Software for Analyzing Data

Tutorial – Day 5

November 18, 2021

John-Oliver Engler

Rieke Baier & Chân Lê

Quantitative Methods of Sustainability Science Group

Today

- Preliminaries:
  - Who am I and what (not) to expect from me
  - Data science and statistics
- Theoretical input on hypothesis testing and the Central Limit Theorem (CLT)
- Excercises:
  1. Type-I and type-II-errors
  2. Simple tests with a real-world data set
  3. Exploring the limits of the CLT

Who am I?

Education and track record:

- M.Sc. Physics (2010, Konstanz University)
- Ph.D. in Economics (2014, Leuphana University)
- Actuary with BNP Paribas Cardif
- Postdoc with Quantitative Methods of Sustainability Science group (since 01/2017)

Research interests:

- Ecological economics / sustainability economics
- Risk theory
- Applied statistics and probability theory (in relation to sustainability)

Consultancy hours: no fixed hours, please contact me at engler@leuphana.de

What (not) to expect from me

- Help with getting started on the basics of applied statistics through:
  - Theoretical input
  - Examples
  - Excercises to be solved during the tutorial and/or at home
- Useful further reading on selected issues
- Advice on how to approach modeling and statistical issues
- However, do not expect:
  - To impress me with super-detailed questions about R programming
  - Help on Python
  - Me to be able to solve any given problem you may have right away

What is data science (not)?

- Data science is not about any of the following:
  - Making ‚complicated‘ models
  - Making ‚awesome‘ visualizations
  - Writing code

- Data science is about …
  - … using data …
  - … to create as much ‚impact‘ as possible …
  - … for yourself and/or the company or institution that you work for.

- ‚Impact‘ can be insights/knowledge, data products or product recommendations.

Things to cover

Technicalities and programming
Basics of R
Conditionals and control structures
Functions (built-in and custom)
Data wrangling / handling
Data visualization
Modeling syntax in R

Things to cover

Technicalities and programming	Statistics
Basics of R	Basic concepts of statistical data analysis
Conditionals and control structures	Linear models (incl. ANOVA) and generalized linear models
Functions (built-in and custom)	Generalized linear mixed models
Data wrangling / handling	Model selection and multi-model inference
Data visualization	Multivariate analysis (PCA, ordinations)
Modeling syntax in R	

R and R Studio

- You should familiarize yourself with R and R Studio, e.g. through the following resources

1. Helpful introductory treatments can be found here, among others:
   - Introduction to R and R Studio by Hefin Rhys

   https://www.youtube.com/watch?v=lL0s1coNtRk&t=12s (Part 1)

   https://www.youtube.com/watch?v=ZA28sOmq7nU&t=1s (Part 2)

2. You can also learn the basics of R by using the swirl package

Suggested further reading

For beginners:

Crawley, M. (2005), Statistics: An Introduction Using R

Intermediate to advanced:

Fox, G., et al. (2015), Ecological Statistics: Contemporary Theory and Application

Crawley, M. (2008), The R Book

Stock, J.H. and M.W. Watson (2015), Introduction to Econometrics

Advanced:

