Business Analytics

Descriptive Statistics

Faton Berisha

Chapter 2

Descriptive Statistics

Descriptive Statistics

2.1 Describing the Shape of a Distribution

2.2 Describing Central Tendency

2.3 Measures of Variation

2.4 Percentiles, Quartiles and Box-and-Whiskers Displays

2.5 Describing Qualitative Data

2.6 Using Scatter Plots to Study the Relationship Between Variables (Optional)

Descriptive Statistics Continued

2.7 Misleading Graphs and Charts (Optional)

2.8 Weighted Means and Grouped Data (Optional)

2.9 The Geometric Mean (Optional)

The Shape of a Distribution

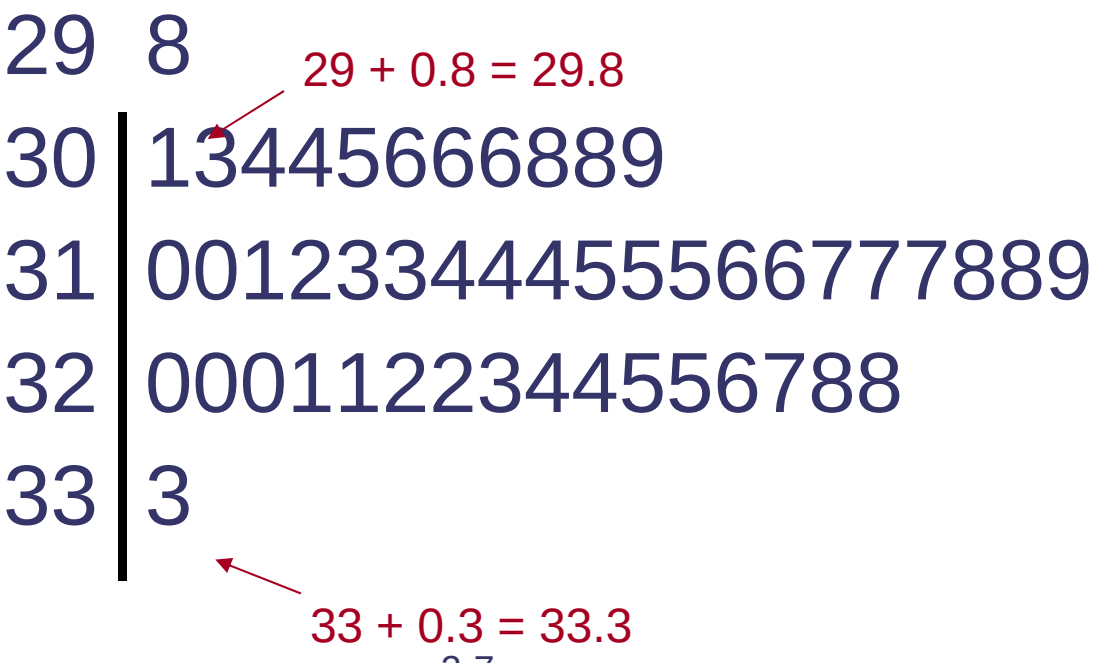
- To know what the population looks like, find the “shape” of its distribution
- Picture the distribution graphically by any of the following methods:
  - Stem-and-leaf display
  - Frequency distributions
  - Histogram
  - Dot plot

Stem-and-Leaf Display

- The purpose of a stem-and-leaf display is to see the overall pattern of the data, by grouping the data into classes
  - To see:
    - the variation from class to class
    - the amount of data in each class
    - the distribution of the data within each class
- Best for small to moderately sized data distributions

Car Mileage Example

- Example 2.1. The Car Mileage Case
  - Data in Table 2.1
- The stem-and-leaf display of car mileages:



Car Mileage Example (Cont.)

- Another display of the same data using more classes
  - Starred classes (*) extend from 0.0 to 0.4
    - E.g., the row labeled 30* contain the data from 30.0 to 30.4 mpg
  - Unstarred classes extend from 0.5 to 09
    - E.g., the row labeled 30 contains the data from 30.5 to 30.9 mpg

29	8
30*	1344
30	5666889
31*	001233444
31	55566777889
32*	0001122344
32	556788
33*	3

Car Mileage: Results

- Looking at the last stem-and-leaf display, the distribution appears almost “symmetrical”
  - The upper portion of the display…
    - Stems 29, 30*, 30, and 31*
  - … is almost a mirror image of the lower portion of the display
    - Stems 31, 32*, 32, and 33*
  - But not exactly a mirror reflection
    - Maybe slightly more data in the lower portion than in the upper portion
      - Later, we will call this a slightly “left-skewed” distribution

Frequency Distribution

- A *frequency distribution* is a list of data classes with the count or "*frequency*" of values that belong to each class
  - "Classify and count"
  - The frequency distribution is a table
- Show the frequency distribution graphically in a *histogram*
  - The histogram is a picture of the frequency distribution
- Examples 2.2 and 2.4, The Payment Time Case

Constructing a Frequency Distribution

- Steps in making a frequency distribution:
  1. Determine the number of classes *K*
  2. Determine the class length
  3. Set the starting value for the classes, that is, the distribution “floor”
  4. Calculate the class limits
  5. Setup all the classes
- Then tally the data into the *K* classes and record the frequencies

Number of Classes

- Group all of the n data into K number of classes
- K is the smallest whole number for which

$$2^K \geq n$$

- In Examples 2.2 and 2.4, $n = 65$
  - For $K = 6$, $2^6 = 64 < n$
  - For $K = 7$, $2^7 = 128 > n$
  - So use $K = 7$ classes

Class Length

- Class length L is the step size from one to the next

$$L = \frac{\text{Largest value} - \text{Smallest value}}{K}$$

- In Examples 2.2 and 2.4, The Payment Time Case, the largest value is 29 days and the smallest value is 10 days, so

$$L = \frac{29 - 10 \text{ days}}{7 \text{ classes}} = \frac{19 \text{ days}}{7 \text{ classes}} = 2.7143 \text{ days/class}$$

- Arbitrarily round the class length up to 3 days/class

Starting the Classes

- The classes start on the smallest data value
  - This is the lower limit of the first class
- The upper limit of the first class is

  smallest value + *L*

  - In the example, the first class starts at 10 days and goes up to 13 days
- The second class starts at the upper limit of the first class and goes up *L* more
  - The second class starts at 13 days and goes up to 16 days
- And so on…

Tallies and Frequencies:
Example 2.4

Classes (days)	Frequency
10 to 13	3
13 to 16	14
16 to 19	23
19 to 22	12
22 to 25	8
25 to 28	4
28 to 31	1
	65

Check: All frequencies must sum to n

Relative Frequency

- The *relative frequency* of a class is the proportion or fraction of data that is contained in that class
  - Calculated by dividing the class frequency by the total number *n* of data values
  - Relative frequency may be expressed as either a decimal or percent
  - A *relative frequency distribution* is a list of all the data classes and their associated relative frequencies

Relative Frequency: Example 2.4

Classes (days)	Frequency	Relative Frequency
10 to 13	3	3/65 = 0.0462
13 to 16	14	14/65 = 0.2154
16 to 19	23	0.3538
19 to 22	12	0.1846
22 to 25	8	0.1231
25 to 28	4	0.0615
28 to 31	1	0.0154
	65	1.0000

