Probability and Statistics

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

CHAPTER 4: IT IS ALL ABOUT DATA

1 An introduction to statistics

1.1 Different flavors of statistics

1.2 Trying to understand the true state of affairs

Parameters and statistics

Populations and samples

1.3 True state of affairs + Chance = Sample data

Random and independent samples

1.4 Sampling distributions

Formal definition of a statistics

Sample moments

Sampling from a finite population

Strategies for variance estimation - The Delta method

1.5 The Standard Error of the Mean: A Measure of Sampling Error

1.6 Making formal inferences about populations: a preview to hypothesis testing

2 Exploring data

2.1 Looking at data

2.2 Outlier detection and influential observations

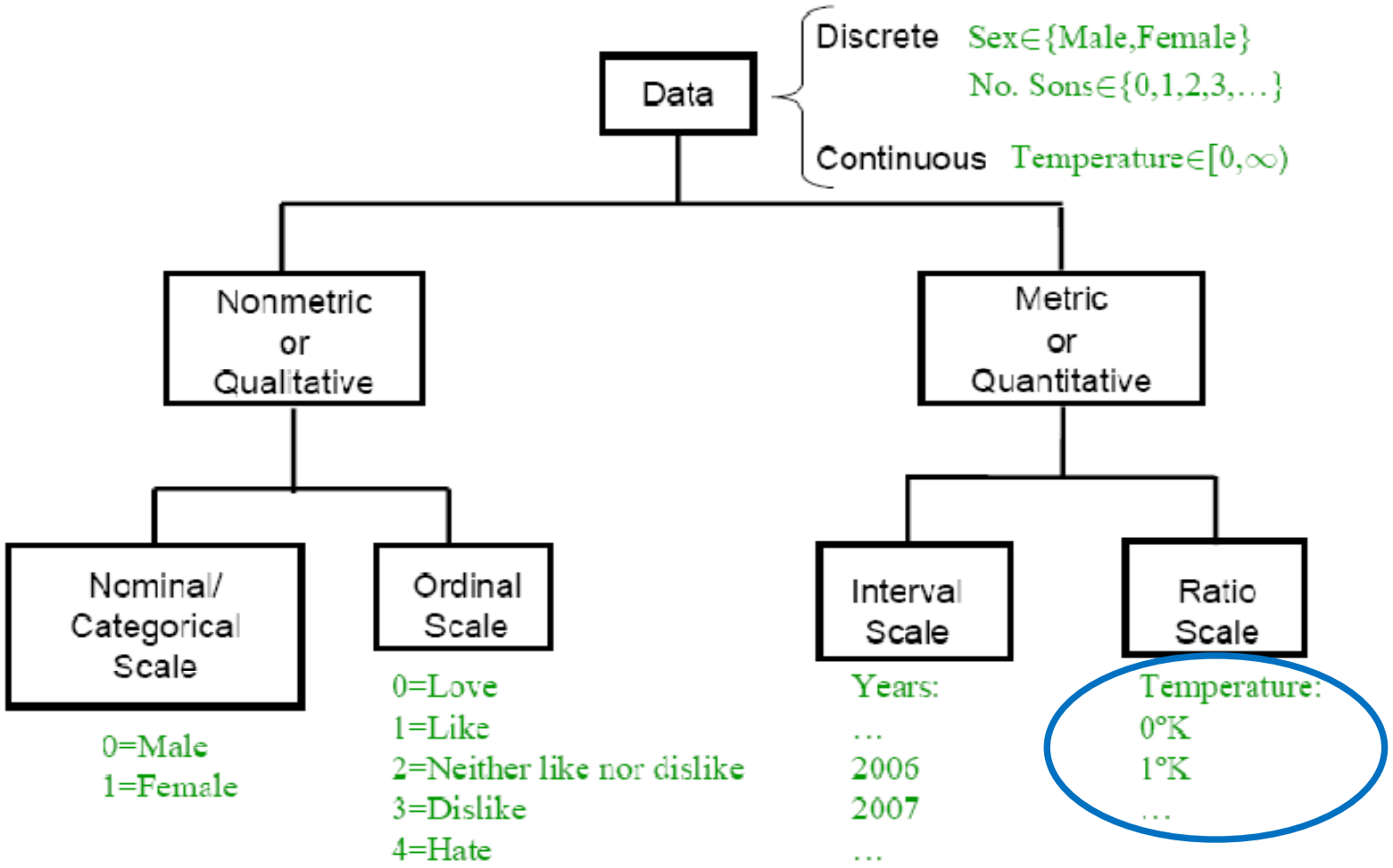
2.3 Exploratory Data Analysis (EDA)

2.4 Box plots and violin plots

2.5 QQ plots

2 Exploring data

2.1 Looking at data



Scales of measurement

- The Celsius-scale is defined by the follow two points:
  - The triple point of water is defined as 0.01 ^{o}C.
  - One degree Celsius equals the change of temperature with one degree on the ideal gas-scale.

The triple point is a theoretical point where the three phases of a matter (for example water) come together. This means that liquid, solid and gas phase from a matter appear at the same time. This is practically impossible.

Set points	**Fahrenheit**	**Celsius**	**Kelvin**
Water boils	212	100	373
Body temperature	98.6	37	310
Water freezes	32	0	273
Absolute zero	-460	-273	0

Metric variables: watch out for "allowable operations"

- **Interval Scale**. You are also allowed to quantify the difference between two interval scale values but there is no natural zero. For example, temperature scales are interval data with 25C warmer than 20C and a 5C difference has some physical meaning. Note that 0C is arbitrary, so that it does not make sense to say that 20C is twice as hot as 10C.

- **Ratio Scale**. You are also allowed to take ratios among ratio scaled variables. Physical measurements of height, weight, length are typically ratio variables. It is now meaningful to say that 10 m is twice as long as 5 m. This ratio hold true regardless of which scale the object is being measured in (e.g. meters or yards). This is because there is a natural zero.

Non-metric variables: need for "coding" them

3 "dummy variables are sufficient !!!

Hair Colour
{Brown, Blond, Black, Red}

No order →

$$\left(x_{Brown}, x_{Blond}, x_{Black}, x_{Red}\right) \in \{0,1\}^4$$

Peter: Black	Peter: $\{0,0,1,0\}$
Molly: Blond	Molly: $\{0,1,0,0\}$
Charles: Brown	Charles: $\{1,0,0,0\}$

Company size
{Small, Medium, Big}

Implicit order →

$$x_{size} \in \{0,1,2\}$$

Company A: Big	Company A: 2
Company B: Small	Company B: 0
Company C: Medium	Company C: 1

Dummy variables

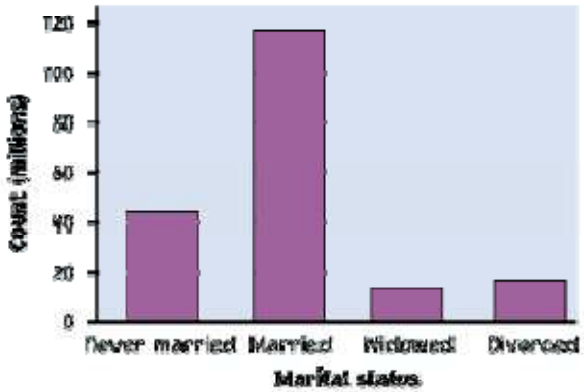
	x_{brown}	x_{blond}	x_{black}	x_{red}	$x_{brown/red}$	$x_{blond/red}$	$X_{black/red}$
Peter	0	0	1	0	**0**	**0**	**1**
Molly	0	1	0	0	**0**	**1**	**0**
Charles	1	0	0	0	**1**	**0**	**0**
Mindy	0	0	0	1	**0**	**0**	**0**

3 "dummy variables are sufficient !!!

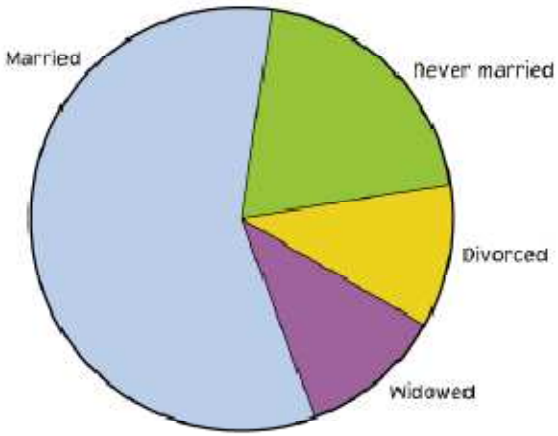
Ways to chart categorical data

Because the variable is categorical, the data in the graph can be ordered any way we want (alphabetical, by increasing value, by year, by personal preference, etc.)

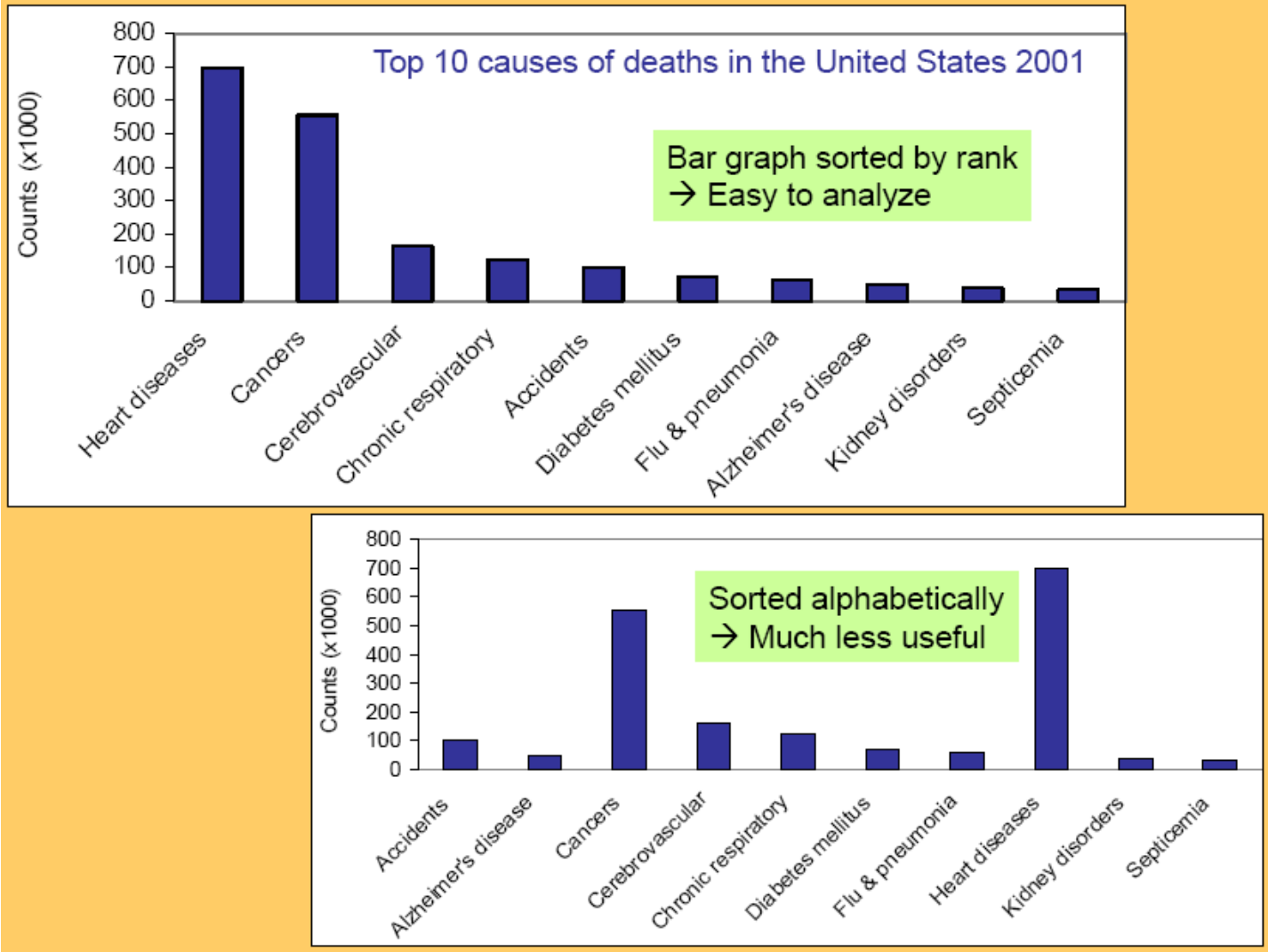
- **Bar graphs**
  Each category is represented by a bar.



