Contents

6.4 Normal Distribution

The normal probability distribution is the most commonly used probability distribution in statistical work. It is the 'bell curve' often used to set test scores, and describe the distribution of variables such as height and weight in human populations. The normal distribution also emerges in mathematical statistics. For example, probabilities in the binomial probability distribution can be estimated by using the normal distribution. Also, if random samples from a population are taken, and if the sample sizes are reasonably large, then the means from these samples are normally distributed. It is these latter uses which are the most important in this textbook, and the normal distribution is applied to these in the next chapter, and throughout the rest of the textbook.

In order to use the normal probability distribution, it is first necessary to understand a few aspects of its construction, and how to determine probabilities using this distribution. These are discussed in this section. A few comments concerning the use of the normal distribution are also given. But what is most important to learn at this point is how to determine areas under the curve of the normal distribution and normal probabilities.

6.4.1 Characteristics of the Normal Distribution

The normal distribution is the **bell curve**, being bell shaped. However, it is not just any bell shaped curve, it is a specific probability distribution, with an exact mathematical formula. Sometimes this distribution is referred to as the **Gaussian** distribution, named after Gauss, one of the mathematicians who developed this distribution.

The distribution is shaped as in Figure 6.3, and several characteristics of the distribution are apparent in this figure. First, note the axes. The horizontal axis represents the values of the variable being described. The vertical axis represents the probability of occurrence of each of the values of the variable. Since the curve is highest near the centre, this means that the probability of occurrence is greatest near the centre. If a population is normally distributed, large proportions of the population are near the centre. The farther away from the centre the value of the variable, the lower is the probability of occurrence of the variable. A population which is normally distributed has most of its cases near the centre, and relatively few values distant from the centre.

A related characteristic evident from the diagram is that the peak of the

Figure 6.3: The Normal or Bell Curve

distribution is at the centre of the distribution. The most common value, the mode, is at the very centre of the normal distribution. In addition, the distribution is symmetric around the centre, with one half being a mirror image of the other. This means that the median also occurs at the peak of the distribution, since one half of the values in the distribution lie on each side of the peak. Finally, the mean is also at the centre of the distribution, since deviations about the mean on one side have the same probability of occurrence of the deviations about the mean on the other side. As a result, the mode, median and mean are equal, and all occur where the distribution reaches its peak value, at the centre of the distribution.

Since the probabilities become smaller, the further from the centre the variable is, it appears as if the distribution touches the X-axis. In fact, the curve never quite touches this axis, getting closer and closer, but never quite touching the X-axis. Such as curve is said to be **asymptotic** to the X-axis, approaching but never quite touching the X-axis. For most practical purposes, it will be seen that the curve becomes very close to the X-axis beyond 3 standard deviations from the centre. If a variable is normally distributed, only a little over 1 case in 1,000 is more than 3 standard deviations from the centre of the distribution.

Figure 6.4: Normal Distribution of Grades

Example 6.4.1 Normal Distribution of Class Grades

In order to make this discussion more meaningful, suppose an instructor grades a class on the basis of a normal distribution. In addition, suppose that the instructor sets the mean grade at 70 per cent, and the standard deviation of grades at 10 per cent. This is the situation depicted in Figure 6.4, where the the mean grade is at the centre of the distribution. In this normal curve, the values along the horizontal axis represent grades of the students, and the vertical axis represents the probability of occurrence of each grade. Recall that the total area under the whole curve must be equal to 1, since this area represents all the cases in the population. The proportion of students receiving each set of grades is then represented by the respective area under the curve associated with that set of grades.

Figure 6.4 geometrically represents the proportion of students who receive grades between each of the limits 50-60, 60-70, 70-80 and 80-90. The proportion of students receiving grades between 60 and 70 is given by the proportion of the area under the curve between 60 and 70. Similarly, the area under the curve between 50 and 60 gives the proportion of students who receive between 50 and 60. The proportion who fail, receiving less than

Grade	Proportion of Students
90+	0.0228
80-90	0.1359
70-80	0.3413
60-70	0.3413
50-60	0.1359
<50	0.0228

Table 6.21: Normal Grade Distribution, $\mu = 70$, $\sigma = 10$

50, is the small area under the curve to the left of 50. The centre of the distribution represents the average grade, and since the curve is highest there, with large areas under the curve, many students receive grades around the average. The farther away from the centre, the smaller the area under the curve, and the fewer the students who receive grades farther from the average. If the instructor does construct grades so that the grades are exactly normally distributed with $\mu = 70$ and $\sigma = 10$, the proportion of students with each set of grades is as given in Table 6.21. You can check these figures once you have studied the methods later in this section.

There are many normal distributions, and each variable X which is normally distributed, has a mean and standard deviation. If the mean is μ and the standard deviation is σ, the probability of occurrence of each value of X can be given as

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

This expression is not likely to be familiar, and being able to use the normal distribution does not depend on understanding this formula. However, this formula shows that there are two parameters for each normal curve, the mean μ, and the standard deviation σ. Once these two values have been specified, this uniquely defines a normal distribution for X. This means that to define a normal distribution, it is necessary to specify a mean and standard deviation. It also means that anytime a normal distribution is given, there will be a particular mean μ, and standard deviation σ associated with this distribution.

