Vocabulary: *Data Distributions*

<table>
<tr><th>Concept</th><th>Examples</th></tr>
<tr><td>

Two Types of Data.

I. **Categorical data:** is data that has been collected and recorded about some non-numerical attribute. For example:

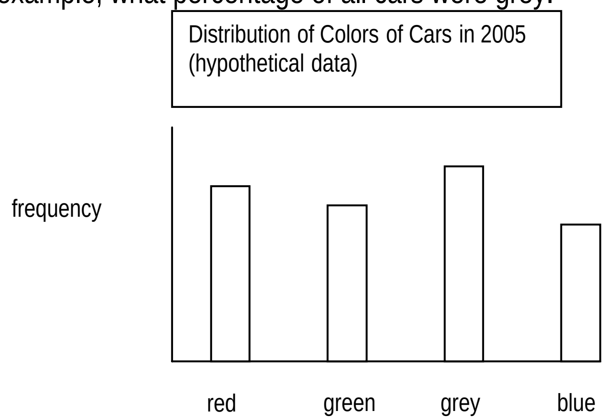
- *color* is an attribute or variable (for which the categories or values could be red, green, blue etc.);
- *gender* is an attribute (for which the categories are male and female);
- *country* is an attribute (for which the categories might be U. S., Germany, etc.);
- *yes/no* are two categories of response.

A **bar graph** that shows categorical data will have category names (red, green, blue etc) along the horizontal or independent axis. These names can be rearranged and sense can still be made of the graph.

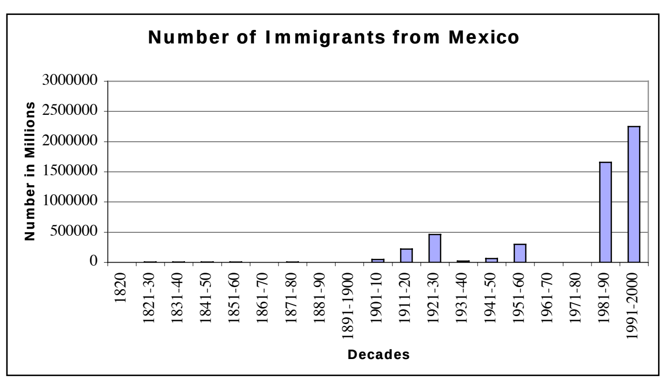
Summarizing Categorical Data: Because the names can be arranged in a different order we do not talk about the "maximum color," for example, or the "mean color." A bar graph may show ***how frequently*** a particular category occurs. For example, suppose that "red" was counted 22 times, and "green" was counted 100 times and that "green" was counted more often than any other category.

</td><td>

Example 1:
Below is a **bar graph of categorical data**. The variable of interest is "Car color in 2005." The **categories** are red, green, grey, blue. The height of the bars indicates the **frequency** with which each color occurs. There is no scale on the vertical axis below, but *there should be*, in order to make sense of the graph. If there were a scale we could say *exactly* how many of each color was observed. Or, if the vertical axis were marked off to show percentages of some total, we would be able to say, for example, what percentage of all cars were grey.



Below are three bar graphs that use the same data about "Immigration from Mexico to U.S." Examples 2 and 3 are actually **graphs over time**, intended to show **trends**.
Example 2: The graph shows numbers of Mexican immigrants in each decade, and we can see that the numbers are increasing quite dramatically.



Example 3: The graph shows the same data, but this time as a **percentage** of total immigration. The graphs are not

</td></tr>
</table>

Therefore we can use the **mode** to summarize the data. Or the bar graph may show information about ***some value** of the variable*, such as population in U.S., population in Germany etc., in which case we can say which country has the largest population, or note the gap between largest and smallest countries.

