# MODULE 8

# NORMAL DISTRIBUTION

## Contents



A MODEL FOR THE DISTRIBUTION of a single quantitative variable can be visualized by "fitting" a smooth curve to a histogram (Figure 8.1-Left), removing the histogram (Figure 8.1-Center), and using the remaining curve (Figure 8.1-Right) as a model for the distribution of the entire population of individuals. The smooth red curve drawn over the histogram serves as a model for the distribution of the **entire population**. If the smooth curve follows a known distribution, then certain calculations are greatly simplified.

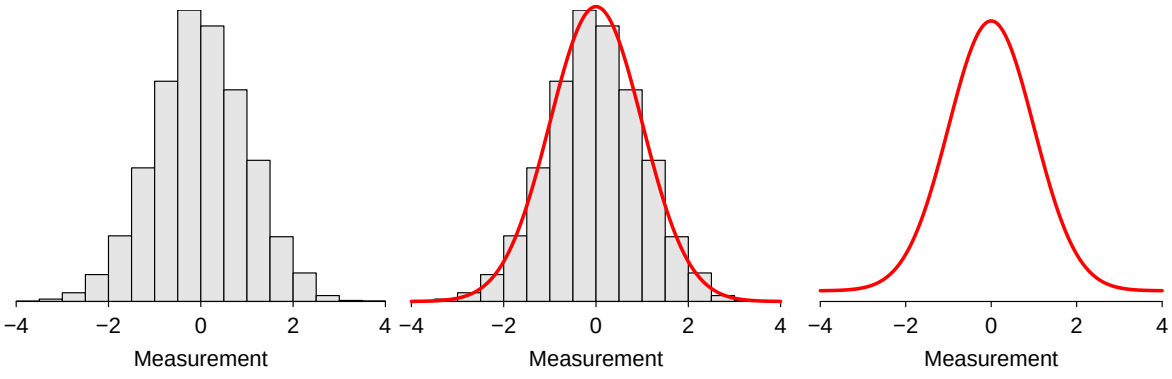


Figure 8.1. Depiction of fitting a smooth curve to a histogram to serve as a model for the distribution.

The normal distribution is one of the most important distributions in statistics because it serves as a model for the distribution of individuals in many natural situations and the distribution of statistics from repeated samplings (i.e., sampling distributions).[1] The use of a normal distribution model to make certain calculations is demonstrated in this module.

## 8.1 Characteristics of a Normal Distribution

The normal distribution is the familiar bell-shaped curve (Figure 8.1-Right). Normal distributions have two parameters – the population mean, $\mu$, and the population standard deviation, $\sigma$ – that control the exact shape and position of the distribution. Specifically, the mean $\mu$ controls the center and the standard deviation $\sigma$ controls the dispersion of the distribution (Figure 8.2).

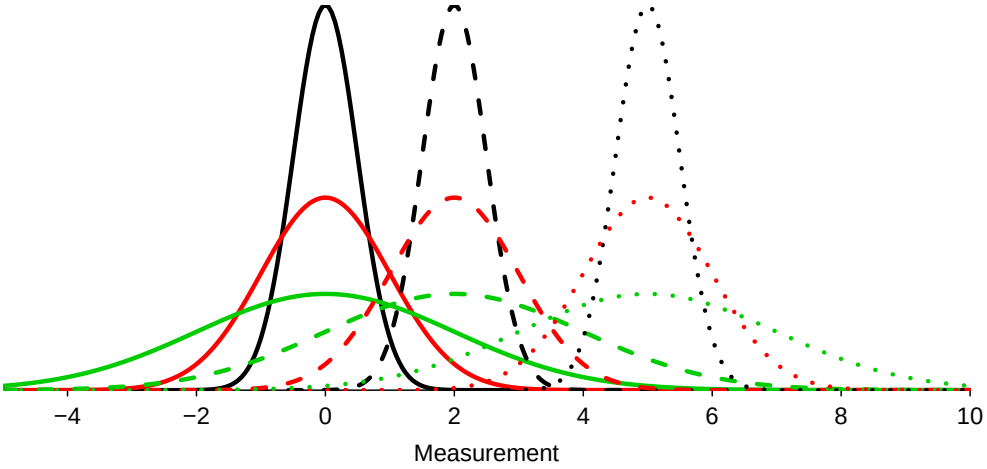


Figure 8.2. Nine normal distributions. Distributions with the same line type have the same value of $\mu$ (solid is $\mu$=0, dashed is $\mu$=2, dotted is $\mu$=5). Distributions with the same color have the same value of $\sigma$ (black is $\sigma$=0.5, red is $\sigma$=1, and green is $\sigma$=2).

