Introduction to Biostatistics and R

Lecture 1

Outline

- Introduction to Biostatistics
- Introduction to R

Introduction to Biostatistics

Datafication

The field of statistics has grown in recent years primarily due to the *datafication* of the world.

Important

98% of all stored information is digital. Collected data is increasing even at this very moment.

Why statistics?

Statistics

Statistics is a field of study concerned with:

- The collection, organization, summarization, and analysis of data
- The drawing of inferences about a body of data when only a part of the data is observed.

Data → Numbers → Information → Investigation → Communication

Statistical Thinking

How is statistical thinking different from mathematical thinking?

Statistical Thinking

Statistical thinking involves understanding and analyzing data while accounting for uncertainty!

Activity

Flip a coin 10 times. If you don't have a coin, search "coin flip" on Google.

> **Note**
>
> How many times did you get heads? Do you think the coin you flipped was fair?

Extensions of Statistics

Biostatistics | Data Science

Biostatistics involves applying statistical concepts to data from the biological sciences, health sciences, and medicine.

Sources of Data

Available data usually come from the following sources:

- Records
- Surveys
- Experiments
- Data Banks
- Prior Literature

Categories of Statistics

Descriptive Statistics | Inferential Statistics

Descriptive statistics are used to describe properties of complex sets of numbers. Summary statistics are a good example of descriptive statistics.

Random Variables

Random variables have values obtained arise as a result of chance factors, so that they cannot be exactly predicted in advance. Values of random variables resulting from measurement procedures are referred to as *observations/measurements*.

Note

Random variables could be classified as qualitative or quantitative.

Random Variable Types

Quantitative Variables | Qualitative Variables

Quantitative variables are variables that can be measured or characterized with a numerical value.

ⓘ Discrete Random Variables

A **discrete variable** is characterized by gaps or interruptions in the values that it can assume.

Example: Customer counts at Cafe Rio, Number of missing teeth, Likert Scale scores

ⓘ Continuous Random Variables

A **continuous variable** is characterized by gaps or interruptions in the values that it can assume.

Example: Speed, Weight, Time

Data Types

Nominal Data | Ordinal Data | Interval Data | Ratio Data

As the name implies, nominal data consist of “naming” observations or classifying them into various mutually exclusive and collectively exhaustive categories.

Examples: Assigned sex at birth (male,female); HHS Regions (HHS Regions 1-10)

> ⓘ **Important**
>
> Nominal data are typically qualitative in nature and does not account for any ordering in variable levels.

Exercise

Question Answers

Identify the type of data/variable for the following:

- BMI
- Satisfaction Scale (Unsatisfied, Moderately Satisfied, Satisfied, Very Satisfied)
- Eye Color
- Credit Rating

Population vs. Sample

ⓘ Population

A population is the largest collection of entities for which we have an interest at a particular time. Measurements of some variable from these entities would generate a population of values for that variable.

