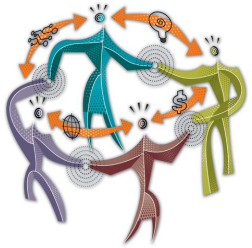
Chapter 5: Exploring Data: Distributions

Lesson Plan

For All Practical Purposes



Mathematical Literacy in Today's World, 7th ed.

- Exploring Data
- Displaying Distributions: Histograms
- Interpreting Histograms
- Displaying Distributions: Stemplots
- Describing Center: Mean and Median
- Describing the Spread: The Quartiles
- The Five-Number Summary and Boxplots
- Describing Spread: The Standard Deviation
- Normal Distributions
- The 68-95-99.7 Rule

Chapter 5: Exploring Data: Distributions

Exploring Data

- Statistics is the science of collecting, organizing, and interpreting data.
- Data
  - Numerical facts that are essential for making decisions in almost every area of life and work.
  - Spreadsheet programs are used to organize data by rows and columns.
- Exploratory data analysis
  1. Examine each variable by itself and then the relationship among them.
  2. Begin with a graph or graphs, then add numerical summaries of specific aspects of the data.

Individual – The objects described by a set of data. May be people or may also be animals or things.

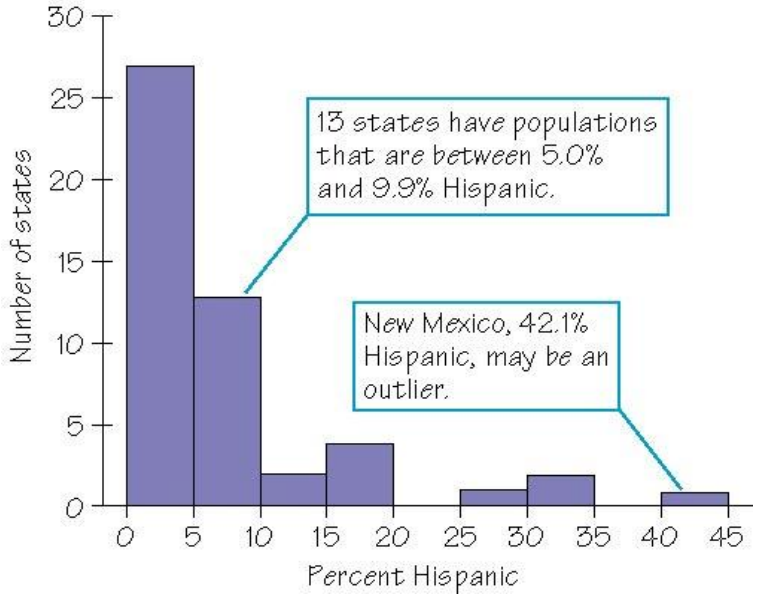
Variable – Any characteristic of an individual. A variable can take different values for different individuals.

Chapter 5: Exploring Data: Distributions

Displaying Distributions: Histograms

- Histogram
  - The graph of the distribution of outcomes (often divided into classes) for a single variable.
- Steps in Making a Histogram
  1. Choose the classes by dividing the range of data into classes of equal width *(individuals fit into one class)*.
  2. Count the individuals in each class *(this is the height of the bar)*.
  3. Draw the histogram:
     - The horizontal axis is marked off into equal class widths.
     - The vertical axis contains the scale of counts *(frequency of occurrences)* for each class.

Distribution – The pattern of outcomes of a variable; it tells us what values the variable takes and how often it takes these values.



Histogram of the percent of Hispanics among the adult residents of the states

Chapter 5: Exploring Data: Distributions

Interpreting Histograms

- Examining a Distribution
  - Overall Pattern *What does the histogram graph look like?*
    - Shape –
      - Single peak (either symmetric or skewed distribution)
        - Symmetric – The right and left sides are mirror images.
        - Skewed to the right – The right side extends much farther out.
        - Skewed to the left – the left side extends much farther out.
      - Irregular distribution of data may appear clustered and may not show a single peak *(due to more than one individual being graphed).*
    - Center – Estimated center or midpoint of the data.
    - Spread – The range of data outcomes *(minimum to maximum).*
  - Deviation *Are there any striking differences from the pattern?*
    - Outlier – An individual value that clearly falls outside the overall pattern; possibly an error or some logical explanation.

