

Elementary Statistics

Brian E. Blank

February 8, 2016

FIRST PRESIDENT UNIVERSITY PRESS

Chapter 3. The Normal Model

How can a university admissions officer compare two students, one with an ACT score and one with an SAT score? How can a calculus professor fairly use an exam from a previous year as a make-up exam for a missed exam if the scores on the exam from the earlier year tended to be lower than the scores on the exam that was missed? For the answers to these and other questions, read on.

3.1 The Standard Normal Distribution

In the first year of a Calculus course, it is shown that there is a (unique) number e such that, for every x , the instantaneous rate of change of e^x is equal to itself, e^x . The calculus notation for expressing this property is

$$\frac{d}{dx}(e^x) = e^x.$$

The function that returns the output value e^x when x is the input value is called the ***natural exponential function*** and is denoted by \exp . (The abbreviation is similar to those of trigonometry in which three character names are used for trigonometric functions: \cos for the cosine, \tan for the tangent, \cot for the cotangent, \sec for the secant, \csc for the cosecant, and \sin for the sine—think of all the time that is saved writing \sin instead of sine .) Thus,

$$\exp(x) = e^x.$$

It is often useful to have the \exp notation when we write mathematics inline, because $\exp(x)$ requires less vertical space than e^x . (Notice that the font for x in the latter expression has had to be reduced.)

The number e has the value 2.718281828459... . The repetition of 1828 is quite a fluke: no pattern to the digits of e exists. In particular, there is no repeating pattern. Nevertheless, *this* author can say that the repetition of the 1828 makes it easy to remember 10 significant digits of e : anybody can recall the two digits 2.7, and 1828 was the year Franz Schubert died.¹

The Bell Curve

The graph of

$$y = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \tag{3.1.1}$$

which is shown in Figure 3.1.1, is called the ***Bell Curve***

¹The author once taught from a calculus book coauthored by Robert McDowell. It contained a sentence that started in the way this memory aid for the digits of e starts. McDowell, a lover of Schubert lieder, wanted to conclude the sentence in the way that the author has. However, McDowell's editor would not permit it: Schubert, in the editor's opinion, was too obscure. So McDowell finished the sentence with "and 1828 was one year after the year Beethoven died." Somehow McDowell's editor was OK with that.

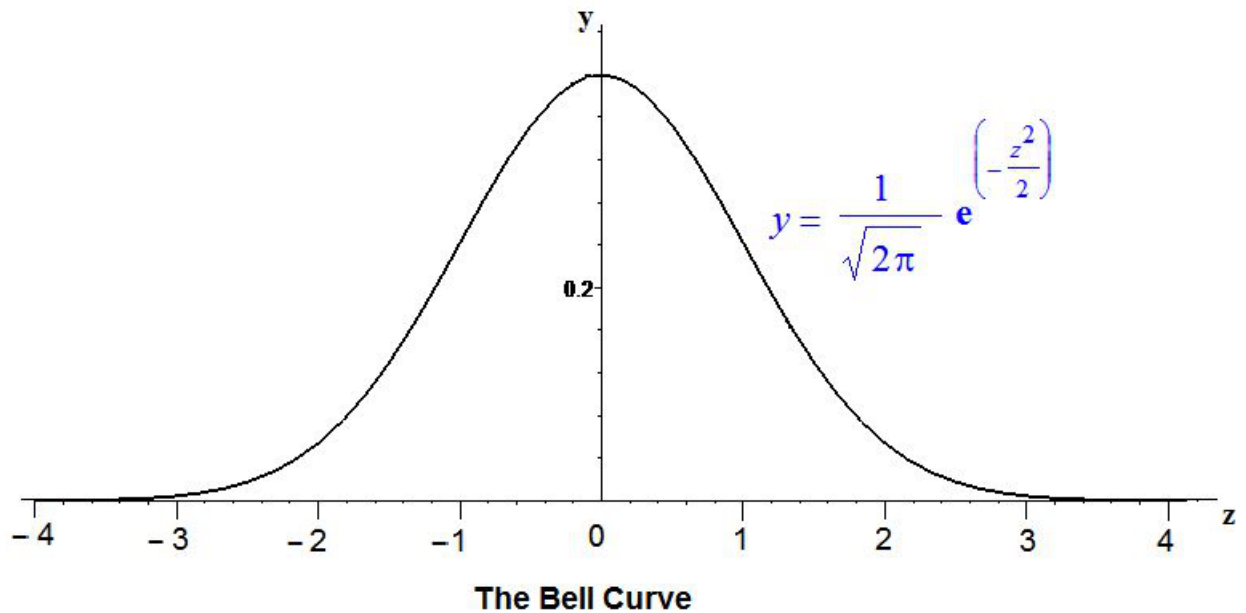


Figure 3.1.1 The Bell Curve

The reason for the name *Bell Curve* is that the curve looks like the outline of a bell.² The truth is, however, that the Bell Curve only looks like a bell when the y -axis is stretched enough to generate a pleasing picture. From the graph, it appears that the y -intercept is 0.4, but it is actually 0.3989422802... In truth, the value of the y -intercept is not terribly significant. We mention it only to draw attention to the different scales that are used for the horizontal and vertical axes: whereas 0.4 units on the y -axis takes you upward from the origin all the way to the y -intercept, starting at the origin on the horizontal axis and moving to the right 0.4 units gets you only as far as the second little tick mark to the right of the origin. If we had plotted the Bell Curve using equal scaling on the axes, so that one unit on the vertical axis had the same length as one unit on the horizontal axis, then the Bell Curve would look like a very flimsy bell run over and flattened by a very heavy truck.

Another possible misconception might be suggested by the figure. It appears that the graph touches the horizontal axis on the left and the right—where exactly is not so evident, but it appears that the ordinate y reaches 0 by the time the abscissa z reaches -4 on the left and $+4$ on the right. That cannot actually be the case because the factors that compose the formula for y are all positive regardless of the value of z . What is happening is that, for the scale used in the figure, the values of y corresponding to $z < -4$ and $z > 4$ are too small to be distinguished from $y = 0$. The reality is that the abscissas extend from negative infinity to positive infinity and, for every such abscissa z , the point (z, y) on the Bell Curve is above the horizontal axis.

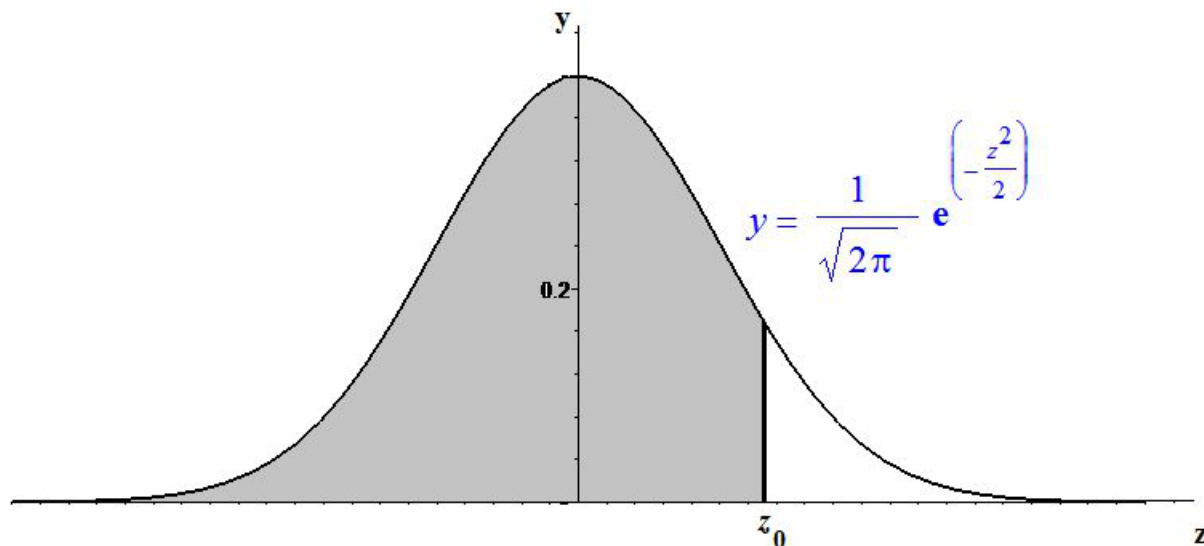
Observe that the graph appears to be symmetric about the y -axis, and it really is. We can see that from the formula for the graph of the Bell Curve. To be specific, because z appears only as a square, the values $z = z_0$ and $z = -z_0$ result in the same y . Consequently, each side of the vertical axis is a mirror reflection of the other side. Other than that, the formula for the Bell Curve will have absolutely no importance for us.

On the other hand, the area under the graph of the Bell Curve will be of paramount importance. Although the graph extends from negative infinity to positive infinity, the area of the region below the Bell Curve and

²The author did not really think the reader would be mystified. Sometimes authors feel compelled to state the obvious for reasons of completeness.

above the horizontal z -axis is finite.³ As Figure 3.1.1 suggests, most of the area under the graph is between $z = -3$ and $z = 3$. In fact, as we will learn, about 99.7% of the area under the Bell Curve is between $z = -3$ and $z = 3$. About 99.994% of the area under the curve is between $z = -4$ and $z = 4$. Because we are interested in the Bell Curve thanks to the area that lies below it, and because 0.006% is an incontrovertibly inconsequential percentage, we will have little use for the points on the Bell Curve with $|z| > 4$.

Areas under the Bell Curve are commonly represented by the Greek letter Φ (pronounced *fly* to rhyme with *why*, except by those who pronounce it *fee* to rhyme with *fee*⁴). The area under the graph of the Bell Curve that is to the *left* of the vertical line $z = z_0$ (extending indefinitely toward $-\infty$) is denoted $\Phi(z_0)$. See Figure 3.1.2.



The area of the shaded region is denoted by $\Phi(z_0)$

Figure 3.1.2 Area Under the Bell Curve

For convenience, we will call the shaded region of Figure 3.1.2 a “left tail” even though it includes the central hump. (Left tails are also called “lower tails”.) Readers with a background in calculus⁵ will recognize that $\Phi(z_0)$ can be expressed as an improper integral:

$$\Phi(z_0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_0} \exp\left(-\frac{1}{2}t^2\right) dt.$$

We mention in passing that the **error function**, commonly denoted by erf , is closely related to the Φ function. The error function is defined by

$$\text{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^x \exp(-t^2) dt.$$

³Readers who have studied second semester calculus will be aware that the concept that underlies this fact is that of a *convergent improper integral*.

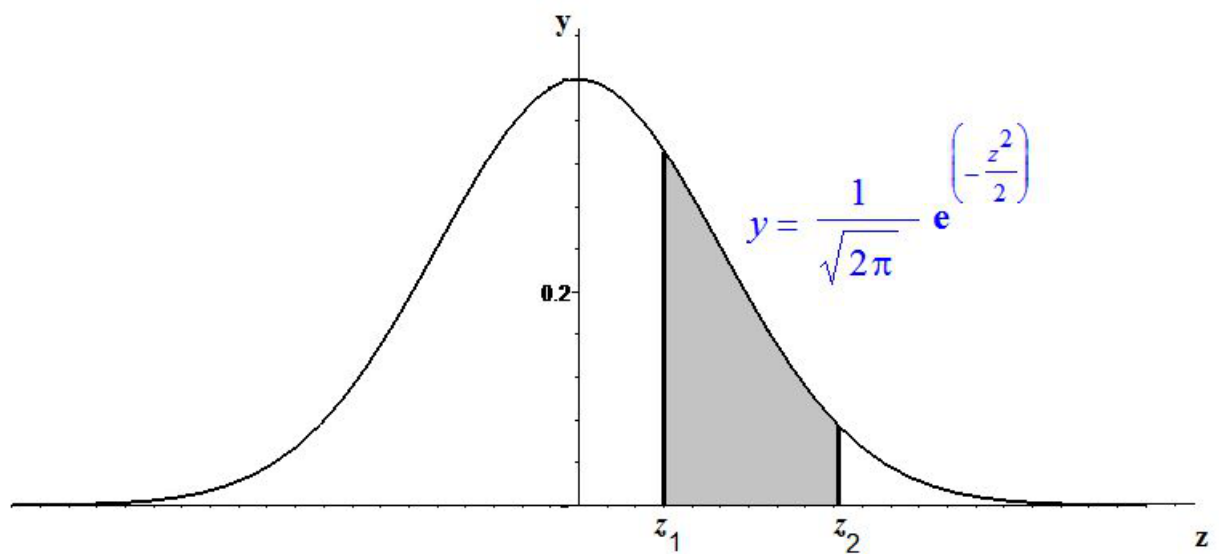
⁴You could write an entire page devoted to the pronunciation of Φ . Actually, someone did: <http://www.foundalis.com/lan/phipro.htm> As an alternative, there is an 11 second Youtube video <http://www.youtube.com/watch?v=ENZub07ORwY> If you watch it, make sure you get beyond the five second mark.

⁵Readers without a background in calculus can, without any qualms, skip over this discussion, including the material involving the error function. We will not make any use of these facts.