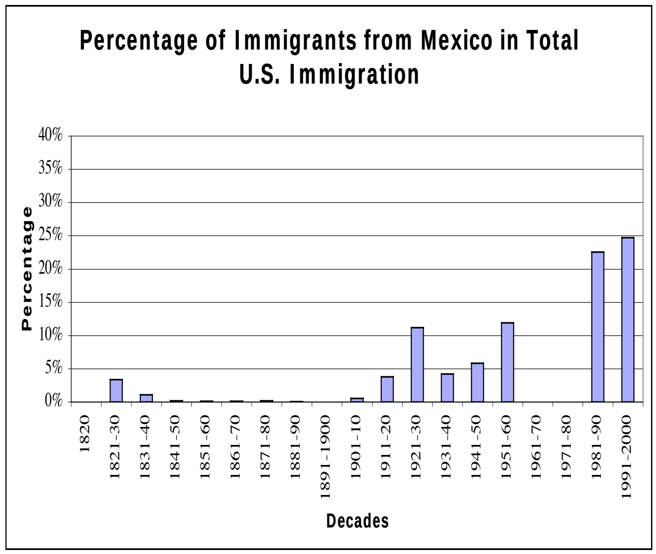
II. Numerical data: is data about some attribute that *must* be organized by numerical order to show how the data varies. For example:

- Number of pets
- Measure of a piece of lumber

A **bar graph** that shows numerical data has the values of the data in a fixed order on the horizontal axis.

Summarizing Numerical Data: The **maximum, minimum** and **center** values of the attribute now make sense. For numerical data, the **dot plot** or **bar graph** always shows ***how frequently*** particular numerical values of the variable of interest occur, for example, how many times a value of "3 pets" was recorded, or how many times a value of 8.1 feet was recorded. We can now compute with the values of the variable, say "number of pets," and say what the **mean** or **median** number of pets might be, as well as the **mode**. (See *Data About Us* for a Dot Plot, also called a Line Plot.)

as a **percentage** of total immigration. The graphs are not exactly the same shape because immigration from other countries is another variable making up the total immigration picture. Thus, the bar for the decade 1821-30 shows about 4% of all immigration was from Mexico. This bar is almost invisible on the first graph because the total immigration numbers were much smaller, so 4% of a smaller number gives a result too small to show up on the scale used. From the second graph we can say that there is an increasing trend in immigration from Mexico, both in actual numbers, and as a percentage of the total. However, the greatest percentage that comprises Mexican immigrants is less than 25% of all immigration.

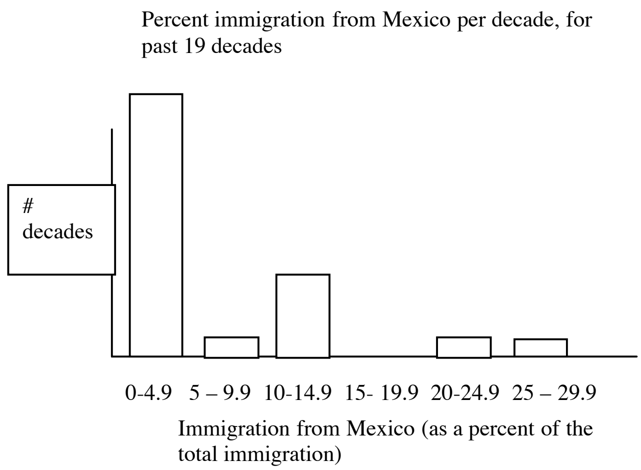


Example 4: The graph is a truly **numerical data** graph. The data is **continuous**; we might have ANY value between 0 and 100 for the immigration percentage. The Mexican immigration percentages have been organized along the horizontal axis to show how this variable, immigration percent, has varied, but not over time in this graph. Each piece of data is a numerical value, the immigration percent for a decade. (There are 19 decades represented in all. The actual percentages are: 0, 0, 0, 0, 0, 0, 0, 1, 1, 3, 4, 4, 6, 11, 12, 14, 14, 23, 25.) There is no scale on the vertical axis, but *there should be*. Without a scale we can not say exactly how often a particular percentage of immigration from Mexico occurred. What we can see from the shape of the graph is which percentages occurred more often and which less often.

