Junior Achievement

$ave, USA

High School Classroom Session

Statistics and Life Choices

Classroom Lesson

Students should analyze a variety of factors before making a decision on career choices. In this lesson, students will determine differences in center and spread and overall shape of the data. They may also determine any statistically significant difference between groups.

Suppose the two data sets represent random samples of home-selling prices in two cities. The chosen variable, home prices, is something to be considered when moving to a particular area to pursue a career.

Ask students to analyze the data below.

Sample Analysis of Home Prices

City A

83000, 83600, 93300, 95000, 120000, 121000, 123000, 150400, 157100, 171400, 182400, 184000, 188600, 193100, 199200, 214300, 222400, 226000, 236100, 247450

City B

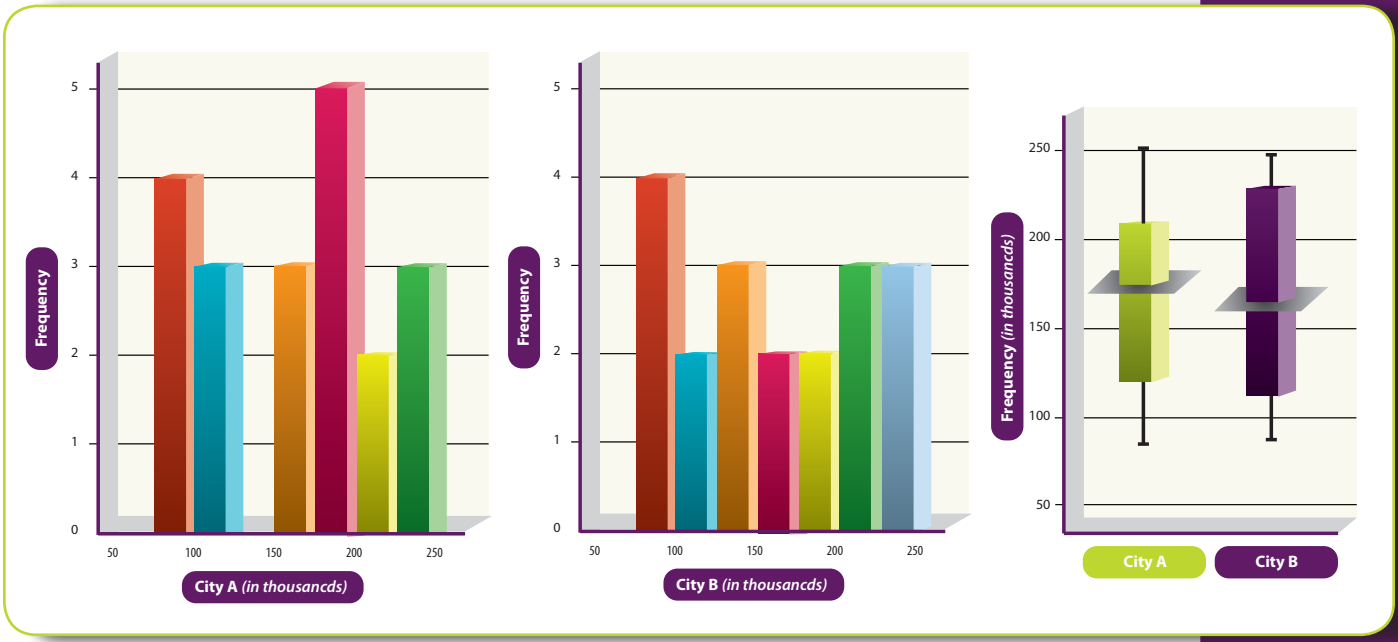
86000, 91700, 94500, 95600, 118900, 120800, 137700, 140600, 146000, 162100, 165800, 182000, 195200, 210000, 216000, 222700, 237300, 239000, 240400, 243800

We will…

1) Create a boxplot and histograms to represent the data.
2) Analyze and describe the shape of the data.
3) Discuss measures of center and spread.
4) Describe which measure of center is most appropriate and provide justification.
5) Determine if a statistically significant difference exists between the groups.

Sample Solutions are shown below:

1)



2) Neither distribution of home prices appears to be normal (mirror the normal, or bell, curve). Also, neither distribution appears to be skewed in one direction.

3) City A has a larger median than City B. The deviations near the center appear to be similar. However, City B appears to have a higher standard deviation, because there is a greater interquartile range. The outliers are similar for both cities.

The descriptive summary is as follows:

Descriptive Statistics

Data Set	Mean	Median	Mode	Range	Standard Deviation
City A	164,567.50	176,900	None	164,450	53,266
City B	167,305	163,950	None	157,800	55,648

Notice that City A has a smaller mean than City B. The mean for City A is smaller because it has fewer homes, selling for around or more than $200,000. The standard deviations are similar, as discussed earlier. In summary, City A has a higher median price, but the standard deviation is smaller.

