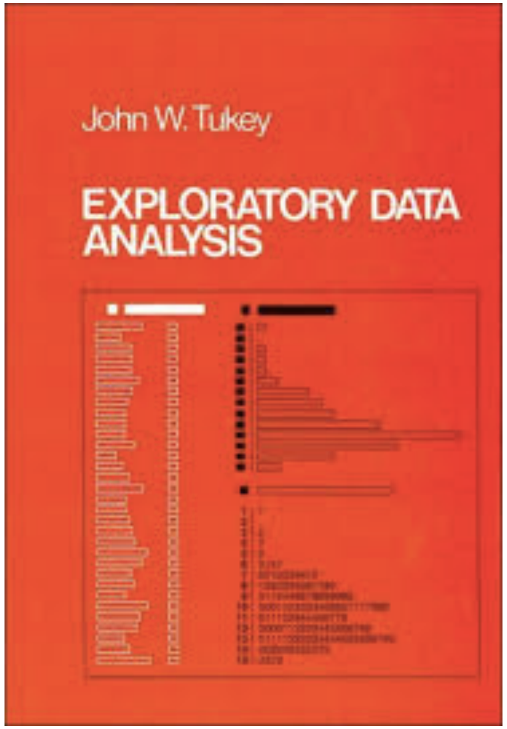
Exploratory Multivariate Data Analysis

EDA and visualization

Instructor: Cheng-Ying Chou

Exploratory Data Analysis 1977

- Based on insights developed at Bell Labs in the 60’s
- Techniques for visualizing and summarizing data
- What can the data tell us? (in contrast to “confirmatory” data analysis)
- Introduced many basic techniques:
  - 5-number summary, box plots, stem and leaf diagrams,…
- 5 Number summary:
  - extremes (min and max)
  - median & quartiles
  - More robust to skewed & longtailed distributions



- Before you begin your analyses, it is imperative that you examine all your variables.
- Why?
  - To listen to the data:
  - To catch mistakes
  - To see patterns in the data
  - To find violations of statistical assumptions
  - …and because if you don’t, you will have trouble later.

What is data?

- Categorical (qualitative)
  - Nominal scales – number is just a symbol that identifies a quality
    - 0=male, 1=female
    - 1=green, 2=blue, 3=red, 4=white
  - Ordinal – rank order

- Quantitative (continuous and discrete)
  - Interval – units are of identical size (i.e. Years)
  - Ratio – distance from an absolute zero (i.e. Age, reaction time)

Organizing your data in a spreadsheet

- Stacked data:
  - Multiple cases (rows) for each subject

Grouping column

Subject	condition	score
1	before	3
1	during	2
1	after	5
2	before	3
2	during	8
2	after	4
3	before	3
3	during	7
3	after	1

- Unstacked data:
  - Only one case (row) per subject

Subject	before	during	after
1	3	2	5
2	3	8	4
3	3	7	1

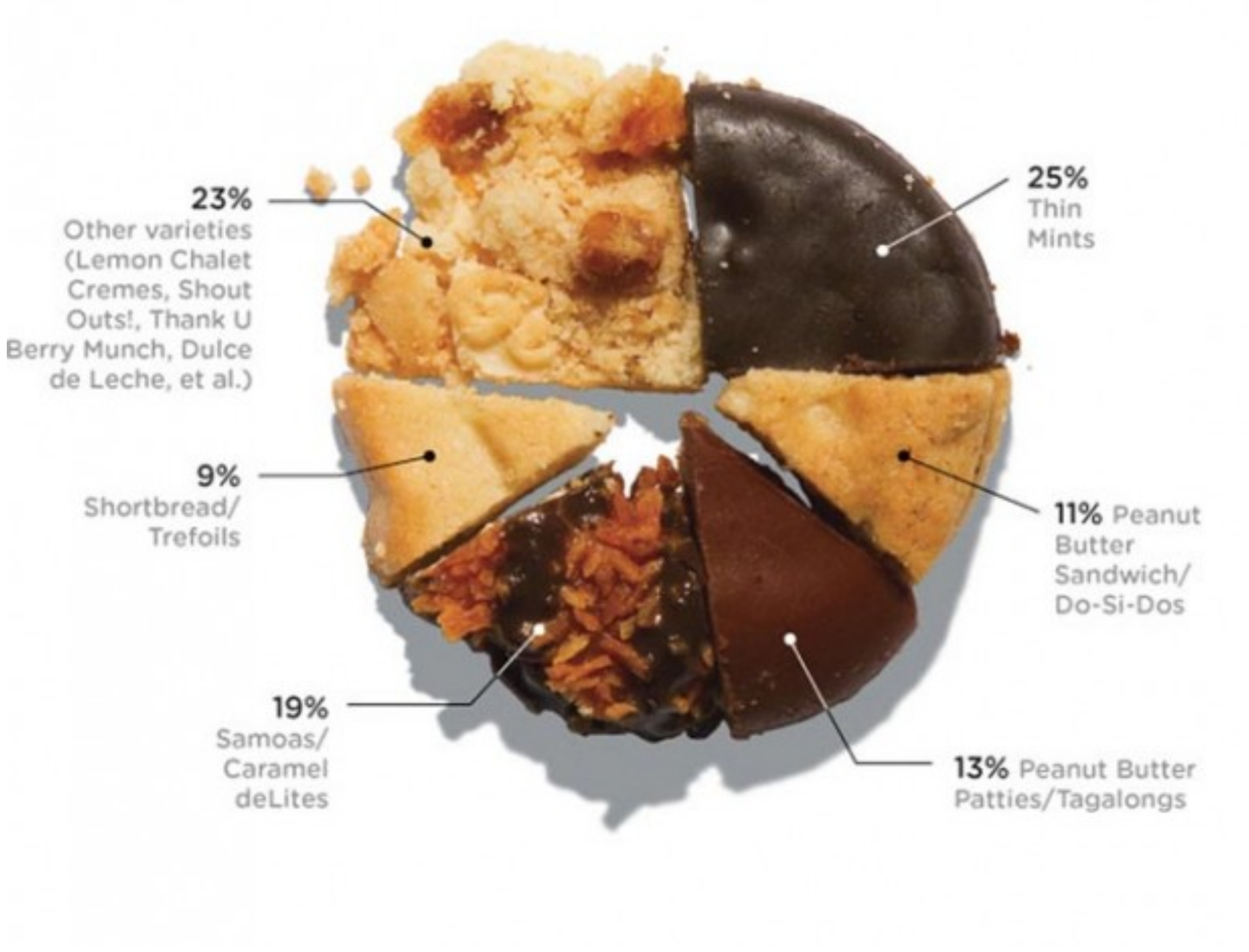
EDA and Visualization

- Exploratory Data Analysis (EDA) and Visualization are very important steps in any analysis task.

- Get to know your data!
  - Distributions (symmetric, normal, skewed)
  - Data quality problems
  - Outliers
  - Correlations and inter-relationships
  - Subsets of interest
  - Suggest functional relationships

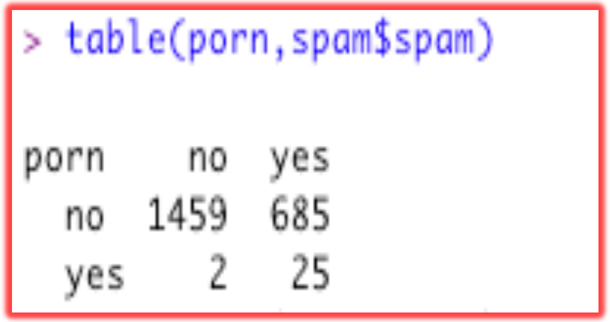
- Sometimes EDA or visualization might be the goal!

Data Visualization – cake bakery



Exploratory Data Analysis (EDA)

- Goal: get a general sense of the data
  - Means, medians, quantiles, histograms, boxplots
  - You should always look at every variable - you will learn something!
- Data-driven (model-free)
- Think interactive and visual
  - Humans are the best pattern recognizers
  - You can use more than 2 dimensions!
    - x,y,z, space, color, time….
- Especially useful in early stages of data mining
  - Detect outliers   (e.g. assess data quality)
  - Test assumptions (e.g. normal distributions or skewed?)
  - Identify useful raw data & transforms (e.g. log(x))
- Bottom line: it is always well worth looking at your data!



```
> table(porn,spam$spam)

porn     no  yes
  no   1459  685
  yes     2   25
```

Descriptive vs. Inferential Statistics

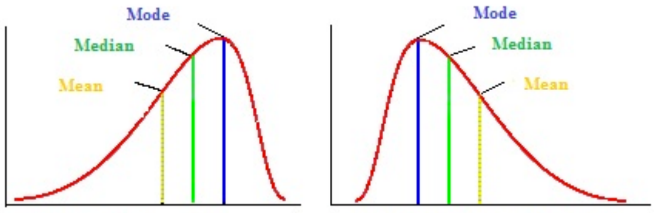
- **Descriptive:** e.g., Median; describes data you have but can't be generalized beyond that.
  - We’ll talk about Exploratory Data Analysis
- **Inferential:** e.g., t-test, that enable inferences about the population beyond our data.
  - These are the techniques we’ll leverage for Machine Learning and Prediction.

Variable Summaries

- Indices of central tendency:
  - Mean – the average value
  - Median – the middle value
  - Mode – the most frequent value
- Indices of Variability:
  - Variance – the spread around the mean
  - Standard deviation
  - Standard error of the mean (estimate)

Summary Statistics

- *Not* visual
- Sample statistics of data X
  - mean: $\mu = \sum_i X_i / n$
  - mode: most common value in X
  - median: **X**=sort(X), median = $\mathbf{X}_{n/2}$ (half below, half above)
  - quartiles of sorted **X**: Q1 value = $\mathbf{X}_{0.25n}$, Q3 value = $\mathbf{X}_{0.75\,n}$
    - interquartile range: value(Q3) - value(Q1)
    - range: max(X) - min(X) = $\mathbf{X}_n - \mathbf{X}_1$
  - variance: $\sigma^2 = \sum_i (X_i - \mu)^2 / n$
  - skewness: $\sum_i (X_i - \mu)^3 / [(\sum_i (X_i - \mu)^2)^{3/2}]$
    - zero if symmetric; right-skewed more common (what kind of data is right skewed?)



Left-Skewed (Negative Skewness)

Right-Skewed (Positive Skewness)

  - Number of distinct values for a variable (see unique() in R)
  - Don’t need to report all of these: Bottom line…do these numbers make sense???

The Mean

Subject	before	during	after
1	3	2	7
2	3	8	4
3	3	7	3
4	3	2	6
5	3	8	4
6	3	1	6
7	3	9	3
8	3	3	6
9	3	9	4
10	3	1	7

Sum = 30 50 50

/n 10 10 10

Mean = 3 5 5

Mean = sum of all scores divided by number of scores

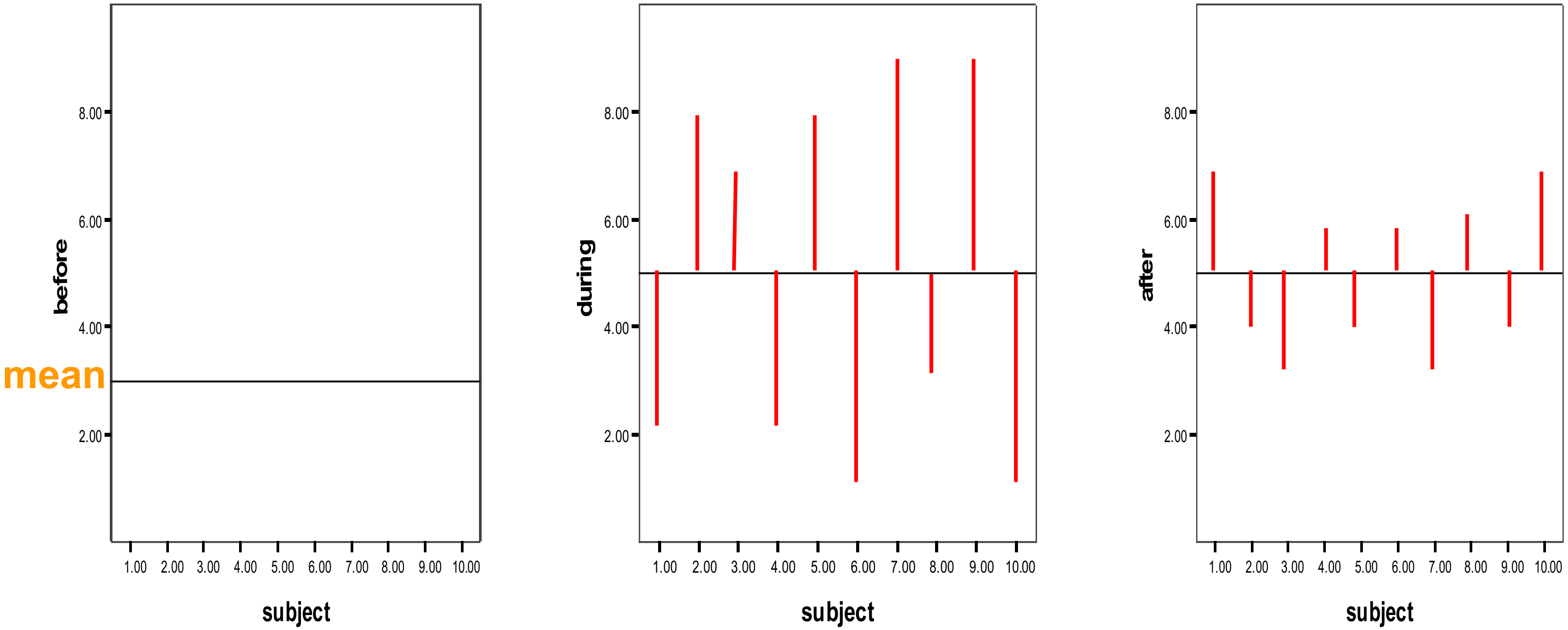
$$\frac{X_1 + X_2 + X_3 + \ldots. X_n}{n}$$

The Variance: Sum of the squared deviations divided by number of scores

Subject	before	during	after	Before –mean	(Before –mean)^2	During – mean	(During – mean)^2	After - mean	(After –mean)^2
1	3	2	7	0	0	-3	9	2	4
2	3	8	4	0	0	3	9	-1	1
3	3	7	3	0	0	2	4	-2	4
4	3	2	6	0	0	-3	9	1	1
5	3	8	4	0	0	3	9	-1	1
6	3	1	6	0	0	-4	16	1	1
7	3	9	3	0	0	4	16	-2	4
8	3	3	6	0	0	-2	4	1	1
9	3	9	4	0	0	4	16	-1	1
10	3	1	7	0	0	-4	16	2	4
Sum =	30	50	50	0	0	0	108	0	22
/n	10	10	10		10*		10		10
Mean =	3	5	5	VAR =	0		10.8		2.2

- Actually you divide by n-1 because it is a sample and not a population, but you get the idea…

Variance continued



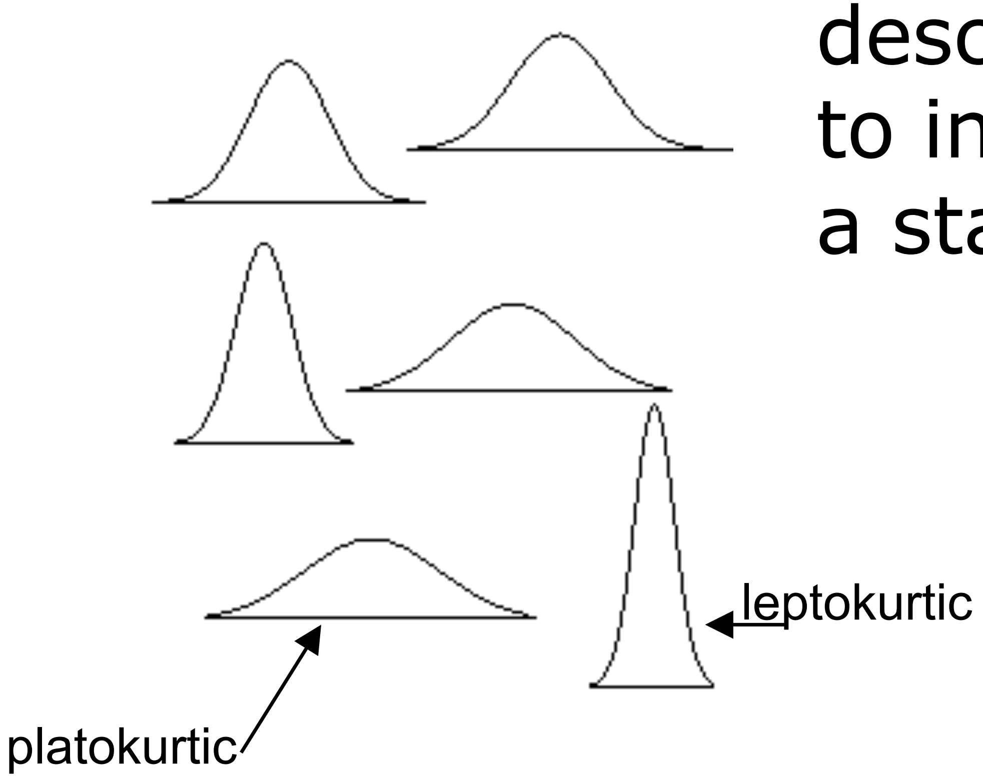
Distribution

- Means and variances are ways to describe *a distribution* of scores.
- Knowing about your distributions is one of the best ways to understand your data.
- A NORMAL (aka Gaussian) distribution is the most common assumption of statistics, thus it is often important to check if your data are normally distributed.

