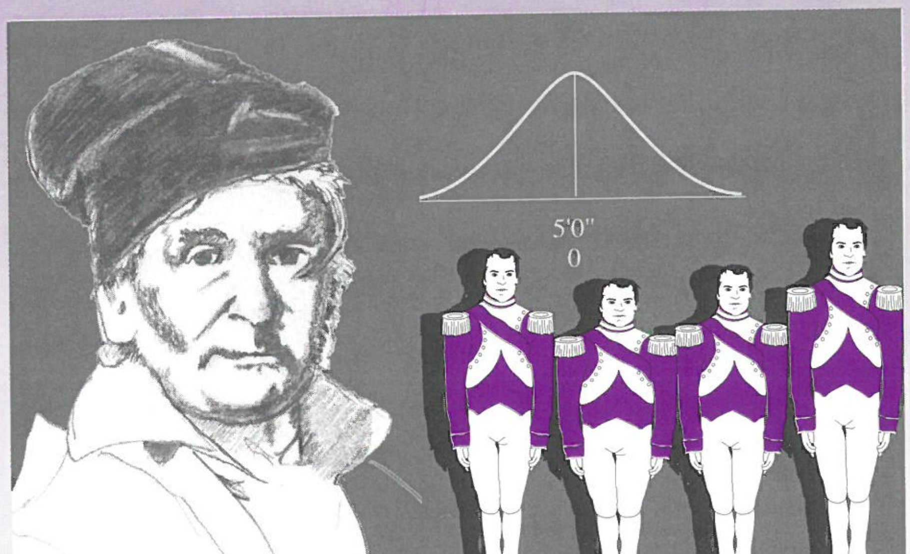
CHAPTER 4

Normal Distribution

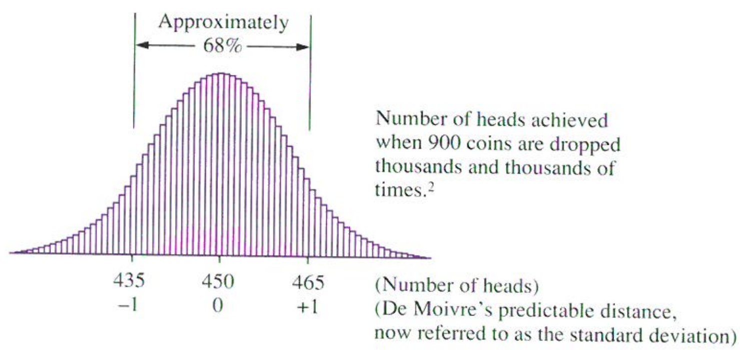


Carl Friedrich Gauss

4.0 Origins of the Concept

Perhaps the single most important distribution in all of statistics is the **normal distribution.** Its discovery dates back to the English mathematician, Abraham De Moivre[1] (1733), and his work on gambling experiments and is perhaps best illustrated with the following example.

Suppose a large number of well-balanced coins, say for instance 900 coins, are dropped on a table and the number of heads counted. How many heads would you expect? Many people would guess approximately 450 (half of 900) and, indeed, experience has shown that if this experiment were repeated thousands and thousands of times, most often you would get approximately 450 heads. However, on many occasions, you would get somewhat more than 450 heads and on many occasions somewhat less. If we were to actually record the results of these thousands and thousands of experiments into a histogram, it might appear as follows.

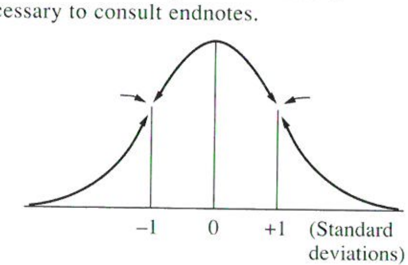


Number of heads achieved when 900 coins are dropped thousands and thousands of times.[2]

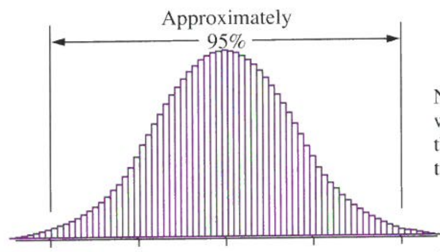
In studying these results, De Moivre noted that approximately 68% of the readings consistently fell within a predictable distance from the mean, denoted by the symbols −1 and +1. In other words, if you dropped 900 coins on a table and counted the number of heads, you would have a 68% probability there would be between 435 and 465 heads. De Moivre's predictable distance is now referred to as the **standard deviation.**[3*]

All information relevant to the understanding of the chapter is presented on each page as footnotes. However, certain information is presented at the end of the chapter as numbered *endnotes*, since they are mostly reference sources and historical fine points that tend to interfere with the flow of the material. It is not necessary to consult endnotes.

*De Moivre used the *inflection points* on the curve as his predictable distance. Inflection points are the points where the steep upward slope of the curve abruptly changes to a more gradual incline. This is useful when sketching the curve to properly estimate where the first standard deviation lines are located.



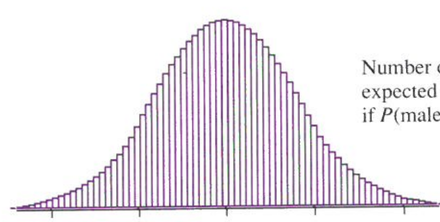
Furthermore, De Moivre noted that approximately 95% of the readings fell within −2 and +2 predictable distances (standard deviations) of the mean, as follows.



De Moivre realized he had discovered something important (he spent over 12 years on it, ultimately deriving the equation for the normal distribution and calculating probabilities associated with its use that to this day would be considered quite accurate) but for reasons[4] was unable to interest others.

Half a century later, the famous French mathematician, Pierre Simon Laplace was working on a probability experiment similar to that of De Moivre, only Laplace's experiment (1781) concerned newborn infants.[5] Laplace was trying to prove that male babies were born with a higher frequency than female babies. Although Laplace's work is quite complex, let's consider the following simplified example.

