STA120 Summary

Course Information

Shared document for the lecture and the exercises

Office Hours: Tuesday from 13:00 to 13:30 in Y27H06 or set up the time slot per email.

10.5 points out of 13 from the exercise sheets have to be obtained to be allowed to take the final exam.

Some problems in the exercise sheets are marked as difficult: at the exam, at least 85% of the points will be possible to get without solving these kinds of problems. Some other problems are marked as long: relevant for the exam, but homework is long.

Corrections of the exercises are not very detailed: feel free to ask your grader for a detailed correction if there are any doubts after seeing the solutions.

Final Exam:

- Part A: 90 minutes, handwritten on paper, online upload (jpg files of photos taken with phone), A4 both-sided cheatsheet allowed. Explain concepts, calculations, explain code, yes/no questions with justification (true or false)
- Part B: 90 minutes, everything is allowed (apart from communication, but ChatGPT allowed). Analyze a given data set, a simulation, write a function

Final exam is not on ACCESS for the B part.

Lecture 1: Exploratory Data Analysis (EDA)

Learning goal: Understand the concept and the need of an exploratory data analysis (EDA) within a statistical data analysis.

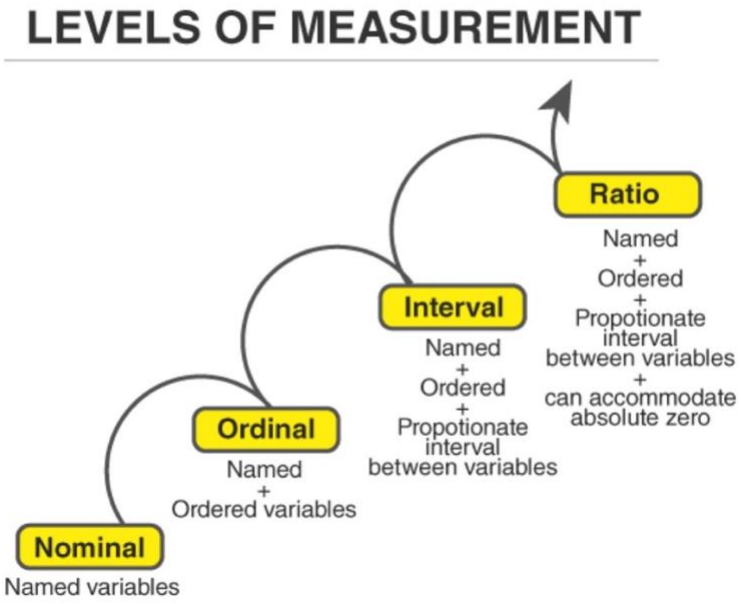
EDA helps to understand what types of data there are in a given dataset, how many missing values and data points there are, what the key summary statistics of the dataset are, and what patterns, features, and clusters exist in the data.

Learning goal: Know different data types and operations we can perform with them.

Qualitative data: consists of categories and are either on nominal scale (e.g., male/female) or on ordinal scale (e.g., weak<average<strong, nominal with an ordering). Non-numerical data is considered qualitative.

Quantitative data: numeric, mathematical operations can be performed with it. Either discrete (taking on only specific values, for example integers), or continuous (taking on any value of the real number set). The ratio scale is characterized by an absolute zero in addition to the characteristics of the interval scale.

		Measurement scale			
		nominal	ordinal	interval	real
Mathematical operators		$=, \neq$	$=, \neq$	$=, \neq$	$=, \neq$
			$<, >$	$<, >$	$<, >$
				$-, +$	$-, +$
					$\times, /$
Statistical measures	location	mode	mode	mode	mode
			median	median	median
				arithmetic mean	arithmetic mean
					geometric mean
	spread			range	range
					studentized range
				standard deviation	standard deviation
					coefficient of variation



	Nominal	Ordinal	Interval	Real
Male/Female/X	x			
Extremely dislike/ dislike/neutral/like/ /extremely like		x		
less than 50K/50K-100K/over 100K		x		
temperature (Celcius)			x	
weight				x
temperature in Kelvin (0.0 Kelvin means "no heat")				x

Natural (intuitive) and original measurement scale:

Natural = intuitive = what kind of measurement scale would you use for a given variable.

"Natural scale" - is not a concept, it is just a question what measurement scale sounds reasonable to you for a given variable.

E.g. we could have a variable gender, with levels "female", "male", "other".

That would be a reasonable/natural choice.

"Original measurement scale" means: what was used in the considered data frame.

But it is technically possible to use numeric values, e.g. 0, 1, 2 to code the genders.

Learning goal: Calculate different descriptive statistics from a dataset.

Be able to calculate number of observations, number and type of variables, number of missing values, (empirical) mean, truncated/trimmed mean, median, quartiles and quantiles, variance, standard deviation, and interquartile range (3rd quartile – 1st quartile).

Statistics for the location (position of the data):

Sample mean:

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

Sample median is the middle number in a sorted list of numbers, or the mean of the two numbers in the middle:

$$med(x_i) = \begin{cases} x_{\left(\frac{n}{2}+\frac{1}{2}\right)} & if\ n\ is\ odd \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & if\ n\ is\ even \end{cases}$$

Trimmed mean, e.g. a 5% trimmed mean, the lowest 5% and highest 5% of the data are excluded, compute the mean from the remaining data.

remove same percentage of obs on both sides and compute the mean with the remaining observations

Why median/trimmed mean? Outliers: observations that are separated far in value from the remaining measurements (extremely high or extremely low values), they drastically influence the value of the mean so it's better to use median or trimmed mean which do not include outliers in the calculation.

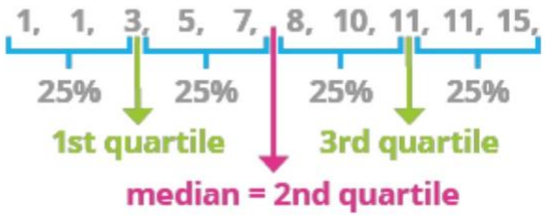
A trimming of 50% is equivalent to the sample median. The sample median is the 50th percentile, where half the data is smaller than the median and the order half is larger. The quartiles divide the data in four equally sized groups.

Quantiles link observations or values with the **position in the ordered data**. → ordered in ascending order!

0.25 quantile = 1st quartile

0.50 quantile = 2nd quartile

0.75 quantile = 3rd quartile



Statistics for the spread (dispersion of the data):

Sample variance:

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

Sample standard deviation:

$$s = \sqrt{s^2}$$

Sample interquartile range: → best for finding outliers

$$IQR = 3^{rd}\ quartile - 1^{st}\ quartile$$

Learning goal: Perform an EDA in R.

read.csv(), head(), str(), summary(), tail(), df[is.na(df$col),], pairs(df) par(mfrowc(1, n)) and then with(df, hist(df$work), hist(df$price)). Relationship between variables.

Subset data : with subset function or df_subset <- df[df$year == 2 & df$class == 1,]

with(df, plot(class ~ year, col = "grey")) to see if there is a relationship in data (for EDA). colSums(is.na(dat)) to see for each column how many observations don't have data

(header=T): information about the variable names is included in dataset.

Learning goal: Sketch schematically, plot with R and interpret a histogram, barplot, boxplot.

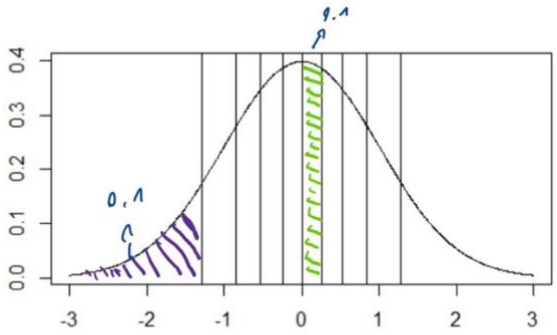
Univariate data: composed of one variable, or a single scalar component, measured at several instances or for several individuals or subjects.

Left-skewed data: more smaller values on the left side of the data.

Right-skewed data; more smaller values on the right side of the data.

Boxplot: graphical representation of five location statistics of the observations: median, lower and upper quartiles, and the minimum and maximum values.

Q-Q plot: has the goal to visually compare empirical data quantiles with the quantiles of a theoretical distribution. The ordered values are compared with the i/(n + 1) quantiles. If the points of the Q-Q plot are aligned along a straight diagonal line, then there is a good match between the sample and theoretical quantiles. A deviation of the points indicates that the sample has either too few or too many small or large points. In a Q-Q plot, the quantiles of a normal distribution are displayed, 0.1 by 0.1, represented by vertical lines.



Q: What can we say about the two marked areas? Are equal (both 10%)

Quantiles 0.1, 0,2, …, 0.9 split the are under the density curve into ten equal areas.

The Q-Q plot is used to check which of the two distributions (chi square or normal) describes the data better.

Learning goal: Visualize multivariate data (e.g., scatterplot) and recognize special features therein.

Multivariate data means that two or more variables are collected for each observation and are of interest. Additionally to the univariate EDA applied for each variable, visualization of multivariate data is often accomplished with scatterplots (plot(x,y)) for two variables or pairs() for several, so that the relationship between pairs of variables is illustrated.

Lecture 2: Random Variables

Learning goal: Describe in own words a cumulative distribution function (cdf), a probability density function (pdf), a probability mass function (pmf), and a quantile.

A random variable X is a measurable function from the sample space omega to the set of real numbers R and represents a possible numerical outcome of an experiment. Random variables are denoted with uppercase letters (e.g., X, Y), while realizations (outcomes) are denoted by the corresponding lowercase letters (x, y).

A random variable is called discrete when it can assume only a finite or countably infinite number of values.

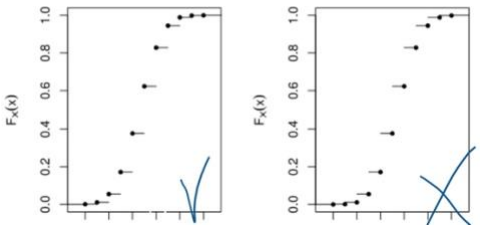
The cdf (cumulative distribution function) gives probabilities that the random variable takes that or any smaller value.

- The distribution function (**cumulative distribution function, cdf**) of a random variable X is

$$F(x) = F_X(x) = \mathrm{P}(X \leq x), \qquad \text{for all } x.$$

Property 2.1 A distribution function $F_X(x)$ is

1. *Monotonically increasing, i.e., for* $x < y$, $F_X(x) \leq F_X(y)$
2. *Normalized, i.e.,* $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$
3. *Right-continuous, i.e.,* $\lim_{\epsilon \to o^+} F_X(x + \epsilon) = F_X(x)$ *for all* $x \in R$



Right-continuous it means that it is a step function. The cdf can take values greater than 1. The cdf of a continuous random variable is always continuous.

The pmf (probability mass function) gives probabilities that the random variable takes a precise single value.

The **probability mass function (pmf)** of a **discrete random variable** X is defined by

$$f_X(x) = P(X = x)$$

A random experiment with exactly two possible outcomes (for example: heads/tails, male/female, success/failure) is called a Bernoulli trial or Bernoulli random variable.

Property 2.2. *Let X be a discrete random variable with probability mass function $f_X(x)$ and cumulative distribution function $F_X(x)$. Then:*

1. *The probability mass function satisfies $f_X(x) \geq 0$ for all $x \in \mathbb{R}$.*
2. $\sum_i f_X(x_i) = 1.$
3. $F_X(x) = \sum_{i:x_i \leq x} f_X(x_i).$
4. *The values $f_X(x_i) > 0$ are the "jumps" in x_i of $F_X(x)$.*
5. *The cumulative distribution function is a right-continuous step function.*

Learning goal: Describe a binomial, Poisson and Gaussian random variable and recognize their pmf/pdf.

Discrete	Continuous
Bernoulli X~Be(p) → 1 experiment, 2 outcomes p: probability of success	**Uniform** X~Unif(a,b) → eg. Equal probability in a given time frame a,b: beginning and end points
Binomial X~Bin(n,p) → Bernoulli trials with a number of successes n: number of trials p: probability of success	**Normal** X~$\mathcal{N}(\mu, \sigma^2)$ → eg. Distribution of height in the population μ: mean σ^2: variance
Poisson X~Pois(λ) → Number of events in a fixed time frame λ: parameter	**Exponential** X~Exp(λ) → Time until next event eg. Radioactive decay λ: rate

name	description	Example: normal distribution
d*name*( )	density or probability function	dnorm()
p*name*( )	cumulative density function	pnorm()
q*name*( )	quantile function	qnorm()
r*name*( )	random sample	rnorm()

distribution	R name
Normal	norm
Uniform	unif
Binomial	binom
Exponential	exp
Chisquare	quisq
F	f
Student t	T
Poisson	pois
Geometric	geom

Bernoulli's pmf: P(X = 1) = p, P(X = 0), 1 – p, where 0 < p < 1

If a Bernoulli experiment is repeated n times (resulting in an n-tuple of zeros and ones), the distribution of X is called the binomial distribution.

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad 0 < p < 1, \quad k = 0, 1, \ldots, n.$$

The Poisson distribution gives the probability of a given number of events occurring in a fixed interval of time if these events occur with a known and constant rate over time. Examples: number of earthquakes in a given period of time, number of likes of Fb in a given period of time.

Definition 2.3. A random variable X, whose probability function is given by

$$\mathrm{P}(X = k) = \frac{\lambda^k}{k!} \exp(-\lambda), \qquad 0 < \lambda, \quad k = 0, 1, \ldots,$$

is said to follow a Poisson distribution with parameter λ, denoted by $X \sim \mathcal{P}ois(\lambda)$.

The Poisson distribution is also a good approximation for the binomial distribution with large n and small p (as a rule of thumb if n > 20 and np < 10).

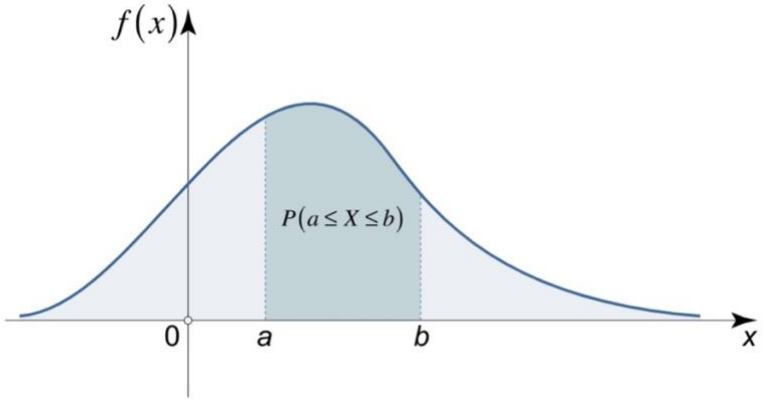
A random variable is called continuous if it can (theoretically) assume any value within one or several intervals. This means that the number of possible values in the sample space is uncountable infinite. We need to consider outcomes that are contained in a specific interval, since if we consider an infinite number of possible outcomes, the likeliness of one particular value being the outcome becomes zero. The probability is described by an integral, as an area under the probability density function.

