Práctica 3: Variabilidad muestral e Intervalos de confianza

Introducción

En esta práctica, estudiamos la distribución en el muestreo asociada a estadísticos, estimadores de los parámetros de la población. [1]

Los datos

Trabajamos con un banco de datos reales procedente del registro oficial de Ames, una ciudad del estado americano de Iowa. En concreto, nos centraremos en la superficie de las viviendas vendidas en esa ciudad entre 2006 y 2010. Estos datos constituyen nuestra población de interés por lo que no tendría mucho sentido desarrollar un análisis estadístico de estos datos. En esta práctica, nuestro interés se centra en ilustrar el concepto de estadístico, entender que diferentes muestras del mismo tamaño suelen generar diferentes valores del estadístico y que podemos evaluar esa variabilidad muestral a través de su distribución en el muestreo. Empezamos cargando los datos.

```
#ames<- read.table(file="./ames.txt",header=T, sep = ",", dec = ".")
#download.file("http://www.openintro.org/stat/data/ames.RData", destfile = #"ames.RData")
set.seed(1)
load("./data/ames.RData")
```

Nuestro banco de datos tiene suficientes variables y datos para realizar un estudio que cumpla con los objetivos propuestos en la práctica. Nos centramos únicamente en dos variables: la superficie construida de la vivienda, en pies cuadrados, (*Gr.Liv.Area*) y el precio de venta, en dólares, (`SalePrice`), a las que por sencillez renombraremos como *superficie* y *precio*, respectivamente. Aprovechamos también para trabajar con m^2 y no pies cuadrados como unidad de superficie (1 pie cuadrado= 0.09290304 m^2) que es la unidad de medida utilizada en los datos originales.

```
superficie <- ames$Gr.Liv.Area*0.09290304
precio <- ames$SalePric
```

Ejercicio 1 Describe la población de datos correspondiente a la variable superficie.

```
library(fitdistrplus)
```

```
## Warning: package 'fitdistrplus' was built under R version 4.1.3
```

```
## Loading required package: MASS
```

```
## Warning: package 'MASS' was built under R version 4.1.3
```

```
## Loading required package: survival
```

```
summary(superficie)
```

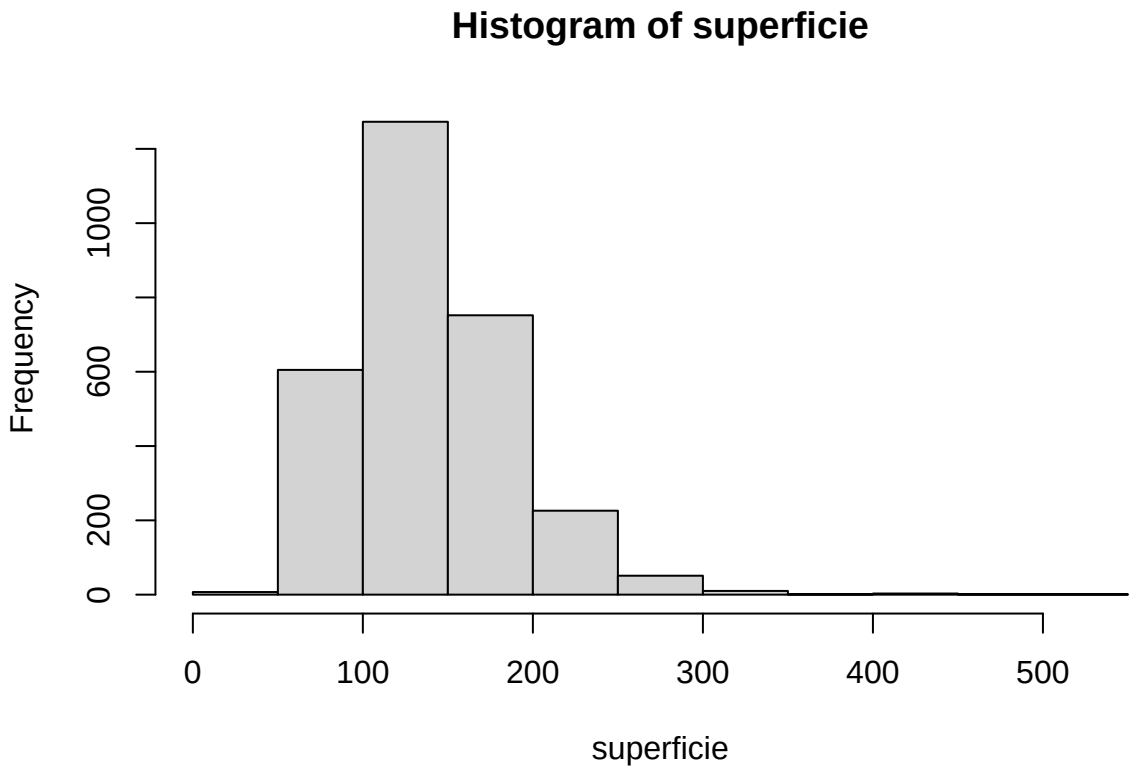
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   31.03  104.61  133.97  139.33  161.91  524.16
```

```
sd(superficie)
```

[1] Esta práctica está basada en el producto OpenIntro that is released under a Creative Commons Attribution-ShareAlike 3.0 Unported (http://creativecommons.org/licenses/by-sa/3.0/). This lab was written for OpenIntro by Andrew Bray and Mine Çetinkaya-Rundel y adaptada por las Profesoras Carmen Armero y Anna Martínez-Gavara.

```
## [1] 46.96331
```

```
hist(superficie)
```




```
dist_norm <- fitdist(superficie, distr = "norm")
dist_gamma <- fitdist(superficie, distr = "gamma")

summary(dist_norm)
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##      estimate Std. Error
## mean 139.3258  0.8674628
## sd    46.9553  0.6133887
## Loglikelihood:  -15435.63   AIC:  30875.27   BIC:  30887.23
## Correlation matrix:
##      mean sd
## mean    1  0
## sd      0  1
```

```
summary(dist_gamma)
```

```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##         estimate  Std. Error
## shape 9.61761000 0.246398572
## rate  0.06902672 0.001814986
## Loglikelihood:  -15201.93   AIC:  30407.86   BIC:  30419.82
```