Suppose we assume the probability of a male birth is $\frac{1}{2}$ (50%). If we equate the probability of having a male baby to the probability of achieving a head when a coin is tossed (which is also $\frac{1}{2}$),[6] then the resulting distribution of heads achieved when dropping 500,000 coins on a table, thousands and thousands of times, should be equivalent to the resulting distribution of male births achieved when 500,000 babies are born, in thousands and thousands of cities. In other words, we would expect approximately 250,000 heads (or approximately 250,000 male births) each time. However, on many occasions we would get somewhat more than 250,000 and on many occasions somewhat less. According to the laws of probability, the resulting distribution should look as follows.[7]



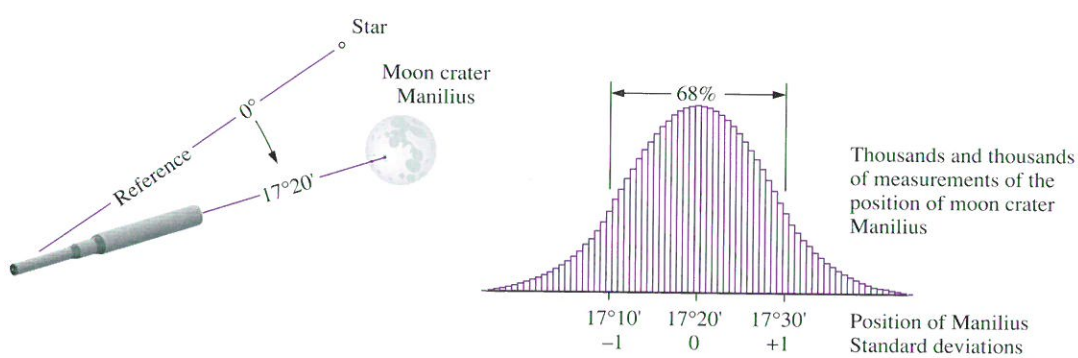
As it turned out, in the years 1745–1770 in Paris, there were equivalent to 254,856 male births out of 500,000.[8] According to the previous histogram, the likelihood of this occurring is almost 0. Just look at the histogram. The probability of getting over 251,000 male births is negligible. To get over 254,000 would be considered nearly impossible.

Laplace correctly reasoned that the probability of a male birth must then be greater than 50% and more likely closer to 51% (254,856/500,000 ≈ 51%). Laplace substantiated these findings using birth records from other European cities such as London and Naples, which also produced similar ratios with only minor variation that Laplace attributed to climate, food, or custom.[9] To this day, these probabilities hold true worldwide. The probability of a male birth is known to be nearly 51%, a female birth 49%.[10]

Although Laplace republished his findings in 1786, the work attracted only minor attention, probably due to the extreme complexity of his mathematical development.

The next stage in this unfolding discovery had to wait an additional 30 years for the work of the famous German mathematician, Carl Gauss. Interest in planetary motion dominated Europe in the late 1700s. Astronomers and mathematicians were encouraged with national grants and contests to correctly measure the position of certain stars and other celestial bodies, which were to be used to determine precise longitudinal measurements for sea navigation. However, imperfections in telescopic lenses (along with the imperfections and variations in the human eye) produced measurement errors that interfered with determining exact positions. These errors and how to deal with them perplexed astronomers for more than half a century until Carl Gauss in 1809 correctly reasoned that the errors of observation had to be distributed much like the heads in a coin experiment and thus created a minor revolution, at least in the field of astronomy.[11] Although the actual mathematics involve trigonometric equations, the underlying principle is quite simple. Suppose we use the following example.

Say the *true* position of the Moon's crater, Manilius,[12] at a certain place and time was known to be precisely 17°20′ from a known reference star. If we were to take thousands and thousands of measurements, the average of all these measurements might indeed be 17°20′, however many measurements would be greater than 17°20′ and many less. If we were to record all these thousands and thousands of measurements, the results might appear as follows.



Notice how the distribution takes on that familiar bell-like shape, with the measurements symmetrical about the mean, and that approximately 68% of the errors fall within a certain predictable distance, which we now refer to as the standard deviation.[13]

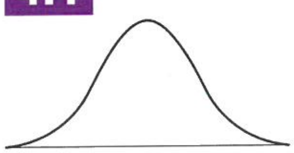
The next two decades following Gauss's discovery were spent mostly gathering large bodies of data related to measurement error and working out mathematical theories associated with its use. However, outside the fields of astronomy and probability theory, the distribution we now call *normal* was relatively unknown.[14] Starting in the 1830s this started to change. In fact, much of the remainder of the 1800s was spent exploring applications to other fields.[15]

By the end of the century, the normal distribution had been successfully applied to such fields as experimental psychology (reaction times, stimuli-sensation measurement, memory), physics (molecular motion), biology (human height and chest measurements, size of fruit and other characteristics of plant and animal life), and education (talent and abilities as demonstrated by examination scores). By the 1930s, statistical techniques based on the normal distribution had become integral parts of the fields of biometric research, factory production, economics, and agriculture, and was on the verge of incorporation into numerous other fields.

Cautionary Note: One of the myths of statistics is that most natural phenomenon, given enough observations, will take on a normal distribution. This is not so. However, still to this day, some of those trained in experimental research cling to this false belief. In fact, much natural phenomenon is skewed, bimodal, or exhibits a variety of distributive forms. Although *some* natural phenomenon can be closely estimated with the normal distribution, the normal distribution's importance derives more from its use in sampling theory, where this distribution reoccurs with uncanny repeatability, which is discussed in section 4.4 and in chapter 5.

Because a full understanding of the normal distribution is so vital for statistical inference (that is, use of samples to estimate population characteristics), we will spend the remainder of this chapter exploring its intricacies.