The density function does not give directly a probability and thus cannot be compared to the probability mass function.

Definition 2.4. The *probability density function* (density function, pdf) $f_X(x)$, or density for short, of a continuous random variable X is defined by

$$\mathrm{P}(a < X \leq b) = \int_a^b f_X(x)dx, \qquad a < b. \tag{2.11}$$

♢



Property 2.3. *Let X be a continuous random variable with density function $f_X(x)$ and distribution function $F_X(x)$. Then:*

i) *The density function satisfies $f_X(x) \geq 0$ for all $x \in \mathbb{R}$ and $f_X(x)$ is continuous almost everywhere.*

ii) $\int_{-\infty}^{\infty} f_X(x)dx = 1.$

iii) $F_X(x) = \int_{-\infty}^{x} f_X(y)dy.$

iv) $f_X(x) = F'_X(x) = \frac{d}{dx}F_X(x).$

v) *The cumulative distribution function $F_X(x)$ is continous everywhere.*

vi) $\mathrm{P}(X = x) = 0.$

Learning goal: Verify that a given function is a pdf/pmf or a cdf. Find a multiplicative constant that makes a given function a pdf/pmf or cdf.

Is $f_X(x) = \begin{cases} \frac{1}{2}x & for\ 0 \leq x \leq 2 \\ 0 & otherwise \end{cases}$ a pdf?

1. $f_X(x) \geq 0$ for any x ✓
2. $\int_0^2 \frac{1}{2}x dx = \left[\frac{1}{4}x^2\right]_0^2 = \frac{1}{4}\cdot 2^2 - \frac{1}{4}\cdot 0^2 = 1$ ✓

Yes, $f_X(x)$ is a pdf.

Consider random variable X with the density $f_X(x) = cx^3, 0 \leq x \leq 1$. Determine c, such that $f(x)$ is a proper density.

$$\int_0^1 f_X(x)\,dx = 1 \Rightarrow \int_0^1 cx^3\,dx = 1$$

$$\Leftrightarrow c \cdot \int_0^1 x^3\,dx = 1$$

$$c\left[\frac{x^4}{4}\right]\Big|_0^1 = 1$$

$$\frac{c}{4} - \frac{c \cdot 0}{4} = 1 \Rightarrow c = 4$$

Learning goal: Pass from a cdf to a quantile function, pdf or pmf and vice versa.

The quantile function is the inverse of the cdf. This means, we are interested in values of x for which $F_X(x) = p$.

Definition 2.5. The quantile function $Q_X(p)$ of a random variable X with (strictly) monotone cumulative distribution function $F_X(x)$ is defined by

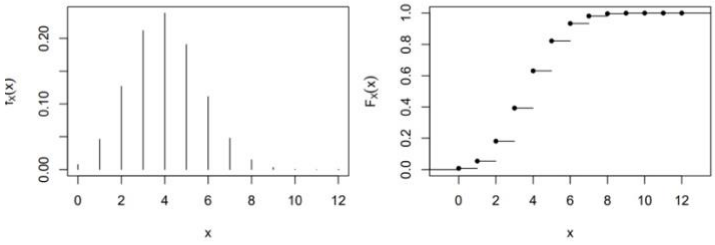
$$Q_X(p) = F_X^{-1}(p), \quad 0 < p < 1, \tag{2.12}$$

i.e., the quantile function is equivalent to the inverse of the distribution function. ♢

In R, the quantile function is specified with the prefix `q` and the corresponding variate. For example, `qunif` is the quantile function for the uniform distribution, which is $Q_X(p) = a + p(b-a)$ for $0 < p < 1$.

Q(0.5) = 0, F(0) = 0.5 -> Q(probability) = x value, F(x value) = probability

Does the inverse function of the cdf below exist? No (multiple values for some argument)



Remark 2.3. For discrete random variables the cdf is not continuous (see the plateaus in the left panel of Figure 2.2) and the inverse does not exist. The quantile function returns the minimum value of x from among all those values with probability $p \leq \mathrm{P}(X \leq x) = F_X(x)$, more formally,

$$Q_X(p) = \inf_{x \in \mathbb{R}} \{p \leq F_X(x)\}, \quad 0 < p < 1, \tag{2.13}$$

Uniform distribution, example quantile function

$$f_X(x) = \begin{cases} \frac{1}{2} & for\ 0 \leq x \leq 2 \\ 0 & otherwise \end{cases}$$

Then $F_X(x) = \int_0^x \frac{1}{2} dy = \frac{1}{2}x$

To find $Q(p)$, we need to find an inverse function to $F_X(x)$.

$$p = \frac{1}{2}x$$
$$x = 2p$$
$$Q(p) = 2p$$

$Q(0.5) =?$ Median $2 \cdot 0.5 = 1$

$Q(0.3) =?$ $2 \cdot 0.3 = 0.6$

Definition 2.6. The median ν of a continuous random variable X with cumulative distribution function $F_X(x)$ is defined by $\nu = Q_X(1/2)$. Accordingly, the lower and upper quartiles of X are $Q_X(1/4)$ and $Q_X(3/4)$. ♢

Learning goal: Give the definition and intuition of an expected value (E), variance (Var), know the basic properties of E, Var used for calculation.

Definition 2.7. The expectation of a discrete random variable X is defined by

$$\mathrm{E}(X) = \sum_i x_i \, \mathrm{P}(X = x_i) \, .$$

The expectation of a continuous random variable X is defined by

$$\mathrm{E}(X) = \int_{\mathbb{R}} x f_X(x) dx \, ,$$

where $f_X(x)$ denotes the density of X.

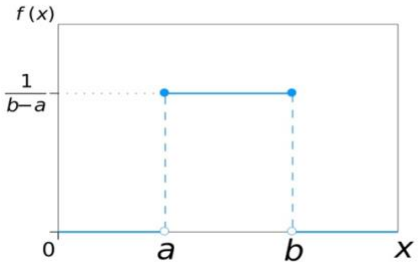
Properties of expectation and variance:

Property 2.5. *For random variables X and Y, regardless of whether discrete or continuous, and for a and b given constants, we have*

1. $\mathrm{Var}(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2$;
2. $\mathrm{E}(a + bX) = a + b\,\mathrm{E}(X)$;
3. $\mathrm{Var}(a + bX) = b^2 \,\mathrm{Var}(X)$,
4. $\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$

Uniform distribution: also known as a rectangular distribution, is a type of probability distribution in which all outcomes are equally likely. Cdf of uniform distribution has slop of 1 and is line y = x.

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & for\ a \le x \le b \\ 0 & otherwise \end{cases}$$



The sum of Gaussian random variables is distributed as a Gaussian random variable. The sum of normal distributions also follows a normal distribution, the mean being the

sum of the means, and the standard deviation being the square root of the variance of the sum, which is equal to the sum of the variances of the individual distributions. Example: Let X and Y be two independent normal random variables with means μ1 and μ2, and variances σ1^2 and σ2^2, respectively. Then the sum Z = X + Y follows a normal distribution with mean μ1 + μ2 and variance σ1^2 + σ2^2.

Definition 2.9. The random variable X is said to be normally distributed if the cumulative distribution function is given by

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx \tag{2.23}$$

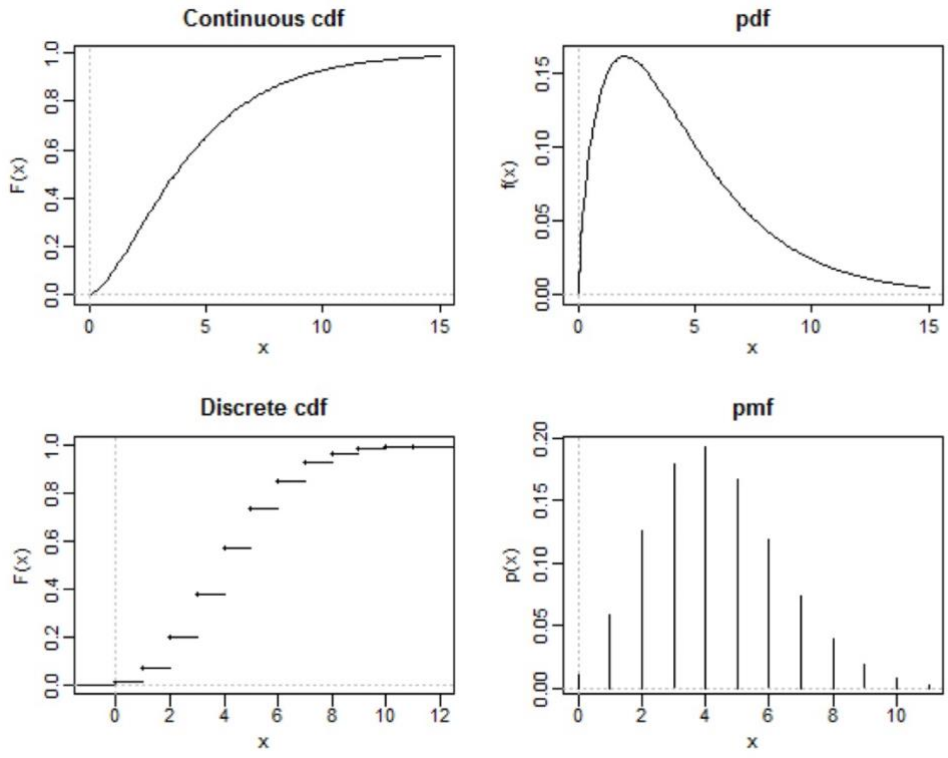
with density function

$$f(x) = f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right), \tag{2.24}$$

for all x ($\mu \in \mathbb{R}$, $\sigma_x > 0$). We denote this with $X \sim \mathcal{N}(\mu, \sigma^2)$.

The random variable $Z = (X - \mu)/\sigma$ (the so-called *z-transformation*) is standard normal and its density and distribution function are usually denoted with $\varphi(z)$ and $\Phi(z)$, respectively. ♢

Property 2.7. *Let* $X \sim \mathcal{N}(\mu, \sigma^2)$, *then* $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$ *and* $F_X(x) = \Phi\left(\frac{X-\mu}{\sigma}\right)$. *Conversely, if* $Z \sim \mathcal{N}(0,1)$, *then* $\sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2)$, $\sigma > 0$.



Questions on Lecture 2:

- The probability of having 2 heads when tossing a coin 3 times is < 0.5 -> True (calculating bin(3, 2) results in 3/8)
- The probability of having at least 2 heads when tossing a fair coin 3 times is = 0.5 -> True (calculating bin(3,2) + bin(3,3) results in 0.5)
- For any cdf F, F(100) > F(0) -> False, F(100) >= F(0) (= e.g., for uniform distribution)
- If a random variable X is normally distributed, then P(X < x) = P(X <= x) -> True
- We consider X a random variable following the binomial distribution. If X ~ Bin(n,p), where n > 5, then P(X <= 5) >= P(X < 5) -> True
- The value of the density function can be greater than 1 -> True
- The value of the pmf can be greater than 1 -> False
- If Q is a quantile function, Q(0.5) is a median of the corresponding distribution

Lecture 3: Functions of Random Variables

Learning goal: know the definition and properties of independent and identically distributed (iid) random variables.

Two events A and B are independent if $P(A \cap B) = P(A)P(B)$ -> knowing anything about the event B does not change the probability (knowledge) of the event A.

Two random variables are independent, if the realization of one does not affect the probability distribution of the other.

Examples of independency: number on the alphabetical list and weight

	Y=0	Y=3	Y=4	
X=5	1/7	1/7	1/7	3/7
X=8	3/7	0	1/7	4/7
	4/7	1/7	2/7	

By the definition, random variable X and Y are independent, if for any events A, B P(X \in A and Y \in B)=P(X \in A)*P(Y \in B)
So in the table above, we would need the following conditions to be fulfilled:

- P(X = 5 and Y =0)=P(X=5)*P(Y=0) <->3/7*4/7=1/7. Not true, so here X and Y are not independent. We don't need to check the remaining ones.
- P(X = 5 and Y =3)=P(X=5)*P(Y=3)
- P(X = 5 and Y =4)=P(X=5)*P(Y=4)
- P(X = 8 and Y =0)=P(X=8)*P(Y=0)
- P(X = 8 and Y =3)=P(X=8)*P(Y=3)
- P(X = 8 and Y =4)=P(X=8)*P(Y=4)

	$Y=0$	$Y=1$	
X=0	1/9	2/9	1/3
X=1	2/9	4/9	2/3
	1/3	2/3	1

In the second table, to check if X and Y are independent, we would need to check if

- P(X=0, Y=0)=P(X=0)*P(Y=0) ⅓*⅓=1/9
- P(X=0, Y=1)=P(X=0)*P(Y=1) ⅓ * ⅔ =2/9
- P(X=1, Y=0)=P(X=1)*P(Y=0) ⅔*⅓ =2/9
- P(X=1, Y=1)=P(X=1)*P(Y=1) ⅔ * ⅔ =4/9

So X and Y are independent.

Questions: Dependent or independent? Identically or not identically distributed?

- Your body height and my body height are …? iid
- Your body height and your siblings body height are …? dependent, iid
- Your body height and your feet size …? dependent and not iid
- Your body height and your siblings feet size …? dependent, not iid
- Your body height and my feet size …? rodiff, not iid

Properties of independent random variables:

Property 3.1. *Let X and Y be two independent random variables. Then*

1. $\mathrm{Var}(aX+bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y)$.

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$ with $\mathrm{E}(X_1) = \mu$ and $\mathrm{Var}(X_1) = \sigma^2$. Denote $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

2. $\mathrm{E}(\overline{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mathrm{E}(X_1) = \mu$ (sample mean) → $\frac{1}{n}\cdot n\cdot E(X_1) = \mu$

3. $\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathrm{Var}(X_1) = \frac{\sigma^2}{n}$. (sample variance)

$= \frac{1}{n^2}\mathrm{var}(\bar{x}) = \frac{n\cdot \mathrm{var}(\bar{x})}{n^2} = \frac{\mathrm{var}(\bar{x})}{n} = \frac{\sigma^2}{n}$

Here: F means any distribution. E.g. $X_1, X_2, \ldots, X_n$ describe outco of throwing a die.