What is “normal” anyway?

- With enough measurements, most variables are distributed normally.

But in order to fully describe data we need to introduce the idea of a standard deviation.



Standard deviation

- Variance, as calculated earlier, is arbitrary.
- What does it mean to have a variance of 10.8? Or 2.2? Or 1459.092? Or 0.000001?
- Nothing.
  - But if you could "standardize" that value, you could talk about any variance (i.e. deviation) in equivalent terms.
- Standard Deviations are simply the square root of the variance.

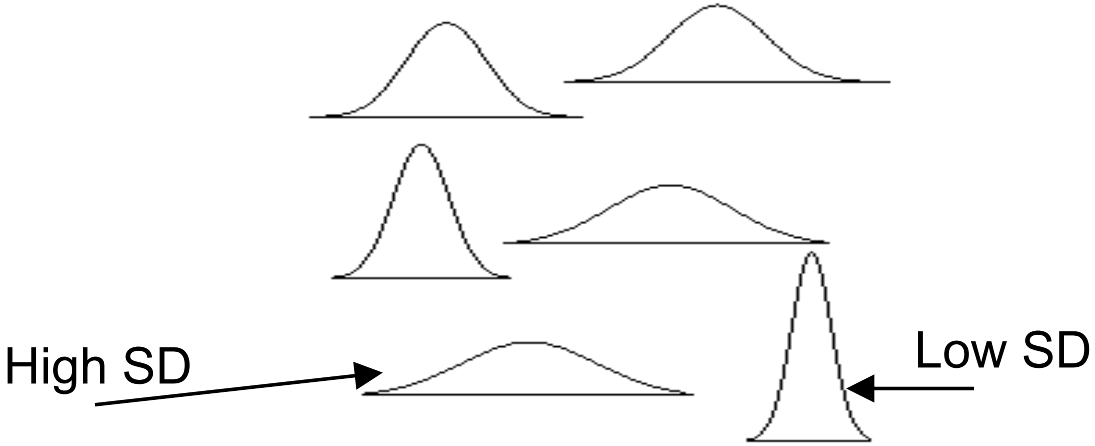
Standard deviation

The process of standardizing deviations goes like this:

1. Score (in the units that are meaningful)
2. Mean
3. Each score’s deviation from the mean
4. Square that deviation
5. Sum all the squared deviations (Sum of Squares)
6. Divide by n (if population) or n-1 (if sample)
7. Square root – now the value is in the units we started with!!!

Interpreting standard deviation (SD)

- First, the SD will let you know about the distribution of scores around the mean.
- High SDs (relative to the mean) indicate the scores are spread out
- Low SDs tell you that most scores are very near the mean.



Interpreting standard deviation (SD)

- Second, you can then interpret any individual score in terms of the SD.
  - For example: mean = 50, SD = 10

    vs. mean = 50, SD = 1
- A score of 55 is:
  - 0.5 Standard deviation units from the mean (not much) OR
  - 5 standard deviation units from mean (a lot!)

Standardized scores (Z)

- Third, you can use SDs to create *standardized* scores
  - Force the scores onto a normal distribution by putting each score into units of SD.
- Subtract the mean from each score and divide by SD

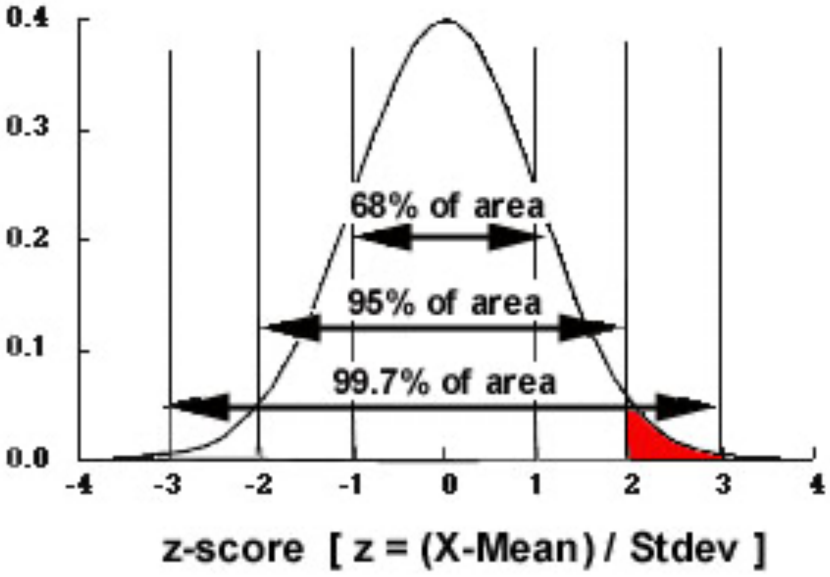
$$Z = (X - \text{mean})/SD$$

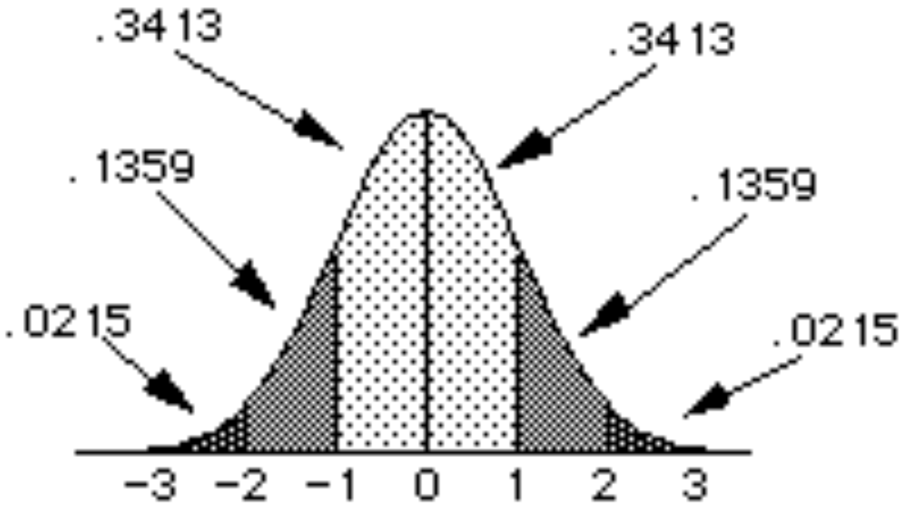
This is truly an amazing thing

Standardized normal distribution

ALL Z-scores have a mean of 0 and SD of 1. Nice and simple.

From this we can get the proportion of scores anywhere in the distribution.





The trouble with normal

- We violate assumptions about statistical tests if the distributions of our variables are not approximately normal.
- Thus, we must first examine each variable’s distribution and make adjustments when necessary so that assumptions are met.

PART II

Examine every variable for:

- Out of range values
- Normality
- Outliers

Visual display of univariate data

- Now the example data from before has decimals
- (what kind of data is that?)
- Precision has increased

Subject	before	during	after
1	3.1	2.3	7
2	3.2	8.8	4.2
3	2.8	7.1	3.2
4	3.3	2.3	6.7
5	3.3	8.6	4.5
6	3.3	1.5	6.6
7	2.8	9.1	3.4
8	3	3.3	6.5
9	3.1	9.5	4.1
10	3	1	7.3

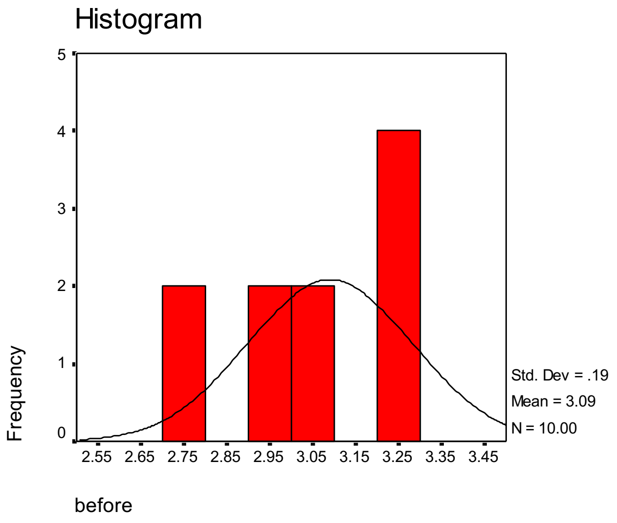
Visual display of univariate data

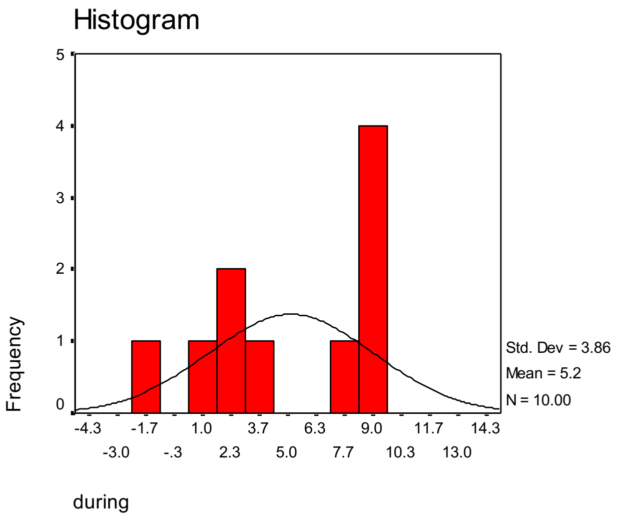
- Histograms
- Stem and Leaf plots
- Boxplots
- QQ Plots
- …and many many more

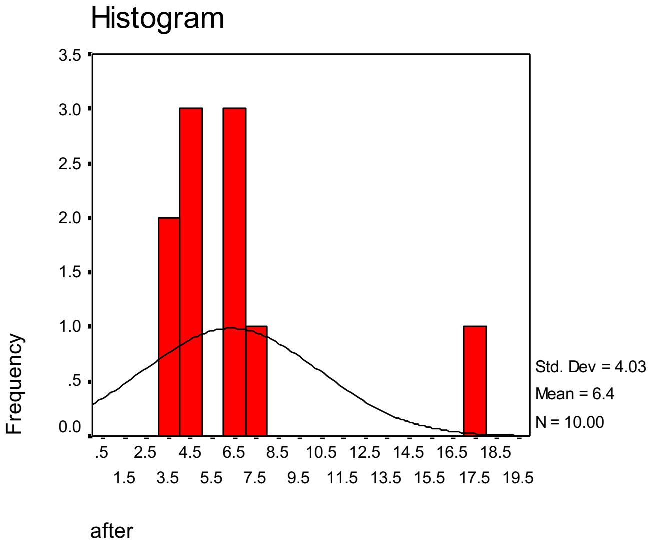
Subject	before	during	after
1	3.1	2.3	7
2	3.2	8.8	4.2
3	2.8	7.1	3.2
4	3.3	2.3	6.7
5	3.3	8.6	4.5
6	3.3	1.5	6.6
7	2.8	9.1	3.4
8	3	3.3	6.5
9	3.1	9.5	4.1
10	3	1	7.3

Histograms

- # of bins is very important:

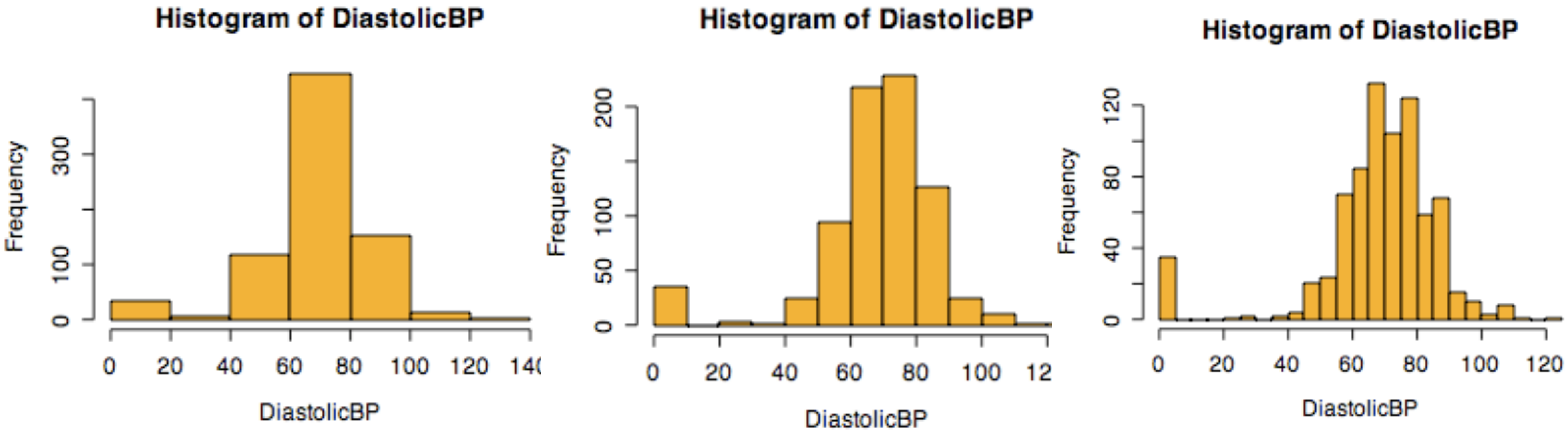






Single Variable Visualization

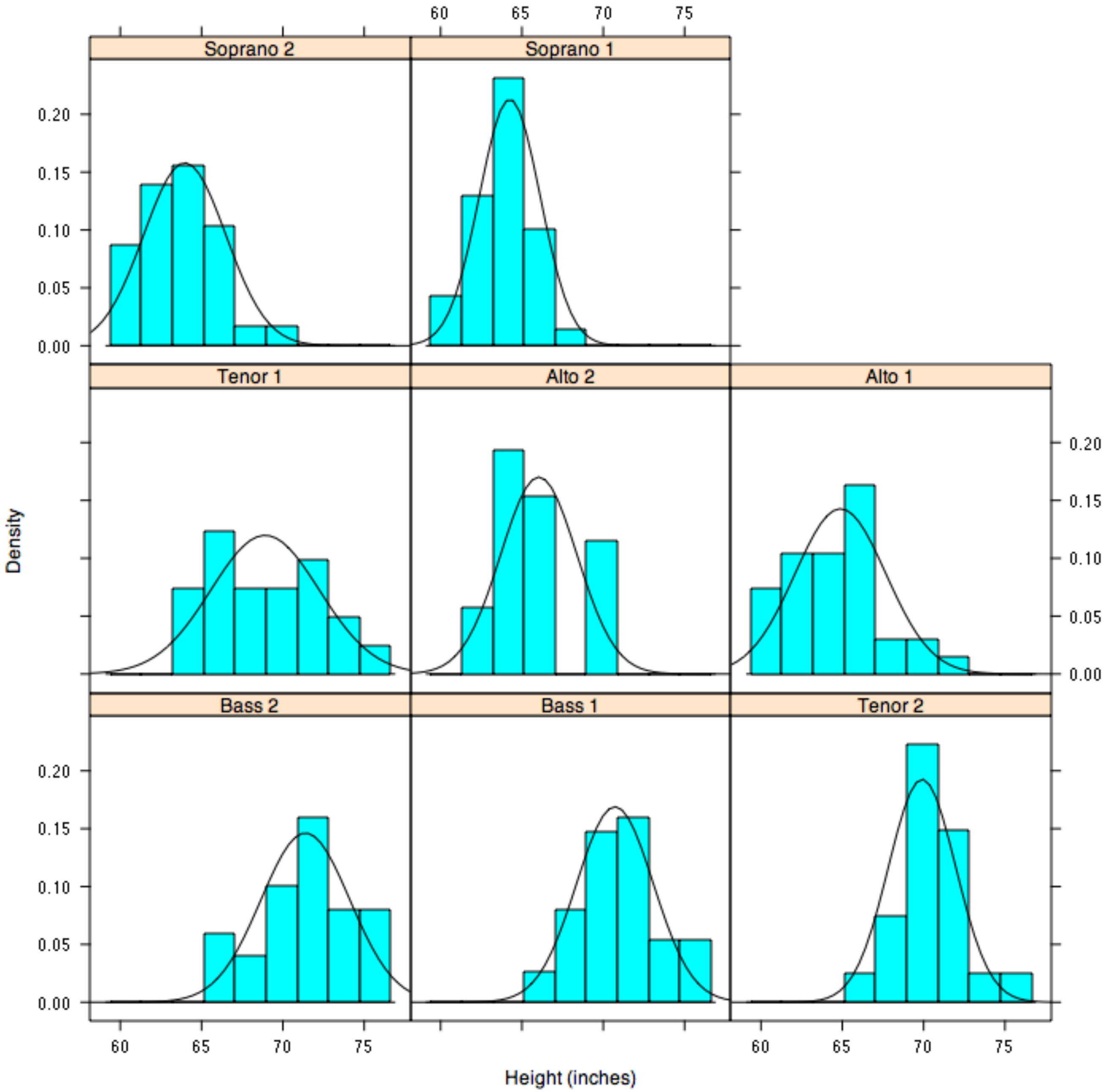
- Histogram:
  - Shows center, variability, skewness, modality,
  - Outliers, or strange patterns.
  - Bin width and position matter



Issues with Histograms

- For small data sets, histograms can be misleading.
  - Small changes in the data, bins, or anchor can deceive.
- For large data sets, histograms can be quite effective at illustrating general properties of the distribution.
- Histograms effectively only work with 1 variable at a time.
  - But ‘small multiples’ can be effective.

