Introduction to Data and the Analysis of Data: Instructor's Guide

Suggested Responses to Investigations & Supplementary Materials

This module introduces students to ways of thinking about and working with data using, as a case study, the analysis of 1.69-oz packages of plain M&Ms. The module is divided into six parts:

Part I. Ways to Describe Data
Part II. Ways to Visualize Data
Part III. Ways to Summarize Data
Part IV. Ways to Model Data
Part V. Ways to Draw Conclusions From Data
Part VI. Now It's Your Turn!

Interspersed within the module's narrative is a series of investigations, each of which asks students to stop and consider one or more important issues. Some of these investigations include data sets for students to analyze; for the data in the module's figures, you may wish to have students use the interactive, on-line version available at http://bit.ly/1ZokZKf.

Many of the investigations draw on a data set that consists of 30 samples of 1.69-oz packages of plain M&Ms, the general structure of which is shown here

TABLE 2. SOURCE, DISTRIBUTION, AND NET WEIGHT OF PLAIN M&MS IN 1.69-OZ BAGS

bag	store	blue	brown	green	orange	red	yellow	net weight (g)
27	CVS	5	17	6	4	8	19	50.802
28	Kroger	1	21	6	5	10	14	49.055
29	Target	4	12	6	5	13	14	46.577
30	Kroger	15	8	9	6	10	8	48.317

The counts for the different colors of M&Ms were collected by students at DePauw University in the fall 1996 semester as part of an in-class exercise; this was the only data collected at that time. To allow for a consideration of grouping in this case study, samples were assigned randomly to one of three hypothetical sources. Because the original data did not include a consideration of mass, the net weights included in the data set were generated specifically for this case study by simulating the random sampling of single plain M&Ms from a normally distributed population of weights. The values of μ and σ for this population were derived using the mean, $\bar{x}$, and the standard deviation, s, for the published weights of 462 plain M&Ms available through the Puget Sound Data Hoard (http://stat.pugetsound.edu/hoard/Default.aspx). The appendix to this Instructor's Guide includes the R code used to generate these net weights, as well as the R code used to generate the figures that accompany some of the suggested responses.

This case study is meant to serve as an introduction to data and to data analysis and, as with any introduction, it considers a small number of topics, principally those covered in Chapter 4 of *Analytical Chemistry 2.0*; additional resources that provide a deeper introduction to data and to data analysis are listed in Appendix 1 of the case study.

Suggested responses are presented in normal font; additional comments, suggestions, and supplementary materials are in *italic* font.

Part I: Ways to Describe Data

Investigation 1. Of the variables included in Table 1, some are categorical and some are numerical. Define these terms and assign each of the variables in Table 1 to one of these terms.

A categorical variable provides qualitative information that we can use to describe the samples relative to each other, or that we can use to place the samples into groups. For the data in Table 1, "bag id," "type," and "rank" are categorical variables.

A numerical variable provides quantitative information on which we can perform a meaningful calculation; for example, we can use "# yellow M&Ms" and "total M&Ms" to calculate the new variable "% yellow M&Ms." For the data in Table 1, "year," "weight (oz)," "# yellow M&Ms," "% red M&Ms," and "total M&Ms" are numerical variables.

Some students will include "year" as a categorical variable, which is not an unreasonable choice as it might serve as a useful way to group samples; however, it is listed here as a numerical variable because it can serve as a useful predictive variable in a regression analysis. Some students will include "rank" as a numerical variable, essentially rewriting the entries as numerals; however, there are no meaningful calculations that we can complete using this variable.

Investigation 2. Suppose we decide to code the type of M&M using 1 for plain and 2 for peanut. Does this change your answer to Investigation 1? Why or why not?

No. Although it is tempting to assume that a number must imply a numerical variable, we need to remember that we can convert any descriptive phrase into a number even if the number does not convey quantitative information. For example, although we might choose to code samples of plain M&Ms using the integer 1 and code samples of peanut M&Ms using the integer 2, we would never report that the average sample is of type $\{(4)(2) + (2)(2)\}/6 = 1.33$ as this does not have any meaningful interpretation.

Not all students are familiar with databases or with coding, and may ask why we might choose to code a variable if replacing a descriptive phrase with an integer provides us with no advantage and if it comes at the cost of making it more difficult for others to read our table. When this question arises, it is helpful to note that there are several reasons we might choose to replace a descriptive phrase with an integer when creating a computerized database, particularly if the database has many records: storage space (it takes less space to store an integer than it does to store a character string); search speed (it takes less time to search for an integer than it does to search for a character string); and fewer errors when entering data (consider how easy it is to type penut for peanut).

Investigation 3. Categorical variables are described as nominal or ordinal. Define the terms nominal and ordinal and assign each of the categorical variables in Table 1 to one of these terms.

A nominal categorical variable does not carry with it any implied order; an ordinal categorical variable, on the other hand, coveys a meaningful sense of order. For the categorical variables in Table 1, "bag id" and "type" are nominal variables, and "rank" is an ordinal variable.

Some students may interpret the use of consecutive alphabetical letters for "bag id" as implying order, but there is nothing to suggest that this order is meaningful.

Investigation 4. A numerical variable is described as either ratio or interval depending on whether it has (ratio) or does not have (interval) an absolute reference. Explain what it means for a variable to have an absolute reference and assign each of the numerical variables in Table 1 as either a ratio variable or an interval variable. Why might this difference be important?

A numerical variable has an absolute reference if it has a meaningful zero—that is, a zero that means a measured quantity of none—against which we reference all other measurements of that variable.

For the numerical variables in Table 1, "year" is an interval variable because our scale for time is referenced to an arbitrary point in time, 1 CE, and not to the beginning of time; "weight (oz)," "# yellow M&Ms," "% red M&Ms," and "total M&Ms" are ratio variables because each has a meaningful zero.

For a ratio variable, we can make meaningful absolute and relative comparisons between two results, but only meaningful absolute comparisons for an interval variable. For example, consider **sample e**, which was collected in 1994 and which has 331 M&Ms, and **sample d**, which was collected in 2000 and which has 24 M&Ms. We can report a meaningful absolute comparison for both variables: **sample e** is six years older than **sample d** and **sample e** has 307 more M&Ms than **sample d**. We also can report a meaningful relative comparison for the total number of M&Ms—there are $331 \div 24 = 13.8$ times as many M&Ms in **sample e** as in **sample d**—but we cannot report a meaningful relative comparison for year because a sample collected in 2000 is not $2000 \div 1994 = 1.003$ times older than a sample collected in 1994.

Investigation 5. Numerical variables also are described as discrete or continuous. Define the terms discrete and continuous and assign each of the numerical variables in Table 1 to one of these terms.

