Statistics

Raw Data

- **Raw data is random and unranked data.**
- **Organizing Data**
  - **Frequency distributions list all the categories and the numbers of elements that belong to each category**

Frequency Distributions for Qualitative Data

- **Status of 50 Students (p. 27)**

Status	Tally	f

Relative Frequency Distribution

- **Status of 50 Students (p. 27)**

Status	Tally	f	Relative Freq
F		12	
So		12	
Jr		15	
Sr		11	

Displaying info from a freq. distribution

Stress on the Job for 30 Employees

Stress	f	Rel Freq
Very	10	
Somewhat	14	
None	6	

- **Construct a bar graph of the frequencies.**
- **Construct a bar graph of the relative frequencies**

Frequency Distribution for Quantitative Data

Table 2-2 Pulse Rates of Females

Pulse Rate	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0
120–129	1

The *frequency* for a particular class is the number of original values that fall into that class.

Lower Class Limits

are the smallest numbers that can actually belong to different classes

Table 2-2 Pulse Rates of Females

Pulse Rate	Frequency
60-69	12
70-79	14
80-89	11
90-99	1
100-109	1
110-119	0
120-129	1

Lower Class Limits

Upper Class Limits

are the largest numbers that can actually belong to different classes

Table 2-2 Pulse Rates of Females

Pulse Rate	Frequency
60-69	12
70-79	14
80-89	11
90-99	1
100-109	1
110-119	0
120-129	1

Upper Class Limits

Class Boundaries

are the numbers used to separate classes, but without the gaps created by class limits

Table 2-2 Pulse Rates of Females

	Pulse Rate	Frequency
59.5	60-69	12
69.5	70-79	14
79.5	80-89	11
89.5	90-99	1
99.5	100-109	1
109.5	110-119	0
119.5	120-129	1
129.5		

Class Boundaries

Class Midpoints

are the values in the middle of the classes and can be found by adding the lower class limit to the upper class limit and dividing the sum by two

Table 2-2 Pulse Rates of Females

	Pulse Rate	Frequency
64.5	60-69	12
74.5	70-79	14
84.5	80-89	11
94.5	90-99	1
104.5	100-109	1
114.5	110-119	0
124.5	120-129	1

Class Midpoints

Class Width

is the difference between two consecutive lower class limits or two consecutive lower class boundaries

Table 2-2 Pulse Rates of Females

	Pulse Rate	Frequency
10	60-69	12
10	70-79	14
10	80-89	11
10	90-99	1
10	100-109	1
10	110-119	0
10	120-129	1

Class Width

relative frequency = $\dfrac{\text{Frequency of that class}}{\sum f}$

Table 2-2 Pulse Rates of Females

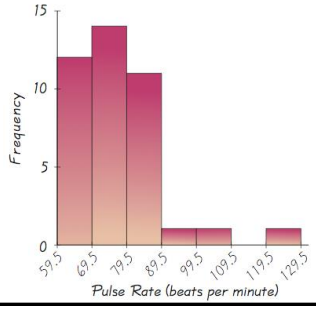
Pulse Rate	Frequency	Relative Frequency
60-69	12	
70-79	14	
80-89	11	
90-99	1	
100-109	1	
110-119	0	
120-129	1	

Histogram (diff from a bar graph)

Basically a graphic version of a frequency distribution.

Table 2-2 Pulse Rates of Females

Pulse Rate	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0
120–129	1



Histogram

The bars on the horizontal scale are labeled with one of the following:

(1) Class boundaries

(2) Class midpoints

(3) Lower class limits (introduces a small error)

Horizontal Scale for Histogram: Use class boundaries or class midpoints.

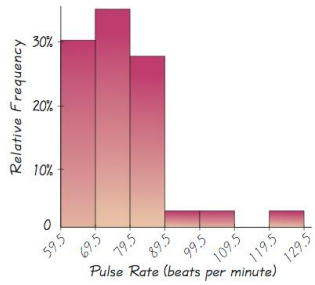
Vertical Scale for Histogram: Use the class frequencies.

