9:00 – 17:00

10:30 – 10:45 Coffee break
12:30 – 13:00 Lunch break
14:30 – 14:45 Coffee break

INTRODUCTION TO STATISTICAL DATA ANALYSIS USING R

ANJA EGGERT

https://github.com/AnjaEggert/StatisticsInR_Intro

Universität Rostock Traditio et Innovatio

Outline

- **Technical R**
  - R, Rstudio, Add-on packages, Rproject, Git
- **Some statistical terms**
  - Sample size, replication, randomization, effect size, degrees of freedom
- **Hypothesis testing**
  - Coin tossing, rejection region, H_0, H_A, Type I error (alpha), Type II error (beta)
- **1F-ANOVA and multiple comparisons**
  - F-test, Bonferroni, Tukey
- **Design types**
  - Blocking, fixed and random effects
  - Completely Randomized Design, Randomized Complete Block Design

R, RSTUDIO, ADDON-PACKAGES

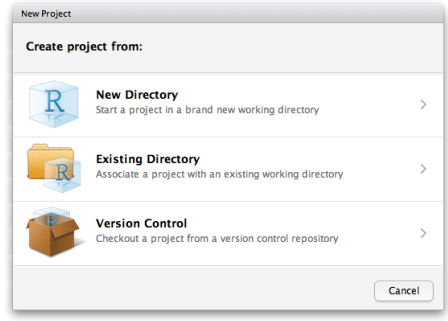
R example

https://www.datacamp.com/community/tutorials/r-packages-guide

```
# all about R packages
```

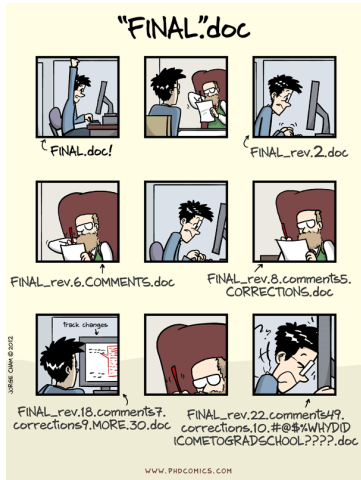
"A package is a like a book, a library is like a library; you use library() to check a package out of the library"— Hadley Wickham

R example




```
# Working with R projects
```


What is version control?

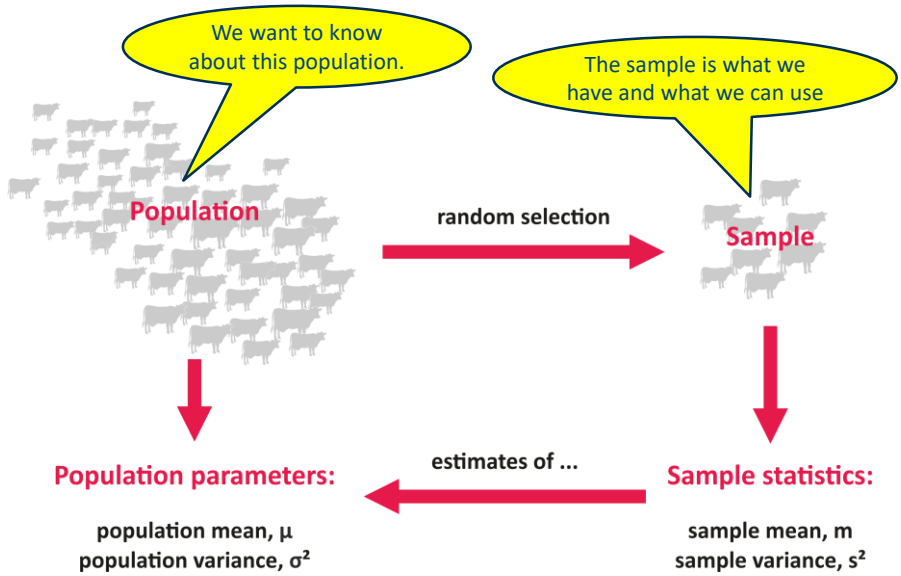


(from phdcomics.com)

https://resources.github.com/whitepapers/github-and-rstudio/

```
# Working with Github
```