But be careful with axes and scales!

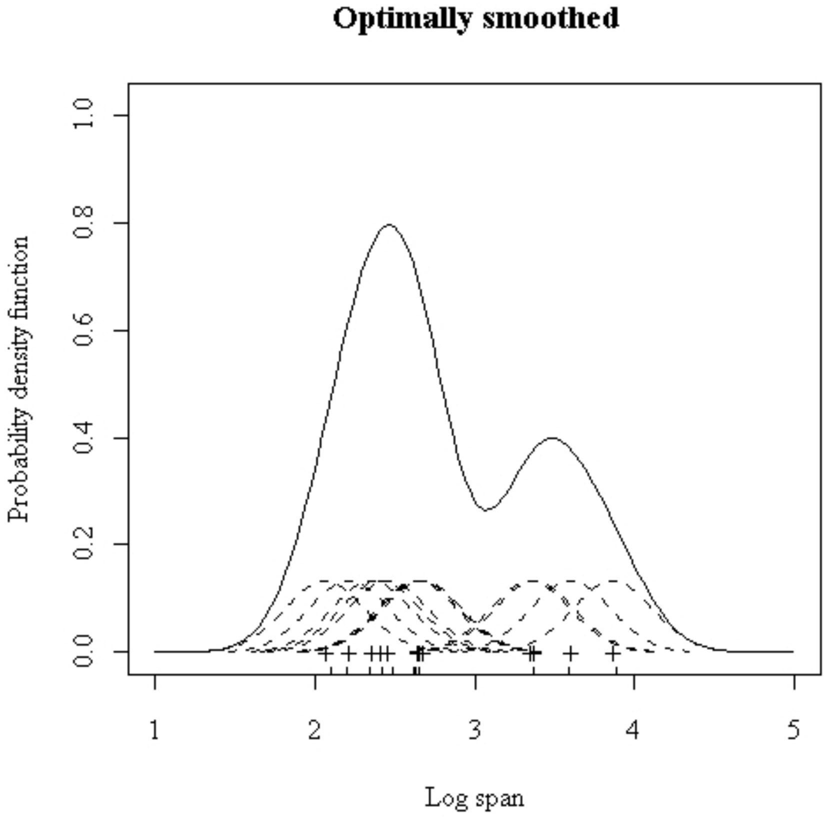


Smoothed Histograms - Density Estimates

- Kernel estimates smooth out the contribution of each datapoint over a local neighborhood of that point.

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

h is the kernel width



- Gaussian kernel is common:

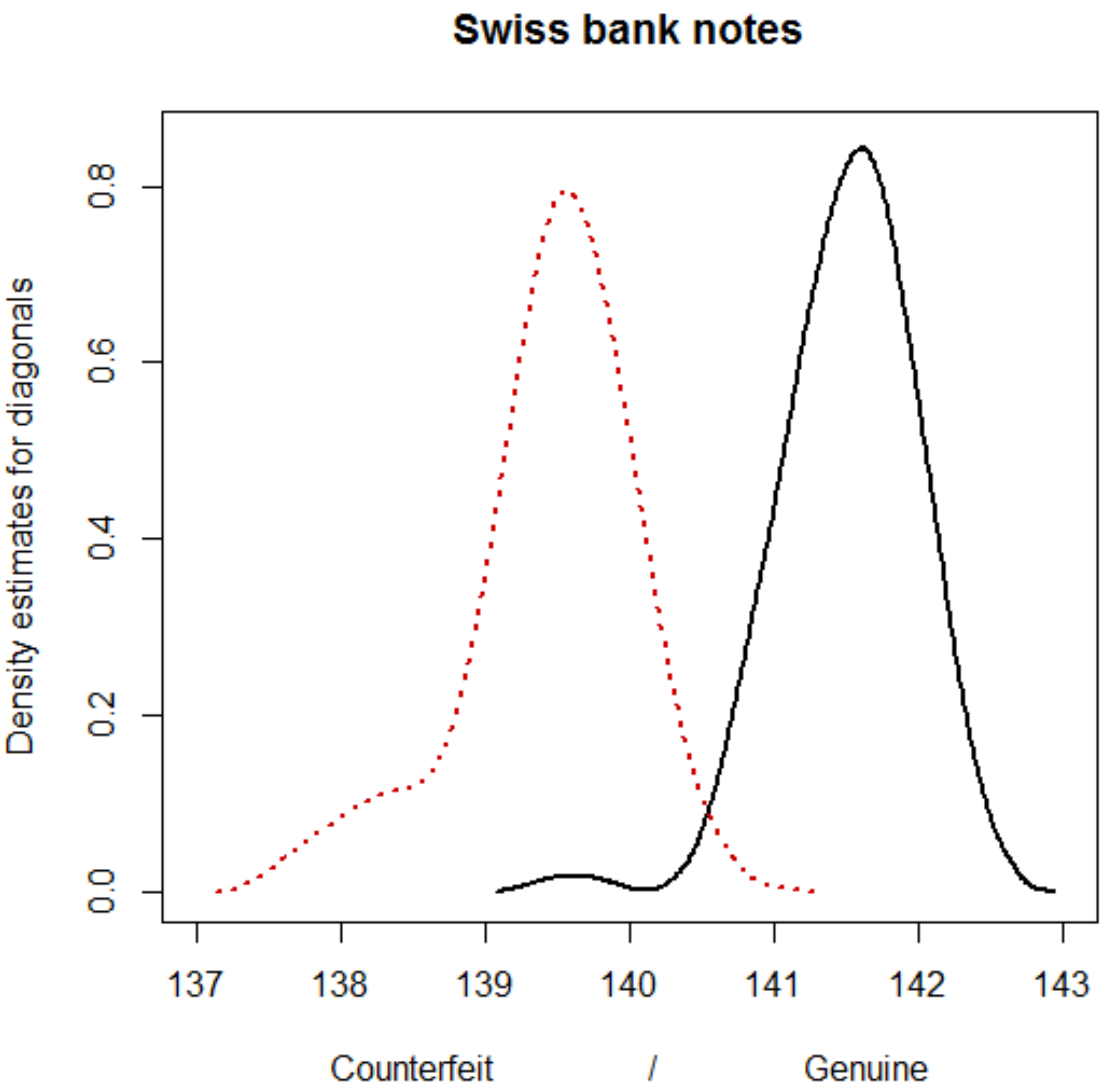
$$Ce^{-\frac{1}{2}\left(\frac{x - x(i)}{h}\right)^2}$$

Kernel functions

$K(\cdot)$	**Kernel**
$K(u) = I(\|u\| \leq \frac{1}{2})$	Uniform
$K(u) = (1 - \|u\|)I(\|u\| \leq 1)$	Triangle
$K(u) = \frac{3}{4}(1 - u^2)I(\|u\| \leq 1)$	Epanechnikov
$K(u) = \frac{15}{16}(1 - u^2)^2 I(\|u\| \leq 1)$	Quartic (Biweight)
$K(u) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{u^2}{2}) = \varphi(u)$	Gaussian

Bandwidth choice is an art

Usually want to try several



Kernel densities

- Kernel densities estimate distribution densities by the kernel method.
- The bandwidth h determines the degree of smoothness of the estimate f.
- A simple (but not necessarily correct) way to find a good bandwidth is to compute the rule of thumb bandwidth h_G=1.06 σ_n-1/5. This bandwidth is to be used only in combination with a Gaussian kernel φ.
- Kernel density estimates are a good descriptive tool for seeing modes, location, skewness, tails, asymmetry etc.

Stem and Leaf plots

Before:

N = 10   Median = 3.1 Quartiles = 3, 3.3

```
2 : 88
3 : 00112333
```

After:

N = 10   Median = 5.5 Quartiles = 4.1, 6.7

```
3 : 24
4 : 125
5 :
6 : 567
7 : 3
```

High: 17

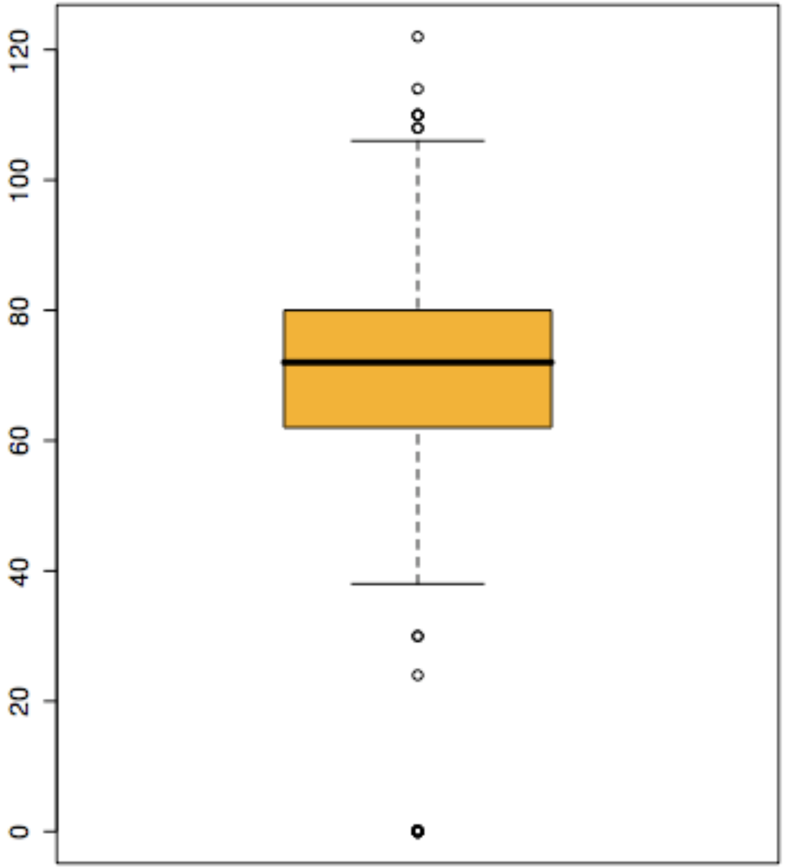
During:

N = 10   Median = 5.2 Quartiles = 2.3, 8.8

```
1 : 05
2 : 33
3 : 3
4 :
5 :
6 :
7 : 1
8 : 68
9 : 15
```

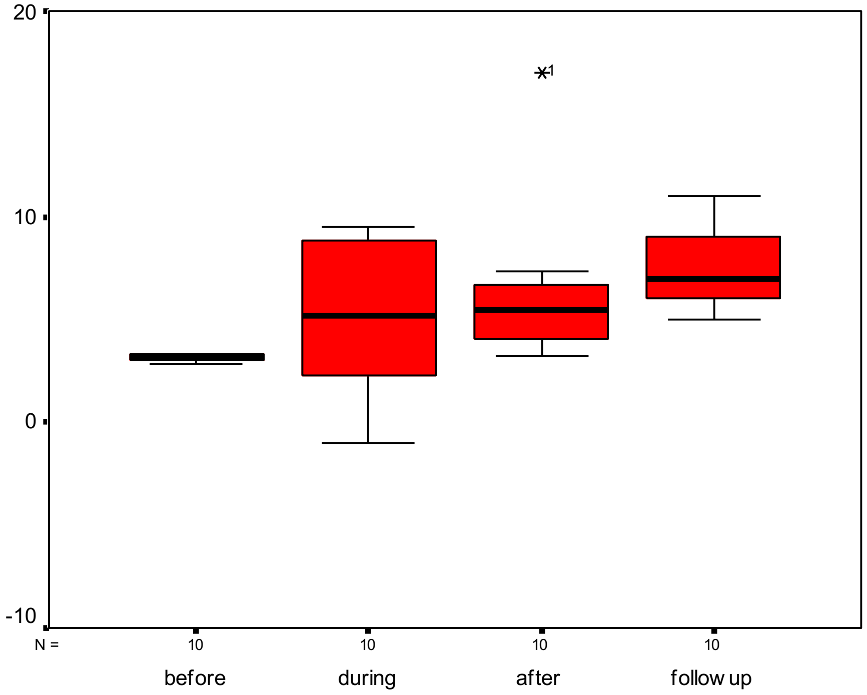
Boxplots

- Shows a lot of information about a variable in one plot
  - Median
  - IQR
  - Outliers
  - Range
  - Skewness
- Negatives
  - Overplotting
  - Hard to tell distributional shape
  - No standard implementation in software (many options for whiskers, outliers)



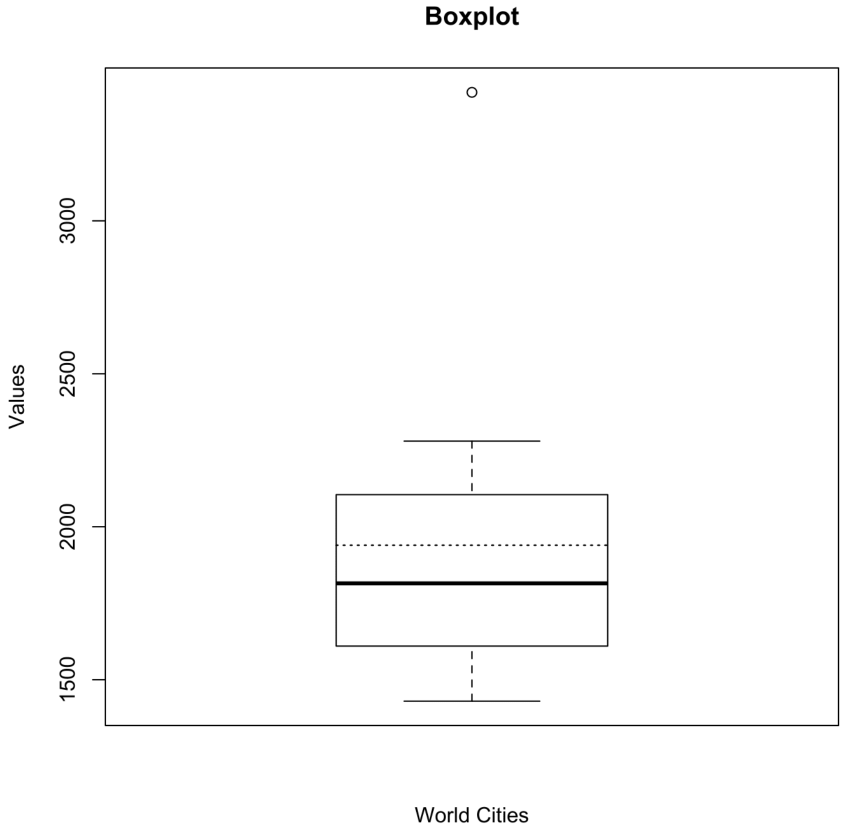
Boxplots

- Upper and lower bounds of boxes are the 25th and 75th percentile (interquartile range)
- Whiskers are min and max value unless there is an outlier
- An outlier is beyond 1.5 times the interquartile range (box length).
- Extreme outlier is beyond 3 IQR.