Check: All relative frequencies must sum to 1

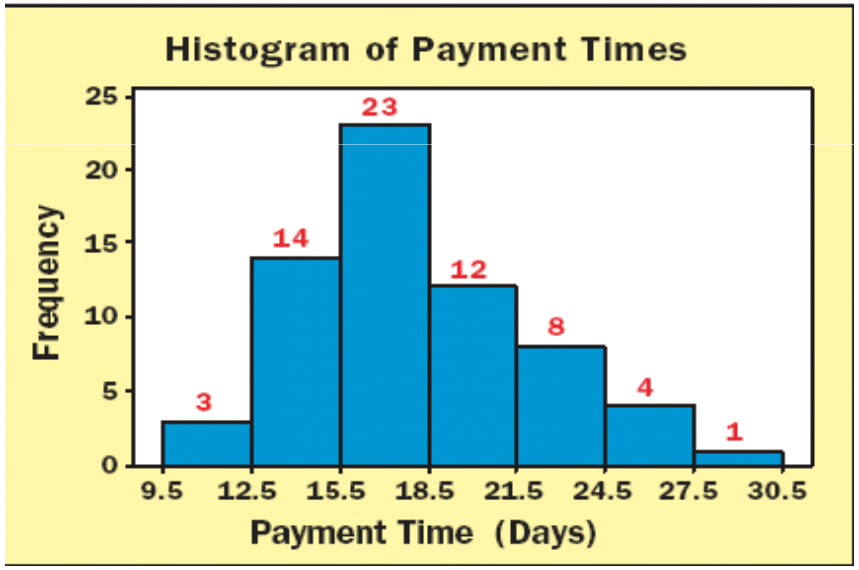
Histogram

- A graph in which rectangles represent the classes
- The base of the rectangle represents the class length
- The height of the rectangle represents
  - the frequency in a frequency histogram, or
  - the relative frequency in a relative frequency histogram

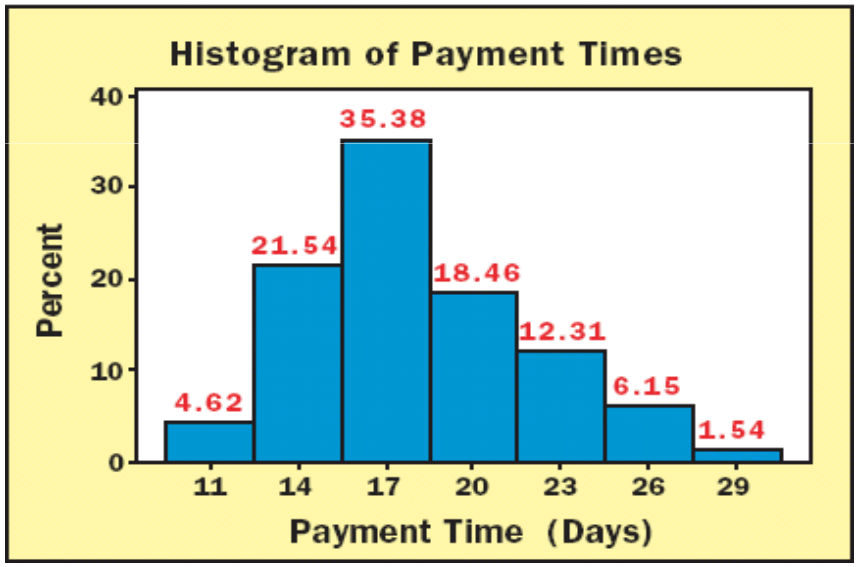
Histograms

Example 2.2: The Payment Times Case

Frequency Histogram

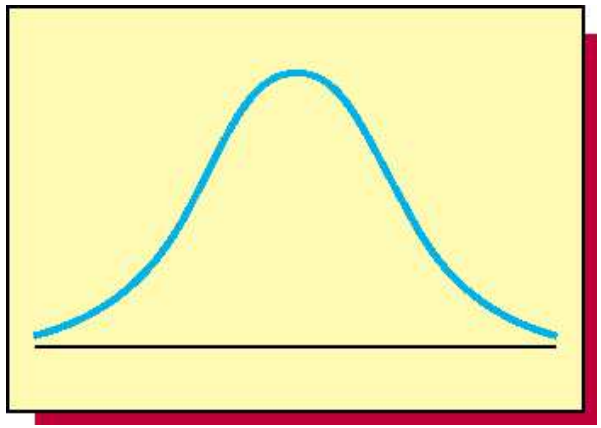


Relative Frequency Histogram



As with the earlier stem-and-leaf display, the tail on the right appears to be longer than the tail on the left

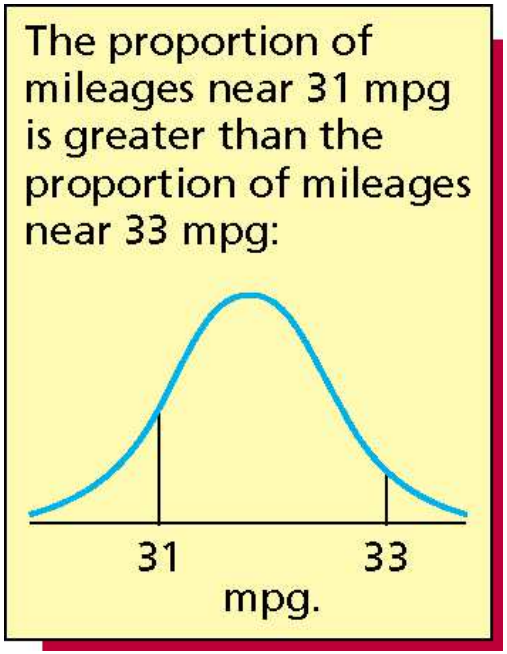
The Normal Curve



Symmetrical and bell-shaped curve for a normally distributed population

The height of the normal over any point represents the relative proportion of values near that point

Example 2.1, The Car Mileages Case

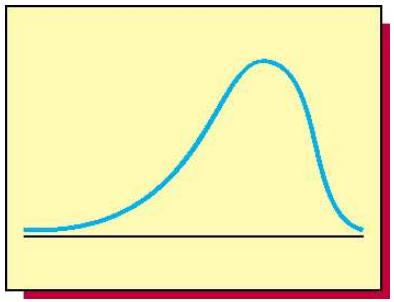


Skewed Distributions

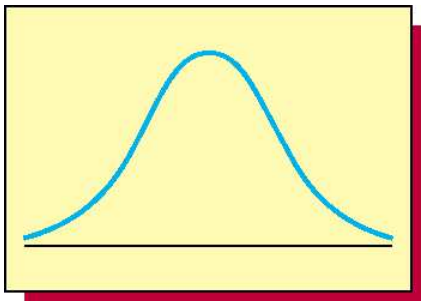
Skewed distributions are not symmetrical about their center. Rather, they are lop-sided with a longer tail on one side or the other.

- A population is distributed according to its relative frequency curve
- The skew is the side with the longer tail

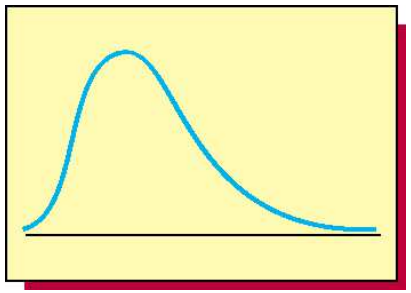
Left Skewed



Symmetric



Right Skewed

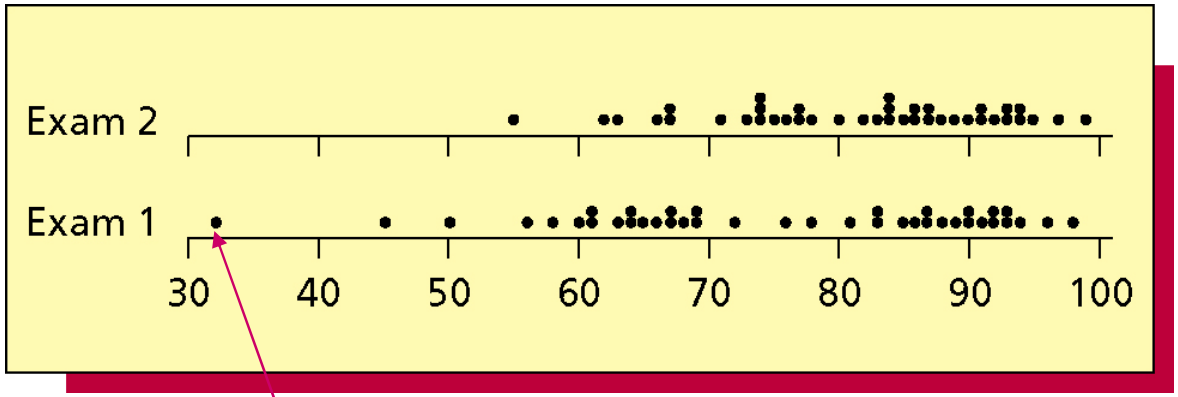


The height of the relative frequency curve over any point represents the relative proportion of values near that point

