MATH 1051H S61: Lecture #04

# Examining Numerical Data

Scatterplots

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

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

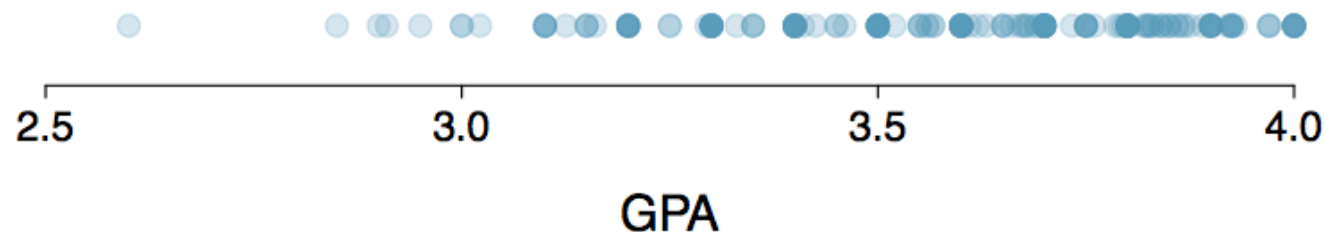They appear to be linearly and negatively associated: as fertility increases, life expectancy decreases.

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

The relationship changed over the years.
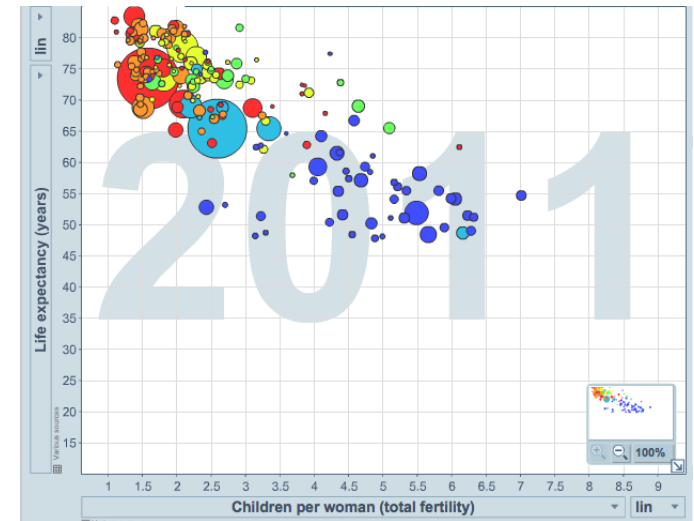
Dot Plots

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

How would you describe the distribution of GPAs in this data set?

Make sure to say something about the center, shape, and spread of the distribution.

Dot Plots and Mean

The **mean**, also called the average (marked with a triangle in the 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

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

where $x_1, x_2, \cdots, x_n$ 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.
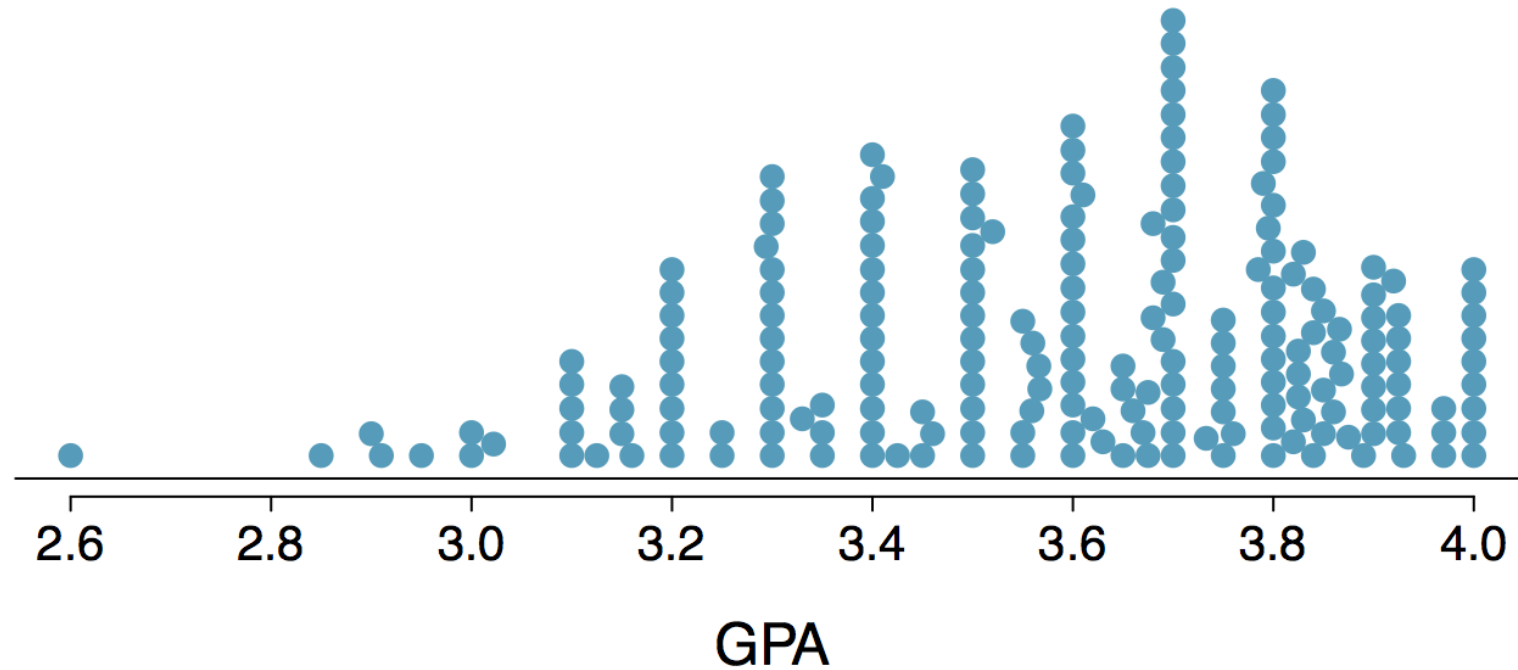
Calculating

I will never expect you to compute a mean without a calculator. In practice, we do this using R almost always. For example,

```
x <- c(1, 3, 8, 14, 21)
mean(x)
```

```
## [1] 9.4
```

Stacked Dot Plot

Higher bars represent areas where there are more observations, makes it a little easier to judge the center and the shape of the distribution.


