Lab 1.4: The Standard Normal Distribution using R

One of the most fundamental distributions in all of statistics is the *Normal Distribution* or the *Gaussian Distribution*. According to Wikipedia, "Carl Friedrich Gauss became associated with this set of distributions when he analyzed astronomical data using them, and defined the equation of its probability density function. It is often called the *bell curve* because the graph of its probability density resembles a bell."

The Probability Density Function

The *probability density function* for the normal distribution having mean μ and standard deviation σ is given by the function in Figure 1.

The Normal Probability Density Function

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

Figure 1. *The probability density function for the normal distribution.*

If we let the mean μ = 0 and the standard deviation σ = 1 in the probability density function in Figure 1, we get the probability density function for the *standard normal distribution* in Figure 2.

The Standard Normal Probability Density Function

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

Figure 2. *The probability density function for the standard normal distribution has mean μ = 0 and standard deviation σ = 1.*

It is a simple matter to produce a plot of the probability density function for the standard normal distribution.

```
> x=seq(-4,4,length=200)
> y=1/sqrt(2*pi)*exp(-x^2/2)
> plot(x,y,type="l",lwd=2,col="red")
```

If you'd like a more detailed introduction to plotting in R, we refer you to the activity Lab 0: Simple Plotting in R. However, these commands are simply explained.

1. The command `x=seq(-4,4,length=200)` produces 200 equally spaced values between -4 and 4 and stores the result in a vector assigned to the variable **x**.

2. The command `y=1/sqrt(2*pi)*exp(-x^2/2)` evaluates the probability density function of Figure 2 at each entry of the vector `x` and stores the result in a vector assigned to the variable `y`.

3. The command `plot(x,y,type="l",lwd=2,col="red")` plots y versus x, using:
   - a solid line type (`type="l"`) --- that's an "el", not an I (eye) or a 1 (one),
   - a line width of 2 points (`lwd=2`), and
   - uses the color red (`col="red"`).

The result is the "bell-shaped" curve shown in Figure 3.

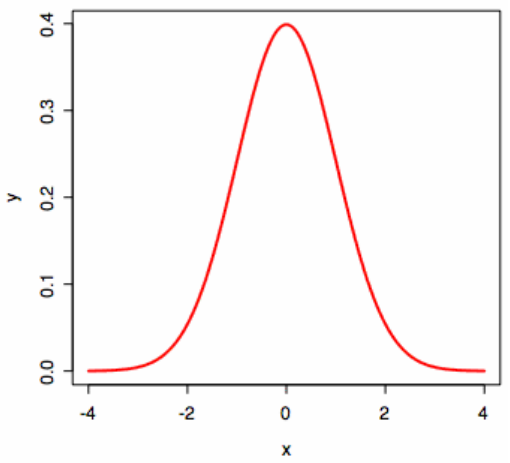

Figure 3. *The bell-shaped curve of the standard normal distribution.*

An Alternate Approach

The command `dnorm` can be used to produce the same result as the probability density function of Figure 2. Indeed, the "d" in `dnorm` stands for "density." Thus, the command `dnorm` is designed to provide values of the probability density function for the normal distribution.

```
> x=seq(-4,4,length=200)
> y=dnorm(x,mean=0,sd=1)
> plot(x,y,type="l",lwd=2,col="red")
```

These commands produce the plot shown in Figure 4. Note that the result is identical to the plot in Fig. 3.

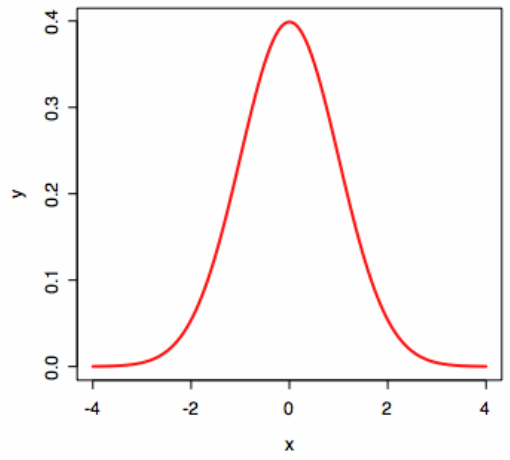

Figure 4. *The bell-shaped curve of the standard normal distribution.*

The Area Under the Probability Density Function

Like all probability density functions, the standard normal curves in Figures 3 and 4 possess three very important properties:

1. The graph of the probability density function lies entirely above the x-axis. That is, $f(x) \geq 0$ for all x.
2. The area under the curve (and above the x-axis) on its full domain is equal to 1.
3. The probability of selecting a number between $x = a$ and $x = b$ is equal to the area under the curve from $x = a$ to $x = b$.