Discrete or Continuous?
Discrete Data: is countable, like number of pets in a household.
Continuous Data: is data that does not have listable values, like the measurement of a piece of lumber. The distinction is that in the first case, where the attribute is "number of pets" we can have 1, 2 or 3 pets, but there is no value of the variable "number of pets" between these whole numbers. Meanwhile, if the attribute is the "measurement of a piece of lumber" we can record any measurements between whole numbers.

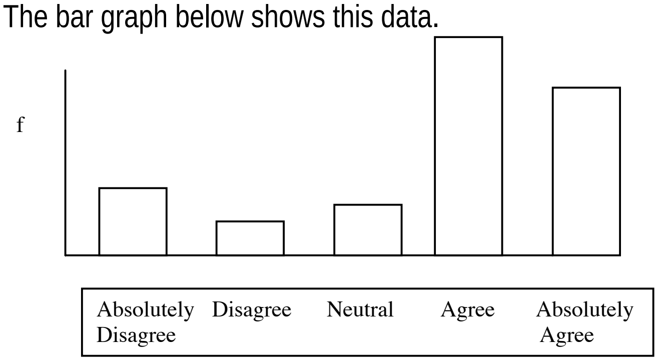
Note: data that must be collected as whole numbers can still have decimal numbers as the mean or median.

III. Graphs over time: are categorical, though they have some of the characteristics of numerical data. For example, the attribute or variable might be "year" and the categories might then be 1900, 1920, 1940 etc., which are indeed numbers. Unlike true categorical data the order does matter, since it tells part of the story. However, it is still not sensible to talk about the mean or maximum value of the variable "year." For example, if we record immigration between 1820 and 2000 then we would not say that "2000" is the maximum value since the distribution arbitrarily ends at 2000. If we collected more information in 2001 then the distribution would be extended, but we would not say that



From this graph we can see **maximum** (25 – 29.9%) and **minimum** (0 – 4.9%) by looking at the horizontal axis. We can use the **mode** (0 – 4.9%) or the **median** (also in the 0-4.9% bar) to talk about what is a typical immigration percentage. (See Measures of Center, below.) The 25% immigration figure in one decade, 1990 – 2000 looks like an **outlier** when viewed in the context of all other immigration from Mexico. This is the same story as examples 2 and 3, but from a different perspective.

Example 5:
Here is some data collected about student opinions on the statement, "Data and Statistics is an interesting part of mathematics." This is **categorical** data.
Absolutely Disagree….4
Disagree……………….2
Neutral………………….3
Agree……………………12
Absolutely agree……….9
The bar graph below shows this data.

