**Department of Statistics,**

**Modern college of Arts, Science and Commerce, Pune-05**

**M.Sc.I (Statistics) Semester II**

**DateST- 28                                                           Submission date:**

**Practical No. 10**

Practical no:-10

**Title:** Exploratory multivariate data analysis.

Q1. Consider the sweat data given below with three variable Sweat rate(X1), Sodium(X2) and Potassium(X3)

|  |  |  |  |
| --- | --- | --- | --- |
| Individual | Sweat rate (X1) | Sodium (X2) | Potassium(X3) |
| 1 | 3.7 | 48.5 | 9.3 |
| 2 | 5.7 | 65.1 | 8 |
| 3 | 3.8 | 47.2 | 10.9 |
| 4 | 3.2 | 53.2 | 12.0 |
| 5 | 3.1 | 55.5 | 9.7 |
| 6 | 4.6 | 36.1 | 7.9 |
| 7 | 2.4 | 24.8 | 14 |
| 8 | 7.2 | 33.1 | 7.6 |
| 9 | 6.7 | 47.4 | 8.5 |
| 10 | 5.4 | 54.1 | 11.3 |
| 11 | 3.9 | 36.9 | 12.7 |
| 12 | 4.5 | 58.8 | 12.3 |
| 13 | 3.5 | 27.8 | 9.8 |
| 14 | 4.5 | 40.2 | 8.4 |
| 15 | 1.5 | 13.5 | 10.1 |
| 16 | 8.5 | 56.4 | 7.1 |
| 17 | 4.5 | 71.6 | 8.2 |
| 18 | 6.5 | 52.8 | 10.9 |
| 19 | 4.1 | 44.1 | 11.2 |
| 20 | 5.5 | 40.9 | 9.4 |

1. Compute the mean and variance for the above data without assuming multivariate normality.
2. Compute correlation matrix for the above data
3. Check the three variates Normality of given data- x1,x2, and x3

Q2. Consider the sales data given below with three variable Sales in Millions of Dollars(X1), Profits in Millions of Dollars(X2) and Assets in Millions of Dollars(X3)

|  |  |  |  |
| --- | --- | --- | --- |
| Sr. No. | Sales (X1) | Profits (X2) | Assets (X3) |
| 1 | 126974 | 4224 | 173297 |
| 2 | 96933 | 3835 | 160893 |
| 3 | 86656 | 3510 | 83219 |
| 4 | 63438 | 3758 | 77734 |
| 5 | 55264 | 3939 | 128344 |
| 6 | 50976 | 1809 | 39080 |
| 7 | 39069 | 2946 | 38528 |
| 8 | 36156 | 359 | 51038 |
| 9 | 35209 | 2480 | 34715 |
| 10 | 32416 | 2413 | 25636 |

1. Compute the mean and variance for the above data without assuming multivariate normality.
2. Compute correlation matrix for the above data
3. Check the three variates Normality of given data- x1,x2, and x3

**Algorithm**

Mean=Xbar(p×1)=(1/n)\*X’(p×n)\*E(n×1)

Sample variance-covariance matrix S(n×n) =(1/(n-1))\* X’(p×n)\*(In-(1/n)\*E(n×n))\* X(p×n)

n= number of observation.

p= number o variables.

In=Identity matrix of order n.

E(n×1)=vector of ones .

E(n×n)=matrix of ones .

Sample correlation matrix =R=D-1/2\*S\* D-1/2

D-1/2= (Note that order of matrix =p)

P-variate normality:-.

Joint normality of the data is based on squared generalized distances.

dj2=(Xj-Xbar)’\*S-1\*(Xj-Xbar) j=1,2…n

Calculate the dj2 compare the result with χ2(p,0.50) quintiles .

i.e. for p-variate normality is indicated if Roughly half of the dj2 are less than or equal to χ2(p,0.50).

**Solution**

**Q1)**

>> X=[ 3.7 48.5 9.3; 5.7 65.1 8; 3.8 47.2 10.9;3.2 53.2 12.0;3.1 55.5 9.7; 4.6 36.1 7.9;2.4 24.8 14;7.2 33.1 7.6;6.7 47.4 8

.5;5.4 54.1 11.3;3.9 36.9 12.7;4.5 58.8 12.3;3.5 27.8 9.8;4.5 40.2 8.4;1.5 13.5 10.1;8.5 56.4 7.1;4.5 71.6 8.2;6.5 52.8 10.

