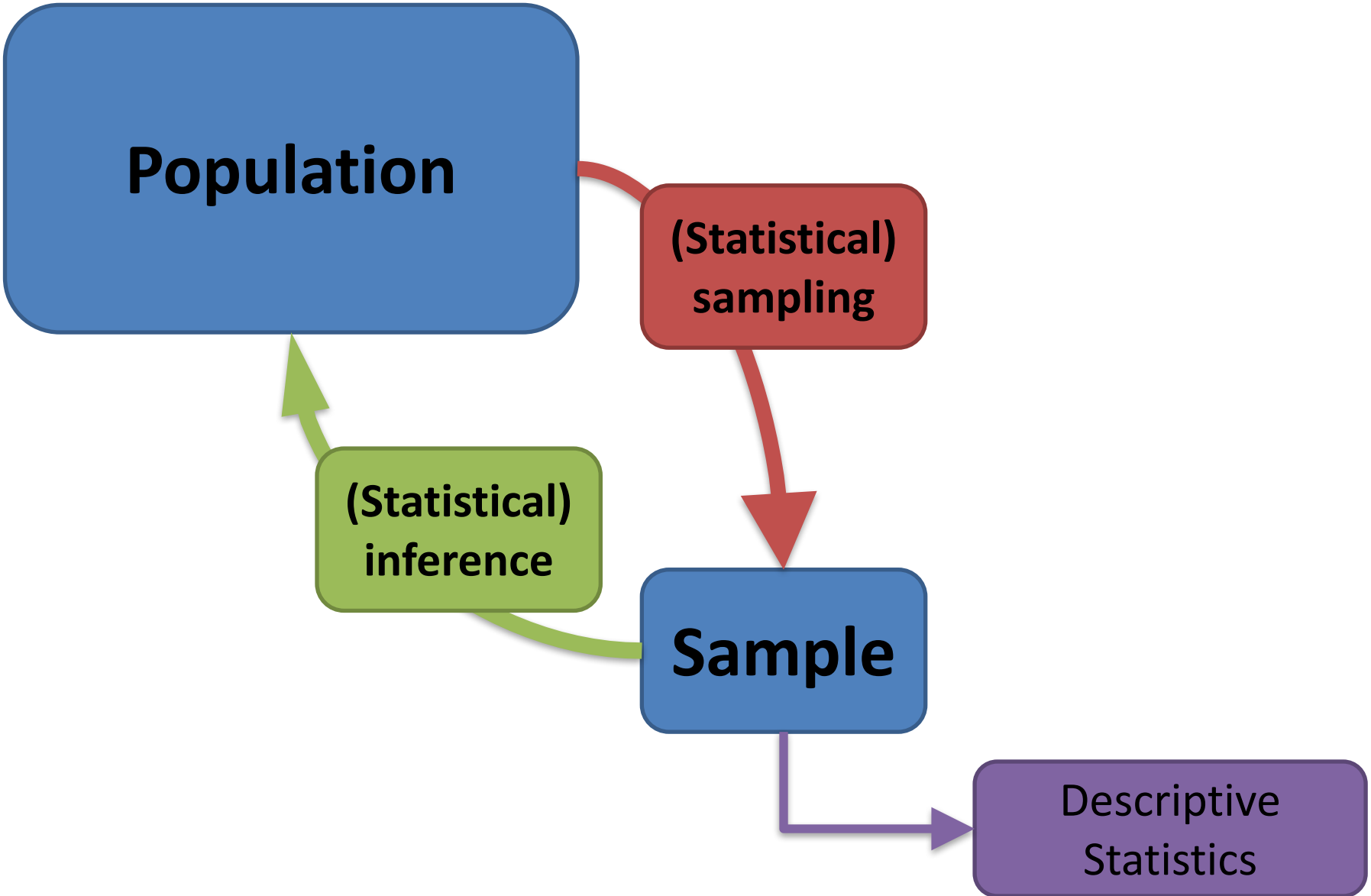
Descriptive Statistics

Prof. Christian Wallraven

wallraven@korea.ac.kr

The Big Picture



Descriptive statistics – your data

- I have asked you a few questions in the survey
  - gender
  - age
  - height
  - Favorite
    - Day and Number
  - Hours spent per day on
    - Computer Games, Studying, Sleeping
  - Your opinion on the course so far

Sampling and generalization

- Can we generalize from this sample to all KU students?

- (a) Yes
- (b) No
- (c) It depends on the data

Class Survey Data

58x10 table

	1 Gender	2 Studying	3 Sleeping	4 CGames	5 Height	6 Course	7 Age	8 FavDay	9 FavNumber	10 ShoeSize
1	"TRUE"	5	5	2	166.5000	"Strongly...	21	"Friday"	25	250
2	"FALSE"	4	6	12	175	"Agree"	25	"Friday"	1	290
3	"TRUE"	2	6	0	160	"Agree"	20	"Friday"	3	235
4	"FALSE"	7	8	2	181	"Strongly...	24	"Monday"	12	280
5	"TRUE"	4	7	0	163	"Agree"	25	"Saturday"	5	245
6	"FALSE"	6	8	6	172	"Agree"	24	"Saturday"	1337	255
7	"TRUE"	2	7	3	160	"Agree"	21	"Friday"	17	240
8	"FALSE"	4	6	20	176	"Strongly...	19	"Tuesday"	9	270
9	"TRUE"	4	7	0	165	"Strongly...	21	"Thursday"	8	230
10	"TRUE"	2	6	0	160	"Agree"	20	"Friday"	28	235
11	"FALSE"	2	7	28	167.8000	"Strongly...	23	"Friday"	21	265
12	"FALSE"	3	10	1	168	"Agree"	22	"Thursday"	10	275
13	"FALSE"	2	7	0	177	"Neither ...	23	"Saturday"	1	275
14	"FALSE"	4	6	4	174	"Strongly...	21	"Friday"	7	280
15	"FALSE"	1	7	8	175	"Disagree"	25	"Friday"	4	265
16	"TRUE"	170	6	1	170	"Strongly...	22	"Friday"	32	245
17	"FALSE"	4	8	2	174	"Neither ...	18	"Friday"	14	265
18	"TRUE"	0	8	30	163	"Strongly...	19	"Friday"	8	250
19	"FALSE"	3	10	12	170	"Strongly...	21	"Friday"	97	260
20	"TRUE"	3	6	0.5000	160	"Strongly...	19	"Friday"	6	250
21	"TRUE"	1.5000	6	0	165	"Strongly...	21	"Wednes...	4	245
22	"FALSE"	2	7	20	165	"Agree"	24	"Thursday"	3	265
23	"FALSE"	1	9	14	180	"Agree"	20	"Friday"	7	260
24	"FALSE"	6	8	4	173	"Agree"	24	"Friday"	7	270

- Your data from 58 responders

Descriptive statistics

- In order to make sense of this data, we need ways to summarize and visualize it
- Summarizing and visualizing variables and relationships between two variables is often as descriptive statistics
- Type of summary statistics and visualization methods depend on the type of variable(s) being analyzed (categorical or quantitative)

What variable types are there?

CognitiveSystems

- Nominal
  - categories
- Ordinal (magnitude)
  - also called rank-ordered variables
- Ratio (magnitude, interval, rational zero)
  - reaction time data, accuracy, confidence

Categorical

Numerical

One Categorical Variable

- Display the number or proportion of cases that fall in each category

- “What is your favorite day of the week?”

Frequency Table

- A frequency table shows the number of cases that fall in each category:

{'Friday' }	{[33]}
{'Monday' }	{[1]}
{'Saturday' }	{[10]}
{'Tuesday' }	{[2]}
{'Thursday' }	{[7]}
{'Wednesday'}	{[5]}

Proportion

- The sample proportion of students in each category is

$$\hat{p} = \frac{\text{number of cases in category}}{\text{total number of cases}}$$

Proportion

- The sample proportion of students in this class who prefer Friday is

$$p=33/58=0.57$$

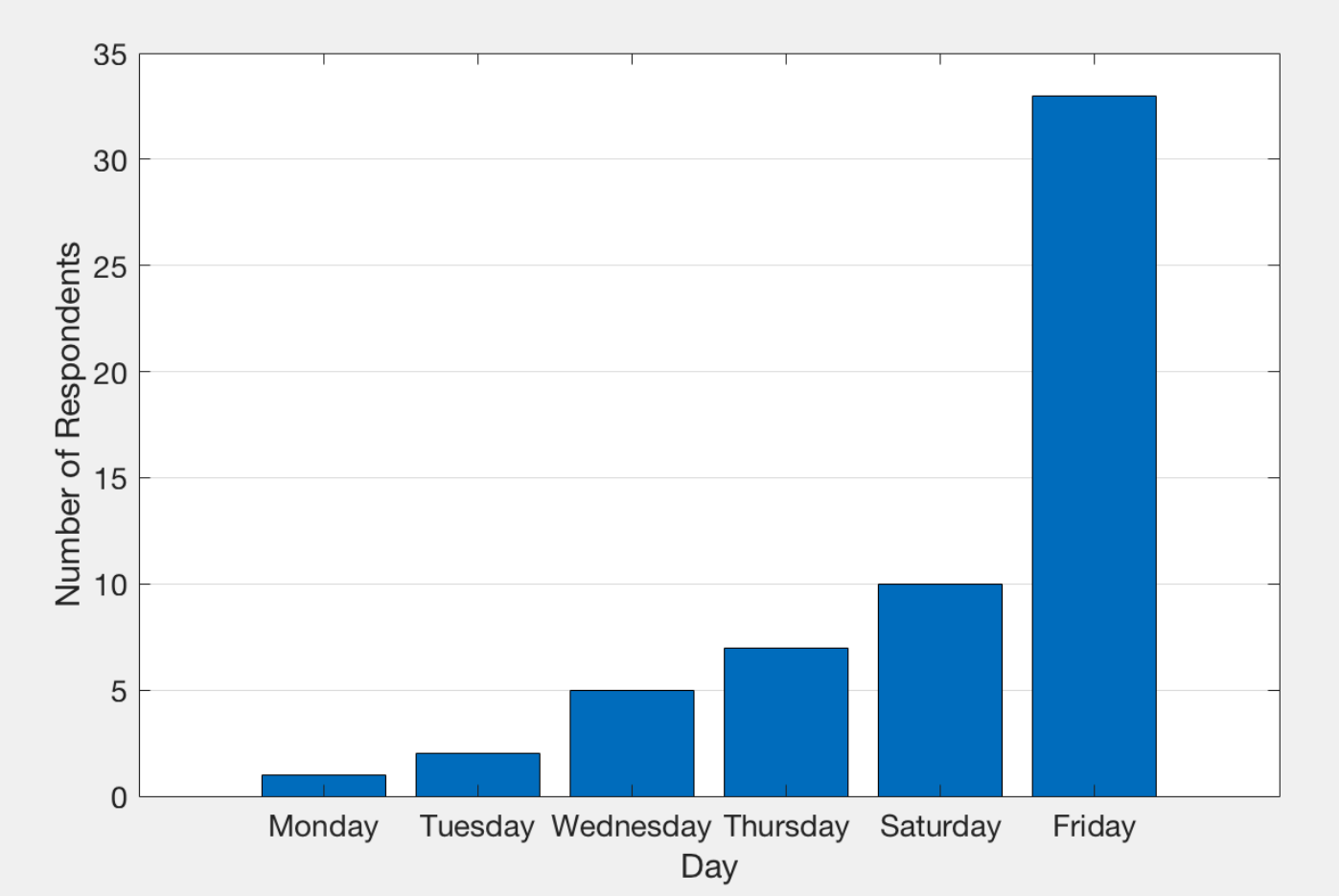
- Proportion and percent can be used interchangeably: 0.57 or 57%

```
{'Friday'   }    {[33]}    {[56.8966]}
{'Monday'   }    {[ 1]}    {[ 1.7241]}
{'Saturday' }    {[10]}    {[17.2414]}
{'Tuesday'  }    {[ 2]}    {[ 3.4483]}
{'Thursday' }    {[ 7]}    {[12.0690]}
{'Wednesday'}    {[ 5]}    {[ 8.6207]}
```

Careful with percentages!!!

