Applied Statistics and Data Analysis

2. Exploratory Data Analysis

Paolo Vidoni

Department of Economics and Statistics
University of Udine
via Tomadini 30/a - Udine
paolo.vidoni@uniud.it

Based mainly on Chapter 2 of the course textbook, *Styles of data analysis*

Table of contents

Summary

- Introduction to Exploratory Data Analysis
- Data and variables
- Graphical summary
- Data summary
- Aims and strategies of statistical analysis

Basic Points

To begin investigation of a new set of data the following points have to be considered.

- **Graphs** are one of the most important tools.
- **Numerical summaries** are also important, but they don't go very far without an appropriate graph.
- **Statistical models** can clarify the information content of data and make prediction possible. But models require **assumptions**, which must be checked, often via graphical methods.
- One should not over-analyze the data, as "under torture the data may yield false confessions".
- The way data were collected should always kept into account.

Exploratory Data Analysis

- The use of graphs (and of numerical summaries) to display and get insight into data has a long tradition.
- John Tukey made it an art, giving it the name of *Exploratory Data Analysis (EDA)*.
- A key point is that data should reveal their information content prior to (or as a part of) a formal analysis.
- Modern computing tools have greatly improved on traditional techniques.

EDA has at least **four roles**.

1. It may suggest ideas and understandings not previously considered.
2. It may challenge the theoretical knowledge that guided the initial collection of the data.
3. It allow the data to cast doubt on an intended analysis and to facilitate checks on assumptions.
4. It may reveal additional information and further lines of research.

- "Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there."
(John M. Tukey)

- "Exploratory data analysis isolates patterns and features of the data and reveals these forcefully to the analyst."
(D.C. Hoaglin, F. Mosteller and J.M. Tukey)

Table of contents

Statistical units and variables

Based on Chapter 2 of *Applied Data Mining for Business and Industry* by P. Giudici and S. Figini

- Data analysis requires the data to be organised into an ordered database.
- It is important to identify the **statistical units**, i.e. the elements in the population (or in the sample) that are considered in the analysis, and the **variables**, i.e. the characteristics measured for each unit.
- Two different types of variables may be defined: **categorical** (qualitative) variables and **numerical** (quantitative) variables.
- Categorical variables can be classified into **nominal**, if the categories do not follow any particular order (i.e. gender, religion professed) or **ordinal**, if the different categories are ordered (i.e. computing skills of a person, credit rate of a company).
- Numerical variables are classified into **discrete**, if they have finite or countable potential values (i.e. family size, number of car accidents) or **continuous**, if they can take any value between two real numbers (i.e. height, time spent waiting in a queue).

Data structures

- A simple way to express a database is in terms of a data matrix, where the rows represent the statistical units and the columns represent the variables.
- An analysis focusing on a single variable is called univariate, whereas it is called multivariate whenever two or more variables are jointly considered.
- Data structures, more complex than a matrix, arise in a number of frameworks as, for example, when temporal and/or spatial information are relevant (longitudinal and/or spatial data).
- Further relevant examples are related to text data (a database of text documents, usually related to each other), web data (e.g. log files on the behavior during a web session) and multimedia data (e.g. texts and audio-visual information).
- Another challenging situation concerns data structures obtained from the integration of different databases.

Table of contents

Views of an univariate data set

Categorical data

- Simple characterizations of the data, and in particular of the frequency distribution, in terms of graphs.
- Data related to **categorical variables** are usually described in the form of tables and thus summarized by calculating frequencies.
- For presentation purposes, table of counts or percentages may be described using the following graphical summaries:
  - **barplots**: the height of each bar is proportional to the (relative) frequency of the associated category;
  - **piecharts**: the arc length of each slice (and consequently its area) is proportional to the frequency of the associated category.

Example: caffeine and marital status

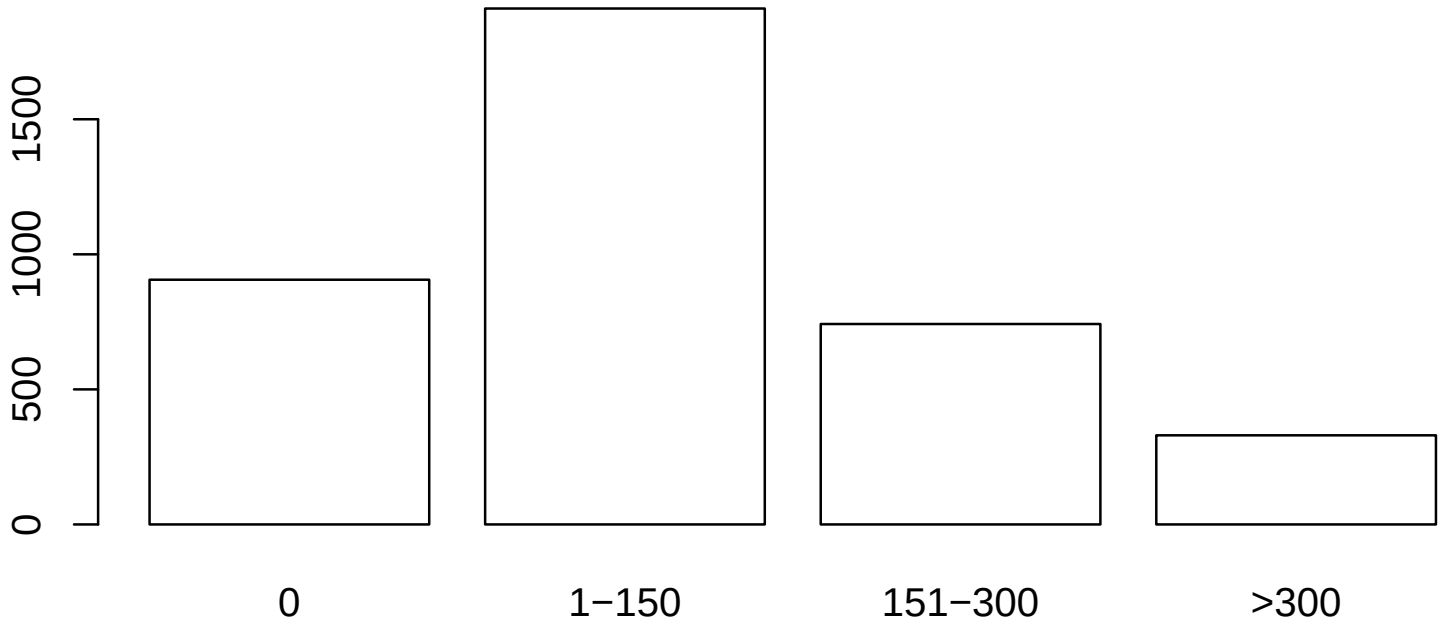
Chapter 4 of *Introductory Statistics with R* by P. Dalgaard

A two-way table containing data on caffeine consumption (mg/day) by marital status (married, previously married, single) among 3888 pregnant women: observed frequencies and row percentages

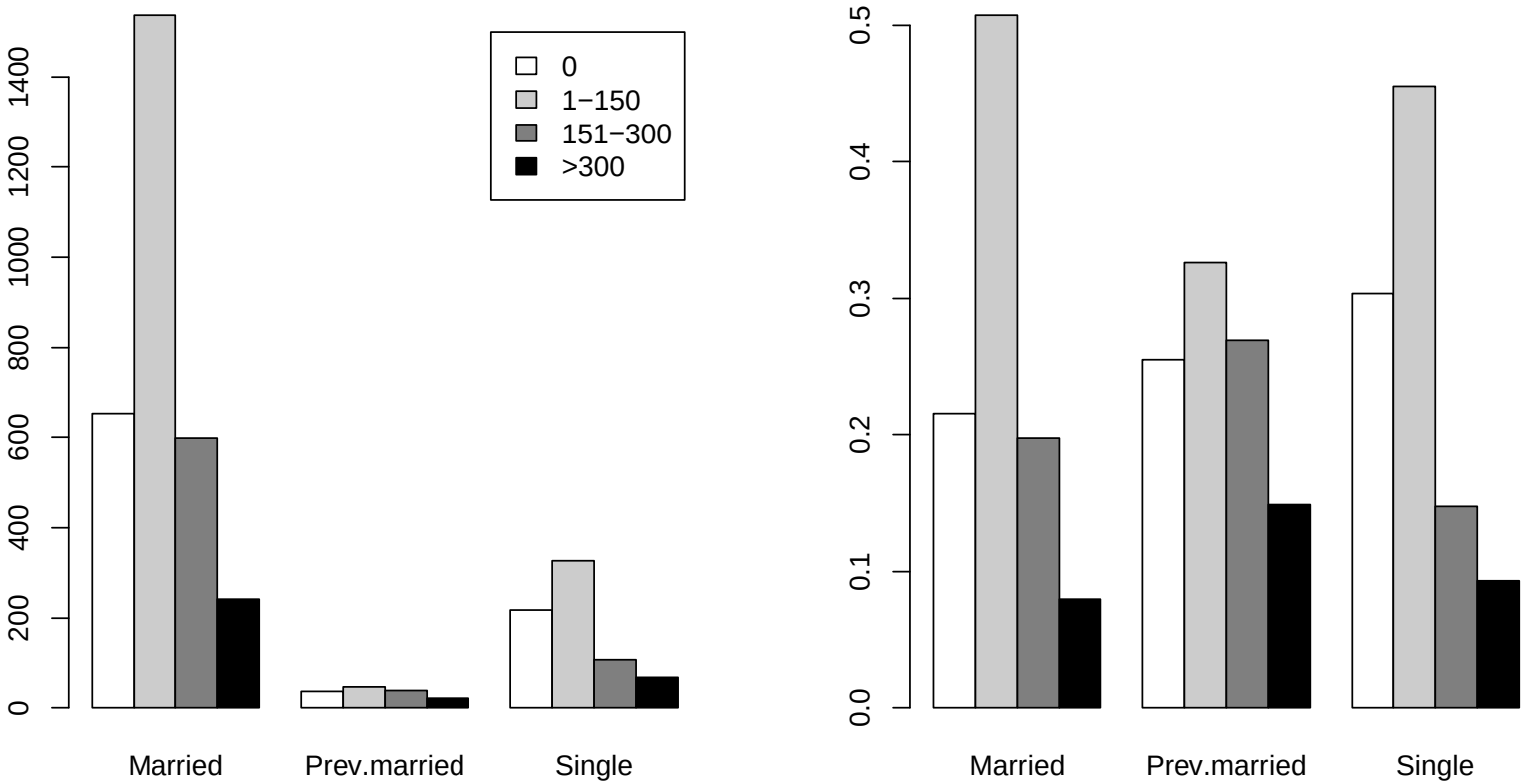
	caffeine consumption (mg/day)				
marital status	0	1-50	151-300	> 300	Total
married	652	1537	598	242	3029
prev. married	36	46	38	21	141
single	218	327	106	67	718
Total	906	1910	742	330	3888

