Distribuciones en R

Computo de probabilidades

Para todas las variables aleatorias que hemos mencionado hasta ahora (¡y muchas más!), R ha incorporado capacidades de cálculo de probabilidades. La sintaxis se divide en dos partes: la raíz y el prefijo. La raíz determina de qué variable aleatoria estamos hablando, y aquí están los nombres de las que hemos cubierto hasta ahora:

- `binom` binomial
- `geom` geometrica
- `pois` Poisson
- `unif` uniforme
- `exp` nexponencial
- `norm` normal

Y los prefijos disponibles son los siguientes:

- `p` computa la **distribución acumulada**
- `d` computa la **distribución de probabilidad** o **distribución de densidad**

- `r` muestrea
- `q` es la distribución cuantil

Por el momento nos vamos a enfocar en `p` y `d`

Por ejemplo si X es una variable aleatoria que se distribuye binomial con $n = 10$ y $p = 0.3$ y queremos computar $P(X = 5)$ entonces:

```
dbinom(5, 10, 0.3)
```

```
## [1] 0.1029193
```

Recordemos que siempre podemos pedir la ayuda de la función para ver cuales son sus parámetros

Si queremos ahora calcular $P(1 \leq X \leq 5)$ podemos hacerlo de dos maneras:

Opción 1

$$P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$

```
dbinom(5, 10, 0.3) +
  dbinom(4, 10, 0.3) +
  dbinom(3, 10, 0.3) +
```

```
  dbinom(2, 10, 0.3) +
  dbinom(1, 10, 0.3)

## [1] 0.9244035
```

Para no volvernos viejos tan rápido podemos usar la opción de `dbinom` también permite vectores como argumentos, por lo que podemos calcular todas las probabilidades necesarias de esta manera:

Opción 2

```
dbinom(1:5, 10, 0.3)

## [1] 0.1210608 0.2334744 0.2668279 0.2001209 0.1029193

sum(dbinom(1:5, 10 , 0.3))

## [1] 0.9244035
```

Algunos ejemplos con COVID 19

Para comenzar vamos a obtener una serie de datos de la página oficial de la OMS o WHO

Vamos a tratar de responder algunos interrogantes respecto a la incidencia de la enfermedad

Trabajaremos con los datos totales y de los últimos 7 días, debemos elegir Argentina y un país de cada region

Definimos Prevalencia como la proporción de individuos de un grupo o una población (en medicina, persona), que presentan una característica o evento determinado (en medicina, enfermedades). Por lo general, se expresa como una fracción, un porcentaje o un número de casos por cada 10 000 o 100.000 personas.

Mientras que la incidencia es el número de casos nuevos de una enfermedad en una población determinada y en un periodo determinado.

Definimos la Incidencia para un periodo de tiempo dado como:

$$Incidencia = \frac{numero.de.personas.enfermas}{numero.de.habitantes}$$

La prevalencia es un término que significa estar extendido y es distinto de la incidencia .

La prevalencia es una medida de todos los individuos (casos) afectados por la enfermedad en un momento determinado, mientras que la incidencia es una medida del número de nuevos individuos (casos) que contraen una enfermedad durante un período de tiempo particular.

La prevalencia responde a “¿Cuántas personas tienen esta enfermedad en este momento?” o “¿Cuántas personas han tenido esta enfermedad durante este período de tiempo?”. La incidencia responde a “¿Cuántas personas adquirieron la enfermedad durante un período

de tiempo específico?". Sin embargo, matemáticamente, la prevalencia es proporcional al producto de la incidencia y la duración promedio de la enfermedad.

$$Prevalencia = Incidencia * duración$$

¿Cuál es la probabilidad de que una persona contraiga la enfermdad en los últimos 7 días?

Asumimos una distribución Bernoulli o binomial con n=1

$$P(X = x) = p^x(1-p)^{(n-x)}$$

con

$$x = 0{,}1$$

Aqui p tomará el valor de la incidencia media y x sera 1

Entonces por ejemplo para el mundo con una incidencia de $\frac{58.848}{100000} = 0.00058848$:

```
dbinom(x=1, size=1, prob=58.848/100000)
```

```
## [1] 0.00058848
```

Obviamente observando la formula podiamos darnos cuenta que ese valor iba a ser igual a p

¿Ahora, cuál es la probabilidad de que un integrante de una burbuja de 15 personas contraiga la enfermdad en los últimos 7 días?

```
dbinom(x=1, size=15, prob=(58.848/100000))
```

```
## [1] 0.008754753
```

y dos?

```
dbinom(x=2, size=15, prob=58.848/100000)
```

```
## [1] 3.608521e-05
```

¿Qué pasa en Argentina y en los paises seleccionados de cada región?

