Classical and Bayesian Statistics - Notes

Table of contents

Introduction

Statistic is the discipline that concerns collection, organisation, analysis, interpretation, and presentation of data. Applied statistics applies to real everyday problems.

> Note: There is no cooking recipes how to solve problems.

Classical Statistics

Classical statistics is a set of tools for decision making using hypothesis.

Bayesien Statistic

Bayesian statistics is a statistical theory that interprets probability as a degree of belief in an event, which can be updated as new evidence is obtained.

> Note: Bayesian statistics is not the same as the Bayesian theorem.

Models

Models are used to simplify things and are essential for statistics. However, models are only useful in a certain context. Models have their limitations. They are statements about how nature operates that deliberately omit many details, thus achieving insight that would otherwise be obscured.

Simulations

Simulations are used to approximate quantities for which an exact solution would be very difficult, if not impossible, to determine. They rely heavily on computer power.

Data

One-Dimensional

List are the simplest kind od datasets. Lists are heterogeneous data structures.

Two-Dimentinal

Tabels are the most common form of datasets

Exploratory Datat Analysis (EDA)

The aim of EDA ist to summarize data by numerical parameters and graphical representation of data. Data should if possible always be graphically displayed and compared with corresponding key figures.

> Note: Note: whenever a dataset is reduced by key figures or graphics, information gets lost.

Key Figures

Location parameters

- Arithemtic mean (average)
- Median
- Quantil

Spread parameters

- Empirical variance
- Stadnard deviation
- Interquantile range

Artithmetic Mean

Mean tells a lot about a dataset: "Center" of data. But average does not tell whole story about (quantitative) datasets. Datasets can have a different spread around mean.

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

```
x <- c(4.3, 5.2, 2.7, 3.1)
mean(x)
```

```
[1] 3.825
```

Note: The arithmetic mean is not robust to outliers.

Empirical Variancehave

The value of empirical variance has no physical interpretation.

$$Var(x) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}^2)$$

```
x <- c(4.3, 5.2, 2.7, 3.1)
var(x)
```

```
[1] 1.3025
```

Standard Deviation

Standard deviation is root of variance. Standard deviation has same unit as data itself.

$$s_x = \sqrt{Var(x)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}^2)}$$

```
x <- c(4.3, 5.2, 2.7, 3.1)
sd(x)
```

```
[1] 1.141271
```

Median

Also called central value or average value. Median is much less influenced by extreme observations than mean.

```
x <- c(4.3, 5.2, 2.7, 3.1)
median(x)
```

```
[1] 3.7
```

Note: Consider Mean and Median simultaneously instead of choosing one.

Quantiles

Quantiles are values that divide a dataset into equal parts, allowing for the analysis of the distribution of data. Common types of quantiles include quartiles which split data into four parts.

Note: Most of the time there is no exact 25 % of observations.

```
x <- c(4.3, 5.2, 2.7, 3.1)
quantile(x) # Default quartil
```

```
   0%   25%   50%   75%  100%
2.700 3.000 3.700 4.525 5.200
```

```
quantile(x, p = 0.7) # Individual value
```

```
 70%
4.39
```

Interquantile Range

Measure for spread of data.

$$\text{upper quartile} - \text{lower quartile}$$

```r
x <- c(4.3, 5.2, 2.7, 3.1)
IQR(x)
```

```
[1] 1.525
```

Graphical Representation of Data

Plotting data is a very important aspect of statistical data analysis. It often points to patterns that are not recognizable from key figures.

> Note: Choosing the "wrong" graphical representation is not useful.

Boxplot

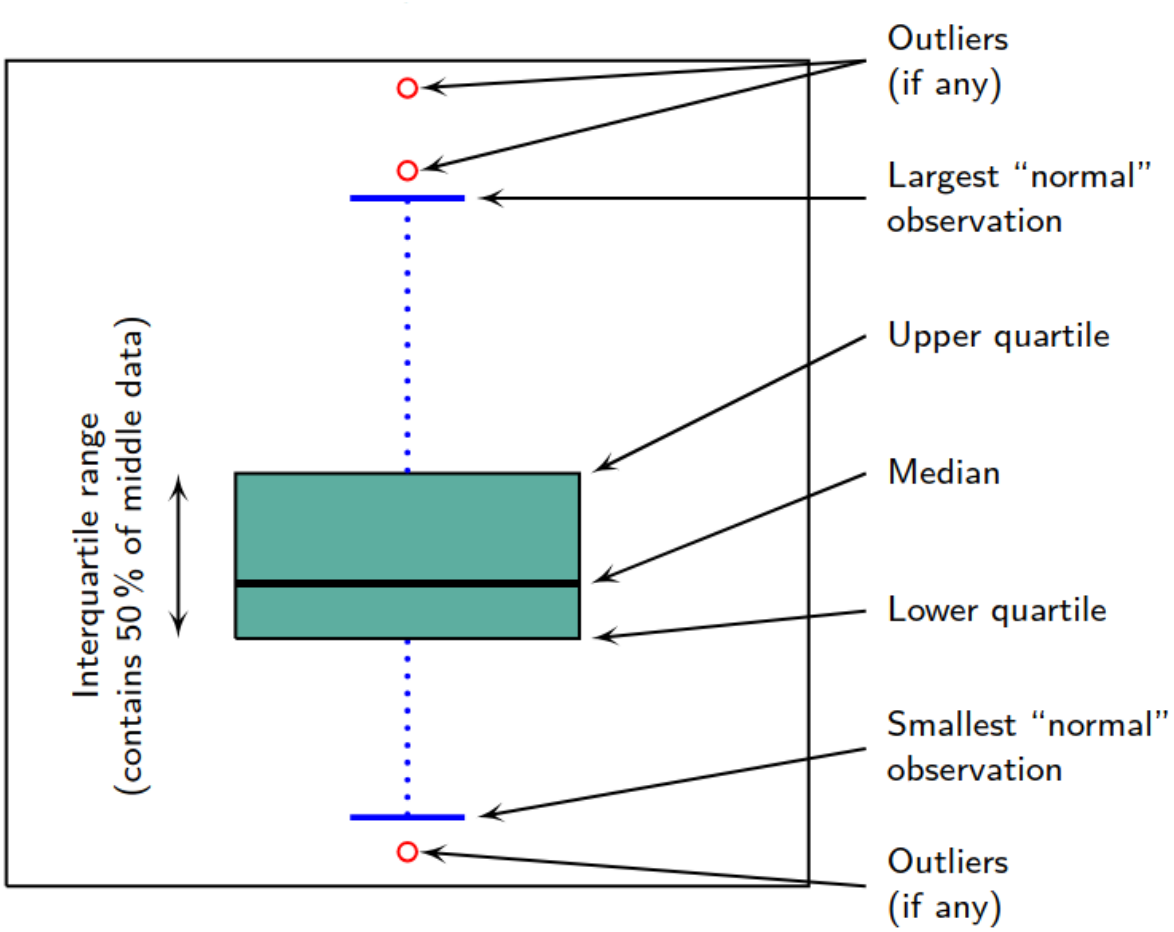


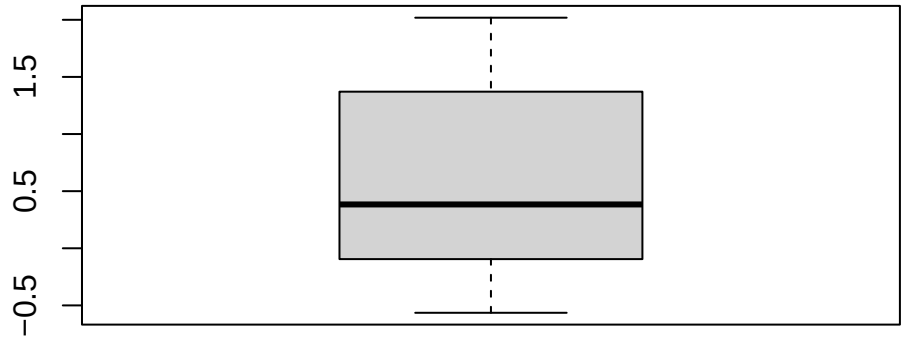
Figure 1: Boxplot: Schematically

- Box: Height is bounded by the lower and upper quartiles. The height of the box is the interquartile range.
- Horizontal line in box: Median
- Whisker: $1.5 \times$ IQR, defined by inventor John Tukey.
- Points: Outliers

