Introduction to Statistics

Exploratory Statistics

Alexandra Posekany
alexandra.posekany@gmail.com

Frequentist statistics

Frequency distribution / table

A **frequency distribution** (or **frequency table**) shows how a data set is partitioned among all of several *categories* (or classes) by listing all of the categories along with the number of data values in each of the categories.

Differentiate between:

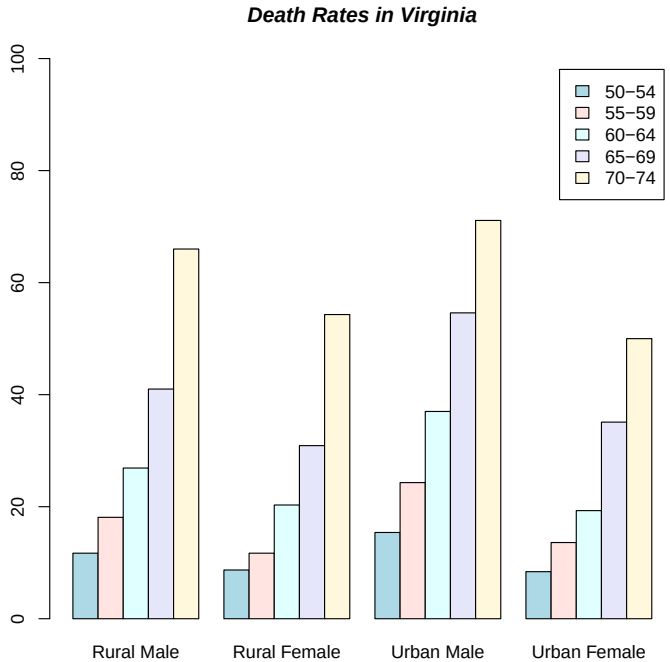
- Relative frequency distribution
- Absolute frequency distribution
- Cumulative (absolute/relative) frequency distribution

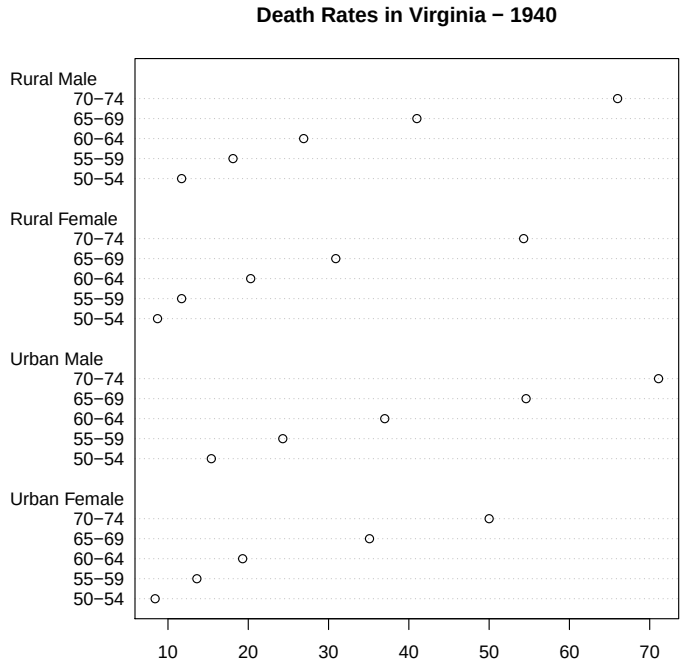
Visualising frequencies

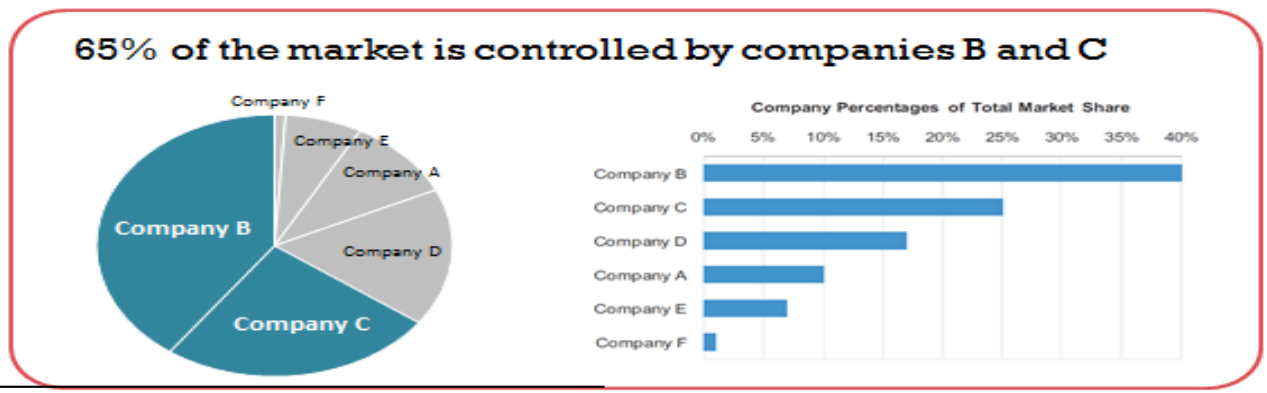
Frequency plots

- **bar charts** visualise absolute or relative frequencies of categories (recommendable in 90% of cases)
  human eye and brain discern lengths better than angles
  R: barplot
- **Cleveland dot charts** visualise absolute or relative frequencies of categories
  alternative to bar charts
  R: dotchart
- **pie charts** visualise relative frequencies
  only when visualising majorities (in order to make a pie chart fully interpretable, always add all percentages)
  R: pie

Frequency Plots



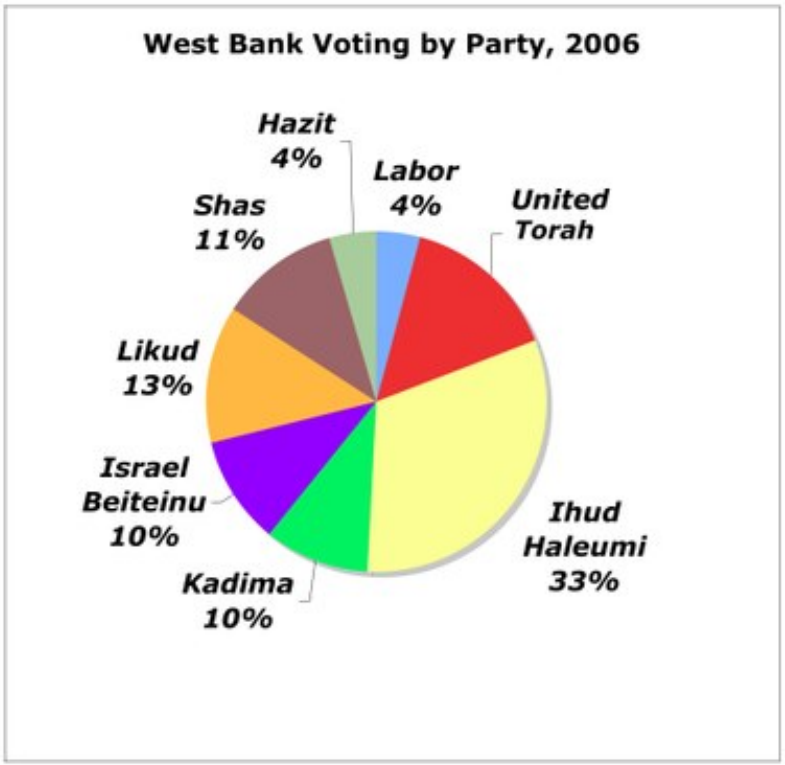


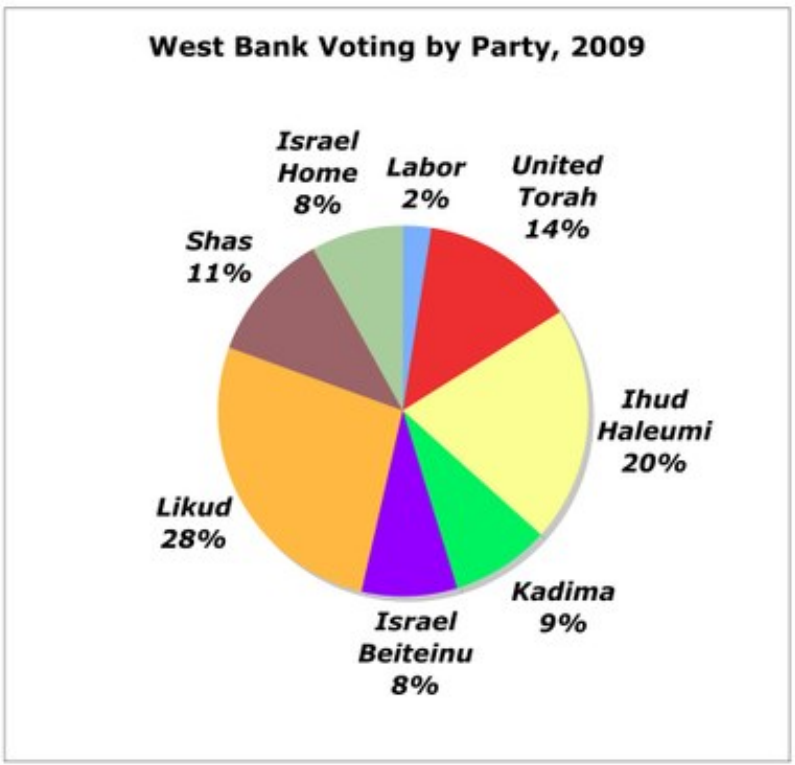


[1]

[1] http://speakingppt.com/2013/03/18/why-tufte-is-flat-out-wrong-about-pie-charts/

When pie charts are useful





[2]

[2] Nissan Ratzlav-Katz, "Judea and Samaria Switch Over to Likud," Arutz 7, February 12, 2009; Central Electoral Committee, "Results of the 17th and 18th Knesset Elections."

Cross tabulation / contingency tables

Example: Skin color and death sentence

Data from *New York Times Magazine*, March 11, 1979, concerning the frequency of death sentences in Florida.

Case no.	Skin color of accused	Death sentence
1	b	0
2	b	0
3	w	0
4	b	1
⋮	⋮	⋮
4764	w	0

Structure of a table

Table

	Variable 1	
Var. 2	F_{11} $\cdots$ $\vdots$	Marg. 1
	Margin 2	observations

R: table(data.frame)

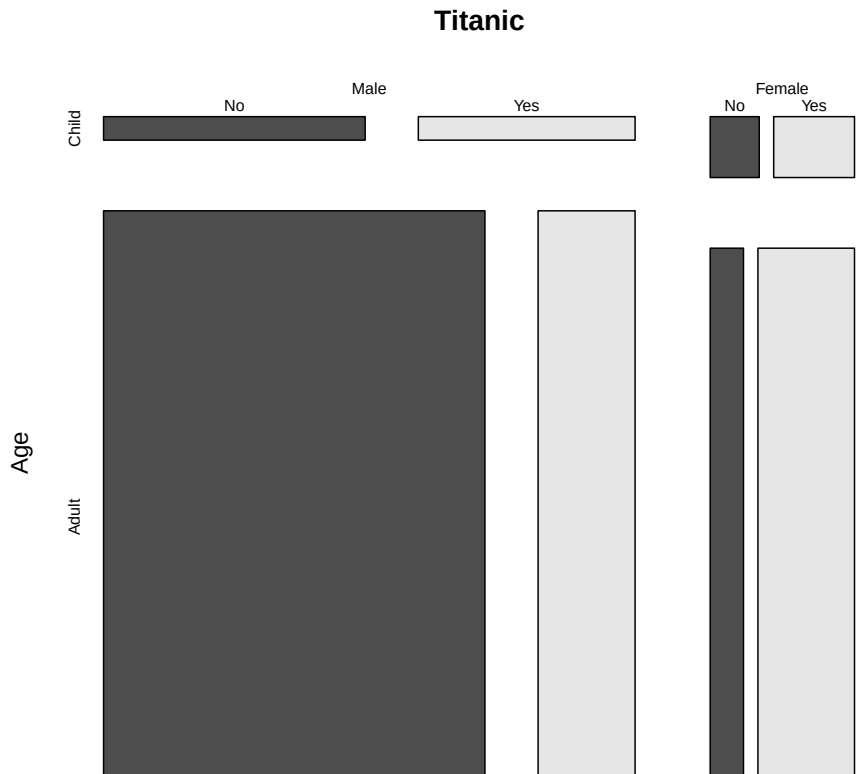
We observe:

- absolute frequencies of all subcategories
- marginal frequencies (relative w. r. t. one margin, i. e. keeping one parameter fixed)

Visualising contingency tables

Mosaic plots

Visualises contingency tables as blocks in matrix where the area corresponds to the absolute frequency of the category, defined by co-occurrence of two or more events.
R: mosaicplot



Cross tabulation / contingency tables

Note that we have two nominal variables. A common way of displaying such data is a (two way) cross tabulation a.k.a. contingency table:

Example: Skin color of accused and death sentence

	black skin	white skin	Σ
Death sentence	59	72	131
No death sentence	2448	2185	4633
Σ	2507	2257	4764
Proportion in %	2.4	3.3	2.8

Cross tabulation / contingency tables

Caveat! Interpret conditional probabilities (frequencies) with care! Let's now also look at the skin-color of the *victim* and construct a three way cross tab:

Example: Skin color and death sentence

Skin-color of victim	black		white	
Skin-color of accused	b	w	b	w
Death sentence	11	0	48	72
No death sentence	2209	111	239	2074
Sum	2220	111	287	2146
Proportion in %	0.5	0.0	20.1	3.5

Simpson's Paradox

Describing ordinal and metric data

- Order statistics and ranks
  examples:
- Moments and moment estimators
  examples:
- Comparing to a distribution
  what distributions could be of interest?

Rank and Order Statistics

Ranks

In the increasingly ordered sample, the *rank* is the position of the observation when starting to count from the smallest one. R. rank(x)

Order statistics

Special values of the ordered sample or values derived from them (sometimes indexed by their rank).

Most important order statistics: minimum (rank=1), maximum (rank=sample size); range=(maximum-minimum); quantiles; percentiles

Percentiles/Quantiles

Quantiles and *percentiles* are important measures for distribution of the data. They are calculated based on the ordered sample $x_{(1)} = min, x_{(2)}, \ldots, x_{(n-1)}, x_{(n)} = max$.

$$\text{percentile of value } x = 100 \times \frac{\text{number of values less than } x}{\text{total number of values}}$$

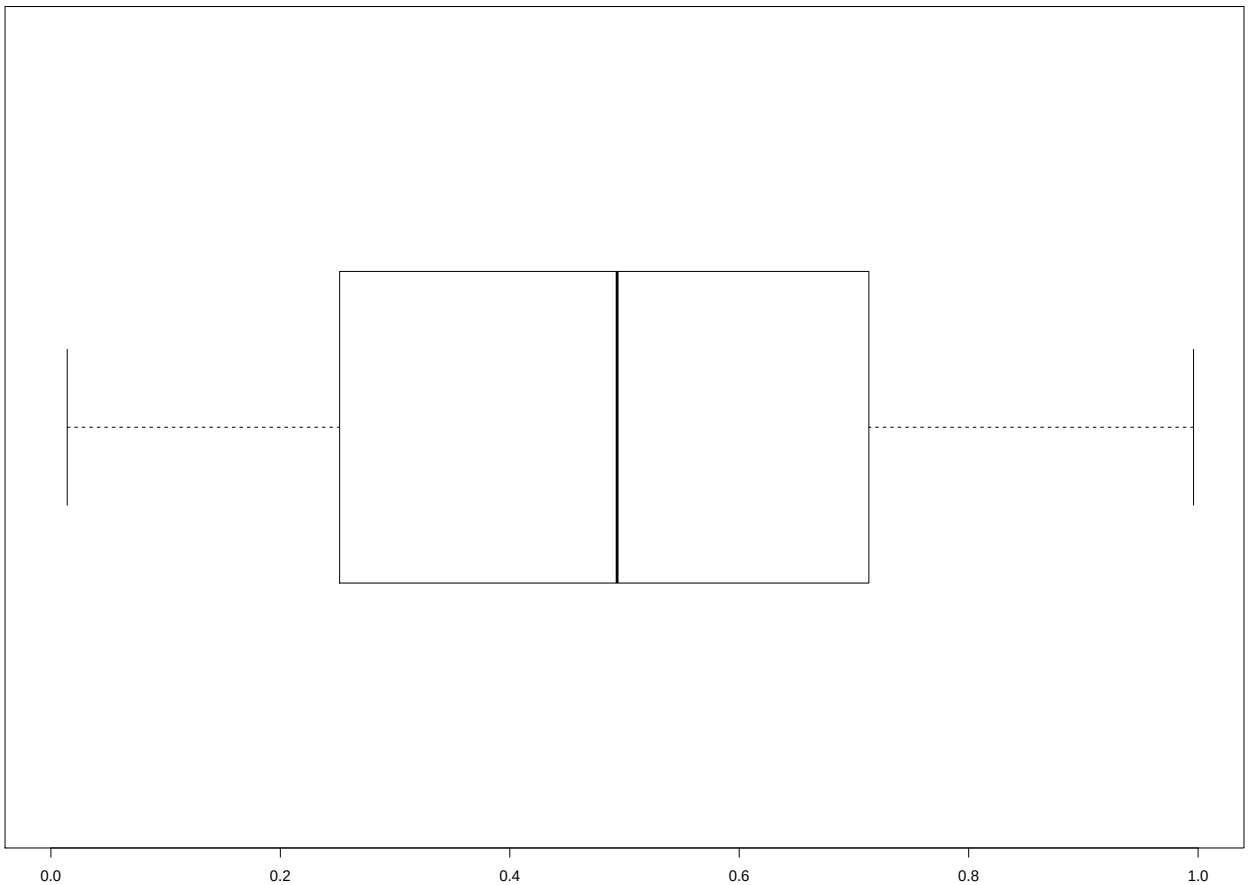
$$\text{quantile of value } x = \frac{\text{number of values less than } x}{\text{total number of values}}$$

- The best known quantile is the *median*, which is the 50%-quantile.
- The *quartiles* refer to the 25% and 75% which in addition to the median spread the data in quarters. A robust measure for variation is based on the quartile, the inter-quartile range.

