BioMath

Biostatistics in Practice: Using Statistics To Discover Links Between Eating Behaviors And Overweight Children

Student Edition

COMAP

Funded by the National Science Foundation,
Proposal No. ESI-06-28091

This material was prepared with the support of the National Science Foundation. However, any opinions, findings, conclusions, and/or recommendations herein are those of the authors and do not necessarily reflect the views of the NSF.

At the time of publishing, all included URLs were checked and active. We make every effort to make sure all links stay active, but we cannot make any guaranties that they will remain so. If you find a URL that is inactive, please inform us at info@comap.com.

Published by COMAP, Inc. in conjunction with DIMACS, Rutgers University.
©2015 COMAP, Inc. Printed in the U.S.A.

COMAP, Inc.
175 Middlesex Turnpike, Suite 3B
Bedford, MA 01730
www.comap.com

ISBN: 1 933223 54 5

Front Cover Photograph: EPA GULF BREEZE LABORATORY, PATHO-BIOLOGY LAB. LINDA SHARP ASSISTANT

This work is in the public domain in the United States because it is a work prepared by an officer or employee of the United States Government as part of that person's official duties.

Biostatistics in Practice: Using Statistics To Discover Links Between Eating Behaviors And Overweight Children

Biostatistics is a field that uses mathematics and statistics to provide answers to questions asked by researchers who study the environment, biology, medicine, and public health. The science of biostatistics includes

- designing biological experiments,
- collecting, summarizing, and analyzing data from these experiments, and
- interpreting the results.

Biostatisticians work in universities, health organizations, government agencies, pharmaceutical companies, and other types of private companies. They might be involved with monitoring the spread of diseases (in human, animal, or plant populations), determining risk factors for certain diseases (such as diabetes and heart disease), or conducting research on genetically altered food.

In this unit, the subjects being studied are people. No two of us are exactly alike. Instead, we vary in all sorts of ways. Statistics provides tools to describe the variability in data. It helps us find patterns in data and determine relationships between variables. This unit explores two research questions about people. The first question is what factors are related to estimating a person's age. The second question is what factors are related to being overweight.

Before beginning this unit, you need to understand the use of two words, **factor** and **variable**. A variable is anything that varies. For example, peoples' weights vary. So, weight is a variable. The average number of times a person eats at a fast food restaurant per week varies from person to person. Hence, it is also a variable. In a study investigating the rising weight of Americans (weight is the *variable* under study), the average frequency of eating at fast food restaurants might be investigated as a *factor* related to weight and, as a *factor*, could help explain increases in weight.

Unit Goals and Objectives

Unit Goal: Students will gain a better understanding of the scientific method.
Objectives: Students have some experience with the following:

- Formulate research questions and hypotheses.
- Gather data in a scientific fashion.
- Use appropriate statistical tools to analyze data and extract information relevant to research questions and hypotheses.
- Use information from data analysis to support or refute hypotheses.
- Suggest further areas for research or suggest different approaches to analysis of the data.

Goal: Students will gain an understanding of variability.
Objectives: Students will be able to:

- Collect data on age estimates for five people.
- Create dotplots to display data.

- Summarize the center of age-estimate data for each person using mean and median and notice that these summaries differ from person to person.
- Summarize spread (variability) in age-estimate data for each person using range.
- Evaluate estimation process using bias.

Goal: Students will gain an understanding of how to analyze quantitative data on one variable.
Objectives: Students will be able to:

- Create five-number summaries.
- Use the 1.5×iqr rule to identify outliers.
- Display data using modified boxplots.
- Form a new variable by combining two (or more) other variables.

Goal: Students will gain an understanding of how to analyze quantitative data on two variables in order to investigate relationships between the variables.

- Display data using scatterplots.
- Determine the direction of a relationship – positive association, negative association.
- Summarize linear patterns in data with a line of best fit (regression line).
- Use the regression equation to make a prediction.
- Calculate a residual error.
- Interpret the strength (amount of variability explained) of a linear relationship using R^2.

Goal: Students will gain an understanding of how to analyze data on two binary (categorical) variables in order to investigate relationships between the variables.
Objective:

- Create binary variables such as overweight (yes/no).
- Organize data on two binary variables using two-way tables.
- Calculate row and column percentages and interpret results.

Lesson 1 Identifying Factors Related to Age

In the "Guess My Age" activity, you will be a part of a small group or team. Five pictures of different people will be displayed. As a team, it is your job to decide the age (variable) of each person. Think of yourself as a scientist using clues (factors) in the pictures to provide the best estimate of age for each person. As you work through this activity, keep in mind the first research question: What factors are related to estimating a person's age?

Before the activity, let's review some of the statistical measures you have previously learned.

Measures of Center and Variability

Two common measures of centrality are **mean** and **median**. The average or mean is probably the most commonly used measure of center. The mean is denoted as $\overline{x}$. To calculate the mean, sum the data and divide by the number of data.

$$\overline{x} = \frac{\text{sum of data}}{\text{number of data}}$$

Another measure of the center is the median, which gives the middle number of a data set. To calculate the median of a data set:

- Arrange the data in order from smallest to largest.
- If the number of data (n) is odd, the median is the middle number. To find it, start at one end of the ordered data and count $(n + 1)/2$ observations.
- If the number of data (n) is even, the median is the average (mean) of the middle two numbers. Start at each end of the ordered data and count in $n/2$ observations. Find the mean of these data values.

One way to describe the spread of a set of observations is by calculating the **range**. The range measures how far apart the smallest data value is from the largest data value. The range is the difference between the minimum value and the maximum value in a data set.

ACTIVITY 1-1 Guess My Age

Objective: Make estimates to collect and analyze data.
Materials:
Handout BS-H1: Guess My Age Activity Worksheet

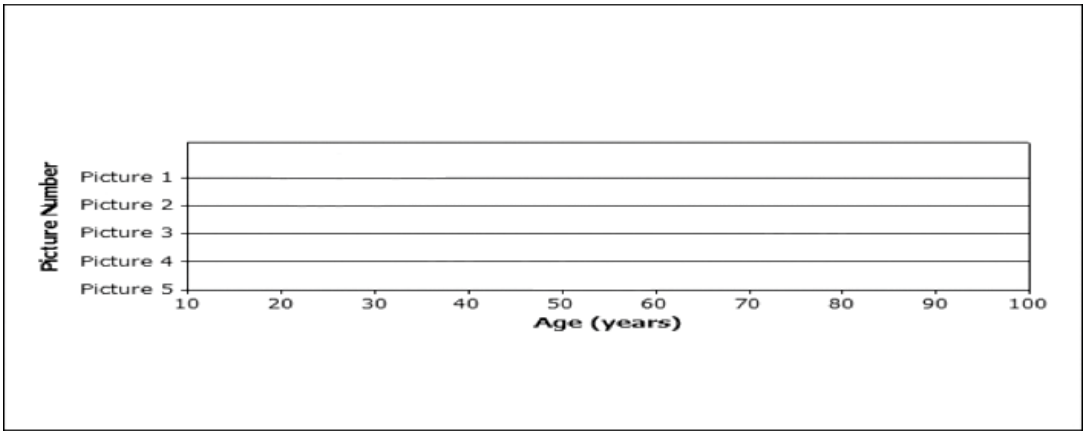
1. You will see pictures of 5 different people. After each picture is shown, take 3 minutes to estimate the age of the person in the picture. Talk with the members of your group and agree on an estimate. Record your team estimate on the corresponding section of the last page of your activity handout and also in your team's column on the chart below. Hand in the team's age estimate. Repeat for each picture.

Class Data – Estimated Age in Years

Picture Number	Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	True Age
Picture 1								
Picture 2								
Picture 3								
Picture 4								
Picture 5								

2. When all teams have completed estimates fill in the entire chart above.

3. For each picture, make a **dotplot** of all of the estimates by placing a dot above each number that corresponds to an estimate. (If two estimates are the same, place one of the dots directly above the other.)



4. After completing dotplots for each picture, now summarize the results with a single number that indicates the center (location) of the data. Look at your table of data.

a. For each picture, calculate the mean and median of the teams' age estimates.

b. For which pictures are the means and medians farthest apart?

c. When the mean and median are farther apart from each other, are the age estimates more or less spread out than when the mean and median are closer together? (Your dotplots may help you answer this question.)

5. Estimates for some of the pictures were more spread out than for other pictures.

a. Calculate the range of age estimates for each picture.

b. Identify the picture corresponding to the smallest range. Identify the picture corresponding to the largest range. What do these values tell you?

6. Your teacher will now tell you the true ages of each person in the photos.

a. Record the true ages of the people in the pictures in your chart (part 2).

b. For each picture, calculate your team's difference between estimate and true age for each photo (age estimate – true age), and record the value in the chart below. Once all teams have determined each value, fill in the chart.

Bias = Age Estimate – True Age

Picture Number	Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	True Age
Picture 1								
Picture 2								
Picture 3								
Picture 4								
Picture 5								

c. Look at the completed chart above. Unless you overestimated everyone's age or underestimated everyone's age, you have both positive and negative numbers in your chart. These values indicate **bias**. Bias measures how far off an estimate is from the true value. An overestimate produces a positive bias; an underestimate produces a negative bias. How should teams combine their bias values into a single number that will give the team a score for how well or how poorly they did at estimating the people's ages in the five pictures?

Bias and the Best Answer

In statistics, each statistical method used to answer a question needs to be evaluated to see if the method produces a "good" answer. In this lesson's activity, the true ages of the people in the pictures (at the time each picture was taken) is known. So, it's possible to compare each team's age estimate with the person's true age. In other words, we can compute the bias, the difference between the estimated age and the true age. It is still difficult, however, to know which team was the best estimator overall.

Questions for Discussion

1. In order to decide which team did the best job in estimating the true ages, you will need to give each team a performance score that indicates how well they did. Decide how to use each team's bias values to come up with a score.

2. Calculate your team's score using the method decided in question 1. Compare your score to the scores for other teams. Which team is the winner?

3. For some pictures the bias across all teams is larger than for other pictures. Look at the pictures again one at a time. What factors in the pictures may have led to better estimates? What factors in the pictures may have led to worse estimates?

4. If you were to participate in this activity again, what factors would you want to see in the pictures or taken out of the pictures so that you could provide a better estimate of age?

