

Chapter 1

Statistics - the branch of mathematics concerned with the collection, analysis, and interpretation of data for presentation and decision-making purposes.

Two types of Statistics:

1. **Descriptive Statistics** consists of all methods for organizing, summarizing, and graphically representing information (data) in a clear and effective way.

Examples:

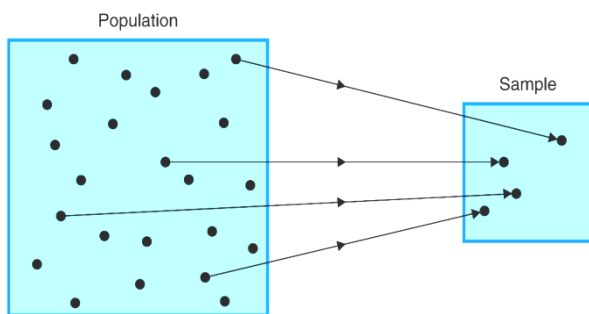
2. **Inferential Statistics** consists of methods of drawing conclusions about a population based on information obtained from a sample of the population.

a) **estimation** - educated guesses about the population.

b) **hypothesis tests** - the testing of a claim about the population.

POPULATION – All subjects (human or otherwise) that are being studied.

SAMPLE – A group of subjects selected from a population.



Raisin Bran Example:

Section 1.2 Variables and Types of Data

DATA are the values (measurements or observations) that the variables can assume. Variables that are determined by chance are called **random variables**.

Two Types of Data:

1. **Qualitative** - gives non-numerical information such as gender, eye color, blood type.
2. **Quantitative** - numerical, measurable.

Discrete data is finite (limited) and only takes on certain values. These CAN be counted.

Continuous data can assume all values between any two specific values. They are obtained by measuring.

The **Nominal level of measurement** classifies data into non-overlapping, exhausting categories in which no order or ranking can be imposed on the data. (small, medium and large, for example)

The **Ordinal level of measurement** classifies data into categories that can be ranked; however, precise differences between the ranks do not exist. (letter grades, for example)

The **Interval level of measurement** ranks data, and precise differences between units of measure do exist; however there is no meaningful zero. (IQ scores, for example)

The **Ratio level of measurement** possesses all the characteristics of interval measurement, and there exists a true zero. Ratio scales have differences between units.

Section 1.3 Data Collection and Sampling Techniques

Types of Surveys

Telephone surveys

Internet surveys

Mailed Questionnaire surveys

Personal Interview surveys

Sampling Techniques (also covered in section 14.1 in more detail)

1. **(Simple) Random Sampling** - each possible sample of a given size is equally likely to be the one selected.

advantages

disadvantages

2. **Samples of Convenience** - based on the people available to be studied

advantages

disadvantages

3. **Cluster Sampling** - randomly selecting *groups* of people to study.
example: Break a city into 1000 precincts. Randomly choose 15 to 25 of them and sample people within these precincts.

advantages

disadvantages

4. **Stratified Sampling** - Break the population into subcategories, called strata, and then sample from each strata.

advantages

disadvantages

5. **Systematic Sampling** - Select every k th member, after the first subject is chosen from 1 to k

advantages

disadvantages

Other sampling Techniques:

In **sequence sampling**, which is used in quality control, successive units taken from production lines are sampled to ensure that the products meet certain standards set by the manufacturing company.

In **double sampling**, a very large population is given a questionnaire to determine those who meet the qualifications for a study. After the questionnaires are reviewed, a second smaller population is defined. Then a sample is selected from this group.

In **multistage sampling**, the researcher uses a combination of sampling methods.

Section 1.4 Observational and Experimental Studies

In an **experimental study**, the researcher manipulates one of the variables and tries to determine how the manipulation influences the other variables.

In an **observational study**, the researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations.

Examples:

Note: Observational studies can reveal only **association**, whereas designed experiments can establish **causation** (cause-and effect.)

Sometimes when random assignment is not possible, researchers use intact groups. These types of studies are done quite often in education where already intact groups are available in the form of existing classrooms. When these groups are used, the study is said to be a **quasi-experimental study**.

An **independent variable** in an experimental study is the one that is being manipulated by the researcher. The independent variable is also called the **explanatory variable**. The resultant variable is called the **dependent variable** or the **outcome variable**.

Vocabulary:

Control Group versus **Treatment Group**

Placebo

The **Hawthorne Effect** – Researchers have found that subjects who know that they are participating in an experiment sometimes actually change their behaviors in ways that affect the results of the study.

A **confounding** (or lurking) **variable** is one that influences the dependent or outcome variable but cannot be separated from the independent variable.

Examples:

Double-blindness occurs when neither the researcher nor the sample subject knows whether the subject is part of a test group.

