Maestría en Computo Estadístico
Inferencia Estadística
Tarea 2
8 de septiembre de 2020
Enrique Santibáñez Cortés
Repositorio de Git: Tarea 2, IE.

1. Cuando una máquina no se ajusta adecuadamente tiene una probabilidad 0.15 de producir un artículo defectuoso. Diariamente, la máquina trabaja hasta que se producen 3 artículos defectuosos. Se detiene la máquina y se revisa para ajustarla. ¿Cuál es la probabilidad de que una máquina mal ajustada produzca 5 o más artículos antes de que sea detenida? ¿Cuál es el número promedio de artículos que la máquina producirá antes de ser detenida?

RESPUESTA
Sea X el número de artículos producidos antes de que se produzcan 3 artículos defectuosos, entonces podemos decir que $X \sim BN(3, 0.15)$. Por lo tanto, **la probabilidad de que una máquina mal ajustada produzca 5 o más artículos antes de que sea detenida** es (ocupamos la función en R pnbinom(1, 3, 0.15)):

$$\mathbb{P}(X \geq 5) = 1 - \mathbb{P}(X \leq 4) = 1 - \sum_{x=3}^{4} \binom{x-1}{3-1}(1-0.15)^{x-3}(0.15)^3 = 1 - 0.01198125 = 0.9880187.$$

Por como se distribuye X podemos decir que **el número promedio de artículos que la máquina producirá antes de ser detenida** es

$$\mathbb{E}(X) = \frac{r}{p} = \frac{3}{0.15} = 20 \blacksquare.$$

2. Los empleados de una compañía de aislantes son sometidos a pruebas para detectar residuos de asbesto en sus pulmones. Se le ha pedido a la compañía que envíe a tres empleados, cuyas pruebas resulten positivas, a un centro médico para realizarles más análisis. Si se sospecha que el 40 % de los empleados tienen residuos de asbesto en sus pulmones, encuentre la probabilidad de que deban ser analizados 10 trabajadores para poder encontrar a 3 con resultado positivo.

RESPUESTA
Sea Y el número de trabajadores que se realizan las pruebas hasta encontrar 3 empleados con resultados positivos. Y como la probabilidad de que algún empleado tenga residuos de asbesto en sus pulmones (dar positivo en la pruebas) es de 0.40. Entonces podemos concluir que $Y \sim BN(3, 0.4)$. Por lo que la **probabilidad de que deban analizar 10 trabajadores para encontrar a 3 con resultado positivo es** (ocupamos la función en dnbinom(10, 3, 0.40) en R):

$$\mathbb{P}(Y = 10) = \binom{10-1}{3-1}(1-0.40)^{10-3}(0.40)^3 = 0.06449725 \quad \blacksquare.$$

3. Para el siguiente ejercicio es necesario usar R.

a) Considere una moneda desequilibrada que tiene probabilidad p de obtener águila. Usando el comando sample, escriba una función que simule N veces lanzamientos de esta moneda hasta obtener un águila. La función deberá recibir como parámetros a la probabilidad p de obtener águila y al número N de veces que se repite el experimento; y tendrá que regresar un vector de longitud N que contenga el número de lanzamientos hasta obtener un águila en cada uno de los N experimentos.

RESPUESTA
Si X es el número de lanzamientos de la modena hasta obtener un águila, con probabilidad p de obtener

águila en un lanzamiento. Entonces, $X \sim Geo(p)$. Por lo que la función que solicitan sería la simulación de X N veces. Ocupando la siguiente notación de 1:águila y 0:sol:

```
moneda_geometrica <- function(p, N){ # p: probabilidad de aguila, N # repeticiones.
  resultados <- c() # Inicializamos un vector.
  for (i in 1:N) { # Repetimos el experimentos N veces.
    contador <- 0 # Inicializamos el número de lanzamientos.
    while(sample(x=c(1,0), size=1, prob=c(p,1-p))!=1){ # si ya se obtuvo águila deterner.
      contador <- contador + 1
    }
    resultados[i] <- contador
  }
  resultados # regresamos los resultados.
}
```

Observamos que en los incisos siguientes se ocupa esta funcipon para N un poco grandes, por lo que vectorizo la función anterior para tener lo mismo en un tiempo más corto. La diferencia entre estas dos funciones radica basicamente en el *sample*, ya que nosotros simularemos por bloques, es decir, como si estuvieramos muchas modenas lanzandose al mismo tiempo.

```
moneda_geometrica_optimizada <- function(p, N, potencia){ # potencia: tamaño del bloques.
  resultados <- c() # Inicializamos un vector.
  while(length(resultados)<N) { # Repetimos el hasta tener N resultados
    contador <- 0 # Inicializamos el número de lanzamientos.
    resultados_preliminar <- c() # Inicializamos los resultados por bloques.
    while(length(resultados_preliminar)<potencia){
      contador_s<- sum(sample(x=c(1,0), size=potencia-length(resultados_preliminar),
                              prob=c(p,1-p), replace=TRUE))
      contador <- contador + 1
     resultados_preliminar<- c(resultados_preliminar, rep(contador, contador_s)) # Concatenamos l
    }
    resultados <- c(resultados, resultados_preliminar) # Concatenamos los resultados por bloques
  }
  resultados # regresamos los resultados.
}
```