Comparar en Argentina p total, de los ultimos 7 días y de las últimas 24 hs.

Variables continuas

Ahora, siendo X una v.a. de media 1 ($\mu = 1$) y desvío estandar 5 ($\sigma = 5$) y queremos calcular $P(X \leq 1)$

Como se trata de una distribución de variables continuas es complejo calcular probabilidades como el ejemplo anterior, aquí tenemos que integrar. Entoncess:

$$\int_{-\infty}^{1} \mathrm{N}(\mathrm{x}, 1, 5)\ dx$$

Pero, la probabilidad que estamos tratando de calcular podemos computarla a traves de $F(x)$ usando pnorm.

```
pnorm(1,1,5)
```

```
## [1] 0.5
```

Si queremos calcular $P(1 \leq X \leq 5)$ Aquí, calculamos la probabilidad de que X sea menor o igual a cinco, y restamos la probabilidad de que X sea menor que 1. Que es lo mismo que:

$$P(X \leq 5) - P(X \leq 1)$$

```
pnorm(5,1,5) - pnorm(1,1,5)
```

```
## [1] 0.2881446
```

Ejemplo

Debemos elegir siete mujeres al azar de una universidad para formar una versión inicial de un video juego de basquet femenino. Las alturas de las mujeres en se distribuyen normalmente con una media de 163.83 cm y una desviación estándar de 5.715 cm. ¿Cuál es la probabilidad de que 3 o más de las mujeres midan 172.72 cm o más?

Para computar esta probabilidad, primero determinamos la probabilidad de que una sola mujer seleccionada al azar mida 172.72 cm o más alta. Sea X una variable aleatoria normal con media 162.56 y desviación estándar 5.715. Calculamos $P(X \geq 172.72)$ usando pnorm:

```
pnorm(172.72, 162.56, 5.715, lower.tail = FALSE)
```

```
## [1] 0.03772018
```

Ahora, tenemos que calcular la probabilidad de que 3 o más de las 7 mujeres midan 172.72 cm o más. Dado que la población de todas las mujeres en una universidad es mucho mayor que 7, el número de mujeres en la configuracón inicial que tienen 172.72 cm o más es binomial con $n = 7$ y 0.03772018, que calculamos en el paso anterior. Entonces, calculamos la probabilidad de que al menos 3 mujeres midan 172.72 cm como:

```
sum(dbinom(3:7, 7, 0.03772018))
```

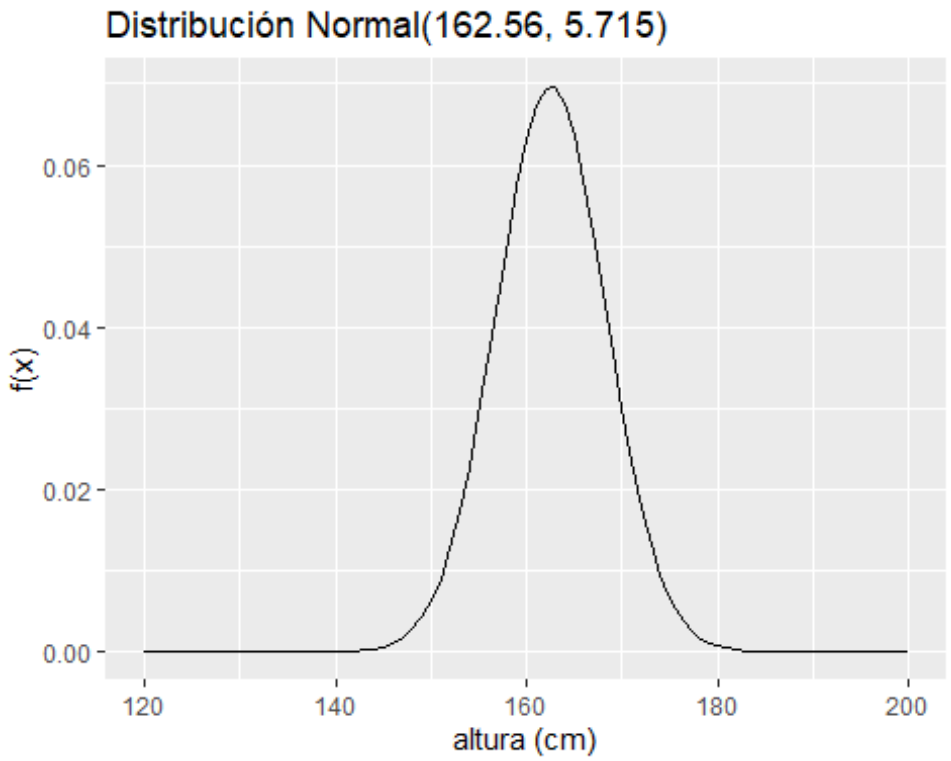
```
## [1] 0.001675265
```

Por lo tanto, hay aproximadamente un 0.17 por ciento de posibilidades de que al menos tres mujeres midan 172.72 cm o más.

