

# The Normal Distribution

Ben Babcock  
University of Minnesota

# Data Transformations

It is sometimes more convenient to work with data in a transformed metric rather than the original metric. When you transform data, you simply perform one or more \_\_\_\_\_  
\_\_\_\_\_ to all of the data points. You must do the same thing to everything.

H, 113

For example, if you subtracted 36 from all of your data points, you have transformed your data.

Two types of transformations: \_\_\_\_\_ and non-\_\_\_\_\_.

# Data Transformations: Linear

Linear transformations involve only \_\_\_\_\_, \_\_\_\_\_, \_\_\_\_\_, or \_\_\_\_\_ by constant values. H, 113.

Linear transformations DO NOT CHANGE THE RELATIVE SHAPE OF THE DISTRIBUTION!!!!!!!!!!!!!!!!!!!!!!

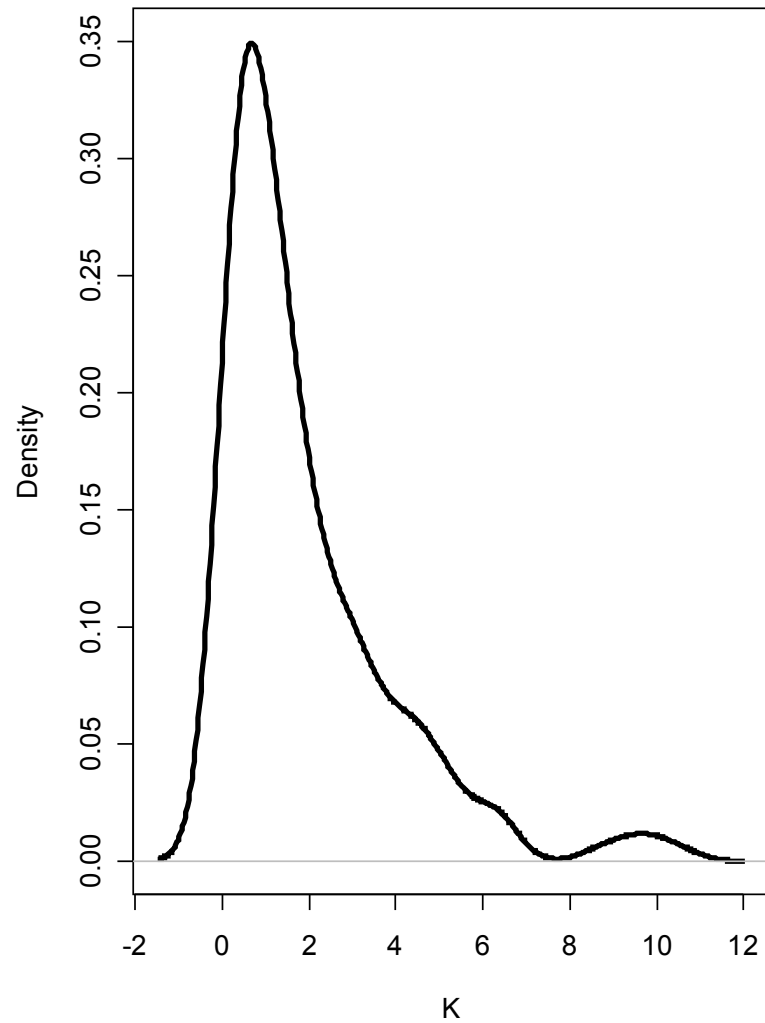
Non-linear transformations involve mathematical operations other than those listed above.

Examples:

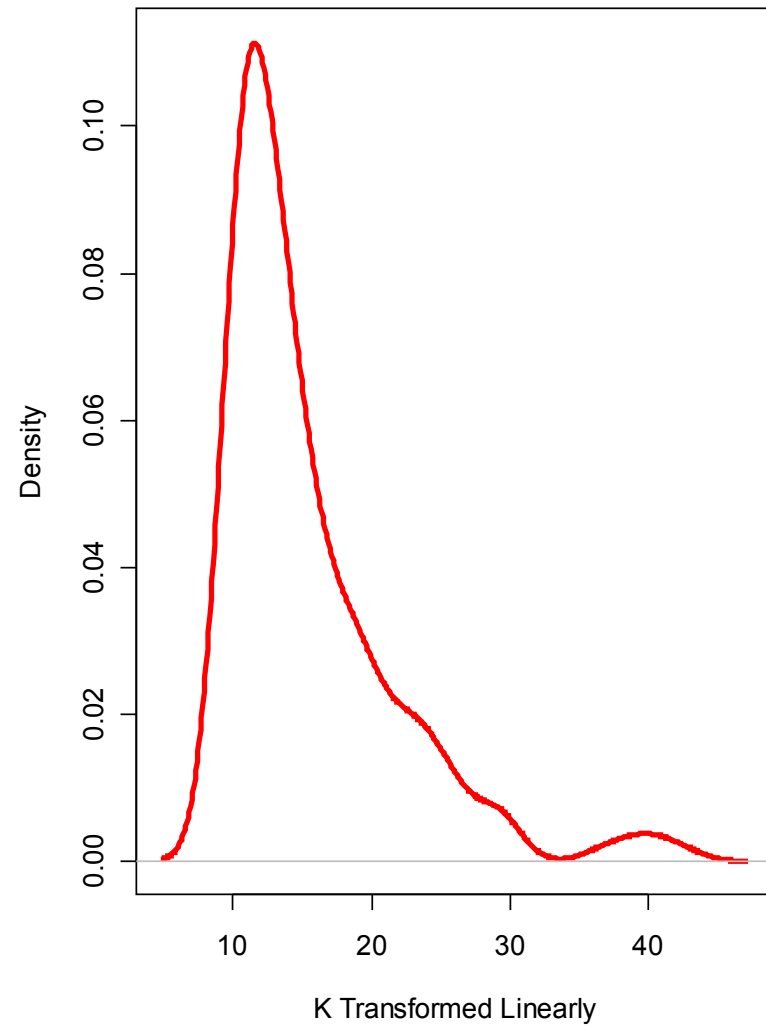
Non-linear transformations change the relative shape of the distribution.

