7. Normal Distributions

A. Introduction

B. History

C. Areas of Normal Distributions

D. Standard Normal

E. Exercises

Most of the statistical analyses presented in this book are based on the bell-shaped or normal distribution. The introductory section defines what it means for a distribution to be normal and presents some important properties of normal distributions. The interesting history of the discovery of the normal distribution is described in the second section. Methods for calculating probabilities based on the normal distribution are described in Areas of Normal Distributions. A frequently used normal distribution is called the Standard Normal distribution and is described in the section with that name. The binomial distribution can be approximated by a normal distribution. The section Normal Approximation to the Binomial shows this approximation.

Introduction to Normal Distributions

by David M. Lane

Prerequisites
• Chapter 1: Distributions
• Chapter 3: Central Tendency
• Chapter 3: Variability

Learning Objectives
1. Describe the shape of normal distributions
2. State 7 features of normal distributions

The normal distribution is the most important and most widely used distribution in statistics. It is sometimes called the "bell curve," although the tonal qualities of such a bell would be less than pleasing. It is also called the "Gaussian curve" after the mathematician Karl Friedrich Gauss. As you will see in the section on the history of the normal distribution, although Gauss played an important role in its history, de Moivre first discovered the normal distribution.

Strictly speaking, it is not correct to talk about "the normal distribution" since there are many normal distributions. Normal distributions can differ in their means and in their standard deviations. Figure 1 shows three normal distributions. The green (left-most) distribution has a mean of -3 and a standard deviation of 0.5, the distribution in red (the middle distribution) has a mean of 0 and a standard deviation of 1, and the distribution in black (right-most) has a mean of 2 and a standard deviation of 3. These as well as all other normal distributions are symmetric with relatively more values at the center of the distribution and relatively few in the tails.
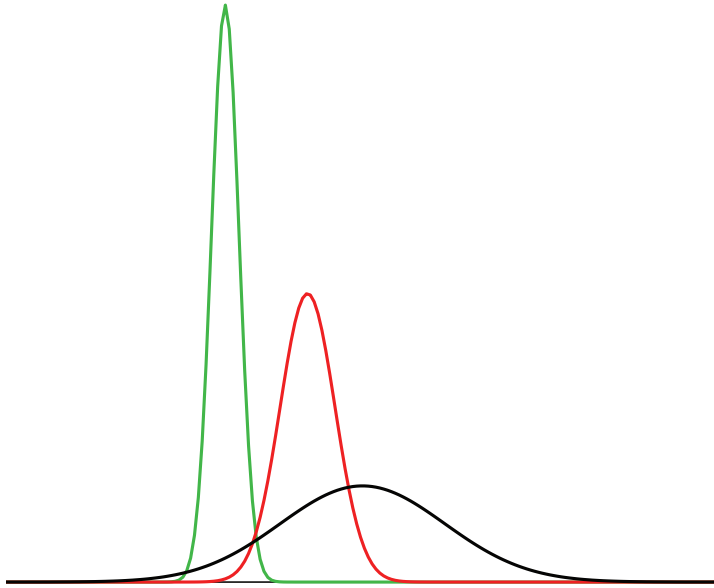
Figure 1. Normal distributions differing in mean and standard deviation.

The density of the normal distribution (the height for a given value on the x-axis) is shown below. The parameters μ and σ are the mean and standard deviation, respectively, and define the normal distribution. The symbol e is the base of the natural logarithm and π is the constant pi.

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

Since this is a non-mathematical treatment of statistics, do not worry if this expression confuses you. We will not be referring back to it in later sections.

Seven features of normal distributions are listed below. These features are illustrated in more detail in the remaining sections of this chapter.

1. Normal distributions are symmetric around their mean.

2. The mean, median, and mode of a normal distribution are equal.

3. The area under the normal curve is equal to 1.0.

4. Normal distributions are denser in the center and less dense in the tails.

5. Normal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).

6. 68% of the area of a normal distribution is within one standard deviation of the mean.

7. Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.

History of the Normal Distribution

by David M. Lane

Prerequisites
- Chapter 1: Distributions
- Chapter 3: Central Tendency
- Chapter 3: Variability
- Chapter 5: Binomial Distribution

Learning Objectives

1. Name the person who discovered the normal distribution and state the problem he applied it to

2. State the relationship between the normal and binomial distributions

3. State who related the normal distribution to errors

4. Describe briefly the central limit theorem

5. State who was the first to prove the central limit theorem

In the chapter on probability, we saw that the binomial distribution could be used to solve problems such as "If a fair coin is flipped 100 times, what is the probability of getting 60 or more heads?" The probability of exactly x heads out of N flips is computed using the formula:

$$P(x) = \frac{N!}{x!\,(N-x)!}\pi^x(1-\pi)^{N-x}$$

where x is the number of heads (60), N is the number of flips (100), and π is the probability of a head (0.5). Therefore, to solve this problem, you compute the probability of 60 heads, then the probability of 61 heads, 62 heads, etc., and add up all these probabilities. Imagine how long it must have taken to compute binomial probabilities before the advent of calculators and computers.

