CAAP Statistics - Lec04

Jul 11, 2022

Review: R Introduction

R an open-source language and environment for statistical computing and graphics

- Download R
- Download RStudio

R as a calculator

```
1+2
## [1] 3
sample(1:5,1)
## [1] 2
names =
c("Ini","Isaias","Luz","Maddie","Ryan","Soraya","Vi
olet")
sample(names, 4, replace = TRUE)
## [1] "Ryan"    "Luz"     "Violet" “Violet"
```

Install and load the packages

```
# Install pacakges first
# install.packages("tidyverse")
# install.packages("openintro")
# install.packages("ggplot2")

library(tidyverse)
library(openintro)
library(ggplot2)
```

Let’s see the actual data

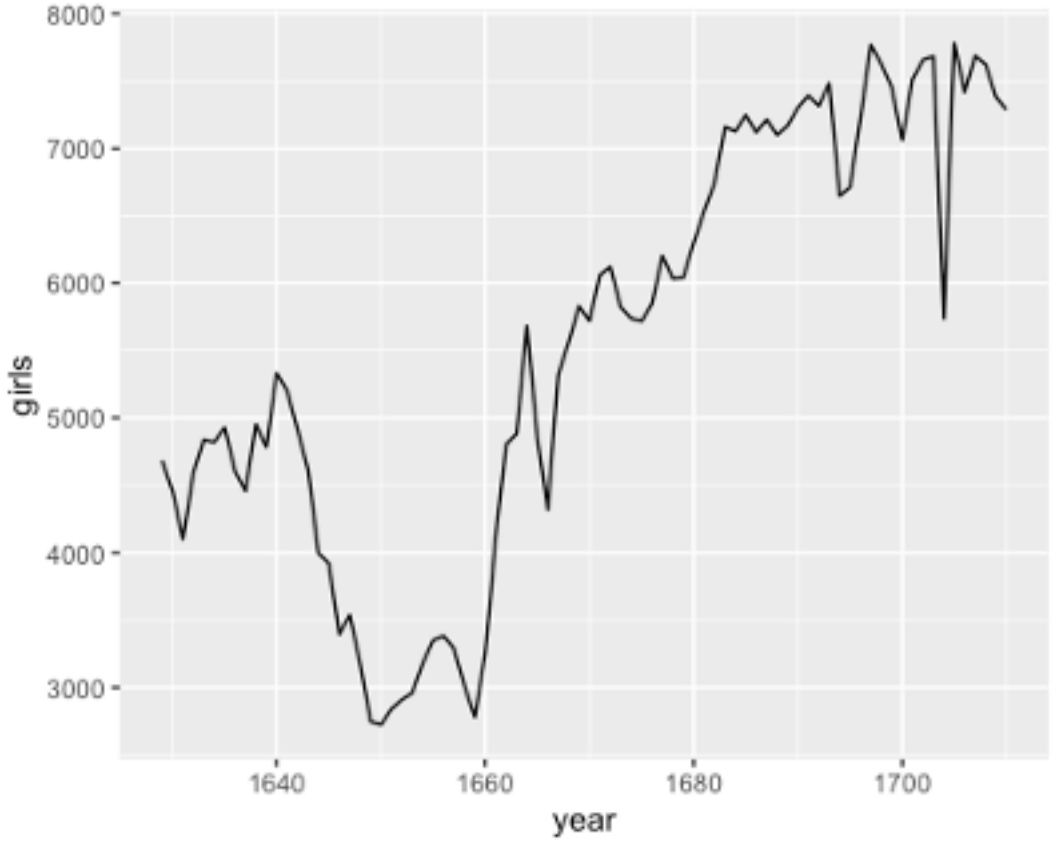
```
arbuthnot # from openintro package
data_web = read.csv("https://www.openintro.org/
book/statdata/arbuthnot.csv") # from web
# getwd() # check for the current working directory
# data = read.csv("arbuthnot.csv") # read from the
working directory
```

How does the data look like?

```
glimpse(arbuthnot)
## Rows: 82
## Columns: 3
## $ year  <int> 1629, 1630, 1631, 1632, 1633,
1634, 1635, 1636, 1637, 1638, 1639…
## $ boys  <int> 5218, 4858, 4422, 4994, 5158,
5035, 5106, 4917, 4703, 5359, 5366…
## $ girls <int> 4683, 4457, 4102, 4590, 4839,
4820, 4928, 4605, 4457, 4952, 4784…
```

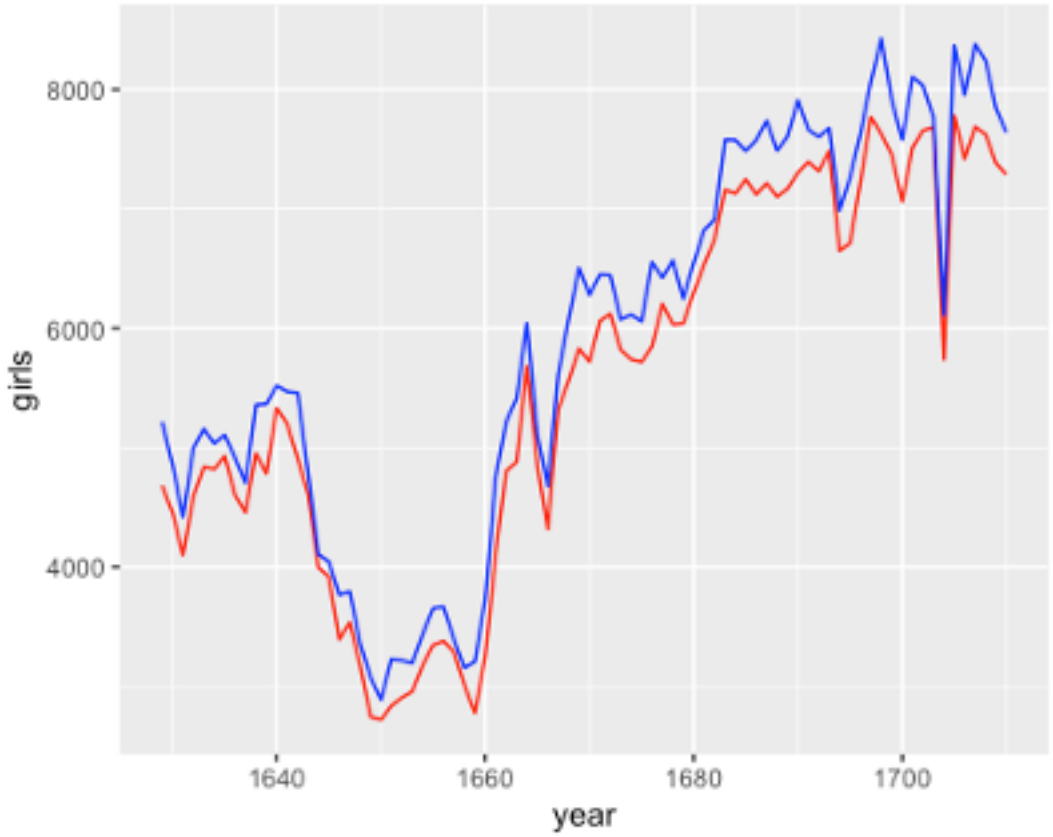
Visualize the Data

```
ggplot(data = arbuthnot,
aes(x=year, y = girls))+
  geom_line()
```



Visualize the Data

```
ggplot(data = arbuthnot)+
  geom_line(aes(x=year, y =
girls),colour = "red")+
  geom_line(aes(x=year, y =
boys),colour="blue")
```



Learning Objectives

- Numerical Data
  - Graphical summary
    - Scatterplot
    - Histogram
    - Boxplot
  - Numerical summary
    - Mean and Variance
- Categorical Data
  - Graphical Summary
    - Contingency tables and Bar plot
    - Mosaic plot(If time allows)

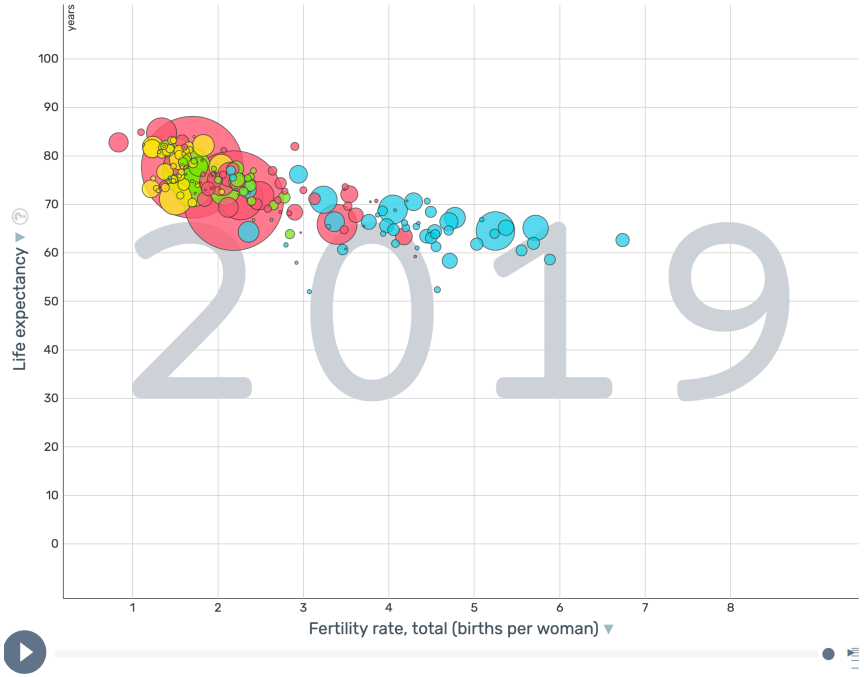
Examining Numerical Data

Scatterplot

Scatterplots are useful for visualizing the relationship between two numerical variables.

Do life expectancy and total fertility appear to be *associated* or *independent*?

Was the relationship the same throughout the years, or did it change?



http://www.gapminder.org/world

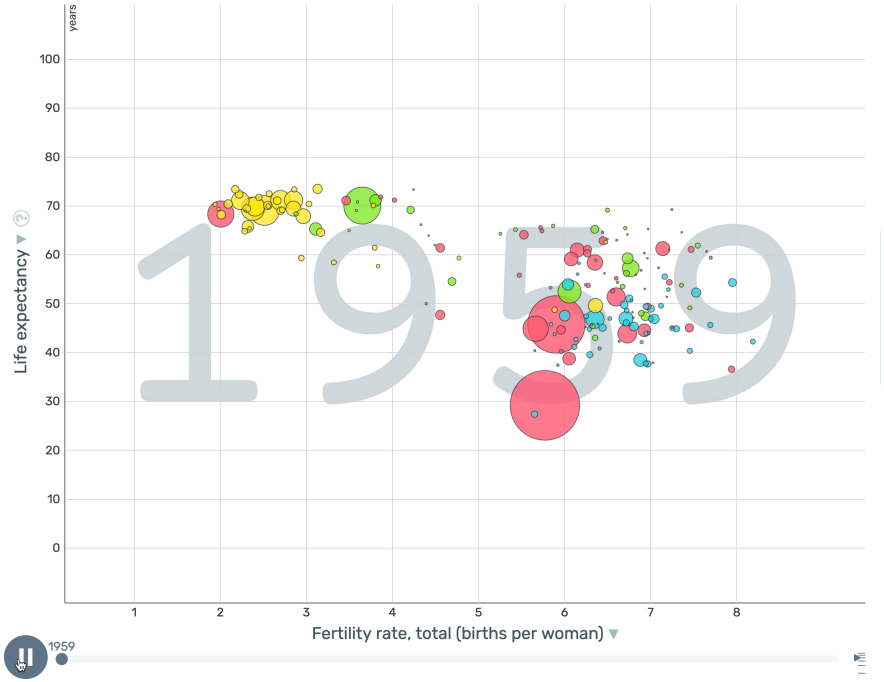
Scatterplot

Scatterplots are useful for visualizing the relationship between two numerical variables.

