

Unit 3: Collecting Data

Statistics 571: Statistical Methods

Ramón V. León

Collecting Data

- Historical data
 - Sometimes this type of data is enough to achieve study goals
 - Relatively inexpensive
- Before data collection begins two questions must be answered
 - What are the specific goals of the study?
 - Which variables should be measured and how?
- Good planning is necessary to collect useful data. Even the most sophisticated statistical analyses cannot salvage a poorly planned study
- Measured characteristics are called **variables**; their measured values constitute the **data**

Classification of Statistical Studies

- A study whose purpose is to compare two or more alternative methods or groups distinguished by some attribute is called a **comparative (analytical)** study
 - e.g. compare different training methods
- A study whose purpose is to learn about characteristics of a group, but not necessarily to make comparisons is called a **noncomparative (descriptive)** study
 - e.g. estimate the number of illiterate Americans

Another Classification of Statistical Studies

- A study may be **observational** or **experimental**
 - An **observational study** records data without interfering with the course of events
 - In an **experimental study** the researcher actively intervenes to control the study conditions and records the responses
 - Either can be comparative or noncomparative

Establishing Causation

- Causation can be established in an experimental study by systematically varying one or more variables (called **predictor** or **explanatory variables**) and measuring their effect on **outcome** or **response variable**
- Causation cannot be established in an observational study

Observational Studies and Causation

- In an observational study only **association** between the predictor and response variables can be established
 - Effects of predictor variables cannot be separated from the effects of the uncontrolled variables; this is called **confounding**
 - The variables that cause confounding of the predictor variables are called **confounding** variables
 - If these variable are not recognized they are called **lurking variables**
- Does listening to Mozart makes you smarter?
- Does drinking wine rather than beer make you smarter?

Importance of a Control Group in a Comparative Study

- The most common comparative experimental study evaluates how a change from a baseline or normal condition affects a response
 - Change condition is called a **treatment** or an **intervention**
 - The normal condition is called a **control**
 - Group that receives the treatment is called the **treatment group**
 - Group that does not receive the treatment is called the **control group**
 - The two groups should be as similar in all other respects to provide a valid comparison (**random allocation** is usually used.)
 - A lurking variable would affect both groups equally so difference in groups can be attributed to the intervention

Experimental Studies that Lack a Control

- **Pretest Only Design:**

Intervention → Observation

- **Pretest-Posttest Design:**

Observation → Intervention → Observation

- Both designs are flawed when there is not a valid basis for comparison. Effect could be the result of a lurking variable rather than of the intervention
- **Gastric Freezing Example**
Cures were attributed to freezing stomach with a balloon. Result disproved when control involving swallowing at room-temperature balloon was used

Counterfactual View of Causality

- The effect of any treatment for a given patient is the difference between what happened to the patient as a result of giving him the treatment and what would have happened had the treatment been denied
- The effect of a treatment is not, for example, the difference between the patient's state after treatment (outcome) and his state before treatment (baseline).

Stephen Senn's *Statistical Issues in Drug Development*. Wiley.

Controlled Studies Issues

- **Placebo effect** (e.g. sugar pill): Feeling of being treated causes a temporary improvement
- **Single blind** study: Patients are kept unaware of the treatment that they receive
- **Double blind** study: Both patient and physician are kept unaware of the treatment that is given
- **Concurrent control group**
 - Best type of control group
- **Historical control group**
 - More dangerous since environmental influences change over time

Observational Studies

- Observational studies are of three basic types:
 - A **sample survey** provides a snapshot of the population based on a sample observed at one point in time
 - A **prospective study** follows a sample forward in time
 - A **retrospective study** follows a sample backwards in time

Sample Surveys

- Examples:
 - Preelection polls
 - Consumer attitude and preference surveys
- A **population** is a collection of all objects, items, or human/animal subjects (referred as **sampling units**) about which information is sought
- A **sample** is a subset of the population that is actually observed
- Emphasis is on some variable associated with the units rather than on the units themselves
 - It is convenient to define the population as the collection of the values of the variable for all the units. A sample is similarly defined.

Sample Surveys (Cont.)

- A numerical characteristic of a population defined for a specific variable is called a **parameter**
 - Parameters are usually unknown quantities
- A numerical function of the sample data, called a **statistics**, is used to estimate an unknown population parameter
- Survey examples:
 - A Gallup poll typically uses a sample of 1000 to 2000 persons drawn from the population of all adults over 18
 - In **acceptance sampling**, a sample of items from a shipment is inspected to estimate the proportion of defectives in the shipment
 - Accountants sample account books entries for audit purposes to estimate accounts receivable that are overdue.