Dot Plots

On a number line, each data value is represented by a dot placed above the corresponding scale value

Scores on Exams 1 and 2



Unusually low score, so an “outlier”

Population Parameters

A *population parameter* is a number calculated from all the population measurements that describes some aspect of the population

The *population mean*, denoted μ, is a population parameter and is the average of the population measurements

Point Estimates and Sample Statistics

A *point estimate* is a one-number estimate of the value of a population parameter

A *sample statistic* is a number calculated using sample measurements that describes some aspect of the sample

- Use sample statistics as point estimates of the population parameters

The sample mean, denoted $\bar{x}$, is a sample statistic and is the average of the sample measurements

- The sample mean is a point estimate of the population mean

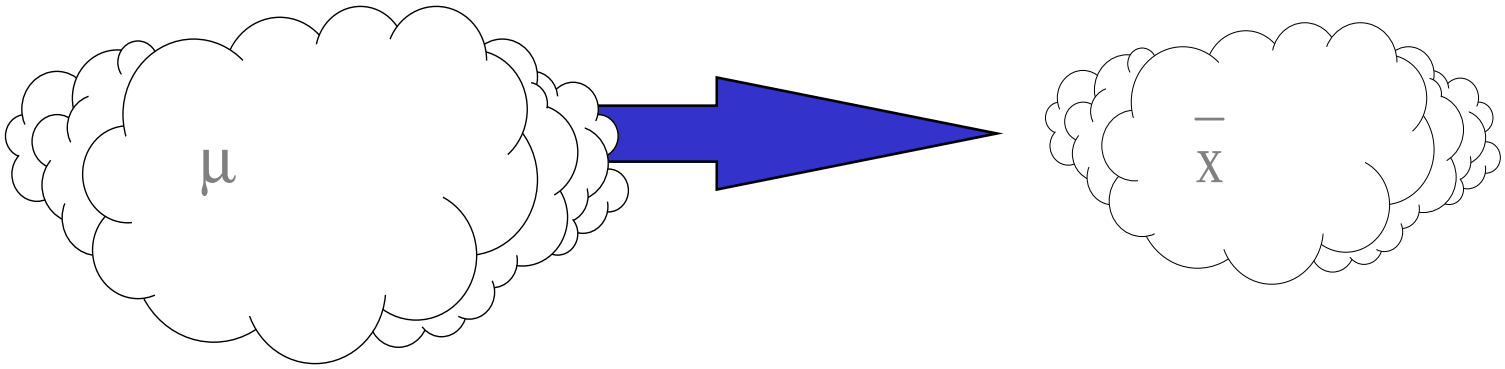
Measures of Central Tendency

***Mean*, μ**	The average or expected value
***Median*, M_d**	The value of the middle point of the ordered measurements
***Mode*, M_o**	The most frequent value (with the largest frequency)

The Mean

Population $X_1, X_2, \ldots, X_N$

Sample $x_1, x_2, \ldots, x_n$



Population Mean

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Sample Mean

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

The Sample Mean

For a sample of size n, the *sample mean* is defined as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

and is a point estimate of the population mean μ

- It is the value to expect, on average and in the long run

Example: Car Mileage Case

Example 2.5: Sample mean for first five car mileages from Table 2.1

30.8, 31.7, 30.1, 31.6, 32.1

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

$$\bar{x} = \frac{30.8 + 31.7 + 30.1 + 31.6 + 32.1}{5} = \frac{156.3}{5} = 31.26$$

The Median

The population or sample *median* M_d is a value such that 50% of all measurements, after having been arranged in numerical order, lie above (or below) it

The median M_d is found as follows:

1. If the number of measurements is odd, the median is the middlemost measurement in the ordered values

2. If the number of measurements is even, the median is the average of the two middlemost measurements in the ordered values

Example: Sample Median

Example 2.6: Internist's Yearly Salaries (x $1000)

127 132 138 141 144 146 **152** 154 165 171 177 192 241

(Note that the values are in ascending numerical order from left to right)

Because n = 13 (odd,) then the median is the middlemost or 7^{th} value of the ordered data, so

$$M_d = 152$$

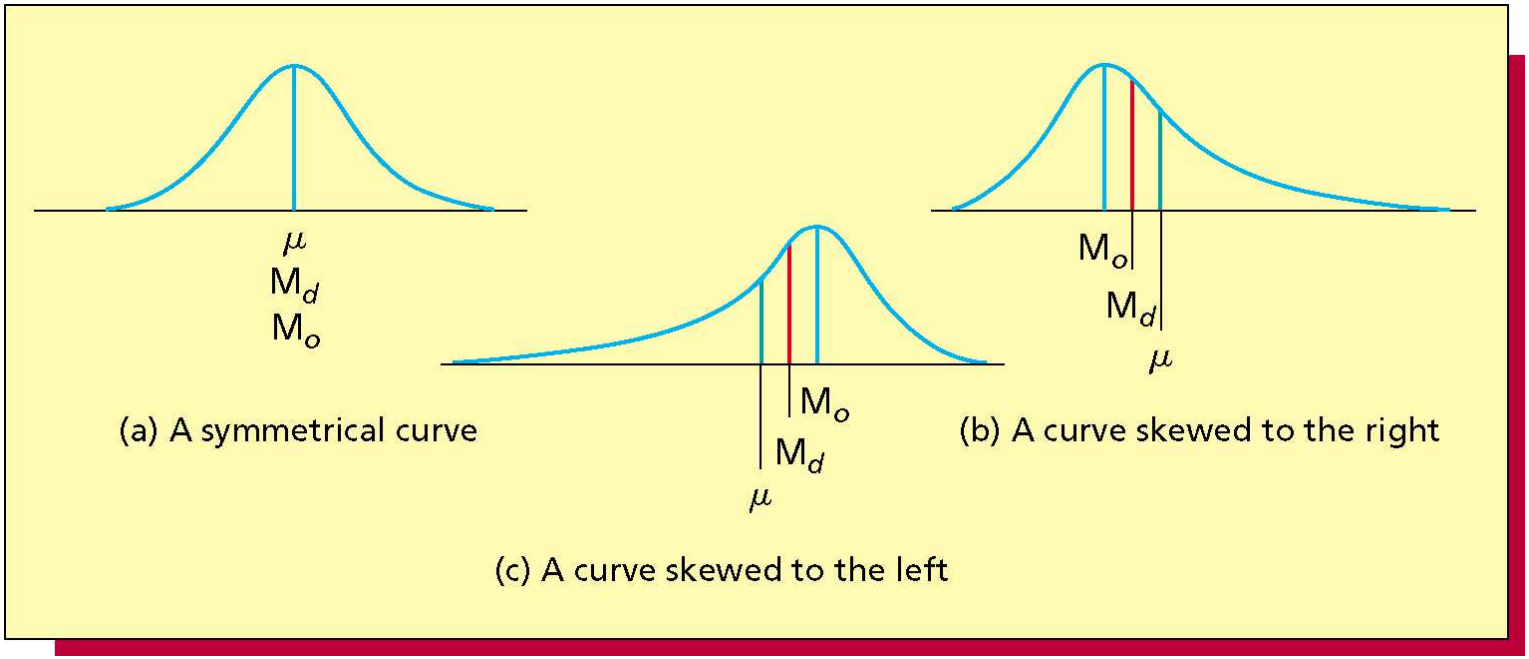
- An annual salary of $180,000 is in the high end, well above the median salary of $152,000
  - In fact, $180,000 is a very high and competitive salary

The Mode

The *mode* M_o of a population or sample of measurements is the measurement that occurs most frequently

