Biostatistics 1BIO43

Pär Villner

Fall 2022

Karolinska Institutet

Table of contents

Lecture 1. Analyzing a data sample

Lecture 2. Probability distributions and discrete variables

Leture 3. The Normal distribution and inference

Lecture 4: The central limit theorem and the t-distribution

Lecture 5: Hypothesis testing

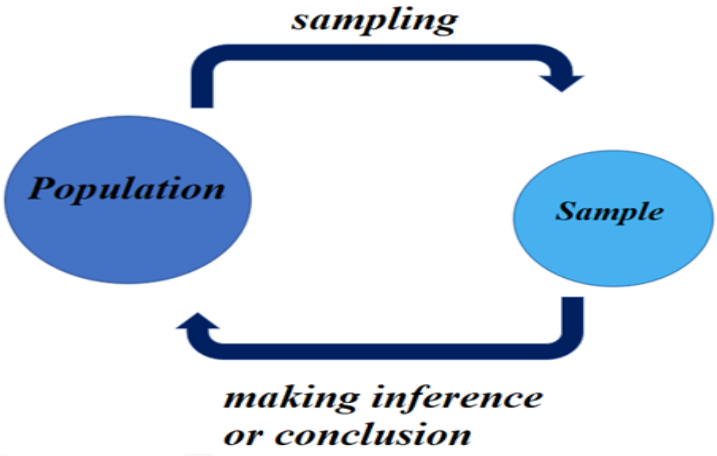
Lecture 6: Hypothesis tests and test errors

Leture 7: Non-parametric tests

Lecture 8: Contingency tables and association

Lecture 1. Analyzing a data sample

Population and sample



Population and sample

- **Population:** all items or individuals of interest.
- **Sample:** a subset of the population.

Example 1 We want to know the effect of a lung cancer treatment. The population is all lung caner patients. We select 100 patients. This is our sample.

Example 2 We want to know the thrombocyte level in my blod. The population is all the blood in my body. The sample is the blood in a syringe.

Randomness and sampling

Samples are taken randomly. This means that chance affects what we observe in our sample.

Example We randomly select 100 people to use a cancer treatment. 80 survive. We repeat the experiment with a new sample of 100 participants. 75 survive.

How can we trust scientific studies if outcomes depend on chance?

The first Randomized Clinical Trial

Studying a sample

We have collected information about ten patients.

sex	age	gene	smoke	sbp
m	56	AA	non-smoker	148
m	78	AG	non-smoker	210
m	47	AA	non-smoker	128
m	60	GG	smoker	159
m	65	AG	smoker	176
f	48	AA	non-smoker	130
f	47	AA	non-smoker	127
f	39	AG	smoker	109
f	58	GG	smoker	153
f	95	AA	smoker	171

Summary statistics describe the data in a compact manner!

Central tendency: Mean

The **mean** is the average.

```
mean(dat1$age)
```

```
## [1] 59.3
```

```
(39 +47 +47 +48 +56 +58 +60 +95 +65 +78)/10
```

```
## [1] 59.3
```

Central tendency: Median

The **median** is a number such that half the sample is smaller and the other half is larger.

```
sort(dat1$age)
```

```
##  [1] 39 47 47 48 56 58 60 65 78 95
```

```
median(dat1$age)
```

```
## [1] 57
```

The mean is **sensitive to outliers** while the median is not.

Central tendency: Quantiles

A quantile is a numerical value such that a given proportion is lower than this values.

The 0.75 quantile of Age is a number such that 75% of the observed Ages are lower.
The 0.25 quantile of Age is an age such that 25% of the observed Ages are lower.

```
quantile(dat1$age,0.75)
```

```
##   75%
## 63.75
```

```
quantile(dat1$age,0.25)
```

```
##   25%
## 47.25
```

The 0.5 quantile is the median.

Variability: Variance and standard deviation.

The **sample variance** is the mean of the squared difference between each observed value and the sample mean. In the case of Age:

```
var(dat1$age)
```

```
## [1] 279.1222
```

```
((39-59.3)^2 + (47-59.3)^2+(47-59.3)^2+(48-59.3)^2 + (56-59.3)^2+
(58-59.3)^2+(60-59.3)^2+(95-59.3)^2+(65-59.3)^2+(78-59.3)^2)/(10-1)
```

```
## [1] 279.1222
```

The **standard deviation** is the square root of the variance.

```
sd(dat1$age)
```

```
## [1] 16.70695
```

Bonus: mathematical details on mean and variance

In general, for sample size n the sample mean of the observed values $\{x_i, x_2, ..., x_n\}$is defined

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

and the variance is defined

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}.$$

Why divide by $n-1$? Because we used information about the sample when we calculated the mean.

Variability: IQR, min and max

The interquartile range (IQR) is the distance between the 0.25 and 0.75 quantiles.

```
IQR(dat1$age)
```

```
## [1] 16.5
```

The min and max are the smallest and largest values.

```
min(dat1$age)
```

```
## [1] 39
```

```
max(dat1$age)
```

```
## [1] 95
```

Association

To describe the association between two variables X and Y, we often use **correlation**.

Correlation is always between -1 and 1.

- If close to 1: if X is high, Y tends to be high.
- If close to -1: if X is high, Y tends to be low.
- If close to 0: X says nothing about Y.

Correlation measures **linear association**.

Bonus: The math behind association

The covariance between X and Y is

$$Cov(X, Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n - 1},$$

with n being the sample size. More often, we use the correlation:

$$Cor(X, Y) = \frac{Cov(X, Y)}{sd(X)sd(Y)}$$

Proportions

With categorical variables, we report **proportions**. E.g. the proportion of males is the number of males divided by the sample size.

The proportion an be seen as a mean value. If we count each male as a 1 and each female as a 0, the proportion of males is the mean of the 1's and 0's.

In R:

```
mean(dat1$sex=="m")
```

```
## [1] 0.5
```

Tables

Frequency tables illustrate how categorical variables are distributed.

Var1	Freq
f	5
m	5

We can create a **contingency table** based on two variables to show their association.

	non-smoker	smoker
f	2	3
m	3	2

Contigency tables

Frequency tables are less useful for numerical variables, such as age:

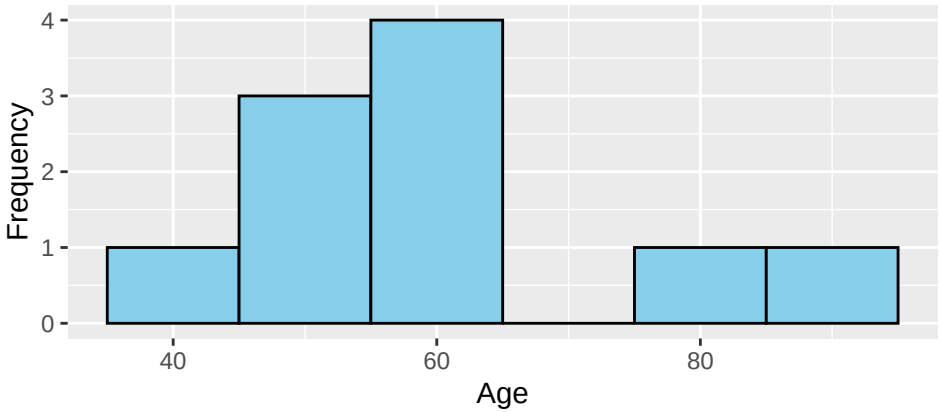
39	47	48	56	58	60	65	78	95
1	2	1	1	1	1	1	1	1

However, we can divide a numerical variable into categories.

Var1	Freq
0-50	4
51-100	6

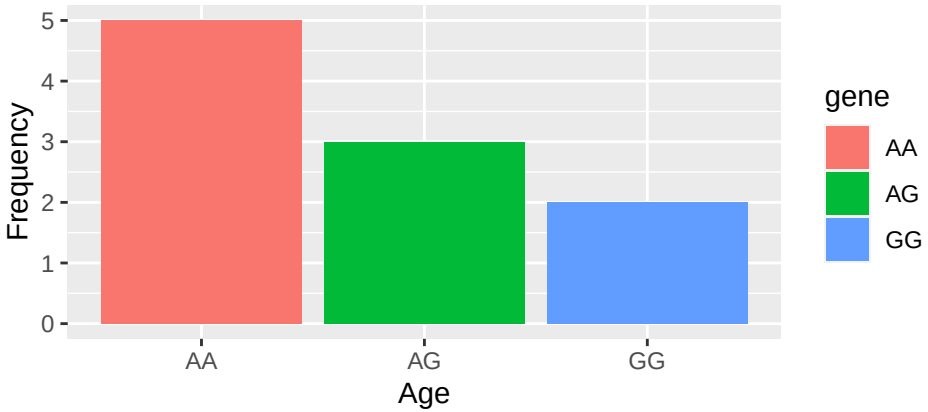
Histograms

Each bar represents the number of participants that are in a particular interval.



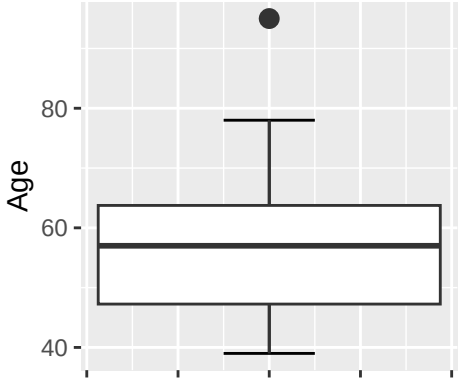
Barplots

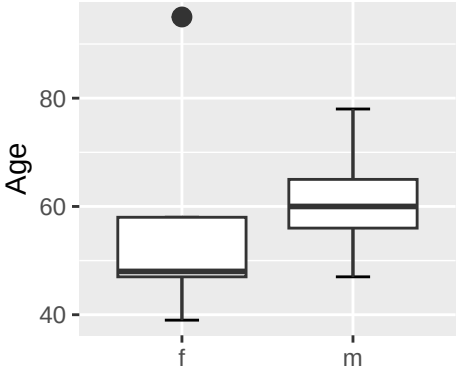
Barplots illustrate frequencty tables. Width of bars is irrelevant. Height is relevant.



Boxplots

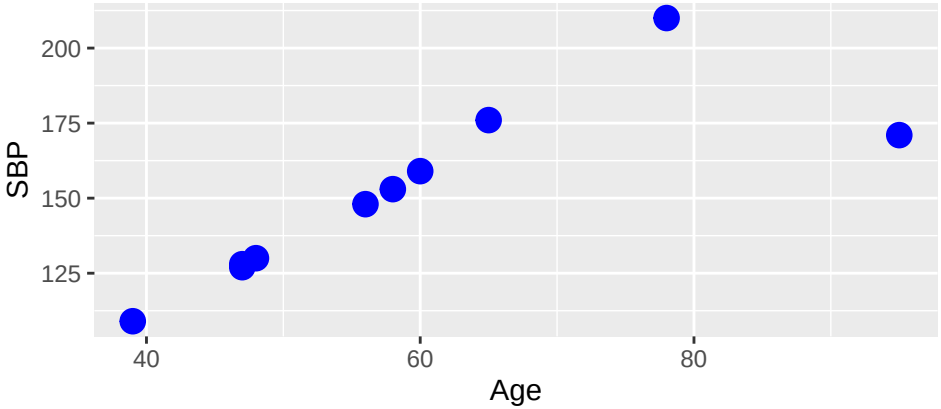
The grey box consists of the 0.25, 0.5 and 0.75 quantiles. The upper and lower whiskers are the min and max within a "reasonable interval". The small circle is an outlier.





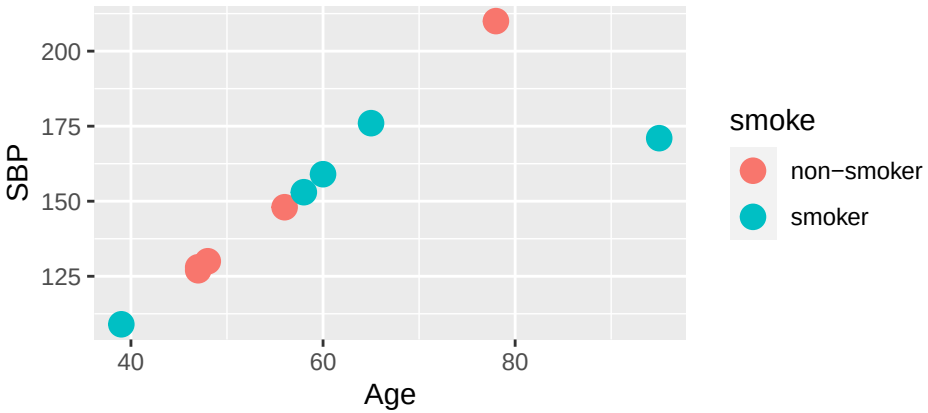
Scatterplots

Scatterplots show the relationship between two numerical variables. If it looks like we can draw a straight line through the cloud of dots, there is a linear association.



Scatterplots

You can add colours to add new dimension to the plot.



Lecture 2. Probability distributions and discrete variables

Last lecture

Last lecture we talked about:

- The role of statistics in medical science
- Populations and samples
- Different ways of analyzing a data sample: mean, variance, histograms, boxplots, scatterplots.

Can we trust research if the outcomes depend on chance, e.g. because of who/what ended up in the sample by chance?

Random variables and observed values

A **random variable** is the outcome of an experiment not yet performed. It is often denoted by an upper-case alphabetical letter

- X is the number of patients who survive in an upcoming experiment.
- Y is the mean height in a sample of Swedes we take tomorrow.

