

Statistics

Energy Technology 2018/19

Krzysztof Arendt

January 17, 2019

Center for Energy Informatics,
University of Southern Denmark

Table of Contents

Fundamentals

Probability

Distributions

Statistical Inference

Linear Regression

Analysis of Variance

Time Series Analysis

Fundamentals

Experiments

- An **experiment**¹ is a procedure carried out to validate or refute a hypothesis.
- Typically, we study the relationships between some variables.
- Variables can be dependent or independent.
- Independent variables² are manipulated by an experimenter. They're usually denoted as X_i ($i \geq 1$).
- Dependent variables³ describe the event expected to change with the independent variables. They're usually denoted as Y_i ($i \geq 1$).

¹In statistics an experiment is sometimes called a trial

²Other names: predictor/explanatory variables, features, covariates

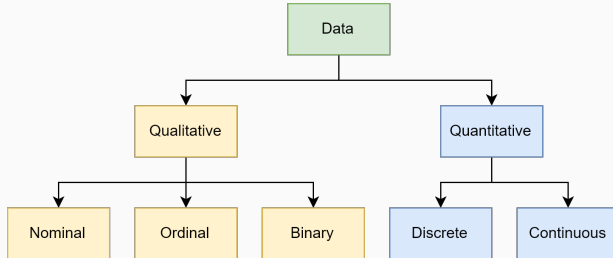
³Other names: response/explained/outcome/output variables, labels

Example

Try to design experiments aimed at:

- Proving that a coin is biased (e.g. it yields more tails than heads in tossing)?
- Finding how much airflow is required to maintain the CO₂ level in this room below 800 ppm.
- Finding how much time devoted to statistics is required to get at least a 75% chance of passing this course.

Data Types



Nominal Variables with inherent order, e.g. gender

Ordinal Variables with ordered series, e.g. performance

Binary Variables with only two options, e.g. coin tossing

Discrete Variables with finite number of possible values, e.g.
number of damaged parts

Continuous Variables with real numbers, e.g. height, length

Data Sample

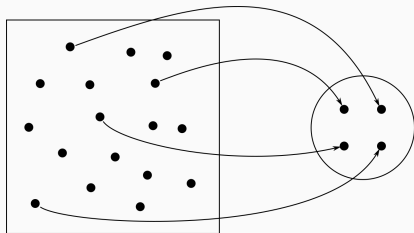
A **data sample** is a subset of a statistical population, selected according to some procedure. The elements of a sample are called **sample points** or **observations**.

Exemplary sampling methods:

- Simple random sampling (SRS)
- Stratified sampling
- Cluster sampling

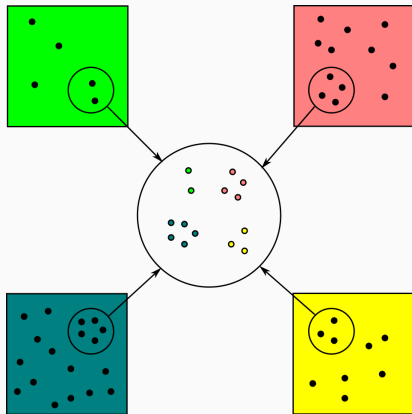
Simple Random Sampling (SRS)

- Each element has an equal probability of selection
- Variance between individual results within the sample is a good indicator of variance in the overall population
- Sample may not reflect the composition of the population (especially true for small samples)



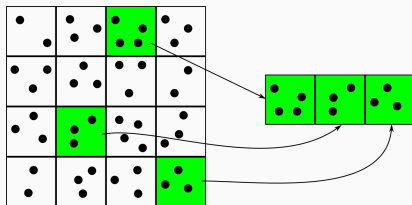
Stratified Sampling

- Population is divided into independent categories, called *strata*
- Random elements are selected from each *stratum* and added to the sample
- Ensures that each *stratum* is represented in the sample
- However, sometimes it is difficult to select relevant *strata*



Cluster Sampling

- Population is divided into groups (*clusters*), e.g. by geography
- Sample is composed of randomly selected *clusters*
- Clustering can reduce travel and administrative costs (e.g. in the case of surveys)
- Larger samples required than in the case of SRS



Probability

- Measure of the likelihood that an event will occur
- Quantified between 0 (impossibility) and 1 (certainty)
- In a random and well-defined experiment it is calculated as the number of desired outcomes divided by the total number of all outcomes

Example

When tossing a fair coin once, the probability P of getting an outcome of "head-head" is 0.25. There are 4 possible outcomes: "head-head", "head-tail", "tail-head", "tail-tail".

Probability

Probability of getting a specific number N in rolling a six-sided die:



N	1	2	3	4	5	6
$P(N)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Probability of getting a specific sum $S = N_1 + N_2$ in rolling two six-sided dice:



S	1	2	3	4	5	6	7	8	9	10	11	12
$P(S)$	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Example

What are the possible outcomes (N_1, N_2) for $S = 7$?

Random Variable

Random Variable

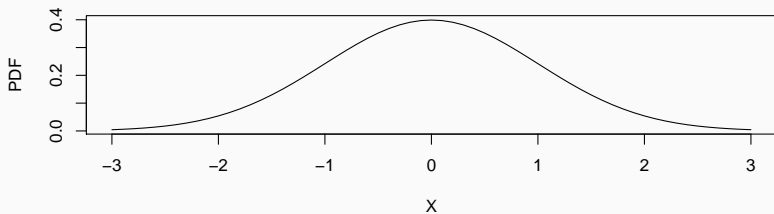
- Outcome of a random process
- More strictly speaking – a function mapping the outcomes of a random process to numerical values
- It has a probability distribution specifying the probability of each outcome

Example

Tossing a coin is a random process. The outcome (H or T) is a random variable. Each outcome has a specific probability of occurring (in this case 0.5, 0.5). The labels (H, T) can be replaced with numerical values, e.g. 0 and 1.

Distribution

Probability distribution function (PDF)



Probability mass function (PMF)



Expected Value

Probability-weighted average for all possible values:

$$E(X) = x_1p_1 + x_2p_2 + \dots + x_np_n \quad (1)$$

Arithmetic average in case of equally probable events:

$$E(X) = \frac{1}{N} \sum x_i \quad (2)$$

Expected value is sometimes called population mean and denoted with μ .

Example

What is the expected value in rolling a six-sided die?

Expected Value

Probability-weighted average for all possible values:

$$E(X) = x_1p_1 + x_2p_2 + \dots + x_np_n \quad (1)$$

Arithmetic average in case of equally probable events:

$$E(X) = \frac{1}{N} \sum x_i \quad (2)$$

Expected value is sometimes called population mean and denoted with μ .

Example

What is the expected value in rolling a six-sided die?

Answer: $(1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$

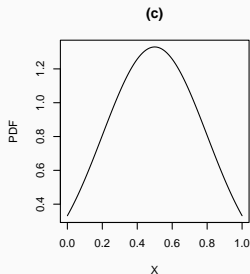
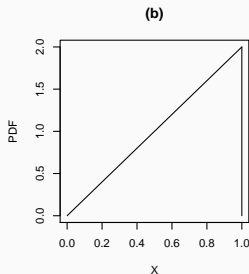
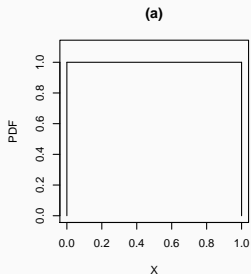
Expected Value for Continuous PDF

If $f(x)$ is the PDF of X , then

$$E(X) = \int x f(x) dx$$

Example

What is the expected value in the following cases?



Test Data Set

Data set to be used when practicing calculations of standard deviation, variance, and covariance:

X	Y
-1.24	-1.84
0.43	0.54
-0.65	-1.00
0.18	0.34
-0.57	-0.88

Standard Deviation

Standard deviation is a measure of the dispersion of the data set, i.e. it quantifies how far the data points are from the mean.

Sample standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}} \quad (3)$$

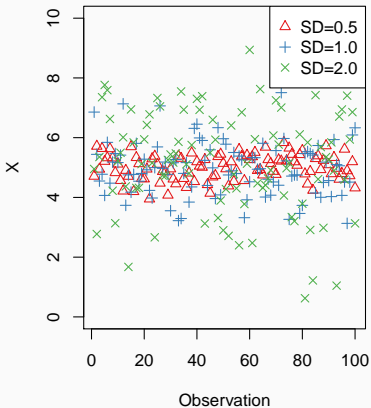
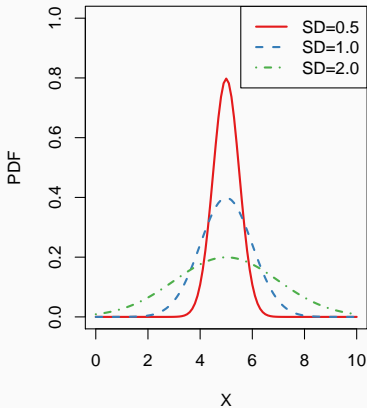
Population standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}} \quad (4)$$

Example

Calculate for X and Y from the test data set.

Standard Deviation



Standard deviation in different data sets with the same mean

Variance

Variance ($\text{Var}(X)$, σ^2 , s^2) is also a measure of the dispersion in the data set. It is the expected value of the squared deviation of a random variable from its mean:

$$\text{Var}(X) = \sigma^2 = E[(X - \mu)^2] = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad (5)$$

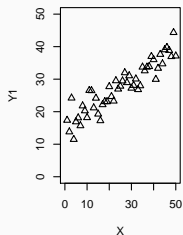
Example

Calculate for X and Y from the test data set.

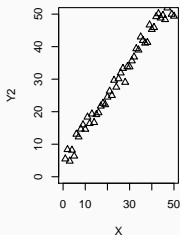
Covariance

Covariance is a measure of the joint variability of two random variables.

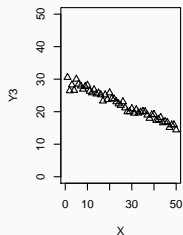
$\text{cov}(Y1, Y1) = 57.52$



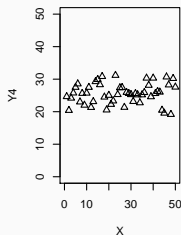
$\text{cov}(Y1, Y2) = 101.41$



$\text{cov}(Y1, Y3) = -28.96$



$\text{cov}(Y1, Y4) = 3.09$



Covariance

Covariance is calculated as the expected product of deviations of X and Y from their individual expected values:

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = \quad (6)$$

$$= \frac{1}{N} \sum_{i=1}^N (x_i - E[X])(y_i - E[Y]) \quad (7)$$

Example

Calculate $\text{Cov}(X, Y)$ based on the test data set.

Part 1: The basics of programming in R, based on Swirl,
<https://swirlstats.com/students.html>

- Basic Building Blocks
- Workspace and Files
- Sequences and Numbers
- Vectors

Part 2: Guided by the teacher

- Calculating the mean, variance, standard deviation, covariance

Read the following chapters of “R for Data Science”:

- <http://r4ds.had.co.nz/introduction.html>
- <http://r4ds.had.co.nz/explore-intro.html>

Install “tidyverse”, as described in Section 1.4.3.

Probability

Complementary Events

The probability of all possible outcomes in an experiment is 1.

If \bar{A} is a complement event to A then

$$P(\bar{A}) = 1 - P(A) \quad (8)$$

Example

When tossing a fair coin the probability of getting “head” (H) is related with the probability of getting “tail” (T):

$$P(H) = 1 - P(T) = 1 - 0.5 = 0.5$$



Independent Events

If events A and B are independent then the joint probability is

$$P(A \text{ and } B) = P(A \cap B) = P(A)P(B) \quad (9)$$

Example

What is the probability of getting an outcome (6, 6) in rolling a six-sided die twice?



Mutually Exclusive Events

If events A and B are mutually exclusive (disjoint) then the probability of **both** occurring is

$$P(A \text{ and } B) = P(A \cap B) = 0 \quad (10)$$

However, the probability of **either** occurring is

$$\begin{aligned} P(A \text{ or } B) &= P(A \cup B) = \\ &= P(A) + P(B) - P(A \cap B) = \\ &= P(A) + P(B) - 0 = \\ &= P(A) + P(B) \end{aligned} \quad (11)$$

Mutually Exclusive Events

Example

What is the probability of getting an outcome 1 or 3 in rolling a six-sided die?



Mutually Exclusive Events

Example

What is the probability of getting an outcome 1 or 3 in rolling a six-sided die?



