Normal Distribution

Larry Ottman, (LarryO)
Ellen Lawsky, (EllenL)

Say Thanks to the Authors
Click http://www.ck12.org/saythanks
(No sign in required)

cK-12 *f*lexbook
next generation textbooks

AUTHORS
Larry Ottman, (LarryO)
Ellen Lawsky, (EllenL)

To access a customizable version of this book, as well as other interactive content, visit www.ck12.org

CK-12 Foundation is a non-profit organization with a mission to reduce the cost of textbook materials for the K-12 market both in the U.S. and worldwide. Using an open-content, web-based collaborative model termed the **FlexBook®**, CK-12 intends to pioneer the generation and distribution of high-quality educational content that will serve both as core text as well as provide an adaptive environment for learning, powered through the **FlexBook Platform®**.

Printed: August 24, 2013

f lexboo*K*
next generation textbooks

CHAPTER **1**

Normal Distribution

CHAPTER OUTLINE

1.1 The Standard Normal Probability Distribution

Learning Objectives

- Identify the characteristics of a normal distribution.
- Identify and use the Empirical Rule (68-95-99.7 Rule) for normal distributions.
- Calculate a *z*-score and relate it to probability.
- Determine if a data set corresponds to a normal distribution.

Introduction

Most high schools have a set amount of time in-between classes during which students must get to their next class. If you were to stand at the door of your statistics class and watch the students coming in, think about how the students would enter. Usually, one or two students enter early, then more students come in, then a large group of students enter, and finally, the number of students entering decreases again, with one or two students barely making it on time, or perhaps even coming in late!

Now consider this. Have you ever popped popcorn in a microwave? Think about what happens in terms of the rate at which the kernels pop. For the first few minutes, nothing happens, and then, after a while, a few kernels start popping. This rate increases to the point at which you hear most of the kernels popping, and then it gradually decreases again until just a kernel or two pops.

Here's something else to think about. Try measuring the height, shoe size, or the width of the hands of the students in your class. In most situations, you will probably find that there are a couple of students with very low measurements and a couple with very high measurements, with the majority of students centered on a particular value.

All of these examples show a typical pattern that seems to be a part of many real-life phenomena. In statistics, because this pattern is so pervasive, it seems to fit to call it normal, or more formally, the normal distribution. The normal distribution is an extremely important concept, because it occurs so often in the data we collect from the natural world, as well as in many of the more theoretical ideas that are the foundation of statistics. This chapter explores the details of the normal distribution.

The Characteristics of a Normal Distribution

Shape

When graphing the data from each of the examples in the introduction, the distributions from each of these situations would be mound-shaped and mostly symmetric. A *normal distribution* is a perfectly symmetric, mound-shaped distribution. It is commonly referred to the as a normal curve, or bell curve.
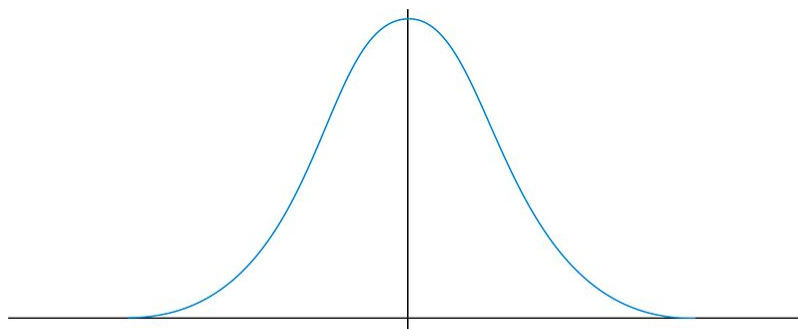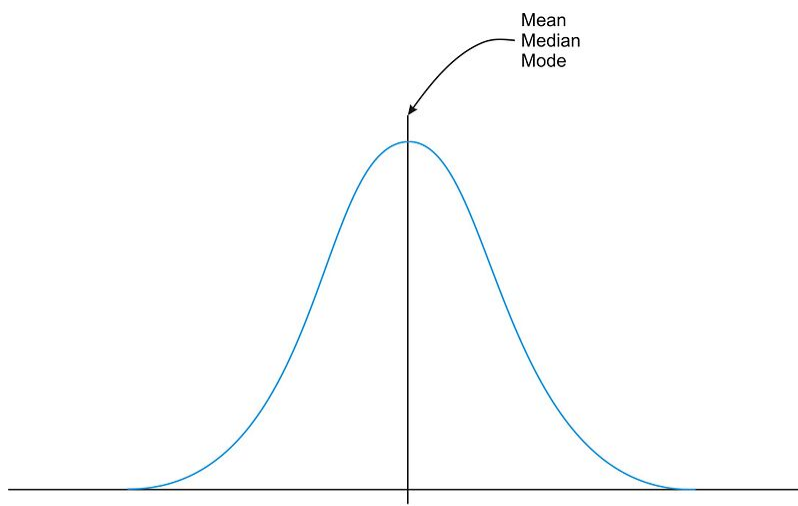
The Normal Distribution

Because so many real data sets closely approximate a normal distribution, we can use the idealized normal curve to learn a great deal about such data. With a practical data collection, the distribution will never be exactly symmetric, so just like situations involving probability, a true normal distribution only results from an infinite collection of data. Also, it is important to note that the normal distribution describes a continuous random variable.

Center

Due to the exact symmetry of a normal curve, the center of a normal distribution, or a data set that approximates a normal distribution, is located at the highest point of the distribution, and all the statistical measures of center we have already studied (the mean, median, and mode) are equal.
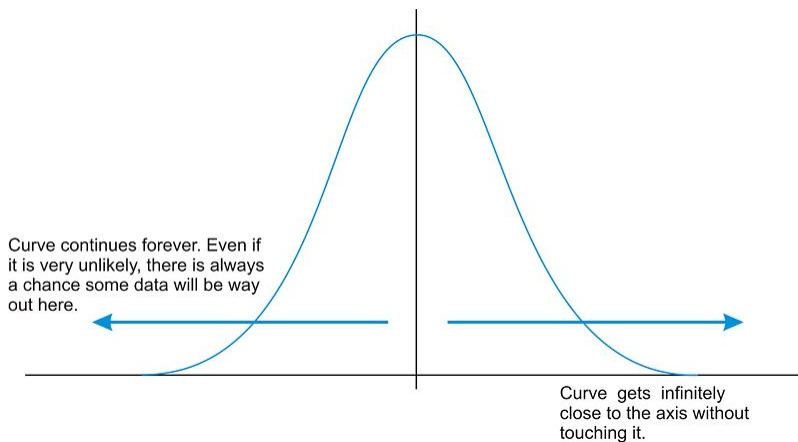
Mean
Median
Mode

Normal Distribution Center

It is also important to realize that this center peak divides the data into two equal parts.

Normal Distribution divided into equal halves by the center
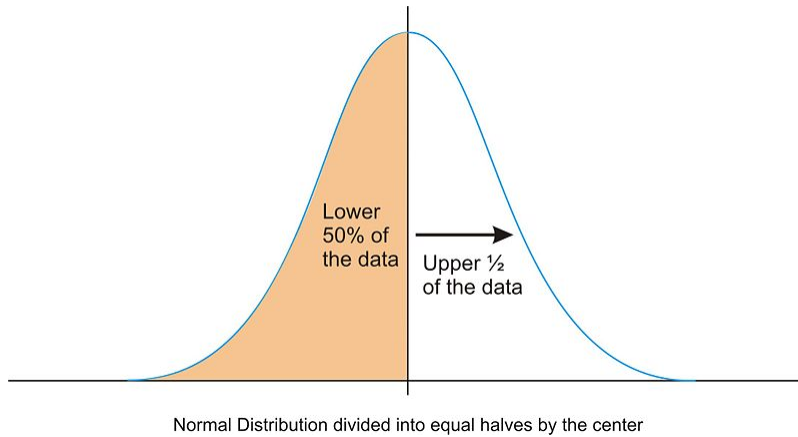
Spread

Let's go back to our popcorn example. The bag advertises a certain time, beyond which you risk burning the popcorn. From experience, the manufacturers know when most of the popcorn will stop popping, but there is still a chance that there are those rare kernels that will require more (or less) time to pop than the time advertised by the manufacturer. The directions usually tell you to stop when the time between popping is a few seconds, but aren't you tempted to keep going so you don't end up with a bag full of un-popped kernels? Because this is a real, and not theoretical, situation, there will be a time when the popcorn will stop popping and start burning, but there is always a chance, no matter how small, that one more kernel will pop if you keep the microwave going. In an idealized normal distribution of a continuous random variable, the distribution continues infinitely in both directions.

Curve continues forever. Even if it is very unlikely, there is always a chance some data will be way out here.

Curve gets infinitely close to the axis without touching it.

Because of this infinite spread, the range would not be a useful statistical measure of spread. The most common way to measure the spread of a normal distribution is with the standard deviation, or the typical distance away from the mean. Because of the symmetry of a normal distribution, the standard deviation indicates how far away from the maximum peak the data will be. Here are two normal distributions with the same center (mean):

normal distribution, standard deviation = 2

normal distribution, standard deviation = 3
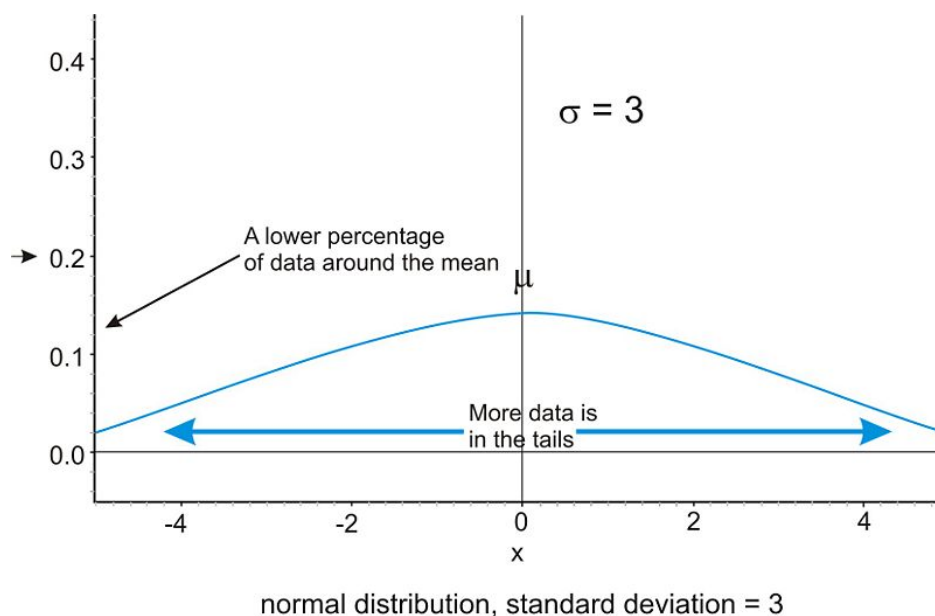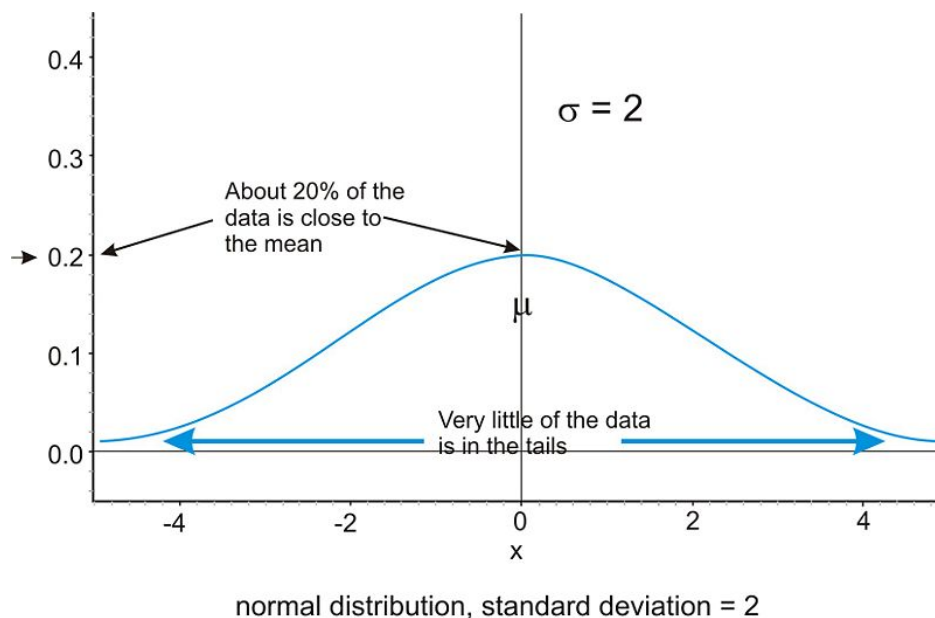
The first distribution pictured above has a smaller standard deviation, and so more of the data are heavily concentrated around the mean than in the second distribution. Also, in the first distribution, there are fewer data values at the extremes than in the second distribution. Because the second distribution has a larger standard deviation, the data are spread farther from the mean value, with more of the data appearing in the tails.

