Normal Distribution

The **normal distribution** ("bell curve" or "Gaussian distribution") is a two-parameter distribution defined by the mean (μ) and standard deviation (σ) and having the following probability density function:

$$p(X|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} exp[-\frac{(X-\mu)^2}{2\sigma^2}]$$

$$X \sim N(\mu,\sigma)$$

Expected value

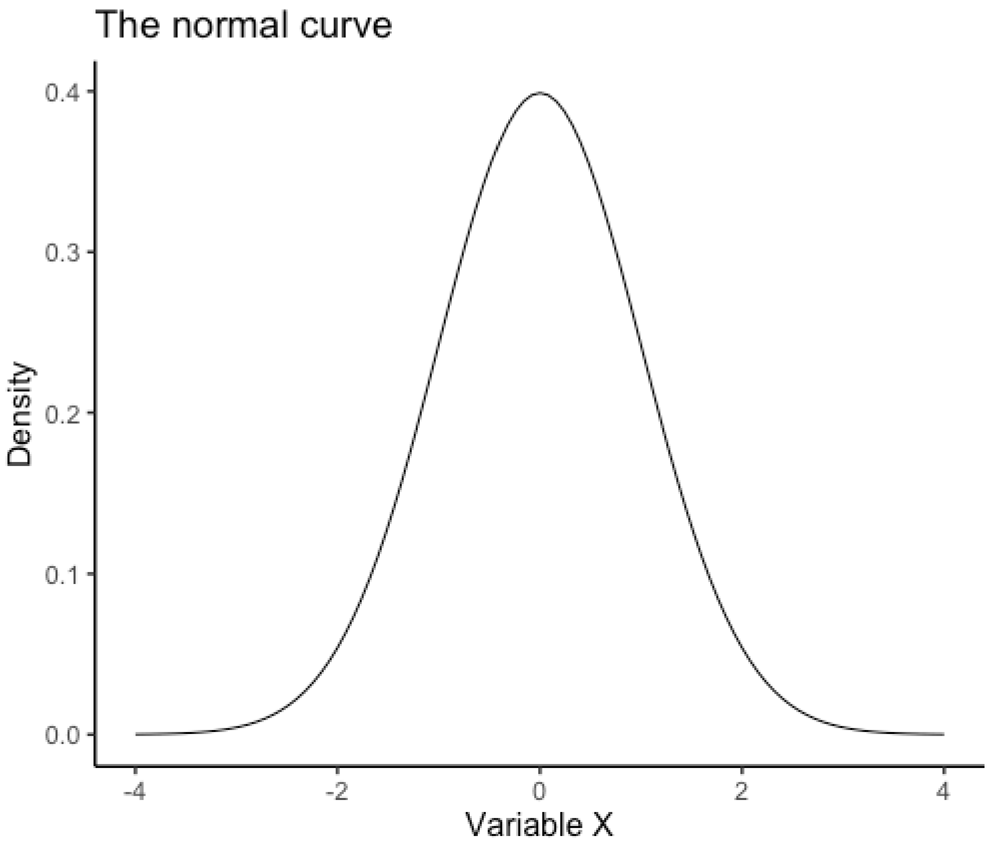
The expected value of the normal distribution is quite easy.

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

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.



Probabilities - same but different

Both the normal distribution and the binomial distribution follow the Law of Total Probability, but in different ways.

In the *binomial distribution*, each outcome in the sample space has a probability and these probabilities **sum to 1**.

In the *normal distribution*, there are an infinite number of values, each having a probability of 0, but the probability that some value will occur is 1. **The area under the curve is 1**.

