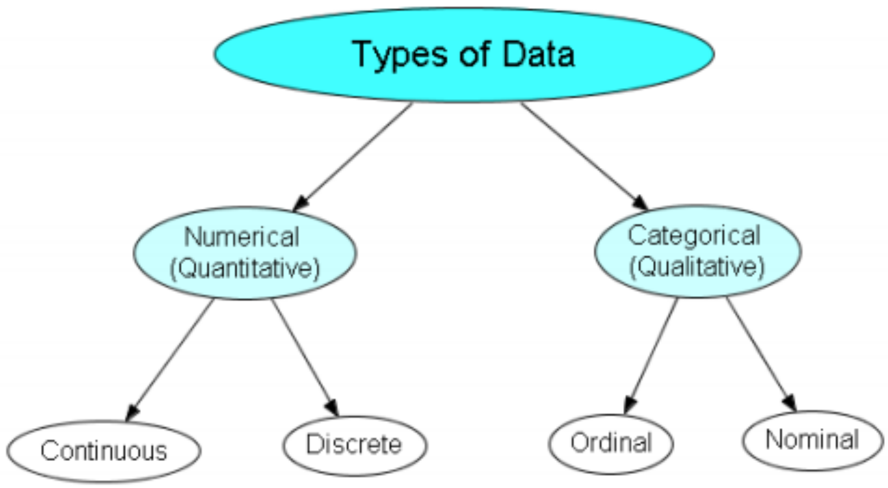
Section 1.2 – Organizing and Displaying Categorical Data

MDM4U
Jensen

Part 1: Types of Variables



Numeric/Quantitative Variable: A variable that takes numerical values for which it makes sense to find an average. These variables can be either continuous or discrete

Continuous Variable: A numeric variable that can have an infinite number of values in a given interval. Measurable with all real numbers.

Examples: temperature, height, weight, speed

Discrete Variable: A numeric variable that can take on only a finite number of values within a given range. Usually measured with integer values only.

Examples: number of dogs, number of goals scored, number of siblings

Categorical/Qualitative Variable: A variable that places an individual into one of several groups or categories. Categorical variables may have categories that are naturally ordered (ordinal variables) or have no natural order (nominal variables).

Ordinal Variable: A categorical variable that has a natural ordering of its possible values, but the distances between the values are undefined.

Example: When asking people to choose between Excellent, Good, Fair and Poor to rate something, the answer is only a category but there is a natural ordering in those categories.

Nominal Variable: Type of categorical variable that describes a name, label, or category with no natural order.

Example: there is no natural order in listing different types of school subjects: "History" does not have to follow "Biology." These subjects can be placed in any order.

Part 2: Frequency Tables

To make an accurate picture of data, the first thing we have to do is make 'piles'. For categorical data, 'piling' is easy. We just count the number of cases corresponding to each category. We can organize these counts into a frequency table, which records the totals and category names.

Frequency tables are used to organize data.

Example 1:

Grade 12's were asked when their spares were and these were the results:

A, B, C, D, A, D, D, B, A, C, A, C, B, B, B, A, D, C, A, A, B, D, C, A, B
B, A, C, C, D, A, B, A, B, B, B, D, D, A, D, D, C, A, D, C, D, A, B, B, A

The problem with data that is presented like this is that you can't 'see' what is going on. Organize the data in to a frequency table to better see the distribution of data.

Spare	Frequency
A	15
B	14
C	9
D	12

Counting the frequency is useful, but sometimes we want to know the proportion of data in each category, so we make a relative-frequency table.

A relative-frequency table shows the frequency of a data group as a fraction or percent of the whole data set.

Spare	Frequency	Relative Frequency
A	15	30%
B	14	28%
C	9	18%
D	12	24%

Part 3: Bar Graphs

Graphs are used to display data. Bar graphs, segmented bar graphs, pie charts, and pictographs are appropriate types of graphs for displaying the data of categorical variables. Bar graphs can also be used for discrete numeric variables.

A bar graph displays the distribution of a categorical variable, showing the counts (frequency) for each category next to each other for easy comparison.

A bar graph consists of parallel bars of equal widths (**with a space between each bar**) with lengths proportional to the frequency of the variables they represent.

