Mathematical Biostatistics Boot Camp: Lecture 1, Introduction

Brian Caffo

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

June 26, 2012

Table of contents

1. Biostatistics
2. Experiments
3. Set notation
4. Probability

Biostatistics defined

From the Johns Hopkins Department of Biostatistics 2007 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: 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 an excess number of negative events.**

See WHI writing group paper JAMA 2002, Vol 288:321 - 333. for the paper and Steinkellner et al.

Menopause 2012, Vol 19:616 621 for a recent discussion of the long term impacts

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.**

For a review of the statistical discussion and discussion, see Royall Statistical Science 1991, Vol 6, No. 1, 52-88

Summary

- These examples illustrate the central role that biostatistics plays in public health and the importance of performing design, analysis and interpretation of statistical data correctly.
- At the Johns Hopkins Bloomberg School of Public Health, the prevailing philosophy for conducting biostatistics includes:
  - A tight coupling of the statistical methods with the ethical and scientific goals.
  - Emphasizing scientific interpretation of statistical evidence to impact policy.
  - Acknowledging and assumptions and evaluating the robustness of conclusions to them.

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

Probability: some discussion

- Useful strategy used in much of science:
  For a given experiment
  - attribute all that is known or theorized to a systematic model (mathematical function)
  - attribute everything else to randomness, *even if the process under study is known not the 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 long proportion of times an event will occur in repeated identical repetitions of an experiment
- Other definitions of probability exists
- 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 differing probabilities
  - Bayesian interpretations of probabilities avoid some of the philosophical difficulties of frequency interpretations

Probability: a useful exercise

In the first few lectures, we'll largely talk about the mathematics and basic uses of probability. However, this is only a small starting point in the use of probability in data analyses. Keep the following questions in the back of your mind as we cover using probability for data analyses.

- What is being modeled as random?
- Where does this attributed randomness arise from?
- Where did the systematic model components arise from?
- How did observational units come to be in the study and is there importance to the missing data points?
- Do the results generalize beyond the study in question?
- Were important variables unaccounted for in the model?
- How drastically would inferences change depending on the answers to the previous questions?