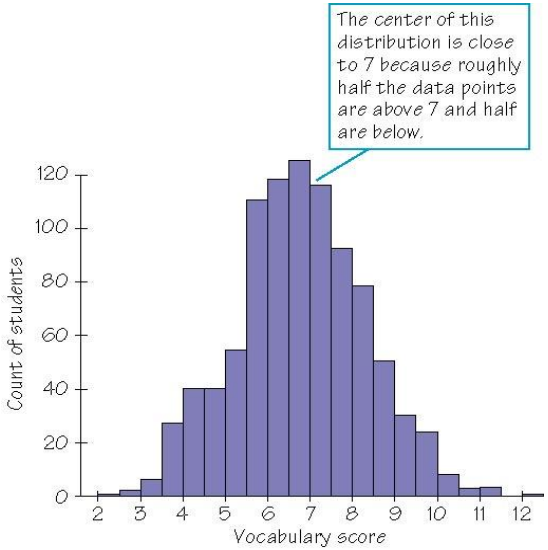
Chapter 5: Exploring Data: Distributions

Interpreting Histograms

- Examples of Distribution Patterns and Deviations
  - Regular Single-Peak Distributions

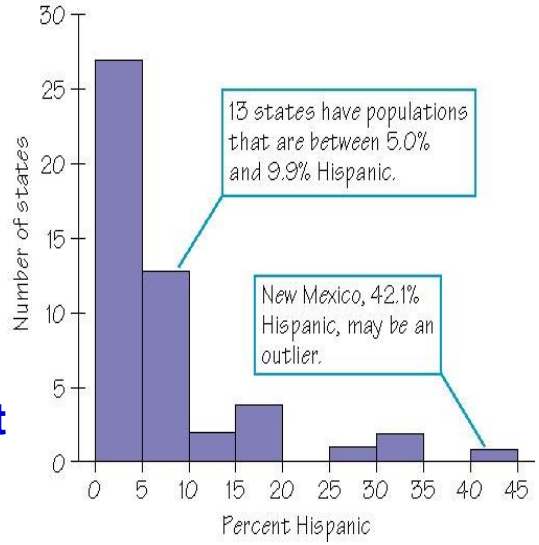
Histogram of Iowa Test of Basic Skills vocabulary scores for 947 seventh-grade students

Single Peak Symmetric →



Histogram of the percent of Hispanics among the adult residents of the states

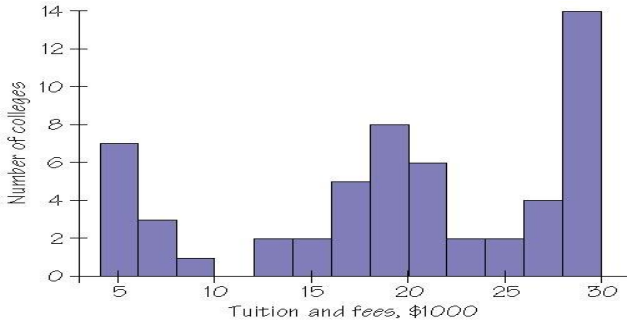
Single Peak Skewed to Right with Outlier →