4.1 Idealized Normal Curve



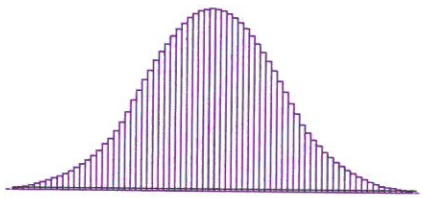
The concept of the idealized normal curve originally stems from the writings and mathematical methods of Laplace (1781, 1786)[16] and is based on the following underlying assumptions.

$n \to \infty$ **a.** The number of observations approaches infinity.

In other words, the number of observations (n) is enormously large (maybe millions or billions of measurements) on the same phenomenon.

$\Delta x \to 0$ **b.** The change in x approaches 0.

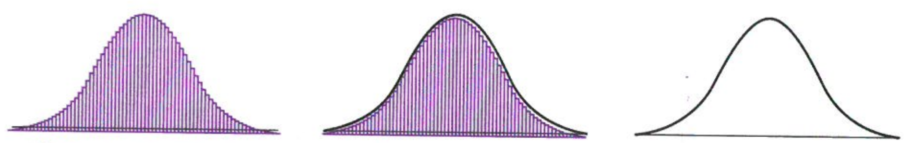
The change in x refers to the *width* of the histogram bars. To say the width "approaches zero" means the histogram bars are exceedingly narrow. In other words, if you were to measure adult height, the data must be grouped in exceedingly narrow categories. Say one histogram bar might represent all women $5'4\frac{1}{100}''$ while the next histogram bar represents all women $5'4\frac{2}{100}''$, the next $5'4\frac{3}{100}''$, which are exceedingly narrow groupings.



c. The resulting histogram contains thousands of histogram bars, which form into the shape of a bell.

If we connect the tops of the histogram bars, it would take on a smooth flowing appearance, which we refer to as the **normal curve.**

These conditions can be summarized pictorially as follows.



Enormous amounts of measurements on one phenomenon are grouped into a histogram with exceedingly narrow intervals and form the shape of a bell.

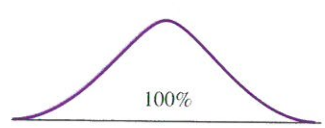
The line connecting the tops of the histogram bars is called a frequency polygon line.

Histogram bars are erased, with only the smooth flowing frequency polygon line remaining. This is the normal curve.

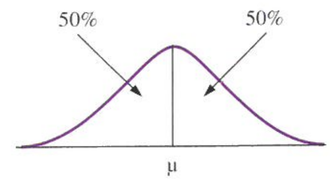
Although the normal curve is somewhat of an idealized construction (it's rare that we can obtain millions or billions of measurements on one phenomenon), experience has shown it to be an indispensable tool in predicting probabilities associated with sampling. Note that the shape of the normal curve can vary somewhat; however, certain characteristics are common to all normal curves, and these are presented next.

Characteristics of the Normal Curve

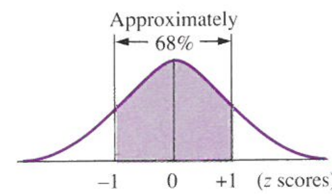
a. Bell-shaped, fading at tails. Theoretically, the distribution continues indefinitely in both directions, approaching but never touching the horizontal axis.



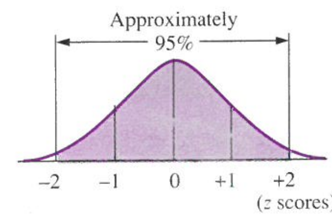
b. The total amount of data is 100%, symmetrical about the mean, μ, with 50% of the data above the value of μ, and 50% below.



c. Approximately 68% of the data lies within −1 and +1 standard deviation of the mean, and approximately 95% of the data lies within −2 and +2 standard deviations of the mean.

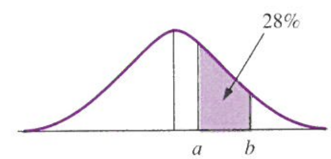


(Recall from chapter 2, section 2.6, a *z* score is defined as the number of standard deviations a value is away from the mean.)



d. The percentage of data between any two points is equal to the probability of randomly selecting a value between those two points.

For example, if 28% of the data lies between points *a* and *b*, and if you randomly select one value from the entire population, the probability this one value will be between *a* and *b* is 28%.



A brief word about terminology: we will often refer to the percentage of data in some part of the normal curve as an **area.** The terminology stems from calculus and pervades much of statistical writing, including this text. Just keep in mind, if we refer to the

Area in some shaded interval	this is equivalent to	% of data in that same interval

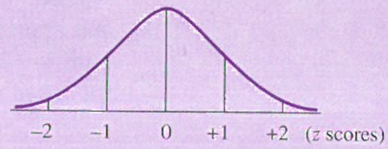
Now let's get down to specifics. How do we obtain precise percentages associated with the normal curve? We merely look them up in the normal curve table in the back of the text.

Use of the Normal Curve Table

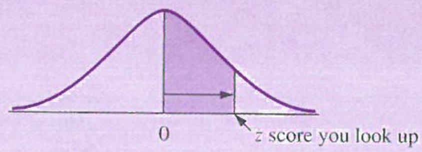
Kramp was the first to tabulate the exact probabilities associated with the normal distribution, which appeared in 1799 in a book concerning the refraction of light. These tables were used for about 100 years and the tables in use today are only slight variations of the original.[17] Let's see how a contemporary table works.

Although contemporary tables vary slightly in structure, our table requires the understanding of three rules, two of which are presented below.

Rule 1
z scores (number of standard deviations from the mean) are used to locate position in the normal curve.



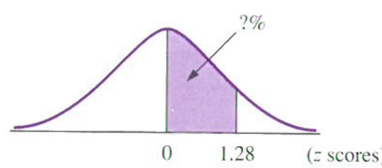
Rule 2
The table only gives the percentage of data *starting from the middle, z* = 0, out to the *z* score you look up.



At the upper right hand corner of the normal curve table (refer to ''Statistical Tables'' in back of book, or for quick reference, see inside cover) is a demonstration example. Let's use it as our first example.

