Interplay Between Science and Statistics

UC Irvine - ISI BUDS 2023

Presented July 24, 2023

Daniel L. Gillen
Chancellor's Professor and Chair
Department of Statistics
University of California, Irvine

Scientific Investigation

First Stage of Scientific Investigation

- Hypothesis generation
  - Observation
    - Measurement of existing populations or systems
  - Disadvantages:
    - Confounding
    - Limited ability to establish cause and effect

Further Stages of Scientific Investigation

- Refinement and confirmation of hypotheses
  - Experiment
  - Intervention

- Elements of experiment
  - Overall goal and specific aims (hypotheses)
  - Materials and methods
  - Collection of data
  - Analysis
  - Interpretation; Refinement of hypotheses

Role of Statistics

- Answering scientific questions in presence of variable response
- Scientific questions often reduce to comparing the magnitude of some measurement across groups
- Outcome measures are rarely constant
  - Inherent randomness
  - Hidden (unmeasured) variables
- Use of probability models for describing variability in the real world
  - Distribution of measurements
  - Summary measure (functional) for scientific tendency
  - Quantification of uncertainty in (contrast of) functional(s) (Signal and noise)

"Statistics means never having to say you're certain."

UCIrvine
University of California, Irvine

Role of Statistics

UCIrvine
University of California, Irvine

Major goals of statistics typically include

1. Quantifying distributions (or functionals of distributions)
2. Comparison of distributions between groups
3. Identification of clusters
4. Factor analysis
5. Prediction

1. Quantifying distributions

- Scientific questions about tendencies for specific measurements within a population
  - Point estimates of summary measures
  - Interval estimates of summary measures
  - Quantifying uncertainty
  - Decisions about hypothesized values

Example 1: Median life expectancy in stage II breast cancer

- General goal: Want to know prognosis
- Follow a cohort of newly diagnosed patients and measure survival time (may be censored)
  - What defines the cohort?
    - All patients from a particular healthcare plan?
    - All patients diagnosed at a particular hospital?
    - All patients in a certain location?
- Best estimate of the median survival (??)
- Quantify uncertainty in that estimate
- Compare to some clinically important time range (e.g., 10 years)

1. Quantifying distributions

UCIrvine
University of California, Irvine

Example 2: Women suffering from primary biliary cirrhosis

- Primary biliary cirrhosis: Serious liver disease often leading to liver failure
- Measure proportion of PBC patients that are women
- Best estimate of the proportion (What does this mean?)
- Quantify uncertainty in that estimate
- Compare to the known proportion of women in the general population (approximately 50%)
- FYI: About 90% of patients with PBC are women

2. Comparing distributions across populations

UCIrvine
University of California, Irvine

Can be classified into two main subgroups

2a. *Quantifying differences in the distribution of some measurement across predefined groups (effects or associations)*

2b. *Quantifying differences in effects across subgroups (interactions or effect modification)*

2a. Quantifying differences in the distribution of some measurement across predefined groups (effects or associations)

- Typically wish to quantify
  - Existence of differences
  - Direction of tendency of effect
  - First, second order relationships in a summary measure
  - Characterization of dose-response in a summary measure

Example: Effect of serum albumin levels on risk of mortality in end-stage renal disease patients

- Possible approaches to the analysis:
  1. Compare incidence of mortality across groups of subjects defined by serum albumin
  2. Compare serum albumin level across groups of subjects defined by mortality
- In either case, comparison can be at many levels of detail regarding nature of differences

2b. Quantifying differences in effects across subgroups (interactions or effect modification)

- Typically wish to determine
  - The existence of interaction
  - The direction of interaction (synergy, antagonism)
  - Quantification of exact relationship of interaction

Example: Differential effect of cholesterol on heart attacks by sex?

- Wish to compare the association between serum albumin level and incidence of mortality between racio-ethnic groups
- Statistical analysis
  - Quantify association in Hispanic White
  - Quantify association in Black
  - Quantify association in Asian
  - Quantify association in Non-Hispanic White
  - Compare measures of association