Graficar funciones con ggplot2

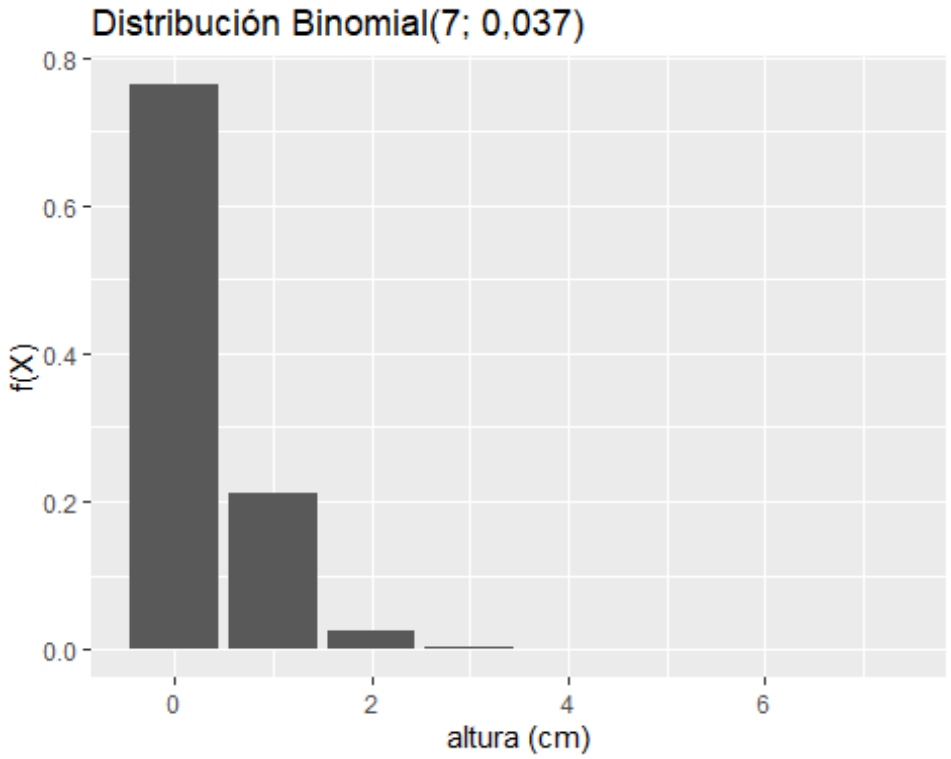
Para graficar la función de densidad del ejemplo anterior, vamos a generar una serie de datos (vector) de datos con los valores de x y los valores de y que queremos representar (usando la función `dnorm()`) y a partir de ese vector vamos a graficar usando `geom_line()`, de la siguiente manera.

```
library(ggplot2)

xvals <- seq(120,200,1)
plotdata <- data.frame(x = xvals, y = dnorm(xvals, 162.56, 5.715))
ggplot(plotdata, aes(x = x, y = y)) +
  geom_line()+
  labs(x="altura (cm)", y= "f(x)", title="Distribución Normal(162.56,
5.715)")
```



