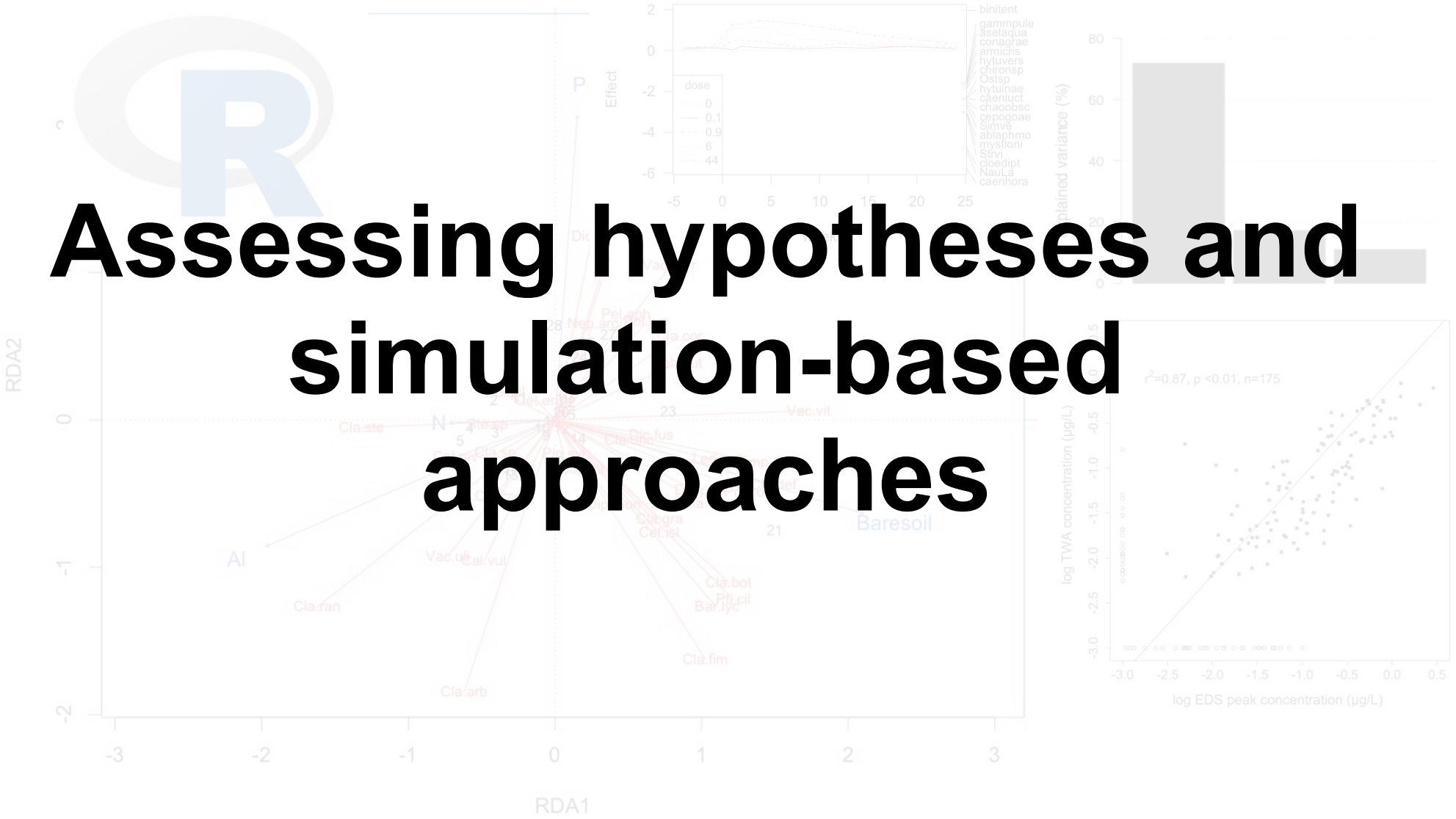
Tools for complex data analysis

University of Koblenz-Landau 2018/19



Assessing hypotheses and simulation-based approaches

Ralf B. Schäfer

Learning targets

- Understand and evaluate the concept of the *p*-value and of assessing hypotheses
- Explain and apply simulation-based approaches to data analysis

Learning targets and study questions

- Understand and evaluate the concept of the *p*-value and of assessing hypotheses
  - Define the *p*-value and explain the rationale for its use.
  - Describe the differences between the different approaches to assess hypotheses.
  - Discuss pros and cons of significance testing.
  - Outline the good practice when assessing hypotheses.
  - Distinguish scientific and statistical hypotheses.
  - What are the assumptions of the *t*-test?

Learning targets and study questions

- Explain and apply simulation-based approaches to data analysis
  - Discuss how simulation-based approaches link the two cultures to data analysis.
  - Explain the purpose and critically discuss permutation tests.
  - Explain the purpose and critically discuss bootstrapping.
  - Explain the main idea of cross-validation and discuss the selection of *k* with respect to the bias-variance trade-off.
  - Discuss the application of bootstrapping and cross-validation in regression analysis.

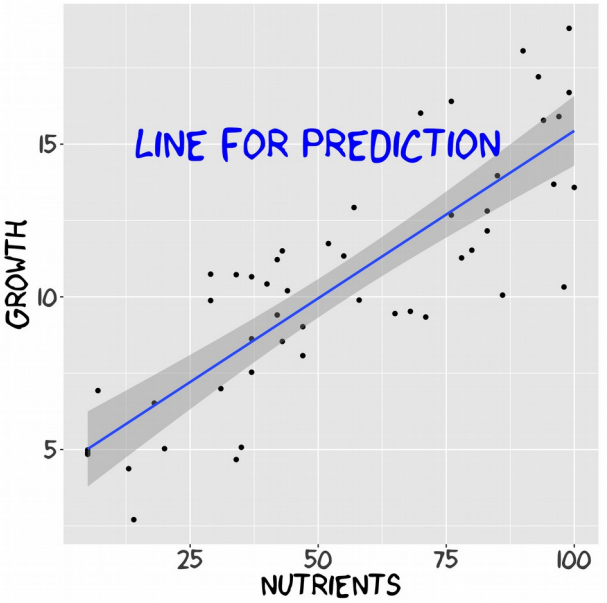
Assessing hypotheses and simulation-based approaches

Contents

Recap: Research goals and model output

1. Prediction
2. (Parameter) estimation
3. Assessing hypotheses

Example: Assessing hypotheses related to the relationship between plant growth and nutrient concentrations.



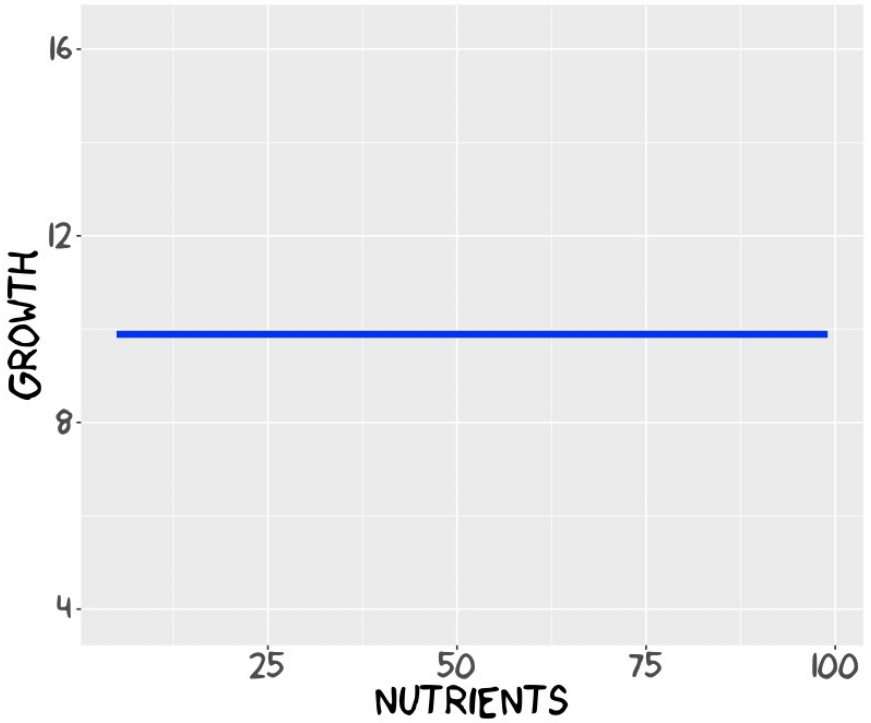
4. Explanation

?