Technology Note: Investigating the Normal Distribution on a TI-83/84 Graphing Calculator

We can graph a normal curve for a probability distribution on the TI-83/84 calculator. To do so, first press **[Y=]**. To create a normal distribution, we will draw an idealized curve using something called a density function. The command is called 'normalpdf(', and it is found by pressing **[2nd][DISTR][1]**. Enter an X to represent the random variable, followed by the mean and the standard deviation, all separated by commas. For this example, choose a mean of 5 and a standard deviation of 1.
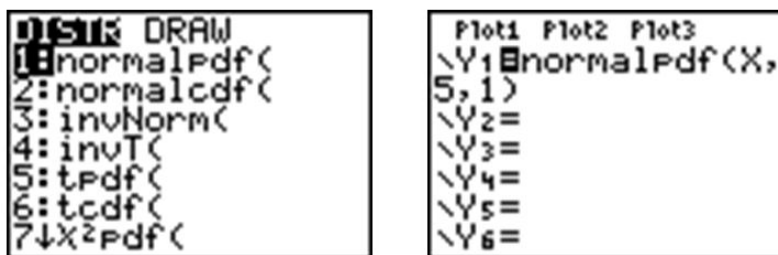
Adjust your window to match the following settings and press **[GRAPH]**.

Press **[2ND][QUIT]** to go to the home screen. We can draw a vertical line at the mean to show it is in the center of the distribution by pressing **[2ND][DRAW]** and choosing 'Vertical'. Enter the mean, which is 5, and press **[ENTER]**.
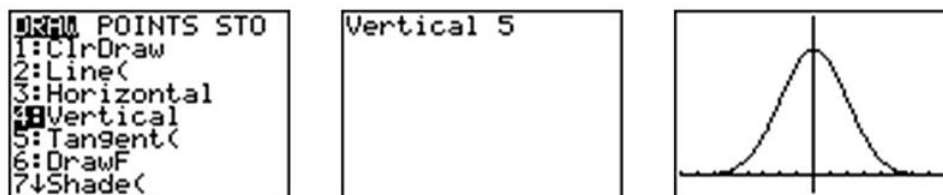
Remember that even though the graph appears to touch the *x*-axis, it is actually just very close to it.

In your **Y=** Menu, enter the following to graph 3 different normal distributions, each with a different standard deviation:
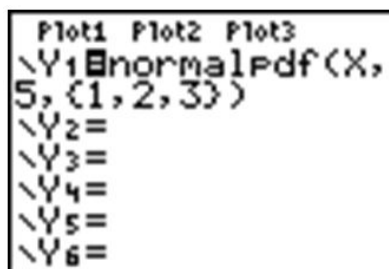
This makes it easy to see the change in spread when the standard deviation changes.

The Empirical Rule

Because of the similar shape of all normal distributions, we can measure the percentage of data that is a certain distance from the mean no matter what the standard deviation of the data set is. The following graph shows a normal distribution 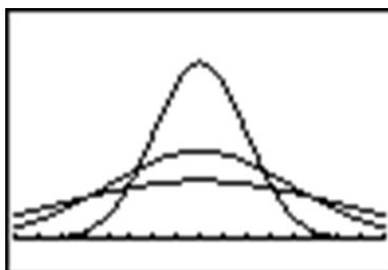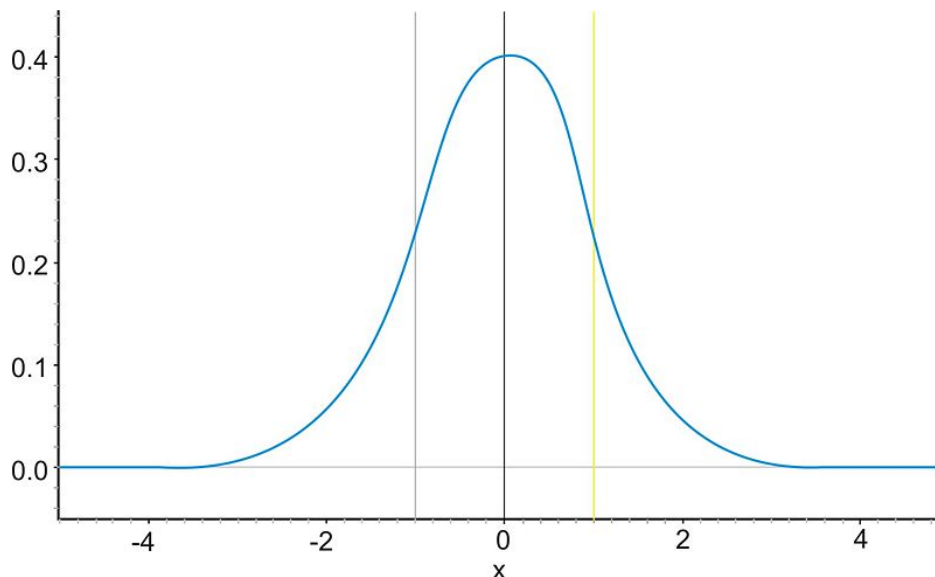with $\mu = 0$ and $\sigma = 1$. This curve is called a *standard normal curve*. In this case, the values of x represent the number of standard deviations away from the mean.

The Standard Normal Distribution

Notice that vertical lines are drawn at points that are exactly one standard deviation to the left and right of the mean. We have consistently described standard deviation as a measure of the typical distance away from the mean. How much of the data is actually within one standard deviation of the mean? To answer this question, think about the space, or area, under the curve. The entire data set, or 100% of it, is contained under the whole curve. What percentage would you estimate is between the two lines? To help estimate the answer, we can use a graphing calculator. Graph a standard normal distribution over an appropriate window.
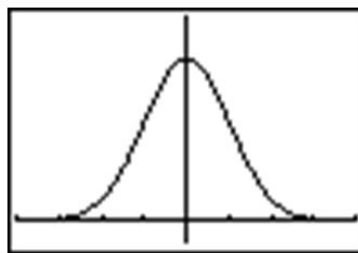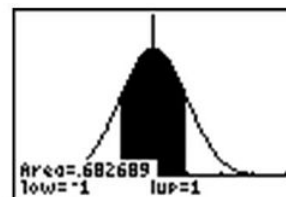

Now press **[2ND][DISTR]**, go to the **DRAW** menu, and choose 'ShadeNorm('. Insert '-1, 1' after the 'Shade-Norm(' command and press **[ENTER]**. It will shade the area within one standard deviation of the mean.

The calculator also gives a very accurate estimate of the area. We can see from the rightmost screenshot above that approximately 68% of the area is within one standard deviation of the mean. If we venture to 2 standard deviations away from the mean, how much of the data should we expect to capture? Make the following changes to the 'ShadeNorm(' command to find out:
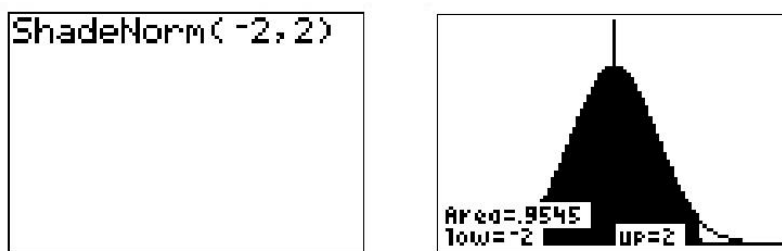
Notice from the shading that almost all of the distribution is shaded, and the percentage of data is close to 95%. If you were to venture to 3 standard deviations from the mean, 99.7%, or virtually all of the data, is captured, which tells us that very little of the data in a normal distribution is more than 3 standard deviations from the mean.
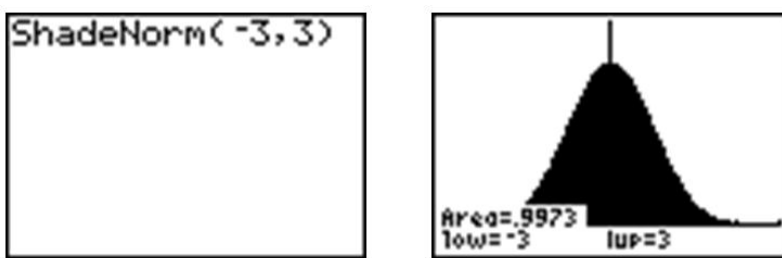
Notice that the calculator actually makes it look like the entire distribution is shaded because of the limitations of the screen resolution, but as we have already discovered, there is still some area under the curve further out than that. These three approximate percentages, 68%, 95%, and 99.7%, are extremely important and are part of what is called the *Empirical Rule*.

The Empirical Rule states that the percentages of data in a normal distribution within 1, 2, and 3 standard deviations of the mean are approximately 68%, 95%, and 99.7%, respectively.

On the Web

http://tinyurl.com/2ue78u Explore the Empirical Rule.
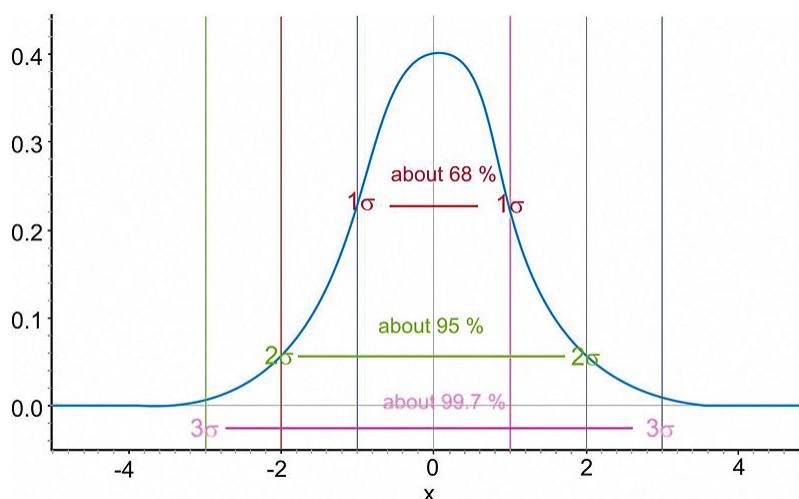
The Empirical Rule

z-Scores

A *z-score* is a measure of the number of standard deviations a particular data point is away from the mean. For example, let's say the mean score on a test for your statistics class was an 82, with a standard deviation of 7 points. If your score was an 89, it is exactly one standard deviation to the right of the mean; therefore, your *z*-score would be 1. If, on the other hand, you scored a 75, your score would be exactly one standard deviation below the mean, and your *z*-score would be -1. All values that are below the mean have negative *z*-scores, while all values that are above the mean have positive *z*-scores. A *z*-score of -2 would represent a value that is exactly 2 standard deviations below the mean, so in this case, the value would be $82 - 14 = 68$.

To calculate a *z*-score for which the numbers are not so obvious, you take the deviation and divide it by the standard deviation.

$$z = \frac{\text{Deviation}}{\text{Standard Deviation}}$$

You may recall that deviation is the mean value of the variable subtracted from the observed value, so in symbolic terms, the *z*-score would be:

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

As previously stated, since σ is always positive, z will be positive when x is greater than μ and negative when x is less than μ. A *z*-score of zero means that the term has the same value as the mean. The value of z represents the number of standard deviations the given value of x is above or below the mean.

Example: What is the *z*-score for an *A* on the test described above, which has a mean score of 82? (Assume that an *A* is a 93.)

The *z*-score can be calculated as follows:

$$z = \frac{x - \mu}{\sigma}$$
$$z = \frac{93 - 82}{7}$$
$$z = \frac{11}{7} \approx 1.57$$

If we know that the test scores from the last example are distributed normally, then a *z*-score can tell us something about how our test score relates to the rest of the class. From the Empirical Rule, we know that about 68% of the students would have scored between a *z*-score of -1 and 1, or between a 75 and an 89, on the test. If 68% of the data is between these two values, then that leaves the remaining 32% in the tail areas. Because of symmetry, half of this, or 16%, would be in each individual tail.