```
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.9740091
## rate  0.9740091 1.0000000
```

Respuesta: en el histograma vemos cómo la distribución se asemeja a una normal y gracias a `summary` *y a* `sd` *podemos decir que, en tal caso, estaría centrada en 139.33 con desviación típica de 46.96331. Parece estar más ladeada hacia la derecha que hacia la izquierda, pero realmente es consecuencia de que no puede haber viviendas con una superficie negativa, por lo que tenemos nuestro conjunto de datos acotado inferiormente. Debido a esto hemos propuesto tamién una distribución gamma y, al ajustar ambas, obtenemos mayor verosimiitud y menor AIC en el caso de la gamma, por lo que nos decantamos finalmente por una distribución Ga(9.61761, 0.0690267), donde 9.61761 nos indica la forma y 0.0690267 es la inversa de la escala.*

La superficie media de todas las viviendas vendidas en Ames entre 2006 y 2010, ¿es un parámetro o un estadístico? *Respuesta: es un parámetro, ya que es una función de los datos de toda la población, no de una muestra.* ¿Conoces su valor? *Respuesta: 139.3258013.* ¿Hay mucha variabilidad entre las viviendas vendidas en Ames entre 2006 y 2010? *Respuesta: Los valores de superficie se alejan, en media, del valor medio 46.9633124.* ¿Podrías proponer una población estadística adecuada para estos datos? *Respuesta: tras intentar ajusta nuestra población a una distribución gamma y a una normal, vemos cómo obtenemos menor AIC en el caso de la gamma, por lo que nos decantaremos por esta.*

Distribución en el muestreo

Suponemos que estamos interesados en estimar la superficie media de las viviendas de Ames vendidas entre 2006 y 2010 a partir de una muestra de tamaño 50. Podemos seleccionar esta muestra en R a través de la función *sample*(): [2]

```
muestra1 <- sample(superficie, 50)
```

Ejercicio 2. Describe la distribución de la muestra *muestra1* y compárala con la distribución poblacional. Como nuestro objetivo es estimar la superficie media de todas las viviendas de Ames vendidas entre 2006 y 2010, ¿te parece que la media muestral sería un buen estimador suyo? Dependiendo de las 50 viviendas seleccionadas tu estimación puede estar un poco por encima o por debajo del verdadero valor de la superficie media (1499.69 pies cuadrados = 139.3258 m^2) pero la media muestral es un buen estimador de la media poblacional con sólo 50 datos (menos del 3% de la población de los datos).

```
x=seq(0,500,by=0.5)
summary(muestra1)
```

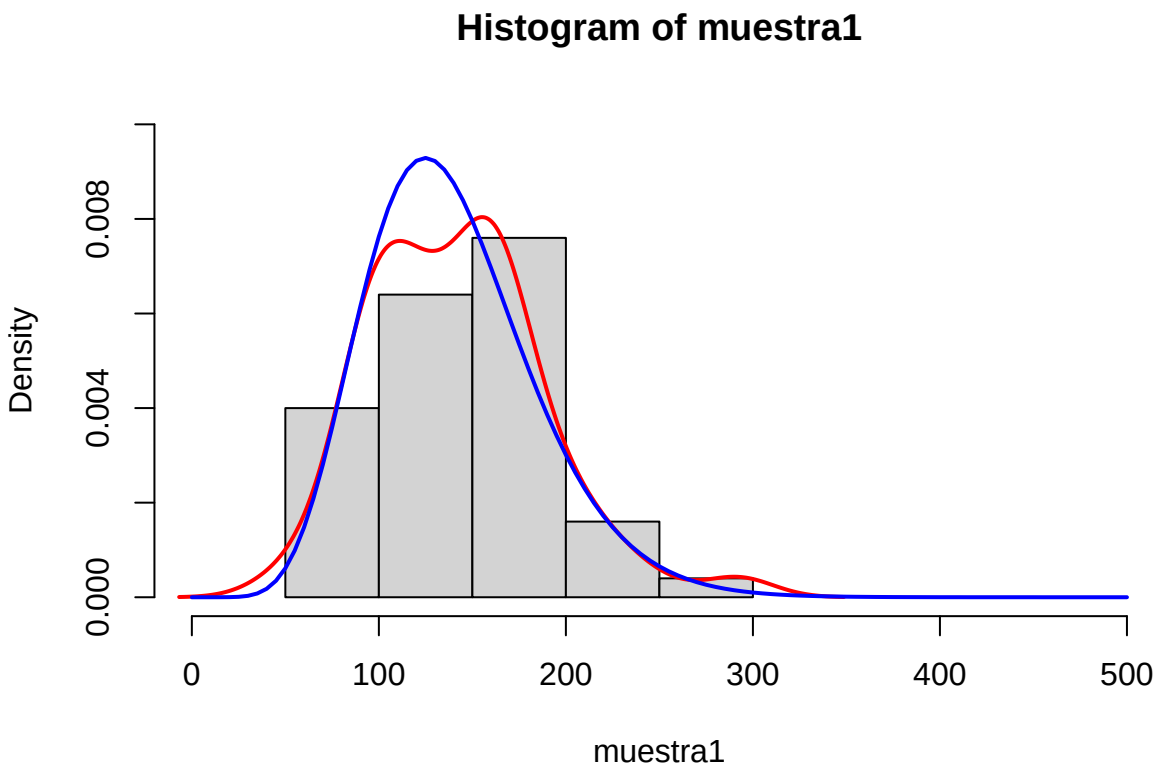
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50.17  102.77  142.14  141.76  165.18  291.72
```

```
sd(muestra1)
```

```
## [1] 46.12712
```

```
hist(muestra1, freq=FALSE, xlim=c(0,500), ylim=c(0,0.01))
lines(density(muestra1), lwd = 2, col = 'red')
curve(dgamma(x,shape=dist_gamma$estimate[[1]],rate=dist_gamma$estimate[[2]]),
xlim=c(0,500),col="blue",lwd=2,add=TRUE)
```

[2] Para reproducir los resultados, se ha empleado 'set.seed(1)' en todos los chuncks.



Respuesta: en azul podemos ver la gráfica de la función de densidad de una distribución $Ga(9.61761, 0.0690267)$. *En rojo la correspondiente a nuestra muestra. Vemos cómo parece ajustarse a ella. Evidentemente con tan solo 50 datos no podemos esperar un ajuste perfecto ni mucho menos, pero sí que parece describir la curva.*

La media de nuestra muestra es de 141.7644649 y la desviación típica de 46.1271165. En ambos casos los estadísticos están un poco por debajo del valor real de los parámetros, pero se aproximan bastante.

