Math Lesson 7: Two Types of Data --
Numerical (Quantitative) and Categorical (Qualitative)

Discussion

So far in our math lessons we have been dealing with data (*numbers in context*) that is of a specific type. It is called **quantitative data**, or **numerical data**. These phrases both mean the same thing. There is another type of data, however, and that is **qualitative data**, often called **categorical data**. The first type, numerical data (which we have dealt with previously), is defined as data that consist of numerical measures or counts. Categorical data, on the other hand, consists of attributes, labels or non-numerical categories.

If you were interested in counting the number of people in each movie theater at Victoria Ward Center on Friday nights at 8:00 p.m., that would be numerical or quantitative data. With these data values it makes sense to calculate meaningful measure of center, examine the spread of the data, etc. The same would be true if you recorded the weights of caught fishes that are brought to the dock in Kewalo Basin each afternoon. That's numerical data.

If, however, you were interested in the types of fishes caught, that would be categorical (qualitative) data. You would record whether the fishes were papio, ono, hinalea, uku, aholehole, ulua, mahimahi, opakapaka, etc. Or, you might want to group the caught fish as to size - small, medium or large. That would be categorical data, as each subject of your study is placed into a category. Does it make sense, when you are looking at types of fish, to talk about an "average type"? No. So we analyze categorical data differently than we analyze numerical data.

In previous lessons, we looked at various ways to graphically represent numerical data. We explored dot plots, stem and leaf plots, histograms created by the calculator, and box plots. In each case we learned something about a representative summary number (or estimated center) for a data set, which included the mode, median and mean. We were also able to make a determination of how the data were spread out – how far away from the center the data entries were. For this we used range, quartiles, inter-quartile range, and standard deviations. Box-plots and histograms were good ways to visually assess the spread of a data set.

In analyzing categorical data, there are really only two ways to graph the data: pie charts and bar charts. The only meaningful measure of "center" would be the mode. Let's now examine the differences between pie charts and bar charts, and see which might be preferable for a given situation.

Activity

DDM-Mathematics in a World of Data
Lesson 4, "Categorical and Measurement Data", pp.21-23

Discussion

All categorical data can be summarized in a frequency distribution. To create a frequency distribution, list the categories on the left with the corresponding frequencies on the right. Let's use the following data set representing the number of Nobel Prize laureates by country during the years 1901-1993.

United States	170	France	24
United Kingdom	69	USSR	10
Germany	59	Other	88

Frequency distribution

Countries	Frequency
United States	170
United Kingdom	69
Germany	59
France	24
USSR	10
Other	88

Pie Charts: Categorical data are represented well in pie charts when the number of categories is relatively small. With a large number of categories, the slices in the pie chart would be too small to give a good visual representation of the distribution of the data. In this case, with six categories, a pie chart is a good choice. To make a pie chart, first find the ***relative frequency*** of each category. Note that the sum of the frequencies (170 + 69 + 59 + 24 + 10 + 88) is 420. That is, n = 420. By dividing each frequency by 420, we obtain the relative frequency for each category.

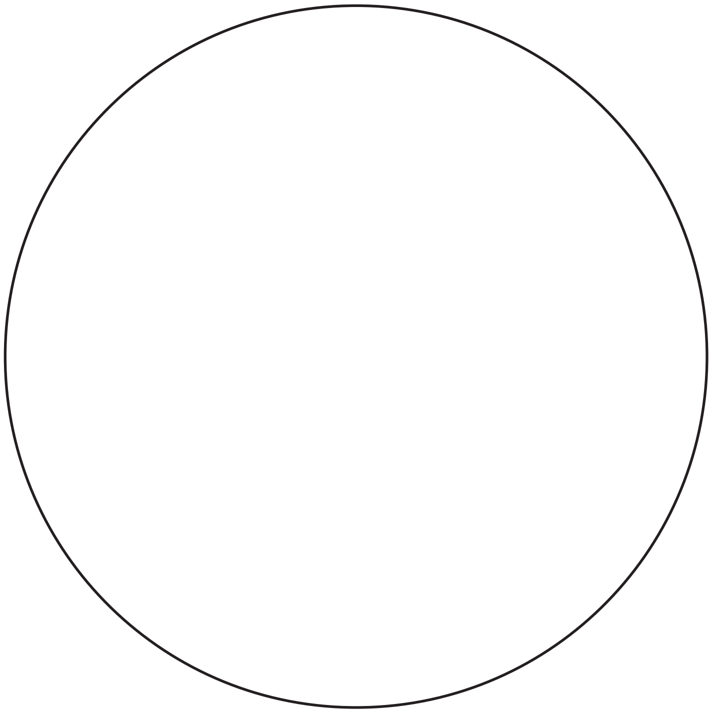
Countries	Frequency	Relative frequency
United States	170	.405
United Kingdom	69	.164
Germany	59	.14
France	24	.057
USSR	10	.024
Other	88	.210

What do you think the sum of the relative frequencies should be? **[1.00]** Sometimes the sum is not exactly 1.00, but that is due to "rounding error."