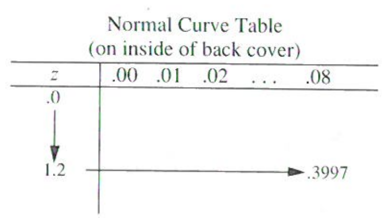
Example

Find the percentage of data (or area) from $z = 0$ to $z = 1.28$.



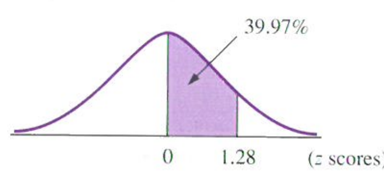
Solution

Look under the z column to 1.2, then across to the .08 column (notice that $1.2 + .08 = 1.28$). Here we find the decimal .3997. To change .3997 to a percentage, we move the decimal two places to the right .39.97, to get 39.97%.



Answer

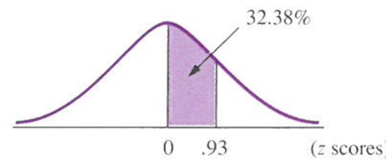
39.97% of the data lies between $z = 0$ and $z = 1.28$.



■

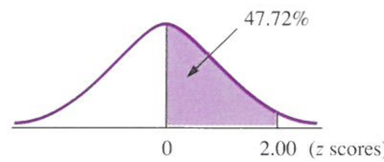
See if you can solve the following two practice problems without looking at the answers.

Practice 1 Find the % of data between $z = 0$ and $z = .93$.



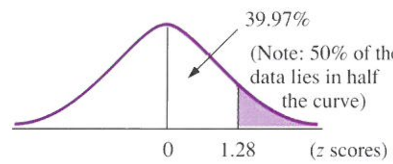
Answer 32.38%. Look under the z column for .9, then across to .03 (note: .9 + .03 = .93).

Practice 2 Find the percentage of data between $z = 0$ and $z = 2.00$.

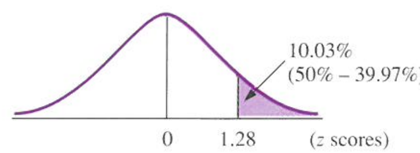


Answer 47.72%. Look under the z column for 2.0, then across to .00 (note: 2.0 + .00 = 2.00).

Example Find the percentage of data *above* $z = 1.28$.



Solution Since the table reads from the middle ($z = 0$) out, when we look up $z = 1.28$, we get the percentage of data from $z = 0$ to $z = 1.28$, which is 39.97%. However, this is not the answer to our question. But, if we remember 50% of the data is in half the curve, then



$$39.97\% + ? = 50.00\%$$

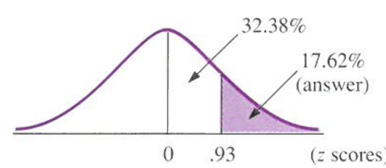
We solve this by subtracting 39.97% from both sides, to get

$$\begin{aligned} ? &= 50.00\% - 39.97\% \\ &= 10.03\% \end{aligned}$$

Answer 10.03% of the data lies *above* $z = 1.28$.

■

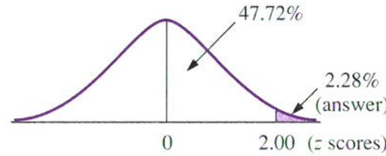
Practice 3 Find the percentage of data *above* $z = .93$.



Answer 17.62% (50% minus 32.38%)

Practice 4

Find the percentage of data *above* $z = 2.00$.



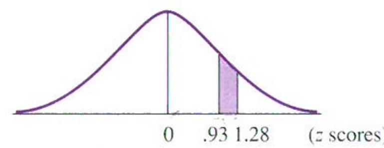
Answer

2.28% (50% minus 47.72%).
Remember: the two percentages (2.28% and 47.72%) must add up to 50%.

The table can also be used to get the percentage of data in any ''slice'' of the normal curve.

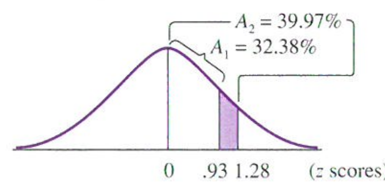
Example

Find the percentage of data from $z = .93$ to $z = 1.28$.



Solution

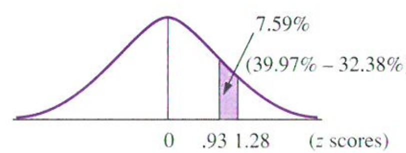
Find the percentage of data from $z = 0$ to 1.28. Then find the percentage of data from $z = 0$ to $z = .93$. Subtract the two percentages to get the answer. Perhaps a formula would be helpful.



$$\begin{aligned} A_{\text{shaded}} &= A_2 \text{ minus } A_1 \\ &= 39.97\% - 32.38\% \\ &= 7.59\% \end{aligned}$$

Answer

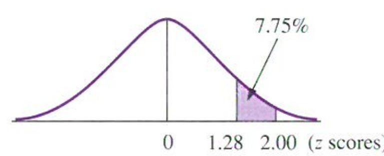
7.59% of the data lies between $z = .93$ and $z = 1.28$.



■

Practice 5

Find the percentage of data from $z = 1.28$ to $z = 2.00$.

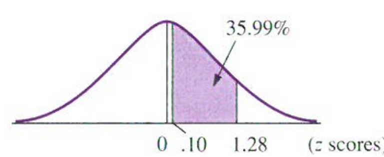


Answer

7.75% (47.72% minus 39.97%)

Practice 6

Find the percentage of data from $z = .10$ to $z = 1.28$.