The formula also means that there are an infinite number of normal distributions, one for each μ and σ. But each normal curve can be transformed into the **standardized normal distribution**. This is the normal distribution which has a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$. A simple mathematical transformation can be used to change any normal distribution to the standardized normal distribution. This is given in Section 6.4.4. The probabilities for each value of the standardized normal variable are given in Appendix **??**. Based on the table in Appendix **??**, and the transformation between any normal variable X, and the standardized normal variable, Z, the probabilities associated with any normal variable can be determined.

The following section shows how to determine probabilities for the standardized normal distribution, and Section 6.4.4 shows how probabilities can be obtained for any normal variable.

6.4.2 The Standardized Normal Distribution

The standardized normal distribution is a particular normal distribution, in that it has a mean of 0 and a standard deviation of 1. In statistical work, the variable which has a standardized normal distribution is termed Z. In statistics, any time the variable Z is encountered, this denotes a variable with a mean of 0 and a standard deviation of 1. For Z, the mean is $\mu = 0$ and the standard deviation of Z is $\sigma = 1$. In Figure 6.5,the distribution of the standardized normal variable Z is given. The Z values are given along the horizontal axis, and the areas associated with this distribution are given in Appendix **??**.

In each of the figures of the normal distribution, notice that the vertical axis is labelled the probability, $P(Z)$. The horizontal axis of the standardized normal distribution is the value of Z. Since the standard deviation of Z is $\sigma = 1$, these Z values also represent the number of standard deviations from centre. For example, when $Z = 2$, this represents a point 2 Z values above the mean, but this also represents a value of the variable that is 2 standard deviations above the mean of the distribution.

The height of the standardized normal curve at each point indicates the probability of particular values of Z occurring. These ordinates are not given in the Appendix **??**. Instead, areas under the normal curve are given as the areas in Appendix **??**. It will be seen that these are the important values for determining the proportions of the population with particular values of the variable. Also recall that the total area under the curve for any distribution is is 1. Since the normal distribution is symmetrical about the centre of the distribution, ($Z = 0$), the area on each side of centre is equal to 0.5. This

Figure 6.5: The Standardized Normal Variable Z

is an important characteristic of the distribution, when determining areas under the curve. These areas will later be interpreted as probabilities, and also as proportions of the population taking on specific sets of values of the variable.

Determining Areas under the Standardized Normal Distribution. In order to use the normal distribution, it is necessary to be able to determine areas under the curve of the normal distribution. These areas are given in Appendix **??**. The diagram attached to the table shows a shaded area between the centre of the curve and a point to the right of centre. As given in the diagram, this is the area under the normal distribution between the centre, $Z = 0$ and the value of Z indicated. The various possible Z values are given in the first column of the table. Note that these are values of Z from 0 to 4, at intervals of 0.01.

The second column of Appendix **??** gives the area under the normal distribution that occurs between $Z = 0$ and the respective values of Z in the first colummn. The values in this column represent the shaded area of the diagram, that is, the area between the centre and the corresponding Z value in the first column. The third column of the table represents the area under the curve from the value of Z indicated, to $+\infty$, that is, the area under the curve beyond the appropriate Z.

Also note that for each Z in the table, the sum of the second and the third column is always 0.5000. This is because one half, or 0.5000, of the area is to the right of centre, with the other half to the left of centre. The half of the area to the right of centre is equal to the area from the centre to Z, plus the area beyond Z.

Since the normal curve is symmetric, only one half of the normal curve need be given in Appendix **??**. Areas to the left of centre are equal to the corresponding areas to the right of centre, but with a negative value for Z. For example, the area under the normal distribution between $Z = 0$ and $Z = -1$ is the same as the area under the distribution between $Z = 0$ and $Z = 1$.

Determining the various areas under the normal distribution will become easier once the following examples are studied.

Example 6.4.2 Areas Associated with a Specific Z

To begin, Appendix **??** can be used to determine the area

1. Between $Z = 0$ and $Z = 1$.
2. Between $Z = 0$ and $Z = -1$.
3. Above $Z = 1.75$.
4. Below -2.31.

These are determined as follows.

1. The area between $Z = 0$ and $Z = 1$ is given in Figure 6.6. This can be read directly from Appendix **??**, by going down the Z column until $Z = 1$ is reached. To the right of this, in column A, the value of 0.3413 is given, and this is the area under the normal distribution between $Z = 0$ and $Z = 1$. Also note the area of 0.1587 in column B. This is the area that lies to the right of $Z = 1$. That is, 0.3413 of the area under the normal distribution is between the centre and $Z = 1$, and 0.1587 of the area under the normal distribution lies to the right of $Z = 1$. In Figure 6.6, also note that one half of the total area lies to the left of centre.

Figure 6.6: Areas Associated with $Z = 1$

2. Since Appendix **??** gives only the area to the right of centre, the only way of determining this area is to recognize that by symmetry, the area between $Z = 0$ and $Z = -1$ is the same as the area between $Z = 0$ and $Z = 1$. Since the latter area has already been determined to be 0.3413, the area under the normal distribution between the centre and $Z = -1$ is 0.3413.