Introduction to the Study of Factors Related to Being Overweight

The number of children, teenagers, and adults who are overweight has more than doubled in the past 20 years among all age groups. From a public health and medical perspective, being overweight affects a person's health. When a person continuously eats more calories than he or she can burn off, the extra calories get stored as fat. The burden on the body of the stored fat may

contribute to the development of diabetes, heart disease, and high blood pressure, which can develop in children or later in adults.

In an effort to combat obesity, many scientists, doctors, and public health advocates are conducting studies to learn more about the factors relating to being overweight. A key to preventing obesity is identifying risk factors that are related to weight in general, and more specifically, to being overweight. Of particular interest for the remainder of this unit, is identifying factors in children that might be indicators that a child may become overweight.

Preparation Homework for Lesson 2

1. Identify factors that you think influence a child's weight, particularly factors that might lead to the child being overweight.

2. Which of the factors listed in part 1 can be controlled?

Practice

1. Use the data in the following table.

Child Number	**1**	**2**	**3**	**4**	**5**
Weight (kg) at age four	19.9	21.0	19.9	14.0	15.2
Weight (kg) at age six	20.4	37.3	25.6	17.5	20.0

a. What is the mean weight (kg) of these children at age 4? At age 6?

b. What is the median weight (kg) of these children at age 4? At age 6?

c. Are the weights more spread out at age 4 or at age 6? How did you determine this?

d. Make dotplots of each set of weight data. Arrange the number lines for the two plots one above the other and use the same scale on both number lines. Interpret what your plots tell you about the weights at age four compared to the weights at age six.

2. The femur is the longest bone in the body. However, femur length varies from person to person. The box below contains data on 26 femur lengths (measured from skeletal remains).

470	490	467	444	470	456	512	524	497	447	472	480	446
455	426	482	435	458	483	476	383	443	437	520	442	432

a. Calculate the mean and median of the femur lengths. Show your calculations.

b. Calculate the range.

c. Make a dotplot of the femur lengths.

d. **Outliers** are data values that tend to depart from the overall pattern of the rest of the data. Do any of the femur lengths appear to be outliers? If so which one(s)?

Lesson 2 The Infant Growth Study

In Lesson 1 you used the mean, median, and range to summarize the center and spread of the team age estimates. In addition, dotplots helped you visualize both where the data were centered and how spread out they were. In this lesson, you will extend your statistics toolbox to help you analyze more complex data set.

At the end of Lesson 1, you determined factors that helped you guess a person's age. You also listed factors that might be related to a child's being overweight. One study, called the Infant Growth Study (IGS), investigated the linkage between eating behavior and being overweight. In the remaining lessons in this unit, you will work with a subset of data collected for the Infant Growth Study. Based on results from analyses of these data, you will determine which factors might be related to childhood obesity.

Introduction to the Infant Growth Study (IGS)

The Infant Growth Study sought to link eating behaviors (risk factors) to children being overweight. Data (such as height, weight, food diaries, etc.) were collected on the children at various times during their childhood. When the children were four years old, they participated in eating a meal. Some of the data collected during this meal is contained in the data set you will work with in the remainder of this unit.

The IGS is a scientific study. Therefore, the conditions present at the test meal had to be carefully controlled. For example, researchers instructed parents not to feed their children for at least four hours prior to the meal. Parents were present at the meal, but did not eat any of the food. Parents were instructed to interact with their children as they would normally during any meal. The food was carefully weighed before it was given to the children. The meals contained familiar foods that were typical of what the children usually ate. However, the sizes of the servings were larger than usual and the children were allowed to choose from a variety of foods (such as spaghetti, hamburgers, chicken, and so forth). A concealed camera was set up to videotape the entire meal. After the meal, research assistants, trained in a standard method for classifying a certain set of behaviors, viewed each video four times and recorded the number of occurrences of each of the following behaviors.

- Parental prompt encouraging the child to eat
- Parental prompt discouraging the child from eating
- Child's request for food
- Child's refusal of food
- Mouthfuls of food eaten
- Duration of meal

Questions for Discussion

1. Why did the researchers tell the parents not to allow their child to eat four hours before the test meal?

2. Why did the researchers ask the parents not to eat during the test meal?

3. Why were the research assistants trained in a standard method of how to classify certain types of eating behaviors?

4. Why would the research assistants record the data based on four viewings of each video.

In addition to the data on eating behaviors, researchers collected a variety of physical measurements from the children both at the time of the meal when the children were four, and again when the children were six. These measurements included height, weight, waist circumference and skinfold thickness.

Lessons 2 – 4 are based on a subset of the data collected for the IGS that includes the following variables:

- GEND: Gender: 1 = female, 2 = male
- HTCM4: Height (cm) at age 4
- WTKG4: Weight (kg) at age 4
- HTCM6: Height (cm) at age 6
- WTKG6: Weight (kg) at age 6
- TSEC: Length of meal (sec)
- MFLS: Mouthfuls of food consumed during meal
- KCAL: Calories consumed during meal

Choice of variable names is very important – the names should help the researcher remember what the data values represent. The variable names above were restricted to five characters because that is the limit for naming calculator lists. For example, the variable named HTCM4 uses an abbreviation for height (HT), the units used to measure the height (CM), and the child's age at the time height was measured (4).

Research questions guide the analysis of data in any scientific study. The remainder of this lesson will focus on the following research question:

What are key factors in identifying children as being overweight or as being at risk for becoming overweight?

This is an important question. Before you can study factors related to being overweight, you first need to define what it means for a child to be overweight.

Question for Discussion

5. Suppose you are told a person weighs 150 pounds. Is the person overweight?

Five-Number Summary

Certainly a child's weight at a particular age is a key factor in determining whether or not the child is overweight. The ISG children's weights (kg) at age four, ordered from largest to

smallest, appear in Table 2.1. Notice how difficult it is to extract much information from 51 numbers even after they have been organized from smallest to largest. Learning to summarize data with a few key numbers and then to display those numbers graphically will provide a useful means of extracting information from these data.

12.13	14.03	14.13	14.27	14.30	14.43	14.60	14.77	15.00	15.17
15.23	15.23	15.27	15.53	15.60	15.67	15.67	15.73	15.77	15.80
15.90	16.07	16.20	16.20	16.37	16.40	16.43	16.47	16.53	16.60
16.73	16.87	17.20	17.20	17.33	17.40	17.80	17.90	17.93	18.30
18.50	18.77	18.77	19.27	19.93	20.30	20.97	21.20	21.87	22.30
24.67									

Table 2.1: Weights (ordered from smallest to largest) of IGS children at age four.

Three of the numeric summaries introduced in Lesson 1 were the minimum, median, and maximum. The minimum weight is 12.13 kg and the maximum weight is 24.67 kg. There are 51 children in this study, so the median is the (51+ 1)/2 or 26th data value, which is 16.40 kg. These three values have been highlighted in Table 2.1.

The median is the middle number. It is also the 50th percentile because, in an ordered list, roughly half of the data fall at or below the median and roughly half fall at or above the median. Additional percentiles can be useful in unlocking more information from the data. The p^{th} **percentile** is a value such that, in an ordered list, p percent of the data fall at or below this value.

Two commonly used percentiles are the 25th and 75th percentiles. These percentiles are referred to as the **first quartile, Q1,** and **third quartile, Q3**, respectively. Note that the median is the second quartile. Roughly one-quarter (25%) of the data will fall at or below Q1 and three-quarters (75%) of the data will fall at or below Q3.

Calculating Q1 and Q3*

1. Arrange the data in order from smallest to largest.
2. Determine the median in the ordered list. Use the median to divide the ordered data into two equally sized groups -- a lower half and an upper half.
3. The first quartile, Q1, is the median of the lower half.
4. The third quartile, Q3, is the median of the upper half.

*The algorithm in this box may produce results that differ somewhat from results computed using software packages such as Excel.

To find the quartiles Q1 and Q3 for the data in Table 2.1, divide the data in half about the median. The data values to the left of the median (12.13 to 16.37) are the lower half and the data values to the right of the median (16.43 to 24.67) are the upper half. To find Q1 and Q3, determine the medians of the lower half and upper half, respectively. Q1 (15.27 kg) and Q3 (17.93 kg) have been underlined in Table 2.1.

Listing the five numbers discussed so far gives a **five-number summary** of the weight data from Table 2.1. A five-number summary for a data set consists of the minimum, the first quartile (Q1), the median, the third quartile (Q3), and the maximum.

Min, Q1, Median, Q3, Max: 12.13, 15.27, 16.40, 17.93, 24.67

So now instead of struggling with 51 data values, we can focus on 5 key values. Notice that the five-number summary divides the data into four parts, the first or lower quarter, the second quarter, the third quarter, and the fourth or upper quarter. Each of these quarters contains roughly the same number of data.

Box-and-Whisker Plot (Boxplot)

A **box-and-whisker plot** or **boxplot** is a graph of the five-number summary. The inner half of the data are represented by a box drawn from Q1 to Q3. The box is divided at the median. Whiskers connect the sides of the box to the minimum and maximum. Figure 2.1 shows a box-and-whisker plot for the weight data from Table 2.1. It lets us visualize the location and spread of each quarter of the data.

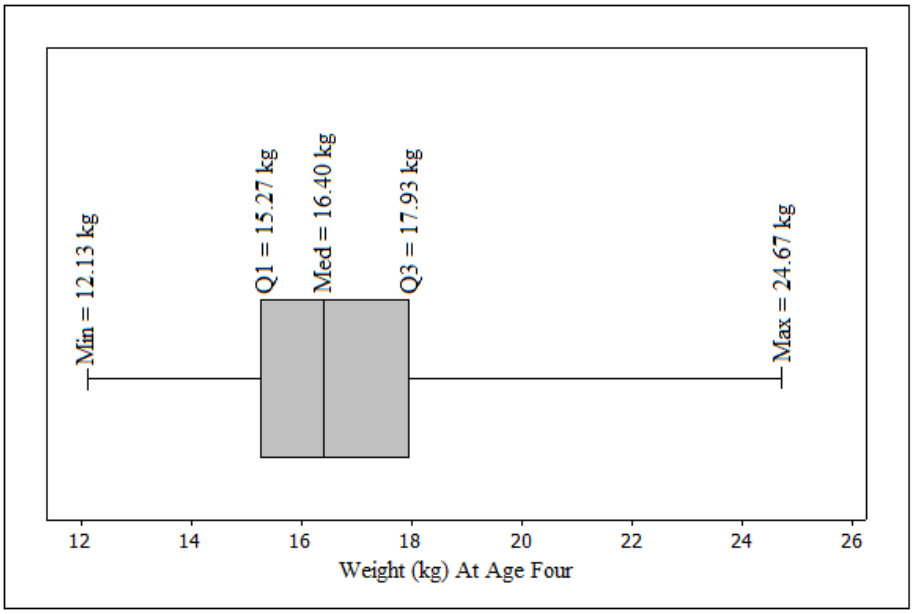


Figure 2.1: Boxplot of weight data – graphic display of five-number summary.

