Chapter 1

Picturing Distributions with Graphs

Statistics

Statistics is a science that involves the extraction of information from numerical data obtained during an experiment or from a sample. It involves the design of the experiment or sampling procedure, the collection and analysis of the data, and making inferences (statements) about the population based upon information in a sample.

Individuals and Variables

- Individuals (outcomes)
  - the objects described by a set of data
  - may be people, animals, or things
- Variable
  - any characteristic of an individual
  - can take different values for different individuals

Variables

- Categorical
  - Places an individual into one of several groups or categories
- Quantitative (Numerical)
  - Takes numerical values for which arithmetic operations such as adding and averaging make sense

Case Study

The Effect of Hypnosis
on the
Immune System

reported in *Science News*, Sept. 4, 1993, p. 153

Case Study

The Effect of Hypnosis
on the
Immune System

Objective:
To determine if hypnosis strengthens the disease-fighting capacity of immune cells.

Case Study

- 65 college students.
  - 33 easily hypnotized
  - 32 not easily hypnotized
- white blood cell counts measured
- all students viewed a brief video about the immune system.

Case Study

- Students randomly assigned to one of three conditions
  - subjects hypnotized, given mental exercise
  - subjects relaxed in sensory deprivation tank
  - control group (no treatment)

Case Study

- white blood cell counts re-measured after one week
- the two white blood cell counts are compared for each group
- results
  - hypnotized group showed larger jump in white blood cells
  - "easily hypnotized" group showed largest immune enhancement

Case Study

Variables measured

categorical	◆ *Easy* or *difficult* to achieve hypnotic trance
	◆ Group assignment
quantitative	◆ Pre-study white blood cell count
	◆ Post-study white blood cell count

Case Study

Weight Gain Spells
Heart Risk for Women

"Weight, weight change, and coronary heart disease in women." W.C. Willett, et. al., vol. 273(6), *Journal of the American Medical Association*, Feb. 8, 1995.

(Reported in *Science News*, Feb. 4, 1995, p. 108)

Case Study

Weight Gain Spells
Heart Risk for Women

Objective:
To recommend a range of body mass index (a function of weight and height) in terms of coronary heart disease (CHD) risk in women.

Case Study

- Study started in 1976 with 115,818 women aged 30 to 55 years and without a history of previous CHD.
- Each woman's weight (body mass) was determined.
- Each woman was asked her weight at age 18.

Case Study

- The cohort of women were followed for 14 years.
- The number of CHD (fatal and nonfatal) cases were counted (1292 cases).

Case Study

Variables measured

quantitative
- Age (in 1976)
- Weight in 1976
- Weight at age 18

categorical
- Incidence of coronary heart disease
- Smoker or nonsmoker
- Family history of heart disease

Distribution

(getting the quantitative picture)

- Tells what values a variable takes and how often it takes these values
- Can be a table, graph, or function

Displaying Distributions

- Categorical variables
  - Pie charts (when categories make a whole)
  - Bar graphs
- Quantitative variables
  - Histograms
  - Stemplots (stem-and-leaf plots)

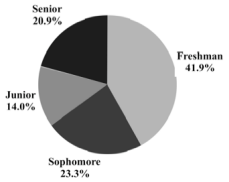
Class Make-up on First Day

Data Table

Year	*Count*	*Percent*
Freshman	18	41.9%
Sophomore	10	23.3%
Junior	6	14.0%
Senior	9	20.9%
Total	43	100.1%

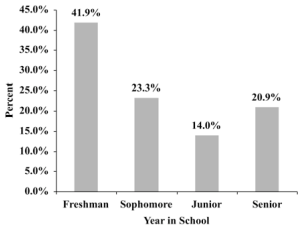
Class Make-up on First Day

Pie Chart



Class Make-up on First Day

Bar Graph



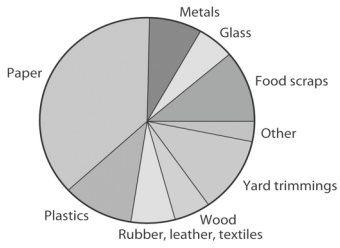
Example: U.S. Solid Waste (2000)

Data Table

Material	Weight (million tons)	Percent of total
Food scraps	25.9	11.2 %
Glass	12.8	5.5 %
Metals	18.0	7.8 %
Paper, paperboard	86.7	37.4 %
Plastics	24.7	10.7 %
Rubber, leather, textiles	15.8	6.8 %
Wood	12.7	5.5 %
Yard trimmings	27.7	11.9 %
Other	7.5	3.2 %
Total	**231.9**	**100.0 %**