An **observed value** is the outcome of a performed experiment. It is often denoted by lower-case alphabetical letters.

- a is the number of patients who survived in our experiment.
- b is the mean height in a sample of Swedes we took today.

Outcomes and probabilities

We can assign **probabilities** to the outcomes of a random variable X.

We do this in every day life. E.g. "The probability the coin lands heads up is 0.5".

Here, our goal is to assign probabilities to outcomes of experiments in such a way that we can make sense of them, despite the randomness.

Laws of probability

Let X be a random variable, e.g X="Result of a dice roll" and a be any of the outcomes, e.g. a="Dice shows 1".

- $0 \leq P(X = a) \leq 1$
- P(At least one outcome happens)=1.
- $P(X = a) = 1 - P(X \neq a)$.

P(Dice=1)= 1/6.

P(Dice=1, 2, 3, 4, 5 or 6)=1.

P(Dice=2 or higher)=1-P(Dice=1)=5/6.

Independence

Two random variables X and Y are **independent** if the outcome of X has nothing to do with the outcome of Y.

In that case P(X=a & Y=b) = P(X=a) P(Y=b).

Example

I flip two coins. X="Result of first coin". Y="Result of second coin".

P(X=Head & Y=Head)=P(X=Head)P(Y=Head)=$0.5 \cdot 0.5 = 0.25$

Probability distribution

A **probability distribution** describes the probabilities of the outcomes of a random variable.

Sometimes the probability distribution is trivial, e.g. when I toss a die and

P(Dice=1)=P(Dice=2)=...=P(Dice=6)=1/6.

Other times, the probability distribution is less obvious. E.g. when I treat 5 cancer patients and count how many survive.

P(0 survivor)=?

P(1 survivors)=?

... and so on.

The Bernoulli distribution

We perform a single experiment with two outcomes: $X = a$ or $X \neq a$.

Example:

- I flip a coin and get Head (X=a) or Tail ($X \neq a$).
- I test for Covid and am positive (X=a) or negative ($X \neq a$).
- I undergo surgery and survive (X=a) or die ($X \neq a$).

$P(X = a) = p$ and $P(X \neq a) = 1 - p$, where $0 \leq p \leq 1$.

The Binomial distribution

Many research studies are **sequences of Bernoulli trials**.

- Out of 100 coin tosses, how many landed head?
- Out of 50 patients, how many survived the surgery?
- Out of 20 covid tests, how many were positive?

These are examples of the Binomial distribution.

The Binomial distribution

If a random variable X follows a Binomial distribution, we write it as:

$$X \sim Bin(n, p)$$

where n is the number of trials and p is the probability that event a happens in each trial.

Important

$P(X = a)$ must be the **same** in all trials, and the trials must be **independent**.

The probability calculations

I flip a coin twice and define X="Number of heads". What is $P(X = 1)$?

Seq.	Toss 1	Toss 2	# Heads	Prob.
1	Head	Head	2	0.25
2	Head	Tail	1	0.25
3	Tail	Head	1	0.25
4	Tail	Tail	0	0.25

If $P(Head) = P(Tail) = 0.5$, all four toss sequences are equally likely. I can get one head in two ways, so P(X=1)=0.25+0.25=0.5.

This is the logic behind the probability calculations. But with larger n, creating tables get tedious.

The probability mass function

We can calculate the probability of an event from the binomial distribution with the **probability mass function**:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x},$$

but you don't need to know it. In the lab, you will calculate probabilities in R with:

```
dbinom(1,2,0.5)
```

```
## [1] 0.5
```

Example: Cancer trial

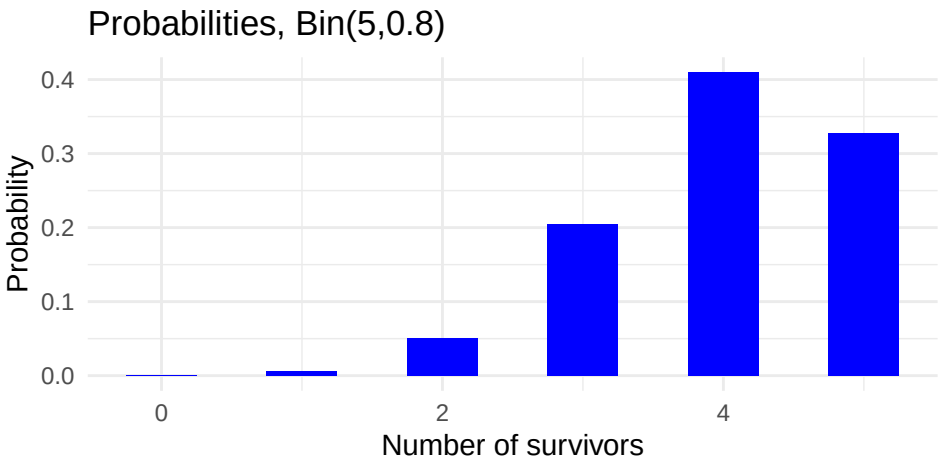
If given a cancer treatment, a patient has a 0.8 probability of survival. Five patients are given the treatment and we define X="Number of patients who survive":

$$X \sim Bin(5, 0.8).$$

Survivors	0	1.000	2.000	3.000	4.00	5.000
Prob	0	0.006	0.051	0.205	0.41	0.328

Example: Cancer trial

We can illustrate the probabilities with a graph.



The cumulative probability function

Sometimes we want to know the probability of observing a number x or smaller:

$$P(X \leq x)$$

E.g. what is the probability that two or fewer out of 5 patients survive?

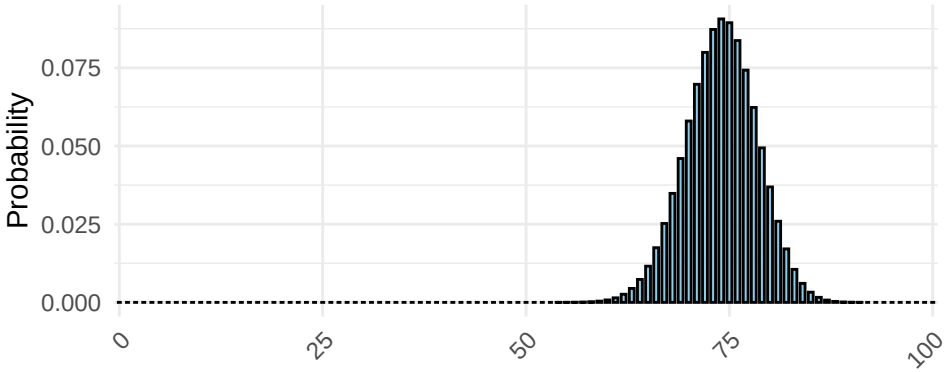
$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

```
pbinom(2,5,0.8)
```

```
## [1] 0.05792
```

Example: Mendel's experiments

Gregor Mendel claimed that if he pollinated plants in a particular way, 74% of the offspring would be tall and the rest short. Define X="Number of tall among 100 offspring" and $X \sim Bin(100, 0.74)$.



Example: Mendel's experiments

What is the probability that 70 or fewer plants are tall?

```
pbinom(70,100,0.74)
```

```
## [1] 0.2105419
```

What is the probability that more than 80 plants are tall?

```
(1-pbinom(80,100,0.74))
```

```
## [1] 0.06579331
```

What is the probability that at most 70 or that more than 80 plants are tall?

```
pbinom(70,100,0.74) + (1-pbinom(80,100,0.74))
```

```
## [1] 0.2763352
```

Estimation

We usually do not know what the probability of the event of interest (e.g. probability that a patient dies) is. We are doing research to find that out!

Based on a data sample, our best guess of probability p is the sample proportion of participants for whom a is the case.

This is an example of **inference**, and we will discuss it in detail tomorrow.

Example: Feet fungus Treatment 1

In a population of patients with feet fungus, we know that 50% recover within a week from diagnose. A treatment is available. Does it have any effect?

We select 10 patients at random and give all of them the treatment. X="Number of patients that recover within a week". If the treatment has no effect at all

$$X \sim Bin(10, 0.5).$$

We observe that all 10 patients recover within a week. If the treatment has no effect, the probability of all 10 patients recovering is

```
dbinom(10,10,0.5)
```

```
## [1] 0.0009765625
```

So either something very unlikely has happened or the treatment does have an effect!

Example: Feet Fungus Treatment 2

10 patients have feet fungus. They are given Treatment 1 on a randomly selected foot and Treatment 2 on the other. X="Number of patients well with Treatment 1 first". If the treatments are equally good

$$X \sim Bin(10, 0.5)$$

We observe that 10 of 10 patients get well first on the foot with Treatment 1. If the treatments are equally good, this has probability

```
dbinom(10,10,0.5)
```

```
## [1] 0.0009765625
```

Either something unlikely happened, or the treatments are not equally good.

Bonus: Expected value and variance

The expected value μ of a random variable is calculated:

$$\mu = E(X) = \sum_{x \in \theta_x} x \cdot p(x)$$

where θ_x is the possible values X can take. The variance σ^2 is calculated:

$$\sigma^2 = Var(X) = \sum_{x \in \theta_x} (x - \mu)^2 \cdot p(x)$$

Intuitively: The true mean and variance are like the mean and variance from an extremely large sample.

Discrete variables

A discrete variable can only take on a countable number of values.

- Integers: ... 1,2,3,4...,10,...
- Factors: Cancer, Diabetes, Covid,...
- Binary: Yes or no; True or false;...

Bonus: discrete distributions

There are many discrete distributions. Some examples:

Poisson: Often used when measuring the number of events that happen during a time-period, e.g. number of patients that visit the hospital in a year.

Multinomial: One of several outcomes can happen, with different probability. E.g. tomorrow it rains, snows or it is sunny.

Uniform: Used when there are several outcomes who all have the same probability of happening, e.g. when I throw a fair die.

Binomial: Used when there are n independent items for which either A or not A, with P(A)=p.

Next lecture

– Continuous variables

– The Normal distribution

– Estimation: guess probabilities and mean values based on data

Leture 3. The Normal distribution and inference

Last lecture

- Before we perform an experiment, many outcomes are possible. What we observe depends on chance.
- A random variable X is the outcome of an experiment yet to be performed. x is an observed value.
- We can assign probabilities to outcomes.
- A probability distribution give the probabilities of every possible outcome of a random variable.
- Bernoulli trial: $P(X = a) = p$ and $P(X \neq a) = 1 - p$
- Binomial distribution: we perform n Bernoulli trials and see how many times X=a happens.

Continuous variables

Discrete variables are countable: categories, integers, Yes/no, etc.

Continuous variables can be measured infinitely precise: time, weight, height, etc

For continuous variables, we cannot assign probability to a precise point. Instead we assign probabilities to intervals on the continuous scale.

Example: We don't say P(Height is 180.10203...), but P(Height is between 179.5 and 180.5).

While there are many continuous probability distributions, we will focus on the Normal distribution.

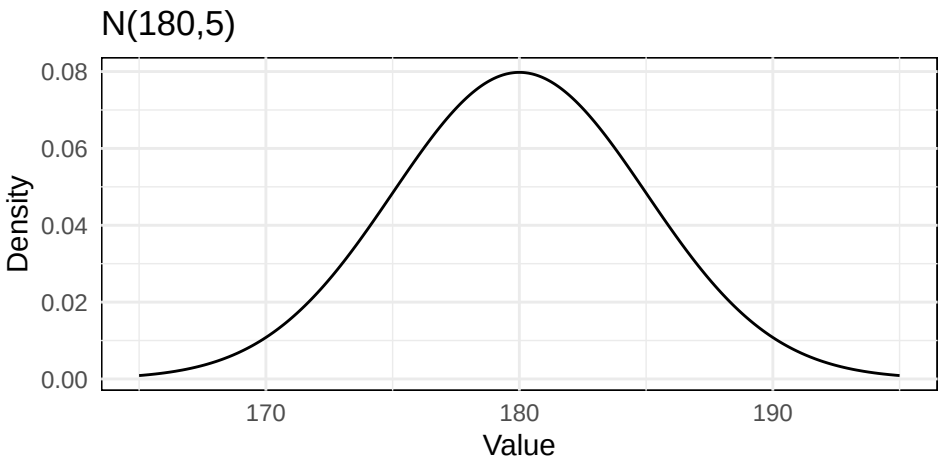
If X is a normal variable, we write $X \sim N(\mu, \sigma)$, where μ is the mean value and σ is the standard deviation.

To calculate the probability of observing values in particular intervals we use the **probability density function**

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

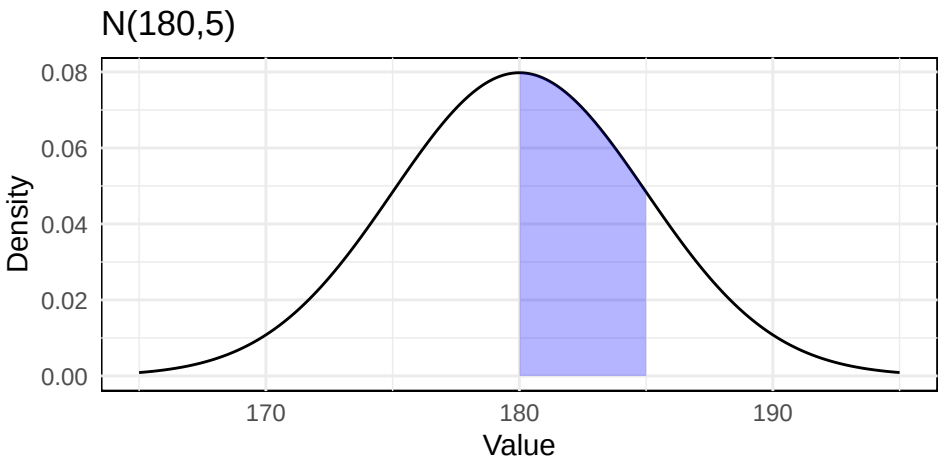
The Normal distribution

The area under the density curve is 1. The probability of observing a value within an interval is the area under the curve in that interval.



