INDRAPRASTHA COLLEGE FOR WOMEN

(31, Shamnath Marg, Civil Lines, New Delhi-110054)

(University of Delhi)

Statistical Methodology Practical File

Submitted By :-

Eva Chaudhary
Bsc(H) Computer Science
14029270015

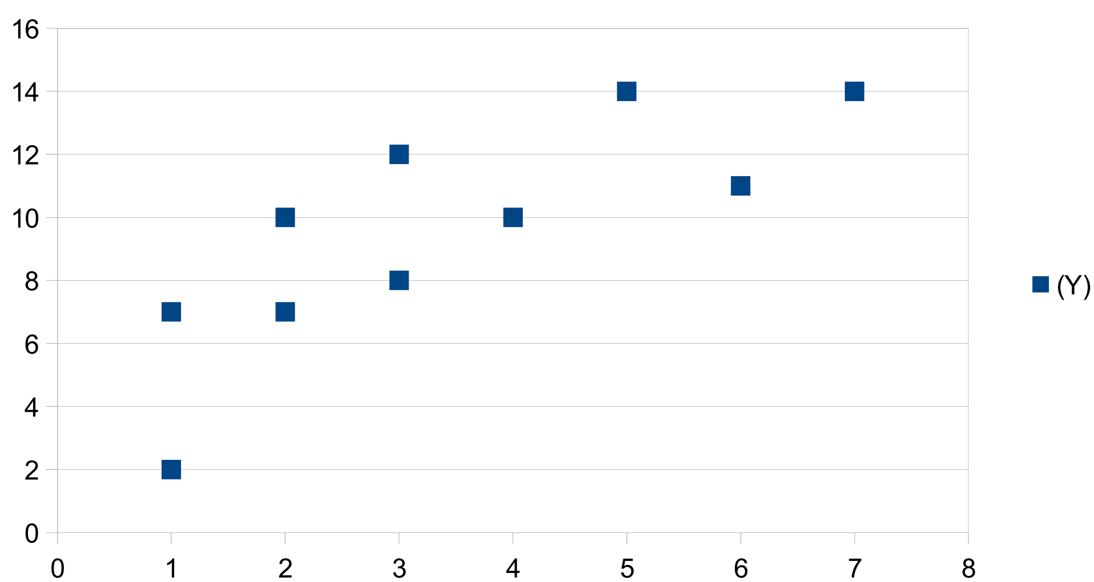
Ques 1) Presentation of Bi-variate data through scatter plot diagram and calculations of covariance .

(X)	(Y)	XY
1	2	2
1	7	7
2	7	14
2	10	20
3	8	24
3	12	36
4	10	40
5	14	70
6	11	66
7	14	98
34	95	377

Mean(X)	3.4	**E(X)**
Mean(Y)	9.5	**E(Y)**
E(XY)	37.7	

Cov(X,Y) = E(XY) – E(X)E(Y)

Cov(X,Y) = 5.4



Ques 2) Fitting of Polynomial using Principle of least square method .
For 10 randomly selected observations the following data were recorded :-

(X)	(Y)	X^2	Y^2	XY	(X^2)Y	X^3	X^4
1	2	1	4	2	2	1	1
1	7	1	49	7	7	1	1
2	7	4	49	14	28	8	16
2	10	4	100	20	40	8	16
3	8	9	64	24	72	27	81
3	12	9	144	36	108	27	81
4	10	16	100	40	160	64	256
5	14	25	196	70	350	125	625
6	11	36	121	66	396	216	1296
7	14	49	196	98	686	343	2401
34	95	154	1023	377	1849	820	4774

Equations :-	**Matrix A**			**Matrix B**
	10	34	154	95
10a+34b+154c=95	34	154	820	377
34a+154b+820c=377	154	820	4774	1849
154a+820b+4774c=1849				

Inverse	1.427441353	-0.8191489362	0.0946536	1.80223677	**(Value of a)**
matrix is :-	-0.8191489362	0.5460992908	-0.067376	3.482269504	**(Value of b)**
A(Inverse)	0.0946535734	-0.0673758865	0.0087289	-0.26895799	**(Value of c)**

The Nonlinear quadratic equation which fits the data is :- $1.80223677+3.482269504X-0.268957992X^2$

Ques 3) Find the correlation coefficient for a bi-variate distribution .
Calculate the correlation coefficient for the heights of fathers(X) and their sons(Y).

