# Linear mixed models in R
## Day 1

JONAS WALTHER

# Outline

Day 1

Statistics background

What is an LME?

Day 2

Assumptions of LMEs

Creating your own models

Day 3

Building the correct model

Day 4

Contrasts and understanding your results

Day 5

Presenting and reporting results

# Exercises

Slides and exercises at:

https://github.com/WaltherJonas/LME2025

1 hour in person each day

Rest as homework

# Summer School on Statistical Methods for Linguistics and Psychology

- Prof. Dr. Shravan Vasishth (organizer)
- Prof. Dr. Douglas Bates (speaker)

# Textbook

https://vasishth.github.io/Freq_CogSci/

# R Package

lme4 – implementation of linear mixed models

# Dependent variables

Used in experiment, surveys, recordings – measurement of interest

Different types of variables:

- Discrete: number of correct responses, head vs. Tail, pregnancy outcome
- Continous: reaction times, amplitude of neuronal response, heart rate

These dependent variables are random variables

# Random variables

Variable whose values are the outcome of a random process

- A function that maps the outcome (w) of a random process on a numeric value (x)
- Flipping a coin, reaction time in a task

The exact outcome cannot be predicted – random phenomenon

- But we can estimate its range

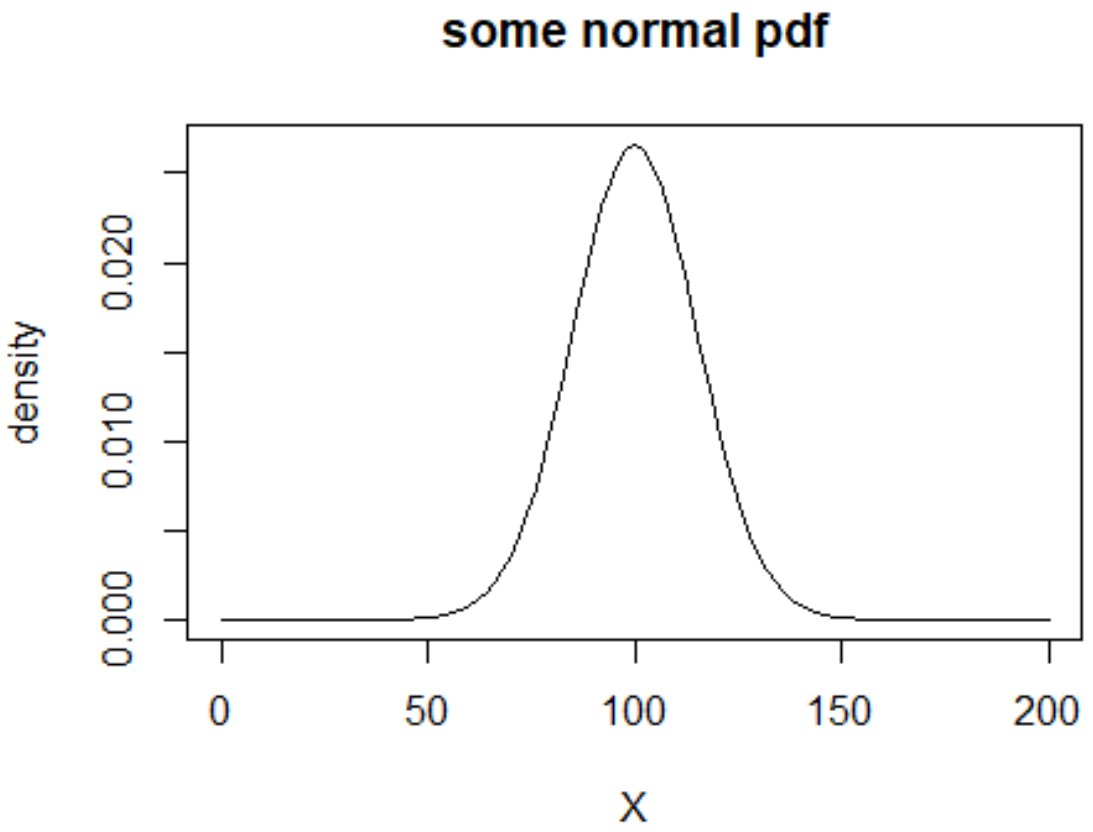
# Probability distribution

Gives probability for each values

Area under the curve is always 1 (all possible values have a probability of 100%)

