Chapter 2
Displaying and Describing Categorical Data

What's It About?

We introduce students to distributions of categorical variables. The mathematics is easy (summaries are just percentages) and the graphs are straightforward (pie charts and bar graphs). We challenge students to uncover the story the data tell, and to write about it in complete sentences in context.

Then we up the ante, asking them to compare distributions in two-way tables. Constructing comparative graphs, discussing conditional distributions, and considering (informally) the idea of independence give students a look at issues that require deeper thought, careful analysis, and clear writing.

Comments

By Chapter 2, it will begin to dawn on your students that this isn't a math class. At the very least, they are going to be expected to write often and clearly. For those who have not yet developed the skill of writing clearly, this may be one of the most valuable things learned in this course. This chapter provides an early challenge to students to write conclusions that are clear, concise, complete, and in context – The Four C's.

1. Continue to emphasize precision of vocabulary (and notation). These are an important part of clear communication, critical to success.

2. Emphasize *Think-Show-Tell* right from the start. The key to doing well in Statistics (and on the AP* Exam) is to think carefully about what each question is asking and what statistical techniques can address those issues before starting to write an answer. And then, after showing some calculations or other work, to write clear and concise explanations of what it all means. Your students may rebel at first at having to write sentences, much less paragraphs, in a course they may have thought was a math class. They are used to just doing the *Show*. *Tell* is at least 50% of each solution. If you make that point consistently right from the start of the course it becomes second nature soon, and puts each student in the right mindset for writing solid AP* answers. Continually remind them: ***Answers are sentences, not numbers.*** (Indeed, on the AP* Exam, clear communication usually accounts for at least 50% of the credit for a problem.)

3. This is the first substantial chapter, and it gets pretty involved right away. We deal with conditional distributions, independence, and confounding (Simpson's paradox). It may seems early to bring up such sophisticated concepts, but our experience is that students can get lulled into a false sense of security in the early part of this course, if all they see is things like means and histograms that they have dealt with since middle school. They think the course is going to be pretty easy, and they may not recognize the level of sophistication that is required until it's too late. The ideas in Chapter 2 are not hard and are introduced only informally, but they do require some thought. Students will find it difficult to make clear explanations. We want these ideas to be interesting, to engage imaginations, and to challenge students. We hope the level of thought required will get their attention and arouse their interest.

Looking Ahead

There are many important skills and ideas here that prepare students for later topics. They need to think about the type of data, checking a condition before plunging ahead. They need to think about what comparisons will answer the questions posed, and write clear explanations in context. They begin to think about independence, one of the most important issues in Statistics. And, in Simpson's paradox, they see the need to think more deeply to avoid being misled by lurking or confounding variables.

Class Do's

Weave the key step of checking the assumptions and conditions into the fabric of doing Statistics. It's easy: have students check that the data are being treated as categorical before they proceed with pie charts, conditional distributions, and the like. As the course goes on, *Thinking* about assumptions and conditions will help students select appropriate statistical procedures – and it's a requirement for a complete solution on the AP* Exam. Start now.

Discuss categorical data and appropriate summaries: numerical (counts/percentages), graphical (pie charts, bar graphs). Discuss *distribution, frequency, relative frequency*.

It gets more interesting when we make comparisons (using bivariate data): e.g., political leanings by gender? Discuss two-way tables, *marginal* and *conditional* distributions. Political views may be interesting, but looking at the differences in political view by gender adds much more to the discussion. You can emphasize the vocabulary by asking things like "What is the marginal frequency distribution of gender?" vs. "What is the conditional relative frequency distribution of gender among Conservatives?"

Make sure students can correctly sort out (*Think-Show-Tell*) answers to similar sounding questions:

1. What percent of the class are women with liberal political views?
2. What percent of the liberals are women?
3. What percent of the women are liberals?

Raise the issue of independence. It's not formal independence yet, just the general idea that if gender and political view were independent, the percentages for either gender would mirror the class as a whole, or the percentages of Liberal, Moderate, and Conservative would be the same for both genders. If they are not, we encounter what the politicians refer to as the "gender gap". Statisticians would say this indicates that voting preference is not independent of gender.

Pay attention in each chapter to the What Can Go Wrong? (WCGW) sections. Helping students avoid common pitfalls is one of the keys to success in this course.

