

Introduction to Data and the Analysis of Data

At the heart of any analysis is data. Sometimes our data is categorical and sometimes it is numerical; sometimes our data conveys order and sometimes it does not; sometimes our data has an absolute reference and sometimes it has an arbitrary reference; and sometimes our data takes on discrete values and sometimes it takes on continuous values. Whatever its form, when we gather data our intent is to extract from it information that can help us solve a problem. In this case study we consider how to find meaning in data, including ways to describe data, to visualize data, to summarize data, to model data, and to draw conclusions from data.

If we are to consider how to describe, to visualize, to summarize, to model, and to draw conclusions from data, then we need some data with which we can work. For the purpose of this case study, we need data that is easy to gather and easy to understand, and that allows us to ask interesting questions; it is helpful, as well, if we can find expected results for at least some of our questions so that we can check our analysis. It also is helpful if you can gather your own data so that you can repeat and verify our work, or so that you can extend our analysis. A simple system that meets these criteria is to analyze the contents of bags of M&Ms. There is a rich history of using M&Ms to introduce or to illustrate the analysis of data in a variety of disciplines; Appendix 1 provides examples of such studies. Although this system may seem trivial, keep in mind that reporting the percentage of yellow M&Ms in a bag is analogous to reporting the concentration of Pb^{2+} in a sample of soil: both express the amount of an analyte present in a unit of its matrix.

Interspersed within the case study's narrative are a series of investigations, each of which asks you to stop and consider one or more important issues. Some of these investigations include data for you to analyze; you can use this link (<http://bit.ly/1ZokZKf>) to access an on-line version of this case study. For additional coverage of the topics in this case study, see Chapter 4 of *Analytical Chemistry 2.0*, which is available using this link (<http://bit.ly/1r3wJoz>), and the additional resources cited within.

Part I: Ways to Describe Data

In the introduction to this case study we identified four contrasting ways to describe data: categorical vs. numerical, ordered vs. unordered, absolute reference vs. arbitrary reference, and discrete vs. continuous. To give meaning to these descriptive terms, let's consider the data in Table 1.

TABLE 1. DISTRIBUTION OF YELLOW AND RED M&Ms

bag id	year	weight (oz)	type	# yellow M&Ms	% red M&Ms	total M&Ms	rank (total M&Ms)
a	2006	1.74	peanut	2	27.8	18	sixth
b	2006	1.74	peanut	3	4.3	23	fourth
c	2000	0.80	plain	1	22.7	22	fifth
d	2000	0.80	plain	5	20.8	24	third
e	1994	10.0	plain	56	23.0	331	second
f	1994	10.0	plain	63	21.9	333	first

The entries in Table 1 are organized by column and by row, with the first row (shaded here for emphasis) identifying the variables used to describe the data. Each additional row is the record for one sample and each entry in a sample's record provides information about one of its variables; thus, the data in the table lists the result for each variable and for each sample.

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.

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?

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.

We also can use a numerical variable to assign samples to groups. For example, we can divide the plain M&Ms in Table 1 into two groups based on the sample's weight. What makes a numerical variable more interesting, however, is that we can use it to make quantitative comparisons between samples; thus, we can report that there are $14.8\times$ as many plain M&Ms in a 10-oz. bag as there are in a 0.8-oz. bag. Although we can complete meaningful calculations using any numerical variable, the type of calculation we can perform depends on whether or not the variable's values have an absolute reference.

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?

Finally, the granularity of a numerical variable provides one more way to describe our data.

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.

Part II: Ways to Visualize Data

Suppose we are interested in characterizing 1.69-oz (47.9-g) packages of plain M&Ms. We obtain 30 bags (ten from each of three stores) and, for each bag, report the number of blue, brown, green, orange, red, and yellow M&Ms—for yellow, the number in parentheses is the number of yellow M&Ms in the first five drawn from the bag—and their combined net weight. Table 2 summarizes the data for the last six samples. The full set of data for all 30 samples is available as a separate spreadsheet or R file.

