Introduction to statistical inference

Statistical inference

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Statistical inference defined

Statistical inference is the process of drawing formal conclusions from data.

In our class, we wil define formal statistical inference as settings where one wants to infer facts about a population using noisy statistical data where uncertainty must be accounted for.

Motivating example: who's going to win the election?

In every major election, pollsters would like to know, ahead of the actual election, who's going to win. Here, the target of estimation (the estimand) is clear, the percentage of people in a particular group (city, state, county, country or other electoral grouping) who will vote for each candidate.

We can not poll everyone. Even if we could, some polled may change their vote by the time the election occurs. How do we collect a reasonable subset of data and quantify the uncertainty in the process to produce a good guess at who will win?

Motivating example: is hormone replacement therapy effective?

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.

Here's there's two inferential problems.

1. Is HRT effective?
2. How long should we continue the trial in the presence of contrary evidence?

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 adiscussion of the long term impacts

Motivating 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 and statistical discussion, see Royall Statistical Science 1991, Vol 6, No. 1, 52-88

Summary

- These examples illustrate many of the difficulties of trying to use data to create general conclusions about a population.
- Paramount among our concerns are:
  - Is the sample representative of the population that we'd like to draw inferences about?
  - Are there known and observed, known and unobserved or unknown and unobserved variables that contaminate our conclusions?
  - Is there systematic bias created by missing data or the design or conduct of the study?
  - What randomness exists in the data and how do we use or adjust for it? Here randomness can either be explicit via randomization or random sampling, or implicit as the aggregation of many complex uknown processes.
  - Are we trying to estimate an underlying mechanistic model of phenomena under study?
- Statistical inference requires navigating the set of assumptions and tools and subsequently thinking about how to draw conclusions from data.

Example goals of inference

1. Estimate and quantify the uncertainty of an estimate of a population quantity (the proportion of people who will vote for a candidate).
2. Determine whether a population quantity is a benchmark value ("is the treatment effective?").
3. Infer a mechanistic relationship when quantities are measured with noise ("What is the slope for Hooke's law?")
4. Determine the impact of a policy? ("If we reduce polution levels, will asthma rates decline?")

Example tools of the trade

1. Randomization: concerned with balancing unobserved variables that may confound inferences of interest
2. Random sampling: concerned with obtaining data that is representative of the population of interest
3. Sampling models: concerned with creating a model for the sampling process, the most common is so called "iid".
4. Hypothesis testing: concerned with decision making in the presence of uncertainty
5. Confidence intervals: concerned with quantifying uncertainty in estimation
6. Probability models: a formal connection between the data and a population of interest. Often probability models are assumed or are approximated.
7. Study design: the process of designing an experiment to minimize biases and variability.
8. Nonparametric bootstrapping: the process of using the data to, with minimal probability model assumptions, create inferences.
9. Permutation, randomization and exchangeability testing: the process of using data permutations to perform inferences.

Different thinking about probability leads to different styles of inference

We won't spend too much time talking about this, but there are several different styles of inference. Two broad categories that get discussed a lot are:

1. Frequency probability: is the long run proportion of times an event occurs in independent, identically distributed repetitions.
2. Frequency inference: uses frequency interpretations of probabilities to control error rates. Answers questions like "What should I decide given my data controlling the long run proportion of mistakes I make at a tolerable level."
3. Bayesian probability: is the probability calculus of beliefs, given that beliefs follow certain rules.
4. Bayesian inference: the use of Bayesian probability representation of beliefs to perform inference. Answers questions like "Given my subjective beliefs and the objective information from the data, what should I believe now?"

Data scientists tend to fall within shades of gray of these and various other schools of inference.

In this class

- In this class, we will primarily focus on basic sampling models, basic probability models and frequency style analyses to create standard inferences.
- Being data scientists, we will also consider some inferential strategies that rely heavily on the observed data, such as permutation testing and bootstrapping.
- As probability modeling will be our starting point, we first build up basic probability.

Where to learn more on the topics not covered

1. Explicit use of random sampling in inferences: look in references on "finite population statistics". Used heavily in polling and sample surveys.
2. Explicit use of randomization in inferences: look in references on "causal inference" especially in clinical trials.
3. Bayesian probability and Bayesian statistics: look for basic itroductory books (there are many).
4. Missing data: well covered in biostatistics and econometric references; look for references to "multiple imputation", a popular tool for addressing missing data.
5. Study design: consider looking in the subject matter area that you are interested in; some examples with rich histories in design:
    1. The epidemiological literature is very focused on using study design to investigate public health.
    2. The classical development of study design in agriculture broadly covers design and design principles.
    3. The industrial quality control literature covers design thoroughly.