Simpson's Paradox is fun, but don't overemphasize it. It's not a critical issue, but it's a good discussion point about making valid comparisons, and not overlooking lurking or confounding variables.

The Importance of What You Don't Say

Probability. You can see that we are patrolling the perimeter of probability. Concepts like relative frequency, conditional relative frequency, and independence cry out for a formal discussion in probabilistic terms. Don't heed the cry. You and we know that we are setting up the habits of thought that students will need for learning about probability. But this isn't the time to discuss the formalities. Or even to say the word "probability" out loud. (Notice that the book doesn't use the

term in this chapter at all – it'll still be a while before we get to it.) Talk about "relative frequency" instead. In this class probability is a relative frequency, so we are encouraging students to think about the concepts correctly. By the time we introduce formal probability, they will have a sound intuitive foundation.

Class Examples

1. Use the class data about gender and political view – liberal, moderate, conservative. Help students develop their *Think-Show-Tell* skills with questions like:
   - What percent of the class are girls with liberal political views?
   - What percent of the liberals are girls?
   - What percent of the girls are liberals?
   - What is the marginal frequency distribution of political views?
   - What is the conditional relative frequency distribution of gender among conservatives?
   - Are gender and political view independent?
2. Use the worksheet about smoking and education level. Students should conclude that smokers tend to have higher education levels than non-smokers. 64% of the smokers had only a high school education compared to 47% overall. And non-smokers were almost twice as likely as smokers (48% to 26%) to have completed at least 4 years of college.

   As a "What Can Go Wrong" exercise, it's fun to see how many incorrect "conclusions" students can think up beyond the one suggested on the sheet. Of course, this does not indicate that kids who smoke in high school will quit when they go off to college. Perhaps smokers can't afford college because of all the money they waste on cigarettes. We simply don't have data to support any such conclusions, and, being a cautious bunch, statisticians should avoid such speculation.
3. Is the color distribution of M&Ms independent of the type of candy? Break open bags of plain and peanut M&Ms and count the colors. Is the color distribution of M&Ms independent of type of candy? (Then eat the data…)
4. Simpson's paradox example:
   It's the last inning of important game. Your team is a run down with the bases loaded and two outs. The pitcher is due up, so you'll be sending in a pinch-hitter. There are 2 batters available on the bench. Whom should you send in to bat? First show the students the overall success history of the two players.

Player	Overall	vs LHP	vs RHP
A	33 for 103	28 for 81	5 for 22
B	45 for 151	12 for 32	33 for 119

   A's batting average is higher than B's (.320 vs. .298), so he looks like the better choice. Someone, though, will raise the issue that it matters whether the pitcher throws right- or left-handed. Now add the rest of the table. It turns out that B has a higher batting average against both right- and left-handed pitching, even though his overall average is lower. Students are stunned.

Here's an explanation. B hits better against both right- and left-handed pitchers. So no matter the pitcher, B is a better choice. So why is his batting "average" lower? Because B sees a lot more right-handed pitchers than A, and (at least for these guys) right-handed pitchers are harder to hit. For some reason, A is used mostly against left-handed pitchers, so A has a higher average.

Suppose all you know is that A bats .227 against righties and .346 against lefties. Ask the students to guess his overall batting average. It could be anywhere between .227 and .346, depending on how many righties and lefties he sees. And B's batting average may slide between .277 and .375. These intervals overlap, so it's quite possible that A's batting average is either higher or lower than B's, depending on the mix of pitchers they see.

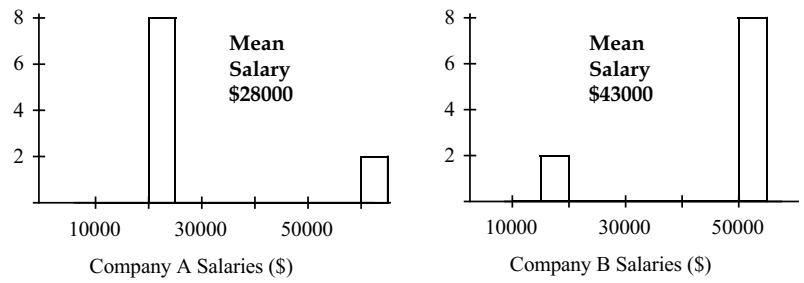
Pooling (nice word to use . . .) the data together loses important information and leads to the wrong conclusion. We always should take into account any factor that might matter.