GPA

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.

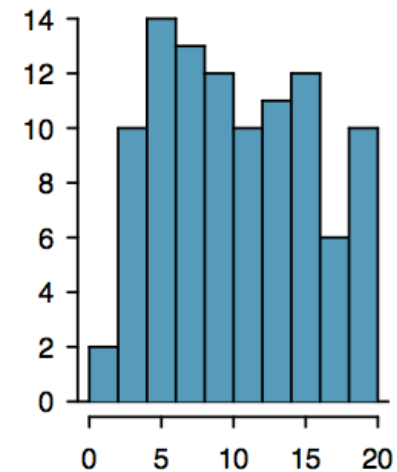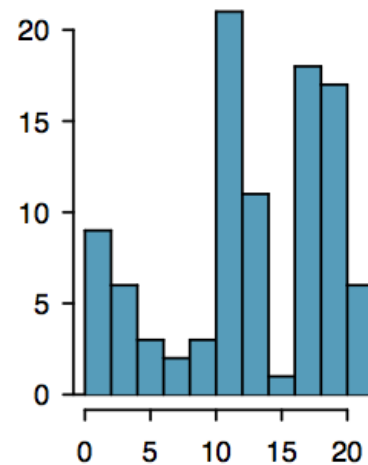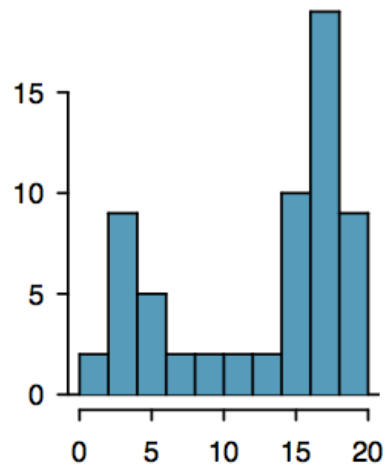Hours / week spent on extracurricular activities
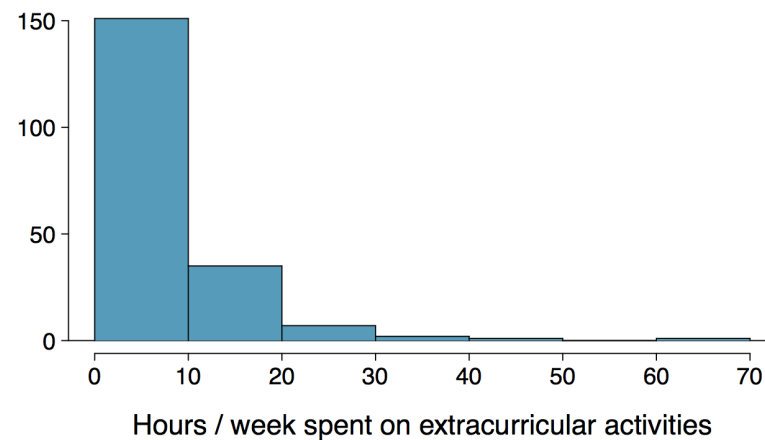
Bin Width

Which one(s) of these histograms are useful? Which reveal too much about the data? Which hide too much?
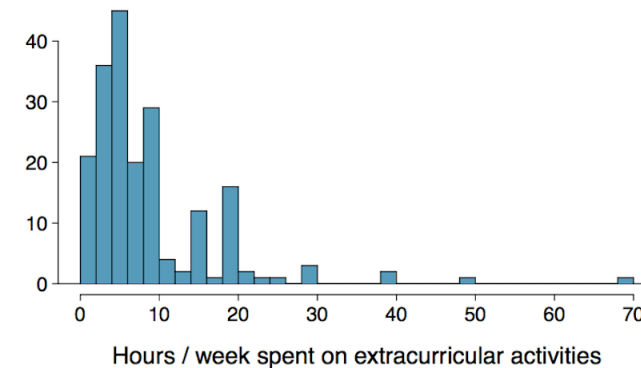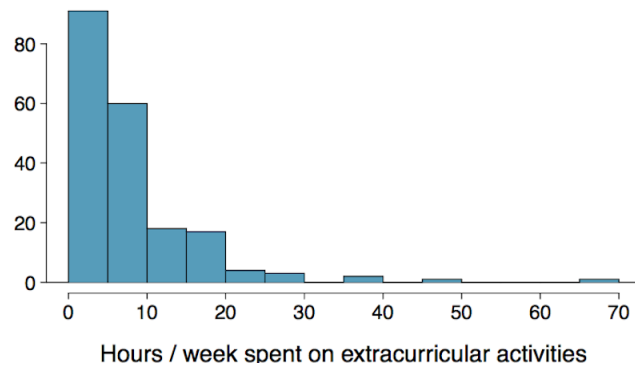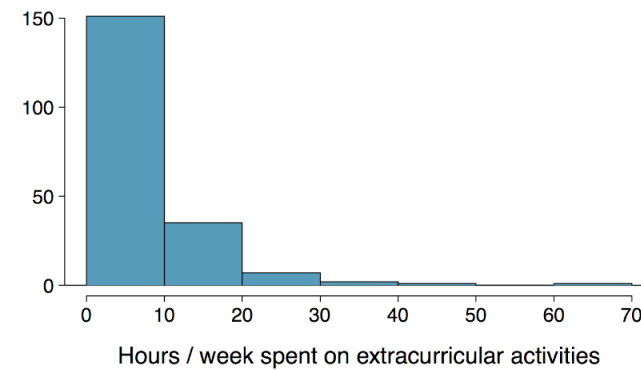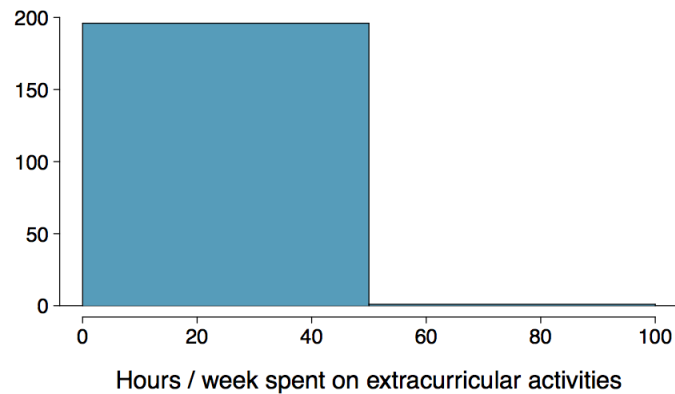
Shape of a Distribution: Modality

