ECO239

Week 4

Describing Data Graphically

Considering Categorical Data

*NOTE: file names listed in this lecture note are different from what we used in the class.
You can use highgpa.csv file to try out the codes.

Describing Data Graphically

Options for Categorical Variables

- Frequency Distribution Table
- Contingency Table
- Bar Chart
- Pie Chart

* Always consider what kind of graphs/tables describe your data the best, answer your question the best.

Tables/Graphs for Categorical Data

Frequency Distribution Table

⇒Summarize Data by Category

Category	Frequency

R-exercise

- Generate Frequency Distribution Table using R

#Data = city_sec2.csv

```
sortcity=sort(table(city_sec1$city),decreasing=T)
sortcity

sortedcity=cbind(sortcity)
sortedcity

attach(gpa_sec1)
sortfam=sort(table(family),decreasing=T)
```

TRY WITH other variable in data = gpa_sec2.csv

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*.

The contingency table below shows the distribution of students' genders and whether or not they are looking for a spouse while in college.

		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

- What kind of contingency table shall we create using our data?

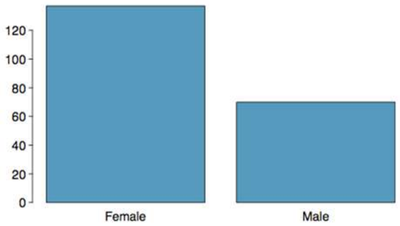
Contingency Table

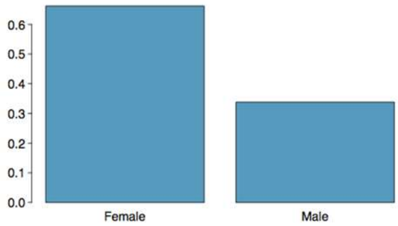
```
# data = gpa_sec1.csv#

attach(gpa_sec1)
with(gpa_sec1,table(gender,partner))


#with nicer table format with chi-square test and fractions info.#
#need to install gmodels package first#

library(gmodels)
with(gpa_sec1,CrossTable(gender,partner))
```

TRY IT WITH OTHER VARIABLES (select a meaningful pair).

Bar Plots

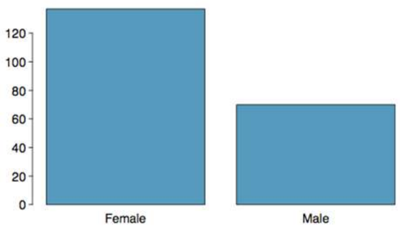
A *bar plot* is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.

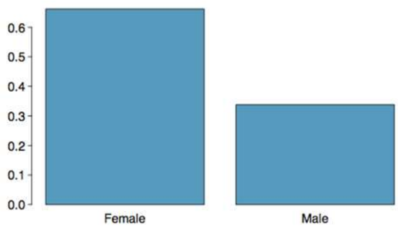




Bar Plots

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*.

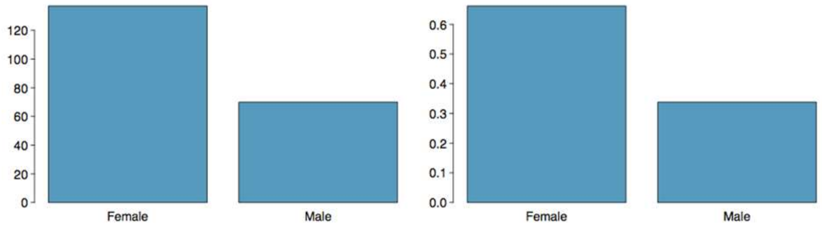




How are bar plots different than histograms?

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.)

Bar Plot

```
#Data=gpa_sec1.csv

attach(gpa_sec1)
cgender=table(gender)
barplot(cgender,main="ECO239(2) Gender",xlab="Gender, 0: Male,
1:Female")
```

TRY IT WITH OTHER VARIABLE.

Other options for Bar Plots

```
#Horizontal version#
barplot(cgender, main="ECO239(2) Gender", horiz=TRUE,
    names.arg=c("Male", "Female"))


# Stacked Bar Plot with Colors and Legend
cgender_partner=(table(partner,gender))
barplot(cgender_partner, main="Gender vs. Partner",
   xlab="Gender", col=c("blue","red"),
  legend = rownames(cgender_partner))
```

Choosing the Appropriate Proportion

Does there appear to be a relationship between gender and whether the student is looking for a spouse in college?

		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

Choosing the Appropriate Proportion

Does there appear to be a relationship between gender and whether the student is looking for a spouse in college?

		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

To answer this question we examine the row proportions:

Choosing the Appropriate Proportion

Does there appear to be a relationship between gender and whether the student is looking for a spouse in college?