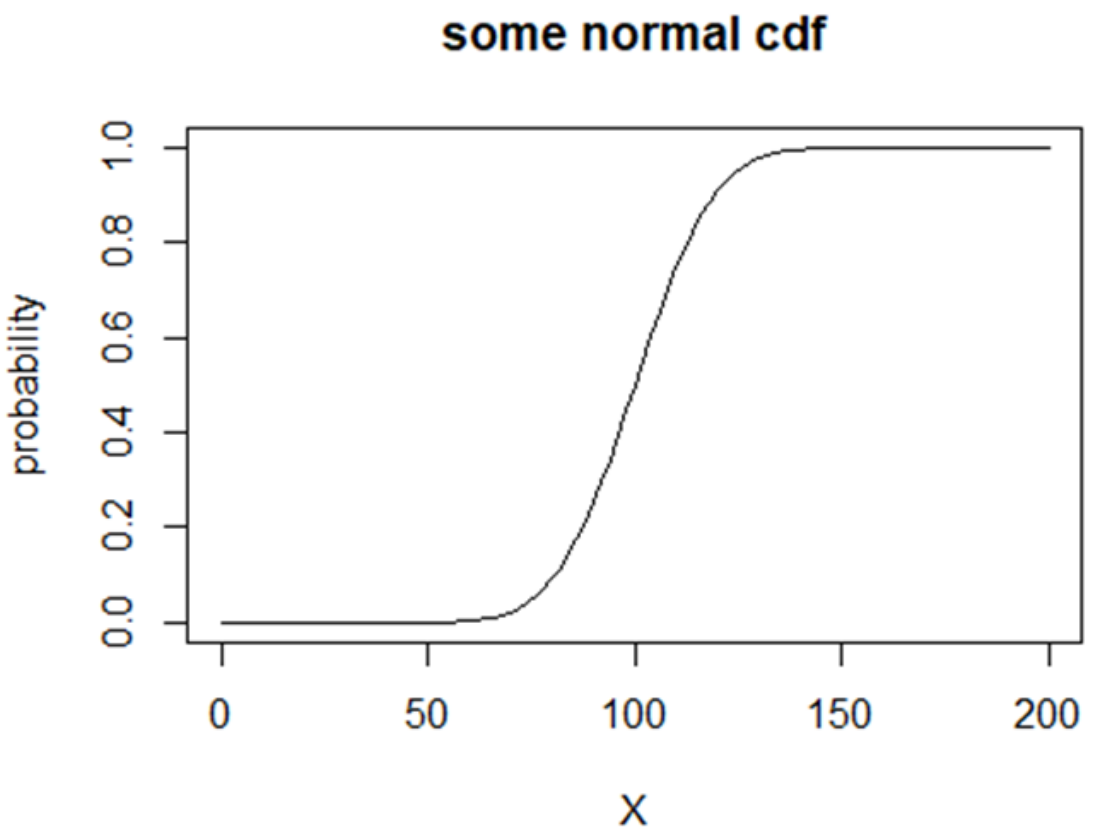
Area under the curve between two values is equal to the probability of getting a value between those two numbers

Probability density function can be described with an equation



# Cumulative distribution function

Gives the probability of a random variable X having a value less than or equal to x, for every value of x



# Normal distribution

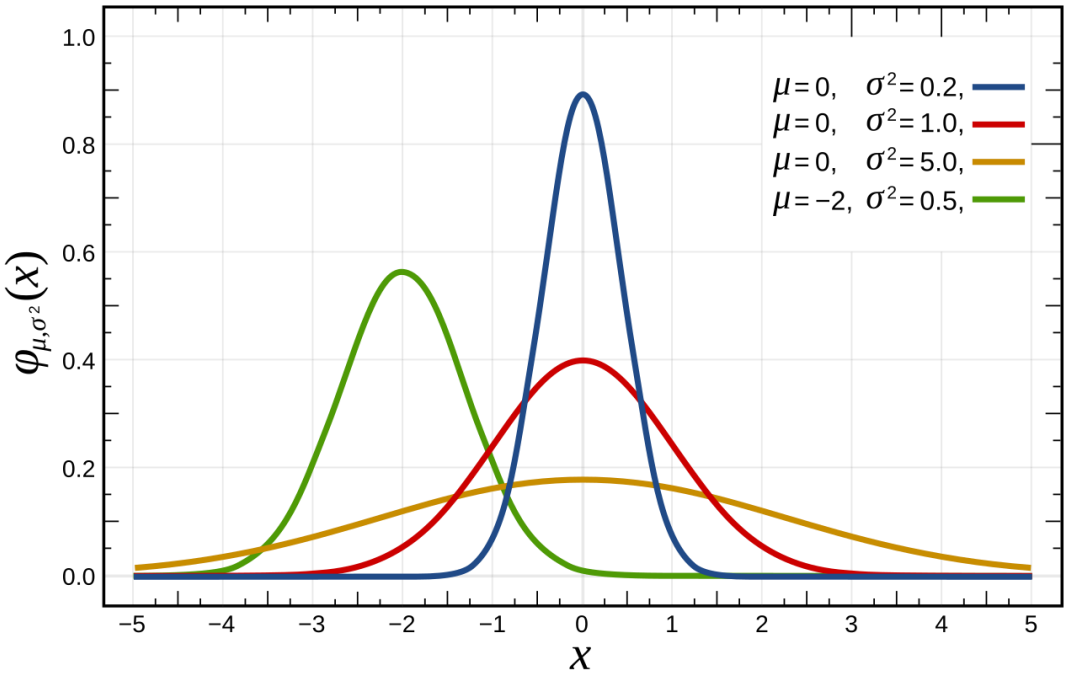
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1(x-\mu)^2}{2\ \sigma^2}}, \quad -\infty < x < \infty$$

Defined by

μ - mean or expectation

$\sigma^2$ – variance (square of standart deviation)

We write $X \sim \mathrm{norm}(\mathrm{mean} = \mu, \mathrm{sd} = \sigma)$.




https://en.wikipedia.org/wiki/Normal_distribution

# Mean, variance and standard deviation

Assume you run an experiment and end up with 25 reaction times.

```
x <- c(350,455, 546, 464, 444, 746, 426, 526, 736, 845, 846, 635,
745, 732, 555, 655, 675, 567, 433, 346, 243, 263, 475, 734, 325)
```

