It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.“

Sherlock Holmes to Dr. Watson

A Study in Scarlet

STATISTICS INTRO

For EDUC/PSY 6600

THE GOAL

When most of you read data analysis sections now, you probably do not understand 90% of what is written

At the end of this course you will probably not believe 90% of what is written

$$\sigma_X = \sqrt{\frac{1}{n}\left\{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right\}}$$

WHAT IS STATISTICS?

- Branch of mathematics
  - Young discipline, emerging in 1920s-1930s
- Problem solving
- Art vs. science
- Deterministic versus stochastic models

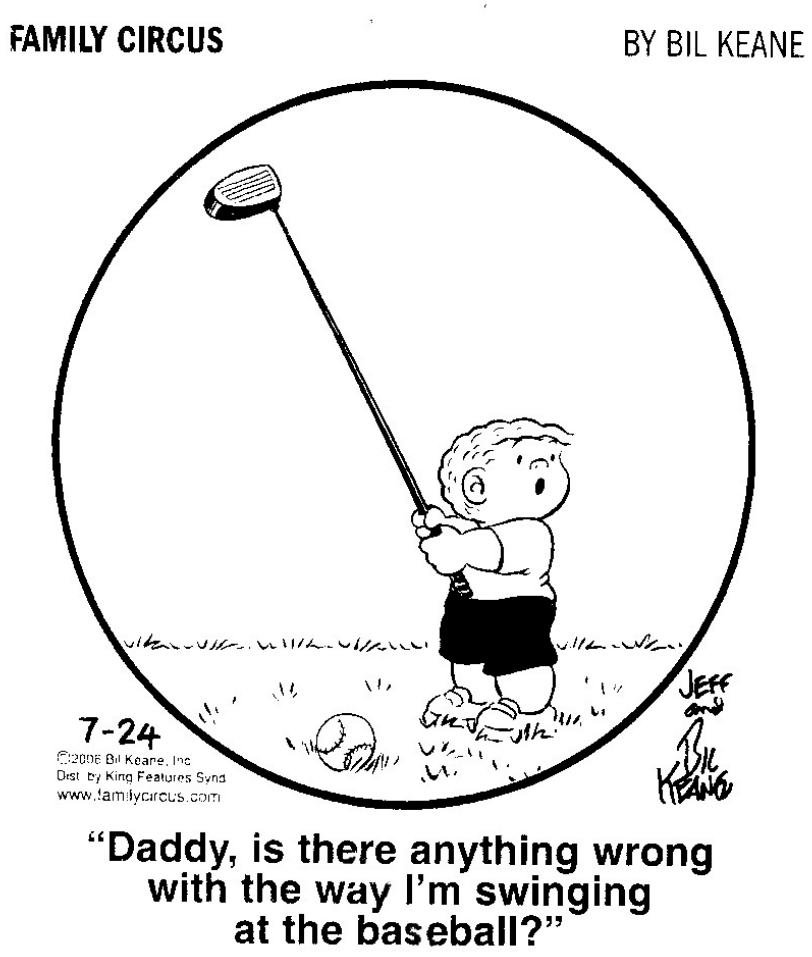
BASIC TERMINOLOGY

- Population vs. Sample
- Parameter vs. Statistic
- Variable vs. Constant Value
- Discrete vs. Continuous Data
- Independent vs. Dependent Variable
- Descriptive vs. Inferential Statistic

EXPECTATIONS

- Expectations
  - 1-2 courses only ‘tip of iceberg’
  - Years to master many subtleties

- Right tool for the job
  - Software package as
  - workshop, analytical
  - methods as tools



EXPECTATIONS

"…How hard one finds it to learn something is a function of:

- *the intrinsic difficulty of the thing,*
- *the person's previous experience,*
- *preferred modes of thinking,*
- *etc., and*
- *how learning is approached."*

-John Fox, PhD

SCIENTIFIC METHOD

- Research Problem → Numerical Form
- Statistical methods apply to samples
  - Infer what's occurring in population from sample
- However, statistical methods only describe what is occurring within data set
  - Inference based on logical judgment of results of statistical analyses and strength of study design
  - Statistics cannot 'make-up' for poor study design (garbage in, garbage out)