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 137.79468   31.92061   4.317 1.95e-05 ***
## X             1.45722    0.03152  46.231  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103 on 448 degrees of freedom
## Multiple R-squared:  0.8267, Adjusted R-squared:  0.8263
## F-statistic:  2137 on 1 and 448 DF,  p-value: < 2.2e-16
```

Case study: Nutrients and plant growth

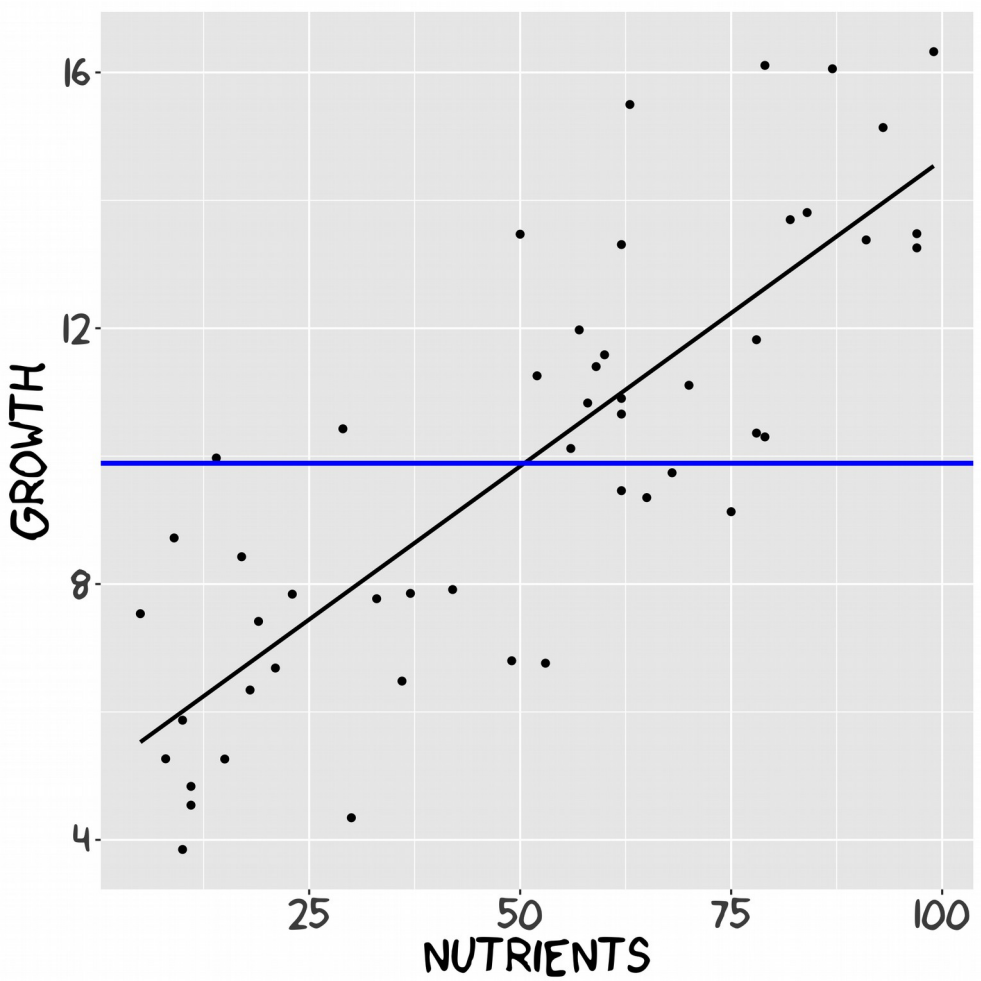
A skeptic claims that some nutrients have no effect on plant growth. This translates to the hypothesis of a slope of 0 in a regression model, i.e. $\beta_1 = 0$.



Our research goal is to produce evidence against this claim, which we define as the so-called *null hypothesis* H_0: $\beta_1 = 0$.

Case study: Nutrients and plant growth

We conduct an experiment and obtain an estimate of 0.1 for the slope, i.e. $\hat{\beta}_1 = b_1 = 0.1$:



We need a statistic that informs on the conformity of our data with H_0.

Recap: Confidence intervals

Can't we use confidence intervals (CIs)? They give upper and lower limits for the true parameter.

General calculation of confidence interval for parameter θ:

$\theta \pm$ value related to confidence level and probability distribution × SE

(e.g. 95%, 99%)

Known or assumed probability distribution of parameter

95% CI for slope:

$b_1 = 0.1$, 95% CI[0.07, 0.12]

$b_1 = 0$ not included $\rightarrow$ If assumptions are met (e.g. probability distribution) and the specific CI is one of those containing the true parameter, then 0 is not a potential true parameter.

Don't we have a more direct way to assess the hypothesis?

The concept of *p*-value

Is the probability p of obtaining such or more extreme data if H_0 is true: $p = 2 \times P(T \geq |t|; H_0)$ where T is a test statistic with the realised value t.

<u>Application to our case study:</u>

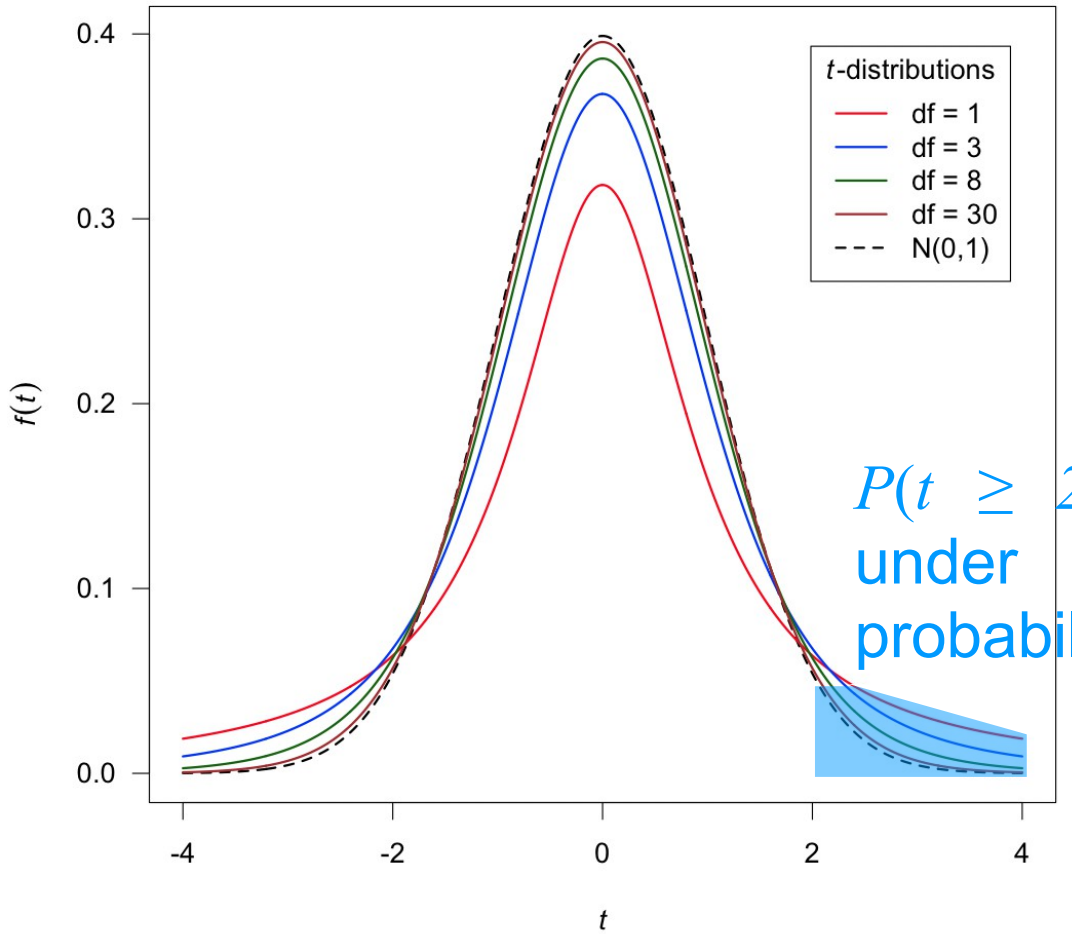
Test statistic is the so-called t-value related to the t-test for regression coefficients:

$$t = \frac{\hat{\beta} - \beta}{\mathrm{SE}_{\hat{\beta}}} \qquad \text{from } H_0\text{: } \beta_1 = 0 \text{ follows:} \qquad t = \frac{\hat{\beta}}{\mathrm{SE}_{\hat{\beta}}}$$

We obtain a value for t = 9.79 → Determine $P(t \geq 9.79; H_0)$ from a so-called *Student's t* distribution.

The *Student's t*-distribution

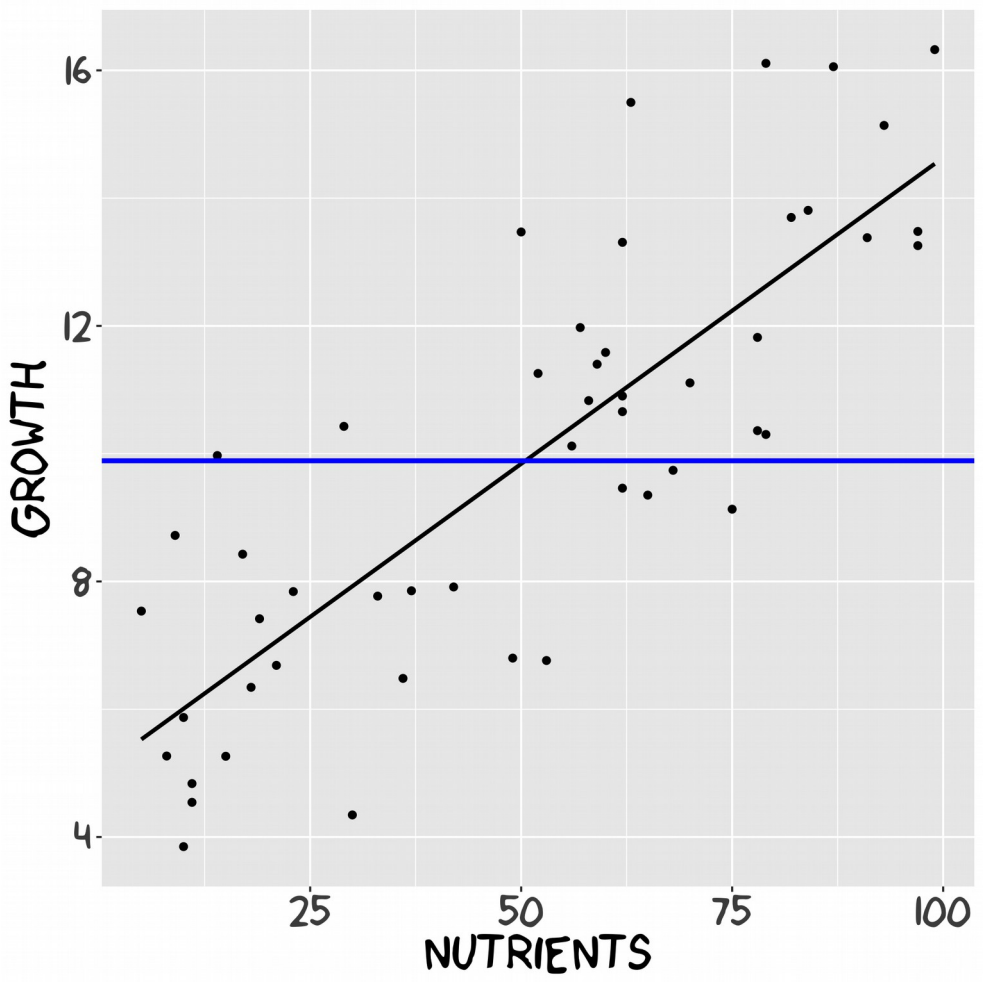
- Symmetric and continuous probability distribution
- Arises when estimating the mean of a normal distribution at small n and unknown σ
- Approximates normal distribution for ≥ 30 degrees of freedom (df)



$P(t \geq 2)$ is the area under the respective probability distribution

Case study: Nutrients and plant growth

p-value for our data: $5 \times 10^{-13} \rightarrow$ probability of obtaining these or more extreme data, if H_0 is true (i.e. no relationship between nutrients and growth), is almost zero.



Assessing hypotheses and simulation-based approaches

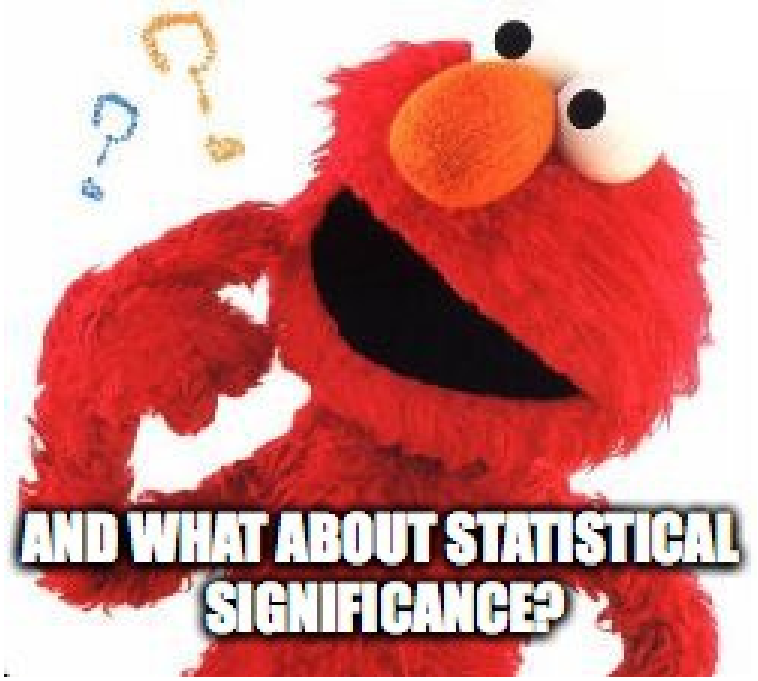
Contents

Interpretation of *p*-values

Provide the degree to which data are compatible with the pattern predicted by a given null hypothesis, conditional that the assumptions of the underlying statistical model are met. (*cf.* Greenland et al. 2016)

→ Can be used to assess the plausibility of H_0

→ This interpretation follows the NeoFisherian approach (*cf.* Hurlbert & Lombardi 2009)



What about statistical significance?

NeoFisherian approach:

- *p*-value is graded measure of evidence against H_0 → Report exact *p*-values
- Possible conclusions from test statistic and related *p*-value:
    1. Effect (e.g. difference, association) seems negative
    2. Effect (e.g. difference, association) seems positive
    3. Neither can be confidently stated, judgement reserved or suspended

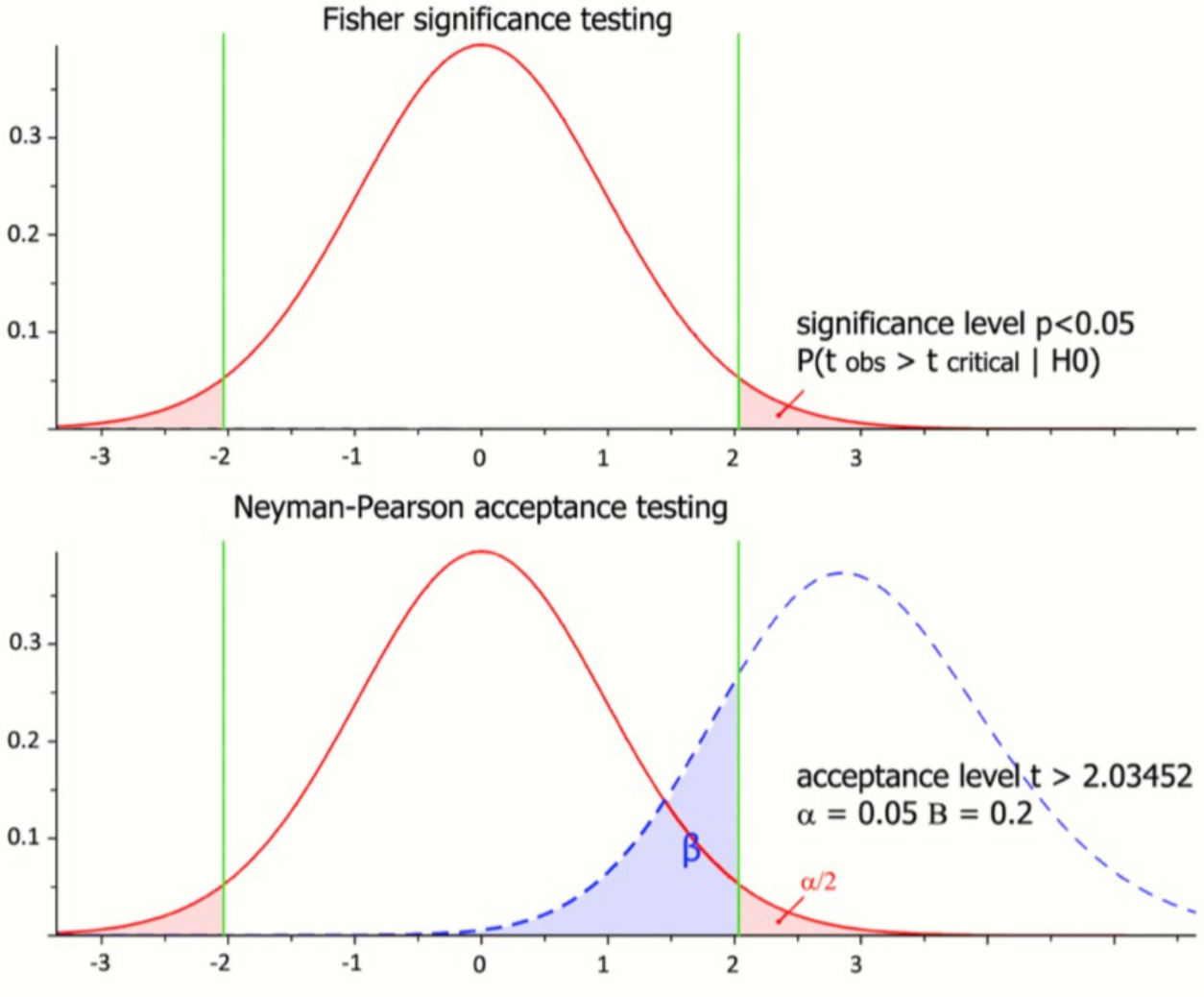
→ Discourages 1) Use of the term "statistical significance" and 2) to assign special status to $p < 0.05$

→ Generally: Evaluate effect size and overall evidence

Hurlbert & Lombardi 2009

Statistical significance: Classical approaches

- Two approaches: Fisherian and Neyman-Pearson (NP)
- Fisher: testing of significance of single (null) hypothesis H_0, NP: accept alternative hypothesis if H_0 can be rejected



Both approaches employ fixed cut-off points and typically interpret $p < 0.05$ as statistically significant

Pernet 2017 *F1000 Research* 4

Problems of significance testing and misinterpretations of *p*-values

- Significance testing dichotomises the *p*-value scale into significant (usually $p < 0.05$) and non-significant. This logic has many flaws:
  - Why is a result with $p = 0.055$ less relevant than $p = 0.045$?
  - *p*-value is sensitive to sample size (increasing sample size *ceterus paribus* will lead to significance)
  - *p*-value has low replicability (see Halsey et al. 2015) → leads to seemingly contradictory (significant and non-significant) results in replicated studies
  - Focus on significant results leads to distortion in scientific literature (ignorance of potentially relevant results)

Problems of significance testing and misinterpretations of *p*-values

- Statistical significance does not imply scientific significance (and vice versa)
- High *p*-value (e.g. above a fixed significance threshold) does not mean that H_0 is true! (see Hurlbert & Lombardi 2009: 321-323 for example)
- Very low *p*-value does not mean that H_0 is incorrect (or H_A is true in NP approach)
- *p*-values do not inform on effect size

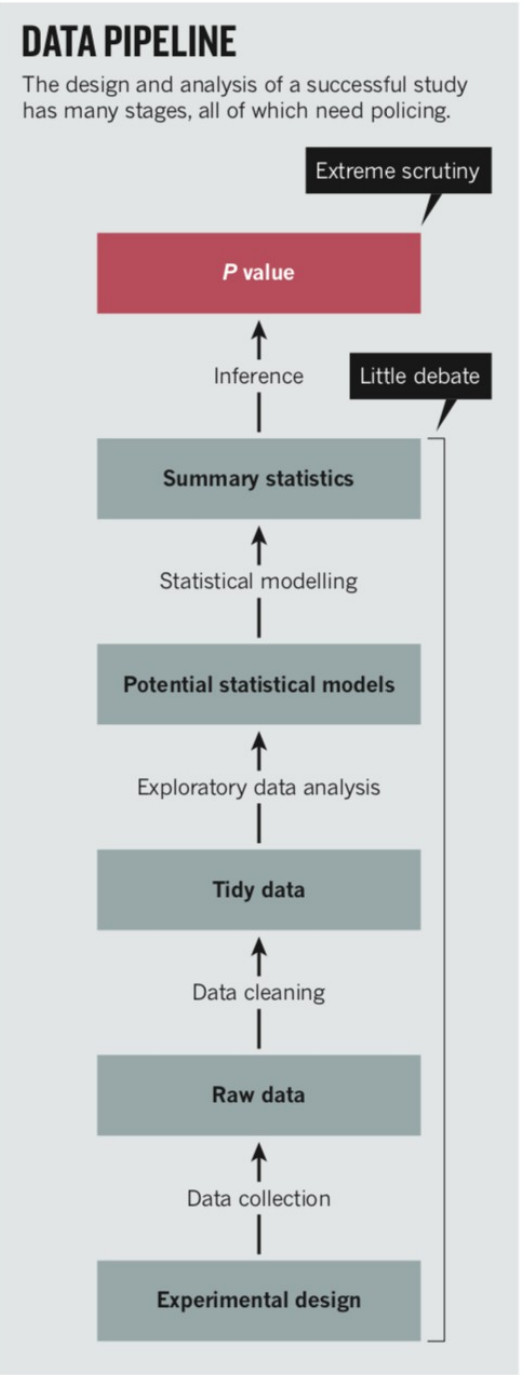
Debate on *p*-values and testing

- Debate since almost 100 years, mainly in statistical community
- Less attention in algorithm-based (machine learning) community
- Relevance of issue may be overstated and alternatives have similar issues

"*Whatever method we use in threshold tests and decision heuristics* […] *we create biases and invalidate the answers we give to our questions*" Arnheim et al. 2017

