Dr. Elizabeth Newton

Slides prepared by Elizabeth Newton (MIT) with some slides by Roy Welsch (MIT) and Gordon Kaufman (MIT).

15.075, Applied Statistics

Lecture: M,W 10-11:30

Recitation: R 4-5

Text: *Statistics and Data Analysis* by Tamhane and Dunlop

Computing: S-Plus

Exams: Mid-term (in class) and Final during exam week

Prerequisites: Calculus, Probability, Linear Algebra

15.075, Applied Statistics, Course Outline

- **Collecting Data**
- **Summarizing and Exploring Data**
- **Review of Probability**
- **Sampling Distributions of Statistics**
- **Inference**
  - **Point and CI Estimation, Hypothesis Testing**
- **Linear Regression**
- **Analysis of Variance**
- **Nonparametric Methods**
- **Special Topics (Data Mining?)**

Statistics

"The science of collecting and analyzing data for the purpose of drawing conclusions and making decisions." **from Tamhane, Ajit C., and Dorothy D. Dunlop. *Statistics and Data Analysis from Elementary to Intermediate*. Prentice Hall, 2000, pp. 1.**

"Statistics are no substitute for judgment." Henry Clay

How is the meter defined?

One ten-millionth of a quarter meridian (distance from pole to equator).

BUT – it isn’t exactly.

Why?

***The Measure of All Things*, by Ken Alder, describes the attempt of 2 French astronomers, Delambre and Mechain, to determine the circumference of the earth during the time of the French Revolution.**

Determined the distance between Barcelona and Dunkirk by triangulation.

Needed to know latitude at each end (by measuring heights of stars).

Seven months stretched to seven years.

Mechain obtained conflicting information and suppressed some of his data.

Page 214 (*Measure of All Things*):

“What counts as an error? Who is to say when you have made a mistake? How close is close enough? Neither Mechain nor his colleagues could have answered these questions with any degree of confidence. They were completely innocent of statistical method.”

- Quote from Alder, Ken. *The Measure of All Things: The Seven-Year Odyssey and Hidden Error that Transformed the World.* Free Press, 2003.

Data: A Set of measurements

Character

Nominal, e.g. color: red, green, blue

Binary e.g. (M,F), (H,T), (0,1)

Ordinal, e.g attitude to war: agree, neutral disagree

Numeric

Discrete, e.g. number of children

Continuous. e.g. distance, time, temperature

also:

Interval, e.g. Fahrenheit temperature

Ratio (real zero), e.g distance, number of children

S-Plus Data Set: cu.summary

Concepts

Population:

The set of all units of interest (finite or infinite).

E.g. all students at MIT

Sample:

A subset of the population actually observed.

E.g. students in this room.

Variable:

A property or attribute of each unit, e.g age, height

Observation:

Values of all variables for an individual unit

A dataset is often organized as a matrix with rows corresponding to observations and columns to variables.

Concepts (continued)

Parameter:

Numerical characteristic of population, defined for each variable, e.g. proportion opposed to war

Statistic:

Numerical function of sample used to estimate population parameter.

Precision:

Spread of estimator of a parameter

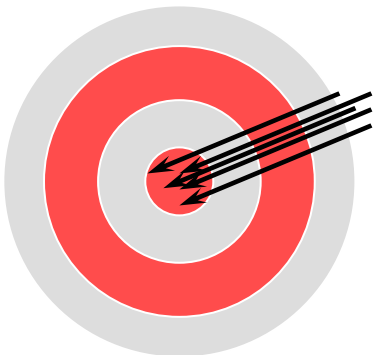
Accuracy:

How close estimator is to true value - opposite of

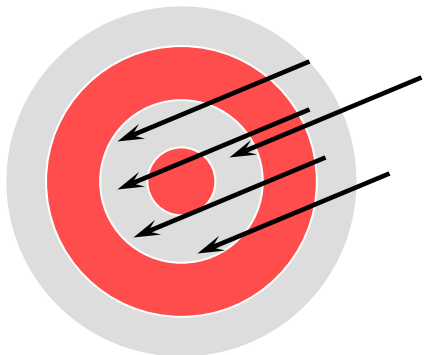
Bias:

Systematic deviation of estimate from true value

Accuracy and Precision

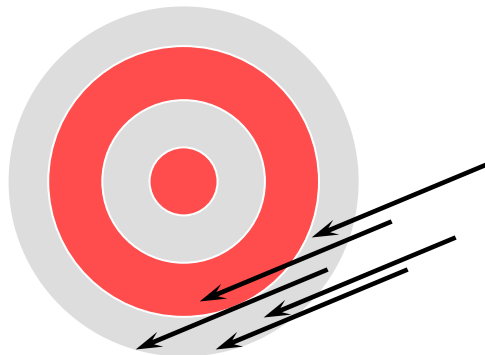


accurate and precise



accurate, not precise

precise, not accurate



not accurate, not precise

Diagram courtesy of MIT OpenCourseWare

Steps in Study Design and Implementation

1. Background research and literature review.

2. Define the goals and specific hypotheses of the study.

3. Determine what variables should be measured and how.

5. Develop a plan to collect the data
   - Sampling design
   - Sample size
   - Inclusions and exclusions

5. Train Personnel

6. Gather Data

7. Analyze Data

8. Report Results

Ethical Issues

For human subjects:

For animal subjects:

(See Hulley & Cummings, *Designing Clinical Research.)*

Statistical Studies

Descriptive:

One group, e.g. survey, poll

Comparative:

2 or more groups, e.g. compare effectiveness of different teaching methods.