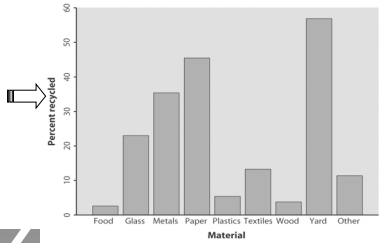
Example: U.S. Solid Waste (2000)

Pie Chart



Example: U.S. Solid Waste (2000)

Bar Graph (cannot use Pie Chart since categories do not make a whole)



Histograms

- For quantitative variables that take many values
- Divide the possible values into *class intervals* (we will only consider equal widths)
- Count how many observations fall in each interval (may change to percents)
- Draw picture representing distribution

Case Study

Weight Data

Introductory Statistics class
Spring, 1997
Virginia Commonwealth University

Weight Data

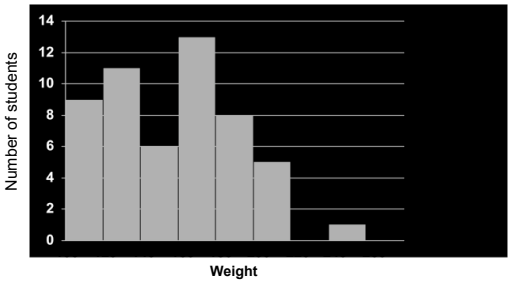
192	110	195	180	170	215
152	120	170	130	130	125
135	185	120	155	101	194
110	165	185	220	180	
128	212	175	140	187	
180	119	203	157	148	
260	165	185	150	106	
170	210	123	172	180	
165	186	139	175	127	
150	100	106	133	124	

Weight Data: Frequency Table

Weight Group	Count
100 - 120	9
120< - 140	11
140< - 160	6
160< - 180	13
180< - 200	8
200< - 220	5
220< - 240	0
240< - 260	1

sqrt(53) = 7.2, or 8 intervals; range (260-100=160) / 8 = 20 = class width

Weight Data: Histogram



* Left endpoint is included in the group, right endpoint is not.

Histograms: Class Intervals

- How many intervals?
  - One rule is to calculate the square root of the sample size, and round up.
- Size of intervals?
  - Divide range of data (max-min) by number of intervals desired, and round to convenient number
- Pick intervals so each observation can only fall in exactly one interval (no overlap)

Examining the Distribution of Quantitative Data

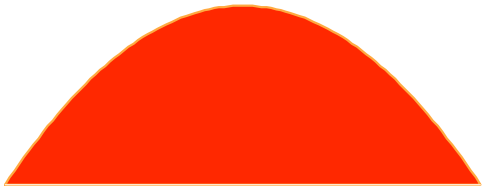
- Overall pattern of graph
- Deviations from overall pattern
- Shape of the data
- Center of the data
- Spread of the data (Variation)
- Outliers

Shape of the Data

- Symmetric
  - bell shaped
  - other symmetric shapes
- Asymmetric
  - right skewed
  - left skewed
- Unimodal, bimodal

Symmetric Bell-Shaped

Symmetric Mound-Shaped



Symmetric Uniform

Asymmetric Skewed to the Left

Asymmetric Skewed to the Right

Outliers

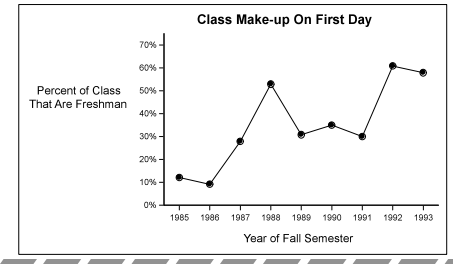
- Extreme values that fall outside the overall pattern (the data point 260 in previous histogram example)
  - May occur naturally
  - May occur due to error in recording
  - May occur due to error in measuring

Time Plots

- A time plot shows behavior over time.
- Time is always on the horizontal axis, and the variable being measured is on the vertical axis.
- Look for an overall pattern (*trend*), and deviations from this trend. Connecting the data points by lines may emphasize this trend.
- Look for patterns that repeat at known regular intervals (seasonal variations).

Class Make-up on First Day

(Fall Semesters: 1985-1993)



Average Tuition (Public vs. Private)