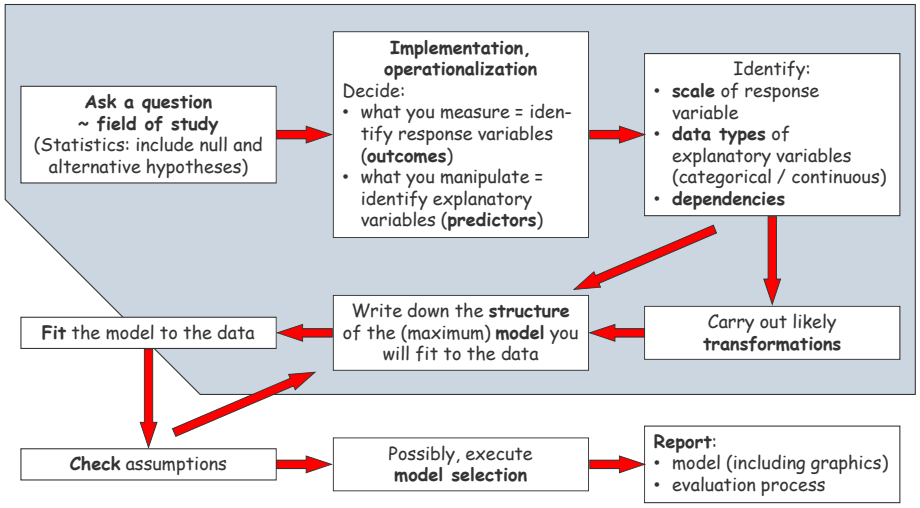
DRS Spring School – week 2

Hypothesis testing and statistical inference

PD Dr. Lorenz Gygax (HU Berlin)

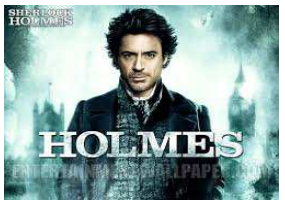
Statistics starts with planning an experiment even long before it is conducted



Two types of statistical procedures

- Descriptive stats
  - Describe
  - Summarise
  - Represent
- Inferential stats
  - Generalise from sample to pop
    - Parameter estimation
  - Draw conclusions
    - Hypothesis testing

vs.



www.zeewallpaper.com

- "Inference is the process of drawing a conclusion by applying clues (statistics) to observations or hypotheses" (Wikipedia)
- The conclusion drawn is also an inference

Core research questions for which statistics is needed

- Testing
  - How strongly is a relationship between an explanatory (predictor) and response (outcome) variable supported statistically?
  - A differences between (treatment) groups?
- Estimation
  - How strong is the relationship, how large are the differences?
- Prediction
  - What response can I expect for a given set of predictors? "It is hard to predict, specifically the future"
- Model selection:
  - What proportion of observed variation can I explain in my response variable? With what explanatory variables?
  - Do some explanatory variables explain more variation than others?

Hypthesis testing & probability

- Population → truth/reality

- Sample → uncertainty
  - Conclusions we draw from sample should be accompanied by a probability

equivocaltruth.tumblr.com

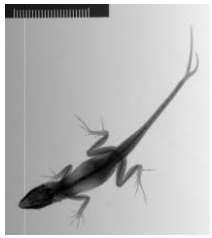
How *scientist*
(non-statisticians)
have thought of
statistical testing
(and many still do)

Hypotheses

- **Scientific method**:
  set-up hypotheses and falsify all but one
- A **null hypothesis** (H_0) is a statement of "no effect"
  Explanatory variable does not have an effect on response variable
- One or several **alternative hypotheses** (H_a) are contrasted with H_0
- Both/all must be specified at the onset before we collect data

Alternative hypotheses

- (More or) less specific
- 2-sided
  A difference exists but direction is unknown
- 1-sided
  Difference can only be in one direction or are expected in one direction only



VS

http://www.anoleannals.org/

- A one-sided hypothesis can be highly meaningful
- Has only occasionally an effect on how to test

Null hypothesis

- Cp. the scientific method
- The null hypothesis is presumed true until (statistical) evidence indicates otherwise
- How probable is the data assuming the null hypothesis is true?
  - Very improbable → reject H_0
  - Not improbable → accept H_0