We would like to compute the expected value and the variance when we throw the die n times.

Var(X + Y) = Var(X) + Var(Y) if X, Y are independent.
Var (X + X) = Var(2X) = 4Var(X) and not Var(X) + Var(X) -> X is not independent with X

Learning goal: know the distribution of the average of iid Gaussian random variables.

Random Variables derived from Gaussian Random Variables:

- **Sum of normal random variables**

E.g. machines distributing blue and red candies into the packages;
$X_R \sim N(\mu_R, \sigma_R^2)$; $X_B \sim N(\mu_B, \sigma_B^2)$; X_B, X_R independent

- **Mean of normal random variables**

E.g. we are considering the distribution of gain in some casino game.
$X_1, \dots, X_n \sim N(\mu, \sigma^2), iid$

- **Sum of squares of normal random variables**

E.g. the measurement error is normally distributed. We look at the sum of squares of errors (otherwise, the negative and positive errors would cancel out)

Standardized mean of normal random variables and ratio of variances of normal random variables (used to compare two sample variances with each other) are also gaussian distributed.

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$ **Population Variance**

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$ **Sample Variance**

Learning goal: Explain in own words how to construct chi-squared, t- and F-distributed random variables.

The distribution of squared standard normal random variables is said to be chi-squared. Formally, let $Z_1, \dots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. The distribution of the random variable

$$X = \sum_{i=1}^{n} Z_i^2 \tag{3.5}$$

is called the chi-square distribution ($\mathcal{X}^2$ distribution) with n degrees of freedom and denoted $X \sim \mathcal{X}_n^2$. The following applies:

$$\mathrm{E}(X) = n; \qquad \mathrm{Var}(X) = 2n. \tag{3.6}$$

The t-distribution is used when standardizing $\bar{x}$. The higher the degrees of freedom, the more it converges to the normal distribution. If you have an independent and identically distributed (iid) normally distributed random variable (RV) with unknown variance, the standardized mean follows a t-distribution with degrees of freedom equal to the sample size minus one (n-1). The t-distribution has better tails than the normal distribution.

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

Here:

$\bar{X}$ - a random variable

μ – a number

$\sqrt{n}$ - a number

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ – a random variable, that's why $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ is not normally distributed, even though $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is normally distributed

t_{n-1} - t-distribution with n-1 degrees of freedom.

For n -> ∞ student converges to Gaussian. For smaller n student distribution has heavier tails

The F-distribution is mainly used to compare two sample variances with each other.

Let $X \sim \mathcal{X}^2_m$ and $Y \sim \mathcal{X}^2_n$ be two independent random variables. The distribution of the random variable

$$W = \frac{X/m}{Y/n} \tag{3.10}$$

is called the F-distribution with m and n degrees of freedom and denoted $W \sim F_{m,n}$. It holds that:

$$\mathrm{E}(W) = \frac{n}{n-2}, \qquad \text{for } n > 2; \tag{3.11}$$

$$\mathrm{Var}(W) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, \qquad \text{for } n > 4. \tag{3.12}$$

Learning goal: Know the central limit theorem (CLT) and being able to approximate binomial random variables.

Using the Central Limit Theorem, it can be shown that the distribution of a binomial random variable converges to the distribution of a normal random variable when n tends to infinity (as a guideline: n > 30 and p approx. = 0.5)

Given a sufficiently large sample size, the sampling distribution of the mean of a variable will approximate a normal distribution regardless of that variable's distribution in a population.

Let $X_1, X_2, X_3, \ldots$ an infinite sequence of iid random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$.

Then

$$\lim_{n\to\infty} P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z\right) = \Phi(z)$$

where $\Phi(z)$ is a cdf of a standard normal distribution $N(0,1)$.

Learning goal: be able to calculate the cdf of transformed random variable, and if applicable, the pdf as well.
Learning goal: be able to approximate the mean and variance of transformed random variables.

Learning goal: R: set.seed(), loops
Set.seed() fixes the randomness of sampling, it's used to have reproducible simulations.

Questions on Lecture 3:

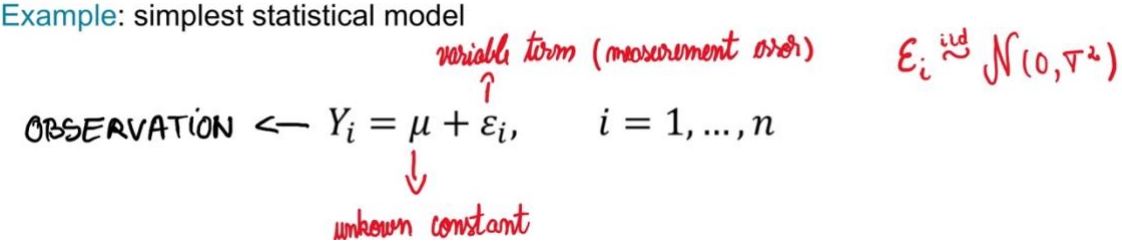
- Var(X – Y) = Var(X) + Var(Y) -> True.
- If X and Y are independent normal random variables, then X*Y is also normally distributed -> False.

Lecture 4: Estimation of Parameters

Inferential statistics: the idea is to draw information from a sample to the whole population and make (infer from these) general statements about the whole population.

Learning goal: Explain what a simple statistical model is (including the role of parameters).

Example: simplest statistical model



Learning goal: Describe the concept of point estimation and interval estimation.
The goal of point estimation is to provide a plausible value for the parameters of the distribution based on the data at hand.

Definition 4.1. A *statistic* is an arbitrary function of a random sample $Y_1, \ldots, Y_n$ and is therefore also a random variable.

An *estimator* for a particular parameter is a statistic used to obtain a plausible value for this parameter based the random sample.

A *point estimate* is the value of the estimator evaluated at $y_1, \ldots, y_n$, the realizations of the random sample.

Estimation (or *estimating*) is the process of finding a (point) estimate. ♢

In order to estimate a parameter, we start from an estimator for that particular parameter and evaluate the estimator at the available data. An estimator is a random variable as well.

Example 4.2. 1. The numerical values shown in R-Code 4.1 are estimates.

2. $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is an estimator.

   $\bar{y} = \frac{1}{100}\sum_{i=1}^{100} y_i = 32.40$ is a point estimate.

3. $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ is an estimator.

   $s^2 = \frac{1}{n-1}\sum_{i=1}^{100}(y_i - \bar{y})^2 = 2.242$ or $s = 1.498$ are a point estimates. ♣

Often, we denote parameters with Greek letters (μ, σ, λ, ...), with θ being the generic one. The estimator and estimate of a parameter θ are denoted by $\widehat{\theta}$. Context makes clear which of the two cases is meant.

Interval Estimation:

Learning goal: Describe the concept of method of moments, least squares and likelihood estimation.

Method of Moments is based on the idea that the parameters of a distribution are expressed as functions of the moments, e.g., E(Y), $E(Y^2)$.

$$\mu := \mathrm{E}(Y)\,, \qquad \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}\,, \tag{4.6}$$

$$\mu_2 := \mathrm{E}(Y^2)\,, \qquad \widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2\,. \tag{4.7}$$

By using the observed values of a random sample in the method of moments estimator, the estimates of the corresponding parameters are obtained. If the parameter is a function of the moments, we need to additionally solve the corresponding equation, as illustrated in the following two examples.

Example 4.3. Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{E}(\lambda)$

$$\mathrm{E}(Y) = 1/\lambda, \qquad \overline{Y} = 1\Big/\widehat{\lambda} \qquad \widehat{\lambda} = \widehat{\lambda}_{\mathrm{MM}} = \frac{1}{\overline{Y}}\,. \tag{4.8}$$

Thus, the method of moment estimate of λ is the value $1/\bar{y}$. ♣

The k-th moment mu_k of a random variable X is defined as $mu_k = E(X^k)$ (provided it exists).
Likelihood Estimation: the likelihood method chooses as estimate the value such that the observed data is most likely to stem from the model (using the estimate). For a given distribution, we call L(θ) the likelihood function, or simply the likelihood. Likelihood must be at least >=0, and can be greater than 1. The likelihood function is the probability mass or density function of the observed data x, viewed as a function of the unknown parameter theta -> find maximum with derivation.

Definition 4.2. The maximum likelihood estimate $\widehat{\theta}_{\mathrm{ML}}$ of the parameter θ is based on maximizing the likelihood, i.e.

$$\widehat{\theta}_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmax}}\, L(\theta). \qquad \diamondsuit \tag{4.12}$$

Least Squares Estimation: aims to minimize the sum of squares of the differences between the random variables and the location parameter. In a linear regression setting, the least squares method minimizes the sum of squares of the differences between observed responses and those predicted by a linear function of the explanatory variables.

Definition 4.3. An estimator $\widehat{\theta}$ of a parameter θ is *unbiased* if

$$\mathrm{E}(\widehat{\theta}) = \theta, \tag{4.16}$$

otherwise it is biased. The value $\mathrm{E}(\widehat{\theta}) - \theta$ is called the *bias*. $\diamondsuit$

$$\mathrm{MSE}(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right) = E\left(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2\right)$$
$$= E(\hat{\theta}^2) - 2E(\hat{\theta}\theta) + E(\theta^2)$$
$$= E(\hat{\theta}^2) - 2\theta E(\hat{\theta}) + \theta^2$$

$$\mathrm{Bias}(\hat{\theta})^2 = E(\hat{\theta})^2 - 2\theta E(\hat{\theta}) + \theta^2$$

$$\mathrm{Var}(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2$$

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Bias}(\hat{\theta})^2 + \mathrm{Var}(\hat{\theta})$$

Learning goal: Construct theoretically and using R confidence intervals.
Learning goal: Interpret point estimates and confidence intervals.

CI 1: Confidence interval for the mean μ

Under the assumption of a normal random sample,

$$\left[\overline{Y} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{Y} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

is an exact $(1-\alpha)$ confidence interval and

$$\left[\overline{Y} - z_{1-\alpha/2}\frac{S}{\sqrt{n}}, \overline{Y} + z_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

an approximate $(1-\alpha)$ confidence interval for μ.

Use t when variance is unknown (most real life cases), and z when variance is known.

CI 2: Confidence interval for the variance σ^2

Under the assumption of a normal random sample,

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right]$$

is an exact $(1-\alpha)$ confidence interval for σ^2.
For a random sample with $n > 50$

$$\left[S^2 - z_{1-\alpha/2}\frac{\sqrt{2}S^2}{\sqrt{n}}, S^2 + z_{1-\alpha/2}\frac{\sqrt{2}S^2}{\sqrt{n}}\right]$$

is an approximate $(1-\alpha)$ confidence interval for σ^2.

Mean Square Error:

A criterion to compare estimators

$$MSE(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right)$$

($\hat{\theta}$: estimator, θ: real parameter)

$$MSE(\hat{\theta}) = bias(\hat{\theta})^2 + Var(\hat{\theta})$$

We aim to construct unbiased estimators with minimal variance.

Lecture 5: Statistical Testing

Learning goal: Explain the concepts of hypothesis and significance test.

The idea of a statistical testing procedure is to formulate a statistical hypothesis and to draw conclusions from them based on the data. Data is compared to the null hypothesis H_0.

The p-value is the probability under the distribution of the null hypothesis of obtaining a result equal to or more extreme than the observed result.

The workflow of a statistical significance test can be summarized as follows. The starting point is a scientific question or hypothesis and data that has been collected to support the scientific claim.

(i) Formulation of the statistical model and statistical assumptions. Formulate the scientific hypothesis in terms of a statistical one.

(ii) Selection of the appropriate test or test statistic and formulation of the null hypothesis H_0 with the parameters of the test.

(iii) Calculation of the p-value,

(iv) Interpretation of the results of the statistical test and conclusion.

Although the workflow is presented in a linear fashion, there are several dependencies. For example the interpretation depends not only on the p-value but also on the null hypothesis in terms of proper statistical formulation and, finally, on the scientific question to be answered, see

Hypothesis testing starts with a null hypothesis H_0 and an alternative hypothesis, denoted by H_1 or H_A.

Learning goal. Define p-value and the significance level.

The p-value gives the probability that the alternative hypothesis is true.

Definition 5.2. The *significance level* α is a threshold determined before the testing, with $0 < \alpha < 1$ but often set to 5% or 1%.

The *rejection region* of a test includes all values of the test statistic for which we reject the null hypothesis. The boundary values of the rejection region are called *critical values*. ♢

If the probability of observing our data under the null hypothesis is below this threshold, we reject the null hypothesis.

Learning goal: Know the difference between one-sample and two-sample t-tests.
The one-tailed test provided more power to detect an effect in one direction by not testing the effect in the other direction. But use it only, when only one direction is of interest.

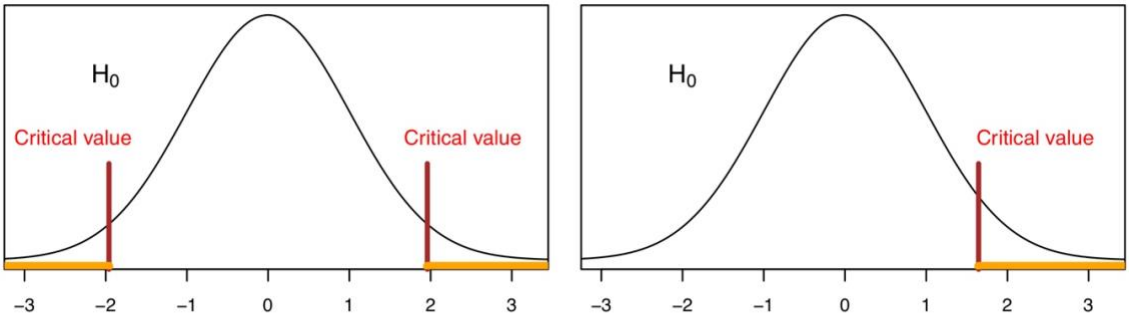
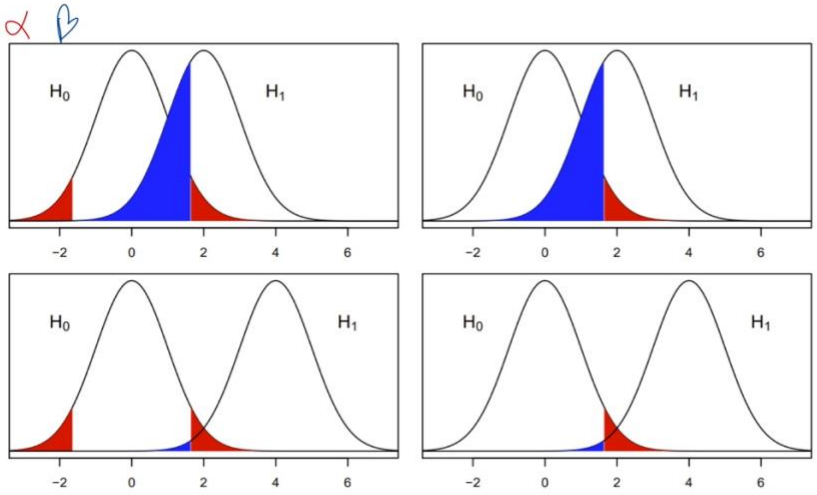


Figure 5.3: Critical values (red) and rejection regions (orange) for two-sided $H_0 : \mu = \mu_0 = 0$ (left) and one-sided $H_0 : \mu \leq \mu_0 = 0$ (right) hypothesis test with significance level $\alpha = 5\%$.