X	Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
544	552	37028	38132	37560

E(X)	68	**Cov(X,Y)**	3
E(Y)	69	**Var(X)**	4.5
E(XY)	4695	**Var(Y)**	5.5
E(X^2)	4628.5	**S.D.(X)**	2.12132
E(Y^2)	4766.5	**S.D.(Y)**	2.345208

Corr(X,Y)=r= $Cov(X,Y)/\sigma(X)\sigma(Y)$

r 0.6030227

Ques 4) Calculation of Spearman's Rank Correlation Coefficients.

a) Without Repetition :-

Discuss which pair of judges has the nearest approach to common liking to music.

Rank by A	Rank by B	Rank by C	D1=(A-B)	D2=(B-C)	D3=(A-C)	D1^2	D2^2
1	3	6	-2	-3	-5	4	9
6	5	4	1	1	2	1	1
5	8	9	-3	-1	-4	9	1
10	4	8	6	-4	2	36	16
3	7	1	-4	6	2	16	36
2	10	2	-8	8	0	64	64
4	2	3	2	-1	1	4	1
9	1	10	8	-9	-1	64	81
7	6	5	1	1	2	1	1
8	9	7	-1	2	1	1	4
					$\Sigma(Di)$	200	214

	D3^2
	25
	4
	16
	4
	4
	0
	1
	1
	4
	1
$\Sigma(Di)$	60
N =	10

Rank Correlation Coefficient = $\rho = 1 - 6\Sigma d^2 / N^3 - N$

Rank correlation between A and B :- -0.21212

Rank correlation between B and C :- -0.29697

Rank correlation between A and C :- 0.636364

Since, the rank correlation coefficient between A and C approaches to 1, hence A and C has the nearest approach to common liking to music.

b) With Repetition :-

Marks in Accounts (X)	Marks in Stats(Y)	Rank in X(R1)	Rank in Y(R2)	Diff(D)	D^2
15	40	2	6	-4	16
20	30	3.5	4	-0.5	0.25
28	50	5	7	-2	4
12	30	1	4	-3	9
40	20	6	2	4	16
60	10	7	1	6	36
20	30	3.5	4	-0.5	0.25
80	60	8	8	0	0
			TOTAL		81.5

T (Correction of Repeated value) $m1(m1^2-1)/12+m2(m2^2-1)/12+----mn(mn^2-1)$

Correlation R = $1-(\Sigma D^2+T)/(N^3-N)$

N 8

T(X) 0.5
T(Y) 2

Rank Correlation(R) 0

Thus the marks in Accountancy and statistics are Uncorrelated.

Ques 5) Fitting of simple linear regression.

For 10 randomly selected observations the following data were recorded :-

Overtime hours(X)	Additional units(Y)	X^2	Y^2	XY
1	2	1	4	2
1	7	1	49	7
2	7	4	49	14
2	10	4	100	20
3	8	9	64	24
3	12	9	144	36
4	10	16	100	40
5	14	25	196	70
6	11	36	121	66
7	14	49	196	98
34	95	154	1023	377

E(X)	3.4			
E(Y)	9.5	**Correlation(r)**	$cov(X,Y)/\sigma(x)\sigma(y)$	
E(XY)	37.7			
E(X^2)	15.4		**Cov(X,Y)**	5.4
E(Y^2)	102.3			
		Correlation(r)		9.565808251
Var(X) = $\sigma^2(x)$	3.84			
Var(Y) = $\sigma^2(y)$	12.05			