Answer

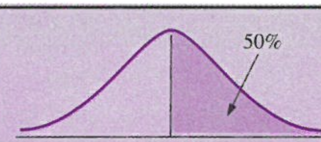
35.99% (39.97% minus 3.98%)

Note: To get the percentage of data for $z = .10$, we look down the z column to .1, then across to the .00 column (.1 + .00 = .10)

The third and last rule in using our normal curve table is as follows:

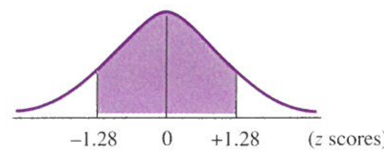
Rule 3

The table only gives percentages for half the curve.

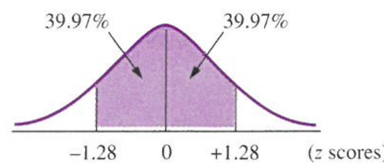


However, since both halves are identical, the percentage of data from, say, $z = 0$ to $z = 1.28$ is identical to the percentage of data from $z = 0$ to $z = -1.28$. In other words, the percentage of data from $z = 0$ to $z = 1.28$ is 39.97% and the percentage of data from $z = 0$ to $z = -1.28$ is 39.97%. This is demonstrated in the following example.

Example Find the percentage of data between $z = -1.28$ and $z = +1.28$.

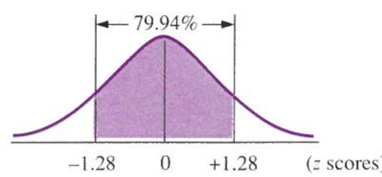


Solution First get the percentage of data from $z = 0$ to $z = +1.28$ by looking up $z = 1.28$. This gives us 39.97%. Next we get the percentage of data from $z = 0$ to $z = -1.28$ by looking up $z = 1.28$. This also gives us 39.97%. Notice in the diagram that we must *add* the two areas to get the total percentage of data from $z = -1.28$ to $z = +1.28$.



$$39.97\% + 39.97\% = 79.94\%$$

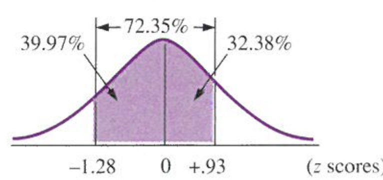
Answer 79.94% of the data lies between $z = -1.28$ and $z = +1.28$.



■

Practice 7 Find the percentage of data between $z = -1.28$ and $z = +.93$.

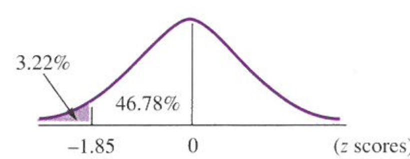
Answer 72.35% (39.97% + 32.38%)



Practice 8 Find the percentage of data *below* $z = -1.85$.

Answer 3.22%

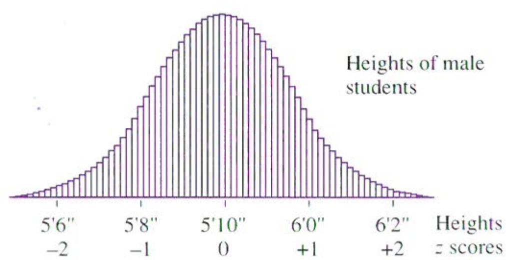
Note that the percentage of data from $z = 0$ to $z = -1.85$ is 46.78%. We then subtract this from 50% to get the answer: 50.00% minus 46.78% = 3.22%.



4.2 Applications: Idealized Normal Curve

It was widely believed in the last century that once enough data is gathered almost all natural phenomenon will be shown to be normally distributed. Although today we know this not to be true, we do find much in nature and life that can be closely approximated with the idealized normal curve. Let's use the following example to demonstrate.

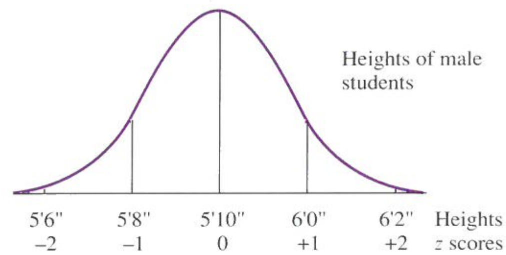
Suppose we measure the height of every male student enrolled at the Community College at Maxwell Airforce Base, Alabama,* and find the *average* height to be $\mu = 5'10''$. Now, what are the chances that every male student at this College will be 5′10″? Of course, this is absurd. Although many will be in the vicinity of 5′10″, the bulk of the students will probably be somewhat shorter than 5′10″ or somewhat taller. Experience has shown that if we were to represent these height measurements in the form of a *histogram*, that chances are the histogram will build into the shape of a normal distribution and might look as follows.†



*Although probably having the largest enrollment of any community college in the country, the College at Maxwell Airforce Base offers no on-campus courses. Instead, the college acts as a "clearing house" for incoming transfer credit from airforce personnel all over the world.

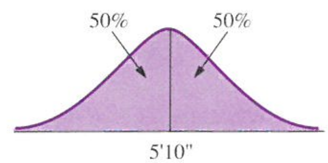
†Adolphe Quetelet (1846) was probably the first to demonstrate a population of male heights as closely fitting a normal distribution. For data, he used the heights of 100,000 French conscripts from the early 1800s. Can you guess the average height of a French soldier in those days? The answer is, $\mu = 5'0''$, for all males. Quetelet also showed chest measurements of nearly 6000 Scottish soldiers to be near normally distributed.

The spread of values would depend on a number of factors,* however let's say for this particular population, we calculated the standard deviation to be $\sigma = 2''$.† If we fit an idealized normal curve over the data, the resulting representation would look as follows.



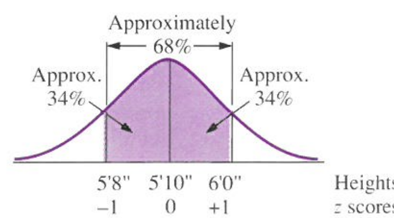
Just from knowing we have a normally distributed population with average height $\mu = 5'10''$ and standard deviation $\sigma = 2''$, we can immediately determine the following.

