**If the Shoe Fits!**

**Overview of Lesson**

In this activity, students explore and use hypothetical data collected on student shoe print lengths, height, and gender in order to help develop a tentative description of a person who entered a school’s grounds over a weekend without permission. Graphs such as comparative boxplots and scatterplots are drawn to illustrate the data. Measures of center (median, mean) and spread (range, Interquartile Range (IQR), Mean Absolute Deviation (MAD)) are computed. Conclusions are drawn based upon the data analysis in the context of question(s) asked.

**GAISE Components**

This investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. The four components are: formulate a question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level B activity.

**Common Core State Standards for Mathematical Practice**

1. Make sense of problems and persevere in solving them.

2. Reason abstractly and quantitatively.

3. Construct viable arguments and critique the reasoning of others.

4. Model with mathematics.

5. Use appropriate tools strategically.

**Common Core State Standard Grade Level Content (High School)**

S-ID. 1. Represent data with plots on the real number line (dot plots, histograms, and box plots).

S-ID. 2. Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (interquartile range, standard deviation) of two or more different data sets.

S-ID. 3. Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

S-ID. 6. Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

S-IC. 3. Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

**NCTM Principles and Standards for School Mathematics**

**Data Analysis and Probability Standards for Grades 9-12**

**Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them:**

* understand the differences among various kinds of studies and which types of inferences can legitimately be drawn from each;
* know the characteristics of well-designed studies, including the role of randomization in surveys and experiments;
* understand the meaning of measurement data and categorical data, of univariate and bivariate data, and of the term variable;
* understand histograms, parallel box plots, and scatterplots and use them to display data;
* compute basic statistics and understand the distinction between a statistic and a parameter.

**Select and use appropriate statistical methods to analyze data:**

* for univariate measurement data, be able to display the distribution, describe its shape, and select and calculate summary statistics;
* for bivariate measurement data, be able to display a scatterplot, describe its shape, and determine regression coefficients, regression equations, and correlation coefficients using technological tools;
* display and discuss bivariate data where at least one variable is categorical.

**Prerequisites**

Students will have knowledge of calculating numerical summaries for one variable (mean, median, Mean Absolute Deviation (MAD), five-number summary). Students will have knowledge of how to construct boxplots and scatterplots. Students will have some familiarity with general concepts from linear regression.

**Learning Targets**

Students will be able to calculate numerical summaries and use them to compare two data sets. Students will be able to determine if any data values are outliers. Students will be able to display data in comparative boxplots and use the plot to compare two data sets. Students will be able to display the relationship between two variables on a scatterplot and interpret the resulting plot.

**Time Required**

1 class period.

**Materials Required**

Pencil and (graphing) paper; graphing calculator or statistical software package.

**Instructional Lesson Plan**

**The GAISE Statistical Problem-Solving Procedure**

**I. Formulate Question(s)**

Begin the lesson by discussing a hypothetical background: Welcome to CSI at School! Over the weekend, a student entered the school grounds without permission. Even though it appears that the culprit was just looking for a quiet place to study undisturbed by friends, school administrators are anxious to identify the offender and have asked for your help. The only available evidence is a suspicious footprint found outside the library door.

Ask students to write some questions that they would be interested in investigating about students’ shoeprint lengths. Some possible questions might be:

1. What kinds of lengths do shoeprints have? What is the shortest shoe print length expected to be? What is the longest shoe print length expected to be?
2. Are there differences in the shoe print lengths for males and females? If so, what are the differences?
3. Are shoe print lengths related to any other variables?

The investigation that follows is based upon some of the questions that might be posed by students and on attempting to answer the overall question of identifying the offender.

**II. Design and Implement a Plan to Collect the Data**

Explain the following hypothetical scenario for data collection: After the incident, school administrators arranged for data to be obtained from a random sample of this high school’s students. The data table shows the shoe print length (in cm), height (in inches), and gender for each individual in the sample.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Shoe Print Length** | **Height** | **Gender** | **Shoe****Print****Length** | **Height** | **Gender** |
| 24 | 71 | F | 24.5 | 68.5 | F |
| 32 | 74 | M | 22.5 | 59 | F |
| 27 | 65 | F | 29 | 74 | M |
| 26 | 64 | F | 24.5 | 61 | F |
| 25.5 | 64 | F | 25 | 66 | F |
| 30 | 65 | M | 37 | 72 | M |
| 31 | 71 | M | 27 | 67 | F |
| 29.5 | 67 | M | 32.5 | 70 | M |
| 29 | 72 | F | 27 | 66 | F |
| 25 | 63 | F | 27.5 | 65 | F |
| 27.5 | 72 | F | 25 | 62 | F |
| 25.5 | 64 | F | 31 | 69 | M |
| 27 | 67 | F | 32 | 72 | M |
| 31 | 69 | M | 27.4 | 67 | F |
| 26 | 64 | F | 30 | 71 | M |
| 27 | 67 | F | 25 | 67 | F |
| 28 | 67 | F | 26.5 | 65.5 | F |
| 26.5 | 64 | F | 30 | 70 | F |
| 22.5 | 61 | F | 31 | 66 | F |
|  |  |  | 27.25 | 67 | F |

