Tarea_06

Ita Santiago

20/9/2020

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.1
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

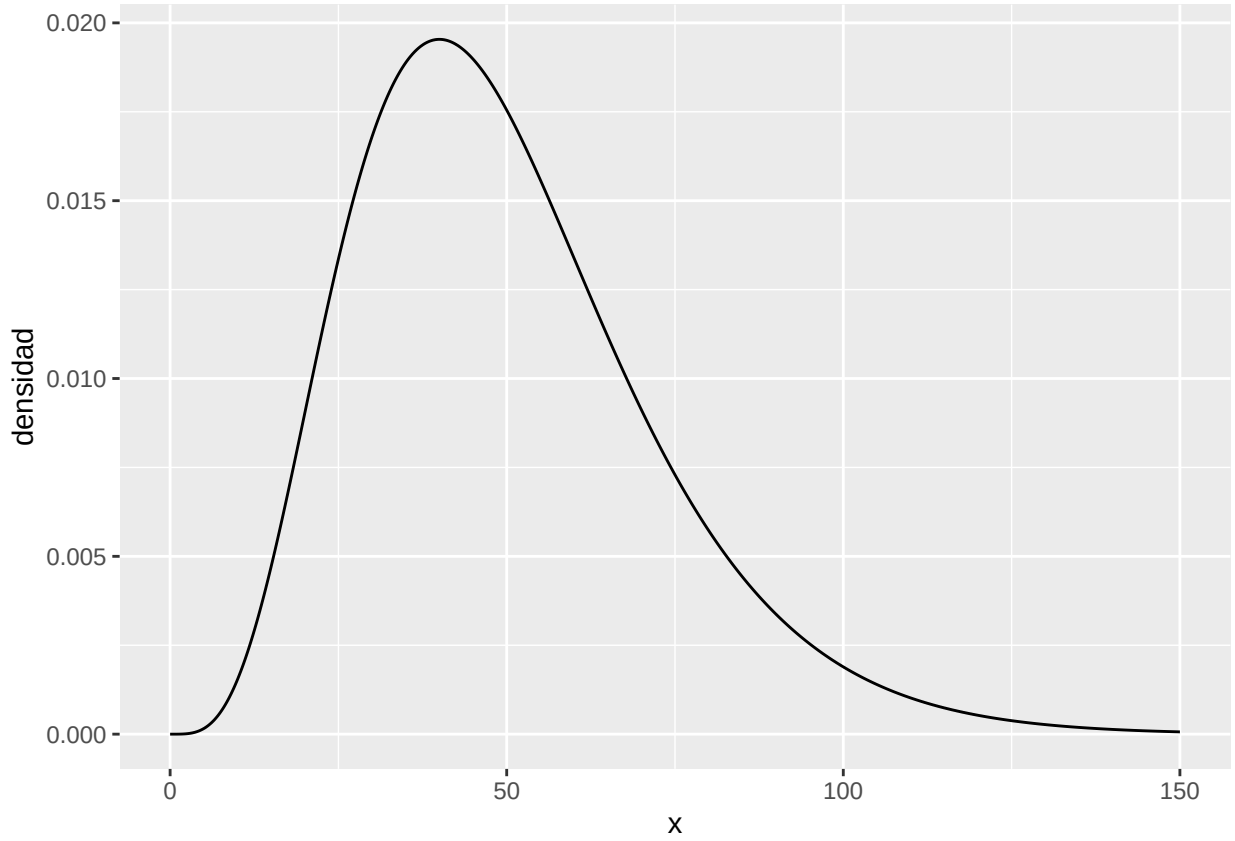
```
library(patchwork)
```

Teorema central del límite

Ejemplo 1

Consideramos la distribución gamma con parÃ¡metro de forma $a = 5$, tasa $\lambda = 0.1$. Su media teórica es 50 $= 5/0.1$ cuya densidad teórica es

```
x <- seq(0, 150, 0.01)
tibble(x = x) %>%
  mutate(densidad = dgamma(x, 5, 0.1)) %>%
  ggplot(aes(x = x, y = densidad)) + geom_line()
```