Donde el parámetro potencia representa el tamaño del bloque, es decir, cuantas monedas se lanzarán al mismo tiempo. Algo curioso de este parámetro por intución entre más grande sea más rápido será, pero no es así aunque no estoy muy seguro por que sucede.

b) Usando la función anterior simule $N = 10^4$ veces una variable aleatoria Geom(p) para $p = 0.5, 0.1, 0.01$. Grafique las frecuencias normalizadas en color azul. Sobre está última figura empalme en rojo la gráfica de la función de masa correspondiente. ¿Qué observa?

RESPUESTA

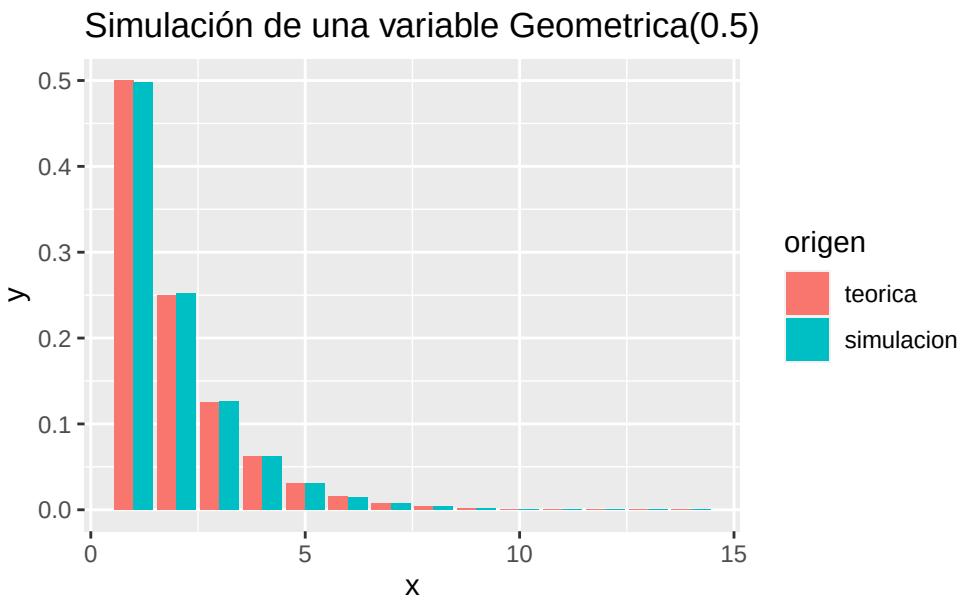
Creemos otra función que utilice la función del inciso a) y que grafique las frecuencias normalizadas en azul y en rojo las frecuencias obtenidas de función de distribución de un variable Geometrica.

```
library(tidyverse) # ggplot and dplyr
geometric_graph_simula_and_teoric <- function(p, N, potencia, titulo, estadisticos=0){
  # Utilizamos la opción del inciso a).
  simular_geometrica <- data.frame(resultado=moneda_geometrica_optimizada(p, N, potencia))
  if(estadisticos==1){
```

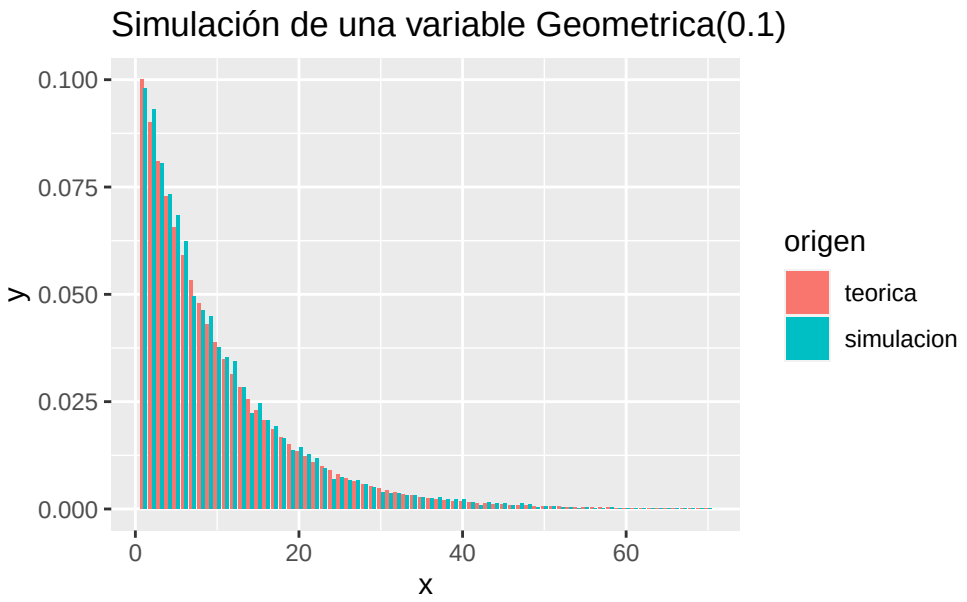
```
    print("La media de las simulaciones es:")
    print(mean(simular_geometrica$resultado))
        print("La desviación estandar de las simulaciones es:")
        print(sqrt(var(simular_geometrica$resultado)))
  }
  # Generamos las frecuenciass normalizadas.
  simular_geometrica <- data.frame(table(simular_geometrica)/N)
  names(simular_geometrica) <- c("x", "y")
  simular_geometrica$x <- as.numeric(simular_geometrica$x)
  # Variable auxiliar.
  simular_geometrica$origen <- "simulacion"
  max_resul <- max(simular_geometrica$x)
  # Función de distribución utilizando la formula.
  teoric_geometrica <- data.frame(x=seq(1,max_resul,1),
                                  y=dgeom(x=seq(0,(max_resul-1),1),1),

  # Concatenamos las frecuencias obtenidas.
  geometrica <- rbind(teoric_geometrica, simular_geometrica)
  # Graficamos
  g <- ggplot(geometrica, mapping=aes(x,y,fill=origen))+
    geom_histogram(position="dodge", stat="identity", bins = max_resul)+
    labs(title=titulo)
  return(g)
}
```