Does the histogram have a single prominent peak (**unimodal**), several prominent peaks (**bimodal/multimodal**), or no apparent peaks (**uniform**)?
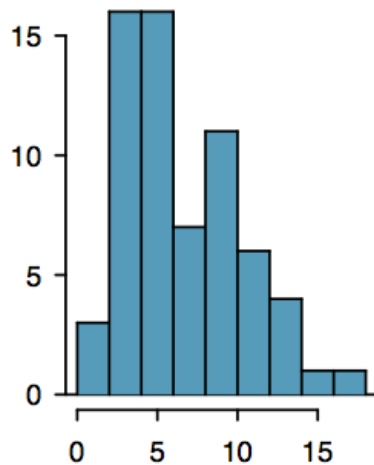
Note: In order to determine modality, step back and imagine a smooth curve over the histogram – imagine that the bars are wooden blocks and you drop a limp spaghetti over them, the shape the spaghetti would take could be viewed as a smooth curve.

Shape of a Distribution: Skewness

Is the histogram **right skewed**, **left skewed** or **symmetric**?
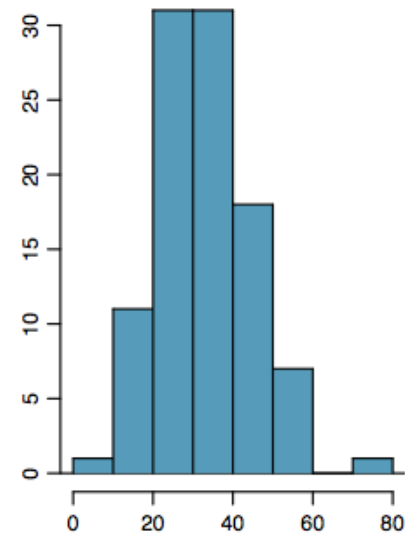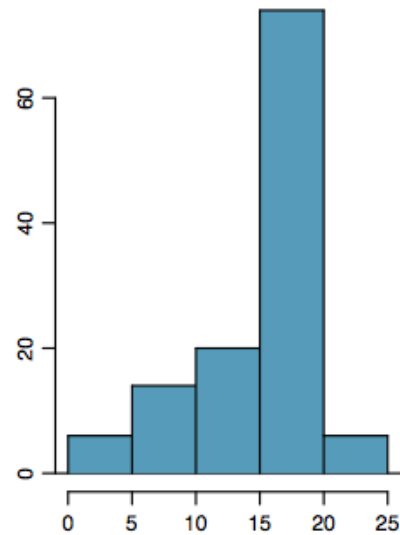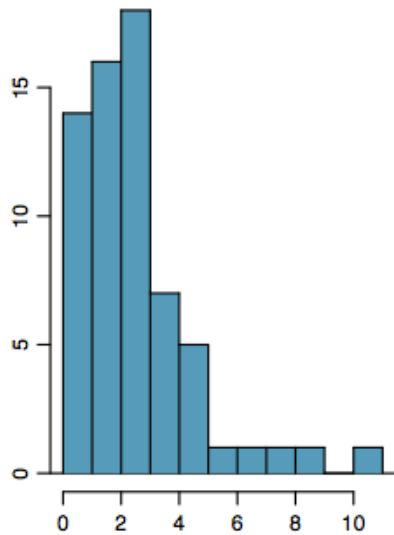
Histograms are said to be skewed to the side of the **long tail**.