3. Cluster analysis

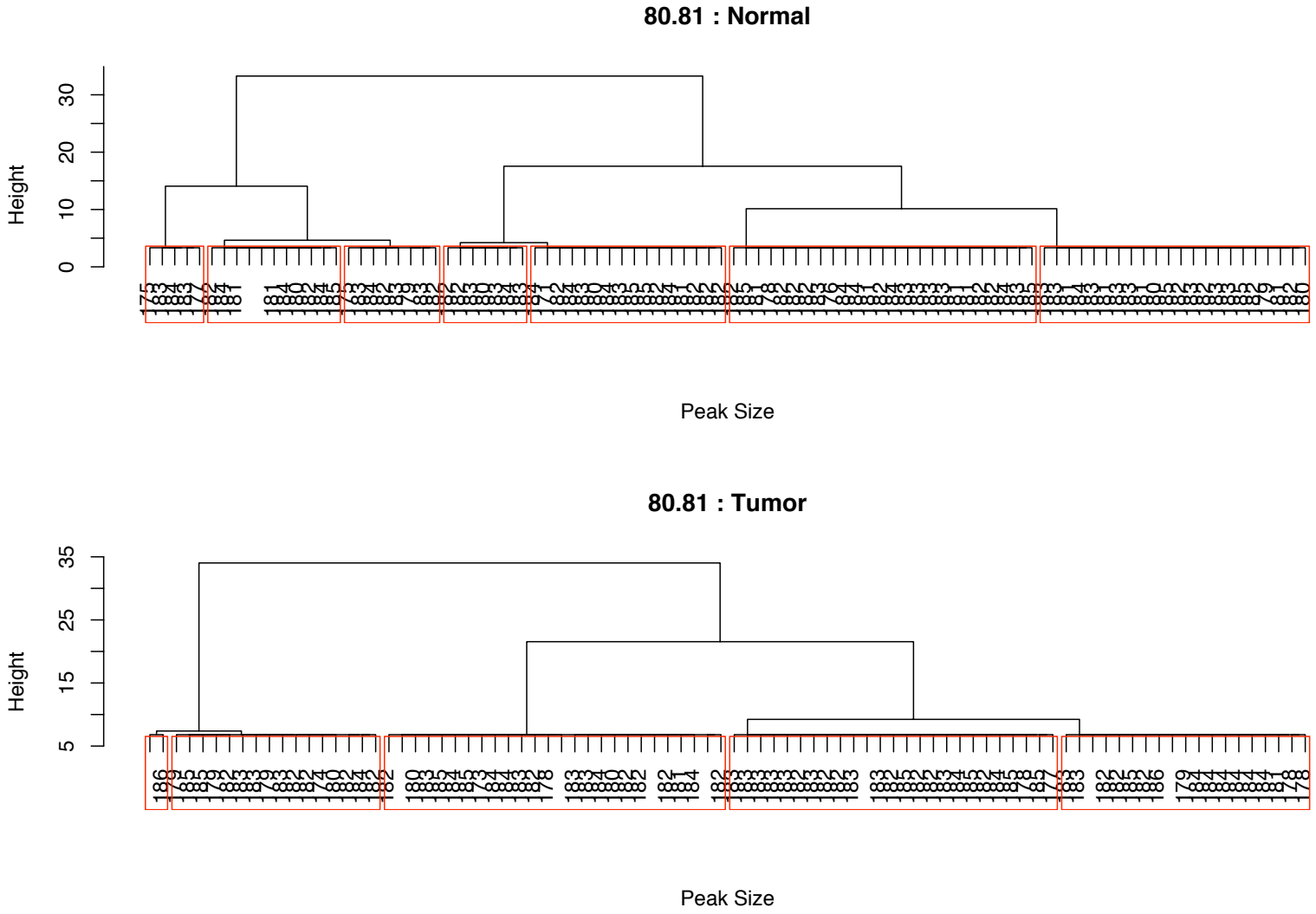
- Focus is on identifying similar groups of observations
- Divide a population into subgroups based on patterns of similar measurements
  - Univariate or multivariate
  - Known or unknown number of clusters
- All variables treated symmetrically with no delineation between outcomes and groups

Example: Gene expression of cancerous tumors

- Dependent upon genetic association, gene expression will be differential between various cancer types
- Goal: Identify gene expression patterns that separate subpopulations of patients (hopefully by cancer type)
  - Array of genes (thousands) with corresponding expression levels
  - Cluster observations into one of K groups to minimize total variability (K-means clustering)
  - Tabulate cancer type by cluster assignment

3. Cluster analysis

Example: Gene expression of cancerous tumors



4. Factor analysis

- Identification of hidden variables indicating groups that tend to have similar measurements of some outcome
- Interest in some particular outcome measurement
- Predictors that imprecisely measure some abstract quality
- Desire to find patterns in predictors that more precisely reflect the abstract quality

Example: Barriers to patient compliance in clinical trials

- In the Health Behavior Questionnaire, multiple variables might be used to measure
  - Self-perceived health
  - Social support
  - Depression
- Goals:
  1. Find subset of questions that would suffice
  2. Identify hidden variables that affect compliance

5. Prediction

- Focus is on future measurements

Point prediction

- Best single estimate for the measurement that would be obtained on a future observation
  - Continuous measurements
  - Binary measurements (discrimination)

Interval prediction

- Range of measurements that might reasonably be observed for a future observation