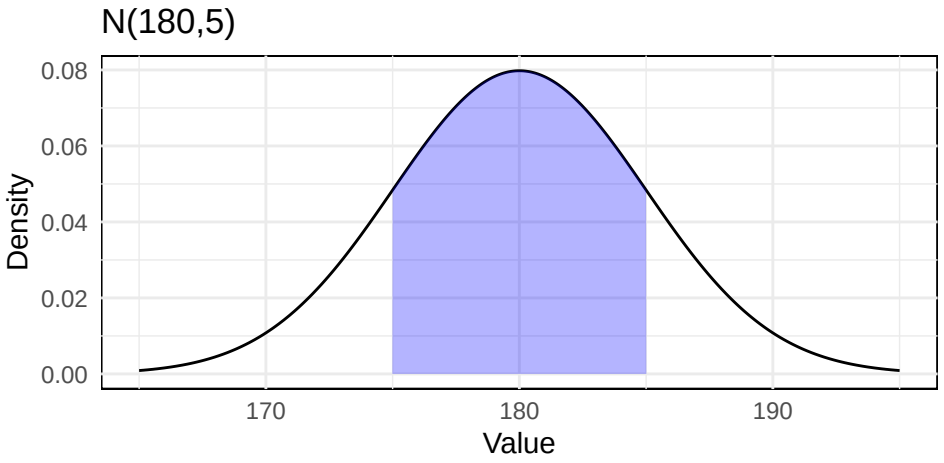
The Normal Distribution

If $X \sim N(180, 5)$, then $P(180 \leq X \leq 185) = 0.34$



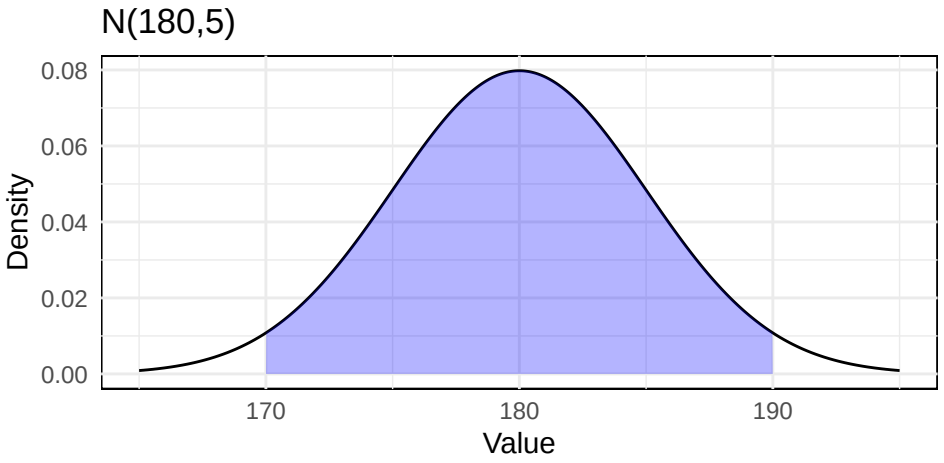
The Normal Distribution

If $X \sim N(180, 5)$, then $P(175 \leq X \leq 185) = 0.68$



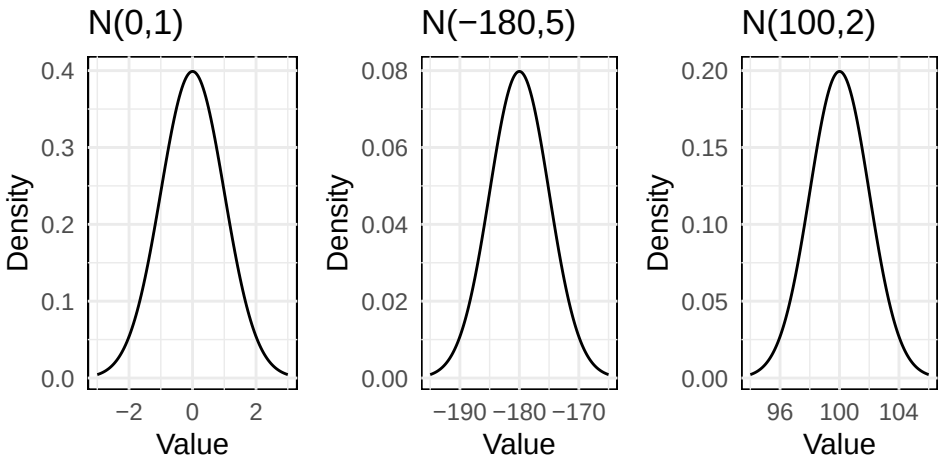
The Normal Distribution

If $X \sim N(180, 5)$, then $P(170 \leq X \leq 190) = 0.95$



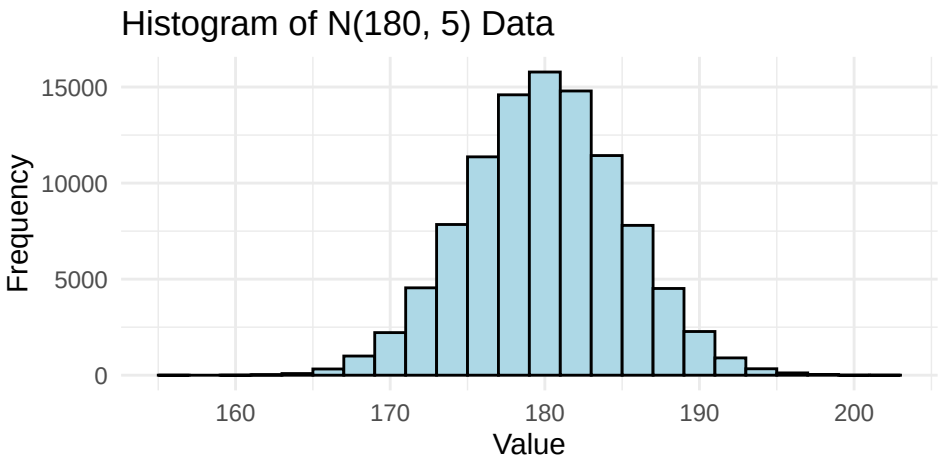
The Normal distribution

The mean and standard deviation can vary.



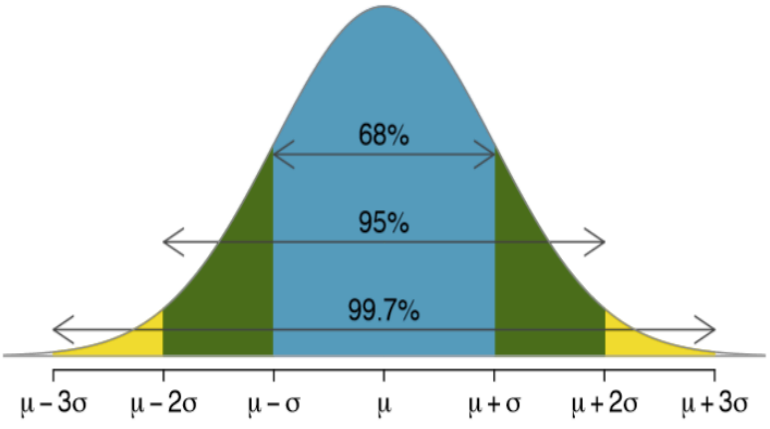
A histogram of normal data

If you have a data from a normal distribution and create a histogram, it should look like the density curve.



Symmetry

The Normal distribution is symmetric. This makes it easy to calculate probabilities in terms of the mean μ and the standard deviation σ.



Some examples

Suppose the height of Swedes follow a normal distribution with mean $\mu = 180$ and standard deviation $\sigma = 5$. We select a Swede at random.

What is the probability she is between 175 and 185?

- $175 = 180 - 5 = \mu - \sigma$ and $185 = 180 + 5 = \mu + \sigma$. By looking at the previous slide, we see that the probability is 0.68.

What is the probability she is between 170 and 190 cm?

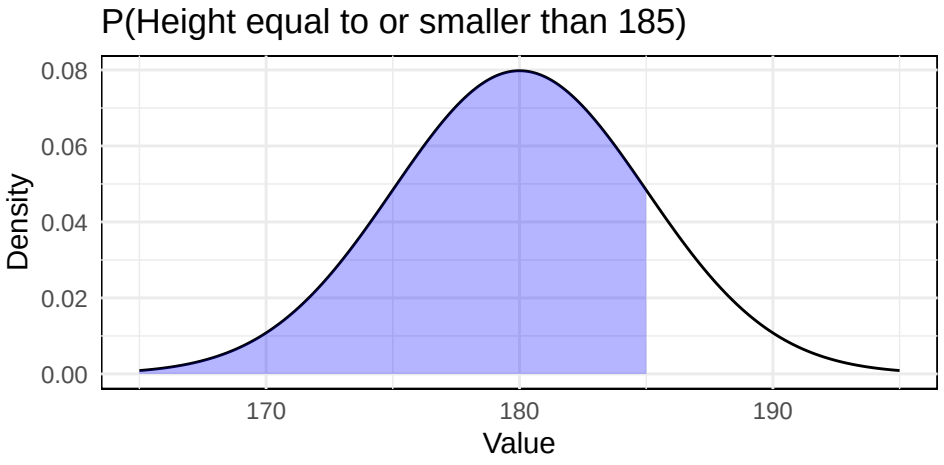
- $170 = 180 - 2 \cdot 5 = \mu - 2\sigma$ and $190 = 180 + 2 \cdot 5 = \mu + 2\sigma$. By looking at the previous slide, we see that the probability is 0.95.

What is the probability she is taller than 190?

- By looking at the previous slide and use symmetry, we realize the probability must be 0.025.

The cumulative distribution function

The cumulative distribution function is $P(X \leq x)$. In R, we use the pnorm() function.



Example

Suppose the height of Swedes follow a normal distribution with mean $\mu = 180$ and standard deviation $\sigma = 5$. We select a Swede at random. What is...

... the probability she is 185 or shorter?

```
pnorm(185,180,5)
```

```
## [1] 0.8413447
```

... the probability she is between 175 and 185?

```
pnorm(185,180,5)-pnorm(175,180,5)
```

```
## [1] 0.6826895
```

Bonus: mathematical details

For any continuous distribution, including the Normal distribution, probabilities are calculated via integrating the density function.

$$P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$$

The cumulative distribution function $F(a) = P(X \leq a)$ is defined:

$$F(a) = \int_{-\infty}^{a} f(x)dx$$

Moreover:

$$E(X) = \mu = \int_{-\infty}^{\infty} xf(x)dx.$$

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx.$$

Bonus: Continuous distributions

There are many probability distributions for continuous variables. Some examples:

Exponential: Is often used to describe the time until something happens. Only positive values are possible: $(0, \infty)$

Uniform: All values within an interval are equally likely, and nothing outside of the interval can happen.

Beta: Only values in the interval $(0, 1)$ are possible. Is often used to describe probabilities.

Estimates and inference

Parameters are numbers that describe a population: mean height, average effect of a treatment, proportion of males etc.

In research, we **estimate** parameters based on samples. This is **inference**.

Example: To estimate the mean height of Swedes, we use the mean of a sample of 10 Swedes.

Example: I take a sample of 100 cancer patients. The sample proportion of survivors is my estimate of survival probability in the population.

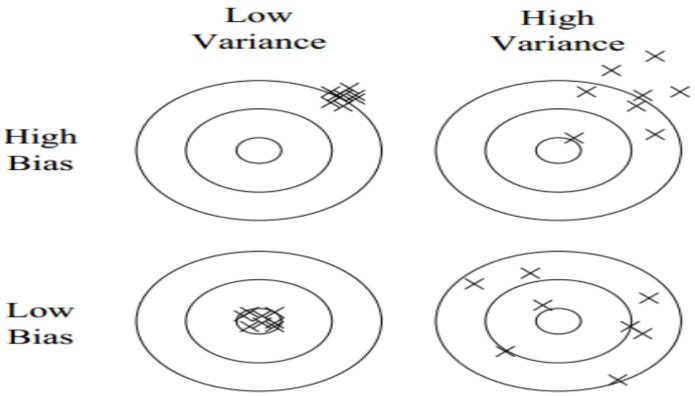
Estimation of means

We can estimate any parameter: variance, median, treatment effects, etc. We will focus on means and proportions.

The mean of a particular sample is denoted $\bar{x}$. If I take a new sample, I will get a new $\bar{x}$. Therefore, the sample mean can be considered a random variable:

$\bar{X}$= "The mean I will calculate from I sample I have yet to take".

Random samples and **large sample size** yields **low bias** and **low variance**.



Estimates and inference

We want to estimate $\mu = E(X)$ and $X \sim N(\mu, \sigma)$.

