Basic of inference - implementation

E.Pastucha

2024-08-28

```
library(tidyverse)
```

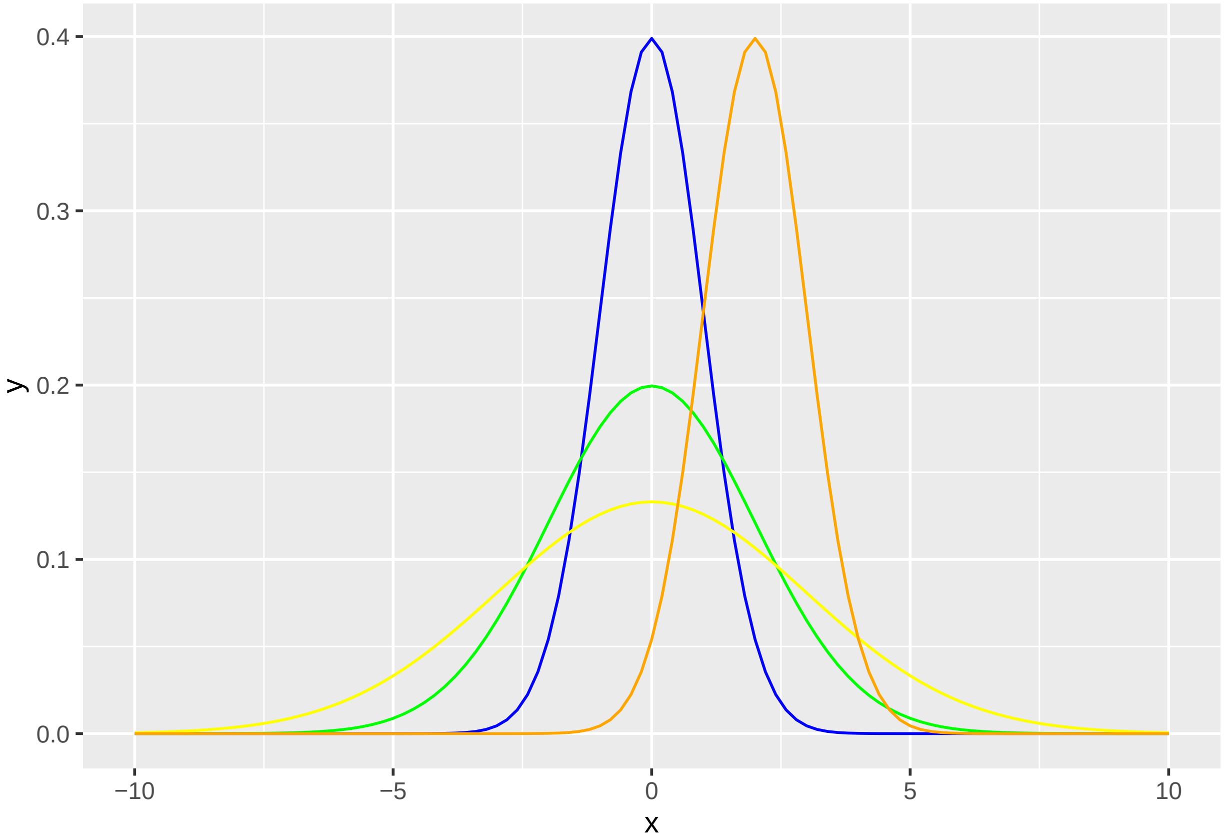
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ 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 become error
```

Normal distribution

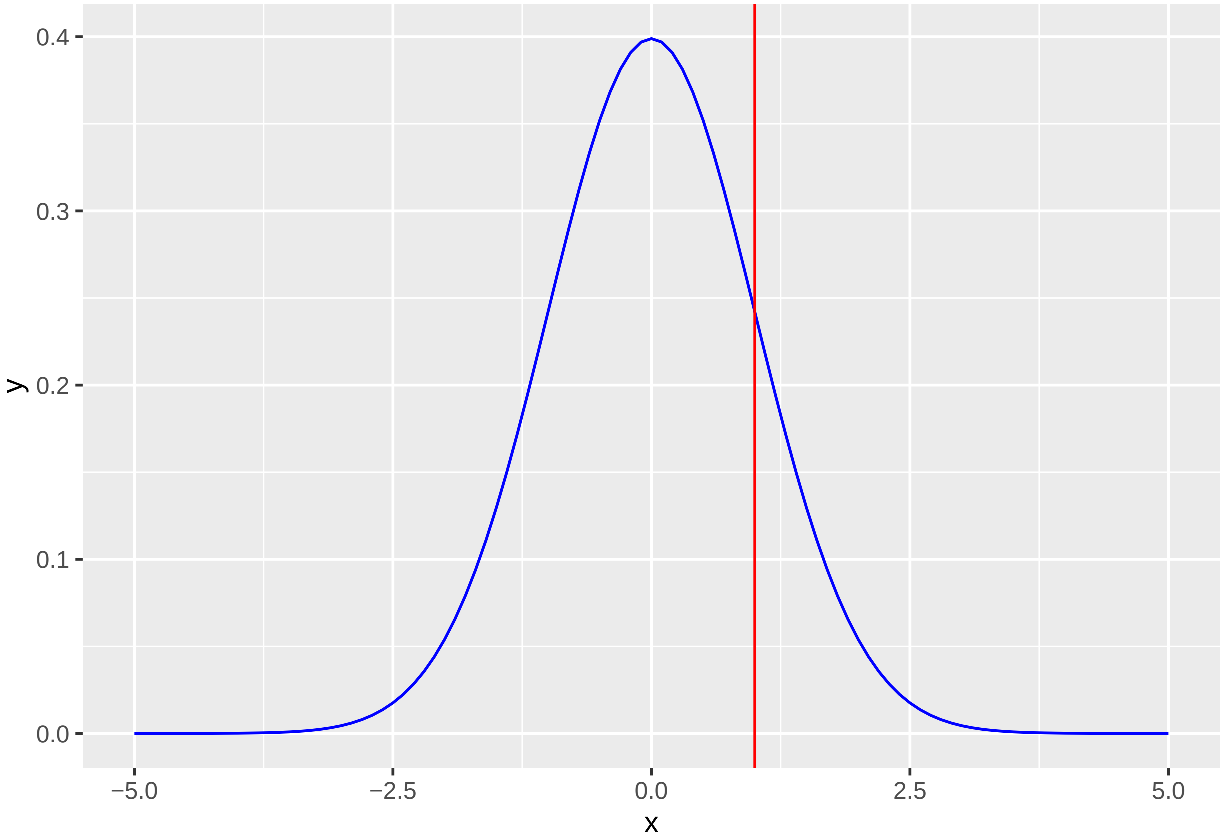
dnorm(x, mean, sd) - distribution function pnorm(x, mean, sd) - probability, qnorm(probability, mean, sd) - given probability what is the value rnorm(how_many, mean, sd) - generate data

default values: mean = 0, sd = 1

