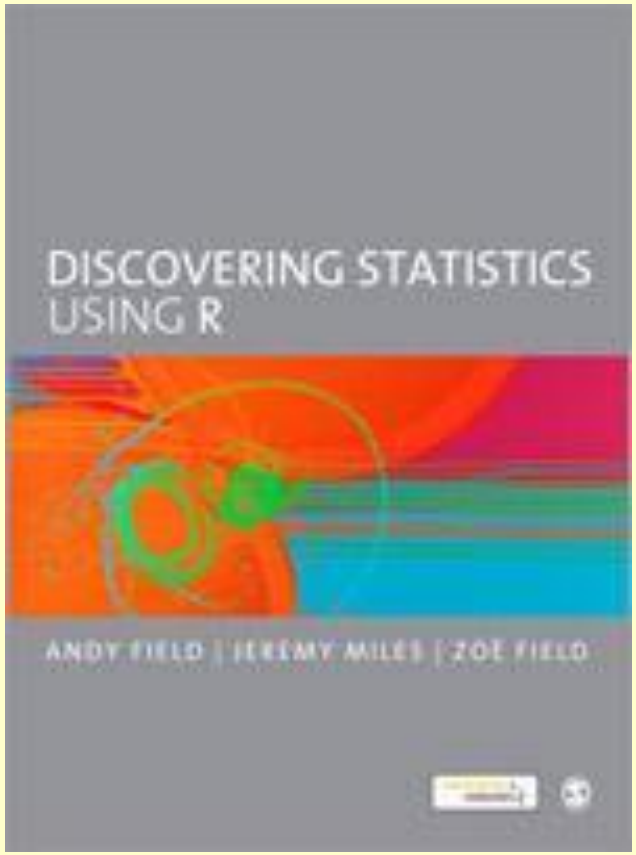
Graphical Exploration I:

R Commander

http://www.pelagicos.net/classes_biometry_fa17.htm

Exploring Data With Graphs



- Aim: Provide overview of Rcmdr graphs

- The Basics:
  - Histograms
  - Density plots
  - Boxplots
  - Scatterplots
  - Line graphs

Chapter 4

Graphs in Rcmdr

Package Rcmdr provides basic graphing tools (Fox 2005):

Journal of Statistical Software

September 2005, Volume 14, Issue 9. *http://www.jstatsoft.org/*

The R Commander: A Basic-Statistics Graphical User Interface to R

John Fox
McMaster University

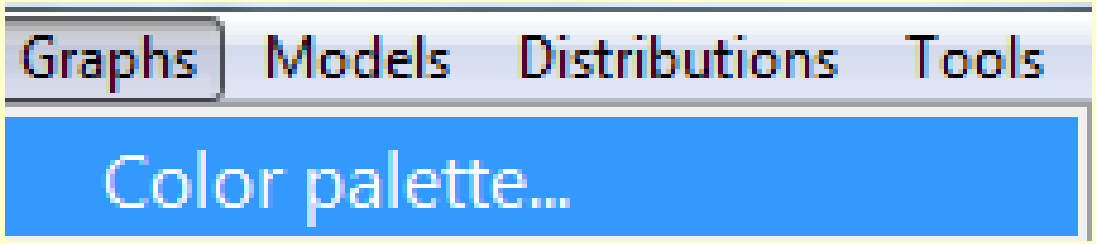
- Index plot
- Histogram
- Density plot
- Box plot
- Q-Q plots
- Scatterplots
- Line plots
- 3D Scatterplots

Remember: The help files for the current version of the Rcmdr package are available on the CRAN website at http://CRAN.R-project.org/doc/packages/Rcmdr.pdf.

