Module 1:

Part 1 - Course introduction

Manuela Zucknick
Oslo Center for Biostatistics and Epidemiology, UiO
manuela.zucknick@medisin.uio.no

MF9130E – Introductory Statistics
April 24, 2023

Course Introduction

Course schedule

Week 1	Monday (24-04)	Tuesday (25-04)	Wednesday (26-04)	Thursday (27-04)	Friday (28-04)
Location	DM Lille auditorium	DM Auditorium 13	DM Store auditorium	DM Runde auditor.	DM Store auditorium
08:30-11:45		C and M SEM*	A SEM*	C SEM*	C SEM*
12:45-16:00	M FOR*	A FOR*	V FOR*	V FOR*	

Week 2	Monday (08-05)	Tuesday (09-05)	Wednesday (10-05)	Thursday (11-05)	Friday (12-05)
Location	DM Lille auditorium	Helga Engs hus Aud 3	DM Auditorium 13	DM Auditorium 13	
08:30-10:00	C FOR	J FOR	M FOR	M FOR	
10:15-11:45	C SEM	J FOR	M and C SEM	M and C SEM	
12:45-14:15	V FOR	M FOR	M FOR	M FOR	
14:30-16:00	V and C SEM	M and C SEM	M and C SEM	M SEM	

Course schedule

Week 1	Monday (24-04)	Tuesday (25-04)	Wednesday (26-04)	Thursday (27-04)	Friday (28-04)
Location	DM Lille auditorium	DM Auditorium 13	DM Store auditorium	DM Runde auditor.	DM Store auditorium
08:30-11:45		C and M SEM*	A SEM*	C SEM*	C SEM*
12:45-16:00	M FOR*	A FOR*	V FOR*	V FOR*	

Week 2	Monday (08-05)	Tuesday (09-05)	Wednesday (10-05)	Thursday (11-05)	Friday (12-05)
Location	DM Lille auditorium	Helga Engs hus Aud 3	DM Auditorium 13	DM Auditorium 13	
08:30-10:00	C FOR	J FOR	M FOR	M FOR	
10:15-11:45	C SEM	J FOR	M and C SEM	M and C SEM	
12:45-14:15	V FOR	M FOR	M FOR	M FOR	
14:30-16:00	V and C SEM	M and C SEM	M and C SEM	M SEM	

Topic
Course introduction; Data and descriptive statistics
Foundations: probability, Bayes law and diagnostic tests, statistical distributions (normal and binomial distribution)
Statistical inference: hypothesis testing and confidence intervals, t-tests, tests for contingency tables
Transformations, non-parametric methods
Sample size and statistical power
Study designs: epidemiological designs and concepts, principles of clinical trials
Regressions: simple and multiple regression, confounding and interactions, linear and logistic regression
Survival analysis; Course summary

Lecturer
M = Manuela Zucknick
A = Alvaro Köhn-Luque
V = Valeria Vitelli
J = Jo S Stenhjem
C = Chi Zhang

Lecture or Lab
FOR = Lecture (classical format)
SEM = Lab (classical format)
FOR* = Lecture (flipped classroom setup)
SEM* = Lab (flipped classroom setup)

Overview for Module 1 "Data and Descriptive Statistics"

This afternoon: Lectures in flipped classroom style (FOR*)

- **Introduction** to this course
- **Data and statistics** in medicine
  - Introduction and motivation
- **Descriptive statistics**
  - Data presentation
  - Central measures
  - Measures of variation
  - Graphical presentation of data

Tomorrow morning: Labs in flipped classroom style (SEM)

- Introduction to statistical computing with **R**

- **Course textbook chapters:**
  - Kirkwood and Sterne chapters 2-4
  - Aalen chapters 1 and 2

Links and Course Material

- **Course webpages:** `https://ocbe-uio.github.io/teaching_mf9130e/`
- We will mainly use the course webpages for all information and access to material. The webpages will be continuously updated throughout the course.

- **Canvas room:** We will not use the Canvas room a lot, but Canvas is used for **emails** and general communication. Please let us know asap, if you do not have access to Canvas!

- **Official UiO course pages** with schedule, literature and details on admission rules, exam etc: `https://www.uio.no/studier/emner/medisin/med/MF9130E/`

Computer exercises in R (starting tomorrow morning)

- You will need to have a laptop computer with access to R and RStudio for the labs.
- We advise that you install R/RStudio on your own laptop.
- Alternatively, you could register for a (free trial) account on a Posit Cloud server.
- See here for instructions: `ocbe-uio.github.io/teaching_mf9130e/get_started/get_started.html`

- Note: You can also access R/RStudio through the UiO Programkiosk: `www.uio.no/english/services/it/home-away/kiosk`.

Homework for tomorrow morning

- Go through the instructions above to get working access to R and RStudio. There will be a detailed introduction to R and RStudio tomorrow morning.

Note

It is recommended to have R and Rstudio installed on your laptop, this is because you have a better control of where you prefer to download data and course material. This is also useful when you want to analyse your own datasets. For example, you might have to upload datafiles to the server for Posit Cloud to work.