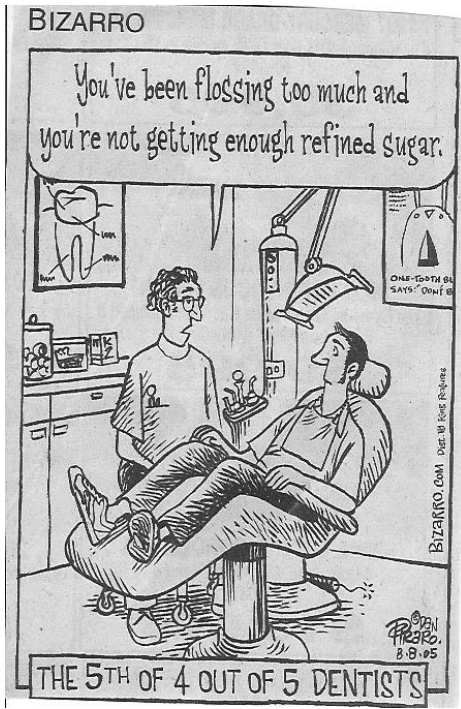
Example:

Golf Ball Example:

Section 1.5 – Uses and Abuses of Statistics

A. Suspect Samples

1. See comic below....



2. Voluntary Response Bias

B. Ambiguous Averages

The word “average” is very vague. It can actually refer to the mean, median, mode or midrange. We will talk about this at length in chapter 3.

C. Changing the Subject

Sometimes different values are used to represent the same data.

D. Detached Statistics

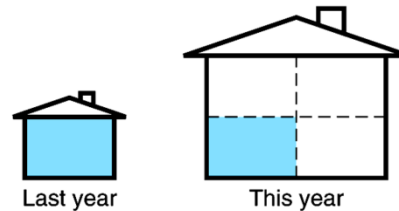
A claim is detached when no comparison is made. For example, you may hear a claim that “our brand has half the calories of the leading brand” or “our brand works 3 times faster.”

E. Implied Connections

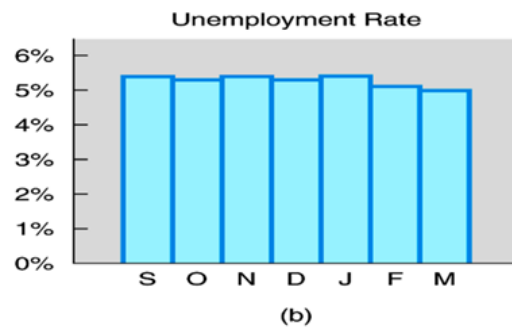
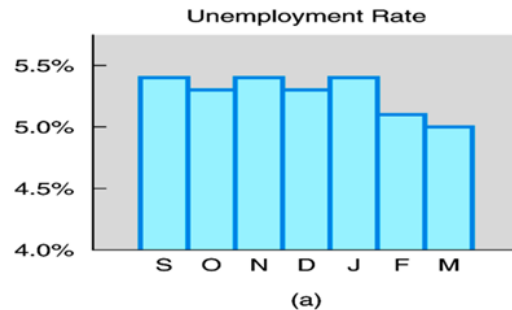
Many claims attempt to imply connections between variables that may not actually exist. Use of the phrases “may help” or “suggest” or “for some people” are phrases to watch out for.

F. Misleading Graphs

1. Improper Scaling



2. Lack of Scaling



3. Truncated Graphs

Grocery Example:

G. Faulty Survey Questions (see section 14.2 also)

1. Wording of questions

Daily Press Headline (Feb 5, 2011)

Unemployment Drop Quickens

The unemployment rate is suddenly sinking at the fastest pace in a half-century, falling to 9 percent from 9.8 percent in just two months—the most encouraging sign for the job market since the recent recession.

2. Ordering of questions

3. Types of questions (open, closed, scaled), etc.

For more information, there is extra info on this in section 14.2. This is a favorite topic of mine. I have several articles that you can download on my website, located on my “handouts page.” The URL is <http://www.stevetoner.com/handouts.html> I recommend the first two in the Statistics column.

Suggested homework:

Section 14.2, page 738, #1-10

Assigned Reading:

Survey Design Article (found at the webpage described above).

HIGH DESERT

DAILY PRESS

Monday, March 30, 2009 www.dailypress.com Victor Valley & The High Desert 50¢ plus tax

Smoke break gets more expensive

Cigarette makers raise prices, per-pack tax climbs to \$1.01

BY RICARDO ALONSO-ZALDIVAR
ASSOCIATED PRESS WRITER

WASHINGTON • However they satisfy their nicotine cravings, tobacco users are facing a big hit as the single largest federal tobacco tax increase ever takes effect Wednesday.

