Rensselaer

why not change the world?®

Introductory Statistics/ Refresher & Intro to Labs

Ahmed Eleish

Data Analytics ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960

Group 1, Week1 - Module 2, September 6th, 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute

Admin information

- Class: ITWS-4600/ 6600/MATP-4450/CSCI-4960/MGMT 4962/6962/BCBP 4960
- Hours:
  Section 1: 10:00am ET - 11:50am ET on Tues/Fri – Troy 2018
  Section 2: 02:00pm ET - 03:50pm ET on Tues/Fri – Troy 2018
- Instructor: Ahmed Eleish
- Instructor contact: eleisa2@rpi.edu
- Instructor office hours: Wed from 01:30 PM - 3:30 PM ET/ Thursday from 02:30 PM - 4:00 PM ET or by appointment/email
- Instructor office location: Amos Eaton 134
- TA: Alyssa Bigness - bignea2@rpi.edu
- TA office hours: Mon 12-2pm ET / Wed 2-3pm
- TA office: Lally 205
- Web site: https://tw.rpi.edu/classes/data-analytics-fall-2024
- LMS (http://lms.rpi.edu/)

Your Github Repository

- Your Github Repository for this class
- Please create a Github repo for the Data Analytics class labs
- Do not share the Assignment codes in your Data Analytics course Github repo, you share only your lab work and your individual project work/code.
- We will collect your Github repo URLs next week.
- Example: https://github.com/tYourGithub/DataAnalytics2024_YOUR_NAME

Definitions/ topics

- Statistic
- Statistics
- Population and Samples
- Sampling
- Distributions and parameters
- Central Tendencies
- Frequency
- Probability
- Significance tests
- Hypothesis (null and alternate)
- P-value
- Density and cumulative distributions

Definitions/ topics

- Statistic
- Statistics
- Population and Samples
- Sampling
- Distributions and parameters
- Central Tendencies
- Frequency

Today's class

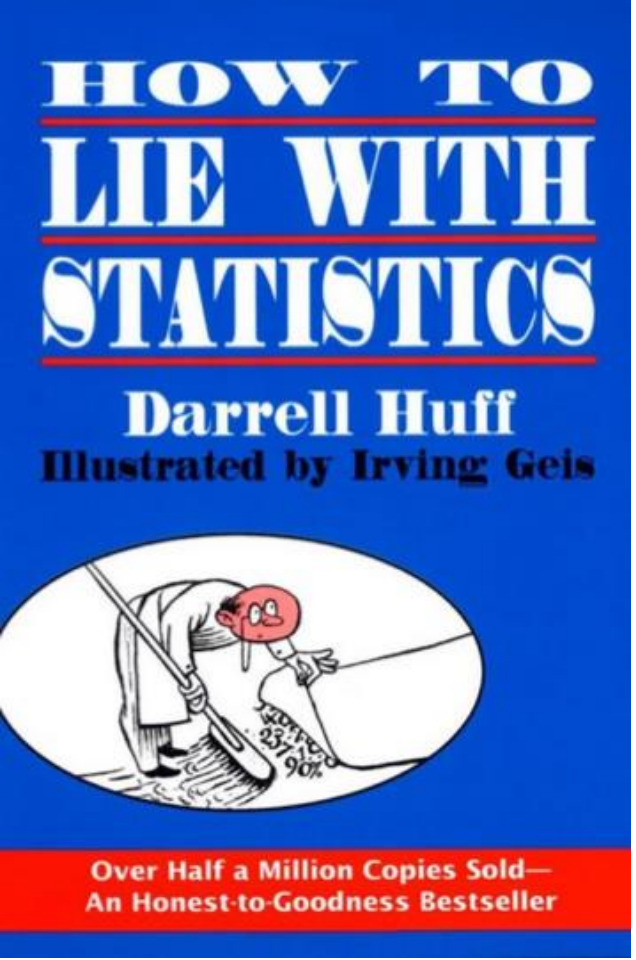
- Probability
- Significance tests
- Hypothesis (null and alternate)
- P-value
- Density and cumulative distributions

Tuesday's class

Statistic and Statistics

- Statistic (not to be confused with Statistics)
  – Characteristic or measure obtained from a sample.

- Statistics
  – Collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions.