Relative Frequency Histogram

Has the same shape and horizontal scale as a histogram, but the vertical scale is marked with relative frequencies instead of actual frequencies

Table 2-3 Relative Frequency Distribution of Pulse Rates of Females

Pulse Rate	Relative Frequency
60–69	30%
70–79	35%
80–89	27.5%
90–99	2.5%
100–109	2.5%
110–119	0
120–129	2.5%



Frequency Polygon

Uses line segments connected to points directly above class midpoint values

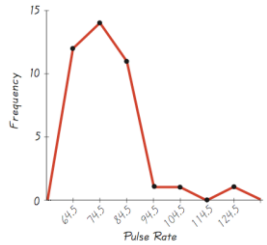
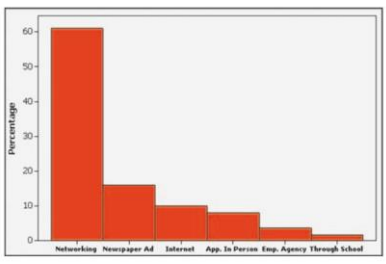


Figure 2-5 Frequency Polygon: Pulse Rates of Women

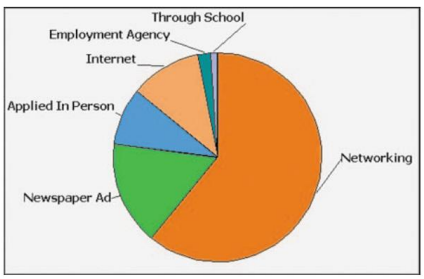
Pareto Chart

A bar graph for qualitative data, with the bars arranged in descending order according to frequencies



Pie Chart

A graph depicting qualitative data as slices of a circle, size of slice is proportional to frequency count



Important Principles Suggested by Edward Tufte

For small data sets of 20 values or fewer, use a table instead of a graph.

A graph of data should make the viewer focus on the true nature of the data, not on other elements, such as eye-catching but distracting design features.

Do not distort data, construct a graph to reveal the true nature of the data.

Almost all of the ink in a graph should be used for the data, not the other design elements.

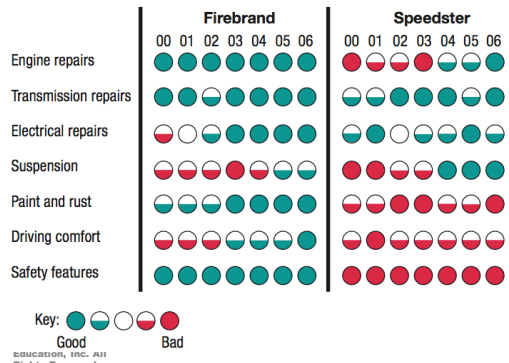
Important Principles Suggested by Edward Tufte

Don't use screening consisting of features such as slanted lines, dots, cross-hatching, because they create the uncomfortable illusion of movement.

Don't use area or volumes for data that are actually one-dimensional in nature. (Don't use drawings of dollar bills to represent budget amounts for different years.)

Never publish pie charts, because they waste ink on nondata components, and they lack appropriate scale.

Car Reliability Data



Nonzero Axis

Are misleading because one or both of the axes begin at some value other than zero, so that differences are exaggerated.

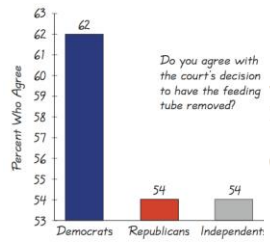


Figure 2-1 Survey Results by Party

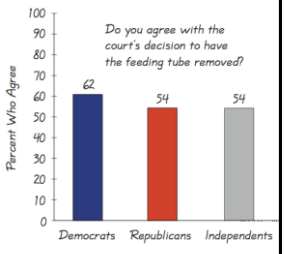
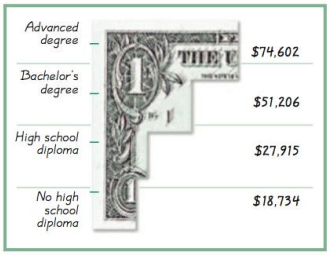


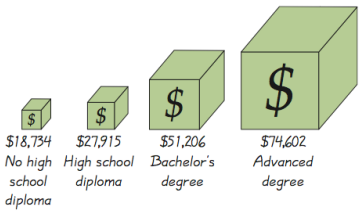
Figure 2-9 Survey Results by Party

Annual Incomes of Groups with Different Education Levels



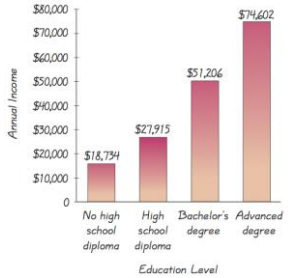
Bars have same width, too busy, too difficult to understand.

Annual Incomes of Groups with Different Education Levels



Misleading. Depicts one-dimensional data with three-dimensional boxes. Last box is 64 times as large as first box, but income is only 4 times as large.

Annual Incomes of Groups with Different Education Levels



Fair, objective, unencumbered by distracting features.