STA220H1: The Practice of Statistics I

Elena Tuzhilina

January 10, 2023

Instructor

Elena Tuzhilina *elena.tuzhilina@utoronto.ca*

- Assistant Professor, Department of Statistical Sciences, U of T (since 2022)
- Major in Mathematics, Moscow State University (2015)
- PhD in Statistics, Stanford (2022)

Research interests: applied statistics, especially, with applications in biology and medicine

Industry experience: ABBYY Lingvo (computer linguistics) and Microsoft (data science)

Agenda for today

- Class logistics
- Course overview: what is statistics?
- Data
- Summary statistics
- Types of variables

Class logisitcs

Please review the course page.

- The course will closely follow the modules
- My office hours will be held in a **hybrid format**
- We will have **two in-person midterms**
- Grading policy is **quizzes (20%) + midterm 1 (20%) + midterm 2 (20%) + final (40%)**
- All communications with the TAs and instructor should be done through *sta220-win23-staff-l@listserv.utoronto.ca*

What is statistics?

There are *three major things* that we can do with statistics.

- **Describe** - the world is complex and we often need to describe it in a simplified way that we can understand
- **Decide** - we often need to make decisions based on data, usually in the face of uncertainty
- **Predict** - we often wish to make predictions about new situations based on our knowledge of previous situations

Example

Why do you think eating saturated fat is unhealthy?

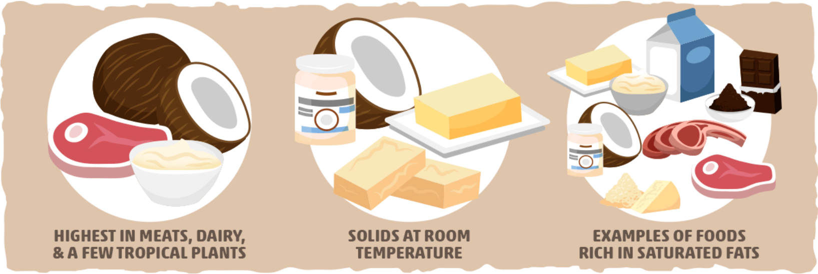


Figure 1: [picture source]

Example

Option 1: use common sense.

- If we eat fat, then it's going to turn straight into fat in our bodies
- We have all seen photos of arteries clogged with fat, so eating fat is going to clog our arteries

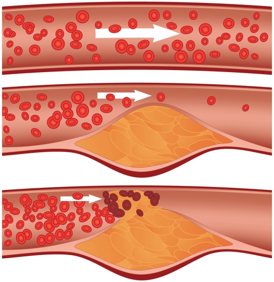


Figure 2: [picture source]

What do the data tell us?

Option 2: use data (the PURE study by Dehghan et al., 2017).

- Investigates how intake of various classes of macronutrients was related to the likelihood of dying
- Includes more than 135,000 people from 18 different countries
- People followed for median 7.4 years

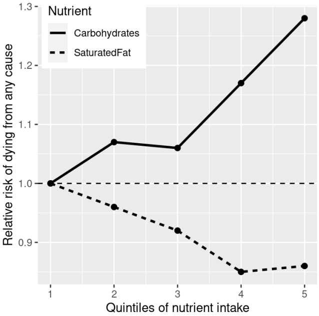


Figure 3: Intake of saturated fats and carbohydrates vs. the risk of dying

What can statistics do for us?

- **Describe** - provide a summary of the PURE data set (135,000 points!)
- **Predict** - predict how many years you will live
- **Decide** - is there a relationship between fat intake and health?

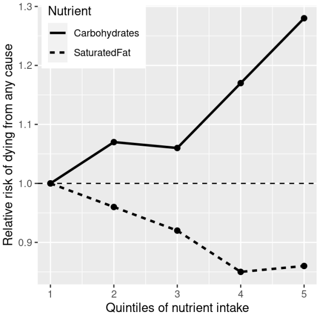


Figure 4: Intake of saturated fats and carbohydrates vs. the risk of dying

Population vs. sample

- We want to determine the value of a **statistic** for an entire **population** of interest
- Here **statistic** refers to a "number" representing certain features of a population
- We cannot investigate each population member, so we pick a **sample** (small subset) of the population

Figure 5: [picture source]

Representative sample

- We hope that small sample is sufficient to accurately estimate the statistic of interest
- **Representative sample** is one in which every member of the population has an equal chance of being selected
- When this fails, the statistic we compute on the sample may be **biased** (i.e. its value is systematically different from the population value)

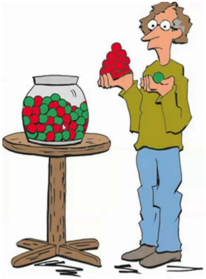


Figure 6: [picture source]

Example

Non-representative sample

- Define prescription dosage for a drug using male only sample
- Compute average income of a country using people with high education only

Figure 7: [picture source]

What are data?