4) Due to the large range in prices, the median is the most appropriate measure of center. It gives a more realistic estimate of the central home price.

5) Two-sample t-tests results are shown below:

$p > 0.05, t \approx -0.16$

The p-value is greater than 0.05, the common threshold for the level of significance. Because it is greater, the null hypothesis of no difference should not be rejected. Thus, we can declare that there is no significant difference between the home prices for the two cities. Students may conclude that this factor is not a concern, because there really isn't a difference in home-selling prices for the cities. They may then look at other factors.

Give each student the Statistics and Life Choices Worksheet. Ask students to analyze data that show reported job satisfaction from a Likert scale survey. Students should analyze the data for the two groups and determine any significant differences between the groups. The activity will guide students through the process of analyzing relevant data and making inferences—a process that will help them analyze other important criteria while making a career decision.

Summary and Review

After students have analyzed the data, provide a time for discussion and review. During this time, students may ask any questions about the shape and spread of data, analyses, interpretations, etc. Encourage students to compare the shapes of data as represented by a boxplot and histogram.

How do they compare? How are they different? Is the measure of center easier to discern from one representation than the other? Which one best represents the spread of the data?

You may also take this time to go through one or more t-test calculations. A random-number generator may be used to create data, while students may choose to come to the front of the class to show how to perform a one-sample t-test and two-sample t-test, using a graphing calculator. This time should be interactive.

Statistics and Life Choices

Worksheet

Job satisfaction varies from career to career and from person to person. An individual's personality must be a good fit for a particular career. Factors that influence job satisfaction may include salary, vacation time, hours, bonuses, work environment, and stress. Brainstorm some factors you associate with a high level of job satisfaction.

In this activity, you will analyze data from the Likert scale that show reported satisfaction for two careers. (A score of 1 shows lowest job satisfaction; a score of 10 shows highest job satisfaction.) Analyze the data for the two groups and determine any significant differences.

Directions: Using the table below

1) Create a boxplot and histograms to represent the data.
2) Analyze and describe the shape of the data.
3) Discuss measures of center and spread.
4) Describe which measure of center is most appropriate and provide justification.
5) Determine if a statistically significant difference exists between the groups.

Possible correct responses are shown below:

Sample Analysis

Career A

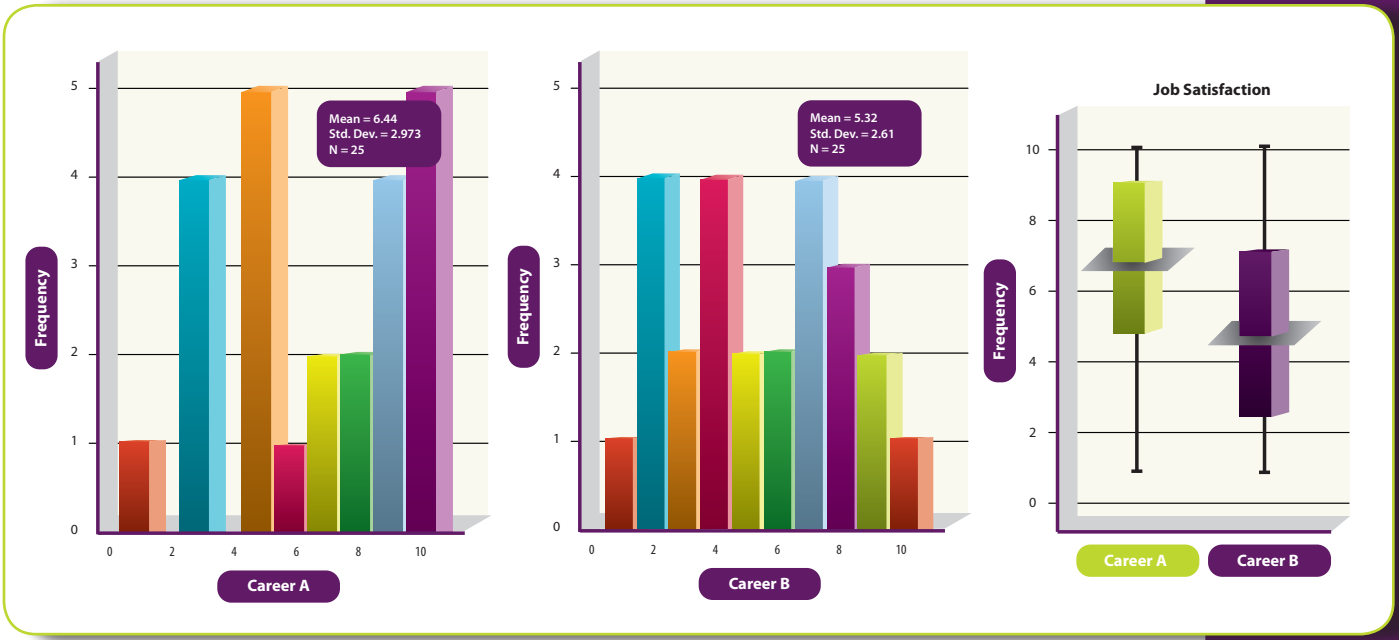
1, 1, 3, 3, 3, 3, 5, 5, 5, 5, 5, 6, 7, 7, 8, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10