Ejercicio 3. Obtén ahora una segunda muestra del mismo tamaño y llámala *muestra2*. ¿Cómo es la distribución de esta segunda muestra en comparación con la anterior? ¿Y con respecto a la distribución de datos poblacional? ¿Y la media muestral en relación a la media de la primera muestra y a la media poblacional?

```
muestra2 <- sample(superficie, 50)
summary(muestra2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   71.35  108.37  139.17  143.01  164.55  341.14
```

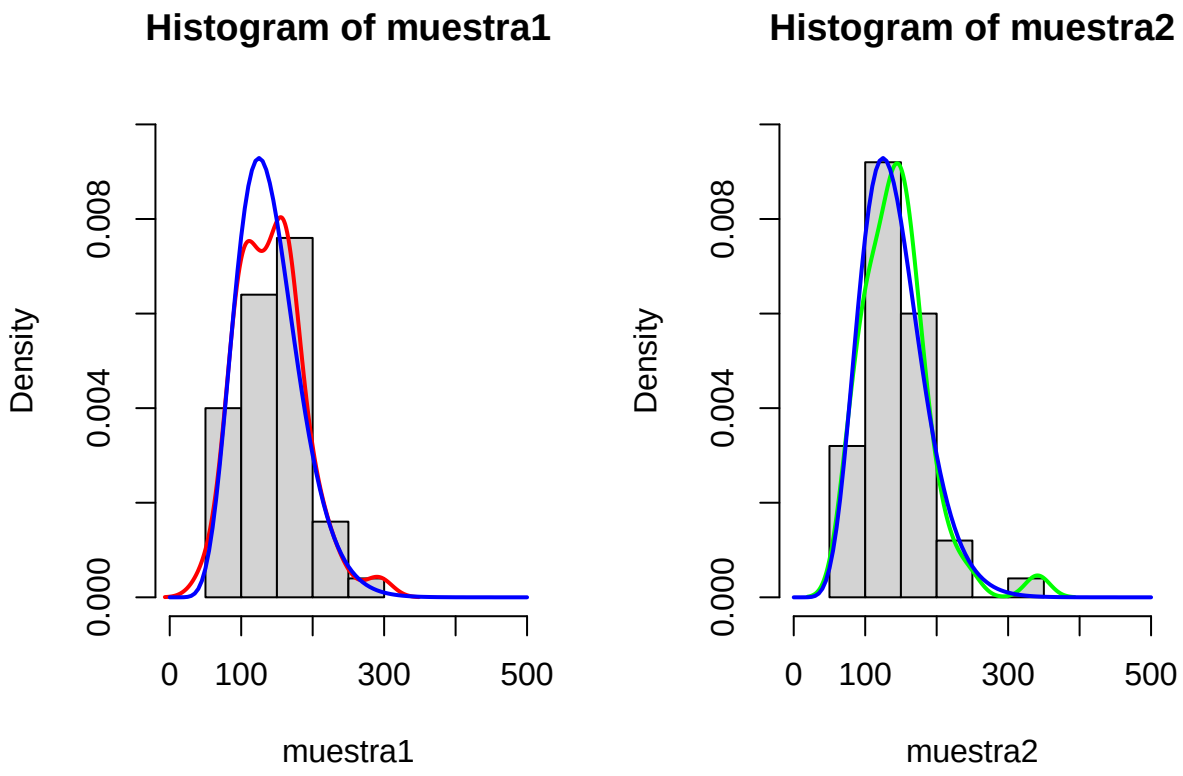
```
sd(muestra2)
```

```
## [1] 48.16077
```

```
par(mfrow = c(1, 2))
hist(muestra1, freq=FALSE, xlim=c(0,500), ylim=c(0, 0.01))
lines(density(muestra1), lwd = 2, col = 'red')
curve(dgamma(x,shape=dist_gamma$estimate[[1]],rate=dist_gamma$estimate[[2]]),
xlim=c(0,500),col="blue",lwd=2,add=TRUE)

hist(muestra2, freq=FALSE,xlim=c(0,500), ylim=c(0, 0.01))
```

```
lines(density(muestra2), lwd = 2, col = 'green')
curve(dgamma(x,shape=dist_gamma$estimate[[1]],rate=dist_gamma$estimate[[2]]),
xlim=c(0,500),col="blue",lwd=2,add=TRUE)
```



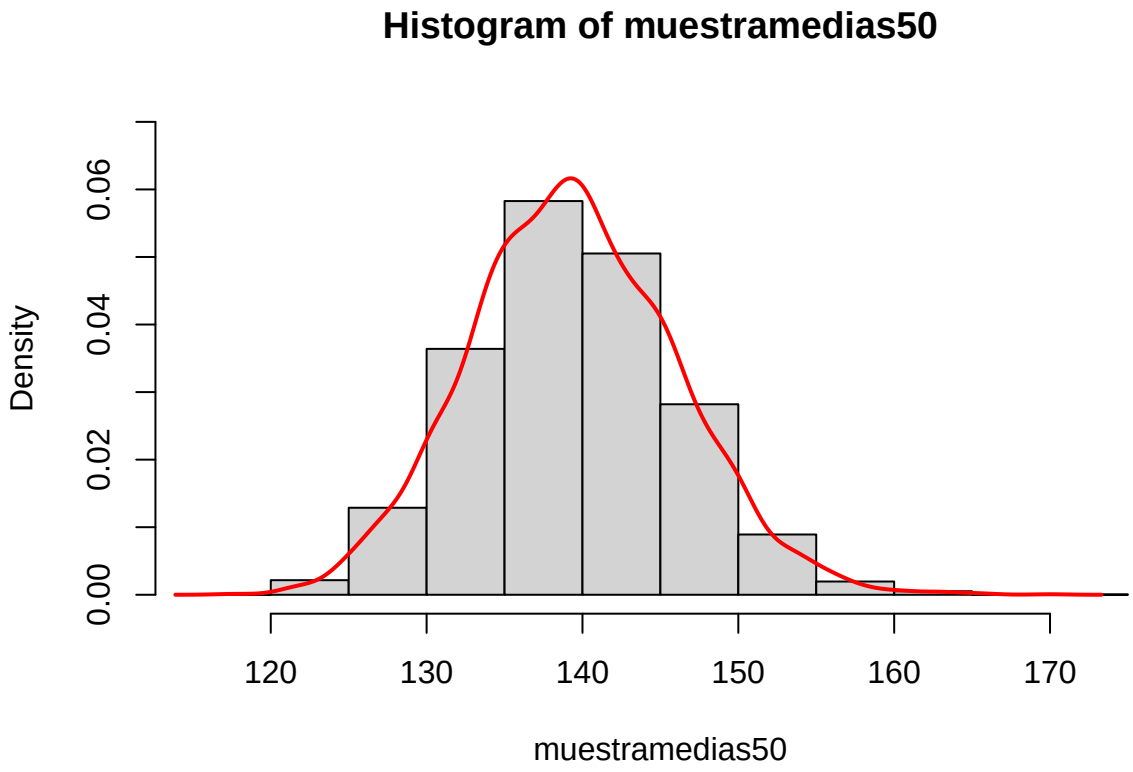

```
par(mfrow = c(1, 1))
```

Respuesta: en este caso, la media que obtenemos par ala segunda muestra es 143.0112237, que como podemos ver queda un poco por encima de la media poblacional, al contrario que nos pasaba en ***muestra1****. La desviación típica es 48.1607715, muy por debajo de la desviación típica de nuestra población y también por debajo de la desviación típica de la muestra uno. Es decir, en esta nueva muestra los datos están menos dispersos, más agrupados alrededor de la media.*