- **Data** contain information about a sample and usually come in the form of a table

Example: an experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens

weight	feed
141	linseed
216	casein
392	sunflower
179	horsebean
171	soybean
320	sunflower
332	casein
169	linseed
258	meatmeal

Data are composed of

- **Variables** contain information about some specific thing (columns of the table) 2 var = (weight, feed)
- **Observational units** are things on which measurements are taken Chicken
- **Observations** are actual values of variables for a selected observational unit (rows of the table) 9 observation

weight	feed
141	linseed
216	casein
392	sunflower
179	horsebean
171	soybean
320	sunflower
332	casein
169	linseed
258	meatmeal

Exercise

Example: fuel consumption and 10 other aspects of automobile design and performance for 6 automobiles

What are the observational units? How many observations and variables are there?

(car) (6) (11)

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
15.2	8	275.8	180	3.07	3.780	18.00	V	A	3	3
19.2	8	400.0	175	3.08	3.845	17.05	V	A	3	2
21.4	6	258.0	110	3.08	3.215	19.44	S	A	3	1
14.3	8	360.0	245	3.21	3.570	15.84	V	A	3	4
21.0	6	160.0	110	3.90	2.620	16.46	V	M	4	4
21.0	6	160.0	110	3.90	2.875	17.02	V	M	4	4

Type of variables: quantitative

- Most commonly in statistics we will work with **quantitative** data, meaning data that are numerical

How tall are you in inches?

74.0, 64.0, 65.0, 64.0, 64.0, 72.5, ...

How many hours per weeks do you spend on HWs at U of T?

10.0, 5.5, 2.0, 13.5, 8.0, ...

Type of variables: qualitative

- Some variables are **qualitative (categorical)**, meaning that they describe a quality rather than a numeric quantity

What is your favorite food?

Berries, Chocolate, Pasta, Pizza, . . .

Which programming languages do you have experience with?

None, Python, R, Java, . . .

Type of variables: qualitative

- **Qualitative** variables can be **nominal** and **ordinal**
- For **nominal** each number represents a different thing.

What color are your eyes?

1 2 3 4

Blue, Green, Grey, Brown, ...

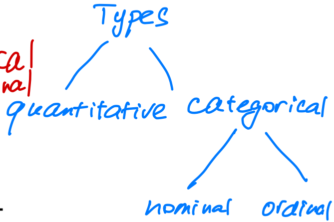
- For **ordinal** values have an ordered relationship to one another

What size is your clothes?

XSmall, Small, Medium, Large, XLarge, ...

1 2 3 4 5

Exercise



What are the types of variables?

quantitative — categorical: nominal, ordinal

weight	feed	protein
141	linseed	low
216	casein	high
392	sunflower	low
179	horsebean	high
171	soybean	high
320	sunflower	low
332	casein	high
169	linseed	high
258	meatmeal	high

low < high

Data summary: one quantitative variable

Why do we summarize data?

- It provides us with a way to **generalize** - that is, to make general statements that extend beyond specific observations

Two ways to summarize the data

- Compute **numerical summary (summary statistics)** - mean, minimum, maximum, range, median, quartiles, IQR, standard deviation
- Summarize using **plots** - histogram, boxplot

What can we say about students grades?

Example: the grades (out of 100) for 9 students of STA220H1

```
sta220.data
```

```
##             student grade
## 1      Jenny Holder    67
## 2        Tammy Snow    88
## 3     Victoria Hall    90
## 4    Saoirse Spence    72
## 5       Raja Cooper    94
## 6 Nicolas Roberson    77
## 7    Finnley Wright    85
## 8      Nate Mcgrath    93
## 9    Joshua Pollard    82
```

Summary statistics

Notations

- n will denote the number of observations
- $x_1, x_2, ..., x_n$ will denote the observations itself

```
sta220.data$grade
```

```
## [1] 67 88 90 72 94 77 85 93 82
```

$n = 9$

$x_1, x_2, x_3, x_4, \ldots, x_9$

x_n

$$\frac{67+88+90+\cdots+82}{9}$$

Summary statistics: mean

$$\sum_{i=2}^{n-1} x_i = x_2 + x_3 + \dots + x_{n-1}$$

*What is the **average** grade in STA220?*

$x_1 + x_2 + \dots + x_n$

sum of obs

$$mean = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

number of obs

$$\sum_{i=1}^{n} x_i = (x_1 + x_2 + \dots + x_n)$$

Summary statistics: mean

*What is the **average** grade in STA220?*

```
mean(sta220.data$grade)
```

```
## [1] 83.11111
```

Summary statistics

Notations

- $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ will denote sorted observations, i.e. $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

```
sort(sta220.data$grade)
```

```
## [1] 67 72 77 82 85 88 90 93 94
```

