Descriptive statistics

Maarten Jansen

Overview

1. Classification of observational values
2. Visualisation methods
3. Guidelines for good visualisation

1.Classification of observational values

The representation (statistical description) of observations and the statistical analysis depend on the nature of the measurement and the possible outcomes.

Types (“Levels”, “Scales”) of measurements

Observations can be classified into 4 groups accoring to the type of measurements

1. Nominal = categorical scale
2. Ordinal scale
3. Metric, numeric or quantitative data
   (a) on an interval scale
   (b) on a ratio scale

Nominal/categorical data

1. **Definition**
   Observations whose outcomes can be classified into groups (sets) without inherent order
2. **Examples**
   - Dichotomy: yes/no, man/wife (two classes)
   - More than two classes: colors, trade marks, ...
   - Postal codes: numbers but not numerical data (see below)
3. **Operations**
   - Variables can be encoded with a number
   - Only a limited number of operations make sense. E.g., if yes = 1 and no = 0, then sum of observations is number of yes’s
4. **Mathematical structure** = **sets**
5. **Descriptive statistical analysis**
   - Central tendency: through **mode** as measure for (no mean or median)
   - Dispersion: not possible

6. **Visualisation (of the distribution)**
   - Bar charts (see below)
   - Pie charts (see below)

The same charts can be used to visualise the mean, median, standard deviation of numerical variable as a function of nominal/ordinal variable

The values of a numerical variable as a function of nominal/ordinal variable can also be summarized as multiple boxplots: one boxplot for each nominal value

7. **Statistical inference**:**chi-squared** tests (ANOVA on frequencies)

$$\chi_0^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

ordinal data

1. **Ordinal scale**: natural order, but not quantifiable.
2. **Examples**
   - “very bad”, “bad”, “medium”, “good”, “very good”, “excellent”
   - “very unsatisfied”, “unsatisfied”,“
   - Beaufort scale for wind speed
   - IQ score

   Operations
   - Variables can be encoded with a number that reflects order
   - No arithmetic operations (such as sums/means, differences). One cannot define differences in IQ, as IQ 120 minus IQ 100 is not the same as IQ 80 minus IQ 60.
3. **Mathematical structure** = **ordered set**
4. **Descriptive statistical analysis**
   - Central tendency: median or mode, no average

   - Quantiles, range measure dispersion
     Median absolute deviation (= based on difference between observations and mean or median) are not allowed
5. **Visualisation**
   - Bar charts (see below)
   - Pie charts (see below)
6. **Statistical inference** through rank order tests

Interval scale

1. **Interval scale**: numerical data, buth no natural zero point. As a consquence, ratios of data make no sense.
2. **Examples**
   - Degrees Celsius: the zero point is defined artificially. 20 degrees is not twice as warm as 10 degrees.
   - The first day in the year with a temperature above $20°\mathrm{C}$. Even if such a day is denoted as Day/Month, it is still numerical.

   Operations
   - Sums (means), differences
   - No ratios (“twice as warm”), hence no logarithms
3. **Mathematical structure** = **affine line**
4. **Descriptive statistical analysis**
   - Central tendency: mean (average), median or mode
   - Dispersion: standard deviation, (empirical) variance, median absolute deviation

5. **Visualisation**
   - Histogram
   - Box (and whisker) plot
6. **Statistical inference** t-test for means, F test for variances
   Note that we cannot say that $20°\mathrm{C}$ is twice as warm as $10°\mathrm{C}$, but we can say that a standard deviation of $2°\mathrm{C}$ is twice as large as a standard deviation of $1°\mathrm{C}$, so we can test whether one climate has a standard deviation in temperatures that is twice the standard deviation in another climate.