"[…] *look for a magic alternative to NHST [null hypothesis significance testing], some other objective mechanical ritual to replace it. It doesn't exist*" Cohen 1994

"*I thought greater use of Bayesian methods would reduce that misuse [RBS: of statistics]. Now I am less convinced. And the debates seem to distract from more important issues.*" McCarthy 2015



Leek & Peng 2015

Good practice in assessing hypotheses

- Report effect sizes
- If calculated, report exact *p*-values (for any approach)
- If following NP approach: report power
- Consider journal requirements (example: *Nature*)

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☐	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☐	A description of all covariates tested
☐	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☐	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

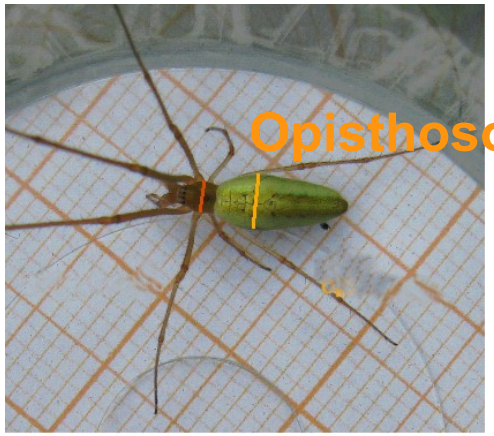
https://www.nature.com/authors/policies/availability.html

Assessing hypotheses and simulation-based approaches

Contents

Case study: Pesticides and spiders

Research question: Does environmental pesticide exposure reduce the body size of spiders?



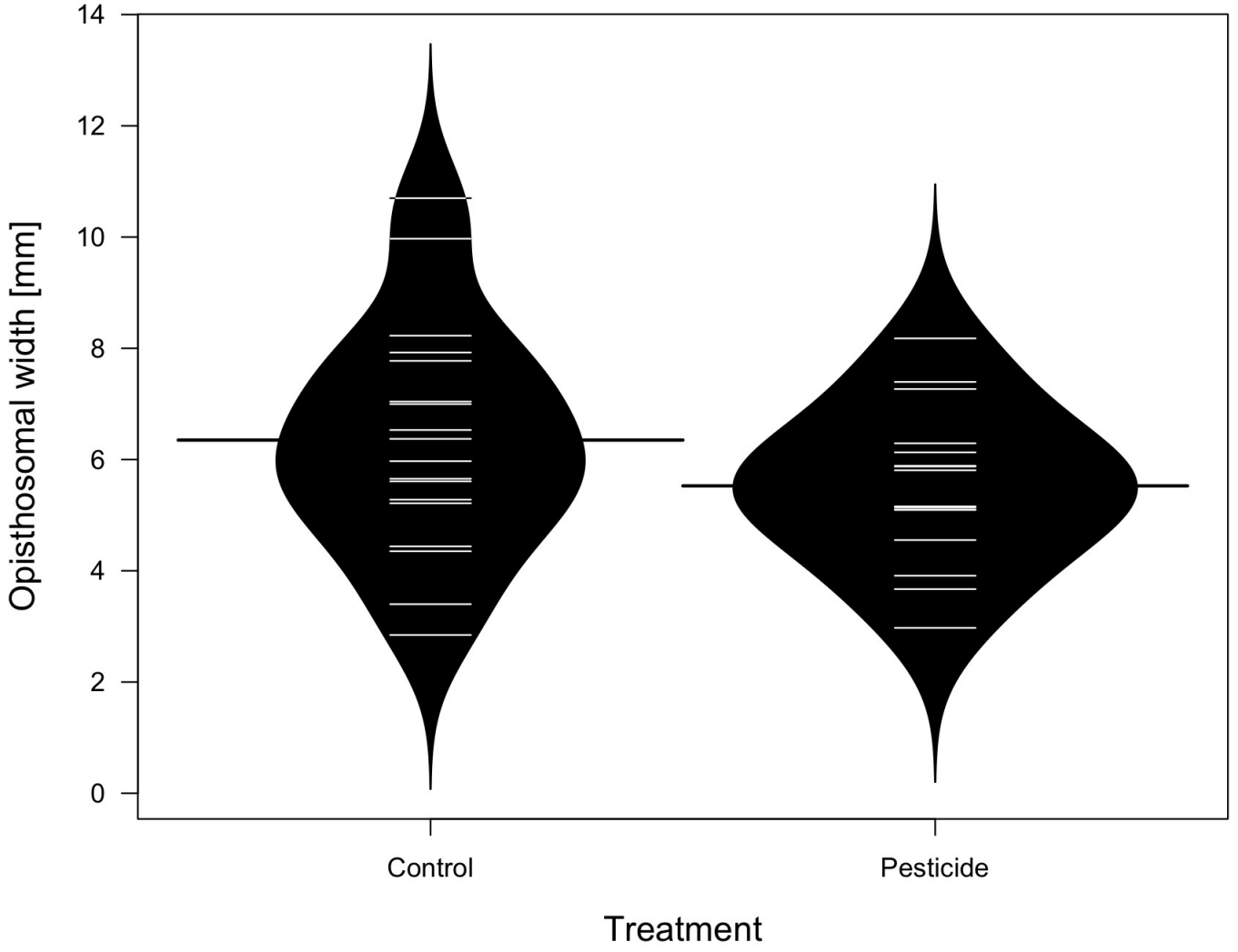
Scientific hypothesis: The concentration of pesticides that are taken up in the environment by spiders require the activation of energetically costly detoxication processes. This reduces the energy for growth and consequently the body size.

Laboratory experiment: Treatment of a group of spiders "*a*" with a typical environmental exposure concentration and measurement of a body size metric (i.e. opisthosomal width) and comparison to a control group "*b*" kept under the same conditions, except for pesticide exposure.

Case study: Pesticides and spiders

<u>Statistical null hypothesis:</u> The population mean of opisthosomal width μ is equal for the groups a and $b \rightarrow$ H_0: $\mu_a = \mu_b$

$\rightarrow$ Do the experimental data provide evidence against H_0?



Comparing two means with the t-test

Recall: p-value is the probability of obtaining such or more extreme data if H_0 is true, with $p = 2 \times P(T \geq |t|; H_0)$ where T is a test statistic with the realised value t.

→ Test statistic for comparison of two means is provided by the **two-sample *t*-test** (recall the one-sample t-test for regression coefficients)

Null hypothesis: The samples have been drawn from populations with equal μ. → $H_0: \mu_1 = \mu_2$

Calculation of the realised value t:

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_{x_1 x_2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{with} \quad s_{x_1 x_2} = \sqrt{\frac{(n_1 - 1) s_{x_1}^2 + (n_2 - 1) s_{x_2}^2}{n_1 + n_2 - 2}}$$

„Difference of means divided by standard error of difference“

Assumptions of the t-test

- Independent samples from the population
  - Both populations have been randomly sampled
  - Evaluation of assumption requires knowledge on sampling
- Normal distribution
  - Both populations that have been sampled (e.g. X, Y) should follow a normal distribution: $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$
  - Evaluation of assumption through graphical inspection (*QQ*-plot)
- Variance homogeneity
  - The variances of both populations that have been sampled (e.g. X, Y) are equal: $\sigma_X^2 = \sigma_Y^2$
  - Evaluation of assumption through graphical inspection (Conditional boxplot, beanplot)

Assessing hypotheses and simulation-based approaches

Contents

1. Assessing hypotheses: The concept of *p*-value
2. Interpretation of *p*-values and statistical significance
3. Example for a hypothesis test: *t*-test
4. **Permutation test**
5. Bootstrapping
6. Cross-Validation and Bias-variance trade-off

Simulation-based approaches in data analysis

- Compatible with both data modelling (classical statistics) and algorithmic modelling (machine learning) cultures
- Infuses algorithm-based thinking into classical statistics
- Examples for simulation-based approaches for estimation, inference or model diagnosis in classical statistics:

1. **Permutation test** → Permuting (shuffling) the data to derive null distribution. Mainly used for inference
2. **Bootstrapping** → Randomly sampling subsets from the data with replacement. Mainly used for estimation of parameter distribution
3. **Cross-validation** (CV) → Splitting data into sets (i.e. sampling without replacement). Mainly used for validation of predictive models

Permutation test: Algorithm

Repeat k times

1) Permute values in data set
2) Compute test statistic t^* for permuted data

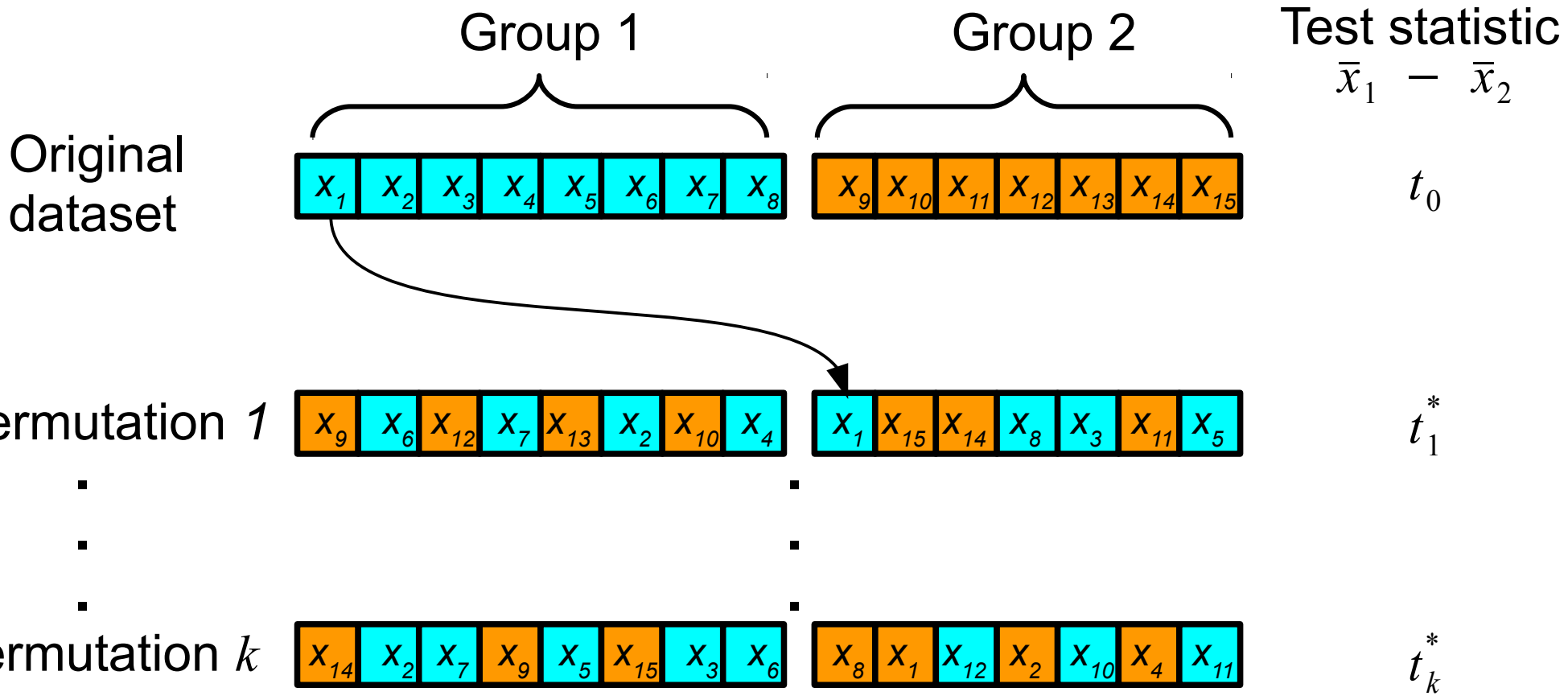
3) Compare test statistic t_0 to generated null distribution

Permutation test: Algorithm

Repeat k times:

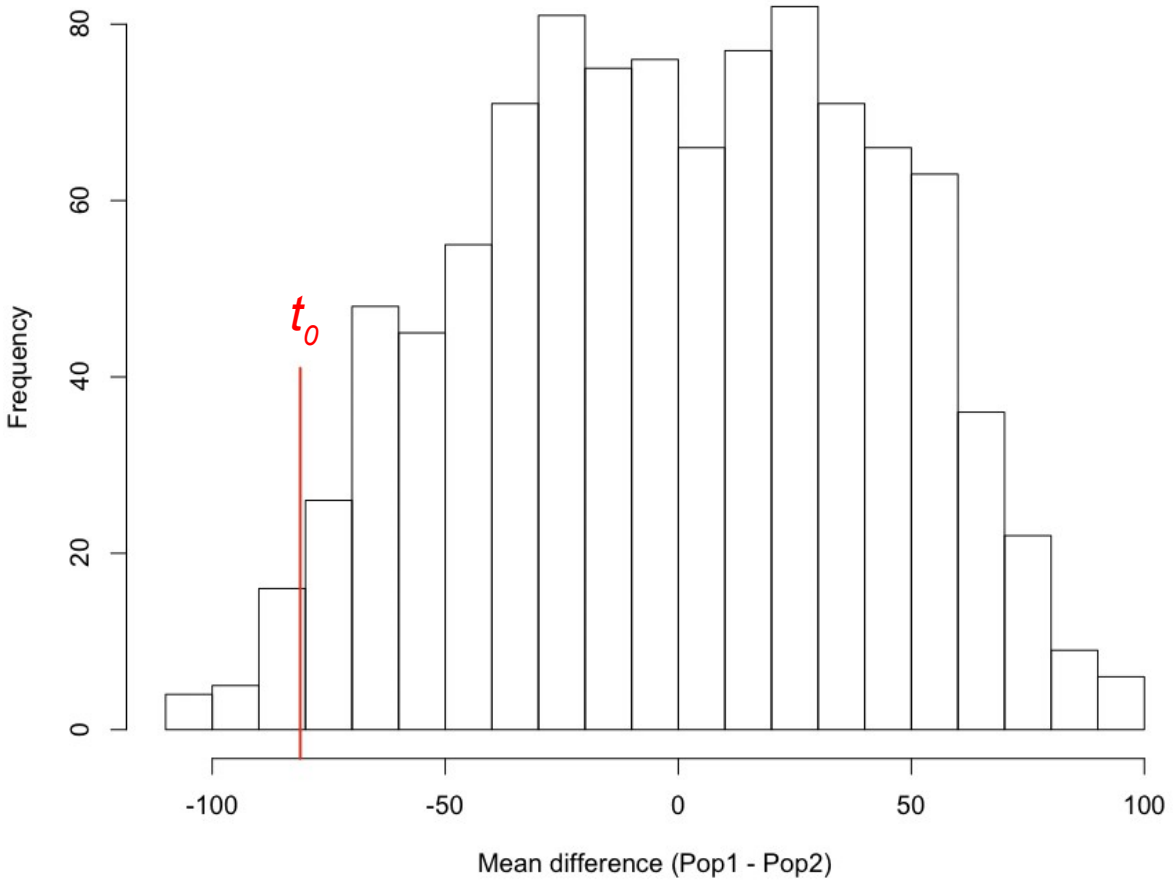
1) Permute values in data set
2) Compute test statistic t^* for permuted data
3) Compare test statistic t_0 to generated null distribution

<u>Example:</u> Permutation test for comparing group means



Permutation test: Generated distribution

$$p = \frac{\sum_{i=1}^{k} 1 \text{ if } t_i^* \leq t_0, else\, 0}{k+1}$$



- p gives probability that pattern is produced by pure chance
- Inference regarding statistical population only valid if distribution of sample data matches actual distribution of statistical population → particularly problematic for small n

Permutation test: Advantages and limitations

- Advantages
  - Applicable to any distribution
  - Applicable to complex designs through restricting permutations
- Limitations
  - Generalisation to statistical population requires matching distribution
  - Assessing statistical hypotheses can imply distributional assumptions that also apply to the permutation test, if aiming to infer to the statistical population (e.g. comparing means is affected by unequal variance)
  - Computationally intensive: Number of all possible permutations for a dataset is factorial *n*, i.e. *n*! (e.g. $35! \approx 10^{40}$) → Monte Carlo simulation

