ORIGIN ≡ 1

Multivariate Data: Summary Statistics and Diagnostic Plots

The first step in analysing multivariate data is input and calculation of summary measures. Often this is followed by graphical analysis of distributions and covariance of individual variables over multiple cases. For more complete information, see RA Johnson & DW Wichern *Applied Multivariate Statistical Analysis 4th Edition* 1998, or AC. Rencher *Methods of Multivariate Analysis* 1995.

Read Data:

M := READPRN("C:\DATA\Multivariate\iris.txt")

M =

	1	2	3	4	5
1	1	5.1	3.5	1.4	0.2
2	2	4.9	3	1.4	0.2
3	3	4.7	3.2	1.3	0.2
4	4	4.6	3.1	1.5	0.2
5	5	5	3.6	1.4	0.2
6	6	5.4	3.9	1.7	0.4
7	7	4.6	3.4	1.4	0.3
8	8	5	3.4	1.5	0.2
9	9	4.4	2.9	1.4	0.2
10	10	4.9	3.1	1.5	0.1
11	11	5.4	3.7	1.5	0.2
12	12	4.8	3.4	1.6	0.2
13	13	4.8	3	1.4	0.1
14	14	4.3	3	1.1	0.1
15	15	5.8	4	1.2	0.2
16	16	5.7	4.4	1.5	0.4
17	17	5.4	3.9	1.3	0.4
18	18	5.1	3.5	1.4	0.3
19	19	5.7	3.8	1.7	0.3
20	20	5.1	3.8	1.5	0.3
21	21	5.4	3.4	1.7	0.2
22	22	5.1	3.7	1.5	0.4
23	23	4.6	3.6	1	0.2
24	24	5.1	3.3	1.7	0.5

Prototype in R:

Anderson's Iris dataset - 50 cases for each of 3 species:

```
#MULTIVARIATE DATA SUMMARY STATISTICS & DIAGNOSTICS
#READ DATA:
M=read.table("c:/DATA/Multivariate/iris.txt")
M
```

```
>M
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
…
51           7.0         3.2          4.7         1.4 versicolor
52           6.4         3.2          4.5         1.5 versicolor
53           6.9         3.1          4.9         1.5 versicolor
54           5.5         2.3          4.0         1.3 versicolor
55           6.5         2.8          4.6         1.5 versicolor
56           5.7         2.8          4.5         1.3 versicolor
57           6.3         3.3          4.7         1.6 versicolor
58           4.9         2.4          3.3         1.0 versicolor
59           6.6         2.9          4.6         1.3 versicolor
60           5.2         2.7          3.9         1.4 versicolor
…
101          6.3         3.3          6.0         2.5  virginica
102          5.8         2.7          5.1         1.9  virginica
103          7.1         3.0          5.9         2.1  virginica
104          6.3         2.9          5.6         1.8  virginica
105          6.5         3.0          5.8         2.2  virginica
106          7.6         3.0          6.6         2.1  virginica
107          4.9         2.5          4.5         1.7  virginica
108          7.3         2.9          6.3         1.8  virginica
109          6.7         2.5          5.8         1.8  virginica
110          7.2         3.6          6.1         2.5  virginica
...
```

Reading data is highly efficient in R allowing for both numeric values and names of factor levels (Species) as well as labels for each variable (column). I typically create the first unlabeled column of sequential numbers (for each case) in Excel prior to inputting into R. MathCad, by comparison reads only numbers, requiring more work (only the first 24 cases shown).

In multivariate statistics, a dataset is a case by variable matrix. Because the number of cases often exceeds the number of variables, cases are commonly placed in rows and variables in columns. In addition, some variables (such as the variable Species here) may serve as classifiers allowing for different meaningful groups of cases. This is termed the "long form" data structure and is standard in computer-based data analysis. Whereas long-form data structure is common, it is important to note that textbooks in multivariate statistics typically treat each case in the dataset (row) as a *column* vector, so the dataset as a whole enters into published matrix algebra in transposed form. By contrast, standard software packages, including R, expect to see datasets in the untransposed long-form.

Indices:

$M_{5,3} = 3.6$

Because "iris.txt" contains a first column of numbers convient for data input in R as row labels, I have to strip matrix M here in order to match output.