Sabemos que si tomáramos más muestras del mismo tamaño de la población obtendríamos medias muestrales diferentes y, por lo tanto, sería conveniente conocer si la variabilidad entre las diferentes medias muestrales es muy grande o muy pequeña y si están o no alrededor de la media poblacional. La distribución de probabilidad de las medias muestrales que conocemos como distribución en el muestreo de la media muestral, nos podría informar sobre dicha variabilidad. Como en esta práctica tenemos una situación privilegiada porque conocemos la población de los datos, podemos simular dicha distribución generando muchas muestras de tamaño 50. Vamos a hacerlo 5000 veces y calculamos la media muestral de cada muestra generada. Tendremos una muestra aleatoria de tamaño 5000 de la distribución de la media muestral.

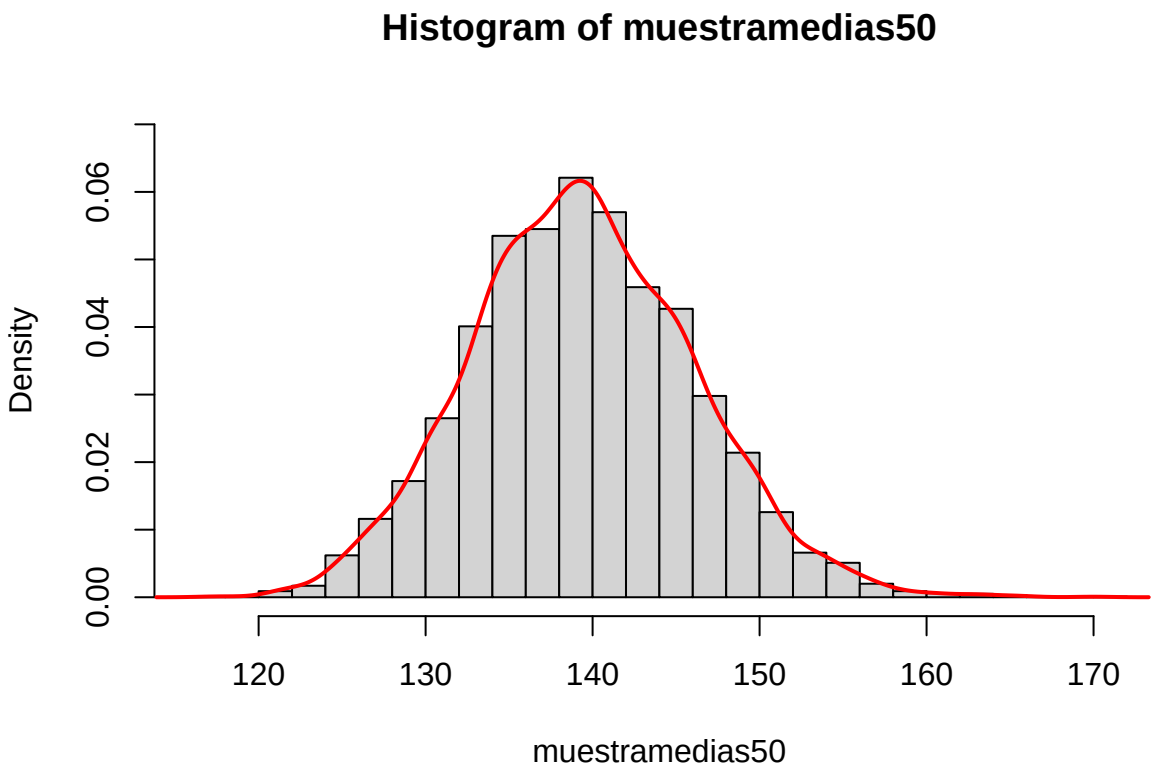
```
muestramedias50 <- rep(NA, 5000)
for(i in 1:5000){
muestra50 <- sample(superficie, 50)
  muestramedias50[i] <- mean(muestra50)
 }
hist(muestramedias50, freq=FALSE, ylim=c(0, 0.07))
```

```
lines(density(muestramedias50), col="red", lwd=2)
```



Recuerda que si quieres ajustar la longitud de los intervalos del histograma para visualizar la información con más detalle puedes modificar el argumento breaks de la función *hist*.

```
hist(muestramedias50, breaks = 25, freq=FALSE, ylim=c(0, 0.07))
lines(density(muestramedias50), lwd=2, col="red")
```