HOW TO LIE WITH STATISTICS
(Huff, D. 1954)

There are three kinds of lies: lies, damned lies, and statistics.
—Disraeli

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.
—H. G. Wells

It ain't so much the things we don't know that get us in trouble. It's the things we know that ain't so.
—Artemus Ward

Round numbers are always false. —Samuel Johnson

I have a great subject [statistics] to write upon, but feel keenly my literary incapacity to make it easily intelligible without sacrificing accuracy and thoroughness.
—Sir Francis Galton

What is "statistics"?

• The term "statistics" has two common meanings, which we want to clearly separate: **descriptive** and **inferential** statistics.

• But to understand the difference between descriptive and inferential statistics, we must first be clear on the difference between populations and samples.

Populations and samples

- A population is a set of well-defined objects
  - We must be able to say, for every object, if it is in the population or not
  - We must be able, in principle, to find every individual of the population

  A geographic example of a population is all pixels in a multi-spectral satellite image
- A sample is a subset of a population
  - We must be able to say, for every object in the population, if it is in the sample or not
  - Sampling is the process of selecting a sample from a population

  Continuing the example, a sample from this population could be a set of pixels from known ground truth points

Courtesy Marshall Ma (and prior sources)

Populations and samples

Definitions

- Population : The complete set of actual or potential elements about which inferences are made

- Sample : A subset of the population selected using some sampling methods.

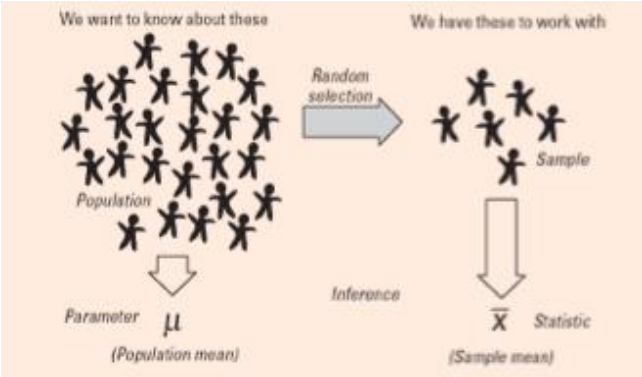
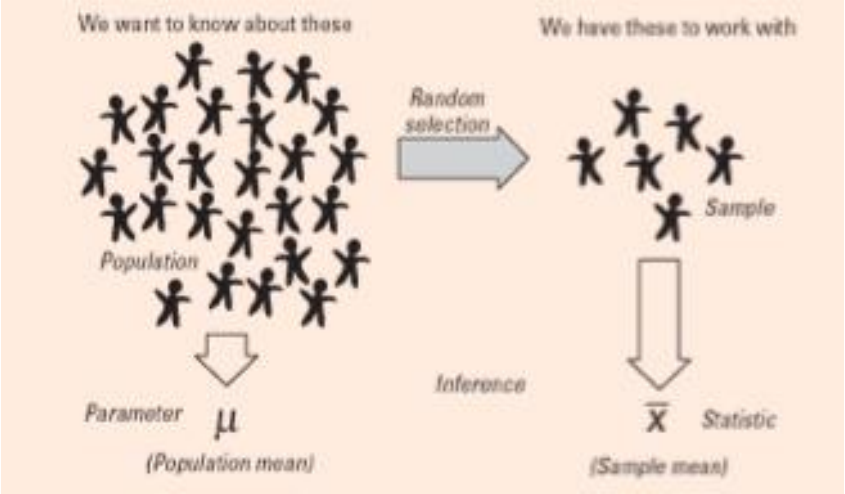


Image Credit/Reference: Quick Study Statistics

Populations and samples

- Population : The complete set of actual or potential elements about which inferences are made
- Sample : A subset of the population selected using some sampling methods.



Reference: Quick Study Statistics
Image Courtesy: https://www.cliffsnotes.com/study-guides/statistics/sampling/populations-samples-parameters-and-statistics

Populations and samples

- A population = “all” of the data, if you can get it (BIG Data)
  – This is about the different methods you use

- A sample = “some” of the data, and you may not know how representative it is
  – This is what limits analysis but certainly the development of models

Courtesy Marshall Ma (and prior sources)

Sampling Types (basic)