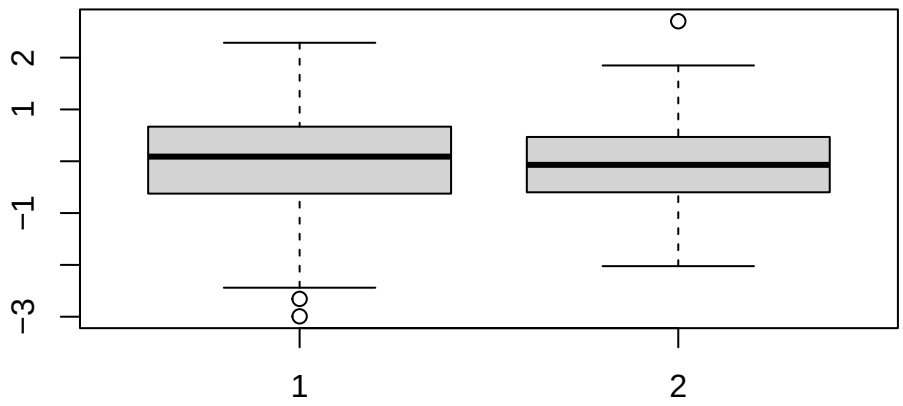
  Note: The upper and lower whiskers do not have to be $1.5 \times$ IQR in length.

```
# Single boxplot
set.seed(42) # Set seed for reproducibility
x <- rnorm(10) # Generate 10 random samples
```

```r
boxplot(x) # Plot data
```



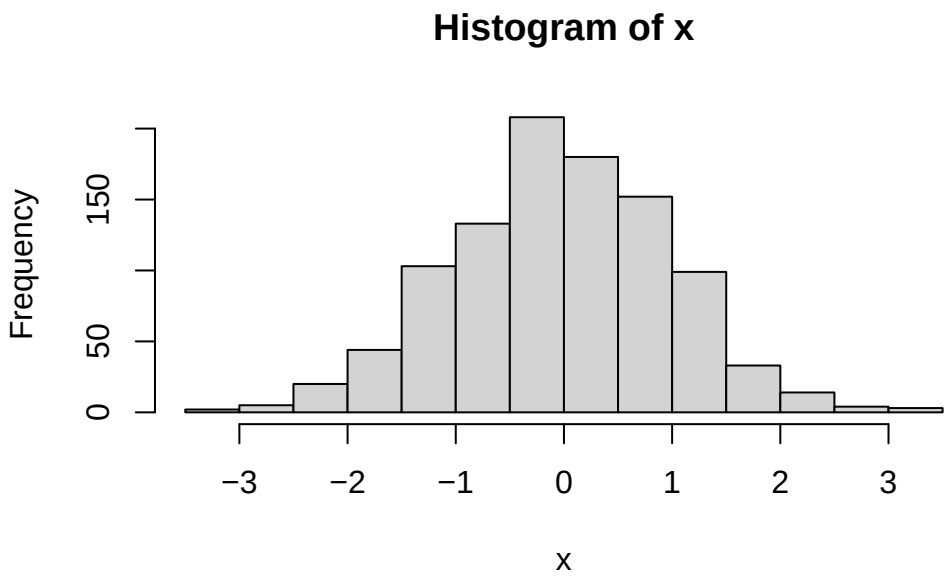

```r
# Compare two sample groups
set.seed(42)
x <- rnorm(100)
y <- rnorm(100)

boxplot(x, y)
```



Histogram

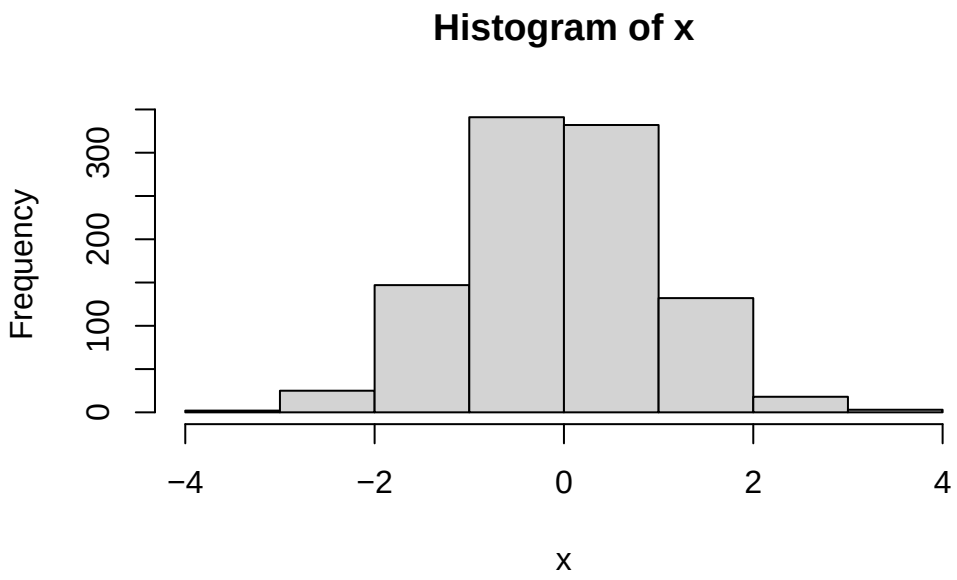
Graphical overview of occurring values. Draw a bar for each class, with the height proportional to the number of observations in that class.

```
# Simple histogram
set.seed(42)
x <- rnorm(1000)

hist(x)
```

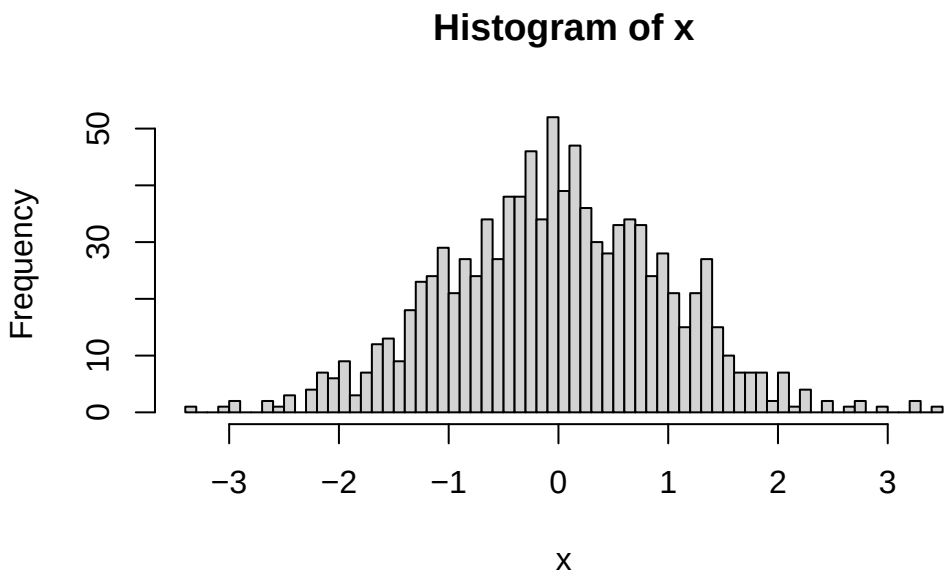


The selection of the number of classes is relevant for the interpretation of a histogram.

```
# Big breaks
set.seed(42)
x <- rnorm(1000)

hist(x, breaks=5)
```




```r
# Small breaks
set.seed(42)
x <- rnorm(1000)

hist(x, breaks=50)
```