# Effect of a Linear Transformation

Density Plot of Distribution K  
n=200

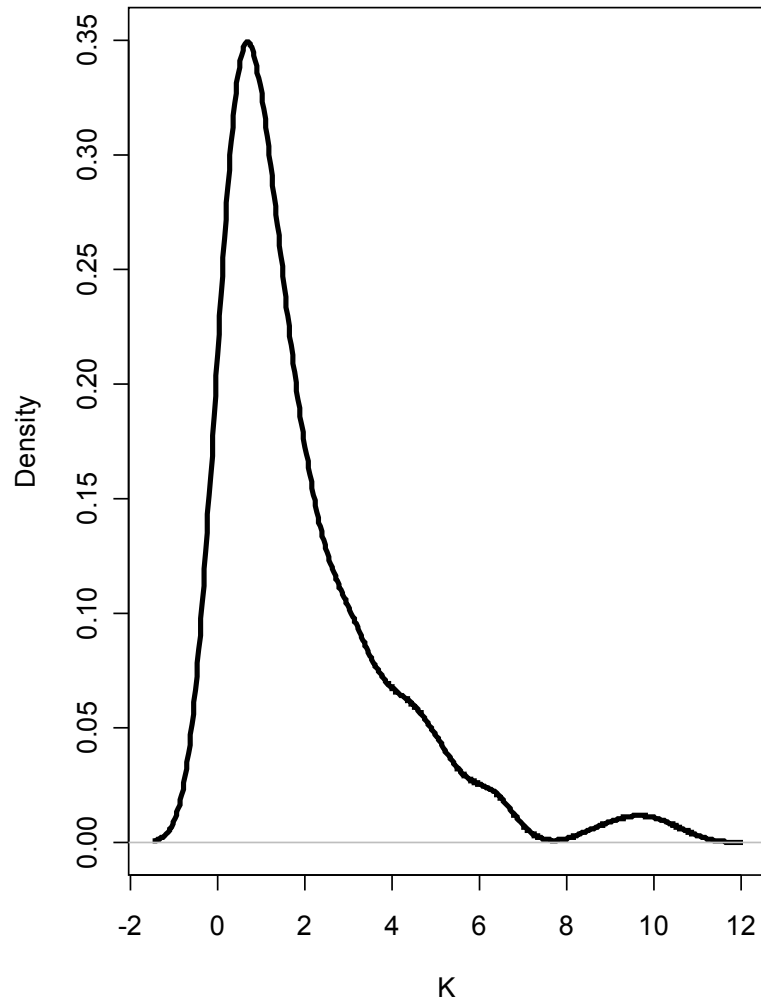


Density Plot of K Transformed  
by  $(K + 3) * \pi$

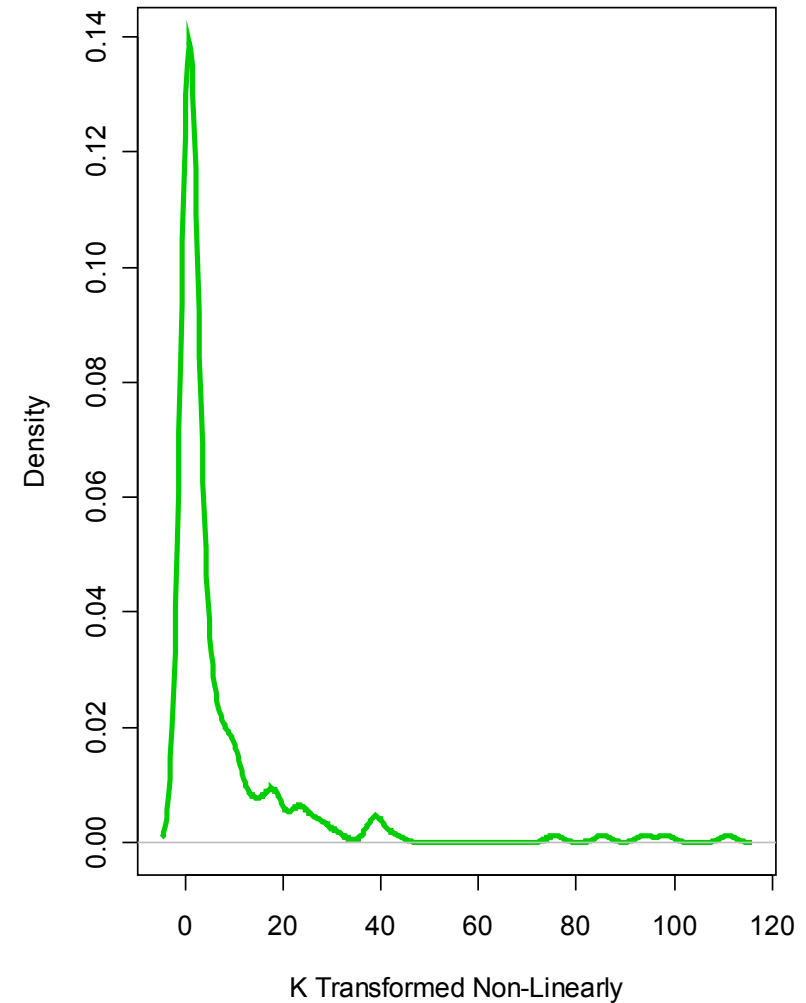


# Effect of a Non-Linear Transformation

Density Plot of Distribution K  
n=200

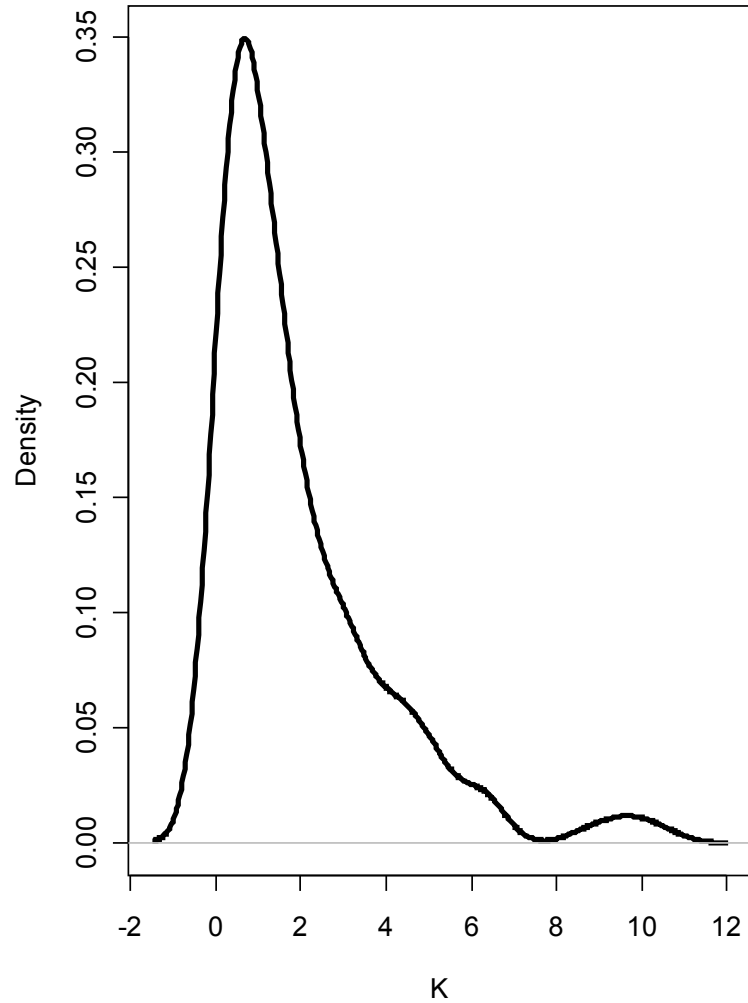


Density Plot of K Transformed  
by  $K^2$

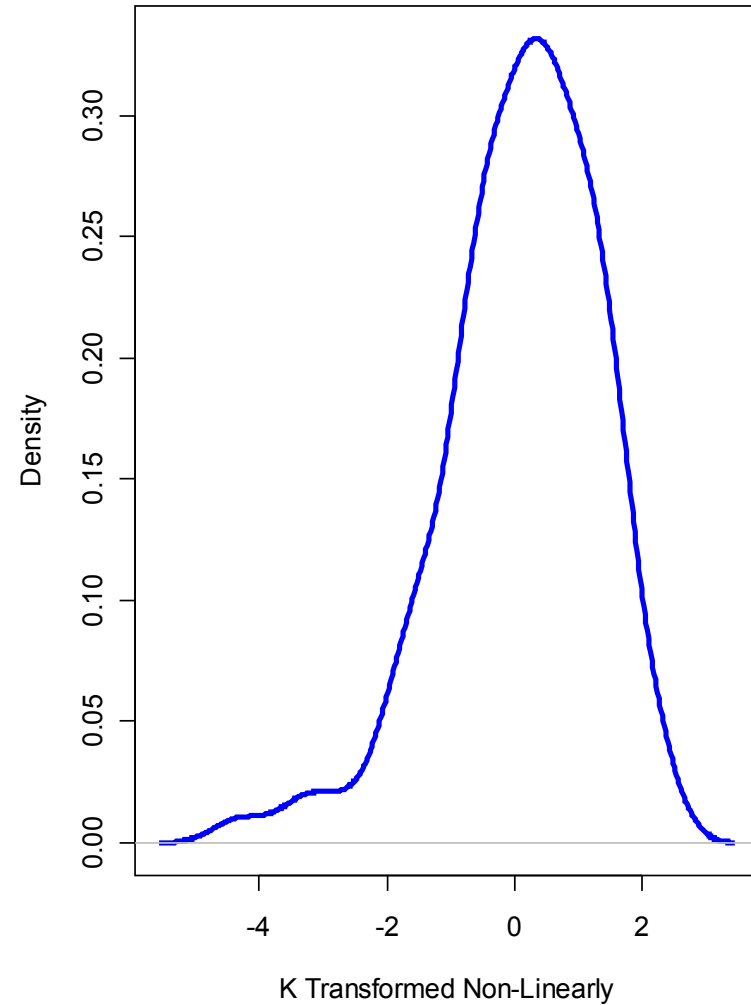


# Effect of a Non-Linear Transformation

Density Plot of Distribution K  
n=200



Density Plot of K Transformed  
by log(K)