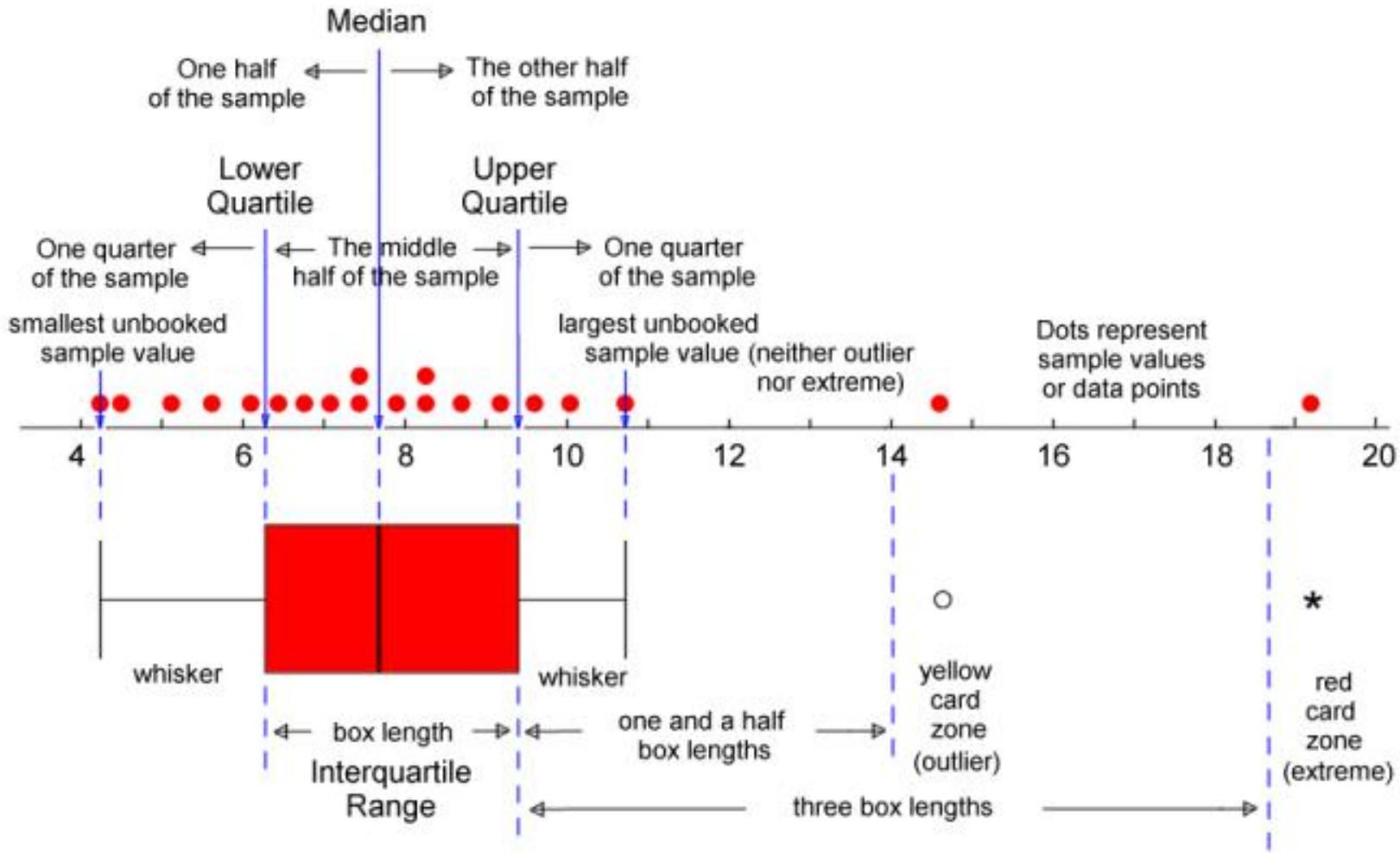
$$IQR = x_{0.75} - x_{0.25}$$

Important Quantiles = Five Number Summary

Minimum	1st Quartile	Median	3rd Quartile	Maximum
x_0	$x_{0.25}$	$x_{0.5}$	$x_{0.75}$	x_1
0.0142	0.2565	0.4936	0.7110	0.9960



(Extended) Boxplot



Quantiles: formal definitions

Quantiles

Quantiles are the minimal value such that the area under the density curve integrates to the given probability.

$$\begin{aligned} \int_{-\infty}^{x_{(p*100\%)}} f(x)dx &= p \\ F(x_{(p*100\%)}) &= p \end{aligned}$$

R: quantile(x,probability)

Moments of a distribution

Moments

Moments of a distribution describe important properties and allow to characterise and define distribution functions.
The **m-th moment** of a random variable X is

$$\begin{aligned} \mathbb{E}[X^m] &= \sum_{n=1}^{\infty} \mathbb{P}[X = x_n] x_n^m \quad \text{(discrete case)} \\ &= \int_{-\infty}^{\infty} f(x) x^m dx \quad \text{(continuous case)} \end{aligned}$$

The **m-th central moment** of a random variable X is

$$\mathbb{E}[(X - \mathbb{E}[X])^m]$$

Empirical Moments of a distribution

Important empirical Moments

Expectation	$\bar{x}$	$= \frac{1}{n}\sum_{i=1}^{n} x_i$
Variance	s_n^2	$= \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$
Skewness	$skew$	$= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$
(Excess) Kurtosis	$kurt$	$= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{2}} - 3$

Location

(More or less) important measures of location

	Sample	R
(Arithmetic) mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	mean
weighted mean	$\bar{x} = \sum_{i=1}^{n} w_i x_i$	weighted.mean
trimmed mean	$\bar{x} = \frac{1}{n}\sum_{i=q_{trim}}^{q_{1-trim}} x_{(i)}$	mean(,trim=p)
Geometric mean	$\sqrt[n]{\prod_{i=1}^{n} x_i}$	-

Location

(More or less) important measures of location

	Sample	R
Harmonic mean	$\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$	-
Median	middle value of ordered data	median
Mode	value with the greatest frequency	mode
Midrange	$\frac{\max x_i + \min x_i}{2}$	-

Location

Which location measure do we use in which situation?

- A biologist wants to calculate the average growth rate of a bacteria culture on different plates sequentially which showed the following grath rates, 3%, 2%, 1.2%, 0.3%, 0.9%, 1.6%.
- What is the average BMI of 5 persons (22.21,19.45,23.65,20.40,21.37)?
- A company wants to know about its average losses due to their employees' sickness leaves.
- An employees argues with his personnel manager about his average days on leave (due to sickness, vacation, maternity leave, etc.).
- Among several candidates the on average most influential gene driving a certain disease is searched based on some numeric score.

Variation

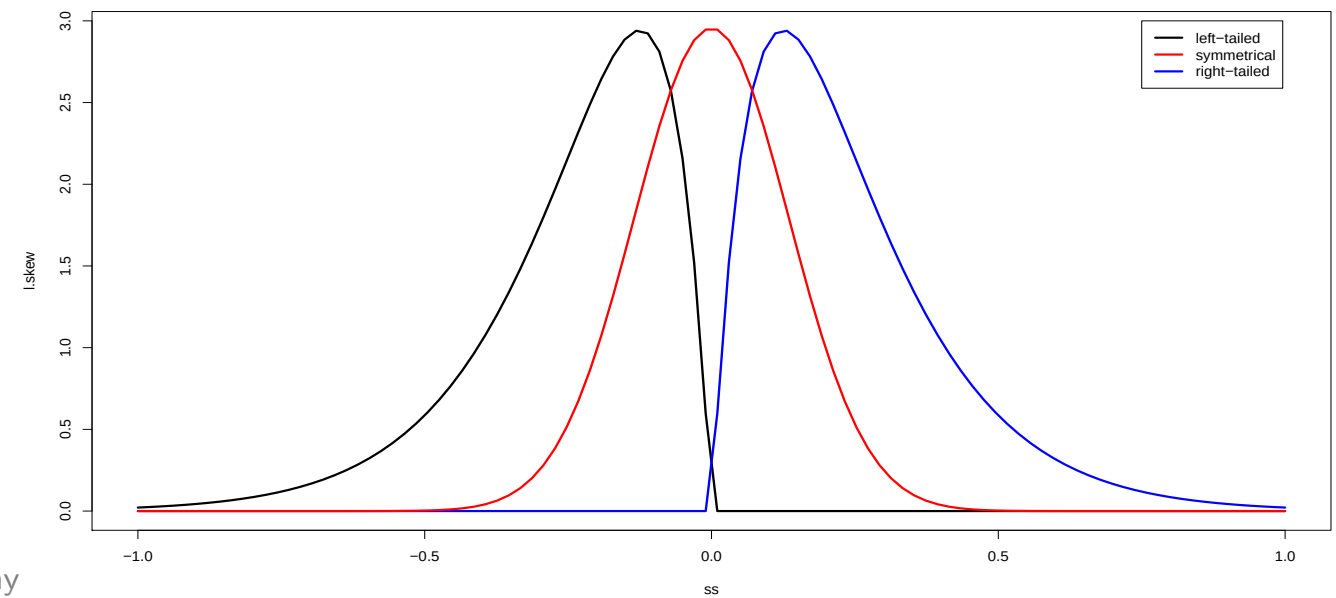
(More or less) important measures of variation

	Sample	Population	R
Range	$\max x_i - \min x_i$		range
Variance	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	var
Standard deviation	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$	sd
MAD (from the mean)	$\frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \bar{x}\rvert$	$\frac{1}{N}\sum_{i=1}^{N}\lvert x_i - \bar{x}\rvert$	mad
MAD from the median	$\frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \operatorname{med} x_i\rvert$	$\frac{1}{N}\sum_{i=1}^{N}\lvert x_i - \operatorname{med} x_i\rvert$	-
MedMed	$\operatorname{med}\lvert x_i - \operatorname{med} x_i\rvert$		-
IQR	$Q_{0.75} - Q_{0.25} =$ Box length		IQR
Coefficient of variation	$\mathrm{CV} = \frac{s}{\bar{x}}$	$\mathrm{CV} = \frac{\sigma}{\mu}$	-

Symmetry and skewness

$$\text{Skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

$$\text{skewness coefficient} = \frac{(x_{0.75} - x_{0.5}) - (x_{0.5} - x_{0.25})}{(x_{0.75} - x_{0.25})},$$

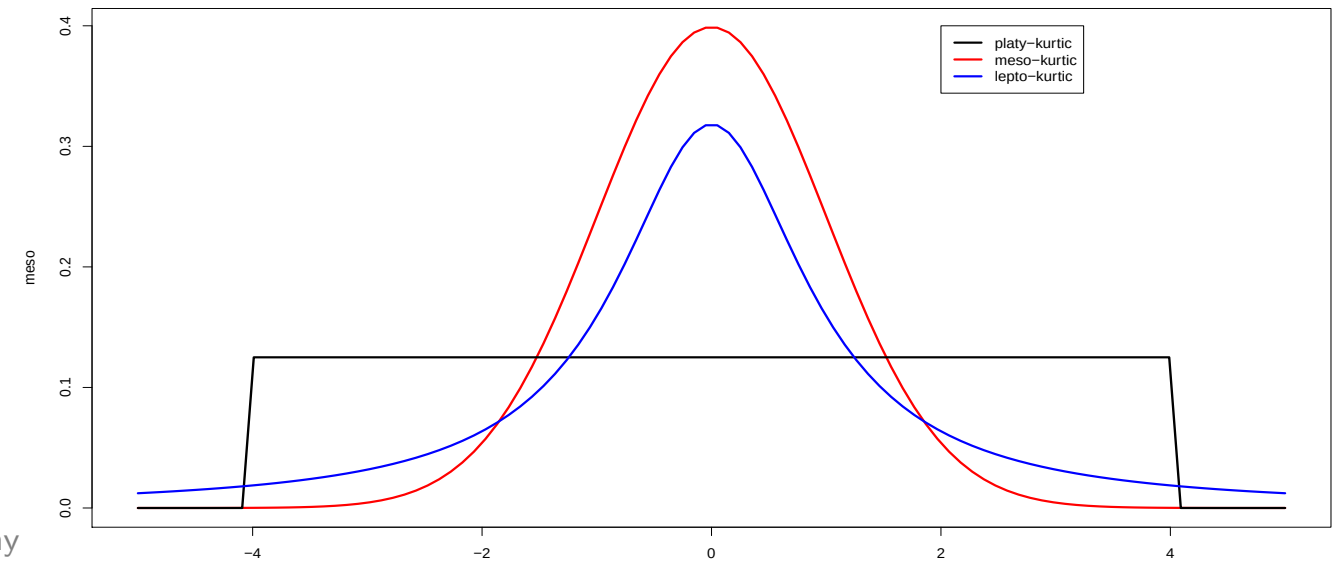


"Peakedness" and heavy tails

$$\text{(Excess) Kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$

$$\text{peakedness coefficient} = \frac{QD}{s}$$

$$\text{peakedness T} = \frac{x_{0.875} - x_{0.125}}{x_{0.75} - x_{0.25}}$$



Example

Summarising 10 measurements

Data: 1.234, 0.341, 89.14, 6.981, 2.152, 1.921, 7.312, 4.983, 0.118, -0.015

- Calculate and compare all (possibly applicable) location estimates from the table above
- Calculate and all (possibly applicable) variation estimates from the table above
- Calculate and compare all (possibly applicable) skewness estimates from above
- Calculate and compare all (possibly applicable) peakedness/tail weight estimates from above

Do the data have large/small variablility?
Are the data symmetric?
Are there any outliers?

Characteristics of data

1. **Center.** The location of the middle of the data illustrated by a representative or average value.
2. **Variation.** A measure of the variability of measurements.
3. **Distribution.** How the spread of the data is shaped and behaves. ((a-)symmetry and "peakedness")
4. **Outliers.** Values of the sample showing a behaviour which differs from the rest of the sample
5. **Time.** Changing characteristics of the data over time.

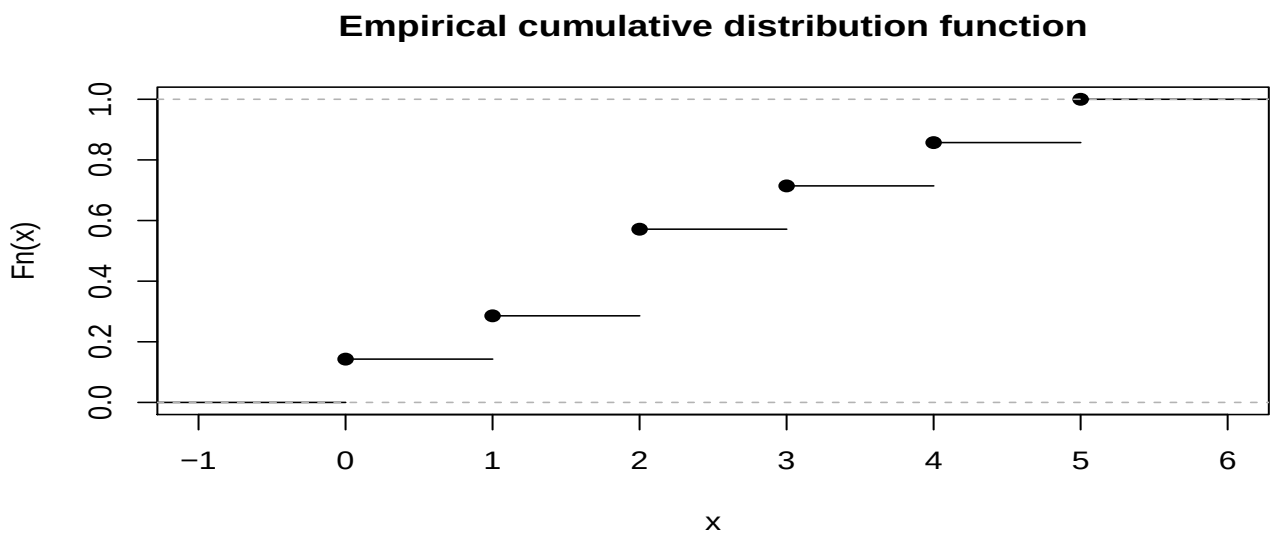
ECDF

Empirical cumulative distribution function

Represents the distribution, as defined by the sample without any assumptions

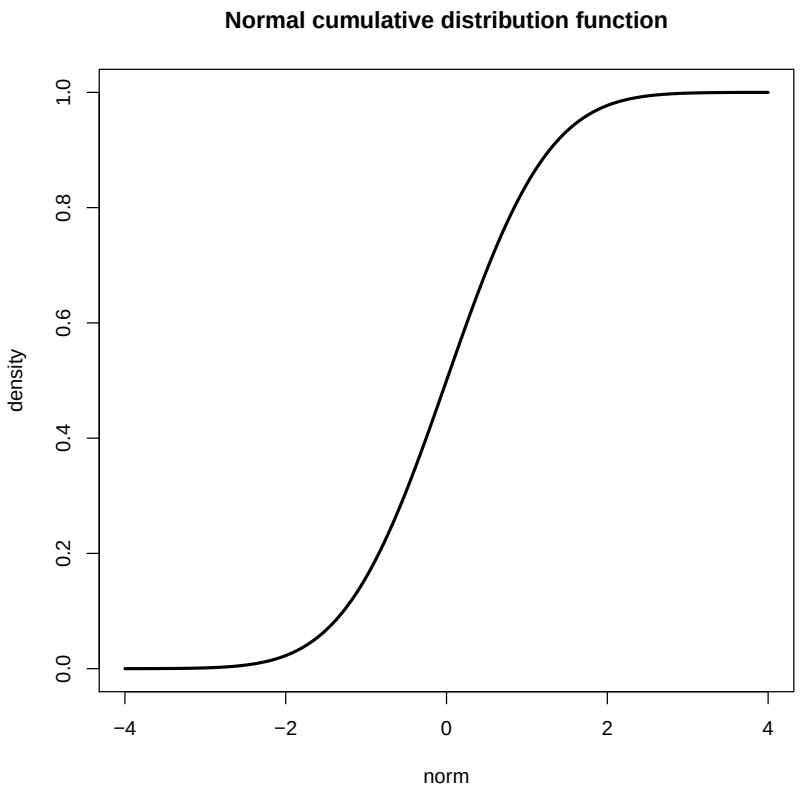
$$\hat{F}_n(x) = \frac{\text{number of observations in the sample} \leq x}{\text{sample size}}$$