3. Areas above particular values of Z are given in column B of Appendix **??**. The area above $Z = 1.75$ can be read directly from Appendix **??**, by looking at the value of $Z = 1.75$. In column B, the associated area is 0.0401. This means that a proportion 0.0401 of the area, or 0.0401% = 4.01% of the area lies to the right of $Z = 1.75$.

4. The area below $Z = -2.31$ is not given directly in the table. But again, by symmetry, this is the same as the area above $Z = 2.31$. Looking up the latter value inAppendix **??**, the area associated with this is 0.0104. This is also the area below $Z = -2.31$. Also note that the area in column B is 0.4896. This is the area between the centre and $Z = 2.31$, and by symmetry, this is also the area under the normal distribution that lies between the centre and $Z = -2.31$.

Example 6.4.3 Areas Between Non Central Z Values

Determine the areas under the normal curve between the following Z values.

1. Between -2.31 and +1.75.

2. Between 0.50 and 2.65.

3. Between -1.89 and -0.13.

These areas cannot be determined directly from Appendix **??** without some calculation. The method of obtaining these areas is as follows.

1. Figure 6.7 may be helpful in determining this probability. The area between -2.31 and +1.75 can be broken into two parts, the area between -2.31 and 0, and the area between 0 and 1.75. This is necessary because Appendix **??** only gives areas between the centre and specific Z values. Once this area has been broken into two parts, it becomes relatively easy to compute this area. From Figure 6.7, the area between $Z = 0$ and $Z = -2.31$ is 0.4896. As shown in Example 6.4.2, this area is the same as the area between $Z = 0$ and $Z = 2.31$. By symmetry, and from Appendix **??**, this area is 0.4896. The area between 0 and 1.75 can be determined from Appendix **??** by looking at column A for $Z = 1.75$. This area is 0.4599.

   The total area requested is thus $0.4896 + 0.4599 = 0.9495$. Almost 95% of the area under the normal distribution is between $Z = -2.31$ and $Z = +1.75$.

2. The area between $Z = 0.5$ and $Z = 2.65$ is given in Figure 6.8. Again this area must be calculated. The area requested can be obtained by recognizing that it equals the area between the centre and $Z = 2.65$, minus the area between the centre and $Z = 0.5$. The whole area between the centre and $Z = 2.65$ is 0.4960, obtained in column A of Appendix **??**. This is more than the area requested, and from this, the area between the centre and $Z = 0.5$ must be subtracted. From Appendix **??**, the latter area is 0.1915.

   The area requested is thus $0.4960 - 0.1915 = 0.3045$. This is the shaded area in Figure 6.8.

Figure 6.7: Area between -2.31 and +1.75

Figure 6.8: Area between 0.5 and 2.65

3. The area between $Z = -1.89$ and $Z = -0.13$ must again be calculated in a manner similar to that used in the last part. The area between $Z = -1.89$ and the centre can be determined by noting that this is the same as the area between the centre and $Z = 1.89$. This area is 0.4706. From this, the area between the centre and -0.13 must be subtracted. The latter is the same as the area between $Z = 0$ and $Z = 0.13$, and for the latter Z, the area given in column A of Appendix **??** is 0.0517.

   The required area is the area between the centre and -1.89 minus the area between the centre and -0.13. This is $0.4706 - 0.0517 = 0.4189$. This is the requested area.

Example 6.4.4 Obtaining Z values from Specified Areas

Areas are frequently given first, and from this the values of Z are to be determined. The areas given in columns A and B of Appendix **??** form the basis for this. However, it may not be possible to find an area in these columns which exactly matches the area requested. In most cases, the area closest to that requested is used, and the Z value corresponding to this is used. Find the Z values associated with each of the following.

1. The Z so that only 0.2500 of the area is above this.
2. The 30th percentile.
3. The middle 0.95 of the normal distribution.
4. The Z values for the middle 90% of the distribution, so that the bottom and top 5% of the distribution are excluded.

These all require examination of columns A and B first, and then determination of the Z values associated with the appropriate area. These are determined as follows. Figure 6.9 is used to illustrate the first two questions.

1. The Z so that only 0.25 of the area is above this, must lie somewhat to the right of centre, since over one half of the area is to the right of centre. Column B of Appendix **??** gives areas above different values of Z, and the value closest to 0.25 in this column is the appropriate Z. The two areas closest to 0.25 are 0.2514, associated with $Z = 0.67$ and 0.2483, associated with $Z = 0.68$. The former of these is closer to 0.25 than the latter, so to two decimal places, the Z which is closest is $Z = 0.67$. Since 0.25 is approximately half way between these two

Figure 6.9: Z for 75th and 30th Percentiles

areas in the table, $Z = 0.675$ might be reported as the appropriate Z. Note that $Z = 0.67$ is the 75th percentile, since a proportion of only 0.25, or 25% the area is above this Z. This means that 75% of the area under the curve is below this. Thus in the standardized normal distribution, $P_{75} = 0.67$.