With sample size n:

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n}).$$

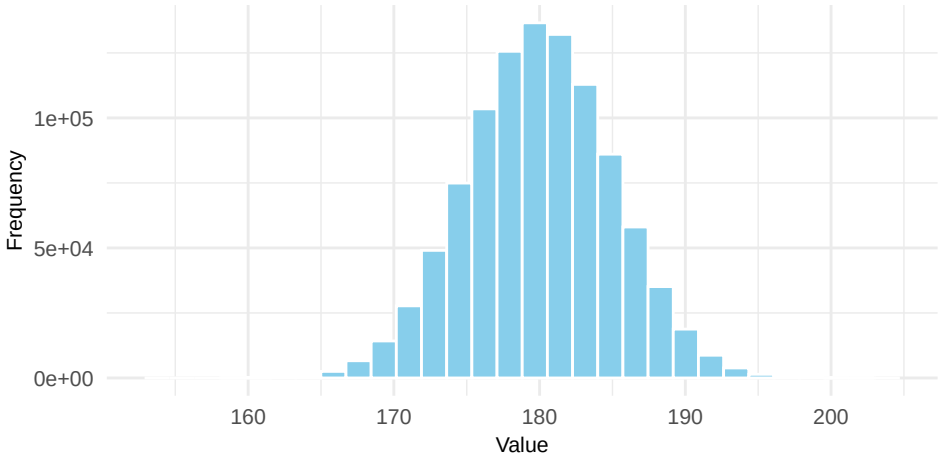
Example: We want to estimate the mean height of Swedes. The true standard deviation is 5. The sample size is $n = 100$.

$$\bar{X} \sim N(\mu, 5/\sqrt{100}) = N(\mu, 0.5).$$

If I repeated my study 100 times, I would expect 95 of the estimates to be in the interval $\mu \pm 1$. Why?

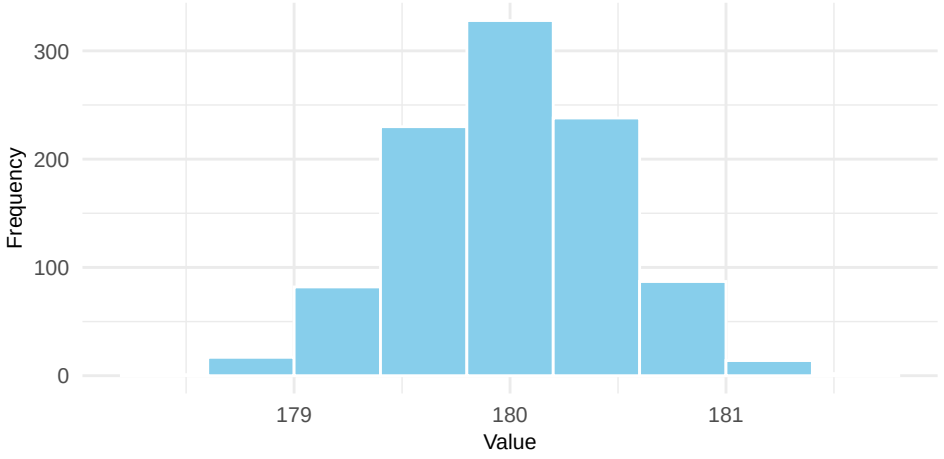
Population and sample

Here is the histogram of the height of 1,000,000 people, following a $N(180, 5)$ distribution.



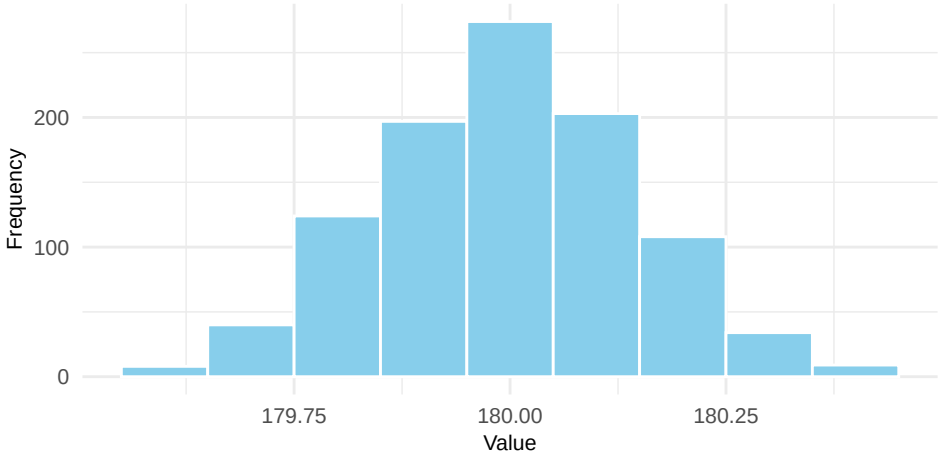
Sample size 100

We take 1000 samples of size 100 and calculate their means.



Sample size 1000

We take 1000 samples of size 1000 and calculate their means.



Important caveat

$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ is only true if we actually take a random sample from the whole population.

Example: If we only sample participants from basketball teams, the expected value of our estimate of the height of the whole population is most likely not μ anymore.

Example: We want to estimate the effect of a new cancer treatment. Will we get a random sample from all cancer patients? Or will the sickest/healthiest patients be over-represented?

Confidence intervals

$\bar{x}$ is an estimate of μ.

Most likely, $\bar{x}$ is not equal to μ.

Often, we give an interval of values that we feel confident that μ is in, based on our $\bar{x}$.

If $X \sim N(\mu, \sigma)$, and we know σ, a 95% confidence interval of μ is

$$[\bar{x} - 2 \cdot \frac{\sigma}{\sqrt{n}}; \bar{x} + 2 \cdot \frac{\sigma}{\sqrt{n}}]$$

In the above interval, -2 and 2 are the approximate 0.025 and 0.975 quantiles from the $N(0, 1)$ distribution.

Interpretation: We feel 95% certain that the true μ is inside this interval.

Confidence intervals

Why do we say that we are "95% confident"? Why don't we say there is a 95% probability?

Before we perform a study, we know that the 95% confidence interval will contain the true parameter with probability 0.95. This simply means: if we perform many studies and calculate confidence intervals, 95% of them will contain the true parameter.

But for any particular study, the confidence interval either contains or does not contain the true parameter. We just don't know which is the case.

Example of confidence interval

We know that the height of Swedes follow a normal distribution with an unknown mean μ and standard deviation $\sigma = 5$. We randomly select 100 Swedes. The sample mean height is $\bar{x} = 180$. A 95% confidence interval is then:

$$[180 - 2\frac{5}{\sqrt{100}}; 180 + 2\frac{5}{\sqrt{100}}] = [180 - 1; 180 + 1] = [179; 181]$$

What if there had been 10 000 people in our sample?

$$[180 - 2\frac{5}{\sqrt{10000}}; 180 + 2\frac{5}{\sqrt{10000}}] = [180 - 0.1; 180 + 0.1] = [179.9; 180.1]$$

Central Limit Theorem

What if we want to estimmate E(X) for an X which does not follow a Normal distribution?

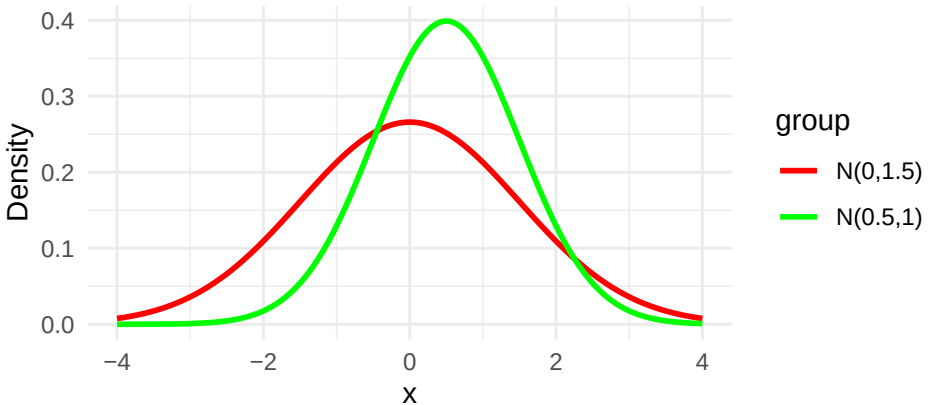
The Central Limit Theorem says that if our sample size n is large, then approximately

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$$

Our confidence intervals will look as when X is normal.

Beyond the mean

The Greens are better on average. But among the top-performers, the Reds dominate. This is due to the Reds having more variance.



Next time

What if our variable is not normally distributed? What if we do not know σ?

Lecture 4: The central limit theorem and the t-distribution

Last lecture

- If $X \sim N(\mu, \sigma)$, then μ determines where the mean is and σ determines how spread out the data is.
- $\bar{x}$ is the sample mean from a particular dataset.
- $\bar{X}$ is a random variable.
- If $X \sim N(\mu, \sigma)$, then $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

$\bar{x}$ is seldom exactly μ. But we can calculate a 95% confidence interval:

$$[\bar{x} - 2\frac{\sigma}{\sqrt{n}}; \bar{x} + 2\frac{\sigma}{\sqrt{n}}]$$

Before we perform a study, there is a 0.95 probability that the 95% confidence interval will contain the true mean. Once we have calculated the confidence interval, we say that were are 95% confident that the true mean is inside the interval.

We are 95% confident that μ is in this interval.

But what if the data is not normal or if σ is unknown?

The central limit theorem

For a large sample of observations from a random variable X, we have that

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}).$$

This is true regardless of what distribution X follows.

A very important consequence is that, assuming we know σ and a large sample size n, we can calculate a 95% confidence interval as

$$[\bar{x} - 2\frac{\sigma}{\sqrt{n}}; \bar{x} + 2\frac{\sigma}{\sqrt{n}}]$$

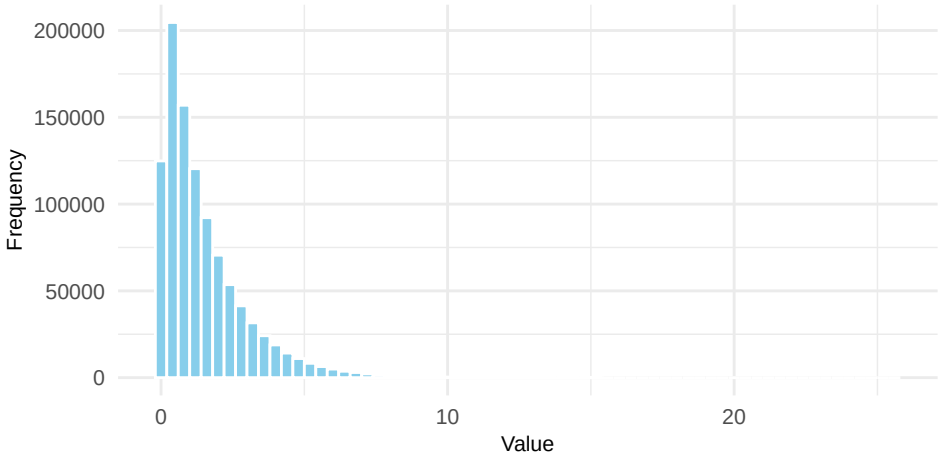
The central limit theorem

- CLT does NOT say that if I take a large sample, then the sample follows a normal distribution.
- CLT does NOT say that if I take a large sample, then the variable I take the sample from follows a normal distribution.

It is the sample mean that follows a normal distribution. This means that if I perform many studies, calculate the sample means and plot them in a histogram, it should look like a normal distribution.

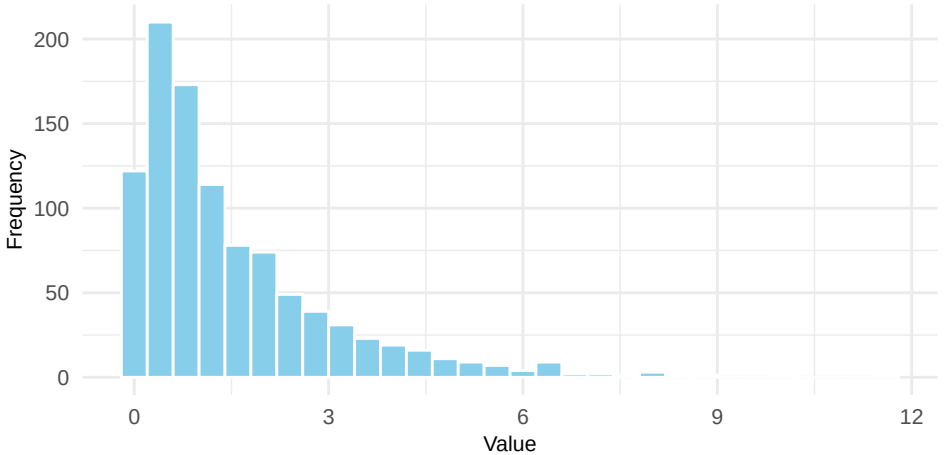
CLT in action

Below we see a histogram of the length of legs in a population of 1,000,000 insects. Clearly, their heights do not follow a normal distribution.



CLT in action

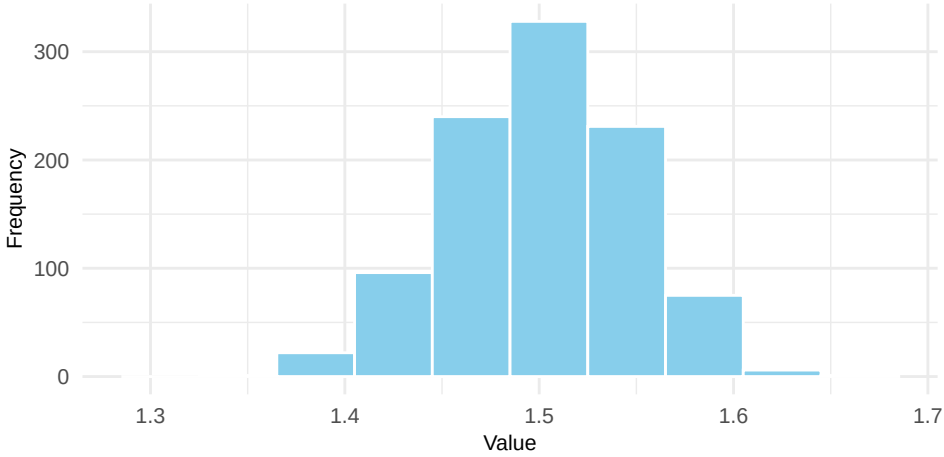
We take a sample of 1000 insects from the population of 1,000,000 insects. The sample is still not normal. The mean of the sample based on 1000 insects is 1.49.