A numerical variable is discrete if it can take on only specific values—typically, but not always, an integer value—between its limits; a continuous variable can take on any possible value within its limits. For the numerical data in Table 1, "year," "# yellow M&Ms," and "total M&Ms" are discrete in that each is limited to integer values. The numerical variables "weight (oz)" and "% red M&Ms," on the other hand, are continuous variables.

Students will sometime ask why weight is not a discrete variable given that a balance records the weight to a set number of decimal points. Here it is helpful to remind students that what makes a variable discrete is not our ability to measure it, but a property inherent in the variable itself. In the context of this data, each M&M is an indivisible unit and the number of units is discrete; however, two M&Ms with masses of 0.8561 g and 0.8559 g have different weights even if our balance reads, and we report, both as 0.856 g.

Part II: Ways to Visualize Data

Investigation 6. Use the dot plot in Figure 1 to deduce the general structure of a box and whisker plot, giving particular attention to the position along the *x*-axis of the three vertical lines that make up the yellow box and the two vertical lines that make up the whiskers on either side of the yellow box. You might begin by tabulating the number of samples that fall to the left of the box, that fall within the box, including its boundaries, and that fall to the right of the box, and the number of samples that lie to the left and to the right of line inside the box.

Of the 30 samples, seven are on the left side of the box, 17 are within the box, and six are on the right side of the box; relative to the box's middle line, 14 lie to the left and 13 lie to the right. One reasonable interpretation of these observations is that the box contains approximately the middle 50% of the data (17 of 30 samples, or 57%) and that the line inside the box divides the data approximately in half (14 of 30 samples, or 47%, are left of the line and 13 of 30 samples, or 43%, are right of the line).

The two whiskers extend to encompass all but one of the 30 samples. Clearly the whiskers convey information about the overall variability of the data, but there is insufficient information in this one example to suggest exactly how the length of the whiskers are determined (although, at least for this example, the whiskers do not include the one sample that lies at a distance of more than $1.5\times w$, where w is the width of the box.

For students who have difficulty accepting 57%, 47%, and 43% as being suggestive of 50%, it helps to have them consider the effect on the percentages of the limited number of samples (30) and the fact that multiple samples have the same result. For further details on box and whisker plots, see https://en.wikipedia.org/wiki/Box_plot.

There are a variety of ways to define the whiskers and to handle points that fall outside of a whisker. The method used here is to draw a whisker to the data point whose value is no greater than $+1.5\times w$ of the box's largest value (in this case $17+1.5\times 4 = 23$), and to draw a whisker to the data point whose value is no less than $-1.5\times w$ of the box's smallest value (in this case $13 - 1.5\times 4 = 7$). Results that fall outside of the whiskers are flagged using a dot (•), even when individual results are not shown using a dot plot.

Investigation 7. The box and whisker plot in Figure 1 is perfectly symmetrical in that each side of the box is two units from the box's middle line, and each whisker is six units from the box's nearest edge. What does this symmetry suggest about how the results are distributed? Is the actual distribution of the 30 results perfectly symmetrical? If no, is this a problem?

The symmetry of the box and the whiskers suggests that there is a symmetrical distribution of the data set's individual results around its middle. The data itself is not perfectly symmetrical—for example, there are five samples within ±2 of the left whisker, but just three samples within ±2 of the right whisker. This difference between the symmetry of the data and the symmetry of the box and whisker plot is not a problem as we use a box and whisker plot simply to develop a general understanding of our data's structure.

Investigation 8. In Figure 1 we see that the result for sample 22 falls outside the range of values included within the whiskers. Why might a result that falls outside the whiskers concern us? Does the presence of this particular point suggest a problem? How might your response change if this sample's reported value is 0 yellow M&Ms? How might your response change if this sample's reported value is 45 yellow M&Ms?

If we assume that the box and the whiskers should include all samples for which the results are not subject to an error—then we might wish to look more closely at a sample that falls outside of the

whiskers as it may suggest a problem with our data, either in the counting of M&Ms, in the recording of that count, or in the manufacturing process. In this case, the result for sample 22 does not bother us as it is not that different from the next lowest value and, more important, an error in counting M&Ms does seem not likely when the bag contains just 55 M&Ms (a counting error is more likely if a bag has 550 M&Ms). For the same reason, we are not likely to question a result of 0. A result of 45 yellow M&Ms, however, seems unreasonable as it is almost twice as many as the next highest value; in this case we might suspect that an error was made when entering the result into the data table.

Investigation 15 introduces the difference between samples and populations, so this language is not used here; if you wish to discuss this difference here, you may wish to begin the case study with a discussion of samples and populations.

Investigation 9. Figure 2 shows box and whisker plots and dot plots for all six colors of M&Ms included in Table 2 (note: even with jittering, you will not be able to see all 30 samples in these dot plots). Based on these plots, where do you see similarities and where do you see differences in the distribution of M&Ms? What do these similarities and differences suggest to you? For those distributions that do not appear symmetrical, suggest one or more reasons for the lack of symmetry. What do the relative positions of the data for brown and for green M&Ms suggest about their relative abundance in 1.69-oz packages of plain M&Ms?

There are many observations we can make using this data, a few of which are gathered here. One observation is that finding a sample outside of the whiskers is a rare event as it happens just once in 180 measurements (sample 22, yellow). Another observation is that the boxes for brown M&Ms and for yellow M&Ms overlap each other but do not overlap with the other four colors of M&Ms (although the upper edge of the box for red abuts the lower edge of the box for yellow); this suggests that yellow M&Ms and brown M&Ms are much more common than the other four colors. Another interesting difference is that the lower whiskers for blue, green, and orange M&Ms are much shorter than their respective upper whiskers; this suggests that their distributions are not symmetrical, a result that is not surprising given that the we cannot have fewer than zero M&Ms with any particular color. Finally, the relative positions of the box and whisker plots for green M&Ms and for brown M&Ms suggests that it is a rare bag that has more green M&Ms than brown M&Ms, which places a hard limit on the data's lower boundary; indeed, this happens just once (sample 30, which has 9 green M&Ms and 8 brown M&Ms).

Investigation 10. Figure 3 shows box and whisker plots and dot plots for yellow M&Ms grouped by the store where the packages of M&Ms were purchased. Based on these plots, where do you see similarities and where do you see differences in the distribution of yellow M&Ms? What do these similarities and differences suggest to you? In what ways might this data be reassuring to us? Give an example of a result that might suggest we look more closely at our data.