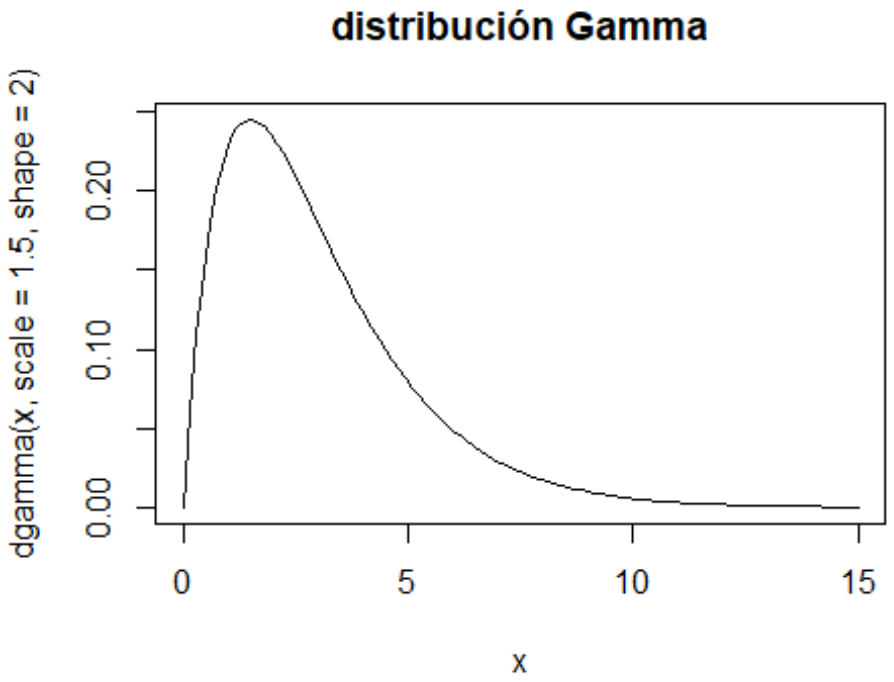
De la misma manera si se trata de una distribución binomial

```
xvals <- 0:7
plotdata <- data.frame(x = xvals, y = dbinom(xvals, 7, 0.03772018))
ggplot(plotdata, aes(x, y)) + geom_bar(stat = "identity")+
  labs(x="altura (cm)", y= "f(X)", title="Distribución Binomial(7; 0,037)")
```

