Chapter 6

The Chi – Square test

Definition :

* Results obtained in samples denote always agree exactly with theoretical results, expected according to rules of probability.

- Suppose that a set of possible events, E_1 , E_2 , … , E_K occure with observed frequencies, O_1 , O_2 , … , O_K, and according to prob. rules the expected frequencies e_1 , e_2 , …, e_K.

- Chi – square measure the discrepancy existing between observed and expected frequencies.

$$\chi^2 = \frac{(O_1 - e_1)^2}{e_1} + \frac{(O_2 - e_2)^2}{e_2} + \ldots + \frac{(O_k - e_k)^2}{e_k}$$

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - e_j)^2}{e_j} \quad \ldots\ldots (1)$$

If the total freq. is N :

$$\sum O_j = \sum e_j = N \quad \ldots\ldots (2)$$

$$\therefore \quad \chi^2 = \sum_{j=1}^{k} \left(\frac{O_j^2}{e_j}\right) - N \quad \ldots\ldots (3)$$

Where K : number of possible events.

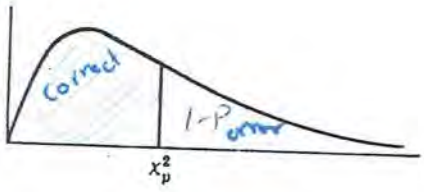
$$\chi^2 = \sum \left(\frac{(O-e)^2}{e}\right)$$

$$= \sum \left(\frac{O^2 - 2Oe + e^2}{e}\right) = \sum \left(\frac{O^2}{e} - \frac{2Oe}{e} + \frac{e^2}{e}\right)$$

Appendix IV

صدول جديد للقيم الحرجة Chi-sq χ^2_p

PERCENTILE VALUES (χ^2_p)
for
THE CHI-SQUARE DISTRIBUTION
with ν degrees of freedom
(shaded area = p)



ν	$\chi^2_{0.995}$	$\chi^2_{0.99}$	$\chi^2_{0.975}$	$\chi^2_{0.95}$	$\chi^2_{0.90}$	$\chi^2_{0.75}$	$\chi^2_{0.50}$	$\chi^2_{0.25}$	$\chi^2_{0.10}$	$\chi^2_{0.05}$	$\chi^2_{0.025}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$
1	7·88	6·63	5·02	3·84	2·71	1·32	0·455	0·102	0·0158	0·0039	0·0010	0·0002	0·0000
2	10·6	9·21	7·38	5·99	4·61	2·77	1·39	0·575	0·211	0·103	0·0506	0·0201	0·0100
3	12·8	11·3	9·35	7·81	6·25	4·11	2·37	1·21	0·584	0·352	0·216	0·115	0·072
4	14·9	13·3	11·1	9·49	7·78	5·39	3·36	1·92	1·06	0·711	0·484	0·297	0·207
5	16·7	15·1	12·8	11·1	9·24	6·63	4·35	2·67	1·61	1·15	0·831	0·554	0·412
6	18·5	16·8	14·4	12·6	10·6	7·84	5·35	3·45	2·20	1·64	1·24	0·872	0·676
7	20·3	18·5	16·0	14·1	12·0	9·04	6·35	4·25	2·83	2·17	1·69	1·24	0·989
8	22·0	20·1	17·5	15·5	13·4	10·2	7·34	5·07	3·49	2·73	2·18	1·65	1·34
9	23·6	21·7	19·0	16·9	14·7	11·4	8·34	5·90	4·17	3·33	2·70	2·09	1·73
10	25·2	23·2	20·5	18·3	16·0	12·5	9·34	6·74	4·87	3·94	3·25	2·56	2·16
11	26·8	24·7	21·9	19·7	17·3	13·7	10·3	7·58	5·58	4·57	3·82	3·05	2·60
12	28·3	26·2	23·3	21·0	18·5	14·8	11·3	8·44	6·30	5·23	4·40	3·57	3·07
13	29·8	27·7	24·7	22·4	19·8	16·0	12·3	9·30	7·04	5·89	5·01	4·11	3·57
14	31·3	29·1	26·1	23·7	21·1	17·1	13·3	10·2	7·79	6·57	5·63	4·66	4·07
15	32·8	30·6	27·5	25·0	22·3	18·2	14·3	11·0	8·55	7·26	6·26	5·23	4·60
16	34·3	32·0	28·8	26·3	23·5	19·4	15·3	11·9	9·31	7·96	6·91	5·81	5·14
17	35·7	33·4	30·2	27·6	24·8	20·5	16·3	12·8	10·1	8·67	7·56	6·41	5·70
18	37·2	34·8	31·5	28·9	26·0	21·6	17·3	13·7	10·9	9·39	8·23	7·01	6·26
19	38·6	36·2	32·9	30·1	27·2	22·7	18·3	14·6	11·7	10·1	8·91	7·63	6·84
20	40·0	37·6	34·2	31·4	28·4	23·8	19·3	15·5	12·4	10·9	9·59	8·26	7·43
21	41·4	38·9	35·5	32·7	29·6	24·9	20·3	16·3	13·2	11·6	10·3	8·90	8·03
22	42·8	40·3	36·8	33·9	30·8	26·0	21·3	17·2	14·0	12·3	11·0	9·54	8·64
23	44·2	41·6	38·1	35·2	32·0	27·1	22·3	18·1	14·8	13·1	11·7	10·2	9·26
24	45·6	43·0	39·4	36·4	33·2	28·2	23·3	19·0	15·7	13·8	12·4	10·9	9·89
25	46·9	44·3	40·6	37·7	34·4	29·3	24·3	19·9	16·5	14·6	13·1	11·5	10·5
26	48·3	45·6	41·9	38·9	35·6	30·4	25·3	20·8	17·3	15·4	13·8	12·2	11·2
27	49·6	47·0	43·2	40·1	36·7	31·5	26·3	21·7	18·1	16·2	14·6	12·9	11·8
28	51·0	48·3	44·5	41·3	37·9	32·6	27·3	22·7	18·9	16·9	15·3	13·6	12·5
29	52·3	49·6	45·7	42·6	39·1	33·7	28·3	23·6	19·8	17·7	16·0	14·3	13·1
30	53·7	50·9	47·0	43·8	40·3	34·8	29·3	24·5	20·6	18·5	16·8	15·0	13·8
40	66·8	63·7	59·3	55·8	51·8	45·6	39·3	33·7	29·1	26·5	24·4	22·2	20·7
50	79·5	76·2	71·4	67·5	63·2	56·3	49·3	42·9	37·7	34·8	32·4	29·7	28·0
60	92·0	88·4	83·3	79·1	74·4	67·0	59·3	52·3	46·5	43·2	40·5	37·5	35·5
70	104·2	100·4	95·0	90·5	85·5	77·6	69·3	61·7	55·3	51·7	48·8	45·4	43·3
80	116·3	112·3	106·6	101·9	96·6	88·1	79·3	71·1	64·3	60·4	57·2	53·5	51·2
90	128·3	124·1	118·1	113·1	107·6	98·6	89·3	80·6	73·3	69·1	65·6	61·8	59·2
100	140·2	135·8	129·6	124·3	118·5	109·1	99·3	90·1	82·4	77·9	74·2	70·1	67·3

Source: Catherine M. Thompson, *Table of percentage points of the χ^2 distribution*, Biometrika, Vol. 32 (1941), by permission of the author and publisher.

If $\sum O = \sum e = N$

$$\chi^2 = \sum_{j=1}^{k} \left(\frac{O_j^2}{e_j} \right) - N$$

- If $\chi^2 = 0$, observed and expected freq. agree exactly.
- If $\chi^2 > 0$, Do not agree exactly.
- The greater $\chi^2 \rightarrow$ the greater the discrepancy.

Sampling distribution of χ^2 :

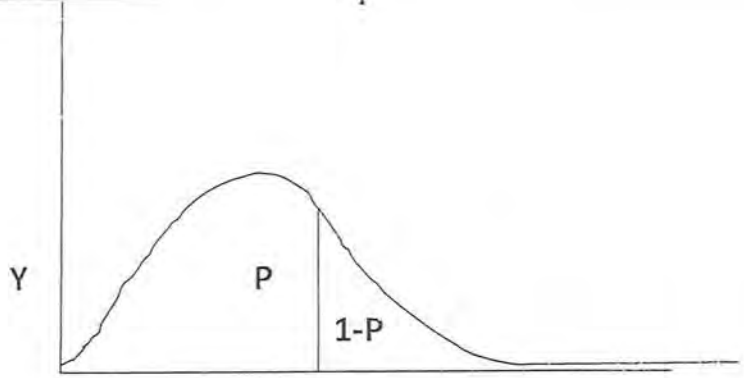
$$Y = Y_0 (\chi)^{\nu-2} e^{-\frac{1}{2}\chi^2}$$

Where the number of degree of freedom (ν) is given by :

a. $\nu = k - 1$ (for hypothesises, where the expected freq. can be computed with out having to estimate pop. parameters from sample statistics). And K: no. of possible events.