```
ggplot(data.frame(x = seq(-10, 10, length = 100)), aes(x = x))+
  stat_function(fun = dnorm, args = list(mean = 0, sd = 1), color = 'blue') +
  stat_function(fun = dnorm, args = list(mean = 0, sd = 2), color = 'green') +
  stat_function(fun = dnorm, args = list(mean = 0, sd = 3), color = 'yellow') +
  stat_function(fun = dnorm, args = list(mean = 2, sd = 1), color = 'orange')
```

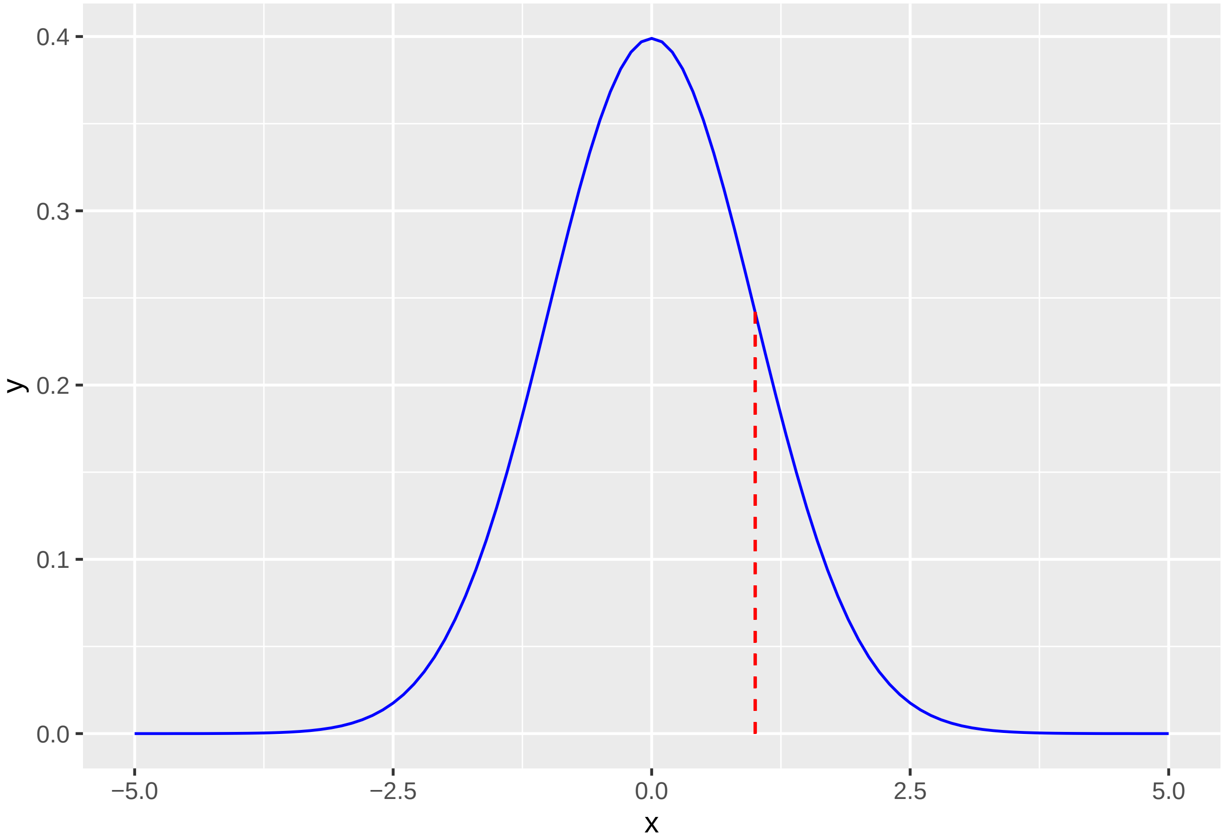


```
ggplot(data.frame(x = seq(-5, 5, length = 100)), aes(x = x))+
  stat_function(fun = dnorm, args = list(mean = 0, sd = 1), color = 'blue') +
  geom_vline(xintercept = 1, color ='red')
```



```
ggplot(data.frame(x = seq(-5, 5, length = 100)), aes(x = x))+
  stat_function(fun = dnorm, args = list(mean = 0, sd = 1), color = 'blue') +
  geom_segment(aes(x = 1, y = 0,
             xend = 1, yend = dnorm(1)), color ='red', linetype = 'dashed')
```