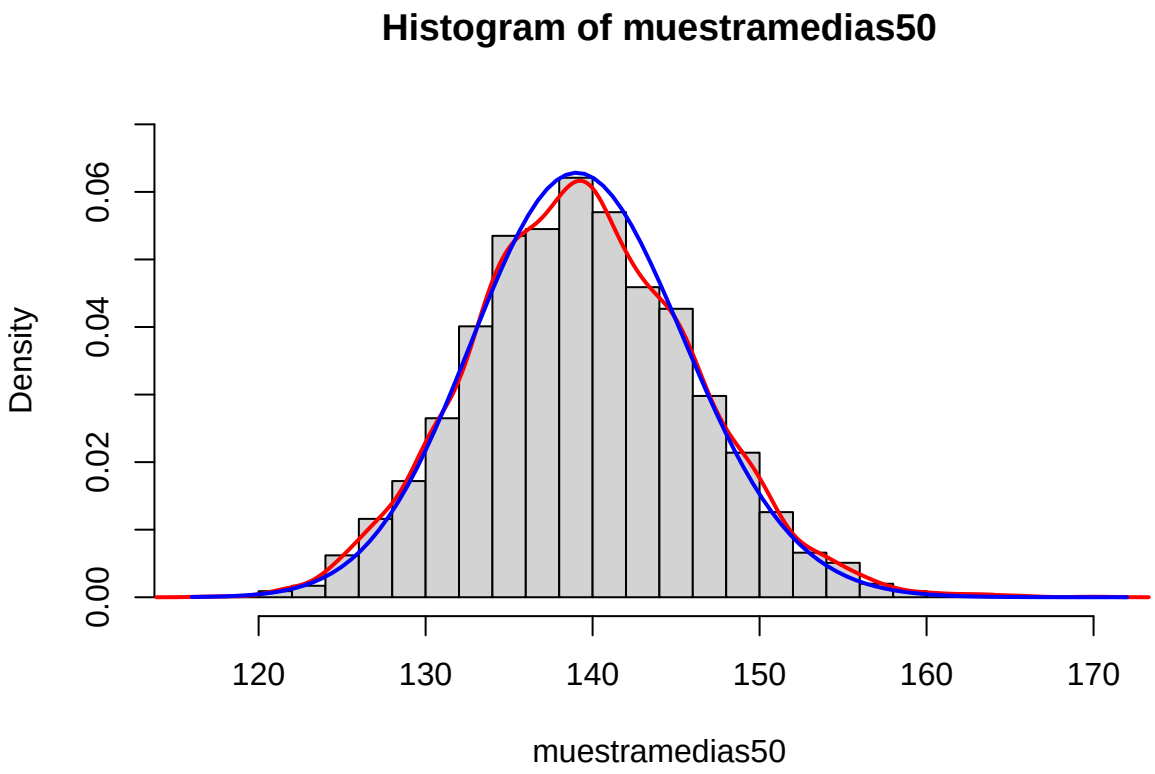
y también añadir color, escribir información en los ejes, etc.

Ejercicio 4. ¿Cuántos elementos hay en `muestramedias50`? Describe la distribución en el muestreo y calcula su media. ¿Cómo esperarías que cambiara la distribución si en vez de 5000 muestras hubiéramos generado 50000?

Respuesta: hay 5000 elementos en `muestramedias50`*. Como podemos observar, la distribución que sigue parece ser una gamma de forma 438.6401995 y rate 0.3178001.*[3]*. La media de nuestras gammas debería de seguir una distribución* $Ga(N \cdot k, N \cdot \theta)$*, donde* N *es el tamaño que hemos empleado para calcular las medias (50 en nuestro caso') y* k *y* θ *son los parámetros forma y rate, respectivamente. Por lo tanto,* $Ga(480.8805, 3.4513359)$*. Volvamos a visualizar el histograma anterior pero esta vez dibujaremos en color azul la curva de la función de densidad de la gamma que en teoría deberían de seguir las medias.*

```
hist(muestramedias50, breaks = 25, freq=FALSE, ylim=c(0, 0.07))
lines(density(muestramedias50), lwd=2, col="red")
curve(dgamma(x,shape=50*dist_gamma$estimate[[1]],rate=50*dist_gamma$estimate[[2]]),
col="blue",lwd=2,add=TRUE)
```

[3] La estimación de los parámetros se basa en el método de los momentos, de donde sabemos que la forma es $\frac{\overline{X}^2}{V}$ y rate es $\frac{V}{\overline{X}}$

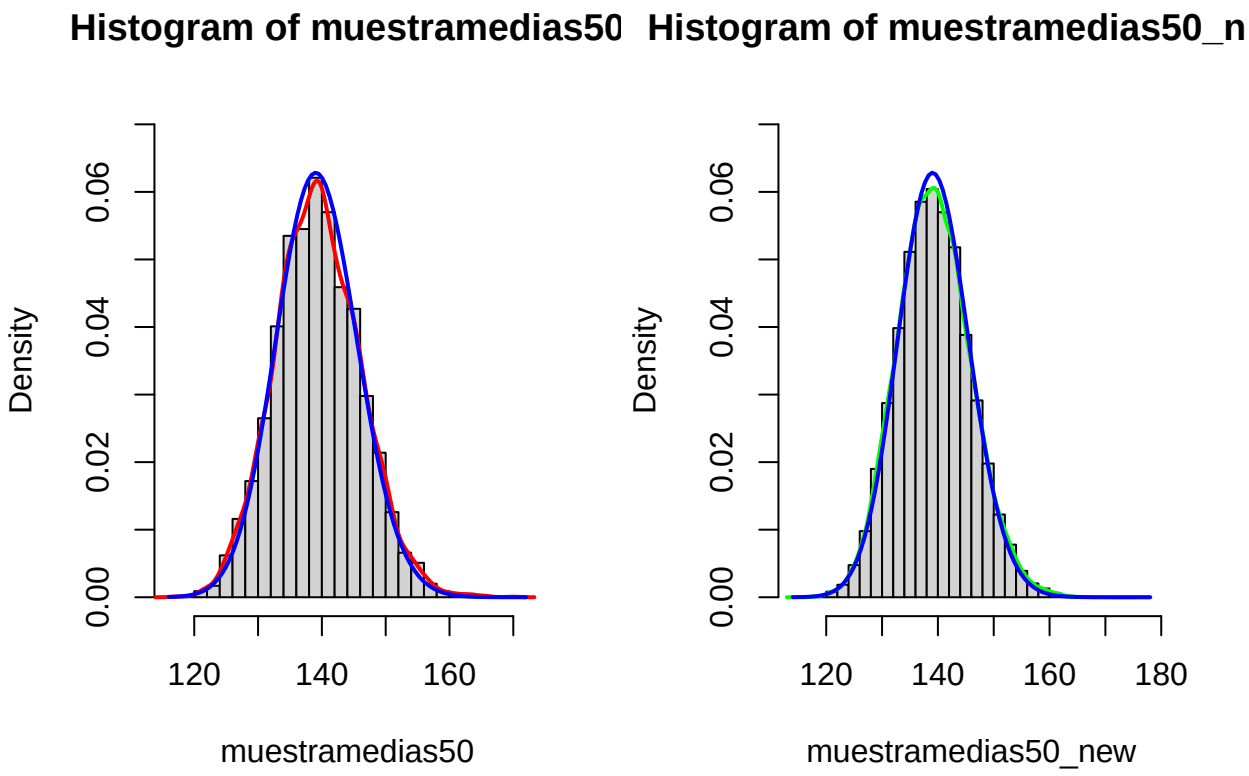


Como vemos, se ajustan muy bien ambas curvas. Si en vez de 5000 hubiéramos generados 50 000, el ajuste debería de ser casi perfecto.

```
muestramedias50_new <- rep(NA, 50000)
for(i in 1:50000){
muestra50 <- sample(superficie, 50)
muestramedias50_new[i] <- mean(muestra50)
 }

par(mfrow = c(1, 2))
hist(muestramedias50, breaks = 25, freq=FALSE, ylim=c(0, 0.07))
lines(density(muestramedias50), lwd=2, col="red")
curve(dgamma(x,shape=50*dist_gamma$estimate[[1]],rate=50*dist_gamma$estimate[[2]]),
col="blue",lwd=2,add=TRUE)

hist(muestramedias50_new,  breaks = 25,freq=FALSE, ylim=c(0, 0.07))
lines(density(muestramedias50_new), lwd = 2, col = 'green')
curve(dgamma(x,shape=50*dist_gamma$estimate[[1]],rate=50*dist_gamma$estimate[[2]]),
col="blue",lwd=2,add=TRUE)
```




```
par(mfrow = c(1, 1))
```