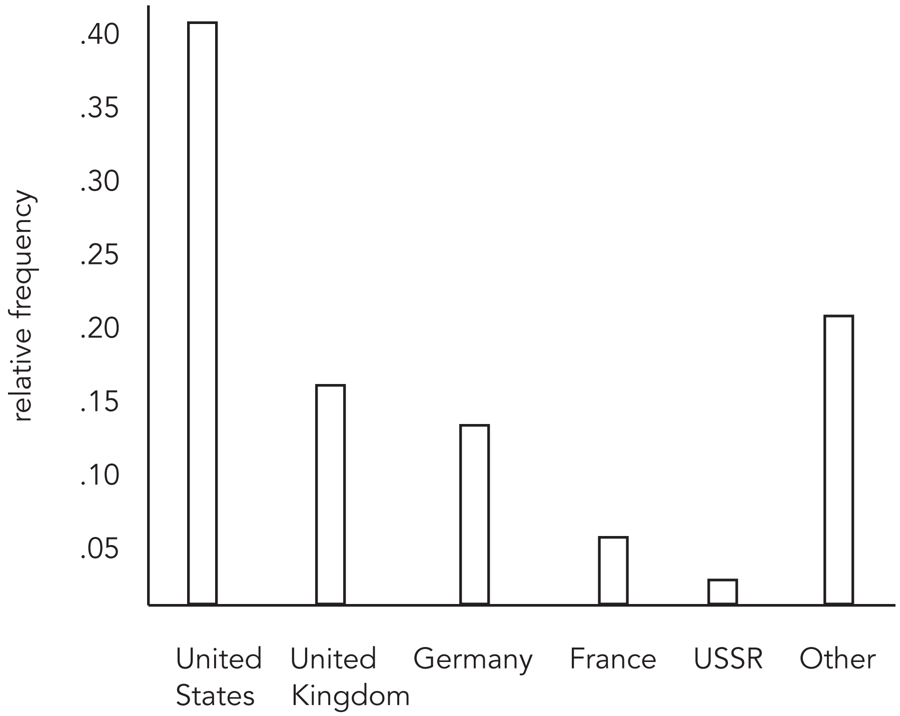
The relative frequencies correspond to the percentages of the data that fall into each category. To construct a pie chart we would need to know how many degrees are needed for the central angle of each "slice" or "wedge" of the circle. Since there are 360° in a circle, we multiply the relative frequency by 360° to obtain the number of degrees.

Countries	Frequency	Relative Frequency	Degrees
United States	170	.405	145.8
United Kingdom	69	.164	59.0
Germany	59	.14	50.4
France	24	.057	20.5
USSR	10	.024	8.6
Other	88	.210	75.6

Note that by adding the degrees in the last column we get 359.9 instead of 360. This is because of rounding error. Now, using a protractor and a compass, we can construct a pie chart, and by coloring the individual segments differently and labeling them, we can get a clear picture of the amount of how the Nobel Prizes were awarded to different countries.



Bar Charts: Another way to represent categorical data is with a bar chart. Unlike histograms, bar charts have bars that are separated from one another. In a histogram, the bars touch each other. Along one axis, the categories are listed. Along the other (perpendicular) axis, the frequencies or relative frequencies are noted.



Activity

Use the summary data from our exploration on the reef in July given below to construct a bar chart showing the types of reef coverage. Why would a bar chart be preferable to a pie chart?

Site name:	Fringe reef	Country/Island:	kaneohe bay
Depth:	3 meters	Date:	7/7/2005
TS/TL:		Data recorded by:	teachers group 1
Time:	1000h		

Substrate Code

HC	hard coral	**SC**	soft coral	**RKC**	recently killed coral
NIA	nutrient indicator algae	**SP**	sponge	**RC**	rock
RB	rubble	**SD**	sand	**SI**	silt/clay
OT	other				