Por lo que las gráficas variando el parámetro p son

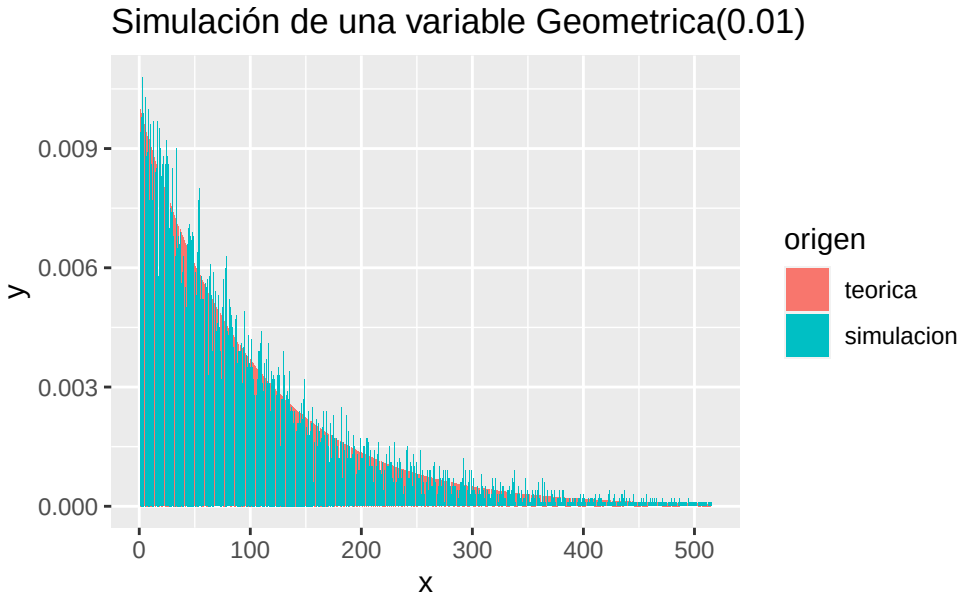
```
set.seed(08081997)
geometric_graph_simula_and_teoric(0.5, 10^4, 10^4,
                                  "Simulación de una variable Geometrica(0.5)")
```




```
geometric_graph_simula_and_teoric(0.1, 10^4, 10^4,
                                  "Simulación de una variable Geometrica(0.1)")
```




```
geometric_graph_simula_and_teoric(0.01, 10^4, 10^4,
                                  "Simulación de una variable Geometrica(0.01)")
```