Intervalos de Confianza

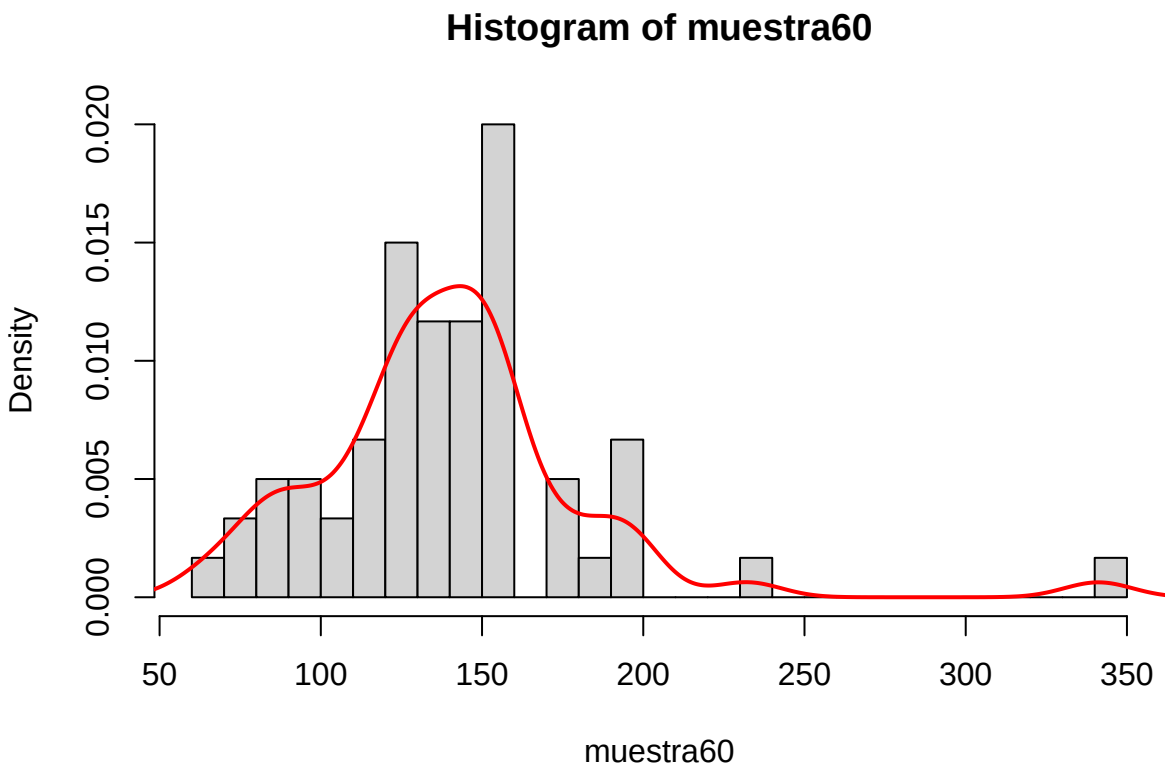
Si tenemos acceso a los datos sobre una población completa, por ejemplo el tamaño de cada casa en Ames, Iowa, es sencillo responder a preguntas como, ¿Qué tamaño es una casa típica en Ames? y ¿Cuánta variabilidad hay en los tamaños de las casas? En cambio, si solo tenemos acceso a una muestra de la población, como suele ser el caso, la tarea se vuelve más complicada. ¿Cuál es su mejor estimación del tamaño típico si solo conoce el tamaño de varias docenas de casas? Este tipo de situación requiere que se use la muestra para hacer inferencia sobre el aspecto de la población.

En esta sección usaremos una muestra aleatoria simple de tamaño 60 de la población ,en particular sólo nos centraremos en el tamaño de la casa, representada por la variable *Gr.Liv.Area* (análogamente a lo que hemos hecho anteriormente usaremos m^2).

```
superficie <- ames$Gr.Liv.Area*0.09290304
muestra60 <- sample(superficie, 60)
```

Ejercicio 5. Describe la distribución de la muestra. ¿Cuál dirías que es el tamaño "típico" de las casa de Ames? ¿Qué interpretas por "típico"? ¿Esperarías que la distribución de otro compañero sea idéntica a la tuya? ¿Esperarías que sea similar? ¿Por qué o por qué no?

```
hist(muestra60, freq=FALSE, breaks=25)
lines(density(muestra60), col="red", lwd=2)
```




```
summary(muestra60)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   62.43  119.43  138.24  140.02  155.01  341.14
```

```
sd(muestra60)
```

```
## [1] 42.63377
```

Respuesta: vemos que la muestra no parece seguir una distribución normal a simple vista. Su media es 140.0172684 y su desviación típica es 42.6337739. En aplicación del teorema central del límite, asumiremos la normalidad de la muestra. El tamaño típico de una casa en Ames es el de la media. Por típico entendemos el valor esperado, en este caso. La distribución de otro compañero no debería de ser igual, ya que cada uno ha obtenido una muestra aleatoria y las supondremos distintas, por lo tanto los estimadores de la media poblacional y la desviación típica poblacional serán diferentes. Sin embargo, no deberían de diferir en exceso de las nuestras.

Una de las formas más comunes de describir el valor típico o central de una distribución es usar la media. En este caso podemos calcular la media de la muestra usando,

```
sample_mean <- mean(muestra60)
```

Si volvemos a la pregunta inicial, en base a nuestra muestra ¿qué podemos inferir sobre la población? La mejor estimación de la superficie promedio habitable de las casa en Ames sería la media de la muestra. Esta es una buena *estimación puntual*. Proporcionar una estimación puntual es dar una información incompleta si no se acompaña de alguna medida del error que podemos estar cometiendo. Esto se puede capturar utilizando los *intervalos de confianza*.