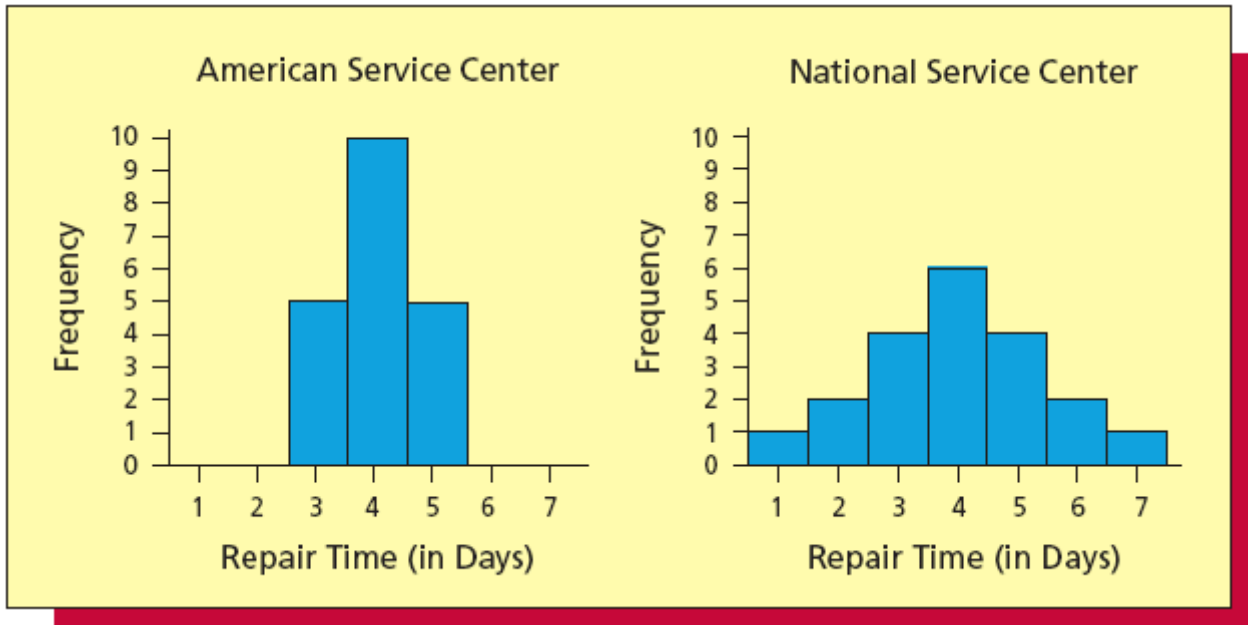
- Modes are the values that are observed "most typically"
- Sometimes higher frequencies at two or more values
  - If there are two modes, the data is bimodal
  - If more than two modes, the data is multimodal
- When data are in classes, the class with the highest frequency is the *modal class*
  - The tallest box in the histogram

Relationships Among Mean, Median and Mode



Central Tendency Not Enough

- Knowing the measures of central tendency is not enough
- Both of the distributions shown below have identical measures of central tendency



Measures of Variation

Range

Largest (maximal) minus the smallest (minimal) measurement

Variance

The average of the squared deviations of all the population measurements from the population mean

Standard Deviation

The square root of the variance

The Range

Range = Largest measurement – Smallest measurement

The range measures the interval spanned by all the data

Example:

Internist's Salaries (in thousands of dollars)

127 132 138 141 144 146 152 154 165 171 177 192 241

Range = 241 - 127 = 114 ($114,000)

Variance

For a population of size N, the *population variance* σ^2 is defined as

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}$$

For a sample of size n, the *sample variance* s^2 is defined as

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

and is a point estimate for σ^2

The Standard Deviation

Population *standard deviation*, σ: $\sigma = \sqrt{\sigma^2}$

Sample *standard deviation*, s: $s = \sqrt{s^2}$

Example: Population Variance
and Standard Deviation

Population of profit margins for five big American companies:

8%, 10%, 15%, 12%, 5%

$$\mu = \frac{8+10+15+12+5}{5} = \frac{50}{5} = 10\,\%$$

$$\sigma^2 = \frac{(8-10)^2 + (10-10)^2 + (15-10)^2 + (12-10)^2 + (5-10)^2}{5}$$

$$= \frac{(-2)^2 + 0^2 + 5^2 + 2^2 + (-5)^2}{5}$$

$$= \frac{4+0+25+4+25}{5} = \frac{58}{5} = 11.6$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{11.6} = 3.406\,\%$$

Example: Sample Variance
and Standard Deviation

Example 2.12: Sample variance and standard deviation for first five car mileages from Table 2.1

30.8, 31.7, 30.1, 31.6, 32.1 so $\bar{x} = 31.26$

$$s^2 = \frac{\sum_{i=1}^{5}(x_i - \bar{x})^2}{5-1}$$

$$= \frac{(30.8-31.26)^2 + (31.7-31.26)^2 + (30.1-31.26)^2 + (31.6-31.26)^2 + (32.1-31.26)^2}{4}$$

$$s^2 = 2.572 \div 4 = 0.643$$

$$s = \sqrt{s^2} = \sqrt{.643} = 0.8019$$

The Empirical Rule for
Normal Populations

If a population has mean μ and standard deviation σ and is described by a normal curve, then

68.26% of the population measurements lie within one standard deviation of the mean: $[\mu-\sigma, \mu+\sigma]$

95.44% of the population measurements lie within two standard deviations of the mean: $[\mu-2\sigma, \mu+2\sigma]$

99.73% of the population measurements lie within three standard deviations of the mean: $[\mu-3\sigma, \mu+3\sigma]$

Chebyshev's Theorem

Let μ and σ be a population's mean and standard deviation, then for any value $k > 1$,

At least $100(1 - 1/k^2)\%$ of the population measurements lie in the interval:

$$[\mu - k\sigma, \mu + k\sigma]$$

Use only for non-mound distributions

- Skewed
  - But not extremely skewed
- Bimodal

z-Scores

- For any *x* in a population or sample, the associated *z*-score is

$$z = \frac{x - \text{mean}}{\text{Standard deviation}}$$

- The *z*-score is the number of standard deviations that *x* is from the mean
  - A positive *z*-score is for *x* above (greater than) the mean
  - A negative *z*-score is for *x* below (less than) the mean

Coefficient of Variation

- Measures the size of the standard deviation relative to the size of the mean
- Coefficient of Variation = $\frac{\text{Standard deviation}}{\text{mean}} \times 100\%$
- Used to:
  - Compare the relative variabilities of values about the mean
  - Compare the relative variability of populations or samples with different means and different standard deviations
  - Measure risk

Percentiles and Quartiles

For a set of measurements arranged in increasing order, the p^{th} percentile is a value such that p percent of the measurements fall at or below the value and (100-p) percent of the measurements fall at or above the value

The *first quartile* Q_1 is the 25th percentile

The *second quartile* (or median) M_d is the 50th percentile

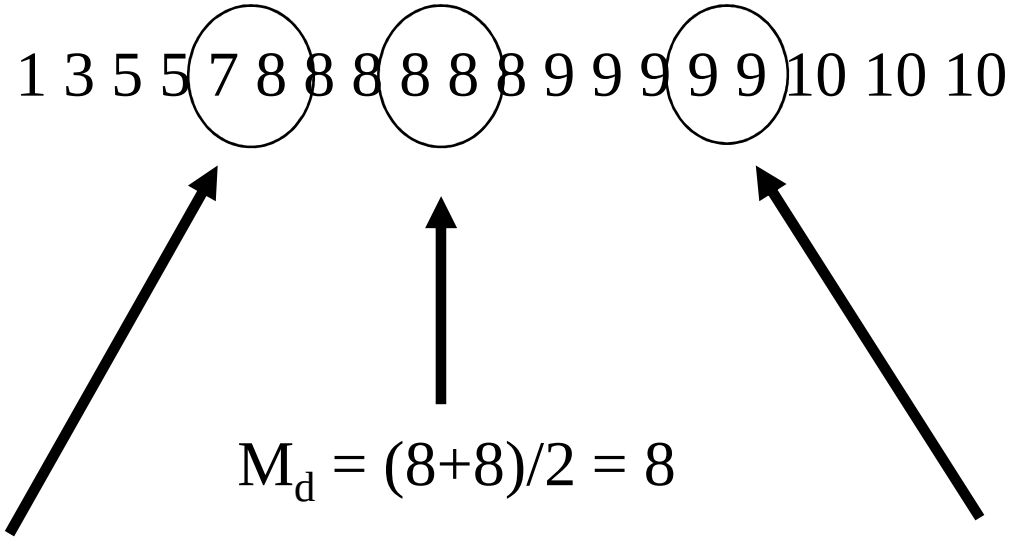
The *third quartile* Q_3 is the 75th percentile