Do life expectancy and total fertility appear to be *associated* or *independent*?

They appear to be linearly and negatively associated: as fertility increases, life expectancy decreases.

Was the relationship the same throughout the years, or did it change?



http://www.gapminder.org/world

Scatterplot

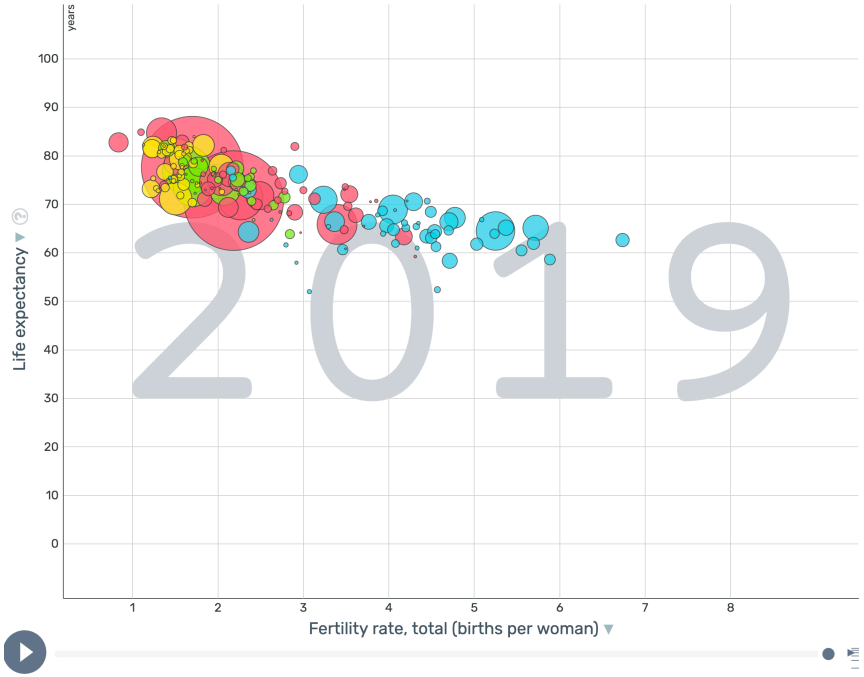
Scatterplots are useful for visualizing the relationship between two numerical variables.

Do life expectancy and total fertility appear to be *associated* or *independent*?

They appear to be linearly and negatively associated: as fertility increases, life expectancy decreases.

Was the relationship the same throughout the years, or did it change?

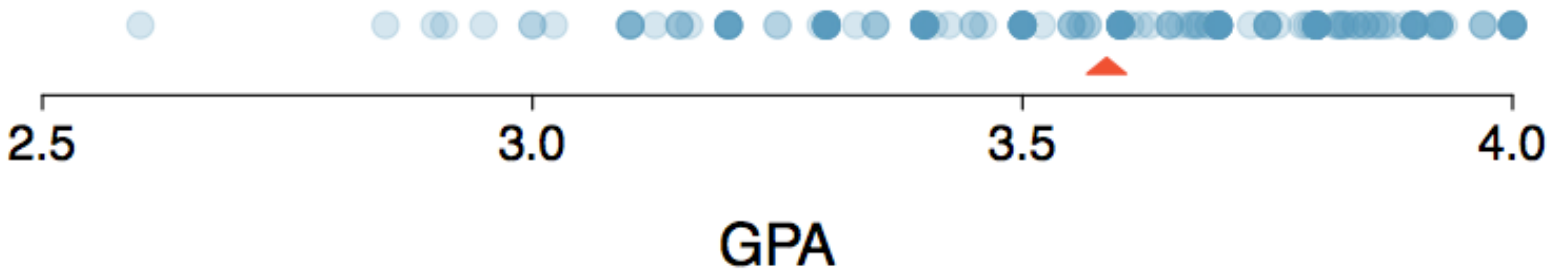
Yes and no



http://www.gapminder.org/world

Dot Plots & Mean

Useful for visualizing one numerical variable. Darker colors represent areas where there are more observations.



The *mean*, also called the *average* (marked with a triangle in the above plot), is one way to measure the center of a *distribution* of data.

The mean GPA is 3.59.

Mean

The *sample mean*, denoted as $\bar{x}$, can be calculated as

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

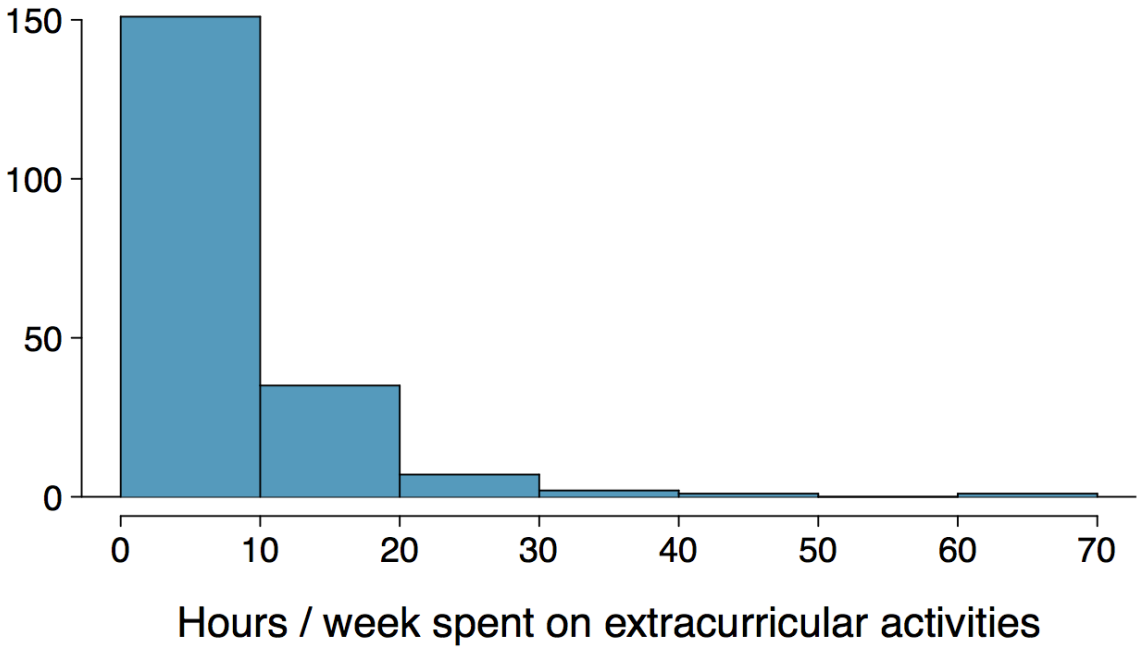
where x_1, x_2, ..., x_n represent the *n* observed values.

The *population mean* is also computed the same way but is denoted as *μ*. It is often not possible to calculate μ since population data are rarely available.

The sample mean is a *sample statistic*, and serves as a *point estimate* of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population), it is usually a pretty good estimate.

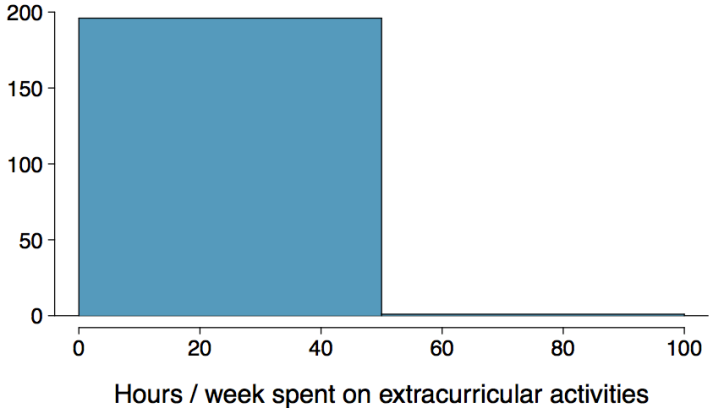
Histograms - Extracurricular Hours

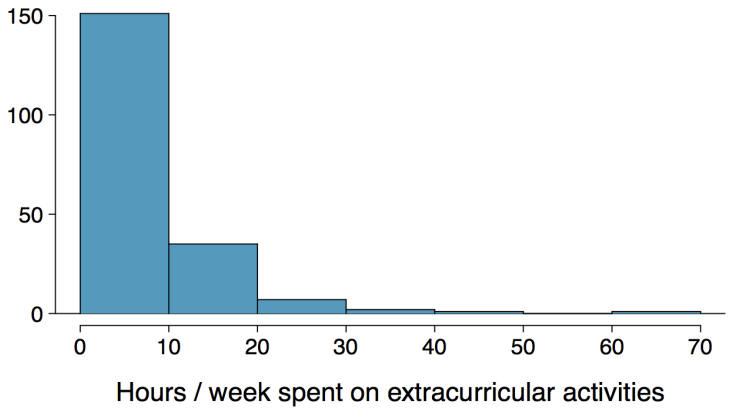
- Histograms provide a view of the *data density*. Higher bars represent where the data are relatively more common.
- Histograms are especially convenient for describing the *shape* of the data distribution.
- The chosen *bin width* can alter the story the histogram is telling.

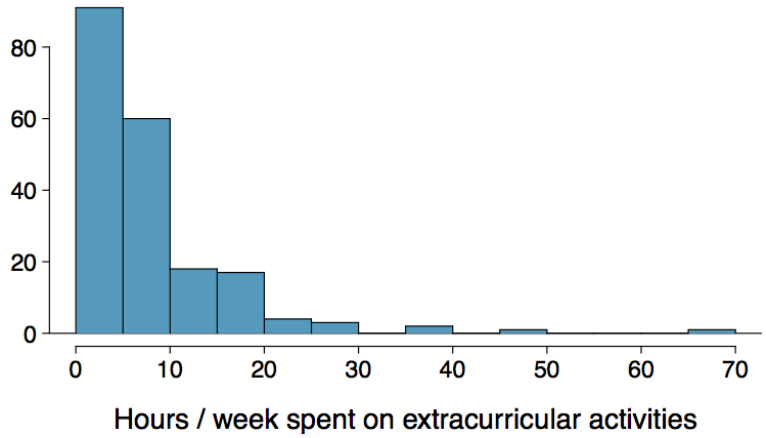


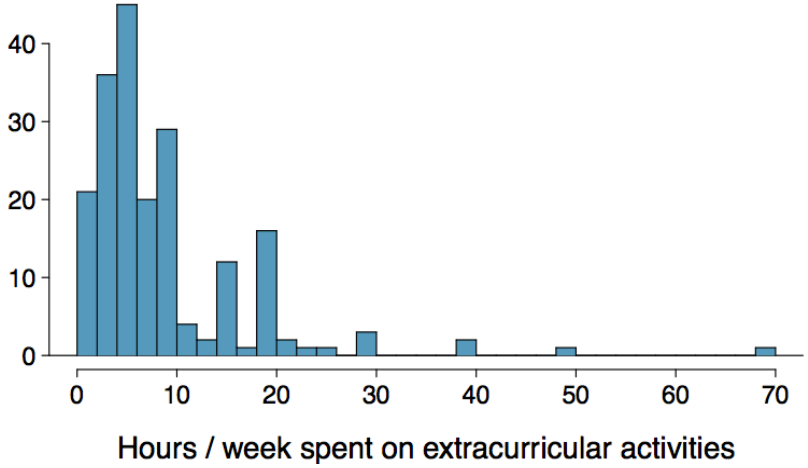
Bin Width

Which one(s) of these histograms are useful? Which reveal too much about the data? Which hide too much?



