Biostatistics

Mats J. Stensrud

Spring 2024

Plan for today

- Clarify aims and expectations.
- Give a high-level introduction to important topics in biostatistics.
- Key topics
  - causal inference
  - survival analysis

Section 1

Structure of the course

Structure

- 90 minutes lectures every Tuesday 10h15
  I will use the iPad as a digital blackboard
- Moodle is our platform
  - Announcements
  - Links to relevant literature
  - Link to Ed Discussions
  - All questions about the course should be asked on Ed Discussions
- Slides and problem sheets will be uploaded every Tuesday

Exam

- Midterm exam.
- Final exam.

Features of the course

- This is a *statistics* course.
- We will study theory and methods that are relevant to solve common practical problems.
- The course will contain proofs,
  but *all* the results we are using will *not* be proved.
  That said, I will strive to motivate all the results.
- I will also spend time on discussing the interpretation of the results:
  We will take interpretation seriously, and we try to think formally about interpretation.

Goal

After the course, you should be able to:

- Understand mathematical and statistical theory for event history analysis and longitudinal data analysis.
- Furthermore, understand the concepts and ideas that this theory expresses.
- Apply these methods to data (there are ubiquitous applications!).
- Critically evaluate how these methods are used in practice.
- Build on the material in this course to derive new results yourself.

Outline of the course

- Core topics and principles in Biostatistics
- **Time-to-event outcomes** ("survival analysis")
- Longitudinal data analysis

We need a motivation

- Why are you interested in biostatistics?
- What types of questions are you interested in?
- Why do you ask them?

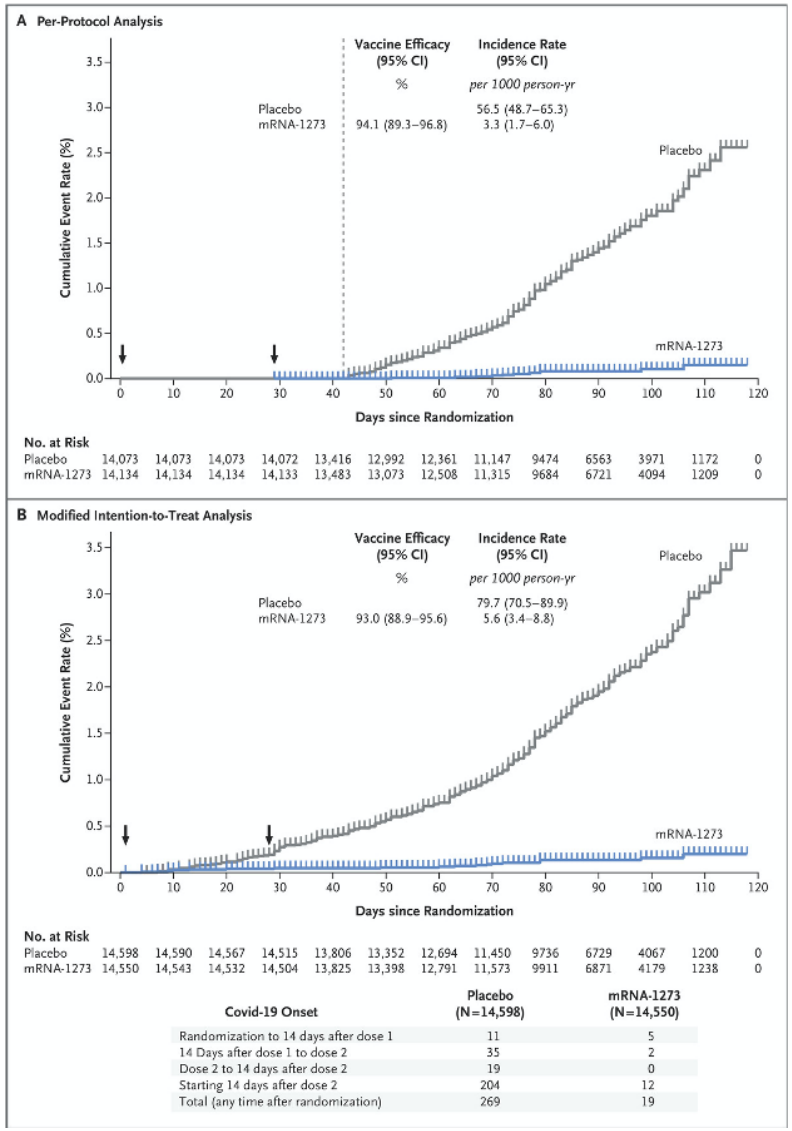
Statistical science (including biostatistics)

