

2AMS10: Longitudinal Data Analysis 2022-2023 Introduction

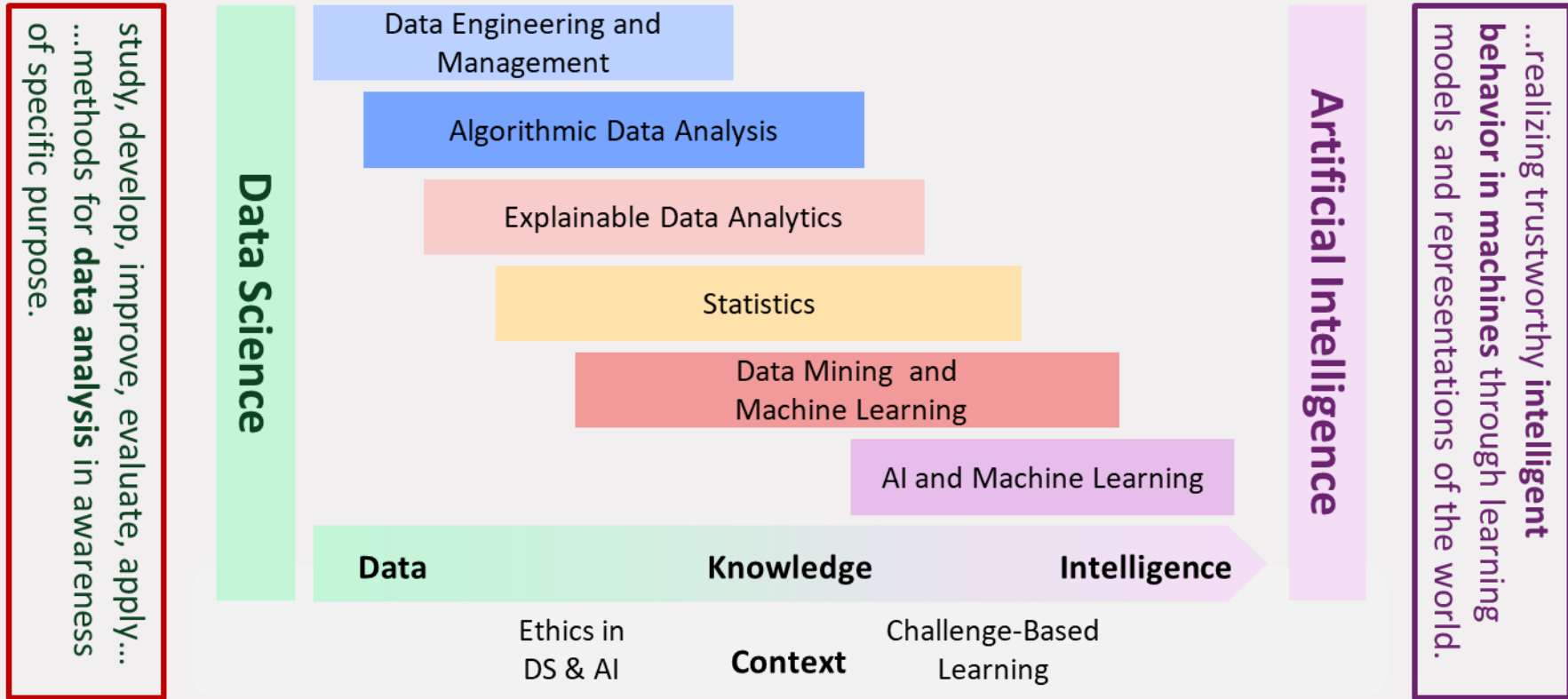
September 6, 2023

Edwin van den Heuvel

e.v.d.heuvel@tue.nl

Professor in Statistics

Master DS&AI



Master DS&AI

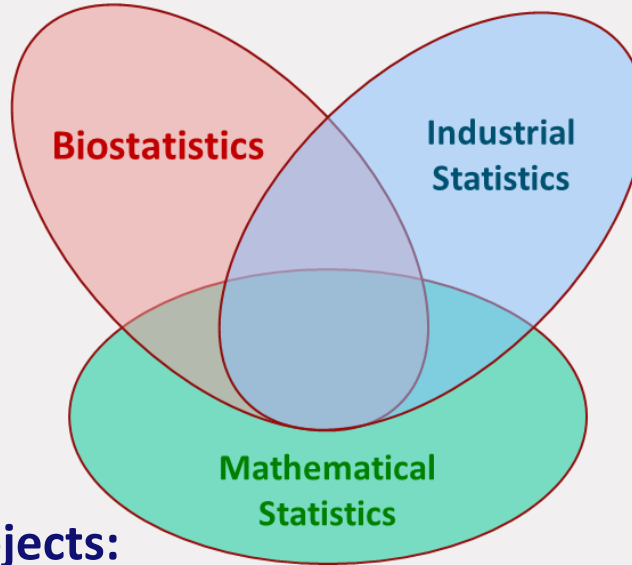
Statistics as learning line

Statistics learning line:

- Longitudinal data
- Statistical learning
- Statistics for big data
- Time series analysis

Internships and master projects:

- Many connections with society
 - Industrial (e.g., Pharma, Insurance, Philips, ASML, DSM)
 - Institutes (e.g., RIVM, CBS, IKNL, ICMS, eMTIC, TNO, Hospitals)
 - Universities (e.g., BU, McMaster, UMCg)
- Both mathematical and data analytic research topics
- Research consultancy activities



Statistical Topics:

- Causality
- Clinical Trials
- Clustering
- Control charts
- Copulas
- Distributed/federated analysis
- Equivalence testing
- Experimental designs
- Hypothesis testing
- Item response theory
- Measurement reliability
- Meta-analysis
- Mixed models
- Missing data analysis
- Non-parametric statistics
- Outlier detection
- Regression
- Reliability Analysis
- Sampling
- Signal process analysis
- Statistical learning
- Statistical process control
- Survival analysis
- Time series analysis

Course LDA

Learning goals

- Analyzing and modeling longitudinal data
[Analysis of Variance, Latent Variable Models, (Linear) Mixed Models]
- Investigate goodness-of-fit of a model
[Outlier Detection, Residual Analysis, Influential Data/Units]
- Estimate & select models
[ML, REML, Likelihood Ratio Test, Information Criteria]
- Formulate and test hypotheses
[Model Parameters; Trajectory Changes; Multiplicity]
- Quantify different sources of variation
[Correlation, Confidence Intervals, Variance Components, Time Changes]

Course LDA

Learning goals

Mathematical Skills:

- Deriving mathematical aspects of estimators and test statistics
 - Moments (at least the first two)
 - Asymptotics (large samples)
 - Measures of effect sizes
- Simulation approaches to mimic real data sets with known probabilistic characteristics
- Studying and comparing performances of methods on simulated data

Data Analytic Skills:

- Comparing performances of *multiple* methods on real data
 - Resampling and data elimination
- Balancing the (potentially conflicting) results of multiple methods on data to draw proper conclusions
- Evaluating underlying statistical assumptions for the applied methods
- Evaluating the fit of models on real data (using different measures)

Course LDA

Learning goals

Working with SAS Software:

- Many companies make use of SAS
 - Pharmaceutical companies
 - Insurance companies
 - Banks
- SAS is very strong in mixed effects models for longitudinal data
 - R and Python are less rich
- SAS has excellent documentation and user groups
- Data scientists should be able to work with SAS, R, and Python






Computational Statistics & Data Analysis

Volume 77, September 2014, Pages 70-83



GEE for longitudinal ordinal data: Comparing R-geepack, R-multgee, R-repolr, SAS-GENMOD, SPSS-GENLIN

Nazanin Noorae^a , Geert Molenberghs^{b, c} , Edwin R. van den Heuvel^a 

- A form of validation study to better understand the performance of packages
- R-multgee, R-repolr, and SPSS performed best (bias, MSE, coverage probability)
- R-geepack failed completely