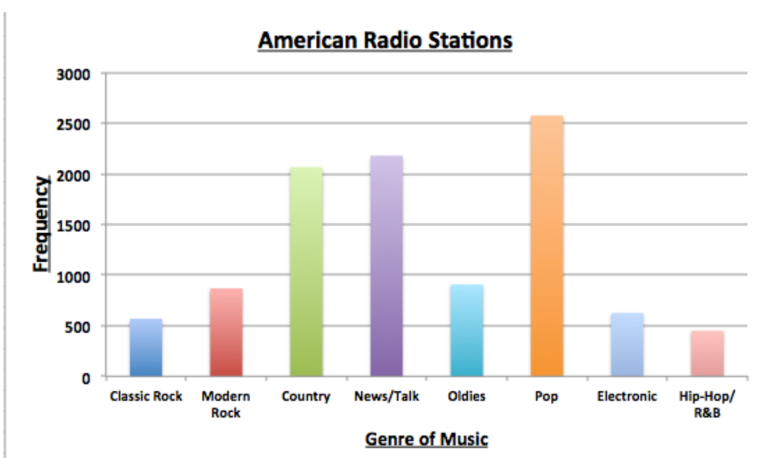
Example 2:

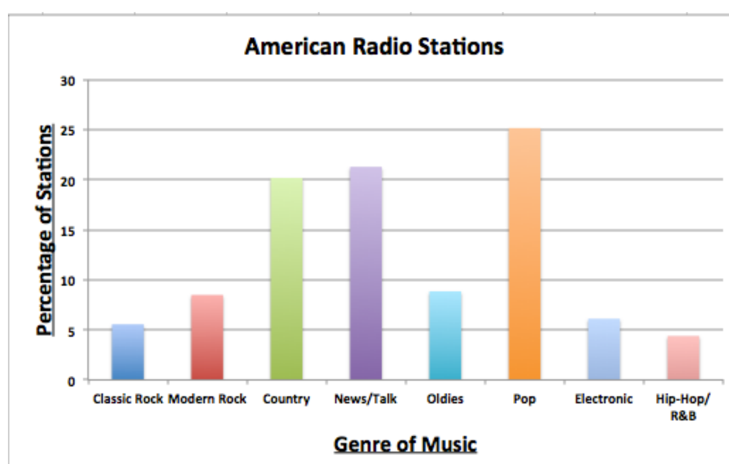
The following frequency table shows the number of different U.S radio stations broken up by category based on the kind of music they broadcast.

I. Complete the relative frequency column

Genre	Frequency	Relative Frequency
Classic Rock	569	5.56
Modern Rock	869	8.49
Country	2066	20.18
News/Talk	2179	21.28
Oldies	906	8.85
Pop	2575	25.15
Electronic	626	6.11
Hip-Hop/R&B	450	4.39
Total	10240	~100

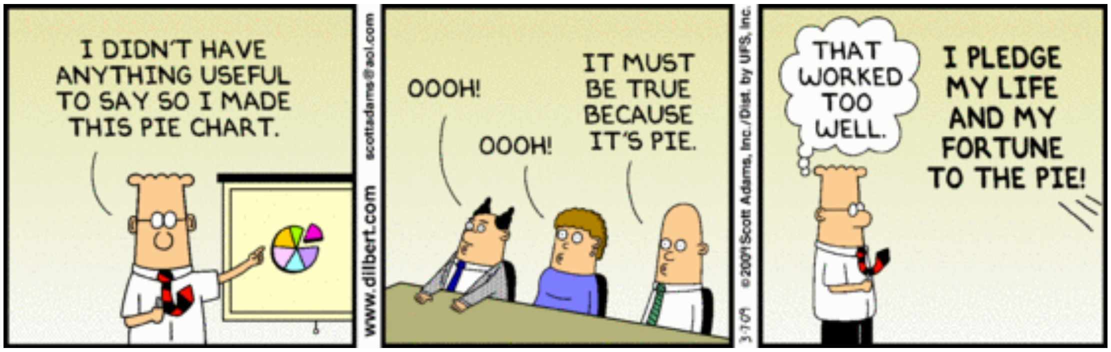
II. Use the table to create two bar graphs. The first showing frequencies and the second showing relative frequencies of each category.





