BIOSTATISTICS

. . . wait. What!?

Erik Kusch

erik.kusch@au.dk

Section for Ecoinformatics & Biodiversity
Center for Biodiversity and Dynamics in a Changing World (BIOCHANGE)
Aarhus University

28/10/2020

The Big Question

Should you care about biostatistics?

The Big Question

YES!

The Big Question

YES!

Thank you for attending my TED talk.

Biological Terminology

No, biostatistics are **not just for math nerds**.

Her: I'm a stats major
Me: [trying to think of something to impress her] **yea I'm bad at math too**

Statisticians don't know important **biological background:**

- *Population* vs. *Sample*
- *Species*, *Family*, *Taxon*, etc.
- *Interpretation* of results

Biologists don't know important **statistical background:**

- *Unsupervised* vs. *Supervised Approaches*
- *Statistical Assumptions*
- *Parametric* vs. *Non-Parametric* Tests

Biological Terminology

No, biostatistics are **not just for math nerds**.

Her: I'm a stats major
Me: [trying to think of something to impress her] **yea I'm bad at math too**

Statisticians don't know important **biological background:**

- *Population* vs. *Sample*
- *Species*, *Family*, *Taxon*, etc.
- *Interpretation* of results

Biologists don't know important **statistical background:**

- *Unsupervised* vs. *Supervised Approaches*
- *Statistical Assumptions*
- *Parametric* vs. *Non-Parametric* Tests

Biological Terminology

No, biostatistics are **not just for math nerds**.

Her: I'm a stats major
Me: [trying to think of something to impress her] **yea I'm bad at math too**

Statisticians don't know important **biological background:**

- *Population* vs. *Sample*
- *Species*, *Family*, *Taxon*, etc.
- *Interpretation* of results

Biologists don't know important **statistical background:**

- *Unsupervised* vs. *Supervised Approaches*
- *Statistical Assumptions*
- *Parametric* vs. *Non-Parametric* Tests

Basic Statistics

How often **do you** actually **check assumptions**?

- *Assumptions*:
  - Normality
  - Independence
  - Homogeneity of variances

→ Testing? Remedies?

- *Scales* and *Distributions*:
  - Continuous, Categorical
  - Nominal, Binary, Ordinal, Interval, Relation/Ratio, Integer
  - Gaussian Normal, Binomial, Poisson

→ Distinguish them?

Basic Statistics

How often **do you** actually **check assumptions**?

- *Assumptions*:
  - Normality
  - Independence
  - Homogeneity of variances

→ Testing? Remedies?

- *Scales* and *Distributions*:
  - Continuous, Categorical
  - Nominal, Binary, Ordinal, Interval, Relation/Ratio, Integer
  - Gaussian Normal, Binomial, Poisson

→ Distinguish them?

Basic Statistics

How often **do you** actually **check assumptions**?

- *Assumptions*:
  - Normality
  - Independence
  - Homogeneity of variances

→ Testing? Remedies?

- *Scales* and *Distributions*:
  - Continuous, Categorical
  - Nominal, Binary, Ordinal, Interval, Relation/Ratio, Integer
  - Gaussian Normal, Binomial, Poisson

→ Distinguish them?

Basic Statistics

How often **do you** actually **check assumptions**?

- *Assumptions*:
  - Normality
  - Independence
  - Homogeneity of variances

→ Testing? Remedies?

- *Scales* and *Distributions*:
  - Continuous, Categorical
  - Nominal, Binary, Ordinal, Interval, Relation/Ratio, Integer
  - Gaussian Normal, Binomial, Poisson

→ Distinguish them?

Basic Statistics

How often **do you** actually **check assumptions**?

- *Assumptions*:
  - Normality
  - Independence
  - Homogeneity of variances

→ Testing? Remedies?

- *Scales* and *Distributions*:
  - Continuous, Categorical
  - Nominal, Binary, Ordinal, Interval, Relation/Ratio, Integer
  - Gaussian Normal, Binomial, Poisson