Ask students to explain why this is an observational study and not an experiment. Students should note that in this context, nothing has been done deliberately to the students in order to measure their responses. From direct observation and measurement, data values were recorded for each student’s height, shoe print length, and gender.

Ask students to think about why the school’s administrators chose to collect data on a random sample of students from the school. What benefit might a random sample offer? Students should note that whenever possible, random selectionshould be used to choose samples for an observational study. By using random selection, chance determines which individuals are included in the sample. This helps ensure that a sample is representative of the population from which it was chosen. Random selection allows the researchers to generalize sample results to a larger population of interest.

**III. Analyze the Data**

Have students begin the investigation by performing data analyses to help determine the gender of the offender. By using appropriate graphs and numerical calculations students can compare shoe print lengths for males and females.

Ask students to suggest graphs that might be used to use to compare the shoe print length data distributions for females and males. Comparative graphs such as dotplots or boxplots are appropriate for displaying these data. Students describe one advantage of using comparative dotplots instead of comparative boxplots to display these data. Students describe one advantage of using comparative boxplots instead of comparative dotplots to display these data. Comparative dotplots have the advantage of showing each individual data value while comparative boxplots are useful for comparing the percentiles of the two distributions.

After a discussion, ask students to construct comparative boxplots for the shoe print lengths for males and females. First, have students calculate five-number summaries of the shoe print length data for males and females. Additionally, have students determine if there are any outlying shoeprint length values. The five-number summaries of the shoe print length data are shown in the Table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Minimum | Quartile 1 (Q1) | Median (Q2) | Quartile 3 (Q3) | Maximum |
| Male | 29 | 30 | 31 | 32 | 37 |
| Female | 22.5 | 25 | 26.5 | 27.4 | 31 |

Note that the median is the halfway point in the data set. Also note that the first quartile is the median of the data points strictly below the median of the data set, and the third quartile is the median of the data points strictly above the median of the data set.

Demonstrate to students that in order to check for outlying shoe print lengths for the males:

(1) the interquartile range (IQR) is calculated: Q3 – Q1 = 32 – 30 = 2 cm

(2) the IQR is multiplied by 1.5: 1.5(2) = 3 cm

(3) 3 is subtracted from Q1: 30 – 3 = 27 cm

(4) 3 is added to Q3: 32 + 3 = 35 cm

Any shoe print length values smaller than 27 cm or greater than 35 cm are outlying values. There is one male student with an outlying shoe print length of 37 cm.

In order to check for outlying shoe print lengths for the females:

(1) the interquartile range (IQR) is calculated: Q3 – Q1 = 27.4 – 25 = 2.4 cm

(2) 1.5(IQR) is 1.5(2.4) = 3.6 cm

(3) 25 – 3.6 = 21.4 cm

(4) 27.4 + 3.6 = 31 cm

Any shoe print length values smaller than 21.4 cm or greater than 31 cm are outlying values. There is one female student with an outlying shoe print length of 31 cm.

Now ask students to construct comparative boxplots to display the shoe print length distributions. The Figure below illustrates the comparative boxplots.



Discuss with students how to interpret the boxplots. Students should understand that there are about the same number of shoe print lengths between the minimum and Q1, Q1 to Q2, Q2 to Q3, and Q3 to the maximum; or approximately 25% of the data will lie in each of these four intervals.

Ask students to describe the similarities and differences in the shoe print length distributions for the males and females in the sample. The box plots show that the females tended to have shorter shoe print lengths than the males. The median shoe print length for females was much lower than for males. The interquartile range for females is much lower than the interquartile range for males. All shoe print lengths for males are longer than at least 75% of the shoe print lengths for females. There is one high outlier in the female group and one high outlier in the male group.

The next step in the analysis of the data is to focus on measures of center. The mean and median shoe print lengths for males are 31.36 and 31 cm; respectively. For females, the mean and median lengths were 26.31 and 26.5 cm; respectively. As expected, it seems that a typical male shoe print length was longer than a typical female shoe print length by about 5 cm.