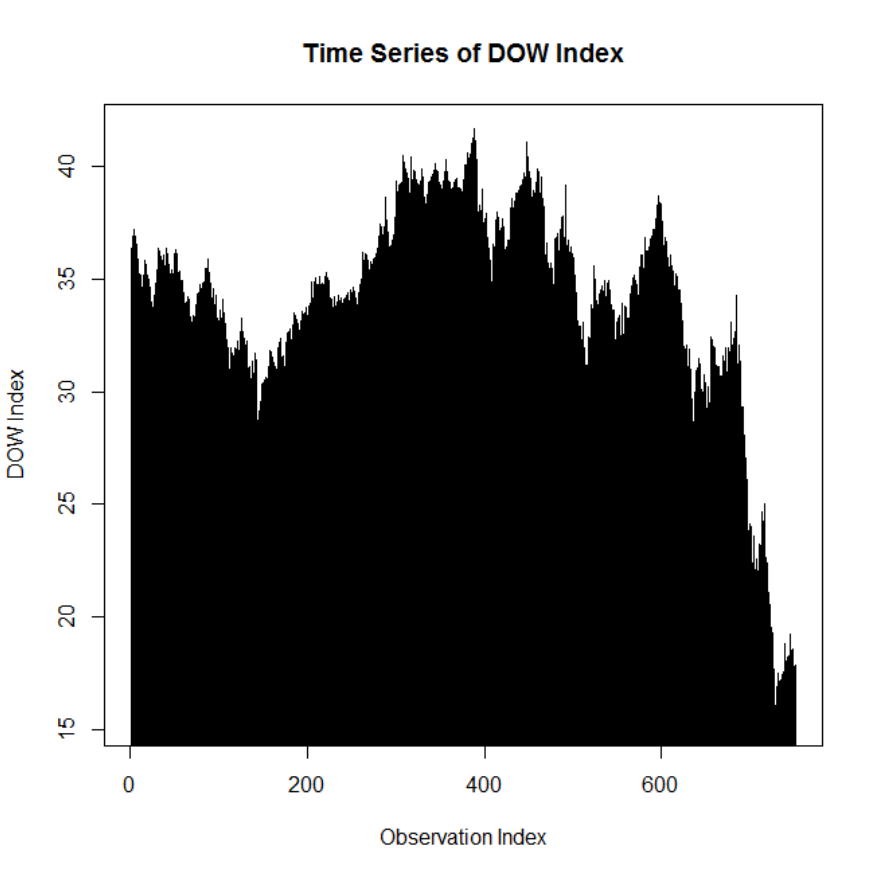
Graphs in Rcmdr – Palette

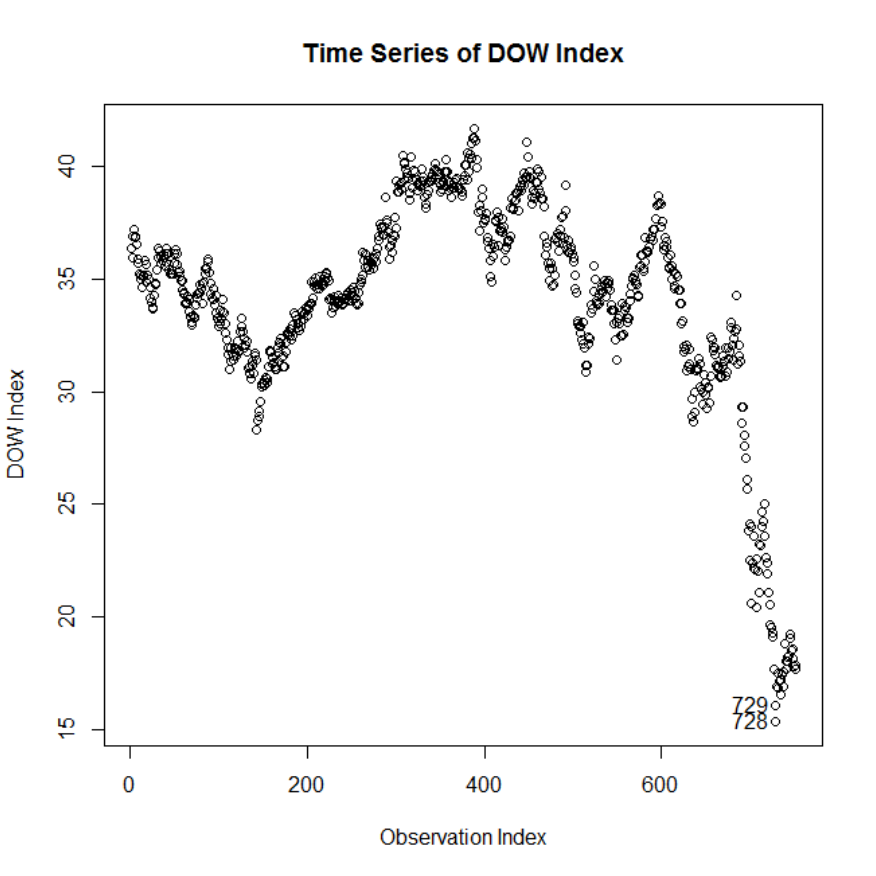
Graphs Tab Menu

Change the color options used in the figures



Graphs in Rcmdr – Index Plot





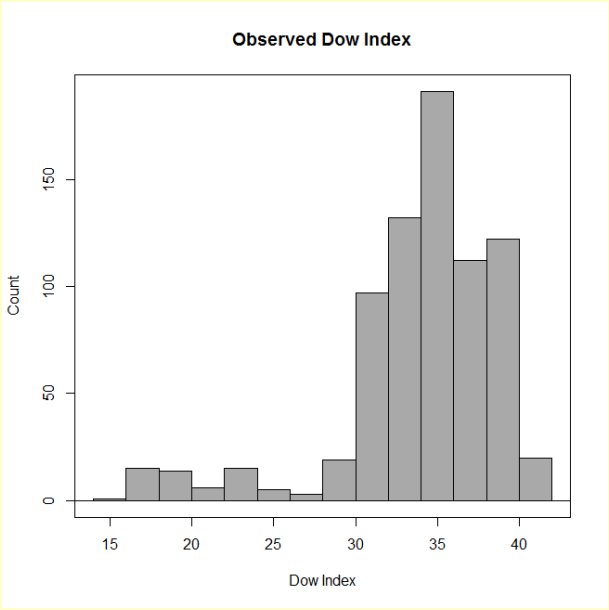
Plot all data values (as spikes or dots) sequentially

Graphs in Rcmdr – Histogram

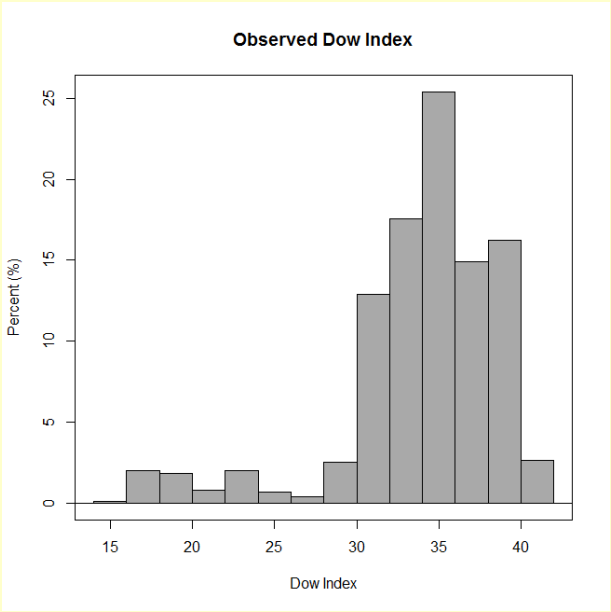
Plot all data values (binned) in histograms

Define the number of bins (or use "auto" option)