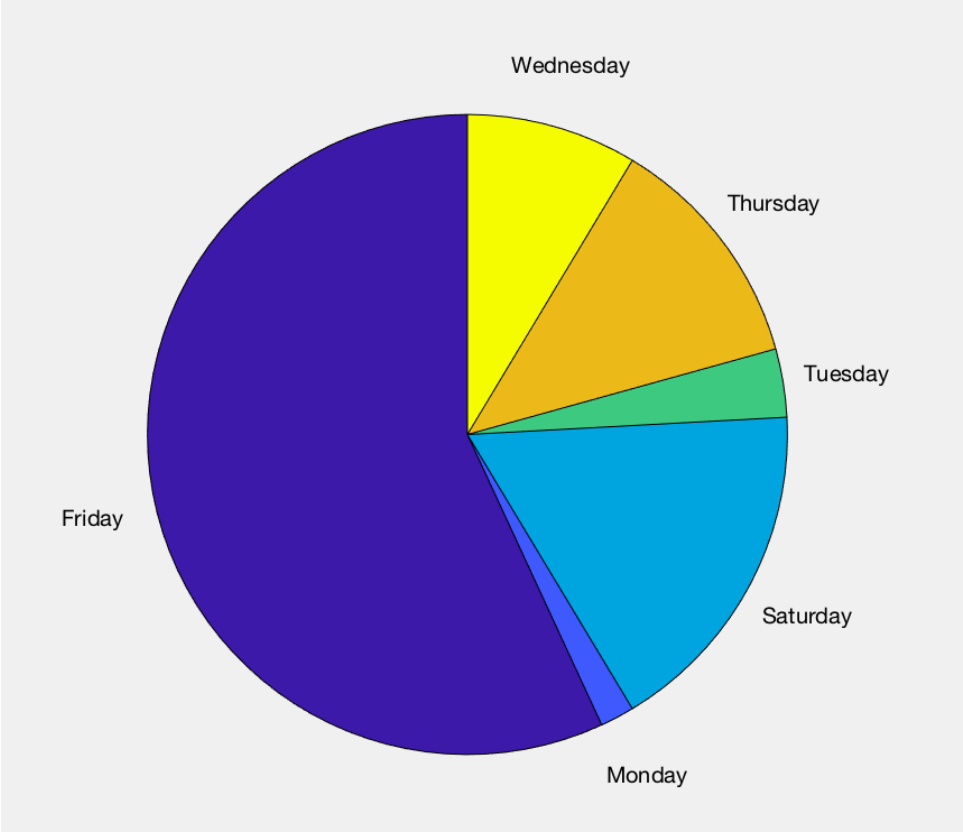
Bar Chart/Plot/Graph

- In a barplot, the height of the bar corresponds to the number of cases falling in each category

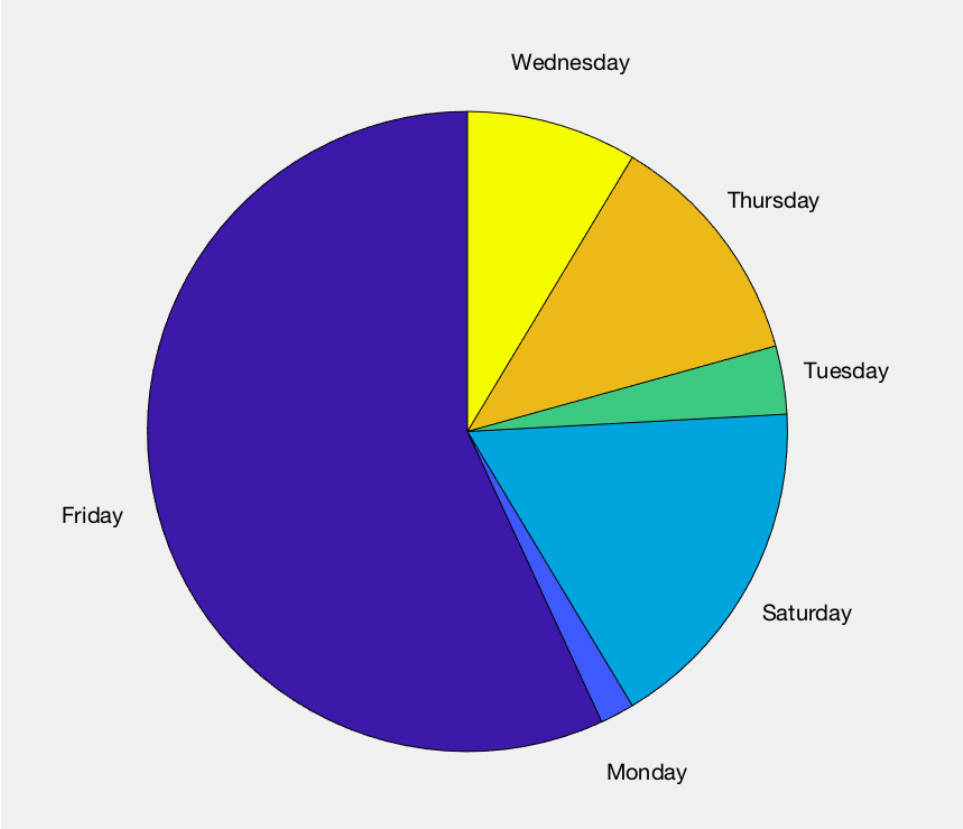


Pie Chart

- In a pie chart, the relative area of each slice of the pie corresponds to the proportion in each category



Pie Chart



Summary: One Categorical Variable

- Summary Statistics
  - Proportion
  - Frequency table
- Visualization
  - Barplot

Two Categorical Variables

- Given two categorical variables, we want to look at the relationship between them

- Favorite day
- Gender

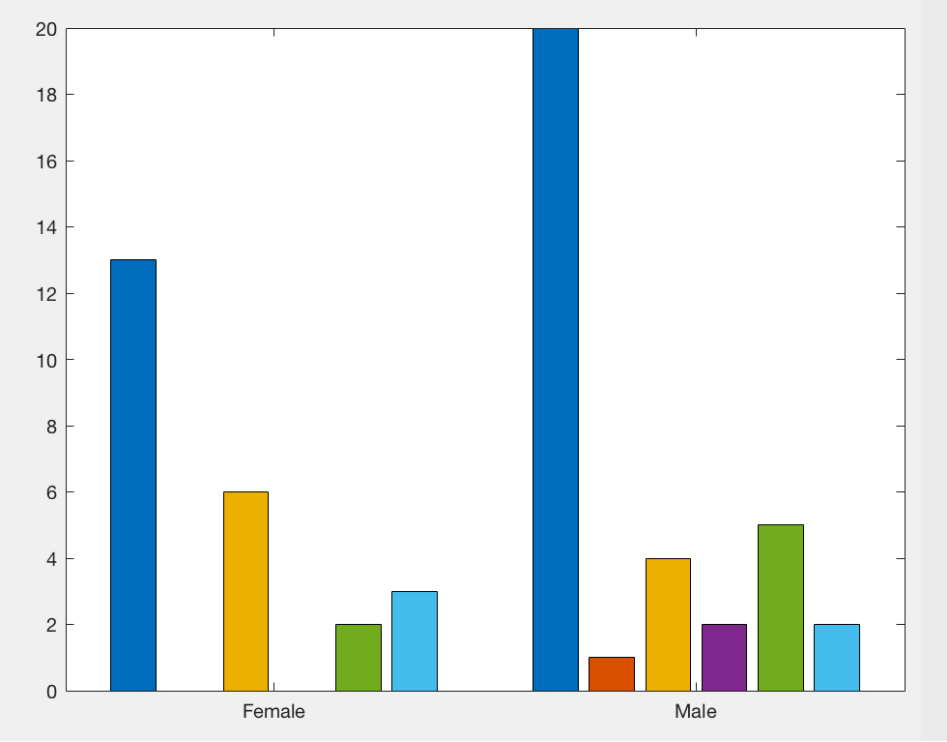
Two-Way Table

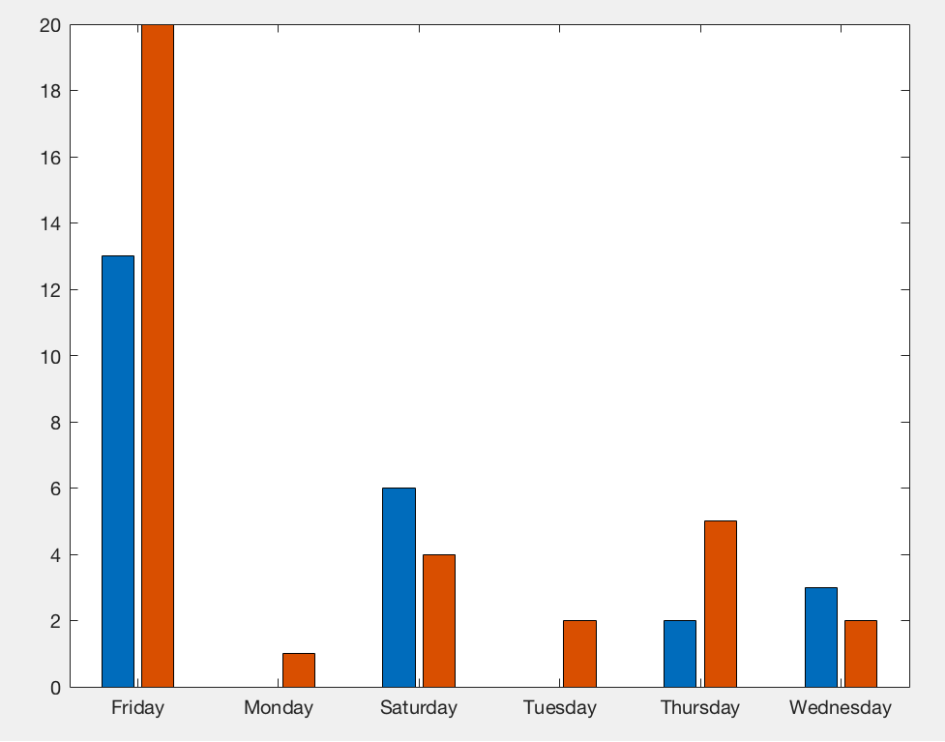
- It doesn't matter which variable is displayed in the rows and which in the columns

	Friday	Monday	Saturday	Tuesday	Thursday	Wednesday
Female	13	0	6	0	2	3
Male	20	1	4	2	5	2

Side-by-Side Bar Chart

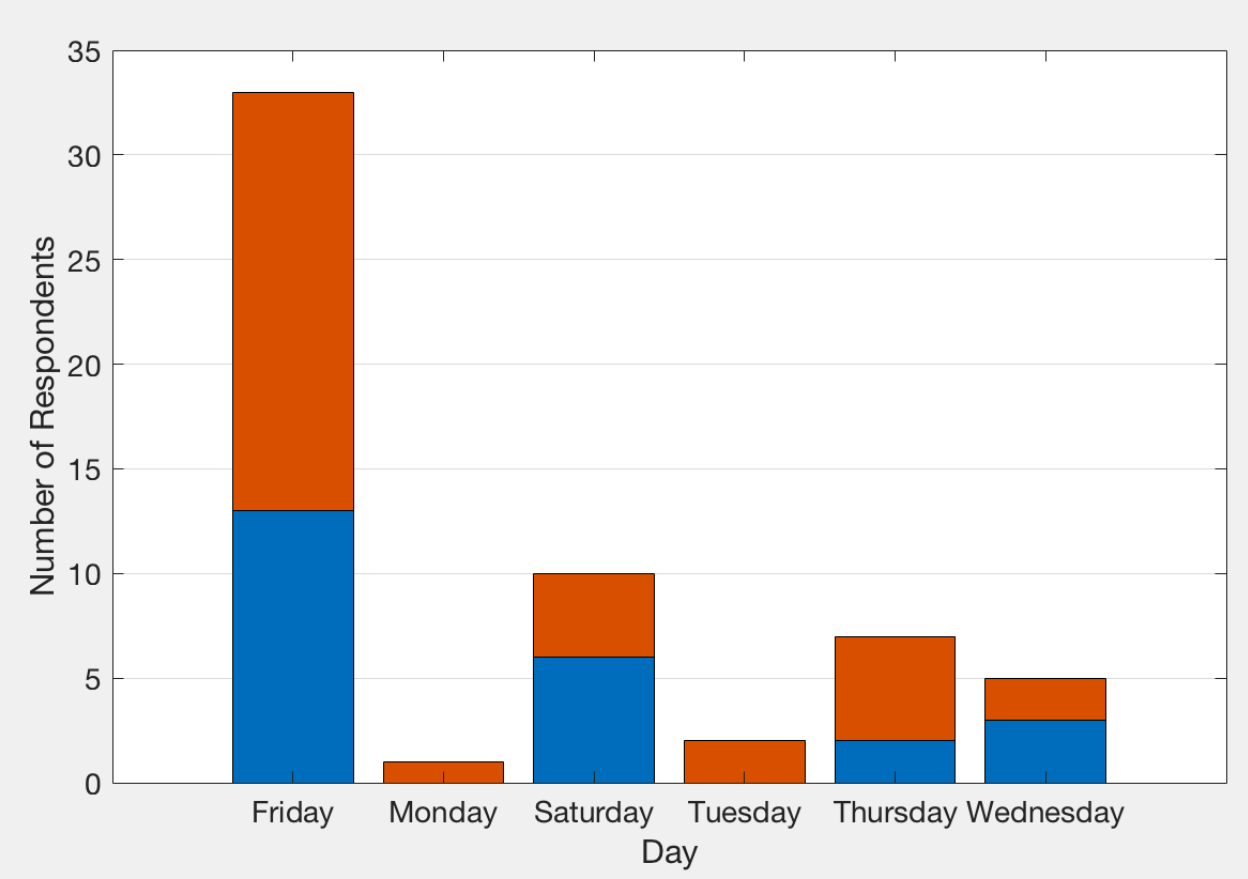
- The height of each bar is the number of the corresponding cell in the two-way table





Stacked Bar Chart

- The height of each bar is the number of the sums of the corresponding cells for all elements in the two-way table



Summary: Two Categorical Variables

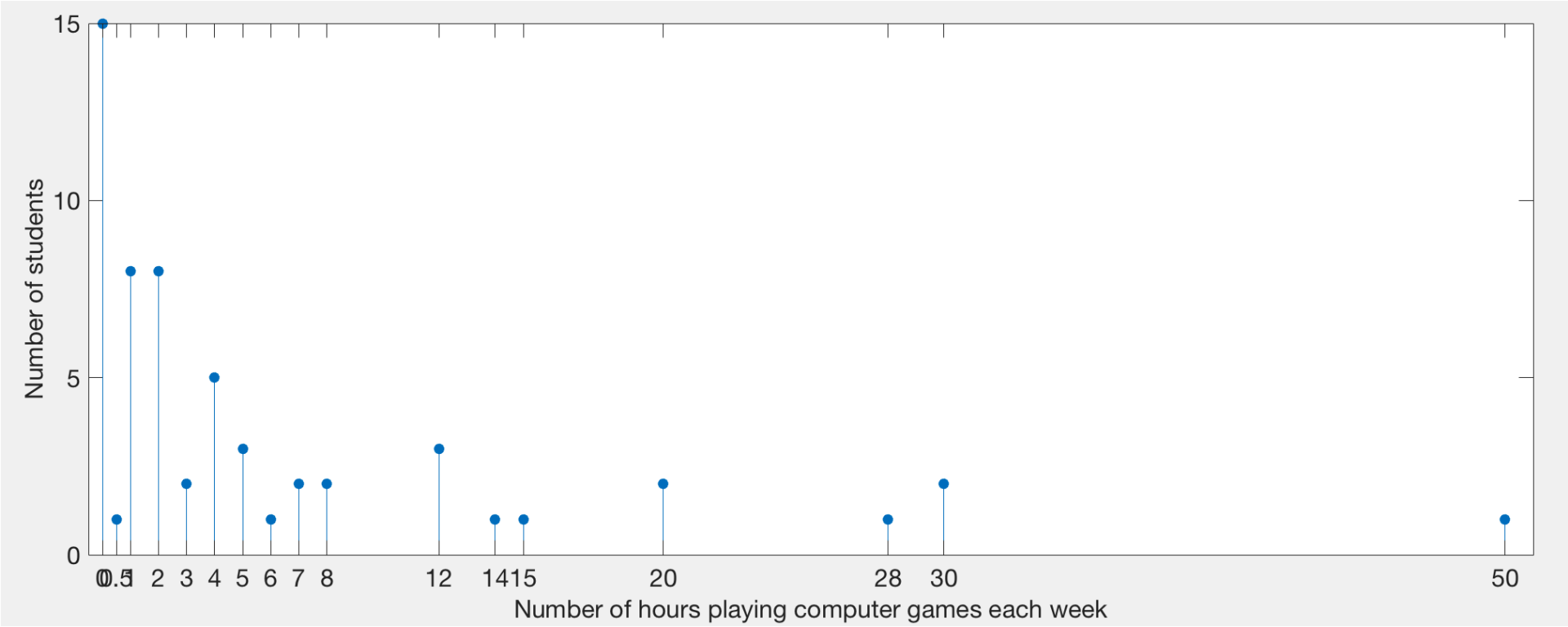
- Summary Statistics
  - Two-way table
- Visualization
  - Side-by-side bar chart
  - Stacked bar chart