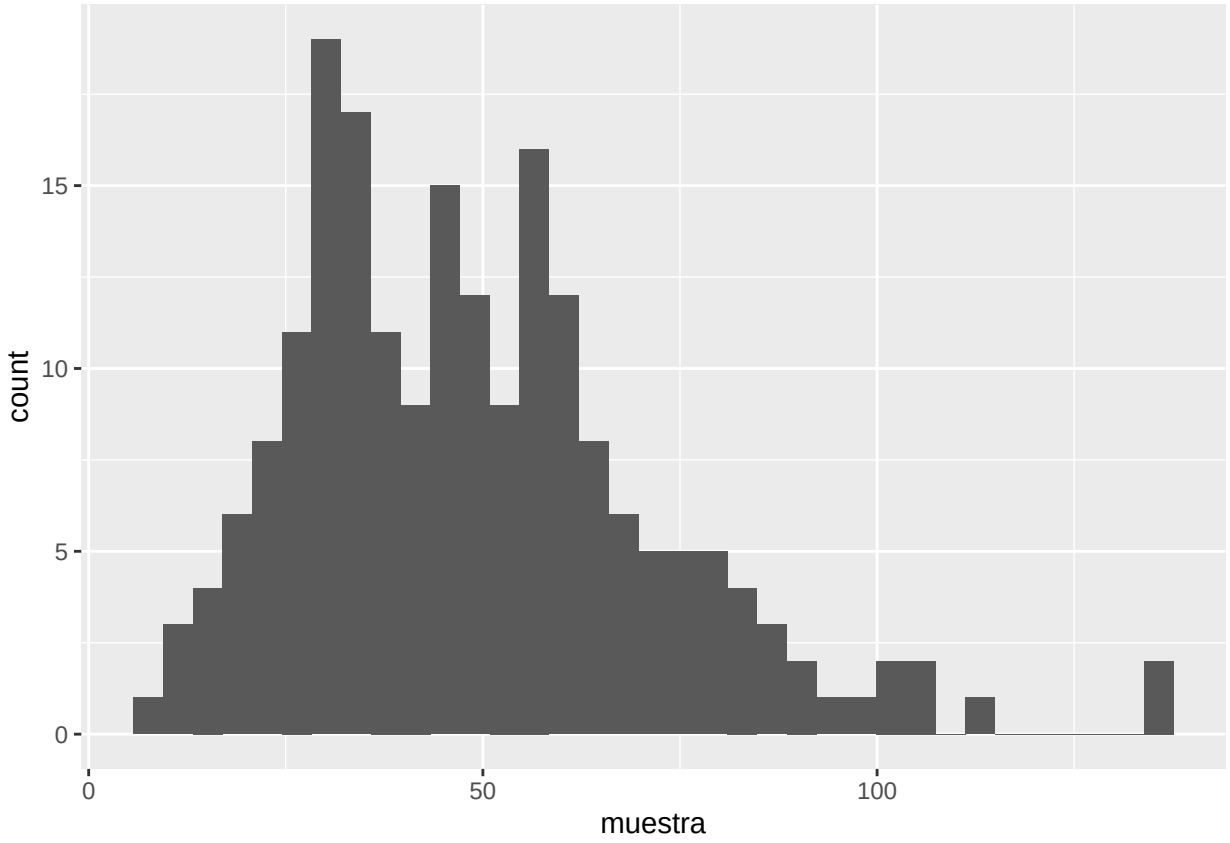
tomamos una muestra:

```
set.seed(232)
n <- 200
muestra <- rgamma(n, 5, 0.1)
```

La distribución de los datos se ve como sigue (haz un histograma de la muestra)

histograma

```
ggplot() + geom_histogram(aes(muestra), bins = 35)
```



- ¿Parece tener distribución normal?