As an example, consider the area under the standard normal curve shown in Figure 5.

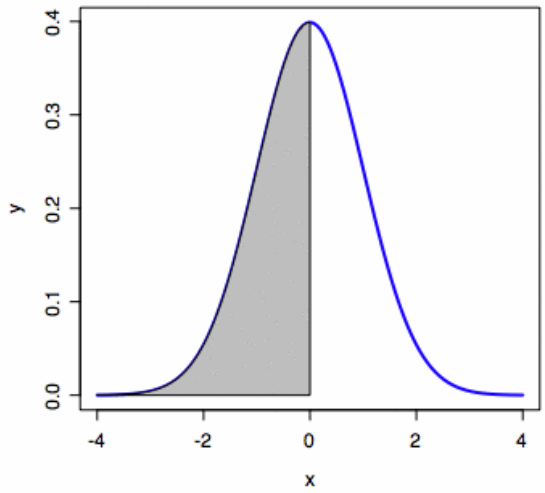

Figure 5. *The area under the standard normal curve to the left of $x = 0$.*

If the total area under the curve equals 1, then by symmetry one would expect that the area under the curve to the left of $x = 0$ would equal 0.5. R has a command called pnorm (the "p" is for "probability") which is designed to capture this probability (area under the curve).

```
> pnorm(0, mean=0, sd=1)
[1] 0.5
```

Note that the syntax is strikingly similar to the syntax for the density function. The command `pnorm(x, mean = , sd = )` will find the area under the normal curve to the left of the number x. Note that we use `mean=0` and `sd=1`, the mean and density of the standard normal distribution.

As a second example, suppose that we wish to find the area under the standard normal curve (mean = 0, standard deviation = 1) that is to the left of the value of x that is one standard deviation to the right of the mean, pictured for the reader in Figure 6.

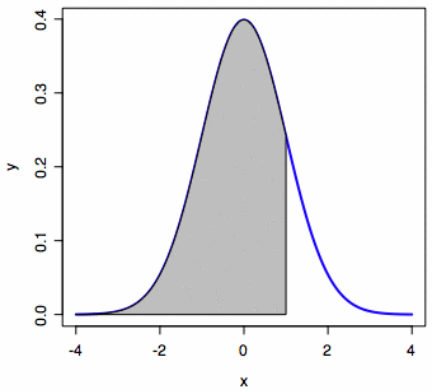

Figure 6. *The area under the curve to the left of $x = 1$.*

In this case, we wish to find the area under the curve to the left of $x = 1$.

```
> pnorm(1, mean=0, sd=1)
[1] 0.8413447
```

The interpretation of area as a probability is all-important. This result indicates that if we draw a number at random from the standard normal distribution, the probability that we draw a number that is less than or equal to 1 is 0.8413447.

If readers are interested, to produce the image in Figure 6, we used the following code.

```
> x=seq(-4,4,length=200)
> y=dnorm(x)
> plot(x,y,type="l", lwd=2, col="blue")
> x=seq(-4,1,length=200)
> y=dnorm(x)
> polygon(c(-4,x,1),c(0,y,0),col="gray")
```

For help on the polygon command enter `?polygon` and read the resulting help file. However, the basic idea is pretty simple. In the syntax `polygon(x,y)`, the argument *x* contains the *x*-coordinates of the vertices of the polygon you wish to draw. Similarly, the argument *y* contains the *y*-coordinates of the vertices of the desired polygon.

68%-95%-99.7% Rule

The 68% - 95% - 99.7% is a rule of thumb that allows practitioners of statistics to estimate the probability that a randomly selected number from the standard normal distribution occurs within 1, 2, and 3 standard deviations of the mean at zero.

Let's first examine the probability that a randomly selected number from the standard normal distribution occurs within one standard deviation of the mean. This probability is represented by the area under the standard normal curve between $x = -1$ and $x = 1$, pictured in Figure 7.

```
> x=seq(-4,4,length=200)
> y=dnorm(x)
> plot(x,y,type="l", lwd=2, col="blue")
> x=seq(-1,1,length=100)
> y=dnorm(x)
> polygon(c(-1,x,1),c(0,y,0),col="gray")
```

The code above was used to produce the shaded region shown in Figure 7.