5. Refer to Simpson, again. Here's a nice thought problem to pose to the class; give them a few minutes to work it out. Two companies have labor and management classifications of employees. Company A's laborers have a higher average salary than company B's, as do Company A's managers. But overall company B pays a higher average salary. How can that be? And which is the better way to compare earning potential at the two companies?

Solution:

First of all, make sure you point out that this example deals with quantitative variables, not categorical. The paradox can be explained when you realize that Company A must employ a greater percentage of laborers than Company B. Also, Company A must employ a smaller percentage of managers than Company B. If laborers earn salaries that are considerably lower than managers, the salaries of Company A's laborers will pull the company average down, and the salaries of Company B's managers will pull the company average up.

The proper way to compare the companies is to use the salaries that are broken down by job type. Using the overall average salary leads to a misleading conclusion.



Resources

TI-Nspire™ Demonstration

- Conditional distributions and association
  - Explore the Titanic data to see which passengers were most likely to survive.
  - Demonstrate conditional distributions through interactive graphical displays.
  - Look for associations between variables.
  - Introduce the concept of independence.

  For more information, see page 0-5 of this guide or the DVD accompanying the Teacher's Edition.

Workshop Statistics

- Topic 7 – "Comparing Distributions: Categorical Variables" has many investigations for class room use.

Web Links

- The AP* Statistics Teacher Community archives contain a discussion about a private school where the administration announced that their school had the highest average faculty salary among eight similar schools surveyed. Further analysis revealed that the school ranked no higher than third for the average salary in every age group (20 – 29, 30 – 39, and so on).
- Gallery of Data Visualization: The Best and Worst of Statistical Graphs. Available at AP* Statistics Teacher Community Resources Library. An homage to John Tukey, this site collects current and historical graphical displays of data. It claims to have the "best statistical graphic ever drawn" as well as the "current record holder for the worst." The site is a large collection of such images, and encourages contribution. A really fun yet sophisticated collection to share with students when discussing graphical displays of data. Attribution: Michael Friendly, York University. http://www.datavis.ca/gallery/index.php

Assignments

This chapter is about 3 days work. Have students read the whole chapter in three segments. About 5 or 6 exercises a night seems to be plenty. (Remember to add one or two from Chapter 1 until those issues seem clear.)

Emphasize that the "*Tell*" is important. Be sure to ask students to read sentences they have written about the data. Take time to revise those sentences until they can be described by the 4 C's: clear, complete, concise, and in context.

Four chapter quizzes are provided.

Investigative Task

We recommend assigning written analyses from time to time, called Investigative Tasks. The first one appears here. The intent is to get students to examine data, to reach some conclusions, to then create graphical and numerical analyses, and interpret in writing what it all means (*Think, Show, Tell*). This would be the only assignment one night, due the next day. In the best of all possible worlds it works like this:

- Wednesday: hand out the task. Along with the usual reading and handful of problems for homework, ask students to read the Task to see if they understand it.
- Thursday in class: answer their questions about the Task (without giving it all away, of course).
- Thursday night: they do the Task (no other homework).
- Friday: collect the written work.
- Weekend: read and grade the Tasks. Use the rubric leading to a score on a 4-point scale, just like the AP* exam. Set the bar high - the AP* exam certainly does! Scores on this first Task will probably be pretty low.
- Monday: return the Tasks, with rubrics attached.
- Tuesday (or later): discuss the Task and the rubric.

The grading rubric we give for the Task is modeled on the AP* grading rubrics. This is typical of the way students will be graded on the exam. By showing them the scoring guidelines we help them understand the expectations and provide them with valuable feedback about how to improve their performance.

We also provide a model solution. Please keep in mind that these solutions are *not* scoring keys. In other words, students should not be graded on how closely their solutions parallel ours, rather on how well their solutions reflect the requirements set forth in the rubric. Our solution is only one example of what a student might write.

AP Statistics – Classwork* *Chapter 2*

Smoking and Education

200 adults shopping at a supermarket were asked about the highest level of education they had completed and whether or not they smoke cigarettes. Results are summarized in the table.