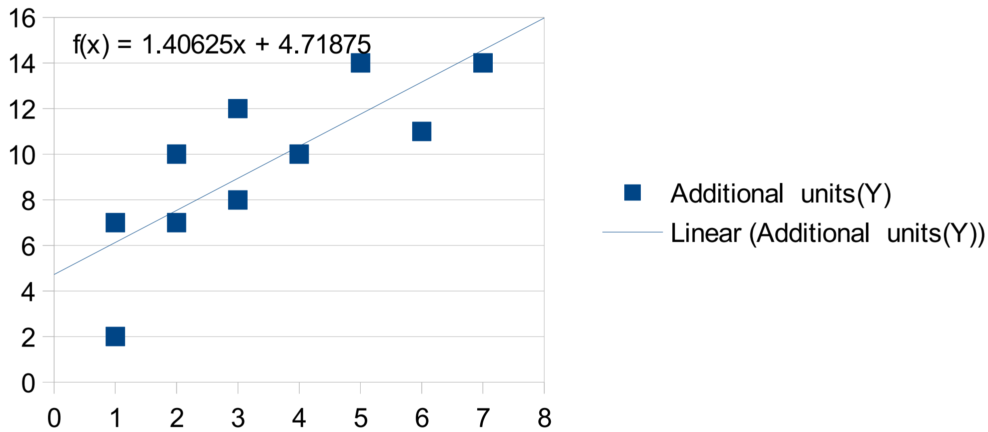
Standard Deviation $\sigma(x)$ 1.9595917942
Standard Deviation $\sigma(y)$ 3.4713109915

Equation of line Y on X through (E(X),E(Y)) is given by :- $Y-E(Y)=(r\sigma(y)/\sigma(x))(X-E(X))$

Y-9.5 = 16.95(X-3.4) **Y-16.95X + 48.13 = 0** $r\sigma(y)/\sigma(x)$ 16.94531

Equation of line X on Y through (E(X),E(Y)) is given by :- $X-E(X)=(r\sigma(x)/\sigma(y))(Y-E(Y))$

X-3.4 = 5.4(Y-9.5) **X – 5.4Y + 47.9 = 0** $r\sigma(x)/\sigma(y)$ 5.4



Ques 6) Fitting of multiple linear regression (3 variates) .

Find the regression equation of X1 on X2 and X3 for the following data and find estimated values of X1.

Weight(X1)	Height(X2)	Age(X3)	X1*X2	X2^2	X2*X3	X1*X3	X3^2
64	57	8	3648	3249	456	512	64
71	59	10	4189	3481	590	710	100
53	49	6	2597	2401	294	318	36
67	62	11	4154	3844	682	737	121
55	51	8	2805	2601	408	440	64
58	50	7	2900	2500	350	406	49
77	55	10	4235	3025	550	770	100
57	48	9	2736	2304	432	513	81
56	52	10	2912	2704	520	560	100
76	61	12	4636	3721	732	912	144
68	57	9	3876	3249	513	612	81
702	601	100	38688	33079	5527	6490	940

Est(X1)
4140
4366
3530
4607
3756
3643
4110
3613
3918
4592
4189

$X1 = aX2 + bX3 + c$ **Eqn 1.**

$\Sigma X1 = a\Sigma X2 + b\Sigma X3 + cn$

$\Sigma X1X2 = a\Sigma X2^2 + b\Sigma X2X3 + c\Sigma X2$

$\Sigma X1X3 = a\Sigma X2X3 + b\Sigma X3^2 + c\Sigma X3$

601	100	11
33079	5527	601
5527	940	100
-0.3195739	0.008877	-0.018198
0.361008851	-0.018198	0.0696588
14.26936634	-0.319574	0.3610089

702 = 601a+100b + 11c
38688 = 33079a + 5527b + 601c
6490 = 5527a + 940b + 100c

Inverse of the matrix is

-0.3195739015	0.0088770528	-0.0181979583	702	0.989791389	**value of a**
0.3610088509	-0.0181979583	0.0696587556	38688	1.470927652	**value of b**
14.2693663351	-0.3195739015	0.3610088509	6490	-3.63249001	**value of c**

Putting values of a,b and c in Eqn 1.

X1 = 0.99X2 + 1.47X3 – 3.63

Ques 7) Calculation of multiple and partial correlation (3 variables only)

Find all the first order partial correlation coefficients from the data relating to three variables given below.