	caffeine consumption (mg/day)				
marital status	0	1-50	151-300	> 300	Total
married	0.22	0.51	0.20	0.08	1
prev. married	0.26	0.33	0.27	0.15	1
single	0.30	0.46	0.15	0.09	1
Total	0.23	0.49	0.19	0.08	1

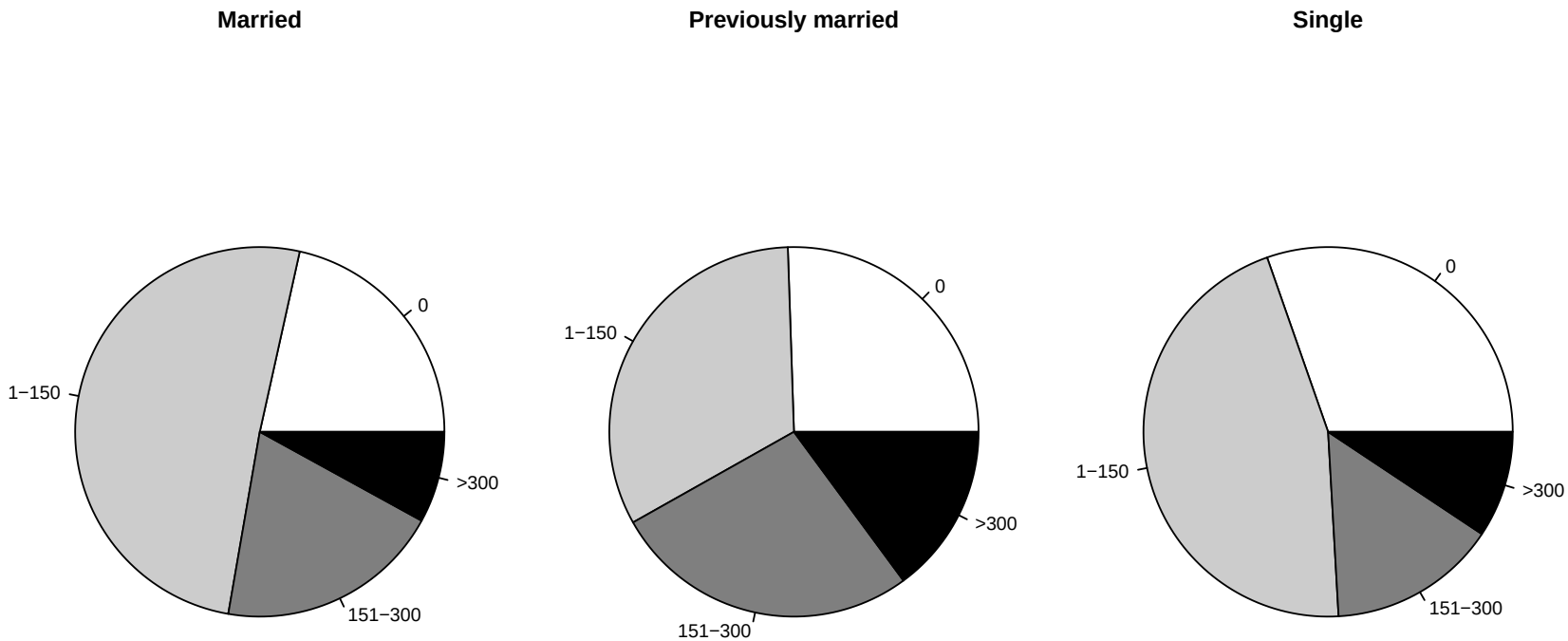
Simple barplot of total `caffeine consumption` (observed frequencies)



Multiple barplots: observed frequencies (left), row percentages (right). Row percentages enable the comparison of caffeine consumption among the three groups.



Piecharts of `caffeine consumption` according to `marital status`



Views of an univariate data set

Numerical data

Simple graphical representations of **numerical variables** (both continuous and discrete), and in particular of the frequency distribution of the associated data set, include:

- **histograms**: the area of each rectangle is proportional to the frequency of the observations that lie within the base of the rectangle;
- **density estimate**: for continuous data, an histogram is rough form of density estimate; a better, smooth alternative is a (kernel) density estimate;
- **boxplot**: a notable graphical summary of the data set emphasizing median, quartiles and potential outliers.

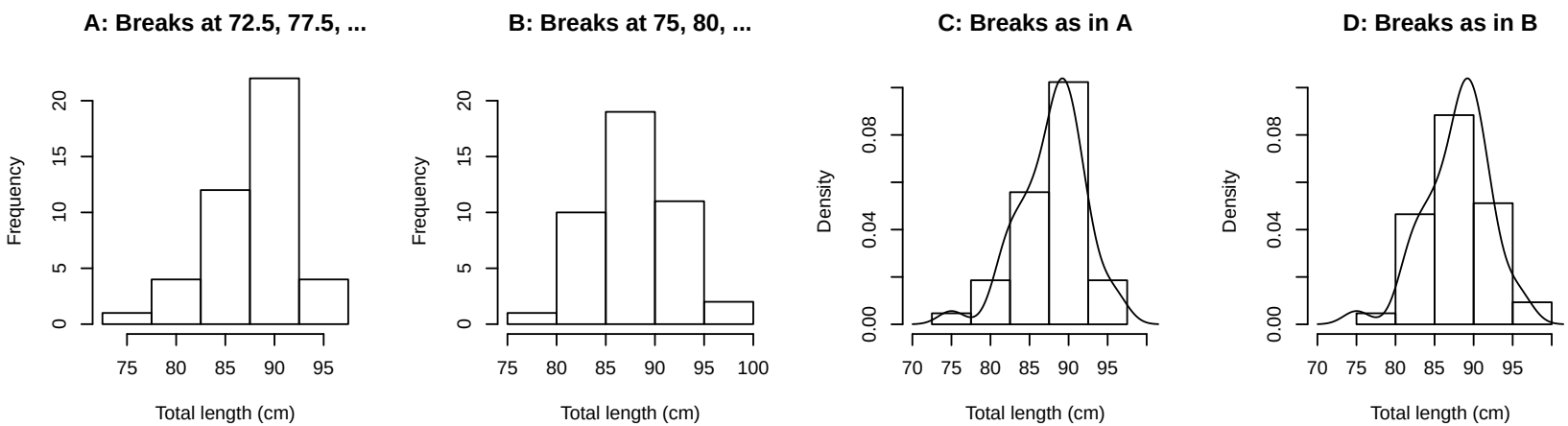
Histograms and density plots

- Histograms are basic EDA tools for displaying the frequency distribution of a data set.
- A symmetric and regular histogram often hints to a normal distribution for the data (the "bell curve").
- In small samples, the shape can be highly irregular, and the appearance may depend on the choice of breakpoints.
- A better alternative is, often, a smooth density estimate.
- Like **width of histogram bars**, that have to be chosen subjectively, density estimates require the choice of a **bandwidth parameter** that tunes the amount of smoothing. Software default choices often work well.
- Density curves are preferable to histograms for drawing attention to particular forms of non-normality.

Example: possum data set

The complete data set has nine morphometric measurements on 104 mountain brushtail possums. Here, attention will be limited to the `length` (cm) measurements for the 43 females.

