Chapter 1.1 Lecture Notes & Examples

Section 1.1 – Analyzing Categorical Data (pp. 7-24)

Review

- Definition of Individual and Variable
- Types of Variables
- Statistics: Collect data, analyze it, make inferences

Distributions/Frequency Tables/Relative Frequency Tables

The **distribution** of the values of a **categorical** variable lists the count or percent of the individuals that fall into each category.

Example, page 8

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
nguage	5.4
Other Form	11.4
Total	**99.9**

Variable

Values

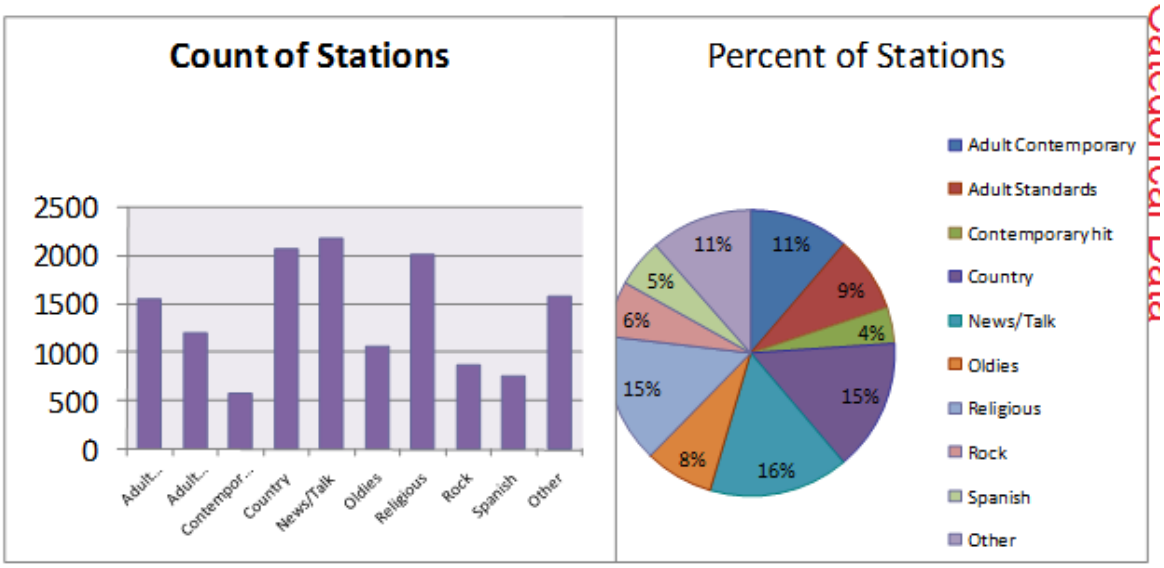
Count

Percent

- Discuss individual data points.
- Discuss how to build relative frequency table from frequency table.
- Discuss rounding errors

Bar Graphs and Pie Charts

A picture is worth a thousand words…….
(Page 9)



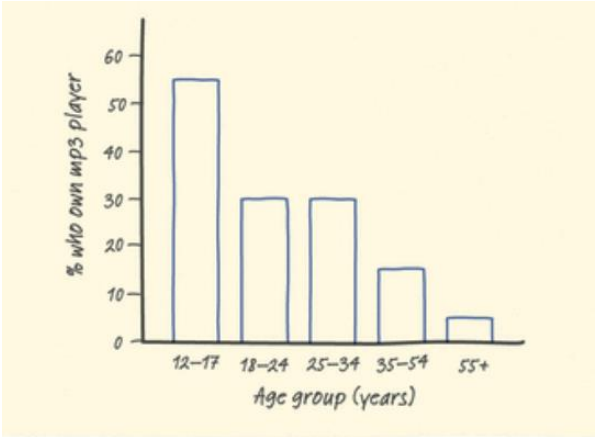
- Discuss tables
- Pie charts must contain all of the categories that make up the whole
- Bar charts are easier to make and are also more flexible than pie charts – a bar chart can display any set of quantities that are measured in the same units (do not have to add to 100%)

EXAMPLE : Who Owns an MP3 Player? (Page 10)

Choosing the best graph to display the data

Portable MP3 music players, such as the Apple iPod, are popular—but not equally popular with people of all ages. Here are the percents of people in various age groups who own a portable MP3 player, according to an Arbitron survey of 1112 randomly selected people.