2. The 30th percentile is the value of Z such that only 0.3000 of the area is less than this Z and the other 70% or 0.7000 of the distribution is above this. This means that the Z so that only 0.3000 of the area is beyond this, must be found. This is obtained by looking down column B until a value close to 0.3000 is found. The closest Z is $Z = 0.52$, where the area beyond this is 0.3015. The next closest value is $Z = 0.53$ associated with an area of 0.2981 beyond this. The former is a little closer so it will be used. Since we are looking for a Z to the left of centre, the appropriate Z is -0.52. That is $P_{30} = -0.52$ in the standardized normal distribution. Only 30 per cent of the area under the curve is below $Z = -0.52$.

3. The middle 0.95 of the distribution is the area around the centre that accounts for 0.95 of the total area. This are cannot be directly deter-

Figure 6.10: Middle 0.95 of the Standardized Normal Variable Z

mined, and the method of determing this is illustrated in Figure 6.10. Since the distribution and area requested are both symmetrical, the total area of 0.95 in the middle of the distribution can be divided into two halves of 0.475 on each side of centre. That is, this amounts to an area of $0.95/2 = 0.475$ on each side of centre. Looking down column A of Appendix **??** shows that an area of 0.475 between the centre and Z gives a Z of exactly 1.96. That is, $Z = 1.96$ is associated with an area of 0.4750 between the centre and this Z. Since the situation is symmetrical around the centre, the appropriate Z on the left is -1.96. That is, going out from centre a Z of 1.96 on either side of centre gives an area of 0.475 on each side of centre, or a total of 0.95 in the middle. The interval requested is from $Z = -1.96$ to $Z = +1.96$.

Also note what is left in the tails of the distribution. Below $Z = -1.96$ is an area of 0.025, and above $Z = 1.96$ is an area of 0.025. That is, there is only 2.5% of the distribution beyond each of these individual Z values. This is a total of 0.05, or 5% per cent, of the distribution outside these limits. This property will be used extensively in Chapter 7. There the interval estimates are often 95% interval estimates,

Figure 6.11: Middle 90% of the Standardized Normal Distribution

requesting the middle 95% of the distribution, leaving out only the extreme 5% of the distribution.

4. This part is a variant of the last part, except looking for the middle 90%, rather than the middle 95% of the distribution. The method of obtaining this is given in Figure 6.11. In this case, the middle 90middle 0.9000 of the area is composed of 0.4500 on each side of centre. This leaves a total of 0.10 in the two tails of the distribution, or an area of 0.05 in each tail of the distribution. Looking down column B of Appendix **??** shows that $Z = 1.64$ has an area of 0.0505 associated with it. The next Z of 1.65 has an area of 0.0495 associated with it. An area of 0.0500 is exactly mdiway between these two values. As a result, the Z value of 1.645 is used in this case. Above $Z = 1.645$, there is exactly 0.05 of the area. By symmetry, there is also an area of 0.05 below $Z = -1.645$. The required interval is from -1.645 to +1.645. This interval contains the middle 90% of the cases in the standardized normal distribution.

   Again, this interval is widely used in the following chapters. In Chapter 7, the 90% interval is used, and if the distribution is normal, this is associated with $Z = \pm 1.645$. Later, when conducting hypothesis tests, one of the extremes of 0.05 of the distribution is often excluded, and again this is associated with a Z of 1.645 from the centre.

6.4.3 Meaning of Areas under the Normal Curve.

The areas under the normal curve have so far been interpreted as nothing more than areas. Depending on how the normal distribution is used, there are several possible interpretations of these areas. They may be simply areas, or they may be interpreted as proportions, percentages, probabilities, or even numbers people. A short discussion of these interpretations follows.

Areas. This requires little further discussion because this is how the normal curve has been interpreted so far. The total area under the whole curve is 1, and each pair of Z values defines an area under the curve. These are simply areas under the curve, or fractions of the total area under the curve.

Proportions. If a population has a standardized normal distribution, the areas under the curve also correspond to **proportions**. That is, each pair of Z values defines a range of values of the variable. The area under the curve between these two Z values is the proportion of the population which takes on values within this range. In Example 6.4.3 suppose that the distribution represents the distribution of a variable in a population. The total area of 1 under the whole curve is equivalent to accounting for the whole population. That is, the sum of all the proportions in a proportional distribution is 1. The proportion of the population between the limits of $Z = 0.5$ and $Z = 2.65$ is equal to the area within these limits, and this was shown to be 0.3045. Thus 0.3045 of the population is between $Z = 0.5$ and $Z = 2.65$.

Percentages. If the proportions are multiplied by 100%, the proportions become percentages. Similarly, the areas multiplied by 100% become percentages of the population within the respective limits. As in the last paragraph, the percentage of the population between $Z = 0.5$ and $Z = 2.65$ is 30.45%, if the population has a standardized normal distribution. Also note that in Example 6.4.4 the middle 95% of the area was associated with the range from $Z = -1.96$ to $Z = 1.96$. The middle 95% of cases in a standardized normal population is within these limits.