The p-value

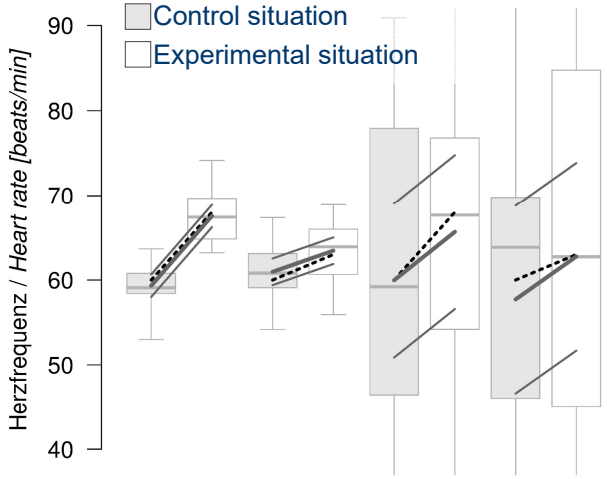
- From the data we calculate the value of a test statistic (t, F, x^2, other)
- Attached to each value of the test statistic is a probability called the p value.
- It describes the chance of getting the observed test statistic (or one more extreme) if the null hypothesis is true
- Rejection rule / criterion: How probable is the data assuming the null hypothesis is true?
  - P < 0.05 → reject H_0
  - P > 0.05 → accept H_0

Errare humanum est

	True Status	
Study decision	H_0	H_A
H_0		
H_A		

Effect
Fire

example



P-value	<0.0001	0.030	0.36	0.47
Size of effect	large	small	large	small
Variability	small		large	

What a p-value can(not) do

(The ASA statement)

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. The American Statistician **70,** 129-133; http://dx.doi.org/10.1080/00031305.2016.1154108 (and the follow-up articles in the same issue)

Special issue of The American Statistician (2019, 73, Suppl1) on "Statistical Inference in the 21st Century: A World Beyond p < 0.05"; https://www.tandfonline.com/toc/utas20/73/sup1?nav=tocList

- Sampling scheme
- Assignment subj/treat
- Experimental design
- Choice of stat. model
- Assumptions of model
- H_0, N, Δ, σ

$P(SSD \geq SSD_{obs} \mid H_0, model)$

TS= test statistic

SSD= Statistical summary of data

H_0= null hypothesis, H_A= alternative hypothesis

No direct information
results from the p-value alone
in respect to
the **scientific** hypothesis or
the biological/clinical **relevance**
of relationships or group differences

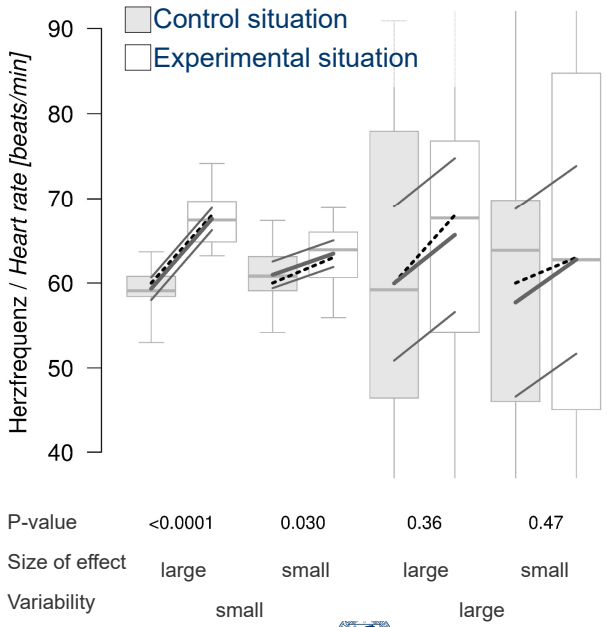
And now?

- Should we not use p-values? Statistical models?
- A low p-value indicates improbable data given the model
- Models are more than p-values:
  - Estimate the **strength** of relationshisps and differences (model estimates, predictions, fitted values)
  - level of **uncertainty** (confidence intervals)
  - identify patters in multi-dimensional space
- Necessary: discuss and interpret results in the context of the subject matter and previous knowledge
- An implausible result remains implausible even with a low p-value

What can we do immediately?

