistical population which is used as a basis of making estimates and inferences about the population.

If the selection takes place so that every unit has an equal chance of being selected, the sample is called random sample.

A small sample usually refers to one which has fewer than 30 sampling units. The value of the distinction between small and large samples lies in the fact that, although certain statistical operations can be applied to large random samples, they cannot be applied to small samples.

Statistic (S):- it refers to a value based on sample values and sometimes known as an estimate. It varies from one sample to another.

Parameter (P): it refers to a value calculated from the population. It is constant for this population.

Sampling error (E): it is the difference between the statistic and the corresponding parameter i.e. $E = S-P$.

If the expected value of the statistic (or the estimate) is equal to the corresponding parameter, the statistic is unbiased. Otherwise it is biased.

Sampling distribution: The frequency distribution of a statistic (an estimate) derived from an indefinitely large number of random samples each of size n is called a sampling distribution.

The standard deviation of this distribution is the standard error of the estimate. The less the variance of this distribution, the greater the precision of the estimate. The variance of a sampling distribution can be reduced by increasing the size of the sample or by using a more efficient sample design.

(5-2) Estimation: assume that the parent population is distributed in a form which world be completely determinate if we knew the value of some parameter θ. We are given a sample of values $x_1, x_2, \ldots, x_n$.

We require to determine with the aid of the x's, a number which can be taken to be the value of θ, or a range of numbers which can be taken to include that value. We cannot expect to find any method of estimation which can be generated to give a close estimate of θ on every occasion and for every sample. We must derive a rule, a method or a formula which will give good results "in the long run" or on the average. The method or rule of estimation is called an Estimator and the value to which it gives rise in particular cases, the Estimate. It is itself a random variable and has a distribution.

In general:-

A population with a density function $f(x;\theta_1,\theta_2,\ldots,\theta_k)$ where x is the variate and $\theta_1,\theta_2,\ldots,\theta_k$ are parameters. On the basis of a random sample of observations say $x_1, x_2, \ldots, x_n$, wish to estimate one or more of the parameters $\theta_1,\theta_2,\ldots,\theta_k$. The question is to find functions of the observations which may be represented by

$$\hat{\theta}_1(x_1,x_2,\ldots,x_n),\ \hat{\theta}_2(x_1,x_2,\ldots,x_n),\ldots\ldots\ldots\ldots$$

such that the distribution of these of functions will be concentrated as closely as possible near the true values of the parameters. The properties of good estimators are :-

(1) Consistent "t is a consistent estimator of θ if it converges to θ in probability"

i.e. $P_r\{t \to \theta\} \to 1$ as $n \to \infty$.

(2) unbiased: " t is said to be an unbiased estimator of θ if $\mathcal{E}(t) = \theta$."

(3) efficient: If for two consistent estimators t_1, t_2, we have var $(t_1) <$ var (t_2) for all n, then t_1 is more efficient than t_2 for all sample sizes.

(4) sufficient: If the joint density of the sample is expressible in the form

$$L(x_1,x_2,\ldots,x_n,\theta) = \prod_{i=1}^{n} f(x_i,\theta)$$

$$= L_1(t,\theta).L_2(x_1,x_2,\ldots,x_n)$$

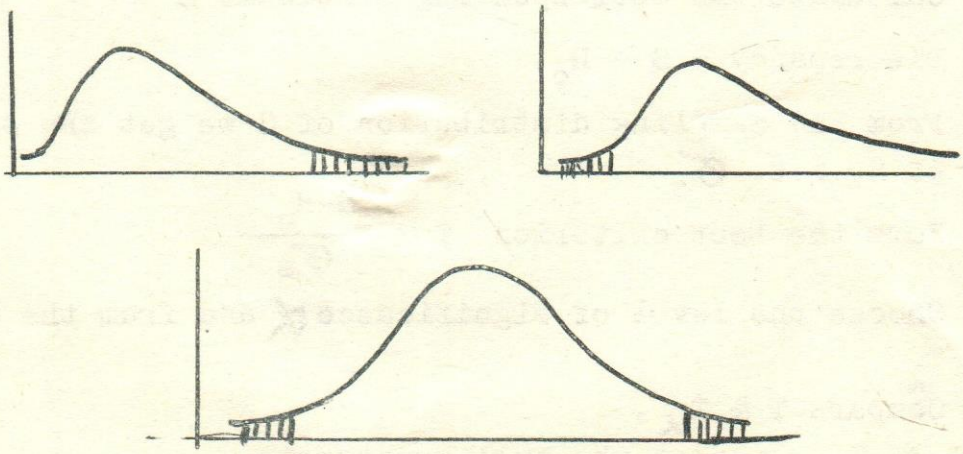
where L_1 does not contain the x's and L_2 is independent of θ, then t is said to be a sufficient estimator of θ.

(5-3) Testing statistical hypotheses

An inference: is any conclusion made on the basis of experience.

Hypothesis: refers to the more systematic and formal statement of an inference usually in the form of a proposition

Statistical hypotheses: are usually classified into null and alternative hypotheses. A test of a such a hypothesis is consisting of testing the null hypothesis against the alternatives. In other words, the testing of a statistical hypothesis consists of formulating and applying an objective criterion for the purpose of accepting or rejecting the hypothesis. This criterion (test-statistic) will have a sampling distribution which indicated the probability of obtaining a value larger than the one obtained. If the obtained value of the test statistic is very likely value, then this is taken to be statistical evidence that the hypothesis is a plausible one. This process consists of selecting a critical region (a portion or portions of the sampling distribution of the statistic being used) and agreeing to reject the hypothesis under test whenever a random sample gives rise to a value of the statistic which falls in this critical region. Hence the sampling distribution of the test statistic must be known when the hypothesis is true



In some problems, the critical region will be only at one end either the upper or the lower and the test in such case is called a one-tailed test. In other problems, the critical region (regions of rejection) will be at both ends and the test is called a two-tailed test. This will depend upon whether we have one alternative hypothesis or two respectively.

The proportion of the probability distribution contained in the critical region is the level of significance of the test. In practice two levels are commonly used 1% or 5%

The size of the critical region is related to the risk one wants to accept in testing a statistical hypothesis. If the region is made too large, we reject too many true hypotheses. Rejecting true hypothesis is called an error of the 1st kind or Type I error. If the region is too small, we accept too many false hypotheses.
Accepting false hypotheses is called an error of the 2nd kind or Type II error

Now for testing a statistical hypotheses $H = H_0$ the procedure is summerized as follows:-

(1) Null hypothesis $H = H_0$
Alternative hypotheses $H \neq H_0$

(2) Calculate the corresponding statistic S

(3) Discrepancy = $S - H_0$

(4) From the sampling distribution of S we get the standard error of S, i.e. σ_s

(5) Form the test criterion $T = \dfrac{S-H_0}{\sigma_s}$

(6) Choose the level of significance α and from the tables find T_α

(7) Compare T & T_α.
If $T > T_\alpha$ reject the null hypothesis
If $T < T_\alpha$ accept the null hypothesis

(5.4) Sampling Distributions

(1) Sampling distribution of the mean $\bar{x}$

A random sample of n independent values $x_1, x_2, \ldots, x_n$ is drawn from a normal population with mean ξ and variance σ^2.

The characteristic function for the normal distribution is

$$\varphi_x(t) = e^{it\xi + \frac{(it)^2}{2}\sigma^2}$$

But

$$\varphi_{\bar{x}}(t) = \left\{\varphi_x\left(\frac{t}{n}\right)\right\}^n = \left\{e^{\frac{it}{n}\xi + \frac{(it)^2}{2}\frac{\sigma^2}{n^2}}\right\}^n$$

$$= e^{it\xi + \frac{(it)^2}{2}\frac{\sigma^2}{n}}$$

This is the characteristic function of a normal distribution with mean ξ and variance $\frac{\sigma^2}{n}$.

$\therefore$ The sampling distribution of the means of indep. random samples of size n drawn from a normal population with mean ξ and variance σ^2 is also a normal distribution with the same mean and variance $\frac{\sigma^2}{n}$.

(2) The sampling distribution of s^2

If x_i are n independent normal variates with mean ξ and variance σ^2 then

$$u_i = \frac{x_i - \xi}{\sigma} \qquad (1)$$

is a unit normal variate.

Consequently $\chi_n^2 = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} \frac{(x_i - \xi)^2}{\sigma^2}$

If we replace ξ by $\bar{x}$ then we get χ^2_{n-1}

i.e. $$\chi^2_{n-1} = \frac{\sum(x_i - \bar{x})^2}{\sigma^2}$$

$$= \frac{ns^2}{\sigma^2} \qquad (2)$$

where

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

$$\therefore \quad s^2 = \chi^2_{n-1}\,\sigma^2/n$$

$$\therefore \quad p(s^2) = p(\chi^2_{n-1}) \left| \frac{d\,\chi^2_{n-1}}{d\,s^2} \right|$$

But

$$p(\chi^2_{n-1}) = \frac{1}{\Gamma(\frac{n-1}{2})\; 2^{\frac{n-1}{2}}}\; e^{-\frac{1}{2}\chi^2_{n-1}}\; (\chi^2_{n-1})^{\frac{n-3}{2}}$$

$$\therefore \quad p(s^2) = \frac{1}{\Gamma(\frac{n-1}{2})}\left(\frac{n}{2\sigma^2}\right)^{\frac{n-1}{2}} e^{-\frac{ns^2}{2\sigma^2}}\; (s^2)^{\frac{n-3}{2}} \qquad (3)$$

<u>To get p(s)</u>

Let $y = s = \sqrt{s^2}$ i.e. $y^2 = s^2$

$$\therefore \quad p(s) = \frac{2}{\Gamma(\frac{n-1}{2})}\left(\frac{n}{2\sigma^2}\right)^{\frac{n-1}{2}} e^{-\frac{ns^2}{2\sigma^2}}\; s^{n-2} \qquad (4)$$

To prove that $\bar{x}$ & s^2 are independent

Normal variates $x_1, x_2, \ldots, x_n$, each has expected value ξ and s.d. σ.

$$\therefore p(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum(x_i-\xi)^2}$$

But

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

$$\bar{x} = \frac{\sum x}{n}, \quad s^2 = \frac{1}{n}\sum(x_i-\bar{x})^2$$

$$\therefore p(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{ns^2}{2\sigma^2}} e^{-\frac{n}{2\sigma^2}(\bar{x}-\xi)^2}$$

To use a transformation

$$x_i = \bar{x} + a_{i2}y_2 + a_{i3}y_3 + \ldots + a_{in}y_n \qquad (i=1,2,\ldots,n)$$

where $\sum a_{ij}=0$, $\sum a_{ij}a_{ik} = 0$, $\sum a_{ik}^2 = 1$

$$j = 2,3,\ldots, n \;\&\; j \neq k$$

$$x_1 = \bar{x} + \frac{1}{\sqrt{1.2}}y_2 + \frac{1}{\sqrt{2.3}}y_3 + \ldots + \frac{1}{\sqrt{(n-1)n}}y_n$$

$$x_2 = \bar{x} - \frac{1}{\sqrt{1.2}}y_2 + \frac{1}{\sqrt{2.3}}y_3 + \ldots + \frac{1}{\sqrt{(n-1)n}}y_n$$

$$x_3 = \bar{x} - \frac{2}{\sqrt{2.3}}y_3 + \ldots + \frac{1}{\sqrt{(n-1)n}}y_n$$

$$\vdots$$

$$x_{n-1} = \bar{x} - \frac{n-2}{\sqrt{(n-2)(n-1)}}y_{n-1} + \frac{1}{\sqrt{(n-1)n}}y_n$$

$$x_n = \bar{x} - \frac{n-1}{\sqrt{(n-1)n}}y_n$$

$\therefore \sum x_i = n\,\bar{x}$

$$\sum (x_i - \bar{x})^2 = \sum_{j=2}^{n} y_j^2 \qquad \text{i.e. } n s^2 = \sum_{j=2}^{n} y_j^2$$

$$\therefore p(\bar{x}, y_2, y_3, \ldots, y_n) = p(x_1, x_2, \ldots, x_n)\,|J|$$

$$= c\, e^{-\frac{1}{2\sigma^2}\sum_{j=2}^{n} y_j^2}\, e^{-\frac{n}{2\sigma^2}(\bar{x}-\xi)^2}$$

$$\therefore p(\bar{x}) = \int\int \cdots \int_{-\infty}^{\infty} p(\bar{x}, y_2, y_3, \ldots, y_n)\, dy_2\, dy_3 \ldots dy_n$$

$$= c_1\, e^{-\frac{n(\bar{x}-\xi)^2}{2\sigma^2}}$$

$$p(\bar{x}) = \frac{\sqrt{n}}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{n(\bar{x}-\xi)^2}{2\sigma^2}}$$

Also

$$p(y_2, y_1, \ldots, y_n) = \int_{-\infty}^{\infty} p(\bar{x}, y_2, y_3, \ldots y_n)\, d\bar{x}$$

$$= C_2\, e^{-\frac{1}{2\sigma^2}\sum_{j=2}^{n} y_j^2}$$

$$p(y_i) = C_3\, e^{-\frac{1}{2\sigma^2} y_j^2}$$

i.e. $y_j \sim N(0, \sigma)$

$\therefore$ $\bar{x}, y_2, y_3, \cdots, y_n$ are mutually independent set of random variables

$\therefore$ $\bar{x}, s^2 = \sum_{j=2}^{n} y_j^2$ are mutually independent

(3) The sampling distribution of "t"

Given a normal population with mean ξ and variance σ^2, the sampling distribution of the mean for independent random samples drawn from such a population is normal with mean ξ & variance $\frac{\sigma^2}{n}$

i.e. $\bar{x} \sim N(\xi, \frac{\sigma^2}{n})$

($\sim$ is distributed as)

$\therefore$ $\frac{\bar{x} - \xi}{\sigma/\sqrt{n}} \sim N(0, 1)$

i.e. the test statistic $u = \frac{\bar{x} - \xi}{\sigma/\sqrt{n}}$ is distributed as unit normal distribution. Usually σ for the population is unknown. An estimate for σ based on sample values can be derived:-

$\therefore$ $\frac{n s^2}{\sigma^2} = \chi^2_{n-1}$

$\therefore$ $\mathcal{E}(\frac{ns^2}{\sigma^2}) = n - 1$

$\therefore$ $\mathcal{E}(\frac{n}{n-1} s^2) = \sigma^2$

$\therefore$ $\hat{\sigma}^2 = \frac{n}{n-1} s^2$

i.e. $\hat{\sigma} = \frac{\sqrt{n}}{\sqrt{n-1}} s$

$\therefore$ The test statistic becomes

$$t = \frac{\bar{x} - \xi}{\sigma/\sqrt{n}} = \frac{\bar{x} - \xi}{s/\sqrt{n-1}} = \frac{\frac{\bar{x} - \xi}{\sigma/\sqrt{n}}}{\chi_{n-1}/\sqrt{n-1}}$$

$$= \frac{u}{\chi_{n-1}/\sqrt{n-1}} = \frac{u}{\chi_{\nu}/\sqrt{\nu}} \qquad (1)$$

Also

Given two normal populations $N(\xi_1, \sigma_1^2)$ & $N(\xi_2, \sigma_2^2)$ and two independent random samples of size n_1, n_2 are drawn. Then the sampling distribution of the difference between the two means i.e. $\bar{x}_1 - \bar{x}_2$ is normally distributed with mean $\xi_1 - \xi_2$ and variance $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

Let the two populations have common variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$ which is usually unknown. An estimate for σ^2 based on the two sample observations can be derived.

$$n_1 s_1^2 = \chi^2_{n_1-1}\, \sigma^2$$

$$n_2 s_2^2 = \chi^2_{n_2-1}\, \sigma^2$$

$$n_1 s_1^2 + n_2 s_2^2 = (\chi^2_{n_1-1} + \chi^2_{n_2-1})\sigma^2$$

$$= \chi^2_{n_1+n_2-2}\, \sigma^2$$

$$\therefore \mathcal{E}(n_1 s_1^2 + n_2 s_2^2) = (n_1+n_2-2)\sigma^2$$

$$\therefore \mathcal{E}\left(\frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2}\right) = \sigma^2$$

$$\therefore \quad \hat{\sigma}^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} = \frac{\chi^2_{n_1+n_2-2}\,\sigma^2}{n_1+n_2-2}$$

But

$$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ is distributed as unit normal.

Since $\sigma^2 = \sigma_1^2 = \sigma_2^2$ and σ^2 is unknown, then the statistic becomes

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\dfrac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\chi_{n_1+n_2-2} \Big/ \sqrt{n_1+n_2-2}}$$

$$= \frac{u}{\chi_\nu / \sqrt{\nu}} \qquad (2)$$

To get $p(t)$

$$p(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

&

$$p(\chi_\nu) = \frac{1}{2^{\frac{1}{2}\nu - 1}\,\Gamma(\frac{\nu}{2})}\, \chi^{\nu-1} e^{-\frac{1}{2}\chi^2}$$

$\therefore$ u & χ_ν are independent

$$\therefore \quad p(u, \chi_\nu) = p(u)\, p(\chi_\nu) = c\, \chi_\nu^{\nu-1} e^{-\frac{1}{2}(u^2+\chi^2)}$$

Make the transformation

$$t = \frac{u}{\chi_\nu / \sqrt{\nu}} \quad , \quad y = \chi_\nu$$

i.e. $\chi_\nu = y \quad , \quad u = ty / \sqrt{\nu}$

$$\therefore \left|\frac{\partial(x,\,u)}{\partial(t,\,y)}\right| = y/\sqrt{\nu}$$

$$\therefore p(t,y) = c_1 y^{\nu} e^{-\frac{1}{2}y^2(1+\frac{t^2}{\nu})}$$

$$\therefore p(t) = c_1 \int_0^{\infty} y^{\nu} e^{-\frac{1}{2}y^2(1+\frac{t^2}{\nu})}\, dy \qquad (y > 0)$$

write $z = \frac{1}{2}y^2(1+\frac{t^2}{\nu})$

$$\therefore p(t) = \frac{c_2}{(1+\frac{t^2}{\nu})^{\frac{1}{2}(\nu+1)}} \Gamma\left(\frac{\nu+1}{2}\right)$$

where

$$c_2 = \frac{1}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})\sqrt{\nu}}$$

$$\therefore p(t) = \frac{1}{\sqrt{\nu}\, B(\frac{1}{2},\frac{\nu}{2})} (1+\frac{t^2}{\nu})^{-\frac{1}{2}(\nu+1)}$$

As $\nu \to \infty$, $p(t_\nu) \longrightarrow N(0,1)$.