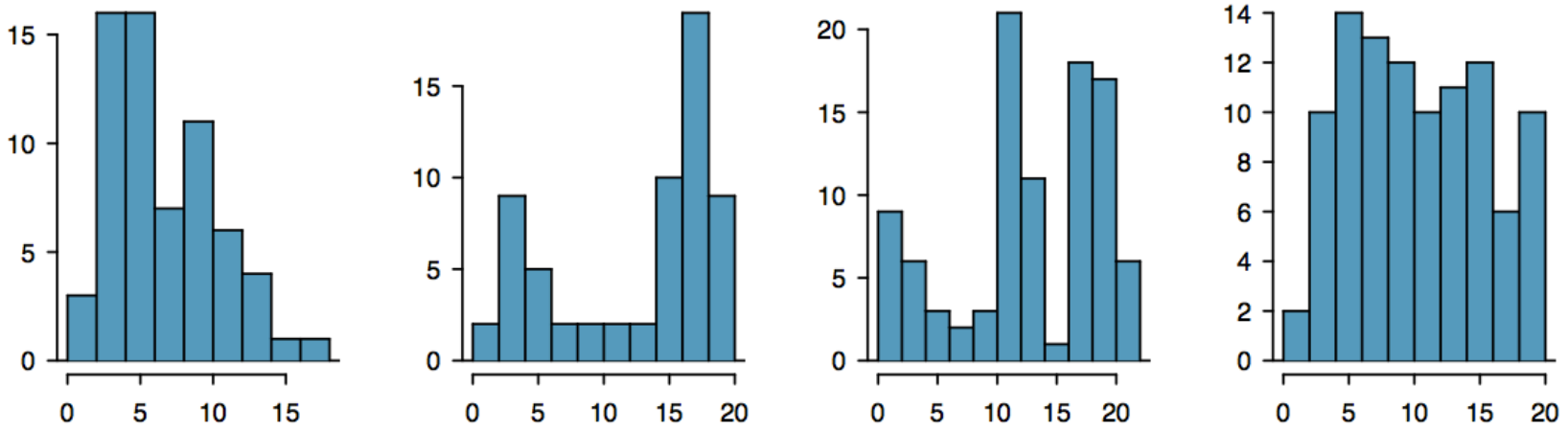






Shape of a Distribution: Modality

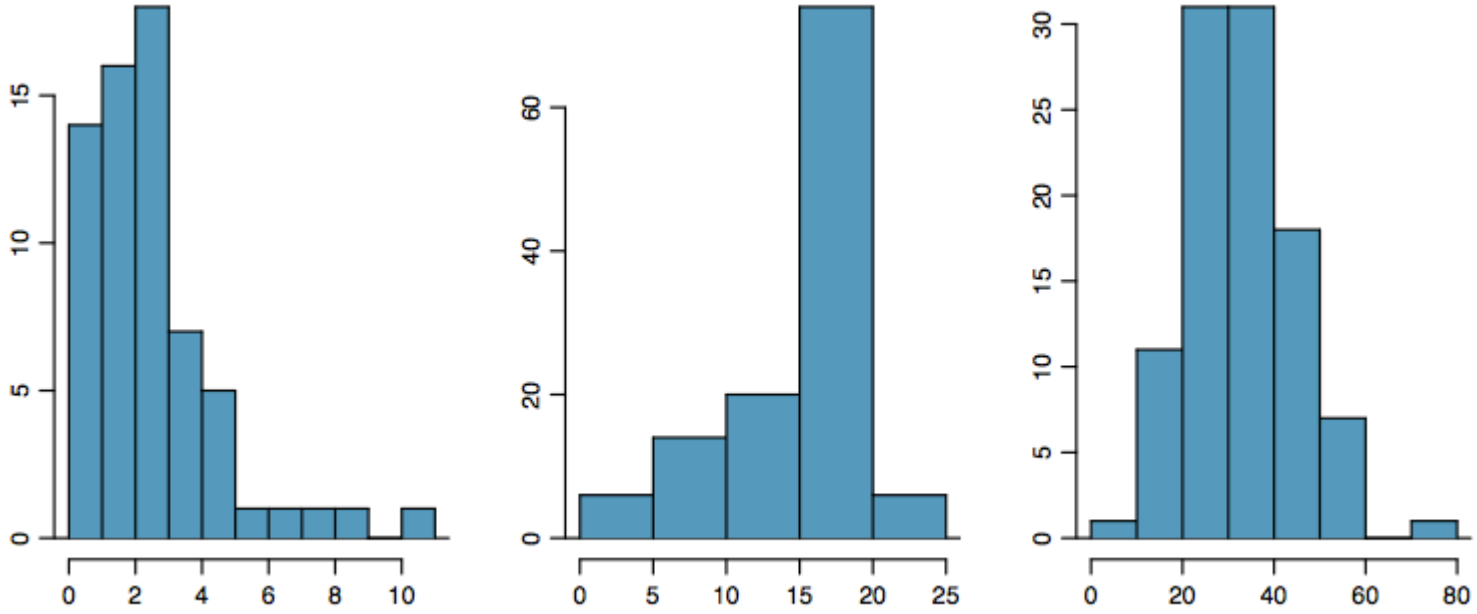
Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?



Note: In order to determine modality, step back and imagine a smooth curve over the histogram -- imagine that the bars are wooden blocks and you drop a limp spaghetti over them, the shape the spaghetti would take could be viewed as a smooth curve.

Shape of a Distribution: Skewness

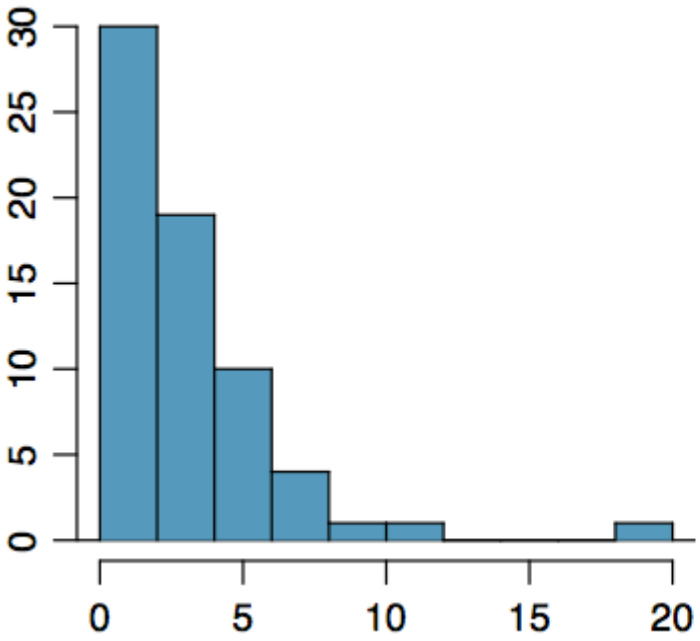
Is the histogram *right skewed*, *left skewed*, or *symmetric*?

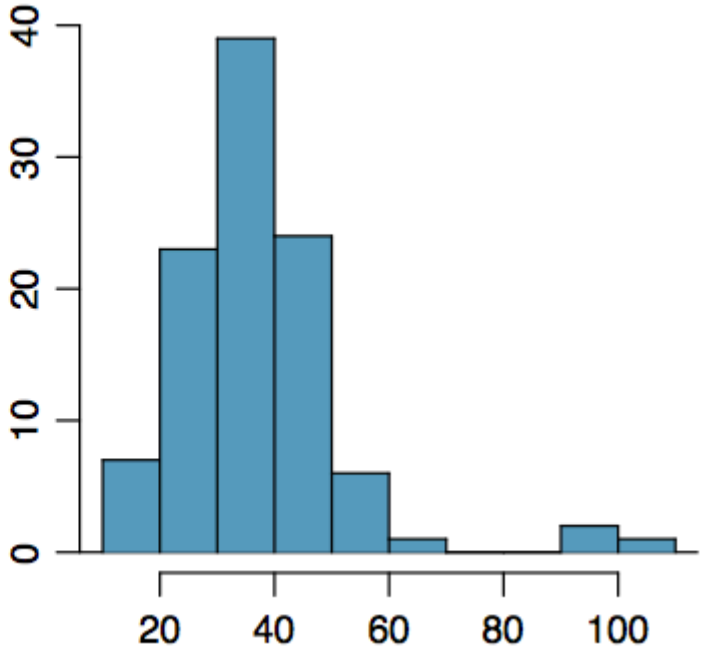


Note: Histograms are said to be skewed to the side of the long tail.

Shape of a Distribution: Unusual Observations

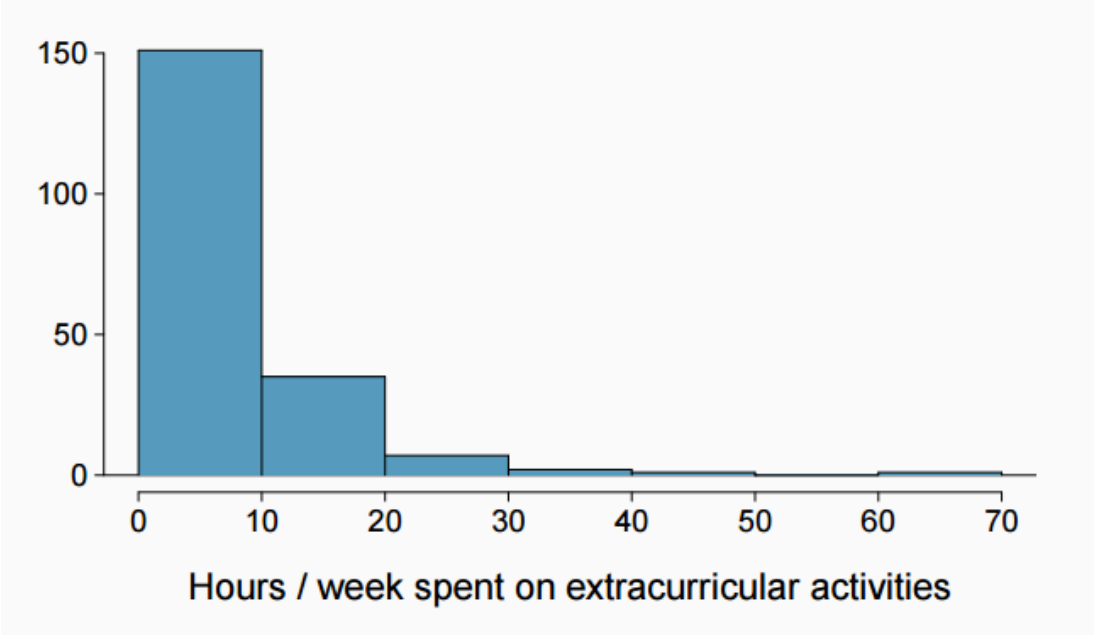
Are there any unusual observations or potential *outliers*?





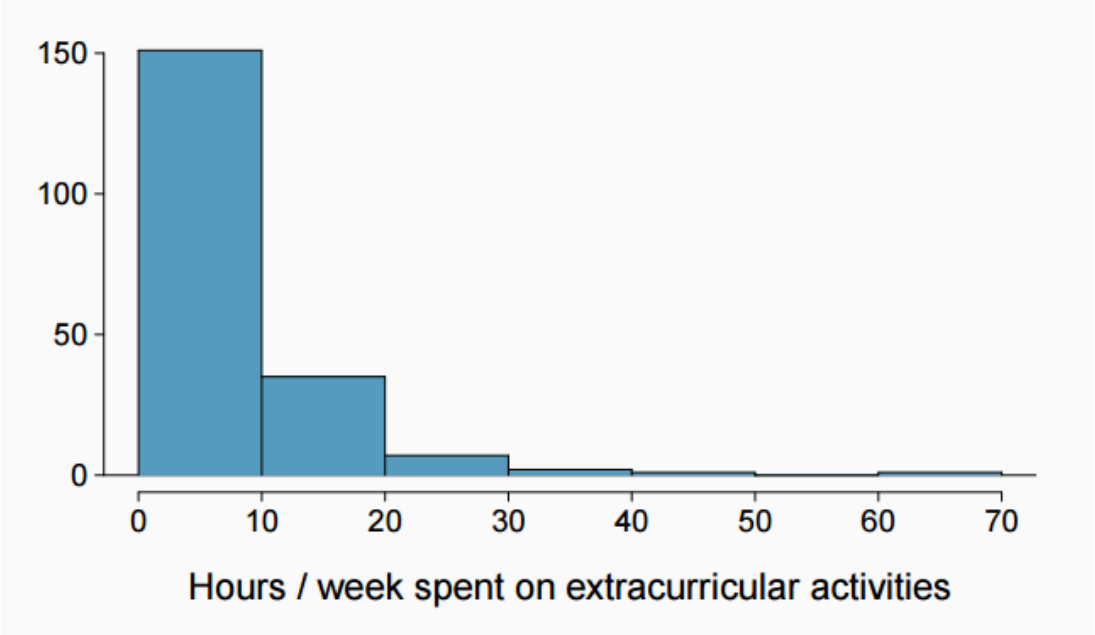
Extracurricular activities

How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?