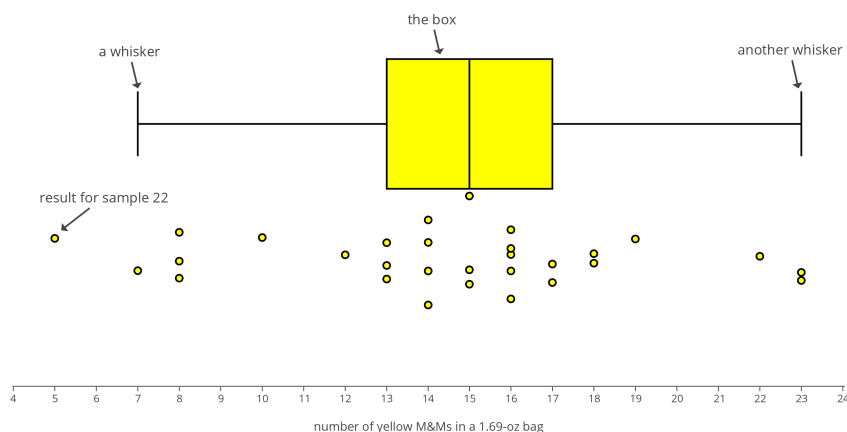
TABLE 2. SOURCE, DISTRIBUTION, AND NET WEIGHT OF PLAIN M&Ms IN 1.69-OZ BAGS

bag	store	blue	brown	green	orange	red	yellow	net weight (g)
25	CVS	7	13	0	4	15	16 (2)	48.212
26	Target	6	15	1	13	10	14 (1)	51.682
27	CVS	5	17	6	4	8	19 (1)	50.802
28	Kroger	1	21	6	5	10	14 (0)	49.055
29	Target	4	12	6	5	13	14 (2)	46.577
30	Kroger	15	8	9	6	10	8 (1)	48.317

Having collected some data, our next step is to examine it for possible problems, such as missing values or errors introduced when we recorded the data, or to identify important variables and interesting patterns or trends within or between these variables. Although this information is embedded within the data itself, often it is difficult to see it when the data is displayed as a table, particularly if the data set is large in size. Instead, we use one or more simple visualizations of the data.

Two simple visualizations are box and whisker plots and dot plots, examples of which are shown in Figure 1 using the data for yellow M&Ms. Note that neither plot has meaningful information along the y -axis as the vertical dimension simply helps us visualize the data. The vertical distribution of points in the dot plot, for example, is the result of jittering, which offsets samples that share a common number of yellow M&Ms so that, we hope, each appears as a distinct point.

Figure 1: Box and Whisker Plot and Dot Plot for Yellow M&Ms



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.

As suggested by the next two investigations, one way to use a box and whisker plot is to look for unexpected features in our data that merit attention, such as an oddly shaped distribution of results or an unusually large or an unusually small result for a variable.

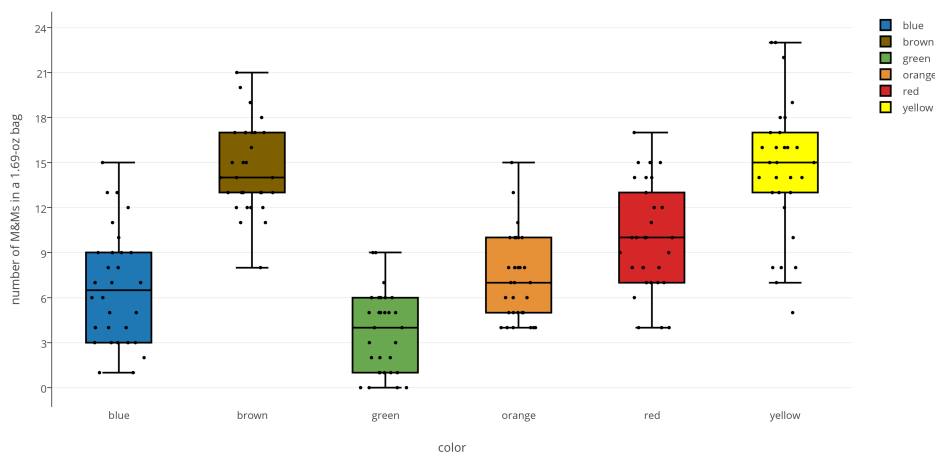
Investigation 7. The box and whisker plot in Figure 1 is perfectly symmetrical in that each side of the box is two units from the box's middle line, and each whisker is six units from the box's nearest edge. What does this symmetry suggest about how the results are distributed? Is the actual distribution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

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?

In addition to providing us with insight into the results for a single variable, we can use box and whisker plots and dot plots to examine differences between variables and differences within a single variable when we can divide that variable into different groups.