9 ;4.1 44.1 11.2;5.5 40.9 9.4 ]

X =

3.7000 48.5000 9.3000

5.7000 65.1000 8.0000

3.8000 47.2000 10.9000

3.2000 53.2000 12.0000

3.1000 55.5000 9.7000

4.6000 36.1000 7.9000

2.4000 24.8000 14.0000

7.2000 33.1000 7.6000

6.7000 47.4000 8.5000

5.4000 54.1000 11.3000

3.9000 36.9000 12.7000

4.5000 58.8000 12.3000

3.5000 27.8000 9.8000

4.5000 40.2000 8.4000

1.5000 13.5000 10.1000

8.5000 56.4000 7.1000

4.5000 71.6000 8.2000

6.5000 52.8000 10.9000

4.1000 44.1000 11.2000

5.5000 40.9000 9.4000

>> e=[1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1]

e =

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

**a)Mean & variance of above data without assuming multivariate.**

>> mean=(1/20)\*X'\*e

mean =

4.6400

45.4000

9.9650

>> S=(1/19)\*(X'\*(eye(20)-(1/20)\*ones(20))\*X)

S =

2.8794 10.0100 -1.8091

10.0100 199.7884 -5.6400

-1.8091 -5.6400 3.6277

>> Sinv=inv(S)

Sinv =

5.8616e-01 -2.2086e-02 2.5797e-01

-2.2086e-02 6.0672e-03 -1.5809e-03

2.5797e-01 -1.5809e-03 4.0185e-01

**b)Correlation matrix for above data.**

>> D=[1/sqrt(S(1,1)) 0 0;0 1/sqrt(S(2,2)) 0 ; 0 0 1/sqrt(S(3,3))]

D =

0.5893 0 0

0 0.0707 0

0 0 0.5250

>> R=D\*S\*D

R =

1.0000 0.4173 -0.5597

0.4173 1.0000 -0.2095

-0.5597 -0.2095 1.0000

**c)Checking the three variates normality of above data –X1,X2 & X3.**

>> for i=1:20

di=(X(i,:)'-mean)'\*Sinv\*(X(i,:)'-mean)

end

di = 1.2117

di = 2.6902

di = 0.4408

di = 2.1828

di = 2.9433

di = 2.2047

di = 5.6197

di = 5.1822

di = 1.6444

di = 1.7087

di = 2.5167

di = 3.1072

di = 1.8537

di = 1.2149

di = 7.3311

di = 5.2845

di = 5.8638

di = 2.9788

di = 0.4241

di = 0.5969

>> for i=1:20

di=(X(i,:)'-mean)'\*Sinv\*(X(i,:)'-mean);

if di>2.366

d=di

end

end

d = 2.6902

d = 2.9433

d = 5.6197

d = 5.1822

d = 2.5167

d = 3.1072

d = 7.3311

d = 5.2845

d = 5.8638

d = 2.9788

**%%% Since ten of these distances are less than 2.366. A proportion 0.50 of the data fall within the 50% contour.**

**%%% This difference in proportion may ordinarily provide evidence for Accepting the assumption of multivariate normality.**

**Q2)**

>> X=[126974 4224 173297;96933 3835 160893;86656 3510 83219;63438 3758 77734;55264 3939 128344;50976 1809 39080;39069 2946

38528 ; 36156 359 51038;35209 2480 34715;32416 2413 25636]

X =

126974 4224 173297

96933 3835 160893

86656 3510 83219

63438 3758 77734

55264 3939 128344

50976 1809 39080

39069 2946 38528

36156 359 51038

35209 2480 34715

32416 2413 25636

>> e=[1;1;1;1;1;1;1;1;1;1]

e =

1

1

1

1

1

1

1

1

1

1

**a) Mean & variance of above data without assuming multivariate.**

>> mean=(1/10)\*X'\*e

mean =

6.2309e+04

2.9273e+03

8.1248e+04

>> S=(1/9)\*(X'\*(eye(10)-(1/10)\*ones(10))\*X)

S =

1.0005e+09 2.5576e+07 1.5118e+09

2.5576e+07 1.4300e+06 4.5655e+07

1.5118e+09 4.5655e+07 2.9805e+09

>> Sinv=inv(S)

Sinv =

4.4317e-09 -1.4663e-08 -2.0233e-09

-1.4663e-08 1.4171e-06 -1.4269e-08

-2.0233e-09 -1.4269e-08 1.5804e-09

**b) Correlation matrix for above data.**

>> D=[1/sqrt(S(1,1)) 0 0;0 1/sqrt(S(2,2)) 0 ; 0 0 1/sqrt(S(3,3))]

D =

3.1615e-05 0 0

0 8.3624e-04 0

0 0 1.8317e-05

>> R=D\*S\*D

R =

1.0000 0.6762 0.8755

0.6762 1.0000 0.6993

0.8755 0.6993 1.0000

**c) Checking the three variates normality of above data –X1,X2 & X3.**

>> for i=1:10

di=(X(i,:)'-mean)'\*Sinv\*(X(i,:)'-mean)

end

di = 4.3524

di = 2.3614

di = 2.4713

di = 1.0749

di = 5.3677

di = 1.5003

di = 1.2963

di = 6.4396

di = 0.9078

di = 1.2284

>> for i=1:10

di=(X(i,:)'-mean)'\*Sinv\*(X(i,:)'-mean);

if di>2.366

d=di

end

end

d = 4.3524

d = 2.4713

d = 5.3677

d = 6.4396

**%%% Since ten of these distances are less than 2.366. A proportion 0.60 of the data fall within the 50% contour.**

**%%% This difference in proportion may ordinarily provide evidence for Accepting the assumption of multivariate normality.**