Porqué la cuasivarianza

Equipo docente

27 de octubre de 2016

El problema

Ya hemos visto que para calcular un intervalo de confianza para la media es necesario tener una idea de la dispersión de la población, en concreto, si la población de estudio se distribuye de acuerdo con una normal, es preciso disponer de la varianza.

Pero es “raro” conocer la varianza sin conocer la media (la media es un ingrediente de la fórmula con la que se calcula la varianza). Así es que lo habitual es no conocer la varianza y, en ese caso, tendremos que *estimar* su valor, aproximarlo.

Así como para estimar la media poblacional usamos la media muestral, lo primero que puede venirnos a la cabeza es usar la varianza de la muestra para estimar la varianza de la población. Y lo que dice la teoría al respecto es que no es tan buena idea como parece. La demostración rigurosa (matemática) es bastante aparatosa y no aporta nada alos objetivos de este curso. Sin embargo, hacer simulaciones sí permite cubrir varios objetivos (además del de convencernos de que la cuasivarianzaes mejor que la varianza).

Pensemos en las características *deseables* al estimar un parámetro

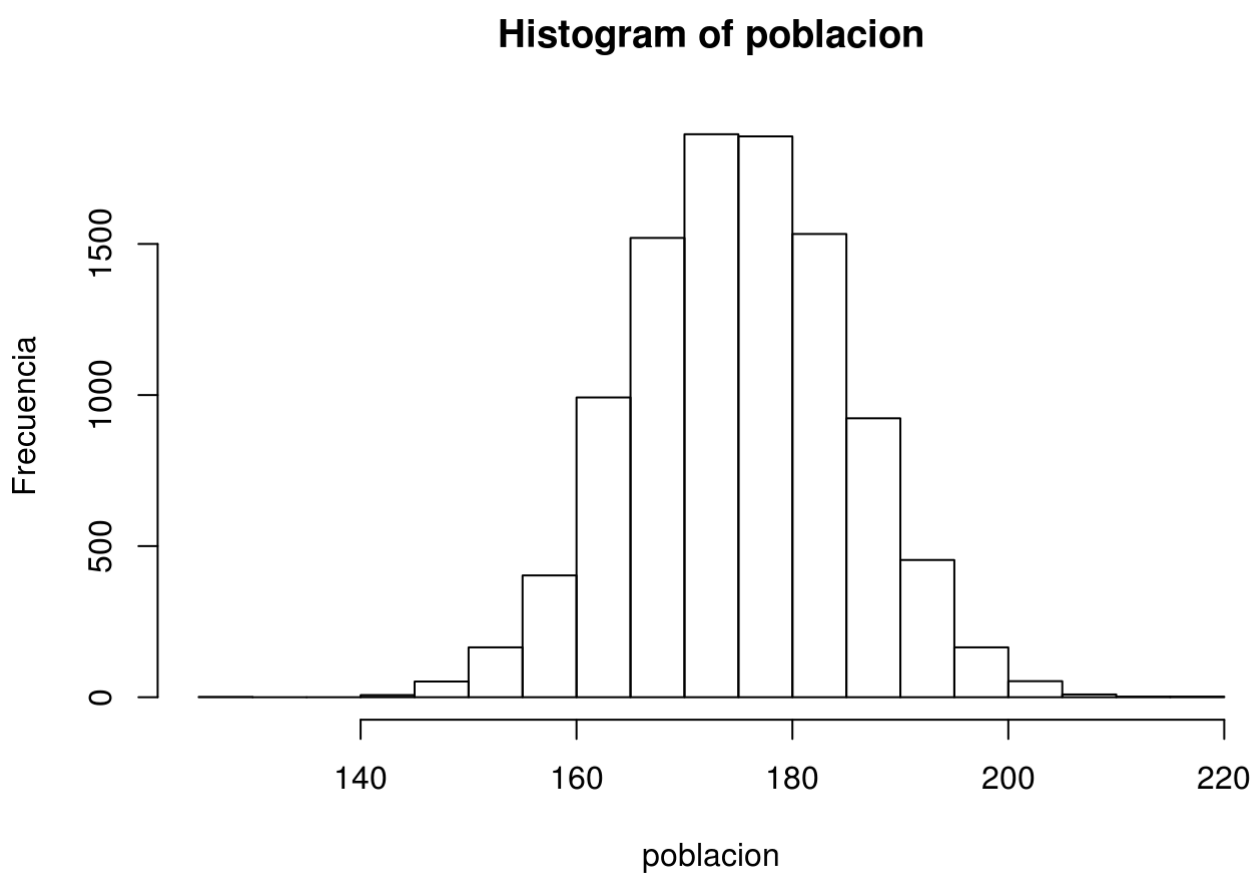
- Si se estima muchas veces ese valor estemos tan cerca como sea posible su verdadero valor como sea posible
- Ya que aproximarlo (casi) seguro que no obtenemos el valor exacto, lo mejor es que sea tan probable que la estimación exceda/no exceda el verdadero valor del parámetro.