However, if there is a problem with the installation, you can use Posit Cloud as an alternative.

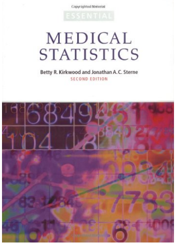
On Tuesday morning we will see if most people can successfully make R run on their laptop and make necessary adjustments.

Exam

- Take-home exam.
- Will be published via Inspera at the end of the course.
- To be submitted within a specified deadline (4 weeks after the end of the course).
- A passed exam is required to get the course approved.

- More details on the last day of this course.

Main course textbook: Kirkwood and Sterne (2003)

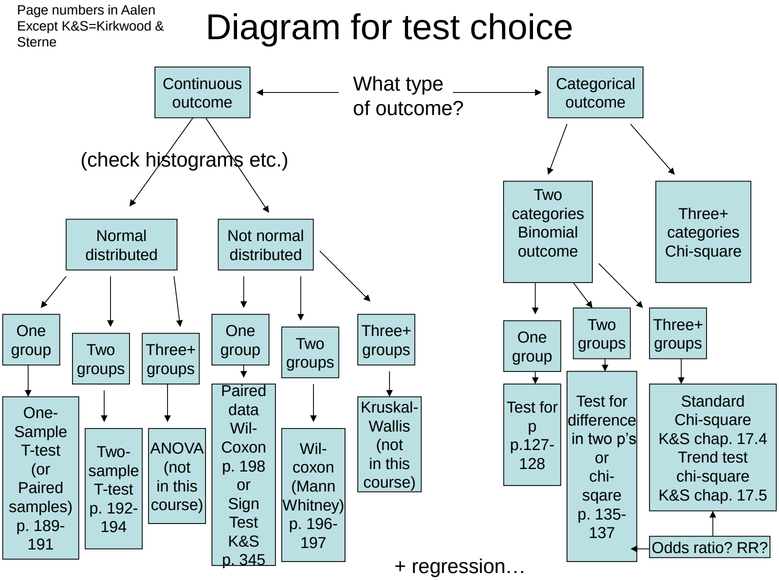


- Betty R. Kirkwood and Jonathan A. C. Sterne. **Essential Medical Statistics**. Second edition, Blackwell Science Ltd, 2003
- `www.blackwellpublishing.com/essentialmedstats/`

Norwegian alternative: Aalen (ed) *et al* (2006)

- Odd O. Aalen (red), Arnoldo Frigessi, Tron Anders Moger, Ida Scheel, Eva Skovlund, Marit B. Veierød. **Statistiske metoder i medisin og helsefag**. Gyldendal Akademisk 2006
- `www.med.uio.no/imb/studier/ressurser/statistikk/statistikkressurser-shs/aalen.html`

Methods in this course



Many of the methods we cover can be seen as **linear models**.

- `https://lindeloev.github.io/tests-as-linear/`
- Regression models as well as most statistical tests:

Common statistical tests are linear models

Last updated: 28 June, 2019.Also check out the *Python version!*

See worked examples and more details at the accompanying notebook: https://lindeloev.github.io/tests-as-linear

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: lm(y ~ 1 + x)	**y is independent of x** P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y_1, y_2, paired=TRUE) wilcox.test(y_1, y_2, paired=TRUE)	lm(y_2 - y_1 ~ 1) lm(signed_rank(y_2 - y_1) ~ 1)	✓ for N >14	One intercept predicts the pairwise y_2-y_1 differences. - (Same, but it predicts the *signed rank* of y_2-y_1.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y_1, y_2, var.equal=TRUE) t.test(y_1, y_2, var.equal=FALSE) wilcox.test(y_1, y_2)	lm(y ~ 1 + G_2)[A] gls(y ~ 1 + G_2, weights=...[B])[A] lm(signed_rank(y) ~ 1 + G_2)[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
Multiple regression: lm(y ~ 1 + x_1 + x_2 + ...)	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	lm(y ~ 1 + G_2 + G_3 +...+ G_N)[A] lm(rank(y) ~ 1 + G_2 + G_3 +...+ G_N)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G_2 + G_3 +...+ G_N + x)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	aov(y ~ group * sex)	lm(y ~ 1 + G_2 + G_3 +...+ G_N + S_2 + S_3 +...+ S_K + G_2*S_2 + G_3*S_3 + ... + G_N*S_K)	✓	Interaction term: changing **sex** changes the **y ~ group** parameters. *Note: **$G_{2 \text{ to } N}$** is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for **$S_{2 \text{ to } K}$** for sex. The first line (with G_i) is main effect of group, the second (with S_i) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.*	*[Coming]*
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	**Equivalent log-linear model** glm(y ~ 1 + G_2 + G_3 + ... + G_N + S_2 + S_3+ ... + S_K + G_2*S_2 + G_3*S_3 +...+ G_N*S_K, family=...)[A]	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments:* `glm(model, family=poisson())` *As linear-model, the Chi-square test is $\log(y_i) = \log(N) + \log(\alpha_i) + \log(\beta_j) + \log(\alpha_i\beta_j)$ where α_i and β_j are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	chisq.test(y)	glm(y ~ 1 + G_2 + G_3 +...+ G_N, family=...)[A]	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation y ~ 1 + x is R shorthand for $y = 1 \cdot b + a \cdot x$ which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_i and S_i are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when $\Delta x = 1$ between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. lm requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at https://lindeloev.github.io/tests-as-linear.

