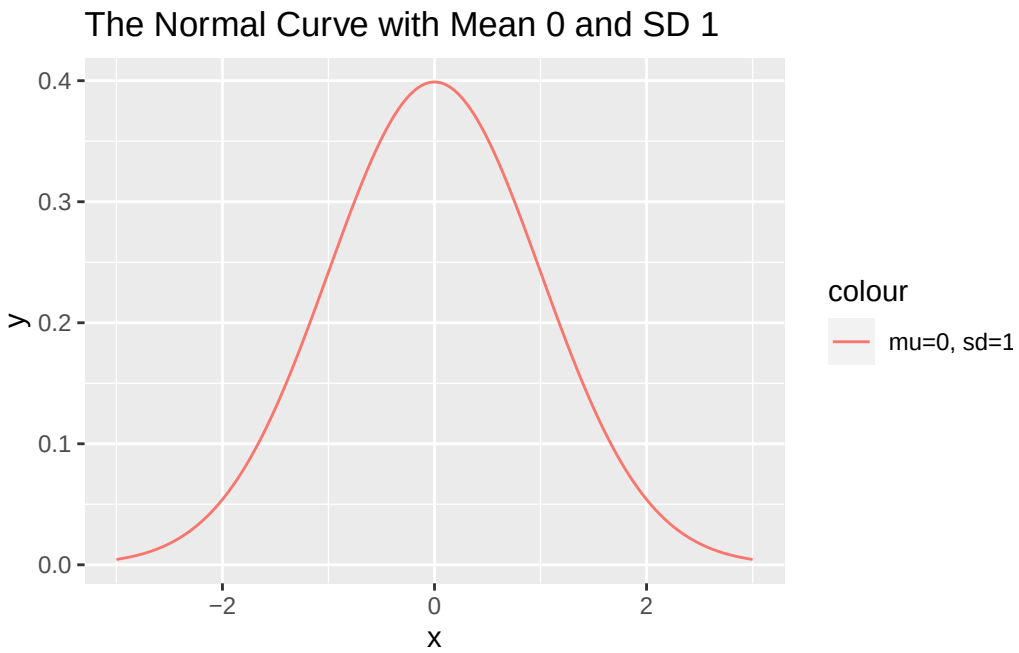
The normal distribution

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

The normal distribution is governed by the famous bell-shaped curve.

```
normal <- tibble(x = seq(-3, 3, .01), y = dnorm(x, 0, 1), y2 = dnorm(x, 0, .5))
plt <- ggplot(data = normal) +
    geom_line(aes(x = x, y = y, color = "mu=0, sd=1")) +
    labs(title = "The Normal Curve with Mean 0 and SD 1")
plt
```