		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

To answer this question we examine the row proportions:

- % Females looking for a spouse: 51 / 137 ~ 0.37

Choosing the Appropriate Proportion

Does there appear to be a relationship between gender and whether the student is looking for a spouse in college?

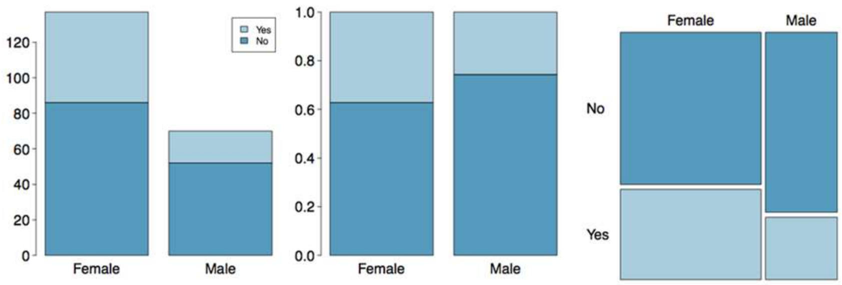
		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

To answer this question we examine the row proportions:

- % Females looking for a spouse: 51 / 137 ~ 0.37
- % Males looking for a spouse: 18 / 70 ~ 0.26

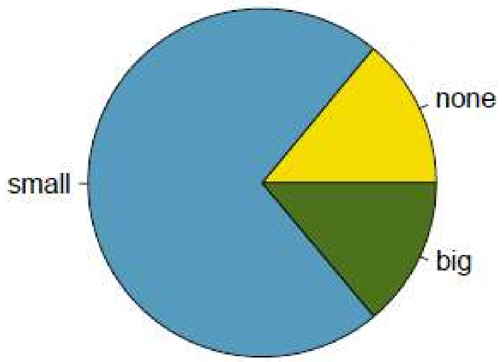
Segmented Bar and Mosaic Plots

What are the differences between the three visualizations shown below?



Graphs for Categorical Data
Pie Chart

- Calculate % of each category and place them in a PIE according to their share.



Pie Chart

```
#Data = city_sec2.csv

# Simple Pie Chart
sortcity=sort(table(city_sec2$city),decreasing=T)
sortcity
sorted=cbind(sortcity)
Sorted

lbls=c("ankara","istanbul","antalya","bursa","amasya","artvin","aydin","bitlis","cankli","cyprus",
"edirne","kars","konya","sakarya","samsun","shymken","sivas","tekirdag","tokat")
slices=sortcity
pct <- round(slices/sum(slices)*100)
lbls<-paste(lbls,pct)
lbls <- paste(lbls,"%",sep="")
pie(slices,labels = lbls, col=rainbow(length(lbls)),main="Pie Chart of Cities")
```

TRY IT WITH Variables in GPA_SEC2 data set.

- When do you think the pie charts are useful???

=> When the share of each category is your interest.

Examining Numerical Data

Tables and Graphs for Numerical Data

Options for Numerical Variables

- Frequency Distribution & Cumulative Distribution
- Histogram
- Scatter Plot
- Box Plot

Table for Numerical Data

- Frequency Distributions

Interval (Class)	Frequency

Procedure for Creating Frequency Distribution Table for Numerical Variables.

1. Find the range of the variable by sorting in ascending order.
2. Round the min (downward) and max (upward) values as necessary.
3. Decide the number of intervals (k).
4. Calculate the interval width as W = (rmax – rmin) / k.
5. Identify [ ) (left closed – right open) intervals.
6. Count the number of observations belonging to each interval.

Create Frequency Distribution Table for study

study
10
3
4
3.5
7
8
3
1
2
15
8
3
0
2
7
2
1
4
10
0
9
5
4
10
2
0
4
15

Frequency Distribution Table for Numerical Variable

```
attach(gpa_sec2)

rstudy=range(study)
rstudy

breaks=seq(0,18,by=3)
breaks

study.cut = cut(study, breaks, right=FALSE)
study.freq = table(study.cut)

cbind(study.freq)
```

Cumulative Distribution

Include

✓Frequency (Count)

✓Relative Frequency (% of Count)

✓Cumulative Frequency (Cumulative Count)

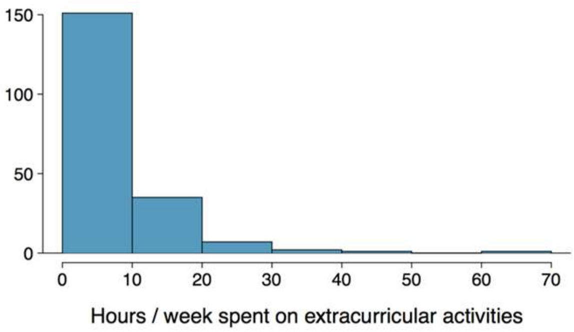
✓Relative Cumulative Frequency (% of Cumulative Count)

