Lecture 1

Ciprian M. Crainiceanu

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

September 8, 2020

Table of contents

1. Table of contents
2. Outline
3. Biostatistics
4. Set notations and experiments
5. Probability

Outline

- Discuss biostatistics
- Cover syllabus
  - mathematical prerequisites
  - website
  - quiz and homework schedule
  - test schedule
  - R
- Abstract the idea of an experiment
- Basic set theory for probability
- Probability

Biostatistics defined

Johns Hopkins Department of Biostatistics self study:

Biostatistics is a theory and methodology for the acquisition and use of quantitative evidence in biomedical research. Biostatisticians develop innovative designs and analytic methods targeted at increasing available information, improving the relevance and validity of statistical analyses, making best use of available information, and communicating relevant uncertainties.

Example: cancer screening

The Canadian National Breast Cancer Screening study found no benefit of early tumor detection via digital mammography for women aged 40-49 in a large randomized screening trial, contradicting standard practice of radiology at the time. Gray, in a 2003 Canadian Medical Association commentary, states that **criticisms of the study focus on design, methodology and conclusions.**

Example: Harvard first-borns

Between 75% and 80% of students at Harvard are first-borns. Do first-born children work harder academically, and so end up over represented at top universities? So claims noted philosopher Michael Sandel. But Antony Millner and Raphael Calel find a simple fault in the statistical reasoning and give a more plausible explanation. **Wealthy and well-educated parents tend to have fewer children.**

1. A. Millner, R. Calel. Are first-borns more likely to attend Harvard? Significance, June 2012

2. Sandel, M. (2009) Justice. New York: Farrar, Straus and Giroux.

Example: Oscar winners

Redelmeier and Singh identified all actors and actresses ever nominated for an academy award in a leading or a supporting role up to the time of the study ($n = 762$). Among these there were 235 Oscar winners. For each nominee another cast member of the same sex who was in the same film and was born in the same era was identified ($n = 887$) and these were used as controls. Overall difference in life expectancy was 3.9 years (79.7 vs. 75.8 years; p-value = .003). To avoid the possible selection bias, an analysis using time-dependent covariates (winners counted as controls until they won the Oscar) did not find significant differences. **This is called a selection bias or Immortal bias**. Popes also live longer.

1. D. Redelmeier and S. Singh. Survival in Academy Award-winning actors and actresses. Annals of Internal medicine, 134(10), 955-962, 2001.
2. J. Hanley, M.-P. Sylvestre and E. Huszti. Do Oscar winners live longer than less successful peers? A reanalysis of the evidence. Annals of Internal medicine, 145(5), 361-363, 2006

Example: hormone replacement therapy

A large clinical trial (the Women's Health Initiative) published results in 2002 that contradicted prior evidence on the efficacy of hormone replacement therapy for post menopausal women and suggested a negative impact of HRT for several key health outcomes. **Based on a statistically based protocol, the study was stopped early due to excess number of negative events.**

Example: ECMO

In 1985 a group at a major neonatal intensive care center published the results of a trial comparing a standard treatment and a promising new extracorporeal membrane oxygenation treatment (ECMO) for newborn infants with severe respiratory failure. **Ethical considerations lead to a statistical randomization scheme whereby one infant received the control therapy, thereby opening the study to sample-size based criticisms.**

Summary

- Biostatistics plays a central role in public health and provides a platform for correct design, analysis and interpretation of data
- At the Johns Hopkins BSPH the research philosophy is:
  - A tight coupling of the statistical methods with the ethical and scientific goals
  - Emphasizing scientific interpretation of statistical evidence to impact policy
  - Acknowledging assumptions and evaluating the robustness of conclusions

Experiments

Consider the outcome of an **experiment** such as:

- a collection of measurements from a sampled population
- measurements from a laboratory experiment
- the result of a clinical trial
- the result from a simulated (computer) experiment
- values from hospital records sampled retrospectively
- ...

Notation

- The **sample space**, Ω, is the collection of possible outcomes of an experiment

  Example: die roll $\Omega = \{1, 2, 3, 4, 5, 6\}$
- An **event**, say E, is a subset of Ω

  Example: die roll is even $E = \{2, 4, 6\}$
- An **elementary** or **simple** event is a particular result of an experiment

  Example: die roll is a four, $\omega = 4$
- $\emptyset$ is called the **null event** or the **empty set**

Interpretation of set operations

Normal set operations have particular interpretations in this setting

1. $\omega \in E$ implies that E occurs when ω occurs
2. $\omega \notin E$ implies that E does not occur when ω occurs
3. $E \subset F$ implies that the occurrence of E implies the occurrence of F
4. $E \cap F$ implies the event that both E and F occur
5. $E \cup F$ implies the event that at least one of E or F occur
6. $E \cap F = \emptyset$ means that E and F are **mutually exclusive**, or cannot both occur
7. E^c or $\bar{E}$ is the event that E does not occur

Set theory facts

- DeMorgan's laws

  $(A \cap B)^c = A^c \cup B^c$

  $(A \cup B)^c = A^c \cap B^c$

  Example: If an alligator or a turtle you are not $[(A \cup B)^c]$ then you are not an alligator and you are also not a turtle $(A^c \cap B^c)$

  Example: If your car is not both hybrid and diesel $[(A \cap B)^c]$ then your car is either not hybrid or not diesel $(A^c \cup B^c)$
- $(A^c)^c = A$
- $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

Proving that two sets are equal

- $(A \cup B)^c = A^c \cap B^c$
- $X = Y$ iff $X \subset Y$ and $Y \subset X$

Set theory: an example

PID	BMI	SEX	AGE
1	22	1	45
2	27	0	57
3	31	1	66
4	24	1	49
5	23	0	33
6	18	0	40
7	21	0	65
8	26	1	59
9	34	1	65
10	20	0	42

A=Subjects with BMI$<$26=$\{$1,4,5,6,7,10$\}$

B=Subjects with AGE$>$45=$\{$2,3,4,7,8,9$\}$

$(A \cap B)^c = ?$, $A^c \cup B^c = ?$, $(A \cup B)^c = ?$, $A^c = ?$

Probability: some discussion

- Useful strategy used in much of science:
  For a given experiment
  - attribute all that is known or theorized to a mechanistic model (mathematical function)
  - attribute everything else to randomness, *even if the process under study is known not to be "random" in any sense of the word*
  - use probability to quantify the uncertainty in your conclusions
  - evaluate the sensitivity of your conclusions to the assumptions of your model

Probability: some discussion

- Probability has been found extraordinarily useful, even if true *randomness* is an elusive, undefined, quantity
- *Frequentist* interpretation of probability
  - A probability is the proportion of times an event occurs in an infinite number of identical repetitions of an experiment
- Other definitions of probability exist
- There is not agreement, at all, in how probabilities should be interpreted
- There is (nearly) complete agreement on the mathematical rules probability must follow

Probability: some discussion

- An alternative interpretation of probability is so-called "Bayesian"
- Named after the 18^{th} century Presbyterian Minister / mathematician Thomas Bayes
- A Bayesian interprets probability as a subjective degree of belief
  - For the same event, two separate people could have different probabilities
  - Bayesian interpretations of probabilities avoid some of the philosophical difficulties of frequency interpretations