Tobacco companies and public health advocates, longtime foes in the nicotine battles, are trying to turn the situation to their advantage. The major cigarette makers raised prices a couple of weeks ago, partly to offset any drop in profits once the per-pack tax climbs from 39 cents to \$1.01.

Medical groups see a tax increase right in the middle of a recession as a great incentive to help persuade smokers to quit.

Tobacco taxes are soaring to finance a major expansion of health insurance for children.

Tobacco tax up more than twofold

The cost of smoking a cigarette, pipe or chewing tobacco will rise Apr. 1 to pay for an expansion in health coverage for children.

Federal tax rate on cigarettes

Period	Tax Rate
1951-82	16¢
'82-'90	20¢
'91-'92	24¢
'93-'99	34¢
'00-'01	39¢
'02-'09	\$1.01

April 1, 2009

SOURCES: Congressional Research Service; U.S. Treasury

FILE PHOTO, THE ASSOCIATED PRESS

A REAL DRAG: An employee takes a drag on a cigarette at Morgan's Place bar and restaurant in Harrisburg, Pa.

>>FINDER

HIGH DESERT

Assemblyman Knight honors Otwell as 'Woman of the Year'

Page B1

STATE•NATION•WORLD

Mexico's war on drugs

REYNOSA, MEXICO • The

SEE TOBACCO TAX • PAGE 5

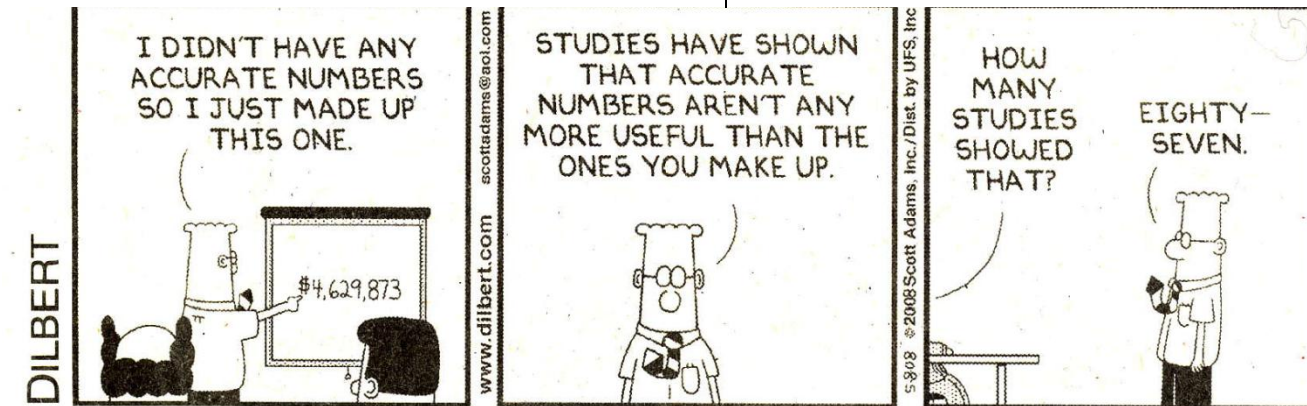
Discussion Questions

Sampling Bias

1. The author Shere Hite undertook a study of women’s attitudes toward sex and love by distributing 100,000 questionnaires through women’s groups. Only 4.5% of the questionnaires were returned. Based on this sample of women, Hite wrote *Women and Love*, a best-selling book claiming that women are fed up with men. For example, 91% of the divorced women in the sample said that they had initiated the divorce, and 70% of the married women said that they had committed adultery. Explain briefly why Hite’s sampling method is nearly certain to produce a strong bias. Are the sampling results cited (91% and 70%) much higher or much lower than the truth about the population of all American women?
2. Some television stations take quick polls of public opinion by announcing a question on the air and asking viewers to call one of two telephone numbers to register their opinion as “Yes” or “No.” Telephone companies make available “900” numbers for this purpose; dialing such a number results in a small charge to your telephone bill. One such call-in poll finds that 73% of those who called are opposed to a proposed local gun control ordinance. Explain why this sampling method is biased. Is the percent of the population who oppose gun control probably higher or lower than the 73% of the sample who are opposed?
3. Identify the source of the bias and specify the direction of the bias (that is, whether the sample result will be systematically above or below the true population result). The Miami Police Department wants to know if black residents of Miami are satisfied with police service in their neighborhood. A questionnaire is prepared. An SRS of 300 mailing addresses in predominantly black neighborhoods is chosen, and a uniformed police officer is sent to each address to interview an adult resident.
4. Comment on each of the following as a potential sample survey question. Is the question clear? Is it slanted toward a desired response?
 - a) “Does your family use food stamps?”
 - b) “Which of the following best represents your opinion on gun control?”
 1. The government should confiscate our guns.
 2. We have the right to keep and bear arms.”
 - c) “A freeze in nuclear weapons should be favored because it would begin a much-needed process to stop everyone in the world from building nuclear weapons now and reduce the possibility of nuclear war in the future. Do you agree or disagree?”
 - d) “In view of escalating environmental degradation and incipient resource depletion, would you favor economic incentives for recycling of resource-intensive consumer goods?”