Create Relative/Cumulative Distribution Table for study

```
# Cumulative Frequency Table#
study.cumsum=cumsum(study.freq)
cbind(study.cumsum)

# Relative Frequency Table#
study.relfreq=study.freq/nrow(gpa_sec2)
cbind(study.relfreq)

# Relative Cumulative Frequency Table#
study.relcum=cumsum(study.relfreq)
cbind(study.relcum)
```

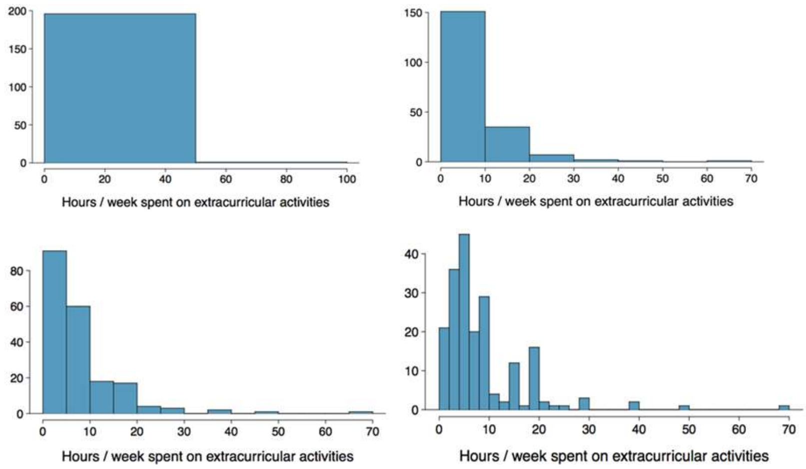
Histograms - Extracurricular Hours

- 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.



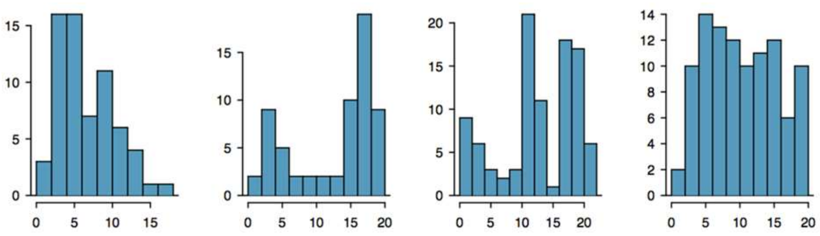
Bin Width

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



Shape of a Distribution: Modality

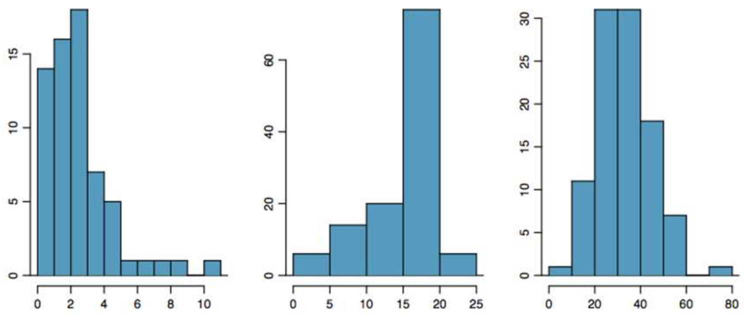
Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?



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*?

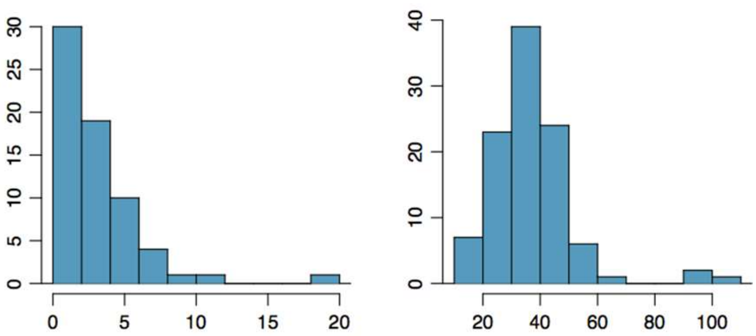


Note: Histograms are said to be skewed to the side of the long tail.