# Deviation Scores

Instead of subtracting any 'ole number from all of our data, we should try subtracting the \_\_\_\_\_ of a dataset from all of the data points. These are called deviation scores.

The mean of deviation scores is  
always 0!

However, the sd is the same as the original distribution.

## Z-SCORES

The z-score transformation is one of the most (if not the most) common transformation in all of statistics.

$$z = \frac{x - \bar{x}}{s_x}$$

where  $x$  is a score from a sample  
 $\bar{x}$  is the mean of the sample and  
 $s_x$  is the sample standard deviation.

We first subtract the \_\_\_\_\_ from every score. After that, we divide by the \_\_\_\_\_. A z-score for one data point is the number of standard deviations it is away from the mean.

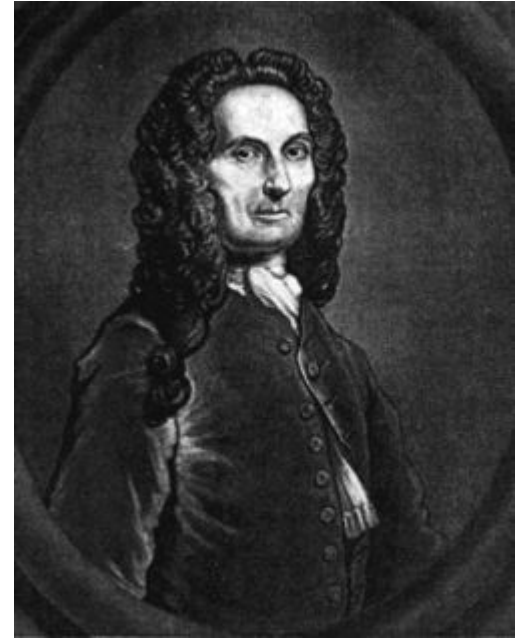
These scores will have a mean of 0 and a standard deviation and variance of 1! Cool!

# Food for Thought

Why would we want to use  $z$ -scores?

# Going Back in Time

Imagine for a moment that you are living in the year 1734. You want to come up with some sort of approximation for the binomial distribution when sample size is very large. This was the problem of Abraham de Moivre.