Ratio scale

1. **Ratio scale**: the same as interval scale, but with absolute zero. Taking logarithms, computing ratios make sense.
2. **Examples**
   - Proportions p
   - Body length, body weight
   - Waiting time at a bus station

Sets of outcomes in numerical data

Numerical data: Interval or ratio scale

Other subdivision: Continuous, discrete (finite, countable) or mixture (=point mass + continuous distribution; e.g.: precipitation: point mass at zero)

Dimension: univariate, multivariate (also possible with non-numerical data)

Longitudinal data

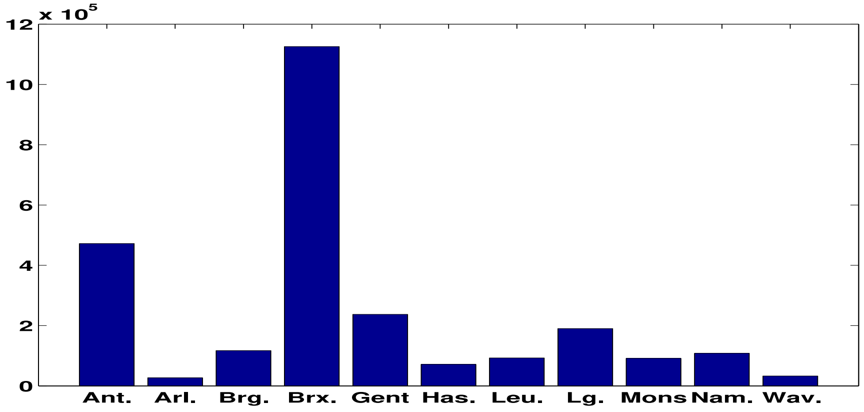
Visualisation methods for several types of data

- Nominal/ordinal data
  - Bar chart
  - Pie chart
- Numerical data
  - Stemplot (stem-and-leaf plot)
  - Histogram
  - Box-whisker plot
- Multivariate data
  - Scatter plot
- Longitudinal data
  - Profiles

2. Visualisation methods

Bar chart

Example: population of Belgian provincial capital cities

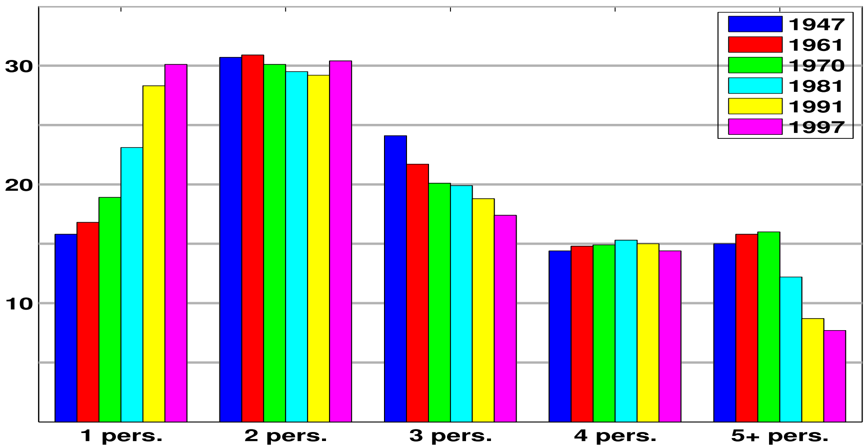


Bar chart

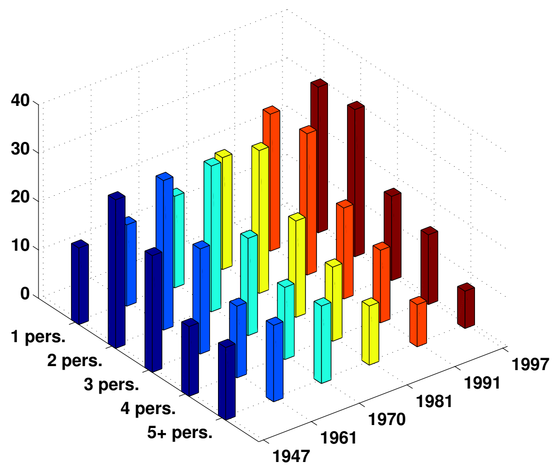
- Absolute or relative frequency
- For every outcome
- Heights of bars represents frequencies, never the area (because area = width $\times$ height, and width = difference, which cannot possibly defined on nominal/ordinal variables)
- Nominal (categorical) or ordinal data. Sometimes also for discrete data
- In principle, not for numerical data, as the bar width is a dimension that could be (falsely) interpreted as a difference between knots or observations. Use histograms instead.
- Multivariate data: 2-d graph or 1-d (see next slides)
- Do not create unnecesary dimensions
- Be aware of visual manipulations, e.g., if horizontal axis is not at height zero (Sometimes, the horizontal axis should not be at height zero; e.g.: temperatures in Kelvin; it all depends on the goal of the graph)

