Descriptive Statistics, pt. II

ECON 3640–001

Marcio Santetti

Spring 2022

Motivation

Another way of describing our data

Last time, we started the process of knowing our data better by looking at **visual** descriptive techniques.

Another way of describing complex sets of data is through **descriptive numerical techniques**.

These techniques are divided in **two** main categories:

- Measures of *central location*;
- Measures of *variability*.

Measures of central location

Measures of central location

Among the **most popular** measures of central location, we will study the following:

1. *Mean*;
2. *Median*;
3. *Mode*.

Measures of central location

Among the most popular measures of central location, we will study the following:

1. *Mean*:

The arithmetic mean, also known as the *average*, is simply the sum of all observations in a data set, divided by the total number of observations.

- **Population mean** (μ):

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

- **Sample mean** ($\bar{x}$):

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

Measures of central location

1. *Mean*:

- **Population mean** (μ):

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

where the numerator is the sum of each observation contained in the data set (x_i), from the first $(i = 1)$ until the N^{th} data point. The denominator is the total population size (N), i.e., the total number of observations within this population.

- **Sample mean** $(\bar{x})$:

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

where n is the *sample* size.

Notice the difference in notation when referring to sample and population measures.

Measures of central location

Among the most popular measures of central location, we will study the following:

1. *Mean*;
2. *Median*:

To calculate the **median**, we need to place all observations *in order* (it does not matter whether ascending or descending).

The observation that lies in the **middle** is the median.

It serves for both population and sample medians.

Measures of central location

Among the most popular measures of central location, we will study the following:

1. *Mean*;
2. *Median*;
3. *Mode*:

The **mode** of a data set is (are) the observation(s) that occur(s) with the *highest frequency*.

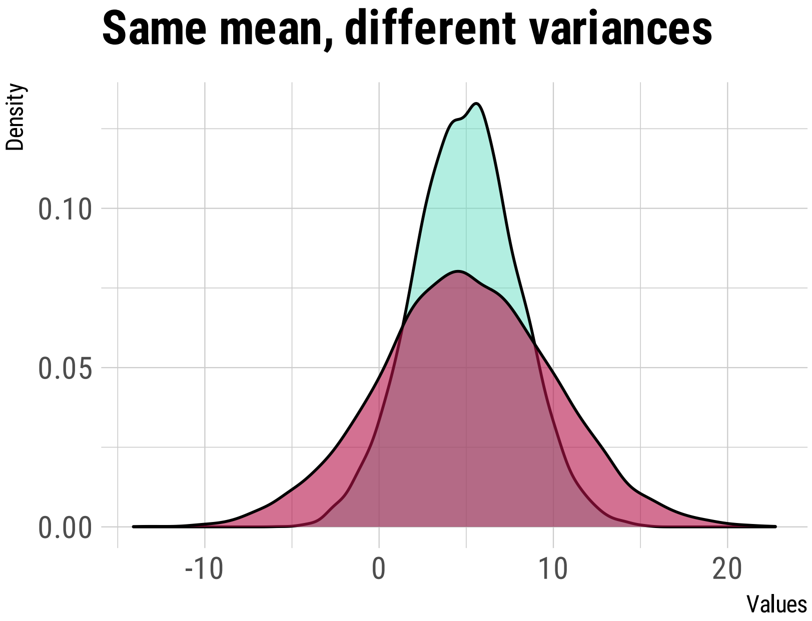
It works in the same way with populations and samples.

Measures of variability

Measures of variability

Up until now, we were interested in information about the central location of a data set.

However, these measures do not tell us anything about **how spread out**, or how **concentrated** are the data.



To address this issue, we will study the following:

1. *Range*;
2. *Variance*;
3. *Standard deviation*.

Measures of variability

To address the variability issue, we will study the following:

1. *Range*:

The **range** of a data set is simply its *largest* observation minus the *smallest*.

Measures of variability

To address the variability issue, we will study the following:

1. *Range*;
2. *Variance*:

The **variance** is the average of the *squared deviations* of each observation within a data set from its mean.

- **Population variance** (σ^2):

$$\sigma_x^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- **Sample variance** (s^2):

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Measures of variability

An *alternative* formula for the sample variance, s^2:

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

Measures of variability

To address the variability issue, we will study the following:

1. *Range*;
2. *Variance*;
3. *Standard deviation*:

The **standard deviation** is simply the *squared root* of the variance.

- **Population standard deviation** (σ):

$$\sigma_x = \sqrt{\sigma_x^2}$$

- **Sample standard deviation** (s):

$$s_x = \sqrt{s_x^2}$$

Measures of variability

The **intuition** behind the variance and standard deviation measures:

> *The idea behind the variance and the standard deviation as measures of spread is as follows: the deviations (x_i - $\bar{x}$) display the spread of the values x_i about their mean $\bar{x}$. Some of these deviations will be positive and some negative because some of the observations fall on each side of the mean. In fact, the sum of the deviations of the observations from their mean will always be zero. Squaring the deviations makes them all positive, so that observations far from the mean in either direction have large positive squared deviations. The variance is the average squared deviation. Therefore, s^2 and s will be large if the observations are widely spread about their mean, and small if the observations are all close to the mean.*

(Moore, McCabe, and Craig, 2009, p. 41)

Measures of relative standing

Measures of relative standing

Measures of **relative standing** are designed to provide information about the *position* of particular values, relative to the entire data set.

The **median** can also be interpreted as one of these measures, since it locates the *central point* of a data set, relative to its entirety.

Let us see other measures in more detail:

- **Percentile**: the p^{th} percentile is the value for which p percent are *less* than that value.
  - And $(100 - p)\%$ are *above* that value.
- **Quartile**: the 25^{th}, 50^{th}, and 75^{th} percentiles are called *quartiles.*
  - Any guesses about what the 50^{th} percentile is equal to?

