The Normal Distribution

Recap

Measures of Central Tendency

- Mean (average)
- Median (middle)
- Mode (most)

Measures of Dispersion

- Variance
- Standard deviation

Standardized Scores

The Normal Distribution

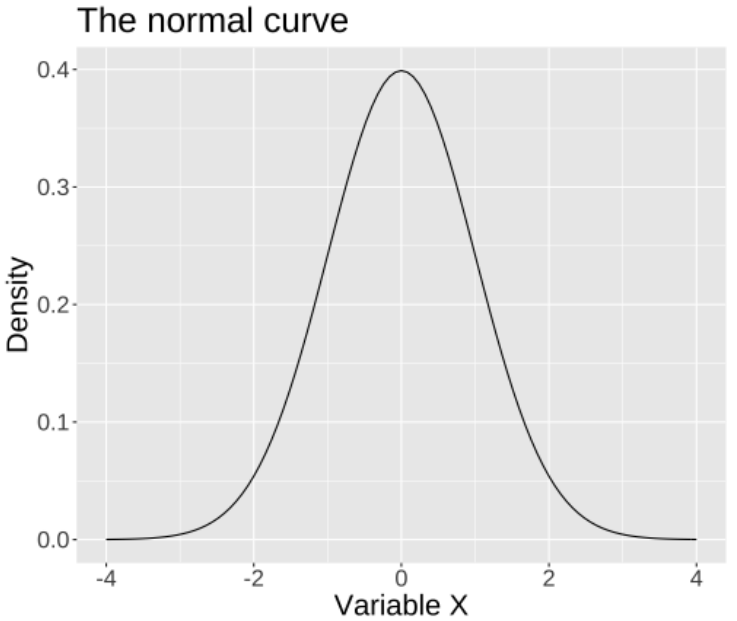
The **normal distribution**

- aka "bell curve" or "Gaussian distribution"
- Two-parameter distribution defined by the mean (μ) and standard deviation (σ)

The Normal Distribution

The **probability density function** gives the height of the curve at a particular value for X.

Although these values communicate information about probability or likelihood, they are not probabilities.



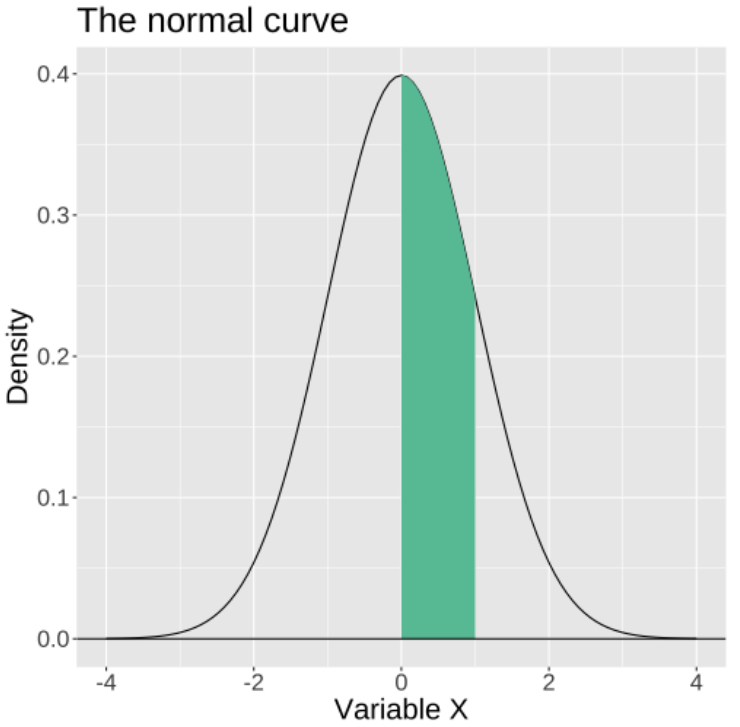
Probability Density Functions

You can take an entire semester long course on probability. Getting into the details is beyond the scope of this class, sadly. What you should know:

- The total area under the curve of a proability density function is **1**
- For a given continuous random variable, the probability of getting any single value is basically **0**

Probability of a single value

The area under the curve that lies between the mean (here 0) and a value of 1 is the probability of a score between 0 and 1.