It is not difficult to show that the relationship between Φ and the error function is given by

$$\Phi(z) = \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{z}{\sqrt{2}} \right) \right),$$

Although the error function has an important role to play in science and engineering, that is the last you will see of it in these notes. Let us return our attention to Φ . By definition, $\Phi(z)$ gives us the area of a left tail. What if we do not want the area of an entire tail? What if two numbers z_1 and z_2 satisfy $z_1 < z_2$ and we want the area of the bounded region to the right of z_1 and to the left of z_2 ? No problem! We start with $\Phi(z_2)$ for the area of the region to the left of z_2 . This includes the region to the left of z_1 , which we don't want included. We must therefore subtract the area $\Phi(z_1)$ of the unwanted region. Thus, $\Phi(z_2) - \Phi(z_1)$ is the area of the region between z_1 and z_2 . See Figure 3.1.3.



The area of the shaded region is $\Phi(z_2) - \Phi(z_1)$

Figure 3.1.3 Area Under the Bell Curve Between Two Vertical Line Segments

Two Useful Identities, One Useful Simplification, and A Crucial Fact

Using techniques from Calculus, the value of $\Phi(z)$ can be calculated accurately for any value of z . Here is a table for $\Phi(1)$, $\Phi(2)$, \dots , $\Phi(10)$:

n	$\Phi(n)$
1	0.84134474606854294860
2	0.97724986805182079280
3	0.99865010196836990545
4	0.99996832875816688005
5	0.99999971334842812080
6	0.9999999901341235500
7	0.9999999999872018745
8	0.9999999999999937790
9	0.999999999999999985
10	1.000000000000000000

As can be seen from the table, $\Phi(n)$ increases to 1 as n increases to $+\infty$. The fact that the difference between $\Phi(4)$ and 1 shows up in a decimal place we hardly care about is another reflection of our previous assertion that we will have little interest in the Bell Curve for abscissas z with $|z| > 4$. The value of $\Phi(10)$ is not exactly 1, but the amount by which it falls short is so small that the deficit shows up beyond the 19th decimal place. A common calculation in a third semester calculus class is that

$$\lim_{z \rightarrow \infty} \Phi(z) = 1.$$

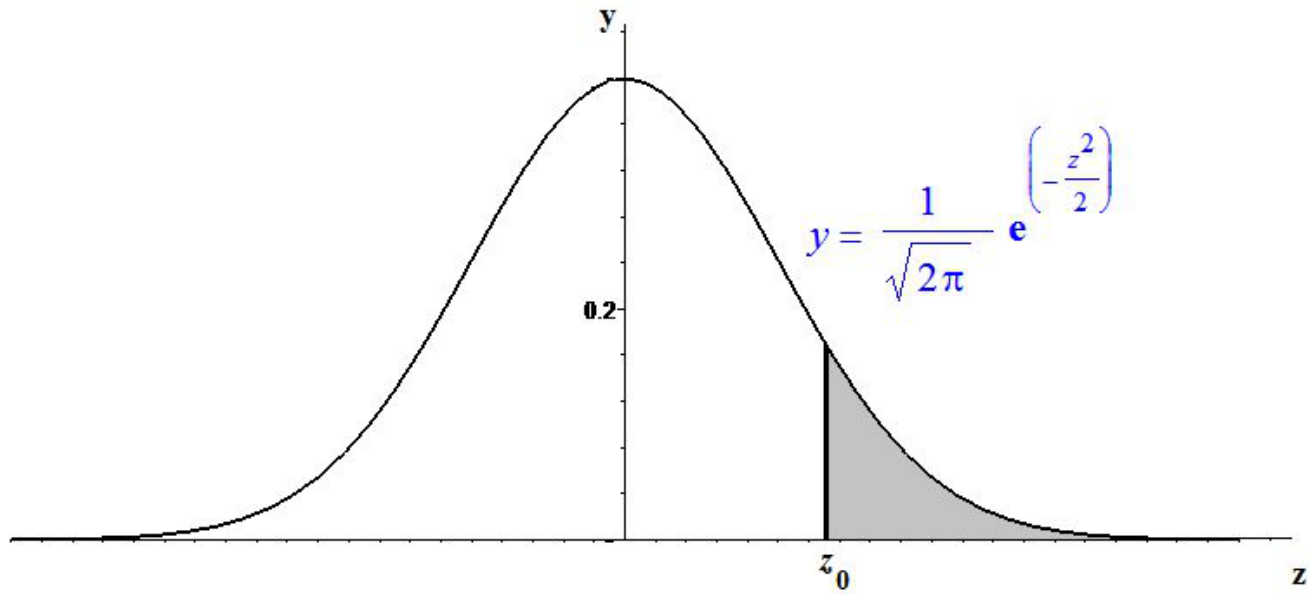
In other words, *The area under the Bell Curve is exactly 1.*⁶

As we have already emphasized, the Φ function is designed to give the area of a left tail. What if we want the area of the tail that is to the *right* of z_0 ? No Problem! Because the area of the entire region under the Bell Curve is 1, the sum of the area to the left of z_0 , namely $\Phi(z_0)$, and the area to the right of z_0 equals 1. Therefore,

$$\text{The area of the region to the right of } z_0 \text{ is } 1 - \Phi(z_0).$$

See Figure 3.1.4.

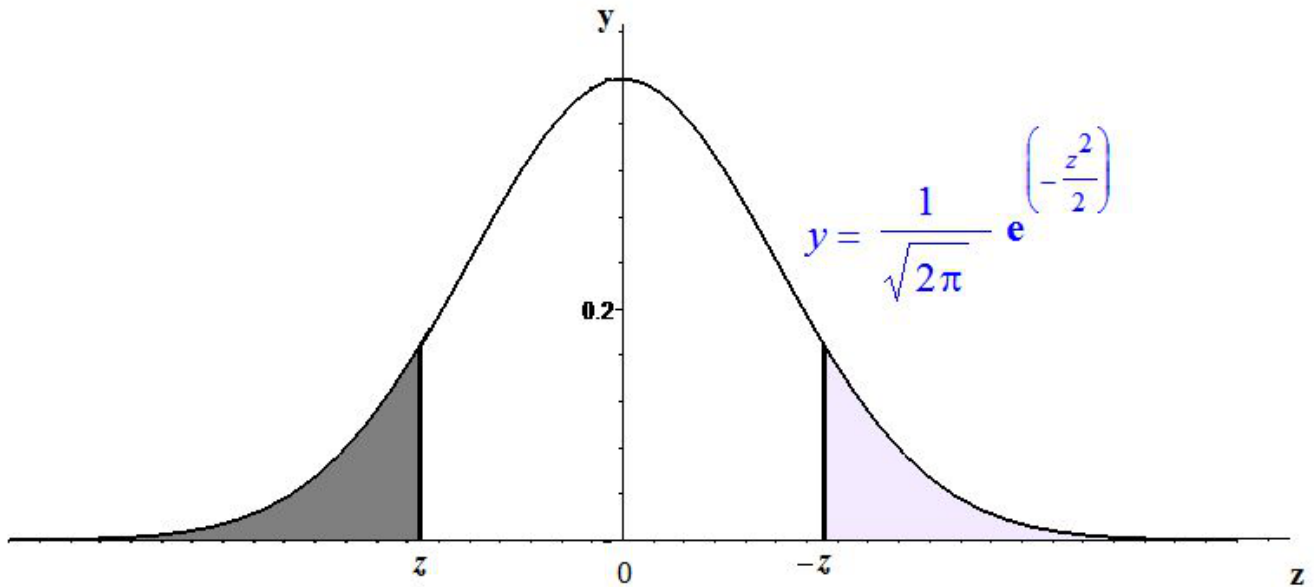
⁶This fact was first proved by the French mathematician Pierre-Simon de Laplace in 1782.



The area of the shaded region is $1 - \Phi(z_0)$

Figure 3.1.4. Area Under the Bell Curve to the right of z_0

Because of the symmetry of the Bell Curve, we can express information about the left half of the curve (the part of the curve where the values of z along the horizontal axis are negative) in terms of information about the right half, where the values of z are positive. First observe that if $z < 0$, then $-z > 0$. By symmetry, the area of the region to the left of $-z$, which is darkly shaded in Figure 3.1.5, is the same as the area to the right of $-z$, which is lightly shaded in Figure 3.1.5.



The area of the darkly shaded region is $1 - \Phi(-z)$

Figure 3.1.5. Area Under the Bell Curve to the left of z with $z < 0$

In view of the formula that we just deduced for the area of a right tail, we conclude that

$$\Phi(z) = 1 - \Phi(-z). \quad (3.1.2)$$

Although our discussion leading up to equation (3.1.2) assumed that $z < 0$, by rewriting the equation as $\Phi(z) + \Phi(-z) = 1$, we realize it is symmetric in z and $-z$. Therefore equation (3.1.2) is valid for all z , positive or negative or zero. We will put equation (3.1.2) to good use in the next subsection. One application that we will consider immediately is a special case of the formula $\Phi(z_2) - \Phi(z_1)$ for the area of the region to the right of z_1 and to the left of z_2 . Let z_0 be any positive number and set $z_1 = -z_0$ and $z_2 = z_0$. Then the area between $-z_0$ and z_0 is equal to $\Phi(z_0) - \Phi(-z_0)$, or, in view of formula (3.1.2), $\Phi(z_0) - (1 - \Phi(z_0))$, or $2\Phi(z_0) - 1$. See Figure 3.1.6.

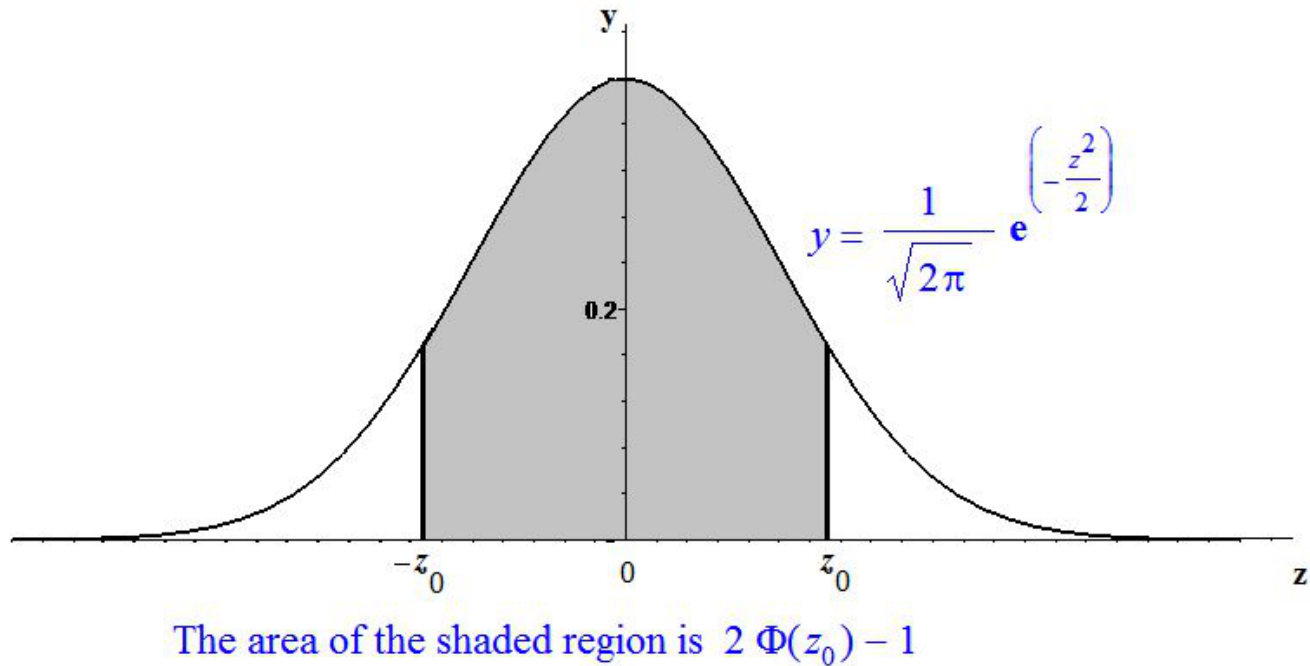


Figure 3.1.6. Area Under the Bell Curve within z_0 units of the central axis.

In words,

The area of the region that is within z_0 units of the central axis (on either side) is equal to $2 \Phi(z_0) - 1$.

The Standard Normal Distribution

Up to this point, we have been referring to areas under the Bell Curve. Therefore, we have been dealing with fractions between 0 and 1. If we multiply such a fraction by 100, then we obtain a number between 0 and 100. Such a number has meaning as a percentage. We can use that idea to define a certain type of distribution: If X is a numerical variable, and if, for any number z , the percentage of values of X that are less than z is equal to $100 \Phi(z)$, then we say that X has a **standard normal** distribution. In this context we refer to Φ as a **cumulative distribution function**, or CDF. Specifically, Φ is the **standard normal cumulative distribution function**. The reason for the term “cumulative” is that $100 \Phi(z)$ does not represent the percentage of values in one bin: instead, it tells us about the accumulation of all values less than or equal to z . That said, we can, by repeating a geometric idea already used for areas, represent the percentage of values in the bin $[z_0, z_1)$ as a difference of two values of the Φ function. To be specific, by subtracting the area of a left tail from that of a larger left tail, we see that, for any pair of numbers z_0 and z_1 with $z_0 < z_1$, the percentage of values of a standard normal variable X that lie in the interval $[z_0, z_1)$ is equal to $100 (\Phi(z_1) - \Phi(z_0))$.

