5.2: The Normal Distribution Curve

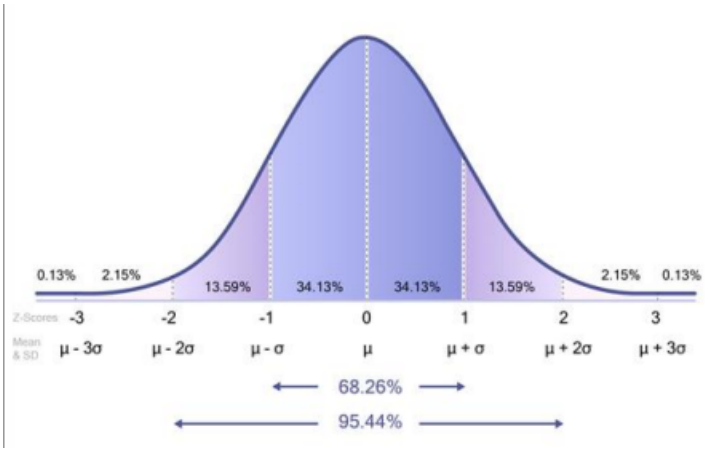

Figure 1.

You might have heard someone refer to data as “normal” before and assumed it meant nothing was unusual or problematic about the data. This is partially true but we must distinguish between what the word “normal” means in general compared to in the field of statistics, specifically. In general, when people say normal, they often mean some version of “This is as expected” or sometimes they mean “This isn’t weird.” The latter can have a negative tone when someone is using the word normal to mean good and abnormal to mean bad. However, for a statistician, normal and abnormal are not synonymous with good or bad. Instead, in statistics, saying data are normal means that the data follow a specific, expected pattern that is visually represented by the normal distribution.

The normal distribution is a probability graph which is commonly referred to in statistics. A probability graph is one which is used to represent how common (likely) or rare (unlikely) various scores are. You will notice that the word “scores” was used; this is because the normal distribution is used to represent the probabilities of various scores or score intervals for quantitative variables rather than qualitative variables. This graph is created univariately meaning only one variable is expected to be represented with the graph at a time, though multiple normal curves can be shown together to allow for comparisons of different variables or groups (which we will see in our later chapters about independent samples *t*-tests and ANOVA).

Key Features of the Normal Curve

There are key features of the normal distribution that make it easy to visually distinguish from other graphs: the peak, asymptotic nature, and symmetry of the graph. The normal curve is essentially a frequency polygon which is tallest (peaks) at the center and gets progressively shorter as you move further into the tails. The word “tails” refers to the outer portions of the graph where the curve takes on a noticeably flatter (more horizontal) appearance or slope. The height of the curve corresponds to the y-axis and represents the frequency with which scores occurred. This means that scores at the middle of the distribution (represented as the apex of the curve) are those that are most frequently occurring. This also means that scores become less common as we move farther from the middle in either direction. Another way to say this is that scores are more common in the tall parts of the graph and less common in the shorter (tail) parts of the graph. When data are perfectly normally distributed, the mode, median, and mean will be the same number and will be on the x-axis directly under the apex or peak of the curve. The normal distribution is also symmetrical. This means that the left side of the graph is a mirror image of the right side of the graph.

The Structure of the Normal Curve

The Axes. The x-axis of the normal curve is used to show quantitative raw scores going from lower scores on the left to higher scores on the right. Because the graph is asymptotic, it theoretically extends to infinity in both directions, however, it would be impossible to actually draw the graph out to infinity. Therefore, the graph usually depicted extending only far enough to show that the tails are getting quite close to the x-axis. Essentially, then, the section of x-axis that is usually shown represents the most relevant segment of a number line which extends from -3 to 3. The y-axis is used to represent the frequencies with which each score occurred with the theoretically lowest possible y-value being 0 to represent the absence of a particular score. These are the same things which are represented on the x-axes and y-axes of histograms and frequency polygons (which were covered in Chapter 2). In fact, you might notice that the normal curve is actually just a very smooth frequency polygon!

