R course
–
Experimental design and hypothesis testing

12. – 16.09.2016

Max Planck Institute for Marine Microbiology

Christiane Hassenrück

Experimental design

Always know how to analyze the data before collecting it!

→ Be aware what kind of data you are collecting: continuous, discrete, percentages, binary, etc.

→ Be aware of the assumptions of the statistical tests suitable for your kind of data

→ Be aware of the limitations of field, laboratory, and statistical techniques

→ Always collect back-up samples

→ Always have a plan B (or C, or D, …)

Experimental design

Always know how to analyze the data before collecting it!

Ideal scenario

- > 10 replicates for each treatment
- Additional technical replicates
- Balanced sampling design
- Normally distributed data
- No missing data
- No outliers
- No (observer) bias
- No confounding variables

Reality

- ~ 3 replicates for each treatment because of logistic constraints
- Technical replicates not comparable
- Unbalanced sampling design due to site inaccessibility
- Irregular data distribution
- Missing data due to failed measurements
- Many outliers
- Strong biases
- Highly confounded environmental data
- …

Experimental design

Always know how to analyze the data before collecting it!

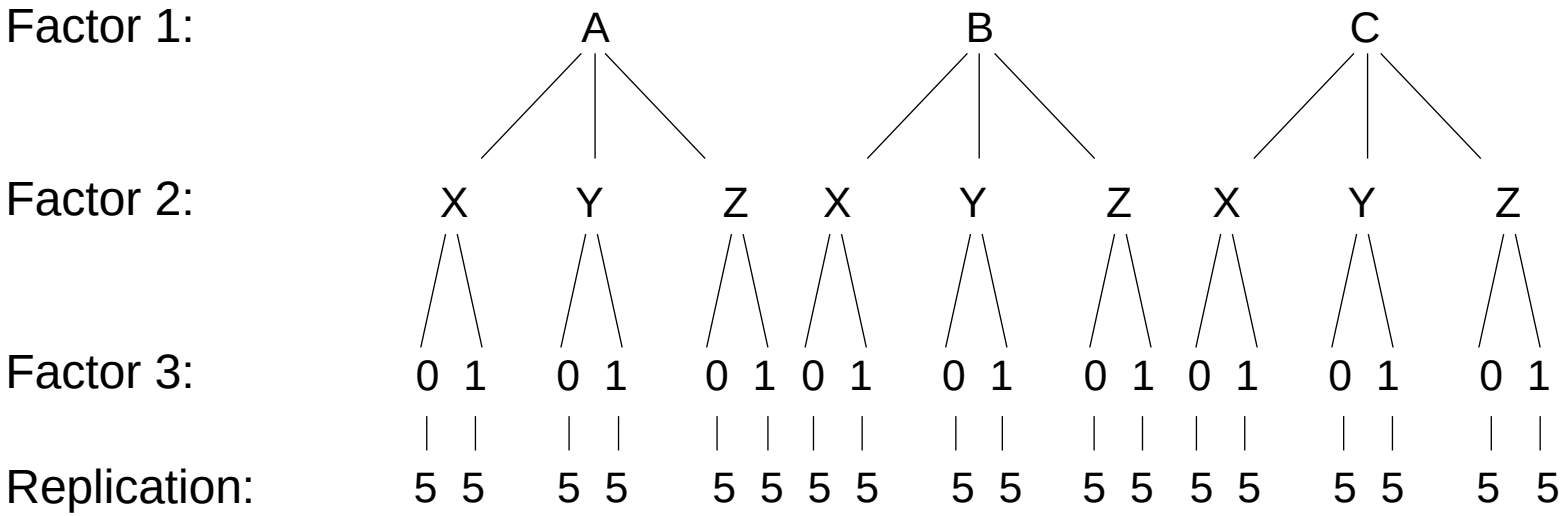
Compromise

- Irregular data distributions:
  - Simple sampling design
  - Consider non-parametric tests and/or permutaion tests
- Missing data:
  - Collect more samples than necessary, even if they are not going to be analyzed
- Logistic contraints:
  - Instead of reducing the number of replicates, reduce the number of treatments
  - Avoid pseudoreplication

Experimental design

Example:

- 16S amplicon sequencing: ~ 50 € per sample



Estimated costs:

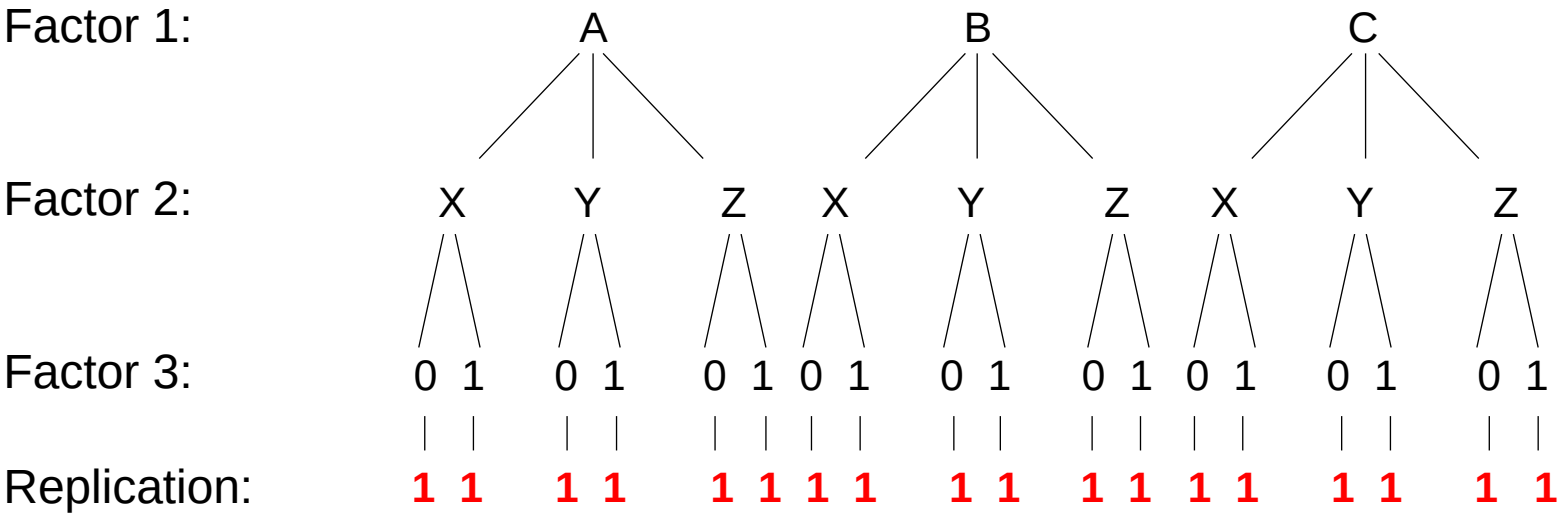
- 16S amplicon sequencing:

3 x 3 x 2 x 5 = 90 replicates x 50 € = ~ 4,500 €

Experimental design

Example:

- 16S amplicon sequencing: ~ 50 € per sample



Estimated costs:

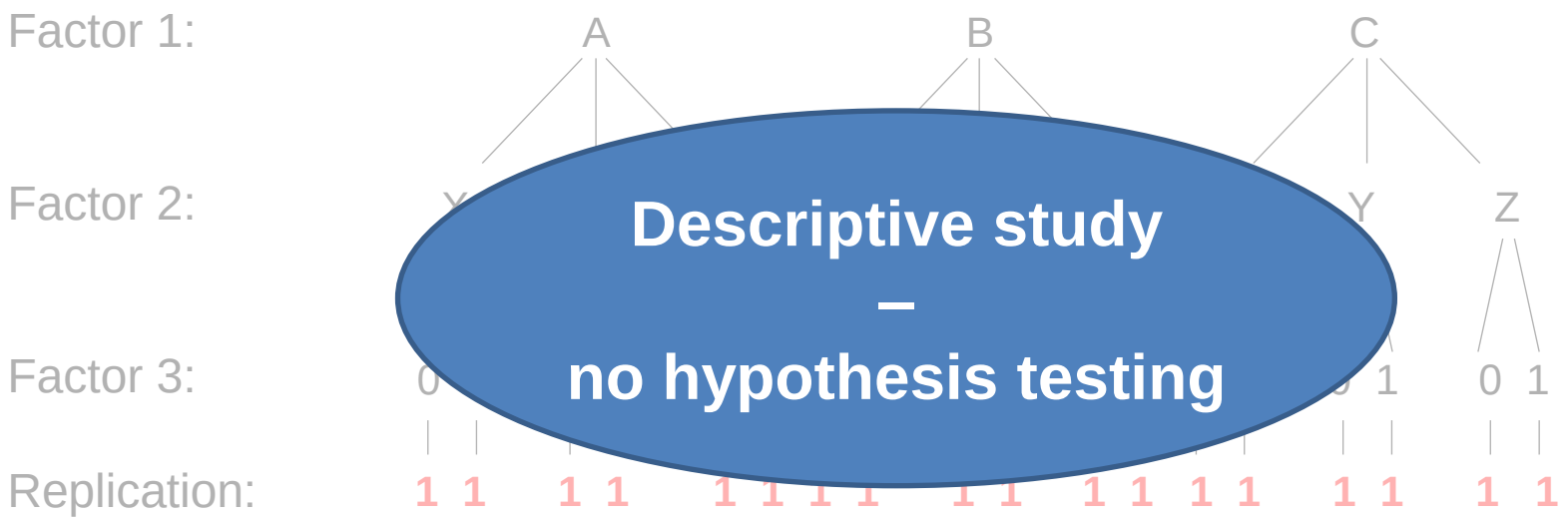
- 16S amplicon sequencing:

3 x 3 x 2 x **1** = **18** replicates x 50 € = ~ **900** €

Experimental design

Example:

- 16S amplicon sequencing: ~ 50 € per sample



Estimated costs:

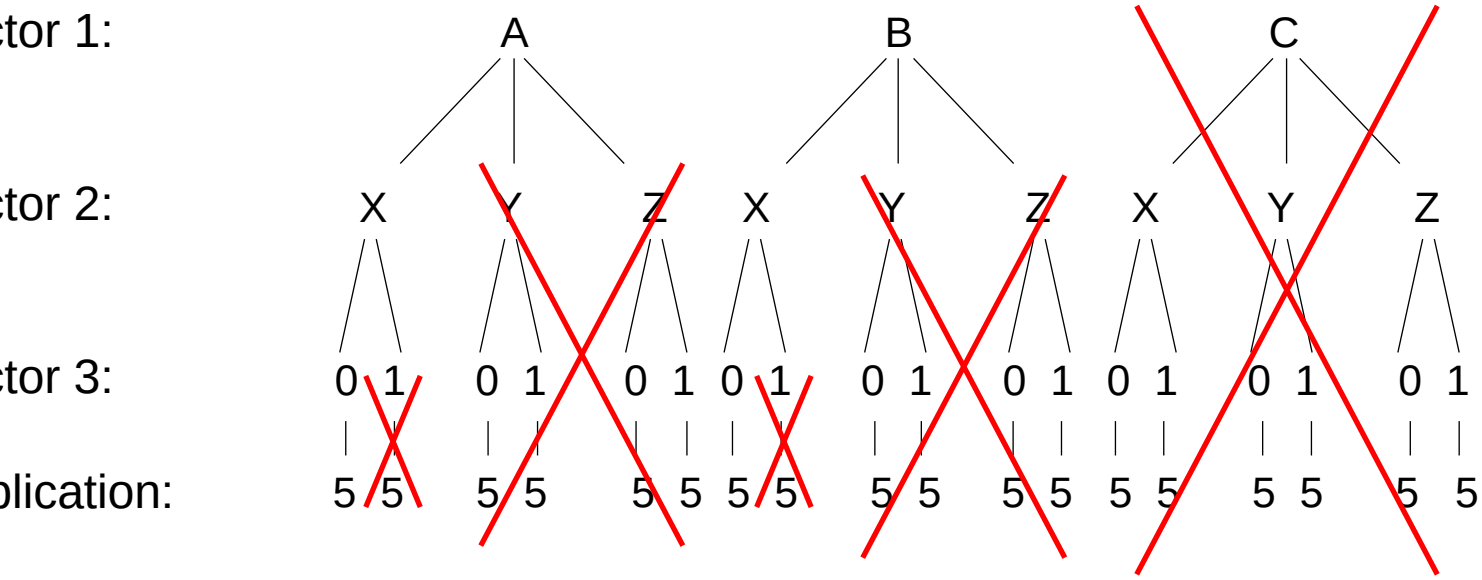
- 16S amplicon sequencing:

3 x 3 x 2 x **1** = **18** replicates x 50 € = ~ **900** €

Experimental design

Example:

- 16S amplicon sequencing: ~ 50 € per sample



Estimated costs:

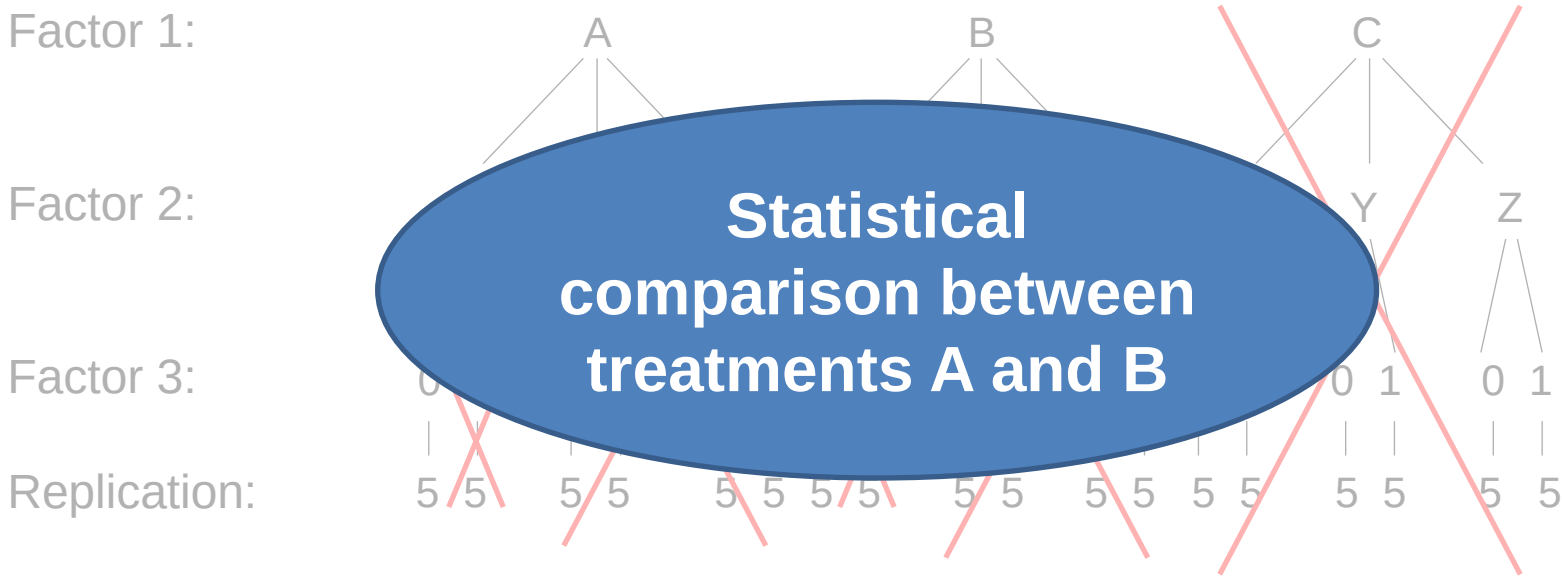
- 16S amplicon sequencing:

2 x **1** x **1** x 5 = **10** replicates x 50 € = ~ **500** €

Experimental design

Example:

- 16S amplicon sequencing: ~ 50 € per sample



Estimated costs:

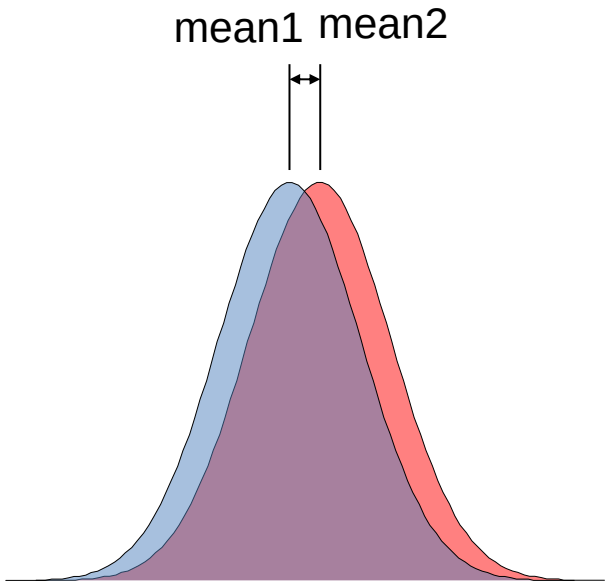
- 16S amplicon sequencing:

2 x **1** x **1** x 5 = **10** replicates x 50 € = ~ **500** €

Hypothesis testing

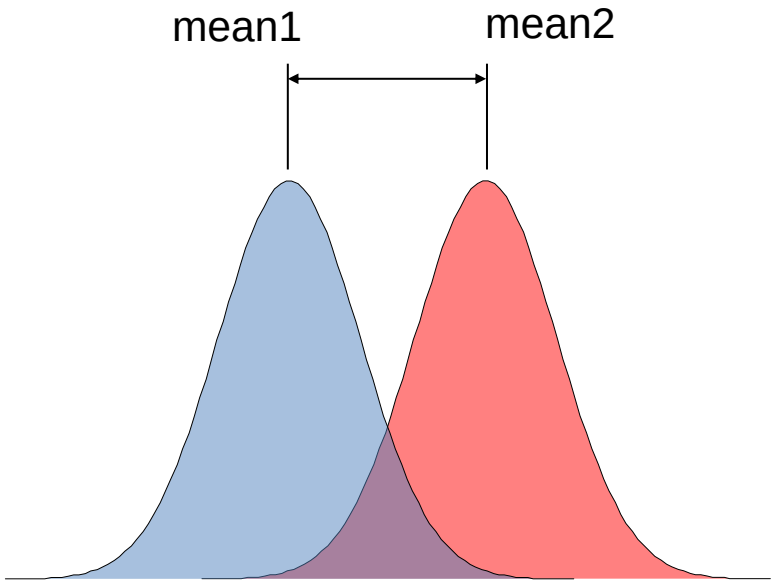
H_0: Null hypothesis

There is no significant difference between treatments



H_A: alternative hypothesis

There is a significant difference between treatments



R cannot tell you which test to use!
This course: only implementation in R

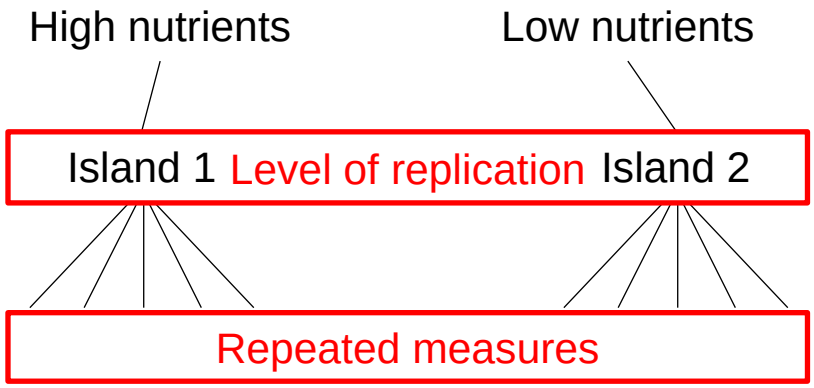
Assumptions of statistical tests

- Independence of observations (name examples)
  - No pseudoreplication
- Balanced sampling design (preferred)
  - Same number of observations per treatment
- Conformity to theoretical distribution (parametric)
  - Otherwise test statistic (and p-value) meaningless
- Homoscedacity (parametric)
  - E.g. variance should not increase with the mean

- Check assumptions in R:
  - `leveneTest {car}`
  - `shapiro.test {stats}`
  - `qqnorm {stats}`
  - `plot(residuals(model) ~ fitted(model))`

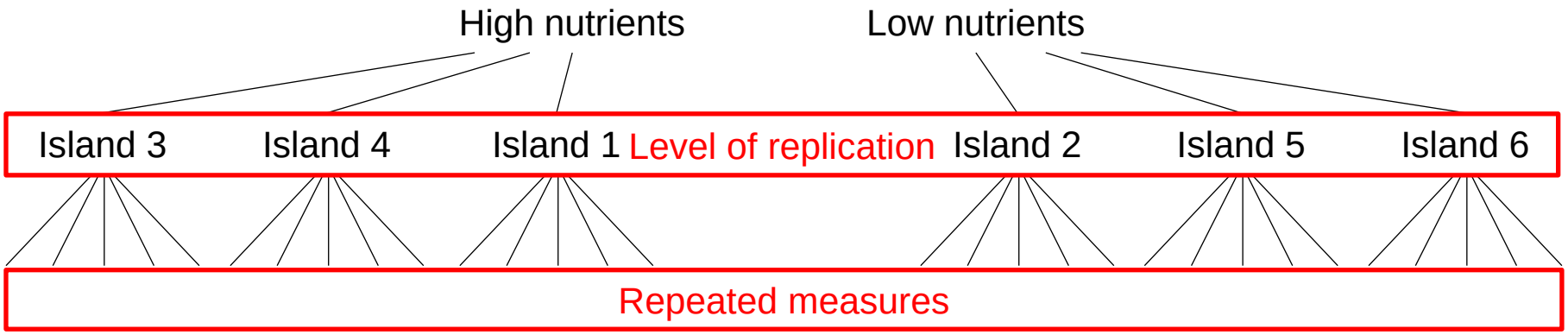
- Alternatives:
  - Data tranformations (e.g. log, sqrt, asin)
  - Permutation test

Pseudoreplication



- Conclusions only about difference between islands, not nutrient effect

Pseudoreplication



Repeated measures can be a good thing…

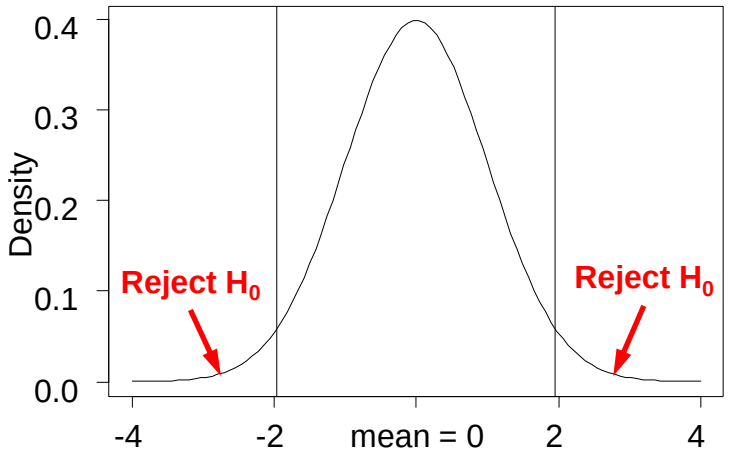
… if they are accounted for in the statistical model!