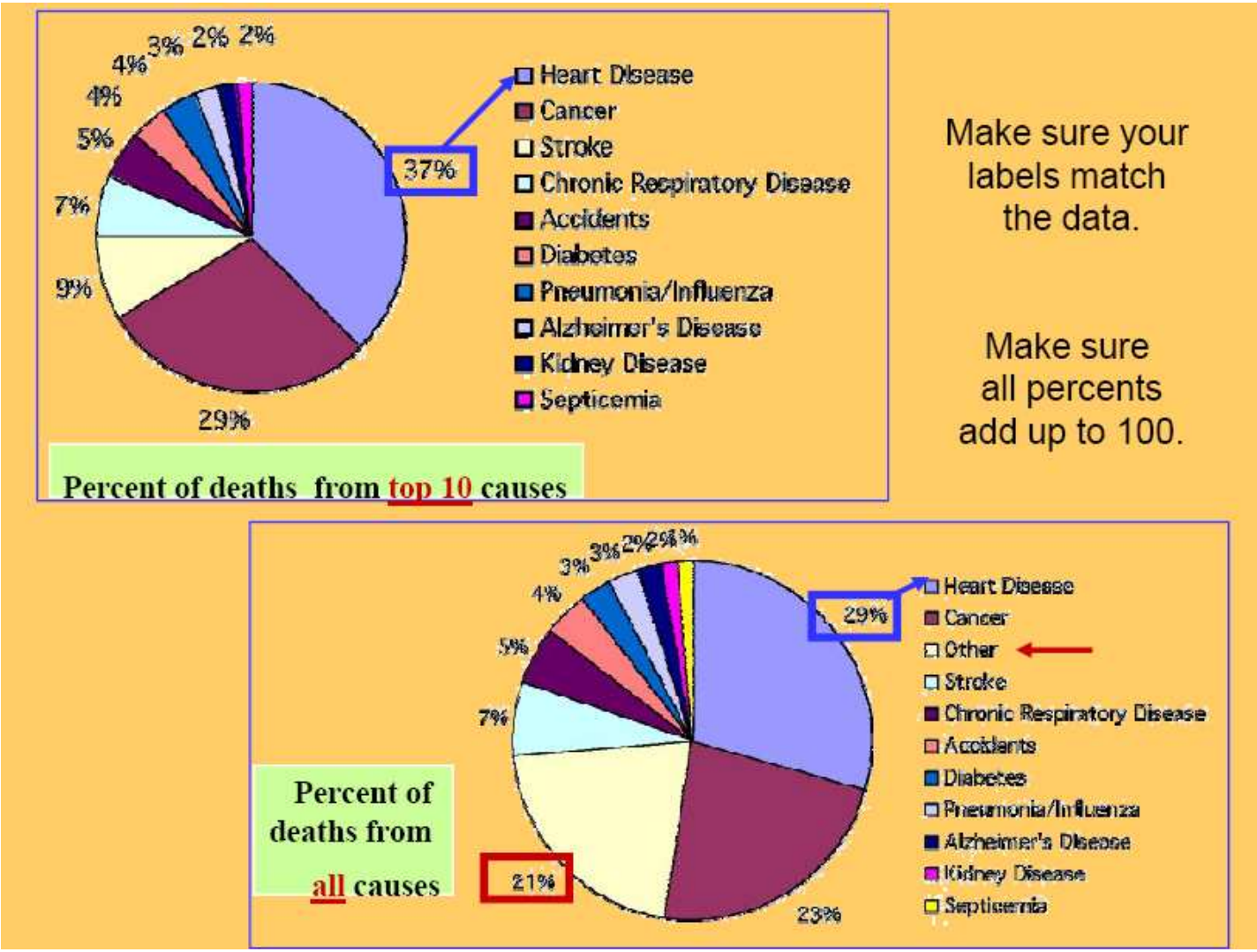
- **Pie charts**
  The slices must represent the parts of one whole.



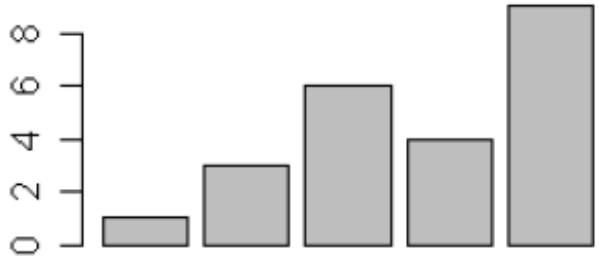
Bar graphs: no meaningful ordering in the categories for categorical data!



Pie charts: clearly define the "whole" pie

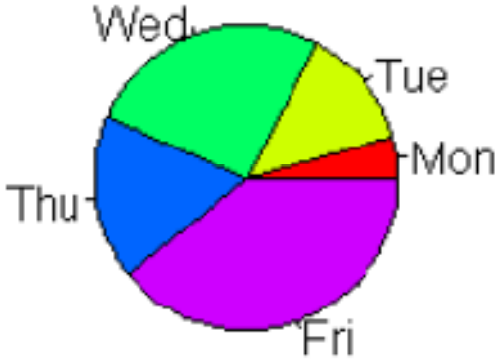


Easy R code examples (http://www.harding.edu/fmccown/R/)



Define the cars vector with 5 values
cars <- c(1, 3, 6, 4, 9)

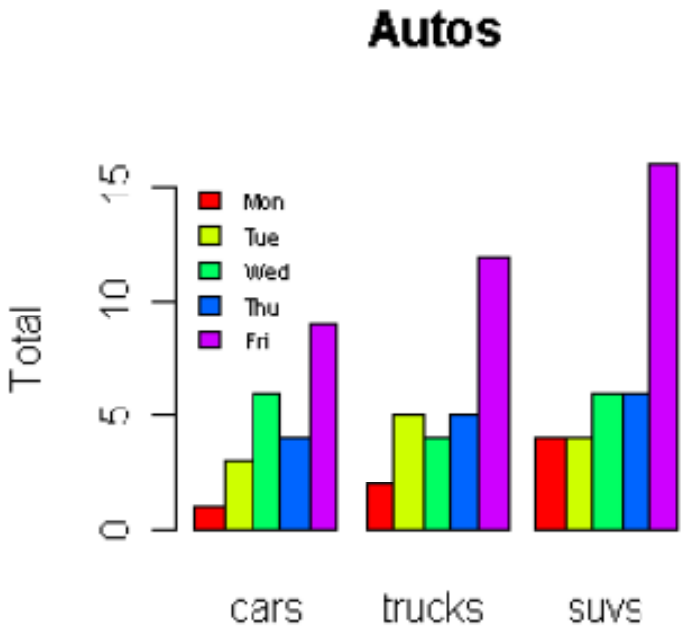
Graph cars
barplot(cars)



Define cars vector with 5 values
cars <- c(1, 3, 6, 4, 9)

Create a pie chart with defined heading and
custom colors and labels
pie(cars, main="Cars",
col=rainbow(length(cars)),
labels=c("Mon","Tue","Wed","Thu","Fri"))

R code examples (http://www.harding.edu/fmccown/R/)



Read values from tab-delimited autos.dat
autos_data <- read.table("C:/R/autos.dat", header=T, sep="\t")

# Expand right side of clipping rect to make
# room for the legend
par(xpd=T, mar=par()$mar+c(0,0,0,4))

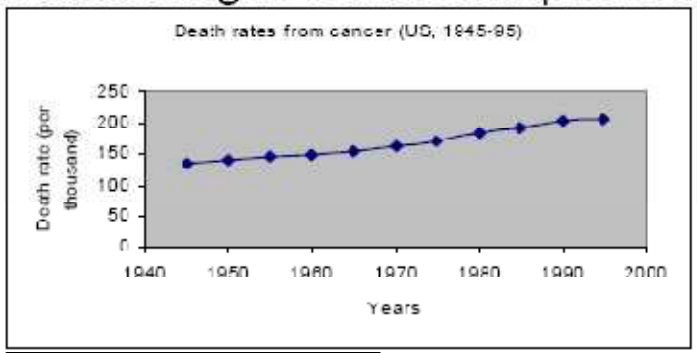
Graph autos (transposing the matrix) using
heat colors, put 10% of the space between
each bar, and make labels smaller with
horizontal y-axis labels
barplot(t(autos_data), main="Autos", ylab="Total", col=heat.colors(3), space=0.1, cex.axis=0.8, las=1, names.arg=c("Mon","Tue","Wed","Thu","Fri"), cex=0.8)

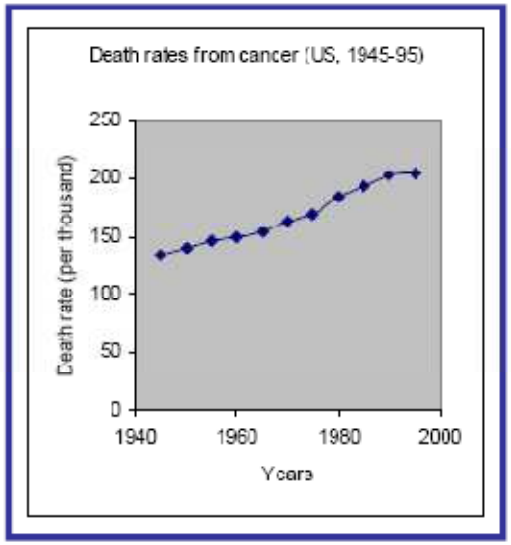
Place the legend at (6,30) using heat colors
legend(6, 30, names(autos_data), cex=0.8, fill=heat.colors(3));

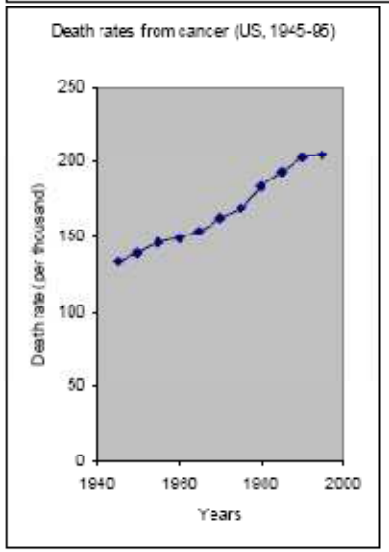
# Restore default clipping rect
par(mar=c(5, 4, 4, 2) + 0.1)

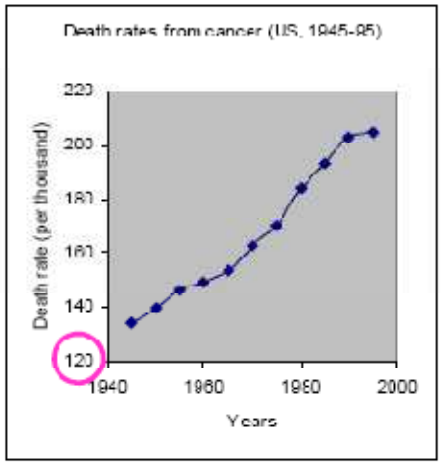
Ways to chart quantitative data: Line graphs

How you stretch the axes and choose your scales can give a different impression.









A picture is worth a thousand words,

BUT

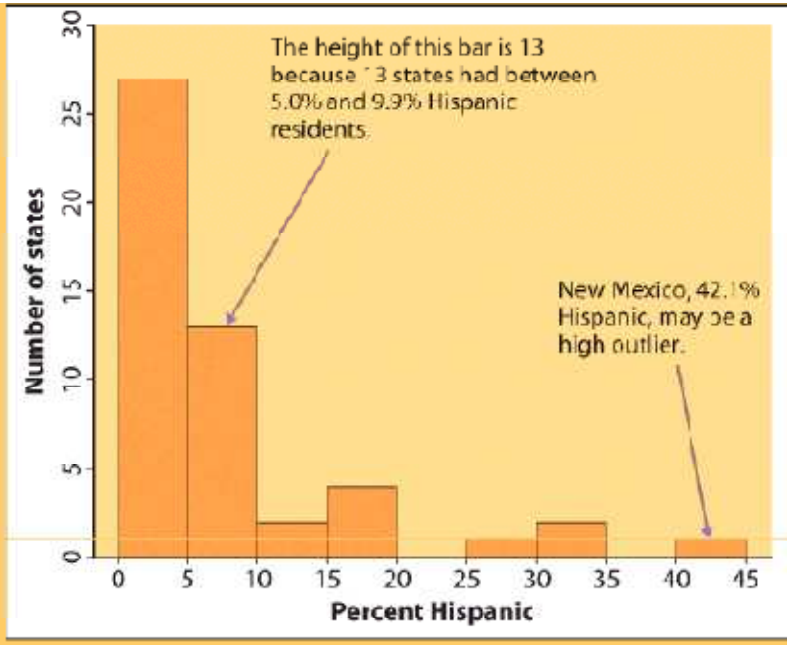
There is nothing like hard numbers.