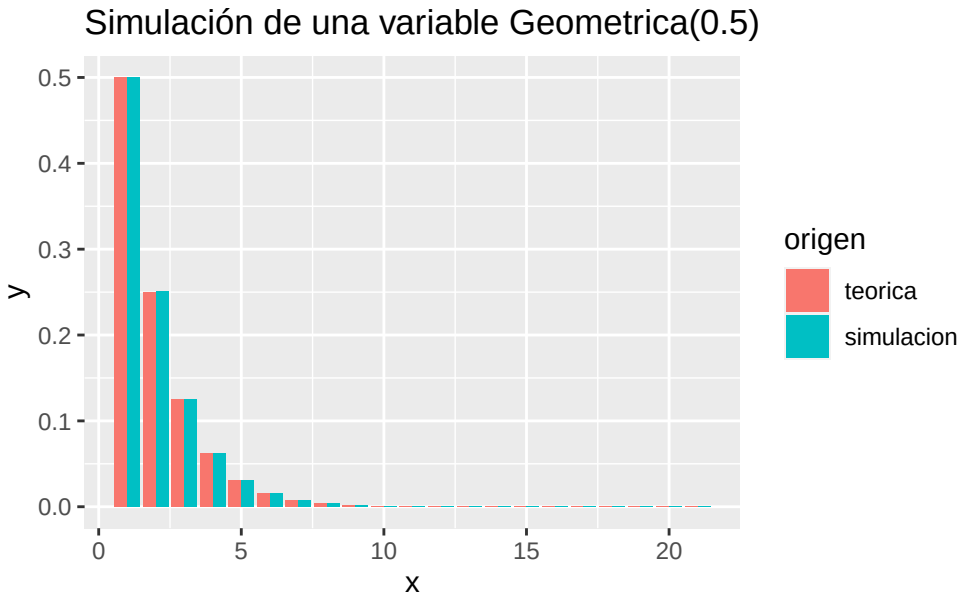


Observemos que si comparamos las frecuencias de las simulaciones y las frecuencias obtenidas de la función de probabilidad de una geometrica se ven muy cercanas. Pero conforme p se acerca a 0 la comparaciones entre estas frecuencias son más notorias. Esto se puede explicar debido a que cuando p es más chico la $\mathbb{P}(X = x)$ se va hacieno más pequeña, por lo que x toma un rango más amplio de valores posibles. No hay que confundirse por el hecho de que como la función de distribución de una variable aleatoria geometrica esta defina en todos los naturales. Ya que si p es cercano a 1, las probabilidades convergen más rapido a 0, y viceversa, si p es cercano a 0 las probabilidad convergen más lento a 0.

c) Repita el inciso anterior para $N = 10^6$. Además calcule el promedio y la desviación estándar de las simulaciones que realizó ¿Qué observa?

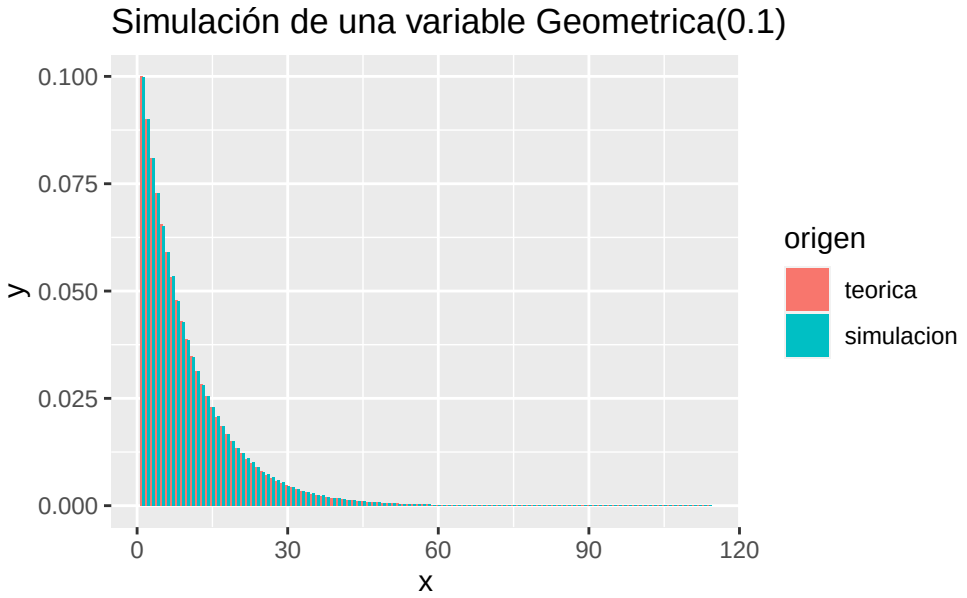
```
set.seed(08081997)
geometric_graph_simula_and_teoric(0.5, 10^6, 10^5,
"Simulación de una variable Geometrica(0.5)",1)
```

```
## [1] "La media de las simulaciones es:"
## [1] 1.998708
## [1] "La desviación estandar de las simulaciones es:"
## [1] 1.41284
```




```
geometric_graph_simula_and_teoric(0.1, 10^6, 10^5,
"Simulación de una variable Geometrica(0.1)",1)
```