City	Pop (10,000)	Order statistics
Tokyo	3,420	X(15)
Mexico city	2,280	X(14)
Seoul	2,230	X(13)
New York	2,190	X(12)
Sao Paulo	2,020	X(11)
Bombay	1,985	X(10)
Delhi	1,970	X(9)
Shanghai	1,815	X(8)
Los Angeles	1,800	X(7)
Osaka	1,680	X(6)
Jakarta	1,655	X(5)
Calcutta	1,565	X(4)
Cairo	1,560	X(3)
Manila	1,495	X(2)
Karachi	1,430	X(1)

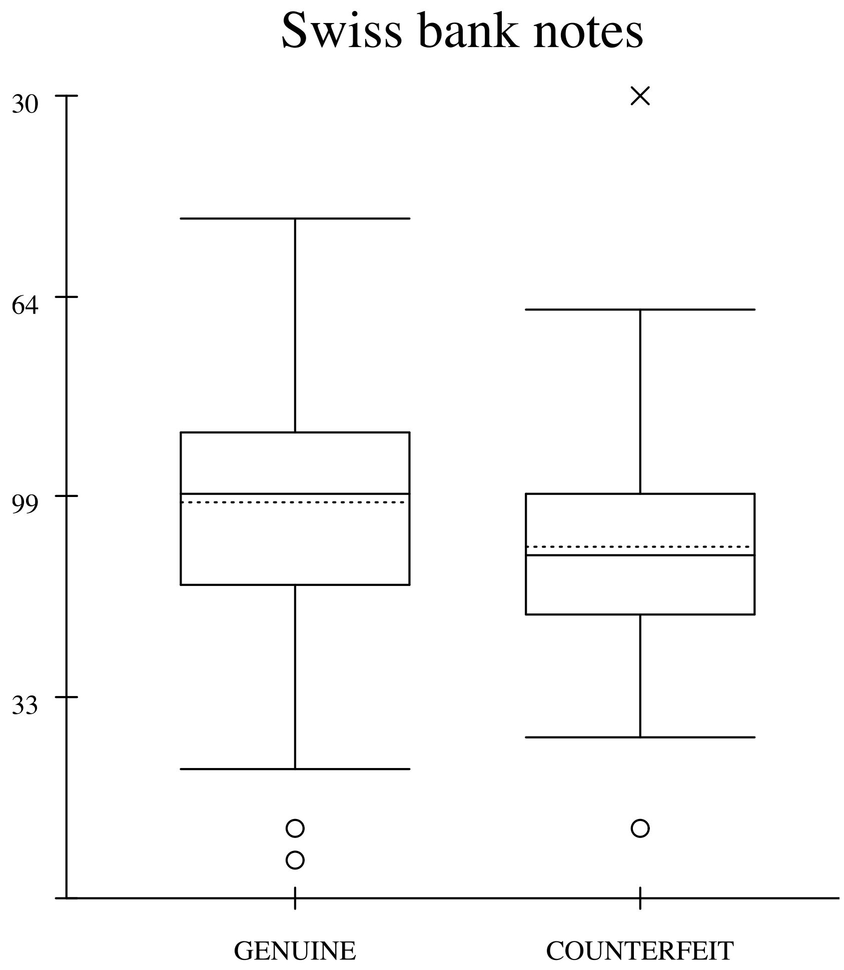
- Median: 1,815
- F_L: ½(1,565+1,655)=1,610
- F_U: ½(2,020+2,190)=2,105
- IQR=F_u-F_L=495
- F_L-1.5*IQR=867.5
- F_U+1.5*IQR=2847.5

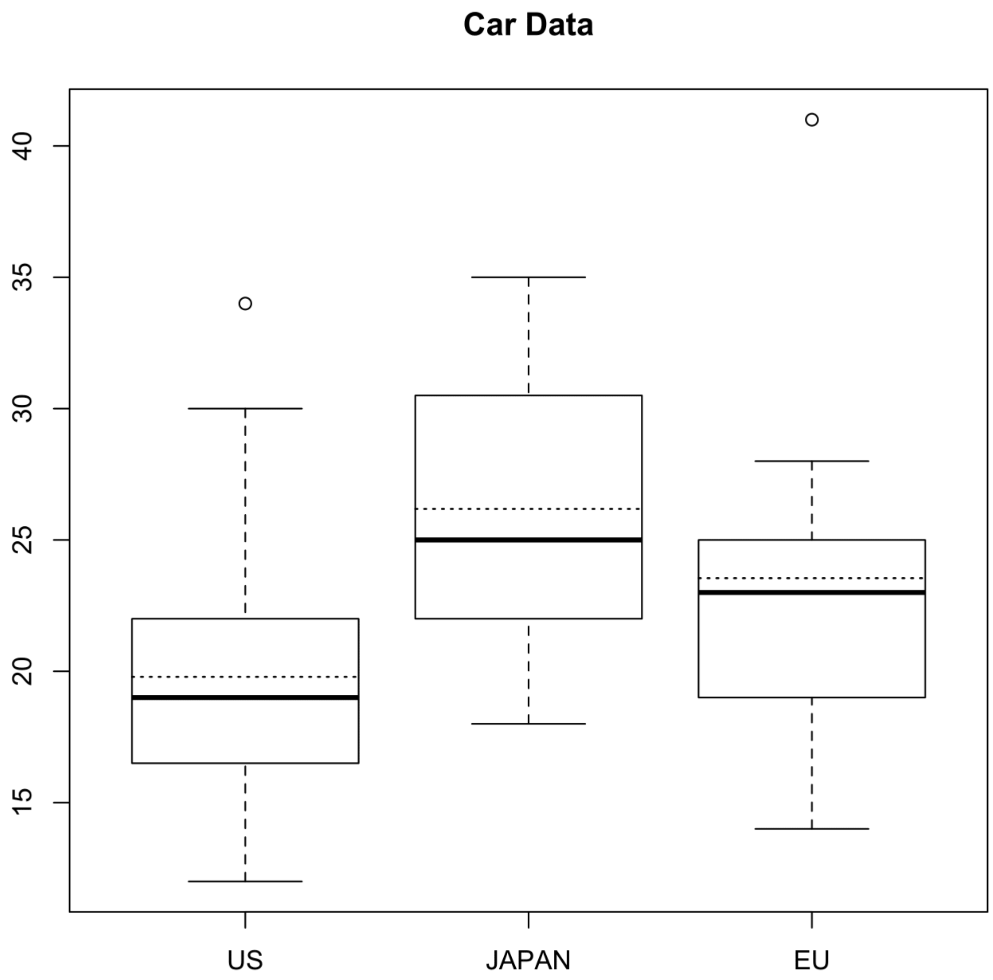


```
# clear variables and close windowsrm(list = ls(all = TRUE))
graphics.off()
# load data
x  = read.table("cities.txt")
m1 = mean(as.matrix(x))
# Plot box plot
boxplot(x, xlab = "World Cities", ylab = "Values")lines(c(0.8, 1.2), c(m1, m1), col = "black", lty = "dotted", lwd = 1.2)
title("Boxplot")
# Five Number Summary R 'quantile' function uses a different algorithm to
# calculate the sample quantiles than in the MVA book Therefore, the values# using Matlab could differ from the Book values, but the difference is not
# great, and should not be significant.  Easiest way to calculate Five Number
# Summary is quantile(population,[.025 .25 .50 .75 .975])
five = quantile(x[, 1], c(0.025, 0.25, 0.5, 0.75, 0.975))
# Display results
print("Five number summary, in millions")print(five/100)
```

MVAboxcity

Boxplots

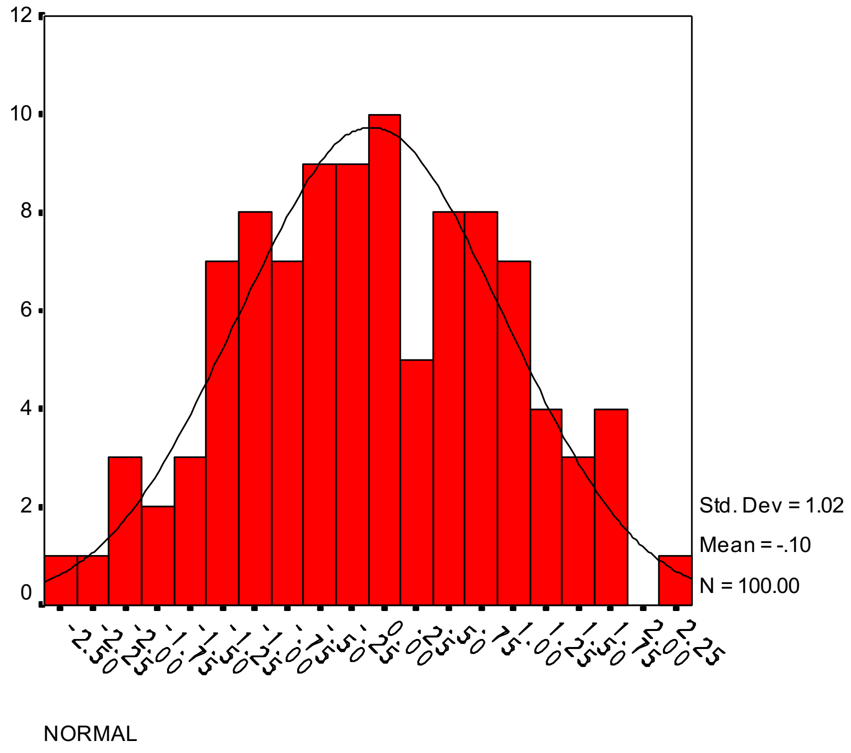

Swiss bank notes
30
64
99
33
GENUINE
COUNTERFEIT


Car Data
40
35
30
25
20
15
US
JAPAN
EU

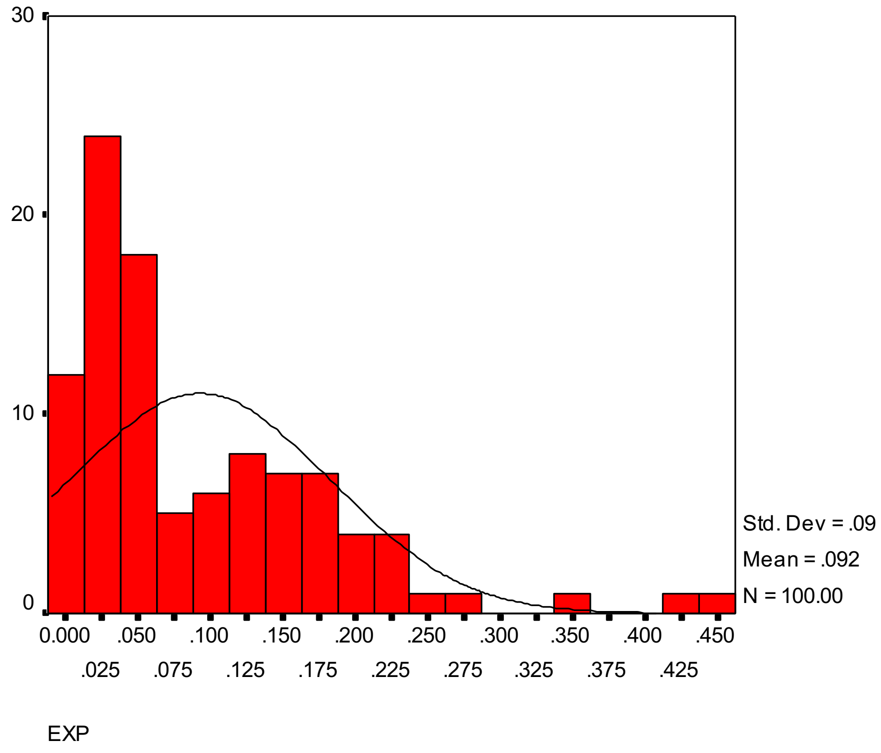
Boxplots

- Median and mean bar indicate the central locations.
- The relative location of median (and mean) in the box is a measure of skewness.
- The length of box and whiskers is a measure of spread.
- The length of whiskers indicate the tail length of the distribution.
- The outlying points are marked as ”○” or‟   ” outside the outside bars.
- If we compare the relative size and location of the boxes we are comparing distributions.

Quantile-Quantile (Q-Q) Plots

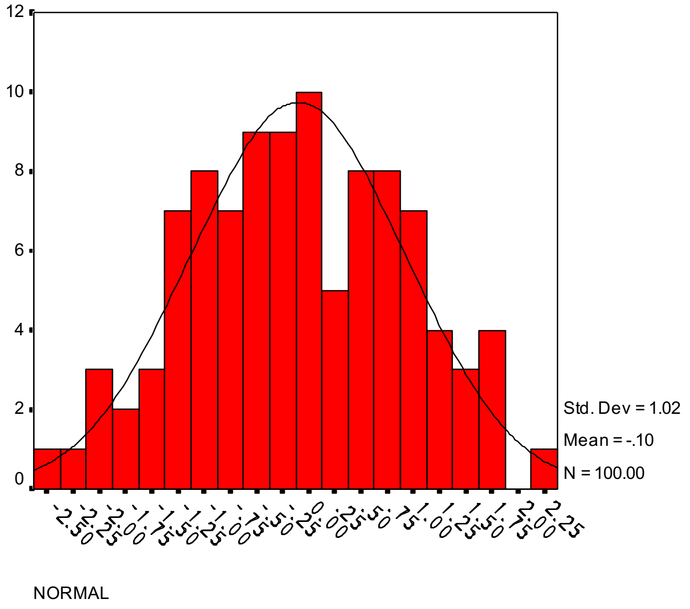


Random Normal Distribution

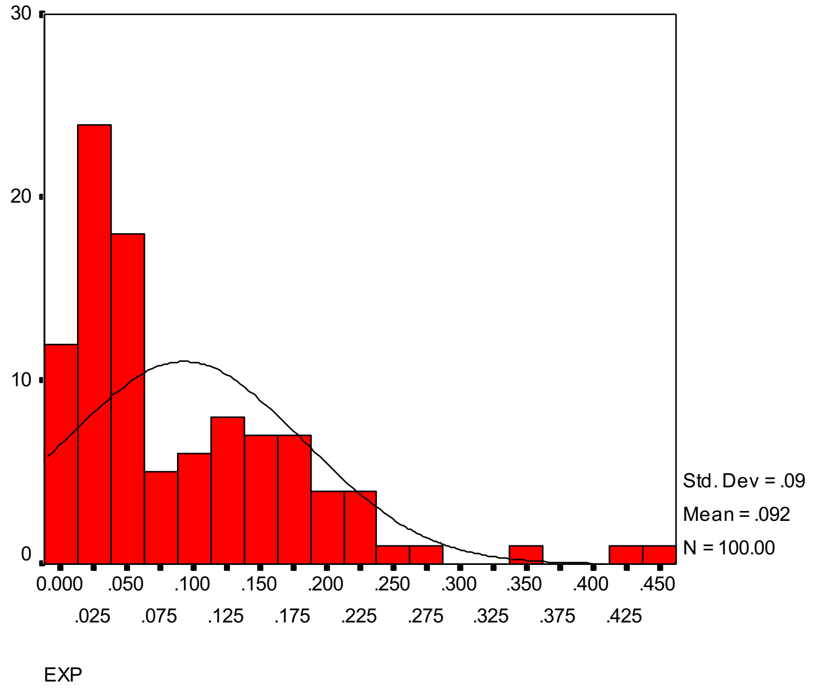


Random Exponential Distribution

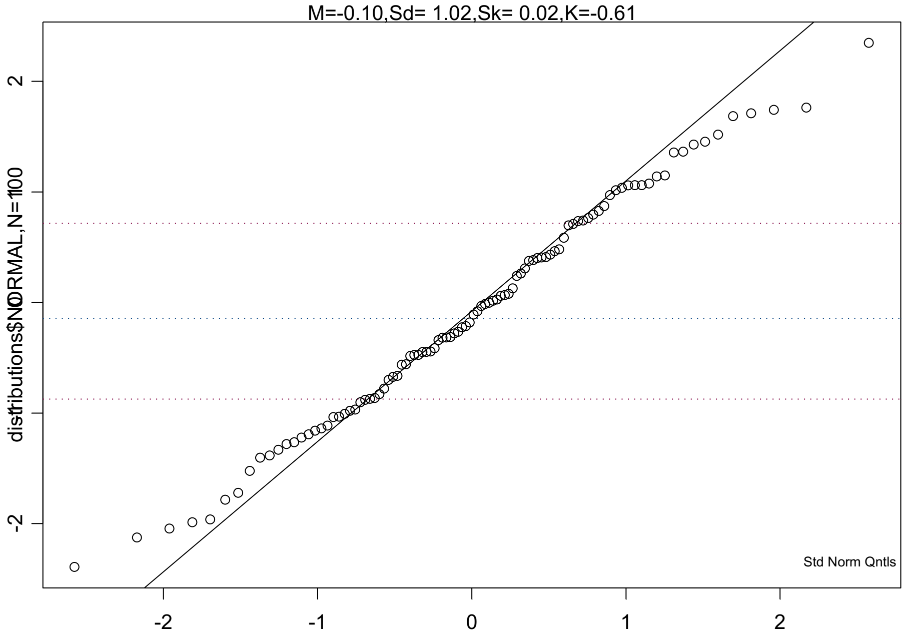
Q-Q Plots



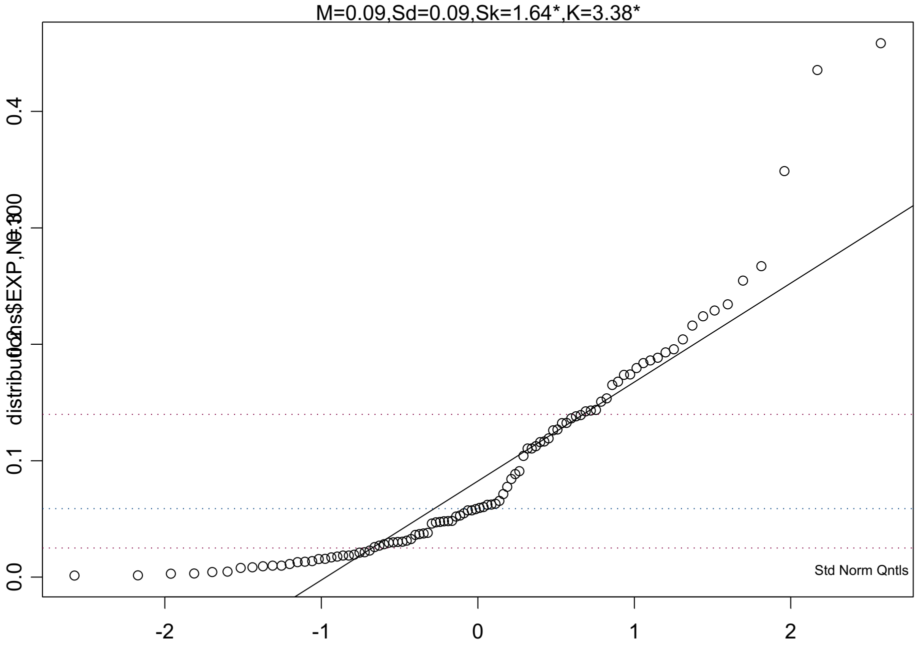
12
10
8
6
4
2
0
-2.50
-2.25
-2.00
-1.75
-1.50
-1.25
-1.00
-.75
-.50
-.25
0.00
.25
.50
.75
1.00
1.25
1.50
1.75
2.00
2.25
Std. Dev = 1.02
Mean = -.10
N = 100.00
NORMAL