The length of the left whisker represents the spread of the lower quarter and the length of the right whisker the spread of the upper quarter. In Figure 2.1 notice that the length of the right whisker is long in comparison to the length of the left whisker. That means that the upper quarter is more spread out than the lower quarter. Keep in mind that the long right whisker might be due to a single outlier. In other words, one four-year old who is considerably heavier than other four year olds could be responsible for the long right whisker. If that outlier were removed, the two whiskers might have similar lengths. We need a way of identifying outliers. That way we can separate out the overall pattern of the data and data values that depart from the pattern of the rest of the data.

In a boxplot, classification of data points as outliers depends on the width of the box or the spread of the inner half of the data. The **interquartile range** (iqr) is the difference between the

first and third quartiles: iqr = Q3 – Q1. The iqr is also the width of the box in a box-and-whisker plot. The width of the box for the boxplot in Figure 1 is 17.93kg – 15.27 kg = 2.66 kg.

1.5×iqr Rule for Identification of Outliers

A data value is identified as an outlier if it falls more than 1.5 × iqr below the first quartile or above the third quartile.

For the weight data, 1.5×iqr = 1.5×2.66 kg = 3.99 kg. The cutoffs for the outliers are:

Q1 – 1.5×iqr = 15.27 kg – 3.99 kg = 11.28 kg
Q3 + 1.5×iqr = 17.93kg + 3.99 kg = 21.92 kg.

Any weight less than 11.28 kg or any weight greater than 21.92 kg is an outlier. Looking at the ordered data in Table 2.1, there are no weights that fall below 11.28 kg. However, there are two weights that fall above 21.92 kg: 22.30 kg and 24.67 kg.

Modified Boxplot

In a modified boxplot, the whiskers extend out from the box to the smallest and largest data values that are not identified as outliers by the 1.5 × iqr rule. Outliers are plotted as individual points. For the weight data, the upper whisker will be shortened so that it ends at 21.87, the largest weight that is not flagged as an outlier by the 1.5 × iqr rule. Figure 2.2 shows the modified boxplot for the weights in Table 2.1. The modified boxplot displays the overall pattern of the data and the outliers that depart from that pattern.

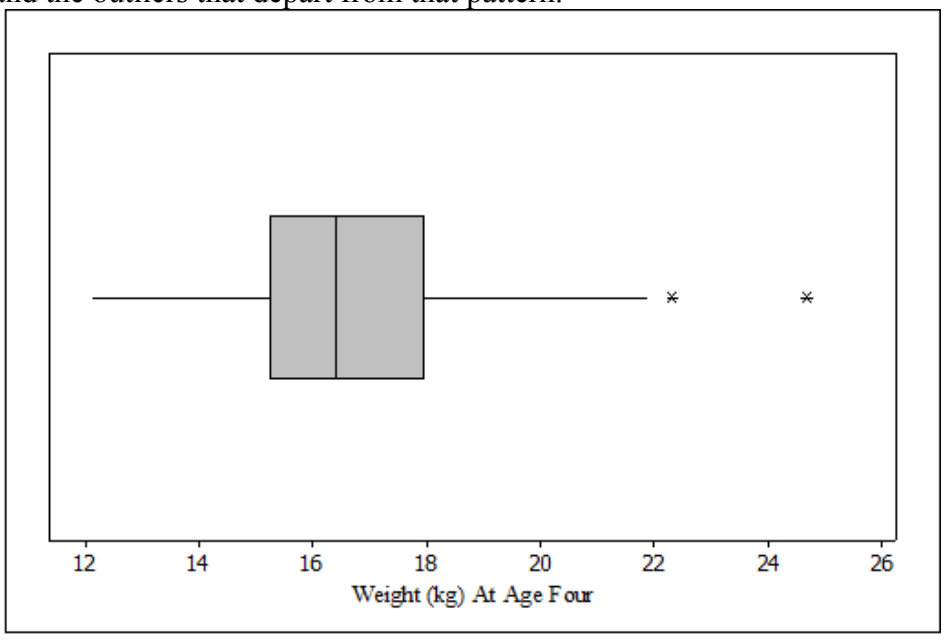


Figure 2.2: Modified boxplot of weight at age 4.

Searching for Overweight Factors

Armed with new statistical tools for summarizing data and for identifying and displaying outliers, you are now ready to return to this lesson's research question. Using weight at age four as a factor and the boxplot in Figure 2.2, two four-year-old children (with IDs 137 and 160) have weights that are flagged as outliers. These children are at risk of being overweight because their weights depart from the overall pattern of the weights of the other four year olds. In the activity Weighty Factors, you will continue the search for other, perhaps better, factors.

ACTIVITY 2-1 Weighty Factors

Objective: Use statistical tools to summarize and display data and look for other factors to explain overweight risk.

Materials:

Handout BS-H5: Weighty Factors Worksheet

Part I: Analyzing Weight-At-Age-Six Data

Table 2.2 contains the weights at age six for the IGS children. As you analyze these data, keep in mind the following two additional questions concerning trends related to being overweight.

- Will children who are overweight at age four remain overweight at age six?
- Will a higher proportion of the IGS children be overweight at age six compared to age four? (In other words, does being overweight become more prevalent as children get older?)

16.30	17.03	17.10	17.50	17.70	17.90	18.00	18.10	18.30	18.80	18.90	19.17
19.40	19.50	19.60	19.75	19.80	19.83	19.90	20.03	20.03	20.20	20.27	20.37
20.47	20.83	21.00	21.00	21.27	21.30	21.80	21.87	21.90	22.20	22.47	22.50
22.73	22.77	22.93	23.00	23.40	23.50	23.70	25.00	25.10	25.60	26.47	34.23
37.27	38.20	44.77									

Table 2.2: Weights (ordered from smallest to largest) of IGS children at age six

1. The weights at age six for the IGS participants appear in Table 4. These data have been arranged from smallest to largest.

a. Determine a five-number summary for the age-six weights. Show any calculations needed to arrive at your answer.

b. Use the 1.5×iqr rule to determine which of the weights should be identified as outliers.

c. Draw a modified box-and-whisker plot. Describe any interesting features in the plot. For example, does the boxplot appear roughly symmetric? Does one quarter of the data tend to be more spread out than another quarter of the data?

d. Find the ID numbers of the children whose weights at age 6 were flagged as outliers. Recall that children with ID 137 and ID 160 had weights at age four that were flagged as outliers. Were their weights at age six among the outliers?

For the next part of this activity, you will need to use technology – a graphing calculator or spreadsheet software. You will need to access the Infant Group Study Data. Before moving to Part II, check your answers to Part I using your calculator or spreadsheet.

Part II: Analyzing Change-In-Weight Data

2. The IGS participants were weighed at ages four and six. Based on the analysis of the weight data, the number of outliers doubled from age four to age six. This result points to another factor that might identify those at risk for becoming overweight, namely the amount of weight gained from age four to age six.

a. What is the average weight of the IGS children at age four?

b. What is the average weight of the IGS children at age six?

c. Find the difference between the average weights from age four to age six.

3. In addition to a child's weight, doctors also consider the change in a child's weight between visits when assessing the child's health. Explain why weight gain might be an important factor linked to being overweight or to becoming overweight.

4. Form a new variable CWTKG that gives the change in weight from age four to age six in kg.

a. Determine the values of CWTKG for the data set and then find the mean of CWTKG. Compare this result to your answer to 2(c).

b. Determine a five number summary for the CWTKG data (the change in weight from age four to age six). Hint: You will have to order the CWTKG data first.

c. Use the 1.5×iqr rule to identify weight gains that are outliers.

d. Create a modified boxplot to display the information from parts (a and b).

e. Find the ID numbers corresponding to the weight gains that you determined to be outliers.

f. Are the outliers using the change in weight data the same children whose weights at age six were outliers? Did this analysis identify any additional children who might be at risk of becoming overweight? Explain.

Part III: Analyzing Test-Meal Data

Recall that the IGS children were treated to a meal when they were four years old. The purpose of the meal was to collect data on eating behaviors of the children so that researchers could study the linkage between eating behaviors and being overweight. You will now focus on the amount eaten during the meal, measured in mouthfuls, and the length of time the children spent eating.

5. The data on the number of mouthfuls is in the column or list MFLS.

a. What is the average (mean) number of mouthfuls of food eaten by the children at the test meal?

b. Determine Q1 and Q3. Use the 1.5×iqr rule to identify any outliers.

c. Find the ID number(s) corresponding to any outliers that you found. Check to see if these IDs appear on any of the previous lists of IDs for outliers.

6. The researchers in the Infant Growth Study hypothesized that the rate at which the children ate at the test meal might be more closely linked to being overweight than the amount eaten.

a. Form the variable MPM, the number of mouthfuls eaten per minute. Note that the duration of the meal is in *seconds*, but the requested rate, MPM, is mouthfuls per *minute*. Determine the values for MPM for the data set.

b. Determine Q1 and Q3. Then use the 1.5×iqr rule to identify outliers in the MPM data.

c. Find the IDs corresponding to the outliers. Were any of these children identified as outliers for weight gain from ages four to six?

Now that you have completed this lesson, you have a better idea of factors related to children being overweight.

Question for Discussion

6. What relationships have you identified between the variables in the IGS data and children who may be overweight or at risk of being overweight?

Preparation Homework for Lesson 3

1. Do you think there is a connection between a child's height and their weight? Explain.

2. Figure 2.3 shows a plot of the ordered pairs (height, weight) for each of the IGS children at age six. This plot is called a **scatterplot** of weight versus height. Does the pattern of dots in the scatterplot support your answer to question 1? Explain.