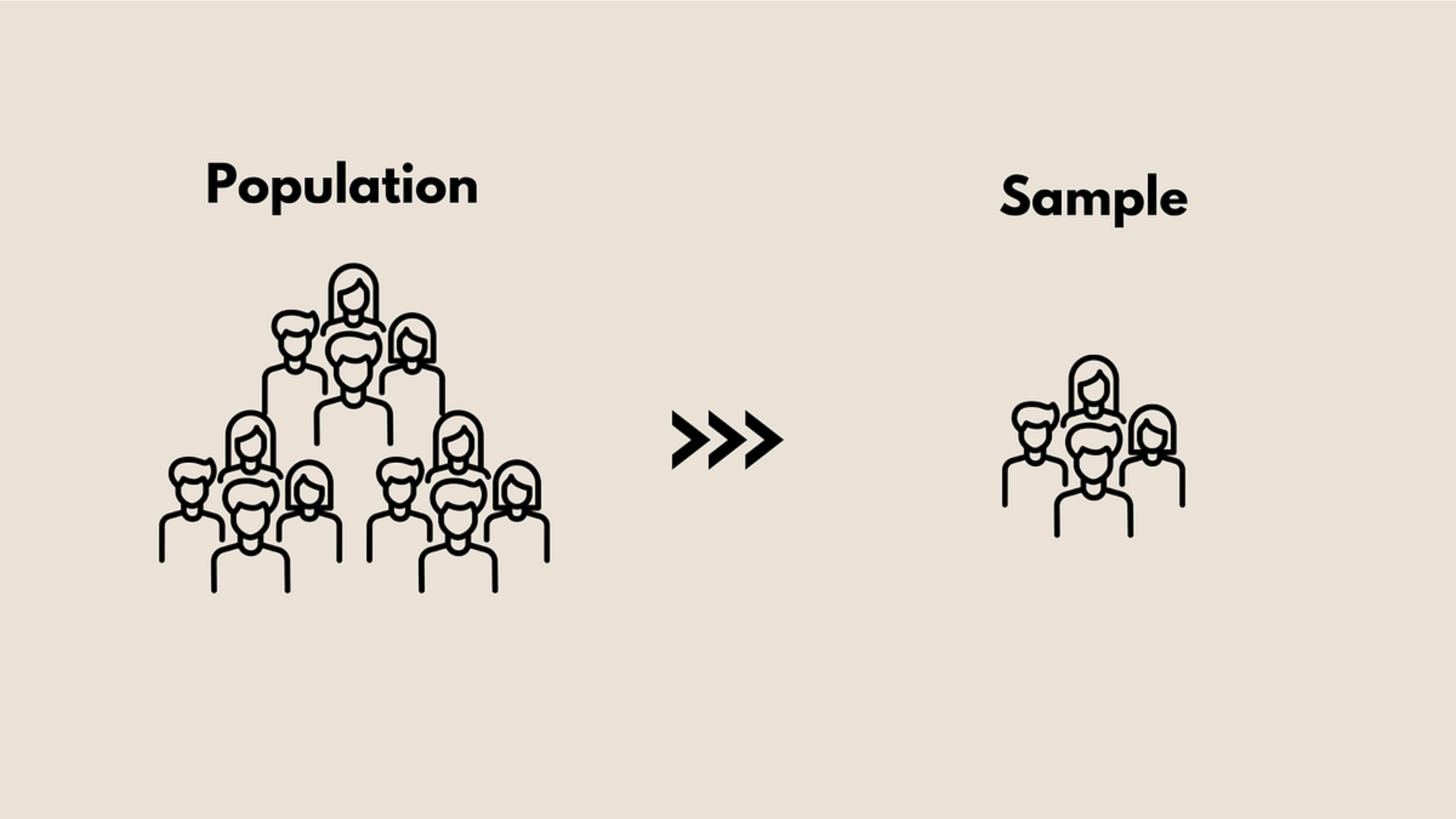
An exact value calculated from a population is referred to as a **parameter**.

ⓘ Sample

A sample is a part of the population.

An exact value calculated from a sample is referred to as a **statistic**.

Population vs. Sample



Taken from: https://medium.com/@ritusantra/population-v-s-sample-f17c40967257

How to Sample

Sampling can be grouped into two broad categories: probability-based/random sampling and convenience sampling.

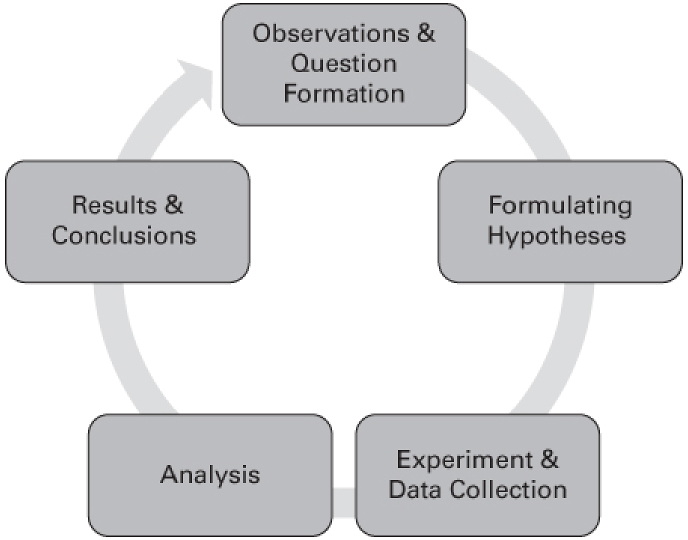
Random Samples | Convenience Samples

A sample is a random sample when the probability with which every respondent was sampled is known. These probabilities are not necessarily equal. Types of random sampling include:

- Simple random sampling
- Stratified random sampling
- Cluster sampling

Scientific Method

The scientific method is a process by which scientific information is collected, analyzed, and reported in order to produce unbiased and replicable results in an effort to provide an accurate representation of observable phenomena.