Example

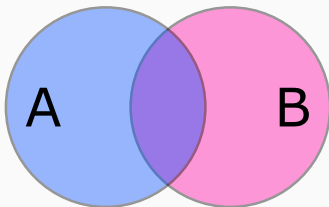
What is the probability of getting a sum of 11 in rolling two six-sided dice?



Not Mutually Exclusive Events

If two events are not mutually exclusive then $P(A \cap B) \neq 0$ and

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (12)$$























































Not Mutually Exclusive Events

Example

For example, when drawing a single card at random from a regular deck of cards, the chance of getting a heart or a face card (J, Q, K) (or one that is both) is $\frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{11}{26}$.

Example set of 52 playing cards; 13 of each suit clubs, diamonds, hearts, and spades

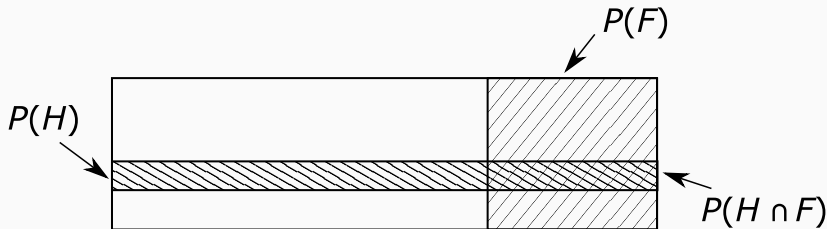
	Ace	2	3	4	5	6	7	8	9	10	Jack	Queen	King
Clubs													
Diamonds													
Hearts													
Spades													

Source: https://en.wikipedia.org/wiki/Standard_52-card_deck

Not Mutually Exclusive Events

Example

For example, when drawing a single card at random from a regular deck of cards, the chance of getting a heart or a face card (J, Q, K) (or one that is both) is $\frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{11}{26}$.



Not Mutually Exclusive Events

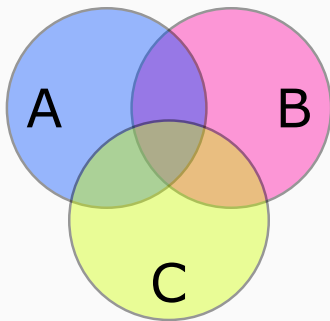
Example

When drawing a single card from a regular deck of cards, what is the probability of getting a spade OR an ace?

Example

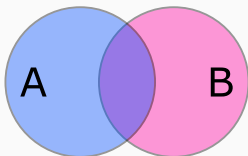
When drawing a single card from a regular deck of cards, what is the probability of getting a spade OR a diamond OR an ace?

Joint Probability of Three Sets



$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \cap B) - P(A \cap C) - P(C \cap B) \quad (13) \\ &\quad + P(A \cap B \cap C) \end{aligned}$$

Conditional Probability



If A depends on B then the probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (14)$$

Note that if A and B are independent then

$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A), \quad (15)$$

confirming that B has no effect on A .

Conditional Probability

Example

Meteorological observations for wind speed w and atmospheric pressure p_{atm} were taken in a period of 1000 days (one measurement per day).

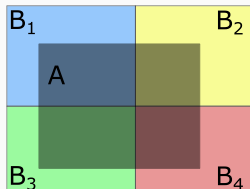
Let A stand for $w < 6$ m/s, \bar{A} for $w \geq 6$ m/s, B for $p_{atm} < 1000$ hPa, and \bar{B} for $p_{atm} \geq 1000$ hPa. The measurements were as follows:

	A	\bar{A}	Sum
B	400	100	500
\bar{B}	200	300	500
Sum	600	400	1000

Based on the measurements, is wind speed dependent on the atmospheric pressure?

Hint: Show that $P(A) \neq P(A|B)$.

The Law of Total Probability



The total (marginal) probability can be calculated from all possible conditional probabilities:

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i) \quad (16)$$

Another useful form:

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$$

The Law of Total Probability: Graphical Explanation

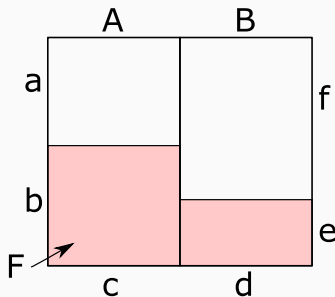
$$P(A) = \frac{(a+b)c}{(a+b)(c+d)} = \frac{c}{c+d}$$

$$P(A|F) = \frac{bc}{bc+de}$$

$$P(F) = \frac{bc+de}{(a+b)(c+d)}$$

$$P(A|\bar{F}) = \frac{ac}{ac+df}$$

$$P(\bar{F}) = \frac{ac+df}{(a+b)(c+d)}$$



$$\begin{aligned} P(A) &= P(A|F)P(F) + P(A|\bar{F})P(\bar{F}) = \\ &= \frac{bc}{(a+b)(c+d)} + \frac{ac}{(a+b)(c+d)} = \\ &= \frac{ac+bc}{(a+b)(c+d)} = \frac{(a+b)c}{(a+b)(c+d)} = \frac{c}{c+d} \end{aligned}$$

Bayes' Theorem

Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. Mathematically, it can be described as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)} \quad (17)$$

where A and B are events and $P(B) \neq 0$.

- $P(A|B)$ – the likelihood of A given that B is true
- $P(B|A)$ – the likelihood of B given that A is true
- $P(A)$ and $P(B)$ – marginal probabilities \rightarrow Eq. (16)

Bayes' Theorem

Example

There are two factories (A and B) manufacturing the same type of a device and both releasing 5000 devices per year to the market. 2 out of 100 devices manufactured in factory A are faulty. 5 out of 1000 devices manufactured in factory B are faulty. If you buy a faulty device, what is the probability it was produced in the factory A?

	A	B	Σ
F	2	5	7
\bar{F}	98	995	1093
Σ	100	1000	1100

$$\begin{aligned}P(A|F) &= \frac{P(F|A)P(A)}{P(F)} = \\&= \frac{P(F|A)P(A)}{P(F \cap A) + P(F \cap B)} = \\&= \frac{P(F|A)P(A)}{P(F|A)P(A) + P(F|B)P(B)} = \\&= \frac{\frac{2}{100} \times \frac{1}{2}}{\frac{2}{100} \times \frac{1}{2} + \frac{5}{1000} \times \frac{1}{2}} = \frac{4}{5}\end{aligned}$$

Probability: Summary

$$P(A) \in [0, 1]$$

$$P(\bar{A}) = 1 - P(A)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = P(A) + P(B) \quad \leftarrow \text{if A and B mut. exclusive}$$

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

$$P(A \cap B) = P(A)P(B) \quad \leftarrow \text{if A and B independent}$$

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

Example

Assume that 5% of students (group 1) can answer to all exam questions, 30% (group 2) can answer 70% questions, 40% (group 3) can answer 60% questions, 25% (group 4) can answer 50% questions.

1. What is the probability that a randomly selected student can answer a question?
2. What is the probability that a student who correctly answered a question belongs to the group no. 2?

Example

Within 200 sensors, 8 are faulty. We have to randomly select 3 sensors. What is the probability that all of them will be faulty?

Example

The numbers $1, 2, 3, \dots, n$ are put in a random sequence.

1. What is the probability that the numbers 1, 2 are put next to each other and exactly in this order?
2. What is the probability that the numbers 1, 2, 3 are put next to each other and exactly in this order?

Part 1: Swirl

- Missing Values
- Subsetting Vectors
- Matrices and Data Frames

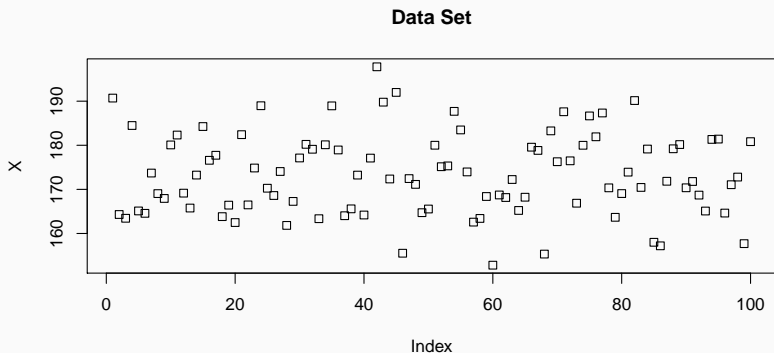
Part 2: R for Data Science

- <http://r4ds.had.co.nz/data-visualisation.html>

Distributions

Data Set

Let's consider some fundamental concepts and visualization techniques related to probability distributions using the following data set:

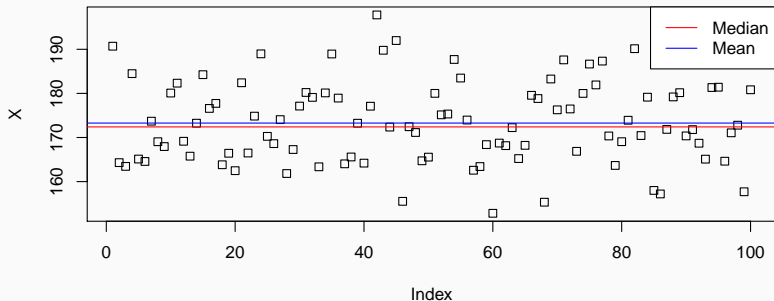


The data can represent e.g. the height (cm) of 100 randomly selected people.

Median

The **median** is the value separating an ordered data set into two equal parts. E.g. for a data set $\{1, 3, 5, 6, 7, 10, 11\}$ the median is 6.

Let's see where the median and mean are located in our data set:



In the case the number of elements in the set is even, the median is calculated as the mean of the two middlemost numbers.

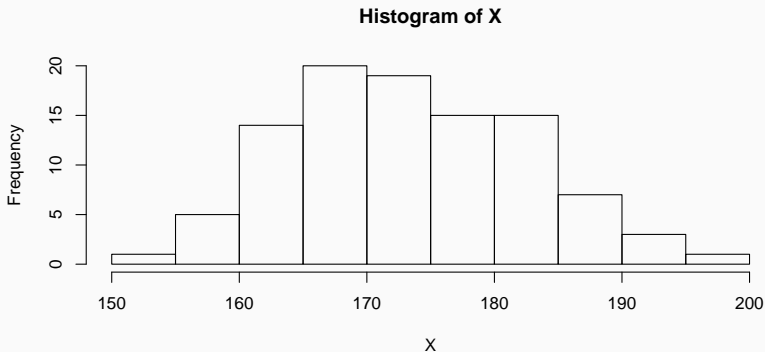
Example

- What is the median for the set $\{1, 2, 5\}$?
- What is the median for the set $\{0, 1, 2, 5\}$?
- What is the median for the set $\{0, 1, 2, 100\}$?

Answers: 2, 1.5, 1.5

Histogram

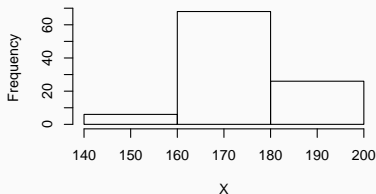
A histogram is an accurate graphical representation of the distribution of numerical data. E.g. for our data set it looks as follows:



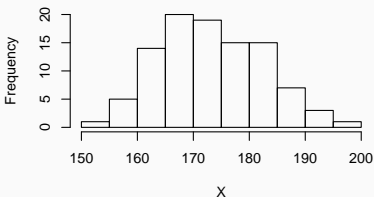
Histogram

The histogram plot depends on the chosen number of breaks:

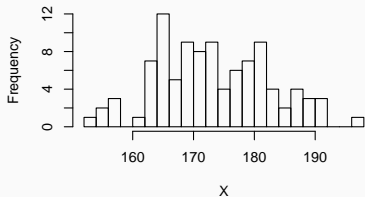
hist(X, breaks=3)



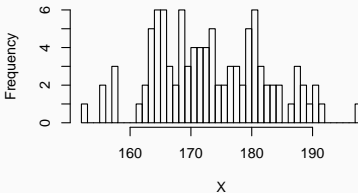
hist(X, breaks=10)



hist(X, breaks=20)



hist(X, breaks=40)



Quantile

The data set can be divided into groups of equal probability of occurring. The cut points used to generate these segments are called **quantiles**.

n -quantiles divide the data set into n groups. There are $n - 1$ n -quantiles.

E.g. the 4-quantiles (called **quartiles**) for our data set are $Q1 = 165.72$, $Q2 = 172.40$, $Q3 = 180.02$.