5. You are on the staff of a member of Congress who is considering a controversial bill that would provide for government-sponsored insurance to cover care in nursing homes. You report that 1128 letters dealing with the issue have been received, of which 871 oppose the legislation. “I’m surprised that most of my constituents oppose the bill. I thought it would be quite popular,” says the congresswoman. Are you convinced that a majority of the voters oppose the bill? State briefly how you would explain the statistical issue to the congresswoman.

6. Suppose that simple random sample of students on a college campus are questioned about a proposed policy to ban smoking in all campus buildings. If one group is interviewed by a person wearing a T-shirt and jeans and smoking a cigarette while another group is interviewed by a non-smoker wearing a business suit, would you expect that the proportion declaring agreement with the policy might differ between the two groups? Explain.



“In the future please keep your opinions to yourself.”



“I’m sick and tired of research polls ... what do the rest of you think?”

Methods of Sampling

1. **Random Sampling** - Each member of the population has an equal chance of being selected
2. **Stratified Sampling** - Classify the population into at least two strata, and then draw a sample from each.
3. **Systematic Sampling** - Select every k th member.
4. **Cluster Sampling** - Divide the population area into sections, randomly select a few of these sections, and then choose or sample all members in each.
5. **Convenience Sampling** - Use results that are readily available.

Directions: Identify the type of sampling used.

1. Motorola selects every 50th pager from the assembly line for careful testing and analysis. _____
2. A reporter writes the name of each U.S. Senator on a separate card, shuffles the cards, and then draws 5 names. _____
3. A dean at Ohio State University surveys all students from each of 12 randomly selected classes. _____
4. A dean at Menlo College selects 15 men and 15 women from each of 4 classes. _____
5. *Glamour* magazine obtains sample data from readers who decide to mail in a questionnaire printed in the latest issue. _____
6. An IRS auditor randomly selects 15 taxpayers with less than \$25,000 in gross income and 15 taxpayers with gross income of at least \$25,000. _____
7. CBS News polls 750 men and 750 women about their use of credit cards. _____
8. A market researcher for the Ford Motor Company interviews all drivers on each of 15 randomly selected city blocks. _____
9. A medical researcher from John Hopkins University interviews all leukemia patients in each of 20 randomly selected counties. _____
10. A reporter for *Business Week* magazine interviews every 50th chief executive officer in that magazine’s listing of CEO’s of the 1000 companies with the highest stock market values. _____
11. A reporter for Business Week magazine obtains a numbered listing of the 1000 companies with the highest stock market values, uses a computer to generate 20 random numbers between 1 and 1000, and then interviews the chief executive officers of companies corresponding to these numbers. _____
12. In conducting research for a psychology course, a student at Boston College interviews 40 students who are leaving the cafeteria. _____

Answers: 1. Systematic, 2. Random, 3. Cluster, 4. Stratified, 5. Convenience, 6. Stratified, 7. Stratified, 8. Cluster, 9. Cluster, 10. Systematic, 11. Random, 12. Convenience.

Chapter 2 – Frequency Distributions and Graphs

When data are collected in original form, they are called **raw data**. The **frequency** is the number of values in a specific class of the distribution. (*Note:* “data” is both a singular and plural word.)

A **frequency distribution** is the organization of raw data in table form using classes and frequencies.

GUIDELINES for grouping data into classes (categories):

1. There should be between 5 and 20 classes.
2. The class width should be an odd number. (Your book’s suggestion only, as this ensures that the class midpoint has the same place value as the data.)
3. The classes should be mutually exclusive. (No data value should belong to more than one class.)
4. The classes should be continuous, even if there are no values in a class.
5. The classes must be exhaustive.
6. The classes must be of equal in width. (This avoids a distorted view of the data.)

METHODS FOR GROUPING INFORMATION:

1. **Tally Method** – use “notch marks” with a “cross-bar” for every fifth item.