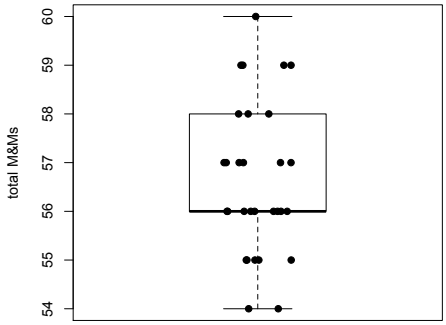
Although the box and whisker plots are quite different in terms of the relative sizes of the boxes and the relative length of the whiskers, the dot plots suggest that the distribution of the underlying data is relatively similar in that most values are in the range of 12–18 yellow M&Ms with a maximum of 22 or 23 yellow M&Ms and a minimum of eight yellow M&Ms (setting aside sample 22, which, as noted in the response to Investigation 9, is the only result in 180 measurements that does not fall within the span of its whiskers). These observations are reassuring because we do not expect the source of the bags of M&Ms to affect the composition of their contents. If we saw evidence that the source did affect our results, then we would need to look more closely at the bags themselves for evidence of a poorly controlled variable, such as type (Did we accidently purchase bags of peanut

butter M&Ms from one store?) or the product's lot number (Did the manufacturer change the composition of colors between lots?).

As a reminder, the division of the 30 samples among these three sources is artificial and is done solely to illustrate the concept of grouping and the analysis of a common variable (yellow M&Ms) between different groups.

Investigation 11. Draw a box and whisker plot and an accompanying dot plot for the total number of M&Ms. Compare your plots to those in Figure 2 and discuss any similarities and differences.

The total number of M&Ms in the 30 samples are, in order 57, 56, 59, 56, 57, 54, 57, 57, 56, 55, 59, 58, 55, 56, 55, 58, 56, 56, 56, 60, 58, 55, 57, 56, 55, 59, 59, 57, 54, and 56. A box & whisker plot and a dot plot are shown below.



The most interesting observation for this data is that the box does not appear to have a middle line. Of course, it actually does have a middle line, but it simply is the same as the box's lower limit. We already know from Investigation 7 that the box's middle line divides the data in half, so we know that half of the bags have 56 or fewer M&Ms and that half have 56 or more M&Ms. We also know that we have a greater number of unique results above the middle value (57, 58, 59, and 60 M&Ms) than below the middle value (54 and 55 M&Ms). Although the box and the whisker plot looks symmetrical, the results are skewed somewhat toward larger numbers of M&Ms.

Students will benefit from drawing this plot by hand. Although Excel does not include a command for drawing a box and whisker plot, an on-line search will yield methods for creating a plot that will mimic the traditional box and whisker plot. The statistical program R has a built in boxplot command.

Investigation 12. For the histograms in Figure 4, where do you see similarities and where do you see differences in the distribution of M&Ms? How do the results seen here compare with your interpretation of the box and whisker plots and the dot plots in Figure 2?

The information here is very similar to what we saw in the box and whisker plots. In particular, the colors of M&Ms with the least symmetrical whiskers—blue, green, and orange—have histograms that are not symmetrical and that tend to decrease in value more slowly when moving from the bin that contains the greatest number of samples to bins that contain samples with greater numbers of M&M. The lack of symmetry for yellow M&Ms, which decreases in value more slowly as we move from the bin that contains the greatest number of samples toward bins that contains samples with a smaller number of M&Ms, is more evident here than in the box and whisker plots.

Investigation 13. The histograms in Figure 5, from left-to-right, use bins widths of 1, 2, and 3 units, respectively. Note that the *x*-axis shows the specific results gathered into each bin. How does the choice of bin size affect your understanding of this data? Which of these histograms provides the best representation of the data? As part of your answer, identify what you see as the limitations of the other two histograms.

Using a bin size of 1 unit makes it easy to see that there were no bags with 9, 12, or 14 M&Ms; this information is not available when the bins have sizes of 2 units or of 3 units. The histograms using bins of 1 unit and of 2 units are similar in shape: if we draw a smooth curve through the data—ignoring the noise due to our limited number of samples—both histograms suggest that the frequency of an outcome decreases as we move from the smallest number of M&Ms (four) to the largest number of M&Ms (15); a smooth curve through the histogram using a bin of 3 units, however, suggests that the frequency increases and then decreases as we move from the smallest number of M&Ms (four) to the largest number of M&Ms (15).

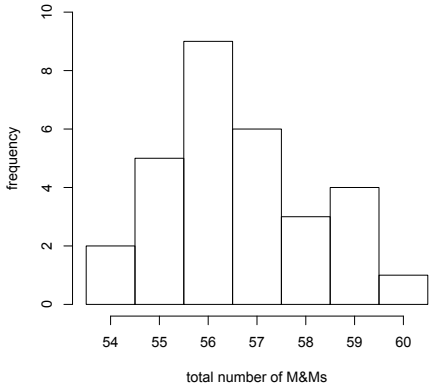
Of the three options, the best representation of the data is the one with a bin size of 2 units. Although the histogram using a bin size of 1 unit does show us possible outcomes that did not occur, the resulting histogram is much noisier. The histogram using a bin size of 3 units is the least noisy, but its first bin includes as possible outcomes results that are not in our data (two and three orange M&Ms), which is somewhat misleading.

Investigation 14. Draw a histogram for the total number of M&Ms and explain the reason(s) for your choice of bin size. Compare your plots to those in Figure 4 and discuss any similarities and any differences.

The total number of M&Ms in the 30 samples are, in order 57, 56, 59, 56, 57, 54, 57, 57, 56, 55, 59, 58, 55, 56, 55, 58, 56, 56, 56, 60, 58, 55, 57, 56, 55, 59, 59, 57, 54, and 56. Gathering these into a frequency table

number of M&Ms	54	55	56	57	58	59	60
frequency	2	5	9	6	3	4	1

suggests that a bin size of 1 unit is a good option as a bin size of 2 units has just four total bins, one of which must include a result either less than 54 or greater than 60; the resulting histogram is shown below.



This histogram is consistent with our observations from the box and whisker plot in Investigation 12, but it presents us with a much clearer picture of the data.

Students will benefit from drawing this plot by hand. Excel's Data Analysis tools provides a method for creating histograms, as does R *using the hist command.*

Part III: Ways to Summarize Data

Investigation 15. Before we consider ways to summarize our data, we need to draw a distinction between a sample and a population. We collect and analyze samples with the hope that we can deduce something about the properties of the population. Using our data for M&Ms as an example, define the terms sample and population.

For our data, the structure of which is in Table 2, each row is single sample of plain M&Ms in the form of individual 1.69-oz packages. These samples are drawn from the much larger population of all plain M&Ms (or, at least all plain M&Ms manufactured at the time the samples were packaged).

Investigation 16. Using the data for yellow M&Ms, calculate the mean and the median for each store and discuss your results. If the mean and the median are equal to each other, what might you reasonably conclude about your data? If the mean is larger than the median, or if the mean is smaller than the median, what might you reasonably conclude about your data? A measure of central tendency is considered robust when it is not changed by one or more results that differ substantially from the remaining results. Which measure of central tendency is more robust? Why?