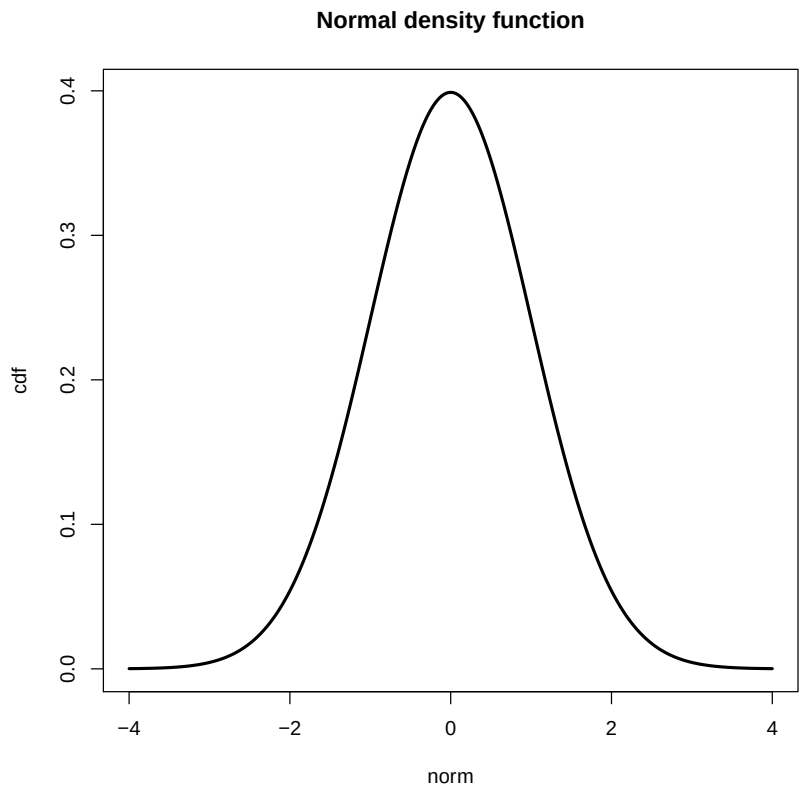
R: ecdf(x)



Functions describing a distribution

- density function
- cumulative distribution function





Stem-and-leaf plot

Stem-and-leaf plot

visualisation of the density by frequency of last digits
R: stem(x)

```
The decimal point is at the |
  -8 |  494
  -6 |  931
  -4 |  998828765531
  -2 |  7773766664433
  -0 |  9533216432
   0 |  2566896669
   2 |  00147824678
   4 |  01134672
   6 |  033445823335
   8 |  23722346
  10 |  0788012
  12 |  834
```

Histograms and Boxplots

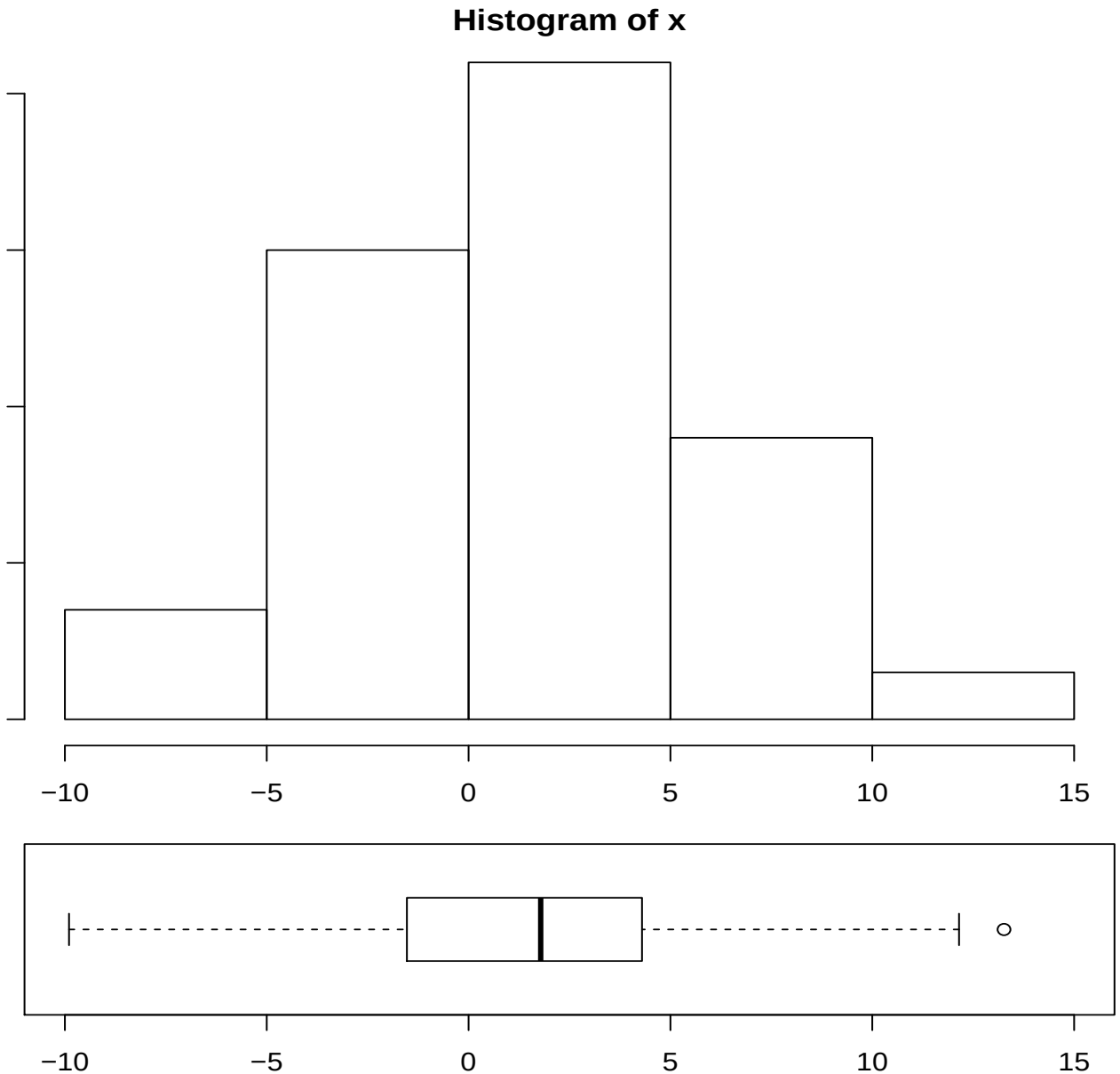
Histogram

A **histogram** is a graph consisting of bars of (by standard equal) width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to the frequency values. R: hist(x)

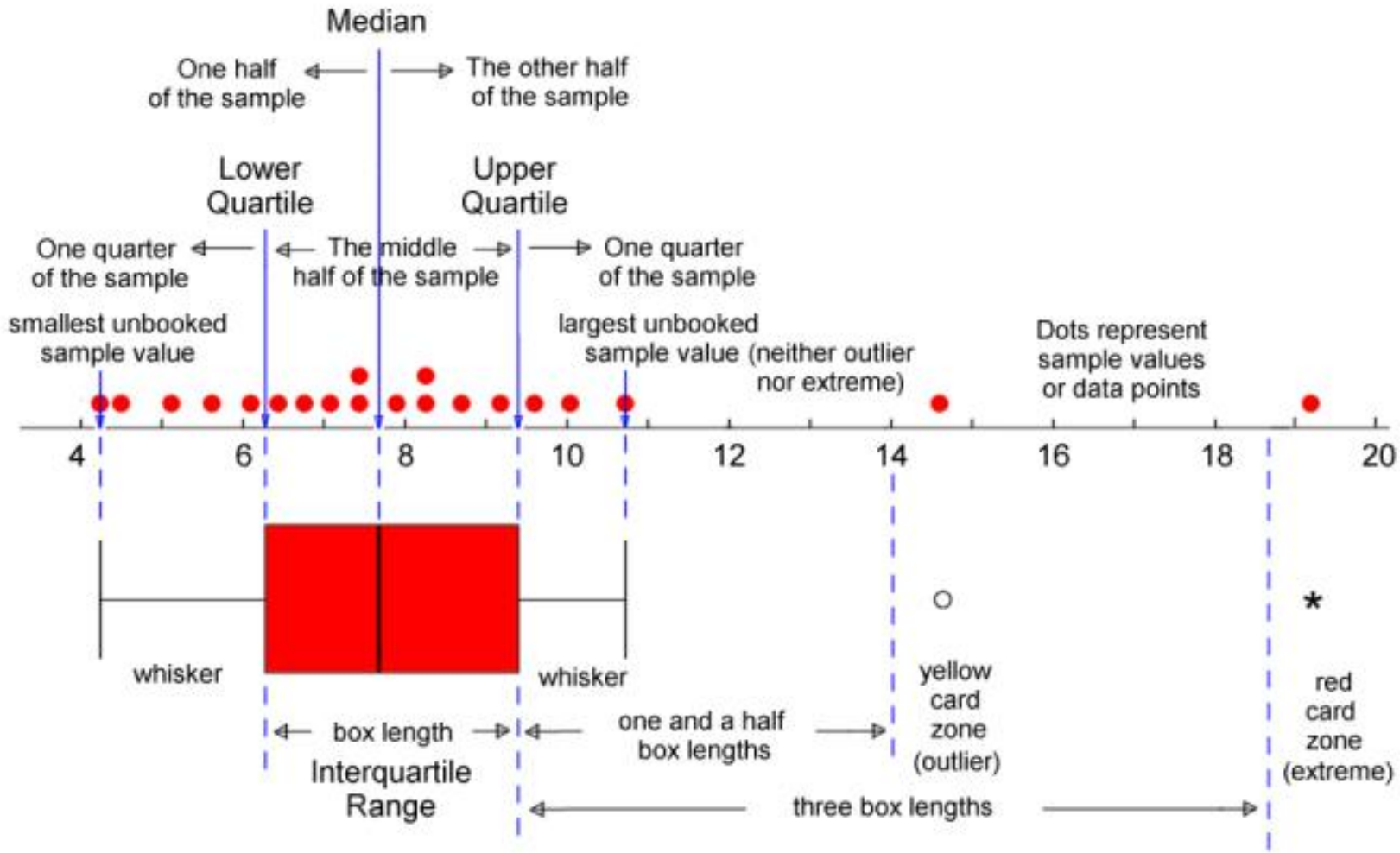
Boxplot

A **boxplot** is a graph summarising the data based on its 5-number statistics, i.e. min, 25%-quantile, Median, 75%-quantile,max. The hinges represent the spread, which is 1.5 times the inter-quartile range. R: boxplot(x)

Histogram and Boxplot



(Extended) Boxplot



Kernel density estimation

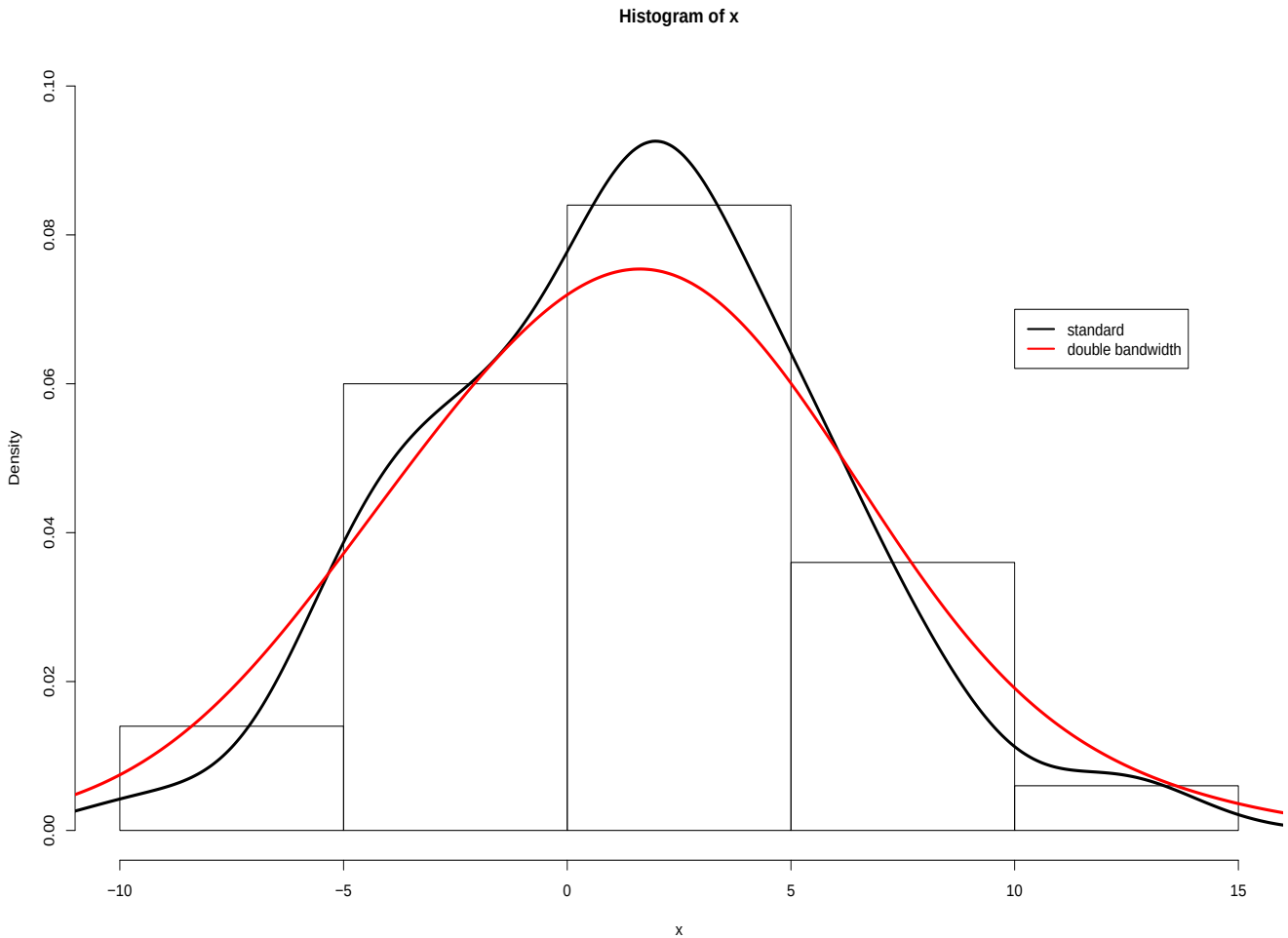
Kernel density estimator

provide a smooth estimate of the density ($\neq$ histogram is discretised) with a smoothing parameter h and kernel K_h

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

R: density(x)

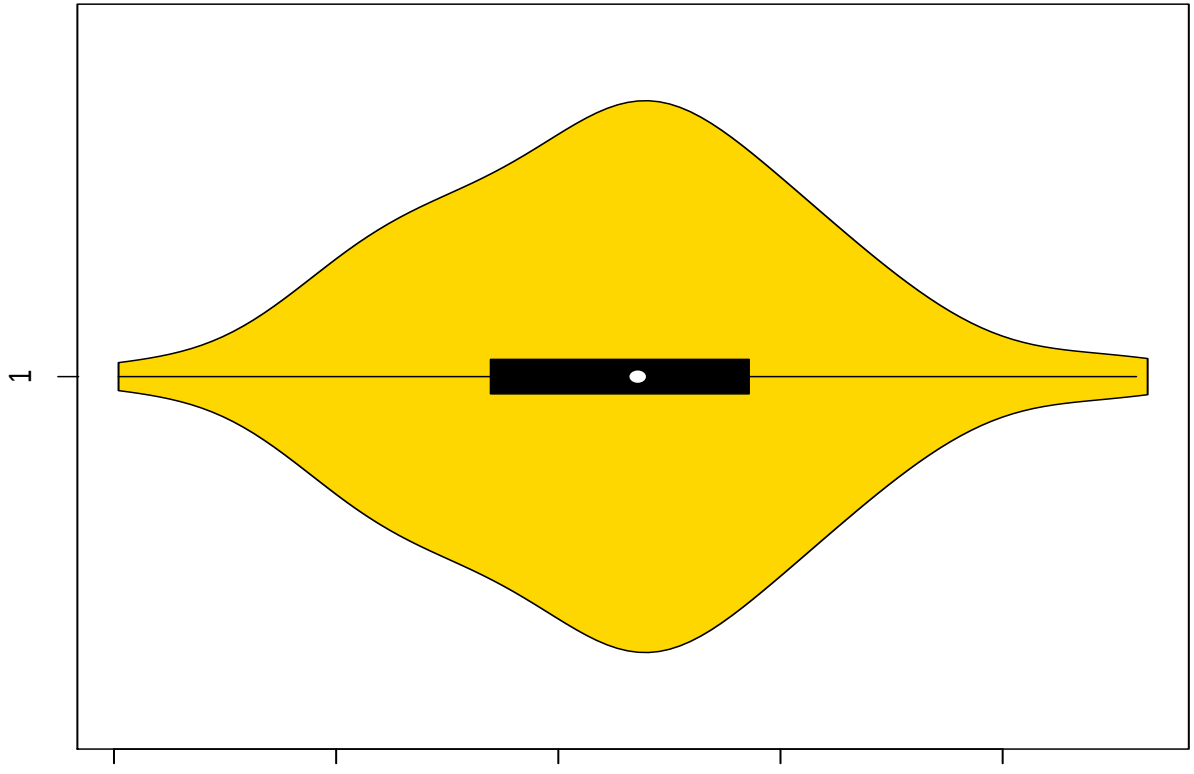
Kernel density estimation



Violin Plots

Violin Plot

combination of boxplot and kernel density estimation in a single plot
R: library(vioplot); vioplot(x)

