What is Data?

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

Before we jump into Exploratory Data Analysis, and really appreciate its importance in the process of statistical analysis, let's take a step back for a minute and ask:

What do we really mean by **data**?

Learning Objectives

LO 1.4: Define basic terms regarding data and recognize common variations in terminology.

Video

What is Data? (2:49)

Data are pieces of information about **individuals** organized into **variables**.

- By an **individual**, we mean a particular person or object.
- By a **variable**, we mean a particular characteristic of the individual.

A **dataset** is a set of data identified with a particular experiment, scenario, or circumstance.

Datasets are typically displayed in tables, in which rows represent individuals and columns represent variables.

EXAMPLE: Medical Records

The following dataset shows medical records for a sample of patients.

	Variables					
Individuals	Gender (M/F)	Age	Weight (lbs.)	Height (in.)	Smoking (1=No, 2=Yes)	Race
Patient #1	M	59	175	69	1	White
Patient #2	F	67	140	62	2	Black
Patient #3	F	73	155	59	1	Asian
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
Patient #75	M	48	90	72	1	White

In this example,

- the **individuals** are patients,
- and the **variables** are Gender, Age, Weight, Height, Smoking, and Race.

Each **row**, then, gives us all of the information about a particular **individual** (in this case, patient), and each **column** gives us information about a particular **characteristic** of all of the patients.

Individuals, Observations, or Cases

Note

The rows in a dataset (representing **individuals**) might also be called **observations**, **cases**, or a description that is specific to the individuals and the scenario.

For example, if we were interested in studying flu vaccinations in school children across the U.S., we could collect data where each observation was a

- student

- school
- school district
- city
- county
- state

Each of these would result in a different way to investigate questions about flu vaccinations in school children.

Independent Observations

> **Note**
>
> In our course, we will present methods which can be used when the **observations** being analyzed are **independent of each other**. If the observations (rows in our dataset) are not independent, a more complex analysis is needed.Clear violations of independent observations occur when
>
> - we have more than one row for a given individual such as if we gather the same measurements at many different times for individuals in our study
> - individuals are paired or matched in some way.

As we begin this course, you should start with an awareness of the types of data we will be working with and learn to recognize situations which are more complex than those covered in this course.

Variables

> **Note**
>
> The columns in a dataset (representing **variables**) are often grouped and labeled by their role in our analysis.

For example, in many studies involving people, we often collect **demographic** variables such as gender, age, race, ethnicity, socioeconomic status, marital status, and many more.

> **Note**
>
> The **role** a variable plays in our analysis must also be considered.
>
> - In studies where we wish to predict one variable using one or more of the remaining variables, the variable we wish to predict is commonly called the **response** variable, the **outcome** variable, or the **dependent variable**.
> - Any variable we are using to predict or explain differences in the outcome is commonly called an **explanatory variable**, an **independent variable**, a **predictor** variable, or a **covariate**.

Various Uses of the Term INDEPENDENT in Statistics

Note: The word "**independent**" is used in statistics in numerous ways. Be careful to understand in what way the words "independent" or "independence" (as well as dependent or dependence) are used when you see them used in the materials.

- Here we have discussed **independent observations** (also called cases, individuals, or subjects).
- We have also used the term **independent variable** as another term for our explanatory variables.
- Later we will learn the formal probability definitions of **independent events** and **dependent events**.
- And when comparing groups we will define **independent samples** and **dependent samples**.