The Position of the Mean and Standard Deviation

The mean and standard deviation are used to define and differentiate areas of the normal curve. The mean falls on the x-axis directly under the center, or peak, of the curve, causing half of the graph to be to the left of the mean and the other half to be to the right of the mean. This means that, proportionally, half of the raw scores are lower than the value of the mean (corresponding to the area to the left of the curve's center) and half of the raw scores are greater than the value of the mean (corresponding to the area to the right of the curve's center).

The standard deviation (*SD*) is then used to distinguish various locations to the left or right of the center of the normal curve. Recall that the standard deviation refers to how far individual raw scores within a sample tended to fall from their sample mean (see Chapter 4 for review). Thus, the standard deviation is a way of describing how much individual scores for a variable

tended to differ from the mean. This is the same way it is used in the normal curve. The mean is the center and standard deviations are subtracted from the mean as we move to the left (meaning we are moving to scores that deviate increasingly more by being lesser than the mean) and added to the mean as we move to the right (meaning we are moving to scores that deviate increasing more by being greater than the mean).

Seven specific locations are generally marked along the x-axis using the mean and *SD*. These locations, from left to right, represent scores that are 3 *SD*s below the mean, 2 *SD*s below the mean, 1 *SD* below the mean, at the mean (i.e. 0 *SD*s away from the mean), 1 *SD* above the mean, 2 *SD*s above the mean, and 3 *SD*s above the mean. Scores which are more than 3 standard deviations from the mean are extremely rare in the normal curve and, thus, it is often sufficient to only show these 7 specific locations to accommodate the vast majority raw scores on the x-axis.

These seven locations offer convenient markers to help us see that scores at the mean are most common and that scores become less common the more standard deviations they are away from the mean. For example, when data are normally distributed, scores at the mean are the most frequently occurring. A raw score that is one standard deviation lower than the mean is less frequently occurring than a raw score which is at the mean; however, a raw score that is one standard deviation below the mean is just as frequently occurring as a raw score which is one standard deviation above the mean. This is because the decrease in height on each side of the graph is symmetrical to the other. Further, a raw score which is two standard deviations below the mean is more common than a raw score that is three standard deviations below the mean, is just as common as a raw score which is two standard deviations above the mean, and is less common that raw scores that are one standard deviation below the mean, those which are at the mean, and those which are one standard deviation above the mean.

Reading Review 5.1

Try to answer the following questions regarding the Normal Curve:

1. What are the three distinguishing features of the Normal Curve?
2. What is represented on the y-axis of the Normal Curve?
3. What is represented on the x-axis of the Normal Curve?
4. Where is the mean found in the Normal Curve?
5. What are the seven positions which are usually shown on the x-axis of the Normal Curve?

Using z-Scores for the x-Axis

You may have noticed that this section focused heavily on where the mean and standard deviations are in the normal curve and how to identify those locations. Because of this focus, statisticians often use something known as z-scores to represent the locations of the mean and the number of standard deviations various raw scores are from the mean.

Raw scores are often converted to z-scores to make their location on the x-axis and their corresponding probabilities easier to understand and compare. A z-score tells us how many standard deviations a raw score is from the sample mean. If a raw score is equal to the mean it does not deviate from the mean at all and, thus, its z-score would be equal to 0. This raw score would be at the center of the curve. If you look at Figure 2, you will see that there are two rows of labels under the x-axis: one refers to the location of the mean and raw scores and the other refers to their corresponding z-scores. Thus, z-scores align with, and indicate, how many standard deviations each raw score is from the mean. Notice how much easier it is to say "corresponds to the z-score of 1" compared to how cumbersome it is to say "corresponds to a raw score which is one standard deviation above the mean." These actually refer to the same locations under the curve yet using z-scores to identify locations is much simpler. For this, and other reasons, statisticians prefer to use z-scores to refer to locations under the curve.

Take a moment to compare the two rows of information under the x-axis to familiarize yourself with what z-scores indicate. When a score is equal to the mean, it is at the center and labelled as a z-score of 0. When a score is below the mean, it is to the left and gets a negative z-score. The negative sign indicates that the corresponding raw score for that location is lower in value than the mean. The value of the z-score indicates how many standard deviations the score is from the mean. Therefore, when z = -1 it refers to where a raw score which is one standard deviation less than the mean is located in the normal curve. When z = -2 it refers to where a raw score which is two standard deviations less than the mean is located in the normal curve. Consistent with this, when a score is above the mean, it is to the right and gets either a positive sign or no sign (because the absence of a sign also indicates that a value is positive in mathematics). Therefore, when z = 1 it refers to where a raw score which is one standard deviation greater than the mean is located in the normal curve. When z = 2 it refers to where a raw score which is two standard deviations greater than the mean is located in the normal curve, and so on.