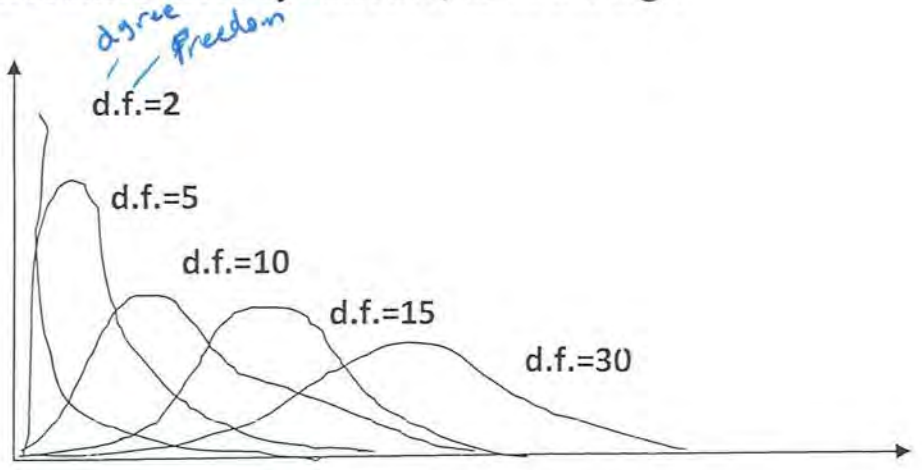
b. $\nu = k - 1 - m$ (if the expected freq. can be computed only by estimating (m) parameters).
m : no. of pop. parameters required to calculate the expected freq.

* The critical values of χ_p^2 from the following table :



Characteristic of the chi-square distribution :

1. It is not symmetric.
2. The shape of χ^2 depends on the degree of freedom.
3. As the no. of degree of freedom increases the χ^2 dist. Becomes more symmetric, as in the fig :



4. The values of χ^2 are nonnegative. That is the values of χ^2 are greater than or equal to 0.

Significance tests using χ^2 :

1. Expected freq. are computed on the basis of hypothesis, or theoretical distributions.

2. Determine the χ^2.

3. From table determine χ_p^2 (the critical values)

$\chi_{0.95}^2 \Rightarrow 0.05$ significance level

$\chi_{0.99}^2 \Rightarrow 0.01$ significance level

$\chi_p^2 \Rightarrow 1-p$ significance level

4. Compare the calculated χ^2 with the critical values χ_p^2 .

- If $\chi^2 > \chi_p^2 \rightarrow$ hypo. Or theo. Prob. dist. Is rejected at (1-p) sig. level.

Where :

P : prob. of being correct.

1-p : prob. of being error (sig. level)

5. If χ^2 is too small (too close to zero).

$\chi^2 > \chi_{0.95}^2 \rightarrow$ suspicious data.

Test if $\chi^2 < \chi_{0.95}^2$ or $\chi_{0.01}^2$, if so can not depend on data. .

Example 1 : (test of hypothesis)

In 200 tosses of a coin, 115 heads, 85 tails, were observed. Test fairness of coin at a sig. level of

a) 0.05 b) 0.01

Soln. : the observed freq. of heads, and tails

O_1=115 , O_2=85

Expected freq. of heads and tails if the coin is fair are

e_1=100 , e_2=100

$$\therefore \quad \chi^2 = \frac{(115-100)^2}{100} + \frac{(85-100)^2}{100} = 4.5$$

ν=k-1 = 2-1 =1

$$\therefore \quad \left.\begin{array}{l} at \ \ \nu = 1 \rightarrow \chi_{0.95}^2 = 3.84 \\ at \ \ \nu = 1 \rightarrow \chi_{0.99}^2 = 6.63 \end{array}\right\} \ from\ table$$

$\therefore$ a) $\chi^2 > \chi_{0.95}^2 \rightarrow$ reject hypothesis

b) $\chi^2 \leq \chi^2_{0.99} \rightarrow$ accept hypothesis

Example 2 : (test of hypothesis)

In 120 throws of a die the following data were observed :

events :	1	2	3	4	5	6
obs. freq.	25	17	15	23	24	16

Test fairness of die at a sig. level of 0.05.

Soln. :

If the die is fair, then the expected freq. are :

$$e_1 = e_2 = e_3 = \cdots = e_6 = \frac{\sum O_f}{6} = \frac{120}{6} = 20$$

$$\therefore \quad \chi^2 = \sum \frac{(O_f - e_f)^2}{e_f} = \frac{(25-20)^2}{20} + \frac{(17-20)^2}{20} + \cdots$$

$$\therefore \quad \chi^2 = 5.0$$

ν=k-1 = 6-1 =5

$at \quad \nu = 5 \rightarrow \chi^2_{0.95} = 11.1$ (from table)

$\therefore \quad \chi^2 < \chi^2_{0.95} \rightarrow$ accept the hypothesis but test of data is required.

$at \quad \nu = 5 \rightarrow \chi^2_{0.05} = 1.15$ (from table)

$\chi^2 > \chi^2_{0.95}$, $\therefore$ data is acceptable, and the die is fair at 0.05 sig. level

Example 3 : (test of hypothesis)

A survey of 320 families with 5 children revealed the dist. below; is the result consistent with the hypothesis that the male and female births are equally probable:

No. of boys & girls	5 boys 0 girls	4 1	3 2	2 3	1 4	0 5	Total
No. of families	18	56	110	88	40	8	320
Exp.	10	50	100	100	50	10	320

Soln. :

If $p = q = \frac{1}{2}$, then :

$$pr\binom{5\ boys}{0\ girls} = \frac{5!}{5!\,(5-5)!}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^0 = \frac{1}{32}$$

$$pr\binom{4}{1} = \frac{5}{32} \quad , \quad pr\binom{3}{2} = \frac{10}{32} \quad , \quad pr\binom{2}{3} = \frac{10}{32}$$

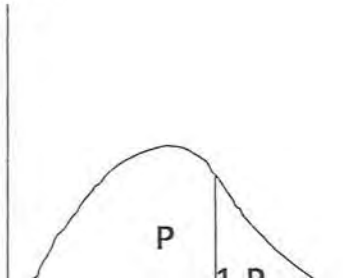
$$pr\binom{1}{4} = \frac{5}{32} \quad , \quad pr\binom{0}{5} = \frac{1}{32}$$

Then the expected no. of families : 10,50,100,100,50,10, hence

$$\chi^2 = \sum \frac{(O-e)^2}{e} = 12.0$$

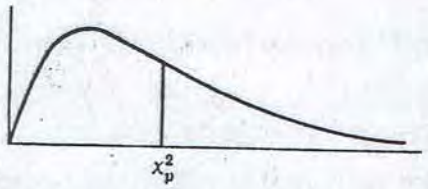
Since $\chi^2_{0.95} = 11.1$ at ν=k-1 = 6-1 =5

We can reject the hypo. at 0.05 significance level, (We conclude that male and female births are not equally).



Appendix IV

PERCENTILE VALUES (χ_p^2)
for
THE CHI-SQUARE DISTRIBUTION
with ν degrees of freedom
(shaded area = p)