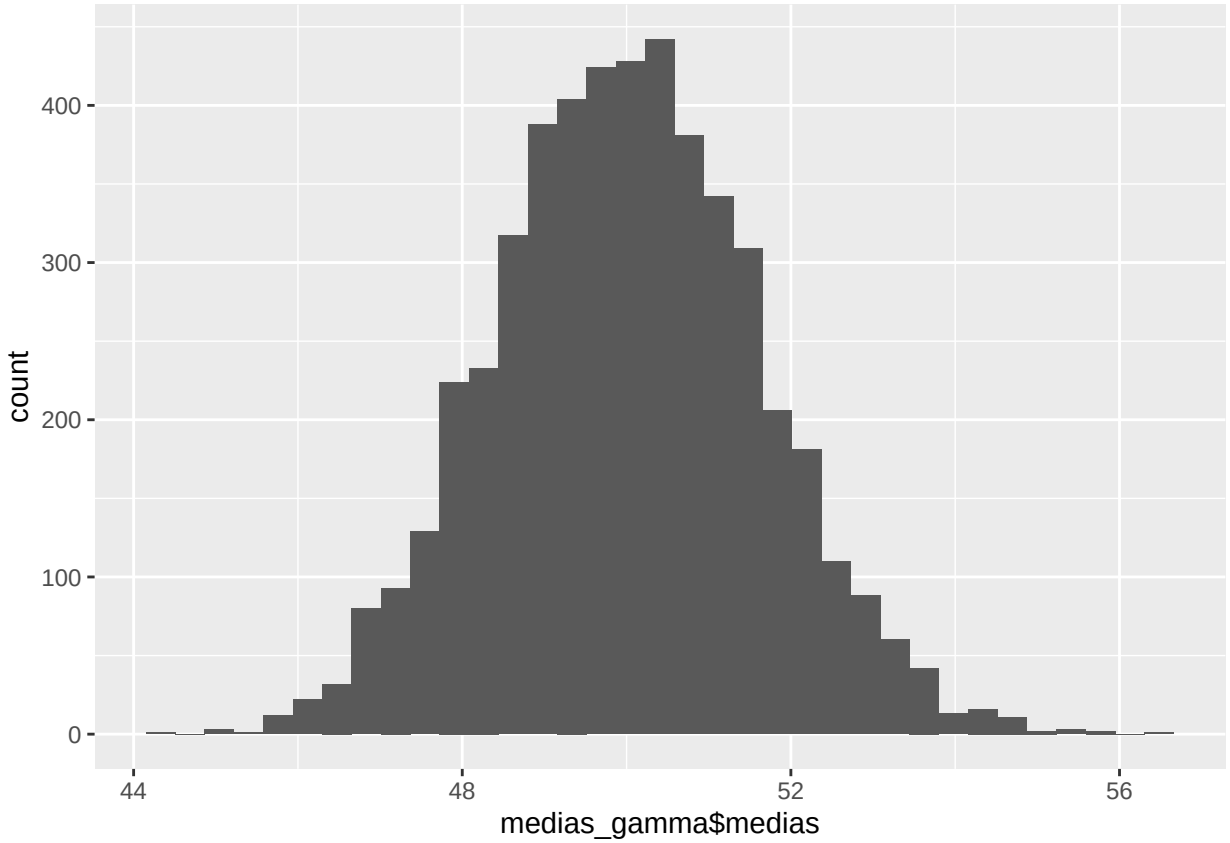
Podemos ver que la distribución no se asemeja a una distribución normal.

Ahora consideramos la distribución de muestreo de la media de esta distribución, con tamaño de muestra fijo n

```
medias <- map_dbl(1:5000,  ~ mean(rgamma(n, 5, 0.1)))
medias_gamma <- tibble(medias = medias)
```

Checa un histograma ¿se ve normal?

```
ggplot() + geom_histogram(aes(medias_gamma$medias), bins = 35)
```