The *interquartile range* IQR is $Q_3 - Q_1$

Example: Quartiles

20 customer satisfaction ratings:

1 3 5 5 7 8 8 8 8 8 8 9 9 9 9 9 10 10 10 10



$M_d = (8+8)/2 = 8$

$Q_1 = (7+8)/2 = 7.5$

$Q_3 = (9+9)/2 = 9$

$IQR = Q_3 - Q_1 = 9 - 7.5 = 1.5$

Box-and-Whiskers Plots

- The box plots the:
  - first quartile, Q_1
  - median, M_d
  - third quartile, Q_3
  - inner fences, located 1.5×IQR away from the quartiles:
    - = $Q_1 - (1.5 \times IQR)$
    - = $Q_3 + (1.5 \times IQR)$
  - outer fences, located 3×IQR away from the quartiles:
    - = $Q_1 - (3 \times IQR)$
    - = $Q_3 + (3 \times IQR)$

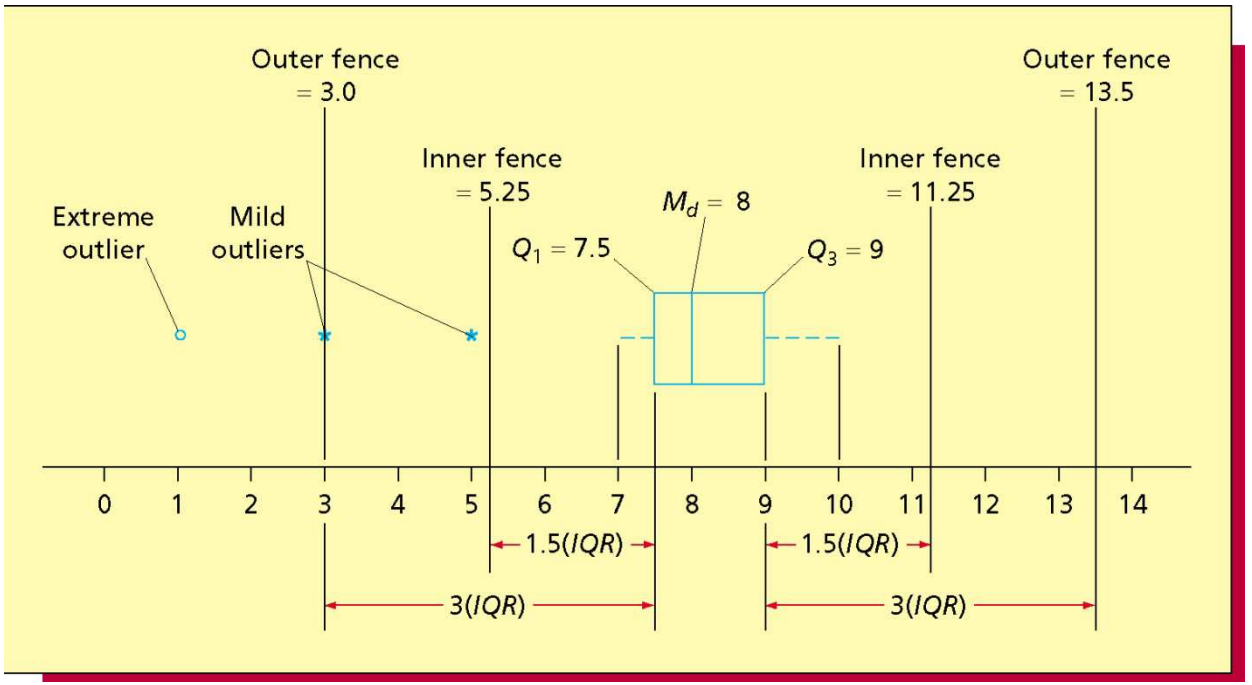
Box-and-Whiskers Plots Continued

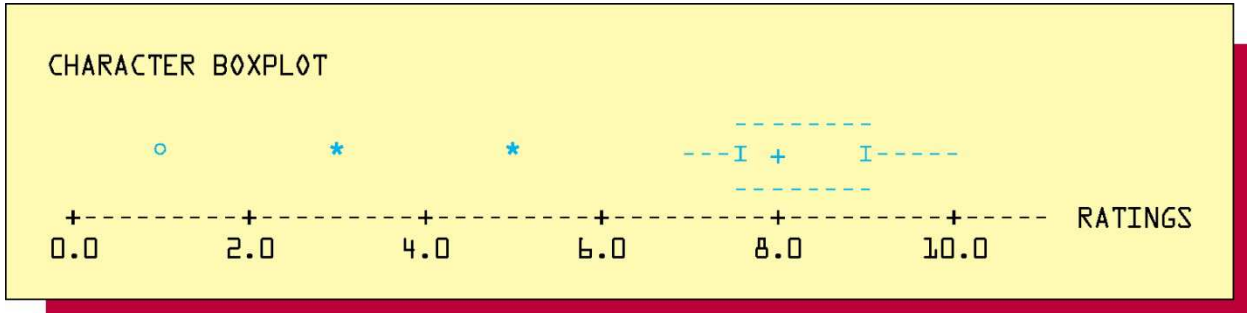
- The “whiskers” are dashed lines that plot the range of the data
  - A dashed line drawn from the box below Q_1 down to the smallest measurement
  - Another dashed line drawn from the box above Q_3 up to the largest measurement
- Note: Q_1, M_d, Q_3, the smallest value, and the largest value are sometimes referred to as the five number summary

Outliers

- Outliers are measurements that are very different from most of the other measurements
  - They are either very much larger or very much smaller than most of the other measurements
- Outliers lie beyond the fences of the box-and-whiskers plot
  - Measurements between the inner and outer fences are **mild** outliers
  - Measurements beyond the outer fences are **severe** outliers