"Statistical science involves far more than data – it requires realistic causal models for the generation of that data and the deduction of their empirical consequences."
Sander Greenland. *The causal foundations of applied probability and statistics*. 2020. arXiv: 2011.02677 [stat.OT]

Statistical science (including biostatistics)

"I will argue that realistic and thus scientifically relevant statistical theory is best viewed as a subdomain of causality theory, not a separate entity or an extension of probability. In particular, the application of statistics (and indeed most technology) must deal with causation if it is to represent adequately the underlying reality of how we came to observe what was seen – that is, the causal network leading to the data."
Greenland, *The causal foundations of applied probability and statistics*

The Moderna vaccine



Covid-19 Onset	Placebo (N=14,598)	mRNA-1273 (N=14,550)
Randomization to 14 days after dose 1	11	5
14 Days after dose 1 to dose 2	35	2
Dose 2 to 14 days after dose 2	19	0
Starting 14 days after dose 2	204	12
Total (any time after randomization)	269	19

Figure 1: Classical pharmaceutical study.

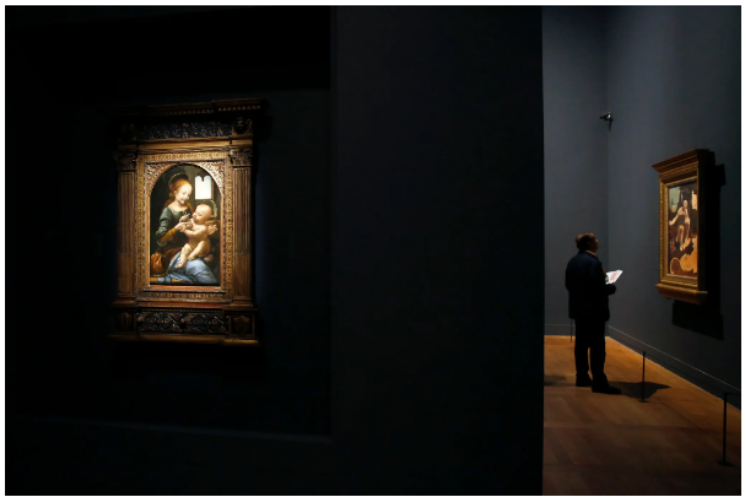
From the New York Times

The New York Times

Another Benefit to Going to Museums? You May Live Longer

Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.

Give this article



Leonardo da Vinci's "The Virgin and Child" at the Louvre in Paris. A British study found that people who engaged with the arts had a lower likelihood of dying during the study period than those who did not. Thibault Camus/Associated Press

Figure 2

Some features of causal inference

- I teach another course, Randomization and Causation (MATH-336), which solely concerns causal inference.
- Understanding causal inference is important in biostatistics.
- I will introduce you to some basic ideas in the beginning of this course. The causal material is related to MATH-336.

Questions

- Descriptive / predictive:
  - "Is this patient at high risk of developing complications during surgery?"

  Causal:
  - "Which type of anaesthetic should this patient receive to reduce the risk of complications during surgery?"
  - "How does the amount of anaesthetic affect the risk of complications during surgery?"
  - "What can be done to reduce the risk of complications during surgery for an average / a particular type of patient?"

Questions

- Descriptive / predictive:
  - "Which type of client will buy which kind of product?"

  Causal:
  - "Should advert be at the top or bottom of website to increase the probability of viewing product?"
  - "How does the size of advert affect the probability of viewing product?"
  - "How can I get a client to buy my product?"

Questions

- Descriptive / predictive:
  - "Who is most likely to become long-term unemployed?"

  Causal:
  - "Will a minimum wage legislation increase the unemployment rate of a country?"
  - "How does the size of advert affect the probability of viewing product?"
  - "What can be done to prevent someone from becoming unemployed?"

Motivation: Consider a table....

Table 1: Data from a study of A (an *exposure*) and Y (an *outcome*).

	$Y = 1$	$Y = 0$
$A = 1$	10	90
$A = 0$	5	95

From the table we could draw *statistical* conclusions.

- Units with $A = 1$ more often have $Y = 1$ than those with $A = 0$.
- We could compute the proportions, p-value etc.
- But we can't draw any *causal* conclusions without more information.

Suppose now that the units are individuals.

Table 2: Data from a study of A (an *exposure*) and Y (an *outcome*).

