Summary & Graphic Display of Data

ORIGIN ≡ 0

It is standard practice to summarize data with *descriptive statistics*, and to display a dataset in the form of *anotated graphs* prior to, or supplementing, *inferential statistics* such as employing a statistical test. Since one tends to do this frequently, it is very useful to set up a page as here in MathCad, or a *script* in a statistical program such as R, that does things automatically in a preferred format using previously prototyped functions.

iris := READPRN("c:/DATA/Biostatistics/iris.txt") **< Input iris same dataset as before**

$SL := iris^{\langle 1 \rangle}$ **< Sepal Length**
$SW := iris^{\langle 2 \rangle}$ **< Sepal Width**
$PL := iris^{\langle 3 \rangle}$ **< Petal Length**
$PW := iris^{\langle 4 \rangle}$ **< Petal Width**

< Labeled Column variables

n := length(SL) n = 150

^ n = number of replicates

Evaluation of variable iris. >

iris =

	0	1	2	3	4
0	1	5.1	3.5	1.4	0.2
1	2	4.9	3	1.4	0.2
2	3	4.7	3.2	1.3	0.2
3	4	4.6	3.1	1.5	0.2
4	5	5	3.6	1.4	0.2
5	6	5.4	3.9	1.7	0.4
6	7	4.6	3.4	1.4	0.3
7	8	5	3.4	1.5	0.2
8	9	4.4	2.9	1.4	0.2
9	10	4.9	3.1	1.5	0.1
10	11	5.4	3.7	1.5	0.2
11	12	4.8	3.4	1.6	0.2
12	13	4.8	3	1.4	0.1
13	14	4.3	3	1.1	0.1
14	15	5.8	4	1.2	0.2
15	16	5.7	4.4	1.5	0.4

Descriptive Statistics in Matrix format:

means:

$$Means := \begin{pmatrix} mean(SL) \\ mean(SW) \\ mean(PL) \\ mean(PW) \end{pmatrix} \qquad Means = \begin{pmatrix} 5.843 \\ 3.057 \\ 3.758 \\ 1.199 \end{pmatrix}$$

medians:

$$Medians := \begin{pmatrix} median(SL) \\ median(SW) \\ median(PL) \\ median(PW) \end{pmatrix} \qquad Medians = \begin{pmatrix} 5.8 \\ 3 \\ 4.35 \\ 1.3 \end{pmatrix}$$

sample variances:

$$Variances_s := \frac{n}{n-1} \cdot \begin{pmatrix} var(SL) \\ var(SW) \\ var(PL) \\ var(PW) \end{pmatrix} \qquad Variances_s = \begin{pmatrix} 0.6856935 \\ 0.1899794 \\ 3.1162779 \\ 0.5810063 \end{pmatrix}$$

sample standard deviations:

$$SD_s := \sqrt{Variances_s} \qquad SD_s = \begin{pmatrix} 0.828 \\ 0.436 \\ 1.765 \\ 0.762 \end{pmatrix} \qquad \begin{pmatrix} \sqrt{0.686} \\ \sqrt{0.19} \\ \sqrt{3.116} \\ \sqrt{0.581} \end{pmatrix} = \begin{pmatrix} 0.828 \\ 0.436 \\ 1.765 \\ 0.762 \end{pmatrix}$$

< square root function for matrix verified.

range:

$$Mins := \begin{pmatrix} min(SL) \\ min(SW) \\ min(PL) \\ min(PW) \end{pmatrix} \quad Mins = \begin{pmatrix} 4.3 \\ 2 \\ 1 \\ 0.1 \end{pmatrix} \quad Maxs := \begin{pmatrix} max(SL) \\ max(SW) \\ max(PL) \\ max(PW) \end{pmatrix} \quad Maxs = \begin{pmatrix} 7.9 \\ 4.4 \\ 6.9 \\ 2.5 \end{pmatrix}$$

< Note: Prototype for these functions can be verified by examining each column of the iris dataset.

^ using MathCad minimum and maximum functions.