Example: On a nationwide math test, the mean was 65 and the standard deviation was 10. If Robert scored 81, what was his *z*-score?

$$z = \frac{x - \mu}{\sigma}$$
$$z = \frac{81 - 65}{10}$$
$$z = \frac{16}{10}$$
$$z = 1.6$$

Example: On a college entrance exam, the mean was 70, and the standard deviation was 8. If Helen's *z*-score was -1.5, what was her exam score?

$$z = \frac{x - \mu}{\sigma}$$
$$\therefore z \cdot \sigma = x - \mu$$
$$x = \mu + z \cdot \sigma$$
$$x = 70 + (-1.5)(8)$$
$$x = 58$$

Assessing Normality

The best way to determine if a data set approximates a normal distribution is to look at a visual representation. Histograms and box plots can be useful indicators of normality, but they are not always definitive. It is often easier to tell if a data set is *not* normal from these plots.
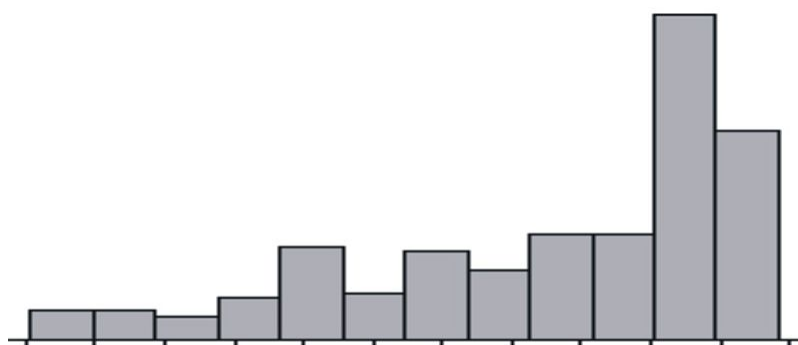
Skewed left distribution

Skewed right distribution with outliers
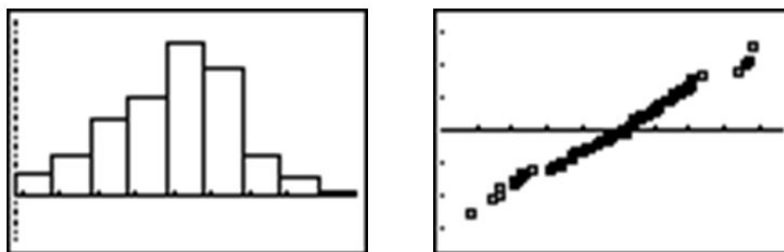
Bimodal Distribution
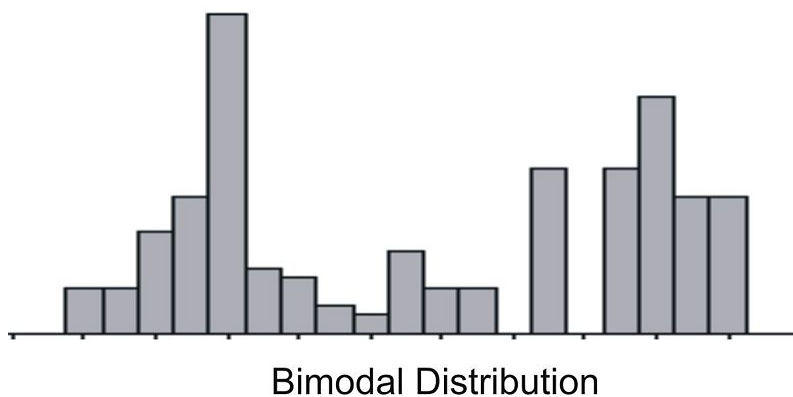
If a data set is skewed right, it means that the right tail is significantly longer than the left. Similarly, skewed left means the left tail has more weight than the right. A bimodal distribution, on the other hand, has two modes, or peaks. For instance, with a histogram of the heights of American 30-year-old adults, you will see a bimodal distribution−one mode for males and one mode for females.

There is a plot we can use to determine if a distribution is normal called a *normal probability plot* or *normal quantile plot*. To make this plot by hand, first order your data from smallest to largest. Then, determine the quantile of each data point. Finally, using a table of standard normal probabilities, determine the closest z-score for each quantile. Plot these z-scores against the actual data values. To make a normal probability plot using your calculator, enter your data into a list, then use the last type of graph in the **STAT PLOT** menu, as shown below:

If the data set is normal, then this plot will be perfectly linear. The closer to being linear the normal probability plot is, the more closely the data set approximates a normal distribution.

Look below at the histogram and the normal probability plot for the same data.

The histogram is fairly symmetric and mound-shaped and appears to display the characteristics of a normal distribution. When the *z*-scores of the quantiles of the data are plotted against the actual data values, the normal probability plot appears strongly linear, indicating that the data set closely approximates a normal distribution. The following example will allow you to see how a normal probability plot is made in more detail.

Example: The following data set tracked high school seniors' involvement in traffic accidents. The participants were asked the following question: "During the last 12 months, how many accidents have you had while you were driving (whether or not you were responsible)?"

TABLE 1.1:

Year	Percentage of high school seniors who said they were involved in no traffic accidents
1991	75.7
1992	76.9
1993	76.1
1994	75.7
1995	75.3
1996	74.1
1997	74.4
1998	74.4
1999	75.1
2000	75.1
2001	75.5
2002	75.5
2003	75.8

Figure: Percentage of high school seniors who said they were involved in no traffic accidents. *Source*: Sourcebook of Criminal Justice Statistics: http://www.albany.edu/sourcebook/pdf/t352.pdf

Here is a histogram and a box plot of this data:
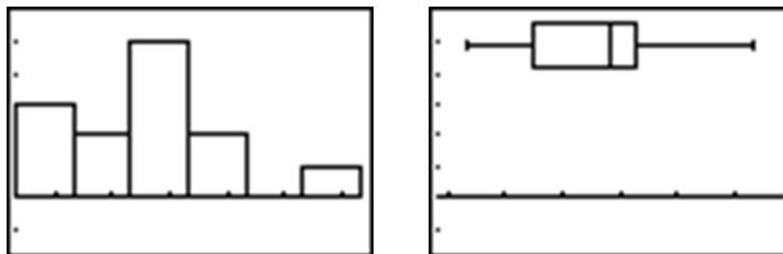
The histogram appears to show a roughly mound-shaped and symmetric distribution. The box plot does not appear to be significantly skewed, but the various sections of the plot also do not appear to be overly symmetric, either. In the following chart, the data has been reordered from smallest to largest, the quantiles have been determined, and the closest corresponding z-scores have been found using a table of standard normal probabilities.

TABLE 1.2:

Year	Percentage	Quantile	z-score
1996	74.1	$\frac{1}{13} = 0.078$	-1.42
1997	74.4	$\frac{2}{13} = 0.154$	-1.02
1998	74.4	$\frac{3}{13} = 0.231$	-0.74
1999	75.1	$\frac{4}{13} = 0.286$	-0.56
2000	75.1	$\frac{5}{13} = 0.385$	-0.29
1995	75.3	$\frac{6}{13} = 0.462$	-0.09
2001	75.5	$\frac{7}{13} = 0.538$	0.1
2002	75.5	$\frac{8}{13} = 0.615$	0.29
1991	75.7	$\frac{9}{13} = 0.692$	0.50
1994	75.7	$\frac{10}{13} = 0.769$	0.74
2003	75.8	$\frac{11}{13} = 0.846$	1.02
1993	76.1	$\frac{12}{13} = 0.923$	1.43
1992	76.9	$\frac{13}{13} = 1$	3.49

Figure: Table of quantiles and corresponding *z*-scores for senior no-accident data.

Here is a plot of the percentages versus the *z*-scores of their quantiles, or the normal probability plot:
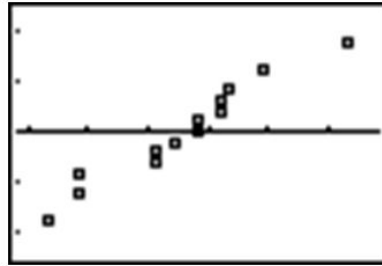
Remember that you can simplify this process by simply entering the percentages into a L_1 in your calculator and selecting the normal probability plot option (the last type of plot) in **STAT PLOT**.

While not perfectly linear, this plot does have a strong linear pattern, and we would, therefore, conclude that the distribution is reasonably normal.

Lesson Summary

A normal distribution is a perfectly symmetric, mound-shaped distribution that appears in many practical and real data sets. It is an especially important foundation for making conclusions, or inferences, about data. A standard normal distribution is a normal distribution for which the mean is 0 and the standard deviation is 1.

A *z*-score is a measure of the number of standard deviations a particular data value is away from the mean. The formula for calculating a *z*-score is:

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

z-scores are useful for comparing two distributions with different centers and/or spreads. When you convert an entire distribution to *z*-scores, you are actually changing it to a standardized distribution. *z*-scores can be calculated for data, even if the underlying population does not follow a normal distribution.

The Empirical Rule is the name given to the observation that approximately 68% of a normally distributed data set is within 1 standard deviation of the mean, about 95% is within 2 standard deviations of the mean, and about 99.7% is within 3 standard deviations of the mean. Some refer to this as the 68-95-99.7 Rule.

You should learn to recognize the normality of a distribution by examining the shape and symmetry of its visual display. A **normal probability plot**, or **normal quantile plot**, is a useful tool to help check the normality of a distribution. This graph is a plot of the *z*-scores of the data as quantiles against the actual data values. If a distribution is normal, this plot will be linear.

Points to Consider

- How can we use normal distributions to make meaningful conclusions about samples and experiments?
- How do we calculate probabilities and areas under the normal curve that are not covered by the Empirical Rule?
- What are the other types of distributions that can occur in different probability situations?

Multimedia Links

For an explanation of a standardized normal distribution **(4.0)(7.0)**, see APUS07, StandardNormal Distribution (4:22).
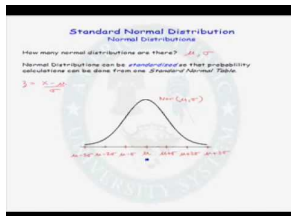
MEDIA
Click image to the left for more content.

Review Questions

Sample explanations for some of the practice exercises below are available by viewing the following videos. Khan Academy: NormalDistribution Problems (10:52)
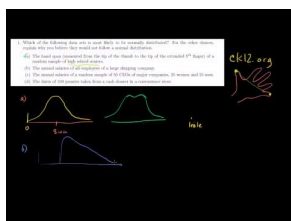
MEDIA
Click image to the left for more content.

KhanAcademy: NormalDistribution Problems-zscore (7:48)
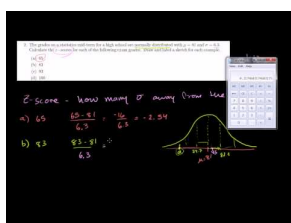
MEDIA
Click image to the left for more content.

KhanAcademy: NormalDistribution Problems (Empirical Rule) (10:25)
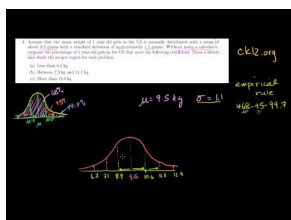
MEDIA
Click image to the left for more content.

KhanAcademy: Standard Normal Distribution and the EmpiricalRule (8:15)

> **MEDIA**
> Click image to the left for more content.

KhanAcademy: More EmpiricalRule and Z-score practice (5:57)

> **MEDIA**
> Click image to the left for more content.

1. Which of the following data sets is most likely to be normally distributed? For the other choices, explain why you believe they would not follow a normal distribution.

    a. The hand span (measured from the tip of the thumb to the tip of the extended 5th finger) of a random sample of high school seniors
    b. The annual salaries of all employees of a large shipping company
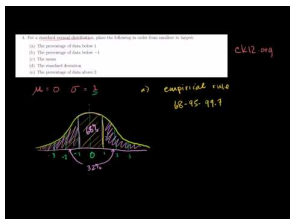   c. The annual salaries of a random sample of 50 CEOs of major companies, 25 women and 25 men
    d. The dates of 100 pennies taken from a cash drawer in a convenience store

2. The grades on a statistics mid-term for a high school are normally distributed, with $\mu = 81$ and $\sigma = 6.3$. Calculate the z-scores for each of the following exam grades. Draw and label a sketch for each example. 65, 83, 93, 100

3. Assume that the mean weight of 1-year-old girls in the USA is normally distributed, with a mean of about 9.5 kilograms and a standard deviation of approximately 1.1 kilograms. Without using a calculator, estimate the percentage of 1-year-old girls who meet the following conditions. Draw a sketch and shade the proper region for each problem.

    a. Less than 8.4 kg
    b. Between 7.3 kg and 11.7 kg
    c. More than 12.8 kg

4. For a standard normal distribution, place the following in order from smallest to largest.

    a. The percentage of data below 1
    b. The percentage of data below -1
    c. The mean
    d. The standard deviation
    e. The percentage of data above 2

5. The 2007 AP Statistics examination scores were not normally distributed, with $\mu = 2.8$ and $\sigma = 1.34$. What is the approximate z-score that corresponds to an exam score of 5? (The scores range from 1 to 5.)

    a. 0.786
    b. 1.46
    c. 1.64
    d. 2.20
    e. A z-score cannot be calculated because the distribution is not normal.