There are an infinite number of normal distributions because there are an infinite number of combinations of $\mu$ and $\sigma$. However, each normal distribution will

1. be bell-shaped and symmetric,
2. centered at $\mu$,
3. have inflection points at $\mu \pm \sigma$, and
4. have a total area under the curve equal to 1.

If a generic variable $X$ follows a normal distribution with a mean of $\mu$ and a standard deviation of $\sigma$, then it is said that $X \sim N(\mu, \sigma)$. For example, if the heights of students ($H$) follows a normal distribution with a $\mu$ of 66 and a $\sigma$ of 3, then it is said that $H \sim N(66, 3)$. As another example, $Z \sim N(0, 1)$ means that the variable $Z$ follows a normal distribution with a mean of $\mu$=0 and a standard deviation of $\sigma$=1.

[1] See Module **??**.

## 8.2 Simple Areas Under the Curve

A common problem is to determine the proportion of individuals with a value of the variable between two numbers. For example, you might be faced with determining the proportion of all sites that have lead concentrations between 1.2 and 1.5 $\mu g \cdot m^{-3}$, the proportion of students that scored higher than 700 on the SAT, or the proportion of Least Weasels that are shorter than 150 mm. Before considering these more realistic situations, we explore calculations for the generic variable $X$ shown in Figure 8.3.

Let's consider finding the proportion of individuals in a *sample* with values between 0 and 2. A histogram can be used to answer this question because it is about the individuals in a sample (Figure 8.3-Left). In this case, the proportion of individuals with values between 0 and 2 is computed by dividing the number of individuals in the red shaded bars by the total number of individuals in the histogram. The analogous computation on the superimposed smooth curve is to find the area under the curve between 0 and 2 (Figure 8.3-Right). The area under the curve is a "proportion of the total" because, as stated above, the area under the entire curve is equal to 1. The actual calculations on the normal curve are shown in the following sections. However, at this point, note that the calculation of an area on a normal curve is analogous to summing the number of individuals in the appropriate classes of the histogram and dividing by $n$.

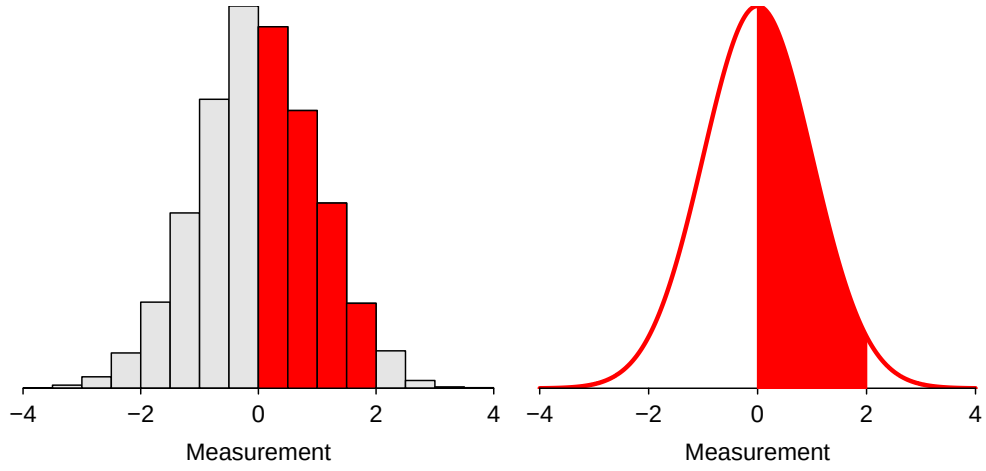


Figure 8.3. Depiction of finding the proportion of individuals between 0 and 2 on a histogram (**Left**) and on a standard normal distribution (**Right**).

◇ **The proportion of individuals between two values of a variable that is normally distributed is the area under the normal distribution between those two values.**

The 68-95-99.7 (or Empirical) Rule states that 68% of individuals that follow a normal distribution have values between $\mu - 1\sigma$ and $\mu + 1\sigma$, 95% have values between $\mu - 2\sigma$ and $\mu + 2\sigma$, and 99.7% have values between $\mu - 3\sigma$ and $\mu + 3\sigma$ (Figure 8.4).