Shape of a Distribution: Unusual Observations

Are there any unusual observations or potential **outliers**?
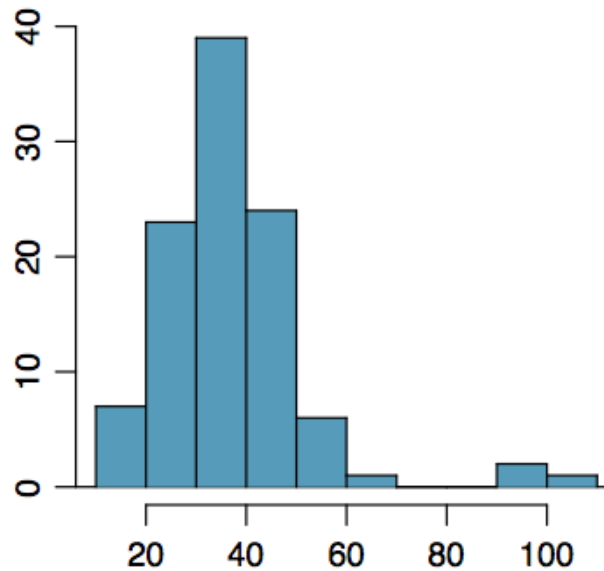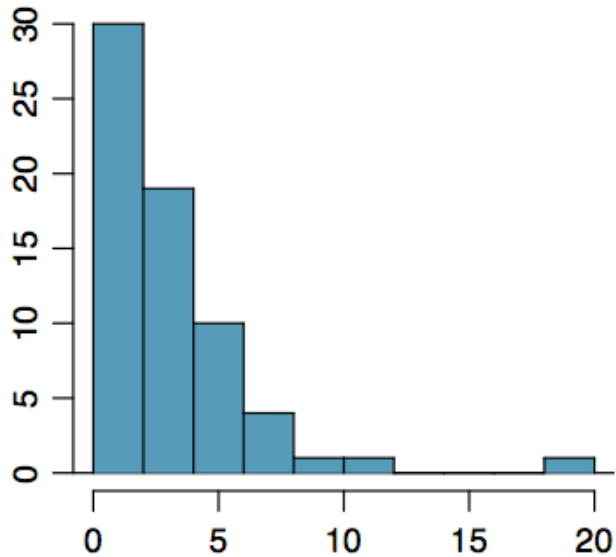
Commonly observed shapes of distributions

Modality

unimodal
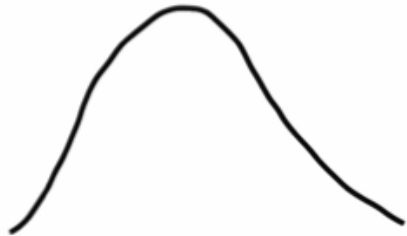
bimodal

multimodal

uniform

Skewness

right skew

left skew

symmetric
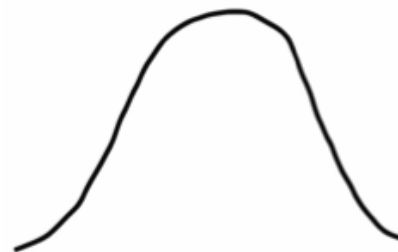
Practice

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

1. weights of adult females

2. salaries of a random sample of people from North Carolina

3. house prices

4. birthdays of classmates (day of the month)

Practice

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

1. weights of adult females

2. salaries of a random sample of people from North Carolina

3. house prices

4. *birthdays of classmates (day of the month)*

Application Activity: Shapes of Distributions

Sketch the expected distributions of the following variables:

- number of piercings

- scores on an exam

- IQ scores

Come up with a concise way (1-2 sentences) to teach someone how to determine the expected distribution of any variable.

Are you typical?

Variance

Variance is roughly the average squared deviation from the mean.

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

Hours of sleep / night

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

- The variance of the amount of sleep students get per night can be calculated as:

$$s^2 = \frac{1}{217-1} \left[(5 - 6.71)^2 + (9 - 6.71)^2 + \cdots + (7 - 6.71)^2 \right]$$

$$= 4.11 \text{ hours}^2$$
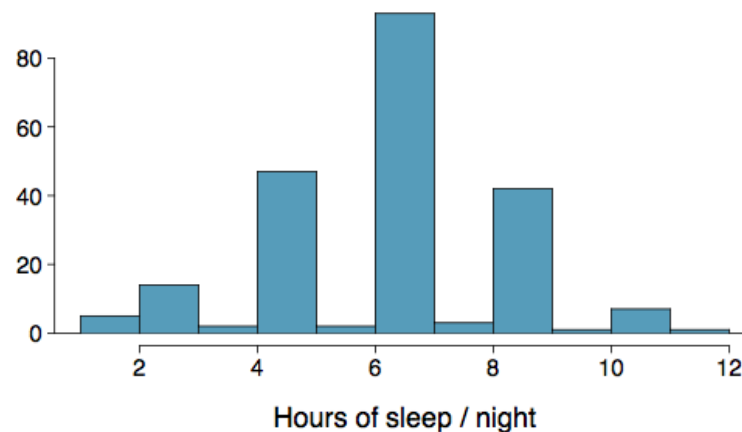
Variance (continued)

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.

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

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

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

We can see that all of the data are within 3 standard deviations of the mean of $\bar{x} = 6.17$.

Hours of sleep / night
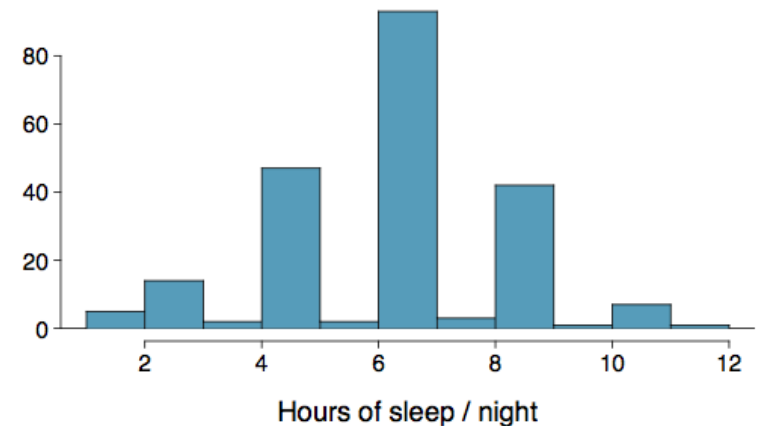
Calculating

I will never expect you to compute a variance or SD without a calculator. In practice, we do this using R almost always. For example,

```
x <- c(1, 3, 8, 14, 21)
var(x)
```

```
## [1] 67.3
```

```
sd(x)
```

```
## [1] 8.203658
```

```
sqrt(var(x))
```

```
## [1] 8.203658
```

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**.