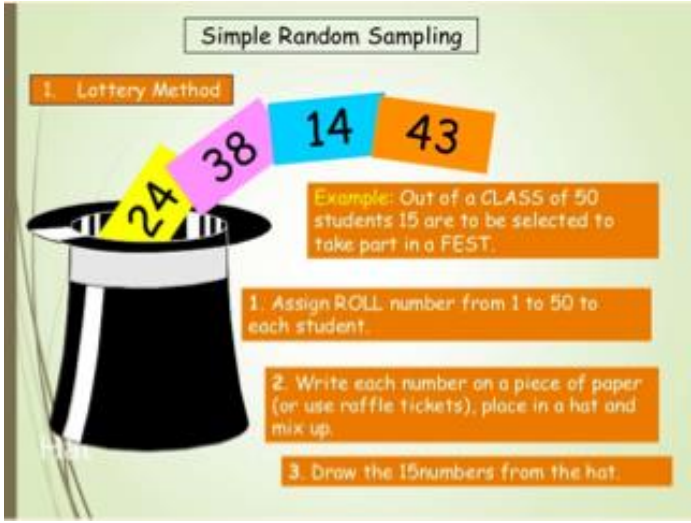
• Random Sampling
– Sampling in which the data is collected using chance methods or random numbers.
• Systematic Sampling
– Sampling in which data is obtained by selecting every *k*th object.
• Convenience Sampling
– Sampling in which data that is readily available is used.
• Stratified Sampling
– Sampling in which the population is divided into groups (called strata) according to some characteristic. Each of these strata is then sampled using one of the other sampling techniques.
• Cluster Sampling
– Sampling in which the population is divided into groups (usually geographically). Some of these groups are randomly selected, and then all of the elements in those groups are selected.

Sampling Methods

- Simple Random Sampling
- Cluster Sampling
- Stratified Sampling

Sampling Methods

Simple Random Sampling: A sample is selected so that each possible sample of the same size has an equal probability of being selected; used for most elementary inference



Courtesy: Quick Study Academic – Statistics www.quickstudy.com
Reference: Quick Study Statistics
Image Courtesy:
https://www.slideshare.net/mohammedzuhairy1/sampling-techniques-64917617

Sampling Methods

- Stratified Sampling: The population is divided into strata, and a fixed number of elements of each stratum are selected for the sample.

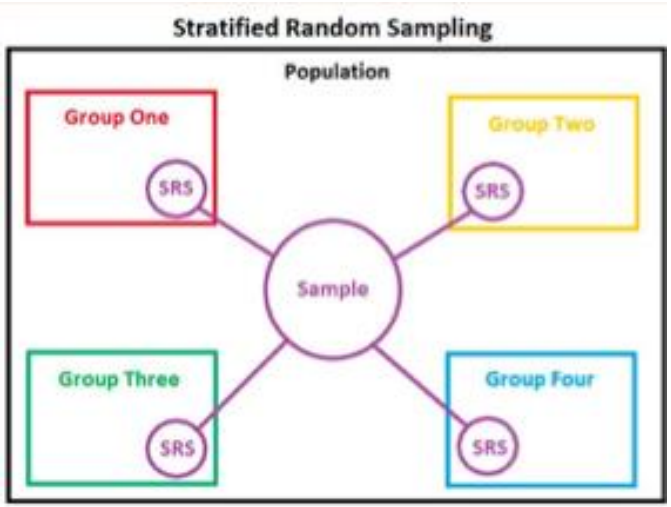
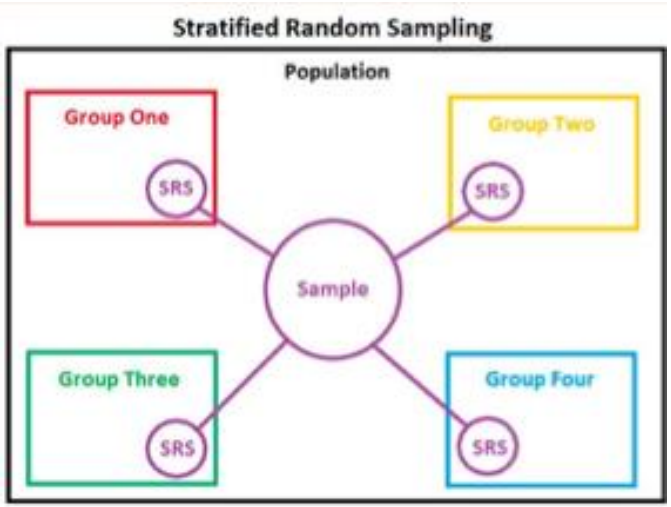


Image Credit:
https://en.wikipedia.org/wiki/Stratified_sampling

Sampling Methods

• Cluster Sampling: The population is divided into groups called clusters; some clusters are randomly selected, and every member in them observed.



https://en.wikipedia.org/wiki/Cluster_sampling

Probability Distributions

• You should be familiar with classical probability distributions.

