STAT 243Z

Introduction to Statistics

N. C. Phillips, Winter 2026

Statistics is the science of collecting and interpreting data.

Data is *everywhere* in our lives:

- medical studies
- sports
- elections
- education
- traffic planning

Even if you don't end up in a career where you analyze data yourself, it is very useful if you understand the basics about how other people are doing it.

It almost never happens that we have access to ALL the data on a certain topic. For example, we will never have data about how EVERYONE in the U.S. reacts to a certain drug—or about how EVERYONE feels about a certain political question.

Instead, the data that we have access to is a **sample**: a portion of the complete data.

The main topic of this course is **statistical inference**: making deductions based on samples.

- How does one differentiate between real evidence for something and random noise?
- How does one decide whether a sample is reliable or not?
- What are the dangers and pitfalls with all this?

Sample scenario:

It has been widely determined that the average height for American men in their 20's is 5 feet 9 inches. The town of Appaloosa, Nebraska (population 10,531) claims that their men are taller than average. You go to investigate.

You randomly select 50 men in their 20's, measure their height, and find that the average is 5 feet 11 inches.

What can you conclude from this?

A. This evidence supports the town's claim.

B. This evidence does not support the town's claim: there is not enough information to draw any conclusions.

Sample scenario:

It has been widely determined that the average height for American men in their 20's is 5 feet 9 inches. The town of Appaloosa, Nebraska (population 10,531) claims that their men are taller than average. You go to investigate.

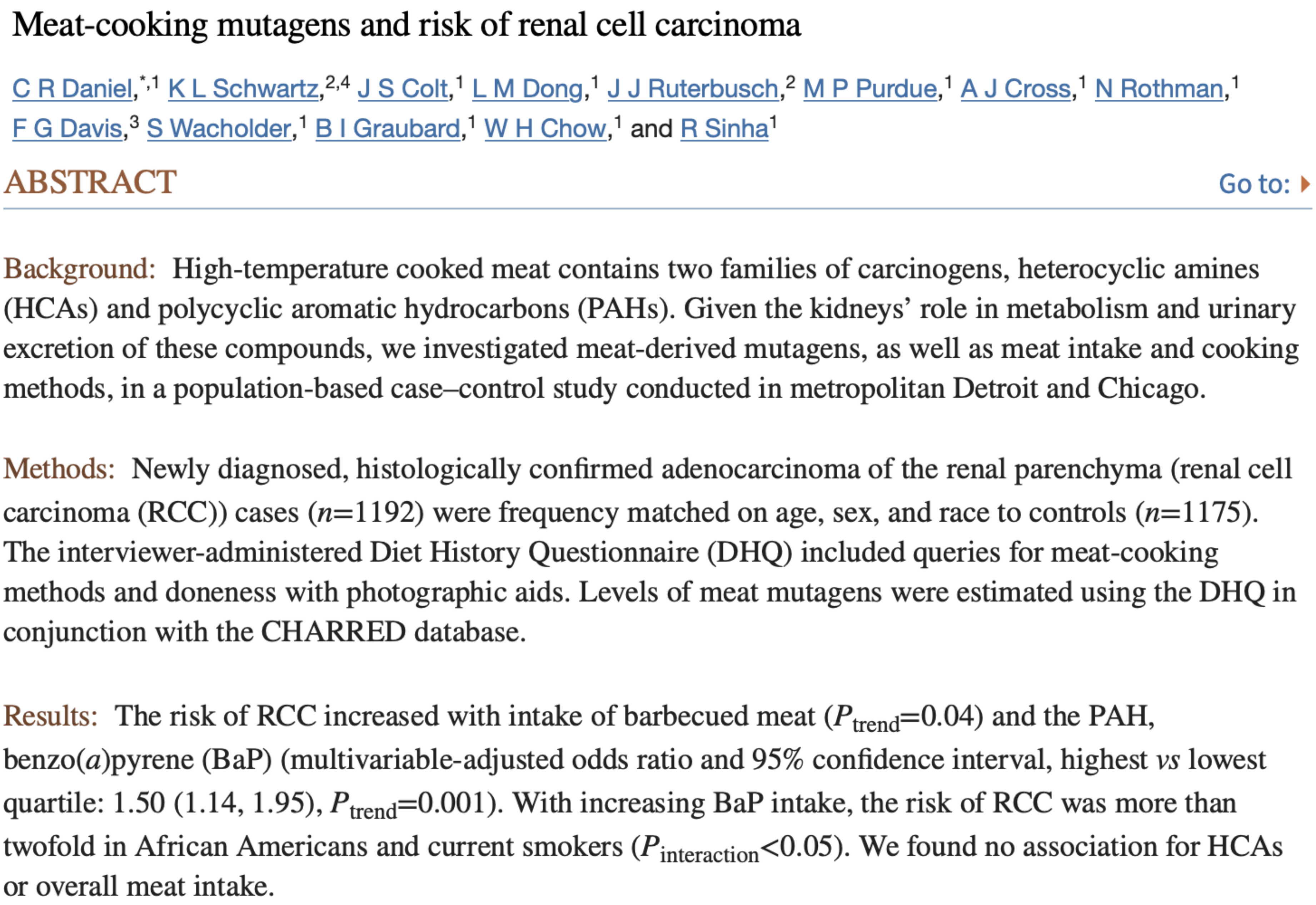
You randomly select 50 men in their 20's, measure their height, and find that the average is 5 feet 11 inches.