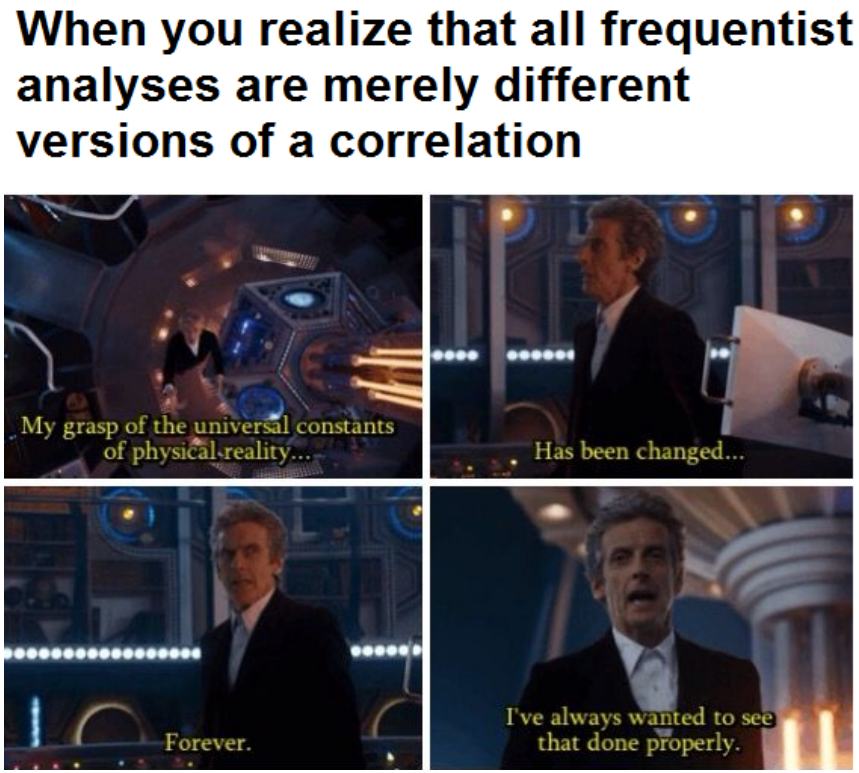
→ Distinguish them?

Correlations

Correlation is **not** necessarily **causation**.

Correlation tests yield two measurements:

- *r* value (measure of correlation)
  - $r \approx 1$ (strong, positive correlation)
  - $r \approx 0$ (no correlation)
  - $r \approx -1$ (strong, negative correlation)
- *p* value (measure of statistical significance)



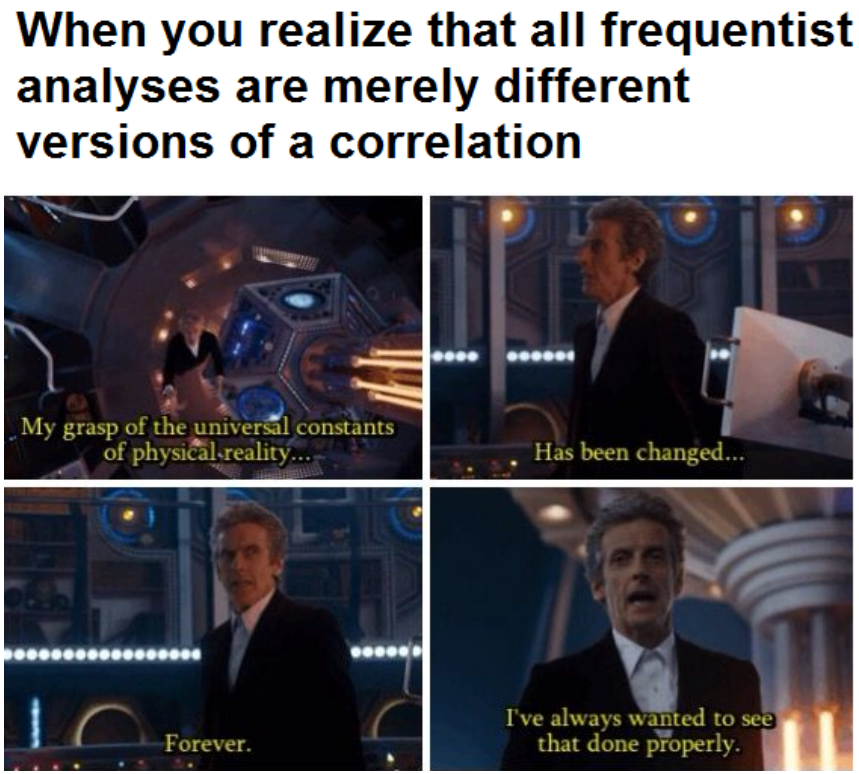
→ Get a feeling for it here http://guessthecorrelation.com/

Correlations

Correlation is **not** necessarily **causation**.

Correlation tests yield two measurements:

- *r* value (measure of correlation)
  - $r \approx 1$ (strong, positive correlation)
  - $r \approx 0$ (no correlation)
  - $r \approx -1$ (strong, negative correlation)
- *p* value (measure of statistical significance)



