8: Statistics

- **Statistics:** Method of collecting, organizing, analyzing, and interpreting data, as well as drawing conclusions based on the data. Methodology is divided into two main areas.
  - ***Descriptive Statistics:*** Collecting, organizing, summarizing, and presenting data.
  - ***Inferential Statistics:*** Making generalizations about and drawing conclusions from sample.
- **Population:** Data Set containing all the objects whose properties are to be described and analyzed
  - ***Sample:*** Subset or subgroup of the population.
  - ***Representative Sample:*** Exhibits characteristics typical of those possessed by the target population.

Populations and Samples

- A ***random sample*** is a sample obtained in such a way that every element in the population has an equal chance of being selected
- A group of hotel owners in a large city decide to conduct a survey among citizens of the city to discover their opinions about casino gambling.
  - Describe the population.
    - Set of all the citizens of the city.
  - Which of the following is the best way to select a random sample to find out how the city's citizens feel about casino gambling?
    - Randomly survey people who live in the oceanfront condominiums in the city.
    - Randomly select neighborhoods of the city and then randomly survey people within neighborhoods selected.

8.1: Frequency Distributions and Histograms

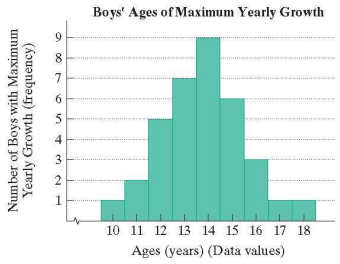
- Construct a frequency distribution for the data of the age of maximum yearly growth for 35 boys:
  12, 14, 13, 14, 16, 14, 14, 17, 13, 10, 13, 18, 12, 15, 14, 15, 15, 14, 14, 13, 15, 16, 15, 12, 13, 16, 11, 15, 12, 13, 12, 11, 13, 14, 14.
- What are some of the conclusions we can draw from this example?

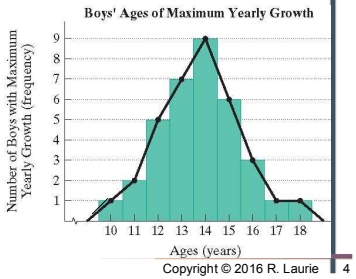
Frequency Distribution for a Boy's Age of Maximum Yearly Growth

Age of Maximum Growth	Number of Boys (Frequency)
10	1
11	2
12	5
13	7
14	9
15	6
16	3
17	1
18	1
Total:	$n = 35$

Histograms and Frequency Polygons

- **Histogram:** A bar graph with bars that touch can be used to visually display the data.
- **Frequency Polygon:** A line graph formed by connecting dots in the midpoint of each bar of the histogram.

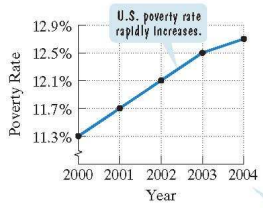




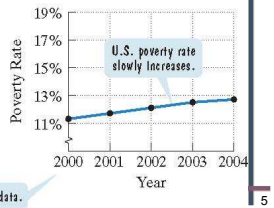
Deceptions in Visual Displays of Data

- Graphs can be used to distort the underlying data
  - The graph on the left stretches the scale on the vertical axis to create an impression of a rapidly increasing poverty rate.
  - The graph on the right compresses the scale on the vertical axis to create impression of a slowly increasing poverty rate.

Percentage of People in the United States Living below the Poverty Level, 2000-2004



Year	Poverty Rate
2000	11.3%
2001	11.7%
2002	12.1%
2003	12.5%
2004	12.7%



The graph in both figures present this data.

