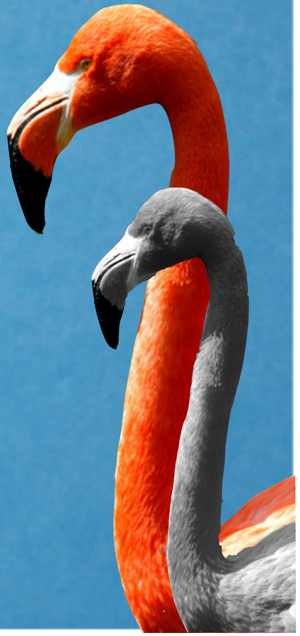
# CHAPTER 1
# Exploring Data

## 1.1
## Analyzing Categorical Data

The Practice of Statistics, 5th Edition
Starnes, Tabor, Yates, Moore



Bedford Freeman Worth Publishers

# Analyzing Categorical Data

## Learning Objectives

After this section, you should be able to:

- ✓ DISPLAY categorical data with a bar graph
- ✓ IDENTIFY what makes some graphs of categorical data deceptive
- ✓ CALCULATE and DISPLAY the marginal distribution of a categorical variable from a two-way table
- ✓ CALCULATE and DISPLAY the conditional distribution of a categorical variable for a particular value of the other categorical variable in a two-way table
- ✓ DESCRIBE the association between two categorical variables

# Categorical Variables

**Categorical variables** place individuals into one of several groups or categories.

| Frequency Table | |
|---|---|
| **Format** | **Count of Stations** |
| Adult Contemporary | 1556 |
| Adult Standards | 1196 |
| Contemporary Hit | 569 |
| Country | 2066 |
| News/Talk | 2179 |
| Oldies | 1060 |
| Religious | 2014 |
| Rock | 869 |
| Spanish Language | 750 |
| Other Formats | 1579 |
| **Total** | **13838** |

| Relative Frequency Table | |
|---|---|
| **Format** | **Percent of Stations** |
| Adult Contemporary | 11.2 |
| Adult Standards | 8.6 |
| Contemporary Hit | 4.1 |
| Country | 14.9 |
| News/Talk | 15.7 |
| Oldies | 7.7 |
| Religious | 14.6 |
| Rock | 6.3 |
| Spanish Language | 5.4 |
| Other Formats | 11.4 |
| **Total** | **99.9** |