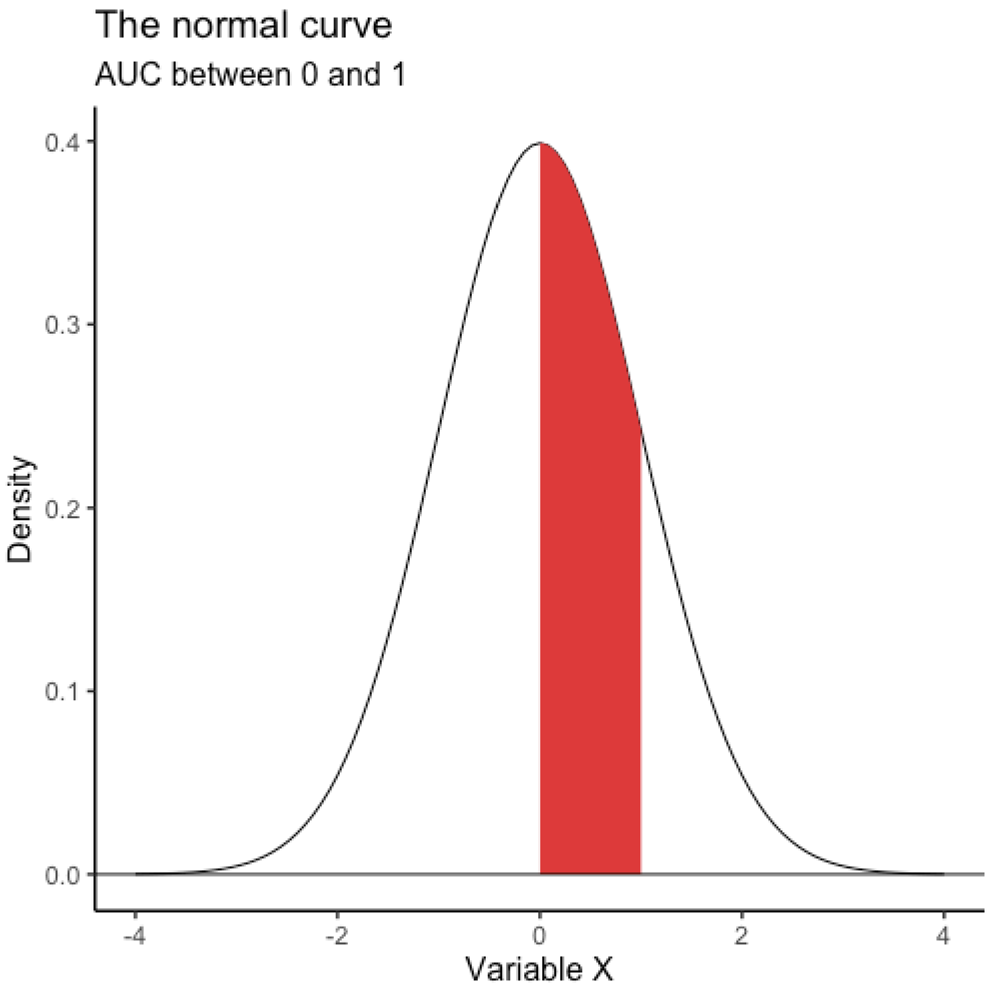
The density values are derived to insure that the area under the density curve is 1. They have no inherent meaning beyond that.

Probabilities - same but different

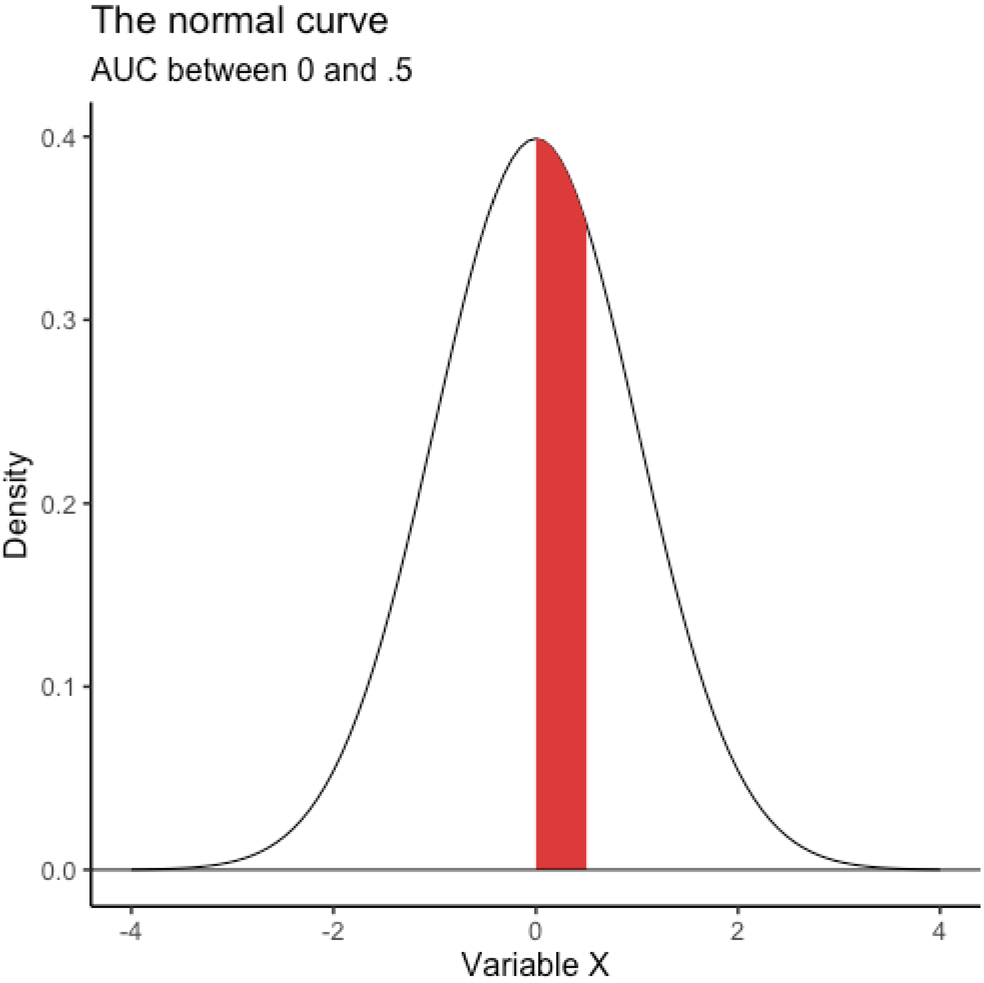
Density is mass per volume. In this context (curve in two dimensions) density is mass per area. The total density in the normal curve is 1.