One Quantitative Variable

- We'll look at how to analyze a quantitative variable such as
  - Average hours of playing computer games
  - Average hours of sleep per night

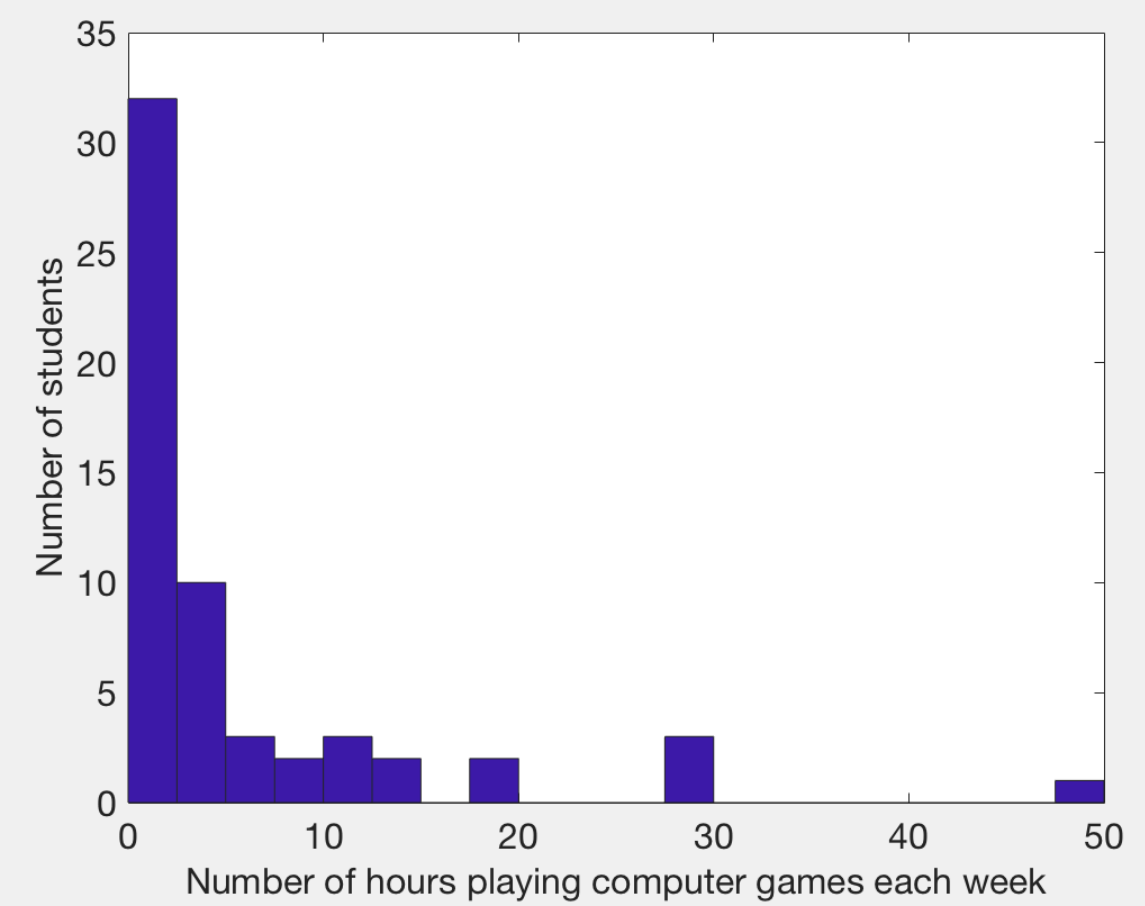
Full histogram

- In a full histogram, every case is represented, and the number of cases for each case are counted and tallied as bar heights.



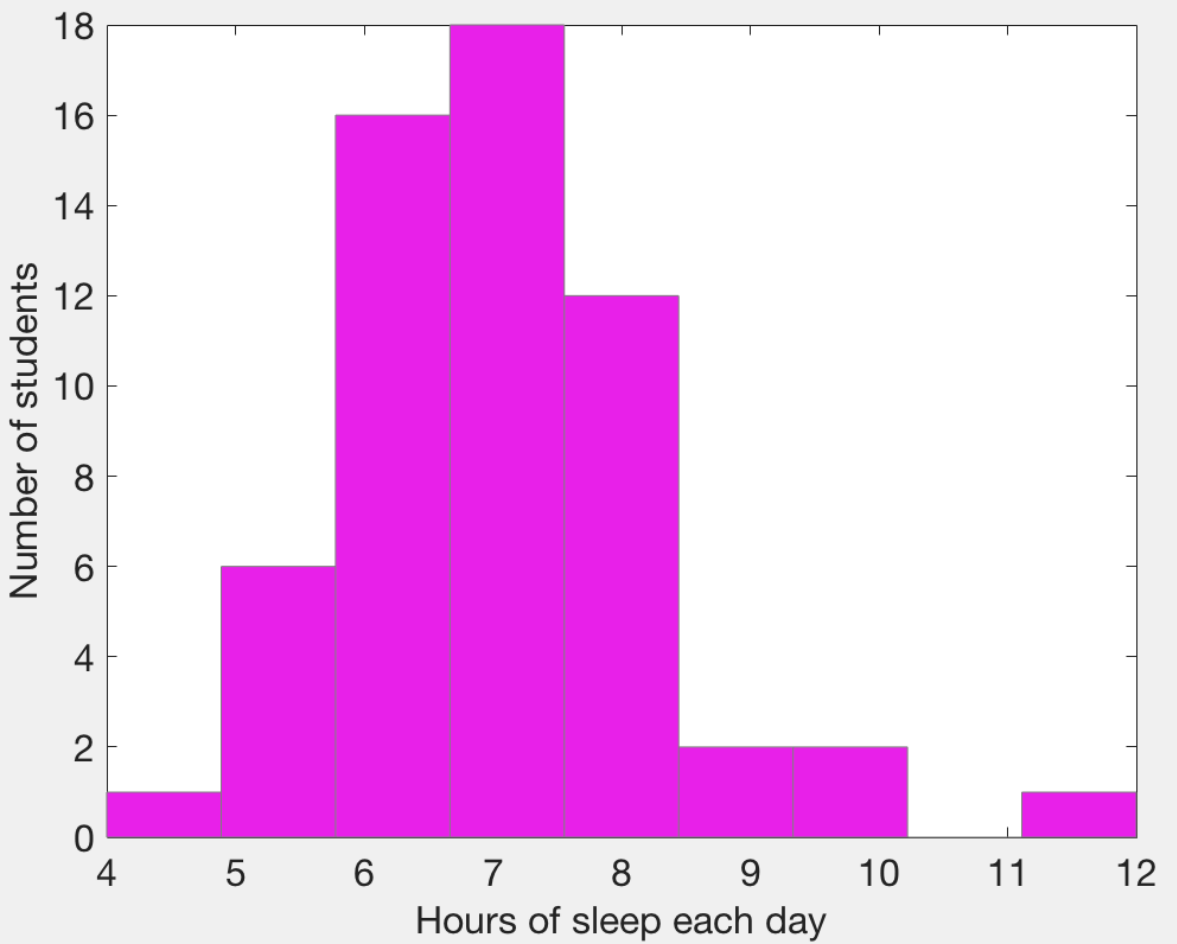
Histogram

- The height of the each bar corresponds to the number of cases within that range of the variable



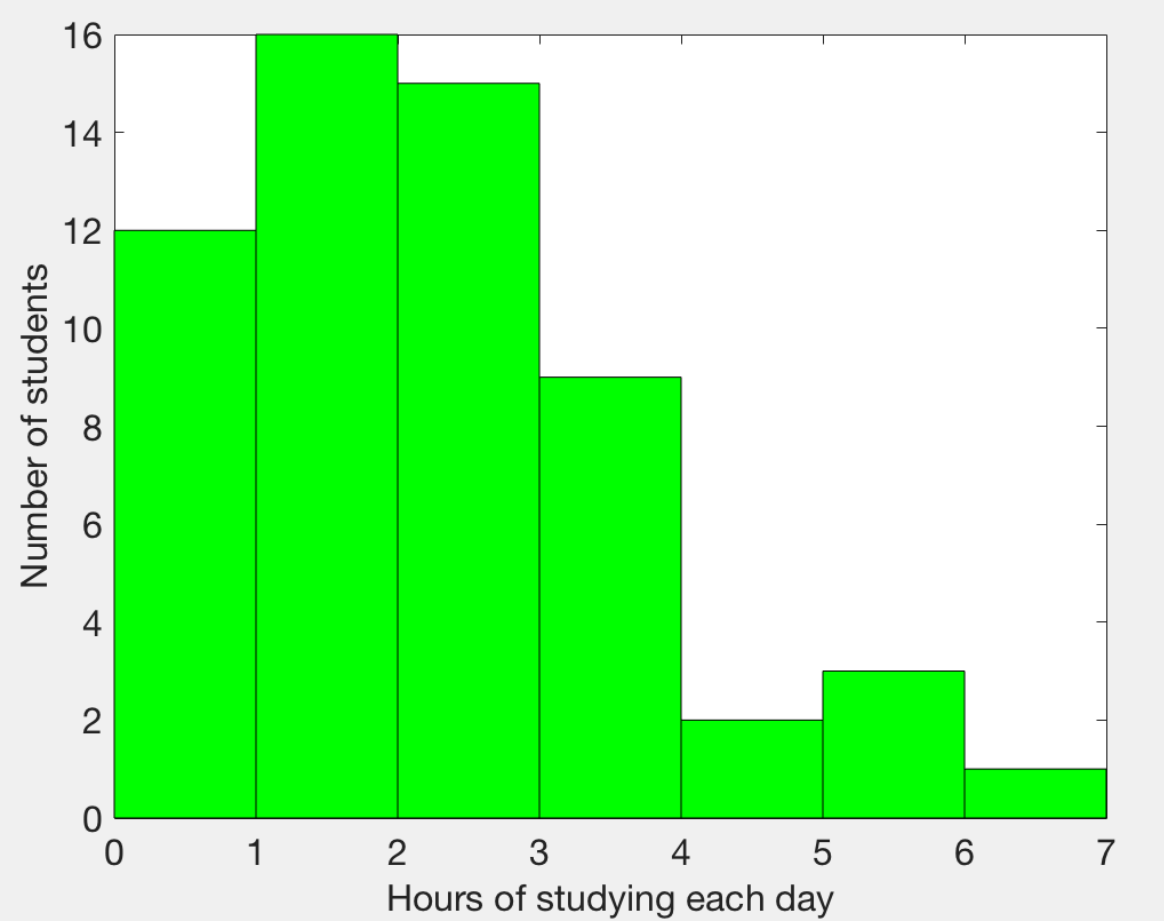
Histogram

- The height of the each bar corresponds to the number of cases within that range of the variable



Histogram

- The height of the each bar corresponds to the number of cases within that range of the variable