```
## Warning in geom_segment(aes(x = 1, y = 0, xend = 1, yend = dnorm(1)), color = "red", : All aesthetics
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
```



Sharks

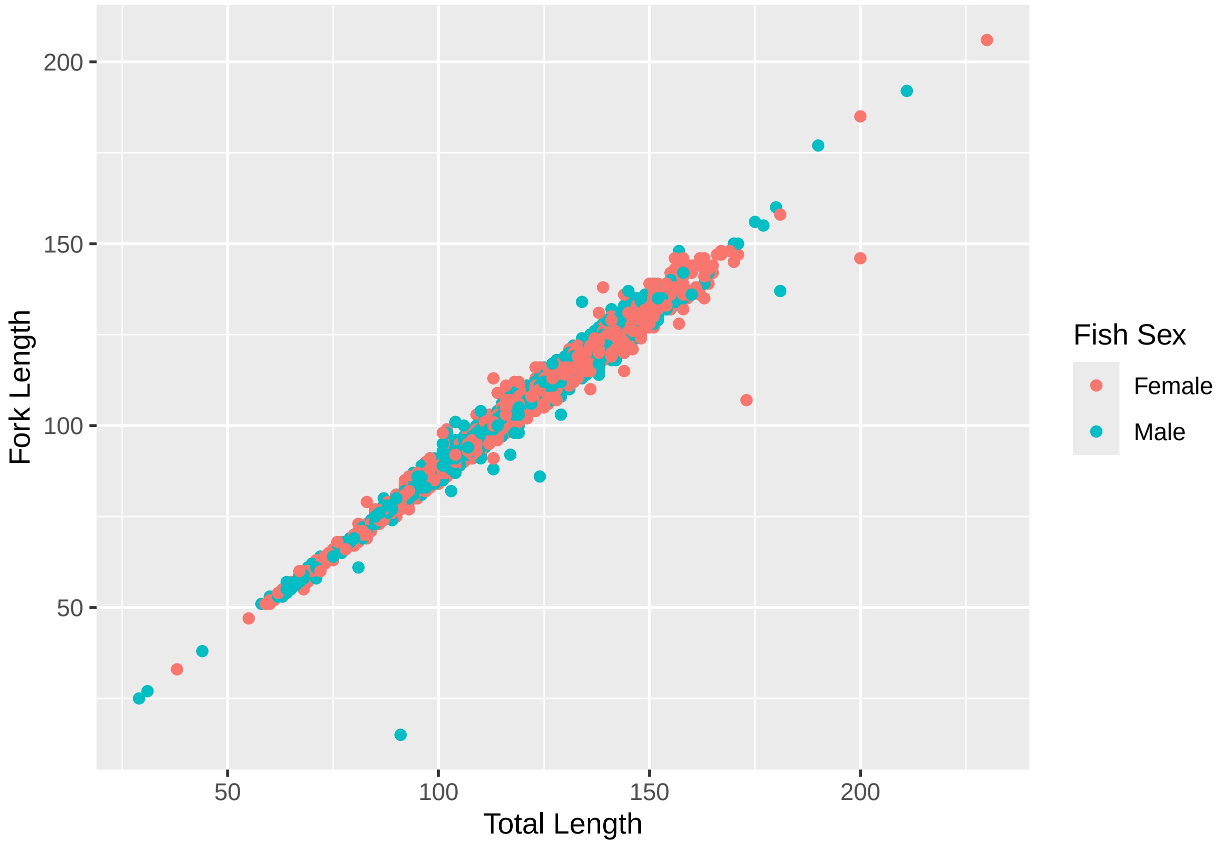
```
sharks <- readr::read_csv('sharks.csv')
```

```
## Rows: 2510 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Fish Sex
## dbl (3): Calendar Year, Total Length, Fork Length
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

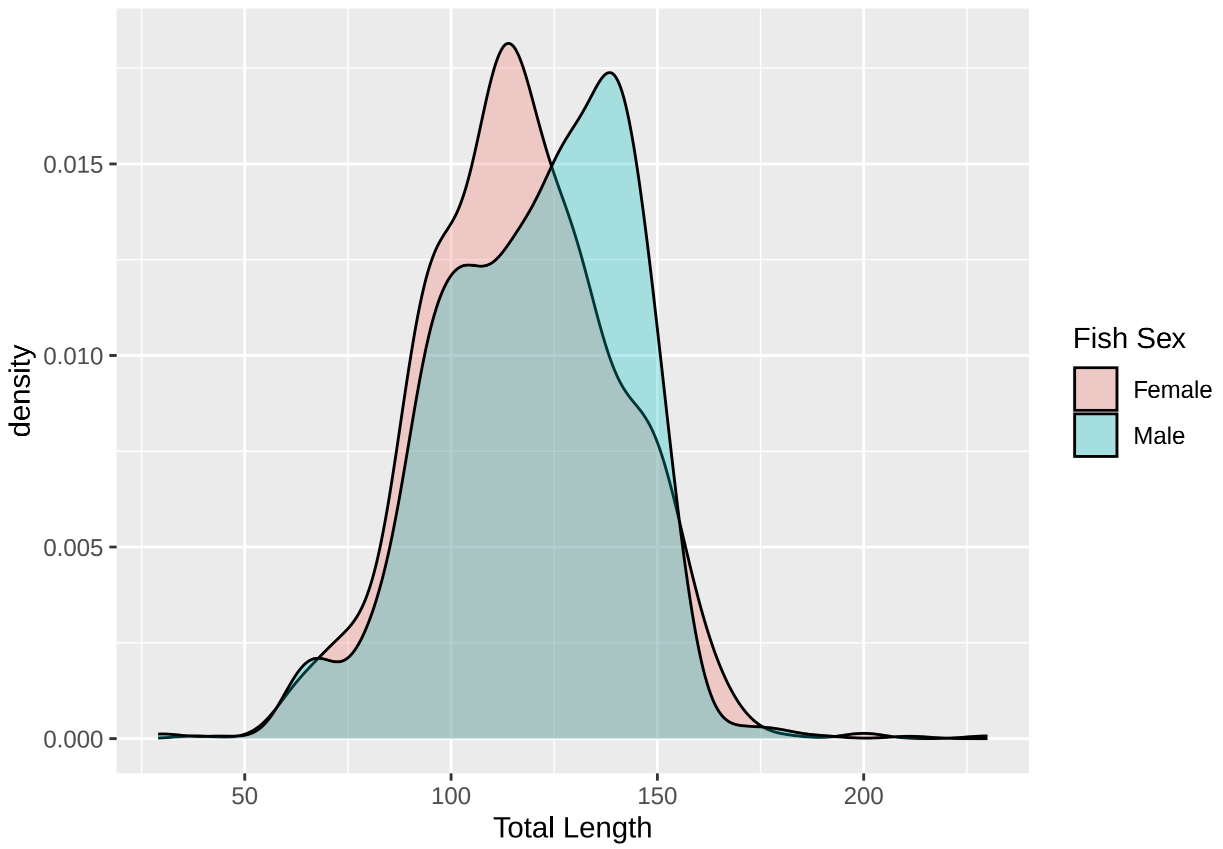
```
head(sharks, 5)
```

```
## # A tibble: 5 x 4
##   ‘Calendar Year‘ ‘Fish Sex‘ ‘Total Length‘ ‘Fork Length‘
##             <dbl> <chr>               <dbl>         <dbl>
## 1            2007 Male                  106            94
## 2            2007 Female                102            92
## 3            2007 Female                 87            75
## 4            2007 Female                133           116
## 5            2007 Female                 84            71
```

```
ggplot(sharks) +
  geom_point(aes(x = `Total Length`, y = `Fork Length`, color = `Fish Sex`))
```