Es lo anterior hay una pequeña trampa, porque en general no conocemos el verdadero valor del parámetro (¡por eso necesitamos estimarlo!), pero en una simulación sí lo podemos conocer, porque somosnosotros los qu definimos toda la población y, por tanto, calcularlo.

Definir la población

Usaremos esta población; puedes cambiar de población comentando/descomentando lineas

```
tamanno.pob = 10000
poblacion = sample(1:100, size = tamanno.pob, replace = TRUE)
 poblacion = rnorm(tamanno.pob, mean = 175, sd = 10)
# poblacion = rexp(tamanno.pob, rate = 1)
# poblacion = rbeta(tamanno.pob, shape1 = .6, shape2 = .7)
hist(poblacion, ylab = "Frecuencia")
```



Al conocer toda la población (los 10^{4} valores), es posible calcular su varianza

```
media = mean(poblacion)
(varianza.pob = sum((poblacion-media)^2)/length(poblacion))
```

```
## [1] 101.2158
```

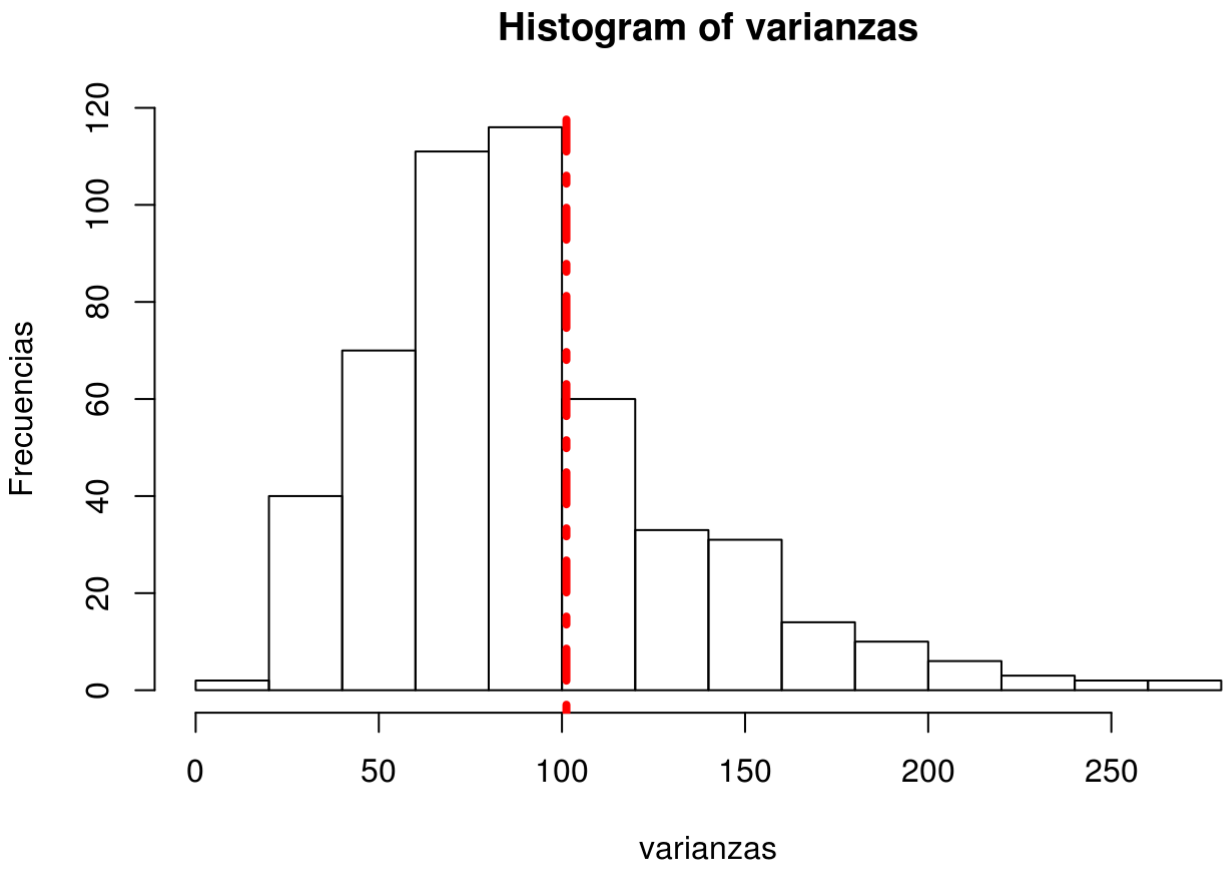
Ahora vamos a tomar `num.muestras` muestras de tamaño ` `tamanno.muestra` . Esto lose hace con un `blucle for` Para cada muestra, se calcula su varianzay su covarianza, y se almacenan en sendos vectores llamados, respectivamente, `varianzas` y `cuasivarianzas` .

```
# definir tamanno de cada muestra y el numero de muestras
tamanno.muestra = 10
num.muestras = 500

varianzas = c()   # vector donde guardar la varianza de cada muestra
cuasivarianzas = c()   # vector donde guardar la cuasivarianza de cada muestra
for(i in 1:num.muestras){ # para cada valor de i entre 1 y num.muestras, ejecuta el siguiente codigo
     muestra.aux = sample(poblacion, size = tamanno.muestra, replace = TRUE) # toma una muestra
     media.aux = mean(muestra.aux)
     varianzas = c(varianzas,  sum((muestra.aux-media.aux)^2)/length(muestra.aux)) # calcula su varianza y annade al vector varianzas
     cuasivarianzas = c(cuasivarianzas, var(muestra.aux))  # calcula su cuasivarianza y annade al vector cuasivarianzas
}
```

