Introduction to Quantitative Science Seminar
Week 2

ricardo.mellado.19@ucl.ac.uk

Ricardo Mellado Labbe

Brief recap Week 1 (1)

- **Statistics** is the art and science of collecting, presenting and analyzing data to answer an investigative question. Its ultimate goal is translating data into knowledge and an understanding of the world around us.

Example: Using a Survey to investigate People Beliefs

The General Social Survey (GSS) is a survey of a few thousand adult Americans provides data about opinions and behaviors of the American public (**"Would you be willing to pay higher prices in order to protect the environment?" "How much TV do you watch per day**)

Elements of Statistical Analysis

a) **Design:** Stating the goal and/or statistical question of interesting and planning how to obtain data that will adress them
b) **Description:** Summarizing and analyzing the data that are obtained
c) **Inference:** Making decisions and predictions based on the data for answering the statistical question

Brief recap Week 1 (2)

- The first step in analyzing data collected in a variable(s) is to describe key features of the distribution of a variable
- **Distribution:** describes how the observations fall (are distributed) across the range of possible values
- The distribution for a **categorical** variable (e.g. school type) shows all possible categories and the number (or proportions) of observations falling in those categories
- For a **numerical** variable, the entire range of possible values is split up into separate intervals, and the number (or proportions) falling in each interval is given

Measuring the CENTER of Quantitative Data

- **Mean**: is the sum of the observations divided by the number of observations

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

- **Median**: is the middle value of the observations when the observations are ordered from the smallest to the largest (or the other way around!)
- **Mode**: the value that occurs most frequently. It describes a typical observation in terms of the most common outcome (more useful when describing categorical variables)

Measuring the VARIABILITY of Quantitative Data

Variability: the extent to which individual data points in a dataset deviate from the central tendency(e.g. mean)

- **Range**: is the difference between the largest and the smallest observation

$$\text{Range} = \text{Max} - \text{Min}$$

where: - Max is the maximum value in the dataset. - Min is the minimum value in the dataset.

- **Standard deviation**: Standard deviation is a statistical measure that quantifies the amount of variation or dispersion in a set of values. It provides a way to understand how spread out the values in a data set are from the mean (average). A higher standard deviation indicates greater variability, while a lower standard deviation suggests that the values are closer to the mean

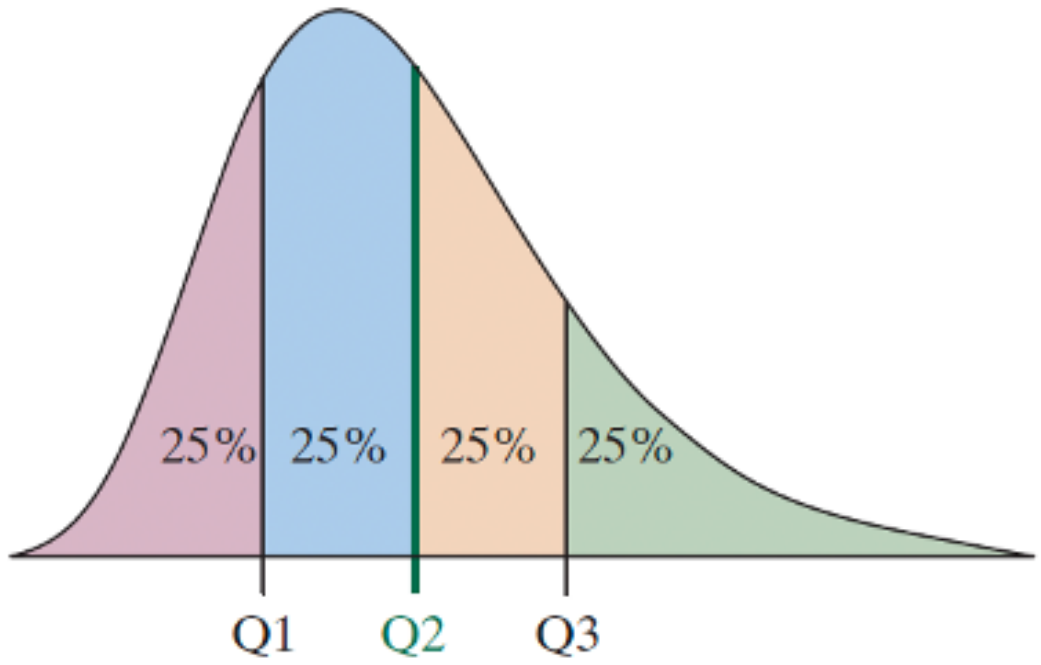
$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Measures of Position