Course LDA

Academic year 2022-2023

Studying for course:

- Join lectures and ask questions:
 - Tuesday's: 08:45 – 10:30
 - Friday's: 13:30 – 15:15
- Make use of instructions
 - Tuesday's: 10:45 – 12:30
 - Friday's: 15:30 – 17:15
- Study available material
 - Three sets of lecture slides with dense information
 - Additional reading (tutorial papers, lecture notes, books)
 - Exercises (not all worked out in detail)

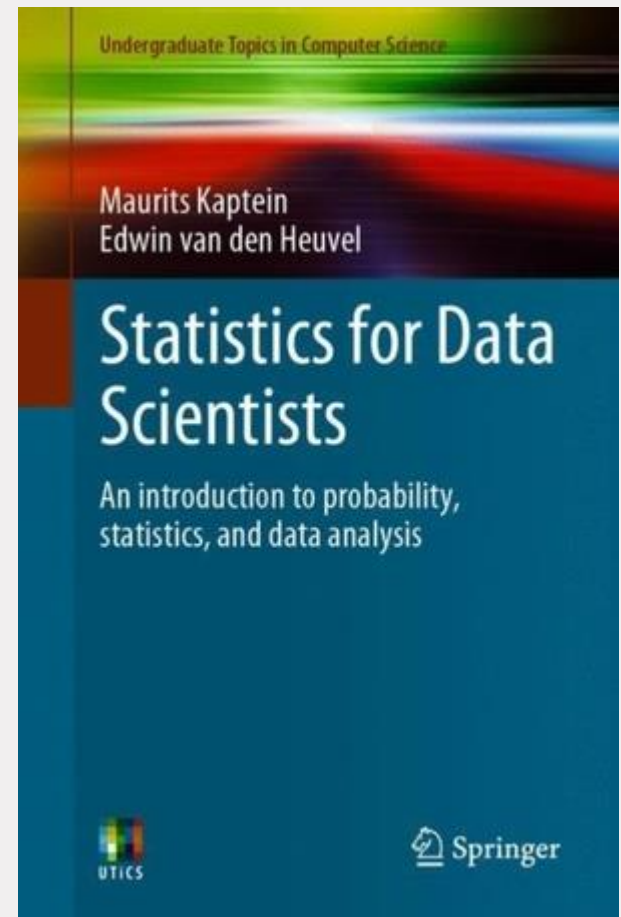
Date	Topic
06/09	Introduction + Background
09/09	One-way random effects ANOVA
13/09	One-way random effects ANOVA
16/09	One-way random effects ANOVA
20/09	Higher-order ANOVA
23/09	Higher-order ANOVA
27/09	Higher-order ANOVA
30/09	Momentum (No teaching)
04/10	Linear mixed effects models
07/10	Subject-specific models
11/10	Subject-specific models
14/10	Marginal models
18/10	Marginal models
21/10	Model selection
25/10	Model selection

Course LDA

Academic year 2022-2023

Prior knowledge: Basic statistical theory

- A first look at data
 - Sampling plans and estimates
 - Probability theory
 - Random variables and distributions
 - Estimation
 - Multiple random variables
 - Making decisions in uncertainty
 - Bayesian statistics
-
- Book can be downloaded from TU/e
 - Lecture notes contain a short summary



Course LDA

Academic year 2022-2023

Expectations from students:

- Assignment Data Analysis
 - Groups of 4 to 6 students:
<http://canvas.tue.nl> ► People ► Groups
 - Evaluation: Scientific reporting
 - Topic and deadline to be determined
 - Will be reported soon
- Exam – written (theory and data)
 - November 4: 13:30 – 16:30
- SAS Software – OnDemand
 - <https://welcome.oda.sas.com/home>
 - SAS tutorial document available
 - Lecture notes contain basic information

Grading:

- Assignment: $A_1 \in [1,10]$
 - **No possibility for retake of assignment!**
 - Expected that everybody contributes, don't let your team members hanging
- Written examination:
 - Based on longitudinal data set
 - Questions related to analysis and theory
 - Most likely requires computer
 - Open book: Slides and notes can be used
 - $E_1 \in [1,10]$
- Calculation grade: $[0.3A_1 + 0.7E_1]$
 - Conditionally on $E_1 > 5.0$

Teacher team

Introducing ourselves



Zhuozhao Zhan

Assistant professor in statistics at TU/e with focus on survival analysis

Has been a teacher at TU/e for several data statistics courses at BSc and MSc level



Sandra Keizer

PhD student in statistics with focus on joint models and causality

Has been a tutor at TU/e for several statistics courses similar to LDA



Tim Engels

PhD student in Operations Research with focus on warehousing

Has been a student assistant at TU/e for several courses similar to LDA



Edwin van den Heuvel

Professor in statistics with focus on meta-analysis, harmonization, LDA, and causality
Responsible for LDA, Statistics for Big Data, Advanced Statistical Modeling, Data Statistics



Hans de Ferrante

PhD student with focus on LDA and casual modeling in organ transplantation

Has been a tutor on bioinformatics, data science, and statistics courses at TU/e and VU



Ray Rounak

PhD student in probability with focus on percolation on dynamic random graph

Is looking for his first experience in tutoring data science courses



Benoit Corsini

Postdoc focusing on random models of graphs, trees and permutations

He has tutored various undergraduate level courses (statistics, probability) outside TU/e

Statistics

The oldest data science field

Historical perspective:

- Originated in 17th and 18th century
 - Collection and analysis of data for understanding developments of countries
 - Founded on probability theory
 - Focus on addressing societal problems
- Matured in the 19th and 20th century
 - Developed as a mathematical discipline
 - Focus on mathematical characterization and optimality given certain assumptions
- Statistics for 21st century is
 - About societal problems and data,
 - Maintaining its mathematical strength

Brief History of Statistics



- ✧ The Word statistics have been derived from Latin word “**Status**” or the Italian word “**Statista**”, meaning of these words is “**Political State**” or a Government.
- ✧ Shakespeare used a word Statist in his drama Hamlet (1602). In the past, the statistics was used by rulers.
- ✧ The application of statistics was very limited but rulers and kings needed information about lands, agriculture, commerce, population of their states to assess their military potential, their wealth, taxation and other aspects of government.

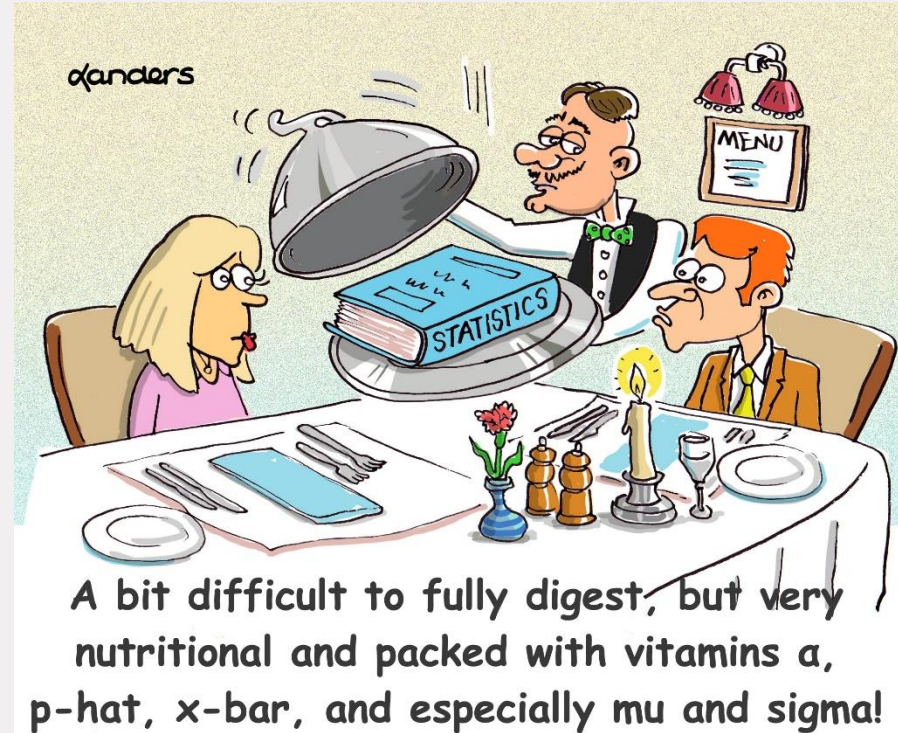
Statistical inference: Bridging data and society

Statistics

The oldest data science field

What is statistics?