Abraham de Moivre, an 18th century statistician and consultant to gamblers, was often called upon to make these lengthy computations. de Moivre noted that when the number of events (coin flips) increased, the shape of the binomial

distribution approached a very smooth curve. Binomial distributions for 2, 4, 12, and 24 flips are shown in Figure 1.
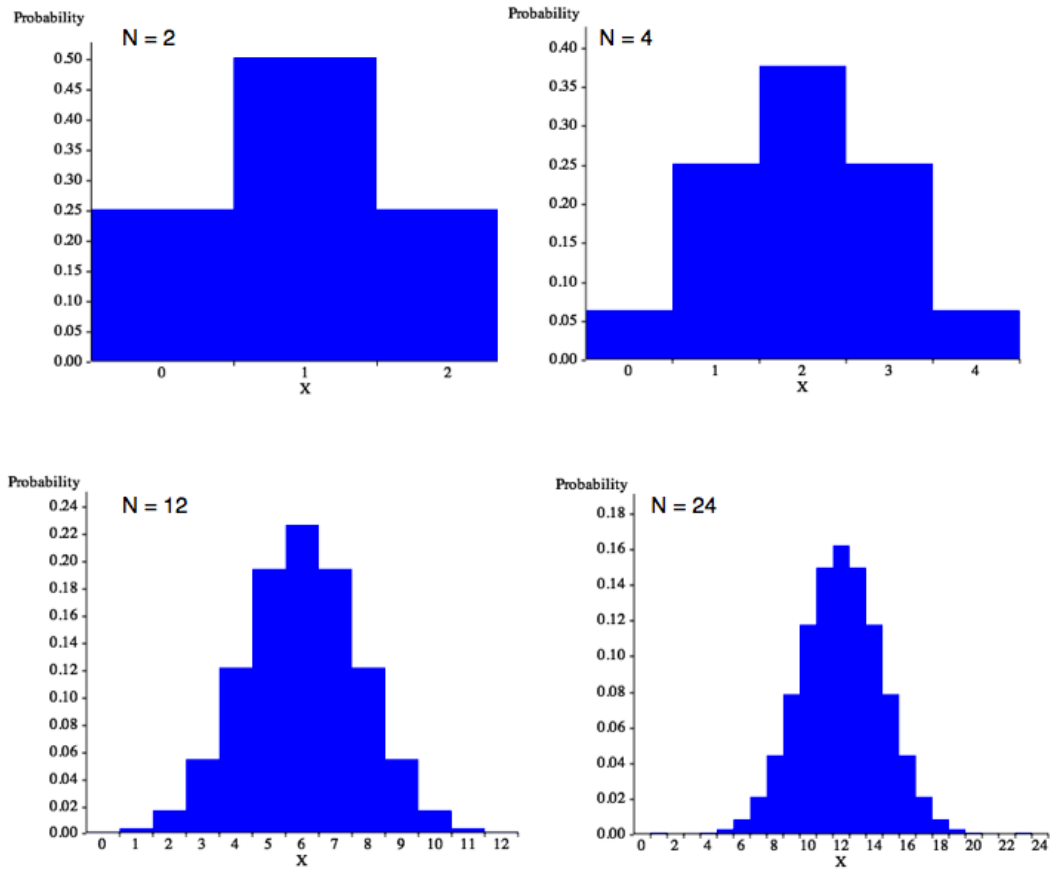

Figure 1. Examples of binomial distributions. The heights of the blue bars represent the probabilities.

de Moivre reasoned that if he could find a mathematical expression for this curve, he would be able to solve problems such as finding the probability of 60 or more heads out of 100 coin flips much more easily. This is exactly what he did, and the curve he discovered is now called the "normal curve."
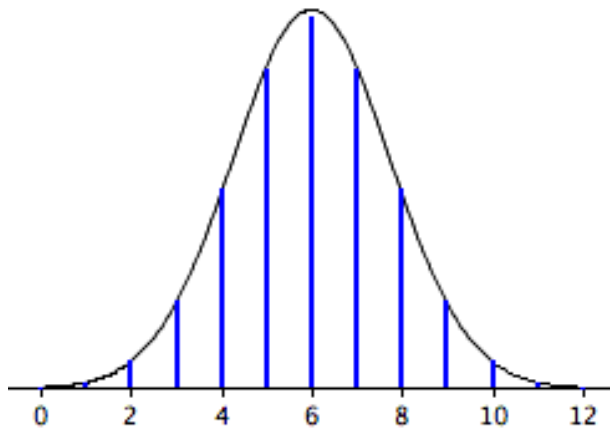
Figure 2. The normal approximation to the binomial distribution for 12 coin flips. The smooth curve is the normal distribution. Note how well it approximates the binomial probabilities represented by the heights of the blue lines.

