Statistical Methods for Plant Biology

PBIO 3150/5150

Anirudh V. S. Ruhil

January 14, 2016

The Voinovich School of Leadership and Public Affairs

Table of Contents

Overview of PBIO 3150/5150

PBIO 3150/5150

- What are we going to do this semester?
  1. Course Map: Basic to intermediate statistics
  2. Distribution of Course Materials: Course website will contain all slide-decks, assignments, answer keys, R scripts with worked examples, miscellaneous handouts
  3. Assignments: Almost weekly, and the weekly labs will help you get set for assignments.
     - Assignments must be submitted via Blackboard as MS Word documents generated with RMarkdown and showing all code.
     - You can submit assignment drafts to me for feedback (see the deadlines specified in the syllabus).
  4. Exams: Three, not cumulative
- Grade Requirements: See the grading scale in the syllabus
  - Easy to do well if you (a) read before class, and (b) practice problem-solving
- Miscellany: No make-ups without prior approval. No extra credit
- Office Hours: Set hours in Porter right after class.
  - You can also request a meeting (through Outlook).

Introduction to Statistics

Statistics

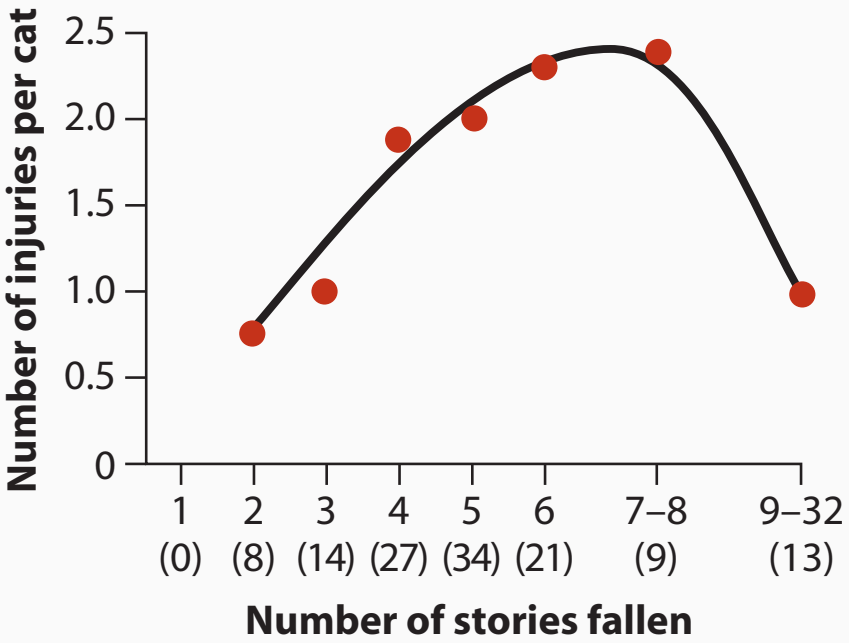
Definition

... involves methods for describing and analyzing data and for drawing inferences about phenomena represented by the data

- technology (thermometers, buoys in the ocean, air quality monitors, etc.) that describes and measures aspects of nature from samples
- allows us to quantify the uncertainty around what we can measure from samples
- all about estimation: inferring an unknown quantity of a target population from sample data
- involves hypothesis testing unless we are only interested in exploratory data analysis

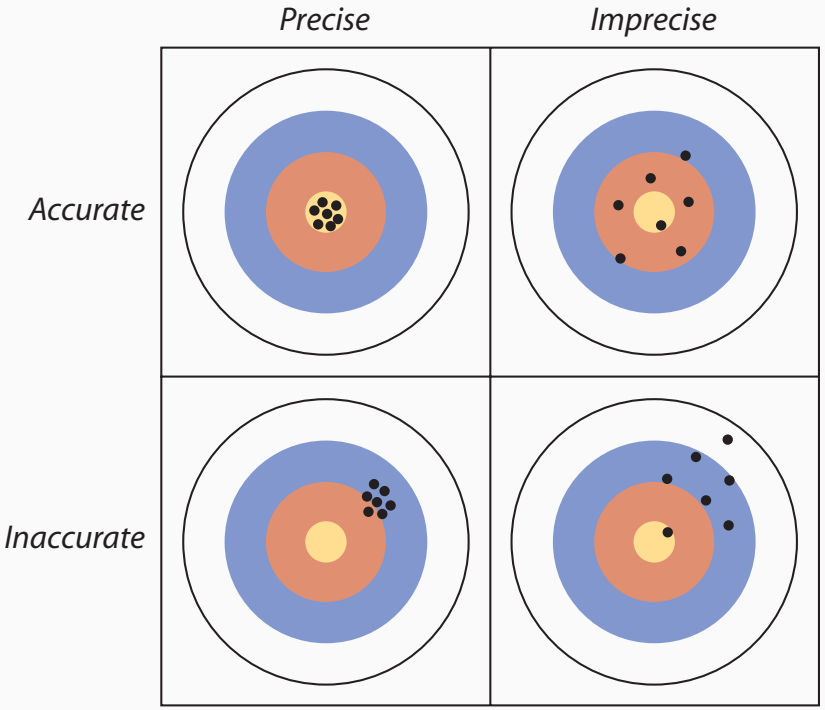
Sampling Populations

- Sampling is the lifeblood of statistics; your work is only as good as your sample
- Population: Universe (or set) of all elements (units) of interest in a particular study
- Sample: Subset of cases (units) drawn for analysis from the population
- Example shown is of a 1987 study (published). Question: No cat from first floor? What about injuries from the 9-32 floors? Suspicious sample?



