Module A: The Foundations of Statistics

Slide Deck A3:

Conducting a Statistical Study

The section in which we explore how to properly conduct statistical studies. Here, we will explore the steps, the importance of categorizing variables, observational studies, and experimental studies.

Today's Objectives

By the end of this slidedeck, you should

- classify variables as categorical and numeric, as nominal, ordinal, interval, and ratio
- know the process of conducting a statistical study
- identify types of observational studies and experimental studies
- properly use the terminology of studies
- know the three keys to a successful experiment
- explain how to protect the experiment and the subject

Example

In the following experiment, please identify the following based on this description:

- the target population,
- the sampled population,
- the sample, and
- the variables.

Neurologists want to study the effect of Vitamin C on nerve pain. The goal of the study is to see if taking an intravenous dose of Vitamin C will reduce the amount of nerve pain reported by patients.

To achieve this goal, the researchers advertise at their local university for volunteers with nerve pain. Each member of the study is put into one of two groups: control (no Vitamin C) and treatment (5000mg Vitamin C daily). At the start of the study, everyone is given a pain-scale analysis (wherein the patient identifies the level of nerve pain from a scale of 0 to 10). At 90 days, everyone repeats the analysis.

Types of Variables

Variables are at the heart of statistics. Different types of variables contain different levels of information. While there are several ways to categorize variables based on their characteristics, here are three different classification systems:

- qualitative, quantitative
- discrete, continuous
- nominal, ordinal, interval, ratio

All three schemes are useful in helpging us better understand what can (and cannot) be done with certain variables.

Qualitative vs. Quantitative

Qualitative data, also known as *categorical* data, consist of labels or descriptions of traits.

- hair color
- phone type
- car make
- government type

Quantitative data, also known as *numeric* data, consist of counts or measurements.

- temperature
- height
- number of pages read in the past week
- amount of ice cream eaten in a month

Continuous vs. Discrete

Discrete variables are quantitative variables that can take on only particular values and are usually counts.

- number of dents on a car
- number of students in the room
- GPA

Continuous variables are quantitative variables that can take on any value in a given interval and are usually measurements.

- temperature
- height
- weight

Note that these refer to the *variable*, not to how its values are recorded.

The Four Levels of Measurement

The level of measurement of a variable describes the *amount of information that variable contains.* The four levels are

- Nominal — description
- Ordinal — nominal plus ordering
- Interval — numeric plus differences between levels are equal
- Ratio — interval plus a true zero value

To remember these, I use the first letter of each as a mnemonic device: NOIR means "black" in French.

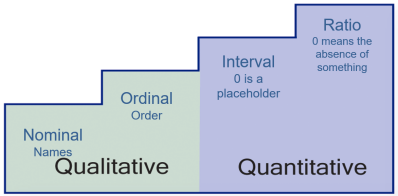
The Four Levels of Measurement

The level of measurement of a variable describes the *amount of information that variable contains.* The four levels are

- Data at the **nominal level** of measurement are qualitative data consisting of labels or names.
- Data at the **ordinal level** of measurement are qualitative data that can be arranged in a meaningful order, but calculations such as addition or division do not make sense.
- Data at the **interval level** of measurement are quantitative data that can be arranged in a meaningful order, and differences between data entries are meaningful.
- Data at the **ratio level** of measurement are quantitative data that can be ordered, differences between data entries are meaningful, and the zero point indicates the absence of something.

The Four Levels of Measurement

An illustration of the four levels:



Examples

Identify each of the following as **numeric** or **categorical** (quantitative or qualitative):

- temperature outside on a Saturday
- whether the student wears a mask
- amount of money carried by the person
- number of credit cards carried by a person
- favorite grocery store in Galesburg
- number of allergies in a person

Examples

Identify each of the following as **continuous** or **not continuous**:

- temperature outside on a Saturday
- whether the student wears a mask
- amount of money carried by the person
- number of credit cards carried by a person
- favorite grocery store in Galesburg
- number of allergies in a person

Examples

Identify the level of each of the following variables (**nominal**, **ordinal**, **interval**, or **ratio**):

- temperature outside on a Saturday
- whether the student wears a mask
- amount of money carried by the person
- number of credit cards carried by a person
- favorite grocery store in Galesburg
- number of allergies in a person

Conducting a Statistical Study

These are the general steps to conduct a *legitimate* scientific study:

- Determine the design of the study.
  - State the question to be studied.
  - Determine the population and variables.
  - Determine the sampling method.
- Collect the data.
- Organize the data.
- Analyze the data to answer the question.

Two Study Types

There are two basic types of scientific experiments. They are

- An **observational study** observes data that already exist.
- An **experiment** generates data to help identify cause-and-effect relationships.

Note: These are the "proper" definitions as used by scientists. A statistician will refer to any 'theoretical' data collection as an experiment. This difference in terminology comes from the fact that statisticians will conduct experiments to better understand the randomness in the data.