Legendre & Legendre 2012: 25ff

Monte-Carlo simulation

- Uses repeated random sampling to solve problems probabilistically (even though they can be deterministic in reality)
- Permutation tests use random numbers to randomly permute data → approximate with MC simulation
- Legendre & Legendre (2012): use at least 10,000 permutations for inference

Entrance of casino in Monte Carlo, Monaco

Edvard Munch - At the Roulette Table in Monte Carlo

Assessing hypotheses and simulation-based approaches

Contents

1. Assessing hypotheses: The concept of *p*-value
2. Interpretation of *p*-values and statistical significance
3. Example for a hypothesis test: *t*-test
4. Permutation test
5. **Bootstrapping**
6. Cross-Validation and Bias-variance trade-off

Bootstrapping: Idea and algorithm

- Inference on statistic *t* is based on sampling distribution
  - Ideally: Draw all or many samples from statistical population
  - Reality: Most frequently only one sample available
  - → **Idea:** Draw samples from an estimate of the statistical population (i.e. the sample) and use these to estimate a property (e.g. variance) of the statistic *t*
- Algorithm:

  1) Draw random sample with replacement from data
  2) Compute statistic *t** for bootstrap sample

  } Repeat *k* times

  3) Use the *k* estimates to derive property of statistic
- Exhaustive bootstrapping ($k = n^n$) computationally demanding → approximate with Monte Carlo simulation
- With current computer power 10^4-10^5 simulations often viable

Bootstrapping: Example

Example: Bootstrap to the mean (to derive variance)

t (here: mean)

Original dataset

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

$\bar{x} = 8$

Sampling with replacement

BS sample 1

15	7	8	4	15	1	11	9	1	3	6	14	2	11	12

$\bar{x}^* = 7.93$

BS sample 2

6	13	2	10	12	5	7	10	1	13	8	8	15	3	10

$\bar{x}^* = 8.2$

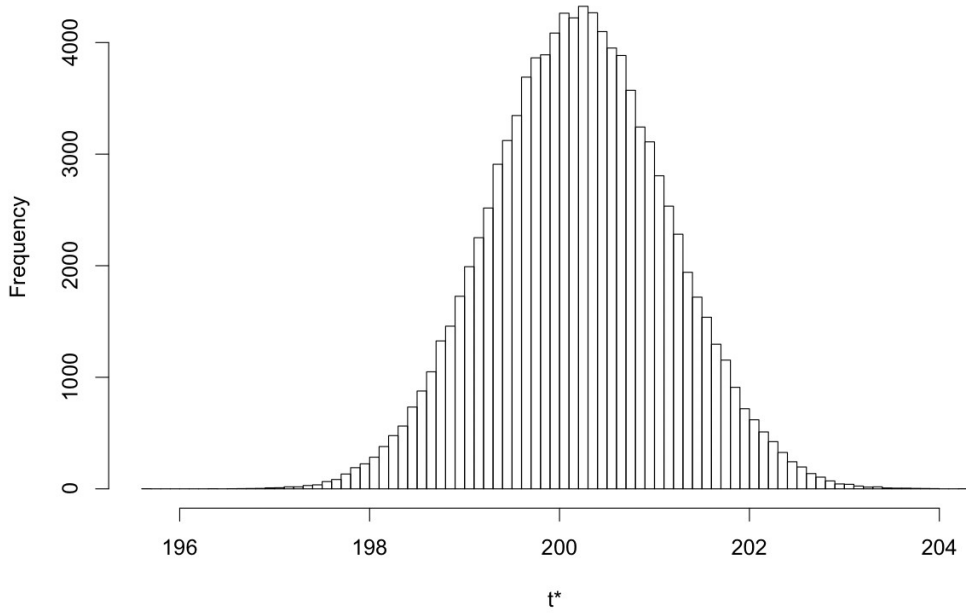
⋮

BS sample *k*

12	7	7	15	5	10	8	13	6	11	8	1	14	2	12

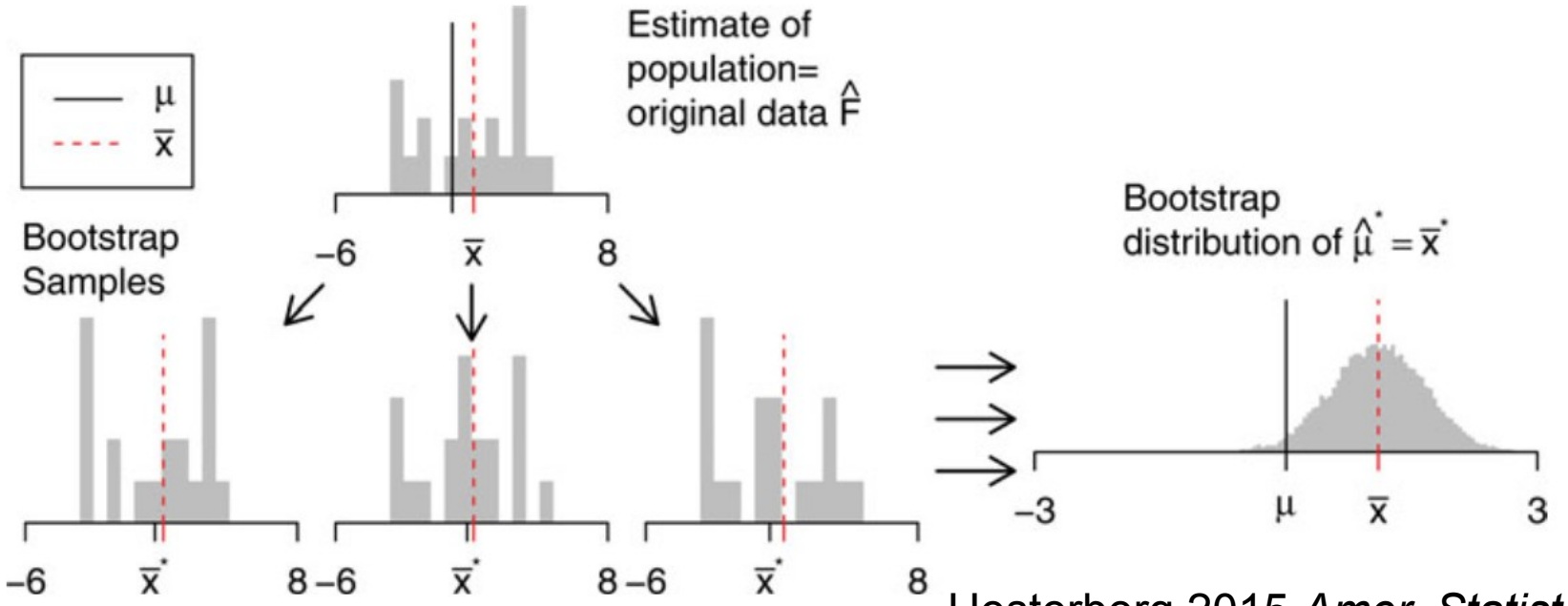
$\bar{x}^* = 8.73$

Distribution of statistic *t*



Bootstrapping: Limitations

- Do not use for assessing hypotheses
- No distributional assumptions implied, but not reliable for all distributions, particularly at small *n* (see Hesterberg 2015)
- Small *n*: use adjusted bootstrap percentiles (Bca) or switch to parametric statistics (allow for additional assumptions)
- Bootstrap does not improve estimate of population parameter μ, centred at $\bar{x}$



Hesterberg 2015 *Amer. Statist.* 69:371

Bootstrapping in regression analysis

Recall the residual definition e_i as: $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$

- Of residuals: Bootstrapping residuals, add to $\hat{y}$ to generate new y^* and calculate regression coefficients $\rightarrow$ x fixed
- Of cases: Bootstrapping complete cases and calculate regression coefficients $\rightarrow$ x random
- If x and y random sample (e.g. x not fixed in experiment), residuals correlated or exhibit non-constant variance $\rightarrow$ Bootstrapping cases

Assessing hypotheses and simulation-based approaches

Contents

1. Assessing hypotheses: The concept of *p*-value
2. Interpretation of *p*-values and statistical significance
3. Example for a hypothesis test: *t*-test
4. Permutation test
5. Bootstrapping
6. **Cross-Validation and Bias-variance trade-off**

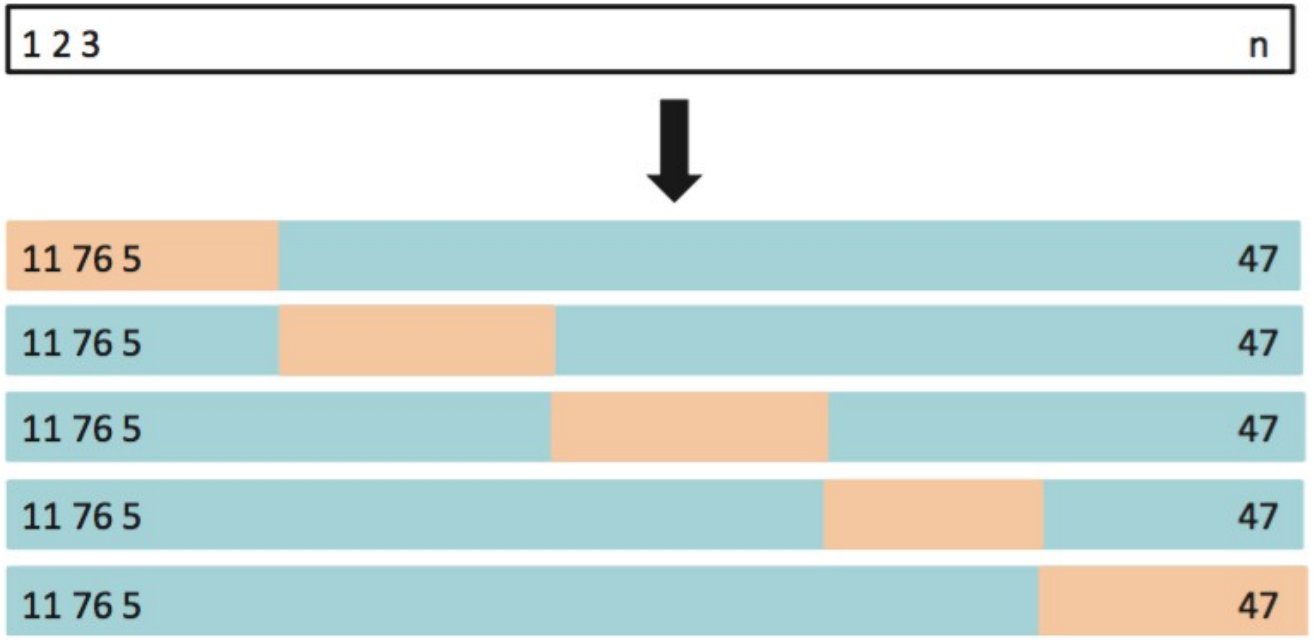
Cross-validation (CV)

- **Aim:** Evaluate predictive accuracy of a fitted model
- Can be checked by predicting (known) responses from independent data sets (that were not used in model fitting) → Rare case
- **Idea:** Split the available data into training and test set and predict (known) observations in test set with a model fitted on the training data
- Algorithm:

1. Draw *k* random samples without replacement from data
2. For each *k:*
    1. Fit the model to the other *k*-1 parts
    2. Predict *k* from model and calculate the prediction error
3. Calculate mean prediction error over the *k* estimates

Cross-validation (CV)

Example: $k = 5$



- Problem of choosing *k*:
  - *k* = n (Leave-one-out CV predicts each observation from all others) → low bias, but high variance
  - *k* = 2 (split data into half) → low variance, but high bias
- *k* typically set to 5 or 10

Taken from James *et al.* 2013: 181

Bias-variance trade-off

Definition in context of model validation:

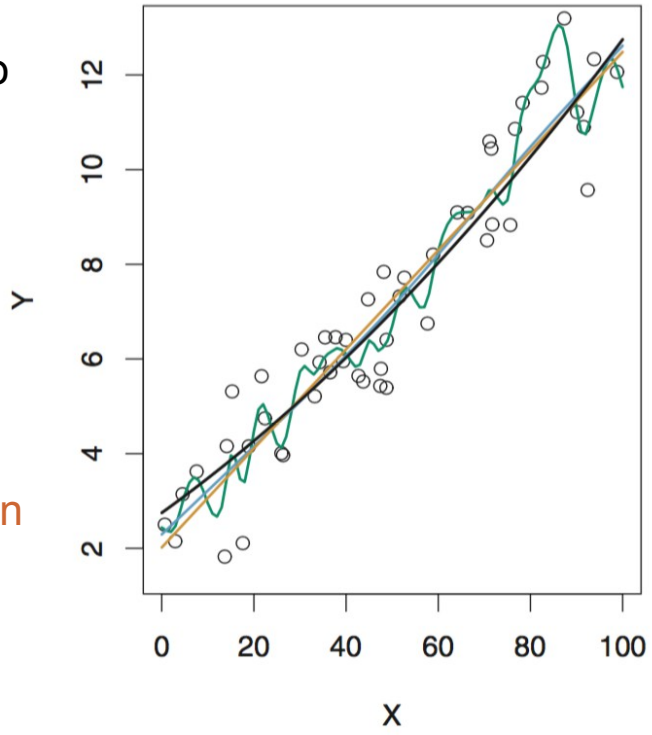
- **Bias:** error when approximating training data
- **Variance:** variability in error when approximating test data

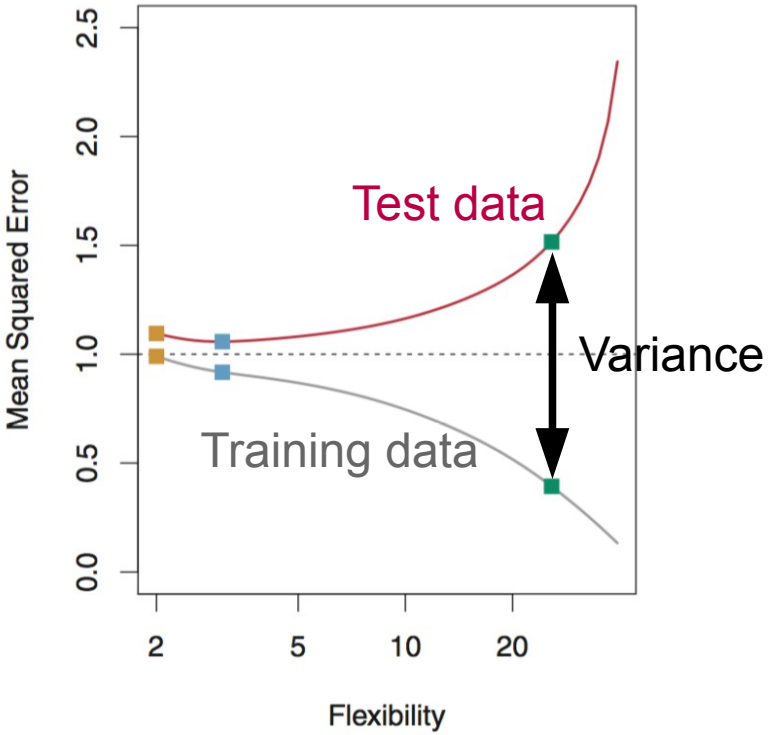
function used to simulate data

highly flexible smoother

little flexible smoother

linear regression



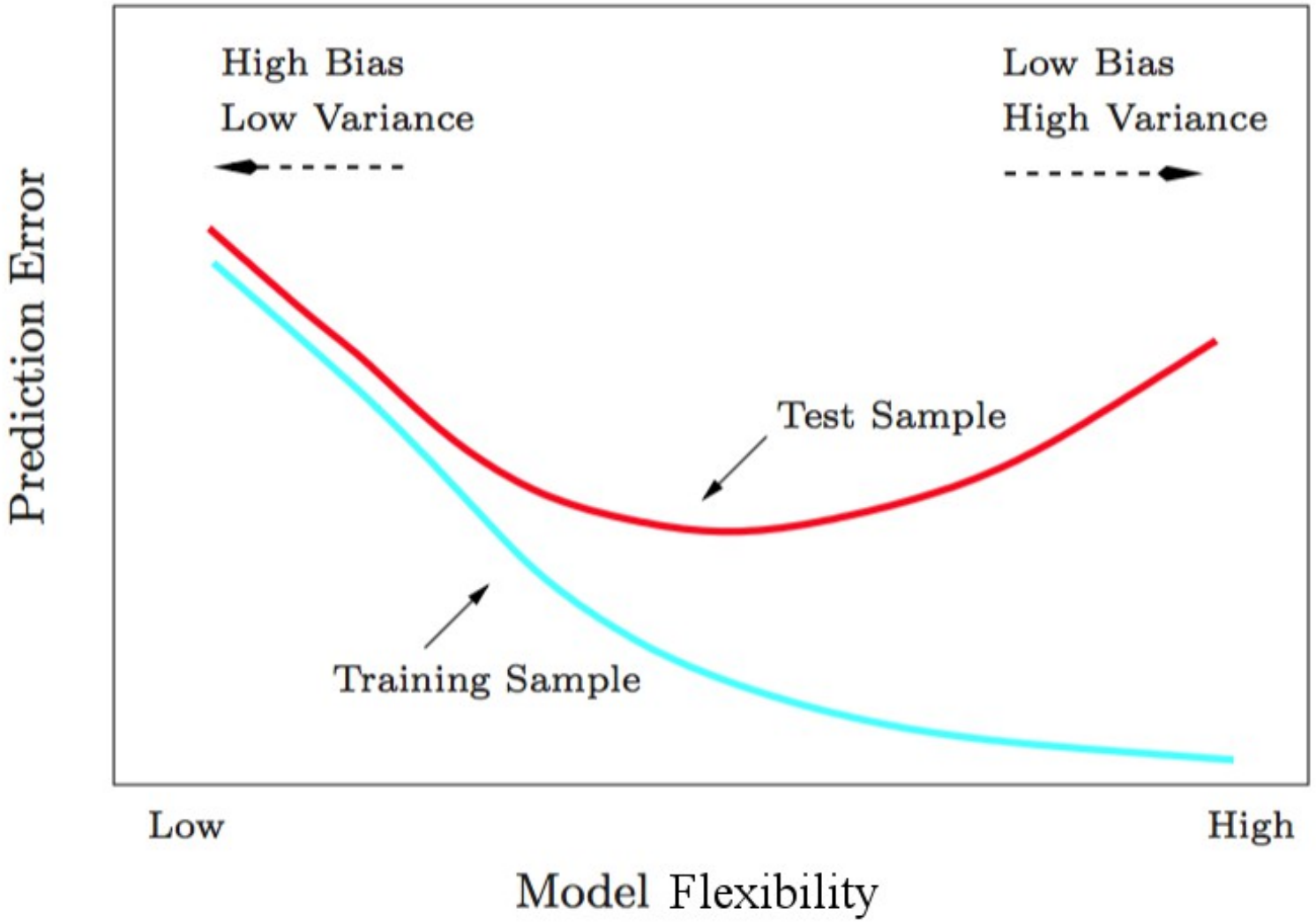


Higher flexibility (higher *k* in CV) → lower error for training data (i.e. lower bias), but variance will start to increase from some point

Taken from James *et al.* 2013: 33

Bias-variance trade-off

Higher flexibility (higher *k* in CV) → lower error for training data (i.e. lower bias), but variance will start to increase from some point → Optimise combined error



Taken from Hastie, Tibshirani and Friedman 2011: 38

CV in regression analysis

- Predictive accuracy measured with estimate of Mean square prediction error (MSPE):

$$\widehat{\mathrm{MSPE}} = \frac{1}{m}\sum_{i=1}^{m}\left(y_{\mathrm{new},i} - \hat{y}_i\right)^2$$ for new observations (y_{new}) 1 to m

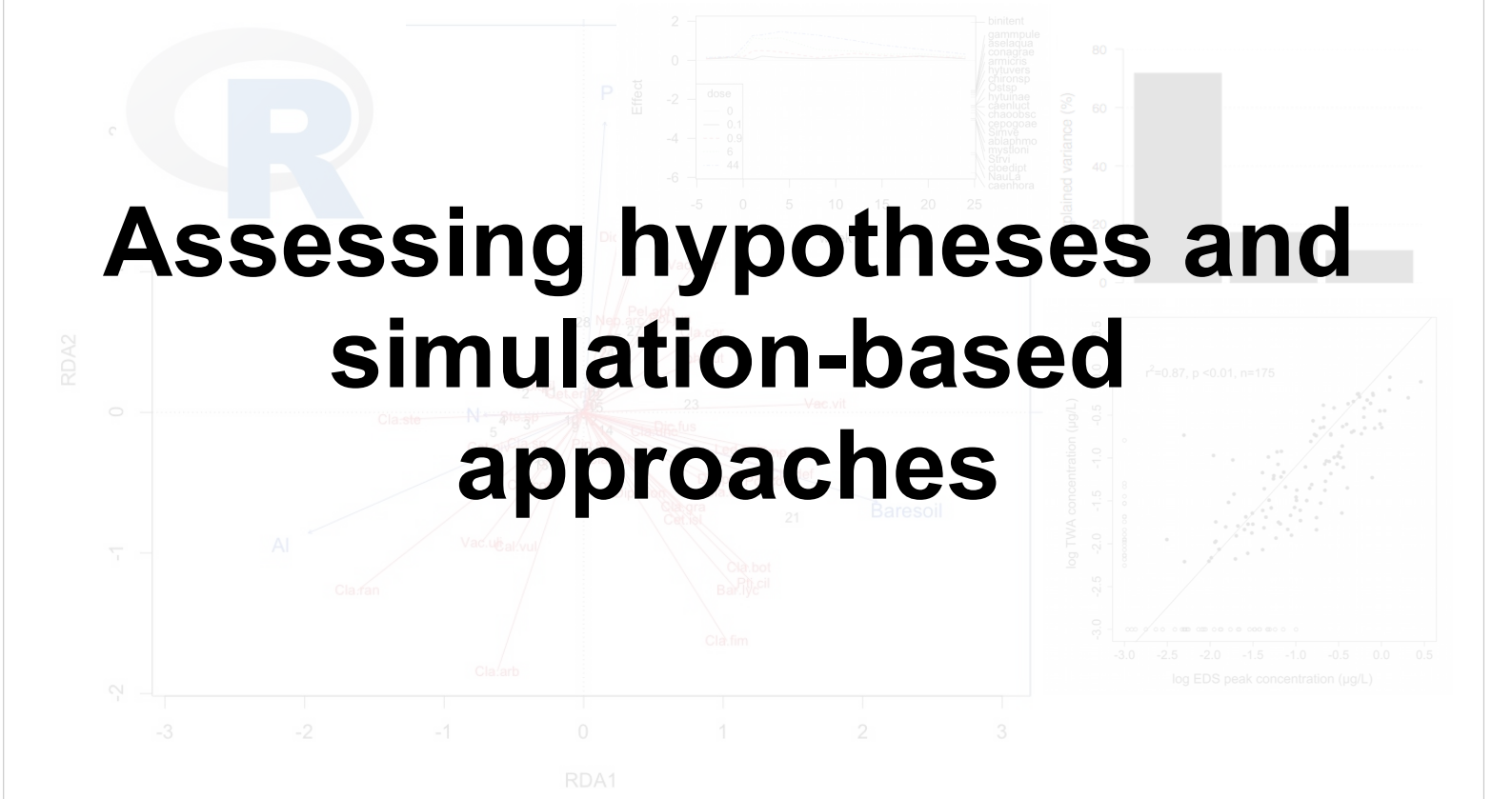
- Recall:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

- Application of CV: Calculate CV-R^2 and $\mathrm{CV} - \widehat{\mathrm{MSPE}}$

Tools for complex data analysis

University of Koblenz-Landau 2018/19



Assessing hypotheses and simulation-based approaches

Ralf B. Schäfer