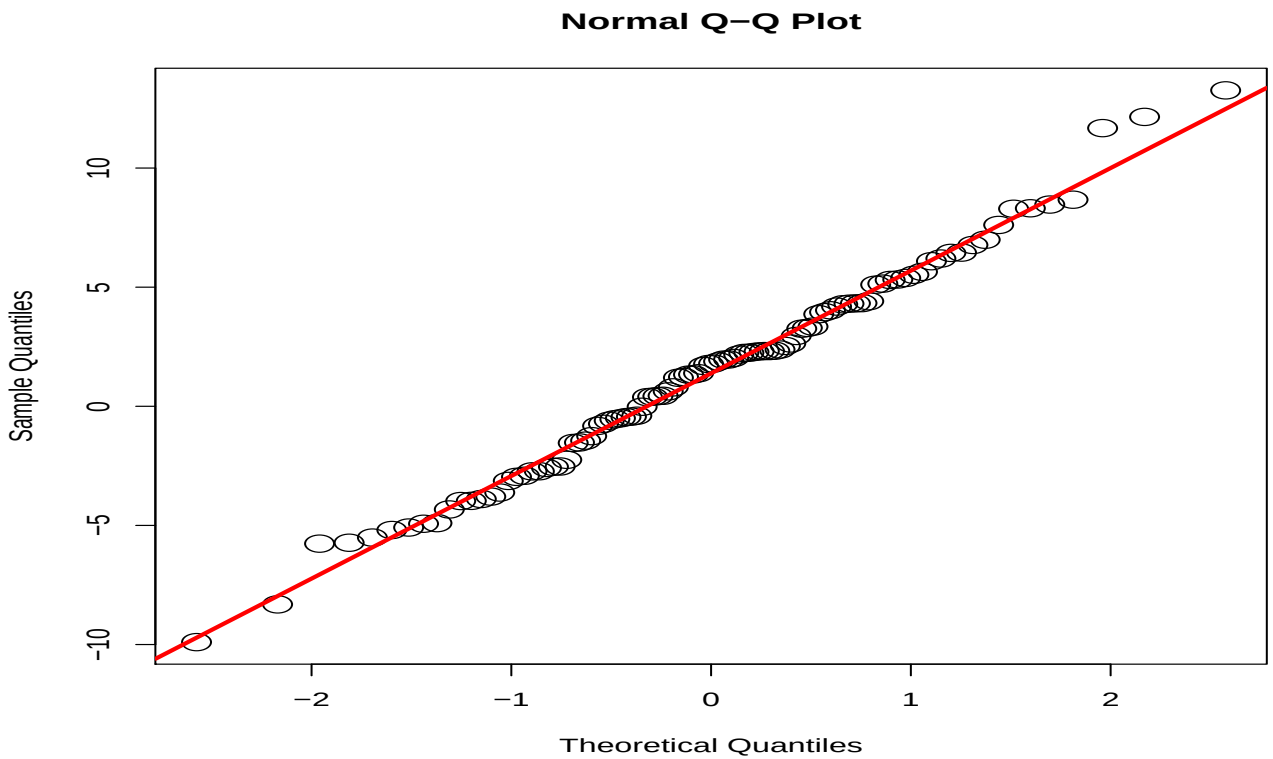


Quantile-Quantile (QQ) Plots

Quantile-Quantile Plot

compare quantiles of the data against quantiles of a theoretical distribution (most typical: normal distribution) or second sample data distribution

R: qqplot(x,y); qqnorm(x)



Normal (Gaussian) distribution

Normal distribution

The probability density of the normal distribution is defined as

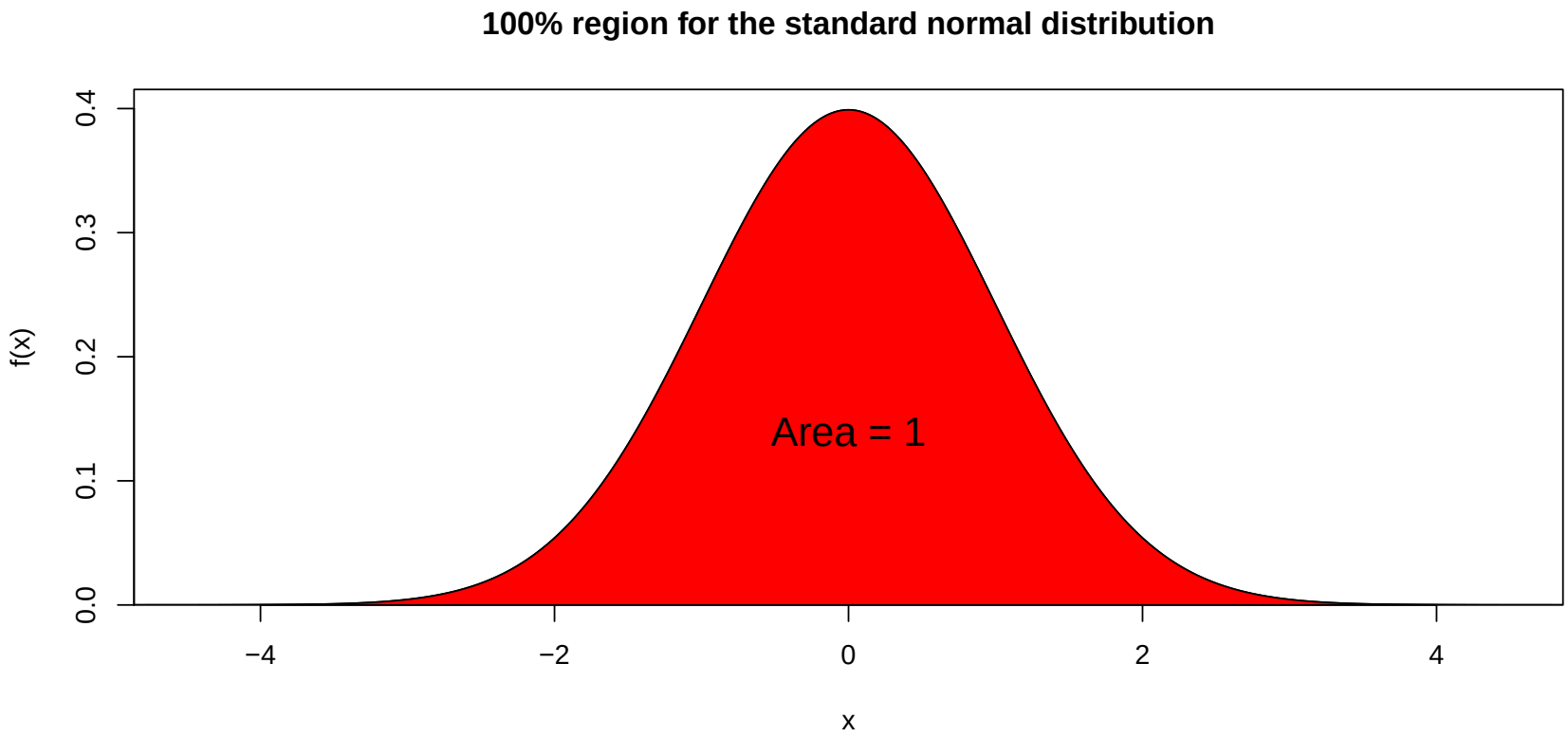
$$\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}.$$

The distribution has expected value and Variance

$$\begin{aligned} \mathbb{E}[X] &= \mu \\ \mathbb{V}[X] &= \sigma^2 \end{aligned}$$

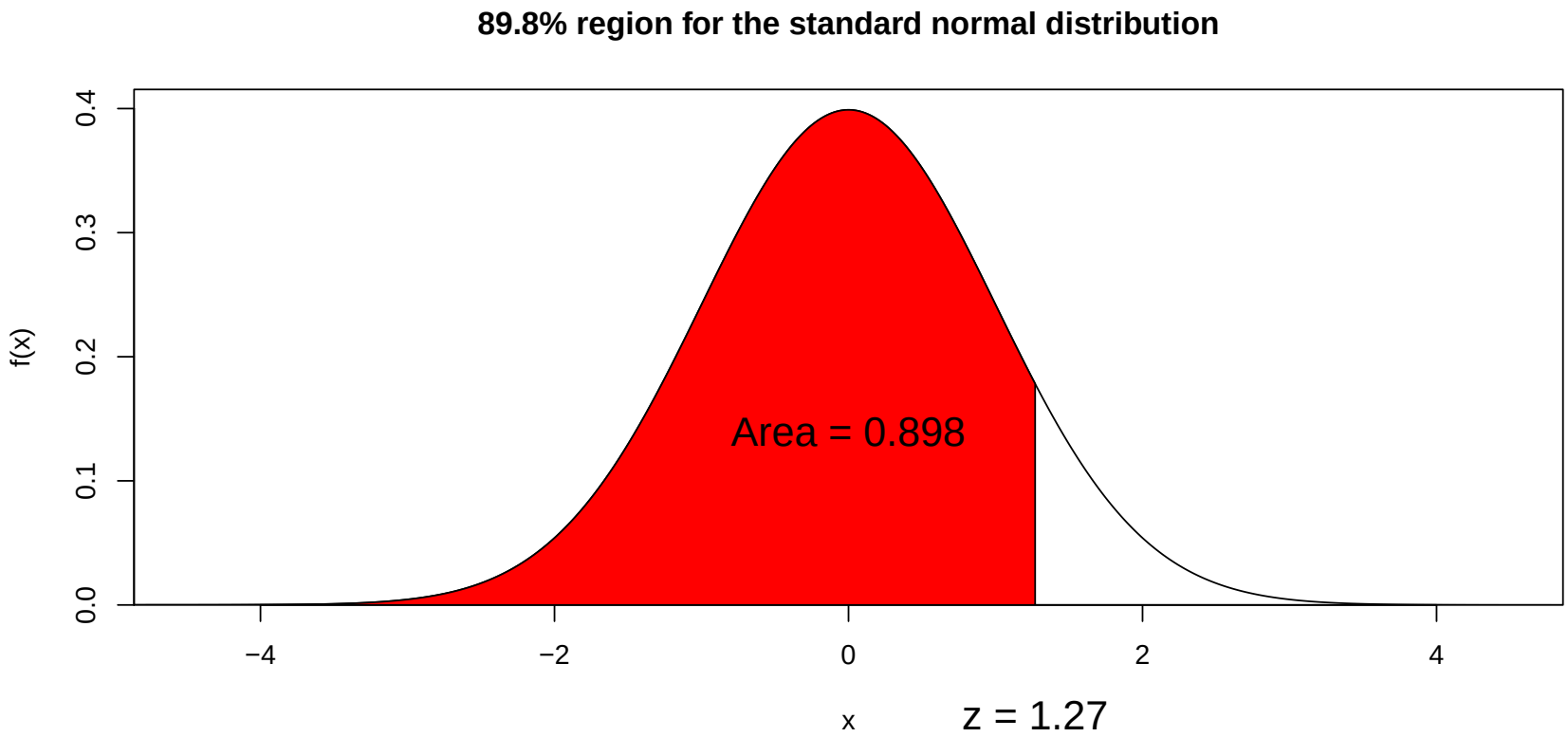
The standard normal distribution has mean $\mu = 0$ and variance $\sigma^2 = 1$.

Standard normal (Gaussian) distribution



$$\mathbb{P}[-\infty \leq X \leq \infty] = \int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx =$$
$$= F(\infty) - F(-\infty) = 1 - 0 = 1$$

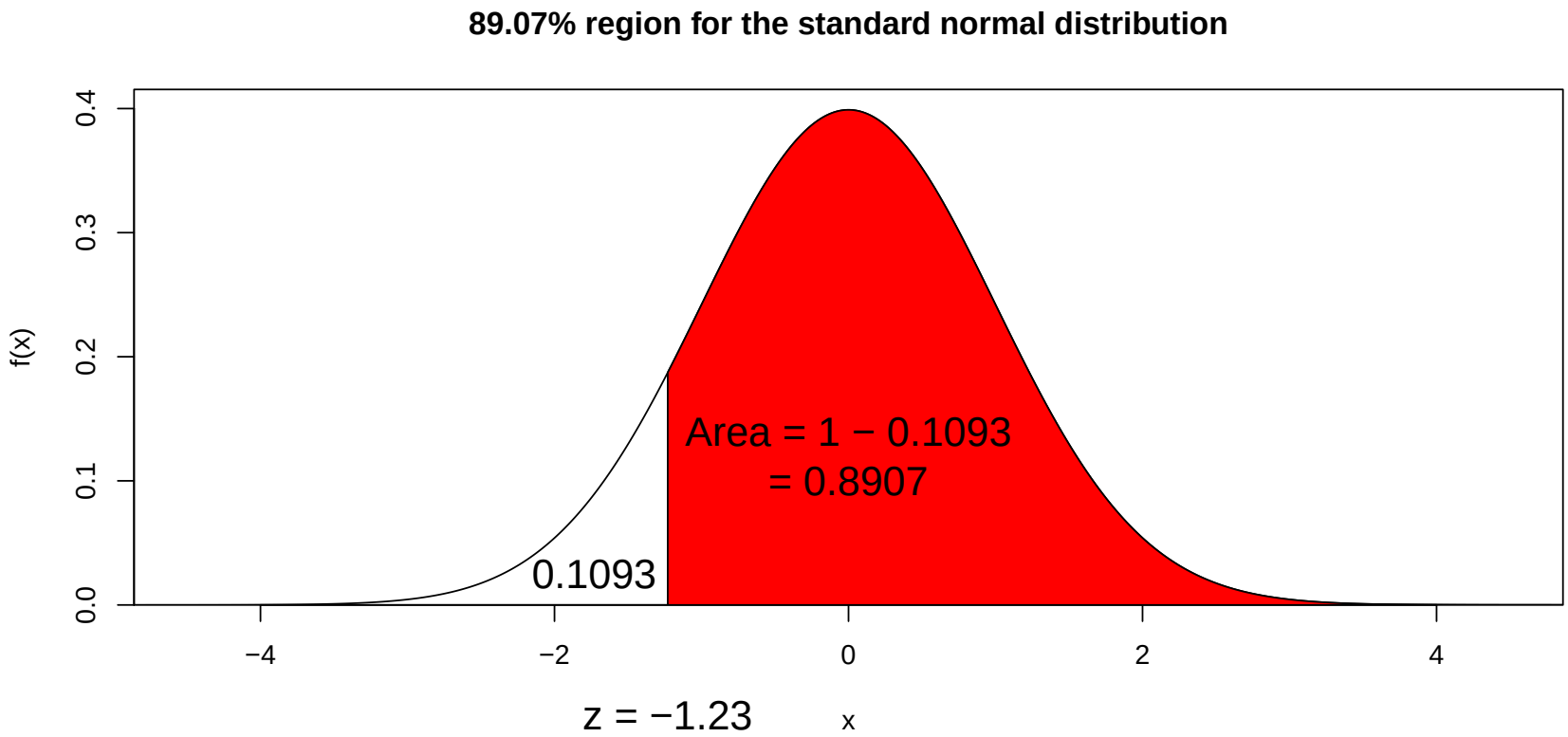
Standard normal (Gaussian) distribution



$$\mathbb{P}[-\infty \leq X \leq 1.27] = \int_{-\infty}^{1.27} f(x)dx = \int_{-\infty}^{1.27} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx =$$