Probabilities. The areas or proportions can also be interpreted as probabilities. If a case is randomly selected from a standardized normal population, then the probability of its being between two specific Z values is the same as the area between these Z values. Example 6.4.4 shows that 90% of the area under the normal curve is between $Z = -1.645$ and $Z = +1.645$. As a probability, this could be interpreted as saying that the probability that Z is between -1.645 and $+1.645$ is 0.90. In symbols, this can be stated as

$$P(-1.645 < Z < +1.645) = 0.90.$$

This also means that the probability is 0.05 that Z is either less than -1.645 or greater than $+1.645$. That is,

$$P(Z < -1.645) = 0.05 \text{ and } P(Z > +1.645) = 0.05.$$

Number of Cases. If the proportion or area is multiplied by the number of cases in the sample or population, this results in the number of cases that lie within specific limits. For example, suppose that a population has 500 people in it. The middle 0.90 of the area has limits for Z of -1.645 and +1.645. This means that $0.90 \times 500 = 450$ people in this population have Z values between these $Z = -1.645$ and $Z = +1.645$.

Meaning of Z. The values of Z can be interpreted as merely being values of Z, distances along the horizontal axis in the diagram of the standardized normal probability distribution. Remember though that the standardized normal has a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$. Thus each Z value not only represents a distance from the centre, but it also represents the number of standard deviations from centre. For example, since the standardized normal distribution has a standard deviation of 1, if $Z = 1$, then this Z is one standard deviation above the mean. $Z = 2$ represents a point which is two standard deviations above the mean. The point on the horizontal axis where $Z = -1.6$ is 1.6 standard deviations to the left of centre, or below the mean.

Some of the rules concerning the interpretation of the standard deviation given on page 254 of Part I of this text can be illustrated using this interpretation of the Z value. In Example 6.4.2 above, the area between $Z = 0$ and $Z = 1$ was shown to be 0.3413, with the area to the right of $Z = 1$ being 0.1587. If a population has a standardized normal distribution, then a proportion 0.3413 of the population is between the mean and 1 standard deviation above centre. This is $0.3413 \times 100\% = 34.13\%$, or rounded off, 34% of the total cases. On the basis of the symmetry of the curve, another 0.3413, or 34% are between the centre and 1 standard deviation below centre. In total, there are thus $0.3413 + 0.3413 = 0.6826$ or 68.26% of the total cases within one standard deviation of the mean. Rounded off, this is 68% of the cases, or just over two thirds of the cases within one standard deviation of the mean.

The proportion of 0.1587 being to the right of $Z = 1$ means that only 0.1587, or 16% of the cases in the standardized normal lie more than 1 standard deviation above the mean. Similarly, another 16% of the cases are more than one standard deviation below the mean. Together this means that about 32% of the population is farther than one standard deviation away from the mean.

Example 6.4.4 and Table 6.10 showed that the middle 95% of the normal distribution was between $Z = -1.96$ and $Z = 1.96$. That is, within 1.96 standard deviations on either side of the mean, there are 95% of all the cases in the standardized normal distribution. While this is an exact determination, this produces the rough rule of page 254 of Part I of this text that 95% of all cases are within two standard deviations of the mean. Also note that this implies that approximately 5% of of the cases are farther than two Z values, or two standard deviations away from the mean.

Finally, you can use the normal probabilities in Appendix **??** to show that within three standard deviations of the mean, the standardized normal has 99.74% of all the cases. Only 0.26% of all the cases are farther than three standard deviations from the mean in the case of the standardized normal.

6.4.4 The General Normal Distribution

The standardized normal distribution is a special case of the normal distribution. In general, a normal variable X has mean μ and standard deviation σ. This section shows how this general normal distribution can be transformed into the standardized normal distribution, and how the transformation can go the other way as well. This means that Appendix **??** can always be used to determine areas or probabilities associated with the standardized normal distribution, and the transformation used to associate these with any other distribution.

Figure 6.12: The General Normal Variable X with Mean μ and Standard Deviation σ

Notation. In order to keep track of each normal curve, it is useful to have a general notation. Suppose the variable X is a normally distributed variable, with a mean of μ and a standard deviation of σ. Let this be denoted by $\text{Nor}(\mu, \sigma)$. That is, . That is,

$$X \ \text{ is } \ \text{Nor}(\mu, \sigma)$$

is a short form for saying that variable X is normally distributed with mean μ and standard deviation σ. Given this notation, the standardized normal variable Z with mean 0 and standard deviation 1 can be denoted by

$$Z \ \text{ is } \ \text{Nor}(0, 1).$$

The general normal distribution is pictured in Figure 6.12. The variable X is normally distributed with the mean μ at the centre of the distribution. The standard deviation is σ. This is a little more difficult to show in the diagram, but it can be seen what the approximate distance from the centre to one standard deviation on each side of the mean is.

In order to transform this variable X into the standardized normal distribution, the following transformation is used.

$$Z = \frac{X - \mu}{\sigma}$$

That is, if the mean μ is subtracted from X, and this difference is divided by σ, the result is to produce a new variable Z. This Z is the same Z as encountered earlier, that is, it has mean 0 and standard deviation 1.

This particular transformation of Z is referred to as standardization, and is widely used in statistical work. Standardization refers to the process of taking a variable with any mean μ and any standard deviation σ and producing a variable with a mean of 0 and a standard deviation of 1. The variable with the mean of 0 and standard deviation of 1 is referred to as a **standardized variable**. If the distribution of X is normal, then the corresponding standardized variable is given the symbol Z, and this is the standardized normal variable of the last section.