Another way to look at it:

Min.	Q1	Q2	Q3	Max.
0%	25%	50%	75%	100 %
152.81	165.72	172.40	180.02	197.80

Note that $Q2$ is equivalent to median.

Interquartile Range (IQR)

The **interquartile range (IQR)** is a measure of statistical dispersion, calculated as the difference of the upper and lower quartiles:

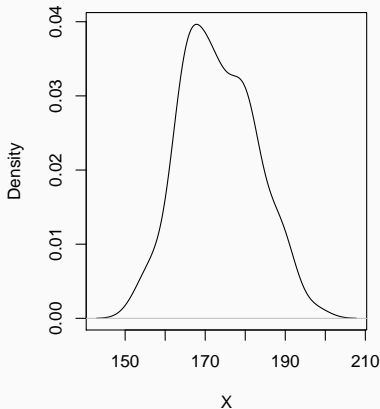
$$IQR = Q3 - Q1 = 180.02 - 165.72 = 14.30$$

It is sometimes called middle 50%.

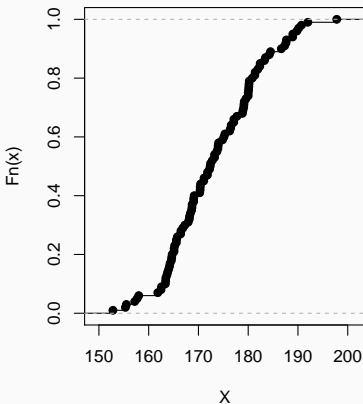
PDF vs. CDF

If we integrate PDF over X , we will get a cumulative distribution function (CDF):

Estimated PDF



Empirical CDF



Quantile Function

The quantile function is the inverse of the cumulative distribution function:

$$Q(p) = F(x)^{-1}$$

In other words, for a given p it gives x for which $P(X \leq x) = p$.

Plot a histogram, PDF, CDF and calculate the quartiles for the following data set:

```
X <- rnorm(100)
```

You can use either R Base Graphics or ggplot2.

You may need these functions: `plot`, `abline`, `hist`, `density`, `ecdf`, `quantile`, `summary`, `IQR`.

To draw subplots, you should first define the grid for subplots, e.g.:

```
par(mfrow = c(2, 1)) # two rows, one column
```

Saving plots in R:

<https://www.stat.berkeley.edu/~s133/saving.html>

- All data must be put into data frames or tibbles
- Find the corresponding plotting functions in the cheat sheet:
<https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf>

$E(X)$ and $Var(X)$ for Discrete Distribution

Example

A random variable X has the following distribution:

x	0	1	2	3	4
$P(X = x)$	0.2	0.3	0.1	0.3	0.1

- Draw the probability mass function.
- Calculate the expected value $E(X)$ and variance $Var(X)$.

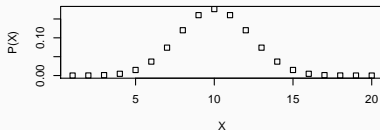
$$E(X) = \sum_i p_i x_i$$

$$Var(X) = E[(X - E(X))^2] = \sum_i p_i (x_i - \mu)^2$$

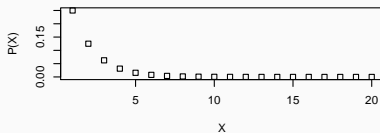
Parametric Family of Distributions

- Probabilities often depend on some numerical constants
- In such cases, it is possible to parametrize the PDF
- In other cases, parametric distributions can be used as good approximations
- Examples of parametric distributions that are used very often:
 1. binomial distribution
 2. geometric distribution
 3. normal distribution

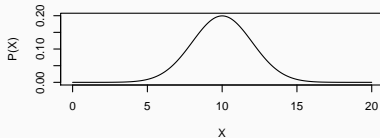
1) Binomial distribution



2) Geometric distribution



3) Normal distribution



Bernoulli Distribution

It is a distribution with just two possible outcomes, 0 and 1.

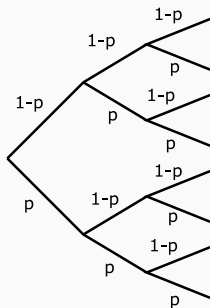
Its PMF can be expressed as follows:

$$f(k, p) = P(X = k) = \begin{cases} q = (1 - p) & \text{for } k = 0 \\ p & \text{for } k = 1 \end{cases} \quad (18)$$

Alternatively:

$$f(k, p) = P(X = k) = p^k(1 - p)^{1-k} \text{ for } k \in \{0, 1\} \quad (19)$$

Binomial Distribution



Example

The probability of a success in some trial is $p = 0.3$. Consider an experiment with 3 independent trials ($n = 3$). Calculate the probabilities that there will be exactly:

- 1 success $\rightarrow P(1) = 3p(1 - p)^2$
- 2 successes $\rightarrow P(2) = 3p^2(1 - p)$
- 3 successes $\rightarrow P(3) = p^3$

How would you calculate it for $n = 100$?

Binomial Distribution

The probability of getting k successes in n trials is given by:

$$f(k, n, p) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (20)$$

where p is the probability of a success in a single trial.

$\binom{n}{k}$ is the binomial coefficient, read as “ n choose k ”, and calculated as follows:

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \quad (21)$$

$\binom{n}{k}$ is equal to the number of subsets of size k that can be formed from a group of n distinct items.

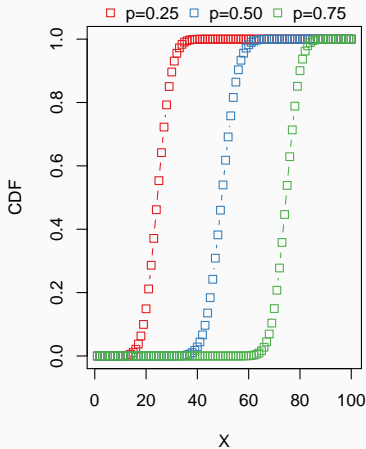
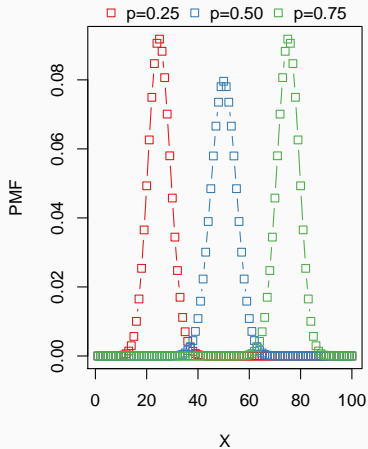
Example

How many subsets of size 2 can be selected from the set $\{1, 2, 3, 4\}$? List all subsets, then calculate by hand and compare with R.

Answer: 6

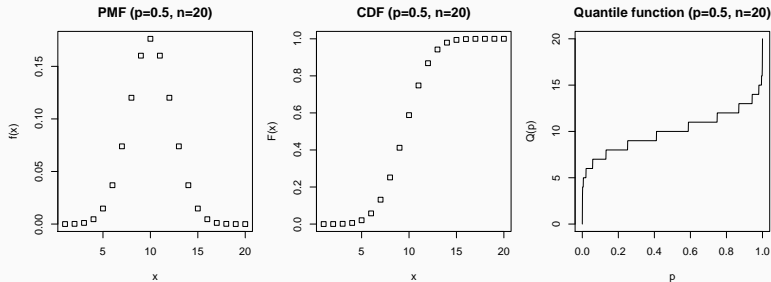
R command: `choose(n, k)`

Binomial Distribution



R commands: `dbinom`, `pbinom`, `qbinom`, `rbinom`

Binomial Distribution



- **PMF**: What is the probability that there will be $X = k$ successes? $\rightarrow f(x) = P(X = k)$
- **CDF**: What is the probability that there will be less or equal than $X = n$ successes? $\rightarrow F(x) = P(X \leq k)$
- **QF**: What is the number of successes such that the probability of getting this number or less is p ? $\rightarrow Q(p) = F^{-1}(x)$

Example

- Plot binomial distribution density experimenting with different n (number of trials) and p (probability of success per trial).
- How to calculate the most probable value of k (expected number of successes) in each case?

Binomial Distribution

Example

Draw a target on the blackboard and play the “paper cannon” game (rules to be explained during the lecture). Based on 10 trials, estimate the chance of success in hitting the target.

- Calculate the probability that a person would hit the target at least 20 times in 100 trials.
- Calculate the probability that a person would miss the target more than 20 times in 100 trials.
- Calculate the probability $P(X = 100)$ for $n = 100$.
- What is the number of successes k in 100 trials, so that $P(X > k) = 90\%$.

Hint: Use CDF and the quantile function of the binomial distribution.

Example

Draw a target on the blackboard and play the “paper cannon” game (rules to be explained during the lecture). Based on 10 trials, estimate the chance of success in hitting the target. Subsequently, calculate the probability that a person would miss the target more than 20 times in 100 trials.

Hint: Use CDF of the binomial distribution

Geometric Distribution

The **geometric distribution** describes the probability that a given trial is the first one successful in a series of trials.

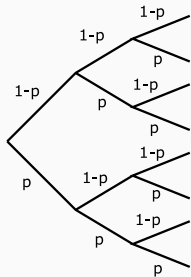
There are two possible definitions. The probability that the k th trial is the first success is given by Eq. (22). The probability that there will be k failures before the first success is given by Eq. (23).

$$P(X = k) = (1 - p)^{k-1}p, \quad k = 1, 2, 3, \dots \quad (22)$$

$$P(X = k) = (1 - p)^k p, \quad k = 0, 1, 2, \dots \quad (23)$$

Eq. (23) is used in R functions: `dgeom`, `pgeom`, `qgeom`, `rgeom`.

Geometric Distribution



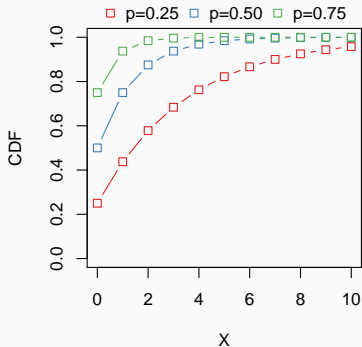
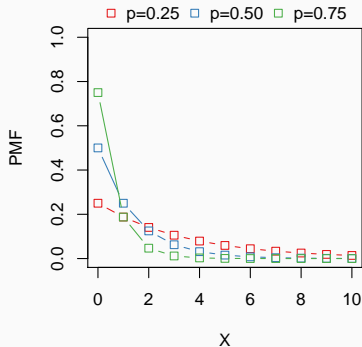
Example

Calculate the probabilities that:

- the 1st trial is the first success
- the 2nd trial is the first success
- the 3rd trial is the first success

Geometric Distribution

Number of failures before the first success, based on Eq. (23):



R commands: `dgeom`, `pgeom`, `qgeom`, `rgeom`

Example

Given $p = 0.3$, calculate the following:

- $P(X \leq 4)$
- $P(X = 10)$
- Find k such that $P(X \leq k) = 0.5$.
- How many trials are needed to be 90% sure we get at least one success?

Hint: The number of trials needed to get at least one success is one larger than the number of preceding failures.

Poisson Distribution

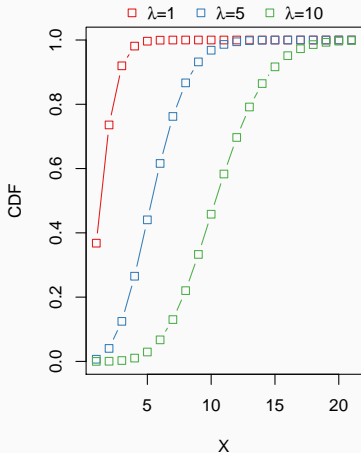
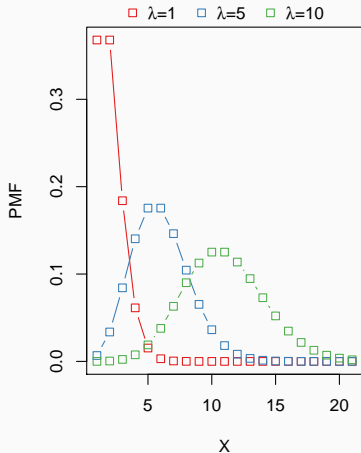
The **Poisson distribution** is a discrete probability distribution (like binomial and geometric), expressing the probability of a given number of events k occurring in a fixed interval of time. It's assumed that the events occur at a constant rate λ .