To help us understand how we arrive at each value, we will use the data in Figure 3 for yellow M&Ms in bags purchased at CVS. To begin, let's construct a frequency table, which shows the eight unique results, ordered from smallest-to-largest, and the number of bags with each unique result.

number (N)	5	8	13	15	16	17	19	23
frequency (f)	1	1	2	2	1	1	1	1

To calculate the mean using a frequency table, we multiply each unique result by its frequency, sum up the values, and divide by the number of samples; thus

number (N)	5	8	13	15	16	17	19	23
frequency (f)	1	1	2	2	1	1	1	1
$N \times f$	5	8	26	30	16	17	19	23

The sum of the values in the last row is 144, which gives the mean as

$$\bar{x} = \frac{144}{10} = 14.4 \text{ yellow M\&Ms}$$

For the 10 samples from CVS, the median is the average of the 5th and the 6th values when ordered by rank. Using the frequency data, the 5th value is 15 and the 6th value is 15, which gives the median as 15 yellow M&Ms. The following table summarizes the means and the medians for yellow M&Ms by store.

store	mean	median
CVS	14.4	15.0
Kroger	14.2	15.0
Target	14.9	14.5

For each store, we see that the mean and the median are similar in value; we also see that the means and the medians between the three stores are similar. Both are reasonable results as, discussed in the response to Investigation 10.

If the mean and the median are equal to each other, then the distribution of the individual values must be perfectly symmetrical about the mean and median. If the mean is larger than the median, then the data likely is skewed toward the right, and if the mean is smaller than the median, then the data likely is skewed toward the left.

The median is more robust than the mean because the median uses the rank, not the value, of each data point, which makes it relatively insensitive to an unusually large or small result. For example, if the sample from CVS with 19 yellow M&Ms has, instead, 29 yellow M&Ms, then the mean increases from 14.4 to 15.4, but the median remains unchanged.

Students should, of course, calculate the mean (and other statistics) using a calculator, a spreadsheet, such as Excel, or a statistical program, such as R. There is benefit, however, in seeing how the sample's data comes together to give the mean, which is the reason for detailing the calculation using a frequency table; the same approach is used in the next investigation.

Although generally it is true that data is skewed to the right when the mean is greater than the median and skewed to the left when the mean is less than the median, this 'rule' does not hold true in all cases. In particular, it may not hold for a discrete distribution when the areas to the left and to the right of the median are not equal (because many samples share the median's value). It also fails with multimodal distributions and in distributions where there is a long tail in the direction of the skew, but a heavy tail in the other direction. See von Hippel, P. T. "Mean, Median, and Skew," J. Statistics Education, 2005, 13(2) (www.amstat.org/publications/jse/v13n2/vonhippel.html) for additional details.

Investigation 17. Using the data for yellow M&Ms, calculate the variance, the standard deviation, the range, and the *IQR* for each store and discuss your results. Is there a relationship between the standard deviation, the range, or the *IQR*? A result is considered robust when its value is not changed by one or more values that differ substantially from the remaining values. Which measure of spread—the variance, the standard deviation, the range, or the *IQR*—is the most robust? Why? Which is the least robust? Why?

To help us understand how we arrive at each value, we will use the data in Figure 3 for yellow M&Ms in bags purchased at CVS. To begin, let's use the same frequency table from Investigation 16, which shows the eight unique results, ordered from smallest-to-largest, and the number of bags with each unique result.

number (N)	5	8	13	15	16	17	19	23
frequency (f)	1	1	2	2	1	1	1	1

To calculate the variance, we first calculate each unique difference relative to the mean $(x_i - \bar{x})$, square these unique differences, multiply each unique squared difference by its frequency, sum up the values, and divide by $n - 1$; thus

number (N)	5	8	13	15	16	17	19	23
frequency (f)	1	1	2	2	1	1	1	1
$(x_i - \bar{x})$	–9.4	–6.4	–1.4	0.6	1.6	2.6	4.6	5.6
$(x_i - \bar{x})^2$	88.36	40.96	1.96	0.36	2.56	6.76	21.16	73.96
$f \times (x_i - \bar{x})^2$	88.36	40.96	3.92	0.72	2.56	6.76	21.16	73.96

The sum of the values in the last row is 238.40, which gives the variance as

$$s^2 = \frac{238.40}{10 - 1} = 26.49$$

and the standard deviation as 5.15 yellow M&Ms. To find the range, we subtract the smallest value (5 yellow M&Ms) from the largest value (23 yellow M&Ms), which makes the range 18 yellow M&Ms. To find the *IQR*, we use the median to divide the 10 samples into a lower half with values of 5, 8, 13, 13, and 15 yellow M&Ms, and an upper half of 15, 16, 17, 19, and 23 yellow M&Ms; the median of the upper half is 17 yellow M&Ms and the median of the lower half is 13 yellow M&Ms,

which makes the *IQR* 4 yellow M&Ms. The following table summarizes the variance, the standard deviation, the range, and the *IQR* for yellow M&Ms by store.

store	variance	standard deviation	range	*IQR*
CVS	26.49	5.15	18	4
Kroger	21.96	4.69	15	3
Target	15.43	3.93	15	7

These results are consistent with our observations from Investigation 10.

This is a nice set of data to show that there is no general relationship between the variance, the standard deviation, the range, and the *IQR* as measures of spread. For example, the store with the smallest standard deviation (Target) is the store with the largest *IQR*.

For the reasons outline in the response to Investigation 16, the *IQR* is the most robust measure of spread as it uses the rank, not the value, of each data point. The least robust measure of spread is the range. For example, if the sample from CVS with 19 yellow M&Ms has, instead, 29 yellow M&Ms, then the range increases from 18 to 24, but the *IQR* remains unchanged.

Students often ask why we divide by $n - 1$ *instead of by* n. *Although a rigorous explanation is beyond the scope of this case study, here is an intuitive way for them to think about this. In the numerator of the equation for variance we sum up the squared differences between the result for each sample,* x_i, *and the mean of these samples,* $\bar{x}$. *Because the sample's mean is calculated from the individual samples, we reasonably might expect that this sum is smaller than the result if we used the population's mean (which is unknown to us and which might be quite different from sample's mean); dividing by* $n - 1$ *instead of by* n *compensates for this difference. For further details on what is called Bessel's correction, see https://en.wikipedia.org/wiki/Bessel%27s_correction.*

Part IV: Ways to Model Data

Investigation 18. So, what does it mean to build a model? Consider the histograms in Figure 4. What property of the population are we attempting to model? What do your responses imply about the model's general mathematical form? What does it mean to test a model and how might we accomplish this?

For the histograms in Figure 4, we wish to model the number of each color of M&M in a 1.69-oz bag of plain M&Ms. A suitable mathematical model will need to predict the probability, p, of drawing X M&Ms of a particular color when we select a sample of N M&Ms from the population of all M&Ms; thus, we expect the equation to be a function of the form $P = f(X, N)$.