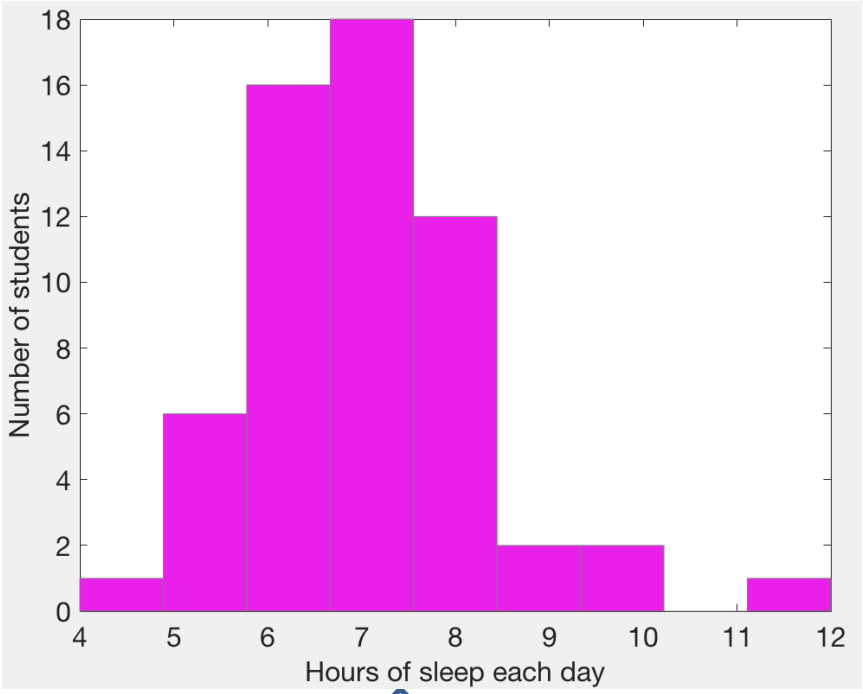


Histogram

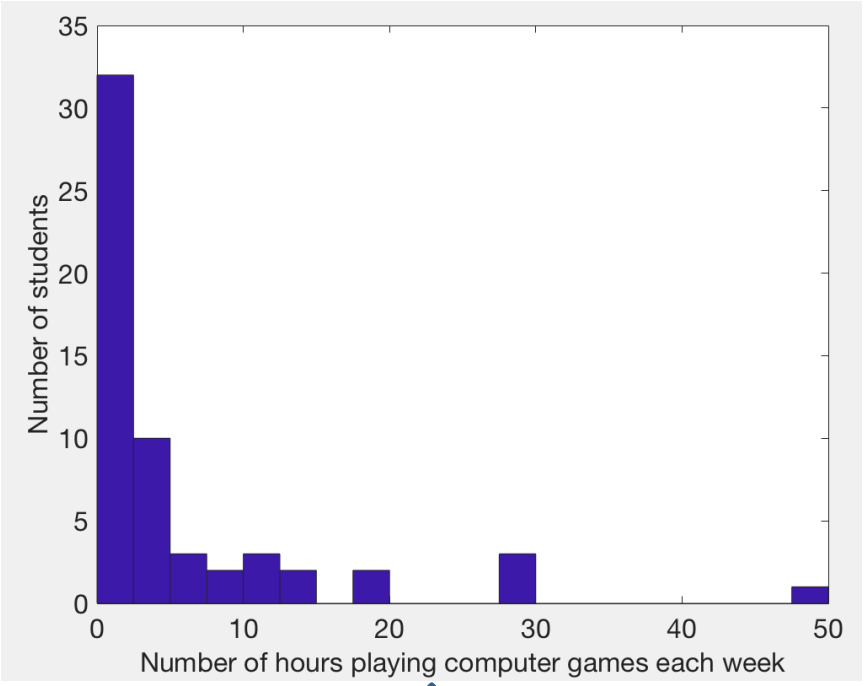
- Although they look similar, a histogram is not the same as a bar plot

- For a **categorical** variable, the number of bars equals the number of categories, and the number in each category is fixed
- For a **quantitative** variable, the number of bars in a histogram is up to a parameter (number of bins or ranges), and the appearance can differ with different number of bars

Shape

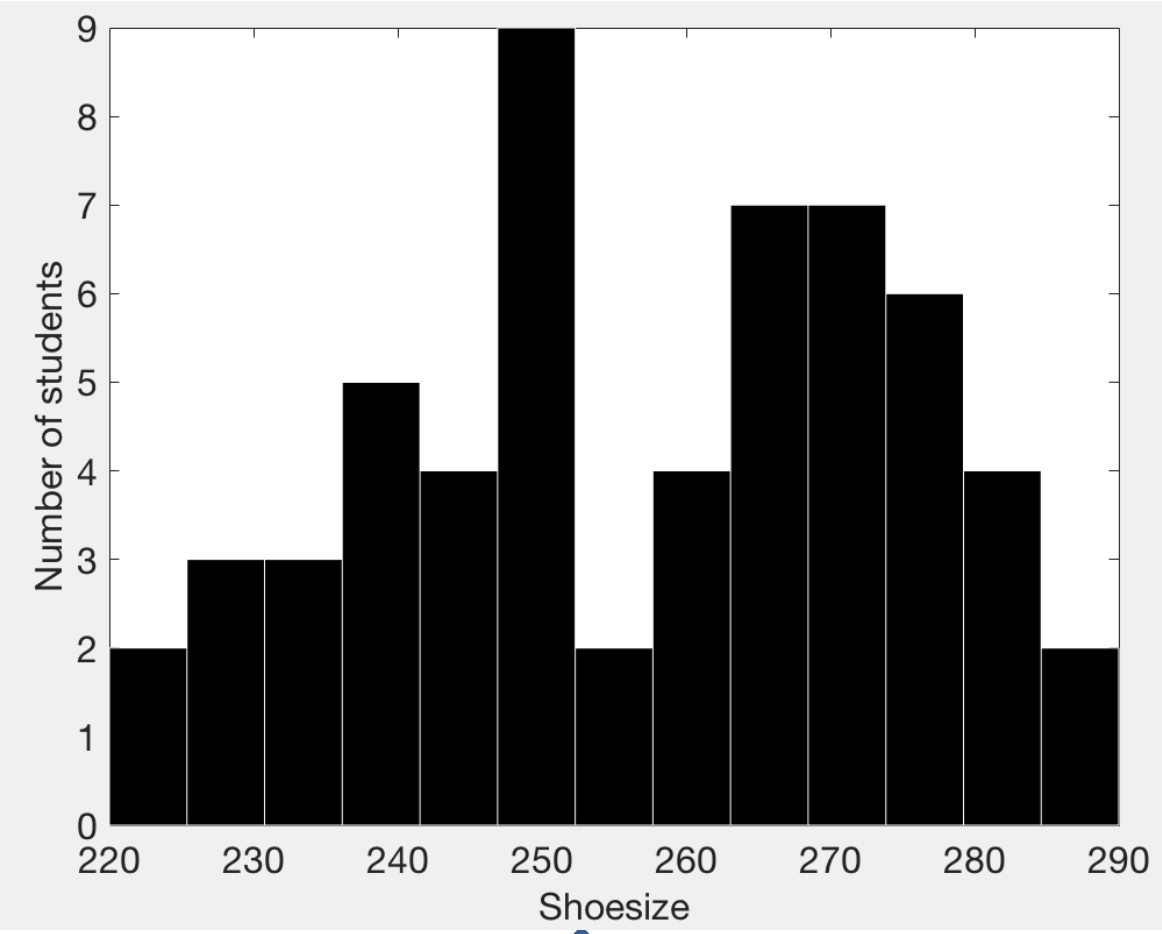


Symmetric



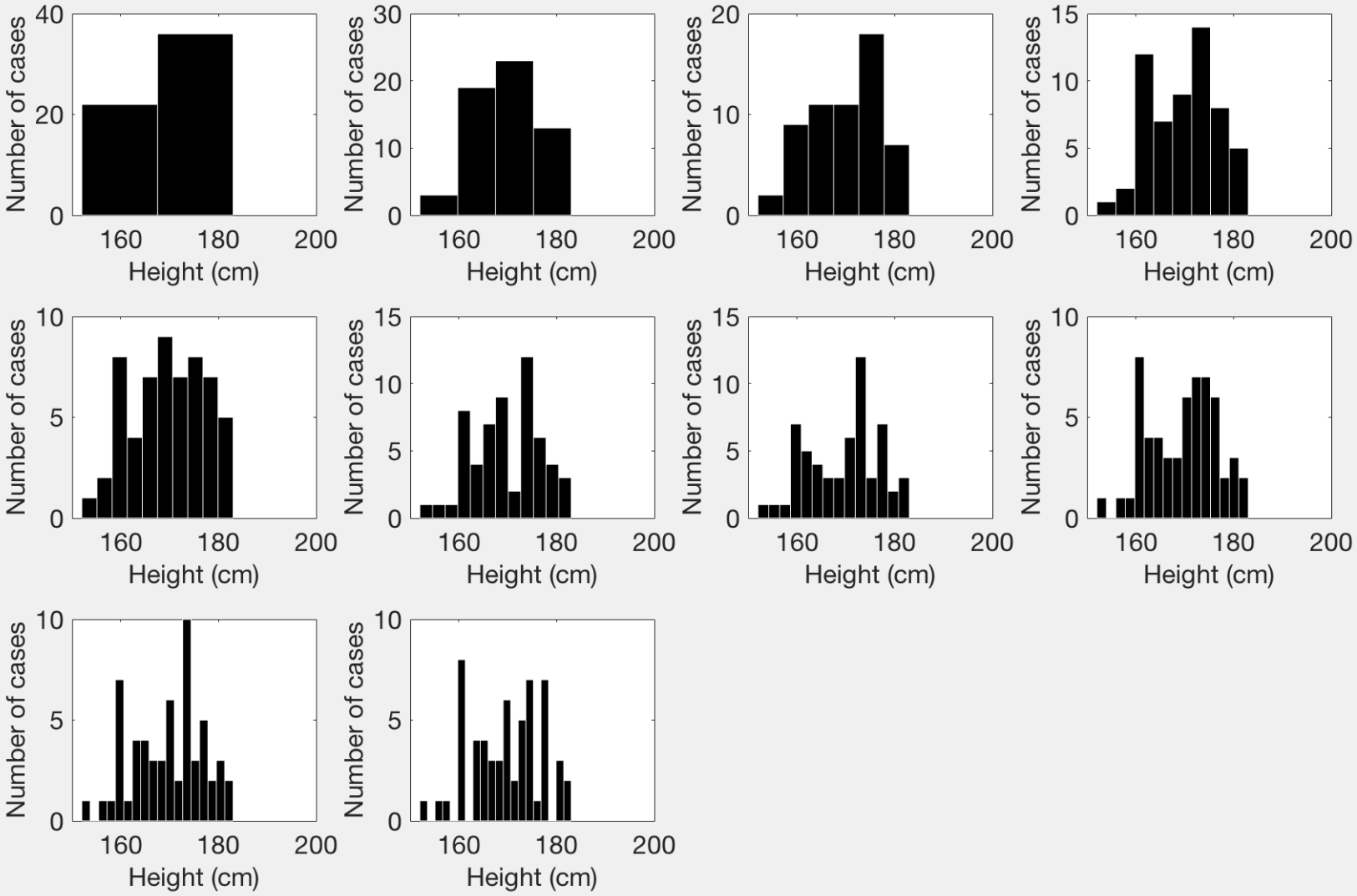
Right-Skewed

Shape



Bi-modal

Shape as a function of bins



How to characterize data

- Now that we have talked about the shape of distributions, let us try to characterize and summarize them with numbers
- Specifically, how to characterize the distribution with one or more numbers!

Mean

- The sample mean is the average, and is computed by adding up all the numbers and dividing by the number of cases

$$\text{Sample Mean: } \bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Mean

- The sample mean is the average, and is computed by adding up all the numbers and dividing by the number of cases

$$\text{Sample Mean: } \bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Age	Height	ShoeSize	Sleep	Study	Games
21.655172	169.591379	256.982759	6.931034	2.693103	5.939655

Median

- The sample median is the middle value when the data is ordered
- If there are an even number of values, the median is the average of the two middle values
- The sample median is denoted as m

Age	Height	ShoeSize	Sleep	Study	Games
21.000000	170.000000	260.000000	7.000000	3.000000	2.000000

Mode

- The mode is the most often occurring value

- The mode is always a “member” of the data

Age	Height	ShoeSize	Sleep	Study	Games
21.000000	160.000000	250.000000	7.000000	3.000000	0.000000

Mean versus Median versus Mode

- These are measures of **“central tendency”**
- For some measures, the three values give very similar results
- For other measures, however, the values change rather dramatically

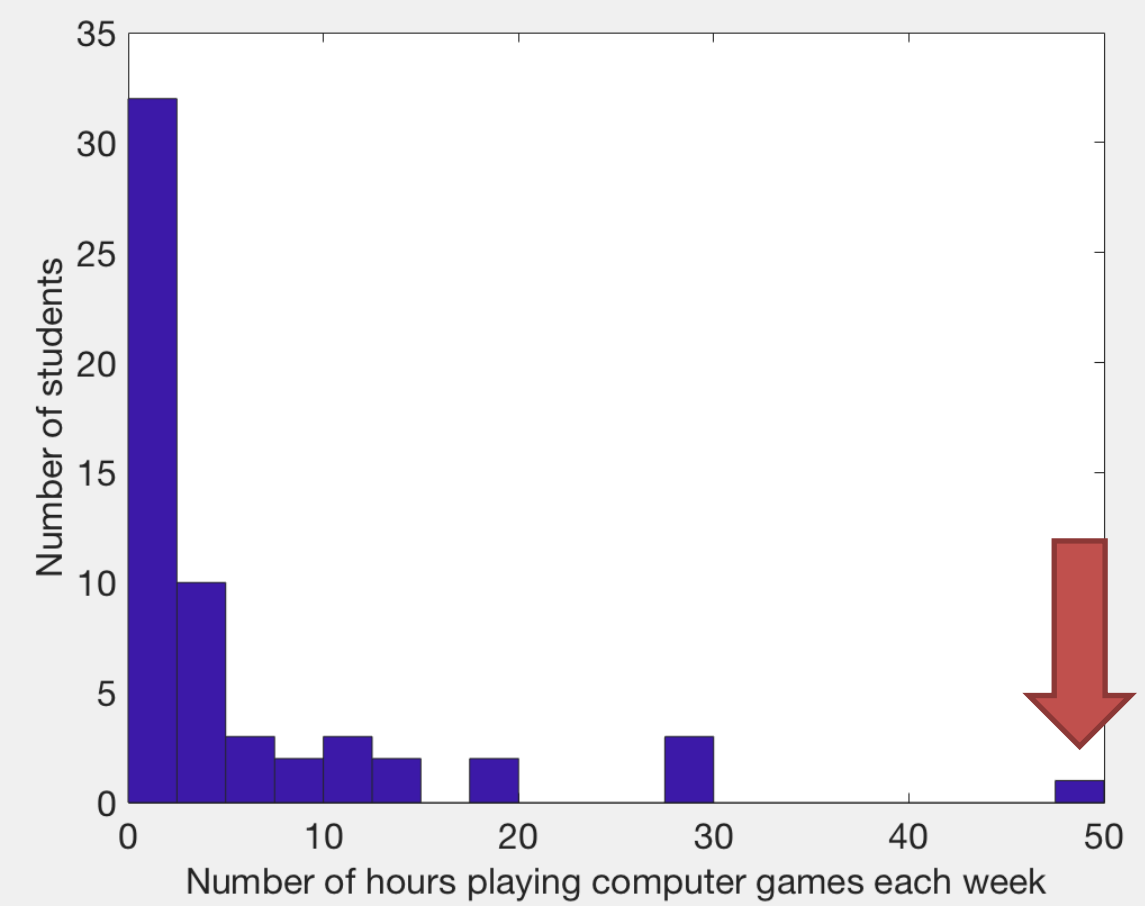
Age	Height	ShoeSize	Sleep	Study	Games
21.655172	169.591379	256.982759	6.931034	2.693103	5.939655