- Uses mathematical models with a probabilistic element to
 - Describe co-relations between variables
 - Quantify variability within data
 - Describe populations beyond the data
- Statistical methods and models provide data features that
 - Extract information present in data
 - Are estimated from the data
- May provide a systemic view which is underrepresented in present day

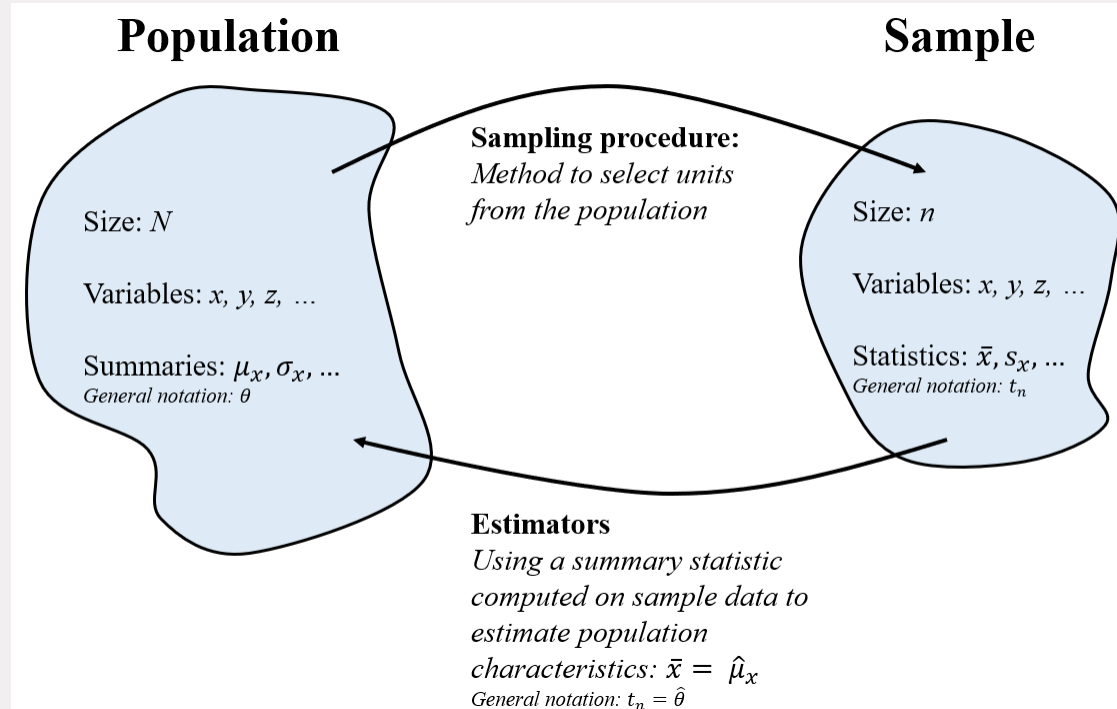


Statistics

Principles of Statistics

Probability Sampling:

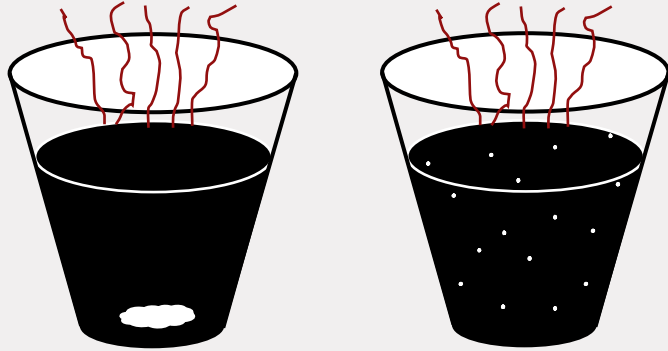
- A known probability mechanism is used to collect units from the population for the sample
- Sampling makes sure that the data is representative
- Then the sample can be used to make statements about the population
- Different sampling procedures require different analysis approaches



Statistics

Principles of Statistics

Example: Sugar in coffee:



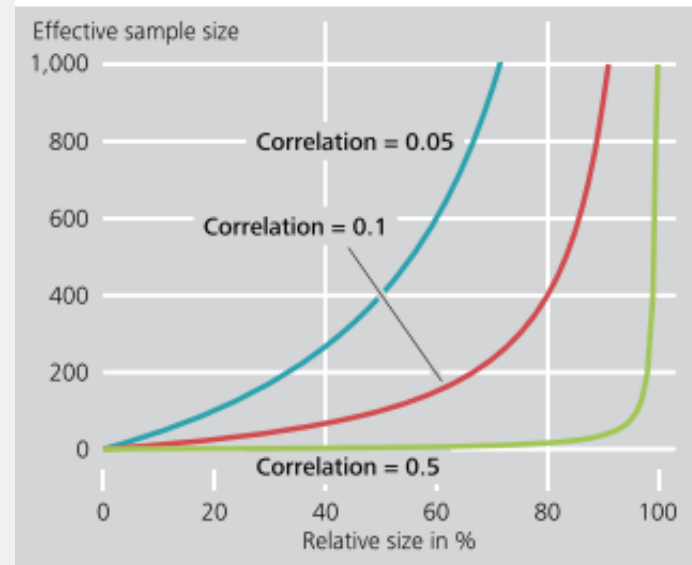
Not stirred

Stirred

- Take one sip of the hot coffee, then for which cup could you taste the sugar?
- Stirring is like probability sampling

Non-Probability Sampling:

- The role of the correlation between outcome and selection mechanism

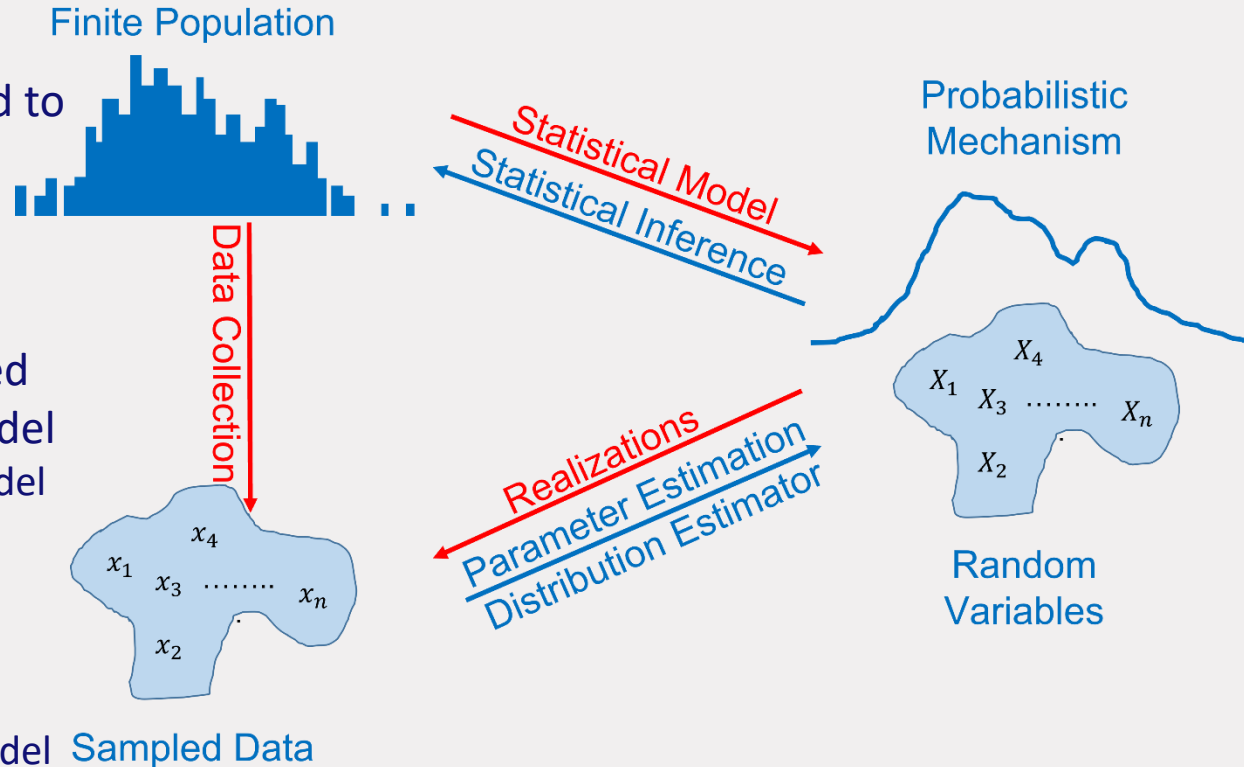


Statistics

Principles of Statistics

Data Mechanisms:

- Statistical models are used to
 - Describe populations
 - Possible mechanisms that has generated the data
- Statements about the population are transitioned through the statistical model
 - Estimate the statistical model
 - Conduct inference
- Ways of evaluating the selected statistical model
 - Goodness-of-Fit
 - Prediction of data with model



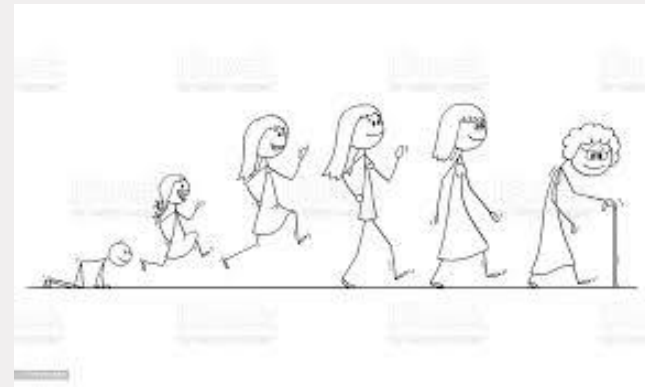
Longitudinal data

The analysis of change



Longitudinal data

The analysis of change: heterogeneity



Within group
variability

Between
group
variability

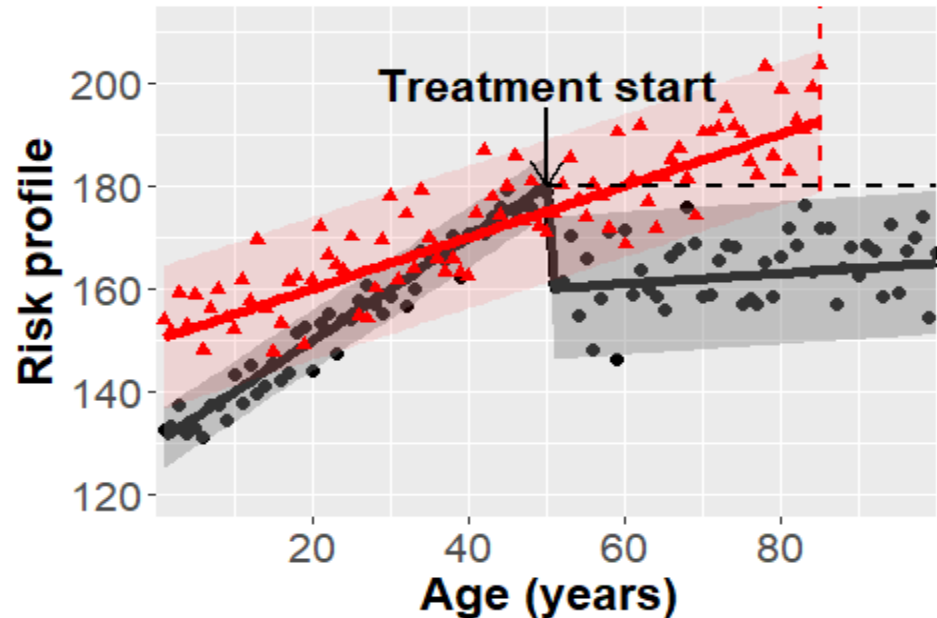


Longitudinal data

Definition and goals

Definition: Data that is collected over time on the same (set of) unit(s)

- Units could be people, animals, processes, networks, etc.
- Data could be measurements, images, text messages, etc.
- The goal of longitudinal data is to understand change over time
 - At what moment did change happen?
 - Is the change gradual or abrupt?
 - Does change happen at different moments for different units?
 - What factors are related to change?
 - Are there differences in time profiles between groups of units?
 - How are different time profiles related to certain events?



Longitudinal data

Questions in medical and epidemiological sciences

What leads to disease? Time-varying covariates & long-term outcome

- Blood pressure profiles with a cardiovascular disease outcome
- Cognitive performance profiles and the risk of dementia and Parkinson
- Social economic status patterns and long-term health status (e.g., quality of life, healthcare utilization)
- Modeling Question: *How do we connect time-dynamic risk factors with disease or health outcomes?*

How does disease progress? Baseline covariates with a time-varying outcome

- Progression of Dupuytren disease (total passive extension deficit) and its relation to (early-life) risk factors
- Number of psychological episodes and treatment modalities
- Recurrent respiratory papillomatosis and the type of virus

Modeling Question: *What is causing or predicting worse disease patterns or profiles?*

Longitudinal data

Type of studies

<u>Individual</u>	<u>Baseline variables</u>	<u>Time-varying variable</u>	<u>End point</u>
	t_0	$t_0 \quad t_1 \quad \dots \quad t_m$	
1	$X_{11} \quad X_{12} \quad \dots \quad X_{1p}$	$Y_{10} \quad Y_{11} \dots Y_{1m}$	D_1/T_1
2	$X_{21} \quad X_{22} \quad \dots \quad X_{2p}$	$Y_{20} \quad Y_{21} \dots Y_{2m}$	D_2/T_2
n	$X_{n1} \quad X_{n2} \quad \dots \quad X_{np}$	$Y_{n0} \quad Y_{n1} \dots Y_{nm}$	D_n/T_n
How does disease progress?		What leads to disease event?	