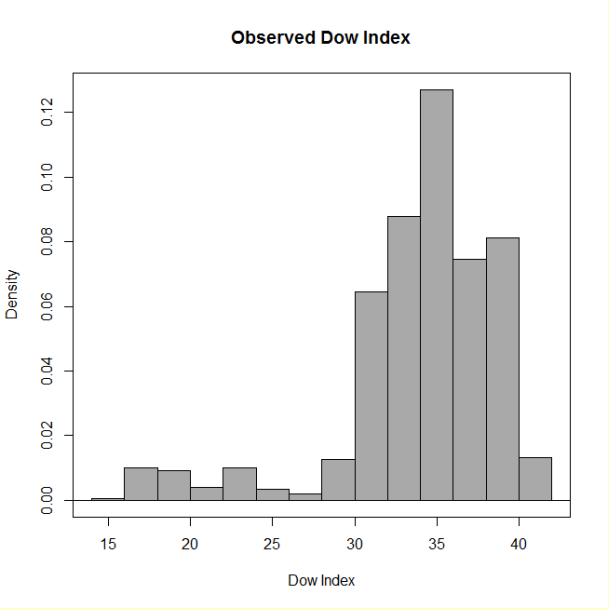
Count



Percent



Density

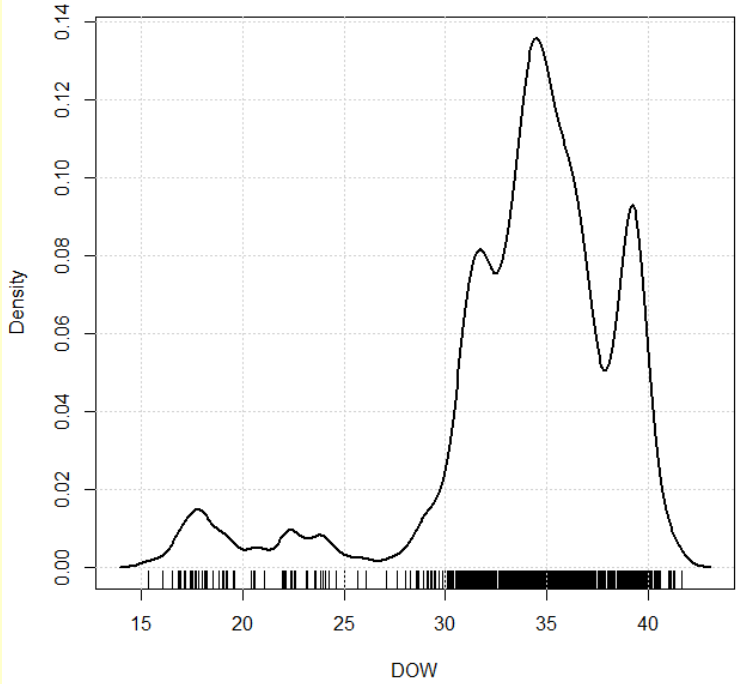


Frequency

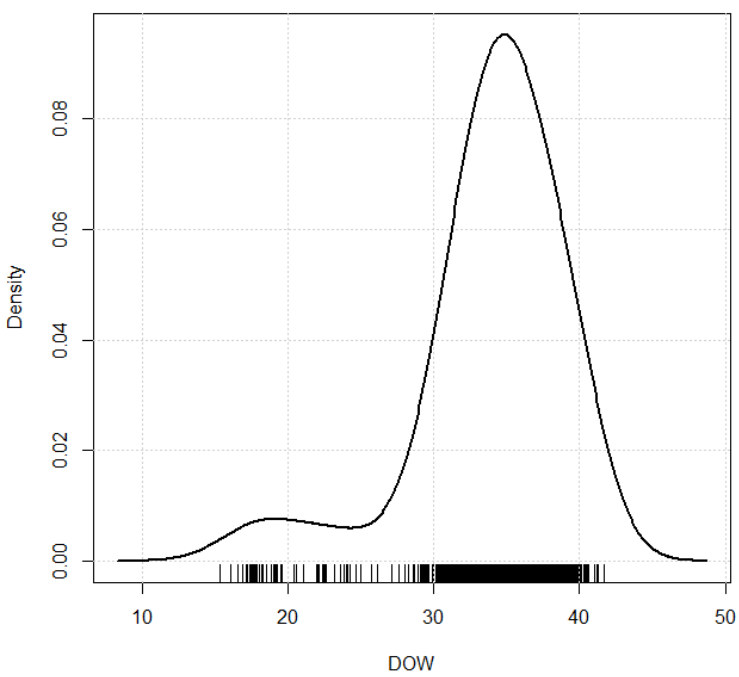
Graphs in Rcmdr – Density

Plot a non-parametric estimate of the data, based on different methods and various smoothing parameters