Right Skewed   Left Skewed   Symmetric

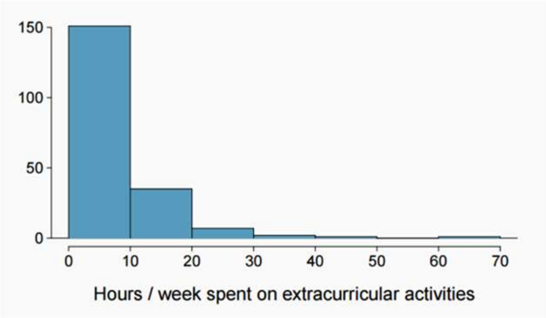
Shape of a Distribution: Unusual Observations

Are there any unusual observations or potential *outliers*?



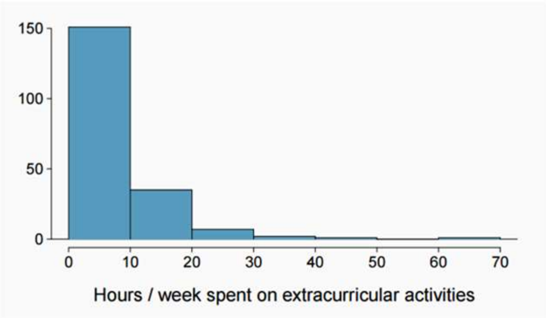
Extracurricular activities

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.

Histogram

```
hist(gpa_sec2$coffee,breaks=10,col=blues9, xlab="cups of coffee/week",
main="Histgram for coffee consumption")
```

TRY IT WITH OTHER VARIABLE

Commonly observed shapes of distributions

Modality

Commonly observed shapes of distributions

Modality

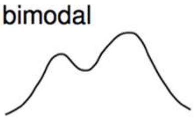
unimodal

Commonly observed shapes of distributions

Modality

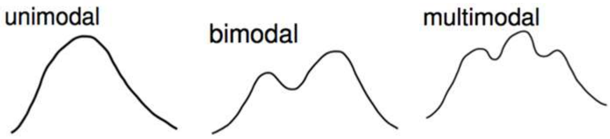
unimodal

bimodal



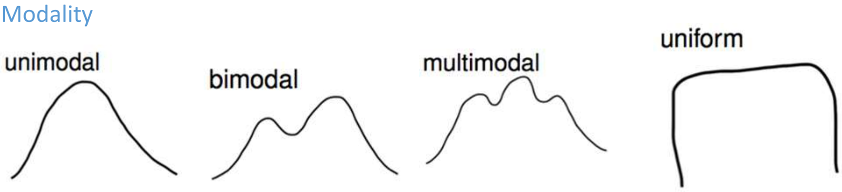
Commonly observed shapes of distributions

Modality



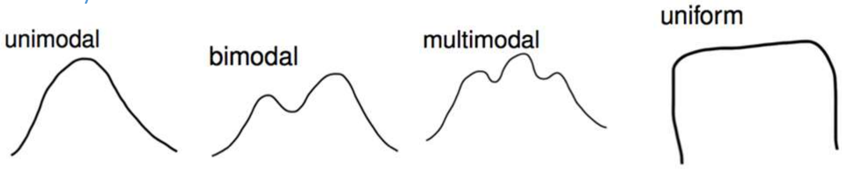
Commonly observed shapes of distributions

