First Day Statistics Activity – Grouping Qualitative Data

R.B. Campbell
Department of Mathematics
University of Northern Iowa
campbell@math.uni.edu

Published: April 2014

Overview of Lesson
This is suggested as a first day activity because it is a nice way to get the students involved, but also because it provides a foundation for distinguishing between qualitative and quantitative data, and how they can be summarized and analyzed (a topic that is encountered early in a statistics course). This lesson illustrates how categorical data can be grouped to make the data comprehensible. It also allows for discussion of the validity of such groupings. Students will select four adjectives which they believe characterize themselves from a list which is provided, in order to characterize the class as a whole. They will then discover that there are too many adjectives to comprehend the data and try to group the words into categories so that the data is easier to comprehend. Finally, they will be given categories which are provided with this exercise, but need to decide whether those are meaningful groupings and whether any important information has been lost by the grouping. This also illustrates how important it is to carefully pose questions in a survey if you want to analyze the results, in particular that it may be necessary to have subjects choose from a list of responses rather than allow free response.

GAISE Components
This investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. The four components are: formulate a question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level B activity.

Common Core State Standards for Mathematical Practice
1. Make sense of problems and persevere in solving them.
3. Construct viable arguments and critique the reasoning of others.

Common Core Standards Grade Level Content (High School)
S-ID. 5. Summarize categorical data for two categories in two way frequency tables. Interpret relative frequencies in the context of the data (including joint, marginal, and conditional relative frequencies). Recognize possible associations and trends in the data.

NCTM Principles and Standards for School Mathematics
Data Analysis and Probability Standards for Grades 9-12

Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them:

- understand the meaning of measurement data and categorical data.

Select and use appropriate statistical methods to analyze data:

- for univariate measurement data, be able to display the distribution.

Develop and evaluate inferences and predictions that are based on data:

- understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference.

Prerequisites

There are no prerequisites other than being able to count. Students' vocabulary should include the twenty words and four grouping words provided (on the Activity Sheet at the end of this lesson), but nobody knows what *prudent* really means.

Learning Targets

Students should learn i) that it may be necessary to lose some information (details) in order to be able to understand/communicate the essence of the information (one can be overwhelmed with data; the first paper in descriptive statistics (John Graunt, Natural and Political Observations Mentioned in a following Index, and made upon the Bills of Mortality) "reduced several confused Volumes into a few perspicuous [another vocabulary word for the students] Tables, and abridged such Observations as naturally flowed from them, into a few succinct Paragraphs …") and ii) that it may be possible to group categorical data in a meaningful way which facilitates communication/understanding, but that may be difficult. Students will also learn that it may be necessary to specify the responses subjects may choose in order to be able to analyze the resultant data.

Time Required

This activity could easily consume a class period, but it could also be abbreviated if the instructor collected the data one class and processed it before the next class. Or the time allowed for students to reflect on the data and make suggestions could be abbreviated. I would allow at least 15 minutes for having students choose from the 20 provided words, tallying their choices, acknowledging that the tally is hard to comprehend, showing the grouped tally, acknowledging that it is easier to comprehend, yet they may be wary of the groupings.

Materials Required

Slips of paper so students can provide their choices (the choices could be projected at the front of the room, or listed on the slips of paper), tally sheets if the students are going to compile responses in groups before reporting to the class as a whole, and a way to display the ungrouped and grouped results (chalkboard, whiteboard, document camera).

Instructional Lesson Plan

The GAISE Statistical Problem-Solving Procedure

I. Formulate Questions(s)

Discuss that people have different social styles, and ask if it is possible to characterize the general nature of people in the class, i.e., whether there are characteristics shared by many people in the class. This is a modification of a get acquainted exercise, and can be used as an excuse to discuss that different people act and think differently, hence we must accommodate those differences when interacting with our peers. You will also need to discuss that a general description of a group may not be an accurate description of everyone in the group. The goal becomes to find a few adjectives which describe the general nature of the people in the class, recognizing that it will not be an accurate description of every individual. A second question can be whether there is a difference in the social styles of the boys and girls in the class, or perhaps between the people in two different classes.