Podemos calcular un intervalo de confianza del 95% para una media muestral sumando y restando $t_{0.975}$

errores estándar a la estimación puntual.

```
a<-qt(0.975,60-1,lower.tail = TRUE)
se <- sd(muestra60) / sqrt (60)
lower <- sample_mean - a * se
upper <- sample_mean + a * se
c(lower, upper)
```

```
## [1] 129.0038 151.0307
```

Para que el intervalo de confianza sea válido, la media muestral debe estar normalmente distribuida y tener un error estándar de $s/\sqrt{(n)}$. Para ello es necesario que la población tenga una distribución Normal con media μ y desviación típica σ. O bien, por el **Teorema central del límite**:

Dada una población con media μ y desviación típica σ. Entonces, si el tamaño de la muestra n es suficientemente grande, la distribución de $\bar{X}$ es normal con media μ y desviación típica $\sigma/\sqrt{n}$.

Si μ es desconocida, σ también lo será pero sabemos que la distribución de

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1},$$

es una t-Student con $n-1$ grados de libertad. El intervalo de confianza para μ es:

$$IC_{100(1-\alpha)\%}(\mu) = \left[\bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

donde $t_{1-\alpha/2}$ es el cuantil $1-\alpha/2$ de una distribución t-Student con $n-1$ grados de libertad.

Ejercicio 6 ¿Qué significa "95% de confianza"?

Respuesta: no significa que con una probabilidad del 95% nuestra media se encuentra en el intervalo descrito. Nuestra media es un parámetro poblacional fijo, y una vez contruido el intervalo para la media, al estar haciéndolo con estadísticos muestrales que NO son fijos, porque dependen de la muestra que tomemos, el 95% de las veces estará dentro del intervalo y habrá un 5% que no. Realmente es un pequeño matiz debido a la definición determinista de nuestra media. La media es la que es, ya esta determinada, y una vez dado un intervalo estará o no dentro de él pero no con una probabilidad porque realmente ese suceso ya está determinado. De ahí surge el concepto de confianza.

En este caso, nos damos el lujo de conocer la media real de la población, ya que tenemos datos sobre toda la población. Este valor se puede calcular con el siguiente comando:

```
mean(superficie)
```

Ejercicio 7 ¿Tu intervalo de confianza captura el tamaño promedio real de las casas en Ames?

Respuesta: en este caso, la media poblacional (139.3258013) está dentro del intervalo de confianza al 95% ([129.0037967, 151.03074]).

Con R, vamos a recrear muchas muestras para obtener más información sobre cómo las medias de las muestras y los intervalos de confianza varían de una muestra a otra. Para ello:

(1) Obtén una muestra aleatoria.
(2) Calcula la media y la desviación estándar de la muestra.
(3) Usa los estadísticos para calcular un intervalo de confianza.
(4) Repite los pasos (1) - (3) 50 veces.

Pero antes de hacer todo esto, primero debemos crear vectores vacíos donde podamos guardar las medias y las desviaciones estándar que se calcularán a partir de cada muestra. Y mientras lo hacemos, también almacenemos el tamaño de muestra deseado como n.

```
samp_mean <- rep (NA, 50)
samp_sd <- rep (NA, 50)
n <- 60
```

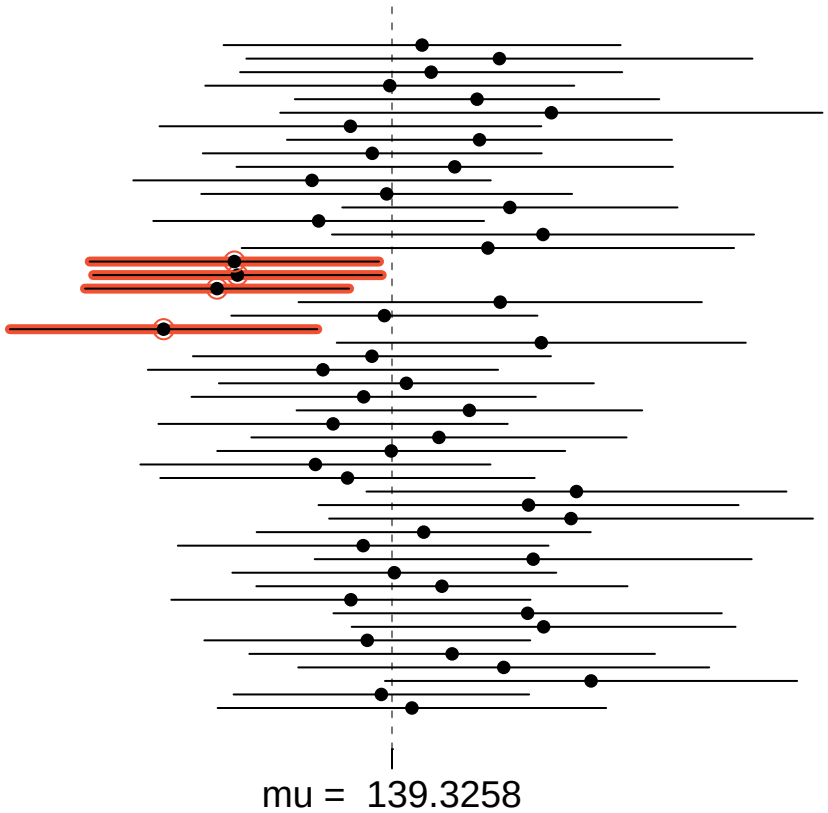
Ahora estamos listos para el ciclo donde calculamos las medias y las desviaciones estándar de 50 muestras aleatorias, y calculamos el intervalo de confianza para cada muestra.

```
for(i in 1:50) {
    # obtener una muestra de tamaño n = 60 de la población
    samp <- sample(superficie, n)
    # guardar la media de la muestra en el elemento i-ésimo de samp_mean
    samp_mean[i] <- mean(samp)
    # guardar muestra sd en el i-ésimo elemento de samp_sd
    samp_sd[i] <- sd(samp)
    a<-qt(0.975,60-1,lower.tail = TRUE)
    lower[i] <- samp_mean[i] - a*samp_sd[i]/sqrt (n)
    upper[i] <- samp_mean[i] + a*samp_sd[i]/sqrt (n)
}
```

TAREA

Dibuja todos los intervalos ¿Qué proporción de tus intervalos de confianza incluyen la media real de la población? ¿Es esta proporción exactamente igual al nivel de confianza? Si no, explica por qué. Para dibujar los intervalos de confianza puedes usar el comando:

```
plot_ci(lower,upper, mean(superficie))
```



```
1-sum(lower>mean(superficie) | upper<mean(superficie))/(50)
```

```
## [1] 0.92
```

Respuesta: como vemos aquí, un 0.92 de las veces nuestro intervalo de confianza contiene a la media. No es exactamente igual al nivel de confianza (95%) debido a que 50 intervalos de confianza son pocos, pero vemos cómo se aproxima.