Calculating

Medians can't really be computed with calculators, but we'll use R for this anyway
...

```
x <- c(1, 3, 8, 14, 21)
median(x)
```

```
## [1] 8
```

Percentile

A **percentile** is the the smallest value from an ordered list of numbers which is greater than or equal to that percentage of list elements.

Example: The 42nd percentile of the numbers $\{1, 2, 3, \cdots, 99, 100\}$ is 42.

It can become quite complicated when there aren't an even multiple of 100 items!

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.

$$IQR = Q3 - Q1$$

Calculating

Quantiles are actually hard to do with calculators, but R makes it easy.

```r
x <- sample(1:100, size = 20, replace = TRUE)
quantile(x, probs = c(0.4, 0.6))
```
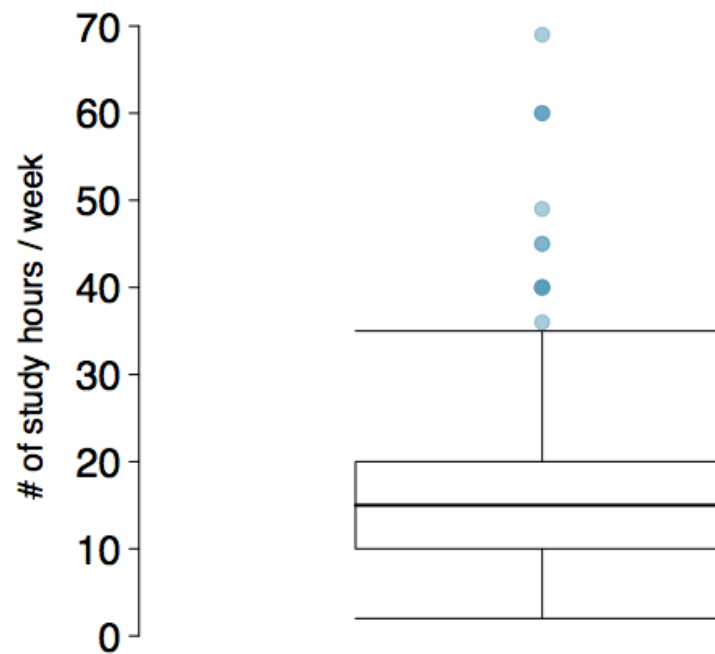
```
## 40% 60%
##  61  72
```

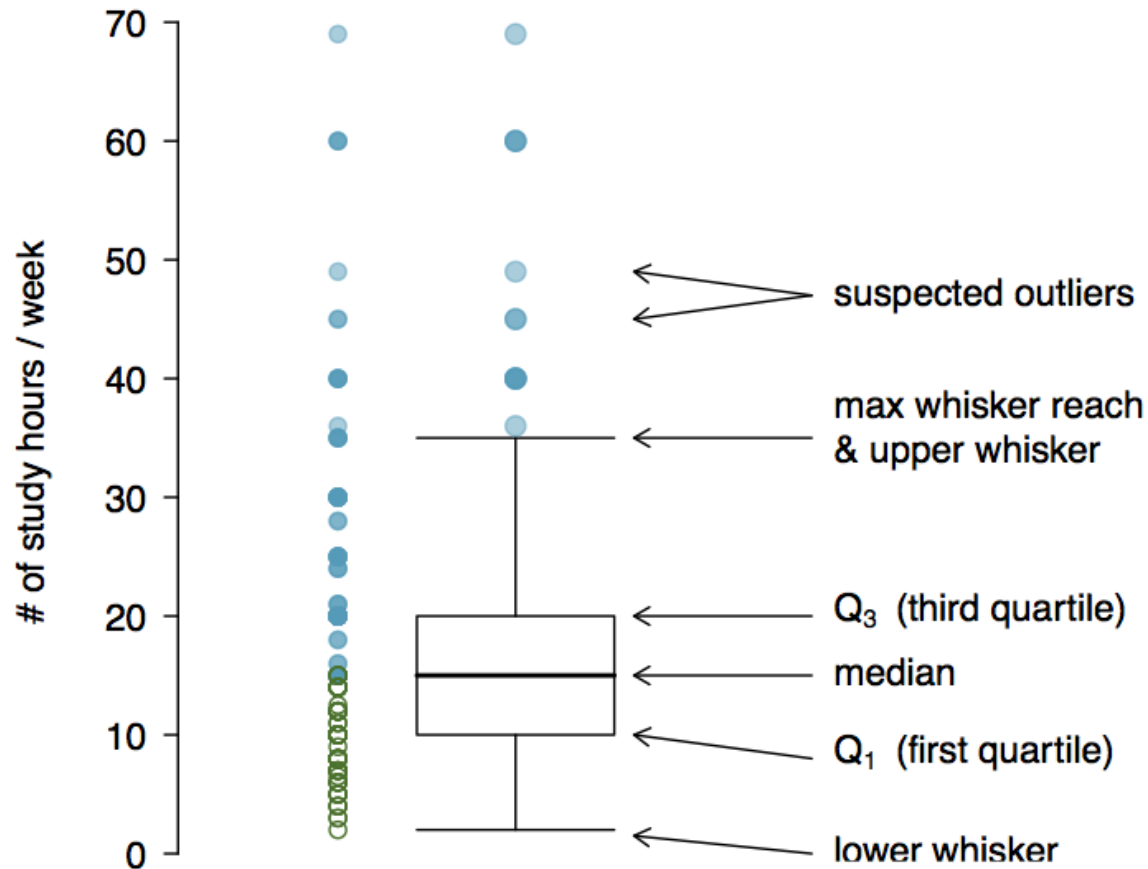This computes the 40th and 60th percentiles.