[1]Data available on the College Board Website: http://professionals.collegeboard.com/data-reports-research/ap/archived/2007

6. The heights of 5th grade boys in the USA is approximately normally distributed, with a mean height of 143.5 cm and a standard deviation of about 7.1 cm. What is the probability that a randomly chosen 5th grade boy would be taller than 157.7 cm?

7. A statistics class bought some sprinkle (or jimmies) doughnuts for a treat and noticed that the number of sprinkles seemed to vary from doughnut to doughnut, so they counted the sprinkles on each doughnut. Here are the results: 241, 282, 258, 223, 133, 335, 322, 323, 354, 194, 332, 274, 233, 147, 213, 262, 227, and 366. (a) Create a histogram, dot plot, or box plot for this data. Comment on the shape, center and spread of the distribution. (b) Find the mean and standard deviation of the distribution of sprinkles. Complete the following chart by standardizing all the values:

$$\mu = \quad \sigma =$$

TABLE 1.3:

Number of Sprinkles	Quantile	z-score
241		
282		
258		
223		
133		
335		
322		
323		
354		
194		
332		
274		
233		
147		
213		
262		
227		
366		

Figure: A table to be filled in for the sprinkles question.

(c) Create a normal probability plot from your results.

(d) Based on this plot, comment on the normality of the distribution of sprinkle counts on these doughnuts.

References

[1]http://www.albany.edu/sourcebook/pdf/t352.pdf

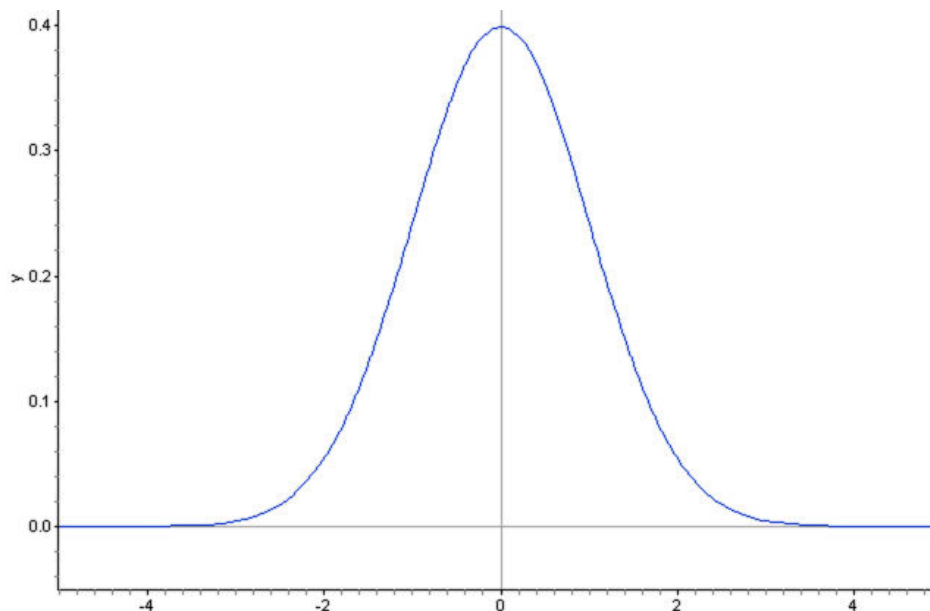## 1.2 The Density Curve of the Normal Distribution

Learning Objectives

- Identify the properties of a normal density curve and the relationship between concavity and standard deviation.
- Convert between *z*-scores and areas under a normal probability curve.
- Calculate probabilities that correspond to left, right, and middle areas from a *z*-score table.
- Calculate probabilities that correspond to left, right, and middle areas using a graphing calculator.

Introduction

In this section, we will continue our investigation of normal distributions to include density curves and learn various methods for calculating probabilities from the normal density curve.

Density Curves

A *density curve* is an idealized representation of a distribution in which the area under the curve is defined to be 1. Density curves need not be normal, but the normal density curve will be the most useful to us.

Inflection Points on a Normal Density Curve

We already know from the Empirical Rule that approximately $\frac{2}{3}$ of the data in a normal distribution lies within 1 standard deviation of the mean. With a normal density curve, this means that about 68% of the total area under the

curve is within *z*-scores of ±1. Look at the following three density curves:
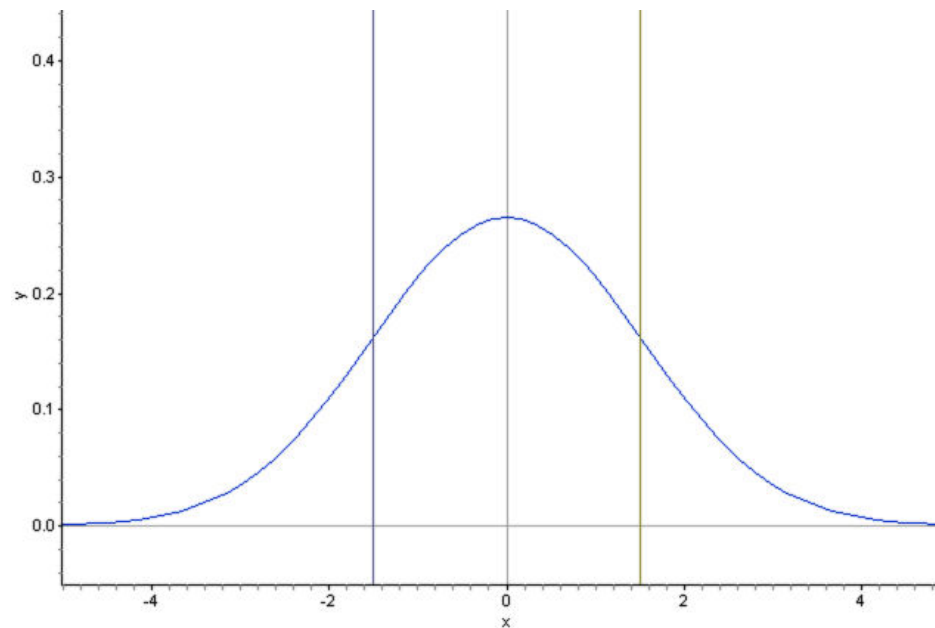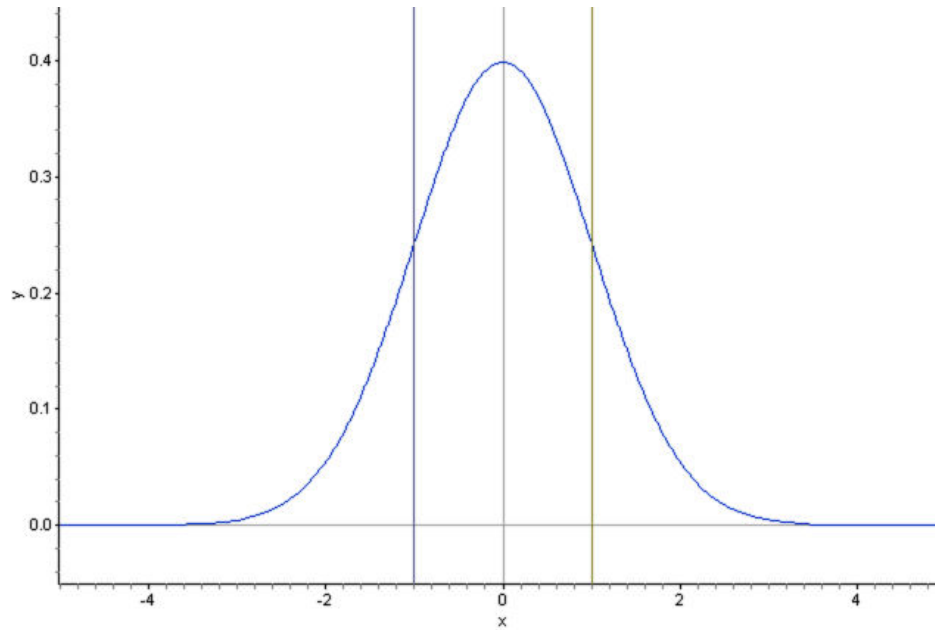
Notice that the curves are spread increasingly wider. Lines have been drawn to show the points that are one standard deviation on either side of the mean. Look at where this happens on each density curve. Here is a normal distribution with an even larger standard deviation.

Is it possible to predict the standard deviation of this distribution by estimating the *x*-coordinate of a point on the density curve? Read on to find out!
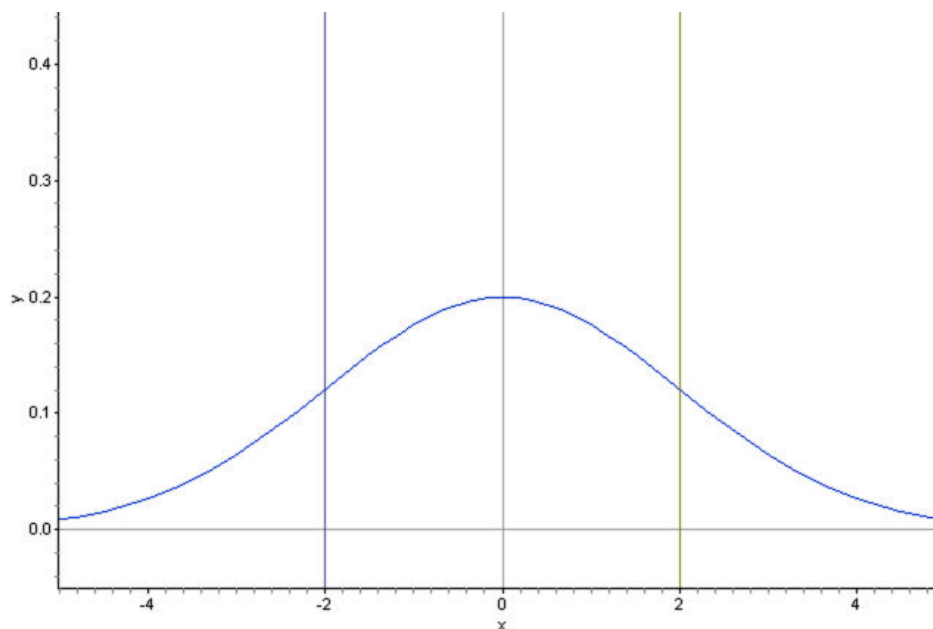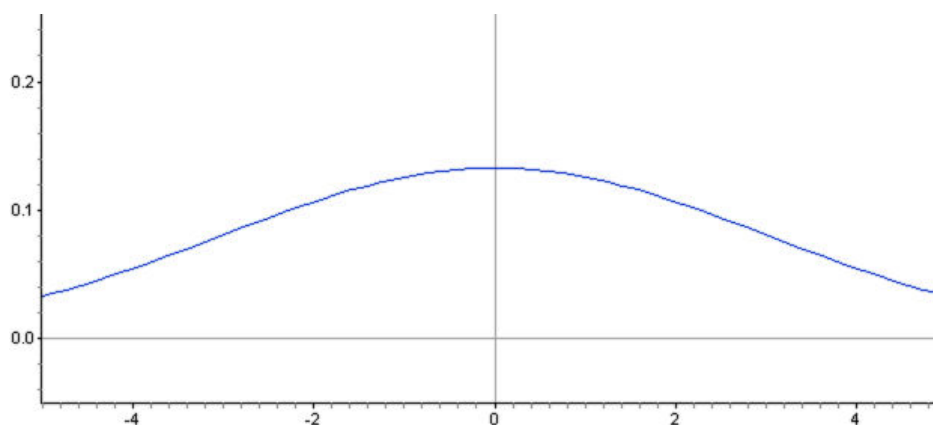
You may have noticed that the density curve changes shape at two points in each of our examples. These are the points where the curve changes concavity. Starting from the mean and heading outward to the left and right, the curve is *concave down*. (It looks like a mountain, or 'n' shape.) After passing these points, the curve is *concave up*. (It looks like a valley, or 'u' shape.) The points at which the curve changes from being concave up to being concave down are called the *inflection points*. On a normal density curve, these inflection points are always exactly one standard deviation away from the mean.
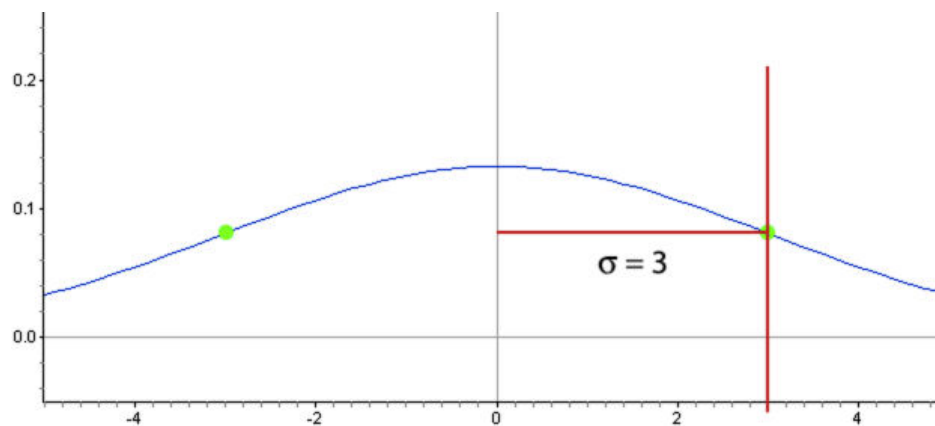
In this example, the standard deviation is 3 units. We can use this concept to estimate the standard deviation of a normally distributed data set.