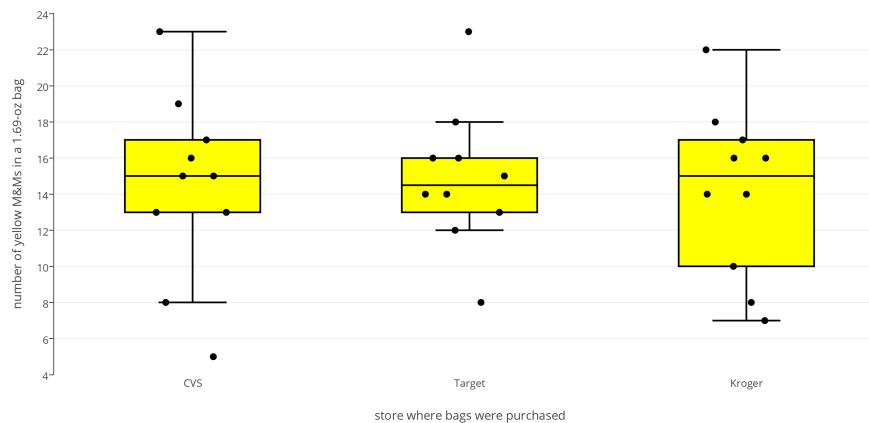
Investigation 9. Figure 2 shows box and whisker plots and dot plots for all six colors of M&Ms included in Table 2 (note: even with jittering, you will not be able to see all 30 samples in these dot plots). Based on these plots, where do you see similarities and where do you see differences in the distribution 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?

Figure 2: Box and Whisker Plots for all Colors of 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.

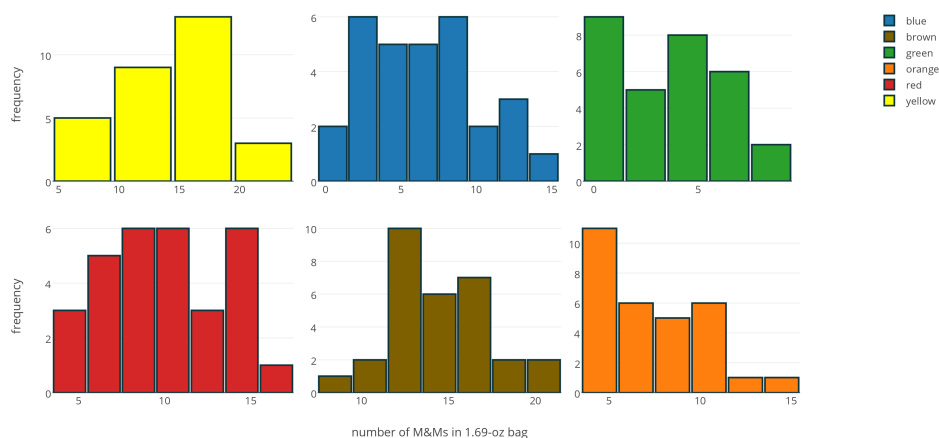
Figure 3: Box and Whisker Plot for Yellow M&Ms Grouped by Store



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.

Although a box and whisker plot provides some evidence of how a variable's values are distributed, it is not particularly easy to see the shape of that distribution. For this we use a histogram, which displays the number of results that fall within a sequence of (usually) equally spaced bins. Figure 4, for example, shows histograms for each color of M&Ms in our data set.

Figure 4: Histograms by Color in 1.69-oz Bags of M&Ms



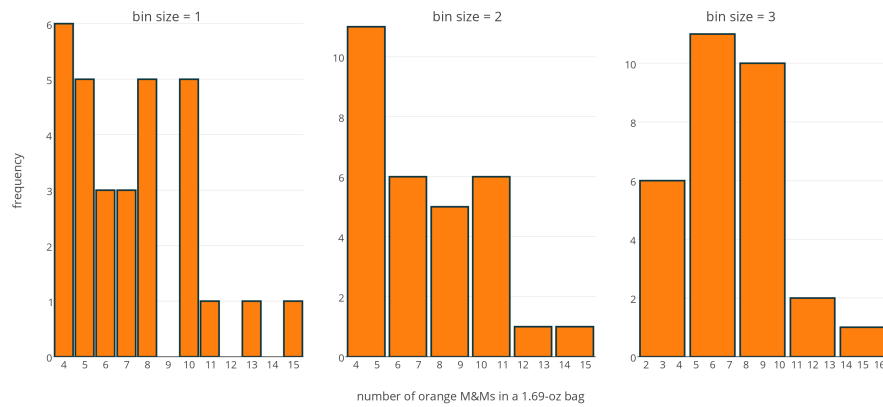
Investigation 12. For the histograms in Figure 4, where do you see similarities and where do you see differences in the distribution of M&Ms? How do the results seen here compare with your interpretation of the box and whisker plots and the dot plots in Figure 2?