Modality



Commonly observed shapes of distributions

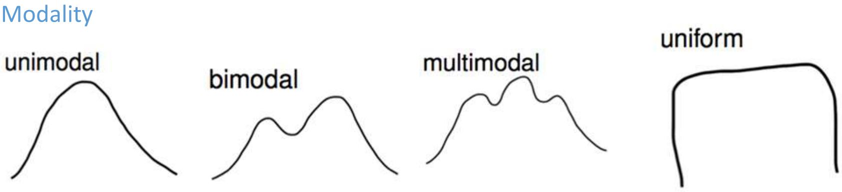
Modality



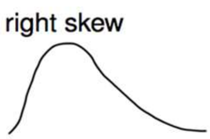
Skewness

Commonly observed shapes of distributions

Modality

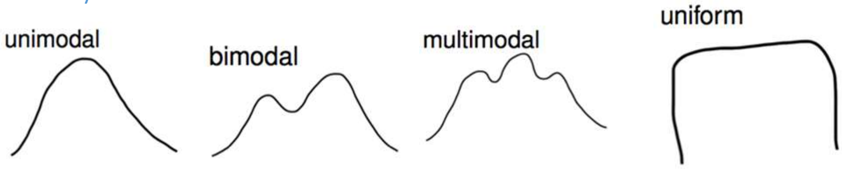


Skewness

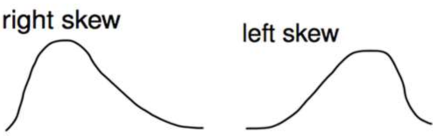


Commonly observed shapes of distributions

Modality

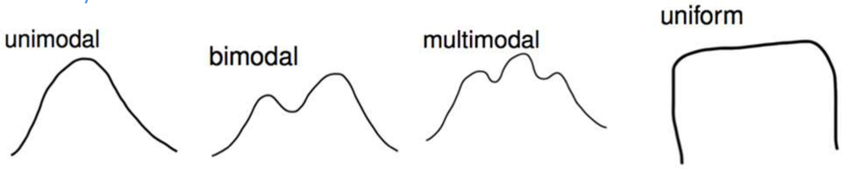


Skewness

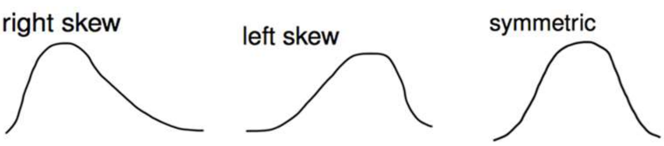


Commonly observed shapes of distributions

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

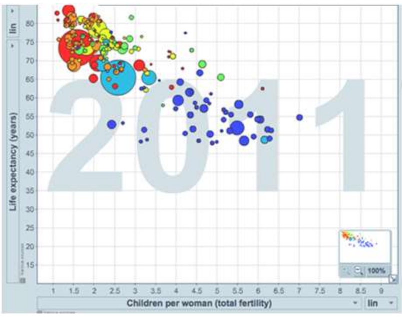
(d) birthdays of classmates (day of the month)

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?



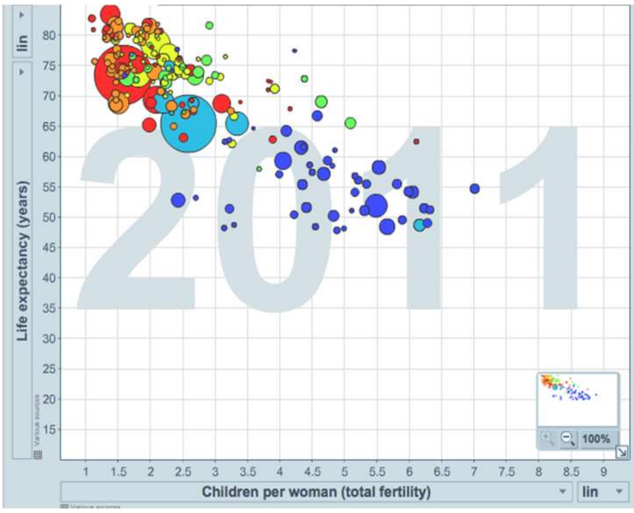
http://www.gapminder.org/world

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.



http://www.gapminder.org/world

Scatter Plot

```
# Relationship between Study Hours and GPA#

plot(study, gpa, main="GPA vs. Study Hours",
    xlab="Study Hours ", ylab="GPA ", pch=19)
```

TRY IT WITH OTHER VARIABLE.

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.

Variance

Variance is roughly the average squared deviation from the mean.

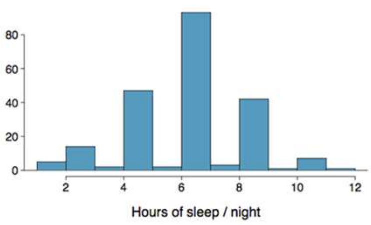
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

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.

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

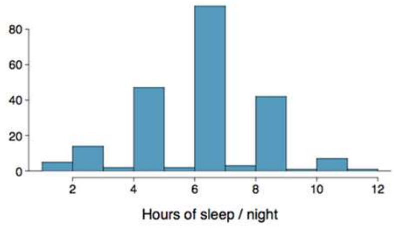
Standard Deviation

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:

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



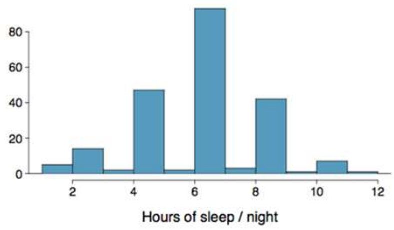
Standard Deviation

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

Summary Statistics in R

min, Q1, median, mean, Q3, max for all variables included in the file.

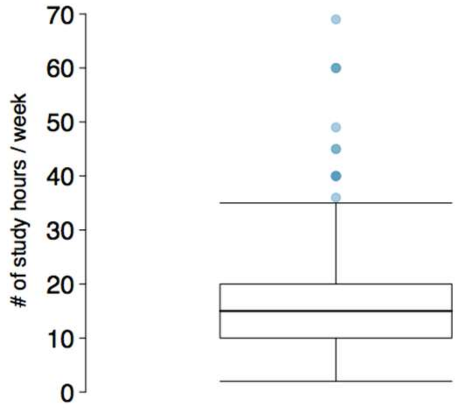
summary(gpa_sec2)

same summary for just one of the variables in gpa_sec2.csv file.