The importance of the normal curve stems primarily from the fact that the distributions of many natural phenomena are at least approximately normally distributed. One of the first applications of the normal distribution was to the analysis of errors of measurement made in astronomical observations, errors that occurred because of imperfect instruments and imperfect observers. Galileo in the 17th century noted that these errors were symmetric and that small errors occurred more frequently than large errors. This led to several hypothesized distributions of errors, but it was not until the early 19th century that it was discovered that these errors followed a normal distribution. Independently, the mathematicians Adrian in 1808 and Gauss in 1809 developed the formula for the normal distribution and showed that errors were fit well by this distribution.

This same distribution had been discovered by Laplace in 1778 when he derived the extremely important central limit theorem, the topic of a later section of this chapter. Laplace showed that even if a distribution is not normally distributed, the means of repeated samples from the distribution would be very nearly normally distributed, and that the larger the sample size, the closer the distribution of means would be to a normal distribution.

Most statistical procedures for testing differences between means assume normal distributions. Because the distribution of means is very close to normal, these tests work well even if the original distribution is only roughly normal.

Quételet was the first to apply the normal distribution to human characteristics. He noted that characteristics such as height, weight, and strength were normally distributed.

Areas Under Normal Distributions

by David M. Lane

Prerequisites
- Chapter 1: Distributions
- Chapter 3: Central Tendency
- Chapter 3: Variability
- Chapter 7: Introduction to Normal Distributions

Learning Objectives
1. State the proportion of a normal distribution within 1 standard deviation of the mean
2. State the proportion of a normal distribution that is more than 1.96 standard deviations from the mean
3. Use the normal calculator to calculate an area for a given X"
4. Use the normal calculator to calculate X for a given area

Areas under portions of a normal distribution can be computed by using calculus. Since this is a non-mathematical treatment of statistics, we will rely on computer programs and tables to determine these areas. Figure 1 shows a normal distribution with a mean of 50 and a standard deviation of 10. The shaded area between 40 and 60 contains 68% of the distribution.
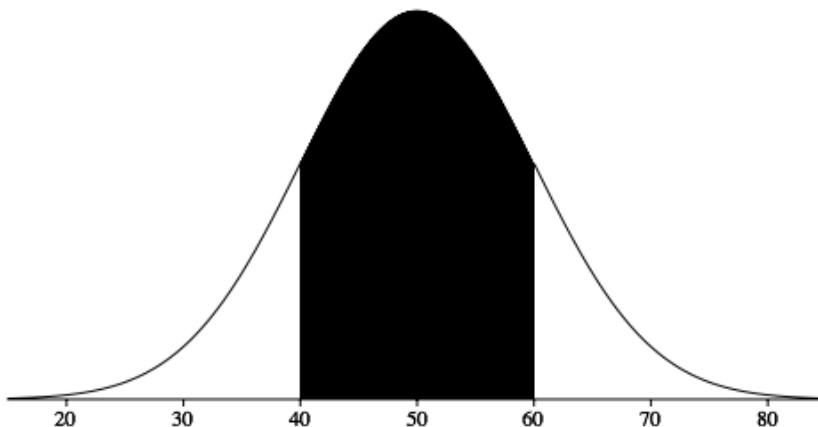

Figure 1. Normal distribution with a mean of 50 and standard deviation of 10. 68% of the area is within one standard deviation (10) of the mean (50).

Figure 2 shows a normal distribution with a mean of 100 and a standard deviation of 20. As in Figure 1, 68% of the distribution is within one standard deviation of the mean.
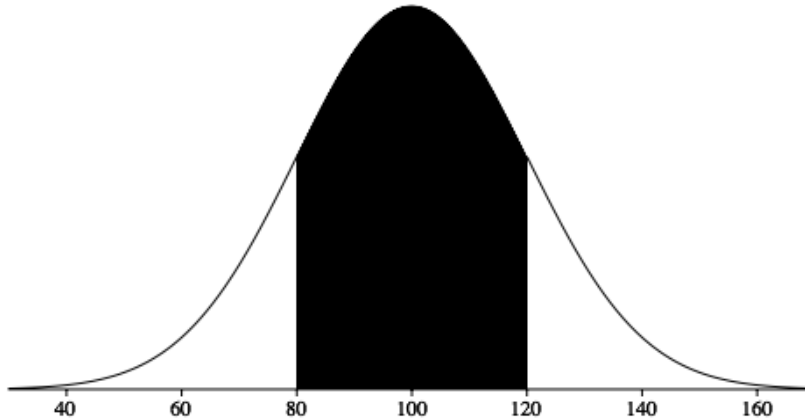

Figure 2. Normal distribution with a mean of 100 and standard deviation of 20. 68% of the area is within one standard deviation (20) of the mean (100).

The normal distributions shown in Figures 1 and 2 are specific examples of the general rule that 68% of the area of any normal distribution is within one standard deviation of the mean.