→ **Look at the scales.**

Ways to chart quantitative data: Histograms

The range of values that a variable can take is divided into equal size intervals.

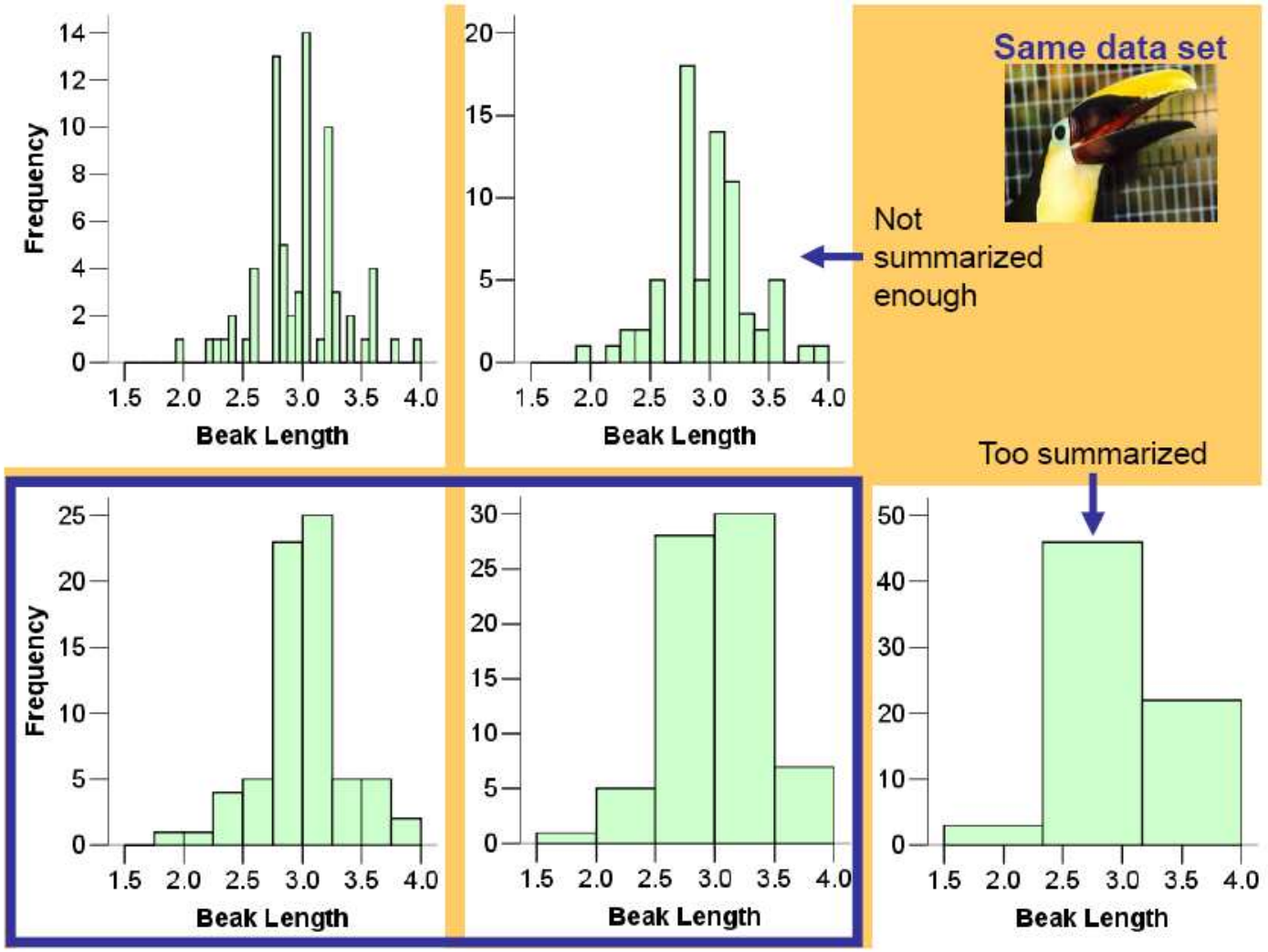
The histogram shows the number of individual data points that fall in each interval.



The first column represents all states with a Hispanic percent in their population between 0% and 4.99%. The height of the column shows how many states (27) have a percent in this range.

The last column represents all states with a Hispanic percent in their population between 40% and 44.99%. There is only one such state: New Mexico, at 42.1% Hispanics.

How to create a histogram?



How to create a histogram?

It is an iterative process – try and try again.

What bin size should you use?

- Not too many bins with either 0 or 1 counts
- Not overly summarized that you loose all the information
- Not so detailed that it is no longer summary

➔ rule of thumb: start with 5 to 10 bins

Look at the distribution and refine your bins

(There isn't a unique or "perfect" solution)

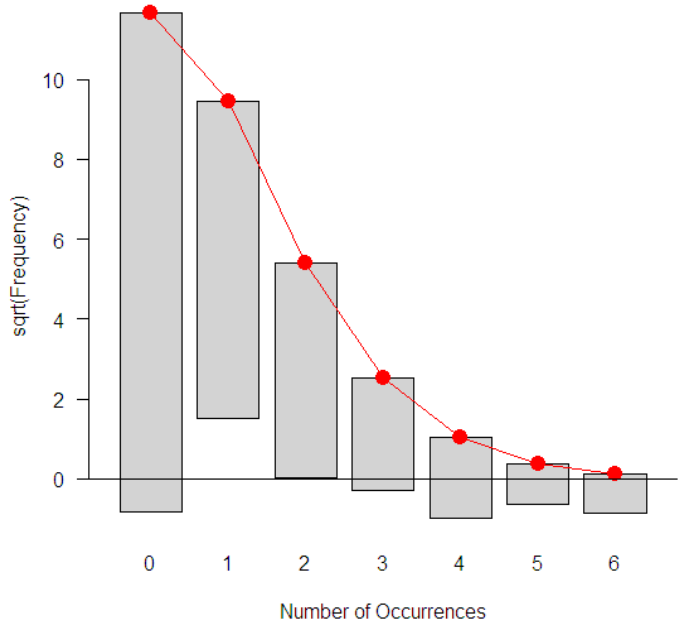
From histograms to probability distribution: “hanging rootograms”

- A ***hanging rootogram*** was suggested by Tukey in 1971, as a graphical means to better compare an observed bar chart or histogram (with equal-width categories) with a theoretical probability distribution.

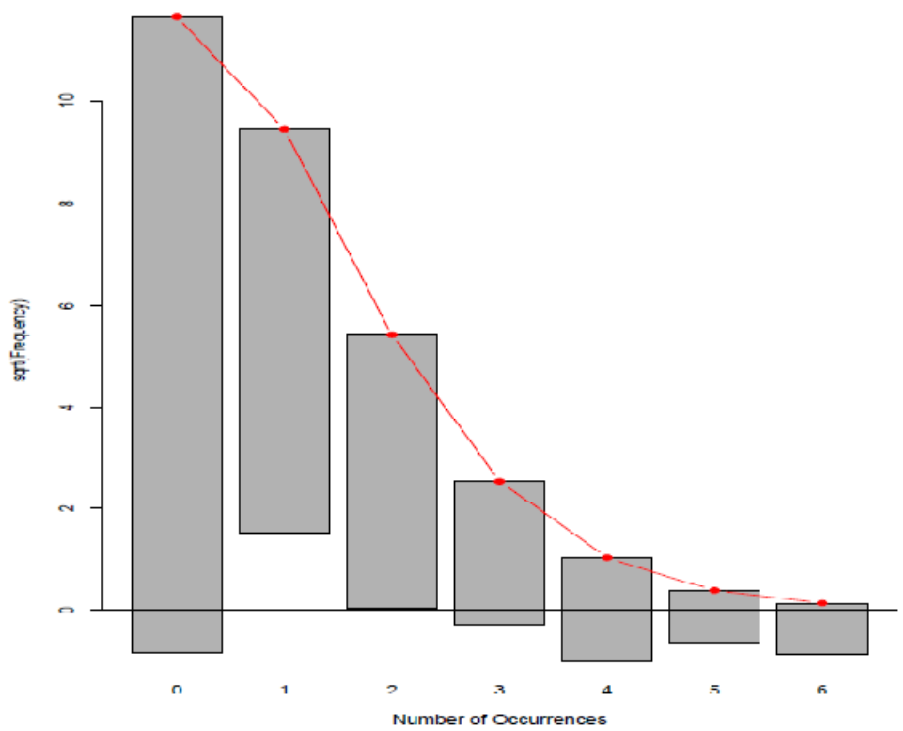
```
library("vcd")
# create data
madison=table(rep(0:6,c(156,63,29,8,4,1,1))
)

# fit a poisson distribution
madisonPoisson=goodfit(madison,"poisson
")

#create rootogram
rootogram(madisonPoisson)
```



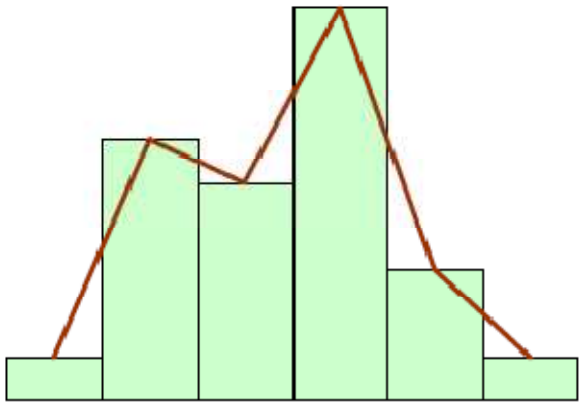
- The comparison is made easier by 'hanging' the observed results from the theoretical curve, so that the discrepancies are seen by comparison with the horizontal axis rather than a sloping curve.
- The vertical axis is scaled to the square-root of the frequencies so as to draw attention to discrepancies in the tails of the distribution.



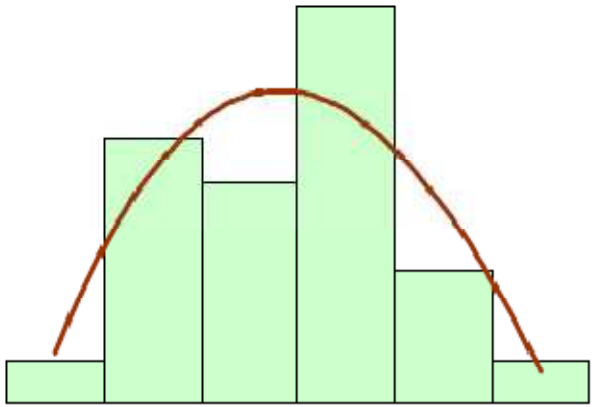
Here: Observed frequencies clearly differ systematically from those predicted under a Poisson model.