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (24)$$

Examples:

- the number of phone calls received by a call center per hour
- the number of computer technical failures in a data center per week

Poisson Distribution



R commands: `dpois`, `ppois`, `qpois`, `rpois`

Example

In a data center computer failures occur at an average rate of 100 per week.

(a) What is the probability that the number of failures next week will be less than 80?

$$P(X < 80) = P(0, 1, 2, \dots, 79) = \sum_{i=0}^{79} P(i)$$

Calculate in R using: (1) for loop, (2) ppois function.

(b) What is the probability $P(X = 100)$?

Example

A one-hundred-year flood is a flood event that has a 1% probability of occurring in any given year.

- What is the probability that there will be no one-hundred-year flood within the next 100 years?
- What is the number of floods k so that $P(X > k) = 1\%$?

Normal Distribution

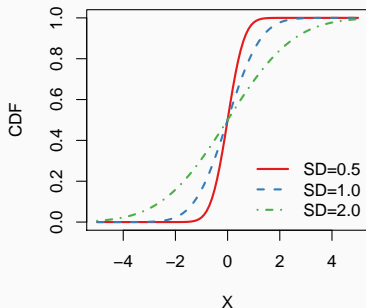
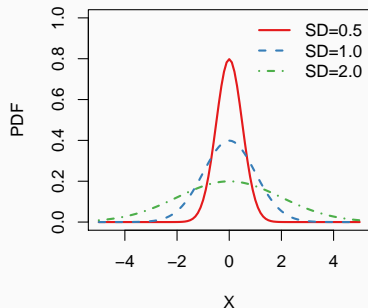
The **normal (Gaussian) distribution** is arguably the most often used probability distribution in statistics. It is a continuous distribution.

It is defined with two parameters: mean μ and standard deviation σ (or variance σ^2). If a random variable X is normally distributed, we write $X \sim \mathcal{N}(\mu, \sigma)$.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (25)$$

If $\mu = 0$ and $\sigma = 1$, the function is often referred to as the standard normal distribution: $f(x|\mu = 0, \sigma = 1) = \frac{1}{\sqrt{2\pi}} e^{-0.5x^2}$

Normal Distribution



Example

Calculate $P(X > 0.5)$ and $P((X > 0.5) \cup (X < -0.5))$ if $X \sim \mathcal{N}(\mu = 0, \sigma = 0.5)$

Normal Distribution

Example

Run the following code:

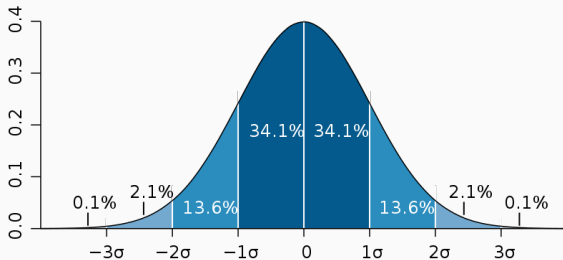
```
curve(exp(-x^2), from = -3, to = 3)
```

Discuss the meaning of the parameters and scaling factors in the normal distribution: $\sqrt{\pi}$, $\frac{1}{2}$, σ^2 , μ .

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

$$f(x|\mu = 0, \sigma = 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

Normal Distribution: 68-95-99.7 Rule



$$P(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 0.68$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997$$

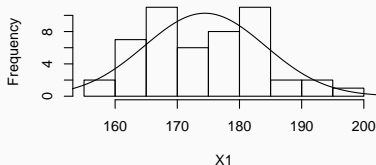
Example

Consider a random variable $X \sim \mathcal{N}(\mu = 0, \sigma = 1)$.

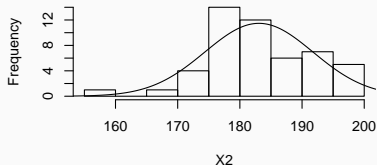
- What is the probability $P(X = 0)$?
- What is the probability $P(-1 < X < 1)$?
- What is the probability $P((X < -3) \cup (X > 3))$?
- What is the probability $P(X > 3)$?

Evaluating the Normal Approximation

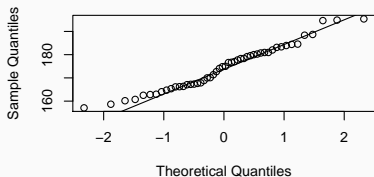
Histogram of X1



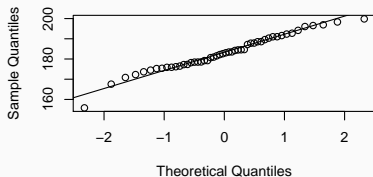
Histogram of X2



Normal Q-Q Plot



Normal Q-Q Plot



R commands: `hist`, `qqnorm`, `qqline`

Evaluating the Normal Approximation: R Code

```
# Exemplary data sets
X1 <- rbeta(50, shape1 = 3, shape2 = 5) * 60 + 150
X2 <- rnorm(50, mean = 180, sd = 10)

# Plot
windows(width = 8, height = 5)
par(mfrow = c(2, 2))

h <- hist(X1)
mlp <- mean(h$counts / h$density) # Multiplier for the norm. dist. func.
xn <- seq(150, 210, by = 0.1)
yn <- dnorm(xn, mean = mean(X1), sd = sd(X1))
lines(xn, yn * mlp, type = 'l')

hist(X2)
mlp <- mean(h$counts / h$density) # Multiplier for the norm. dist. func.
xn <- seq(150, 210, by = 0.1)
yn <- dnorm(xn, mean = mean(X2), sd = sd(X2))
lines(xn, yn * mlp, type = 'l')

qqnorm(X1)
qqline(X1)

qqnorm(X2)
qqline(X2)
```

Example

Take three samples from a normal distribution: $N_1 = 10$, $N_2 = 30$, $N_3 = 100$. Evaluate the normal approximation for each sample. What is the conclusion with respect to the sample size?

Example

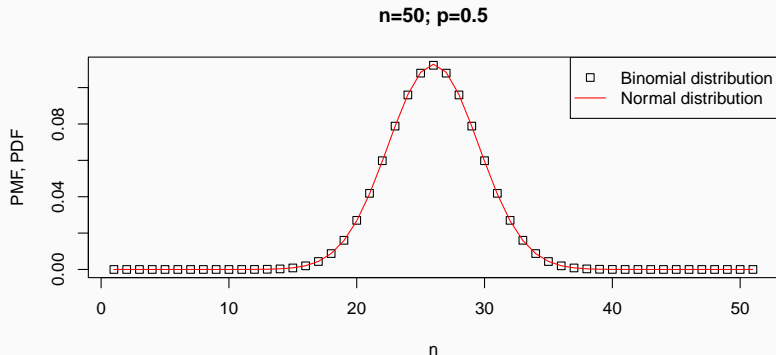
Take an anonymous sample of the height of SDU students:

<http://bit.ly/sta2018-height>

Tasks:

- Evaluate the normal approximation.
- What is the probability to observe a person higher than 2m?

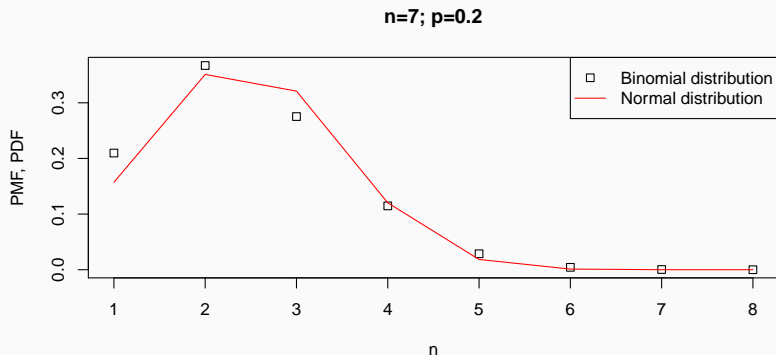
Binomial Distribution: Normal Approximation



For large n and p not too far from 0.5 the binomial distribution can be approximated with the normal distribution:

$$\mathcal{N}\left(\mu = np, \sigma = \sqrt{np(1-p)}\right).$$

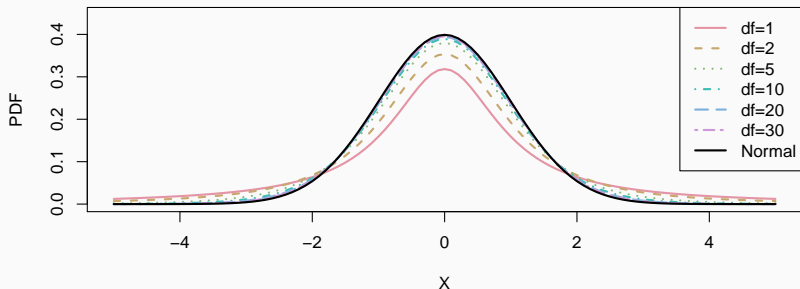
Binomial Distribution: Normal Approximation



The approximation is not accurate for small n , especially if p is far from 0.5.

t-Distribution

If the sample size is small ($n < 30$) and the population standard deviation is unknown, we often use the t-distribution instead of the normal distribution. t-distribution depends only on the number of degrees of freedom $df = n - 1$. For $df \geq 30$ it's almost indistinguishable from the standard normal distribution.



R commands: `dt`, `pt`, `qt`, `rt`

Degrees of Freedom: Intuitive Explanation

Imagine you have 4 unknown numbers (a, b, c, d) with the mean equal to 20:

$$\frac{a + b + c + d}{4} = 20.$$

We can arbitrarily suggest 3 numbers (a, b, c), but the last number (d) is not free - there is only one specific d satisfying the overall mean of 20.

In general, the degrees of freedom is equal to the number of observations minus the number of parameters to be estimated (e.g. means).

Example

If $x_1 = 0$, choose x_2 and x_3 so that $\mu_x = 0$ and $\sigma_x^2 = 1$.

Swirl:

- Logic
- Functions

R for Data Science:

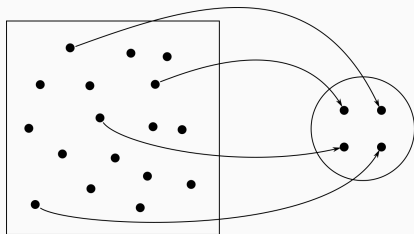
- `http://r4ds.had.co.nz/transform.html`

Statistical Inference

Statistical inference is the process of deducing the population characteristics based on the sample.

Consider the following:

- Is the sample mean and SD the true population mean and SD?
- What is the underlying distribution?
- Does the sample distribution represent well the population?
- How likely is it to observe value x ?
- What, most likely, the future observations will be?



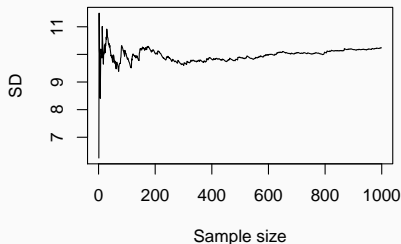
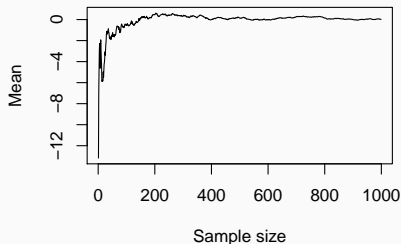
Point Estimate

A **point estimate** is the “best guess” for a population parameter based on a sample.

Example

Point estimates for the mean and SD of a population following a normal distribution $\mathcal{N}(\mu = 0, \sigma = 10)$ depending on the sample size:

Samples increased incrementally



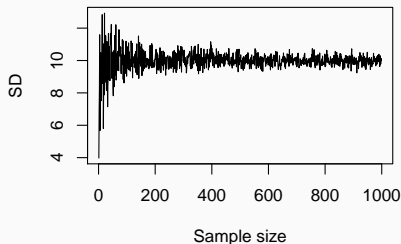
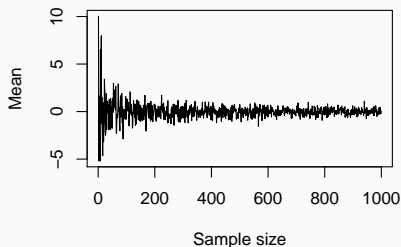
Point Estimate

A **point estimate** is the “best guess” for a population parameter based on a sample.

