STAT011 Statistical Methods I

Lecture 2 Exploratory Data Analysis I

Lu Chen
Swarthmore College
1/24/2019

Data set structure

```r
load("Stat11SurveyF16_part.RData")  # Import the data set
ls()  # Check list of objects in the current R environment
```

```
## [1] "Survey"
```

```r
head(Survey)  # Look at the first 6 rows of Survey
```

```
##   Gender ClassYear Handedness Height ShoeLength Coffee
## 1      M        So          L     70      12.50      0
## 2      F        Fr          R     65       9.25      5
## 3      M        Sr          R     74      12.00      5
## 4      M        Jr          R     72      11.50     16
## 5      F        So          R     63      10.25      0
## 6      M        So          R     66      11.00      3
```

```r
tail(Survey, 3)  # Look at the last 3 rows of Survey
```

```
##     Gender ClassYear Handedness Height ShoeLength Coffee
## 120      M        So          R     72      11.75      1
## 121      F        So          R     62       9.50      0
## 122      F        So          R     63      10.25      1
```

Data set structure

- **Cases/Observations**: usually by rows
  - The objects described by a set of data. They can be customers, companies, subjects in a study, units in an experiment, or other objects.
- **Variables**: usually by columns
  - Characteristics of cases/observations
    - **Quantitative variable**: numerical values
    - **Categorical variable**: several groups or categories
    - **Label variable**: A special variable used in some data sets to distinguish different cases/observations. For example, names, IDs. Each observation has a **unique** value.

Outline

- Data set structure
- Exploratory data analysis for
  - Categorical variables
    - Table of counts and proportions
    - Bar plot and pie chart
  - Quantitative variables
    - Mean and median
    - Histogram

Exploratory data analysis

> The examination of the main features of data is called **exploratory data analysis**.

- Promoted by John Tukey in his book *Exploratory Data Analysis* in 1977
- Gives us the first impression of data
- A must-have procedure of data analysis but easily ignored

> **Two approaches to exploratory data analysis**
>
> - Summary statistics (Chapter 1.3)
>   - Use numerical values to decribe data
> - Data visualization (Chapter 1.2)
>   - Use graphs to visualize data

Exploratory data analysis

	Summary statistics	Data visualization
Categorical variables		
Quantitative variables		

Statistics versus statistic

- **Statistics**: a dicipline
  - We are studying Statistics.
- **statistic**: a numerical value decribing/summarizing the data
  - The average height of STAT 011 students is a statistic.
- **statistics**: plural form of statistic
  - The statistics that decribe and summarize the data are called summary statistics.

Categorical variables

Gender and *ClassYear* variable in the Survey data

```
Survey$Gender
```

```
##   [1] M F M M F M M M F F F F F M M M M F F F M F F F M M M M F M F F M F F F
##  [37] M F F M M M F M F F F F M M M M F F F F F M M F F M M F F F M M M M M M
##  [73] M F F M O M F F M M F F F M M M F M F F F F F F F F M M M F F F M M M F
## [109] F F F F F F M F M F M M F F
## Levels: F M O
```

```
Survey$ClassYear
```

```
##   [1] So Fr Sr Jr So So So Fr Fr So Fr So Fr Sr Jr Jr So Fr Fr Fr Fr Fr Fr So
##  [25] So Fr So So So Jr So So So So So So So So So Sr Jr Sr Jr So So Jr So So
##  [49] So So So Jr So Sr Jr So So Sr Jr So So So So So So Jr So So Jr So So So
##  [73] So So So So So Sr Jr So So So So So So So Jr Jr So So So So So So So So
##  [97] So So So So So So So So So So So So Jr So Jr So So So So Jr So So So So
## [121] So So
## Levels: Fr So Jr Sr
```

Categorical variables - Summary statistics

Summarizing the *Gender* variable by counts and proportions

```
tab.gen <- table(Survey$Gender)
tab.gen
```

```
## 
##  F  M  O 
## 66 55  1
```

```
prop.table(tab.gen)
```

```
## 
##           F           M           O 
## 0.540983607 0.450819672 0.008196721
```

Categorical variables - Summary statistics

> The **distribution** of a categorical variable lists the categories and gives either the count or the percent of observations that fall in each category.