(4) The sampling distribution of F

Given two normal populations with variances σ_1^2, σ_2^2 respectively. (Usually σ_1^2, σ_2^2 are unknown)

Two independent random samples of size n_1, n_2 are drawn. If their variances are s_1^2, s_2^2 then

$$\hat{\sigma}_1^2 = \frac{n_1 s_1^2}{n_1 - 1}$$

$$\hat{\sigma}_2^2 = \frac{n_2 s_2^2}{n_2 - 1}$$

$$\therefore \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{n_1 s_1^2/(n_1-1)}{n_2 s_2^2/(n_2-1)}$$

Also $$n_1 s_1^2 = \chi^2_{n_1-1}\,\sigma_1^2$$

$$n_2 s_2^2 = \chi^2_{n_2-1}\,\sigma_2^2$$

and if $\sigma_1^2 = \sigma_2^2$, then

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{\chi^2_{n_1-1} / (n_1-1)}{\chi^2_{n_2-1} / (n_2-1)}$$

$$\therefore F_{\nu_1,\nu_2} = \frac{\chi^2_{\nu_1} / \nu_1}{\chi^2_{\nu_2} / \nu_2}$$

$\because \chi^2_{\nu_1}, \chi^2_{\nu_2}$ are independent

$$\therefore p(\chi^2_1, \chi^2_2) = p(\chi^2_1)\, p(\chi^2_2)$$

$$= C(\chi^2_{\nu_1})^{\frac{1}{2}(\nu_1-2)} (\chi^2_{\nu_2})^{\frac{1}{2}(\nu_2-2)} e^{-\frac{1}{2}(\chi^2_{\nu_1} + \chi^2_{\nu_2})}$$

where

$$C = \frac{1}{2^{\frac{1}{2}(\nu_1+\nu_2)}\,\Gamma(\frac{\nu_1}{2})\,\Gamma(\frac{\nu_2}{2})}$$

Make the transformation

$$F = \frac{\nu_2}{\nu_1}\,\frac{\chi^2_{\nu_1}}{\chi^2_{\nu_2}} \quad , \quad y = \chi^2_{\nu_2}$$

$$\therefore \left| \frac{\partial(\chi^2_1, \chi^2_2)}{\partial(F, y)} \right| = \frac{\gamma_1}{\gamma_2} y$$

$$\therefore p(F,y) = C_1 F^{\frac{\gamma_1 - 2}{2}} y^{\frac{\gamma_1}{2} + \frac{\gamma_2}{2} - 1} e^{-y\left(1 + \frac{\gamma_1}{\gamma_2} F\right)}$$

$$\therefore p(F) = \int_0^\infty p(F,y)\, dy$$

$$\therefore p(F) = \frac{\gamma_1^{\frac{1}{2}\gamma_1}\, \gamma_2^{\frac{1}{2}\gamma_2}}{B\left(\frac{\gamma_1}{2}, \frac{\gamma_2}{2}\right)} \cdot \frac{F^{\frac{1}{2}\gamma_1 - 1}}{(\gamma_2 + \gamma_1 F)^{\frac{1}{2}(\gamma_1 + \gamma_2)}}$$

$$0 < F < \infty$$

The moments

$$\mu'_r = C \int_0^\infty F^{\frac{1}{2}\gamma_1 + r - 1} (\gamma_2 + \gamma_1 F)^{-\frac{1}{2}(\gamma_1 + \gamma_2)}\, dF$$

where

$$C = \frac{\gamma_1^{\frac{1}{2}\gamma_1}\, \gamma_2^{\frac{1}{2}\gamma_2}}{B\left(\frac{\gamma_1}{2}, \frac{\gamma_2}{2}\right)}$$

write $z = \frac{\gamma_1 F}{\gamma_2 + \gamma_1 F}$

$$\therefore 1 - Z = \frac{\gamma_2}{\gamma_2 + \gamma_1 F}$$

$$\therefore \quad F = \frac{\nu_2}{\nu_1} \frac{Z}{1-Z}$$

$$\therefore \mu'_r = \left(\frac{\nu_2}{\nu_1}\right)^r \frac{B\left(\frac{\nu_1}{2} + r, \frac{\nu_2}{2} - r\right)}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)}$$

Exercises

(1) Prove that

$$\Pr\left\{t_\nu > t_0\right\} = \frac{1}{2}\left\{1 - I_k\left(\frac{1}{2}, \frac{1}{2}\nu\right)\right\}$$

where

$$k = \frac{t_0^2}{1 + t_0^2/\nu}$$

(2) Find β_1 & β_2 for F-distribution

(3) Show that

$$p(t_\nu^2) = p(F_{1,\nu})$$

(VIII) Testing statistical hypothesis and Confidence interval

Testing Statistical Hypothesis and Confidence Limits.

1) Testing statistical hypothesis around the mean

(i) Large samples

Theorem:- The sampling distribution of the means of independent random samples drawn from a normal population $N(\mu, \sigma^2)$ is also normal with expected value μ and variance σ^2/n.

Cor.(1) Usually the population is unknown and hence the above theorem can be stated again as:

The sampling distribution of the means of independent large samples drawn at random from an infinite population (μ, σ^2) approaches a normal distribution with expected value μ & variance σ^2/n as n increases.

Cor.(2) If σ is unknown, then the sampling distribution of the means of independent large samples drawn at random from infinite population (μ, σ^2) approaches a normal distribution with expected value μ & variance $s^2/(n-1)$ as n increases.

(Remember $\hat{\sigma}^2 = \frac{n}{n-1} s^2$).

Hence to test the null hypothesis $\mu = \mu_0$ against $\mu \neq \mu_0$, we use the test criterion

$$u = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$$

where $\sigma_{\bar{x}}$ is the standard error of the mean and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad \text{if } \sigma \text{ is known}$$

or $$= \frac{s}{\sqrt{n-1}} \quad \text{if } \sigma \text{ " unknown}$$

The above statistic is distributed as unit normal N(o,1) and hence we compare u with u_α looked in the tables at α% level of significance

(ii) Small samples

To test the null hypothesis $\mu = \mu_0$ against $\mu \neq \mu_0$ we use the test criterion

$$t = \frac{\bar{x} - \mu_0}{\tilde{\sigma}_{\bar{x}}}$$

where $\tilde{\sigma}_{\bar{x}} = s/\sqrt{n-1}$. This statistic is distributed as t-distribution with $n-1 = \gamma$ degrees of freedom. Hence we compare t with $t_{\gamma,\alpha}$ at $\alpha\%$ level of significance

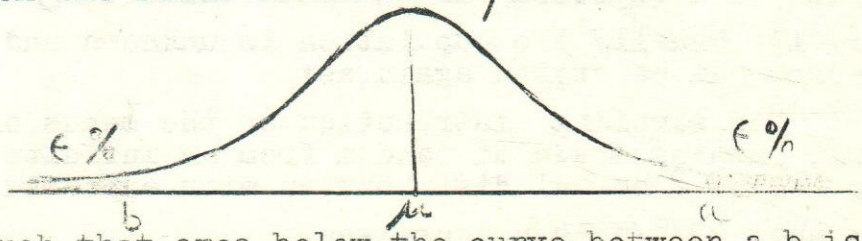
2) Confidence limits for the population mean

It is required to find two values between which we expect that the population mean will lie, with confidence coefficient $\gamma = (100 - 2\epsilon)\%$

Consider the sampling distribution of the mean and that a, b are two points on the sides of the mean μ where

$a = \mu + k\sigma_{\bar{x}}$

$b = \mu - k\sigma_{\bar{x}}$



k is determined such that area below the curve between a,b is $(100 - 2\epsilon)\%$. Then the probability that $\bar{x}$ lies between a,b is given as

(1) $\Pr\left\{\mu - k\sigma_{\bar{x}} \leq \bar{x} \leq \mu + k\sigma_{\bar{x}}\right\} = (100 - 2\epsilon)\%$

Now $\because \mu - k\sigma_{\bar{x}} \leq \bar{x}$, $\therefore \mu \leq \bar{x} + k\sigma_{\bar{x}}$

$\because \mu + k\sigma_{\bar{x}} \geq \bar{x}$, $\therefore \mu \geq \bar{x} - k\sigma_{\bar{x}}$

Hence $\mu - k\sigma_{\bar{x}} \leq \bar{x} \leq \mu + k\sigma_{\bar{x}}$ is equivalent

to $\bar{x} - k\sigma_{\bar{x}} \leq \mu \leq \bar{x} + k\sigma_{\bar{x}}$

Consequently

(2) $\Pr\left\{\bar{x} - k\sigma_{\bar{x}} \leq \mu \leq \bar{x} + k\sigma_{\bar{x}}\right\} = (100 - 2\epsilon)\%$

$\therefore$ The upper Confidence limit for μ is

$$\bar{x} + k\sigma_{\bar{x}}$$

and the lower Confidence limit is

$$\bar{x} - k\sigma_{\bar{x}}$$

In large samples, $k = u_{\alpha}$ where

$u_{\alpha} = 1.96$ at 95 % confidence coefficient

& $u_{\alpha} = 2.58$ at 99 % confidence coefficient

$\sigma_{\bar{x}} = \sigma/\sqrt{n}$ or $s/\sqrt{n-1}$ according to as σ is known or unknown.

In small samples , $k = t_{\alpha,\nu}$ and we need to look in tables of t at $\nu = n-1$ degrees of freedom and at $\alpha = 100 - \gamma$ level of significance. But $\tilde{\sigma}_{\bar{x}} = s/\sqrt{n-1}$ where $s^2 = \Sigma(x_i - \bar{x})^2/n$.

Example

A standardized intelligence examination has been given for serveral years with an average score of 80 and a standard deviation 7. A group of 36 students are taught with special emphasis on reading skill. If the 36 students obtained a mean grade of 83 on this examination, is there reason to believe that the special emphasis changes the results on the test?

1- The null hypothesis is $\mu_o = 80$

The alternative " " $\mu_o \neq 80$

2- Test criterion

$$u = z = \frac{-\mu_o}{\sigma/\sqrt{n}} = \frac{83 - 80}{7/\sqrt{35}} = 2.53$$

3- Comparing this $|u| = 2.53$ with $|u_{1\%}| = 2.58$ we accept the null hypothesis. In other words, there is no evidence that special emphasis on reading skill changes the results on the test.

Example:

A fertilizer mixing machine is set to give 10 pounds of nitrate for every 100 pounds of fertilizer Ten bags (each of 100 pounds) are examined. The per centages of nitrate are as follows:

9, 12, 11, 10, 11, 9, 11, 12, 9, 10

Is there reason to believe that the mean is not equal to 10%.

Mean $(\bar{x}) = 10.4$

Variance $(s^2) = 1.24$

$$\tilde{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n-1}} = 0.37$$

Null hypothesis $\mu_o = 10$

alternative hypothesis $\mu_o \neq 10$

The test criterion is

$$t = \frac{\bar{x} - \mu_o}{\tilde{\sigma}_{\bar{x}}} = \frac{10.4 - 10}{0.37} = 1.11$$

From the tables $t_{9,\ 1\%} = 3.25$

$\therefore$ We accept the null hypothesis
There is an evidence that the pupulation mean is 10.

3) Testing statistical hypothesis around the difference between two means

3-1) Independent samples

(i) Large samples: Given two independent random samples. The 1st sample is of size n_1, has a mean $\bar{x}_1$, a standard deviation s_1 and is drawn from a normal population $N(\mu_1, \sigma_1^2)$. The 2nd sample is of size n_2, has a mean $\bar{x}_2$, a standard deviation s_2 and is drawn from a normal population $N(\mu_2, \sigma_2^2)$. Then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is exactly normal with expected value $\mu_1 - \mu_2$ and standard deviation

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

If the populations are infinite and the samples are large, then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ approaches normal with mean $\mu_1 - \mu_2$ and standard diviation

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Usually σ_1^2 & σ_2^2 are unknown and hence the sampling distribution of $\bar{x}_1 - \bar{x}_2$ approaches normal with mean $\mu_1 - \mu_2$ and standard deviation $\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}$.

Now to test the null hypothesis $\mu_1 - \mu_2 = 0$ against the alternative hypothesis $\mu_1 - \mu_2 \neq 0$ we use the test criterion

$$u = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} \sim N(0,1)$$

where

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

$$\text{or } \tilde{\sigma}_{\bar{x}_1-\bar{x}_2} = \frac{s_1^2}{n_1-1} + \frac{s_2^2}{n_2-1}$$

according to σ_1^2, σ_2^2 are known or unknown. Hence we compare u with u_α at α % level of significance

(ii) Small samples:- In large samples we know that the standard error of x_1-x_2, i.e.

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

Assuming that the two populations have the same variance i.e.

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

$$\therefore \sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})}$$

But σ^2 is usually unknown and can be replaced by its estimate $\tilde{\sigma}^2$ then

$$\tilde{\sigma}_{\bar{x}_1-\bar{x}_2} = \sqrt{\tilde{\sigma}^2(\frac{1}{n_1} + \frac{1}{n_2})}$$

where

$$\tilde{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{\sum(x_1-\bar{x}_1)^2 + \sum(x_2-\bar{x}_2)^2}{n_1 + n_2 - 2}$$

Hence, the test criterion

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\tilde{\sigma}_{\bar{x}_1 - x_2}}$$

is distributed as t-distribution with $\gamma = n_1+n_2-2$ degrees of freedom. The calculated t will be compared with $t_{\gamma,\alpha}$ at α % level of significance.

(.3-2) Related samples

In some problems we have to take one random sample and call its values-before treatment - the control group and its values-after treatment - the experimental group. For examples, to study the effect of a certain diet on the weight of Children, the effect of a certain trainning course on the achievement of students and the effect of a certain medecine on the recovery from a given disease.

Let us denote the observations for the control group by x_c and those for the experimental group by x_e. Then the differences will be $d = x_e - x_c$.

Now the null hypothesis is $D = 0$
the alternative " is $D \neq 0$

The test critinon

for large samples is $u = \dfrac{\bar{d}}{s_d/\sqrt{n-1}}$

& for small samples is $t = \dfrac{\bar{d}}{s_d/\sqrt{n-1}}$

with $\nu = n-1$ degrees of freedom

Example Two independent random samples are drawn from large populations A,B respectively. The following information is known

	n	$\bar{x}$	s
Sample A	122	105	33
Sample B	50	115	28

Test that the two populations have identical means.

$$\tilde{\sigma}^2_{\bar{x}_1-\bar{x}_2} = \frac{s_1^2}{n_1-1} + \frac{s_2^2}{n_2-1} = \frac{33^2}{121} + \frac{28^2}{49} = 25$$

$$\therefore \quad \tilde{\sigma}_{\bar{x}_1-\bar{x}_2} = 5$$

The null hypothesis is $\mu_1 - \mu_2 = 0$
The alternative " is $\mu_1 - \mu_2 \neq 0$

The test criterion is

$$u = \frac{\bar{x}_1-\bar{x}_2}{\tilde{\sigma}_{\bar{x}_1-\bar{x}_2}} = \frac{115 - 105}{5} = 2$$

Comparing with $|u_{5\%}| = 1.96$, we reject the null hypothesis. In other words, there is no evidence that the two samples are drawn from populations having the same mean.

Example:- A certain stimulus is to be tested for its effect on blood pressure. Twelve men have their blood pressure measured before and after the stimulus. The results are shown below:

Men:-	1	2	3	4	5	6	7	8	9	10	11	12
Before:	120	124	130	118	140	128	140	135	126	130	126	127
After:	128	131	131	127	132	125	141	137	118	132	129	135

Is there reason to believe that the stimulus world on the average raise the blood pressure.

The differences d_i are :-

8, 7, 1, 9, -8, -3, 1, 2, -8, 2, 5, 8

$$\bar{d} = \frac{22}{12} = 1.83$$

$$s_d^2 = \frac{434}{12} - (\frac{11}{6})^2 = 32.8056$$

$$s_d = 5.18$$

$$t = \frac{\bar{d}}{s_d / \sqrt{n-1}} = \frac{1.83 \times 11}{5.18} = 1.17$$

From the table $t_{11,5\%} = 2.20$

$\therefore$ We accept the null hypothesis that $\mu_1 - \mu_2 = 0$. In other words there is no evidence that the stimulas would on the average raise the blood pressure.

Exercise The following table shows the distribution of households according to the annual consumption expenditure of two random samples drawn from the rural and urban parts of the country.

Intervals	Frequencies	
	Urban	Rural
15-	6	43
25-	66	280
50-	167	474
75-	261	453
100-	621	760
150-	520	446
200-	367	238
250-	262	127
300-	373	126
400-	295	68
600-	125	15
800-	31	3
1000-1400	51	4
Total	3145	3037

Is the difference between the averages of annual consumption expenditure in rural and urban parts significant?

Exercise :- We have the weights ($\bar{x}$) of 10 childre before eating a certain food and their weights ($\bar{y}$) after eating this food.

Children:	1	2	3	4	5	6	7	8	9	10
x :	142	140	143	158	149	140	134	124	116	157
y :	146	139	148	161	151	138	135	127	115	162

Is there reason to believe that the food would on the average increase the weights of the Children?

-4) Testing the hypothesis around the proportion and the difference between two proportions

(6-4.1) If we draw a large sample of size n from an infinite dichotomus population and if the proportion of successes is p, then, to test the null hypothesis that the sample is drawn from a population where the proportion is $\emptyset = \emptyset_o$ against $\emptyset \neq \emptyset_o$ we use the test criterion

$$u = \frac{p - \emptyset_o}{\tilde{\sigma}_p}$$

where

$$\tilde{\sigma}_p = \sqrt{\frac{\emptyset_o(1-\emptyset_o)}{n}}$$

Example: A superint ndent of schools has stated that at least 60% of high school seniors expect to attend college. In a random sample of 200 cases only 96 say they are planning for college. Does this refute the superintendent's statement?

$$\tilde{\sigma}_p = \frac{0.6 \times 0.4}{200} = 0.1096$$

The null hypothesis is $\emptyset_o = 0.6$

The alternative " " $\emptyset_o \neq 0.6$

$$p = \frac{96}{200} = 0.48$$

The test criterion is

$$u = \frac{0.48 - 0.60}{0.1096} = 1.09$$

Comparing this with $|u_{5\%}| = 1.96$, we accept the null hypothesis. In other words there is an evidence that the superintendent's statement is accepted

-4.2) A large random sample of size n_1 with proportion of successes p_1, is drawn from an infinite dichotomus population where the proportion $\emptyset_1$ is unknown. Another large random sample of size n_2 with proportion of successes p_2 is drawn from another infinite dichotomus population where the proportion $\emptyset_2$ is unknown. It is required to test the null hypothesis $\emptyset_1 = \emptyset_2 = \emptyset$ against the alternative $\emptyset_1 \neq \emptyset_2$.