In significance testing two types of errors can occur. Type I errors: we reject H_0 if we should have not and Type II errors: we fail to reject H_0 if we should have. The framework of hypothesis testing allows us to express the probabilities of committing these two errors. The probability of committing a Type I error is exactly $\alpha = \mathrm{P}(\text{reject } H_0 | H_0)$. This probability is often called the *size* of the test. To calculate the probability of committing a Type II error, we need to assume a specific value for our parameter within the alternative hypothesis, e.g., a simple alternative. The probability of Type II error is often denoted with $\beta = \mathrm{P}(\text{not rejecting } H_0 | H_1)$. Table 5.1 summarizes the errors in a classical 2×2 layout.

Table 5.1: Probabilities of Type I and Type II errors in the setting of a significance test.

		True but unknown state	
		H_0 true	H_1 true
Test result	do not reject H_0	$1-\alpha$	β
	reject H_0	α	$1-\beta$



When reducing the significance level α, the critical values move further from the center of the density under H_0 and thus to an increase of the Type II error β -> intuition: if the data stems from H_1 which is far from H_0, the chance that we reject is large. The type II error depends on the distance of the distribution of H0 and H1. The Type I error does not directly depend on the distance between the distributions of H0 (null hypothesis) and H1 (alternative hypothesis). Instead, it depends on the significance level (α) chosen for the hypothesis test.

As a summary, the Type I error

- is defined a priori by selection of the significance level (often 5%, 1%),
- is not influenced by sample size,
- is increased with multiple testing of the same data (we discuss this in Section 5.5.2)

and the Type II error

- depends on sample size and significance level α,
- is a function of the alternative hypothesis.

The value 1 – β is called the power of a test. High power of a test is desirable in an experiment: we want to detect small effects with a large probability.

What do we use a t-test for?

→ To test whether a given sample has a certain mean (or to compare it to the mean of another sample)

Why is it a t-test?

→ Because the test statistics follows a Student-t distribution. Why? → difficult mathematical derivation

When is a t-test two sample?

→ Comparing two samples in terms of their means (instead of just checking the mean of a single sample)

When is a t-test two-sided?

→ Test a hypothesis of equality (instead a hypothesis of "greater than" or "smaller than", which is one-sided)

Test 1: Comparing a sample mean with a theoretical value

Question: Does the sample mean deviate significantly from the postulated but unknown mean?

Assumptions: The observed data is a realization of a Gaussian random sample with unknown mean and variance.

Calculation: $t_{\text{obs}} = \dfrac{|\bar{x} - \mu_0|}{s/\sqrt{n}}$.

Decision: Reject $H_0 : \mu = \mu_0$, if $t_{\text{obs}} > t_{\text{crit}} = t_{n-1,1-\alpha/2}$.

Calculation in R: `t.test( x, mu=mu0, conf.level=1-alpha)`

Assumptions for t-tests:

1) Normality: Q-Q plot, otherwise
2) Outliers
3) When comparing two groups: equal variances
4) Independence

Comparing means from two independent samples:

Question: Are the means and of two samples significantly different?

Assumptions: Both samples are normally distributed with the same unknown variance. The samples are independent.

Calculation: $t_{\text{obs}} = \dfrac{|\bar{x} - \bar{y}|}{s_p / \sqrt{1/n_x + 1/n_y}} = \dfrac{|\bar{x} - \bar{y}|}{s_p} \cdot \sqrt{\dfrac{n_x \cdot n_y}{n_x + n_y}}$,

where $s_p^2 = \dfrac{1}{n_x + n_y - 2} \cdot \left((n_x - 1)s_x^2 + (n_y - 1)s_y^2\right)$.

Decision: Reject $H_0 : \mu_x = \mu_y$ if $t_{\text{obs}} > t_{\text{crit}} = t_{n_x+n_y-2,1-\alpha/2}$.

Calculation in R: `t.test( x, y, var=TRUE, conf.level=1-alpha)`

Comparing means from two paired samples: E.g., observation before and after an intervention. The researcher records the happiness level of each subject in the study, before and after a drug is administered. These measurements would be paired data, since each before measure is related only to the after measure from the same subject.

Question: Are the means $\bar{x}$ and $\bar{y}$ of two paired samples significantly different?

Assumptions: The samples are paired. The differences are normally distributed with unknown mean δ. The variance is unknown.

Calculation: $t_{\text{obs}} = \dfrac{|\bar{d}|}{s_d/\sqrt{n}}$, where

- $d_i = x_i - y_i$ is the i-th observed difference,
- $\bar{d}$ and s_d are the mean and the standard deviation of the differences d_i.

Decision: Reject $H_0 : \delta = 0$ if $t_{\text{obs}} > t_{\text{crit}} = t_{n-1,1-\alpha/2}$.

Calculation in R: `t.test(x, y, paired=TRUE, conf.level=1-alpha)` or

`t.test(x-y, conf.level=1-alpha)`

Multiple testing: P(at least 1 false significant results) = 1 – P(no false significant results = $1 - (1 - \alpha)^m$

Questions on Lecture 5:

- We consider the dish washing example: the more patient you are, the smaller significance level alpha you choose -> True.
- In the paired two sample test, we always have the same number of observations in the two groups that we are comparing -> True.
- When $(\bar{x} - \mu) / s / \sqrt{n}$ has a larger absolute value, the p-value gets smaller -> True

Lecture 6: Estimating and Testing Proportions

If one of the assumptions of normal distributed observations or equal variances between two samples is violated, non-parametric tests should be used; since these do not have any assumption about the distribution of the data.

- Paired Tests:
  - Sign Test
  - Wilcoxon Signed Rank Test
- Unpaired Tests:
  - Wilcoxon-Mann-Whitney Test
  - Permutation Test

Estimating p in a binomial distribution: Method of moments:

$$P_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$E(X) = np$ (the first moment) (see slide 9 for derivation)

We have only one observation (number of successes), which is the estimate of the mean.
For the random variable X, describing the number of successes we have:

$$n\hat{p} = X$$

The estimator of p (based on the method of moments),

$$\hat{p}_{MM} = \frac{X}{n}$$

Estimation of p in a binomial distribution: Maximum likelihood:

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

It is often easier to work with log-likelihood. log() is a monotonic function, so the $\hat{p}_{ML}$ based on the log-likelihood will be for the same as $\hat{p}_{ML}$ based on the likelihood

$$l(p) = \log\left(\binom{n}{x}\right) + x \log p + (n-x)\log(1-p)$$

To find the maximum, we compute the derivative of the log-likelihood:

$$\frac{\mathrm{d}l(p)}{\mathrm{d}p} = \frac{x}{p} - \frac{n-x}{1-p}$$

And set the derivative to zero.

10

Table 6.1: Example of a two-dimensional contingency table displaying frequencies. The first index refers to the presence of the risk factor, the second to the diagnosis.

		Diagnosis		
		positive	negative	Total
Risk	with factor	h_{11}	h_{12}	n_1
	without	h_{21}	h_{22}	n_2

with each other. To simplify the exhibition, we discuss the estimation using one of the two risk factors only.

The goal of the test of proportion is to check if the proportions of people with the disease are the same or different in the two groups (with and without some factor). If the proportion of people with the disease is the same in both groups, we would expect:

$$\frac{total_{positive}}{total_{total}} \cdot with.factor$$

Or equivalently:

$$\frac{h_{11} + h_{21}}{h_{11} + h_{12} + h_{21} + h_{22}} \cdot (h_{11} + h_{12})$$

to be a number of people with the disease in the group with the factor.

CI 3: Confidence intervals for proportions

An approximate $(1-\alpha)$ Wald confidence interval for a proportion is

$$B_{l,u} = \widehat{p} \pm q \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \tag{6.10}$$

with estimator $\widehat{p} = X/n$ and quantile $q = z_{1-\alpha/2}$. An approximate $(1-\alpha)$ Wilson confidence interval for a proportion is

$$B_{l,u} = \frac{1}{1+q^2/n} \cdot \left(\widehat{p} + \frac{q^2}{2n} \pm q \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{q^2}{4n^2}} \right). \tag{6.11}$$

In R an exact $(1-\alpha)$ confidence interval for a proportion is computed with `binom.test( x, n)$conf.int`.

Wilson's confidence interval is more complex but also more precise than Wald's.

For small and very large values x, the Wald confidence interval has a way too small coverage and thus wider intervals are desired.

Test 5: Test of proportions

Question: Are the proportions in the two groups the same?

Assumptions: Two sufficiently large, independent samples of binary data.

Calculation: We use the notation for cells of a contingency table, as in Table 6.1. The test statistic is

$$\chi^2_{\text{obs}} = \frac{(h_{11}h_{22} - h_{12}h_{21})^2(h_{11}+h_{12}+h_{21}+h_{22})}{(h_{11}+h_{12})(h_{21}+h_{22})(h_{12}+h_{22})(h_{11}+h_{21})}$$

and, under the null hypothesis that the proportions are the same, is $\mathcal{X}^2$ distributed with one degree of freedom.

Decision: Reject if $\chi^2_{\text{obs}} > \chi^2_{\text{crit}} = \chi^2_{1,1-\alpha}$.

Calculation in R: `prop.test( tab)` or `chisq.test(tab)`

Test 6: Comparison of observations with expected frequencies

Question: Do the observed frequencies o_i of a sample deviate significantly from the expected frequencies e_i of a certain distribution?

Assumptions: Sample with data from any scale.

Calculation: Calculate the observed values o_i and the expected frequencies e_i (with the help of the expected distribution) and then compute

$$\chi^2_{\text{obs}} = \sum_{i=1}^{N} \frac{(o_i - e_i)^2}{e_i}$$

where N is the number of categories.

Decision: Reject H_0: "no deviation between the observed and expected" if $\chi^2_{\text{obs}} > \chi^2_{\text{crit}} = \chi^2_{N-1-k,1-\alpha}$, where k is the number of parameters estimated from the data to calculate the expected counts.

Calculation in R: `chisq.test( obs, p=expect/sum(obs))` or `chisq.test( obs, p=expect, rescale.p=TRUE)`

Learning goal: Explain and apply different methods of comparing proportions (Difference Between Proportions, Odds Ratio, Relative Risk).

Relative Risk: estimates the size of the effect of a risk factor compared with the effect size when the risk factor is not present -> RR = P(Positive diagnosis with risk factor) / P(Positive diagnosis without risk factor). The relative risk assumes positive values, a value of 1 means that the risk is the same in both groups and there is no evidence of an association between the diagnosis/disease/event and the risk factor. A value greater than 1 is evidence of a possible positive association between a risk factor and a diagnosis/disease. If the relative risk is less than one, the exposure has a protective effect, as is the case, for example, for vaccinations.

$$\widehat{\text{RR}} = \frac{\widehat{p}_1}{\widehat{p}_2} = \frac{\dfrac{h_{11}}{h_{11}+h_{12}}}{\dfrac{h_{21}}{h_{21}+h_{22}}}.$$

Odds Ratio:

$$\mathrm{OR} = \frac{\frac{\mathrm{P}(\text{Positive diagnosis with risk factor})}{\mathrm{P}(\text{Negative Diagnosis with risk factor})}}{\frac{\mathrm{P}(\text{Positive diagnosis without risk factor})}{\mathrm{P}(\text{Negative diagnosis without risk factor})}}$$

$$= \frac{\frac{\mathrm{P}(A)}{1-\mathrm{P}(A)}}{\frac{\mathrm{P}(B)}{1-\mathrm{P}(B)}} = \frac{\mathrm{P}(A)(1-\mathrm{P}(B))}{\mathrm{P}(B)(1-\mathrm{P}(A))}$$

with A and B the positive diagnosis with and without risk factors. The odds ratio indicates the strength of an association between factors. When a disease is rare (very low probability of disease), the odds ratio and relative risk are approximately equal. For odds ratio, if 1 is not in the CI, the null hypothesis that states that the odds are the same in both groups is rejected.

Odds ratio

Text book definition: the strength of an association between factors

In our example: just ratio between the odds of Schumacher winning and the odds of Vettel winning.

! Important: a value greater than 1, the event is more likely to occur in the first group; the converse is true. value = 1: doesn't matter who drives, odds are the same

$$\mathrm{OR} = \frac{\frac{\mathrm{P}(\text{Schumacher wins})}{\mathrm{P}(\text{Schumacher doesn't win})}}{\frac{\mathrm{P}(\text{Vettel wins})}{\mathrm{P}(\text{Vettel doesn't win})}} = \frac{\frac{91}{214}}{\frac{53}{227}} \simeq 1.82$$

(Schumacher has a higher odd)

Fisher's test: good to use for small datasets, the larger the dataset the more exact the normal approximation will be so it's better to use Fisher' exact test for small datasets. The test is based on calculating the exact probability of observing a particular set of frequencies in the contingency table under the null hypothesis of independence between the variables. A low p-value (usually less than 0.05) suggests that the null hypothesis can be rejected, and there is a significant association between the variables. A high p-value indicates that there is insufficient evidence to reject the null hypothesis, and the association between the variables may be due to chance.

Learning goal: Compare different CI for proportions (Wald, Wilson).
Wilson should be used if you have a higher probability (more exact). Wald uses basic assumption that data is normally distributed. In R it's better to use Wilson.

Questions on Lecture 6:

- The confidence interval for the relative risk is [0.6, 1.1], will you reject the null hypothesis, that the risk in the two groups is the same? Since 1 is in the CI, we do not reject it.