The value of bar charts and histograms … be observant

When describing the distribution of a quantitative variable, we look for the overall pattern and for striking deviations from that pattern. We can describe the *overall* pattern of a histogram by its **shape, center**, and **spread.**



Histogram with a line connecting each column → too detailed



Histogram with a smoothed curve highlighting the overall pattern of the distribution

2.2 Outlier detection and influential observations

Outlier detection and influential observations

- Definition of Hawkins [Hawkins 1980]:

"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"

- Statistics-based intuition
  - "Normal data" objects follow a "generating mechanisms", e.g. some given statistical process
  - "Abnormal objects" deviate from this generating mechanism

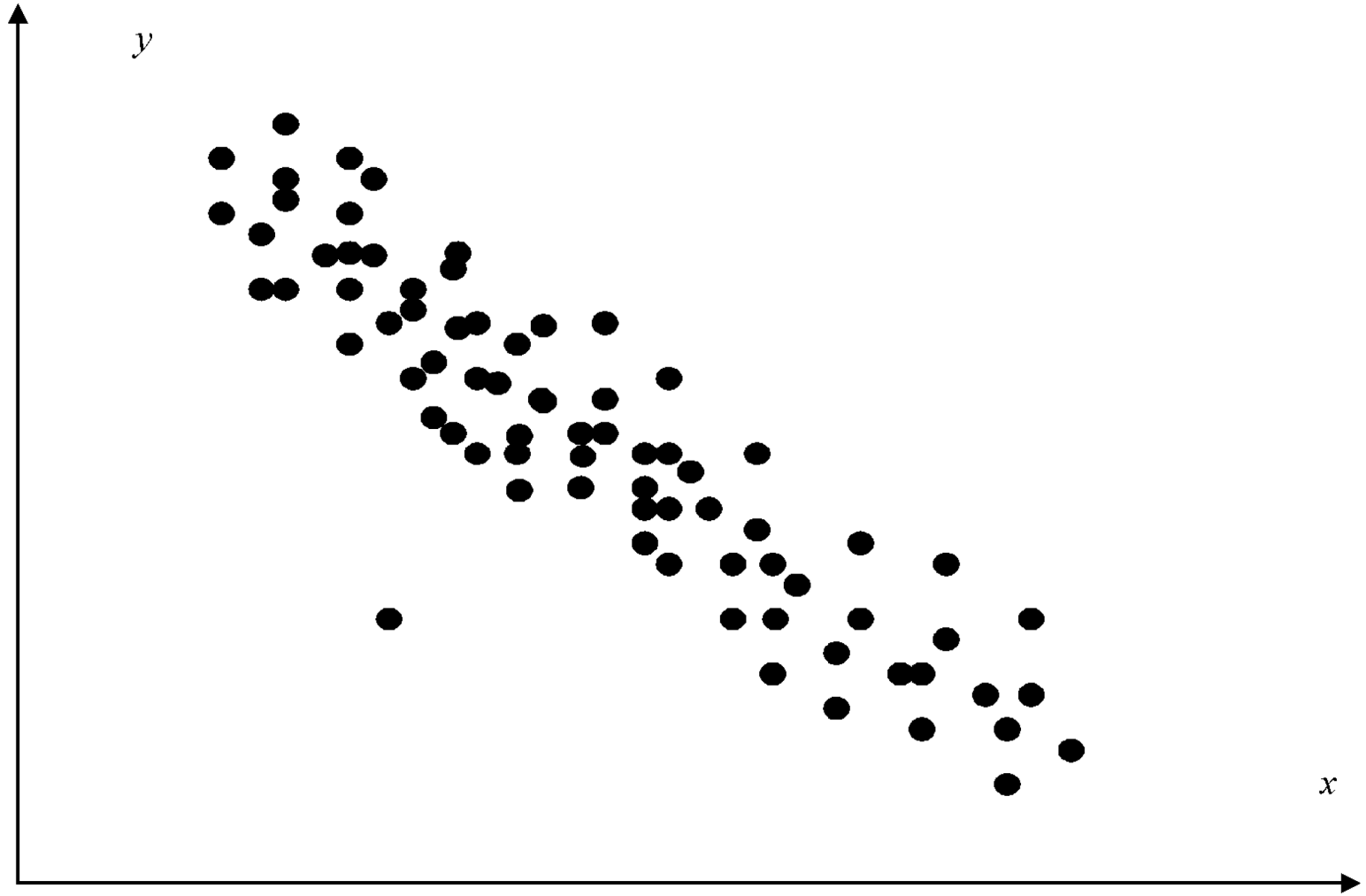
Whether an occurrence is abnormal depends on different aspects like frequency, spatial correlation, etc.

Sample applications of outlier detection

- Fraud detection
  - Purchasing behavior of a credit card owner usually changes when the card is stolen
  - Abnormal buying patterns can characterize credit card abuse
- Medicine
  - Unusual symptoms or test results may indicate potential health problems of a patient
  - Whether a particular test result is abnormal may depend on other characteristics of the patients (e.g. gender, age, …)
- Public health
  - The occurrence of a particular disease, e.g. tetanus, scattered across various hospitals of a city indicate problems with the corresponding vaccination program in that city

- Sports statistics
  - In many sports, various parameters are recorded for players in order to evaluate the players' performances
  - Outstanding (in a positive as well as a negative sense) players may be identified as having abnormal parameter values
  - Sometimes, players show abnormal values only on a subset or a special combination of the recorded parameters
- Detecting measurement errors
  - Data derived from sensors (e.g. in a given scientific experiment) may contain measurement errors
  - Abnormal values could provide an indication of a measurement error
  - Removing such errors can be important in other data mining and data analysis tasks

"One person's noise could be another person's signal."


y
x

Food for thought using a "basic model" for outlier detection

- Data is usually multivariate, i.e., multi-dimensional
  - basic model is univariate, i.e., 1-dimensional (see previous plot!!!)
- There is usually more than one generating mechanism/statistical process underlying the "normal" data
  - basic model assumes only one "normal" generating mechanism
- Anomalies may represent a different class (generating mechanism) of objects, so there may be a large class of similar objects that are the outliers
  - basic model assumes that outliers are rare observations
- A lot of models and approaches have evolved in the past years in order to exceed these assumptions
- It is not easy to keep track with this evolution: often involve typical, sometimes new, though usually hidden assumptions and restrictions, which should be verified for their validity

2.3 Exploratory Data Analysis (EDA)

Introduction

- Three popular data analysis approaches are:
  - **Classical**
  - **Bayesian**
  - **Exploratory** (Exploratory Data Analysis)

- These three approaches are similar in that they all start with a general science/engineering problem and all yield science/engineering conclusions. The difference is the sequence and focus of the intermediate steps.

- For classical analysis, the sequence is
  Problem => Data => Model => Analysis => Conclusions

- For Bayesian, the sequence is
  Problem => Data => Model => Prior Distribution
  => Analysis => Conclusions
- For EDA, the sequence is
  Problem => Data => Analysis => Model => Conclusions

Introduction

- For **classical analysis**, the data collection is followed by proposing a model (normality, linearity, etc.) and the analysis, estimation, and testing that follows are focused on the parameters of that model.
- For a **Bayesian analysis**, the analyst attempts to incorporate scientific/engineering knowledge/expertise into the analysis by imposing a *data-independent* distribution on the parameters of the selected model: the analysis formally combines both the prior distribution on the parameters and the collected data to jointly make inferences and/or test assumptions about the model parameters.
- For **EDA**, the data collection is not followed by a model imposition: it is followed immediately by analysis with a goal of inferring what model would be appropriate.

So what does EDA involve and what does it not involve?

- Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

  - maximize insight into a data set;
  - uncover underlying structure;
  - extract important variables;
  - detect outliers and anomalies;
  - test underlying assumptions;
  - develop parsimonious models

- Some common questions that exploratory data analysis is used to answer are:
  - What is a typical value?
  - What is the uncertainty for a typical value?
  - What is a good distributional fit for a set of numbers?
  - What is a percentile?
  - Does an engineer modification have an effect?
  - Does a factor have an effect?
  - What are the most important factors?
  - Are measurements coming from different laboratories equivalent?
  - What is the best function for relating a response variable to a set of factor variables?
  - What are the best settings for factors?
  - Can we separate signal from noise in time dependent data?
  - Can we extract any structure from multivariate data?
  - Does the data have outliers?

So what does EDA involve and what does it not involve?

- The EDA approach is precisely that - an approach/philosophy - . It is *not* just a set of techniques or toolbox
- So also: EDA is *not* identical to statistical graphics (although the two terms are used almost interchangeably) … It is much more.
  - Statistical graphics is a collection of graphically-based techniques. They are all focusing on data characterization aspects.
  - EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model.

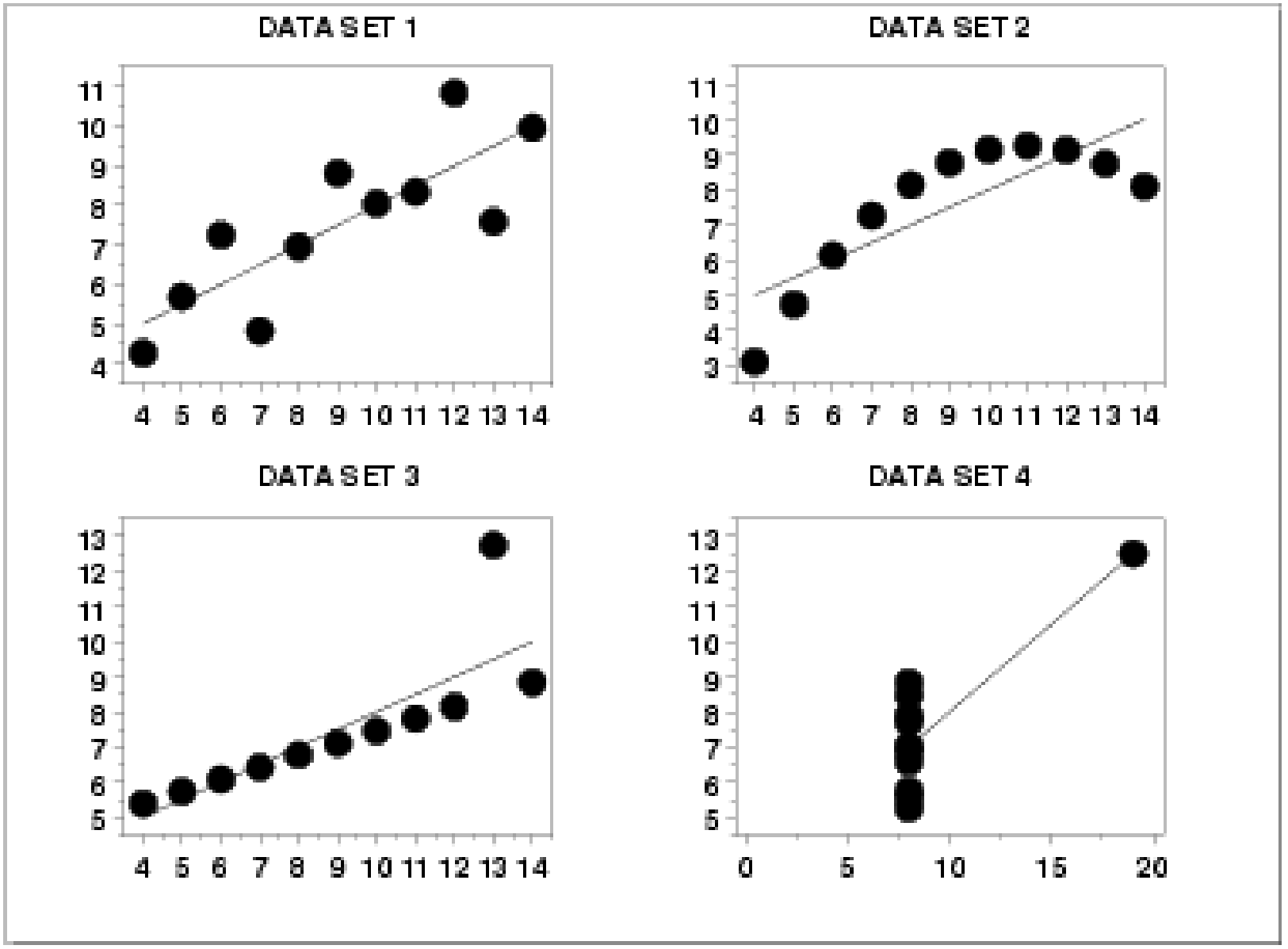
The main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so

Motivating example

- Given 4 data sets (actual data omitted), for which

  N = 11
  Mean of X = 9.0
  Mean of Y = 7.5
  Intercept = 3
  Slope = 0.5
  Residual standard deviation = 1.236
  Correlation = 0.816 (0.817 for data set 4)

- This implies that in some quantitative sense, all four of the data sets are "equivalent".
- In fact, the four data sets are far from "equivalent"!
- A “scatter plot” of each data set (i.e., plotting Y values versus corresponding X values in a plane), would be the first step of any EDA approach … and would immediately reveal non-equivalence!