Since ϕ is unknown (common proportion), an estimate based on the sample values is given as $\hat{\phi}$ where

$$\hat{\phi} = (n_1 p_1 + n_2 p_2)/(n_1 + n_2)$$

Now, if the null hypothesis is true, the test criterion is

$$u = \frac{p_1 - p_2}{\sigma_{p_1 - p_2}}$$

where

$$\tilde{\sigma}_{p_1 - p_2} = \sqrt{\hat{\phi}(1 - \hat{\phi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Example In an experiment, it is found that 56 out of 6815 inoculated persons were attacked by a certain disease while 272 out of 11668 not inoculated persons were attacked by the same disease. Test the significance of the difference between the proportions of the attacked persons in the two populations.

$$p_1 = 56/6815 = 0.0082 = 0.008$$

$$p_2 = 272/11668 = 0.0233 = 0.023$$

$$\phi = \frac{56 + 272}{18483} = 0.0177 = 0.018$$

$$\tilde{\sigma}^2_{p_1 - p_2} = 0.18(0.82)\left(\frac{1}{6815} + \frac{1}{11668}\right) = 0.000034$$

$$\tilde{\sigma}_{p_1 - p_2} = 0.0059$$

$$u = \frac{0.023 - 0.008}{0.0059} = \frac{.015}{.0059} = 2.54$$

Comparing with $|u_{5\%}| = 1.96$, we reject the null hypothesis. In other words there is an evidence that the difference between the proportions of dichotomus in the two populations are significant

Exercise:- A random sample of size 450 is drawn from an infinite dichotomus population and we find that 50 of them were attacked by a certain disease. Test the null hypothesis that the proportion of attacked persons in the population is 0.5.?

Exercise:- A random sample of 40 individuals is drawn from an infinite population and we find that in the sample 10 individuals were unemployed. Another random sample of 30 individuals is drawn from another infinite population and we find that in this sample 10 were unemployed. Test the hypothesis that the proportions of unemployment in the two populations are the same?

-5) Testing the hypothesis around the population variance

(6-5.1) Given a random variable x_i having a normal distribution with mean μ and standard deviation σ, then the standardized variate

$$u_i = \frac{x_i - \mu}{\sigma} \qquad (1)$$

will have a normal distribution with mean 0 and unit standard deviation. Then by definition $\sum_{i=1}^{n} u_i^2$ will be distributed as χ^2 with n degrees of freedom i.e.

$$\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2} = \chi_n^2 \qquad (2)$$

If we replace μ by $\bar{x}$, then

$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2} = \chi_{n-1}^2 \qquad (3)$$

i.e. $$\frac{ns^2}{\sigma^2} = \chi_{n-1}^2 \qquad (4)$$

(4) is used as a test for the population variance, where s^2 is the sample variance.

-5.2) we can also determine the confidence limits between which we expect the variance of the normal population will lie with a given confidence coefficient $\gamma = 100\ (1-\alpha)\%$ where α is the level of significance.

The lower limit $\underline{\sigma^2}$ is determined from

$$\chi^2_{\text{upper}} = \frac{ns^2}{\underline{\sigma^2}}$$

The upper limit $\overline{\sigma^2}$ is determined from

$$\chi^2_{\text{lower}} = \frac{ns^2}{\overline{\sigma^2}}$$

Example (1) A random sample of size 5 and variance 40 is drawn from a normal population. Test the hypothesis that the population variance is 25.

The null hypothesis is $\sigma^2 = 25$

The alternative " " $\sigma^2 \neq 25$.

$$\therefore \chi^2 = \frac{ns^2}{\sigma^2} = \frac{5 \times 40}{25} = 8$$

From the tables $\chi^2_{0.975,\ 4} = 0.484$

$$\chi^2_{0.025,\ 4} = 11.143$$

$\therefore$ The obtained value of χ^2 lies between the two tabulated values. Hence we accept the null hypothesis. In other words there is an evidence that the population variance is 25.

Example (2) A random sample of size 10 and variance 30, is drawn from a normal population, determine the 99% confidence limits of the population variance.

$$\chi^2_{0.995,9} = 1.735$$

$$\chi^2_{0.005,9} = 23.589.$$

The lower limit $\underline{\sigma^2} = 10 \times 30/23.589 = 12.72$

The upper limit $\overline{\sigma^2} = 10 \times 30/1.735 = 172.91$

Remark When the sample size is large, we know that the sampling distribution of the statistic

$$x = \sqrt{2\chi^2} - \sqrt{2\nu - 1}$$

- where ν is the no. of d.f. - is very near to a normal curve $N(0,1)$.

Example (3) A random sample of size 146 & $s^2 = 197.26$ is drawn from a normal population. Test the hypothesis that population variance is 256.

$$\chi^2 = ns^2/\sigma^2 = \frac{146 \times 197.26}{256} = 112.5$$

$$\therefore x = -2$$

But at 5% level of significance $|u| = 1.96$.

$\therefore$ We reject the null hypothesis

Exercise (1) A random sample of 65 individuals is drawn from an infinite population. If the mean of the sample is 95 and its standard deviation is 16. Can we say that this sample is drawn from a population with mean equals to 100?

Exercise (2) A certain type of rat shows a mean gain in weight of 65 grams during the first 3 months of life. Twelve rats were fed a particular diet from birth until age 3 months and the following weight gains were observed:

55, 62, 54, 58, 65, 64, 60, 62, 59, 67, 62, 61.

Is there reason to believe at the 5% level of significance that the diet causes a change in the amount of weight gained?

Exercise (3) Two astronomers recorded observations on a certain star. The 120 observations obtained by the first astronomer have a mean reading of 1.20. The 80 observations obtained by the second astronomer have a mean 1.15. Past experience has indicated that these astronomers obtain readings with a variance of about 0.40. Does the difference between the two results seem reasonable?

Exercise (4) Two new types of rations are fed to pigs. It is desired to test whether one or the other of the types is better. A sample of 12 pigs is fed type (A) ration, and another sample of 12 pigs is fed type (B) ration. The gains in weight are recorded below

Type A	31	34	29	26	32	35	38	34	30	29	32	31
Type B	26	24	28	29	30	29	32	26	31	29	32	28

Exercise (5) A random sample of 10 observations

11, 11, 10, 12, 8, 10, 15, 8, 10, 8, is drawn from a normal population. Test that the population variance is 2.5.

(IX) Chi-Square and its Applications

Chi-Square & its application

In the field of research, the experimenter may face one of the following situations:-

1- The experimenter may believe that the students should be distributed equally regarding the preference of one of the three different subjects A,B,C. He draws a sample of 90 students and asking them which subject they mostly prefer, he get the following result

29 students mostly prefer subject A
28 " " " " B
33 " " " " C

According to what he believes, the expected number of students who will prefer any of the subjects should be 30.

The data can be represented in the following table

Subjects	observed frequencies	Expected frequencies
A	29	30
B	28	30
C	33	30

Are the observed frequencies Consistent with the expected frequencies? The null hypothesis to be tested is "there is consistencey".

2- In fitting a certain theoritical model to a given oberved frequency distribution, the experimenter may wish to test whether the model is a good fit or not. In other words, it is required to test the goodness of fit. For example, given the distribution of students according to their scores, the experimenter would like to test that this distribution will follow a normal curve. The null hypothesis to be tested is "there is a good fit".

3- Sometimes the experimenter may be faced with the data arranged in a bivariate tables where the two variates are graded into categories e.g.

Educational Status \ Marital Status	Single	married	divorced	widow
illiterate				
read & write				
meduim culificate				
higher "				

Such tables are called "Association tables" or "contingency tables"

The hypothesis to be tested is "there is no association" or "there is independence"

Now, in the above situations, we have:

Observed frequencies denoted by f_o
Expected frequencies denoted by f_e

The test criterion is

$$T = \sum \frac{(f_o - f_e)^2}{f_e} \qquad (1)$$

This test statistic is a discrete variable and has a sampling distribution which was proved to be very near to that of the continuous variate χ^2 defined as

$$\chi^2 = \sum u^2,$$

where u is a standardized normal variate. Hence we refer to T as χ^2 i.e.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \qquad (2)$$

and we use the corresponding tables for χ^2 defined by (2)

Note f_e in any cell should not be less than 5

Applications

Example (1) In a questionnair distributed to 200 students it is asked, "which subject A,B,C, or D do you prefer?" It is found that 45 prefer A, 54 prefer B, 46 prefer C and 55 prefer D. Test the hypothesis that there is consistency with what is believed that the four subjects are equally preferable.

Now the data can be represented in the following table

Subjects	f_o	f_e
A	45	50
B	54	50
C	46	50
D	55	50

$$\therefore \chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(45-50)^2}{50} + \frac{(54-50)^2}{50} + \frac{(46-50)^2}{50} + \frac{(55-50)^2}{50}$$

$$= \frac{82}{50} = 1.64$$

From the tables

$$\chi^2_{3,0.05} = 7.82$$

$\therefore$ We accept the null hypothesis. In other words, there is an evidence of excstence of consistency.

Example (2) A random sample of 2970 students is drawn from the 3rd grade secondary schools. They are given an achievement test. The observed frequency distribution of their seores is as follows:-

Intervals:	40	40-	50-	60-	70-	80-	90-
f :	1	4	7	48	130	338	592

Intervals :	100-	110-	120-	130-	140-	150-	160-	200
f :	719	610	330	120	55	13		

Do we consider this sample as drawn from a normal population having the same mean and standard deviation?

The mean ($\bar{x}$) = 104.56
The standard deviation (s) = 16.99

The following table summerises the results of fitting a normal

Upper limits	f_o	Standardized Values	Area below	Δ	f_e
39.5	1	- 3.83	0.0001	0.0001	0.30
49.5	4	- 3.24	0.0004	0.0005	1.49
59.5	7	- 2.65	0.0040	0.0036	10.10
69.5	48	- 2.06	0.0197	0.0157	46.63
79.5	140	- 1.47	0.0708	0.0511	151.77
89.5	338	- 0.89	0.1867	0.1159	344.22
99.5	522	- 0.30	0.3821	0.1954	580.34
109.5	719	0.29	0.6141	0.2320	689.04
119.5	610	0.88	0.8106	0.1965	583.61
129.5	330	1.47	0.9292	0.1186	352.24
139.5	120	2.06	0.9803	0.0511	151.77
149.5	55	2.65	0.9960	0.0157	46.63
159.5	13	3.23	0.9994	0.0034	10.10
199.5	3	5.59	1.0000	0.0006	1.78
	2970				2970.02

In order that none of the expected frequencies should be less than 5, we add the 1st three expected frequencies (i.e.: 0.30+1.49+10.10 = 11.89) and the corresponding observed frequencies (i.e. 1+4+7 = 12) and consider each as one class. Similarly we add the last two expected frequencies (i.e. 10.10+1.78 = 11.88) and the corresponding observed frequencies (i.e. 13+3 = 16). Calculating χ^2 we find that

$$\chi^2 = \frac{\sum(f_o - f_e)^2}{f_e} = 16.97$$

The number of degrees of freedom $\nu = 11-3 = 8$ where 11 is the number of classes and 3 is the number of constraints (the total, the mean and the standard deviation).

At 1% level of significance we find from the table

$$\chi^2_{8,1\%} = 20.09$$

$\therefore$ We accept the null hypothesis. In other words there is an evidence that the distribution of the scores in the population is normal with mean 104.56 and standard deviation 16.99

Example (3) Five coins were tossed 100 times. The frequency distribution for the number of coins on which the head occurs is given below

x :	0	1	2	3	4	5	Total
f :	2	14	20	34	22	8	100

Fit a binomial distribution and test for the goodness of fit.

$$\bar{x} = 2.84 = np$$
$$5p = 2.84 \qquad \therefore\ p = 0.57$$

The probability binomial distribution is given by

$$p(x) = \binom{n}{x} p^x q^{n-x}$$

In our example

$$p(x) = \binom{5}{x}(0.57)^x(0.43)^{5-x}$$

for $x = 0,1,2,3,4,5$.

The table below gives the observed frequencies and the expected frequencies calculated from the above formula.

x	f_o	Probabilities	Expected freq. f_e
0	2	0.0147	1.47
1	14	0.0974	9.744
2	20	0.25832	25.832
3	34	0.34242	34.242
4	22	0.22695	22.695
5	8	0.06017	6.017
	100		100.000

$\chi^2 = 4.035$

Degrees of freedom = 5 - 2 = 3

From the tables $\chi^2_{3,5\%} = 7.815$

$\therefore$ We accept the null hypothesis. In other words, there is an evidence that the distribution of the number of the coins on which the head occurs (in the population) is a binomial distribution with mean 2.84.

Contingency tables: In general the table can be represented as follows where we have " " raws and "s" columns

A \ B	1	...	j	. . .	s	
1	f_{11}	. . .	f_{ij}	. . .	f_{1s}	$f_{1.}$
. .	. .		. .		. .	. .
i	f_{i1}	. . .	f_{ij}	. . .	f_{is}	$f_{i.}$
. .	. .		. .		. .	. .
r	f_{r1}	. . .	f_{rj}	. . .	f_{rs}	. .
	$f_{.1}$	. . .	$f_{.j}$	. . .	$f_{.s}$	$f_{..}$

f_{ij} denotes the frequency in the (i,j) cell

$f_{i.} = \sum_{j=1}^{s} f_{ij}$ denotes the frequency in the $i^{\underline{th}}$ row

$f_{.j} = \sum_{i=1}^{r} f_{ij}$ denotes the frequency in the $j^{\underline{th}}$ column

$f_{..} = \sum_{i} \sum_{j} f_{ij} = \sum_{i} f_{i.} = \sum_{j} f_{j.}$ denotes the total frequency.

It should be noticed that no association or independence occurs in two cases:-

(i) if the frequencies in the cells are all equal

or (ii) if the proportions of the corresponding frequencies in any two rows (or columns) do not differ from column to column (or row to row).

The expected frequency f_e for the observed $f_o = f_{ij}$ is calculated as

$$f_e = f_{i.} \, f_{.j} / f_{..}$$

Now $\sum_i f_e = \sum_i f_{i.} f_{.j} / f_{..} = f_{.j}$

& $\sum_j f_e = \sum_j f_{i.} f_{.j} / f_{..} = f_{i.}$

This means that the marginal frequencies for the observed and expected frequencies are the same. Consequently the number of degrees of freedom γ is given as

$$\gamma = rs - (s-1) - (r-1) = (r-1)(s-1)$$

Example:- In a questionnaire filled by a random sample of 2026 students, we ask, "Do you like to work in the givernment or on your own? "The students are also classified into three groups A,B,C according to their father's job. The results are given in the following table:-

		Father's Job			Total
		A	B	C	
Attitude toward working in the government	In the government	120 (58.9)	208 (242.0)	13 (40.1)	341
	On their own	230 (291.1)	1230 (1196.0)	225 (197.9)	1685
	Total	350	1438	238	2026

The expected frequencies are given between brackets

Now $$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 103.904$$

The number of degrees of freedom is $\gamma = (2-1)(3-1) = 2$. At 1% level of significance we find that $\chi^2 = 9.21$, with 2 d.f. Hence we rejict the null hypothesis. In other words there is an evidence that there is an association between attitude towards working in the government and father's Job.

2X2 tables

Let us consider the case where each of the two variates A&B is classified into two categories. The tables are known as 2X2 tables and take the following form:-

		Variate B		Total
		B_1	B_2	
Variate A	A_1	a	b	a+b
	A_2	c	d	c+d
	Total	a+c	b+d	N

where $N = a + b + c + d$

The expected frequency for "a" is $(a+b)(a+c)/N$

$\therefore$ The difference between "a" and its expected values will be

$$a - \frac{(a+b)(a+c)}{N} = \frac{ad - bc}{N}$$

Similarly the differences for b, c, d and their expected values are

$$-\frac{ad - bc}{N}, -\frac{ad - bc}{N}, \frac{ad - bc}{N}$$

respectively

$$\therefore \chi^2 = \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(ad-bc)^2}{N^2}\left[\frac{1}{(a+b)(a+c} + \frac{1}{(a+b)(b+d)} + \frac{1}{(c+d)(a+c)} + \frac{1}{(c+d)(b+d)}\right]$$

$$= \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Example

The distribution of a sample of 1464 individuals according to sex and attitude towards smoking is given in the following table

	Smoke	Do not Smoke	Total
Men	295	462	757
Women	183	524	707
Total	478	966	1464

The hypothesis to be tested is that there is no association between smoking and sex. Using the above formula the value of χ^2 is 28.5. For γ =1, and at 1% level of significance we have from the tables χ^2 =6.6. Hence we reject the hypothesis. In other words there is an evidence that in the population, attitude towards smoking depends upon sex.

Note (1) when the sample size is < 500, we use Yate's correction which is only appropriate for χ^2 with 1 degree of freedom. The procedure is to change the frequency in each cell by 0.5, keeping the marginal total unchanged and reducing the size of χ^2. Hence

$$\chi^2 = \frac{N(|ad - bc| - N/2)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Note (2) It is known that for a large sample if χ^2 has 1 degree of freedom, $\sqrt{\chi^2}$ has a distribution which is the right hand half of a normal distribution. Hence, we can use χ^2 with 1 degree of freedom as a test for the hypothesis that the proportions are the same in the two populations. In the above example, the hypothesis to be tested may be stated as follows: the proportion of persons who smoke is the same among men as among women. The outcome of the test gives an evidence that equal proportions in the two populations is untenable.