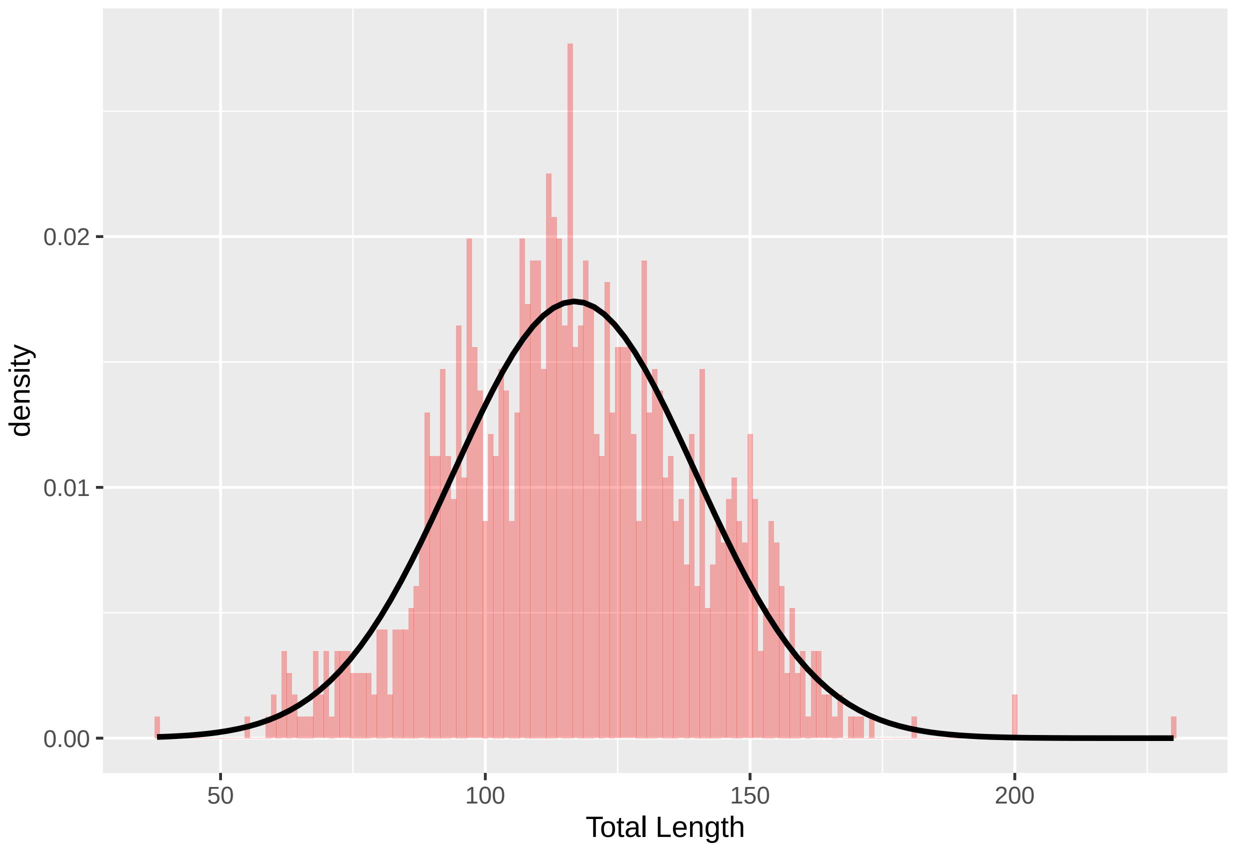
```
ggplot(sharks) +
  geom_density(aes(x = `Total Length`, fill = `Fish Sex`), alpha = 0.3)
```



```
females <- filter(sharks, `Fish Sex` == 'Female')
```