Age	Height	ShoeSize	Sleep	Study	Games
21.000000	170.000000	260.000000	7.000000	3.000000	2.000000

Age	Height	ShoeSize	Sleep	Study	Games
21.000000	160.000000	250.000000	7.000000	3.000000	0.000000

Outliers

- An outlier is a value that is notably different from the other values



Resistance

- Average hours of computer games each day:

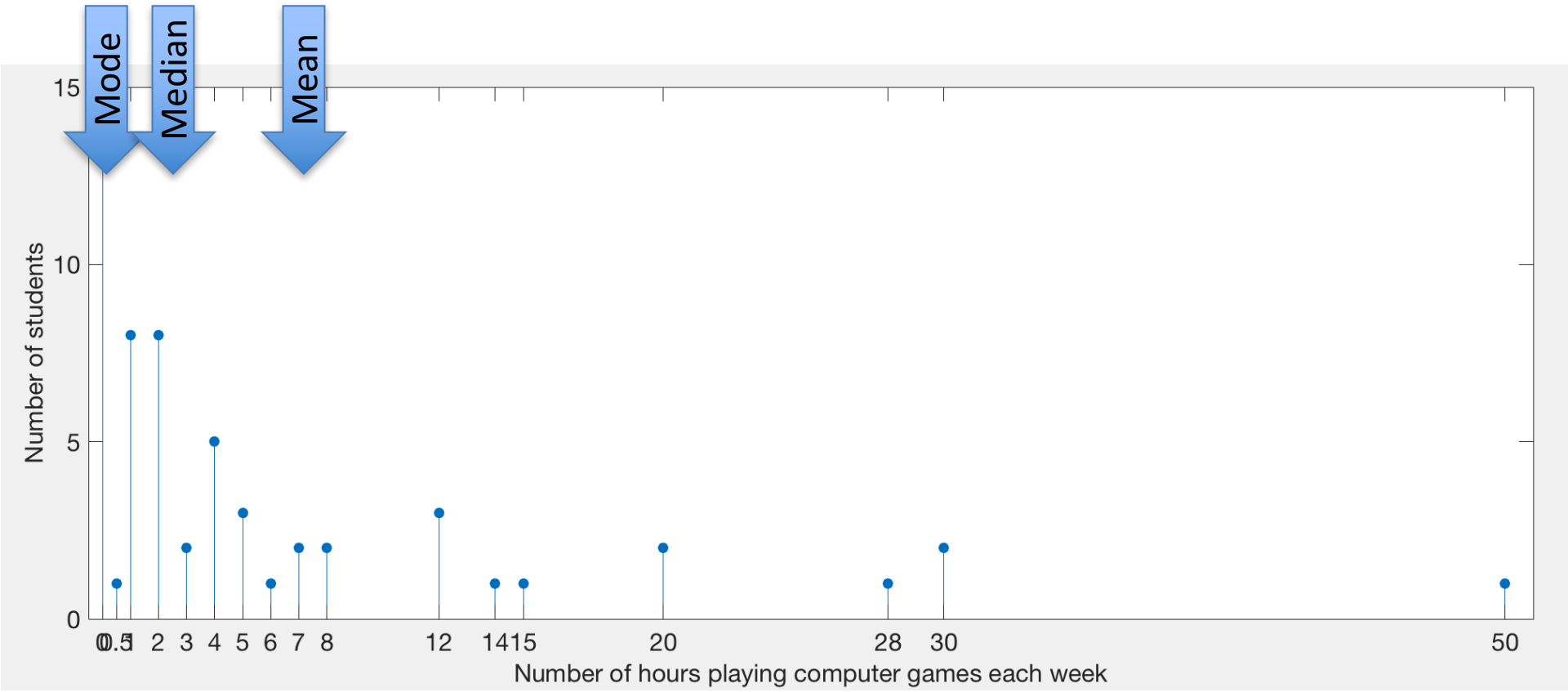
	Mean	Median	Mode
With Outlier	6.7	2	0
Without Outlier	5.9	2	0

Outliers

- When using statistics that are not resistant to outliers, stop and think about whether the outlier is a mistake
- If not, you have to decide whether the outlier is part of your population of interest or not
- Usually, for outliers that are not a mistake, it's best to run the analysis twice, once with the outlier(s) and once without, to see how much the outlier(s) are affecting the results

Robustness versus generalizability

- How to best characterize the data with one number??



Robustness versus generalizability

- Generalizability means to take into account all data when calculating a statistic

- This is just one of the many examples of the compromises we have to make when summarizing data!

Measures of spread

- Previously, we thought about characterizing data with just one value
  - measures of “central tendency”: mode, median, mean
- How can we add more information?
  - give information about variability in data!
  - measures of spread: standard deviation, IQR, range

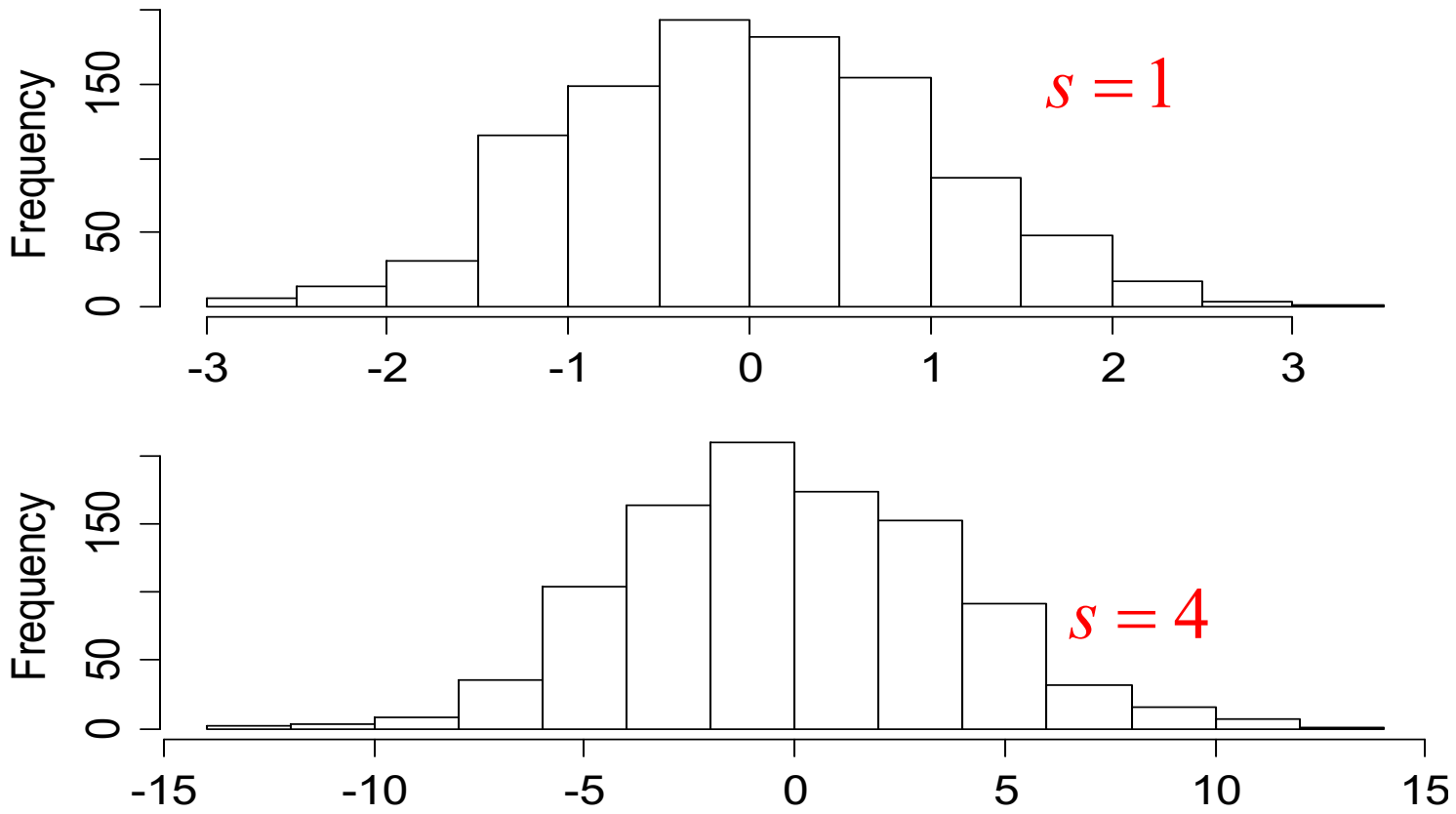
Standard Deviation

- The sample standard deviation, s, measures the spread of a distribution. The larger s is, the more spread out the distribution is

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

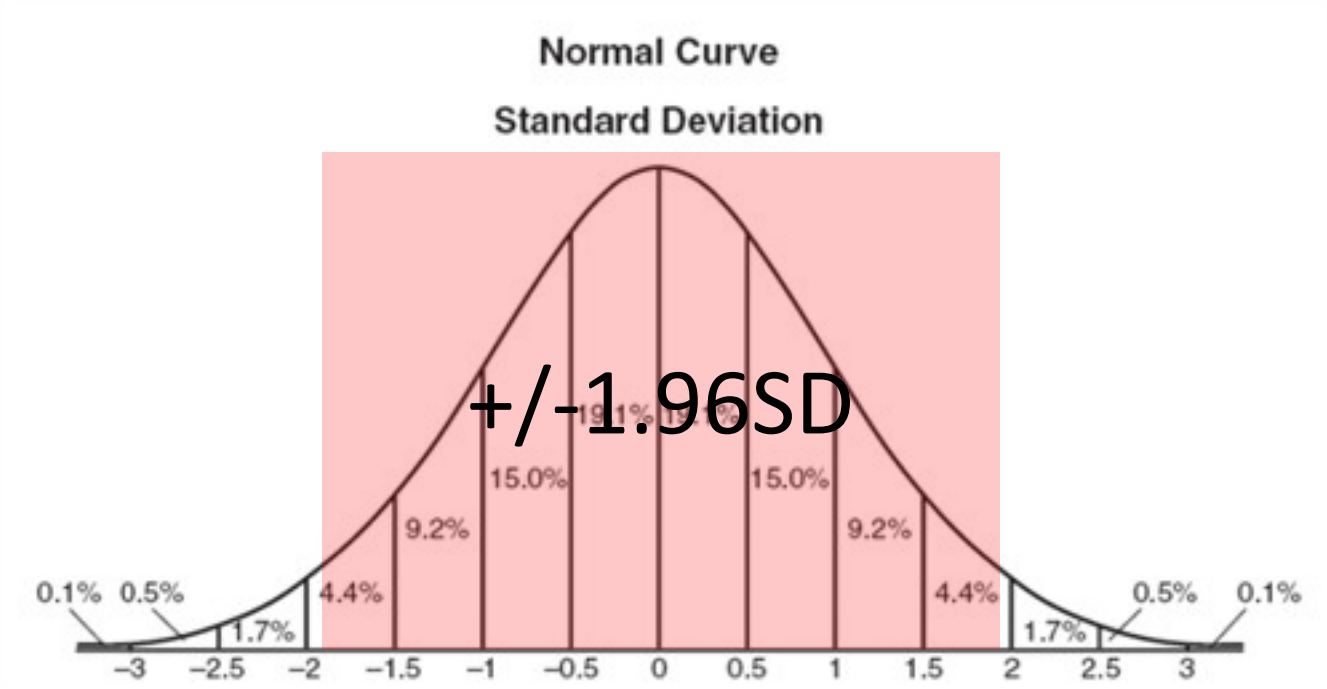
- Standard deviation is always ≥ 0.