  - Irregular Clustered Distributions

Histogram of the tuition and fees charged by four-year colleges in Massachusetts

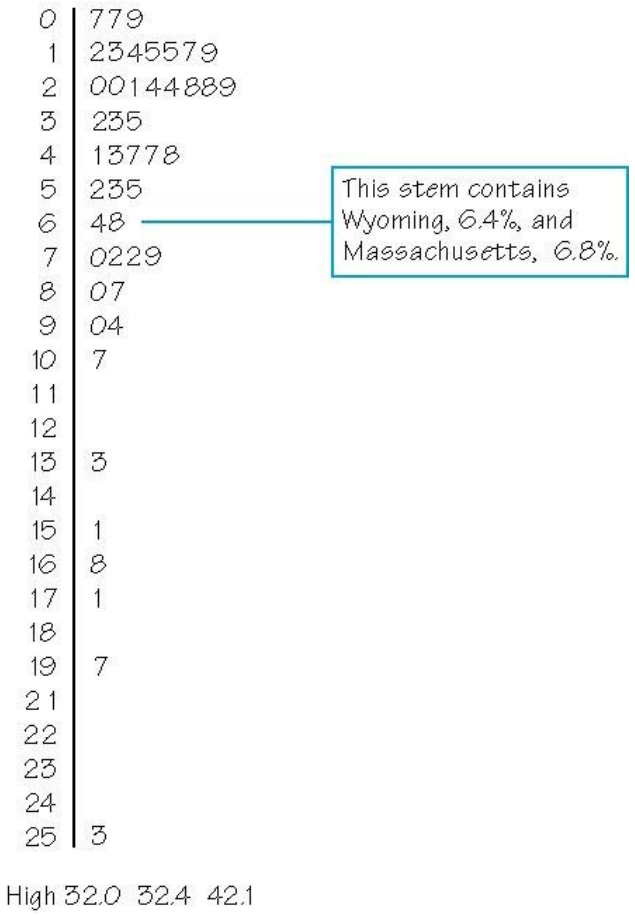
Two separate distributions, graphing two individuals (state and private schools)→



Chapter 5: Exploring Data: Distributions

Displaying Distributions: Stemplots

- **Stemplot**
  - A display of the distribution of a variable that attaches the final digits of the observation as leaves on stems made up of all but the final digit, usually for small sets of data only. Stemplots look like histograms on the side.
- **How to Make a Stemplot**
  1. Separate each observation into a stem (all but the final rightmost digit) and a leaf (the final rightmost digit).
  2. Write the stems in a vertical column, smallest at top, sequentially down to the largest value. Draw a vertical line to the right of this column.
  3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.



Stemplot of the percent of Hispanics among the adult residents of the states

Chapter 5: Exploring Data: Distributions

Describing Center: Mean and Medians

- Two Ways to Describe the Center: Mean and Median
  - Mean "average value"
    - Ordinary arithmetic average of a set of observations, *average value.*
    - To find mean of a set of observations, add their values, (x_1, x_2, …, x_n) and divide by the number of observations, *n.*
    - *x-bar,* $\bar{x} = (x_1 + x_2 + \ldots x_n)/n$
  - Median "middle value"
    - The midpoint or center of an ordered list; *middle value* of a set of observations; half fall below the median and half fall above.
    - Arrange observations in order (smallest to largest).
    - If observations are odd, location of the median is $(n + 1)/2$.
    - If observations are even, average the two center observations, find $(n + 1)/2$, then average the two values on either side of this value.

Chapter 5: Exploring Data: Distributions

Describing Center: Mean and Medians

- Finding the Mean and Median
  - Mean *average value,* $\bar{x}$ {*x-bar*}

    Mean, $\bar{x} = (x_1 + x_2 + \ldots x_n)/n$

For two-seater cars, highway, the mean is:

$\bar{x} = (24 + 28 + 28 + \ldots + 23 + 32)/21$

$\bar{x} = 518/21 = 24.7$ miles per gallon

  - Median *middle value, M*

    Arrange observations in order, then choose the middle value: 13 15 16 16 17 19 20 22 23 23 **23** 24 25 25 26 28 28 28 29 32 66.

For two-seater cars, highway, median is:

For 21 cars (odd): $(n + 1)/2 = (21 + 1)/2 = 11$

11th observations is 23, median.

