22S:30 and 22S:105
Statistical Methods and Computing

Graphical Depiction of Qualitative and Quantitative Data and Measures of Central Tendency

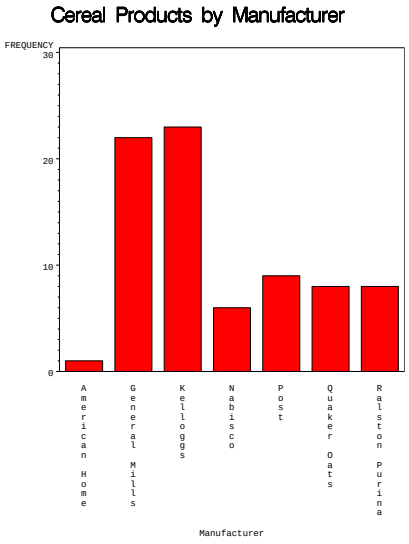
Lecture 2
January 20, 2017

Kate Cowles
374 SH, 335-0727
kate-cowles@uiowa.edu

Bar charts for nominal and ordinal data

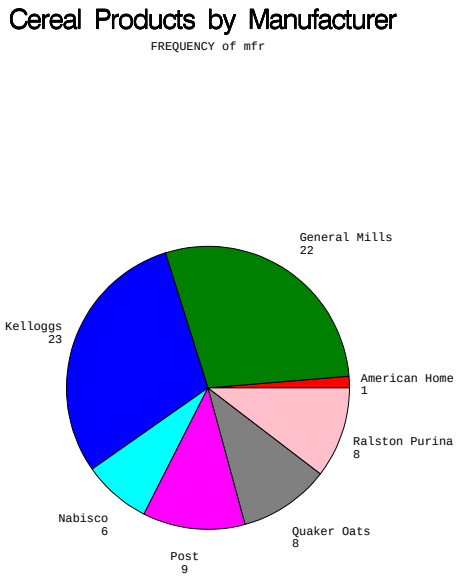
- present a frequency distribution in visual form
- categories that are possible values of the variable are listed on horizontal axis
- bar heights represent either frequency or relative frequency of observations in that class

Continuing example of cereal data



Pie charts

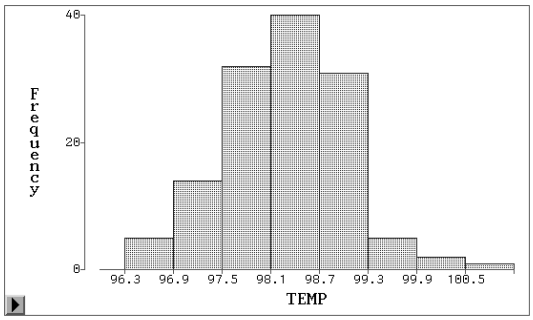
- a "slice" for each possible value of the variable
- area of slice represents the proportion of the whole that the category makes up
- all categories must be included



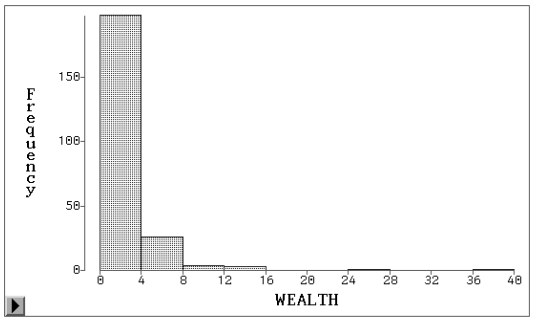
Histograms for quantitative data

- presents a frequency distribution of discrete or continuous data in visual form
- range of possible values must be divided into intervals
  - easiest to work with if intervals are of equal width
  - limits of intervals are shown on horizontal axis
- vertical bar centered at midpoint of each interval
  - area of each bar represents frequency associated with corresponding interval