Figure 3 shows a normal distribution with a mean of 75 and a standard deviation of 10. The shaded area contains 95% of the area and extends from 55.4 to 94.6. For all normal distributions, 95% of the area is within 1.96 standard deviations of the mean. For quick approximations, it is sometimes useful to round off and use 2 rather than 1.96 as the number of standard deviations you need to extend from the mean so as to include 95% of the area.
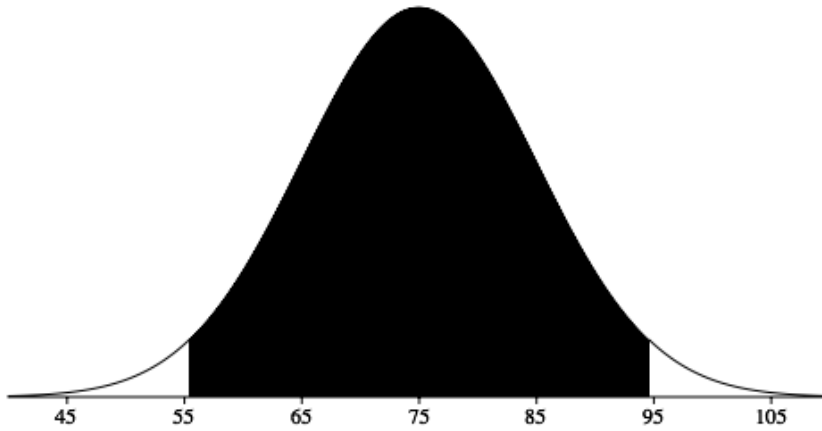
Figure 3. A normal distribution with a mean of 75 and a standard deviation of 10. 95% of the area is within 1.96 standard deviations of the mean.

Areas under the normal distribution can be calculated with this online calculator.

Standard Normal Distribution

by David M. Lane

Prerequisites
- Chapter 3: Effects of Linear Transformations
- Chapter 7: Introduction to Normal Distributions

Learning Objectives
1. State the mean and standard deviation of the standard normal distribution
2. Use a Z table
3. Use the normal calculator
4. Transform raw data to Z scores

As discussed in the introductory section, normal distributions do not necessarily have the same means and standard deviations. A normal distribution with a mean of 0 and a standard deviation of 1 is called a standard normal distribution.

Areas of the normal distribution are often represented by tables of the standard normal distribution. A portion of a table of the standard normal distribution is shown in Table 1.

Table 1. A portion of a table of the standard normal distribution.

Z	Area below
-2.5	0.0062
-2.49	0.0064
-2.48	0.0066
-2.47	0.0068
-2.46	0.0069
-2.45	0.0071
-2.44	0.0073
-2.43	0.0075
-2.42	0.0078
-2.41	0.008
-2.4	0.0082

-2.39	0.0084
-2.38	0.0087
-2.37	0.0089
-2.36	0.0091
-2.35	0.0094
-2.34	0.0096
-2.33	0.0099
-2.32	0.0102

The first column titled "Z" contains values of the standard normal distribution; the second column contains the area below Z. Since the distribution has a mean of 0 and a standard deviation of 1, the Z column is equal to the number of standard deviations below (or above) the mean. For example, a Z of -2.5 represents a value 2.5 standard deviations below the mean. The area below Z is 0.0062.

The same information can be obtained using the following Java applet. Figure 1 shows how it can be used to compute the area below a value of -2.5 on the standard normal distribution. Note that the mean is set to 0 and the standard deviation is set to 1.