30
20
10
0
0.000
.025
.050
.075
.100
.125
.150
.175
.200
.225
.250
.275
.300
.325
.350
.375
.400
.425
.450
Std. Dev = .09
Mean = .092
N = 100.00
EXP



M=-0.10,Sd= 1.02,Sk= 0.02,K=-0.61
distributions$NORMAL,N=100
2
-2
Std Norm Qntls
-2
-1
0
1
2



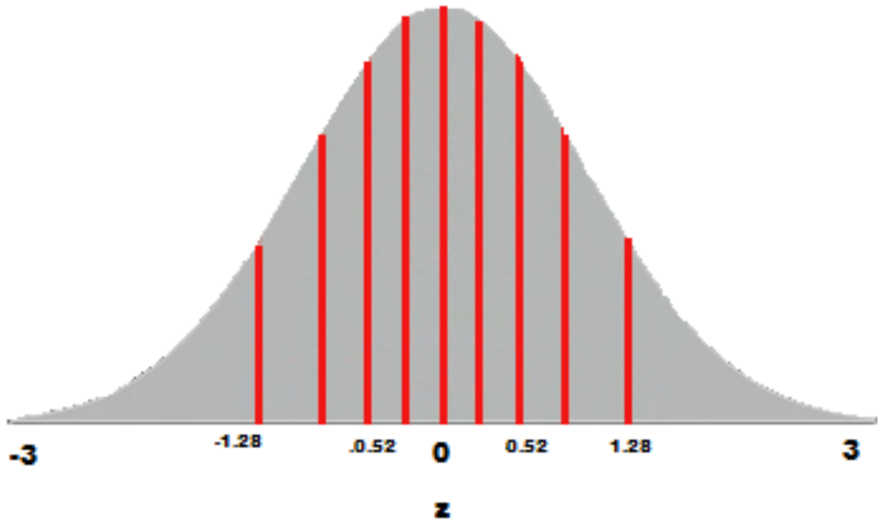
M=0.09,Sd=0.09,Sk=1.64*,K=3.38*
distributions$EXP,N=100
0.4
0.1
0.0
Std Norm Qntls
-2
-1
0
1
2

How to make a QQ plot?

- Do the following values come from a normal distribution?
  7.19, 6.31, 5.89, 4.5, 3.77, 4.25, 5.19, 5.79, 6.79.

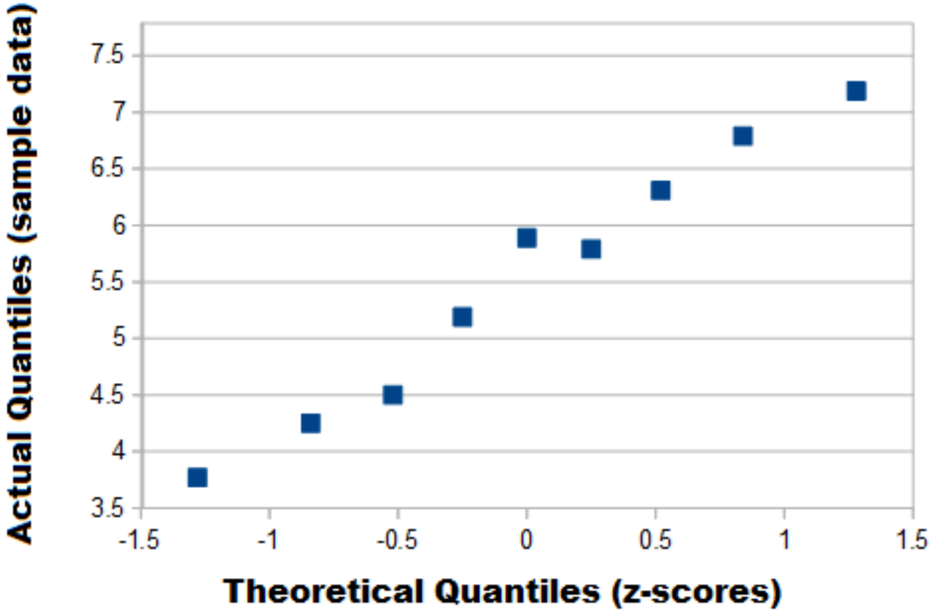
Make a QQ plot

- Step 1: Order the items from smallest to largest.
  - 3.77, 4.25, 4.50, 5.19, 5.79, 5.89, 6.31, 6.79, 7.19
- Step 2: Draw a normal distribution curve.
  - Divide the curve into n+1 segments. We have 9 values, so divide the curve into 10 equally-sized areas. For this example, each segment is 10% of the area (because 100% / 10 = 10%).
- Step 3: Find the z-value (cut-off point) for each segment (z-table).
  - 10% = -1.28, 20% = -0.84, 30% = -0.52, 40% = -0.25, 50% = 0
  - 60% = 0.25, 70% = 0.52, 80% = 0.84, 90% = 1.28, 100% = 3.0



- Step 4: Plot your data set values (Step 1) against your normal distribution cut-off points (Step 3).

How to make qq plot in R?



So…what do you do?

- If you find a mistake, fix it.
- If you find an outlier, trim it or delete it.
- If your distributions are askew, transform the data.

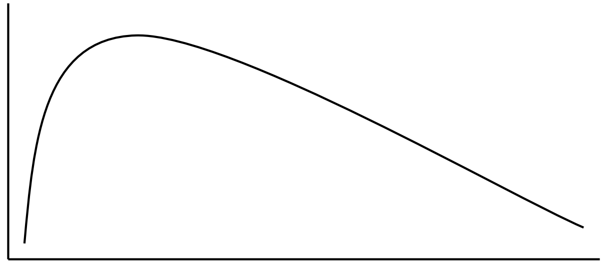
Dealing with Outliers

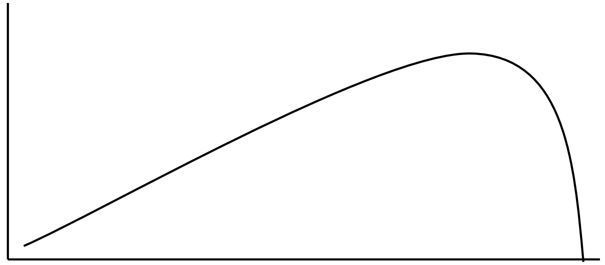
- First, try to explain it.
- In a normal distribution 0.4% are outliers (>2.7 SD) and 1 in a million is an extreme outlier (>4.72 SD).
- For analyses you can:
  - Delete the value – crude but effective.
  - Change the outlier to value ~3 SD from mean.
  - "Winsorize" it (make = to next highest value).
  - "Trim" the mean – recalculate mean from data within interquartile range.

Dealing with skewed distributions

(Skewness and kurtosis greater than +/- 2)

- Positive skew is reduced by using the square root or log.
- Negative skew is reduced by taking square root of data values.





Skewness of data & data transformation

- Ex: river water turbidity
  - 1.0, 1.2, 1.1, 1.1, 2.4, 2.2, 2.6, 4.1, 5.0, 10.0, 4.0, 4.1, 4.2, 4.1, 5.1, 4.5, 5.0, 15.2, 10.0, 20.0, 1.1, 1.1, 1.2, 1.6, 2.2, 3.0, 4.0, 10.5

```
Turbidity = c(1.0, 1.2, 1.1, 1.1, 2.4,
2.2, 2.6, 4.1, 5.0, 10.0, 4.0, 4.1,
4.2, 4.1, 5.1, 4.5, 5.0, 15.2, 10.0,
20.0, 1.1, 1.1, 1.2, 1.6, 2.2, 3.0,
4.0, 10.5)
```

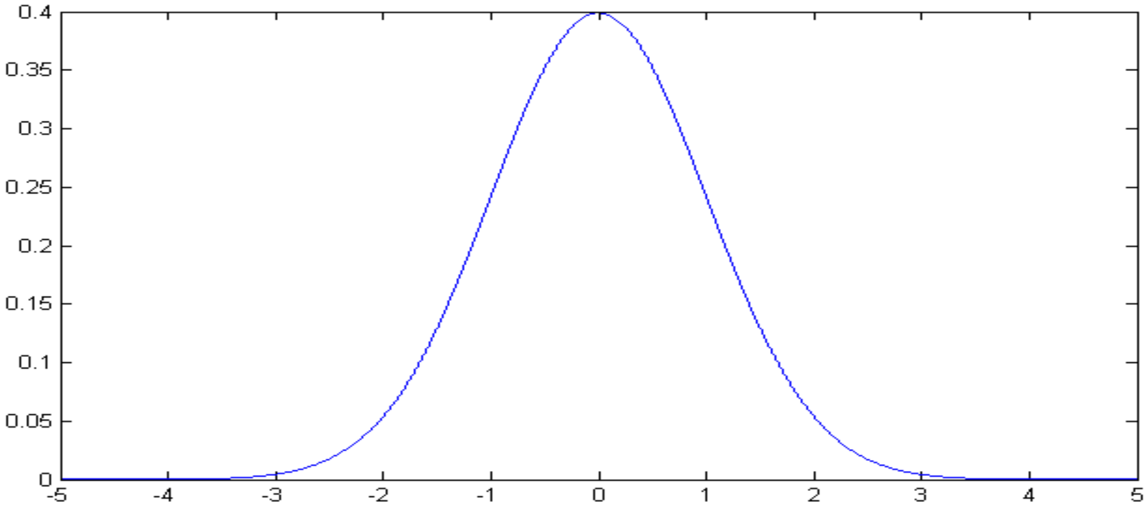
```
# QQ plot



# Square root transformation



# Cube root transformation
# Avoid complex numbers
```

Central Limit Theorem

The distribution of the sum (or mean) of a set of n identically-distributed random variables Xi approaches a normal distribution as n → ∞.

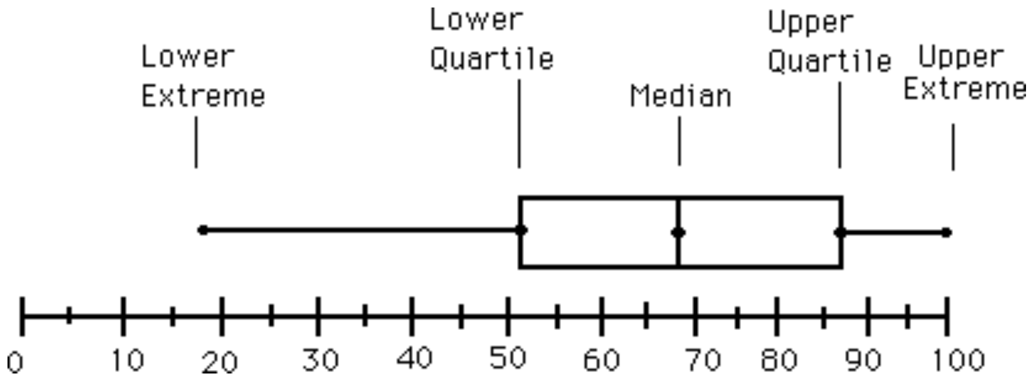
The common parametric statistical tests, like t-test and ANOVA assume normally-distributed data, but depend on sample mean and variance measures of the data.

They typically work reasonably well for data that are not normally distributed as long as the samples are not too

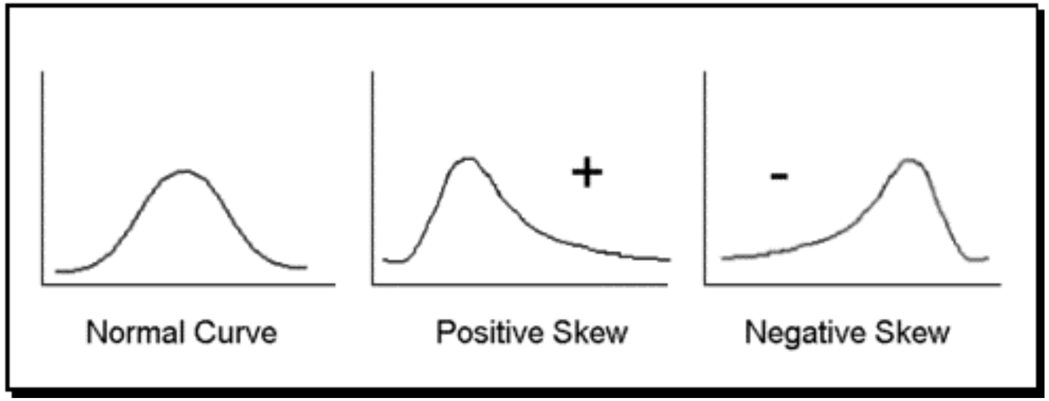


Correcting distributions

- Many statistical tools, including mean and variance, t-test, ANOVA etc. assume data are normally distributed.
- Very often this is not true. The box-and-whisker plot is a good clue

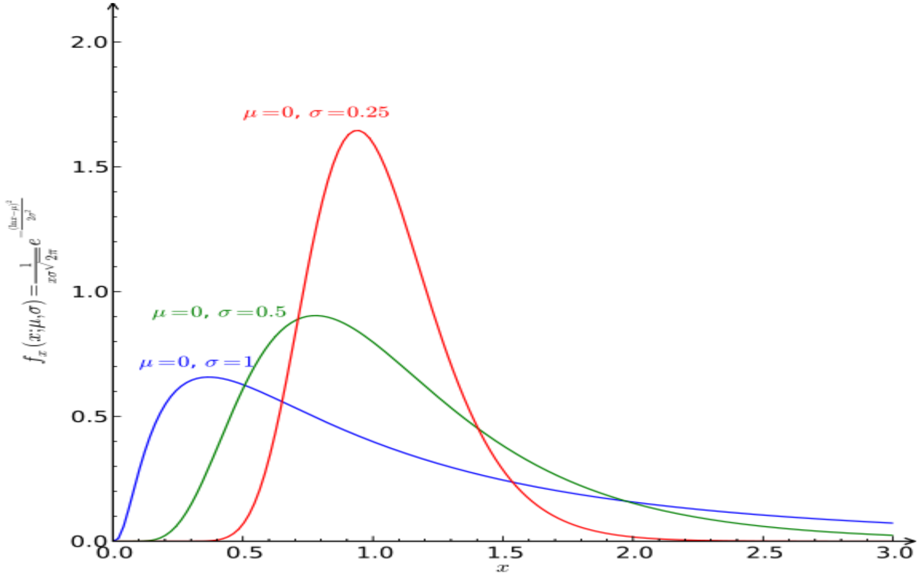


- Whenever its asymmetric, the data cannot be normal. The histogram gives even more information



Correcting distributions

- In many cases these distribution can be corrected before any other processing.
- Examples:
- X satisfies a log-normal distribution, Y=log(X) has a normal dist.



- X Poisson with mean k and sdev. sqrt(k). Then sqrt(X) is approximately normally distributed with sdev 1.