**Variable**

**Values**

**Count**

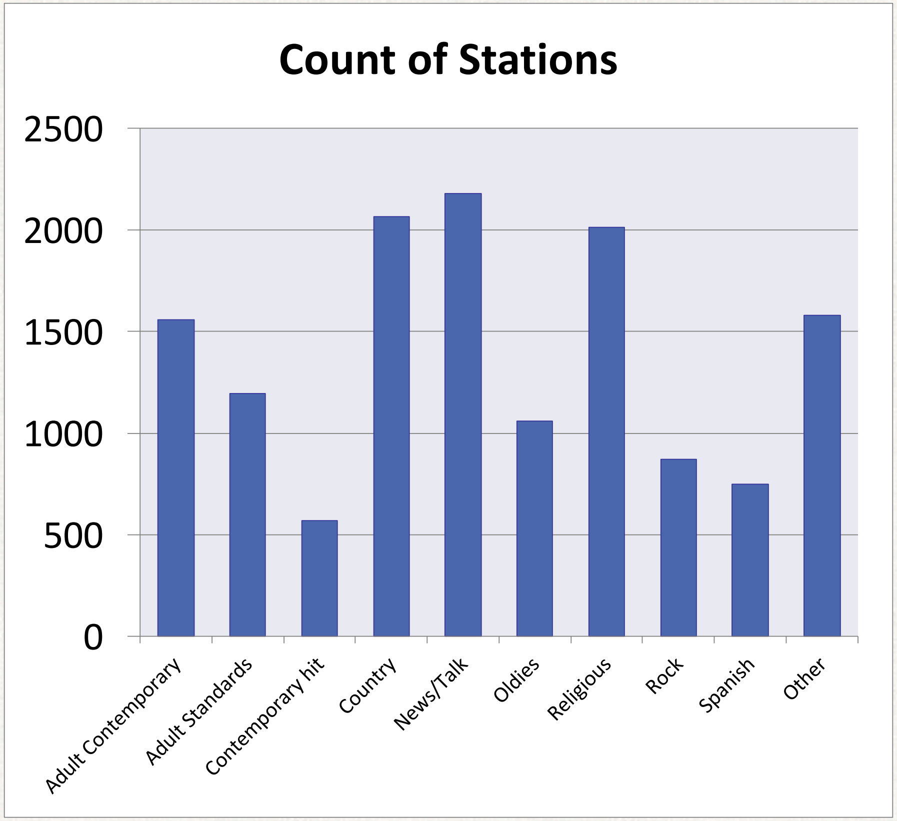
**Percent**

# Displaying Categorical Data

Frequency tables can be difficult to read.

Sometimes is is easier to analyze a distribution by displaying it with a **bar graph** or **pie chart**.

| Frequency Table | |
|---|---|
| **Format** | **Count of Stations** |
| Adult Contemporary | 1556 |
| Adult Standards | 1196 |
| Contemporary Hit | 569 |
| Country | 2066 |
| News/Talk | 2179 |
| Oldies | 1060 |
| Religious | 2014 |
| Rock | 869 |
| Spanish Language | 750 |
| Other Formats | 1579 |
| **Total** | **13838** |

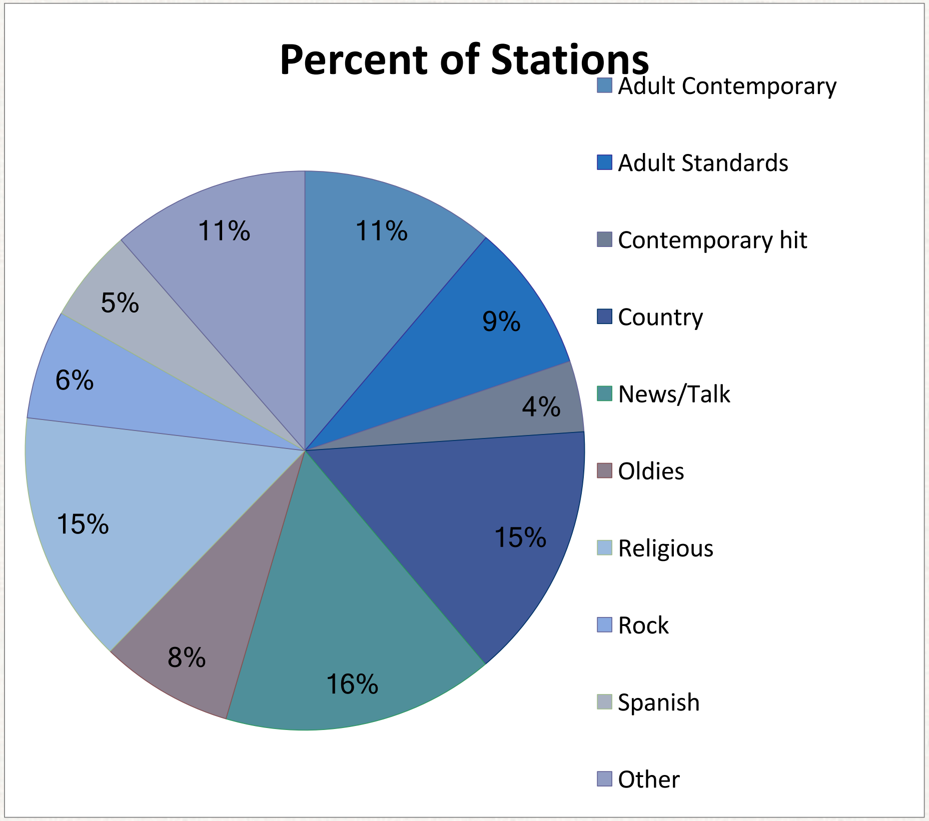


# Displaying Categorical Data

Frequency tables can be difficult to read.

