Probability in Data Science

A Case Study

C. Durso

Polio Trial Example

- Poliomyelitis: serious illness, comparatively rare
- Salk vaccine trials 1954 (Sabin vaccine later)
- Contributions to complexity of trial
  - Rare disease
  - Annual rates vary

Two Study Designs

1. Observed control study
   - Inoculate second graders at participating schools
   - Observe first, second, and third graders for cases of polio
2. Placebo control study
   - Combine participating first, second, and third graders
   - Assign half of participants to treatment, half to control

Major Effort

Study	Group	Count
Observed control	Vaccinated	221,998
Observed control	Unvaccinated, other grades	725,173
Observed control	Unvaccinated, second grade	123,605
Placebo control	Vaccinated	200,745
Placebo control	Placebo	201,229

Placebo Control Study

- Randomized
- Double blind

Results

Study	Group	Count	Paralytic	Nonparalytic
Observed control	Vaccinated	221,998	38	18
Observed control	Unvaccinated, other grades	725,173	330	61
Observed control	Unvaccinated, second grade	123,605	43	11
Placebo control	Vaccinated	200,745	33	24
Placebo control	Placebo	201,229	115	27

Significant Reduction

What does this mean?

- Could mean reduction unlikely to be due to chance (statistically significant)
- Could mean reduction important in human terms, relative to resources required and to risk

Randomized Control

C. Durso

Control: Compared to What?

- Make a claim about a phenomenon
  - Effective
  - Profitable
  - Safe
- May need or want comparison group or groups

Treatment/Control

- Treatment/no treatment
- New treatment/best existing treatment
- Dose levels

Goal: Treatment Effect

- Compare outcomes
  - Statistically significant difference
  - Clinically significant difference
- Causal inference: treatment caused difference

Risk: Systematic Difference

- Treatment and control group should differ systematically only in treatment effect.
- Problems:
  - Observational studies
  - Opt in to treatment (observed control)
  - Effect of treatment experience vs. treatment
  - Non-response, missing data

Randomized Control

- Randomize assignment to treatment(s) and control
- No systematic difference (check randomization)
- “Gold standard” for causal inference

Double Blind

- Placebo and treatment
  - Treatment experience doesn’t differ
- Assessment
  - Interpretation of outcome doesn’t differ

Probability Model

Harness probability theory

- Systematic difference confined to the treatment
- Chance model used to assess statistical significance
- Chance model used to estimate effect size

Sketch of Statistical Analyses

C. Durso

Statistics: The Numbers, Not the Subject

Definition

Statistic: A number calculated from data without use of unknown parameters.

Examples of Statistics

- Sum
- Mean
- Median
- Maximum
- Minimum
- Interquartile range
- And many more

What Do We Know?

- A statistic summarizes aspects of the data.
- Example:
  - Proportion of subjects contracting disease
    - Treatment group
    - Control group
  - Mean amount of medication in sample of manufactured doses (capsules)
  - Slope of “best” line relating nitrogen level of fertilizer to crop yield
  - Average date of full flower for orchard trees

How Well Do We Know It?

- Reproducibility of results
- In another sample: predictive power
  - Another treatment and control group
  - Another sample of capsules
  - Another collection of fields
- Precision of results, the "±"
- Use probability theory

Applications

- Analysis of experiments
  - Relation of outcome to manipulation
  - Natural science
  - Social science
- Process control
  - Are results consistent with manufacturing parameters
- Data exploration
  - Summary descriptions
  - Relations among measurements

Intuitive Recognition

- Outcome of the same process may vary.
- Who is the better at thumb wrestling?
  - One bout?
  - Best two out of three?

Confidence Intervals

- Example: The poll's 95% confidence interval for proportion of the population supporting the referendum is (0.48,0.52).
- (Convoluted) interpretation: The interval was computed from the polling data using a particular process. If the poll and the process were repeated many times, 95% of intervals would include the true population proportion of support for the referendum.

Hypothesis Test: Polio Trial Example

Does the treatment group disease outcome differ from the control group disease outcome?

Null Hypothesis

- **Often:** no change, no relation, no difference (treatment has no effect)
- **Sometimes:** hypothesis against which you want evidence of known quality (example: election results hacked)

A Process for Testing Null Hypothesis

- Define question via null hypothesis and alternative hypothesis
- Collect data (how much?)
- Define and calculate statistic
- Estimate probability of a statistic as unlikely or more unlikely under null hypothesis

Vaccine Has No Effect

- Safety of vaccine is thought to be established before large clinical trial.
  - Alternative hypothesis is that vaccine reduces incidence of polio.
- Are case counts consistent with null?
- Think of all cases as predetermined, then randomly assigned to (irrelevant) treatment category. Are data consistent with this?

Group	Total	Paralytic polio count
Vaccinated	200,754	33
Placebo	201,229	115

Shuffle Model

Simulate outcome unrelated to treatment
by permuting treatment and control labels.
What is the probability of a difference in control
and treatment counts this large in shuffled group
assignments of all participants?

Sample Proportion Model

Assume equal probability of cases in treatment and control. Under null, probability that observed number would be assigned to treatment given proportion of treatment to control population.

Challenges

- Model assumes no relation among subjects. Geographic clusters, family clusters present.
- Model assumes no missing data. Possible differential non-response.
- Other issues?

Other Ideas?

Introduction to R and RStudio

C. Durso

Introduction to R and RStudio

- Data analysis software
- Statistical programming environment
- Free
- Open source, large contributor base
- Hub for multiple tools

Strengths of R

- Fully integrated statistical software and programming
- Reproducibility
- Implementation of cutting edge methods
- Active online community

R or X: Challenges of R

- No menu-driven analyses
- Few default diagnostics (opt-in instead)
- Many ways to do a task
- Lack of standardization
- Some weaker libraries
- Lack of speed

Back to Strengths

Fully integrated statistical software and programming

- Standard analysis tools built in
- Libraries for extended functionality
  - No decision to buy required
- Task views for more extended functionality
- Full service programming language
- Model fit output accessible to programmer
  - Collect parameters from multiple models
  - Use parameters as data

Reproducibility

- In house
  - Easily documented
- By others
  - Easily documented
  - Platform free, simple, and readily available
  - Work done on single platform
  - Backward compatible versions

Cutting Edge Methods

- First publication opportunity for new methods
- Quick publication (minimal vetting)
- Opportunity to test tools before buying for other platforms

RStudio

- Integrated Development Environment for R
- Convenient interaction for rapid development
- Free single-user version
- Active development

Quick Startup

- Download R
- LaTeX optional
  - MacTeX
  - MiKTeX
  - install.packages("tinytex")
- Download Rstudio: IDE for R