ν	$\chi^2_{0.995}$	$\chi^2_{0.99}$	$\chi^2_{0.975}$	$\chi^2_{0.95}$	$\chi^2_{0.90}$	$\chi^2_{0.75}$	$\chi^2_{0.50}$	$\chi^2_{0.25}$	$\chi^2_{0.10}$	$\chi^2_{0.05}$	$\chi^2_{0.025}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$
1	7·88	6·63	5·02	3·84	2·71	1·32	0·455	0·102	0·0158	0·0039	0·0010	0·0002	0·0000
2	10·6	9·21	7·38	5·99	4·61	2·77	1·39	0·575	0·211	0·103	0·0506	0·0201	0·0100
3	12·8	11·3	9·35	7·81	6·25	4·11	2·37	1·21	0·584	0·352	0·216	0·115	0·072
4	14·9	13·3	11·1	9·49	7·78	5·39	3·36	1·92	1·06	0·711	0·484	0·297	0·207
5	16·7	15·1	12·8	11·1	9·24	6·63	4·35	2·67	1·61	1·15	0·831	0·554	0·412
6	18·5	16·8	14·4	12·6	10·6	7·84	5·35	3·45	2·20	1·64	1·24	0·872	0·676
7	20·3	18·5	16·0	14·1	12·0	9·04	6·35	4·25	2·83	2·17	1·69	1·24	0·989
8	22·0	20·1	17·5	15·5	13·4	10·2	7·34	5·07	3·49	2·73	2·18	1·65	1·34
9	23·6	21·7	19·0	16·9	14·7	11·4	8·34	5·90	4·17	3·33	2·70	2·09	1·73
10	25·2	23·2	20·5	18·3	16·0	12·5	9·34	6·74	4·87	3·94	3·25	2·56	2·16
11	26·8	24·7	21·9	19·7	17·3	13·7	10·3	7·58	5·58	4·57	3·82	3·05	2·60
12	28·3	26·2	23·3	21·0	18·5	14·8	11·3	8·44	6·30	5·23	4·40	3·57	3·07
13	29·8	27·7	24·7	22·4	19·8	16·0	12·3	9·30	7·04	5·89	5·01	4·11	3·57
14	31·3	29·1	26·1	23·7	21·1	17·1	13·3	10·2	7·79	6·57	5·63	4·66	4·07
15	32·8	30·6	27·5	25·0	22·3	18·2	14·3	11·0	8·55	7·26	6·26	5·23	4·60
16	34·3	32·0	28·8	26·3	23·5	19·4	15·3	11·9	9·31	7·96	6·91	5·81	5·14
17	35·7	33·4	30·2	27·6	24·8	20·5	16·3	12·8	10·1	8·67	7·56	6·41	5·70
18	37·2	34·8	31·5	28·9	26·0	21·6	17·3	13·7	10·9	9·39	8·23	7·01	6·26
19	38·6	36·2	32·9	30·1	27·2	22·7	18·3	14·6	11·7	10·1	8·91	7·63	6·84
20	40·0	37·6	34·2	31·4	28·4	23·8	19·3	15·5	12·4	10·9	9·59	8·26	7·43
21	41·4	38·9	35·5	32·7	29·6	24·9	20·3	16·3	13·2	11·6	10·3	8·90	8·03
22	42·8	40·3	36·8	33·9	30·8	26·0	21·3	17·2	14·0	12·3	11·0	9·54	8·64
23	44·2	41·6	38·1	35·2	32·0	27·1	22·3	18·1	14·8	13·1	11·7	10·2	9·26
24	45·6	43·0	39·4	36·4	33·2	28·2	23·3	19·0	15·7	13·8	12·4	10·9	9·89
25	46·9	44·3	40·6	37·7	34·4	29·3	24·3	19·9	16·5	14·6	13·1	11·5	10·5
26	48·3	45·6	41·9	38·9	35·6	30·4	25·3	20·8	17·3	15·4	13·8	12·2	11·2
27	49·6	47·0	43·2	40·1	36·7	31·5	26·3	21·7	18·1	16·2	14·6	12·9	11·8
28	51·0	48·3	44·5	41·3	37·9	32·6	27·3	22·7	18·9	16·9	15·3	13·6	12·5
29	52·3	49·6	45·7	42·6	39·1	33·7	28·3	23·6	19·8	17·7	16·0	14·3	13·1
30	53·7	50·9	47·0	43·8	40·3	34·8	29·3	24·5	20·6	18·5	16·8	15·0	13·8
40	66·8	63·7	59·3	55·8	51·8	45·6	39·3	33·7	29·1	26·5	24·4	22·2	20·7
50	79·5	76·2	71·4	67·5	63·2	56·3	49·3	42·9	37·7	34·8	32·4	29·7	28·0
60	92·0	88·4	83·3	79·1	74·4	67·0	59·3	52·3	46·5	43·2	40·5	37·5	35·5
70	104·2	100·4	95·0	90·5	85·5	77·6	69·3	61·7	55·3	51·7	48·8	45·4	43·3
80	116·3	112·3	106·6	101·9	96·6	88·1	79·3	71·1	64·3	60·4	57·2	53·5	51·2
90	128·3	124·1	118·1	113·1	107·6	98·6	89·3	80·6	73·3	69·1	65·6	61·8	59·2
100	140·2	135·8	129·6	124·3	118·5	109·1	99·3	90·1	82·4	77·9	74·2	70·1	67·3

Source: Catherine M. Thompson, *Table of percentage points of the χ^2 distribution*, Biometrika, Vol. 32 (1941), by permission of the author and publisher.

χ^2 test for goodness of fit :

χ^2 test can be used to determine how well theoretical distributions, such as (normal, binomial, poisson), fit distributions which obtained from sample data.

Example 3 : (test goodness of fit of Binomial dist.)

(5) pennies were tossed 1000 times, and at each toss the no. of heads was observed. Determine the goodness of fit of binomial dist. Of the following data at a sig. level of 0.05

No. of heads:	0	1	2	3	4	5
O_f :	38	144	342	287	164	25

$\sum O_f = 1000$

Soln. :

The expected freq. is obtained from the binomial dist. :

$$p(x) = \frac{N!}{X!\,(N-X)!} p^X q^{N-X}$$

p : is obtained as fellows :

the true mean $\mu = \frac{\sum f_i x_i}{\sum f_i} = \frac{0*38+1*144+2*342+\cdots}{1000} = 2.47$

$\mu_{true} = \mu_{binomial} = N * P = 2.47 = 5 * p \rightarrow p = 0.494$

p+q=1 → q = 0.506

∴ The binomial dist. Eqn. is given by :

$$p(x) = \underset{5}{C}\, x\,(0.494)^x (0.506)^{5-x}$$

Or

$$p(x) = \frac{5!}{x!\,(5-x)!}(0.494)^x (0.506)^{5-x}$$

$$\sum O_f = \sum e_f = N = 1000$$

No. of heads x :	0	1	2	3	4	5
Pr(x) :	0.0332	0.1619	0.13162	0.3087	0.1507	0.0294
e_f :	33.2	161.9	316.2	308.7	150.7	29.4
O_f :	38	144	342	287	164	25

$$\chi^2 = \sum \frac{(O_f - e_f)^2}{e_f} = 7.54$$

ν=k-1-m = 6-1-1 =4

$at \quad \nu = 4 \rightarrow \chi^2_{0.95} = 9.49$ (from table)

$\chi^2 < \chi^2_{0.95} \rightarrow$ thefit is good.

$at \quad \nu = 4 \rightarrow \chi^2_{0.05} = 0.711$ (from table)

$\chi^2 > \chi^2_{0.05} \rightarrow$ can depend on data.

Example 4 : (test goodness of fit of normal dist.)

The distribution of masses with $\mu = 67.45$ kg and $= 2.92$, were observed as fellows :

Mass (class limits)	Observed freq. O_f
60-62	5
63-65	18
66-68	42
69-71	27
72-74	8
	$\sum O_f = 100$

Determine the goodness of fit of normal dist. At a sig. level of 0.05.

Soln. :

The class boundaries converted to standard scores $(Z)^s$, then the area of each class is obtained as fraction between Z_1 and Z_2 :

Boundaries	Standard scroses Z_1 Z_2	Fraction
59.5 – 62.5	-2.72 to -1.7	0.0413
62.5 – 65.5	-1.7 to -0.67	0.2078 68
65.5 – 68.5	-0.67 to 0.36	0.3892
68.5 – 71.5	0.36 1.39	0.2771
71.5 – 74.5	1.39 2.41	0.0743

Then each fraction is converted to e_f by multiplying each fraction by $\sum O_f$

Event :	1	2	3	4	5
O_f :	5	8	42	27	8
e_f :	4.13	20.68	38.92	27.71	7.43

$$\left(\sum O_f = \sum e_f = N\right)$$

$$\therefore \quad \chi^2 = \sum \frac{(O_f - e_f)^2}{e_f} = 0.959$$

ν=k-1-m =5-1-2 =2

$$\therefore \quad at \quad \nu = 2 \rightarrow \chi^2_{0.95} = 5.99 \quad \text{(from table)}$$

$\chi^2 < \chi^2_{0.95} \rightarrow$ the fit is good at a sig. level of 0.05.

$$at \quad \nu = 2 \rightarrow \chi^2_{0.05} = 0.103 \quad \text{(from table)}$$

$\chi^2 > \chi^2_{0.05} \rightarrow$ we can depend on data.

معلوب مل

Tutorial sheet (6)

The Chi – square test

Q. 1) The number of book borrowed from a public library during a particular week is given below, Test the hypothesis that the number of books borrowed does not depend on the day of the week, using a significance level of

a) 0.05 , b) 0.01

	Mon.	Tues.	Wed.	Thur.	Fri.
Number of book borrowed	135	108	120	114	146

Q. 2) Two hundred bolts were selected at random from the production of each (4) machines. The numbers of defective bolts found were 2 , 9 , 10 , 3 . determine whether there is no a significant difference between the machines using a significance level of 0.05.