The 68-95-99.7 Rule is true no matter what $\mu$ and $\sigma$ are as long as the distribution is normal. For example, if $A \sim N(3, 1)$, then 68% of the individuals will fall between 2 (i.e., 3-1*1) and 4 (i.e., 3+1*1) and 99.7% will fall between 0 (i.e., 3-3*1) and 6 (i.e., 3+3*1). Alternatively, if $B \sim N(9, 3)$, then 68% of the individuals will fall between 6 (i.e., 9-1*3) and 12 (i.e., 9+1*3) and 95% will be between 3 (i.e., 9-2*3) and 15 (i.e., 9+2*3). Similar calculations can be made for any normal distribution.

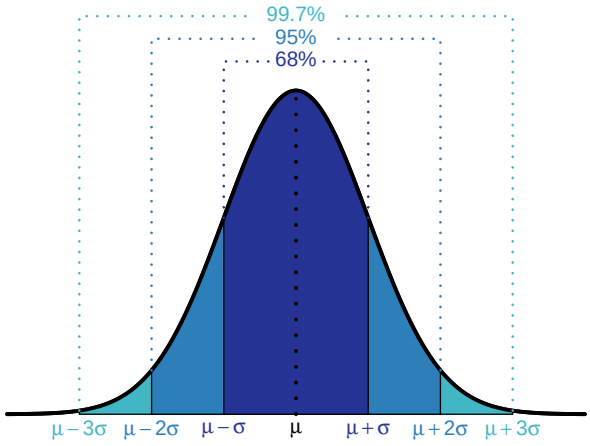


Figure 8.4. Depiction of the 68-95-99.7 (or Empirical) Rule on a normal distribution.

The 68-95-99.7 Rule is used to find areas under the normal curve as long as the value of interest is an **integer** number of standard deviations away from the mean. For example, the proportion of individuals that have a value of A greater than 5 (Figure 8.5) is found by first realizing that 95% of the individuals on this distribution fall between 1 and 5 (i.e., $\pm 2\sigma$ from $\mu$). By subtraction this means that 5% of the individuals must be less than 1 **AND** greater than 5. Finally, because normal distributions are symmetric, the same percentage of individuals must be less than 1 as are greater than 5. Thus, half of 5%, or 2.5%, of the individuals have a value of A greater than 5.

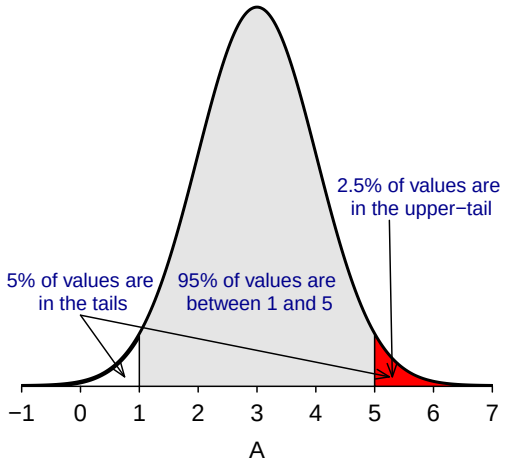


Figure 8.5. The N(3,1) distribution depicting how the 68-95-99.7 Rule is used to compute the percentage of individuals with values greater than 5.

◇ **The 68-95-99.7 Rule can only be used for questions involving integer standard deviations away from the mean.**

## 8.3 More Complex Areas (Forward Calculations)

Areas under the curve relative to non-integer numbers of standard deviations away from the mean can only be found with the help of special tables or computer software. In this course, we will use R.

The area under a normal curve relative to a particular value is computed in R with `distrib()`. This function requires the *particular value* as the first argument and the mean and standard deviation of the normal distribution in the `mean=` and `sd=` arguments, respectively. The `distrib()` function defaults to finding the area under the curve to the **left of** the particular value, but it can find the area under the curve to the right of the particular value by including `lower.tail=FALSE`.