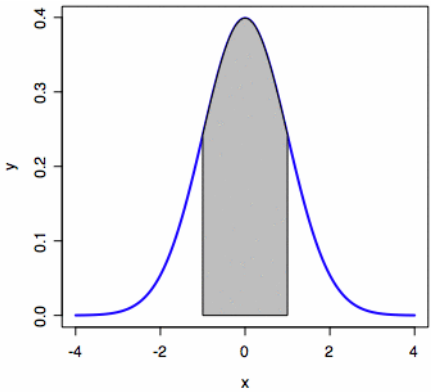

Figure 7. *The shaded area represents the probability of drawing a number from the standard normal distribution that falls within one standard deviation of the mean.*

Remember that R's pnorm command finds the area to the left of a given value of *x*. Thus, to find the area between $x = -1$ and $x = 1$, we must subtract the area to the left of $x = -1$ from the area to the left of $x = 1$.

```
> pnorm(1,mean=0,sd=1)-pnorm(-1,mean=0,sd=1)
[1] 0.6826895
```

There's the promised 68%!
In similar fashion, we can get the area within two and three standard deviations.

```
> x=seq(-4,4,length=200)
> y=dnorm(x)
> plot(x,y,type="l", lwd=2, col="blue")
> x=seq(-2,2,length=200)
> y=dnorm(x)
> polygon(c(-2,x,2),c(0,y,0),col="gray")
```

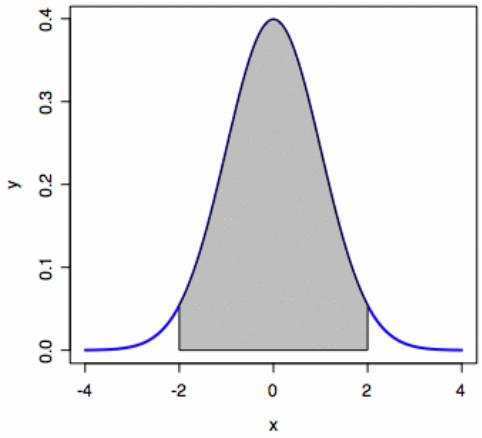

Figure 8. *The shaded area represents the probability of drawing a number from the standard normal distribution that falls within two standard deviations of the mean.*

To find the area between $x = -2$ and $x = 2$, we must subtract the area to the left of $x = -2$ from the area to the left of $x = 2$.

```
> pnorm(2,mean=0,sd=1)-pnorm(-2,mean=0,sd=1)
[1] 0.9544997
```

There is a 95% chance that the number drawn falls within two standard deviations of the mean.

For 3 standard deviations:

```
> x=seq(-4,4,length=200)
> y=dnorm(x)
> plot(x,y,type="l",lwd=2,col="blue")
> x=seq(-3,3,length=200)
> y=dnorm(x)
> polygon(c(-3,x,3),c(0,y,0),col="gray")
```

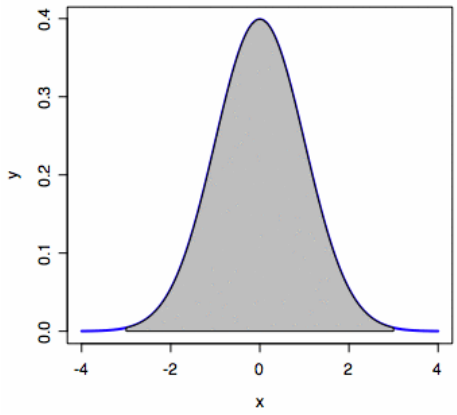

Figure 9. *The shaded area represents the probability of drawing a number from the standard normal distribution that falls within three standard deviations of the mean.*

To find the area between x = -3 and x = 3, we must subtract the area to the left of x = -3 from the area to the left of x = 3.

```
> pnorm(3,mean=0,sd=1)-pnorm(-3,mean=0,sd=1)
[1] 0.9973002
```

Therefore, the chance that a number drawn randomly from the standard normal distribution falls within three standard deviations of the mean is 99.7%!