DATA SET 1
DATA SET 2
DATA SET 3
DATA SET 4

The role of graphics in EDA is ONE important component, indeed

- Statistics and data analysis procedures can broadly be split into two parts:
  - Quantitative procedures
  - Graphical procedures

- **Quantitative techniques** are the set of statistical procedures that yield numeric or tabular output:
  - hypothesis testing (see next chapters)
  - analysis of variance (is there more variation within groups of observations than between groups of observations?)
  - point estimates and confidence intervals (see next chapters)
  - least squares regression

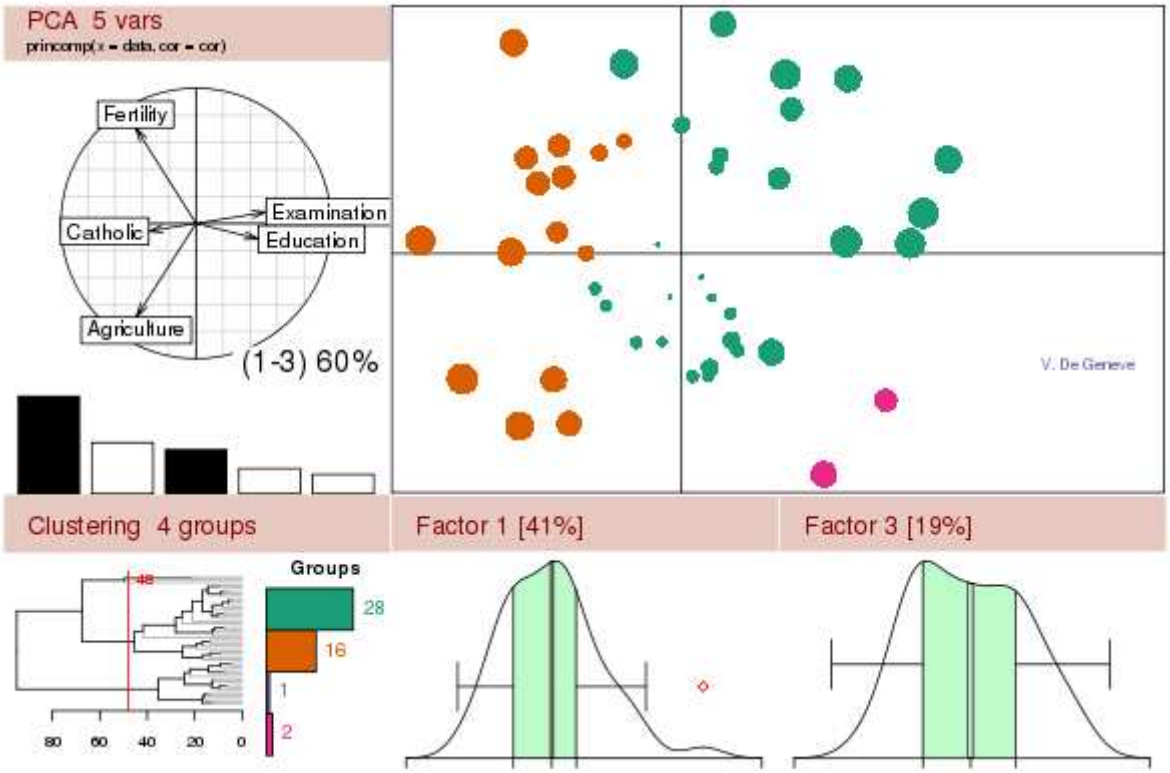
These and similar techniques are all valuable and are mainstream in terms of classical analysis.

- The **graphical techniques** are for a large part employed in an EDA framework. They are often quite simple:

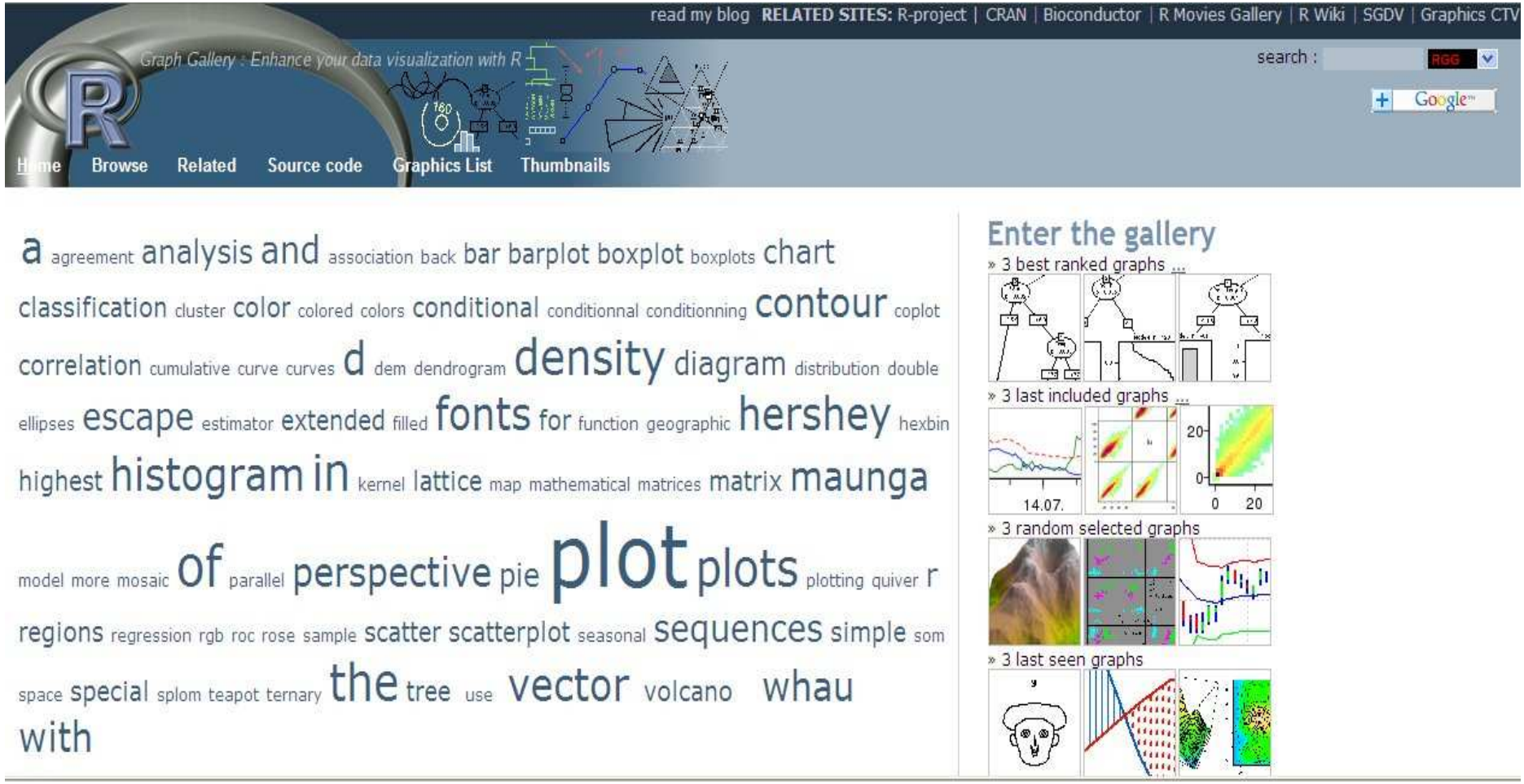
  - plotting the raw data such as via histograms, probability plots
  - plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
  - positioning such plots so as to maximize our natural pattern-recognition abilities (multiple plots, when grouped together, may give a more complete picture of what is going on in the data – see later)

Easy (see before) and more complicated graphical approaches for EDA

The R Project for Statistical Computing



(http://www.r-project.org/)



(http://addictedtor.free.fr/graphiques/)

Assumptions of EDA

- Virtually any data analysis approach relies on assumptions that need to be verified
- There are four assumptions that typically underlie all measurement processes; namely, that the data from the process at hand "behave like":

    - random drawings,
    - from a fixed distribution,
    - with the distribution having fixed location and
    - with the distribution having fixed variation

Assumptions of EDA

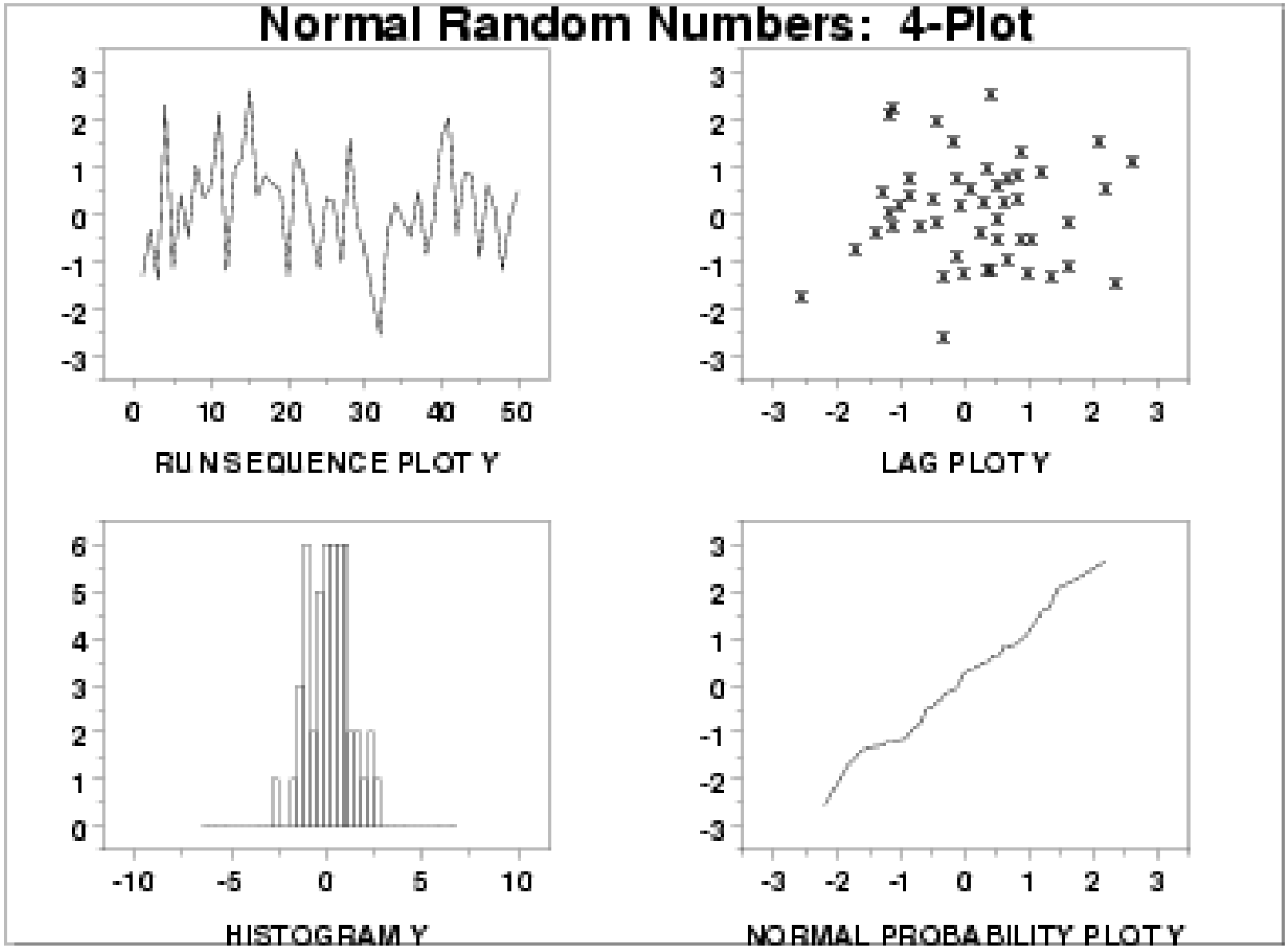
- The most common assumption is that the differences between the raw response data and the predicted values from a fitted model (these are called residuals) should themselves behave like a **univariate process**
- Under a univariate process we understand that the data following such a process behaves like:
  - random drawings,
  - from a fixed distribution,
  - with fixed location (namely, 0 in the case of “residuals” above) and
  - with fixed variation.

Assumptions of EDA

- If the residuals from the fitted model do in fact behave like the ideal, then testing of these underlying assumptions for univariate processes becomes a tool for the validation and quality of fit of the chosen model.
- On the other hand, if the residuals from the chosen fitted model violate one or more of the aforementioned univariate assumptions, then we can say that the chosen fitted model is inadequate and an opportunity exists for arriving at an improved model.

