Práctica 6

Alejandro Cáceres
UPC - Statistics 2019/2020

Objetivo

- Distribuciones de muestreo
- Teorema central del límite

Distribuciones de muestreo

Un contador Geiger mide el número de partículas radioactivas por segundo. Es escencial para saber que tan expuesta está una persona en un lugar con alta radiación (central nuclear, Fukushima, Chernovil).

Distribuciones de muestreo

en `http://mightyohm.com/files/geiger/capture.txt`

hay una muestra real de este tipo de datos

```
> geiger <- read.table("capture.txt", sep=",")
> dim(geiger)
[1] 1238    7
> head(geiger)
   V1 V2   V3 V4       V5   V6    V7
1 CPS  1  CPM  7  uSv/hr 0.03  SLOW
2 CPS  0  CPM  7  uSv/hr 0.03  SLOW
3 CPS  0  CPM  7  uSv/hr 0.03  SLOW
4 CPS  1  CPM  8  uSv/hr 0.04  SLOW
5 CPS  1  CPM  9  uSv/hr 0.05  SLOW
6 CPS  0  CPM  9  uSv/hr 0.05  SLOW
```

Distribuciones de muestreo

Nos interesa saber cual es el promedio de detecciones por segundo (segunda columna), para saber si un trabajador está expuesto a niveles demasiado altos.

Imaginemos que seguimos en tiempo real a un trabajador que entra a limpiar un reactor nuclear.

Nuestro trabajo es garantizar que el trabajador no esté expuesto a niveles radiación excesivos.

El promedio de partículas detectadas por segundo debe ser menor de 0.4 para la salud del trabajador.

Distribuciones de muestreo

Empezamos observando las detecciones del contados Geiger en los primeros 10 segundos depués de que el trabajador entra al reactor.

```
detecciones <- geiger[,2]
detecciones10 <- detecciones[1:10]
> detecciones10
 [1] 1 0 0 1 1 0 1 1 0 0
xbar <- mean(detecciones10)
> xbar
[1] 0.5
```

Distribuciones de muestreo

El promedio de detección debe es mayor que 0.4, pero:

Qué tan confiados estamos de que el trabajador está en una zona de peligro?

Al fin y al cabo si esperamos y obervamos las detecciones de los siguientes 10 segundos (11,...20) nos dará otro valor ($xbar = 0.2$).

Distribuciones de muestreo

- $\bar{X}$ (con X mayúscula) es una variable aleatoria.
- $\bar{x} = 0.5$ (con x minúscula) es el resultado de un experimento sobre $\bar{X}$.
- Si conocemos cómo se distribuye $\bar{X}$ podemos **predecir** qué tan probable es obtener $\bar{x} < 0.4$ en un futuros muestreos de 10 segundos.

la función de probabilidad para $\bar{X}$ se conoce como una función de **distribución se muestreo** para la media (en nuestro caso para una muestra de tamaño $n = 10$).

Distribuciones de muestreo

Supongamos un **modelo** de probabillidad para estos datos y estudiemos **teóricamente** cómo se comporta $\bar{X}$ bajo el modelo.

O sea: Cómo sería la distribución de valores de $\bar{x}$ cuando repetimos muchas veces el experimento de contar el número de partículas readioactivas por segundo durante 10 segundos y calculamos la media?

Después volveremos a los datos...

Poisson

Primero recordemos que de las funciones de distribución para el conteo de eventos en un intervalo determinado es la distribución de Poisson.

$$f(x;\lambda) = Pr(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Donde λ, en nuestro caso, es el promedio en el número de detecciones de partículas radioactivas (x) por segundo. En R: **dpoiss(x, lambda)**

