Lecture 24: AB Testing 2 and Wrap-up

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner

ANNOUNCEMENTS

- Homework 8:
  - Completely on ED. Will be posted tonight. Partners allowed.
  - It will be about half as long as a typical HW.
- Project:
  - Website (45 points) and notebook (30 points) due on Wed, 12/11
  - Individual peer evaluations (5 points) due on Thurs, 12/12
  - Details: https://github.com/Harvard-IACS/2019-CS109A/blob/master/content/projects/ProjectGuidelines.pdf

Outline

- AB Testing: a Brief Review
- Adaptive Experimental Design
- Course Wrap-up

AB Testing: a Brief Review

Assessing Causal Effects

Most data are collected **observationally**, without intervention into what measurements the predictors take on.

It is difficult to assess **causality** in an observational study and may even be impossible. You never know if all **confounders** are accounted and controlled for properly.

An experiment (called **AB test** in the world of Data Science) can be conducted to determine causal relationships between a **treatment** and a **response**, but they come with their own drawbacks (artificial, expensive, etc.).

AB Testing and Experimental Design

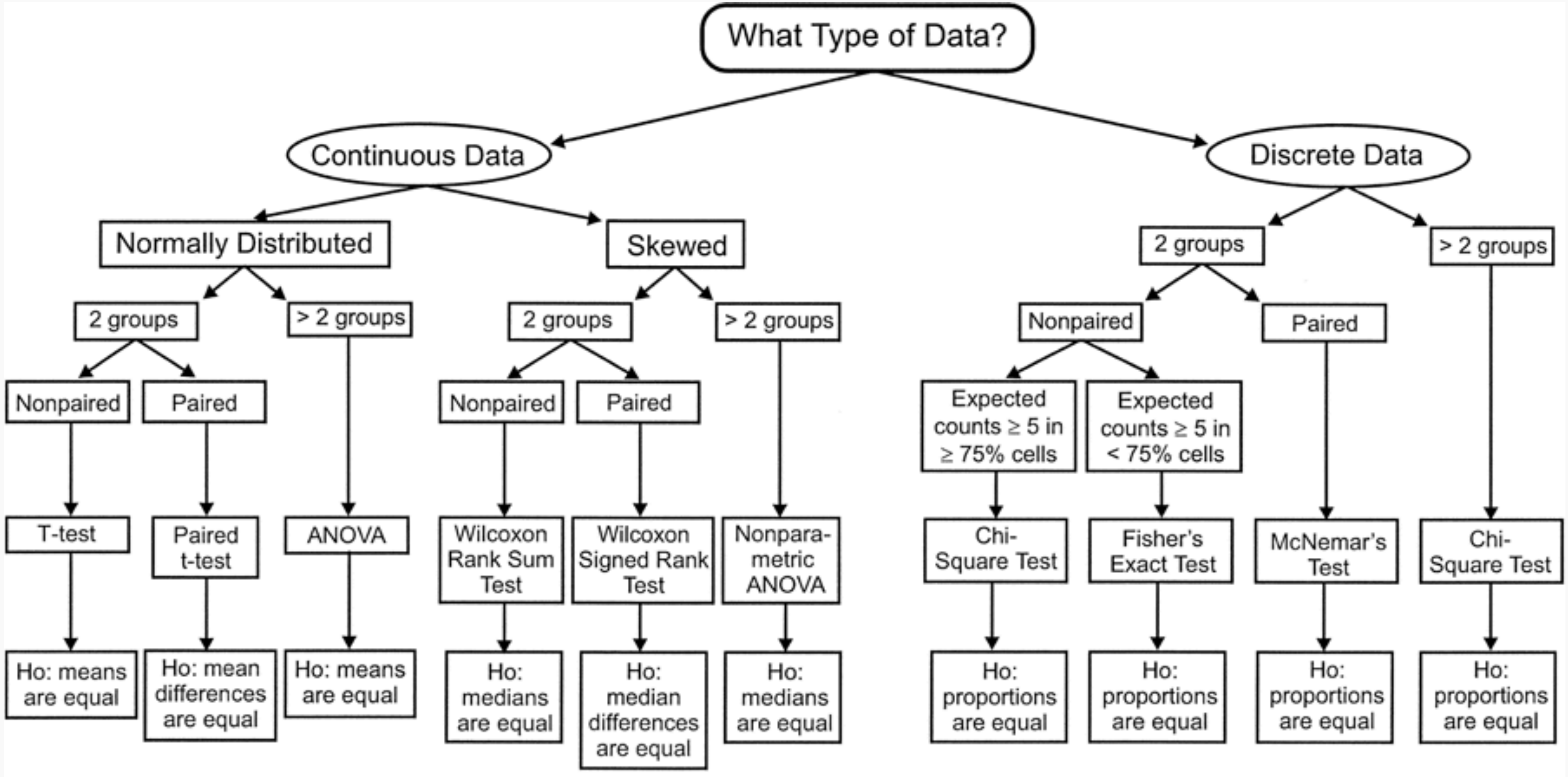
Many flavors of AB tests. 3 key characteristics:

1. Comparison/**control group**
2. **Random assignment** of treatment to subjects
3. **Repetition** (to *ensure* balance).

Completely Randomized Design (CRD) is like pulling names out of a hat. **Stratified Randomized Design** performs a CRD within **strata**.

The **multivariate experimental design** generalizes this approach. If there are two treatment types (font color, and website layout), then both treatments' effects can (and should) be tested simultaneously.

Analyzing the results: should be easy



Source: Waning B, Montagne M: *Pharmacoepidemiology: Principles and Practice*: http://www.accesspharmacy.com

Adaptive Experimental Design

Beyond CRD designs

The approaches we have seen to experiments all rely on the completely randomized design (CRD) approach. There are many extensions to the CRD approach depending on the setting. For example:

- If there are more than two types of treatments (for example: (i) font type and (ii) old vs. new layout), then a *factorial* approach can be used to test both types of treatments at the same time.
- If the treatment effect is expected to be different across different subgroups (for example possibly different for men vs. women), then a stratified/cluster randomized design should be used.

Beyond CRD designs (cont.)

These different experimental designs will need to have adjusted analysis approaches to analyze them appropriately.

Examples:

1. **factorial design:** a multi-way ANOVA when the response variable is quantitative.
2. **stratified analysis**: the Mantel-Haenszel test for cluster randomized design with a categorical response variable.

Beyond CRD designs (cont.)

But all of these procedures rely on the fact that there is a fixed sample size for the experiment. This has a glaring limitation: you have to wait to analyze until *n* is recruited/reached.

If you peak at the results before *n* is reached, then this is a form of *multiple comparisons* and thus overall Type I error rate is inflated.

Bandit Designs

A **sequential** or **adaptive** procedure can be used if you would like to intermittently check the results as subjects are recruited (or want to look at the results after each and every new subject is enrolled).

One example of a sequential test/procedure is a **bandit-armed** design. In this design, after a burn-in period based on a CRD, then the treatment that is performing better is chosen more often to be administered to the subjects.

Bandit Design Example

For example, in the **play the winner** approach for a binary outcome, if treatment *A* is successful for a subject, then you continue to administer this treatment to the next subject until it fails, and then you skip to treatment *B*, and vice versa.

The advantage to this approach is that if one treatment is truly better, then the number of subjects exposed to the worse treatment is lessened.

What is a major disadvantage?

Bayesian Bandit Designs

Our friend Bayes’ theorem comes into play again if we would like to have a bandit design for a quantitative outcome.

The randomization to treatment for each subject is based on a biased coin, where the probability of being assigned to treatment *A* is based on the poster probability that treatment *A* is a better treatment.

Bayesian Bandit Designs (cont.)

This probability can be calculated based on the Bayes theorem as follows:

$$P\left(\mu_{Y|trt_A} > \mu_{Y|trt_B} \middle| Data\right)$$
$$\propto P\left(Data \middle| \mu_{Y|trt_A} > \mu_{Y|trt_B}\right) P(\mu_{Y|trt_A} > \mu_{Y|trt_B})$$

where $P(\mu_{Y|trt_A} > \mu_{Y|trt_B})$ is the prior belief (can be set to 0.5). It is a little more complicated than that.