Summary statistics: min and max

*What are the **minimum and maximum** grade in STA220?*

$minimum = x_{(1)}$

$maximum = x_{(n)}$

Summary statistics: min and max

*What are the **minimum and maximum** grade in STA220?*

```
sort(sta220.data$grade)
```

```
## [1] 67 72 77 82 85 88 90 93 94
```

min max

```
min(sta220.data$grade)
```

```
## [1] 67
```

```
max(sta220.data$grade)
```

```
## [1] 94
```

Summary statistics: median

What is the ***median*** *grade in STA220?*

- If we were to sort all of the values in order of their magnitude, then the **median** is the value in the middle

```
sort(sta220.data$grade)
```

median

```
## [1] 67 72 77 82 85 88 90 93 94
```

```
median(sta220.data$grade)
```

```
## [1] 85
```

Summary statistics: median

*What is the **median** grade in STA220?*

- If there is an even number of values then there will be two values tied for the middle place, in which case we take **the average (i.e. the halfway point) of those two numbers**

If we had 8 grades:

```
sort(sta220.data$grade[1:8])
```

```
## [1] 67 72 77 85 88 90 93 94
```

$\frac{85+88}{2}$

```
median(sta220.data$grade[1:8])
```

```
## [1] 86.5
```

Summary statistics: median

*What is the **median** grade in STA220?*

If n is odd, then $median = x_{(r)}$ where $r = \frac{n+1}{2}$

If n is even, then $median = \frac{x_{(r)} + x_{(r+1)}}{2}$ where $r = \frac{n}{2}$

$n=9 \quad r = \frac{9+1}{2} = 5$

$x_{(5)}$

$n=8 \quad r=4$

$\frac{x_{(4)} + x_{(5)}}{2}$

Summary statistics: median vs. mean

- Both mean and median measure **central tendency** of a data set – that is, what value are the data centered around
- However, median tends to be more **robust** (less sensitive) to bad values (outlies)

```
grades
```

```
## [1] 67 88 90 72 94 77 85 93 82
```

1 4 4 3 3 1

1 1 3 3 4 4

```
median(grades)
```

```
## [1] 85
```

```
mean(grades)
```

```
## [1] 83.11111
```

Summary statistics: median vs. mean

- Both mean and median measure **central tendency** of a data set – that is, what value are the data centered around
- However, median tends to be more **robust** (less sensitive) to **outlies** (values that are much larger or smaller than the rest of the data)

```
grades.corrupted
```

```
## [1]   67   88   90   72   94   77   85 9300   82
```

```
median(grades.corrupted)
```

```
## [1] 85
```

```
mean(grades.corrupted)
```

```
## [1] 1106.111
```

Exercise

If n is odd, then $median = x_{(r)}$ where $r = \frac{n+1}{2}$

If n is even, then $median = \frac{x_{(r)} + x_{(r+1)}}{2}$ where $r = \frac{n}{2}$

Compute mean and median of the following values:

3, 10, 5, 6, 10, 9?

3 5 6 9 10 10 $\quad n = 6 \Rightarrow r = \frac{n}{2} = \frac{6}{2} = 3$

$$\frac{x_{(3)} + x_{(4)}}{2}$$

3 -1 -2 4 -2 -1 3 4

Summary statistics: first and third quartiles

- Median is the **second quartile**: to find median we sort the values and travel half way (1/2) through the sorted list
- To find the **first quartile** we travel quarter (1/4) way through the sorted list
- To find the **third quartile** we travel three quarters (3/4) way through the sorted list

```
sort(sta220.data$grade)
```

```
## [1] 67 72 77 82 85 88 90 93 94
```

min Q1 median Q3 max

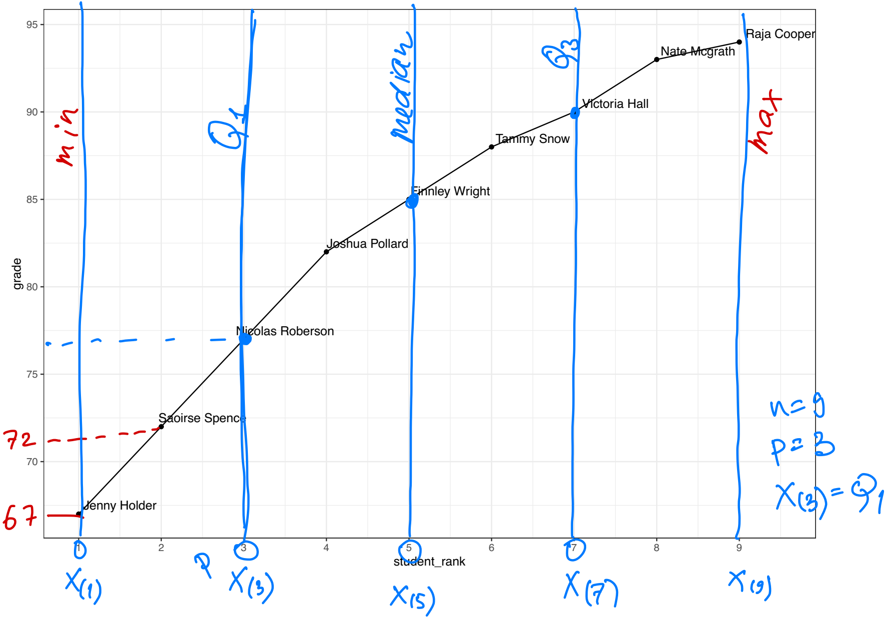
```
quantile(sta220.data$grade, 0.25)
```

```
## 25%
##  77
```