	Smoker	Non-smoker	Total
High school	32	61	93
2 yr college	5	17	22
4+ yr college	13	72	85
Total	50	150	200

1. Discuss the W's.

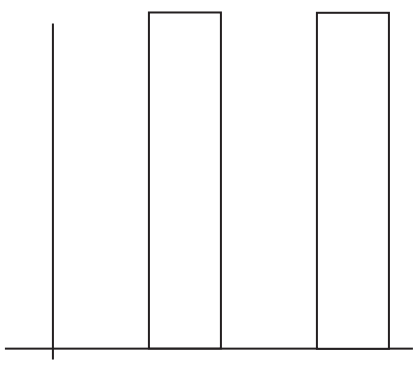
2. Identify the variables.

3. a) What percent of the shoppers were smokers with only high school educations? _____

   b) What percent of the shoppers with only high school educations were smokers? _____

   c) What percent of the smokers had only high school educations? _____

4. Create a segmented bar graph comparing education level among smokers and non-smokers. Label your graph clearly

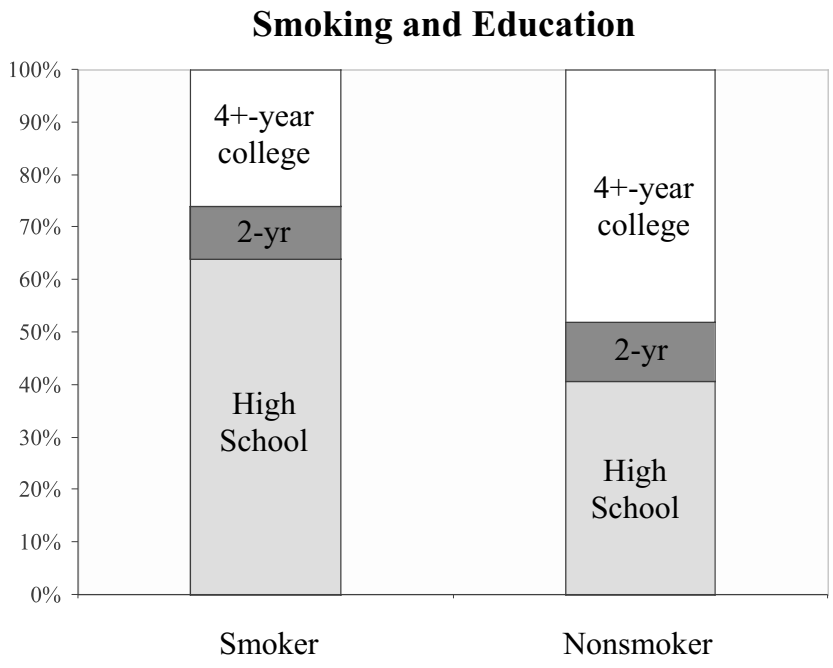


5. Do these data suggest there is an association between smoking and education level? Give statistical evidence to support your conclusion.

6. Follow-up question: Does this indicate that students who start smoking while in high school tend to give up the habit if they complete college? Explain.

AP* Statistics Classwork – Smoking and Education Key

1. Who: 200 adults
   What: education level and smoking habits
   When: not specified
   Where: shopping mall
   How: not specified. Was this a random sample, or were some people simply asked?
   Why: to examine possible links between smoking and education level
2. Categorical variables: Education level, and whether or not the person was a smoker.
3. a) $\frac{32}{200} = 16\%$
   b) $\frac{32}{93} \approx 34.4\%$
   c) $\frac{32}{50} = 64\%$
4. The segmented bar graph comparing education level among smokers and non-smokers is at the right.



5. These data provide evidence of an association between smoking and education level. 64% of smokers had only a high school diploma, while only 40.7% of non-smoker had only high school diplomas. Only 26% of smokers had four or more years of college, compared to 48% of smokers.
6. These data do not indicate that students who start smoking in high school tend to give up the habit if they complete college. These data were gathered at one time, about two different groups, smokers and non-smokers. We have no idea if smoking behavior changes over time.