Plotting as Numerical Summary

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

The **whiskers** of a box plot can extend up to $1.5 \times \mathrm{IQR}$ away from the quartiles.

- max upper whisker reach = $\mathrm{Q3} + 1.5 \times \mathrm{IQR}$

- max lower whisker reach = $\mathrm{Q1} - 1.5 \times \mathrm{IQR}$

Example: IQR: 20 - 10 = 10

- max upper whisker reach = $20 + 1.5 \times 10 = 35$

- max lower whisker reach = $10 - 1.5 \times 10 = -5$

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 (continued)

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

Annual Household Income

Robust Statistics

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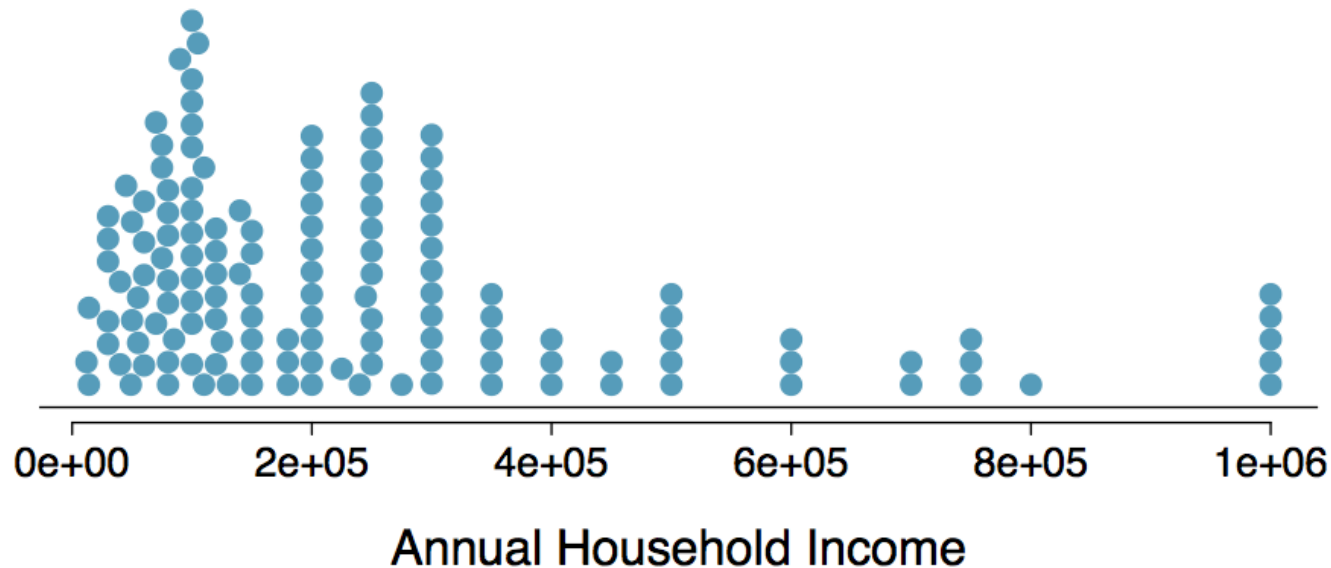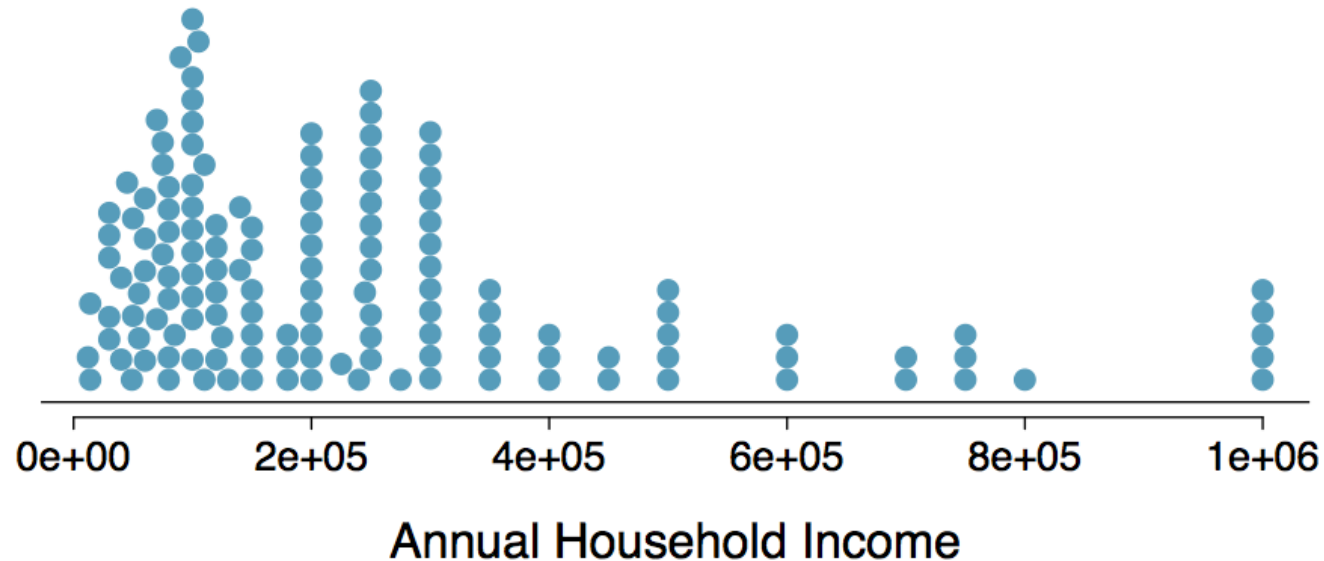
scenario	robust		not robust	
	median	IQR	$\bar{x}$	s
original data	190K	200K	245K	226K
move largest to $10 million	190K	200K	309K	853K
move smallest to $10 million	200K	200K	316K	854K