- Complexities in Modeling:

- Y_{it} is considered a disease response and it is non-normally distributed (e.g., binary)
- Y_{it} is high-frequent (e.g., wearables) and is meant as input for the end point
- Y_{it} is multivariate distributed (diastolic and systolic blood pressure)
- Disease event D_i or time to event T_i is repeated over time
- Sparsity: many baseline covariates (e.g., genetics, metabolites, proteomics)
- Missing data issues, in particular when missingness is not at random

Longitudinal data

Type of studies

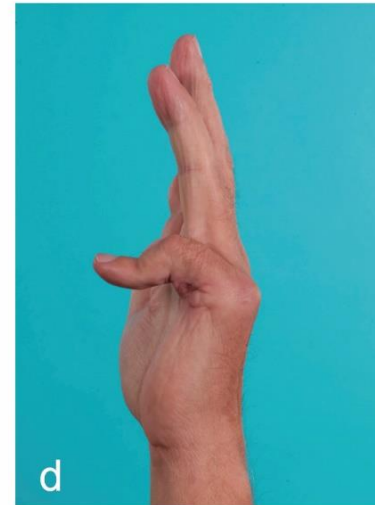
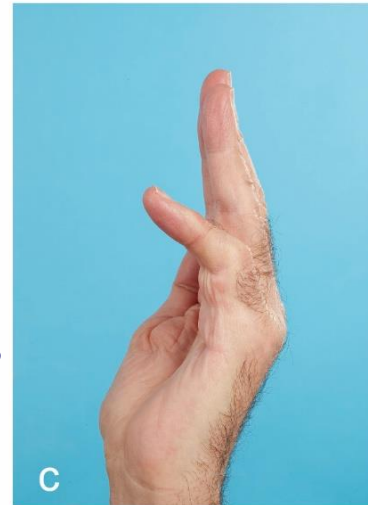
<u>Individual</u>	<u>Time-varying covariates</u>	<u>Time-varying response</u>	<u>Repeated end points</u>
1	$t_0 \quad t_1 \quad \dots \quad t_m$ $X_{10} \quad X_{11} \quad \dots \quad X_{1m}$	$t_0 \quad t_1 \quad \dots \quad t_m$ $Y_{10} \quad Y_{11} \quad \dots \quad Y_{1m}$	$T_{11}, T_{12}, \dots, T_{1r_1}$
2	$X_{20} \quad X_{21} \quad \dots \quad X_{2m}$	$Y_{20} \quad Y_{21} \quad \dots \quad Y_{2m}$	$T_{21}, T_{22}, \dots, T_{2r_2}$
n	$X_{n0} \quad X_{n1} \quad \dots \quad X_{nm}$	$Y_{n0} \quad Y_{n1} \quad \dots \quad Y_{nm}$	$T_{n1}, T_{n2}, \dots, T_{nr_n}$
How do risk factors and disease status interact?			
How do risk factors and disease events interact?			

- Example: BMI patterns with blood pressure profiles and repeated CVD events
- Complexities in Modeling:
 - Time to events (T_{ir}) is repeated in combinations with dynamic variables
 - Covariates (X_{it}) and/or Response (Y_{it}) is multi-dimensional
 - Missing data issues, in particular when missingness is not at random

Longitudinal data

Example Dupuytren Disease (DD)

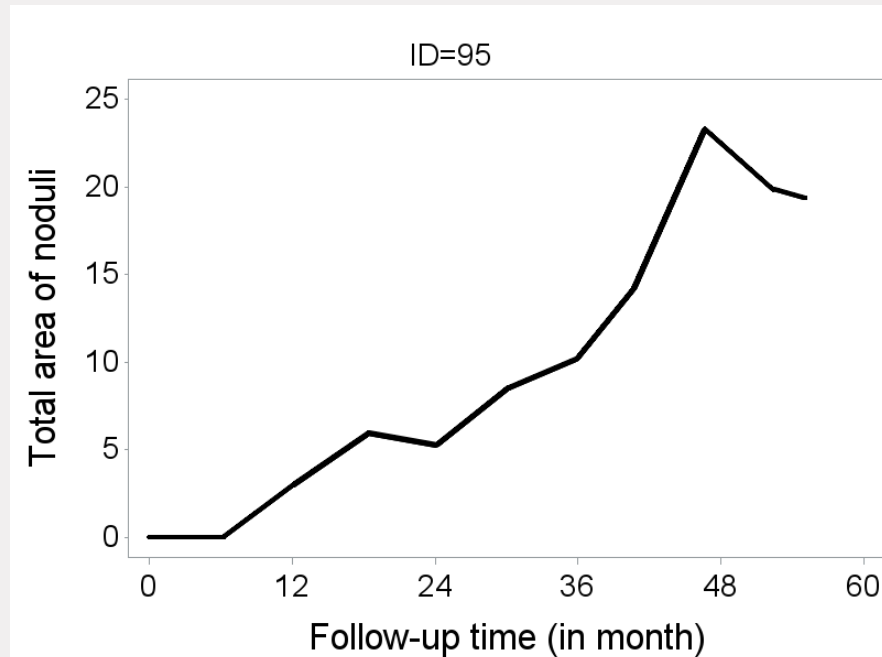
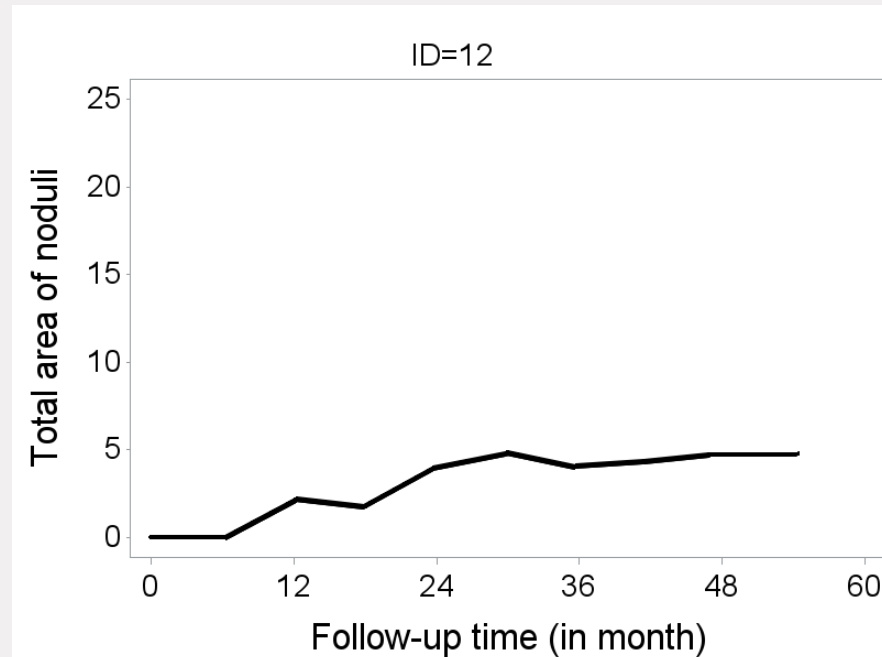
- DD is a fibroproliferative disease of the palmar fascias of the hand.
- DD causes formation of nodules that can
 - Progress into cords
 - Lead to flexion contractures of affected fingers
- It is mainly an inconvenient disease
- DD is associated with several risk factors:
e.g., Age, Sex, Smoking, Alcohol, Diabetes
- A longitudinal study is used to investigate disease progression in individuals
 - Surface area of noduli in the hand in cm^2
 - TPED: Angle for flexion contractures in degrees



Longitudinal data

Example Dupuytren Disease (DD)

