How to collect DATA

Collecting Data Sensibly

Dr. Tanujit Chakraborty

Assistant Professor of Statistics

Sorbonne University

Introduction

- A primary goal of statistical studies is to collect data that can be used to make informed decisions.
- The ability to make good decisions depends on the quality of the information available.
- The data collection step is critical to obtaining reliable information.
- The conclusions that can be drawn depend on how the data are collected.

Observational Study

- A study is an **observational study** if the investigator observes characteristics of a sample selected from one or more existing populations.
- The goal of an observational study is usually to draw conclusions about the corresponding population or about differences between two or more populations.
- In a well designed observational study, the sample is selected in a way that is designed to produce a sample that is representative of the population.
- Example: an ecologist might be interested in estimating the average shell thickness of bald eagle eggs.
- In an observational study, it is impossible to draw clear cause-and-effect conclusions because we cannot rule out the possibility that the observed effect is due to some variable other than the explanatory variable being studied. Such variables are called confounding variables.

Experimental Study

- A study is an **experiment** if the investigator observes how a response variable behaves when one or more explanatory variables, also called factors, are manipulated.
- The usual goal of an experiment is to determine the effect of the manipulated explanatory variables (factors) on the response variable.
  - Example: An educator may wonder what would happen to test scores if the required lab time for a data science course were increased from 3 hours to 6 hours per week. To answer such questions, the researcher conducts an experiment to collect relevant data. The value of some response variable (test score) is recorded under different experimental conditions (3-hour lab and 6-hour lab). In an experiment, the researcher manipulates one or more explanatory variables to create the experimental conditions.

Examples:

- The article **“Television’s Value to Kids: It’s All in How They Use It” (*Seattle Times,* July 6, 2005)** described a study in which researchers analysed standardized test results and television viewing habits of 1700 children. They found that children who averaged more than 2 hours of television viewing per day when they were younger than 3 tended to score lower on measures of reading ability and short-term memory.

  a. Is the study described an observational study or an experiment?

  b. Is it reasonable to conclude that watching two or more hours of television is the cause of lower reading scores? Explain.

Sampling

- Many studies are conducted in order to generalize from a sample to the corresponding population.
- It is important that the sample be representative of the population.
- We must carefully consider the way in which the sample is selected.
- There are many reasons for selecting a sample rather than obtaining information from an entire population (a **census**).

Bias in Sampling

- **Selection bias** is introduced when the way the sample is selected systematically excludes some part of the population of interest.

  Example:

  - A researcher may wish to generalize from the results of a study to the population consisting of all residents of a particular city, but the method of selecting individuals may exclude the homeless or those without telephones.
  - If those who are excluded from the sampling process differ in some systematic way from those who are included, the sample is virtually guaranteed to be unrepresentative of the population.

- If this difference between the included and the excluded occurs on a variable that is important to the study, conclusions based on the sample data may not be valid for the population of interest.

Bias in Sampling

- Response Bias: Tendency for samples to differ from the corresponding population because the method of observation tends to produce values that differ from the true value.

  Example:

  - In a town, 25% of car accidents among 18-20 year olds were alcohol-related. Do you support lowering the legal drinking age to 18?

- Non-Response Bias: Tendency for samples to differ from the corresponding population because data are not obtained from all individuals selected for inclusion in the sample.

  Example:

  - A polling company is conducting a study in a certain city on people's attitudes to ward their occupation. The company calls random phone numbers each day between the hours of 6.00 pm and 9.00 pm. Those work during the evening hours will be unable to take part in the study.

Random Sampling

- A **simple random sample of size *n*** is a sample that is selected from a population in a way that ensures that every different possible sample of the desired size has the same chance of being selected.
- **Simple random sampling methods**
  - The following steps are involved in selecting simple random sampling:
  - A list of all the members of the population is prepared initially and then each member is marked with a specific number (for example, there are nth members then they will be numbered from 1 to N).
  - From this population, random samples are chosen using two ways: random number tables and random number generator software. A random number generator software is preferred more as the sample numbers can be generated randomly without human interference.