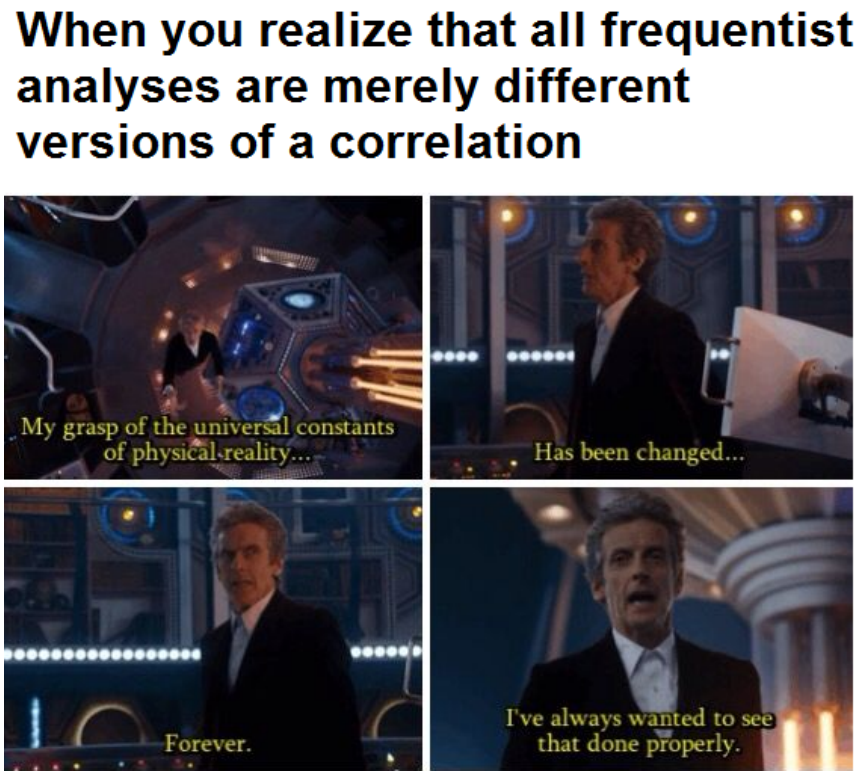
→ Get a feeling for it here http://guessthecorrelation.com/

Correlations

Correlation is **not** necessarily **causation**.

Correlation tests yield two measurements:

- *r* value (measure of correlation)
  - $r \approx 1$ (strong, positive correlation)
  - $r \approx 0$ (no correlation)
  - $r \approx -1$ (strong, negative correlation)
- *p* value (measure of statistical significance)



→ Get a feeling for it here http://guessthecorrelation.com/

Advanced Statistics

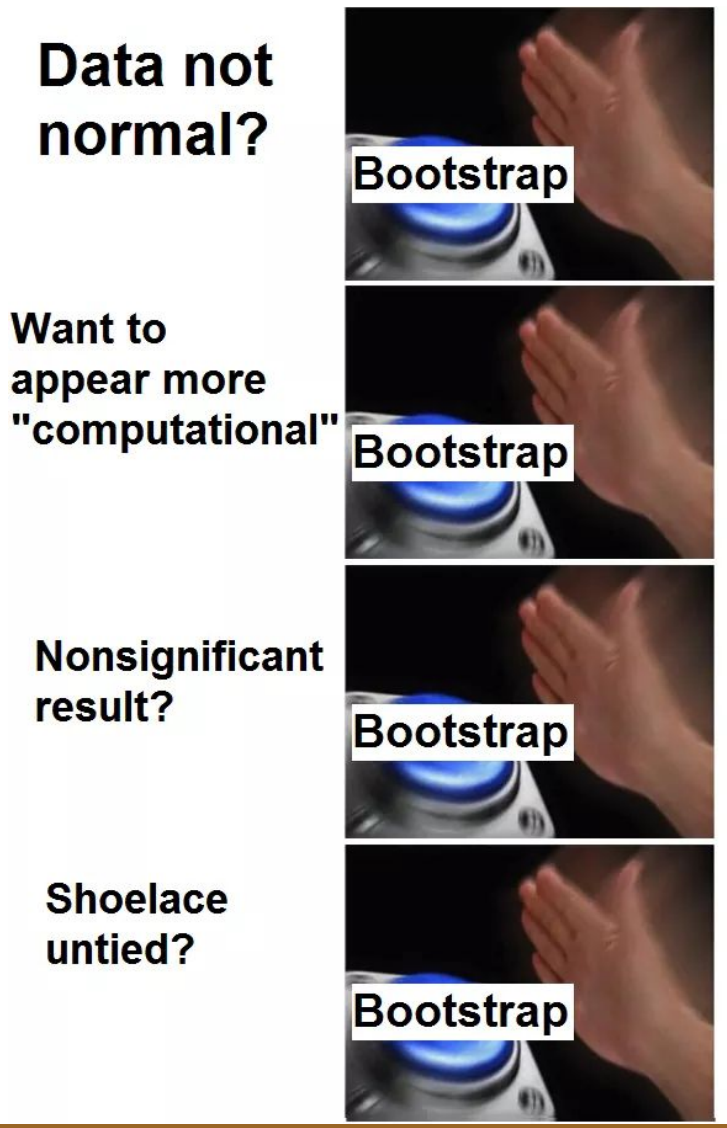
What do you want to **analyse** and **predict**?

- *Classifications*:
  - K-Means
  - Support-Vector Machines
  - Hierarchies
  - Networks

→ When to use which one?