Q. 3) Determine the goodness of fit of a binomial distribution to the following data, using a significance level of 0.05. Is the fit :too good" :

X :	0	1	2	3	4
f :	30	62	46	10	2

Where X : is no. of heads in tossing 4 coins 150 throws.

Q. 4) Determine the goodness of fit of normal distribution to the following data, using a significance level of 0.05. Is the fit is "too good".

Class limit	f
93-97	2
98-102	5

103-107	12
108-112	17
113-117	14
118-122	6
123-127	3
128-132	1
	Total = 60

Q. 5) Determine the goodness of fit of poisson distribution to the following data, using a significance level of 0.05.

X :	0	1	2	3	4
f :	109	65	22	3	1

Where X : no. of heads in tossing 4 coins 200 throws :

Chi – Square test for independence :

To test the association between (or independence) two variables in a table, we use the steps that follow :

1. Determine the null and alternative hypotheses .

Null hypo. = H_0 : The row variable and column variable are independent.

Alter. Hypo. = H_1 : The row and column variables are dependent

2. Determine the critical value at a specified level of significance (1-p).

χ^2_p calculated from the table of χ^2_p at a degree of freedom $\nu = (r-1)(c-1)$

Where : r : no. of rows

c : no. of columns

3. calculate the expected frequency for each cell in the table

4. compute the test statistic (χ^2).

5. compare the critical value to the test statistic.

If $\chi^2 > \chi^2_p$ →reject the null hypothesis

Example : (test the independence)

The following table contains observed frequency for two variables. X and Y.

	X_1	X_2	X_3	total
y_1	87	74	34	195
y_2	12	32	18	62
total	99	106	52	257

a) compute the value of χ^2

b) Test the hypothesis that X and Y are independent at the 0.05 of significance level (null hypothesis).

soln. : (R Total)(C Total) / Table Total

1. $expected\ freq. = \frac{(row\ total)(column\ total)}{table\ total}$

	X_1	X_2	X_3
y_1	75.12	80.43	39.46
y_2	23.88	25.57	12.54

For y_1 :

$$x_1 = \frac{195 * 99}{257} = 75.12$$

$$x_2 = \frac{195 * 106}{257} = 80.43$$

$$x_3 = \frac{195 * 52}{257} = 39.46$$

For y_2 :

$$x_1 = \frac{62 * 99}{257} = 23.88$$

$$x_2 = \frac{62 * 106}{257} = 25.57$$

$$x_3 = \frac{62 * 52}{257} = 12.54$$

2. calculate χ^2

O	e	$(O-e)^2/e$
87	75.12	1.878
74	80.43	0.514
34	39.46	0.755
12	23.88	5.91
32	25.57	1.62
18	12.54	2.377

$\chi^2 = 13.054$

Compute $\chi^2_{0.95}$ at $\nu = (r-1)(c-1) = (2-1)(3-1) = 2$

$\chi^2_{0.95} = 0.103$ at $\nu = 2$

5.99

4. compare χ^2 with $\chi^2_{0.95}$

$\chi^2 > \chi^2_p$ reject the null hypothesis H_0

(H_0) → the row and column variables are dependent. (H_1)

Ex :

	X_1	X_2	X_3
y_1	34	43	52
y_2	18	21	17

Test the null hy pothesis (H_0) at the significance level of 0.05

Chapter (7)

Curve fitting and method of Least – squares

* Relation ship between variables :

Very often in practice a relationship is found to exist between two (or more) variables. For example circumferences of circles depend on their radii; and the pressure of a given mass of gas depends on its temp. and volume.

It is frequently desirable to express this relationship in mathematical form by determining an equation connecting the variables.

Curve fitting procedure :

1. Plot set of data points (x,y).
2. Suggest a form of relation defining y=f(x)
   From : a. Theoretical considerations.
   b. observation of the trend of data points.
3. Evalute constants in the suggested function, so that the deviations of data points from the function are minimized.
4. Calculate statistical measures of the degree of fit.
5. Others functions may be proposed, and procedure is repeated.

Method of Least Squares :

The simplest situation is a linear or straight – line relation between a single input and the response :

$$E(y) = \alpha + \beta x$$

Where α and β are constants parameters that we want to estimate, (regression coefficients). For a sample of n pairs of data (x_i , y_i) we calculate a , for α and b for β .

If at $x = x_i$, $\hat{y}_i$ is the estimated value of (y) , we have the fitted regression line :

$$\hat{y}_i = a + bx_i$$

Let $e_i = y_i - \hat{y}_i$ be the deviation in the Y – direction of any data pt^s. from the fitted regression line. Then the estimates a and b are chosen so that the sum of the squares of deviations of all the $pt^s. \sum e_i^2$ is smaller than for any other choice of a and b . so that :

$\sum e_i^2 = \sum(y_i - \hat{y}_i)^2$ has a min. value.

This is called the **method of Least Squares** and the resulting eqn. Called the **regression line** of y on x , where y is the response (dependent) and x is the input (independent variable).

If the estimated eqn. $\hat{y} = a + bx$ then $e_i = y_i - (a + bx)$ these deviation called **residuals**

$$e_i^2 = [y_i - (a + bx)]^2$$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (a + bx)]^2$$

This sum of the squares of the deviations or errors or residuals for all n pt^s. is abbreviated as SSE. So the principle of L.S.M. is to minimize

$$\text{SSE} = \sum e_i^2 = \sum(y_i - \hat{y}_i)^2 = \sum[y_i - (a + bx_i)]^2$$

To minimize a quantity we take the derivative with respect to the independent variable and set it equal to zero.

$$\frac{\partial}{\partial a}(SSE) = \frac{\partial}{\partial a} \sum [y_i - (a + bx_i)]^2$$

$$= -2[\sum y_i - n\,a - b \sum x_i] = 0 \quad \ldots\ldots (1)$$

And

$$\frac{\partial}{\partial b}(SSE) = \frac{\partial}{\partial b}\sum[y_i - (a + bx_i)]^2$$

$$= -2\left[\sum x_i y_i - a\sum x_i - b\sum x_i^2\right] = 0 \quad \ldots\ldots (2)$$

Eqn^s (1) and (2) are called the least squares eqn^s. (or normal equations).

Eqn. (1) and (2) can be solved simultaneously, the results are :

$$b = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}[\sum x_i]^2} \quad \ldots\ldots (3)$$

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$a = \frac{\sum y_i - b\sum x_i}{n} = \bar{y} - b\bar{x} \quad \ldots\ldots (4)$$

Then we have :

The sum of squares for $x = S_{xx} = \sum(x_i - \bar{x})^2$

1. $S_{xx} = \sum x_i^2 - \frac{1}{n}[\sum x_i]^2 \quad \ldots\ldots (5)$

2. $S_{yy} = \sum(y_i - \bar{y})^2 = \sum y_i^2 - \frac{1}{n}[\sum y_i]^2 \quad \ldots\ldots (6)$

3. $S_{x,y} = \sum(x - \bar{x})(y - \bar{y}) = \sum x_i y_i - \frac{1}{n}[\sum x_i][\sum y_i] \quad \ldots (7)$

Eqn^s (3) and (4) can be written compactly as :

$$b = \frac{S_{x,y}}{S_{xx}} \quad \ldots\ldots (8)$$

And

$$a = \bar{y} - b\bar{x} \quad \ldots\ldots (9)$$

If we subst. in the eqn. (9)

$$\hat{y} = a + b\,x_i$$

We get

$$(\hat{y}_i - \bar{y}) = b(x_i - \bar{x})$$

This is indicates that the best-fit line passes through the pt. $(\bar{x}, \bar{y})$, which is called the centroidal pt. and is the centre of mass of the data pt^s.