Class limits	Class boundaries	Tally	Frequency	Cumulative frequency
24–30	23.5–30.5	///	3	3
31–37	30.5–37.5	/	1	4
38–44	37.5–44.5	///	5	9
45–51	44.5–51.5	///	9	18
52–58	51.5–58.5	///	6	24
59–65	58.5–65.5	/	1	25
			25	

2. **Frequency Distribution** - the tabular summary listing all classes and their frequencies.
3. **Relative Frequency** - the *percentage* of items falling into a class, expressed as a *decimal*.
4. **Cumulative Frequency Distribution** - the *frequency* of items falling in each class that is less than or equal to a particular value.

5. An **ogive** is a graph that represents the cumulative relative frequencies for the classes in a frequency distribution.

Lower Class Limit - the smallest data value that can go into a class.

Upper Class Limit - the largest data value that can go into a class.

The lower class limit for a frequency distribution with the class $0 \leq x < 5$ is 0, while the upper class limit for the class is 4. (We assume that the data consists of integers.)

If we include the right endpoint of the class, that is, $0 \leq x \leq 5$, then the **upper class limit is 5**.

Class boundaries are used to separate the classes so that there are no gaps in the frequency distribution.

Class midpoints - of a class found by taking the average of the LCL and UCL.

$$class\ midpoint = \frac{(LCL + UCL)}{2}$$

Class width - the difference between the LCL of a given class and the LCL of the next higher class.

Our Class Data:

of Fast Food Meals Eaten In the Past 14 Days

Create a frequency distribution from the data:

Class (# of meals)	Frequency
$0 \leq x < 5$	
$5 \leq x < 10$	
$10 \leq x < 15$	
$15 \leq x < 20$	
$20 \leq x < 25$	
$25 \leq x < 30$	
$30 \leq x < 35$	
$35 \leq x < 40$	
$40 \leq x < 45$	
$45 \leq x < 50$	
$50 \leq x < 55$	
$55 \leq x < 60$	

Create a relative frequency distribution from the data:

Class (# of meals)	Relative Frequency
$0 \leq x < 5$	
$5 \leq x < 10$	
$10 \leq x < 15$	
$15 \leq x < 20$	
$20 \leq x < 25$	
$25 \leq x < 30$	
$30 \leq x < 35$	
$35 \leq x < 40$	
$40 \leq x < 45$	
$45 \leq x < 50$	
$50 \leq x < 55$	
$55 \leq x < 60$	

Create a cumulative frequency distribution from the data:

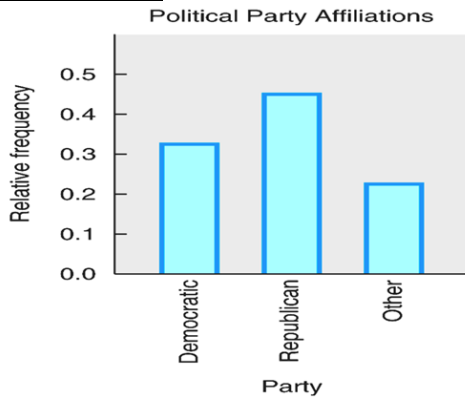
Class (# of meals)	Cumulative Frequency
$x < 5$	
$x < 10$	
$x < 15$	
$x < 20$	
$x < 25$	
$x < 30$	
$x < 35$	
$x < 40$	
$x < 45$	
$x < 50$	
$x < 55$	
$x < 60$	

Create a cumulative relative frequency distribution from the data:

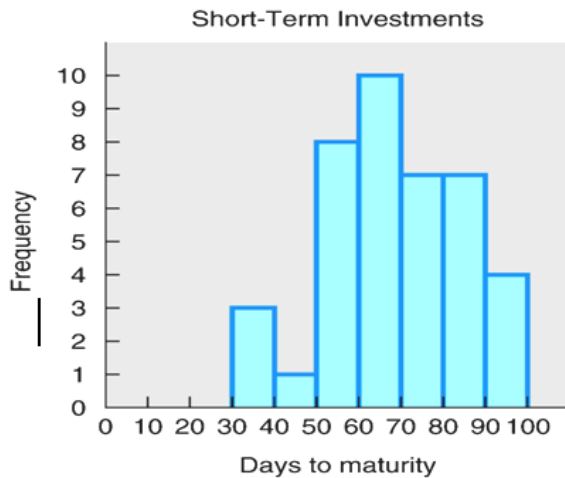
Class (# of meals)	Cumulative Relative Frequency
$x < 5$	
$x < 10$	
$x < 15$	
$x < 20$	
$x < 25$	
$x < 30$	
$x < 35$	
$x < 40$	
$x < 45$	
$x < 50$	
$x < 55$	
$x < 60$	

A **histogram** is a graphical representation of the frequency distribution. It differs from a bar chart (also known as a Pareto chart) in that there is a numerical scaling on the horizontal axis. A Pareto chart contains categories, or other non-numerical classes.