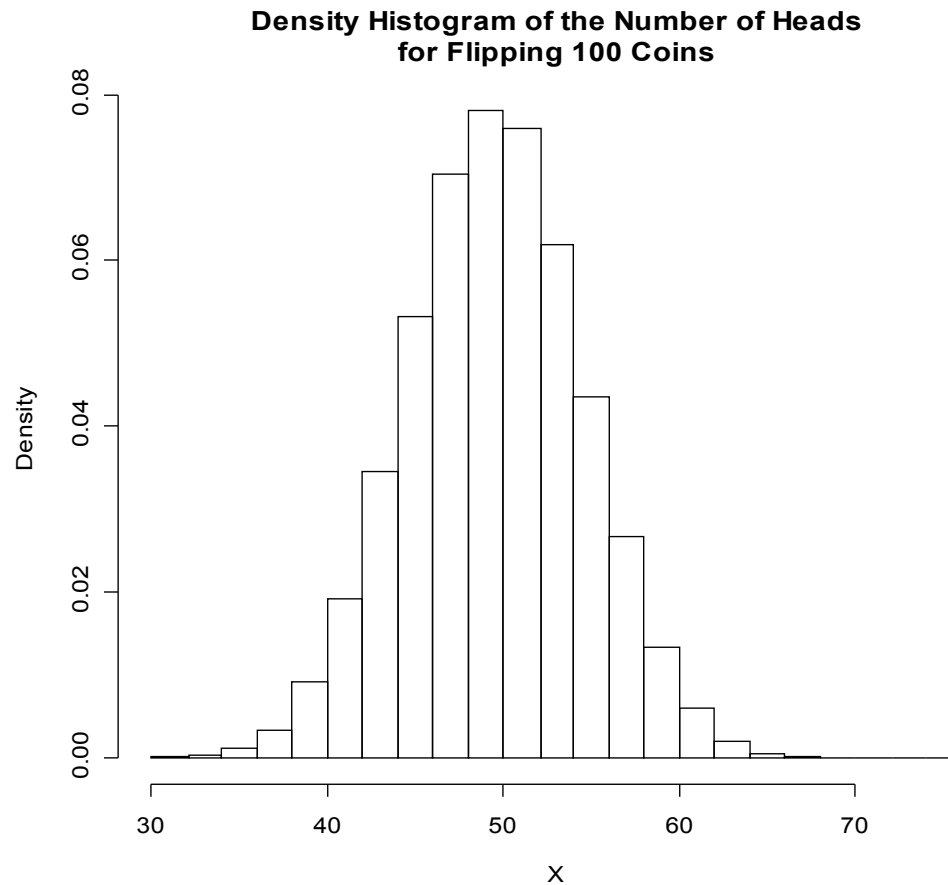


Binomial distribution example: flipping a set of 10 coins a whole lot of times and counting the number of heads.

Fact: When Jack Bauer flips a coin, it always comes up heads. Even dead presidents know never to turn your back on Jack Bauer.

# A Binomial Distribution

If we would have flipped 100 coins 100,000 times, the histogram would look like this:



# A Binomial Distribution

If you want a cont's function that is a good approximation of the binomial distribution, Abraham de Moivre and later Carl Friedrich Gauss found that the following function will do fine:

$$\frac{1}{\sigma_x \sqrt{2\pi}} \exp \left\{ \frac{-(x-\mu)^2}{2\sigma_x^2} \right\}$$

where  $x$  represents a variable

$\sigma_x$  is the standard deviation of the variable

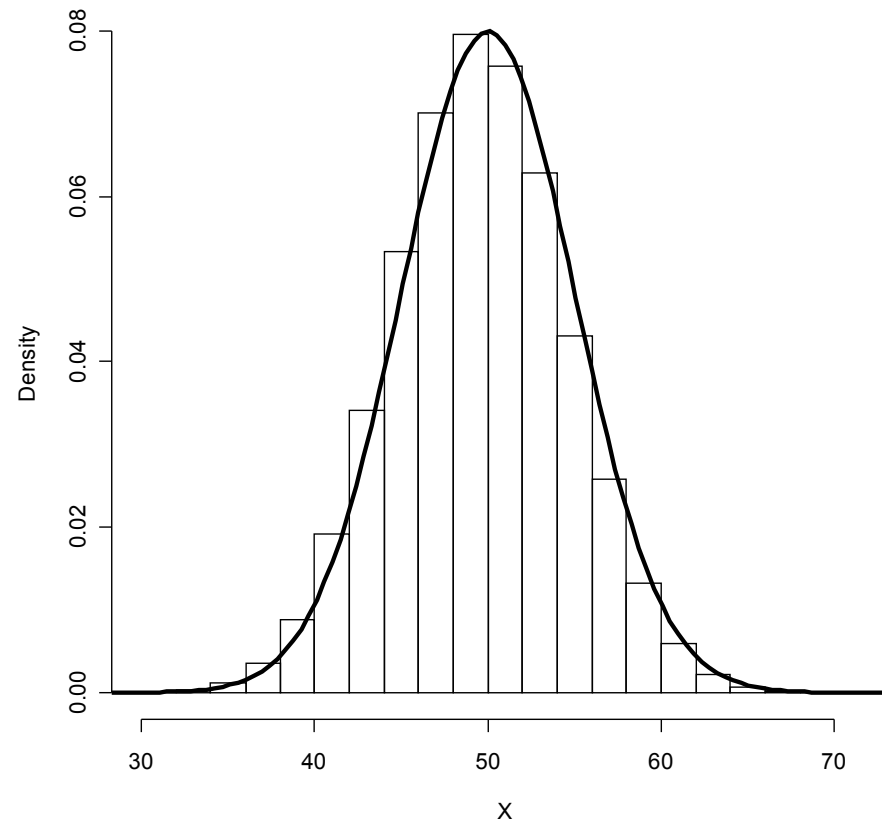
$\mu$  is the mean of the variable and

$\exp\{\}$  is notation for raising Euler's number ( $\approx 2.718281828$ ) to a power. H, 109. You don't have to memorize the equation.