Obtaining Numerical Values for Areas of Regions Under the Bell Curve

It is one thing to use the Φ function to obtain formulas for areas of regions under the Bell Curve, but how do we turn these formulas into numbers? The answer is that there are three ways: calculators with statistical

functions, software programs such as R, and tables.⁷ Our focus in these notes will be on the low-tech method in which tables are employed. To that end, Figure 3.1.7 shows the table of values of $\Phi(z)$ that we will use.

⁷The importance of the Φ function was first recognized by Laplace. In a publication of 1783, he suggested the value of tabulating $\Phi(x)$.

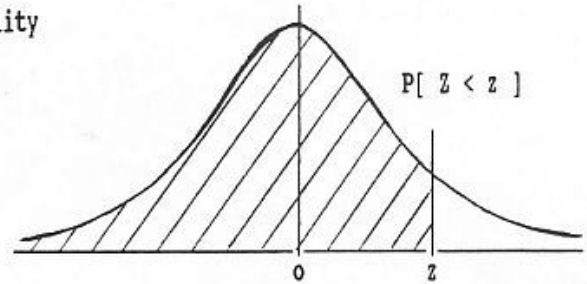
STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z

i.e.

$$P[Z < z] = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7020	0.7054	0.7089	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

Figure 3.1.7 Table of $\Phi(z)$ for Nonnegative z

Look at the top left corner of the table. You will see the letter “z”. The column under that z and the row to the right of it should be thought of as margins. The numbers not found in one of these margins are all values of Φ . By contrast, the numbers in the margins combine to form values of z with two decimal places. The detail of Figure 3.1.7, shown below, illustrates what we mean.

z	0.00	0.01	0.02	0.03
0.0	0.5000	0.5040	0.5080	0.5120
0.1	0.5398	0.5438	0.5478	0.5517
0.2	0.5793	0.5832	0.5871	0.5910
0.3	0.6179	0.6217	0.6255	0.6293
0.4	0.6554	0.6591	0.6628	0.6664

Figure 3.1.7 (detail) Top Left Corner of $\Phi(z)$ Table

In the detail, look at the cell entry highlighted in blue. It lies in the row beginning with marginal entry 0.3 and in the column headed by marginal entry 0.02. Together, these marginal entries sum to give $z = 0.32$. The highlighted cell entry, 0.6255, is the Φ -value corresponding to this value of z . That is, $\Phi(0.32) = 0.6255$.

Notice that Table 3.1.7 does not provide any values of $\Phi(z)$ for negative z . That is because equation (3.1.2) tells us that we do not need to have these numbers tabulated: if we need to know $\Phi(z)$ for z negative, then we use $1 - \Phi(-z)$, which can be determined from the Φ -table because $-z$ is positive. Look at Table 3.1.7 again. This time focus on the Φ -values. You will see that they range from 0.5 in the top left corner to 1 in the bottom right corner. Because of symmetry, half the area under the Bell Curve is to the left of $z = 0$. The equation $\Phi(0) = 0.5$ is a result of this observation. As z increases from 0, the area under the Bell Curve and over the interval $(-\infty, z]$ increases as well. As z increases from 0 to ∞ , the value of $\Phi(z)$ increases from 0.5 to 1. Many of these values are tabulated in Figure 3.1.7. The value $\Phi(3.90)$ is actually 0.9999519... but this value rounds to 1.0000 when we carry only four decimal places. For all values of z greater than 3.90, the values of $\Phi(z)$ also round to 1.0000, so there is no point in displaying them. Besides the values of Φ between 0.5 and 1, the Φ function also takes on all values between 0 and 0.5, but these occur for $\Phi(z)$ with z negative. That is the range of z -values that are not tabulated.

The following example will illustrate how to use the Φ -table to calculate $\Phi(z)$ for negative values of z .

Example 1. Direct Lookup What is the area of the tail of the Bell Curve that lies to the left of $z = -0.84$?

Solution. The answer is $\Phi(-0.84)$, but we require a numerical evaluation of this expression. However, negative values of z are not in Table 3.1.7, so we use formula (3.1.2): $\Phi(-0.84) = 1 - \Phi(0.84)$. We will obtain the answer to the problem once we calculate $\Phi(0.84)$. In Table 3.1.7, scan down the first column, which is headed by z and which consists of z -values. Stop at 0.8. In the row beginning with 0.8, locate the entry, 0.9775, under 0.04. This entry is the Φ -value of 0.84: $\Phi(0.84) = 0.7995$. Thus, the area of the tail of the Bell Curve that lies to the left of $z = -0.84$ is $1 - 0.7995$, or 0.2005. (Calculators and software, it must be admitted, provide greater accuracy. R gives 0.2004542 for the value of $\Phi(-0.84)$, and Maple gives 0.2004541933. Even greater accuracy is available from these software packages by overriding their default digits settings.)

We described the preceding example as a direct lookup because we had only to pluck out one value from the table. The next example illustrates the technique of **interpolation**, in which $\Phi(z)$ is approximated for a value z that is between two z -values of the table.

Example 2. Interpolation What is the area of the tail of the Bell Curve that lies to the left of $z = -0.843$? To the left of $z = -0.8437$?

Solution. Proceeding as in the preceding example, we see that the requested area is $1 - \Phi(0.843)$. The z -value 0.84 is too little. That means that the Φ -value $\Phi(0.84)$ will be too little. The next z -value in the table, 0.85, is too big. That means that the Φ -value $\Phi(0.85)$ will be too big. We want a Φ -value that is just right. The problem is, we have run out of z -values in the table. The solution is to use a simple arithmetic process called interpolation. We need one more digit in the z -value so let us divide the gap between 0.84 and 0.85 into 10 units. We want to add 3 of those units, which constitutes 3 tenths of the gap, to the z -value 0.84 to obtain 0.843. Now the gap between the Φ -values that correspond to 0.84 and 0.85 is $\Phi(0.85) - \Phi(0.84)$, or $0.8023 - 0.7995$, or 0.0028. We make the approximation

$$\Phi(0.843) = \Phi\left(0.84 + \frac{3}{10} \times 0.01\right) \approx \Phi(0.84) + \frac{3}{10} \times 0.0028 = 0.7995 + 0.00084 = 0.80034.$$

For comparison, the value of $\Phi(0.843)$ obtained by software, accurate to a passel of decimal places, is 0.8003857780. Our error of approximation is a mere 0.0000457780, which means that it is as accurate as the numbers found in the table. The answer to the first question is, The area of the tail of the Bell Curve that lies to the left of $z = -0.843$ is $1 - 0.80034$, or 0.19966.

For the z -value -0.8437, we change the preceding calculation in only one way. Because we have 0.8437 instead of 0.84 or 0.85, we need two more digits in the z -value, so let us divide the gap between 0.84 and 0.85 into 100 units. We want to add 37 of those units, which constitutes 37 hundredths of the gap, to the z -value 0.84 to obtain 0.8437. As in the first calculation, the gap between the Φ -values that correspond to 0.84 and 0.85 is $\Phi(0.85) - \Phi(0.84)$, or $0.8023 - 0.7995$, or 0.0028. So our approximation will be

$$\Phi(0.8437) = \Phi\left(0.84 + \frac{37}{100} \times 0.01\right) \approx \Phi(0.84) + \frac{37}{100} \times 0.0028 = 0.7995 + 0.001036 = 0.800536.$$

The area of the tail of the Bell Curve that lies to the left of $z = -0.8437$ is $1 - 0.800536$, or 0.199464. Using software, we find that the requested Φ -value, $\Phi(-0.8437)$, accurate to a scad of decimal places is, 0.1994185338. We are satisfied with our approximative interpolations.

Who Phoned Me? – Reverse Lookups

Table 3.1.7 has z -values and Φ -values. The direct way of using the table is to start with a z -value z and look up the corresponding Φ -value, $\Phi(z)$. Often, however, we must do the reverse: starting with a given Φ -value p , we must find the corresponding z -value z . Mathematically, that task amounts to solving the equation $\Phi(z) = p$ for z , given a value p .

It might be useful to first consider reverse lookups from a more general perspective. Suppose that S and T are sets of numbers, and that f is a function such that $t = f(s)$ is a number in T for every number s in S . An example would be $S = [4, 9]$, $T = [2, 3]$, and $f(s) = \sqrt{s}$. Suppose also that, for every t in T there is a unique s in S for which $f(s) = t$. In other words, for every t in T , the equation $f(s) = t$ can be solved, and there is exactly one solution s in S . In this case we say the function $f : S \rightarrow T$ is invertible, and, for a given t in T , we write the unique solution s of the equation $f(s) = t$ as $s = f^{-1}(t)$. We call $f^{-1} : T \rightarrow S$ the **inverse function** of f . In our illustration, we solve the equation $\sqrt{s} = f(s) = t$ as $s = t^2$, which tells us that $f^{-1}(t) = t^2$.

In the example that is our real interest, we take the intervals $S = (-\infty, \infty)$ and $T = (0, 1)$ to be our sets of numbers and $\Phi : S \rightarrow T$ to be the function that maps S onto T . For every value p in the interval $T = (0, 1)$, there is a unique z in the interval $S = (-\infty, \infty)$ for which $\Phi(z) = p$. In accordance with the general theory of invertible functions, we write $z = \Phi^{-1}(p)$.

Just as the function Φ has a name, the standard normal cumulative distribution function, its inverse function Φ^{-1} has a name, the **standard normal quantile** function. Unlike our toy illustrative example, $f(s) = \sqrt{s}$ in which $f^{-1}(t) = t^2$, we cannot use algebra to find a formula for $z = \Phi^{-1}(p)$, given a value of p .

Instead, we will use Table 3.1.7 in a reverse direction. This approach is analogous to a graphical procedure for understanding the inverse function. For our toy, illustrative example, we have plotted the graph of $f(s) = \sqrt{s}$ in the left panel of Figure 3.1.8. We could have also plotted the graph of the inverse function $f^{-1}(t) = t^2$ had we wanted, but we did not because it was not necessary. The right panel shows that all we have to do to get f^{-1} is reverse the arrows of the left panel.

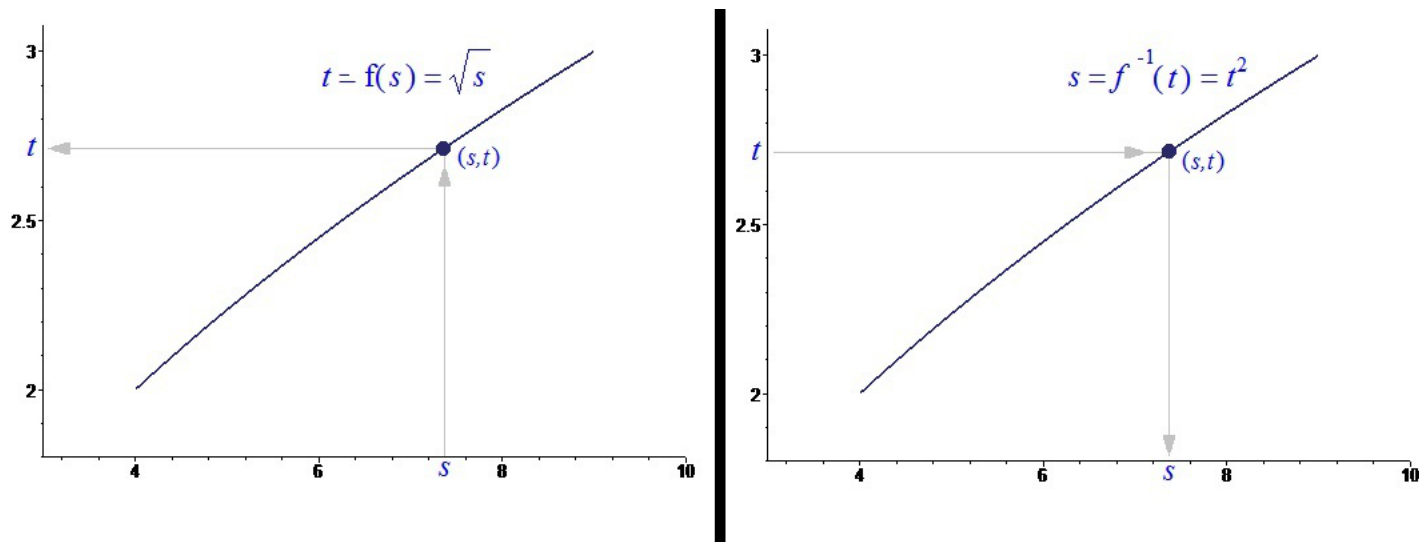


Figure 3.1.8 The Graphical interpretation of an Inverse Function

Using a table in reverse is similar to using a graph in reverse. We simply reverse direction of the arrows. Instead of locating a z -value in the margins and then identifying the corresponding Φ -value in the table, we use a tabulated Φ -value to identify the corresponding z -value in the margins. The left panel of Figure 3.1.9 shows the direct lookup of $0.6255 = \Phi(0.32)$, whereas the right panel shows the reverse lookup of $0.32 = \Phi^{-1}(0.6255)$.

z	0.00	0.01	0.02	0.03
0.0	0.5000	0.5040		0.5120
0.1	0.5398	0.5438		0.5517
0.2	0.5793	0.5832		0.5910
0.3			0.6255	0.6293
0.4	0.6554	0.6591	0.6628	0.6664

$0.6255 = \Phi(0.32)$

z	0.00	0.01	0.02	0.03
0.0	0.5000	0.5040		0.5120
0.1	0.5398	0.5438		0.5517
0.2	0.5793	0.5832		0.5910
0.3			0.6255	0.6293
0.4	0.6554	0.6591	0.6628	0.6664

$0.32 = \Phi^{-1}(0.6255)$

Figure 3.1.9 Reverse Lookup of $z = \Phi^{-1}(p)$

Example 3. For what value of z is the area of the tail of the Bell Curve to the left of z equal to 0.9015?