- *Regression*:
  - Linear Models
  - Least Squares vs. Maximum Likelihood
  - Mixed Effect Models
  - GLS/GLM, and GAM

→ How do you select the best model?



Advanced Statistics

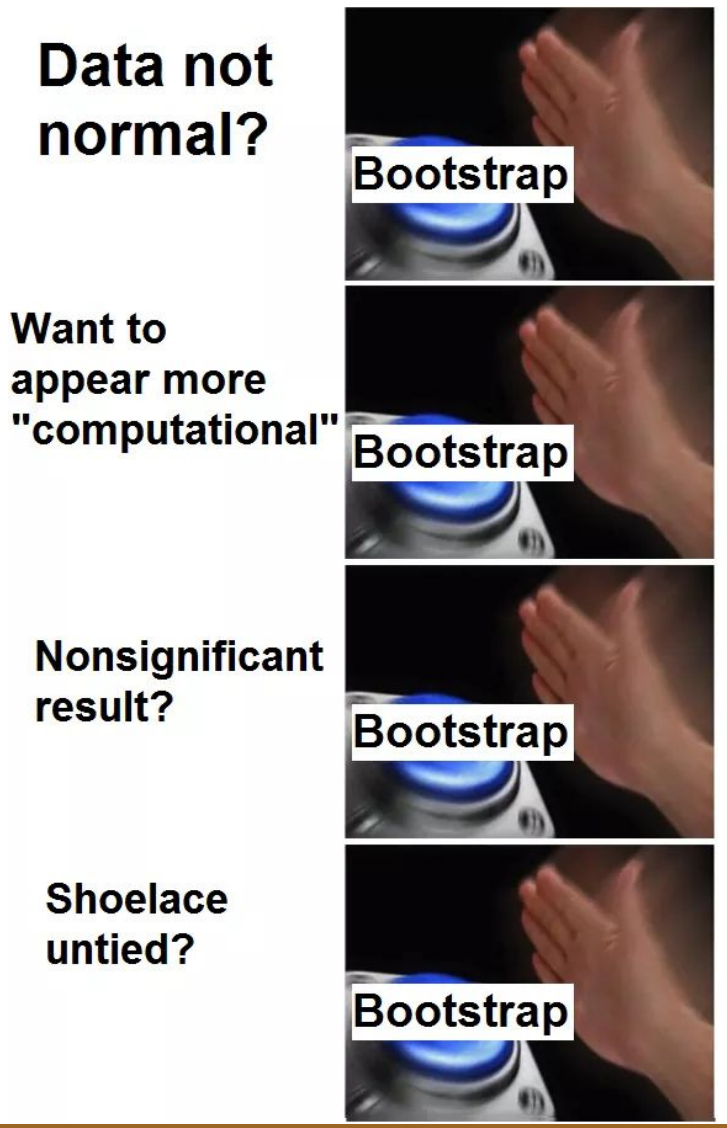
What do you want to **analyse** and **predict**?

- *Classifications*:
  - K-Means
  - Support-Vector Machines
  - Hierarchies
  - Networks

→ When to use which one?

- *Regression*:
  - Linear Models
  - Least Squares vs. Maximum Likelihood
  - Mixed Effect Models
  - GLS/GLM, and GAM

→ How do you select the best model?



Advanced Statistics

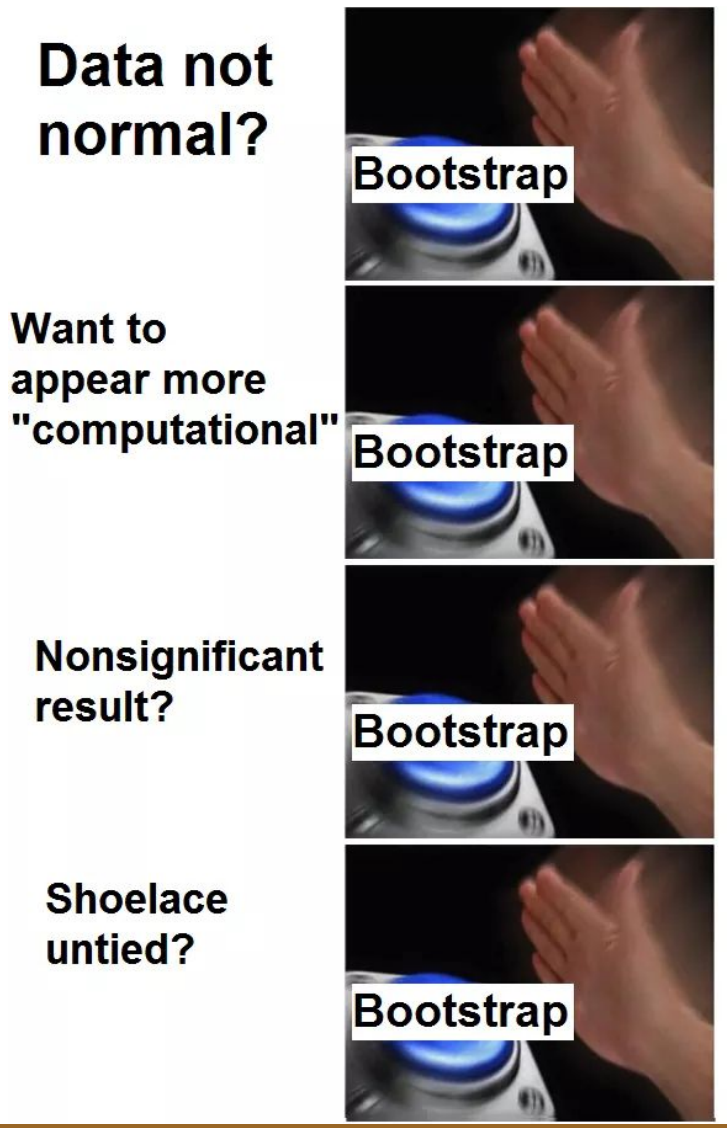
What do you want to **analyse** and **predict**?