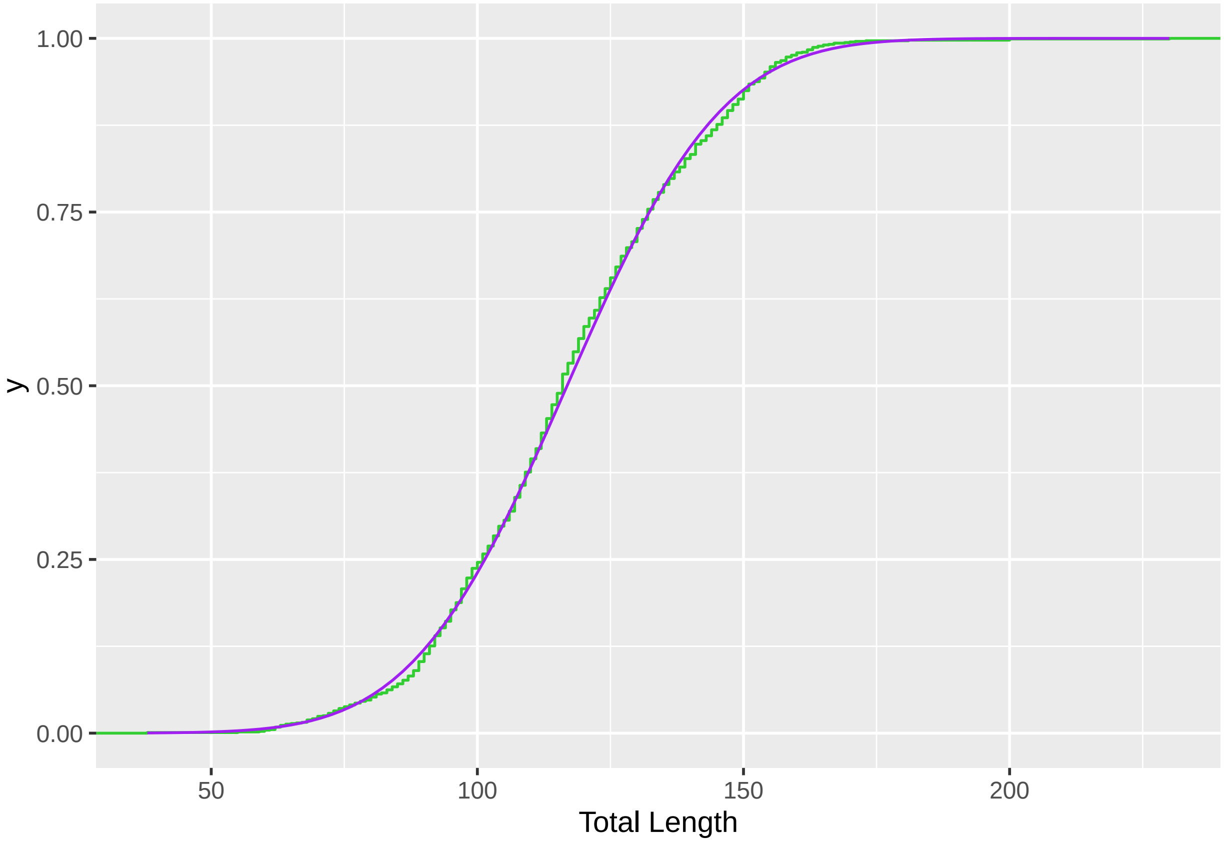
```
mean_fs <- mean(females$`Total Length`)
sd_fs <- sd(females$`Total Length`)
ggplot(females) +
  geom_histogram(aes(x = `Total Length`, y = after_stat(density)),
                 fill = 'red', alpha = 0.3, binwidth = 1) +
  stat_function(fun = dnorm, args = list(mean = mean_fs, sd = sd_fs), size = 1)
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



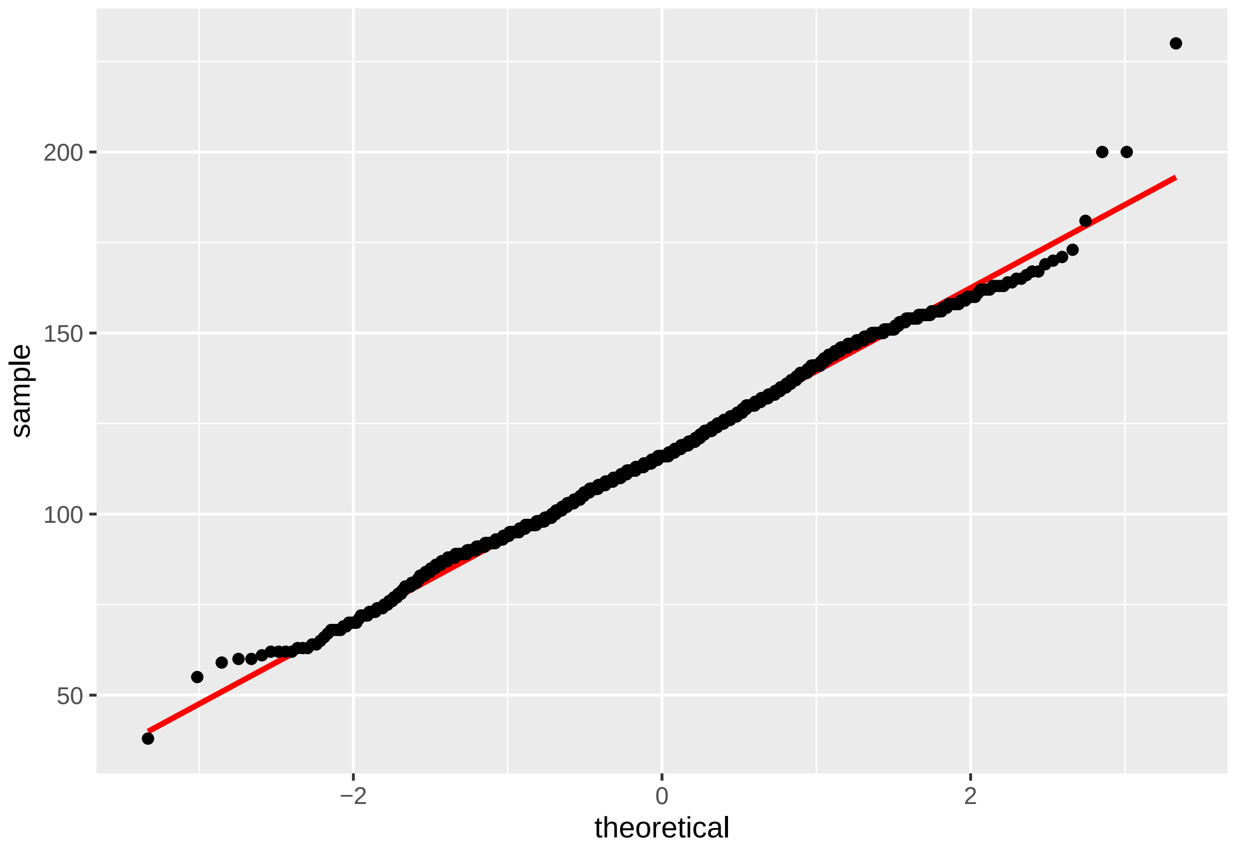
# ECDF

```r
ggplot(females) +
  stat_ecdf(aes(x = `Total Length`), color = 'limegreen') +
  geom_line(stat = 'function', fun = pnorm, args = list(mean = mean_fs, sd = sd_fs),
            color = 'purple')
```



Q-Q plot

```
ggplot(females) +
  stat_qq_line(aes(sample = `Total Length`), color = 'red', size = 1) +
  stat_qq(aes(sample = `Total Length`))
```



Z-scores

150 inches long female shark.