Later investigators claimed that non-coin flipping phenomena were distributed approximately normally.

# An Approximation to the Binomial Distribution

Density Histogram of the Number of Heads for Flipping 100 Coins, Normal Curve Added



$$\text{Mean} = 100 \times 0.5 = 50$$
$$\text{Sd} = (100 \times 0.5 \times 0.5)^{1/2} = 5$$

# The Standard Normal Distribution

The standard normal distribution is a special case of the normal distribution where the mean is  $0$  and the variance (standard deviation) is  $1$ . H, 111.

The previously nasty normal equation equation simplifies from

$$\frac{1}{\sigma_x \sqrt{2\pi}} \exp \left\{ \frac{-(x-\mu)^2}{2\sigma_x^2} \right\}$$

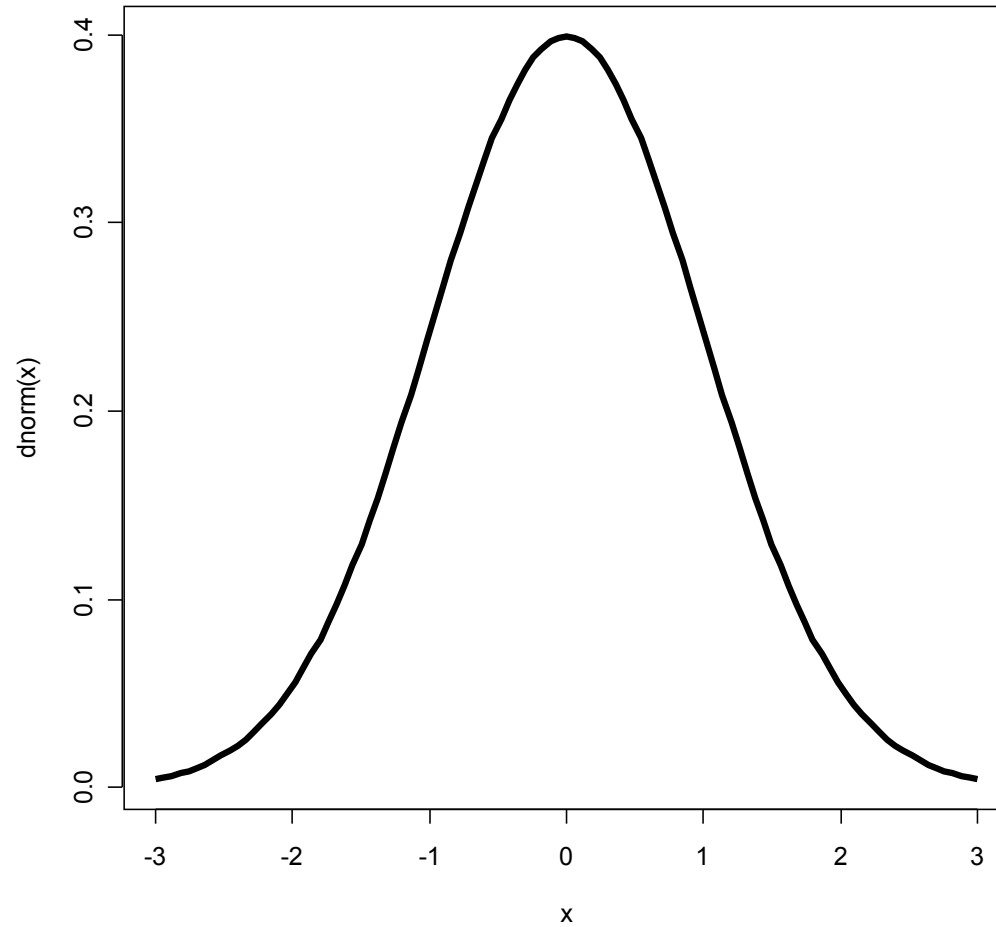
to

$$\frac{1}{\sqrt{2\pi}} \exp \left\{ \frac{-x^2}{2} \right\}$$

The standard normal distribution is on the same scale (same mean and standard deviation) as z-scores.