Testing underlying assumptions using an EDA approach

- The following EDA techniques are simple, efficient, and powerful for the routine testing of underlying assumptions:
  - **run sequence plot** (Y_i versus i) -- upper left on next slide
  - **lag plot** (Y_i versus Y_{i-1}) -- upper right on next slide
  - **histogram** (counts versus subgroups of Y) -- lower left
  - **normal probability plot** (ordered Y versus theoretical ordered Y) – lower right on next slide
- Together they form what is often called a **4-plot** of the data**.**


Normal Random Numbers: 4-Plot
RUNSEQUENCE PLOT Y
LAG PLOT Y
HISTOGRAM Y
NORMAL PROBABILITY PLOT Y

Interpretation of 4-plots

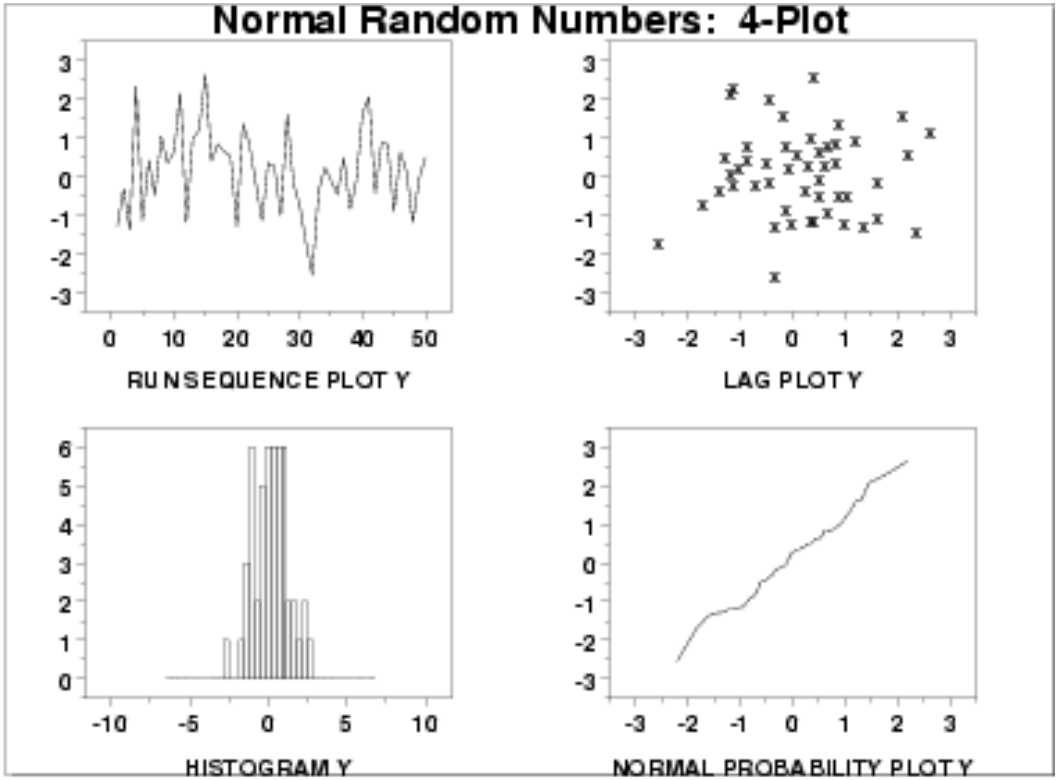
- **Randomness:**
  If the randomness assumption holds, then the lag plot (Y_i versus Y_{i-1}) will be without any apparent structure and random.
- **Fixed Distribution:**
  If the fixed distribution assumption holds, in particular if the fixed normal distribution holds, then the histogram will be bell-shaped, and the normal probability plot will be linear.
- **Fixed Location:**
  If the fixed location assumption holds, then the run sequence plot (Y_i versus i) will be flat and non-drifting.
- **Fixed Variation:**
  If the fixed variation assumption holds, then the vertical spread in the run sequence plot (Y_i versus i) will be the approximately the same over the entire horizontal axis.

Interpretation of 4-plots

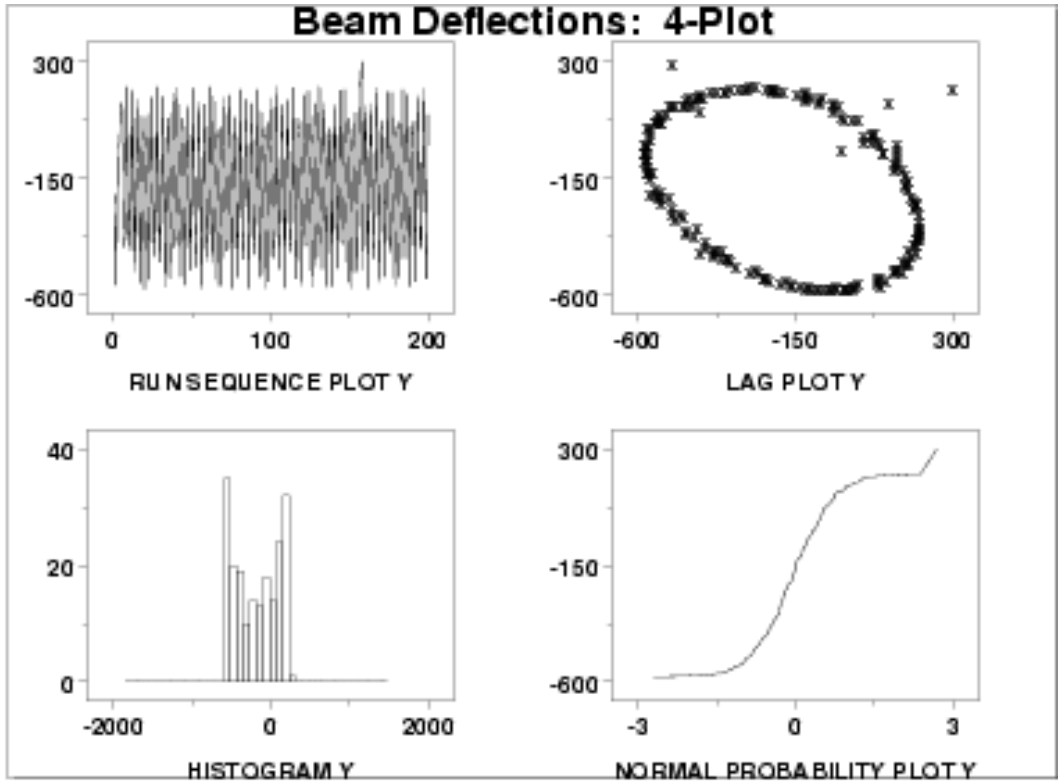
- **Run Sequence Plot:**
  If the run sequence plot (Y_i versus i) is flat and non-drifting, the fixed-location assumption holds. If the run sequence plot has a vertical spread that is about the same over the entire plot, then the fixed-variation assumption holds.
- **Lag Plot:**
  If the lag plot is without structure, then the randomness assumption holds.
- **Histogram:**
  If the histogram is bell-shaped, the underlying distribution is symmetric and *perhaps* approximately normal.
- **Normal Probability Plot:**
  If the normal probability plot is linear, the underlying distribution is approximately normal.

If all 4 assumptions hold, then the process is said to be "in statistical control".

Two examples of 4-plots



- The following example of a 4-plot reveals a process that has fixed location, fixed variation, is random, apparently has a fixed approximately normal distribution, and has no outliers.



- This 4-plot reveals a process that has fixed location, fixed variation, is non-random (oscillatory), has a non-normal, U-shaped distribution
- There seem to be several outliers

(interpretation = qcm material!!!)

Consequences of non-randomness

- If the randomness assumption does not hold, then
  - All of the usual statistical tests are invalid.
  - The calculated uncertainties for commonly used statistics become meaningless.
  - The calculated minimal sample size required for a pre-specified tolerance becomes meaningless.
  - The simple model (linear regression line): y = constant + error becomes invalid.
  - The parameter estimates become suspect and non-supportable
  - …

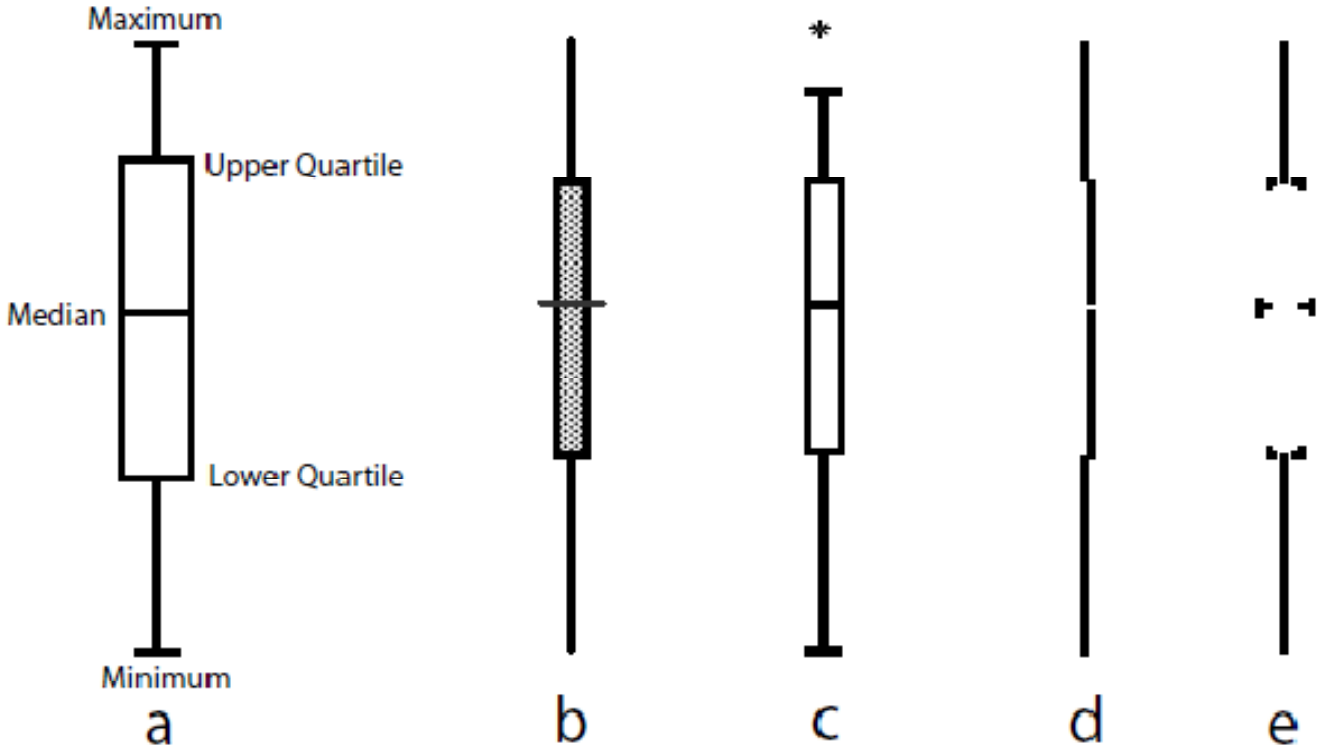
When violations cannot be “corrected” in some sense, usually a more complicated analysis strategy needs to be adopted.

2.4 Box plots

Introduction

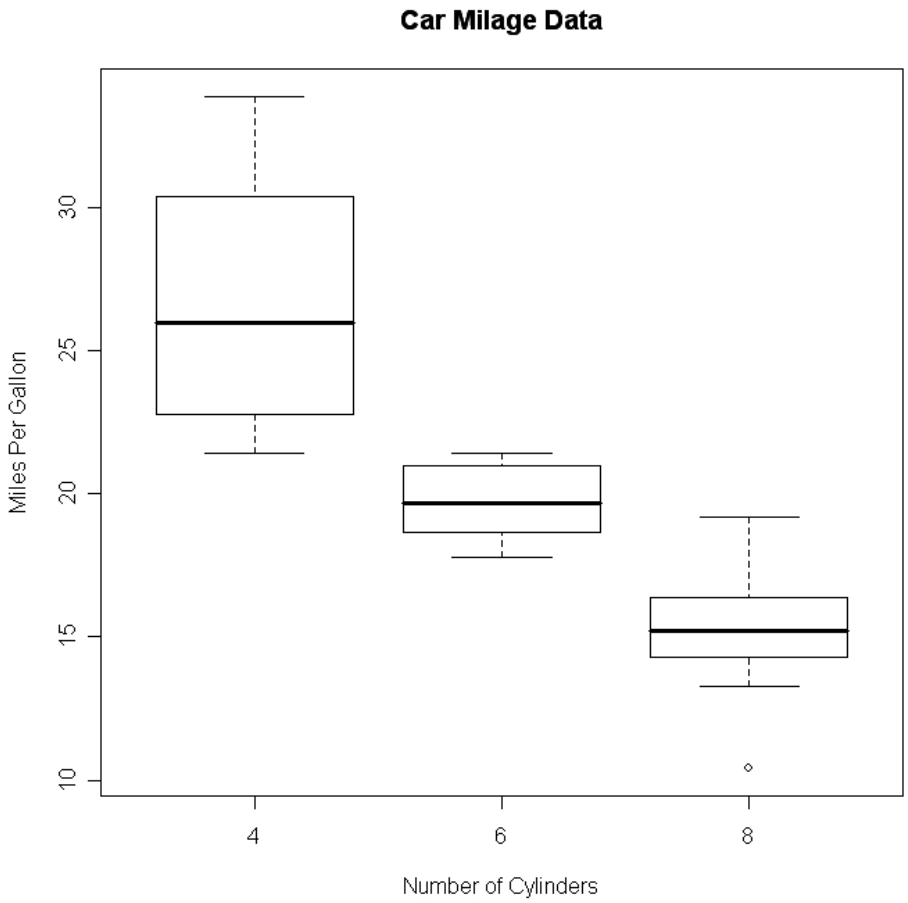
- The box-plot, also known as the box and whisker plot, is a graphical method of displaying 5 descriptive statistics:
  - the median,
  - the upper and lower quartiles (the lower quartile is the 25th percentile and the upper quartile is the 75th percentile),
  - and the minimum and maximum data values.
- First created by John Tukey in a 1977 publication, box plots have evolved into a familiar and useful standard in data interpretation.