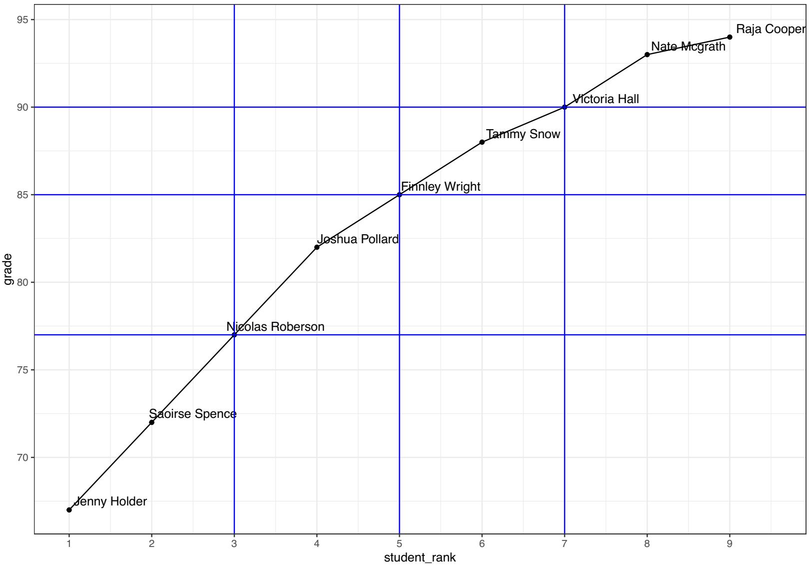
```
quantile(sta220.data$grade, 0.75)
```

```
## 75%
##  90
```

Summary statistics: first and third quartiles

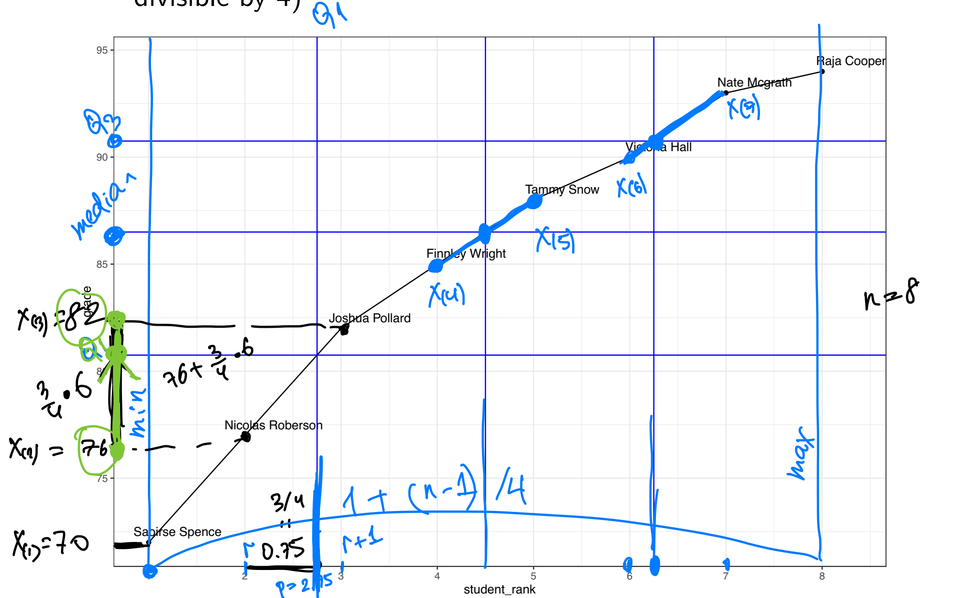


Summary statistics: first and third quartiles



Summary statistics: first and third quartiles

- Sometimes we need to use **interpolation** (when $n - 1$ is not divisible by 4)



Summary statistics: first quartiles (Q_1)

Step 1: find position $p = 1 + 0.25 \cdot (n - 1)$

Step 2: check if p is an integer

- If yes, then set the **first quartile** as $Q_1 = x_{(p)}$
- If no, we interpolate
  - Find an integer r such that $r < p < r + 1$
  - Take $Q_1 = x_{(r)} + (x_{(r+1)} - x_{(r)}) \cdot (p - r)$