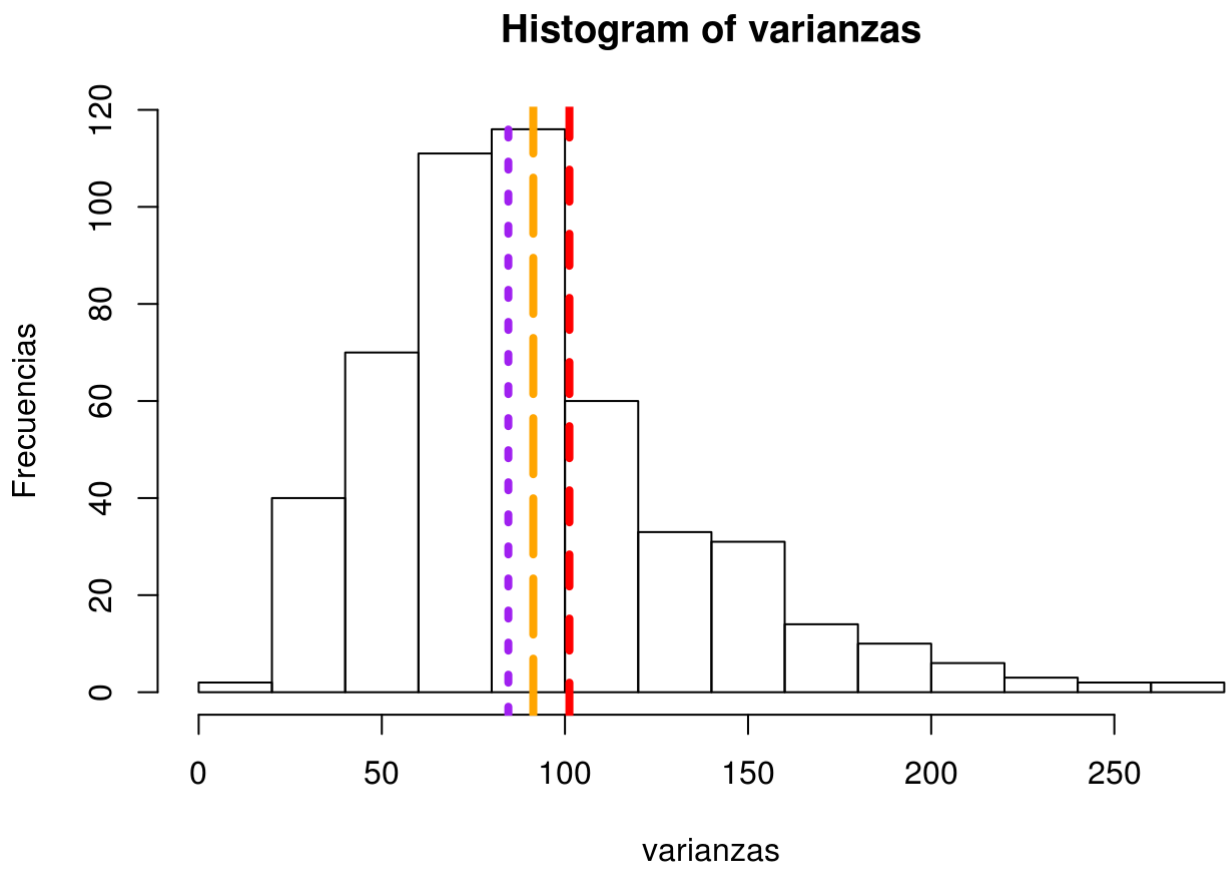
A continuación, representamos en un histograma las 500 varianzas muestrales (estimadas) y marcamos también la varianza poblacional 101.2158 con una linea a trazos discuntinuos roja

```
hist(varianzas, ylab = "Frecuencias")
abline(v = varianza.pob, lty = 4, lwd = 4, col = "red")
```



Aunque es cierto que la mayoría de las varianzas están “cerca”, en torno a, la verdadera varianza, ¿no te da la sensación de que hay más datos a la izquierda que a la derecha de la linea roja? Podemos conprobar esa impresión representando también la mediana de las varianzas calculadas (¿ves por qué?) en este caso con una linea a puntos morada

```
hist(varianzas, ylab = "Frecuencias")
abline(v = varianza.pob, lty = 2, lwd = 4, col = "red")
abline(v = mean(varianzas), lty = 5, lwd = 4, col = "orange")
abline(v = median(varianzas), lty = 3, lwd = 4, col = "purple")
```




```
par(mfrow=c(1,2))
```

Efectivamente, casi siempre (repite la simulción)

- Como la mediana (en morado) queda por debajo de la varianza poblacional (en rojo) más de la mitad de las estimaciones de la varianza poblacional hechas con la varianza de la muestra están por debajo del verdadero valor de la varianza.
- La media de las varianzas muestrales (en naranja) queda por debajo del valor de la varianza poblacional (en rojo). Eso es lo que significa