Random Sampling: Example

- Suppose a list containing the names of the 427 customers who purchased a new car during 2009 at a large dealership is available.
- The owner of the dealership wants to interview a sample of these customers to learn about customer satisfaction.
- She plans to select a simple random sample of 20 customers.
- Because it would be tedious to write all 427 names on slips of paper, random numbers can be used to select the sample.
- To do this, we can use three-digit numbers, starting with 001 and ending with 427, to represent the individuals on the list.

Random Sampling

- **Sampling without replacement:** Once an individual from the population is selected for inclusion in the sample, it may not be selected again in the sampling process. A sample selected without replacement includes *n* distinct individuals from the population.

- **Sampling with replacement:** After an individual from the population is selected for inclusion in the sample and the corresponding data are recorded, the individual is placed back in the population and can be selected again in the sampling process. A sample selected with replacement might include any particular individual from the population more than once.

Stratified Sampling

- When the entire population can be divided into a set of non-overlapping subgroups, a method known as **stratified sampling** often proves easier to implement and more cost-effective than simple random sampling.
- In stratified random sampling, separate simple random samples are independently selected from each subgroup.
- the subgroups are called **strata** and each individual subgroup is called a **stratum.**

Stratified Sampling: Example

- To estimate the average cost of malpractice insurance, a researcher might find it convenient to view the population of all doctors practicing in a particular metropolitan area as being made up of four subpopulations: (1) surgeons, (2) family practitioners, (3) obstetricians, and (4) a group that includes all other areas of specialization.
- Rather than taking a random simple sample from the population of all doctors, the researcher could take four separate simple random samples — one from the group of surgeons, another from the family practitioners, and so on.
- These four samples would provide information about the four subgroups as well as information about the overall population of doctors.

Examples

- During the previous calendar year, a county's small claims court processed 870 cases. Describe how a simple random sample of size $n = 20$ might be selected from the case files to obtain information regarding the average award in such cases.

- Suppose that you were asked to help design a survey of adult city residents in order to estimate the proportion that would support a sales tax increase. The plan is to use a stratified random sample, and three stratification schemes have been proposed.

Scheme 1: Stratify adult residents into four strata based on the first letter of their last name (A–G, H–N, O–T, U–Z).

Scheme 2: Stratify adult residents into three strata: college students, nonstudents who work full time, nonstudents who do not work full time.

Scheme 3: Stratify adult residents into five strata by randomly assigning residents into one of the five strata.

Which of the three stratification schemes would be best in this situation?

Cluster Sampling

- Sometimes it is easier to select groups of individuals from a population than it is to select individuals themselves.
- **Cluster sampling** involves dividing the population of interest into non-overlapping subgroups, called **clusters**.
- Clusters are then selected at random, and then *all* individuals in the selected clusters are included in the sample.

Cluster Sampling: Example

- Suppose that a large urban high school has 600 senior students, all of whom are enrolled in a first period homeroom.
- There are 24 senior homerooms, each with approximately 25 students.
- If school administrators wanted to select a sample of roughly 75 seniors to participate in an evaluation of the college and career placement advising available to students, they might find it much easier to select three of the senior homerooms at random and then include all the students in the selected homerooms in the sample.
- In this way, an evaluation survey could be administered to all students in the selected homerooms at the same time—certainly easier logistically than randomly selecting 75 students and then administering the survey to those individual seniors.

Systematic Sampling

- **Systematic sampling** is a procedure that can be used when it is possible to view the population of interest as consisting of a list or some other sequential arrangement.
- A value *k* is specified (for example, *k*=50 or *k*=200).
- Then one of the first *k* individuals is selected at random, after which every *k*th individual in the sequence is included in the sample.
- A sample selected in this way is called a **1 in *k* systematic sample**.
- For example, a sample of faculty members at a university might be selected from the faculty phone directory. One of the first *k*=20 faculty members listed could be selected at random, and then every 20th faculty member after that on the list would also be included in the sample. This would result in a 1 in 20 systematic sample.