Career B

1, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 5, 5, 6, 6, 7, 7, 7, 7, 8, 8, 8, 9, 9, 10

1)

2) Career A appears to be negatively skewed, or skewed left, indicating



a higher frequency of high job satisfaction scores. Career B appears to be more normal.

3) Career A has a larger median score for job satisfaction. It can also be discerned that Career A has a larger mean. The standard deviations appear to be similar.

The summary statistics are given below:

Descriptive Statistics

City	Mean	Median	Mode	Range	Standard Deviation
Career A	6.44	7	5	9	2.97
Career B	5.32	5	2	9	2.61

4) Because the range is small, the measures of mean or median would be appropriate for the center.

5) Two-sample t-tests results are shown below:

$p > 0.05, t \approx 1.4$

The p-value is greater than 0.05, the common threshold for the level of significance. Because it is greater, the null hypothesis of no difference should not be rejected. Thus, we can declare that there is no significant difference between job satisfaction for the two careers.

Statistics Reference Sheet

Take Home

Definitions

A ***sample*** is a subset of a population. A ***population*** is simply the entire group from which a sample is drawn. The sample is used to approximate the population.

A ***random sample*** is a sample in which each participant has an equal chance of being chosen. Random samples may be created by such methods as choosing the 15th name from a list, drawing numbers from a hat, using a random-number generator, or assigning numbers to subjects. Random samples also may be drawn from stratified groups. There are several approaches that may be used to create a random sample. Such samples allow the researcher to use the mean to approximate a population.

A ***boxplot*** is a graphical representation that shows the median, first and third quartiles, and outliers of a set of data.

A ***histogram*** is another type of graphical representation that shows the frequencies of values for intervals of data. Thus, each bar represents a range (or interval) of data.

A histogram shows the ***shape*** of a distribution of data. Normally distributed data show a bell, or normal, curve. These data show a constant rate of increase and then decrease in scores. More normally distributed data allow the researcher to make estimates with a larger degree of certainty. The shape of data may be discussed in terms of ***normality***, as well as skewness, kurtosis, and modality. The ***data*** of a data set describes whether there are more low scores or high scores. Data skewed to the left (or negatively skewed) means there are more high scores. Data skewed to the right

Key Terms
sample
random sample
boxplot
histogram
mean
median
mode

(or positively skewed) means there are more low scores. ***Kurtosis*** refers to the flatness or steepness of a distribution. The names of the types of kurtosis need not be described here. ***Modality*** refers to the number of peaks in a distribution. For example, a distribution may be unimodal (one peak) or bimodal (two peaks).

Key Terms
range
interquartile range
standard deviation
t-test
two-sample t-test
one-sample t-test

Measures of center include the mean, median, and mode. The ***mean*** of a data set is the average of the values in the set. It is a measure of center. The ***median*** of a data set is the middle value (or average of the two middle values) of the set. When outliers are present, the median is a more desired measure of center. The ***mode*** is simply the value that occurs most often. Some data sets will have no mode or more than one mode.

Some measures of spread are range, interquartile range, and standard deviation. The ***range*** of a data set is the difference between the maximum and minimum value. The ***interquartile range*** is the difference between the first and third quartiles, or the median of the lower half of the scores and the median of the upper half of the scores. The ***standard deviation*** measures the deviation of scores about the mean. Mathematically speaking, standard deviation is the square root of the ratio of the sum of the squared deviations and the difference of the sample size and 1. This formula is

$$s_x = \sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{N-1}}$$

written as:

(*The standard deviation may be computed very quickly using a calculator or spreadsheet.*)

A ***t-test*** is a statistical test that compares two groups of interval data or compares the mean of a sample to a population mean (claimed mean). A ***two-sample t-test*** compares the means and standard deviations of two

groups, in order to determine if a statistically significant difference exists between the groups. A ***one-sample t-test*** compares a sample mean to some score. Either test may be performed using a graphing calculator, an Excel spreadsheet, or some other software.