To test a model, we need to compare the results predicted by our model to the results of our experiments. For example, Table 2 contains results for the distribution of colors and the net weight of plain M&Ms in 1.69-oz bags. If we develop a model that predicts successfully the average number of yellow M&Ms in a 1.69-oz bag, then we have some confidence that the model is reasonable. Of course, we need to define how we decide if a model's predictions agree with our data, which we will explore in greater detail in Part V.

Investigation 19. The box and whisker plot in Figure 1 includes data from the analysis of 30 samples of 1.69-oz bags of plain M&Ms. Collectively, the samples have 1699 M&Ms, of which 435 are yellow. If you pick one M&M at random from these 1699 M&Ms, what is the probability, p, that it is yellow? Suppose that this probability applies to the population of all plain M&Ms. If we draw a sample of five M&Ms from this population, what is the probability that the sample contains no yellow M&Ms? Repeat for each of 1–5 yellow M&Ms. Construct a histogram of your results and report the mean and the variance. Repeat this analysis for green M&Ms. Compare your two histograms and discuss their similarities and their differences. Using the data in Table 2, comment on the suitability of the binomial distribution for modeling the number of yellow M&Ms in samples of five M&Ms.

Given the data from our samples, the probability of selecting a single yellow M&M is

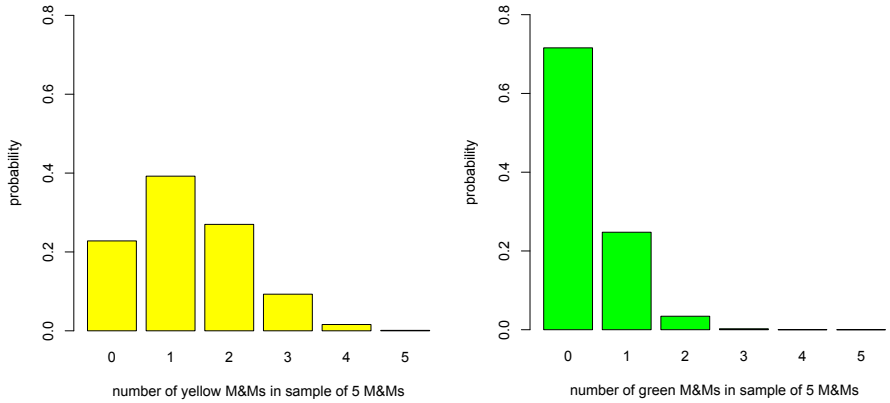
$$p = \frac{435}{1699} = 0.256$$

If we assume that this probability applies to the population of plain M&Ms and if we assume a binomial distribution, then the probability of drawing zero yellow M&Ms in a sample of five M&Ms is

$$P(0, 5) = \frac{5!}{0!(5-0)!} \times (0.256)^0 \times (1 - 0.256)^{5-0} = 0.228$$

The remaining probabilities are 0.392, 0.270, 0.093, 0.016, and 0.001 for $P(1, 5)$ to $P(5, 5)$, respectively; the resulting distribution is shown below on the left. The mean and the variance are

$$\mu = Np = 5 \times 0.256 = 1.28 \qquad \sigma^2 = Np(1-p) = 5 \times 0.256 \times (1 - 0.256) = 0.95$$



For green M&Ms, there are 110 in the combined sample of 1699 plain M&Ms, or a probability of 0.0647. The probabilities for drawing zero to five green M&Ms in a sample of five M&Ms are, respectively, 0.716, 0.248, 0.034, 0.002, 0.000, and 0.000; the resulting histogram is shown above to the right. The mean and the variance are 0.32 and 0.30, respectively.

In terms of similarities, both histograms encompass a net probability of 1.00 as they span the six possible outcomes when drawing a sample of five M&Ms, and neither histogram is symmetrical around its most probable outcome. The most importance difference between the two histograms is the relative frequencies of the possible outcomes; in particular, a sample of five M&Ms is more than 3× as likely to have no green M&Ms than to have no yellow M&Ms, and is more than 3× as likely to have three or more yellow M&Ms than three or more green M&Ms. Given their relative abundances—25.6% of the 1699 total M&Ms are yellow versus just 6.47% for green—these differences make sense.

From Table 2, we know that actual distribution of results for yellow M&Ms in the first five sampled is seven with no yellow M&Ms, 13 with one yellow M&M, eight with two yellow M&Ms, two with three yellow M&Ms, and zero with four and with five yellow M&Ms. The following table compares the results of our experiment and the predicted results from our model.

$P(X,N)$	experiment	model	absolute error
$P(0,5)$	0.233	0.228	0.005
$P(1,5)$	0.433	0.392	0.041
$P(2,5)$	0.267	0.270	–0.003
$P(3,5)$	0.067	0.093	–0.026
$P(4,5)$	0.000	0.016	–0.016
$P(5,5)$	0.000	0.001	–0.001

As one sample is $1/30^{\text{th}}$, or 0.033, of the 30 samples, the absolute errors represent an oversampling of approximately one for $P(1,5)$ and an undersampling of approximately one for $P(3,5)$, an experimental uncertainty that seems reasonable given the relatively small number of samples and the relatively small value for N.

Investigation 20. Explain why we cannot use the binomial distribution to model the distribution of yellow M&Ms in 1.69-oz bags of plain M&Ms.

The binomial distribution predicts the probability of a particular event, X, such as drawing five yellow M&Ms, in samples of fixed size, N, where $X \leq N$. Because the number of M&Ms varies from bag-to-bag, the value of N varies from bag-to-bag and we cannot model the distribution of M&Ms

in 1.69-oz bags of plain M&Ms; we could, however, model the distribution of yellow M&Ms in all 1.69-oz bags that contain the same total number of M&Ms.

Investigation 21. The histograms in Figure 4 include data from the analysis of 30 samples of 1.69-oz bags of plain M&Ms. Collectively, the samples have an average of 14.5 yellow M&Ms per bag. Suppose this rate applies to the population of all 1.69-oz bags of plain M&Ms. If you pick a 1.69-oz bag of plain M&Ms at random, what is the probability that it contains exactly 11 yellow M&Ms? Repeat for each of 0–29 yellow M&Ms. Construct a histogram that shows the actual distribution of bags of M&Ms for each of 0–29 yellow M&Ms, using a bin size of 1 unit, and overlay a line plot that shows the predicted distribution of bags; be sure to you use the same scale for each plot's *y*-axis. Comment on your results.