Quantiles:

Quantiles are set fractions of the dataset. Commonly reported quantiles include:

The Median (and median value) - where half of the replicate observations have values above and/or below the median value.
Quartiles (and quartile values) - where the replicate observations are ranked and divided into 4 bins of equal count
Percentiles (and percentile values) - where replicate observations are ranked and divided into 100 bins of equal count

if the dataset has a total number of replicate observations that do not divide evenly into equal bins, then some rule is usually employed by hand or in software to set nearest boundaries. Different approaches to what is, in essence, an arbitrary decision sometimes yield slightly different results. Of course, the larger the dataset, the less of a problem this becomes.

Quartiles:

$SL_{sort} := \text{sort}(SL)$ $n = 150$

^ number of replicates

$$Q := \begin{pmatrix} \frac{n}{4} \\ \frac{n}{2} \\ n \cdot \frac{3}{4} \\ n \end{pmatrix} \quad Q = \begin{pmatrix} 37.5 \\ 75 \\ 112.5 \\ 150 \end{pmatrix} \quad \text{Index} := \begin{pmatrix} 37 \\ 75 \\ 112 \\ 149 \end{pmatrix} \quad \begin{pmatrix} 37-0 \\ 74-37 \\ 111-74 \\ 149-111 \end{pmatrix} = \begin{pmatrix} 37 \\ 37 \\ 37 \\ 38 \end{pmatrix}$$

^ bins **^ index starts at 0** **^ count in each bin**

$SL_{sort} =$

	0
0	4.3
1	4.4
2	4.4
3	4.4
4	4.5
5	4.6
6	4.6
7	4.6
8	4.6
9	4.7
10	4.7
11	4.8
12	4.8
13	4.8
14	4.8
15	4.8

quartile data values:

$$\begin{pmatrix} SL_{sort_{37}} \\ SL_{sort_{75}} \\ SL_{sort_{112}} \\ SL_{sort_{149}} \end{pmatrix} = \begin{pmatrix} 5.1 \\ 5.8 \\ 6.4 \\ 7.9 \end{pmatrix}$$

< Quartile data values may be plotted along with mean, median, etc.

Deciles & Percentiles

$$\begin{pmatrix} n \cdot 0.1 \\ n \cdot 0.2 \\ n \cdot 0.3 \\ n \cdot 0.4 \\ n \cdot 0.5 \\ n \cdot 0.6 \\ n \cdot 0.7 \\ n \cdot 0.8 \\ n \cdot 0.9 \\ n \end{pmatrix} = \begin{pmatrix} 15 \\ 30 \\ 45 \\ 60 \\ 75 \\ 90 \\ 105 \\ 120 \\ 135 \\ 150 \end{pmatrix} \quad \text{Index} := \begin{pmatrix} 14 \\ 29 \\ 44 \\ 59 \\ 74 \\ 89 \\ 104 \\ 119 \\ 134 \\ 149 \end{pmatrix} \quad \begin{pmatrix} SL_{sort_{14}} \\ SL_{sort_{29}} \\ SL_{sort_{44}} \\ SL_{sort_{59}} \\ SL_{sort_{74}} \\ SL_{sort_{89}} \\ SL_{sort_{104}} \\ SL_{sort_{119}} \\ SL_{sort_{134}} \\ SL_{sort_{149}} \end{pmatrix} = \begin{pmatrix} 4.8 \\ 5 \\ 5.2 \\ 5.6 \\ 5.8 \\ 6.1 \\ 6.3 \\ 6.5 \\ 6.9 \\ 7.9 \end{pmatrix}$$

< Decile data values (10 percentiles or units of 10%) may be plotted along with mean, median, etc.

coefficient of variation:

$$CVs := \begin{pmatrix} \dfrac{SDs_0}{Means_0} \\ \dfrac{SDs_1}{Means_1} \\ \dfrac{SDs_2}{Means_2} \\ \dfrac{SDs_3}{Means_3} \end{pmatrix} \qquad CVs = \begin{pmatrix} 0.142 \\ 0.143 \\ 0.47 \\ 0.636 \end{pmatrix}$$

< Sample Standard Deviation divided by mean for each variable

Note: individual values in a matrix are called by index starting with 0 since ORIGIN is globally defined above as zero.