In testing the difference between two proportions, the test criterion is

$$u = \frac{p_1 - p_2}{\tilde{\sigma}_{p_1 - p_2}}$$

where $$\tilde{\sigma}_{p_1 - p_2} = \sqrt{\hat{\phi}\,(1 - \hat{\phi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

In terms of the 2 X 2 table,

$$p_1 = \frac{a}{a+b}\,,\quad p_2 = \frac{c}{c+d}\,,\quad \hat{\phi} = \frac{a+c}{N}$$

$$\therefore (p_1-p_2)^2 = \left(\frac{a}{a+b} - \frac{c}{c+d}\right)^2 = \frac{(ad - bc)^2}{(a+b)^2(c+d)^2}$$

$$\hat{\phi}(1-\hat{\phi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{a+c}{N}\,\frac{b+d}{N}\left(\frac{1}{a+b} + \frac{1}{c+d}\right)$$

$$= \frac{(a+c)(b+d)}{N(a+b)(c+d)}$$

$$\therefore u^2 = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+b)(c+d)} = \chi_1^2$$

Note (3) In 2 X k tables, χ^2_{k-1} is used to test the null hypothesis that the k proportions are identical.

(X) Simple Analysis of Variance

Simple Analysis of Variance

1. Introduction: Analysis of variance is a technique applicable to testing the hypothesis that several independent samples have been drawn at random from a common normal population. The development of the analysis of variance as a powerful test in experimental and research work is largely the responsibility of R.A. Fisher and his Co-workers.

The analysis of variance has proved to be not only convenient method but also a powerful method of analysis for the research worker is testified to by the extent to which it is being used in the planning design and analysis of research in a variety of fields.

Basically the analysis of variance is a simple arithmetical method of sorting out the components of variation in a given set of results. Whenever there is heterogeneity of variation, more than one component is present. Suppose that we have a population in which the people are all of the some race and the variable studied is height in inches. Four groups are drawn from the population. One group of men, another group of women, a third group of boys from 13-15 years of age and a fourth group of girls of a similar age. When the four groups are combined, the frequency distribution might appear reasonably normal but, we know that two components of variation are actually present, one representing variation within groups and another between the groups. The arithmetical procedure of the analysis of variance enables us to sort out and evaluate the components of variation for such mixed population.

In the present chapter, we are going to deal with analysis of variance in its simplest form, namely, the on-way classification or sometimes called simple randomized design. The importance of this design in experimental work cannot be overemphasized. Not only is the design widely employed by itself, but it constitutes a basic unit in nearly all of the more complex designs employed in experimental research. In this design each treatment is independently administered to a different group of subjects, all groups having been originally drawn at random from the same parent population. After the treatments have been administered, these groups may be regarded as random samples from a single population, only if the treatments all have identical effects on the distribution of criterion measures for the population. Otherwise, the group receiving treatment T_1 say, may be regarded as a random sample from a hypothetical population which is like the parent population except that all its members have received treatment T_1. The sample that received treatment T_2 may, likevise, be regarded as a random sample from a population like the original except that all members of this population have received T_2 etc.

The hypothesis to be tested is that the criterion means of these treatment populations are identical. The design can be represented as follows:-

T_1	T_2	....	T_j	....	T_k
x_{11}	x_{12}	....	x_{1j}	....	x_{1k}
x_{21}	x_{22}	....	x_{2j}	....	x_{2k}
.	.	....	.	....	.
.	.	....	.	....	.
.	.	....	.	....	.
x_{i1}	x_{i2}	....	x_{ij}	....	x_{ik}
.	.	....	.	....	.
.	.	....	.	....	.
.	.	....	.	....	.
x_{n1}	x_{n2}	....	x_{nj}	....	x_{nk}
$\bar{x}_{.1}$	$\bar{x}_{.2}$	....	$\bar{x}_{.j}$	....	$\bar{x}_{.k}$

$i = 1,2, \ldots, n$ & $j = 1,2, \ldots, k$

x_{ij} denotes the value of the $i^{\underline{th}}$ observation in the group of the $j^{\underline{th}}$ treatment

$\bar{x}_{.j}$ denotes the mean of the values of the $j^{\underline{th}}$ treatment

$\bar{x}_{..}$ denotes the over all mean

n denotes the number of observations in each group. It may differ from one treatment to another i.e. n_j where $j = 1,2, \ldots, k$.

If all the treatments have identical effects on the distribution of criterion measures for the parent population, that is, if the distribution of criterion measures is the same for all treatment populations, these populations may be regarded as just one population. In this case the various treatment groups may be all regarded as simple random samples from the same population whose variance we denote by σ^2.

2. Estimates for σ^2 and measure of discrepancy

From the experimental data we can derive two independent estimates for σ^2. One estimate is based on the differences among the observed treatment (group) means, the other upon the variance of the measures within the individual treatment groups. We can form the ratio of the first of these estimates to the second. If the treatments have identical effects, the 1st of these estimates will exceed the other only by chance. If the treatments really differ in effectiveness, then the differences among the observed treatment means will be larger than they would be and the ratio will then tend to be larger than 1.0. Accordingly we can use the ratio as the measure of the discrepancy between hypothesis and observations. If we find the sampling distribution of this ratio, we can use it in a statistical test of the hypothesis.

The 1st estimate (based on the means of the groups) If the k treatments have identical effects on the distribution of the criterion measures for the parent population the various treatment groups are all random samples from the same population and the means of these treatment groups are a simple random sample of k values from a population of such means. Hence an unbiased estimate is given as

$$\tilde{\sigma}^2_{\bar{x}} = \sum_{j=1}^{k} (\bar{x}_{\cdot j} - \bar{x})^2 / (k-1)$$

But

$$\tilde{\sigma}^2_{\bar{x}} = \tilde{\sigma}^2 / n$$

$$\tilde{\sigma}^2_{(1)} = n \tilde{\sigma}^2_{\bar{x}} = \frac{n \sum (\bar{x}_{\cdot j} - \bar{x})^2}{k - 1} \qquad (1)$$

The 2nd estimate (based on the individuals within groups) If the treatments have identical effects on the criterion distribution, each treatment group is a simple random sample from the same population. For the jth treatment group, an unbiased estimate for σ^2 is given as

$$\tilde{\sigma}^2_{j} = \sum_{j} (x_{ij} - \bar{x}_{\cdot j})2 / (n - 1)$$

A better estimate can be secured by averaging

$$\tilde{\sigma}_1^2 , \quad \tilde{\sigma}_2^2 , \cdots , \quad \tilde{\sigma}_k^2 , \quad \text{i.e.,}$$

$$\tilde{\sigma}_{(2)}^2 = \sum_{j=1}^{k} \tilde{\sigma}_j^2 = \frac{\sum_j \sum_i (x_{ij} - \bar{x}_{.j})^2}{k(n-1)} \qquad (2)$$

3. Inter pretation of $\dfrac{\tilde{\sigma}_{(1)}^2}{\tilde{\sigma}_{(2)}^2}$

$$\frac{\tilde{\sigma}_{(1)}^2}{\tilde{\sigma}_{(2)}^2} = \frac{n \sum_j (\bar{x}_{.j} - \bar{x})^2 / (k-1)}{\sum_j \sum_i (x_{ij} - \bar{x}_{.j})^2 / \; k(n-1)} \qquad (3)$$

We know that if the population has a normal distribution, then the sampling distribution of the means of random samples drawn from this population is also normal having the same mean as the population mean and variance σ^2/n.

$$u = \frac{(\bar{x}_{.j} - \bar{x})}{\sigma / \sqrt{n}} \sim N(0,1)$$

$$\therefore \sum_{j=1}^{k} \frac{n(\bar{x}_{.j} - \bar{x})^2}{\sigma^2} = \sum_{j=1}^{k} u^2 = \chi^2_{k-1} \qquad (4)$$

Also, for the j^{th} group

$$u = \frac{x_{ij} - \bar{x}_{.j}}{\sigma} \sim N(0,1)$$

$$\sum_{i=1}^{n} \frac{(x_{ij} - \bar{x}_{.j})^2}{\sigma^2} = \sum_{i=1}^{n} u^2 = \chi^2_{n-1}$$

Hence

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \frac{(x_{ij} - \bar{x}_{.j})^2}{\sigma^2} = \sum_{j=1}^{k} \chi^2_{n-1} = \chi^2_{k(n-1)} \qquad (5)$$

(applying the additive property of χ^2)

Dividing both numerator and denominator by σ^2 and using (4),(5) we get

$$\frac{\tilde{\sigma}^2_{(1)}}{\tilde{\sigma}^2_{(2)}} = \frac{\chi^2_{k-1} / (k-1)}{\chi^2_{k(n-1)} / k(n-1)} \sim F_{k-1,k(n-1)} \qquad (6)$$

Note: The above derivation is valid even if the number of observations (n_j) differs from one group to another

1. $$\frac{x_{ij} - \bar{x}_{.j}}{\sigma} = u$$

$$\sum_{i=1}^{n_j} \frac{(x_{ij} - \bar{x}_{.j})^2}{\sigma^2} = \sum_{i=1}^{n_j} u^2 = \chi^2_{n_j-1}$$

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} \frac{(x_{ij} - \bar{x}_{.j})^2}{2} = \sum_{j} \chi^2_{n_j-k} = \chi^2_{N-k}$$

where $$N = \sum_{j=1}^{k} n_j$$

2. Also $$\frac{(\bar{x}_{.j} - \bar{x})}{\sigma / \sqrt{n_j}} = u_j$$

$$\therefore \sum_{j=1}^{k} \frac{n_j(\bar{x}_{.j} - \bar{x})^2}{\sigma^2} = \sum_{j=1}^{k} u_j^2 = \chi^2_{k-1}$$

$$\frac{\tilde{\sigma}^2_{(1)}}{\tilde{\sigma}^2_{(2)}} = \frac{\sum_{j=1}^{k} n_j \frac{(\bar{x}_{.j} - \bar{x})^2}{\sigma^2} / (k-1)}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} \frac{(x_{ij} - \bar{x}_{.j})^2}{\sigma^2} / (N-k)} = \frac{\chi^2_{k-1} / (k-1)}{\chi^2_{N-k} / (N-k)}$$

$$= F_{k-1,N-k}$$

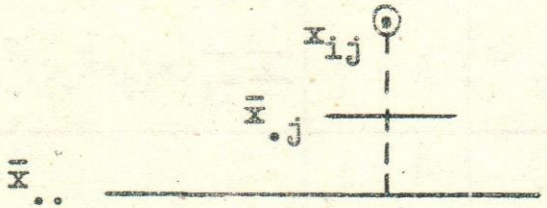
4. Analysis of Variance table

Partition of total sum of squares

Again let X_{ij} denote the i^{th} observation in the j^{th} group or (sample)

$\bar{x}_{.j}$ denote the mean of the j^{th} group

& $\bar{x}_{..}$ denote the over all mean



Hence

$x_{ij} - \bar{x}_{..}$ represents the deviation of the i^{th} observation in the j^{th} group from the over all mean.

$x_{ij} - \bar{x}_{.j}$ represents the deviation of the i^{th} observation in the j^{th} group from the mean of the j^{th} group.

$\bar{x}_{.j} - \bar{x}_{..}$ represents the deviation of the mean of the j^{th} group from the over all mean

Now $x_{ij} - \bar{x}_{..} = (x_{ij} - \bar{x}_{.j}) + (\bar{x}_{.j} - \bar{x}_{..})$

Squaring both sides

$$\sum_j \sum_i (x_{ij} - \bar{x}_{..})^2 = \sum_j \sum_i (x_{ij} - \bar{x}_{.j})^2 + \sum_j n_j (\bar{x}_{.j} - \bar{x}_{..})^2$$

i.e.

Total sum of squares (SST) = Sum of squares within groups (SSW) + Sum of Squares between groups (SSB)

Where SST has N-1 degrees of freedem
SSW has N-k " " "
SSB has k-1 " " "

For numerical calculations of the sums of Squares we have

$$SST = \sum_j \sum_i (x_{ij} - \bar{x}_{..})^2 = \sum_j \sum_i x_{ij}^2 - \frac{(\sum_j \sum_i x_{ij})^2}{N}$$

$$SSB = \sum_j n_j (\bar{x}_{.j} - \bar{x}_{..})^2 = \sum_j \frac{(\sum_i x_{ij})^2}{n_j} - \frac{(\sum_j \sum_i x_{ij})^2}{N}$$

and by subtraction we get SSW.

$(\sum_j \sum_i x_{ij})^2 / N$ is called the correction term.

A summary can be arranged in the following analysis of variance table

Source of variation	Sum of Squares	Degrees of freedom	Mean Sum of Squares
Between groups	$\sum_j n_j(\bar{x}_j - \bar{x}_{..})^2$	k - 1	$\sum_j n_j(\bar{x}_j - \bar{x}_{..})^2/(k-1)$
Within groups	$\sum_j \sum_i (x_{ij} - \bar{x}_{.j})^2$	N - k	$\sum_j \sum_i (x_{ij} - \bar{x}_{.j})^2/(N-k)$
Total	$\sum_j \sum_i (x_{ij} - \bar{x}_{..})^2$	N - 1	

5- Expected values of sums of squares

Let x_{ij} denote the value of the character x for the i^{th} observation in the j^{th} sample (or group) we may write the model as

$$x_{ij} = \mu_j + z_{ij}$$

where z_{ij} is a random variable with expected value o & standard deviation σ_j. (j=1,2,...,k).

We suppose that the sampling is random within each group and x_{ij} from different group are independent. Further we assume that the x_{ij}'s are independent in the same group. We also assume that $\sigma_1 = \sigma_2 = \ldots = \sigma_k$ i.e. the groups from which the samples are drawn have common variance.

In the above model the statistical hypothesis to be tested is

$$\mu_1 = \mu_2 = \ldots = \mu_j = \ldots = \mu_k.$$

This model can be modified replacing μ_j by $\mu + \lambda_j$ where μ is a parameter representing the over all average level of the character measured. λ_j represents the diviations of the average of each group from the whole average. Without any loss of generality we can assume that $\sum n_j \lambda_j = 0$

The model becomes

$$x_{ij} = \mu + \lambda_j + z_{ij} \qquad (1)$$

This can be presented in another way. Assume that the x_{ij} are independently $N(\mu_j, \sigma^2)$ Consider the identity

$$x_{ij} = \mu + (\mu_j - \mu) + (x_{ij} - \mu_j)$$

Let $\lambda_j = \mu_j - \mu$ and $\mu = \sum_{j=1}^{k} n_j \mu_j / N$

where $N = \sum_{j=1}^{k} n_j$

so that $\sum_{j=1}^{k} n_j \lambda_j = \sum_{j=1}^{k} n_j (\mu_j - \mu) = \sum_{j=1}^{k} n_j \mu_j - N$

$= N - N = 0$

Also define $z_{ij} = x_{ij} - \mu_j$ which has mean 0 since the mean of x_{ij} is μ_j. But both z_{ij} & x_{ij} have the same variance σ^2. Finally z_{ij} are normally distributed since the x_{ij} came from a normal distribution.

$$\therefore \quad x_{ij} = \mu + \lambda_j + z_{ij} \qquad \begin{cases} i = 1,2,\ldots,n_j \\ j = 1,2,\ldots,k \end{cases}$$

z_{ij} are independently $N(0, \sigma^2)$

$$\sum_{j=1}^{h} n_j \lambda_j = 0$$

Now

$$\bar{x}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} = \mu + \lambda_j + \bar{z}_{.j} \qquad (2)$$

$$\bar{x}_{..} = \frac{1}{N} \sum_{j} \sum_{i} x_{ij} = \frac{1}{N} \sum_{j=1}^{k} n_j \bar{x}_{.j} = \mu + \bar{z}_{..} \qquad (3)$$

$$\therefore \quad \bar{x}_{.j} - \bar{x}_{..} = \lambda_j + (\bar{z}_{.j} - \bar{z}_{..}) \quad \& \quad x_{ij} - \bar{x}_{.j} = z_{ij} - \bar{z}_{.j}$$

Hence

1- Sum of Squares between groups (SSB) $= \sum_{j=1}^{k} n_j (\bar{x}_{.j} - \bar{x}_{..})^2$

$$= \sum_{j=1}^{k} n_j \left[\lambda_j + (\bar{z}_j - \bar{z}_{..}) \right]^2$$

$$\therefore \mathcal{E}(SSB) = \sum_{j=1}^{k} n_j \lambda_j^2 + \mathcal{E}\left\{ \sum_{j=1}^{k} n_j (\bar{z}_{.j} - \bar{z}_{..})^2 \right\}$$

$$= \sum_{j=1}^{k} n_j \lambda_j^2 + (k-1) n_j \frac{\sigma^2}{n_j}$$

$$= (k-1)\sigma^2 + \sum_{j=1}^{k} n_j \lambda_j^2 \qquad (4)$$

2- Sum of squares within groups (SSW) $= \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{.j})^2$

$$= \sum_{j=1}^{k} \sum_{i=1}^{n_j} (z_{ij} - \bar{z}_{.j})^2$$

$$\therefore \quad (SSW) = \sum_{j=1}^{k} (n_j - 1)\sigma^2 = (N-k)\sigma^2 \qquad (5)$$

These can be represented in the following summary table

Source of variation	Sum of Squares	Degrees of freedom	Expected value of mean sum of Squares
Between groups	$\sum_{j=1}^{k} n_j (\bar{x}_{.j} - \bar{x}_{..})^2$	k-1	$\sigma^2 + \sum_{j=1} n_j \sigma_j^2/(k-1)$
Within groups	$\sum_j \sum_i (x_{ij} - \bar{x}_{.j})^2$	N-k	σ^2
Total	$\sum_j \sum_i (x_{ij} - \bar{x}_{..})^2$	N-1	

If the hypothesis tested is valid i.e. $\lambda_1 = \lambda_2 = \ldots = \lambda_k = 0$, (4) becomes $(k-1)\sigma^2$, while (5) is still $(N-k)\sigma^2$. A large value of the mean square ratio obtained from (4), (5) by (k-1), (N-k) the corresponding degrees of freedom respectively may be regarded as evidence of difference between groups independently of any assumption about the distribution of the z_{ij}'s. That is we have no system of significance limits against which the observed values may be judged. To obtain such significance limits, we assume that each of the z_{ij} is normally distributed i.e. $z_{ij} \sim N(0,\sigma)$. On the null hypothesis i.e. $\lambda_j = 0$, the mean square ratio is

$$\frac{\sum_j n_j(\bar{z}_{.j}-\bar{z}_{..})^2/(k-1)}{\sum_j \sum_i (z_{ij}-\bar{z}_{.j})^2/(N-k)}$$

But $\frac{1}{k-1}\sum_{j=1}^{k} n_j(\bar{z}_{.j}-\bar{z}_{..})^2 \sim \chi^2_{k-1}\sigma^2/(k-1)$

$\frac{1}{N-k}\sum_j\sum_i (z_{ij}-\bar{z}_{.j})^2 \sim \chi^2_{N-k}\sigma^2/(N-k)$

∴ Sum of squares within groups independent of sum of squares between groups, then

Mean square ratio = $\frac{MSSB}{MSSW} = \frac{\chi^2_{k-1}/(k-1)}{\chi^2_{N-k}/(N-k)} \sim F_{k-1,N-k}$

Example Random samples of dixs are drawn from 4 different machines. The values of their diameters are given as follows

machines

A	B	C	D
25	24	23	24
20	26	23	23
28	30	32	33
15	25	24	26
	35	28	29
	34		

Taking 25 as an arbitrary origin we have

	A	B	C	D		
	0	-1	-2	-1		
	-5	1	-2	-2		
	3	5	7	8		
	-10	0	-1	1		
		10	+3	4		
		9				
n_j	4	6	5	5	20	$= N$
$\sum_i x_{ij}$	-12	24	5	10	27	$= \sum_j\sum_i x_{ij}$
$\sum_i x^2_{ij}$	134	208	67	86	495	$= \sum_j\sum_i x^2_{ij}$
$(\sum x_{ij})^2/n_j$	36	96	5	20	157	$= \sum_j \frac{(\sum x_{ij})^2}{n_j}$

Correction term $= \frac{(27)^2}{20}$.