One challenge when we draw a histogram is choosing the width for the bins or the number of bins. In Figure 4, for example, the bins for yellow M&Ms are five units wide—the first bin, for example, includes samples with 5, 6, 7, 8, and 9 yellow M&Ms—but the bins are two units wide for all other colors of M&Ms. There are no simple rules for determining the number or the width of bins, so it is a good idea to try several bin sizes before we settle on a final choice.

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.

Figure 5: Histogram for Orange M&Ms Using Different Bin Sizes



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.

Part III: Ways to Summarize Data

Although box and whisker plots, dot plots, and histograms help us see qualitative patterns in our data, they do not allow us to express this information in a quantitative way. For example, in Figure 3 and in Investigation 10 we learned that the distribution of yellow M&Ms in 1.69-oz bags is relatively similar between the three different sources, although the plot for samples purchased from Target has much shorter whiskers and the individual results seem more tightly clustered than is the case for samples purchased at CVS and at Kroger, and the box for the samples purchased from Kroger is quite a bit wider than is the case for the samples from CVS and Target.

Qualitative phrases such as “relatively similar,” “much shorter,” “more tightly clustered,” and “quite a bit wider” are, frankly, fuzzy, but in the absence of a more quantitative way to characterize our data, we have little choice but to adopt such fuzzy terms. When we summarize data, our goal is to report quantitative characteristics, or statistics, that we can use to provide clearer statements about the differences and the similarities between results for different variables, or between the results for a variable and an expected result already known to us. In this part of the case study we consider several useful statistics that we can use to summarize the data for our samples.

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.

Both a box and whisker plot and a histogram suggests that the distribution of results for a single variable has two important features: its center, which presumably lies somewhere in the middle of the data, and its spread, which is suggested by the length of the whiskers in a box and whisker plot, or how quickly or how slowly the counts in a histogram’s bins decrease as we move away from the bin that has the most counts. For our purposes, we will consider two quantitative measures of central tendency—the mean and the median—and four quantitative measures of spread: the variance, the standard deviation, the range, and the interquartile range.

Central Tendency. The mean, $\bar{x}$, is the arithmetic average of all n of a variable’s results; thus

$$\bar{x} = \frac{\sum x_i}{n}$$

where x_i is the result for an individual sample. The median is the middle value when the n results are ordered by rank from smallest-to-largest. If n is odd, then the median is the $(n + 1)/2^{\text{th}}$ value; if n is even, then the median is the average of the $(n/2)^{\text{th}}$ value and the $((n/2) + 1)^{\text{th}}$ value.

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?

Spread. A sample’s variance, s^2 , provides an estimate of the average squared deviation of its n results relative to its mean; thus

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

where x_i is the result for an individual sample and $\bar{x}$ is the variable's mean value. The standard deviation, s , is the square root of the variance.

The range is the difference between the sample's largest value and its smallest value. A variable's interquartile range, IQR , spans the middle 50% of its values. To find the IQR , we order the data from smallest-to-largest, and separate it into two equal parts; if the sample has an odd number of values, then we do not include the median in either part. Next, we find the median for each of the two parts. The IQR is the difference between these two medians. Note: There actually are several methods for calculating the IQR , which differ in how they divide the data into four parts. As you might expect, different methods may result in different values for the IQR . The method described here was used to create the box and whisker plots in Figures 1–3, where the width of the box is the interquartile range.

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 measure of spread 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?

Part IV: Ways to Model Data

In Part III we made a distinction between a sample and a population, noting that a population is every member of a system that we could analyze and that a sample is the discrete subset of a population that we actually analyze. We collect and analyze samples with the hope that we can use their properties to deduce something about the population's properties. We accomplish this by using suitable mathematical models.

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?

There are a variety of ways in which we might model our data, three of which we consider in this section: the binomial distribution, the Poisson distribution, and the normal distribution.