$n = 8$

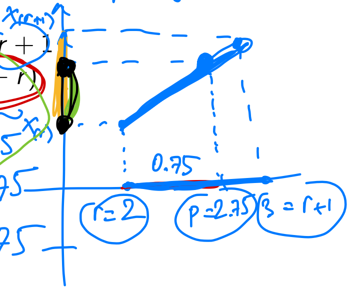
$p = 1 + \frac{7}{4} = 2.75$

$r = 2 \quad r + 1 = 3$

$Q_1 = x_{(2)} + (x_{(3)} - x_{(2)}) \cdot 0.75$

$76 + (82 - 76) \cdot 0.75$

$= 80.5$



Summary statistics: third quartiles

0.5 – median

0.25

0.75

0.3

0.7

Step 1: find position $p = 1 + 0.75 \cdot (n - 1)$

Step 2: check if p is an integer

- if yes, then set the **third quartile** as $Q_3 = x_{(p)}$
- if no, we interpolate
  - find an integer r such that $r < p < r + 1$
  - take $Q_3 = x_{(r)} + (x_{(r+1)} - x_{(r)}) \cdot (p - r)$

Summary statistics: k-th percentile

- **Percentile** is generalization of quartile
- Median is **50-th percentile**
- Q_1 is **25-th percentile**, , Q_3 is **75-th percentile**

General formula for the position is $p = 1 + \frac{k}{100} \cdot (n - 1)$

```
sort(sta220.data$grade)
```

```
## [1] 67 72 77 82 85 88 90 93 94
```

```
quantile(sta220.data$grade, 0.3)
```

```
## 30%
##  79
```

Exercise

Compute the first and the second quartiles of the following values:

3, 10, 5, 6, 10, 9?

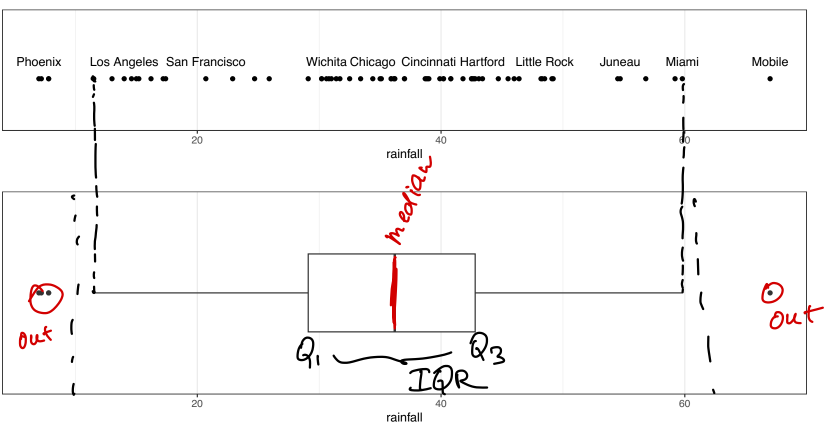
What can we say about precipitation level in the US?

Example: the precipitation (rainfall) level in inches for 69 United States cities

```
precip.data
```

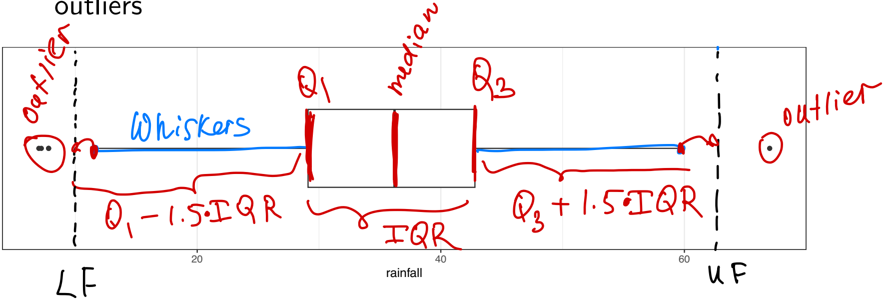
```
##                     rainfall
## Mobile                  67.0
## Juneau                  54.7
## Phoenix                  7.0
## Little Rock             48.5
## Los Angeles             14.0
## Sacramento              17.2
## San Francisco           20.7
## Denver                  13.0
## Hartford                43.4
## Wilmington              40.2
## Washington              38.9
## Jacksonville            54.5
## Miami                   59.8
## Atlanta                 48.3
## Honolulu                22.9
## Boise                   11.5
```