2-dimensional data

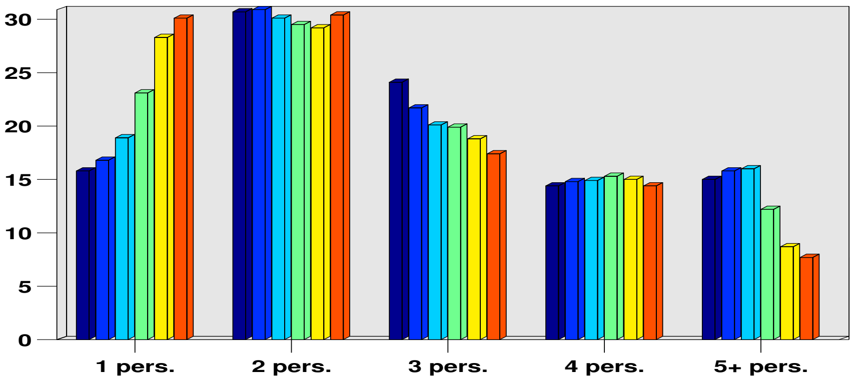
E.g.: number of persons in households

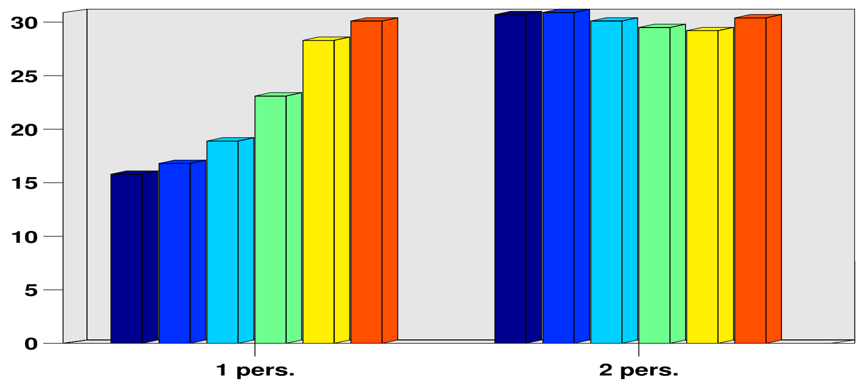


2-dimensional representation



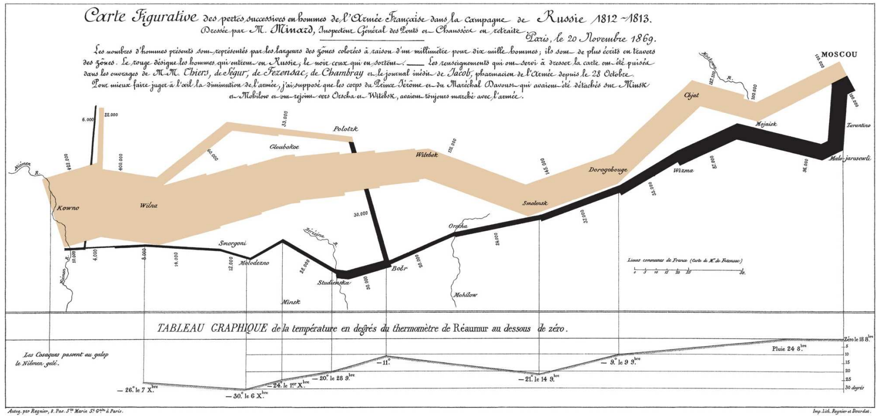
Unnecessary dimensions





Example: Charles Joseph Minard

(ref.: Edward Tufte, The Visual Display of Quantitative Information)



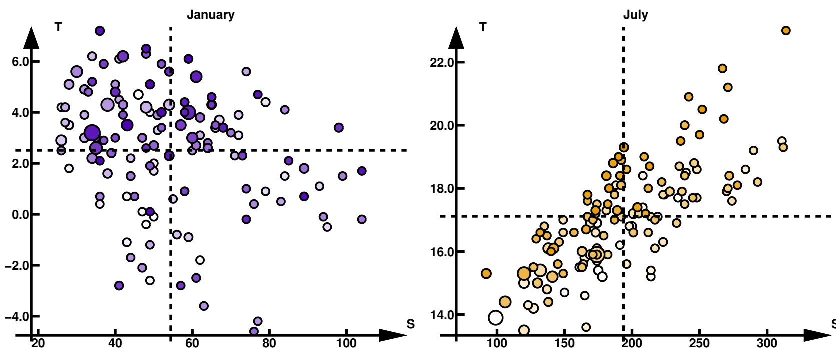
This is a flow map, representing six variables: the location of the army (2D), size of the army, temperatures, direction of movement, time