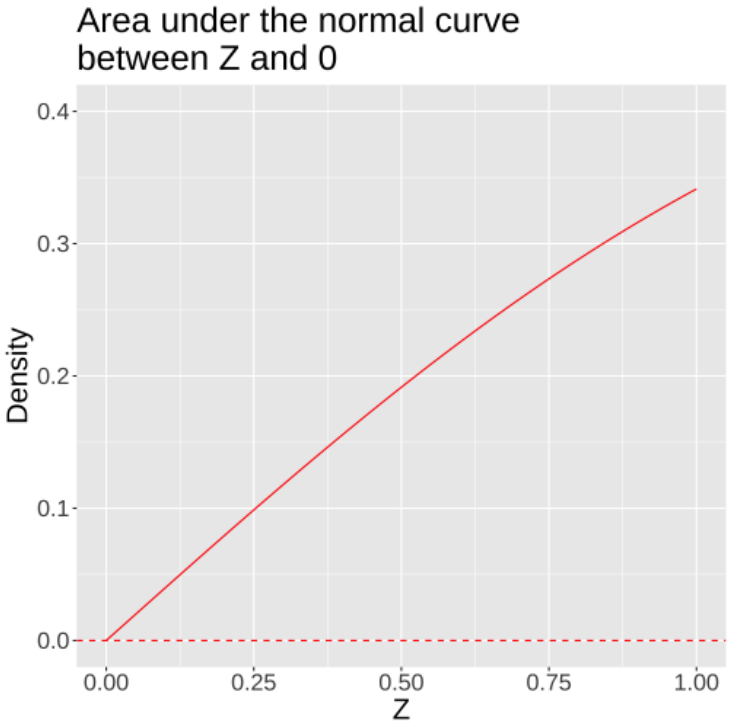


Probability of a single value

As our interval shrinks closer and closer to 0, our area (probability) shrinks as well.

It can get vanishingly close to 0—essentially a point rather than an area.

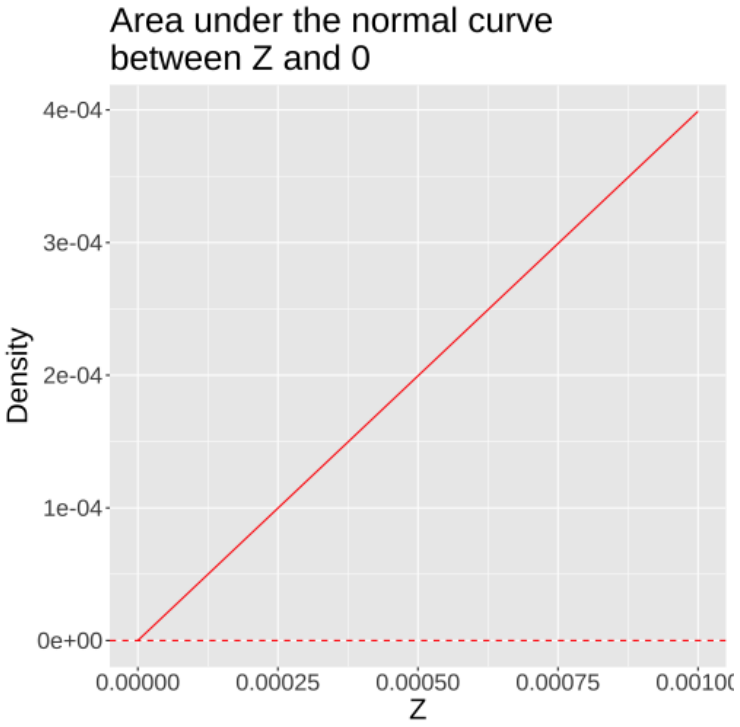
The probability of that "point" is 0.



Probability of a single value

We can keep shrinking the distance between Z and 0, never reaching 0, and still calculate an area.