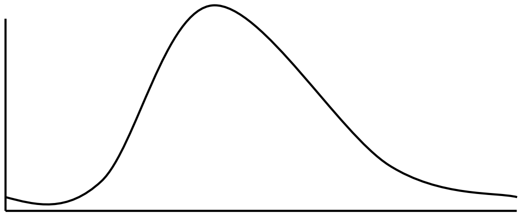
Data transformation

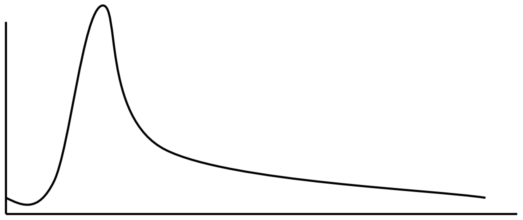
Data distribution	Transformation method
Moderately positive skewness	Square-Root NEWX = sqrt(X)
Substantially positive skewness	Logarithmic (Log 10) NEWX = log10(X)
Substantially positive skewness (with zero values)	Logarithmic (Log 10) NEWX = log10(X + C)
Moderately negative skewness	Square-Root NEWX = sqrt(K – X)
Substantially negative skewness	Logarithmic (Log 10) NEWX = log10(K – X)

C = a constant added to each score so that the smallest score is 1.
K = a constant from which each score is subtracted so that the smallest score is 1; usually equal to the largest score + 1.

Visual Display of Bivariate Data

- So, you have examined each variable for mistakes, outliers and distribution and made any necessary alterations. Now what?
- Look at the relationship between 2 (or more) variables at a time

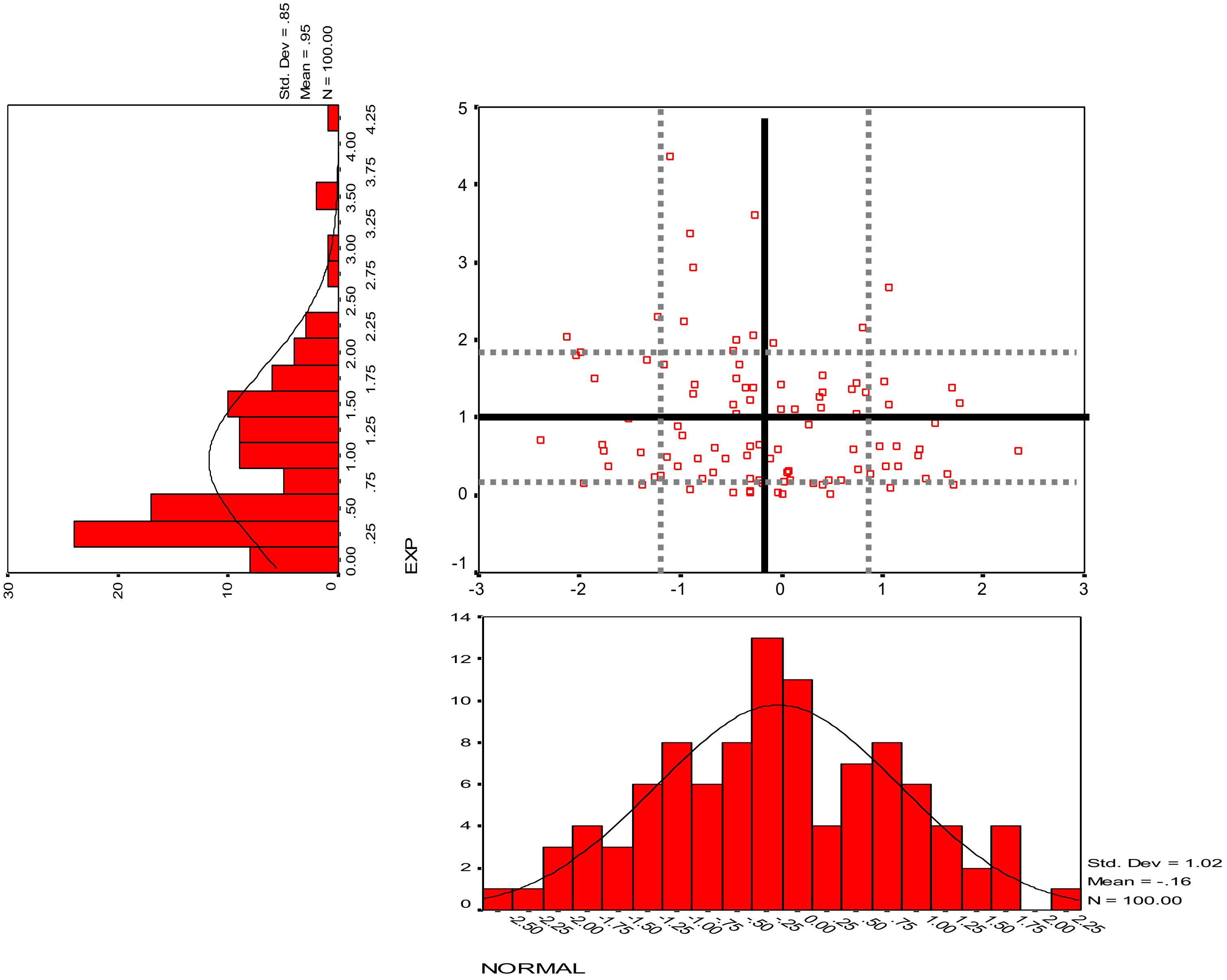




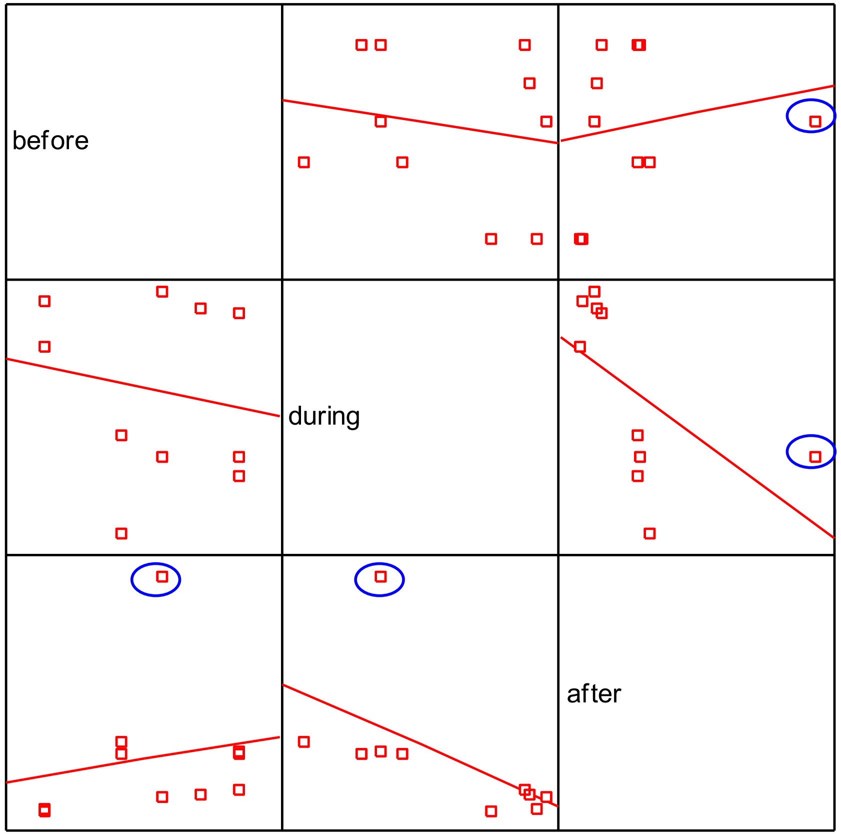
Visual Displays of Bivariate Data

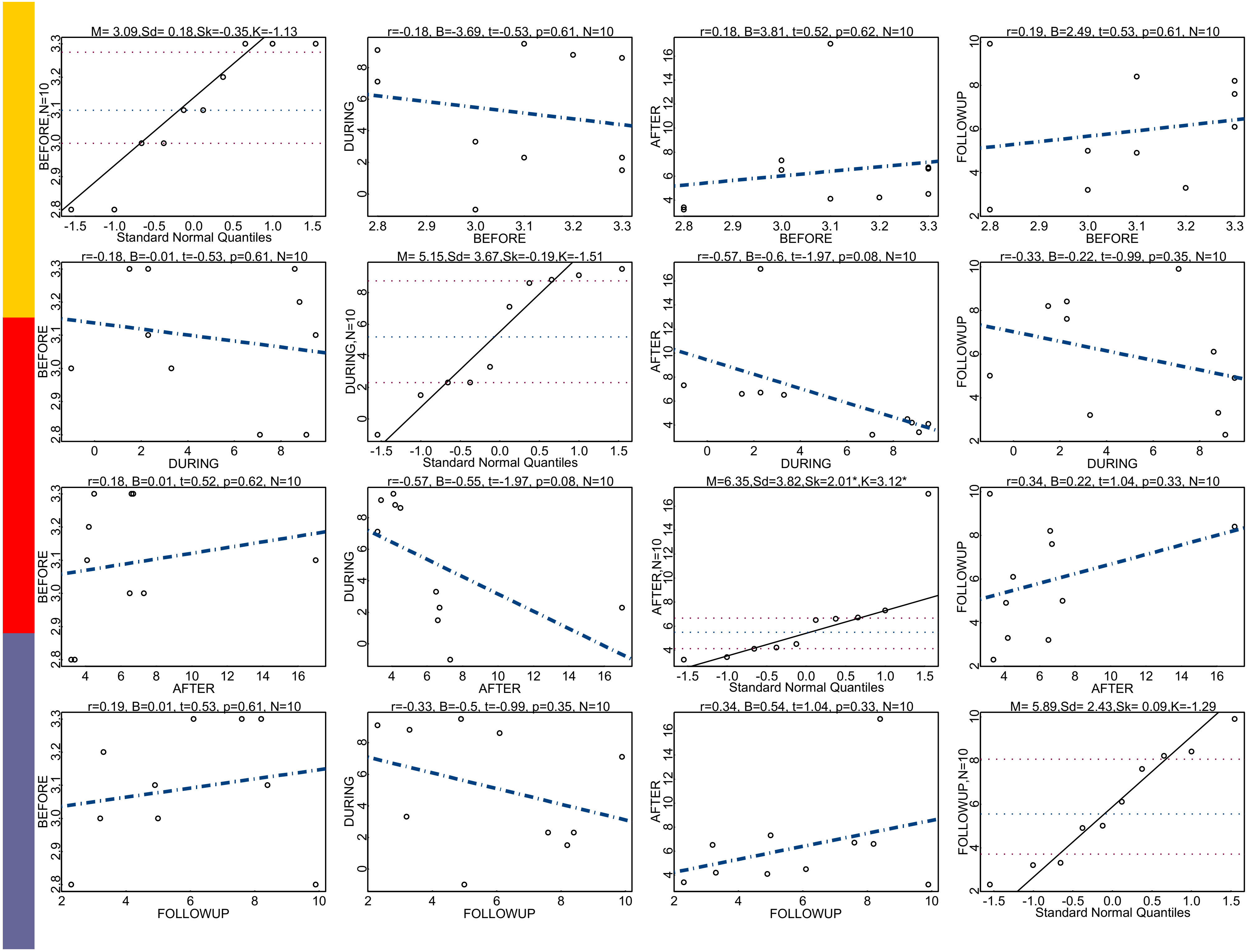
Variable 1	Variable 2	Display Example
Categorical	Categorical	
Categorical	Continuous	
Continuous	Continuous	