Normal Distribution

Mean: 0　　　　Sd: 1

○ Above　　1

◉ Below　　-2.5

○ Between　-1　　and　1
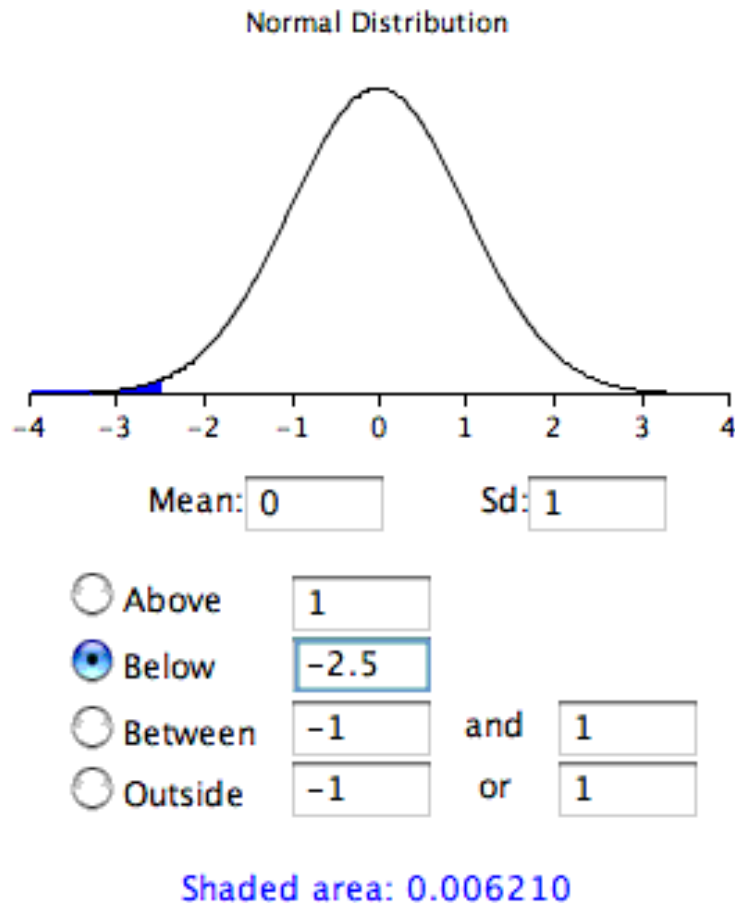
○ Outside　-1　　or　1

Shaded area: 0.006210

Figure 1. An example from the applet.

A value from any normal distribution can be transformed into its corresponding value on a standard normal distribution using the following formula:

$$Z = (X - \mu)/\sigma$$

where Z is the value on the standard normal distribution, X is the value on the original distribution, μ is the mean of the original distribution, and σ is the standard deviation of the original distribution.

As a simple application, what portion of a normal distribution with a mean of 50 and a standard deviation of 10 is below 26? Applying the formula, we obtain

$$Z = (26 - 50)/10 = -2.4.$$

From Table 1, we can see that 0.0082 of the distribution is below -2.4. There is no need to transform to Z if you use the applet as shown in Figure 2.

Normal Distribution

Mean: 50   Sd: 10

○ Above   1
● Below   26
○ Between   -1   and   1
○ Outside   -1   or   1
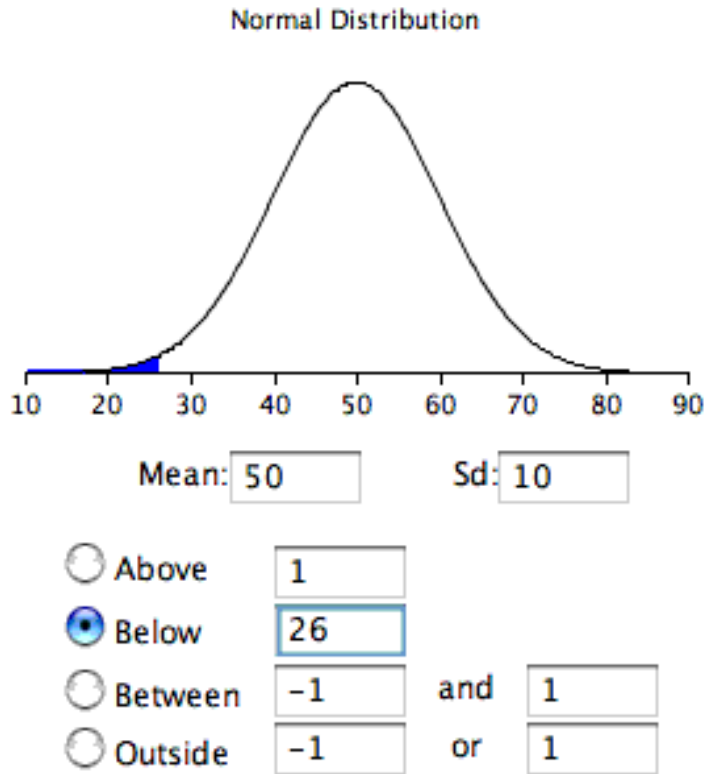
Shaded area: 0.008198

Figure 2. Area below 26 in a normal distribution with a mean of 50 and a standard deviation of 10.

If all the values in a distribution are transformed to Z scores, then the distribution will have a mean of 0 and a standard deviation of 1. This process of transforming a distribution to one with a mean of 0 and a standard deviation of 1 is called standardizing the distribution.

Normal Approximation to the Binomial

by David M. Lane

Prerequisites
• Chapter 5: Binomial Distribution
• Chapter 7: History of the Normal Distribution
• Chapter 7: Areas of Normal Distributions

Learning Objectives
1. State the relationship between the normal distribution and the binomial distribution
2. Use the normal distribution to approximate the binomial distribution
3. State when the approximation is adequate

In the section on the history of the normal distribution, we saw that the normal distribution can be used to approximate the binomial distribution. This section shows how to compute these approximations.