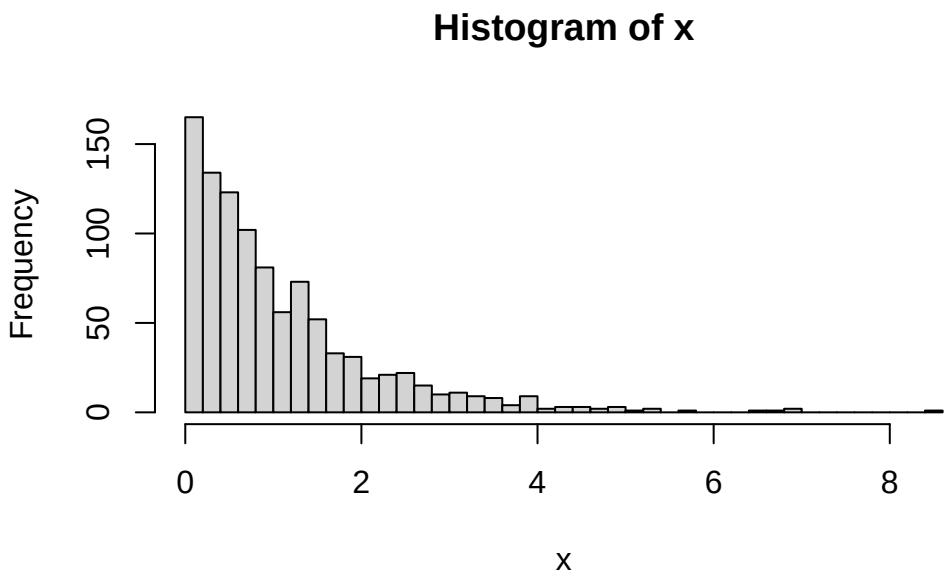


Note: Since we used `set.seed(42)`, both plots show the same data.

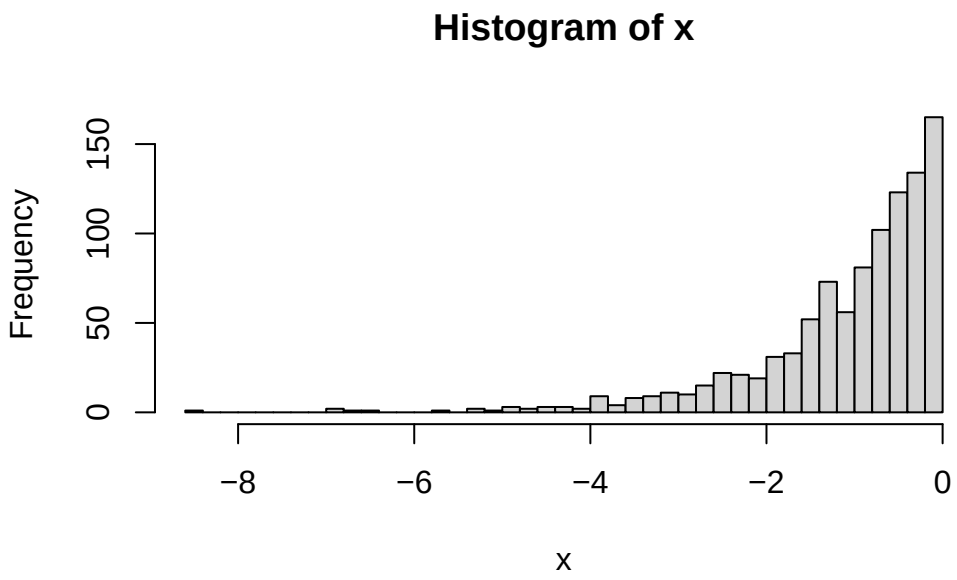
Skweness of Hostograms

Histograms can have a skewness depending on the data.

```
# Right skewed data
set.seed(42)
x <- rexp(1000, rate = 1) # Using expontial distribution

hist(x, breaks=50)
```




```r
# Left left data
set.seed(42)
x <- rexp(1000, rate = 1)
x <- -x # Trun data positivity

hist(x, breaks=50)
```



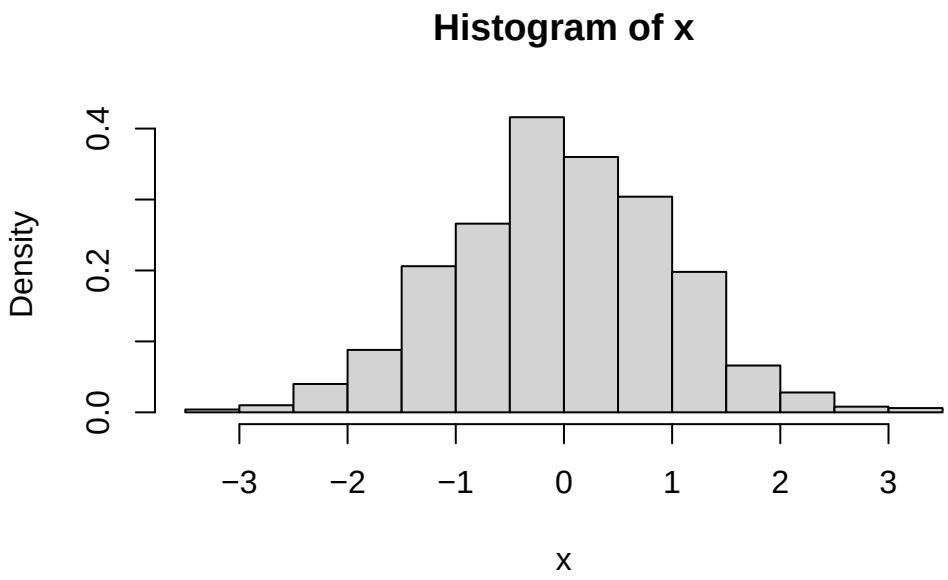
Note: The terms “right” and “left” always refer to the direction where there is less data (the tail of the distribution).

Normalised Histograms

Select the bar height such that the bar area corresponds to the proportion of respective observations in the total number of observations.