SOME STATISTICAL TERMS

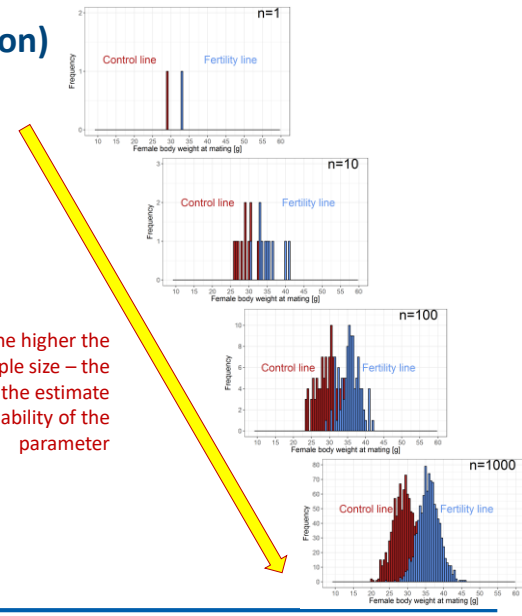


Sample size (= replication)

- How to assess the degree of variability in a response variable

```
mu <- 29 (or 36)
sigma <- 3

rnorm(n,mean=mu,sd
=sigma)
```

The higher the sample size – the better the estimate of variability of the parameter



Randomization

- **Randomization = randomly assigning treatments to the experimental units and to distribute the units randomly in a room.**
  - We could draw cards, throw dice, …
  - Or we use software and random number generator

Example:
- 3 treatments (Control, LIPO, CON)
- 18 animals (6 per treatment)
- Completely randomize in 18 cages in a room

CON	Control	CON	Control	LIPO	CON
LIPO	CON	Control	CON	Control	Control
CON	LIPO	LIPO	LIPO	CON	CON
Control	Control	LIPO	Control	CON	LIPO
CON	Control	CON	LIPO	Control	LIPO
LIPO	LIPO	Control	Control	CON	LIPO

```
library(agricolae)
design <-design.crd(trt, repeat, serie=2, seed=42, method="Super-Duper")
library(desplot)
desplot(data=book, form=trt~col+row, text=trt, shorten="no", main="", show.key=F)
```

Using Effect Size—or Why the P Value Is Not Enough ?

- **Are the obtained differences biologically relevant?**

$$Effect\ size = \frac{Mean\ of\ group\ 1 - Mean\ of\ group\ 2}{standard\ deviation}$$

- *p* value → does an effect exists
- effect size → the size of the effect

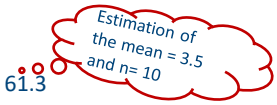
- Both are essential results to be reported.

- An effect size is exactly equivalent to a 'Z-score' of a standard normal distribution.
- Effect size = 0.8 → the score of the average person in group 1 is 0.8 standard deviations above the average person in group 2, i.e. it exceeds the scores of 79% of group 2.

Degrees of freedom

- **Imagine you're into funny hats:** Unfortunately, you have constraints. You have only 7 hats. Yet you want to wear a different hat every day of the week.
- You had 7-1 = 6 days of "hat" freedom, but on Day 7 you must wear the one remaining hat.

- **Imagine you're into data analysis:** Degrees of freedom are the number of "observations" (pieces of information) in the data that are free to vary when estimating statistical parameters.



34, -8.3, -37, -92, -1, 0, 1, -22, 99 -----> 10th value must be 61.3

0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 ----> 10th value must be 30.5

HYPOTHESIS TESTING

An example: coin tossing

- Let's first review the mechanics of single hypothesis testing. For example, suppose we are flipping a coin to see if it is fair. We flip the coin 100 times and each time record whether it came up heads or tails. So, we have a record that could look something like this:

```
# H T H T H H H T T H T H H T T T H T H …
```

```
# do it in R
coinFlips
 H  T
59 41

?dbinom
```

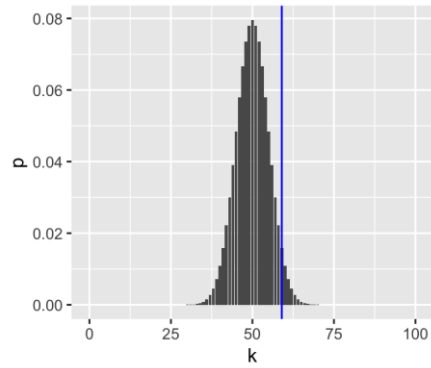
- Binomial distribution:

$$P(K = k|\mathrm{n}, \mathrm{p}) = \binom{n}{k} p^k (1-p)^{n-k}$$

The probability that the observed value for K is k, given the values of n and p.