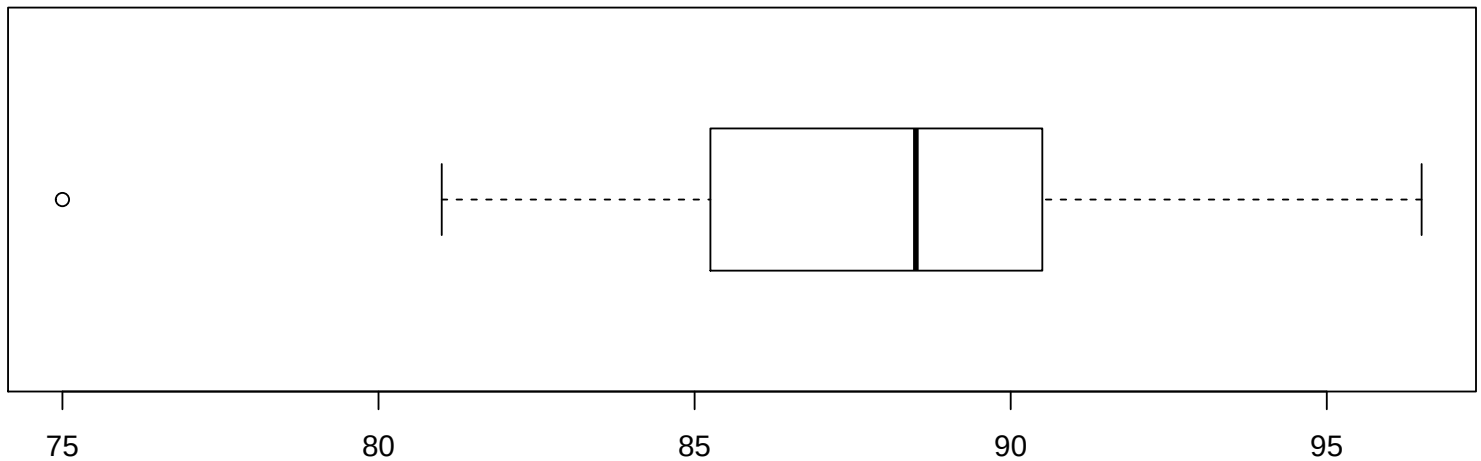
Alternative breakpoints for the histograms suggest different conclusions for the frequency distribution.



Boxplots

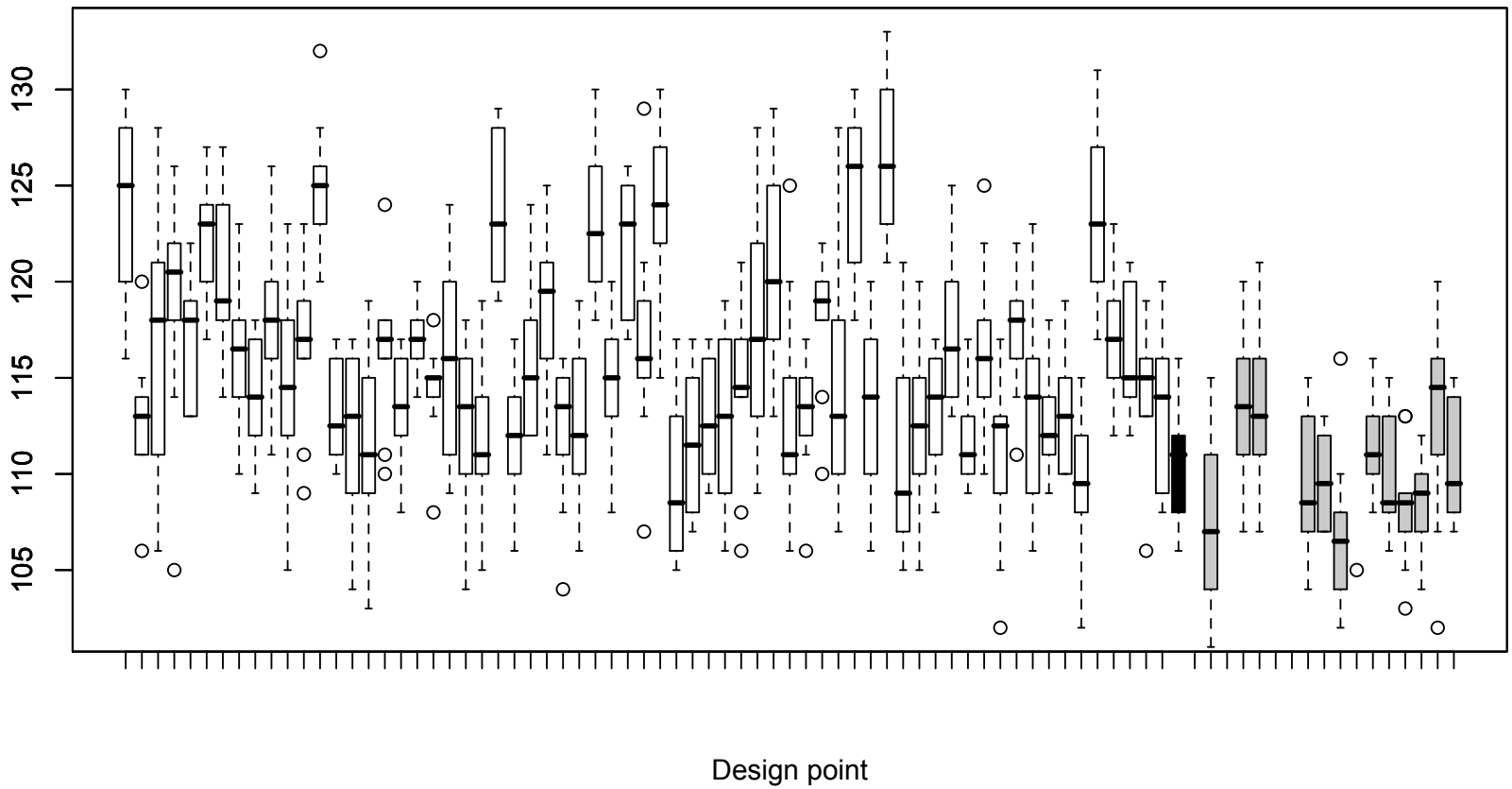
Boxplots allow a comprehension of specific important features of the data at a glance. Indeed, they are useful to identify **outliers**.

With regard to the possum data set and the `length` measurement



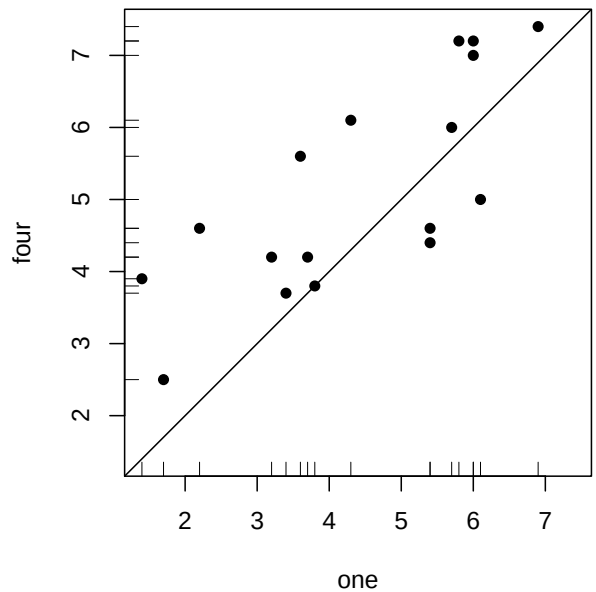
The **whiskers** range from the smallest to the largest value (outliers excepted), while the **box** is defined by the lower quartile and the upper quartile, with indication of the **median**.

Boxplots are very useful to compare different data sets.



Patterns in bivariate numerical data: scatterplot

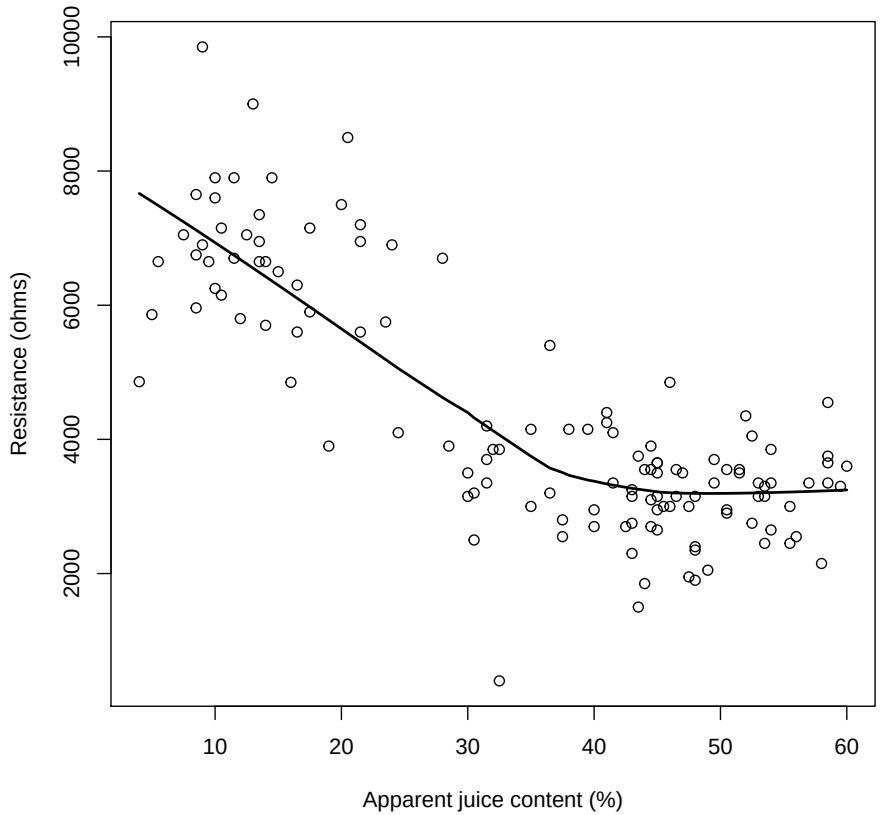
Data from a tasting session where each of 17 panelists assessed the `sweetness` of two milk samples, one with four units of additive, the other with one unit of additive



The line $y = x$ assists in comparing the two samples: most panelists rated the sample with four units as sweeter than the sample with one unit.