Note: Density values are not percentages.

```r
# Normalized histogram
set.seed(42)
x <- rnorm(1000)

hist(x, freq=FALSE) # Disable frequency
```


Histogram of x
Density
0.0
0.2
0.4
−3
−2
−1
0
1
2
3
x

Boxplot X Histogram

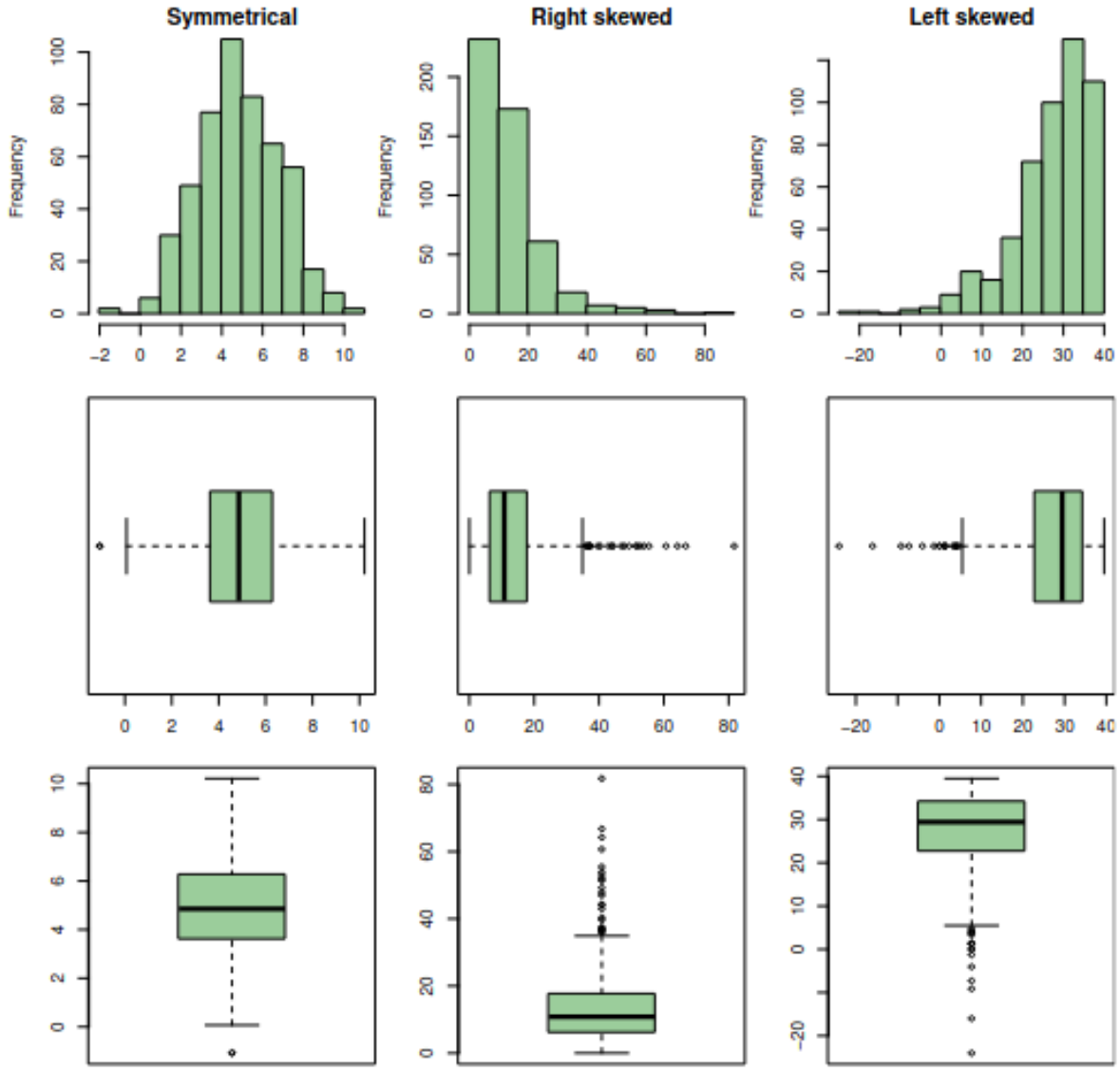


Figure 2: Boxplot X Histogram

Probability

- A random element: $x = 2$
- Set of elements: $X = 1, 2, 3, 4$
- x is an element of the set X: $x \in X$

- Subset: If $B = 2, 3$, then $B \subset A$
- Empty set: $\emptyset$. The empty set is a subset of every set.

Probability Model

Probability models use set theory as a language. For random experiments, the outcome is not predictable. A probability model consists of events that are possible in such an experiment and probabilities for different results occurring. Probability models have the following components:

- Sample space Ω: Contains all possible elementary events ω.
- Events A, B, C: Subsets of the sample space.
- Probabilities p associated with events A, B, C.
- Event: More general and more important than elementary events, but consists of them.

Probability formula:

$$p(E) = \frac{|E|}{|\Omega|}, \text{ where } E \subseteq \Omega$$

Set Theory

- Union: $A \cup B$
- Intersection: $A \cap B$
- Complement: $\bar{A}$
- Difference: $A \setminus B$

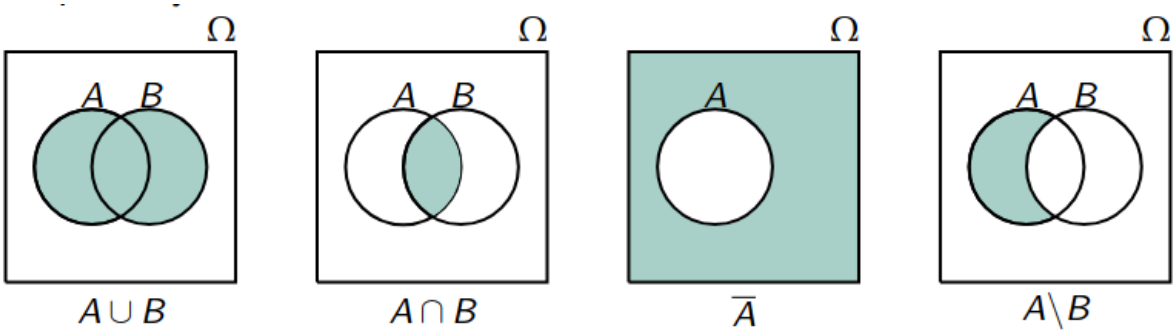


Figure 3: Set theory

Axioms of Probability

Each event A probability $P(A)$ is assigned, with properties:

- $p(A) \geq 0$
- $p(\Omega) = 1$
- $p(A \cup B) = p(A) + p(B)$, if $A \cap B = \emptyset$

  Note: Kolmogorov Axioms of Probability

Laws for Calculating Probabilities

If A, B and $A_1, \ldots, A_n$ events, then:

- $p(\bar{A}) = 1 - p(A)$, for all A
- $p(A \cup B) = p(A) + p(B) - p(A \cap B)$, for all A and B
- $p(A_1 \cup \ldots A_n) \leq p(A_1) + \cdots + p(A_n)$, for all $A_1, \ldots, A_n$
- $p(B) \leq p(A)$, for all A and B with $B \subseteq A$
- $p(A \setminus B) = p(A) - p(B)$, for all A and B with $B \subseteq A$

Discrete Probability Models

A sample space can be finite or infinite and discrete. It can also be infinite but still discrete.

Probabilities for Discrete Models

The probability of event $A = \omega_1, \omega_2, \ldots, \omega_n$ is determined by the sum of the probabilities $p(\omega)$ of the corresponding elementary events.

$$p(A) = \sum_{\omega_i \in A} p(\omega_i)$$

Laplace Model

Probabilities of all elementary elements add up to 1.

$$p(E) = \frac{f}{p} = \sum_{k:\omega_i \in E} p(\omega_k)$$

Divides number of "favorable" elementary events by number of "possible" elementary events.

