2.1

- variable : any characteristic that is recorded for subjects in a study
- observation : observed data values for a variable
- categorical variable : each observation belongs to one of a set of categories
- quantitative variable : observations take on numerical values
  1. discrete : possible values take form a set of separate numbers
  2. continuous : if possible values form an interval
- examples :
  1. Weather station : county where station is located, daily observation of whether it rained or not, daily high temperature, amount of precipitation.
  2. Demographics : gender, religious affiliation, place of residence, age, number of siblings, annual income.
- frequency tables : listing of possible values for a variable, together with the number of observations for each value.
  1. proportion : frequency count of observations in that category divided by the total number of observations.
  2. percentage : is the proportion multiplied by 100
  3. proportions and percentages are also called relative frequencies
- mode : category with the highest frequency

2.2

- pie chart : a way to summarize data graphically. when two slices are about the same size, we have difficulty determining which is actually larger. this makes the bar graph more precise.
- bar graph : use vertical (or horizontal) bars, height of the bar is the frequency (or percentage) of the different categories.
- histogram : a graph that uses bars to portray the frequencies or the relative frequencies of the possible outcomes for a quantitative variable.

  how to : the histogram is a bar graph for quantitative variables. To construct a histogram do the following
  1. Divide the range of the data into intervals of equal width. (For a discrete variable with few values, use the actual possible values.)
  2. Count the number of observations that fall in each interval, forming a frequency table.

3. on the horizontal axis, label the values or the endpoints of the intervals. Draw a bar over each value or interval with height equal to its frequency (values are marked on the vertical axis).

- shape of distribution (what information is there?)

  1. overall pattern : do the data cluster together, or do one or more observations noticeably deviate from the rest?
  2. mode
     (a) unimodal : one single mound (one distinct mode within the histogram)
     (b) bimodal : two distinct mounds
  3. skewed vs symmetric : a distribution is skewed if one side of the distribution stretches out longer than the other. it is called symmetric if this is not the case.
     (a) skewed to the left : if the left tail is longer than the right tail
     (b) skewed to the right : if the right tail is longer than the left tail
  4. tails of distribution : the parts of the curve for the lowest and highest values.

- time series : data collected over time

- time plot : chart each observation on the y-axis against the time it was collected on the x-axis.

- trend : a common pattern (rising or falling) over time.

- what is the difference in the information given between the histograms and the time plots?

  1. cannot see changes over time in the histogram.
  2. it is difficult to see how many years had a given average temperature in a time plot. we can also see the distribution of the data in the histogram, but not in the time plot

2.3

- graphical summaries give a good idea of the shape of the distribution.

- numerical summaries (statistics) give a good indication of central tendenciy and spread

- mean (average) : sum of observations divided by the number of observations. if there are n observations, the mean is denoted

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- median : the middle observation. half the observations are smaller and half are larger. to find the median

  1. order all n observations
  2. when n is odd, the median is the middle observation
  3. when n is even, take the average of the two middle observations

- example : 37, 55, 62, 18, 51, 91, 35, 60, 58, 20, 81.

  1. find the mean (48.7 w/o 81) (51.63 w/ 81)
  2. find the median (53 w/o 81) (55 w/ 81)

- properties of the mean

  - the mean is the balance point of the data. if we were to put weights on a line representing where the observations occur, then the line would balanceby placing a fulcrum at the mean.
  - for a skewed distribution, the mean is pulled in the direction of the skew.
  - the mean can be heavily influenced by an outlier (an observation that falls well above or well below the bulk of the data)

- the shape of a distribution influences whether the mean is larger or smaller than the median. the mean lies toward the direction of the skew relative to the median.

  - perfectly symmetric : the mean equals the median
  - skewed to the right : mean is larger than the median
  - skewed to the left : mean is smaller than the median

- example : colleges data

	EN	SB
n	2533	9540
mean ($\bar{x}$)	45.521	13.663
median	43.490	6.340
std dev (s)	22.848	16.784

- (why is it important to understand the skew of a distribution?)

- go over 2.38 homework

- resistant : the median is said to be resistent to outliers, the mean can be heavily influenced by outliers. how is this evident in the colleges example?

- the median can be too resistant. use the example on pg 53.