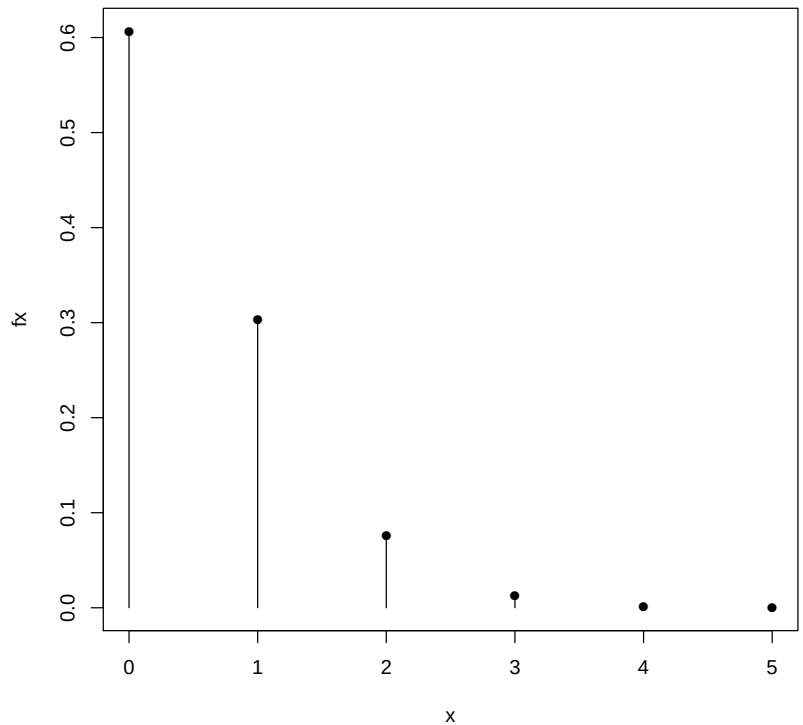
Vamos a asumir que $\bar{X}$ es un estimador de λ, es decir que en últimas podemos remplazar λ por el resulatdo de nuestro experimento $\bar{x} = 0.5$.

Poisson

dibujemos la distribución:

```
x <- 0:5

fx <- dpois(x, lambda=0.5)
names(fx) <- x

plot(x,fx, type="p", pch=16)
for(i in 1:11)
lines(c(x[i], x[i]), c(0, fx[i]))
```

Poisson



Poisson

Queremos saber cómo se distribuye el promedio de detecciones ($\bar{X}$) en 10 mediciones (una cada segundo), cuando cada detección (X) se distribuye $Pois(\lambda = 0.5)$

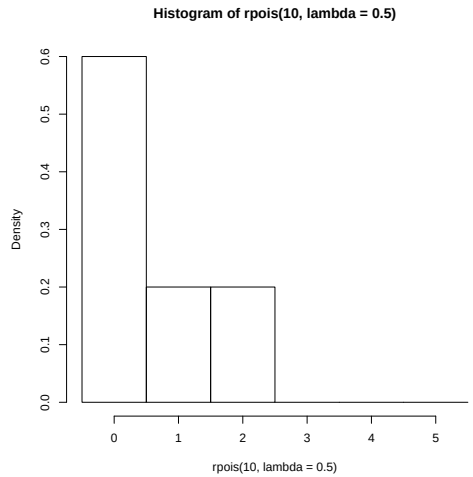
La generación de **10** valores aleatorios que siguen una Poisson se hace con **rpois(10, lambda=0.5)**

```
conteos <-  rpois(10, lambda=0.5)
> conteos
 [1] 0 0 1 0 0 1 0 0 1 0
```

hagamos el histograma de un experimento (muestra) de $n = 10$ mediciones.

Distribuciones de muestreo: Media muestral

```
hist(rpois(10, lambda=0.5), freq=FALSE, breaks=seq(-0.5,5.5)
```



Este es un experimento con 10 mediciones con su media $\bar{x} = 0.33$

Cada nuevo experimento con 10 mediciones tendrá su propia $\bar{x}$.

Distribuciones de muestreo: Media muestral

Queremos generar muchos valores de $\bar{x}$ para ver su distribución.

Podemos empezar haciendo una función general que compute $\bar{x}$ en un experimento de 10 mediciones.

Como harías la función **Xbar** tal que tome n=10 valores de una variable de Poisson con $\lambda = 5$ y compute su media

```
> Xbar(10)
[1] 0.7
```

Distribuciones de muestreo: Media muestral

```
Xbar <- function(n)
{
  conteos <- rpois(n, lambda=0.5)
  mean(conteos)
}

> Xbar(10)
[1] 0.7
```

Queremos ver un histograma de muchos valores de Xbar ($\bar{X}$) para ver como se distribuye. Hagamos 600 valores para $\bar{X}$ recordando la función **sapply**.