Example 1)

Data for simple linear regression :

x :	0	1	2	3	4	5	6	7	8	9	10	11	12
y :	3.85	0.03	3.50	6.13	4.07	7.07	8.66	11.65	15.23	12.29	14.74	16.02	16.86

Soln. :

N = 13 , $\sum x_i = 78$, $\sum y_i = 120.1$

$$\sum x_i^2 = 650 \quad , \quad \sum y_i^2 = 1483.0828 \quad , \quad \sum x_i y_i = 968.95$$

The centroidal pt. $(\bar{x}, \bar{y}) = (6\,, 9.23846)$

$$S_{xx} = \sum x_i^2 - \frac{1}{n}[\sum x_i]^2 = 650 - \frac{1}{13}(78)^2$$

$$S_{xx} = 182$$

$$S_{yy} = \sum y_i^2 - \frac{1}{n}\left[\sum y_i\right]^2 = 1483.08 - \frac{1}{13}(120.1)^2$$

$$S_{yy} = 373.5436$$

$$S_{x,y} = \sum x_i y_i - \frac{1}{n}\left[\sum x_i\right]\left[\sum y_i\right]$$

$$= 968.95 - \frac{1}{3}(78)(120.1)$$

$$S_{x,y} = 248.35$$

$$b = \frac{S_{xy}}{S_{xx}} = \frac{284.35}{182} = 1.36456$$

$$a = \bar{y} - b\bar{x} = 9.23846 - (1.36456)(6) = 1.0511$$

$\therefore$ the best – fit regression eqn. Is

$$y = 1.0511 + 1.36456x$$

Variance of experimentel pt^s. around the line :

This must be found from the residuals,

$$e_i = y_i - \hat{y} = y_i - (a + bx_i) = y_i - a - bx_i$$

$$SSE = \sum (y_i - a - bx_i)^2$$

Since $a = \bar{y} - b\bar{x}$ --- (9)

$$\bar{y} = a + b\bar{x} \rightarrow a = \bar{y} - b\bar{x}$$

$$\therefore \quad SSE = \sum [(y_i - \bar{y}) - b(x_i - \bar{x})]^2$$

$$= \sum (y_i - \bar{y})^2 - 2b \sum (x_i - \bar{x})(y_i - \bar{y}) + b^2 \sum (x_i - \bar{x})^2$$

$$SSE = S_{yy} - 2bS_{xy} + b^2 S_{xx}$$

$$\because \quad b = \frac{S_{xy}}{S_{xx}}$$

$$\therefore \quad SSE = S_{yy} - 2bS_{xy} + \frac{(S_{xy})(S_{xy})}{(S_{xx})(S_{xx})} \cdot S_{xx}$$

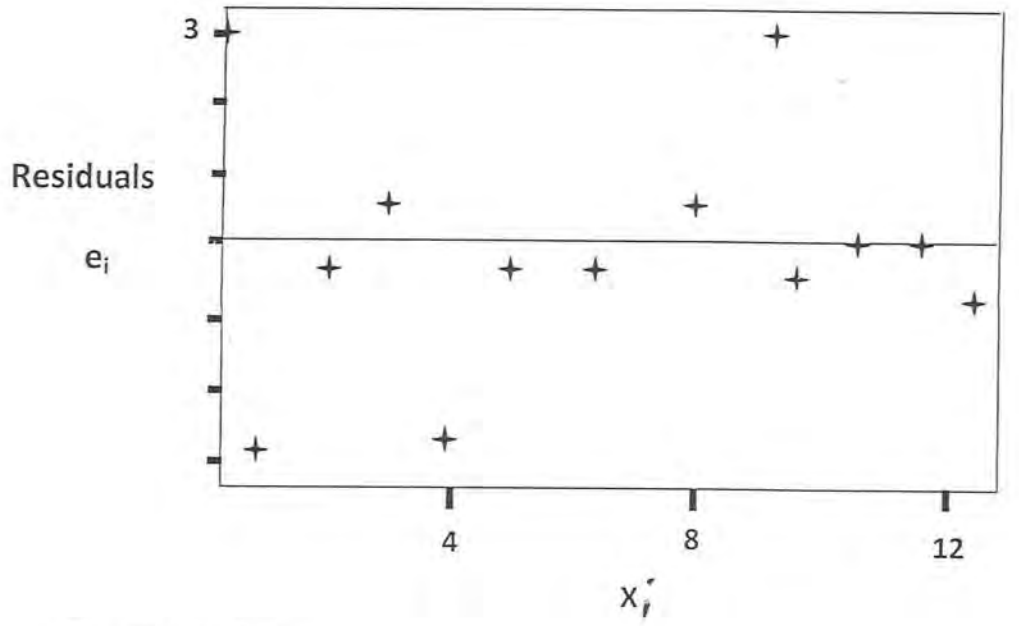
$$= S_{yy} - 2bS_{xy} + bS_{xy}$$

$$\therefore \quad SSE = S_{yy} - bS_{xy}$$

The estimate of the variance of the pt^s. about the line is :

$$S^2_{y\backslash x} = \frac{SSE}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}$$

This quantity is a measure of the scatter of experimental pt^s. around the line.



Example 2)

For the data of example (1) calculate the standard deviation of pt^s. about the regression line, then plot residuals against x.

$S^2_{y/x} = \frac{SSE}{n-2} = \frac{S_{yy} - b\,S_{xy}}{n-2}$

$S^2_{y/x}$ = variance

$\sqrt{S^2_{y/x}}$ = standard div

$S_{y/x} = \sqrt{\frac{SSE}{n-2}}$

Soln. :

$\hat{y} = a + bx$

$\hat{y} = 1.0511 + 1.36456\,x$ (from ex. (1))

Residual $e_i = y_i - \hat{y}$

$$S_{y\backslash x} = \sqrt{\frac{SSE}{n-2}}$$

x_i	y_i	$\hat{y}$	e_i
0	3.85	1.05	+ 2.8
1	0.03	2.41	- 2.38
2	3.5	3.77	- 0.27
3	6.13	5.13	+ 1.0
4	4.07	6.49	- 2.42

$$b = \frac{S_{xy}}{S_{xx}} = \frac{284.35}{182} = 1.36456$$

$$a = \bar{y} - b\bar{x} = 9.23846 - (1.36456)(6) = 1.0511$$

$\therefore$ the best – fit regression eqn. Is

$$y = 1.0511 + 1.36456x$$

Variance of experimentel pt^s. around the line :

This must be found from the residuals,

$$e_i = y_i - \hat{y} = y_i - (a + bx_i) = y_i - a - bx_i$$

$$SSE = \sum (y_i - a - bx_i)^2$$

Since $a = \bar{y} - b\bar{x}$ — (9)

$$\bar{y} = a + b\bar{x} \rightarrow a = \bar{y} - b\bar{x}$$

$$\therefore \quad SSE = \sum [(y_i - \bar{y}) - b(x_i - \bar{x})]^2$$

$$= \sum (y_i - \bar{y})^2 - 2b \sum (x_i - \bar{x})(y_i - \bar{y}) + b^2 \sum (x_i - \bar{x})^2$$

$$SSE = S_{yy} - 2bS_{xy} + b^2 S_{xx}$$

$$\because \quad b = \frac{S_{xy}}{S_{xx}}$$

$$\therefore \quad SSE = S_{yy} - 2bS_{xy} + \frac{(S_{xy})(S_{xy})}{(S_{xx})(S_{xx})} \cdot S_{xx}$$

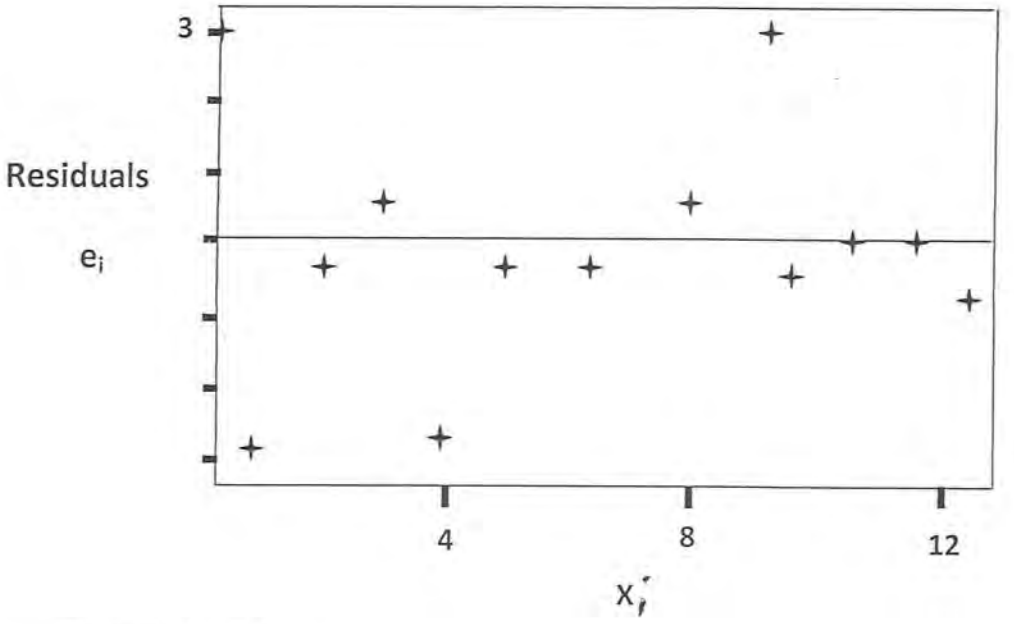
$$= S_{yy} - 2bS_{xy} + bS_{xy}$$

$$\therefore \quad SSE = S_{yy} - bS_{xy}$$

The estimate of the variance of the pt^s. about the line is :

$$S^2_{y\backslash x} = \frac{SSE}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}$$

This quantity is a measure of the scatter of experimental pt^s. around the line.