Prototype in R:

```
#LOAD DATA:
Iris=read.table("iris.txt")
#DESCRIPTIVE STATISTICS:
summary(Iris)
I=Iris[,1:4]
I
quantile(I$Sepal.Length,probs=seq(0,1,0.1))
variances=diag(var(I))
variances
sd=sqrt(variances)
sd
coefvar=sd/colMeans(I)
coefvar
```

```
> summary(Iris)
  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
```

```
> quantile(I$Sepal.Length,probs=seq(0,1,0.1))
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
4.30 4.80 5.00 5.27 5.60 5.80 6.10 6.30 6.52 6.90 7.90
```

```
> variances
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
   0.6856935    0.1899794    3.1162779    0.5810063
```

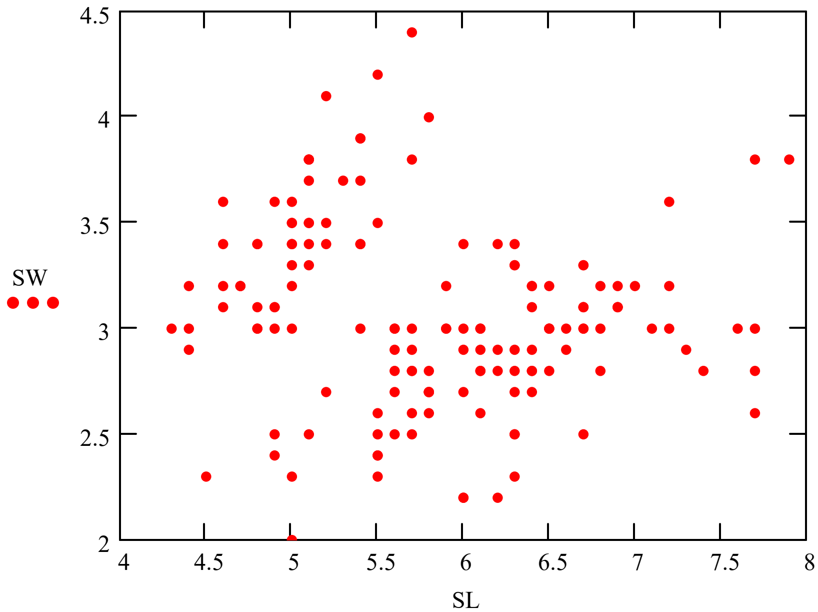
```
> sd
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
   0.8280661    0.4358663    1.7652982    0.7622377
```

```
> coefvar
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
   0.1417113    0.1425642    0.4697441    0.6355511
```

Scatter Plots:

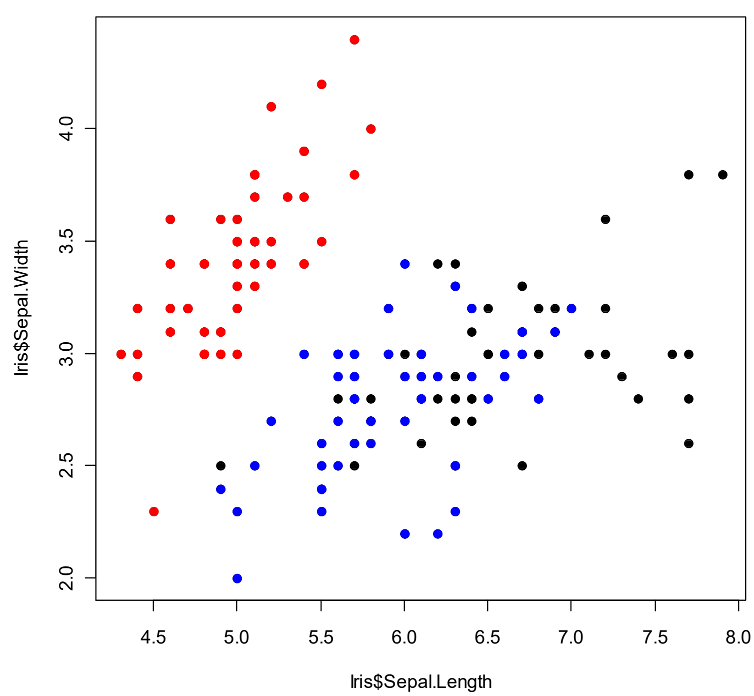
2-D scatter plot:

Any pair of variables may be plotted to look for patterns...