$$E[Var(X)] = \frac{n-1}{n}\sigma_X^2$$

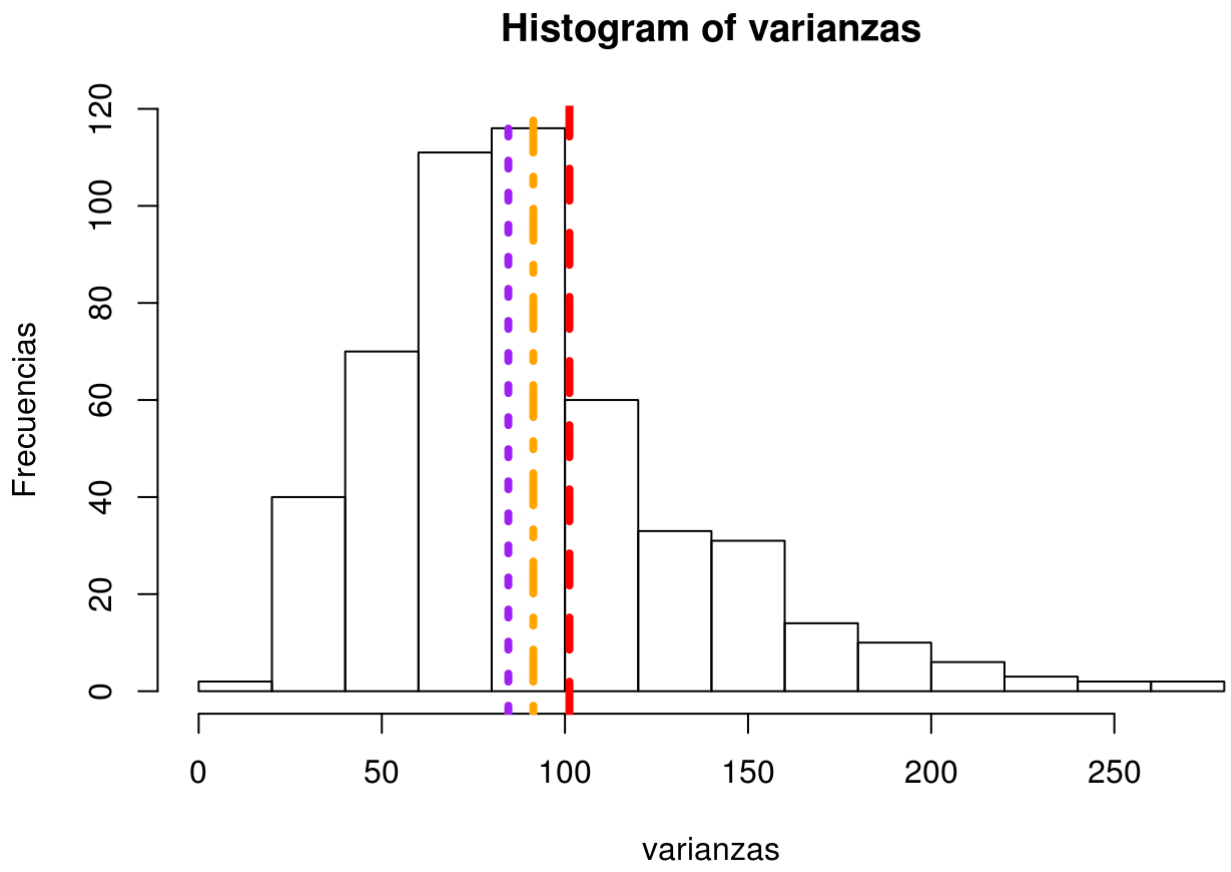
que el valor esperado de la varianza muestral es más pequeño que la varianza poblacinal. En concreto, (n-1)/n veces menor.

Puedes ejecutar varias veces el documento reproducible para cambiar los datos, o incluso cambiar la poblacion de base. La teoría nos dice que lo que comentamos en este párrafo es precisamente lo cabe esperar. A eso se le llama sesgo: cabe esperar que la varianza muestral infravalore el valor de la varianza poblacional