The density for a part of the curve (a smaller area) will necessarily be less than 1.

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

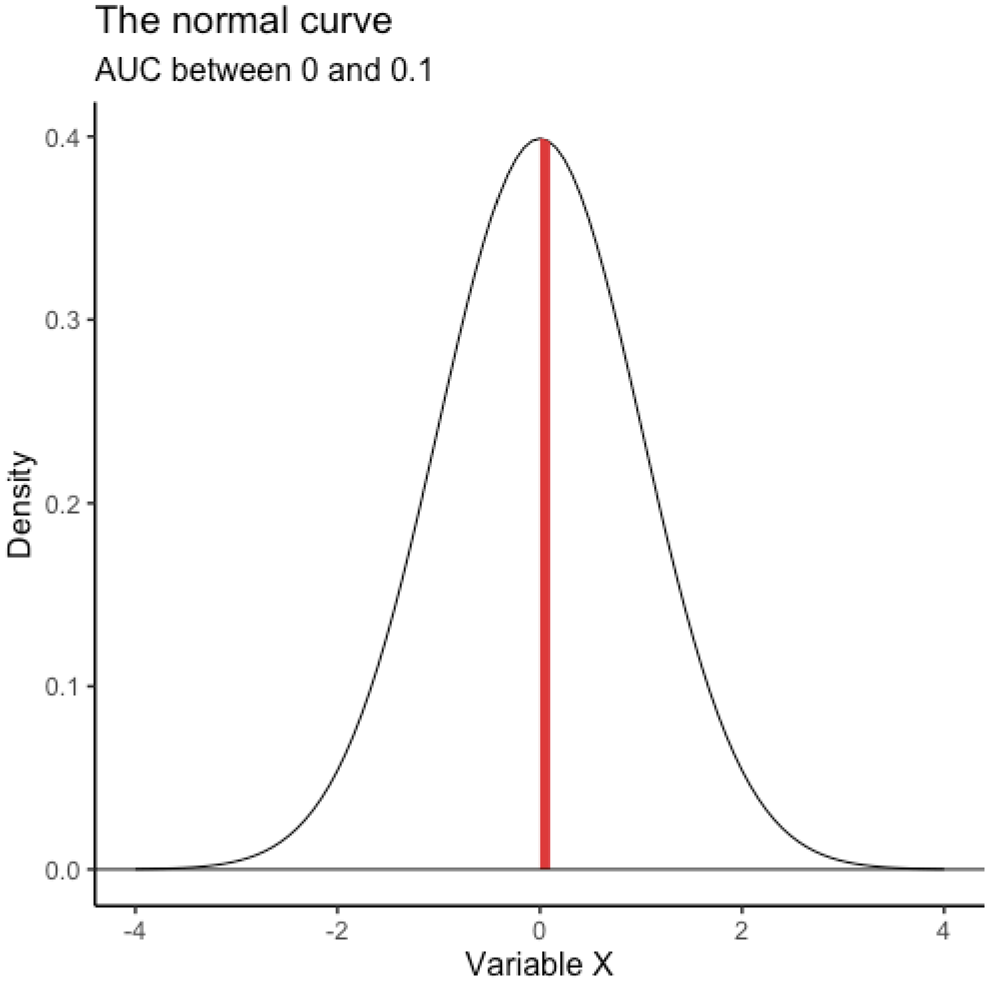


As we shrink that area by moving X closer to the mean, that probability interpretatic holds.