Percentile: is a value such that p percent of the observations fall at or below that value

- **Example**: You are informed that your score 1,200 (out of 1,600) falls in the 90th percentile. Then, 90% of those who took the exam scored 1,200 or below
- **Quartiles**
- The first quartile has $p = 25$, therefore 25% falls it
- The second quartile has $p = 50$, therefore 50% falls below it (**median**!)
- The third quartile has $p = 75$, so the highest 25% of the data balls above it.

Measures of Position 2



Measures of Position to Measure Variability

- The **Interquatile range** summarizes the range of the **middle half** of the data. The middle 50% of the observations fall between the first quartile and the third quartile - 25% from Q1 to Q2 and 25% from Q2 and Q3.

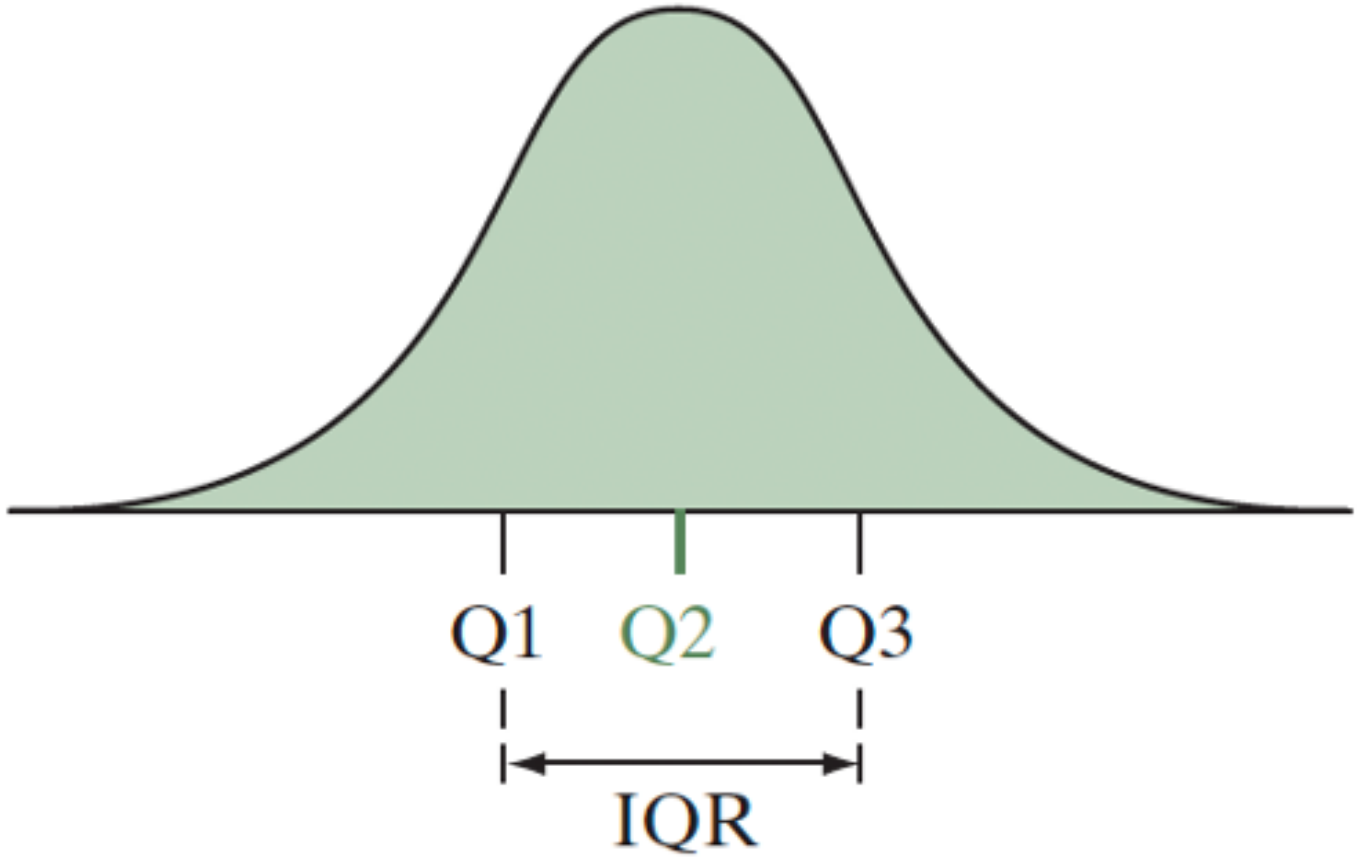
The Interquartile Range (IQR) is calculated using the formula:

$$\text{IQR} = Q3 - Q1$$

Where:

- $Q1$ is the first quartile (25th percentile).
- $Q3$ is the third quartile (75th percentile).

Measures of Position to Measure Variability



Today seminar database

- Contains the number of assaults incidents in which an ambulance has been called in London between 2009 and 2011

Filter

	BorCode	WardName	WardCode	WardType	Assault_09_11
1	00AA	Aldersgate	00AAFA	Prospering Metropolitan	10
2	00AA	Aldgate	00AAFB	Prospering Metropolitan	0
3	00AA	Bassishaw	00AAFC	Prospering Metropolitan	0
4	00AA	Billingsgate	00AAFD	Prospering Metropolitan	0
5	00AA	Bishopsgate	00AAFE	Prospering Metropolitan	188
6	00AA	Bread Street	00AAFF	Prospering Metropolitan	0
7	00AA	Bridge & Bridge Without	00AAFG	Prospering Metropolitan	0
8	00AA	Broad Street	00AAFH	Prospering Metropolitan	0
9	00AA	Candlewick	00AAFJ	Prospering Metropolitan	0
10	00AA	Castle Baynard	00AAFK	Prospering Metropolitan	0

Showing 1 to 10 of 649 entries, 5 total columns

Seminar tasks and questions

Seminar Task: Use the seminar, you will be continuing with the data set Ambulance and Assault Incidents data.csv. Still work with the data frame object named as London.Ambulance.

- Create a new object / data set that only contains data for ward type Suburbs and Small Towns. Hint: Try to subset the data by filtering it based on WardType.
- Calculate the mode, median, mean, range, interquartile range, and standard deviation for the Assault_09_11 variable for Suburbs and Small Towns
- Produce a boxplot() that provides a visual description of it's distribution

Subsetting

```
Subs_towns <- London.Ambulance[London.Ambulance$WardType == 'Suburbs and Small Towns',]
ProsMetro <- London.Ambulance[London.Ambulance$WardType == 'Prospering Metropolitan',]
```

Summary statistics

- Summary statistics of "number of assaults incidents" for wards categorize as "Prospering Metropolitan"

```
summary(ProsMetro$Assault_09_11)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    75.0   119.0   153.3   184.0  1582.0
```

```
sd(ProsMetro$Assault_09_11)
```

```
## [1] 179.8195
```

- Summary statistics of "number of assaults incidents" for wards categorize as "Suburbs and Small Towns"

```
summary(Subs_towns$Assault_09_11)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    19.0    67.0    98.0   110.0   142.8   579.0
```

```
sd(Subs_towns$Assault_09_11)
```

```
## [1] 64.5271
```

Mode

```
# create a function to calculate the mode
get_mode <- function(x) {
  # get unique values of the input vector
  uniqv <- unique(x)
  # select the values with the highest number of occurrences
  uniqv[which.max(tabulate(match(x, uniqv)))]
}
# calculate the mode of the 2011 population variable
get_mode(ProsMetro$Assault_09_11)
```

```
## [1] 0
```

```
# calculate the mode of the 2011 population variable
get_mode(Subs_towns$Assault_09_11)
```