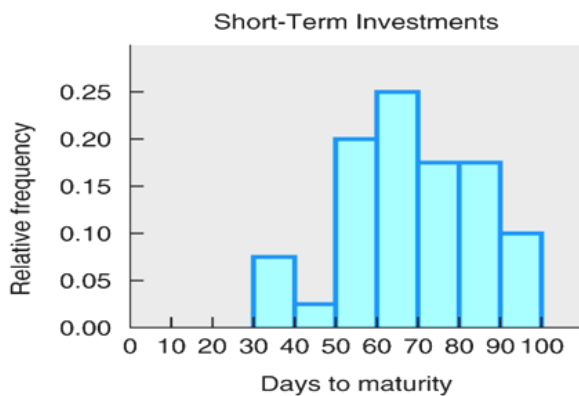
Pareto Chart:



Frequency Histogram

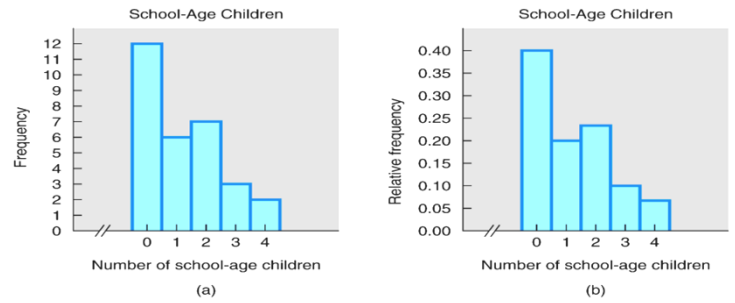


Relative Frequency Histogram using the same data:

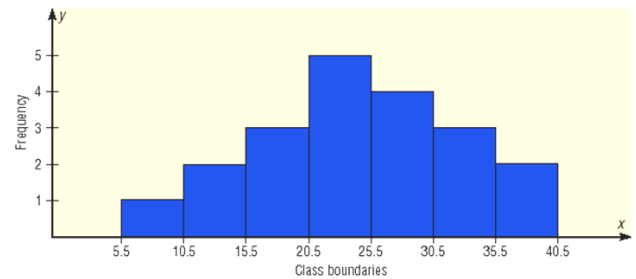


Note:

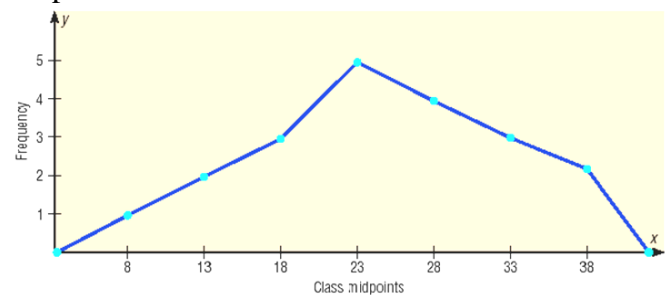
Truncation marks:



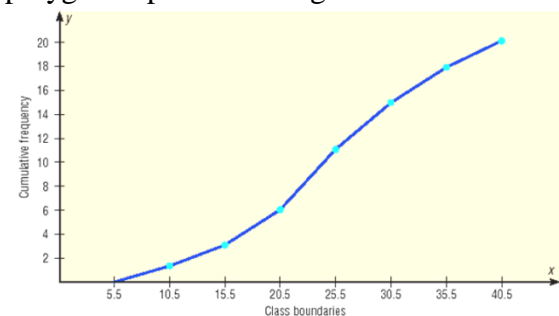
Look at the following frequency histogram. Why is the horizontal axis labeled as it is?



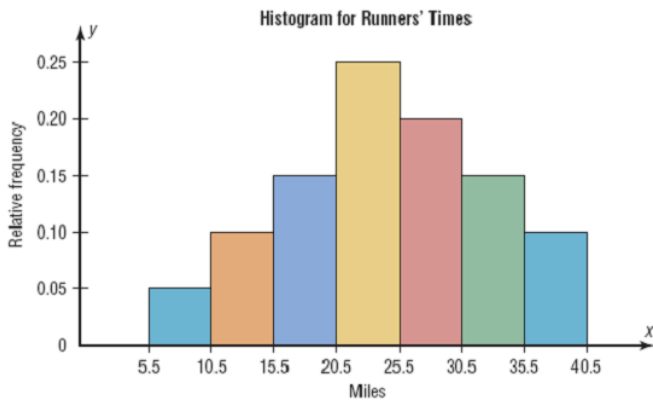
A **frequency polygon** is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of their classes. The frequencies are represented by their heights of the points.