Examples

- A sample of pages from a book is to be obtained, and the number of words on each selected page will be determined. For the purposes of this exercise, equations are not counted as words and a number is counted as a word only if it is spelled out—that is, *ten* is counted as a word, but *10* is not.

a. Describe a sampling procedure that would result in a simple random sample of pages from this book.

b. Describe a sampling procedure that would result in a stratified random sample. Explain why you chose the specific strata used in your sampling plan.

c. Describe a sampling procedure that would result in a systematic sample.

d. Describe a sampling procedure that would result in a cluster sample.

e. Using the process you gave in Part (a), select a simple random sample of at least 20 pages, and record the number of words on each of the selected pages. Construct a dotplot of the resulting sample values, and write a sentence or two commenting on what it reveals about the number of words on a page.

f. Using the process you gave in Part (b), select a stratified random sample that includes a total of at least 20 selected pages, and record the number of words on each of the selected pages. Construct a dotplot of the resulting sample values, and write a sentence or two commenting on what it reveals about the number of words on a page.

Simple Comparative Experiments

- An **experiment** is a study in which one or more explanatory variables are manipulated in order to observe the effect on a response variable.
- The **explanatory variables** are those variables that have values that are controlled by the experimenter. Explanatory variables are also called **factors.**
- The **response variable** is a variable that is not controlled by the experimenter and that is measured as part of the experiment.
- An **experimental condition** is any particular combination of values for the explanatory variables. Experimental conditions are also called **treatments.**

Simple Comparative Experiments

- An **extraneous variable** is one that is not one of the explanatory variables in the study but is thought to affect the response variable.
- Two variables are **confounded** if their effects on the response variable cannot be distinguished from one another.
- **Direct Control:** Holding extraneous variables constant so that their effects are not confounded with those of the experimental conditions (treatments).
- **Random Assignment** to ensure that the experiment does not systematically favour one experimental condition (treatment) over another.
- **Blocking:** Using extraneous variables to create groups (blocks) that are similar. All experimental conditions (treatments) are then tried in each block.
- **Replication:** Ensuring that there is an adequate number of observations for each experimental condition.
- **Experimental Units:** An experimental unit is the smallest unit to which a treatment is applied.

More on Experimental Design

- If the purpose of an experiment is to determine whether some treatment has an effect, it is important to include an experimental group that does not receive the treatment. Such a group is called a **control group**.
  - Not all experiments require the use of a control group.
- A **placebo** is something that is identical (in appearance, taste, feel, etc.) to the treatment received by the treatment group, except that it contains no active ingredients.
  - The placebo group would provide a better basis for comparison and would allow the researchers to determine whether the treatment had any real effect over and above the “placebo effect.”

Examples

- The head of the quality control department at a printing company would like to carry out an experiment to determine which of three different glues results in the greatest binding strength. Although they are not of interest in the current investigation, other factors thought to affect binding strength are the number of pages in the book and whether the book is being bound as a paperback or a hardback.

  a. What is the response variable in this experiment?

  b. What explanatory variable will determine the experimental conditions?

  c. What two extraneous variables are mentioned in the problem description? Are there other extraneous variables that should be considered?

Examples

- Red wine contains flavonol, an antioxidant thought to have beneficial health effects. But to have an effect, the antioxidant must be absorbed into the blood. The article **"Red Wine is a Poor Source of Bioavailable Flavonols in Men" (*The Journal of Nutrition* [2001]: 745–748)** describes a study to investigate three sources of dietary flavonol—red wine, yellow onions, and black tea—to determine the effect of source on absorption. The article included the following statement:

We recruited subjects via posters and local newspapers. To ensure that subjects could tolerate the alcohol in the wine, we only allowed men with a consumption of at least seven drinks per week to participate ... Throughout the study, the subjects consumed a diet that was low in flavonols.

a. What are the three treatments in this experiment?

b. What is the response variable?

c. What are three extraneous variables that the researchers chose to control in the experiment?

Single-Blind and Double-Blind Experiments

- A **single-blind** experiment is one in which the subjects do not know which treatment was received but the individuals measuring the response do know which treatment was received, or one in which the subjects do know which treatment was received but the individuals measuring the response do not know which treatment was received.

- A **double-blind** experiment is one in which neither the subjects nor the individuals who measure the response know which treatment was received.

Example:

Give an example of an experiment for each of the following:

a. Single-blind experiment with the subjects blinded

b. Single-blind experiment with the individuals measuring the response blinded

c. Double-blind experiment

d. An experiment for which it is not possible to blind the subjects