Prototype in R:

```
#SCATTER PLOTS:
plot(Iris)
plot(Iris$Sepal.Length,Iris$Sepal.Width,col='red,pch=19)
plot(Iris$Sepal.Length,Iris$Sepal.Width,pch=19)
points(Iris$Sepal.Length[Iris$Species=='setosa'],Iris$Sepal.Width[Iris$Species=='setosa'],col='red',pch=19)
points(Iris$Sepal.Length[Iris$Species=='versicolor'],Iris$Sepal.Width[Iris$Species=='versicolor'],col='blue',pch=19)
```

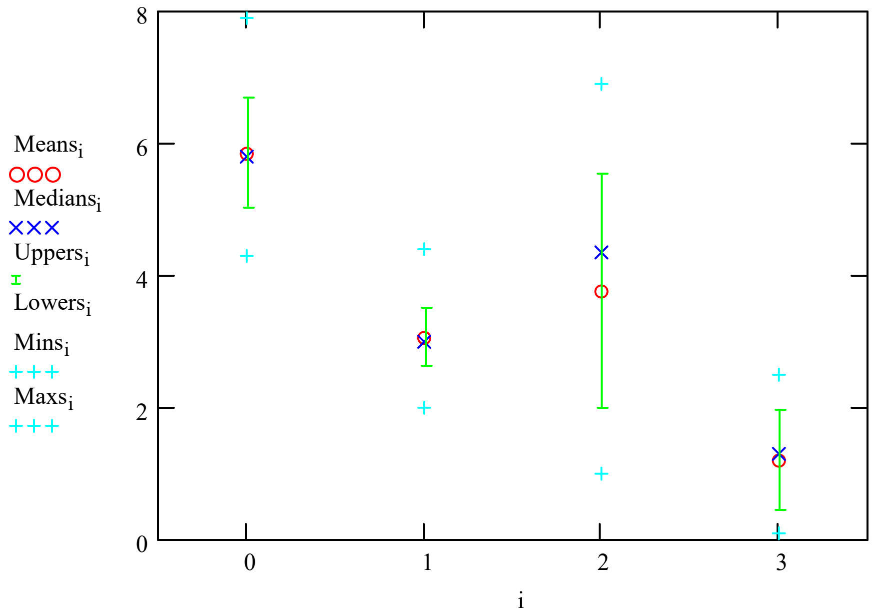


Stem & Leaf Plots and Box plots:

$\text{Uppers} := \text{Means} + \text{SD}_s$

$\text{Lowers} := \text{Means} - \text{SD}_s$

$i := 0 .. 3$



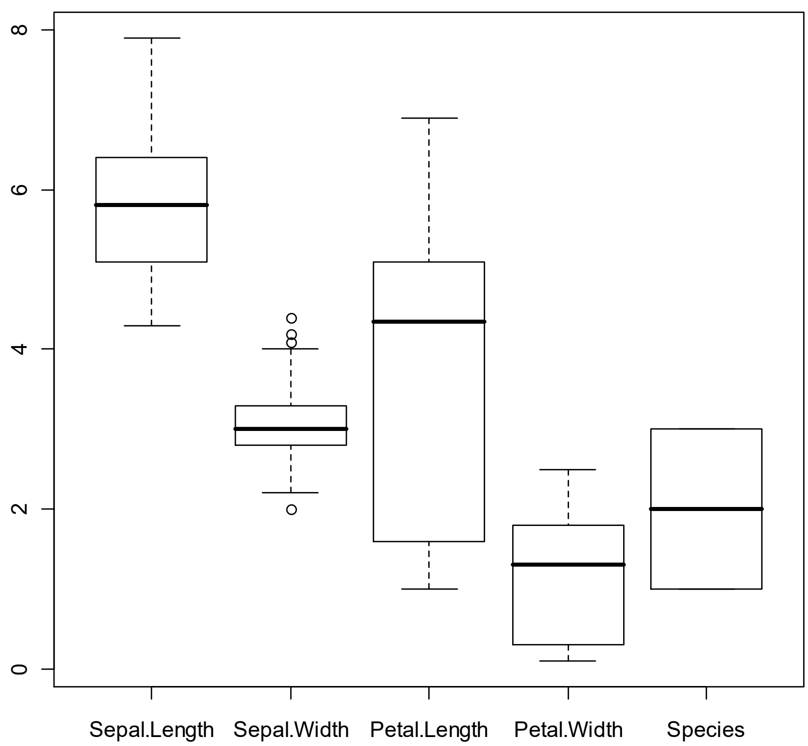
$$\text{CVs} = \begin{pmatrix} 0.142 \\ 0.143 \\ 0.47 \\ 0.636 \end{pmatrix}$$