What can you conclude from this?

A. This evidence supports the town's claim.

B. This evidence does not support the town's claim: there is not enough information to draw any conclusions.

The correct answer is B!!!
In this course we will learn why, and also what extra information one needs in order to draw valid statistical conclusions.



Meat-cooking mutagens and risk of renal cell carcinoma

C R Daniel,[*,1] K L Schwartz,[2,4] J S Colt,[1] L M Dong,[1] J J Ruterbusch,[2] M P Purdue,[1] A J Cross,[1] N Rothman,[1] F G Davis,[3] S Wacholder,[1] B I Graubard,[1] W H Chow,[1] and R Sinha[1]

ABSTRACT

Background: High-temperature cooked meat contains two families of carcinogens, heterocyclic amines (HCAs) and polycyclic aromatic hydrocarbons (PAHs). Given the kidneys' role in metabolism and urinary excretion of these compounds, we investigated meat-derived mutagens, as well as meat intake and cooking methods, in a population-based case–control study conducted in metropolitan Detroit and Chicago.

Methods: Newly diagnosed, histologically confirmed adenocarcinoma of the renal parenchyma (renal cell carcinoma (RCC)) cases (n=1192) were frequency matched on age, sex, and race to controls (n=1175). The interviewer-administered Diet History Questionnaire (DHQ) included queries for meat-cooking methods and doneness with photographic aids. Levels of meat mutagens were estimated using the DHQ in conjunction with the CHARRED database.

Results: The risk of RCC increased with intake of barbecued meat ($P_{\text{trend}}=0.04$) and the PAH, benzo(*a*)pyrene (BaP) (multivariable-adjusted odds ratio and 95% confidence interval, highest *vs* lowest quartile: 1.50 (1.14, 1.95), $P_{\text{trend}}=0.001$). With increasing BaP intake, the risk of RCC was more than twofold in African Americans and current smokers ($P_{\text{interaction}}<0.05$). We found no association for HCAs or overall meat intake.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3185955

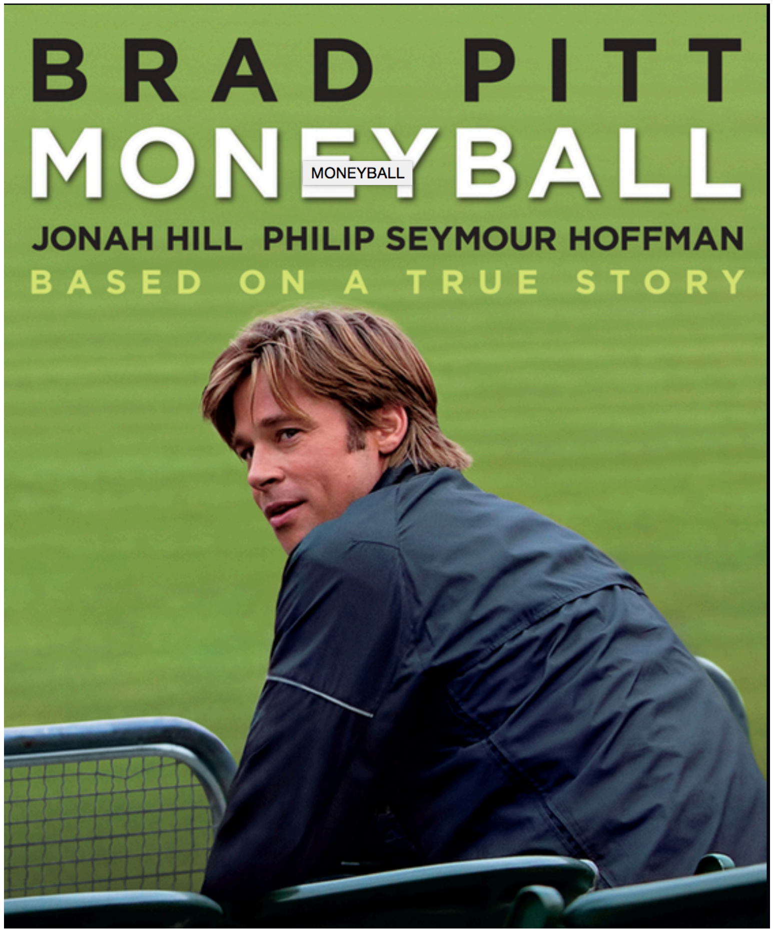
Main goals for the course:

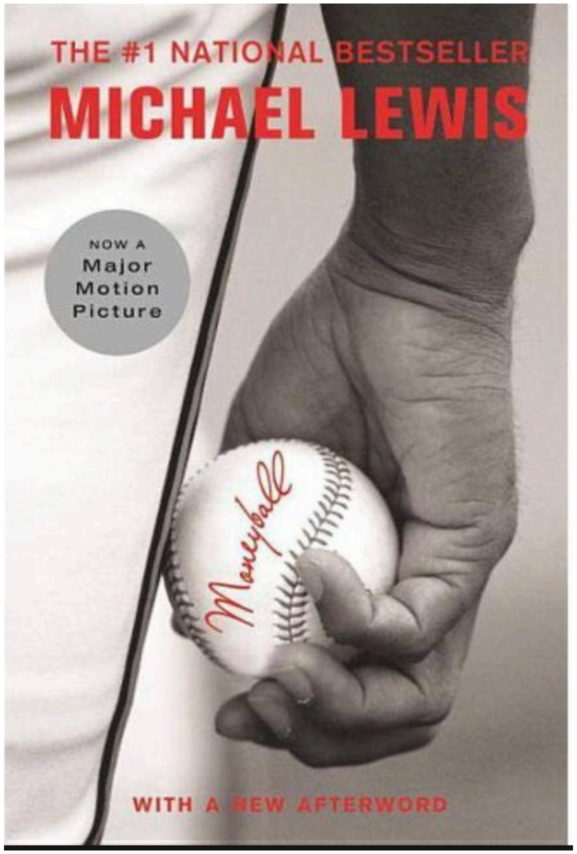
Confidence Intervals

P-values

Statistical significance tests

Sports analytics





Businesses, hospitals, sports teams, universities, and institutions of all kinds use statistics to drive their decisions.

This course won’t make you an expert on statistics, but it will teach you enough to carry out basic statistical studies on your own and to understand studies by others that you may encounter in the future.

Review of Syllabus

Course components:

- Lectures Mondays, Tuesdays, and Fridays
- Online homework (WeBWorK) due Wednesdays and Fridays by 8:00 pm.
- Weekly worksheet for extra practice, Wednesdays except exam weeks
- In class quizzes Fridays except exam weeks, closely related to the worksheets.

Warning from a previous instructor: tack all of this via the Modules page (don't rely on the Canvas calendar).

Major assessments:

- Two midterms (Wednesdays of Weeks 4 and 8)

- Final exam: 10:15 am–12:15 pm on Friday 20 March 2026.

Chapter 1: Picturing Distributions (with graphs)

Chapter 2: Describing Distributions (with numbers)

A better title: “dealing with data”

Definitions

Individuals are the objects described by a set of data.

A variable is any characteristic of an individual.

A categorical variable places an individual into one of several groups.

A categorical variable places an individual into one of several groups.

A quantitative variable takes numerical values for which arithmetic operations make sense. (Often there is a unit of measurement.)

Example: Grade data for a course.
The students are the individuals.

A numerical exam score is an example of a quantitative variable.

The letter grade for the course is an example of a categorical variable.

Estimated number of births of seals on St. Paul Island:

Year	Births (in thousands)
1979	245.93
1980	203.82
1981	179.44
1982	203.58
1983	165.94
1984	173.27
1985	182.26

What are the “individuals”? The “population”?

What are the “variables”?

The distribution of a variable tells us what values it takes and how often it takes these values.

- The values of a categorical variable are labels for the categories.

  A distribution lists the categories and gives the count or the percent of individuals who fall in each category.