```
(z_score <- (150 - mean_fs)/sd_fs)
```

```
## [1] 1.446787
```

```
pnorm(z_score, mean = 0, sd = 1)
```

```
## [1] 0.9260217
```

```
pnorm(z_score)
```

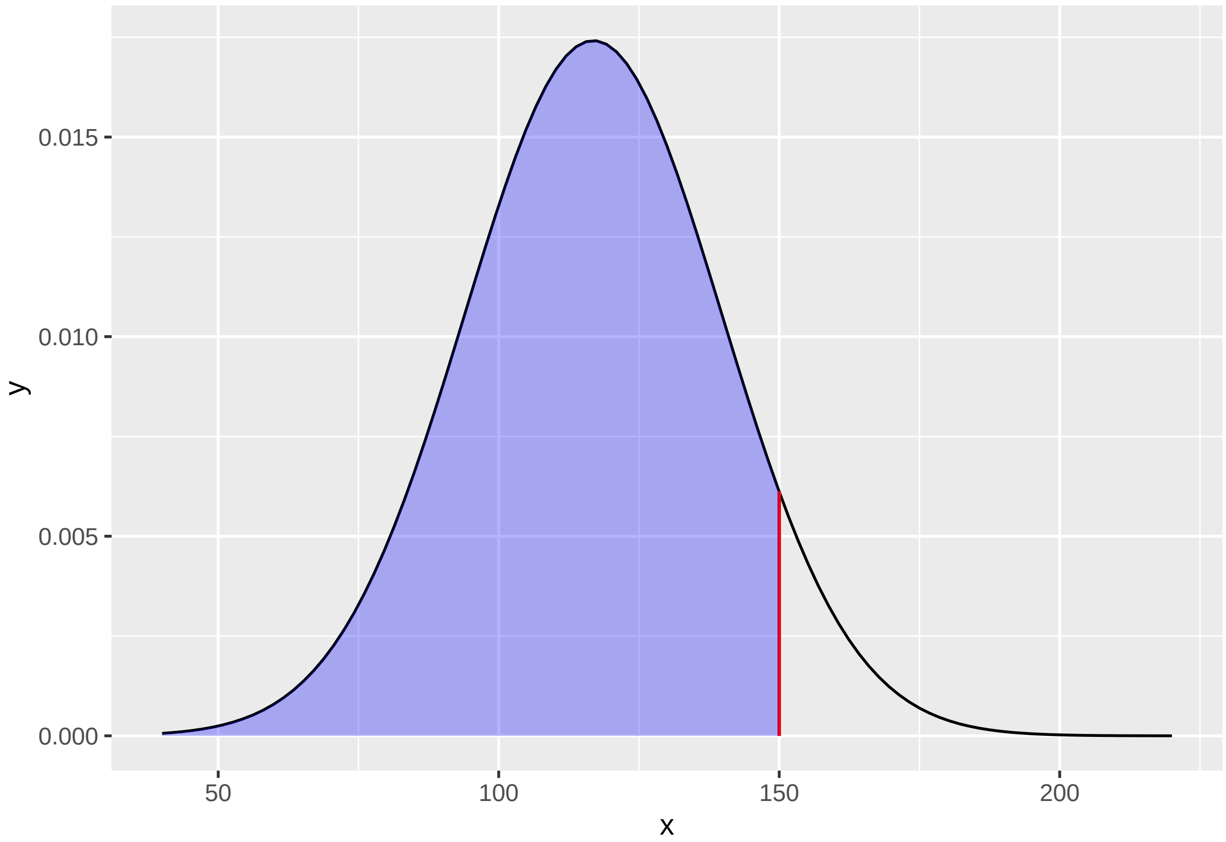
```
## [1] 0.9260217
```

```
pnorm(150, mean = mean_fs, sd = sd_fs)
```

```
## [1] 0.9260217
```

```r
ggplot(data.frame(x = seq(40, 220, length = 500)), aes(x=x)) +
  stat_function(fun = dnorm, args = list(mean = mean_fs, sd = sd_fs)) +
  geom_segment(aes(x = 150, y = 0,
                   xend = 150, yend = dnorm(150, mean = mean_fs, sd = sd_fs)), color = 'red') +
  geom_area(stat = 'function', fun = dnorm, args = list(mean = mean_fs, sd = sd_fs),
            fill = 'blue', xlim = c (40, 150), alpha = 0.3)
```

```
## Warning in geom_segment(aes(x = 150, y = 0, xend = 150, yend = dnorm(150, : All aesthetics have leng
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
```



SE

Mean female shark Total Length. 1155 cases, 40-220 inches.