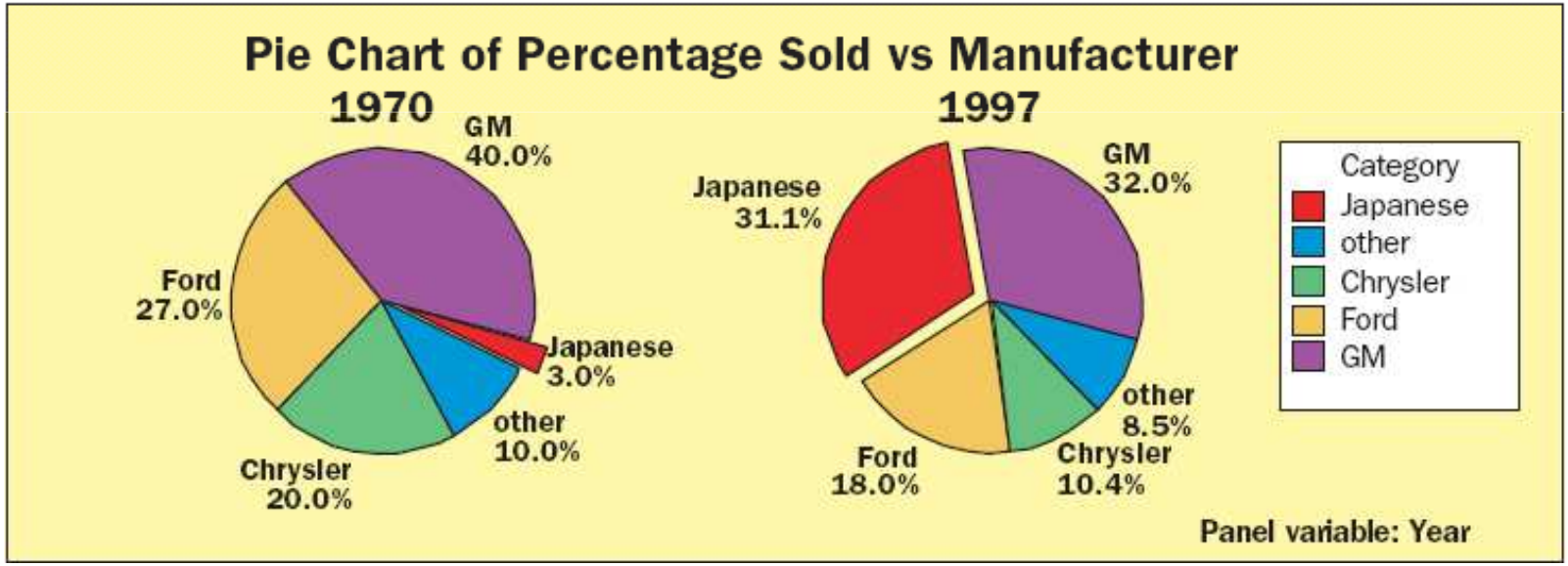
Box-and-Whiskers Plots





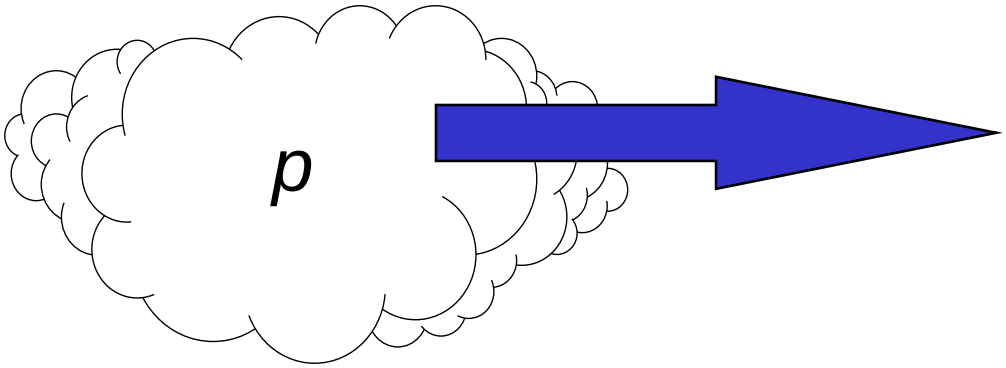
Describing Qualitative Data

Pie charts of the proportion (as percent) of all cars sold in the United States by different manufacturers, 1970 versus 1997



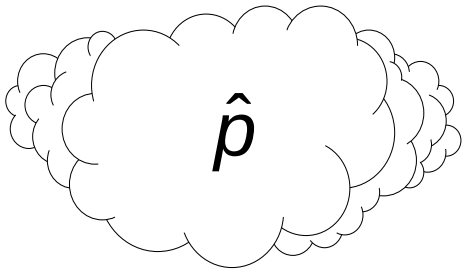
Population and Sample Proportions

Population $X_1, X_2, \ldots, X_N$



Population Proportion

Sample $x_1, x_2, \ldots, x_n$



Sample Proportion

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$\hat{p}$ is the point estimate of p

Example: Sample Proportion

Example 2.16: Marketing Ethics Case

117 out of 205 marketing researchers disapproved of action taken in a hypothetical scenario

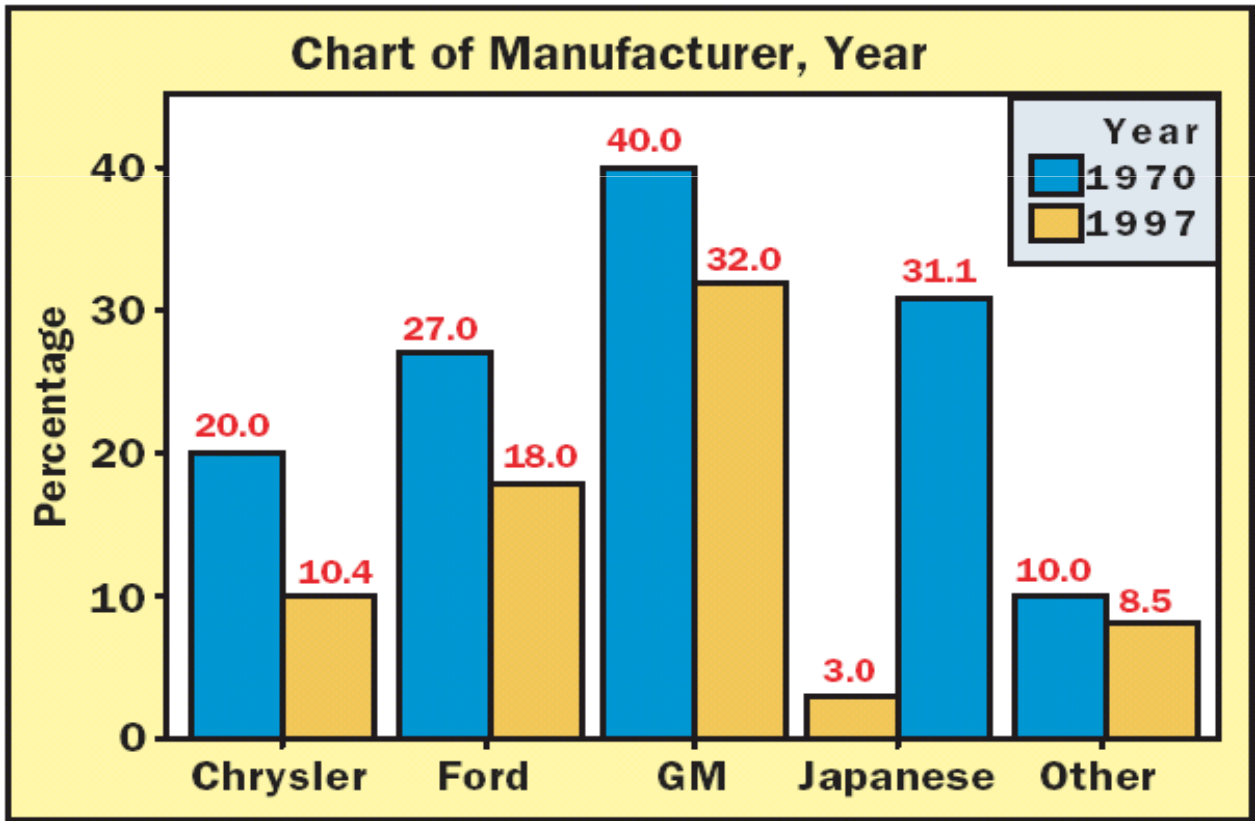
X = 117, number of researches who disapprove