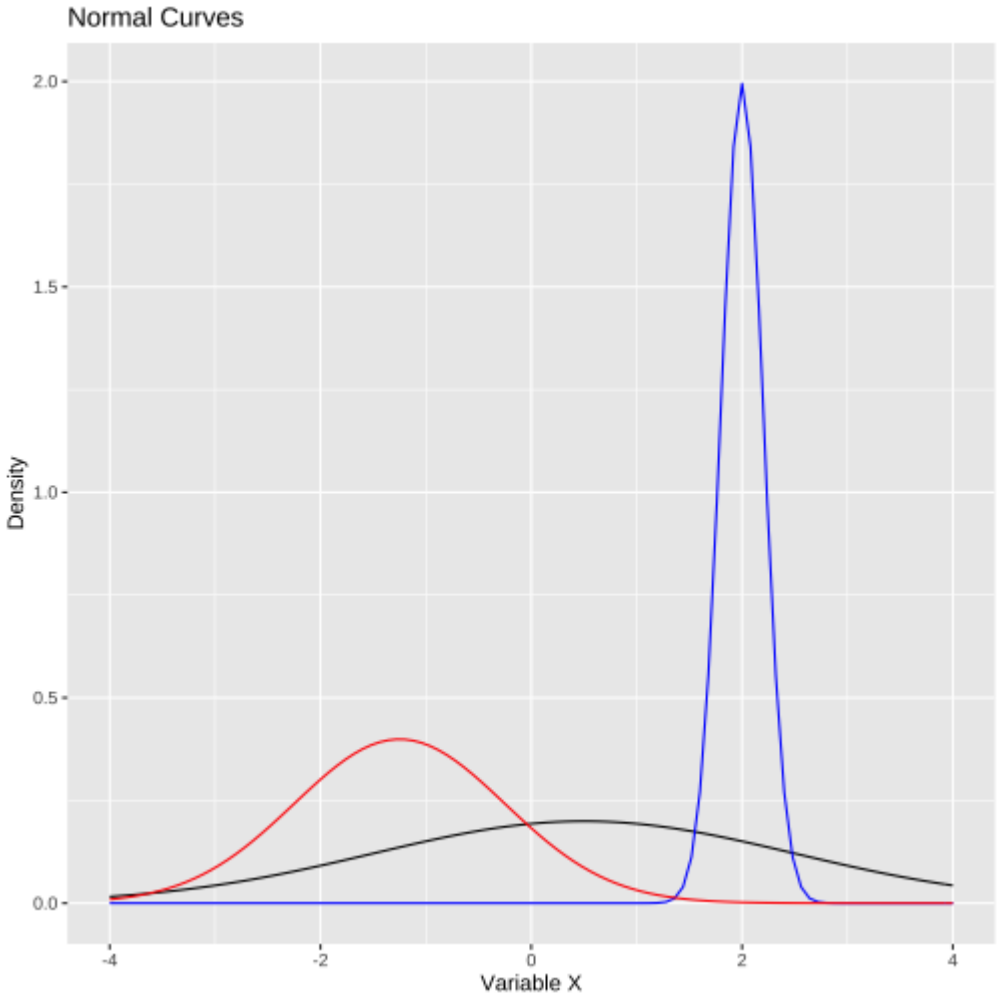
It will be very, very small.