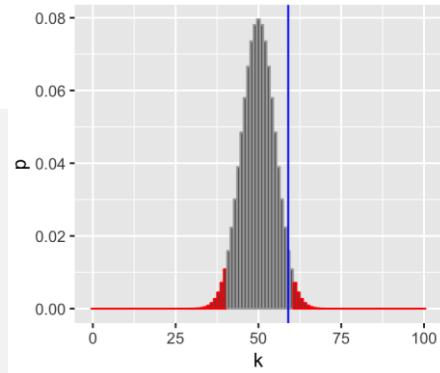
The most likely number of heads is 50

- How do we quantify whether the observed value, 59, is among those values that we are likely to see from a fair coin, or whether its deviation from the expected value is already large enough for us to conclude with enough confidence that the coin is biased?



Colour the rejection region

```
arrange () from the dplyr package

    sort the p-values from lowest to
    highest

%>% pass the result to mutate

    it adds another dataframe column

logical vector reject: TRUE/FALSE
```



The observed value, 59, lies in the grey shaded area, so we would not reject the H_0 of a fair coin from these data at a significance level of α=0.05 !!

The five steps of hypothesis testing

1. Decide on the effect that you are interested in, design a suitable experiment or study, pick a data summary function and test statistic.
2. Set up a null hypothesis, which is a simple, computationally tractable model of reality that lets you compute the null distribution, i.e., the possible outcomes of the test statistic and their probabilities under the assumption that the null hypothesis is true.
3. Decide on the rejection region, i.e., a subset of possible outcomes whose total probability is small.
4. Do the experiment and collect the data; compute the test statistic.
5. Make a decision: reject the null hypothesis if the test statistic is in the rejection region.

Hypothesis testing

- Hypothesis tests use **sample data** to make inferences about the properties of a population - it is usually impossible to measure the entire population.
- However, there are tradeoffs when you use samples. The samples we use are typically a **miniscule percentage of the entire population**. Consequently, they occasionally **misrepresent the population** severely enough to cause hypothesis tests to make errors!
- For a generic hypothesis test, the two hypotheses are as follows:
  - Null hypothesis H_0: There is **no effect**
  - Alternative hypothesis H_a: There is an **effect**
- The sample data must provide sufficient evidence to reject the null hypothesis and conclude that the effect exists in the population.

Two types of error in hypothesis testing

- Ideally, a hypothesis test fails to reject the null hypothesis when the effect is not present in the population, and it rejects the null hypothesis when the effect exists.
- BUT! There are two types of errors in hypothesis testing, Type I and Type II errors. Both types of error relate to incorrect conclusions about the null hypothesis.

If the fire alarm rings when there is no fire.

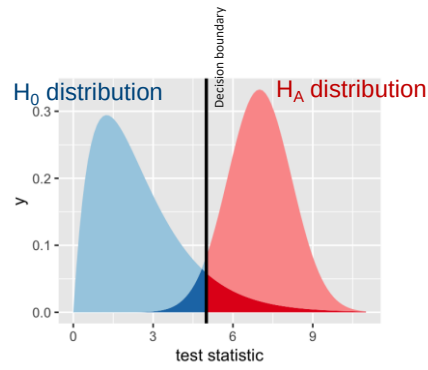
	Test rejects H_0	Test fails to reject H_0
H_0 is true, **no effect** ($\mu_1 = \mu_2$)	**Type I error: false positive**	**Correct decision: effect exists**
H_0 is false, **effect exists** ($\mu_1 \neq \mu_2$)	**Correct decision: effect exists**	**Type II error: false negative**

The fire alarm fails to ring when there is a fire.