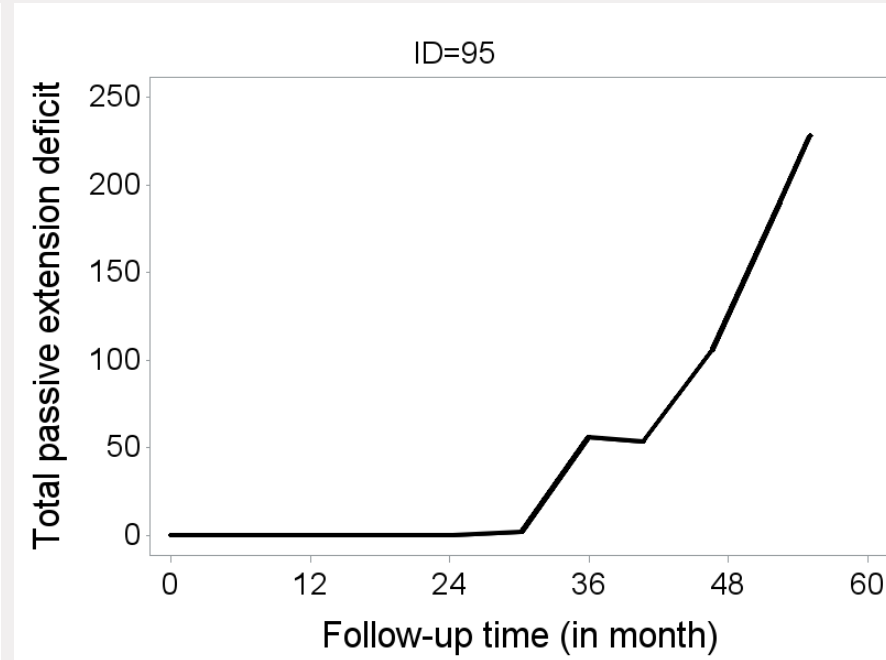
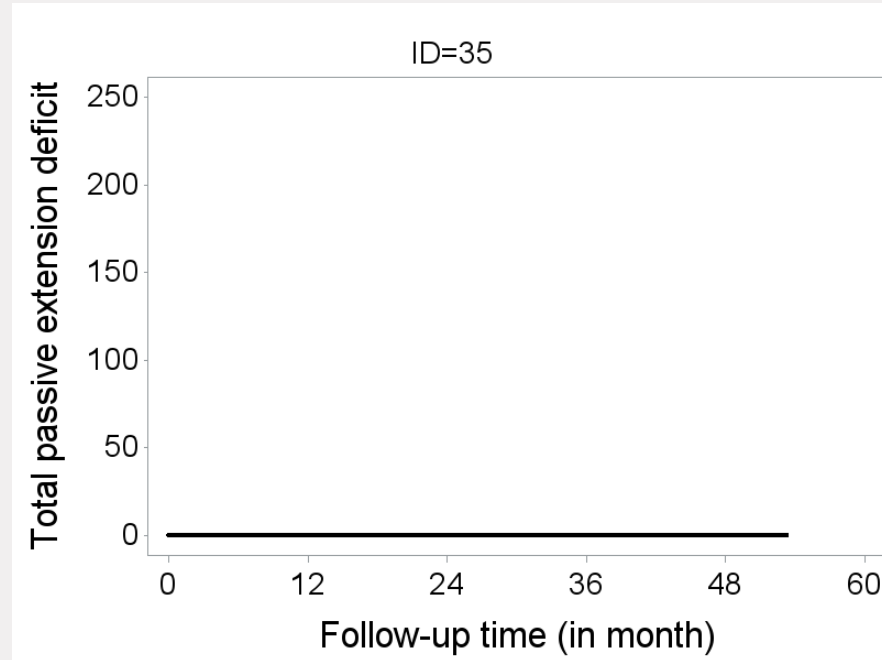
Surface area of noduli:



Longitudinal data

Example Dupuytren Disease (DD)

Total Passive Extension Deficit:



Longitudinal data

Course content academic year 2022-2023

- Core course for new master on Data Science & Artificial Intelligence

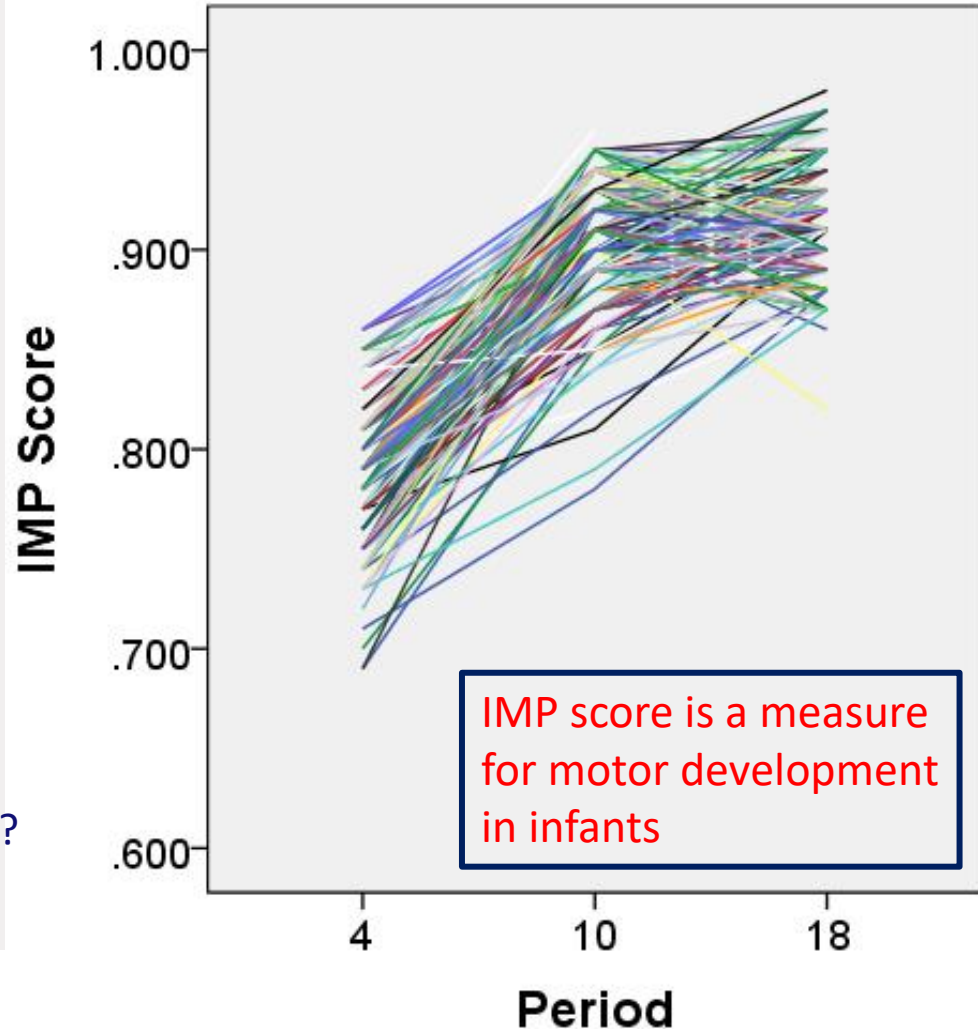
X_{11} Y_{10} X_{21} Y_{20} X_{n1} Y_{n0}	Two-sample t-tests Kruskal-Wallis One-way ANOVA Regression analysis	Y_{10} Y_{11} Y_{20} Y_{21} Y_{n0} Y_{n1}	Paired t-tests Correlations (P,S,K) Confidence intervals Bivariate models	X_{11} Y_{10} Y_{11} X_{21} Y_{20} Y_{21} X_{n1} Y_{n0} Y_{n1}	Mixed effects ANOVA ICC Confidence intervals Heterogeneity
Y_{10} Y_{11} Y_{1m} Y_{20} Y_{21} Y_{2m} Y_{n0} Y_{n1} Y_{nm}	Latent variable models Marginal models Diagnostics Model fit & Goodness-of-fit ML, REML	X_{11} Y_{10} Y_{11} Y_{1m} X_{21} Y_{20} Y_{21} Y_{2m} X_{n1} Y_{n0} Y_{n1} Y_{nm}	Testing time profiles Model selection Non-linear mixed models General linear mixed models Generalized estimating equations		
X_{11} X_{12} X_{1p} Y_{10} Y_{11} Y_{1m} X_{21} X_{22} X_{2p} Y_{20} Y_{21} Y_{2m} X_{n1} X_{n2} X_{np} Y_{n0} Y_{n1} Y_{nm}	Missing data Confounding Causality inference Propensity score	Black is expected background knowledge (study if needed) Blue will be discussed in this course Green is out of scope			

Longitudinal data analysis

Longitudinal time profiles

Definition: *A variable observed over time is described as mathematical function of time*

- Each individual has its own trajectory (**individual-specific**)
- Features of profile can be related to factors and real-life phenomena
- Research questions:
 - Group and quantify differences in the individual trajectories
 - At what time point do we see a change?
 - What factors drive the differences?