- *Classifications*:
  - K-Means
  - Support-Vector Machines
  - Hierarchies
  - Networks

→ When to use which one?

- *Regression*:
  - Linear Models
  - Least Squares vs. Maximum Likelihood
  - Mixed Effect Models
  - GLS/GLM, and GAM

→ How do you select the best model?



Advanced Statistics

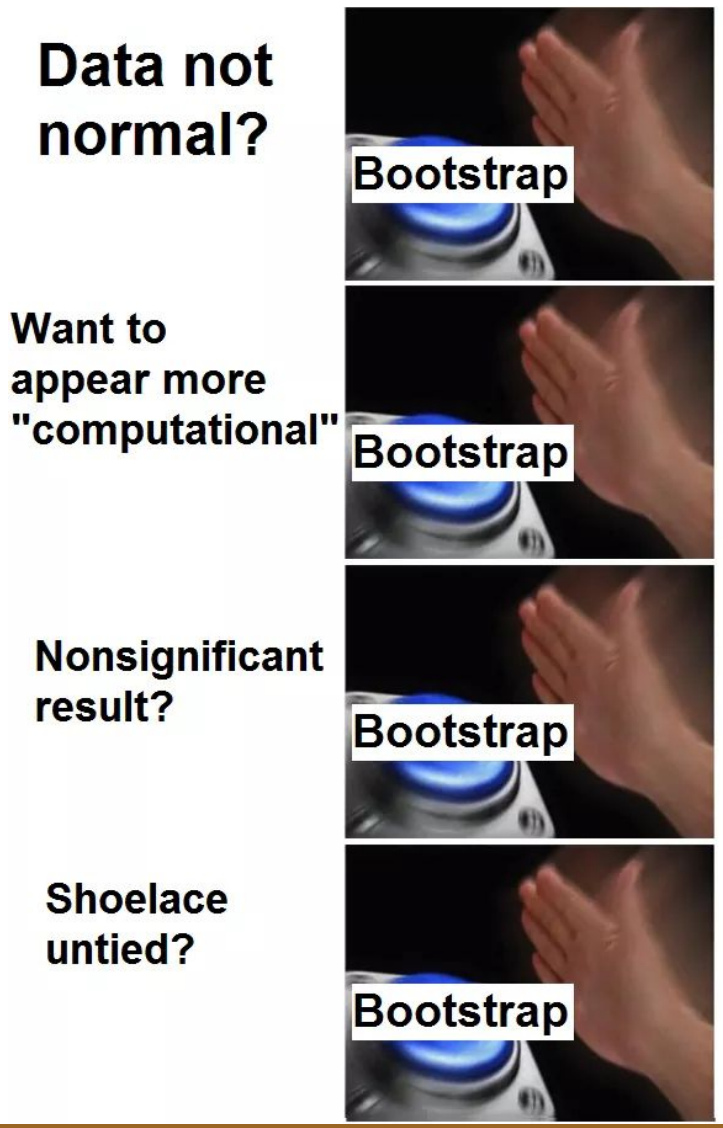
What do you want to **analyse** and **predict**?

- *Classifications*:
  - K-Means
  - Support-Vector Machines
  - Hierarchies
  - Networks

→ When to use which one?

- *Regression*:
  - Linear Models
  - Least Squares vs. Maximum Likelihood
  - Mixed Effect Models
  - GLS/GLM, and GAM

→ How do you select the best model?