Sometimes is is easier to analyze a distribution by displaying it with a **bar graph** or **pie chart**.

| Relative Frequency Table | |
|---|---|
| **Format** | **Percent of Stations** |
| Adult Contemporary | 11.2 |
| Adult Standards | 8.6 |
| Contemporary Hit | 4.1 |
| Country | 14.9 |
| News/Talk | 15.7 |
| Oldies | 7.7 |
| Religious | 14.6 |
| Rock | 6.3 |
| Spanish Language | 5.4 |
| Other Formats | 11.4 |
| **Total** | **99.9** |



# Graphs: Good and Bad

Bar graphs compare several quantities by comparing the heights of bars that represent those quantities. Our eyes, however, react to the area of the bars as well as to their height.

✓**When you draw a bar graph, make the bars equally wide.**

It is tempting to replace the bars with pictures for greater eye appeal.

✓**Don’t do it!**

There are two important lessons  to keep in mind:

(1)beware the pictograph, and
(2)watch those scales.

# Two-Way Tables and Marginal Distributions

When a dataset involves two categorical variables, we begin by examining the counts or percents in various categories for one of the variables.

A **two-way table** describes two categorical variables, organizing counts according to a *row variable* and a *column variable*.

**Young adults by gender and chance of getting rich**

| | Female | Male | **Total** |
|---|---|---|---|
| Almost no chance | 96 | 98 | **194** |
| Some chance, but probably not | 426 | 286 | **712** |
| A 50-50 chance | 696 | 720 | **1416** |
| A good chance | 663 | 758 | **1421** |
| Almost certain | 486 | 597 | **1083** |
| **Total** | **2367** | **2459** | **4826** |

What are the variables described by this two-way table?

How many young adults were surveyed?

# Two-Way Tables and Marginal Distributions

The **marginal distribution** of one of the categorical variables in a two-way table of counts is the distribution of values of that variable among all individuals described by the table.

**Note**: Percents are often more informative than counts, especially when comparing groups of different sizes.

**How to examine a marginal distribution:**

1)Use the data in the table to calculate the marginal distribution (in percents) of the row or column totals.

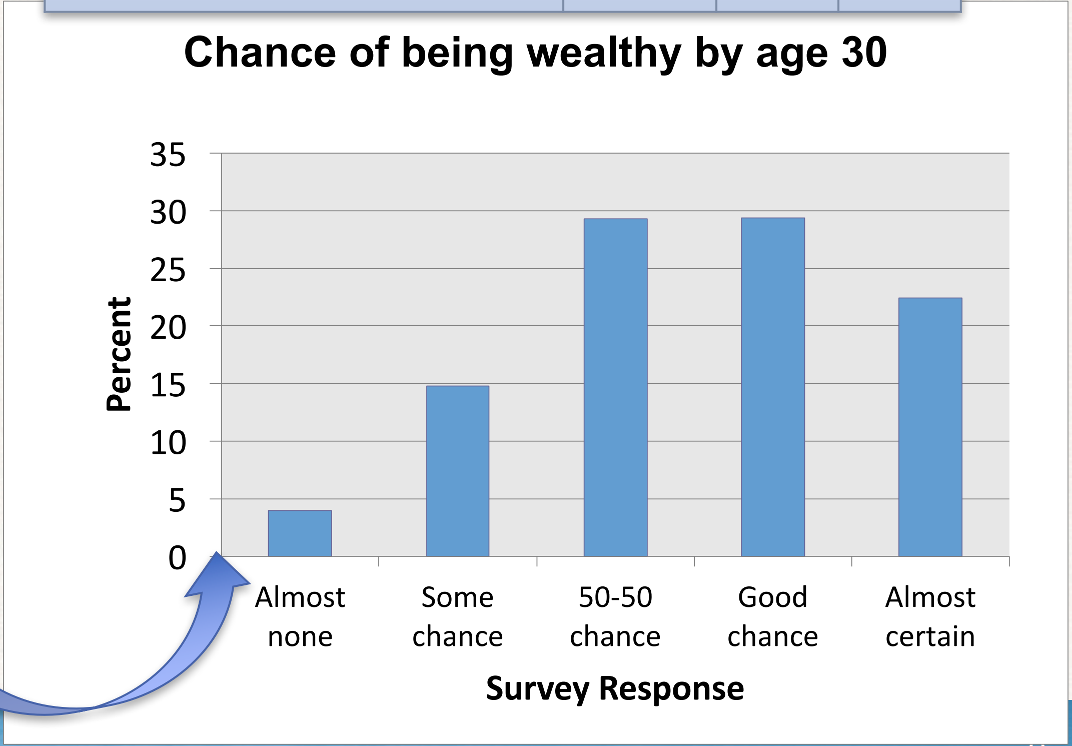
2)Make a graph to display the marginal distribution.