	$Y = 1$	$Y = 0$
$A = 0$	10	90
$A = 1$	5	95

$A = 1$ is getting surgery, $Y = 1$ indicates survival after 1 year.

- What does the table tell us now?
- Could we infer that surgery reduces the risk of death?

Suppose that we say this was a randomized controlled trial, where A was randomized?

Observational data

Definition (Observational data)

A sample from a population where the treatment (exposure) is not under the control of the researcher.

That is, the treatment (exposure) of interest is not randomly assigned.

Back to our simple table

Table 3: Data from a study of A (an *exposure*) and Y (an *outcome*).

	$Y = 1$	$Y = 0$
$A = 1$	10	90
$A = 0$	5	95

More on the table

	$Y = 1$	$Y = 0$
$A = 1$	10	90
$A = 0$	5	95

a row for each unit, labeled *id*:

id	A	Y
1	1	1
2	1	0
3	0	1
4	1	0
5	0	0
⋮	⋮	⋮
200	0	1

Counterfactuals a.k.a. potential outcomes

- We will posit unobserved fixed potential or counterfactual outcomes[1] for each unit[2] under different treatments[3]

  Hint: It is helpful to think about a counterfactual random variable as a variable that **does exist** in this world, even before interventions take place, but we are not able to observe it.

- We will use superscripts to indicate that a random variable is counterfactual. For example consider a random variable Y. A counterfactual version Y^g is the value Y would have had under an intervention g (also called treatment regime or treatment strategy).
- To get started, in the first lectures, we will consider some simple interventions g which only fixes a binary treatment A to a value $a \in \{0, 1\}$.

[1] I will use the terms "counterfactuals" and "potential outcomes" interchangeably.
[2] I will use the terms "unit", "subject" and "individual" interchangeably.
[3] I will use the terms "treatment" and "exposure" interchangeably.

Let's go further: Potential outcomes

Each potential outcome Y^a answers the question "What would have happened to outcome Y, if A had taken on a specific value a?" We can expand the table to include the potential outcome:

id	A	Y	Y^1	Y^0
1	1	1	1	?
2	1	0	0	?
3	0	1	?	1
4	1	0	0	?
5	0	0	?	0
⋮	⋮	⋮	⋮	⋮
200	0	1	?	1

Table 4: Data from the same study of A (an *exposure*) and Y (an *outcome*), with the addition of the potential outcome variables.

That is, $Y = I(A = 0)Y^{a=0} + I(A = 1)Y^{a=1}$.

Why randomisation is the gold standard

- In a randomised experiment, the treatment is assigned independently of all other factors (e.g. by a coin flip or a random number generator).
- In a randomised experiment one of the counterfactual outcomes $Y^{a=0}$ or $Y^{a=1}$ is unobserved.
- However, randomisation ensures that it is *random* whether $Y^{a=0}$ or $Y^{a=1}$ is unobserved, that is,

$$P(Y^a = y \mid A = 1) = P(Y^a = y \mid A = 0), \forall a \in \{0, 1\}, \forall y \in \mathcal{Y}.$$

because the treatment assignment is independent of all other factors, including the counterfactual outcomes (Y^a). This conditional independence is called **exchangeability**.

Following Robins, let's reflect on why we analyse data

- "A dataset is a string of numbers...
- ...These data represent empirical measurements...
- ...In an analysis, calculations are performed on these numbers...
- ...Based on the calculations, (causal) inference is drawn...
- ...Since the numerical strings and the computer algorithm applied to them are well-defined mathematical objects, it would be important to provide formal mathematical definitions for **the English sentences expressing the investigator's causal inferences** that agree well with our informal intuitive understanding", Robins (1987)

What we talk about when we talk about a formal framework

Want to be precise about

- Target of inference: What is our **causal estimand**? What possible decisions do we want to compare for what population?
- Assumptions: Under what **assumptions** linking the data to the causal question do our methods give valid conclusions (**identifiability**)? and can we **justify** these assumptions?
- Methods: what makes a method **suitable** to answer a particular causal question?

Terminology

Remember the difference between the following terms:

- **Estimand** (a quantity of interest).
- **Estimator** (an algorithm / function / rule that can be applied to data).
- **Estimate** (an output from applying the estimator to data).

We talk about bias of an estimator with respect to an estimand.
That is, the term *bias* (biased / unbiased) is defined with respect to an estimand.

Terminology

estimand

estimator

estimate

Section 2

Prediction vs. causal inference

Prediction and causal inference are different exercises

- Prediction: Learn about Y after observing $A = a$.
  That is, infer properties of the law P that generated the observations Y.
- Causal inference: Learn about Y after ~~observing~~ fixing $A = a$.
  That is, infer properties of a *counterfactual* law, say, P^a,that would generate data when a is fixed.