Characteristics of the normal distribution

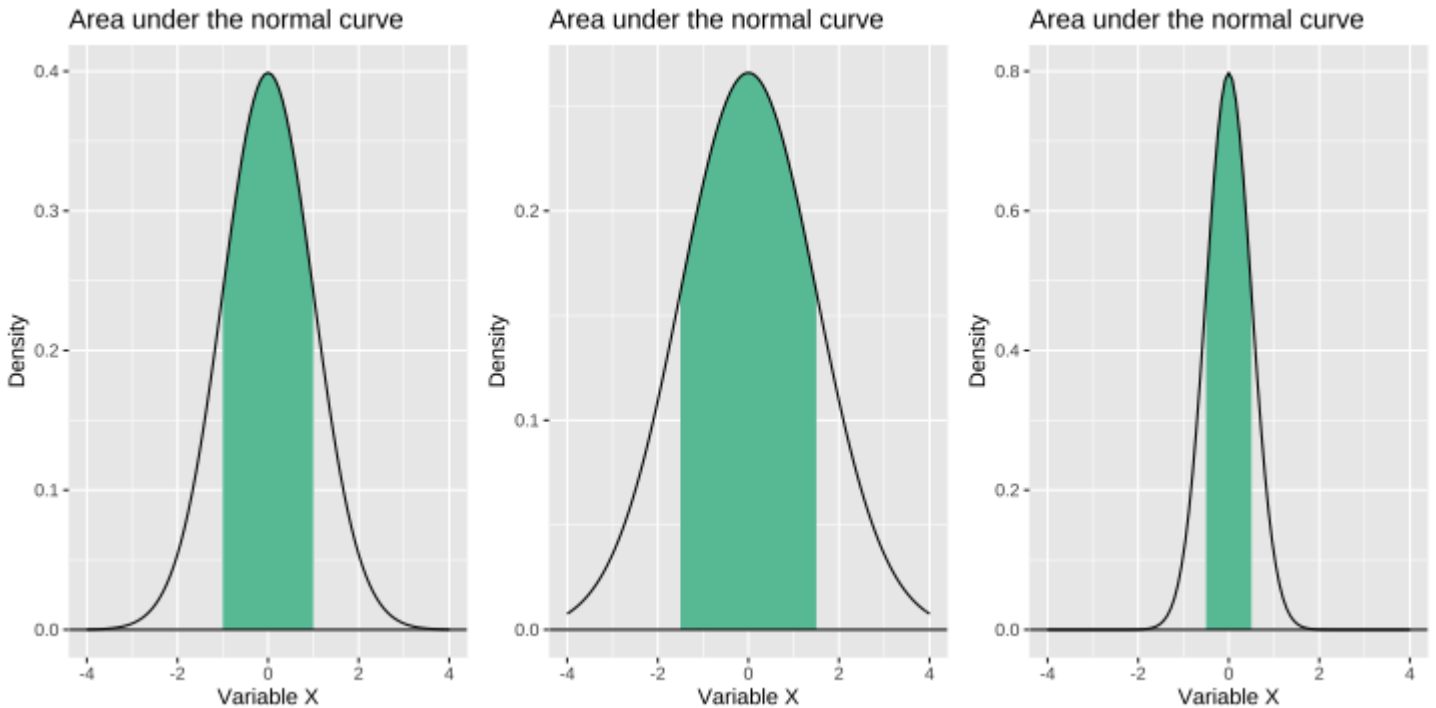
- The mean and standard deviation are independent
- The distribution is unimodal and symmetrical
- For two normal distributions, the area under the curve between corresponding locations in standard deviation units is the same regardless of μ and σ

Family of Normal Distributions



Family of Normal Distributions

All of these distributions are normal and have an equivalent area (proportion) that falls between a standard deviation below and above their respective means.



Characteristics of the normal distribution

- About 68.3% of the data will be within one standard deviation of the mean.
- About 95% of the data will be within two standard deviations of the mean.
- About 99.7% of the data will be within three standard deviations of the mean.

In other words, nearly **all** of the data will fall within 3 standard deviations of the mean in a normal distribution.

Standard normal distribution

A normal distribution with μ=0 and σ=1 is called **standard normal**. It is one specific distribution that comes from the larger family of normal distributions.

Variables with quite different means and standard deviations can be standardized so that they can be compared in the same metric (standard deviation units). This allows statements such as "relative to the mean, I am more conscientious (e.g., $z = 2$) than I am extraverted (e.g., $z = 1$)."

All continuous distributions can be standardized, but if they are not normal to begin with, standardization will not make them so. *Standardization does not alter distribution shape.*

Standard normal distribution

There is only one (1) standard normal distribution.

How is this useful?

- Given any score, we can calculate the probability of getting a value greater than that z-score. *(Or less than that z-score.)*
- Given any two z-scores, we can calculate the probability of getting a value between these scores. *(Or outside those z-scores)*
- Given a probability p, we can identify the z-score at which the proportion of scores below *(or above)* p falls.
- Given a probability p, we can identify the z-score at which the proportion of scores that fall above $-z$ and below z is equal to p.

Standardized scores (z-scores)

> Distance from the mean in standard deviation units

$$z = \frac{x_i - \bar{x}}{s}$$

Properties of z-scores:

- $\mu_z = 0$
- $\sigma_z = 1$
- Compare across scales and units of measures
- More easily identify extreme data

Using z-scores

```
## # A tibble: 6 x 2
##   name           height
##   <chr>           <int>
## 1 Luke Skywalker    172
## 2 C-3PO             167
## 3 R2-D2              96
## 4 Darth Vader       202
## 5 Leia Organa       150
## 6 Owen Lars         178
```

```
starwars %>%
  select(1:2) %>%
  mutate_at(2, ~round(x = scale(.
  head(.) %>%
  print(.)
```

```
## # A tibble: 6 x 2
##   name           height[,1]
##   <chr>               <dbl>
## 1 Luke Skywalker      -0.07
## 2 C-3PO               -0.21
## 3 R2-D2               -2.25
## 4 Darth Vader          0.79
## 5 Leia Organa         -0.7
## 6 Owen Lars            0.1
```

Using z-scores

Given any score, we can calculate the probability of getting a value greater than that z-score. *(Or less than that z-score.)*

You can look up tables that give you the probability value that corresponds to any given z-score. Or, you can use R code.