It should be possible to see that

$$\frac{X - \mu}{\sigma}$$

has a mean of 0. Note that the mean of X is μ, so that when μ is subtracted from each value of X, the mean of these differences will be 0, with the negative and positive deviations cancelling. This is essentially the same as the property that

$$\sum(X_i - \bar{X}) = 0$$

as shown in Chapter 5, except that $\bar{X}$ is replaced by the true mean μ. The proof that the standard deviation of the above expression is 1 requires a little more algebra than has been introduced in this textbook. However, this property can be recognized in an intuitive manner by noting that each difference $X - \mu$ is divided by the standard deviation σ. Since the standard deviation of X is σ, this is like dividing σ by σ, producing a value of 1.

The transformation can be used to go in the other direction as well. Solving the first transformation for X gives the following:

$$Z = \frac{X - \mu}{\sigma}$$

$$Z\sigma = \frac{X - \mu}{\sigma}\sigma$$

$$X = \mu + Z\sigma$$

$$X = \mu + Z\sigma$$

The latter expression is the important one. In that expression, if Z is given, and if μ and σ are known, then the value of X can be calculated. This expression will be used when areas or probabilities under the normal distribution are given first, and the values of X that corresponds to these are required. From the original areas which are specified, the Z value can be obtained. Then the values of X can be determined by using

$$X = \mu + Z\sigma$$

The use of these transformations should become clearer in the following examples.

Example 6.4.5 Distribution of Socioeconomic Status

The Regina Labour Force Survey asked respondents their occupation. The Blishen scale of socioeconomic status was used to determine the socioeconomic status of each occupation. The Blishen scale is an interval level scale with a range of approximately 100. It is based on a combination of the education level and income of the occupation of the respondent. A low number on the scale indicates low socioeconomic status, and a high value indicates an occupation with high status or prestige. The mean socioeconomic status of those surveyed was 48 and the standard deviation was 14. Assuming that socioeconomic status is a normally distributed variable, determine

1. The proportion of the population with socioeconomic status of 65 or more.

2. The percentage of the population with socioeconomic status of less than 40.

3. The proportion of the population with socioeconomic status between 40 and 65.

4. The socioeconomic status so that only 10% of the population has greater socioeconomic status.

5. The 35th percentile of socioeconomic status.

Let X represent the distribution of socioeconomic status (SES). The distribution of X is normal with $\mu = 48$ and $\sigma = 14$, that is,

$$X \text{ is } \text{Nor}(48, 14).$$

Figure 6.13 illustrates this distribution diagramatically, with SES along the horizontal axis, and the mean at 48. The first three parts of this question use this figure. In obtaining the required proportions and percentages, it is useful to place both the X and Z values on the same horizontal axis. The normal curve is the same in both situations, all that the transformation from X to Z really does is change the units on the horizontal axis. When the horizontal axis is X, the units are units of socioeconomic status, with the mean status being at $X = 48$. Transforming the X into Z transforms the horizontal axis into Z values. Since the standard deviation of Z is 1, these Z values are also represent the number of standard deviations for centre. The following description shows how the transformation can be used to determine the required probabilities.

1. The proportion of the population with SES of 65 or more is represented by the area under the normal curve in Figure 6.13 that lies to the right of an SES of 65. This area can be determined by calculating the Z associated with $X = 65$, and then using the normal table. Since

   $$Z = \frac{X - \mu}{\sigma}$$

   for $X = 65$,

   $$Z = \frac{65 - 48}{14} = \frac{17}{14} = 1.21$$

   to two decimal places. In Appendix **??**, when $Z = 1.21$, the area in column B is 0.1131. This is the area under the normal distribution that lies to the right of $Z = 1.21$ or to the right of $X = 65$. Thus the proportion of the population with SES of 65 or more is 0.1131.

   Note that SES of 65 is associated with $Z = 1.21$. This is equivalent to saying that SES of 65 is 1.21 standard deviations above average. There is thus a proportion 0.1131 of the population with SES of 1.21 standard deviations or more above the mean.

2. The percentage of the population with SES of less than 40 is associated with an X to the left of centre, as indicated in Figure 6.13. For $X = 40$,

Figure 6.13: Distribution of SES

the Z value is

$$Z = \frac{X - \mu}{\sigma} = \frac{40 - 48}{14} = \frac{-8}{14} = -0.57$$

Note that Z is negative here, since this value is to the left of centre. Since the curve is symmetrical, the required area under the curve to the left of $Z = -0.57$ is the same as the area under the curve to the right ot $Z = +0.57$. From column B of Appendix **??** this area is 0.2843. The percentage of the population with SES of less than 40 is thus $0.2843 \times 100\% = 28.43\%$, or rounded to the nearest tenth of a percentage point, this is 28.4% of the population.