- report p-values as continuous numbers
- do not relate a p-value to an arbitrary criterion; delete «(non-)significant» from the vocabulary
- report measures of the size of the effect and ist variability (CI, credibility interval); integrate estimates / predictions
- report all evaluations that have been applied (transparency)
- differentiate explorative versus inferential experiments:
  - explorative: «anything goes»; generating hypotheses
  - inferential: «only as planned» (pre-registration)
- interpretation in the context of the subject area (e.g. plausibility; clincal, biological relevance)
- replicability is only achieved by repeating experiments

example



Outlook & conclusions

- Formal consideration of former knowledge: Bayesian statistics
- Other measures for statistical evidence (evidence ratio, false discovery rate)

- Be modest: The world cannot be saved in one experiment
- Courageously include expert knowledge

Descriptive Statistics

What you should know before the course

- We assume that all participants are somewhat familiar with:

- The concept of distribution
  (mostly Normal/Gaussian)

- Ways to describe numerical variables
  (central tendency and measures of dispersion)

Recap on distributions

- A **random variable** is a quantity whose value is not known (cp. outcome variables)

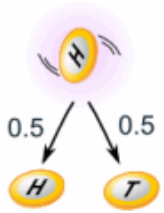
- A **probability distribution** comprises all the values that the random variable can take, with their associated probabilities

- Discrete distributions: finite set of integers
  e.g. binomial distribution

- Continuous distributions: infinite set of non-discrete numbers
  e.g. Normal distribution

Binomial distribution

- ...is the discrete probability distribution of the number of successes in a series of n independent trials in which each trial can result in either a success (with probability *p*) or a failure (with probability 1-*p*)
- possible values of the random variable are 0, 1, …, n successes



Binomial distribution

- ...is the discrete probability distribution of the number of *Leptospira* seropositive samples in a series of 30 independent serological samples in which each test can result in either a seroconversion (with probability *p*) or not (with probability 1-*p*)
- The binomial distribution is particularly important in statistics to analyse proportions.

Normal distribution

- also sometimes called Gaussian distribution
- ...is a good approximation to the distribution of many naturally occuring variables.
- ...has several useful properties

Normal distribution: properties

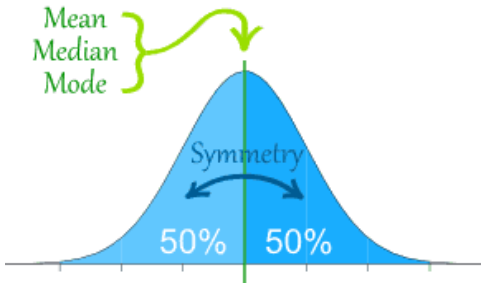
- Is completely described by 2 parameters:

$$X \sim N(\mu, \sigma)$$

- X comes from a Normal distribution with mean μ and standard deviation σ.

Normal distribution: properties

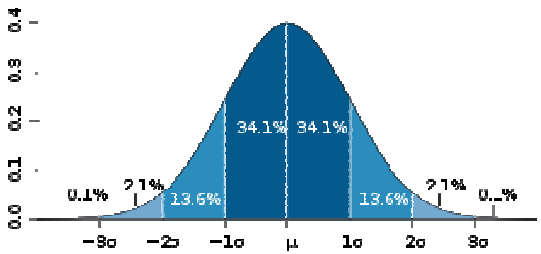
- It is unimodal
- It is symmetrical about its mean («bell-shaped»)
- Its mean, median and mode are all equal



www.mathsisfun.com

Normal distribution: properties

- 68% of values drawn from a Normal distribution are within one σ away from the mean μ
- about 95% of the values are within two σ.
- about 99.7% lie within three σ.



How can we quantitatively summarise a set of observations?

Measures of central tendency

- **Arithmetic mean**: the sum of all measurements divided by the number of observations in the data set

- **Median**: the middle value that separates the higher half from the lower half of the data set

- **Mode**: the most frequent value in the data set

Example

Data: 10, 10, 9, 5, 7, 8, 6, 10, 4, 12, 16, 10, 8
Mean = 8.85 ; Median = 9; Mode = 10;

How can we quantitatively summarise a set of observations?