Other example: climate data (Brussels)

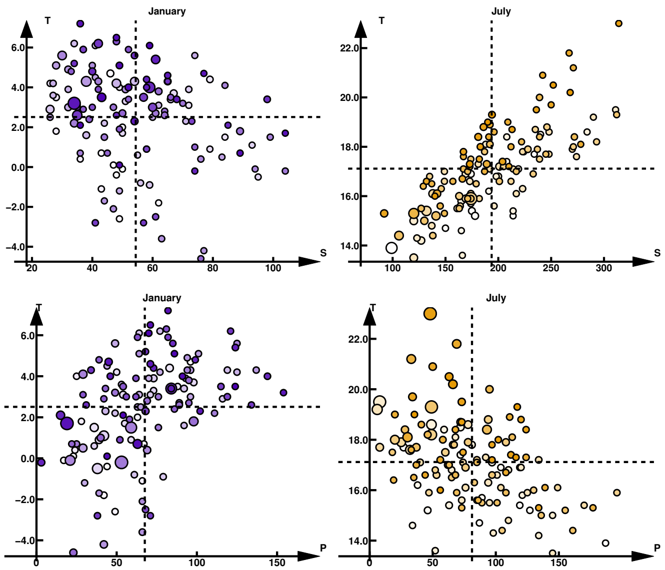
Scatter plots representing four variables: (montly/seasonal/yearly values)

1. Sunshine (x-axis)
2. Temperature (y-axis)
3. Precipitation (size of circles)
4. Year (color intensity)

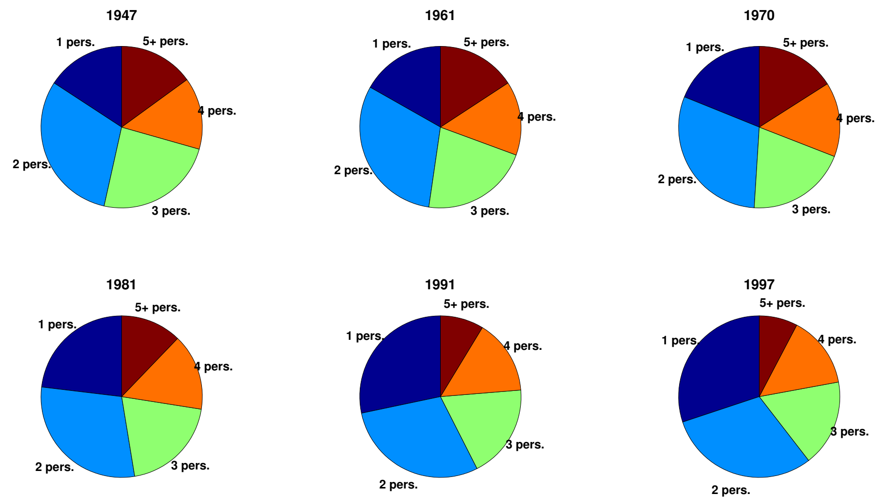
Not all dimensions are represented with the same precision



Think about how to display relations between variables



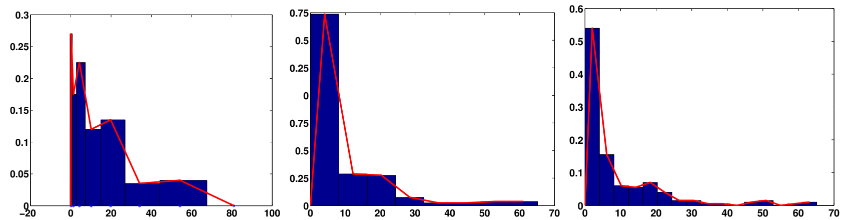
Pie Chart



Histogram

- = bar chart for (disjoint) **bins** of continuous data
  (also for large samples of discrete data)
- Loss of information
- Bins: number of bins, bin width, bin centers
- Rules of thumb: $k = \sqrt{n}$, $k = 1 + 3.3\log(n)$, $7 \leq k \leq 20$, $k = \lceil \log_2(n) + 1 \rceil$
  (for $n > 30$)
- Height represents (absolute) numbers (frequency histograms) or proportions (relative frequency histograms)
- **Frequency polygons** connect bin centers at bin heights.
  Area under frequency polygon = area of histogram