$M := submatrix(M, 1, 150, 2, 5)$

$M_{5,2} = 3.6$ **^ stripping the first column from M**

```
> #INDICES:
> M[5,2]
[1] 3.6
```

Summary Statistics:

Matrix Dimensions:

$n := rows(M) = (150)$

$p := cols(M) = (4)$

```
> #MATRIX DIMENSIONS:
> n=nrow(M)
>n
[1] 150
> p=ncol(M)
>p
[1] 5
```

Mean Vector:

$i := 1 .. 150$ $j := 1 .. 4$ **< index variables**

$$Mbar_j := \frac{1}{n} \cdot \sum_i M_{i,j}$$

$$Mbar = \begin{pmatrix} 5.843 \\ 3.057 \\ 3.758 \\ 1.199 \end{pmatrix}$$

```
> summary(M)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

```
#MEAN VECTOR
M=subset(M,select=c("Sepal.Length","Sepal.Width","Petal.Length","Petal.Width"))
> Mbar=mean(M)
```

In Matrix Algebra Form: **^ stripping "Species" column from matrix M**

$lvec_i := 1$ **< I_{vec} is the identity vector (vector of 1's with n rows)**

$$Mbar := \frac{1}{n} \cdot M^T \cdot lvec \qquad Mbar = \begin{pmatrix} 5.843 \\ 3.057 \\ 3.758 \\ 1.199 \end{pmatrix}$$

```
> Mbar
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
    5.843333     3.057333     3.758000     1.199333
```

Covariance Matrix:

$k := 1 .. 4$

$$S_{j,k} := \frac{1}{n-1} \cdot \sum_i \left(M_{i,j} - Mbar_j\right) \cdot \left(M_{i,k} - Mbar_k\right)$$

```
> #COVARIANCE MATRIX:
> cov(M)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
```

$$S = \begin{pmatrix} 0.6856935 & -0.042434 & 1.2743154 & 0.5162707 \\ -0.042434 & 0.1899794 & -0.3296564 & -0.1216394 \\ 1.2743154 & -0.3296564 & 3.1162779 & 1.2956094 \\ 0.5162707 & -0.1216394 & 1.2956094 & 0.5810063 \end{pmatrix}$$

In Matrix Algebra Form:

$I := identity(n)$ **< I is the identity matrix size n**

$$S := \frac{1}{n-1} \cdot M^T \cdot \left(I - \frac{1}{n} \cdot lvec \cdot lvec^T\right) \cdot M \qquad S = \begin{pmatrix} 0.6856935 & -0.042434 & 1.2743154 & 0.5162707 \\ -0.042434 & 0.1899794 & -0.3296564 & -0.1216394 \\ 1.2743154 & -0.3296564 & 3.1162779 & 1.2956094 \\ 0.5162707 & -0.1216394 & 1.2956094 & 0.5810063 \end{pmatrix}$$

Correlation Matrix:

In Matrix Algebra Form:

$$D_{j,j} := \sqrt{S_{j,j}}$$

$$D = \begin{pmatrix} 0.828 & 0 & 0 & 0 \\ 0 & 0.436 & 0 & 0 \\ 0 & 0 & 1.765 & 0 \\ 0 & 0 & 0 & 0.762 \end{pmatrix}$$

> #CORRELATION MATRIX:
> cor(M)

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

$$R := D^{-1} \cdot S \cdot D^{-1}$$

$$R = \begin{pmatrix} 1 & -0.1175698 & 0.8717538 & 0.8179411 \\ -0.1175698 & 1 & -0.4284401 & -0.3661259 \\ 0.8717538 & -0.4284401 & 1 & 0.9628654 \\ 0.8179411 & -0.3661259 & 0.9628654 & 1 \end{pmatrix}$$

Standardized Data Matrix:

$$M_{S_{i,j}} := \frac{M_{i,j} - M_{bar_j}}{\sqrt{S_{j,j}}}$$

$M_S =$

	1	2	3	4
1	-0.898	1.016	-1.336	-1.311
2	-1.139	-0.132	-1.336	-1.311
3	-1.381	0.327	-1.392	-1.311
4	-1.501	0.098	-1.279	-1.311
5	-1.018	1.245	-1.336	-1.311
6	-0.535	1.933	-1.166	-1.049
7	-1.501	0.786	-1.336	-1.18
8	-1.018	0.786	-1.279	-1.311
9	-1.743	-0.361	-1.336	-1.311
10	-1.139	0.098	-1.279	-1.442
11	-0.535	1.474	-1.279	-1.311
12	-1.26	0.786	-1.222	-1.311
13	-1.26	-0.132	-1.336	-1.442
14	-1.864	-0.132	-1.506	-1.442
15	-0.052	2.163	-1.449	-1.311
16	-0.173	3.08	-1.279	-1.049

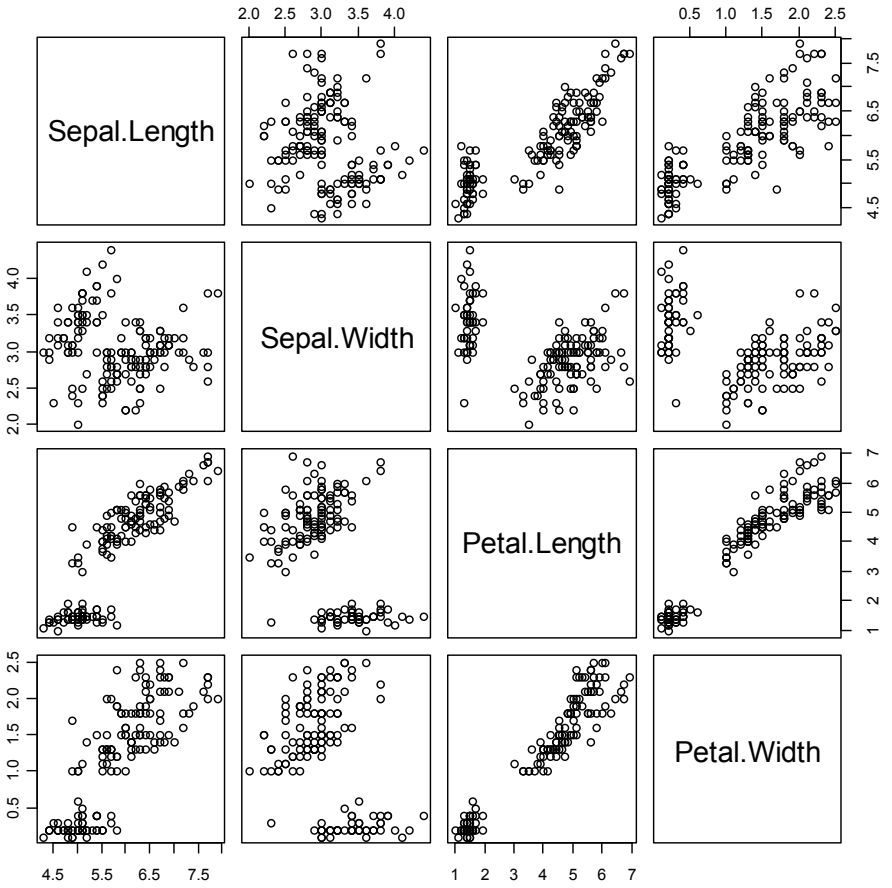
Standardized data subtracts the mean from each observation and divides by the standard deviation. Standardization differs from Normalization. The latter scales data values to range between 0 and 1.

> #STANDARDIZED DATA MATRIX:
> Ms=scale(M)
> Ms