Age group (years)	Percent owning an MP3 player
12 to 17	54
18 to 24	30
25 to 34	30
35 to 54	13
55 and older	5



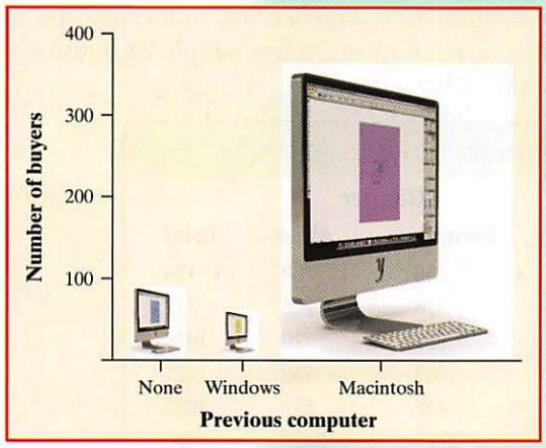
A) Describe what you see in bar graph.

B) Would it be appropriate to make a pie chart for this data? Why or why not?

Graphs: Good and Bad

- Bars should be the same width
- Bars should not be pictographs
- Y-axis should start at 0 and not be compressed.

Examples on page 11.

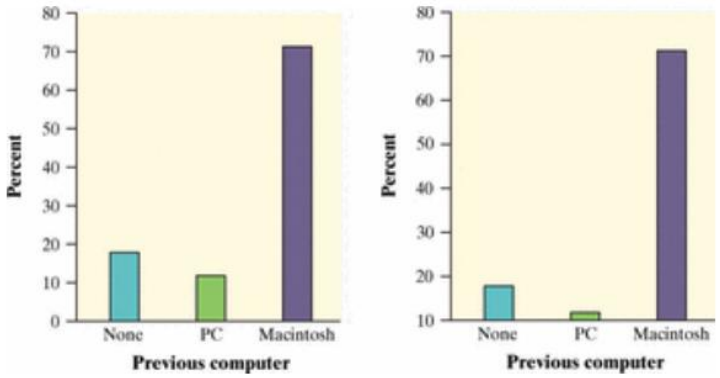


Previous ownership	Count	Percent
None	85	17.0
PC	60	12.0
Macintosh	355	71.0
Total	500	100.0

PROBLEM:

(a) Here's a clever graph of the data that uses pictures instead of the more traditional bars. How is this graph misleading?

(b) Two possible bar graphs of the data are shown on the next page. Which one could be considered deceptive? Why?

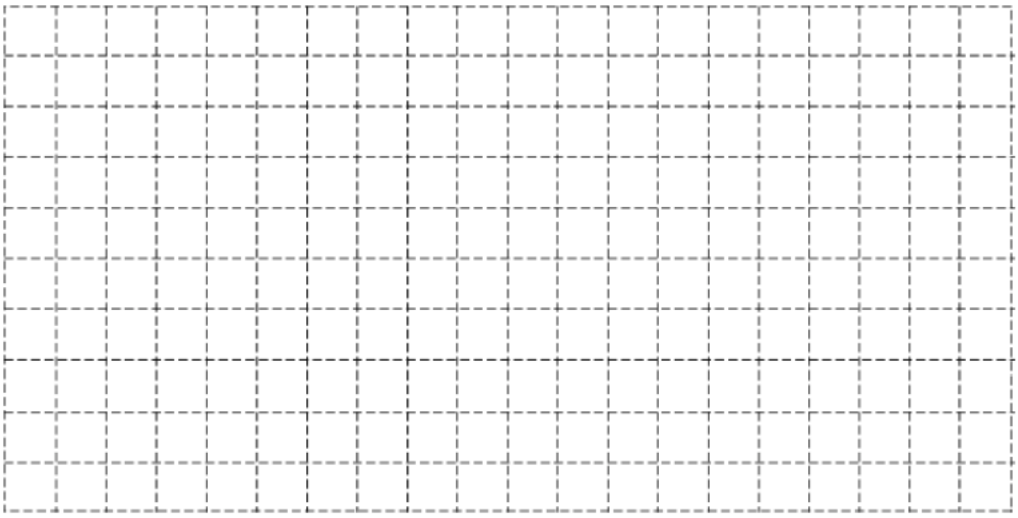


Teams – Do problem 16 on pp 23-24

16. The audience for movies – Here are data on the percent of people in several age groups who attended a movie in the past 12 months:

Age Group	Movie Attendance
18-24	83%
25-34	73%
35-44	68%
45-54	60%
55-64	47%
65-74	32%
75 and up	20%

(a) Display these data in a bar graph. Describe what you see.



(b) Would it be correct to make a pie chart of these data? Why or why not?

(c) A movie studio wants to know what percent of the total audience for movies is 18-24 year olds. Explain why these data do not answer this question.

Two-Way Tables

A two-way table is a table that describes two categorical variables. They have a *row variable* and a *column variable*.

Example on page 12

Example, p. 12

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?