Bandwidth = 1



Bandwidth = 5

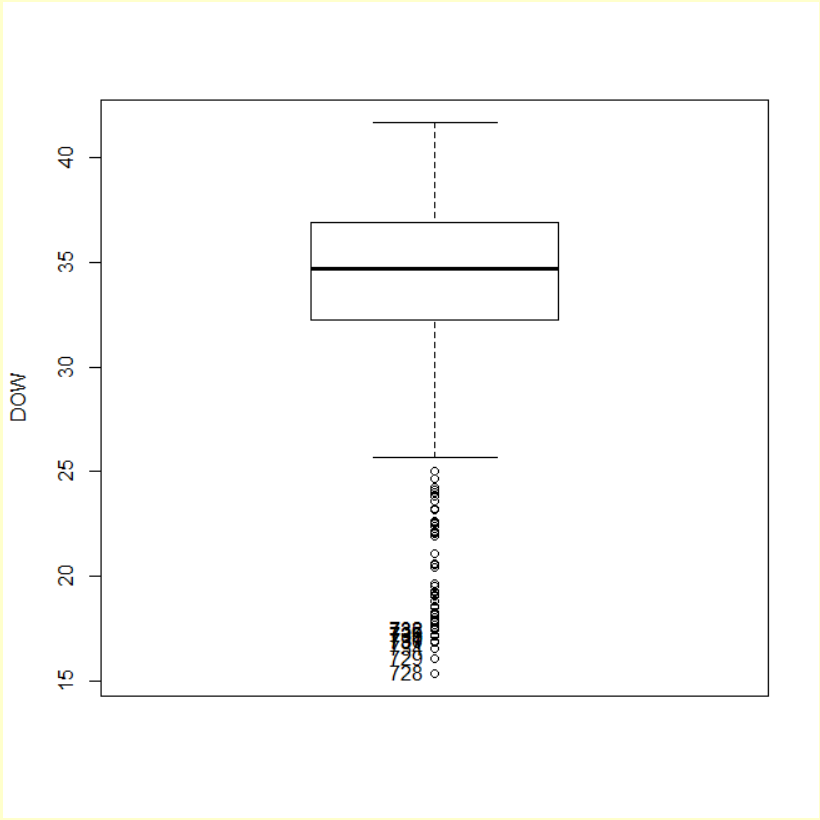


The Rug = shows location of observed X values

Graphs in Rcmdr – Box Plots

Plots quantiles of the distribution: 5, 25, 50, 75, 95

Identifies outliers: values too far from the median



Identify Points: Automatic or Interactively

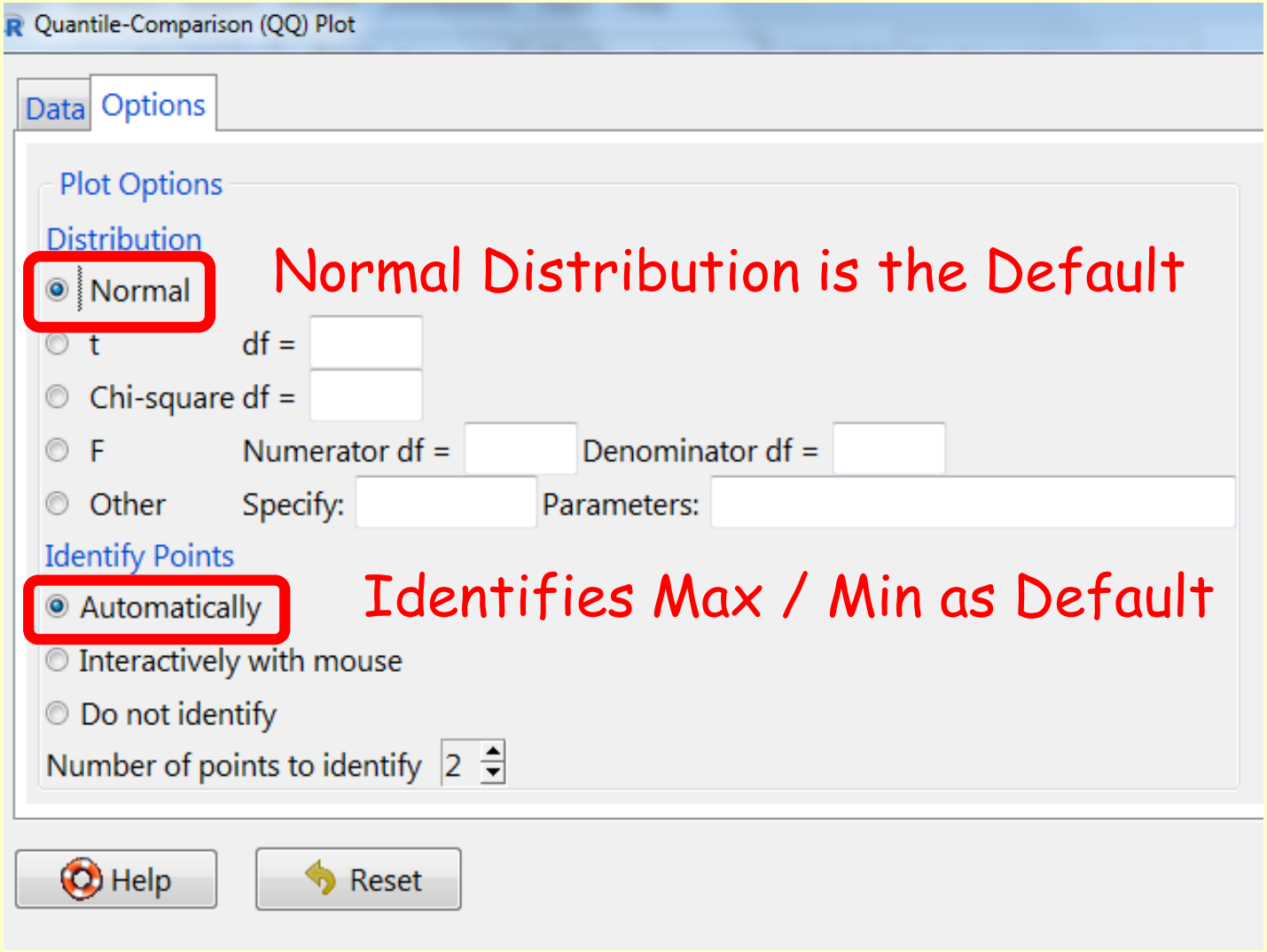
Graphs in Rcmdr – Quantiles

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. It shows the quantiles of the first data set against the quantiles of the second data set.

By a quantile, we mean the percent of data points below the given value. That is, the 30% quantile is the point at which 30% percent of the data fall below and 70% fall above that value. The median is the 50% quantile.

If the two sets come from a population with the same distribution, the points will fall along a 45 degree line. The greater the departure from this reference line, the stronger the evidence that the two data sets come from populations with different distributions.

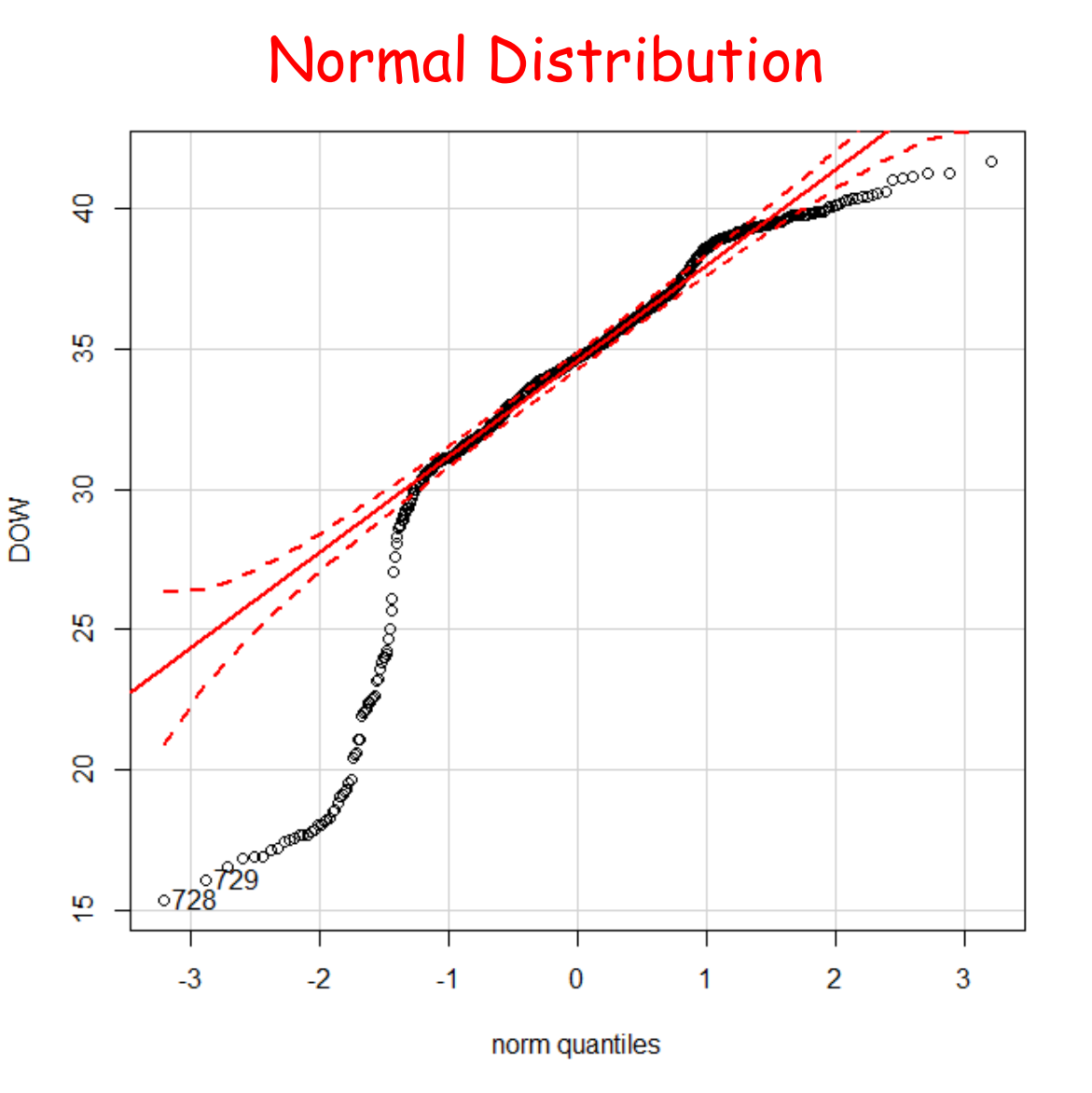
Graphs in Rcmdr – Quantiles



Graphically compares an observed (empirical) distribution (points) with a chosen theoretical expectation (line)