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

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?
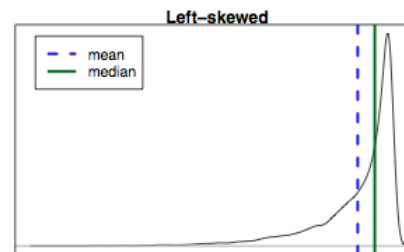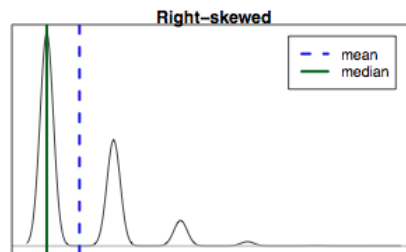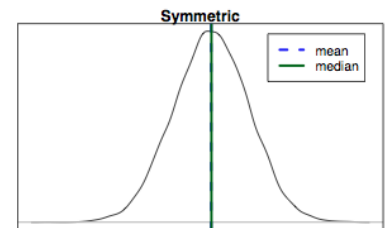
Median

Mean versus Median

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

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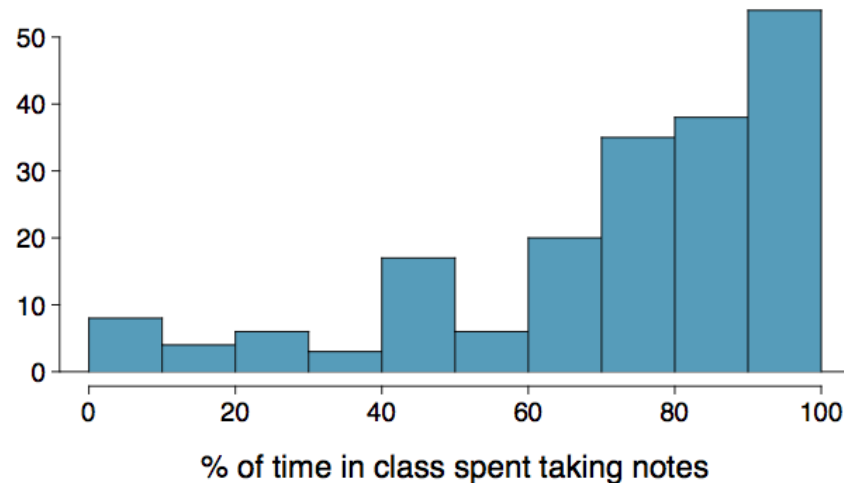
If the distribution is skewed or has extreme outliers, center is often defined as the median

- 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.?

% of time in class spent taking notes

1. mean > median   3. mean ≈ median

2. mean < median   4. 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.?
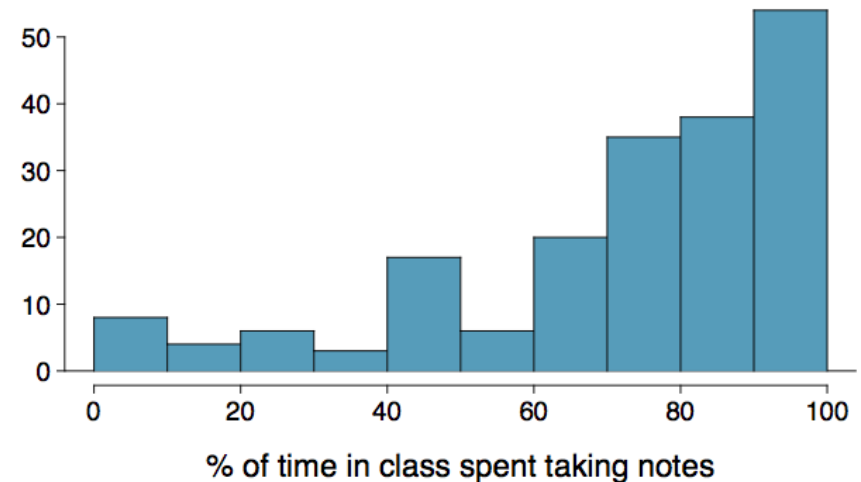
If we compute, the mean = 80% and the median = 76%. So …

1. mean > median   3. mean ≈ median

2. *mean < median*   4. impossible to tell

% of time in class spent taking notes

Considering Categorical Data

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

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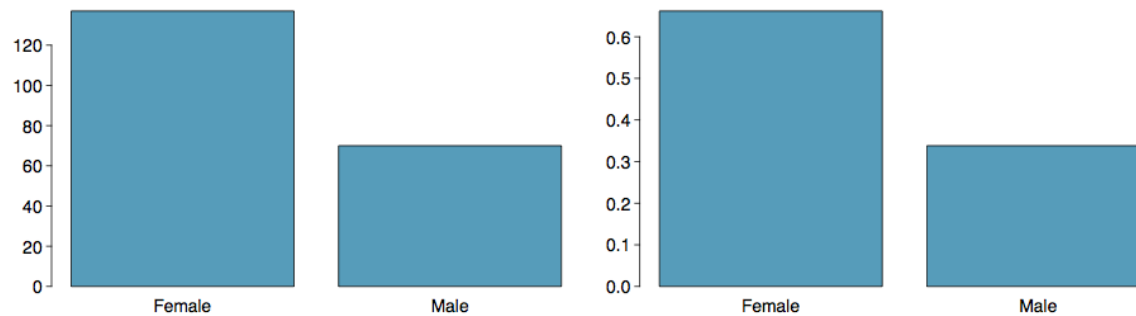
How are bar plots different than histograms? Bar plots are used for displaying distributions of categorical variables, while histograms are used for numerical variables. The x-axis in a histogram is a number line, hence the order of the bars cannot be changed, while in a bar plot the categories can be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