Adding a smooth trend curve

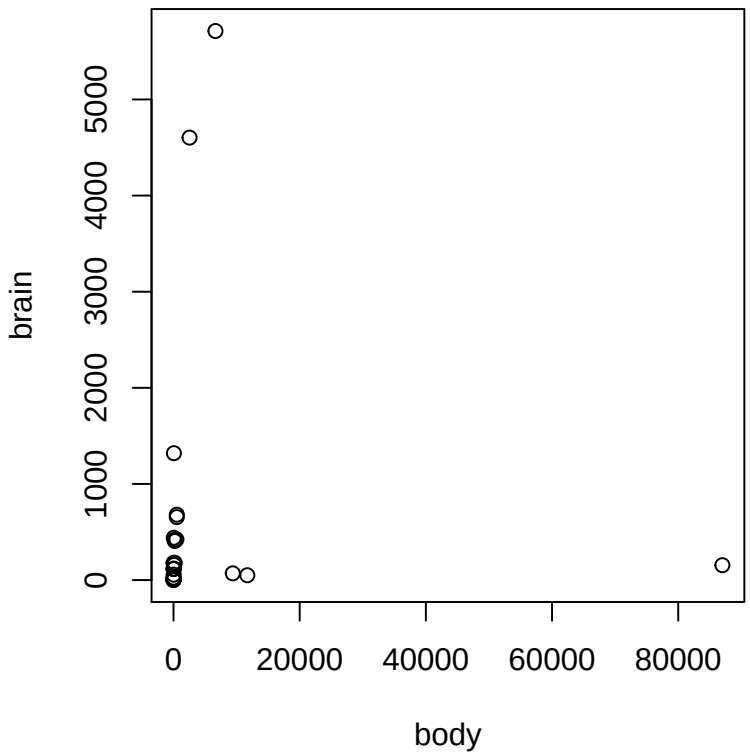
Data from a study that measured both `electrical resistance` (ohms) and `apparent juice content` (%) for a number of slabs of kiwifruit

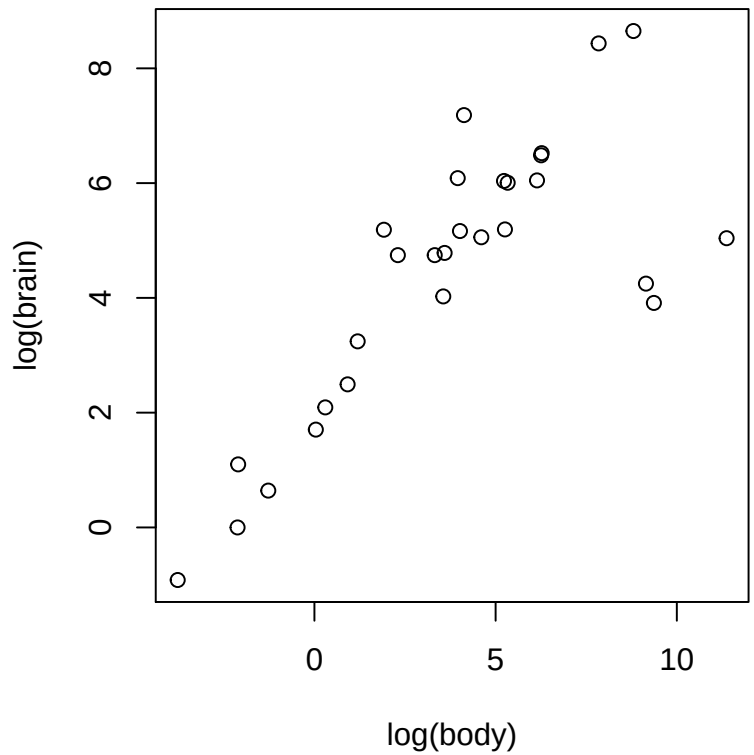


The curve estimates the relationship between electrical resistance and apparent juice content, which is nonlinear.

What is the appropriate scale?

Data of `brain weight` (g) against `body weight` (kg), for a number of different animals: untransformed scale vs logarithmic scale

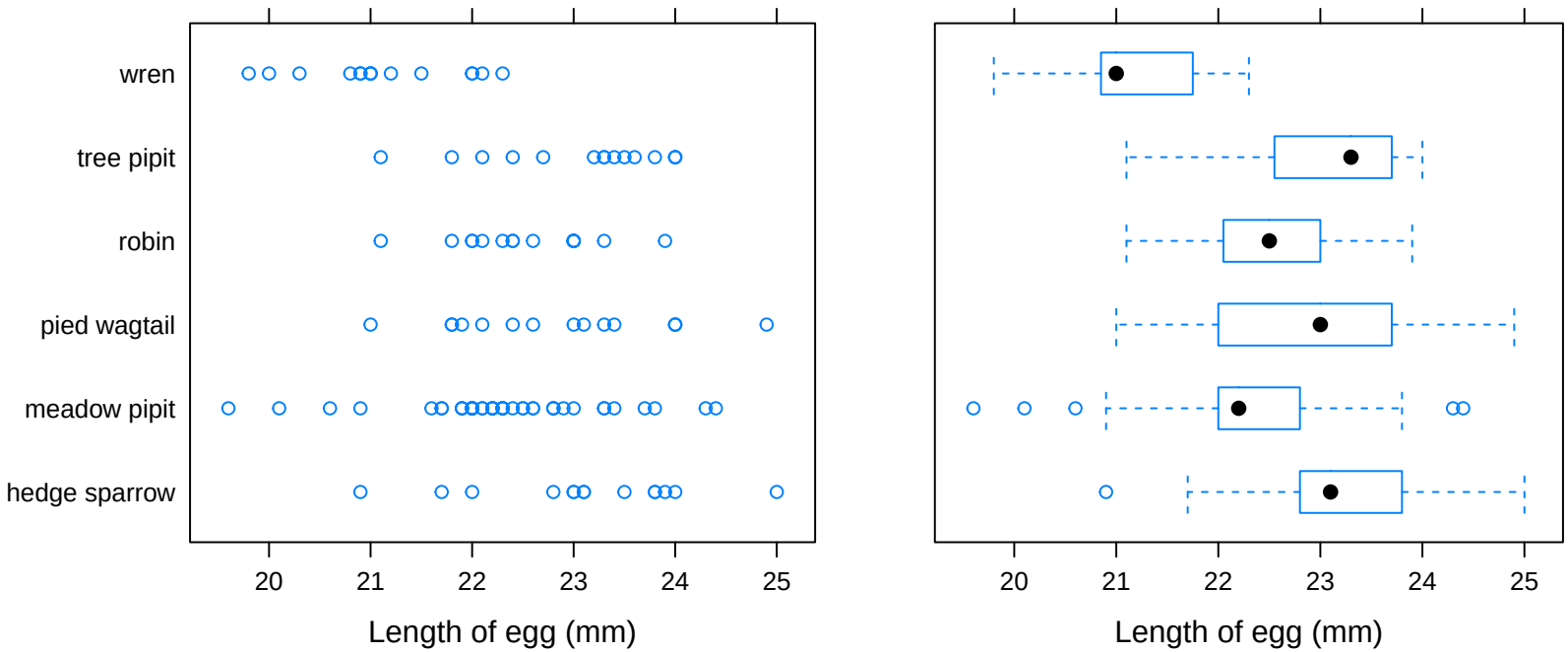




Logarithmic scale is appropriate for quantities that change multiplicatively, like cells in growing organisms.