Example

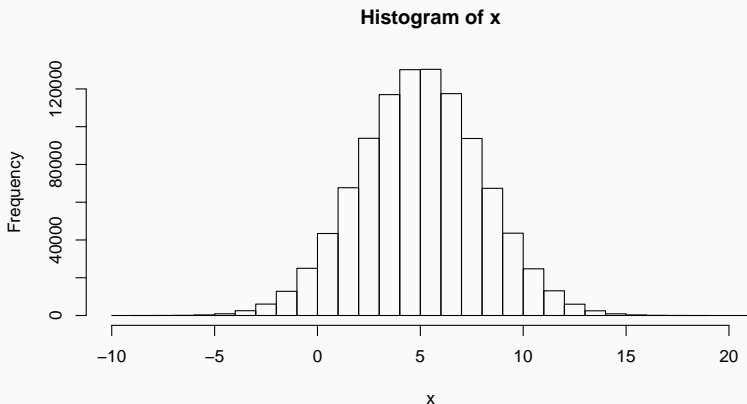
Point estimates for the mean and SD of a population following a normal distribution $\mathcal{N}(\mu = 0, \sigma = 10)$ depending on the sample size:

Samples drawn independently



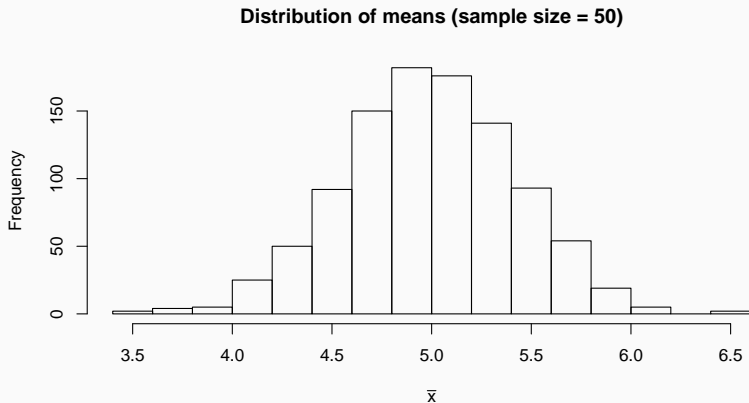
Standard Error of the Mean

Let's assume we have a population of 1 million individuals following the normal distribution $\mathcal{N}(\mu = 5, \sigma = 3)$. This is the true histogram of the population:



Standard Error of the Mean

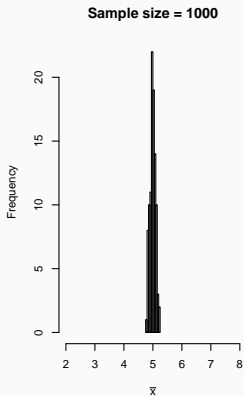
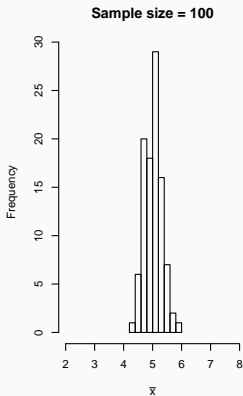
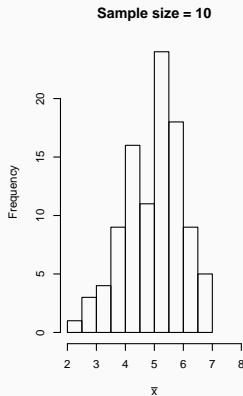
We draw 1000 independent samples of size 50, and for each sample calculate the mean. This is the distribution of the means:



Standard Error of the Mean

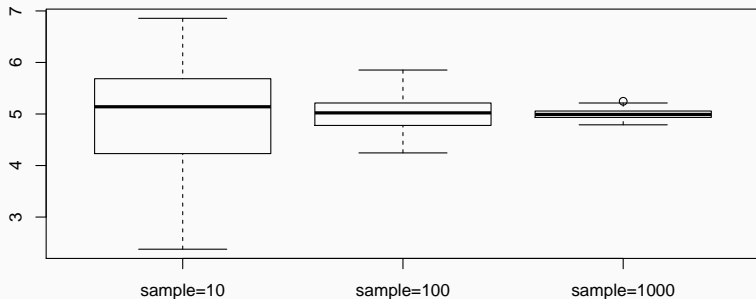
The **sampling distribution** of the mean depends on the number of samples and the sample size:

Sampling distribution of the mean (100 samples)



Standard Error of the Mean

The **sampling distribution** of the mean depends on the number of samples and the sample size:



The obtained means are similar (5.12, 4.93, 4.99), but our level of confidence increases with the increasing sample size.

Standard Error of the Mean

The **sampling distribution** of the mean depends on the number of samples and the sample size:

- the distribution becomes more “normal” with more samples
- the distribution becomes more dense (narrow) with larger samples

Standard Error of the Mean

The **standard error of the mean** ($SE_{\bar{x}}$) is the standard deviation of the sampling distribution of the mean:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} \quad (26)$$

where:

σ – population standard deviation (often unknown),

s – sample standard deviation,

n – sample size.

The samples should be independent. A reliable method to ensure that is by choosing samples consisting of less than 10% of the population.

Central Limit Theorem

Central Limit Theorem (classical definition)

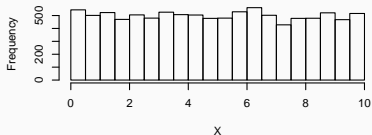
The sum of independent and identically distributed random variables tends toward a normal distribution.

Central Limit Theorem (informal description w.r.t. the mean)

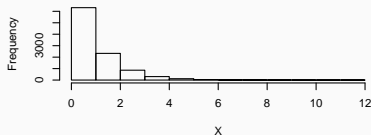
The sampling distribution of the mean \bar{x} is approximately normal, even if X is not normally distributed. The approximation improves with larger sample sizes.

Examining the Central Limit Theorem

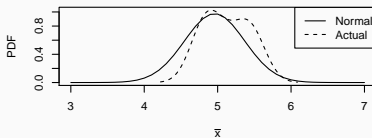
Uniform distribution



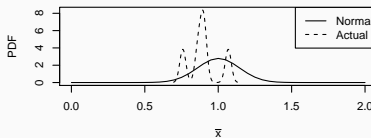
Exponential distribution



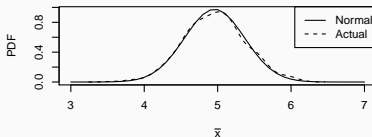
5 samples of size 50



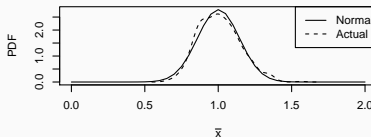
5 samples of size 50 from an exponential distribution



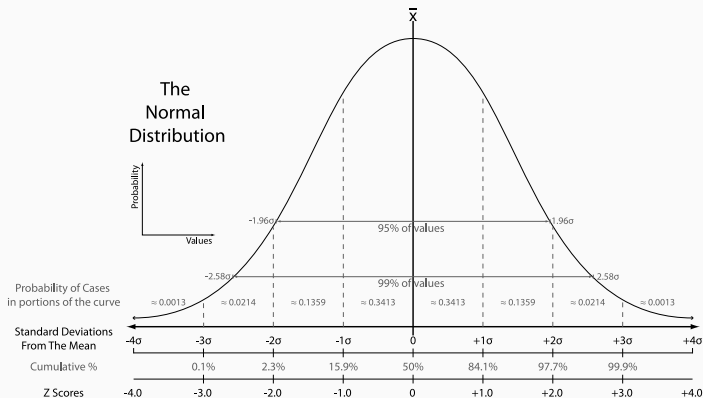
500 samples of size 50



500 samples of size 50 from an exponential distribution



Standardized Random Variable (Z-score)



$$Z = \frac{X - \mu}{\sigma} \quad (27)$$

Image based on: https://commons.wikimedia.org/wiki/File:The_Normal_Distribution.svg

Standardized Random Variable (Z-score)

Example

Run the following code:

```
pop_size = 1000
X <- rnorm(pop_size, mean = 10, sd = 2)
Z <- (X - mean(X)) / sd(X)
plot(density(Z))
xgrid <- seq(-4, 4, by = 0.1, lty=1)
lines(xgrid, dnorm(xgrid), lty=2)
```

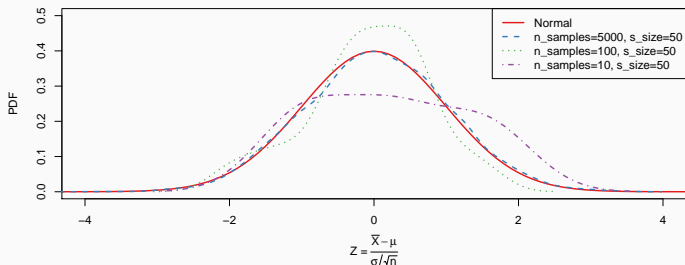
Try with different population sizes (`pop_size`).

Standardized Sampling Distribution of the Mean

Let \bar{x} be a sampling distribution of the mean from a population $X \sim \mathcal{N}(\mu, \sigma)$. The standardized version of \bar{x} is:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - \mu}{SE_{\bar{x}}} \quad (28)$$

Z converges in distribution to $\mathcal{N}(\mu = 0, \sigma = 1)$ with the increasing number of samples. The sample size n does not affect the distribution of Z (because it's normalized).



Confidence Intervals

A plausible range of values for the population parameter is called a **confidence interval (CI)**. It's calculated using the point estimate and the standard error. The $c\%$ confidence interval is given by:

$$\text{point estimate} \pm z^* \times SE \quad (29)$$

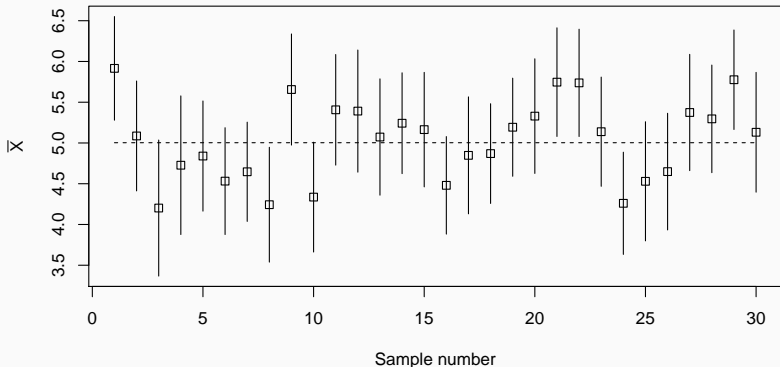
c	90%	95%	99%
z^*	1.6	2.0	2.6

$z^* \times SE$ is called **the margin of error**.

Note: The given values for z^* are approximate. You can calculate the exact ones for any $c\%$ in R.

Confidence Intervals

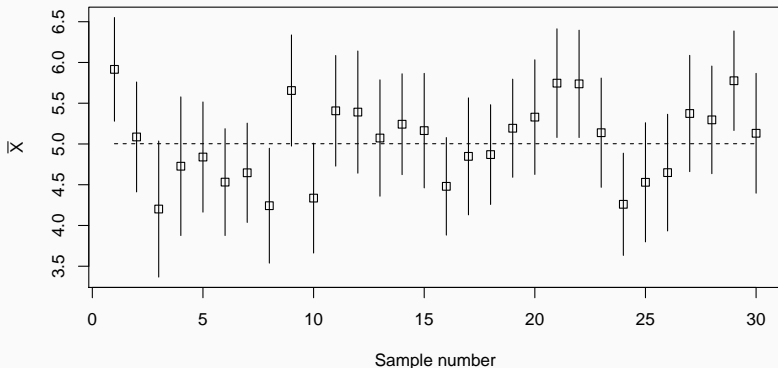
An example of 90% CI calculated for 30 samples (dashed line – true mean):



If the sampling is repeated many times, in 90% of cases the CI will encompass the true mean.