Total S1		Total S2		Total S3		Total S4		Grand total	
HC	19	HC	17	HC	26	HC	12	HC	74
SC	0	SC	0	SC	0	SC	0	SC	0
RKC	2	RKC	10	RKC	3	RKC	5	RKC	20
NIA	4	NIA	0	NIA	1	NIA	6	NIA	11
SP	0	SP	0	SP	0	SP	0	SP	0
RC	0	RC	0	RC	1	RC	0	RC	1
RB	0	RB	0	RB	0	RB	2	RB	2
SD	15	SD	13	SD	8	SD	2	SD	38
SI	0	SI	0	SI	0	SI	0	SI	0
OT	0	OT	0	OT	1	OT	0	OT	1
#	40	#	40	#	40	#	27		

The data above are for the fringe reef. Now suppose we also had data for the patch reef, and we wanted to compare the coverage on the two types of reef. A good way to graphically investigate this is to make a comparative bar chart, where we have two bars situated next to each other (and touching) above each category—one for the fringe reef and another for the patch reef—in differing colors. Be sure to include a color legend! Always label your graphs so that anyone reading them will know exactly what they represent. Here's the data for the patch reef: Construct a comparative bar chart for the fringe reef and the patch reef below.

Patch Reef Data

Total S1		Total S2		Total S3		Total S4		Grand total	
HC	13	HC	8	HC	19	HC	11	HC	51
SC	0	SC	0	SC	0	SC	0	SC	0
RKC	5	RKC	12	RKC	6	RKC	3	RKC	26
NIA	14	NIA	0	NIA	6	NIA	9	NIA	29
SP	0	SP	0	SP	0	SP	0	SP	0
RC	0	RC	0	RC	4	RC	0	RC	4
RB	0	RB	0	RB	0	RB	4	RB	4
SD	23	SD	11	SD	1	SD	8	SD	43
SI	0	SI	0	SI	0	SI	0	SI	0
OT	7	OT	0	OT	1	OT	0	OT	8
#	62	#	31	#	37	#	35		165

When we want to see if there is some sort of association between two different categorical variables, we can make a two-way table for the categorical data as follows.

Does political party affiliation have anything to do with gender? Are males more likely than females to be democrats? Suppose we visit a gathering of 100 people at an OHA (Office of Hawaiian Affairs) reception one evening. After asking each person his or her political party affiliation and noting the gender, we come up with the following numbers which we represent in a two-way table.

	Democrat	Republican	Independent	**Total Counts**
Male	29	10	3	42
Female	36	22	0	58
Total Counts	65	32	3	100

Note that the grand total in the bottom right hand corner should be the sum of the margin totals for both the rows and the columns. What is the proportion of democrats at the gathering? **[65 + 100 = .65]** What is the proportion of males at the gathering? **[42 + 100 = .42]** These results can similarly be interpreted as probabilities. In other words, based on this data, the probability of being a democrat at the OHA gathering is .65. The probability of being a male at the OHA gathering is .42. How can we determine if being a male means you are more likely to be a democrat? That is like asking, "What is the probability that, given that a person is a male, he is a democrat?" Since we are only looking at the males, we take the margin total for males (42) and use it as the denominator of our fraction. Then we use the number of male democrats (intersection of the male row and democrat column – 29) and use it as the denominator of our fraction. **[29 + 42 = .69]** So, the answer to the question "What is the probability that, given that a person is a male, he is a democrat?" is .69. Now we can ask "What is the probability that, given that a person is a female, that she is a democrat?" We would calculate 36 ÷ 58 = .62. What is your conclusion?

We can answer a different question by using the column totals as our denominators. The question would then be: "What is the probability that, given that a person is a democrat, he is male?" Now our calculation would be 29 ÷ 65 = .45. Can you find the probability that a person is a female given that she is a republican?

Practice 1. Suppose we are interested in learning about the effects of parents' smoking habits on their children. If parents smoke, are their children more likely to become smokers? What if only one parent smokes? Do children of non-smoking parents ever become smokers? Below are data from eight Arizona high schools.

	Student smokes	Student does not smoke	Total counts	Percent of students who smoke
Both parents smoke	400	1380		
One parent smokes	416	1823		
Neither parent smokes	188	1168		
Total counts				
Percent of parents' smoking habits				

a) How many students do the data describe? [**5375]**

b) What percent of these students smoke? **[18.7%]**

c) Give the marginal distribution of parents' smoking behavior, both in counts and percents. **[1780 – 33.1%; 2239 – 41.7%; 1356 – 25.2%]**

d) What percent of students smoke among those with two smoking parents, among those with one smoking parent, and among those with neither parent smoking? **[22.5%, 18.6%, 13.9%]**

e) Draw a bar graph that compares the three percents you found in part (c).

f) Briefly describe the relationship between parents' smoking and students' smoking.

2. Now investigate the relationship between location of the reef (fringe or patch) and one or more of the coverages, e.g. hard coral, rock and sand. How can you determine if reef location is associated with type of coverage?