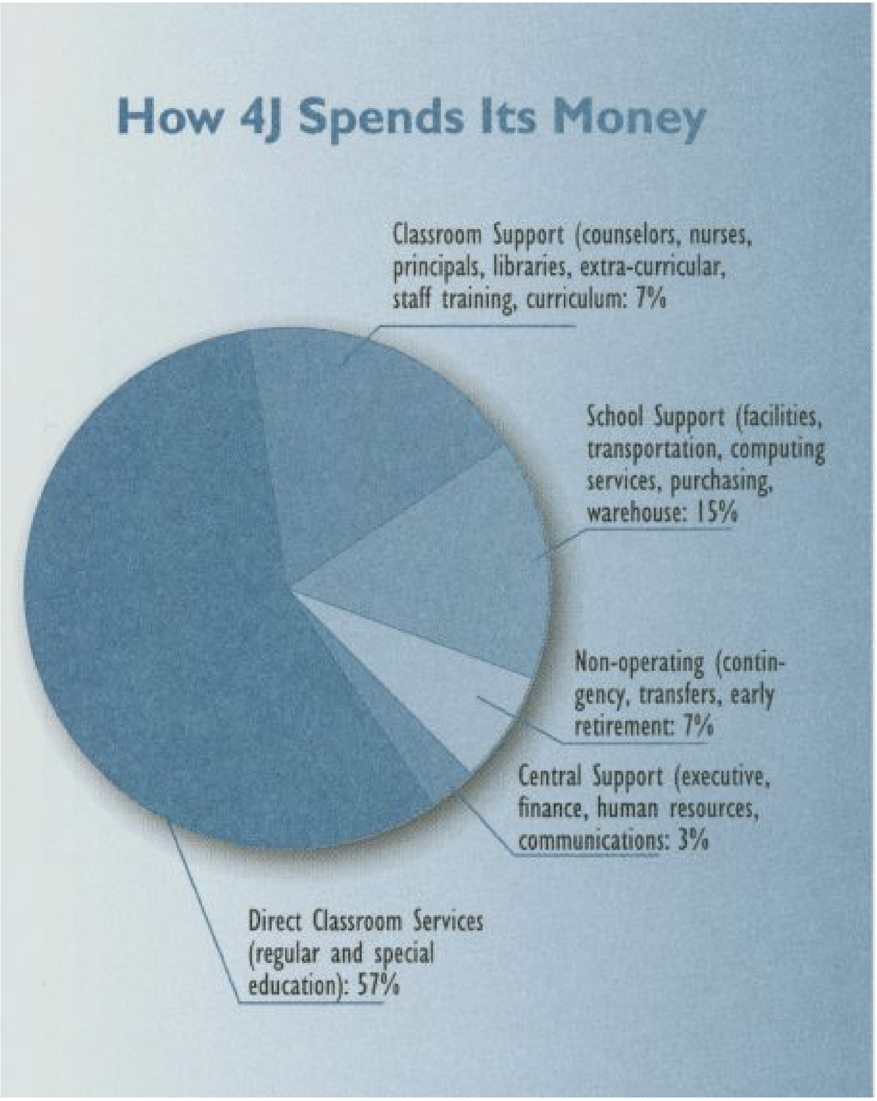
- The values of a quantitative variable are the possible numbers assigned to an individual.

- An outlier is an observation that subjectively doesn’t fit with the rest of the data.

There are different ways to graphically represent data:

- pie chart
- bar graph
- histogram
- stemplot

What is wrong with this pie chart? (It was in a mailing by the Eugene school some years ago.)



The Centers for Disease Control (CDC) tracks rates of drug poisoning mortality[1] by year, county, and age range. For 2016 across all Oregon counties, a portion of the rates for deaths related to drug poisoning is given below. Express the distribution as a histogram with classes of width 2.

Number (deaths per 100k)	Frequency
2 – 3.9	1
4 – 5.9	0
6 – 7.9	1
8 – 9.9	2
10 – 11.9	5
12 – 13.9	6
14 – 15.9	4

[1] https://catalog.data.gov/dataset/drug-poisoning-mortality-by-county-united-states

Number (deaths per 100k)	Frequency
2 – 3.9	1
4 – 5.9	0
6 – 7.9	1
8 – 9.9	2
10 – 11.9	5
12 – 13.9	6
14 – 15.9	4

We can describe the overall pattern of a distribution by its shape, center, and spread.

Some basic terms about shape: symmetric, skewed-left, and skewed-right

Center

The mean is the average, denoted $\bar{x}$. If your n values are $x_1, x_2, \ldots, x_n$, then

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

OR

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The median is another measure of center, denoted M. It is the middle value.

If the number of observations is even, M is midway between the two center observations.