Let's begin with an example. Assume you have a fair coin and wish to know the probability that you would get 8 heads out of 10 flips. The binomial distribution has a mean of $\mu = N\pi = (10)(0.5) = 5$ and a variance of $\sigma^2 = N\pi(1-\pi) = (10)(0.5)(0.5) = 2.5$. The standard deviation is therefore 1.5811. A total of 8 heads is $(8 - 5)/1.5811 = 1.897$ standard deviations above the mean of the distribution. The question then is, "What is the probability of getting a value exactly 1.897 standard deviations above the mean?" You may be surprised to learn that the answer is 0: The probability of any one specific point is 0. The problem is that the binomial distribution is a discrete probability distribution, whereas the normal distribution is a continuous distribution.

The solution is to round off and consider any value from 7.5 to 8.5 to represent an outcome of 8 heads. Using this approach, we figure out the area under a normal curve from 7.5 to 8.5. The area in green in Figure 1 is an approximation of the probability of obtaining 8 heads.
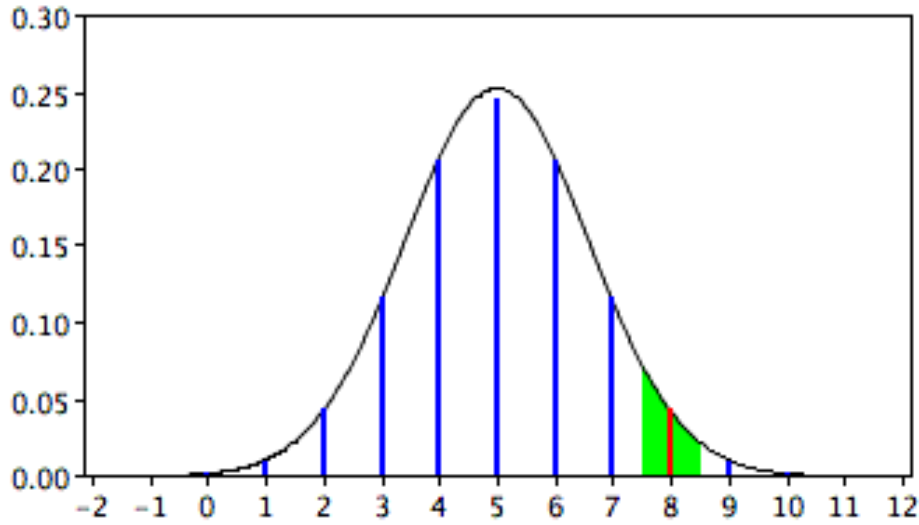
Figure 1. Approximating the probability of 8 heads with the normal distribution.

The solution is therefore to compute this area. First we compute the area below 8.5, and then subtract the area below 7.5.

The results of using the normal area calculator to find the area below 8.5 are shown in Figure 2. The results for 7.5 are shown in Figure 3.
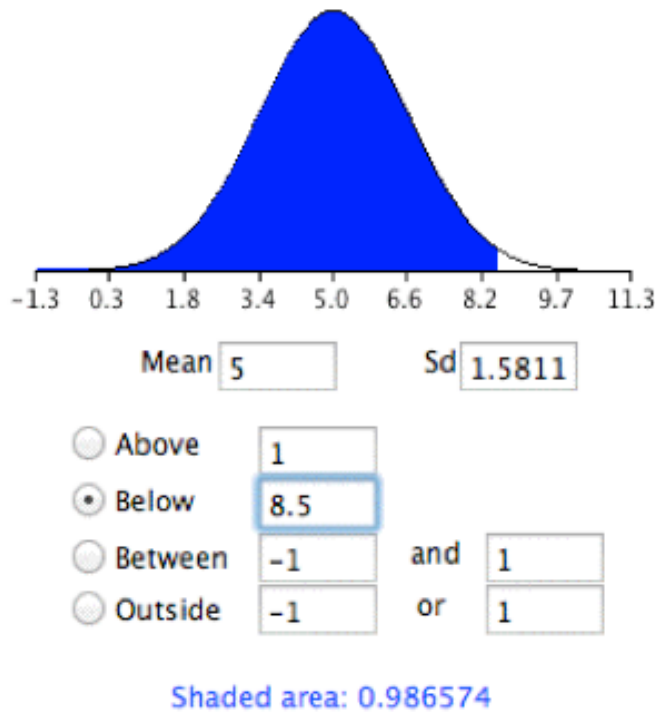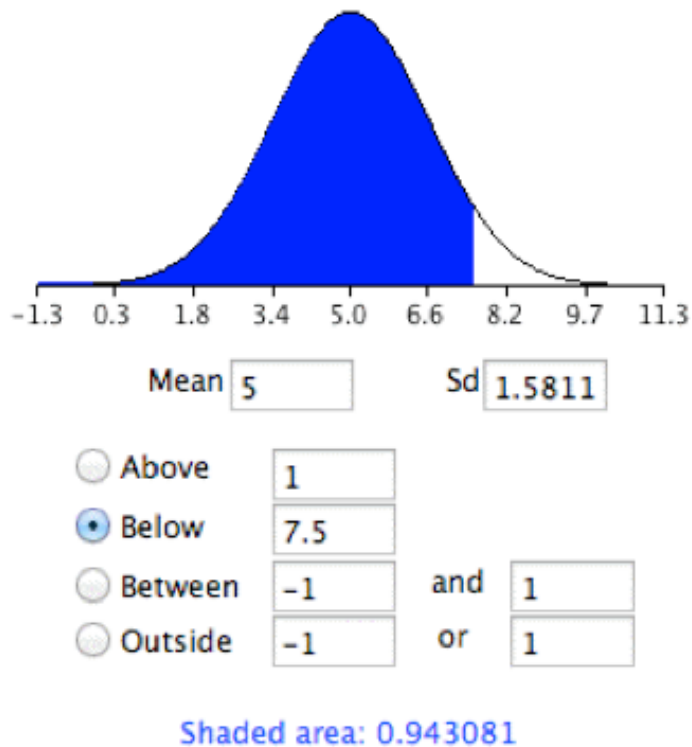