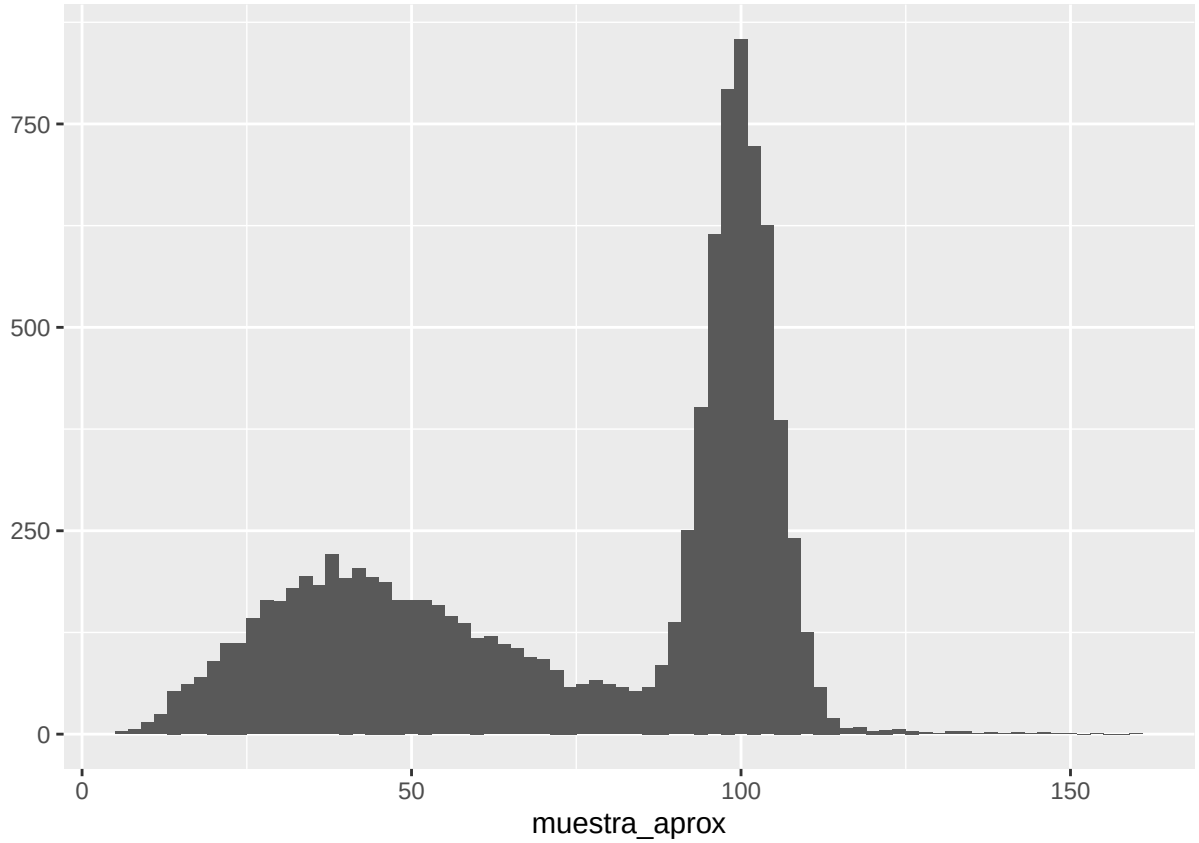
Ejemplo: mezcla de distribuciones

Este ejemplo es más complicado. Imaginemos que nuestro modelo teórico es una mezcla de dos poblaciones, una gamma y una normal

```
muestrear_pob <- function(n){
  u <- runif(n) # número aleatorio
  map_dbl(u, ~ ifelse(.x < 1/2, rgamma(1, 5, 0.1), rnorm(1, 100, 5)))
}
```

El modelo teórico se puede graficar, pero también podemos obetener una aproximación buena haciendo una cantidad grande de simulaciones

```
muestra_aprox <- muestrear_pob(10000)
qplot(muestra_aprox, binwidth = 2)
```