Binomial Distribution. A binomial distribution describes the probability, P , of a particular event, X , during a fixed number of trials, N , given the probability, p , that the event happens during a single trial. Mathematically, we express the binomial distribution as

$$P(X, N) = \frac{N!}{X!(N - X)!} \times p^X \times (1 - p)^{N - X}$$

where $!$ is the factorial symbol. The theoretical mean, μ , and the theoretical variance, σ^2 , for a binomial distribution are

$$\mu = Np \quad \sigma^2 = Np(1 - p)$$

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.

Poisson Distribution. The binomial distribution is useful if we wish to model the probability of finding a fixed number of yellow M&Ms in a sample of M&Ms of fixed size, but not the probability of finding a fixed number of yellow M&Ms in a single bag.

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.

To model the number of yellow M&Ms in packages of M&Ms, we use the Poisson distribution, which gives the probability of a particular event, X , given an average rate, λ , for that event. Mathematically, we express the Poisson distribution as

$$P(X, \lambda) = \frac{e^{-\lambda} \lambda^X}{X!}$$

The theoretical mean, μ , and the theoretical variance, σ^2 , are both equal to λ .

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.

Normal Distribution. The binomial distribution and the Poisson distribution are examples of discrete functions in that they predict the probability of a discrete event, such as finding exactly two green M&Ms in the next bag of M&Ms that we open. Not all data we might collect on M&Ms, however, is discrete.

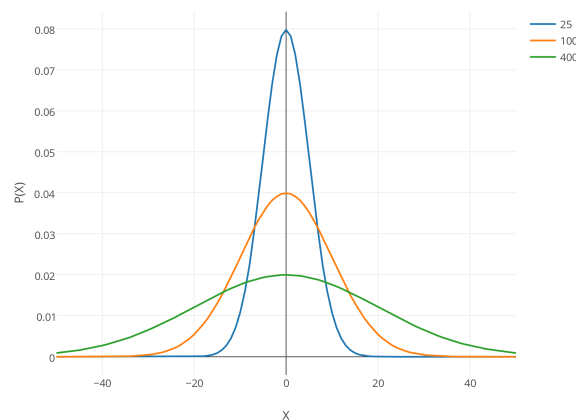
Investigation 22. Explain why we cannot use the binomial distribution or the Poisson distribution to model data for the net weight of M&Ms in Table 2.

To model the net weight of packages of M&Ms, we use the normal distribution, which gives the probability of obtaining a particular outcome from a population with a known mean, μ , and a known variance, σ^2 . Mathematically, we express the normal distribution as

$$P(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

Figure 6 shows the normal distribution curves for $\mu = 0$ and for variances of 25, 100, and 400.

Figure 6: Normal Distribution Curves for Mean of 0 and Variances of 25, 100, and 400



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?

Because the equation for a normal distribution depends solely on the population's mean, μ , and variance, σ^2 , the probability that a sample drawn from a population has a value between any two arbitrary limits is the same for all populations. For example, 68.26% of all samples drawn from a normally distributed population will have values within the range $\mu \pm \sigma$, and only 0.621% will have values greater than $\mu + 2.5\sigma$; see Appendix 2 for further details.

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 a binomial distribution, if $N \times p \geq 5$ and $N \times (1 - p) \geq 5$, then a normal distribution closely approximates a binomial distribution; the same is true for a Poisson distribution when $\lambda \geq 20$.

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?

Part V: Ways to Draw Conclusions From Data

In Part IV we noted that when a population is normally distributed, the probability of obtaining a particular result for any single sample is determined by that result's area under the normal distribution curve defined by the population's mean and standard deviation. For example, in Investigation 24 we showed that for 1.69-oz bags of plain M&Ms, 22.8% have a net weight less than 1.69 oz if the population's mean is 48.98 g and its standard deviation is 1.433 g.

Suppose we select a single sample from this population: What can we predict about the net weight of M&Ms in that sample? Rearranging our equation for z , we find that

$$x = \mu \pm z\sigma$$

We call this equation a confidence interval because the value we choose for z defines the probability (our confidence) that the result for a single sample is in the range $\mu \pm z\sigma$.

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 interval, and what value of z corresponds to a 99% confidence interval? 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?

In Investigation 26 we calculated the confidence interval for a single sample based on the properties of the population from which we obtained the sample. If we draw several replicate samples from this population and calculate their mean, $\bar{x}$, then the confidence interval becomes

$$\bar{x} = \mu \pm \frac{z\sigma}{\sqrt{n}}$$

where n is the number of 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?