- The confidence interval for the difference between proportions is [0.6, 1.1], will you reject the null hypothesis, that the risk in the two groups is the same? 0 is not in the CI, so we reject it.
- The confidence interval for the Odds Ratio is [0.2, 0.8], will you reject the null hypothesis, that the risk in the two groups is the same? 1 is not in the CI, so we reject it.
- If the chi-square observed/measured value is greater than the chi-square critical value, the null hypothesis gets rejected -> True.
- The X^2 random variable with df=n is a sum of n iid squared standard normal variables -> True
- For the maximum likelihood method we (can) take the log of the likelihood because the log function is monotonically increasing, and not because it is easier to take the derivative of a product nor because only the likelihood yields an optimal parameter.
- In the discrete case P(X<=20) are the same -> False, would be true for continuous case.
- The Wald and the Wilson CI use a normal approximation -> True
- The Wald confidence interval has a better coverage -> False
- The Wilson confidence interval is more precise -> True
- If there is no difference between two proportions the relative risk is 1 -> True.

Lecture 7: Rank-Based Methods

Learning goal: Explain robust estimators and their properties.

A robust estimator is an estimator of a parameter that is not sensitive to one or possibly several outliers. Sensitive means that if we replace one or more values of the sample with arbitrarily values, the corresponding estimate does not or only marginally change -> often no specific distribution assumption.

Disadvantages of robust estimators: not an easy distribution of the estimator (but CI can be approximated) and low efficiency (these estimators have large variances) when t-test assumptions are fulfilled. A more efficient estimator needs fewer observations to achieve a given performance. The efficiency is the ratio of the variance of one estimator to the variance of the second estimator.

The rank of a value in a set of values is the position of that value in the ordered sequence from smallest to largest. The smallest rank has rank 1 and the largest rank n (in case of ties, the arithmetic mean of the ranks is used).

Test 7: Comparing the locations of two paired samples with the sign test

Question: Are the medians of two paired samples significantly different?

Assumptions: The samples are paired and from continuous distributions.

Calculation: (1) Calculate the differences $d_i = x_i - y_i$. Ignore all differences $d_i = 0$ and consider only the $n_\star$ differences $d_i \neq 0$.

(2) Categorize each difference d_i by its sign (+ or −) and count the number of positive signs: $S^+ = \sum_{i=1}^{n_\star} \mathbb{I}_{\{d_i>0\}}$. $S_{\text{obs}} = \min(S^+, n_\star - S^+)$.

Decision: Reject H_0 : "the medians are the same", if $S_{\text{obs}} < S_{\text{crit}}$, where S_{crit} is the $\alpha/2$-quantile of a $\mathcal{B}in(n_\star, 1/2)$ distribution.

Calculation in R:

```
binom.test( sum( d>0), sum( d!=0), conf.level=1-alpha)
```

The Wilcoxon signed rank test is used to test an effect in paired samples, i.e., two matched samples or two repeated measurements on the same subject.

H_0: the distributions are the same.
H_1: the distributions are not the same.

Before (X)	After (Y)	$Y - X$	$\lvert Y - X \rvert$	Rank of $\lvert Y - X \rvert$
2	2.6	0.6	0.6	2
4	4.8	0.8	0.8	3
3	2.8	-0.2	0.2	1
6	9.3	3.3	3.3	5
9	8	-1	1	4

H_0: the distributions are the same

H_1: the distributions are not the same

W_+ = sum of ranks of positive differences

W_- = sum of ranks of negative differences

$W_+ = 2 + 3 + 5 = 10$

$W_- = 1 + 4 = 5$

$W = \min(W_-, W_+) = 5$

Under H_0 W follows signrank distribution. (dsignrank, psignrank,... in R)

3

The sign test takes in consideration the number of positive/negative differences. It is used to test the median of paired differences, where the null hypothesis states that the median of the paired differences is 0. The Wilcoxon rank sign test takes in consideration the sum of ranks of positive/negative differences. It is used to test if the distributions of two paired samples are different, where the null hypothesis states that the distributions of the paired samples are the same. The Wilcoxon makes more assumptions about the nature of the distributions under test, but has more statistical power than the sign test, which makes fewer assumptions.

Test 8: Comparing the distribution of two paired samples

Question: Are the distributions of two paired samples significantly different?

Assumptions: Both samples are from continuous distributions of the same shape, the samples are paired.

Calculation: (1) Calculate the differences $d_i = x_i - y_i$. Ignore all differences $d_i = 0$ and consider only the remaining $n_\star$ differences $d_i \neq 0$.

(2) Order the $n_\star$ differences d_i by their absolute differences $|d_i|$.

(3) Calculate the sum of the ranks of the positive differences:
$W^+ = \sum_{i=1}^{n_\star} \mathbb{I}_{\{d_i>0\}}$.
$W^- = \frac{n_\star(n_\star+1)}{2} - W^+$ (sum of the ranks of the negative differences).

$$W_{\text{obs}} = \min(W^+, W^-) \qquad (W^+ + W^- = \frac{n_\star(n_\star+1)}{2})$$

Decision: Reject H_0 : "the distributions are the same" if $W_{\text{obs}} < W_{\text{crit}}(n_\star; \alpha/2)$, where W_{crit} is the critical value.

Calculation in R: `wilcox.test(x-y, conf.level=1-alpha)` or

`wilcox.test(x, y, paired=TRUE, conf.level=1-alpha)`

The Wilcoxon-Mann-Whitney test represents the two-sample version of the Wilcoxon signed rank test. It calculates the rank sums of the two samples and corrects these based on the corresponding sample size. It is used to compare two unpaired samples, where the null hypothesis states that the medians of the two unpaired samples are the same. It follows the same idea as the t-test to compare two unpaired samples, but the assumptions of normality and equal variance are dropped.

Question: Are the medians of two independent samples significantly different?

Assumptions: Both samples are from continuous distributions of the same shape, the samples are independent and the data are at least ordinally scaled.

Calculation: Let $n_x \leq n_y$, otherwise switch the samples. Assemble the $(n_x + n_y)$ sample values into a *single* set ordered by rank and calculate the sum R_x and R_y of the sample ranks. Let

$$U_x = R_x - \frac{n_x(n_x+1)}{2} \qquad U_y = R_y - \frac{n_y(n_y+1)}{2}$$
$$U_{\text{obs}} = \min(U_x, U_y) \qquad \text{Note: } U_x + U_y = n_x n_y.$$

Decision: Reject H_0: "medians are the same" if $U_{\text{obs}} < U_{\text{crit}}(n_x, n_y; \alpha/2)$, where U_{crit} is the critical value.

Calculation in R: `wilcox.test(x, y, conf.level=1-alpha)`

Learning goal: Explain the idea of a permutation test.

Permutation test: follows the idea to reassign the group of treatment to the observations and recalculate the test statistic. Under the null hypothesis, the observed test statistic will be similar compared to the ones obtained by reassigning the groups. The resulting p-value is the proportion of cases that the permutation yielded a more extreme observation that the data.

H_0: difference in median between two groups is 0.

H_1: difference in median between two groups is not 0.

Test 13: Comparing the locations of two independent samples

Question: Are the locations of two independent samples significantly different?

Assumptions: The null hypothesis is formulated, such that under H_0 the groups are exchangeable.

Calculation: (1) Calculate the difference t_{obs} in the means of the two groups to be compared (m observations in group 1, n observations in group 2).

(2) Form a random permutation of the values of both groups by randomly allocating the observed values to the two groups (m observations in group 1, n observations in group 2).

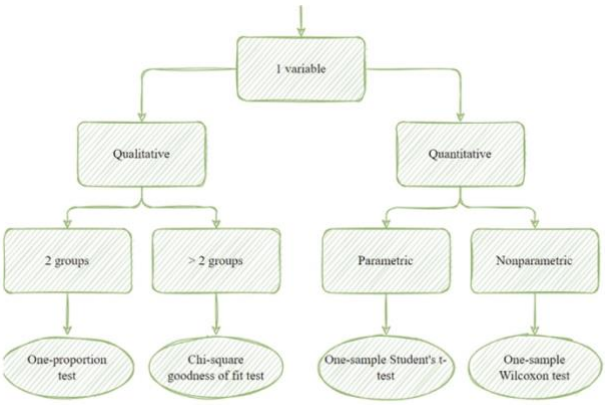
(3) Calculate the difference in means of the two new groups.

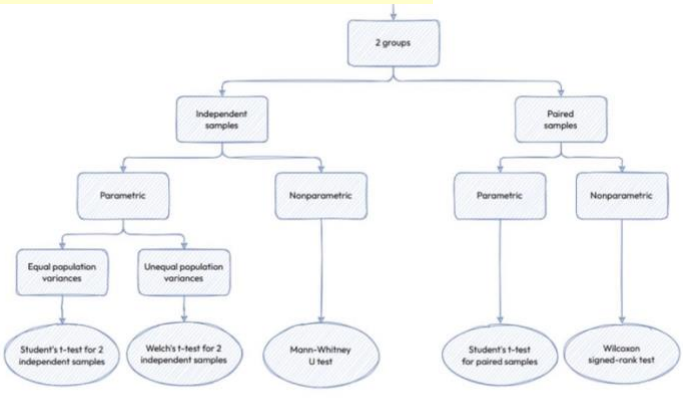
(4) Repeat this procedure R times (R large).

Decision: Compare the selected significance level with the p-value:

$$\frac{1}{R}(\text{number of permuted differences more extreme than } t_{\text{obs}})$$

Calculation in R: `require(coin); oneway_test( formula, data)`



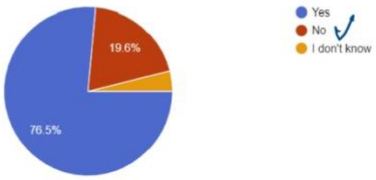


Test	Sign	Wilcoxon Signed Rank	Wilcoxon-Mann-Whitney	Permutation
Paired	Yes	Yes	No	No
Comparison	Medians	Distributions	Medians	Mean or Median
H_0	Difference of X and Y has Median 0	The Distributions of both samples are the same	The Medians in both samples are the same	Difference of the Means/Medians between the permuted groups is 0

Questions on Lecture 7:

We are considering happiness levels in 5 persons before and after treatment. Levels before are (1,3,2,4,2), levels after are (102, 101, 103, 101, 104). We are using sign test to compare the happiness before and after treatment. Will we reject the H0, that the median of the difference of happiness level before and after treatment is 0, at the significance level 1%?

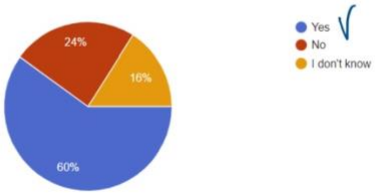
51 responses



If $X \sim Bin(5, 0.5)$, $P(X = 5) = \left(\frac{1}{2}\right)^5 = \frac{1}{32}$
Even if we consider one-sided alternative hypothesis, $\frac{1}{32} > 1\%$, hence we cannot reject the H_0

We are considering the same data set. But here we are using t-test to compare the means. Will we reject the H0, that the means are the same at the significance level 1%?

50 responses



The difference $= (101{,}98{,}101{,}97{,}102)$
$\hat{S} < 5$, $\hat{\mu} \approx 100$
$n = 5$ hence the chance of getting such data when the true mean is 0, is very low.

Can the following be covariance matrices?

- $\begin{pmatrix} 3 & 1 \\ 0 & 1 \end{pmatrix}$ No (not symmetric, $Cov(X_1, X_2) = Cov(X_2, X_1)$)
- $\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$ No (variance cannot be negative)
- $\begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$ Yes
- $\begin{pmatrix} 1 & 10 \\ 10 & 1 \end{pmatrix}$ No, because of the relationship between cov and var

(for the same reason, the correlation is between -1 and 1)

The covariance matrix is symmatric and (*) positive definte.

- If I want to look at the distribution of weight of people that have a height of 1.75m. I use the conditional distribution -> True
- If I want to look at the distribution of height only, I use the marginal distribution -> True
- The joint distribution would be used when you are interested in studying the relationship between height and weight in a population -> True, the joint distribution allows you to analyze how the two variables (height and weight) are distributed together. It can be used to calculate the probability of a specific combination of height and weight occurring within the population, or to determine if there is a correlation between height and weight in your population, and much more.
- In the bivariate case: when computing a probability we are looking for the volume under the surface -> True
- The covariance can take negative values -> True
- If two variables are indepent, their covariance is 0 -> True
- The correlation takes values between -1 and 1 -> True. A correlation coefficient of 1 indicates a perfect positive linear relationship between the two variables, meaning as one variable increases, the other variable also increases proportionally. A correlation coefficient of -1 indicates a perfect negative linear relationship, meaning as one variable increases, the other variable decreases proportionally. A correlation coefficient of 0 signifies that there is no linear relationship between the two variables.
- The correlation can be 0 even if there is a relationship between the two random variable -> True
- The covariance is independent from the scale -> False
- Cov(X,X)=Var(X) -> True

Lecture 8: Multivariate Normal Distribution

Learning goal: Describe a random vector, cdf, pdf of a random vector and its properties.

Random vector - a vector $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)^T = \begin{pmatrix} X_1 \\ X_2 \\ \cdots \\ X_p \end{pmatrix}$ with p random variables as components.

(From lecture 2) The distribution function (**cumulative distribution function, cdf**) of a random variable X is defined as:

$$F_X(x) = P(X \leq x)$$

The multivariate (or multidimensional) distribution function of a random vector $\boldsymbol{X}$ is defined as:

each random variable's smaller (equal) than the corresponding element in the vector

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = P(\boldsymbol{X} \leq \boldsymbol{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots X_p \leq x_p)$$

not random variable vec, but random vector (bold font)

6

Learning goal: Explain and work with conditional and marginal distributions.

Joint Distribution: observation of multiple characteristics at once

$X = (X_1, X_2)$. Pmf: $P(X_1 = x_1, X_2 = x_2)$

Conditional distribution:

$X_1 \mid X_2 = x_2$, e.g., the distribution of X_1 if $X_2 = 1$. Pmf: $P(X_1 = x_1 \mid X_2 = x_2)$

The multivariate (or multidimensional) distribution function of a random vector $\boldsymbol{X}$ is defined as:

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = P(\boldsymbol{X} \leq \boldsymbol{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \dots X_p \leq x_p)$$

Properties of a continuous bivariate random vector:

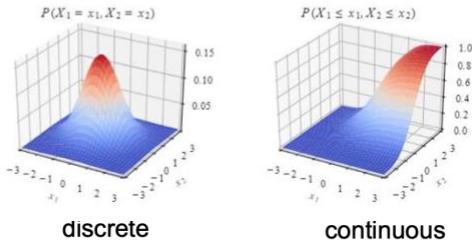
Let $(X, Y)^T$ be a bivariate continuous random vector with joint density function $f_{X,Y}(x, y)$ and joint distribution function $F_{X,Y}(x, y)$.