SST $= 495 - 36.45 = 458.55$

SSB $= 157 - 36.45 = 120.55$

Analysis of variance table

Source of Variation	Sum of Squares	Degrees of freedom	Mean sum of Squares.
Between groups	120.55	3	40.183
Within groups	338.00	16	21.125
Total	458.55	19	

Mean Ratio $= \frac{40.183}{21.125} = 1.90$

From the tables we have $F_{3,16} = 3.24$

∴ We accept the null hypothesis, i.e., there is an evidence that the population means of the diameters of the discs produced by the 4 machines are identical

Note If F proves to be significant this means that $\lambda_j \neq 0$. In other words the population means are not identical. To test the significance of the difference between the means of any two populations from which samples h,g are drawn we apply

$$t = \frac{\bar{x}_h - \bar{x}_g}{\sqrt{\tilde{\sigma}^2 \left(\frac{1}{n_h} + \frac{1}{n_g}\right)}}$$

n_h, $\bar{x}_h$ are the size and the mean of the sample "h", n_g, $\bar{x}_g$ are the size and the mean of the sample "g" & $\tilde{\sigma}^2$ is the mean sum or squares of within groups.

Bartlett's test for homogeneity of variances

In analysis of variance it is necessary to assume that a number of population variances are equal. If we have reason to doubt that this is the case, we may want to test the null hypothesis

$$H_o : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

gainst the alternative

H_1 : at least two variances differ Under the assumptions:-

(i) k random samples are drawn from k populations

(ii) the k populations are normal, the statistic

$$\frac{2.3026}{C}\left[(N-k)\ \log_{10} s_p^2 - \sum_{j=1}^{k}(n_j-1)\log_{10} s_j^2\right]$$

is approximately distributed as X^2 with k-1 degrees of freedom if H_o is true. Here s_1^2 , s_2^2 , ... , s_k^2 are the k samples variances. Also

$$s_p^2 = \frac{1}{N-k}\sum_{j=1}^{k}(n_j-1)\ s_j^2 ,$$

$$s_j^2 = \frac{1}{n_j-1}\sum(x_{ij} - \bar{x}_j)^2$$

and

$$C = 1 + \frac{1}{3(k-1)}\left[\sum_{j=1}^{k}\frac{1}{n_j-1} - \frac{1}{N-k}\right]$$

Example To test the homogeneity of the populations from which the following samples are drawn

a. 8, 4, 7, 5, 6

b. 6,12,14, 8

c. 4, 2, 2, 1, 2, 1

d. 3, 2, 4

e. 10, 7,15,17,13, 8,14

a	b	c	d	e
8	6	4	3	10
4	12	2	2	7
7	14	2	4	15
5	8	1		17
6		2		13
		1		8
				14

n_j	5	4	6	3	7	$25 = N$
$\sum_i x_{ij}$	30	40	12	9	84	$175 = \sum_j \sum_i x_{ij}$
$\sum_i x_{ij}^2$	190	440	30	29	1092	$1781 = \sum_j \sum_i x_{ij}^2$
$(\sum x_{ij})^2/n_j$	180	400	24	27	1008	$1639 = \sum_j \frac{(\sum_i x_{ij})^2}{n_j}$
$\sum_i (x_{ij}-\bar{x}_{.j})^2$	10	40	6	2	84	

Samples	$\sum_i (x_{ij}-\bar{x}_{.j})^2$	n_j-1	s_j^2	$\log_{10} s_j^2$	$(n_j-1)\log_{10} s_j^2$
a	10	4	2.5	0.3979	1.5916
b	40	3	13.3	1.1239	3.3717
c	6	5	1.2	0.0792	0.3960
d	2	2	1.0	0.0000	0.0000
e	84	6	14.0	1.1461	6.8766
	142	20			$12.2359 = \sum_j (n_j-1)\log_{10} s_j^2$

$$s_p^2 = 142/20 = 7.1$$

$$\log_{10} s_p^2 = \log_{10} 7.1 = 0.8513$$

$$\therefore (N-k) \log_{10} s_p^2 = 20 \times 0.8513 = 17.026$$

$$= 1.117$$

$$\therefore \chi_4^2 = \frac{2.3026}{1.117} \left[17.0260 - 12.2359 \right] = 9.88$$

But from the tables $\chi^2_{4,5\%} = 9.49$

$\therefore$ We reject the null hypothesis $H_o: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

In other words there is no evidence that the populations from which these samples are drawn - have common variance

Exercises

1- A professor is trying to select a good text book for his elementary course in statistics from 4 different ones which are available. He has 37 students whom he distributed at random to 4 different groups. The assignment of textbooks to groups is also done at random. After the course is over all students who are still enrolled take the same examination. The results are

Textbook

A	B	C	D
68	41	54	44
68	47	44	51
69	54	51	69
60	65	56	59
73	32	47	59
64	73	61	55
71	44	59	66
67	48	49	
75	64	41	
	54	61	
		73	

Assuming that differences attributed to using a number of textbooks are the only variables that need be considered, what conclusions can be drawn?

2- Worms are classified into three groups by a structural characteristic. Three random samples of 11 are taken from each group and the length of each worm is measured. These data are recorded as follows

Group (1) 8.9,9.7,11.5,8.2,10.5,10.8,11.0,8.0,9.9,11.0,11.0

Group (2) 12.2,12.0,11.5,8.7,10.5,9.0,10.5,13.0,13.0,11.0,11.1

Group (3) 9.5,8.0,8.3,10.0,9.5,10.0,11.3,10.5,8.0,8.0,9.2

Test the hypothesis that the mean length of each group is the same

3- Jenkins & Snedecor compared the yields of a number of varieties of corn. Each variety being represented by several inbred lines. Six varieties with yields (bushels per acre) of their inbred lines are follows:

V_1 : 7.4, 7.3,4.5,7.4,5.0,5.9,6.4,6.3,5.0,6.1,7.9,5.7

V_2 : 7.7,5.4,5.2,4.0

V_3 : 6.9,6.8,7.6,8.1,9.4,12.0,15.9,7.4,9.0,5.2,9.2,8.6

V_4 : 9.6,7.8,9.6,7.7,8.2,7.3,11.3,9.5,8.8,8.4,6.8

V_5 : 4.8,9.2,8.5,8.8,7.9,5.9,9.2

V_6 : 4.3,8.4,6.6,4.9,5.8,7.6,5.7

(XI) SIMPLE LINEAR REGRESSION

Given n pairs of values for the two variates X & Y, i.e.

$X : x_1 , x_2 , \ldots\ldots , x_n$

$Y : y_1 , y_2 , \ldots\ldots , y_n ,$

it is required to find the best fitted line to the above data. The equation for such a line is not just a mathematical form where we can calculate the value of Y given the value of X and vice versa. In fact we are going to have two equations. From one equation we can predict the value of Y given the value of X with the least possible error in Y. This is called the regression line of Y on X and usually written in the form

$$Y = a + b(X - \bar{X}) .$$

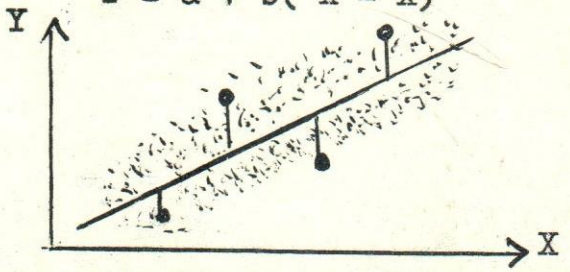
From the the other equation ,we can predict the value of X given the value of Y with the least possible error in X. This is called the regression line of X on Y and usually written in the form

$$X = a_1 + b_1(Y - \bar{Y}) .$$

1- The regression line of Y on X :

Let the required equation be in the form

$$Y = a + b(X - \bar{X}) \qquad (1)$$



If y_i denotes the observed value and $\hat{y}_i$ denotes the predicted value then the error is represented by

$$e_i = y_i - \hat{y}_i \qquad (i=1,2,\ldots,n) \qquad (2)$$

The over all error around the fitted line can be measured by

$$E = \sum e_i^2 /n = \sum (y_i - \hat{y}_i)^2 /n$$

But

$$\hat{y}_i = a + b (x_i - \bar{x})$$

then

$$E = \sum (y_i - a - b(x_i - \bar{x}))^2 /n$$

Hence E is minimum if the quantity

$$Q = \sum (y_i - a - b(x_i - \bar{x}))^2 \qquad (3)$$

is minimum.(n is a given value)
It is obvious that this quantity depends on "a" & "b". Now the question is : what are the estimates of "a" & "b" which make Q minimum ? Using the Least Square method ,the required conditions are

$$\partial Q/\partial a = o \quad , \quad \partial Q/\partial b = o \qquad (4)$$

From (3) we have

$$\left.\begin{aligned} \partial Q/\partial a &= -2 \sum (y_i - a - b(x_i - \bar{x})) \\ \partial Q/\partial b &= -2 \sum (x_i - \bar{x})(y_i - a - b(x_i - \bar{x})) \end{aligned}\right\} \qquad (5)$$

From (4) & (5) the required conditions become

$$\left.\begin{aligned} \sum (y_i - a - b(x_i - \bar{x})) &= o \\ \sum (x_i - \bar{x})(y_i - a - b(x_i - \bar{x})) &= o \end{aligned}\right\} \qquad (6)$$

Solving (6) ,we get

$$a = \sum y_i / n = \bar{y} \qquad (7)$$

&

$$b = \sum (x_i - \bar{x})(y_i - \bar{y}) / \sum (x_i - \bar{x})^2 \qquad (8)$$

Substituting in (1) we get the regression line of Y on X, and "b" is called the coefficient of regression.

Similarly it can be shown that the regression line of X on Y is

$$X = a_1 + b_1 (Y - \bar{Y}),$$

where

$$a_1 = \bar{x}$$

&

$$b_1 = \sum (x_i - \bar{x})(y_i - \bar{y}) / \sum (y_i - \bar{y})^2$$

Remarks :

1- The two rgression lines of Y on X and of X on Y intersect at the point ($\bar{x}$, $\bar{y}$)

2- The regression coefficients of Y on X and of X on Y can be rewritten as

$$b = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

and $$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum y_i^2 - (\sum y_i)^2}$$

Both formulae are preferable for numerical calculations.

3-Again the coefficients of regression can be shown to have the following forms:

$$b = r\,\frac{s_y}{s_x} \quad \text{and} \quad b_1 = r\,\frac{s_x}{s_y}$$

4- Introducing the above values for b & b_1 in the regression lines the two equations become

$$\frac{y-\bar{y}}{s_y} = r\,\frac{x-\bar{x}}{s_x}$$

and

$$\frac{x-\bar{x}}{s_x} = r\,\frac{y-\bar{y}}{s_y}$$

respectively.

5- From remark 3 ,we get

$$b\,b_1 = r^2$$

Hence b , b_1 and r have the same sign, i.e. they may be all positive or may be all negative.

6- If θ is the angle between the two regression lines then

$$\tan\theta = \frac{s_x\,s_y\,(r^2-1)}{r(s_x^2 + s_y^2)}$$

Consequently, if the two regression line coincide then r = 1 and if they are perpendicular then r = o

7- If

$$d_1 = (x-c_1)/l_1 \quad \& \quad d_2 = (y-c_2)/l_2$$

then

$$b_{y.x} = b_{d_2.d_1}\,\frac{l_2}{l_1}$$

and

$$b_{x.y} = b_{d_1.d_2}\,\frac{l_1}{l_2}$$

2- Testing the significance of "a" &"b"

Let the fitted line be of the form

$$y_i = a + b\,(x_i - \bar{x}) \tag{9}$$

where "a" &"b" are given by (7)&(8)

Let also the theoritical model be of the form

$$y_i = \alpha + \beta\,(x_i - \bar{x}) + z_i \tag{10}$$

where z_i are independent random variates with

$$\mathcal{E}(z_i) = 0 \quad \& \quad \text{Var.}(z_i) = \sigma^2 \tag{11}$$

From (9) &(10)

$$\bar{y} = a \tag{12}$$

$$\bar{y} = \alpha + \bar{z} \tag{13}$$

$$\therefore \quad a = \alpha + \bar{z}$$

Henc (a)= (14)

i.e. "a" is an unbiassed estimator for α

Also $\text{Var}(a) = \mathcal{E}(a - \mathcal{E}(a))^2 = \mathcal{E}(a-\alpha)^2$

$= \mathcal{E}(\bar{z}^2) = \sigma^2/n$ (15)

Now

$$y_i = \alpha + \beta(x_i - \bar{x}) + z_i$$

$$\bar{y} = \alpha + \bar{z}$$

$$\therefore \quad y_i - \bar{y} = \beta(x_i - \bar{x}) + z_i - \bar{z}$$

Hence

$$b = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}$$

$$= \frac{\sum (x_i-\bar{x})(\beta(x_i-\bar{x}) + z_i - \bar{z})}{\sum (x_i-\bar{x})^2}$$

$$= \beta + Z \tag{16}$$

where

$$Z = \sum (x_i-\bar{x}) z_i \,/ \sum (x_i-\bar{x})^2$$

$$= \sum c_i z_i \tag{17}$$

The c_i are constants and given as

$$c_i = (x_i-\bar{x}) / \sum (x_i-\bar{x})^2 \tag{18}$$

$$\therefore \quad \mathcal{E}(Z) = \mathcal{E}\left(\sum c_i z_i\right) = 0$$

$$\& \quad \mathrm{Var}(Z) = \sum c_i^2 \, \mathrm{Var}(z_i) = \sigma^2 \sum c_i^2$$

$$= \sigma^2 / \sum (x_i - \bar{x})^2 \qquad (19)$$

$$\therefore \quad \mathcal{E}(b) = \mathcal{E}(\beta + Z) = \beta \qquad (20)$$

$$\mathrm{Var}(b) = (b - \mathcal{E}(b))^2$$

$$= \mathcal{E}(b - \beta)^2 = \mathcal{E}(Z^2)$$

$$= \mathrm{Var}(Z) = \sigma^2 / \sum (x_i - \bar{x})^2 \qquad (21)$$

Now , to test the significance of "a" ,we test $\alpha = \alpha_0$ against $\alpha \neq \alpha_0$. The test criterion is

$$T = \frac{a - \alpha_0}{\sigma / \sqrt{n}}$$

To test the significance of "b" ,we test $\beta = 0$ against $\beta \neq 0$. The test criterion is

$$T = \frac{b}{\sigma / \sqrt{\sum (x_i - \bar{x})^2}}$$

In the above two test criteria , σ is usually unknown. An estimate for σ^2 is discussed in the next section.