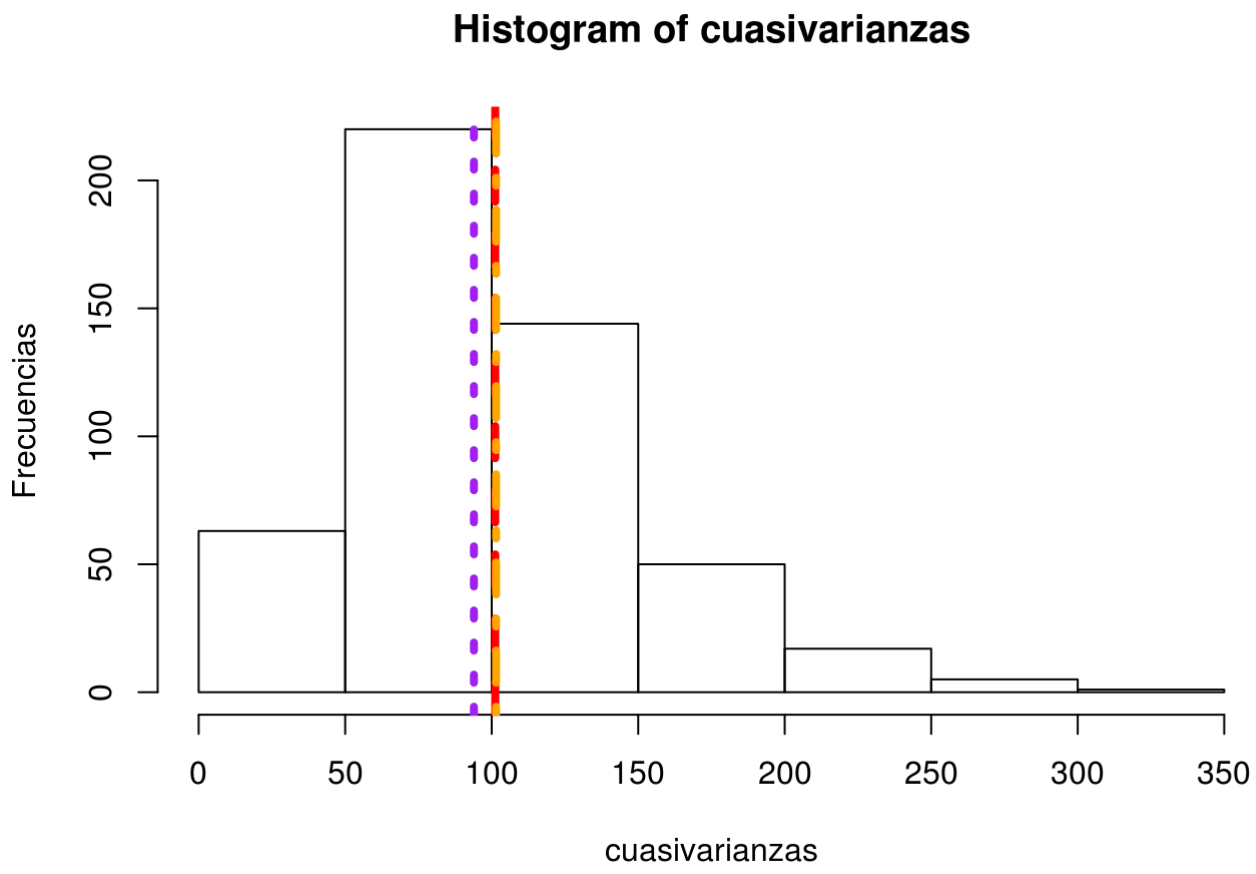
La cuasivarianza entra en escena

Lo que también dice la teoría es que si en lugar de la varianza muestral se usa la cuasivarianza para estimar la varianza, ese sesgo desparece. Vamos a simular las dos situaciones conjuntamente

```
hist(varianzas, ylab = "Frecuencias")
abline(v = varianza.pob, lty = 2, lwd = 4, col = "red")
abline(v = mean(varianzas), lty = 4, lwd = 4, col = "orange")
abline(v = median(varianzas), lty = 3, lwd = 4, col = "purple")
```




```
hist(cuasivarianzas, ylab = "Frecuencias")
abline(v = varianza.pob, lty = 2, lwd = 4, col = "red")
abline(v = median(cuasivarianzas), lty = 3, lwd = 4, col = "purple")
abline(v = mean(cuasivarianzas), lty = 4, lwd = 4, col = "orange")
```




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

varianza.pob-mean(varianzas)
```

```
## [1] 9.864913
```

```
varianza.pob-mean(cuasivarianzas)
```

```
## [1] -0.2851885
```