CLT in action

We take new samples of size 1000. For every sample, we calculate the sample mean. When we have done this 1000 times, we have 1000 sample means – each based on 1000 insects. The histogram on the next slide shows the distribution of the 1000 sample means. Looks pretty normal!

CLT in action



Example

We want to know the mean height of Swedes. We take a random sample of $n = 100$ Swedes. The sample mean height is $\bar{x} = 180$.

For some reason, we know that the standard deviation $\sigma = 5$.

We do not know if the height of Swedes is normally distributed or not. The approximate 95% confidence interval is still

$$[180 - 2 \cdot \frac{5}{\sqrt{100}}; 180 + 2 \cdot \frac{5}{\sqrt{100}}] = [179; 181]$$

Example

The survival time of infants born with a particular disease has standard deviation 50 days, but we are not sure what the mean value is and what distribution the variable follows. Based on a sample of 100 infants, the sample mean surivival time is 300 days. An approximate 95% confidence interval is given by

$$[\bar{x} - 2 \cdot \frac{\sigma}{\sqrt{n}}; \bar{x} + 2 \cdot \frac{\sigma}{\sqrt{n}}] =$$

$$[300 - 2 \cdot \frac{50}{\sqrt{100}}; 300 + 2 \cdot \frac{50}{\sqrt{100}}] =$$

$$[300 - 2 \cdot \frac{50}{10}; 300 + 2 \cdot \frac{50}{10}] = [290; 310]$$

The t-distribution

So far, we assumed that we know σ. Usually, this is not the case. Instead, we estimate σ based on the sample standard deviation:

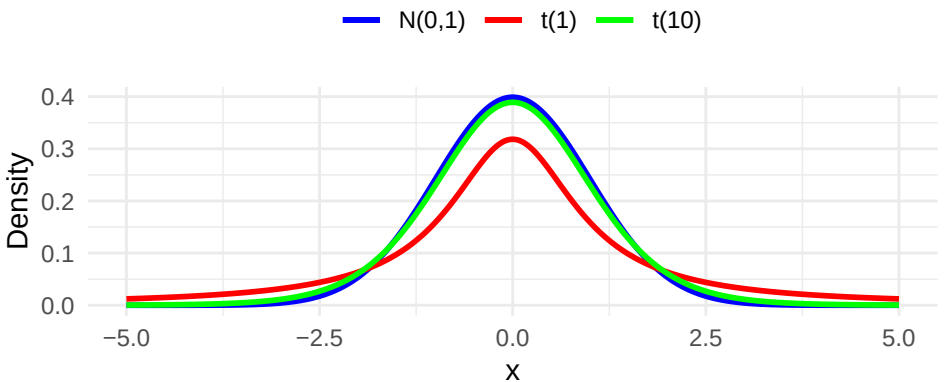
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

We can use s instead of σ to calculate confidence intervals. However, we should be less confident in our interval when we are using s instead of σ. Therefore, we do not construct a confidence interval with quantiles from the $N(0, 1)$ distribution. Instead we use quantiles from the t-distribution.

This assumes that we are estimating the mean of a variable following a normal distribution.

The t-distribution

We write t(n-1) where n is the sample size. The large n is, the more similar t(n-1) is to N(0,1). **n-1 is called the degrees of freedom.**



Confidence intervals with the t-distribution

When we know σ and the variable of interest is normal, we use the 0.025 and 0.975 quantiles from the N(0,1) distribution. If we don't know σ, we use 0.025 and 0.975 quantiles from the t(n-1) distribution.

Distribution	Q_0.025	Q_0.975
N(0,1)	-2.0	2.0
t(1)	-12.7	12.7
t(10)	-2.2	2.2
t(30)	-2.0	2.0

Example

The survival time of infants born with a particular disease follws a normal distribution, but we don't know the mean or standard deviation. Based on a sample of 11 infants, the sample mean surivival time is 300 days and the standard deviation is 50. An approximate 95% confidence interval is given by

$$[\bar{x} - t_{0.975}(11-1) \cdot \frac{s}{\sqrt{n}}; \bar{x} + t_{0.975}(11-1) \cdot \frac{s}{\sqrt{n}}] =$$

$$[300 - 2.2 \cdot \frac{50}{\sqrt{11}}; 300 + 2.2 \cdot \frac{50}{\sqrt{11}}] = [267; 333]$$

Example

We want to know the proportion of Swedes who have had covid, p. We take a sample of 100 Swedes. 40 of them have had covid, so $\bar{p} = 0.4$. The standard deviation is $\sqrt{0.4 \cdot 0.6} \approx 0.5$. We get an approximate 95% confidence interval by:

$$[\bar{p} - 2 \cdot \frac{s}{\sqrt{n}}; \bar{p} + 2 \cdot \frac{s}{\sqrt{n}}] =$$

$$[0.4 - 2 \cdot \frac{0.5}{\sqrt{100}}; 0.4 + 2 \cdot \frac{0.5}{\sqrt{100}}] =$$

$$[0.4 - 2 \cdot \frac{0.5}{10}; 0.4 + 2 \cdot \frac{0.5}{10}] = [0.3; 0.5]$$

Confidence intervals

With a large sample size, we often combine the CLT and the t-distribution. This way, we can construct confidence intervals of mean values no matter what distribution the variable is from and we don't know the standard deviation.

This is an incredible result!

What does "many" mean? Common rule of thumb is that many od 30 or more. But the situation is more complicated...

When to use what confidence interval

- **For all sample sizes**
- X is normal and you know σ: $[\bar{x} - 2\frac{\sigma}{\sqrt{n}}; \bar{x} + 2\frac{\sigma}{\sqrt{n}}]$.
- X is normal and you don't know σ: $[\bar{x} - t_{0.975}(n-1)\frac{s}{\sqrt{n}}; \bar{x} + t_{0.975}(n-1)\frac{s}{\sqrt{n}}]$.
- **For large sample sizes**
- X is not normal and you don't know σ: $[\bar{x} - t_{0.975}(n-1)\frac{s}{\sqrt{n}}; \bar{x} + t_{0.975}(n-1)\frac{s}{\sqrt{n}}]$.
- X is not normal and you know σ: $[\bar{x} - 2\frac{\sigma}{\sqrt{n}}; \bar{x} + 2\frac{\sigma}{\sqrt{n}}]$.

If sample size is small and X is not normal, you can still calculate a confidence interval, but it is not covered in this course.

One-sided confidence intervals

A one-sided lower-bound 0.95 confidence interval:

$$[\bar{x} - q(0.95) \cdot \frac{\sigma}{\sqrt{n}}; \infty]$$

A one-sided upper-bound 0.95 confidence interval:

$$[-\infty; \bar{x} + q(0.95) \cdot \frac{\sigma}{\sqrt{n}}]$$

Here, $q(0.95)$ is the 0.95 quantiles from the N(0,1) distribution. If we estimate σ with the sample standard deviation, we use the same quantile from the t(n-1) distribution.

Example

In a study with 100 participants, the mean height was 180. The standard deviation is known to be 5.

An lower-bound 0.95 confidence interval is

$$[180 - q(0.95)\frac{5}{10}; \infty] = [180 - 1.65 \cdot 0.5; \infty] = [179.2; \infty]$$

Interpretation: we are 0.95 confident that the mean height is 179.2 or greater.

An upper-bound 0.95 confidence interval is

$$[-\infty; 180 + q(0.95)\frac{5}{10}] = [-\infty; 180 + 1.65 \cdot 0.5] = [-\infty; 180.8]$$

Interpretation: we are 0.95 confident that the mean height is 180.8 or smaller.

0.9, 0.95 and 0.99 confidence intervals

Calculating 90 and 99% intervals is easy. We just use different quantiles from the N(0,1) or t(n-1) distributions. The table shows what quantiles to use depending on the type of interval and confidence level.

Interval	CI_0.99	CI_0.95	CI_0.9
One-sided	0.990	0.950	0.90
Two-sided	0.995	0.975	0.95

Lecture 5: Hypothesis testing

Last week

- Population and sample
- The Binomial distribution
- The Normal distribution
- Population parameters and sample estimates
- Confidence intervals, e.g. $[\bar{x} - 2\frac{\sigma}{\sqrt{n}}; \bar{x} - 2\frac{\sigma}{\sqrt{n}}]$
- The Central Limit Theorem: as the sample size increases, the sample mean approximates a normal distribution.
- The T-distribution lets us calculate confidence intervals with estimates of σ.

Hypothesis testing

In a hypothesis test, we have a hypothesis that we test by checking if the observed result of a experiment is consistent with the hypothesis.

If the result we observe is unlikely to have occurred under our hypothesis, we reject the hypothesis.

Null and alternative hypothesis

Null hypothesis: The hypothesis we wish to test. It is denoted H_0.

Alternative hypothesis: Says H_0 is wrong. It is denoted H_a.

One-sided and two-sided hypotheses

Example 1

H_0 : "At least half the people have anti-bodies", $P(Positive) \geq 0.5$.

H_a : "Less than half have anti-bodies", $P(Positive) < 0.5$.

Example 2

H_0 : "The mean SBP of Swedes is 130", $\mu = 130$.

H_a : "The mean SBP of Swedes is not 130", $\mu \neq 130$.

In Example 1, we have a **one-sided** null hypothesis: it can only be wrong in one way.
In Example 2, we have a **two-sided** null hypothesis: it can be wrong in two ways.

P-value and significance level

The probability of the observed outcome or something more unlikely, assuming H_0 is true, is called the **p-value**.

H_0 is rejected if the p-value is lower than **the significance level**, often denoted α. Common values on α are 0.1, 0.05 and 0.01.

The significance level α is related to the confidence level. The confidence level is $1 - \alpha$.

Hypothesis test with a p-value

- Set H_0 and H_a
- Set the significance significance level α, often 0.1, 0.05 or 0.01.
- Collect data/perform the experiment.
- Calculate the p-value, i.e. the probability of the observed outcome or something even more unlikely, assuming H_0 is true.
- If the p-value is lower than α, reject H_0. Otherwise, do not reject H_0.

Hypothesis test with a confidence interval

- Set $H_0 : \mu = \mu_0$ or $\mu \leq \mu_0$ or $\mu \geq \mu_0$, and a corresponding H_a.
- Set the significance significance level α, often 0.1, 0.05 or 0.01.
- Collect data/perform experiment.
- See if μ_0 is included in the $1 - \alpha$ confidence interval, where this interval is two- or one-sided depending on H_0.
- If μ_0 is not in the confidence interval, reject H_0. Otherwise, do not reject H_0.

Experimental design

To calulate a p-value or a confidence interval, we must design the experiment so that its outcomes follow a probability distribution that we are familiar with.

- The binomial distribution
- The normal distribution

Hypothesis test with the Binomial distribution

We want to test a new cancer treatment.

H_0: Survival probability is at least 0.8.

H_a: Survival probability is less than 0.8.

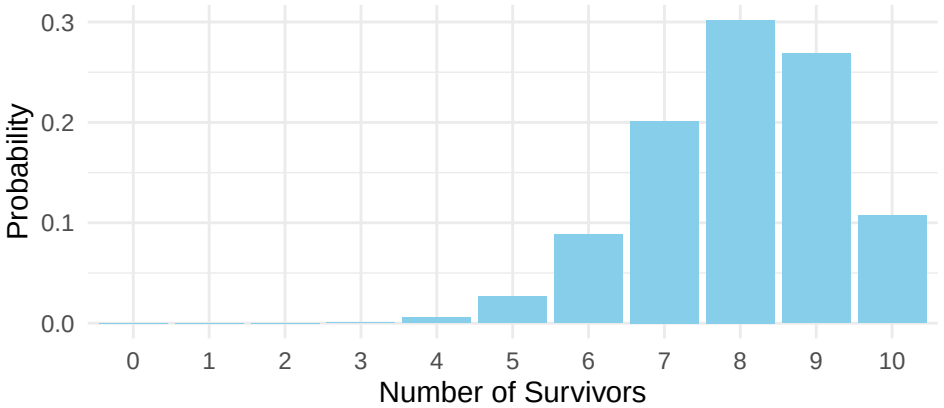
We set $\alpha = 0.1$ and give the treatment to 10 patients and see how many survive.

X="Number of survivors". By H_0, $X \sim Bin(10, 0.8)$.

Outcome: 5 persons survive. How unlikely is at this or something even more unlikely would happen, if H_0 is true?

Hypothesis test with the Binomial distribution

If H_0 is true, the probability of 5 or fewer surviving is 0.033. I.e. the p-value is lower than $\alpha = 0.1$. We reject H_0.