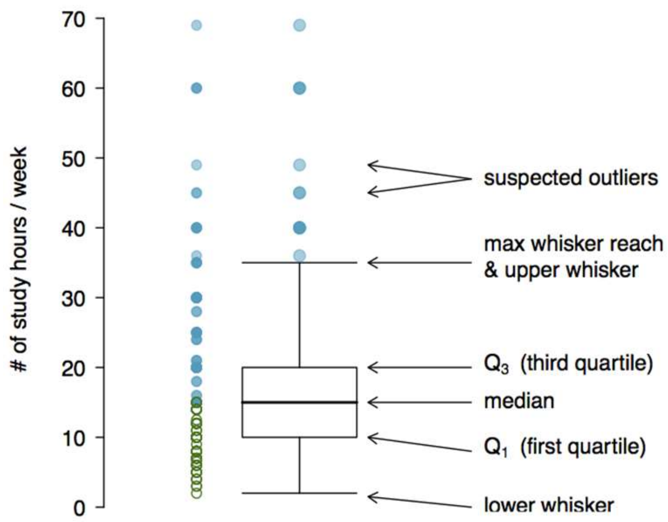
summary(gpa_sec2$gpa)

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.

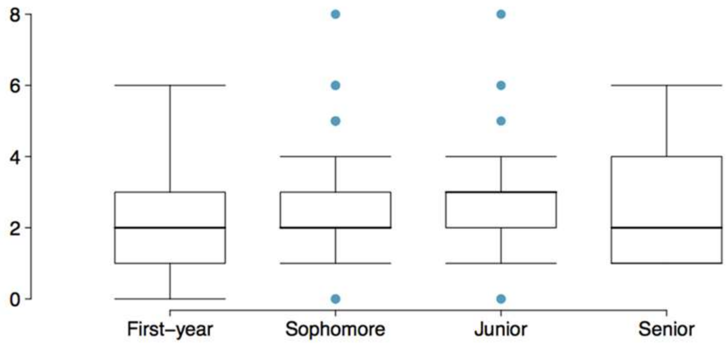
Box Plot

```
attach(gpa_sec2)
boxplot(gpa~gender,data=gpa_sec2,names=c("male","female"),
main="GPA by gender",    xlab="Gender", ylab="GPA")
```

TRY IT WITH OTHER VARIABLE.

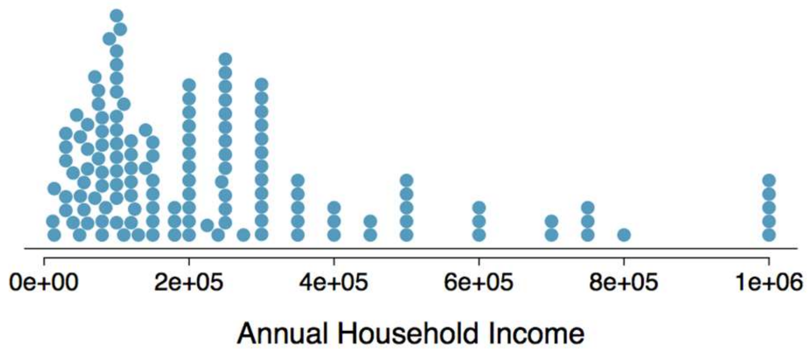
Comparing Numerical Data Across Groups

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

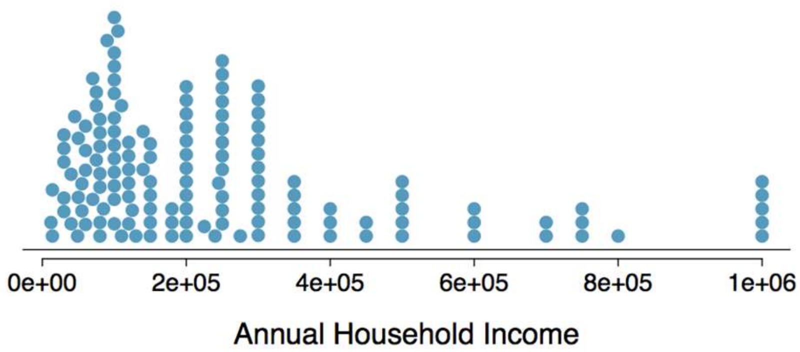


Extreme Observations

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? What if the smallest value was replaced with \$10 million?



Robust Statistics



	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original data	190K	200K	245K	226K
move largest to \$10 million	190K	200K	309K	853K
move smallest to \$10 million	200K	200K	316K	854K

Robust Statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

Robust Statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

If you would like to estimate the typical household income for a student, would you be more interested in the mean or median income?