Solution. First thing to notice: we have been given the value, 0.9015, of an area. In other words, we have been bestowed with a Φ -value, not a z -value. To be specific, we are to find z such that $\Phi(z) = 0.9015$. In this case we are able to find the Φ -value 0.9015 in the table: it is in the row for 1.2 and in the column for 0.09. Therefore, $\Phi(1.29) = 0.9015$, or, equivalently, $1.29 = \Phi^{-1}(0.9015)$.

When we use the Φ table for reverse look-ups, we do not expect to have such an easy time as in the preceding example. There are gaps between entries of the table, and, as a general rule, a specified Φ -value will fall in one of the gaps. That is what happens in the next example.

Example 4. For what value of z is the area of the tail of the Bell Curve to the right of z equal to 0.10?

Solution. If the area to the right of a value of z is 0.10, then the area to the left of z , that is, the area of the left tail terminating at z , is $1 - 0.10$, or 0.90. Before continuing with the calculation, let us pause to consider the step that has been taken. Table 3.1.7 of Φ -values is all about areas of left tails. If we are given information about a right tail that is to the right of z , then we must convert that information to equivalent information about the left tail on the other side of z . To do that, we subtract the given area of the right tail from 1. Pay attention to the following point because it may prevent a common error: We must often subtract an area or Φ -value from 1, but we never have to subtract a z -value from 1: if you ever find yourself calculating $1 - z$, then stop and reflect on your objective.

Let us resume. We seek z such that $\Phi(z) = 0.9$. We scan the entries of Table 3.1.7, but the Φ -value 0.9 does not appear. The closest two are $\Phi(1.28) = 0.8997$ and $\Phi(1.29) = 0.9015$. Because 0.9 is between 0.8997 and 0.9015, we need a value of z between 1.28 and 1.29. Interpolation time! We calculate $9000 - 8997 = 3$ and $9015 - 8997 = 18$. The Φ value we have been given is 3 eighteenths of the way between $\Phi(1.28)$ and $\Phi(1.29)$. Remember that we are seeking a z -value. The gap between the two z -values 1.28 and 1.29 is 0.01. So we make the approximation

$$\Phi\left(1.28 + \frac{3}{18} \times \underbrace{0.01}_{\text{gap in } z\text{-values}}\right) \approx \Phi(1.28) + \frac{3}{18} \times \underbrace{(0.9015 - 0.8997)}_{\text{gap in } \Phi\text{-values}} = 0.8997 + \frac{3}{18} \times 0.0018 = 0.9.$$

In other words, the z -value we are seeking is $1.28 + \frac{3}{18} \times 0.01$, or 1.2817. The actual value, correct to gobs of decimal places, is 1.281551566. The error of our approximation is a scant 0.00015.

Example 5. For what value of z is the area of the tail of the Bell Curve to the left of z equal to 0.10?

Solution. Here we are asked for the area of a left tail, and that is what the Φ function is all about. However, the table that we have provides Φ -values that are greater or equal to 0.5, and, alas, we have been presented with the Φ -value 0.1, which is too small to be in the table. So we must adjust. The left tail with area 0.1 terminates on the right at a negative value. Let us call that value $-z$, where z is positive. The area of the left tail terminating at the positive number z is $1 - 0.1$, or 0.9. So we solve $\Phi(z) = 0.9$. In fact, we did that in the preceding example and found $z = 1.2817$. Consequently, the answer to our current problem is $z = -1.2817$.

The Empirical Rule, also known as The 68-95-99.7 Rule

In this subsection we will study the special values $\Phi(\pm 1)$, $\Phi(\pm 2)$, and $\Phi(\pm 3)$. Additionally, we will look at $\Phi(\pm 2/3)$.

Example 6. What is the area of the left tail of the Bell Curve that terminates on the right at $z = -1$?

Solution. The answer is $\Phi(-1)$, which equals $1 - \Phi(1)$ by equation (3.1.2). From the table, we see that $\Phi(1) = 0.8413$. Therefore, the answer we seek is

$$\Phi(-1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587.$$

Example 7. What is the area of symmetric region under the Bell Curve that is within 1 unit of the central axis? Within 2 units? Within 3 units?

Solution. According to the discussion in the preceding section, the answer is $2\Phi(1) - 1$, or $2 \times 0.8413 - 1$, or 0.6826. Similarly, the area within 2 units is $2\Phi(2) - 1$, or $2 \times 0.9773 - 1$, or 0.9546. Finally, the area within 3 units is $2\Phi(3) - 1$, or $2 \times 0.9986 - 1$, or 0.9972.

The three answers of the preceding question are often rounded down to 0.68, 0.95, and 0.997. As a result, we obtain the **Empirical Rule**, also known as the **68-95-99.7 Rule**:

About 68% of the area under the Bell Curve is within 1 unit of the central axis, about 95% of the area is within 2 units of the central axis, and about 99.7% of the area is within three units of the central axis.

Example 8. For what value z is 25% of the area under the Bell Curve to the left of z ?

Solution. Because $\Phi(0) = 0.5$, we see that the z we seek is negative. Instead, we will find the positive value of z such that the right tail beginning at z has area 0.25. In other words, we want to identify the z such that $\Phi(z) = 0.75$. The entry in the row for 0.6 and the column for 0.07 is 0.7486. That means, $\Phi(0.67) = 0.7486$. Looking at the next entry to the right, we see that $\Phi(0.68) = 0.7517$. We want the value, 0.7500 in between. This value is $(0.7500 - 0.7486)/(0.7517 - 0.7486)$, or $14/31$ of the distance of 0.7486 to 0.7517. So we set $z = 0.67 + 14/3100$ (because adding $14/3100$ brings us $14/31$ of the way from 0.67 to 0.68). Thus, we obtain $z = 0.6745$. The answer to the question is therefore -0.6745 .

From the result of the preceding example, we see that the area under the Bell Curve between $z = -0.6745$ and $z = 0.6745$ is 0.5. Because 0.6745 is approximately $2/3$, a fraction that is easy to remember, we have another approximate rule that is similar in spirit to the empirical rule:

50% of the area under the Bell Curve is within $2/3$ of one unit from the central axis. The IQR of the standard normal distribution is therefore about $4/3$.

Normal Distributions, or Repetitious Redundancy?

standard: (adjective) used or accepted as normal or average.

normal: (adjective) conforming to a standard; usual, typical, or expected.

So, if *standard* is *normal* and *normal* is *standard*, then isn't *standard normal* repetitiously redundant? Hold that thought!

Suppose that μ is any number and that σ is a positive number. Given that the area under the Bell Curve is 1, it is not difficult to show that the area under the graph of

$$y = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2\right) \quad -\infty < x < \infty \quad (3.1.3)$$

is also 1. Therefore, for any numbers a and b with $a < b$, the quantity $100 \times \frac{1}{\sqrt{2\pi}\sigma} \int_a^b \exp\left(-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2\right) dx$,

which is 100 times the area of the region that is under the graph of equation (3.1.2) and between the two vertical lines $x = a$ and $x = b$, is a number between 0 and 100—in other words, a number that makes sense as a percentage. If X is a numerical variable, and if for every a and b with $a < b$, the percentage

$$100 \times \frac{1}{\sqrt{2\pi}\sigma} \int_a^b \exp\left(-\frac{1}{2}\left(\frac{(x-\mu)}{\sigma}\right)^2\right) dx$$

is equal to the percentage of values of X between $x = a$ and $x = b$, then we say that X has a **normal** distribution with mean μ and standard deviation σ . We are not using \bar{x} and s on purpose. The numbers μ and σ are called the **parameters** of the distribution. They are not the sample mean and sample standard deviation—no sampling has been done. They are the numbers that refer to an entire population and that satisfy theoretical definitions for mean and standard deviation. A common notation for a variable that is normal with mean μ and standard deviation σ is $N(\mu, \sigma)$. Thus, *the* standard normal distribution, having as it does mean 0 and standard deviation 1, is denoted $N(0, 1)$. To answer the question with which we opened this subsection, the “standard” of *standard normal* refers to mean 0 and standard deviation 1. The “normal” of *standard normal* refers to a distribution whose subpopulation percentages can be expressed as areas under the graph of equation (3.1.3).

The graphs of equation (3.1.3) for three pairs of μ and σ are shown in Figure 3.1.10. You will note that, as the standard deviation increases, the graph becomes flatter, less peaked, and has tails that do not tail off so quickly. When the standard deviation decreases, the graph becomes more peaked about the center, and the tails die away more quickly. These behaviors are in agreement with our interpretation of the standard deviation as a measure of the spread of a distribution from the center.

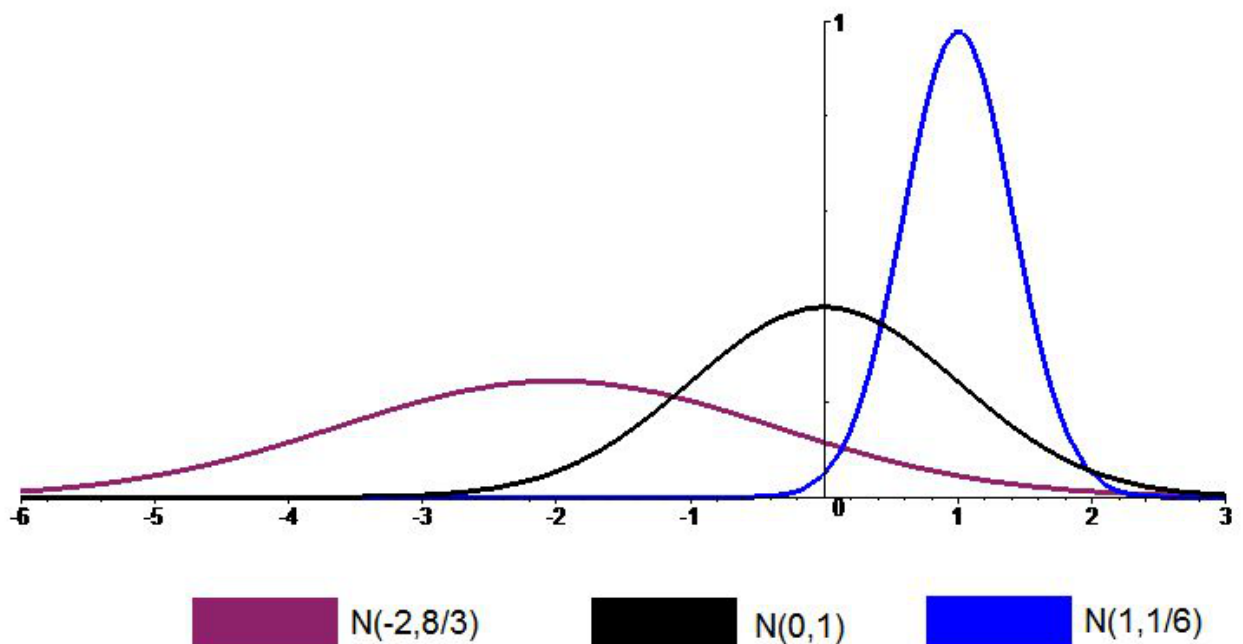


Figure 3.1.10 Curves Giving Rise to Three Normal Distributions

The Standard Normal Distribution in R (Optional)

There are several useful functions in R that concern the normal distribution. Remember that the Bell Curve is the graph of the equation

$$y = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \quad (3.1.2)$$

The right side of equation (3.1.2) is called the *probability density function* of the standard normal distribution. In R, we write equation (3.1.2) as $y = \text{dnorm}(x)$, the d standing for “density.” We will not need the R-function `dnorm` for the sort of question we investigate in these notes, but it may be helpful to use it to

illustrate a plotting capability of R. Consider the R code

```
> x = seq(-4, 4, by = 0.01)
> y = dnorm(x)
> plot(x, y, type = 'l',
+ main = "The Bell Curve (Graph of the Standard Normal Probability Density Function)")
```

The first line creates the sequence $-4.00, -3.99, -3.98, \dots, 3.99, 4.00$ and names it `x`. The second line creates the sequence `dnorm(-4.00), dnorm(-3.99), dnorm(-3.98), ..., dnorm(3.98), dnorm(3.99), dnorm(4.00)` and names it `y`. The third, split line plots the ordered pairs (ξ, η) for every value ξ in `x` and corresponding $\eta = \text{dnorm}(\xi)$ in `y`. The first optional argument of `plot` calls for the plot to consist of line segments joining consecutive plotted points. The second optional argument, following R's line continuation character (`+`), adds a title to the plot. The result is Figure 3.1.11.