Confidence Intervals

An example of 90% CI calculated for 30 samples (dashed line – true mean):



~~There is a 90% probability that the true mean lies within a specific CI~~
→ NOT TRUE (the true mean is not a random variable)

Example

Calculate a 95% CI for the mean based on the following observations *:

14.6, 9.0, 11.5, 17.6, 13.2

Use the standard normal distribution. Consider two cases:

1. Estimate σ using the sample standard deviation s **.
2. Use the true standard deviation $\sigma = 5$.

* The sample was generated with the following command: `rnorm(5, mean = 10, sd = 5)`

** Read the docs: `?sd`

Confidence Intervals When σ is Unknown

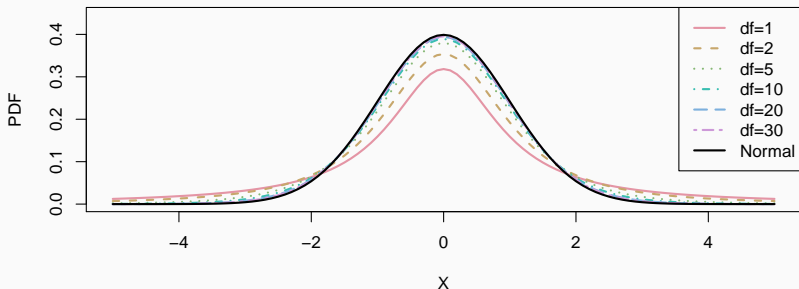
In the case when we don't know the population standard deviation σ and the sample size is small (< 100), we can't use the normal distribution. Instead, we use the t -distribution:

$$\text{point estimate} \pm t_{df}^* \times SE \quad (30)$$

where the degrees of freedom $df = n - 1$, and n is the sample size. The standard error SE is calculated as previously.

t-Distribution (Repetition)

If the sample size is small ($n < 30$) and the population standard deviation is unknown, we often use the t-distribution instead of the normal distribution. t-distribution depends only on the number of degrees of freedom $df = n - 1$. For $df \geq 30$ it's almost indistinguishable from the standard normal distribution.



R commands: `dt`, `pt`, `qt`, `rt`

Confidence Intervals

Example

Calculate the exact z^* values for the following confidence levels: 90%, 95%, 99%, 99.9%. Use the standard normal distribution and the t-distribution. Assume the sample size $s = 5$.

$c\%$	90%	95%	99%	99.9%
z^*				
t_{df}^*				

Confidence Intervals Based on the t -Distribution

Example

Calculate 95% and 99% CIs for the mean based on the following observations:

14.6, 9.0, 11.5, 17.6, 13.2

Use the t -distribution.

Hint: Use `qt(p, df)`

Confidence Intervals: Practical Example

Example

Using the previously collected data about the height of SDU students (<http://bit.ly/sta2018-height>), calculate 95% and 99% CIs for the mean height.

Compare the CIs based on the t-distribution and the normal distribution (assume that the sample standard deviation is the population standard deviation). Interpret the results.

Example

True or false?

- Confidence intervals describe the probability that the true mean lie within the interval.
- 99% confidence interval is wider than a 95% confidence interval.
- When a sample follows the normal distribution, we should use the normal distribution to calculate the confidence interval.
- Confidence interval gets narrower with the increasing size of the samples.
- Confidence interval gets narrower with the increasing number of samples.

Hypothesis testing is a statistical technique for comparing two exclusive hypotheses: a null hypothesis H_0 and an alternative hypothesis H_A . E.g.:

- H_0 : the new drug and the old drug are equally effective
- H_A : the new drug is better than the old

The null hypothesis H_0 often represents a skeptical perspective or a claim to be tested. The null hypothesis H_0 is not rejected unless the evidence in favor of the alternative hypothesis H_A is strong.

Hypothesis Testing

Hypotheses can be described mathematically, e.g.:

$$\left. \begin{array}{l} H_0 : \mu = x \\ H_A : \mu \neq x \end{array} \right\} \text{double-sided test}$$

or

$$\left. \begin{array}{l} H_0 : \mu = x \\ H_A : \mu > x \end{array} \right\} \text{single-sided test}$$

Hypotheses can be tested with:

1. Confidence intervals (suitable for double-sided tests)
2. p-values (suitable for both single- and double-sided tests)

Example

Why confidence intervals are not suitable for single-sided tests?

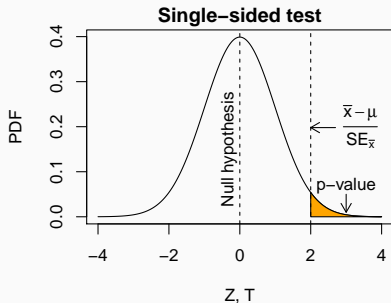
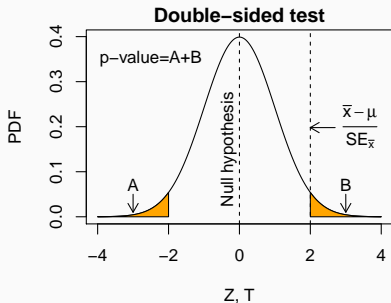
Give an example of a single-sided test.

Hypothesis Testing Using Confidence Intervals

1. Formulate H_0 , H_A , and the confidence level (90%, 95%, 99%)
2. Assume the model (normal distribution or t -distribution)
3. Find z^* or t_{df}^*
4. Calculate the confidence interval:
 - normal distribution $\rightarrow \mu \pm z^* \times SE_{\bar{x}}$
 - t -distribution: $\rightarrow \mu \pm t_{df}^* \times SE_{\bar{x}}$
5. Reject H_0 if the sample mean \bar{x} lies outside the confidence interval

Hypothesis Testing Using p-Value

1. Formulate H_0 , H_A , and the significance level α (e.g. 0.05)
2. Assume the model (normal distribution or t -distribution)
3. Calculate the test statistic (Z -score or t -score)
4. Calculate the p-value
5. Reject H_0 if $p\text{-value} < \alpha$



Hypothesis Testing: Factory Example #1

Example

A factory claims that the average lifetime of the sensors it produces is 3 years. A random sample of 100 sensors was taken and they worked for 2.8 years on average with the standard deviation of 0.9 years.

Construct the null and alternative hypotheses and evaluate whether there is a sufficient evidence to reject the null hypothesis.

Use both: (a) confidence intervals, (b) p-values.

Factory Example #1: Hypotheses

The hypotheses could be as follows:

$$H_0 : \mu = 3$$

$$H_A : \mu \neq 3$$

Alternatively, if we expect that the actual sensor lifetime is shorter than 3 years:

$$H_0 : \mu = 3$$

$$H_A : \mu < 3$$

Factory Example #1: Confidence Intervals

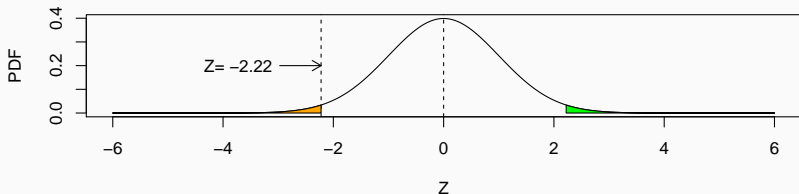
1. The sample is sufficiently large to use the normal model.
2. The population standard deviation can be approximated with the sample standard deviation.
3. The sample likely consists of less than 10% of the population, so we can assume that the observations are independent.
4. Let's assume a 95% confidence interval $\rightarrow z^* = 2$.

$$3 \pm z^* \times SE_{\bar{x}} \rightarrow 3 \pm 2 \times \frac{s}{\sqrt{n}} \rightarrow 3 \pm 2 \times \frac{0.9}{\sqrt{100}} \rightarrow 3 \pm 0.18$$

The sample mean $\bar{x} = 2.8$ is outside the confidence interval, so we have **sufficient evidence to reject the null hypothesis**.

Factory Example #1: p-Values

Let's keep assumptions 1-3 from the previous slide. We'll use the significance level $\alpha = 0.05$.



Single-sided test ($H_A : \mu < 3$):

$$p\text{-value} = 0.013 < \alpha = 0.05 \rightarrow \text{Reject } H_0$$

Double-sided test ($H_A : \mu \neq 3$):

$$p\text{-value} = 0.026 < \alpha = 0.05 \rightarrow \text{Reject } H_0$$

Hypothesis Testing: Factory Example #2

Example

A factory produces 100 thousands sensors per year. A sample of 200 sensors was taken for extensive field tests and it turned out the sensors worked with a sufficient accuracy for 20 months on average, with a standard deviation of 5 months. After that time they needed a recalibration.

After introducing stricter quality assurance (QA) procedures in the factory, another sample of 200 sensors was taken and this time the average time of accurate work was 27 months with a standard deviation of 3 months.

Determine whether the results give a statistically significant evidence that the new QA procedures improved the average sensor lifetime. Perform the test based on p-values.

- `http:`
`//r4ds.had.co.nz/exploratory-data-analysis.html`
- `http://r4ds.had.co.nz/workflow-projects.html`

Linear Regression

Pearson Correlation Coefficient

The Pearson correlation coefficient for a population:

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (31)$$

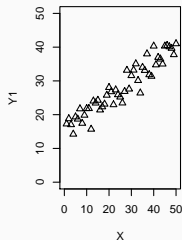
The Pearson correlation coefficient for a sample:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (32)$$

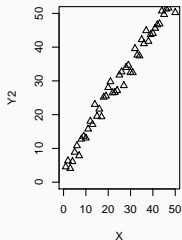
ρ and r have values between +1 and -1, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

Pearson Correlation Coefficient

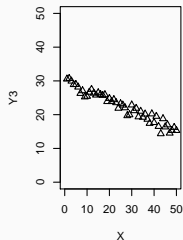
$r(Y1, Y1) = 1.00$
 $cov(Y1, Y1) = 59.16$



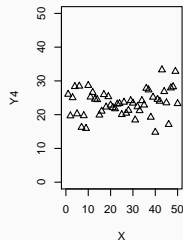
$r(Y1, Y2) = 0.94$
 $cov(Y1, Y2) = 108.02$



$r(Y1, Y3) = -0.94$
 $cov(Y1, Y3) = -32.52$



$r(Y1, Y4) = 0.09$
 $cov(Y1, Y4) = 2.60$



Pearson correlation coefficient vs. covariance

R commands: `cor`, `cov`

Linear Regression

Linear regression is a simple approach for predicting quantitative responses using linear models.

E.g. linear regression can be used to answer the following questions:

1. Is there a relationship between advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media (TV, radio, newspapers) contribute to sales?
4. How accurately can we estimate the effect of each medium on sales?
5. How accurately can we predict future sales?

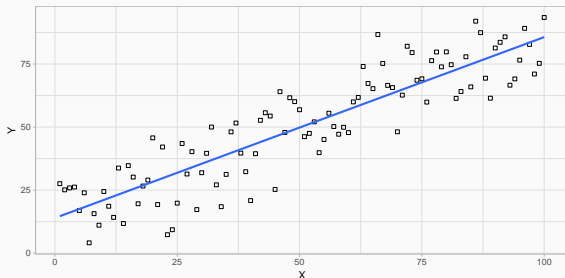
Simple Linear Regression

For the prediction of a single quantitative response Y based on a single predictor variable X , we can assume a linear relationship between them:

$$Y \approx \beta_0 + \beta_1 X \quad (33)$$

There are two coefficients (or parameters) in this model:

β_0 – intercept, β_1 – slope. The true coefficients are unknown, but we can estimate them.



Estimating the Coefficients: Ordinary Least Squares

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the i th value of X . Then the i th residual is given by:

$$e_i = y_i - \hat{y}_i,$$

and the residual sum of squares (RSS) is as follows:

$$\begin{aligned} \text{RSS} = S(\hat{\beta}_0, \hat{\beta}_1) &= e_1^2 + e_2^2 + \dots + e_n^2 = \\ &= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 = \\ &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{aligned}$$

In order to get the coefficients estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ we need to minimize $S(\hat{\beta}_0, \hat{\beta}_1)$.