- Option 1: use means
- Option 2: build mixed models (introduce random factor ‘Island’)

P-values

- P-values are (usually) calculated based on a theoretical mathematical distribution

  → Reason for normality assumption

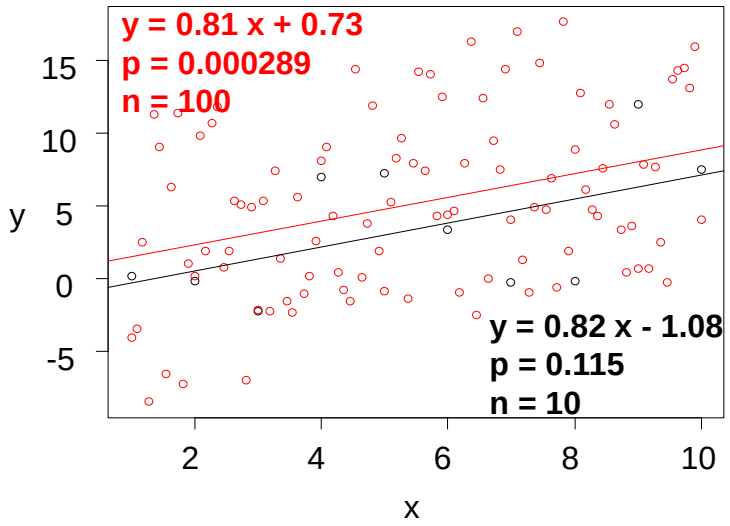


P-values

- P-values are (usually) calculated based on a theoretical mathematical distribution

  → Reason for normality assumption

- P-values depend on effect size (test statistic) and sampling size (degrees of freedom)



P-values

- P-values are (usually) calculated based on a theoretical mathematical distribution

  → Reason for normality assumption

- P-values depend on effect size (test statistic) and sampling size (degrees of freedom)

- The widely used significance threshold of $\alpha = 0.05$ is an arbitrary number, chosen as compromise between type I and type II errors

	Accept H_0	Reject H_0
$p > 0.05$	Correct	Type II (false negative)
$p < 0.05$	Type I (false positive)	Correct

P-values

- P-values are (usually) calculated based on a theoretical mathematical distribution

  → Reason for normality assumption

- P-values depend on effect size (test statistic) and sampling size (degrees of freedom)

- The widely used significance threshold of $\alpha = 0.05$ is an arbitrary number, chosen as compromise between type I and type II errors

- Running several tests on the same data set leads to p-value inflation

$$FWER = 1 - (1 - alpha)^n$$

Family-wise error rate

Significance threshold per comparison

Number of comparisons

n	FWER
1	0.05
3	0.14
1000	~ 1

	Parametric	Non-parametric
Univariate		
Compare 2 treatments	**T-test**	Mann-Whitney-U (**Wilcoxon**)
Compare > 2 treatments	**ANOVA** **Post-hoc: TukeyHSD**	**Kruskal-Wallis**
Continuous explanatory variable	**Linear regression**	Spearman correlation (not regression anymore)
Mixed effects models	**GLMMs** **Post-hoc**	
Multivariate		
Ordination	**PCA** PCoA CA	**NMDS**
Hypothesis testing	**RDA** CCA	**ANOSIM** **PERMANOVA**
Comparing 2 multivariate objects	**Mantel-test** **Procrustes-test**	
Causal interactions	Path analysis	

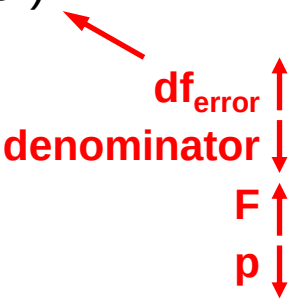
Mixed-effects models

- Extension of GLMs
- Additional feature: include random effects

- GLM: $$F = \frac{explained\ variation}{unexplained\ variation} = \frac{SS_{fixed}/df_{fixed}}{SS_{error}/df_{error}}$$

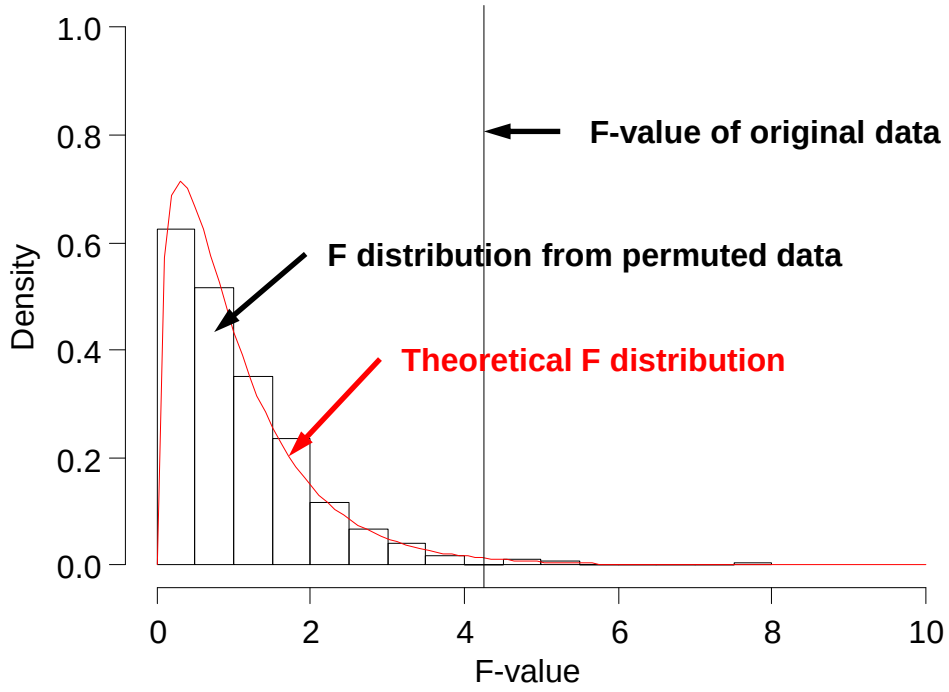
- GLMM: $$F = \frac{explained\ variation}{unexplained\ variation} = \frac{SS_{fixed}/df_{fixed}}{SS_{random}/df_{random}}$$