Gender	F	M	O	Total
Counts	66	55	1	122
Proportions	.54	.45	.01	1.00

Categorical variables - Summary statistics

Summarizing the *ClassYear* variable by counts and proportions

```
tab.cy <- table(Survey$ClassYear)
tab.cy
```

```
##
## Fr So Jr Sr
## 12 85 18  7
```

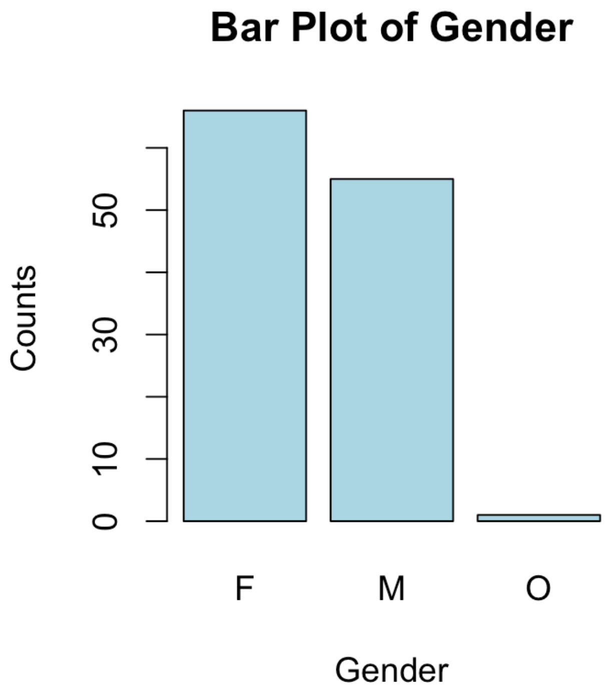
```
prop.table(tab.cy)
```

```
##
##         Fr         So         Jr         Sr
## 0.09836066 0.69672131 0.14754098 0.05737705
```

Categorical variables - Visualization

Bar plot of counts

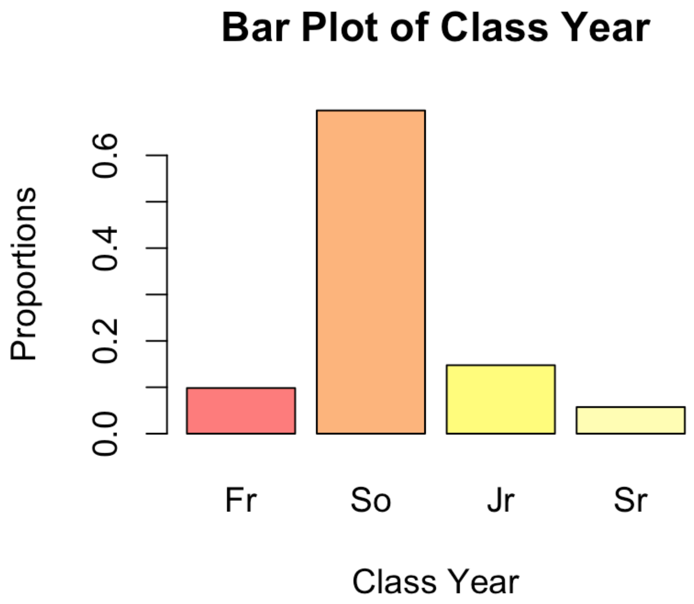
```
barplot(tab.gen, xlab="Gender", ylab="Counts", main="Bar Plot of Gender",
        col="lightblue")
```



Categorical variables - Visualization

Bar plot of proportions

```r
barplot(prop.table(tab.cy), xlab="Class Year", ylab="Proportions",
        main="Bar Plot of Class Year", col=heat.colors(4, alpha=0.6))
```



Categorical variables - Visualization

Pie chart

```r
pie(tab.gen, main="Pie Chart of Gender")
pie(tab.cy, main="Pie Chart of Class Year", border=FALSE,
    labels=c("Fr: 10%","So: 75%","Jr: 15%","Sr: 6%"),
    col=terrain.colors(4, alpha=0.7), clockwise=TRUE)
```

Exploratory data analysis

	Summary statistics	Data visualization
Categorical variables	Table of counts `table()` and proportions `prop.table()`	Bar plot `barplot()` Pie chart `pie()`
Quantitative variables		