Choosing the Appropriate Proportion

Does there appear to be a relationship between gender and whether the student is looking for a spouse in college?
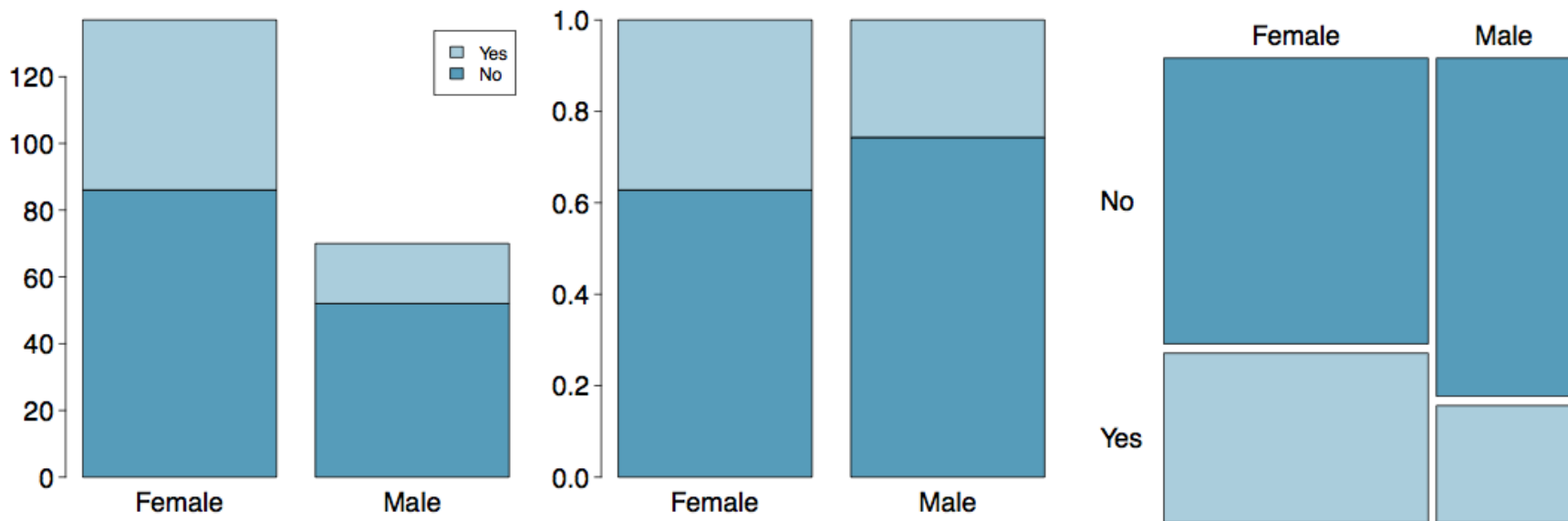
	looking for spouse		
	No	Yes	Total
gender			
Female	86	51	137
Male	52	18	70
Total			
	138	69	207

To answer this question we examine the row proportions:

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

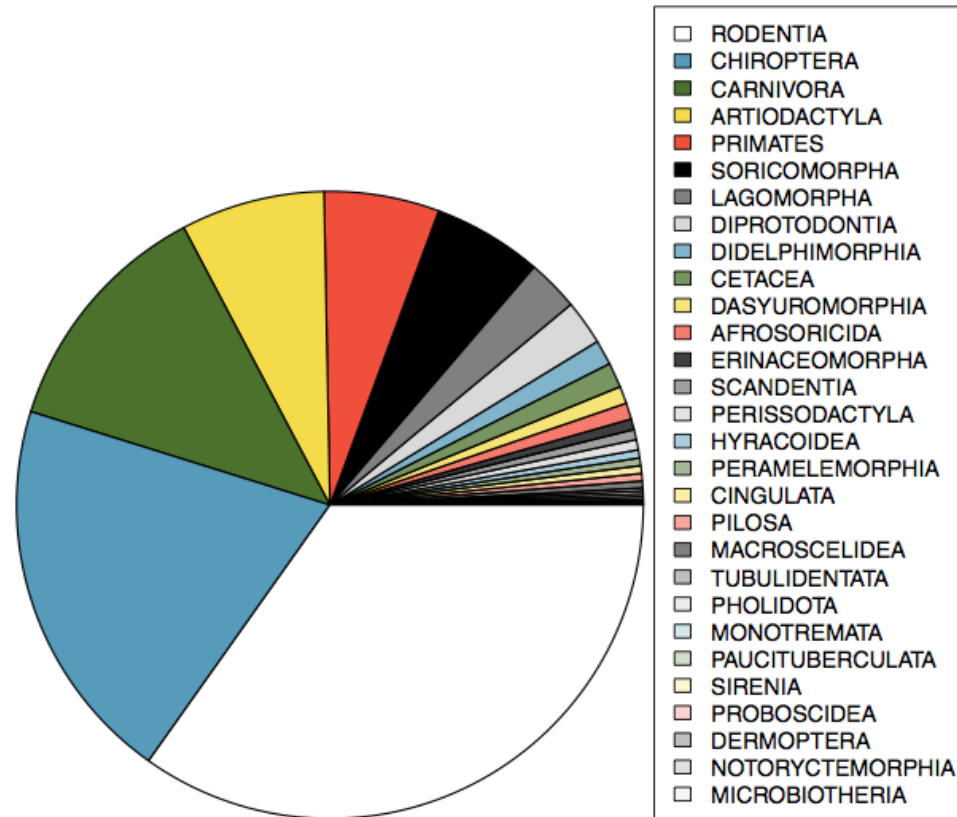- % Males looking for a spouse: 18 / 70 ~ 0.26

Segmented Bar and Mosaic Plots

What are the differences between the three visualizations shown below?

Pie Charts

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

Comparing Numerical Data Across Groups

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