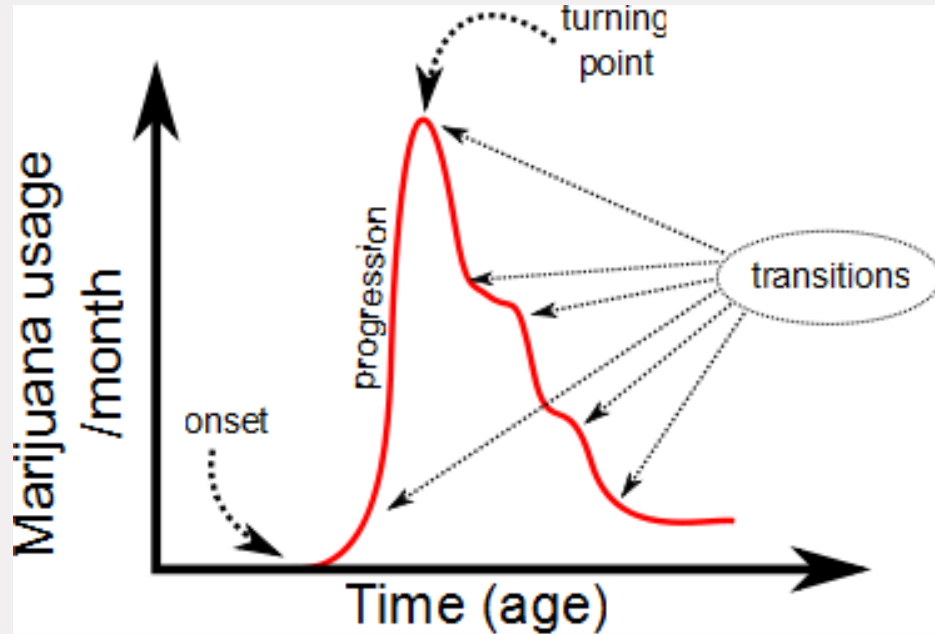


Longitudinal data analysis

Longitudinal time profiles

Critical and sensitive period:

- Identify the timing in which the health outcome changes:
 - Turning points
 - Sudden shifts
 - Progression
 - Regression
- Identify events or risk factors that relate to profile characteristics
- Identify subgroups of participants with similar trajectories
- Can be used with thresholds (BMI)

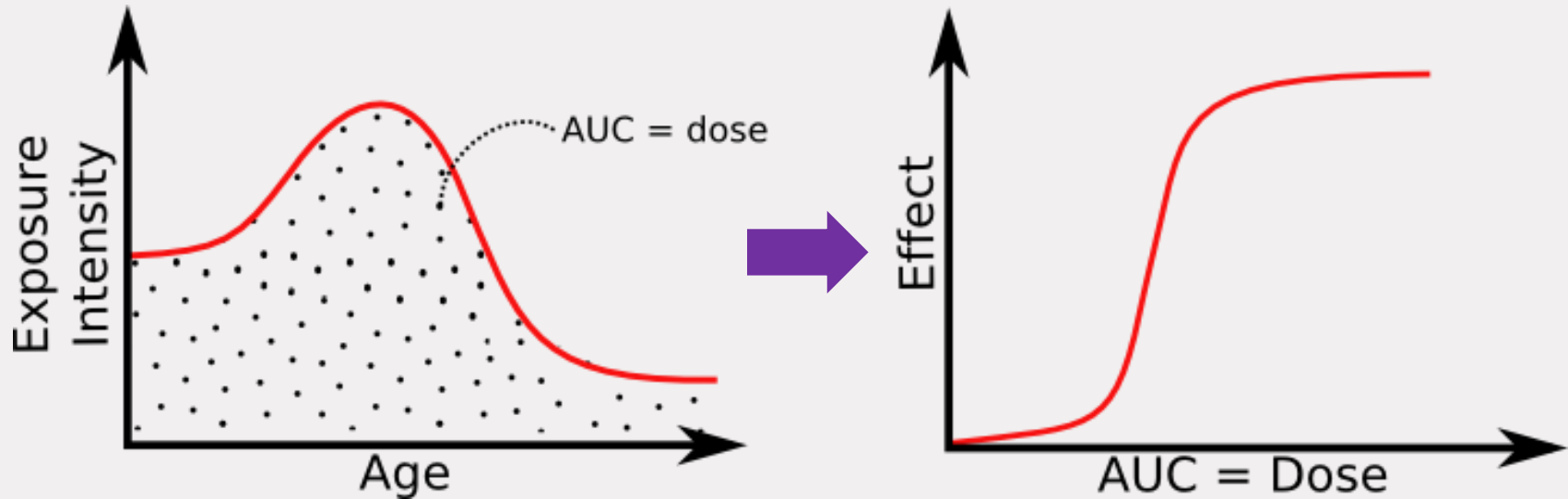


Longitudinal data analysis

Longitudinal time profiles

Accumulation of risk:

- Continuous dose response model
- Area under the curve (Example: smoking/alcohol)

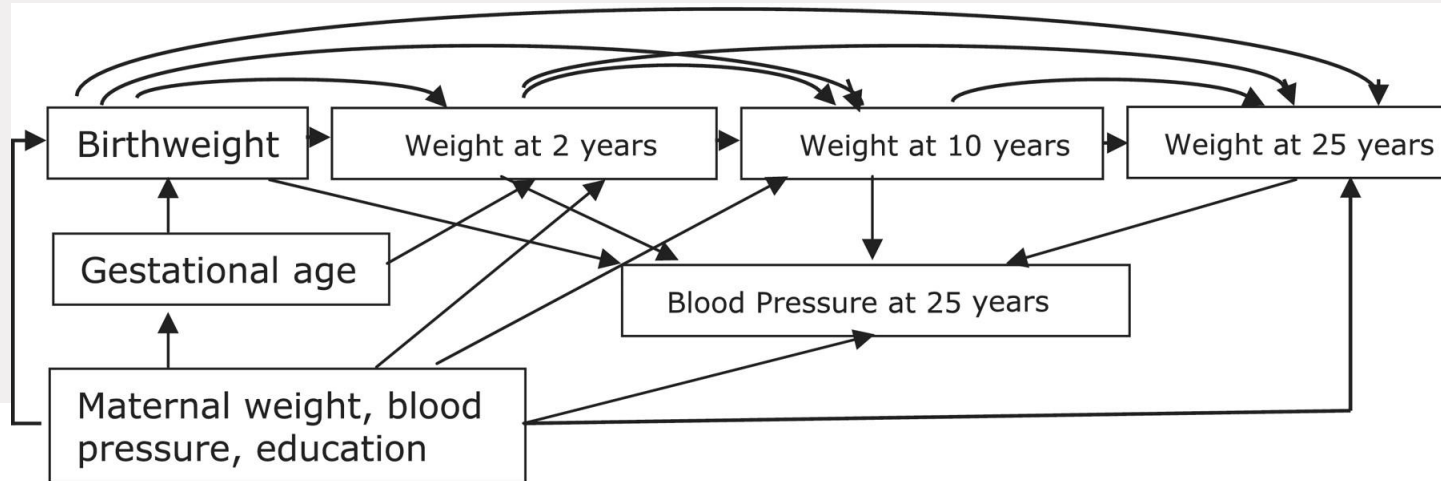


Longitudinal data analysis

Longitudinal pathways

Definition: A variable observed over time is described as mathematical function of variables from previous time points

- Analysis is at a population level (**marginal modeling**)
- Research questions:
 - What paths lead to disease?
 - What factors moderate and mediate the risk of disease?