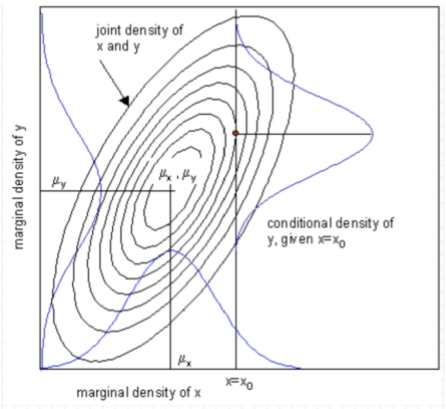
1. The distribution function is monotonically increasing:
for $x_1 \leq x_2$ and $y_1 \leq y_2$, $F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$.

2. The distribution function is normalized:

$$\lim_{x,y\to\infty} F_{X,Y}(x, y) = 1 \qquad \lim_{x,y\to-\infty} F_{X,Y}(x, y) = 0$$

$$\lim_{x\to-\infty} F_{X,Y}(x, y) = 0 \qquad \lim_{y\to-\infty} F_{X,Y}(x, y) = 0$$





- Joint $P(X = x, Y = y)$

$$P(a < X \leq b, c < Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x, y)dxdy$$

- Marginal $P(X = x), P(Y = y)$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx$$

- Conditional $P(X|Y = y)$

Learning goal: Give the definition and intuition of E, Var and Cov for a random vector.

Expected value of a random vector:

$$\mathrm{E}(\mathbf{X}) = \mathrm{E}\left(\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}\right) = \begin{pmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_p) \end{pmatrix}$$

Covariance: measure of the joint variability of two random variables. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for lesser variables (i.e., the variables tend to show similar behavior), the covariance is positive.

Learning goal: Know basic properties of E, Var and Cov for a random vector.

$$\mathrm{Cov}(X_1, X_2) = \mathrm{E}\big((X_1 - \mathrm{E}(X_1))(X_2 - \mathrm{E}(X_2))\big) = \mathrm{E}(X_1 X_2) - \mathrm{E}(X_1)\,\mathrm{E}(X_2).$$

Property 8.2. *We have for arbitrary random variables X_1, X_2 and X_3:*

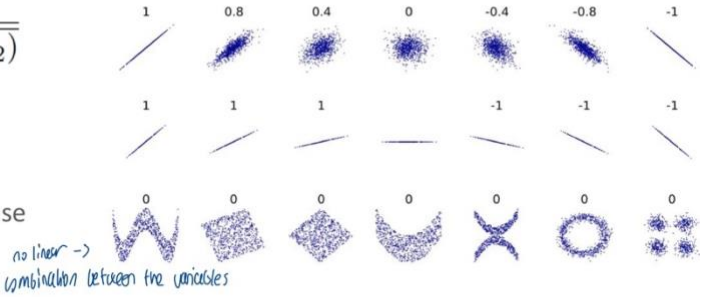
1. $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$, Symmetry
2. $\mathrm{Cov}(X_1, X_1) = \mathrm{Var}(X_1)$, only if we work on only one random variable
3. $\mathrm{Cov}(a + bX_1, c + dX_2) = bd\,\mathrm{Cov}(X_1, X_2)$, *for arbitrary values a, b, c and d,* Covariance scales linearly (not quadratic like with variance)
4. $\mathrm{Cov}(X_1, X_2 + X_3) = \mathrm{Cov}(X_1, X_2) + \mathrm{Cov}(X_1, X_3)$. → covariance of sum: sum of covariances

The correlation between two random variables X1 and X2 is defined as:

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)}}$$

For any random variables X_1, X_2,
$$-1 \leq Corr(X_1, X_2) \leq 1,$$
with equality only in the degenerate case $X_2 = a + bX_1$ for some a and $b \neq 0$.

1 or -1: one variable is linear combination of the other one



no linear → combination between the variables

Q: We are considering the correlation between people's height and weight.
Once we express the height in centimeters, once in meters. Will the correlation be the same? Yes and, not with covariance

The variance of a p-variate random vector $\boldsymbol{X} = \big(X_1, X_2, \cdots, X_p\big)^T$ is defined as:

$$\mathrm{Var}(\mathbf{X}) = \mathrm{E}\big((\mathbf{X} - \mathrm{E}(\mathbf{X}))(\mathbf{X} - \mathrm{E}(\mathbf{X}))^{\mathsf{T}}\big)$$

$$= \mathrm{Var}\left(\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}\right) = \begin{pmatrix} \mathrm{Var}(X_1) & \dots & \mathrm{Cov}(X_i, X_j) \\ & \ddots & \\ \mathrm{Cov}(X_j, X_i) & \dots & \mathrm{Var}(X_p) \end{pmatrix}$$

Variance – covariance matrix, covariance matrix

Learning goal: Recognize the density of Gaussian random vector and know properties of Gaussian random vector.

Bivariate normal distribution:

Property 8.4. *For the bivariate normal random vector as specified by (8.18), we have: (i) The marginal distributions are* $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ *and* $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ *and (ii)*

$$\mathrm{E}\left(\begin{pmatrix} X \\ Y \end{pmatrix}\right) = \begin{pmatrix} \mu_x \\ \mu_y, \end{pmatrix} \qquad \mathrm{Var}\left(\begin{pmatrix} X \\ Y \end{pmatrix}\right) = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}. \tag{8.19}$$

Thus,

$$\mathrm{Cov}(X, Y) = \rho\sigma_x\sigma_y, \qquad \mathrm{Corr}(X, Y) = \rho. \tag{8.20}$$

(iii) If $\rho = 0$, X *and* Y *are independent and vice versa.*

Example:

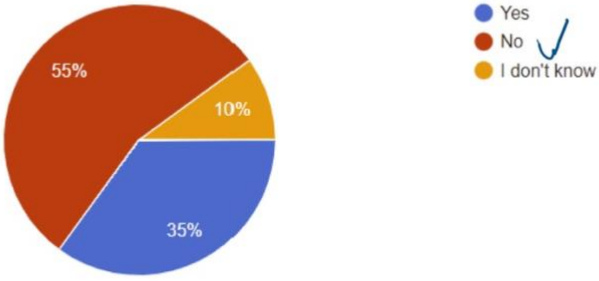
$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N_3(\mu, \Sigma) \qquad \mu = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 2 \end{bmatrix}$$

1. What is the marginal distribution of X_3?
   $X_3 \sim N(-1, 2)$
2. What is $\mathrm{Cov}(X_1, X_2)$? 0
   What is $\mathrm{Cov}(X_1, X_3)$? 1
3. What is the variance of X_1? 3

Questions on Lecture 8:

We are considering a random vector X=(X1, X2, X3, X4). We know the marginal distributions X1, X2, X3, X4. There exists exactly one possible joint distribution of X.

20 responses

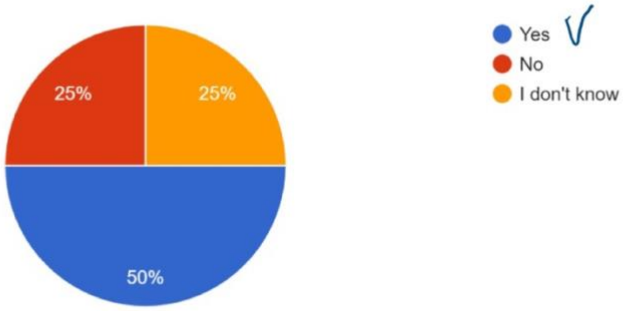


E.g $\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & ? & ? & ? \\ ? & 2 & ? & ? \\ ? & ? & 1 & ? \\ ? & ? & ? & 4 \end{pmatrix} \right)$

There are many possible ways of filling in the covariance matrix. The marginal distributions remain the same.

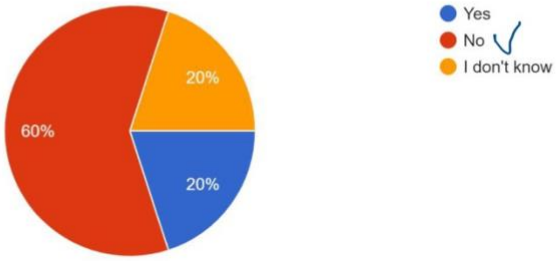
X is multivariate normally distributed. The covariance matrix of X=(X1, X2, X3) is shown below. The expected value is mu = (2, 3, 4). Then (X1, X2) and X3 are independent.

20 responses



$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N_3 \left(\begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}, \begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix} \right)$$

We are considering the same distribution as in the questions above. E(X3*X3) = 16.
20 responses



$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N_3\left(\begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}, \begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{pmatrix}\right)$$

$$Var(X_3) = 2 \ , E(X_3) = 4$$
$$Var(X_3) = E(X_3^2) - E(X_3)^2, \quad hence \quad E(X_3^2) = Var(X_3) + E(X_3)^2 = 2 + 16 = 18 \neq 16$$

Lecture 9: Estimation of Correlation and Simple Regression

Learning goal: Explain the concept of correlation.

Correlation is a measure of linear dependency. Correlation is a symmetric measure, meaning that Corr(X, Y) = Corr(Y, X).

Learning goal: Explain the statistical model of the linear regression.

More formally, in simple linear regression a (dependent) variable is explained linearly through a single independent variable. The statistical model writes as

$$Y_i = \mu_i + \varepsilon_i \tag{9.9}$$
$$= \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{9.10}$$

with

- Y_i: dependent variable, measured values, or observations;
- x_i: independent variable, predictor, assumed known or observed and not stochastic;
- β_0, β_1: parameters (unknown);
- ε_i: measurement error, error term, noise (unknown), with symmetric distribution around zero.

It is often also assumed that $\text{Var}(\varepsilon_i) = \sigma^2$ and that the errors are independent of each other.

Learning goal: Explain different correlation coefficient estimates.

$$r = \widehat{\mathrm{Corr}}(X,Y) = \frac{\widehat{\mathrm{Cov}}(X,Y)}{\sqrt{\widehat{\mathrm{Var}}(X)\widehat{\mathrm{Var}}(Y)}} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_j (y_j - \overline{y})^2}},$$

The estimate r is called the **Pearson correlation coefficient**

Test 14: Test of correlation

Question: Is the correlation between two paired samples significant?

Assumptions: The pairs of values stem from a bivariate normal distribution.

Calculation: $t_{\text{obs}} = |r| \frac{\sqrt{n-2}}{\sqrt{1-r^2}}$ where r is the Pearson correlation coefficient (9.1).

Decision: Reject H_0: $\rho = 0$ if $t_{\text{obs}} > t_{\text{crit}} = t_{n-2,1-\alpha/2}$.

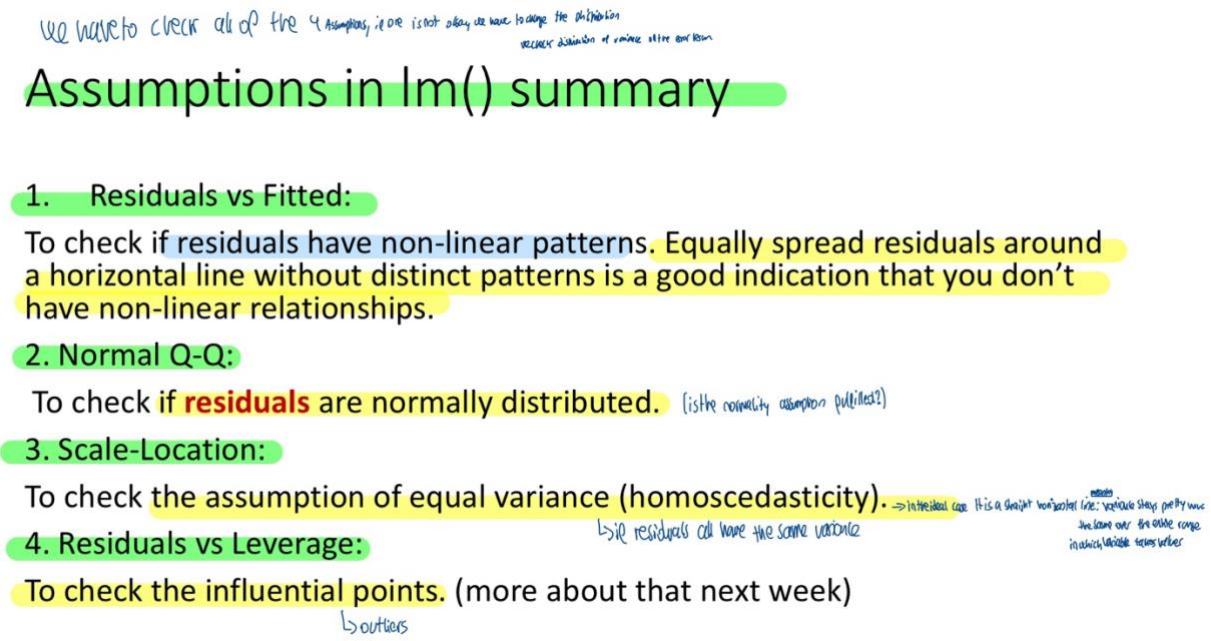
Calculation in R: `cor.test( x, y, conf.level=1-alpha)`

Correlation: find a numerical value expressing the relationship between variables.
Regression: estimate values of random variable on the basis of the values of fixed variable.
Indicates the impact of a unit change in the known variable on the estimated variable.

The **estimated regression line** $y = \widehat{\beta_0} + \widehat{\beta_1} x$

The **predicted values** $\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$

The **residuals** $y_i - \hat{y}_i = y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i)$



Lecture 10: Multiple Regression

Take from sta121 summary (Multiple Regression + ANOVA)

ANOVA checks whether different groups have the same mean.

One-Way ANOVA is used to compare different groups where the variability of the observations around the mean is the same in all groups. If the variability between the means is relatively large compared to the variance within the groups, then the result is much larger than 1. The samples most likely do not come from a common distribution -> in such case, H0 that the means are equal gets rejected. Assumption: samples are independent.

In a balanced design, the group sizes are the same. In ANOVA, if the variance within groups is smaller than the variance between groups, it means that the division in two separate groups is significant.

Multiple linear regression is used to predict the value of a dependent variable based on one or more independent variables.