Illustration: histograms and frequency polygons



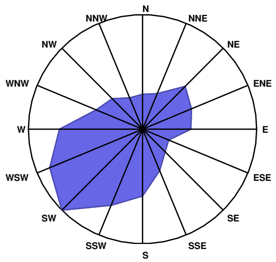
Three histograms for the same data, with corresponding frequency polygons

A “circular” frequency polygon

Example: distribution of wind direction (in Brussels, 1971-2000).

Although wind directions have names, these data are numerical, and cyclic/periodic.

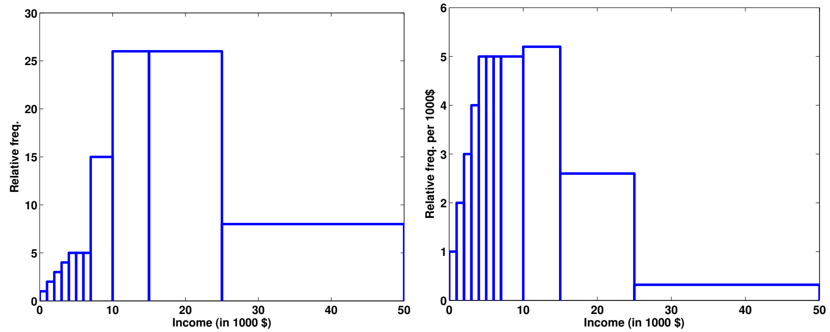
N	NNE	NE	ENE	E	ESE	SE	SSE
3.8	4.2	6.6	5.8	5.3	3.1	3.5	4.9
S	SSW	SW	WSW	W	WNW	NW	NNW
7.3	9.0	12.5	11.0	9.1	5.5	4.7	3.7



Density histogram

Classical histogram may be misleading is not all bins have equal width.

Density histogram: <u>area</u> represents frequency

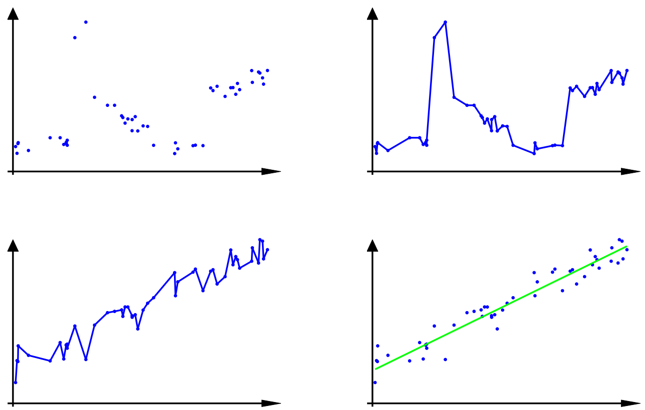


(Yearly income 5000 American families in 1973)

Line charts or line graphs

- Successive observations, connected by line segments (= polyline)
- Display the data in order of measurement, e.g., in time series
- Connecting with lines emphasizes order of measurement, can be useful even if line segments themselves have no physical meaning
  - Visualisation
  - May reveal trends
- If order of measurement is unimportant or nonexistent, avoid line segments (see next slide)
- Only ONE random variable (this is not to be confused with scatter plots)

Use and abuse of line charts

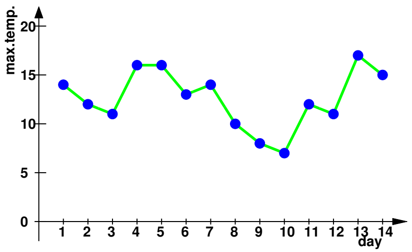


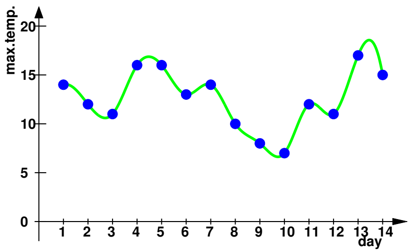
- The figs on top show a time series: order of measurement is important; lines reveal tendency
- The figs on the bottom show linear model: order of observation is irrelevant, or even nonexistent

Use and abuse of interpolating splines

Do not use (interpolating) splines for discrete time series.