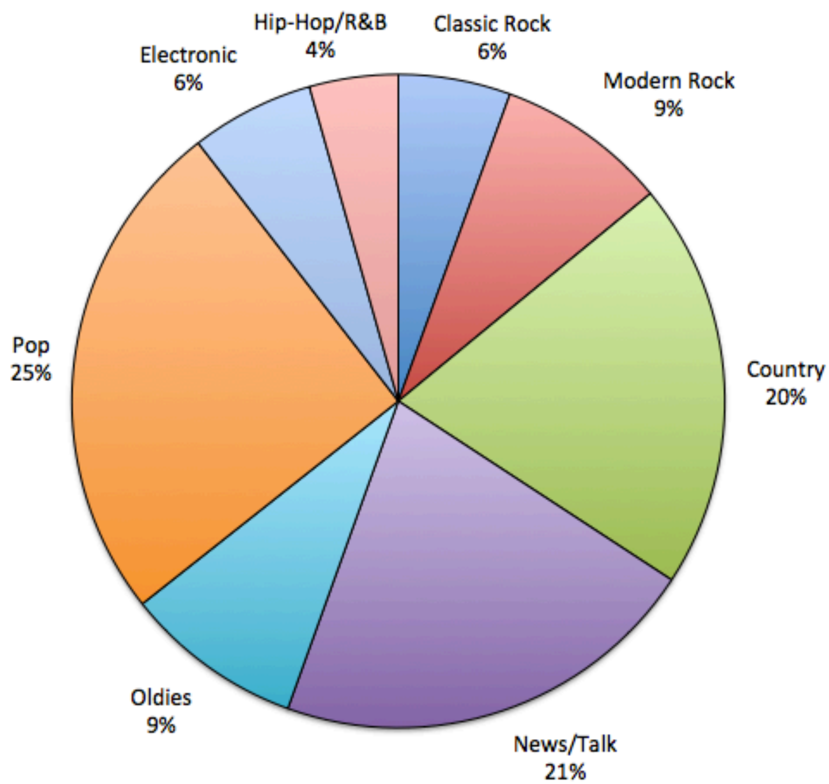
What do you notice about the shapes of the distributions? *They are the same*

Part 4: Pie Charts



A pie chart shows the distribution of a categorical variable as a 'pie'. They slice the circle into pieces whose sizes are proportional to the fraction of the whole in each category. A pie chart is best used when trying to show a category's relation to the whole. Pie charts are awkward to make by hand, but technology will do the job for you.

Here is a pie chart showing the data for the U.S radio stations from the previous example:

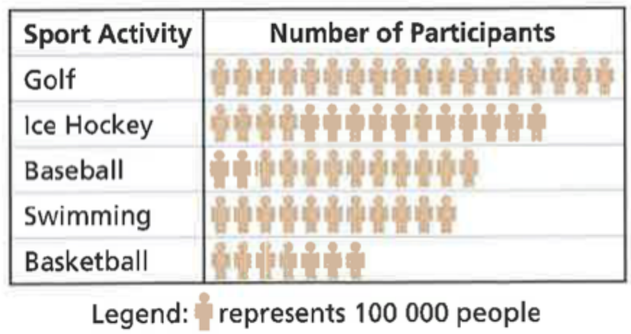


Part 5: Pictographs

A pictograph is a symbolic representation of data. The following pictograph displays the number of participants, aged 15 and older, in the five most popular sports activities in Canada.

How many people aged 15 and older play hockey?

1 500 000



Example 3:

a) How many red delicious apples are in the store?

30

b) How would you represent 11 apples?

Based on the scale it would be very difficult to accurately represent 11 apples

Problems with Pictographs:

- Pictographs can make a graph more interesting but…
  - Legends are often missing or confusing
  - Graphics are sometimes distorted or confusing
  - Relative frequencies are sometimes hard to determine (how would you represent 11 apples?)

Part 6: Contingency Tables and Segmented Bar Graphs

We have learned some techniques for analyzing the distribution of a single categorical variable. If a data set involves two categorical variables, we use a two-way table (contingency table). A two-way table of counts organizes data about two categorical variables measured from the same set of individuals.

Example 4: Only 32% of those aboard the Titanic survived. Was that survival rate the same for all ticket classes? To answer that question, we can arrange the counts for the two categorical variables, survival and ticket class, in a two-way table.

Survival	Class: First	Second	Third	Crew	Total
Alive	203	118	178	212	**711**
Dead	122	167	528	673	**1490**
Total	**325**	**285**	**706**	**885**	**2201**

In this case, survival is our row variable and class is our column variable. The margins of the table give totals. When analyzing a contingency table, the goal is to see if the variables depend on each other. This can be done by looking at the two possible conditional distributions (row and column).

If we think that class might depend survival, then we should look at the distribution of the row percentages. This is the conditional distribution for class based on survival.