- Example: repeated measurements
  - 3 treatments x 10 replicates x 3 measurements = 90 values
  - df_{fixed} ~ number of treatments
  - df_{error} ~ total sample size (without random factor)
  - df_{random} ~ number of levels in random factor



Permutations tests

- Create your own theoretical distribution of the test statistic
- Randomly reshuffle the response variable



- Implemented in R as default for several tests (mostly multivariate tests)

Dissimilarity and distance

- Community (dis)similarity between samples

	OTU1	OTU2	OTU3	OTU4
S1	14	2	14	14
S2	10	14	0	8
S3	0	5	0	2
S4	0	0	1	0

presence/ absence →

	OTU1	OTU2	OTU3	OTU4
S1	1	1	1	1
S2	1	1	0	1
S3	0	1	0	1
S4	0	0	1	0

Asymmetrical vs. symmetrical
Bray-Curtis vs. euclidean

Jaccard

	S1	S2	S3	S4
S1	0			
S2	0.5	0		
S3	0.8	0.6	0	
S4	1.0	1	1	0

	S1	S2	S3	S4
S1	0			
S2	19.8	0		
S3	23.3	14.7	0	
S4	23.8	19	5.5	0

	S1	S2	S3	S4
S1	0			
S2	0.25	0		
S3	0.5	0.33	0	
S4	0.75	1	1	0

- Zeros in ecology: Is this species really not there or did we just not find it?

→ double zeros not relevant

PCA

- Continuous environmental data
- Metric ordination based on euclidean distances
- Create new axes (principal components) along direction of highest variability ($N_{PC} = N_{variables}$)

NMDS

- Species abundance data
- Non-metric ordination based on any kind of distance/dissimilarity measure
- Show maximum variation in 2 (or 3) dimensions

Visualization of higher dimensional data

More information:

GUide to STatistical Analysis in Microbial Ecology

http://mb3is.megx.net/gustame

RDA

- Linear technique
- Variation in response matrix (e.g. microbial communities) explained by explanatory variables (e.g. environmental parameters)

PERMANOVA

Hypothesis testing

- Non-parametric multivariate ANOVA
- ANOVA based on on ranked dissimilarities
- Multifactorial design

ANOSIM

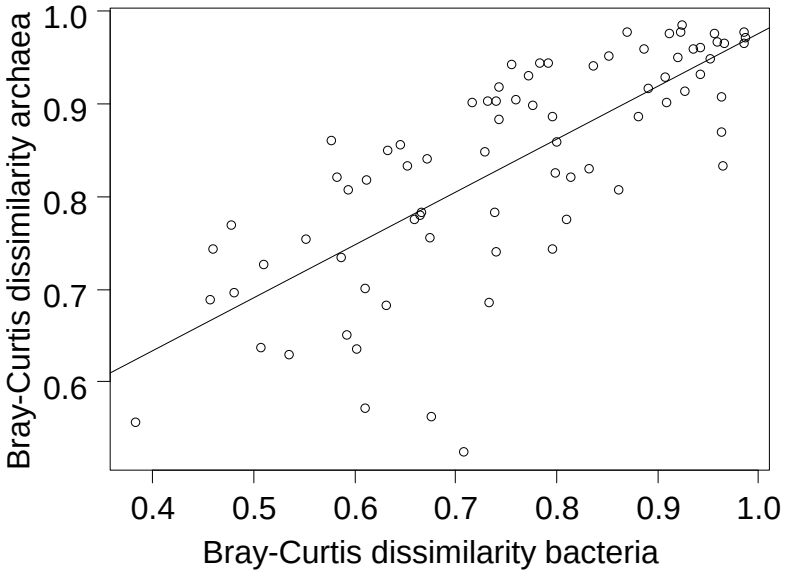
- Hypothesis testing
- based on on ranked dissimilarities
- Unifactorial design