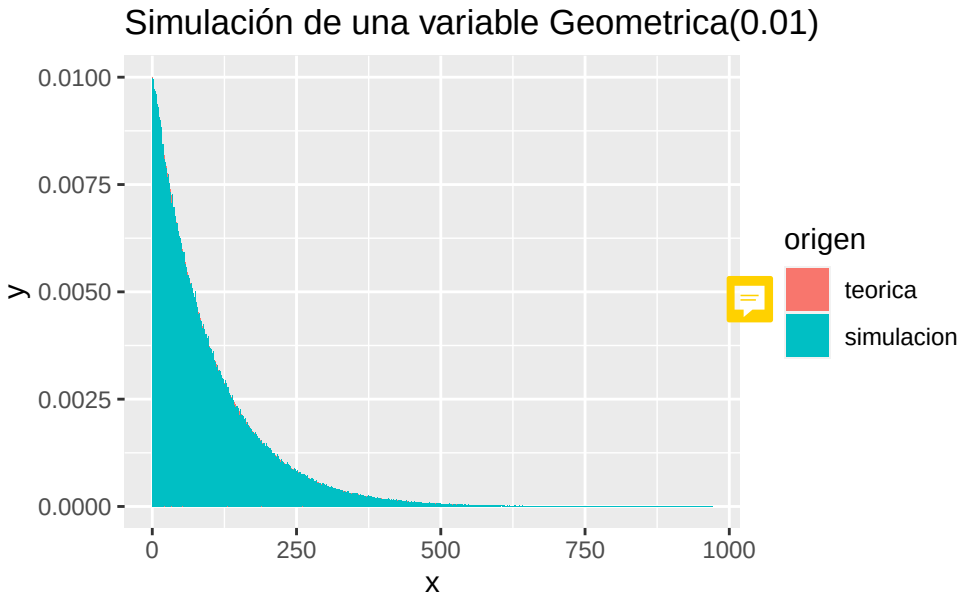
```
## [1] "La media de las simulaciones es:"
## [1] 10.01586
## [1] "La desviación estandar de las simulaciones es:"
## [1] 9.495429
```




```
geometric_graph_simula_and_teoric(0.01, 10^6, 10^5,
"Simulación de una variable Geometrica(0.01)",1)
```

```
## [1] "La media de las simulaciones es:"
## [1] 99.97975
```

```
## [1] "La desviación estandar de las simulaciones es:"
## [1] 99.41808
```



Cómo el número de simulaciones son mayores que el inciso anterior, observamos que las diferencias se entre las frecuencias simuladas y frecuencias calculados son muy cercanas "casi nulas". Y esto incita a concluir que la distribución Geometrica modela bien este esperimento de lanzamiento de monedas. Ahora analizando los promedios y desviaciones de las contra la experanza de X para cada P:

■.

4. Usando las ideas del inciso anterior escriba una función en R que simule N veces los lanzamientos de moneda hasta obtener r águilas. La función deberá recibir como parámetros a la probabilidad p de obtener águila, al número r de águilas a observar antes de detener el experimento y al número N de veces que se repite el experimento; y tendrá que regresar un vector de longitud N que contenga el número de lanzamientos hasta obtener las r águilas en cada uno de los N experimentos. Grafique las frecuencias normalizadas de los experimentos para $N = 10^6$, $p = 0.2, 0.1$ y $r = 2, 7$ y compárelos contra la función de masa de la distribución más adecuada para modelar este tipo de experimentos.

RESPUESTA

Sea X el número de lanzamientos hasta obtener r aguilas. Esto implica que $X \sim BN(r, p)$, donde p es la probabilidad de obtener águila en un lanzamiento. Entonces la función que simula este experimento sería:

```r
moneda_nbinom <- function(r, p, N){
  resultados <- c()
  for(i in 1:N){
    contador <- 0
    lanzamiento <- ""
    num_aguilas <- 0
    while(num_aguilas<r){
      lanzamiento <- sample(x=c("aguila", "sol"), size=1, prob=c(p,1-p))
      contador <- contador + 1
      if(lanzamiento=="aguila"){
        num_aguilas<-num_aguilas+1
      }
    }
```

```
    resultados[i] <- contador
  }
  resultados
}
```

La función anterior tiene un problema, ya que es muy lenta. Por lo que se vectorizo para tener un mejor rendimiento.

```
moneda_nbinom_optimizada <- function(r, p, N, potencia){
  resultados <- c()
  while(length(resultados)<N) { # Repetimos el experimentos N veces.
    contador <- 0 # Inicializamos el número de lanzamientos.
    resultados_preliminar <- c()
    inicial <- rep(0, potencia)
    while(length(resultados_preliminar)<potencia){ # si ya se obtuvo águila deterner.
      inicial <- inicial + sample(x=c(1,0), size=potencia-length(resultados_preliminar),
                                  prob=c(p,1-p), replace=TRUE)
      contador_s <- sum(inicial==r)
      contador <- contador + 1
     resultados_preliminar<- c(resultados_preliminar, rep(contador, contador_s))
     inicial <- inicial[inicial<r]
    }
    resultados <- c(resultados, resultados_preliminar)
  }
  resultados # regresamos los resultados.
}
```