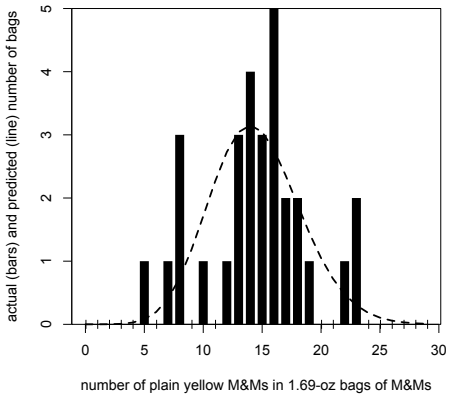
If we assume that the average rate of 14.5 yellow M&Ms per bag applies to the population of plain M&Ms, and assume a Poisson distribution, then the probability of finding 11 yellow M&Ms in a 1.69-oz package of plain M&Ms is

$$P(11,14.5) = \frac{e^{-14.5}14.5^{0}}{11!} = 0.0753$$

or 2.3 out of 30 bags of M&Ms. The actual number of bags of M&Ms that contain each of 0–29 yellow M&Ms and the predicted probabilities are gathered in the following table; the predicated probabilities from the Poisson equation are multiplied by 30 so that the two results are on the same scale.

X	Actual	$P(X,14.5)$	X	Actual	$P(X,14.5)$	X	Actual	$P(X,14.5)$
0	0	0.000	10	1	1.713	20	0	1.050
1	0	0.000	11	0	2.258	21	0	0.725
2	0	0.002	12	1	2.729	22	1	0.478
3	0	0.008	13	3	3.043	23	2	0.301
4	0	0.028	14	4	3.152	24	0	0.182
5	1	0.081	15	3	3.047	25	0	0.106
6	0	0.195	16	5	2.761	26	0	0.059
7	1	0.405	17	2	2.355	27	0	0.032
8	3	0.733	18	2	1.897	28	0	0.016
9	0	1.181	19	1	1.448	29	0	0.008

The resulting histogram for the actual distribution of yellow M&Ms in the 30 samples and the predicted distribution are shown here



Two factors make difficult any comparison of the actual counts to the predicted counts: the actual counts are discrete (we can have 0 or 1 bag with five yellow M&Ms, but we cannot have 0.08 bags with five yellow M&Ms), and the number of samples, at 30, is too small to allow for predicted counts of at least one bag for outcomes with a small probability (we would need to sample 370 bags of M&Ms to have a predicted count of one bag with five yellow M&Ms). Still, the overlap of the actual and the predicted values, and the general shape of the actual distribution suggests that the Poisson distribution provides a reasonable model of our data.

Investigation 22. Explain why we cannot use the binomial distribution or the Poisson distribution to model data for the net weight of M&Ms in Table 2.

The binomial distribution and the Poisson distribution are useful for modeling discrete events, such as the number of yellow M&Ms in a sample of fixed size, or the number of green M&Ms in bags of a particular size. The net weight of a sample of M&Ms, however, is a continuous variable, which requires a different type of mathematical model.

Investigation 23. Using the curves in Figure 6 as an example, discuss the general features of a normal distribution, giving particular attention to the importance of variance. How do you think the areas under the three curves from $-\infty$ to $+\infty$ are related to each other? Why might this be important?

Here are three observations based on these three examples of normal distribution curves: (a) a normal distribution is symmetric about μ, with half of its outcomes on either side of μ; (b) the most likely outcome in a normal distribution is when $x = \mu$; and (c) as the variance increases, the normal distribution's maximum value decreases and the spread of its distribution on either side of μ becomes wider.

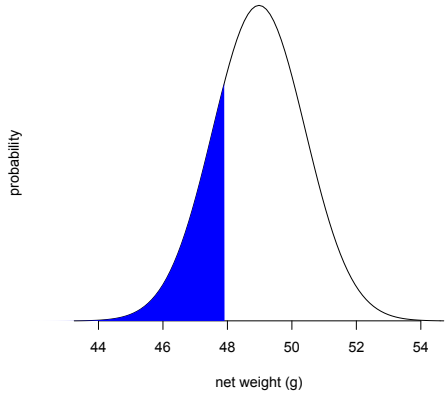
The area under the curve must equal the total probability of obtaining the outcome x; thus, the area under all three curves is the same and is equal to 1. This is important because it means that the area between any two limits defined in terms of μ and σ is the same for any value of μ and σ.

Investigation 24. Assuming that the mean, $\bar{x}$, and the standard deviation, s, for the net weight of our samples of M&Ms are good estimates for the population's mean, μ, and standard deviation, σ, what is the probability that the contents of a 1.69-oz bag of plain M&Ms selected at random will weigh less than the stated net weight of 1.69 oz? Suppose the manufacturer wants to reduce this probability to no more than 5%: How might they accomplish this?

For our 30 samples, the mean net weight is 48.98 g with a standard deviation of 1.433 g. The stated net weight of 1.69 oz is equivalent to 47.9 g. To find the probability that the M&Ms in a randomly selected 1.69-oz bag have a mass less than 47.9 g, we first calculate the deviation, z, taking the mean and the standard deviation for our samples as estimates for μ and σ

$$z = \frac{x - \mu}{\sigma} = \frac{47.9 - 48.98}{1.433} = -0.747$$

and then use the table in Appendix 3 to find the area under the normal curve to the left of 47.9, finding that it is 0.228, or 22.8%; the figure below shows this area highlighted in blue



To decrease this probability to 5%, or 0.050, we use Appendix 3 to find that this corresponds to a z of –1.645. Substituting this into the equation for z gives

$$z = \frac{x - \mu}{\sigma} = \frac{47.9 - \mu}{\sigma} = -1.645$$

With one equation and two unknowns, there are many possible combinations of mu and sigma that will work. We can place an upper limit on each by maintaining σ as 1.433 and calculating μ

$$\frac{47.9 - \mu}{1.433} = -1.645$$

and calculating μ as 50.27 g, or by maintaining μ as 48.98

$$\frac{47.9 - 48.98}{\sigma} = -1.645$$

and calculating σ as 0.650 g. Given that the average bag has a mean net weight of 48.98 g and a mean number of M&Ms of 56.63, the average plain M&M has a mass of 0.865 g. To increase the mean net weight from 48.98 g to 50.27 g, we need to increase the mean number of M&Ms per bag to 57.92, or we need to decrease the standard deviation by the equivalent of ±0.91 M&Ms.

Investigation 25. Suppose we arrange to collect samples of plain M&Ms such that each sample contains 330 M&Ms—an amount roughly equivalent to a 10-oz bag of plain M&Ms—drawn from the same population as the data in Table 2. Can we model this data using a normal distribution in place of the binomial distribution or the Poisson distribution? What advantages are there in being able to use the normal distribution? How might you apply this to more practical analytical problems, such as determining the concentration of Pb^{2+} in soil?