Properties of Good Samples

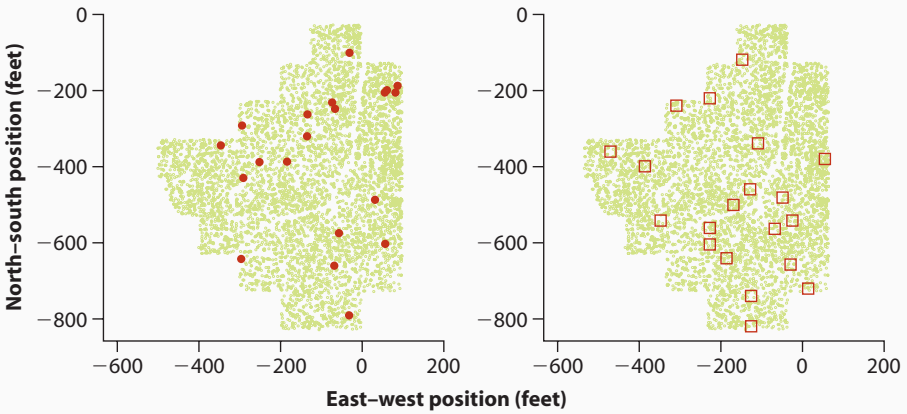
- Samples should $\approx$ Population
- Chance (and other factors) can lead sample estimates to differ from population parameters $=$ sampling error
- Estimates ought to be best (you can't do any better) and unbiased (shouldn't consistently overestimate/underestimate)
- Random Sampling requires
  1. Every unit in the population have an equal chance of being sampled
  2. Every unit be sampled independently of all other units



- Violated? $=$ bias
- Violated? $=$ imprecision

Taking a Random Sample - Harvard Forest (MA)

1. Assign pseudo-ID to every population unit
2. Choose sample size (n)
3. Let random-number generator give you the n pseudo-IDs $1 \ldots 5699$
4. More realistic? – Sample from equal-size plots that are themselves randomly selected
5. Convenience Samples $\rightarrow$ bias
6. Many sampling schemes $\Rightarrow$
7. Get as large a sample as you can



Probability Sampling

- Simple (pure random sampling)
- Stratified (split units into homogenous groups and sample within all groups)
- Cluster (identify clusters and sample within clusters)
- Systematic/Interval (pick every k^{th} person to get desired n)

Random or not?

U.S. Army wants to test stress levels in recruits stationed in Helmand province. All recruits (1,000) are given random ID. Researchers pick 100 at random.

1. What is the population of interest?
2. Could this sample have sampling error?
3. What benefits does random sampling give these researchers?
4. Would a large sample size help?

Typology of Data and Variables

Data & Variables

- Data can be ...
  1. Cross-Sectional – MANY finches observed at ONE point in time
  2. Time-Series – ONE finch observed over time
  3. Panel data – MANY finches observed over time (best)
- Variables broadly classified as ...
  1. Categorical – characteristics/attributes without a numeric scale.
     Examples: Sex, language, Species type, race/ethnicity, method of disease transmission
  2. Numerical – characteristics/attributes with a numeric scale.
     1. Continuous – divisible units (temperature, landmass, weight, etc.)
     2. Discrete – indivisible units (number of trees, number of kids, etc.)
- Variables can be sub-classified into
  1. Nominal – categorical, no hierarchy of levels (e.g., Sex, Seasons, etc.)
  2. Ordinal – categorical, hierarchy of levels (e.g., Poor, Middle-class, etc.)
  3. Interval – numerical, without natural zero point (e.g., degrees Celsius)
  4. Ratio – numerical, with natural zero point (e.g., Kelvin scale)

Variable Type?

Which of these is discrete? Which is continuous?

1. Number of injuries sustained in a fall
2. Fraction of birds infected with the avian flu virus
3. Number of crimes committed by juveniles in Athens County
4. Body mass
5. Survival time after accidental poisoning

Which is nominal, which ordinal?

1. The 260 known species of monkeys
2. Four seasons (Fall, Winter, Spring, Summer)
3. Saffir-Simpson Hurricane scale [1 (weak) ... 5 (major)]
4. Freshman/Sophomore/Junior/Senior

Types of Studies/Research Designs

Types of Studies

- Our goal is almost always to assess how one or more *explanatory* (aka covariate(s), independent, etc.) variable(s) influences the *response* (aka dependent, outcome, etc.) variable
- Experiment - intervention deliberately introduced to observe its effect
- Randomized Experiment - units are assigned to the treatment via a random process
- Quasi-Experiment - units are not randomly assigned but instead assigned via self-selection or administrative selection
- Natural Experiment - involves a rare, naturally occurring event
- Correlational - involves merely exploring the strength and direction of a correlation between likely cause and likely effect

Frequency Distributions & Probability Distributions

Frequency & Probability Distributions

- Frequency – count of unique "values" of a variable
- Frequency Distribution – how often does each unique value occur in the sample?
- Probability Distribution – how is this variable distributed in the population?
- Example: Distribution of beak depths in $n = 100$ finches from a Galápagos Island ... see here for the Boag & Grant (1984) study, and here for the data
- Ideally: Frequency distribution $\approx$ probability distribution