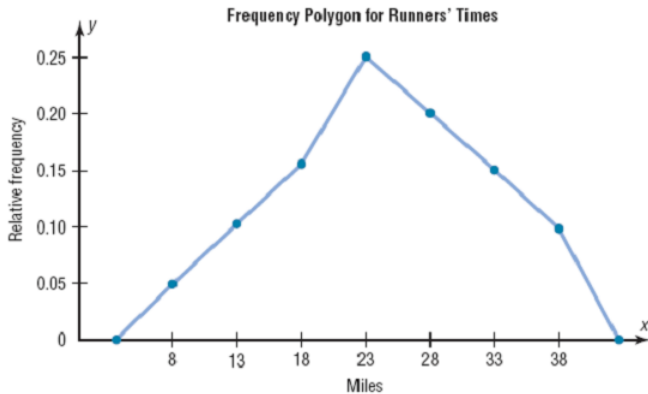
The **cumulative frequency** is the sum of the frequencies accumulated up to the upper boundary of a class in a distribution. A cumulative frequency polygon is pictured at right.



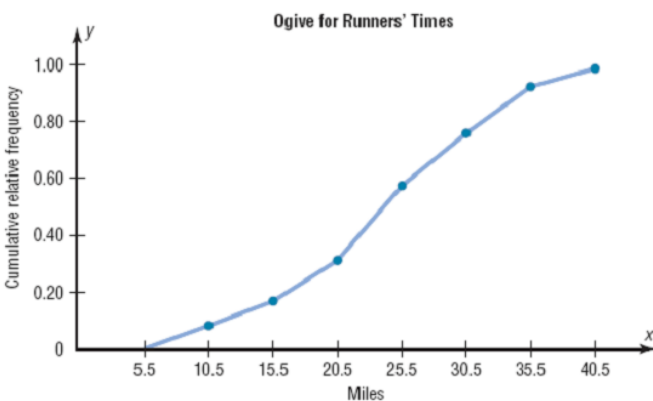
The following is a **relative frequency histogram**:



A **relative frequency polygon** is pictured next. Note that the midpoints of the classes are used on the horizontal axis.



An **ogive** is a graph that represents the cumulative relative frequencies for the classes in a frequency distribution.



Note:

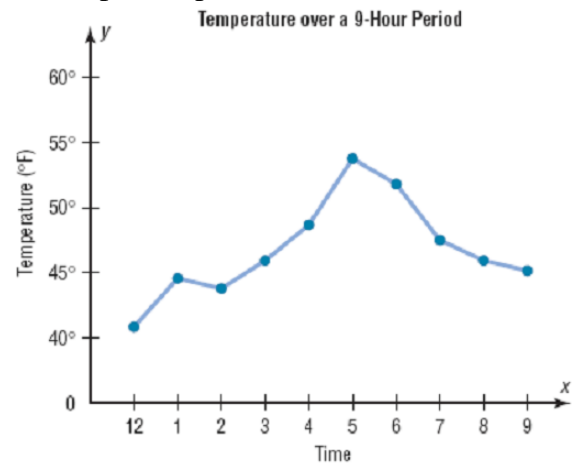
2.3 Other Types of Charts and Graphs

Web tutorial on preparing effective graphs and charts:
<http://bcs.bedfordstmartins.com/techcomm8e/tutorials/chartsgraphs/1b.html>

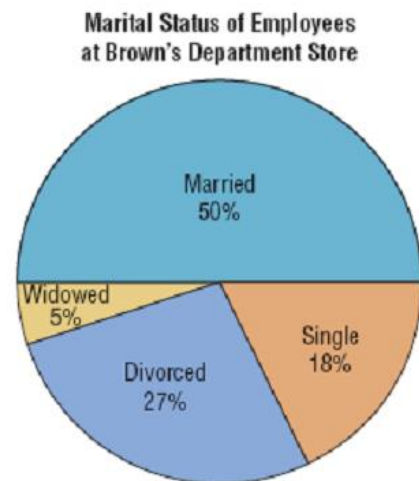
A **bar graph** represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies of the data.

A **Pareto chart** is used to represent a frequency distribution for a categorical variable and the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest.

A **time series graph** represents the data that occur over a specific period of time.



A **pie chart** is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.



A **stem and leaf plot** is a data plot that uses part of the data value as the stem and part of the data value as the leaf to form groups or classes.