Plots: boxplot



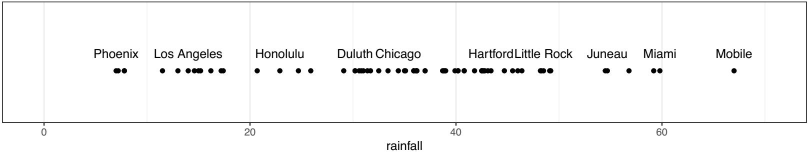
Plots: boxpolot

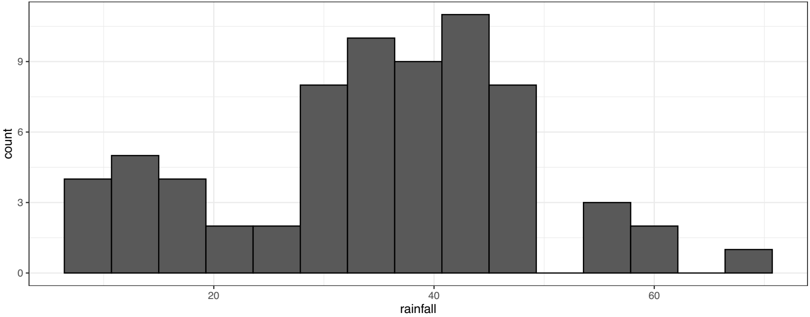
- Box represents $[Q_1, Q_3]$ range
- Thick line is median
- Box size is **interquartile range** $IQR = Q_3 - Q_1$
- **Lower and upper fences** $LF = Q_1 - 1.5 \cdot IQR$ and $UF = Q_3 + 1.5 \cdot IQR$ are not present
- **Outliers** are dots that lie outside the $[LF, UF]$ range
- **Whiskers** represent the $[min, max]$ range after excluding outliers



Plots: histogram

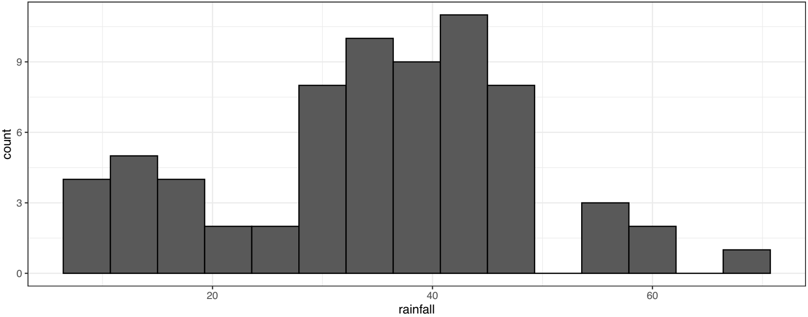
- **Histogram** is used for visualizing data **distibution**





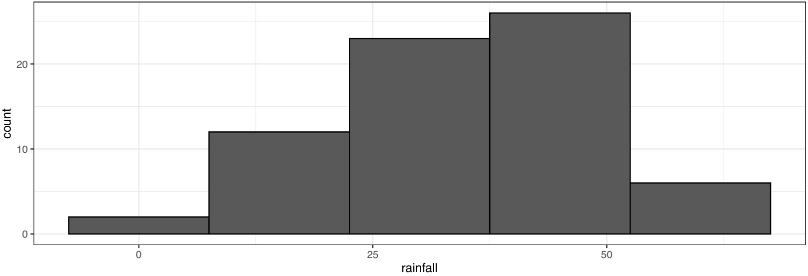
Plots: histogram

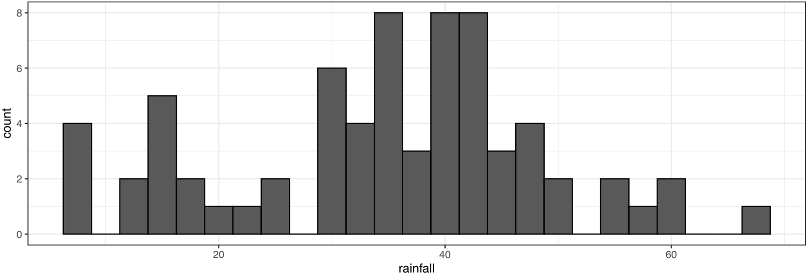
- **Bins** - x-axis is split in intervals, they should be mutually exclusive and exhaustive
- **Breaks (cutpoints)** - the values that define the beginnings and the ends of the bins
- **Counts (frequencies)** - number of data points in each bin (height of each bar)