There is a tradeoff between Type I and Type II errors !

- It's always possible to reduce one of the two error types at the cost of increasing the other one. The real challenge is to find an acceptable trade-off between both of them.

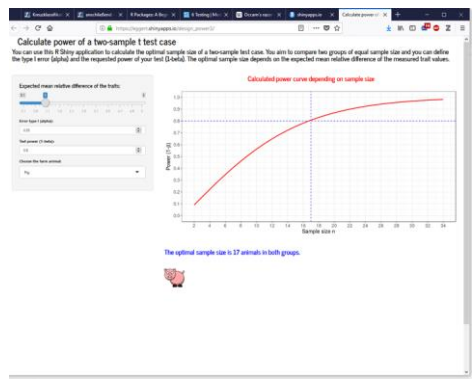
- The decision boundary is the black line
- The hypothesis is rejected if the statistic falls to the right.
- The probability of a Type I error is then simply the dark red area.
- The probability of a Type II error is the dark blue area.



There is a tradeoff between Type I and Type II errors !

- If you hold everything else constant, as you decrease alpha from 5% to 1%, you increase the opportunity for a Type II error.
- Many **possible reasons for Type II errors:**
  - small effect sizes
  - small sample sizes
  - high data variability
- Unlike Type I errors, you can't set the Type II error rate for your analysis. Instead, the best that you can do is estimate it before you begin your study by approximating properties of the alternative hypothesis that you're studying. When you do this type of estimation, it's called **power analysis**.

R Shiny Power



R example

r_script_power.R

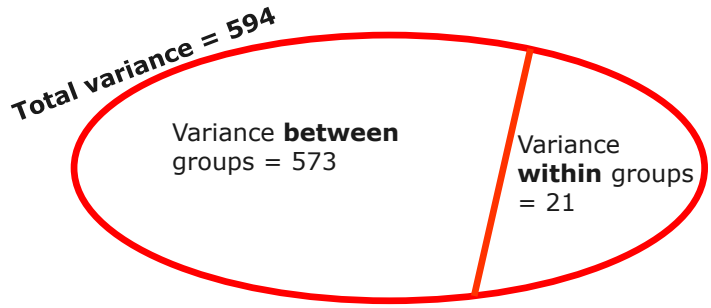
 # apply() functions

```
?apply
# Construct a 5x6 matrix
x <- matrix(rnorm(30), nrow=5, ncol=6)

# Sum the values of each column with `apply()`
apply(x, 2, sum)

sapply(1:3, function(x) x^2)

lapply(1:3, function(x) x^2)
```

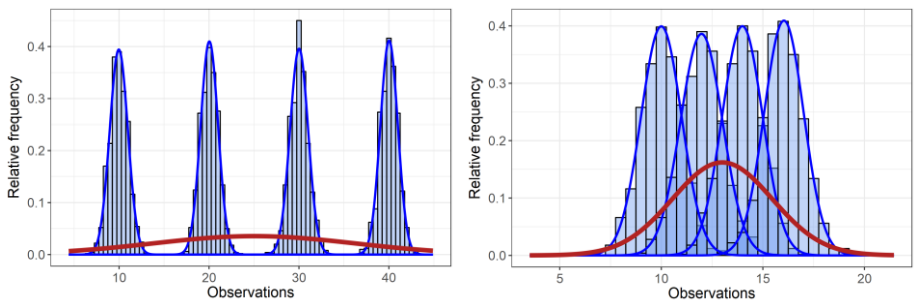

ANOVA & MULTIPLE COMPARISONS

Splitting up the variance

- In the ANOVA, the observed variance in a particular variable is partitioned into components attributable to different sources of variation.
- Total variance = Variance between groups + Variance within groups



Variance between groups



Introduction to analysis of variance (ANOVA)

- **Research question:**
  - Do sea urchins influence shoot density of seagrass?