Students are now asked to characterize the spreadof the shoe print length distributions. Ask students to begin by calculating the range in shoe print lengths. For males the data cover the interval from the minimum of 29 cm to the maximum of 37 cm so the range in lengths is 8 cm. For females, the range in lengths is 8.5 cm. Thus, the ranges of shoe print lengths for males and females are very close.

In addition to the range the boxplot suggests another measure of spread, the interquartile range (IQR). The IQR provides a measure of spread of the middle 50% of the shoe print lengths. The interquartile ranges for males and females are 2 cm and 2.4 cm, respectively. These IQRs are very comparable.

Yet another measure of spread can be calculated by incorporating how far the data are from the mean, on average. This measure, called the mean absolute deviation (MAD), is the arithmetic average of the absolute deviations of the data values from their mean. Students find the MAD using the following steps:

1. Find the deviations from the mean (by subtracting the mean from each shoe print length).

2. Find the absolute value of each deviation.

3. Find the mean of the absolute values.

The table below shows the calculation of the MAD for the male shoe print lengths. For the males the MAD = 16.08/11 = 1.46 cm. In words, on average the male shoe print lengths are 1.46 cm away from the mean male shoe print length of 31.36 cm.

Shoe Print Length Length – Mean |Length – Mean|

29 29 – 31.36 2.36

29.5 29.5 – 31.36 1.86

30 30 – 31.36 1.36

30 30 – 31.36 1.36

31 31 – 31.36 .36

31 31 – 31.36 .36

31 31 – 31.36 .36

32 32 – 31.36 .64

32 32 – 31.36 .64

32.5 32.5 – 31.36 1.14

37 37 – 31.36 5.64

 Sum  0 Sum = 16.08

Mention to students that the sum of the Length – Mean column is approximately zero and that the sum of the deviations from the mean for any data set will be zero. Ask students to explain why this sum is zero. The negative and positive deviations will ‘cancel each other out’. This is the motivation for finding the mean *absolute* deviation.

A similar calculation for the female shoe print lengths gives the MAD = 42.03/28 = 1.50 cm. The female shoe print lengths are, on average, 1.50 cm away from the mean female shoe print length of 26.31 cm.

After completing a data analysis of the shoe print lengths for males and females ask students to think about the shoe print length of the offender and how this length might be used to help determine if the shoe print was left by a male or a female. Ask students: If the length of the offending student’s shoe print was 32 cm would you think that the print was made by a male or a female? How sure are you that you are correct? Ask students to explain their reasoning. Students will say that a shoe print length of 32 cm is beyond the maximum shoe print length for the 28 females in the sample. Further the maximum female shoe print length of 31 cm was determined to be an outlier. It is highly unlikely that a shoe print 32 cm long was left by a female. Thus the offending student is most likely not a female. Note that the mean shoe print length for the 11 sampled males is 31.36 cm so a shoe print length of 32 cm long would not be unexpected for a male.

Now, ask students: how would you answer if the suspect’s shoe print length was 27 cm? Students will note that a shoe print length of 27 cm is 2 cm below the minimum shoe print length for the 11 sampled males. It is highly unlikely that a shoe print 27 cm long was left by a male. The offending student is most likely not a male. The mean shoe print length for the 28 sampled females is 26.31 cm with a MAD of 1.50 cm. Further; students will note that a shoe print length of 27 cm is only .4 cm below the third quartile of the shoe print lengths for females. We would expect close to 25% of female students to have a shoe print length of 27 cm or longer. It is much more likely for a female student to leave a shoe print 27 cm long than it is for a male student.

In the next phase of analysis have students explore the relationship between height (in inches) and length of shoe print (in cm). Discuss which variable should be the explanatory (independent) variable and which variable should be the response (dependent) variable. Here we would expect a student’s shoe print length to help explain the height of the student (since the shoe print length of the offender is assumed to be known). Ask students to construct a scatterplot of height versus shoe print length using different colors or different plotting symbols to represent the data for males and females. The Figure below illustrates the scatterplot.



Ask students if it looks like there is an overall linear relationship between height and shoe print length. The scatterplot does indicate a somewhat strong, positive, linear association between shoe print length and height. We can see that as the shoe print length increases the height also tends to increase. Ask students to draw a line through the “center of the data”.