Extracurricular activities

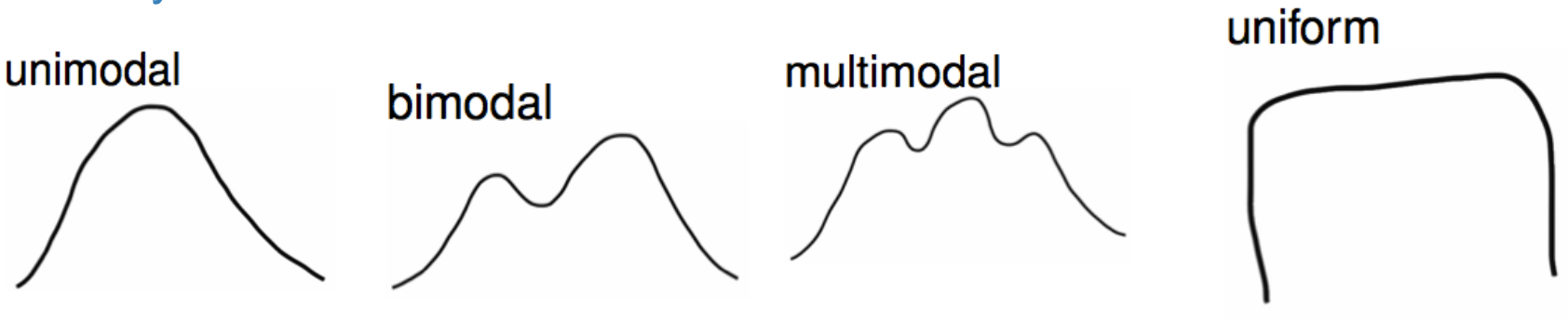
How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?



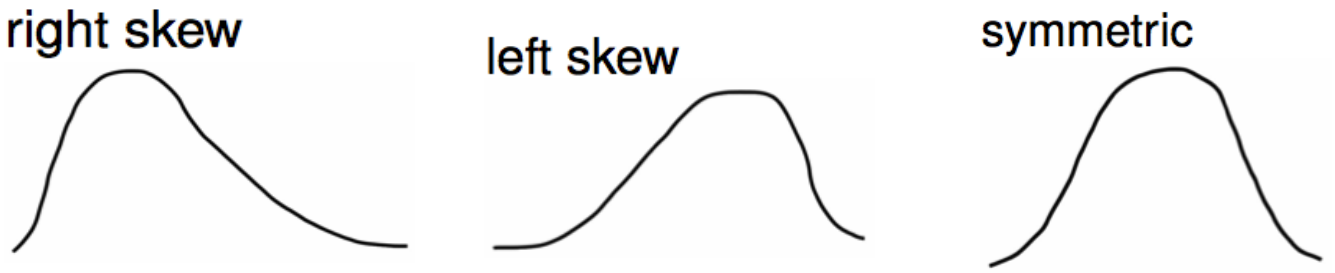
Unimodal and right skewed, with a potentially unusual observation at 60 hours/week.

Commonly observed shapes of distributions

Modality



Skewness



Practice

Which of these variables do you expect to be uniformly distributed?

(a) weights of adult females

(b) salaries of a random sample of people from North Carolina

(c) house prices

(d) birthdays of classmates (day of the month)

Practice

Which of these variables do you expect to be uniformly distributed?

(a) weights of adult females

(b) salaries of a random sample of people from North Carolina

(c) house prices

(d) birthdays of classmates (day of the month)

Are you typical?

http://www.youtube.com/watch?v=4B2xOvKFFz4

How useful are centers alone for conveying the true characteristics of a distribution?

Variance

Variance is roughly the average squared deviation from the mean.

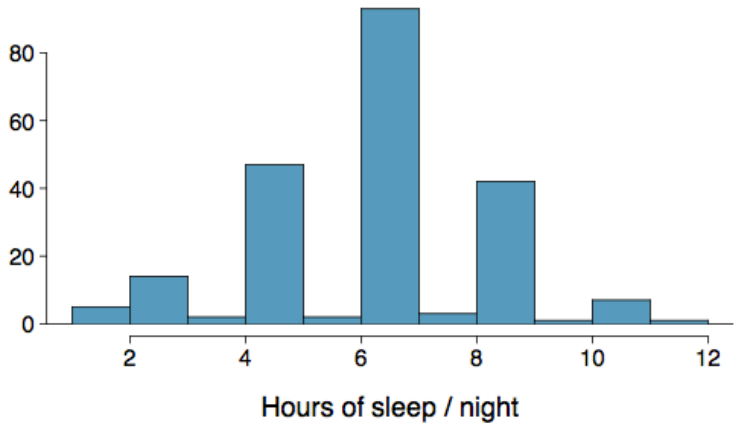
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Variance

Variance is roughly the average squared deviation from the mean.

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

- The sample mean is $\bar{x} = 6.71$, and the sample size is n = 217.

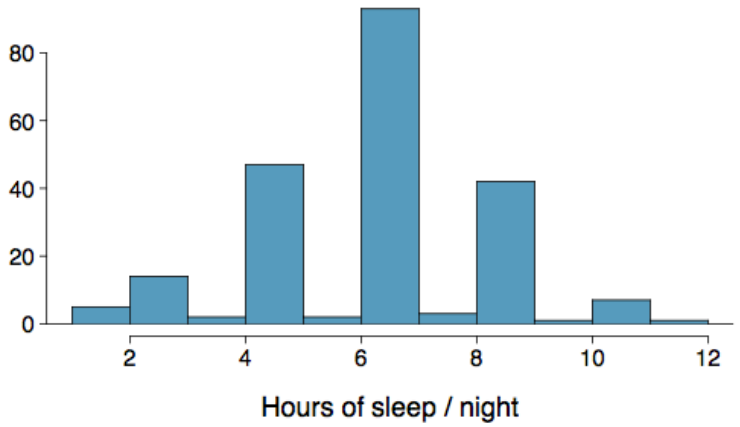


Variance

Variance is roughly the average squared deviation from the mean.

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

- The sample mean is $\bar{x} = 6.71$, and the sample size is n = 217.
- The variance of amount of sleep students get per night can be calculated as:



$$s^2 = \frac{(5-6.71)^2 + (9-6.71)^2 + \cdots + (7-6.71)^2}{217-1} = 4.11\ hours^2$$

Variance (cont.)

Why do we use the squared deviation in the calculation of variance?

Variance (cont.)

Why do we use the squared deviation in the calculation of variance?

- To get rid of negatives so that observations equally distant from the mean are weighed equally.
- To weigh larger deviations more heavily.

Standard Deviation

The *standard deviation* is the square root of the variance, and has the same units as the data.

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

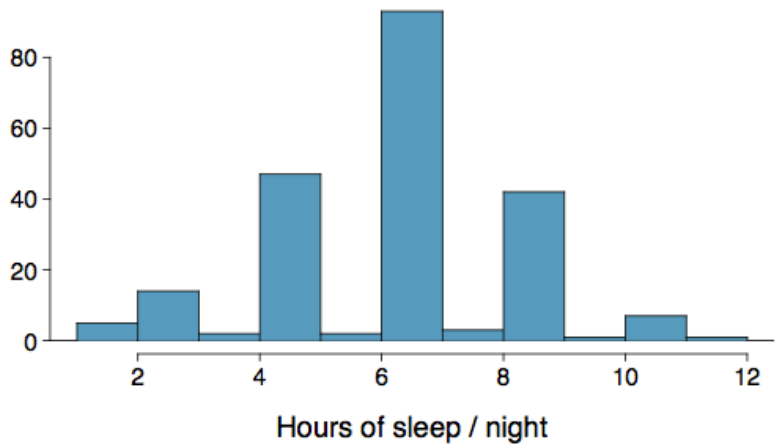
Standard Deviation

The *standard deviation* is the square root of the variance, and has the same units as the data.

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

- The standard deviation of amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03 \; hours$$



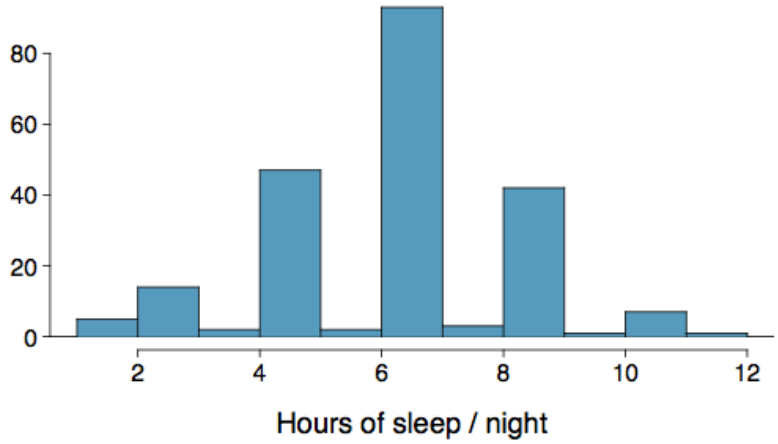
Standard Deviation

The *standard deviation* is the square root of the variance, and has the same units as the data.

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

- The standard deviation of amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03 \; hours$$



- We can see that all of the data are within 3 standard deviations of the mean.

Median

The *median* is the value that splits the data in half when ordered in ascending order.

$$0, 1, 2, 3, 4$$

If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50th percentile.

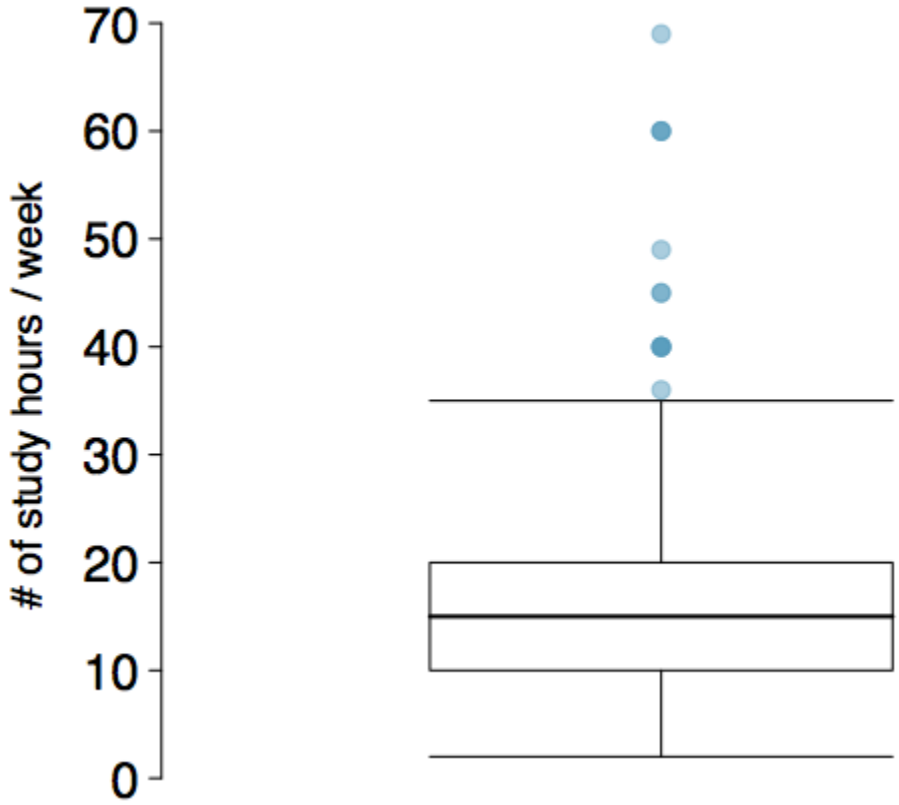
Q1, Q3, and IQR

- The 25th percentile is also called the first quartile, *Q1*.
- The 50th percentile is also called the median.
- The 75th percentile is also called the third quartile, *Q3*.
- Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the *interquartile range*, or the *IQR*.