A measure of statistical dispersion:

- **Range**: difference between largest and smallest observation

- the **quartiles** of a **ranked** set of data values are the three points that divide the data set into four equal groups, each group comprising a quarter of the data

Example

Data: 10, 10 ,9, 5, 7, 8, 6, 10, 4, 12, 16, 10, 8

Range: 12; Quartiles: 7,9,10

How can we quantitatively summarise a set of observations?

A measure of statistical dispersion:

- **Variance** $\sigma^2 = E\left[(X-\mu)^2\right]$
  - is the expected, or mean, value of the square of the deviation of that variable from its expected value or mean
- **Standard deviation** $\sigma = \sqrt{E\left[(X-\mu)^2\right]}$
  - is the square root of its variance.
- **Standard error** $SE = \frac{\sigma}{\sqrt{n}}$
  - Is the standard deviation divided by square root of the population size
  - Corresponds to the expected variation of the mean

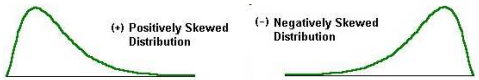
Example

Data: 10, 10, 9, 5, 7, 8, 6, 10, 4, 12, 16, 10, 8
Variance: 9.81 ; Standard deviation: 3.13 ; Standard error: 0.87

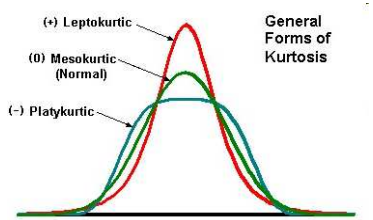
How can we quantitatively summarise a set of observations?

Measure of the shape of the distribution

- **Skewness**: a measure of asymmetry



- **Kurtosis**: a measure of the "peakedness"



Are simple descriptive measures good?

- Most commonly, the mean and standard deviation are reported as descriptive statistics
- These are only good = representative measures if the (raw) data is (approximately) normally distributed
  cp. barplot with error bar
- Otherwise, median, quartiles and the data range are much better
  cp. boxplot

Institute for Veterinary Epidemiology and Biostatistics (IVEB)
Department of Veterinary Medicine

Freie Universität Berlin

Inferential Statistics
Statistical Modelling

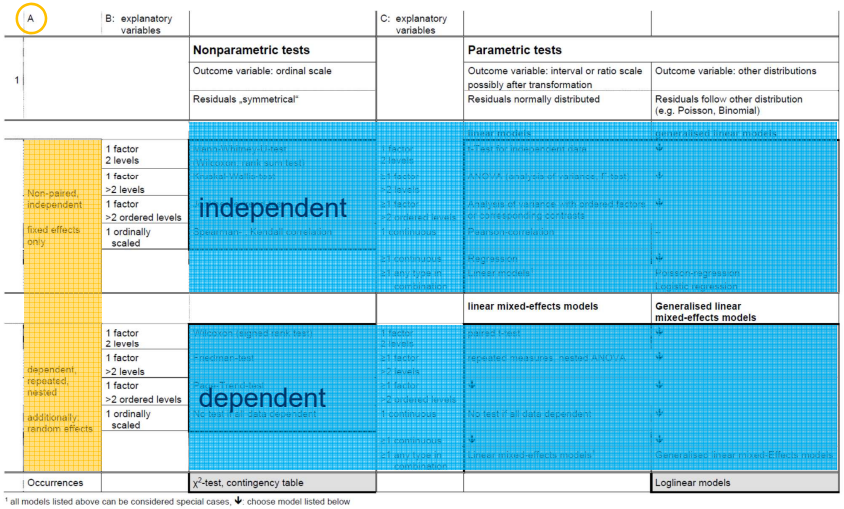
Choice of a statistical model

(1) What is the unit of observation? Are there dependencies in a data set (repeated measurements on the same unit, clustering of units in groups, etc.)?

(2) How is the **response** variable (outcome) distributed?
(cp. check of assumptions – residual analysis)

(3) How many explanatory variables (predictors) are on hand, what data type do they have?

