What is biostatistics?

STA 102: Introduction to Biostatistics

Yue Jiang

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.
Who are you?
Who are you?
Who are you?
STA 102 is a rigorous introduction to biostatistics that...

- ...provides a tour of basic statistical methods commonly encountered in public health and biomedical research,
- ...emphasizes understanding of the methods, using them to arrive at data-driven decisions, effective communication of such results, and critically assessing existing evidence,
- ...is motivated by timely, relevant examples from current research in biomedicine, epidemiology, and public health policy, and
- ...utilizes modern software such as RStudio and R Markdown documents to reproducibly examine and manipulate data to make sound scientific conclusions.
What IS biostatistics?

A process that converts data into useful information, whereby practitioners

1. form a question of interest,
2. collect and summarize data,
3. and interpret the results
What is biostatistics good for?

The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL ARTICLE

Mortality in Puerto Rico after Hurricane Maria


NEJM (July, 2018)
What is biostatistics good for?

Decrees in global beer supply due to extreme drought and heat

Wei Xie\textsuperscript{1*}, Wei Xiong\textsuperscript{2,3,4}, Jie Pan\textsuperscript{2}, Tariq Ali\textsuperscript{1}, Qi Cui\textsuperscript{5}, Dabo Guan\textsuperscript{6,7*}, Jing Meng\textsuperscript{8}, Nathaniel D. Mueller\textsuperscript{9}, Erda Lin\textsuperscript{2*} and Steven J. Davis\textsuperscript{9,10}

Nature \textit{Plants} (October, 2018)
What is biostatistics good for?

Science Translational Medicine (April, 2019)
What is biostatistics good for?

ORIGINAL ARTICLE

Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists

H. A. Haenssle¹*,¹, C. Fink¹†, R. Schneiderbauer¹, F. Toberer¹, T. Buhl², A. Blum³, A. Kalloo⁴, A. Ben Hadj Hassen⁵, L. Thomas⁶, A. Enk¹ & L. Uhlmann⁷

Annals of Oncology (August, 2018)
What is biostatistics good for?

Ramucirumab plus erlotinib in patients with untreated, EGFR-mutated, advanced non-small-cell lung cancer (RELAY): a randomised, double-blind, placebo-controlled, phase 3 trial

Kazuhiko Nakagawa, Edward B Garon, Takashi Seto, Makoto Nishio, Santiago Ponce Aix, Luis Paz-Ares, Chao-Hua Chiu, Keunchil Park, Silvia Novello, Ernest Nadal, Fumio Imamura, Kiyotaka Yah, Jin-Yuan Shih, Kwok Hung Au, Denis Moro-Sibilot, Sotaro Enatsu, Annamaria Zimmermann, Bente Frimodt-Moller, Carla Visseren-Grul, Martin Reck, for the RELAY Study Investigators

Lancet Oncology (October, 2019)
What is biostatistics good for?

Research

JAMA Internal Medicine | Original Investigation
Comparison of Hospital Mortality and Readmission Rates for Medicare Patients Treated by Male vs Female Physicians

Yusuke Tsugawa, MD, MPH, PhD; Anupam B. Jena, MD, PhD; Jose F. Figueroa, MD, MPH; E. John Orav, PhD; Daniel M. Blumenthal, MD, MBA; Ashish K. Jha, MD, MPH

JAMA Internal Medicine (February, 2017)
Back to the statistical process

1. Forming a question of interest
2. Collecting and summarizing data
3. Interpreting the results
Identifying the population and question of interest

The **population** is the group we’d like to learn something about:

- What is the prevalence of diabetes among **U.S. adults**, and has it changed over time?
- Is there a relationship between tumor type and five-year mortality in **breast cancer patients**?
- Does the average amount of caffeine vary by vendor in **12 oz. cups of coffee at Duke coffee shops**?

If we had data from every unit in the population, we could just calculate what we wanted and be done!
Sampling from the population

Unfortunately, we (usually) have to settle with a sample from the population.

Ideally, the sample is representative, allowing us to use probability and statistical inference to make conclusions that are generalizable to the broader population of interest.
Sampling methods

**Probability sampling** (e.g., simple random sampling, stratified, cluster, or multi-stage sampling)
- All units have a known chance of being selected
- More likely to be generalizable
- Can be more expensive and time-consuming

**Non-probability sampling** (e.g., quota, convenience, or snowball sampling)
- Some units unable to be selected, with no way of knowing size or effect of sampling errors
- Less generalizable to population of interest
- More convenient and less costly
Study design

**Experimental studies (e.g., RCTs)**
- Researchers directly control exposures or treatments
- Ability to make causal statements
- Less real-world applicability and generalizability

**Observational studies (e.g., surveys)**
- Researchers do not assign exposures or treatments
- Real-world setting with lower burden on participants
- Inability to prove causality
What can go wrong?

Selection bias, reporting bias, non-response bias, attrition bias, spin bias, confounding, detection bias, lack of blinding, straight up falsified data (this happens), ...

Catalogue of Bias

...and so much more.
In recent news...

**COVID-19 Antibody Seroprevalence in Santa Clara County, California**

Eran Bendavid, Bianca Mulaney, Neeraj Sood, Soleil Shah, Emilia Ling, Rebecca Bromley-Dulfano, Cara Lai, Zoe Weissberg, Rodrigo Saavedra-Walker, James Tedrow, Dona Tversky, Andrew Bogan, Thomas Kupiec, Daniel Eichner, Ribhav Gupta, John Ioannidis, Jay Bhattacharya

**doi:** https://doi.org/10.1101/2020.04.14.20062463

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.
In recent news...

Peer Review of “COVID-19 Antibody Seroprevalence in Santa Clara County, California”

The high reported positive rate in this serosurvey may be explained by the false positive rate of the test and/or by sample recruitment issues.

What do some other people have to say about this study...?
Reproducibility and replicability

**Reproducibility**: being able to take the original data and code to reproduce all numerical findings

**Replicability**: being able to independently repeat an entire study without use of the original data (generally with the same methods)

Some best practices from the American Statistical Association:

- End-to-end scripting of research
- Use of version control and documentation
- Publication of code along with data
The current replication crisis
The course syllabus is the official document regarding all policies and guidelines and serves as the course syllabus. It is available on the course website here.