Advanced Statistics

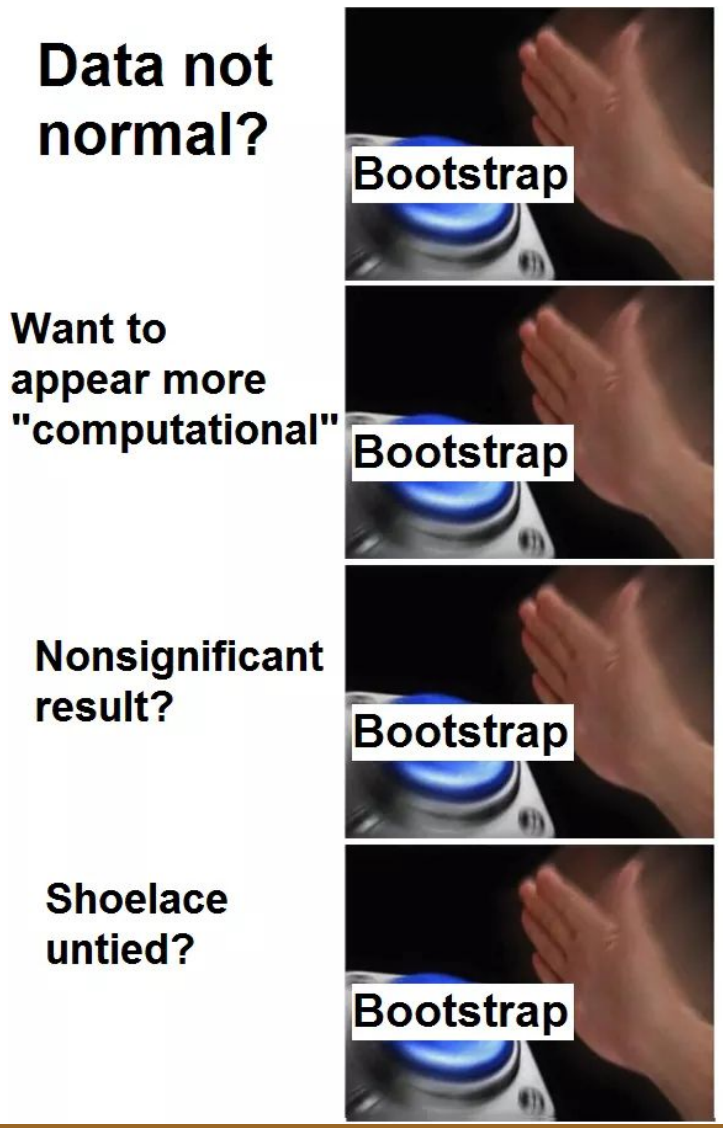
What do you want to **analyse** and **predict**?

- *Classifications*:
  - K-Means
  - Support-Vector Machines
  - Hierarchies
  - Networks

→ When to use which one?

- *Regression*:
  - Linear Models
  - Least Squares vs. Maximum Likelihood
  - Mixed Effect Models
  - GLS/GLM, and GAM

→ How do you select the best model?



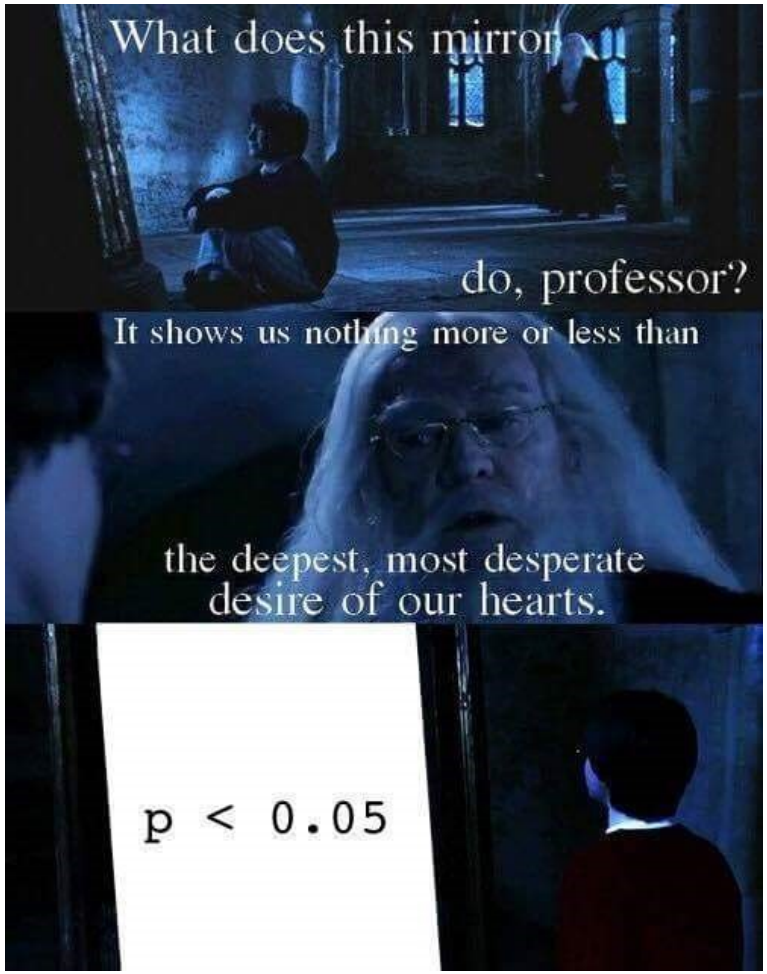
Statistical Significance - the p-value

Misconceptions

- The p-value is not designed to tell us whether something is strictly true or false
- It is not the probability of the null hypothesis being true
- The size of $p \neq$ strength of an observed effect