Learning goal: Statistical model of multiple regression

$$\begin{aligned} Y_i &= \mu_i + \varepsilon_i, \\ &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \\ &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad\qquad i = 1, \ldots, n, \quad n > p, \end{aligned}$$

Information criterion balances the goodness of fit of the estimated model with its complexity, measured by the number of parameters. IN MLE, the larger the likelihood function, or, equivalently, the smaller the negative log-likelihood function is, the better the model is (but: maximizing likelihood can lead to overfitting).

Lower AIC scores are better, and AIC penalizes models that use more parameter, Lower BIC also better. BIC penalizes in a stronger way than AIC if more parameters are added to the model.

Learning goal: Explain the least squares criterion.

The least squares criterion is a method for determining the line of best fit that minimizes the sum of the squares of the differences between the fitted values (calculated by the model) and the actual values (i.e., the values in the data set). This criterion is based on the assumption that the errors are normally distributed and independent from each other. The residual is equal to the distance between the observed value and the predicted/fitted value (value estimated by the model). The best line is the one that minimized the sum of the squared residuals (by using the least squares criterion).

Linear Regression is sensitive to outliers.
Sample data is used to find the estimated regression line.

Y = beta0 + beta1*x + e -> equation of the regression line. Beta0 is the intercept of the line with the y-axis and beta1 the slope of the regression line.

4 assumptions (to be checked with diagnostic plots): residuals have linear patterns, residuals are normally distributed, they have equal variance (homoscedasticity), check influential points (outliers).

We don't have to test for the normality of the data, but we have to for the normality of the residuals.

Linear regression is sensitive to outliers. Estimate/std.error = t.value.

We calculate the sum of errors squared so the errors only have positive values and won't cancel each other out. The goal is to find a vector of estimates such that the sum of squared errors is minimized.

$Var(B\_Hat) = \sigma^2(X^TX)^{-1}$. σ^2 can be substituted by s^2 if sample variance is given.

The hat matrix, , is also known as projection matrix because it projects the vector of observations y onto the vector of prediction y hat.

The vector of the estimated coefficients can be calculated with the formula: $B\_Hat = (X^TX)^{-1}X^Ty$ -> least squares solution

Error terms have a different distribution that residuals because an error term represents the way observed data differs from the actual population, whereas a residual represents the way observed data differs from sample population data.

The variance of the vector of estimated Betas is unknown, so we need to estimate it with a t- distribution. We use the t-distribution and not the normal distribution since the variance

is unknown (we work with estimated variance) and it looks closer to the variance of a t-distribution rather than a normal distribution.

Sample variance estimation (1/n-1 etc.) vs Regression variance estimation (1/n-p-1, because we estimate p parameters and not just one):

When we estimate the variance for a sample, it makes sense only when we have at least 2 points. When we estimate the variance in linear regression, it makes sense only if we have more points than the parameters (p+1), otherwise, the sum of squared residuals would be 0.

Assumptions of linear models:

1) Linear relationship between the response and predictor variables (equal spreadness).

2) Normality: Error terms are normally distributed

3) Equal variance between the error terms (homoscedasticity) and zero mean

4) The errors are uncorrelated (no or little multicollinearity). To check it we need extra info:

    time sequence in which the data was collected (not checked in this course).

5) Independency of Observations:

    violated if there are:

        a) Repeated measures: several observations from the same individual (all other parameters

        kept fixed).
        Problems with repeated measures:

            1) Observations of the same (random) individual are more correlated than observations

            of different individuals.

            2) Parameters of each individual or level i might be slightly different.

        b) Hierarchical models: observations belong to subpopulations.

        c) Longitudinal settings: observations follow the same individual or subject over time. How

        do the observations change over time? Longitudinal data is a data that consists of repeated measurements of several individuals over time.

The observations in a dataset are not independent if the groups of observations are more correlated to each other than to the other observations.

Learning goal: Articulate assumptions for multiple linear regression, understand why we need to check the assumptions of the model, check the assumptions using diagnostic plots

Why is violating independency assumptions an issue? you run the risk that all of your results will be wrong.

In R 4 diagnostic plots are returned:

- Residual vs Fitted values: to check for Linear relationship between residuals.

- Normal Q-Q Plot: to check for normality.

- Scale Location Plot: to check for homoscedasticity.

- Residuals vs Leverage Plot: to check for outliers/influential points (high-leverage points).

We assume that the rank of the matrix X (with dimensions n x (p+1)) containing the rows x_i^T has rank p+1.

T-value = Estimate/Std. Error

The error term is iid and normally distributed with mean 0 and variance o^2_i

As part of assessing the adequacy of the linear model, we should check if the error distribution is adequate and check if there is any evidence against the iid assumption. To do it we can use the 4 diagnostic plots.

Multiple-R squared is the percentage of the response variable variation that is explained by a linear model. The r-squared increases with every predictor added to a model. Adjusted R-squared penalizes for adding more predictors.

The F-distribution is mainly used to compare two sample variances with each other.

The F value in regression is the result of a test where the null hypothesis is that all of the regression coefficients are equal to zero. The F-test compares your model with zero predictor variable (intercept only model) and decides whether your added coefficients improved the model. If you get a significant result, then whatever coefficients got included in the model, they improved the model's fit (it does not mean that all predictors are needed, but it means that the current model is better than the intercept only model).

Overfitting: the model has residuals with low variance, but it is more complex than it should be, it is overcomplicated, it fits too precisely (fits noise too). Such models have a high-variance and very low-bias tradeoff.
If we remove on point from the set the model would change a lot (See slide 21). Such

models don't generalize to new and unobserved cases. (e.g. interpolation that connects lines through all the points).

Multicollinearity: refers to predictors that are correlated with other predictors (can make the model worse, because predictors explain the same thing).

Variance-Inflation factor: measures how much the variance of an estimated regression coefficient is increased because of collinearity (how much it can be explained of one predictor using another predictor).

An influential point is an outlier that greatly affects the regression line.

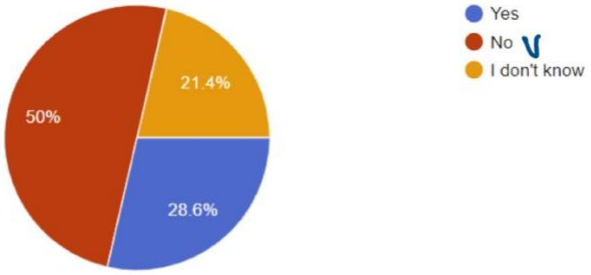
Learning goal: Be aware of nonlinear regression – examples
Many real-life examples do not have linear relationship in data but rather other types, for example an exponential growth e.g., population growth.

Questions on Lecture 10:

- Individually R squared cannot tell about variance importance -> True. R squared never decreases.
- Complicated models with a lot of parameters are better for prediction than simple models with just a few parameters -> False.
- Number of variables + degrees of freedom = number of observations in the data set -> True.

We consider the model y = β0 + β1x + e. Let [−0.01, 1.5] be the 95%-confidence interval for β1. In this case, a t-Test with significance level 1% rejects the null hypothesis H0 : β1 = 0. (Hint: would a t-Test with significance level 5% reject the null hypothesis? Is a 99% confidence interval wider or narrower than a 95% confidence interval)

28 responses



- 0 belongs to [-0.01,1.5], so we would not reject H_0 at 5% significance level. 99% CI is wider than 95% CI, hence we would also not reject H_0 at 1% significance level

Lecture 11: Analysis of Variance

Learning goal: Define an ANOVA model.

ANOVA is used to gain information about the relationship between the dependent and independent variables (i.e., to see if there is a difference between groups). If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1. ANOVA is a method for comparing models.

One way ANOVA is used to investigate if there are at least two mean values which show a significant difference. One way ANOVA has one independent variable, whereas two-way one has two independent variables.

Learning goal: Understand the concept of sums of squares decomposition.

Source	Sums of squares	Degrees of freedom	Mean squares	F-value
Model	$\mathrm{SS}_{\mathrm{model}} = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2$	$\mathrm{df}_{\mathrm{model}} = p$	$\mathrm{MS}_{\mathrm{model}} = \frac{\mathrm{SS}_{\mathrm{model}}}{\mathrm{df}_{\mathrm{model}}}$	$F_{\mathrm{obs}} = \frac{\mathrm{MS}_{\mathrm{model}}}{\mathrm{MS}_{\mathrm{error}}}$
Error	$\mathrm{SS}_{\mathrm{error}} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$	$\mathrm{df}_{\mathrm{error}} = n - p - 1$	$\mathrm{MS}_{\mathrm{error}} = \frac{\mathrm{SS}_{\mathrm{error}}}{\mathrm{df}_{\mathrm{error}}}$	
Total	$\mathrm{SS}_{\mathrm{total}} = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$\mathrm{df}_{\mathrm{total}} = n - 1$		

To use ANOVA to compare models, you would first fit each model to the data, and then calculate the following statistics:

Total Sum of Squares (TSS): The total variation in the dependent variable.

Model Sum of Squares (SSM): The variation in the dependent variable that can be explained by the model.

Error Sum of Squares (SSE): The variation in the dependent variable that cannot be explained by the model.

Degrees of freedom (df): The number of observations in the data set minus the number of parameters in the model.

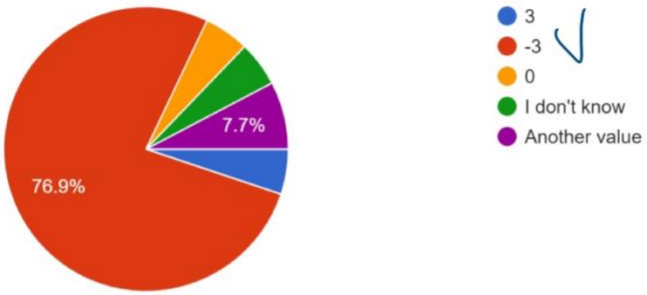
Then, you can use these statistics to calculate the following: Mean Square Error (MSE) = SSE/df
Mean Square Model (MSM) = SSM/df
F-Statistic = MSM/MSE

It's important to note that ANOVA assumes that the error terms across the different models are normally and independently distributed with equal variances (homoscedasticity). If these assumptions are not met, other statistical tests like likelihood ratio test or AIC should be considered.

Questions on Lecture 11:

We set contrasts to sum to zero. We are considering one Way ANOVA, three groups. We performed lm(). The intercept = 2, beta1 =5, beta2 = -2. beta3 =

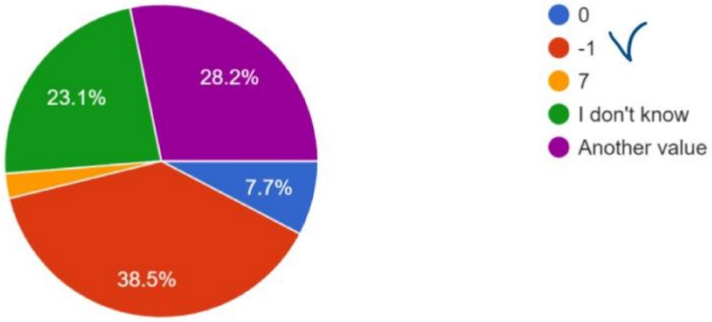
39 responses



$$beta1 + beta2 + beta3 = 0, \text{hence } beta3 = 0 - 5 + 2 = -3$$

We set contrasts to sum to zero. We are considering one Way ANOVA, three groups. We performed lm(). The intercept = 2, beta1 =5, beta2 = -2. The mean of the third group =

39 responses



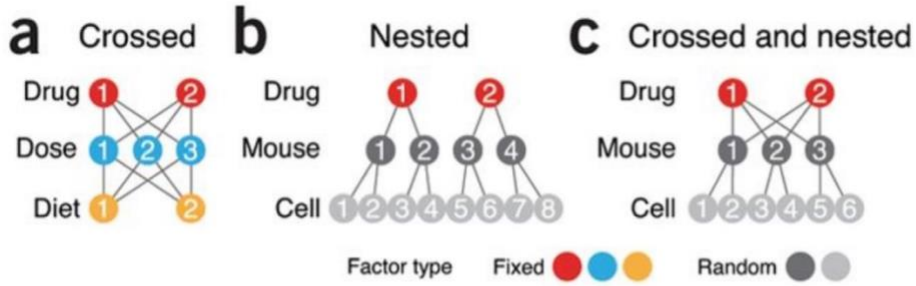
$$beta1 + beta2 + beta3 = 0, \text{hence } beta3 = 0 - 5 + 2 = -3$$
$$\text{The mean of the third group } intercept + beta3 = 2 - 3 = -1$$

Lecture 12: Design of Experiments

Learning goal: Understand the issues and principles of Design of Experiments.

Aim: maximize primary variance, minimize error variance and control for secondary variance -> maximize the signal we are investigating, minimize the noise we are not modeling and control for uncertainties with carefully chosen independent variables.

Exploratory experiments: look for ideas -> valuable for discovering new insights, generating hypotheses.
Confirmatory experiments: used to validate hypotheses.
Learning goal: Describe different setups of an experiment.
Blocking: helps dealing with heterogeneity in the data that masks the effect we would like to study.



(a) A crossed design examines every combination of levels for each fixed factor
(b) Nested design can progressively sub-replicate a fixed factor with nested levels of a random factor that are unique to the level within which they are nested.
(c) If a random factor can be reused for different levels of the treatment, it can be crossed with the treatment and modeled as a block.

Randomization: assign a treatment to an experiment unit by pure chance. Randomization allows the use of probability theory and statistical analysis. It minimizes the differences among groups by equally distributing people with particular characteristics among all the trial arms.
A confounder is a variable that influences both the dependent variable and the independent variable. For example, smoking causes lung cancer and yellow finger. It is not possible to conclude that yellow fingers cause lung cancer. Smoking here is the confounder. Not taking confounders into account will lead to bias in the results. In the case of discrete confounders, it is possible to split your sample into subgroups according to these pre-defined factors.
Example: randomized complete block design (RCBD): each block receives the same number of subjects (i.e., there is the same number of subjects for each combination of factors).
Selection bias: occurs when groups differ systematically next to the effect that is analyzed (e.g., smoke vs dementia). Confounding occurs due to another factor that distorts the relationship between treatment and outcome (e.g., coffee drinkers smoke more and have more lungs cancer). Performance bias occurs when a care giver or analyzertreats the subjects of the two groups differently.
Learning goal: Compute sample size for an experiment.

$$n \approx 4z_{\alpha/2}^2 \frac{\hat{p}(1-\hat{p})}{\omega^2},$$

Of course, $\hat{p}$ is not known a priori and we often take the conservative choice of $\hat{p} = 1/2$ as the function $x(1-x)$ is maximized over $(0,1)$ at $x = 1/2$.