```r
mean_fs
```

```
## [1] 116.8606
```

Central Limit Theorem: * Samples are independent * Sample size is bigger than 30 * Population distribution is not strongly skewed

YES!

$$SE = \frac{\sigma}{\sqrt{n}}$$

```
(SE <- sd_fs/sqrt(nrow(females)))
```

```
## [1] 0.6739831
```

95% confidence interval.

1.96 qnorm()

```
mean_fs - 1.96 *SE
```

```
## [1] 115.5396
```

```
mean_fs + qnorm(0.025)*SE
```

```
## [1] 115.5396
```

```
mean_fs + 1.96*SE
```

```
## [1] 118.1816
```

```
mean_fs + qnorm(0.975)*SE
```

```
## [1] 118.1816
```

We are 95% confident that population mean female shark total length is in between 115.54 inch and 118.18 inch.

99% confidence interval

```
mean_fs - 2.58 *SE
```

```
## [1] 115.1217
```

```
mean_fs + qnorm(0.005)*SE
```

```
## [1] 115.1245
```

```
mean_fs + 2.58*SE
```

```
## [1] 118.5995
```

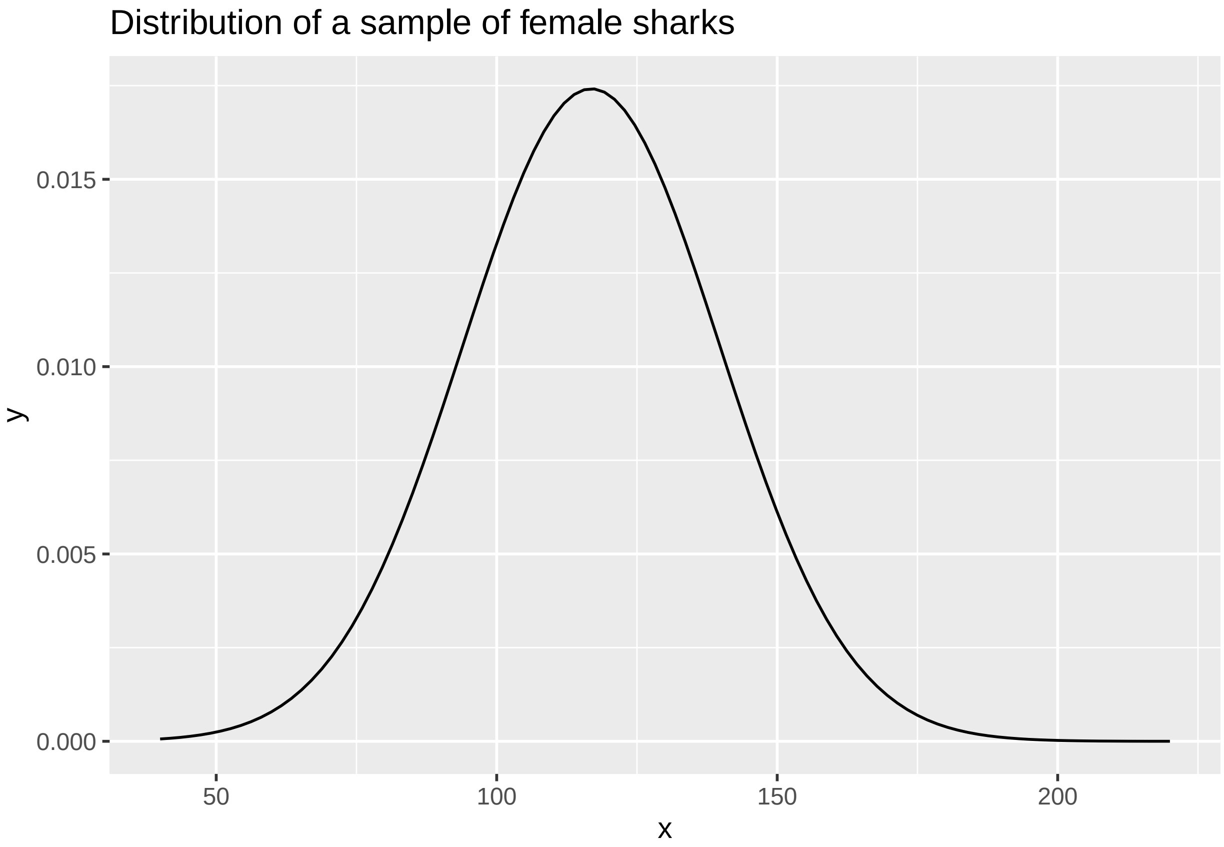
```
mean_fs + qnorm(0.995)*SE
```

```
## [1] 118.5967
```

We are 99% confident that population mean total length of a female shark is between 115.12 inch and 118.6 inch.

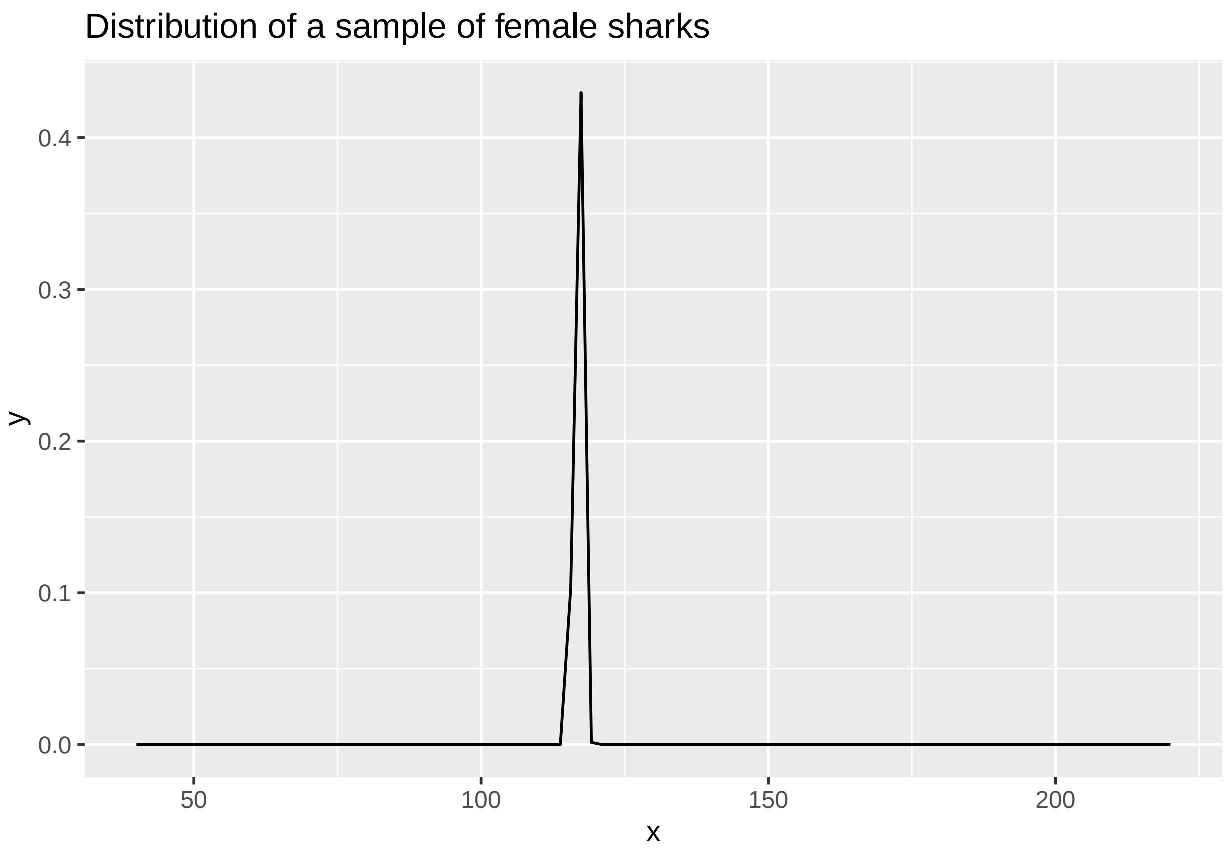
Distribution of a sample - one sample (ONE!)