These slides and notes complement the lecture with exercises “Tools for complex data analysis” for ecotoxicologists and environmental scientists. Do not hesitate to contact me if you have any comments or you find any errors (slides, slide notes, or code): schaefer-ralf@uni-landau.de

While I made notes below the slides, some aspects are only mentioned in the R demonstration associated with the lecture.

Learning targets

- Understand and evaluate the concept of the *p*-value and of assessing hypotheses
- Explain and apply simulation-based approaches to data analysis

2

Learning targets and study questions

- Understand and evaluate the concept of the *p*-value and of assessing hypotheses
  - Define the *p*-value and explain the rationale for its use.
  - Describe the differences between the different approaches to assess hypotheses.
  - Discuss pros and cons of significance testing.
  - Outline the good practice when assessing hypotheses.
  - Distinguish scientific and statistical hypotheses.
  - What are the assumptions of the *t*-test?

3

Learning targets and study questions

- Explain and apply simulation-based approaches to data analysis
  - Discuss how simulation-based approaches link the two cultures to data analysis.
  - Explain the purpose and critically discuss permutation tests.
  - Explain the purpose and critically discuss bootstrapping.
  - Explain the main idea of cross-validation and discuss the selection of *k* with respect to the bias-variance trade-off.
  - Discuss the application of bootstrapping and cross-validation in regression analysis.

4

Assessing hypotheses and simulation-based approaches

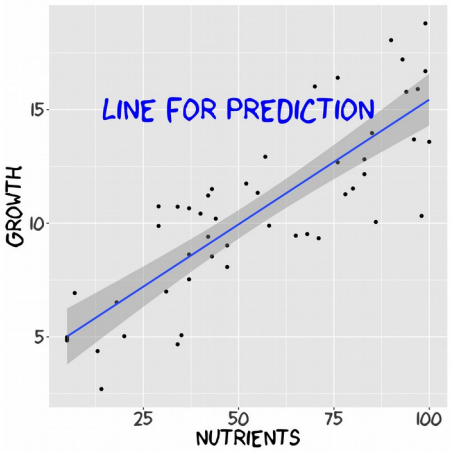
Contents

5

Recap: Research goals and model output

1. Prediction
2. (Parameter) estimation
3. Assessing hypotheses

Example: Assessing hypotheses related to the relationship between plant growth and nutrient concentrations.



4. Explanation

?

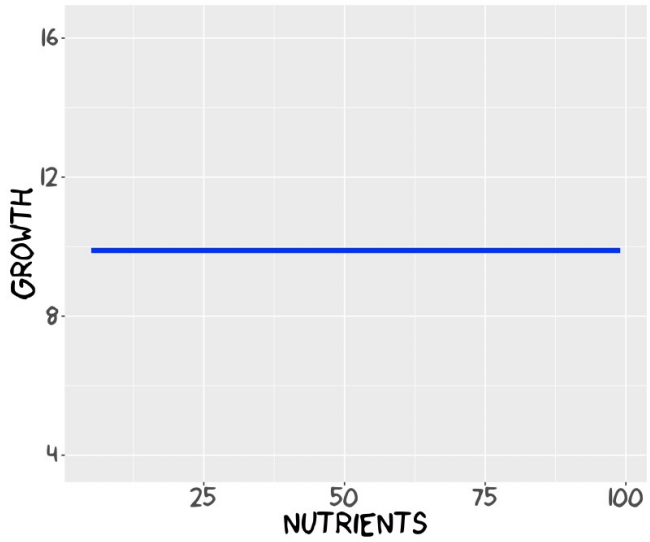
```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 137.79468   31.92061   4.317 1.95e-05 ***
## X             1.45722    0.03152  46.231  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103 on 448 degrees of freedom
## Multiple R-squared:  0.8267, Adjusted R-squared:  0.8263
## F-statistic:  2137 on 1 and 448 DF,  p-value: < 2.2e-16
```

6

We will discuss the *F*-statistic in the context of multiple linear regression in the next session. For the case of simple linear regression, it does not contain much additional information to those related to the *t*-value, which will be discussed hereafter.

Case study: Nutrients and plant growth

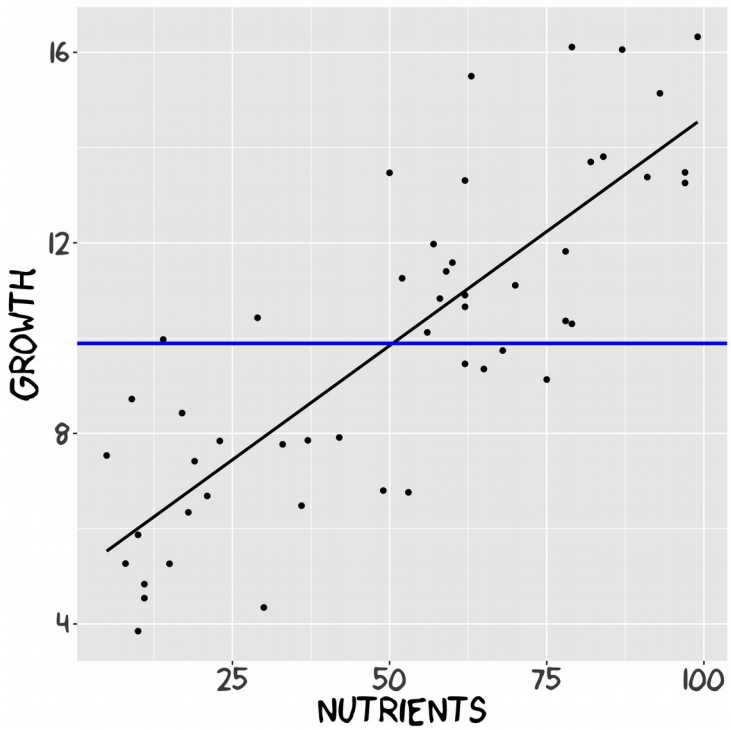
A skeptic claims that some nutrients have no effect on plant growth. This translates to the hypothesis of a slope of 0 in a regression model, i.e. $\beta_1 = 0$.



Our research goal is to produce evidence against this claim, which we define as the so-called *null hypothesis* H_0: $\beta_1 = 0$.

Case study: Nutrients and plant growth

We conduct an experiment and obtain an estimate of 0.1 for the slope, i.e. $\hat{\beta}_1 = b_1 = 0.1$:



We need a statistic that informs on the conformity of our data with H_0.

8

Recap: Confidence intervals

Can't we use confidence intervals (CIs)? They give upper and lower limits for the true parameter.

General calculation of confidence interval for parameter θ:

$\theta \pm$ value related to confidence level and probability distribution × SE

(e.g. 95%, 99%)

Known or assumed probability distribution of parameter

95% CI for slope:

$b_1 = 0.1$, 95% CI[0.07, 0.12]

$b_1 = 0$ not included → If assumptions are met (e.g. probability distribution) and the specific CI is one of those containing the true parameter, then 0 is not a potential true parameter.

Don't we have a more direct way to assess the hypothesis?

9

Remember that the confidence level defines the frequency the interval contains the true parameter in (hypothetical) repeated studies.

The statement related to whether the CI contains 0 may sound somehow discontenting, but is owed to the fact that a single experiment or study can not achieve absolute certainty. Or in the words of one of the most important statisticians of the 20th century: “No isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon” (Fisher 1935:16).

Generally, CIs should not be used to assess hypotheses. We only do this here for the purpose of illustration and to lay the basis for the later recognition of the relationship between CIs and significance testing.

Fisher, R.A. (1935): The design of experiments. London: Oliver & Boyd.

The concept of *p*-value

Is the probability p of obtaining such or more extreme data if H_0 is true: $p = 2 \times P(T \geq |t|; H_0)$ where T is a test statistic with the realised value t.

<u>Application to our case study:</u>

Test statistic is the so-called t-value related to the t-test for regression coefficients:

$$t = \frac{\hat{\beta} - \beta}{\mathrm{SE}_{\hat{\beta}}} \qquad \text{from } H_0\text{: } \beta_1 = 0 \text{ follows:} \qquad t = \frac{\hat{\beta}}{\mathrm{SE}_{\hat{\beta}}}$$

We obtain a value for t = 9.79 → Determine $P(t \geq 9.79; H_0)$ from a so-called *Student's t* distribution.

10

We provide the mathematical formulation for the so-called two sided *p*-values. They represent the default *p*-values and so-called one sided *p*-values are only of interest in very specific situations discussed in Lombardi & Hurlbert (2009).

Note that the T and t in the definition of the *p*-value refer to any test statistic, whereas the *t*-test and *t*-value for our case study refer to the so-called *Student's t* distribution, which will be discussed hereafter. The *t*-test will be further discussed later in the session. The *t*-test for regression coefficients is the so-called one-sample *t*-test.

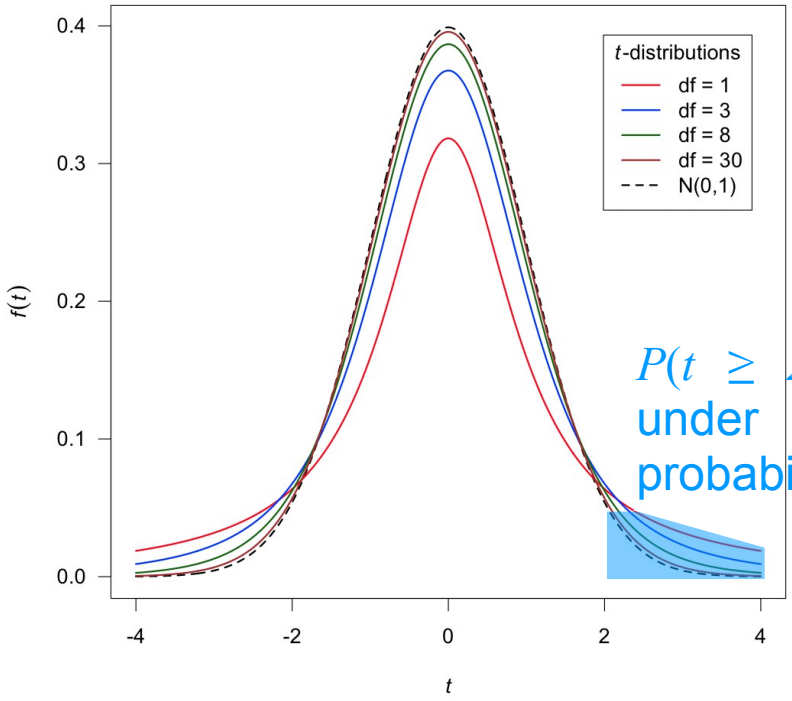
Consider the below R output for a regression model. Indeed, the *t*-value is simply the division of our Estimate by its Std. Error:

```
Coefficients:
            Estimate Std. Error t value
(Intercept) 5.048892   0.567110   8.903
nut         0.095850   0.009791   9.790
```

Lombardi C.M. & Hurlbert S.H. (2009) Misprescription and misuse of one-tailed tests. *Austral Ecology* 34, 447–468.

The *Student's t*-distribution

- Symmetric and continuous probability distribution
- Arises when estimating the mean of a normal distribution at small n and unknown σ
- Approximates normal distribution for ≥ 30 degrees of freedom (df)



$P(t \geq 2)$ is the area under the respective probability distribution

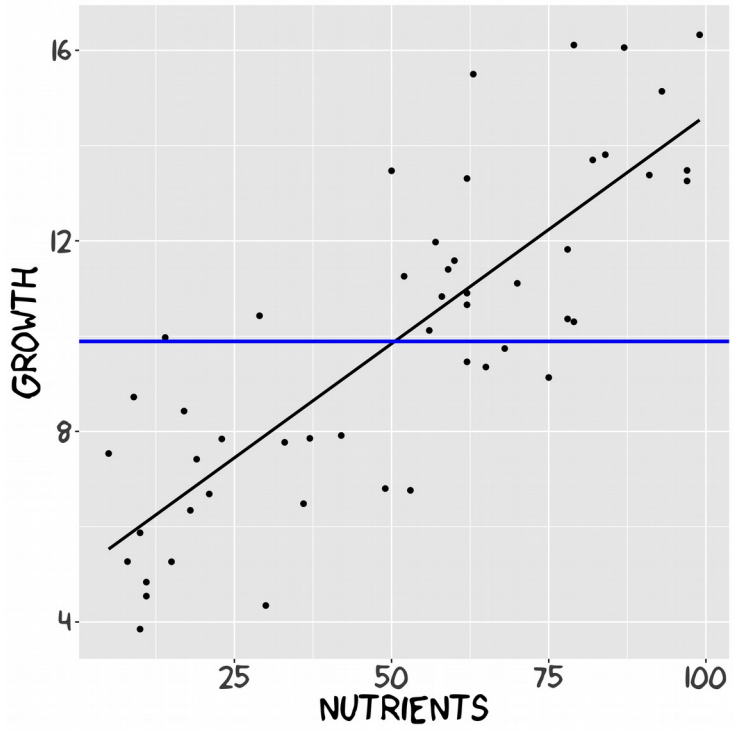
11

For details on the distribution see the excellent Wikipedia page: https://en.wikipedia.org/wiki/Student%27s_t-distribution

Note that we choose $P(t \geq 2)$ for the sole purpose of illustration (the *t*-value from our case study is outside the boundaries of this figure).

Case study: Nutrients and plant growth

p-value for our data: $5 \times 10^{-13} \rightarrow$ probability of obtaining these or more extreme data, if H_0 is true (i.e. no relationship between nutrients and growth), is almost zero.



12

In other words, the probability of obtaining a regression slope of 0.1 for the given degrees of freedom and standard error, if the true slope is 0, is almost zero.

Assessing hypotheses and simulation-based approaches

Contents

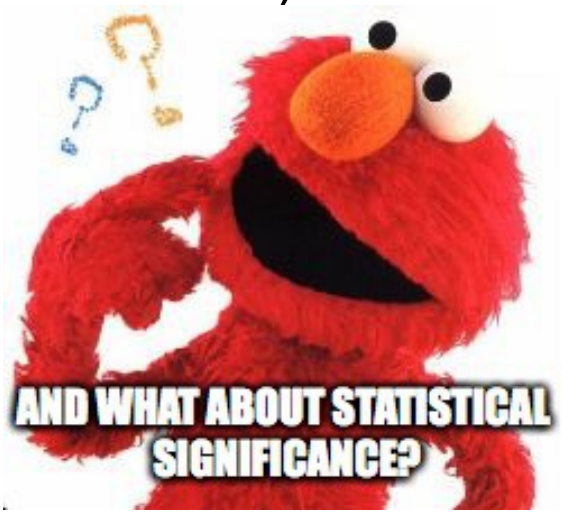
13

Interpretation of *p*-values

Provide the degree to which data are compatible with the pattern predicted by a given null hypothesis, conditional that the assumptions of the underlying statistical model are met. (*cf.* Greenland et al. 2016)

→ Can be used to assess the plausibility of H_0

→ This interpretation follows the NeoFisherian approach (*cf.* Hurlbert & Lombardi 2009)



14

Our presentation and interpretation of *p*-values follows the approach by Hurlbert & Lombardi (2009) called NeoFisherian significance assessment (NFSA).

Greenland S., Senn S.J., Rothman K.J., Carlin J.B., Poole C., Goodman S.N., et al. (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology 31, 337–350. Free to download at: https://link.springer.com/article/10.1007/s10654-016-0149-3

Hurlbert S.H. & Lombardi C.M. (2009) Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. Annales Zoologici Fennici 46, 311–349. Free to download at: http://cescos.fau.edu/gawliklab/papers/HurlbertLombardi.pdf

What about statistical significance?

NeoFisherian approach:

- *p*-value is graded measure of evidence against H_0 → Report exact *p*-values
- Possible conclusions from test statistic and related *p*-value:
    1. Effect (e.g. difference, association) seems negative
    2. Effect (e.g. difference, association) seems positive
    3. Neither can be confidently stated, judgement reserved or suspended

→ Discourages 1) Use of the term "statistical significance" and 2) to assign special status to $p < 0.05$

→ Generally: Evaluate effect size and overall evidence

Hurlbert & Lombardi 2009

The approach of Hurlbert & Lombardi (2009) is largely compatible with other contributions to the debate (e.g. Greenland 2016, Amrhein et al. 2017) including the statement of the American Statistical Association (ASA) (Wasserstein & Lazar 2016). However, Amrhein et al. (2017) rather suggest to suspend firm decisions than judgments and to be cautious when interpreting the *p*-value. They emphasise that in any case we should discuss how strong we judge the overall evidence (including other studies), and how practically important the effect is.

Criticism includes that without clear thresholds for significance (such as $p = 0.05$) this increases subjectivity. In the end, everybody needs to make a judgment and without a firm rule, this would add a layer of variability and increase the confirmation bias (that researchers tend to favor interpretations and judgments that confirm their prior beliefs or hypotheses). See Amrhein et al. (2017) for a discussion of this argument.

Note that a confidence level of 95% in the context of CIs is related to the significance threshold of $p = 0.05$.

A nice simulation that allows the user to identify the *p*-value that he or she regards as relevant in a regression setting is provided here: www.openintro.org/why05

Amrhein V., Korner-Nievergelt F. & Roth T. (2017) The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ 5, e3544. Free to download: https://peerj.com/articles/3544/

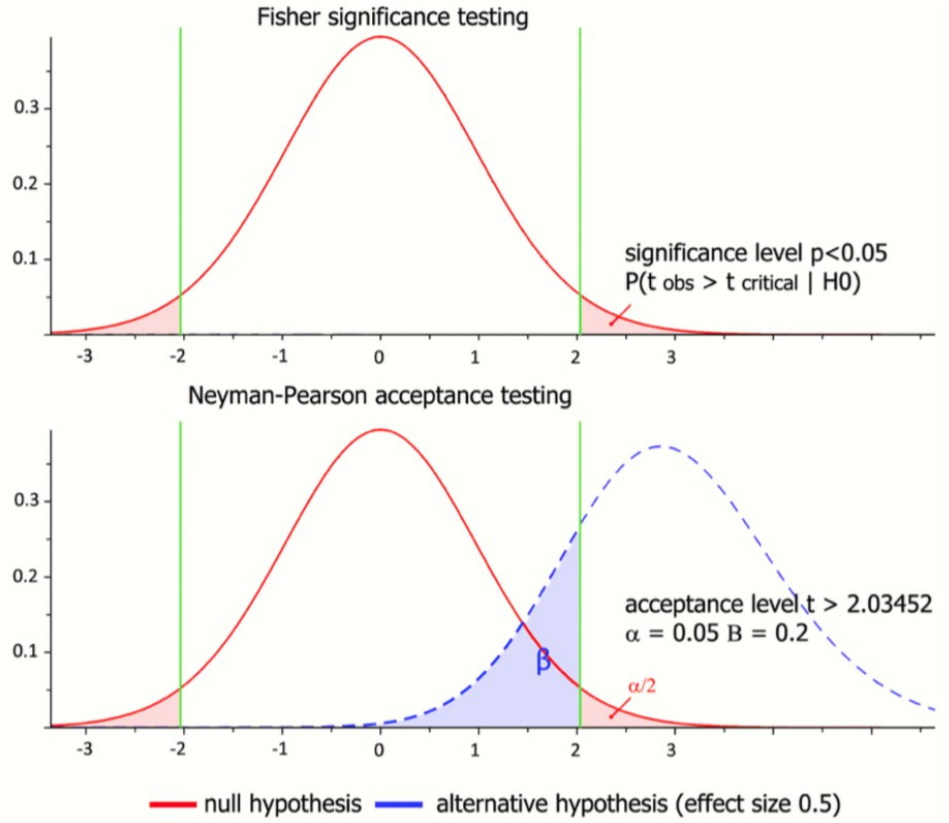
Hurlbert S.H. & Lombardi C.M. (2009) Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. Annales Zoologici Fennici 46, 311–349. Free to download at: http://cescos.fau.edu/gawliklab/papers/HurlbertLombardi.pdf

Greenland S., Senn S.J., Rothman K.J., Carlin J.B., Poole C., Goodman S.N., et al. (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology 31, 337–350. Free to download at: https://link.springer.com/article/10.1007/s10654-016-0149-3

Wasserstein R.L. & Lazar N.A. (2016) The ASA's Statement on *p*-Values: Context, Process, and Purpose. The American Statistician 70, 129–133.

Statistical significance: Classical approaches

- Two approaches: Fisherian and Neyman-Pearson (NP)
- Fisher: testing of significance of single (null) hypothesis H_0, NP: accept alternative hypothesis if H_0 can be rejected



Both approaches employ fixed cut-off points and typically interpret $p < 0.05$ as statistically significant

Pernet 2017 *F1000 Research* 4

To avoid confusion, Hurlbert and Lombardi (2009) suggest to use the term *PaleoFisherian* approach for Fisher's significance testing. This approach is in line with the early writings of Fisher, whereas the proposed NeoFisherian approach is in line with the later writings of Fisher.

The Neyman-Pearson approach introduces a so-called alternative hypothesis H_A and focuses on the acceptance of hypotheses. H_A postulates a difference (i.e. an effect size) between the two statistical populations from which is sampled. In this framework, the main focus is on the acceptance/rejection of hypotheses rather than on *p*-values. They are only a tool to decide on the acceptance or rejection. The framework describes two errors: First, the so-called *Type I error* that we make when rejecting H_0, which states that there is no effect (i.e. effect size = 0), although H_0 is true. The so-called *Type II error* is to accept H_0 although it is wrong. The probability of making such errors under repeated experiments (*cf.* rationale of CIs) should be minimised by selecting appropriate sample sizes. The α gives the frequency of Type I errors, whereas β gives the frequency of *Type II errors*, indicated by the overlapping area in the figure. Within the NP framework, both α and β should be set before a study (for guidance on setting see Mudge et al. (2012)). The power of a test is 1-β and can be used to either set the sample size before conducting a study or, afterwards, to inform on the probability of finding a (true) effect. For an introduction to power analysis see Crawley (2012): 317-319 and Qian (2017): 109-116. Hurlbert and Lombardi (2009) and Amrhein et al. (2017) argue that this approach should be reserved for very specific situations and discourage its general use.

In the scientific literature, many studies have used some form of synthesis of both approaches (PaleoFisherian and NP). For an overview see McCarthy (2015) and Pernet (2017).

Amrhein V., Korner-Nievergelt F. & Roth T. (2017) The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ 5, e3544. Free to download: https://peerj.com/articles/3544/

Hurlbert S.H. & Lombardi C.M. (2009) Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. Annales Zoologici Fennici 46, 311–349. Free to download at: http://cescos.fau.edu/gawliklab/papers/HurlbertLombardi.pdf

McCarthy M.A. (2015) Approaches to statistical inference in: Fox G.A., Negrete-Yankelevich S. & Sosa V.J. eds (2015) Ecological statistics: contemporary theory and application. Oxford University Press, Oxford. p. 15-43.

Mudge J.F., Baker L.F., Edge C.B. & Houlahan J.E. (2012) Setting an Optimal α That Minimizes Errors in Null Hypothesis Significance Tests. PLoS ONE 7, e32734. Free to download: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0032734

Pernet C. (2017) Null hypothesis significance testing: a guide to commonly misunderstood concepts and recommendations for good practice. F1000 Research 4. Free to download at: https://f1000research.com/articles/4-621/v5

Perezgonzalez J.D. (2015) Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Frontiers in Psychology 6. Free to download at: http://journal.frontiersin.org/Article/10.3389/fpsyg.2015.00223/abstract

Qian S.S. (2017) Environmental and ecological statistics with R, 2nd edn. Chapman & Hall/CRC, Boca Raton, Fla.

Problems of significance testing and misinterpretations of *p*-values

- Significance testing dichotomises the *p*-value scale into significant (usually $p < 0.05$) and non-significant. This logic has many flaws:
  - Why is a result with $p = 0.055$ less relevant than $p = 0.045$?
  - *p*-value is sensitive to sample size (increasing sample size *ceterus paribus* will lead to significance)
  - *p*-value has low replicability (see Halsey et al. 2015) → leads to seemingly contradictory (significant and non-significant) results in replicated studies
  - Focus on significant results leads to distortion in scientific literature (ignorance of potentially relevant results)

If these arguments do not put you off from significance testing, give Arnheim et al. (2017) a read before using it, to be aware of the flaws.

Note that the same issues apply to CIs when using them for significance testing. They are also no salvation from the low replicability of *p*-values (see van Helden 2016).

Amrhein V., Korner-Nievergelt F. & Roth T. (2017) The earth is flat (p > 0.05): significance thresholds and the crisis of unreplicable research. PeerJ 5, e3544. Free to download: https://peerj.com/articles/3544/

Halsey L.G., Curran-Everett D., Vowler S.L. & Drummond G.B. (2015) The fickle P value generates irreproducible results. Nature Methods 12, 179–185.

van Helden J. (2016) Confidence intervals are no salvation from the alleged fickleness of the *P* value. Nature Methods 13, 605–606.

Problems of significance testing and misinterpretations of *p*-values

- Statistical significance does not imply scientific significance (and vice versa)
- High *p*-value (e.g. above a fixed significance threshold) does not mean that H_0 is true! (see Hurlbert & Lombardi 2009: 321-323 for example)
- Very low *p*-value does not mean that H_0 is incorrect (or H_A is true in NP approach)
- *p*-values do not inform on effect size

18

Ceterus paribus (i.e. when everything else remains the same such as sample size, statistical test and variance) an increase in the effect size is associated with a decrease in the *p*-value. However, the *p*-value does not allow for general judgments – the *p*-value can be high despite a high effect size and low despite a low effect size.

Greenland et al. 2016 provide many more examples of misinterpretations of *p*-values and significance testing.

Scientific hypotheses are often quite complex and can rarely be directly translated into the framework of the PaleoFisherian or NP approach. In addition, evaluation of scientific hypotheses requires weighing of all evidence, i.e. also considering other studies. Thus, a high or low *p*-value from a single study may not be that relevant in the light of the overall evidence. Moreover, at least PaleoFisherian significance testing can also be conducted to protect against overinterpretation of detected effects, albeit the statistical null hypothesis may not be regarded as likely or relevant from a scientific perspective.

Greenland S., Senn S.J., Rothman K.J., Carlin J.B., Poole C., Goodman S.N., et al. (2016) Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology 31, 337–350. Free to download at: https://link.springer.com/article/10.1007/s10654-016-0149-3