Standard Deviation



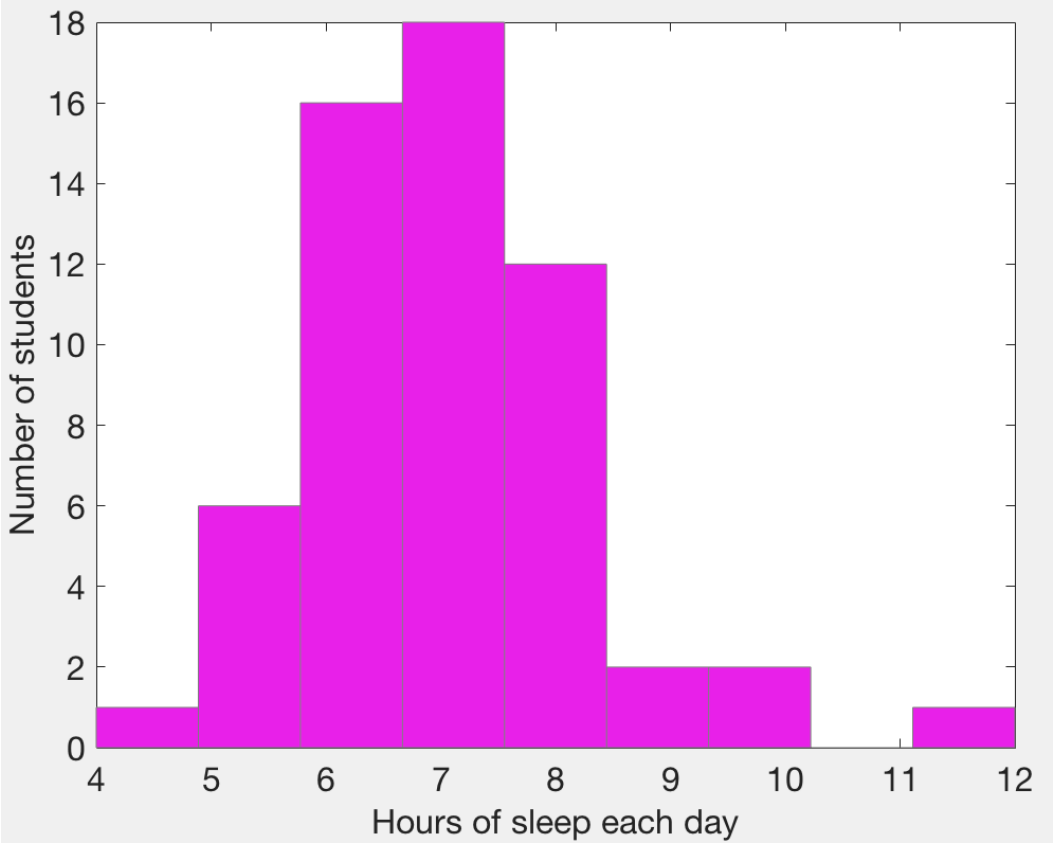
The 95% Rule

- If a distribution is symmetric and bell-shaped, then approximately 95% of the data values will lie within 1.96 standard deviations of the mean



The 95% Rule

- The standard deviation for hours of sleep per night is a bit larger than 1 hour (=1.4 hours)

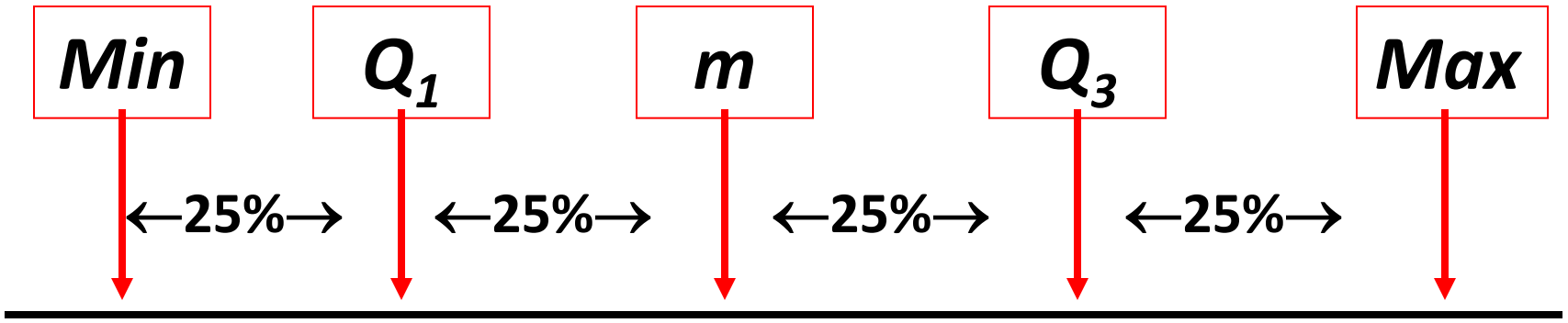


Other measures of spread

- Maximum = largest data value
- Minimum = smallest data value
- Quartiles:
- Q1 = median of the values below m.
- Q3 = median of the values above m.

Five Number Summary

- Five Number Summary:



Five Number Summary

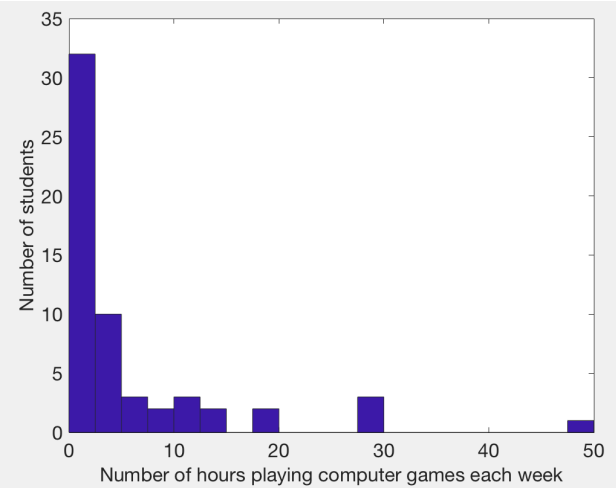
Studying

min	25%	Median	75%	Max
0	2.0000	3.0000	3.5000	7.0000

- The distribution of number of hours you study each day is

- (a) Symmetric
- (b) Right-skewed
- (c) Left-skewed
- (d) Impossible to tell

There is a much larger range between the center and the max than between the center and the min.



Measures of Spread

CognitiveSystems

- Range = Max – Min
- Interquartile Range (IQR) = Q3 – Q1

- Is the range resistant to outliers?
- Yes
- No

The range is only determined by the most extreme values, so it will be affected by outliers.

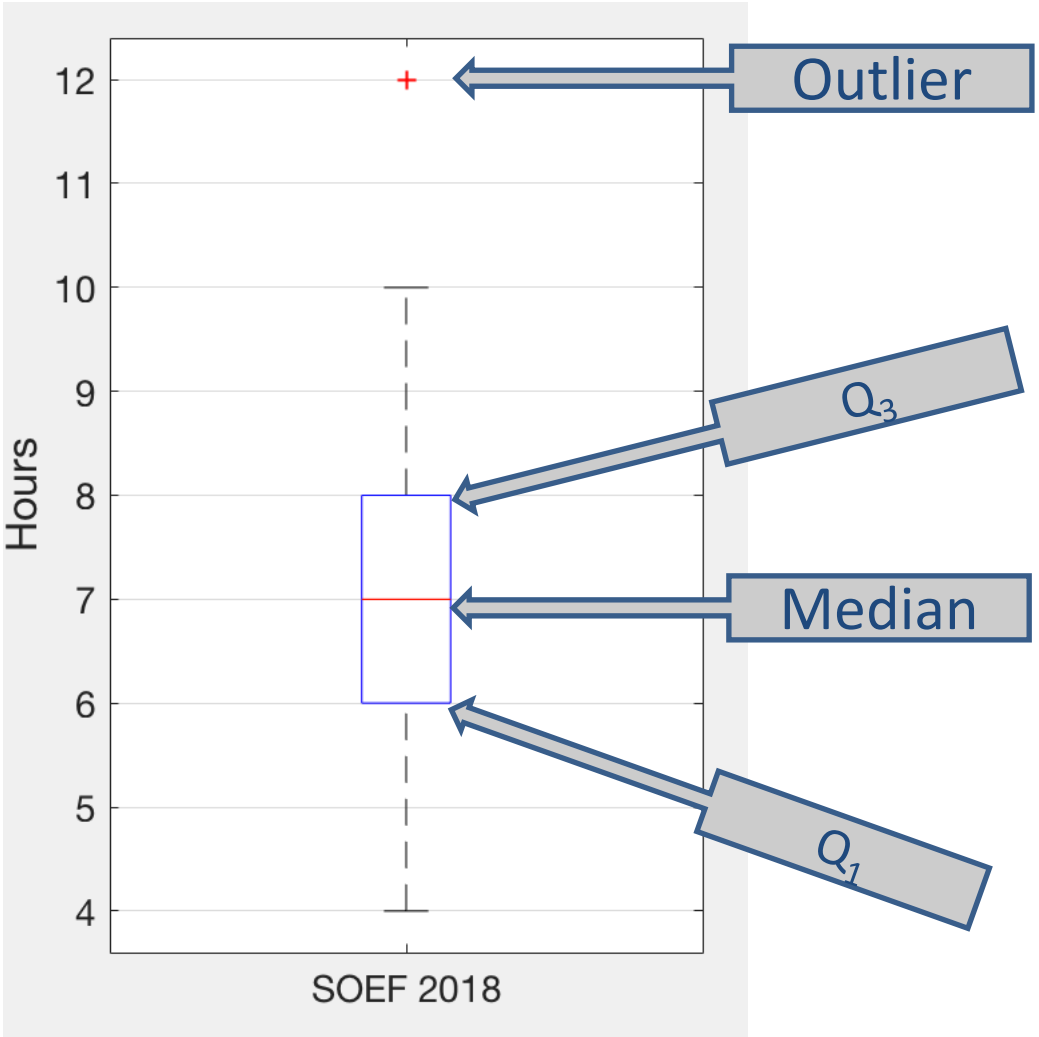
- Is the IQR resistant to outliers?
- Yes
- No

The IQR is not very affected by outliers, because it ignores the most extreme data (the +/- 25% maximum values).

Outliers

- Outliers can be informally identified by looking at a plot, but one rule of thumb for identifying outliers is data values more than 1.5 IQRs beyond the quartiles
- A data value is an outlier if it is
- Smaller than Q1 – 1.5(IQR)
- or
- Larger than Q3 + 1.5(IQR)

Boxplot



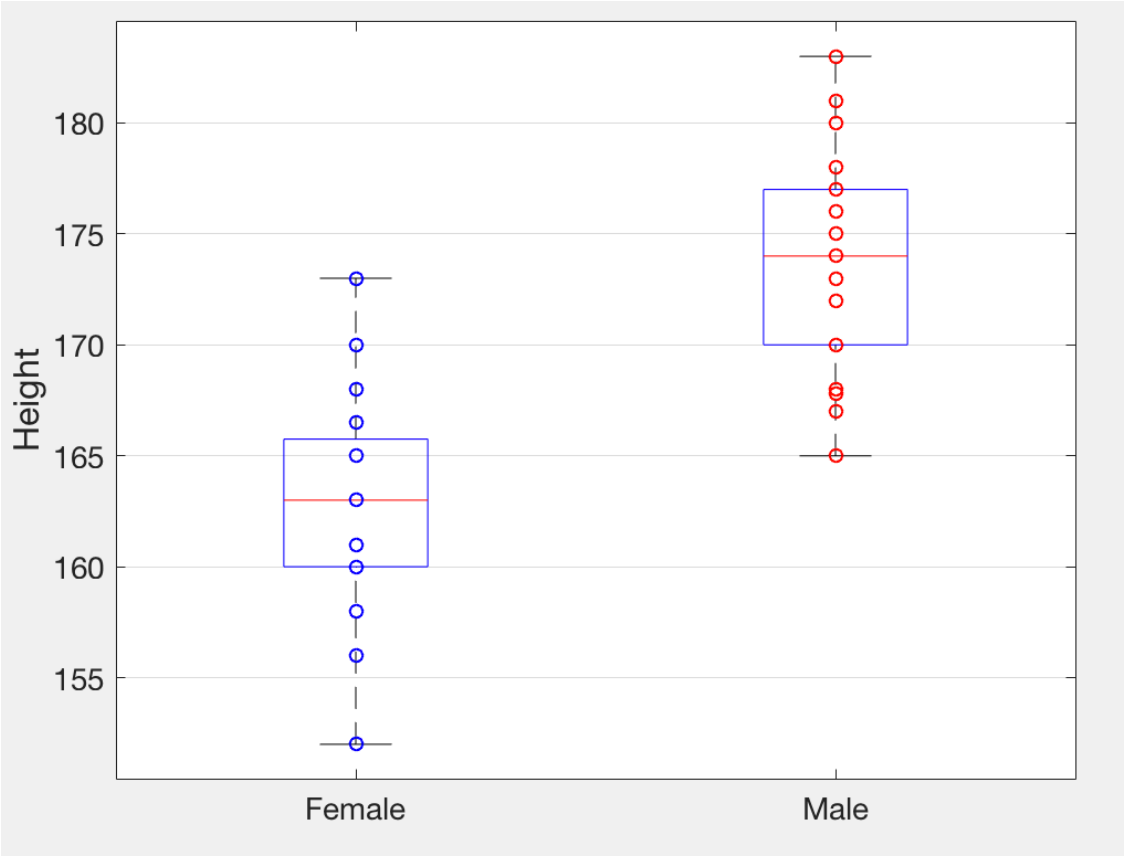
- Lines ("whiskers") extend from each quartile to the most extreme value that is not an outlier

Summary: One Quantitative Variable

- Summary Statistics
  - Center: mean, median, mode
  - Spread/Dispersion: standard deviation, range, IQR
  - Percentiles
  - 5 number summary
- Visualization
  - Histogram
  - Boxplot
- Other concepts
  - Shape: symmetric, skewed, bell-shaped, bi-modal
  - Outliers, resistance

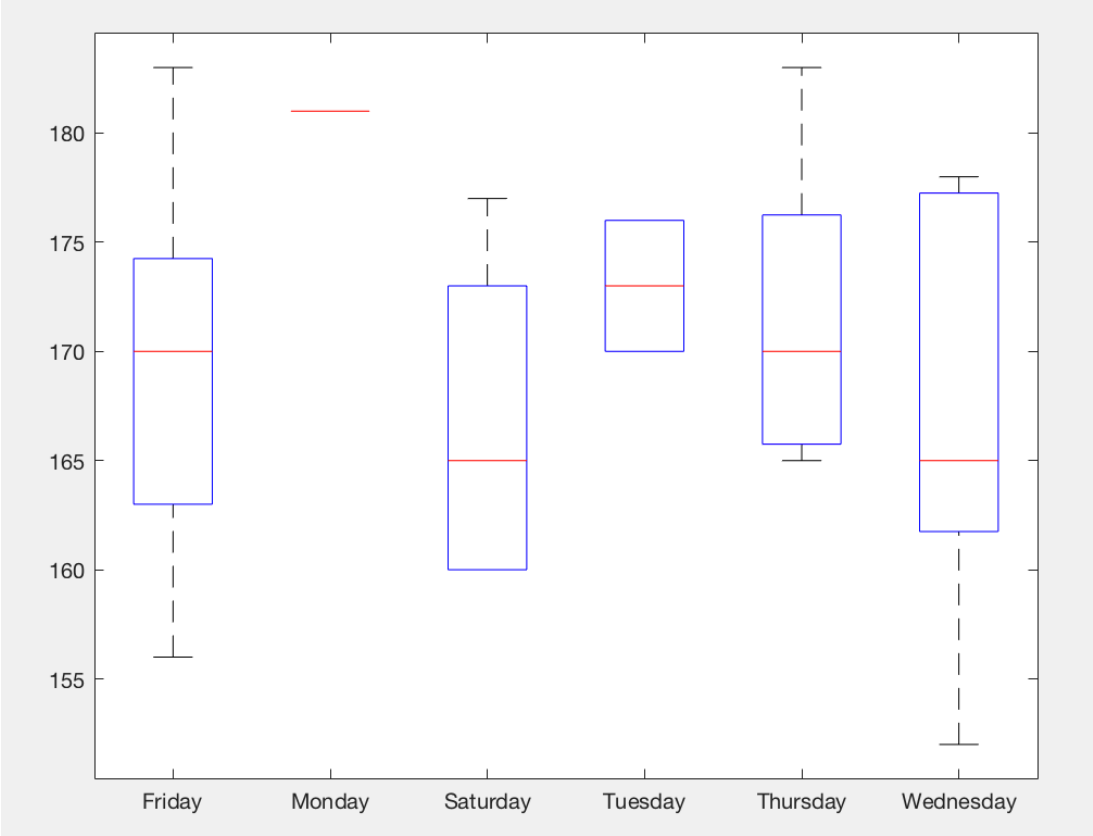
Quantitative and Categorical Relationships