II. Design and Implement a Plan to Collect the Data

A preliminary plan could be to have each student give a couple of adjectives which describe their social style (i.e., the nature of their actions and interactions with others). The teacher will write the adjectives on the board, or perhaps just a couple dozen of them. This should result in a lot of words, and the realization that a concise description of the people is not evident. This illustrates that one should consider how the data is going to be analyzed before it is collected. The teacher can recite the quote from R.A. Fisher: To call in the statistician after the experiment is done may well be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.

The teacher will then announce that someone who has given serious thought to this question has selected 20 words (provided below) which people can choose from to describe their social style. This will allow discussion that sometimes it is necessary to restrict the responses that subjects may choose in order to analyze the data. Further, the subjects (students in the class) are asked to choose four adjectives which best describe themselves. Choosing four will force them to identify their most important characteristics, and having everyone choose four will give equal weight to everyone in the class.

Ask students to choose four of the following twenty social style adjectives which they believe best describe them: agreeable, cautious, conceptual, competitive, creative, decisive, dependable, dramatic, enthusiastic, efficient, independent, imaginative, logical, loyal, organized, patient, persistent, practical, prudent, and supportive. Have the students write their four choices on a slip of paper. If you want to investigate whether responses differ by gender, have the students also write their gender on the slips of paper.

III. Analyze the Data

Tally the number of students who chose each adjective. You may have students read off their choices and tally them as they read them, have students compile sub-tallies in groups, or collect the sheets of paper and compile the tally overnight; your choice may be dictated by the size of your class and how much time you have allotted for the activity. The result may be something like:

2	agreeable	5	creative	2	enthusiastic	2	logical	4	persistent
1	cautious	1	decisive	5	efficient	8	loyal	4	practical
0	conceptual	8	dependable	5	independent	4	organized	0	prudent
10	competitive	4	dramatic	5	imaginative	4	patient	2	supportive

If you collect gender, the results might look something like

1/1	agreeable	2/3	creative	1/1	enthusiastic	2/0	logical	3/1	persistent
0/1	cautious	0/1	decisive	4/1	efficient	5/3	loyal	4/0	practical
0/0	conceptual	1/7	dependable	5/0	independent	3/1	organized	0/0	prudent
5/5	competitive	4/0	dramatic	4/1	imaginative	3/1	patient	1/1	supportive

with the number of males before the solidus and the number of females after the solidus.

Ask the students what they see. They may say they see competitive, dependable, and loyal; with no prudent. But they can easily be overwhelmed by twenty different adjectives and unable to comprehend an overall characterization of the class. If you collected gender, the differences between males and females will be confounded by different numbers of males and females providing their selections. Further comments on gender differences (or differences between two classes) are under Possible Extensions below.

You can ask them if they can group the adjectives together into a few categories based on similarity of meaning. This may be done in small groups. The result will probably be mainly frustration, but active frustration. Then tell them that this exercise was designed by someone who believed the adjectives could be grouped into four categories: Amiable (agreeable, dependable, loyal, patient, supportive), Analytical (cautious, logical, organized, persistent, and prudent), Driver (competitive, decisive, efficient, independent, and practical), and Expressive (conceptual, creative, dramatic, enthusiastic, and imaginative). There may be disagreement about whether these groupings are meaningful, but there should be agreement that creating such groupings is difficult. Then the data will then look something like:

24	Amiable	11	Analytical	25	Driver	16	Expressive
2	agreeable	1	cautious	10	competitive	0	conceptual
8	dependable	2	logical	1	decisive	5	creative
8	loyal	4	organized	5	efficient	4	dramatic
4	patient	4	persistent	5	independent	2	enthusiastic
2	supportive	0	prudent	4	practical	5	imaginative

or

11/13	Amiable	8/3	Analytical	18/7	Driver	11/5	Expressive
1/1	agreeable	0/1	cautious	5/5	competitive	0/0	conceptual
1/7	dependable	2/0	logical	0/1	decisive	2/3	creative
5/3	loyal	3/1	organized	4/1	efficient	4/0	dramatic
3/1	patient	3/1	persistent	5/0	independent	1/1	enthusiastic
1/1	supportive	0/0	prudent	4/0	practical	4/1	imaginative

This table provides an easy to comprehend presentation of the information which was collected. Further analysis is suggested under Possible Extensions below.

IV. Interpret the Results

It will be clear to the students that they are amiable and drivers, but less expressive and analytical (the latter being qualities we want in mathematics students). But it will be less clear that they understand, or believe, the four groupings. If they understand and believe the groupings, they will have a concise and easy to communicate description of the students in the class. If they do not understand or believe the groupings, they will be left with the 20 adjectives which is an overwhelming amount of data to easily comprehend.

Differences between genders is a more subtle matter to eyeball from the data. The above data suggests that the women are more dependable when you look at the twenty adjectives. For the four groupings, the men appear to be more analytical, driver, and expressive by calculation of the ratio of males to females for each category, but this reflects the ratio of 12 boys to 7 girls in the class. Adjusting for the number of boys and girls will still have the males with an excess of analytical, driver, and expressive, and the females with an excess of amiable. Although this might be consistent with some sex stereotypes, in fact gender was randomly assigned to this data after it was collected (i.e., is not the gender of the person who supplied the data). Trying to eyeball whether there is a sex difference in social styles illustrates the need for inferential statistics. Although the ratios .49, 1.56, 1.55, and 1.28 (11/13, 8/3, 18/7, and 11/5 adjusted by the 12 to 7 sex ratio of respondents) may seem quite different from 1, the chi-square test does not show a significant association between gender and the four categories for this data set. Errors of judging too little deviation from expected values have also occurred; John Arbuthnott in the first paper employing inferential statistics (An Argument for Divine Providence, taken from the Constant Regularity observed in the Births of both Sexes) correctly demonstrated that more boys than girls are born, but asserted that there was less variation in the sex ratio than would occur by chance, when in fact there was more variation (he developed a test of hypothesis for the mean of the sex ratio, but a test for the variation of the sex ratio had not been developed yet).

Assessment

1. Which of the 20 adjectives were chosen by the most students? Did half of the students choose any of the adjectives?

2. Which group was the most common choice? Was the most common adjective in that group?

3. Is there a synonym for prudent people may have been more likely to choose (did the particular choice of words effect whether people chose them)?

4. How can you group models of cars into fewer categories?

5. What information was lost when the tally of the twenty adjectives was created?

6. What further information was lost when the data was summarized as four categories?

7. Give another example when it is appropriate/necessary to provide choices for the subjects rather than allow free response.

8. Does providing answers for this exercise (the 20 adjectives) clarify the question or bias the responses?

Answers

1. Competitive, dependable, and loyal were the most common choices, only competitive was chosen by half the students (there were 19 students).

2. Driver was the most common group, and in this case the most common choice, competitive, was in that group. But the next two choices (dependable and loyal) were both in Amiable.

3. Perhaps 'careful', but 'circumspect' and 'discreet' may not be more likely to be chosen. 'Cautious' only got one vote so that the choice to list 'prudent' may not be why few people chose that nature of behavior, it may be a behavior that people do not identify with.

4. Models of cars could be grouped by manufacturer (Ford, GM, Toyota, etc.), country of origin (US, Germany, Japan, etc.), style (sub-compact, sedan, SUV, etc.), or with quantitative descriptions such as engine size, miles per gallon, price. Whether a grouping is appropriate will depend on why you are interested in car models.