n = 205, number of researchers surveyed

Sample Proportion: $\hat{p} = \frac{X}{n} = \frac{117}{205} = 0.57$

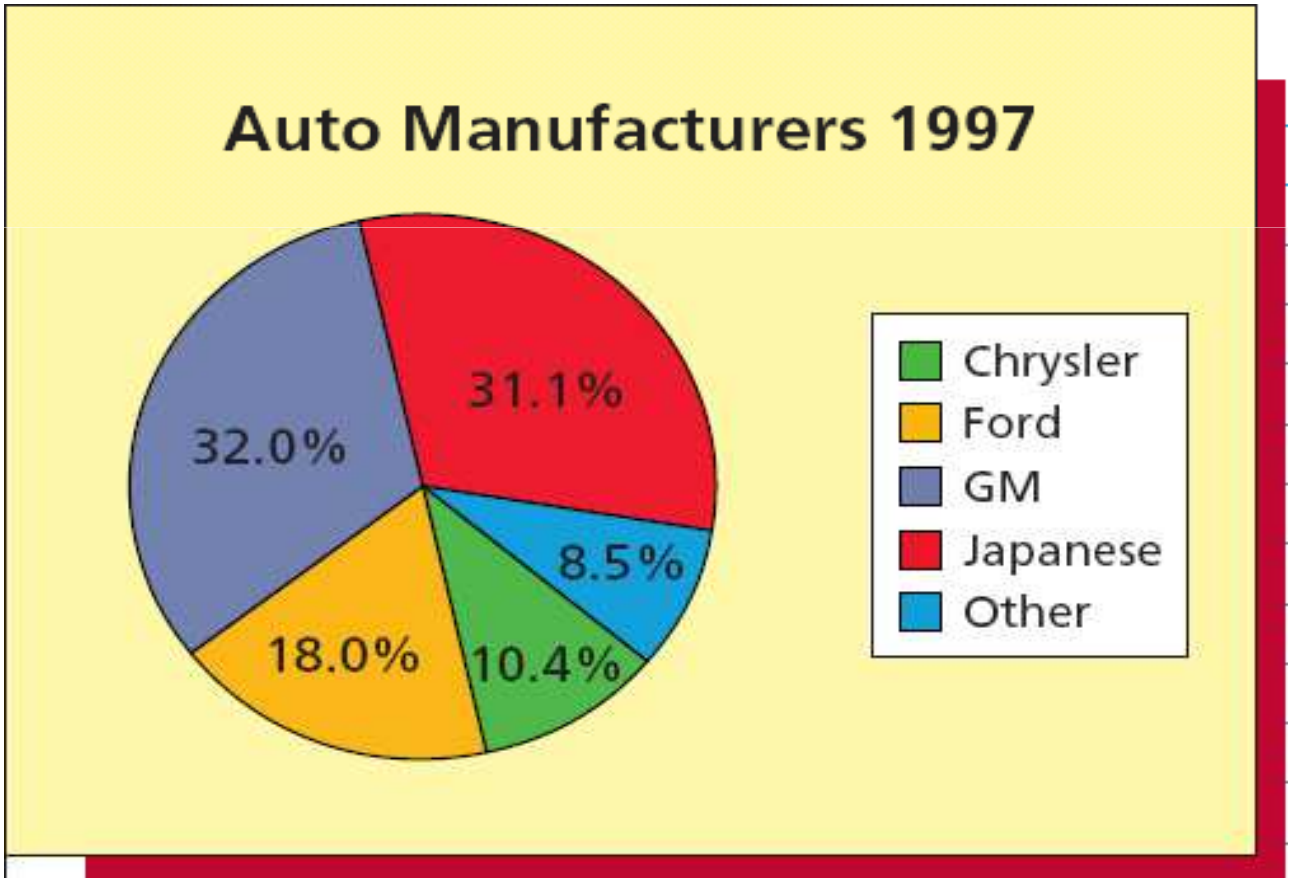
Bar Chart

Percentage of Automobiles Sold by Manufacturer, 1970 versus 1997



Pie Chart

Percentage of Automobiles Sold by Manufacturer,1997

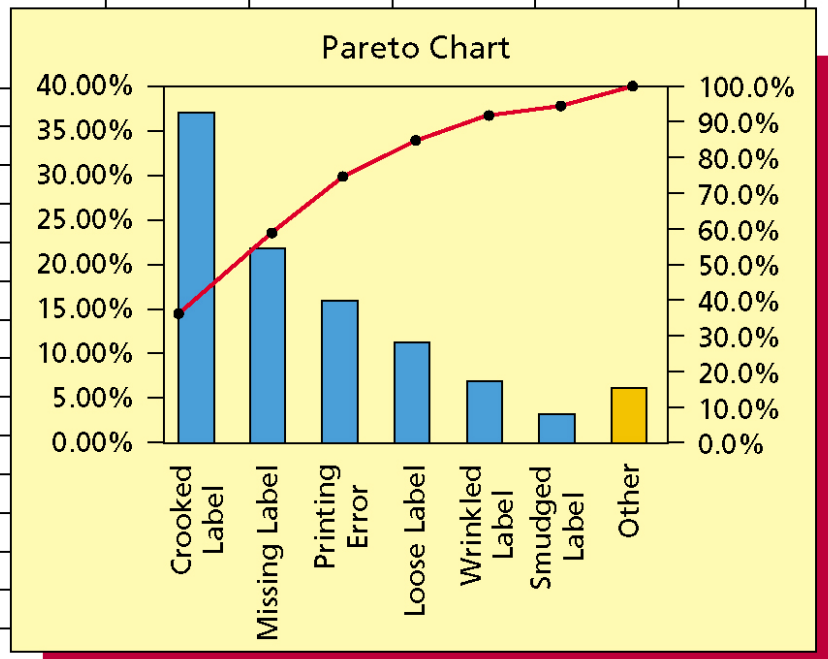


Pareto Chart

Pareto Principle: The “vital few” versus the “trivial many”

Pareto Chart of Labeling Defects

	A	B	C	D
1	Frequency Table			
3	Type of Defect	Freq.	Percent	Cum. Percent
4	Crooked Label	78	36.97%	37.0%
5	Missing Label	45	21.33%	58.3%
6	Printing Error	33	15.64%	73.9%
7	Loose Label	23	10.90%	84.8%
8	Wrinkled Label	14	6.64%	91.5%
9	Smudged Label	6	2.84%	94.3%
10	Other	12	5.69%	100.0%
11	*Total*	211	100%	

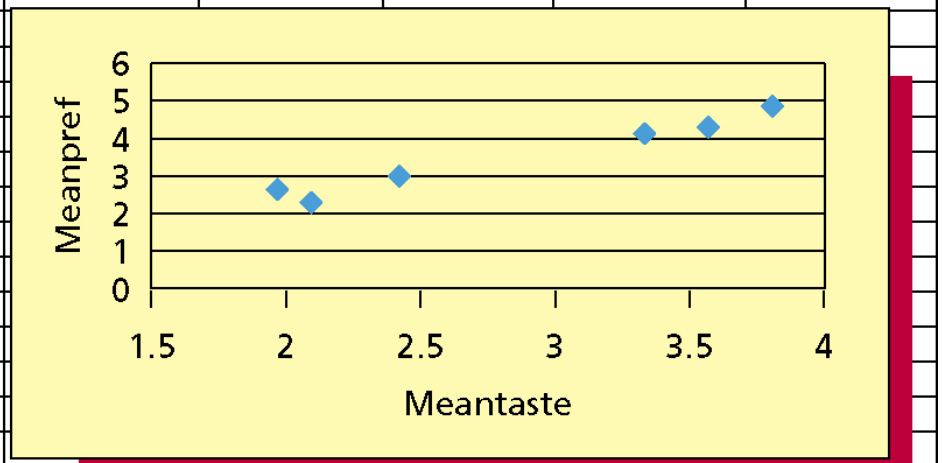


Scatter Plots

Visualize the data to see patterns, especially “trends”

Restaurant Ratings: Mean Preference vs. Mean Taste

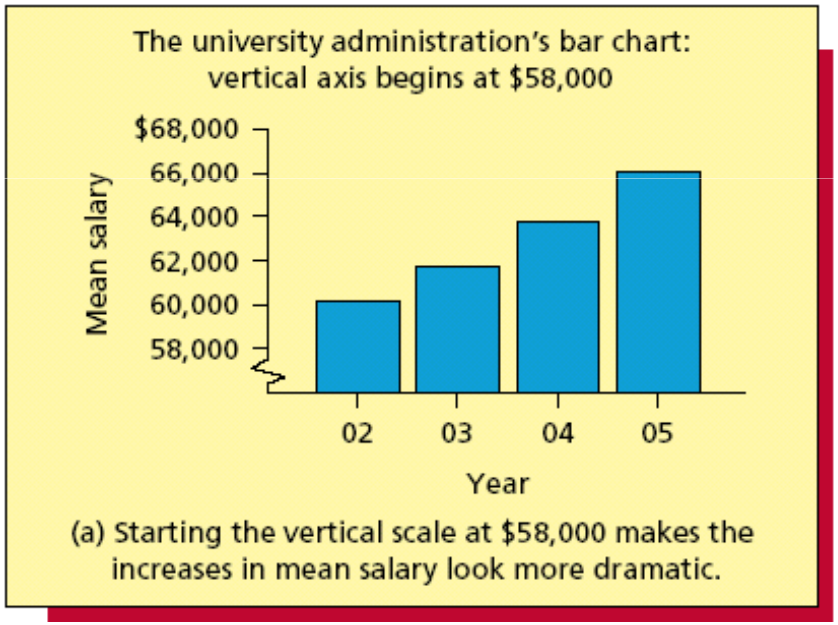
	A	B	C	D	E	F
1	Restaurant	Meantaste	Meanconv	Meanfam	Meanprice	Meanpref
2	Borden Burger	3.5659	2.7005	2.5282	2.9372	4.2552
3	Hardee’s	3.329	3.3483	2.7345	2.7513	4.0911
4	Burger King	2.4231	2.7377	2.3368	3.0761	3.0052
5	McDonald’s	2.0895	1.938	1.4619	2.4884	2.2429
6	Wendy’s	1.9661	2.892	2.3376	4.0814	2.5351
7	White Castle	3.8061	3.7242	2.6515	1.708	4.7812
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						