Note: If Honda Insight, 66mpg, is removed there are 20 observations (even): $(n + 1)/2 = (20 + 1)/2 = 10.5$

Median = Average of 10th and 11th value $(23 + 23)/2 = 23$

Mean = Average = 22.6 mpg (mean is affected by outliers)

Fuel Economy (Miles per Gallon) for Two-Seater Cars

Model	City	Highway
Acura NSX	17	24
Audi TT Roadster	20	28
BMW Z4 Roadster	20	28
Cadillac XLR	17	25
Chevrolet Corvette	18	25
Dodge Viper	12	20
Ferrari 360 Modena	11	16
Ferrari Maranello	10	16
Ford Thunderbird	17	23
Honda Insight	60	66
Lamborghini Gallardo	9	15
Lamborghini Murcielago	9	13
Lotus Esprit	15	22
Maserati Spyder	12	17
Mazda Miata	22	28
Mercedes-Benz SL500	16	23
Mercedes-Benz SL600	13	19
Nissan 350Z	20	26
Porsche Boxster	20	29
Porsche Carrera 911	15	23
Toyota MR2	26	32

Chapter 5: Exploring Data: Distributions

Describing Spread: The Quartiles

- Include Spread and Center to Better Describe a Distribution
  - Range – States the smallest and largest observations.
  - Quartiles – The center and the middle of the top and bottom halves.

- Calculating the Quartiles

1. Arrange the observations in increasing order and locate the median ***M*** in the ordered list of observations.
   - If *n* = even, split group in half and use all the numbers.
   - If *n* = odd, circle the median and do not use it in finding quartiles.
2. The first quartile, Q_1 is the median of the observations whose position in the ordered list is to the left of the overall median *(midpoint of lower half).*
3. The third quartile, Q_3 is the median of the observations whose position in the ordered list is to the right of the overall median *(midpoint of upper half).*
   - First quartile, Q_1 is larger than 25% of the observation.
   - Third quartile, Q_3 is larger than 75% of the observations.
   - Second quartile, Q_2 is the median, and larger than 50% of observations.

The Five-Number Summary and Boxplots

The Five-Number Summary

- A summary of a distribution that gives the median, the first and third quartiles, and the largest and smallest observations.
- These five numbers offer a reasonably complete description of center and spread.
- In symbols, the five-number summary is:

Minimum Q_1 M Q_3 Maximum

Examples

Five-number summary for the highway gas mileages:

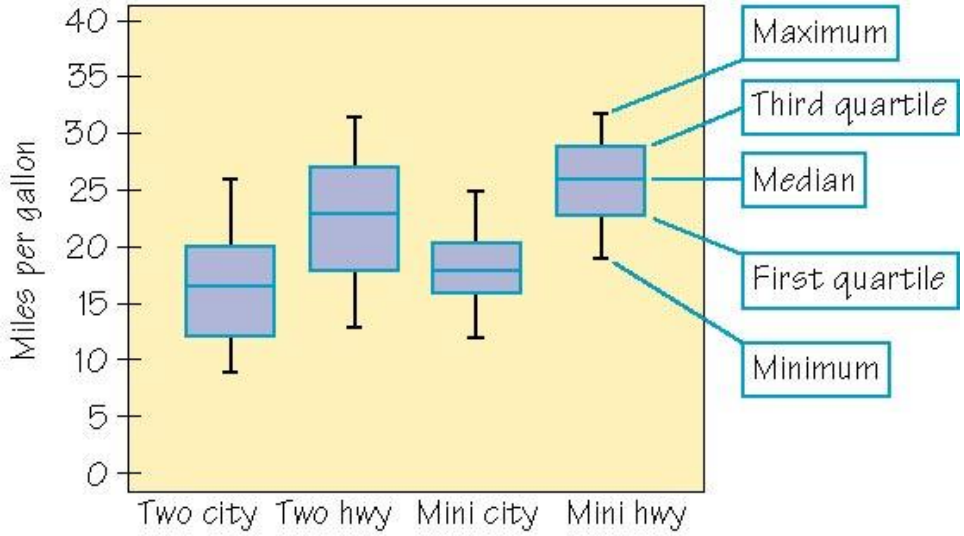
For the two-seaters: 13 18 23 27 32

For the minicompacts: 19 23 26 29 32

Boxplots

- A boxplot is a graph of the five-number summary.
- Boxplots are often used for side-by-side comparison of one or more distributions (they show less detail than histograms or stemplots).
  - A box spans the quartiles, with an interior line marking the median.
  - Lines extend out from this box to the extreme high and low observations (maximum and minimum).