For example, suppose that the heights of a population of students is known to be $H \sim N(66, 3)$. The proportion of students in this population that have a height less than 71 inches is computed below. Thus, approximately 95.2% of students in this population have a height less than 71 inches (Figure 8.6).

```
> ( distrib(71,mean=66,sd=3) )
[1] 0.9522096
```

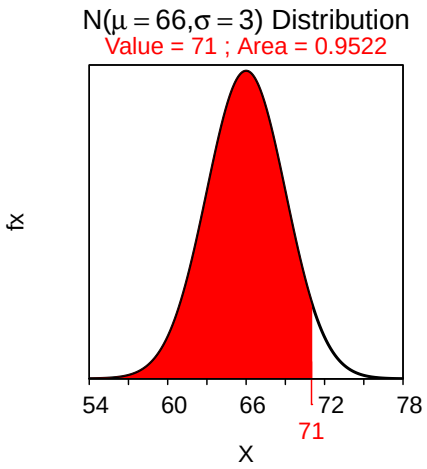


Figure 8.6. Calculation of the proportion of individuals on a $N(66, 3)$ with a value less than 71.

The proportion of students in this population that have a height *greater* than 68 inches is computed below (note use of `lower.tail=FALSE`). Thus, approximately 25.2% of students in this population have a height greater than 68 inches (Figure 8.7).

```
> ( distrib(68,mean=66,sd=3,lower.tail=FALSE) )
[1] 0.2524925
```

Finding the area between two particular values is a bit more work. To answer "between"-type questions, the area less than the smaller of the two values is subtracted from the area less than the larger of the two values. This is illustrated by noting that two values split the area under the normal curve into three parts – A, B, and C in Figure 8.8. The area between the two values is B. The area to the left of the larger value corresponds to the area A+B. The area to the left of the smaller value corresponds to the area A. Thus, subtracting the latter from the former leaves the "in-between" area B (i.e., (A+B)-A = B).

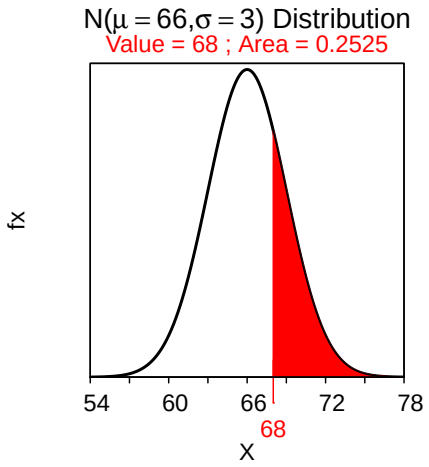


Figure 8.7. Calculation of the proportion of individuals on a $N(66, 3)$ with a value greater than 68.

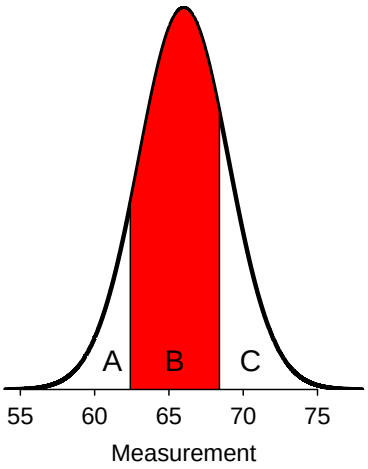


Figure 8.8. Schematic representation of how to find the area between two $Z$ values.

For example, the area between 62 and 70 inches of height is found below. Thus, 81.8% of students in this population have a height between 62 and 70 inches.

```
> ( AB <- distrib(70,mean=66,sd=3) )  # left-of 70
[1] 0.9087888
> ( A <- distrib(62,mean=66,sd=3) )   # left-of 62
[1] 0.09121122
> AB-A                                # between 62 and 70
[1] 0.8175776
```

◇ **The area between two values is found by subtracting the area less than the smaller value from the area less than the larger value.**

## 8.4 Values from Areas (Reverse Calculations)

Another important calculation with normal distributions is finding the value or values of $X$ with a given proportion of individuals less than, greater than, or between. For example, it may be necessary to find the test score such that 90% (or 0.90 as a proportion) of the students scored lower. In contrast to the calculations in the previous section (where the value of $X$ was given and a proportion of individuals was asked for), the calculations in this section give a proportion and ask for a value of $X$. These types of questions are called **"reverse" normal distribution questions** to contrast them with questions from the previous section.