- 50% of the heights will be over 5′10′′, and 50% of the heights will be under 5′10′′.



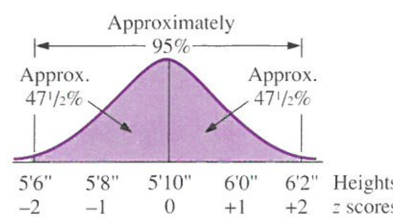
- Approximately 68% of the heights will be within ±1 standard deviation of the mean, μ—that is, approximately 68% of the heights will be from 5′8′′ to 6′0′′.

  Since the curve is symmetrical, approximately 34% (half 68%) will be from 5′8′′ to 5′10′′ *and* approximately 34% will be from 5′10′′ to 6′0′′.



- Approximately 95% of the heights will be within ±2 standard deviation of the mean, μ—that is, approximately 95% of the heights will be from 5′6′′ to 6′2′′.

  Since the curve is symmetrical, approximately $47\frac{1}{2}$% (half of 95%) will be from 5′6′′ to 5′10′′ *and* approximately $47\frac{1}{2}$% will be from 5′10′′ to 6′2′′.



*Certain normal populations have been shown to be comprised of a number of smaller normal populations. In other words, several cultural groups in this case might mix (each with a different average height and normal distribution) into one larger composite normal distribution.

†The standard deviation for height is generally closer to 2.4″; 2″ was used for demonstration purposes.

To determine more precise percentages, we must refer to the normal curve table, which requires z scores to be calculated to two decimal places. For this, we use the following formula.

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

- x: point at which we wish to know the z score
- μ: the mean of the normal curve
- σ: the standard deviation of the normal curve

Recall, a z score is the number of standard deviations a value is away from the mean. Now let's look at an example.

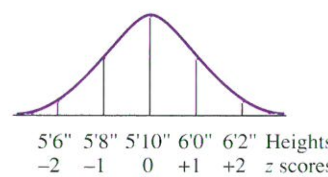
Example

Suppose the heights of all male students at the Community College at Maxwell Airforce Base are known to be normally distributed with $\mu = 5'10''$ and $\sigma = 2''$, find the percentage of male students over $6'0''$.

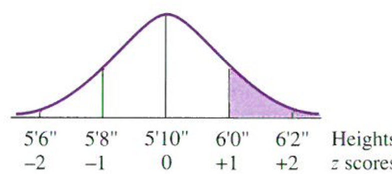
Solution

We proceed in four steps.

a. Draw normal curve, listing real data and z scores for at least ±2 standard deviations.



b. Shade the area in question, in this case, over $6'0''$.



c. Calculate the z score at the cutoff ($6'0''$). This point is represented by the symbol, x, in the z formula.

$$Z = \frac{x - \mu}{\sigma} = \frac{6'0'' - 5'10''}{2''}$$

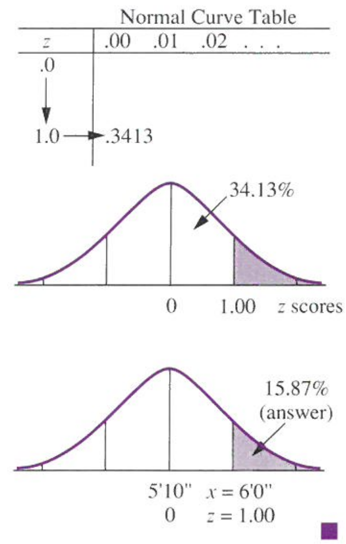
$$= \frac{2''}{2''} = +1.00$$

d. Now we must stop and think a moment as to how the normal curve table can be used to get the information we wish. If we look up $z = 1.00$, we get 34.13% (.3413), but this is *not* the answer.

However, if we subtract 34.13% from 50%, we get the percentage of data in the shaded region.

$$A_{\text{shaded}} = 50\% - 34.13\%$$
$$= 15.87\%$$

Note: 34.13% plus 15.87% equals 50% (half the curve).



Answer 15.87% of the male students registered at the Community College at Maxwell Airforce Base are expected to be over 6′0″. ■

Example Referring to the above problem: if we were to randomly select *one* male student from this Community College, what is the probability this one student would be over 6′0″?

Answer Since 15.87% of the male students are over 6′0″ (according to the above problem), then the probability of randomly selecting a male student over 6′0″ is 15.87%. ■

Example Again referring to the above problem: what percentage of the area under the curve is in the shaded region?

Answer The words, area, percentage of data, and probability, have much the same meaning when discussing the normal curve. Thus, we can state,

$$\begin{array}{c}\text{Area of}\\ \text{shaded region}\end{array} = \begin{array}{c}\text{Percentage of data}\\ \text{in shaded region}\end{array} = \begin{array}{l}\text{Probability of}\\ \text{selecting a male} = 15.87\%\\ \text{in shaded region}\end{array}$$
■

To Solve a Normal Curve Problem

1. Draw a separate normal curve for each example, indicating values for at least ±2 standard deviations.
2. Shade the area asked for in the question.
3. Calculate the *z* score(s) at the cutoff(s).
4. Stop and think a moment as to how the normal curve table can best be used to get the specific information we request.

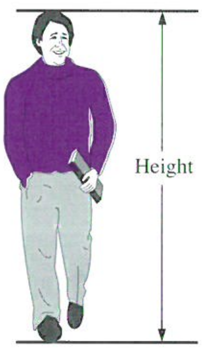
Using the same population, now let's ask a different question.

Example

Suppose the heights of male students at the Community College at Maxwell Airforce Base are known to be normally distributed with $\mu = 5'10''$ and $\sigma = 2''$. Find the percentage of male students who are over $5'7\frac{1}{2}''$.