3- Regression analysis table

Partition of total sum of squares i.e. $(y_i - \bar{y})^2$

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y} + \hat{y} - \bar{y})^2$$

$$= \sum (y_i - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2 + 2 \sum (y_i - \hat{y})(\hat{y} - \bar{y})$$

But $$\sum (y_i - \hat{y})(\hat{y} - \bar{y}) = \sum (y_i - \bar{y} - b(x_i - \bar{x}))\, b(x_i - \bar{x})$$

$$= b \left[\sum (y_i - \bar{y})(x_i - \bar{x}) - b \sum (x_i - \bar{x})^2 \right]$$

$$= 0$$

$$\therefore \quad \sum (y_i-\bar{y})^2 = \sum (y_i-\hat{y})^2 + \sum (\hat{y}_i-\bar{y})^2$$

In other words the total sum of squares ($\sum (y_i-\bar{y})^2$) is partitioned into
sum of squares due to regression ($\sum (\hat{y}-\bar{y})^2$) = $b^2 \sum (x_i-\bar{x})^2$
and
sum of squares of residuals ($\sum (y_i-\hat{y})^2$

<u>Expected values for sum of squares due to regression and sum of squares of residuals</u>

1- Sum of squares due to regression = $b^2 \sum (x_i-\bar{x})^2$

$$= (\beta + Z)^2 \sum (x_i-\bar{x})^2$$

Expected value of SS(Reg.) = $\mathcal{E}\left[(\beta^2 + 2\beta Z + Z^2)(x_i-\bar{x})^2\right]$

$$= \beta^2 \sum (x_i-\bar{x})^2 + \sum (x_i-\bar{x})^2 \mathcal{E}(Z^2)$$

$$= \sigma^2 + \beta^2 \sum (x_i-\bar{x})^2$$

2- Sum of squares of residuals = $\sum (y_i-\hat{y})^2$

$$= \sum (y_i-\bar{y} - b(x_i-\bar{x}))^2$$

$$= \sum \left[(z_i-\bar{z}) - (b-\beta)(x_i-\bar{x})\right]^2$$

$$= \sum (z_i-\bar{z})^2 + \sum Z^2 (x_i-\bar{x})^2 - 2\sum (z_i-\bar{z}) Z (x_i-\bar{x})$$

$$= \sum (z_i-\bar{z})^2 - \sum Z^2 (x_i-\bar{x})^2$$

Expected value of SS(Resd.) = $\mathcal{E}\left[\sum (z_i-\bar{z})^2 - \sum Z^2 (x_i-\bar{x})^2\right]$

$$= (n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2$$

In other words

$$\mathcal{E}\left\{\sum (y_i-\hat{y})^2/(n-2)\right\} = \sigma^2$$

i.e. $\sum (y_i-\hat{y})^2/(n-2)$ is an unbiassed estimator for σ^2

$$\therefore \quad \hat{\sigma}^2 = \sum (y_i-\hat{y})^2/(n-2)$$

= Mean sum of squares of residual.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean sum of Squares	Expected value of Mean-Sum of Squares
Due to regression	$b^2 \sum (x_i-\bar{x})^2$	1	$b^2 \sum (x_i-\bar{x})^2$	$\sigma^2 + \beta^2 \sum (x_i-\bar{x})^2$
Residual	$\sum (y_i-\hat{y})^2$	n- 2	$\sum (y_i-\hat{y})^2/(n-2)$	σ^2
Total	$\sum (y_i-\bar{y})^2$	n- 1		

(7)

(135)

Regression Analysis Table

To test the significance wheather =o ,we have the criterion

$$\frac{b^2\sum(x_i-\bar{x})^2}{\sum(y_i-\hat{y})^2/(n-2)} = F_{1,n-2} = t^2_{n-2}$$

Note : From the above criterion, we can deduce the distribution of the coefficient correlation r given as

$$r = \sum(x_i-\bar{x})(y_i-\bar{y})/ns_xs_y$$

Sum of squares due to linear regression $= b^2\sum(x_i-\bar{x})^2$

$= r^2\sum(y_i-\bar{y})^2$

Sum of squares of residual $= \sum(y_i-\hat{y})^2$

$= (1-r^2)\sum(y_i-\bar{y})^2$

$$\therefore \frac{b^2\sum(x_i-\bar{x})^2}{\sum(y_i-\hat{y})^2/(n-2)} = \frac{r^2(n-2)}{(1-r^2)}$$

i.e. $\frac{r^2(n-2)}{(1-r^2)} = t^2_{n-2}$ $\therefore \frac{r\sqrt{n-2}}{\sqrt{(1-r^2)}} = t_{n-2}$

$\therefore$ If $u = r\sqrt{n-2}/\sqrt{1-r^2}$ & $\rho = o$, then $u = t_{n-2}$

But $p(t_{n-2}) = c\ (1 + \frac{t^2}{n-2})^{-(n-1)/2}$

Hence $p(r) = p(t)\left|\frac{dt}{dr}\right|$

$= C\ (1 - r^2)^{(n-4)/2}$

where $C = 2^{3-n} / B\left[\frac{1}{2}(n-2), \frac{1}{2}(n-2)\right]$

EXAMPLES

1-

x	y	x	y	x	y	x	y	x	y	x	y	x	y
3	2	3	1	3	1	3	1	3	1	5	3	7	4
5	5	5	4	5	4	5	4	5	3	7	5	9	3
7	5	7	5	7	5	7	5	7	5	9	4	5	3
9	5	9	4	9	4	9	4	9	4	5	3	7	4
11	3	11	2	11	2	11	2	11	2	7	4	7	4
7	4	9	3	5	2	3	7	7	4	9	3	5	2
7	4	7	4	7	4	7	3	9	2	7	3	4	3
7	3												

$n = 50$ $\quad \Sigma x = 350$ $\quad \Sigma y = 168$

$\Sigma xy = 1202$ $\quad \Sigma x^2 = 2690$ $\quad \Sigma y^2 = 632$

$\bar{x} = 7$ $\quad \bar{y} = 3.36$

$\therefore a = \bar{y} = 3.36$ & $b = 0.108$

$\therefore y = 3.36 + 0.108 (x - 7)$

$= 2.60 + 0.11 x$

Total sum of squares $= \Sigma (y-\bar{y})^2 = \Sigma y^2 - (\Sigma y)^2/n$

$= 67.52$

Due to regression sum of squares

$= b^2 \quad (x-\bar{x})^2 = b^2 \left[\Sigma x^2 - (\Sigma x)^2/n \right.$

$= 2.82$

Regression analysis table

Source of Variation	Sum of Squares	Degrees of freedom	Mean SSQ
Due to regression	2.82	1	2.82
Residual	64.70	48	1.35
Total	67.52	49	

$R = 2.82/1.35 = 2.09$

From the tables $F_{1,48,\ 5\%} = 4.00$

$\therefore$ We accept the null hypothesis that $=0$. In other words the regression is not significant

2- Distribution of 100 labours according to their age and number of working hours

No. of working hours (Y) \ Age (X)	25-	35-	45-	Total
27-	10	10	10	20
32-	20	40		50
37-		10		30
Total	30	60	10	100

X-table

Intervals	f	d_1	fd_1	fd_1^2
25-	30	-1	-30	30
35-	60	0	0	0
45-55	10	1	10	10
Total	100		-20	40

Y-table

Intervals	f	d_2	fd_2	fd_2^2
27-	20	-1	-20	20
32-	50	0	0	0
37-42	30	1	30	30
Total	100		10	50

$N = 100 \qquad \sum fd_1d_2 = -30$

$\sum f(x-\bar{x})^2 = 100(40-4) = 3600$

$\sum f(y-\bar{y})^2 = 25(50-1) = 1225$

$\sum f(x-\bar{x})(y-\bar{y}) = 50(-30+2) = -1400$

$\therefore \quad b = -1400/3600 = -0.389$

$\bar{x} = 10(-20/100) + 40 = 38$

$\bar{y} = 5(10/100) + 34.5 = 35$

$\therefore$ The regression line is

$$y = 35 - 0.389\,(x-38)$$

Now Total sum of squares $\sum f(y-\bar{y})^2 = 1225$

Due to regression SSQ $b^2 \sum f(x-\bar{x})^2 = 544.44$

Residual sum of squares $= 680.56$

$\therefore \quad R = 544.44 / (680.56/98) = 78.4$

Comparing the value of R with that of $F_{1,98,5\%} = 4.00$, we reject the null hypothesis(=0).

In other words the regression is significant.

Exercise

The following is the bivariate distribution of 311 skulls according to their length & breadth. Carry out the required regression analysis

y \ x	Breadth in mms.										
	133-	136-	139-	142-	145-	148-	151-	154-	157-	160-	163-
160-				1	2						
165-	2		1	1		2	1				
170-			2	2	2	1		2			
175-	2	2	2	7	9	4	5	3			
180-		1	1	10	15	11	9	2	1		
185-		1	4	13	12	18	7	4	1		
190-				4	13	17	14	8	3	2	1
195-			3	1	13	14	5	6	2	1	
200-				2	6	4	4	3			1
205-				1	3	2	1	2	1	1	
210-				1	1	3	1	2			
215-							1	1			
Total	4	4	13	43	76	76	48	33	8	4	2

Exercise

The followings are the measures of the wing length and tongue length both in millimeters for 25 bees:-

Wing :	9.68	9.81	9.59	9.68	9.84	9.59	9.61	
Tongue:	6.53	6.71	6.70	6.69	6.70	6.62	6.59	
Wing :	9.55	9.25	9.08	9.70	9.60	9.50	9.74	9.72
Tongue:	6.55	6.35	6.25	6.61	6.51	6.55	6.74	6.75
Wing :	9.64	9.73	9.77	9.72	9.54	9.65	9.74	9.59
Tongue:	6.45	6.75	6.70	6.65	6.68	6.77	6.44	6.54
Wing :	9.71	9.56						
Tongue:	6.64	6.55						

Carry out the regression analysis required.

(XII) Multiple Regression

(1). Introduction: This is an extension for simple linear regression, where we have more than two characters. We are going to deal with the Case where we can express one of our variables (dependent) in terms of the other variables (independent). Suppose X_o is the dependent variable & $X_1, X_2, \dots, X_k$ are the independent variables. Let X_o take the values x_{oi}, $X_1, X_2, \dots, X_k$ take the values $x_{1i}, x_{2i}, \dots, x_{ki}$ and $i = 1,2, \dots, n$.

The data can be arranged in the following form

X_o	X_1	X_2	. . . .	X_k
x_{o1}	x_{11}	x_{21}	. . . .	x_{k1}
x_{o2}	x_{12}	x_{22}	. . . .	x_{k2}
.	.	.	. . . .	.
.	.	.	. . . .	.
.	.	.	. . . .	.
x_{on}	x_{1n}	x_{2n}	. . . .	x_{kn}

The expected value of x_{oi} given the values of $x_{1i}, x_{2i}, \dots, k_{ki}$ is

$$x_{oi} = a + b_1(x_{1i}-\bar{x}_1) + \dots + b_r(x_{ri}-\bar{x}_r) + \dots + b_k(x_{ki}-\bar{x}_k) \quad (1)$$

The mathematical model is

$$x_{oi} = \alpha + \beta_1(x_{1i}-\bar{x}_1) + \dots + \beta_r(x_{ri}-\bar{x}_r) + \dots + \beta_k(x_{ki}-\bar{x}_k) + \gamma_i \quad (2)$$

We are going to treat these independent variables as constants Moreover, we assume that the γ_i are independent random variables, $E(\gamma_i) = 0$ & $\text{Var}(\gamma_i) = \sigma^2$. The β's are noticed to be the partial regression Coefficients i.e. $\beta_{or.12\dots(r)\dots k}$ The question is to derive estimators for β_r and test the mull hypothesis $\beta_r = 0$ against the alternative $\beta_r \neq 0$.

(2-1) Let us try first as an example the simplest case where we have two independent variates. The mathematical model is

$$x_{oi} = \alpha + \beta_1 (x_{1i}-\bar{x}_1) + \beta_2(x_{2i}-\bar{x}_2) + \gamma_i \tag{3}$$

Let the best fitted line be of the form

$$x_{oi} = a + b_1 (x_{1i}-\bar{x}_1) + b_2 (x_{2i}-\bar{x}_2) \tag{4}$$

Using least square method, we minimize

$$S = \sum_{i=1}^{n} \left[x_{oi}-a-b_1(x_{1i}-\bar{x}_1) - b_2(x_{2i}-\bar{x}_2) \right]^2 \tag{5}$$

To get a, b_1, b_2 we should have

$$\left.\begin{aligned} \frac{\partial S}{\partial a} &= 0 \\ \frac{\partial S}{\partial b_1} &= 0 \\ \frac{\partial S}{\partial b_2} &= 0 \end{aligned}\right\} \tag{6}$$

In other words

$$a = \bar{x}_o \tag{7}$$

$$b_1 \sum_i (x_{1i}-\bar{x}_1)^2 + b_2 \sum_i (x_{2i}-\bar{x}_2)(x_{1i}-\bar{x}_1) = \sum_i (x_{oi}-x_o)(x_{1i}-\bar{x}_1) \tag{8}$$

$$\& \quad b_1 \sum_i (x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2) + b_2 \sum_i (x_{2i}-\bar{x}_2)^2 = \sum_i (x_{oi}-\bar{x}_o)(x_{2i}-\bar{x}_2) \tag{9}$$

If we write $\lambda_{rs} = \sum_{i=1}^{n} (x_{ri}-\bar{x}_r)(x_{si}-\bar{x}_s)$, then (8),(9) become

$$b_1 \lambda_{11} + b_2 \lambda_{21} = \lambda_{o1} \tag{10}$$

$$b_1 \lambda_{12} + b_2 \lambda_{22} = \lambda_{o2} \tag{11}$$

Solving (10), (11) we get

$$b_1 = \frac{\lambda_{o1}\lambda_{22} - \lambda_{o2}\lambda_{12}}{\lambda_{11}\lambda_{22} - \lambda_{12}^2} \tag{12}$$

$$b_2 = \frac{\lambda_{o2}\,\lambda_{11} - \lambda_{o1}\,\lambda_{21}}{\lambda_{11}\,\lambda_{22} - \lambda_{12}^2} \tag{13}$$

(2-2) We want to express b_1, b_2 in terms of our theortical model In other words, what are λ_{o1}, λ_{o2} in terms of the ζ's ?

$$x_{oi} = \alpha + \beta_1 (x_{1i} - \bar{x}_1) + \beta_2 (x_{2i} - \bar{x}_2) + \zeta_i \tag{14}$$

$$\bar{x}_o = \alpha + \bar{\zeta} \tag{15}$$

$$\therefore\ x_{oi} - \bar{x}_o = \beta_1 (x_{1i} - \bar{x}_1) + \beta_2 (x_{2i} - \bar{x}_2) + \zeta_i - \bar{\zeta} \tag{16}$$

Now

$$\lambda_{o1} = \sum_i (x_{oi} - \bar{x}_o)(x_{1i} - \bar{x}_1)$$

$$= \sum_{i=1}^{n} \left[\beta_1 (x_{1i} - \bar{x}_1) + \beta_2 (x_{2i} - \bar{x}_2) + (\zeta_i - \bar{\zeta}) \right] (x_{1i} - \bar{x}_1)$$

$$= \beta_1 \lambda_{11} + \beta_2 \lambda_{21} + \lambda_{\zeta 1} \tag{17}$$

Similarly

$$\lambda_{o2} = \sum_i (x_{oi} - \bar{x}_o)(x_{2i} - \bar{x}_2) = \beta_1 \lambda_{12} + \beta_2 \lambda_{22} + \lambda_{\zeta 2} \tag{18}$$

But $\lambda_{\zeta 1} = \sum_i (\zeta_i - \bar{\zeta})(x_{1i} - \bar{x}_1) = \sum_i (x_{1i} - \bar{x}_1)\,\zeta_i$

$$\therefore\ E(\lambda_{\zeta 1}) = 0 \tag{19}$$

Also $\lambda_{\zeta 2} = \sum_i (\zeta_i - \bar{\zeta})(x_{2i} - \bar{x}_2) = \sum_i (x_{2i} - \bar{x}_2)\,\zeta_i$

$$\therefore\ E(\lambda_{\zeta 2}) = 0 \tag{20}$$

Again $\mathcal{E}(\lambda^2_{z1}) = \mathcal{E}\left[\sum_i z_i^2 (x_{1i}-\bar{x}_1)^2\right]$

$= \sum_i (x_{1i}-\bar{x}_1)^2 = \lambda_{11}\sigma^2$ (21)

Similarly $\mathcal{E}(\lambda^2_{z2}) = \lambda_{22}\sigma^2$ (22)

& $\mathcal{E}(\lambda_{z1}\lambda_{z2}) = \lambda_{12}\sigma^2$ (23)

But $$b_1 = \frac{\lambda_{01}\lambda_{22} - \lambda_{02}\lambda_{12}}{\lambda_{11}\lambda_{22} - \lambda_{12}^2}$$

$$= \beta_1 + \frac{\lambda_{22}\lambda_{z1} - \lambda_{12}\lambda_{z2}}{\lambda_{11}\lambda_{22} - \lambda_{12}^2} = \beta_1 + Z_1$$

Similarly $$b_2 = \beta_2 + \frac{\lambda_{11}\lambda_{z2} - \lambda_{12}\lambda_{z1}}{\lambda_{11}\lambda_{22} - \lambda_{12}^2} = \beta_2 + Z_2$$

$\mathcal{E}(b_1) = \beta_1$

& $\mathcal{E}(b_2) = \beta_2$

Also $\text{Var}(b_1) = \mathcal{E}(b_1-\beta_1)^2 = \mathcal{E}(Z_1^2) = \frac{\lambda_{22}}{\lambda_{11}\lambda_{22}-\lambda_{12}^2}\sigma^2$ (24)

& $\text{Var}(b_2) = \mathcal{E}(b_2-\beta_2)^2 = \mathcal{E}(Z_2^2) = \frac{\lambda_{11}}{\lambda_{11}\lambda_{22}-\lambda_{12}^2}\sigma^2$ (25)

and $$\rho_{(b_1,b_2)} = \frac{\mathcal{E}[(b_1-\beta_1)(b_2-\beta_2)]}{\sigma_{b_1}\sigma_{b_2}} = -\frac{\lambda_{12}}{\lambda_{11}\lambda_{22}-\lambda_{12}^2}\sigma^2$$

$$= -\rho(x_1, x_2) \quad (26)$$

(2-3) Partition of Total Sum of Squares

Total sum of squares $= \sum_i (x_{oi}-\bar{x}_o)^2 = \lambda_{oo}$

$$= \sum_{i=1}^{n}\left[\left\{x_{oi}-\bar{x}_o-b_1(x_{1i}-\bar{x}_1)-b_2(x_{2i}-\bar{x}_2)\right\}+\left\{b_1(x_{1i}-\bar{x}_1)+b_2(x_{2i}-\bar{x}_2)\right\}\right]^2$$

$$= \sum_{i=1}^{n}\left\{x_{oi}-\bar{x}_o-b_1(x_{1i}-\bar{x}_1)-b_2(x_{2i}-\bar{x}_2)\right\}^2$$

$$+ \sum_{i=1}^{n}\left\{b_1(x_{1i}-\bar{x}_1)+b_2(x_{2i}-\bar{x}_2)\right\}^2$$

$$+ 2\sum_{i=1}^{n}\left\{x_{oi}-\bar{x}_o-b_1(x_{1i}-\bar{x}_1)-b_2(x_{2i}-\bar{x}_2)\right\}\left\{b_1(x_{1i}-\bar{x}_1)+b_2(x_{2i}-\bar{x}_2)\right\}$$

The 1st term is about multiple regression sum of squares (residual sum of squares) i.e.