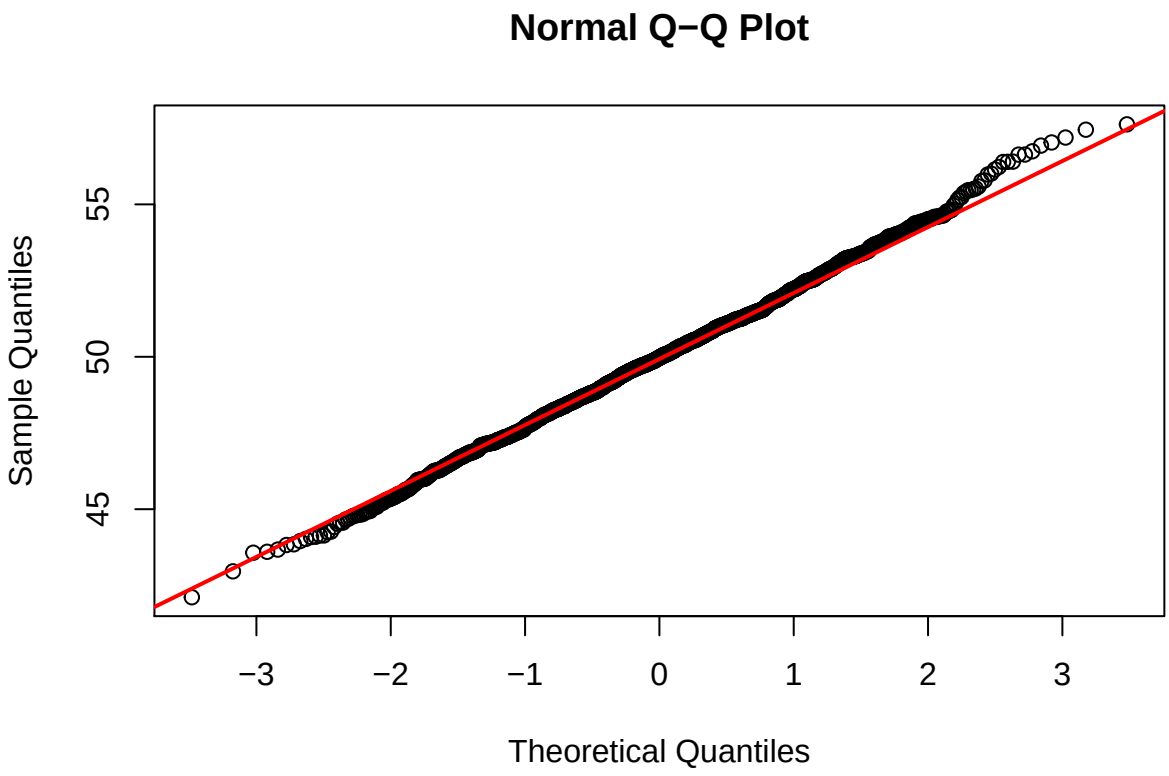


Ahora consideramos estimar la media de esta distribución con un muestra de tamaño 100 ¿Cómo se ve la distribución de muestreo de la media?

```
medias <- map_dbl(1:2000,  ~ mean(rgamma(100, 5, 0.1)))
medias_muestra <- tibble(medias = medias)
```