Hypothesis test with the Binomial distribution

We can calculate a confidence interval for a the probability **p** used in the binomial distribution. We do not cover how this confidence interval is calculated in this course.

```
binom.test(5,10,0.8,conf.level=0.9,alternative="less")
```

```
##
##  Exact binomial test
##
## data:  5 and 10
## number of successes = 5, number of trials = 10, p-value = 0.03279
## alternative hypothesis: true probability of success is less than 0.8
## 90 percent confidence interval:
##  0.0000000 0.7326819
## sample estimates:
## probability of success
##                    0.5
```

Hypothesis test with a Normal distribution

We want to know if the average survival time of cancer patients is 5.7 years.

H_0 : "Mean survival time is 5.7 years", i.e. $\mu = 5.7$

H_a : "Mean survival time is not 5.7 years", i.e. $\mu \neq 5.7$

We set $\alpha = 0.05$. Based on a sample of 100 patients, the sample mean is 6.2 and the sample standard deviation is 2. The dataset is large, so thanks to CLT and the t-distribution, no matter what distribution the survival times follow, we can construct a 0.95 CI as approximately

$$[\bar{x} - 2\frac{s}{\sqrt{n}}; \bar{x} + 2\frac{s}{\sqrt{n}}] = [6.2 - 2\frac{2}{10}; 6.2 + 2\frac{2}{10}] = [5.8; 6.6]$$

We reject H_0 since 5.7 is not inside this interval.

The T-statistic

When the variable follows a normal distribution and we estimate the standard deviation, we can calculate a p-value with the **T-statistic**.

Assuming a $H_0 : \mu = \mu_0$ or $\mu \geq \mu_0$ or $\mu \leq \mu_0$, the T-statistic is defined

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

It can be proved that if H_0 is true, then for sample size n

$$T \sim t(n-1)$$

We get a p-value by comparing the observed t-statistic with the t(n-1) distribution.

Intuition behind the T-statistic

$\bar{X} - \mu_0$ is the difference between the sample mean and the mean according to H_0. If this difference is large, the T score gets large. We divide by $s/\sqrt{n}$ to take the standard deviation of $\bar{X}$ into account.

Bonus: The Z-statistic

Same H_0 and H_a as in the previous slide, but the true σ is known. If H_0 is true and if $\bar{X}$ follows a normal distribution it can be proved that:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

We will not use this result in the course, but in the Biostatistics course you take next autumn, you will use this result in **logistic regression**.

Hypothesis test with the T-statistic

H_0 : Mean survival time for cancer patients is 5.7 years.

H_a : Mean survival time for cancer patients is **not** 5.7 years.

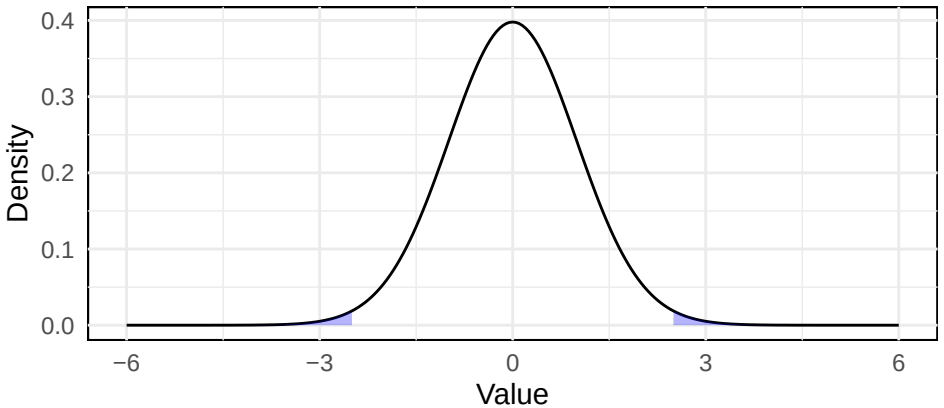
Among 100 patients, the sample mean survival time is 6.2 and sample standard deviation is 2.

$$t = \frac{6.2 - 5.7}{2/\sqrt{100}} = 0.5/0.2 = 2.5$$

To see if this is unlikely, we need to compare with the t(99) distribution, which is almost identical to N(0,1).

Hypothesis test with the T-statistic

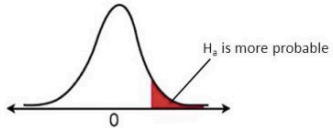
The probability of observing $t = 2.5$ or something even more unlikely under H_0 is **0.014**, which is lower than 0.05. We reject.



The function t.test() in R gives you both a confidence interval and a p-value.

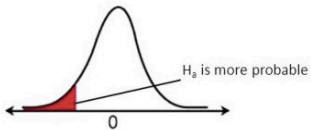
Note that what value of the T-statistic that counts as evidence against H_0 depends on if H_0 is one-sided or two-sided.

The T-statistic



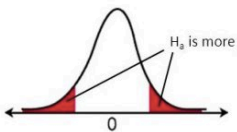
Right-tail test

H_a: $\mu >$ value



Left-tail test

H_a: $\mu <$ value



Two-tail test

H_a: $\mu \neq$ value

Two-sample t-test

We have measurements on a variable from individuals in two populations. Are the means of the populations different?

If the data in both populations follow a normal distribution or if there are many observations (CLT), and we estimate the standard deviation, we can test this with a two-sample t-test.

Intuition: Compare the sample means and divide by the standard deviation.

Two-sample t-test

We have 10 patients with feet fungus. We randomly divide them into two groups. One group gets Treatment A, the other Treatment B.

H_0 : The mean reovery time is the same with the two treatments.

H_a : The mean recovery time is different with the treatments.

```
tA <- c(5.1,4.9,1,2.1,7.8); tB <- c(4.1,2,4.2,1.1,5.2)
t.test(tA,tB,type="two.sample")$p.value
```

```
## [1] 0.5653084
```

The p-value is high so we do not reject.

Bonus: the math behind the two-sample t-test

Say that we have observations $X_1, X_2, ..., X_n$ which $N(\mu_x, \sigma_x^2)$ and $Y_1, Y_2, ..., Y_n$ which $N(\mu_y, \sigma_y^2)$, and $\sigma_x^2 = \sigma_y^2$. Assume that the X and Y variables are independent of each other.

$H_0 : \mu_x = \mu_y$

The test statistic is

$$\frac{\bar{x} - \bar{y}}{s_p\sqrt{2/n}} \sim t_{2n-2}$$

where $s_p^2 = (n-1)(s_x^2 + s_y^2)/(2n-2)$.

If there are different numbers of observations in the two groups, or if $\sigma_x^2 \neq \sigma_y^2$, the calculations get slightly different.

Paired t-test

We have two measurements from each individual, e.g. measure before and after a treatment. We want to test if the two measurements are on average the same.

If the variables follow a normal distribution or if there are many observations (CLT), we can test this with a **paired t-test**.

Intuition: Compare the mean difference between the two measurements and divide by the sample standard deviation.

Paired t-test

We have 5 patients with feet fungus. For each patient, we randomly assign salve A to one foot and salve B to the other foot.

H_0 : The mean reovery time is the same for both treatments.

H_a : The mean reovery time is not the same.

```
sA <- c(5.1,4.9,1,2.1,7.8); sB <- c(4.1,2,4.2,1.1,5.2)
t.test(sA,sB,paired=TRUE)$p.value
```

```
## [1] 0.4738173
```

The p-value is high so we do not reject.

The math behind a Paired t-test

Sometimes we have two measurements from each individual in our sample: $\{X_1, Y_1\}, \{X_2, Y_2\}, ..., \{X_n, Y_n\}$. If we want to test if $E(X) = E(Y)$, we can create a new variable $D_i = X_i - Y_i$ and test whether $E(D) = 0$. A suitable test-statistic is:

$$\frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{n-1}$$

where $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$

Where probabilities can lead us wrong

Hypothesis tests **only** make sense if you first set H_0 and **then** collect data and calculate probabilities.

Collecting data and performing many null hypotheses is bad practice. More on this tomorrow.

Lecture 6: Hypothesis tests and test errors

Hypothesis test with a p-value

- Set H_0 and H_a
- Set the significance significance level α, often 0.1, 0.05 or 0.01.
- Collect data/perform the experiment.
- Calculate the p-value.
- If the p-value is lower than α, reject H_0. Otherwise, do not reject H_0.

Hypothesis test with a confidence interval

- Set $H_0 : \mu = \mu_0$ or $\mu \leq \mu_0$ or $\mu \geq \mu_0$.
- Set the significance significance level α, often 0.1, 0.05 or 0.01.
- Collect data/perform the experiment.
- See if μ_0 is included in the $1 - \alpha$ confidence interval, where this interval is two- or one-sided depending on H_0.
- If μ_0 is not in the confidence interval, reject H_0. Otherwise, do not reject H_0.

Intuition: Suppose we have $k \geq 3$ populations and we want to test if they all have the same mean. We can look at how much variation there is between the sample means of the k different groups. If there is a large variance, the means are different. We should also take the variance within the groups into account.

ANOVA

Assume the k populations all follow normal distributions, with potentially different mean values but similar standard deviation.

$H_0 : \mu_1 = \mu_2 = ... = \mu_k$.

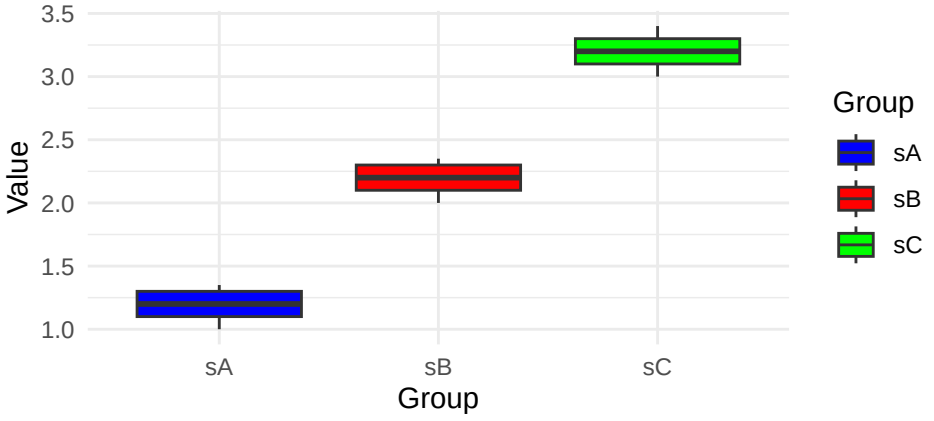
H_a : At least one μ is different.

$$F = \frac{\text{Variance between group sample means}}{\text{Variance within groups}} \sim F_{n-k,k-1}$$

This ratio is the F statistic. It follows and $F_{n-k,k-1}$ distribution, where n is the total number of participants and k is the number of populations. We can calculate a p-value based on the F-statistic. R will do this for you.

ANOVA, example 1

We have 15 patients with feet fungus. We randomly divide them into three groups, and assign salves A, B and C to the groups.



ANOVA, example 1

H_0 : Salve A, B and C have the same mean effect.

H_a : The salves are not having the same mean effect.

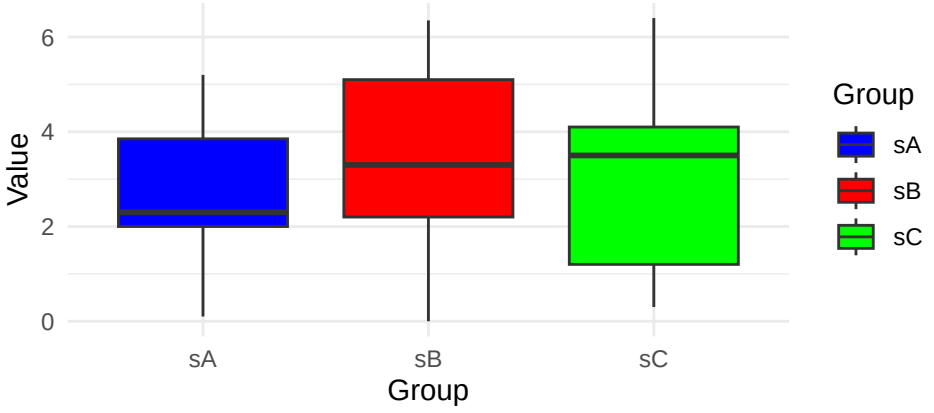
```
summary(aov(Value ~ Group, data = fungus))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Group         2 10.100   5.050   229.6 2.73e-10 ***
## Residuals    12  0.264   0.022
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is very low, so we reject H_0 on any reasonable significance level.

ANOVA, example 2

We have 15 patients with feet fungus. We randomly divide them into three groups, and assign salves A, B and C to the groups.



ANOVA,example 2

H_0 : Salve A, B and C have the same mean effect.

H_a : The salves are not having the same mean effect.