Patterns in grouped data

Cuckoos lay eggs in the nest of other birds, which adopt and hatch the eggs; the egg `lengths` (mm) are grouped by the `species` of the host bird and strip plots (left) and boxplots (right) are useful for side-by-side comparisons

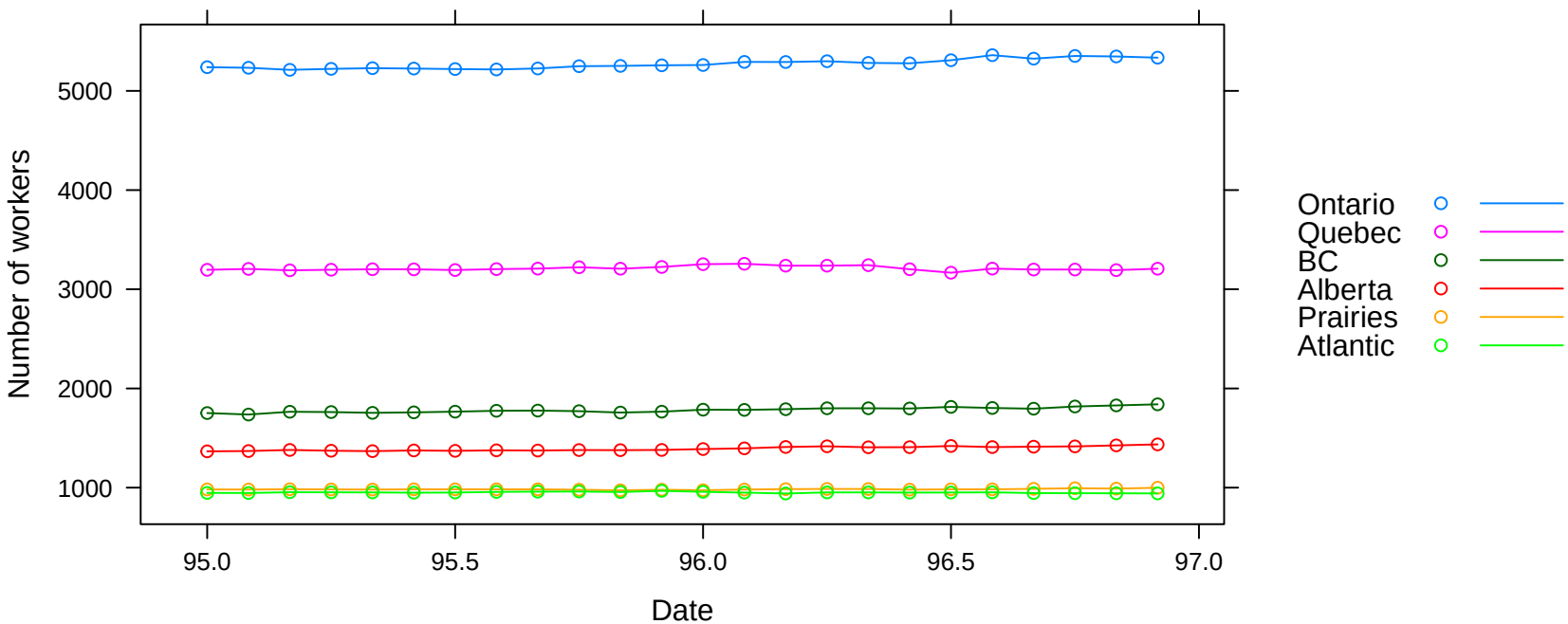


Eggs in wrens' nests appear smaller than eggs planted in other birds' nests; there are several outlying egg lengths in the meadow pipit nests.

Temporal data: comparing several time series

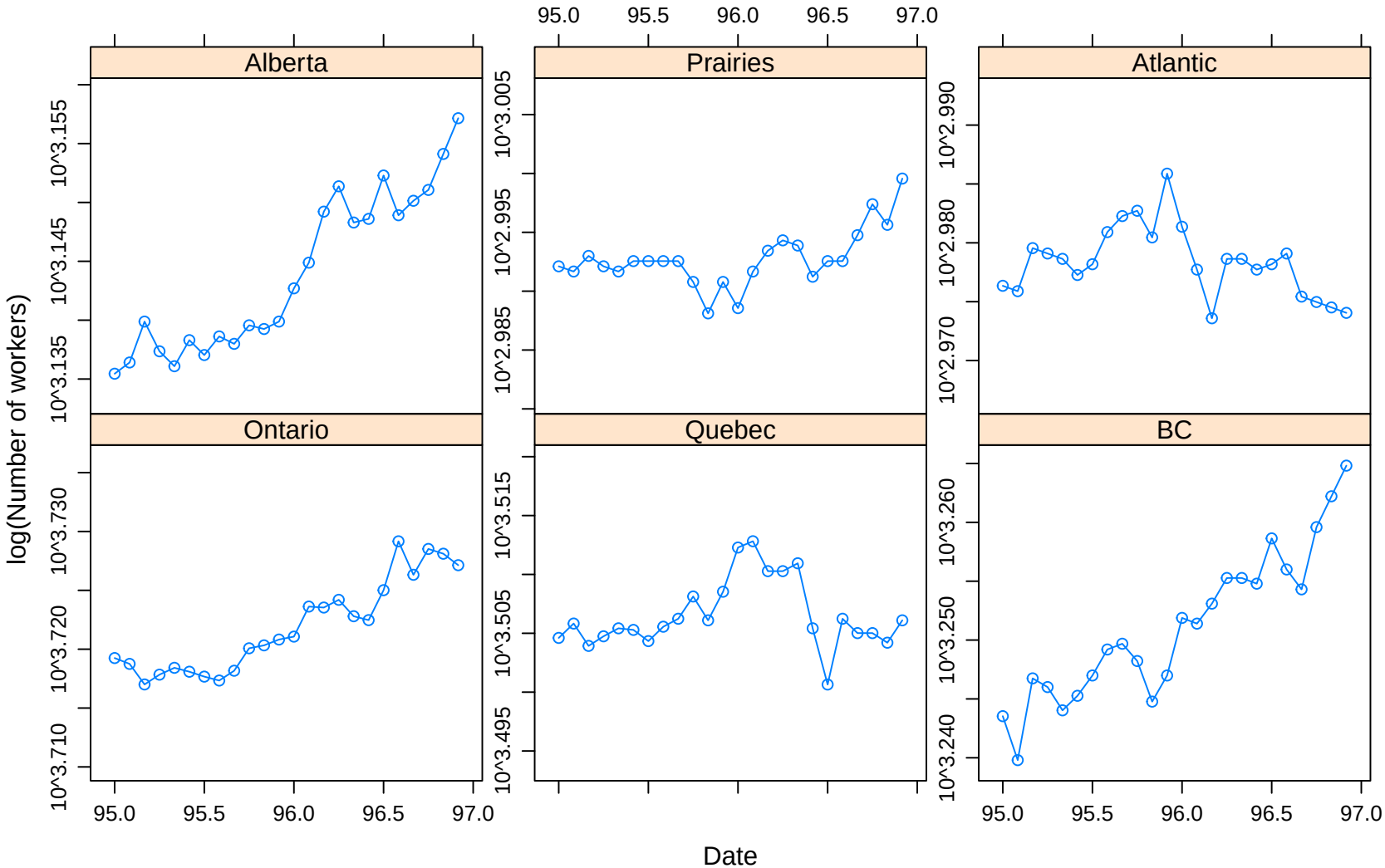
Overlaying plots of the time series might seem appropriate for making direct comparisons, provided that the scales are similar for the different series.

Multiple time series of `number of workers` (in thousands) in Canadian labor force, from 6 regions, from January 1995 to December 1996



The labor forces in the various regions do not have similar sizes so that it is impossible to discern any differences among the regions.

In order to account for different sizes and to consider relative changes: six different panels using the logarithmic scale



Alberta and BC experienced the most rapid job growth during the period.

Scatterplots, broken down by multiple factors

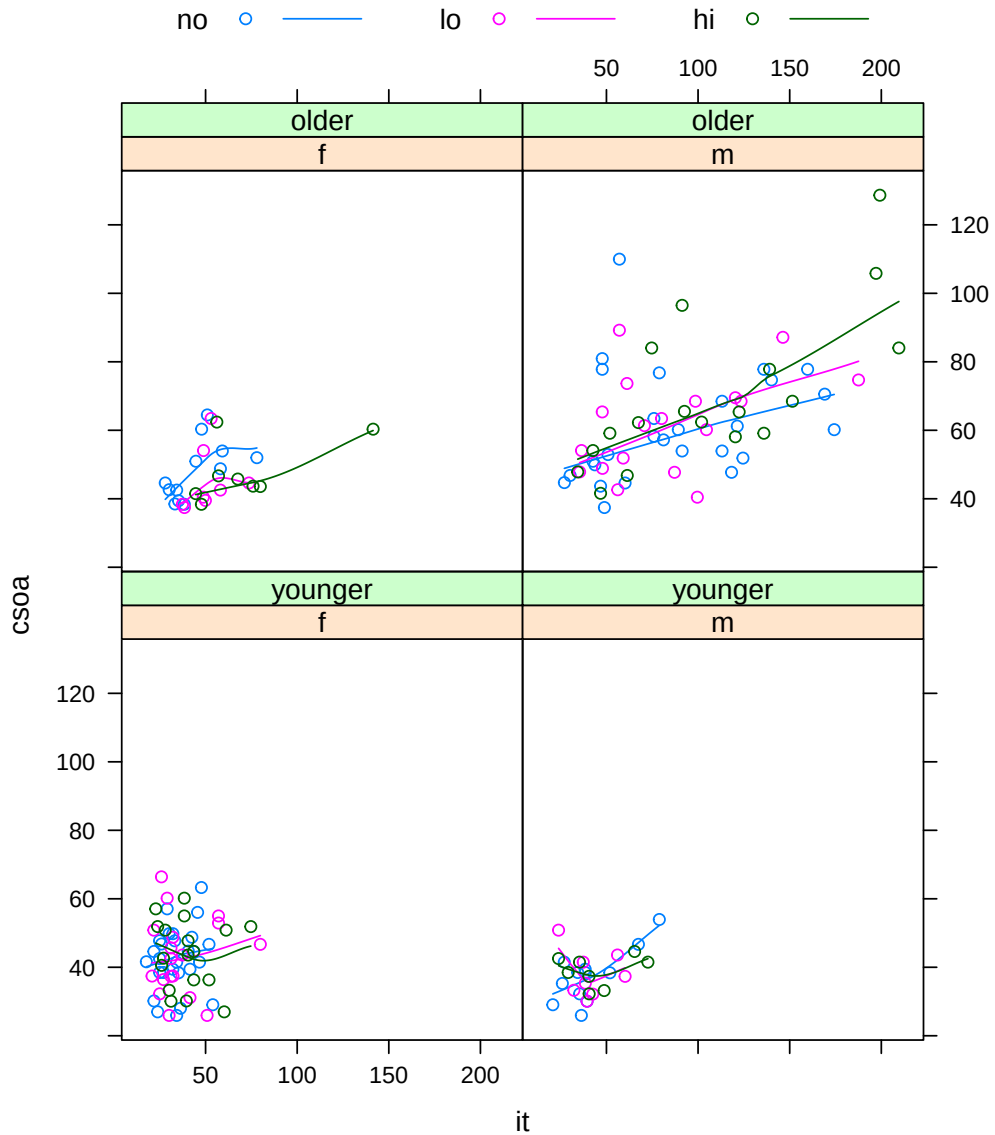
Data from an experiment on the effect of car window tinting on visual performance.