• Classical distributions have two properties:
1. They describe shapes of frequency distributions that rise often in real world applications.
2. They can often be described mathematically with few parameters.

Statistical Tests

• Let's say, your observed (collected data) not necessary arise from a particular theoretical distribution just because of its shape is familiar.

• There are "Statistical Tests" that can be used to rigorously test whether your obtained data reflects samples drawn from a particular distribution.

• We are going to talk about some of these statistical tests during the next couple of week... ☺

Variable

- Variable: An attribute of a population or sample that can be measure.

Example: weight, height, eye-color, pulse rate are some of the many variables that can be measured for people.

Data

- Types of Data:
  - Qualitative (or Categorical)
  - Quantitative (data like numeric values)

Data

- Qualitative:

Qualitative (or Categorical) data are descriptive, but not numeric.

Example: your eye-color, you gender, color of a vehicle, your birthplace

Data

- Quantitative:

Quantitative data take on numeric values

- Discrete data take counting numbers (0,1,2,3..) this is used to represent things that can be counted. Example: Number of times an employee is late to work. Number of cars parking lot in the parking garage.

- Continuous data can take a range of numeric values, not just the counting numbers (fractions, decimals are included..) Example: height of a person, weight of an apple, number of times an employee late to work.

Courtesy: Quick Study Academic – Statistics www.quickstudy.com

Levels of Measurement

Qualitative (or Categorical) data can be measured at the:

• Nominal Level: Values are just names, without any order (example: eye-color)

• Ordinal Level: Values have some natural order, example: high school class (freshman, sophomore, ..) military rank

Quantitative (data like numeric values) can be measured at the:

• Interval Level: Numeric data with no natural zero point; intervals (differences) are meaningful but ratios are not, example: Temperature in Fahrenheit degrees 80F is 20F hotter than 60F, but it is not 150% as hot.

• Ratio Level: Numeric data which there is true zero, both intervals and ratios are meaningful; Example: weight, length, duration

Types of Data

Type of data	Level of measurement	Examples
Categorical	**Nominal** (no inherent order in categories)	Eye colour, ethnicity, diagnosis
	Ordinal (categories have inherent order)	Job grade, age groups
	Binary (2 categories – special case of above)	Gender
Quantitative (Interval/Ratio) (NB units of measurement used)	Discrete (usually whole numbers)	Size of household **(ratio)**
	Continuous (can, in theory, take any value in a range, although necessarily recorded to a predetermined degree of precision)	Temperature °C/°F (no absolute zero) **(interval)** Height, age **(ratio)**

Parameter

- Parameter: A numeric measure that describe a population: parameters are usually not computed; but are inferred from sample statistics.
- Parameters are normally denoted using Greek symbols, whereas the corresponding statistics are denoted using Latin letters. Below you find a list of the most important parameters and the corresponding statistics.

	Parameter	Statistic
mean	μ	m
standard deviation	σ	s
correlation coefficient	ρ	r

1.Courtesy: Quick Study Academic – Statistics www.quickstudy.com
Resources & Image Credit: http://www.statistics4u.info/fundstat_eng/cc_parameter.html

Something to remember...