Ahora modificamos la función del problema 3 para adaptarla a este problema,

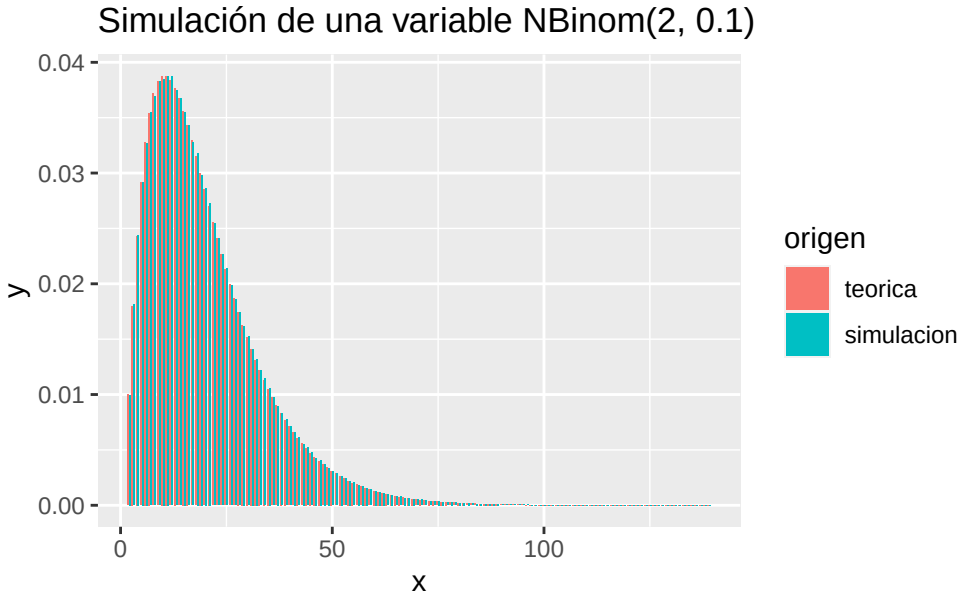
```
bimneg_graph_simula_and_teoric <- function(r, p, N, potencia, titulo, estadisticos=0){
  # Utilizamos la opción del inciso a).
  simular_geometrica <- data.frame(resultado=moneda_nbinom_optimizada(r, p, N, potencia))
  if(estadisticos==1){
    print("La media de las simulaciones es:")
    print(mean(simular_geometrica$resultado))
        print("La desviación estandar de las simulaciones es:")
        print(sqrt(var(simular_geometrica$resultado)))
  }
  # Generamos las frecuenciass normalizadas.
  simular_geometrica <- data.frame(table(simular_geometrica)/N)
  names(simular_geometrica) <- c("x", "y")
  simular_geometrica$x <- as.numeric(simular_geometrica$x)+r-1
  # Variable auxiliar.
  simular_geometrica$origen <- "simulacion"
  max_resul <- max(simular_geometrica$x)
  # Función de distribución utilizando la formula.
  teoric_geometrica <- data.frame(x=seq(r,max_resul,1),
                                  y=dnbinom(x=seq(0,(max_resul-r),1), size=r,prob = p), origen=re

  # Concatenamos las frecuencias obtenidas.
  geometrica <- rbind(teoric_geometrica, simular_geometrica)
```

```
  # Graficamos
  g <- ggplot(geometrica, mapping=aes(x,y,fill=origen))+
    geom_histogram(position="dodge", stat="identity", bins = max_resul)+
    labs(title=titulo)
  return(g)
}
```

Por lo que las gráficas variando el parámetro p y r son:

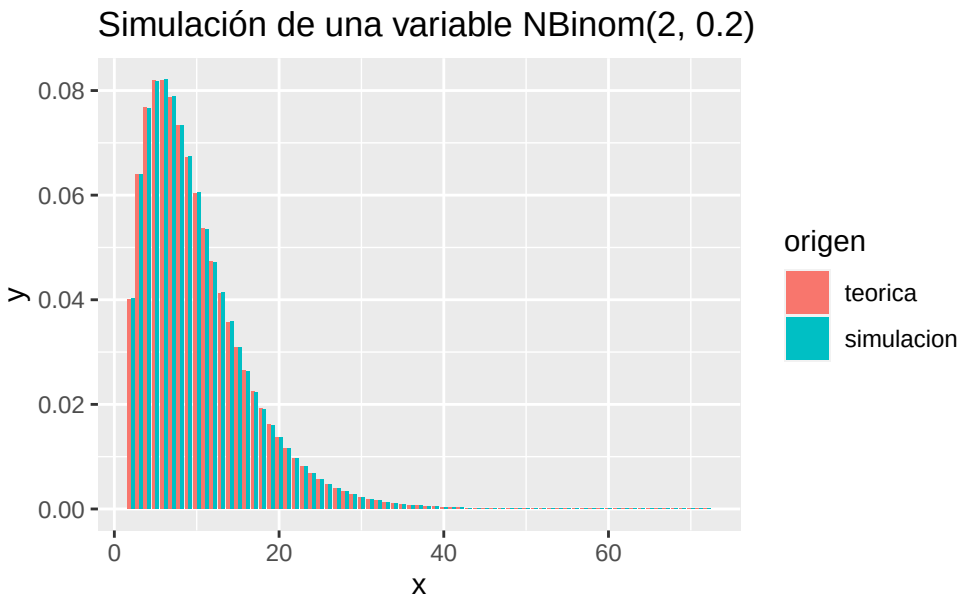
```
set.seed(08081997)
bimneg_graph_simula_and_teoric(2, 0.1, 10^6, 10^5,
                               "Simulación de una variable NBinom(2, 0.1)", 1)
```

```
## [1] "La media de las simulaciones es:"
## [1] 20.00012
## [1] "La desviación estandar de las simulaciones es:"
## [1] 13.4112
```