AP Statistics Quiz A – Chapter 2* *Name* _____

Has the percentage of young girls drinking milk changed over time? The following table is consistent with the results from "Beverage Choices of Young Females: Changes and Impact on Nutrient Intakes" (Shanthy A. Bowman, *Journal of the American Dietetic Association*, *102*(9), pp. 1234-1239):

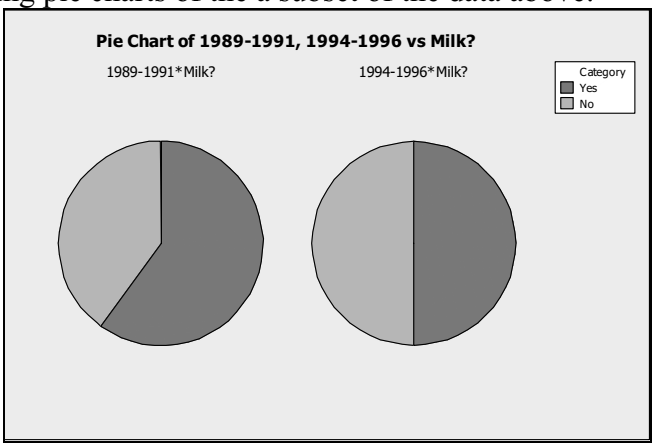
		Nationwide Food Survey Years			
		1987-1988	1989-1991	1994-1996	**Total**
Drinks Fluid Milk	Yes	354	502	366	**1222**
	No	226	335	366	**927**
	Total	**580**	**837**	**732**	**2149**

1. Find the following:
   a. What percent of the young girls reported that they drink milk? _____
   b. What percent of the young girls were in the 1989-1991 survey? _____
   c. What percent of the young girls who reported that they drink milk were in the 1989-1991 survey? _____
   d. What percent of the young girls in 1989-1991 reported that they drink milk? _____

2. What is the marginal distribution of milk consumption?

3. Do you think that milk consumption by young girls is independent of the nationwide survey year? Use statistics to justify your reasoning.

4. Consider the following pie charts of the a subset of the data above:



Do the pie charts above indicate that milk consumption by young girls is independent of the nationwide survey year? Explain.

AP Statistics Quiz A – Chapter 2 - Key*

Has the percentage of young girls drinking milk changed over time? The following table is consistent with the results from "Beverage Choices of Young Females: Changes and Impact on Nutrient Intakes" (Shanthy A. Bowman, *Journal of the American Dietetic Association*, *102*(9), pp. 1234-1239):

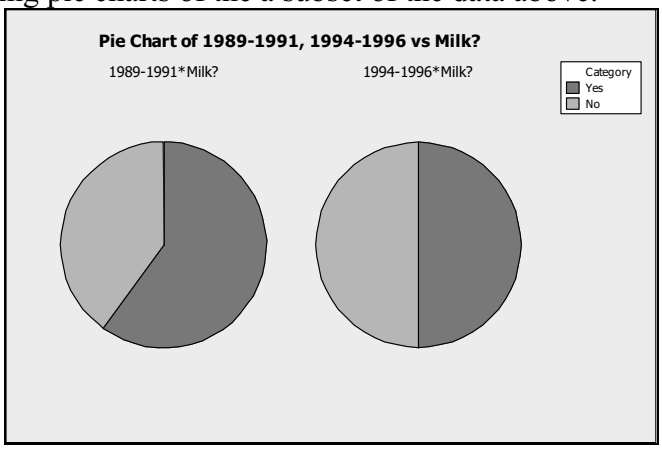
		Nationwide Food Survey Years			
		1987-1988	1989-1991	1994-1996	**Total**
Drinks Fluid Milk	Yes	354	502	366	**1222**
	No	226	335	366	**927**
	Total	**580**	**837**	**732**	**2149**

1. Find the following:
   a. What percent of the young girls reported that they drink milk? 56.9%
   b. What percent of the young girls were in the 1989-1991 survey? 38.9%
   c. What percent of the young girls who reported that they drink milk were in the 1989-1991 survey? 41.1%
   d. What percent of the young girls in 1989-1991 reported that they drink milk? 60.0%

2. What is the marginal distribution of milk consumption?
   Yes: 1,222; No: 927

3. Do you think that milk consumption by young girls is independent of the nationwide survey year? Use statistics to justify your reasoning.
   No. 56.9% of all young girls surveyed reported drinking milk, but 60% of the young girls reported drinking milk in the 1989-1991 survey. Since these percentages differ, milk consumption and year are not independent.