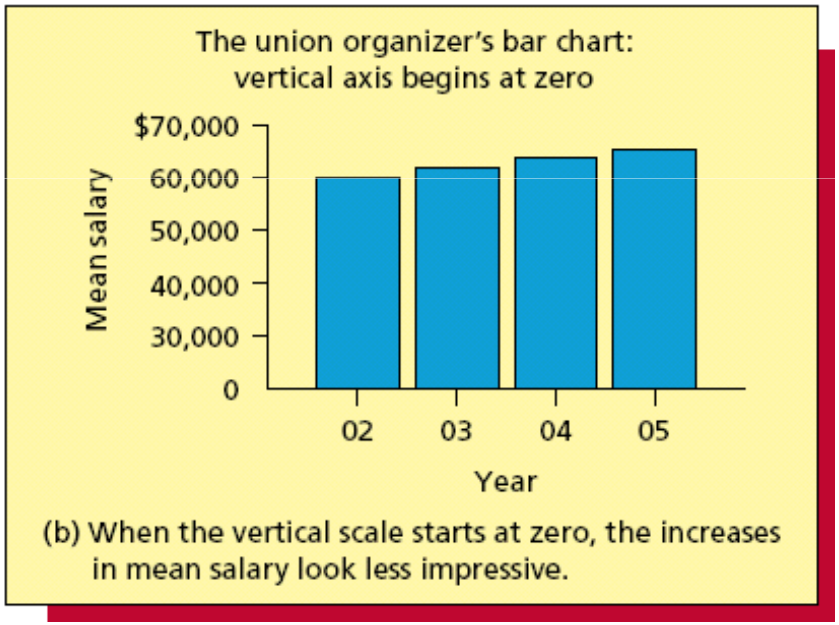
Misleading Graphs and Charts:
Scale Break

Break the vertical scale to exaggerate effect

Mean Salaries at a Major University, 2002 - 2005



(a) Starting the vertical scale at $58,000 makes the increases in mean salary look more dramatic.

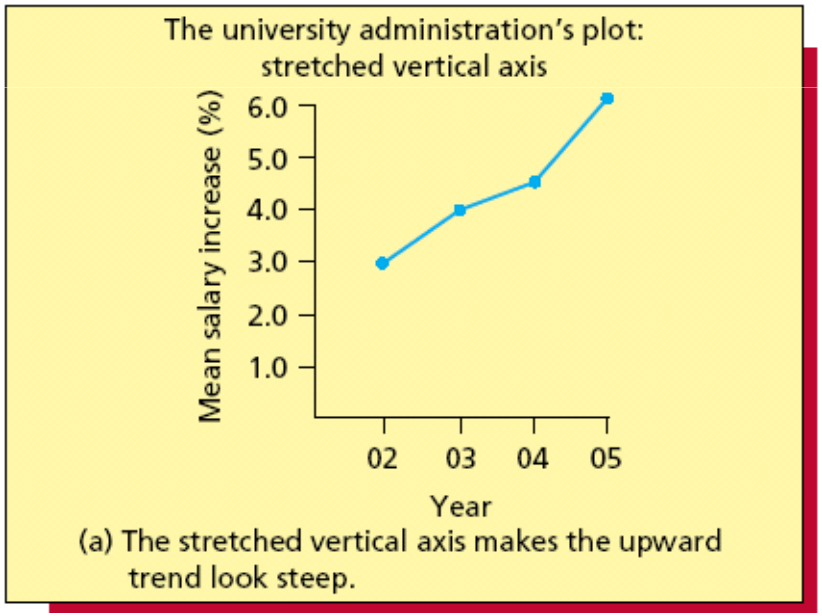


(b) When the vertical scale starts at zero, the increases in mean salary look less impressive.

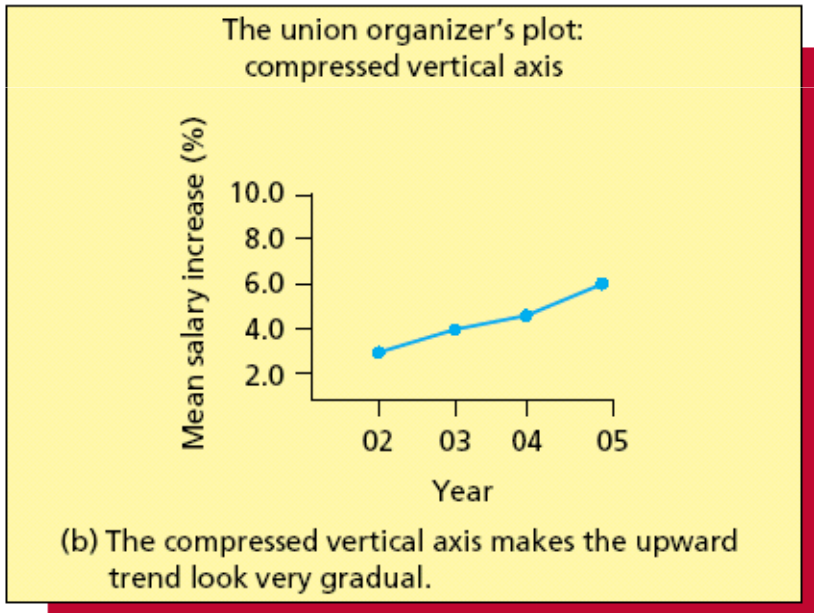
Misleading Graphs and Charts:
Horizontal Scale Effects

Compress vs. stretch the horizontal scales to exaggerate or minimize the effect

Mean Salary Increases at a Major University, 2002 - 2005



(a) The stretched vertical axis makes the upward trend look steep.



(b) The compressed vertical axis makes the upward trend look very gradual.

Weighted Means

- Sometimes, some measurements are more important than others
  - Assign numerical “weights” to the data
    - Weights measure relative importance of the value
- Calculate weighted mean as

$$\frac{\sum w_i x_i}{\sum w_i}$$

- where w_i is the weight assigned to the i^{th} measurement x_i

Descriptive Statistics for
Grouped Data (Sample)

- Data already categorized into a frequency distribution or a histogram is called grouped data
- Sample mean for grouped data:

$$\bar{x} = \frac{\sum f_i M_i}{\sum f_i} = \frac{\sum f_i M_i}{n}$$

- Sample variance for grouped data:

$$s^2 = \frac{\sum f_i (M_i - \bar{x})^2}{n - 1}$$

- where f_i is the frequency for class i
- M_i is the midpoint of class i
- $n = \Sigma f_i$ = sample size

Descriptive Statistics for
Grouped Data (Population)

- Population mean for grouped data:

$$\mu = \frac{\sum f_i M_i}{\sum f_i} = \frac{\sum f_i M_i}{N}$$

- Population variance for grouped data:

$$\sigma^2 = \frac{\sum f_i (M_i - \bar{x})^2}{N}$$

- where f_i is the frequency for class i
- M_i is the midpoint of class i
- $N = \Sigma f_i$ = population size

Geometric Mean

- For percent rates of return of an investment, use the *geometric mean* to give the correct terminal wealth at the end of the investment period
- Suppose the rates of return (expressed as decimal fractions) are R_1, R_2, …, R_n for periods 1, 2, …, n
- The mean of all these returns is the calculated as the geometric mean:

$$R_g = \sqrt[n]{(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)} - 1$$