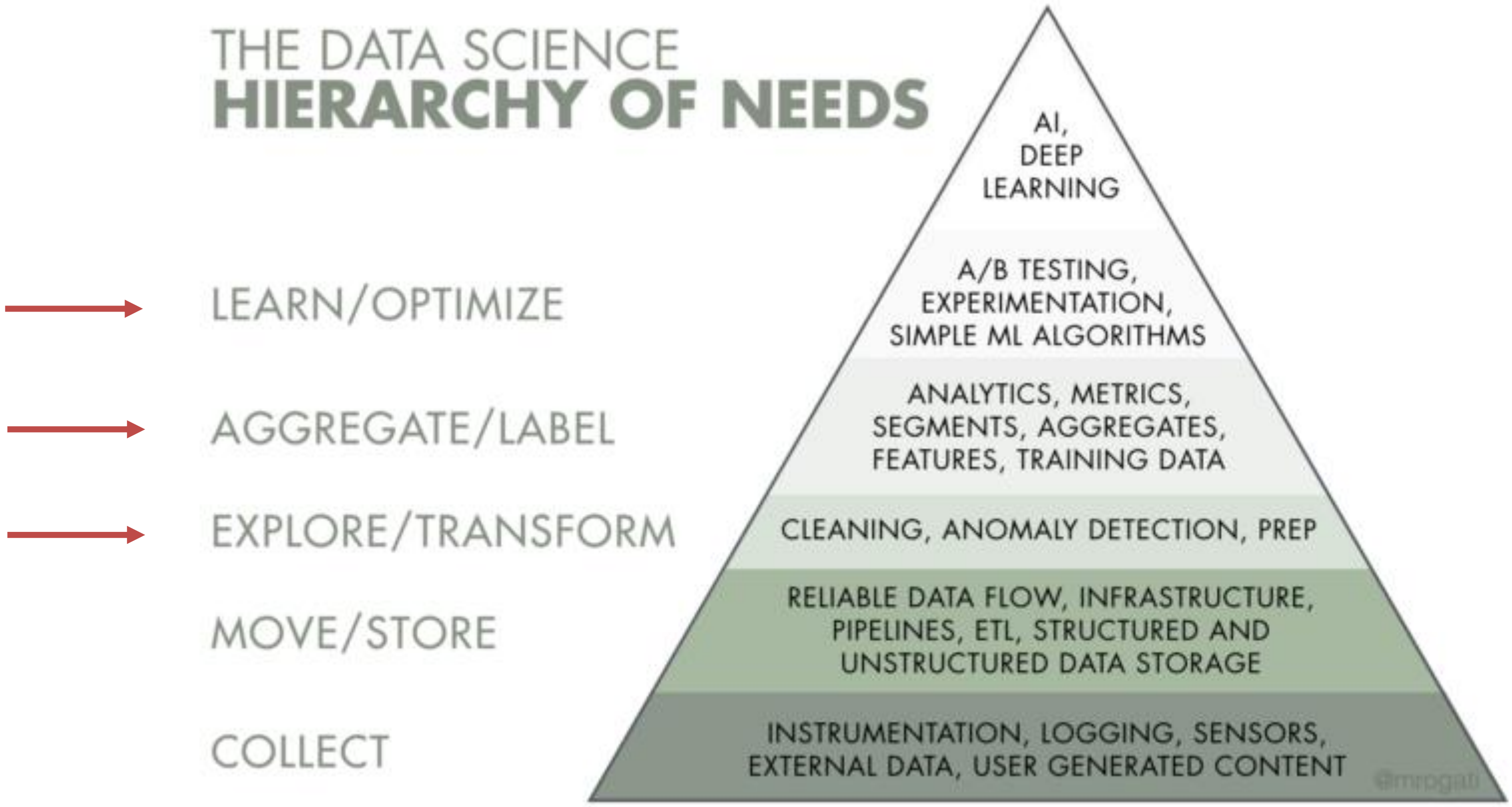
Harrel, F.E. (2005), Regression Modeling Strategies

Zuur, A.F, et al. (2008), Mixed Effects Models and Extensions in Ecology with R

Statistical fundamentals

Why statistics? | Excercises

Why statistics?


THE DATA SCIENCE
HIERARCHY OF NEEDS
LEARN/OPTIMIZE
AGGREGATE/LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
AI, DEEP LEARNING
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
@mrogati

Why statistics?

Definition (Statistics):

Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data.

Statistics is a sub-discipline of mathematics. It consists of three sub-parts:

1. Descriptive statistics
2. Inferential statistics
3. Probability theory

→ Statistics is essential to data science.

Hypothesis testing

Any statistical hypothesis test follows the following scheme:

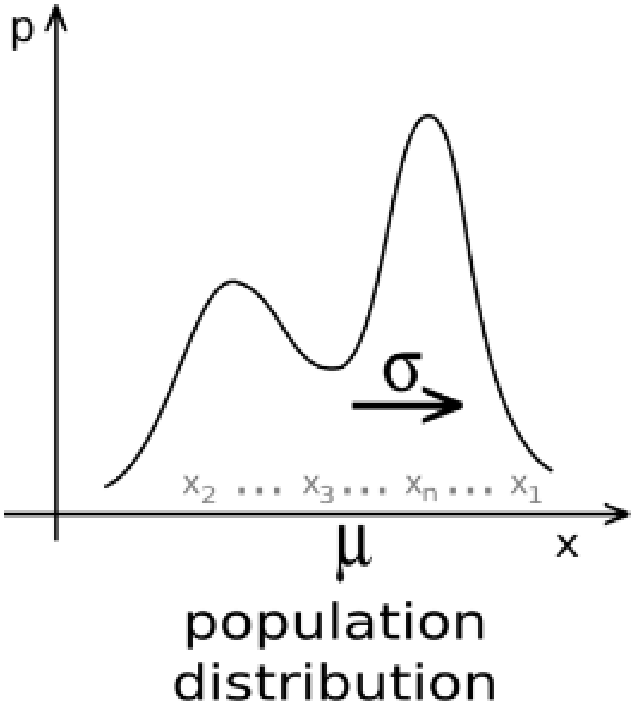
1. Formulate H_0 and H_1.