In both Investigation 26 and Investigation 27 we attempt to predict a property of a sample based on a population with known values of μ and σ . For most practical analytical problems, however, we need to work in the opposite direction, using the sample's mean, $\bar{x}$, and its standard deviation, s , to predict the population's mean, μ . To do this, we make three modifications to our equation for the confidence interval: we rewrite the equation so that it expresses μ in terms of $\bar{x}$; we replace the population's standard deviation, σ , with the sample's standard deviation, s ; and we replace z with the variable t , where we define t such that, for any confidence level, $t \geq z$ and the value of t approaches z as the number of samples, n , increases.

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}}$$

Clearly the value of t depends on the confidence interval and the number of samples; see Appendix 3 for further details.

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?

Our approach in Investigation 28 suggests we can use a confidence interval to decide whether a known value is consistent with our results, a process that we call significance testing and that we carry out a bit more formally than suggested by Investigation 28. To illustrate the process, we will use the data from Table 2 for the bags of M&Ms purchased at Target and evaluate whether the mean net weight for these samples is consistent with the stated net weight of 1.69 oz (47.9 g).

To begin, we summarize the experimental results for our sample, which in this case is a mean of 49.52 g and a standard deviation of 1.649 g for $n = 10$ samples. Next, we state our problem in the form of a yes/no question, the answers to which we define using a null hypothesis (H_0) and an alternative hypothesis (H_A); for example, for this problem our yes/no question is “Is the mean of the samples consistent with the stated net weight of 1.69 oz?,” which we define as

$$H_0: \bar{x} = \mu \text{ (yes)}$$

$$H_A: \bar{x} \neq \mu \text{ (no)}$$

where $\bar{x}$ is 49.52 g and μ is 47.9 g. To evaluate the two hypotheses, we rewrite the equation for the confidence interval so that we can solve for t

$$t = \frac{|\bar{x} - \mu|\sqrt{n}}{s} = \frac{|49.52 - 47.9|\sqrt{10}}{1.649} = 3.087$$

Finally, we compare this experimental value of t to the critical values of t for the correct number of degrees of freedom (in this case, $\nu = n - 1 = 10 - 1 = 9$). From Appendix 3 we see that $t(\alpha, \nu)$ is 1.833 for a 90% confidence interval (an α of 0.10), 2.262 for a 95% confidence interval (an α of 0.05), 2.821 for a 98% confidence interval (an α of 0.02), and 3.250 for a 99% confidence interval (an α of 0.01). Our experimental value for t of 3.087 falls between the critical values for the 98% and the 99% confidence interval; if we are willing to accept an uncertainty of 1–2%, then we can reject the null hypothesis and accept the alternative hypothesis, concluding that the mean of 49.52 g is not consistent with the stated net weight of 1.69 oz. We call this a t -test of $\bar{x}$ vs. μ .

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.

Part VI: Now Its Your Turn!

Using the data from Table 2, data from the references in Appendix 1, data from one or more of the four data sets listed below, and/or data you collect on your own, pose a question, gather relevant data, complete an analysis of that data, and propose an answer to your question. Although this case study's introduction to data and to data analysis is limited to just a few types of visualizations, a few ways to summarize data, a few ways to model data, and one way to draw conclusions from data, your background is sufficient to use the texts in Appendix 1 and your textbook to explore other ways to visualize, to summarize, to model, and to draw conclusions from data. Be adventurous!

Data Set 1

source: Math Department, University of Alabama at Huntsville

website: <http://www.math.uah.edu/stat/data/MM.html>

structure: color distribution and net weight for 30 samples of plain M&Ms in 1.69-oz bags

Data Set 2

source: University of Puget Sound Data Hoard

website: <http://stat.pugetsound.edu/hoard/datasetDetails.aspx?id=1>

structure: type, color, diameter, and mass for M&Ms in a 14.0-oz bag of plain M&Ms, a 12.7-oz bag of peanut M&Ms, and a 12.7-oz bag of peanut butter M&Ms

Data Set 3

source: Stats Monkey

website: http://apstatsmonkey.com/StatsMonkey/m%26m_Activities.html

structure: type and color (all data is simulated)

Data Set 4

source: various

Excel spreadsheet: [link](#)

structure: color distribution of plain M&Ms attributed to Mars (1996–2008)

Appendix 1: External Resources for Teaching and Learning Statistics

The papers and websites gathered here provide examples of studies using M&Ms (and other similar items) to illustrate concepts in the broad area of data analysis.