```
summary(aov(Value ~ Group, data = fungus))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Group        2   1.24   0.618   0.118   0.89
## Residuals   12  63.08   5.257
```

The p-value is high, so we do not reject H_0 on any reasonable significance level.

Types of error

Type I error: Reject H_0 when H_0 is true.

- The true mean height is $\mu = 180$. My $H_0 : \mu = 180$. My 95% confidence interval is $(181; 185)$. I reject H_0 even though H_0 was true.

Type II error: Do not reject H_0 when H_0 is false.

- The true mean height is $\mu = 180$. My $H_0 : \mu_0 = 177$. My confidence interval is $(175; 179)$. So I do not reject H_0 even though it is false.

Type I error

The significance level α is the probability that we reject H_0 when H_0 is true.

$\alpha = \text{P(Type I error)}$

I.e. if $\alpha = 0.1$ then we expect to reject H_0 10% of the studies where H_0 is correct.

Therefore, the smaller α, to smaller is the risk of Type I error.

The impact of α

Example: We calculate a $1 - \alpha$ confidence interval. We set α low.

Effect on Type II error: We get a broad confidence interval. We will accept $H_0 : \mu = \mu_0$ even when the true μ is far from μ_0. Risk of Type II error increases.

Effect on Type I error: Since the confidence interval is broad, there is a smaller risk that we reject H_0 when H_0 is true.

What if we set $\alpha = 0$? We always accept H_0, no matter what data we observe. The hypothesis test is pointless!

The impact of α

Example: We calculate a $1 - \alpha$ confidence interval. We set α high.

Effect on Type II error: When α is high, we get a narrower confidence interval. Risk of Type II error decreases.

Effect on Type I error: Since the confidence interval is narrower, there is a larger risk that we reject H_0 when H_0 is true.

What happens when we set $\alpha = 1$? We always reject H_0, no matter what data we observe. In that case, it is pointless to perform a hypothesis test!

The Power of a test

Once we have set α, we know P(Type I error).

The **power** of a test is 1-P(Type II error)=P(Not Type II error).

The power is the probability of rejecting the null hypothesis when it is false.

For some reason, researchers almost always aim for a power of 0.8.

Four aspects that impact power

Here are four important aspect of a test that affects the power:

Significance level α: The higher α is, the easier it is to reject H_0, including when H_0 is false.

Sample size: More participants increases the power of the test.

The type of test statistic: Different test statistics lead to different power.

The true parameter value: If the true parameter value is far from what H_0 stipulates, it is easier to reject a faulty H_0.

Calculating power in R

With the power.t.test() command we can calculate the power of a t-test, given **plausible assumptions** on e.g. standard deviation and difference in treatment effect.

We will look at this in the R Session.

The malaria parasite

Researchers think that the base-pair length of a gene in a malaria parasite is different in two different parasite populations. If this can be established, it may further the development of a malaria vaccine.

Idea: Collect n samples from each parasite population and compare the mean basepair length between the groups.

Analyzing n=20 parasites per population costs 500,000 SEK. Analyzing n=100 parasites per population costs 1,000,000 SEK.

H_0 : "No difference in mean base-pair length between the populations".

20 parasites: If the distribution of base-pair length is approximately normal in both populations, we can perform a two-sample t-test. The standard deviation is believed to be 15. With 20 per group and $\alpha = 0.05$, the power is 0.8 if the true difference in means is 13.6.

100 parasites: Thanks to CLT, we can definitely perform a t-test. Assuming a standard deviation of 15 and $\alpha = 0.05$, the power is 0.8 if the true difference in means is 6.

20 parasites, without normality

If the distribution of base-pair lengths is far from normal, it may be unwise to use a t-test with only 20 parasites per group.

Then, we can only look at whether it is more common among parasites in one population to have a larger proportion of members with a base-pair length of 243 or longer. With only 20 parasites per group, this would require that there is a big difference in proportion between the populations (about 0.4) for the power to be 0.8.

What differences are important

Whether or not it is worth spending extra money to order more samples, depends on how important it is for us to detect a particular difference between the two groups.

If the true difference in mean base-pair length is 0.00001, would it matter to vaccine development? If yes and if we think the difference is likely to be this small, it may be worth collecting a huge sample.

If a mean difference smaller than 13.6 is useless to know about, because it will not further vaccine development, 20 samples will probably do.

Important regarding α

α is the probability of rejecting H_0 when it is true.

If $\alpha = 0.05$ and we perform 20 studies, we expect 1 of them to reject H_0, even if H_0 is true.

P-hacking

If we perform several hypothesis tests based on a single dataset, we expect to find a H_0 that can be rejected even if all of them are false.

This is called p-hacking.

Leture 7: Non-parametric tests

Introduction

The hypothesis tests we have looked at so far assume that the data follows a normal distribution, or that we have a large sample size so that the sample mean follow a normal distribution in virtue of CLT.

Non-parametric tests make fewer assumptions about the distribution of the variable of interest. They only require that the data is **ordinal**.

Example 1 Patients report their level of pain as low, medium or high.

Example 2: The GAD-7 score is given on a scale from 0 to 21. It is based on patients saying how often they suffer from certain anxieties.

When to use non-parametric tests

- The data is ordinal, but not on a ratio scale.
- The data cannot be assumed to follow a normal distribution, and sample size is too low to use CLT.
- There are outliers that make the mean value unreliable.

You **can** use non-parametric tests with normally distributed data, but usually this leads to lower power than if you used a parametric test.

The median

The median is a value such that at least half the observations are lower than or equal to this value, and at least half are higher than or equal to this value.

In the special case when there are an even number of observations and the two middle values on a sorted list of the values are not equal, this simply means that half the observations are lower than the median.

We can use the median for any ordinal variable, whereas the mean makes no sense in many situations.

The sign test

We want to test the following hypothesis regarding a population:

H_0 : The median is μ_0

H_a : The median is not μ_0.

If I take a sample from the population and remove any values equal to μ_0, so that n observations remain. I calculate X="Number of observations lower than μ_0", then under H_0

$$X \sim Bin(n, 0.5).$$

Example: Feet fungus

H_0: The median recovery time from feet fungus is one week.

I study 10 patients and define X="Number of patients that recover within a week".

$$X \sim Bin(10, 0.5)$$

If all 10 patients recover within a week:

```
binom.test(10,10,0.5)$p.value
```

```
## [1] 0.001953125
```

We reject H_0 on e.g. $\alpha = 0.01$.

The sign test, paired values

We have paired values, e.g. measurements from two different treatments on the same persons: X_b and X_a

$H_0 : P(X_a < X_b) = P(X_a > X_b)$

$H_a : P(X_a < X_b) \neq P(X_a > X_b)$

If H_0 is true, then for each patient there is a 0.5 probability that $X_a < X_b$. Define $Y =$"Number of participants out of n for whom $X_a < X_b$".

$$Y \sim Bin(n, 0.5)$$

If there are observations for which $x_a = x_b$, we remove them before we perform the test. n is then the number of participants without ties.

Example: Feet fungus

H_0: Treatment A and B have the same effect on recovery time from feet fungus.

I study 10 patients. Each patient has Treatment A on one foot and Treatment B on the other, which foot gets which treatment is decided by coin-flip. Define X="Number of patients who recover on the foot with Treatment A first". Assuming no ties:

$$X \sim Bin(10, 0.5)$$

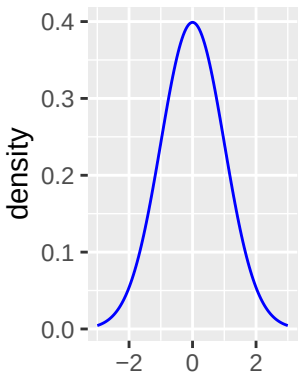
If all 10 recover on the foot with Treatment A first, the p-value is lower than any sensible α:

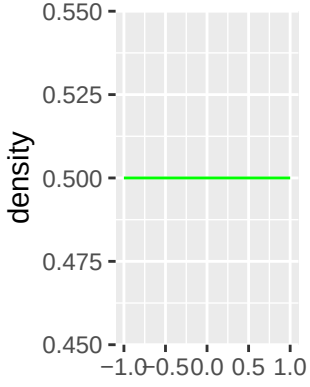
```
binom.test(10,10,0.5)$p.value
```

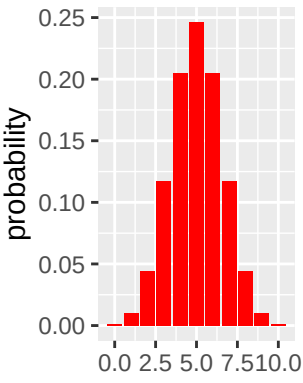
```
## [1] 0.001953125
```

Symmetry

The normal, uniform and the Bin(n,0.5) distribution are examples of symmetric distributions.







Wilcoxon signed rank test

Assuming that the variable is symmetrically distributed around its median, we test:

H_0 : The median is μ_0.

H_a: The median is not μ_0.

Very few people are aware that the variable should be symmetrically distributed.

Wilcoxon signed ranked test

We take a sample of size n from the population.

For every observation, we calculate $x_i - \mu_0$, and rank the difference from highest to lowest in terms of absolute magnitude: $|x_i - m_0|$.

We then sum the ranks for all x_i that are larger than μ_0. If the rank sum is **high** or **low**, then we reject the null hypothesis.

You can look at the sum of ranks for x_i smaller than μ_0 as well. It gives equivalent results.

Wilcoxon signed ranked test

H_0 : The median recovery time from feet fungus is 7 days.

Recovery	mu_0	difference	abs_difference	rank
1.1	7	-5.9	5.9	4
7.4	7	0.4	0.4	1
0.3	7	-6.7	6.7	5
8.3	7	1.3	1.3	3
6.4	7	-0.6	0.6	2

The rank sum of the values smaller than 7 is $4 + 5 + 2 = 11$ and the rank sum of the values larger than 2.9 is $3 + 1 = 4$.

Paired Wilcoxon signed rank test

We have paired observations $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$. Assuming the differences between X and Y are symmetric:

H_0 : The median of the differences between X and Y is 0.

H_0 : The median of the differences between X and Y is not 0.

For every observation, we calculate $x_i - y_i$, and rank the differences from highest to lowest in terms of absolute magnitude: $|x_i - y_i|$.

We reject H_0 if the rank sum corresponding to the positive differences is **high** or **low**.

Paired Wilcoxon signed rank test

Five patients have Treatment A on one foot and Treatment B on the other foot.

H_0 : The median of the differences between the recovery time from feet fungus with Treatment A and B is 0.

A	B	difference	abs_difference	rank
1.1	2.2	-1.1	1.1	2
8.2	4.2	4.0	4.0	4
3.4	3.0	0.4	0.4	1
5.1	9.2	-4.1	4.1	5
9.2	5.4	3.8	3.8	3

The rank sum corresponding to the positive differences is $4 + 1 + 3 = 8$.

Ties

The standard Wilcoxon signed rank tests assume that no value is equal to μ_0 and that there are no tied values in the paired test. There are ways of handling ties and values equal to μ_0. R can do this for us.

Wilcoxon rank sum test

With the rank sum test, we test if two independent variables have the same distribution **up to a location shift**.

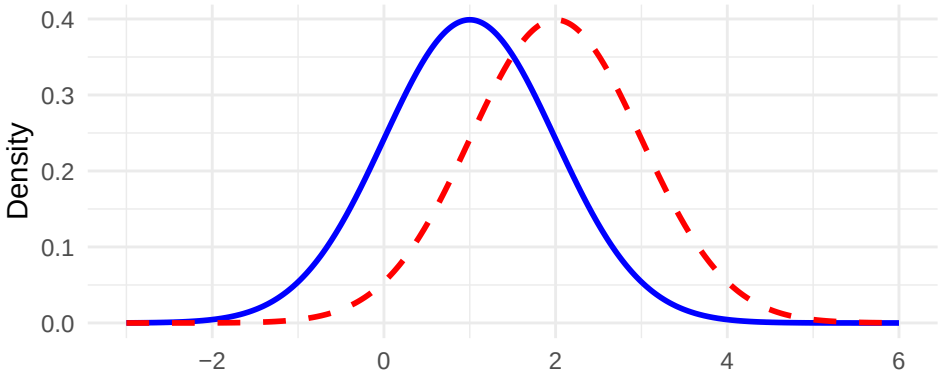
$H_0 : P(X \leq a) = P(Y \leq a + \mu_0)$ for all a.

$H_a : P(X \leq a) \neq P(Y \leq a + \mu_0)$ for all a.

If H_0 is true, then the medians differ by μ_0.

Location shift

Two Normal distributions with a location shift of 1.



Wilcoxon rank sum test

We rank all observations from highest to lowest. We calculate the rank sum of each group. If the difference in rank sum is **high**, we reject H_0.

Wilcoxon rank sum test

H_0 : The recovery time from feet fungus follow the same distribution with Treatment A as with Treatment B. I.e. the location shift $\mu_0 = 0$.

We randomly assign the Treatment A and B to 10 patients. The rank sum is 29 (A) and 26 (B).

Treatment	Recovery	rank
A	1.1	1
A	8.2	8
A	3.4	4
A	5.1	6
A	9.2	10
B	2.2	2
B	4.2	5
B	3	3
B	9.1	9
B	5.4	7

How to evaluate the ranks

We have said that we reject H_0 if ranks are **high** or **low**. What does this mean?

In all cases we have discussed, we would expect the ranks of the positive and negative differences (or the ranks of both treatment groups) to be same, if H_0 is true. If H_0 is true and we observe high or low ranks in one group (or for positive or negative differences), this must have happened by chance in the sampling/randomization procedure.

We can calculate the probability of this happening by listing all possible rankings and calculating what proportion of them give more extreme rank sums than what we observed.

In the Wilcoxon rank sum test, we also take into account that there can be different number of observations from the populations we are comparing.

Kruskall-Wallice

We have more than two variables $X, Y, Z, ...$ and we want to know if they follow the same distribution, that is if $P(X \leq a) = P(Y \leq a) = P(Z \leq a)...$ for all a.

We calculate the rank of all observed values. For each group, we calculate the mean rank and also the variance of ranks within the group.

Compare the difference in mean rank between the groups with regard taken to the variance of ranks within the groups.

Lecture 8: Contingency tables and association

Association

Often, two variables are **associated** with each other. An association can be positive or negative. If there is not association, the variables are **independent**.

Positive: Taller people tend to weigh more than shorter people.

Negative: People who exercise a lot have a lower comorbidity score.

Independence: Socioeconomic status is independent of height.

Contingency tables

If two variables are categorical, we can show the association with a contingency table.

	non-smoker	smoker
f	6	9
m	9	6

There is a positive association between sex and smoking habits in our sample. Is this true for the whole population?

Idea: Compare the observed numbers with what we would expect to see if sex and smoking habits are independent!

Reminder: Independence

X and Y are independent if for all a and b:

$$P(X = a \ \& \ Y = b) = P(X = a) \cdot P(Y = b)$$

Independence

If sex and smoking are independent, then

$$P(Sex = Male \ \& \ Smoker = Yes) = P(Male)P(Yes)$$

and so forth for all other values of Sex and Smoker.

Assuming independence, the expected number of male smokers in the dataset would be

$$n \cdot P(Male)P(Smoker)$$

where n is the sample size. We don't know P(Male) and P(sex), but we can estimate them with the sample proportions.

The proportion of males is $\frac{15}{30} = 0.5$ and the proportion of smokers is $\frac{15}{30} = 0.5$.

Our estimate of the expected number of male smokers would then be

$$30 \cdot 0.5 \cdot 0.5 = 30 \cdot 0.25 = 7.5$$

By similar reasoning, we can calculate the expected number of all combinations of sex and smoking habits.

Expected cell counts

If sex and smoking habits are independent variables, we would expect the following table:

	Smoker	Non-Smoker
m	7.5	7.5
f	7.5	7.5

If the actual cell counts are sufficiently different from the expected cell counts, sex and smoking are probably not independent. **So, what is "sufficiently different"?**

Comparing actual and expected counts

We compare the expected and the observed tables in the following way:

- For every cell, we square the difference between expected and actual counts, and we divide by the expected counts.
- We sum the squared differences divided by the expected cell counts, and call this sum Q.
- The larger Q is, the larger the difference between expected and observed cell counts.
- It can be proved that if the variables are independent, Q follows a χ^2 distribution with $(n_r - 1) \cdot (n_c - 1)$ degrees of freedom,where n_r is the number of row categories and n_c is the number of column categories

The χ^2 test

H_0 : The variables are independent.

H_a : The variables are not independent.

We reject H_0 if Q is larger than the $1 - \alpha$ quantile of the χ^2 distribution with $(n_r - 1)(n_c - 1)$ degrees of freedom, .