Stochastic Independence

If events A and B are stochastically independent, then:

$$P(A \cap B) = P(A) \cdot P(B)$$

Note: Formula applies only if events A and B are stochastically independent.

Outcome of event A has no influence on outcome of event B and vice versa.

Random Variable & Probability Distribution

Random Variable

A Random variable X is a function:

$$\Omega \to W_X \subset \mathbb{R}$$
$$\omega \to X(\omega)$$

Random variable denoted by capital letters X (or Y, Z). Corresponding lowercase letters x (or y, z) represents specific value that random variable can take. Once ω is chosen: $X(\omega)$ is fixed, not random.

Note: x also called a realisation of random variable X.

Probability Distribution of Random Variable

Values of random variable X (possible realisations of X) occur with certain probabilities that X takes value x:

$$P(X = x) = P(\{\omega \mid X(\omega) = x\}) = \sum_{\omega; X(\omega)=x} P(\omega)$$

Probability Distribution

Associated probability is determined for all realisations of random variable. List of $P(X = x)$ for all possible values $x_1, x_2, \ldots, x_n$ is called discrete probability distribution of discrete random variable X.All values of probability distribution has to sum up to 1.

$$\sum_{\text{For all } x} P(X = x) = 1$$

Note: For finite sample space the probability distribution is a table.

Key Figures of Distribution

Expected Value $E(X)$

Central location of distribution. Weighted mean of all possible values, weighted by their probability of occurring.

$$\mu = \sum_{\text{all possible } x} x \times P(X = x)$$

Note: $E(X)$ is a theoretical value, which results from a model, i.e. distribution.

Variance

Variance is square of spread of value of random variable from expected value weighted with respective weight.

$$\sigma = \sum_{\text{all possible } x} (x - E(X))^2 \times P(X = x)$$

Standard deviation σ

Spread of distribution about $E(X)$. Standard deviation has same unit as X.

$$\sigma(X) = \sqrt{Var(X)}$$

Note: $\sigma(x)$ is a theoretical value, which results from a model, i.e. distribution.

```r
x <- 1 : 6
p <- 1 / 6
E_X <- sum(x * p)
var_X <- sum((x - E_X)^2 * p)
sd_X <- sqrt(var_X)

sd_X
```

```
[1] 1.707825
```

Note: Means: Deviation on "average" 1.7 from 3.5.

Week 04: Continuous Distributions & Normal Distribution

Continuous Probability Distribution

For continuous probability distributions, probabilities correspond to areas under density function. Range W_X of a random variable: Set of all values X can take. Random variable X is continuous, if its range W_X is continuous. For continuous random variable X for all $x \in W_X$:

$$P(X = x) = 0$$

Note: Probability distribution of X can not be described by $P(X = x)$.

Probability Density

Probability density function f(x) has the properties:

- Function is not negative: $f(x) \geq 0$
- Probability corresponds to area between a and b under $f(x) = P(a < X \leq b)$
- Total area under curve is 1

Note: Values of f(x) are not probabilities, only areas are.

Quantiles of Continuous Probability Distribution

For continuous distributions, the α quantile q_α is value where area (probability) under density function from $-\infty$ to q_α is just α.

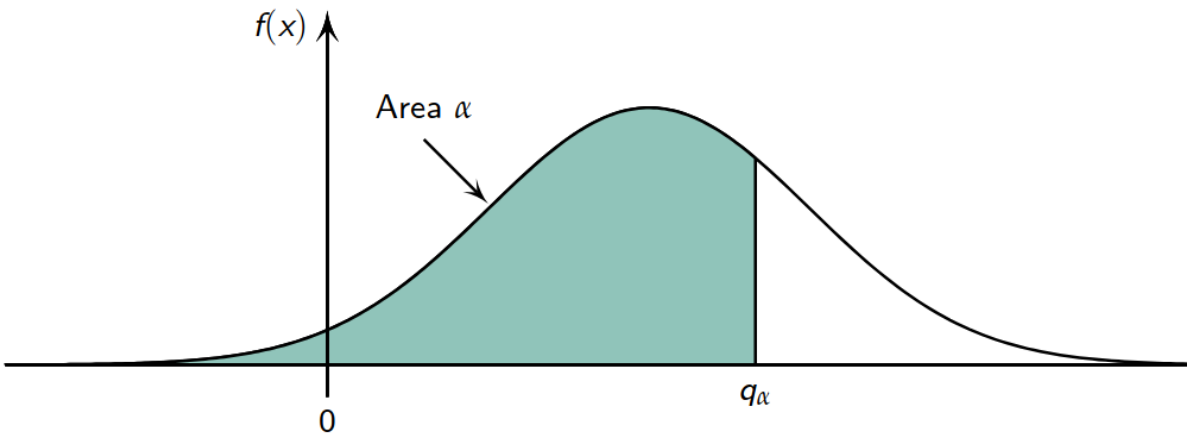


Figure 4: Quantiles of Continuous Probability Distribution

Normal Distribution

The normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ have the desity function:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

With the expected value $E(X) = \mu$ and variance $Var(X) = \sigma^2$. The parameter μ shifts the curve horizontally from origin while σ defines the shape of the curve.

Properties

- Approx. $\frac{2}{3}$ of are between $\mu \pm \sigma$
- Approx. 95% of area between $\mu \pm 2\sigma$

Calculate Normal Distributed Probability

Calculation of IQ $P(X \leq 130)$:

```
# mean and sd are predefined in this task
pnorm(q = 130, mean = 100, sd = 15)
```

```
[1] 0.9772499
```

since total area under curve is 1, $P(X > 130)$:

```
1 - pnorm(q = 130, mean = 100, sd = 15)
```

```
[1] 0.02275013
```

Determine quantiles:

```
qnorm(p = 0.025, mean = 100, sd = 15)
```

```
[1] 70.60054
```

```
qnorm(p = 0.975, mean = 100, sd = 15)
```

```
[1] 129.3995
```

i.i.d Assumption

The i.i.d. assumption means a set of random variables are Independent (the value of one doesn't affect the others) and Identically Distributed (all come from the exact same probability distribution).

Key Figures

S_n

- $E(S_n) = n\mu$
- $\text{Var}(S_n) = n\text{Var}(X_i)$
- $\sigma(S_n) = \sqrt{n}\sigma_X$

$\bar{X}_n$

- $E(\bar{X}_n) = \mu$
- $\mathrm{Var}(\bar{X}_n) = \frac{\sigma_X^2}{n}$
- $\sigma(\bar{X}_n) = \frac{\sigma_X}{\sqrt{n}}$

  Note: Standard deviation of X_n is called standard error of arithmetic mean.

Central Limit Theorem

The Central Limit Theorem (CLT) states that, given a sufficiently large sample size (n), the distribution of the sample mean ($\bar{X}_n$) will be approximately normally distributed, regardless of the original population's distribution. This is true as long as the population has a finite mean (μ) and variance (σ^2). In short: For large samples, the sample mean is normally distributed.

Week 05: Hypothesis Tests

Hypothesis testing are a important statistical tool to decide whether observations "fits" a certain parameter. Ro introduce a standardised, reproducible procedure to decide whether mean of observations does (or not) match a certain true mean μ.

> Note: We can only show that this quantity does not fit to the observations with high probability.

Estimation

Point estimates for expected value.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Point estimates for variance.

$$\hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

> Note: Hat $\hat{x}$ denotes estimate of a quantity.

Procedure Hypothesis Test

Using a observation, check whether, under assumption $\mu = x$, mean of observations is probable or not.