Distribuciones de muestreo: Media muestral

Recordemos: la aplicación de Xbar sobre dos experimentos de 10 mediciones se puede hacer explicitamente con

```
c(Xbar(10), Xbar(10))
```

o con **sapply**:

```
sapply(c(10,10), Xbar)
```

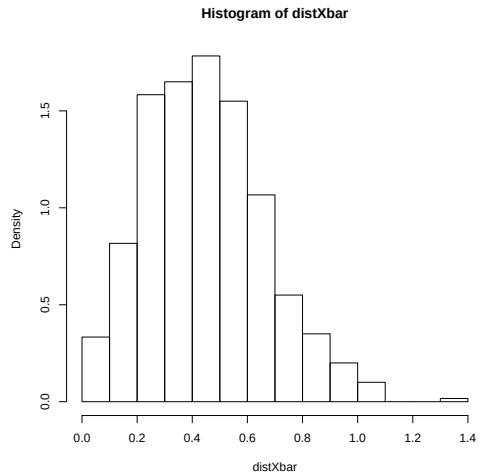
De tal forma que:

```
sapply(rep(10,600), Xbar)
```

es la aplicación de Xbar 600 veces. Cuál es el histograma de estos 600 promedions $(\bar{x}_1, \bar{x}_2 ... \bar{x}_{600})$?

Distribuciones de muestreo: Media muestral

```
distXbar <- sapply(rep(10,600), Xbar)
hist(distXbar, freq=FALSE)
```



Bajo el modelo $Pois(x, lambda = 0.5)$ no es improbable $\bar{X} < 0.4$, pero si muy improbable obtener $\bar{X} > 1$.

Cuál es la probabilidad $Pr(\bar{X} < 0.4)$?

Distribuciones de muestreo: Media muestral

Cuando n es grande (> 30) sabemos que por el TCL

$$\bar{X} \sim N(\mu_{\bar{X}}, \sigma_{\bar{X}})$$

$$\mu_{\bar{X}} = E(X) = \mu = 0.5$$

$$\sigma^2_{\bar{X}} = Var(X)/\sqrt{n} = \sigma^2_X/\sqrt{n} = 0.5/\sqrt{n}$$

Recordemos que para una distribucion de Poisson $\mu = \sigma^2 = \lambda$ ($=$0.5 para nuestros datos).

Entonces: Bajo el TCL podemos usar la distibución normal para calcular

$$P(\bar{X} < 0.4)$$

Distribuciones de muestreo: Media muestral

Pero con n=10 no podemos usar este teorema:
Veamos por qué.

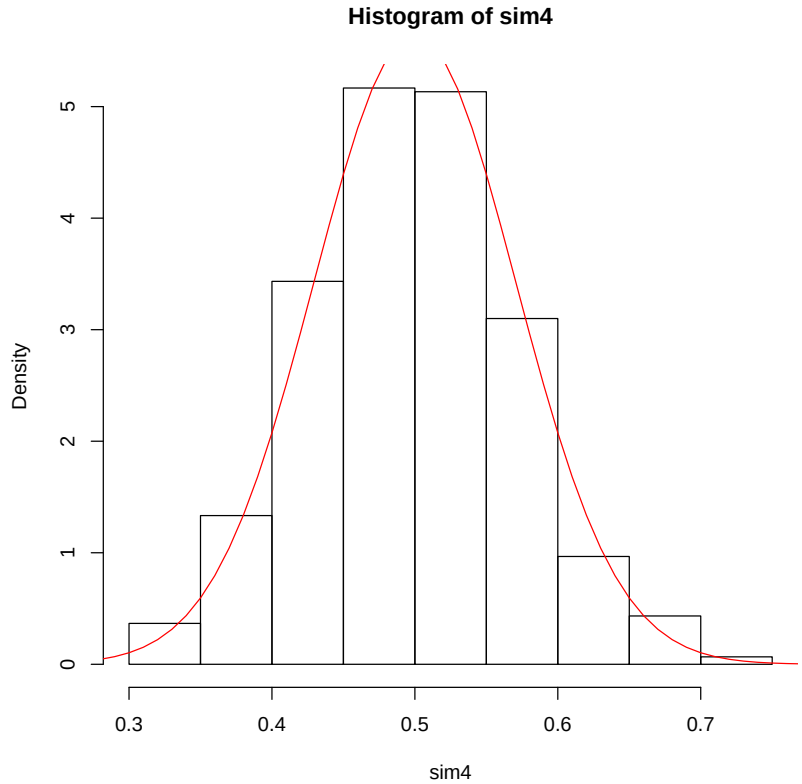
Pinta la distribución normal; `dnrom(x,mu,sigma)` que le corresponde a la distribución de `Xbar(100)` en el intervalo:

```
xprom<-seq(0,1.5,0.01)
```