Luke Skywalker's height is z = -.07

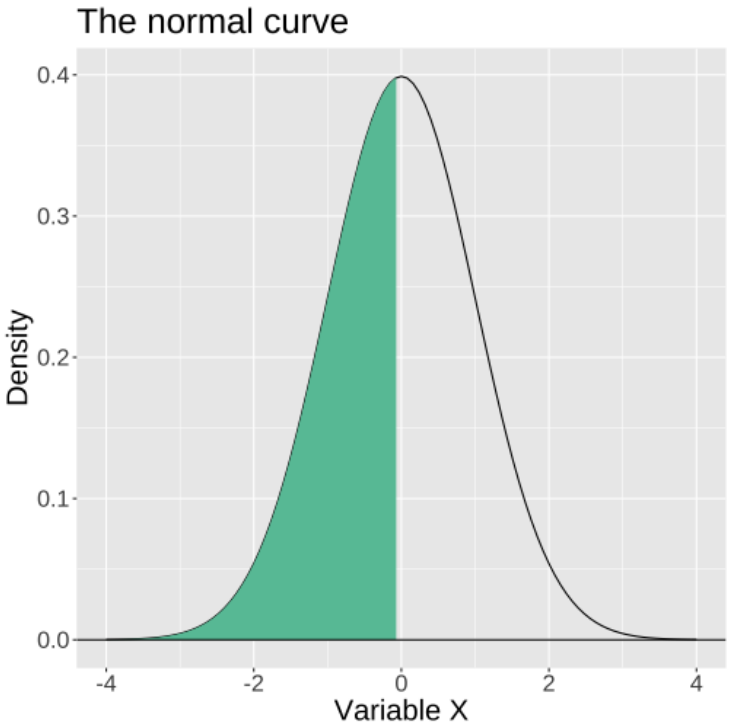
```
pnorm(q = -.07, mean = 0, sd = 1)
```

```
## [1] 0.4720968
```

p-values

What does $p = .4721$ mean?

- The probability of obtaining a z-score less than -.07
- The area under the curve from -.07, moving left



p-values

A p-value is the probability of getting a particular test statistic or more extreme given the null hypothesis is true

What is a p-value *NOT*:

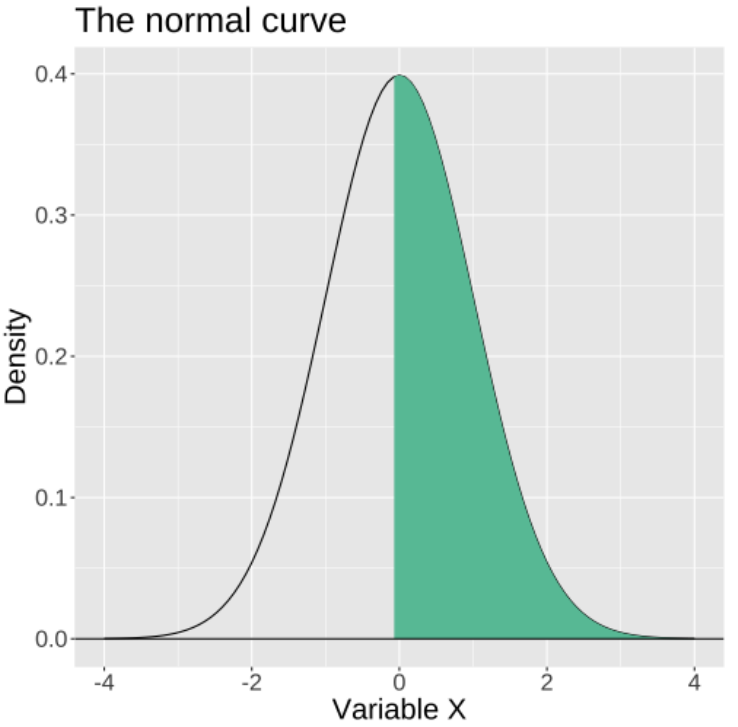
- p is not the probability that H0 is true
- p is not the probability of a Type I error
- p is not the probability that the data are due to chance
- p is not the probability of making a wrong decision
- the complement of p, which is (1-p), is not the probability that the alternative hypothesis is true

