Announcement
- From now to Oct 11, you should carefully design your data collection protocol. We will cover topics on study designs in the next several lectures.
- Midterm is approaching faster than I expected.
- I will read your project forum and post comments.

Did you pick 3?
- Almost 75% of all people pick the number 3.
- 20% pick either 2 or 4
- Only 5% pick 1
- We are lousy random number generators!

Random pick one:
1 2 3 4

Define Randomness
- Completely random: equal chance to each outcome to be sampled.
- Random (generalized): each outcome has a chance (10% or 40%, etc), each trial of the random phenomenon (lottery, poker game, draw candy) stick to the same specification of chance.

Everyday randomness
- New York Lottery
- Housing Lottery
- Course Lottery
**How to randomize using the random digit table**
- Use the random digits table in appendix E (A-97):
  - Say I want 3 random numbers between 1 and 20.
- Scenario: I numbered my population of 20 students and I need to sample 3 randomly.
  - Line 33 in random digits table:
    - 28541 01029 08068 17795 21484 ...
  - Two digits at a time
    - 28 54 10 10 29 08 06 89 66 56 17 79 52 14 84 ...
  - Write down those within the targeted range and ignore repeats
- You can also use computer software or calculator

**Learn about randomness using simulations**
- Assume 30 random numbers between 0 and 10.
- Use it as the data of a variable with 30 observations.
- What the histogram will look like?
- How about 3000?

**Weird hands in computer simulated poker game**
See computer screen

**Types of data analysis**
- Exploratory data analysis
  - “Here is a set of data on NCAA basketball games over the past 10 years”
  - “Anything interesting?”
- Statistical inference
  - “I want to know which team will be national champion this year”
  - “I have data from the past 10 years”
  - “So what is your prediction based on these data?”

**Type of available data**
- Anecdotal evidence: Bad source
  - Do you think New York City is a safe place to live?
- Published data (library or Internet)
  - You should read the collecting procedure carefully.
- Official statistics
  - Studies on sports
  - Studies on public health

**Terms**
- **Population**: collection of individuals of interests
- **Sample**: a (small) part of population, through which one can gain information about the population.
- **Sampling**: random action of selecting individuals out of a population
- **Individuals, Variables**
- **Model, Parameter, Statistic**
Why sample?

- Why not?
- Study the population as a whole (census)?
  - expensive
  - time-consuming
  - the whole population may be hard to reach
  - the population may be a self-evolving one
- Statisticians say that we can study the population by examining a part of it, in a cheaper, faster, feasible, unbiased way.

Why study sampling?

- You may need to do sampling yourself.
- You may need to understand the study design when analyzing data produced by others.
- You may need to understand the study design when evaluating results from a study based on data collected.
- It is always good to know…

Sample what?

A part of the population that represents the population.

Random sampling: give each individual in the population an equal chance to be sampled.

Representative sample

- What is the bad consequences if we don’t have a good (representative) sample?
  - Misleading outcomes
  - Results can not be applied to the rest of the population
  - Difficult to analyze
  - Waste of time and $ $ $  
- No statistical methods can save a bad sample.

How to get a representative sample?

- A good sampling design produces samples carrying correct images of the population
- Biased design: sampling method with systematic errors; The sampling chance of an individual is associated with some characteristic of this individual related to the outcome of interests.
  - Example: if you had it to do over again, would you have children? (70% said ‘NO’)
  - Why not matching?
• 1 cup butter flavored shortening
• 3/4 cup white sugar
• 3/4 cup brown sugar
• 2 eggs
• 1 teaspoon salt

Did you mix your ingredients well so that each cookie can be considered as a random sample of your ingredients?

Randomization is the solution!!!

Making approximately 24 cookies.

Logic behind Randomization

Randomized sampling

population

Chance, Laws of probability

Statistical inference

Statistical inference

• Population
• Sample is a part of population and carries information about population

Parameters: the Greek letters

• What is IN the population that one is interested in?

Statistics: the ones with bar or hat, etc

• What is IN the sample that one can use to study the population?

Sampling variability

• Sample carries information about the population: similarities across samples
• Different samples may differ due to chance
• Sampling variability: the variation of a statistic on many different random samples out of one population
• By understanding the sampling variability, one can judge the reliability of a statistic when inferring the population parameters.

Sampling probability

• Define population size: \( N \)
• For a simple random sample: any individual has \( n/N \) probability of being selected
• 140 students, random select 3 students → every student has 3/140 of being selected
• In addition to that, every combination of 3 students has an equal chance → SRS

Simple random sample

• Sample size \( n \): the number of individuals to be sampled
• Design: Every set of \( n \) individuals has an equal chance to be the sample actually sampled
• A sample produced by such a design is called Simple Random Sample (SRS)
• Example: shuffling cards
**Sampling Scheme (I)**

- Stratified sampling
  - **Stratum (strata):** group(s) of similar individuals; eg. CC students, BC students, SEAS students, GS students, CE students, GSAS students, and other.
  - Stratified sampling: select an SRS out of each stratum and merge the individual SRS' to make a final sample
  - Example: selecting a group of students representing every school
- **Sampling probability**
  - Same SRS size for each stratum: unequal probabilities for individuals from different strata
  - Solution: SRS size proportional to strata size

**Sampling Scheme (II)**

- **Multistage sampling:** divide the population into subpopulations; sample an random sample of the subpopulations first and then sample an SRS out of each sampled subpopulation
- Can be extended to more stages
- Sampling probability calculation: sampling 10 dorm rooms on campus
- In a multistage sampling: how to make every individual has the same chance of being selected—sampling probability proportional to the subgroup size.

**Discussion on sampling schemes**

- Does stratified sampling lead to SRS? **NOPE!**
- Does multistage sampling lead to SRS? **NOPE!**

**Cautions about Sampling**

- **Sampling bias (under-coverage)**
  - Convenience sample
  - Telephone surveys: random digits dialing
  - Daytime versus evening
  - Multiple phone lines
  - Large household
- **Non-responses**
  - Long surveys
  - Sensitive topic
  - Incentives might encourage participants, while may also cause sampling bias

**Cautions about sample surveys**

- **Response bias**
  - ‘Lie’ to sensitive questions
  - Make up desirable answers:
    - Following hints such as wording, tone, facial expression, etc
- **Wording of questions**
  - Misleading information
  - Do you know one out of … students admitted they once cheated in an exam? Do you think …
  - Confusing definition
  - Do you regard ‘cheating’ a normal phenomenon?
  - Revealing desirable answers
  - We are conducting a survey on honesty, do you think…

**Sensitive questions**

- Carefully design questions
- During the interview, assure the interviewee the confidentiality of the study.
- A special probability ‘game’.
Reading (required!)

- Chapters 11-12
- Online guidelines on "statistics projected" website.