Marginal Distributions

In order to grasp how the variables compare we will compute a *marginal distribution*. 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*. It will be in the form of percents.

Percents are better than counts to make comparisons especially when comparing groups of different sizes.

Example: Steps: (1) Use the data in the table to calculate the marginal distribution; (2) make a graph of the marginal distribution.

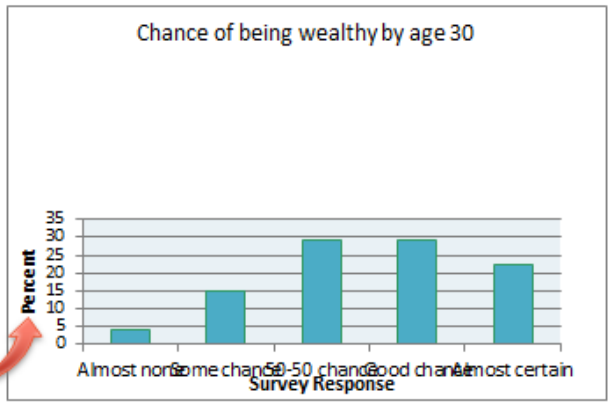
Example, p. 13

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

Examine the marginal distribution of chance of getting rich.

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%**



Teams – Check your understanding

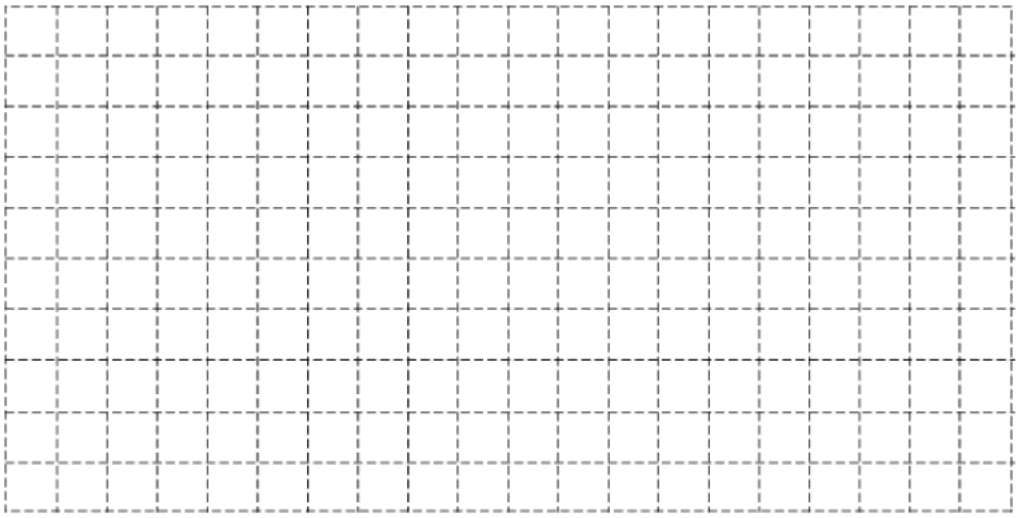
Sample: Super Powers

A sample of 200 children from the United Kingdom ages 9-17 was selected from the CensusAtSchool website (www.censusatschool.com). The gender of each student was recorded along with which super power they would most like to have: invisibility, super strength, telepathy (ability to read minds), ability to fly, or ability to freeze time.

Problem: (a) Use the data in the two-way table to calculate the marginal distribution (as a proportion) of superpower preferences.

Responses	Female	Male	Total
Invisibility	17	13	30
Super Strength	3	17	20
Telepathy	39	5	44
Fly	36	18	54
Freeze Time	20	32	52
Total	115	85	200

(b) Make a graph to display the marginal distribution. Describe what you see.



Young adults by gender and chance of getting rich			
	Gender		
Opinion	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

Relationships between categorical variables: Conditional Distributions

Marginal distributions do not tell us anything about the relationship between two variables. To do this we must calculate some well-chosen percents.

Look at females alone in the table. Now we are only looking at 2367 individuals.

Conditional distribution of opinion among women **Response** — **Female** Almost no chance — $\frac{96}{2367} = 4.1\%$ Some chance — $\frac{426}{2367} = 18.0\%$ A 50-50 chance — $\frac{696}{2367} = 29.4\%$ A good chance — $\frac{663}{2367} = 28.0\%$ Almost certain — $\frac{486}{2367} = 20.5\%$ p. 14	This gives us the **conditional distribution for females**. A **conditional distribution** of a variable describes the values of that variable among individuals who have a specific value of another variable. There is a separate conditional distribution for each value of the other variable.
Example. Conditional dist for men: Conditional distribution of opinion among men **Response** — **Male** Almost no chance — $\frac{98}{2459} = 4.0\%$ Some chance — $\frac{286}{2459} = 11.6\%$ A 50-50 chance — $\frac{720}{2459} = 29.3\%$ A good chance — $\frac{758}{2459} = 30.8\%$ Almost certain — $\frac{597}{2459} = 24.3\%$ p. 15	

Organizing a Statistical Problem – 4 Step Process

1. **State**: What is the question that you are trying to answer?