# What Does It Look Like?

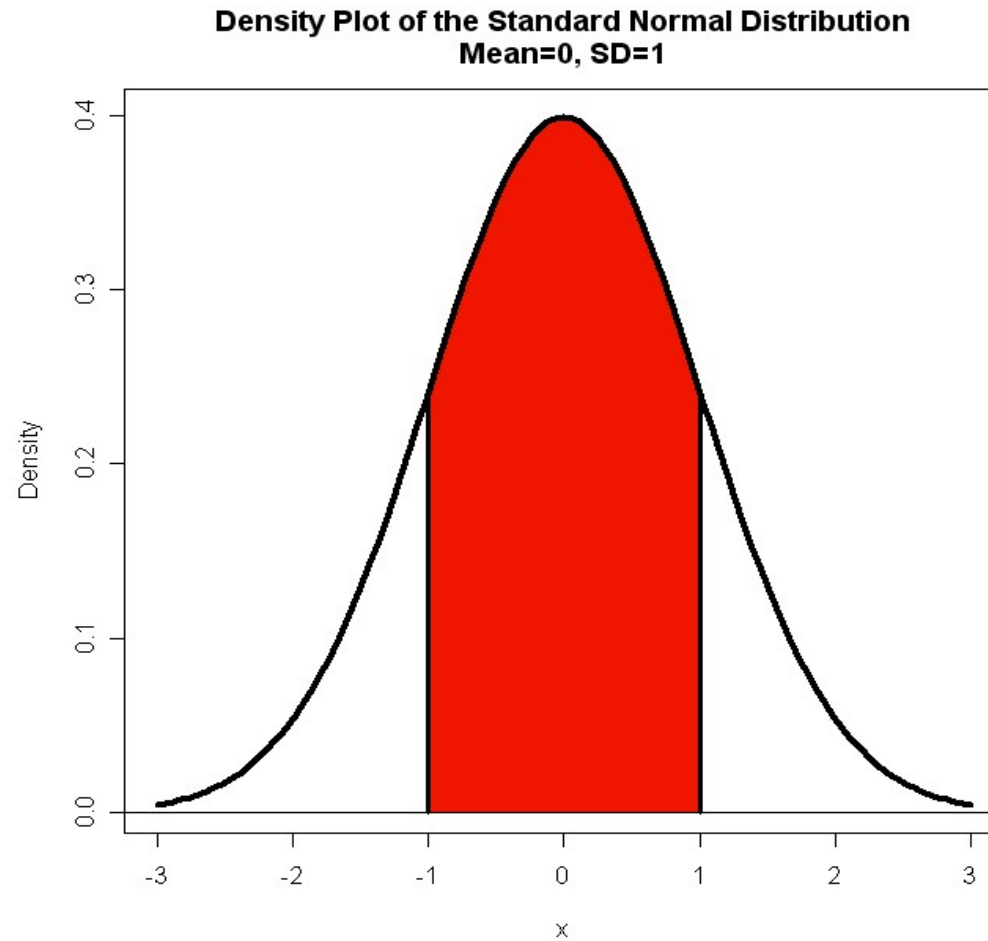
Density Plot of the Standard Normal Distribution  
Mean=0, SD=1



Think of the x-axis like z-scores

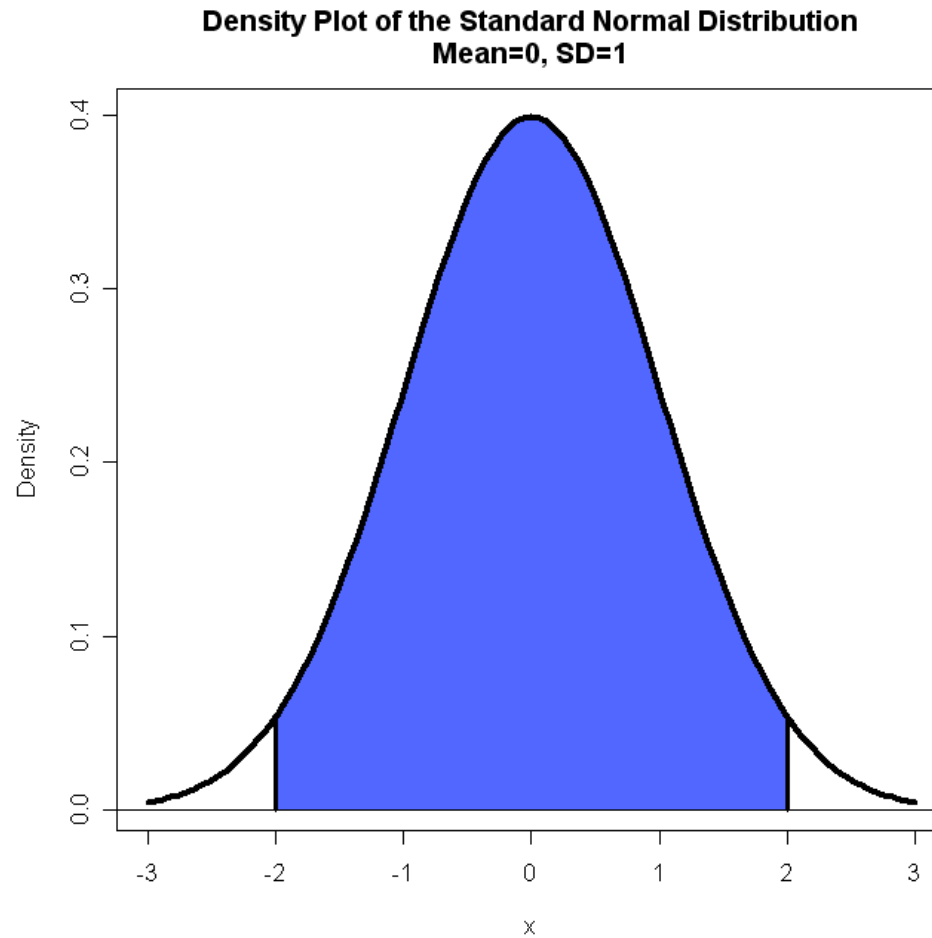
# Normal Distribution: Area Under the Curve

The area between -1 and 1 standard deviations is about 68% of the area in the normal distribution.