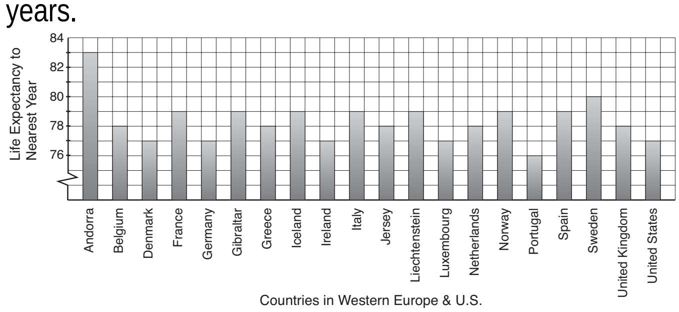


When you compare this bar graph to the example 4 you

the data has become more variable solely because of this additional data. Graphs over time are good for indicating *trends*, rather than investigating variability or what is typical.	can see how the maximum and minimum values are not read from the horizontal axis. You can use the **mode** to say what a typical answer is. You can not calculate a mean opinion.
Types of graphs*:*: • **Dot plot (or line plot*)***: Each single piece of data is represented as a dot (or an "x") positioned over a labeled number line. (See *Data About Us* for examples.) • **Value bar graph**: Each category is represented by a separate bar whose relative length corresponds to the magnitude or value of that category. • **Frequency bar graph:** A bar's height is not the value of an individual category but rather the number (frequency) of pieces of data that all fit that category, or all have the same numerical value.	*Example 1* above is a **frequency bar graph** of categorical data. The frequency of the occurrence of each color is recorded. If the vertical axis were marked as percentages of all cars then this would be a **relative frequency graph.** *Example 2* above is a **value bar graph** over time. The value or amount of the immigration from Mexico is recorded for each decade. *Example 3* above is also a **value bar graph** over time. In this case changing the data to percentages did not signal a change to recording a frequency with which any category on the x-axis occurred. The graph does not say that a particular decade occurred with more frequency than another. Nor does it say (in this format) that a particular percentage of immigration occurred with more frequency than another. *Example 4* above is a **frequency bar graph** of numerical data. The frequency with which different percentages of immigration occurred is recorded. Low percentages occurred often. *Example 5* above is a **frequency bar graph** of categorical data. The frequency of different responses is recorded.
Distributions: of data are just tables or graphs showing how the collected data varies. We are interested in the picture of the situation conveyed by having lots of data to examine, and	*Example 4* above can be summarized by saying that **distribution** of percentage immigration from Mexico is very variable; the **minimum** value is 0% and the **maximum** percentage is 25%. (The **range** is 0 – 25%.)

<table>
<tr>
<td>having lots of data to examine, and we summarize the distributions of numerical data by finding measures of central tendency (i.e., mean, median, mode) or spread or variability (e.g., outliers, range) or shape (e.g., clumps, gaps). Most of these summarizing descriptions are easier to see on a graph that in a table, though the details are lost in moving from a table to a graph.</td>
<td>However, the range or variability of the data is most influenced by the two percentages, 23% and 25%, which occurred only in the most recent decades. These two pieces of data are so unlike the rest of the data that we might call them outliers. (Technical definition of an outlier is given in Samples and Populations.) The most typical percentage immigration from Mexico is in the 0 – 4.9% interval, which is the largest cluster of data. If we use the graph alone we can say that this interval occurs most often, so is the mode. We can also say that the median (10^{th} decade from top or bottom of the ordered list of percentages by decade) will be in this interval also. (The mean is hard to estimate from the graph, because calculating a mean involves adding in ALL the percentages and dividing, or looking for a balance point. The unusually high percentages of 23 and 25% will have an influence on this calculation, so the mean may be in an interval higher than 0 – 4.9%.)

See Example 7 below (Life Expectancy data). This can be summarized by noting that the range is small, only 8 years between the minimum and maximum values, that Andorra’s life expectancy of 83 years is an outlier which skews the graph to the right. Most of the data clusters between 76 and 79 years. The most typical life expectancy is 78-78.9, using the mode, or 77 – 77.9, using the median. The mean would not be such a good representative of what is typical since it is affected by the unusual value of 83.</td>
</tr>
<tr>
<td>Variability: This property of numerical data distributions can be seen visually on a graph. There are several ways to see, describe or measure variability, but in every case what we are looking for is how the recorded values of the attribute vary. For example, measurements of an object vary, not because of mistakes or inaccuracies in the process but because measurement results are inherently variable. A saw might cut a piece of lumber intended to be 8 feet long, but someone measuring it will say it is in fact 8 feet and _ inch, or 8 feet and $\frac{5}{16}$ of an inch etc. Another example is how samples</td>
<td>Example 6. Below is a value bar graph of categorical data. The categories are countries The values are life expectancies. Measures of variability make no sense for this kind of graph. Notice that the scale on the y-axis MIGHT be misleading. It makes it look like Andorrans live twice as long as Belgians, when in fact the difference is 5 years.