4. Consider the following pie charts of the a subset of the data above:



   Do the pie charts above indicate that milk consumption by young girls is independent of the nationwide survey year? Explain.
   No. It looks like there is some sort of relationship between milk consumption and nationwide survey year, since the percentage of young girls who reported drinking milk is a larger slice of the pie chart for the 1989-1991 survey than the same response for the 1994-1996 survey.

AP Statistics Quiz B – Chapter 2* *Name* _____

To determine if people's preference in dogs had changed in the recent years, organizers of a local dog show asked people who attended the show to indicate which breed was their favorite. This information was compiled by dog breed and gender of the people who responded. The table summarizes the responses.

	Female	Male	Total
Yorkshire Terrier	73	59	**132**
Dachshund	49	47	**96**
Golden Retriever	58	33	**91**
Labrador	37	41	**78**
Dalmatian	45	28	**73**
Other breeds	86	67	**153**
Total	**348**	**275**	**623**

1. Identify the variables and tell whether each is categorical or quantitative.

2. Which of the W's are unknown for these data?

3. Find each percent.
   a. What percent of the responses were from males who favor Labradors? _____
   b. What percent of the male responses favor Labradors? _____
   c. What percent of the people who choose Labradors were males? _____

4. What is the marginal distribution of breeds?

5. Write a sentence or two about the conditional relative frequency distribution of the breeds among female respondents.

6. Do you think the breed selection is independent of gender? Give statistical evidence to support your conclusion.

AP Statistics Quiz B – Chapter 2 – Key*

To determine if people's preference in dogs had changed in the recent years, organizers of a local dog show asked people who attended the show to indicate which breed was their favorite. This information was compiled by dog breed and gender of the people who responded. The table summarizes the responses.

	Female	Male	Total
Yorkshire Terrier	73	59	**132**
Dachshund	49	47	**96**
Golden Retriever	58	33	**91**
Labrador	37	41	**78**
Dalmatian	45	28	**73**
Other breeds	86	67	**153**
Total	**348**	**275**	**623**

1. Identify the variables and tell whether each is categorical or quantitative.

   Gender and Breed; both categorical.

2. Which of the W's are unknown for these data?

   We do not know how or when the people were surveyed, or where the local dog show was located.

3. Find each percent.
   a. What percent of the responses were from males who favor Labradors? 6.6%
   b. What percent of the male responses favor Labradors? 14.9%
   c. What percent of the people who choose Labradors were males? 52.6%

4. What is the marginal distribution of breeds?

   There were 132 Yorkshire terrier responses, 96 Dachshund responses, 91 Golden Retriever responses, 78 Labrador responses, 73 Dalmatian responses, and 153 Other responses.

5. Write a sentence or two about the conditional relative frequency distribution of the breeds among female respondents.

   Among females, 20.9% chose Yorkshire Terriers, 14.2% Dachshunds, 16.7% Golden Retrievers, 10.6% Labs, and 12.9% Dalmatians. The remaining 24.7% of females preferred other breeds.

6. Do you think the breed selection is independent of gender? Give statistical evidence to support your conclusion.

   The breed selection does not appear to be independent of gender. Overall, 56% of the respondents were females, but females were over-represented among those who favored Golden Retrievers (64%) and Dalmatians (62%), yet a much lower percentage (47%) among those who chose Labradors.

AP Statistics Quiz C – Chapter 2* *Name* _____

In order to plan transportation and parking needs at a private high school, administrators asked students how they get to school. Some rode a school bus, some rode in with parents or friends, and others used "personal" transportation – bikes, skateboards, or just walked. The table summarizes the responses from boys and girls.

	Male	Female	Total
Bus	30	34	**64**
Ride	37	45	**82**
Personal	19	23	**42**
Total	**86**	**102**	**188**

1. Identify the variables and tell whether each is categorical or quantitative.

2. Which of the W's are unknown for these data?

3. Find each percent.
   a) What percent of the students are girls who ride the bus? _____
   b) What percent of the girls ride the bus? _____
   c) What percent of the bus riders are girls? _____

4. What is the marginal distribution of gender?

5. Write a sentence or two about the conditional relative frequency distribution of modes of transportation for the boys.