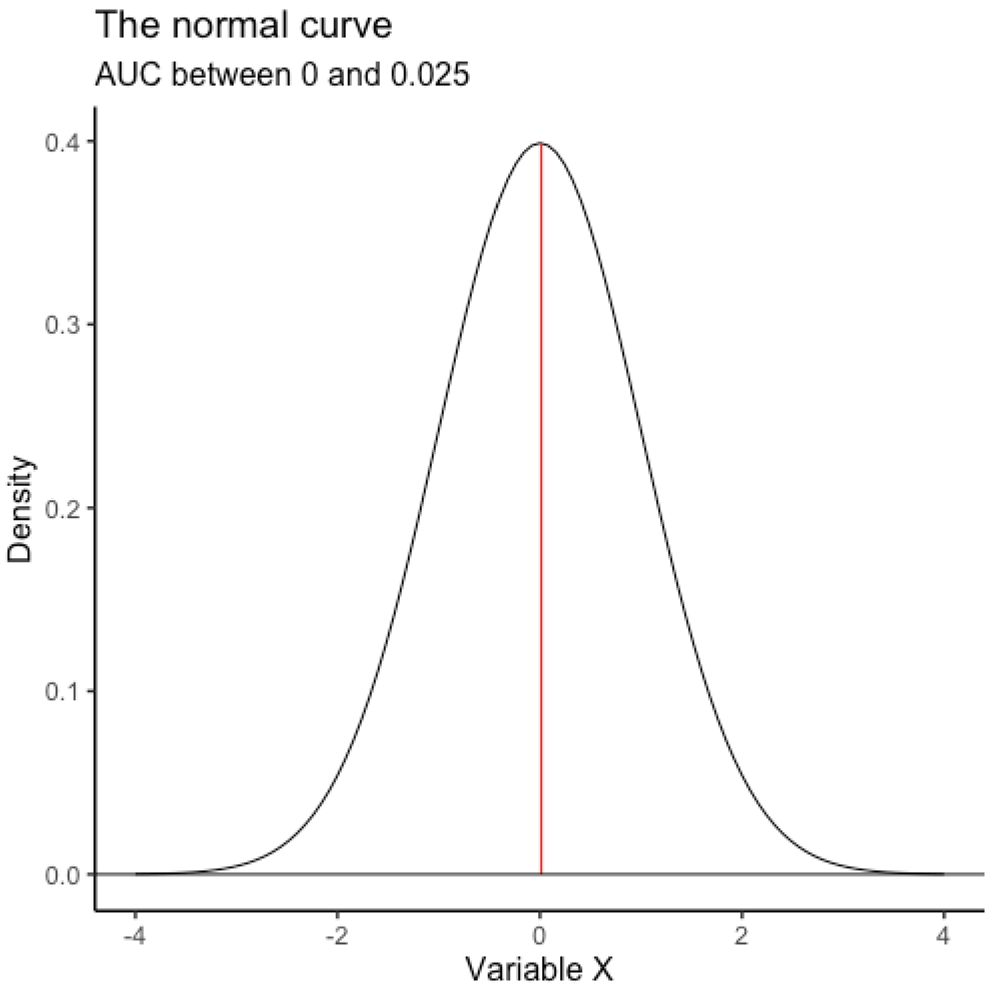
As our interval shrinks towards 0, our area shrinks as well.

It can get *very* close to 0—essentially a point rather than an area. The probability of that "point" is 0.



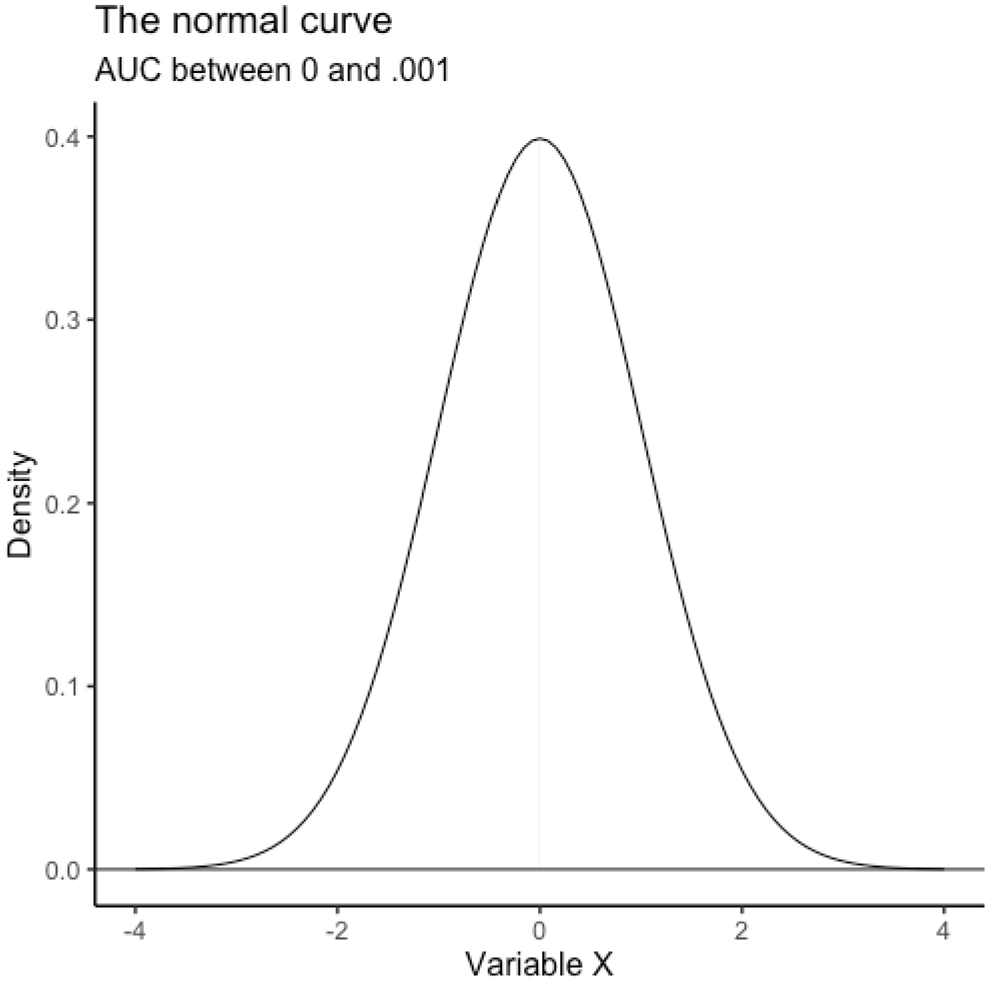
We can keep shrinking the distance between Z and 0, never reaching 0, and still calculate an area.