$$\sum_{i=1}^{n}\left\{x_{io}-\bar{x}_o-b_1(x_{1i}-\bar{x}_1)-b_2(x_{2i}-\bar{x}_2)\right\}^2$$

$$=\lambda_{oo}+b_1^2\lambda_{11}+b_2^2\lambda_{22}-2b_1\lambda_{o1}-2b_2\lambda_{o2}+2b_1b_2\lambda_{12}$$

$$=\lambda_{oo}+b_1(b_1\lambda_{11}+b_2\lambda_{12}-\lambda_{o1})+b_2(b_1\lambda_{12}+b_2\lambda_{22}-\lambda_{o2})$$

$$-\lambda_{o1}b_1-\lambda_{o2}b_2$$

$$=\lambda_{oo}-b_1\lambda_{o1}-b_2\lambda_{o2}$$

The 2nd term is due to regression sum of squares i.e.

$$\sum_{i=1}^{n}\left\{b_1(x_{1i}-\bar{x}_1)+b_2(x_{2i}-\bar{x}_2)\right\}^2$$

$$=b_1^2\lambda_{11}+b_2^2\lambda_{22}+2b_1b_2\lambda_{12}$$

$$=b_1(b_1\lambda_{11}+b_2\lambda_{12})+b_2(b_2\lambda_{22}+b_1\lambda_{12})$$

$$=b_1\lambda_{o1}-b_2\lambda_{o2}$$

The last term i.e.

$$2\sum_{i=1}^{n}\left\{x_{oi}-\bar{x}_o-b_1(x_{1i}-\bar{x}_1)-b_2(x_{2i}-\bar{x}_2)\right\}\left\{b_1(x_{1i}-\bar{x}_1)+b_2(x_{2i}-\bar{x}_2)\right\}$$

$$=2(b_1\lambda_{o1}-b_1^2\lambda_{11}-b_1b_2\lambda_{12}+b_2\lambda_{o2}-b_1b_2\lambda_{12}-b_2^2\lambda_{22})$$

$$=2\,b_1(\lambda_{o1}-b_1\lambda_{11}-b_2\lambda_{12})+b_2(\lambda_{o2}-b_1\lambda_{12}-b_2\lambda_{22})$$

$$=0$$

In other words the last term vanishes and we are left with

Total sum of squares = Residual sum of squares + Due to regression sum of squares.

i.e.

$$\lambda_{oo}=(\lambda_{oo}-b_1\lambda_{o1}-b_2\lambda_{o2})+(b_1\lambda_{o1}+b_2\lambda_{o2})$$

(2-4) The Expected Values

$$\because \quad x_{oi}-\bar{x}_o = \beta_1(x_{1i}-\bar{x}_1)+ \beta_2(x_{2i}-\bar{x}_2)+ z_i-\bar{z}$$

$$\therefore \quad \lambda_{oo}= \sum_{i=1}^{n} (x_{oi}-\bar{x}_o)^2$$

$$= \beta_1^2\lambda_{11}+ \beta_2^2\lambda_{22}+ \lambda_{zz}+2\beta_1\lambda_{31}+2\beta_2\lambda_{32}+ 2\beta_1\beta_2\lambda_{12}$$

$$\therefore \mathcal{E}(\lambda_{oo}) = \beta_1^2\lambda_{11}+\beta_2^2\lambda_{22}+2\beta_1\beta_2\lambda_{12}+(n-1)\sigma^2$$

(i) Due to regression sum of squares $=b_1\lambda_{o1}+b_2\lambda_{02}$

$$= (\lambda_{22}\lambda_{o1}^2- 2\lambda_{12}\lambda_{o1}\lambda_{o2}+\lambda_{11}\lambda_{o2}^2)/(\lambda_{11}\lambda_{22}-\lambda_{12}^2)$$

But $\mathcal{E}(\lambda_{o1}^2) = \mathcal{E}(\beta_1\lambda_{11}+\beta_2\lambda_{21}+\lambda_{z1})^2$

$$= (\beta_1\lambda_{11}+\beta_2\lambda_{21})^2+\lambda_{11}\sigma^2$$

Also $\mathcal{E}(\lambda_{o2}^2) = (\beta_1\lambda_{21}+\beta_2\lambda_{22})^2+\lambda_{22}\sigma^2$

& $\mathcal{E}(\lambda_{o1}\lambda_{o2}) = (\beta_1\lambda_{11}+\beta_2\lambda_{21})(\beta_1\lambda_{21}+\beta_2\lambda_{22})+\lambda_{12}\sigma^2$

$\therefore$ Expected value of due to regression sum of squares

$$= \mathcal{E}(b_1\lambda_{o1}+b_2\lambda_{o2}) = \beta_1^2\lambda_{11}+\beta_2^2\lambda_{22}+2\beta_1\beta_2\lambda_{12}+2\sigma^2$$

(ii) Residual sum of squares $= \lambda_{oo}-(b_1\lambda_{o1}+b_2\lambda_{o2})$

$\therefore$ Expected value of Residual sum of squares $= (n-3)\sigma^2$

Regression analysis table

Source of Variation	Sum of Squares	Degrees of freedom	Expected value of mean sum of squares
Due to multiple regression	$b_1\lambda_{01}+b_2\lambda_{02}$	2	$\sigma^2+\frac{1}{2}(\beta_1^2\lambda_{11}+\beta_2^2\ 22+2\beta_1\beta_2\lambda_{12})$
Residual	$\lambda_{00}-b_1\lambda_{01}-b_2\lambda_{02}$	n-3	σ^2
Total	λ_{00}	n-1	

From the above table, the mean sum of squares of residual is an unbiased estimator for σ^2 i.e.

$$\tilde{\sigma}^2 = ms_{resid.}$$

(2-5) Tests for the significance of b_1, b_2

Again the normal equations (10),(11) can be written in a matrix form

$$\begin{bmatrix} \lambda_{11} & \lambda_{21} \\ \lambda_{12} & \lambda_{22} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \lambda_{01} \\ \lambda_{02} \end{bmatrix} \qquad (27)$$

$$\therefore \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{21} \\ \lambda_{12} & \lambda_{22} \end{bmatrix}^{-1} \begin{bmatrix} \lambda_{01} \\ \lambda_{02} \end{bmatrix}$$

$$= \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \begin{bmatrix} \lambda_{01} \\ \lambda_{02} \end{bmatrix}$$

where $c_{11} = \dfrac{\lambda_{22}}{\Delta}$.

$$c_{22} = \lambda_{11}/\Delta$$

$$c_{12} = c_{21} = -\lambda_{12}/\Delta$$

& $\Delta = \lambda_{11}\lambda_{22} - \lambda_{12}^2$

Hence equations (24), (25), (26) can be rewritten as

$$\text{Var}(b_1) = C_{11}\sigma^2$$
$$\text{Var}(b_2) = C_{22}\sigma^2 \qquad (28)$$

Now to test the null hypothesis $\beta_1 = 0$ against the alternative $\beta_1 \neq 0$, we apply the test criterion

$$t_{b_1} = b_1 / \tilde{\sigma}_{b_1}$$

where $\tilde{\sigma}_{b_1} = C_{11}\, ms_{resd.}$

Similarly to test the null hypothesis $\beta_2 = 0$ against the alternative $\beta_2 \neq 0$ we apply

$$t_{b2} = b_2 / \tilde{\sigma}_{b2}$$

where $\tilde{\sigma}_{b2} = C_{22}\, ms_{resd.}$

To test for the significance of the difference between b_1, b_2 we have

$$\tilde{\sigma}^2_{b_1-b_2} = (C_{11}+C_{22} - 2C_{12})\sigma^2$$
$$= (C_{11}+C_{22} - 2C_{12})\, ms_{resd.}$$
$$\therefore \quad t = (b_1-b_2)/\tilde{\sigma}_{b_1-b_2}$$

The no. of degrees of freedom for the statistic "t" is n-3, the same as that for the residual sum of squares.

(3) The above procedure can be generalized to more than two independent variables.

Let

$$x_{oi} = a + b_1(x_{1i}-\bar{x}_1) + \ldots + b_r(x_{ri}-\bar{x}_r) + \ldots + b_k(x_{ki}-\bar{x}_k) \quad (1)$$

where $i = 1,2,\ldots,n$ denotes the number of observations
and $r = 1,2,\ldots,k$ denotes the number of independent variables.

The mathematical model is

$$x_{oi} = \alpha + \beta_1(x_{1i}-\bar{x}_1) + \ldots + \beta_r(x_{ri}-\bar{x}_r) + \ldots + \beta_k(x_{ki}-\bar{x}_k) + \zeta_i \quad (2)$$

where $\mathcal{E}(\zeta_i) = 0$ & Var $(\zeta_i) = \sigma^2$

Minimizing

$$S = \sum_{i=1}^{n} \left[x_{oi} - a - b_1(x_{1i}-\bar{x}_1) - b_2(x_{2i}-\bar{x}_2) - \ldots - b_k(x_{ki}-\bar{x}_k) \right]^2$$

we get

$$a = \bar{x}_o \quad (3)$$

$$b_1\lambda_{11} + b_2\lambda_{21} + \ldots + b_r\lambda_{r1} + \ldots + b_k\lambda_{k1} = \lambda_{o1}$$

$$b_1\lambda_{12} + b_2\lambda_{22} + \ldots + b_r\lambda_{r2} + \ldots + b_k\lambda_{k2} = \lambda_{o2}$$

$$\vdots \qquad \vdots \qquad \cdots \qquad \vdots \qquad \cdots \qquad \vdots \qquad \vdots$$

$$b_1\lambda_{1k} + b_2\lambda_{2k} + \ldots \quad b_r\lambda_{rk} \quad \ldots \quad b_k\lambda_{kk} = \lambda_{ok}$$

Where λ_{rs} is as defined before

Now the above set of equations (4) can be rewritten in a matrix form as

$$\begin{bmatrix} \lambda_{11} & \lambda_{21} & \cdots & \lambda_{r1} & \cdots & \lambda_{k1} \\ \lambda_{12} & \lambda_{22} & \cdots & \lambda_{r2} & \cdots & \lambda_{k2} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \lambda_{1k} & \lambda_{2k} & \cdots & \lambda_{rk} & \cdots & \lambda_{kk} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} \lambda_{o1} \\ \lambda_{o2} \\ \vdots \\ \lambda_{ok} \end{bmatrix} \quad (5)$$

This is simply written

$$[\lambda](b) = (\lambda_o) \quad (6)$$

where $[\lambda]$ is a square matrix, $(b), (\lambda_o)$ are column matrices.

$$\therefore (b) = [\lambda]^{-1}(\lambda_o) \quad (7)$$

From (2) $\bar{x}_o = \alpha + \bar{\zeta}$

$$\therefore x_{oi} - \bar{x}_o = \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \ldots + \beta_k(x_{ki} - \bar{x}_k) + (\zeta_i - \bar{\zeta})$$

Multiplying both sides by $x_{1i} - \bar{x}_1, x_{2i} - \bar{x}_2, \ldots, (\bar{x}_{ki} - \bar{x}_k)$

successively and summing over i in each case, we get

$$\begin{aligned} \lambda_{o1} &= \beta_1\lambda_{11} + \beta_2\lambda_{21} + \ldots\ldots + \beta_k\lambda_{k1} + \lambda_{\zeta 1} \\ \lambda_{o2} &= \beta_1\lambda_{12} + \beta_2\lambda_{22} + \ldots\ldots + \beta_k\lambda_{k2} + \lambda_{\zeta 2} \\ &\vdots \\ \lambda_{ok} &= \beta_1\lambda_{1k} + \beta_2\lambda_{2k} + \ldots\ldots + \beta_k\lambda_{kk} + \lambda_{\zeta k} \end{aligned} \quad (8)$$

These set of equations can be written in a matrix form

$$\begin{bmatrix} \lambda_{o1} \\ \lambda_{o2} \\ \vdots \\ \lambda_{ok} \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{21} & \cdots & \lambda_{k1} \\ \lambda_{12} & \lambda_{22} & \cdots & \lambda_{k2} \\ \vdots & \vdots & \cdots & \vdots \\ \lambda_{1k} & \lambda_{2k} & \cdots & \lambda_{kk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \lambda_{z1} \\ \lambda_{z2} \\ \vdots \\ \lambda_{zk} \end{bmatrix} \qquad (9)$$

This is simply written as

$$(\lambda_o) = [\lambda](\beta) + (\lambda_z) \qquad (10)$$

where $$E(\lambda_z) = 0 \qquad (11)$$

Substituting from (10) in (7)

$$(b) = [\lambda]^{-1}\{[\lambda](\beta) + \lambda_z\}$$
$$= (\beta) + [\lambda]^{-1}(\lambda_z) \qquad (12)$$

Hence $$E(b_r) = \beta_r \qquad (13)$$

From (12)

$$(b - \beta) = [\lambda]^{-1}(\lambda_z)$$
$$\therefore (b - \beta)' = (\lambda_z)'[\lambda]^{-1} \qquad (14)$$

$$(b - \beta)(b - \beta)' = [\lambda]^{-1}\lambda_z\lambda_z'[\lambda]^{-1} \qquad (15)$$

But

$$\lambda_z\lambda_z' = \begin{pmatrix} \lambda_{z1} \\ \lambda_{z2} \\ \vdots \\ \lambda_{zk} \end{pmatrix} (\lambda_{z1} \quad \lambda_{z2} \quad \cdots \quad \lambda_{zk})$$

$$= \begin{bmatrix} \lambda_{z1}^2 & \lambda_{z1}\lambda_{z2} & \cdots & \lambda_{z1}\lambda_{zk} \\ \lambda_{z2}\lambda_{z1} & \lambda_{z2}^2 & \cdots & \lambda_{z2}\lambda_{zk} \\ \vdots & \vdots & \cdots & \vdots \\ \lambda_{zk}\lambda_{z1} & \lambda_{zk}\lambda_{z2} & \cdots & \lambda_{zk}^2 \end{bmatrix} \qquad (16)$$

Since $\lambda_{zr} = \sum_{i=1}^{n} (z_i - \bar{z})(x_{ri} - \bar{x}_r)$

$$= \sum_{i=1}^{n} (x_{ri} - \bar{x}_r) z_i$$

$$\therefore \mathcal{E}(\lambda_{zr}^2) = \lambda_{rr}\sigma^2 \qquad (17)$$

Also $\mathcal{E}(\lambda_{zr}\lambda_{zs}) = \lambda_{rs}\sigma^2 \qquad (18)$

Consequently

$$\mathcal{E}(\lambda_z \lambda_z') = \begin{bmatrix} \lambda_{11}\sigma^2 & \lambda_{12}\sigma^2 & \cdots & \lambda_{1k}\sigma^2 \\ \lambda_{21}\sigma^2 & \lambda_{22}\sigma^2 & \cdots & \lambda_{2k}\sigma^2 \\ \vdots & \vdots & \cdots & \vdots \\ \lambda_{k1}\sigma^2 & \lambda_{k2}\sigma^2 & \cdots & \lambda_{kk}\sigma^2 \end{bmatrix} = [\lambda]\sigma^2 \qquad (19)$$

$$\mathcal{E}(b - \beta)(b - \beta)' = \mathcal{E}[\lambda^{-1}\lambda_z\lambda_z'\lambda^{-1}]$$

$$= \lambda^{-1}\lambda\sigma^2\lambda^{-1} = [\lambda]^{-1}\sigma^2 \qquad (20)$$

But

$$(b-\beta)(b-\beta)' = \begin{pmatrix} b_1-\beta_1 \\ b_2-\beta_2 \\ \vdots \\ b_k-\beta_k \end{pmatrix} (b_1-\beta_1 \quad b_2-\beta_2 \cdots b_k-\beta_k)$$

$$= \begin{bmatrix} (b_1-\beta_1)^2 & (b_1-\beta_1)(b_2-\beta_2) & \cdots & (b_1-\beta_1)(b_k-\beta_k) \\ (b_2-\beta_2)(b_1-\beta_1) & (b_2-\beta_2)^2 & \cdots & (b_2-\beta_2)(b_k-\beta_k) \\ \vdots & \vdots & \cdots & \vdots \\ (b_k-\beta_k)(b_1-\beta_1) & (b_k-\beta_k)(b_2-\beta_2) & \cdots & (b_k-\beta_k)^2 \end{bmatrix} \quad (21)$$

If $[\lambda]^{-1}$ is written as

$$[\lambda]^{-1} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & c_{22} & \cdots & c_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ c_{k1} & c_{k2} & \cdots & c_{kk} \end{bmatrix} \quad (22)$$

then, substituting in (20) and comparing with the expected values of (21), we get

$$\mathcal{E}\,(b_r-\beta_r)^2 = c_{rr}\,\sigma^2$$

i.e. $$\text{Var}\,(b_r) = c_{rr}\,\sigma^2 \quad (23)$$

Also $$\mathcal{E}\,(b_r - \beta_r)(b_s - \beta_s) = C_{rs}\,\sigma^2 \qquad (24)$$

Similarly

$$\text{Var}\,(b_r - b_s) = \mathcal{E}\left[b_r - b_s - (\beta_r - \beta_s)\right]^2$$

$$= \mathcal{E}\left[(b_r - \beta_r) - (b_s - \beta_s)\right]^2$$

$$= \mathcal{E}\,(b_r - \beta_r)^2 + \mathcal{E}(b_s - \beta_s)^2 - 2\mathcal{E}(b_r - \beta_r)(b_s - \beta_s)$$

$$= (C_{rr} + C_{ss} - 2\,C_{rs})\,\sigma^2 \qquad (25)$$

The regression analysis table

Source of variation	Sum of squares	Degrees of feedom
Due to multiple Regression	$b_1\lambda_{o1} + b_2\lambda_{o2} + \ldots + b_k\lambda_{ok}$	k
Residual	$\lambda_{oo} - (b_1\lambda_{o1} + b_2\lambda_{o2} + \ldots + b_k\lambda_{ok})$	n - k - 1
Total	oo	n-1

Following the same procedure as in the case of two independent variates, it can be shown that

$$\mathcal{E}\ (\text{Mean sum of Squares of residuals}) = \sigma^2$$

$$\text{i.e. } \tilde{\sigma}^2 = MS_{\text{resid.}} = \frac{\text{Sum of Squares of residuals}}{n - k - 1} \qquad (26)$$

To test the null hipothesis $\beta_r = 0$ against the alternative $\beta_r \neq 0$, we apply the test criterion

$$t = \frac{b_r - \beta_r}{\tilde{\sigma}_{b_r}} = b_r \,/\, \tilde{\sigma}_{b_r} \qquad (27)$$

where $\tilde{\sigma}_{b_r} = \sqrt{C_{rr} \; MS_{resid}}$ (28)

and with $n - k - 1$ degrees of freedom

Also to test the null hypothesis $\beta_r - \beta_s = 0$ against the alternative $\beta_r - \beta_s \neq 0$ we apply the test criterion.

$$t = \frac{b_r - b_s}{\tilde{\sigma}_{b_r - b_s}} \tag{29}$$

where $\tilde{\sigma}_{b_r - b_s} = \sqrt{(C_{rr} + C_{ss} - 2C_{rs}) \; MS_{resid}}$ (30)

Example (1) Consider the following data

x_0	:	12	1	2	4	12	13	3	1
x_1	:	15	1	6	23	26	27	20	2
x_2	:	6	11	9	6	4	2	7	11

$$\sum x_0 = 48 \quad , \sum x_1 = 120, \quad \sum x_2 = 56$$

Sum of Squares and Sum of products in deviate form

	x_0	x_1	x_2
x_0	200	290	-105
x_1		800	-227
x_2			72

$$\Delta = \begin{vmatrix} 800 & -227 \\ -227 & 72 \end{vmatrix} = 6071$$

$$C_{11} = 72/6071 = 0.01186$$
$$C_{22} = 800/6071 = 0.13177$$
$$C_{12} = 227/6071 = 0.03739$$

$$b_1 = C_{11}\ o1 + C_{12}\ o2 = -0.4866$$
$$b_2 = C_{21}\ o1 + C_{22}\ o2 = -2.9927$$

Sum of squares due to multiple regression

$$= b_1 \lambda_{o1} + b_2 \lambda_{o2} = 173.1195$$

Total Sum of squares = 200.