Example: maximum temperatures: only one temperature per day, interpolation is meaningless

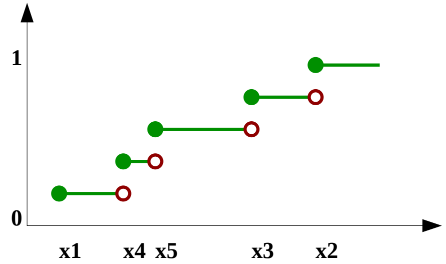


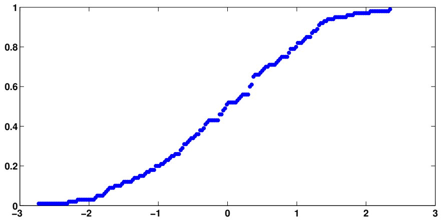


Cumulative distribution

$$\widehat{F}(x) = \frac{\#\{i \in \{1, \ldots, n\} | x_i \leq x\}}{n}$$

Order statistics $x_{(k-1)} \leq x_{(k)} \leq x_{(k+1)}$ $\widehat{F}(x_{(k)}) = k/n$

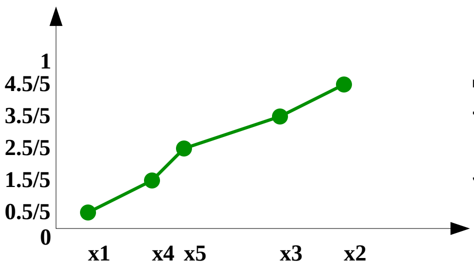


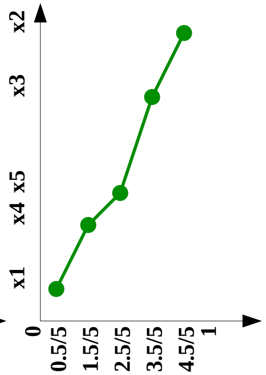


Empirical quantiles

= a polyline (= continuously connected line segments) with knots: $\widehat{Q}\left(\frac{k-1/2}{n}\right) = x_{(k)}$

where $x_{(k)}$ is the order statistic.





Central tendency

Sample mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

If x_i can take k discrete values with absolute frequency f_j, then $\overline{x} = \frac{1}{n}\sum_{j=1}^{k} m_j f_j$

If x_i can take continuous values in k bins around centers m_j, then, approximately: $\overline{x} \approx \frac{1}{n}\sum_{j=1}^{k} m_j f_j$

Linearity: if $y_i = ax_i + b$, then $\overline{y} = a\overline{x} + b$

(e.g. Celsius — Fahrenheit)

Central tendency (2)

Sample median

$\boxed{\tilde{x} = \widehat{Q}(0.5)}$

If n odd: $\boxed{\tilde{x} = x_{(\frac{n+1}{2})}}$

If n even: $\boxed{\tilde{x} = \frac{1}{2}\left[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right]}$

Linearity: if $y_i = ax_i + b$, then $\tilde{y} = a\tilde{x} + b$

Median is (more) **robust** against **outliers**

Trimmed/truncated mean — Drawback: choice trimming percentage, loss of information

Mode (concepts unimodal, multimodal)

Dispersion (1)

ADM: average distance to mean

$\boxed{\mathrm{AMD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|}$

MAD: median absolute deviation

$\boxed{\mathrm{MAD} = \mathrm{med}(|x_i - \mathrm{med}(x_i)|}$

Sample variance $\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$

Factor $n-1$ in denominator

1. because of mathematical-statistical properties: unbiased estimator
2. case $n=1$:
   If $n=1$, then $s^2 = 0/0 = \texttt{Not}-\texttt{a}-\texttt{Number}$, which can be interpreted as: "we don't know".
3. case $n=2$:
   $\overline{x}$ is the mean; $x_1 - \overline{x}$ is the opposite of $x_2 - \overline{x}$, hence, $(x_1-\overline{x})^2 = (x_2-\overline{x})^2$. We only have one independent observation of the deviation from the mean. The mean deviation is thus a mean over one observation.