Identify Points: Automatic or Interactively

Graphs in Rcmdr – Quantiles



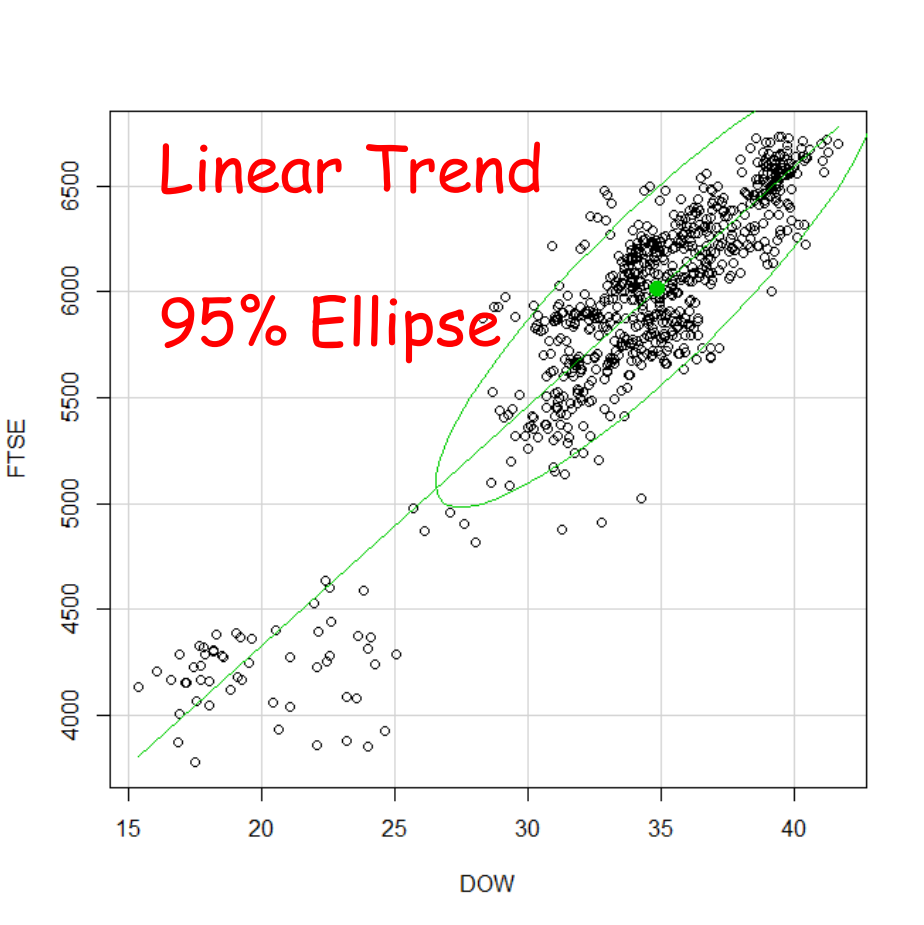
The solid red line is the expected pattern a normal distribution with the same mean and SD and the sampled data.

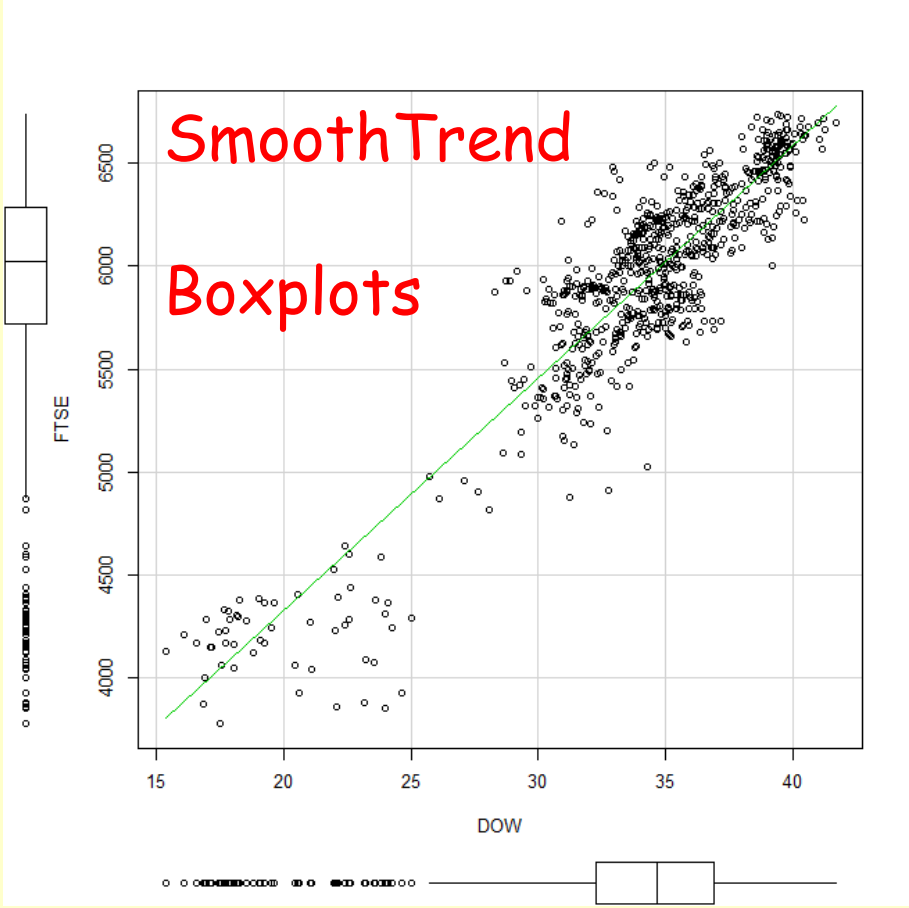
Points outside of the dashed line envelope suggest significant deviations