2. Calculate the value of the test statistic.

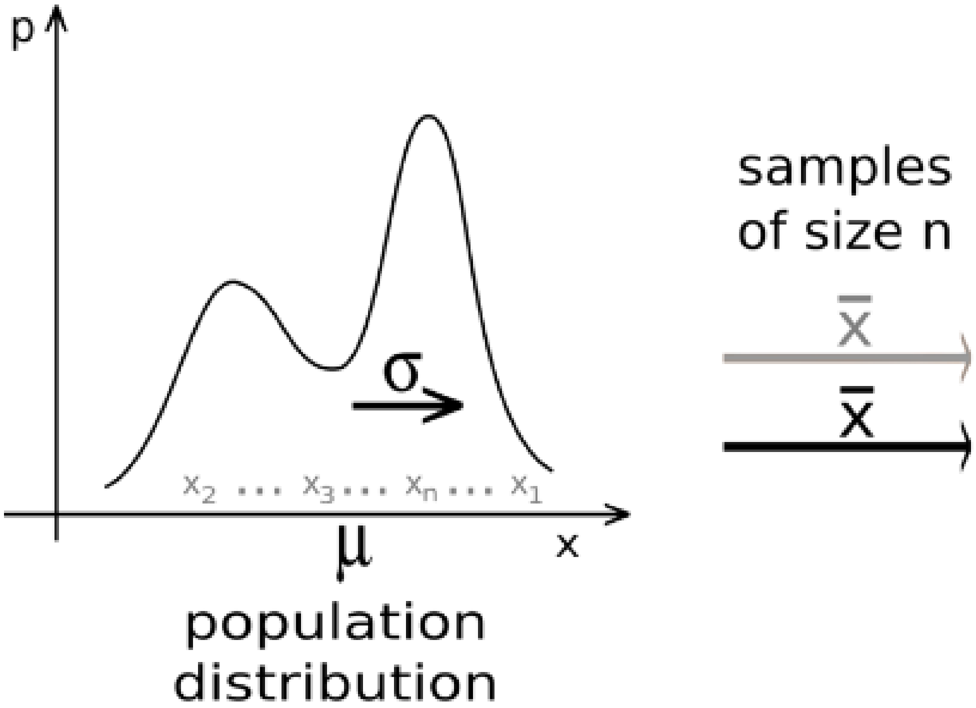
3. Check whether the test statistic is smaller or larger than some threshold value (‚critical value of the test statistic‘), which is determined by the level of significance (type-I-error probability) and the degrees of freedom.

4. Reject H_0 or fail to reject it.

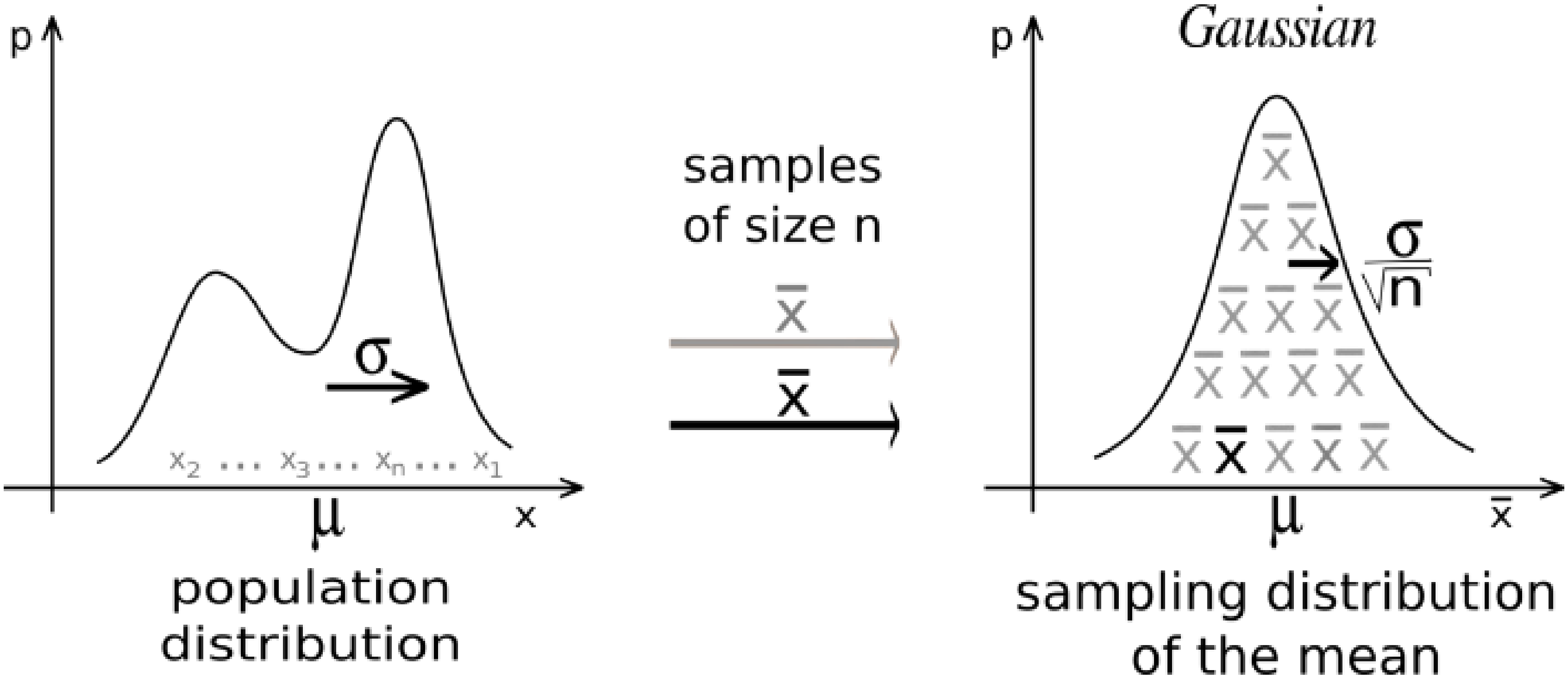
The Central Limit Theorem (CLT)