3. The proportion of the population between SES of 40 and 65 can be quickly determined on the basis of the Z values already calculated. The required proportion is the area under the normal curve between $X = 40$ and $X = 65$. This is equal to the area between $X = 40$ and the centre plus the area between the centre of the distribution and $X = 65$. Since $X = 40$ is associated with $Z = -0.57$, column A of Appendix **??** shows that the area between this value and centre is 0.2157. The area between centre and $X = 65$ or $Z = 1.21$ is 0.3869. The required area is

Figure 6.14: 90th Percentile of SES

thus $0.2157 + 0.3869 = 0.6026$. Just over six tenths of the population have SES between 40 and 65.

4. The SES so that only 10% of the population have greater status is given on the right side of Figure 6.14. If only 10% of the population have greater status, then this is equivalent to the upper 10% of the distribution, or the upper 0.10 of the area under the normal curve. What is given here is the area under the curve, and from this, the X value of SES must be determined. The first step in doing this is to recognize that when an area is given, the Z value associated with this area can be determined. Since the area given is in the tail of the distribution, column B of Appendix **??** is used to find this area. The Z value which comes closest to producing an area of 0.1000 in column B is $Z = 1.28$. This is considerably closer to the required area than the next Z of 1.29. This means that only 10% of the population have status more than 1.28 standard deviations above the mean.

   The next step is to determine the exact SES associated with this. This is done by using the reverse transformation, going from Z to X. It was shown that

$$X = \mu + Z\sigma$$

and since the mean is $\mu = 48$ and the standard deviation is $\sigma = 14$,

$$X = \mu + Z\sigma = 48 + (1.28 \times 14) = 48 + 17.92 = 65.92$$

Rounding this off to the nearest integer, the SES so that only 10 per cent of the population has greater status is 66. Note that this is the 90th percentile of SES. That is, P_{90} is the value of SES so that 90% of the population has lower status, and only 10% have greater status. Thus $P_{90} = 66$ in this distribution.

5. Similar considerations are used to obtain the 35th percentile. This is the value of SES so that 35%, or 0.3500 of the distribution is less than this. Looking down column B in Appendix **??**, the Z which comes closest to producing an area of 0.3500 in the tail of the distribution is $Z = 0.39$, although $Z = 0.38$ is almost as close. Since this is to the left of centre, the appropriate Z is $Z = -0.39$. Using the reverse transformation to find the value of X gives

$$X = \mu + Z\sigma = 48 + (-0.39 \times 14) = 48 - 5.46 = 42.54$$

Note the importance of remembering Z is negative in this case, because the required SES is below the mean. The 35th percentile of SES is 42.5, or 43. That is $P_{35} = 43$.

Example 6.4.6 Graduate Record Examination Test Scores

The verbal test scores for U.S. college seniors taking the Graduate Record Examinations (GRE) in Clinical Psychology and Computer Science over the years 1984-1987 are contained in Table 6.22.

1. For each of the two groups, assume the distribution of test scores is normal. Use the mean and standard deviation listed to work out the percentage of students who would be expected to have test scores in each interval, based on the normal curve.

2. Compare the percentages you have with the percentages from the respective frequency distributions in Table 6.22. Explain whether the actual test scores appear to be normally distributed or not.

	Distribution of Test Scores (Per Cent by Discipline)	
Test Score	Clinical Psychology	Computer Science
300 or less	1.5	8.2
300-400	12.9	20.0
400-500	34.2	25.9
500-600	32.9	24.6
600-700	15.1	15.0
700 and over	3.4	6.3
Number of Students	15,169	16,593
Mean Test Score	502	482
Standard Deviation	102	133

Table 6.22: GRE Test Scores

3. Assuming the test scores are normally distributed, what test score should a student taking the Clinical Psychology test obtain in order to assure that he or she is in the 85th percentile?

Answer.

1. For Clinical Psychology, the normal distribution has $\mu = 502$ and $\sigma = 102$, and for Computer Science, $\mu = 482$ and $\sigma = 133$. For each of the values 300, 400, etc., Z values must be calculated in order to determine the appropriate areas and percentages. For example, for Clinical Psychology, the Z value for $X = 300$ is

$$Z = \frac{X - \mu}{\sigma} = \frac{300 - 502}{102} = -1.98.$$

and for $X = 400$ is

$$Z = \frac{X - \mu}{\sigma} = \frac{400 - 502}{102} = -1.00.$$

Using this same method, Table 6.23 gives the values of Z for each of the X values in Table 6.22. In addition, based on column A of

Appendix **??**, this table gives the areas between each Z and the mean of the normal distribution.

	Psychology		Computer Science	
X	Z	Area	Z	Area
300	-1.98	0.4761	-1.37	0.4147
400	-1.00	0.3413	-0.62	0.2324
500	-0.02	0.0080	0.14	0.0557
600	0.96	0.3315	0.89	0.3133
700	1.94	0.4738	1.64	0.4495

Table 6.23: Z Values and Areas for GRE Test Scores

For Clinical Psychology, the area below $X = 300$ is the area below $Z = -1.98$, and from column B of Appendix **??**, this is 0.0239. The area between $X = 300$ and $X = 400$ is the area between $Z = -1.98$ and $Z = -1.00$. From Table 6.23, this is $0.4761 - 0.3413 = 0.1348$. Similarly, the area between $X = 500$ or $Z = -0.02$ and $X = 400$ or $Z = -1.00$ is $0.3413 - 0.0080 = 0.3333$.