Numerical variables: `csoa` (time in msec to recognize a target), `it` (time in msec for a simple discrimination task) and `age` (to the nearest year).

Categorical variables (factors): `tint` (3 ordered levels, `no, lo, hi`), `target` (2 levels, `locon, hicon`), `sex` (2 levels, `f,m`) and agegp (2 levels, `younger, older`).

Each of 28 subjects was tested at each level of `tint` for each of the two levels of `target`, and all in all there are four factors that may influence two response variables `csoa` and `it`, and also the relationship existing between them.

Plot of `csoa` against `it` for each combination of `sex` and agegp, with different colors depending on whether the `tint` is absent, low or high.



The longest times are usually for the high level of tinting, the relationship between `csoa` and `it` seems much the same for different levels of tinting.

What plots may reveal

- **Outliers**: points that look isolated from the main body of the data; they may convey quite useful information, but are not easy to detect in high-dimensional data sets; they may depend on the scale.
- **Asymmetry of the distribution**: positive or negative skewness is often encountered with socio-economic data; symmetry is preferable, and it is typically obtained by transforming the data; **kurtosis**, i.e. heaviness of the tails of the distribution, is a further key feature.
- **Changes in variability**: usually not difficult to detect by graphical summaries; when variability increases as the data values increase, the log transformation (or the square root one) is usually a good idea.
- **Clustering**: clusters of separate points may be quite informative, and often correspond to the values of some relevant variable; it may also happen that such variable is not recorded.
- **Nonlinearity**: linear relationships are often a good approximation to some more complex forms; choosing the right scale may be crucial, as multiplicative relationships become linear on the log scale; sometimes the relationships are inherently non-linear.

Table of contents

The importance of data summaries

There are at least three reasons to value (numerical) data summaries.

1. They might be important *per se*.
2. They may give insight into aspects of data, that can be relevant for subsequent analysis.
3. They may be used as data for further analysis, though this requires some caution to avoid information loss.

Appropriate data summaries depend on the nature of the data at hand and the focus here is on:

- summary statistics used to describe the distribution of univariate (numerical) variables with regard to **location**, **variability**, **symmetry** and **kurtosis**;
- summaries for multivariate and, in particular, for **bivariate numerical data sets**;
- summaries for **counts data** (categorical variables).

Some basic summaries for univariate data

Measures of location

- The most common measure of location is the (**sample**) **mean** $\bar{y}$, which can be computed only for numerical variables.

  Let $y = \{y_1, \ldots, y_n\}$ represent the (sample) values of a variable Y

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

  It can be influenced by outlying observations.

- Another measure of position is the (**sample**) **median** $y_{0.5}$, which can be computed for both numerical and ordered categorical data

$$y_{0.5} = \begin{cases} \text{the middle obs.} & \text{if } n \text{ is odd} \\ \text{the average of the two middle obs.} & \text{if } n \text{ is even} \end{cases}$$

  where the sample is sorted in ascending order.

  It is less influenced by outliers, then it is more robust than $\bar{y}$.

- A further simple measure of location is the **mode**, which can be computed for all kind of variables.

  It is the value associated with the greatest frequency.
- The notion of (**sample**) **quantile** generalizes the notion of median.

  The α-quantile y_α, with $\alpha \in (0, 1)$, is the value which splits the frequency distribution into two parts, corresponding (approximately) to $\alpha 100\%$ observations (on the left) and $(1 - \alpha)100\%$ observations (on the right).
- The three **quartiles**, $y_{0.25}$, $y_{0.5}$, $y_{0.75}$, divide the distribution into four equal parts; similarly, the **percentiles** divide the distribution into one hundred equal parts.

Some basic summaries for univariate data

Measures of variability

- The **range** $R = \max(y) - \min(y)$ and the **interquartile range** $IQR = y_{0.75} - y_{0.25}$ are simple measures for numerical data.
- The most commonly used measure of variation for numerical data is the (**sample**) **variance**

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

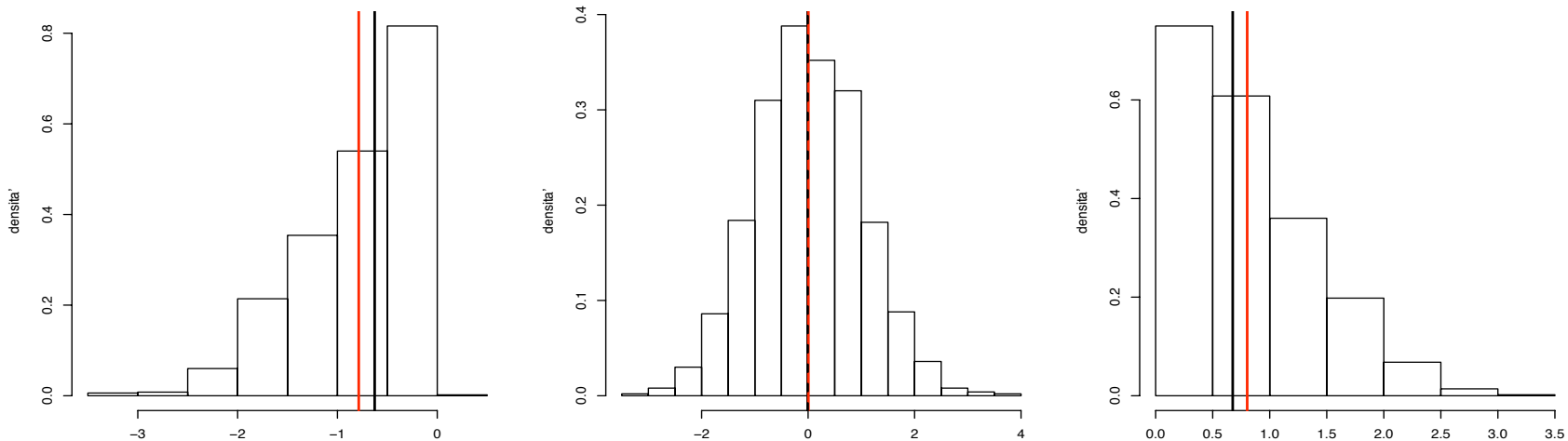
  where the denominator $n-1$ is the number of **degrees of freedom** remaining after estimating the mean with $\bar{y}$.

  The positive square root s is called (**sample**) **standard deviation**.
- The median of the transformed data set $|y - y_{0.5}|$ is called **median absolute deviation** (MAD) and it is the robust counterpart of s.
- The **coefficient of variation** $CV = s/|\bar{y}|$ is a standardized measure of variability, useful for data comparison.
- For categorical data: indices of heterogeneity and of concentration.

Some basic summaries for univariate data

Measures of asymmetry and kurtosis

- Graphs (barplots, histograms, density estimates and boxplots) are useful for investigating the shape of the frequency distribution.
- A preliminary indication of the **asymmetry** (**skewness**) of a distribution may be obtained by comparing the mean and the **median**:
  - negative skew: $\bar{y} < y_{0.5}$ (left panel);
  - symmetric distribution: $\bar{y} \approx y_{0.5}$ (central panel);
  - positive skew: $\bar{y} > y_{0.5}$ (right panel).



- The most common **index of skewness** is

$$\gamma = \frac{n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^3}{s^3}$$

  If the distribution is symmetric, $\gamma \approx 0$; if it is skewed to the left, $\gamma < 0$; if it is skewed to the right, $\gamma > 0$.
- The analysis of the **kurtosis** concerns the shape of a frequency distribution, focusing on tail weight and peakedness.
- The kurtosis is evaluated with respect to the normal distribution (the "bell curve"), considered as the reference standard.
- The most common **index of kurtosis** (it measures tails heaviness) is

$$\beta = \frac{n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^4}{s^4}$$

  If the distribution is normal shaped, $\beta \approx 3$; if it is hyponormal (thinner tails), $\beta < 3$; if it hypernormal (fatter tails), $\beta > 3$.