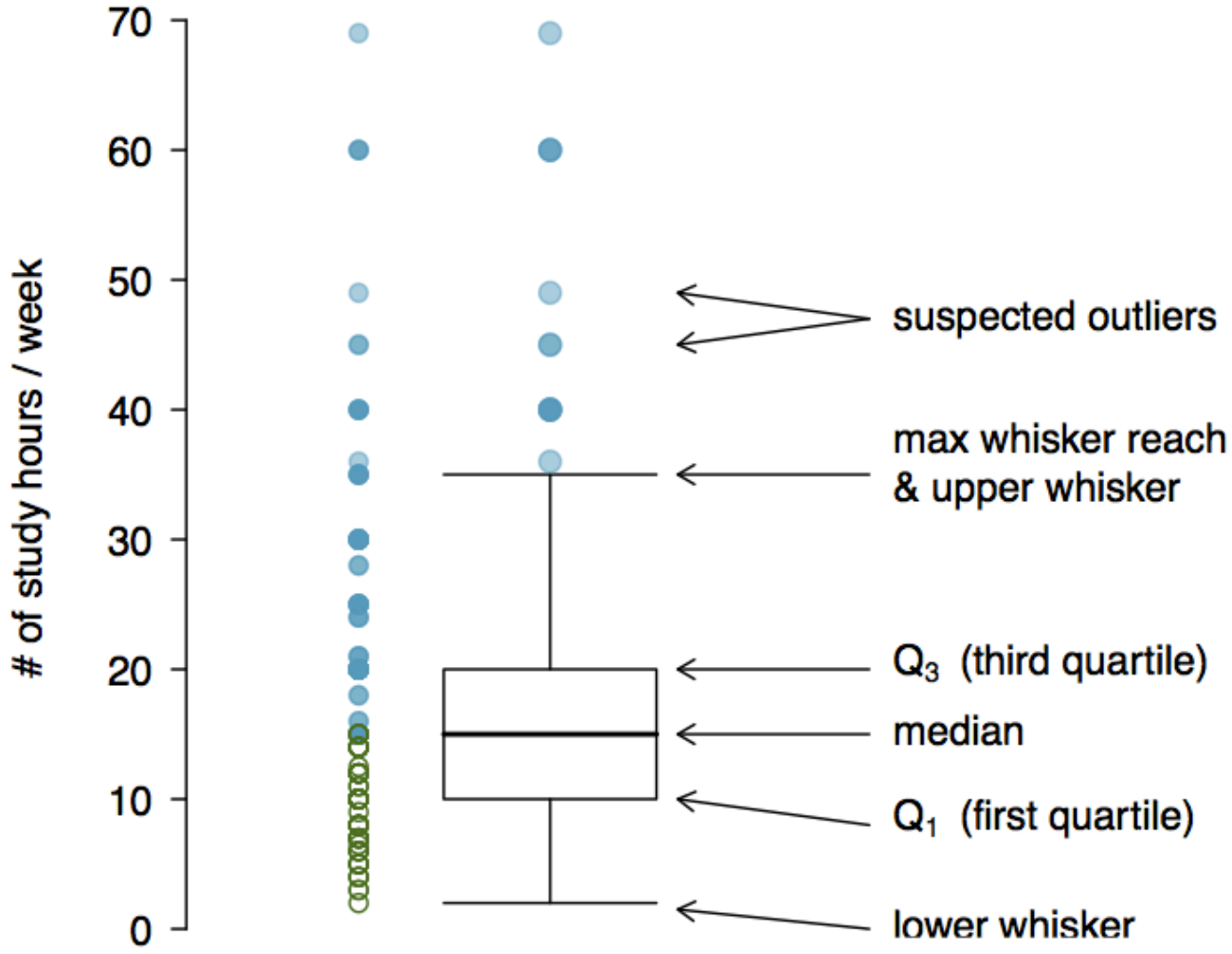
IQR = Q3 - Q1

Box Plot

The box in a *box plot* represents the middle 50% of the data, and the thick line in the box is the median.



Anatomy of a Box Plot



Whiskers and Outliers

Whiskers of a box plot can extend up to 1.5 x IQR away from the quartiles.

max upper whisker reach = Q3 + 1.5 x IQR

max lower whisker reach = Q1 - 1.5 x IQR

Whiskers and Outliers

Whiskers of a box plot can extend up to 1.5 x IQR away from the quartiles.

max upper whisker reach = Q3 + 1.5 x IQR

max lower whisker reach = Q1 - 1.5 x IQR

IQR: 20 - 10 = 10

max upper whisker reach = 20 + 1.5 x 10 = 35

max lower whisker reach = 10 - 1.5 x 10 = -5

Whiskers and Outliers

Whiskers of a box plot can extend up to 1.5 x IQR away from the quartiles.

max upper whisker reach = Q3 + 1.5 x IQR

max lower whisker reach = Q1 - 1.5 x IQR

IQR: 20 - 10 = 10

max upper whisker reach = 20 + 1.5 x 10 = 35

max lower whisker reach = 10 - 1.5 x 10 = -5

A potential *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

Outliers (cont.)

Why is it important to look for outliers?

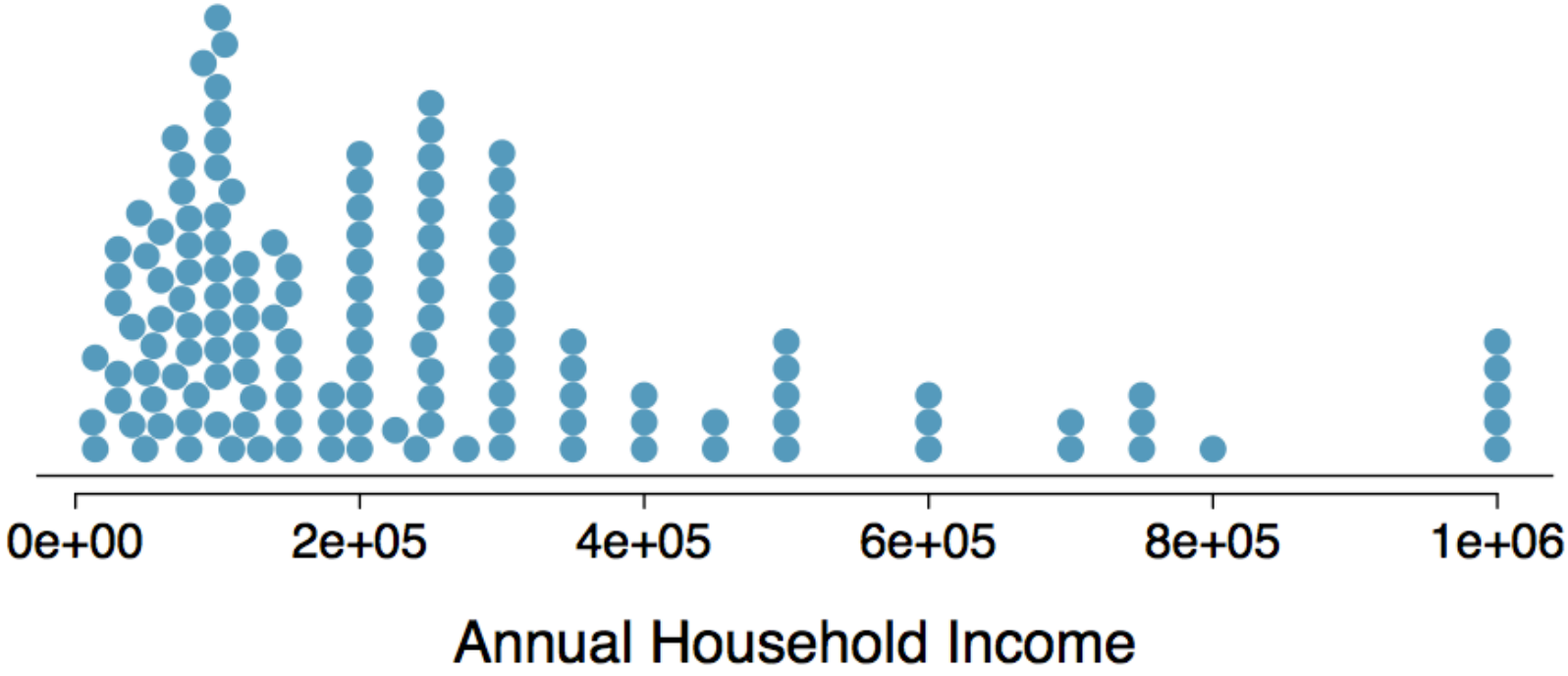
Outliers (cont.)

Why is it important to look for outliers?

- Identify extreme skew in the distribution.
- Identify data collection and entry errors.
- Provide insight into interesting features of the data.

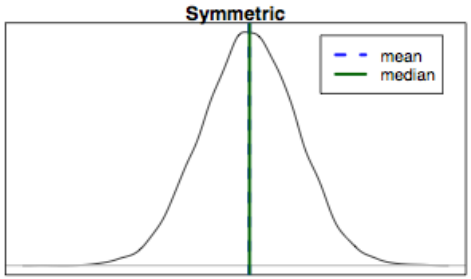
Extreme Observations

How would sample statistics such as mean, median, SD, and IQR of household income be affected if the largest value was replaced with $10 million?