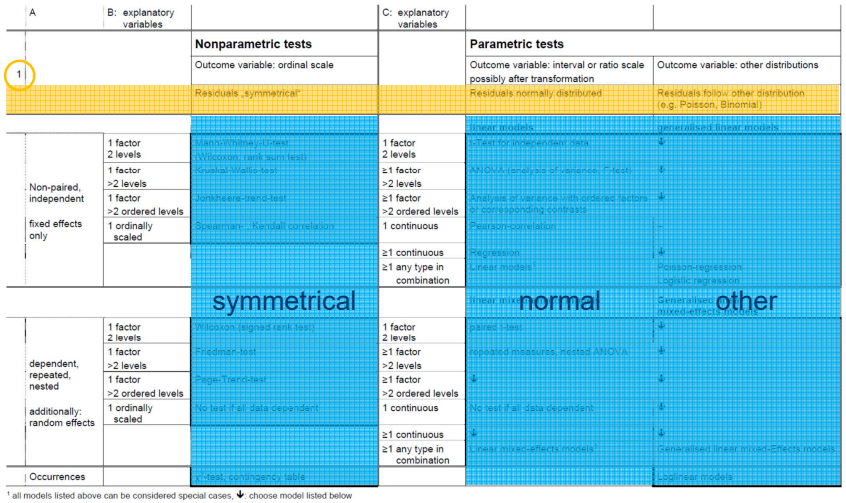
Dependecies?



A	B: explanatory variables		C: explanatory variables		
		Nonparametric tests		**Parametric tests**	
1		Outcome variable: ordinal scale		Outcome variable: interval or ratio scale possibly after transformation	Outcome variable: other distributions
		Residuals „symmetrical"		Residuals normally distributed	Residuals follow other distribution (e.g. Poisson, Binomial)
				linear models	**generalised linear models**
Non-paired, independent; fixed effects only	1 factor, 2 levels	Mann-Whitney-U-test (Wilcoxon, rank sum test)	1 factor, 2 levels	t-Test for independent data	↓
	1 factor, >2 levels	Kruskal-Wallis-test	≥1 factor, >2 levels	ANOVA (analysis of variance, F-test)	↓
	1 factor, >2 ordered levels	Jonkheere-trend-test	≥1 factor, >2 ordered levels	Analysis of variance with ordered factors or corresponding contrasts	↓
	1 ordinally scaled	Spearman- , Kendall correlation	1 continuous	Pearson-correlation	–
			≥1 continuous	Regression	↓
			≥1 any type in combination	Linear models[1]	Poisson-regression Logistic regression
				linear mixed-effects models	**Generalised linear mixed-effects models**
dependent, repeated, nested; additionally: random effects	1 factor, 2 levels	Wilcoxon (signed rank test)	1 factor, 2 levels	paired t-test	↓
	1 factor, >2 levels	Friedman-test	≥1 factor, >2 levels	repeated measures, nested ANOVA	↓
	1 factor, >2 ordered levels	Page-Trend-test	≥1 factor, >2 ordered levels	↓	↓
	1 ordinally scaled	No test if all data dependent	1 continuous	No test if all data dependent	↓
			≥1 continuous	↓	↓
			≥1 any type in combination	Linear mixed-effects models[1]	Generalised linear mixed-Effects models
Occurrences		χ^2-test, contingency table			Loglinear models

[1] all models listed above can be considered special cases, ↓: choose model listed below

Outcome(s)?



A	B: explanatory variables		C: explanatory variables		
		Nonparametric tests		**Parametric tests**	
1		Outcome variable: ordinal scale		Outcome variable: interval or ratio scale possibly after transformation	Outcome variable: other distributions
		Residuals „symmetrical"		Residuals normally distributed	Residuals follow other distribution (e.g. Poisson, Binomial)
				linear models	**generalised linear models**
Non-paired, independent; fixed effects only	1 factor, 2 levels	Mann-Whitney-U-test (Wilcoxon, rank sum test)	1 factor, 2 levels	t-Test for independent data	↓
	1 factor, >2 levels	Kruskal-Wallis-test	≥1 factor, >2 levels	ANOVA (analysis of variance, F-test)	↓
	1 factor, >2 ordered levels	Jonkheere-trend-test	≥1 factor, >2 ordered levels	Analysis of variance with ordered factors or corresponding contrasts	↓
	1 ordinally scaled	Spearman- , Kendall correlation	1 continuous	Pearson-correlation	–
			≥1 continuous	Regression	↓
			≥1 any type in combination	Linear models[1]	Poisson-regression Logistic regression
				linear mixed-effects models	**Generalised linear mixed-effects models**
dependent, repeated, nested; additionally: random effects	1 factor, 2 levels	Wilcoxon (signed rank test)	1 factor, 2 levels	paired t-test	↓
	1 factor, >2 levels	Friedman-test	≥1 factor, >2 levels	repeated measures, nested ANOVA	↓
	1 factor, >2 ordered levels	Page-Trend-test	≥1 factor, >2 ordered levels	↓	↓
	1 ordinally scaled	No test if all data dependent	1 continuous	No test if all data dependent	↓
			≥1 continuous	↓	↓
			≥1 any type in combination	Linear mixed-effects models[1]	Generalised linear mixed-Effects models
Occurrences		χ^2-test, contingency table			Loglinear models