- Boxplots are particularly useful for comparing distributions of a quantitative variable across different levels of a categorical variable



Quantitative and Categorical Relationships

- Boxplots are particularly useful for comparing distributions of a quantitative variable across different levels of a categorical variable

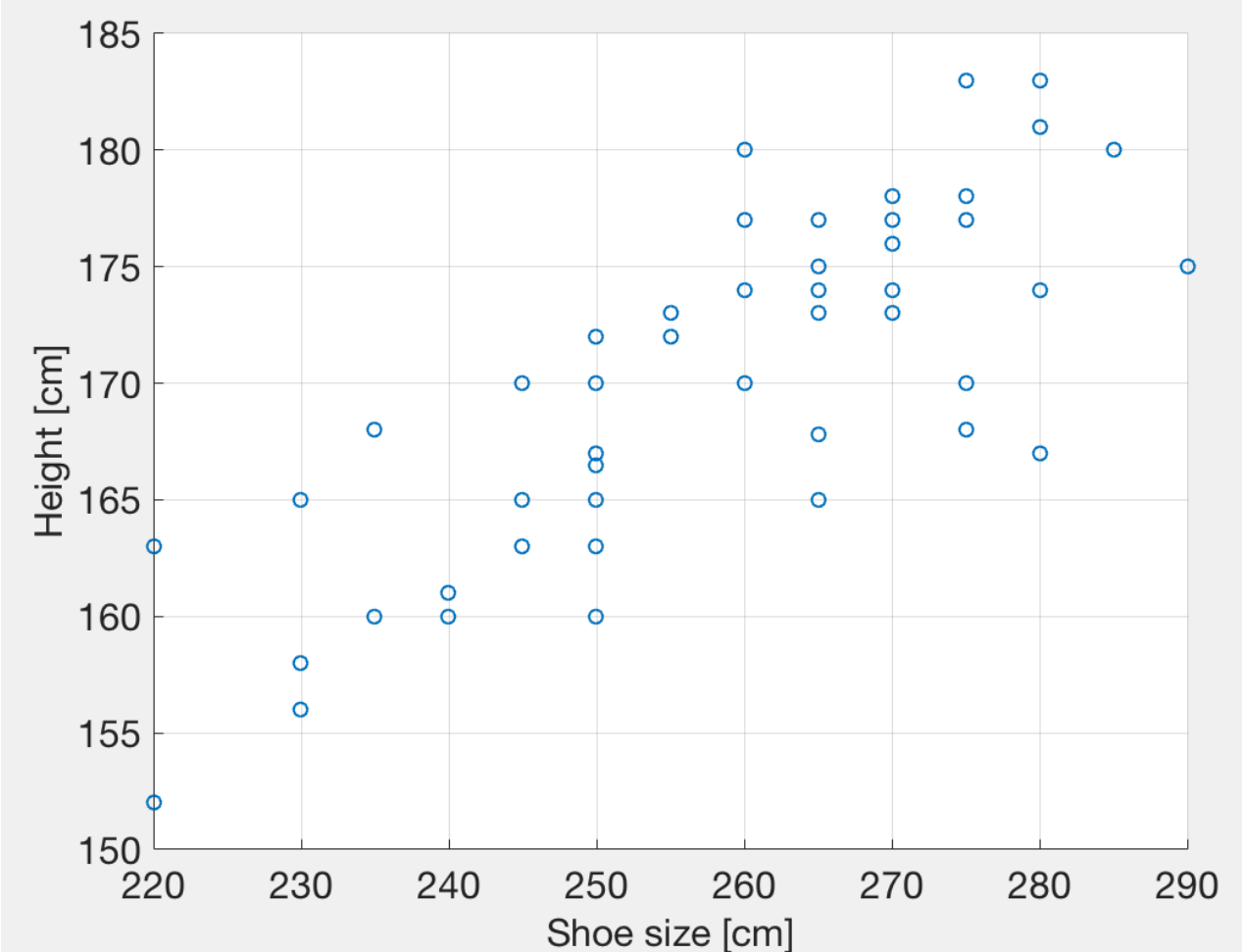


Summary: One Quantitative and One Categorical

- Summary Statistics
  - Any summary statistics for quantitative variables, broken down by each level of the categorical variable
- Visualization
  - Side-by-side boxplots

Scatterplot

- A scatterplot is a graph of the relationship between two quantitative variables. Each dot represents one case.



Direction of Association

- A positive association means that values of one variable tend to be higher when values of the other variable are higher

- A negative association means that values of one variable tend to be lower when values of the other variable are higher

- Two variables are not associated if knowing the value of one variable does not give you any information about the value of the other variable

Pearson correlation

- The Pearson [sample] correlation, r, measures the strength and direction of **linear** association between two quantitative variables

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_X} \right) \left(\frac{y_i - \bar{y}}{s_Y} \right)$$

- sX : sample standard deviation of X
- sY : sample standard deviation of Y

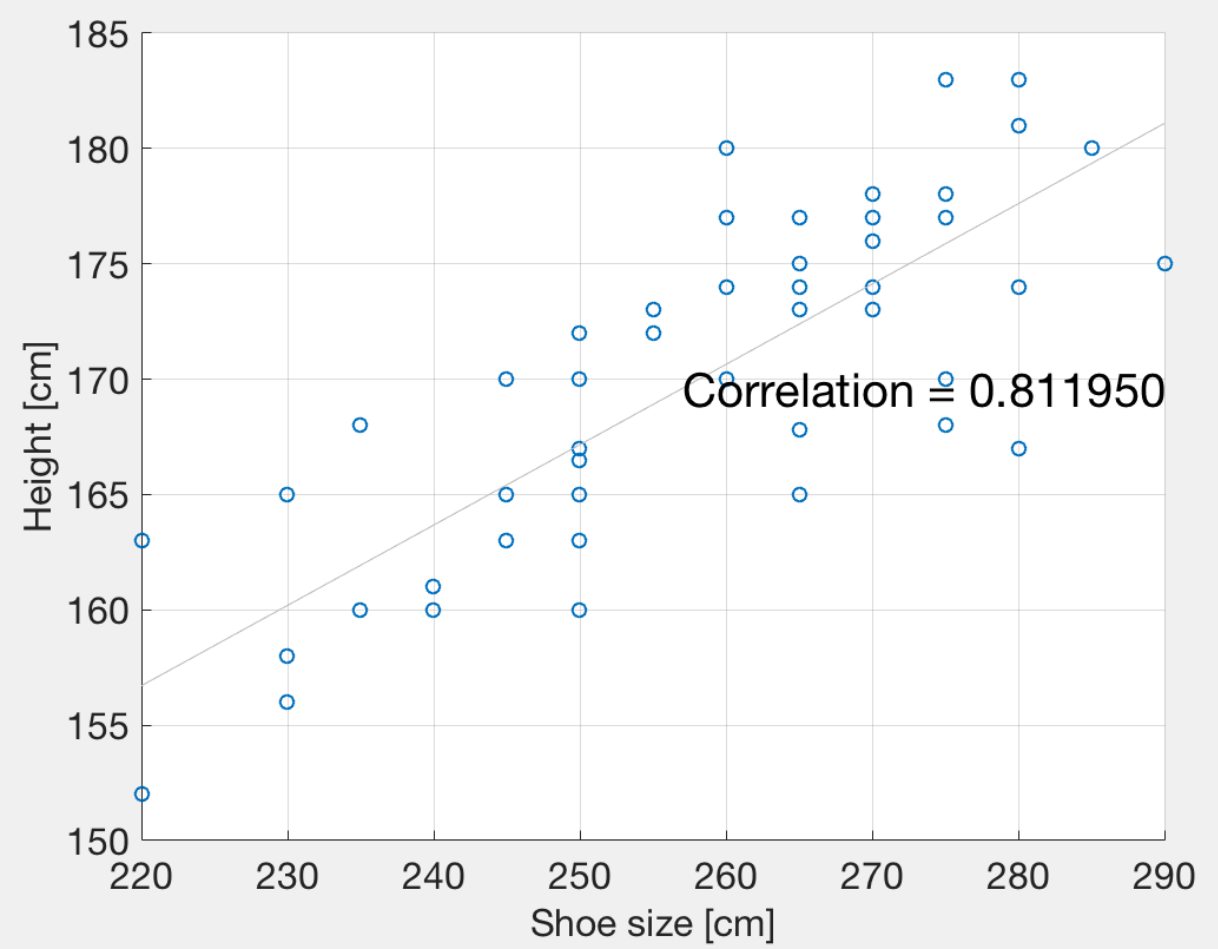
Properties of Pearson correlation

- $-1 \leq r \leq 1$

- positive association: $r > 0$
- negative association: $r < 0$
- no linear association: $r \approx 0$

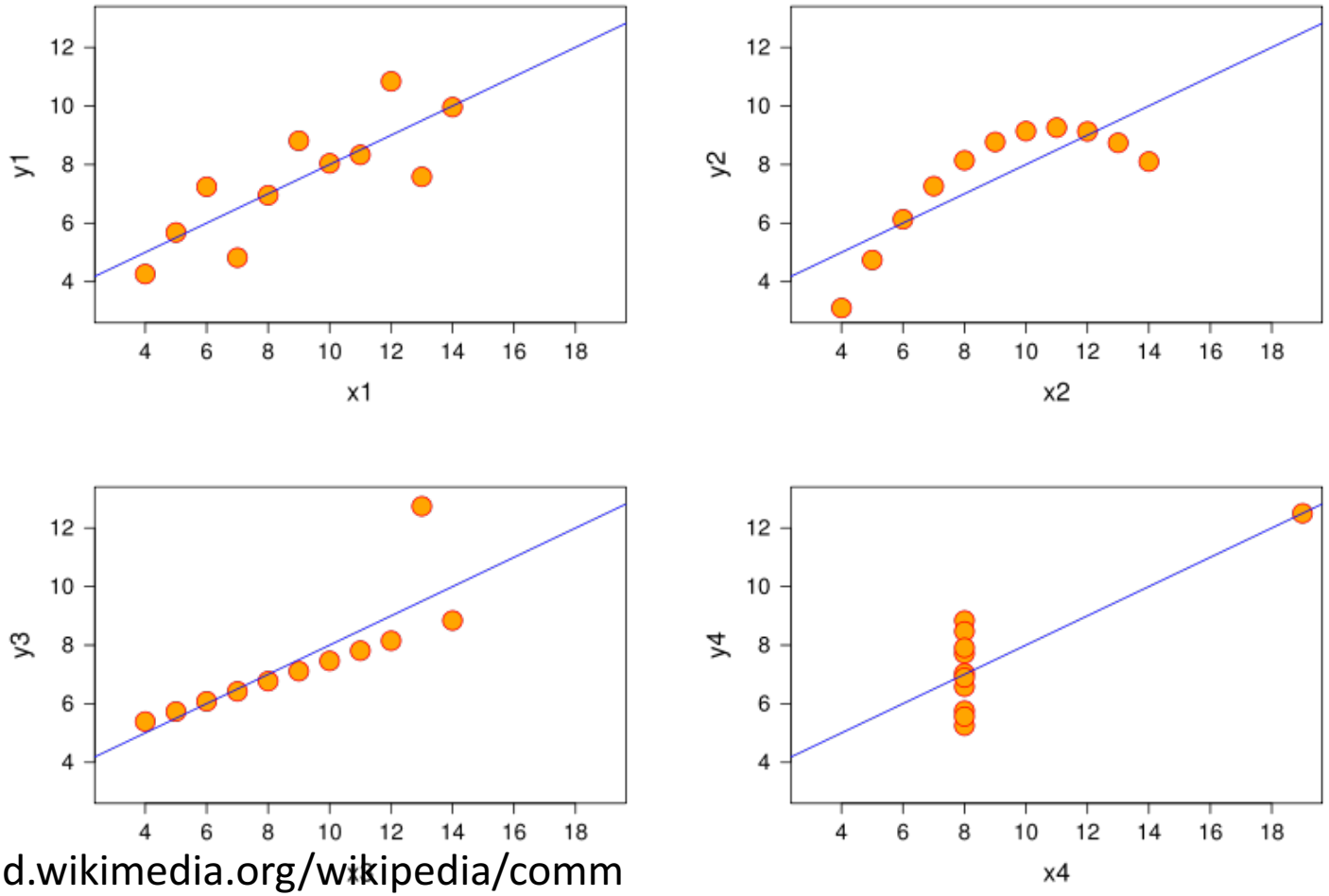
Correlation

- For the data relating shoe size to height, the correlation is (not surprisingly) very high: r=0.812



Dangers of correlation

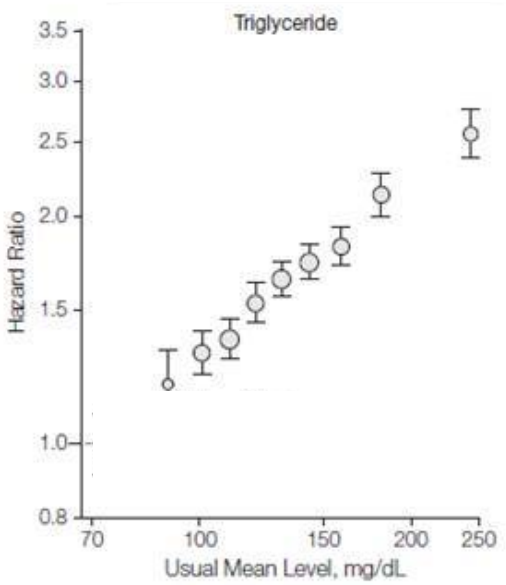
- All of these datasets have r=0.816



http://upload.wikimedia.org/wikipedia/commons/b/b6/Anscombe.svg

Correlation: determining causation

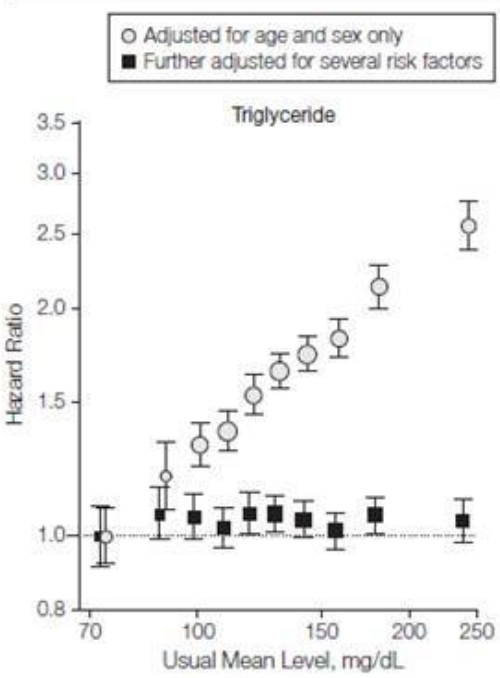
- Cardiovascular events are the leading cause of death in the US
- US people enjoy a very fatty diet – hence people started to look at analyzing markers in the blood and correlation it to prevalence of heart diseases



Getting the levels of triglycerides down would reduce your risk of heart diseases dramatically!