ISSUES IN STUDY DESIGN

Trade-off between internal and external validity

- Validity lies on continuum, rather than yes-no

Samples should be…

- **Large** enough to limit random error & provide power
- **Representative** enough to…
  - Limit systematic error (bias): Internal validity
  - Offer generalizability: External validity
- **Obtainable** given time and cost

SAMPLE SELECTION

Specify well-defined target population

- Sample is subset of target population

Define inclusion/exclusion criteria in *Methods*

- Inclusion criteria: 1^0 characteristics of target pop.
  - Specification of demo. characteristics, disease status, etc.
  - Limits threats to internal validity
- Exclusion criteria: Characteristics threaten data quality
- Use sparingly: Ultra-specific sample ↓ generalizability

SAMPLING

Many cases meet inclusion/exclusion criteria

Sampling unit: Basic unit around which a sampling procedure is planned

- Individuals, animals, households, classrooms, neighborhoods

Sampling frame: List of ALL possible units in pop.

Sample: Units chosen from eligible pop.

SAMPLING

- Non-probability sampling:

  ↑ risk of bias

  - Convenience sampling
  - Purposive or judgmental sampling
  - Quota sampling

- Probability (random) sampling:

  ↓ risk of bias &↑ likelihood results can be replicated

  - Simple random sampling
  - Systematic sampling
  - Stratified random sampling
  - Cluster sampling

SAMPLING

- Descriptive and inferential statistics based on laws of probability
- Laws of probability apply to random events
- Random events → Random sampling
- Measurements from random sample → Random variable

SUPPLEMENTAL SLIDES – TYPES OF SAMPLING

NON-PROBABILITY SAMPLING

In most research, probability sampling is not practical, possible, or even ethical

Subjective judgment as to whether convenience sample represents target population

- For given research question, would conclusions be similar if a true probability sample was obtained?
- Burden is on researcher

CONVENIENCE SAMPLING

Non-random, cases selected due to ‘convenience’

- As they appear or agree to participate

Every unit meeting criteria selected for study

Easily accessible in terms of costs and logistics

PURPOSIVE OR JUDGMENTAL SAMPLING

Sample selected on basis of investigator's knowledge of target population

Used when shown that a certain subset is representative of target population

- Often used for pretest or pilot study
- Specific disease, behavioral, or educational states

Reliance on available subjects

- e.g., survey every student enrolled in a large class

QUOTA SAMPLING

Proportions in population are reflected in sample

- Used to obtain a representative sample

Cases meeting study inclusion/exclusion criteria enrolled to 'fill' quotas

- Sometimes randomly, sometimes not

SIMPLE RANDOM SAMPLING

Sampling units

- Known and equal (non-zero) probability of selection
- Assigned a number
- Selected by random process (computer)

Simple to do, but requires knowledge and access to complete sampling frame in advance

Generally sampling WITHOUT replacement

SYSTEMATIC SAMPLING

Units are sampled via periodic, consistent process
- e.g., selecting every 4th patient admitted to clinic

Randomly select starting point

If end is reached, consider sampling frame as circular

Rarely better choice than simple random sampling

Disadvantages
- Cyclical sampling may lead to consistent highs, lows, or lack of variability (periodicity), or duplicates
- Researchers may predict cases to be enrolled, manipulate inclusion/exclusion

STRATIFIED RANDOM SAMPLING

Attempt to...

- Obtain more representative sample
- Ensure sampling of less frequent cases
- Achieve more precise estimates of population parameters

Target population divided into strata/subgroups with common characteristics

- e.g., sex, race, age, educational groups

Simple random sampling applied within each stratum

Use only 3-10 strata: Negligible improvement in precision thereafter

CLUSTER SAMPLING

Entire groups of naturally occurring units selected via simple random sampling

- e.g., classrooms, schools, clinics, sports teams, neighborhoods, work crews

All individual units within cluster are measured

Use when impossible/impractical to…

- Obtain exhaustive list of all sampling units
- Sample individuals from sampling frame

Clusters should be as similar as possible

Units within clusters tend to be more homogeneous than units drawn from simple random sampling

MULTI-STAGE OR 2-STAGE CLUSTER SAMPLING

1st stage: Primary sampling unit (PSU)
- Clusters are sampled randomly

2nd stage: Secondary sampling units (SSU)
- Individuals are sampled randomly within each cluster

More than 2 stages possible
- School districts → schools → classrooms/teachers → students

Potential loss in accuracy, sampling error at each stage

Cluster sampling adds complexity to data analysis