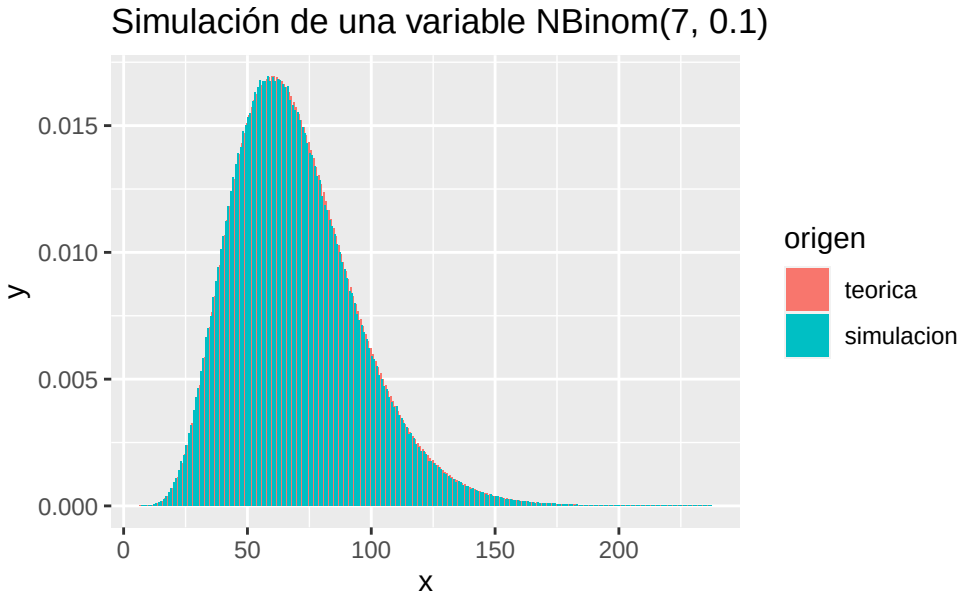

```
bimneg_graph_simula_and_teoric(2, 0.2, 10^6, 10^5,
                               "Simulación de una variable NBinom(2, 0.2)", 1)
```

```
## [1] "La media de las simulaciones es:"
## [1] 10.00252
## [1] "La desviación estandar de las simulaciones es:"
## [1] 6.338914
```



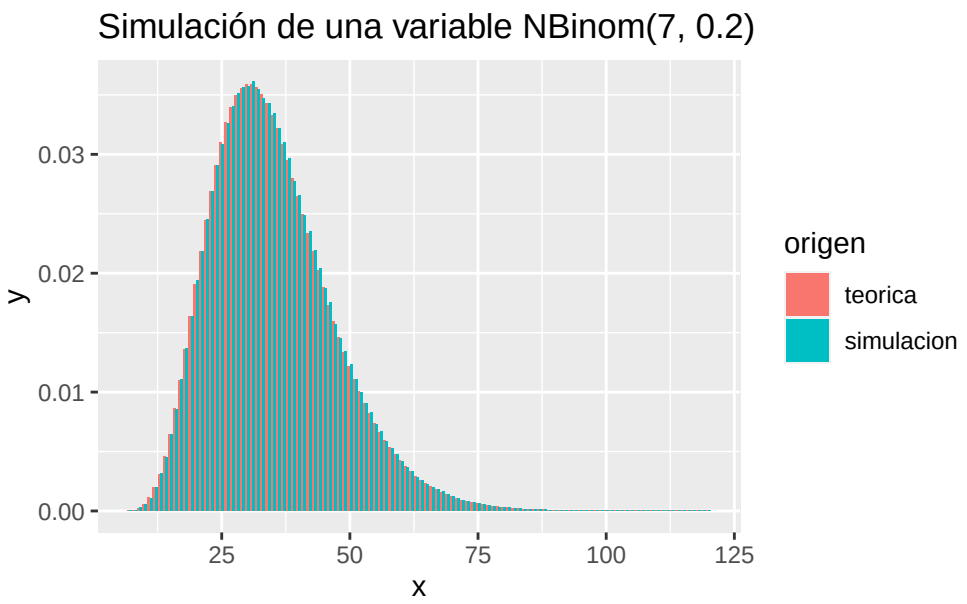

```
bimneg_graph_simula_and_teoric(7, 0.1, 10^6, 10^5,
                               "Simulación de una variable NBinom(7, 0.1)", 1)
```

```
## [1] "La media de las simulaciones es:"
## [1] 70.00665
## [1] "La desviación estandar de las simulaciones es:"
## [1] 25.09928
```




```
bimneg_graph_simula_and_teoric(7, 0.2, 10^6, 10^5,
                               "Simulación de una variable NBinom(7, 0.2)", 1)
```

```
## [1] "La media de las simulaciones es:"
## [1] 34.98512
## [1] "La desviación estandar de las simulaciones es:"
## [1] 11.8182
```