# Two-Way Tables and Marginal Distributions

Examine the **marginal distribution** of chance of getting rich.

| Young adults by gender and chance of getting rich | | | |
|---|---|---|---|
| | Female | Male | **Total** |
| Almost no chance | 96 | 98 | **194** |
| Some chance, but probably not | 426 | 286 | **712** |
| A 50-50 chance | 696 | 720 | **1416** |
| A good chance | 663 | 758 | **1421** |
| Almost certain | 486 | 597 | **1083** |
| **Total** | **2367** | **2459** | **4826** |

| Response | Percent |
|---|---|
| Almost no chance | 194/4826 = 4.0% |
| Some chance | 712/4826 = 14.8% |
| A 50-50 chance | 1416/4826 = 29.3% |
| A good chance | 1421/4826 = 29.4% |
| Almost certain | 1083/4826 = 22.4% |



# Relationships Between Categorical Variables

A **conditional distribution** of a variable describes the values of that variable among individuals who have a specific value of another variable.

**How to examine or compare conditional distributions:**

1) Select the row(s) or column(s) of interest.

2) Use the data in the table to calculate the conditional distribution (in percents)  of the row(s) or column(s).

3) Make a graph to display the conditional distribution.

- Use a **side-by-side bar graph** or **segmented bar graph** to compare distributions.

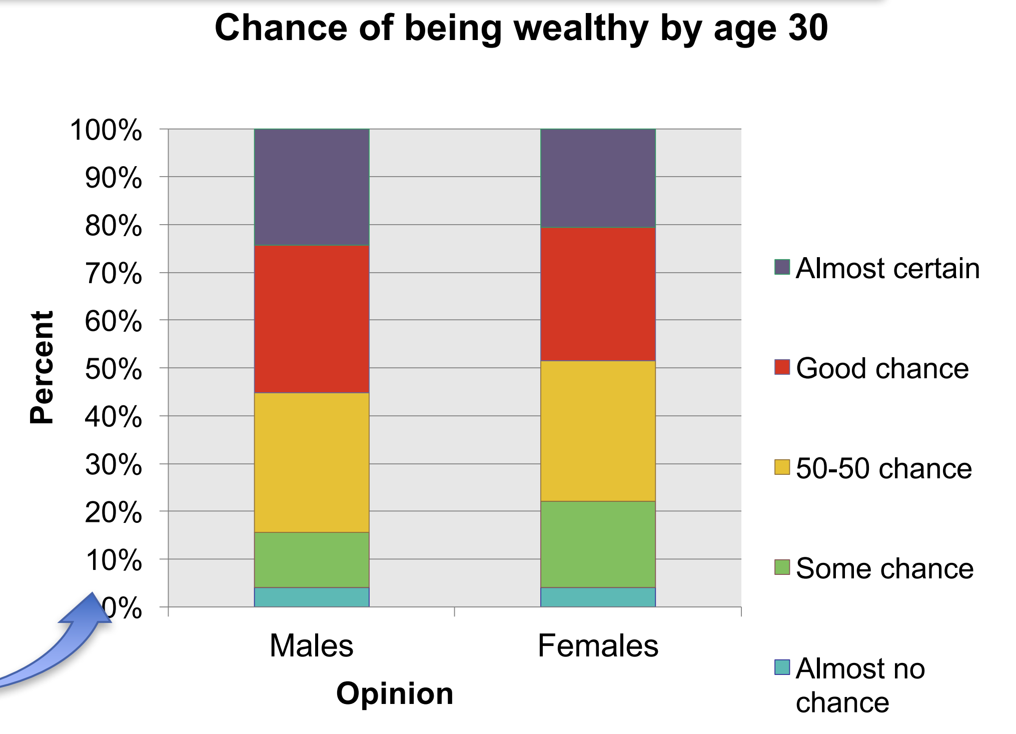
# Relationships Between Categorical Variables

Calculate the **conditional distribution** of opinion among males. Examine the relationship between gender and opinion.