Constructing a Grouped Frequency Distribution

- Here are the statistics test scores for class of 40 students:

82 47 75 64 57 82 63 93
76 68 84 54 88 77 79 80
94 92 94 80 94 66 81 67
75 73 66 87 76 45 43 56
57 74 50 78 71 84 59 76

Class	Grade	Frequency
40-49	F-	3
50-59	F	6
60-69	D	6
70-79	C	11
80-89	B	9
90-100	A	5
	Total:	$n = 40$

- Group the frequencies into classes meaningful for data. Since letter grades are given based on 10-point ranges, use classes: 40–49, 50–59, 60–69, 70–79, 80–89, 90–99

8.2: Measures of Central Tendency

- Four measures of central tendancy
  - Determine the *mean* for a data set
  - Determine the *median* for a data set
  - Determine the *mode* for a data set
  - Determine the *midrange* for a data set
- *Mean*: Sum of the data items divided by the number of items.

$$\text{Mean} = \bar{x} = \frac{\sum x}{n}$$

where Σx represents the sum of all the data items and n represents the number of items.

Example: Young Male Singer Age Data Mean

U.S. Male Singers to Have a Number 1 Single with Age < 18

Artist/Year	Title	Age
Stevie Wonder, 1963	"Fingertips"	13
Donny Osmond, 1971	"Go Away Little Girl"	13
Michael Jackson, 1972	"Ben"	14
Laurie London, 1958	"He's Got the Whole World in His Hands"	14
Chris Brown, 2005	"Run It!"	15
Paul Anka, 1957	"Diana"	16
Brian Hyland, 1960	"Itsy Bitsy Teenie Weenie Yellow Polkadot Bikini"	16
Shaun Cassidy, 1977	"Da Doo Ron Ron"	17
Soulja Boy, 2007	"Crank that Soulja Boy"	17
Sean Kingston, 2007	"Beautiful Girls"	17

$$\text{Mean} = \bar{x} = \frac{\sum x}{n} = \frac{13+13+14+14+15+16+16+17+17+17}{10} = 15.2$$

Mean for a Frequency Distribution

- When many data values occur more than once and a frequency distribution is used to organize the data, we can use the following formula to calculate the mean:

$$\text{Mean} = \bar{x} = \frac{\sum x_m \cdot f}{n}$$

  - where
    - x_m represents each data value
    - f represents the frequency of that data value
    - $\Sigma x_m \cdot f$ represents the sum of all products obtained by multiplying each data value by its frequency
    - n represents the *total frequency* of the distribution

- Example: Young Male Singer Age Data Mean

$$\bar{x} = \frac{\sum x_m \cdot f}{n} = \frac{13 \cdot 2 + 14 \cdot 2 + 15 \cdot 1 + 16 \cdot 2 + 17 \cdot 3}{10} = 15.2$$

Example: Stress Frequency Distribution Mean

- The table to the right shows the students' responses to the question "How stressed have you felt in the last 2½ weeks, on a scale of 0 to 10, with 0 being not stressed at all and 10 being as stressed as possible?"
- Use the frequency distribution to find the mean of the stress-level ratings.

Stress Rating x	Frequency f
0	2
1	1
2	3
3	12
4	16
5	18
6	13
7	31
8	26
9	15
10	14

Solution: Mean for a Frequency Distribution

Solution:

1) First find xf, obtained by multiplying each data value, x, by its frequency, f as shown in third column.
2) Then use the formula to find the mean.

$$\text{Mean} = \bar{x} = \frac{\sum f \cdot x}{n} = \frac{975}{151} \approx 6.46$$

Stress Rating x	Frequency f	Data value x frequency xf
0	2	$0 \cdot 2 = 0$
1	1	$1 \cdot 1 = 1$
2	3	$2 \cdot 3 = 6$
3	12	$3 \cdot 12 = 36$
4	16	$4 \cdot 16 = 64$
5	18	$5 \cdot 18 = 90$
6	13	$6 \cdot 13 = 78$
7	31	$7 \cdot 31 = 217$
8	26	$8 \cdot 26 = 208$
9	15	$9 \cdot 15 = 135$
10	14	$10 \cdot 14 = 140$
TOTALS	151	975

The Median

- *Median* is the data item in the middle of each set of ranked, or ordered, data.
- To find the median of a group of data items,
  1. Arrange the data items in order, from smallest to largest.
  2. If the number of data items is odd, the median is the data item in the middle of the list.
  3. If the number of data items is even, the median is the mean of the two middle data items.
- Find the median for each of the following groups of data:
  84, 90, 98, 95, 88
  1. Arrange the data items in order from smallest to largest.
  2. The number of data items in the list, five, is odd.
  3. Thus, the median is the middle number.