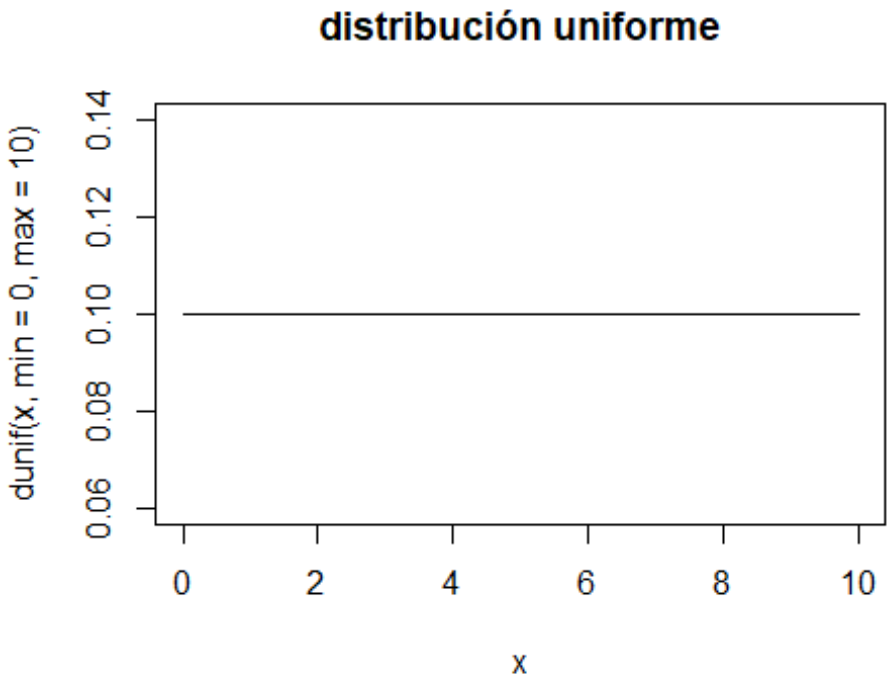


Otras opciones para variables continuas

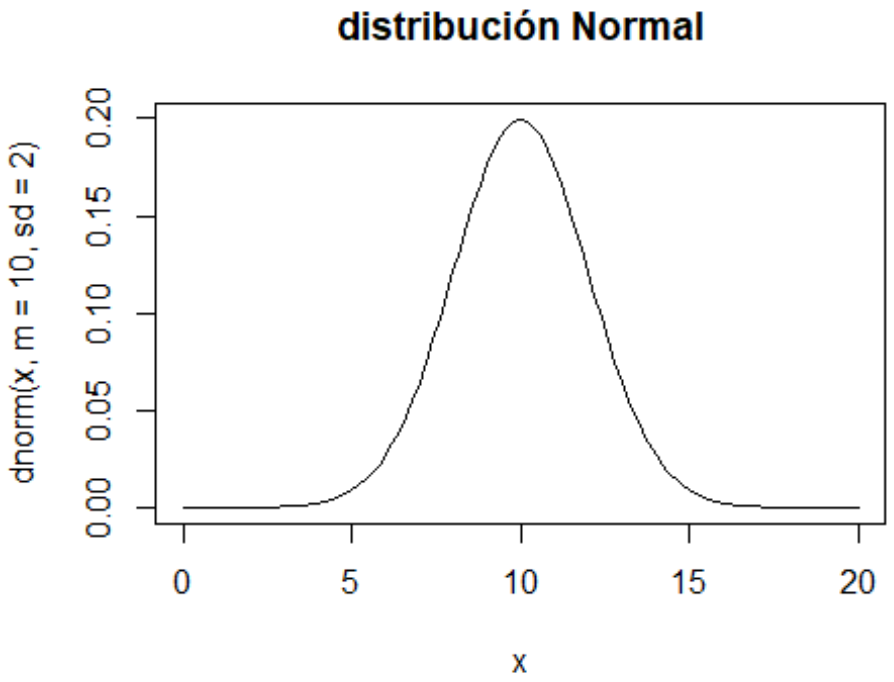
```
curve(dgamma(x, scale=1.5, shape=2),from=0, to=15, main="distribución Gamma")
```




```
curve(dunif(x,min=0,max=10),from=0,to=10, main="distribución uniforme")
```




```
curve(dnorm(x,m=10,sd=2),from=0,to=20,main="distribución Normal")
```



Ejemplo Salario neto mensual

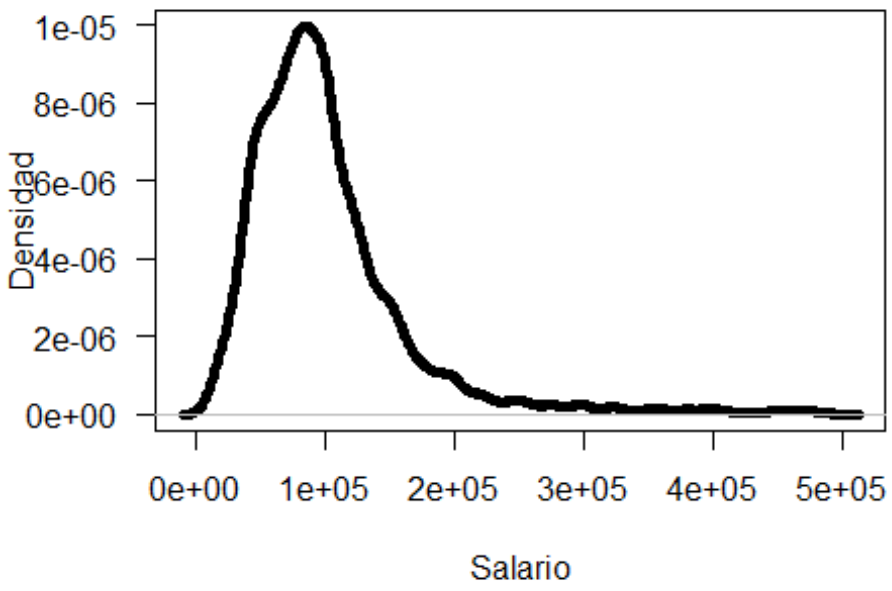
Veamos el comportamiento de la variable salario neto

```
head(salario_neto_gen)
```

```
## # A tibble: 6 x 3
##     neto `Me identifico` genero
##    <dbl> <chr>           <fct>
## 1  90000 Varón Cis       Varón Cis
## 2 109000 Varón Cis       Varón Cis
## 3  39259 Varón Cis       Varón Cis
## 4  91713 Varón Cis       Varón Cis
## 5 137700 Varón Cis       Varón Cis
## 6  38500 Varón Cis       Varón Cis
```