Elige otro nivel de confianza distinto al 95% ¿Cuál es el valor crítico apropiado? Calcula 50 intervalos de confianza en el nivel de confianza que eligiste en la pregunta anterior. No necesitas obtener nuevas muestras, simplemente calcula los nuevos intervalos en función de las medias de la muestra y las desviaciones estándar que ya has calculado. Dibuja los intervalos de confianza y calcula la proporción de intervalos que incluyen la media real de la población. ¿Cómo se compara este porcentaje con el nivel de confianza seleccionado para los intervalos?

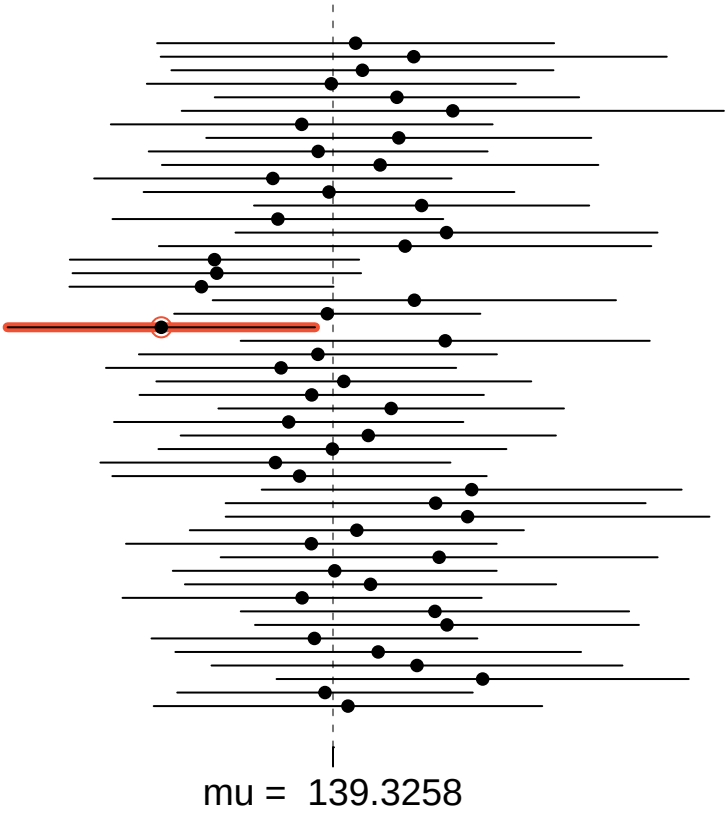
Respuesta: fijaremos la confianza al 99%. Calculamos a continuación el valor crítico:

```
alpha<-0.01
a<-qt(1-alpha/2,60-1,lower.tail = TRUE)
a
```

```
## [1] 2.661759
```

Repetimos ahora el proceso anterior:

```
lower<- samp_mean- a*samp_sd/sqrt (n)
upper<- samp_mean+ a*samp_sd/sqrt (n)
plot_ci(lower,upper, mean(superficie))
```

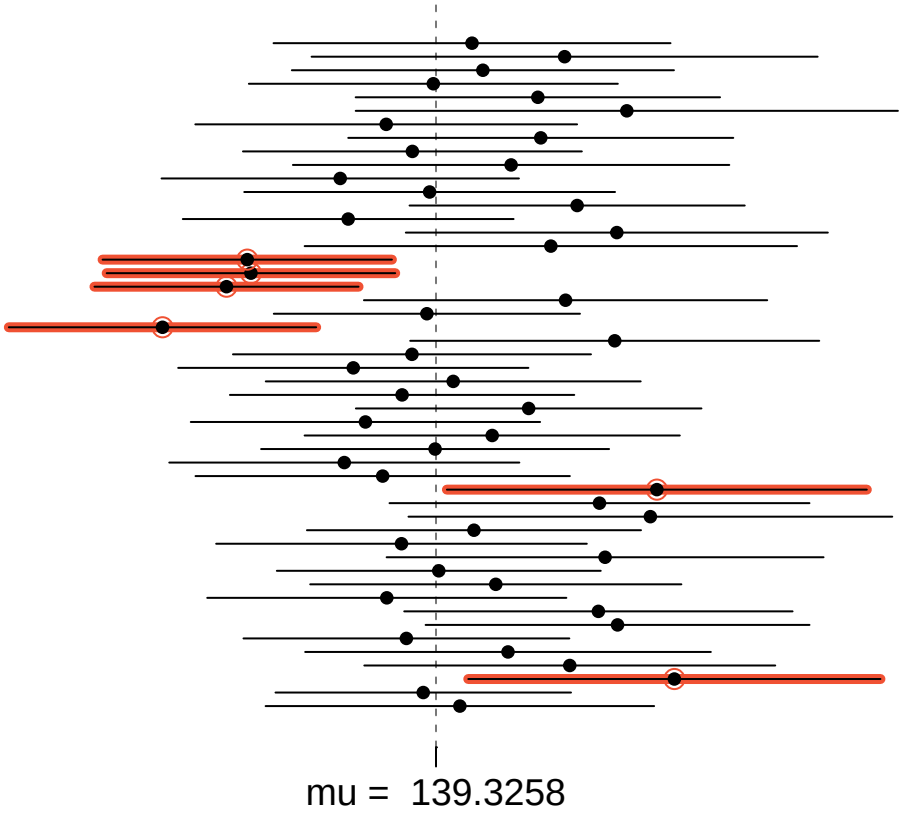


```
1-sum(lower>mean(superficie) | upper<mean(superficie))/(50)
```

```
## [1] 0.98
```

Vemos como ahora el 0.98 de los intervalos contienen a la media. Reharemos el proceso con una confianza del 90%.

```
alpha<-0.1
# Calculamos el valor crítico
a<-qt(1-alpha/2,60-1,lower.tail = TRUE)
lower<- samp_mean- a*samp_sd/sqrt (n)
upper<- samp_mean+ a*samp_sd/sqrt (n)
plot_ci(lower,upper, mean(superficie))
```




```
1-sum(lower>mean(superficie) | upper<mean(superficie))/(50)
```

```
## [1] 0.88
```

En este caso, el 88% de los intervalos contienen a la media.