Chapter 5
The Normal Distribution

Objectives

After completing this chapter, you will be able to:

- To understand the topic on Normal Distribution and its importance in different disciplines.
- Describe the characteristics of a standard normal distribution.
- Find the probability of some range of z values in a standard normal distribution.
- Find z scores corresponding to regions under the curve representing a standard normal distribution

INTRODUCTION:

If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed.

The normal, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world.

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them.

The Normal curve

What is the Normal Curve? The normal curve is the beautiful bell shaped curve shown in Figure 1. It is a very useful curve in statistics because many attributes, when a large number of measurements are taken, are approximately distributed in this pattern. For example, the distribution of the wingspans of a large colony of butterflies, of the errors made in repeatedly measuring a 1 kilogram weight and of the amount of sleep you get per night are approximately normal. Many human characteristics, such as height, IQ or examination scores of a large number of people, follow the normal distribution.

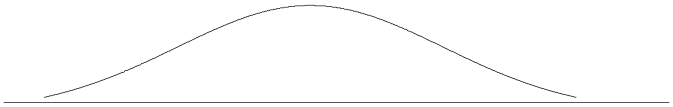


Figure 1: A normal curve.

You may be wondering what is “normal” about the normal distribution. The name arose from the historical derivation of this distribution as a model for the errors made in astronomical observations and other scientific observations. In this model the “average” represents the true or normal value of the measurement and deviations from this are errors. Small errors would occur more frequently than large errors.

The model probably originated in 1733 in the work of the mathematician Abraham Demoivre, who was interested in laws of chance governing gambling, and it was also independently derived in 1786 by Pierre Laplace, an astronomer and mathematician. However, the normal curve as a model for error distribution in scientific theory is most commonly associated with a German astronomer and mathematician, Karl Friedrich Gauss, who found a new derivation of the formula for the curve in 1809. For this reason, the normal curve is sometimes referred to as the “Gaussian” curve. In 1835 another mathematician and astronomer, Lambert Qutelet, used the model to describe human physiological and social traits. Qutelet believed that “normal” meant average and that deviations from the average were nature’s mistakes.

When we draw a normal distribution for some variable, the values of the variable are represented on the horizontal axis called the X axis. We will refer to these values as scores or observations. The area under the curve over any interval represents the proportion of scores in that interval. The height of the curve over an interval from a to b, is the density or crowdedness of that interval; the higher the curve over an interval the more “crowded” that interval. This is illustrated in Figure 2.

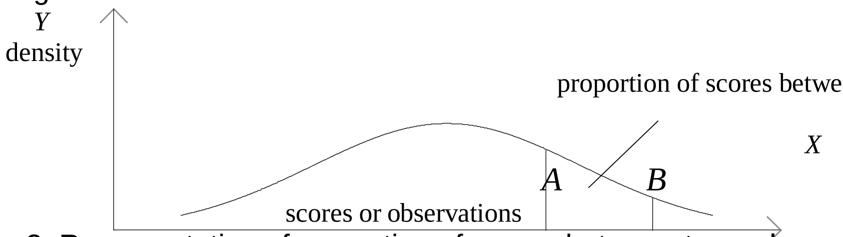


Figure 2: Representation of proportion of scores between two values of variable X.

Can you see where the normal distribution is most crowded or dense?

The scores or observations are most crowded (dense) in intervals around the mean, where the curve is highest. Towards the ends of the curve, the height is lower; the scores become less crowded the further from the mean we go. This tells us that observations around the mean are more likely to occur than observations further from the centre. In a random selection from the normal distribution, scores around the mean have a higher likelihood or probability of being selected than scores far away from the mean.

The normal distribution is not really the normal distribution but a family of distributions. Each of them has these properties:

1. the total area under the curve is 1;
2. the curve is symmetrical so that the mean, median and mode fall together;
3. the curve is bell shaped;
4. the greatest proportion of scores lies close to the mean. The further from the mean one goes (in either direction) the fewer the scores;
5. almost all the scores (0.997 of them) lie within 3 standard deviations of the mean.

The reason for these common properties is that all normal curves are based on the scary looking equation below. If we are measuring values (x) of a variable, such as height, then the distribution of these heights is given by f (x) where

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

This equation does not need to concern us other than to note that it involves μ, the mean of the population, and σ, the standard deviation of the population. The value of the mean fixes the location of the normal curve, where it is centred. In all normal curves half the scores lie to the left of the mean and half to the right. The value of the standard deviation determines the spread; the bigger σ, the more spread out or flat the curve.

Shapes of distributions

Although many variables are approximately normal in distribution, many are not. For example, Figure 5 shows the hypothetical distribution of income for adults in Australia. As you can see this is not symmetrical in shape but has a “tail” of high earners. This is called skewed to the right.

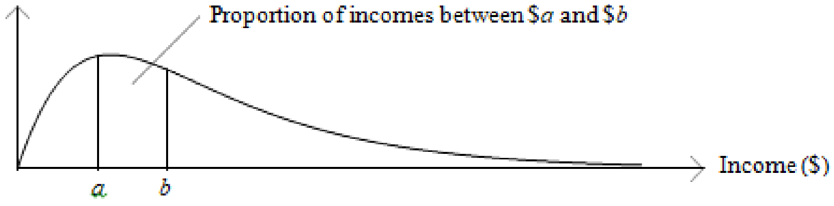


Figure 5: Example of a skewed distribution.