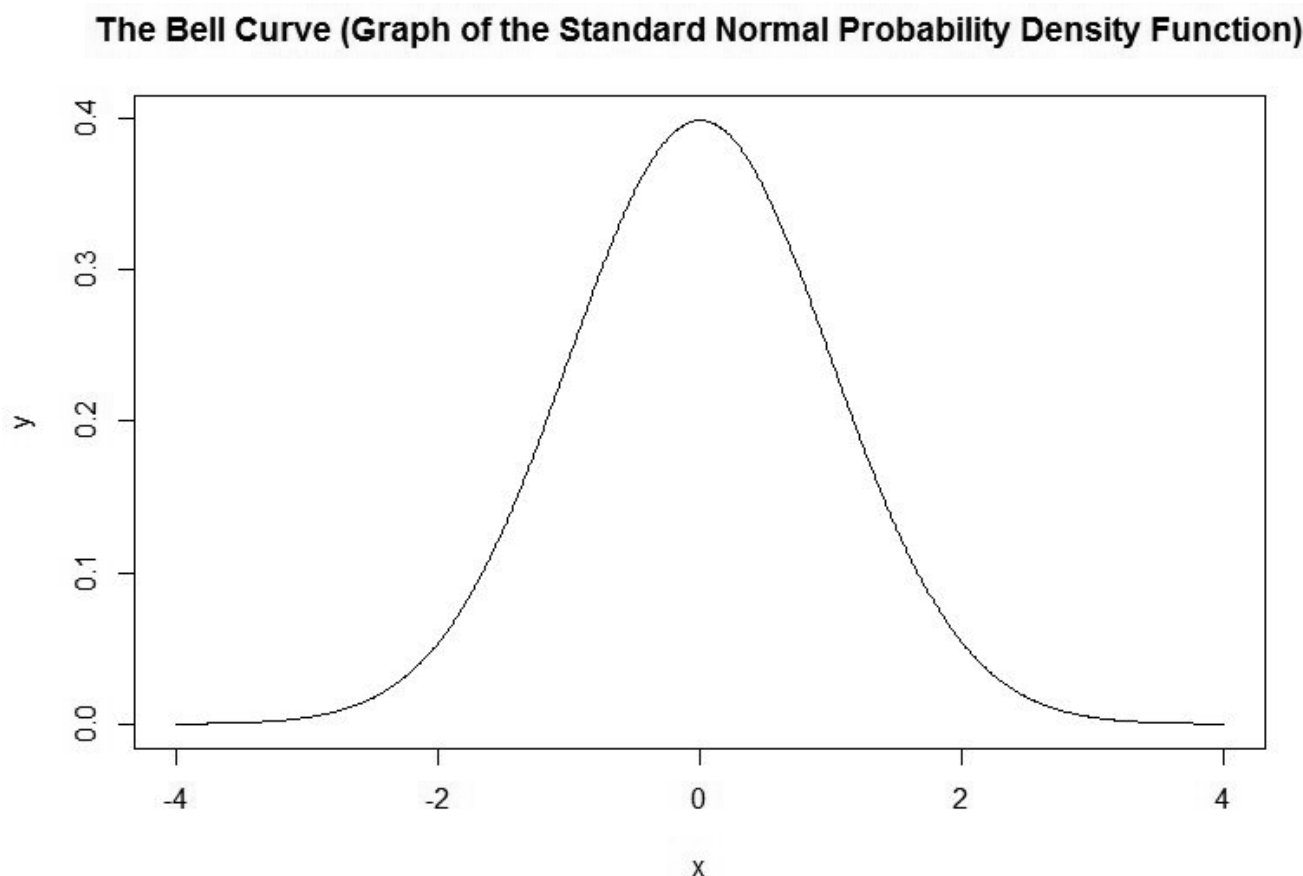


Figure 3.1.11 The Standard Normal Probability Density Function

An R function that is extremely useful for many of the calculations needed in these notes is `pnorm`. Think of the “p” before “`pnorm`” as standing for “probability” or “percentage”, which is what you obtain when you multiply a probability by 100, or “Phi”. In R, the value of $\Phi(x)$ is obtained by the call `pnorm(x)`. Thus, `pnorm(2) - pnorm(-2)` will result in the value 0.9544997, which is a little more accurate than the approxi-

mation 0.95 of the Empirical Rule. The code

```
> x = seq(-4, 4, by = 0.01)
> y = pnorm(x)
> plot(x, y, type = 'l',
+ main = "The Standard Normal Cumulative Distribution Function: y = Phi(x)")
```

results in the plot of the equation $y = \Phi(x)$ shown in Figure 3.1.12.

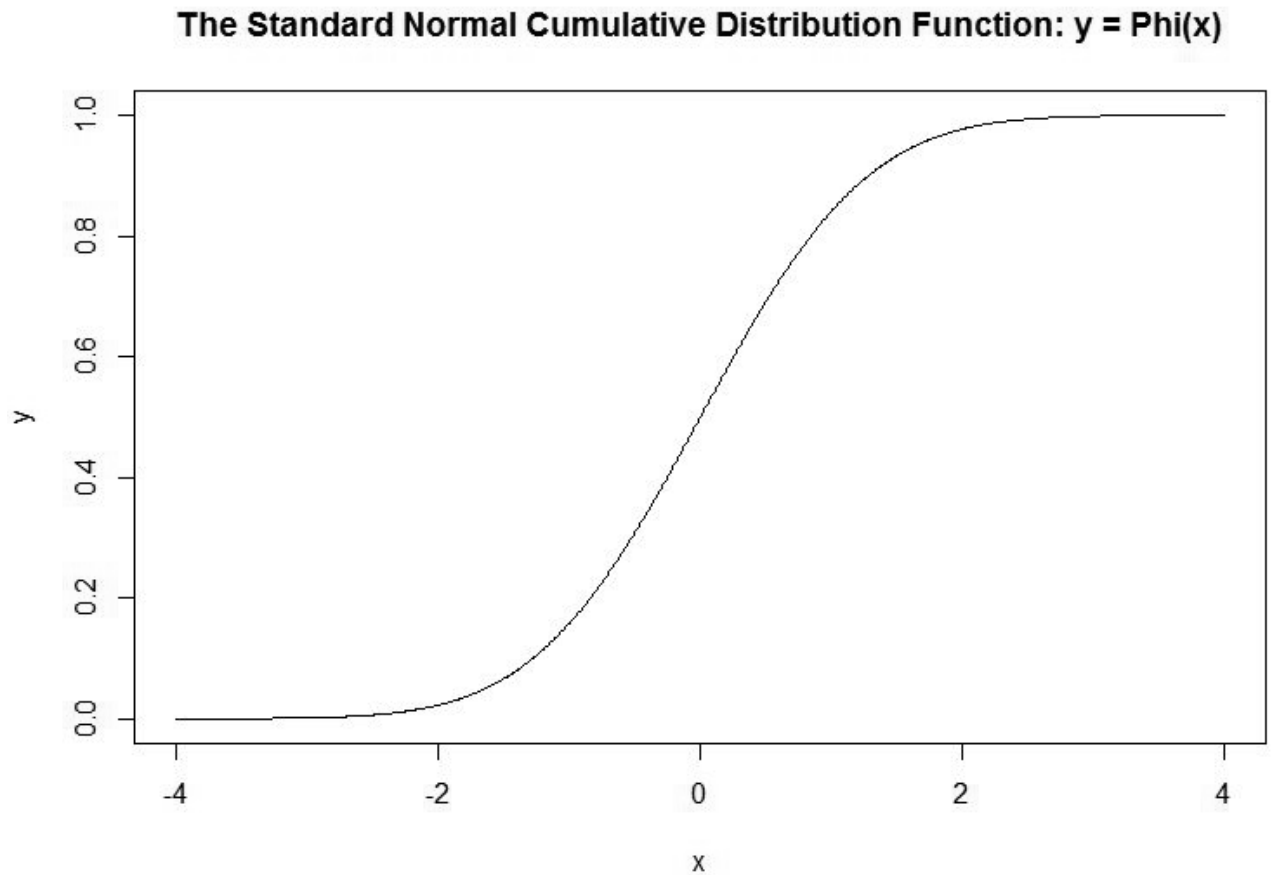


Figure 3.1.12 Plot of $y = \Phi(x)$

From the plot we see that $\Phi(0) = 0.5$, that $\lim_{z \rightarrow -\infty} \Phi(z) = 0$, and $\lim_{z \rightarrow \infty} \Phi(z) = 1$.

A useful feature of `pnorm` is that it can be applied not only to single values but also to all members of a list. For example, in order to derive the Empirical Rule, we calculated $\Phi(1)$, $\Phi(2)$, and $\Phi(3)$ one at a time. The R code

```
> X = c(1, 2, 3)
> P = pnorm(X)
> P
```

creates a list `X` with entries 1, 2, 3, calculates the Φ -value of each of the three entries of `X`, makes these Φ -values entries of a list named `P`, and then prints `P`: the R output is:

```
[1] 0.8413447 0.9772499 0.9986501
```

Another essential R function is `qnorm`. The “q” prepended to “norm” stands for “quantile.” The call `qnorm(p)` will return the value of z corresponding to the Φ -value `p`. In other words, `qnorm` calculates reverse lookups: for a given value of p , `qnorm(p)` is the value of z such that $\Phi(z) = p$. Thus, if we seek the unique z -value that exceeds precisely 75% of z -values, we call on `qnorm(0.75)` and R returns 0.6744898. If we ask, for what number λ does the interval $[-\lambda, \lambda]$ contain 75% of all observations of a standard normal distribution, then we call on `qnorm(0.875)` and obtain 1.150349 for the value of λ . Of course, our request was really for the number λ such that 87.5% of all observations do not exceed λ . But that means that the right tail of values exceeding λ contains 12.5% of all observations. By symmetry, the tail to the left of $-\lambda$ also contains 12.5% of all observations. In total, for $\lambda = 1.150349$, 25% of all observations fall outside of the z -interval $[\lambda, \lambda]$.

A useful feature of `qnorm` is that it can be applied not only to single values but also to all members of a list. For example, earlier in this subsection, we saw that `X = c(1, 2, 3)` followed by `P = pnorm(X)` creates a list named `P` consisting of the Φ -values of 1, 2, 3. If we follow this with `Q = qnorm(P)`, then `Q` is the list 1, 2, 3, that we started with.

Other Normal Distribution in R (Optional)

The R functions described in the preceding subsection, as described there, pertain to $N(0, 1)$. In fact, those functions can be called with additional parameters so that they apply to general normal distributions. Thus, if m is any real number and s is positive, the call `pnorm(x, m, s)` refers to the fraction of observations from a $N(m, s)$ distribution that are less than x . For example, `pnorm(23, 17, 3) - pnorm(11, 17, 2)` represents the fraction of observations in a $N(17, 2)$ distribution in the interval from 13 to 21. Because 11 and 23 are two standard deviations from the mean of $N(17, 3)$, it is no surprise (if one remembers the Empirical Rule) that R responds with 0.9544997.

Here is an example using `dnorm`. The code

```
> x = seq(-4, 6, by = 0.01)
> y1 = dnorm(x, 4, 0.5)
> y2 = dnorm(x)
> plot(x, y1, type = 'l',
+ main = main = "The Probability Density Functions of N(0,1) and N(4, 0.5)")
> lines(x, y2, col = "purple")
```

produces the plots of $N(0, 1)$ and $N(4, 0.5)$ shown in Figure 3.1.13.