84, 88, 90, 95, 98 The median is 90.

Median for Even Number of Data Items

68, 74, 7, 13,15, 25, 28, 59, 34, 47

Solution:

1. Arrange the data items in order from smallest to largest.
2. The number of data items in the list, ten, is even.
3. Thus, the median is the mean of the two middle numbers.

7, 13,15 ,25, 28, 34, 47, 59, 68, 74

$$\text{Median} = \frac{28+34}{2} = \frac{62}{2} = 31$$

Finding the Median using Position Formula

Listed below are the points scored per season by the 13 top point scorers in the National Football League. Find the median points scored per season for the top 13 scorers. The data items are arranged from smallest to largest:

144, 144, 145, 145, 145, 146, 147, 149, 150, 155, 161, 164, 176

Solution:

$$\text{Median Position Formula} = \frac{n+1}{2}$$

$$\text{Median Position} = \frac{13+1}{2} = 7^{\text{th}} \text{ position}$$

The median is 147

Median for a Frequency Distribution

Find the median stress-level rating.

Solution:

There are 151 data items in table so n = 151

$$\text{Median} = \frac{151+1}{2} = \frac{152}{2} = 76^{\text{th}} \text{ position}$$

Count down the frequency column in the distribution until we identify the 76th data item:

2+1+3+12+16+18+13= 65

65+31 = 96

Therefore, 76th data item = 7

Median = 7

Stress Rating x	Frequency f
0	2
1	1
2	3
3	12
4	16
5	18
6	13
7	31
8	26
9	15
10	14
TOTALS	151

Comparing the Median and the Mean

Five employees in a manufacturing plant earn salaries of: $19,700, $20,400, $21,500, $22,600 and $23,000 annually. The section manager has an annual salary of $95,000. Find the median and mean annual salary for the six people

Solution: First arrange the salaries in order.

$19,700 $20,400 $21,500 $22,600 $23,000 $95,000

$$\text{Median Position} = \frac{6+1}{2} = 3.5$$

$$\text{Median} = \frac{21,500+22,600}{2} = \frac{44,100}{2} = \$22,050$$

Since even number of data items Median is average of 3rd and 4th items

Calculating Mean

$$\text{Mean} = \frac{19,700+20,400+21,500+22,600+23,000+95,000}{6} = \frac{202,200}{6} = \$33,700$$

Which is a better measure of central tendency? Why?

The Mode

- *Mode* is the data value that occurs most often in a data set
  - If more than one data value has the highest frequency, then each of these data values is a mode
  - If no data items are repeated, then the data set has no mode
- Find the mode for the following groups of data:

7, 2, 4, 7, 8, 10

Solution:

The mode is 7.

The Midrange

- *Midrange* is found by adding the lowest and highest data values and dividing the sum by 2

$$\text{Midrange} = \frac{\text{Lowest data value} + \text{Highest data value}}{2}$$

In 2009, the New York Yankees had the greatest payroll, a record $201,449,181. The Florida Marlins were the worst paid team, with a payroll of $36,834,000. Find the midrange for the annual payroll of major league baseball teams in 2009.

$$\text{Midrange} = \frac{36{,}834{,}000 + 201{,}449{,}181}{2} = \frac{238{,}283{,}181}{2}$$

Solution: = $119,141,590.50

8.3: Measures of Dispersion

- Two of the most common measures of dispersion are *Range* and *Standard Deviation*
- *Range* is used to describe the spread of data items in a data set between highest and lowest data values
  - Range = highest data value – lowest data value
- **Example:**

  Honolulu's hottest day is 89° and its coldest day is 61°. What is its Range of Temperatures?
  - **Solution:**