Survival		Class: First	Second	Third	Crew	Total
Alive	Count	203	118	178	212	**711**
	%	28.6%	16.6%	25.0%	29.8%	**100%**
Dead	Count	122	167	528	673	**1490**
	%	8.2%	11.2%	35.4%	45.2%	**100%**

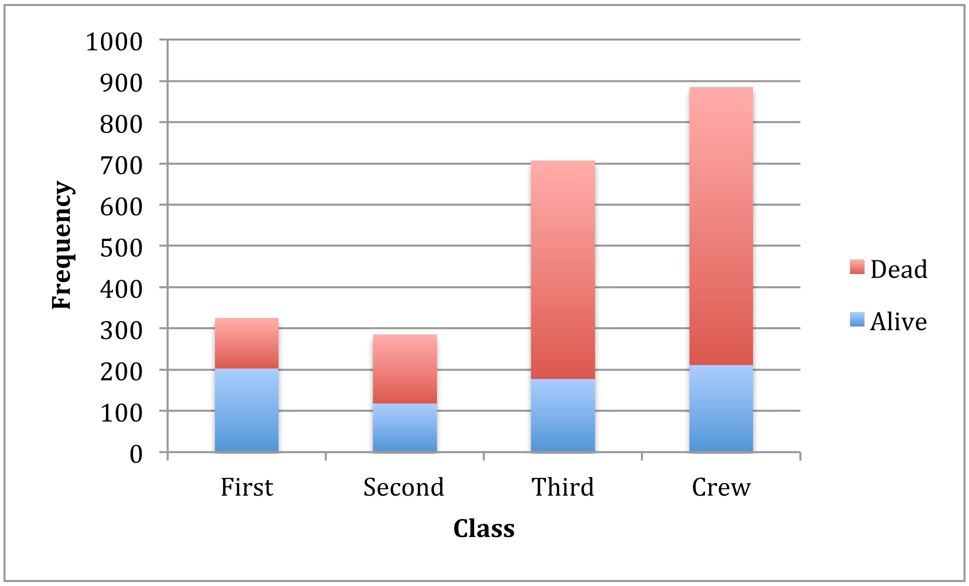
However, in this scenario it would make more sense to determine if survival depends on class. To do this, we should look at the column percentages. This is the conditional distribution for survival based on class.

Survival		Class: First	Second	Third	Crew	Total
Alive	Count	203	118	178	212	**711**
	% of Column	62.5%	41.4%	25.2%	24.0%	**32.3%**
Dead	Count	122	167	528	673	**1490**
	% of Column	37.5%	58.6%	74.8%	76.0%	**67.7%**
Total	Count	**325**	**285**	**706**	**885**	**2201**
		100%	**100%**	**100%**	**100%**	**100%**

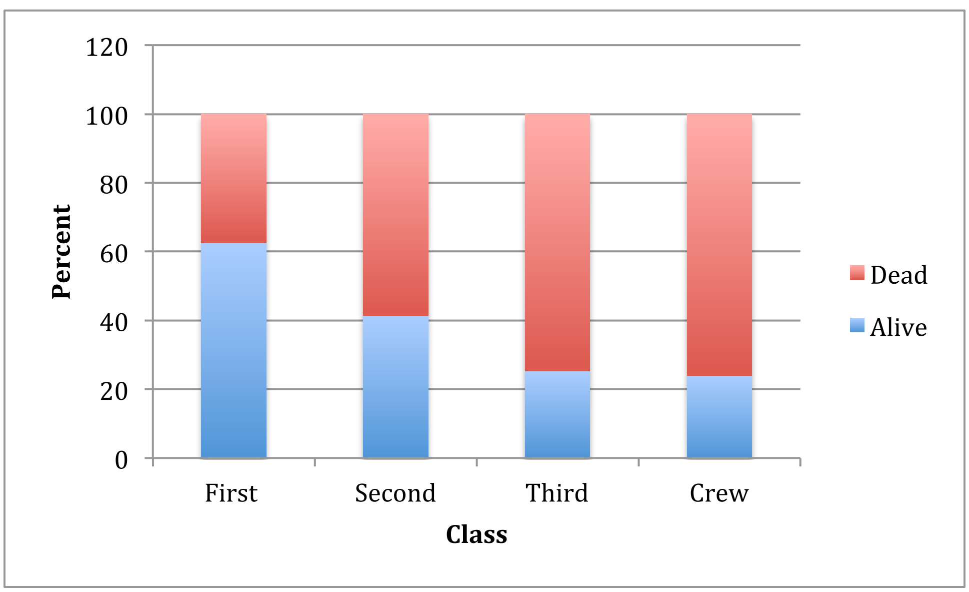
Looking at how the percentages change across the row, it sure seems that class influenced whether a persons survived or not. 62.5% of first class passengers survived while only 25.2% of third class passengers survived.

Two-way tables are often displayed using segmented bar graphs.

Example: Segmented bar graph of survival based on class using frequencies



Example: Segmented bar graph of survival based on class using conditional percentages



Note: *The bars of each graph have the same proportions but it is easier to see in the second graph that first class passengers had the highest proportion of survivors.*