STA 1013 : Statistics through Examples

Lecture 4: Sampling methods

Hwiyoung Lee

September 6, 2019

Department of Statistics, Florida State University

Overview

1. Review
2. Sampling Methods
3. Calculator

Review

Review

Types of Statistical study

- **Observational studies**
    - Cross-sectional study
    - Retrospective study
    - Prospective study
- **Experiments**
    - Treatment group
    - Control group

Review

Examples

Drinking and Driving Study : In order to study the seriousness of drinking and driving, a researcher obtains records from past car crashes. Drivers are partitioned into a group that had no alcohol consumption and another group that did have evidence of alcohol consumption at the time of the crash. → **Retrospective Study**

Meat and Mortality : Researchers at the National Cancer Institute studied meat consumption and its relationship to mortality. Approximately one-half million people were surveyed, and they were then followed for a period of 10 years. → **Prospective Study**

Smoking Study : Researchers from the National Institutes of Health want to determine the current rates of smoking among adult males and adult females. They conduct a survey of 500 adults of each gender.→ **Cross-sectional study**

Review

Variables of Interest

- Explanatory variable (Independent variable)
  : variable that may explain or cause the effect
- Response variable (Dependent variable)
  : variable that responds to changes in the explanatory variable

Confounding variable : variables that lead to the confusion.

- Hidden variable
- Cannot determine the specific effects → Wrong conclusion

Review

Large daily doses of vitamin C help prevent colds

Consider a medical study designed to test whether large daily doses of vitamin C help prevent colds. The researchers conduct their experiment with two groups of subjects: One group takes large doses of vitamin C daily and another group does not.

- Treatment group : group takes large doses of vitamin C daily
- Control group : another group does not take vitamin C
- Explanatory variable : Large doses of vitamin C (Yes or No)
- Response variable : Cold (Yes of No)

Review

Bad research design : Randomization

In testing the **effectiveness of a new vaccine**, suppose that researchers used females for the treatment group and males for the control group.

- **Treatment group** : group takes a new vaccine
- **Control group** : group does not take a new vaccine
- **Confounding variable** : Gender
  - **Treatment group ← Female**
  - **Control group ← Male**

Cannot determine the true effect : Vaccine vs Gender ?

Control and Treatment groups have to be randomnly assigned.

Sampling Methods

Sampling

- The only way to know the true value of a parameter is to observe every member of the population.
- A collection of data from every member of a population is called a **census**

Example : To learn the exact mean height of all students at your school, you'd need to measure the height of every student

- Sample is used to estimate the unknown parameter of population
- **How can we draw a sample ?**

Why Sampling is important ? (sample selection)

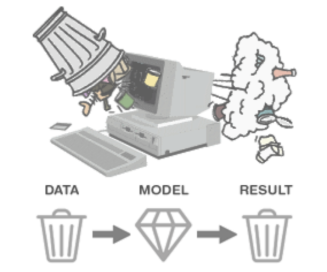
Bad sampling design leads to Wrong conclusion

> Imagine that, for the 5,000 homes in its sample, **Nielsen chose only homes in which the primary wage earners worked a late-night shift**. Because late-night workers aren't home to watch late-night television, Nielsen would find late-night shows to be unpopular among the homes in this sample.

- This sample would not be representative of all American homes
- It would be wrong to conclude that late-night shows were unpopular among all Americans.
- We say that such a sample is biased because the homes in the sample differed in a specific way from "typical" American homes.

Why Sampling is important ? (sample selection)

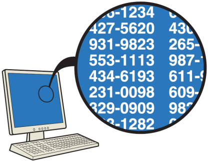
GIGO : Garbage in, Garbage out



- The quality of the output is determined by the quality of the input
- When we use the garbage data, we must get the garbage result

Simple Random Samples

- In most cases, the best way to obtain a representative sample is by choosing **randomly from the population.**
- A random sample is one in which every member of the population has an equal chance of being selected to be part of the sample.



Simple Random Sampling:
Every sample of the same size has an equal chance of being selected. Computers are often used to generate random numbers.

How can we draw the sample randomly ?

Use the **random number generator**.

#	1	2	3	4	5	6	7	8	9	10
Obs	3.5	4.1	5.4	4.6	9.3	2.7	3.5	6.3	5.1	8.3
#	11	12	13	14	15	16	17	18	19	20
Obs	4.4	5.3	6.7	4.6	2.8	8.8	1.5	3.9	7.2	5.8

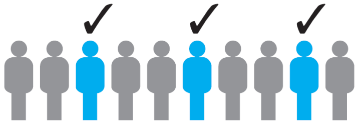
- The above table is a population data with $N = 20$
- We want to draw random sample with $n = 5$
- Compare the population mean (5.19) and the sample mean ?