Plots of this kind generally require a more sophisticated graphics system more directly related to statistical analysis than MathCad worksheets. See R!

Prototype in R:

```
#BOX PLOTS:
plot(Iris$Species,Iris$Sepal.Length)

library(lattice)
boxplot(Iris)
```



Histograms:

$plot_{SL} := histogram(10, SL)$

$plot_{SW} := histogram(10, SW)$

$plot_{PL} := histogram(10, PL)$

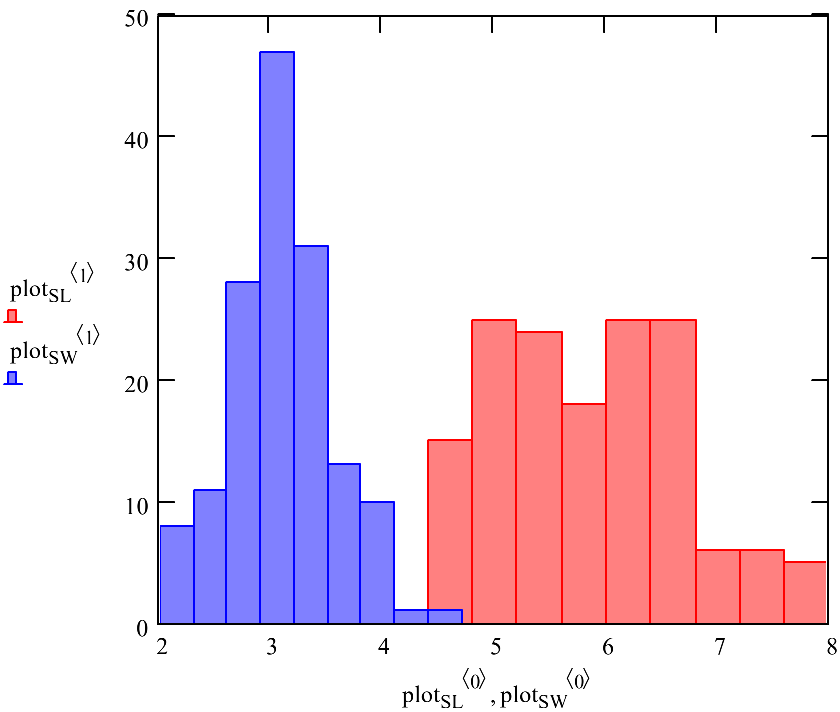
$plot_{PW} := histogram(10, PW)$

< variable plot contains two columns:

column 0: x axis = number of bins
column 1: y axis = count in each bin

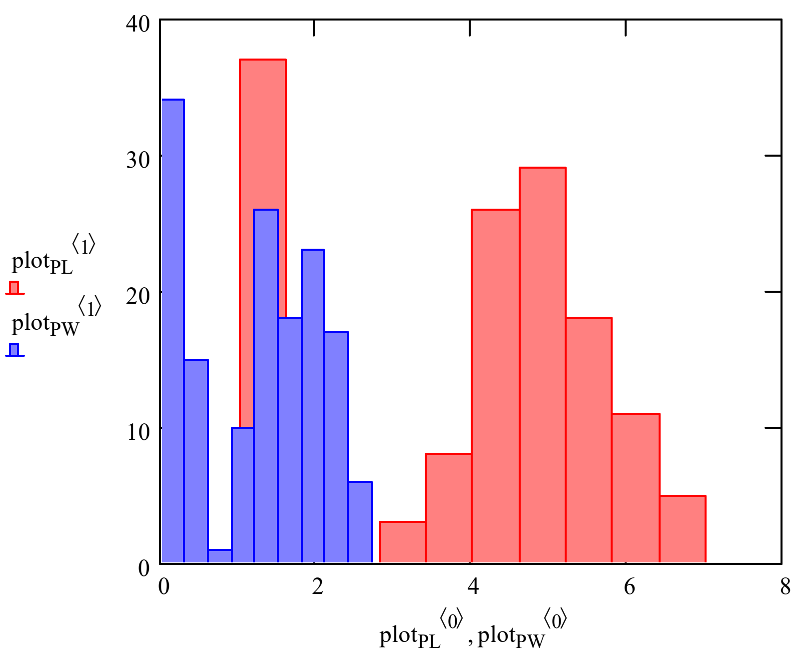
$plot_{SL} =$

	0	1
0	4.2	1
1	4.6	15
2	5	25
3	5.4	24
4	5.8	18
5	6.2	25
6	6.6	25
7	7	6
8	7.4	6
9	7.8	5



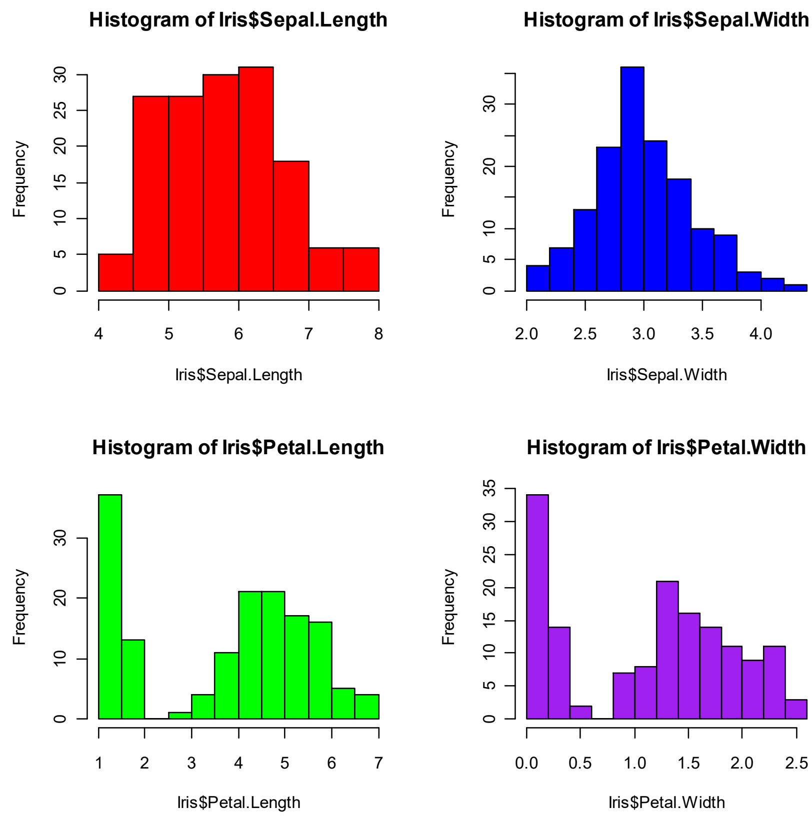
$plot_{SW} =$

	0	1
0	2.15	8
1	2.45	11
2	2.75	28
3	3.05	47
4	3.35	31
5	3.65	13
6	3.95	10
7	4.25	1
8	4.55	1
9	4.85	0



Prototype in R:

```
#HISTOGRAMS:
op=par(mfrow=c(2, 2))
hist(Iris$Sepal.Length,col='red')
hist(Iris$Sepal.Width,col='blue')
hist(Iris$Petal.Length,col='green')
hist(Iris$Petal.Width,col='purple')
par(op)
```




```
histogram(~ Sepal.Length | Species,data=Iris)
histogram(~ Sepal.Width | Species,data=Iris)
histogram(~ Petal.Length | Species,data=Iris)
histogram(~ Petal.Width | Species,data=Iris)
```