Bivariate Distribution

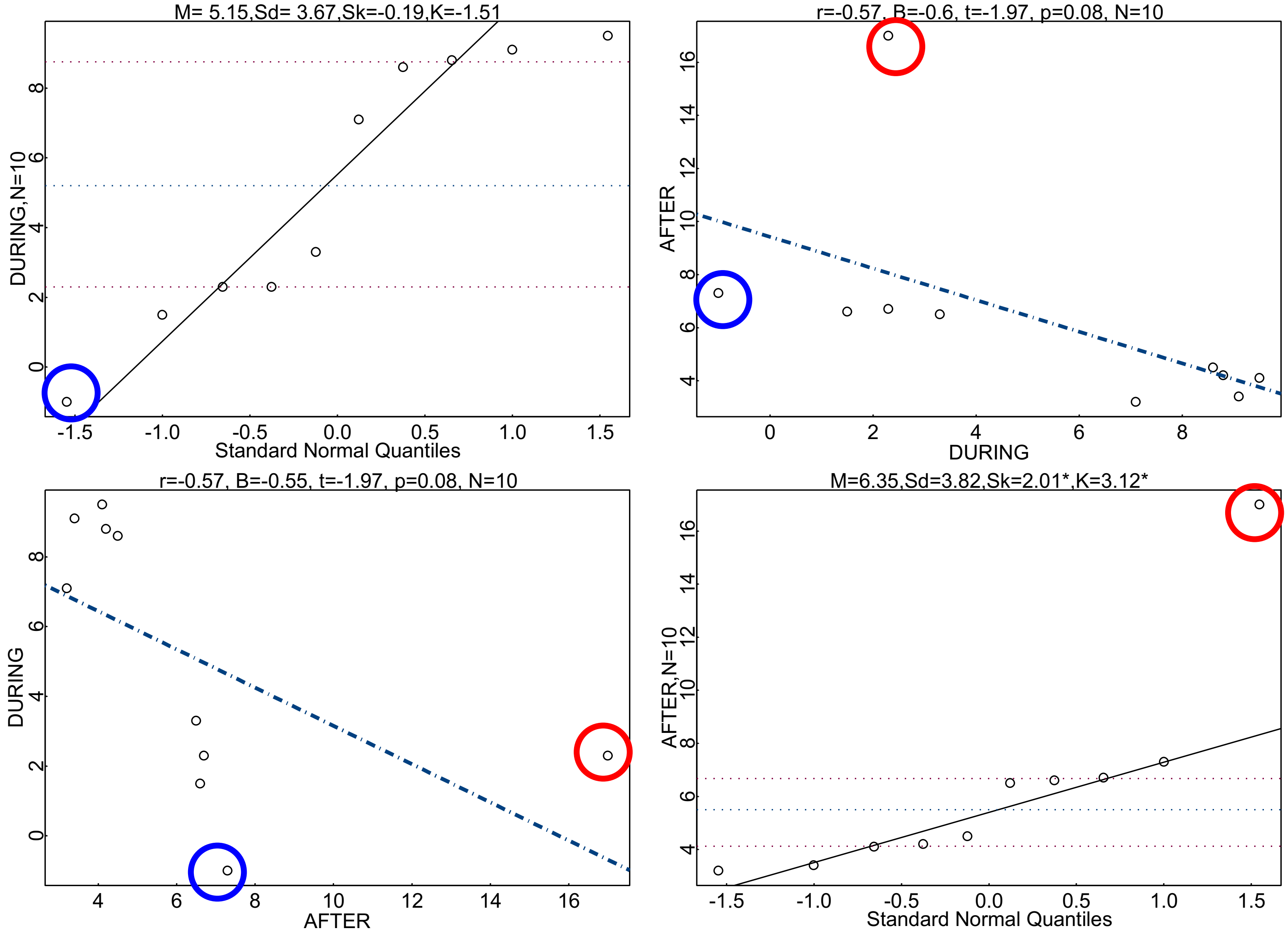


Intro to Scatter plots

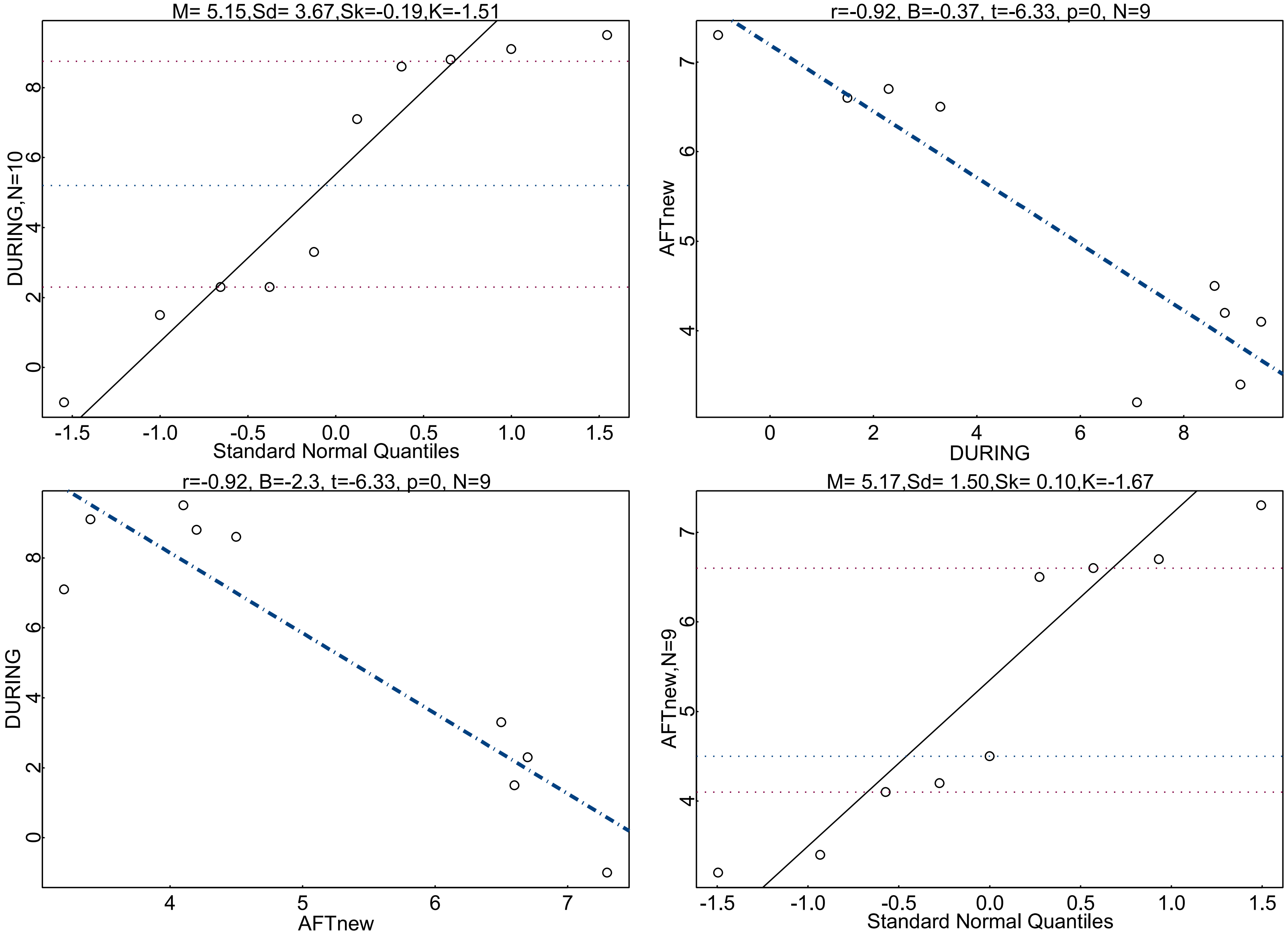



M= 3.09,Sd= 0.18,Sk=-0.35,K=-1.13
r=-0.18, B=-3.69, t=-0.53, p=0.61, N=10
r=0.18, B=3.81, t=0.52, p=0.62, N=10
r=0.19, B=2.49, t=0.53, p=0.61, N=10
BEFORE,N=10
Standard Normal Quantiles
DURING
AFTER
FOLLOWUP
BEFORE
r=-0.18, B=-0.01, t=-0.53, p=0.61, N=10
M= 5.15,Sd= 3.67,Sk=-0.19,K=-1.51
r=-0.57, B=-0.6, t=-1.97, p=0.08, N=10
r=-0.33, B=-0.22, t=-0.99, p=0.35, N=10
DURING,N=10
r=0.18, B=0.01, t=0.52, p=0.62, N=10
r=-0.57, B=-0.55, t=-1.97, p=0.08, N=10
M=6.35,Sd=3.82,Sk=2.01*,K=3.12*
r=0.34, B=0.22, t=1.04, p=0.33, N=10
AFTER,N=10
r=0.19, B=0.01, t=0.53, p=0.61, N=10
r=-0.33, B=-0.5, t=-0.99, p=0.35, N=10
r=0.34, B=0.54, t=1.04, p=0.33, N=10
M= 5.89,Sd= 2.43,Sk= 0.09,K=-1.29
FOLLOWUP,N=10
-1.5 -1.0 -0.5 0.0 0.5 1.0 1.5
2.8 2.9 3.0 3.1 3.2 3.3
0 2 4 6 8
4 6 8 10 12 14 16
2 4 6 8 10

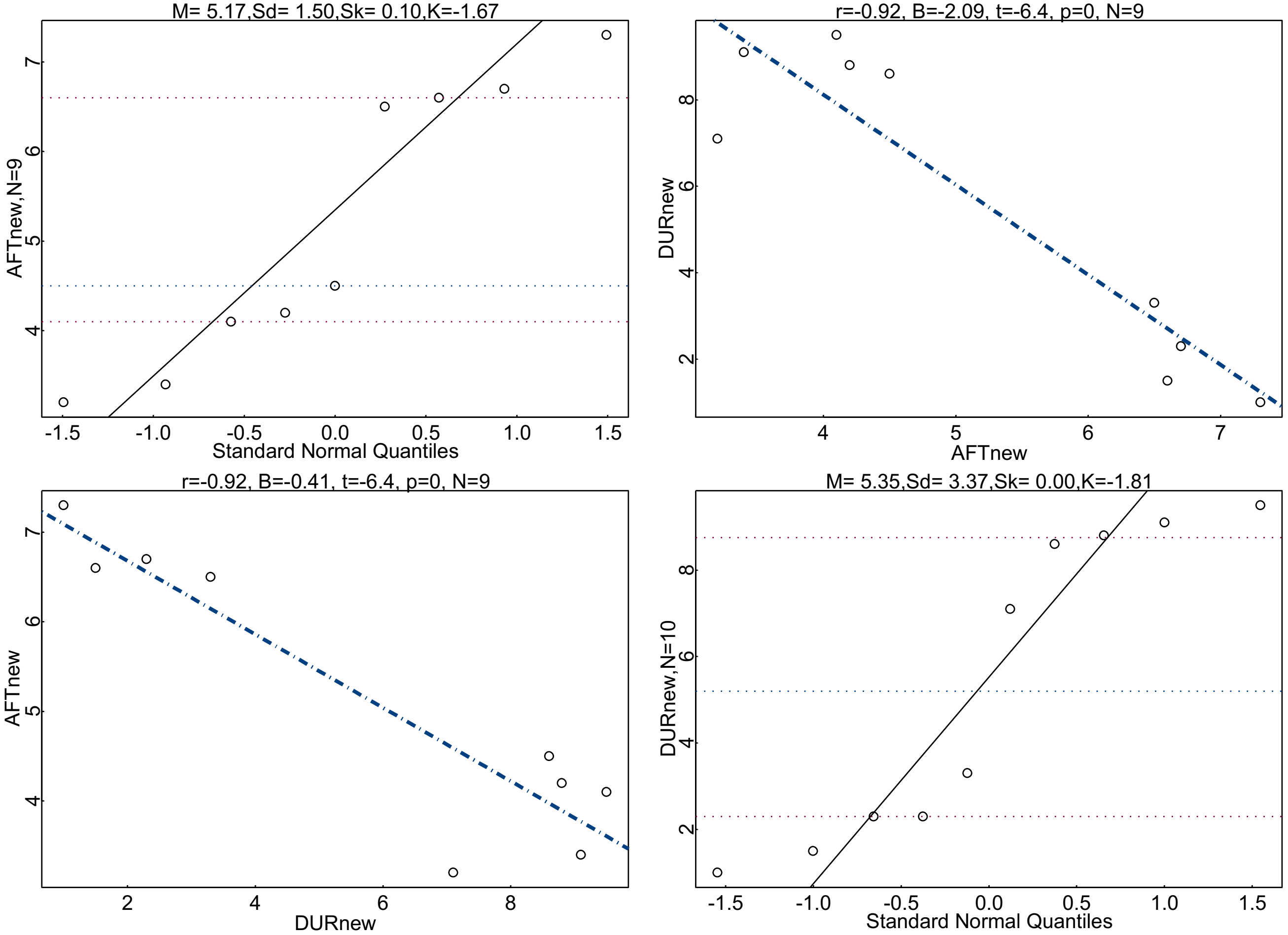
With Outlier and Out of Range Value



Without Outlier

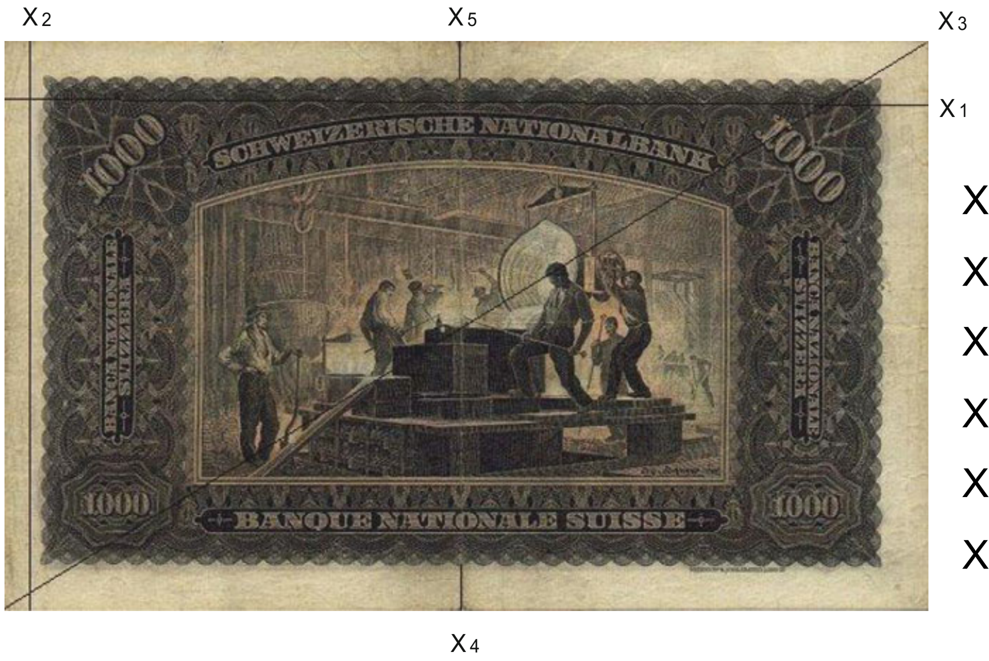


With Corrected Out of Range Value



Scatterplots

- Scatterplots in two and three dimensions help us in seeing separated points or clouds.
- They help us in judging positive or negative dependence.
- Draftman scatterplot matrices are useful for detecting structures conditioned on values of certain variables.
- As the brush of a scatterplot matrix is moving in the point cloud we can study conditional dependence.



X1= length of the bill

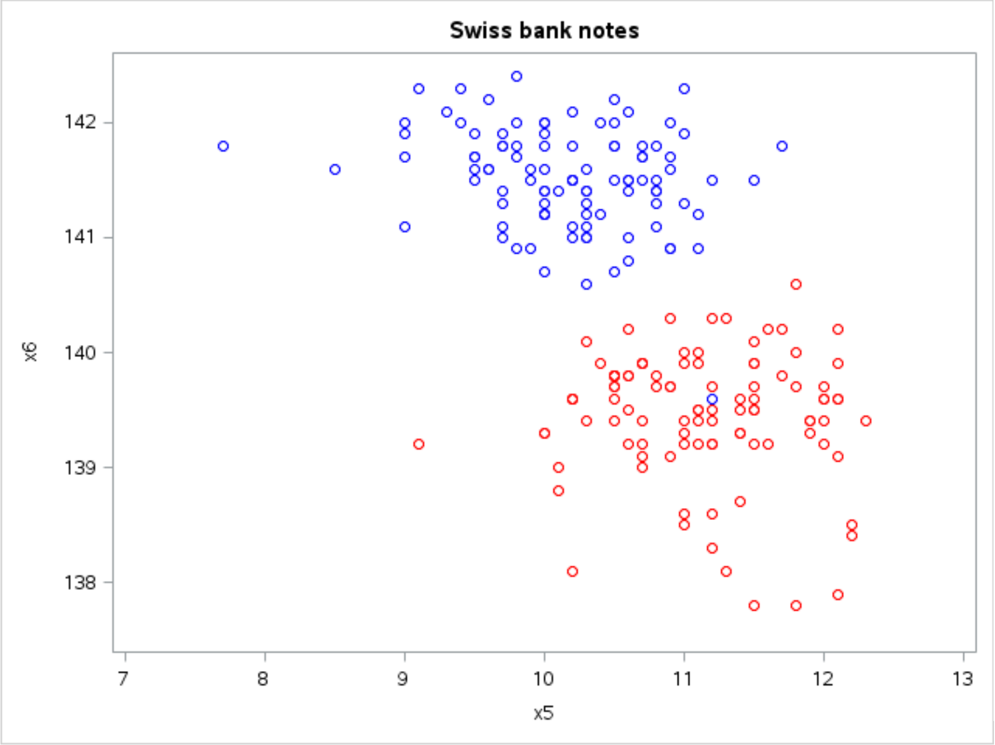
X2= height of the bill (left)

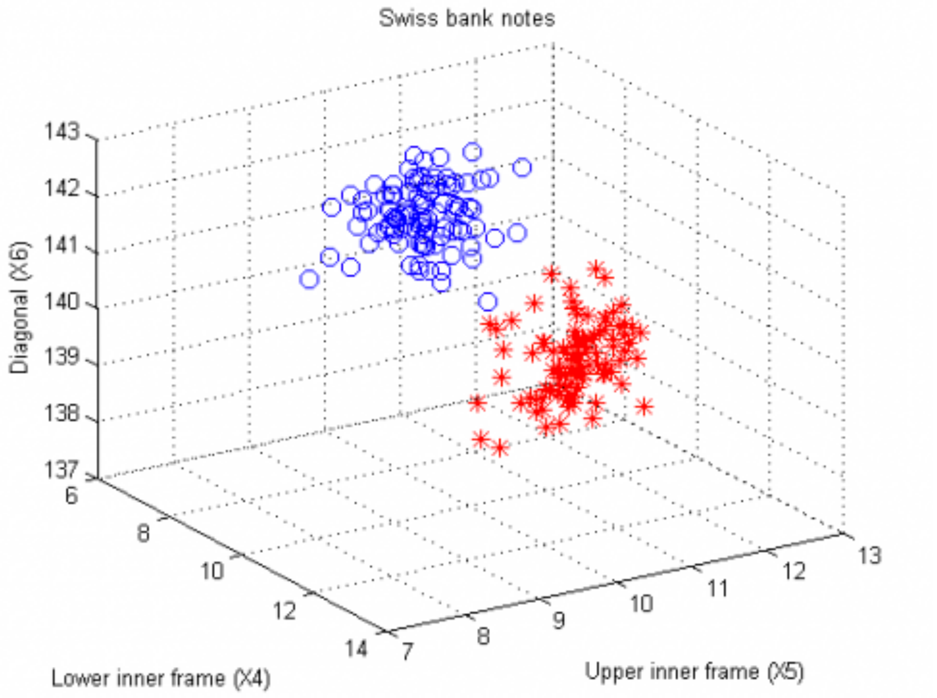
X3=height of the bill (right)

X4=distance of the inner frame to the lower border

X5=distance of the inner frame to the upper border

X6=length of the diagonal of the central picture.