Graphs in Rcmdr – Scatterplots

Plots two samples on the same coordinates

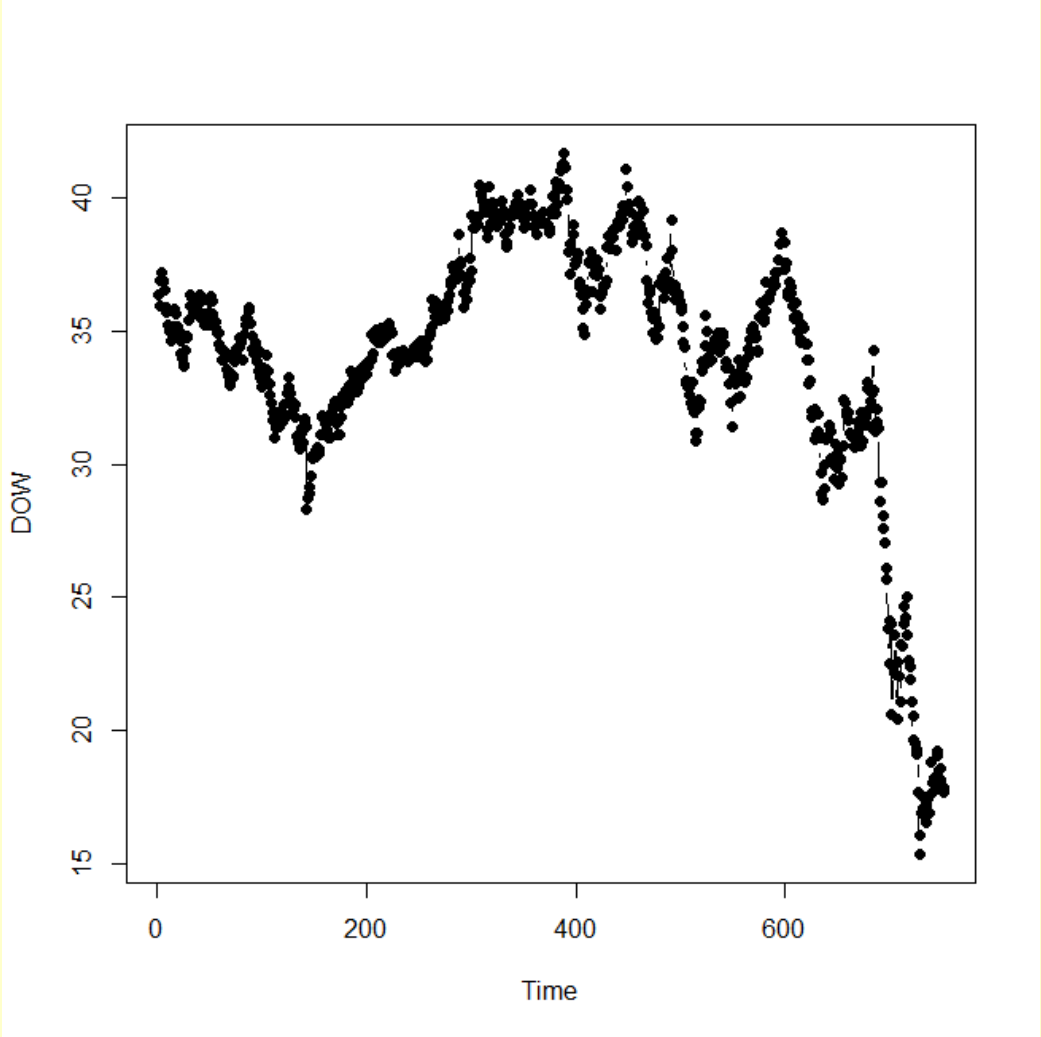
Can add a linear relationship or a smooth trend





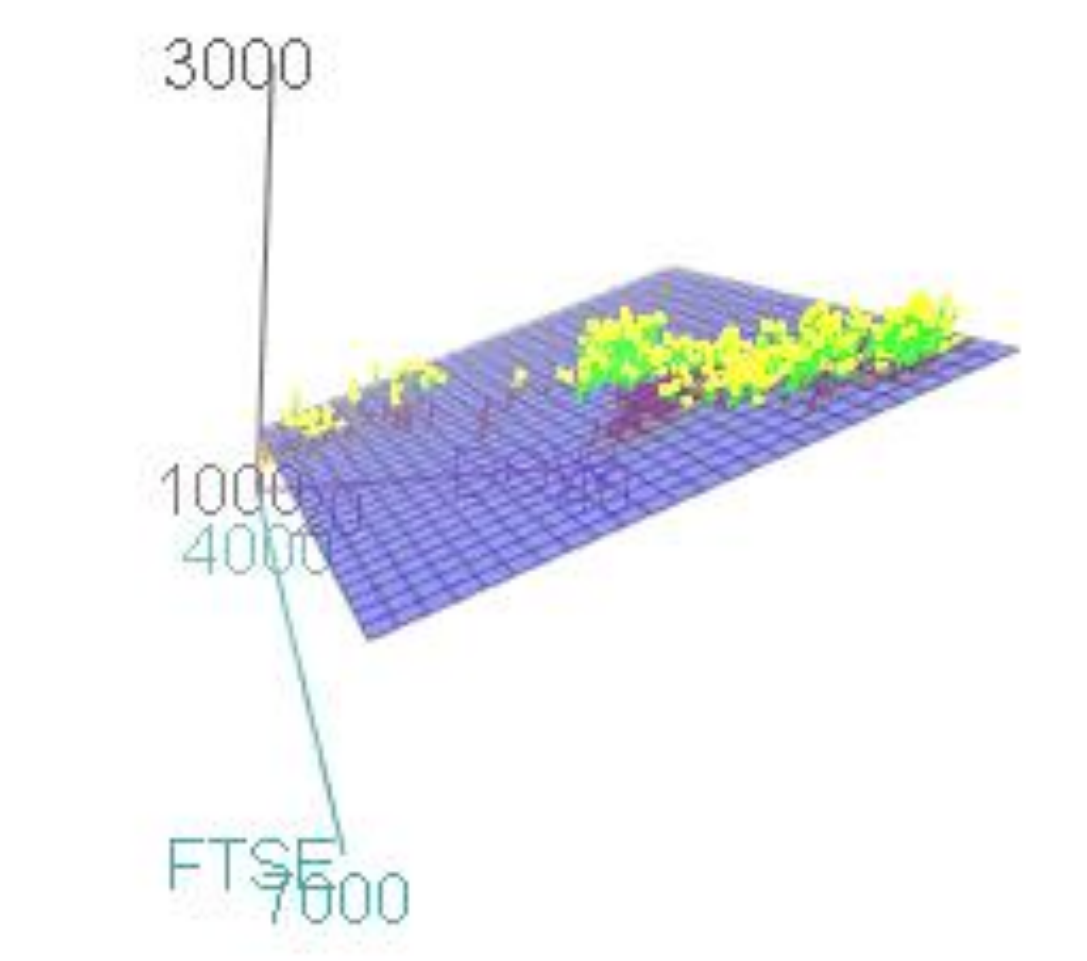
Graphs in Rcmdr – Line Plots

Plots data points linked with a line



Graphs in Rcmdr – 3D Scatterplots

Plots three variables (one response and two drivers) and adds a response surface and the deviations