Alternatives

- Effect Sizes
- Confidence Intervals
- Akaike Information Criterion (AIC)
- Bayes Factor
- Credible Intervals



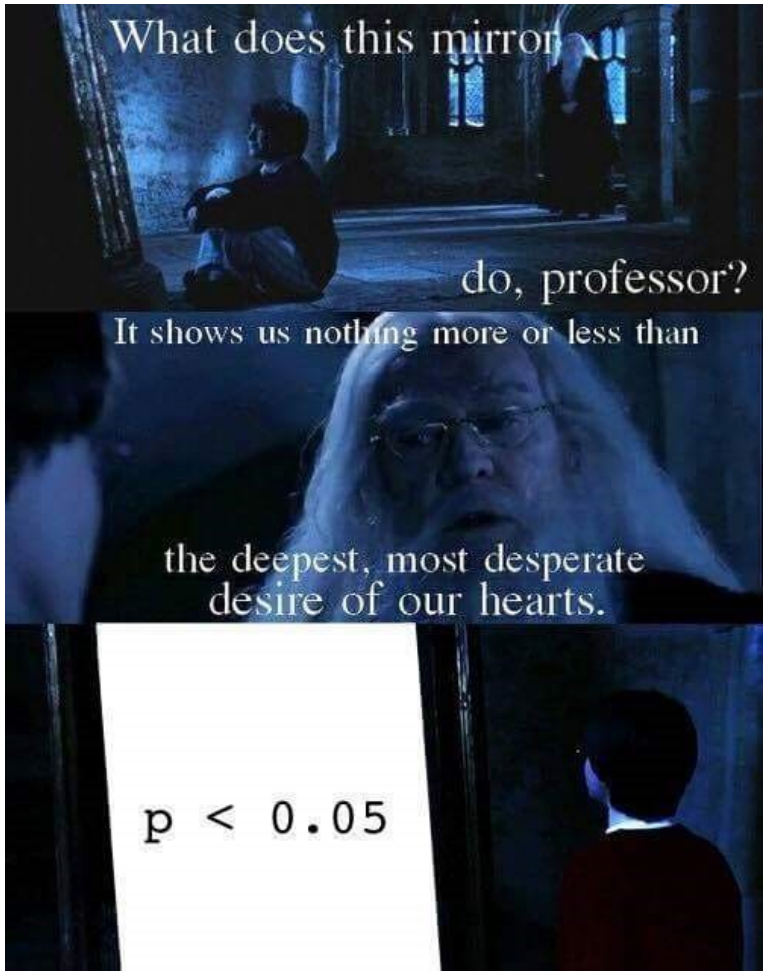
Statistical Significance - the p-value

Misconceptions

- The p-value is not designed to tell us whether something is strictly true or false
- It is not the probability of the null hypothesis being true
- The size of $p \neq$ strength of an observed effect

Alternatives

- Effect Sizes
- Confidence Intervals
- Akaike Information Criterion (AIC)
- Bayes Factor
- Credible Intervals