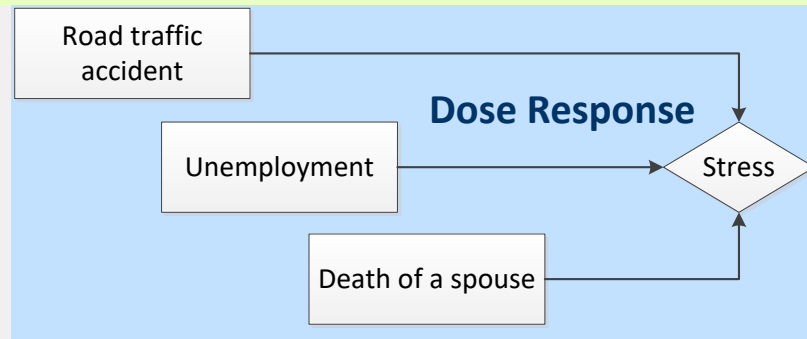
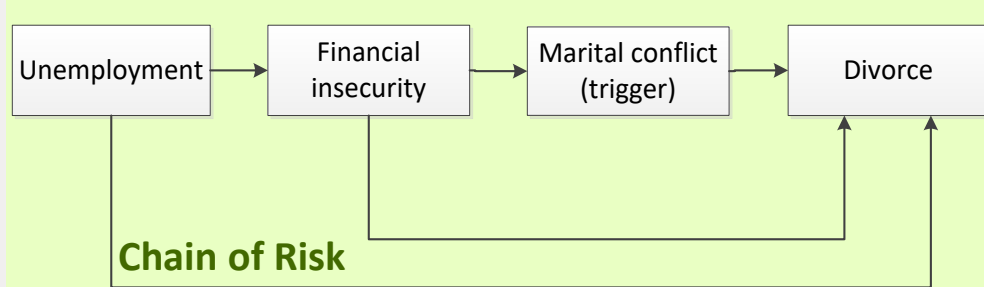


Longitudinal data analysis

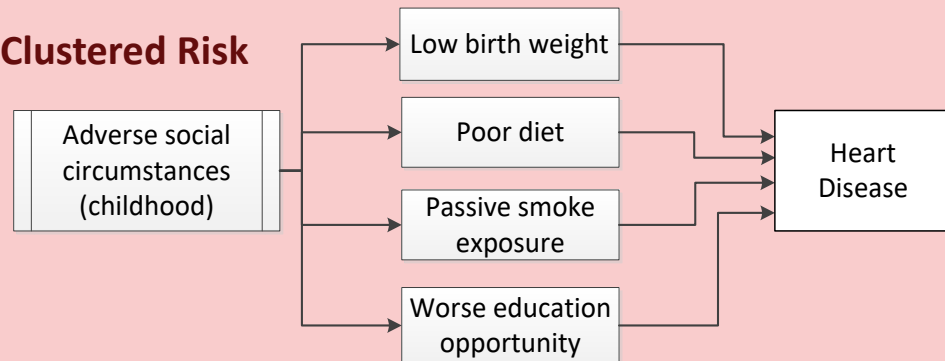
Longitudinal pathways

Accumulation of risk:

- Chain of risk model
 - Risk factors trigger each other
 - Factors may affect disease as well
- Dose response model
 - Risk factors add up
 - Risk factors act typically independently
- Clustered risk model
 - Risk factors add up
 - Risk factors typically occur in a cluster simultaneously



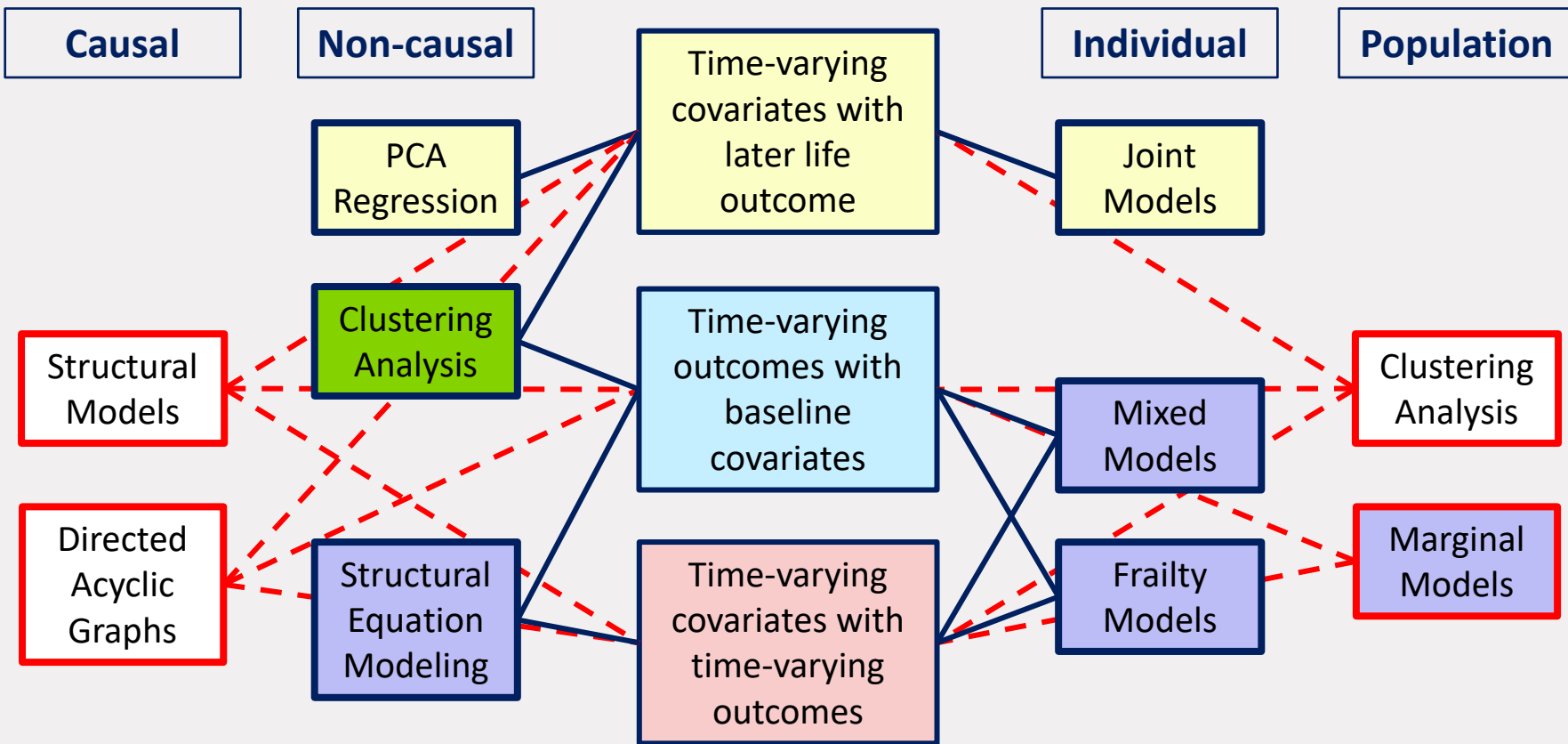
Clustered Risk



Pathway analysis

Study design

Time profile analysis



Longitudinal data analysis

Exercise S1

Get prepared for the course on LDA:

- Read chapter 1 of the lecture notes
 - Overview of statistical theory you should already know
 - Basic introduction to SAS software
 - If needed study the book “Statistics for Data Scientists”
 - If needed study the slides of “Applied Statistics” in CANVAS that uses SAS
- Read the SAS tutorial document
 - Use the IQ_Data_SAS_Tutorial [which is a small part of the school data set we will use for ANOVA]
 - Procedures you should practice with: SORT, MEANS, FREQ, UNIVARIATE
 - Learn how to create and manipulate data sets
- *Use the instruction after the lecture to get help if you need help*

QUESTIONS