This is how you get the mean: Sum all values and divide by the number of values. You can do that in R

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

```
sum(x)/ 25
```

```
## [1] 550.68
```

```
#or use
mean(x)
```

```
## [1] 550.68
```

# Mean, variance and standard deviation

Variance (how much do the different data points differ from the mean?): for each value, substract the mean and square the difference. Divide by the number of values (-1 if this is a sample)

$$v = \frac{\sum(x_i - \mu)^2}{n - 1}$$

```
sum((x-mean(x))^2)/(length(x)-1)
```

```
## [1] 31609.73
```

```
#or use
var(x)
```

```
## [1] 31609.73
```

# Mean, variance and standard deviation

And this is how you get the standard deviation: square root of the variance

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n - 1}}$$

```
sqrt(sum((x-mean(x))^2)/(length(x)-1))
```

```
## [1] 177.7912
```

```
#or use
sd(x)
```

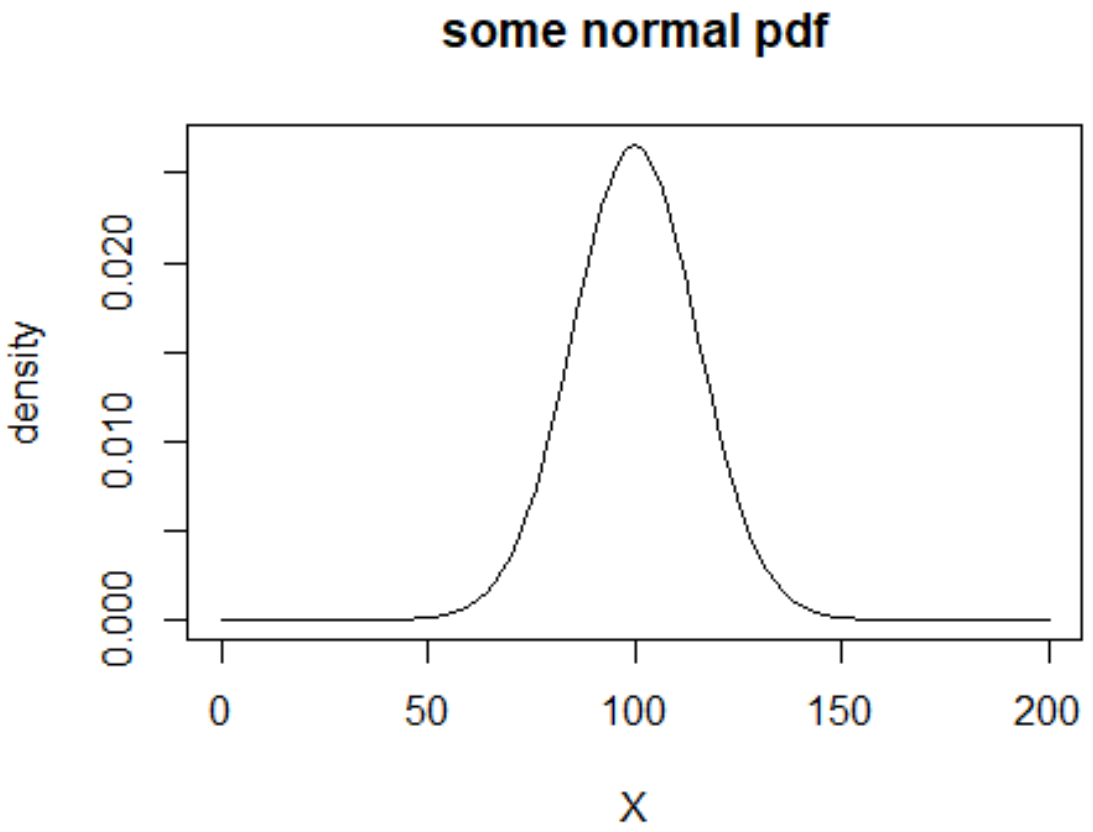
```
## [1] 177.7912
```

# R built-in functions

Surface under the curve = probability

How do we get this surface? What we need to do is compute the integral from e.g., -Inf to a given value or between two values.

Alternatively, we can use R built-in functions



# dnorm() – probability density function

The R function for the pdf of the normal distribution is dnorm(x, mean, sd).

With this function we can get the value of the probability density function for the normal distribution given parameters for $\mu$ , and $\sigma$ for a given value of this distribution (x). It returns the height of the pdf for a given value (or quantile)

x must be specified (this is the value for which you want the density).

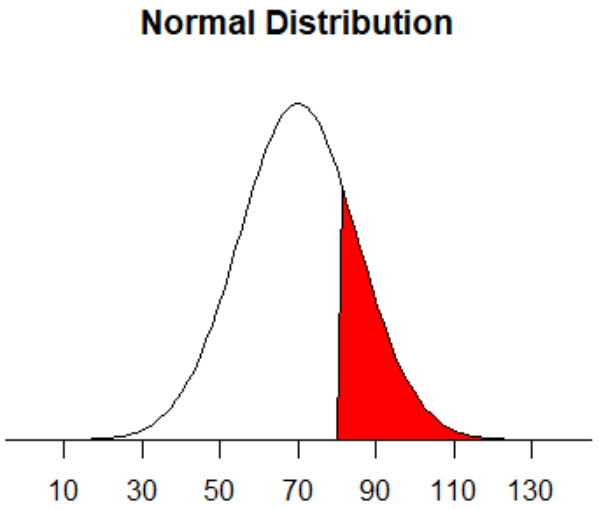
mean and sd can be specified. If not, the values of the standardized normal distribution are used (mean =0, sd =1)

Examples:

The following command returns the density for the value 80, on a normal distribution with a mean of 70 and a standard deviation of 20

```
round(dnorm(80,mean=70,sd=20),3)
```

```
## [1] 0.018
```




# pnorm() for cumulative density function

The function pnorm returns the value of the probability density function for the normal distribution given parameters for x, $\mu$ , and $\sigma$ .