Example 2)

For the data of example (1) calculate the standard deviation of pt^s. about the regression line, then plot residuals against x.

$S^2_{y/x} = \frac{SSE}{n-2} = \frac{S_{yy} - b\,S_{xy}}{n-2}$

$S^2_{y/x}$ = variance

$\sqrt{S^2_{y/x}}$ = standard div

$S_{y/x} = \sqrt{\frac{SSE}{n-2}}$

Soln. :

$\hat{y} = a + bx$

$\hat{y} = 1.0511 + 1.36456\,x$ (from ex. (1))

Residual $e_i = y_i - \hat{y}$

$$S_{y\backslash x} = \sqrt{\frac{SSE}{n-2}}$$

x_i	y_i	$\hat{y}$	e_i
0	3.85	1.05	+ 2.8
1	0.03	2.41	- 2.38
2	3.5	3.77	- 0.27
3	6.13	5.13	+ 1.0
4	4.07	6.49	- 2.42

5	7.07	7.85	- 0.78
6	8.66	9.21	- 0.55
7	11.65	10.57	+ 1.08
8	15.23	11.93	+ 3.3
9	12.29	13.56	- 1.27
10	14.74	14.65	+ 0.09
11	16.02	16.01	+ 0.01
12	16.86	17.37	- 0.51

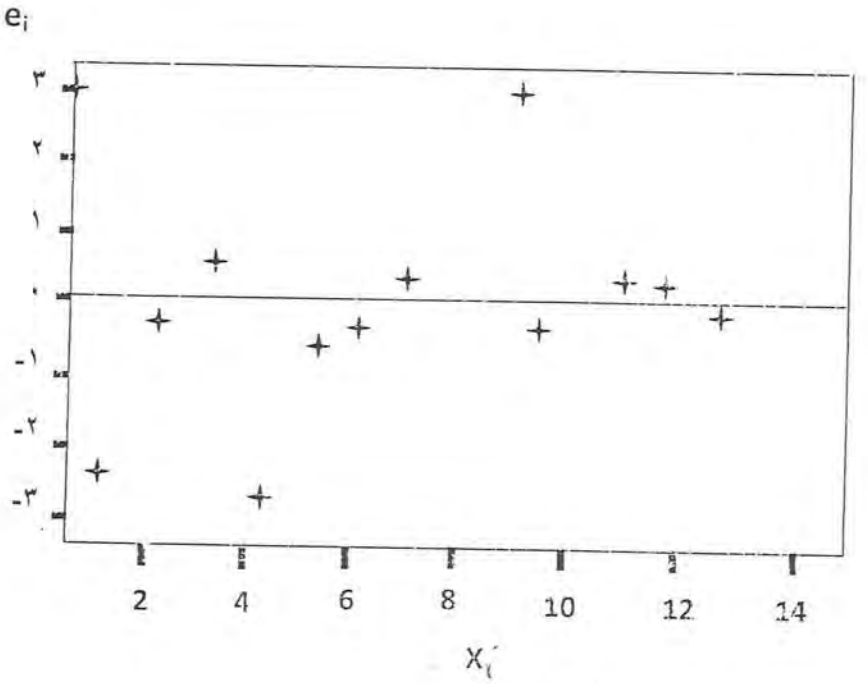
$$SSE = S_{yy} - bS_{xy}$$

$$= 373.5436 - 1.36456(248.35)$$

$$= 34.655$$

$$S_{y\backslash x} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{34.655}{13-2}}$$

$$S_{y\backslash x} = 1.775$$



Relation forms :

1. Straight line through origin $\rightarrow y = m\,x$

2. Other single constant forms are all transformable to straight line through origin.

$y = me^{x}$ define $Y = y$, $X = e^{x}$ $\rightarrow Y = mX$

3. straight line $\rightarrow y = a_0 + a_1 x$

4. straight line forms :

- Two – constant relations may be transformed to straight line :

$y = ae^{bx}$ exponential $\rightarrow \ln y = \ln a + bx$

$y = ax^{b}$ power $\rightarrow \ln y = \ln a + b \ln x$

$y = \frac{1}{a_0 + a_1 x}$ hyporbola $\rightarrow \frac{1}{y} = a_0 + a_1 x$

5. Higher constant relations :

These may be polynomials or other forms that contain more than two constants. It is not usually possible to transform them in to st. line forms.

e.g. /

$y = a_0 + a_1 x + a_2 x^2$ 2^{nd} degree polynomial

(*) $y = a + be^{cx}$ modified exponential

$\frac{a}{y} = b + cx$ $\rightarrow \frac{1}{y} = \frac{b}{a} + \frac{c}{a} x$ (st. line form).

Example)

Transform $p = exp.(a + \frac{1}{bx})$ and define parameters :

$$\ln p = a + \frac{1}{bx} \qquad \therefore\ Y = \ln p \quad , X = \frac{1}{x}$$

$$A_0 = a \quad , A_1 = \frac{1}{b}$$

Example 1)

Fit the following data to a straight line :

Time :	0	3	5	8	10	12
Speed :	0.28	11.2	18.3	29.1	36.2	43.4

Solu. :

$$N = 6 \quad , \sum x_i = 38 \quad , \sum y_i = 138.48$$

$$\sum x_i^2 = 342 \quad , \quad \sum y_i^2 = 4501.2 \quad , \quad \sum x_i y_i = 1240.7$$

$$\bar{y} = 23.08 \quad , \quad \bar{x} = 6.33$$

$$b = \frac{S_{xy}}{S_{xx}} \quad , \quad a = \bar{y} - b\bar{x}$$

$$S_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right)$$

$$= 1240.7 - \tfrac{1}{6}(38)(138.48) = 363.7$$

$$S_{xx} = \sum x_i^2 - \frac{1}{n}\left(\sum x\right)^2 = 342 - \frac{1}{6}(38)^2 = 101.3$$

$$b = \frac{363.7}{101.3} = 3.59 \quad , \qquad a = 23.08 - 3.59 * 6.33 = 0.34$$

$b = 3.59 \quad , \quad a = 0.34$

The relation :

$S = 0.34 + 3.59t$

Example 2)

Fit the following data to $p = exp.\left[a + \frac{1}{bx}\right]$

x :	22	23	24	25	26
p :	0.368	0.223	0.134	0.082	0.05

Solu. :

Transform relation to st. line form :

$$\ln p = a + \frac{1}{bx} \quad , Y = \ln p \qquad , \; X = \frac{1}{x}$$

$$A_0 = a \quad , A_1 = \frac{1}{b}$$

$$N = 5 \quad , \sum x_i = 0.2091 \quad , \sum y_i = -10.007$$

$$\sum x_i^2 = 8.77X10^{-3} \qquad , \; \sum xy = -0.4097$$

$$S_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right)$$

$$= -0.4097 - \tfrac{1}{5}(-0.2091)(-10.007)$$

$$= 8.79X10^{-3}$$

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

$$= 8.77X10^{-3} - \frac{1}{5}(0.2091)^2$$

$$= 2.54X10^{-5}$$

$$A_1 = \frac{S_{xy}}{S_{xx}} = \frac{8.79X10^{-3}}{2.54X10^{-5}} = 3.46X10^2 = 346$$

$$\bar{x} = 0.0418 \quad , \quad \bar{y} = -2.0014$$

$$A_0 = -2.0014 - 346 * 0.0418 = -16.5$$

$$A_1 = 346 \quad , \quad A_0 = -16.5$$

$$\therefore \quad a = A_0 = -16.5$$

$$A_1 = \frac{1}{b} \quad \rightarrow \quad b = 2.89X10^{-3}$$

$$\therefore \quad p = exp.\left[-16.5 + \frac{1}{2.89X10^{-3}x}\right]$$

Example 3)

Fit the data in (1) above to a st. line that passes through the origin.

$$y = m\,x$$

$$\frac{\partial \sum(y_i - \hat{y})^2}{\partial m} = 0 \quad \rightarrow \quad \frac{\partial \sum(y_i - m\,x_i)^2}{\partial\, m} = 0$$

$$-2\sum(y_i - m\,x_i)\,(x_i) = 0 \quad \rightarrow \quad \sum y_i x_i = m\sum x_i^2$$

$$\therefore \quad m = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{1240.7}{342} = 3.628$$

$$\therefore \quad Relation\ is \quad y = 3.628\,x$$

The Least square parabola :

The least square parabola approximating the set of pt^s. (X_1,Y_1) , (X_2,Y_2), … , (X_n,Y_n) has the eqn.

$$Y = a_0 + a_1X + a_2X^2$$

Where the constants a_0 , a_1 and a_2 are determined by solving simultaneously the eqn^s.

$$\left\{ \begin{array}{c} \sum Y = a_0N + a_1\sum X + a_2\sum X^2 \\ \sum XY = a_0\sum X + a_1\sum X^2 + a_2\sum X^3 \\ \sum X^2Y = a_0\sum X^2 + a_1\sum X^3 + a_2\sum X^4 \end{array} \right\}$$

Called the normal eqn^s. for the least square parabola.