The area between $X = 500$ and $X = 600$ is the same as the area between $Z = -0.02$ and $Z = 0.96$. These two areas are added to produce this total area, so this is $0.0080 + 0.3315 = 0.3395$. These areas must be added together because they are on opposite sides of the centre. The area between $X = 600$ or $Z = 0.96$ and $X = 700$ or $Z = 1.94$ is $0.4738 - 0.3315 = 0.1423$. Finally, the area above $X = 700$ is the area above $Z = 1.94$. From Appendix **??** this is 0.0262.

For the Computer Science distribution, the method is the same. The area below $X = 300$ is the area below $Z = -1.37$, and from Appendix **??**, this is 0.0853. The area between $X = 400$ or $Z = -0.62$ and $X = 300$ or $Z = -1.37$ is $0.4147 - 0.2324 = 0.1823$. The area between $X = 400$ or $Z = -0.62$ and $X = 500$ or $Z = 0.14$ is $0.2324 + 0.0557 = 0.2881$. The area between $X = 500$ or $Z = 0.14$ and $X = 600$ or $Z = 0.89$ is $0.3133 - 0.0557 = 0.2576$. The area between $X = 600$ or $Z = 0.89$ and $X = 700$ or $Z = 1.64$ is $0.4495 - 0.3133 = 0.1362$. The area above $X = 700$ or $Z = 1.64$ is 0.0505.

All these areas under the normal curve are converted into percentages and are reported in Table 6.24. There the percentages from the normal distribution can be compared with the actual percentages.

Distribution of Test Scores
(Per Cent by Discipline)

	Clinical Psychology		Computer Science	
Test Score	Actual	Normal	Actual	Normal
300 or less	1.5	2.4	8.2	8.5
300-400	12.9	13.5	20.0	18.2
400-500	34.2	33.3	25.9	28.8
500-600	32.9	34.0	24.6	25.8
600-700	15.1	14.2	15.0	13.6
700 and over	3.4	2.6	6.3	5.1

Table 6.24: GRE Test Scores - Actual and Normal

2. Based on the distributions in Table 6.24, it can be seen that both the actual Clinical Psychology and Computer Science test scores are very similar to the percentages obtained from the normal distribution. As a result, it could be claimed that the actual test scores are normally distributed. Later in the textbook, the chi square test will be used to test the hypothesis that the grades are normally distributed. For now, close inspection of the table shows that the actual test scores are very close to normal.

   In terms of the differences that do appear, for Clinical Psychology, the actual scores have only $12.9 + 1.5 = 14.4\%$ below 400, while the normal curve would indicate there should be $2.4 + 13.5 = 15.9\%$. The normal curve has fewer above 600 (16.8%) than the 18.5% who actually scored above 600. There are also minor differences in the middle of the distribution.

   For Computer Science, there are more test scores in the 300-400 range (20.0%) than in this range in the normal curve (18.2%). Also, there are considerably more test scores above 600 (21.3%) than indicated by the normal distribution (18.7%). As a result, there are fewer test scores in the middle range, between 400 and 600, than in the case of the normal distribution.

3. In order to be in the 85th percentile in Clinical Psychology, the Z value would be 1.03 or 1.04. That is, when $Z = 1.03$, this leaves 0.1515 of the area above it, so that 0.8485, or just under 85% of the area under the curve is below this. For $Z = 1.04$, there is 0.1492 of the area above this, or 0.8508, just over 0.85 below this. The latter is a little closer to 0.85 or 85%, so the 86th percentile can be considered to be at $Z = 1.04$. Since $\mu = 502$ and $\sigma = 102$, for this Z,

$$X = \mu + Z\sigma = 502 + (1.04)(102) = 502 + 106.08 = 608.08$$

In order to ensure being in the 85th percentile, it would be necessary to obtain a test score of 608 or 609 in the Clinical Psychology Graduate Record Examination.

6.4.5 Conclusion

The normal distribution can be used to describe the distribution in actual populations, as in Example 6.4.6. Test scores, such as the GRE, LSAT or other standardized tests, are often assumed to be normally distributed. The implication of this is that most of those who take such tests will obtain results near the mean. There will be some who score very high, more than 2 or 3 standard deviations above the mean. However, there will be very few of these. At the other end, there will always be a few who do very poorly. Some instructors even take this normal distribution so seriously that grades are assigned on the basis of a normal distribution. This may be justified when there are an extremely large number of students in a class. But in general, anyone using the normal distribution should be very cautious concerning this form of application. Unless there is strong evidence that distributions of actual populations are normal, it may be best to assume that actual distributions are not normal.

Where the normal distribution can be more legitimately used is in sampling. It will be seen in Chapter 7 that, under certain circumstances, the normal distribution describes the sampling distribution of sample means. That is, if several random samples are taken from a population, each having a large sample size, the distribution of these sample means is very close to being normally distributed. This happens regardless of the nature of the population from which the sample is drawn. In the next section, it will be seen that the normal distribution can also be used to provide approximations of normal probabilities. It is these latter uses which are more legitimate, and in this textbook, these are the main uses for the normal distribution.