In this course, the laboratory activities are experiments. They are designed to give you better understanding of randomness and its effects on your estimates.

The Goal of the Sample

There is only one purpose/goal of the sample: The be **representative** of the population.

- A representative sample has the same relevant characteristics as the population and does not favor one group from the population over another.

Note: A sample can be representative for one characteristic of the population (parameter) but not for another. In fact, it can be representative in every single parameter, but not for the relationships between them.

An important question:
— How does one know if a sample is representative of the population?

Types of Observational Studies

Broadly speaking, a statistical study is either experimental or observational. There are a few main types of observational studies:

- In a **cross-sectional** study, data are collected at a single point in time.
- In a **time-series** study (ts), data are gathered by following a single person over a period of time.
- In a **longitudinal** study, data are gathered by following a particular group over a period of time.
- In a **cross-sectional-time-series** study (xt), data are gathered by following a larger group over a period of time.

Yes, the last two categories differ only in terms of "particular group" vs. "larger group." There is no hard-and-fast separator between these two. As a result, the statistical analysis is quite similar between the two.

Terminology of Experimental Studies

Here is some terminology related to scientific experiments. Some of these we have already used, others not.

- The **subjects** (participants) are people or things being studied in an experiment.
- A **treatment** is some condition that is applied to a group of subjects in an experiment.
- The **response variable** is the variable in an experiment that *responds* to the treatment. This type of variable is also termed the **dependent** variable.
- The **explanatory variable** is the variable (or variables) in an experiment that *causes* the change in the response variable. This type of variable is also termed the **independent** variable.

Terminology of Experimental Studies

Here is some terminology related to scientific experiments. Some of these we have already used, others not.

- A **control group** is a group of subjects to which no treatment is applied in an experiment.
  - A **placebo** is a substance that appears identical to the actual treatment but contains no intrinsic beneficial elements.
  - The **placebo effect** is a response to the power of suggestion, rather than the treatment itself, by participants of an experiment.
- A **treatment group** is a group of subjects to which researchers apply a treatment in an experiment.
- The **confounding variables** are *unmeasured* factors that cause an effect on the subjects of an experiment.

The Fundamental Principle of Experiments

To properly perform a scientific experiment, one should. . .

- completely randomize the control and treatment groups
  - or, for reasons that improve the statistics (case-control study), explicitly tie each member of the control group to the a member of the treatment group
- control for outside effects on the response variable
- analyze a sufficiently large sample
- replicate the experiment a significant number of times to see meaningful patterns

For more information on case-control studies: Please read the two pages that is "Epidemiology in Practice: Case-Control Studies" (Lewallen and Courtright, 1998) located at `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1706071/`.

The Three Keys for a Successful Experiment

There are reasons behind these fundamental principles. They delve into the goal of ensuring the sample is representative of the population.

- The purpose of the above principles is to ensure that *any* difference in the response variable is due *only* to the treatment (dependent variable) and *not* to differences in the groups.
- One needs to have a sufficiently large sample size to detect the "signal" and not be confused by the "noise."
- Repeat, using proper sampling methods to help ensure the average of the results are close to the parameter of interest.

Protecting the Experiment

Now, let us look at how we can ensure experiments are properly performed (beyond the research report):

- **Blinding**
  - In a **single-blind experiment**, subjects do not know if they are in the control group or the treatment group, but the people interacting with the subjects in the experiment know in which group each subject has been placed.
  - In a **double-blind experiment**, neither the subjects nor the people interacting with the subjects know to which group each subject belongs.

Protecting the Subject

An **Institutional Review Board** is a group of people who review the design of a study to make sure that it is appropriate and that no unnecessary harm will come to the subjects.

> *The Institutional Review Board (IRB) at Knox College is responsible for overseeing all research at Knox College that involves the use of human subjects. The ultimate goal of the IRB is to safeguard the well-being of individuals who participate in research conducted by Knox College and/or its faculty, staff, and students and to ensure that research occurring at Knox College meets accepted ethical standards. The Knox IRB also reviews projects conducted by outside researchers who wish to collect data from members of the Knox College community.*

This information — and more — about the Knox College IRB can be found at

`https://www.knox.edu/offices/academic-affairs/institutional-review-board`

— **Informed consent** involves completely disclosing to participants the goals and procedures involved in a study and obtaining their agreement to participate.

Today's Objectives

Now that we have concluded this lecture, you should be able to

- classify variables as categorical and numeric, as nominal, ordinal, interval, and ratio
- know the process of conducting a statistical study
- identify types of observational studies and experimental studies
- properly use the terminology of studies
- know the three keys to a successful experiment
- explain how to protect the experiment and the subject

Supplemental Readings

The following are some readings that may be of interest to you in terms of today's topics:

- Hawkes Learning: Section 1.2
- Intro to Modern Statistics: Sections 2.2–4
- R for Starters: Section 4.1