- μ: True (unknown) mean of data
- μ_0: Assumed true mean of data

Null Hypothesis

$$H_0 : \mu = \mu_0 = x$$

Alternative Hypothesis

$$H_A : \mu \neq \mu_0 = x$$

Test with this distribution whether assumption = x is justified. Distribution of test statistic T under the null hypothesis H_0

$$T : \bar{X}_{10} \sim \mathcal{N}(x, \frac{1^2}{x})$$

```
# Assume real mean of 500
pnorm(q = 499.22, mean = 500, sd = 1/sqrt(10))
```

```
[1] 0.006820578
```

It has proven practical to set this limit of what is too small and what is not at 2.5 %. So assume that given mean of $\mu_0 = 500$ is not plausible. We reject null hypothesis.

Significance Level

Significance level α, indicates how high a risk one is willing to take of making a wrong decision. For most tests α value of 0.05 or 0.01. Boundary rejection range 0.025- and 0.975-quantiles.

```
qnorm(p = c(0.025, 0.975), mean = 500, sd = 1/sqrt(10))
```

```
[1] 499.3802 500.6198
```

If observed mean lies in red area of Figure, null hypothesis is rejected, also called rejection range.

p-Value

p-value is probability of observing an event under null hypothesis that is at least as extreme (in direction of alternative) as currently observed event. The smaller p-value, the more result argues against null hypothesis. Values smaller than a predetermined limit, such as 5 %, 1 % or 0.1 % are reason to reject the null hypothesis.

Week 06: t-Test, Wilcoxon-Test, Confidence Interval

t-Distribution

Distribution of test statistics for t-test under null hypothesis. Similar to normal distribution, but flatter, due to greater uncertainty. Depends on number of observations. Symmetric distribution around 0, but flattens out slower than standard normal distribution $\mathcal{N}(0, 1)$

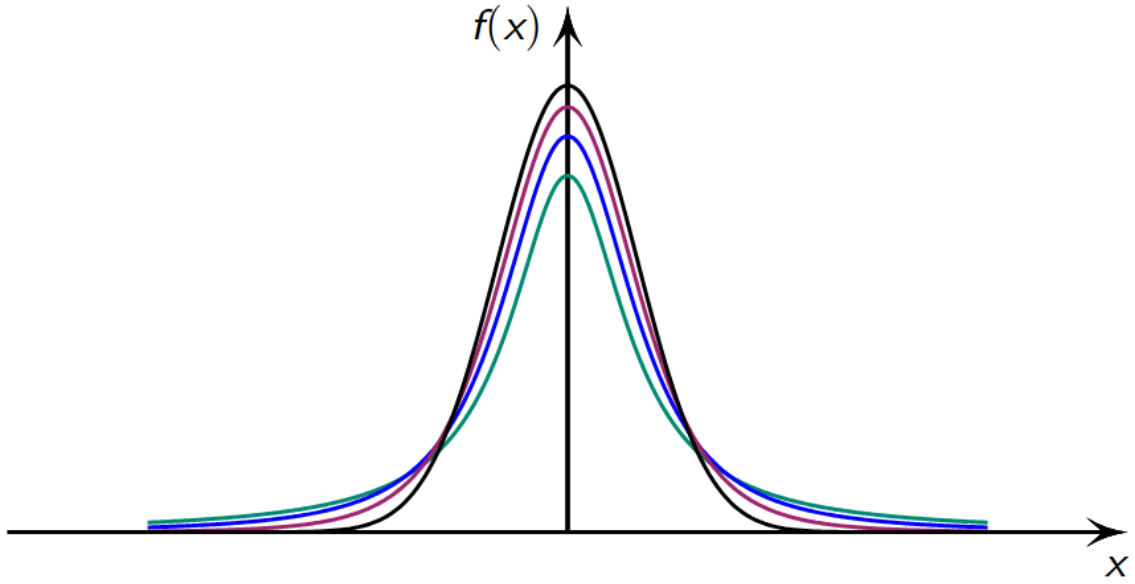


Figure 5: t-Distribution

$$H_0 = \mu = \mu_o$$

is given by

$$T = \bar{X}_n \sim t_{n-1}\left(\mu, \frac{\hat{\sigma}^2_{\bar{X}}}{n}\right)$$

where t_{n-1} is a t-distribution with $n-1$ degrees of freedom.

```
x <- c(5.9, 3.4, 6.6, 6.3, 4.2, 2.0, 6.0, 4.8, 4.2, 2.1, 8.7, 4.4, 5.1, 2.7, 8.5, 5.8, 4.9, 5
mean(x)
```

```
[1] 5.215
```

```
sd(x)
```

```
[1] 1.883802
```

```
t.test(x, mu = 5)
```

```
    One Sample t-test

data:  x
t = 0.51041, df = 19, p-value = 0.6156
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 4.333353 6.096647
sample estimates:
mean of x
    5.215
```

Confidence Interval

Interval indicating where, roughly speaking, true mean lies with a certain predefined probability

Test Decision with Confidence Interval

If μ_0 of null hypothesis does not lie within confidence interval of $\bar{X}_n$, H_0 is rejected.

Non-Normally Distributed Data: Wilcoxon Test

Assumes less than t-test. Distribution under null hypothesis is symmetrical with respect to median μ_0.

```
x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97, 80.05, 80.03, 80.02, 80.00, 80

wilcox.test(x, mu = 80.00, alternative = "two.sided")
```

```
Warning in wilcox.test.default(x, mu = 80, alternative = "two.sided"): cannot
compute exact p-value with ties
```

```
Warning in wilcox.test.default(x, mu = 80, alternative = "two.sided"): cannot
compute exact p-value with zeroes
```

```
    Wilcoxon signed rank test with continuity correction

data:  x
V = 69, p-value = 0.0195
alternative hypothesis: true location is not equal to 80
```

Wilcoxon test vs. t-test

Wilcoxon test is in the vast majority of cases preferable to the t-test: It often has much greater power in many situations (probability of correctly rejecting the null hypothesis). Even in the most extreme cases it is never much worse.

Paired Samples

Both observations are not independent, because same experimental unit is measured twice. Each observation of one group can be clearly assigned to an observation of the other group. Sample size is inevitably same in both groups

```
before <- c(25, 25, 27, 44, 30, 67, 53, 53, 52, 60, 28)
after <- c(27, 29, 37, 56, 46, 82, 57, 80, 61, 59, 43)

t.test(x = before, y = after, paired=TRUE)
```

```
    Paired t-test

data:  before and after
t = -4.2716, df = 10, p-value = 0.001633
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -15.63114  -4.91431
sample estimates:
mean difference
      -10.27273
```

Unpaired (Independent) Samples

So-called unpaired (or independent) samples. No assignment of observations possible. Sample sizes can be differen.

Mann-Whitney U-Test (aka Wilcoxon Rank-sum Test)

If data are non-normally distributed.

```
x <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04,79.97, 80.05, 80.03, 80.02, 80.00, 80
y <- c(80.02, 79.94, 79.98, 79.97, 80.03, 79.95, 79.97)

wilcox.test(x, y, alternative="two.sided", mu=0, paired=FALSE, conf.level=0.95)
```

```
Warning in wilcox.test.default(x, y, alternative = "two.sided", mu = 0, :
cannot compute exact p-value with ties

    Wilcoxon rank sum test with continuity correction

data:  x and y
W = 76.5, p-value = 0.01454
alternative hypothesis: true location shift is not equal to 0
```

Interpreting p-Values

Determine whether hypothesis test results are statistically significant. If p-value is less than significance level, reject null hypothesis and conclude that effect or relationship exists. The p-values are directly connected to null hypothesis.