It will be very, very small.



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

It will be very, very small.

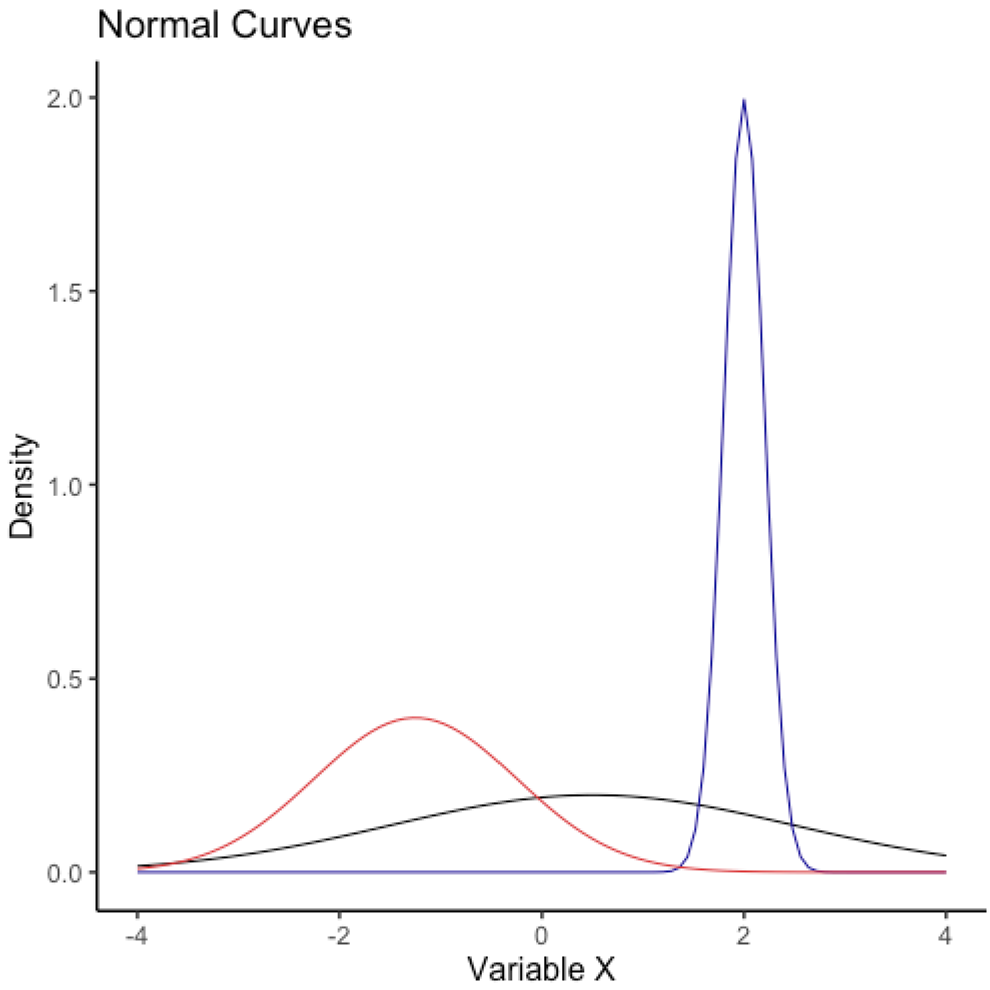


Characteristics of the normal distribution

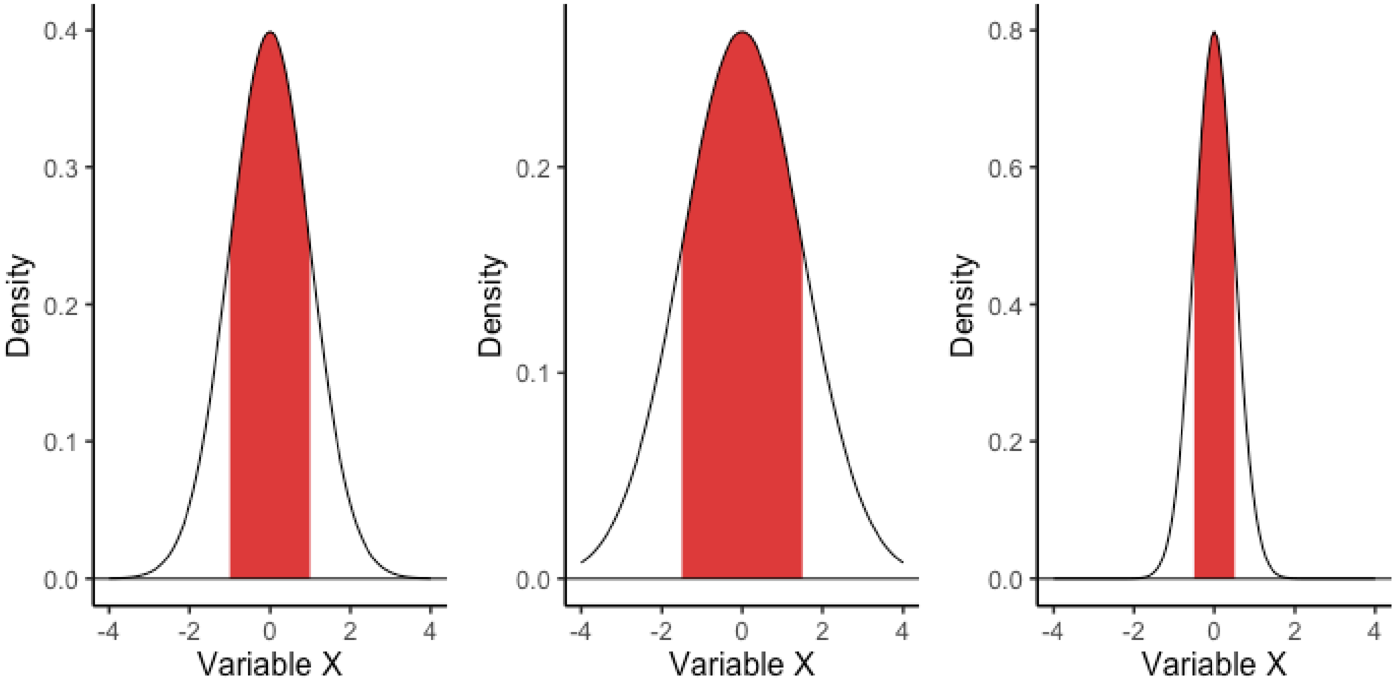
- The mean and standard deviation are independent
- The distribution is unimodal and symmetrical.

- For two normal distributions, the area under the curve between corresponding locations in standard deviation units is the same regardless of μ and σ.

    - For example, 68.3% of the area under a normal curve falls between 1 σ below the mean and 1 σ above mean—for every normal curve. This general characteristic is called the **Empirical Rule**.


Normal Curves
Density
2.0
1.5
1.0
0.5
0.0
-4
-2
0
2
4
Variable X

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



The Normal Distribution Is Cool

- Approximately 68% of the data in a normal distribution will be within one standard deviation of the mean.
- Approximately 95% of the data will be within two standard deviations of the mean.
- Approximately 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.

These benchmarks are convenient for determining if a score (and later, a mean) is rare or unusual in the context of a particular distribution.

Standard normal distribution

The normal distribution with μ=0 and σ=1 is called the **standard normal distribution** or the **Z distribution**.

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 extroverted (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

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)

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

Scores interpreted as distance from the mean, in standard deviations.

Properties of z-scores

- $\Sigma z = 0$
- $\Sigma z^2 = N$
- $s_z = \frac{\Sigma z^2}{n}$

Standardized scores (z-scores)

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

Why is this useful?

- Compare across scales and unit of measures
- More easily identify extreme data
- mean = 0, s = 1!