Hurlbert S.H. & Lombardi C.M. (2009) Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. Annales Zoologici Fennici 46, 311–349. Free to download at: http://cescos.fau.edu/gawliklab/papers/HurlbertLombardi.pdf

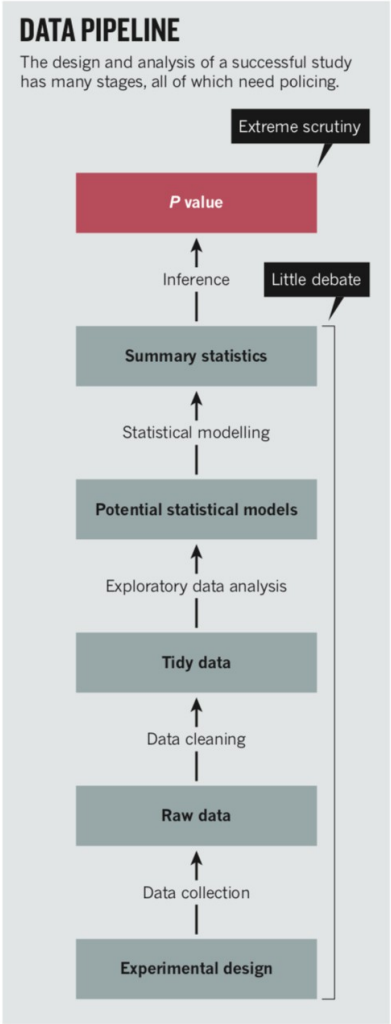
Debate on *p*-values and testing

- Debate since almost 100 years, mainly in statistical community
- Less attention in algorithm-based (machine learning) community
- Relevance of issue may be overstated and alternatives have similar issues

"*Whatever method we use in threshold tests and decision heuristics* […] *we create biases and invalidate the answers we give to our questions*" Arnheim et al. 2017

"[…] *look for a magic alternative to NHST [null hypothesis significance testing], some other objective mechanical ritual to replace it. It doesn't exist*" Cohen 1994

"*I thought greater use of Bayesian methods would reduce that misuse [RBS: of statistics]. Now I am less convinced. And the debates seem to distract from more important issues.*" McCarthy 2015



Leek & Peng 2015

We will later discuss other approaches to statistical inference. For an overview see McCarthy (2015). Note that for multimodel inference (e.g. model selection) other approaches such as likelihood or Bayesian approaches are typically more appropriate.

Amrhein V., Korner-Nievergelt F. & Roth T. (2017) The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. PeerJ 5, e3544. Free to download: https://peerj.com/articles/3544/

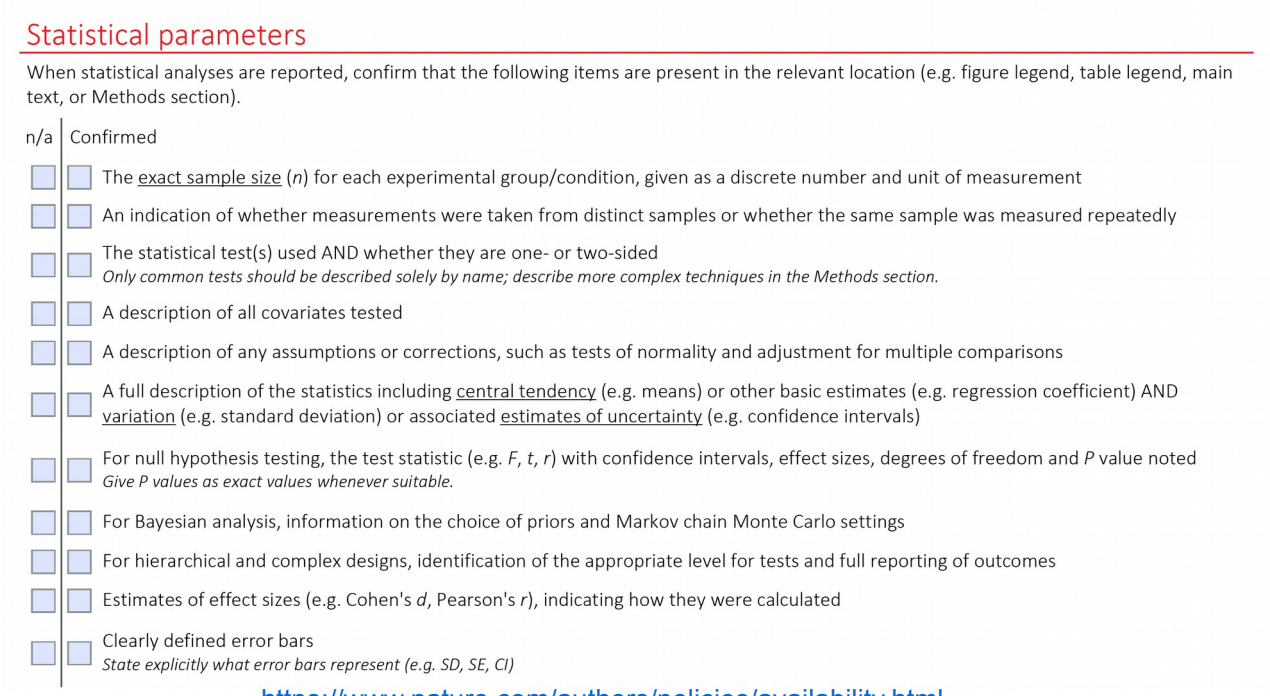
Cohen J. (1994) The Earth Is Round (P-Less-Than.05). American Psychologist 49, 997–1003.

Leek J.T. & Peng R. (2015) P values are just the tip of the iceberg. Nature 520, 1.

McCarthy M.A. (2015) Approaches to statistical inference in: Fox G.A., Negrete-Yankelevich S. & Sosa V.J. eds (2015) Ecological statistics: contemporary theory and application. Oxford University Press, Oxford. p. 15-43.

Good practice in assessing hypotheses

- Report effect sizes
- If calculated, report exact *p*-values (for any approach)
- If following NP approach: report power
- Consider journal requirements (example: *Nature*)



Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☐	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☐	A description of all covariates tested
☐	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☐	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

20 https://www.nature.com/authors/policies/availability.html

For an introduction to power analysis see Crawley (2012): 317-319 and Qian (2017): 109-116.

Only a few journals provide as detailed guidelines on statistics as *Nature*. They can be used to also guide submissions to other journals.

Qian S.S. (2017) Environmental and ecological statistics with R, 2nd edn. Chapman & Hall/CRC, Boca Raton, Fla.

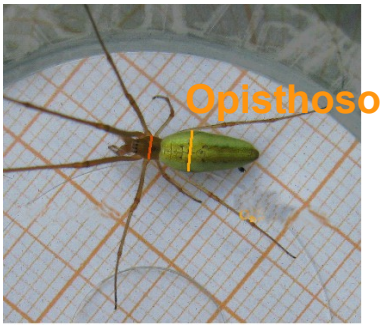
Assessing hypotheses and simulation-based approaches

Contents

21

Case study: Pesticides and spiders

Research question: Does environmental pesticide exposure reduce the body size of spiders?



Scientific hypothesis: The concentration of pesticides that are taken up in the environment by spiders require the activation of energetically costly detoxication processes. This reduces the energy for growth and consequently the body size.

Laboratory experiment: Treatment of a group of spiders "*a*" with a typical environmental exposure concentration and measurement of a body size metric (i.e. opisthosomal width) and comparison to a control group "*b*" kept under the same conditions, except for pesticide exposure.

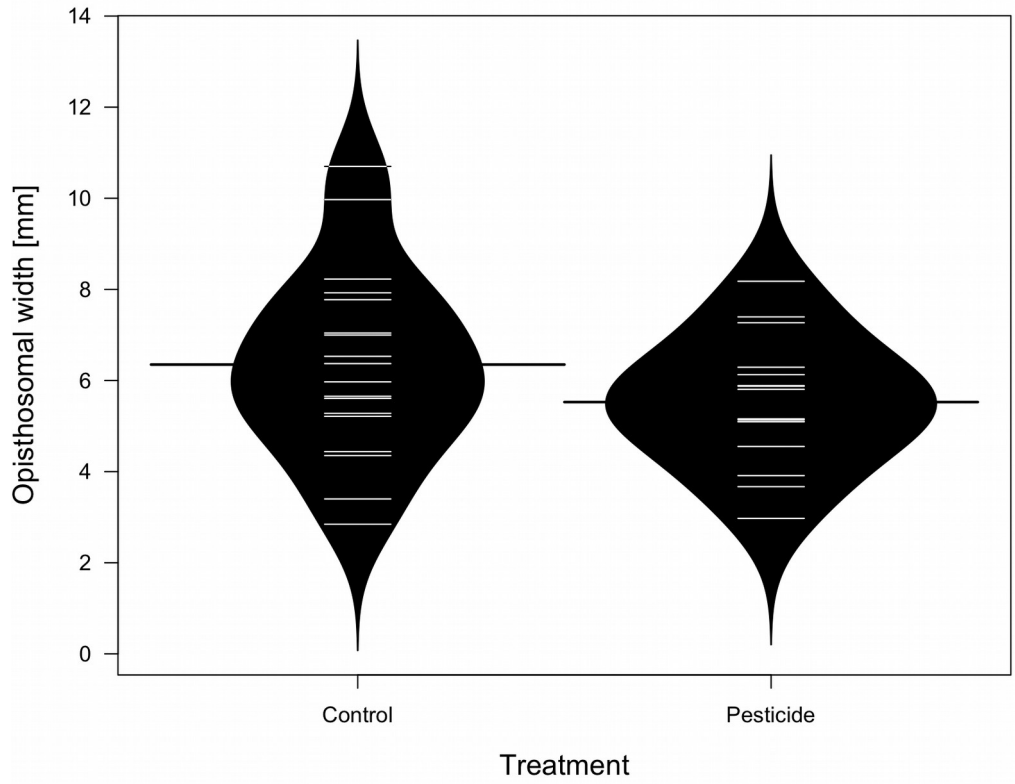
22

Note that this case study is of course very simplified, the scientific hypothesis would require more specification (e.g. which environment, which spider species, which pesticides) and evaluation would require more data than is produced in this study (e.g. data on environmental exposure and internal concentrations, measurement of detoxication processes). Moreover, laboratory conditions (e.g. single species, different conditions than in the real world, exposure likely too high) are likely biased compared to the real world situation. Nevertheless, this case study exemplifies the difference between the scientific hypothesis and the statistical (null) hypothesis that is assessed.

Case study: Pesticides and spiders

Statistical null hypothesis: The population mean of opisthosomal width μ is equal for the groups a and $b \rightarrow$ H_0: $\mu_a = \mu_b$

→ Do the experimental data provide evidence against H_0?



23

Note the difference between the scientific hypothesis and the statistical (null) hypothesis that is assessed.

Again, we want to refute the sceptic's claim, which is that the pesticide has no effect. This translates to a H_0 that the samples for the groups a and b have been drawn from statistical populations with equal mean.

Comparing two means with the t-test

Recall: p-value is the probability of obtaining such or more extreme data if H_0 is true, with $p = 2 \times P(T \geq |t|; H_0)$ where T is a test statistic with the realised value t.

→ Test statistic for comparison of two means is provided by the **two-sample *t*-test** (recall the one-sample t-test for regression coefficients)

Null hypothesis: The samples have been drawn from populations with equal μ. → $H_0: \mu_1 = \mu_2$

Calculation of the realised value t:

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_{x_1 x_2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{with} \quad s_{x_1 x_2} = \sqrt{\frac{(n_1 - 1)s_{x_1}^2 + (n_2 - 1)s_{x_2}^2}{n_1 + n_2 - 2}}$$

24 „Difference of means divided by standard error of difference“

The t-test is an example for a classical statistical test. We discuss it to exemplify the classical procedure in statistical testing: A test statistic is calculated to assess the probability of obtaining such or more extreme data if H_0 is true, using the p-value. Typically, H_0 is rejected if $p < 0.05$ (but remember our discussion on statistical significance).

As stated before, note that the T and t in the definition of the *p*-value refer to any test statistic, whereas the t-test and t-value for our case study refer to the so-called *Student's t* distribution, introduced before.

The *t*-test should be used when the standard deviation s is estimated from the data. If the standard deviation of the population is known, i.e. σ, the z-test can be used .

For definitions in the formula see the *Key terms and definitions* document.

Generally, the higher the value for a test statistic, in this case t, the lower the p-value. The t value increases with a higher difference between the sample means. Hence, the higher the difference between the sample means, the lower the probability to obtain such or more extreme data, if the samples originate from populations with equal mean. Moreover, the t value decreases with an increase in the standard error of difference. Hence, the higher the variability around the sample means, the higher the probability to obtain such or more extreme data, if the samples originate from populations with equal mean.

Assumptions of the *t*-test

- Independent samples from the population
  - Both populations have been randomly sampled
  - Evaluation of assumption requires knowledge on sampling
- Normal distribution
  - Both populations that have been sampled (e.g. X, Y) should follow a normal distribution: $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$
  - Evaluation of assumption through graphical inspection (*QQ*-plot)
- Variance homogeneity
  - The variances of both populations that have been sampled (e.g. X, Y) are equal: $\sigma_X^2 = \sigma_Y^2$
  - Evaluation of assumption through graphical inspection (Conditional boxplot, beanplot)

25

The *t*-test should be used when the standard deviation was estimated from the data. If the standard deviation of the population is known, the *z*-test can be used (see Aho 2016, p. 206ff).

Graphical diagnostics should be preferred to hypothesis tests for checking the assumptions of normal distribution and variance homogeneity, which is discussed in detail in Quinn and Keough (2002) p.193ff. Graphical checking for normal distribution is conducted using *QQ*-plots. The two sample *t*-test is relatively robust against violation of the assumption of normal distribution as long as the variances are equal. In case of strong violations of this assumption, a non-parametric test should be applied such as the Wilcoxon rank sum test or the permutational *t*-test, which is briefly discussed later. If the assumption of variance homogeneity is violated (i.e. variances are unequal), the Welch`s *t*-test should be selected. See Crawley (2012, p. 293 ff) and Hothorn & Everitt (2014, p. 49 ff) for further information on the *t*-test as well as on the alternative tests if specific assumptions are not met.

Aho, K.A., 2016. Foundational and Applied Statistics for Biologists Using R. CRC Press, Boca Raton, Fl.

Assessing hypotheses and simulation-based approaches

Contents

26

Simulation-based approaches in data analysis

- Compatible with both data modelling (classical statistics) and algorithmic modelling (machine learning) cultures
- Infuses algorithm-based thinking into classical statistics
- Examples for simulation-based approaches for estimation, inference or model diagnosis in classical statistics:

1. **Permutation test** → Permuting (shuffling) the data to derive null distribution. Mainly used for inference
2. **Bootstrapping** → Randomly sampling subsets from the data with replacement. Mainly used for estimation of parameter distribution
3. **Cross-validation** (CV) → Splitting data into sets (i.e. sampling without replacement). Mainly used for validation of predictive models

27

For a more advanced overview on approaches that build on both classical statistics and algorithmic modelling see Ryo & Rillig (2017).

Ryo M. & Rillig M.C. (2017) Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere* 8, e01976.

Permutation test: Algorithm

Repeat k times:

1) Permute values in data set
2) Compute test statistic t^* for permuted data

3) Compare test statistic t_0 to generated null distribution

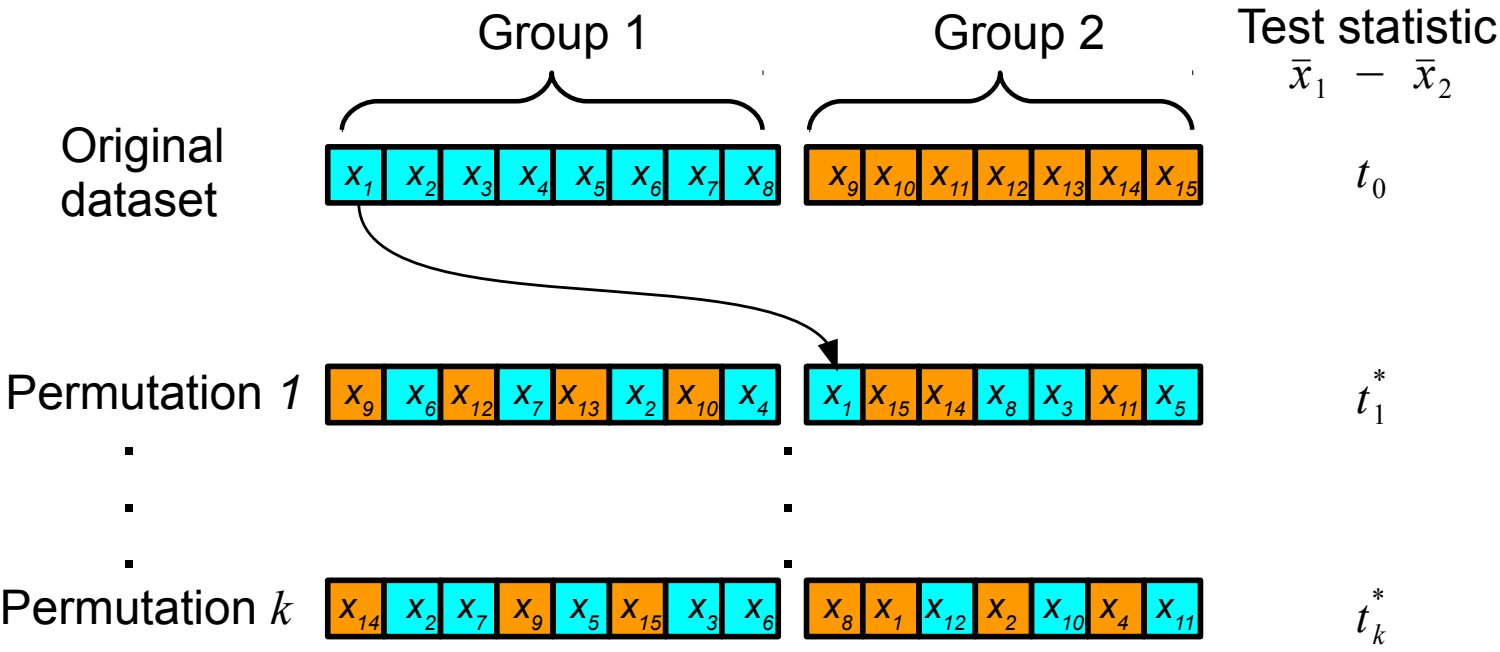
28

Permutation test: Algorithm

Repeat k times:

1) Permute values in data set
2) Compute test statistic t^* for permuted data
3) Compare test statistic t_0 to generated null distribution

Example: Permutation test for comparing group means

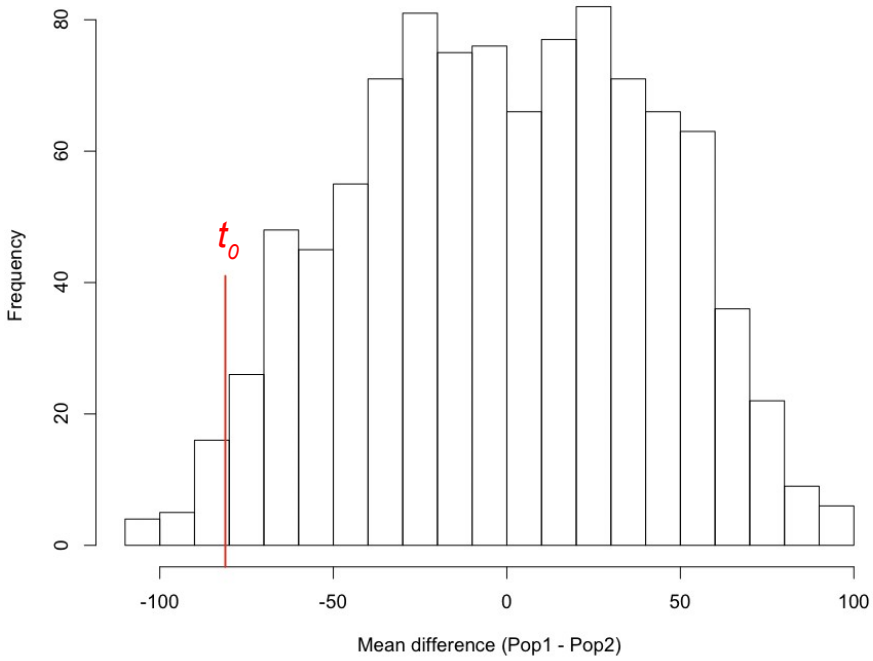


29

The permutation test for comparing group means represents an alternative to the two-sample *t*-test. This will be discussed in more detail in one of the following slides.

Permutation test: Generated distribution

$$p = \frac{\sum_{i=1}^{k} 1 \text{ if } t_i^* \leq t_0, else\, 0}{k+1}$$



- p gives probability that pattern is produced by pure chance
- Inference regarding statistical population only valid if distribution of sample data matches actual distribution of statistical population → particularly problematic for small *n*

30

The *p*-value is computed as the fraction of test statistics t^*, which are based on permuted data, that are more extreme (lower or higher depending on the hypothesis) than the non-permuted test statistic.

If the sample distribution deviates from the actual distribution of the statistical population, the permutation test only allows to infer conclusions that apply to the data at hand. These may not be very interesting. For the example of the mean comparison, this would translate to being unable to assess the null hypothesis:

$$H_0: \quad \mu_{group1} = \mu_{group2}$$

What a small sample size n is, depends on the context and no single number applies to all situations. For example, it will depend on the statistical distribution, statistical test etc. However, as a rule of thumb, sample sizes < 30 for a population are small. Still, much larger sample sizes can be required to reliably generalize from the permutation test to the statistical population.