It informs on the probability of a value of x or smaller (to do so, it computes the integral from $-\infty$ to x of the pdf of the normal distribution)

$\mu$ and $\sigma$ can be specified. If not, the values of the standardized normal distribution are used ($\mu = 0, \sigma = 1$)

What is the probability of x being equal to 80 or higher on a normal distribution with a mean of 70 and a standard deviation of 20?

```
round(1-pnorm(80,mean=70,sd=20),3)
```

```
## [1] 0.309
```

# Central limit theorem

The distribution of a sample will approximate a normal distribution as the sample size becomes larger, regardless of the population's actual distribution shape.

- If we repeat an experiment 1000 times, the observed effect means will be normally distributed.
- From SINGLE sample with a finite number of data points we can estimate the underlying sampling distribution.

# Frequentist statistics in research

Probability can be seen as frequency

Any experiment can be infinitely repeated, with each run drawing statistically independent results

**Point estimation:** calculating a single value as the best estimate of an unknown population parameter

**Confidence intervals:** range of values within which the parameter is likely to fall, offering a measure of uncertainty

**Hypothesis testing:** evaluates a claim about a population parameter by comparing observed data against a null hypothesis

# Frequentist statistics in research

Null hypothesis statistical significance testing:

1. The researcher sets a hypothesis: e.g., Answering in your native language is faster than in your second language
2. The researcher sets the corresponding null hypothesis: e.g., naming latencies in the first and second language are identical
3. The researcher collects data on a sample of the population s/he is interested in: e.g., picture naming in Polish and English for 50 bilingual speakers
4. The researcher uses a statistical test to determine how likely it is to get the value s/he obtained if the null hypothesis is true (or, under H0)
5. If it is likely, s/he does not reject H0, if unlikely, s/he rejects H0 (and claims s/he has some evidence in favour of H1)

# Example experiment

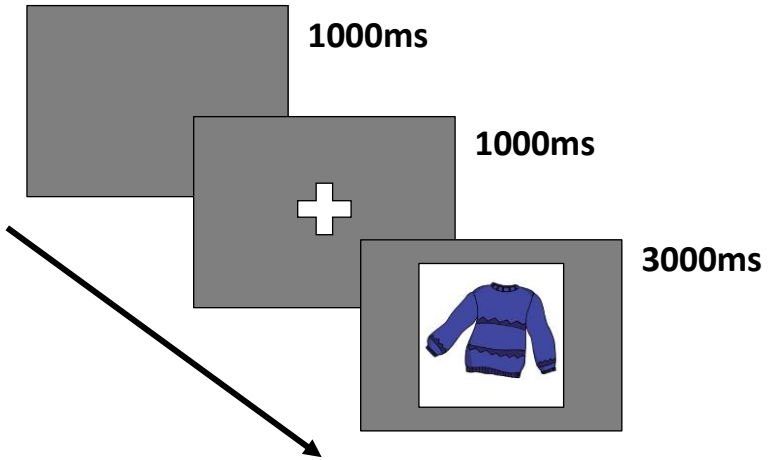
Bilingual speaker name pictures in a Polish and an English Context

Every speaker saw every picture in both contexts

Experimental variable: Context - UK vs PL

Grouping factors: Subject and Item

Dependent variable: naming latency



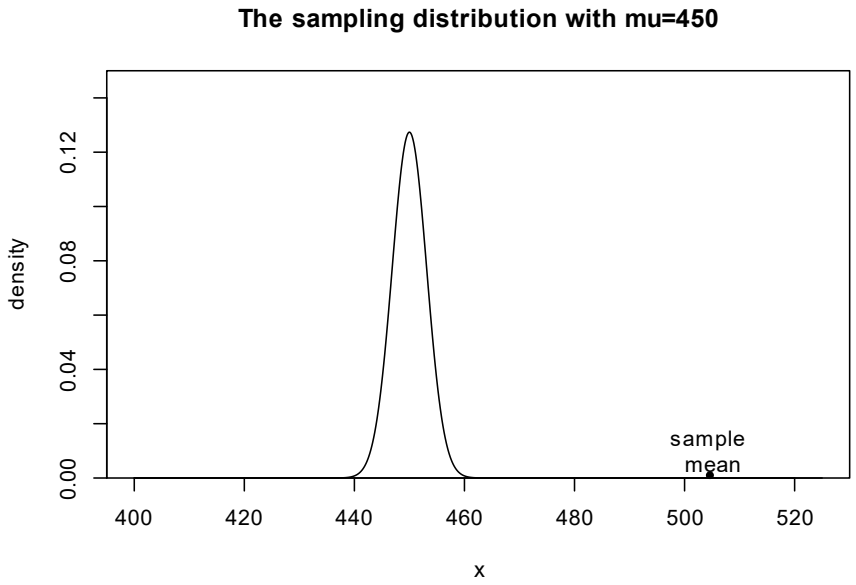
# One-sample t-test

Testing experimental hypotheses:

- Null hypothesis $H_0$: observed sample mean $\overline{y}$ is „near“ the hypothesised μ
- Distance between sample mean and hypothesised mean can be written as:

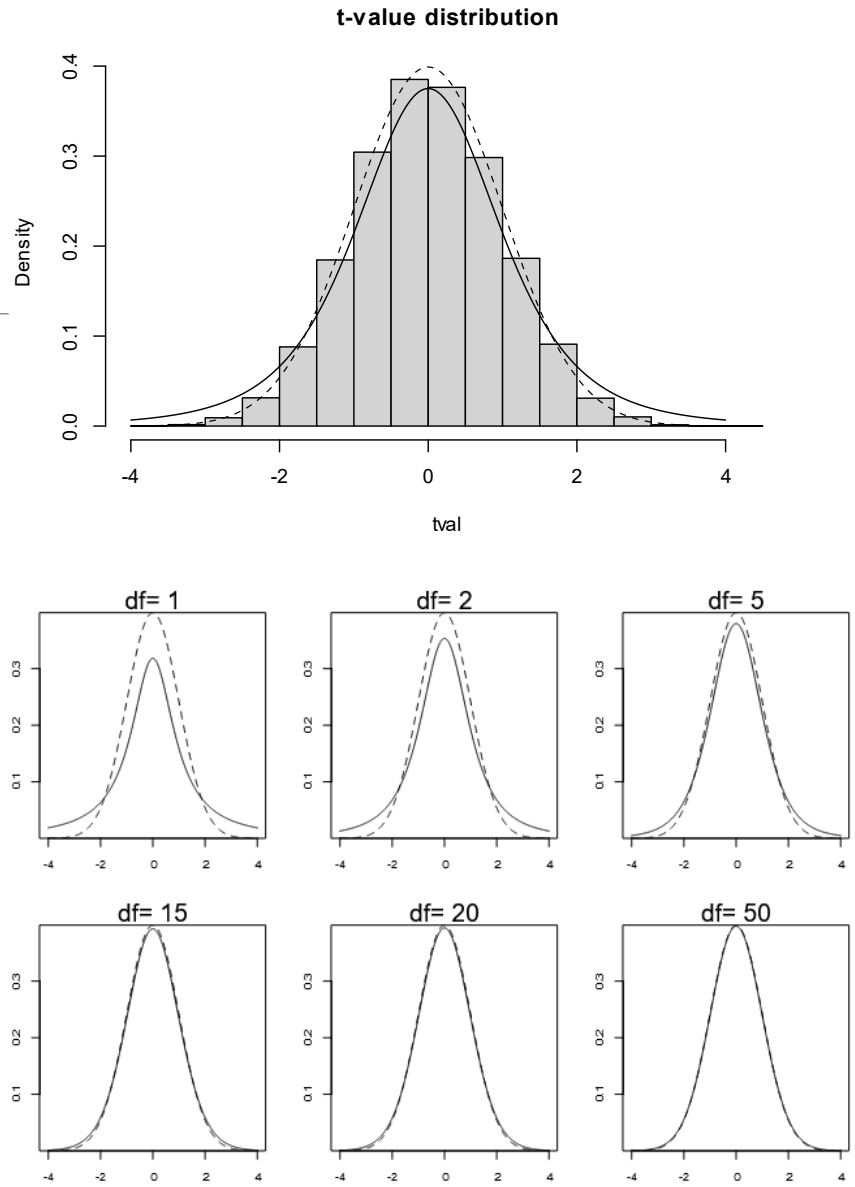
$$t \times SE = \overline{x} - \mu$$

$$T = \frac{\overline{X} - \mu}{SE}$$



# One-sample t-test

- T is a random variable derived from the random variable $\overline{x}$
- T-distribution approaches normal distribution, but has heavier tails
  (t – straight line, normal – broken line)
- T-distribution depends on degrees of freedom
  - Higher df approach normal distribution