[1] all models listed above can be considered special cases, ↓: choose model listed below

Explanatory variables?

A	B: explanatory variables		C: explanatory variables		
		Nonparametric tests		**Parametric tests**	
1		Outcome variable: ordinal scale		Outcome variable: interval or ratio scale possibly after transformation	Outcome variable: other distributions
		Residuals „symmetrical"		Residuals normally distributed	Residuals follow other distribution (e.g. Poisson, Binomial)
				linear models	**generalised linear models**
Non-paired, independent; fixed effects only	1 factor, 2 levels	Mann-Whitney-U-test (Wilcoxon, rank sum test)	1 factor, 2 levels	t-Test for independent data	↓
	1 factor, >2 levels	Kruskal-Wallis-test	≥1 factor, >2 levels	ANOVA (analysis of variance, F-test)	↓
	1 factor, >2 ordered levels	Jonkheere-trend-test	≥1 factor, >2 ordered levels	Analysis of variance with ordered factors or corresponding contrasts	↓
	1 ordinally scaled	Spearman- , Kendall correlation	1 continuous	Pearson-correlation	–
			≥1 continuous	Regression	↓
			≥1 any type in combination	Linear models[1]	Poisson-regression Logistic regression
				linear mixed-effects models	**Generalised linear mixed-effects models**
dependent, repeated, nested; additionally: random effects	1 factor, 2 levels	Wilcoxon (signed rank test)	1 factor, 2 levels	paired t-test	↓
	1 factor, >2 levels	Friedman-test	≥1 factor, >2 levels	repeated measures, nested ANOVA	↓
	1 factor, >2 ordered levels	Page-Trend-test	≥1 factor, >2 ordered levels	↓	↓
	1 ordinally scaled	No test if all data dependent	1 continuous	No test if all data dependent	↓
			≥1 continuous	↓	↓
			≥1 any type in combination	Linear mixed-effects models[1]	Generalised linear mixed-Effects models
Occurrences		χ^2-test, contingency table			Loglinear models

[1] all models listed above can be considered special cases, ↓: choose model listed below

The most general model (all others are special cases)

A	B: explanatory variables		C: explanatory variables		
		Nonparametric tests		**Parametric tests**	
1		Outcome variable: ordinal scale		Outcome variable: interval or ratio scale possibly after transformation	Outcome variable: other distributions
		Residuals „symmetrical"		Residuals normally distributed	Residuals follow other distribution (e.g. Poisson, Binomial)
				linear models	**generalised linear models**
Non-paired, independent; fixed effects only	1 factor, 2 levels	Mann-Whitney-U-test (Wilcoxon, rank sum test)	1 factor, 2 levels	t-Test for independent data	↓
	1 factor, >2 levels	Kruskal-Wallis-test	≥1 factor, >2 levels	ANOVA (analysis of variance, F-test)	↓
	1 factor, >2 ordered levels	Jonkheere-trend-test	≥1 factor, >2 ordered levels	Analysis of variance with ordered factors or corresponding contrasts	↓
	1 ordinally scaled	Spearman- , Kendall correlation	1 continuous	Pearson-correlation	–
			≥1 continuous	Regression	↓
			≥1 any type in combination	Linear models[1]	Poisson-regression Logistic regression
				linear mixed-effects models	**Generalised linear mixed-effects models**
dependent, repeated, nested; additionally: random effects	1 factor, 2 levels	Wilcoxon (signed rank test)	1 factor, 2 levels	paired t-test	↓
	1 factor, >2 levels	Friedman-test	≥1 factor, >2 levels	repeated measures, nested ANOVA	↓
	1 factor, >2 ordered levels	Page-Trend-test	≥1 factor, >2 ordered levels	↓	↓
	1 ordinally scaled	No test if all data dependent	1 continuous	No test if all data dependent	↓
			≥1 continuous	↓	↓
			≥1 any type in combination	Linear mixed-effects models[1]	Generalised linear mixed-Effects models
Occurrences		χ^2-test, contingency table			Loglinear models