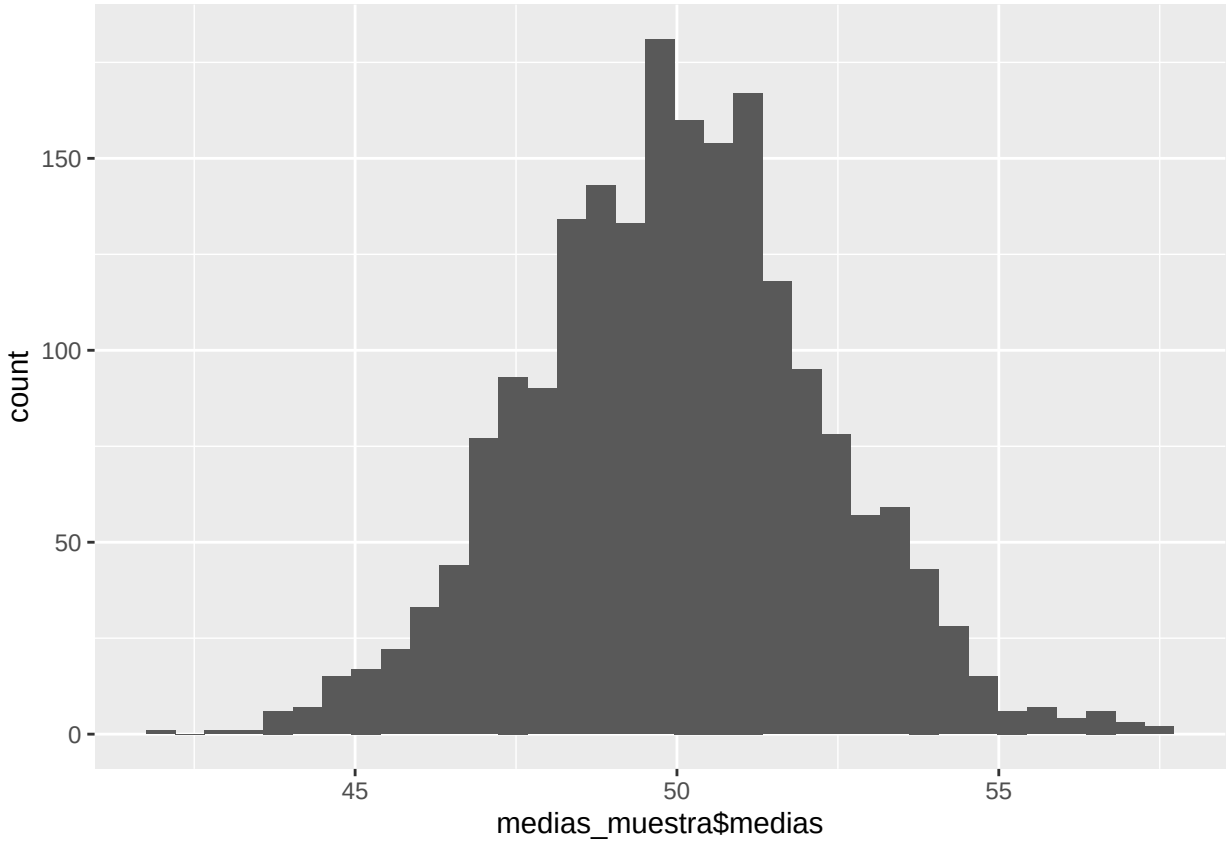
Grafica un histograma y una gráfica cuantil-cuantil normal

```
qqnorm(medias, pch = 1)
qqline(medias,col = "red", lwd = 2)
```




```
g_histograma <- ggplot() + geom_histogram(aes(medias_muestra$medias), bins = 35)

g_histograma
```



Ejemplo discreto

Tomaremos muestra de unos y ceros

```
set.seed(1212)
n_volados <- 200
muestra <- rbinom(n_volados, 1, prob = 0.7)
head(muestra)
```

```
## [1] 1 1 0 1 1 1
```

La media es la proporción de unos en la muestra, o la proporción de "soles":

```
mean(muestra)
```