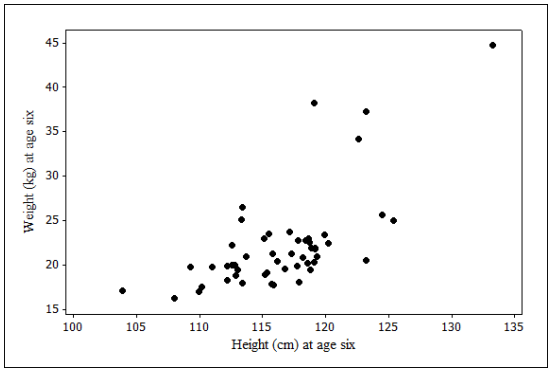


Figure 2.3: Plot of weight versus height.

3. The **body mass index**, **BMI**, provides one measure of a person's "thickness." The calculation of a person's BMI involves both the person's weight and height. The formula for BMI in metric units is given below:

$$\text{BMI} = \frac{\text{weight (kg)}}{(\text{height (m)})^2}$$

a. Form two new variables BMI4 and BMI6, the body mass index for the IGS children at age four and at age six, respectively. Notice that in the formula for BMI, height is in meters. Height for the IGS study is reported in centimeters. Determine the values of these variables for the children in the IGS study. You will need these values in Lesson 3.

b. Determine Q1 and Q3 for BMI6. Use the 1.5×iqr rule to identify potential outliers.

c. Determine the ID numbers that correspond to the outliers that you found in (b). In addition, write the child's gender.

d. Compare IDs of the six-year-old IGS children who were identified as outliers using BMI to the IDs of the six-year-old IGS children who were identified as outliers using weight gain. Is there much overlap in these two groups?

e. Finally, make a modified boxplot for BMI at age 6. Label each of the outliers with its corresponding BMI.

4. In Part III of Activity 2-1, you investigated the amount of food consumed by the children at the test meal (measured by mouthfuls eaten). However, weight may also be associated with the types of food that the children chose. Did they choose high Calorie foods or low Calorie foods?

a. Determine a five number summary for KCAL.

b. Using the 1.5×iqr rule, would any children have been identified as outliers?

Practice

1. A number of studies have described the health benefits of eating fruit. The Penn State Young Women's Health Study researched the connection between eating fruit, fitness, and heart health. Table 2.3 contains data on the average number of daily of servings of fruit eaten by 74 seventeen-year-olds girls who were part of this study. (Data have been truncated to whole servings.)[1]

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4
5	5	5	5	5	6	6	6	7	7	7	8	8	8						

Table 2.3: Average servings of fruit eaten per day.

a. Make a dotplot for these data. Describe its shape.

b. Based on the shape of the data, would you expect the mean or median to be larger? Explain.

c. Calculate the mean and median. Do your calculations confirm your answer to (b)?

2. Use Table 2.3

a. Determine a five-number summary for the data on average servings of fruit per day. Explain how to get the answer by hand.

b. Represent your five-number summary with a modified boxplot.

c. Based on your boxplot, what can you say about the variability of each of the four quarters of the data? Explain.

3. A man in a nursing home has his pulse taken every day. His pulse readings over a one-month period appear in Table 6.

72	56	56	68	78	72	70	70	60	72	68	74
76	64	70	62	74	70	72	74	72	78	76	74
72	68	70	72	68	74	70					

Table 2.4: Data on pulse beats per minute.

a. Draw a dotplot for these data.

b. Based on the dotplot do you think the mean or median will be larger? Explain.

c. Calculate the mean and median. (Be sure to include the units in your answer.) Do these calculations confirm your answer to (b)?

d. Based on these data, which better describes the center of the pulse data, the mean or the median? Explain your reasoning.

4. a. Determine a five-number summary for the pulse data in Table 2.4.

b. Use the 1.5×iqr rule to determine if any of the pulse values should be considered outliers. How many of the pulse readings are outliers?

c. Represent the pulse data with a modified box-and-whisker plot.

d. Which whisker is longer, the right or left whisker? What does that tell you about these data?

Lesson 3 Exploring Relationships Using Regression

As you discovered in Lesson 2, identifying a child as being overweight or at risk of becoming overweight is more complicated than simply checking the child's weight. Finding a factor useful in identifying overweight children often means combining more than one variable. For example, the factor CWTKG, change in weight from age four to age six, is a combination of two variables. This factor might prove useful in identifying children who are at risk of becoming overweight. The preparation homework for Lesson 3 introduced a new factor, **body mass index** (BMI), which is a combination of weight and height.

Displaying a Relationship Between Two Variables

In Lesson 2 you observed that four-year olds whose weights were flagged as outliers also had weights that were flagged as outliers when they were six. So, there may be a relationship between the two variables weight at age four (WTKG4) and weight at age six (WTKG6). If there is a relationship, it could help answer the following question: Do heavier four year olds tend to become heavier six year olds? Figure 3.1 shows a **scatterplot** that displays this relationship. Notice that weight at age four is on the horizontal axis. It's on the horizontal axis because we believe that children's weights at age four might be useful in "explaining" children's weights when they are six. It is mathematical convention that the **explanatory variable** is on the horizontal axis and the **response variable** on the vertical axis.

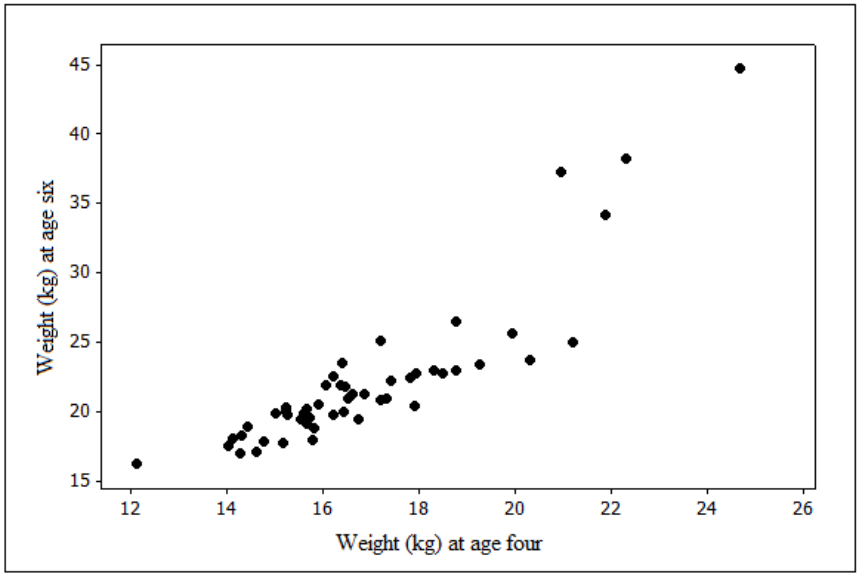


Figure 3.1: Scatterplot of weight at age six versus weight at age four.

In a scatterplot of data, it is customary to put the explanatory variable on the horizontal axis. This variable helps "explain" what values we might expect from the variable on the vertical axis, which is called the response variable.

Notice that as your eye scans the scatterplot in Figure 3.1, the dots tend to be lower on the left side of the plot and higher on the right. This pattern indicates that children who are lighter at age

four tend to be lighter at age six and children who are heavier at age four tend to be heavier at age six. Because of the direction of this pattern, the variables weight at age four and weight at age six are said to have a **positive association**. Two variables are positively associated if, as one variable increases, the other tends to increase. Two variables are **negatively associated** if, as one variable increases, the other tends to decrease. Figure 3.2 shows examples of positive and negative associations.

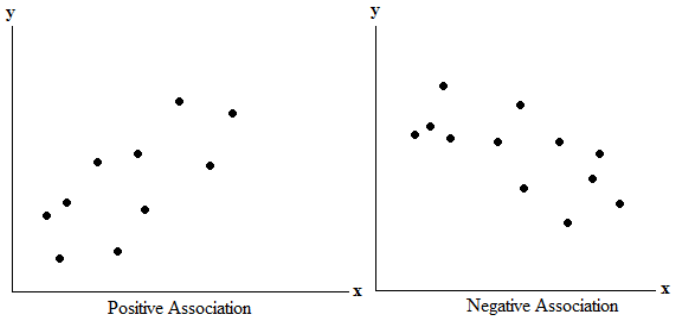


Figure 3.2: Examples of Positive and Negative Association

Questions for Discussion

Figure 3.3 shows a scatterplot of BMI at age six versus BMI at age four. Recall that BMI is $\text{BMI} = \frac{\text{weight (kg)}}{(\text{height (m)})^2}$.

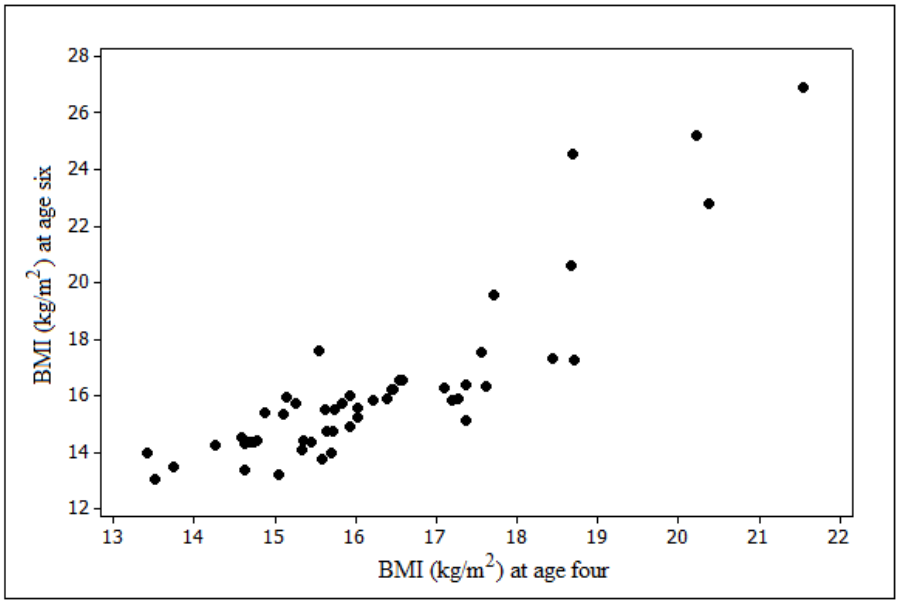


Figure 3.3: Scatterplot of BMI at age six versus BMI at age four.

1. Explain why the body mass index, BMI, is considered a better factor for identifying children who are overweight than weight.

2. In Figure 6 why do you think the researchers placed BMI at age four on the horizontal axis?

3. Is the relationship between BMI at age four and BMI at age six an example of positive association or negative association? Explain.

