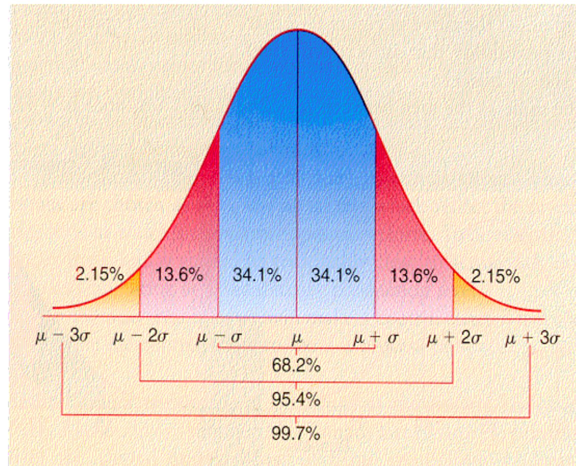
11 – Normal Distributions

The normal distribution is a theoretical *continuous* distribution that should be thought to spawn all frequency distributions that can be characterized as normal: symmetric, bell-shaped, and with mean, median, and mode all equal and in the center of the frequency distribution.

The normal distribution should be thought of as a curve (rather than a histogram).



The normal distribution has many important properties. Here are some that we will use:

- The normal distribution is symmetrically centered at its mean (μ).
- The area under the normal curve that is within one standard deviation of the mean (between $\mu - \sigma$ and $\mu + \sigma$) is 68.2% (or just about 2/3) of the total area under the curve.
- The area under the normal curve that is within two standard deviations of the mean (between $\mu - 2\sigma$ and $\mu + 2\sigma$) is 95.4% of the total area under the curve.
- The area under the normal curve that is within three standard deviations of the mean (between $\mu - 3\sigma$ and $\mu + 3\sigma$) is 99.7% (pretty much all) of the total area under the curve.

The properties are all shown in the image on the right.

> Example 1: The average commute to work in PA is 25.5 minutes, a normally distributed variable with a standard deviation of 5.8 minutes. The properties above indicate that 68.2% of PA workers commute between 19.7 minutes and 31.3 minutes. 95.4% commute between 13.9 and 37.1 minutes, and nearly all PA workers (99.7% of them) commute between 8.1 and 42.9 minutes.

The standard normal distribution is a specific normal distribution: it is the normal distribution with mean equal to 0 and standard deviation equal to 1.

Any normally distributed data can be converted to standard normal by converting each data value to its *z*-score. This value is calculated as follows (from Chapter 3, Notes 5):

> A data point's *z*-score is computed by first subtracting the mean and then dividing the result by the standard deviation. If the data point is *x*, then $z = \frac{x-\bar{x}}{s}$ for sample data or $z = \frac{x-\mu}{\sigma}$ for population data.

The reason we bother converting normally distributed data to standard normal is so that we can use one table for all normal distributions (Table E, on textbook pages 646-647 and the inside front cover) to find the portion of the area that lies beneath the curve to the left or right of a value, or in between values.

Table E give the portion of the area under the standard normal curve that lies *to the left* of a given value by looking up the value at the top and left margins of the table. The following example shows how.

> Example 2, part A: To find the portion of the area beneath the standard normal curve that is less than z = 1.76, we look for 1.7 on the left column (second page of the table for positive *z*-scores) and for 0.06 on the top. The number that appears in the corresponding row and column is the area. In this case, that number is 0.9608. That is, 96.08% of the area under the standard normal curve is to the left of z = 1.76.

Using the fact that the total area is 1 (or 100%), we can indirectly find the area to the right of values.

> Example 2, part B: Since the portion of the area beneath the standard normal curve that is less than z = 1.76 is 0.9608, we can just subtract this number from 1 to get the area to the right. That figure is 0.0392. So, 3.92% of the area under the standard normal curve is to the right of z = 1.76.

Further, we can find the area between values.

Example 2, part C: The area to the left of $z = 1.76$ is 0.9608. Similarly, we can find the area to the left of $z = -0.94$, which is 0.1736. So, the area under the standard normal distribution between $z = -0.94$ and $z = 1.76$ is 0.9608 – 0.1736 = 0.7872. That means that 78.72% of the area under the standard normal curve is between the values $z = -0.94$ and $z = 1.76$.

These areas can have several different interpretations:

- The portion of data points in a normal distribution that are less than, greater than, or in between given values.
- The probability of randomly selecting a value less than, greater than, or in between given values.

Example 3, part I: The average American adult consumes 1.64 cups of coffee in a day with a standard deviation of 0.24 cups. If the distribution of cups of coffee American adults drink each day is known to be normal, then we can use the standard normal distribution find the probability that a randomly selected person drinks less than 2 cups of coffee per day by converting to a *z*-score: $z = \frac{2-1.64}{0.24} = 1.5$. Finding the *z*-score of 1.5 on Table E gives us an area of .9332. So, 93.32 percent of American adults drink 2 or fewer cups of coffee each day. This is also the probability that a randomly selected American adult drinks 2 or fewer cups of coffee each day.

Example 3, part II: To find the portion of American adults who drink more than 3 cups of coffee in a day, we first must compute the *z*-score for 3 cups of coffee: $z = \frac{3-1.64}{0.24} \approx 5.67$. If we look for 5.67 on Table E, it does not appear (the highest *z*-score is 3.49). So, if we use 3.49 as the nearest value to 5.67, the only way to express a solution is that some portion lower than 1 – 0.9998, or fewer than 0.02% of Americans, drink more than 3 cups of coffee each day.

It is often the case that we need to reverse the look-up process in Table E. An example of such a scenario appears below.

Example 4 (adapted from problem 28 in 6 – 2 exercises): An advertising company plans to market a product to lower-income families. A study stated that for a particular area, family income is distributed normally with mean $42,395 and standard deviation $6256. If the company wishes to target the bottom 20% of the families in that area, find the cutoff income that defines that bottom 20%.

Solution: To find the bottom 20% of any normally distributed variable, we start by finding the *z*-score that corresponds to an area of 0.20. Looking up this value within Table E, the number that is closest to it is 0.2005. Using the reverse of the look-up process for Table E, we find that this corresponds to $z = -0.84$. So, to find the value in the income distribution that this corresponds to we must solve the equation

$-0.84 = \frac{x-42395}{6256}$, which has the solution $x =$ $37,140. This is the maximum income a family can earn to be in the bottom 20% of families from that area.