Since different plotting symbols were used for males and females ask students if it appears that the same straight line could be used to summarize the relationship between height and shoe print length for males and females. On the scatterplot the male heights and shoe print lengths appear to both have larger values than the values for the females. For both genders as was mentioned above the height tends to increase as the shoe print length increases. As far as the linear association within each gender it appears that the females have a larger increase in height for a unit (1 cm) increase in the shoe print length. For females it appears as though the height increases by about 1 inch for every 1 cm increase in shoe print length. For males it appears as though the height increases by about .5 inches for every 1 cm increase in shoe print length. Additionally, if students were asked to draw a line through the “center of the data” for males and then do the same for females; the y-intercept of the female line would be roughly 20 inches lower than the y-intercept of the male line.

Once again referring to the scatterplot for both genders combined ask students: Based on the scatterplot if a student’s shoe print length was 30 cm, approximately what height would you predict for the person who made the shoe print? Ask students to explain how they arrived at their prediction. Looking at the scatterplot if we locate 30 cm on the horizontal scale and we visualize having drawn a line through the center of the data, we see that for a 30 cm shoe print length the predicted student height is between 65 and 70 inches.

**IV. Interpret the Results**

Now that students have compared and contrasted the distributions of shoe print lengths for males and females and examined the relationship between shoe print length (in cm) and height (in inches) both overall and within the genders; ask students to answer the hypothetical question that follows. Suppose that the mystery shoe print found outside the library actually had a length of 30 cm. Based on the given data and the analysis of the data write a description of the person who you think may have left the print. Explain the reasoning that led to your description and give some indication of how confident you are that your description is correct.

Students will note that a shoe print length of 30 cm is equal to the first quartile of the shoe print lengths for the 11 sampled males. So, about 25% of the males have a shoe print length of 30 cm or shorter. Recall that the mean shoe print length for the 11 sampled males is 31.36 cm with a MAD of 1.46 cm. A shoe print length of 30 cm is only 1 cm below the longest female shoe print length. The third quartile of the female shoe print lengths is 27.4 cm. 30 cm is 2.6 cm above this value. Additionally, the mean shoe print length for the 28 sampled females is 26.31 cm with a MAD of 1.50 cm. A female shoe print as long as 30 cm is highly unlikely. Further if we consider the scatterplot recall that the predicted height corresponding to a shoe print length of 30 cm is between 65 and 70 inches – or between 5 foot 5 inches and 5 foot 10 inches; a height range that more likely corresponds to a male versus a female.

Lastly discuss the fact that this sample of students may or may not be representative of the larger population of students at this high school and ask students to answer: Would you feel comfortable generalizing this result to the population of students at the school? Why or why not? Students should acknowledge that since a random sample of student data was utilized in the analysis; looking beyond the data is feasible; however, we always need to be mindful of sampling error and sampling variability.

**Assessment**

1. A question in the 2002 General Social Survey (GSS) conducted by the national Opinion Research Center asked participants how long they spend on e-mail each week. A summary of responses (hours) for 1881 respondents follows.

  ****

(a) Explain how the summary statistics show us that at least 25% of the respondents said that they do not use e-mail.

(b) What is the interval that contains the lower 50% of the responses?

(c) What is the interval that contains the upper 50% of the responses?

(d) Explain whether the maximum value, 70 hours, would be marked as an outlier on a boxplot.

**Answer:**

(a)The lower quartile Q1 = 0 hours. By definition, 25% of the values are at or below the lower quartile.

(b)0 to 2 hours (Minimum to Median).

(c) 2 to 70 hours (Median to Maximum).

(d) Yes, 70 hours would be marked as an outlier.

The boundary defining an outlier on the high side is Q3 + 1.5×IQR = 5 + 1.5 (5-0) = 12.5 hours.

2. The figure below is a boxplot for comparing tip percentages for a male and a female restaurant server, each of whom drew happy faces on the checks of randomly selected dining parties. Discuss the ways in which the tip percentage distributions for the two servers differ.



**Answer:** Generally females tended to have higher tip percentages. The median is clearly greater for females than it is for males. Note that the first quartile for females is above the third quartile for the males. As can be seen from the overall range and the interquartile range, the data for the females shows greater spread than the data for the males. There is one high outlier for the females. The males do not have any outliers.

**Possible Extensions (to GAISE Level C)**

1. Determine the estimated regression equation for predicting a student’s shoe print length (in cm) from the student’s height (in inches).

(a) Interpret the slope of the estimated regression equation in the context of this problem.

(b) Predict the height of a student who has a shoe print length of 30 cm.

(c) Calculate the residual for a student who has a shoe print length of 30 cm.

2. Give the numerical value and interpret the Pearson correlation coefficient between a student’s shoe print length (in cm) and a student’s height (in inches).