```
chisq.test(smokesex)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  smokesex
## X-squared = 0.53333, df = 1, p-value = 0.4652
```

The math behind the χ^2 test

We use e_{ij} to denote the expected value for the cell on row i and column j, under the H_0 of independence. We use o_{ij} to denote the observed value on row i and column j. Then

$$Q = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}.$$

Under H_0, $Q \sim \chi_2((n_r - 1)(n_c - 1))$.

In the case of the smoke-sex table:

$$Q = \frac{(9 - 7.5)^2}{7.5} + \frac{(6 - 7.5)^2}{7.5} + \frac{(6 - 7.5)^2}{7.5} + \frac{(9 - 7.5)^2}{7.5}$$

Fisher's exact test

The χ^2 is unreliable with few observations or unbalanced data. Rule of thumb: there should be at least 5 observed values in every cell. Otherwise, use Fisher's test. It also has H_0: "The variables are independent".

```
fisher.test(smokesex)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  smokesex
## p-value = 0.4661
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.0807593 2.3758837
## sample estimates:
## odds ratio
##  0.4569634
```

Correlation

If two variables are continuous, we can measure the **linear association** with Pearson's correlation coefficient ρ. It is between -1 and 1.

- ρ close to 1: Strong positive association.
- ρ close to -1: Strong negative association.
- ρ close to 0: Independence.

ρ is a measure of linear association, but says nothing about causality.

Bonus: The math behind correlation

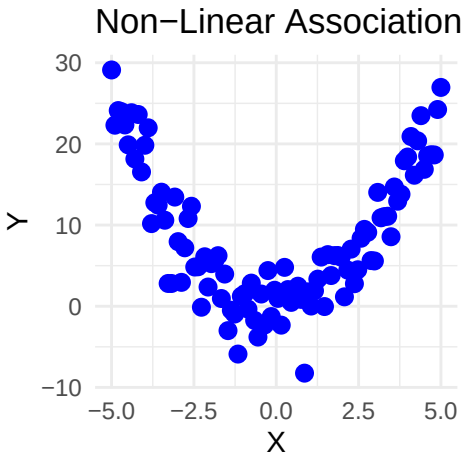
The covariance between X and Y is

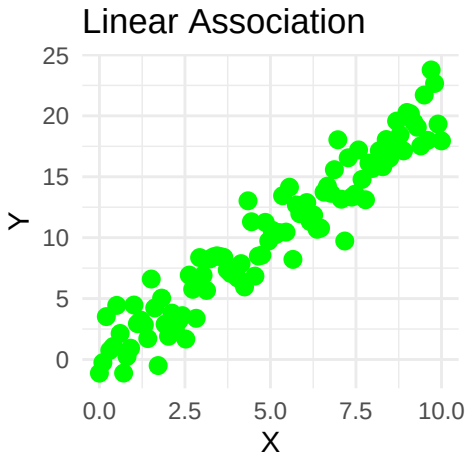
$$Cov(X, Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n - 1},$$

with n being the sample size. Pearson's correlation coefficient is:

$$Cor(X, Y) = \frac{Cov(X, Y)}{sd(X)sd(Y)}$$

Linear and non-linear association





Modelling

In science, we build models of phenomena we are interested in. A model is a caricature of the world. Hopefully it is useful.

Galileo's formula

$$s = \frac{1}{2}gt^2$$

where s is the movement of an object, t is time and g is gravital force.

In medicine, we often develop statistical models that are simplifications – but helpful. E.g. no real life phenomenon is perfectly normally distributed, but it may be useful to assume that they are.

Linear regression

We can build a statistical model around the linear association between two variables.

$$E(Weight|Height) = \beta_0 + \beta_1 Height$$

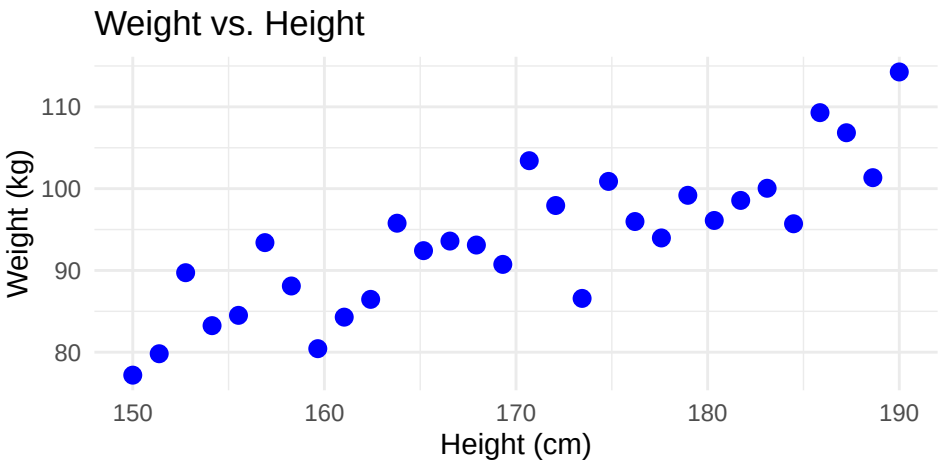
This is the equation of a straight line, and we can think of the model as a way of fitting a line to observed data.

More generally

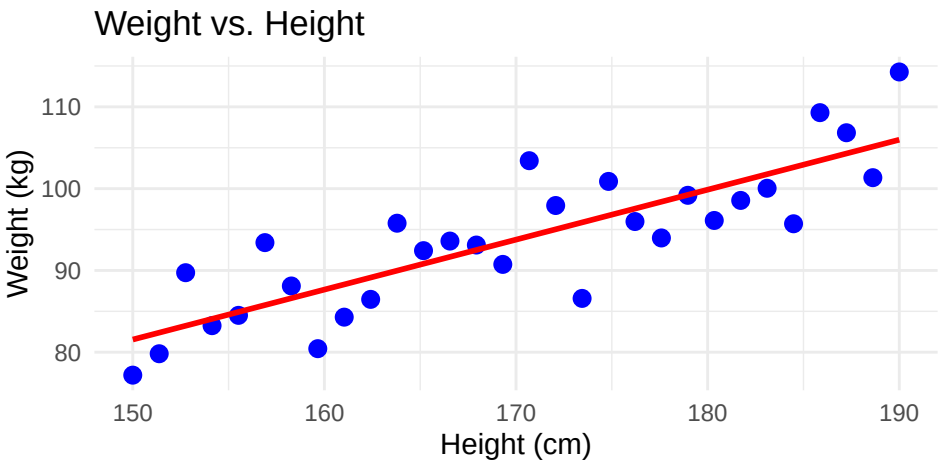
$$E(Y|x) = \beta_0 + \beta_1 x$$

where we often call Y the dependent variable and x the independent variable.

Weight and height data



Weight and height data, with regression line



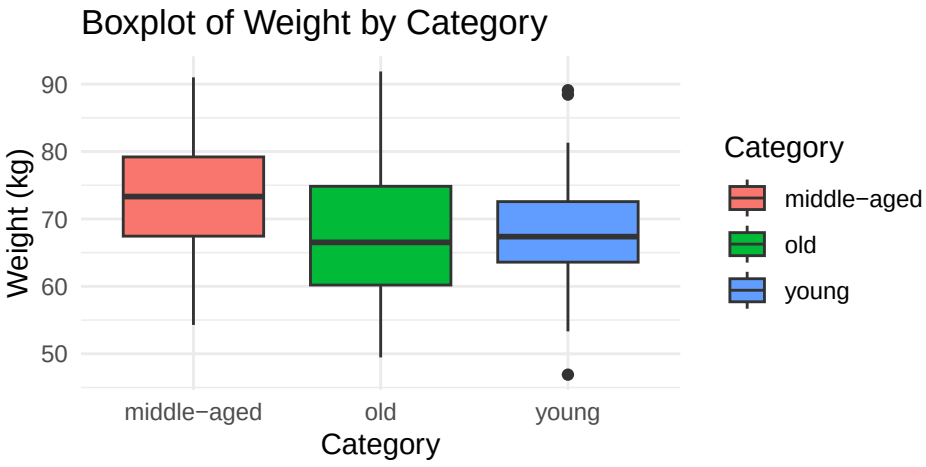
The intercept corresponds to β_0 and it is the expected weight of a person who is 0 cm tall. Obviously, such a person does not exist. In this case, the purpose of β_0 is simply to act as a constant necessary for our equation to work.

The height parameter corresponds to β_1. It describes the expected weight gain of a person who gains a cm of height.

```
lm(weight~height)
```

```
##
## Call:
## lm(formula = weight ~ height)
##
## Coefficients:
## (Intercept)       height
##    -10.0744       0.6108
```

We can have a categorical variable as the independent variable.



ANOVA

We can test if the three age categories are having the same mean weight with an ANOVA.

```
summary(aov(Weight ~Category,data=df))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Category      2    394  197.00   2.201  0.116
## Residuals    97   8681   89.49
```

Linear regression

With linear regression, we get precise estimates of the mean value of each group. The intercept is the mean of the middle-aged. The paramters corresponding to old and young, describe how their expected weight deviates from the expected weight of middle-aged.

```
lm(Weight ~Category,data=df)
```

```
##
## Call:
## lm(formula = Weight ~ Category, data = df)
##
## Coefficients:
##   (Intercept)    Categoryold  Categoryyoung
##        72.273         -4.191         -4.320
```

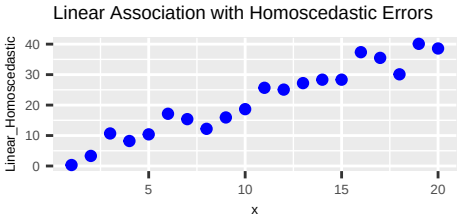
Assumptions

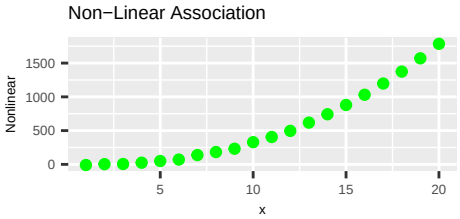
$E(Y|x) = \beta_0 + \beta_1 x$ is equivalent to the model $Y|x = \beta_0 + \beta_1 x + \epsilon$.

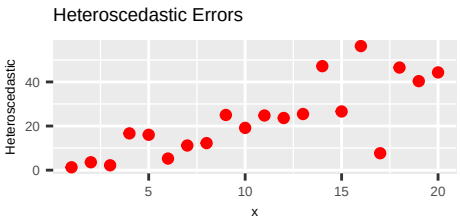
Assumptions behind linear regression:

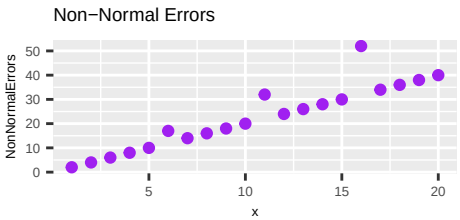
- The error terms $\epsilon \sim N(0, \eta)$. With a large sample size, normality of the error terms is approximated thanks to CLT.
- The error terms are homoscedastic: they are independent of the value of the independent variable.
- Linear relationship between the independent and the dependent variable.

Assumptions









Final words

Whenever you are doing statistics, stay humble – and slightly paranoid.

"All who drink of of this treatment recover within a short time. Except those whom it does not help, and who will all die. It is obvious therefore, that it fails only in incurable cases."

- Galen (129-216 A.D.)