Distribuciones de muestreo: Media muestral

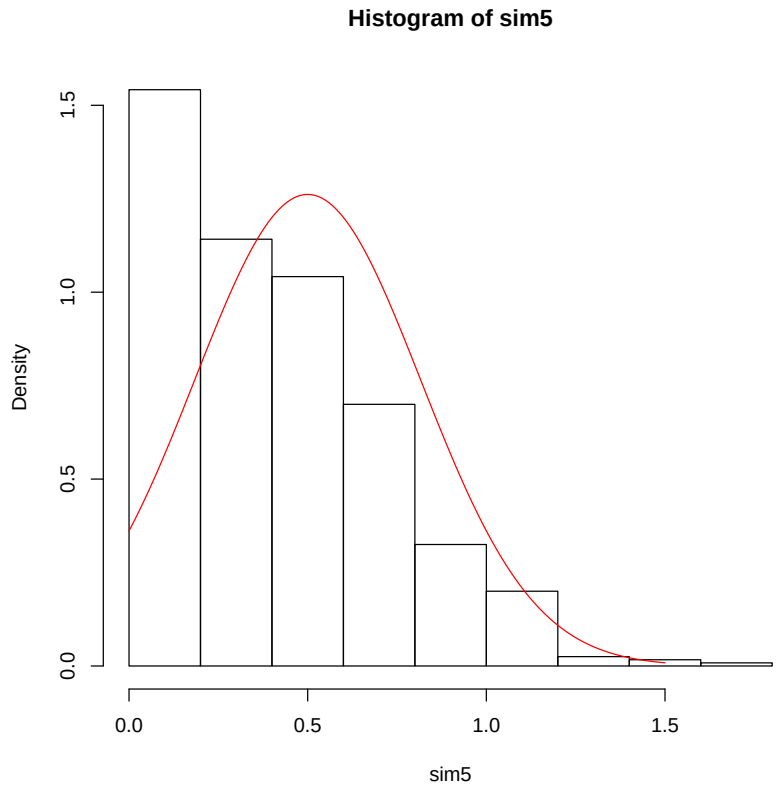
```
sim <- sapply(rep(100,600), Xbar)
hist(sim, freq=FALSE)

xprom <- seq(0,1.5,0.01)
normvals <- dnorm(xprom, mean=0.5, sd=sqrt(0.5)/sqrt(100))
lines(xprom, normvals, col="red")
```

Distribuciones de muestreo: Media muestral



$$\mu = 0.5,\ \sigma = \frac{\sqrt{0.5}}{\sqrt{n}}$$

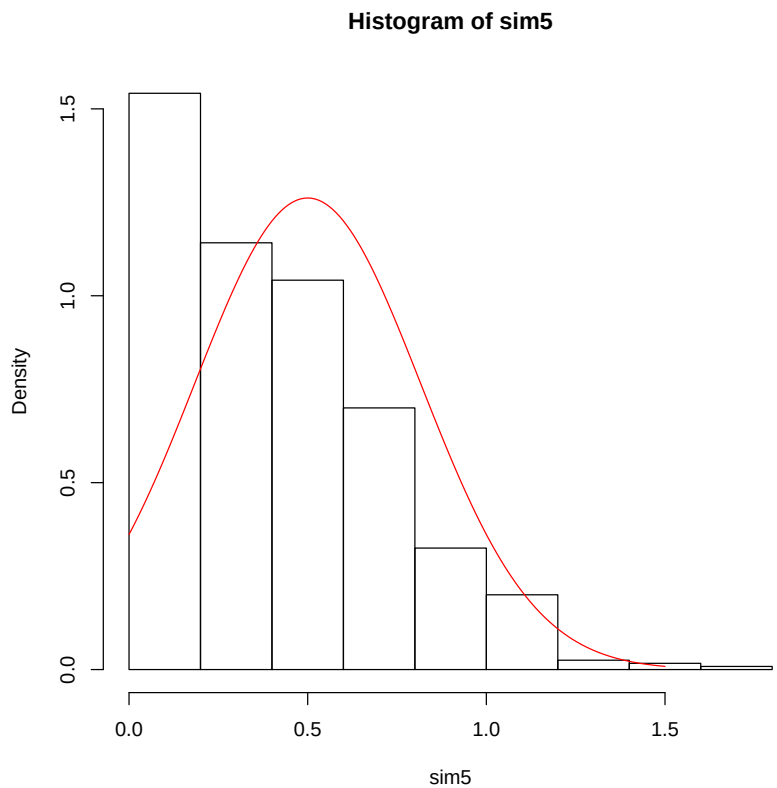
Distribuciones de muestreo: Media muestral