Remember that the normal curve is just a polygon for a population histogram and, thus, the height of the line reflects frequencies of occurrences at points along the x-axis. In the normal distribution, scores nearest the mean are more common and occurrences are expected to get less and less common the further we move away from the mean along the x-axis. This corresponds to the probability that each person who is randomly selected will have a score at various places along the x-axis.

Probability and the Normal Curve

The area under the curve at any given place or section is equal to the probability that a score is in that area. Therefore, the likelihoods that scores exist is computed using their corresponding areas under the curve. Some of these probabilities are easy to see while others are a bit more challenging. For example, we know that the mean, median, and mode are all at the center which divides the graph symmetrically in half. Therefore, we know that half of the raw scores are expected to be to the left of mean and the other half are expected to be to the right of the mean. Because the tails of the graph are asymptotic, meaning they get progressively closer to the x-axis without ever touching or crossing it, we know that 100% of scores are expected to be somewhere under the curve. Think about it this way: If the tails are asymptotes that, theoretically, can continue out to negative and positive infinity, every value of the number line is being represented on the x-axis. Therefore, every possible value of X (and every possible case when measuring X) is somewhere in the curve. This would mean that 100% of all values of X are somewhere in the curve. However, finding the probabilities of other segments of the curve requires us to consider the sloping nature of the curve, the role of the standard deviation, and/or the corresponding z scores.

The standard deviation is an important part of the normal distribution. When data are normally distributed, 34.13% of cases are expected to fall within one standard deviation below the mean to the mean. When data are normally distributed, 34.13% of cases are expected to fall within one standard deviation above the mean to the mean. This means that 68.26% of the population is expected to be within one standard deviation of the mean (half of whom are expected to be lower than the mean an half of whom are expected to be higher than the mean; see Figure 2).

What happens to the scores that are at the mean?

These are theoretically divided in half such that half of the scores at the mean are treated as being very slightly below the mean and half are treated as being very slightly above the mean when estimating the proportions of scores to the left and right of the mean line.

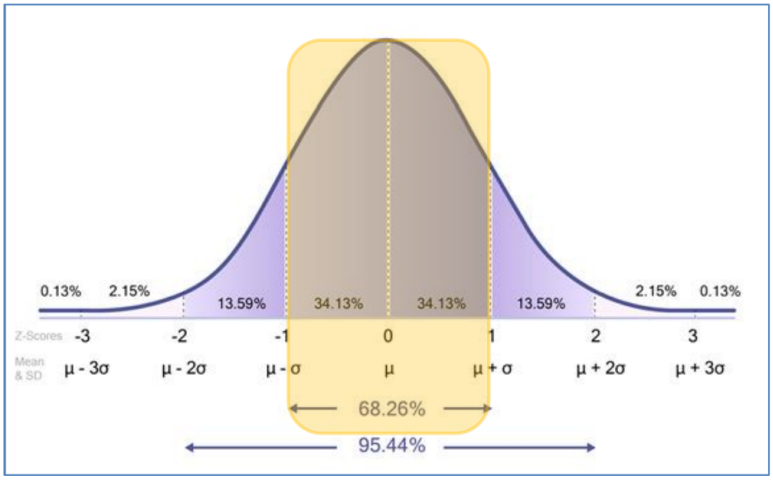

Figure 2.

You can see these two portions of the curve highlighted by the yellow box in Figure 2. If you were to add those two areas within one standard deviation below the mean and one standard deviation above the mean you would get the total of 68.26% that you see reflected on the graph. Note that in the graph, the percentages are rounded to two decimal places.

Reading Review 5.2

Try to answer the following questions using the graph:

1. What proportion of cases are expected to fall below the mean?
2. What proportion of cases are expected to fall above the mean?
3. What proportion of cases are expected to fall between one and two standard deviations below the mean?
4. What proportion of cases are expected to fall between one and two standard deviations above the mean?