Using z-scores

The probability of getting a z-score of -.07 or greater?

```
1-pnorm(q = -.07, mean = 0, sd =
```

```
## [1] 0.5279032
```



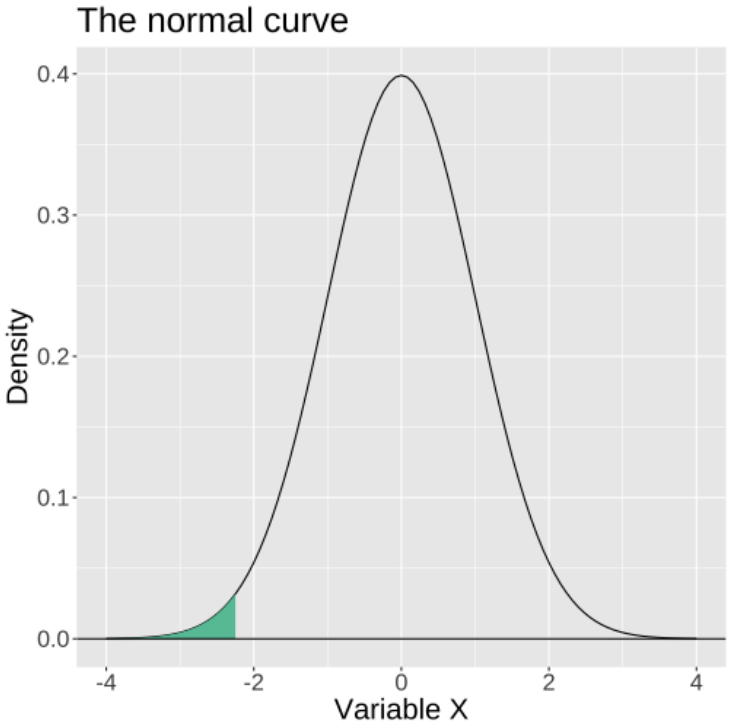
Using z-scores

What about R2D2? (z-score of -2.25)

- Probability of getting a z-score of -2.25 or something even smaller

```
pnorm(q = -2.25, mean = 0, sd = 1
```

```
## [1] 0.01222447
```

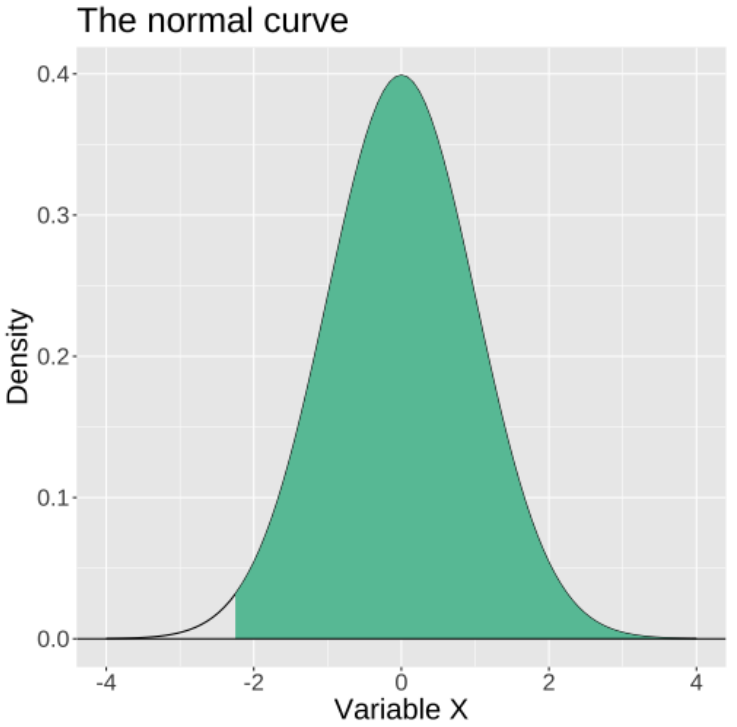


Using z-scores

Probability of getting a z-score of -2.25 or something larger

```
1-pnorm(q = -2.25, mean = 0, sd =
```

```
## [1] 0.9877755
```



Some z-scores of note

- $z = 1.64$; most extreme 5% of the standard normal distribution (the very far tail)
- $z = 1.96$; most extreme 2.5% of the standard normal distribution (used when splitting the difference of most positive and most negative extremes)