This can easily extend to more than just 2 treatment groups.

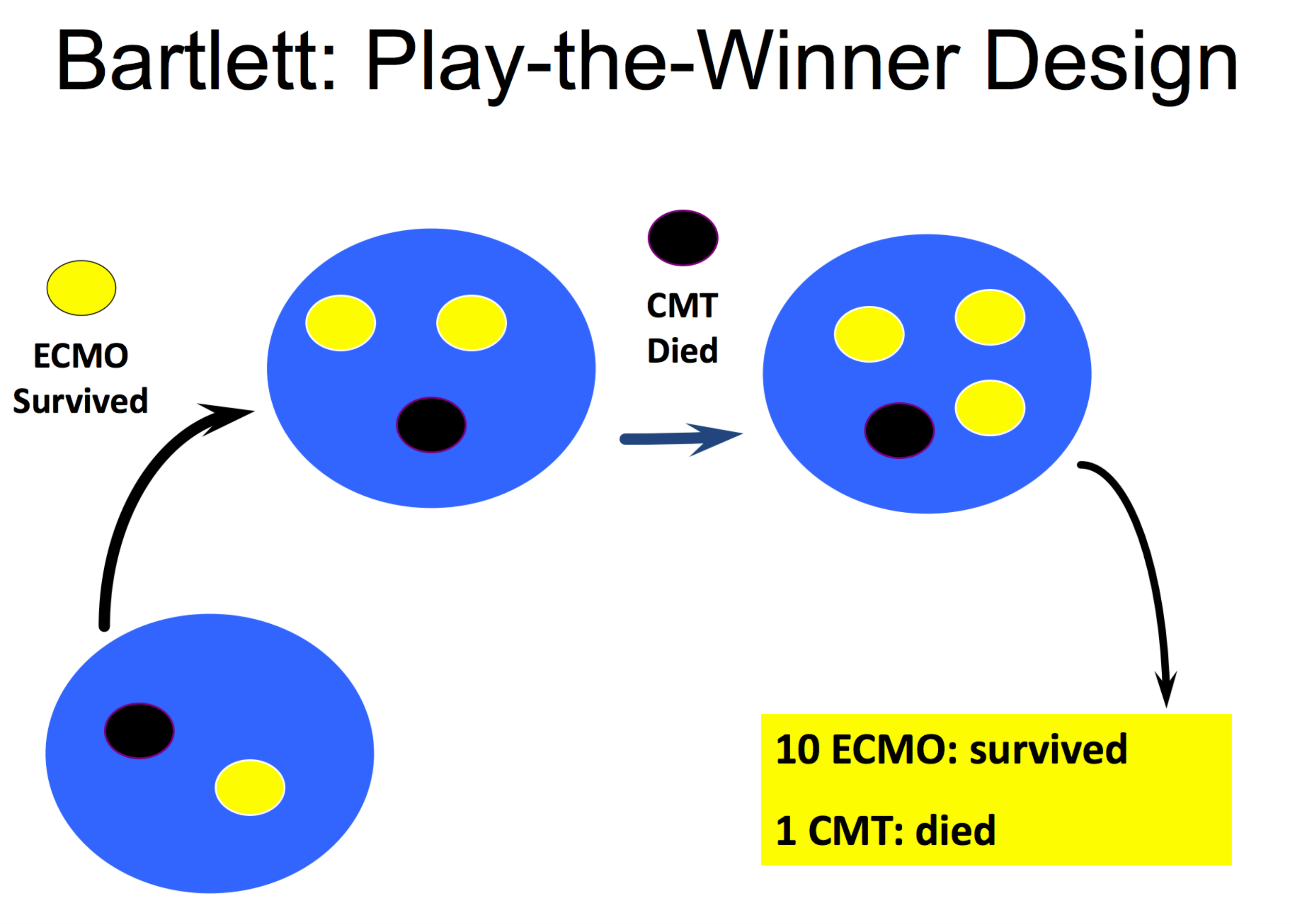
ECMO Trial: Bayesian Bandit Trial Example

In the 80’s a bandit-armed design (Bartlett, et al.) was used to determine whether or not Extracorporeal Membrane Oxygenation (ECMO) would improve survival (compared to ‘standard of care’) of neonatal patients (premature babies) experiencing respiratory failure.