Sample Surveys (Cont.)

- If the sample equals the population so that all members of the population are observed, then we have a **census**
- Why is sampling used so widely instead of a census?
 - Populations that are too large
 - Destructive testing
 - Less work lead to higher measuring accuracy
- There are two types of populations
 - Finite populations
 - Infinite populations
- An example of an infinite population is all patients with a certain disease (includes patients that might get the disease in the future)
- A list of all units in a finite population is called a **sampling frame**

Sample Surveys (Cont.)

- The population of interest, called the **target population** should be carefully defined in study goals
- Sometimes the target population is difficult to study so sampling is done using another easily accessible population, called the **sampled population**
 - Telephone directories
 - Study of almost 17,000 blood specimens tested for HIV in 19 universities led to overestimate of students with HIV rate. (Sampled population – those who took blood test – had a higher HIV incidence than target population of all university students)

Sampling and Nonsampling Errors

- The deviation between a sample estimate and the true parameter value is called **sampling error**.
 - The sampling error can be driven down to zero, at least in theory, by taking a census
- But **nonsampling errors** often are present, even if a census is taken. These errors cause **bias** which is a systematic deviation between the sample estimate and the population parameter
 - Bias is more serious than sampling error because it does not go away just by increasing the sample size
 - Measurement bias (e.g. poor wording)
 - Self-selection bias – people who participate in surveys are often different from those that don't
 - Response bias (Untruthful response)
- To generalize from a sample to a population one needs a **representative sample**

Measuring Public Opinion



Figure 3.1 The Princess and the Poll (STEVE BENSON reprinted by permission of United Feature Syndicate, Inc.)

Prospective Studies

- A **prospective study** begins in the present with a sample and follows the sample forward in time to record the occurrence of specific outcomes over the study period
 - **Cohort studies** follow a group of people forward in time (usually) to observe and compare characteristics of people who do and who do not develop a disease (Some times some questions about the past are asked too.)

Prospective and Retrospective Studies

- A **retrospective study** begins and ends in the present but takes a look back in time to collect data about things in the past
 - A **Case-control study** identifies a case group of people with the disease and a control group of people without the disease, then looks back in time for risk factors associated with the disease
 - To make this comparison more accurate, cases and controls are usually matched on confounding variables such as smoking, exercise, and stress
 - Among smokers compare the rate of hypertension among obese and non-obese people

Prospective and Retrospective Studies Compared

- Prospective studies are more costly than retrospective studies but usually provide better information
- Both types of studies are observational so there can be problems with confounding variables
 - In hypertension-obesity study, one could argue that a common gene causes both obesity and hypertension

Basic Sampling Designs

- Representative samples – those that do not differ in systematic and important ways from the population – are needed to help assure valid and unbiased conclusions about the population
- Sampling methods subject to bias result in nonrepresentative samples
 - Example: haphazard sampling or convenience sampling (e.g. people in the street)

Basic Sampling Designs

- In **judgment sampling** interviewers are trained to exercise their judgment in selecting a representative sample
 - In **quota sampling** an interviewer is given quotas of interviews for specified categories. Categories and quotas are selected to reflect proportion in the general population. Within the categories interviewers use their judgment to fill the quota. (Possibility of interviewer bias)
 - No guaranty of representative sample but better than convenience sampling
 - Another problem is that there are always other categories not specified in advance

Simple Random Samples

- What method should one use to avoid bias?
 - Use a chance mechanism!
- Simple random sampling (SRS):
 - *Random* does not mean *haphazard*
 - Each sample of size n from a population of size N has the same chance of being selected:

$$\frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}$$

- Each unit has an equal chance n/N of being selected, called the **sampling rate or fraction**

Simple Random Samples

- All statistical inference methods discussed in the textbook
 - Assume simple random sampling;
 - Ignore the effect of finiteness of the population, i.e., assume the sampling fraction is negligibly small.
 - Sampling fraction is less than 10%
- How to select a random sample?
 - Need a sampling frame
 - Samples are usually without replacement
 - Table of random numbers or computer random generator

Common Misconception About Simple Random Samples

- **Misconception:** Information in the sample depends mostly on the proportion of the population sampled
- **The truth:** The amount of information in a sample depends more on the absolute sample size and less on the relative sample size n/N , as long as n is not too large relative to N (e.g. $n/N < 0.10$)
 - Whether the population size is 500 or 5000, an SRS of size 50 is almost equally informative