4. If a child's BMI at age four indicates that the child is overweight, his or her BMI can't be instantly changed. However, researchers can investigate what factors might have an impact on the child's future BMI measurements. Consider the following research question: Are eating behaviors at age four related to BMI at age six?

Generally when researchers pose a question, they have an idea of the answer, a **hypothesis**. The researchers look to the data to see if the data support their hypothesis. What is your idea for an answer to this research question? Explain.

In discussion question 1, you explained why BMI might be a better factor than weight for identifying children who are overweight. Next, we use the data to check some of the assumptions you may have made in answering question 1. Return to Figure 3.3 that shows graphically the positive association between weight at age six and weight at age four. Most people would expect that taller children tend to weigh more than shorter children. Now consider the question: Do taller four year olds also tend to become taller six year olds?

If this is true, then the relationship between weights at ages four and six might be due to the relationship between heights at ages four and six. So, the heavier children might not be overweight – instead, they might be a normal weight for a tall child.

Summarizing a Pattern in a Scatterplot with a Line

Figure 3.4 shows a scatterplot of height at age six, *y*, versus height at age four, *x*. Since the pattern of dots in the scatterplot appears to be linear, a **line of best fit**, also known as a **regression line,** has been added to the scatterplot. This line summarizes the relationship between the two variables.

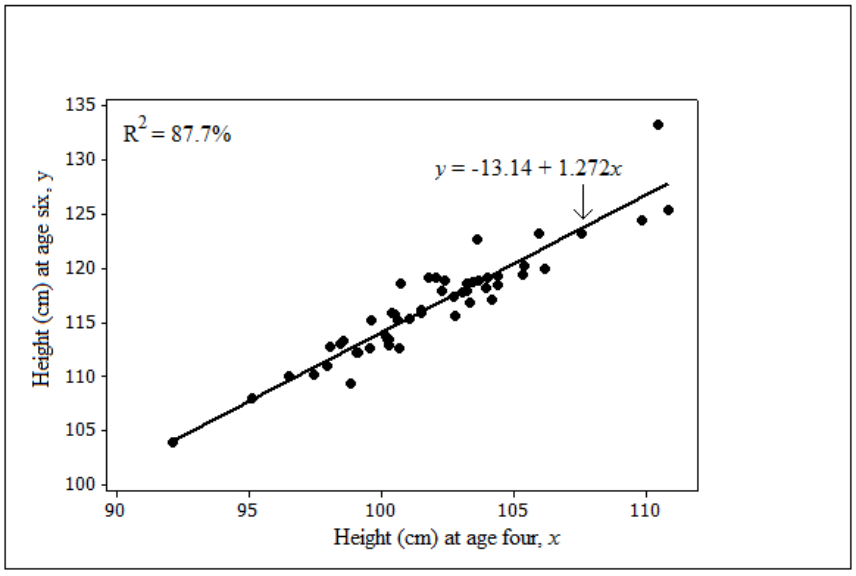


Figure 3.4: Height at age six versus height at age four.

The regression equation, $y = -13.14 + 1.272x$, can be used to predict children's heights at age six as soon as their heights at age four are known. For example, the child with ID 106 was 98.07 cm tall at age four. Substituting 98.07 in for x in the regression equation gives the following estimate of her height at age six:

$$y = -13.14 + 1.272(98.07) \text{ or about } 111.61 \text{ cm.}$$

In this case, we know that the child's true height at age six was 112.80 cm. So, we can calculate how far off our estimate is from the actual value. This is called the **residual error** or simply the **residual**. A residual is the difference between the observed y-value and the y-value predicted by the regression line.

$$\text{residual} = \text{observed } y - \text{predicted } y$$

So, in this case:

$$\text{residual} = 112.80 \text{ cm} - 111.61\text{cm} = 1.19 \text{ cm}$$

Graphically, the residual is the directed vertical distance from the observed point to the line.

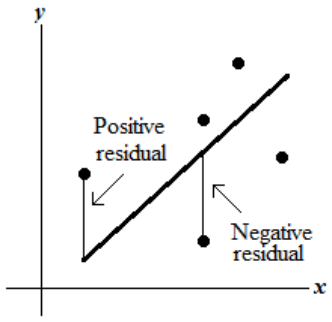


Figure 3.5: Geometric description of residuals.

The line of best fit in Figure 3.4 can be calculated using technology such as graphing calculators, spreadsheets, or statistics software. It is a best-fitting line because its sum of squares of the residuals is as small as possible. In other words, it is the line that makes the sum of the squares of the vertical distances of the data points to the line as small as possible.

Look again at Figure 3.4. The IGS children's heights at age six vary quite a lot – from a minimum of around 104 cm to a maximum of around 133. The range is 29.4 cm. Notice that the dots in the scatterplot form a very tight pattern about the line. Hence, heights at age four explain much about heights at age six. The value of the **coefficient of determination**, R^2 (reported in the upper left corner of Figure 3.4), means that 87.7% of the variability in height at age six can be explained, using this best-fit line, based on height at age four.

The coefficient of determination, R^2, is the percentage of the variation in one variable that is explained by the regression line on the other variable. If R is 100%, then the data points fall exactly on a line. Any time R^2 is greater than around 64%, the relationship between the two variables is considered strong. Therefore, the variables height at age four and height at age six have a strongly positive relationship.

Biological connections, such as height at two different ages, often tend to be quite strong. However, don't expect strong relationships when one of the factors is behavioral, such as an eating quickly.

Relationships Between Eating Behaviors and BMI

Suppose a four year old is diagnosed as being overweight based on his BMI. The child can't change his BMI overnight. However, there may be some interventions that might help reduce his BMI when he is six. The Infant Growth Study focused its research on factors related to being overweight *that could be controlled* – such as eating behavior. In the next activity, you will be the biostatistician in charge of investigating relationships between eating behavioral factors and being overweight. It will be your job to determine which, if any, of the factors might explain children's being overweight or being at risk for becoming overweight.

Activity 3-1 Eating Behavior and BMI

Objectives: Investigate relationships between variables to identify explanatory factors.
Materials:

Handout BS-H7: Eating Behavior and BMI Worksheet
Graphing calculator or spreadsheet

By the end of this activity, you will have made some progress in answering the following research question:

Are eating behaviors at age four related to being overweight at age six? More specifically, are any factors associated with the test meal related to being overweight at age six?

Part I: Examining the relationship between BMI at age four and BMI at age six.

Before tackling the research question, first consider the relationship between BMI at age four and BMI at age six. Recall that BMI is considered a good factor for determining if a person is overweight because it takes into account both the person's height and weight.

1. Figure 3.4 shows a scatterplot of BMI at age six versus BMI at age four.

   a. Using technology (a graphing calculator or spreadsheet software), make a scatterplot of BMI at age six versus BMI at age four. Make sure that BMI at age four is on the horizontal axis.

   b. Determine the equation of the best-fit line (the regression equation).

   c. Report the value of R^2 and interpret its value.

   d. Use information from the regression equation to fill in the blank: For each 1 kg/m^2 increase in BMI at age four, the BMI at age six is expected to increase by ____ kg/m^2.

2. Return to the regression equation from question 1.

   a. The Child with ID 106 had a BMI at age four of 15.8353 kg/m^2. Use the regression equation to estimate her BMI at age six. Show your calculations.

   b. This child's actual BMI at age six was 15.7421 kg/m^2. Calculate the residual.

   c. In this case, was the estimate from the regression equation an overestimate or an underestimate of the actual BMI at age six? How could you answer this question from looking at the value of the residual?

   d. Suppose a residual turned out to be negative. Did the regression equation overestimate or underestimate the actual BMI at age six. Explain.

Part II: Relationships between eating behaviors and BMI

Data was collected on the IGS children at a test meal when they were four. The unit data set includes three of the test meal variables: duration of meal (TSEC), mouthfuls eaten during meal (MFLS), and Calories consumed during meal (KCAL). You created a fourth variable in Lesson 2 that measured how fast the children ate, measured in mouthfuls per minute (MPM).

KCAL stands for kilocalories, which is the same as Calories (with a capital C). If a lowercase c is used, calories, it means $1/1000^{th}$ of a Calorie. When we diet, it is Calories that we count and Calories that are listed on food items.

You will investigate which, if any, of these test-meal factors can be used to explain the variability in BMI at age six.

3. Focus first on the factor MFLS, the number of mouthfuls consumed during the test meal. This factor is related to the quantity of food that was consumed during the test meal.

a. Make a scatterplot of BMI6 versus MFLS. (Since you want to know if MFLS is helpful in explaining BMI, MFLS is the explanatory variable (a factor) and should be on the horizontal axis.) Would you describe the relationship between MFLS and BMI6 as positive, negative, or neither? Explain.

b. Fit a line of best fit the variables MFL and BMI6 (in other words to the variables mouthfuls of food eaten during the test meal and BMI at age six). Add a graph of this line to your plot in (a). Write its equation.

c. Does this line appear to do a good job in describing the overall pattern in the data? Explain.

d. What percentage of the variability in BMI at age six can be explain by the amount of food that children ate at the test meal? Do you think that MFLS is a useful factor in explaining BMI at age six? Explain.

4. Next, consider the rate at which the children ate during the test meal, MPM.

a. Make a scatterplot of BMI6 versus MPM. (Hint: Since, you want to know if MPM is helpful in explaining BMI, MPM should be on the horizontal axis.) Would you describe the relationship between MPM and BMI6 as positive, negative, or neither? Explain.

b. Fit a line of best fit MPM, mouthfuls per minute, and BMI6, BMI at age six. Add a graph of the line to your plot in (a). Write its equation.

c. What percentage of the variability in BMI at age six can be explain by the rate at which the children ate (MPM)?

5. In helping children with high BMIs at age four achieve a healthy BMI at age six, which of the following interventions might be better. Support your choice using information from your answers to questions 2 and 3.

Intervention 1: Tell children to eat less (eat fewer mouthfuls at a meal).
Intervention 2: Tell children to eat slower (change how fast they eat).

Practice

Continue your investigation into which factors from the test meal might be related to BMI at age six. Use technology, either a graphing calculator or spreadsheet software, to complete these practice problems.