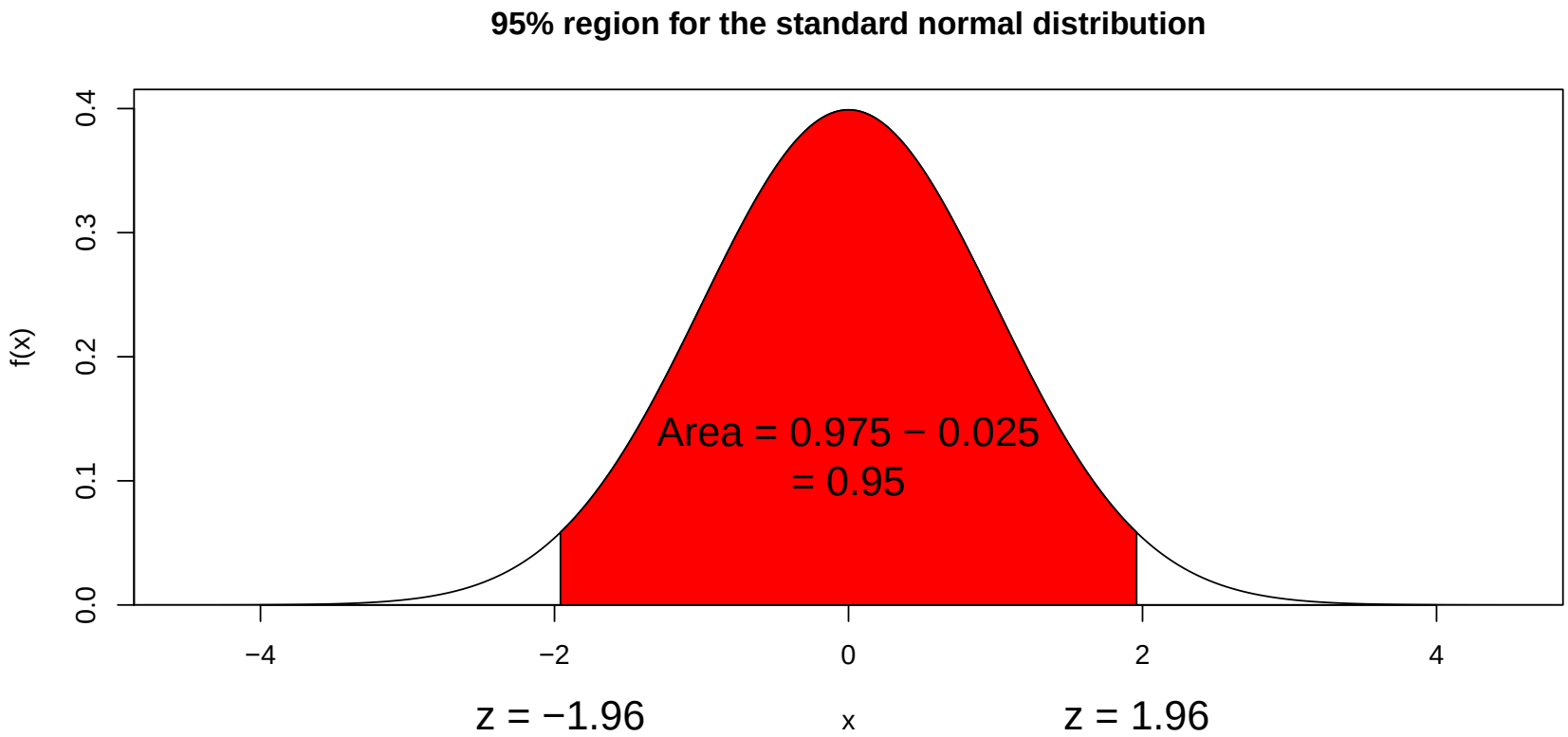
$$= F(1.27) - F(-\infty) = 0.8980 - 0 = 0.8980$$

Standard normal (Gaussian) distribution



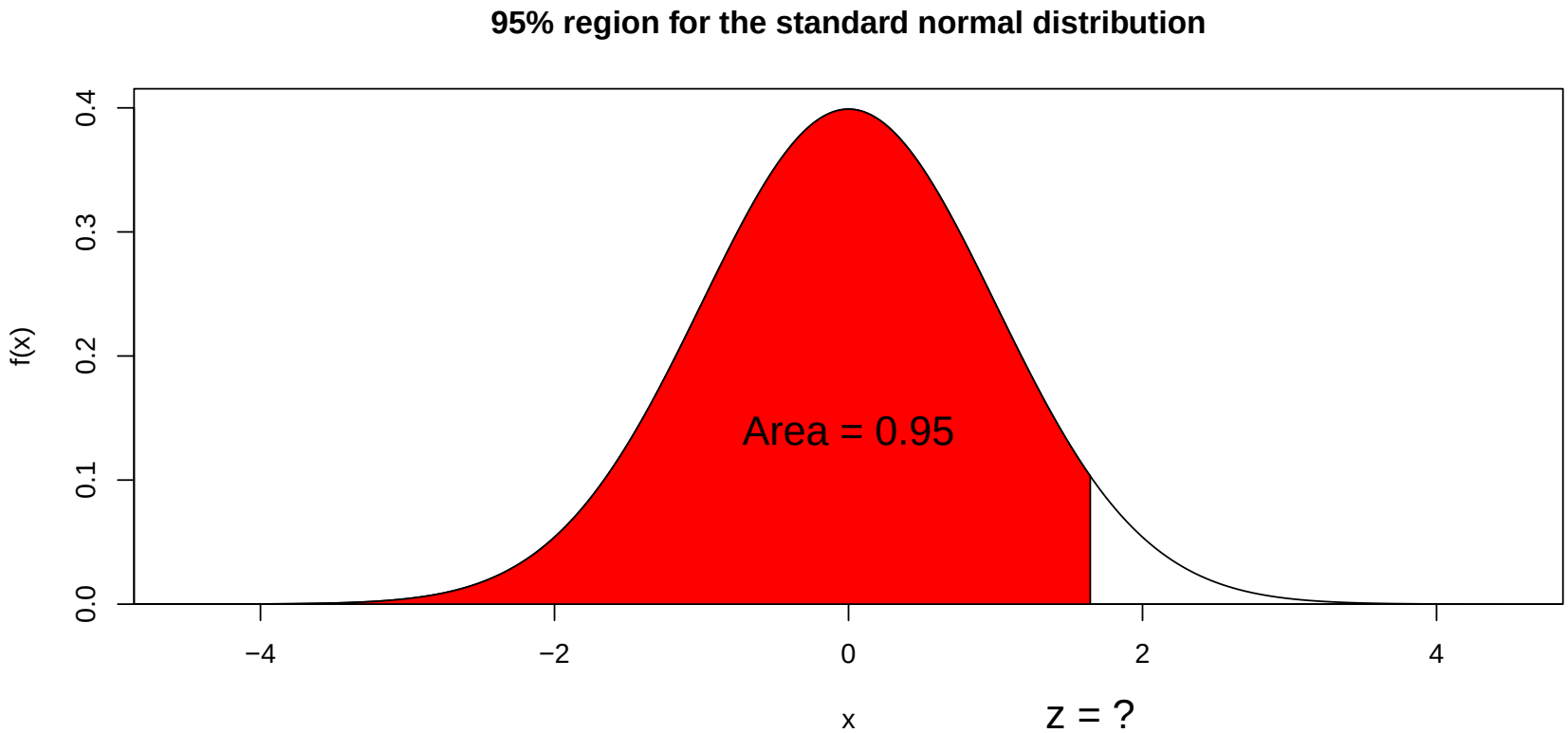
$$\mathbb{P}[-1.23 \leq X \leq \infty] = \int_{-1.23}^{\infty} f(x)dx = \int_{-1.23}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx =$$
$$= F(\infty) - F(-1.23) = 1 - 0.1093 = 0.8907$$

Standard normal (Gaussian) distribution



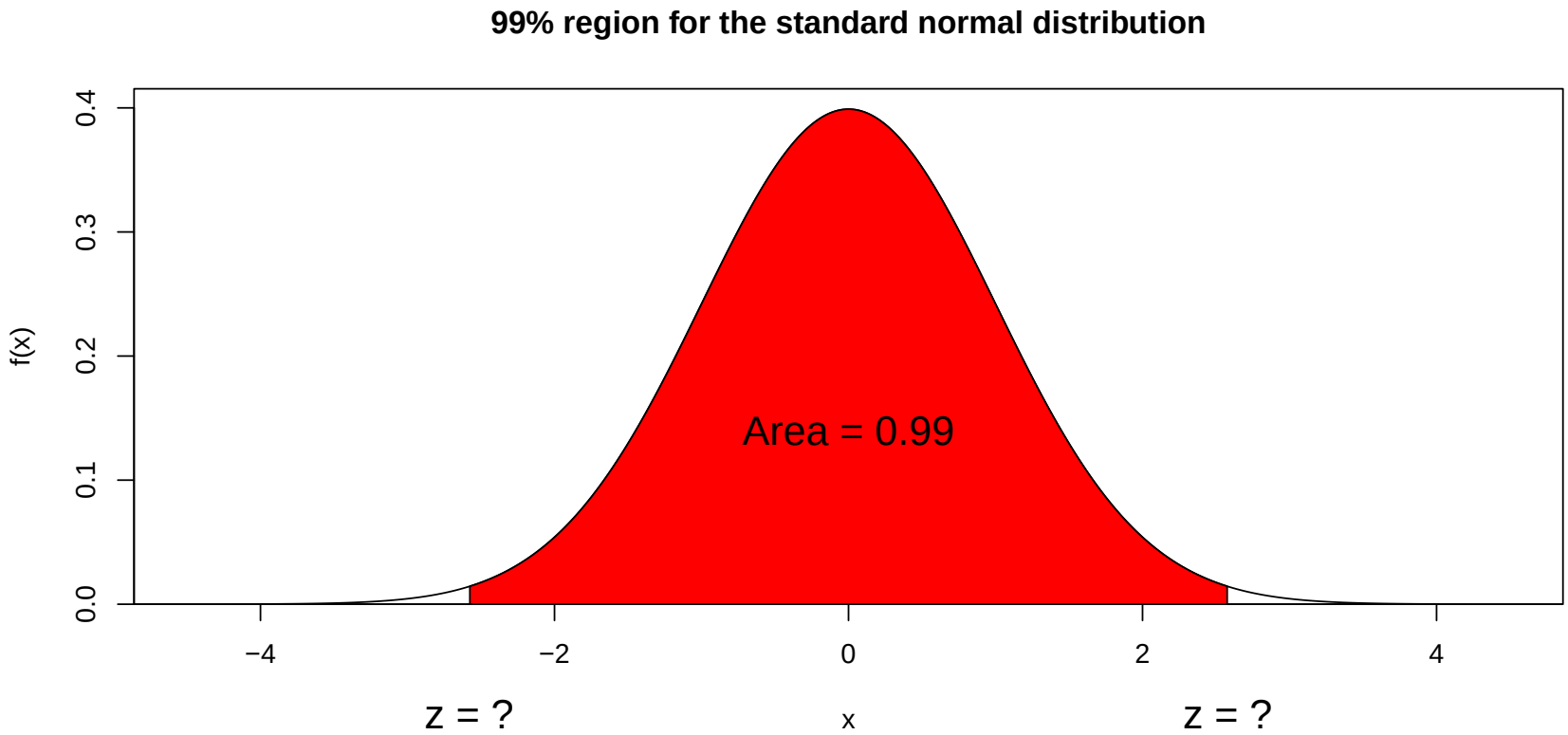
$$\mathbb{P}[-1.96 \leq X \leq 1.96] = \int_{-1.96}^{1.96} f(x)dx = \int_{-1.96}^{1.96} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx =$$
$$= F(1.96) - F(-1.96) = 0.975 - 0.025 = 0.95$$

Standard normal (Gaussian) distribution



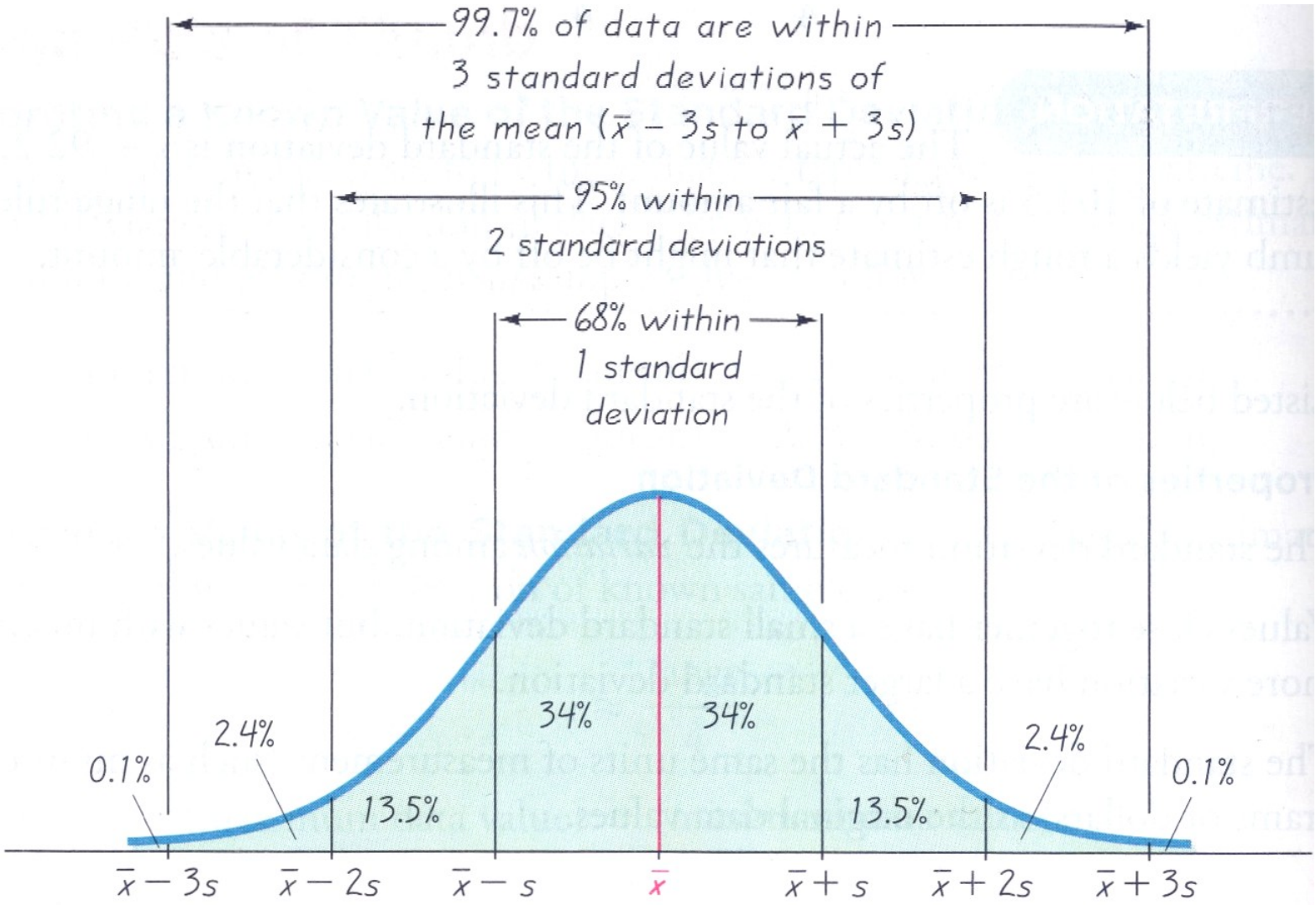
$$\mathbb{P}[-\infty \leq X \leq z] = 0.95 \Longrightarrow z = 1.645$$

Standard normal (Gaussian) distribution



$$\mathbb{P}[p_{0.5} \leq X \leq p_{99.5}] = 0.99 \Longrightarrow z = \pm 2.575$$

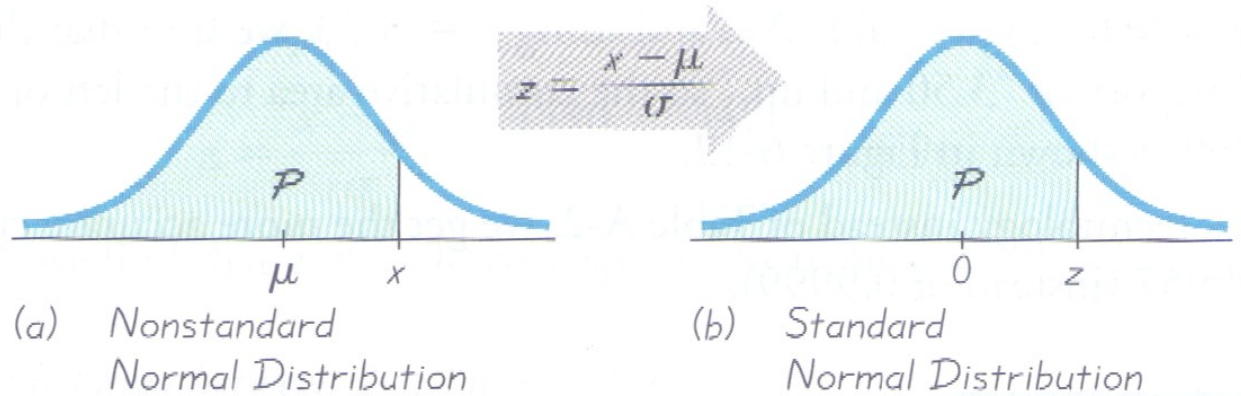
Empirical (or 68-95-99.7) rule for Gaussian data



Converting from a nonstandard to a standard normal distribution

z-Transformation

$$z = \frac{x - \mu}{\sigma}$$



(a) Nonstandard Normal Distribution

(b) Standard Normal Distribution

Example: Finding values from known areas

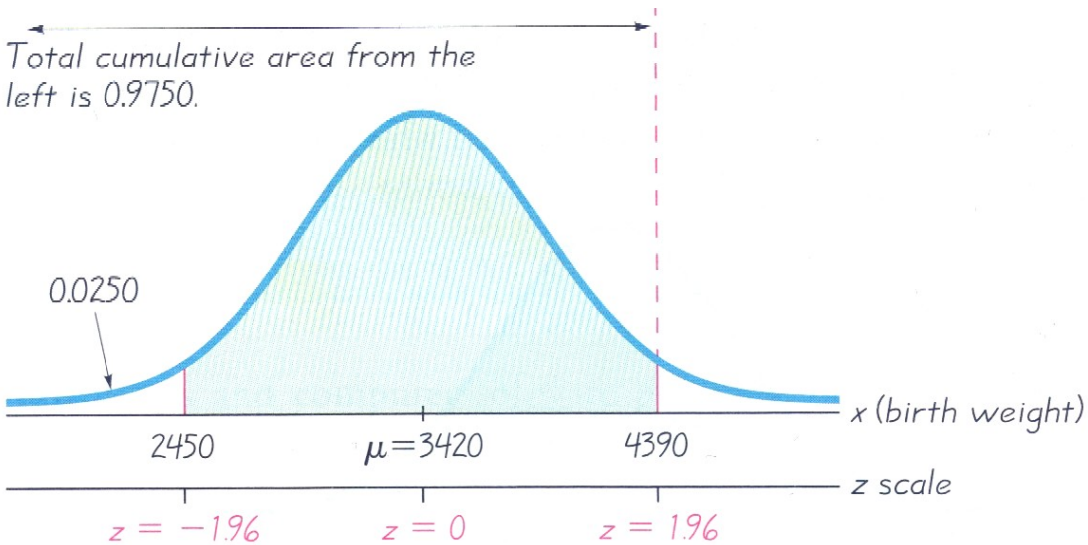
Birth weights

Birth weights in the US are normally distributed with mean 3420g and a standard deviation of 495g. The Newport General Hospital requires special treatment for babies that are less than 2450g (unusually light) or more than 4390g (unusually heavy).