Systematic Sampling

Select some starting point, then select every kth element in the population.

Example

Testing the quality of microchips produced by Intel. As the chips roll off the assembly line, you might decide to test every 50th chip. there's no reason to believe that every 50th chip has any special characteristics compared with other chips.



Systematic Sampling:
Select every *k*th member.

When Systematic Sampling Fails ?

Sample may be biased if **hidden periodicity** in population coincides with that of selection.

Example

You are conducting a survey of students in a co-ed dormitory in which males are assigned to odd-numbered rooms and females are assigned to even-numbered rooms. Can you obtain a representative sample when you choose every 10th room? **No.**

- If you start with an odd-numbered room, every 10th room will also be oddnumbered (Ex : 3, 13, 23, $\cdots$).
- If you start with an even-numbered room, every 10th room will also be even-numbered.
- How do we sample ?

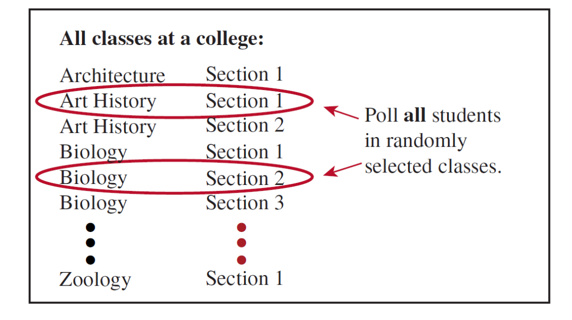
Cluster Sampling

Divide the population into sections (or clusters), then randomly select some of those clusters, and then choose all members from those selected clusters

- Population divided into clusters of homogeneous units, usually based on geographical contiguity.
- Sampling units are groups rather than individuals.
- A sample of such clusters is then selected. All units from the selected clusters are studied.

Cluster Sampling

Want to estimate average height of students in FSU



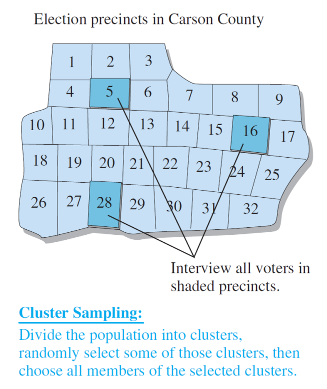
All classes at a college:

Architecture	Section 1
Art History	Section 1
Art History	Section 2
Biology	Section 1
Biology	Section 2
Biology	Section 3
⋮	⋮
Zoology	Section 1

Poll **all** students in randomly selected classes.

Cluster Sampling

This can reduce travel and other administrative costs.



Election precincts in Carson County

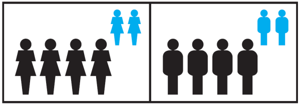
Interview all voters in shaded precincts.

Cluster Sampling:

Divide the population into clusters, randomly select some of those clusters, then choose all members of the selected clusters.

Stratified Sampling

- We use this method when we are concerned about differences among subgroups, or strata, within a population.
- We first identify the strata and then draw a random sample within each stratum.
- The total sample consists of all the samples from the individual strata



Stratified Sampling:
Partition the population into at least two strata, then draw a sample from each.

Want to estimate average weight of students in our class (N=12) with sample size (n=4)

Stratified Sampling

Average height and weight of dogs?

- Suppose we randomly select 6 dogs for sample
- Accidentally samples are chosen only from large breed dogs

Solution : Draw 2 sample at each stratum (small, medium, large)

Stratified Sampling

Example

The U.S. Labor Department surveys 60,000 households each month to compile its unemployment report. To select these households, the department first groups cities and counties into about 2,000 geographic areas. It then randomly selects households to survey within these geographic areas.

What are the strata? 2,000 geographic regions

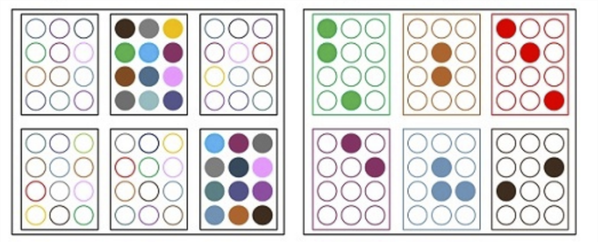
Why stratified sampling is imporant ?

- Unemployment rates are likely to differ in different geographic regions.
- For example, unemployment rates in rural Kansas may be very different from those in Silicon Valley.