Thus, we may choose $n \approx \left(z_{\alpha/2}/\omega\right)^2$

Lecture 13: Bayesian Approach

Approaches in statistics:

- **frequentist** -> confidence interval

The parameter Θ is fixed, inference based on the **data only**

- **Bayesian** -> credible interval

We consider the parameter Θ as a random variable, inference based on the **data** and prior **knowledge**

Example:

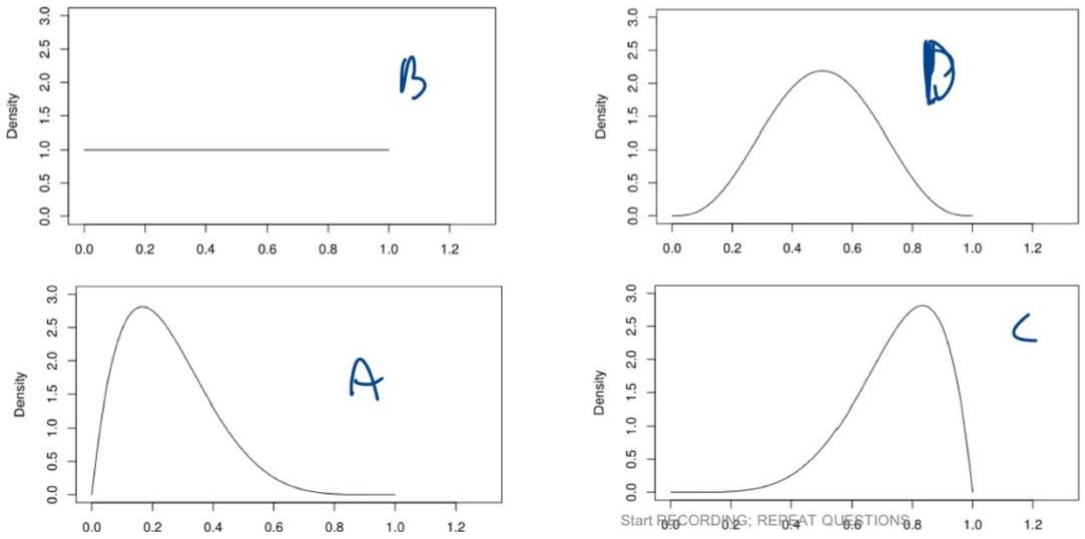
Washing the dishes based on tossing a coin. So far – 3 times heads.

Heads – you do the dishes, tails – person A, B, C or D does the dishes.

A – an altruistic dishwashing lover, B - a complete stranger,

C – a dishwashing hater with a history of cheating in a casino,

D – the most fair person you have ever met.



Beta-distribution: defined between 0 and 1 -> used as prior knowledge. Binomial models the number of successes, beta models the probability of successes.

Learning goal: Explain how to compute posterior probability.
Posterior Distribution:

$$P(\boldsymbol{\Theta}|data) = \frac{P(data|\boldsymbol{\Theta}) \times P(\boldsymbol{\Theta})}{P(data)}$$

These are just different names compared with: $\mathrm{P}(A \mid B) = \frac{\mathrm{P}(B \mid A)\,\mathrm{P}(A)}{\mathrm{P}(B)}, \quad \text{for } \mathrm{P}(B) \neq 0,$

We consider the parameter θ as a random variable.
$P(\Theta)$ is the **prior distribution**. It represents our beliefs about the true value of the parameters (just like we had distributions representing our belief about the p chosen by A, B, C or D).

$P(\Theta \mid data)$ is called as the **posterior distribution**. This is the distribution representing our belief about the parameter values after we have calculated everything taking the observed data into account.

$\mathrm{P}(data|\Theta)$ is the **likelihood**.

$\mathrm{P}(data)$ is the marginal distribution of the data. We are only interested in the parameter values. But $\mathrm{P}(data)$ doesn't have any reference to them. In fact, $P(data)$ is just a number. Computing this number can be hard. $P(data)$ is a normalising constant. 21

Posterior density is proportional to Likelihood x prior density.
Learning goal: Interpret posterior probability and a posterior credible interval.

Definition 13.1. The interval R with

$$\int_R f(\theta \mid y_1, \ldots, y_n)\, \mathrm{d}\theta = 1 - \alpha \tag{13.20}$$

is called a $(1-\alpha)\%$ credible interval for θ with respect to the posterior density $f(\theta \mid y_1, \ldots, y_n)$ and $1-\alpha$ is the credible level of the interval. ♢

Credible interval is an interval within which an unobserved parameter value falls with a particular probability

Learning goal: Explain the idea of the Bayes factor and link it to model selection.
Bayes Factors: A bayes factor is the ratio of the likelihood of one particular hypothesis to the likelihood of another hypothesis.

The Bayesian counterpart to hypothesis testing is done through a comparison of posterior probabilities. For example, consider two models specified by two hypotheses H_0 and H_1.

$$\underbrace{\frac{\mathrm{P}(H_0 \mid y_1, \ldots, y_n)}{\mathrm{P}(H_1 \mid y_1, \ldots, y_n)}}_{\text{Posterior odds}} = \underbrace{\frac{\mathrm{P}(y_1, \ldots, y_n \mid H_0)}{\mathrm{P}(y_1, \ldots, y_n \mid H_1)}}_{\text{Bayes factor } (\mathrm{BF}_{01})} \times \underbrace{\frac{\mathrm{P}(H_0)}{\mathrm{P}(H_1)}}_{\text{Prior odds}},$$

The Bayes factor BF01 summarizes the evidence of the data for the hypothesis H_0 versus the hypothesis H_1.

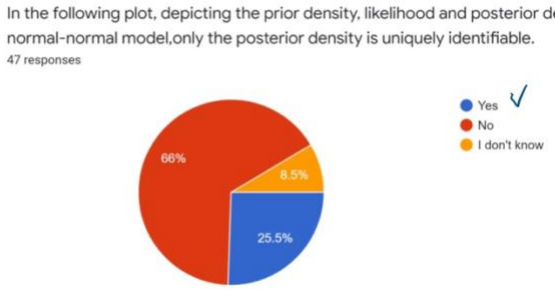
0.01 bayes factor: equal to 1/0.01 = 100 more evidence for H_0

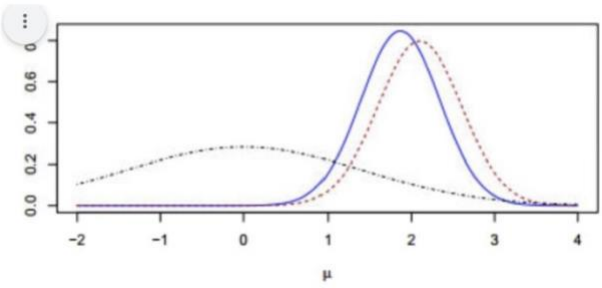
Factor scale

$$1 < \begin{matrix}\text{barely worth} \\ \text{mentioning}\end{matrix} < 3 < \text{substantial} < 10 < \text{strong} < 30 < \begin{matrix}\text{very} \\ \text{strong}\end{matrix} < 100 < \text{decisive}$$

Start RECORDING; REPEAT QUESTIONS 39

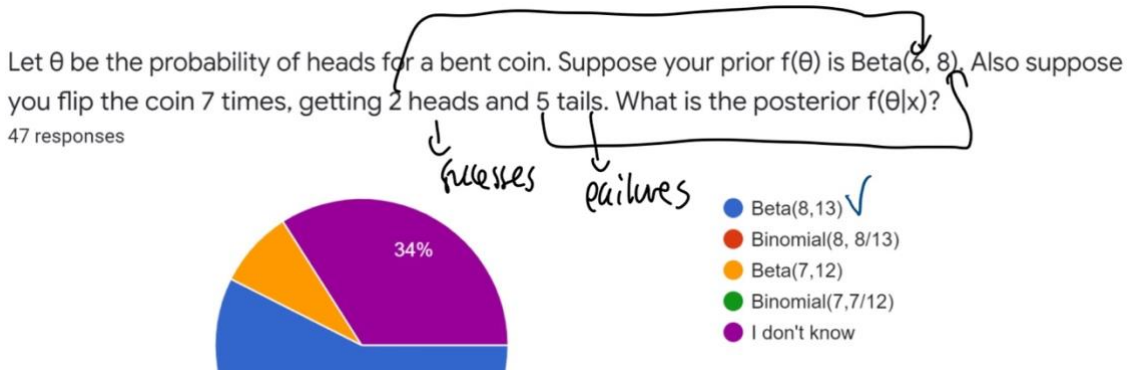
Choose a prior distribution, that the posterior distribution belongs to the same family.
Questions on Lecture 13:





We cannot say which of the dashed lines represents likelihood, which one prior. So only the posterior is uniquely identifiable.

4



In Bayesian statistics we would always get the same posterior distribution of the parameter of interest, no matter which prior is used -> False, depending on the prior, the posterior may change.

Learning goal: Describe the fundamental differences between the Bayesian and frequentist approach.

Frequentist	Bayesian
• The parameter to be estimated is fixed • We infer it based on data only	• The parameter to be estimated is considered a random variable, we have **prior knowledge** on the distribution it's coming from • We infer it based on the data and the prior knowledge

Lecture 14: Monte Carlo Methods

Motivation: posterior in the Bayesian methods is not always a known and implemented distribution. Monte Carlo simulation is used to numerically solve a complex problem through repeated random sampling. Monte Carlo estimates the probability based on rando, numbers. With enough simulated random numbers, the estimate is very good, but it is still inherently random.

Monte Carlo Integration: instead of looking at the analytical solution which can be complicated, sampling is performed.

- Let X be a random variable and $f_X(x)$ be its density function and $g(x)$ an arbitrary (sufficiently "well behaved") function.

- We are looking for the expected value of $g(x)$, i.e.,

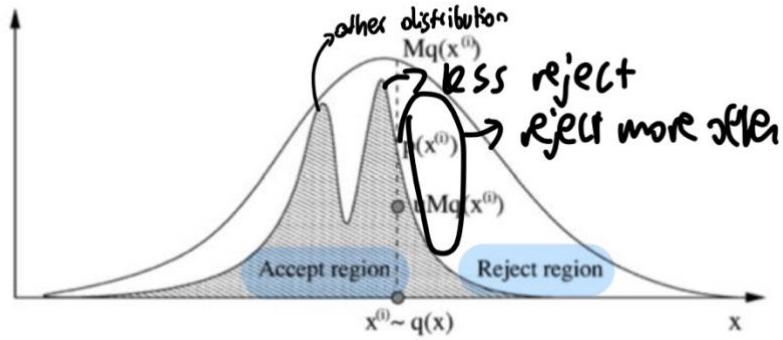
$$E\big(g(x)\big) = \int_{\mathbb{R}} g(x) f_X(x)\,\mathrm{d}x$$

An approximation of this integral is (along the idea of method of moments):

$$E\big(g(x)\big) = \int_{\mathbb{R}} g(x) f_X(x)\,\mathrm{d}x \approx E\big(g(x)\big) = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

Where $x_1, \ldots, x_n$ is a random sample of $f_X(x)$. The method relies on the law of large numbers

Rejection sampling: Has the aim to sample from a distribution with given density when no direct method exists. The values from a known density are drawn and through rejection of unsuitable values, the observations of the other distribution are generated -> we sample from a complex distribution by sampling with another distribution.



Algorithm:

$$f(y) = c \cdot f^*(y)$$

Step 0: Find an $m < \infty$, so that $f^*(y) \le m f_Z(y)$ -> density is smaller than some other density

Step 1: draw a realization $\tilde{y}$ from $f_Z(y)$ and a realization u from a standard uniform distribution.

Step 2: if $u \le f^*(\tilde{y})/m\, f_Z(\tilde{y})$ then $\tilde{y}$ is accepted as a simulated value from $f_Y(y)$, otherwise $\tilde{y}$ is discarded and no longer considered.

We cycle along Steps 1 and 2 until a sufficiently large sample has been obtained

Learning goal: Describe qualitatively Gibbs sampling.
Gibbs sampling: used when we know how to easily sample from conditional distribution but not the joint distribution (because it is difficult to sample directly from it). Doing it stepwise allows the sample to be most similar to the one of the joint distributions. If independent samples are required, one may take samples that are far apart from each other in sequence. The initial samples in the sequence may not accurately represent the joint distribution (initial burn-in period) -> solution: drop first few samples. The samples may not converge, there is no guarantee that it will always converge.
When we can limit the distribution with some function, use rejection sampling. For the posterior in Bayesian setting, use JAGS (just another gibbs sampling) based on the prior and the distribution of the data.

Learning goal: Explain the idea of a Bayesian hierarchical model.
A hierarchical Bayesian model is a Bayesian model in which the prior distribution of some of the parameters depends on further parameters to which we also assign priors.

$$Y_{ij} \mid \boldsymbol{\beta}_i, \kappa \overset{\text{indep}}{\sim} \mathcal{N}(\boldsymbol{x}_{ij}\boldsymbol{\beta}_i, 1/\kappa), \qquad i = 1, \ldots, n, j = 1, \ldots, n_i, \tag{14.15}$$

$$\boldsymbol{\beta}_i, \mid \boldsymbol{\eta}, \lambda \overset{\text{iid}}{\sim} \mathcal{N}_p(\boldsymbol{\eta}, \mathbf{I}/\lambda), \tag{14.16}$$

$$\boldsymbol{\eta} \sim \mathcal{N}_p(\boldsymbol{\eta}_0, \mathbf{I}/\tau), \quad \lambda \sim \mathcal{G}am(\alpha_\lambda, \beta_\lambda), \quad \kappa \sim \mathcal{G}am(\alpha_\kappa, \beta_\kappa). \tag{14.17}$$

where τ, α_λ, β_λ, α_κ and β_κ are hyperparameters. The three levels (14.15) to (14.17) are often referred to as observation level, state or process level and prior level.

Questions on Lecture 14:

- In rejection sampling it is always best to use the uniform distribution as the known distribution -> False, the t-distribution is defined between -∞ and ∞
- It is possible to use a uniform distribution as known distribution if we wanted to sample from t-distribution using rejection sampling -> False, it would be tricky to get values outside the uniform distribution.

We would like to sample from a distribution. Its density has a shape of a triangle - see on the plot. We would like to use rejection sampling, as a known distribution we use the uniform distribution [-1,1] with m =2. We sample y1 = -0.9. In the rejection step, we sampled u =0.9 (the notation as in slide 22). Would you use y1 as a sample from the "triangle distribution"? Copy

42 responses