- Canaes, L. S.; Brancalion, M. L.; Rossi, A. V.; Rath, S. “Using Candy Samples to Learn about Sampling Techniques and Statistical Data Evaluation,” *J. Chem. Educ.* **2008**, *85*, 1083–1088 (<http://pubs.acs.org/doi/abs/10.1021/ed085p1083>).
- Diamond, J. J. “Using Peanut M&M’s in an Introductory Statistics Class Illustrate Binomial Properties,” *The Statistics Teacher Network*, **2010**, *75*, 2–4 (<http://www.amstat.org/education/stn/pdfs/stn75.pdf>).
- Downey, A. B. *Think Bayes*, Green Tea Press (<http://www.greenteapress.com/thinkbayes/thinkbayes.pdf>).
- Duncan, D. R.; Litwiller, B. H. “Milk Chocolate M&M Color Distribution: A Chi-Square Experience,” *Illinois Mathematics Teacher*, **2008**, *Spring*, 32–33 (<http://ictm.org/journal/index.php/imt/article/view/46/44>).
- Fricker, R. D. Jr. “The Mysterious Case of the Blue M&Ms®,” *Chance* **1996**, *9*(4), 19–22 (<http://faculty.nps.edu/rdfricke/docs/MandM.pdf>).
- Froelich, A. G.; Stephenson, W. R. “How Much do M&M’s Weigh?” *Teaching Statistics*, **2012**, *35*, 14–20 (<http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9639.2012.00515.x/abstract>).
- Holland, E. T.; Manley, G.; Chiba, T.; Ramos, R.; Mochrie, S.; Frederick, J. “Infectious Chocolate Joy with a Side of Poisson Statistics: An Activity Connecting Life Science Students with Subtle Physics Concepts,” (<http://www.coursesource.org/courses/infectious-chocolate-joy-with-a-side-of-poissonian-statistics-an-activity-connecting-life#tabs-0-content=1>).
- Johnson, R. W. “Testing Colour Proportion of M&Ms,” *Teaching Statistics*, **1993**, *15*, 2–4 (<http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9639.1993.tb00243.x/abstract>).
- Juneau, J.; Juneau, K. R.; Coates, E. R. “A Sweet Demonstration of Statistical Hypothesis Testing,” (<http://se.asee.org/proceedings/ASEE1998/Juneau02.pdf>).
- Lee, H. K. H. “Chocolate Chip Cookies as a Teaching Aid,” *The American Statistician* **2007**, *61*(4), 1–5 (<http://amstat.tandfonline.com/doi/abs/10.1198/000313007X246905>).
- Lin, T.; Sanders, M. S. “A Sweet Way to Learn DoE,” *Quality Progress*, **2006**, *February*, 88 (<http://asq.org/quality-progress/2006/02/one-good-idea/a-sweet-way-to-learn-doe.html>).
- Peterson, I. “A Taste for M&M’s,” (http://www.rci.rutgers.edu/~mmm431/quant_methods_F13/MnMs_web_exercise.pdf).
- Ross, M. R. “A Classroom Exercise in Sampling Techniques,” *J. Chem. Educ.* **2000**, *77*, 1015–1016 (<http://pubs.acs.org/doi/abs/10.1021/ed077p1015>).
- Schwartz, T. A. “Teaching Principles of One-Way Analysis of Variance Using M&M’s Candy,” *J. Stats. Educ.* **2013**, *21* (www.amstat.org/publications/jse/v21n1/schwartz.pdf).
- Stat Monkey “m&m Statistics Activities,” (http://apstatsmonkey.com/StatsMonkey/m%26m_Activities.html).
- Staub, N. L. “Teaching Evolutionary Mechanisms: Genetic Drift and M&M’s®,” *BioScience*, **2002**, *52*, 373–377 (<http://bioscience.oxfordjournals.org/content/52/4/373.short>).
- University of Puget Sound Data Hoard (<http://stat.pugetsound.edu/hoard/datasetDetails.aspx?id=1>).
- Xu-Friedman, M. A. “Illustrating Concepts of Quantal Analysis with an Intuitive Classroom Model,” *Adv. Physiol. Educ.* **2013**, *37*, 112–116 (<http://advan.physiology.org/content/37/1/112>).