Cluster vs Stratified sampling

- It is easy to confuse stratified and cluster sampling

Cluster sampling	Stratified sampling
Both use subgroups	
Clusters are homogeneous	Strata are not homogeneous
Uses all members from a sample of clusters	Uses a sample of members from all strata



Exerciese (Sampling Methods)

Identify which of these types of sampling is used: random, systematic, stratified, or cluster

- When collecting data from different sample locations in a lake, a researcher uses the "line transect method" by stretching a rope across the lake and collecting samples at every interval of 5 meters.

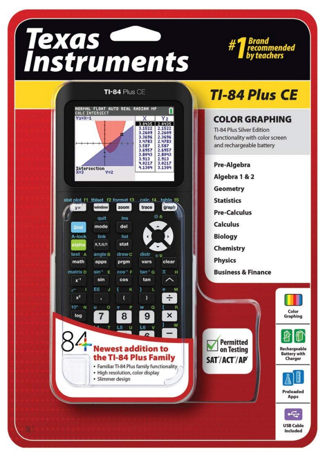
- The apple harvest from an orchard is collected in 1,200 baskets. An agricultural inspector randomly selects 25 baskets and then checks every apple in each of these baskets for worms.

- An educational researcher wants to know whether, at a particular college, men or women tend to ask more questions in class. Of the 10,000 students at the college, she interviews 50 randomly selected men and 50 randomly selected women.

Calculator

Calculator

TI 84 plus series


Texas Instruments
#1 Brand recommended by teachers
TI-84 Plus CE
TI-84 Plus CE
COLOR GRAPHING
TI-84 Plus Silver Edition functionality with color screen and rechargeable battery
Pre-Algebra
Algebra 1 & 2
Geometry
Statistics
Pre-Calculus
Calculus
Biology
Chemistry
Physics
Business & Finance
Color Graphing
Rechargeable Battery with Charger
Preloaded Apps
USB Cable Included
Permitted on Testing SAT/ACT/AP
84+
Newest addition to the TI-84 Plus Family
• Familiar TI-84 Plus family functionality
• High resolution, color display
• Slimmer design

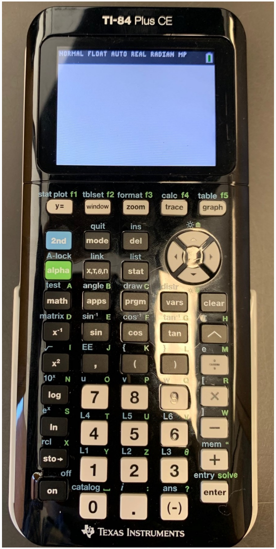
Calculator

Required functions

- Box plot, Histogram
- Summary Statistics (Mean, Variance, $\cdots$)
- Normal Distributions
- Simple Linear Regression
- One-sample Confidence Intervals
- One-sample Hypothesis Tests

Note : The TI-30X calculators will not suffice for this class.

Calculator



- Blue letter : First Press $\boxed{\text{2nd}}$

  Example : stat plot : $\boxed{\text{2nd}} + \boxed{\text{y=}}$

- Green letter : First Press $\boxed{\text{alpha}}$

  Example : X : $\boxed{\text{alpha}} + \boxed{\text{sto} \rightarrow}$

Example : Random number generator & Sequence generator

#	1	2	3	4	5	6	7	8	9	10
Obs	3.5	4.1	5.4	4.6	9.3	2.7	3.5	6.3	5.1	8.3
#	11	12	13	14	15	16	17	18	19	20
Obs	4.4	5.3	6.7	4.6	2.8	8.8	1.5	3.9	7.2	5.8

Simple Random Sample

- Need a random number generator
- $\boxed{\text{math}}$ → PROB ($\triangleright$ 4times) → 8:randIntNoRep → lower : 1, upper : 20, n : 5 , Paste + $\boxed{\text{enter}}$x2

Systematic Sample (start from 3 every 4th observation)

- Need a sequence generator
- $\boxed{\text{list}}$ (press $\boxed{\text{2nd}}$ + $\boxed{\text{stat}}$) → OPS → 5:seq + $\boxed{\text{enter}}$ → Expr : $3 + 4 * X$, Variable : X, start : 0, end : 4, Paste + $\boxed{\text{enter}}$x2

Notice

1st quiz :

- September 13 (Fri)
- You can use your calculator
- Bring one piece of hand written cheat sheet (both side allowed)
- Questions types
  - Multiple choice
  - True / False
  - Matching
  - Short answer
  - Drawing plot, Simple calculation