Example: Finding values from known areas

$$z_{\text{low}} = \frac{2450 - 3420}{495} \approx -1.96$$

$$z_{\text{high}} = \frac{4390 - 3420}{495} \approx 1.96$$



Thus,

- 95% of the babies don't require special treatment
- 2.5% of the babies are considered to be too light
- 2.5% of the babies are considered to be too heavy

Finding values from known areas

Some hints:

- Don't confuse z scores and areas.
- Choose the correct (right/left) side of the graph.
- A z score must be *negative* whenever it is located in the *left* half of the normal distribution.
- Areas (or probabilities) are positive or zero values, but they are never negative.

Procedure:

- Sketch!
- Use table/technology to find the z score corresponding to the cumulative left area bounded by x.
- Use the (inverse) z-transformation to solve for $x = \mu + z\sigma$.
- Check for plausibility!

Example: Finding areas from known values

Birth weights

The Newport General Hospital wants to redefine the minimum and maximum birth weights that require special treatment because they are unusually low or unusually high. After considering relevant factors, a committee recommends special treatment for birth weights in the lowest 3% and highest 1%.

$$\begin{aligned} x_{\text{low}} &= \mu + z\sigma = 3420 + (-1.88) \times 495 = 2489.4 \\ x_{\text{high}} &= \mu + z\sigma = 3420 + 2.33 \times 495 = 4573.35 \end{aligned}$$

Thus,

- 2489g separates the lowest 3% of birth weights,
- 4573g separates the highest 1% of birth weights.

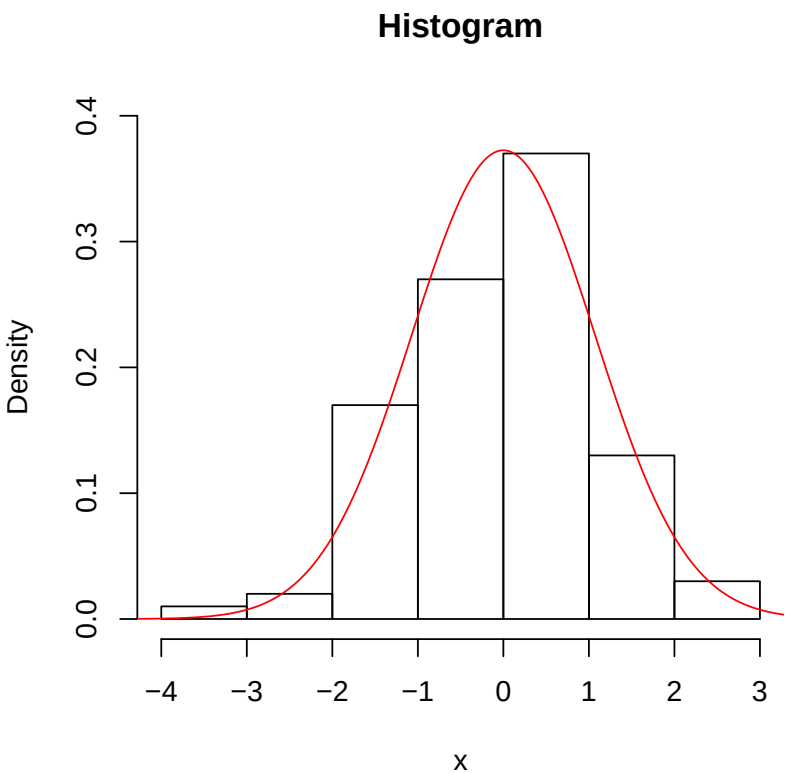
Assessing normality

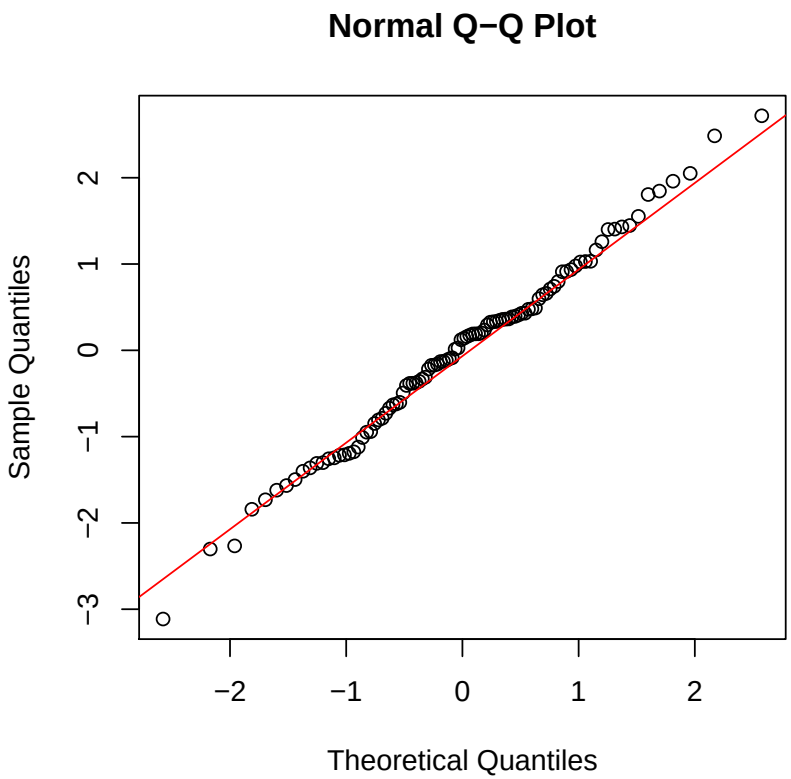
Procedure for determining whether it is reasonable to assume that sample data are from a normally distributed population:

1. Histogram: Construct a histogram. Reject normality if the histogram departs dramatically from a bell shape.
2. Outliers: Reject normality if there are obvious outliers present.
3. Normal quantile-quantile plot: Do the points lie reasonably close to a straight line or is there a *systematic pattern* that is not a straight-line pattern? (R: qqnorm(x) , qqplot(x))
4. Testing for normality: Shapiro-Wilk test; measures whether the shape of the sample distribution is reasonably close to the normal distribution (R: shapiro.test(x))

Example: Assessing normality

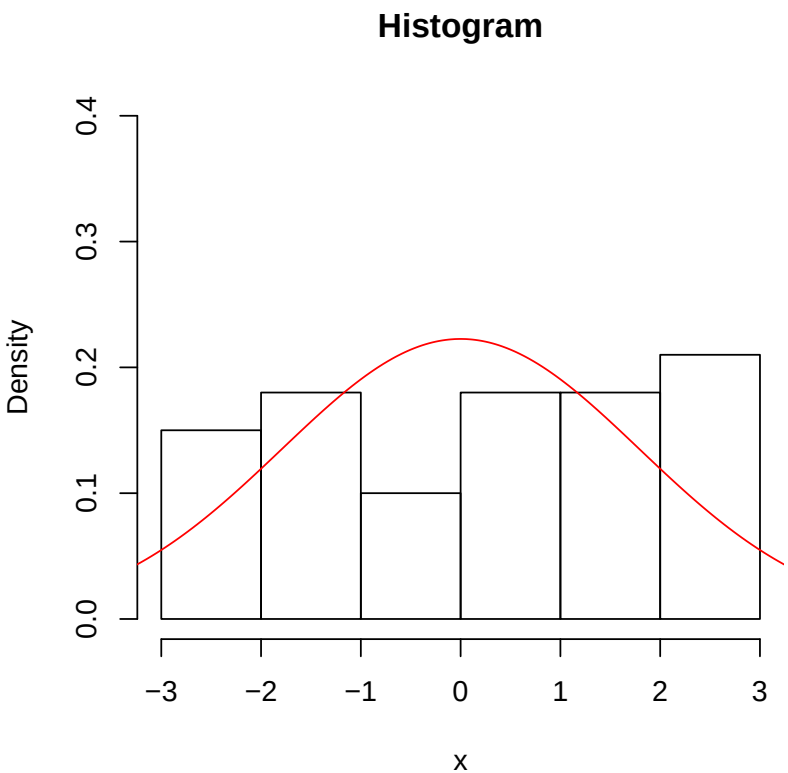
100 draws from a standard normal distribution:

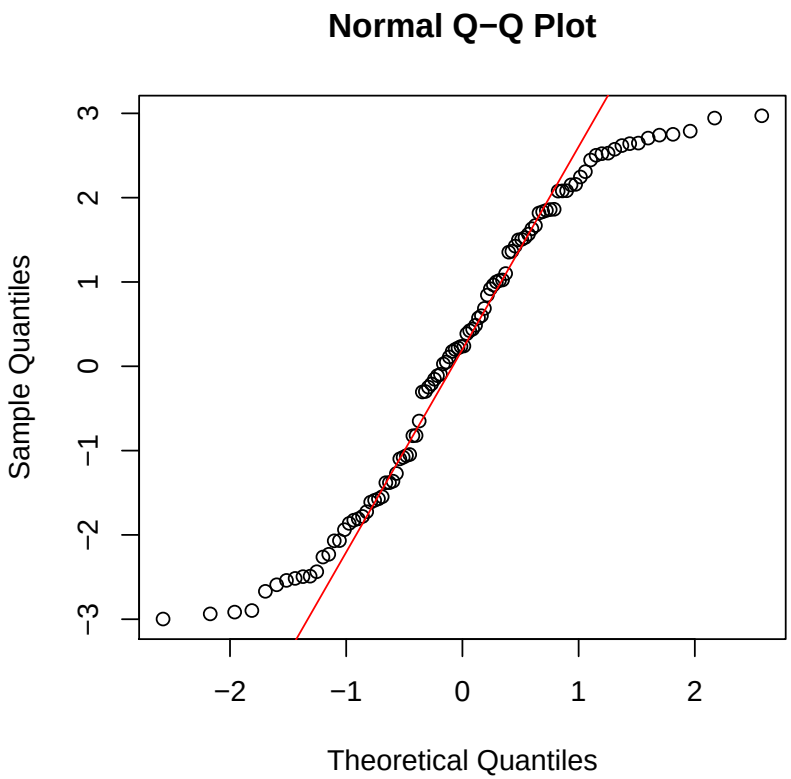




Example: Assessing normality

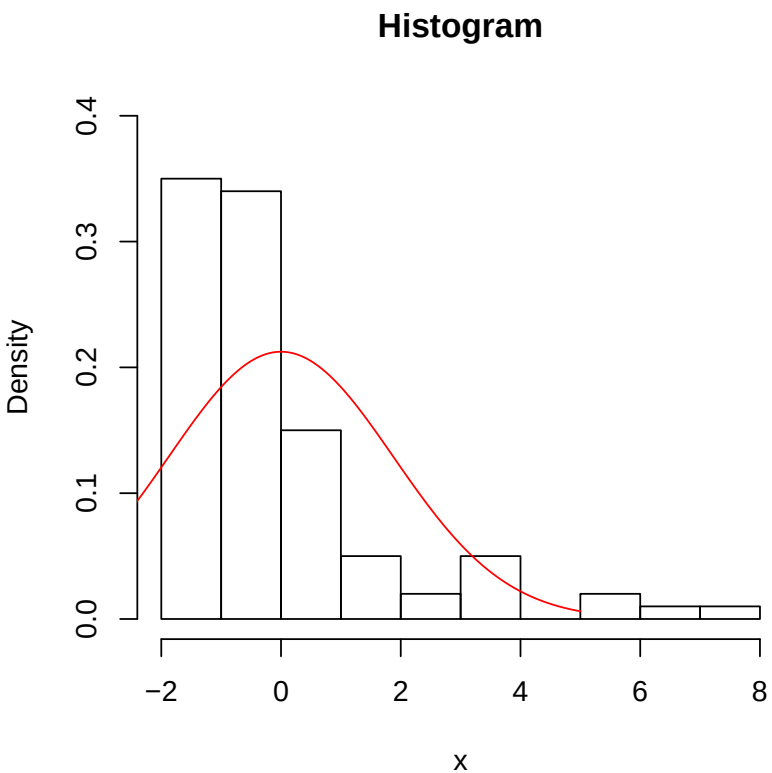
100 draws from a uniform distribution on [-3,3]:

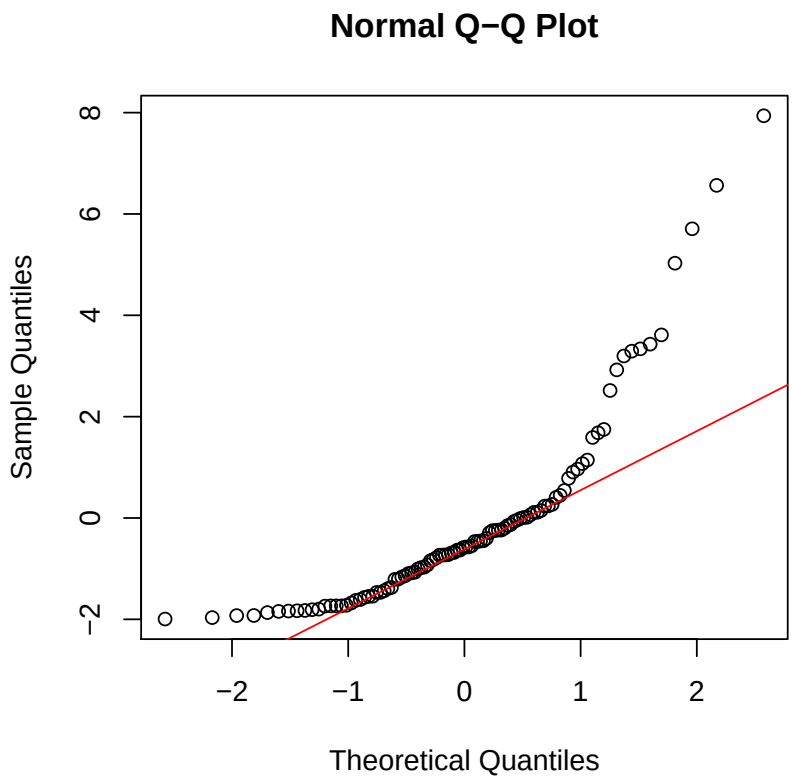




Example: Assessing normality

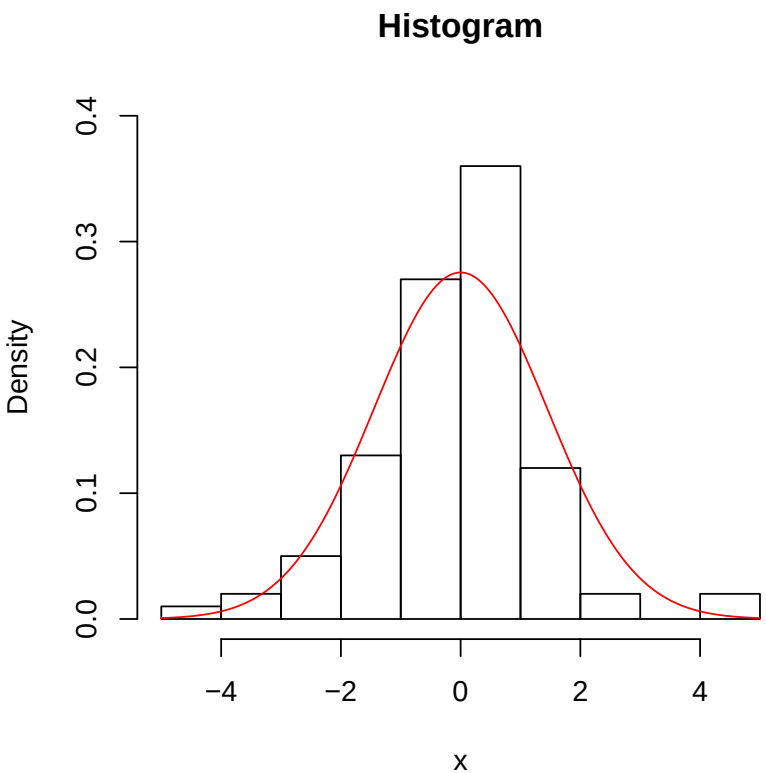
100 draws from an exponential distribution:

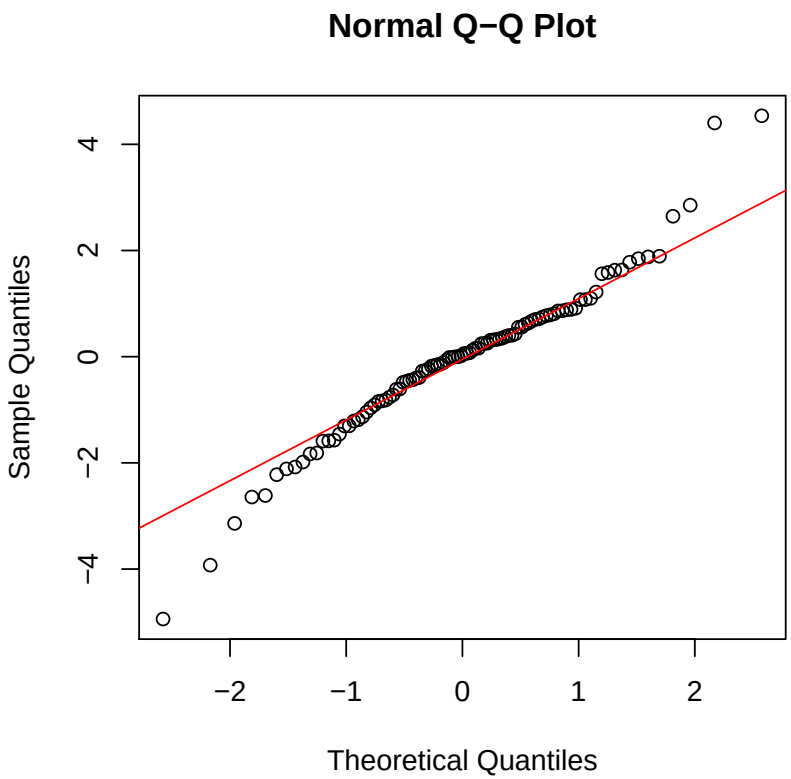




Example: Assessing normality

100 draws from a t distribution with 2 degrees of freedom:



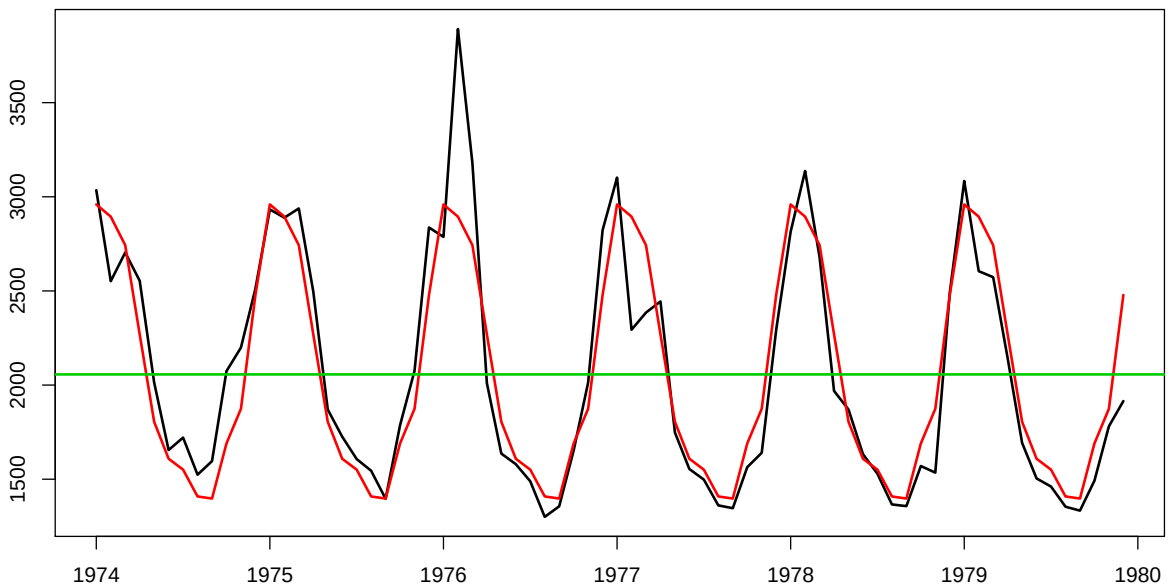


Time Series

Time Series

Time series always have *time* as *independent* variable. R: ts, ts.plot
Time series consist of:

- a trend component
- a cyclic (seasonal) component
- a random error.



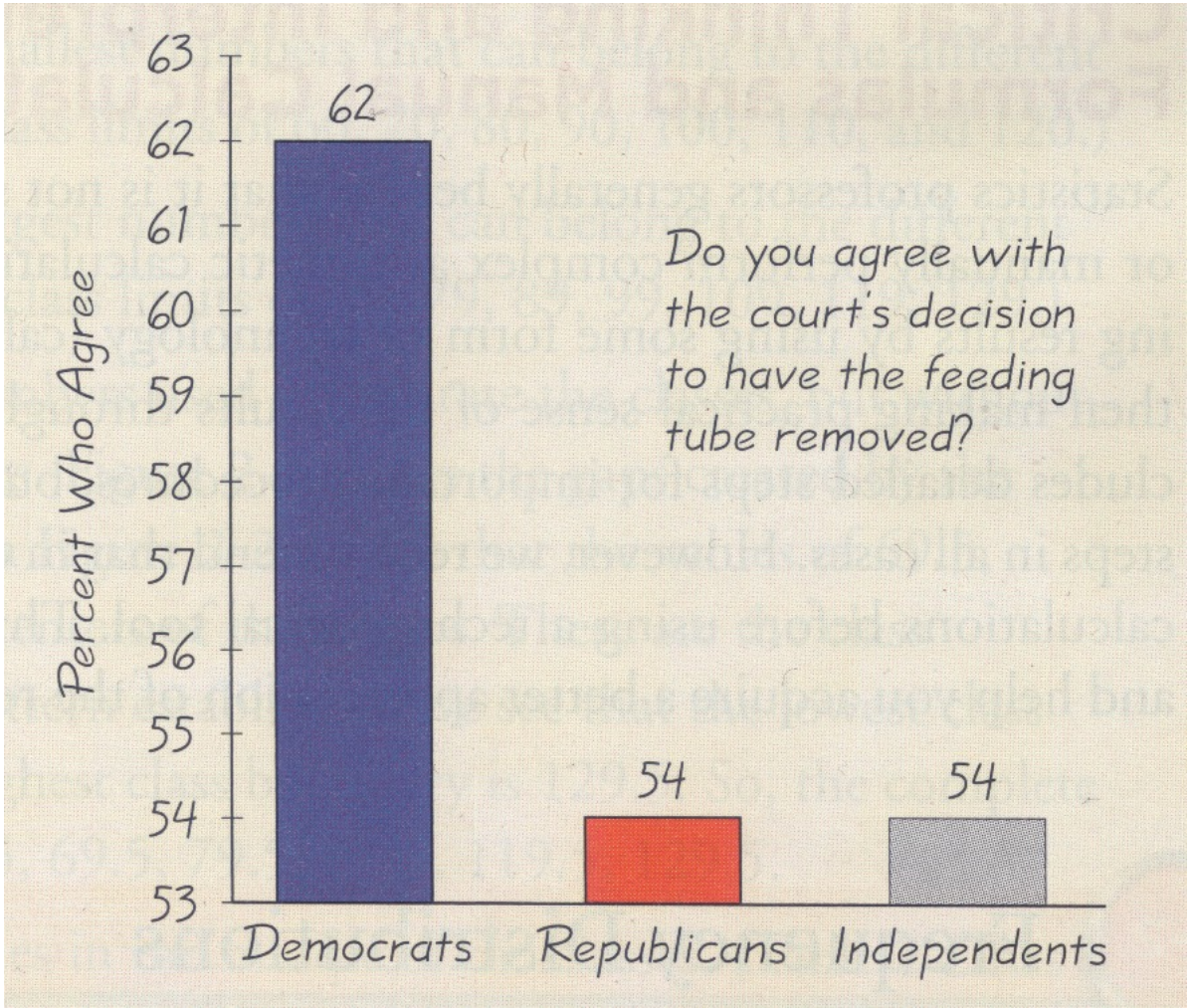
How not to present your data[3]

- Display as little information as possible
- Obscure what you do show (with chart junk)
- Use pseudo-3d and color gratuitously
- Make a pie chart (preferably in color and 3d)
- Use a poorly chosen scale

[3] www.biostat.wisc.edu/ kbroman/

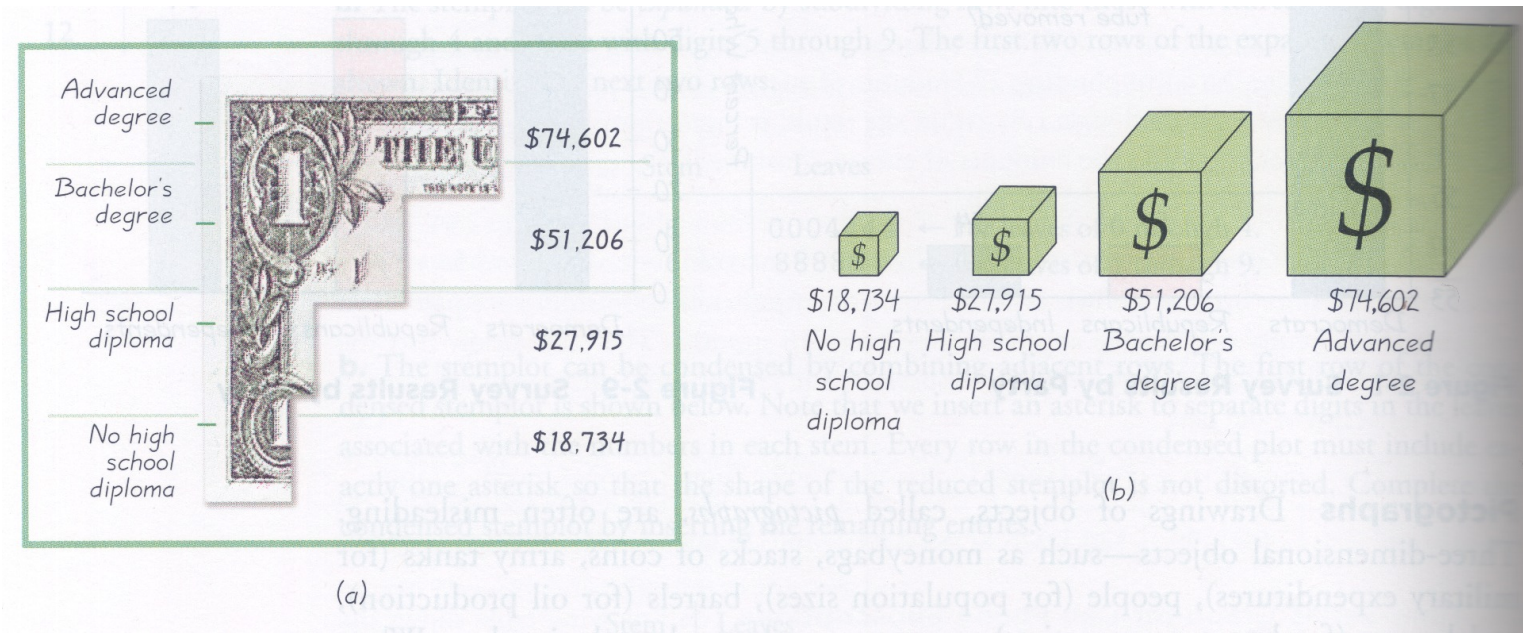
How not to present your data

Terri Shiavo and the CNN/USA Today/Gallup poll:



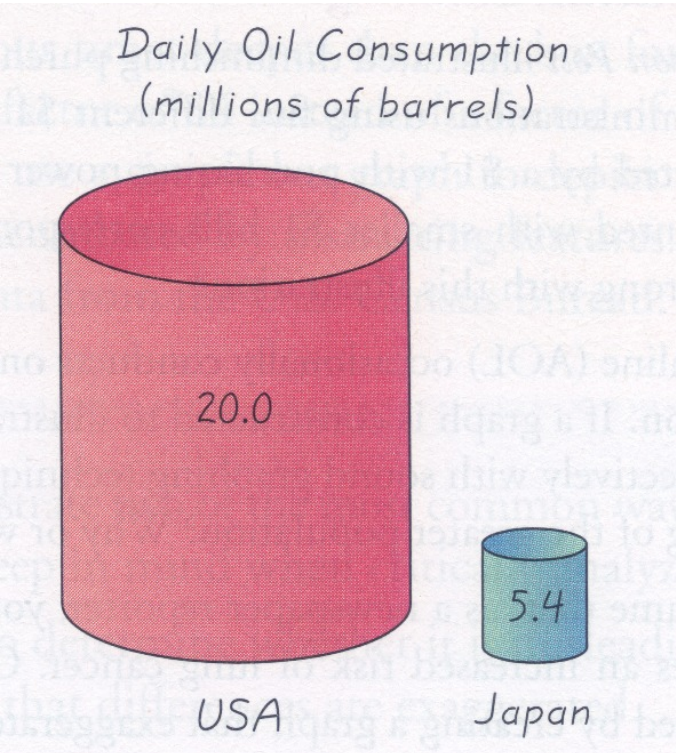
from Triola: Essentials of Statistics (2011)

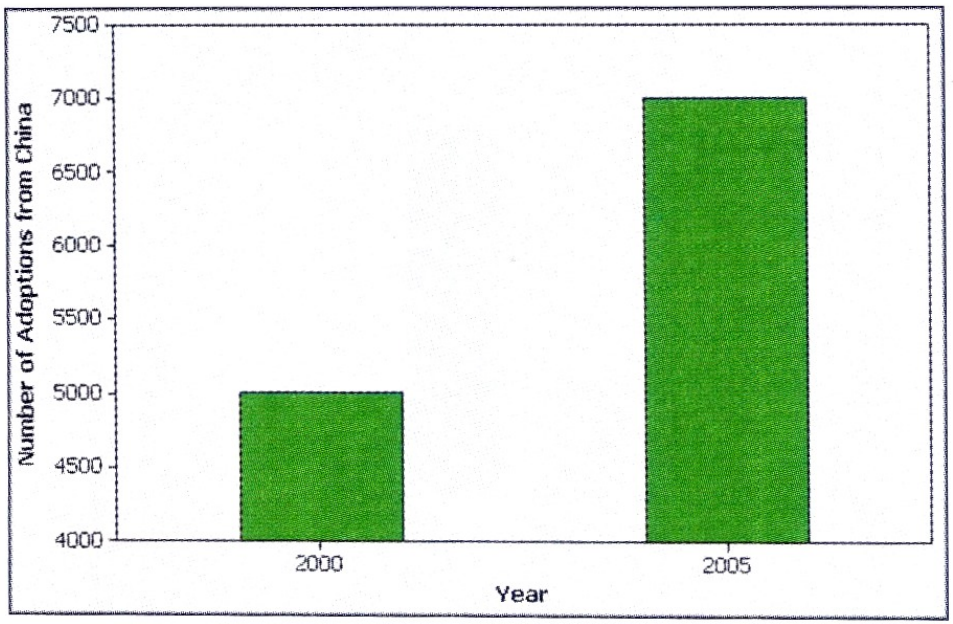
How not to present your data II



from Triola: Essentials of Statistics (2011)

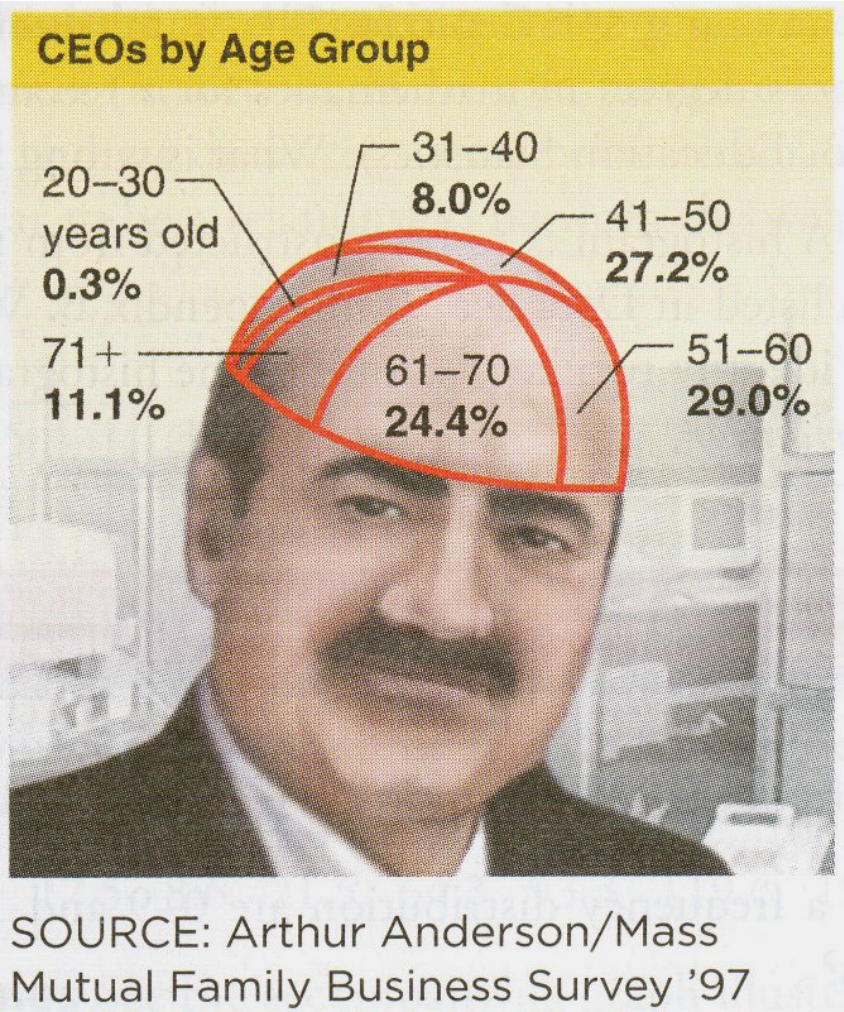
How not to present your data III





from Triola: Essentials of Statistics (2011)