Example

What are the mean and median of the worker commute times (in minutes) in North Carolina given below?

5, 10, 60, 40, 30, 10, 20, 15, 30, 10, 10, 40, 12, 20, 25

Example continued

Suppose the person with the 60 minute commute moves for away and now has a 180 minute commute time, so the list of times would be

5, 10, 180, 40, 30, 10, 20, 15, 30, 10, 10, 40, 12, 20, 25

Compute the mean and median.

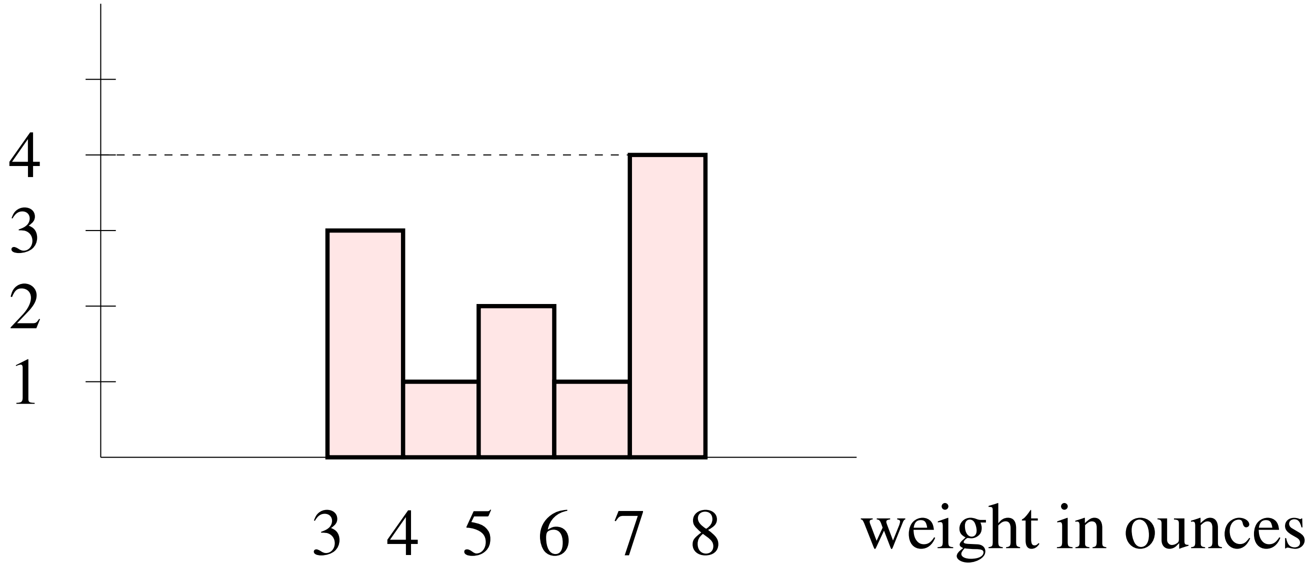
Mean and Median

- The mean and median are close together if the distribution is symmetric.
- The median is a resistant measure—it is resistant to the influence of outliers. The mean is not resistant.
- In a skewed distribution, the mean is typically farther out the extended side. “The mean moves away from the median in the direction of the skew.”

- The strength of the mean is that it is easy to calculate, even for large values of n. Its primary weakness is that it is strongly affected by outliers.

- The strength of the median is that it is resistant to outliers. Its primary weakness is that it is time-consuming to order a very large number of observations.

Sample problem: weights of hamsters



What is the total number of hamsters?

What percentage of the population weighs 6 ounces or more?

Another way to measure spread is via the **five-number summary**.

Example:

Consider the NC commute times:
5, 10, 60, 40, 30, 10, 20, 15, 30, 10, 10, 40, 12, 20, 25

Here they are in order:
5, 10, 10, 10, 10, 12, 15, 20, 20, 25, 30, 30, 40, 40, 60

Min = 5, $Q_1 = 10$, $Q_2 = M = 20$, $Q_3 = 30$, Max = 60

A box plot is a graph of the five-number summary.

In the last example we had

Min = 5, $Q_1 = 10$, $Q_2 = M = 20$, $Q_3 = 30$, Max = 60

WARNING:

Some software packages calculate the quartiles a little differently from the way the textbook does it. For example, when computing $Q3$ do we **include** the median in the data or not? There are competing conventions about this.

This won’t produce much of a difference for large data sets, but it can make a difference for small data sets.

In this course, when asked to compute quartiles by hand always do it the way we did in the previous examples. This is also the way that most calculators do it.