Plots: histogram

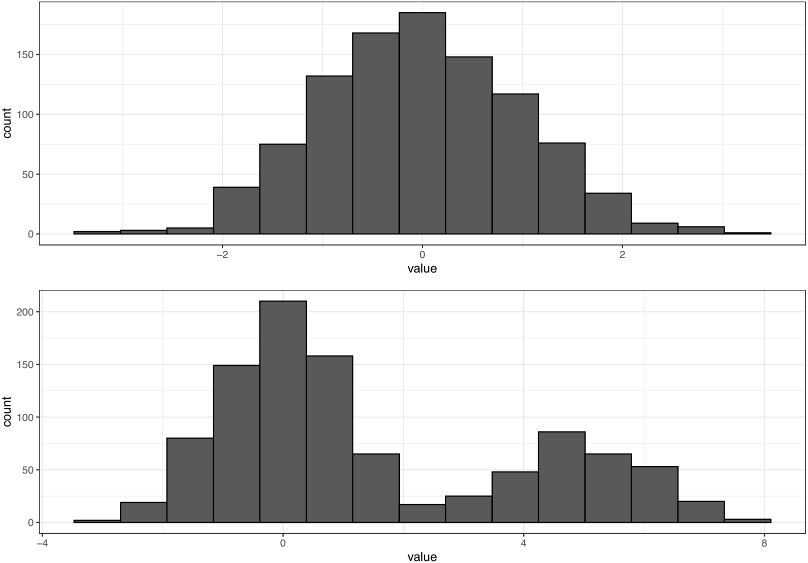
- The appearance of histogram **depends on the cutpoints**





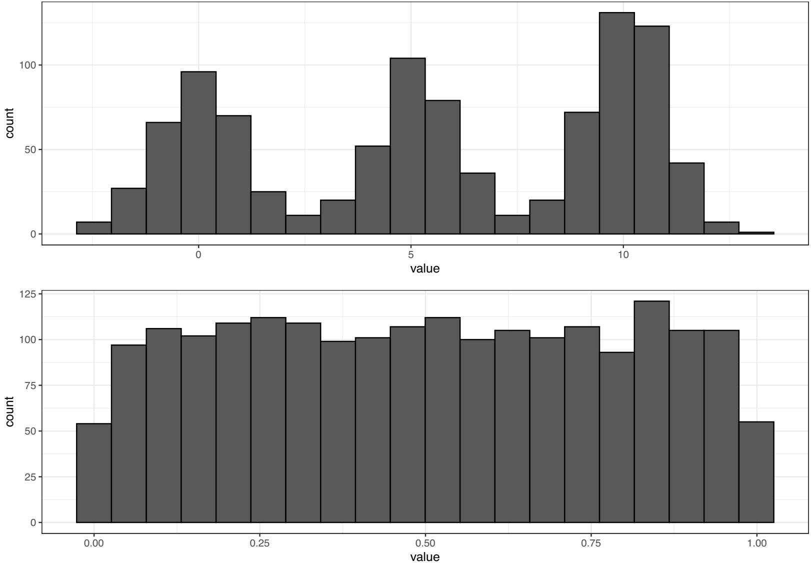
Plots: histogram

- **Mode** - the peak of the distribution
- Histogram can be **unimodal**, **bimodal**, **multimodal**, **uniform**



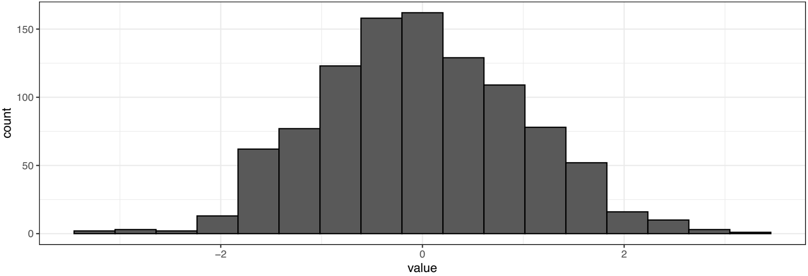
Plots: histogram

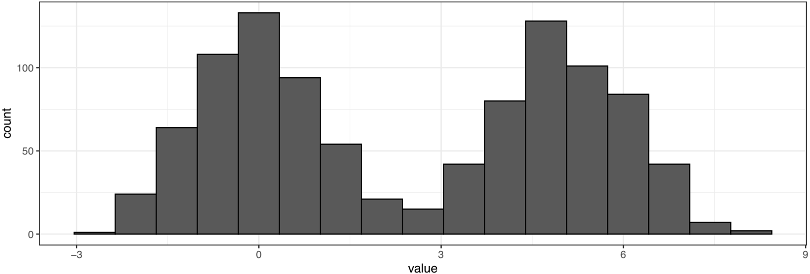
- **Mode** - the peak of the distribution
- Histogram can be **unimodal**, **bimodal**, **multimodal**, **uniform**



Plots: histogram

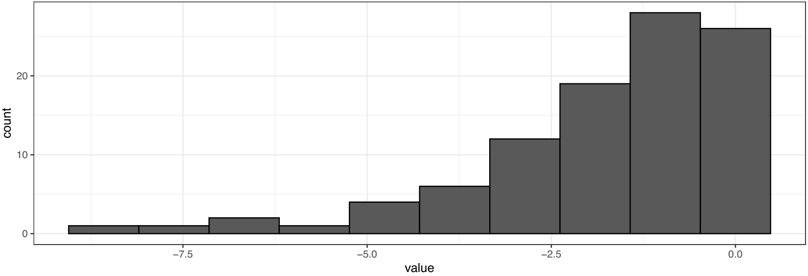
- Histogram can be **symmetric**, **left-skewed** (long left tail), **right-skewed** (long right tail)