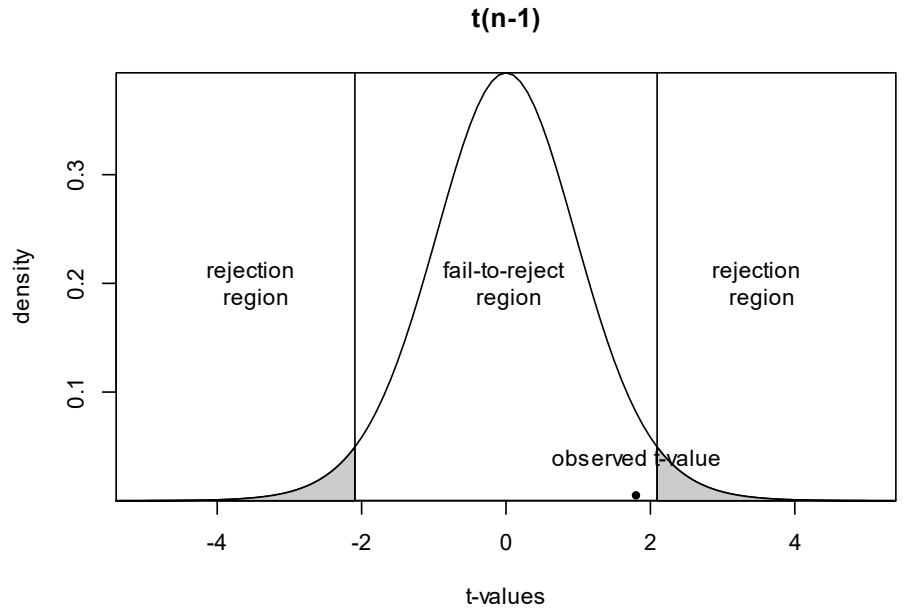


# One-sample t-test

$$t = \frac{\overline{y} - \mu}{s/\sqrt{n}}$$

- $\overline{y}$ - estimate
- s – standart deviation
- n – sample size

- „Small t-values“ confirm $H_0$



# How do we decide whether a value is likely or unlikely

- The cut-off value is called alpha.
- Deciding which value alpha should take is a matter of convention.
- The convention in science is often to set alpha at 0.05

What does it mean?

It means that when there is 5% chances of getting the results we obtained under the null hypothesis we consider that this is low probability, low enough to reject the null hypothesis.

N.B. it also means that if we do the same experiment 100 times under the null Hypothesis, we will wrongly reject the null in 5% of the cases.

**But:**

- Several authors suggest that alpha should be decreased or discarded for other tools (Bayesian, etc)

# Paired t-test: RT in Polish and English

```
head(paired_PN_Data)
```

```
## # A tibble: 6 × 3
## # Groups:   Subject [6]
##   Subject    PL    UK
##   <chr>   <dbl> <dbl>
## 1 AF1310  1064. 1097.
## 2 AG0712   974.  934.
## 3 AG0911   709.  913.
## 4 AJ1312   956.  949.
## 5 AJ1611   448. 1070.
## 6 AJ3007   767.  728.
```

$$H_0: \mu_{UK} - \mu_{PL} = 0$$

$$T = \frac{\mu_{UK} - \mu_{PL} - 0}{SE}$$

```
t.test(paired_PN_Data$PL, paired_PN_Data$UK, paired = TRUE)
```

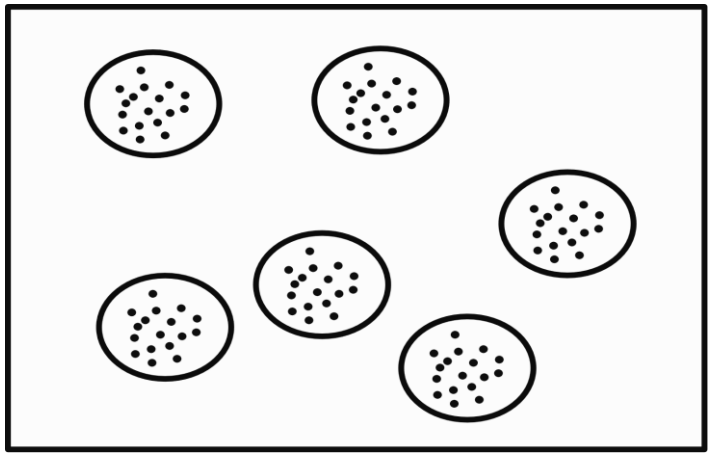
```
## t = -2.4871, df = 36, p-value = 0.01765
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -67.783400  -6.890745
## sample estimates:
## mean difference
##       -37.33707
```

# T-test assumes independence of samples

Experimental sample are often not independent (multiple trials per subject, different recording sites etc.)