Experimental:

Investigator actively intervenes to control study conditions
Look at relationship between predictor (explanatory) and response (outcome) variables
Establish causation, e.g. drug trial

Observational:

Investigator records data without intervening
Difficult to distinguish effects of predictors and confounding variables (lurking variables)
Establish association, e.g. Framingham Heart Study

Observational Studies:

Cross-sectional

- **Look at sample at a single point in time**
- **E.g. Census, Sample survey**

Prospective (expensive!)

- **Follow sample (cohort) forward in time.**
- **E.g. Framingham heart study, Nurses' Health Study**

Retrospective (case-control)

- **Look back in time**

Sources of Error in Observational Studies

Sampling Error – sample differs from population

Measurement Bias – poorly worded questions

Self-Selection Bias – refusal to participate

Response Bias – incorrect or untruthful responses

Types of Samples

Probability Sample (every element in population has known non-zero probability of inclusion)

- **Simple Random Sample (SRS)**
- **Stratified Random Sample**
- **Multi-Stage Cluster Sample**
- **Systematic Sample**

Non-Probability Sample (estimates may be biased, but frequently used as only feasible method)

- **Convenience Sample e.g. supermarket survey**
- **Judgment Sample – chosen by investigator**

Simple Random Sample (SRS)

Requires a Sampling Frame, a list of all the units in a finite population

Sample of size n is drawn without replacement from population of size N, such that each sample (there are $\binom{N}{n}$ of them) has same chance of being chosen.

Each unit in population has same chance of being chosen: n/N (the sampling fraction).

Generate random numbers to select from sampling frame.

Stratified Random Sample

Divide a diverse population into homogeneous subpopulations (strata).

Draw simple random sample from each one.

Advantages:

Separate estimates for strata obtained in addition to overall estimates.

Precision of estimates higher than for simple random sample

Disadvantage: Requires sampling frame

Multistage Cluster Sampling

Used to survey large populations when sampling frame not available, e.g. USA

For instance, in an educational survey, draw a sample of states, then towns within states, then schools within towns.

Prepare a sampling frame of students from selected schools and use SRS.

Systematic Sampling

Useful when list of units exists or when units arrive sequentially (cars through a toll booth).

Select first unit at random, then every kth unit.

In finite population, each unit has same probability of selection (n/N)
(however not all samples are equally likely).

Must avoid choosing k to coincide with regular cyclic variations in the data

Questionnaire Design

Structured questions: responses should be mutually exclusive and collectively exhaustive.

E.g. How many glasses of water do you drink per day?
---------- 0 to 2
---------- 3 to 5
---------- 6 or more

Non-structured:
E.g. How many glasses of water do you drink per day?
Allow more individualized response, but more prone to data entry errors.

Attitude questions

1. The homework load in this course is reasonable.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
_____	_____	_____	_____	_____

Usually 5 to 9 categories.
(Should we assign numbers to these categories?)
(High to low or low to high?)

Problems with Question Wording

Double-barreled question

Leading question

One-sided question

Ambiguous question

Pretest! Pretest! Pretest!

(For more information, see Johnson & Wichern, *Business Statistics*)

Sensitive Questions

E.G Have you ever used heroin?

Randomized Response may elicit more accurate responses.
Interviewer does not know what question respondent is answering.

E.g. Roll a die. If less than 3 then say whether statement 1 is true or false. Otherwise say whether statement 2 is true of false.

Statement 1: I have used heroin.
Statement 2: I have not used heroin.

Let p=proportion of people who have used heroin
q=proportion of people answering question 1 (can't be 0.5).

$$P(True)=P(True|1)P(1) + P(True|2)P(2) = p\,q + (1-p)\,(1-q)$$

Solve for p.

Question Sequencing

1. **Demographics at end**
2. **Sensitive questions nearer to end**
3. **Same topic questions appear together**
4. **Go from general to specific**
5. **Avoid skipping around.**

Experimental Studies

Purpose: Evaluate how a set of predictor variables (factors) affect a response variable.

Treatment Factors are of primary interest. Values (Levels) are controlled.

Nuisance Factors also affect response.

Treatment: particular combination of levels of treatment factors.

Experimental units (EU's): subjects to which treatments applied.

Treatment group: all EU's receiving same treatment

Run: observation on an EU under particular treatment condition.

Replicate: another independent run.

Sources of Error in Experimental Studies

Systematic Error: differences among EU's caused by Confounding Factors

Random Error: inherent variability in responses of EU's.

Measurement Error: due to imprecision of measuring instruments.

Strategies to Control Error in Experimental Studies

Blocking: Divide sample into groups of similar EU's (same value for nuisance factors).
E.g. In agricultural trials effect of nutrient and moisture gradients can be controlled for by blocking on agricultural plots

Matching: EU's can be matched on nuisance factors, then each member of match can be randomly assigned to different treatment (each match is a block).

Regression Analysis: If value of nuisance factor is known can include as covariate in final model.

Randomization: Randomly assign EU's to treatments.

Basic Idea: Block over those nuisance factors that can be easily controlled and randomize over the rest

Basic Experimental Designs

Completely Randomized Design (CRD)
EU's assigned at random to treatments

Randomized Block Design (RBD)
EU's divided into homogeneous blocks
Treatments assigned randomly within blocks.

Randomized Complete Block Design (RCBD):
Blocks contain all treatments.

Randomized Incomplete Block Design (RIBD)
Blocks do not contain all treatments.