**References**

1. *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) Report, ASA, Franklin et al., ASA, 2007 <http://www.amstat.org/education/gaise/>

2. Adapted from an activity appearing in *Making Sense of Statistical Studies* by Peck and Starnes (with Kranendonk and Morita), ASA, 2009 <http://www.amstat.org/education/msss>

3. Adapted from the Investigation: How Long are our Shoes? In Bridging the Gap Between the Common Core State Standards and Teaching Statistics (2012, in press).

Authors: Pat Hopfensperger, Tim Jacobbe, Deborah Lurie, and Jerry Moreno.

4. CSI Stats: Helping Students Become Data Detectives with the GAISE Framework presented by Daren Starnes. Recorded web-based seminar. <http://www.amstat.org/education/webinars/>

5. Assessment questions from: *Mind on Statistics*, Third Edition by Utts/Heckard, 2006. Cengage Learning.

**If the Shoe Fits! Activity Sheet**

Welcome to CSI at School! Over the weekend, a student entered the school grounds without permission. Even though it appears that the culprit was just looking for a quiet place to study undisturbed by friends, school administrators are anxious to identify the offender and have asked for your help. The only available evidence is a suspicious footprint outside the library door.

After the incident, school administrators arranged for the data in the table below to be obtained from a random sample of this high school’s students. The table shows the shoe print length (in cm), height (in inches), and gender for each individual in the sample.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Shoe Print Length** | **Height** | **Gender** | **Shoe****Print****Length** | **Height** | **Gender** |
| 24 | 71 | F | 24.5 | 68.5 | F |
| 32 | 74 | M | 22.5 | 59 | F |
| 27 | 65 | F | 29 | 74 | M |
| 26 | 64 | F | 24.5 | 61 | F |
| 25.5 | 64 | F | 25 | 66 | F |
| 30 | 65 | M | 37 | 72 | M |
| 31 | 71 | M | 27 | 67 | F |
| 29.5 | 67 | M | 32.5 | 70 | M |
| 29 | 72 | F | 27 | 66 | F |
| 25 | 63 | F | 27.5 | 65 | F |
| 27.5 | 72 | F | 25 | 62 | F |
| 25.5 | 64 | F | 31 | 69 | M |
| 27 | 67 | F | 32 | 72 | M |
| 31 | 69 | M | 27.4 | 67 | F |
| 26 | 64 | F | 30 | 71 | M |
| 27 | 67 | F | 25 | 67 | F |
| 28 | 67 | F | 26.5 | 65.5 | F |
| 26.5 | 64 | F | 30 | 70 | F |
| 22.5 | 61 | F | 31 | 66 | F |
|  |  |  | 27.25 | 67 | F |

**1.** Explain why this study was an observational study and not an experiment.

**2.** Why do you think the school’s administrators chose to collect data on a random sample of students from the school? What benefit might a random sample offer?

**3.** Suggest a graph that might be used to use to compare the shoe print length data distributions for females and males.

**4.** Describe one advantage of using comparative boxplots instead of comparative dotplots to display these data.

**5.** For each gender calculate the five-number summary for the shoe print lengths. Additionally, for each gender, determine if there are any outlying shoe print length values.

**6.** Construct comparative boxplots for the shoe print lengths of males and females. Discuss the similarities and differences in the shoe print length distributions for the males and females in this sample.

**7.** For each gender calculate the mean shoe print length. What information does the mean shoe print length provide?

**8.** The mean will give us an indication of a typical shoe print length. In addition to knowing a typical length we would also like to know how much variability to expect around this length. For each gender calculate the Range; Interquartile Range; and Mean Absolute Deviation of the shoe print lengths. Interpret each of the calculated values.

**9.** If the length of a student’s shoe print was 32 cm, would you think that the print was made by a male or a female? How sure are you that you are correct? Explain your reasoning. Use results from Questions 5 through 8 in your explanation.

**10.** How would you answer Question 9 if the suspect’s shoe print length was 27 cm?

**11.** Construct a scatterplot of height (vertical scale) versus shoe print length horizontal scale) using different colors or different plotting symbols to represent the data for males and females.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

(a) Interpret the scatterplot. Does it look like there is a linear relationship between height and shoe print length? Explain.

(b) Does it look like the same straight line could be used to summarize the relationship between shoe print length and height for both males and females? Explain.

(c) Based on the scatterplot, if a student’s shoe print length was 30 cm, approximately what height would you predict for the person who made the shoe print? Explain how you arrived at your prediction.

**12.**  Suppose that the shoe print found outside the library actually had a length of 30 cm. Based on the given data and your answers to Questions 5 through 11, write a description of the person who you think may have left the shoe print. Explain the reasoning that led to your description and give some indication of how confident you are that your description is correct.