Chapter 3
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

This is the first substantial chapter, and it gets pretty involved right away. Most texts do not deal with conditional distributions, independence, and confounding (Simpson's paradox) so soon. 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 3 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.

It is probably beginning 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.

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

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

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.)

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.