- Difference between “N” and “n”
- Population → N
- Sample → n

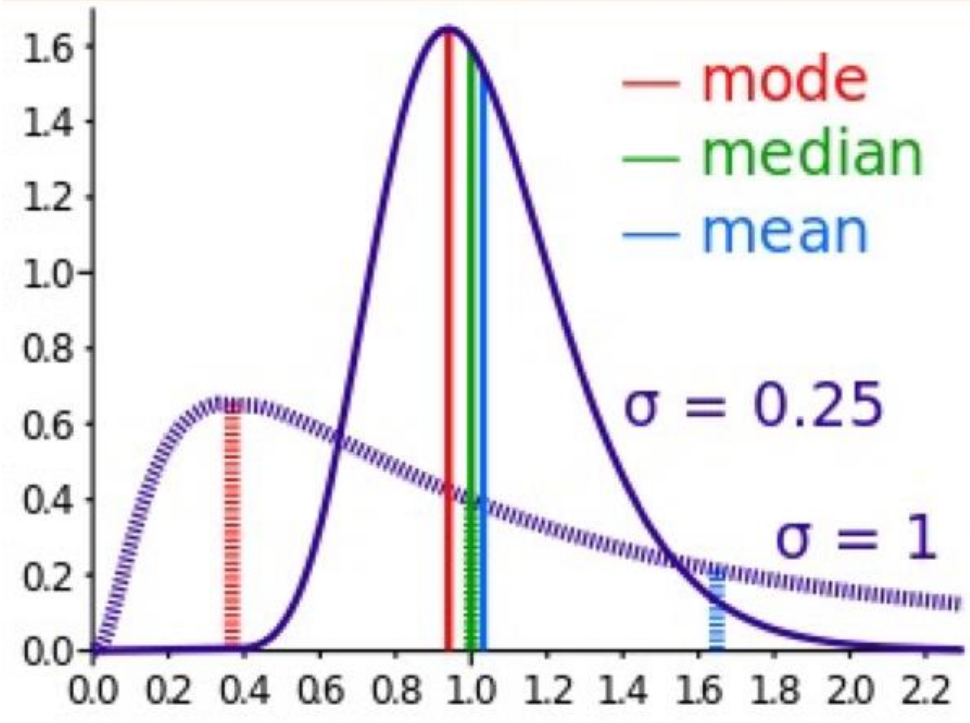
Special values in data

- Fill value
- Error value
- Missing value
- Not-a-number (NAN)
- Infinity
- Default
- Null
- Rational numbers

Outlier

- An extreme, or atypical, data value(s) in a sample.
- They should be considered carefully, before exclusion from analysis.
- For example, data values maybe recorded erroneously, and hence they may be corrected.
- However, in other cases they may just be surprisingly different, but not necessarily 'wrong'.

Central tendency – median, mean, mode



Measure of Central Tendency

• Mean: The most commonly used measure of central tendency, commonly referred to as “Average”, sensitive to extreme values (sensitive to outliers)

– Population Mean
– Sample Mean

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

https://www.youtube.com/watch?v=k5EbijWu-Ss

Image Resource: https://www.onlinemathlearning.com/population-mean.html Courtesy: Quick Study Academic – Statistics www.quickstudy.com

Measure of Central Tendency

- **Median**: Value that divides the set in 2 so the same number of observations lie on each side of it.
- ***The median is less sensitive to extreme values***
- For an even number, it is the average of middle two values.

1, 3, 3, **6**, 7, 8, 9

Median = 6

1, 2, 3, **4**, **5**, 6, 8, 9

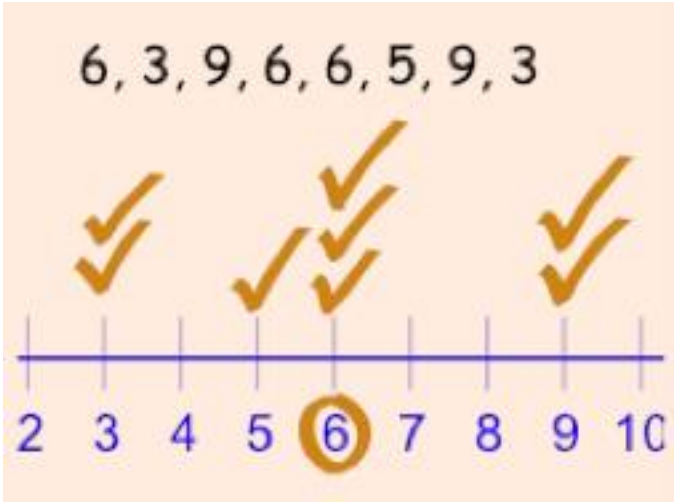
Median = (4 + 5) ÷ 2

= 4.5

Courtesy: Quick Study Academic – Statistics www.quickstudy.com Image Resource: https://en.wikipedia.org/wiki/Median

Measure of Central Tendency

- **Mode**: Observation that occurs with the greatest frequency.



Courtesy: Quick Study Academic – Statistics
www.quickstudy.com Image Resource:
https://www.mathsisfun.com/definitions/mode.html

Mean; Median; Outlier

Found this on Twitter... Credit: Anna .J. Egalite

Frequencies...

- The **Absolute frequency** n_i is the number of observations belonging to a category a_i or falling into a particular class c_i. The sum of all frequencies of all categories/classes is equal to N, the total number of observations:

$$\Sigma n_i = N$$

- **Relative frequencies** f_i are obtained by normalizing the individual frequencies to a total sum of 1.0 (or 100%, respectively). This way the frequencies become independent of the sample size and will be comparable to each other.
- Frequencies are usually delineated in a **frequency table** or displayed as a **histogram**.