Example: Estimate the standard deviation of the distribution represented by the following histogram.

Approximately normal data set
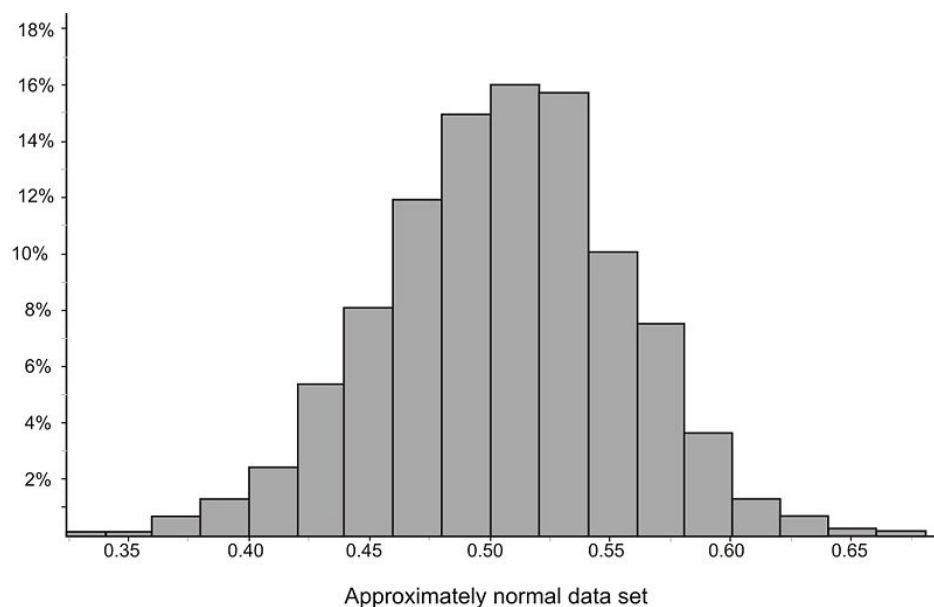
This distribution is fairly normal, so we could draw a density curve to approximate it as follows:

Density Curve approximating histogram
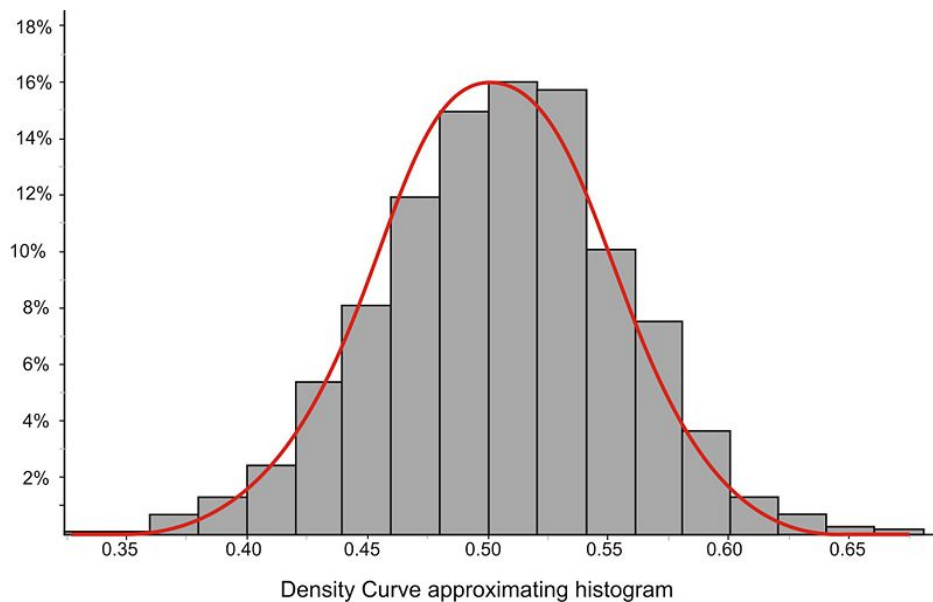
Now estimate the inflection points as shown below:

Density Curve with Inflection Points
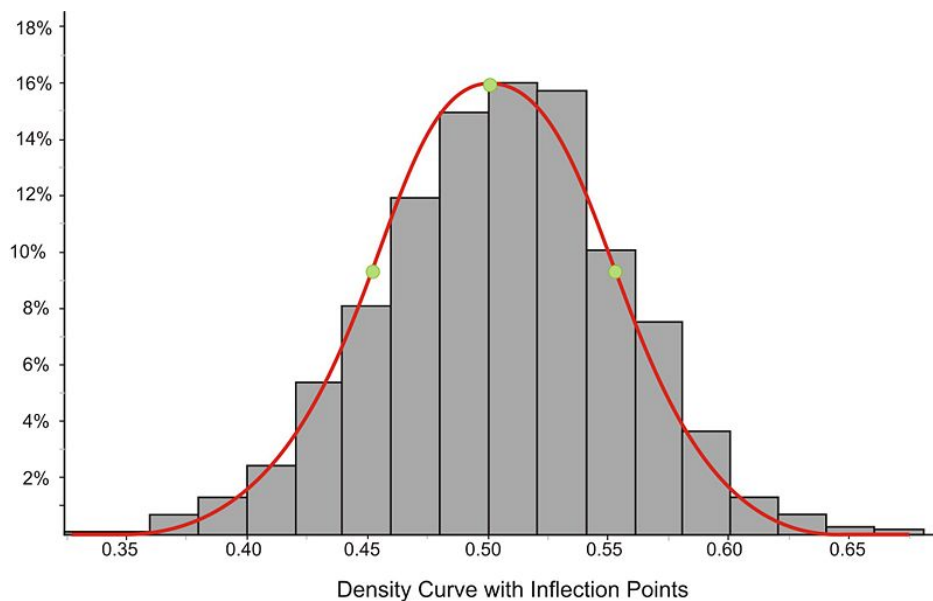
It appears that the mean is about 0.5 and that the *x*-coordinates of the inflection points are about 0.45 and 0.55, respectively. This would lead to an estimate of about 0.05 for the standard deviation.

The actual statistics for this distribution are as follows:

$$s \approx 0.04988$$
$$\bar{x} \approx 0.4997$$

We can verify these figures by using the expectations from the Empirical Rule. In the following graph, we have highlighted the bins that are contained within one standard deviation of the mean.

Area within 1 σ of the mean
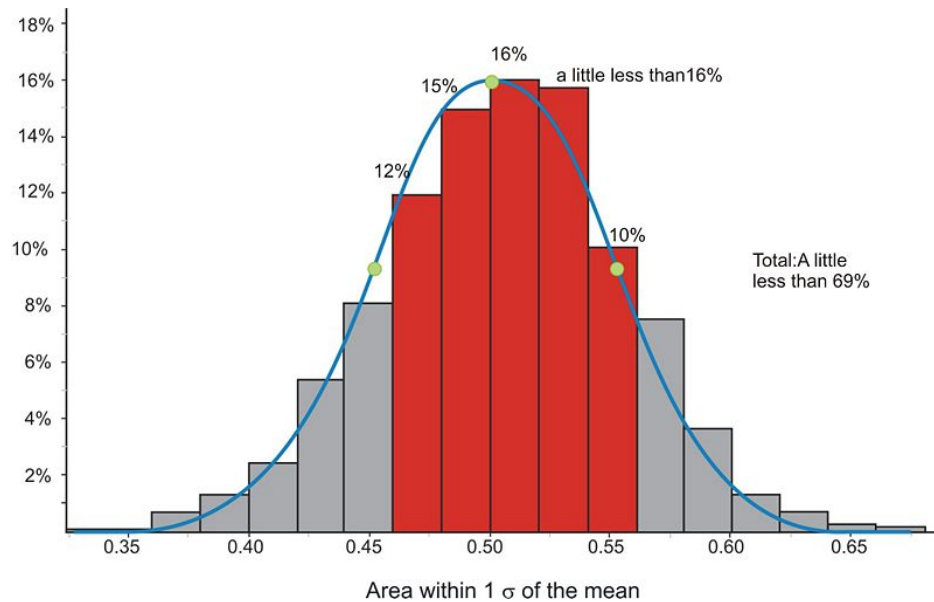
If you estimate the relative frequencies from each bin, their total is remarkably close to 68%. Make sure to divide the relative frequencies from the bins on the ends by 2 when performing your calculation.

Calculating Density Curve Areas

While it is convenient to estimate areas under a normal curve using the Empirical Rule, we often need more precise methods to calculate these areas. Luckily, we can use formulas or technology to help us with the calculations.

z-Tables

All normal distributions have the same basic shape, and therefore, rescaling and re-centering can be implemented to change any normal distributions to one with a mean of 0 and a standard deviation of 1. This configuration is referred to as a *standard normal distribution*. In a standard normal distribution, the variable along the horizontal axis is the *z*-score. This score is another measure of the performance of an individual score in a population. To review, the *z*-score measures how many standard deviations a score is away from the mean. The *z*-score of the term *x* in a population distribution whose mean is μ and whose standard deviation is σ is given by: $z = \frac{x-\mu}{\sigma}$. Since σ is always positive, *z* will be positive when *x* is greater than μ and negative when *x* is less than μ. A *z*-score of 0 means that the term has the same value as the mean. The value of *z* is the number of standard deviations the given value of *x* is above or below the mean.

Example: On a nationwide math test, the mean was 65 and the standard deviation was 10. If Robert scored 81, what was his *z*-score?

$$z = \frac{x-\mu}{\sigma}$$
$$z = \frac{81-65}{10}$$
$$z = \frac{16}{10}$$
$$z = 1.6$$

Example: On a college entrance exam, the mean was 70 and the standard deviation was 8. If Helen's *z*-score was -1.5, what was her exam score?

$$z = \frac{x - \mu}{\sigma}$$
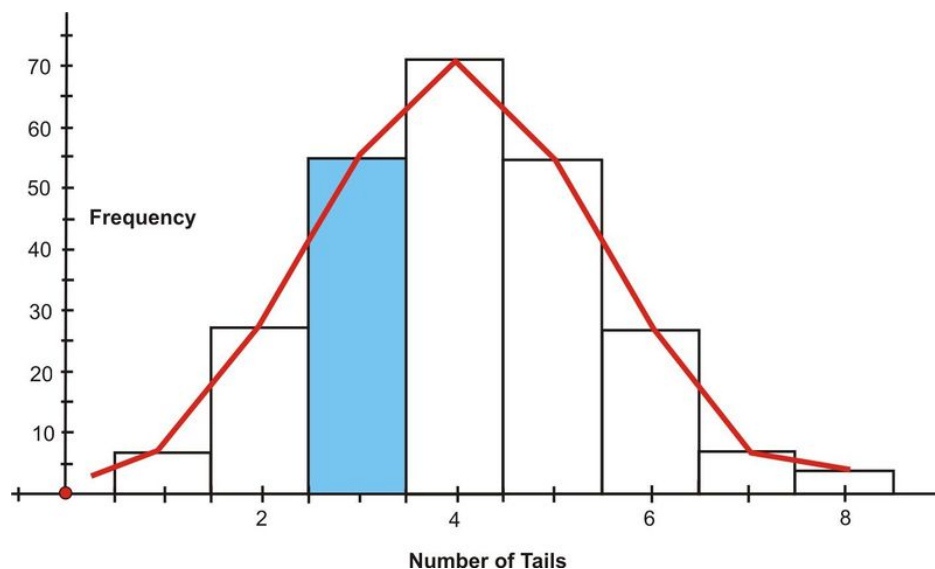$$\therefore z \cdot \sigma = x - \mu$$
$$x = \mu + z \cdot \sigma$$
$$x = (70) + (-1.5)(8)$$
$$x = 58$$

Now you will see how *z*-scores are used to determine the probability of an event.