2. **Plan**: How will you go about answering the question? What statistical techniques does the problem call for? Have you met the conditions and assumptions necessary to use those techniques?

3. **Do**: Make graphs and carry out the calculations

4. **Conclude**: Give your **practical** conclusion in the **context** of the real-world problem.

Example p. 18 – Can we conclude that young men and young women differ in their opinions about the likelihood of future wealth? Give appropriate evidence to support your answer. Follow the four-step process.

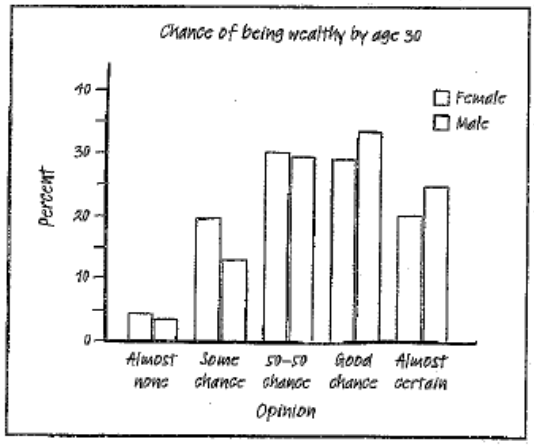


FIGURE 1.7 Side-by-side bar graph comparing the opinions of males and females.

STATE: What is the relationship between gender and responses to the question "What do you think are the chances you will have much more than a middle-class income at age 30?"

PLAN: We suspect that gender might influence a young adult's opinion about the chance of getting rich. So we'll compare the conditional distributions of response for men alone and for women alone.

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

Side-by-side bar graph

DO: We'll make a side-by-side bar graph to compare the opinions of males and females. Figure 1.7 displays the completed graph.

CONCLUDE: Based on the sample data, men seem somewhat more optimistic about their future income than women. Men were less likely to say that they have "some chance but probably not" than women (11.6% vs. 18.0%). Men were more likely to say that they have "a good chance" (30.8% vs. 28.0%) or are "almost certain" (24.3% vs. 20.5%) to have much more than a middle-class income by age 30 than women were.

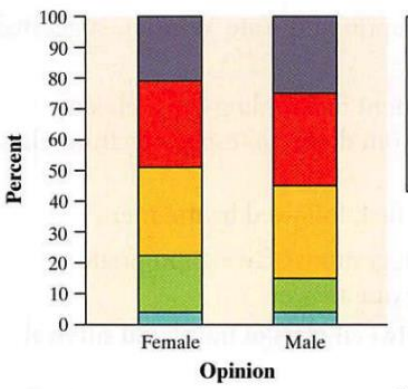


FIGURE 1.8 Segmented bar graph comparing the opinions of males and females.

Association – We say there is **association** between two variables if specific values of one variable tend to occur in common with specific values of the other.

Caution: Just because an association exists does not mean one variable *causes* another variable to act in a certain way. Also, there may be other variables *lurking in the background*.

SUMMARY

- The distribution of a categorical variable lists the categories and gives the count (frequency table) or percent (relative frequency table) of individuals that fall in each category.
- Pie charts and bar graphs display the distribution of a categorical variable. Bar graphs can also compare any set of quantities measured in the same units. When examining any graph, ask yourself, "What do I see?"
- A two-way table of counts organizes data about two categorical variables. Two-way tables are often used to summarize large amounts of information by grouping outcomes into categories.
- The row totals and column totals in a two-way table give the marginal distributions of the two individual variables. It is clearer to present these distributions as percents of the table total. Marginal distributions tell us nothing about the relationship between the variables.
- There are two sets of conditional distributions for a two-way table: the distributions of the row variable for each value of the column variable, and the distributions of the column variable for each value of the row variable. You may want to use a side-by-side bar graph (or possibly a segmented bar graph) to display conditional distributions.
- A statistical problem has a real-world setting. You can organize many problems using the four steps state, plan, do, and conclude.
- To describe the association between the row and column variables, compare an appropriate set of conditional distributions. Remember that even a strong association between two categorical variables can be influenced by other variables lurking in the background.
- An association between two variables that holds for each individual value of a third variable can be changed or even reversed when the data for all values of the third variable are combined. This is Simpson's paradox.