Sampling Theory

Course: Statistics 1

Lecturer: Dr. Courtney Pindling

Sampling Theory

- Sampling theory is a branch of statistics dedicated to providing appropriate mathematical and theoretical foundations for the use of statistics to describe a set of data, or to predict outcomes with data, or make inference about a sample or its population from a sample of the population.

Population - Sample

- A **population** is a collection of data that is alike in one or more characteristics as *defined by the researcher*
  - *All 4th Graders in the USA*
- A **sample** is a group of data selected from the population and is smaller than the size of the population
  - *All 4th Graders in NY, CT, and NJ*

Statistics - Parameters

- **Statistics**: measures that estimate population values or parameters
  - *The mean score of 20 student in a school of 1,500 students*
- **Parameter:** measures obtained from data on the entire population
  - *The mean score of all 1,500 students in a school*
  - ***Census***

Scales of Measurements

- Determines the **amount of information** contained in the data and indicates the **appropriate** statistical **summary or analysis** applicable
- Types of Measurement Scales:
  - **Nominal**
  - **Ordinal**
  - **Interval**
  - **Ratio**

Elements - Variables

- **Elements:** the entities on which data are collected
  - *Cases or row information*
- **Variables:** groups of elements with similar characteristics
  - *Column information*

Cases	GPA
Sam	2.0
Joe	3.1
Sally	4.0
Mary	2.3

Quantitative - Qualitative

- **Quantitative:**
  - **Data**: require numeric values indicating how much or how many – *Interval* or *Ratio* scales
  - **Quantitative Variable**: is a variable with quantitative data
- **Qualitative:**
  - **Data:** include labels or names used to identify an element – Nominal or Ordinal scales / Numeric or nonnumeric
  - **Qualitative Variable:** is a variable with qualitative data

Cases	GPA	Color
Sam	2.0	1
Joe	3.1	2
Sally	4.0	3
Mary	2.3	4

Dependent - Independent

- In regression terminology, the variable being predicted is called the ***dependent variable***
- The variable or variables being used to predict the value of the dependent variable are called the ***independent variable***
- We may which to examine the effects of advertising expense on sales volume
  - **Sales volume** would be our **dependent** variable and **advertising expense** the **independent** variable
    - Sales is dependent upon advertising expense

Cross-Sectional – Time Series Data

- **Cross-sectional data** are data collected at the same or approximately the same time
- **Time series data** are data collected over several time periods

Cross-sec.	GPA	SAT
John	3.2	1120

Time series	GPA
Year 1	3.1
Year 2	2.8

Types of Samples – Part 1

- **Random Samples**
  - Every member of population has **equal chance** of being selected in sample
- **Stratified Random Samples**
  - **Representative** samples that mirrors, in percent, the groups within a population, collected randomly
- **Cluster Samples**
  - Sampling and re-sampling based in information in initial sample of groups (**clusters**) within a population to find a representation of a population often too large to sample

Stratified Samples - Example

Cases	Population	Sample
Group 1	40	4
Group 2	50	5
Group 3	10	1
Total	100	10

Types of Samples – Part 2

- **Systematic Samples**
  - The first *k* element of is selected randomly and then every k^{th} element thereafter
  - *50 out of 5000: 12, 62, 112, 162, etc.*
- **Convenience Samples**
  - Samples are selected from population base on convenience
- **Judgment Samples**
  - Samples taken form population based on the judgment of person doing the study

Incidental Sample

- This is often a **sample of convenience** and often the results cannot be generalized to the population
  - If you collect data based on a sample of college freshmen, you may not be able to use the results to generalize about college seniors.

Bias Sample

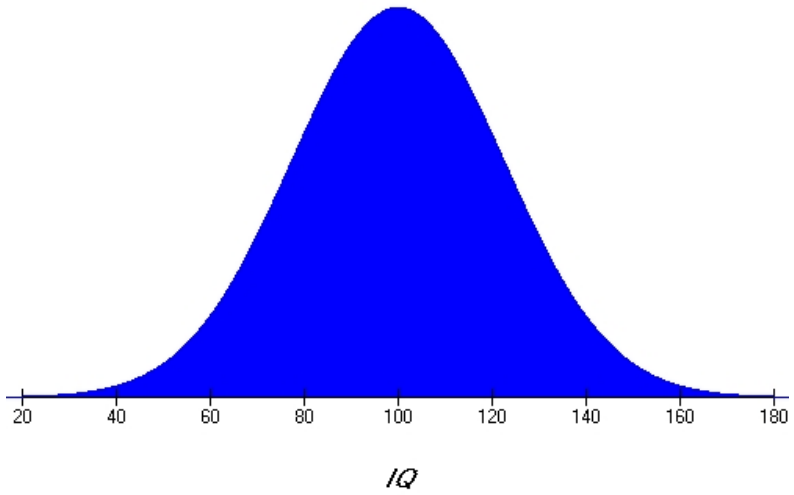
- Any sample containing a **systematic error** is said to be a *biased* sample
  - If you wanted to measure the public views on the consumption of alcoholic beverages, your sample may be biased it you only select the opinions of persons coming out of a bar

Sampling with Replacement

- Sampling that requires repeated selection from a population of which the sample is returned is called **sample with replacement**
  - Non-replacement of samples from a population and replacement sampling has different outcome probabilities and a researcher must be aware of this

Central Limit Theorem

- The means of repeated sampling of the population is normally distributed around the “true” mean or the population with a standard deviation
- The standard deviation of this sampling distribution of means is called the ***standard error of the mean***
- We use the standard deviation of our sample to estimate this standard error



Ethical Sampling Strategy

- Be aware of your institutional requirements for sampling based on human subjects or protected groups
- Obtain proper consent and permission to conduct research
- Collect and make inference about sample data ethically