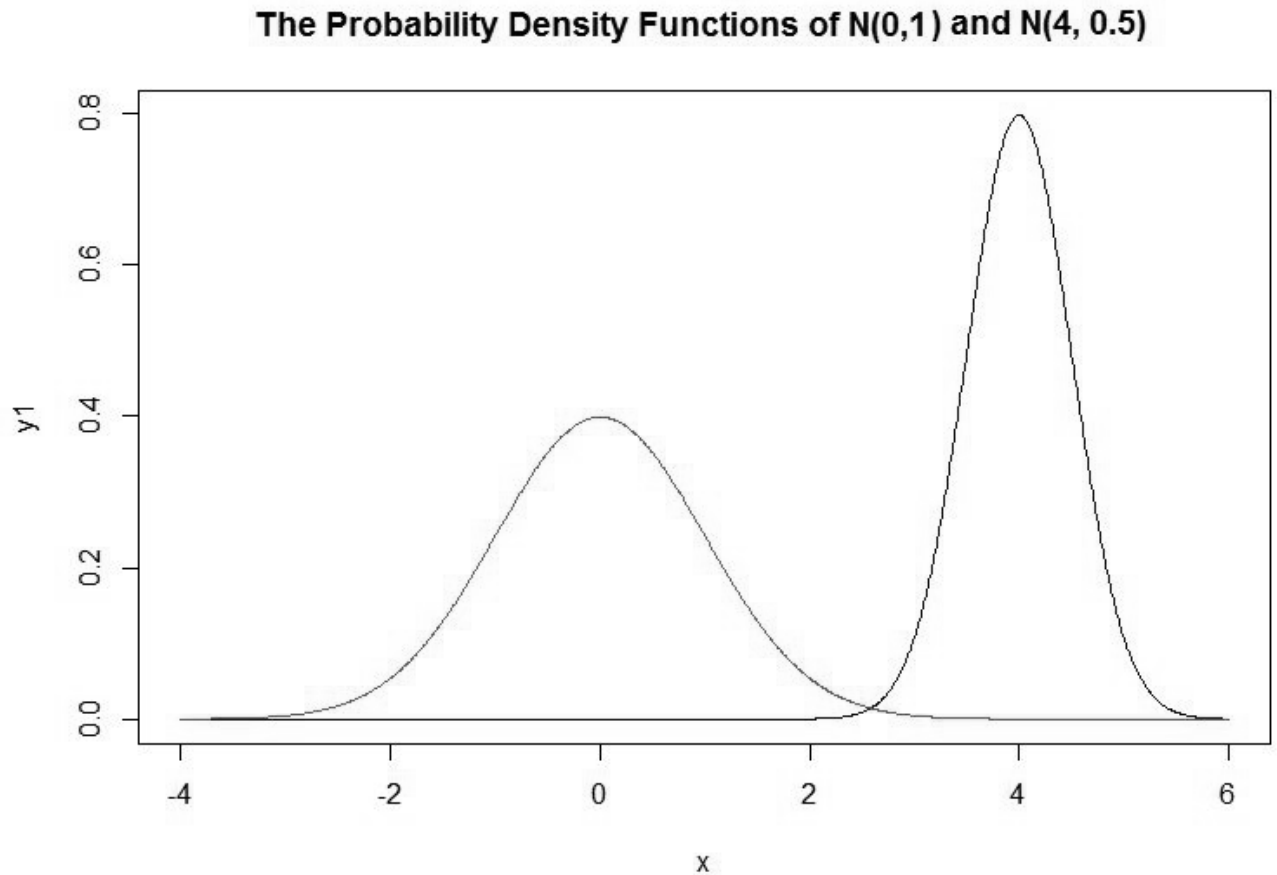


Figure 3.1.13 Plots of Two Normal probability Density Functions—One Standard, One Not

3.2 Standardized Scores

In mid-July, 2011, a massive heat dome settled over Canada. Temperatures were brutal. As CTV News reported, “In Winnipeg, the unrelenting heat hit record-highs on Tuesday, with temperatures hitting 34 C. Next door in Saskatchewan, Regina roasted at 31.9 C.” Eh? What were those unrelenting Winnipeg temperatures expressed in American degrees? And at what Fahrenheit temperature were those weather-oppressed folks from Regina roasting? How can we compare those foreign centigrade measurements to the ones with which we are familiar?

In Figure 3.2.1, we have used two parallel horizontal lines to plot both the centigrade scale and the Fahrenheit scale of temperature measurement.

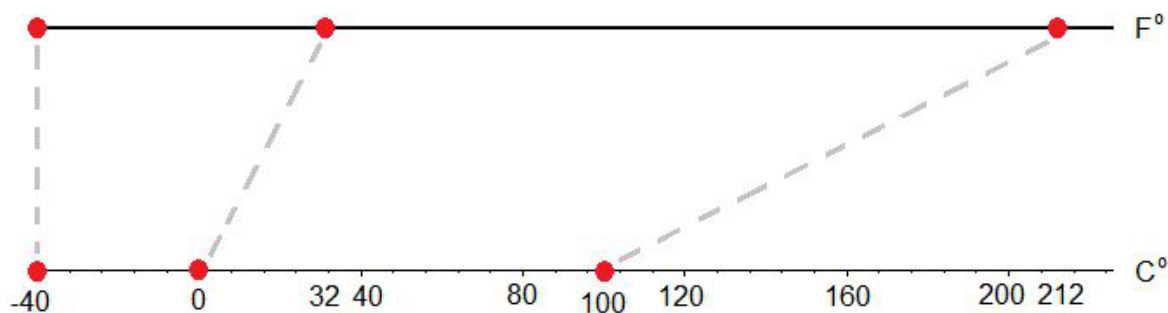


Figure 3.2.1 The Fahrenheit and Centigrade Temperature Scales

The line representing F° is a copy, translated vertically upwards, of the line representing C° . To avoid clutter, we have not drawn tickmarks on the line that represents degrees Fahrenheit but the *number* scale is the same as that of the line below it: each vertical line—one is drawn at -40 —intersects both horizontal lines at the same number. On the other hand, the two *temperature* scales, as you know, are different: equal temperatures are represented by different numbers (with the exception of the temperature that is represented by both -40°C and -40°F).⁸ In Figure 3.2.1, equal temperatures are connected by dotted line segments, but they are *not* vertical (with the exception of the aforementioned unique temperature that has the same numerical measurement in each scale). There are two problems that arise when trying to compare temperatures measured on the two scales, C and F.

The first problem can be seen by considering the temperature at which water freezes, a temperature that is represented by 0 in degrees centigrade but 32 in degrees Fahrenheit.⁹ We can easily fix the shift that results in the freezing point of water being expressed as 0°C and 32°F : we subtract 32°F from a Fahrenheit temperature x to obtain $y = x - 32$. With that shift, or, as it is usually called in mathematics, *translation*, the temperature at which water freezes is the same in both the C and the $Y = F - 32$ scales.

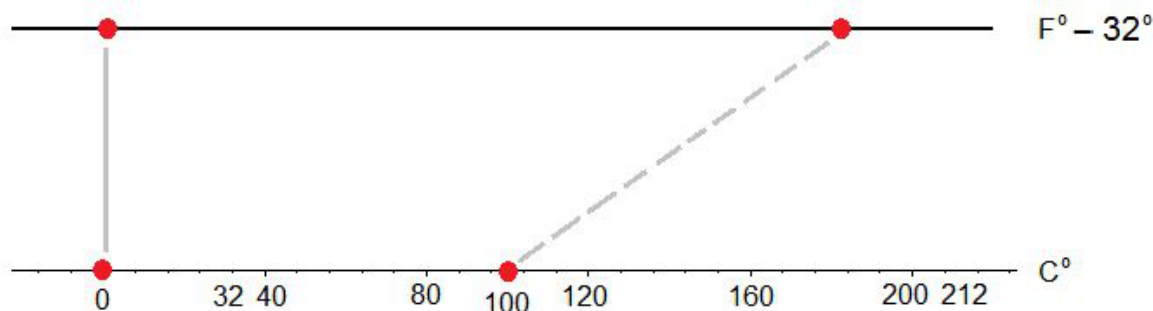


Figure 3.2.2 The Translated (Shifted) Fahrenheit and Centigrade Temperature Scales

⁸That -40°C and -40°F represent the same temperature has no significance for us. For those in Winnipeg, it can be a source of mirth. Winnipeg kid 1: Let's play street hockey tomorrow. It's going to warm up to -40 . Winnipeg kid 2: Fahrenheit or centigrade?

⁹The freezing point of water is used to define the origin of the centigrade scale, a temperature scale based on one introduced by Anders Celsius in 1742. Somewhat earlier, in 1724, Daniel Fahrenheit had used a salt water solution instead. The number 100 represents the boiling point of water in degrees centigrade and the average core temperature of the human body in degrees Fahrenheit. Thanks to the somewhat strange choices of reference points, the freezing and boiling points acquire the somewhat strange values 32 and 212 in degrees Fahrenheit.

The problem of *scale*—the size of one degree—remains. The boiling point of water is at 180°F° but only 100°C°. Fahrenheit degrees are 180/100, or 9/5, as large as centigrade degrees. We can fix this too: we simply divide $F - 32$ by 9/5, obtaining

$$C = \frac{F - 32}{9/5}.$$

We shall see that this idea of translating (shifting) by subtraction and then rescaling by division is a crucial procedure for comparing two distributions.¹⁰

Standardizing Unimodal Symmetric Distributions—An Etruscan Example

Alas, poor Yorick! I knew him, Horatio.

—Hamlet (while gazing into the eye-sockets of Yorick’s skull)

Beginning around 800 BCE but enduring for only a few hundred years, the Etruscan civilization flourished in the region of Italy that now comprises the provinces of Tuscany and Umbria. The origins of the Etruscans are still a puzzle for historians. Recent DNA studies suggest that the Etruscans were an indigenous people. In the years prior to DNA testing, anthropometry was an important archeological tool for investigating such questions. In one anthropometric study, the breadth X of the skull of each of 84 Etruscan skeletons, all males, was measured.¹¹ Rather than report all the data, we will be content with a statistical summary, which will serve our purposes admirably. The mean breadth \bar{x} of X was 143.75 mm with standard deviation s equal to 5.9331 mm. The five number summary was 126.0, 140.0, 143.5, 148.0, 158.0. Figure 3.2.3 shows the boxplot of the data.

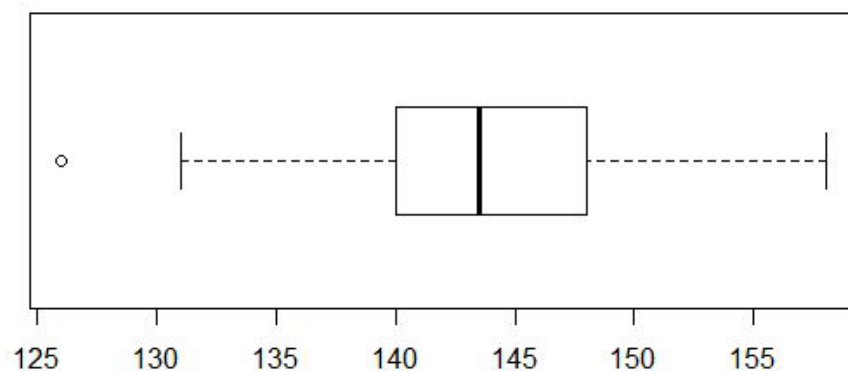


Figure 3.2.3 Boxplot of Etruscan Skull Breadths

The boxplot suggests that the data is approximately symmetric: the median is located centrally in the box and the whiskers are about the same length. There is only one outlier. Furthermore, the mean (143.75) and median (143.5) are very close to each other, indicating little skewness. A histogram (with the outlier excluded) confirms the assessment of symmetry. See Figure 3.2.4

¹⁰For those still wondering, the centigrade-Fahrenheit conversion formula reveals that those unfortunate inhabitants of Regina were roasting at 89.42°F°. What a bummer for those Canadians to be exposed to such a sweltering heat wave. Inhuman, really.

¹¹Barnicot, N.A., and Brothwell, D.R., *The Evaluation of Metrical Data in the Comparison of Ancient and Modern Bones*, in *Medical Biology and Etruscan Origins*, 1959.

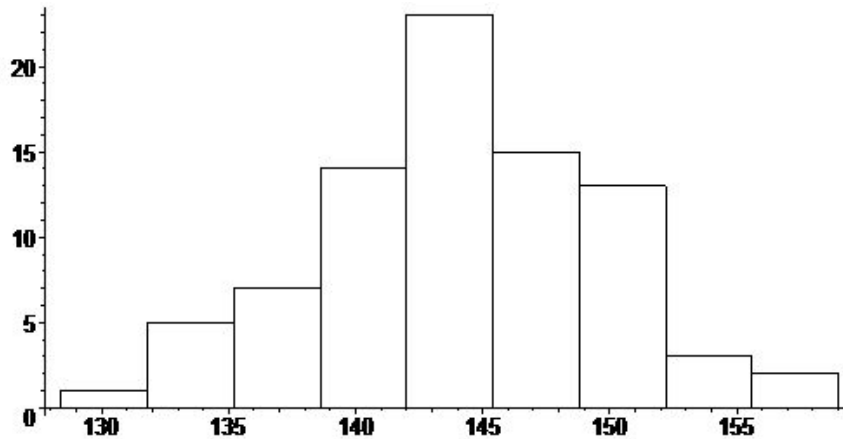


Figure 3.2.4 Histogram of Etruscan Hat Sizes

It is not difficult to imagine a bell-shaped curve superimposed on the data. *The* Bell Curve, however, has its center at 0, not at 143.75. The measure of the spread of the Bell Curve is the standard deviation 1, whereas the skull breadths have standard deviation 5.9331. Following the method used for shifting and rescaling Fahrenheit measures, we translate (shift) the skull breadth measurement X by subtracting the mean 143.75 and then we rescale by dividing by the standard deviation 5.9331. The result is the variable

$$Z = \frac{X - \bar{x}}{s} = \frac{X - 143.75}{5.9331}.$$

By design, the mean of Z is 0 and the standard deviation of Z is 1. The five number summary of Z is -2.99168509, -0.63204615, -0.04213641, 0.71631897, 2.40177535. Recall from Section 3.1 that the IQR for the standard normal distribution is from -0.67 to 0.67. Compare that range with the corresponding range, -0.63204615 to 0.71631897, of Z : our real world old world data set is not far off from the theoretical model. This adds further evidence to support our belief that the transformed data set is approximately standard normal. A histogram of the relative frequencies of Z with the Bell Curve superimposed adds visual evidence. See Figure 3.2.5.