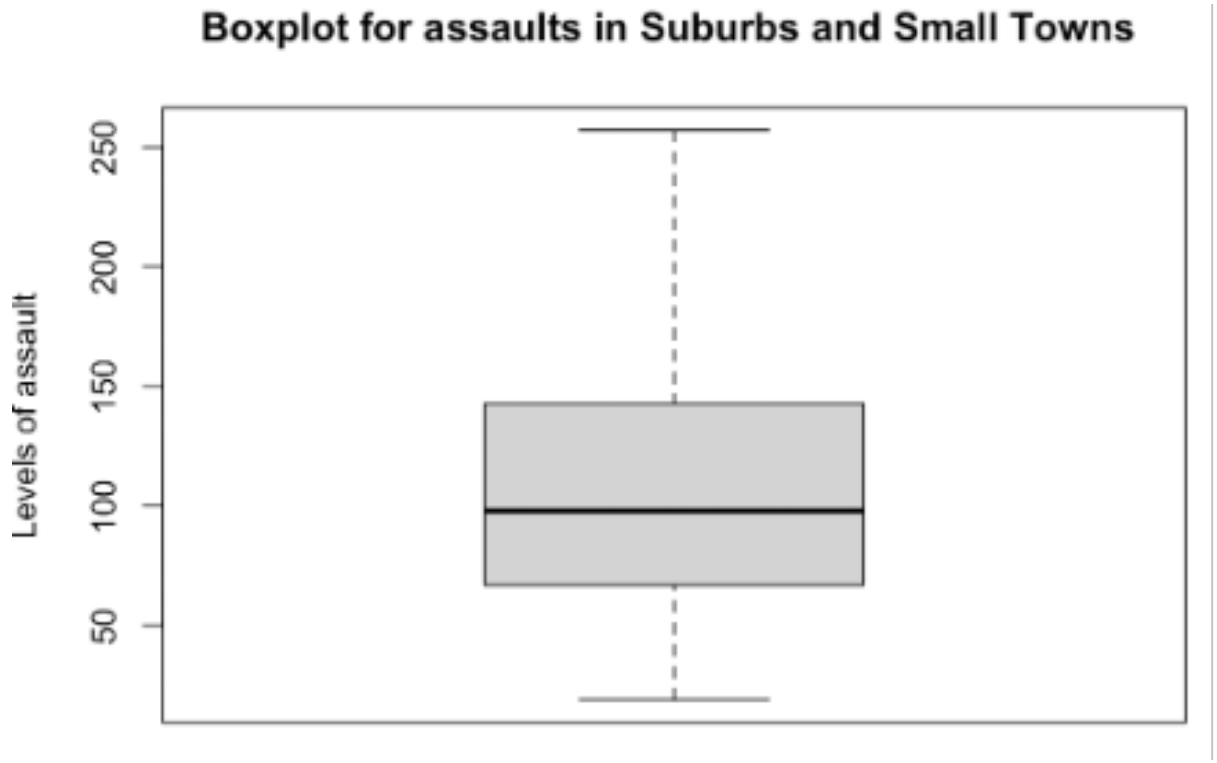
```
## [1] 89
```

(3) Boxplot of assaults for each Ward Type

```
boxplot(Subs_towns$Assault_09_11,
        outline = FALSE,
        xlab = "",
        ylab="Levels of assault", main="Boxplot for assaults in Suburbs and Small Towns ")


boxplot(ProsMetro$Assault_09_11,
        outline = FALSE,
        xlab = "",
        ylab="Levels of assault", main="Boxplot for assaults in Prospering Metropolitans a
```

(3) Boxplot of assaults for each Ward Type



Seminar questions

Seminar questions

Compare the results of the descriptive statistics you have calculated for your Suburbs and Small Towns object with the results of the descriptive statistics you have calculated for you Prospering Metropolitan object / data set.

What do these differences tell us about the levels of violent assaults within these separate environments? Create a dual boxplot to show a visual representation of this comparison.

Produce a boxplot() that provides a visual description of it's distribution

- **Note**: The rbind() function in R is used to combine two or more objects (typically data frames or matrices) by row binding them

```
# To generate the dual boxplot
# use rbind()
data <- rbind(Subs_towns, ProsMetro)
# dual boxplot
#:::: use option outline=FALSE to exclude outliers
boxplot(data$Assault_09_11 ~ data$WardType,
        outline = FALSE,
        xlab = "",
        ylab="Levels of assault", main="Boxplot [Note: Outliers were excluded]")
```

Produce a boxplot() that provides a visual description of it's distribution