```
ggplot(data.frame(x = seq(40, 220, length = 500)), aes(x = x)) +
  stat_function(fun = dnorm, args = list(mean = mean_fs, sd = sd_fs))+
labs(title = 'Distribution of a sample of female sharks ')
```

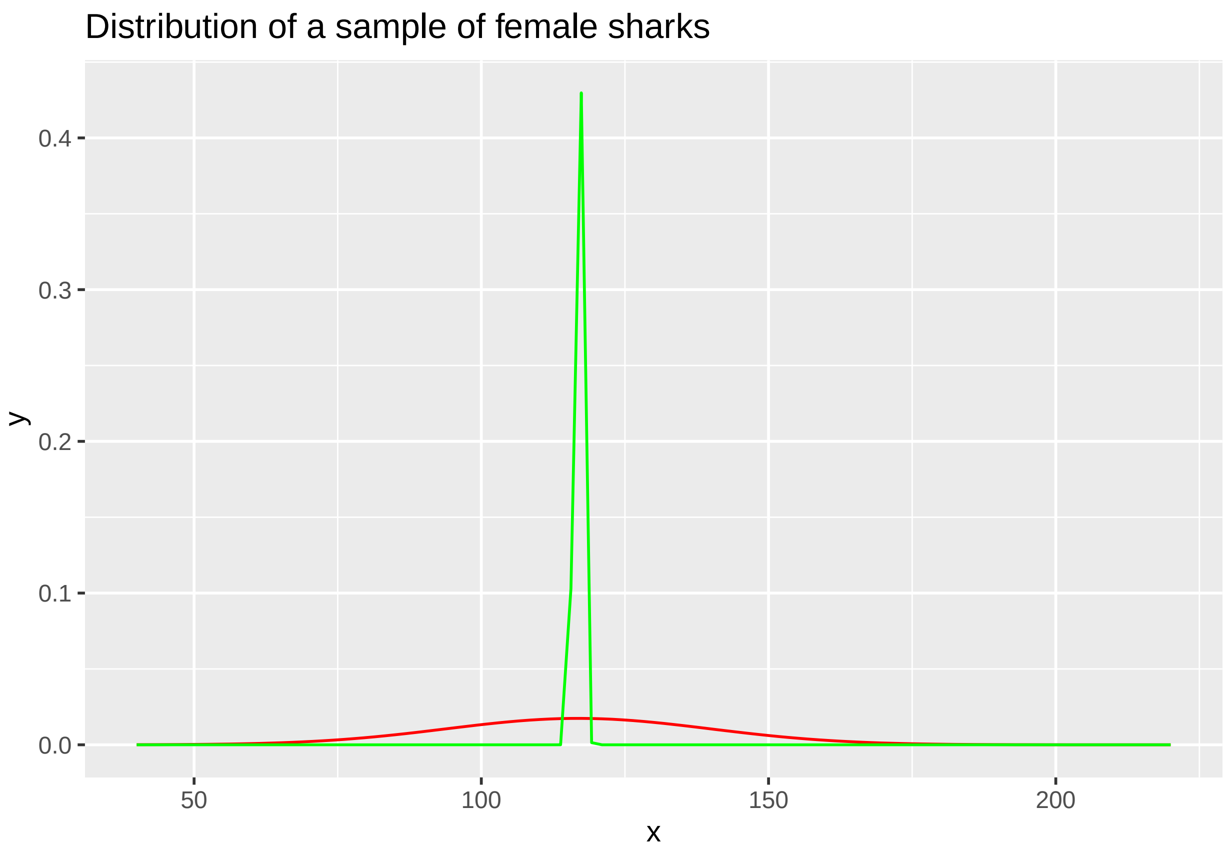


Sampling Distribution - multiple samples.

```
ggplot(data.frame(x = seq(40, 220, length = 500)), aes(x = x)) +
  stat_function(fun = dnorm, args = list(mean = mean_fs, sd = SE))+
labs(title = 'Distribution of a sample of female sharks ')
```



```r
ggplot(data.frame(x = seq(40, 220, length = 500)), aes(x = x)) +
  stat_function(fun = dnorm, args = list(mean = mean_fs, sd = sd_fs), color = 'red')+
  stat_function(fun = dnorm, args = list(mean = mean_fs, sd = SE), color = 'green')+
  labs(title = 'Distribution of a sample of female sharks ')
```


Distribution of a sample of female sharks
y
0.0
0.1
0.2
0.3
0.4
x
50
100
150
200