Observations & Question Formation
Formulating Hypotheses
Experiment & Data Collection
Analysis
Results & Conclusions

Introduction to R

Installation

You can install `R` and `RStudio` on your personal computers and laptops by following the instructions on this page: https://posit.co/download/rstudio-desktop/

`RStudio` is currently installed on the classroom computers.

Basic Programming Terminology

- Source Code: A text listing of commands to be compiled or assembled into an executable computer program.
- Running Code: The act of telling `R` to perform an act by giving it commands through source code.
- Console Pane: Where `R` commands are entered

Important

There are different types of programming data types such as integers, doubles/numerics, logicals, and characters.

- Integers (int) have values like -1,0,2
- Numerics (dbl, num) are numbers including integers and decimals,
- Logicals (logi) are either `TRUE` or `FALSE`
- Characters (chr)are text variables such as "Hello, World", "Female", "Yes"

Basic Programming Terminology

Vectors Variables Factors

Vectors are a series of values. These can be created using the `c()` function, known as the combine/concatenate function.

```r
c(1,2,3)
```

```
[1] 1 2 3
```

```r
c("A","B","C")
```

```
[1] "A" "B" "C"
```

Basic Programming Technology

Data Frames   Conditionals   Functions

Data frames are rectangular spreadsheets. Data are typically imported as data frames.

Note

Rows correspond to observations and the columns correspond to variables.

Example:

```r
head(mtcars)
```

```
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Errors, Warnings, and Messages

Errors | Warnings | Messages

When you input a legitimate error, R will warn you using a sentence starting with “Error in” and includes a sentence explaining what went wrong.

```r
add(1,2,3)
```

```
Error in add(1, 2, 3): could not find function "add"
```

```r
c("A","B")+1
```

```
Error in c("A", "B") + 1: non-numeric argument to binary operator
```

> ⓘ **Note**
>
> If the text starts with “Error”, figure out what’s causing it. Think of errors as a red traffic light: stop and assess for anything wrong in the code (missing parenthesis, adding characters to numbers, non-existent functions, etc.)

> ⓘ **Important**

R packages

R packages extend the functionality of R by providing additional functions, data, and documentation. These packages are written by R users around the world and can be downloaded for free!

ⓘ **Note**

Think of R as a new phone. R packages are apps that you can download to use your phone in many different ways.

Installing and Loading R Packages

Like apps on a phone, R packages also need to be installed. These packages can be installed by running the following code snippet `install.packages("PackageName")`. For example, to install the package `tidyverse` used for data manipulation and cleaning, you can run the following code:

```r
install.packages("tidyverse")
```

To load this package in R, you can use the following syntax:

```r
library(tidyverse)
```

Important

You must have an active internet connection to install R packages to your device.

Exercise

Exercise | Answer

Install and load the following packages: `readxl`, `nycflights23` and `knitr`.

Exploring Data Sets

The `nycflights23` package includes some data sets saved as data frames. These data sets are related to all domestic flights departing from one of New York City’s three main airports in 2023: Newark Liberty International (EWR), John F. Kennedy International (JFK), and LaGuardia Airport (LGA).

One of the data sets in this package is the `flights` data set.

```r
flights
```

```
# A tibble: 435,352 × 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
 1  2023     1     1        1           2038       203      328              3
 2  2023     1     1       18           2300        78      228            135
 3  2023     1     1       31           2344        47      500            426
 4  2023     1     1       33           2140       173      238           2352
 5  2023     1     1       36           2048       228      223           2252
 6  2023     1     1      503            500         3      808            815
 7  2023     1     1      520            510        10      948            949
 8  2023     1     1      524            530        -6      645            710
 9  2023     1     1      537            520        17      926            818
10  2023     1     1      547            545         2      845            852
```

Exploring the `flights` data set.

You can use the following functions to explore a data set.

`View()` `glimpse()` `kable()` `$`

`View()` brings up RStudio's built in data viewer. That is, if you want to view data like an Excel sheet.

```r
View(flights)
```

Exercise

Exercise | Answer

Can you provide me with two qualitative variables and two quantitative variables in the dataset `planes` in the `nycflights23` package?

Exercise

Exercise | Answer

Explore the data set `iris`.

- How many observations does `iris` have?
- How many variables does `iris` have?
- Use `glimpse()` to determine the type of data of each column of `iris`.
- Use the `$` operator to extract the species variable in `iris`

Summary

Summary

- Introduced biostatistics and its importance
- Introduced `R`
- Explored data sets

What’s next?

We will be using R to work with data and perform statistical analysis. We will also explore how to use R to explore and describe data from external sources (.csv files).