Day 4

Outline

1. Statistical Terminology
2. Sampling Distributions

Statistical Terminology

Tidy Data

Each column represents a variable

Header row will contain the name of the variable

Each row represents a case

Each value goes in its own cell.

Good form: left most column contains label variable whose values are unique IDs for the cases

One row could represent:

- a patient
- a particular test for that patient
- all patients seen by a doctor

Merging datasets: you need to pair like data

Data Dictionary

For each variable:

- name of the variable
- type of the variable
- units of measurement
- description

Type of Variable

Numerical (Quantitative): int, float, double

Categorical (Qualitative): string, classes, char, software-specific variable type

Typically, we do not select only one case from the population. We instead select a subset of the population: sample. A sample will always exist in the real world.

Statistical Terminology (Continued)

Variables vary between cases.

Statistics vary between samples

Parameters vary between populations.

Frequentist Statistics

Parameters are constants, but we don't know their value.

Statistics are random variables that describe "randomly select a sample of some fixed size, record values of a variable for each case in the sample, and summarize the value".

In this class

Numerical variables: we use μ to represent population mean and $\bar{X}$ to represent sample mean

Categorical variables: we use "p" to represent population proportion of outcome in a particular category.

We use $\hat{p}$ to represent sample proportion of outcomes in a particular category.

Example

A clinical trial compares two bladder cancer drugs:

- Drug A (Company's "new" drug)
- Drug B (Current best drug)

They recruit 200 subjects with bladder cancer and assign 100 to take Drug A and 100 to take Drug B

Questions

- What is the case
- We can consider this study to have 2 "hypothetical" populations. What are they?
- What are the two samples from those hypothetical populations? (subset of a larger group/population)
- Name an outcome the drug company might be interested in. Is that outcome a numerical or categorical variable?
- What statistic might we use to summarize that outcome in a sample
- What is the corresponding parameter in the hypothetical population.

Answers

- A case is one patient with bladder cancer
- Everyone on Drug A (all bladder cancer patients, if they took Drug A) and everyone on Drug B
- 100 people who took Drug A and 100 people who took Drug B
- How much more effective is Drug A compared to Drug B. This would yield a numerical value. (Reduction in tumor size, cancer remission/not in remission)
- Sample mean ($\bar{X}$) tumor reduction. Sample proportion/sample percent of patients who are in remission.
- Population mean tumor reduction and population proportion in remission

Sampling Distribution

These are things that do not have real world equivalences.

The probability distribution of a statistic is its sampling distribution

Distribution of a statistic over all possible samples of a given size from the sample population. **Must** specify size of sample.

To find a sampling distribution:

1. Simulation: approximate the sampling distribution by simulating samples.
2. Asymptotic behavior: as the number of samples $\to \infty$, what does the distribution look like?

Properties of Sampling Distributions

Let X be the statistic we use to estimate a parameter θ for a population. X is an unbiased estimator of θ if $\mu_{\mathrm{x}} = \theta$. Otherwise X is biased and the amount of **bias** is $\mu_{\mathrm{x}} - \theta$.

The variability of X describes the amount by which individual realizations of X are "spread" about μ_{x}.

We can summarize variability by variance, standard deviation, standard error, margin of error