Reverse questions are also answered with `distrib()`, though the first argument is now the given proportion (or area) of interest. The calculation is treated as a "reverse" question when `type="q"` is given to `distrib()`.[2] For example, the height that has 20% of all students shorter is 63.5 inches, as computed below (Figure 8.9).

```
> ( distrib(0.20,mean=66,sd=3,type="q") )
[1] 63.47514
```

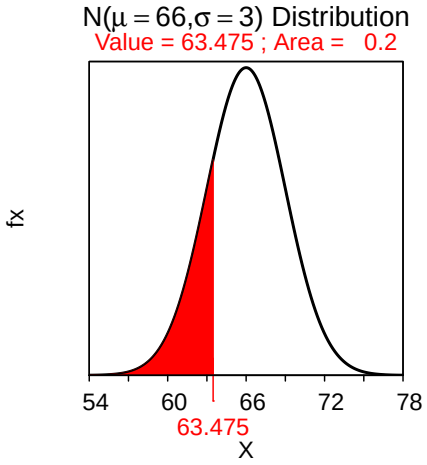


Figure 8.9. Calculation of the height with 20% of all students shorter.

"Greater than" reverse questions are computed by including `lower.tail=FALSE`. For example, 10% of the population of students is taller than 69.8 inches, as computed below (Figure 8.10).

```
> ( distrib(0.10,mean=66,sd=3,type="q",lower.tail=FALSE) )
[1] 69.84465
```

"Between" questions can only be easily handled if the question is looking for endpoint values that are symmetric about $\mu$. In other words, the question must ask for the two values that contain the "most common" proportion of individuals. For example, suppose that you were asked to find the most common 80% of heights. This type of question is handled by converting this "symmetric between" question into two "less than" questions. For example, in Figure 8.11 the area D is the symmetric area of interest. If D is 0.80, then C+E must be 0.20.[3] Because D is symmetric about $mu$, C and E must both equal 0.10. Thus, the

[2] "q" stands for quantile.
[3] Because all three areas must sum to 1.

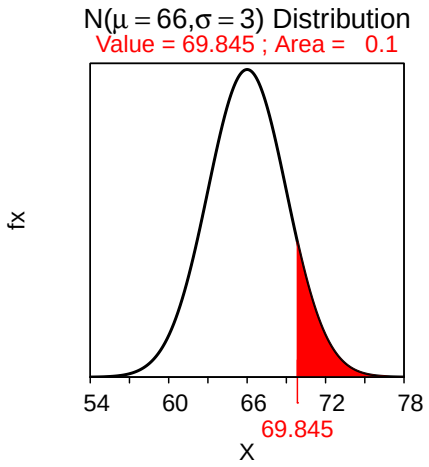


Figure 8.10. Calculation of the height with 10% of all students taller.

lower bound on D is the value that has 10% of all values smaller. Similarly, because the combined area of C and D is 0.90, the upper bound on D is the value that has 90% of all values smaller. This question has now been converted from a "symmetric between" to two "less than" questions that can be answered exactly as shown above. For example, the two heights that have a symmetric 80% of individuals between them are 62.2 and 69.8 as computed below.

```
> ( distrib(0.10,mean=66,sd=3,type="q") )
[1] 62.15535
> ( distrib(0.90,mean=66,sd=3,type="q") )
[1] 69.84465
```

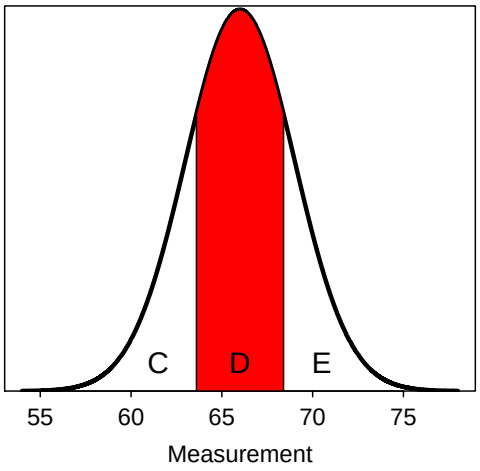


Figure 8.11. Depiction of areas in a reverse between type normal distribution question.

## 8.5 Distinguish Calculation Types

It is critical to be able to distinguish between the two main types of calculations made from normal distributions. The first type of calculation is a "forward" calculation where the area or proportion of individuals relative to a value of the variable must be found. The second type of calculation is a "reverse" calculation where the value of the variable relative to a particular area is calculated.