In addition to the many standard textbooks used in introductory and advanced courses in analytical chemistry, these references introduce additional ways to visualize, to summarize, to model, and to draw conclusions from data.

- Boslaugh, S. *Statistics in a Nutshell*, 2nd Edition, O'Reilly: Sebastopol, CA, 2013.
- Harvey, D. T. *Analytical Chemistry 2.0*, (<http://bit.ly/1r3wJoz>).
- Larose, D. T.; Larose, C. D. *Discovering Knowledge in Data*, 2nd Edition, Wiley: Hoboken, NJ, 2014.
- Miller, J. C.; Miller, J.N. *Statistics and Chemometrics for Analytical Chemistry*, Pearson: Essex, England, 2010.
- Robbins, N. *Creating More Effective Graphs*, Chart House: Wayne, NJ, 2013.
- van Belle, G. *Statistical Rules of Thumb*, 2nd Edition, Wiley: Hoboken, NJ, 2008.

Appendix 2: Single-Sided Normal Distribution

The table below gives the proportion, P , of the area under a normal distribution curve that lies to the right of a deviation, z , which is defined as

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

where μ and σ are the distribution's mean and standard deviation, respectively, and where x is the value for which the deviation is defined. For example, the area under a normal distribution to the right of a deviation of +0.04 is 0.4840 (see entry in red), or 48.40% of the total area. The area to the left of the deviation is $1 - P$. For a deviation of +0.04, this is $1 - 0.4840$, or 51.60%. If x is smaller than μ , then z is negative. In this case, the values in the table give the area to the left of z . For example, if z is -0.04 , then 48.40% of the area lies to the left of the deviation.

To use the single-sided normal distribution table, sketch the normal distribution curve for your problem and shade the area that corresponds to your answer. This divides the normal distribution curve into three regions: the shaded area that corresponds to your answer, the area to the right of this, and the area to the left of this. Calculate the values of z for the limits of the area that corresponds to your answer. Use the table to find the areas to the right and to the left of these deviations, subtract these values from 100% and, voilà, you have your answer.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4365	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4502	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0466	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102		0.00964		0.00914		0.00866	
2.4	0.00820		0.00776		0.00734		0.00695		0.00657	
2.5	0.00621		0.00587		0.00554		0.00523		0.00494	
2.6	0.00466		0.00440		0.00415		0.00391		0.00368	
2.7	0.00347		0.00326		0.00307		0.00289		0.00272	
2.8	0.00256		0.00240		0.00226		0.00212		0.00199	
2.9	0.00187		0.00175		0.00164		0.00154		0.00144	

Appendix 3: Critical Values of t

The table below gives values of $t(\alpha, \nu)$ where α defines the confidence level and ν defines the degrees of freedom. Values for α are defined as follows

$$\alpha = 1 - \text{confidence level (as fraction)}$$

For example, for a 95% confidence level, $\alpha = 1 - 0.95 = 0.05$. The degrees of freedom is the number of independent measurements given any constraints that we place on the measurements. For example, if we have n measurements and we calculate their mean, $\bar{x}$, then we have $n - 1$ degrees of freedom because the mean, $\bar{x}$, and the values for the first four measurements, x_1, x_2, x_3 , and x_4 , removes the independence of the fifth measurement, x_5 , whose value is defined exactly as

$$x_5 = \bar{x} - x_1 - x_2 - x_3 - x_4$$

The values of t in this table are two-tailed in that they define a confidence interval that is symmetrical around the mean. For example, for a 95% confidence interval ($\alpha = 0.05$), half of the area not included within the confidence interval is at the far right of the distribution and half is at the far left of the distribution. For a one-tailed confidence interval, in which the excluded area is on one side of the distribution, divide the values of α in half; thus, for a one-tailed 95% confidence interval, we use values of t from the column where $\alpha = 0.10$.

ν	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.02$	$\alpha = 0.01$
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.255
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
12	1.782	2.179	2.681	3.055
14	1.761	2.145	2.624	2.977
16	1.746	2.120	2.583	2.921
18	1.734	2.101	2.552	2.878
20	1.725	2.086	2.528	2.845
30	1.697	2.042	2.457	2.750
50	1.676	2.009	2.311	2.678
∞	1.645	1.960	2.326	2.576