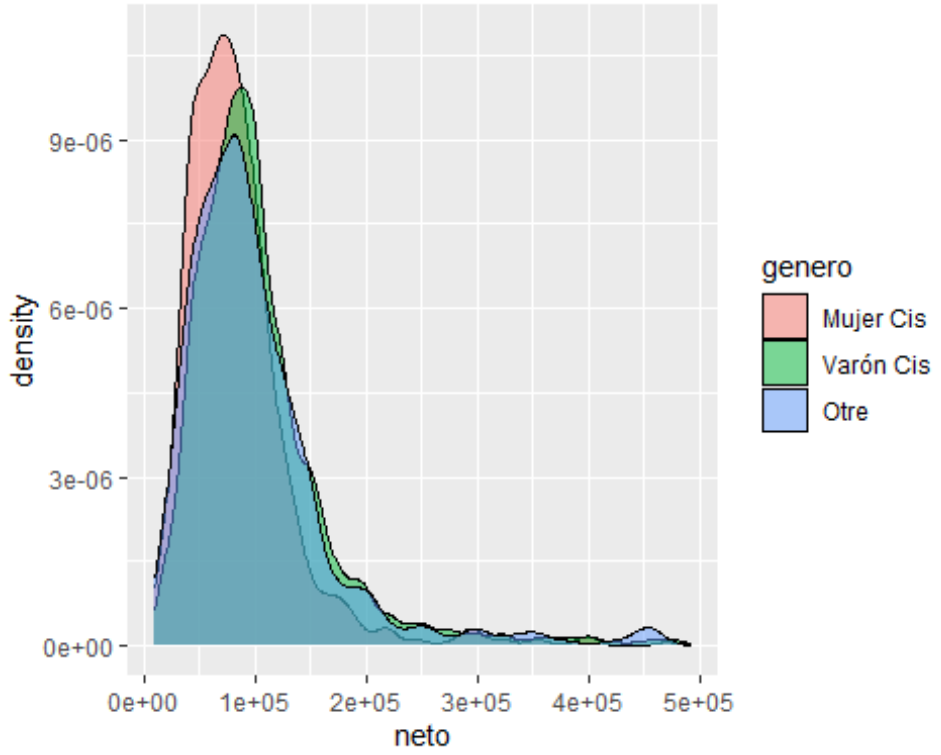
Función de densidad

```
plot(density(salario_neto_gen$neto), main='', lwd=5, las=1,
     xlab='Salario', ylab='Densidad')
```



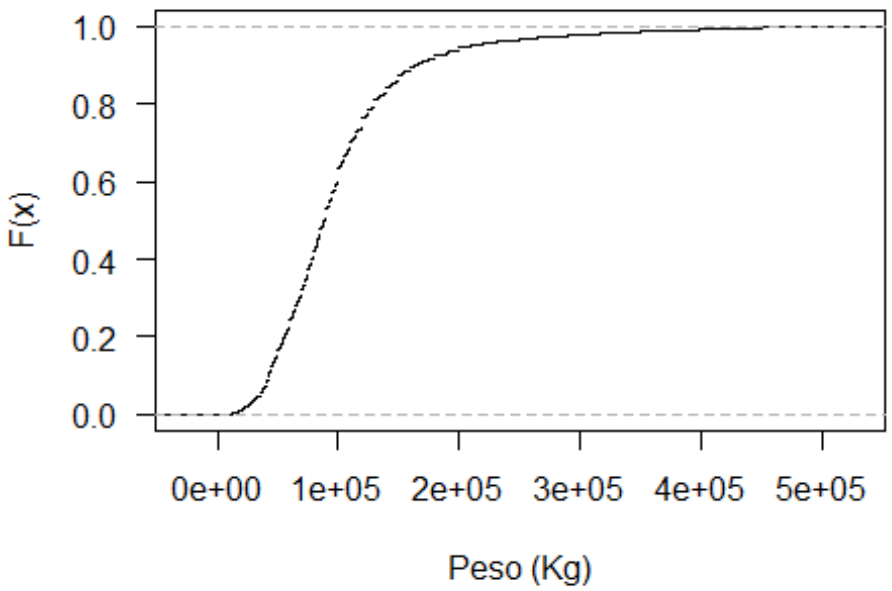
Con ggplot2 por genero

```
ggplot(salario_neto_gen, aes(x=neto)) +
  geom_density(aes(group=genero, fill=genero), alpha=0.5)
```