Intervening is not the same as conditioning

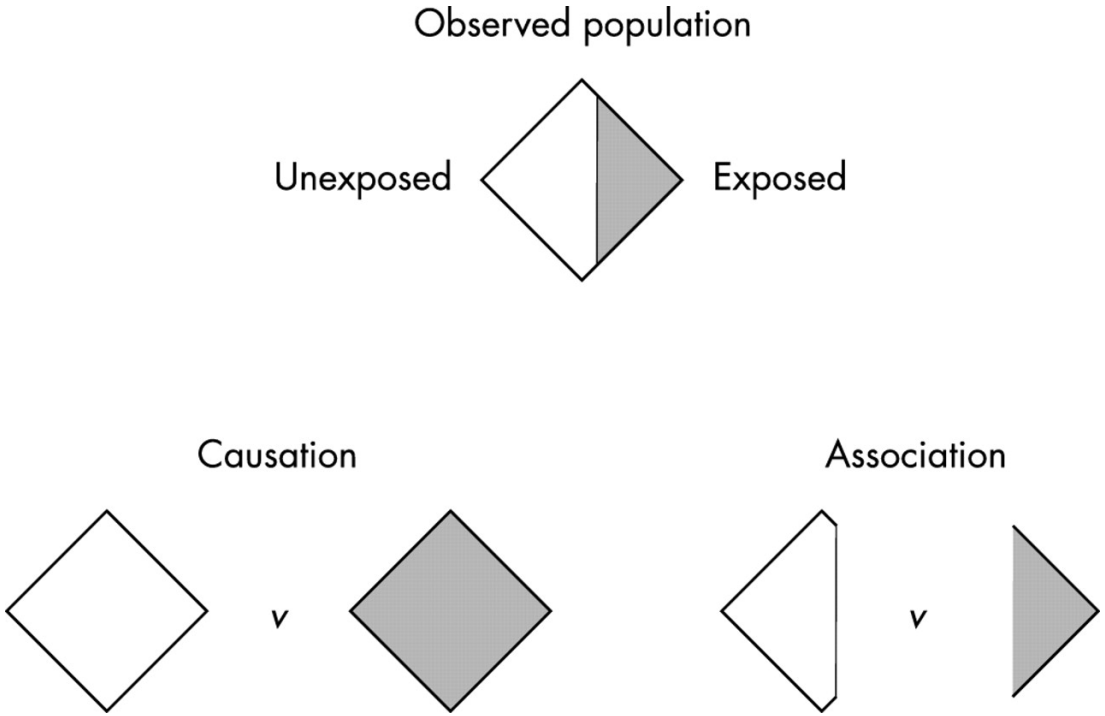


Figure taken from Hernan, 2014, BMJ.

Section 3

Defining a causal effect

What is a causal effect

There is no such thing as 'the' causal effect.
Causal effects are defined as some counterfactuals with the same conditioning set (conditions on the same subset of the population).

We make decisions based on "what if" questions...

- Would starting treatment A prevent a heart attack?
- Is Drug A better than Drug B?
- How would breast cancer rates change if we instituted a policy in which all women took vitamins?
- Would the election campaign increase the number of votes?
- Would university education increase my future earnings?
- What would happen if I went to UNIGE instead of EPFL?

Why so important in (bio)medicine

- We treat patients, we make choices based on counterfactual reasoning...
  What if I accept the surgery? the vaccine? the training program?
- We *do not* have a thorough understanding of the *mechanisms*.
- So we need to the best out of *observations*, combined with the mechanistic assumptions we are willing to make.
- Observational data are increasingly available ...
  Health registries, mobile devices, censors etc etc

What is a causal effect (in a simple setting)

Consider the following observed random variables:

- A binary treatment $A \in \{0, 1\}$.
- An outcome $Y \in \mathcal{Y}$.
- A vector of baseline covariates $L \in \mathcal{L}$.

Define the *counterfactual* or *potential outcome* variables

- $Y^a \in \mathcal{Y}$.
  The outcome variable that would have been observed under the treatment value a (the superscript denotes the counterfactual).
- Often we will specifically instantiate a, i.e. set a to a value:
  $Y^{a=0} \in \mathcal{Y}$.
  The outcome variable that would have been observed under the treatment value $a = 0$.
  $Y^{a=1} \in \mathcal{Y}$.
  The outcome variable that would have been observed under the treatment value $a = 1$.