confirma que la aproximación no es buena para n=5

Distribuciones de muestreo: Media muestral

```
sim <- sapply(rep(5,600), Xbar)
hist(sim, freq=FALSE)

xprom <- seq(0,1.5,0.01)
normvals <- dnorm(xprom, mean=0.5, sd=sqrt(0.5)/sqrt(5))
lines(xprom, normvals, col="red")
```



Distribuciones de muestreo: Media muestral

Tomemos más datos, ahora 50 mediciones (los primeros 50 segundos)

```
detecciones50 <- detecciones[1:50]
xbar <- mean(detecciones50)
> xbar
[1] 0.36
```

La situación cambia! ahora $xbar < 0.4$ y el trabajador estaría en zona segura. Pero con qué probabilidad $Pr(\bar{X} < 0.4)$?

Distribuciones de muestreo: Media muestral

Ahora tenemos mas datos (50) y podemos calcular $Pr(\bar{X} < 0.4)$ con el TCL

```
> pnorm(0.4, mean=0.36, sd=sqrt(0.36/50))
[1] 0.6813241
```

Según los 50 primeros datos el trabajador tiene una probabilidad de 0.68 de estar en zona segura.

Cuántos datos necesitamos para estar muy seguros?

Distribuciones de muestreo: Media muestral

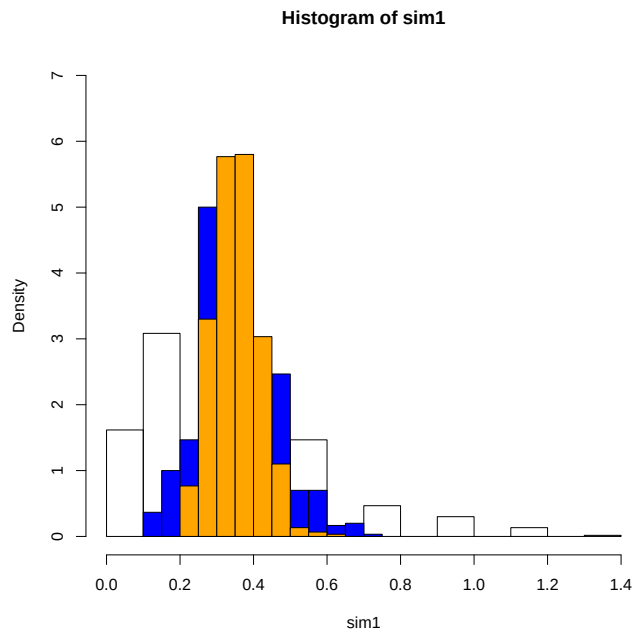
El modelo ha cambiado ahora para estos 50 datos el estimador de $\hat{\lambda} = \bar{x} = 0.36$ (le ponemos el gorro a λ para remarcar que es un valor estimado)

```
Xbar <- function(n)
{
  conteos <- rpois(n, lambda=0.36)
  mean(conteos)
}
```

Hagamos los histogramas para muestras de n=5, n=30 (blue), n=100(orange)