**Young adults by gender and chance of getting rich**

| | Female | Male | **Total** |
|---|---|---|---|
| Almost no chance | 96 | 98 | **194** |
| Some chance, but probably not | 426 | 286 | **712** |
| A 50-50 chance | 696 | 720 | **1416** |
| A good chance | 663 | 758 | **1421** |
| Almost certain | 486 | 597 | **1083** |
| **Total** | **2367** | **2459** | **4826** |

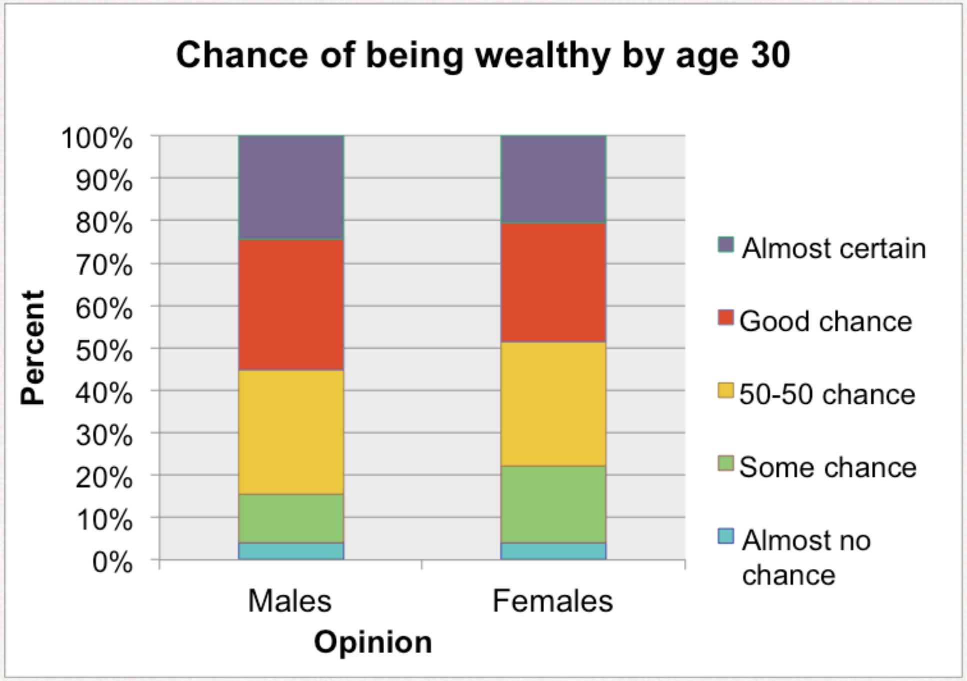
| **Response** | **Male** | **Female** |
|---|---|---|
| Almost no chance | 98/2459 = 4.0% | 96/2367 = 4.1% |
| Some chance | 286/2459 = 11.6% | 426/2367 = 18.0% |
| A 50-50 chance | 720/2459 = 29.3% | 696/2367 = 29.4% |
| A good chance | 758/2459 = 30.8% | 663/2367 = 28.0% |
| Almost certain | 597/2459 = 24.3% | 486/2367 = 20.5% |



# Relationships Between Categorical Variables

Can we say there is an association between gender and opinion in the *population* of young adults?

Making this determination requires formal inference, which will have to wait a few chapters.



**Caution!**
Even a strong association between two categorical variables can be influenced by other variables lurking in the background.

# Data Analysis: Making Sense of Data

## Section Summary

In this section, we learned how to…

- ✓ DISPLAY categorical data with a bar graph
- ✓ IDENTIFY what makes some graphs of categorical data deceptive
- ✓ CALCULATE and DISPLAY the marginal distribution of a categorical variable from a two-way table
- ✓ CALCULATE and DISPLAY the conditional distribution of a categorical variable for a particular value of the other categorical variable in a two-way table
- ✓ DESCRIBE the association between two categorical variables