Quantitative variables

Height variable in the Survey data

```
Survey$Height
```

```
##   [1] 70.0 65.0 74.0 72.0 63.0 66.0 68.0 65.0 67.0 63.0 61.0 66.0 61.0 61.0
##  [15] 64.0 68.0 70.0 61.0 64.0 60.0 71.0 64.0 64.0 64.0 68.0   NA 71.0 73.0
##  [29] 64.0 72.0 62.0 67.0 72.0 68.0 66.5 65.0 70.5 67.0 59.0 71.0 70.0 69.0
##  [43] 63.0 71.0 66.5 65.5 63.0 65.0 75.5 71.0 71.0 71.0 66.5 62.0 66.5 64.0
##  [57] 65.0 75.0 68.0 60.5 63.0 67.0 68.0 65.0 71.0 61.5 73.0 72.0 71.5 77.0
##  [71] 72.0 72.0 69.0   NA 61.0 74.0 65.0 69.5 65.0 66.0 72.0 72.0 63.5 64.5
##  [85] 63.0 66.0 72.0 69.5 63.0 69.0 64.0 70.0 64.0 61.0 65.0 61.5 67.0 65.0
##  [99] 71.0 72.0 70.0 59.0 60.0 64.0 73.0 72.0 71.5 60.0 61.5 62.0   NA 63.0
## [113] 71.5 63.0 74.0 58.5 76.0 67.0 76.0 72.0 62.0 63.0
```

Quantitative variables - Summary statistics

> The **mean** $\bar{x}$ of a set of observations is the summation of their values divided by the number of observations. If the n observations are $x_1, x_2, \cdots, x_n$, the mean is
>
> $$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Mean height of the first 5 observations

$$\bar{x} = \frac{70 + 65 + 74 + 72 + 63}{5}$$

Quantitative variables - Summary statistics

Mean of *Height* in the Survey data

```
Survey$Height[1:5]  # View the first 5 elements of Height
```

```
## [1] 70 65 74 72 63
```

```
(70 + 65 + 74 + 72 + 63)/5
```

```
## [1] 68.8
```

```
mean(Survey$Height[1:5])  # Mean of the first 5 observations
```

```
## [1] 68.8
```

```
mean(Survey$Height)  # Mean of the Height variable
```

```
## [1] NA
```

Quantitative variables - Summary statistics

Mean of *Height* in the Survey data

```r
mean(Survey$Height)  # The variable has NA values
```

```
## [1] NA
```

```r
# Calculate the mean of Height with NA removed. Note: now the mean is NOT the
# average of 122 observations but 119 obs.
mean(Survey$Height, na.rm = TRUE)
```

```
## [1] 67.0042
```

```r
# Mean Height by Gender
mean(Survey$Height[Survey$Gender == "F"], na.rm = TRUE)
```

```
## [1] 63.86719
```

```r
mean(Survey$Height[Survey$Gender == "M"], na.rm = TRUE)
```

```
## [1] 70.75926
```

Quantitative variables - Summary statistics

Mean of *ShoeLength* and *Coffee*

```r
mean(Survey$ShoeLength)
```

```
## [1] NA
```

```r
mean(Survey$ShoeLength, na.rm = TRUE)
```

```
## [1] 10.6813
```

```r
mean(Survey$Coffee)
```

```
## [1] 2.889344
```

Quantitative variables - Summary statistics

The **median** M is the **midpoint** of a set of observations. Half the observations are smaller than the median, and the other half are larger than the median.

Rule for finding the median:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations n is **odd**, the median M is the center observation in the ordered list.
3. If the number of observations n is **even**, the median M is the mean of the two center observations in the ordered list.

Quantitative variables - Summary statistics

Median of *Coffee*, *Height* and *ShoeLength*

```r
sort(Survey$Coffee)
```

```
##   [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##  [25]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##  [49]  0  0  0  0  0  0  0  0  0  1  1  1  1  1  1  1  1  1  1  1  1  2  2  2
##  [73]  2  2  2  2  2  2  2  3  3  3  3  3  3  3  3  3  4  4  4  4  4  5  5  5
##  [97]  5  5  5  5  5  5  5  6  6  6  7  7  7  8 10 10 10 10 12 14 15 16 16 16
## [121] 17 30
```

```r
median(Survey$Coffee)
```

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

```r
median(Survey$Height, na.rm = TRUE)
```

```
## [1] 66.5
```

```r
median(Survey$ShoeLength, na.rm = T)
```

```
## [1] 10.5
```

Quantitative variables - Summary statistics

Comparing Mean and Median: both measure the center of a set of observations.