Estimating the Coefficients: Ordinary Least Squares

We know that $S(\hat{\beta}_0, \hat{\beta}_1)$ is at the minimum when:

$$\frac{\partial S}{\partial \hat{\beta}_0} = 0, \quad \frac{\partial S}{\partial \hat{\beta}_1} = 0$$

We need to solve the following system of equations:

$$\begin{cases} \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0 \\ \frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0 \end{cases}$$

After some calculus we get the following solution:

$$\begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases} \quad (34)$$

Estimating the Coefficients: Ordinary Least Squares

Solving for β_0 :

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^n (y_i^2 - 2y_i\beta_0 - 2y_ix_i\beta_1 + \beta_0^2 + 2x_i\beta_0\beta_1 + x_i^2\beta_1^2) = 0$$

$$\sum_{i=1}^n (-2y_i + 2\beta_0 + 2\beta_1 x_i) = 0$$

$$\sum_{i=1}^n \beta_0 = \sum_{i=1}^n (y_i - \beta_1 x_i)$$

$$\beta_0 = \frac{\sum_{i=1}^n y_i}{n} - \beta_1 \frac{\sum_{i=1}^n x_i}{n}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Estimating the Coefficients: Ordinary Least Squares

Solving for β_1 :

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^n (y_i^2 - 2y_i\beta_0 - 2y_ix_i\beta_1 + \beta_0^2 + 2x_i\beta_0\beta_1 + x_i^2\beta_1^2) = 0$$

After substituting β_0 we get:

$$\begin{aligned} \frac{\partial}{\partial \beta_0} \sum_{i=1}^n (y_i^2 - 2y_i\bar{y} + 2y_i\bar{x}\beta_1 - 2y_ix_i\beta_1 + \bar{y}^2 - 2\bar{x}\bar{y}\beta_1 + \bar{x}^2\beta_1^2 + \\ + 2x_i\bar{y}\beta_1 - 2x_i\bar{x}\beta_1^2 + x_i^2\beta_1^2) = 0 \end{aligned}$$

$$\beta_1 \sum_{i=1}^n (\bar{x}^2 - 2\bar{x}x_i + x_i^2) = \sum_{i=1}^n (y_ix_i + \bar{x}\bar{y} + y_i\bar{x} - x_i\bar{y})$$

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

Linear Regression Example

Example

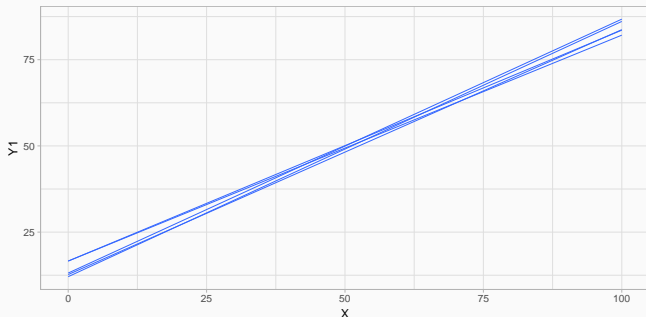
Generate a sample X from the normal distribution $\mathcal{N}(\mu = 10, \sigma = 2)$. Generate a sample $Y = 0.5X + \epsilon$, where $\epsilon \sim \mathcal{N}(\mu = 0, \sigma = 1)$.

Calculate the linear regression coefficients (Eq. 34, slide 123) and make a plot visualizing the correlation.

Assessing the Accuracy of the Coefficient Estimates

All these lines were generated from the same model, using five different samples:

$$Y = 15 + 0.5X + \epsilon, \quad \epsilon \sim \mathcal{N}(\mu = 0, \sigma = 10)$$



Which line describes best the real population?

Assessing the Accuracy of the Coefficient Estimates

$\hat{\beta}_0$ and $\hat{\beta}_1$ are only estimates of the real coefficients β_0 and β_1 :

$$Y = \beta_0 + \beta_1 X + \epsilon$$

If we assume that ϵ is a zero-mean random error with the standard deviation σ_ϵ , then we can derive formulas for the standard errors of $\hat{\beta}_0$, $\hat{\beta}_1$:

$$SE_{\hat{\beta}_0} = \sqrt{\sigma_\epsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} \quad (35)$$

$$SE_{\hat{\beta}_1} = \sqrt{\frac{\sigma_\epsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (36)$$

Assessing the Accuracy of the Coefficient Estimates

In general, σ_ϵ is not known, but can be estimated from the data:

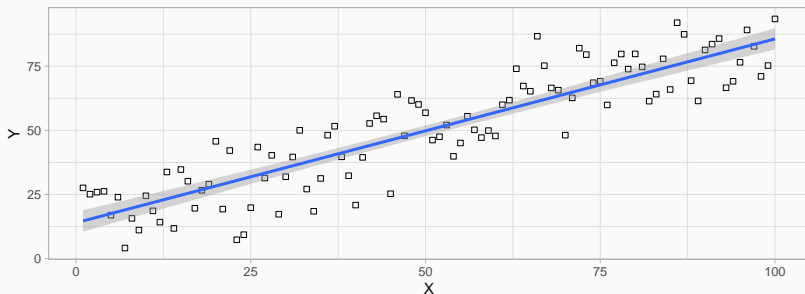
$$\sigma_\epsilon \approx \text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}} = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}$$

RSE is known as the residual standard error and RSS is the residual sum of squares.

Assessing the Accuracy of the Coefficient Estimates

Since we know the standard errors, we can define 95% intervals for the linear regression coefficients:

$$\left(\hat{\beta}_0 \pm 2 \times SE_{\hat{\beta}_0}, \quad \hat{\beta}_1 \pm 2 \times SE_{\hat{\beta}_1} \right)$$



Linear regression line with 95% confidence interval for $\hat{\beta}_0$, $\hat{\beta}_1$

Linear Regression in R

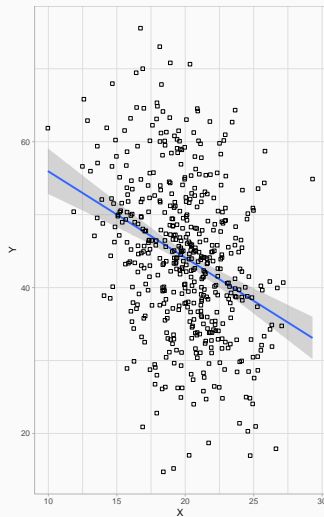
Try the following code:

```
# Define X and Y
X <- rnorm(500, mean = 20, sd = 3)
Y <- -1.3 * X + 70 + rnorm(length(X),
                             mean = 0,
                             sd = 10)

# Put X, Y in a data frame
df <- tibble(X = X, Y = Y)

# Plot the regression line using ggplot2
ggplot(df, mapping = aes(x = X, y = Y)) +
  geom_smooth(method = 'lm', se = T) +
  geom_point()

# Fit a linear model using lm()
linmod <- lm(Y ~ X)
print(linmod)
```



Example

1. Are the estimated coefficients close to the true coefficients?
2. Plot the histogram of residuals for the previously fitted model.
Is the distribution of residuals close to normal? Is the mean of residuals close to zero?
3. What happens if you reduce/increase the noise in Y ?
4. Fit a linear model to a nonlinear relationship:

```
X <- rnorm(500, mean = 1, sd = 3)
```

```
Y <- X^2 + rnorm(length(X), mean = 10, sd = 10)
```

Is the distribution of residuals close to normal? Is the mean of residuals close to zero? Why?

Hypothesis Tests on the Coefficients

Standard errors of the coefficients can be used in hypothesis testing for relationships between variables. E.g.:

- H_0 : There is no relationship between X and $Y \Rightarrow \beta_1 = 0$
- H_A : There is some relationship between X and $Y \Rightarrow \beta_1 \neq 0$

We need to determine whether $\hat{\beta}_1$ is sufficiently far from zero, taking into account the uncertainty $SE_{\hat{\beta}_1}$, before we can be confident that the real β_1 is non-zero.

Hypothesis Tests on the Coefficients

First, we compute the t -statistic:

$$t = \frac{\hat{\beta}_1 - 0}{SE_{\hat{\beta}_1}}$$

Afterwards, we compute the p-value and compare it with the assumed significance level α (e.g. 0.05). The p-value is computed based on the t -distribution with $df = n - \nu$ degrees of freedom, where ν is the number of model coefficients. $\nu = 2$ in the simple linear regression.

If $\text{p-value} < \alpha$, we reject the null hypothesis.

Hypothesis Tests on the Coefficients

Analyze the following summary:

```
> linmod <- lm(Y ~ X)
> summary(linmod)

Call:
lm(formula = Y ~ X)

Residuals:
    Min       1Q   Median       3Q      Max
-31.2473  -6.4663   0.0575   5.8874  27.5949

Coefficients:
              Estimate      Std. Error  t value    Pr(>|t|)
(Intercept)  67.7640       3.0661     22.101    < 2e-16 ***
X            -1.1834       0.1517     -7.802    3.6e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.04 on 498 degrees of freedom
Multiple R-squared:  0.1089, Adjusted R-squared:  0.1071
F-statistic: 60.88 on 1 and 498 DF,  p-value: 3.596e-14
```

Assessing the Accuracy of the Model

There are two main statistics used in assessing the accuracy of a linear model:

- RSE - absolute measure of error
- R^2 - relative measure of error

R^2 describes the proportion of variance explained. It always takes a value between 0 and 1.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \quad (37)$$

where $TSS = \sum (y_i - \bar{y})^2$ is the total sum of squares.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

TSS corresponds to the amount of variability inherent in the response before the regression is performed

RSS corresponds to the amount of variability that is left unexplained after performing the regression

In example, if $Y = f(X) = \beta_0 + \beta_1 X$ and $R^2 = 0.72$,
then 72% of variability in Y can be explained using X .

Multiple Linear Regression

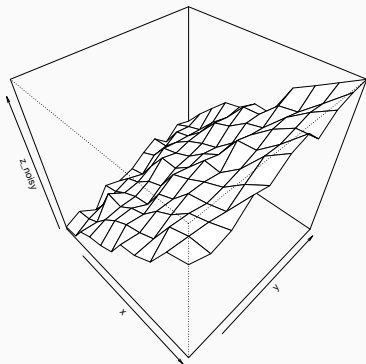
It is possible to include multiple predictors in the linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$$

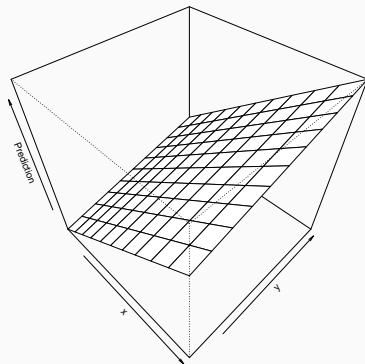
In R:

```
m <- lm(Y ~ X1 + X2 + X3)
```

Multiple Linear Regression: Example With 2 Predictors



Actual data



Fitted linear model

F-statistic

In addition to the p-values calculated for each predictor, the `lm()` function calculates a so-called the F-statistic. The F-statistic is used to test the following hypothesis:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_a : \text{at least one } \beta_j \text{ is non-zero}$$

F-statistic is calculated based on the F-distribution, which is most notably used in the analysis of variance (ANOVA - to be introduced later).

If the p-value associated with the F-statistic is close to zero, we have a strong evidence that at least one coefficient is non-zero. This insight is especially useful if there are many (hundreds or thousands) coefficients.

E.g. if there are 100 predictors and $H_0 : \beta_1 = \beta_2 = \dots = \beta_{100} = 0$ is true, then about 5% of the p-values will be below 0.05 simply by chance!

The F-statistic doesn't suffer from this problem.

Predictor Selection

Most often the response variable is dependent on a subset of predictors.

To find the appropriate predictors:

1. Examine the F-statistic and the associated p-value
2. Use expert knowledge
3. Test different models and compare R^2 (manual testing feasible only for small number of predictors)

Testing of different model configurations can be automated using:

- Forward selection (start with none, add 1-by-1)
- Backward selection (start with all, remove 1-by-1)
- Mixed selection (start with none, add 1-by-1, remove if p-value grows too much)

Making Predictions

Once the model coefficients are estimated, it is straightforward to use the model. In R the function for feeding new data to a model is called `predict.lm()` (see also `predict()`).

```
# Training
X1 <- rnorm(10)
X2 <- rnorm(10)
Y <- 2 * X1 - 0.7 * X2
m <- lm(Y ~ X1 + X2)

# Prediction
df <- data.frame(X1 = 1, X2 = 1)
Y_predict <- predict(m, df) # Returns 1.3
```

Polynomial Regression

Polynomial regression is a special form of regression analysis in which the relationship between the outcome and predictor variables is modelled using a polynomial, e.g.:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1 X_2$$

Even though the polynomial regression fits a non-linear function to data, the coefficients are estimated in the same way as in the multiple linear regression, because the regression function is linear with respect to the parameters β_j .

Polynomial Regression

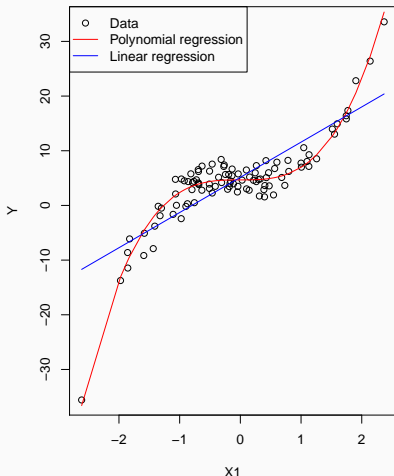
In example, to find the coefficients for the following model:

$$Y = \beta_0 + \beta_1 X_1^3$$

use the following command in R:

```
lm(Y ~ I(X1^3))
```

I() has to be used, because ^ has a special meaning in formulas.
Type ?formula for more info.