Using z-scores

```
## # A tibble: 6 × 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 × 2
##   name               height[,1]
##   <chr>                   <dbl>
## 1 Luke Skywalker          -0.07
## 2 C-3PO                   -0.22
## 3 R2-D2                   -2.26
## 4 Darth Vader              0.79
## 5 Leia Organa             -0.71
## 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.)*

Check out this z-table

- Left column is the z-score to the tenths place in the decimal
- Top is how you can add on to the hundreths place in the decimal

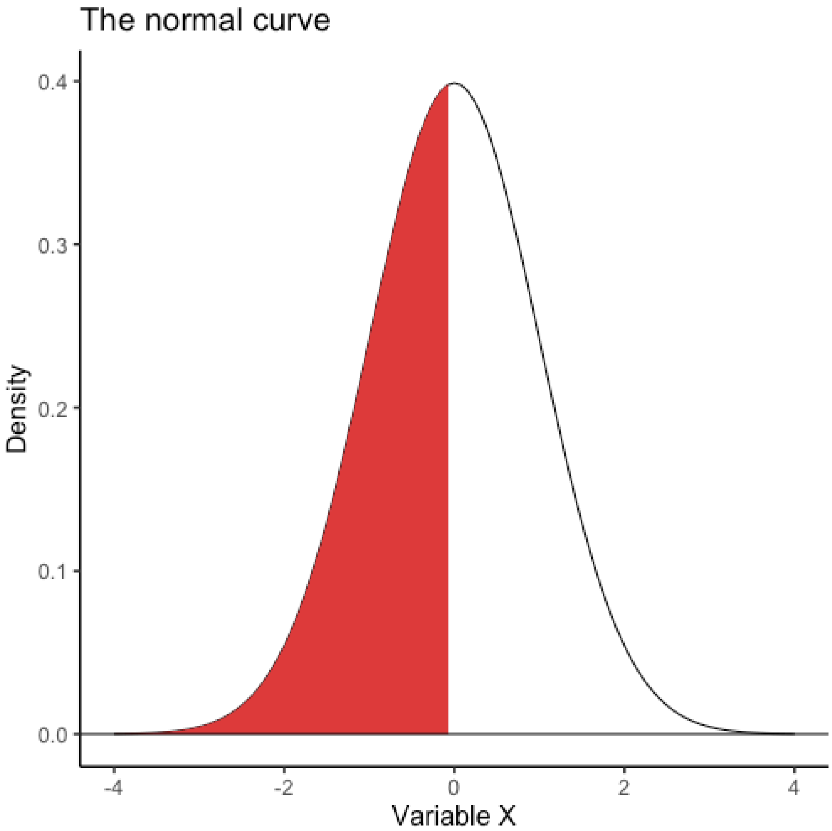
Luke Skywalker's height is z = -.07

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

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

What does .4721 mean?

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



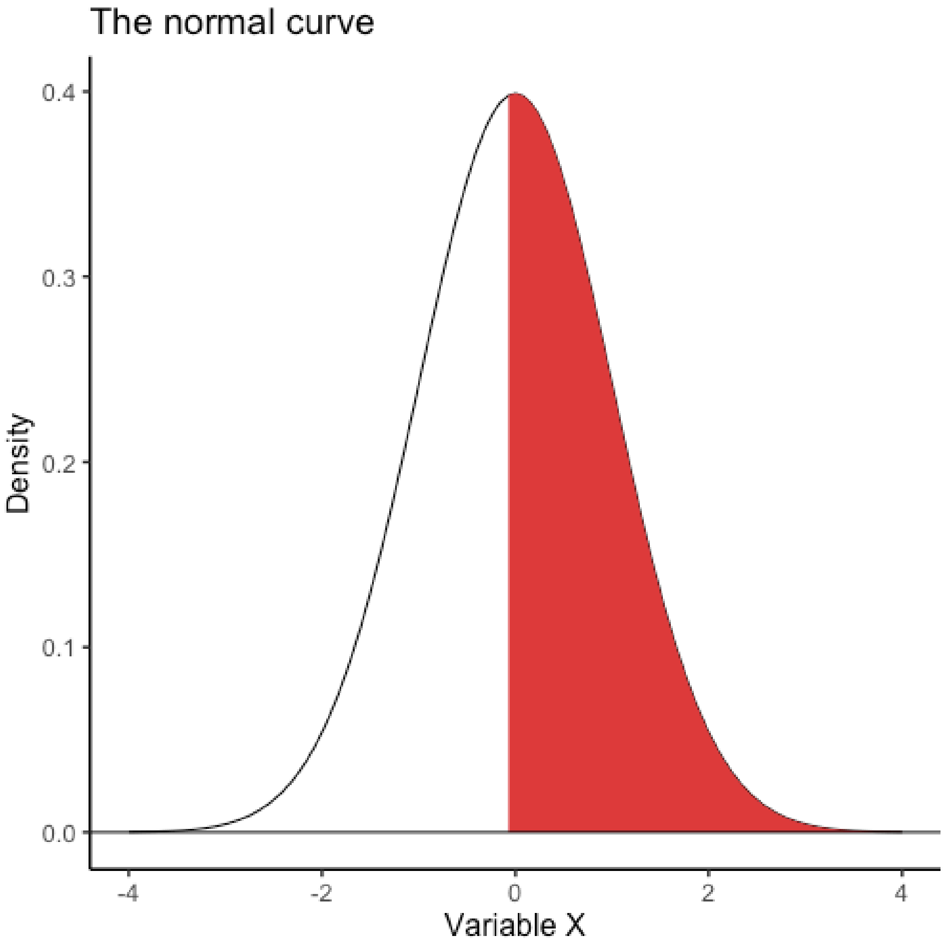
What do we want?

- The probability of obtaining a z-score of -.07 or something more extreme