Days to maturity for 40 short-term investments:

70	64	99	55	64	89	87	65
62	38	67	70	60	69	78	39
75	56	71	51	99	68	95	86
57	53	47	50	55	81	80	98
51	36	63	66	85	79	83	70

Diagrams for days-to-maturity data:

(a) stem-and-leaf

Stems	Leaves
3	8 6 9
4	7
5	7 1 6 3 5 1 0 5
6	2 4 7 3 6 4 0 9 8 5
7	0 5 1 0 9 8 0
8	5 9 1 7 0 3 6
9	9 9 5 8

(b) Ordered stem-and-leaf

3	6 8 9
4	7
5	0 1 1 3 5 5 6 7
6	0 2 3 4 4 5 6 7 8 9
7	0 0 0 1 5 8 9
8	0 1 3 5 6 7 9
9	5 8 9 9

Back-to-back Stem and Leaf Plots

Atlanta			Philadelphia	
	9 8 6	2		5
8 6 4 4 2 2 2 2 2 1		3	0 0 0 0 2 2 3 4 6 6 6 8 8 9 9	
	7 4 4 0 0	4		0 0 0 0
	5 3 2 2 0 0	5		0 3 4 8
	3 0	6		1
	0	7		

Stem-and-leaf diagram for cholesterol levels: (a) using one line per stem (b) using two lines per stem

210	209	212	208	202	218	200	214	218	210
217	207	210	203	215	221	213	210	199	208

19	9
20	8 2 9 7 0 8 3
21	0 7 5 0 8 2 0 0 3 8 4
22	1

(a)

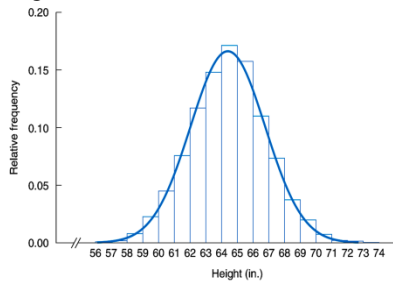
19	9
20	2 0 3
20	8 9 7 8
21	0 0 2 0 0 3 4
21	7 5 8 8
22	1
22	

(b)

Distribution Shapes

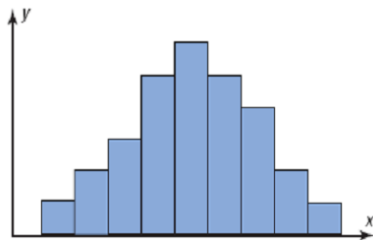
The distribution of a data set is a table, graph, or formula that tells us the values of the observations and how often they occur. An important aspect of the distribution of a quantitative data set is its shape.

Relative-frequency histogram and approximating smooth curve for the distribution of heights

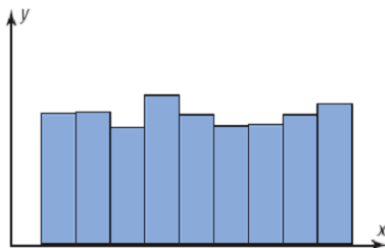


Common distribution shapes

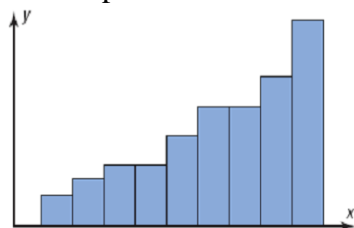
1. Bell-shaped



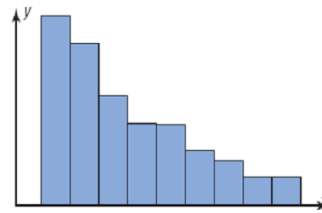
2. Uniform



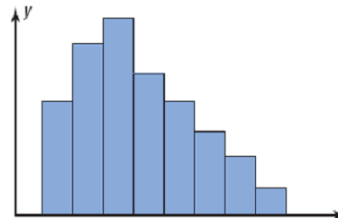
3. J-shaped



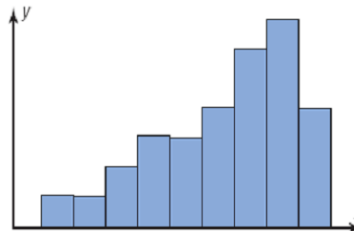
4. Reverse J-shaped



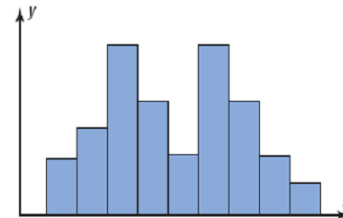
5. Right-Skewed



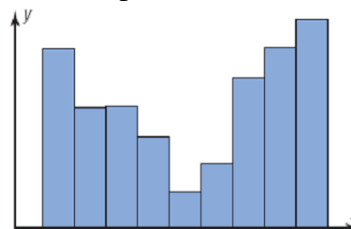
6. Left-Skewed



7. Bimodal



8. U-Shaped



KEY FACT (*paraphrased*)

If a random sample of a "large enough" size is taken from a population, the shape of the distribution of the sample will approximate the shape of the population's distribution.

* The larger the sample size, the better the approximation tends to be.