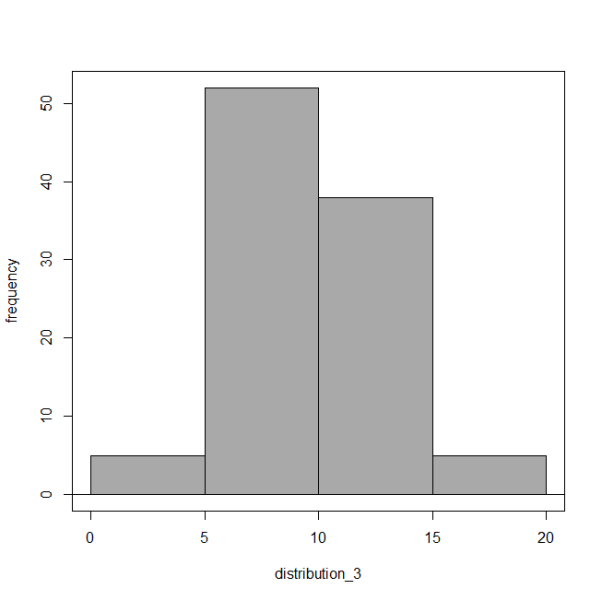


PSA-1: Histograms

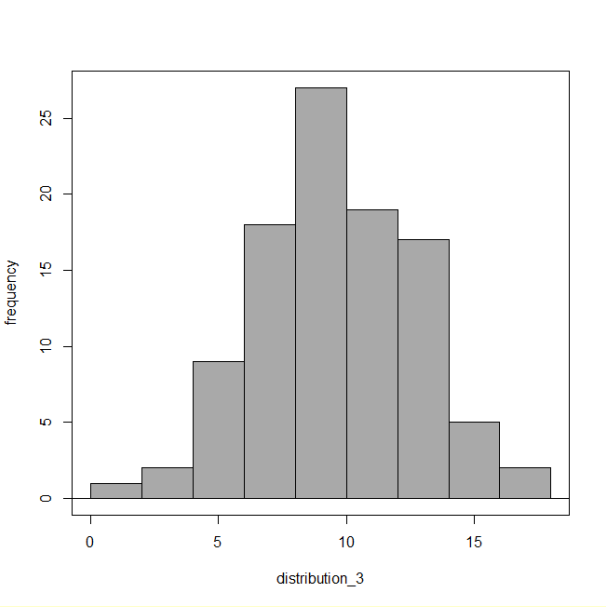
- Number of bins affects:
  "shape" of distribution & number / location of modes
- Even when plotting a Normal Distribution:

	mean	sd	5%	50%	95%	n
dist_3	9.705097	3.033430	5.113512	9.761769	14.40838	100

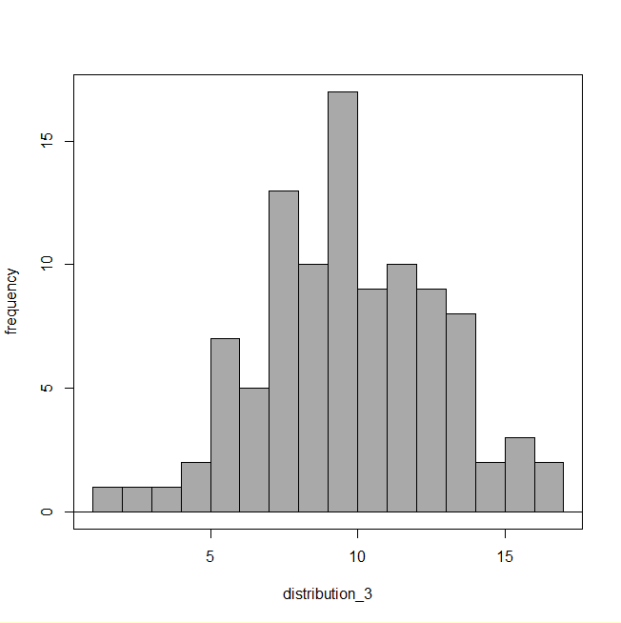
5 bins



9 bins (default)