- This could be more extreme in the positive direction (moving right)
- This could be more extreme in the negative direction (moving left)
- You could split the difference and it can be in either direction

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

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

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

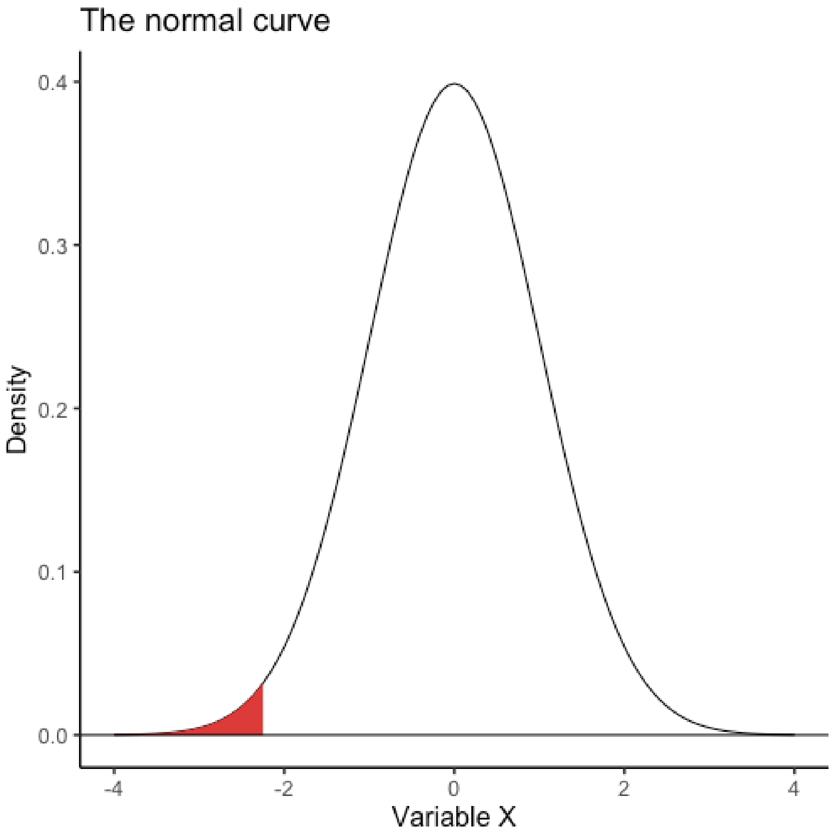


What about R2D2? (z-score of -2.25)

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

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

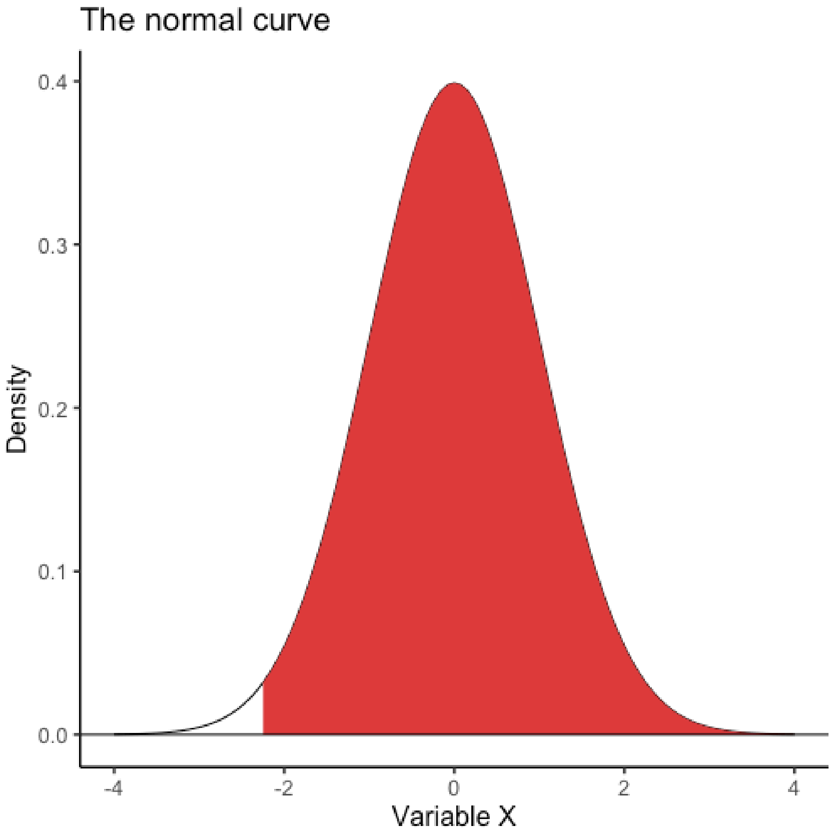
```
## [1] 0.01222447
```



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

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

```
## [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)

Next time...

- Sampling distributions and expansions of the normal distribution