* this technique can be extended to obtain normal eqn^s. for **cubic** and **quartic** curves.

Example 4)

Fit the following data to an eqn. of the form

$y = a_0 + a_1x + a_2x^2$, by the method of least squares.

X	Y	X_{new}	X^2	X^3	X^4	XY	X^2Y
10	157	-5	25	.	625	-785	3925
20	179	-3	9	.	81	-537	1611
30	210	-1	1	.	1	-210	210
40	252	1	1	.	1	252	252
50	302	3	9	.	81	906	2718
60	361	5	25	.	625	1805	9025
	1461	$\sum X = 0$	70	0	1414	1431	17741

Using least square method to obtain the normal eqn^{s}. for 2nd. Order polynomial (parabola).

$$\sum Y = a_0 N + a_1 \sum X + a_2 \sum X^2$$

$$\sum XY = a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3$$

$$\sum X^2 Y = a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4$$

Subst to obtain :

$1461 = 6\ a_0 + 70\ a_2$ .....(1)

$1431 = 70\ a_1$ .....(2) $\rightarrow$ $a_1 = 20.44$

$17741 = 70\ a_0 + 1414\ a_2$ .....(3)

Eqn. (1) * 70 – eqn. (3) * 6

$12270 = 420\ a_0 + 4900\ a_2$

$-$

$106446 = 420\ a_0 + 8484\ a_2$

$-4176 = -3584\ a_2$ $\rightarrow$ $a_2 = 1.165$

$\therefore$ $a_0 = 229.9$

$y = 229.9 + 20.44x + 1.165x^2$

Correlation :

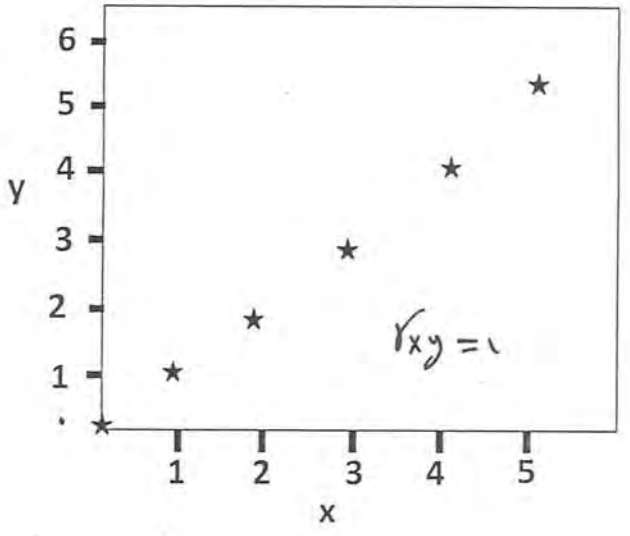
Is a measure of the association between two random variables, both variables are assumed to be varying randomly. We do assume for this analysis that X and Y are related linearly. So the usual correlation coefficient gives a measure of the linear association between X and Y.

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{\sum xy - \frac{1}{n}\sum x \sum y}{\sqrt{\sum x^2 - \frac{1}{n}(\sum x)^2 \cdot \sum y^2 - \frac{1}{n}(\sum y)^2}}$$

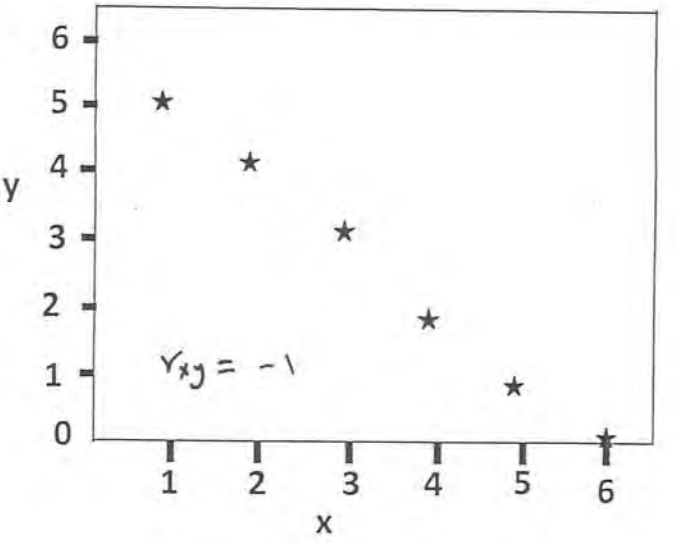
For perfect correlation $\rightarrow r = \pm 1$

If there is no systematic relation between X and Y at all ,

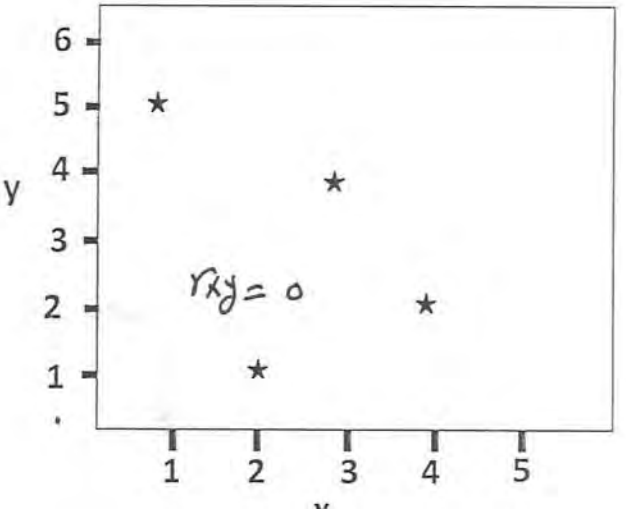
$r_{xy} \simeq 0$

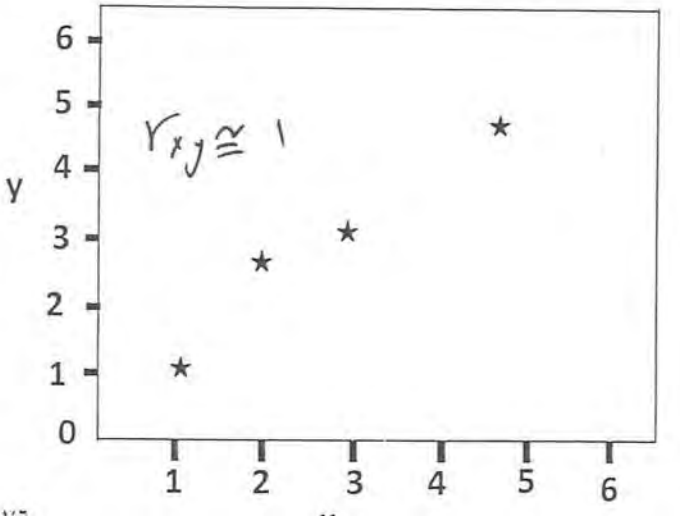


(a)



(b)





This fig. illustrate various correlation coefficients.

A) $r_{xy} = 1$

B) $r_{xy} = -1$

C) $r_{xy} = 0$

D) $r_{xy} \simeq 1$

Tutorial sheet (7)

q. 1) It is required to fit the following eqn^{s}. to a straight line; so determine the constants, then calculate the correlation coefficient. (r_{xy}).

A.

$$y = axe^{x} + bx^{2.2}$$

x :	1	2	3	4	5
y :	37.5	32.0	25.8	28.6	37.6

B.

$$y = ae^{x} + be^{-x}$$

x :	0	0.2	0.4	0.6	0.8	1.0
y :	-1.12	0.026	1.15	2.32	3.59	5.0

C.

$$\ln y = axe^{x} + bx$$

x :	0.21	0.27	0.35	0.38	0.43
y :	10	22	70	100	240

D.

$$y = \frac{x}{a + bx}$$

y :	3.5	7.2	12.6	16.4	20.2
x :	100	200	300	400	500

E .

$$C^2 = \frac{C_i^2}{2C_iKt + 1}$$

C :	2.5	1.65	1.18	0.95	0.88
t :	10	15	20	25	30

F .

$$K = A\, e^{-E/RT}$$

K :	1.22	2.72	4.95	7.39	11.0
T :	316.46	322.58	331.16	336.7	340.14

Multiple and Partial Correlation

- Multiple Correlation :

The degree of relationship existing between three or more variables is called multiple correlation. The fundamental principles involved in problems of multiple correlation are analogous to those of simple correlation.