Limitations of Simple Random Sampling

- Simple random sampling is not a practical method except for rather small populations for which sampling frames are available. However, it forms a basic building block for more complicated sampling schemes used in practice
- If a population is highly diverse, then simple random sampling is not an efficient method
 - Requires a very large sample size in order for a sample to be representative

Stratified Random Sampling

- If a population can be divided into homogeneous subpopulations, a small SRS can be drawn from each, resulting in a sample that is representative of the population
 - The subpopulations are called **strata**
 - The sampling design is called **stratified random sampling**
 - An advantage is that estimates are available for subpopulations
 - Sampling from subpopulations are more economical because of administrative convenience
- Example: IRS stratifies by income to estimate median amount of deductions of different types

Multistage Cluster Sampling

- Usually used to survey large populations
 - Where no frame is available
- Example: National Sample
 1. Draw SRS of states
 2. Draw SRS of counties from each selected state
 3. Draw SRS of towns from each selected county
 4. Draw SRS of wards from each selected town
 5. Prepare a sampling frame of households for each selected ward and draw a SRS of households as the actual sample

Multistage Cluster Sampling

- Notice that the sampling units are different at each stage
- One scheme commonly used is called **probability proportional to size** in which larger units are assigned proportionately larger selection probabilities
 - E.g. Large states have a higher probability of being selected

Systematic Sampling

- Alternative to simple random sampling when
 - A sequential list of sampling units exists
 - Sampling units become available sequentially
- A **1-in-k systematic sample** consists of selecting one unit at random from the first k units and selecting every k^{th} unit thereafter
- Systematic sampling gives a representative sample as long as there is no regular cyclic variation in the population

Sampling Reference

“Sampling: Second Edition”

by Steven K. Thompson. Wiley

Experimental Studies

- The primary purpose of an experiment is to evaluate how a set of predictor variables (called factors in experimental design jargon) affect a response variable
- Other objectives:
 - Screen factors
 - Select best combination of factor values to optimize the response
 - Fit a model that can be used to make predictions

Experimental Studies

- Factors are of two types
 - **Treatment factors** are controlled in an experiment, and their effect on the response are of primary interest
 - **In Heat Treatment of Steel Example:**
furnace temperature and quench bath temperature
 - **Nuisance factors** or **noise factors** are all the other factors that might affect the response
- The different possible values of a factor are called its **levels**
- Each **treatment** is a particular combination of the levels of the different treatment factors
 - The treatments are applied to subjects or items whose responses are then measured
 - These subjects or items are called **experimental units**

Experimental Studies

- All experimental units receiving the same treatment form a **treatment group**
- A **run** is an observation made on a experimental unit under a particular treatment condition
- A **replicate** is another *independent* run carried out under identical treatment conditions – not just a **repeat measurement**

Strategies to Reduce Experimental Error Variation

- The main purpose of an experiment is to evaluate the effects of treatment factors on a response variable
- Experimental error
 - If there is excessive experimental error, then even large treatment effects may be masked
 - One needs to minimize the experimental error to make the experiment sufficiently sensitive to detect factor effects of a practically important magnitude

Main Components of Experimental Error

- **Systematic error** is caused by the differences between experimental units. The **nuisance factors** on which the experimental units differ are said to **confound** or **bias** the treatment comparisons. Nuisance factors are sometime not even known
 - **In the Heat Treatment of Steel Example:** difference between batches is not of primary interest so “batches” is a nuisance factor
- **Random error** is caused by the inherent variability in the responses of similar experimental units given the same treatment
- **Measurement error** is caused by imprecise measuring instruments

Strategies to Reduce Systematic Error

- **Blocking:**

- Divide the sample into groups of similar experimental units (i.e., units having the same or similar values for the nuisance factor). These groups are called **blocks**, and the nuisance factor used to form the blocks is called a **blocking factor**
- All treatments are assigned on the experimental units within each block; thus the effect of the nuisance factor is the same across all treatments within each block and cancels out when these treatments are compared to each other,
 - Heat Treatment of Steel Example: batches can be used as a blocking factor

Strategies to Reduce the Systematic Error

- **Regression Analysis:**

- Some nuisance factors cannot be used as blocking factors because they are not controllable.
- Such nuisance factors are called **covariates**, *assuming they are measured* (e.g. ambient temperature).
- It is possible to model the effects of covariates by a data analysis technique called **regression analysis**
- Covariates should be measured before the treatments are applied because the treatment could also affect the covariates.