Mantel test

Procrustes test

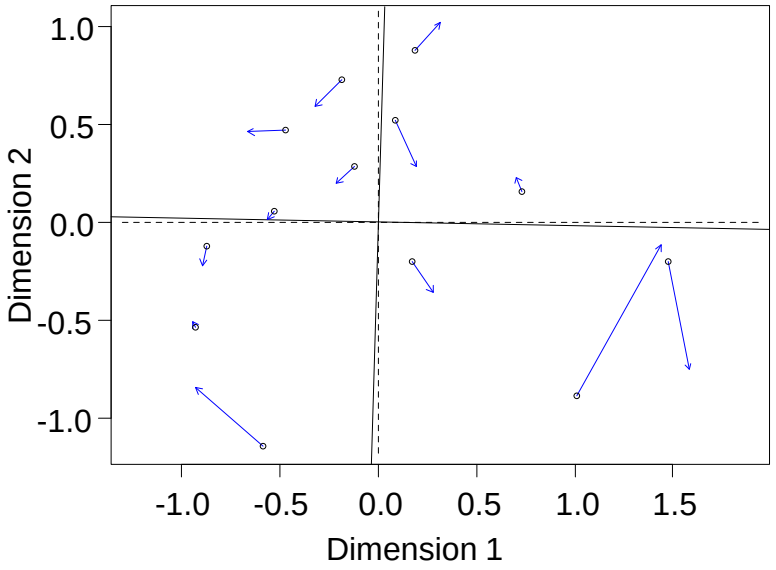
Comparison of 2 multivariate data sets

- Correlation of dissimilarity matrices
- Comparison based on all variation



Mantel statistic r: 0.7511
Significance: 0.001

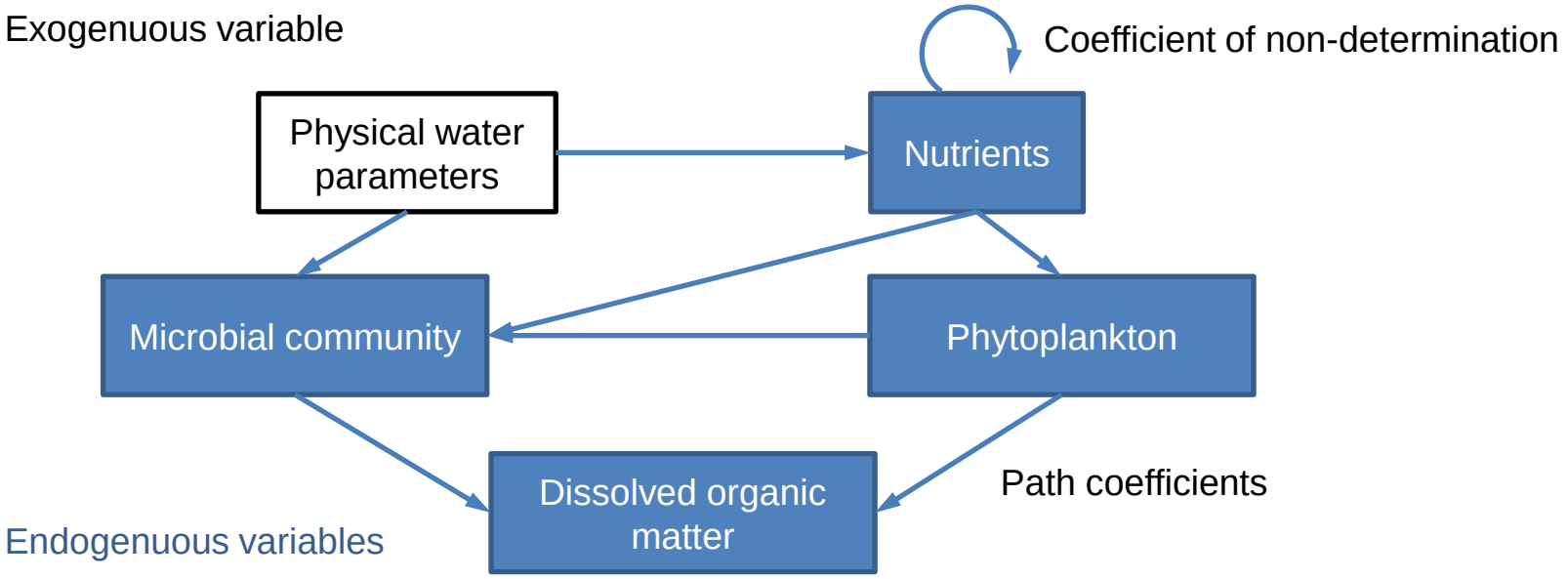
- Correlation of ordination objects
- Comparison based on the majority of the variation



Procrustes SS: 0.1266
Correlation (symmetric rotation): 0.9346
Significance: 0.001

Path analysis

- Causal (directed) relationships between blocks of variables
- Based on *a priori* hypotheses about biotic and abiotic interactions
- Large sample size (> 10 observations per path)
- Relationships between variable blocks:
  - Regression coefficient
  - (Partial) Mantel statistic
  - Multivariate coefficient of correlation (RV)



More examples

GUide to STatistical Analysis in Microbial Ecology

http://mb3is.megx.net/gustame