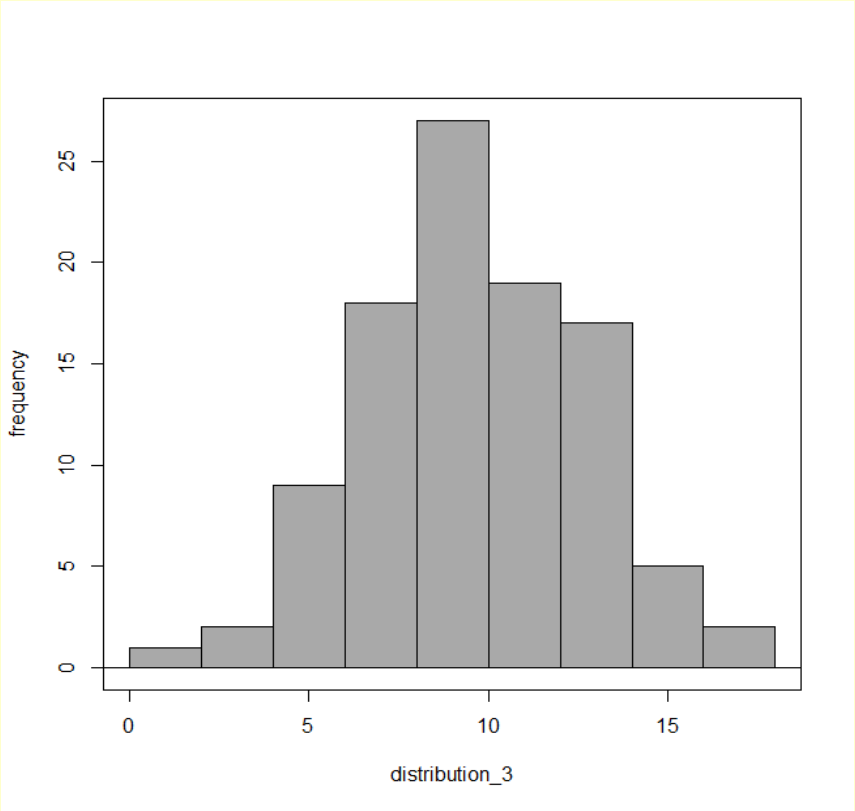


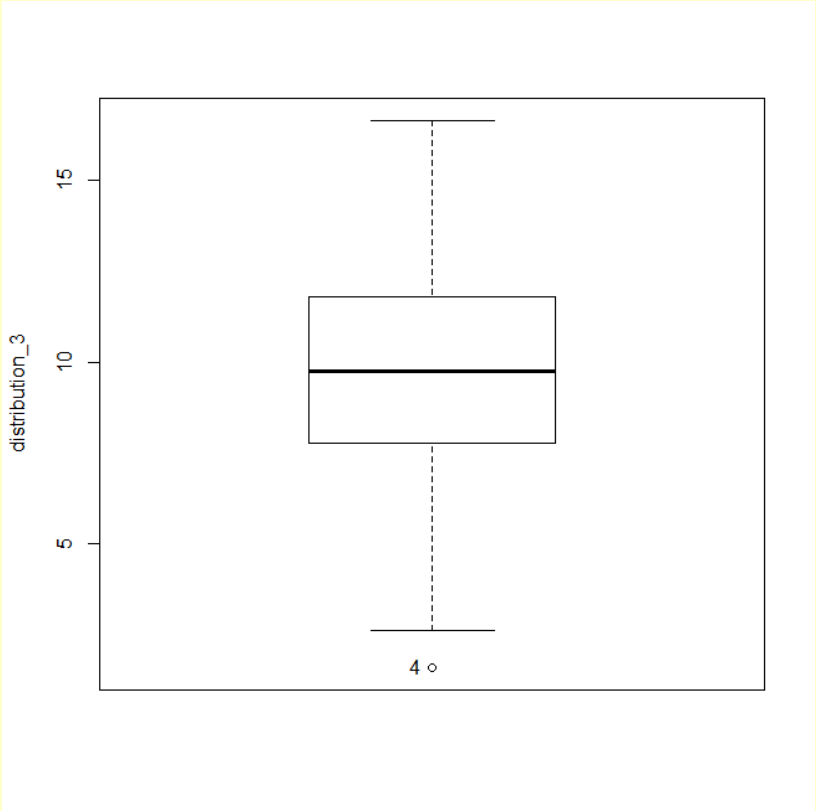
20 bins



PSA-2: Boxplots

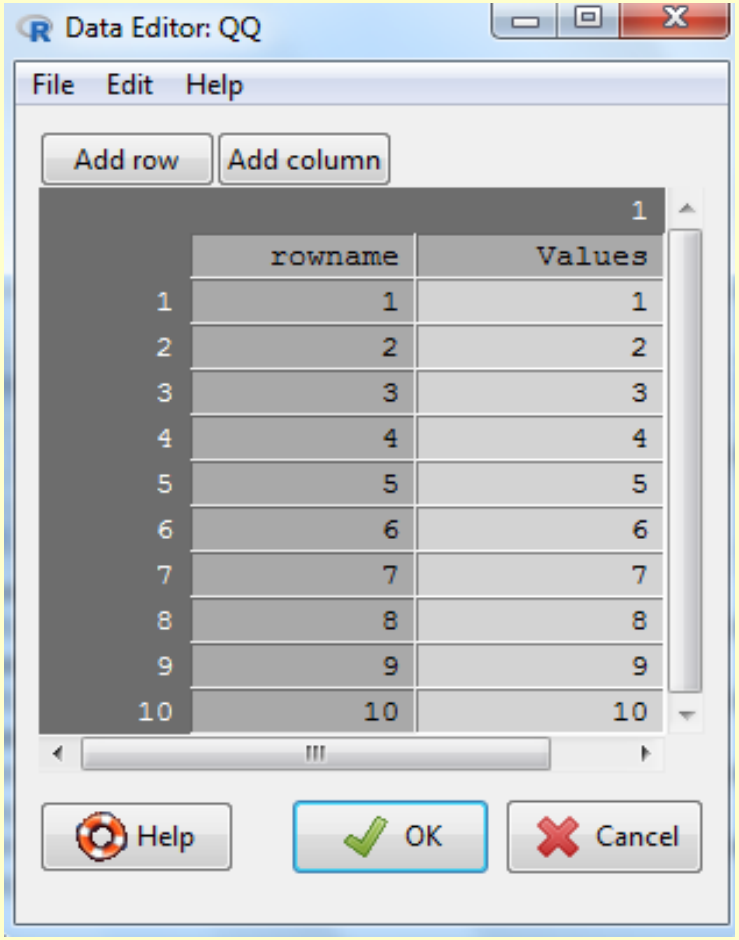
- "Shape" of distribution does not depend on binning
- Highlights the percentiles and outliers

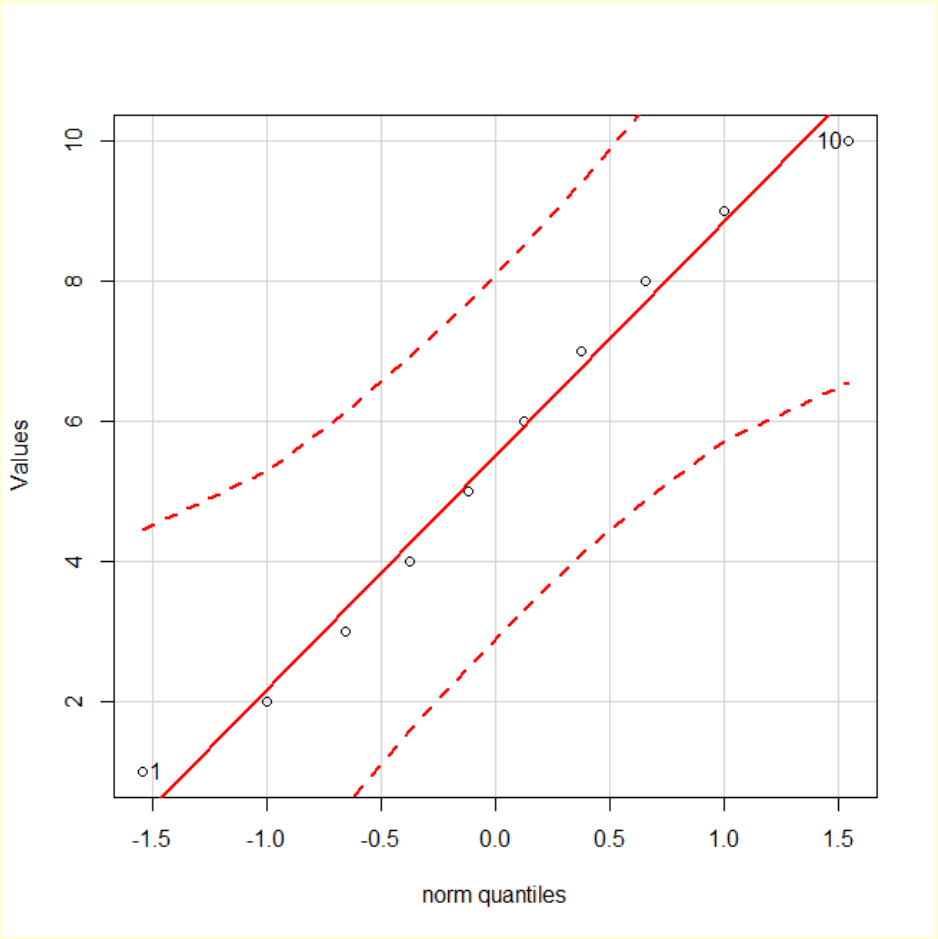




PSA-3: Q-Q Plots

NOTE: Each value in the distribution is plotted.
What are the x and the y coordinates ?
X axis are quartiles of "normal distribution"





PSA-3: Q-Q Plots

NOTE: What are the normalized quantiles?

```
Rcmdr>
summary(QQ)

Values
Min.   : 1.00
1st Qu.: 3.25
Median : 5.50
Mean   : 5.50
3rd Qu.: 7.75
Max.   :10.00
```

```
Rcmdr> numSummary
(QQ[,"Values",drop=FALSE],
statistics=c("IQR","quantiles"),
quantiles=c(.05,0.95))

IQR:   4.5

5%:    1.45
95%:   9.55
N:    10
```

IQR = (3^{rd} Qu) – (1^{st} Qu): 7.75 – 3.25 = 4.5