Correlation: determining causation

- Cardiovascular events are the leading cause of death in the US
- US people enjoy a very fatty diet – hence people started to look at analyzing markers in the blood and correlation it to prevalence of heart diseases

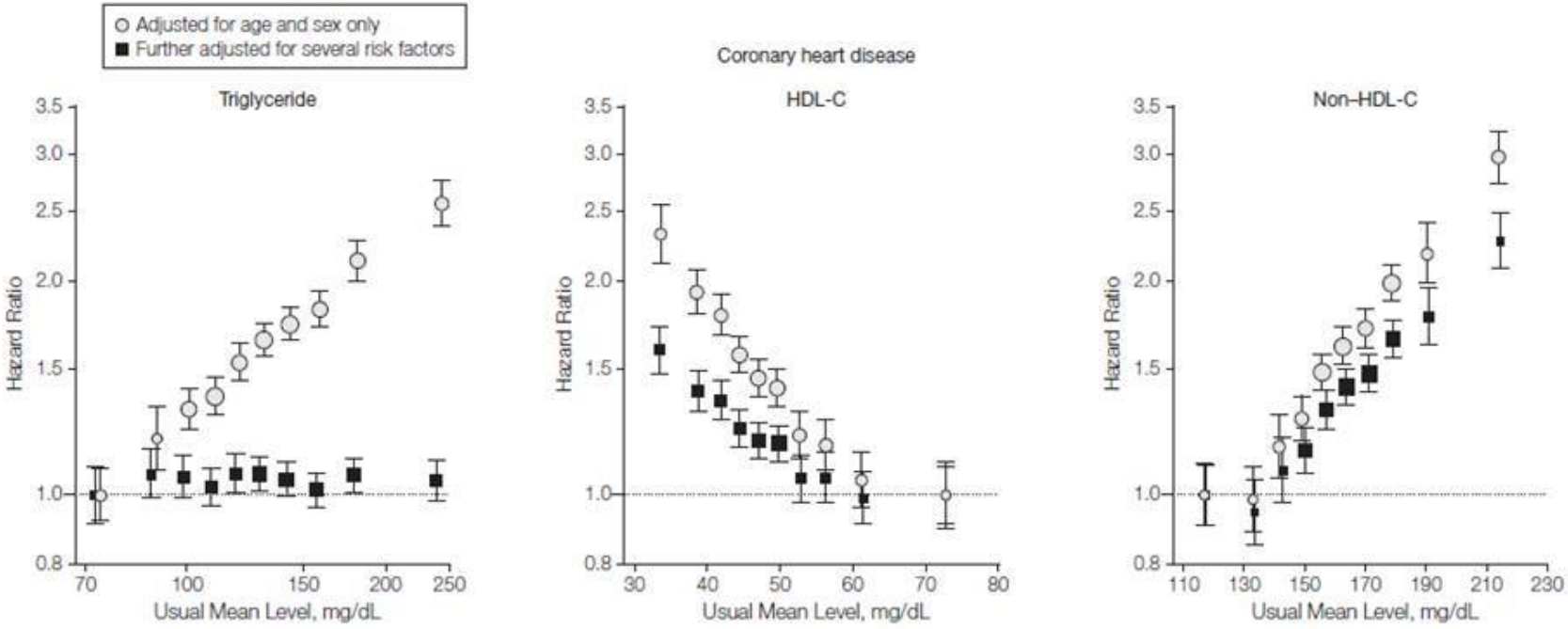


Or maybe not!

Correlation: determining causation

- Cardiovascular events are the leading cause of death in the US
- US people enjoy a very fatty diet – hence people started to look at analyzing markers in the blood and correlation it to prevalence of heart diseases

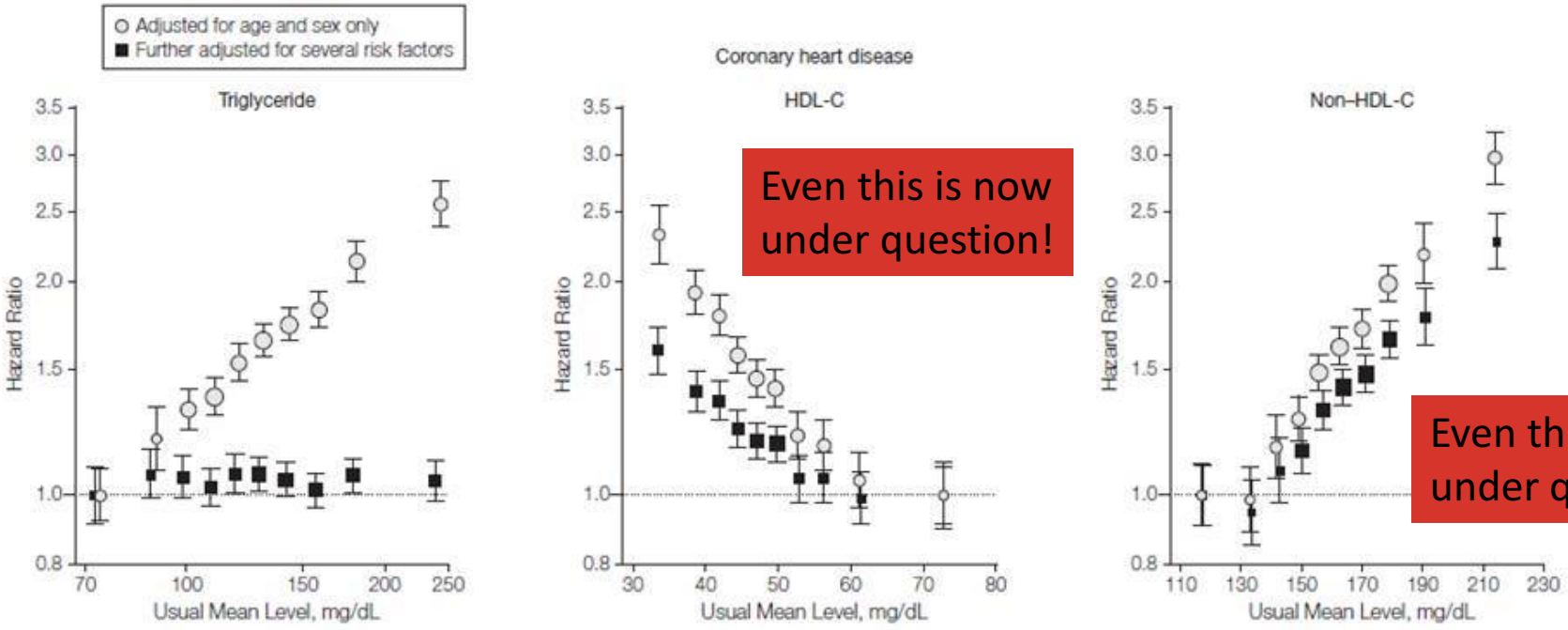
Figure 1. Hazard Ratios for Coronary Heart Disease or Ischemic Stroke Across Quantiles of Usual Triglyceride, HDL-C, and Non–HDL-C Levels



Correlation: determining causation

- Cardiovascular events are the leading cause of death in the US
- US people enjoy a very fatty diet – hence people started to look at analyzing markers in the blood and correlation it to prevalence of heart diseases

Figure 1. Hazard Ratios for Coronary Heart Disease or Ischemic Stroke Across Quantiles of Usual Triglyceride, HDL-C, and Non–HDL-C Levels



Correlation Caveats

CognitiveSystems

- Correlation can be heavily affected by outliers – especially for small sample sizes
  - Always plot your data!

- r = 0 means that there is no **linear** association between your variables - they could still be otherwise associated
  - Always plot your data!

- Correlation does not imply causation
  - Be careful in interpreting your data!

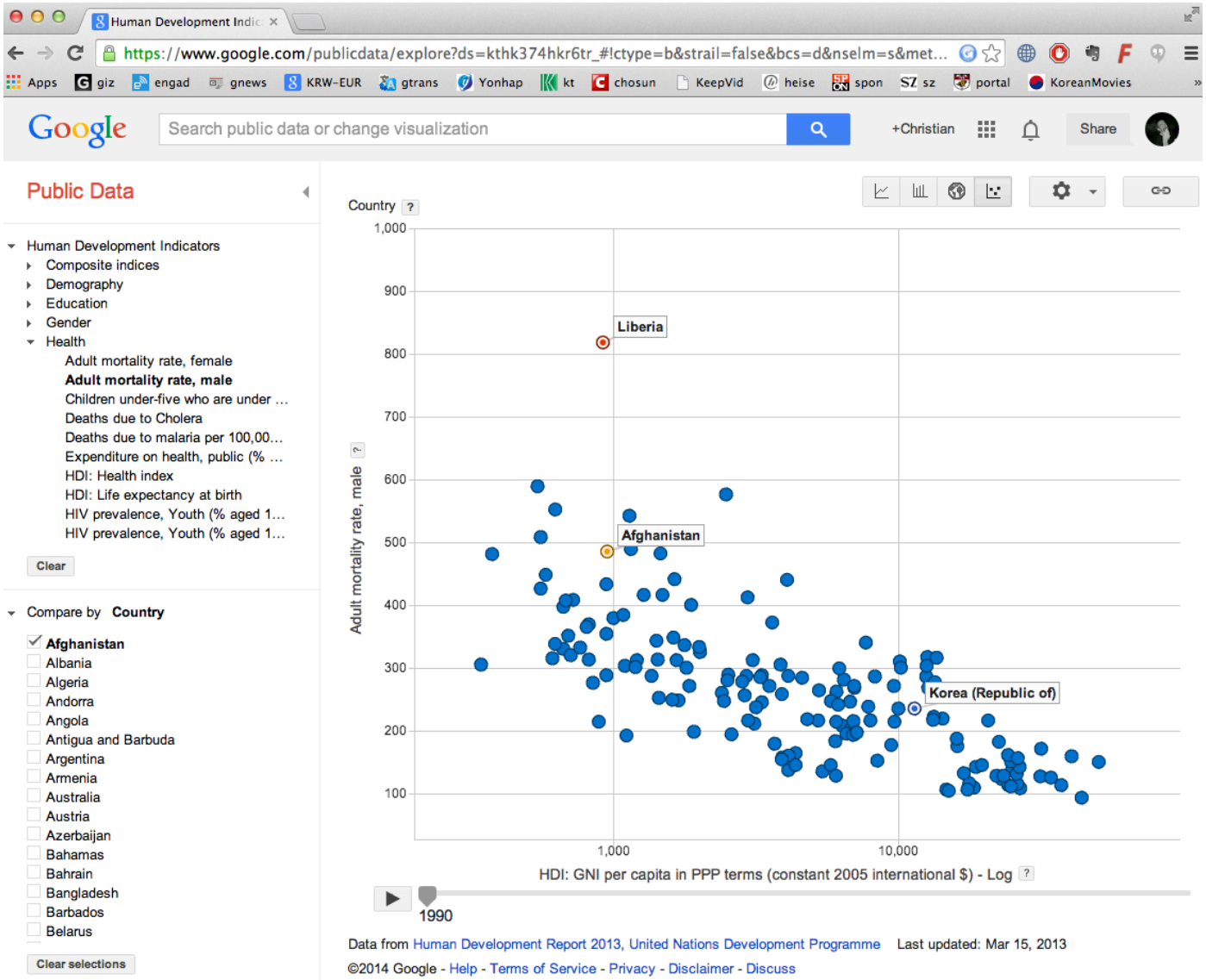
Summary: Two Quantitative Variables

- Summary Statistics
  - Linear Pearson correlation
- Visualization
  - Scatterplot

Variable(s)	Visualization	Summary Statistics
Categorical	bar chart	frequency table, relative frequency table, proportion
Numerical	dotplot, histogram, boxplot	mean, median, max, min, standard deviation, range, IQR, five number summary
Categorical vs Categorical	side-by-side bar chart, stacked bar chart	two-way table, difference in proportions
Numerical vs Categorical	side-by-side boxplots	statistics by group
Numerical vs Numerical	scatter-plot	correlation

Google public data

CognitiveSystems



Statistics in real life!

CognitiveSystems