Measures of relative standing

In order to **locate** a specific percentile within a data set, it may not always be straightforward to do so.

The following formula helps:

$$L_p = (n+1)\frac{p}{100}$$

where n is the sample size, and p is the percentile we would like to find.

Measures of relative standing

The **interquartile range** (*IQR*) is obtained by subtracting the *third* from the *first* quartile ($Q_3 - Q_1$).

It measures the *spread* of the middle 50% of observations.

- Large values will mean that Q_1 and Q_3 are far apart, indicating *high variability* (spread) in the data set.

Measures of relative standing

A *visual* tool that helps the statistics practitioner on measures of relative standing is the **box plot** (*aka* box & whiskers plot).

It shows **5** statistics:

1. The *minimum* value;
2. The *maximum* value;
3. And the first, second, and third *quartiles* (Q_1, Q_2, and Q_3).

The **3** main steps to construct a box plot are the following:

1. Place the observations in *ascending* order;
2. Separate *maximum* and *minimum* observations;
3. Identify the 3 *quartiles* (Q_1, Q_2, and Q_3).

Measures of relative standing

In a data set, we may sometimes find **unusally** *large* or *small* values.

Such values are called **outliers**.

They may appear due to wrong entrances in data spreadsheets, processing errors, or because the data set actually contains unusual observations.

That being said, *how* to identify outliers in a data set?

There are a few techniques, and the one we will cover in this course is by using information used for *box plots*, as well as the *interquartile range*.

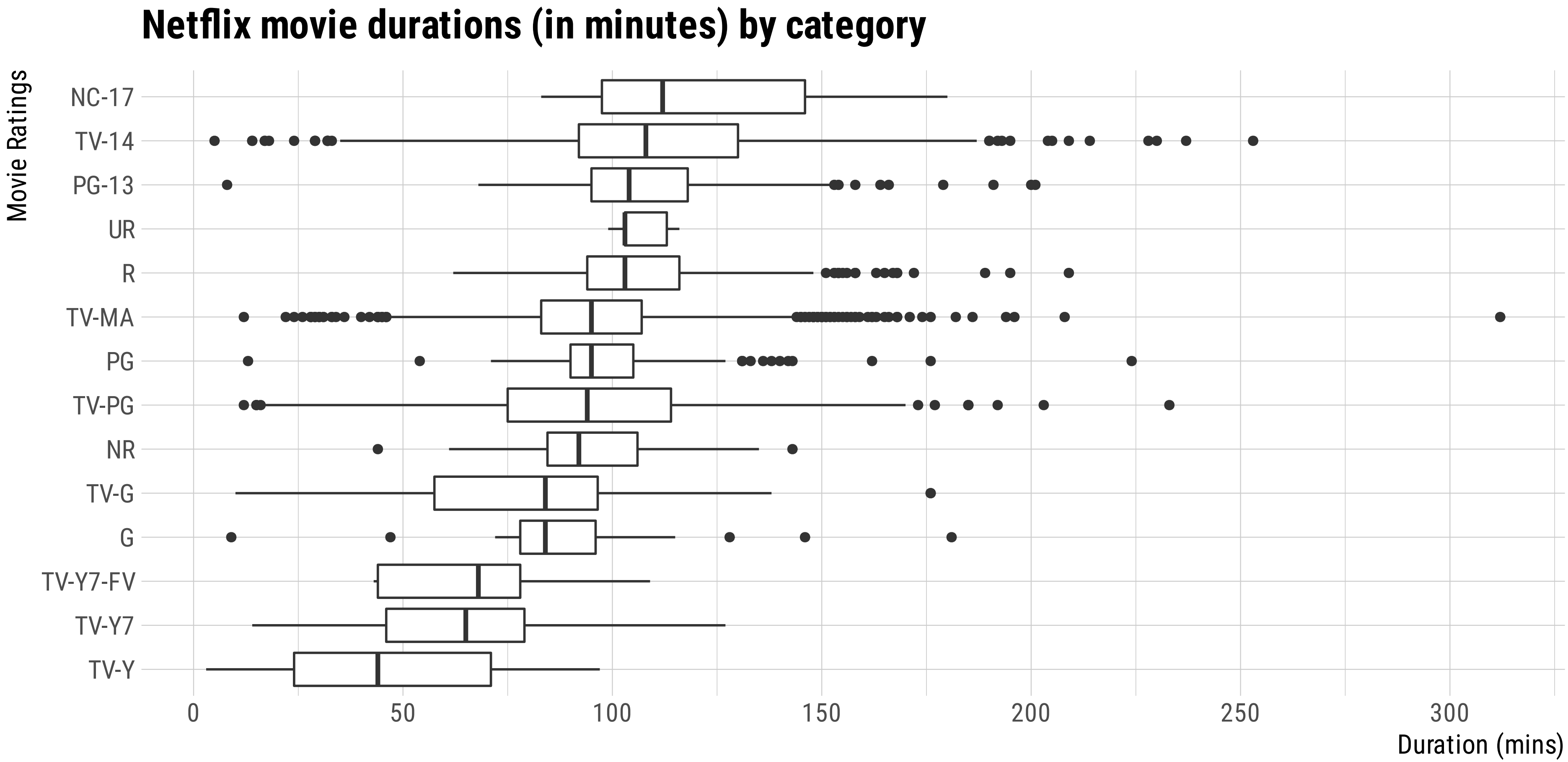
A data point is an *outlier* if it is **smaller** than

$$Q_1 - 1.5(Q_3 - Q_1)$$

A data point is an *outlier* if it is **greater** than

$$Q_3 + 1.5(Q_3 - Q_1)$$

Measures of relative standing



Next time: Descriptive Statistics in R