6. Do you think mode of transportation is independent of gender? Give statistical evidence to support your conclusion.

AP Statistics Quiz C – Chapter 2 – Key*

In order to plan transportation and parking needs at a private high school, administrators asked students how they get to school. Some rode a school bus, some rode in with parents or friends, and others used "personal" transportation – bikes, skateboards, or just walked. The table summarizes the responses from boys and girls.

	Male	Female	Total
Bus	30	34	**64**
Ride	37	45	**82**
Personal	19	23	**42**
Total	**86**	**102**	**188**

1. Identify the variables and tell whether each is categorical or quantitative.

   Gender and mode of transportation, both categorical.

2. Which of the W's are unknown for these data?

   We don't know how or when the students were surveyed, nor where the school is.

3. Find each percent.
   a) What percent of the students are girls who ride the bus? 18.1%
   b) What percent of the girls ride the bus? 33.3%
   c) What percent of the bus riders are girls? 53.1%

4. What is the marginal distribution of gender?

   There are 86 males and 102 females.

5. Write a sentence or two about the conditional relative frequency distribution of modes of transportation for the boys.

   More boys (43%) caught rides to school than any other means of transportation. 35% rode the bus while only 22% used personal transportation like biking, skateboarding, or walking.

6. Do you think mode of transportation is independent of gender? Give statistical evidence to support your conclusion.

   The way students get to school does seem to be independent of gender. Overall, 34% of students ride the bus, compared to 35% of the boys and 33% of the girls. 44% of all students caught rides with someone and 22% used personal transportation, almost the same as the percentages for boys (43% and 22%) or girls (44% and 23%) separately. These data provide little indication of a difference in mode of transportation between boys and girls at this school.

AP* Statistics Quiz D – Chapter2 ***Name*** _____

A research company frequently monitors trends in the use of social media by American Adults. The results of one survey of 1846 randomly selected adults looked at social media use versus age group. The table summarizes the survey results.

		Age Group				
		18-29	30-49	50-64	65+	**Total**
Uses Social Media	Yes	328	417	288	114	**1147**
	No	67	125	265	242	**699**
	Total	**395**	**542**	**553**	**356**	**1846**

1. Identify the variables and tell whether each is categorical or quantitative.

2. Which of the W's are unknown for these data?

3. Find each percent.
   a) What percent of adults surveyed are social media users aged 30-49? _____
   b) What percent of the social media users are aged 30-49? _____
   c) What percent of adults aged 30-49 are social media users? _____

4. What is the marginal distribution of age groups?

5. Write a sentence or two about the conditional relative frequency distribution of ages of social media users.

6. Do you think social media use is independent of age? Give statistical evidence to support your conclusion.

AP Statistics Quiz D – Chapter2* *Name* _____

A research company frequently monitors trends in the use of social media by American Adults. The results of one survey of 1846 randomly selected adults looked at social media use versus age group. The table summarizes the survey results.

		Age Group				
		18-29	30-49	50-64	65+	**Total**
Uses Social Media	Yes	328	417	288	114	**1147**
	No	67	125	265	242	**699**
	Total	**395**	**542**	**553**	**356**	**1846**

1. Identify the variables and tell whether each is categorical or quantitative.

   Age is numerical, but the grouping treats it as categorical, and social media use is categorical.

2. Which of the W's are unknown for these data?

   We don't know when the adults were surveyed.

3. Find each percent.
   a. What percent of adults surveyed are social media users aged 30-49? 22.6%
   b. What percent of the social media users are aged 30-49? 36.4%
   c. What percent of adults aged 30-49 are social media users? 76.9%

4. What is the marginal distribution of age groups?

   There were 395 adults aged 18-29, 542 aged 30-49, 553 aged 50-64, and 356 that were 65 or older.

5. Write a sentence or two about the conditional relative frequency distribution of ages of social media users.

   More social media users in the survey (36.4%) were aged 30-49 than any other age group. Next was the 18-29 age group at 28.5%, then the 50-64 group at 25.1%, and the smallest group of social media users (9.6%) was the 65 and older group.

6. Do you think social media use is independent of age in the population of American adults? Give statistical evidence to support your conclusion.

Social media use does not appear to be independent of age. Overall, 21.4% of adults surveyed were 18-29 years old, but 28.6% of social media users are in that age group. And 19.2% of adults surveyed were 65 and older, but only 9.9% of social media users were. In general, older groups seem to be underrepresented among social media users.