Boxplots of the highway and city gas mileages for cars classified as two-seaters and as minicompacts



Chapter 5: Exploring Data: Distributions

Describing Spread: The Standard Deviation

- Standard Deviation s
  - A measure of the spread of a distribution about its mean as center. It is the square root of the variance.
- Variance s^2
  - The average squared deviation of the observations from their mean.
  - Calculated by computing the sum of the squared deviations divided by 1 less than the number of observations.

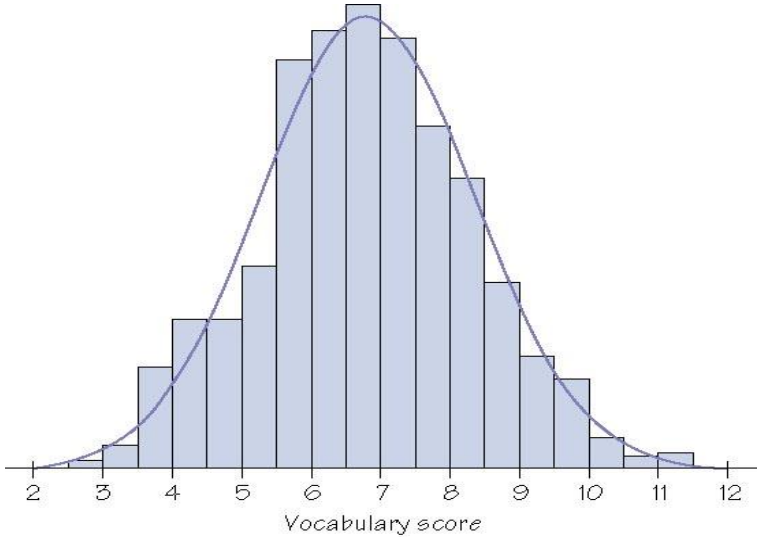
The variance, s^2 of n observation $x_1, x_2, \ldots, x_n$ is

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

The standard deviation, s, is the square root of the variance s^2.

Normal Distributions

- When the overall pattern of a large number of observations is so regular, we can describe it as a smooth curve.
- A family of distributions that describe how often a variable takes its values by areas under a curve.
- Normal curves are symmetric and bell-shaped, *smoothed-out histograms.*
- The total area under the Normal curve is exactly 1 (specific areas under the curve actually are proportions of the observations).



Histogram of the vocabulary scores of all seventh-grade students. The smooth curve shows the overall shape of the distribution.