Robust Statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

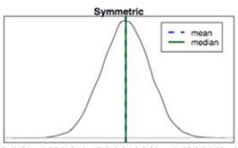
- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

If you would like to estimate the typical household income for a student, would you be more interested in the mean or median income?

Median

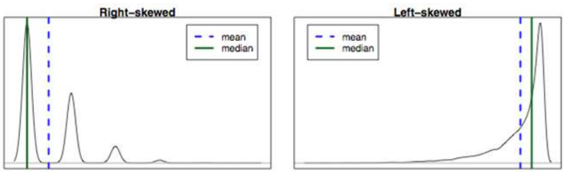
Mean vs. Median

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



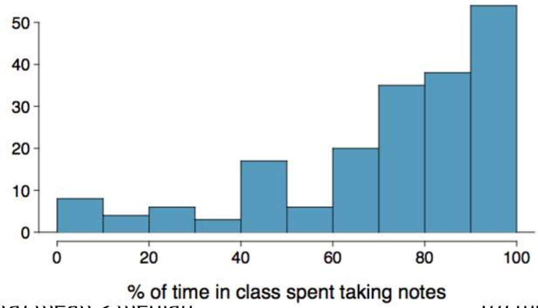
If the distribution is skewed or has extreme outliers, center is often defined as the median

- 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.?

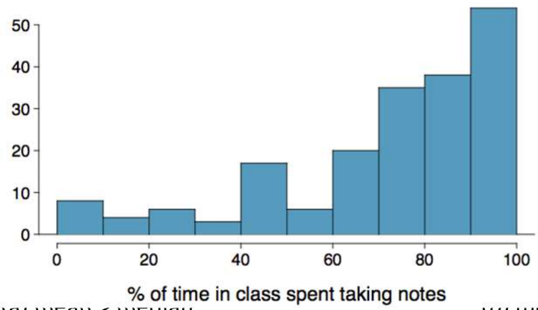


(a) mean > median (b) mean ~ median

(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

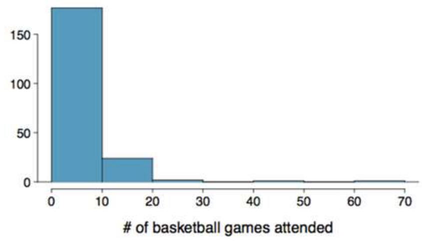
Extremely Skewed Data

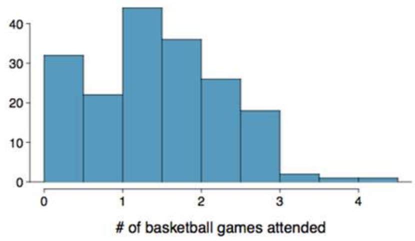
When data are extremely skewed, transforming them might make modeling easier. A common transformation is the *log transformation*.

Extremely Skewed Data

When data are extremely skewed, transforming them might make modeling easier. A common transformation is the log transformation.

The histograms on the left shows the distribution of number of basketball games attended by students. The histogram on the right shows the distribution of log of number of games attended.





Pros and Cons of Transformations

- Skewed data are easier to model with when they are transformed because outliers tend to become far less prominent after an appropriate transformation.

# of games	70	50	25	…
# of games	4.25	3.91	3.22	…

- However, results of an analysis might be difficult to interpret because the log of a measured variable is usually meaningless.

Pros and Cons of Transformations

- Skewed data are easier to model with when they are transformed because outliers tend to become far less prominent after an appropriate transformation.

# of games	70	50	25	…
# of games	4.25	3.91	3.22	…

- However, results of an analysis might be difficult to interpret because the log of a measured variable is usually meaningless.

What other variables would you expect to be extremely skewed?

Pros and Cons of Transformations

- Skewed data are easier to model with when they are transformed because outliers tend to become far less prominent after an appropriate transformation.

  # of games 70 50 25 …

  # of games 4.25 3.91 3.22 …

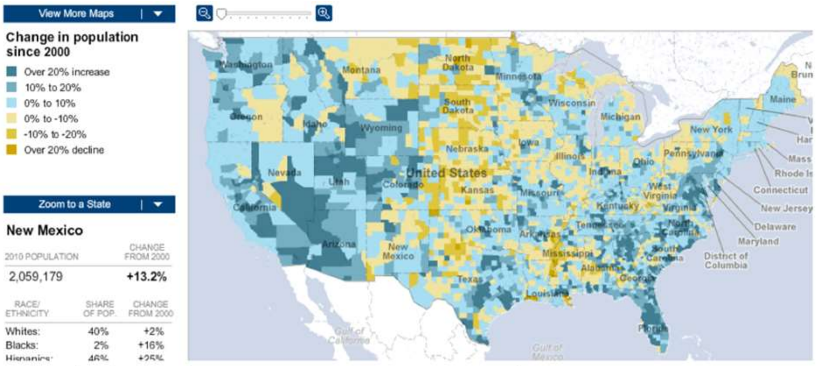
- However, results of an analysis might be difficult to interpret because the log of a measured variable is usually meaningless.