Example 7. Below is the same data as in example 6,</td>
</tr>
</table>

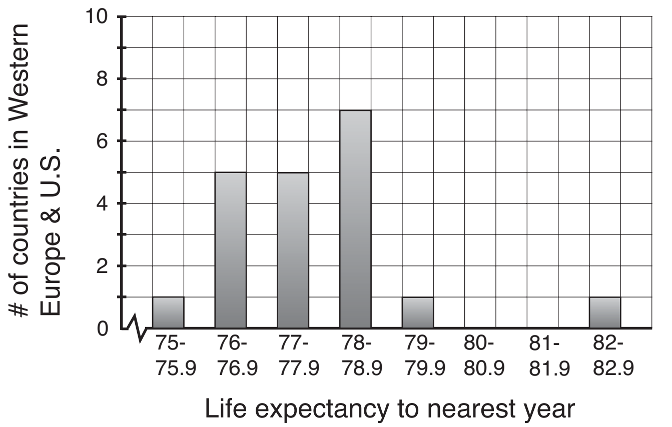
vary: one sample might indicate that 48% will vote for candidate A, while a second sample from the same population might indicate that 51% will vote for candidate A. **Variability is an inherent property of certain kinds of numerical data**, not a sign of inexactness in the process of data collection.

Measures or descriptors of variability: In this unit students look at

- the **range** to determine if the gap from minimum to maximum is wide,
- where data **cluster** to determine if most data lie within a narrow range of values
- **gaps** in a distribution to determine if some values occur much less frequently than others
- **outliers** to determine if these are true representatives of wide variation in the data or perhaps errors

Note: the only *measure* of variability in the above list is *range*. In future units students learn to measure variability by using the *interquartile range*.

organized as a **frequency bar graph of numerical data.** The variable or attribute is life expectancy to nearest year, and the vertical axis indicates how frequently this life expectancy was observed.

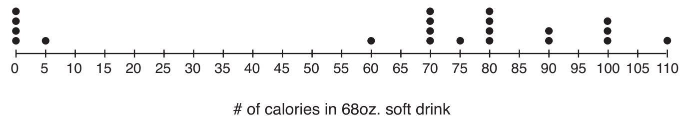


Now it makes sense to say that the **range** for this variable, life expectancy, is quite small, only 8 years or 83 – 75 years, and even smaller if the **outlier**, 83 years (Andorra) is not included in the calculation. There is a **gap** in the graph from 80 – 82 years, making Andorra's 83 years look even more unusual. Most of the data clusters between 76 and 79 years.

The shape of a distribution tells us

- whether the data is *widely spread* with only a relatively few pieces of data at

Example 8:
Below is a **numerical data frequency graph**, in which the variable is # calories in a 6 ounce soft drink.

many different values of the variable of interest, or *clustered* together at a few values of the variable. • Whether there are *gaps* or *outliers* • Whether the distribution is *symmetric* or not • Where the *center* or most typical value of the variable is, and whether this is affected by the shape.	 The **shape** of the distribution tells us that the # calories **varies widely**, with no particular # calories being really typical. There is a **large gap** in the distribution. For some reason there were no soft drinks with calories between 5 and 59.9. If we tried to use **mode** to summarize what is typical we would have to give 3 different answers. If we used **median** we would use 80. If we used mean we would estimate approximately 60 calories, using the idea that the **mean** is the balance point of the distribution. None of these measures of center gives a satisfactory way to summarize this distribution because of the shape. In *Examples 4 and 7* above the distributions were **skewed** right.
Three measures of central tendency (center): • **Mode** is the data value or category occurring with the highest frequency. It sometimes has more than one value and is unstable because a change in one or a few data values can lead to a very large change in the mode value. A distribution may be unimodal, bimodal, or multimodal. As a measure of "typical" mode is applicable to both categorical and numerical data, but not to graphs over time. • **Median** is the numerical value that marks the middle of a distribution. It is not influenced by extreme data values so is a good measure to use when working with distributions that are skewed. Graphically, the median marks the location that	In *Example 1* above, the **mode** car color was grey. In some sense this is "most typical" but the other numbers were so close that it would be misleading to think that "grey" was "typical" but "red" was not. In *Example 5* above, "Agree" was the **mode** response. In *Example 4* above, 0 – 5% was the most frequently occurring, or **mode**, immigration figure for immigrants from Mexico. *In Examples 2 and 3, which were graphs over time, it makes no sense to talk of a particular decade occurring with more frequency than another. So mode is not applicable.* In *Example 4* above there were 19 decades represented, each with a particular immigration value associated. If we place these 19 immigration numbers in order from least to greatest we actually have a list of percentages: 0, 0, 0, 0, 0, 0, 0, 1, 1, 3, 4, 4, 6, 11, 12, 14, 14, 23, 25. The middle of this list is the 10^{th} piece of data from either end, or 3%. So the **median** immigration figure is 3%. We can say that over the past 200 years the typical immigration from Mexico is 3% of the total. (Notice that the