When N is equal to 5—as is the case in Investigation 22—it is impossible to use a normal distribution to approximate a binomial distribution because there is no probability, p, where both $N \times p \geq 5$ and $N \times (1 - p) \geq 5$ are true. If we increase N to 330, then any value of p that is greater than 0.0152 or that is smaller than 0.984 will allow us to approximate the data using a normal distribution; for the data in Table 2, the smallest value of p is for green M&Ms (0.0647, or 6.47%) and the largest value of p is for brown M&Ms (0.258, or 25.8%); thus, we expect that it is possible to model the data using a normal distribution.

The average count per 1.69-oz bag, λ, of each color of M&M ranges from a minimum of 3.67 for green M&Ms to a maximum of 14.8 for brown M&Ms, neither of which meets the criterion of $\lambda \geq 5$ needed to use a normal distribution to approximate the Poisson distribution. If we increase N

from an average of 56.63, for the data in Table 2, to 330, then the average count for any color will increase by 5.83×; thus, the smallest value of λ for any color is 21.3 for green M&Ms, which suggests that it is possible to model the data using a normal distribution.

The primary advantage to us in using the normal distribution is being able to use a single distribution to model diverse types of data, which often simplifies our analysis of data.

A sample of soil consists of many different types of materials—some organic and some inorganic—each of which has a different μ and σ for its concentration of Pb^{2+}. Because individual particles of these materials are small, in any reasonable sample the value of N for each particle is sufficiently large that the concentration of Pb^{2+} likely follows a normal distribution even if the underlying distribution of particles follows a binomial or a Poisson distribution.

Part V: Ways to Draw Conclusions From Data

Investigation 26. A z of 1.96 corresponds to a 95% confidence interval. Using Appendix 2, show that this is correct. What value of z corresponds to a 90% confidence inteval, and what value of z corresponds to a 99% confidence inteval? Report the 90%, the 95% and the 99% confidence intervals for the net weight of a single 1.69-oz bag of plain M&Ms drawn from a population for which μ is 48.98 g and σ is 1.433 g. For the data in Table 2, how many of the 30 samples have net weights that fall outside of the 90% confidence interval? Does this result make sense given your understanding of a confidence interval?

From the table in Appendix 2, we know that the area to the right of a z of 1.96 is 2.5% (a probability of 0.025) of the total area. Because the normal distribution is symmetrical, we know that the area to the left of a z of –1.96 also is 2.5% of the total area. The combined area of 5% is the percentage of samples excluded from the confidence interval; thus, the term $\pm z\sigma$ encompasses 95% of all samples, which is the meaning of a 95% confidence interval.

For a 90% confidence interval, we look for the value of z that corresponds to an area of 5%, which is a z of 1.645. For a 99% confidence interval, we look for the value of z that corresponds to an area of 0.5%, which is a z of 2.576. In both cases, the value of z is interpolated using the two neighboring values from the table. For example, from the table we see that z of 2.56 corresponds an area of 0.523% and a z of 2.58 corresponds to an area of 0.494%; to find the value z that corresponds to an area of 0.500% we set up the following equation

$$\frac{2.56 - z}{2.56 - 2.58} = \frac{0.523 - 0.500}{0.523 - 0.494}$$

and solve for z, obtaining a result of 2.576.

The 90%, 95%, and 99% confidence intervals for a single 1.69-oz bag of plain M&Ms are

90% confidence interval: $48.98\text{ g} \pm (1.645)(1.433\text{ g}) = 48.98\text{ g} \pm 2.36\text{ g}$

95% confidence interval: $48.98\text{ g} \pm (1.96)(1.433\text{ g}) = 48.98\text{ g} \pm 2.81\text{ g}$

99% confidence interval: $48.98\text{ g} \pm (2.576)(1.433\text{ g}) = 48.98\text{ g} \pm 3.69\text{ g}$

For the data in Table 2, four of the samples have a net weight that falls outside of our 90% confidence interval, which extends from 46.62 g to 51.34 g; these samples are number 6 (46.405 g), number 29 (46.577 g), number 26 (51.682 g), and number 20 (51.730 g). With 30 samples and a 90% confidence interval, we expect, on average, to find that three samples have net weights that fall outside the confidence interval; finding four such samples is not an unreasonable outcome given that each sample represents just 3.3% of our pool of 30 samples.

Investigation 27. Suppose we draw four 1.69-oz bags of M&Ms from a population for which μ is 48.98 g and σ is 1.433 g. What are the 90%, the 95% and the 99% confidence intervals for the mean, $\bar{x}$, of these samples? Prepare a plot that shows how n affects the width of the 95% confidence interval, expressed as $\pm z\sigma/\sqrt{n}$, and discuss the significance of your plot. Suppose we wish to decrease the confidence interval by a factor of $3\times$ solely by increasing the number of samples taken. If the original confidence interval is based on the mean of four samples, how many additional samples must we acquire?

The 90%, 95%, and 99% confidence intervals for the mean of four 1.69-oz bag of plain M&Ms are

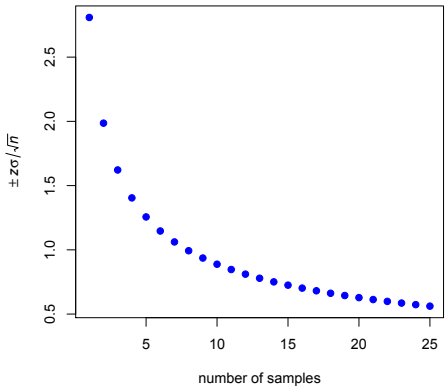
90% confidence interval: $48.98\text{ g} \pm (1.645)(1.433\text{ g})/\sqrt{4} = 48.98\text{ g} \pm 1.18\text{ g}$

95% confidence interval: $48.98\text{ g} \pm (1.96)(1.433\text{ g})/\sqrt{4} = 48.98\text{ g} \pm 1.40\text{ g}$

99% confidence interval: $48.98\text{ g} \pm (12.576)(1.433\text{ g})/\sqrt{4} = 48.98\text{ g} \pm 1.85\text{ g}$

Note that each confidence interval is half of that for the analysis of a single sample because the square root of the number of samples is two.

The figure below shows how the confidence interval changes as we increase the value of n from 1 to 25. The most important feature of this plot is to note how the rate at which the confidence interval's width becomes smaller slows down as we increase the number of samples. For example, increasing the number of samples from one to four, decreases the confidence interval by a factor of 2× from ±2.81 to ±1.40. A further two-fold decrease in the confidence interval to ±0.70 requires 16 total samples, or an additional 12 samples.