Suppose you were to toss 8 coins 256 times. The following figure shows the histogram and the approximating normal curve for the experiment. The random variable represents the number of tails obtained.

The blue section of the graph represents the probability that exactly 3 of the coins turned up tails. One way to determine this is by the following:

$$P(3 \text{ tails}) = \frac{_8C_3}{2^8}$$
$$P(3 \text{ tails}) = \frac{56}{256}$$
$$P(3 \text{ tails}) \cong 0.2188$$

Geometrically, this probability represents the area of the blue shaded bar divided by the total area of the bars. The area of the blue shaded bar is approximately equal to the area under the normal curve from 2.5 to 3.5.

Since areas under normal curves correspond to the probability of an event occurring, a special normal distribution table is used to calculate the probabilities. This table can be found in any statistics book, but it is seldom used today. The following is an example of a table of *z*-scores and a brief explanation of how it works: http://tinyurl.com/2ce9o gv.

The values inside the given table represent the areas under the standard normal curve for values between 0 and the relative *z*-score. For example, to determine the area under the curve between *z*-scores of 0 and 2.36, look in the intersecting cell for the row labeled 2.3 and the column labeled 0.06. The area under the curve is 0.4909. To determine the area between 0 and a negative value, look in the intersecting cell of the row and column which sums

to the absolute value of the number in question. For example, the area under the curve between -1.3 and 0 is equal to the area under the curve between 1.3 and 0, so look at the cell that is the intersection of the 1.3 row and the 0.00 column. (The area is 0.4032.)

It is extremely important, especially when you first start with these calculations, that you get in the habit of relating it to the normal distribution by drawing a sketch of the situation. In this case, simply draw a sketch of a standard normal curve with the appropriate region shaded and labeled.

Sketch for z ≤ -1.58
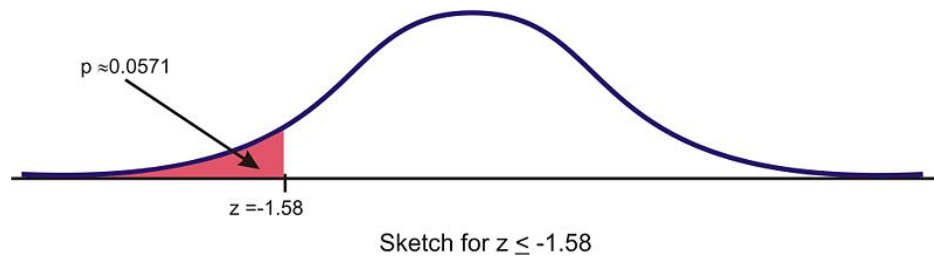
Example: Find the probability of choosing a value that is greater than $z = -0.528$. Before even using the table, first draw a sketch and estimate the probability. This z-score is just below the mean, so the answer should be more than 0.5.

Estimated Sketch for z ≥ -0.528
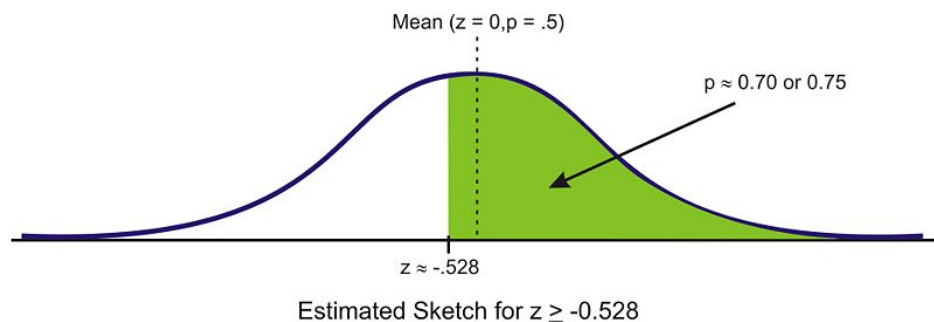
Next, read the table to find the correct probability for the data below this z-score. We must first round this z-score to -0.53, so this will slightly under-estimate the probability, but it is the best we can do using the table. The table returns a value of $0.5 - 0.2019 = 0.2981$ as the area below this z-score. Because the area under the density curve is equal to 1, we can subtract this value from 1 to find the correct probability of about 0.7019.

Final Sketch for z ≥ -0.528
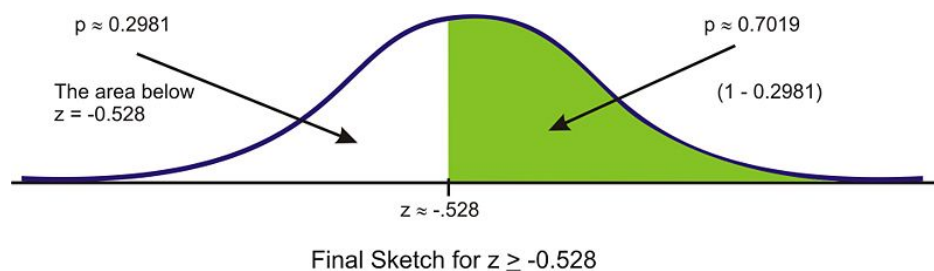
What about values between two z-scores? While it is an interesting and worthwhile exercise to do this using a table, it is so much simpler using software or a graphing calculator.

Example: Find $P(-2.60 < z < 1.30)$

This probability can be calculated as follows:

$$P(-2.60 < z < 1.30) = P(z < 1.30) - P(z < -2.60) = 0.9032 - 0.0047 = 0.8985$$

It can also be found using the TI-83/84 calculator. Use the 'normalcdf(-2.60, 1.30, 0, 1)' command, and the calculator will return the result 0.898538. The syntax for this command is 'normalcdf(min, max, μ, σ)'. When using this command, you do not need to first standardize. You can use the mean and standard deviation of the given distribution.

Technology Note: The 'normalcdf(' Command on the TI-83/84 Calculator

Your graphing calculator has already been programmed to calculate probabilities for a normal density curve using what is called a *cumulative density function*. The command you will use is found in the **DISTR** menu, which you can bring up by pressing **[2ND][DISTR]**.

```
DISTR DRAW
1:normalpdf(
2:normalcdf(
3:invNorm(
4:invT(
5:tpdf(
6:tcdf(
7↓X²pdf(
```

Press **[2]** to select the 'normalcdf(' command, which has a syntax of 'normalcdf(lower bound, upper bound, mean, standard deviation)'.

The command has been programmed so that if you do not specify a mean and standard deviation, it will default to the standard normal curve, with $\mu = 0$ and $\sigma = 1$.

For example, entering 'normalcdf(-1, 1)' will specify the area within one standard deviation of the mean, which we already know to be approximately 0.68.

```
normalcdf(-1,1)
          .6826894809
■
```

Try verifying the other values from the Empirical Rule.

Summary:

'Normalcdf (a, b, μ, σ)' gives values of the cumulative normal density function. In other words, it gives the probability of an event occurring between $x = a$ and $x = b$, or the area under the probability density curve between the vertical lines $x = a$ and $x = b$, where the normal distribution has a mean of μ and a standard deviation of σ. If μ and σ are not specified, it is assumed that $\mu = 0$ and $\sigma = 1$.

```
normalcdf(-2,2)
          .954499876
normalcdf(-3,3)
          .9973000656
```
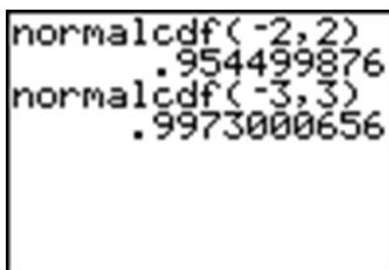
Example: Find the probability that $x < -1.58$.

The calculator command must have both an upper and lower bound. Technically, though, the density curve does not have a lower bound, as it continues infinitely in both directions. We do know, however, that a very small percentage of the data is below 3 standard deviations to the left of the mean. Use -3 as the lower bound and see what answer you get.

```
normalcdf( -3, -1.
58)
        .0557034698
```

The answer is fairly accurate, but you must remember that there is really still some area under the probability density curve, even though it is just a little, that we are leaving out if we stop at -3. If you look at the z-table, you can see that we are, in fact, leaving out about $0.5 - 0.4987 = 0.0013$. Next, try going out to -4 and -5.
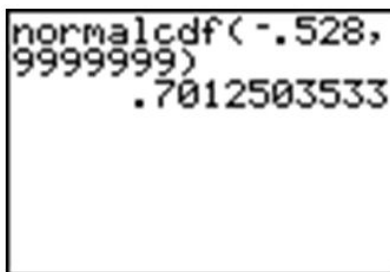
```
normalcdf( -4, -1.
58)
         .057021751
normalcdf( -5, -1.
58)
         .0570531499
```

Once we get to -5, the answer is quite accurate. Since we cannot really capture all the data, entering a sufficiently small value should be enough for any reasonable degree of accuracy. A quick and easy way to handle this is to enter -99999 (or "a bunch of nines"). It really doesn't matter exactly how many nines you enter. The difference between five and six nines will be beyond the accuracy that even your calculator can display.

```
normalcdf( -99999
, -1.58)
          .057053437
normalcdf( -99999
9, -1.58)
          .057053437
```

Example: Find the probability for $x \geq -0.528$.

Right away, we are at an advantage using the calculator, because we do not have to round off the z-score. Enter the 'normalcdf(' command, using -0.528 to "a bunch of nines." The nines represent a ridiculously large upper bound that will insure that the unaccounted-for probability will be so small that it will be virtually undetectable.

```
normalcdf(-.528,
9999999)
         .7012503533
```

Remember that because of rounding, our answer from the table was slightly too small, so when we subtracted it from 1, our final answer was slightly too large. The calculator answer of about 0.70125 is a more accurate approximation than the answer arrived at by using the table.

Standardizing

In most practical problems involving normal distributions, the curve will not be as we have seen so far, with $\mu = 0$ and $\sigma = 1$. When using a z-table, you will first have to *standardize* the distribution by calculating the z-score(s).

Example: A candy company sells small bags of candy and attempts to keep the number of pieces in each bag the same, though small differences due to random variation in the packaging process lead to different amounts in individual packages. A quality control expert from the company has determined that the mean number of pieces in each bag is normally distributed, with a mean of 57.3 and a standard deviation of 1.2. Endy opened a bag of candy and felt he was cheated. His bag contained only 55 candies. Does Endy have reason to complain?

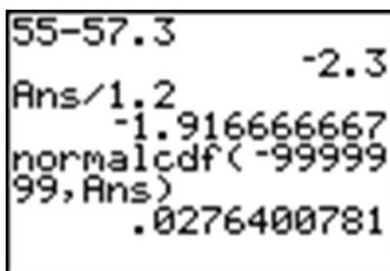To determine if Endy was cheated, first calculate the z-score for 55:

$$z = \frac{x - \mu}{\sigma}$$
$$z = \frac{55 - 57.3}{1.2}$$
$$z \approx -1.911666\ldots$$
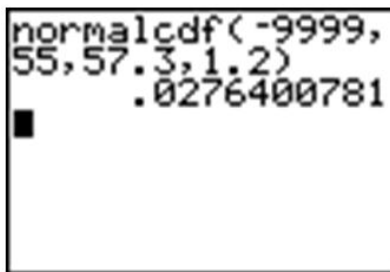
Using a table, the probability of experiencing a value this low is approximately $0.5 - 0.4719 = 0.0281$. In other words, there is about a 3% chance that you would get a bag of candy with 55 or fewer pieces, so Endy should feel cheated.

Using a graphing calculator, the results would look as follows (the 'Ans' function has been used to avoid rounding off the z-score):

```
55-57.3
                -2.3
Ans/1.2
        -1.916666667
normalcdf(-99999
99,Ans)
        .0276400781
```

However, one of the advantages of using a calculator is that it is unnecessary to standardize. We can simply enter the mean and standard deviation from the original population distribution of candy, avoiding the z-score calculation completely.

```
normalcdf(-9999,
55,57.3,1.2)
        .0276400781
█
```

Lesson Summary

A density curve is an idealized representation of a distribution in which the area under the curve is defined as 1, or in terms of percentages, a probability of 100%. A normal density curve is simply a density curve for a normal distribution. Normal density curves have two inflection points, which are the points on the curve where it changes concavity. These points correspond to the points in the normal distribution that are exactly 1 standard deviation away from the mean. Applying the Empirical Rule tells us that the area under the normal density curve between these two points is approximately 0.68. This is most commonly thought of in terms of probability (e.g., the probability of choosing a value at random from this distribution and having it be within 1 standard deviation of the mean is 0.68). Calculating other areas under the curve can be done by using a z-table or by using the 'normalcdf(' command on the TI-83/84 calculator. A z-table often provides the area under the standard normal density curve between the mean and a particular z-score. The calculator command allows you to specify two values, either standardized or not, and will calculate the area under the curve between these values.