- The box plot identifies the middle 50% of the data, the median, and the extreme points.



a) The anatomy of a box plot. b-e) Variations of the Box Plot.

Dissection of a classical box plot



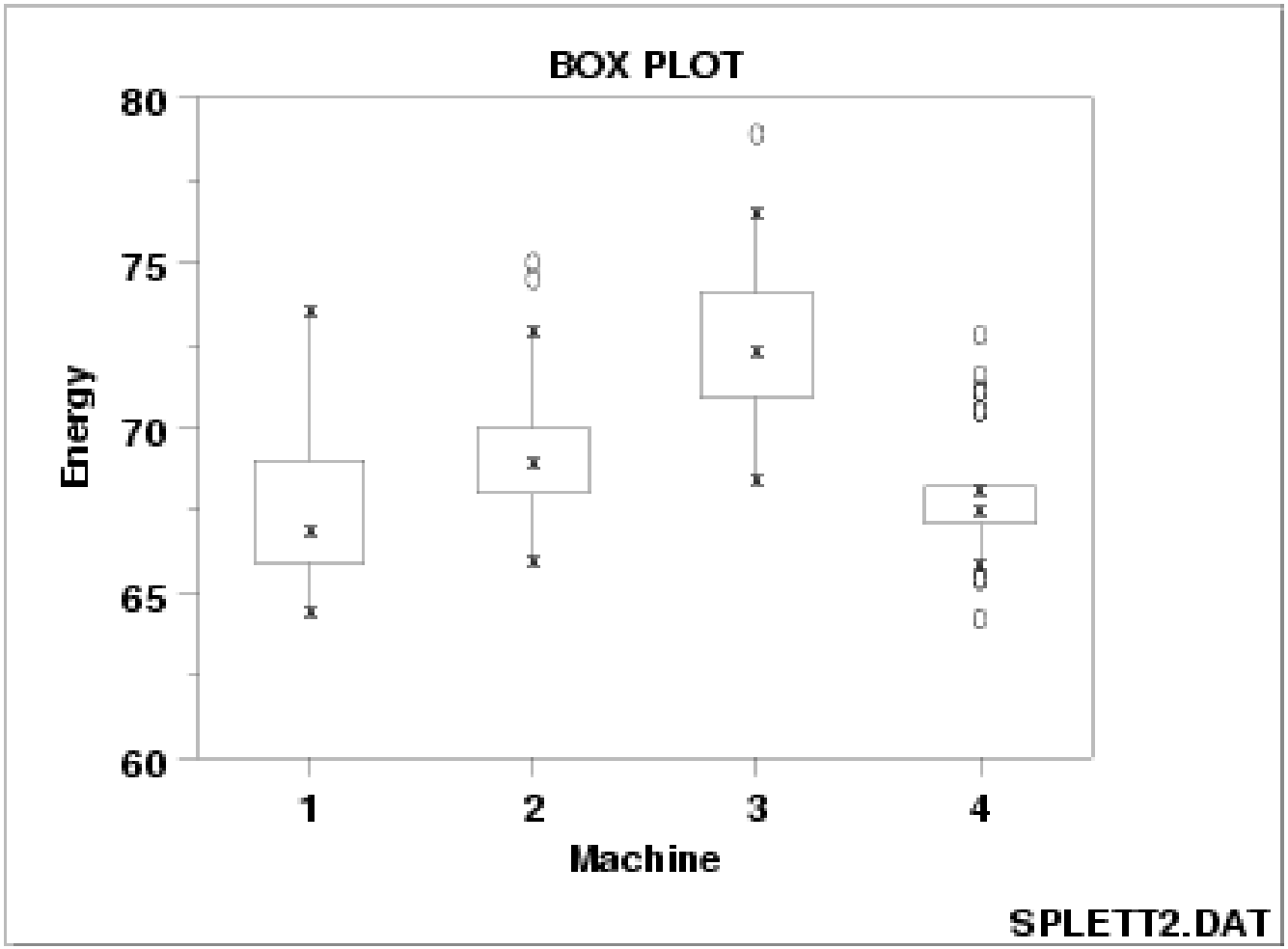
- A symbol is plotted at the median (or a line is drawn) and a box is drawn (hence the name--box plot) between the lower and upper quartiles; this box represents the middle 50% of the data--the "body" of the data.
- A line is drawn from the lower quartile to the minimum point and another line from the upper quartile to the maximum point

Box plots, identifying outliers

- There is a useful variation of the box plot that more specifically identifies outliers.
- To create this variation:
  - Calculate the interquartile range (the difference between the upper and lower quartile) and call it IQ.
  - Calculate the following points:

    L1 = lower quartile - 1.5*IQ
    L2 = lower quartile - 3.0*IQ
    U1 = upper quartile + 1.5*IQ
    U2 = upper quartile + 3.0*IQ

  - The line from the lower (upper) quartile to the minimum (maximum) is now drawn from the lower (upper) quartile to the smallest (largest) point that is greater (smaller) than L1 (U1).
  - Points between L1 and L2 or between U1 and U2 are drawn as small circles. Points less than L2 or greater than U2 are drawn as large circles.



(interpretation = qcm material!!!)

Uses of box plots, as they are most commonly used

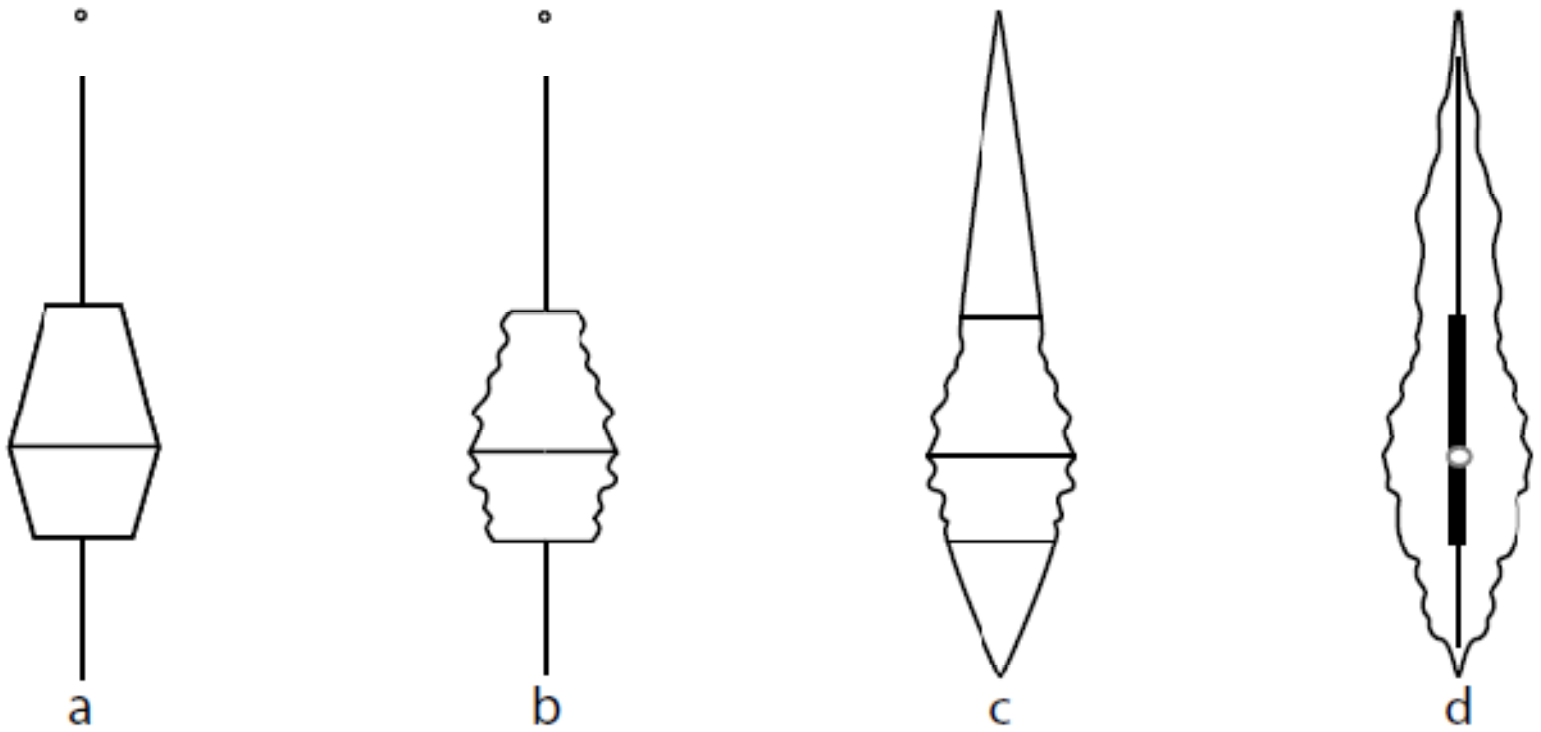
- The strong point of the box plot is its ability to compare two populations without knowing anything about the underlying statistical distributions of those populations.
  - Note that the distribution that defines a population also determines the type of statistical analyses that can be properly applied, so the box plot actually allows you to compare "apples and oranges" graphically that might not be directly comparable statistically.

- The box plot can provide answers to the following questions:
  - Is a factor significant?
  - Does the location differ between subgroups?
  - Does the variation differ between subgroups?
  - Are there any outliers?

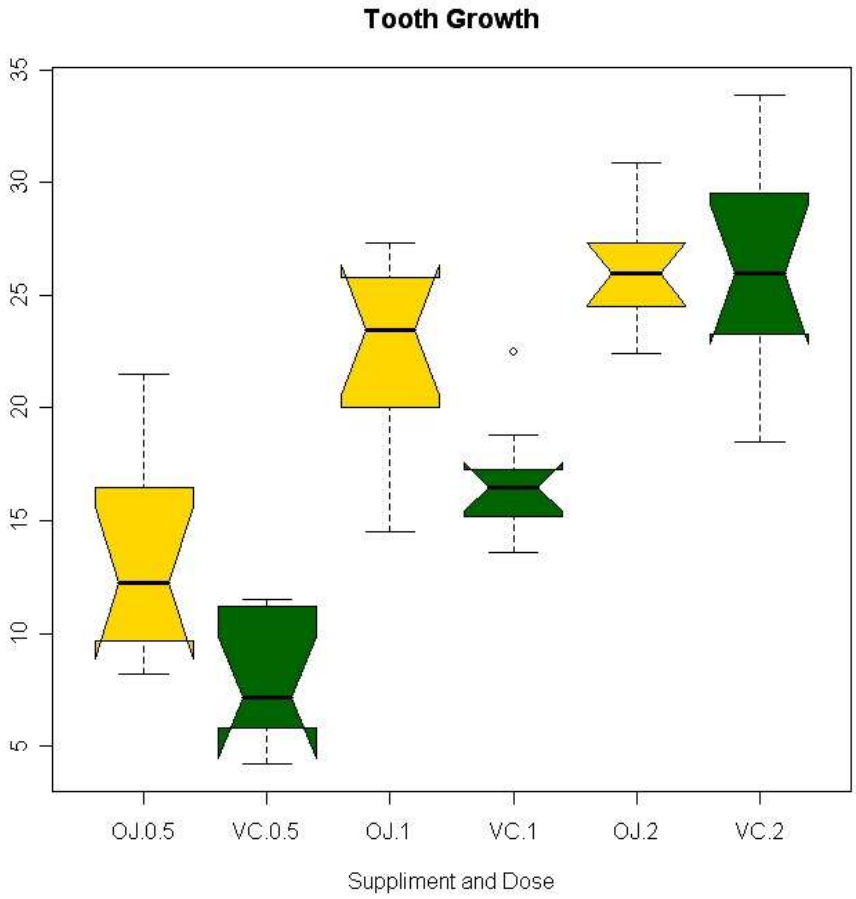
Extensions to the classical box plots – 1 (no exam material)

- One of the most common types of information added to the box plot is a description of the distribution of the data values. The box plot summarizes the distribution using only 5 values, but this overview may hide important characteristics.
- For instance, the modality (or number of most often occurring data values) of a distribution is hidden by the box plot, and distinctive distributions with varying modality may be encoded using similar looking box plots.
- One solution to these types of problems is to add into the box plot indications of the density of underlying distribution.
- Or sometimes, you simply like to have a better grip on the number of contributing observations..