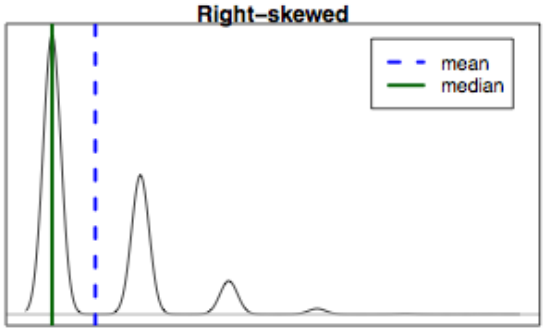
Mean vs. Median

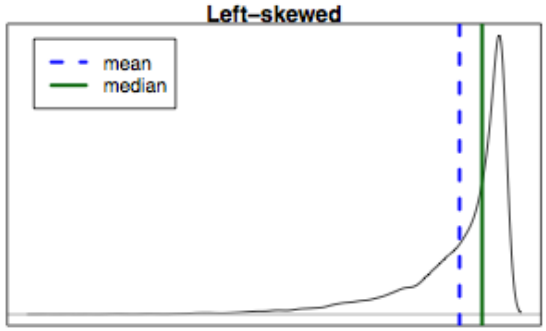
If the distribution is symmetric, center is often defined as the mean:
mean ~ median



If the distribution is skewed or has extreme outliers, center is often defined as the median

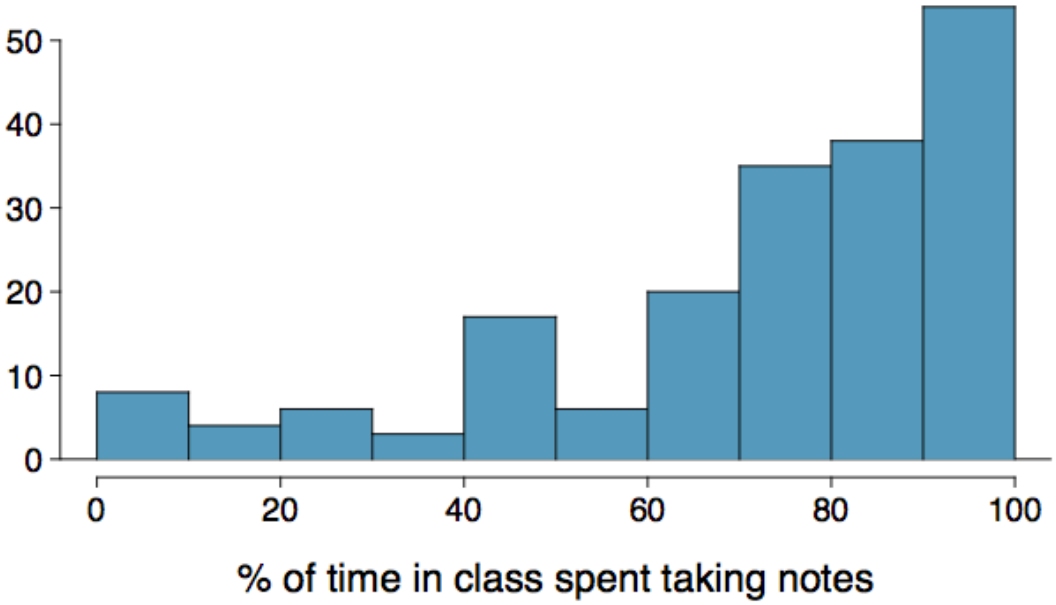
- Right-skewed: mean > median
- Left-skewed: mean < median





Practice

Which is most likely true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter, etc.?



(a) mean > median

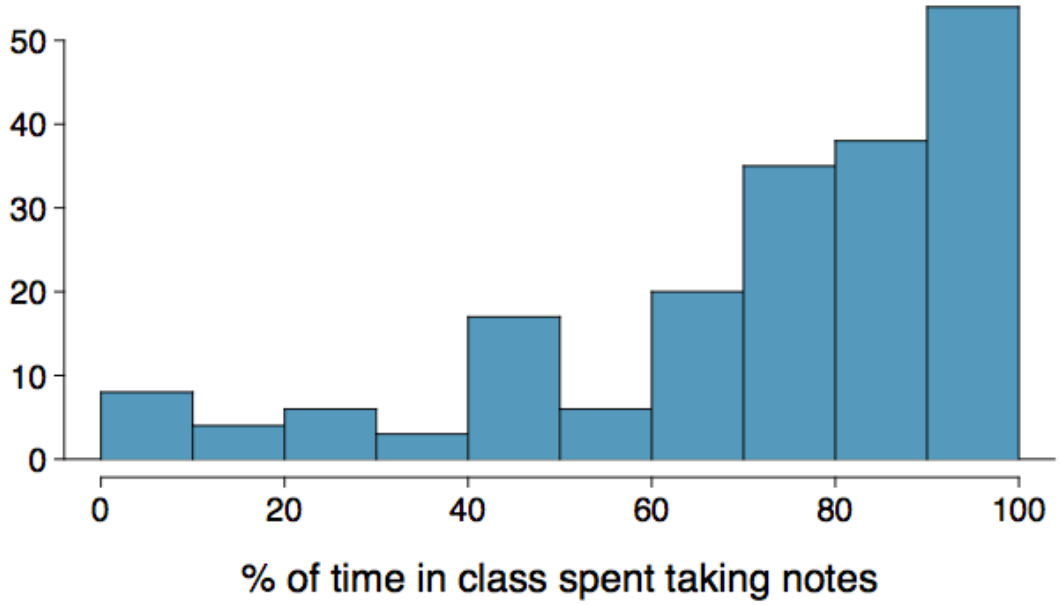
(b) mean ~ median

(c) mean < median

(d) impossible to tell

Practice

Which is most likely true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter, etc.?



median: 80%

mean: 76%

(a) mean > median

(b) mean ~ median

(c) mean < median

(d) impossible to tell

Considering Categorical Data

Contingency Tables

A table that summarizes data for two categorical variables is called a *contingency table*.

Contingency Tables

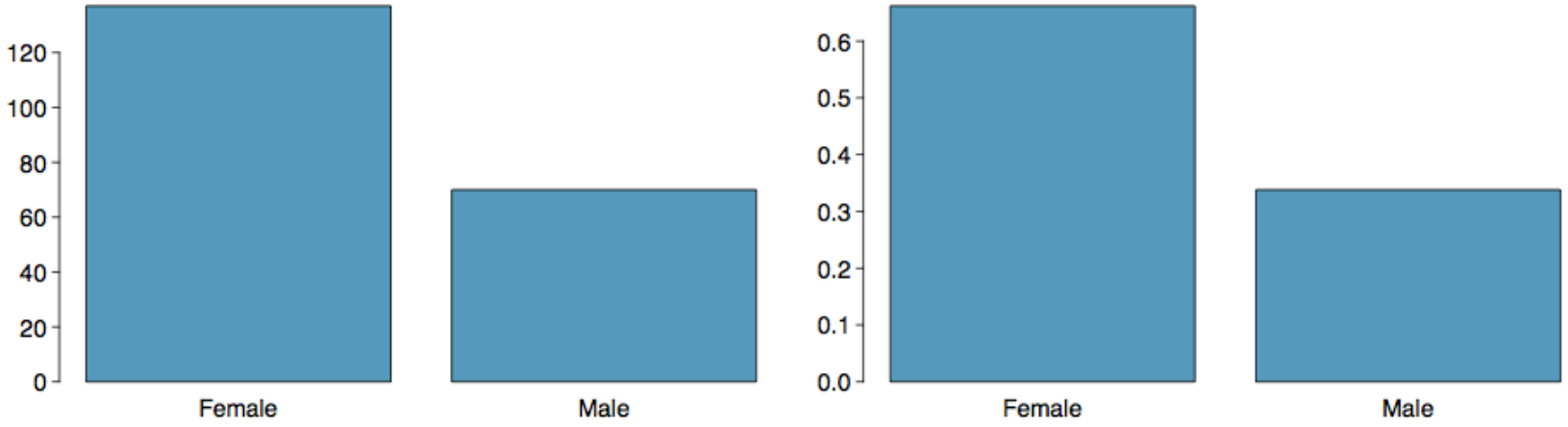
A table that summarizes data for two categorical variables is called a *contingency table*.

Recall the Chronic Fatigue Syndrome experiment. The contingency table below shows the distribution of participants improved after the cognitive-behavioral therapy.

		Good outcome		
		Yes	No	Total
Group	Treatment	19	8	27
	Control	5	21	26
	Total	24	29	53

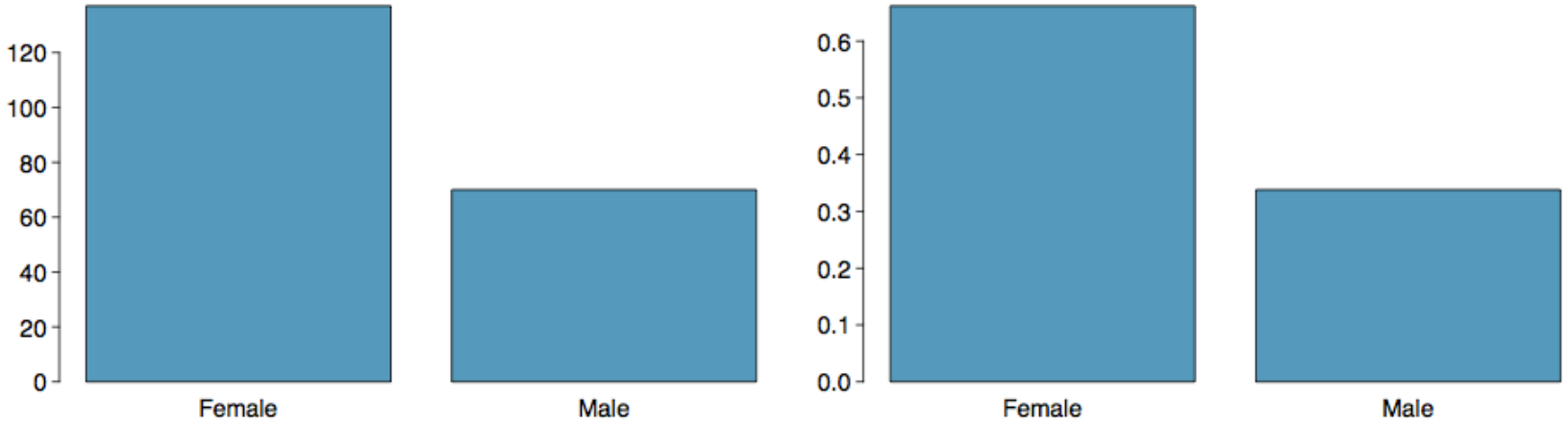
Bar Plots

A *bar plot* is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.



Bar Plots

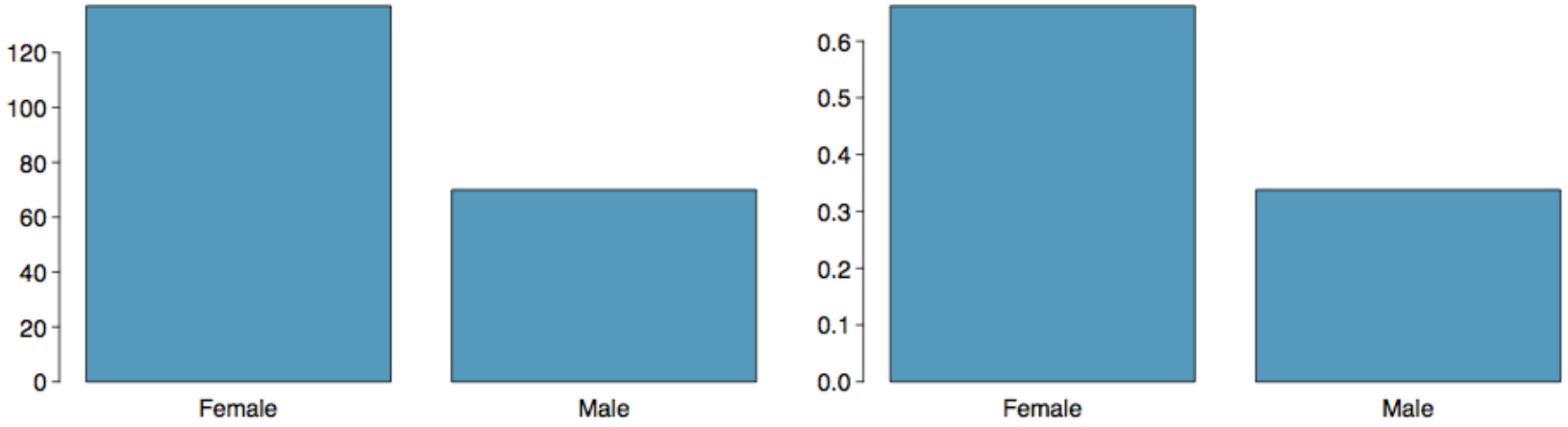
A *bar plot* is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.



How are bar plots different than histograms?

Bar Plots

A *bar plot* is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.