X1	X2	X3			
11	2	2			
17	4	3	**Rank(r23)**		0.979826272
26	6	4	**Rank(r13)**		0.982926676
28	5	5	**Rank(r12)**		0.978976388
31	8	6		**Partial Correlation**	
35	7	7		**Rank(r23.1)**	$r23-r12.r13/\sqrt{(1-r12^2)}*\sqrt{(1-r13^2)}$
41	10	9			
49	11	10	**Rank(r23.1)**		0.467997631
63	13	11	**Rank(r13.2)**		0.581387591
69	14	13	**Rank(r12.3)**		0.431821588

Multiple Correlation

$$R(1.23)=\sqrt{((r12^2+r13^2-2\mathrm{r12r13r23})/1-r23^2)}$$

R(1.23)	0.9861327577
R(2.13)	0.9836194486
R(3.12)	0.9866913729

Ques 8) Test of Hypothesis concerning Population mean.

Ques) Boys of a certain age are known to have a mean weight of 85 pounds. A complaint is made that the boys living in a municipal children's home are underfed. As one bit of evidence,n=25 boys(of same age) are weighed and found to have a mean weight of E(X) = 80.94 pounds. It is known that population standard deviation is 11.6 pounds. Based on the available data, what should be concluded concerning the complaint ?

Null Hypotheis : $\mu = 85$
Alternate Hypothesis : $\mu < 85$

using Z-test , where $Z = (E(X) - \mu)/(\sigma/\sqrt{(n)})$

E(X)	80.94
μ	85
σ	11.6
n	25
Z	-1.75

Now Probability at P(Z<-1.75) = 1 – P(Z<1.75) = 1 – 0.9599

(P-Value) **P(Z<-1.75)** 0.0401

α 0.05(given)

For Null Hypothesis to be accepted , P-Value > α

But the P-Value(0.0401) < α **-0.05**

Hence Null Hypothesis is rejected, and the mean weight of boys is less than 85 pounds.

Ques 9) Test of Significance based on Chi Square distribution.
a)
A die is tossed 120 times and each outcome is recorded as under :-
Is the distribution outcome uniform

Faces	Frequency(Oi)	Expected(Ei)	Oi-Ei	(Oi-Ei)^2
1	20	20	0	0
2	22	20	2	4
3	17	20	-3	9
4	18	20	-2	4
5	19	20	-1	1
6	24	20	4	16
Total	120			34

$$\chi^2 = \Sigma(Oi - Ei)^2 / Ei$$

Chi-square(cal) χ^2 1.7
Chi-square(tab) = 9.236

Since, Chi-square(tab) > Chi-square(cal) , Hence the distribution outcome is uniform.

b)
200 digits when chosen at random from a set of tables,the frequency of digits were

digits	Frequency(Oi)	Expected(Ei)	Oi-Ei	(Oi-Ei)^2
0	18	20	-2	4
1	19	20	-1	1
2	23	20	3	9
3	21	20	1	1
4	16	20	-4	16
5	25	20	5	25
6	22	20	2	4
7	20	20	0	0
8	21	20	1	1
9	15	20	-5	25
Total	200			86

$$\chi^2 = \Sigma(Oi - Ei)^2 / Ei$$

Chi-square(cal) χ^2 4.3
Chi-square(tab) = 16.9

Since, Chi-square(tab) > Chi-square(cal) , Hence Chi-Square fits good to the distribution .

Ques 10) Test of Significance based on f – distribution.
a)
In one sample of 8 observations the sum of the squares of deviation of sample value from the sample mean was 84.4 and in the other sample of 10 observations it was 102.6. Test whether the difference is significant at 5% level of significance. Given F at 5%=3.29

N1	8
N2	10
Xi-E(X)	84.4
Yi-E(Y)	102.6

$$SX^2=\Sigma(Xi-E(X))^2/n1-1$$

$$SY^2=\Sigma(Yi-E(Y))^2/n2-1$$

SX^2	12.0571428571
SY^2	11.4

$$SX^2>SY^2$$

$$F=SX^2/SY^2$$

F(cal) 1.0576441103 **F(tab) = 3.29**

Since F(tab) > F(cal) , therefore this sample is accepted.

b)
The two samples of 8 and 7 items respectively had the following values
Do the two estimates of population variance differ significantly. Given F(tab)=4.21.