Example 1: Histogram of body temperatures of 130 people



Example 2: Wealth in billions of dollars of the 209 billionaires in the world in 1992



Symmetric and skewed distributions

- symmetric – right and left sides of histogram are roughly mirror images
- skewed to the right – long "tail" on right side; some extremely large values
- skewed to the left – some extremely small values

Outliers

- individual values that deviate from the general pattern of the data

Stemplots for quantitative variables

- show overall shape of distribtution
- give more detailed information than histograms
- feasible only for fairly small datasets

```
Stem Leaf
 100 8
 100 0
  99 59
  99 000001112223344                                          1
  98 55566666666667777777888888888899                        3
  98 00000000001112222222223333444444444                     3
  97 556667778888888999999                                   2
  97 0111222344444                                           1
  96 7789
  96 34
     ----+----+----+----+----+----+----+---
```

Example

- Investigators suspected that Benzo(a)pyrene, or BaP, from a pipe foundry in Phillipsburg, NJ, might be contaminating household air.
- This dataset presents data from 14 different days on samples of indoor air from a house near the foundry and samples of outdoor air collected at the same times.
- The measures are concentrations of BaP-containing particles no larger than 10 micrograms.
- The two variables are:
  - indoor air BaP
  - outdoor air BaP

Reference: Lioy, PL, Walman, JM, Greenberg, A, Harkov, R and Pietarninen, C (1988). The total human environmental exposure study (THEES) to Benzo(a)pyrene: Comparison of the inhalation and food pathways. Archives of Environmental Health, 43: 304-312.


```
Variable=OUTDOOR

   Stem Leaf                     #
      7 8                        1
      6 57                       2
      5 06                       2
      4 01                       2
      3 58                       2
      2 04557                    5
        ----+----+----+----+
 Multiply Stem.Leaf by 10**+1
```


Line plots or time plots

- Usually time is plotted on the x-axis.
- Some other variable that changes over time is plotted on y-axis. Points are connected by lines.

Example: High-water mark for Amazon River at Isquitos, Peru, for years 1962-1978

Measures of central tendency for quantitative data

- Before we can use data to draw conclusions, we must summarize the data to get the "overall picture"
  - Number of values may be so large that looking at them all at once loses meaning
  - We may be interested in too many different variables to graph each one.
  - Note: We often refer to the data we have collected as a "sample" because it probably does not include all the possible subjects of the type in which we are interested
- One useful measure is to define the center or middle of the data
- Several different measures of central tendency are useful in different situations

Example: a sample of birthweights of live-born infants born at a private hospital in San Diego during a 1-week period (in grams)

1	3265
2	3260
3	3245
4	3484
5	4146
6	3323
7	3649
8	3200
9	3031
10	2069
11	2581
12	2841
13	3609
14	2838
15	3541
16	2759
17	3248
18	3314
19	3101
20	2834

The mean

- The **arithmetic mean** or **average** of a set of values is calculated by adding up all the values and dividing by the number of values.

If we add up all the birthweights and divide by 20, we find that the mean is 3166.9 g.

Notation

- Generically, we may refer to each value of a particular numeric variable in a dataset as x_i, where i indexes observations.

  In the birthweights data,

  $$x_1 = 3265$$

  $$x_{15} = 3541$$

- So all the values for this variable may be referred to as $x_1, \ldots, x_n$, where n is the total number of observations in the dataset
- We can use the summation sign Σ to indicate a sum. The following notation

  $$\Sigma_{i=1}^{n} x_i$$

  is a short way of writing

  $$x_1 + x_2 + \ldots + x_n$$

- We can write the computation of the mean as

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

- $\bar{x}$ is the standard notation for the mean, if we are referring to the individual data values as x_is

The mean is very sensitive to extreme values in the sample.

Example: the mean of the following numbers is 84.

```
75  82  95  80  88
```

But the mean of the following numbers is 74.

```
25  82  95  80  88
```