The outcomes of random events also do not necessarily follow the normal curve. For example, if you tossed a die over and over again, the long term pattern of outcomes would be uniform. That is, in theory, each number on the die from 1 to 6 would come up about one sixth of the time. The graph of the outcomes would look something like Figure 6.

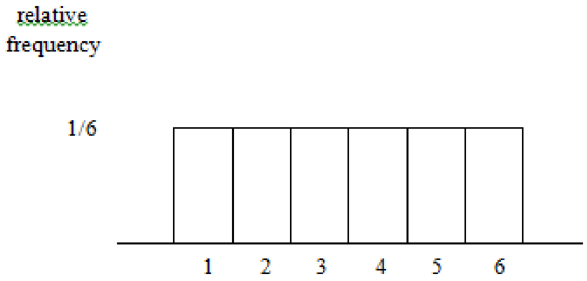


Figure 6: Uniform distribution.

Now here is an amazing fact which explains why the normal curve is so important in statistical investigations. If we take many, many random samples from some population of interest and calculate the sample mean in each case, then the distribution of these sample means will be approximately normal in shape provided the sample size is large.

Central Limit Theorem

Informally, the Central Limit Theorem expresses that if a random variable is the sum of n, independent, identically distributed, non-normal random variables, then its distribution approaches normal as n approaches infinity.

As a consequence of the Central Limit Theorem we have the following corollary: The distribution of the sample mean (X) approaches the normal distribution as the sample size n increases, if the parent distribution from which the samples are drawn is not normal. Let us look at a demonstration of this result. Suppose we have a box containing three tickets marked 1, 2, 3 as illustated in Figure 10.

1	2	3

Figure 10: Box containing tickets marked 1, 2, 3.

If we draw out one ticket at random, record the number then replace the ticket and repeat this process over and over, there would be roughly an equal number of 1s 2s and 3s. Let X = Number on the ticket drawn. This is our parent population. It has a **uniform distribution** which looks something like this.

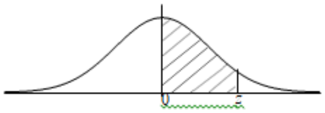
More about finding areas under the standard normal curve

Up to now we have only looked at areas under the normal curve corresponding to 1, 2 or 3 standard deviations above or below the mean. Now we will expand our understanding to a more comprehensive view of areas under the normal curve where the number of standard deviations from the mean may not be whole numbers, for example z = 1.58.

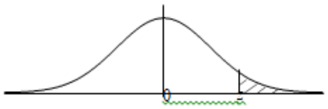
Turn to the end of this booklet to see the table giving areas under the standard normal curve for z scores from 0 to 4.00. Remember that in a standard normal curve the mean is 0 and the standard deviation is 1. Since the normal curve is symmetric we can use the same table to find the areas below the mean corresponding to negative z scores. The purpose of using this table is that we can find probabilities represented by these areas.

This is how the table works. The left hand column shows the z score, that is, the number of standard deviations above the mean. These z scores increase in jumps of 0.01. Notice that this column starts at z = 0 or z = 0.00, that is, the mean itself. The remaining three columns show areas under the normal curve. They are

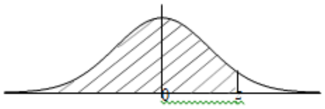
a.
the area between the mean and the z score



b.
the area beyond the z score, called the smaller portion



c.
the area up to the z score, the larger portion.



Remember: The whole area under the curve is 1.

We will start with some examples of finding areas associated with positive and negative z scores and the interpretations of these areas. It is useful to draw a diagram showing the z score and required area.

Note: It is very important that you distinguish between z scores which are represented as points on the horizontal axis and areas under the curve. These areas represent proportions or probabilities.

Transforming raw scores to z scores

In the last chapter we saw that a standard normal curve is well understood and tabulated so that we can find areas associated with intervals of standard scores or z scores.

Furthermore any normally distributed variable, X, can be transformed to a standard normal variable. To do this we shift the mean of the distribution to 0 and shrink or expand the standard deviation to 1. Suppose our population is normally distributed with mean µX and standard deviation σX. To transform raw scores to z scores we must find out how many standard deviations the raw score is from the mean. To see how this is done consider this example.

Let X = score on a nationwide English test. X is normally distributed with µX = 80 and σX = 10. For each student's raw score, termed x, we define the corresponding z score as: z = number of standard deviations from the mean.

Suppose Mike achieved 90 on the test. This is 10 marks above the mean and since the standard deviation is 10, Mike achieved a mark 1 standard deviation above the mean. So the z score for Mike is 1. In short, for x = 90, z = 1.

Video Links:

The Normal Distribution

- https://www.coursera.org/lecture/business-data/2-5-1-normal-distribution-UgvCU
- https://courses.lumenlearning.com/introstats1/chapter/introduction-to-the-normal-distribution/
- https://sites.google.com/a/byron.k12.mn.us/stats4g/elective/normal-distribution

References

- stat.wvu.edu/srs/modules/normal/normal.html
- https://sites.google.com/a/scholarsnyc.com/college-statistics-weekly-assignments/week-by-week-pacing/module-10-the-normal-distribution
- https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module6-RandomError/PH717-Module6-RandomError5.html