1. First, focus on duration of the meal, TSEC. You will investigate whether children who spent more time eating tended to have higher BMIs at age six (BMI6) than children who spent less time eating.

a. Make a scatterplot of BMI6 versus TSEC. Would you describe the relationship between the duration of the meal and BMI at age six as positive, negative, or neither?

b. Does your answer to (a) tend to confirm or refute the hypothesis that children who spent more time eating tended to have higher BMIs at age six than children who spent less time eating? Explain.

c. Fit a line of best fit to duration of meal, *x*, and BMI at age six, *y*. Add a graph of the line to your plot in (a). Write its equation.

d. Does this line appear to do a good job in describing the overall pattern in the data? Explain.

e. What percentage of the variability in BMI at age six can be explain by the duration of the test meal?

2. Next, you will investigate whether children who consumed more Calories (KCAL) during the test meal tended to have higher BMIs at age six (BMI6) than children who consumed fewer Calories. (Save your results for use in Lesson 4, Activity 3, question 4(e).)

a. Make a scatterplot of BMI6 versus KCAL. Fit a line of best fit to your data. Add its graph to your scatterplot. Write the regression equation.

b. Does your answer to (a) tend to confirm or refute the hypothesis that children who consumed more Calories during the test meal tended to have higher BMIs at age six than children who consumed fewer Calories? Explain.

c. What percentage of the variability in BMI at age six can be explain by the Calories consumed during the meal?

3. Look at the rate at which Calories children consumed calories during the test meal. Investigate whether children who consumed Calories at a faster rate tended to have higher BMIs at age six than children who consumed Calories at a lower rate.

a. Create a new variable CALPM, Calories consumed per minute.

b. Make a scatterplot of BMI6 versus CALPM. Fit a regression line to the data. Write its equation.

c. Does the information you gathered from (b) tend to confirm or refute the hypothesis that children who consumed Calories at a faster rate tended to have higher BMIs at age six than children who consumed Calories at a lower rate. Explain.

d. What percent of the variability in BMI at age six can be explained by the regression line based on the factor Calories consumed per minute?

4. In Lesson 2 you observed a positive association between height at age six and weight at age six. That positive association supported the hypothesis that taller children weigh more than lighter children. Consider the two variables, height and weight at age four.

a. Make a scatterplot and fit a least-squares regression line to weight at age four versus height at age four. Write the equation of the regression line.

b. Does your scatterplot from (a) tend to support or refute the hypothesis that taller children tend to be heavier than shorter children? Explain.

c. How variable are IGS children's weights at age four? Answer this question by calculating the range of weight.

d. What percentage of the variability in weight at age four can be explained by the regression line based on height at age four?

e. A child in the IGS data set (ID114) was 96.50 cm tall at age four. Use your regression equation to estimate his weight at age four. Show your calculations.

f. The actual weight at age four for the child in (e) was 14.27 kg. Determine the residual. (In other words, how far off was your estimate in (e)?)

5. A woman in a nursing home takes blood pressure medication. Her blood pressure is taken daily. Both her systolic and diastolic blood pressure readings for the month of February appear below.

Date	2/1	2/2	2/3	2/4	2/5	2/6	2/7	2/8	2/9	2/10
Systolic	150	148	136	120	142	144	130	150	130	142
Diastolic	72	68	72	64	70	72	68	70	70	74
Date	2/11	2/12	2/13	2/14	2/15	2/16	2/17	2/18	2/19	2/20
Systolic	140	130	148	142	120	166	120	136	130	148
Diastolic	72	70	70	72	60	72	60	72	70	70
Date	2/21	2/22	2/23	2/24	2/25	2/26	2/27	2/28		
Systolic	136	172	130	128	140	130	150	152		
Diastolic	72	74	60	60	60	70	70	74		

Systolic and diastolic blood pressure readings (mmHg).

a. Make a scatter plot of the systolic versus the diastolic blood pressure readings. (Put systolic blood pressure on the vertical axis and diastolic blood pressure on the horizontal axis.)

b. Fit a line to the data in (a). What is the equation of your line? Be sure to define what your variables *x* and *y* represent.

c. Suppose on March 10th the woman's systolic blood pressure was not recorded. Her diastolic blood pressure was 62 mmHg. Use your equation from (b) to estimate the woman's systolic blood pressure.

d. Suppose the woman's diastolic blood pressure increased by 10 mmHg from one day to the next. Use your equation in (b) to estimate the rise in her systolic blood pressure. Explain how you got your answer.

Lesson 4 Exploring Relationships Using Two-WayTables

In Lesson 3 you used BMI as a factor for identifying overweight children. Because BMI takes into account two factors, a person's height and weight, it is considered a better factor than weight alone for determining if a person is overweight. However, BMI alone is not a sufficient for determining whether or not a child is overweight. That's because BMI doesn't take into account the child's age or gender.

A Four-Factor Approach

The Centers for Disease Control (CDC) provides growth charts that doctors can use to determine whether a child is overweight. Because there are separate charts for boys and girls, gender is considered as a factor in the CDC growth charts. Using the CDC growth charts for the appropriate gender, a child is considered overweight if his/her BMI is at or above the 85th percentile for his/her age.[2][3]

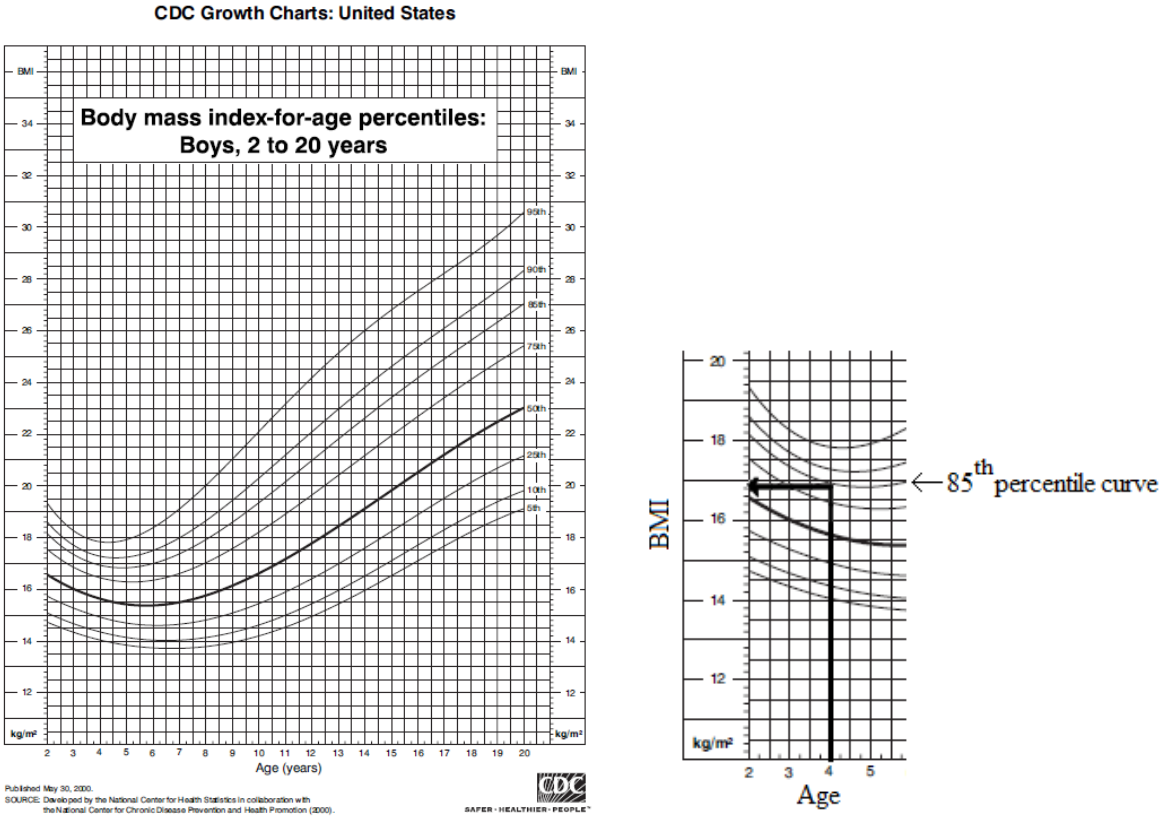


Figure 4.1: CDC BMI Growth Chart for Boys with Enlarged Section of 85th Percentile

The CDC Growth Charts allow the creation of two new variables, overweight at age four (OW4) and overweight at age six (OW6). Each of these is a **binary variable**, meaning there are only two possible outcomes:

- OW4 or OW6 = 1 if the child is overweight at age four or age six respectively
- OW4 or OW6 = 0 if the child is not overweight at age four or age six respectively

Analysis of relationships between binary variables will require new statistical tools involving two-way tables and percent.

Creating Binary Variables OW4 and OW6

Examine the CDC Growth Chart in Figure 4.1. Notice that age (years) is on the horizontal or *x*-axis and that BMI (kg/m^2) is on both sides of the vertical or *y*-axis. In the middle of the chart there is a set of curves. Each curve corresponds to a different percentile of BMI starting with the 5^{th} percentile at the bottom and ending with the 95^{th} percentile at the top. The curve corresponding to the 85^{th} percentile is the third curve down from the top. For a child who is at the 85^{th} percentile it means that there are only 15 out of 100 children (15%) of the same age and gender who have a higher BMI for Age (based on the national representative group from which the CDC developed the charts).

As mentioned above, overweight is defined as "at or above the 85^{th} percentile." According to the CDC chart for boys, a four-year-old boy would be identified as overweight if his BMI is at or above 16.9 kg/m^2, the 85^{th} percentile for four-year-old boys. This cutoff allows us to form a new variable, overweight at age 4 (OW4), which divides four-year-old boys into two groups:

- OW4 = 1 if BMI4 $\geq$ 16.9 (Yes – four-year-old boy is overweight)
- OW4 = 0 if BMI4 < 16.9 (No – four-year-old boy is not overweight)

Because there are only two possible outcomes for this variable, 0 or 1, OW4 is called a **binary variable**.

Questions for Discussion

1. What is the BMI cutoff value for being overweight in boys at age six?

2. What is the BMI cutoff value for being overweight in girls at age four?

3. What is the BMI cutoff value for being overweight in girls at age six?

4. Look at the child with ID 100. At age four, this boy was 101.50 cm tall and weighed 17.90 kg. At age six, he was 116.17 cm tall and weighed 20.37 kg. Determine the value of OW4 and OW6 for this child.

5. Select a girl from the IGS data. Determine the values of the new variables OW4 and OW6 for the selected child.

6. Being overweight at a particular age is defined as having a BMI that is at or above the 85^{th} percentile for age and gender. Assume that the BMI's for today's children are similar to the data on which the CDC Growth Charts (created in 2000) are based. In a sample of 200 randomly chosen six-year-old children, how many of them would you expect to be overweight? Explain.