You need to average over your data

- Usually Subject, Items, Recording sites etc

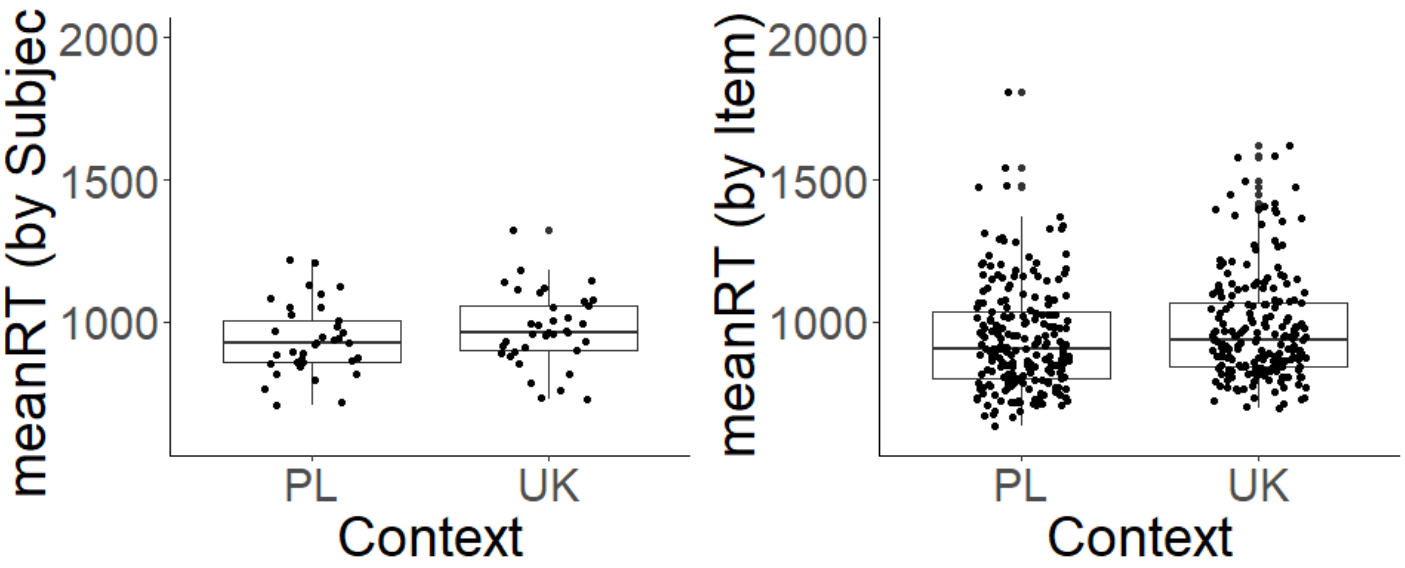


https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/

# T-test assumes independence of samples

```
head(Raw_Data)
```

```
## # A tibble: 6 × 5
##   Subject    RT Context Trial ItemNr
##   <chr>   <dbl> <chr>   <dbl> <chr>
## 1 AS3008   1049 UK          4 127
## 2 AW1912   1007 UK          4 127
## 3 JM2904    794 UK          4 127
## 4 LM1102    826 UK          4 127
## 5 MB0509    842 UK          4 127
## 6 MB2601   1131 UK          4 127
```



# Subject and Item analysis

Both t-tests show significant differences in Context

But you lose data

Interpretation can be difficult

```
t.test(Context_Item_PNData$PL, Context_Item_PNData$UK, paired = TRUE)

##  Paired t-test
## data:  Context_Item_PNData$PL and Context_Item_PNData$UK
## t = -3.6664, df = 207, p-value = 0.0003127
## 95 percent confidence interval: -57.64880 -17.33075
## sample estimates:
## mean difference -37.48977

t.test(Context_Subject_PNData$PL, Context_Subject_PNData$UK, paired = TRUE)
##
##  Paired t-test
## data:  Context_Subject_PNData$PL and Context_Subject_PNData$UK
## t = -2.4871, df = 36, p-value = 0.01765
## 95 percent confidence interval:
##  -67.783400  -6.890745
## sample estimates:
## mean difference
##       -37.33707
## t = -2.4871, df = 36, p-value = 0.01765
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval: -67.783400  -6.890745
## sample estimates:
## mean difference -37.33707
```

# Adding subjects

$$y = \beta_0 + \beta_1 x + \epsilon$$

$$predicted\ naming\ latency = average\ naming\ latency + Language\ effect + noise$$

We can add individual subject information

$$y_{iUK} = \beta_0 + \beta_{Cond} x_{iCond} + \epsilon_{Cond} + u_{0i}$$

$\beta_0$ -average naming latency

$u_{0i}$ -subject dependant adjustment

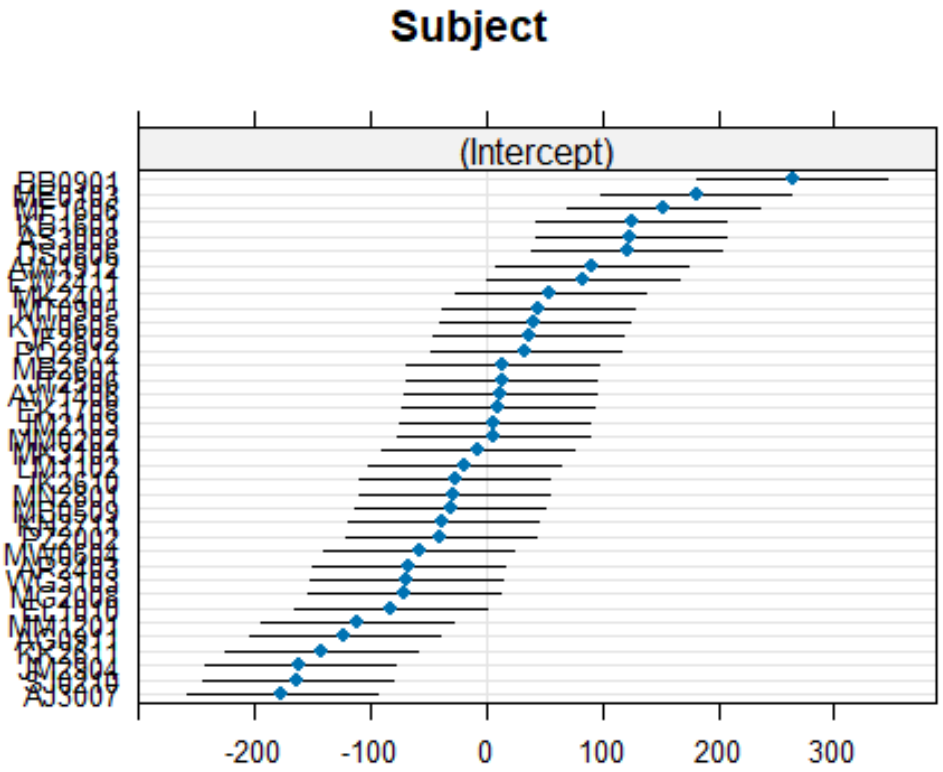
$\epsilon_{\mathrm{Cond}}$ -error/noise term