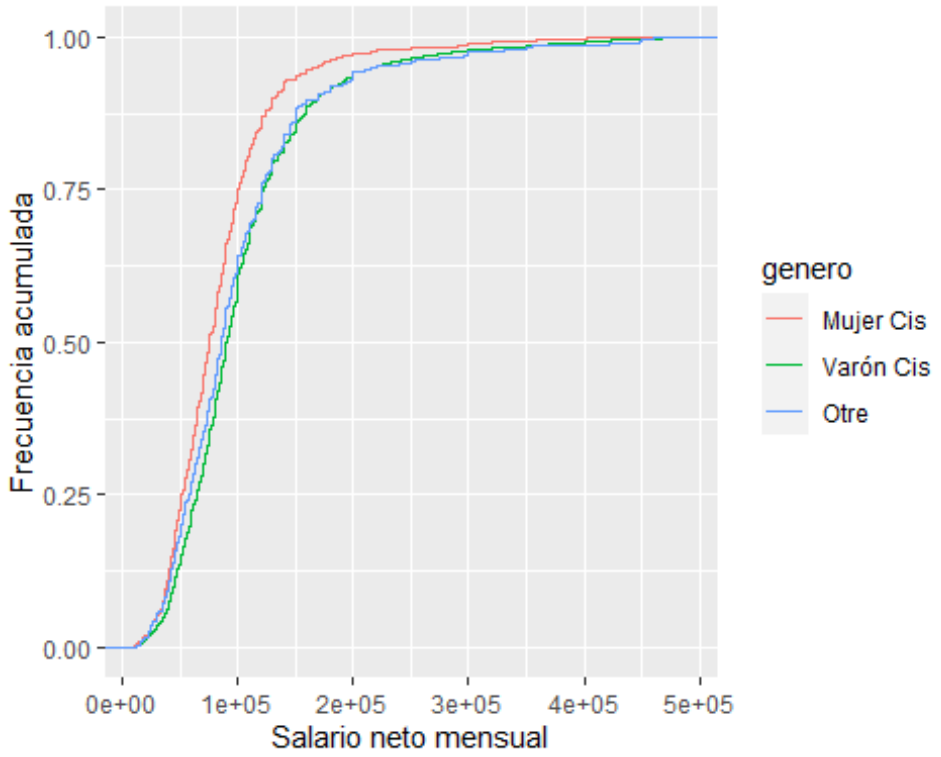
Función de distribución acumulada

```
F <- ecdf(salario_neto_gen$neto)
plot(F, main='', xlab='Peso (Kg)', ylab='F(x)', cex=0.5, las=1)
```



Con ggplot2 por genero

```
ggplot(salario_neto_gen, aes(neto, fill=genero, col=genero))+
  geom_step(stat="ecdf") +
  ylab("Frecuencia acumulada")+
  xlab("Salario neto mensual")
```



Calcular la probabilidad de que una persona de la industria del software cobre un sueldo inferior al salario minimo vital y movil.

SMVM = 21600

```
F(21600)
```

```
## [1] 0.01651572
```

TCL

Media muestral del salario

Ahora muestreamos 10 individuos y calculamos la media del salario sobre esos 10 individuos

Muestreo una vez una muestra de tamaño 10

```
muestra_size10_1 <- sample(salario_neto_gen$neto, size=10)
```

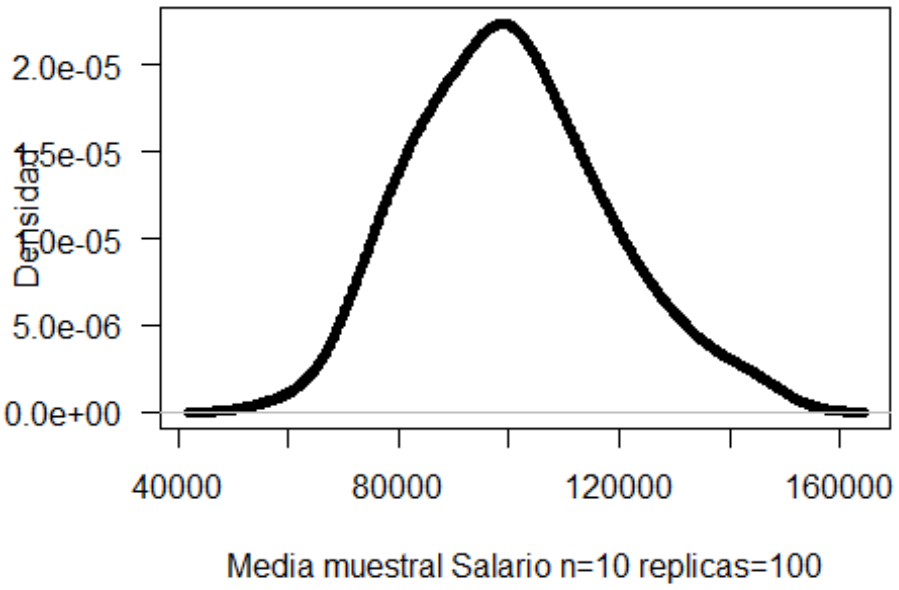
Calculo la media

```
mean(muestra_size10_1)
```

```
## [1] 71362.4
```

