Introduction to Statistics Note

2024 Spring Semester

21 CST H3Art

Chapter 3: Producing Data

3.1 Sources of Data

Anecdotal data represent individual cases that often come to our attention because they are striking in some way. "The plural of **anecdote** is not evidence."

Sample surveys are a special type of **designed experiment** that usually aim to discover the opinions of people on certain topics. In a sample survey, a **sample** of individuals is selected from a larger **population** of individuals.

The **distinction** between **observational study** and **experiment** is one of the most important in statistics:

- An **observational study** observes individuals and measures variables of interest but **does not attempt to influence the responses**.
- An **experiment** deliberately **imposes some treatment** on individuals to measure their responses.

Observational studies of the effect of one variable on another often **fail** because of **confounding** between the explanatory variable and one or more **lurking variables**:

- **Confounding** occurs when two variables are associated in such a way that **their effects** on a response variable **cannot be distinguished** from each other.
- A **lurking variable** is a variable that is **not among the explanatory or response variables** in a study but that may influence the response variable.

3.2 Design of Experiments

An **experiment** is a study in which we actually do something (a **treatment**) to people, animals, or objects (the **experimental units**) to observe the **response**.

An **experimental unit** is the smallest **entity** to which a **treatment is applied**.

- When the units are **human beings**, they are often called **subjects**.

The **explanatory variables** in an experiment are often called **factors**.

A specific condition applied to the individuals in an experiment is called a **treatment**.

Many laboratory experiments operate as follows:

$$\text{Experimental Units} \rightarrow \text{Treatment} \rightarrow \text{Measure Response}$$

Outside the laboratory, **badly designed experiments** often **yield worthless results** because of **confounding**.

In a **comparative experiment**, **comparison alone** isn't enough. If the treatments are given to groups that differ greatly, **bias** will result. The **solution** to the problem of bias is **random assignment**.

In a **completely randomized** design, the treatments are assigned to all the experimental units completely by chance. Some experiments may include a **control group** that **receives an inactive treatment** or an existing **baseline treatment**.

How to randomly choose n individuals from a group of N:

- We first label each of the N individuals with a number (typically from 1 to N, or 0 to $N-1$).
- Imagine writing the whole numbers from 1 to N on separate pieces of paper. Now put all the numbers in a hat.
- Mix up the numbers and randomly select one.
- Mix up the remaining $N-1$ numbers and randomly select one of them.
- Continue in this way until we have our sample of n numbers.

Principles of Experimental Design:

- **Control** for lurking variables that might affect the response, most simply by comparing two or more treatments.
- **Randomize**: Use chance to assign experimental units to treatments.
- **Replication**: Use enough experimental units in each group to reduce chance variation in the results.

An observed effect so large that it would rarely occur by chance is called **statistically significant**.

A **statistically significant association** in data from a **well-designed experiment** does imply **causation**.

In a **double-blind experiment**, neither the **subjects** nor those **who interact with them** and measure the response variable **know which treatment a subject received**.

A **matched pairs** design is a **randomized blocked experiment** in which each block consists of a **matching pair of similar experimental units**.

A **block** is **a group of experimental units** that are known before the experiment to be **similar** in some way that is expected to affect the response to the treatments.

3.3 Sampling Design

The **population** in a statistical study is the **entire group of individuals** about which we want information.

A **sample** is the part of the population from which we actually collect information.

The design of a **sample** is **biased** if it **systematically favors** certain outcomes.

A **voluntary response sample（自愿响应样本）** consists of people who choose themselves by responding to a general appeal. Voluntary response samples often show bias because people with strong opinions (often in the same direction) may be more likely to respond.

A **simple random sample (SRS)** of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.

A **probability sample** is a sample chosen by chance. We must know what samples are possible and what chance, or probability, each possible sample has.

To select a **stratified random sample**, first classify the population into groups of similar individuals, called **strata**. Then choose a separate SRS in each stratum and combine these SRSs to form the full sample.

Good sampling technique includes the art of reducing all sources of **error**:

- **Undercoverage（覆盖不足）** occurs when some groups in the population are left out of the process of choosing the sample.
- **Nonresponse（无响应）** occurs when an individual chosen for the sample can't be contacted or refuses to participate.
- A systematic pattern of incorrect responses in a sample survey leads to **response bias（响应偏差）**.
- The **wording of questions（问题措辞）** is the most important influence on the answers given to a sample survey.

3.4 Toward Statistical Inference

A **parameter** is a number that describes some characteristic of the population.

A **statistic** is a number that describes some characteristic of a sample.

We write μ (the Greek letter mu) for the population mean and σ for the population standard deviation.

We write $\bar{x}$ (x-bar) for the sample mean and s for the sample standard deviation.

The **population distribution** of a variable is the distribution of values of the variable among all individuals in the population.

The **sampling distribution** of a **statistic** is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Bias concerns the **center of the sampling distribution**. A **statistic** used to estimate a **parameter** is **unbiased**.

- To reduce bias, use **random sampling**.

The **variability** of a **statistic** is described by **the spread of its sampling distribution**.

- To reduce variability of a statistic from an SRS, use a **larger sample**.

The process of drawing **conclusions about a population** on the basis of **sample data** is called **inference**.