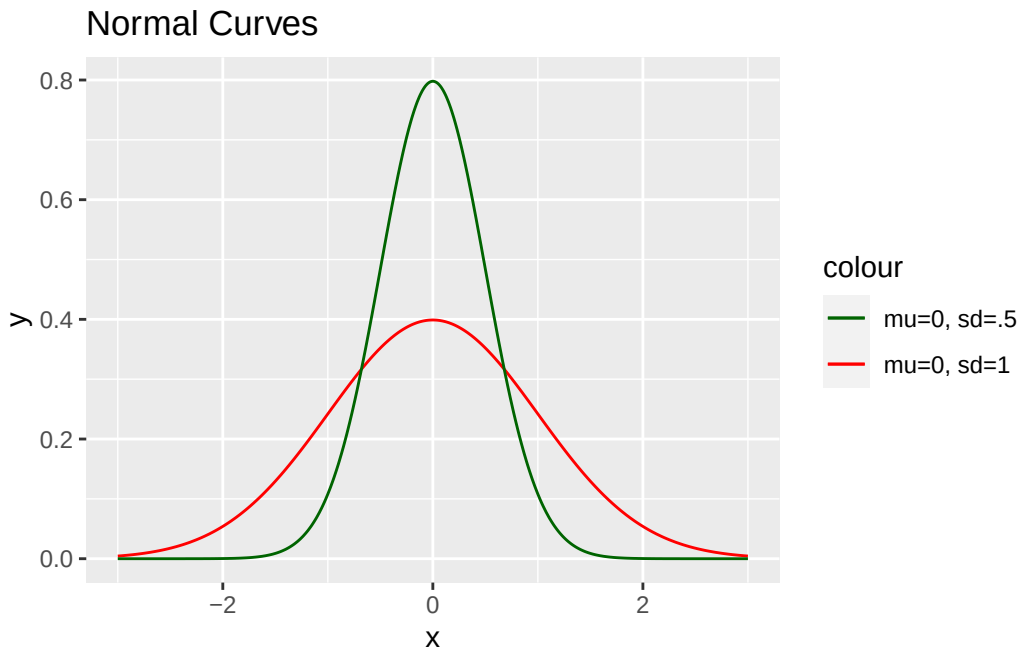


The two key parameters are the mean and the standard deviation, which govern the position of the peak and the "spread" of the curve respectively.

```
normal |> mutate(y2 = dnorm(`x`, 0, .5))
```

```
# A tibble: 601 x 3
       x       y           y2
   <dbl>   <dbl>        <dbl>
 1 -3    0.00443 0.0000000122
 2 -2.99 0.00457 0.0000000137
 3 -2.98 0.00470 0.0000000154
 4 -2.97 0.00485 0.0000000174
 5 -2.96 0.00499 0.0000000196
 6 -2.95 0.00514 0.0000000220
 7 -2.94 0.00530 0.0000000248
 8 -2.93 0.00545 0.0000000279
 9 -2.92 0.00562 0.0000000313
10 -2.91 0.00578 0.0000000352
# i 591 more rows
```

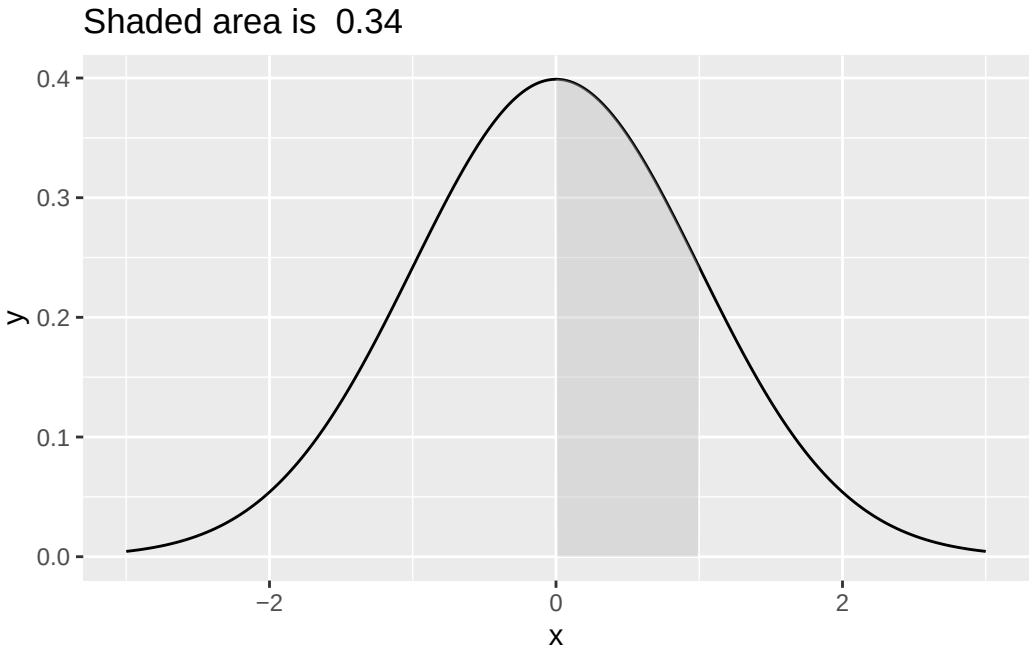
```
plt <- plt + geom_line(aes(x = x, y = y2, color = "mu=0, sd=.5")) +
       scale_color_manual(
            values = c("mu=0, sd=1" = "red", "mu=0, sd=.5" = "darkgreen")
        ) +
        labs(title = "Normal Curves")
plt
```



Probabilities

The normal curve is a *probability density*. This means that the chance that a number x drawn from a population with a normal distribution with mean μ and standard deviation σ lies between a and b is the area under the curve between $x = a$ and $x = b$.

```
area <- pnorm(1) - pnorm(0)
plt <- ggplot(data = normal) +
    geom_line(aes(x = x, y = y)) +
    geom_ribbon(
        aes(x = x, ymin = ifelse(x > 0 & x < 1, 0, y), ymax = y),
        fill = "gray",
        alpha = .4
    ) +
    labs(title = paste("Shaded area is ", format(area, nsmall = 2, digits = 2)))
plt
```



In the example above, the shaded area between 0 and 1 is .34. Let's draw 1000 numbers from the normal distribution.

```
# count draws a 1000 points from a standard normal
# and counts how many lie between 0 and 1
# this is a binomal variable with N=1000 and p=.34

count <- function() {
    x <- rnorm(1000, 0, 1)
    sum(x >= 0 & x <= 1)
}

# now we do the experiment 100 times and record the results
draws <- replicate(100, count())
```