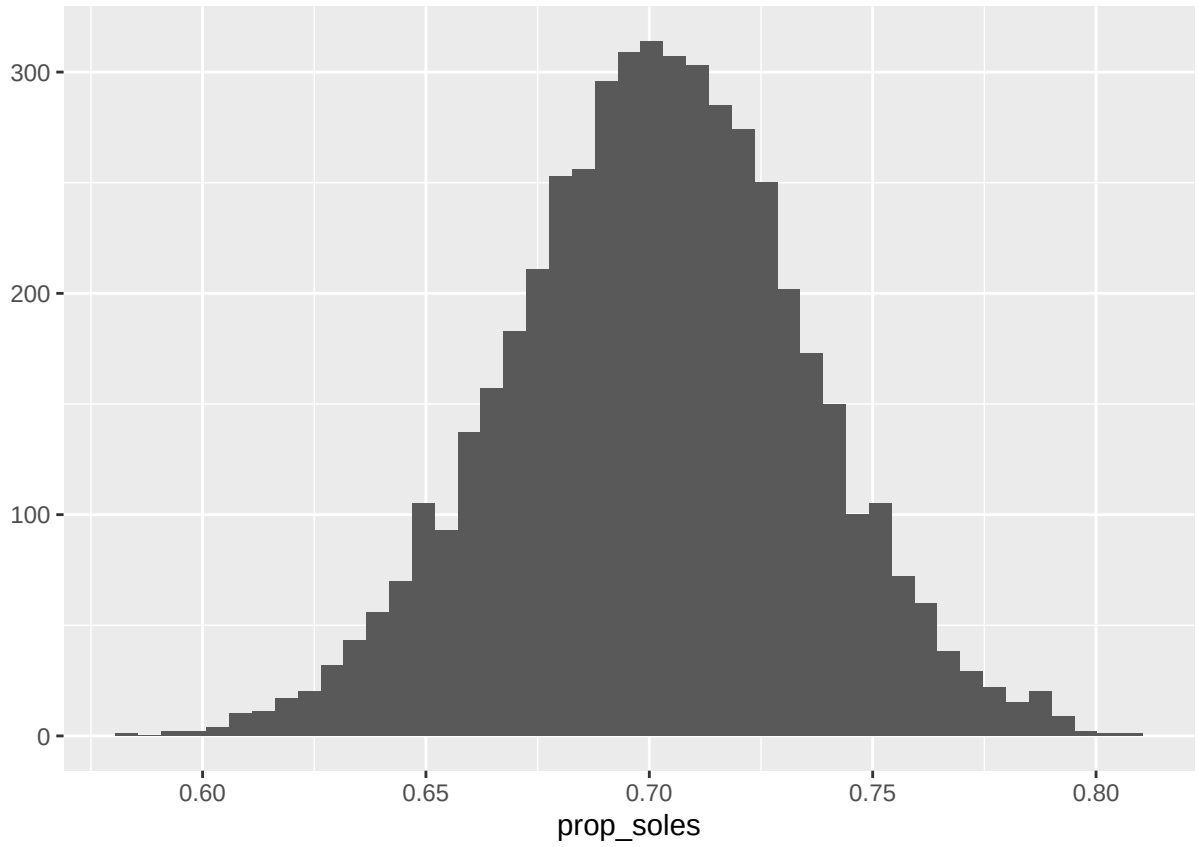
```
## [1] 0.785
```

¿Cuál es la distribución de muestreo para la proporción de soles en la muestra?