When pie Charts are evil



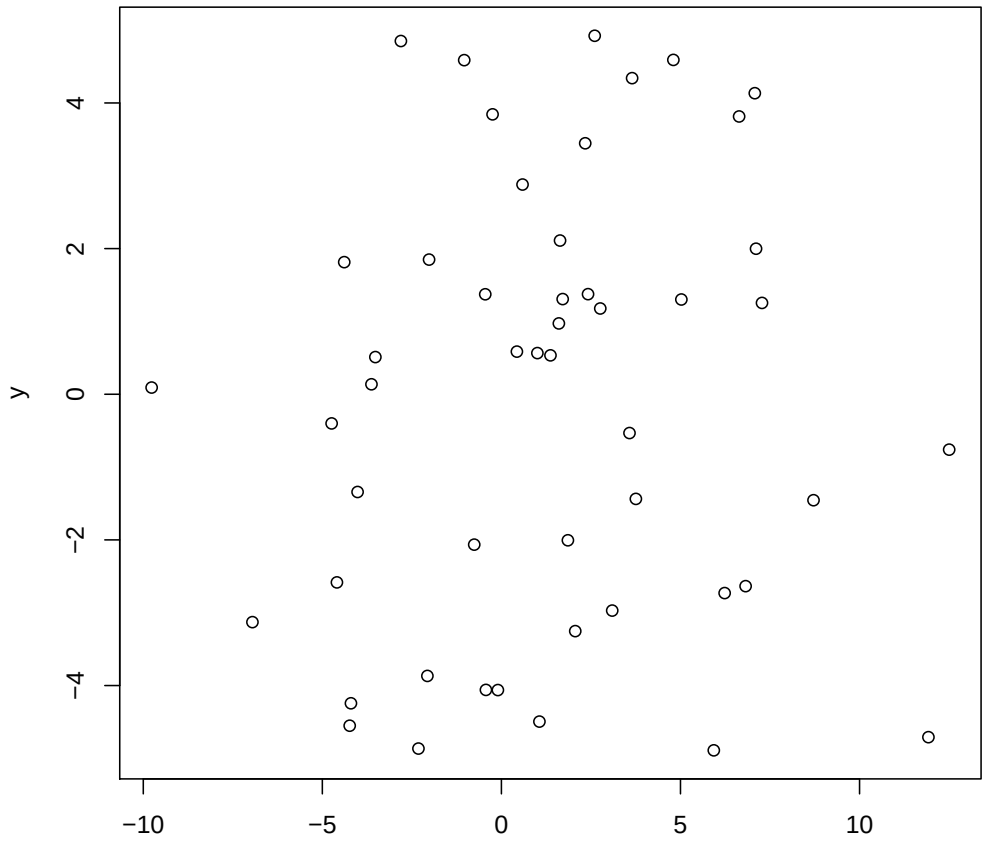
Visualisation - Infographics

http://www.informationisbeautiful.net/

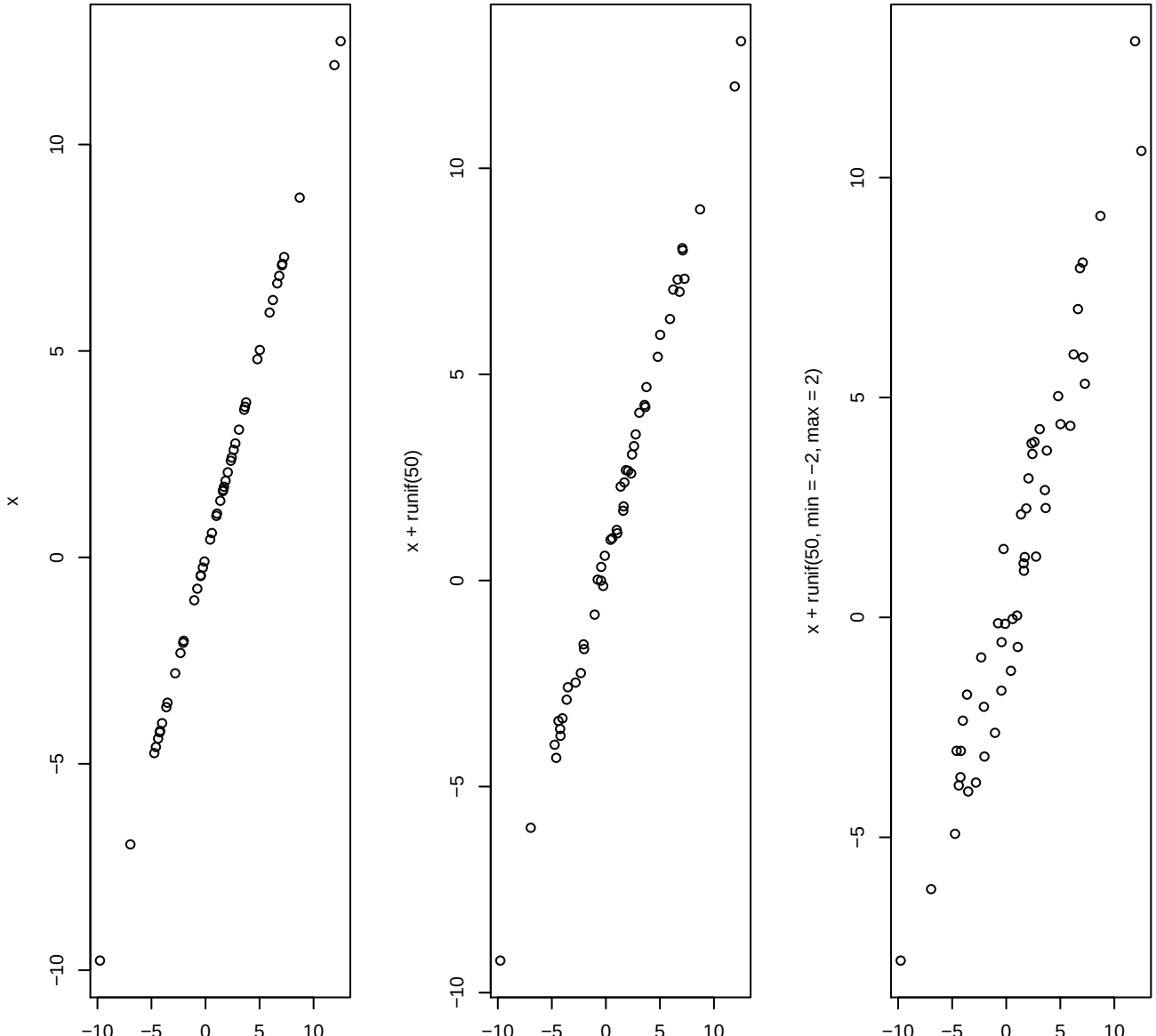
Relations of two numerical variables

Scatterplot

visualising the relationship between two metric variables by plot one variable along the x-axis, the other along the y-axis (R: plot)



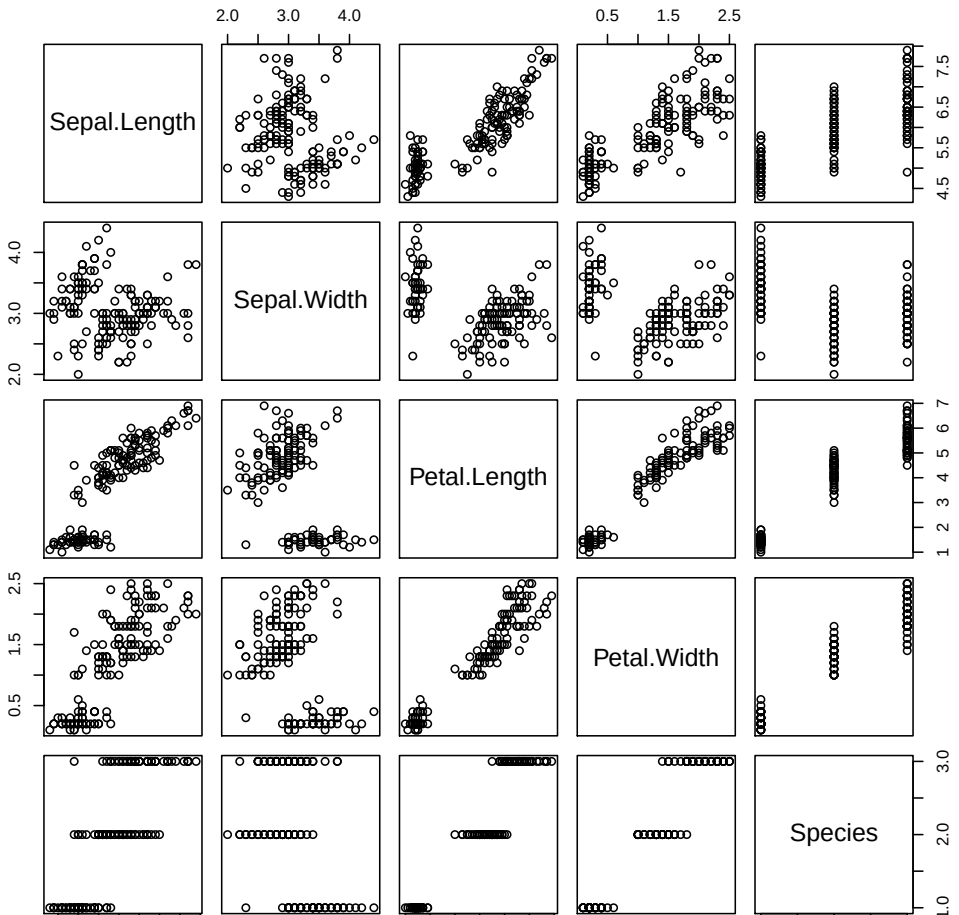
Scatterplot: Seeing relations



Relations of two numerical variables

Pairwise Scatterplot Matrix

for plotting the pairwise scatterplots of more than two possible variables (R: pairs)



Covariance

Covariance and Correlation

The metric variables X and Y have realizations $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. Then the covariance of X and Y can be estimated through

$$\widehat{\text{cov}}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}).$$

We then define the Pearson's correlation coefficient as

$$r = r(X, Y) = \frac{\widehat{\text{cov}}(X, Y)}{s(X) \cdot s(Y)},$$

which only takes values between -1 and 1.

Spearman's rank correlation for *ordinal* data

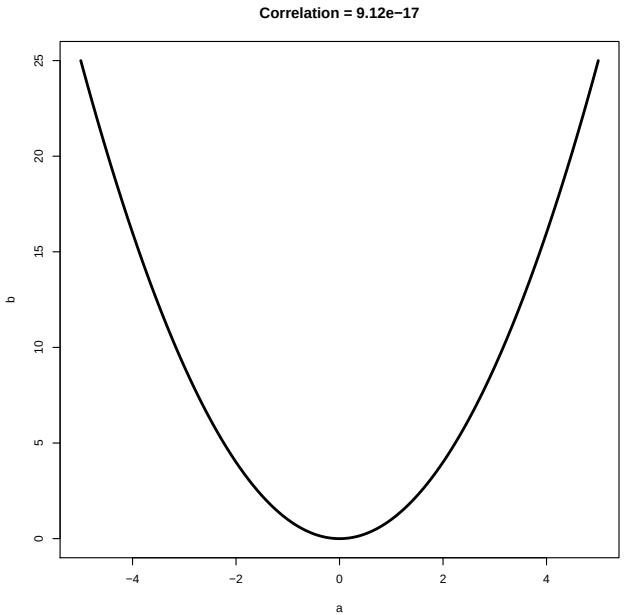
Spearman's rank correlation

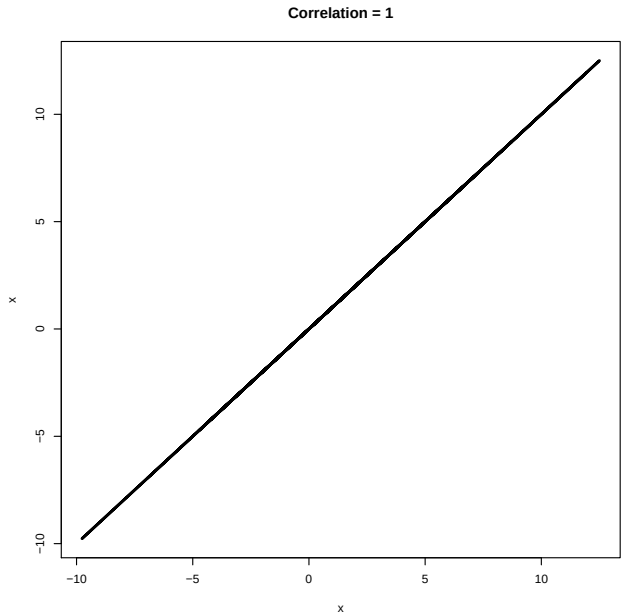
$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)},$$

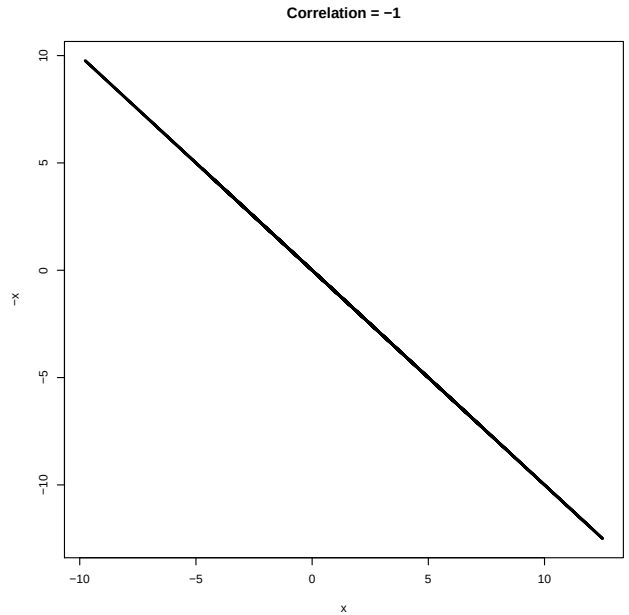
where d_i denotes the difference between ranks for the two values within the ith pair "without ties". If there are ties in the ranks, simply calculate the Pearson correlation r for the ranks. Rank correlation is a *nonparametric* measure of dependence, often considered as more robust.

$$R : cor(x, method = c("pearson", "spearman"))$$

Correlation visualised

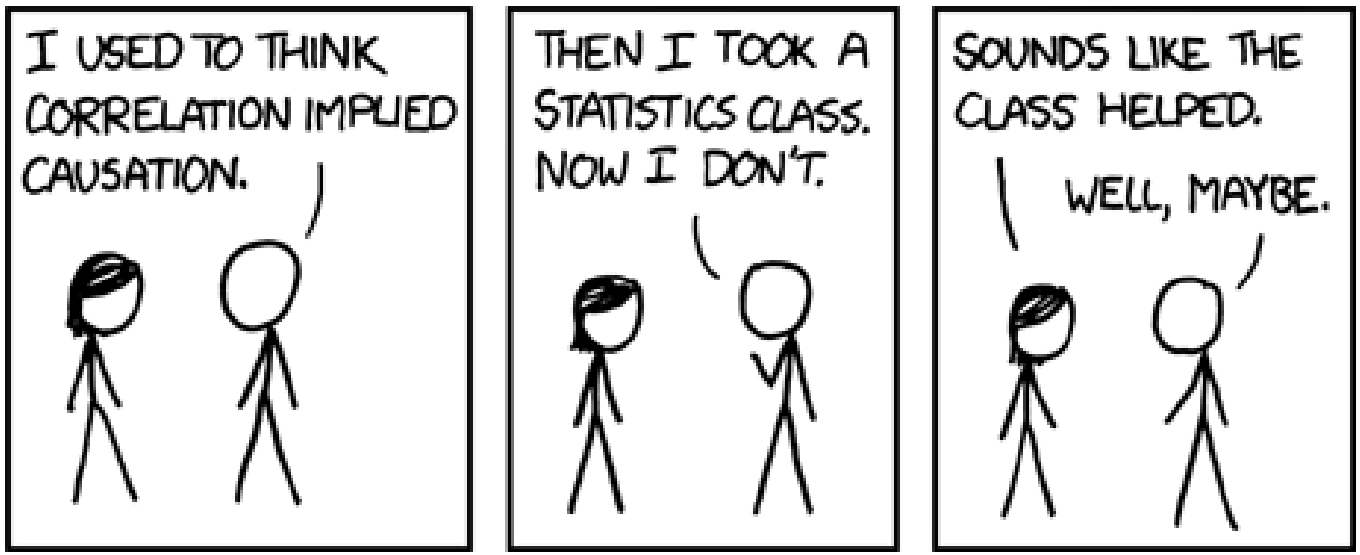






Correlation is not causality!

- Correlation is only a **linear** measure of dependence!
- Correlation **does not imply causality** ("babies and storks")!



Causality in Science

When a scientist says "X causes Y", he/she means:

- "X is only one of a number of possible causes of Y."
- "The occurence of X makes the occurrence of Y more probable."

One can never prove that X is a cause of Y. At best, we can infer that X is a cause of Y.

Can we measure causality at all?

Please be aware that even if we manage to find the ideal survey method, obtain a perfectly representative random sample and measure exactly what we intend to measure, we can not automatically make causal inference – unless we design a suitable experiment guaranteeing:

- concomitant variation,
- time order of occurrence of variables,
- absence of other possible causal factors,
- internal and external validity.

This is impossible in many cases and in general quite cumbersome. If no proper experiment has been conducted and confounding (or extraneous) variables are controlled, the results must be interpreted with great care!

Example for missleading correlation

AIDS infections and cellphone usage in Switzerland

Year	1995	1996	1997	1998	1999
AIDS infections	736	542	565	422	262
Cellphone users (in thous.)	447	663	1044	1699	3058

- Very strong negative correlation between AIDS infections and cellphone users with a correlation coefficient $r = -0.94$
- Do mobile phones protect us from AIDS?
- (Or is time fooling us?)

from Duller: Einführung in die Statistik mit EXCEL und SPSS (2007)

AIDS and mobile phones with rank correlation

Year	1995	1996	1997	1998	1999
AIDS (rank)	736 (5)	542 (3)	565 (4)	422 (2)	262 (1)
Phone (rank)	447 (1)	663 (2)	1044 (3)	1699 (4)	3058 (5)

$$r_s = 1 - \frac{6 \cdot [(5-1)^2 + (3-2)^2 + (4-3)^2 + (2-4)^2 + (1-5)^2]}{5 \cdot (5^2 - 1)} = 0.9$$

A rule of thumb for interpreting correlation:

$r_{(s)} = 0$	no correlation
$0 < \lvert r_{(s)} \rvert \leq 0.3$	weak correlation
$0.3 < \lvert r_{(s)} \rvert \leq 0.7$	medium correlation
$0.7 < \lvert r_{(s)} \rvert < 1$	strong correlation
$\lvert r_{(s)} \rvert = 1$	complete correlation

Fooled by correlation