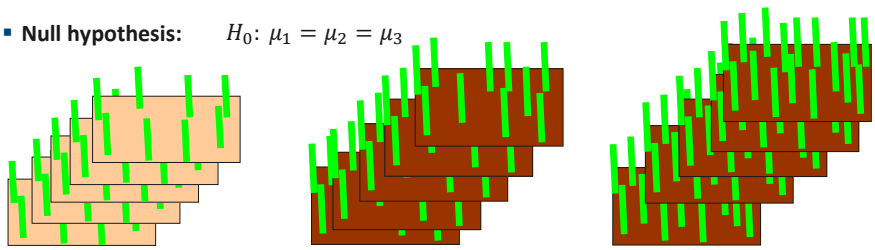
- **Biological background:**
  - Sea urchins feed on seagrass, but could also enhance growth due to excretion of nutrients.

- **Experimental design:**
  - Manipulation of density of sea urchins, i.e. 1 fixed factor with 3 levels
  - Assumption: sea urchins are independent (maybe not? brothers or sisters?)

Principles of ANOVA

- **Question:**
  - Is there a significant difference between the sample mean of the three groups?

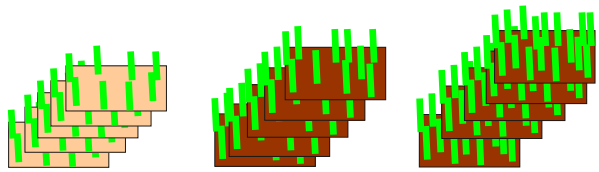
- **Null hypothesis:** $H_0: \mu_1 = \mu_2 = \mu_3$



1: „no" **2: „medium"** **3: „high"**

Experimentel fixed factor: density of sea urchins

Sea urchins: results

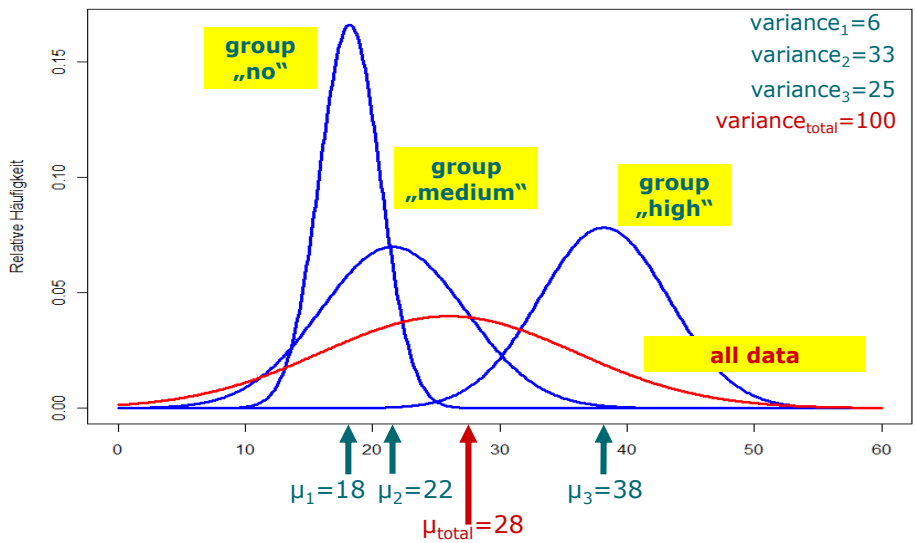


Group	N	Data
No	5	15 ; 17 ; 18 ; 20 ; 21
medium	5	13 ; 20 ; 22 ; 25 ; 28
high	5	31 ; 37 ; 38 ; 40 ; 45