Points to Consider

- How do we calculate areas/probabilities for distributions that are not normal?
- How do we calculate z-scores, means, standard deviations, or actual values given a probability or area?

On the Web

Tables

http://tinyurl.com/2ce9ogv This link leads you to a $z-$table and an explanation of how to use it.

http://tinyurl.com/2aau5zy Investigate the mean and standard deviation of a normal distribution.

http://tinyurl.com/299hsjo The Normal Calculator.

http://www.math.unb.ca/ knight/utility/NormTble.htm Another online normal probability table.
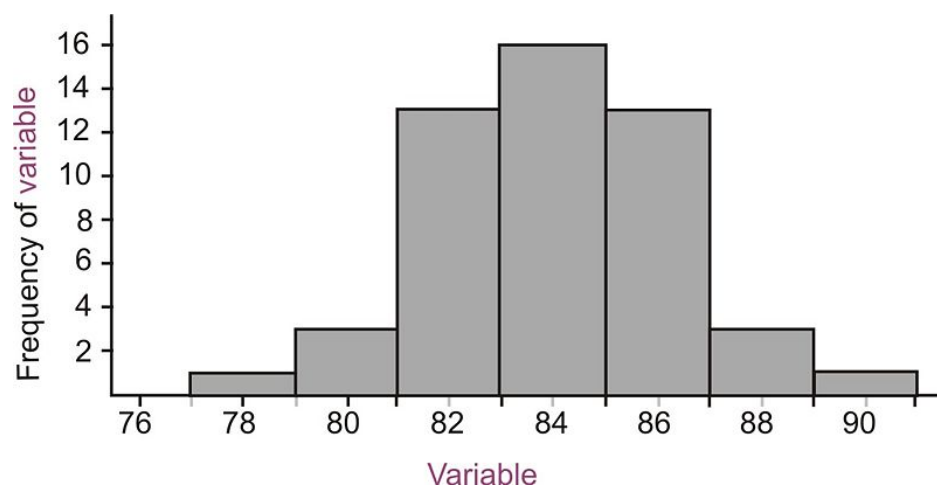

Multimedia Links

For an example showing how to compute probabilities with normal distribution **(8.0)**, see ExamSolutions, Normal Distribution: P(more than x) wherex is less than the mean (8:40).

ExamSolutions

Making Maths Easy

MEDIA

Click image to the left for more content.

Review Questions

1. Estimate the standard deviation of the following distribution.

2. Calculate the following probabilities using only the z-table. Show all your work.

   a. $P(z \geq -0.79)$
   b. $P(-1 \leq z \leq 1)$ Show all work.
   c. $P(-1.56 < z < 0.32)$

3. Brielle's statistics class took a quiz, and the results were normally distributed, with a mean of 85 and a standard deviation of 7. She wanted to calculate the percentage of the class that got a *B* (between 80 and 90). She used her calculator and was puzzled by the result. Here is a screen shot of her calculator:

```
normalcdf(80,90)
                0
```

Explain her mistake and the resulting answer on the calculator, and then calculate the correct answer.

4. Which grade is better: A 78 on a test whose mean is 72 and standard deviation is 6.5, or an 83 on a test whose mean is 77 and standard deviation is 8.4. Justify your answer and draw sketches of each distribution.
5. Teachers A and B have final exam scores that are approximately normally distributed, with the mean for Teacher A equal to 72 and the mean for Teacher B equal to 82. The standard deviation of Teacher A's scores is 10, and the standard deviation of Teacher B's scores is 5.

a. With which teacher is a score of 90 more impressive? Support your answer with appropriate probability calculations and with a sketch.
b. With which teacher is a score of 60 more discouraging? Again, support your answer with appropriate probability calculations and with a sketch.

1.3 Applications of the Normal Distribution

Learning Objective

- Apply the characteristics of a normal distribution to solving problems.

Introduction

The normal distribution is the foundation for statistical inference and will be an essential part of many of those topics in later chapters. In the meantime, this section will cover some of the types of questions that can be answered using the properties of a normal distribution. The first examples deal with more theoretical questions that will help you master basic understandings and computational skills, while the later problems will provide examples with real data, or at least a real context.

Unknown Value Problems

If you understand the relationship between the area under a density curve and mean, standard deviation, and z-scores, you should be able to solve problems in which you are provided all but one of these values and are asked to calculate the remaining value. In the last lesson, we found the probability that a variable is within a particular range, or the area under a density curve within that range. What if you are asked to find a value that gives a particular probability?

Example: Given the normally-distributed random variable X, with $\mu = 35$ and $\sigma = 7.4$, what is the value of X where the probability of experiencing a value less than it is 80%?

As suggested before, it is important and helpful to sketch the distribution.
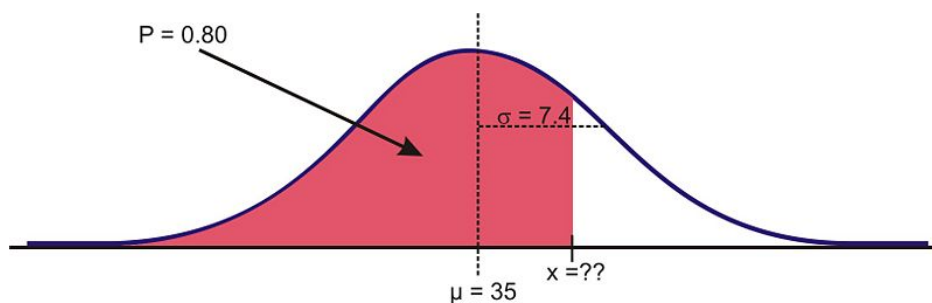
Sketch of distribution

If we had to estimate an actual value first, we know from the Empirical Rule that about 84% of the data is below one standard deviation to the right of the mean.

$$\mu + 1\sigma = 35 + 7.4 = 42.4$$

Therefore, we expect the answer to be slightly below this value.

Estimating using the empirical rule
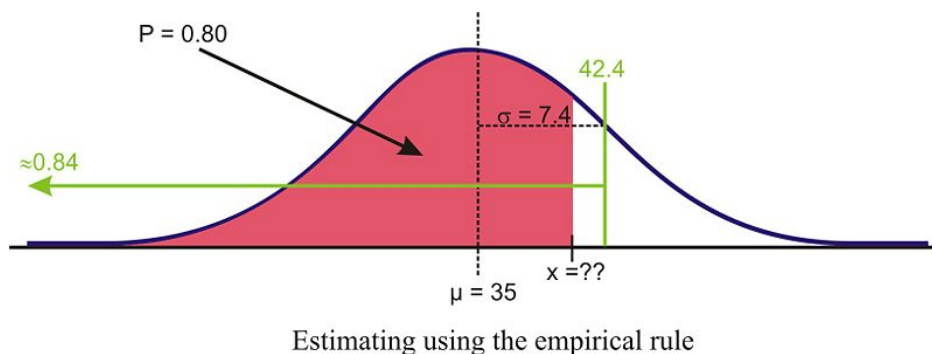
When we were given a value of the variable and were asked to find the percentage or probability, we used a *z*-table or the 'normalcdf(' command on a graphing calculator. But how do we find a value given the percentage? Again, the table has its limitations in this case, and graphing calculators and computer software are much more convenient and accurate. The command on the TI-83/84 calculator is 'invNorm('. You may have seen it already in the **DISTR** menu.

The syntax for this command is as follows:

'InvNorm(percentage or probability to the left, mean, standard deviation)'
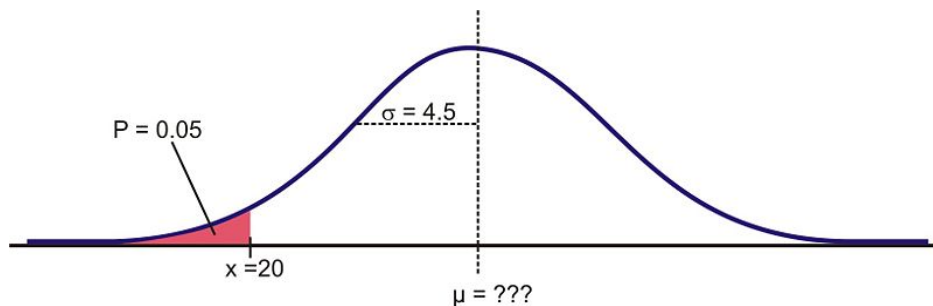
Make sure to enter the values in the correct order, such as in the example below:
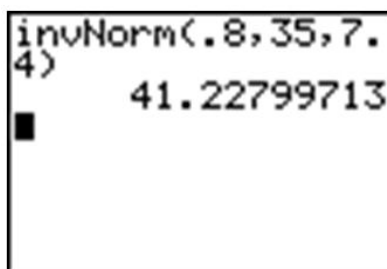

Unknown Mean or Standard Deviation

Example: For a normally distributed random variable, $\sigma = 4.5$, $x = 20$, and $p = 0.05$, Estimate μ.

To solve this problem, first draw a sketch:

Remember that about 95% of the data is within 2 standard deviations of the mean. This would leave 2.5% of the data in the lower tail, so our 5% value must be less than 9 units from the mean.

Because we do not know the mean, we have to use the standard normal curve and calculate a z-score using the 'invNorm(' command. The result, -1.645, confirms the prediction that the value is less than 2 standard deviations from the mean.
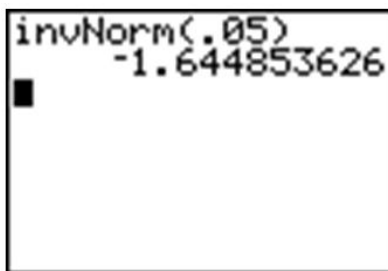
Now, plug in the known quantities into the z-score formula and solve for μ as follows:

$$z = \frac{x - \mu}{\sigma}$$
$$-1.645 \approx \frac{20 - \mu}{4.5}$$
$$(-1.645)(4.5) \approx 20 - \mu$$
$$-7.402 - 20 \approx -\mu$$
$$-27.402 \approx -\mu$$
$$\mu \approx 27.402$$

Example: For a normally-distributed random variable, $\mu = 83$, $x = 94$, and $p = 0.90$. Find σ.

Again, let's first look at a sketch of the distribution.
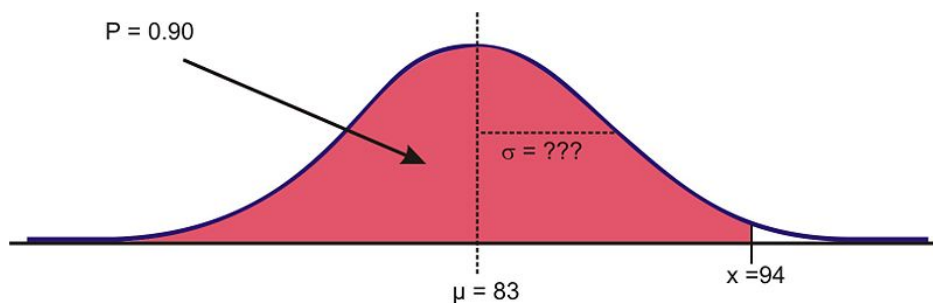
Sketch of Distribution

Since about 97.5% of the data is below 2 standard deviations, it seems reasonable to estimate that the *x* value is less than two standard deviations away from the mean and that σ might be around 7 or 8.

Again, the first step to see if our prediction is right is to use 'invNorm(' to calculate the *z*-score. Remember that since we are not entering a mean or standard deviation, the result is based on the assumption that $\mu = 0$ and $\sigma = 1$.

```
invNorm(.9)
           1.281551567
■
```
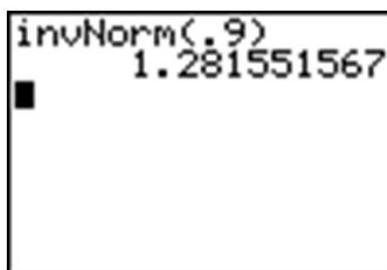
Now, use the *z*-score formula and solve for σ as follows:

$$z = \frac{x - \mu}{\sigma}$$
$$1.282 \approx \frac{94 - 83}{\sigma}$$
$$\sigma \approx \frac{11}{1.282}$$
$$\sigma \approx 8.583$$

Technology Note: Drawing a Distribution on the TI-83/84 Calculator

The TI-83/84 calculator will draw a distribution for you, but before doing so, we need to set an appropriate window (see screen below) and delete or turn off any functions or plots. Let's use the last example and draw the shaded region below 94 under a normal curve with $\mu = 83$ and $\sigma = 8.583$. Remember from the Empirical Rule that we probably want to show about 3 standard deviations away from 83 in either direction. If we use 9 as an estimate for σ, then we should open our window 27 units above and below 83. The *y* settings can be a bit tricky, but with a little practice, you will get used to determining the maximum percentage of area near the mean.

```
WINDOW
 Xmin=56
 Xmax=110
 Xscl=9
 Ymin=-.02
 Ymax=.06
 Yscl=.01
 Xres=1
```
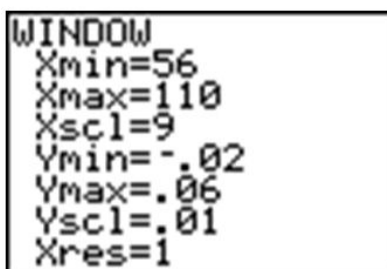
The reason that we went below the *x*-axis is to leave room for the text, as you will see.