Standard deviation

Sample standard deviation $\boxed{s = \sqrt{s^2}}$

Behavior under linear transforms

If $y_i = ax_i + b$, then $s_y = as_x$

Dispersion (2)

Range $\boxed{R = x_{(n)} - x_{(1)}}$

Quartiles

First quartile = 25% - quantile = $\widehat{Q}(0.25)$

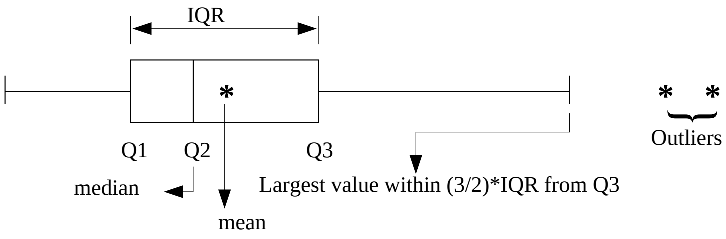
Second quartile = 50% - quantile = $\widehat{Q}(0.5)$ = median

Third quartile = 75% - quantile = $\widehat{Q}(0.75)$

Inter quartile range $\boxed{\mathrm{IQR} = \widehat{Q}(0.75) - \widehat{Q}(0.25)}$

Box-and-whisker plots

Graphical summary of observations through central tendency and dispersion measures for numerical data



Information about

1. central tendency
2. dispersion
3. skewness (asymmetry)
4. curtosis (flatness) and outliers

Little info about bimodality

Multivariate data

Sample covariance

If X and Y are random variables, then $\boxed{\mathrm{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}$

Visualisation by a pairwise scatterplot

Linear transformation

If $u_i = a_1x_i + b_1$ and $v_i = a_2y_i + b_2$, then $\mathrm{cov}(u,v) = a_1a_2\mathrm{cov}(x,y)$

Pearson correlation $\boxed{r(x,y) = \frac{\mathrm{cov}(x,y)}{s_xs_y}}$

We have $r(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right)$

$-1 \leq r(x,y) \leq 1$

Correlation measures linear association

If $y_i = ax_i + b$, then $r(x,y) = \mathrm{sign}(a)$

If $y_i = ax_i^2 + bx_i + c$, then $r(x,y) \neq \pm 1$ although the association between x and y is fixed/deterministic

3. Guidelines for good visualisation

Graphics

1. Graphs should be **convincing** and should have a clear **focus** (i.e., think about you want to illustrate; typicall a relation between variables)
2. Avoid optic effects
   - Surface, volume
   - Useless colors (see below)
   - Perspective (additional dimension)
3. Use colors or grey scales for adding a dimension, e.g., in maps
4. Choice of scale
5. Choice of horizontal/vertical axis: away from zero may be deceiving
6. Horizontal frame, not vertical
7. Captions should be informative

Tables

1. Remove useless information
2. Arrange numbers that are to be compared into columns (not rows)
3. Only show (the most) significant figures
4. Use appropriate units
5. Avoid scientific notation (difficult to compare visually)
6. Order the elements in a quantitative (not an alphabetical or random) way
7. Avoid excessive use of colors (colors may highlight or structure, e.g., alternating colors in rows/columns)

Lines or points

- In observations with a **natural order** (observations ordered by time or distance) it maybe usefull to plot the observations as a broken line (polyline) for reasons of **visualisation** (detecting trends). These line segments should not be interpreted as having a physical meaning on their own: they support the global visual interpretation.
- When the observations are expected to follow a model + noise, it may be better to draw a regression curve through the data
- When there is no natural order, observations should not be connected by line segments

4. Guidelines for exploratory analysis

- A statistical report starts by an **exploratory analysis** of the data
- The **objective** is to discover global trends graphically, **independent from a statistical model**. The statistical model may be inspired by the plots in the exploratory analysis.
- This analysis uses descriptive statistics, but keeping in mind the **objectives** of the study, that is: simple boxplots of observed values do not reveal any of the trends to be investigated. So it is better to present boxplots of categories of observations, corresponding to the trends to be investigated (see also examples of bar charts)
- QQ-plots can be used for visual validation of the normality assumption in a model after the statistical inference. This validation takes place on the **residuals**, never on the response variables before the inference: in principle no QQ-plots/normal probability plots in the exploratory analysis.