The Central Limit Theorem (CLT)



The Central Limit Theorem (CLT)


p
σ
x_2 ... x_3 ... x_n ... x_1
μ
x
population distribution
samples of size n
$\bar{x}$
$\bar{x}$
Gaussian
p
$\frac{\sigma}{\sqrt{n}}$
$\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$
μ
$\bar{x}$
sampling distribution of the mean

The Central Limit Theorem (CLT)

Given a sufficiently large sample size from a population with a finite level of variance and mean, the mean of all samples from the same population will tend towards the mean of the population. All samples (the sampling distribution of the mean) will be normally distributed, with all variances being equal to the variance of the population.

- Works well for any population distribution with well-defined mean and variance
- **Beware:** Not all population distributions are ‚well-behaved' in this sense (see excercise #3)!

The Pareto distribution

- Example for a statistical distribution where the CLT does <u>not</u> hold (as we will see in excercise 3)
- Mean and variance do not exist: there is no ‚typical‘ value for something that follows a Pareto distribution.
- Examples:
  - City sizes in a country
  - Diameters of meteors in space
  - Standardized price returns on individual stocks or the stock market in general

Excercise 1: Type-I and type-II-errors (15‘)

Assume you want to assess with a statistical test whether some data are normally distributed. What test would you use? Briefly discuss with your neighbor how sample size will likely affect the frequency of errors of type-I and type-II.

Let‘s now test our suspicion. Find out how to create a vector of *N* random normal numbers and run a series of tests for normality

a.) on 1000 different vectors of random normal numbers of lengths 10, 100, 1000

b.) on 1000 different vectors of random non-normal numbers of lengths 10, 100, 1000. You may choose any non-normal distribution you like.

Hint: Loops will help, but they are not the focal point of this excercise.

Excercise 2: Simple tests (15‘)

Let us investigate the data set by Nitish Ghosal containing data on 28 variables about 5043 movies

(https://raw.githubusercontent.com/nitishghosal/IMDB-Data-Analysis/master/movie_metadata.csv)

a.) Load and investigate the data set.

b.) Did James Cameron and Gore Verbinski have significantly different budgets at their disposal when directing their movies? What about the variability of their respective budgets? Answer the same questions for the gross of their movies.

c.) Repeat the excercise for directors Stanley Kubrick and Sergio Leone. Is it sensible to run the tests from b.) in this case?

Excercise 3: The limits of the CLT (15‘)

Let‘s test the limits of the CLT.

a.) Take any two non-normal distributions you would like and draw *k* = 1000 samples of size *N* = 100 from it. For each distribution, test the null hypothesis that the sampling distribution of the mean is normally distributed, and take a look at the histogram of the sampling distribution of the mean.

b.) Install and load the package `actuar.` Repeat excercise a.) for samples taken from a Pareto distribution using the `rpareto()` function. Does the prediction of the CLT still hold? Why (not)?

Bonus question: Repeat part b. using random Cauchy numbers using `rcauchy().`