For all of these extra pieces of information, an extension of the classical box plot exists.



Examples of methods for adding density to the box plot. The a) histplot, b) vaseplot, c) box-percentile plot, and d) violin plot (a combination of a boxplot and a kernel density plot).



In the notched boxplot, if two boxes' notches do not overlap this is ‘strong evidence’ their medians differ

2.5QQ plots

Percentiles and quantiles revisited

The k-th *percentile* of a set of values divides them so that $k\,\%$ of the values lie below and $(100 - k)\%$ of the values lie above.

- The 25th percentile is known as the *lower quartile*.
- The 50th percentile is known as the *median*.
- The 75th percentile is known as the *upper quartile*.

It is more common in statistics to refer to *quantiles*. These are the same as percentiles, but are indexed by sample fractions rather than by sample percentages.

The previous definition of quantiles and percentiles is not completely satisfactory. For example, consider the six values:

$$3.7 \quad 2.7 \quad 3.3 \quad 1.3 \quad 2.2 \quad 3.1$$

What is the lower quartile of these values?

There is no value which has 25% of these numbers below it and 75% above.

To overcome this difficulty we will use a definition of percentile which is in the spirit of the above statements, but which (necessarily) makes them hold only approximately.

We define the quantiles for the set of values:

3.7 2.7 3.3 1.3 2.2 3.1

as follows.

First sort the values into order:

1.3 2.2 2.7 3.1 3.3 3.7

Associate the ordered values with sample fractions equally spaced from zero to one.

Sample fraction	0	.2	.4	.6	.8	1
Quantile	1.3	2.2	2.7	3.1	3.3	3.7

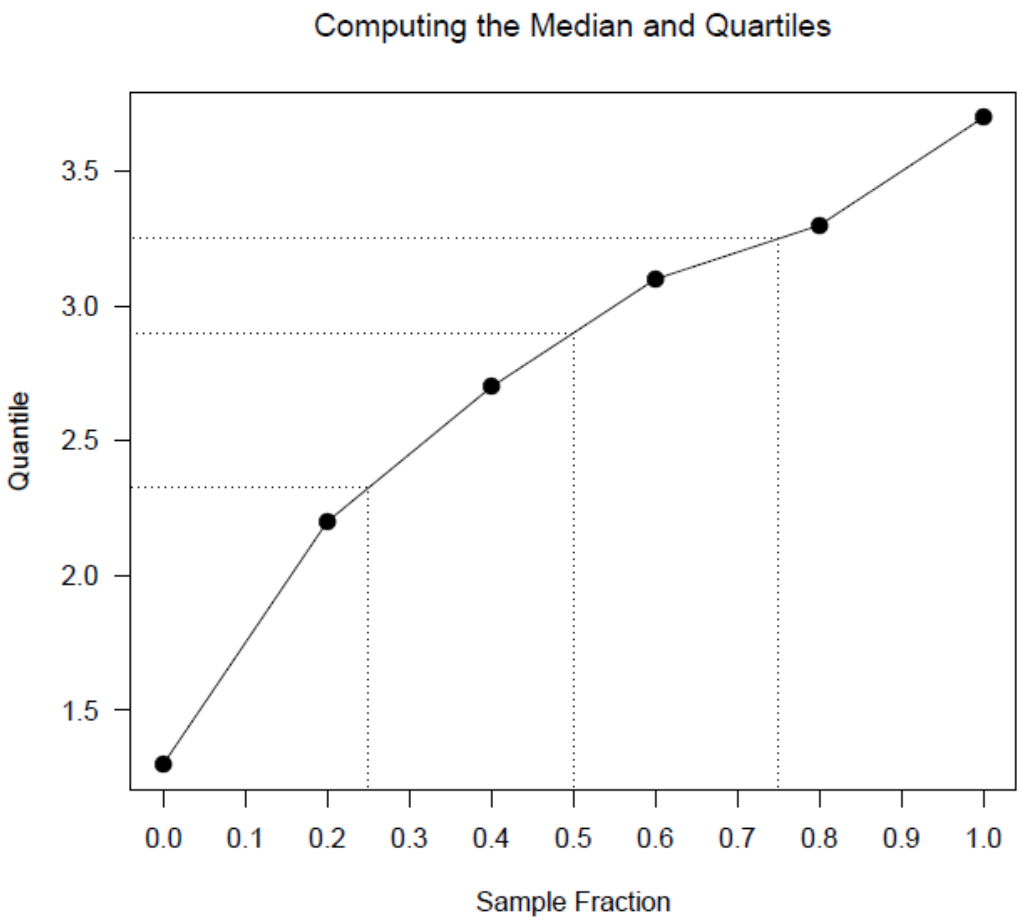
The other quantiles of

$$1.3 \quad 2.2 \quad 2.7 \quad 3.1 \quad 3.3 \quad 3.7$$

can be obtained by linear interpolation between the values of the table.

The median corresponds to a sample fraction of .5. This lies half way between 0.4 and 0.6. The median must thus be $.5 \times 2.7 + .5 \times 3.1 = 2.9$

The lower quartile corresponds to a sample fraction of .25. This lies one quarter of the way between .2 and .4. The lower quartile must then be $.75 \times 2.2 + .25 \times 2.7 = 2.325$.



0.05 = 25% (0.4-0.2)

25%(2.7-2.2) + 2.2 = targeted value

The general case: deriving quantiles from your data

(compare with theoretical quantile function)

Given a set of values $x_1, x_2, \ldots .x_n$ we can define the quantiles for any fraction p as follows.

Sort the values in order

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

The values $x_{(1)}, \ldots, x_{(n)}$ are called the *order statistics* of the original sample.

Take the order statistics to be the quantiles which correspond to the fractions:

$$p_i = \frac{i-1}{n-1}, \quad (i = 1, \ldots, n),$$

The quantile function

In general, to define the quantile which corresponds to the fraction p, use linear interpolation between the two nearest p_i.

If p lies a fraction f of the way from p_i to p_{i+1} define the pth quantile to be:

$$Q(p) = (1 - f)Q(p_i) + fQ(p_{i+1})$$

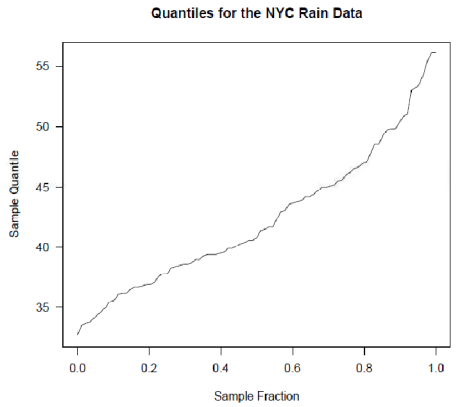
As special cases, define the median and quartiles by:

Median:	$Q(.5)$
Lower Quartile:	$Q(.25)$
Upper Quartile:	$Q(.75)$

The function Q defined in this way is called the *Quantile Function*.

Important note

- We defined the quantile function before: the quantile function of a probability distribution is the inverse of its cumulative distribution function (cdf) *F*
- Now we have seen how to derive the quantile function from the data.
- So, if we consider this to be an estimate (derived from our data) for the truth (at population level), the quantile function estimated from the data will learn us something about the underlying true distribution of the data
- **Quantile plots**:
  - The sample quantiles are plotted against the fraction of the sample they correspond to



Important note

- If we assume a particular “model” or mechanism that could have generated the data, we can compare the quantile function corresponding to this “theoretically proposed distribution” to the quantile function corresponding to our observed data
- Such a plot, comparing observed versus theoretical quantiles, goes further than a simple boxplot (also using quantiles), since it gives a clue about the validity of a proposed model for the data or data generation mechanism.

Quantile-quantile (Q-Q) plots

- In general, QQ plots allow us to compare the quantiles of two sets of numbers
- This kind of comparison is much more detailed than a simple comparison of means or medians
- There is a cost associated with this extra detail. We need more observations than for simple comparisons.
- Important remark:
  - A P-P plot compares the empirical cumulative distribution function of a data set with a specified theoretical cumulative distribution function.
  - A Q-Q plot compares the quantiles of a data distribution with the quantiles of a standardized theoretical distribution from a specified family of distributions

The many uses of Q-Q plots

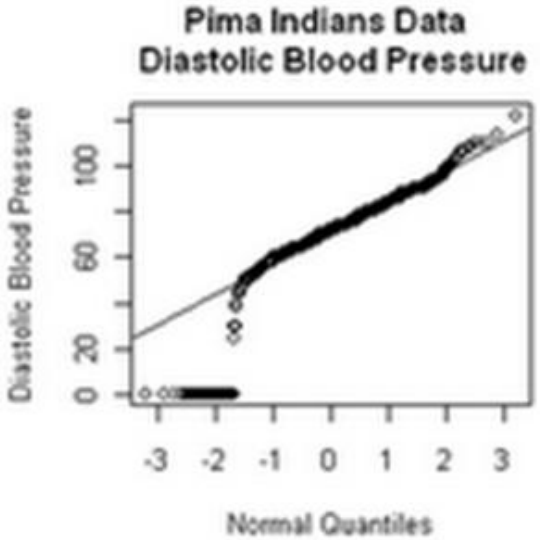
- The most common form of a Q-Q-plot is the normal Q-Q plot, which represents an informal graphical test of the hypothesis that a data sequence is normally distributed.
- That is, if the points on a normal Q-Q plot are reasonably well approximated by a straight line, the popular Gaussian data hypothesis is plausible, while marked deviations from linearity provide evidence against this hypothesis.
- The utility of normal Q-Q plots goes well beyond this informal hypothesis test. In particular, the shape of a normal Q-Q plot can be extremely useful in highlighting distributional asymmetry, heavy tails, outliers, multi-modality, or other data anomalies.

Quantile-quantile plot diagnostics

Description of Point Pattern	Possible Interpretation
all but a few points fall on a line	outliers in the data
left end of pattern is below the line; right end of pattern is above the line	long tails at both ends of the data distribution
left end of pattern is above the line; right end of pattern is below the line	short tails at both ends of the data distribution
curved pattern with slope increasing from left to right	data distribution is skewed to the right
curved pattern with slope decreasing from left to right	data distribution is skewed to the left
staircase pattern (plateaus and gaps)	data have been rounded or are discrete

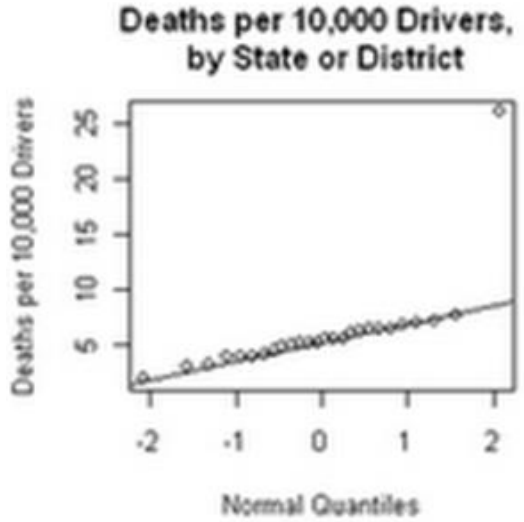
Quantile-quantile plot diagnostics

Description of Point Pattern	Possible Interpretation
all but a few points fall on a line	outliers in the data
left end of pattern is below the line; right end of pattern is above the line	long tails at both ends of the data distribution
left end of pattern is above the line; right end of pattern is below the line	short tails at both ends of the data distribution
curved pattern with slope increasing from left to right	data distribution is skewed to the right
curved pattern with slope decreasing from left to right	data distribution is skewed to the left
staircase pattern (plateaus and gaps)	data have been rounded or are discrete



Quantile-quantile plot diagnostics

Description of Point Pattern	Possible Interpretation
all but a few points fall on a line	outliers in the data
left end of pattern is below the line; right end of pattern is above the line	long tails at both ends of the data distribution
left end of pattern is above the line; right end of pattern is below the line	short tails at both ends of the data distribution
curved pattern with slope increasing from left to right	data distribution is skewed to the right
curved pattern with slope decreasing from left to right	data distribution is skewed to the left
staircase pattern (plateaus and gaps)	data have been rounded or are discrete



Quantile-quantile plot diagnostics

Description of Point Pattern	Possible Interpretation
all but a few points fall on a line	outliers in the data
left end of pattern is below the line; right end of pattern is above the line	long tails at both ends of the data distribution
left end of pattern is above the line; right end of pattern is below the line	short tails at both ends of the data distribution
curved pattern with slope increasing from left to right	data distribution is skewed to the right
curved pattern with slope decreasing from left to right	data distribution is skewed to the left
staircase pattern (plateaus and gaps)	data have been rounded or are discrete