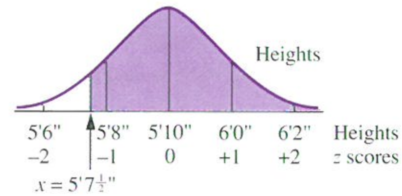
Solution

We proceed using the same four steps.



a. Draw the normal curve, listing real data and z scores for at least ± 2 standard deviations and

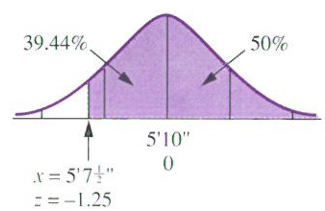
b. Shade the area in question.



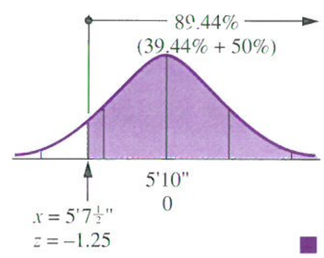
c. Calculate the z score at the cutoff (in this case, at $5'7\frac{1}{2}''$).

$$z = \frac{x-\mu}{\sigma} = \frac{5'7\frac{1}{2} - 5'10}{2''} = \frac{-2\frac{1}{2}}{2}$$
$$= \frac{-2.5}{2} = -1.25$$

d. Look up $z = -1.25$, which gives us 39.44%. However this is not the complete answer. If we examine the diagram, we note that this is the percentage of males from $5'10''$ to $5'7\frac{1}{2}''$ (that is, from $z = 0$ to $z = -1.25$). To get the complete answer we must add 50%, the percentage of males over $5'10''$.



$$A_{\text{shaded}} = 39.44\% + 50\%$$
$$= 89.44\%$$

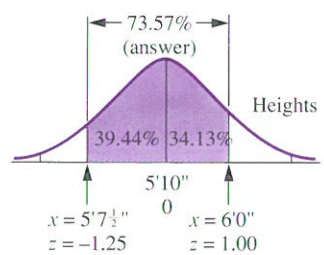


Answer

89.44% of the male students registered at the Community College at Maxwell Airforce Base are expected to be over $5'7\frac{1}{2}''$. ■

Practice 1

For the problem above, find the percentage of male students who are $5'7\frac{1}{2}''$ to $6'0''$.



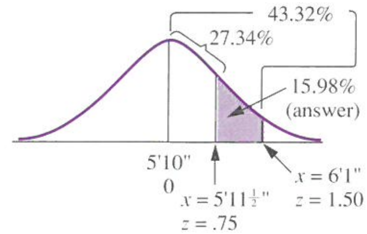
Answer

73.57% (39.44% + 34.13%)

Note: we must look up two %'s of data and then add the two together.

Practice 2

For the problem above, find the percentage of male students who are $5'11\frac{1}{2}''$ to $6'1''$.



Answer

15.98% (43.32% − 27.34%)

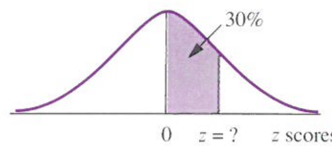
Note: Again, we must look up two %'s of data, only this time we subtract.

4.3 Working Backward with the Normal Curve Table

The normal curve table can also be used in reverse. That is, if we already know the percentage of data in a certain region, we may be able to use the normal curve table to find the z score at the cutoff.

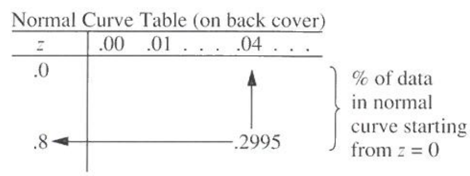
Example

Suppose the percentage of data in the normal curve from $z = 0$ to $z = ?$ is known to be 30%. Find the missing z score.



Solution

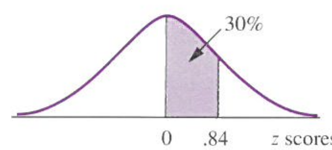
Since the table reads from $z = 0$ out to $z = ?$ and we already know the percentage of data in this region is 30% (.3000 in table), we merely use the table in reverse. First, we find the percentage of data closest to .3000 (30%), which turns out to be .2995.



Next, we look across to the z score indicated, to get .8 and up to get .04. The z score is $z = .84$. (Note: if the percentage of data falls precisely midway between two values, we round to the higher z score.)

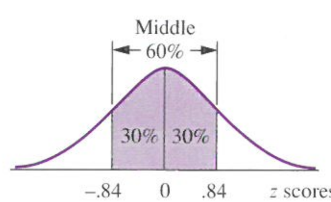
Answer

The missing z score is $z = .84$ (in other words, approximately 30% of the data lies between $z = 0$ and $z = .84$).



Practice 1

Find the z scores associated with the *middle 60%* of the data in the normal curve.



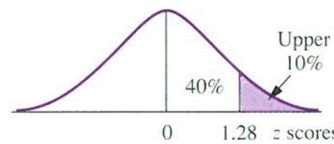
Answer

$z = -.84$ to $z = +.84$

Note: We must split 60% into 30% plus 30%. When we look up 30% (closest value is .2995), we get $z = -.84$ and $z = +.84$.

Practice 2

Find the z score associated with the *upper 10%* of the data.



Answer

$z = +1.28$

Note: We must look up 40% in the table, since the table starts reading from $z = 0$ outward. The closest value to 40% (.4000) is .3997, which gives us $z = 1.28$ at the cutoff.

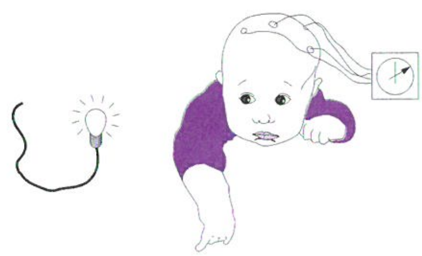
Applications

It's long been known by experimental psychologists that different people react to the same stimuli in different times. For instance, an automobile driver responds to danger by jamming on the brakes but the precise time to react will vary from individual to individual. Similarly, the speed at which a student reacts to (or absorbs) facts in a classroom varies from individual to individual. Gerling (1838) was probably the first to demonstrate reaction times as normally distributed.[18]

Let's demonstrate with the following experiment.