7. As of 2008, using the CDC Growth Charts to identify overweight children, around 35% of American children ages 6 – 11 are overweight. (It may be even higher today!) What, if anything, does this tell you about the 85th percentile cutoffs used to classify children as being overweight?

Creating a Binary Recommended Calorie Variable (RKCAL)

Sometimes we learn more when we condense information from variables. For example, by forming the binary variable overweight at age six (OW6), we condensed information from BMI6 and gender down to two outcomes, being overweight or not being overweight. We can do the same for other variables such as the Calories consumed during the test meal at age four (KCAL). We can create a new binary variable (RKCAL) that puts the children into one of two categories, those who consumed too many Calories and those who did not.

In Practice question 2 of Lesson 3, you examined the relationship between BMI at age six (BMI6) and the number of Calories children consumed during the test meal at age four (KCAL). The R^2 for that relationship was only 3.3%. So, KCAL explained very little of the variability in BMI6. Using the binary variables RKCAL and OW6 provides another approach to study the relationship between eating behaviors and being overweight.

To create a binary variable for Calorie consumption, we need to determine reasonable cutoffs. The CDC charts provided the cutoffs for forming OW4 and OW6. The US Dietary Guidelines for Americans provide the basis for determining cutoffs for the Calories consumed during the test meal (KCAL). Table 4.1 provides the dietary Calorie recommendations for children.[4]

U.S. Dietary Guidelines: Daily Estimated Calorie Recommendation for Children.

	1 Year	2-3 Years	4-8 Years	9-13 Years	14-18 Years
Calories (kcal)	900 kcal	1000 kcal			
Female			1200 kcal	1600 kcal	1800 kcal
Male			1400 kcal	1800 kcal	2200 kcal

Table 4.1: Daily estimated Calorie recommendations for children.

Note that the Calorie estimates are based on a sedentary lifestyle. Increased physical activity will require additional Calories. Caloric intake should increase by 0-200 kcal/day if moderately physically active, and by 200–400 kcal/day if very physically active.

Question for Discussion

8. Based on the Table of Daily Estimated Calorie Recommendation for Children, what is the daily recommended Calorie consumption (kcal) for children in the IGS dataset?

9. Our data is based on one meal, dinner. If we assume that each child eats 3 meals per day and eats approximately the same number of Calories per meal, what is the Calorie recommendation for dinner?

10. In Practice question 3 of Lesson 2, you determined a 5-number summary for Calories (KCAL). Table 4.2 provides a recap of those results and includes the results broken down for boys and girls separately. Based on Table 4.2, estimate the approximate percentage of boys that

have exceeded the recommended dinner Calories. Then do the same for girls. Use your answers from question 9 for the recommended Calories.

Calories (KCAL)	Minimum	25th Percentile	Median	75th Percentile	Maximum
Overall	145	331	410	558	832
Boys Only	171	368	465	591	734
Girls Only	145	325	379	464	832

Table 4.2: Five-number summaries for Calories eaten at a test meal.

11. The IGS children consist of 27 boys and 24 girls. Based on your answer to 10, what is the approximate number of boys who exceed the recommended dinner Calories? What about for girls?

The binary variable RKCAL is defined as follows based on the recommended dinner Calories for the test meal.

- RKCAL = 1 if KCAL > 400 for girls or KCAL > 467 for boys (Yes – child exceeded recommended Calories at test meal)
- RKCAL = 0 if KCAL ≤ 400 for girls or KCAL ≤ 467 for boys (No – child did not exceed recommended Calories at test meal)

12. Create the variable RKCAL and determine its value for the IGS children in your database.

Two-Way Tables

The IGS investigators designed their study to include a test meal. The purpose of this meal was to collect data to study relationships between certain eating behaviors and being overweight. For example, in Lesson 3 you used a scatterplot to visualize the relationship between KCAL and BMI. It turned out that that analysis didn't produce any interesting results. In this activity, you will tackle the same question again – Is there a relationship between the Calories consumed during the test meal at age four and being overweight at age six? This time you will base your analysis on the binary variables exceeded the recommended dinner Calories (RKCAL) and overweight at age six (OW6).

The remainder of this lesson will focus on relationships between three binary variables, gender (GNDR), overweight at age six (OW6), and exceeded the recommended dinner Calories at age four (RKCAL) in an attempt to answer the two research questions that follow.

1. Who are more likely, girls or boys, to exceed the recommended dinner Calories at age four? (Variables: GENDR and RKCAL)

2. Are children who exceeded the recommended dinner Calories at age four more likely to be overweight at age six than children who did not exceed the recommended dinner Calories at age four? (Variables: RKCAL and OW6)

In your investigation into relationships between two binary variables, you will use **two-way tables** (also called **contingency tables** or **cross-tabulation tables**) to classify all possible

combinations of outcomes from the two variables. Because the relationships we are examining are between two binary variables, you will be working with 2 × 2 two-way tables, tables with 2 rows and 2 columns.

ACTIVITY 4-1 The Tables Have Turned

Objective: Use two-way tables to classify and analyze outcomes.

Materials:

Handout BS-H10 The Tables Have Turned Worksheet

To answer the research question 1, focus on the variables gender (GENDR) and exceeded the recommended dinner Calories at age four (RKCAL). There are four possible ways to classify outcomes from these two binary variables. Table 4.3 organizes these outcomes in a 2 × 2 **two-way table** (the portion with the bold outline). The 2 × 2 dimension of the table comes from the 2 rows corresponding to RKCAL = 0 and RKCAL = 1, and the 2 columns corresponding to GENDR = 1 and GENDR = 2. The numbers entered in the 2 x 2 cells indicate the **frequency** (the number of occurrences) of that particular situation in the data.

Number of Girls and Boys who Exceed (RKCAL = 1) or who Do Not Exceed (RKCAL = 0) the Dinner Calorie Consumption Recommendation

	Gender (GENDR)		
RKCAL	**Girls = 1**	**Boys = 2**	**Total**
Not exceed = 0			
Exceed = 1			
Total	24	27	51

Table 4.3: Two-way table of Gender and RKCAL.

1. Use the two variables GENDR and RKCAL in the IGS data set.

   a. Complete the two-way table in Table 4.3.

   b. Once you have completed your 2 ×2 table, check to see that the column totals are correct. Then add the frequencies in each row and enter these row totals in the last column. Notice that both the sum of the row totals and sum of the column totals must equal the number of IGS children, 51.

Now that you have completed question 1, you have a 2 × 2 (two rows, two columns) table for gender and exceeded recommended dinner Calories. You are ready to tackle research question 1.

2. Use your completed table.

   a. Determine the percentage of girls who exceeded the recommended dinner Calories at age four. Show your calculations.

   b. Determine the percentage of boys who exceeded the recommended dinner Calories at age four. Show your calculations.

   c. Based on your 2 × 2 table and your answers to (a) and (b), were girls or boys more likely to exceed the recommended dinner Calories at age four? Justify your answer by reporting

percentages and frequencies from your 2 × 2 table. (The general format for reporting results is to give the percent and then follow it with the frequency in parentheses, for example: 25% (N = 15).)

3. Suppose the question of interest had been: Of the children who exceeded the recommended dinner Calories at age four, was there a higher percentage of boys or girls? That question is considered next.

a. What was the total number of children who exceeded the recommended dinner Calories at age four? What is the total number of children who did not exceed the recommended dinner Calories?

b. How many of the children who exceeded the recommended dinner Calories were boys? How many were girls?

c. What percent of the children who exceeded the recommended dinner Calories were girls? What percent were boys? Show your calculations.

d. What percent of the children who did not exceed the recommended dinner Calories were girls? What percent were boys? Show your calculations.

e. Explain why the percentages that you calculated in (c) and (d) do not directly answer research question 1.

4. The investigators were interested in whether Calorie consumption at age four was associated with being overweight at age 6.

a. Using the binary variables KCALR and OW6, create a 2 × 2 table to explore this relationship. Add the column and row totals to your table.

b. Of the children who exceeded the recommended dinner Calories, what percent was overweight? What percent was not overweight?

c. If an IGS child ate more Calories than recommended during the test meal, is she or he likely to be overweight at age 6? Explain. Report percentages and frequencies to justify your answer. Are you surprised by the answer?

d. If an IGS child ate more Calories than recommended during the test meal is she or he more likely to be overweight at age six than a child who did not eat more than the recommended? Report percentages and frequencies to justify your answer. Are you surprised by the answer?

e. In the preparation homework for lesson 3, you used scatterplots to investigate the relationship between Calories consumed at the test meal and BMI at age 6. In this question, you've tried a different approach to find linkage between Calorie consumption at age four and being overweight. Summarize your findings between the two approaches. What, if anything, have you learned from these analyses?

Practice

1. The investigators were interested in finding out if being overweight at age 4 was a predictor of being overweight at age 6. We've considered this question before when we investigated the relationship between BMI at age four and BMI at age six and found that the relationship was positive. But that analysis didn't exactly answer the question. We revisit the question here this time using the variables overweight at age four (OW4) and overweight at age six (OW6).

a. Create a 2 × 2 table of overweight at age four (OW4) versus overweight at age six (OW6). Add the row and column totals to your table.

b. Answer the following question based on your table from (a). If a child is overweight at age 4 is she or he more likely to be overweight or not be overweight at age 6? Report percentages and frequencies as justify for your answer.

c. What percentage of the overweight children at age six were overweight at age four?

2. A large sample of 12th grade students was asked to rate their intelligence compared to their peers. Researchers were interested in knowing whether gender was a factor in how students responded to this question. The results are summarized below

	Response			
Gender	Above Average	Average	Below Average	Total
Female	4209	2279	465	
Male	4555	1657	416	
Total				

Results on intelligence question from 12th grade students who participated in the 2010 Monitoring the Future (MTF) study.

a. Fill in the row and column totals to complete the table.

	Response			
Gender	Above Average	Average	Below Average	Total
Female	4209	2279	465	
Male	4555	1657	416	
Total				

b. How many people answered the survey question?

c. What percent of the respondents were males? What percent were females? Show your calculations.

d. What percent of the respondents rated themselves as above average intelligence? Show your calculations.

e. Were males or females more likely to rate themselves as above average? Support your answer with appropriate percentages and frequencies.