In the end, only 11 patients were enrolled before “statistical significance” was achieved.

What is an issue with these results?

ECMO Trial



Bartlett: Play-the-Winner Design

10 ECMO: survived

1 CMT: died

Analysis of Bayesian Approaches

So when should you stop an adaptively designed trial?

You could continue the trial until a *p*-value of less than 0.05 is achieved (or until a large sample size is taken without coming to a statistically significant result)?

What is an issue with this “stopping criterion”?

If our *p*-value is determined from a classical method, then this is an example of **multiple comparisons**: you have looked at the data at many points along the timeline, so a significant result is more likely to occur than 0.05 if there is not a *true* difference in the treatments.

We need to adjust how the ‘statistical significance’ is determined!

Course Wrap-up

Things we haven’t discussed

There are lots of topics we have not covered in one semester…some are covered in 109B in the Spring:

- Unsupervised Classification/Clustering
- Smoothers
- Bayesian Data Analysis
- Reinforcement Learning
- Other versions of Neural Networks (and ‘Deep Learning’)
- Interactive Visualizations
- Database Management (SQL, etc.)
- Cloud Computing and Scaling (AWS)
- And much, much more…

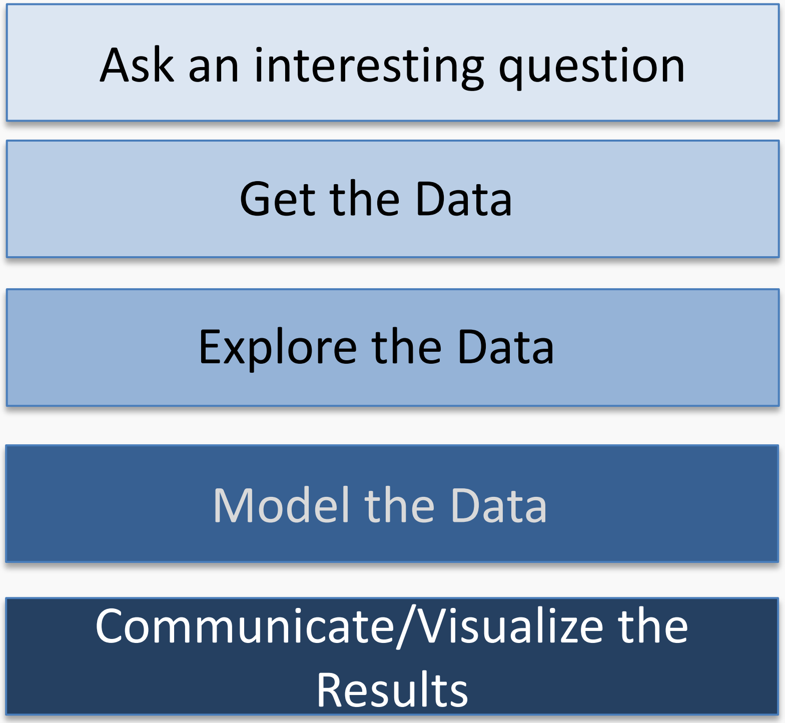
Courses Related to Data Science

- CS 109B: Advanced Topics in Data Science
- CS 109C: Very Advanced Topic in Data Science
- CS 171: Visualizations
- CS 181/281: Machine Learning
- CS 182: Artificial Intelligence (AI)
- CS 205: Distributive Computing
- Stat 110/210: Probability Theory
- Stat 111/211: Statistical Inference
- Stat 139: Linear Models
- Stat 149: Generalized Linear Models
- Stat 195: Intro to Statistical Machine Learning

This list is not exhaustive!

The Data Science Process

Don’t forget what everything is all about:



Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Thanks for all your hard work!

It’s been a long semester for everyone involved. Thank you for your patience, your hard work, and your commitment to data science!

It’s sad to see you go...