5. Prediction

UCIrvine
University of California, Irvine

Example 1 (continuous): Measure of renal function

- Creatinine is produced by breakdown of creatine, which is used by muscles for energy transfer
- Removed by kidneys via filtration with little secretion/reabsorption
- Amount of creatinine cleared by the kidneys in 24 hours used as a measure of blood processed by the kidneys
- Problem: Need to collect urine (and blood creatinine) for 24 hours
- Goal: Find blood/urine measures that can be obtained instantly, yet still provide an accurate estimate of a patient's creatinine clearance

5. Prediction

UCIrvine
University of California, Irvine

Example 2 (categorical): Diagnosis of prostate cancer

- Goal: Use other measurements to predict whether a particular patient might have prostate cancer
  - Demographic: Age, race, (sex)
  - Clinical: Symptoms
  - Biological: Prostate specific antigen (PSA)

Example 3 (interval): Determining normal range for PSA

- Goal: Identify the range of PSA values that would be expected in 95% of males in the “healthy” population
  - Typically consider age and race specific values

Distinctions Without a Difference? No!

- The largest distinctions in these five goals arise between (1-2) and (3-5)

(1) and (2) are often the focus of hypothesis driven science

- Testing a well-defined scientific hypothesis
- Avoidance of data-driven models to guard against inflation of the false positive rate

(3), (4) and (5) necessitate model building and data driven results

- Allows for the use of much more flexible models
- Requires the need for stronger assumptions to make probability statements and/or
- Requires the need for additional *independent* data samples to validate models (this is also the case in hypothesis driven analyses as well)

Study Design

UCIrvine
University of California, Irvine

Methods

- Observational study
  - Data collecting from existing state without intentional interference
- Interventional study (experiment)
  - Covariate levels dictated by study researcher

Statistical Tasks

UCIrvine
University of California, Irvine

Time frame

- Cross-sectional sampling
  - Real (calendar) time
  - Event time (e.g. at diagnosis or birth)
- Longitudinal sampling
  - Prospective
    - Follow subjects for occurrence of a specified event
  - Retrospective
    - Event has already occurred but exposure measured earlier in time

Statistical Tasks

UCIrvine
University of California, Irvine

Subjects

- Independent cohorts
  - Randomly sample groups to be compared
- Matched groups
  - Match on potentially important factors that you are not interested in comparing across groups

UCIrvine
University of California, Irvine

Aside: Why match?

- Usually to control for potential *confounding* factors
  - Suppose we are interested in the relationship between X (a predictor of interest) and Y (the outcome).
  - A *confounder* is a third variable Z that is causally associated with X and Y.
- Lack of adjustment for confounding can lead to a different estimated relationship (unintended) between X and Y.
- Classic example of confounding
  - Consider the relationship between alcohol use and the incidence of lung cancer.
  - Many studies have shown a positive relationship between the two, but why?

Statistical Tasks

Strength of statistical evidence

- Sample size determination
- Statistical power is defined as

$$\Pr[\text{Reject } H_0 | H_0 \text{ false}]$$

  where H_0 is some particular hypothesis (eg. “no treatment effect")
- Because of variability, power is a function of sample size
- It is possible for a study to be either *over-* or *under-powered*
  - A *over-powered* study is designed to detect differences in outcome that are too small to be clinically relevant
    - “Our sample size ensures that we will have 99% power to detect a difference of 1 day in median survival among newly diagnosed pediatric acute lymphoma cases."

Statistical Tasks

UCIrvine
University of California, Irvine

Strength of statistical evidence

- A *under-powered* study has limited ability to reject the null hypothesis for differences in outcome that are clinically relevant
  - “Our sample size ensures that we will have 50% power to detect a difference of 10-years in median survival among newly diagnosed pediatric acute lymphoma cases."
- At best, an over-powered or under-powered study is a waste of resources (money, time, etc.). At worst it can be unethical (consider quality of life, delay in progress of treatment, etc.)

Common Study Designs

UCIrvine
University of California, Irvine

1. Survey studies

- Cross-sectional
- Real (calendar) or event time
- Efficiency for examining:
  1. Common diseases and risk factors
  2. Associations (not cause and effect)
- Often limited ability to control for confounding

Common Study Designs

UCIrvine
University of California, Irvine

1. Survey studies

- Example Is there an association between ethnicity and health care?
  - Exposure (ethnicity) and outcome (health care) are easy to measure
  - How does one choose survey participants?
  - What about those that refuse to participate? Different from others?
  - If association is found, do we attempt to explain it?

