CHAPTER OVERVIEW

Unit 1: Exploratory Data Analysis

CO-1: Describe the roles biostatistics serves in the discipline of public health.

CO-6: Apply basic concepts of probability, random variation, and commonly used statistical probability distributions.

Video

Exploratory Data Analysis Introduction (2 videos, 7:04 total)

The Big Picture

Learning Objectives

LO 1.3: Identify and differentiate between the components of the Big Picture of Statistics

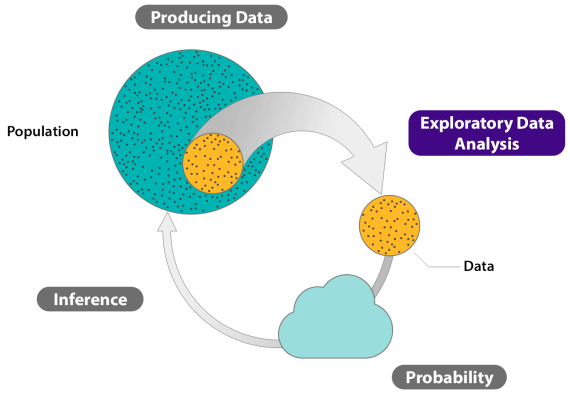
Recall "The Big Picture," the four-step process that encompasses statistics (as it is presented in this course):

1. Producing Data — Choosing a sample from the population of interest and collecting data.

2. Exploratory Data Analysis (EDA) {Descriptive Statistics} — Summarizing the data we've collected.

3. and 4. Probability and Inference — Drawing conclusions about the entire population based on the data collected from the sample.

Even though in practice it is the second step in the process, we are going to look at Exploratory Data Analysis (EDA) first. (If you have forgotten why, review the course structure information at the end of the page on The Big Picture and in the video covering The Big Picture.)



Exploratory Data Analysis

Learning Objectives

LO 1.5: Explain the uses and important features of exploratory data analysis.

As you can tell from the examples of datasets we have seen, raw data are not very informative. **Exploratory Data Analysis (EDA)** is how we make sense of the data by converting them from their raw form to a more informative one.

Note

In particular, **EDA consists of:**

- organizing and summarizing the raw data,
- discovering important features and patterns in the data and any striking deviations from those patterns, and then
- interpreting our findings in the context of the problem

And can be useful for:

- describing the distribution of a single variable (center, spread, shape, outliers)
- checking data (for errors or other problems)
- checking assumptions to more complex statistical analyses
- investigating relationships between variables

Exploratory data analysis (EDA) methods are often called **Descriptive Statistics** due to the fact that they simply describe, or provide estimates based on, the data at hand.

In Unit 4 we will cover methods of **Inferential Statistics** which use the results of a sample to make inferences about the population under study.

Comparisons can be visualized and values of interest estimated using EDA but descriptive statistics alone will provide no information about the certainty of our conclusions.

Important Features of Exploratory Data Analysis

There are two important features to the structure of the EDA unit in this course:

Note

- The material in this unit covers two broad topics:

  Examining Distributions — exploring data **one variable at a time**.

  Examining Relationships — exploring data **two variables at a time**.

Note

- In Exploratory Data Analysis, our exploration of data will always consist of the following two elements:

  visual displays, supplemented by

  numerical measures.

Try to remember these structural themes, as they will help you orient yourself along the path of this unit.

Examining Distributions

Learning Objectives

LO 6.1: Explain the meaning of the term distribution in statistics.

We will begin the EDA part of the course by exploring (or looking at) **one variable at a time**.

As we have seen, the data for each variable consist of a long list of values (whether numerical or not), and are not very informative in that form.

In order to convert these raw data into useful information, we need to summarize and then examine the **distribution** of the variable.

Note

By **distribution** of a variable, we mean:

- what values the variable takes, and
- how often the variable takes those values.

We will first learn how to summarize and examine the distribution of a single categorical variable, and then do the same for a single quantitative variable.

Case C-C
Case C-Q
Case Q-Q
Causation
One Categorical Variable
One Quantitative Variable: Introduction
Role-Type Classification
Summary (Unit 1)