[1] all models listed above can be considered special cases, ↓: choose model listed below

Standardisation
Replicability, Validity

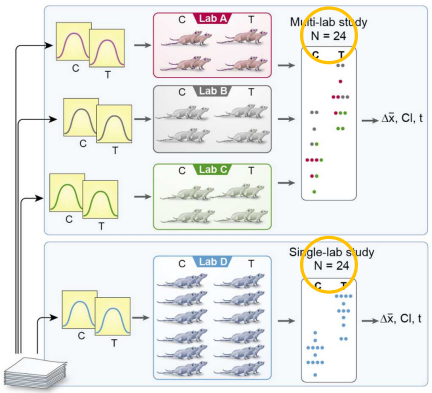
Standardisation & reproducibility

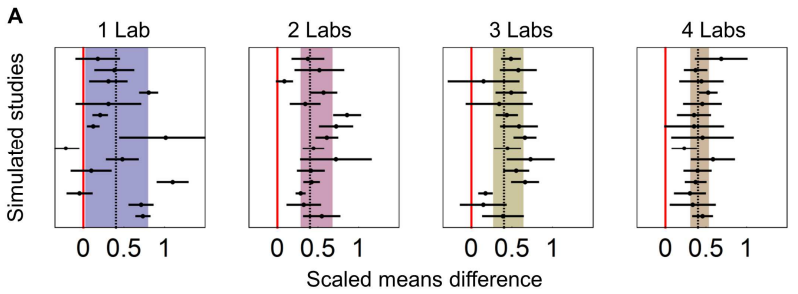
- Voelkl B, Vogt L, Sena ES, Würbel H (2018) Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLoS Biol 16(2): e2003693. https://doi.org/10.1371/journal.pbio.2003693

- Standardisation: keeping constant of as many (potential) nuisance variables as possible (genetics, housing, interaction)
- Is thought to be good praxis
  (≠ unlike of the aspects of the problematic of reproducibility)
- Problem:
  in single lab studes (e.g. pre-clinical research) erratic factors will have an effect (different phenotypes are tested)
- Validity only under the highly specific circumstances of the given laboratory
- (In spite of) **due to** standardisation, deficient reproducibility

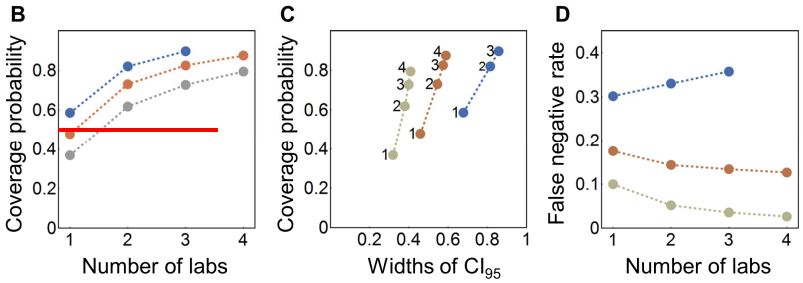
Approach

- 50 independent studies
  on the effect of therapeutic hypothermia on infarct volume in rodent models
- «True» effect (meta-analysis):
  Reduction of infarct volume 47.8% [40.6-55.0%]
- Simulation of single-lab and multi-lab studies





- Showing each 15 random simulation (with N = 24 each)
- ——0-effect; ····· «true» effect
- Coloured areas: CI ~ simulations



- 10^5 simulations
- N = 12; N = 24; N = 48
- replicated with 12 different interventions

Conclusions reproducibility

Improved external validity / reproducibility can be reached

- Multi-lab studies (2-4 labs)
- Heterogenisation of the sample
- **without** enlargement of sample