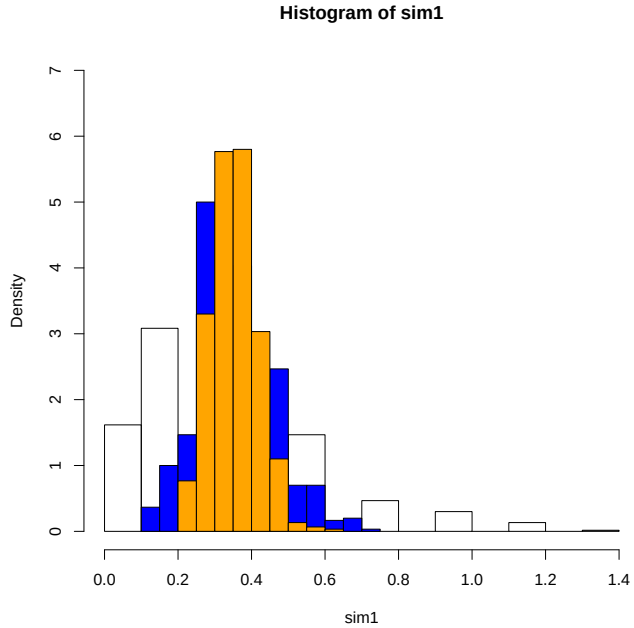
Distribuciones de muestreo: Media muestral

```
sim1 <- sapply(rep(5,600), Xbar)
hist(sim1, freq=FALSE, ylim=c(0,7))

sim2 <- sapply(rep(30,600), Xbar)
hist(sim2, freq=FALSE, add=TRUE, col="blue")

sim3 <- sapply(rep(100,600), Xbar)
hist(sim3, freq=FALSE, add=TRUE, col="orange")
```



Nota: add=TRUE para añadir histogramas.

Distribuciones de muestreo: Media muestral



A medida que las medidas aumentan (n) la varianza de $\bar{X}$, que llamamos $\sigma^2_{\bar{X}}$, es cada vez mas pequeña ($\sqrt{0.5/n}$) y cada vez tenemos mas confianza de que nuestro promedio dado por los datos $\hat{\lambda} = \bar{x}$ está cerca del verdadero valor de λ: $\lambda \sim \hat{\lambda} = 0.36$

Distribuciones de muestreo: Media muestral

De hecho todos nuestros datos constituyen una muestra de n=1238.

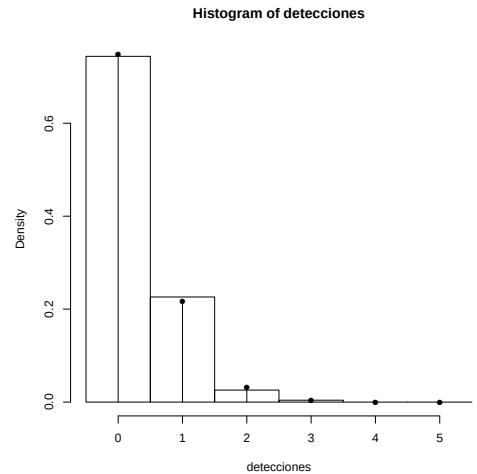
```
> mean(detecciones)
[1] 0.2899838
> pnorm(0.4, mean=0.2899838, sd=sqrt(0.2899838/1238))
[1] 1
```

y muestran una probabilidad de 1 de que nuestro trabajador esta en zona segura.

Distribuciones de muestreo: Media muestral

También podemos comprobar que estos datos reales están muy bien descritos por una distribución de Poisson.

```
hist(detecciones, freq=FALSE, breaks=seq(-0.5,5.5))

x <- 0:5
fx <- dpois(x, lambda= 0.2899838)

points(x,fx, type="p", pch=16)
for(i in 1:6)
lines(c(x[i], x[i]), c(0, fx[i]))
```



Distribuciones de muestreo: Varianza muestral

Veamos la distribución no para la media sino para la variaza muestral S^2 de n mediciones.

Imaginemos que tomamos 100 medidas aleatorias de la altura de hombres. Queremos saber si estas 100 medidas nos dan una medida precisa de qué tan variable es la altura en la población, para por ejemplo diseñar bicicletas que acomoden a la mayoría de la población.

Distribuciones de muestreo: Varianza muestral

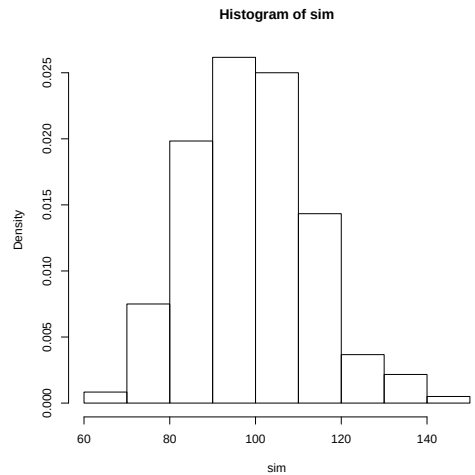
Veamos que pasaría teóricamente con 100 mediciones, si **asumimos** un modelo de probabilidad normal para la altura de los hombres $\mu = 175cm$ y desviación típica $\sigma = 10cm$.