Formulas in Polynomial Regression

R formulas might seem peculiar at first, but they are actually very handy, especially in the cases with many predictor variables:

Function	Regression formula
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$	<code>lm(Y ~ X1 + I(X^2))</code>
$Y = \beta_0 + \beta_1 X_1 X_2$	<code>lm(Y ~ X1:X2)</code>
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$	<code>lm(Y ~ X1*X2)</code>
$Y = \beta_1 X^3$	<code>lm(Y ~ I(X^3) - 1)</code>
$Y = \beta_0 + \sum_i \beta_i X_i$	<code>lm(Y ~ ., data = df)</code>

Energy Efficiency Data Set

Example

- Download the following data set:
http:
`//archive.ics.uci.edu/ml/datasets/Energy+efficiency`
- Choose predictor variables and fit a linear model. Try to minimize R^2 .
- Plot model predictions vs. validation data.
- Optional: divide the data set into training (80%) and validation subsets (20%) (random sample!).

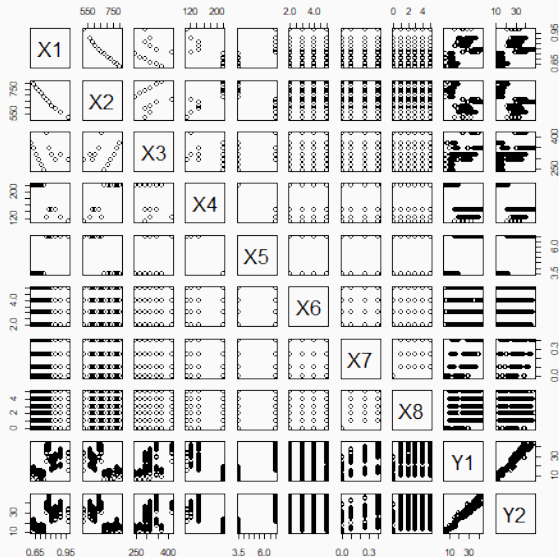
Related paper: A. Tsanas, A. Xifara: 'Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools', Energy and Buildings, Vol. 49, pp. 560-567, 2012

Energy Efficiency Data Set

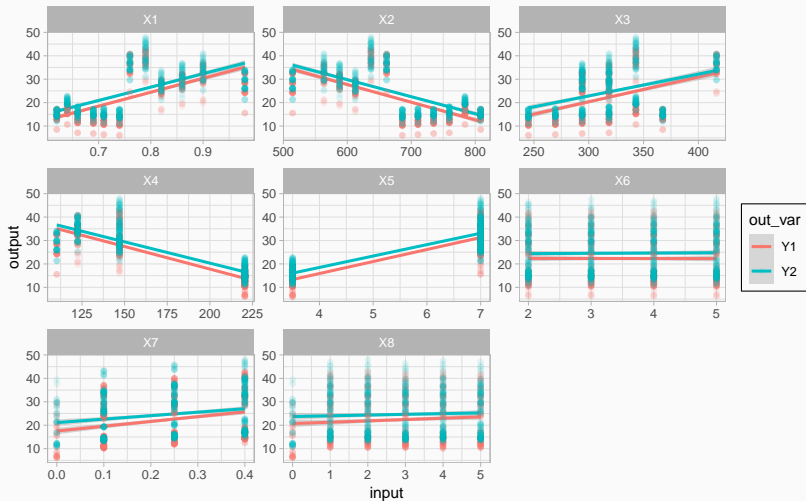
Variables:

- X1** Relative compactness
- X2** Surface area
- X3** Wall area
- X4** Roof area
- X5** Overall height
- X6** Orientation
- X7** Glazing area
- X8** Glazing area distribution
- Y1** Heating load
- Y2** Cooling load

Energy Efficiency Data Set



Energy Efficiency Data Set



Simple linear regression is insufficient. Can you get better results with multiple regression?

Analysis of Variance

This is a brief introduction to a one-way ANOVA test.

For an extensive description of ANOVA please see Chapters 2-3 of *Statistical Analysis with the General Linear Model*, by Miller and Haden,
<https://web.psy.otago.ac.nz/miller/StatsBook.htm>.

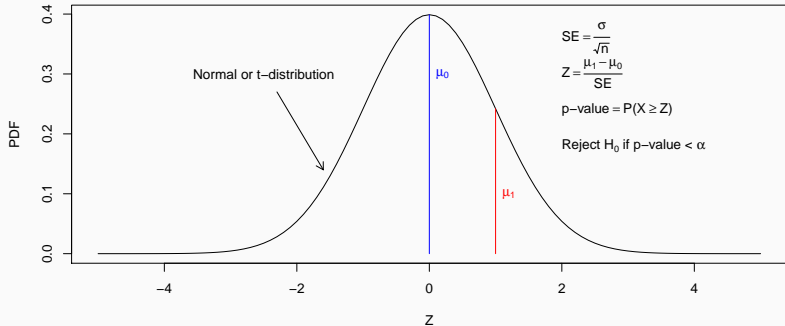
Hypothesis Testing: Repetition

In its simplest form, the hypothesis testing methodology is used to check whether two means are equal or not:

$$H_0 : \mu_0 = \mu_1$$

$$H_1 : \mu_0 \neq \mu_1$$

Summary of the hypothesis testing procedure ($H_0: \mu_0 = \mu_1$)



Analysis of Variance (ANOVA)

ANOVA, in its simplest form (so-called one-way ANOVA), is used to check whether more than two means are equal or not:

$$H_0 : \mu_0 = \mu_1 = \dots = \mu_n$$

$$H_1 : \text{at least one is different}$$

This can be assessed by analyzing and comparing:

- the variability within groups
- the variability between groups

ANOVA: Example With 3 Groups

Data:

```
x <- rnorm(n = 30, mean = 50, sd = 5)
g1 <- x * 1.45 + rnorm(n = length(x), sd = 5)
g2 <- x * 1.50 + rnorm(n = length(x), sd = 5)
g3 <- x * 1.55 + rnorm(n = length(x), sd = 5)
```

The true means (typically unknown) are: $\bar{\mu}_{g1} = 72.5$, $\bar{\mu}_{g2} = 75$, $\bar{\mu}_{g3} = 77.5$.

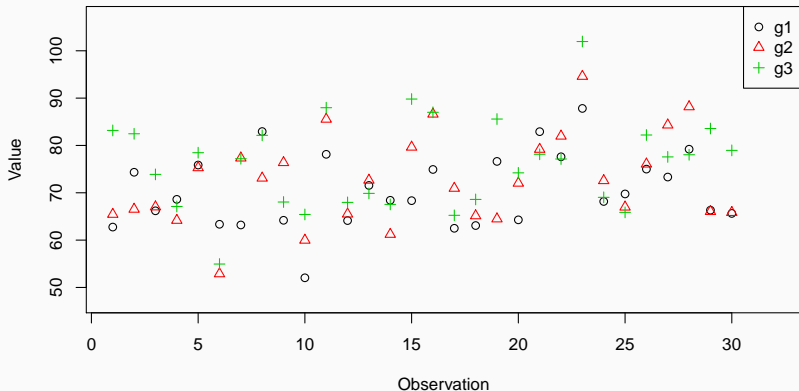
The sample means (30 observations in each group) are:

$\bar{x}_{g1} = 70.4$, $\bar{x}_{g2} = 72.6$, $\bar{x}_{g3} = 76.3$

Each group could represent e.g. a result of a specific medical treatment. In such case ANOVA can be used to check if there is a difference between 3 treatments.

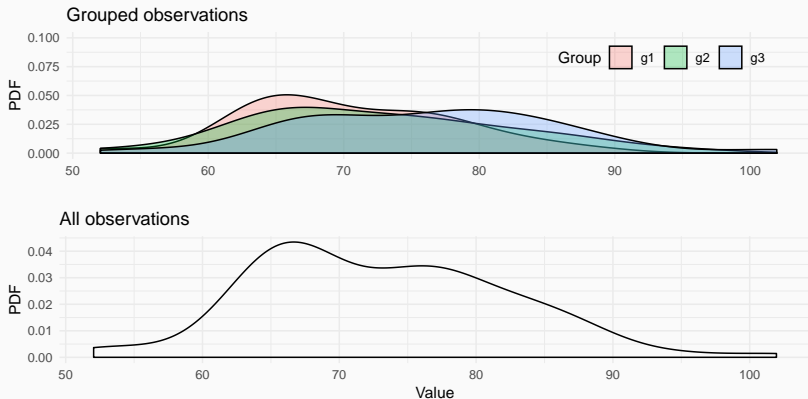
ANOVA: Example With 3 Groups

Are the group means equal?



ANOVA: Example With 3 Groups

Are the group means equal?



ANOVA: F-statistic

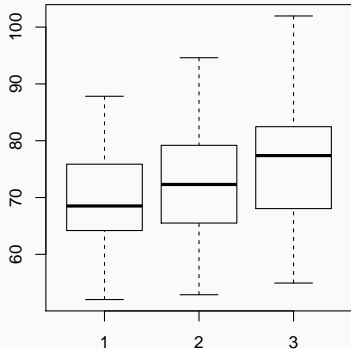
To assess whether the group means are equal, we have to compare the variability between groups with the variability within groups.

$$\begin{aligned} F &= \frac{\text{variability between groups}}{\text{variability within groups}} = \\ &= \frac{\frac{\sum_g (\bar{y}_g - \bar{y})^2}{df_1}}{\frac{\sum_g \sum_i (y_{gi} - \bar{y}_g)^2}{df_2}} \end{aligned}$$

where y_{gi} is the observation i in group g , \bar{y}_g is the mean observation in group g , \bar{y} is the mean observation in general, df_1 and df_2 are the respective degrees of freedom.

Variability Between/Within Groups

- The **variability between groups** describes how far the group means are from one another.
- The **variability within groups** describes how spread the observations are within the groups.



Degrees of Freedom

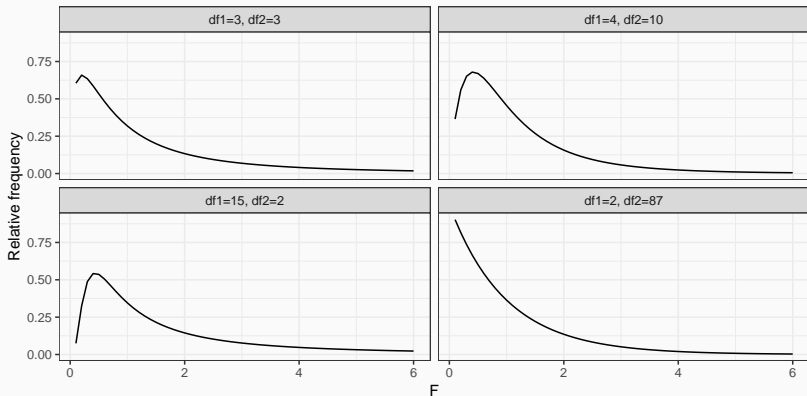
- To be able to compare the variability between and within groups, we need to normalize the nominator and denominator of F
- Both the nominator and denominator represent a sum of squares calculated from different number of elements (\sum_g vs. $\sum_g \sum_i$)
- The degrees of freedom df represents the number of independent values that the associated term can take on
- For 3 groups, the nominator $df_1 = 3 - 1 = 2$
- For 30 observations in each of the 3 groups, the denominator $df_2 = 3 \times 30 - 3 = 87$

ANOVA: Rejecting Null Hypothesis

- The F -statistic is a random variable itself and it follows a special distribution which depends on the degrees of freedom of the nominator (df_1) and the denominator (df_2).
- The null hypothesis $H_0 : \mu_0 = \mu_1 = \dots = \mu_n$ is rejected if $F > F_{critical}$.
- $F_{critical}$ is calculated based on the assumed significance level α (typically 0.05), similarly to the hypothesis testing based on the t -distribution.

F-distribution

The F -distribution depends on the degrees of freedom of the nominator (df_1) and the denominator (df_2) of the F -statistic equation. Exemplary distributions:



ANOVA: Example With 3 Groups

ANOVA summary:

```
> summary(aov(Value ~ Group, data = df))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	2	538	269.18	3.324	0.0406 *
Residuals	87	7046	80.99		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

If we assume α equal to 0.05, we can reject H_0 .

If we assume α equal to 0.01, we cannot reject H_0 .