Chapter 5: Exploring Data: Distributions

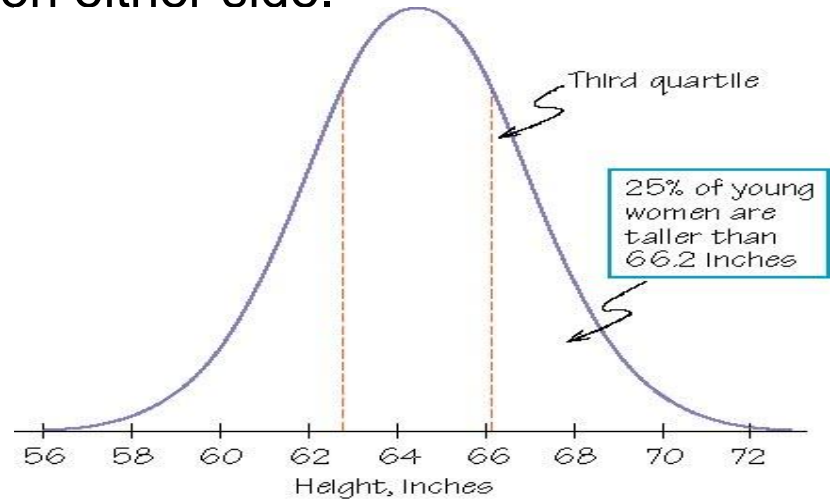
Normal Distributions

- Standard Deviation of a Normal Curve
  - The shape of a Normal distribution is completely described by two numbers, the mean and its standard deviation.
    - The mean is at the center of symmetry of the Normal curve.
    - The standard deviation is the distance from the center to the change-of-curvature points on either side.

- Calculating Quartiles
  - The first quartile of any Normal distribution is located 0.67 standard deviation below the mean.

    Q_1 = Mean − (0.67)(Stand. dev.)
  - The third quartile is 0.67 standard deviation above the mean.

    Q_3 = Mean + (0.67)(Stand. dev.)



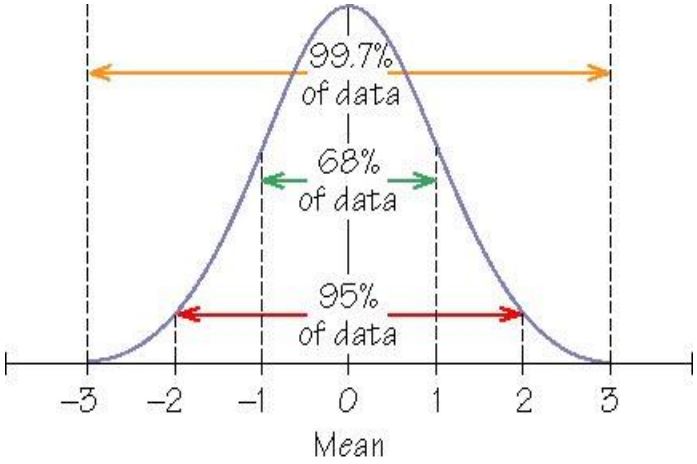
Example: Mean = 64.5, Stand. dev.= 2.5

Q_3 = 64.5 + 0.67(2.5) = 64.5 + 1.7 = 66.2

Chapter 5: Exploring Data: Distributions

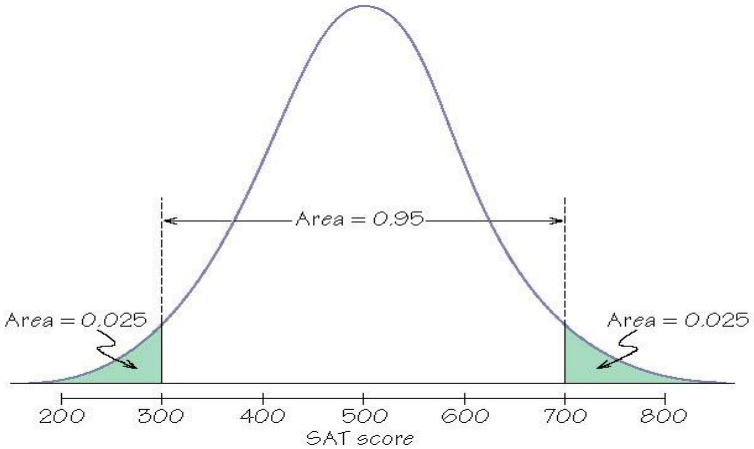
The 68-95-99.7 Rule

- Normal Distributions 68-95-99.7 Rule
  - 68% of the observations fall within 1 standard deviation of the mean.
  - 95% of the observations fall within 2 standard deviations of the mean.
  - 99.7% of the observations fall within 3 standard deviations of the mean.



- Example
  - SAT scores are close to a Normal distribution, with a mean = 500 and a standard deviation = 100.
- What percent of scores are above 700?

  Answer: Score of 700 is +2 stand. dev. Since 95% of data is between +2 and −2 stand. dev., then above 700 is in top 2.5%.



SAT scores have Normal distribution