> **Note**
>
> p-values: Probability of observing a sample statistic that is at least as extreme as sample statistic when assuming that null hypothesis is correct

Note: p-Values Are **not** an Error Rate.

What Is True Error Rate?

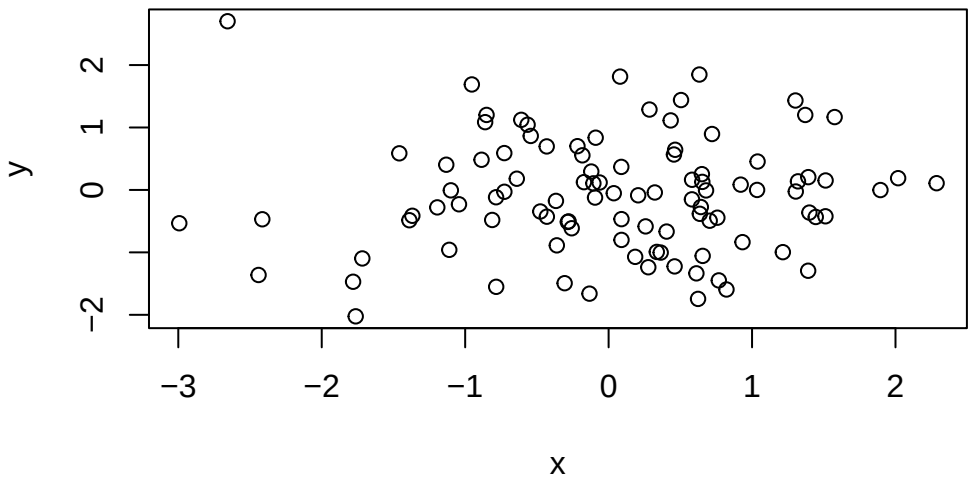
Can't directly calculate error rate based on a p-value.

Week 07: Linear Regression

Scatter plot

Two observations interpreted and displayed as coordinates of points in a coordinate system.

```r
set.seed(42)

# Generate random data
x <- rnorm(100)
y <- rnorm(100)

plot(x, y) # By defaults Scatter Plot
```



Dependence and Causality

Caution regarding scatter plots: Do not confuse dependence with causality. Caution regarding scatter plots: Do not confuse dependence with causality.

Equation of a Line

$$y = a + bx$$

$$b = \frac{\Delta y}{\Delta x}$$

Linear Regression

Formula-based relationship between book x and y. The Problem is to find a line that fits all points as good as possible.

Residuals

Distance between points and line.

> **Note**
>
> A residual r_i is the vertical difference between a data point (x_i , y_i) and the point (x_i , $a + bx_i$) on the sought line:
>
> $$r_i = y_i - (a + bx_i) = y_i - a - bx_i$$

Least Squares Method

Determine a and b so that the sum of the squared residuals becomes minimal.

$$r_1^2 + r_2^2 * \cdots + r_n^2 = \sum_i r_i^2$$

Parameter a, b minimise (least squares method)

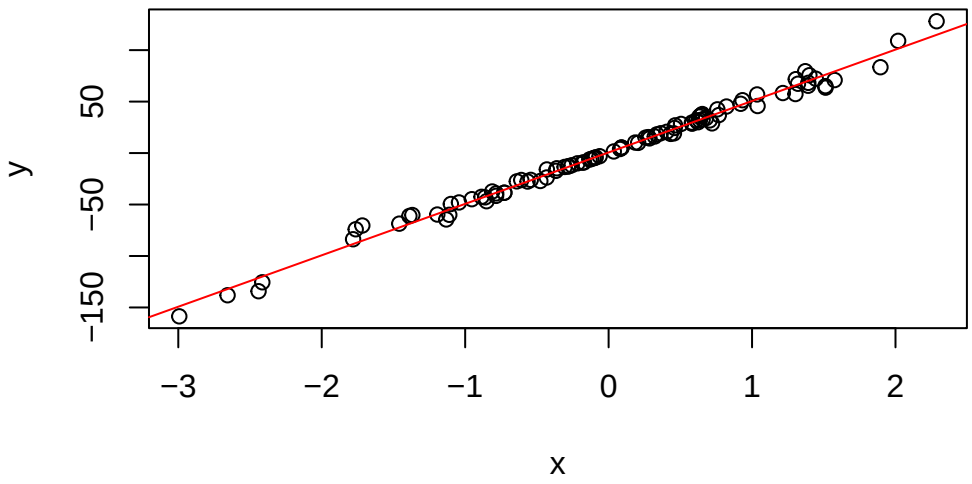
$$\sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

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

$$a = \bar{y} - b\bar{x}$$

Note: $\bar{x}$ and $\bar{y}$ are the mean values of the respective data.

```
set.seed(42)

# Generate random linear data
x <- rnorm(100)
y <- x * rbinom(100, size=100, prob=0.5)

plot(x, y)
abline(lm(y ~ x), col="red") # Regression line
```



Note: It is `lm(y~x)` and not `lm(x~y)`.

Empirical correlation

The empirical correlation is a dimensionless number between -1 and $+1$ and measures strength and direction of the linear dependency between the dates x and y.

```
set.seed(42)

x <- rnorm(100)
y <- x * rbinom(100, size=100, prob=0.5)

cor(x, y) # High value because of pos linear data
```

```
[1] 0.9953896
```

> **Warning**
>
> Empirical correlation only measures the linear correlation.

R^2 Value

R^2-statistics: Value between 0 and 1. It indicates to what proportion of the variability in Y is explained by X using the model. Value close to 1: A large proportion of the variability is explained by the regression. The model therefore describes the data very well.

Note: R^2 can be used for any regression.

```
summary(lm(Sales~TV))$r.squared
```

F-Statistic

The F-statistic is a value resulting from an F-test, which evaluates the overall significance of a regression model. Unlike the t-test, which looks at individual predictors, the F-test looks at the model as a whole.

- High F-value: Indicates that the model explains a significant amount of the variation in the data.
- Low F-value (near 1): Suggests that the observed patterns could likely be due to random chance.

Week 08: Multiple Linear Regression

Simple linear regression are useful procedure to predict output based on one single predictor. In practice the output often depends on more than one predictor. Multiple linear model generalises simple linear model. Calculations and interpretations for multiple model similar, although usually more complicated than linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

```
# Same syntax as in linear models
model <- lm(Sales ~ TV + Radio + Newspaper)
```

Interpretation of Coefficients

The interpretation of the coefficients is similar to the interpretation in a linear regression model.

- β_0 (Intercept): β_0 represents the estimated mean value of the dependent variable Y when all predictor variables (X_1 , $X : 2$, ... , X_p) are equal to zero.
- β_i (Partial Effect): A coefficient β_i represents the estimated change in the mean of Y for a one-unit increase in the predictor variable X_i , while holding all other predictor variables constant.