Example: 28 persons have been asked for their eye colors, resulting in the following frequencies:

eye color	abs. frequency	rel. frequency
brown	14	0.500 (50%)
gray	2	0.071 (7.1%)
blue	9	0.321 (32.1%)
green	3	0.107 (10.7%)

The dataset with 28 observations and one variable exihibts four categories which differ in their frequencies.

Resource & Image Credit:
http://www.statistics4u.info/fundstat_eng/ee_frequency.html

Ranges: z, Percentiles, Quartiles

• The standard score is obtained by subtracting the mean and dividing the difference by the standard deviation. The symbol is z, which is why it's also called a z-score.

• **Percentiles (quantiles) (100 regions)**

– The k^{th} percentile is the number which has k% of the values below it. The data must be ranked.

• **Quartiles (4 regions)**

– The quartiles divide the data into 4 equal regions.

– Note: The 2^{nd} quartile is the same as the median. The 1^{st} quartile is the 25^{th} percentile, the 3^{rd} quartile is the 75^{th} percentile.

Data Prepared for Analysis = Munging

• Missing values, null values, etc.
– Most data applications provide built-ins for these higher-order functions – in R “NA” is used and functions such as is.na(var), etc. provide powerful filtering options (we’ll cover these on next Friday)

• Of course, different variables often are missing “different” values

• In R – higher-order functions such as: Reduce, Filter, Map, Find, Position and Negate will become your enemies and then your friends: http://www.johnmyleswhite.com/notebook/2010/09/2 42 3/higher-order-functions-in-r/

Five-number summary

The five-number summary is a set
of descriptive statistics that provide information
about a dataset. It consists of the five most important sample percentiles:

1. the sample minimum (smallest observation)
2. the lower quartile or first quartile
3. the median (the middle value)
4. the upper quartile or third quartile
5. the sample maximum (largest observation)

Read: https://en.wikipedia.org/wiki/Five-number_summary

Getting started – look at the data

- Visually

– What is the improvement in the understanding of the data as compared to the situation without visualization?

– Which visualization techniques are suitable for one's data?

  - Scatter plot diagrams
  - Box plots (min, 1st quartile, median, 3rd quartile, max) • Stem and leaf plots
  - Frequency plots
  - Group Frequency Distributions plot
  - Cumulative Frequency plots
  - Distribution plots

Why visualization?

- Reducing amount of data
- Patterns
- Features
- Events
- Trends
- Irregularities
- Leading to presentation of data, i.e. information products

Installing R

- http://lib.stat.cmu.edu/R/CRAN/ - install this first
- http://cran.r-project.org/doc/manuals/
- http://cran.r-project.org/doc/manuals/R-lang.html
- R Studio (https://www.rstudio.com/products/rstudio/) (desktop version)

Getting Started : Rstudio – MASS library

```
install.packages("MASS") # installing the MASS package library(MASS) # load the library MASS
attach(Boston) # attaching the dataset
?Boston # help function with "?"
head(Boston) # show the head of the dataset
dim(Boston) # dimensions of the dataset
names(Boston) # column names
str(Boston) # str function shows the structure of the dataset
nrow(Boston) # function shows the number of rows
ncol(Boston) # function shows the number of columns
summary(Boston) # summary() function shows the summary statistics
summary(Boston$crim) # summary of the "crime" column in the Boston dataset
```

Getting Started : Rstudio – ISLR library – Auto dataset

```
install.packages("ISLR") # installing the ISLR package library(ISLR)
data(Auto)
head(Auto)
names(Auto)
summary(Auto)
summary(Auto$mpg)
fivenum(Auto$mpg)
boxplot(Auto$mpg)
hist(Auto$mpg)
summary(Auto$horsepower)
summary(Auto$weight)
fivenum(Auto$weight)
boxplot(Auto$weight)
mean(Auto$weight)
median((Auto$weight))
```

Time to “Play☺” with the data

- ***In class work*** – Explore the EPI dataset – https://tw.rpi.edu/sites/default/files/2024-09/epi2024results06022024.txt

Next class: Tuesday 09/10

- Lecture – Data resources, continue stats review, hypotheses, exploration and distributions

Friday 09/13 – Lab 01

Thanks!
(Have a great weekend)

*** Experiment with R!