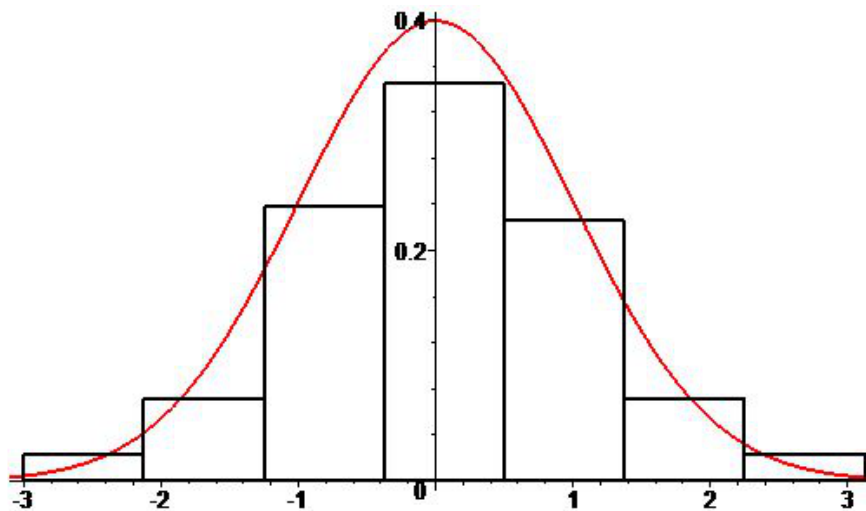


Figure 3.2.5 Relative Frequency Histogram of Standardized Etruscan Skull Breadths

Remember that we are not asserting that Z has a standard normal, only that its distribution can be approximated, to a reasonably good degree, by a standard normal distribution. Furthermore, the 84 skulls under examination constitute one smallish sample from a much larger population—another sample would not be expected to be an exact match of the data we have studied. In this study, the percentage of measurements within one standard deviation of the mean is 71%. Within 2 standard deviations: 96%. Within 3: 100%. Our approximations are decent but not perfect.

Standardizing a Variable

The standardizing procedure we employed with the Etruscan skull data can be applied to any distribution. Suppose that X is a variable with data set $x_1, x_2, x_3, \dots, x_N$, sample mean \bar{x} , and sample variance s . Let $Y = X - \bar{x}$. The values of Y are $y_1, y_2, y_3, \dots, y_N$, where $y_1 = x_1 - \bar{x}$, $y_2 = x_2 - \bar{x}$, and so on. What is the mean \bar{y} of Y ? To calculate it, we need the following identity:

$$x_1 + x_2 + \dots + x_N = N \times \bar{x}. \quad (3.2.1)$$

To see why equation (3.2.1) is true, divide each side of the equation by N : the formula for the sample mean of X is the result of the divisions. We proceed with the calculation of the sample mean of Y , using equation (3.2.1) to obtain the third line in the calculation from the second:

$$\begin{aligned} \bar{y} &= \frac{1}{N} ((x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_N - \bar{x})) \\ &= \frac{1}{N} \left((x_1 + x_2 + \dots + x_N) - N \times \bar{x} \right) \\ &= \frac{1}{N} (N \times \bar{x} - N \times \bar{x}) \\ &= 0. \end{aligned}$$

Next, what can we say about the sample standard deviation s_Y of Y ? The following calculation shows that translation does not change the sample standard deviation. That is, the sample standard deviation s_Y of Y is the same as the sample standard deviation s of X :

$$\begin{aligned} s_y^2 &= \frac{1}{N-1} ((y_1 - 0)^2 + (y_2 - 0)^2 + \dots + (y_N - 0)^2) \\ &= \frac{1}{N-1} (y_1^2 + y_2^2 + \dots + y_N^2) \\ &= \frac{1}{N-1} ((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_N - \bar{x})^2) \\ &= s^2. \end{aligned}$$

In summary, $Y = X - \bar{x}$ has the same standard deviation as X , but its mean is 0.

Next, let

$$Z = \frac{Y}{s} = \frac{X - \bar{x}}{s}.$$

The values of Z are $y_1/s, y_2/s, y_3/s, \dots, y_N/s$, or $(x_1 - \bar{x})/s, (x_2 - \bar{x})/s, \dots, (x_N - \bar{x})/s$. What are the mean

\bar{z} and standard deviation s_Z of Z ? The calculations are not difficult:

$$\begin{aligned}\bar{z} &= \frac{y_1/s + y_2/s + \cdots + y_N/s}{N} \\ &= \frac{1}{s} \frac{y_1 + y_2 + \cdots + y_N}{N} \\ &= \frac{1}{s} \bar{y} \\ &= 0,\end{aligned}$$

and

$$\begin{aligned}s_Z^2 &= \frac{(y_1/s - \bar{z})^2 + (y_2/s - \bar{z})^2 + \cdots + (y_N/s - \bar{z})^2}{N - 1} \\ &= \frac{(y_1/s)^2 + (y_2/s)^2 + \cdots + (y_N/s)^2}{N - 1} \\ &= \frac{1}{s^2} \frac{y_1^2 + y_2^2 + \cdots + y_N^2}{N - 1} \\ &= \frac{1}{s^2} \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{N - 1} \\ &= \frac{1}{s^2} s^2 \\ &= 1.\end{aligned}$$

In summary, if X is any data set with mean \bar{x} and nonzero standard deviation s , then the standardization $Z = \frac{X - \bar{x}}{s}$ of X has mean 0 and standard deviation 1.

Standardizing Unimodal Symmetric Distributions: z-Scores

Suppose that X is any data set with mean \bar{x} and nonzero standard deviation s . Then $Z = (X - \bar{x})/s$ has mean 0 and standard deviation 1, as shown in the preceding subsection. If X has a normal distribution, then Z is standard normal. Even if X is not normal, the standard normal distribution is often a good approximation of the distribution of Z . Because the standard normal approximation is symmetric and unimodal, we can only expect it to be a reasonable approximation of the distribution of Z if the distribution of X is unimodal and symmetric. In this case, if x is a data value of X , then we say that

$$z = \frac{x - \bar{x}}{s}$$

is a **standardized value** or **z-score** of x . We use z-scores to approximate information about the distribution of Z by performing the same calculations with $\Phi(z)$ that we employed in Section 3.1. After these calculations, we use the equation $X = s \times Z + \bar{x}$ to transfer the information to the original variable X .

For the remainder of this section we will assume that all variables under consideration are approximately normal (which is equivalent to their standardizations being approximately standard normal).

Examples

Example 1. *An approximately normal population has mean 70.4 and standard deviation 10. What percentage of the population lies between 56.9 and 90.5?*

Solution. The z -score of 56.9 is $(56.9 - 70.4)/10$, or -1.35 . The z -score of 90.5 is $(90.5 - 70.4)/10$, or 2.01. With these values at hand, the question becomes, What percentage of a population with an approximately normal distribution has a z -score between -1.35 and 2.01? Notice that in this form of the question, there is no need to bring up the mean or standard deviation. Finally, the question may be restated as, What percentage of a standard normal population lies between -1.35 and 2.01? The answer is $100 \times (\Phi(2.01) - \Phi(-1.35))$. To use the table provided in the preceding section, which gives $\Phi(z)$ only for positive z , we must express $\Phi(-1.35)$ as $1 - \Phi(1.35)$. The answer we seek is $100 \times (\Phi(2.01) + \Phi(1.35) - 1)$, or $100 \times (0.9778 + 0.9115 - 1)$, or 88.93.

Example 2. (Washington University Exam, Fall 2010)

Assuming that the weight of girls at birth is approximately normal with mean 7.5 pounds and standard deviation 1, what percent of girls at birth weigh over 9 pounds?

Solution. The z -score of 9 is $(9 - 7.5)/1$, or 1.5. The percent of girls at birth with a z -score greater than 1.5 is $100 \times (1 - \Phi(1.5))$, or $100 \times (1 - 0.9332)$, or 6.68.

Example 3. (Washington University Exam, Fall 2010)

In a population survey of the greater St. Louis area, it was determined that there were 2,700,000 residents of which 100,000 were age 18 or younger. Assuming that the age of a resident is an approximately normal variable with standard deviation 12.1, what is the average age of the population?

Solution. The fraction of the population that is 18 years or younger is $100,000/2,700,000$ or 0.03704. That segment of the population is a left tail of the age variable. We seek a z -value such that $\Phi(z) = 0.03704$. Such a z must be negative because the value of $\Phi(z)$ is less than $1/2$. So, if we are to solve this with a table of positive z -values, we must convert the data so that the right side of the Bell Curve may be used. The identity we use is $\Phi(z) = 1 - \Phi(-z)$. Thus, we must solve $0.03704 = 1 - \Phi(-z)$, or $\Phi(-z) = 1 - 0.03704$, or $\Phi(-z) = 0.96296$. This Φ -value is between 1.78, for which $\Phi = 0.9625$, and 1.79, for which $\Phi = 0.9633$. Because $96296 - 96250 = 46$ and $96330 - 96250 = 80$, the Φ -value of our equation is $46/80$ of the way from 0.9625 to 0.9633. So the z -value we solve for is $46/80$ of the way from 1.78 to 1.79, or $1.78 + 0.01 \times (46/80)$, or 1.7857. In other words, the equation $\Phi(-z) = 0.96296$ has solution $-z = 1.7857$, or $z = -1.7857$. In still other words, the z -score of the age value 18 is -1.7857 . (The command `qnorm(0.03704)` in R results in the more precise value -1.786119 .) In terms of the unknown average \bar{x} of the age variable, we have $(18 - \bar{x})/12.1 = -1.7857$, or $18 - \bar{x} = -12.1 \times 1.7857$, or $\bar{x} = 18 + 12.1 \times 1.7857$, or $\bar{x} = 39.61$.

The Old Familiar Tails

The Empirical Rule of Section 3.1, also known as the 68-95-99.7 Rule is easily translated: If a data set X has an approximately normal distribution, then about 68% of the data values lie within one standard deviation from the mean, about 95% of the data values lie within two standard deviations from the mean, and about 99.7% of the data values lie within three standard deviations from the mean. An approximate rule that is in the same spirit is, If a data set X has an approximately normal distribution, then about 50% of the data values lie within two thirds of a standard deviation from the mean.

Equivalent formulations of these rules for an approximately normal variable are, about 25% of the data values lie in a tail, left or right, that begins $2/3$ of a standard deviation from the mean, about 16% of the data values lie in a tail, left or right, that begins 1 standard deviation from the mean, about 2.5% of the data values lie in a tail, left or right, that begins 2 standard deviations from the mean, and about 0.15% of the data values lie in a tail, left or right, that begins 3 standard deviations from the mean.

Events that are 5 or more standard deviations from the mean are rare indeed. The next example considers two such performances in track and field.

Example 4. At the 1968 Olympic Games in Mexico City, Bob Beamon shattered the world record for the men's long jump. Prior to his jump, since the beginning of record-keeping, the world record had been broken only 13 times. Let X be the variable that measures the amount in meters by which the record is broken. For

example, the first world record, set in 1901, was 7.61 m. That record was broken in 1921 with a jump of 7.69 m, so $x_1 = 0.08$ m. Before Beamon's jump, the mean of X was 0.0564 m and the standard deviation was 0.04199. With his jump of 8.90 m, which bettered the world record by 0.55 m (about 22 inches), by how many standard deviations did Beamon's extension of the record exceed the mean? (Prior to Beamon, the greatest progression of the record was by 0.15 m. In other words, Beamon increased x_{\max} from 0.15 m to 0.55 m. In the nearly 50 years that have gone by, Beamon's record has been bettered only once, and then by only 0.05 m. His jump still stands as the Olympic record.)

Some 20 years after Beamon's jump, at a meet in Indianapolis in 1988 prior to the Seoul Olympics, Florence Griffith Joyner ran the 100 m in 10.49 s. That broke the world record by 0.27 s. Prior to that run, the amount by which each new world record bettered the previous one had a mean of 0.05167 s, a standard deviation of 0.04215 s, and a maximum of 0.13 s. By how many standard deviations did Florence Griffith Joyner's run exceed the mean? (In the quarter century since, nobody has come close. The two closest times were 10.61 s and 10.62 s, both by Griffith Joyner in 1988.)