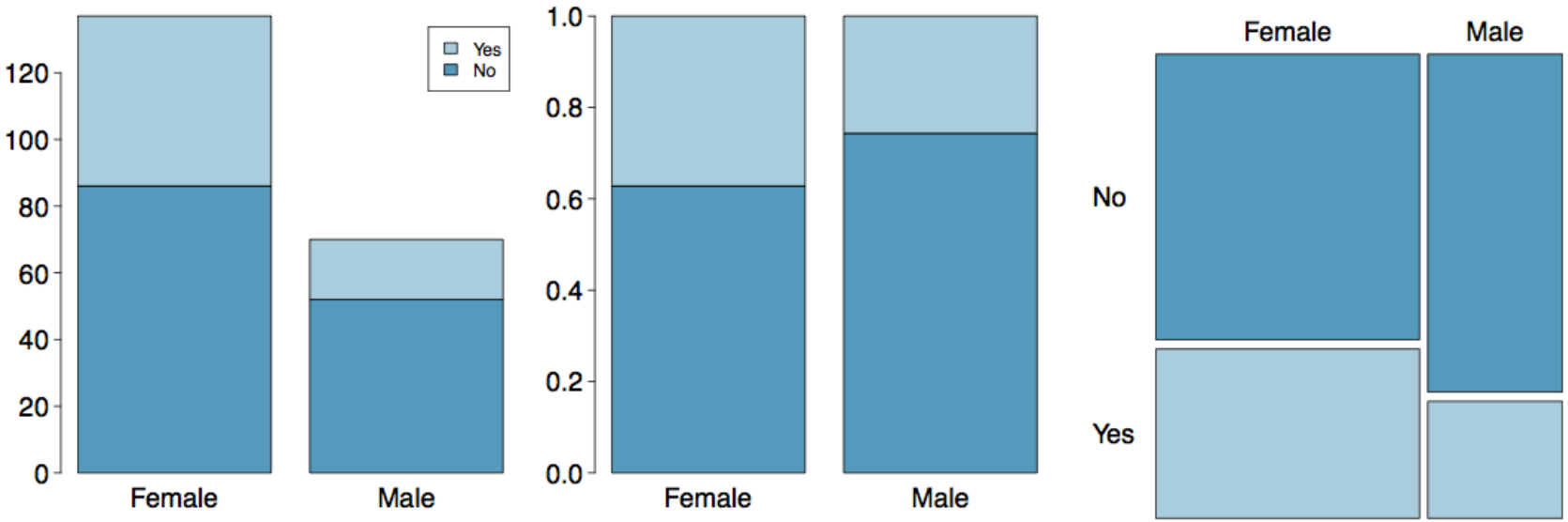


How are bar plots different than histograms?

Bar plots are used for displaying distributions of categorical variables, while histograms are used for numerical variables. The x-axis in a histogram is a number line, hence the order of the bars cannot be changed, while in a bar plot the categories can be listed in any order (though some orderings make more sense than others, especially for ordinal variables.)

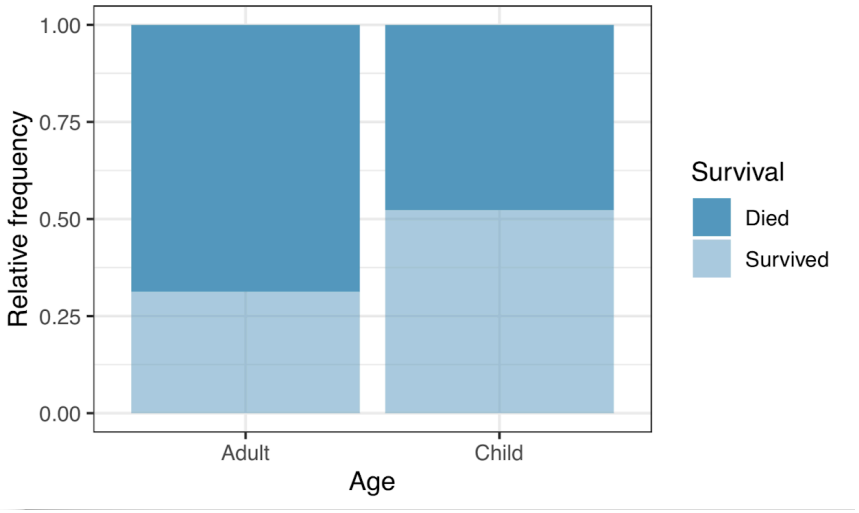
Segmented Bar and Mosaic Plots

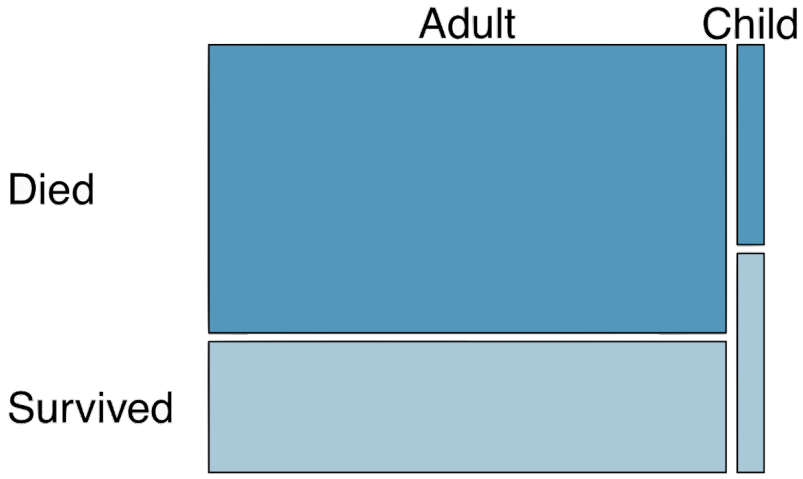
What are the differences between the three visualizations shown below?



Mosaic plots

What are the differences between the two visualizations shown below?

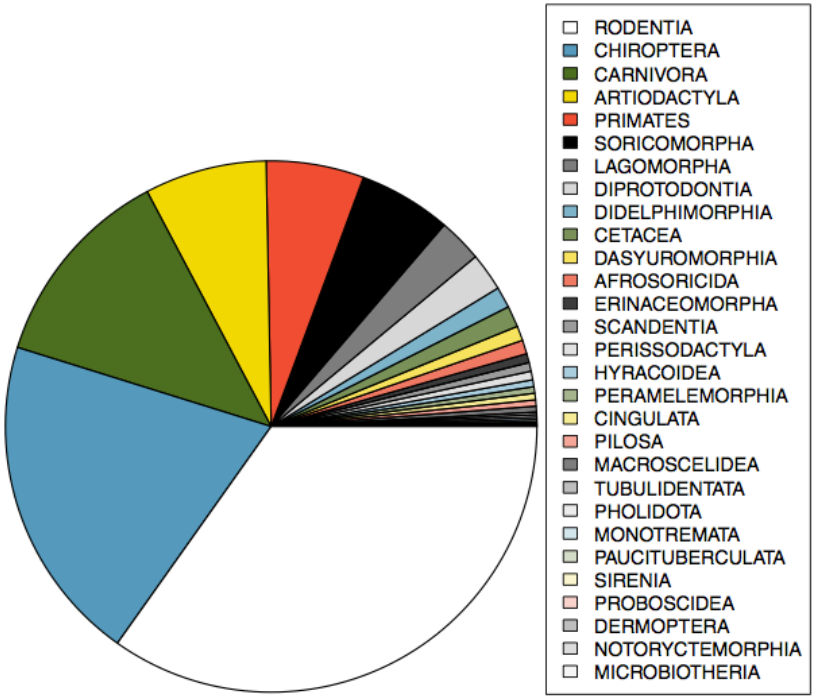




Pie Charts

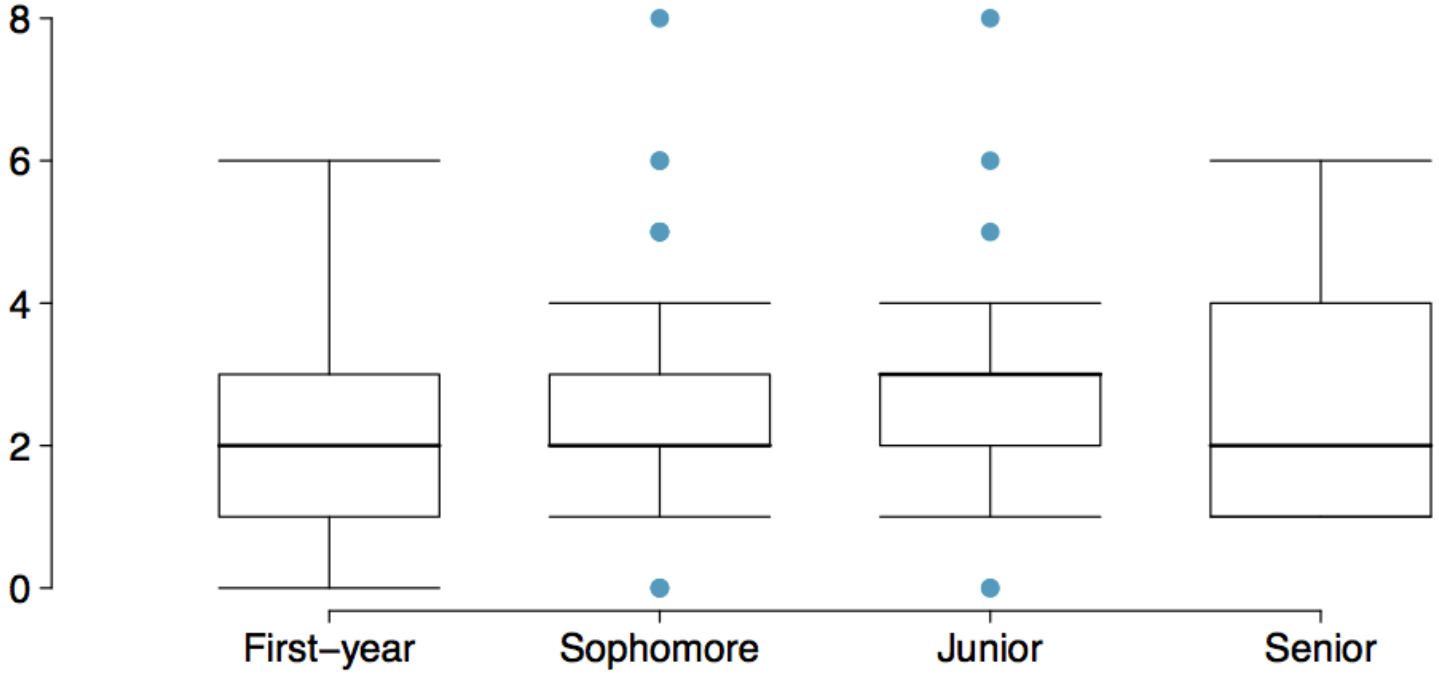
Can you tell which order encompasses the largest percentage of mammal species?

http://www.bucknell.edu/msw3



Comparing Numerical Data Across Groups

Does there appear to be a relationship between class year and number of clubs students are in?



Case Study:
Malaria Vaccine

Malaria Vaccine

- Let's consider a study on a new malaria vaccine called PfSPZ
- Volunteer patients were randomized into one of two experiment groups: 14 patients received an experimental vaccine and 6 patients received a placebo vaccine.
- Nineteen weeks later, all 20 patients were exposed to a drug-sensitive malaria virus strain; the motivation of using a drug-sensitive strain of virus here is for ethical considerations, allowing any infections to be treated effectively

Is this an observational study or an experiment?

Data

At a first glance, does there appear to be a relationship between promotion and gender?

Data

At a first glance, does there appear to be a relationship between vaccine and infection?

		outcome		
		infection	no infection	Total
treatment	vaccine	5	9	14
	placebo	6	0	6
	Total	11	9	20

Figure 2.29: Summary results for the malaria vaccine experiment.

% of treatment group got infected: 5 / 14 = 0.357

% of control group got infected: 6 / 6 = 1.000

Practice

We saw a difference of almost 65% (64.3% to be exact) between the infection rate of treatment group(with vaccine) and the control group. Based on this information, which of the below is true?