```
    Sepal.Length Sepal.Width Petal.Length  Petal.Width
1     -0.89767388  1.01560199  -1.33575163 -1.3110521482
2     -1.13920048 -0.13153881  -1.33575163 -1.3110521482
3     -1.38072709  0.32731751  -1.39239929 -1.3110521482
4     -1.50149039  0.09788935  -1.27910398 -1.3110521482
5     -1.01843718  1.24503015  -1.33575163 -1.3110521482
6     -0.53538397  1.93331463  -1.16580868 -1.0486667950
7     -1.50149039  0.78617383  -1.33575163 -1.1798594716
8     -1.01843718  0.78617383  -1.27910398 -1.3110521482
9     -1.74301699 -0.36096697  -1.33575163 -1.3110521482
10    -1.13920048  0.09788935  -1.27910398 -1.4422448248
…
51     1.39682886  0.32731751   0.53362088  0.2632599711
52     0.67224905  0.32731751   0.42032558  0.3944526477
53     1.27606556  0.09788935   0.64691619  0.3944526477
54    -0.41462067 -1.73753594   0.13708732  0.1320672944
55     0.79301235 -0.59039513   0.47697323  0.3944526477
56    -0.17309407 -0.59039513   0.42032558  0.1320672944
57     0.55148575  0.55674567   0.53362088  0.5256453243
58    -1.13920048 -1.50810778  -0.25944625 -0.2615107354
59     0.91377565 -0.36096697   0.47697323  0.1320672944
60    -0.77691058 -0.81982329   0.08043967  0.2632599711
…
101    0.55148575  0.55674567   1.27004036  1.7063794137
102   -0.05233076 -0.81982329   0.76021149  0.9192233541
103    1.51759216 -0.13153881   1.21339271  1.1816087073
104    0.55148575 -0.36096697   1.04344975  0.7880306775
105    0.79301235 -0.13153881   1.15674505  1.3128013839
106    2.12140867 -0.13153881   1.60992627  1.1816087073
107   -1.13920048 -1.27867961   0.42032558  0.6568380009
108    1.75911877 -0.36096697   1.43998331  0.7880306775
109    1.03453895 -1.27867961   1.15674505  0.7880306775
110    1.63835547  1.24503015   1.32668801  1.7063794137
…
attr(,"scaled:center")
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
    5.843333     3.057333     3.758000     1.199333
attr(,"scaled:scale")
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
   0.8280661    0.4358663    1.7652982    0.7622377
```

Diagnostic Plots:

```
> plot(M)
```



```
> library(car)
> scatterplotMatrix(M)
```