Example

Suppose a researcher concerned with measuring forms of intelligence in newborn infants sets up an experiment to electronically monitor the neurological reaction time when a tiny light is flashed into a baby's eye.



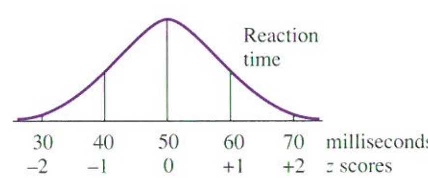
After testing thousands and thousands of newborn infants, it was found that the *average* reaction time was $\mu = 50$ milliseconds (ms) with standard deviation $\sigma = 10$ ms. Assuming the reaction times are normally distributed, below what value would you expect to find the *fastest* 10% of the reaction times?

Solution

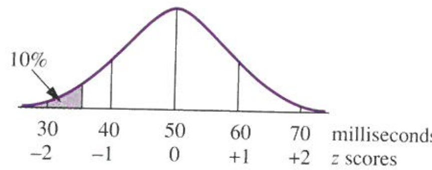
This is a typical working-backward problem where the percentage of data is given (in this case, the fastest 10% of the values), and we use the normal curve table in reverse to determine the z score.

To start, we proceed in much the same way as solving any normal curve problem.

a. Draw the normal curve, listing real data and z scores for at least ± 2 standard deviations.

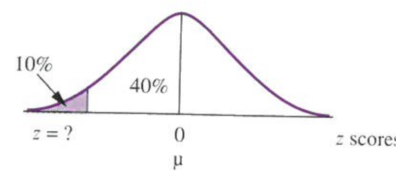


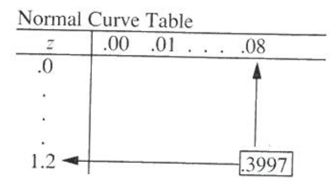
b. Shade the area in question. Since the fastest times would be *less than* 50 ms, we shade the extreme left of the normal curve, estimating 10%.



c. Next obtain the z score at the cutoff. To do this, we look up 40% (.4000) since the table reads data only from the *center* ($z = 0$) out.

Note in the table that the closest value to 40% (.4000) is .3997, which gives us a z score of $z = 1.28$. Since the z value is below μ, we must make the z value negative. Thus, $z = -1.28$.





d. Now use the z formula to solve for x, the real data value at the cutoff. Essentially we know z (-1.28), and we wish to solve for x in the formula:

$$z = \frac{x - \mu}{\sigma} \quad \begin{cases} \mu = 50 \\ \sigma = 10 \end{cases}$$

$$-1.28 = \frac{x - 50}{10}$$

$$(10)(-1.28) = x - 50$$

$$-12.8 = x - 50$$

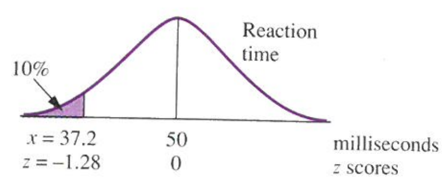
$$37.2 = x$$

$$\text{or } x = 37.2 \text{ ms}$$

This calculation required some algebraic manipulation. First, we multiplied both sides of the equation by 10 to obtain $-12.8 = x - 50$. Second, we added $+50$ to both sides of the equation to get $37.2 = x$. In other words, at the cutoff, $x = 37.2$ ms.

Answer

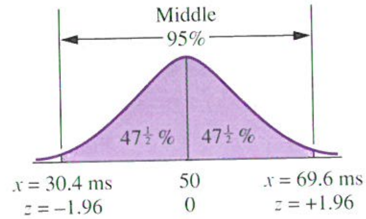
Below 37.2 ms you would expect to find the fastest 10% of the reaction times.



■

Practice 3

For the preceding problem, between what two values would you expect to find the *middle* 95% of the reaction times?



Answer

Between 30.4 and 69.6 ms

Note: We must look up $47\frac{1}{2}$% (half of 95%), or in decimal form .4750, to obtain $z = 1.96$ on both sides. When we substitute $z = -1.96$ and $z = +1.96$ in our z formula, we obtain the following:

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

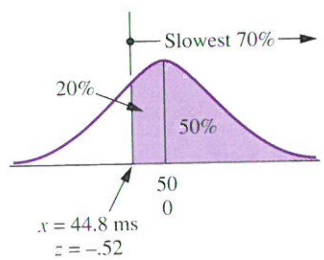
$$-1.96 = \frac{x - 50}{10} \qquad +1.96 = \frac{x - 50}{10}$$

$$x = 30.4 \text{ ms} \qquad x = 69.6 \text{ ms}$$

(To solve, multiply both sides by 10, then add 50 to both sides.)

Practice 4

For the above problem, *above* what value would you expect to find the *slowest 70%* of the reaction times?



Answer

Above 44.8 ms

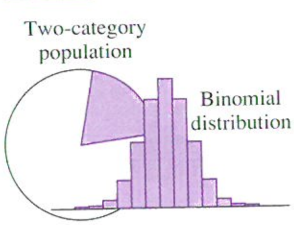
Note: 50% of the data is above $\mu = 50$ ms so we must look up the remaining 20% (.2000). The closest value to .2000 is .1985, which is equivalent to $z = -.52$. Substituting $z = -.52$ in our z formula, we obtain the following:

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

$$-.52 = \frac{x - 50}{10}$$

$$x = 44.8 \text{ ms}$$

4.4 Binomial Distribution: An Introduction to Sampling



Although some natural populations have distributions that can be approximated with the normal curve, the normal curve's importance is derived more from its consistent and uncanny ability to predict the outcomes when we *sample* from a population. Although different "types" of populations exist (from which we may sample), one of the most important in research is the two-category population.

A **two-category population** is a population where every member is classified into exactly one of two categories.