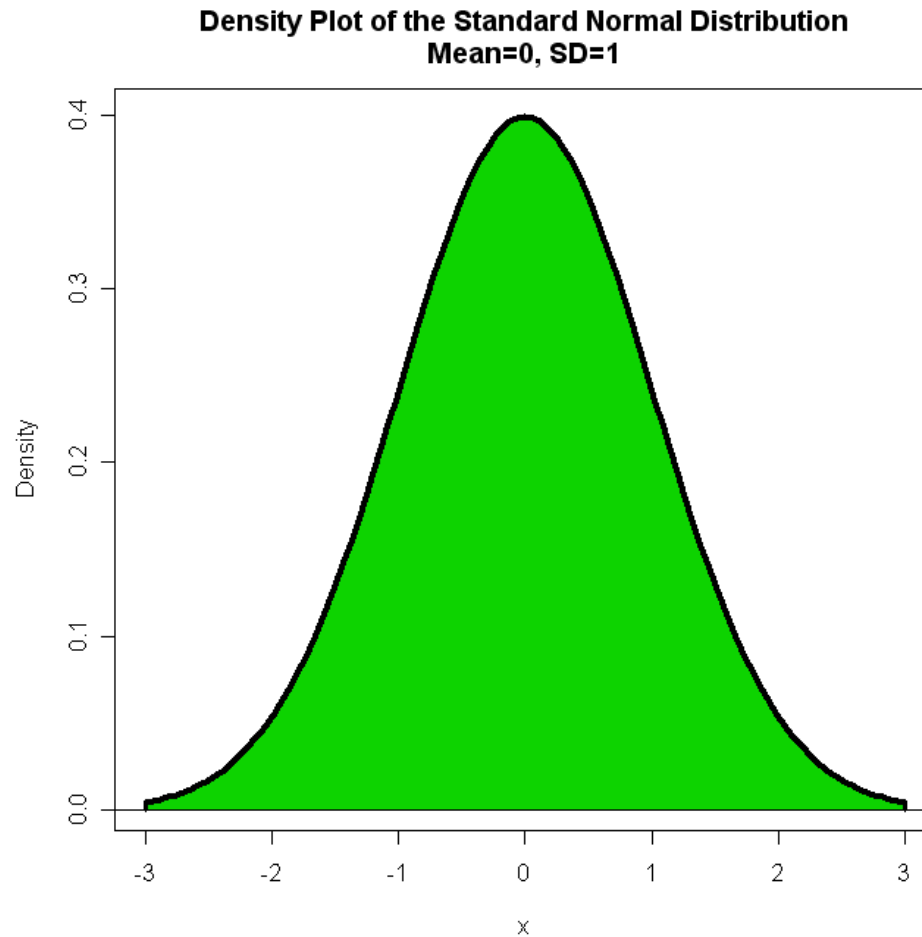
# Normal Distribution: Area Under the Curve

The area between -1.96 and 1.96 (about 2) standard deviations is 95% of the area in the normal distribution.



# Normal Distribution: Area Under the Curve

The area between -3 and 3 standard deviations is 99.5% of the area in the normal distribution. H, 114



# Calculating “Normal” Area Under the Curve

The part of the normal distribution equation in the circle makes calculating area under the curve a real \$!%&# (impossible).

$$\frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-x^2}{2}\right\}$$

The best that we can do is \_\_\_\_\_.

Areas between standard deviation values (like z-scores) and the mean are in tables that can be found in almost any statistics book (like Howell, Appendix D, p. 554). You will discuss how to use these tables in lab.

Computer programs (like R) will compute these areas for you.

# Why Use the Normal Curve?

1. Assuming that data are normally distributed makes numerous mathematical derivations turn out nicely.
2. Some phenomena loosely follow the normal distribution.
3. Before computers and simulation, statisticians had to use something to approximate things.
4. It corresponds to the Central Limit Theorem (CLT). We'll learn about this later.
5. Everyone's used it in the past, so why stop now?\*

There are certainly reasons to NOT use the normal distribution.

See Micceri, 1989. *The Unicorn, the Normal Curve, and Other Improbable Creatures*.

[http://www.indiana.edu/~educy520/sec6342/week\\_08/micceri89.pdf](http://www.indiana.edu/~educy520/sec6342/week_08/micceri89.pdf)

# Statistics Myth

Transforming scores into  $z$ -scores makes any distribution of scores more normal.

**WRONG!**

Why?