`data: urchins.xlsx`

`r_script_anova.R`

Within group and total variance

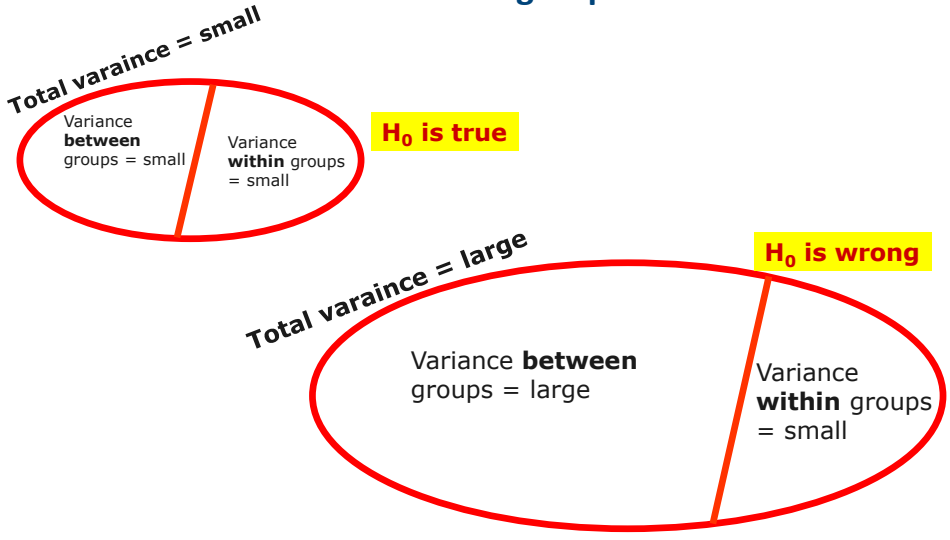


Variance within groups

Group	N	df	Mean	Variance	Mean variance
no	5	4	18,2	6	
medium	5	4	21,6	33	**21**
high	5	4	38,2	25	

(6+33+25) / 3 = 21

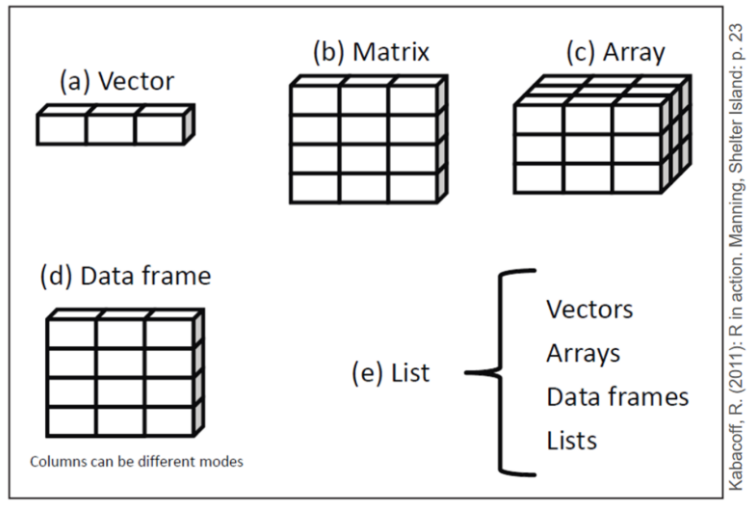
Variance between and within groups



R example

```
r_script_probability_distributions.R

# vector
# matrix
# data frame
# list
```


Data types: vector – matrix - data frame - list

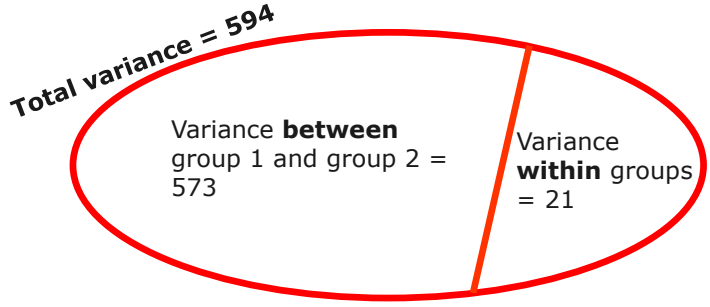


R example

```
r_script_ctd_pre_posec.R

# list
```


Variance between and within groups

Can we reject H_0?



F-test

Due to treatment effect

$$F\ statistics = \frac{Variance\ between\ groups}{Variance\ within\ groups}$$

Residual variance

- The larger F, the higher the probability to reject H0
- F > Fcritical, reject H_0

```
# Draw F distribution
curve(df(x, df1=2, df2=12), from=0, to=5)

# Get the critical F value from R
qf(.95, df1=2, df2=12)
[1] 3.885294
```


ANOVA: sea urchins

```
Analysis of Variance Table

Response: value
          Df Sum Sq Mean Sq F value    Pr(>F)
key        2 1145.2  572.60  26.967 3.634e-05 ***
Residuals 12  254.8   21.23
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple comparisons: sea urchins