To achieve a three-fold improvement in the width of the confidence interval—that is, a decrease in the confidence interval's width by a factor of three—requires that we increase the number of samples from n_1 to n_2, where

$$\frac{\frac{z\sigma}{\sqrt{n_1}}}{\frac{z\sigma}{\sqrt{n_2}}} = \frac{\sqrt{n_2}}{\sqrt{n_1}} = 3$$

Solving shows that the ratio n_2/n_1 is 3^2 or 9; thus, if the original confidence interval is based on four samples, then to achieve the desired smaller confidence interval we need a total of $9 \times 4 = 36$ samples, or an additional 32 samples.

Investigation 28. Our data for 1.69-oz bags of plain M&Ms includes 30 measurements of the net weight. What are the 90%, the 95% and the 99% confidence intervals for the mean, $\bar{x}$, of these samples? Using the 99% confidence interval as an example, explain the meaning of this confidence interval. Is the stated net weight of 1.69 oz a reasonable estimate of the true mean for the population of 1.69-oz bags of plain M&Ms?

The 90%, 95%, and 99% confidence intervals for the mean of 1.69-oz bag of plain M&Ms are

90% confidence interval: $48.98\text{ g} \pm (1.699)(1.433\text{ g})/\sqrt{30} = 48.98\text{ g} \pm 0.45\text{ g}$

95% confidence interval: $48.98\text{ g} \pm (2.045)(1.433\text{ g})/\sqrt{30} = 48.98\text{ g} \pm 0.54\text{ g}$

99% confidence interval: $48.98\text{ g} \pm (2.756)(1.433\text{ g})/\sqrt{30} = 48.98\text{ g} \pm 0.72\text{ g}$

Using the 99% confidence interval as an example, and assuming that there are no errors in our measurements, there is a 99% probability that the confidence interval's range of net weights—from a low of 48.26 g to a high of 49.70 g—includes the true mean for the population of all 1.69-oz bags of plain M&Ms; there is a 1% probability that population's mean falls outside of this range. Given that the 99% confidence interval does not include the stated net weight of 1.69 oz (47.9 g), we can safely conclude that 1.69 oz is not a good estimate for the population's mean net weight.

Investigation 29. In 1996, Mars, the manufacturer of M&Ms, reported the following distribution for the colors of plain M&Ms: 30% brown, 20% red, 20% yellow, 10% blue, 10% green, and 10% orange. Pick any one color of M&Ms and, using the data in Table 2, calculate the percentage of that color in each of the 30 samples. Report the mean and the standard deviation for your color and use a *t*-test to determine whether your sample's mean is consistent with the result reported by Mars. Gather results for the remaining five colors from other students and discuss your pooled results. Assuming that the distribution of colors reported by Mars is correct, what can you conclude about the manufacturing process.

For this problem the null hypothesis is $H_0: \bar{x} = \mu$ and the alternative hypothesis is $H_A: \bar{x} \neq \mu$, where $\bar{x}$ is the mean of the 30 samples for the color of interest and μ is the mean reported by Mars for the color of interest. The following table summarizes the results of the *t*-test, by color

color	$\bar{x}$ (%)	s (%)	μ (%)	t	reject H_0 at
brown	25.81	5.004	30	4.581	$\alpha < 0.01$
red	17.64	6.600	20	1.962	$0.05 < \alpha < 0.10$
yellow	25.52	7.594	20	3.984	$\alpha < 0.01$
blue	11.75	6.667	10	1.440	$\alpha > 0.10$
green	6.52	4.743	10	4.018	$\alpha < 0.01$
orange	12.75	4.983	10	3.024	$\alpha < 0.01$

where s is the standard deviation for the 30 samples, t is the calculated experimental value of t based on $\bar{x}$, s, and μ, and the last column defines the value of α for which we can reject the null hypothesis and accept the alternative hypothesis.

With the exception of the color blue, we have good evidence that the distribution of colors for these 30 samples is not in agreement with the manufacturer's stated distribution. For brown, yellow, green, and orange, the 99% confidence interval does not include μ; for red, the 90% confidence interval does not include μ. These results are not particularly suprising. Although the distribution of colors in a production batch presumably matches the percentages provided by Mars, the mixing of the M&Ms at the level at which individual bags are filled likely is far from homogeneous.

In a response to a query regarding the proportions of colors in bags of M&Ms, the manufacturer noted that "[e]ach large production batch is blended to [this] ratio and mixed thoroughly. However, since the individual packages are filled by weight on high-speed equipment, and not by count, it is possible to have an unusual color distribution." The full response is at https://www.exeter.edu/documents/mandm.pdf (accessed 12/13/15). It seems more likely that the color distribution is a function of sampling uncertainty associated with the population's homogeneity at the level of sampling.

Part VI: Now Its Your Turn!

Answers, of course, will vary.

In addition to duplicating some of the questions considered in this case study using the additional data sets or newly collected data, students should be able to extend the concept of a t-test to include comparisons of different colors of M&Ms or off different types of M&Ms, and a consideration of paired vs. unpaired data. Students also can explore their ability to model data using the binomial distribution to model the distribution of samples of fixed size, or using the Poisson distribution to model the frequency of a rare event, such as the presence of damaged M&Ms in bags of a fixed size.

Other types of analyses that go beyond what is presented in this case study include linear regression (for example, mass as a function of diameter using Data Set 2), analysis of variance, and quality control charts.

Appendix

The following R code, with comments, was used to generate the net weights for the 30 samples. The file data.csv was identical in structure to Table 2, but did not include the final column of net weights. This code reads in the original data, calculates the total number of M&Ms in each sample, draws the appropriate number of M&Ms for each sample and calculates the average weight of an M&M in the sample, calculates the net weight of M&Ms in each sample, and adds the net weights to the original data and saves the data as a new file.

```
# Create data frame 'rawdata' to store data for the samples
   rawdata = read.csv("data.csv")
# Create vector 'total' to store number of M&Ms in each sample, calculated
# by summing, by row (1), the number of M&Ms in columns 3-8 of 'rawdata'
   total = apply(rawdata[ , 3:8], 1, sum)
# Create vector 'avg.weight' to store average weight of M&Ms in each sample
   avg.weight = seq(1:30)
# Create separate vectors for the population mean and the standard deviation
# with values determined using the masses of 462 plain M&Ms available at
# the Puget Sound Data Hoard (http://stat.pugetsound.edu/hoard/Default.aspx)
   mu = 0.86483
   sig = 0.046199
# For each sample, calculate the average mass for its M&Ms, with the number
# of M&Ms defined by the vector 'total,' using a random draw from a normal
# distribution defined by the population's mean standard deviation
   for (i in 1:30) avg.weight[i] = mean(rnorm(total[i], mu, sig))
# Find the net weight for each sample by multiplying the number of M&Ms in
# the sample by the average weight of an M&M in the sample
   net.weight = total * avg.weight
# Create final data set by adding net.weight to the original data and save
   mmdata = data.frame(rawdata, net.weight)
   write.csv = (mmdata, file = "mmdata.csv")
```