	X	Xi-E[X]	(Xi-E(X))^2		Y	Yi-E(Y)	(Yi-E(Y))^2
	9	-2.75	7.5625		10	-0.428571	0.1836735
	11	-0.75	0.5625		12	1.571429	2.4693878
	13	1.25	1.5625		10	-0.428571	0.1836735
	11	-0.75	0.5625		14	3.571429	12.755102
	15	3.25	10.5625		9	-1.428571	2.0408163
	9	-2.75	7.5625		8	-2.428571	5.8979592
	12	0.25	0.0625		10	-0.428571	0.1836735
	14	2.25	5.0625				
Total	94		33.5		73		23.714286
	E(X)	11.75		**E(Y)**		10.42857	

$$SX^2=\Sigma(Xi-E(X))^2/n1-1 \qquad SY^2=\Sigma(Yi-E(Y))^2/n2-1$$

$SX^2>SY^2$ $\quad F=SX^2/SY^2$ **SX^2** **4.7857143** **SY^2** **3.952381**

F(Cal) 1.2108433735 **F(tab)** 4.21

Since F(Tab) > F(Cal) these samples will be accepted.

Ques 11) Test of Significance based on t – distribution .
a)
StudentT-Test T(tab) = 0.711

Find the student T for the following variable values in a sample of 8.

X	**Xi-E(X)**	**(Xi-E(X))^2**
-4	-4.25	18.0625
-2	-2.25	5.0625
-2	-2.25	5.0625
0	-0.25	0.0625
2	1.75	3.0625
2	1.75	3.0625
3	2.75	7.5625
3	2.75	7.5625
2		49.5

E(X) 0.25

Since n=8, therefore degree of freedom(dof)
Is n-1 ,i.e. dof=7.

var(s^2) $s^2=\Sigma(Xi-E(X))^2/(n-1)$ $\mu=0$

s^2	7.0714285714
S.D.(s)	2.6592157813
sqrt(n)=sqrt(8)	2.8284271247

T-test(t) $t=(E(X)-\mu)/(s/\sqrt{(n)})$

t**-test(t)** 0.265908 (**At 7 degree of freedom)**

t(cal) = 0.27(approx) **T(tab) = 0.711** And (0.711>0.27)
Since t(cal) < t(tab) , therefore the above sample will be accepted .

b)

Ques) A Random sample of 16 values from a normal population shows a mean of 41.5 and sum of square of deviations from the mean equal to 135. Show that the assumption of a mean 43.5 for the population is not reasonable.

$\Sigma(Xi - E(X))^2 = 135$

E(X)		41.5
μ		43.5

n=16

var(s^2) $s^2 = \Sigma(Xi - E(X))^2 / (n-1)$

s^2 9

T-test(t) $t = (E(X) - \mu) / (s / \sqrt{(n)})$

T-test(t) -2.6666666667

Since T-test(t) value is in the modulus , therefore T-test(t) = 2.7 (approx)

T(cal) = 2.7 T(tab) = 2.95(at 1% level of significance)
T(tab) = 2.13(at 5% level of significance)

Now at 1% level of significance, T(cal) < T(Tab) , therefore the above sample will be accepted.
But at 5% level of significance, T(cal) > T(tab) , therefore the above sample will be rejected.

Ques 12) Calculation of Type I & Type II error.

Ques) A fabric manufacturers believes that the proportion of orders for raw material arriving late is p=0.6. If random sample of 10 order shows that 3 or fewer arrived late, the hypothesis p=0.6 should be rejected in favor of alternative hypothesis p<0.6 . Find the probability of type I and type II error using binomial distribution.

Null Hypothesis : p=0.6
Alternative hypothesis : p<0.6

α = P{x<=3 | p=0.6}

β = P{x>3 | p=0.3}

Calculation of Type I and Type II Error.