Figure 2. Area below 8.5

Figure 3. Area below 7.5.

The difference between the areas is 0.044, which is the approximation of the binomial probability. For these parameters, the approximation is very accurate. The demonstration in the next section allows you to explore its accuracy with different parameters.

If you did not have the normal area calculator, you could find the solution using a table of the standard normal distribution (a Z table) as follows:
1. Find a Z score for 8.5 using the formula $Z = (8.5 - 5)/1.5811 = 2.21$.
2. Find the area below a Z of $2.21 = 0.987$.
3. Find a Z score for 7.5 using the formula $Z = (7.5 - 5)/1.5811 = 1.58$.
4. Find the area below a Z of $1.58 = 0.943$.
5. Subtract the value in step 4 from the value in step 2 to get 0.044.

The same logic applies when calculating the probability of a range of outcomes. For example, to calculate the probability of 8 to 10 flips, calculate the area from 7.5 to 10.5.

The accuracy of the approximation depends on the values of N and π. A rule of thumb is that the approximation is good if both $N\pi$ and $N(1-\pi)$ are both greater than 10.

Statistical Literacy

by David M. Lane

Prerequisites
- Chapter 7: Areas Under the Normal Distribution
- Chapter 7: Shapes of Distributions

Risk analyses often are based on the assumption of normal distributions. Critics have said that extreme events in reality are more frequent than would be expected assuming normality. The assumption has even been called a "Great Intellectual Fraud."

A recent article discussing how to protect investments against extreme events defined "tail risk" as "A tail risk, or extreme shock to financial markets, is technically defined as an investment that moves more than three standard deviations from the mean of a normal distribution of investment returns."

What do you think?

Tail risk can be evaluated by assuming a normal distribution and computing the probability of such an event. Is that how "tail risk" should be evaluated?

> Events more than three standard deviations from the mean are very rare for normal distributions. However, they are not as rare for other distributions such as highly-skewed distributions. If the normal distribution is used to assess the probability of tail events defined this way, then the "tail risk" will be underestimated.

Exercises

Prerequisites
• All material presented in the Normal Distributions chapter

1. If scores are normally distributed with a mean of 35 and a standard deviation of 10, what percent of the scores is:

   a. greater than 34?

   b. smaller than 42?

   c. between 28 and 34?

2. What are the mean and standard deviation of the standard normal distribution? (b) What would be the mean and standard deviation of a distribution created by multiplying the standard normal distribution by 8 and then adding 75?

3. The normal distribution is defined by two parameters. What are they?

4. What proportion of a normal distribution is within one standard deviation of the mean? (b) What proportion is more than 2.0 standard deviations from the mean? (c) What proportion is between 1.25 and 2.1 standard deviations above the mean?

5. A test is normally distributed with a mean of 70 and a standard deviation of 8. (a) What score would be needed to be in the 85th percentile? (b) What score would be needed to be in the 22nd percentile?

6. Assume a normal distribution with a mean of 70 and a standard deviation of 12. What limits would include the middle 65% of the cases?

7. A normal distribution has a mean of 20 and a standard deviation of 4. Find the Z scores for the following numbers: (a) 28 (b) 18 (c) 10 (d) 23

8. Assume the speed of vehicles along a stretch of I-10 has an approximately normal distribution with a mean of 71 mph and a standard deviation of 8 mph.

   a. The current speed limit is 65 mph. What is the proportion of vehicles less than or equal to the speed limit?

   b. What proportion of the vehicles would be going less than 50 mph?

c. A new speed limit will be initiated such that approximately 10% of vehicles will be over the speed limit. What is the new speed limit based on this criterion?

d. In what way do you think the actual distribution of speeds differs from a normal distribution?

9. A variable is normally distributed with a mean of 120 and a standard deviation of 5. One score is randomly sampled. What is the probability it is above 127?

10. You want to use the normal distribution to approximate the binomial distribution. Explain what you need to do to find the probability of obtaining exactly 7 heads out of 12 flips.