    Range in temperature is: 89° − 61° = 28°

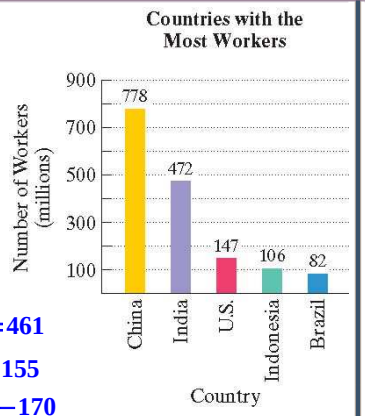
Computing Standard Deviation for a Data Set

- Computing the Standard Deviation for a Data Set requires 6 Steps

1. Find the Mean of the data items $\text{Mean} = \bar{x} = \frac{\sum x}{n}$
2. Find the *Deviation* of each data item from mean $\text{Deviation} = x - \bar{x}$
3. Square each deviation $(x - \bar{x})^2$
4. Sum the squared deviations $\sum (x - \bar{x})^2$
5. Divide the sum of step 4 by n – 1, where n represents number of data items $\frac{\sum (x - \bar{x})^2}{(n-1)}$
6. Take the square root of the quotient from step 5. This value is the *Standard Deviation* for the data set. $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$

Steps 1 & 2: Compute Mean and Deviations

This graph describes the size of the labor force size for five different countries.



The mean is first calculated to be:

$$\bar{x}=\frac{\sum x}{n}=\frac{(1585)}{5}=317$$

Deviation is calculated for each country by subtracting the mean from each data item:

$$\text{Deviation}_{\text{China}}=x-\bar{x}=778-317=461$$

$$\text{Deviation}_{\text{India}}=x-\bar{x}=472-317=155$$

$$\text{Deviation}_{\text{USA}}=x-\bar{x}=147-317=-170$$

This indicates that the labor force in the United States is 170 million workers below the mean.

Steps 3 – 6: Tabulate Deviations, Square, and Sum

Data Item	Deviation	(Deviation)²
778	778 – 317 = 461	461² = 212,521
472	472 – 317 = 155	155² = 24,025
147	147 – 317 = – 170	(–170)² = 28,900
106	106 – 317 = – 211	(–211)² = 44,521
82	82 – 317 = – 235	(– 235)²= 55,225
SUM=	0	365,192

Step 5: Divide the SUM in step 4 by $n-1$, where n represents the number of data items:

$$\frac{\sum(x-\bar{x})^2}{(n-1)}=\frac{365{,}192}{5-1}=\frac{365{,}192}{4}=91{,}298$$

Step 6: The standard deviation is the square root of the quotient in the previous step.

$$s=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}=\sqrt{91{,}298}=302.16$$

The standard deviation is approximately 302.16 million workers.

Standard Deviation Example: Young Male Singer Age Data

Step 1: Calculate Mean

$$\bar{x}=\frac{\sum x\cdot f}{n}=\frac{13\cdot2+14\cdot2+15\cdot1+16\cdot2+17\cdot3}{10}=15.2$$

Step 2: Calculate each Deviation

Step 3: Square each Deviation

Step 4: Multiply each Deviation by each Frequency

Step 5: Sum last two columns

Age	D = Deviation	D²	f	D²·f
13	13-15.2 = -2.2	4.84	2	9.68
14	14-15.2 = -1.2	1.44	2	2.88
15	15-15.2 = -0.2	0.04	1	0.04
16	16-15.2 = 0.8	0.64	2	1.28
17	17-15.2 = 1.8	3.24	3	9.72
		SUM=	10	23.60

Step 6: Divide the SUM in by $n-1$:

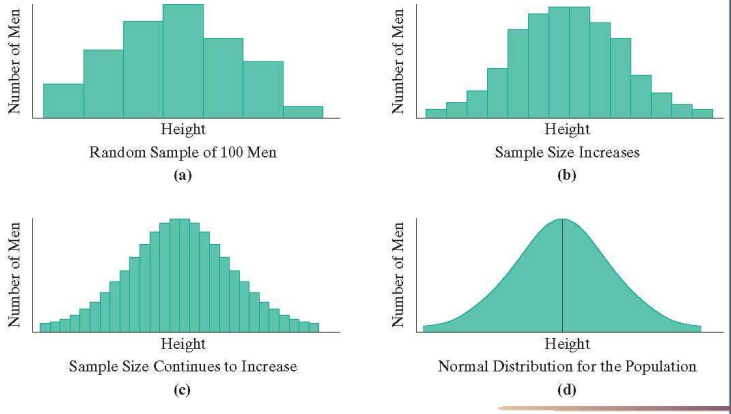
$$\frac{\sum[(x-\bar{x})^2\cdot f]}{(n-1)}=\frac{23.60}{10-1}=\frac{23.60}{9}=2.62$$

Step 7: The standard deviation is the square root of the quotient

$$s=\sqrt{\frac{\sum[(x-\bar{x})^2\cdot f]}{n-1}}=\sqrt{2.62}=1.62$$

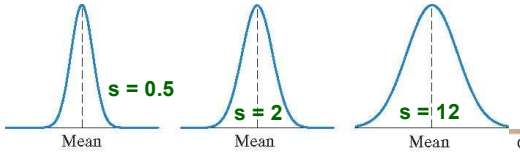
8.4: Normal Distribution and Percentiles

Normal Distribution often become bell shaped for large populations



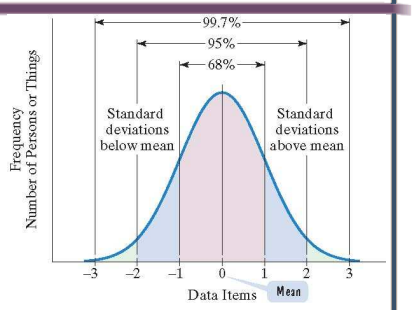
Normal Distribution

- Also called the bell curve or Gaussian distribution
- Normal distribution is bell shaped and symmetric about a vertical line through its center which is at the mean
- Mean, median and mode are all equal and located at the center of the distribution
- The shape of the normal distribution depends on the mean and the standard deviation.
- These three graphs have the same mean but different standard deviations. As the standard deviation increases, the distribution becomes more spread out.



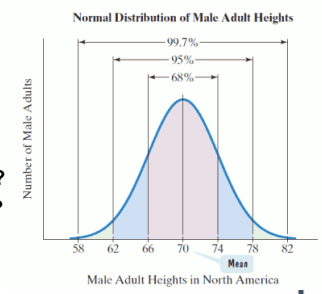
Standard Deviation and the 68-95-99.7 Rule

1. Approximately 68% of the data items fall within 1 standard deviation of the mean (in both directions).
2. Approximately 95% of the data items fall within 2 standard deviations of the mean.
3. Approximately 99.7% of the data items fall within 3 standard deviations of the mean.



Exercise: Finding Male Heights

- Male adult heights in North America are approximately normally distributed with a mean of 70 inches and a standard deviation of 4 inches.
  - What is the height of a Man that is 2 standard deviations above the mean?
  - What percentage of men are shorter?
  - What percentage of men are taller?



- *Solution:*
  - $\text{Height} = \bar{x} + 2 \cdot s = 70 + 2 \cdot 4 = 78 \text{ inches}$
  - $\%\text{Shorter} = 50\,\% + \frac{95\,\%}{2} = 50\,\% + 47.5\,\% = 97.5\,\%$
  - $\%\text{Taller} = 100\,\% - \left(50\,\% + \frac{95\,\%}{2}\right) = 1 - 97.5\,\% = 2.5\,\%$

Computing z-Scores

- A *z-Score* describes how many standard deviations a data item in a normal distribution lies above or below the mean
- The can be obtained using the formula:

$$\text{z-Score} = \frac{x - \bar{x}}{s}$$

  - Data items below mean have negative z-scores
  - Data items above mean have positive z-scores
- Utilizing a z-Score table the percentile of a data item can be determined

Exercise: Understanding z-Scores

- Intelligence quotients (IQs) on the Stanford–Binet intelligence test are normally distributed with a mean of 100 and a standard deviation of 16. What is the IQ corresponding to a z-score of −1.5?