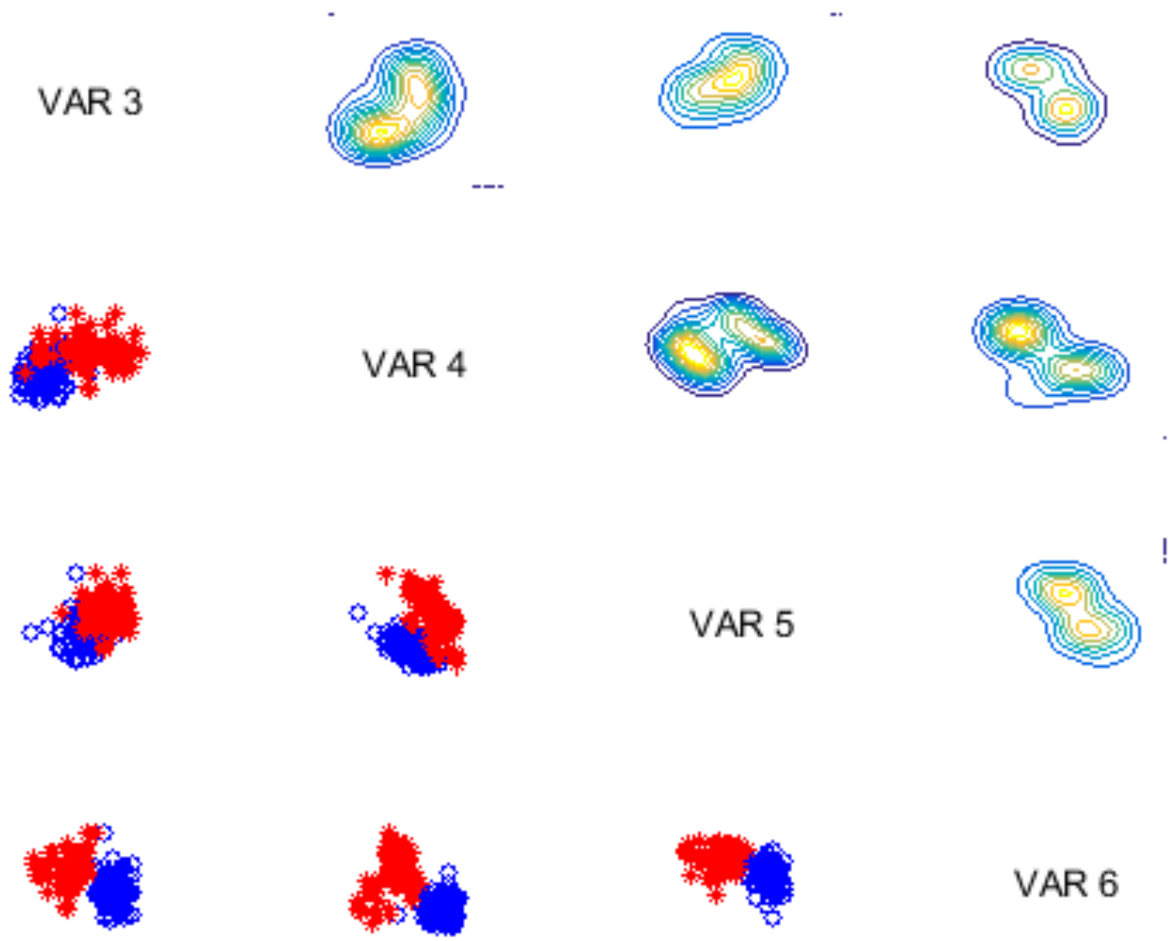




MVAscabank56

MVAscabank456

Draftman plot of the bank notes



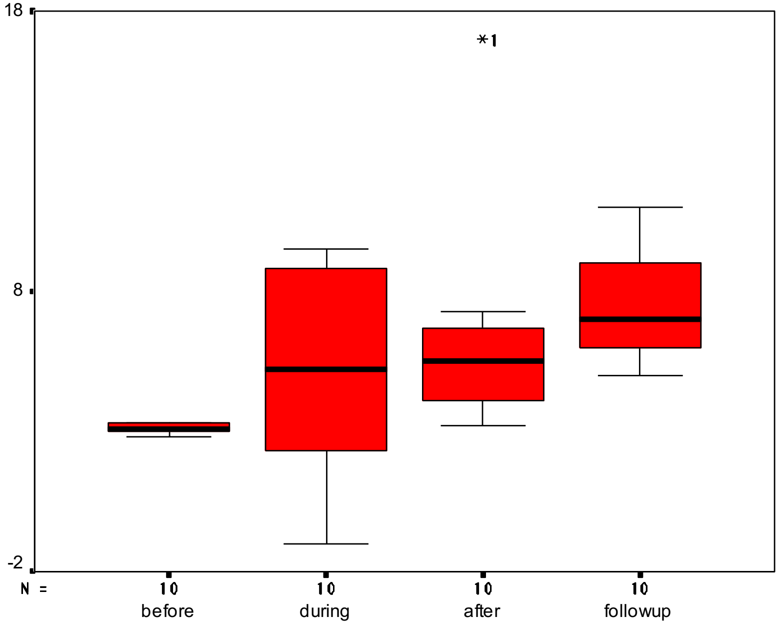
MVAdraftbank4

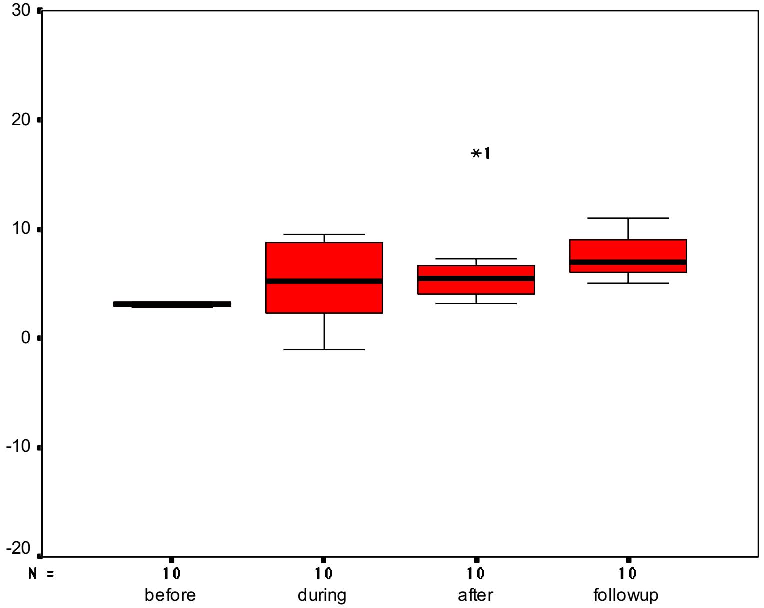
Scatterplots

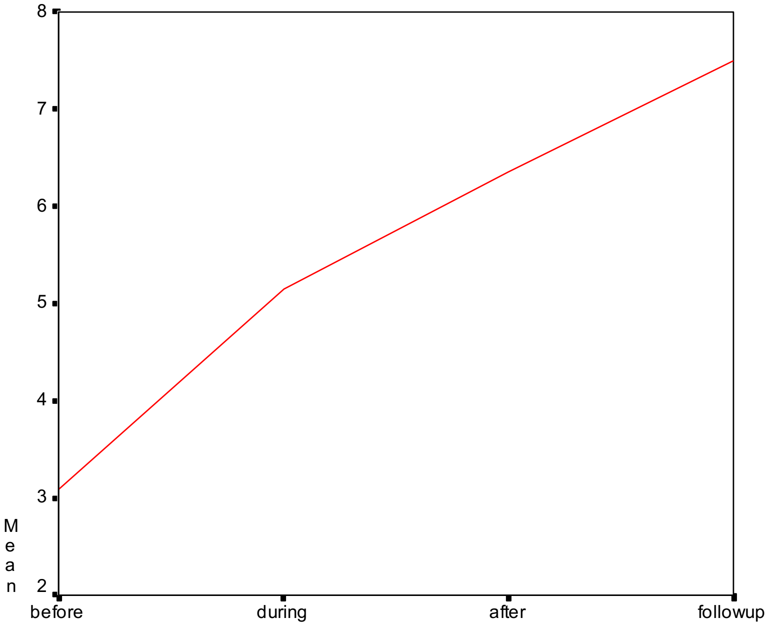
- Scatterplots in two and three dimensions help us in seeing separated points or clouds.
- They help us in judging positive or negative dependence.
- Draftman scatterplot matrices are useful for detecting structures conditioned on values of certain variables.
- As the brush of a scatterplot matrix is moving in the point cloud we can study conditional dependence.

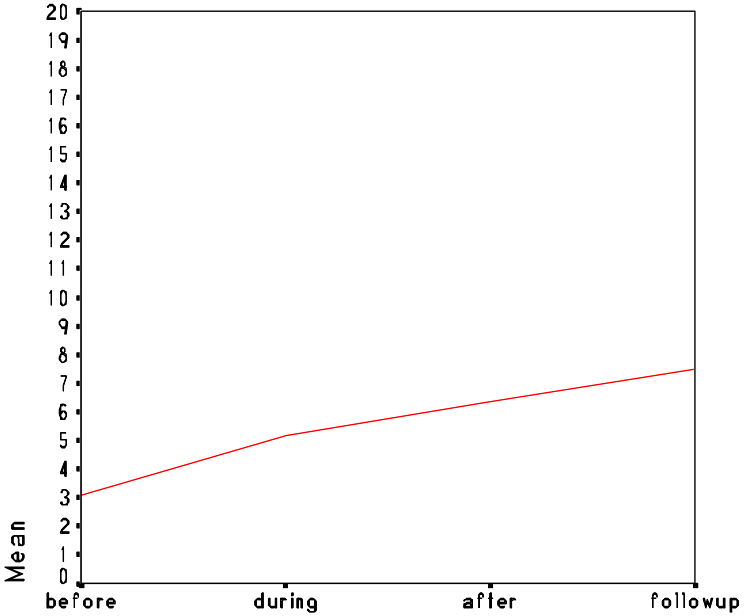
Scales of Graphs

- It is very important to pay attention to the *scale* that you are using when you are plotting.
- Compare the following graphs created from identical data.


18
8
-2
*1
N =
10
10
10
10
before
during
after
followup


30
20
10
0
-10
-20
*1
N =
10
10
10
10
before
during
after
followup


8
7
6
5
4
3
2
Mean
before
during
after
followup


20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Mean
before
during
after
followup

What’s missing?

- pie charts
  - very popular
  - good for showing simple relations of propor
  - Human perception not good at comparing a
  - barplots, histograms usually better (but les

- 3D
  - nice to be able to show three dimensions
  - hard to do well
  - often done poorly
  - 3d best shown through “spinning” in 2D
    - uses various types of projecting into 2D
    - http://www.stat.tamu.edu/~west/bradley/

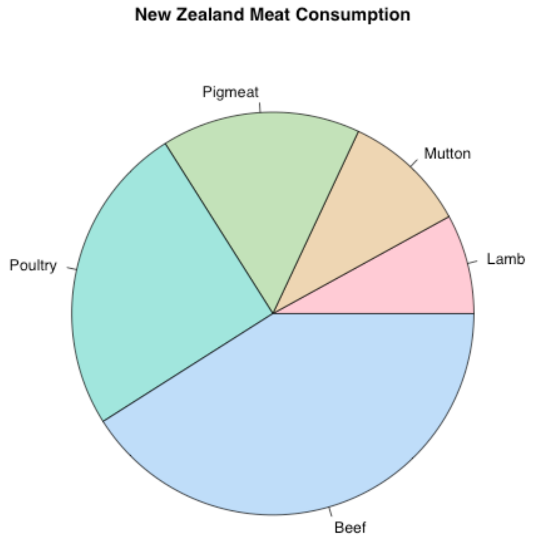
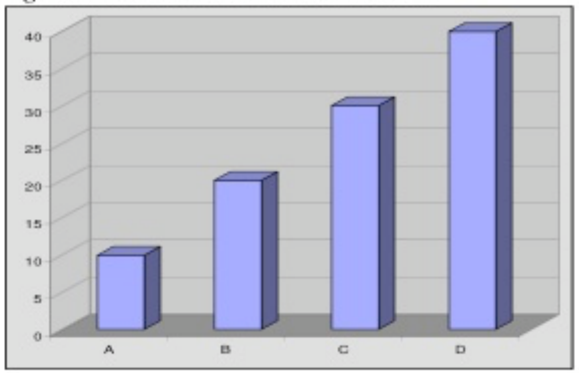
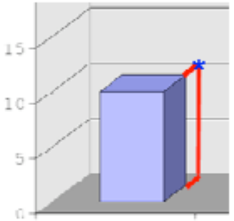


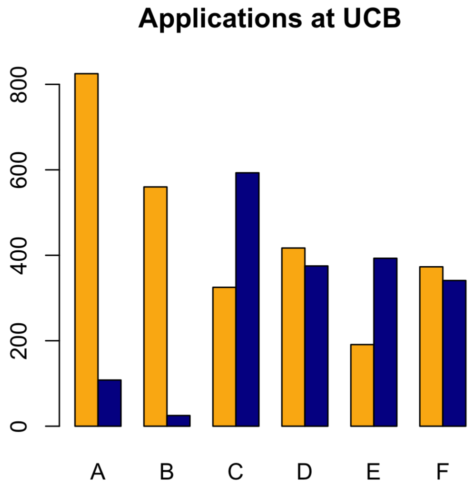
Figure 1. Three-dimensional bar chart.

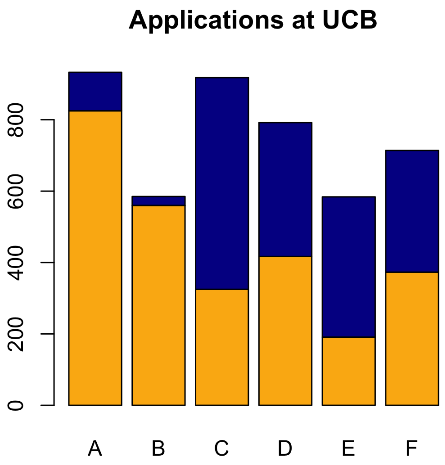




Barcharts and Spineplots

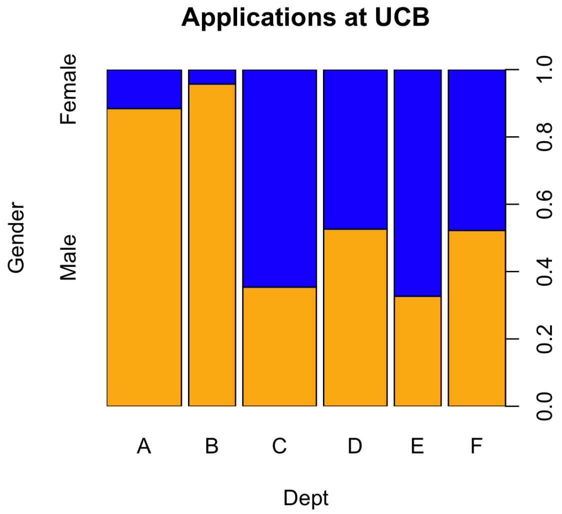
stacked barcharts can be used to compare continuous values across two or more categorical ones.





orange=M blue=F

spineplots show proportions well, but can be hard to interpret

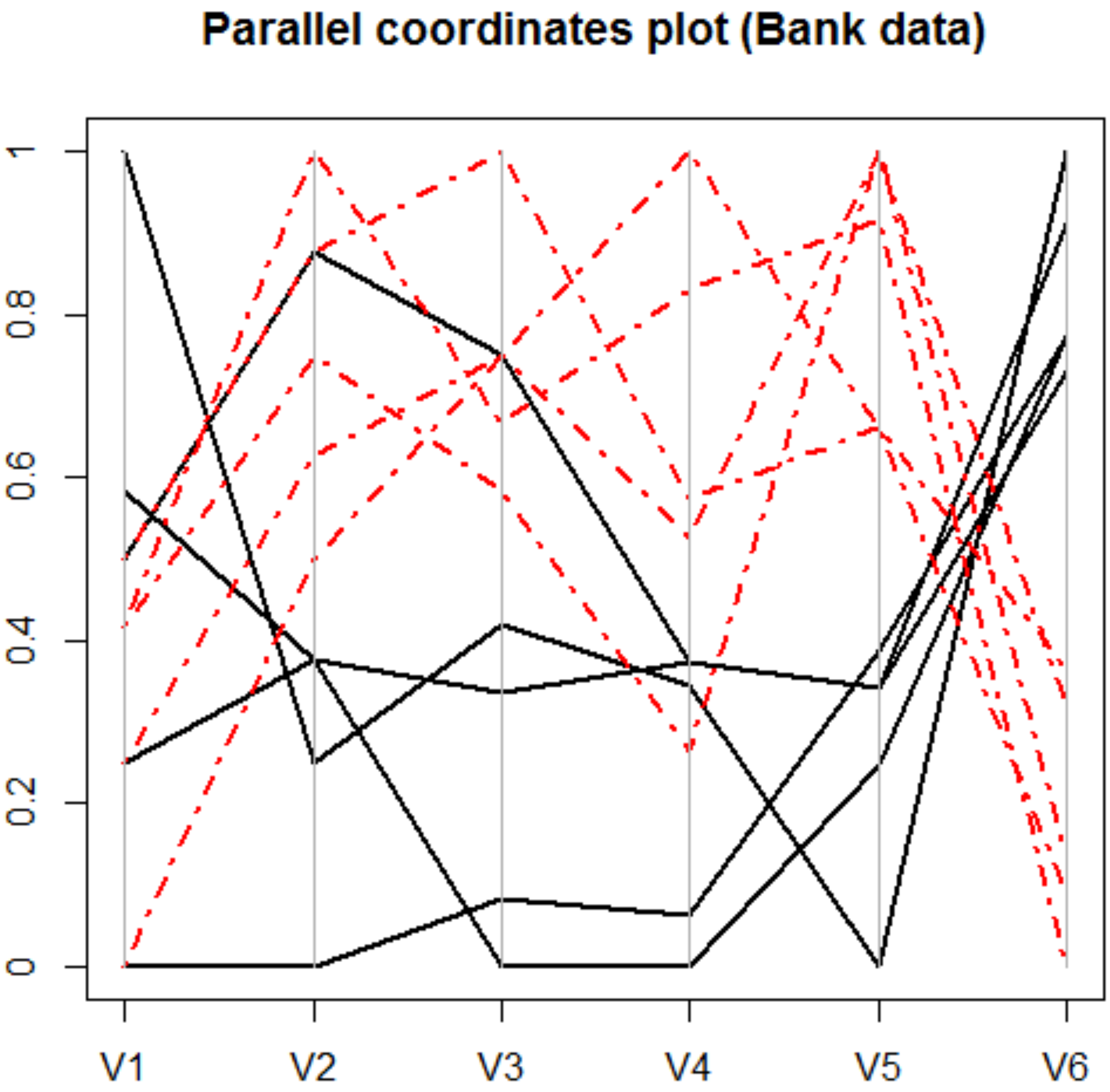


Parallel coordinate plots

- Based on a orthogonal coordinate system.
- Allows to see more than four dimensions.

- Idea:
  - Instead of plotting observations in an orthogonal coordinate system one draws their coordinates in a system of parallel axes. This way of representation is however sensitive to the order of the variables.

observations 96–105



MVAparcoo1

Summary

- Examine all your variables thoroughly and carefully before you begin analysis
- Use visual displays whenever possible
- Transform each variable as necessary to deal with mistakes, outliers, and distributions

- http://www.quantlet.de/
- https://github.com/QuantLet/MVA