11. A group of students at a school takes a history test. The distribution is normal with a mean of 25, and a standard deviation of 4. (a) Everyone who scores in the top 30% of the distribution gets a certificate. What is the lowest score someone can get and still earn a certificate? (b) The top 5% of the scores get to compete in a statewide history contest. What is the lowest score someone can get and still go onto compete with the rest of the state?

12. Use the normal distribution to approximate the binomial distribution and find the probability of getting 15 to 18 heads out of 25 flips. Compare this to what you get when you calculate the probability using the binomial distribution. Write your answers out to four decimal places.

13. True/false: For any normal distribution, the mean, median, and mode will be equal.

14. True/false: In a normal distribution, 11.5% of scores are greater than $Z = 1.2$.

15. True/false: The percentile rank for the mean is 50% for any normal distribution.

16. True/false: The larger the n, the better the normal distribution approximates the binomial distribution.

17. True/false: A Z-score represents the number of standard deviations above or below the mean.

18. True/false: Abraham de Moivre, a consultant to gamblers, discovered the normal distribution when trying to approximate the binomial distribution to make his computations easier.

Answer questions 19 - 21 based on the graph below:
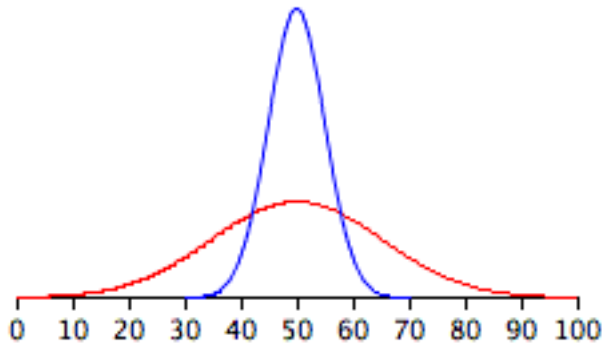
19. True/false: The standard deviation of the blue distribution shown below is about 10.

20. True/false: The red distribution has a larger standard deviation than the blue distribution.

21. True/false: The red distribution has more area underneath the curve than the blue distribution does.

Questions from Case Studies

Angry Moods (AM) case study

22. For this problem, use the Anger Expression (AE) scores.
    a. Compute the mean and standard deviation.
    b. Then, compute what the 25th, 50th and 75th percentiles would be if the distribution were normal.
    c. Compare the estimates to the actual 25th, 50th, and 75th percentiles.

Physicians' Reactions (PR) case study

23. (PR) For this problem, use the time spent with the overweight patients. (a) Compute the mean and standard deviation of this distribution. (b) What is the probability that if you chose an overweight participant at random, the doctor would have spent 31 minutes or longer with this person? (c) Now assume this distribution is normal (and has the same mean and standard deviation). Now what is the probability that if you chose an overweight participant at random, the doctor would have spent 31 minutes or longer with this person?

The following questions are from ARTIST (reproduced with permission)

24. A set of test scores are normally distributed. Their mean is 100 and standard deviation is 20. These scores are converted to standard normal z scores. What would be the mean and median of this distribution?

a. 0

b. 1

c. 50

d. 100

25. Suppose that weights of bags of potato chips coming from a factory follow a normal distribution with mean 12.8 ounces and standard deviation .6 ounces. If the manufacturer wants to keep the mean at 12.8 ounces but adjust the standard deviation so that only 1% of the bags weigh less than 12 ounces, how small does he/she need to make that standard deviation?

26. A student received a standardized (z) score on a test that was -. 57. What does this score tell about how this student scored in relation to the rest of the class? Sketch a graph of the normal curve and shade in the appropriate area.

27. Suppose you take 50 measurements on the speed of cars on Interstate 5, and that these measurements follow roughly a Normal distribution. Do you expect the standard deviation of these 50 measurements to be about 1 mph, 5 mph, 10 mph, or 20 mph? Explain.

28. Suppose that combined verbal and math SAT scores follow a normal distribution with mean 896 and standard deviation 174. Suppose further that Peter finds out that he scored in the top 3% of SAT scores. Determine how high Peter's score must have been.

29. Heights of adult women in the United States are normally distributed with a population mean of $\mu = 63.5$ inches and a population standard deviation of $\sigma = 2.5$. A medical re- searcher is planning to select a large random sample of adult women to participate in a future study. What is the standard value, or z-value, for an adult woman who has a height of 68.5 inches?

30. An automobile manufacturer introduces a new model that averages 27 miles per gallon in the city. A person who plans to purchase one of these new cars wrote the manufacturer for the details of the tests, and found out that the standard deviation is 3 miles per gallon. Assume that in-city mileage is approximately normally distributed.

    a. What is the probability that the person will purchase a car that averages less than 20 miles per gallon for in-city driving?

    b. What is the probability that the person will purchase a car that averages between 25 and 29 miles per gallon for in-city driving?