Randomization to Reduce Systematic Error

- What about additional nuisance factors, whether known or unknown?
 - Fisher's solution is to randomly assign experimental units to different treatments. This makes treatment groups probabilistically equal on all nuisance factors.
- In summary, the basic strategy for dealing with nuisance factors is:
 - **Block over those nuisance factors that can be easily controlled, randomize over the rest**

Reduction of Random and Measurement Errors

- The effect of **random error** can be reduced by replicating the experimental runs
 - Making multiple independent runs, or **replicates**, under identical treatment conditions
 - Average of the reading are used
- The effect of **measurement error** (lack of **precision**) can be reduced by making **repeat measurements**
 - Having different persons make the measurements.
 - Average of the reading are used
 - Lack of **accuracy** cannot be reduced using repeat measurements
- A benefit of of replicates and repeat measurements is that variation within each group of measurements can be used to estimate random and measurement errors, respectively

Four Strategies Useful to Improve the Precision of an Experiment

- Blocking
- Regression analysis on covariates
- Randomization

- Deal with systematic error
- Do not require larger samples

- Replication
- Repeat measurements

- Deal with random error or measurement error
- Require larger samples

CRD Versus RBD: A Simple Example

EXAMPLE 3.14 (CHEMICAL PROCESS YIELD: RANDOMIZED BLOCK DESIGN)

Refer to Example 3.11. Suppose it is known from past experience that the morning (AM) runs are different from the afternoon (PM) runs in terms of the yield obtained. Then the CRD given in Example 3.11, namely

$$\underbrace{\{B, C, B\}}_{\text{AM}}, \underbrace{\{A, A, C\}}_{\text{PM}},$$

is not balanced with respect to this nuisance factor, since both B runs are done in the morning, while both A runs are done in the afternoon. Therefore the A vs. B difference is confounded with the AM vs. PM difference. An RBD that uses the AM and PM runs as blocks provides the required balance. An example is

$$\underbrace{\{B, C, A\}}_{\text{AM}}, \underbrace{\{A, B, C\}}_{\text{PM}}.$$

A, B, C are different operating conditions

Basic Experimental Designs:

Completely Randomized Design (CRD)

- All the experimental units are assigned at random to the treatments
- Example:
 - Study the effect of furnace temperature (High and Low) and quench bath temperature (High and Low) on the surface hardness of steel. **Four treatments.**
 - A CRD consist of randomly assigning 20 samples to the 4 treatments without regard to the batches that the samples came from

(a) Completely Randomized Design

Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
A	C	D	D	B
B	A	A	C	D
C	D	B	A	C
B	C	D	B	A

(b) Randomized Block Design

Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
A	D	C	D	B
C	A	B	C	C
B	C	D	B	A
D	B	A	A	D

Basic Experimental Designs: Randomized Block Design

- A more precise comparison between treatments can be made by
 - Forming blocks of units which are similar (in terms of the blocking factor)
 - Randomly assigning the treatments to the units within each blocks
- The resulting design is called a **Randomized Block Design (RBD)**
 - Randomization is done separately within each block. (In a CRD randomization is done over all experimental units.)
- **Matched pairs design** is a special case of a RBD.
A **cross-over design** is a variation of the matched pair design
- A CRD exploits the benefits of randomization, but fails to use information on any blocking factors. The result is that the experimental error is not minimized.

Iterative Nature of Experiments

Experimentation cycles through the steps below:

1. Recognize and state the problem
2. Review literature and determine what has been previously done
- 3. Plan an experiment to answer the questions**
4. Conduct the experiment and collect data
5. Analyze the data to find the answers
6. Redefine the problem in light of the results of the experiment

Do You Need to Know More?

Reference: "Designing Clinical Research: An Epidemiologic Approach, Second Edition" by Stephen B. Hulley, Steven R. Cummings, Warren S. Browner, Deborah Grady, Norman Hearst, and Thomas B. Newman. Lippincott Williams & Wilkins.

A text on planning and implementing clinical research, for beginning investigators. Section I examines basic ingredients of research, such as selection of study subjects, hypotheses, and estimating sample sizes. Section II presents design options, covering cohort, cross-sectional, case-control and diagnostic test studies, clinical trials, and confounding. Section III looks at additional skills in ethics, data management, and community and international studies. Emphasis is on common sense as the main ingredient of good science..