3. Use the table above to answer the following questions.

a. How many students rated themselves as having below average intelligence compared to their peers?

b. What percent of the group that rated themselves as having below average intelligence were males? What percent were females? Show your calculations.

c. Were males or females more likely to rate themselves as having below average intelligence? Support your answer with percentages and frequencies.

d. Explain the difference between the questions (b) and (c) and the percentages used to answer those questions.

Lesson 5 Project: Investigating Waist Circumference and Being Overweight

In this unit, you have used body mass index (BMI) by itself or with the CDC BMI Growth Charts for Boys and Girls to identify children who are overweight. However, BMI does not fully explain body fat distribution, which is related to metabolic health risks, at least in adults. In adults, waist circumference is a better indicator than BMI of fat distribution in the midsection of the body, which is related to certain health risks such as cardiovascular disease.

The waists of the IGS children were measured when they were four and again when they were six. The variables WCCM4 and WCCM6 in the Unit IGS Data contain the waist circumference measurements in centimeters at ages four and six, respectively.

1. Focus on the waist circumference measurements of the ISG children at age six.
   a. Create a modified boxplot of waist circumference at age six (WCCM6). Describe in words what the boxplot tells you about the waist circumferences of the six-year-old ISG children.

   b. Identify the ID numbers associated with the outliers. Were these children also classified as overweight by the variable OW6?

2. Investigate the relationship between waist circumference at ages four and six.
   a. Make a scatterplot of waist circumference at age six (WCCM6) versus waist circumference at age four (WCCM6). (Which variable belongs on the horizontal axis?)

   b. Do children with larger waist circumferences at age four also tend to have larger waist circumferences at age six? In other words, would you describe the association between the two variables as positive, negative, or neither? Explain.

   c. Fit a regression line to the data in your scatterplot. Write both its equation and the value of R^2.

   d. How much of the variability in waist circumference at age six can be explained using the regression equation based on waist circumference at age four?

   e. If two children's waist circumferences at age four differed by 1 cm, by how much would you expect their waist circumferences to differ at age six? Explain how you determined your answer.

   f. If a four-year-old child's waist circumference is 55 cm, predict the child's waist circumference at age 6. Show your calculations.

3. Just as there are overweight cutoffs for BMI that take into account age and gender, there are also overweight cutoffs for waist circumference that take into account age and gender. For six-year-old boys, the overweight cutoff for waist circumference is 56.3 cm and for six-year-old girls, the overweight cutoff is 57.1 cm.[5]

a. Form a new variable for overweight due to waist circumference at age six, WCOW6 as follows:

WCOW6 = 1: boys with WCCM6 > 56.3 or girls with WCCM6 > 57.1
WCOW6 = 0: boys with WCCM6 ≤ 56.3 or girls with WCCM6 ≤ 57.1

b. What percent of IGS children are classified as overweight by the variable WCOW6? Do you think this is reasonable? Explain.

4. In this unit, you discovered that one problem with using weight alone to identify overweight children is that weight and height are positively associated. Taller children tend to weigh more than shorter children. There may be a similar problem in using waist circumference to identify overweight children.

a. Make a scatterplot of waist circumference at age four and height. Fit a line to these data. Report the regression equation and value of R^2.

b. Describe the relationship between waist circumference and height. Do you think there is a connection between these two variable that might indicate the formation of a new variable for identifying being overweight , one that takes into account both waist circumference and height? Explain.

5. a. Form a new variable, the ratio of waist circumference and height at age six,

$$\text{RWCHT} = \frac{\text{waist circumference at age six}}{\text{height at age six}}.$$

b. Investigate the relationship the waist-to-height ratio RWCHT and BMI for the six-year-old IGS children.

6. A study done in the UK suggested that a waist-to-height ratio above 0.5 put children at increased metabolic risk.[6]

a. Form a binary variable for metabolic risk at age six (MR6) as follows:

MR6 = 1 if RWCHT > 0.5 (yes, increased metabolic risk)
MR6 = 0 if RWCHT ≤ 0.5 (no, not at increased metabolic risk)

b. What percent of the six-year-old IGS children are classified as at increased metabolic risk? Report the percent followed by the frequency in parentheses.

c. Where the boys or girls at age six more likely to be at increased metabolic risk? Create a two-way table and base your answer to this question on percentages calculated from your table.

7. Consider the following research question:

Are eating behaviors at age four related to being overweight at age six? More specifically, are any factors associated with the test meal related to waist-to-height ratios (either RWCHT or RM6)?

Use the Unit IGS Data to explore this question. Summarize the results of your explorations in a written report. Use charts and graphs to support your results. Justify your answer based on analysis of the Unit IGS Data. Feel free to compare your results using waist-to-height ratios in place of BMI at age 6.

Glossary

Bias - a measure of how far off an estimate is from the true value usually based on some non-impartial influence or judgment.

Binary variable - a variable that has only two possible outcomes; 0 and 1.

Biostatistics - the application of statistical methods and reasoning to biological study; the study of living organisms.

Body mass index (BMI) - a weight to height ratio of the individual's body mass divided by the square of his/her height. For metric units, body mass is measured in kilograms and height is measured in meters. Hence, the units for BMI are kg/m^2.

Box-and-whisker plot or boxplot - a graphical way to display the median, quartiles and extremes of a data set.

Cross-tabulation table for two variables - a two-way table.

Contingency table for two variables - a two-way table.

Dotplot - a graph in which a dot for each observation along a number line. If two or more data values are the same, dots corresponding to these values are stacked one above the other.

Explanatory variable - a variable that we think may explain (or even cause) changes in another variable. This variable is also referred to as the independent variable.

Factor - a factor is an explanatory variable that may cause or contribute to a result.

First quartile (Q1): The 25^{th} percentile (calculated as the median of the lower half of the data).

Five-number summary - a list of five important percentiles derived from a set of data: the minimum (0^{th} percentile), the first quartile (Q1, the 25^{th} percentile), the median (50^{th} percentile), the third quartile (Q3, 75^{th} percentile), and the maximum (100^{th} percentile).

Frequency - The number of times an observation occurs in a set of data.

Hypothesis - a tentative explanation or proposed explanation about an outcome, relationship, or scientific problem that can be tested by analysis of data and/or further investigation.

Interquartile range (iqr) - The difference between the first and third quartiles: iqr = Q3 – Q1.

Line of best fit - the line that minimizes the sum of the squares of the residuals (the vertical distances of the data points from the line).

Mean - a measure of center of a data set found by dividing the sum of the data by the number of data. Mathematically, this is the arithmetic average of the data values.

Median - a measure of center of a data set that is the middle number in an ordered list of the data. It is also the 50^{th} percentile or the second quartile.

Modified boxplot - a boxplot in which the 1.5 × iqr rule has been used to identify outliers. Outliers are plotted individually and the whiskers are shortened to extend from the ends of the box to the lowest and highest data values that are not outliers.

Negative association - a relationship between two variables in which smaller values in one variable tend to correspond to larger values in the other. Hence, as one variable increases, the other tends to decrease.

Outlier - a data value that lies outside the overall pattern made by the other data in a data set. Using the 1.5 × iqr rule, a data value is identified as an outlier if it lies more than the 1.5 × iqr units below the first quartile (Q1) or above the third quartile (Q3).

Percentile - a measure of the position in an ordered data set from the bottom (lowest values).

Positive association - a relationship between two variables in which smaller values in one variable tend to correspond to smaller values in the other and larger values in one variable tend to correspond to larger values in the other. Hence, as one variable increases, the other variable also tends to increase.

Range - the difference between the greatest and least data value in a set of data. The range is one measure of the variability of the data.

Regression line - a straight line that describes how a response variable is related to an explanatory variable. The regression line can be used to predict a value of the response variable given a specific value of the explanatory variable by substituting this value into the regression equation.

Residual or residual error - the difference between an actual response variable value and a predicted response variable value obtained from the regression equation.

Response variable - a variable that we think may respond to changes in another variable. This variable is also referred to as the dependent variable.

Scatterplot - a plot of ordered pairs, for example (x, y), that displays the relationship between two quantitative variables measured on the same individuals. Use the horizontal axis for the explanatory variable, x, and the vertical axis for the response variable, y.

Third quartile (Q3) - the 75^{th} percentile as calculated as the median of upper half of data.

Two-way table - a table that displays the frequencies (and/or percentages) of all possible combinations of two categorical variables. If the row variable has m possible outcomes and the column variable has n possible outcomes, then the dimension of the two-way table will be $m \times n$, where m is the number of rows and n is the number of columns.

Variable - a description of some characteristic or attribute of a person, place, thing or idea under study that can vary from one entity to the next.

References

[1] Lloyd, T., Chinchilli, V.M., Rollings, N., Kieselhorst, K., Tregea, D.F., Henerson, N.A. & Sinoway, L.I. (1998). Fruit consumption, fitness and cardiovascular, health in female adolescents: The Penn State Young Women's Health Study. *American Journal of Clinical Nutrition*, *67*, 624-630.

[2] Ogden, C.L., Kuczmarski, R.J., Flegal, K.M., Mei, Z., Guo, S., Wei, R., Grummer-Strawn, L.M., Curtin, L.R., Roche, A.F., & Johnson, C.L. (2002). Centers for Disease Control and Prevention 2000 growth charts for the United States: Improvements to the 1977 National Center for Health Statistics version. *Pediatrics, 109*(1), 45-60.

[3] Ogden, C.L., Carroll, M.D., Curtin, L.R., Lamb, M.M. & Flegal, K.M. (2010). Prevalence of high body mass index in U.S. children and adolescents, 2007-2008. *Journal of the American Medical Association, 303*(3), 242-249.

[4] U.S. Department of Health and Human Services and U.S. Department of Agriculture. (2005). Dietary Guidelines for Americans 2005 (6th Edition). Washington, DC: U.S. Government Printing Office.

[5] Mazicioglu, M.M., Hatipoglu, N., Ozturk, A., Cicek, B., Ustunbas, H.B. and Kurtoglu, S. (2010. Waist circumference and mid-upper arm circumference in evaluation of obesity in children aged between 6 and 17 years. *Journal of Clinical Research in Pediatric Endocrinology*, *2*(4), 144-150. doi: 10.4274/jcrpe.v2i4.144. Found at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3005693/.

[6] McCarthy, H.D. & Ashwell, M. (2006). A study of central fatness using waist-to-height ratios in UK children and adolescents over two decades supports the simple message – 'keep your waist circumference to less than half your height'. *International Association for the Study of Obesity*, *30,* 988-92.