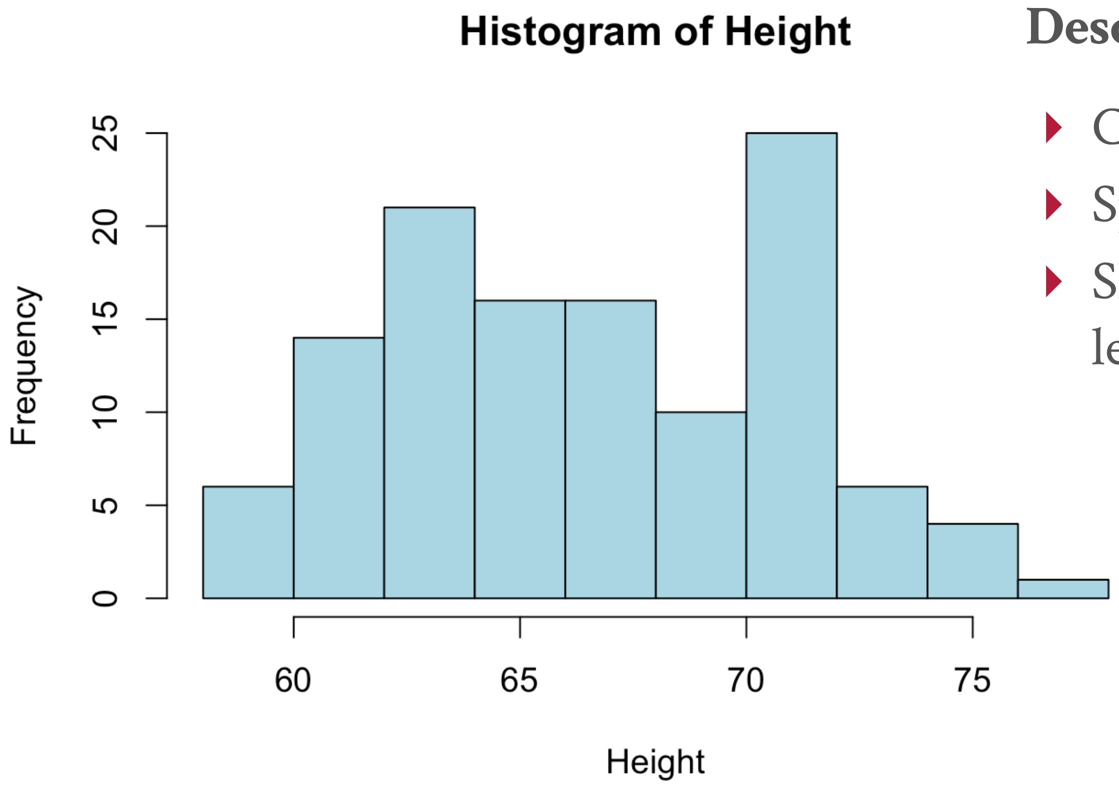
Variable	Height	ShoeLength	Coffee
Mean	67.0	10.7	2.89
Median	66.5	10.5	1

- For *Height* and *ShoeLength*, their mean and median are quite close to each other.
- For *Coffee*, the mean is much larger than the median.

Quantitative variables - Visualization

Histogram displays the **distribution** of a quantitative variable.

```
hist(Survey$Height, xlab="Height", main="Histogram of Height", col="lightblue")
```