5. Which words were chosen together (i.e., on the same slip of paper). If you were interested in which words were chosen together, a different presentation/analysis of the data would have been appropriate.

6. The frequency with which the various words in each category were chosen. If you were interested in which aspects of the four categories people identified with, a different analysis would have been appropriate.

7. Answers will vary, but I once asked students for their hair color intending to make a pie chart, and got about 17 colors from 51 students including light brown, dirty blond, sandy blond, sandy brown, …. I really did not know the distinctions between those colors, or which were darker.

8. Both. I assume many students are confused by 'what social style adjectives' means, and the options are useful for indicating what is desired. But once the genre of adjective is understood, individuals cannot choose adjectives such as cordial, systematic, assertive, or insightful which fit into the categories, yet have different meanings than the choices offered. But offering 5 words per category removes the bias of more choices being available for some categories.

Possible Extensions

Often student heights and weights are collected to make stem-and-leaf plots and histograms. These social style adjectives could be collected at the same time. However, do not ask students to provide information which may be embarrassing to them (in some instances students will be embarrassed by their weight).

If one collected gender with this data, or collected data from two classes, they can return to this data set if they cover using the chi-square distribution for test of independence or test of homogeneity. There is no difference between the mechanics of tests of independence and tests of homogeneity, they are distinguished by the manner in which the data was collected. If one collected gender with the adjectives, a test of independence is performed because two categorical variables were collected about each subject. If one collected data without gender for two classes, a test of homogeneity is performed because one categorical variable is collected within each of

two groups. This can be used to illustrate the difference between a test of homogeneity and a test of independence.

To employ the chi-square test, one should have all or most expected values greater than or equal to 5, which will not be the case if all 20 adjectives are used (unless you have very large classes). Many texts suggest grouping the small classes together to get all or most expected values greater than or equal to 5. This exercise offers a meaningful way to group classes together rather than just a functional method.

One could have a discussion of when information is lost. I tell my classes that although one usually says that no information is lost when you put it into a stem-and-leaf plot, if I asked my students their weights in order from the front of the class to the back, and was interested in whether heavier students sat in the front; that information would be lost when a stem-and-leaf plot was constructed. But if that was not the question of interest, no useful information was lost. It is not important whether information is lost, but whether relevant information is lost.

One could have an extended discussion of when it is necessary/appropriate to group qualitative data. The answer to this, and how to group the data, will depend upon why you want the data. One could group desserts by calories if they were on a diet or by price if they were on a budget. One could also group desserts by qualitative categories such as fruit, dairy, or pastry.

Also one could discuss the trade-off between retaining information and communicating information with quantitative data. The weights of 100 football players would be an overwhelming amount of data, but just the mean or median would miss much of the nature of the distribution. You could compromise with the five number summary or the mean and standard deviation.

References

(These are readily available on the web.)
1. Graunt, J. (1662). Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality. Martyn and Allestry, London.
2. Arbuthnott, J. (1710). An Argument for Divine Providence, Taken from the Constant Regularity Observed in the Births of Both Sexes. Phil. Trans. 27:186-190.

Grouping Qualitative Data Activity Sheet

You will need to prepare (either a single copy to display, several for tally sheets for student groups, and/or individually for the students) an ungrouped (e.g., alphabetical order) list of the 20 words:

	agreeable		creative		enthusiastic		Logical		persistent
	cautious		decisive		efficient		Loyal		practical
	conceptual		dependable		independent		organized		prudent
	competitive		dramatic		imaginative		patient		supportive

You will also need to prepare a grouped list of the words:

	Amiable		Analytical		Driver		Expressive
	agreeable		cautious		Competitive		conceptual
	dependable		logical		Decisive		creative
	loyal		organized		Efficient		dramatic
	patient		persistent		Independent		enthusiastic
	supportive		prudent		Practical		imaginative

You can later write your counts on these sheets (the second day if you compile the data overnight).