Individual level causal effect

Definition (Individual level causal effect)

A causal effect for individual (unit) i is $Y_i^{a=0}$ vs $Y_i^{a=1}$.

From now on, I will often omit the i subscript and assume that individuals are iid.

Definition (Causal effect)

A causal effect is a contrast of functionals of counterfactual outcomes under different treatment interventions but in the same individuals.

Remark on counterfactuals

The definition of counterfactuals presupposes:

- $Y^a = Y$ for every unit with $A = a$. In other words, $Y^{a=A} = Y$.
  "Consistency".

  This "consistency" assumption requires that
  - The intervention on A is well-defined.
    No matter how unit i received treatment a, the outcome Y^a is the same.
  - The counterfactual outcome of unit i does not depend on the treatment values of other units j, that is, "no interference".
    Otherwise Y_i^a is not well-defined.[4]

`We will revisit these assumptions.`

[4] This use of consistency is different from the use in estimation.

Remarks on $Y^{a=0}$ and $Y^{a=1}$

The fundamental problem of causal inference:

- Suppose $A = 1$. Then $Y = Y^{a=1}$ is observed, but $Y^{a=0}$ is unobserved...
- Suppose $A = 0$. Then $Y = Y^{a=0}$ is observed, but $Y^{a=1}$ is unobserved...

The consequence is that individual level effect cannot be identified.

Why randomisation is the gold standard

- In a randomised experiment, the treatment is assigned independently of all other factors (e.g. by a coin flip or a random number generator).
- In a randomised experiment one of the counterfactual outcomes $Y^{a=0}$ or $Y^{a=1}$ is unobserved.
- However, randomisation ensures that it is *random* whether $Y^{a=0}$ or $Y^{a=1}$ is unobserved, that is,

$$P(Y^a = y \mid A = 1) = P(Y^a = y \mid A = 0), \forall a \in \{0, 1\}, \forall y \in \mathcal{Y}.$$

because the treatment assignment is independent of all other factors, including the counterfactual outcomes (Y^a). This conditional independence is called **exchangeability**.

Independence notation

Definition (Conditional independence)

$X \perp\!\!\!\perp Y \mid Z \iff F_{X,Y|Z=z}(x,y) = F_{X|Z=z}(x) \cdot F_{Y|Z=z}(y) \; \forall \; x, y, z,$
where $F_{X,Y|Z=z}(x,y) = P(X \leq x, Y \leq y \mid Z = z)$.

We say that X and Y are conditionally independent given Z.
In other words, when $Z = z$ is known, X provides no additional information that allows us to *predict* Y.

Exchangeability (re-visited)

In particular, we can re-write the condition from Slide 41,

$$P(Y^a = y \mid A = 1) = P(Y^a = y \mid A = 0), \forall a \in \{0, 1\}, \forall y \in \mathcal{Y},$$

as

$$Y^a \perp\!\!\!\perp A, \forall a \in \{0, 1\}.$$

Example conditions that ensure identification of causal effects

Suppose that the following 3 conditions hold:

1. $Y^a \perp\!\!\!\perp A, \forall a \in \{0,1\}$ (exchangeability[5]).
2. $P(A = a) > 0 \ \forall a \in \{0,1\}$ (positivity[6]).
3. $Y^a = Y$ for every unit with $A = a$ (consistency[7]).
   that is, $Y = I(A = 0)Y^{a=0} + I(A = 1)Y^{a=1}$.

From (1)-(3), $\mathbb{E}(Y^a) = \mathbb{E}(Y \mid A = a)$.
That is, we have *identified* $\mathbb{E}(Y^a)$ as a functional of observed data.
Assumptions (1)-(3) are external to the data, but – importantly – they hold by design in a perfectly executed experiment.
Just to be clear: The counterfactual independence $Y^a \perp\!\!\!\perp A, \forall a \in \{0,1\}$ does NOT imply the factual independence $Y \perp\!\!\!\perp A$.

[5] Also called ignorability.
[6] Also called overlap. Note that this is a feature of the distribution, not the sample.
[7] Similar to the condition SUTVA: Stable Unit Treatment Value Assumption.

More on consistency

An example of an ill-defined intervention:
Imagine A is a person's body mass index (BMI). Setting the BMI to a counterfactually different level can happen in many different ways - losing weight by running, loss of appetite due to chain smoking, liposuction etc. Depending on what way the intervention is implemented each time, we will have very different health outcomes, i.e., re-running the experiment will give inconsistent results.