Distinguishing between these two types of calculations is a matter of deciding if (i) the value of the variable is given and the proportion (or area) is to be found or (ii) if the proportion (or area) is given and the value of the variable is to be found. Therefore, distinguishing between the calculation types is as simple as identifying what is given (or known) and what must be found. If the value of the variable is given but not the proportion or area, then a forward calculation is used. If the area or proportion is given, then a reverse calculation to find the value of the variable is used.

## 8.6 Standardization and Z-Scores

An individual that is 59 inches tall is 7 inches shorter than average if heights are $N(66, 3)$. Is this a large or a small difference? Alternatively, this same individual is $\frac{-7}{3} = -2.33$ standard deviations below the mean. Thus, a height of 59 inches is relatively rare in this population because few individuals are more than two standard deviations away from the mean.[4] As seen here, the relative magnitude that an individual differs from the mean is better expressed as the number of standard deviations that the individual is away from the mean.

Values are "standardized" by changing the original scale (inches in this example) to one that counts the number of standard deviations (i.e., $\sigma$) that the value is away from the mean (i.e., $\mu$). For example, with the height variable above, 69 inches is one standard deviation above the mean, which corresponds to +1 on the standardized scale. Similarly, 60 inches is two standard deviations below the mean, which corresponds to -2 on the standardized scale. Finally, 67.5 inches on the original scale is one half standard deviation above the mean or +0.5 on the standardized scale.

The process of computing the number of standard deviations that an individual is away from the mean is called **standardizing**. Standardizing is accomplished with

$$Z = \frac{\text{"value"} - \text{"center"}}{\text{"dispersion"}} \tag{8.6.1}$$

or, more specifically,

$$Z = \frac{x - \mu}{\sigma} \tag{8.6.2}$$

For example, the standardized value of an individual with a height of 59 inches is $z = \frac{59-66}{3} = -2.33$. Thus, this individual's height is 2.33 standard deviations below the average height in the population.

Standardized values ($Z$) follow a $N(0, 1)$. Thus, $N(0, 1)$ is called the "standard normal distribution." The relationship between $X$ and $Z$ is one-to-one meaning that each value of $X$ converts to one and only one value of $Z$. This means that the area to the left of $X$ on a $N(\mu, \sigma)$ is the same as the area to the left of $Z$ on a $N(0, 1)$. This one-to-one relationship is illustrated in Figure 8.12 using the individual with a height of 59 inches and $Z = -2.33$.

◇ **The standardized scale (i.e., z-scores) represents the number of standard deviations that a value is from the mean.**

[4] From the 68-95-99.7% Rule.

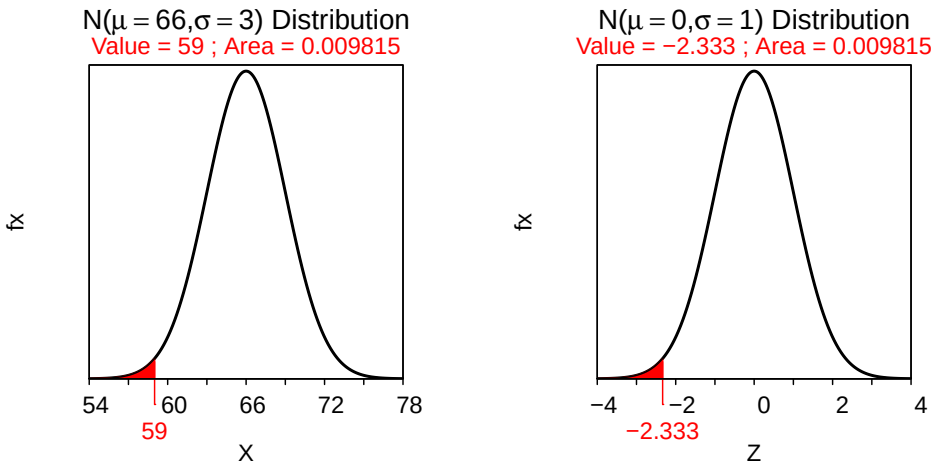


Figure 8.12. Plots depicting the area to the left of 59 on a $N(66, 3)$ (**Left**) and the area to the right of the corresponding Z-score of $Z = -2.33$ on a $N(0, 1)$ (**Right**). Not that the x-axis scales are different.