Solution:

The negative sign in -1.5 tells us that the IQ is 1½ standard deviations below the mean.

$$\text{z-Score} = \frac{x - \bar{x}}{s}$$

$$-1.5 = \frac{x - 100}{16}$$

Score = 100 – 1.5 · 16

= 100 – 24 = **76**

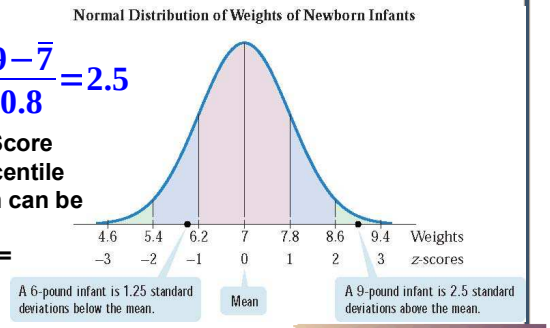
The IQ corresponding to a z-score of −1.5 is 76.

Exercise: Computing z-Scores

- The mean weight of newborn infants is 7 pounds and the standard deviation is 0.8 pound.The weights of newborn infants are normally distributed. Find the z-score for a weight of 9 pounds.
- Solution

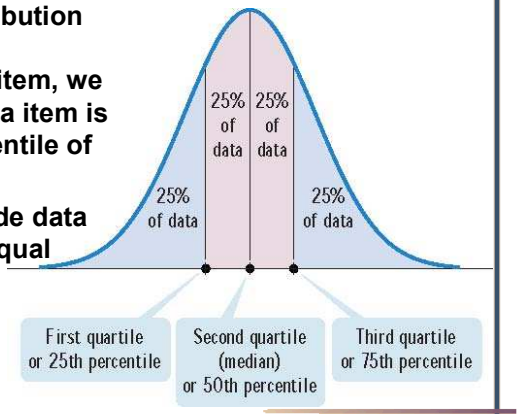
$$\text{z-Score} = \frac{9 - \bar{7}}{0.8} = 2.5$$

- Utilizing a z-Score table the percentile of a data item can be determined: 2.5 → 0.9938 = 99.38%



Percentiles and Quartiles

- *Percentiles:* If *n*% of the items in a distribution are less than a particular data item, we say that the data item is in the *n*th percentile of the distribution
- *Quartiles:* Divide data sets into four equal parts



Percentage of Data Items Less Than an Item

According to the U.S. Department of Health cholesterol levels are normally distributed. For men between 18 and 24 years, the mean is 178.1 and the standard deviation is 40.7. What percentage of men in this age range have a cholesterol level less than 239.15?

Solution:

Compute the z-score for a 239.15 cholesterol level.

$$z_{239.15} = \frac{x - \bar{x}}{s} = \frac{239.15 - 178.1}{40.7} = 1.5$$

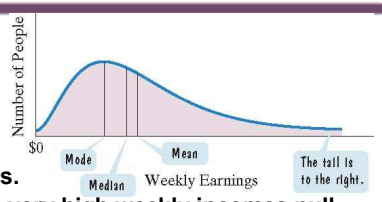
Examine a z-Score Table

z-Score = 1.5 → 93.32 %

z-score	Percentile
1.4	91.92
1.5	93.32
1.6	94.52

Skewed Distributions

This graph represents the population distribution of weekly earnings in the United States. There is no upper limit on weekly earnings. The relatively few people with very high weekly incomes pull the mean income to a value greater than the median.



- The most frequent income, the mode, occurs towards the low end of the data items.
- This is called a skewed distribution because a large number of data items are piled up at one end or the other with a "tail" at the other end.
- This graph is skewed to the right.
- For skewed distributions z-scores cannot be used to find accurate percentiles

Finding Percentage of Data Items Between 2 Items

1. Convert each given data item to a z-score:

$$\text{z-Score} = \frac{x - \bar{x}}{s}$$

2. Use z-score table to determine the percentile corresponding to each z-score in step 1
3. Subtract the lesser percentile from the greater percentile and attach a % sign

The amount of time that self-employed Americans work each week is normally distributed with a mean of 44.6 hours and a standard deviation of 14.4 hours. What percentage of self-employed individuals in the United States work between 37.4 and 80.6 hours per week?