Common Study Designs

UCIrvine
University of California, Irvine

2. Cohort studies

- Groups defined by risk factor
- Identified prospectively or retrospectively
- Followed for some outcome event(s)
- Efficiency for examining
  1. Common diseases
  2. Many different outcomes for same exposure
  3. Associations (not cause and effect)

Common Study Designs

UCIrvine
University of California, Irvine

2. Cohort studies

- Example: Is renal impairment among middle aged individuals associated with an increased risk of cardiovascular disease (CVD)?
  - Measure renal function on subjects between age 35 and 45 with no history of (CVD) and follow until first (CVD) event
  - Could also consider the relationship between renal function and death
  - If an association is found, how could it not be causal?

Common Study Designs

UCIrvine
University of California, Irvine

3. Case-control studies

- Groups defined by some outcome event
- Characterize prior exposures
- Efficiency for examining
  1. Rare diseases
  2. Many different risk factors for same disease
  3. Associations (not cause and effect)

3. Case-control studies

- Example: Risk factors for childhood leukemia
  - Randomly sample children diagnosed with leukemia at time of diagnosis (cases)
  - Randomly sample children without leukemia (controls)
  - Compare multitude of risk factors from environment to genetic profile
  - Careful though must go into the choice of controls in order to make a fair comparison
    - Possibly match to reduce confounding

Common Study Designs

UCIrvine
University of California, Irvine

4. Interventional (clinical trials)

- Ideally controlled and randomized
- Efficiency for examining
  - Common outcomes
  - Cause and effect

4. Interventional (clinical trials)

- Example Does recombinant human thrombin (rhThrombin) reduce incidental bleeding during spinal surgery?
  - Randomize patients to receive rhThrombin (applied via sponge) or a *placebo* consisting of a saline solution
  - *Double-blind* study participants so that the patient nor the surgeon knows what is being applied
  - Measure time from application of sponge until bleeding stops
    - Note: A *single-blind* study would only blind the patient and not the surgeon
    - How could this lead to a *placebo effect*?

Common Study Designs

UCIrvine
University of California, Irvine

What are we able to estimate under each of these study designs?

1. Cross-sectional design
   - Sample from population
   - Measure exposure (E), disease (D)
   - Can estimate:
     1. $\Pr[E, D]$: Joint distn of Exposure, Disease
     2. $\Pr[D|E]$: Conditional distn of D within levels of E
     3. $\Pr[E|D]$: Conditional distn of E within levels of D
     4. $\Pr[D]$: Marginal distn (*prevalence*) of Disease
     5. $\Pr[E]$: Marginal distn of exposure

Common Study Designs

UCIrvine
University of California, Irvine

What are we able to estimate under each of these study designs?

2. Cohort design
   - Fix sample sizes for each level of exposure (E)
   - Measure disease (D)
   - Can estimate:
     1. $\Pr[D|E]$: Conditional distn of D within levels of E

Common Study Designs

UCIrvine
University of California, Irvine

What are we able to estimate under each of these study designs?

3. Case-control design
   - Fix sample sizes for each level of disease (D)
   - Measure exposure (E)
   - Can estimate:
     1. $\Pr[E|D]$: Conditional distn of E within levels of D

Common Study Designs

UCIrvine
University of California, Irvine

What are we able to estimate under each of these study designs?

4. Interventional design (clinical trial)
   - Fix sample sizes for each level of disease (E)
   - Measure disease (D)
   - (so a cohort design)
   - Can estimate:
     1. $\Pr[D|E]$: Conditional distn of D within levels of E

Common Study Designs

UCIrvine
University of California, Irvine

Key point

- Each can be used for testing an association between E and D:
  - H_0 : $\Pr[D, E] = \Pr[D) \times \Pr[E]$ (cross sec)
  - H_0 : $\Pr[D|E] = \Pr[D|\bar{E}]$ (cohort)
  - H_0 : $\Pr[E|D] = \Pr[E|\bar{D}]$ (case-control)
- What about parameter interpretation?

Statistical Tasks

UCIrvine
University of California, Irvine

Description of a sample

- Identification of measurement or data entry errors
- Characterization of materials and methods
- Validity of analysis methods
- Hypothesis generation (for inference and estimation)

Statistical Tasks

UCIrvine
University of California, Irvine

Inference

- Goal: Generalize from sample to population
- Two main components:
  1. Estimation
     - 1a. Point estimation
     - 1b. Interval estimation
  2. Testing

Statistical Tasks

UCIrvine
University of California, Irvine

Inference

1a. Point estimation

- Identification of clusters
- Point estimates
- Individual observations (predictions)
  - Continuous measurements
  - Categorical measurements (discrimination, classification)
- Summary measures of distributions
  - Within a population
  - Across populations

Statistical Tasks

UCIrvine
University of California, Irvine

Inference

1b. Interval estimates (quantifying uncertainty)

- Individual observations
  - Prediction intervals (continuous measurements)
  - Accuracy (discrimination, classification)
- Summary measures
  - Confidence or credible intervals

2. Decisions (hypothesis testing)