X	P(X=x,p=0.6)	P(X=x,p=0.3)
0	0.0001048576	0.0282475249
1	0.001572864	0.121060821
2	0.010616832	0.2334744405
3	0.042467328	0.266827932
4	0.111476736	0.200120949
5	0.2006581248	0.1029193452
6	0.250822656	0.036756909
7	0.214990848	0.009001692
8	0.120932352	0.0014467005
9	0.040310784	0.000137781
10	0.0060466176	5.90490E-006

$\alpha(\mathit{Type\ I\ Error})$ 0.0547618816

$\beta(\mathit{Type\ II\ Error})$ 0.3503892816

Ques 13) Application of Sampling distribution (Central Limit Theorem).
a)
Ques) Let X be the no of times that a fair coin flips 40 times, lands head. Find the probability That X=20 . Use the normal approximation and then compare it to the exact value.

Since, the binomial is discrete random variable and the normal is continuous random variable, so we have to write for better approximation, we write as

P(X=20) = P(19.5<X<20.5)

Central Limit Theorem is defined as:

$$Z = X - \mu / \sigma / \sqrt{(n)}$$

P(X=20) = P(19.5<X<20.5)
P(19.5-20 < X-20 <20.5-20)

$$P\left(-0.5/\sqrt{(npq)} < X - 20/\sqrt{(npq)} < 0.5/\sqrt{(npq)}\right)$$

n=40	**P=0.5**	**Q=0.5**
npq	10	
sqrt(npq)	3.1622776602	

$$P\left(-0.5/\sqrt{(10)} < Z < 0.5/\sqrt{(10)}\right)$$

P(-0.16 < Z< 0.16)
P(Z < 0.16) – P(Z<-0.16)
P(Z < 0.16) – 1 + P(Z < 0.16)
2P(Z < 0.16) – 1
2*(0.5636) – 1
1.1272 – 1
P(X=20) = 0.1272

Exact value 0.1253706876 **(using binomial distribution)**

Both the values are approximately same .

b)

Ques) The life time of a special type of battery is a random variable with mean 40 and Std Deviation 20 hrs. A battery is used until it fails at which point it is replaced by a new one. Assuming a stock pile of 25 such batteries, the life time of which are independent. Approx the probability that over 1100 hrs of use can be obtained.

$\mu=40$ $\sigma=20$ n=25

P(X1 + X2 + X3 + --------- + X25 > 1100)

If we let Xi denote the lifetime of a ith battery to be put in use, then we desire the probability(X1 + X2 + ---- + X25 > 1100)

$$P\left(\left(X1+X2+----X25-n\mu\right)/\sigma/\sqrt{(n)}>1100-n\mu/20*5\right)$$

$$P\left(\left(X1+X2+---+X25-25*40\right)/100>1100-25*40/100\right)$$

$$P(Z>1)=1-P(Z<1)$$

P(Z > 1) = 1 – 0.8413

P(Z>1) = 0.1587

Ques 14) Calculation of sample mean and simple variance of continuous data (grouped data).

Ques) Find the sample mean and sample variance of the following data for the marks obtained in a test by 88 students.

Marks	Mid Interval Value (X)	frequency (f)	(fX)	X^2	f*(X^2)
0<=X<10	5	6	30	25	150
10<=X<20	15	16	240	225	3600
20<=X<30	25	24	600	625	15000
30<=X<40	35	25	875	1225	30625
40<=X<50	45	17	765	2025	34425
Total		88	2510	4125	83800

Mean $\mu=\Sigma(fX)/n$ **n=88**

μ 28.522727273

Variance $\sigma^2=\Sigma(f*(X^2))/n-\mu^2$

σ^2 138.7267562

Sample Mean = $(\mu+\mu+---+(ntimes)\mu)/n$

Hence Sample Mean = μ 28.522727

Sample Variance = $(\sigma^2+\sigma^2+---+(ntimes)\sigma^2)/n^2=\sigma^2/n$

Hence Sample Variance = 1.5764404