divides a distribution into two equal parts. As a measure of "typical," *median* is applicable only to numerical data. • **Mean** is the numerical value that marks the balance point of a distribution; it is influenced by all values of the distribution, including extremes and outliers. It is a good measure to use when working with distributions that are roughly symmetric. As a measure of "typical," *mean* is applicable only to numerical data. **Note**: Students must learn to choose the measure of center that is appropriate to the situation and distribution. The goal is to choose the measure that gives the most useful information about what is "typical."	actual percentages are grouped on the graph for *Example 4*, so this list is not directly accessible from the graph. But even without the specific details we can see that over half the data is in the first bar on the graph, 0 – 4.9%.) *In Examples 1 and 5 the data is categorical so we can not order these to find a middle or median piece of data. We can't give a median color or a median opinion. In the graphs over time in Examples 2 and 3 we can not use a middle year as a typical year for immigration.* Since ***mean*** is applicable only to numerical data we can apply it to *Example 4* above. If we work from the list of percentages (0, 0, 0, 0, 0, 0, 0, 1, 1, 3, 4, 4, 6, 11, 12, 14, 14, 23, 25) we have three ways to think of the mean: *Strategy 1*: we can think of "evening" these numbers out, trying to end up with a list of 19 identical values, yet the same overall value as the original list. There are different ways to effect this "evening" out. One way is to take the numbers in pairs. For example, the lowest and highest numbers are *0* and *25*. If we replace one or these with a *12* and the other with a *13* the overall total value is not affected. We have *shared* out the total value of these two data points. The list would become: ***12***, 0, 0, 0, 0, 0, 0, 1, 1, 3, 4, 4, 6, 11, 12, 14, 14, 23, ***13***. We can do the same with *0* and *23*, making them into *11* and *12*.. The list would then be: ***12, 11***, 0, 0, 0, 0, 0, 1, 1, 3, 4, 4, 6, 11, 12, 14, 14, ***13, 13.*** Now we could "even out" a *0* and a *14*, to make two 7's. ***12, 11, 7***, 0, 0, 0, 0, 1, 1, 3, 4, 4, 6, 11, 14***, 7, 13, 13.*** If we continue like this we will eventually have a list with 6's and 7's. The mean is between 6 and 7. *Strategy 2*: Find the total value by adding all the values, and then "share out" this total value into 19 equal pieces by dividing. (0 + 0 + 0 + 0 + 0 + 0 + 0 +1 + 1 + 3 + 4 + 4 + 6 + 11 + 12 + 14 + 14 + 23 + 25)÷19 = 6.2.. Therefore, the **mean** Mexican immigration over these 19 decades is 6.2% of the total immigration. Notice that this measure of center is influenced by the two unusually high values, and so is not as good a measure of "typical" as the median. Strategy 3: Looking at the graphical representation of the

	data (See Ex 4) we can try to visualize where a balance point would be. This distribution is skewed right, so the high values pull the mean up. But there is a very large cluster of values between 0 and 5. Trying to visualize a fulcrum point that will allow this distribution to balance is challenging.