Entonces, como las mediciones provienen de una distribución normal en la función **Xbar** remplazamos `rpois` por `rnorm` y **mean** por $sd()^2$

```
Scuadrado <- function(n)
{
  medidas <- rnorm(n, mean=175, sd=10)
  sd(medidas)^2
}

sim <- sapply(rep(100,600), Scuadrado)
hist(sim, freq=FALSE)
```

Distribuciones de muestreo: Varianza muestral

```
sim <- sapply(rep(100,600), Scuadrado)
hist(sim, freq=FALSE)
```



Distribuciones de muestreo: Varianza muestral

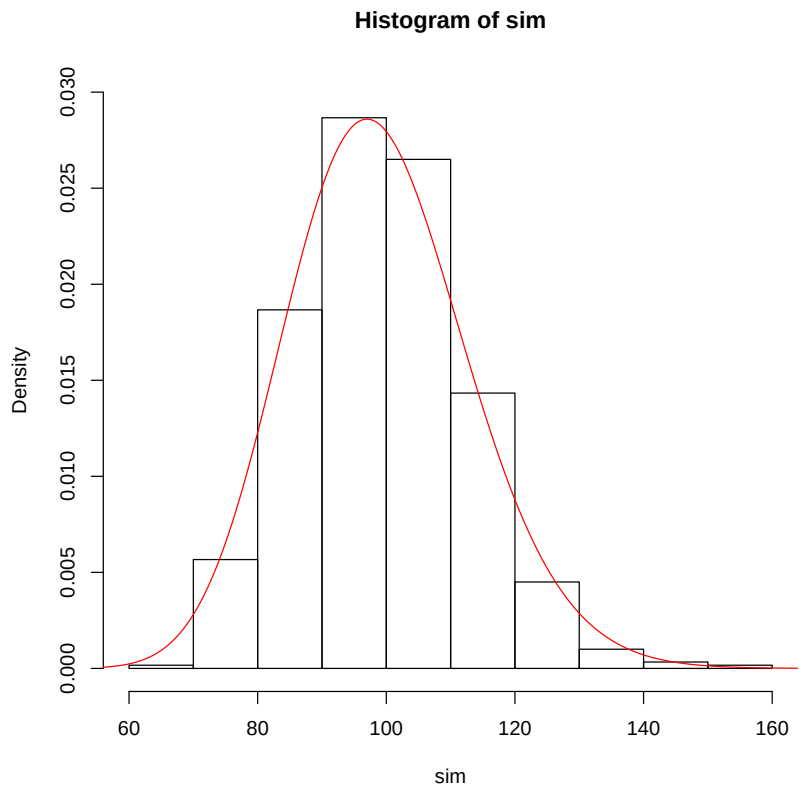
La varianza muestral S^2 es una variable aleatoria que se distribuye como una distribucón χ^2 con n-1 grados de libertad cuando las mediciones provienen de una distribución normal, de tal forma

$$\frac{(n-1)S^2}{\sigma^2} \rightarrow \chi^2_{n-1}$$

en R existe la funcion **dchisq(x,df)** que da la distribución χ^2 con df grados de libertad

Distribuciones de muestreo: Varianza muestral

```
sim <- (100-1)*sapply(rep(100,600), Scuadrado)/10^2
hist(sim, freq=FALSE, ylim=c(0,0.03))

chi.var <- seq(0,200,0.01)
f.chi <- dchisq(chi.var, df=(100-1))
lines(chi.var, f.chi, col="red")
```



Distribuciones de muestreo: Varianza muestral

En este modelo para la altura, el rango intercuartílico para desviación típica se es

```
chi.cuartiles <- qchisq(c(0.25,0.75), df=(100-1))
> s.cuartiles <- sqrt(chi.cuartiles*10^2/(100-1))
> s.cuartiles
[1]  9.491156 10.449168
```

Distribuciones de muestreo: Varianza muestral

```
> s.cuartiles
[1]  9.491156 10.449168
```

o sea que si decidimos usar 100 mediciones para calcular s, el 50% de las veces observaríamos una desviación típica **muestral** s que cae entre (9.49cm, 10.44cm).