Finding Percentage of Data Items Between 2 Items

Solution:

Step 1: Convert each given data item to a z-score.

$$z_{37.4} = \frac{x - \bar{x}}{s} = \frac{37.4 - 44.6}{14.4} = \frac{-7.2}{14.4} = -0.5 \qquad z_{80.6} = \frac{x - \bar{x}}{s} = \frac{80.6 - 44.6}{14.4} = \frac{36}{14.4} = 2.5$$

Step 2: Use the z-Score Table to find the percentile corresponding to these z-scores which is percent below.

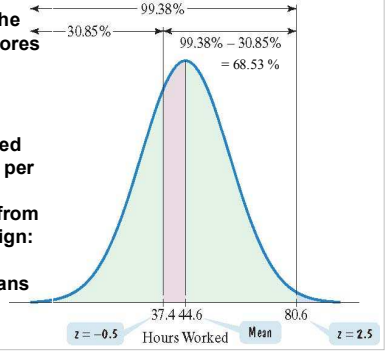
- $z_{37.4} = -0.50$ yields 30.85% percentile
- $Z_{80.6} = 2.5$ yields 99.38% percentile

That means that 99.38% of self employed Americans work fewer than 80.6 hours per week.

Step 3: Subtract the lesser percentile from the greater percentile and attach a % sign:

99.38 – 30.85 = 68.53

Thus, 68.53% of self-employed Americans work between 37.4 and 80.6 hours per week.

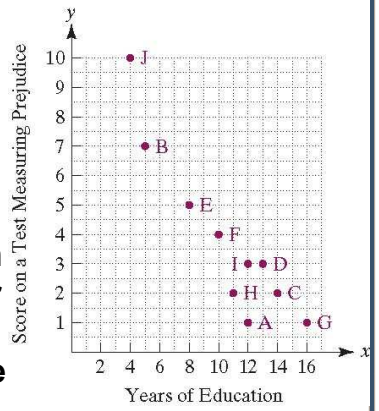


8.5: Correlation and Scatter Plots

- A *scatter plot* is a collection of data points, one data point per person or object.
  - Can be used to determine whether two quantities are related.
- *Correlation*
  - Clear relationship between two quantities
  - Determines if there is a relationship between two variables and, if so, the strength and direction of that relationship
- *Regression line* is a line that best fits the data points in a scatter plot

Scatter Plots and Correlation

- The scatter plot shows a downward trend among the data points, with some exceptions
- People with increased education tend to have a lower score on the test measuring prejudice



Correlation and Causal Connections

- Although the scatter plot shows a correlation between education and prejudice, we cannot conclude that increased education causes a person's level of prejudice to decrease.
- The correlation could be simply a coincidence.
  - Education usually involves classrooms with a variety of different kinds of people.
    - Increased exposure to diversity in the classroom might be an underlying cause.
  - Education requires people to look at new ideas and see things in different ways.
    - Thus, education causes one to be more tolerant and less prejudiced.

Correlation Coefficients

- ***Correlation coefficient***, designated by ***r***, is a measure that is used to describe the strength and direction of a relationship between variables whose data points lie on or near a line. The relationship is:
  - ***Negative correlation*** if one variable decreases while other increases. Slope of regression line is negative.
  - ***Positive correlation*** if they tend to increase or decrease together. Slope of regression line is positive.
  - ***No Correlation*** the points are a random scatter
- Correlation Level
  - ***Perfect correlation*** if all points lie on the regression line
  - ***Strong correlation*** if all points lie close to the regression line
  - ***Weak correlation*** if all points are spread widely but a regression line is observable for data modeling purposes

Scatter Plots and Correlation Coefficients

- Formula to calculate correlation coefficient, r:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \cdot \sqrt{n(\sum y^2) - (\sum y)^2}}$$