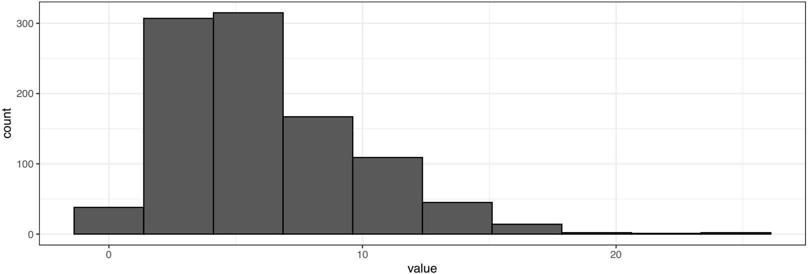




Plots: histogram

- Histogram can be **symmetric**, **left-skewed** (long left tail), **right-skewed** (long right tail)



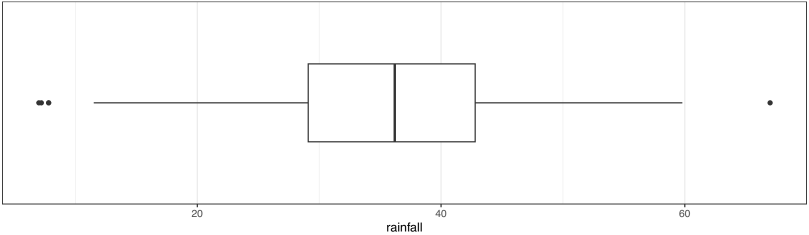


Data spread

There are several ways to measure the **spread of the data**

$$range = x_{(n)} - x_{(1)}$$

$$IQR = Q_3 - Q_1$$




```
max(precip.data$rainfall) - min(precip.data$rainfall)
```

```
## [1] 60
```

```
IQR(precip.data$rainfall)
```

```
## [1] 13.7
```

Summary statistics: standard deviation

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

$$standard\ \ deviation = \sqrt{variance}$$

```
var(precip.data$rainfall)
```

```
## [1] 190.5252
```

```
sd(precip.data$rainfall)
```

```
## [1] 13.80309
```

Exercise

Compute standard deviation of the following values:

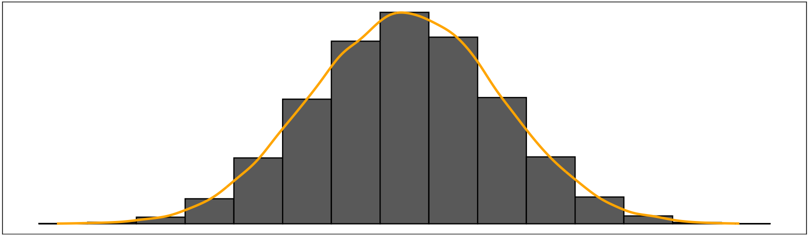
3, 10, 5, 6, 10, 8?

```
vec = c(3, 10, 5, 6, 10, 8)
summary(vec)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    5.25    7.00    7.00    9.50   10.00
```

Summary statistics: standard deviation

There is an **empirical rule** for **symmetric, unimodal, bell-shaped** distributions.



Summary statistics: standard deviation

- **68%** of the data lies in $[mean - sd, mean + sd]$
- **95%** of the data lies in $[mean - 2 \cdot sd, mean + 2 \cdot sd]$
- **99.7%** of the data lies in $[mean - 3 \cdot sd, mean + 3 \cdot sd]$

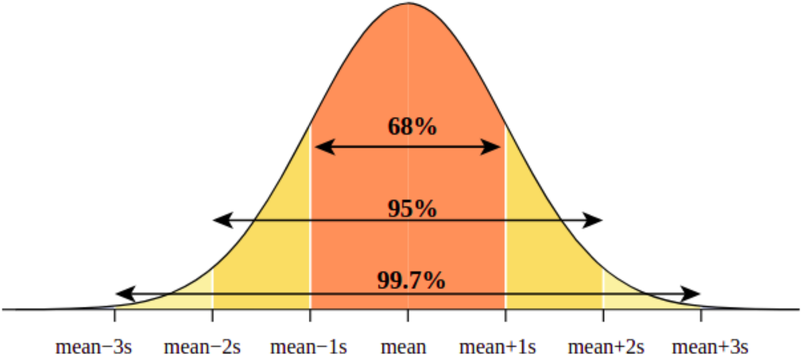


Figure 8: [picture source]

TO DO

1. Module 1. Summarizing Data: One variable and Module 5. Data collection
2. Quiz 1 due Monday (January 16) @ 11:59 PM (EST)
3. Practice Problem Set 1