Repito el proceso 100 veces

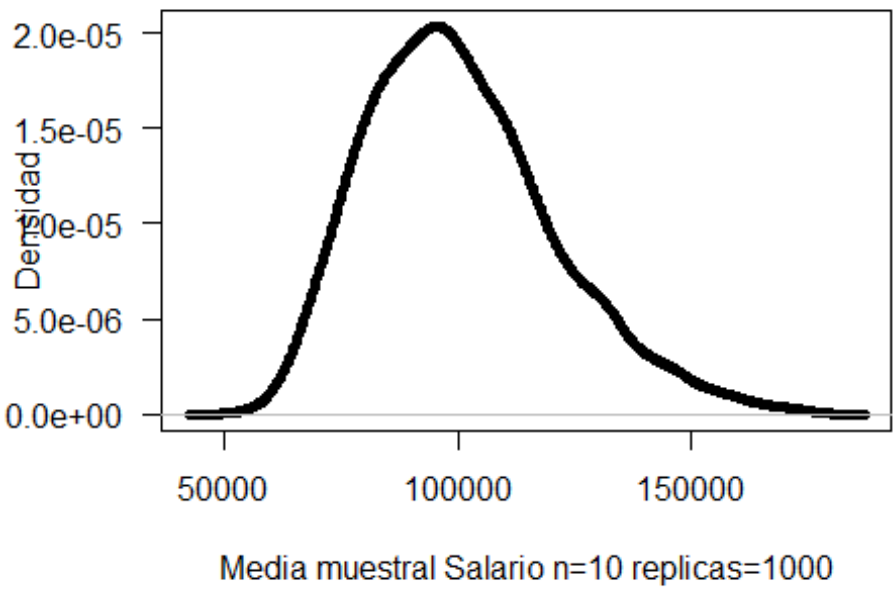
```
medias_muestrales <- replicate(100, mean(sample(salario_neto_gen$neto,
size=10)))

plot(density(medias_muestrales), main='', lwd=5, las=1,
     xlab='Media muestral Salario n=10 replicas=100', ylab='Densidad')
```



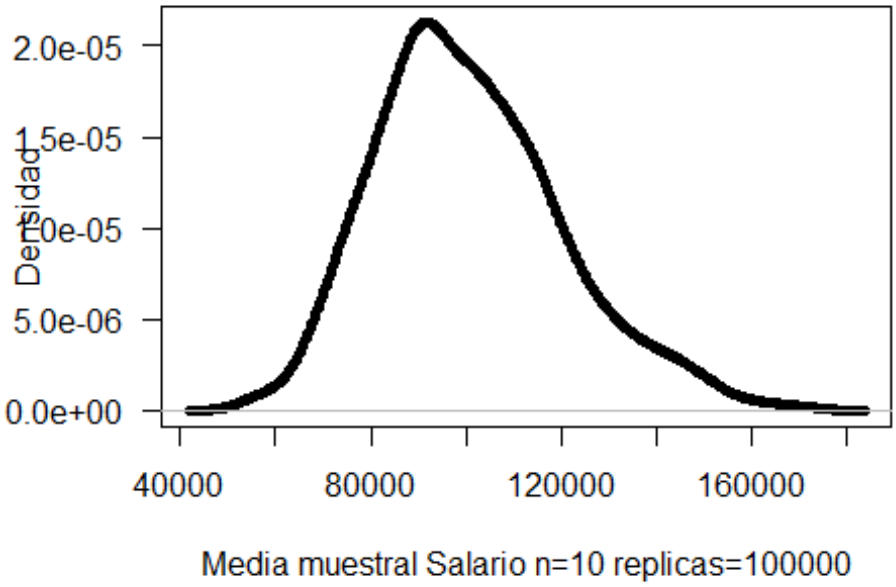

```
mean(medias_muestrales)

## [1] 100412.5

sd(medias_muestrales)

## [1] 17403.1
```

Ahora lo repito 1000 veces y 100000 veces

```
medias_muestrales <- replicate(1000, mean(sample(salario_neto_gen$neto,
size=10)))

plot(density(medias_muestrales), main='', lwd=5, las=1,
     xlab='Media muestral Salario n=10 replicas=1000', ylab='Densidad')
```




```
mean(medias_muestrales)

## [1] 100524.4

sd(medias_muestrales)

## [1] 20167.3

medias_muestrales <- replicate(1000, mean(sample(salario_neto_gen$neto,
size=10)))

plot(density(medias_muestrales), main='', lwd=5, las=1,
     xlab='Media muestral Salario n=10 replicas=100000', ylab='Densidad')
```




```
mean(medias_muestrales)

## [1] 100719.2

sd(medias_muestrales)

## [1] 19713.47
```

Repetir el proceso para un tamaño mestral de 20 y un tamaño mestral de 100, calculando media y desvio estandar para cada caso. Construir una tabla y concluir.