Estimation of Regression Coefficients

The regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ generally unknown. We estimate those based on the data by using a a cost function like RSS.

$$RSS = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

```
coef(lm(Sales ~ TV + Radio + Newspaper))

## (Intercept) TV          Radio       Newspaper
## 2.938889369 0.045764645 0.188530017 -0.001037493
```

> **!** Important
>
> Slope for *Newspaper* describes change in response *Sales* when spending *CHF 1000* more on newspaper advertising, **while other two predictors *TV* and *radio* are hold constant**

Correlation coefficients

```
cor(data.frame(TV, Radio, Newspaper, Sales))

##              TV         Radio      Newspaper  Sales
## TV        1.00000000 0.05480866 0.05664787 0.7822244
## Radio     0.05480866 1.00000000 0.35410375 0.5762226
## Newspaper 0.05664787 0.35410375 1.00000000 0.2282990
## Sales     0.78222442 0.57622257 0.22829903 1.0000000
```

Relationship between Predictors and Response Variable

Multiple linear regression with p predictors: All regression coefficients except β_0 are zero, no variable has influence (Null hypothesis). As a alternative hypothesis: At least one β_i is not equal to 0.

No Linear Regression

The moderated effect, often referred to as an interaction effect, summarizes a statistical relationship where the influence of one variable is not constant but instead depends on the value of a second variable.

```
model -> lm(Sales ~ TV + Radio + TV * Radio)
```

> Note: The asterisk $*$ in R is an abbreviation for the addition of the main effects AND the interaction term.

Week 09: Conditional Probability

Conditional probability is probability that event A occurs when one already knows that B has occurred

General formula:

$$P(A \cap B) = \frac{A \cap B}{\Omega}$$

Formula is used to define conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(S)}$$

> Caution
>
> $P(A|B) \neq P(B|A)$

Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Law of Total Probability

f $A_1, \ldots, A_k$ is a partition of A and B an event, then:

$$\sum_{i=1}^{k} P(B|A_k)P(A_k)$$

Week 10: Bayesian Statistics Introduction

Bayesian Statistics

Bayesian statistics is a theory in the field of statistics based on the Bayesian interpretation of probability, where probability expresses a degree of belief in an event.

In general:

$$P(r \cap c) = P(r|c)P(c)$$

But also:

$$P(r \cap c) = P(c|r)P(r)$$

Therefore:

$$P(c|r)P(r) = P(r|c)P(c)$$

Hence:

$$P(c|r) = \frac{P(r|c)P(c)}{P(r)}$$

Bayes' Theorem: Prior, Likelihood, Posterior

Introduce the following terms:

- $P(M)$: Prior probability
- $P(M|+)$: Posterior probability
- $P(+|M)$: Likelihood function

Week 11: Beta distribution

Evidence `P(D)`

Calculate evidence P(D) with law total probability.

$$\sum_{i=1}^{n} P(D|\theta_i)P(\theta_i)$$

Use probability density functions instead of probabilities and sums become integrals.

$$P(D) = \int P(D \mid \theta^*)p(\theta^*)d\theta^*$$

> Note: Areas under probability density curves correspond to probabilities. Total ares under the curve is 1.

Bernoulli distribution

Likelihood function for coin tosses

If N denotes number of tosses, z number of H and $N - z$ number of T, then probability distribution:

$$P(\{y_i\} \mid \theta) = \theta^Z(1-\theta)^{N-Z}$$

Description of Probabilities: Beta Distribution

The Beta distribution ($Beta(\alpha, \beta)$) is a continuous probability distribution defined over the interval $[0, 1]$. The Beta distribution is controlled by two positive shape parameters, typically denoted as α and β. By changing α and β, the distribution can take on a wide range of shapes, including symmetric, skewed left or right, or even U-shaped.

The ratio of α to β determines where the peak (the mode) lies on the x-axis.

- $a > b$**:** There are more successes than failures. The probability θ is likely high.
  - **Effect:** The peak shifts to the **right** (above 0.5).
- $b > a$**:** There are more failures than successes. The probability θ is likely low.
  - **Effect:** The peak shifts to the **left** (below 0.5).

- $a = b$**:** Successes and failures balance each other out.
  - **Effect:** The peak is exactly in the **middle** (at 0.5).

The sum $\alpha + \beta$ (often called n or concentration) determines how narrow or wide the curve is. This is the "weight" of your experience.

- a **and** b **are small (e.g., 2 and 2):**
  - You have hardly seen any data. Although it's 50/50, you are uncertain.
  - **Shape:** A flat, wide bump. The HDI is huge.
- a **and** b **are large (e.g., 100 and 100):**
  - You have seen a lot of data. You are very sure that it is 50/50.
  - **Shape:** A tall, sharp needle. The HDI is tiny.

Central Tendency

Our goal is to transform a prior belief, which is expressed in terms of tendency and uncertainty, into corresponding parameter values a and b in the beta distribution.

Mean

$$\mu = \frac{a}{a+b}$$

Mode

$$\omega = \frac{a-1}{a+b-2}$$

Spread

$$k = a + b$$

For a and b in terms of mean μ and concentration κ

$$a = \mu\kappa$$

$$b = (1-\mu)\kappa$$

For a and b in terms of mode ω and concentration κ

$$a = \omega(\kappa - 2) + 1$$

$$b = (1 - \omega)(\kappa - 2) + 1 \text{ for } \kappa > 2$$

Posterior Beta

If prior distribution is beta distribution Beta(| a, b), and data show z heads in N tosses, then posterior distribution is again beta distribution.

$$p(\theta \mid z, N) = \text{Beta}(\theta \mid z + a, N - z + b)$$

Highest Density Interval (HDI)

To summarise a distribution, use highest density interval (HDI). It indicates which points of a distribution are most credible and which represent largest part of the distributio.

- Wide HDI (large distance between start and end): There are many different values that could be plausible. The probability is distributed "widely" across the x-axis. We cannot narrow down the true value very well.
- Narrow HDI (small distance): The probability is concentrated in a very small range. We are very certain where the true value lies.

> **Note**
>
> Since this normal distribution is symmetric about zero, the 95 %-HDI extends from -1.96 to $+1.96$.

Note: HDI has nothing to do with confidence interval.

ROPE

The ROPE is the "range of tolerance" that you define before the experiment. It prevents us from misinterpreting tiny, insignificant deviations as "significant" just because we have a lot of data.

Relationship between HDI and ROPE	Interpretation	Decision
HDI is completely inside the ROPE	The entire range of credible values is practically equivalent to the null value.	**Accept** the null value (practical equivalence).
HDI is completely outside the ROPE	The entire range of credible values is different from the null value.	**Reject** the null value.
HDI and ROPE overlap	Some credible values are practically equivalent, others are not.	**Undecided / Inconclusive** (more data needed).

Influence of prior on posterior distribution

Practical relevance: It does not matter whether choice $a = b = 10$ or $a = b = 15$ for prior distribution. It does matter whether $a =$ b $= 1\$$ or $a = b = 10$ or $a = b = 100$. If we have no idea about coin: Choose $a = b = 1$. If we have examined the coin more closely and find that it is very symmetrical, we may choose $a = b = 100$.

Week 12: General Metropolis Algorithm

Markov Chain Monte Carlo (MCMC)

Present methods to generate good approximations of Bayes' posterior distributions. Method described starts from two assumptions:

- Prior distribution: Function that can be easily evaluated by a computer.
- Likelihood function: Function that can be easily evaluated by a computer.

Posterior distribution is estimated by randomly generating a set of θ values from it

Approximation of a Distribution Mean Large Samples

Concept of representing a distribution using a large sample of representative θ values is fundamental to approach to Bayesian analysis of complex models. The larger the sample, the better the estimate.

Metropolis Algorithm

Approximation of continuous posterior distribution. Metropolis algorithm generates many representative θ values whose histogram approximates posterior distribution given a sufficiently large number of sample θ values.

Region of Practical Equivalence (ROPE)

Specifies a small range of parameter values that is considered practically equivalent to null value for purposes of a particular application.