Summaries for bivariate numerical data sets

Correlation

- The relationship between two numerical variables can be graphically represented by a scatterplot; in case of more variables: scatterplot matrix.
- A relevant measure of the *linear* relationship between two numerical variables is the (**sample**) **covariance** s_{xy}.

  Let $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$ represent the (sample) values of the variables X and Y, observed on the same units,

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

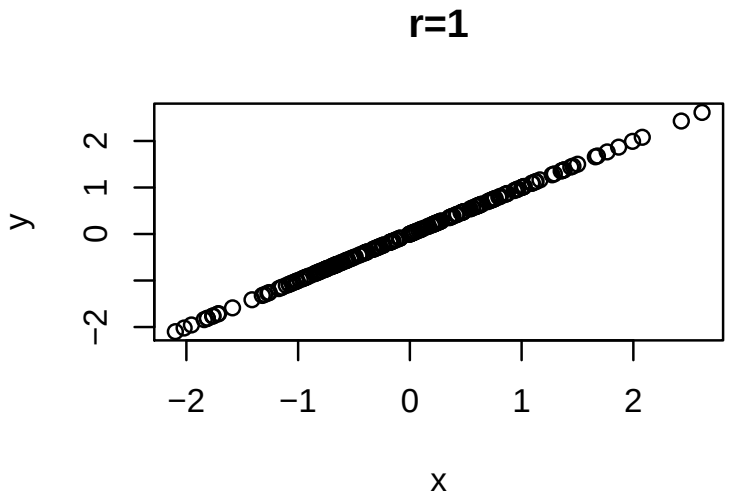
  The sign of the covariance shows the tendency, positive or negative, in the **linear relationship** between the variables; $s_{xy} \approx 0$ indicates absence of linear relationship.

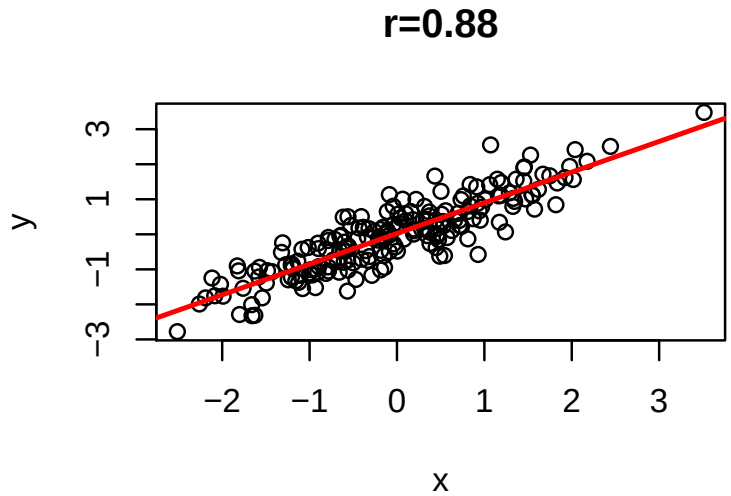
- The **Pearson correlation coefficient** is a standardized summary measure of linear relationship

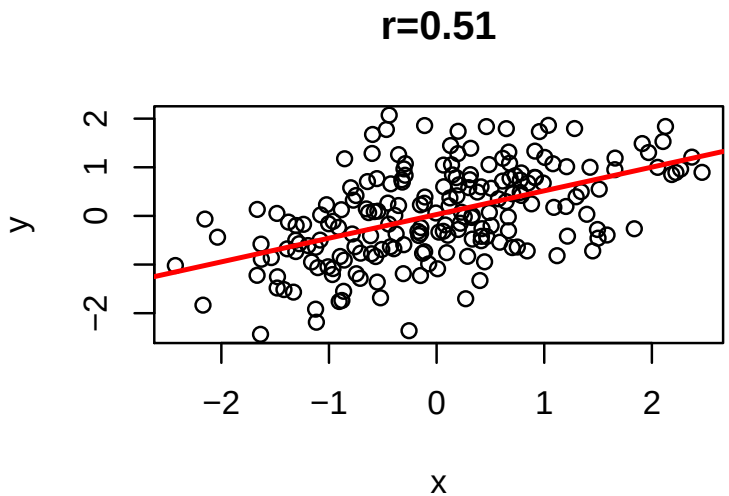
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

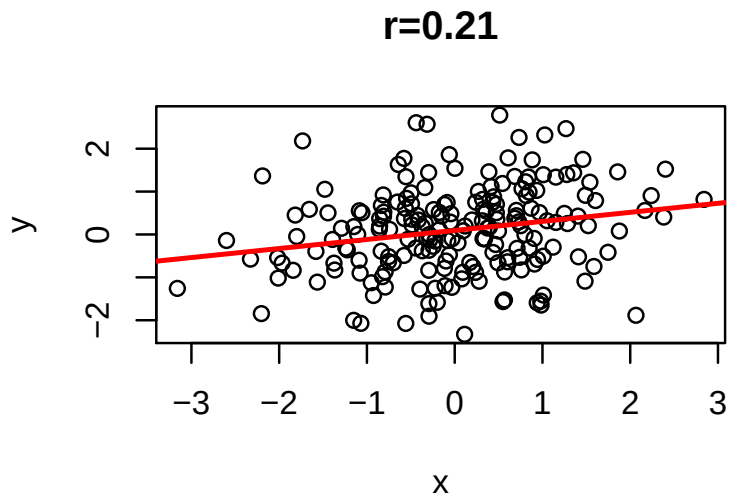
  $r_{xy} \in [-1, 1]$ and when $r_{xy} \approx 1$ ($r_{xy} \approx -1$) the data points tend to lay exactly on a line with positive (negative) slope.
- The calculation should go together with a scatterplot for checking linearity; a smooth trend is usually helpful.
- Another fact to check is whether the marginal distributions of the two variables are roughly normal, or at least not highly skewed.
- For monotonic nonlinear relationships or asymmetric marginal distributions the **Spearman rank correlation coefficient** should be used, namely $r_{x,y}$ computed using the **ranks** of x and y.

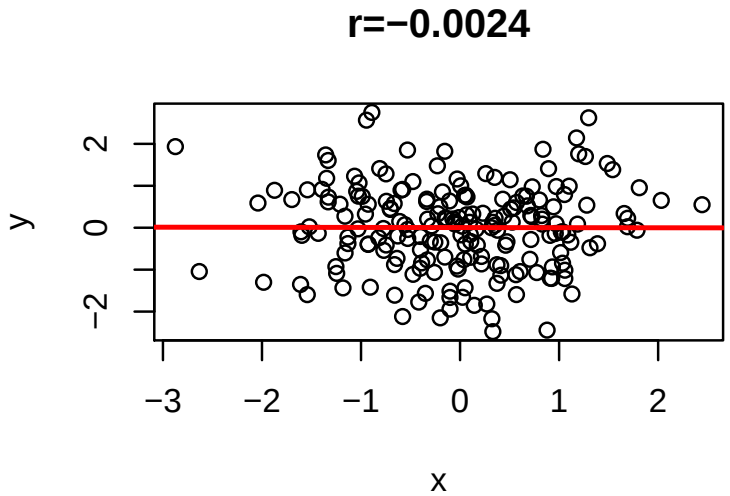
  A further alternative is the **Kendall's tau correlation coefficient**, based on the number of concordant and discordant pairs.