Regression analysis table

Source of variation	Sum of Squares	Degree of freedom	Mean Sum of Squares
Due to multiple regression	173.1195	2	86.5597
Residual	26.8805	5	5.2761
Total	200.0000	7	

$R = 86.5597 / 5.2761 = 16.4$

$F_{2,5,10/0} = 13.27$

The regression is significant

$\text{Var}(b_1) = C_{11} MS_{resid} = 0.0625$

$\text{Var}(b_2) = C_{22} MS_{resid} = 0.6952$

$\tilde{\sigma}_{b_1} = 0.25 \quad \& \tilde{\sigma}_{b_2} = 0.83$

Hence $t_{b_1} = 1.96$

$t_{b_2} = 3.61$

But $t_{5,\ 5\%} = 2.571$

There is an evidence that $\beta_1 = 0$ while $\beta_2 \neq 0$.

Example (2) The following data represent: the score on the final exam (x_0), number of daily study hours (x_1), intelligent quotient (x_2) and the average score through the year (x_3) for a sample of 10 students.

x_0	: 7	8	5	11	17	18	13	11	9	10
x_1	: 5	6	4	7	8	9	7	6	5	4
x_2	: 10	10	11	14	15	13	10	12	14	11
x_3	: 7	10	6	15	16	18	15	10	7	12

$x_0 = 109$, $x_1 = 61$, $x_2 = 120$, $x_3 = 116$

Sum of squares and sum of products in derivate form

	x_0	x_1	x_2	x_3
x_0	154.9	55.1	38.0	143.6
x_1		24.9	13.0	54.4
x_2			32.0	28.0
x_3				162.4

To get the inverse of the matrix formed from the squares and products of x_1, x_2, x_3 we use the Square root method as follows

x_1	x_2	x_3	I		
24.9	13.0	54.4	1	0	0
	32.0	28.0	0	1	0
		162.4	0	0	1
4.98999	2.60522	10.90183	0.20040	0	0
	5.02124	-0.07999	-0.10398	0.19915	0
		6.59877	-0.33234	0.00241	0.15154
			0.16142	-0.02151	-0.05036
		$[\lambda]^{-1}$ =		0.03966	0.00036
					0.02296

$[\lambda]$=

where $[\lambda] = \begin{bmatrix} 24.9 & 13.0 & 54.4 \\ & 32.0 & 28.0 \\ & & 162.4 \end{bmatrix}$

& $[\lambda]^{-1} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} = \begin{bmatrix} 0.16142 & -0.02151 & -0.05036 \\ & 0.03966 & 0.00036 \\ & & 0.02296 \end{bmatrix}$

$b_1 = c_{11}\lambda_{o1} + c_{12}\lambda_{o2} + c_{13}\lambda_{o3} = 0.8451$

$b_2 = c_{21}\lambda_{o1} + c_{22}\lambda_{o2} + c_{23}\lambda_{o3} = 0.3736$

$b_3 = c_{31}\lambda_{o1} + c_{32}\lambda_{o2} + c_{33}\lambda_{o3} = 0.5359$

Sum of Squares due to regression

$$= b_1 \lambda_{o1} + b_2 \lambda_{o2} + b_3 \lambda_{o3} = 137.7197$$

& Total sum of Squares $\lambda_{oo} = 154.9$

Regression analysis table

Source of variation	Sum of Squares	Degrees of freedom	Mean sum of squares
Due to regression	137.7197	3	45.9066
Residual	17.1803	6	2.8634
	154.9000	9	

$R = 45.9066 / 2.8634 = 16.03$

$F_{3,6,1\%} = 9.78$

The regression is Significant

$Var(b_1) = C_{11} MS_{resid} = 0.4622$

$Var(b_2) = C_{22} MS_{resid} = 0.1136$

$Var(b_3) = C_{33} MS_{resid} = 0.0657$

$\tilde{\sigma}_{b_1} = 0.680$, $\tilde{\sigma}_{b_2} = 0.337$, $\tilde{\sigma}_{b_3} = 0.256$

$t_1 = 1.24$, $t_2 = 1.11$, $t_3 = 2.09$

But $t_{6,5\%} = 2.447$

There is an eridence that each of $\beta_1, \beta_2, \beta_3$ is equal to zero.

Exercise In a problem to study the factors which may be used in estimating cotton and cotton seeds production index, the data collected are given in the following table

Y	D	A_r	A_s	L	F_N	F_p
1945	636	504	525	1016	234	65
46	734	585	655	1014	189	1
47	769	627	683	1009	483	177
48	1074	796	805	1002	677	214
49	1051	852	931	999	765	258
50	1928	1154	1116	999	924	703
51	976	1108	1133	1001	999	1003
52	1200	1115	1117	998	997	1264
53	857	759	745	999	1011	1037
54	938	864	889	1003	993	755
55	900	962	992	1010	943	1047
56	875	891	926	1017	840	1254
57	1091	990	1042	1027	934	1427
58	1200	1142	1093	1032	1184	1522
59	1229	1017	996	1040	993	1475
60	1285	1065	1035	1048	981	1941

where Y the years

P Cotton & Cotton seeds production index

A_r Area cultivated with cotton weighted with the relative product of 1 feddan in upper, middle and lower Egypt.

A_s Area cultivated with cotton weighted with the relative product of 1 feddan cultivated with long, long medium and medium Staple cotton

L Labour index

F_N, F_p nitrate and phosphate index

Make the required regression analysis.

Exercise The following represents data on pigs. Thus animals were kept on adequate vations during the period of approximately uniform growth rate. We wish to inquire the amount of information about rate of gain which is furnished in advance by the two independent variates

Pig No.	Initial age x_1 (days)	weight x_2 (pounds)	Rate of gain Y (pounds/day)
1	78	61	1.40
2	90	59	1.79
3	94	76	1.72
4	71	50	1.47
5	99	61	1.26
6	80	54	1.28
7	83	57	1.34
8	75	45	1.57
9	62	41	1.57
10	67	40	1.26
11	78	74	1.61
12	99	75	1.31
13	80	64	1.12
14	75	48	1.35
15	94	62	1.29
16	91	42	1.24
17	75	52	1.29
18	63	43	1.43
19	62	50	1.29
20	67	40	1.26
21	78	80	1.67
22	83	61	1.41
23	79	62	1.73
24	70	47	1.23
25	85	59	1.49
26	83	42	1.22
27	71	47	1.39

Pig No.	Initial age x_1 (days)	weight x_2 (pounds)	Rate of gain Y (pounds/day)
28	66	42	1.39
29	67	40	1.56
30	67	40	1.36
31	77	62	1.40
32	71	55	1.47
33	78	62	1.37
34	70	43	1.15
35	95	57	1.22
36	96	51	1.48
37	71	41	1.31
38	63	40	1.27
39	62	45	1.22
40	67	39	1.36

Exercises

(I) The mean coefficient of volume of expension of water was determined by 21 different observers at 3 different temperatures. It is known that the determination of this coefficient will become more difficult as the temperature increases owing to the formation of bubbles in the water. Fine out from the data given below whether the determinations of the coefficient are significantly more variable as the temp. range of the water increases. Estimate a mean coefficient for each temp. and give confidence limits for the true coefficient

Temp. observer	Coefficient of Volume Expansion of Water (10^6)			Temp Obser-ver			
	0-40	40-60	60-80		0-40	40-60	60-80
1	239	522	517	11	249	386	612
2	274	474	584	12	301	404	437
3	295	512	606	13	175	504	561
4	256	417	447	14	260	361	580
5	354	327	542	15	301	491	285
6	213	361	627	16	445	449	453
7	264	504	614	17	335	436	520
8	236	572	632	18	166	346	372
9	302	390	505	19	270	449	566
10	269	526	667	20	254	560	607
				21	343	468	560

2- Five independent samples of 25 individuals each were drawn from a normal population of unknown mean and variance. The 98% confidence intervals for the population mean were obtained as follows from each sample

Sample No.	Confidence limits
1	105.7 - 116.4
2	104.3 - 113.8
3	107.2 - 116.0
4	100.6 - 115.4
5	101.7 - 118.1

Obtain a 98% confidence interval for the population standard deviation

3- Using the symbols given in the table below, derive expressions for the probabilities that of four men aged exactly 75,80,85 & 90 respectively

(i) all will attain age 95
(ii) all will die before attaining age 95
(iii) at least one will survive 10 years
(iv) none will die between age 90 & 95.

4- Ruther ford and Geiger counted the number of alpha - particles emitted from a disc in 2608 periods of 7.5 seconds duration. The frequencies are given below

Number per period	frequency	Number per period	frequency
0	57	8	45
1	203	9	27
2	383	10	10
3	525	11	4
4	532	12	2
5	408	13	0
6	273	14	0
7	139		

Compare the relative frequencies with the corresponding probabilities of the fitted "Posson distribution"

5- Show that the sum of two Poisson variates is itself a Poisson variate with mean equal to the sum of the seperate means

6- Number of individual incomes in different ranges of net income assessed in 1945-46:

Range of Income after tax (x)	Number of Incomes
150 - 500	13,175,000
500 - 1000	652,000
1000 - 2000	137,500
2000 and over	
Total	140,000,000

@ Exact age

	75	80	85	90	95
Probability of surviving 5 years	p_0	p_1	p_2	p_3	p_4

Assume that this distribution of incomes f(x) is linked with the normal distribution

$$N(t) = \frac{1}{2} \quad e^{-\frac{1}{2} t^2}$$

by the relationship

$$\int_{-\infty}^{t} N(t)\, dt = \int_{150}^{x} f(x)\, dx$$

where $t = a \log (x - 150) + b$

Obtain estimates for "a" & "b" from the data, and find the number of incomes between 250 and 500.

7- The following results were obtained in 4 independent samplings

(1)	6	14	12	6	2	5
(2)	10	17	6	19	19	16
(3)	11	11	19	23	8	17
(4)	19	2	29	16	14	20

Carryout an analysis of variance on these data

8- Twelve dice were thrown 26306 times and a "5" or a "6" was counted as a success. The number of successes in each throw was noted, with the following results

Number of successes	frequency	Number of successes	frequency
0	185	6	3067
1	1149	7	1331
2	3265	8	403
3	5475	9	105
4	6114	10	18
5	5194		
		Total	26306

Is there evidence that the dice are biased?

9- The following table gives the distribution of the length, measured in cm. of 294 eggs

Length	frequency	length	
3.5	1	4.2	54
3.6	1	4.3	34
3.7	6	4.4	12
3.8	20	4.5	6
3.9	35	4.6	1
4.0	53	4.7	2
4.1	69		
		Total	294

Test whether these results are consistent with the hypothesis that egg length is normally distributed

10- A certain type of surgical operation can be performed either with a local anaasthetic or with a general anoesthetic. Results are given below

	alive	Dead
Local	511	24
General	173	21

Test for any difference in the mortabity rates associated with the different types of anoesthetic.

11- A Ministry of Labour Memorandum on carbon Monoxide Poisoning gives the following data on accidents due to gassing by carbon monoxide

	1941	1942	1943	Total
At blast furnaces	24	20	19	63
At gas producers	28	34	41	103
At gas ovens and works	26	26	10	62
In distribution and use gas	80	108	123	311
Miscellaneous sources	68	51	32	151
Total	226	239	225	690

Is there significant association between the site of the accident and the year.

(12) From the following data find the regression equation of x_1 on x_2 & x_3

x_1	x_2	x_3
5	2	21
3	4	21
2	2	15
4	2	17
3	3	20
1	2	13
8	4	32

(13) If s_1^2 , s_2^2 are the variances in two independent samples of the same size taken from a common normal population, determine the distribution of $s_1^2 + s_2^2$

(14) A random variate **x** is known to have the distribution

$$p(x) = C\left(1+\frac{x}{a}\right)^{m-1} e^{-\frac{mx}{a}} \qquad -a \leq x \leq \infty$$

Find the constant C and the 1st four moments of x. **Prove that**

$$8\beta_2 - 9\beta_1 = 4$$

(15) If $u = ax + by$ and $v = bx - ay$, where x,y represent deviations from respective means and if ρ is the correlation coefficient between x,y, but u, v are uncorrelated, show that

$$\sigma_u \sigma_v = (a^2+b^2)\sigma_x \sigma_v \sqrt{1-\rho^2}$$

(16) The following table gives the distribution according to age of becoming a widewer

Age	freq.	Age	freq.
18 - 22	5	53 - 57	88
23 - 27	43	58 - 62	76
28 - 32	108	63 - 67	82
33 - 37	107	68 - 72	59
38 - 42	130	73 - 77	41
43 - 47	109	78 - 82	16
48 - 52	115	83 - 87	8

Calculate β_1 & β_2

(17) Examine the following data showing how age is associated with size of farm

size of farm	Percentage of occupiers whose ages are						
	under 25	25-34	35-44	45-54	55-64	65-74	over 74
	%	%	%	%	%	%	%
20-	2	12	25	28	23	8	2
50-	1	7	21	36	26	7	2
100-	1	10	29	28	24	6	2
150-	1	9	21	36	21	10	2
over 300	0	5	19	39	22	11	4

(18) Four machines A,B,C,D, are producing large numbers of small articles. It is known that the average proportions of defective articles produced by the machines are

A = 1%, B = 1½ %, C = 1½ %, D = 2½ %

In a group of 20 articles known to have been produced by the same machine, one defective article is found. What is the probability that this group was produced by machine D.

State carefully any assumption which you make.

(19) $x_1, x_2, \ldots, x_n$ are random variables each with the same expected value ξ, and the same variance σ^2. The correlation between any two of the x's is . Show that

(i) $\left\{\sigma(\bar{x})\right\}^2 = \frac{\sigma^2}{n} + (1 - \frac{1}{n})\rho\sigma^2$

(ii) $\mathcal{E}\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right] = (n-1)(1-\rho)\sigma^2$

(iii) $\rho > -\frac{1}{n-1}$

(20) Suppose that the probability of death within a year for a person aged exactly x is q_x. If there are n_x persons aged exactly x, what is the probability that r of these persons will die before reaching exact age (x+1) ?

(21) Assuming that in 1000 items ,20 are defective,what is the probability that in a random sample of 100 items x items will be defective.

(22) Adie has n+1 faces ,numbered o,1/n,2/n,;...;(n-1)/n,1 respectively.Assuming it is equally likely to fall with any one face uppermost,find the expected value and standard deviation of the random variable corresponding to the number on the uppermost face.

(23) For a high voltage network,uniform cables of great tensile strength are required.Each cable is composed of wires which are manufactured in one length.In order to examine the tensile strength a sample is taken from each wire and tested.The table shows the tensile strength of each wire of 5 cables each with 10 wires

Cable 1:	345	327	335	338	330	334	335	340	337	342
Cable 2:	329	327	332	348	337	328	328	330	335	334
Cable 3:	340	330	325	328	338	332	335	340	336	339
Cable 4:	328	344	342	350	335	332	328	340	335	337
Cable 5:	347	341	345	340	350	346	345	342	340	339

Carry out the analysis of variance(subtract an arbitrary origin)

(24) The following table gives the number of yeast cells in 400 squares of a haemacytometer:

No. of cells	0	1	2	3	4	5	Total
Frequency	213	128	37	18	3	1	400

Fit a negative binomial

(Hint: mean=0.68, variance=0.81. np=0.68, npq=0.81, then q=1.19 & p=-0.19 and n=-3.59)

(25) Fit a normal distribution to the distribution given in the following table and test for the goodness of fit.

Height	≤59	60-	61-	62-	63-	64-	65-	66-	67-	68-
Freq.	23	169	439	1030	2116	3947	5965	8012	9089	8763

Height	69-	70-	71-	72-	73-	74-	75-	76-	77&over
Freq.	7132	5314	3320	1884	876	383	153	63	25

REFERENCES

1- David,F.N. Probability theory for statistical methods

2- Edwards,Allen L. Experimental design in Psychological research

3- Fisher,R.A. Statistical methods for research workers

4- Gouldin,Cyril H. Methods of statistical analysis

5- Hald,A. Statistical theory with engineering applications

6- Johnson,N.L. & Tetley,H. "Statistics" An intermediate text book Vol. 1 & 11

7- Kendall,M.G. & Stuart,A. The advanced theory of statistics Vol. 1 & 11

8- Lindquist,E.F. Design and analysis of experiments in Psychology & Education.

9- Snedecor,George W. "Statistical methods"

10- Rosander,A.C. Elementary principles of statistics

11- Walker,Halen M. & Lev ,Joef "Statistical inference"

12- Wilks,S.S. Mathematical statistics

Remark

The mareial in this note is still in a tentative form.Corrections,criticisms and expressions of other points of view on the teaching and presentation of the course in this note will be gratefully received.

Moharram W.Mahmoud