Sample Analysis

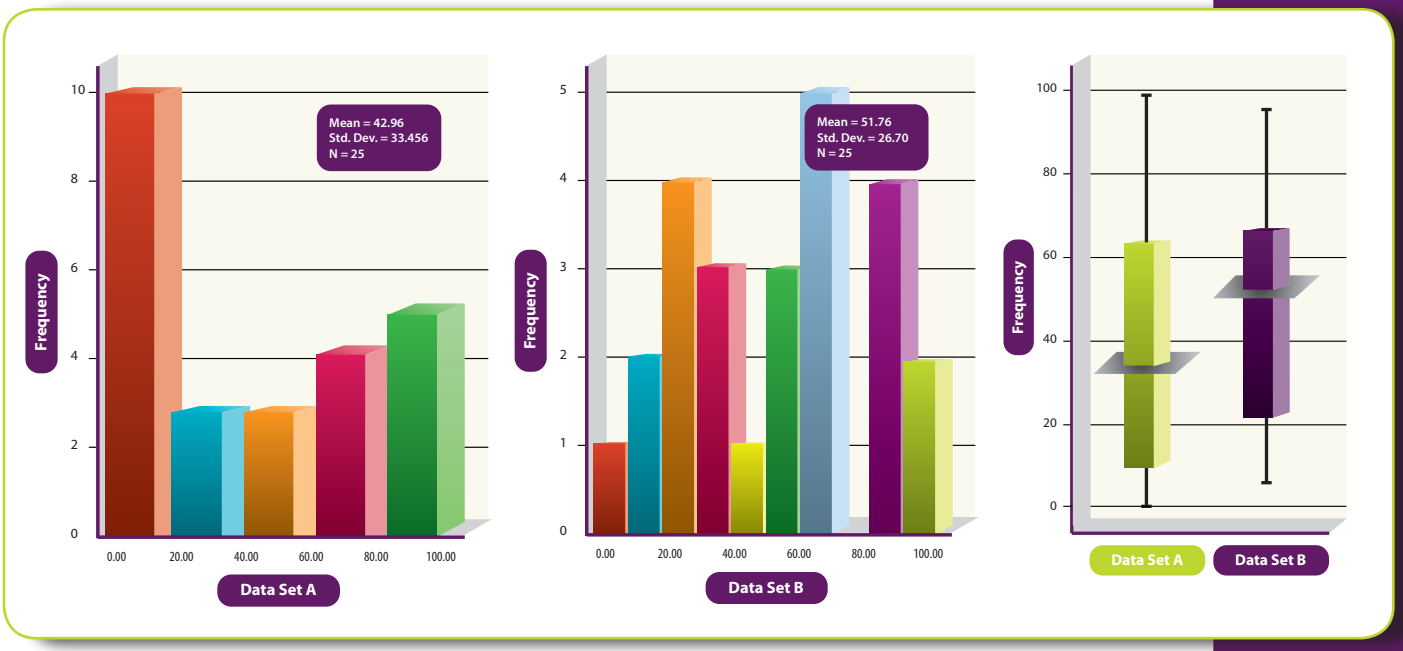
Data Set A

1, 4, 5, 5, 6, 10, 12, 12, 14, 19, 34, 35, 37,
45, 55, 56, 60, 63, 65, 74, 89, 90, 91, 93, 99

Data Set B

9, 10, 17, 21, 21, 26, 26, 37, 37, 37, 40, 55, 56,
59, 61, 62, 64, 68, 69, 83, 84, 84, 84, 90, 94

Consider the data sets below:



The following graphical displays may be created:

Descriptive Statistics

Data Set	Mean	Median	Mode	Range	Standard Deviation
Data Set A	42.96	37	5	98	33.46
Data Set B	51.76	56	37	85	26.7

Descriptive Statistics are as follows:

Data Set B is more normal. It also has a larger median.

Data Set A is skewed right, meaning there is a larger number of smaller scores. The histogram for Data Set A shows the mean will be smaller, since there is a greater number of smaller scores, and the mean is pulled towards the tails. The boxplot for Data Set A shows a greater interquartile range and deviation about the center. Thus, it can be inferred that the standard deviation for Data Set A will be greater.

Because there is a larger range, the median would be a more appropriate measure of center to use.

A two-sample *t*-test shows the following:

$p > 0.05; t \approx \square 0.13$

Because the p-value is greater than 0.05, the commonly used level of significance (alpha), we will declare that no statistically significant difference exists between the sets.

Suppose the population mean is 67. Use a one-sample t-test to see if the random data from Set A supports or refutes this population mean.

The results are:

$p < 0.01; t \approx \square 3.6$

The results show that we may declare the population mean is false (rejecting the null hypothesis of no difference between the population mean and sample mean).