Solution. The z -score for Beamon was $(0.55 - 0.0564)/0.04199$, or 11.76. The z -score for Griffith Joyner was $(0.27 - 0.05167)/0.04215$, or 5.18. Notice that these unitless z -scores, which are scores on the same scale, provide us with a means of comparing measurements that would ordinarily be incomparable because of the different units involved. However, these z -scores are so off-the-chart, that they are, literally, off the chart—tables typically stop at a z -score of 3. The following graphic may be useful for appreciating the frequency with which events occur if they are several “sigs” away from the mean.

Range	Population in range	Expected frequency outside range	Approx. frequency for daily event
$\mu \pm 1\sigma$	0.682689492137086	1 in 3	Twice a week
$\mu \pm 1.5\sigma$	0.866385597462284	1 in 7	Weekly
$\mu \pm 2\sigma$	0.954499736103642	1 in 22	Every three weeks
$\mu \pm 2.5\sigma$	0.987580669348448	1 in 81	Quarterly
$\mu \pm 3\sigma$	0.997300203936740	1 in 370	Yearly
$\mu \pm 3.5\sigma$	0.999534741841929	1 in 2149	Every six years
$\mu \pm 4\sigma$	0.999936657516334	1 in 15 787	Every 43 years (twice in a lifetime)
$\mu \pm 4.5\sigma$	0.999993204653751	1 in 147 160	Every 403 years (once in the modern era)
$\mu \pm 5\sigma$	0.999999426696856	1 in 1 744 278	Every 4 776 years (once in recorded history)
$\mu \pm 5.5\sigma$	0.999999962020875	1 in 26 330 254	Every 72 090 years (thrice in history of modern humankind)
$\mu \pm 6\sigma$	0.999999998026825	1 in 506 797 346	Every 1.38 million years (twice in history of humankind)
$\mu \pm 6.5\sigma$	0.999999999919680	1 in 12 450 197 393	Every 34 million years (twice since the extinction of dinosaurs)
$\mu \pm 7\sigma$	0.99999999997440	1 in 390 682 215 445	Every 1.07 billion years (four or five times in Earth's history)

Figure 3.2.6 Higher Deviations Source: http://en.wikipedia.org/wiki/68-95-99_rule
Retrieved:13 September 2014

Standardizing in R (Optional)

If x is a value of a data set X coded as a list in R, then $z = (x - \text{mean}(X))/\text{sd}(X)$ is the z -score of x . This standardization can be performed on every observation in X by means of a for-loop as follows:

```

> N = length(X)
> Z = numeric(N)
for(j in 1:N) Z[j] = (X[j] - mean(X))/sd(X)

```

Details of this construction have been discussed in connection with Figure 2.2.8 in Section 2.2 and will not be repeated here. To be up front, it must be admitted that the construction just given is not necessary: the code `Zee = (X - mean(X))/sd(X)` results in a list `Zee` that is identical to `Z`.¹² Actually, R has a built-in function, `scale()`, that gets you to the same place even more quickly. The call `Zed = scale(X)` results in a table named `Zed` with one column. The call `Zed[,1]` causes the printing of a list that is identical to the lists `Z` and `Zee` that have already been created. In the new construction, `Zed[j,1]` is the z -score of `X[j]`. By the way, if all we want are the Φ -values of the entries of `X`, then we need not even cause their z -scores `Z` to be calculated as an intermediate step. The call `pnorm(Z)` returns the Φ values of all the z -scores of `X`, but so does `pnorm(X, mean(X), sd(X))`.

Exercises

1. (Nuts and bolts of Φ) This problem should be done by using the table provided in this chapter. If you intend to use a statistics-enabled calculator on course examinations, then it is also advisable that you learn to answer these questions efficiently with your calculator.

In all parts, `X` is a distribution that can be accurately described by the standard normal model.

- a) What fraction of the observations of `X` are smaller than 0.684?
 - b) What fraction of the observations of `X` are smaller than -0.684?
 - c) What fraction of the observations of `X` are greater than 0.553?
 - d) What fraction of the observations of `X` are greater than -1.234?
 - e) What fraction of the observations of `X` are between -0.85 and 0.85?
 - f) What fraction of the observations of `X` are between 0.57 and 0.75?
 - g) What fraction of the observations of `X` are between -0.94 and 0.57?
 - h) What fraction of the observations of `X` are between -1.75 and -0.94?
 - i) What fraction of the observations of `X` are between 0 and 1.87?
 - j) What fraction of the observations of `X` are between -1.66 and 0?
2. (Bolts and nuts of Φ) This problem should be done by using the table provided in this chapter. If you intend to use a statistics-enabled calculator on course examinations, then it is also advisable that you learn to answer these questions efficiently with your calculator.

In all parts, `X` is a distribution that can be accurately described by the standard normal model.

- a) For what z is $\Phi(z) = 0.8$?
- b) For what z is $\Phi(z) = 0.75$?
- c) For what z is $\Phi(z) = 2/3$?
- d) For what z is $\Phi(z) = 0.4$?
- e) For what z is $\Phi(z) = 1/3$?
- f) For what z is $\Phi(z) = 1/4$?

¹²The author admits it grudgingly rather than pronouncing it proudly because the operation looks ugly: the code seems to be subtracting a scalar from a vector, an operation that is not defined in algebra.

3. Use the values found in the preceding exercise to fill in the blanks in the given statements about a normal distribution X .
 - a) The percentage of the observations in X that are above the mean but not more than z standard deviations greater than the mean is 30% for $z = \underline{\hspace{2cm}}$.
 - b) The percentage of the observations in X that are within z standard deviations of the mean is 60% for $z = \underline{\hspace{2cm}}$.
 - c) About half the observations in X are within z standard deviations of the mean for $z = \underline{\hspace{2cm}}$.
 - d) About one-sixth of the observations in X are below the mean but not more than z standard deviations below the mean for $z = \underline{\hspace{2cm}}$.
4. Suppose that an observed value x of a normal distribution X has a given z -score. What is the raw value of x if
 - a) $z = -0.37$, $\bar{X} = 65$ and the standard deviation of X is 4.8.
 - b) $z = 1.43$, $\bar{X} = 3.91$ and the standard deviation of X is $5/7$.
5. Zoe's SAT score gives her a percentile exactly equal to her IQ of 125. What is her SAT score? (SAT exams are scored so that the mean is 500 and the standard deviation is 100. IQ scores have a mean of 100 and a standard deviation of 15 or 16. Use 15 in this problem.)
6. The LSAT scores of the law students of Second President University can be modelled by a normal distribution with mean 700 and standard deviation 40. Zenobia, Zelda, Ziva, and Zuzu are law students at the university. Zenobia's LSAT score is 755. Zelda's LSAT score is 1.34 standard deviations below the school's mean. Ziva's LSAT score gives her a 95 percentile in the university's law school and unlucky Zuzu's percentile is 13.
 - a) What percentage of law students at the university have higher LSAT scores than Zenobia's?
 - b) What percentage of law students at the university have lower LSAT scores than Zelda's?
 - c) What is Ziva's LSAT score? d) What is Zuzu's LSAT score?
7. Assume that SAT exams are graded so that the mean is 500 and the standard deviation is 100, that ACT exams are graded so that the mean is 21 and the standard deviation is 5, and that the scores on both types of exams follow a normal distribution.
 - a) To what ACT score does a 772 on the SAT correspond?
 - b) To what SAT score does a 28 on the ACT correspond?
 - c) For what value of b is it true that the percentage of students with ACT scores between 24 and 30 equals the percentage of students with SAT scores between 540 and b ?
8. Whereas statisticians transform raw data into z -scores so that the mean of the transformed data is 0 and the standard deviation is 1, psychologists often prefer to transform raw data into T -scores so that the mean of the transformed data is 50 and the standard deviation is 10. (The letter " t " brings to mind a particular distribution that has nothing to do with the T -score defined in this problem. Alas, values arising from that distribution are sometimes called t -scores. Adding to the confusion, there is yet another t -score used for reporting results of bone density measurements.) Complete the following table for a distribution X of size 5:

Observation	Raw	z-score	T-score
1	11		
2			45.07634
3		0.246183	
4	20	0.7385489	
5	21	0.9847319	
Mean			
Standard Devtn			

Hint: The two rows with two entries allow you to obtain two linear equations in the unknown mean and unknown standard deviation. Once you find these values, all the other unspecified quantities fall into place.

9. Do newborns recognize their mothers' faces? An experiment reported in 1989 was designed to test this question. The subjects were 48 neonates between 13 and 100 hours old with an average age of 50 hours. The mother of each neonate and a stranger, another woman of similar hair and skin color, stood within the view of the newborn. A screen placed in front of the women allowed only the faces of the women to be seen. The percentage of the time that each newborn looked at its mother was recorded. The observations, sorted for your convenience, were 18, 21, 28, 28, 29, 38, 41, 46, 48, 48, 49, 51, 51, 53, 55, 55, 56, 57, 58, 58, 59, 59, 59, 61, 62, 63, 64, 65, 66, 67, 67, 68, 72, 73, 75, 76, 76, 79, 82, 84, 86, 86, 87, 87, 89, 92, 92, 93. The mean of this distribution is 62.02083 and the standard deviation is 19.21711.

a) What are the z-scores of the quartiles Q_1, Q_2, Q_3 ? (Calculate each score even if the quartile is not an observed value.)

b) If the distribution was accurately modelled by a normal distribution, then what would you expect the z-scores of the quartiles Q_1, Q_2, Q_3 to be?

c) What is the z-score of the largest observation?

d) In a distribution of size 48 that conforms to a normal model reasonably well, about how many observations would you expect there to be with z-scores less than or equal to the value found in part (c)?

10. In an experiment investigating the imitative responses of infants to certain gestures, a psychologist demonstrated "tongue protrusion" to 140 three-month old infants. (The author is not entirely certain but he believes that, in layman's terms, the psychologist stuck out his tongue at the infants.) Seven responses to the gesture were observed. The responses were things like finger movements, mouth openings, and reciprocating tongue protrusions. The frequencies of these responses were 16, 18, 31, 17, 29, 10, 19. Calculate the z-scores of these frequencies. Verify that the distribution of z-scores is a standard distribution (mean 0, standard deviation 1).

A manufacturer asserts that the lifespan (in months) of the copy machine it produces is approximately normal and is modelled by $N(42, 7)$. The next five questions pertain to this machine. (Washington University exam, Spring 2014)

11. In trying to win a contract for the sale of 200 units, a salesman relies on the manufacturer's assertion and states, "190 of the units will last between 28 and n months." What was the asserted value of n ?
12. What is the 3rd quartile of copier lifespans?

13. What percent of the copiers are expected to fail before 36 months?
14. The manufacturer wants to reduce the 36-month failure rate to only 10%. Assuming the standard deviation will stay the same, what mean lifespan must they achieve?
15. A competing manufacturer says that not only will 90% of their copiers last at least 36 months, 65% will last at least 42 months. What normal model parameters is that manufacturer claiming?
16. Suppose a normal model has mean $\mu = 5$ and standard deviation $\sigma = 3$. Calculate to 2 decimal places the percentage of data that you expect to be between -1 and 11. (Washington University exam, Spring 2010)
17. Which of these variables is most likely to follow a Normal model? (Washington University exam, Spring 2014)
 - A) number of cigarettes smoked daily
 - B) number of TV sets at home
 - C) head circumference
 - D) hours of homework last week
 - E) eye color
18. Are women really more talkative than men? You might very well think that, but the author couldn't possibly comment. On the other hand, it is a question that has been investigated. In a study¹³ of American men and women, the mean number of words spoken daily by women was 14,297 with a standard deviation of 6441. For men, the corresponding values were 14,060 and 9065. Suppose that Nadia utters more words daily than 90% of her cohort. Suppose that Aidan, never mind that some see him as backwards, is exactly as chatty as Nadia—that is, his daily word count is the same. How many words per day does Aidan speak and what is his rank in his cohort?
19. Suppose that the age X in months at which babies learn to walk is a normal distribution. By the age of 10 months, 5% of babies have learned to walk. By the age of 13 months, 75% of babies have learned to walk. What are the mean and standard deviation of X ?
20. Dear First President University Professor,
 For question 1, part a in the Chapter 3 exercises, I was wondering why my method of interpolation produced a different result from yours? I have attached an image of my work.
 Best Regards,
 First President University Student

¹³Matthias R. Mehl et al., *Are women really more talkative than men?* Science **317** (2007), p.82.

$$\begin{aligned}
 \Phi(0.684) &= \Phi(0.68 + 0.004) \\
 &= \Phi\left(0.68 + \left(\frac{1}{10}\right)(0.04)\right) \\
 &= \Phi\left(0.68 + \left(\frac{1}{10}\right)(0.09 - 0.05)\right) \\
 &= \Phi(0.68) + \frac{1}{10}(\Phi(0.09) - \Phi(0.05)) \\
 &= 0.7517 + \frac{1}{10}(0.5359 - 0.5199) \\
 &= 0.7533
 \end{aligned}$$

Diagnose what is going wrong with the calculation. Start with the second line, which is not wrong, but which points to an intended wrong path. Continue with a critique of the third line, which is also not wrong, but which is already on the way down the wrong path. Be ruthlessly critical of the fourth line, where the calculation does go very wrong. Use a graph to illustrate what the correct interpolation should have been instead.