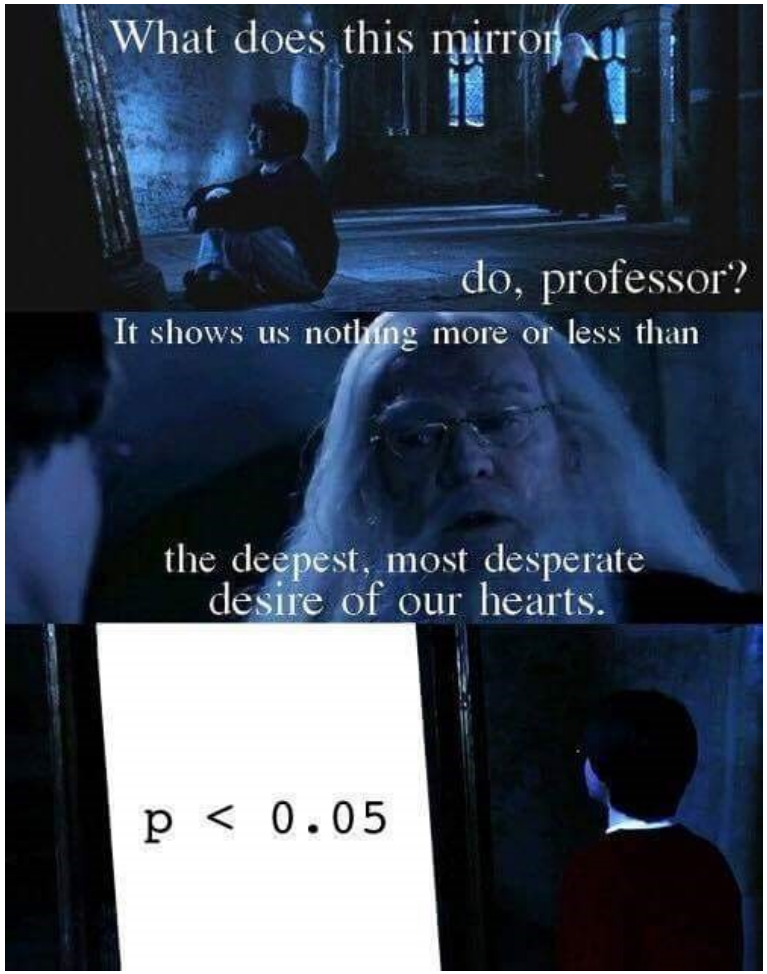
Statistical Significance - the p-value

Misconceptions

- The p-value is not designed to tell us whether something is strictly true or false
- It is not the probability of the null hypothesis being true
- The size of $p \neq$ strength of an observed effect

Alternatives

- Effect Sizes
- Confidence Intervals
- Akaike Information Criterion (AIC)
- Bayes Factor
- Credible Intervals



Coding Etiquette

R Coding

- Object Modes
- Object Types
- Sub-setting
- Vectorisation
- Statements, Loops
- Functions, Packages

Coding Schools

- Hard-coding vs. Soft-coding
- Base `plot` vs. `ggplot2`
- Base `code` vs. `tidyverse`

And what about **Git Hub**?

Coding Etiquette

R Coding

- Object Modes
- Object Types
- Sub-setting
- Vectorisation
- Statements, Loops
- Functions, Packages

Coding Schools

- Hard-coding vs. Soft-coding
- Base `plot` vs. `ggplot2`
- Base `code` vs. `tidyverse`

And what about **Git Hub**?

Coding Etiquette

R Coding

- Object Modes
- Object Types
- Sub-setting
- Vectorisation
- Statements, Loops
- Functions, Packages

Coding Schools

- Hard-coding vs. Soft-coding
- Base `plot` vs. `ggplot2`
- Base `code` vs. `tidyverse`

And what about **Git Hub**?

Coding Etiquette

R Coding

- Object Modes
- Object Types
- Sub-setting
- Vectorisation
- Statements, Loops
- Functions, Packages

Coding Schools

- Hard-coding vs. Soft-coding
- Base `plot` vs. `ggplot2`
- Base `code` vs. `tidyverse`

And what about **Git Hub**?

Manuscript Workflow

Using `Rmarkdown` for your research comes with a multitude of advantages:

1. Entire **workflow in one program** (`RStudio`)
2. **Research** and reports **reproducible** at the click of **one button**
3. **Combines** `R` functionality and LaTeX formatting (if desired)
4. **Consistent formatting**
5. **Clear presentation of code**
6. **Dynamic documents** (you can generate various output document types)
7. Applicable for **almost all document types** you may desire as an output (e.g. manuscripts, presentations, posters, etc.)

Need Statistical Advice?

Erik Kusch

Studies:
PhD @ Aarhus University (currently enrolled)
M.Sc. @ University of Bergen
B.Sc. @ Technical University of Dresden

Experience:
Biostatistics Tutor @ University of Leipzig
Biostatistics Research Assistant @ University of Leipzig
Biostatistics Research Assistant @ University of Kyoto

Research:
- Dryland vegetation memory analyses
- Large-scale vegetation-climate modelling
- Remote sensing approaches in macroecology
- Biostatistical approaches in behavioural ecology
- Statistical downscaling of climate reanalysis data for use in biological analyses

Find me in room 318, building 1540 (Fridays, 09.00-12.00) or via erik.kusch@bio.au.dk.

Need Statistical Advice?

Erik Kusch

Studies:
PhD @ Aarhus University (currently enrolled)
M.Sc. @ University of Bergen
B.Sc. @ Technical University of Dresden

Experience:
Biostatistics Tutor @ University of Leipzig
Biostatistics Research Assistant @ University of Leipzig
Biostatistics Research Assistant @ University of Kyoto

Research:
- Dryland vegetation memory analyses
- Large-scale vegetation-climate modelling
- Remote sensing approaches in macroecology
- Biostatistical approaches in behavioural ecology
- Statistical downscaling of climate reanalysis data for use in biological analyses

Find me in room 318, building 1540 (Fridays, 09.00-12.00) or via erik.kusch@bio.au.dk.