[A] See the note to the two-way ANOVA for explanation of the notation.

[B] Same model, but with one variance per group: `gls(value ~ 1 + G2, weights = varIdent(form = ~1|group), method="ML")`.

Jonas Kristoffer Lindeløv
https://lindeloev.net

Why do we need statistics?

"Statistics is the science of collecting, summarizing, presenting and interpreting data, and of using them to estimate the magnitude of associations and test hypotheses"

Kirkwood and Sterne p. 1

The build-up of a research project

- **Planning**
- **Design**
- **Execution** (data collection)
- **Data analysis**
- **Presentation**
- **Interpretation**
- **Publication**

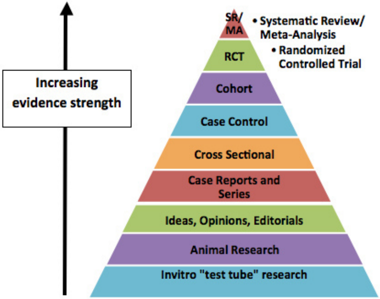
Statistics in <u>all</u> points

Critical reading of publications

- Research **design**
- **Inclusion and exclusion criteria**
- **Sample size**
- **Exposure** (risk factor) and **confounding factors**
- **Outcome** (response)
- Statistical **analysis**
- **Bias**
- **Interpretation** of results

Statistics in all points

Pyramid of evidence



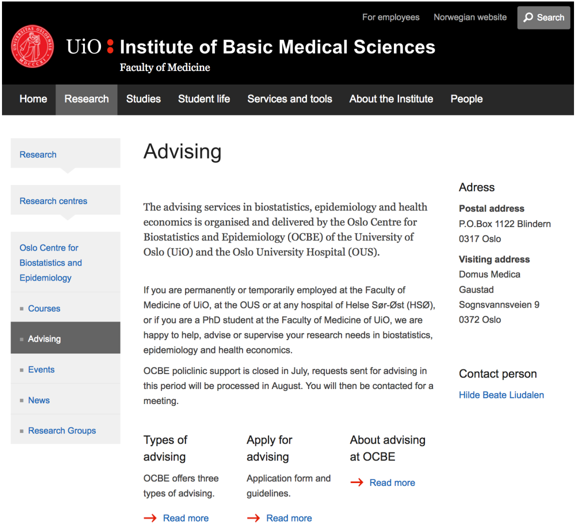
- **Grading the evidence** for practice guidelines after susceptibility of threats to internal validity
- **Health literacy guide** designed to help students find and assess sources of quality health information: `https://libraryguides.unh.edu/c.php?g=326606&p=2191225`

Oslo Centre for Biostatistics and Epidemiology (OCBE)

- ... is a joint centre of UiO (Department of Biostatistics, IMB) and OUS (Biostatistics and Epidemiology group at Forskningsstøtte). Approx. 80 people in total.

- **Research**: Methodological research in several areas, e.g.
  - Statistical genomics
  - Epidemiological research
  - Infectious disease research
  - Clinical Trials Unit (CTU)
- **Statistical advising** for researchers at the Medical Faculty, OUS and Helse-Sør-Øst
- **Teaching** at MedFak: professional study programme for Medicine, Master's programmes in Clinical Nutrition and International Community Health, PhD courses

OCBE Statistical Advising Service


For employees Norwegian website Search

UiO : Institute of Basic Medical Sciences

Faculty of Medicine

Home Research Studies Student life Services and tools About the Institute People

Research

Research centres

Oslo Centre for Biostatistics and Epidemiology

- Courses
- Advising
- Events
- News
- Research Groups

Advising

The advising services in biostatistics, epidemiology and health economics is organised and delivered by the Oslo Centre for Biostatistics and Epidemiology (OCBE) of the University of Oslo (UiO) and the Oslo University Hospital (OUS).

If you are permanently or temporarily employed at the Faculty of Medicine of UiO, at the OUS or at any hospital of Helse Sør-Øst (HSØ), or if you are a PhD student at the Faculty of Medicine of UiO, we are happy to help, advise or supervise your research needs in biostatistics, epidemiology and health economics.

OCBE policlinic support is closed in July, requests sent for advising in this period will be processed in August. You will then be contacted for a meeting.

Types of advising

OCBE offers three types of advising.

→ Read more

Apply for advising

Application form and guidelines.

→ Read more

About advising at OCBE

→ Read more

Adress

Postal address

P.O.Box 1122 Blindern

0317 Oslo

Visiting address

Domus Medica

Gaustad

Sognsvannsveien 9

0372 Oslo

Contact person

Hilde Beate Liudalen

https://www.med.uio.no/imb/english/research/centres/ocbe/advising/