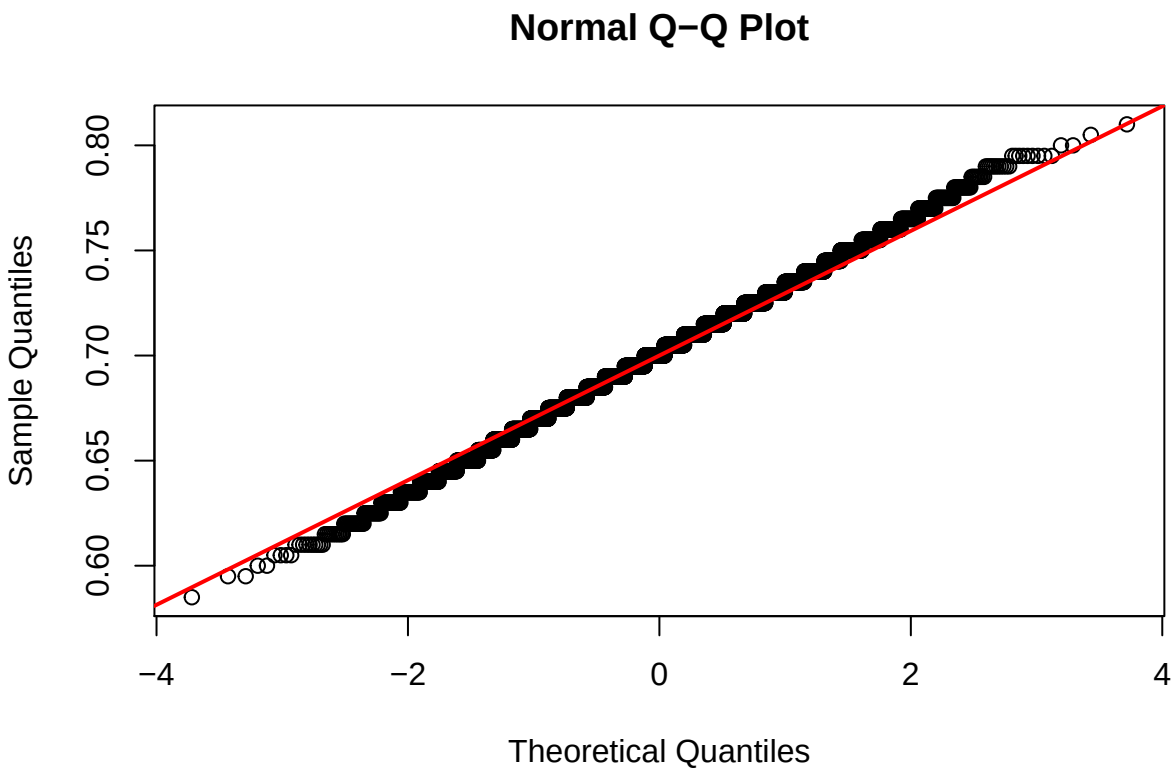
```
prop_soles <- map_dbl(1:5000,  ~ mean(rbinom(n_volados, 1, prob = 0.7)))
prop_soles_tbl <- tibble(prop_soles = prop_soles)
```

Checa un histograma ¿se ve normal? También ve una gráfica qq

```r
qplot(prop_soles, bins = 45)
```




```r
qqnorm(prop_soles)
qqline(prop_soles, col = "red", lwd = 2)
```



Podemos observar que aquí las colas ya no se ven tan pegadas

Error estándar e intervalos bootstrap normales

Ejemplo 1: error estándar de una media

Retomaremos el ejemplo de la prueba ENLACE de la tarea anterior. Para cada tamaño de muestra n = 10, 100, 1000

i) Selecciona una muestra y utilizala para estimar la media de las calificaciones de español 3o de primaria.

```
enlace <- read_csv("enlace_15.csv")
```

```
## Parsed with column specification:
## cols(
##   id = col_double(),
##   cve_ent = col_double(),
##   turno = col_character(),
##   tipo = col_character(),
##   esp_3 = col_double(),
##   esp_6 = col_double(),
##   n_eval_3 = col_double(),
##   n_eval_6 = col_double()
## )
```

```r
set.seed(2020)

muestra_1000 <- sample_n(enlace, 1000, replace = TRUE) %>% select(esp_3)
mean(muestra_1000$esp_3)
```

```
## [1] 549.993
```

ii) Utiliza bootstrap para calcular el error estándar de tu estimador

```r
media_muestra <- map_dbl(1:5000, ~ muestra_1000 %>%
  sample_n(100, replace = TRUE) %>%
  summarise(media_califs = mean(esp_3)) %>%
  pull(media_califs))

ee <- sd(media_muestra)
```

iii) Grafica la distribución bootstrap

Retoma la muestra de tamaño 100, y calcula la correlación entre las calificaciones de español 3o y 6o de primaria. Utiliza bootstrap para calcular el error estandar

```r
set.seed(2020)

muestra_100_3 <- sample_n(enlace, 100, replace = TRUE) %>% select(esp_3)
muestra_100_6 <- sample_n(enlace, 100, replace = TRUE) %>% select(esp_6)
mean(muestra_100_3$esp_3)
```

```
## [1] 556.2
```

```r
mean(muestra_100_6$esp_6)
```

```
## [1] 525.18
```

```r
cor(muestra_100_3, muestra_100_6)
```

```
##             esp_6
## esp_3 0.005382858
```