# Subject intercepts

```
library(lme4)

sub.lmer <-
lmer(RT~Context+(1|Subject),Raw_
Data2)

print(dotplot(ranef(sub.lmer,con
dVar=TRUE)))
```



# Linear mixed model as solution

$$H_0: \mu_{UK} - \mu_{PL} = 0$$
$$H_1: \mu_{UK} - \mu_{PL} = \delta$$
$$T = \frac{\mu_{UK} - \mu_{PL}}{SE}$$
$$y_{iCond} = \beta_0 + u_{0i} + \epsilon_{Cond} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \delta$$
$$y_{iUK} - y_{iPL} = (\beta_0 + u_{0i} + \epsilon_{UK} + \delta) - (\beta_0 + u_{0i} + \epsilon_{Cond})$$
$$y_{iUK} - y_{iPL} = \delta + (\epsilon_{UK} - \epsilon_{Cond})$$
$$T = \frac{(y_{iUK} - y_{iPL})}{SE}$$

$\beta_0$ -average RT

$u_{0i}$ -subject dependant adjustment

$\epsilon_{UK}$ -error/noise term

$\delta$ – Context effect

# Syntax of a crossed mixed-effects model

RT ~ Context + (1 + | Subject)

Dependent variable

Independent variables/fixed effects

Random effects

# What are mixed effects?

**Fixed effects**

- Usually, the effects that are actually of interest

**Random effects**

- Usually, the effects within which measures are being repeated
- Based on the experimental design/data structure
  - In Psychology often just Subjects and Items

Mixed-effect models combine both fixed and random effects into the same model

# Implemenation in R

```
library(lme4)
```

```
head(Raw_Data2)
```

```
## # A tibble: 6 × 3
## # Groups:   Subject [6]
##   Subject Context    RT
##   <chr>   <chr>   <dbl>
## 1 AG0911  PL       709.
## 2 AJ3007  PL       767.
## 3 AP2403  PL       863.
## 4 AS3008  PL      1100.
## 5 AW1406  PL      1006.
## 6 AW1912  PL      1049.
```

```
Context_Model <-
lmer(RT~Context+(1|Subject),Raw_Data2)
```

# Implemenation in R

```
Context_Model <- lmer(RT~Context+(1|Subject),Raw_Data2)


summary(Context_Model)


## Linear mixed model fit by REML ['lmerMod']
## Formula: RT ~ Context + (1 | Subject)
##    Data: Raw_Data2
##
## REML criterion at convergence: 879.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.81576 -0.35532  0.07066  0.35808  2.01444
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 11715    108.24
##  Residual              4169     64.57
## Number of obs: 74, groups:  Subject, 37
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   936.64      20.72  45.205
## ContextUK      37.34      15.01   2.487
##
## Correlation of Fixed Effects:
##           (Intr)
## ContextUK -0.362
```

# LME and paired t-test

```
t.test(Context_Subject_PNData$PL, Context_Subject_PNData$UK,
paired = TRUE)

##  Paired t-test

## t = -2.4871, df = 36, p-value = 0.01765

## 95 percent confidence interval:

##  -67.783400  -6.890745

## sample estimates: mean difference

##       -37.33707
```

```
Context_Model <- lmer(RT~Context+(1|Subject),Raw_Data2)

## Fixed effects:

##             Estimate Std. Error t value

## (Intercept)   936.64      20.72  45.205

## ContextUK      37.34      15.01   2.487

## Correlation of Fixed Effects:

##            (Intr)

## ContextUK -0.362
```

# When to use LMEs

Continuous dependent variable

Clustered data (e.g. participants, recording locations, etc)

- No interest in the differences between clusters

Multiple crossed or nested clusters

- Subject vs Items; students in schools

Flexibility for more complicated designs

# Advantages of LME over ANOVA

Analysis of unbalanced datasets (unbalanced designs or missing values)

Modelling of correlations between observations

Allows unequal variance between groups

But:

ANOVA provides analysis of variance tables

ANOVA is computationally more efficient

# Doing science in the real world

Experimental data is often complex and messy, especially biological or medical data

Different grouping factors like populations, species, sites etc.

Potentially low sample sizes, especially for models with many parameters

Samples are often not truly independent (same/different experimenters, recording devices etc)

Linear mixed models allow the inclusion of messy data, even with low samples sizes, data with complex structures and many covariates

# Exercises

Slides and exercises at:

https://github.com/WaltherJonas/LME2025

1 hour in person each day

Rest as homework

# Mark-down files

**Do not forget to write the HW number and your name in the header!**

**Please use R Markdown for HW submission if possible**

Please work on the exercises in the Mark-down files and answer within. For submission send the finished pdf file to my mail address before the next lecture:

Jonas.Walther@uni.tuebingen.de

# Mark-down files

When answering a HW question, don't copy the question out

Show your computation, so I know that you know how to do it (don't just write the answer!).

```
prob<-pnorm(6,mean=7,sd=1)-pnorm(4,mean=7,sd=1)
```

Sometimes, an R command will print out tons of code, and not all the code needs to be shown always. You can suppress code output using the following settings in R chunks:

warning=FALSE

message=FALSE

echo=FALSE

include=FALSE

eval=FALSE

# Thank you for your attention!