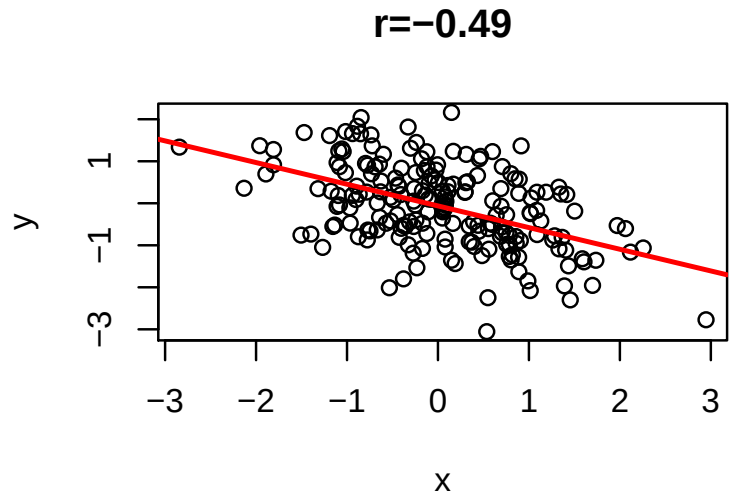

r=1
y
x
−2
−1
0
1
2

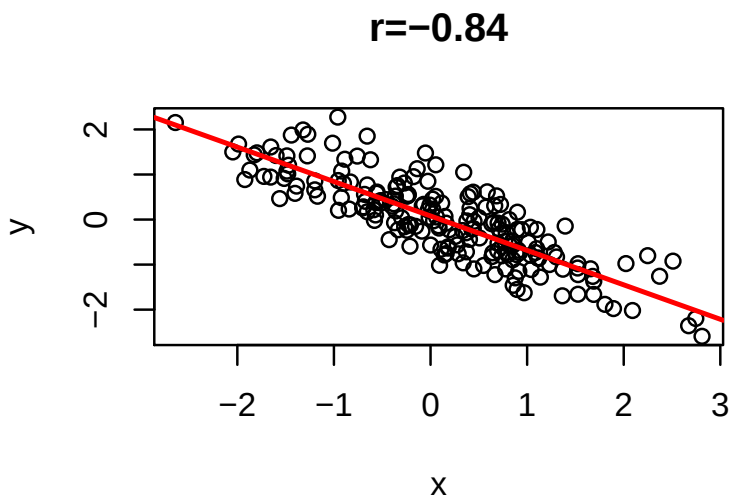

r=0.88
y
x
−3
−1
1
3
−2
−1
0
1
2
3

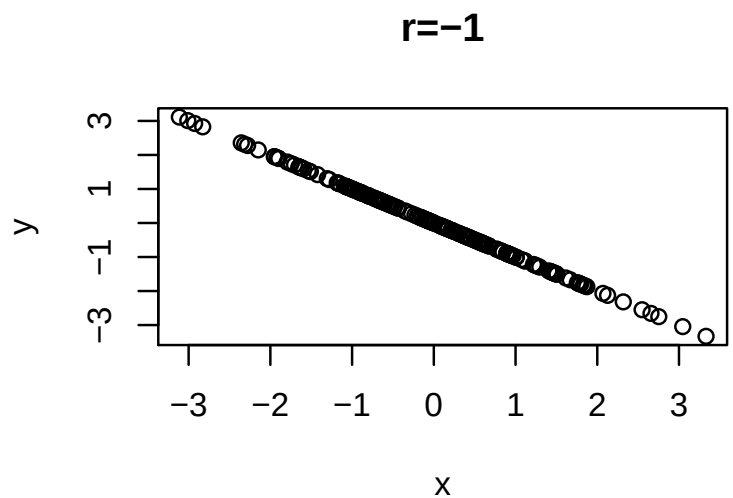

r=0.51
y
x
−2
0
1
2
−2
−1
0
1
2


r=0.21
y
x
−2
0
2
−3
−2
−1
0
1
2
3

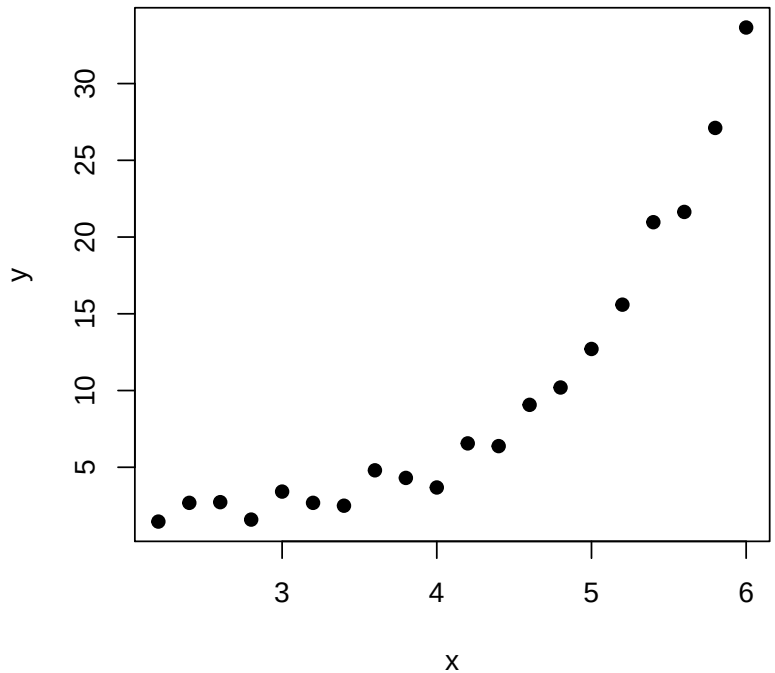

r=−0.0024
y
x
−3 −2 −1 0 1 2
−2 0 2


r=−0.49
y
x
−3 −2 −1 0 1 2 3
−3 −1 1


r=−0.84
y
x
−2 −1 0 1 2 3
−2 0 2


r=−1
y
x
−3 −2 −1 0 1 2 3
−3 −1 1 3

The following scatterplot shows a strong nonlinear pattern



- With the above data set, the Pearson correlation coefficient is 0.887, while the Spearman correlation coefficient, which better captures the strength of the relationship, is 0.958.
- The magnitude of r_{xy} does not of itself indicate whether the fit is adequate; here the linear fit is clearly inappropriate and the graphical representation may guide a suitable numerical analysis.

Summaries for count data

- Count data are usually given in the form of **contingency tables**, collecting the observed frequencies of each combination of variable categories.
- Summaries for count data require some care since information may be lost or obscured, for example when summarizing counts across the margins of multi-way tables.

A famous data frame contains the data from a study on unemployed individuals, focusing in particular on differences between those that followed a training program and those that did not.

The two groups, "treated" and "untreated", are not genuinely comparable; by considering who had completed high school and who had not, it is clear that training group has a much higher proportion of dropouts.

	high school	education	
treatment	completed	dropout	% completed
none	1730	760	68.5
training	80	217	26.9

Example: kidney stones

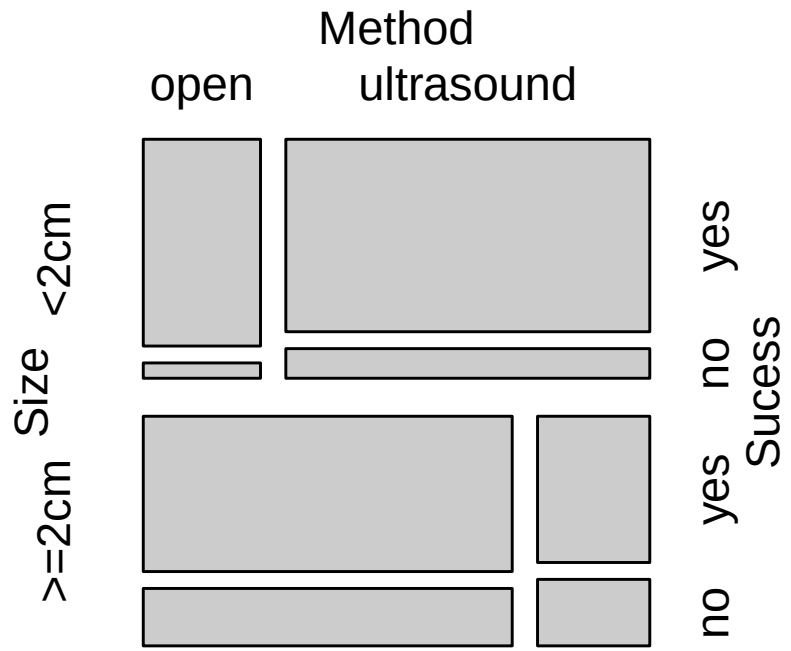
Data are from a study that compares outcomes for two different `methods` of surgery for kidney stones: `open`, which used open surgery, and `ultrasound`, which used a small incision and ultrasound.

Additional information on the `size` of the stones: $<$ 2cm, $\geq$2cm

		success		
method	size	yes	no	% yes
open	<2cm	81	6	93.1
	≥2cm	192	71	73.0
ultrasound	<2cm	234	36	86.7
	≥2cm	55	25	68.8

The success rate for each size of stone separately favors always open surgery.

The multi-way table is summarized using the following **mosaic plot**.



Summarizing the counts, by summing with respect to the size, produces a loss of information leading to the apparent conclusion that the success rates favor ultrasound

	success		
method	yes	no	% yes
open	273	77	78.0
ultrasound	289	61	82.6

Table of contents

Aim of the analysis

- The data available are not suitable for any possible question; ideally, they should be collected (or, even, generated) after the aims of the analysis have been planned.
- Usually, the aims include **scientific understanding** of some critical points (*is a training program effective in reducing unemployment?*) or **prediction** of some key variables (*which is the price that house purchasers may be willing to pay for a certain area and house size?*).
- Statistical data analysis can greatly help in answering scientific questions, but does not stand alone, and it must be interpreted against a background of **subject area knowledge**.
- A critical distinction is between what can be reached by an **experiment** and what can be reached by an **observational study**.

  Well designed experiments give **highly reliable results**, whereas observational studies require a lot of care, and can even be misleading.

Observational versus experimental data: the job training program example

Two different strategies for evaluating a job training program:

1. some enrolled subjects are **randomly assigned** to training and non-training groups, and after some time from the end of the course (for those under training) their job status is assessed;
2. some subjects **freely decide** whether to attend a training course or not, and after some time from the end of the course (for those under training) their job status is assessed.

The problem with 2. is that trained subjects may differ **systematically** from the untrained ones for reasons totally unrelated to the training courses, such as *motivation*; this does not occur, due to **randomization**, in 1.

Yet, in practice, 2. is more common than 1. in economic studies. There are methods to analyse those kind of data, but they are more difficult than those for 1.

Statistical analysis strategies

- A careful initial analysis of the data is essential in any statistical analysis, and should never been neglected.
- EDA techniques are also important to assess the results of more formal analysis, when statistical models are employed.
- For planning a formal analysis on experimental or observational data, all the available information should be considered. At times, **pilot studies**, limited in size, could be performed before designing a more extensive experiment.

  EDA of the pilot study, or of data from previous studies, is then a crucial step.