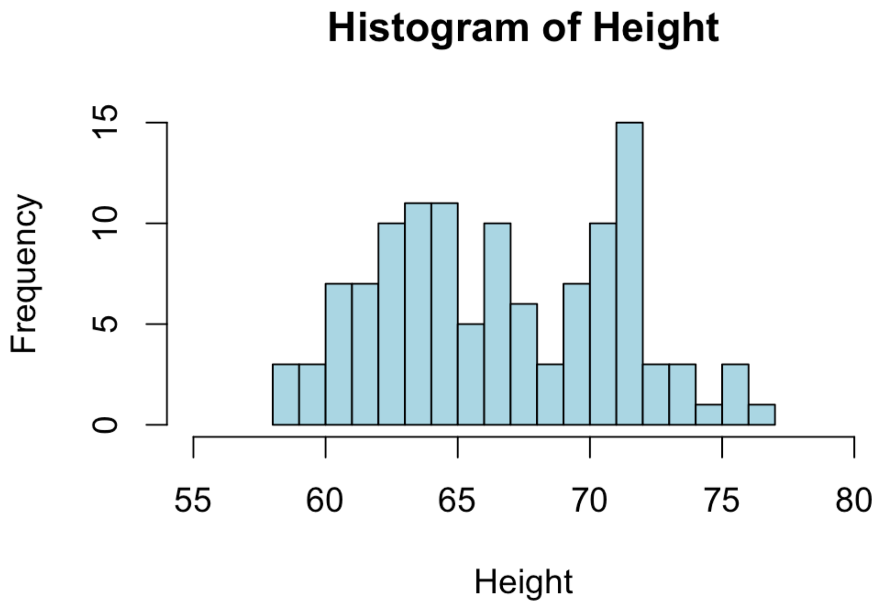
Describe a histogram:

- Center
- Spread
- Shape (unimodal, bimodal, left/right-skewed, etc.)

Quantitative variables - Visualization

A **histogram** breaks the range of values of a variable into classes and displays only the count or percent of the observations that fall into each class.

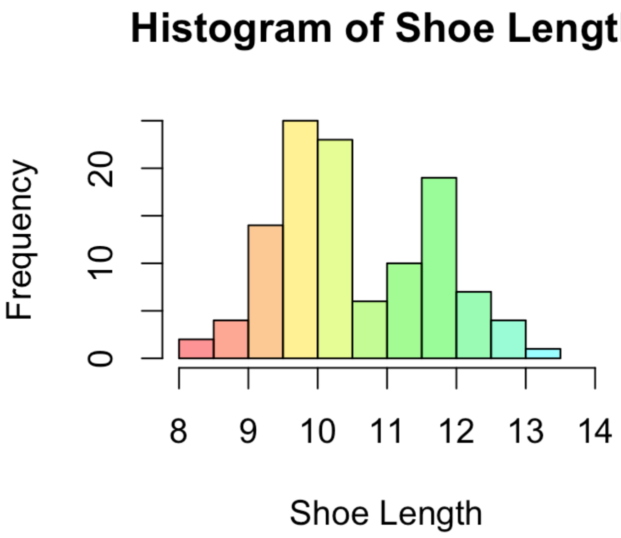
- You can choose any convenient number of classes
- But you should always choose classes of equal width