A. If we were to repeat the experiment we will definitely see that more participants from treatment group infected. This was a fluke.
B. Infection is dependent on vaccine, vaccine could reduce the infection rate, and hence the vaccine is effective.
C. The difference in the infection rate between the groups is due to chance, this is not evidence of effectiveness of vaccine against malaria.
D. The participants in the control group are less heathy than the participants in the treatment group, and this is the reason for the difference in infection rate.

Practice

We saw a difference of almost 65% (64.3% to be exact) between the infection rate of treatment group(with vaccine) and the control group. Based on this information, which of the below is true?

A. If we were to repeat the experiment we will definitely see that more participants from treatment group infected. This was a fluke.
B. Infection is dependent on vaccine, vaccine could reduce the infection rate, and hence the vaccine is effective. Maybe
C. The difference in the infection rate between the groups is due to chance, this is not evidence of effectiveness of vaccine against malaria. Maybe
D. The participants in the control group are less heathy than the participants in the treatment group, and this is the reason for the difference in infection rate.

Two Competing Claims

1. “There is nothing going on.”

   Promotion and gender are *independent*, no gender discrimination, observed difference in proportions is simply due to chance.

   → Null Hypothesis

Two Competing Claims

1. “There is nothing going on.”

   Infection rate and vaccine are *independent*—vaccine is not effective —observed difference in proportions is simply due to chance.

   → Null Hypothesis

2. “There is something going on.”

   Infection rate and vaccine are *dependent*—vaccine is actually effective —observed difference in proportions is not due to chance.

   → Alternative Hypothesis

A Trial as a Hypothesis Test

Hypothesis testing is very much like a court trial.

- H_0 : Defendant is innocent
  H_A : Defendant is guilty
- We then present the evidence - collect data.
- Then we judge the evidence - “Could these data plausibly have happened by chance if the null hypothesis were true?"
  - If they were very unlikely to have occurred, then the evidence raises more than a reasonable doubt in our minds about the null hypothesis.
- Ultimately we must make a decision. How unlikely is unlikely?

Image from http://www.nwherald.com/_internal/cimg!0/oo1il4sf8zzaqbboq25oevvbg99wpot

A Trial as a Hypothesis Test (cont.)

- If the evidence is not strong enough to reject the assumption of innocence, the jury returns with a verdict of “not guilty".
  - The jury does not say that the defendant is innocent, just that there is not enough evidence to convict.
  - The defendant may, in fact, be innocent, but the jury has no way of being sure.
- **Said statistically, we fail to reject the null hypothesis**.
  - We never declare the null hypothesis to be true, because we simply do not know whether it's true or not.
  - **Therefore we never “accept the null hypothesis”**.

A Trial as a Hypothesis Test (cont.)

- In a trial, the burden of proof is on the prosecution.
- In a hypothesis test, **the burden of proof is on the unusual claim.**
- The null hypothesis is the ordinary state of affairs (the status quo), so it's the alternative hypothesis that we consider unusual and for which we must gather evidence.

Recap: Hypothesis Testing Framework

- We start with a *null hypothesis* (H_0) that represents the status quo.
- We also have an *alternative hypothesis* (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation (today) or theoretical methods (later in the course).
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

Simulating the experiment...

... under the assumption of independence, i.e. leave things up to chance.

If results from the simulations based on the *chance model* look like the data, then we can determine that the difference between the proportions of promoted files between males and females was simply *due to chance* (promotion and gender are independent).

If the results from the simulations based on the chance model do not look like the data, then we can determine that the difference between the proportions of promoted files between males and females was not due to chance, but *due to an actual effect of gender* (promotion and gender are dependent).

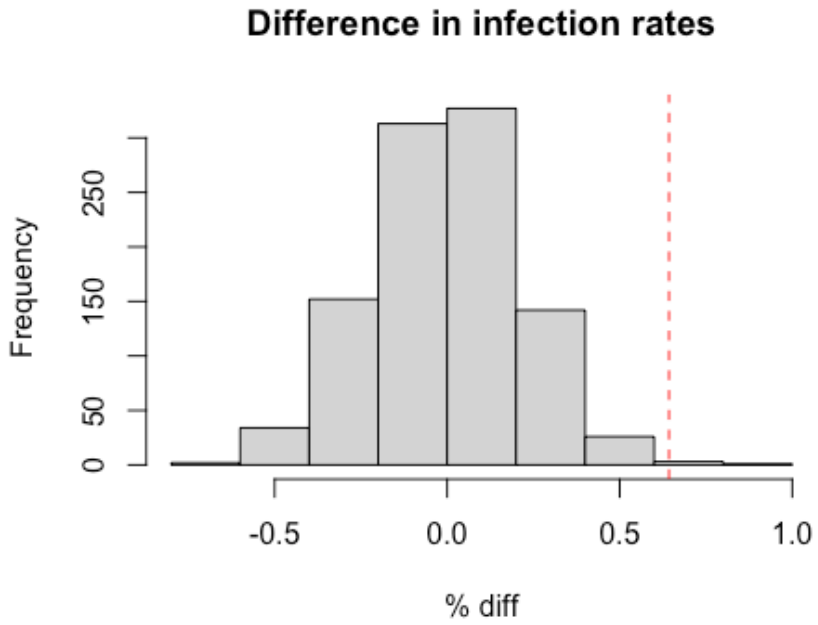
Simulations Using Software

In reality, we use software to generate the simulations.

```
nexp = 1000
res_mal = matrix(0, nrow = nexp, ncol = 3)
colnames(res_mal) = c("trt", "control","diff")
for (i in 1:nexp){
  set.seed(1234+i)
  infected = sample(1:20,11, replace = FALSE)
  trt = sample(1:20,10)
  n = sum(infected %in% trt)
  res_mal[i,1] = n/10
  res_mal[i,2] = (11-n)/10
  res_mal[i,3] = res_mal[i,1]-res_mal[i,2]
}
```

Simulations Using Software

In reality, we use software to generate the simulations. The histogram below shows the distribution of simulated differences in promotion rates based on 1000 simulations.



Practice

Do the results of the simulation you just ran provide convincing evidence that the vaccine is effective, i.e. dependence between the vaccination and infection rate?

A. No, the data do not provide convincing evidence for the alternative hypothesis, therefore we can't reject the null hypothesis of independence between the vaccination and infection rate. The observed difference between the two proportions was due to chance.

B. Yes, the data provide convincing evidence for the alternative hypothesis that the vaccine is effective against the malaria. The observed difference between the two proportions was due to a real effect of vaccination.

Practice

Do the results of the simulation you just ran provide convincing evidence that the vaccine is effective, i.e. dependence between the vaccination and infection rate?

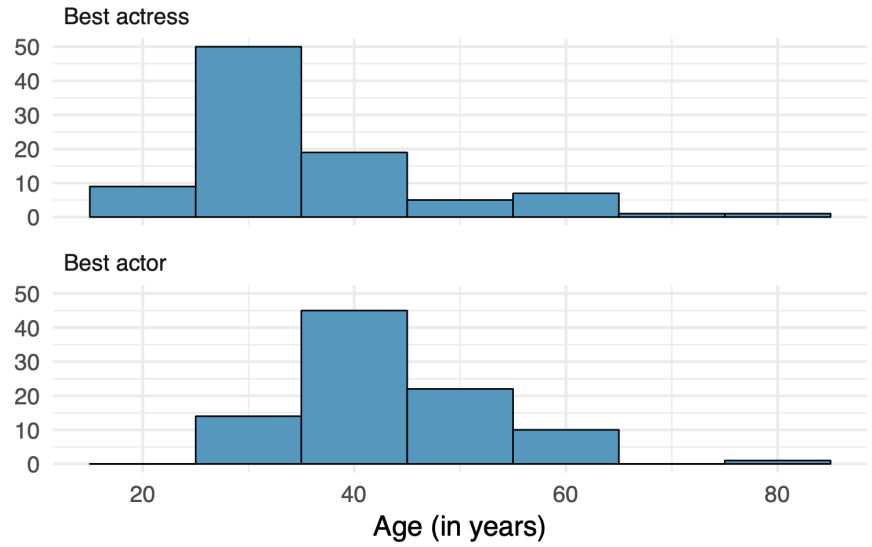
A. No, the data do not provide convincing evidence for the alternative hypothesis, therefore we can't reject the null hypothesis of independence between the vaccination and infection rate. The observed difference between the two proportions was due to chance.

B. Yes, the data provide convincing evidence for the alternative hypothesis that the vaccine is effective against the malaria. The observed difference between the two proportions was due to a real effect of vaccination.

Let’s discuss!

Oscar Winners

The first Oscar awards for best actor and best actress were given out in 1929. The histograms below show the age distribution for all of the best actor and best actress winners from 1929 to 2018. Summary statistics for these distributions are also provided. Compare the distributions of ages of best actor and actress winners

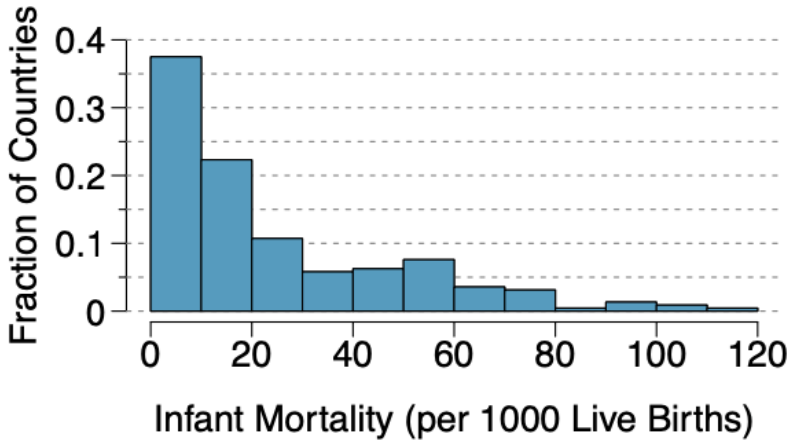


	Best Actress
Mean	36.2
SD	11.9
n	92

	Best Actor
Mean	43.8
SD	8.83
n	92

Infant mortality

The infant mortality rate is defined as the number of infant deaths per 1,000 live births. This rate is often used as an indicator of the level of health in a country. The relative frequency histogram below shows the distribution of estimated infant death rates for 224 countries for which such data were available in 2014.

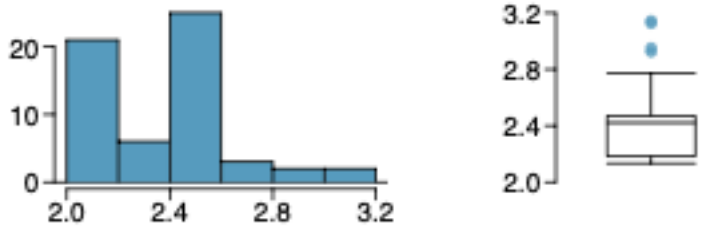


(a) What can you observe from the above histogram?(skewness, mean, median..)

(b) Would you expect the mean of this data set to be smaller or larger than median? Explain your reasoning.

Marathon winners

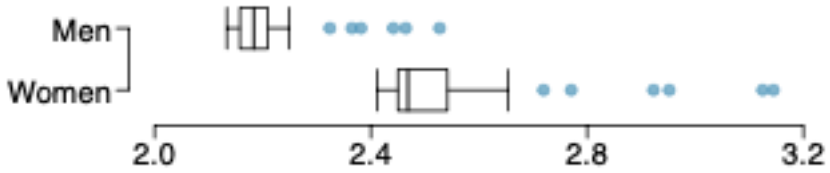
The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1970 and 1999.



(a) What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?

(b) What may be the reason for the bimodal distribution? Explain.

(c) Compare the distribution of marathon times for men and women based on the box plot shown below.



Tomorrow is R Session!

- Don’t forget to bring laptop with you.
- Office hour from **7pm** via Zoom.
  - **This week, make sure to participate at least one office hour!!**
  - 5 out of 50 participation pts