Summarizing Categorical Data

Math 145

Student Survey

`Lock5withR` includes a data set with data from a student survey. Two of the variables it includes are

- `Gender` Student's gender: 'F' or 'M'
- `Award` Prefered award: 'Academy' 'Nobel' 'Olympic'

1. Each student was asked whether they would rather win a Academy Award, a Nobel Prize, or an Olympic Gold medal. `Award` records their answers. If the members of your group were added to the data set (just for these two variables), what would the new rows of data look like?
2. Write down some questions we might answer using the `Sex` and/or `Award` variables. Which of your questions/answers need both variables? Which only require one of the variables?

Our main tools for investigating question like this will be `tally()` for numerical summaries and `gf_bar()` for bar plots.

Answering some questions

Which award is most popular?

1. Run these commands to find out.

```
library(Lock5withR)           # Load the package that contains the data
gf_bar( ~ Award, data = StudentSurvey)
 tally( ~ Award, data = StudentSurvey)
```

Does award preference vary by sex?

2. Run the commands below to make numerical tables of different kinds. Feel free to experiment with other versions as well.

```
tally( ~ Award | Sex, data = StudentSurvey, format = "percent")
tally( Award ~ Sex, data = StudentSurvey, format = "prop")
tally( Award ~ Sex, data = StudentSurvey, margins = TRUE)
tally( Award ~ Sex, data = StudentSurvey, margins = TRUE, format = "percent")
```

3. Which tables do you like best for this question?
4. When you use proportions or percents, be sure to check which things add up to 1 or 100%. (Possible answers: rows, columns, or the whole table.)

`gf_bar()` can create a variety of bar charts.

5. Try these examples.

```
gf_bar( ~ Award, data = StudentSurvey, fill = ~Sex)
gf_bar( ~ Award, data = StudentSurvey, fill = ~Sex, position = "dodge")
gf_bar( ~ Award | Sex, data = StudentSurvey, fill = ~Sex)
gf_bar( ~ Sex, data = StudentSurvey, fill = ~ Award)
```

6. Which do you like best for answering this question?

We can also use `gf_props()` or `gf_percents()` to make bar charts on a proportion or percent scale.

7. Try these (our use `gf_percents()` instead of `gf_props()` if you want percents instead of proportions):

```
gf_props( ~ Award, data = StudentSurvey, fill = ~Sex)
gf_props( ~ Award, data = StudentSurvey, fill = ~Sex, position = "dodge")
gf_props( ~ Award | Sex, data = StudentSurvey, fill = ~Sex)
gf_props( ~ Sex, data = StudentSurvey, fill = ~ Award)
gf_props( ~ Award, data = StudentSurvey, fill = ~Sex, position = "dodge",
          denom = ~fill)
gf_props( ~ Sex, data = StudentSurvey, fill = ~ Award, denom = ~x)
```

8. In each case, determine which segments add to 1 (or 100 percent).
9. What does `denom` do?

What about pie charts?

10. Although common in newspapers, magazines, and TV, pie charts are rarely (but not never) used by statisticians. Why do you think this is? In what situations could a pie chart work reasonably well?

Notation Note

It is important to distinguish between the proportion of people in the **population** who would answer a certain way and the proportion of people in our **sample** who did answer a certain way. We have terminology and notation to distinguish between the two.

summary	parameter	statistic
proportion	p	$\hat{p}$ (read: p hat)
mean	μ (Greek letter "mu")	$\overline{x}$ (read: x bar) or $\hat{\mu}$
standard deviation	σ (Greek letter "sigma")	s or $\hat{\sigma}$

The notation for median and iqr is less standardized.

Pew Study

A nationwide US telephone survey conucted by the Pew Foundation in October 2010 asked 2625 adults ages 18 and older "Some people say there is only one true love for each person. Do you agree or disagree?" The survey participants were selected randomly by landlines and cell phones. In addition to the answer to the question, surveyors recorded the sex of each person surveyed.

5. What is the population for this study?
6. What are some potential sources of bias in this study? Do you expect the bias to be relatively small or potentially large?
7. What are the cases in this study?
8. What are the variables? Are they categorical or quantitative?
9. Write down what the first few rows of the data set would look like if your group members were the first few cases.
10. Of those surveyed, 735 people agreed, 1812 disagreed, and 78 answered "don't know".
    a. Display this information in a **frequency table** that lists each possible response and the number of people who gave that response.
    b. What should the sum of the frequencies equal.
    c. Add an extra row or column to your table that gives the **relative frequency** (i.e., the proportion) who gave each response.
    d. What should the sum of the relative frequencies equal?
    e. Write the R command that computes the frequency table and the relative frequency table. (Note: We don't have access to the raw data for this example, so you can't actually run your command.)
    f. What sort of plot can we use to visualize this data. Write the R command to create such a plot. (Again, you won't be able to run the command because we don't have the raw data.)

Two-way table

Here is the two way table for the Pew study.

answer	Male	Female
agree	372	363
disagree	807	1005
don't know	34	44

11. Use the table to answer the following questions
    a. How many males were called? How many females? How might you add this information to the table? (These numbers are sometimes called **marginal totals**, do you see why?)
    b. What proportion of females agree?
    c. What proportion of those who agree were female?
    d. Which of b and c is a more interesting number in this study? why?
    e. What proportion of males agree?
    f. How would you describe the difference between males and females in this sample?
    g. If you could ask everyone in the population, do you think the difference would be exactly the same as in the Pew study? Pretty close to the same? or Possibly quite different? Explain.
    h. What proportion of the responders were female?
    i. Create a relative frequency table these data. There are at least 3 ways you might do this. Create all three. Which one is the most useful one in this situation? Why?