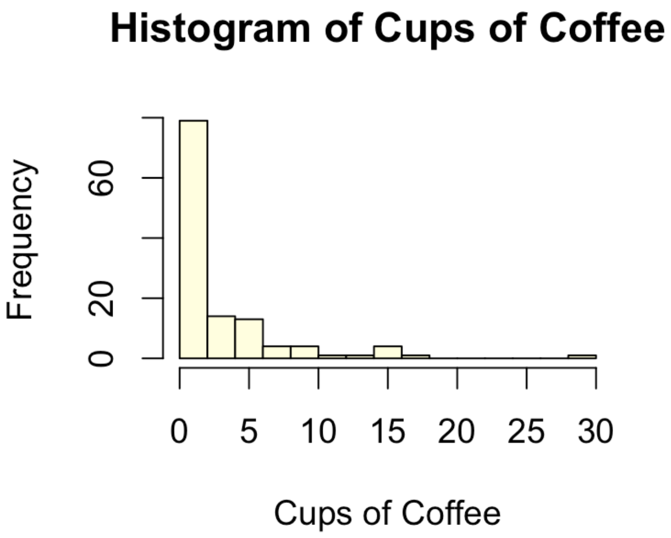


Quantitative variables - Visualization

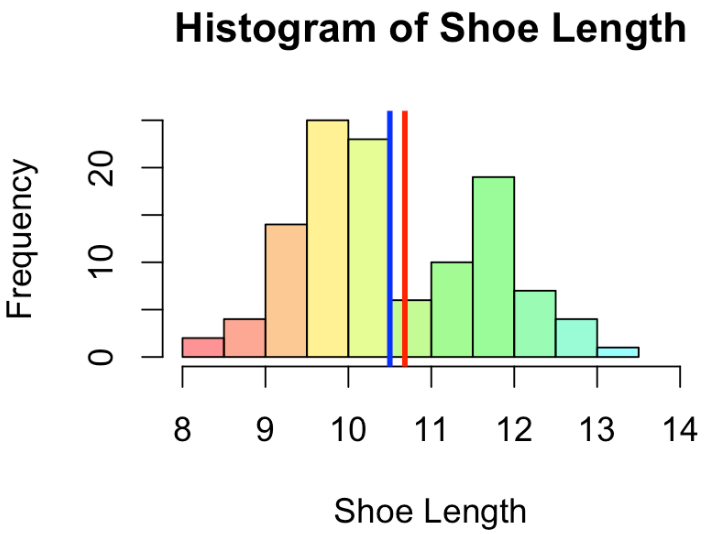
Histogram

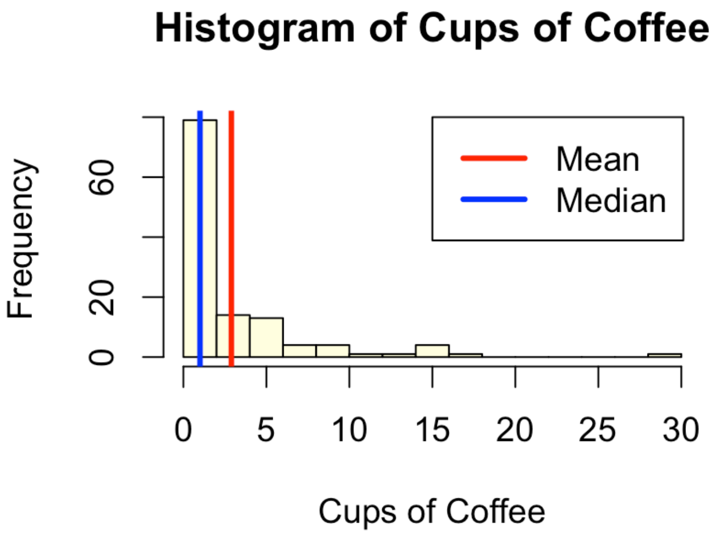
```r
hist(Survey$ShoeLength, xlab="Shoe Length", xlim=c(8,14),
     main="Histogram of Shoe Length", col=rainbow(20,alpha=0.5))
hist(Survey$Coffee, breaks=20, xlab="Cups of Coffee",
     main="Histogram of Cups of Coffee", col="lightyellow")
```





Quantitative variables - Mean and Median





Variable	Height	ShoeLength	Coffee
Mean	67.0	**10.7**	**2.89**
Median	66.5	**10.5**	**1**

Quantitative variables - Mean and Median

Comparing Mean and Median

```r
sort(Survey$Coffee, decreasing = TRUE)
```

```
##   [1] 30 17 16 16 16 15 14 12 10 10 10 10  8  7  7  7  6  6  6  5  5  5  5  5
##  [25]  5  5  5  5  5  4  4  4  4  4  3  3  3  3  3  3  3  3  3  2  2  2  2  2
##  [49]  2  2  2  2  2  1  1  1  1  1  1  1  1  1  1  1  1  0  0  0  0  0  0  0
##  [73]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##  [97]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [121]  0  0
```

Quantitative variables - Mean and Median

Comparing Mean and Median

- Mean is sensitive to the influence of a few extreme observations.
- Because mean cannot resist the influence of extreme observations, we say that it is NOT a **resistant measure** of center.
- However, extreme values usualy have little influence on median because it is determined by position but not values.
- Median is more **resistant** than the mean. We call it a **resistant/robust** measure.
- This is why the U.S. Census Bureau always reports household income using medians.

Quantitative variables - Mean and Median

Comparing Mean and Median

Statistic	Mean	Median
Pros	Taking all the values into account	Resistant to extreme values
Cons	Sensitive (not resistant) to extreme values	Losing information on data values

- There is no definite answer which one is better.
  - Mean is better for roughly symmetric distributions;
  - Median is better for skewed distributions with extreme values.
- In exploratory data analsyis, people usually look at both of them.

Review and preview

	Summary statistics	**Data visualization**
Categorical variables	Table of counts `table()` and proportions `prop.table()`	Bar plot `barplot()` Pie chart `pie()`
Quantitative variables	Mean `mean()` (center) Median `median()` (center) **Standard deviation (spread)** **Interquartile range (spread)**	Histogram `hist()` **Boxplot**