What other variables would you expect to be extremely skewed?

Salary, housing prices, etc.

Intensity Maps

What patterns are apparent in the change in population between 2000 and 2010?



Describing Data Graphically

Options for Categorical Variables

- Frequency Distribution Table
- Bar Chart
- Pie Chart

* Always consider what kind of graphs/tables describe your data the best, answer your question the best.

Tables/Graphs for Categorical Data

Frequency Distribution Table

⇒Summarize Data by Category

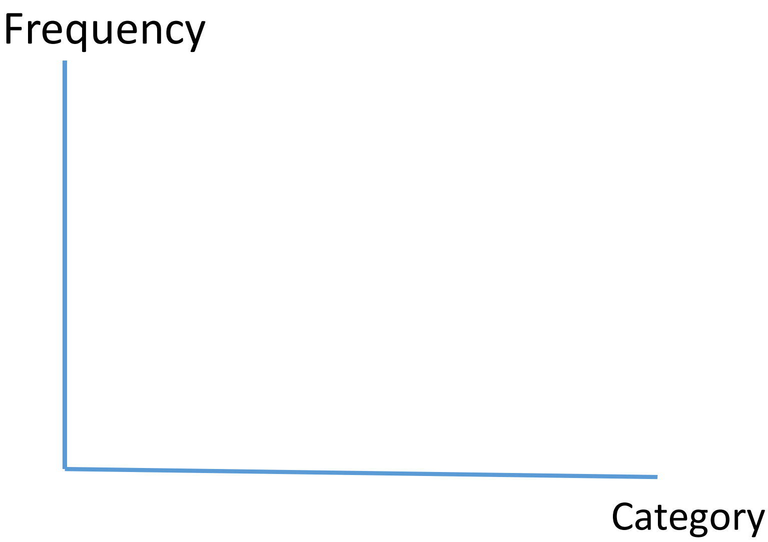
Category	Frequency

Frequency Distribution Table

e.g. Which City are you from?

City

Graph for Categorical Variable
Bar Chart



Bar Chart

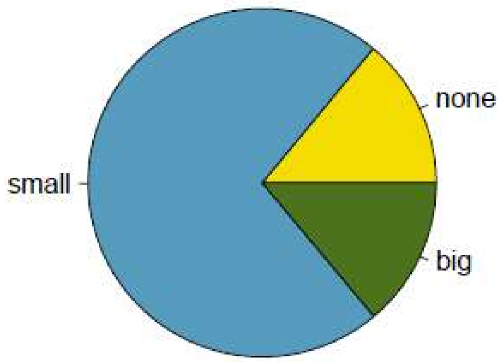
- City

Q: If you have data for 2 years, what do you do?

City

Graphs for Categorical Data
Pie Chart

- Calculate % of each category and place them in a PIE according to their share.



Pie Chart

- City

• When do you think the pie charts are useful???

=> When the share of each category is your interest.

R practice

• Download R Studio from the web.

Step1: Prepare data (save it as .csv)
Step2: Import data to R Studio
Step3: Create a bar chart.
Step4: Create a pie chart.

At home, practice with class$gender, class$partner variables.

Tables and Graphs for Numerical Data

Options for Numerical Variables

- Frequency Distribution & Cumulative Distribution
- Histogram
- Box Plot

Table for Numerical Data

- Frequency Distributions
- Finalscore

Interval (Class)	Frequency

Frequency Distribution Table: How to determine the classes?

Step 1: sort raw data in ascending order
(small-> large)
Step 2: Find the range of data (100-0 = 100)
Step 3: Determine the number of interval (classes) k
Step 4: Compute interval width, w
Step 5: Determine interval boundaries
Step 6: Count observations & assign to each interval.

RULES!

1. Intervals should have the same width "w".

$$w = \frac{\text{(Largest number – Smallest number)}}{\text{(\# of desired intervals,k)}}$$

2. Use at least 5, but no more than 15-20 intervals
3. Intervals NEVER overlap
4. Round up the interval width to get desirable interval endpoints.

Create Frequency Distribution Table for

- Finalscore

Cumulative Distribution

Include
✓Frequency (Count)

✓Relative Frequency (% of Count)

✓Cumulative Frequency (Cumulative Count)

✓Relative Cumulative Frequency (% of Cumulative Count)

Create Cumulative Distribution Table for

- Finalscore