Important Result: We conclude that virtually all numbers from the standard normal distribution occur within three standard deviations of the mean.

Quantiles

Sometimes the opposite question is asked. That is, suppose that the area under the curve to the left of some unknown number is known. What is the unknown number?
For example, suppose that the area under the curve to the left of some unknown x-value is 0.85, as shown in Figure 10.

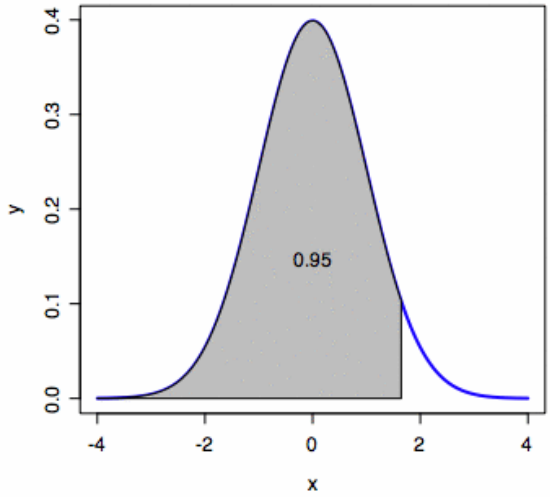

Figure 10. *The area under the curve to the left of some unknown x-value is 0.95.*

To find the unknown value of x we use R's qnorm command (the "q" is for "quantile").

```
> qnorm(0.95,mean=0,sd=1)
[1] 1.644854
```

Hence, there is a 95% probability that a random number less than or equal to 1.644854 is chosen from the standard normal distribution.

In a sense, R's `pnorm` and `qnorm` commands play the roles of inverse functions. On one hand, the command `pnorm` is fed a number and asked to find the probability that a random selection from the standard normal distribution falls to the left of this number. On the other hand, the command `qnorm` is given the probability and asked to find a limiting number so that the area under the curve to the left of that number equals the given probability.

We must emphasize that the area under the curve to the left is used when applying the commands `pnorm` and `qnorm`. If you are given an area to the right, then you must make a simple adjustment before applying the `qnorm` command. Suppose, as shown in Figure 11, that the area to the right of an unknown number is 0.80.

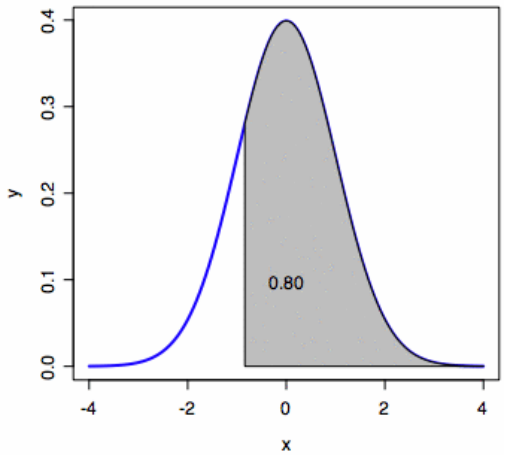

Figure 11. *The area under the curve to the right of some unknown x-value is 0.80.*

In this case, we must subtract the area to the right from the number 1 to obtain 1 - 0.80 = 0.20, which is the area to the left of the unknown value of *x* shown in Figure 12.

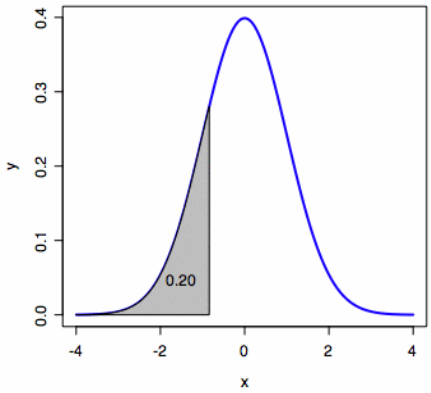

Figure 12. *The area under the curve to the left of some unknown x-value is 1 - 0.80 = 0.20.*

We can now use the qnorm command to find the unknown value of *x*.

```
> qnorm(0.20,mean=0,sd=1)
[1] -0.8416212
```

We now know that the probability of selecting a number from the standard normal distribution that is greater than or equal to -0.8416212 is 0.80.

Enjoy!

We hope you enjoyed this introduction to the standard normal distribution using the R system. We encourage you to explore further.