Permutation test: Advantages and limitations

- Advantages
  - Applicable to any distribution
  - Applicable to complex designs through restricting permutations
- Limitations
  - Generalisation to statistical population requires matching distribution
  - Assessing statistical hypotheses can imply distributional assumptions that also apply to the permutation test, if aiming to infer to the statistical population (e.g. comparing means is affected by unequal variance)
  - Computationally intensive: Number of all possible permutations for a dataset is factorial *n*, i.e. *n*! (e.g. $35! \approx 10^{40}$) → Monte Carlo simulation

31

Legendre & Legendre 2012: 25ff

If the data follow a complex design (e.g. repeated measurements, clustered sampling), simple classical tests are not suitable (e.g. *t*-test), because the assumption of independence would be violated. Permutation tests are very versatile in embracing such designs.

Permutation tests also represent alternatives to classical tests (e.g. *t*-test) if the distributional assumption (e.g. normal distribution of *t*-test) is violated.

Although permutation tests can be used for any distribution, the violation of test assumptions of classical tests such as the equality of variance, will also affect inference from permutation tests. See Legendre & Legendre 2012: 25 for further details.

Monte-Carlo simulation

- Uses repeated random sampling to solve problems probabilistically (even though they can be deterministic in reality)
- Permutation tests use random numbers to randomly permute data → approximate with MC simulation
- Legendre & Legendre (2012): use at least 10,000 permutations for inference

Entrance of casino in Monte Carlo, Monaco

Edvard Munch - At the Roulette Table in Monte Carlo

32

Name refers to the city, it was chosen as code name for a secret project in the context of nuclear weapon research in Los Alamos, USA.

The larger the number of MC-based permutations, the lower is the error when approximating the distribution of all possible permutations with the MC-based permutation.

Picture sources:
Photo of Casino
https://pixabay.com/de/spielbank-casino-monte-carlo-monaco-188882/

Picture of Munch:
https://upload.wikimedia.org/wikipedia/commons/1/1f/Edvard_Munch_-_At_the_Roulette_Table_in_Monte_Carlo_-_Google_Art_Project.jpg

Assessing hypotheses and simulation-based approaches

Contents

33

Bootstrapping: Idea and algorithm

- Inference on statistic *t* is based on sampling distribution
  - Ideally: Draw all or many samples from statistical population
  - Reality: Most frequently only one sample available
  - → **Idea:** Draw samples from an estimate of the statistical population (i.e. the sample) and use these to estimate a property (e.g. variance) of the statistic *t*
- Algorithm:
  1) Draw random sample with replacement from data
  2) Compute statistic *t** for bootstrap sample

  } Repeat *k* times
  3) Use the *k* estimates to derive property of statistic
- Exhaustive bootstrapping ($k = n^n$) computationally demanding → approximate with Monte Carlo simulation
- With current computer power 10^4-10^5 simulations often viable

34

The name bootstrapping alludes to the phrase “pulling oneself up by one’s bootstraps,” which has been voiced by the fictional character Baron Münchhausen.

The core idea of bootstrapping is the following: The sampled data result from a specific data generating process. Samples from the sampled data should be similar to samples from the data generating process.

A sampling distribution describes the distribution of a parameter such as the mean, variance or median. For details see the section on the standard error in the document *Key terms and definitions.*

In analogy to the permutation tests, the following applies to bootstrapping: The larger the number of MC-based bootstrap samples, the lower is the error when approximating the bootstrap distribution with the MC-based samples.

Bootstrapping: Example

Example: Bootstrap to the mean (to derive variance)

t (here: mean)

Original dataset

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

$\bar{x} = 8$

Sampling with replacement

BS sample 1

15	7	8	4	15	1	11	9	1	3	6	14	2	11	12

$\bar{x}^* = 7.93$

BS sample 2

6	13	2	10	12	5	7	10	1	13	8	8	15	3	10

$\bar{x}^* = 8.2$

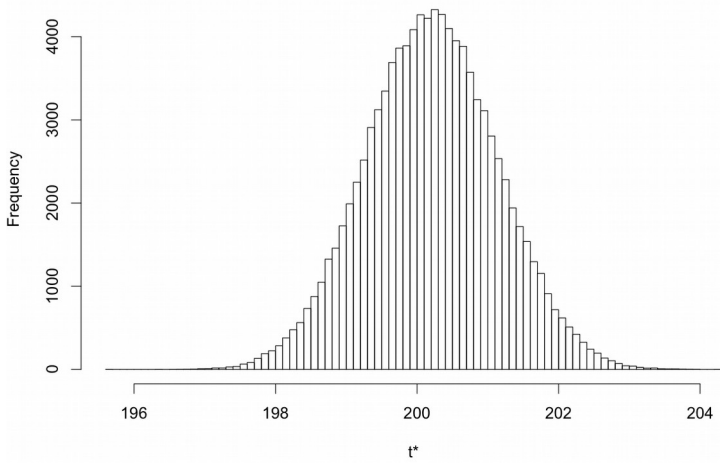
⋮

BS sample *k*

12	7	7	15	5	10	8	13	6	11	8	1	14	2	12

$\bar{x}^* = 8.73$

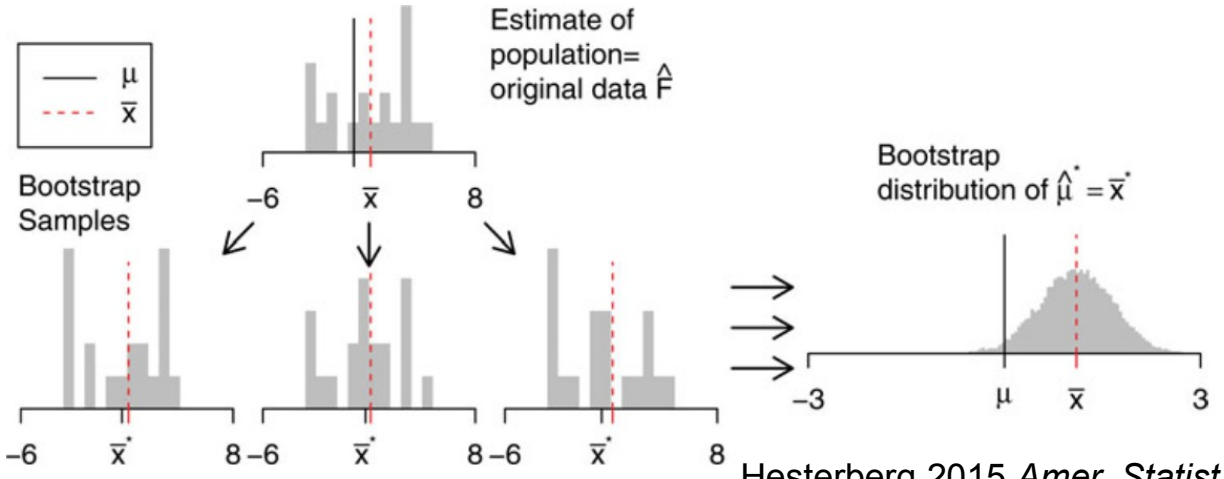
Distribution of statistic *t*



35

Bootstrapping: Limitations

- Do not use for assessing hypotheses
- No distributional assumptions implied, but not reliable for all distributions, particularly at small *n* (see Hesterberg 2015)
- Small *n*: use adjusted bootstrap percentiles (Bca) or switch to parametric statistics (allow for additional assumptions)
- Bootstrap does not improve estimate of population parameter μ, centred at $\bar{x}$



Hesterberg 2015 *Amer. Statist.* 69:371

Bootstrapping is generally less accurate than permutation tests for hypothesis testing.

BCa corrects for bias and skewness in the distribution of bootstrap estimates.

A very nice introduction and overview on bootstrapping is provided by:

Hesterberg T.C. (2015) What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. The American Statistician 69, 371–386.

Freely available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4784504/pdf/utas-69-371.pdf

Bootstrapping in regression analysis

Recall the residual definition e_i as: $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$

- Of residuals: Bootstrapping residuals, add to $\hat{y}$ to generate new y^* and calculate regression coefficients $\rightarrow x$ fixed
- Of cases: Bootstrapping complete cases and calculate regression coefficients $\rightarrow x$ random
- If x and y random sample (e.g. x not fixed in experiment), residuals correlated or exhibit non-constant variance $\rightarrow$ Bootstrapping cases

When bootstrapping residuals, the bootstrap samples (samples with replacement) are drawn from the the n residuals $e_1, e_2, \ldots, e_n$ yielding to a bootstrap sample of residuals $e_1^*, e_2^*, \ldots, e_n^*$

These bootstrapped residuals are added to the vector of fitted responses ($\hat{y}$) to obtain a vector of new responses y^*: $y_i^* = \hat{y}_i + e_i^*$

These new responses are used to calculate new bootstrapped regression coefficients (e.g. b_0^*, b_1^*). The procedure is repeated 1,000 to 10,000 times and, as usual in bootstrapping, delivers the distribution for a test statistic *t** (in simple regression analysis for b_0 and b_1).

For bootstrapping cases, pairs of x and y are bootstrapped and then the regression model is fitted, also providing bootstrapped regression coefficients (i.e. b_0^*, b_1^*).

Now when to use what? In case that the residuals exhibit non-constant variance or are correlated, the bootstrapping of residuals does not preserve the properties of the population sample and bootstrapping of cases should be preferred. However, if the observations for the predictors (x) have not been drawn randomly, but are fixed (for example, fixed concentration levels in an experiment), bootstrapping residuals should be preferred as it preserves these original x. For further details see Fox (2015): 658-660 and Hesterberg (2015) *Americ. Statist.* 69: 371–386.

Assessing hypotheses and simulation-based approaches

Contents

38

Cross-validation (CV)

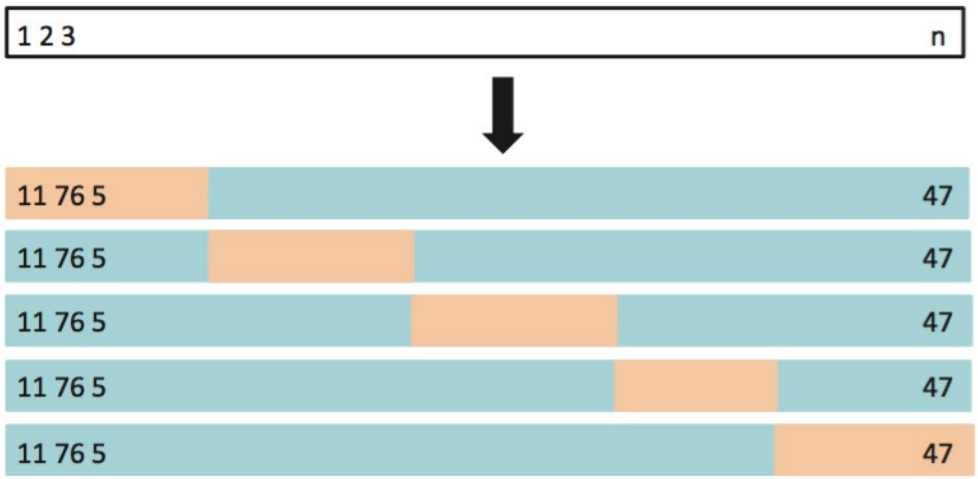
- **Aim:** Evaluate predictive accuracy of a fitted model
- Can be checked by predicting (known) responses from independent data sets (that were not used in model fitting) → Rare case
- **Idea:** Split the available data into training and test set and predict (known) observations in test set with a model fitted on the training data
- Algorithm:
  1. Draw *k* random samples without replacement from data
  2. For each *k:*
     1. Fit the model to the other *k*-1 parts
     2. Predict *k* from model and calculate the prediction error
  3. Calculate mean prediction error over the *k* estimates

Predictive accuracy measures the accuracy of predictions for new data.

CV is typically used in validation, but can also be used as goodness-of-fit measure to guide parameter estimation (see shrinkage methods later).

Cross-validation (CV)

Example: $k = 5$



- Problem of choosing *k*:
  - *k* = n (Leave-one-out CV predicts each observation from all others) → low bias, but high variance
  - *k* = 2 (split data into half) → low variance, but high bias
- *k* typically set to 5 or 10

Taken from James *et al.* 2013: 181

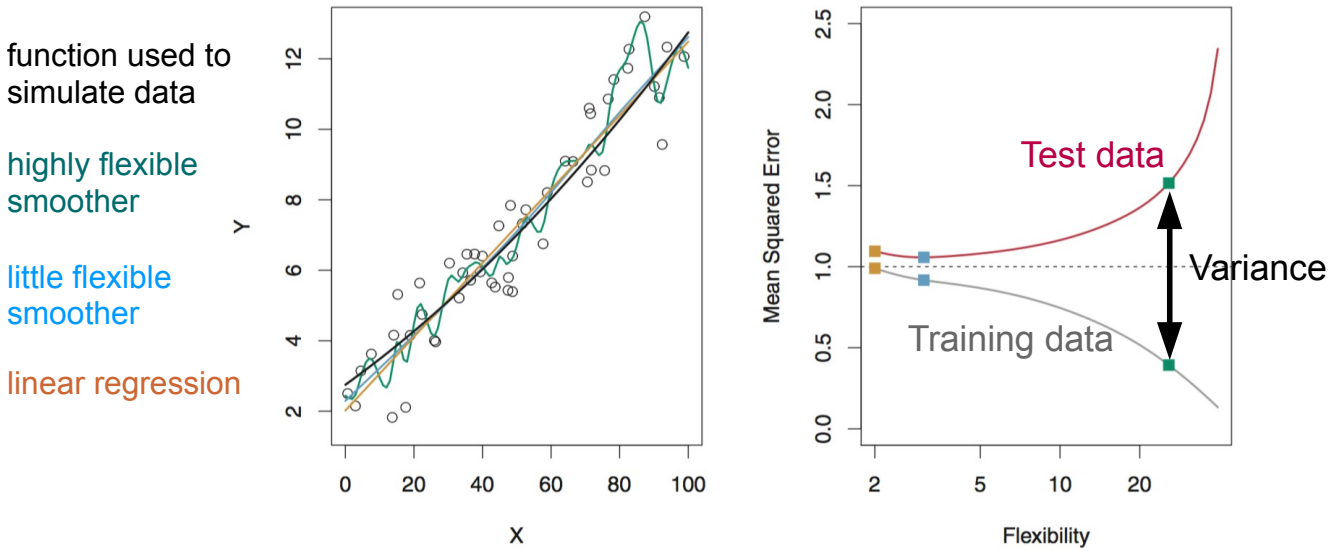
The bias-variance trade-off will be discussed in a more general context hereafter. Here, we discuss it with respect to the prediction accuracy, i.e. using cross validation to estimate the prediction accuracy. There is a trade-off between bias (error when estimating the 'true' prediction accuracy of the sample data) and variance (variability of the error when estimating new data). If we use a major fraction of the data (extreme case: $k = n$, where we use n-1 observations) in model fitting, the error of estimating the prediction accuracy of the full data is probably very low (low bias). However, the variability of the error when predicting a few (or only one for $k = n$) observations from different training sets is most likely high, which translates to a high variance. Conversely, if we use only half of the data ($k = 2$) in model fitting, we decrease the variance. In other words, the error when predicting the test set is most likely similar for the two training sets. But this comes at the cost of bias. In the case of $k = 2$, we are estimating the predictive accuracy from only a fraction of the data, whereas in practice all observations will be used in prediction. The prediction accuracy estimated from the fraction of the data is likely to differ (i.e. lower or higher) from that of the complete data set, i.e. exhibit bias. Thus, the bias increases when the relative size of the training set in CV decreases.

k is typically set to 5 or 10, i.e. the data is partitioned in 5 or 10 groups during CV as a compromise between bias and variance. Leave-one-out CV is considered less reliable than 5- or 10-fold CV (see Harrell 2015: 172).

Bias-variance trade-off

Definition in context of model validation:

- **Bias:** error when approximating training data
- **Variance:** variability in error when approximating test data



Higher flexibility (higher *k* in CV) → lower error for training data (i.e. lower bias), but variance will start to increase from some point

Taken from James *et al.* 2013: 33

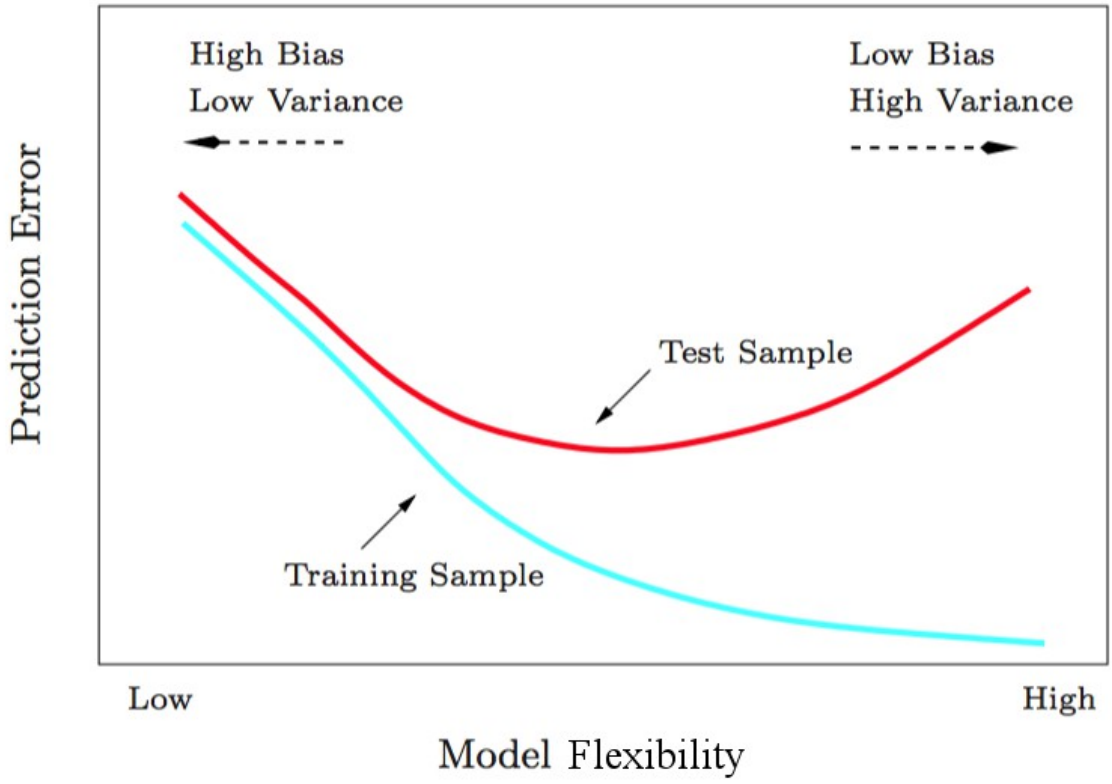
The left figure displays the fit of different models to data originating from the function plotted in black.

The models rank regarding bias: linear regression > little flexible smoother > highly flexible smoother.

Regarding variance (see right figure), the ranking is: highly flexible smoother > little flexible smoother > linear regression.

Bias-variance trade-off

Higher flexibility (higher *k* in CV) → lower error for training data (i.e. lower bias), but variance will start to increase from some point → Optimise combined error



Taken from Hastie, Tibshirani and Friedman 2011: 38

For a mathematical derivation of the bias-variance trade-off see Matloff(2017): 48f.

42

CV in regression analysis

- Predictive accuracy measured with estimate of Mean square prediction error (MSPE):

$$\widehat{\text{MSPE}} = \frac{1}{m}\sum_{i=1}^{m}\left(y_{\text{new},i} - \hat{y}_i\right)^2$$ for new observations (y_{new}) 1 to m

- Recall:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

- Application of CV: Calculate CV-R^2 and $\text{CV} - \widehat{\text{MSPE}}$

43

The square root of the the estimate of the MSPE, is termed RMSPE (Root mean square prediction error) and frequently used for comparison across models.

Remember that: $\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} e_1^2$