Now, press **[2ND][DISTR]** and arrow over to the **DRAW** menu.

Choose the 'ShadeNorm(' command. With this command, you enter the values just as if you were doing a 'normalcdf(' calculation. The syntax for the 'ShadeNorm(' command is as follows:

'ShadeNorm(lower bound, upper bound, mean, standard deviation)'

Enter the values shown in the following screenshot:

```
ShadeNorm( -99999
,94,83,8.583)
```

Next, press **[ENTER]** to see the result. It should appear as follows:

```
Area=.900009
low= -99999   up=94
```
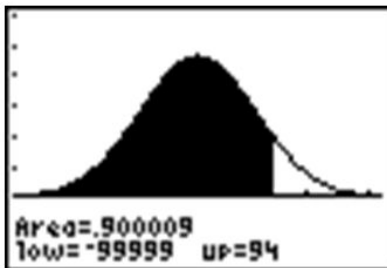
Technology Note: The 'normalpdf(' Command on the TI-83/84 Calculator

You may have noticed that the first option in the **DISTR** menu is 'normalpdf(', which stands for a normal probability density function. It is the option you used in lesson 5.1 to draw the graph of a normal distribution. Many students wonder what this function is for and occasionally even use it by mistake to calculate what they think are cumulative probabilities, but this function is actually the mathematical formula for drawing a normal distribution. You can find this formula in the resources at the end of the lesson if you are interested. The numbers this function returns are not really useful to us statistically. The primary purpose for this function is to draw the normal curve.

To do this, first be sure to turn off any plots and clear out any functions. Then press **[Y=]**, insert 'normalpdf(', enter 'X', and close the parentheses as shown. Because we did not specify a mean and standard deviation, the standard normal curve will be drawn. Finally, enter the following window settings, which are necessary to fit most of the curve on the screen (think about the Empirical Rule when deciding on settings), and press **[GRAPH]**. The normal curve below should appear on your screen.

```
Plot1 Plot2 Plot3        WINDOW
\Y1 normalpdf(X)          Xmin=-3
                          Xmax=3
\Y2=                      Xscl=1
\Y3=                      Ymin= -.01
\Y4=                      Ymax=.5
\Y5=                      Yscl=.1
\Y6=                      Xres=1
```

Normal Distributions with Real Data

The foundation of performing experiments by collecting surveys and samples is most often based on the normal distribution, as you will learn in greater detail in later chapters. Here are two examples to get you started.

Example: The Information Centre of the National Health Service in Britain collects and publishes a great deal of information and statistics on health issues affecting the population. One such comprehensive data set tracks information about the health of children[1]. According to its statistics, in 2006, the mean height of 12-year-old boys was 152.9 cm, with a standard deviation estimate of approximately 8.5 cm. (These are not the exact figures for the

population, and in later chapters, we will learn how they are calculated and how accurate they may be, but for now, we will assume that they are a reasonable estimate of the true parameters.)

If 12-year-old Cecil is 158 cm, approximately what percentage of all 12-year-old boys in Britain is he taller than?

We first must assume that the height of 12-year-old boys in Britain is normally distributed, and this seems like a reasonable assumption to make. As always, draw a sketch and estimate a reasonable answer prior to calculating the percentage. In this case, let's use the calculator to sketch the distribution and the shading. First decide on an appropriate window that includes about 3 standard deviations on either side of the mean. In this case, 3 standard deviations is about 25.5 cm, so add and subtract this value to/from the mean to find the horizontal extremes. Then enter the appropriate 'ShadeNorm(' command as shown:

```
WINDOW
 Xmin=127.4
 Xmax=178.4
 Xscl=5
 Ymin=-.025
 Ymax=.05
 Yscl=.05
 Xres=1
```

```
ShadeNorm(-9999,
158,152.9,8.5)
```

```
Area=.725747
low=-9999    up=158
```
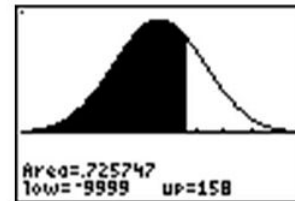
From this data, we would estimate that Cecil is taller than about 73% of 12-year-old boys. We could also phrase our assumption this way: the probability of a randomly selected British 12-year-old boy being shorter than Cecil is about 0.73. Often with data like this, we use percentiles. We would say that Cecil is in the 73rd percentile for height among 12-year-old boys in Britain.

How tall would Cecil need to be in order to be in the top 1% of all 12-year-old boys in Britain?

Here is a sketch:

P = 0.99

σ = 8.5cm

μ = 152.9cm

x =??
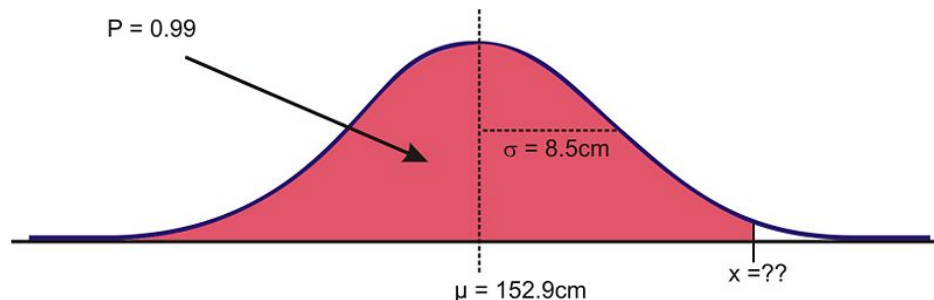
Sketch for Height Problem

In this case, we are given the percentage, so we need to use the 'invNorm(' command as shown.

```
invNorm(.99,152.
9,8.5)
        172.673957
```

Our results indicate that Cecil would need to be about 173 cm tall to be in the top 1% of 12-year-old boys in Britain.

Example: Suppose that the distribution of the masses of female marine iguanas in Puerto Villamil in the Galapagos Islands is approximately normal, with a mean mass of 950 g and a standard deviation of 325 g. There are very few

young marine iguanas in the populated areas of the islands, because feral cats tend to kill them. How rare is it that we would find a female marine iguana with a mass less than 400 g in this area?
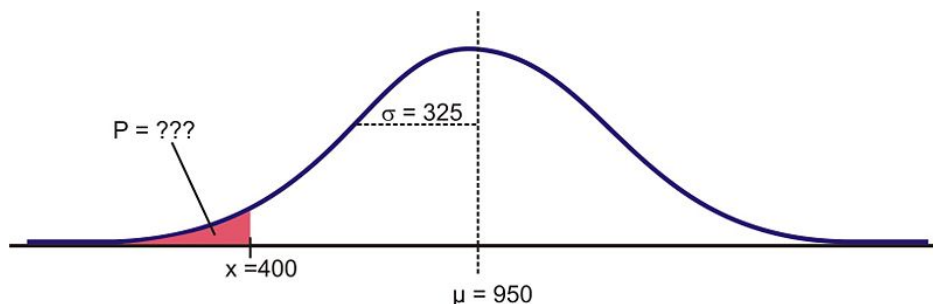
Using a graphing calculator, we can approximate the probability of a female marine iguana being less than 400 grams as follows:

With a probability of approximately 0.045, or only about 5%, we could say it is rather unlikely that we would find an iguana this small.

Lesson Summary

In order to find the percentage of data in-between two values (or the probability of a randomly chosen value being between those values) in a normal distribution, we can use the 'normalcdf(' command on the TI-83/84 calculator. When you know the percentage or probability, use the 'invNorm(' command to find a z-score or value of the variable. In order to use these tools in real situations, we need to know that the distribution of the variable in question is approximately normal. When solving problems using normal probabilities, it helps to draw a sketch of the distribution and shade the appropriate region.

Point to Consider

- How do the probabilities of a standard normal curve apply to making decisions about unknown parameters for a population given a sample?

Multimedia Links

For an example of finding the probability between values in a normal distribution **(4.0)(7.0)**, see EducatorVids, Statistics: Applications of the NormalDistribution (1:45).
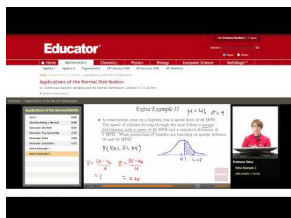
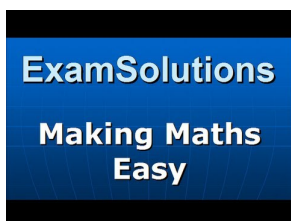|	MEDIA
	Click image to the left for more content.

For an example showing how to find the mean and standard deviation of a normal distribution **(8.0)**, see ExamSolutions, Normal Distribution: Finding the MeanandStandard Deviation (6:01).
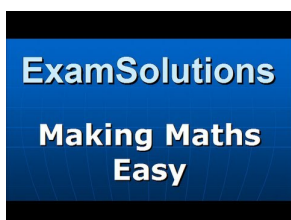
	MEDIA
	Click image to the left for more content.

For the continuation of finding the mean and standard deviation of a normal distribution **(8.0)**, see ExamSolutions, Normal Distribution: Finding the MeanandStandard Deviation (Part 2) (8:09).

	MEDIA
	Click image to the left for more content.

Review Questions

1. Which of the following intervals contains the middle 95% of the data in a standard normal distribution?

   a. $z < 2$
   b. $z \leq 1.645$
   c. $z \leq 1.96$
   d. $-1.645 \leq z \leq 1.645$
   e. $-1.96 \leq z \leq 1.96$

2. For each of the following problems, X is a continuous random variable with a normal distribution and the given mean and standard deviation. P is the probability of a value of the distribution being less than x. Find

the missing value and sketch and shade the distribution.

mean	Standard deviation	x	P
85	4.5		0.68
mean	Standard deviation	x	P
	1	16	0.05
mean	Standard deviation	x	P
73		85	0.91
mean	Standard deviation	x	P
93	5		0.90

3. What is the z-score for the lower quartile in a standard normal distribution?
4. The manufacturing process at a metal-parts factory produces some slight variation in the diameter of metal ball bearings. The quality control experts claim that the bearings produced have a mean diameter of 1.4 cm. If the diameter is more than 0.0035 cm too wide or too narrow, they will not work properly. In order to maintain its reliable reputation, the company wishes to insure that no more than one-tenth of 1% of the bearings that are made are ineffective. What would the standard deviation of the manufactured bearings need to be in order to meet this goal?
5. Suppose that the wrapper of a certain candy bar lists its weight as 2.13 ounces. Naturally, the weights of individual bars vary somewhat. Suppose that the weights of these candy bars vary according to a normal distribution, with $\mu = 2.2$ ounces and $\sigma = 0.04$ ounces.

   a. What proportion of the candy bars weigh less than the advertised weight?
   b. What proportion of the candy bars weight between 2.2 and 2.3 ounces?
   c. A candy bar of what weight would be heavier than all but 1% of the candy bars out there?
   d. If the manufacturer wants to adjust the production process so that no more than 1 candy bar in 1000 weighs less than the advertised weight, what would the mean of the actual weights need to be? (Assume the standard deviation remains the same.)
   e. If the manufacturer wants to adjust the production process so that the mean remains at 2.2 ounces and no more than 1 candy bar in 1000 weighs less than the advertised weight, how small does the standard deviation of the weights need to be?

References

http://www.ic.nhs.uk/default.asp?sID=1198755531686

http://www.nytimes.com/2008/04/04/us/04poll.html

On the Web

http://davidmlane.com/hyperstat/A25726.html Contains the formula for the normal probability density function.

http://www.willamette.edu/ mjaneba/help/normalcurve.html Contains background on the normal distribution, including a picture of Carl Friedrich Gauss, a German mathematician who first used the function.

http://en.wikipedia.org/wiki/Normal_distribution Is highly mathematical.

Keywords

Concave down

Concave up

Cumulative density function

Density curve

Empirical Rule

Inflection Points

Normal distribution

Normal probability plot

Normal quantile plot

Probability density function

Standard normal curve

Standard normal distribution

Standardize

z-score