- Subscript Notation :

- Regression equation. Regression plane :

A regression eqn. Is eqn. For estimating a dependent variable, X_1 , from the independent variables X_2 , X_3 , ... and is called a regression eqn. Of X_1 on X_2 , X_3 ,... and can be written as $X_1=F(X_2 , X_3 , \ldots)$.

- The simplest reg. eqn. Of X_1 on X_2 and X_3 :

$X_1 = b_{1.23} + b_{12.3} X_2 + b_{13.2} X_3$ .....(1)

If we keep X_3 constant in eqn. (1) , the graph of X_1 vs. X_2 is a straight line with slop $b_{12.3}$. if we keep X_2 constant, the graph of X_1 vs. X_3 is a straight line with slop $b_{13.2}$.

The subscript after the dot (.) indicate the variables held constant in each case.

- X_1 varies partially because of variation in X_2 and partially because of variation in X_3 , so $b_{12.3}$ and $b_{13.2}$ called the partial regression coefficients of X_1 on X_2 keeping X_3 constant and of X_1 on X_3 keeping X_2 constant respectively eqn. (1) is called a **linear regression** of X_1 on X_2 and X_3 . In a three dimensional

rectangular co-ordinate system it represents a plane called a **regression plane**.

- Normal eqn^s. for the least square reg. plane :

The least square reg. plane of X_1 on X_2 and X_3 has the eqn. (1) where $b_{1.23}$, $b_{12.3}$ and $b_{13.2}$ are determined by solving simultaneously the normal eqn^s.

$$\left\{\begin{array}{l} \sum X_1 = b_{1.23}N + b_{12.3}\sum X_2 + b_{13.2}\sum X_3 \\ \sum X_1X_2 = b_{1.23}\sum X_2 + b_{12.3}\sum X_2^2 + b_{13.2}\sum X_2X_3 \\ \sum X_1X_3 = b_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X_3^2 \end{array}\right\} \ldots eqn.(2)$$

These can be obtained by multiplying both sides of eqn. (1) by 1 , X_2 and X_3 and summing on both sides.

If $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$, $x_3 = X_3 - \bar{X}_3$

The lst. Eqn. of (2) divided by N , to get :

$$\bar{X}_1 = b_{1.23} + b_{12.3}\bar{X}_2 + b_{13.2}\bar{X}_3 \longrightarrow$$

Subtracting this eqn. from eqn. (1) $X_1 = b_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$

$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

Or :

$$x_1 = b_{12.3}\,x_2 + b_{13.2}\,x_3 \qquad \ldots\ldots (3)$$

Where $b_{12.3}$ and $b_{13.2}$ are obtained by solving simultaneously the eqn^s.

$$\left\{ \begin{array}{l} \sum x_1 x_2 = b_{12.3} \sum x_2^2 + b_{13.2} \sum x_2 x_3 \\ \sum x_1 x_3 = b_{12.3} \sum x_2 x_3 + b_{13.2} \sum x_3^2 \end{array} \right\} \quad \ldots\ldots (4)$$

These eqn^s. which are equivalent to the normal equ^s. (2) and can be obtained by multiplying both sides of (3) by x_2 and x_3 and summing both sides.

Example 1)

The following table shows the corresponding values of three variables X_1 , X_2 and X_3 . find the least square reg. eqn. of X_3 on X_1 and X_2 .

X_1	X_2	X_3
3	16	90
5	10	72
6	7	54
8	4	42
12	3	30
14	2	12
$\sum = 48$	$\sum = 42$	$\sum = 300$

$X_3 = f(X_1 , X_2)$

$X_3 = b_{3.12} + b_{31.2} X_1 + b_{32.1} X_2$

The normal eqn^s. of the least square reg. line :

$$\left\{ \begin{array}{c} \sum X_3 = b_{3.12} N + b_{31.2} \sum X_1 + b_{32.1} \sum X_2 \\ \sum X_3 X_1 = b_{3.12} \sum X_1 + b_{31.2} \sum X_1^2 + b_{32.1} \sum X_2 X_1 \\ \sum X_3 X_2 = b_{3.12} \sum X_2 + b_{31.2} \sum X_1 X_1 + b_{32.1} \sum X_2^2 \end{array} \right\}$$

Or by using eqn. (4)

$x_3 = b_{31.2}\, x_1 + b_{32.1}\, x_2$

$$\sum x_3 x_1 = b_{31.2} \sum x_1^2 + b_{32.1} \sum x_1 x_2$$

$$\sum x_3 x_2 = b_{31.2} \sum x_1 x_2 + b_{32.1} \sum x_2^2$$

$\bar{X}_1 = 8$, $\bar{X}_2 = 7$, $\bar{X}_3 = 50$

x_1	x_2	x_3	x_3x_1	x_3x_2	x_2x_1	x_1^2	x_2^2
-5	9	40	-200	360	-45	25	81
-3	3	22	-66	66	-9	9	9
–2	0	4	-8	0	0	4	0
0	-3	-8	0	24	0	0	9
4	-4	-20	-80	80	-16	16	16
6	-5	-38	-228	190	-30	36	25
$\sum = 0$	$\sum = 0$	$\sum = 0$	$\sum = -582$	$\sum = 720$	$\sum = -10($	$\sum = 90$	$\sum = 140$

-582 = $b_{31.2}$ (90) + $b_{32.1}$ (-100) .....(1)

720 = $b_{31.2}$ (-100) + $b_{32.1}$ (140) .....(2)

Eqn. (1) * (100) + eqn. (2) * (90) yield :

6600 = $b_{32.1}$ (2600)

∴ $b_{32.1}$ = 2.54

-582 = $b_{31.2}$ (90) + 2.54 (-100)

$b_{31.2}$ = -3.64

subst. the above constant in following eqn.

$\bar{X}_3 = b_{3.12} + b_{31.2}\bar{X}_1 + b_{32.1}\bar{X}_2$

$50 = b_{3.12} + (-3.64)\,(8) + (2.54)\,(7)$

$b_{3.12} = 61.34$

- Standard error of estimate : Can be defined as :

$$S_{1.23} = \sqrt{\frac{\sum(X_1 - X_{1.est})^2}{N}}$$

$X_{1.est.}$ calculated from reg. eqn. :

$X_1 = b_{1.23} + b_{12.3}\, X_2 + b_{13.2}\, X_3$

- Coefficient of Multiple correlation :

$$R_{1.23} = \sqrt{1 - \frac{S_{1.23}^2}{S_1^2}}$$

$$S_1 = \sqrt{\frac{\sum(X_1 - \bar{X}_1)^2}{N}} = \sqrt{\frac{\sum x_1^2}{N}}$$

OR :

$$R_{.123} = \sqrt{1 - \frac{\sum(X_1 - X_{1.est})^2}{\sum(X_1 - \bar{X}_1)^2}}$$

Example 2)

Find the standard error of estimate of X_3 on X_1 and X_2 of example (1) :

Soln. :

- The reg. eqn. of X_3 on X_1 and X_2 :

$X_3 = 61.34 - 3.64\ X_1 + 2.54\ X_2$

$$S_{3.12} = \sqrt{\frac{\sum(X_3 - X_{3.est})^2}{N}}$$

X_3	$X_{3.est.}$	$(X_3 - X_{3.est})^2$
90	91.06	1.124
72	68.54	11.97
54	57.3	10.89
42	42.4	0.16
30	25.3	22.1
12	15.43	11.76
		$\sum$ =58.0

$$\therefore\ S_{3.12} = \sqrt{\frac{58}{6}}$$

$S_{3.12} = 3.11$

- The correlation coefficient of X_3 on X_2 and X_1

$$R_{3.12} = \sqrt{1 - \frac{S_{3.12}^2}{S_3^2}}$$

$$S_3 = \sqrt{\frac{\sum(X_3 - \bar{X}_3)^2}{N}} = \sqrt{\frac{\sum x_3^2}{N}}$$

x_3	x_3^2
40	1600
22	484
4	16
-8	64
-20	400
-38	1444
$\sum x_3 = 0$	$\sum = 4008$

$$S_3 = \sqrt{\frac{4008}{6}} = 25.85$$

$$\therefore \quad R_{3.12} = \sqrt{1 - \frac{(3.11)^2}{(25.85)^2}}$$

$R_{3.21} = 0.9927$