Podemos concluir que la distribución Binomial Negativa ajusta muy bien este experimento. Observando el promedio y la desviación estandar calculados, si las comparamos con la esperanza y desviación de la distribución Binomial Negativa.

$$E[X] = \frac{r}{p} \quad \sigma = \sqrt{Var[X]} = \sqrt{\frac{r(1-p)}{p^2}}$$

Comparando cada una de estas observamos que son muy parecidas, es decir, podemos dicer que son buenos estadisticos para estimar la esperanza y la desviación estandar. ■.

5. Considera X una v.a. con función de distribución F y función de densidad f, y sea A un intervalo de la línea real $\mathbb{R}$. Definamos la función indicadora $1_A(x)$:

$$1_A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{en otro caso} \end{cases}$$

Sea $Y = 1_A(x)$. Encuentre una expresión para la distribución acumulada y el valor esperado de Y.
RESPUESTA
Para calcular la función de distribución de distribución acumulada primero calculemos la función distribución:

$$f(Y = y) = \mathbb{P}(1_A(x) = y) = \mathbb{P}(\{x : 1_A(x) = y\}).$$

$$f(y) = \begin{cases} \mathbb{P}(x \in A) & \text{para y=1} \\ \mathbb{P}(x \notin A) & \text{para y=0} \\ 0 & \text{en otro caso} \end{cases}$$

Utilizando como $A = \{\{x_1\}, \{x_2\}, ...\{x_n\}\}$, donde x_i representa los particiones del intervalor A. Por lo que la función de distribución acumulada sería:

$$F(y) = \begin{cases} \frac{\sum_1^i \mathbb{P}(\ x_i \in A)}{\sum_i^n \mathbb{P}(x_i \in A)} & x_i \\ 0 & \text{en otro caso} \end{cases}$$

El valor esperado es

$$\mathbb{E}[Y] = \sum yf(y) = 1 \cdot f(1) + 0 \cdot f(0) = 1 \cdot f(1) = \mathbb{P}(x \in A) \quad ■.$$

6. Las calificaciones de un estudiante de primer semestre en un examen de química se describen por la densidad de probabilidad

$$f_y(y) = 6y(1-y) \qquad 0 \leq y \leq 1,$$

donde y representa la proporción de preguntas que el estudiante contesta correctamente. Cualquier calificación menor a 0.4 es reprobatoria. Responda lo siguiente:

a) ¿Cuál es la probabilidad de que un estudiante repruebe?

b) Si 6 estudiantes toman el examen, ¿cuál es la probabilidad de exactamente 2 reprueben?

RESPUESTA

Por como esta definida la función de probabilidad podemos decir que Y es es una variable continua. Ahora, solo para comprobación veamos que realmente sea una función de probabilidad, para ello observemos que

$$\int_{-\infty}^{\infty} f_y(y) = \int_0^1 6y(1-y) = 3y^2 - 2y^3|_0^1 = 1.$$

Por lo tanto observamos que si es una función de probabilidad.

Entonces **la probabilidad de que un estudiante repruebe es**

$$f_y(Y < 0.4) = \int_0^{0.4} f_y(y) = \int_0^{0.4} 6y(1-y) = 3y^2 - 2y^3|_0^{0.4} = 0.352$$

Ahora, sea X el número de estudiantes de reprueban el examen de un conjunto de 6 estudiantes que realizaron el examen. Por definición podemos decir que $X \sim Bin(6, p)$ donde p es la probabilidad de reprobar, pero si consideramos que las calificaiones de los estudiantes se distribuye como la variable Y, entonces podemos concluir que $X \sim Bin(6, 0.352)$. Por lo tanto, **la probabilidad de que exactamente 2 estudiantes reprueben es** (usamos la función dbinom(x=4, size = 6, prob = 0.352):

$$\mathbb{P}(X = 2) = \binom{6}{2} 0.352^2 (1 - 0.352)^4 = 0.328907 \quad \blacksquare.$$

7. Escriba una función en R que simule una aproximación al proceso Poisson a partir de las 5 hipótesis que usamos en clase para construir tal proceso. Usando esta función, simule tres trayectorias de un proceso Poisson $\lambda = 2$ sobre el intervalo $[0, 10]$ y grafíquelas. Además simule 10^4 veces un proceso de Poisson N con $\lambda 1/2$ y hasta el tiempo $t = 1$. Haga un histograma de $N(1)$ en su simulación anterior y compare contra la distribución de Poisson correspondiente.

RESPUESTA

Utilizando las definiciones y el archivo compartido en clase:

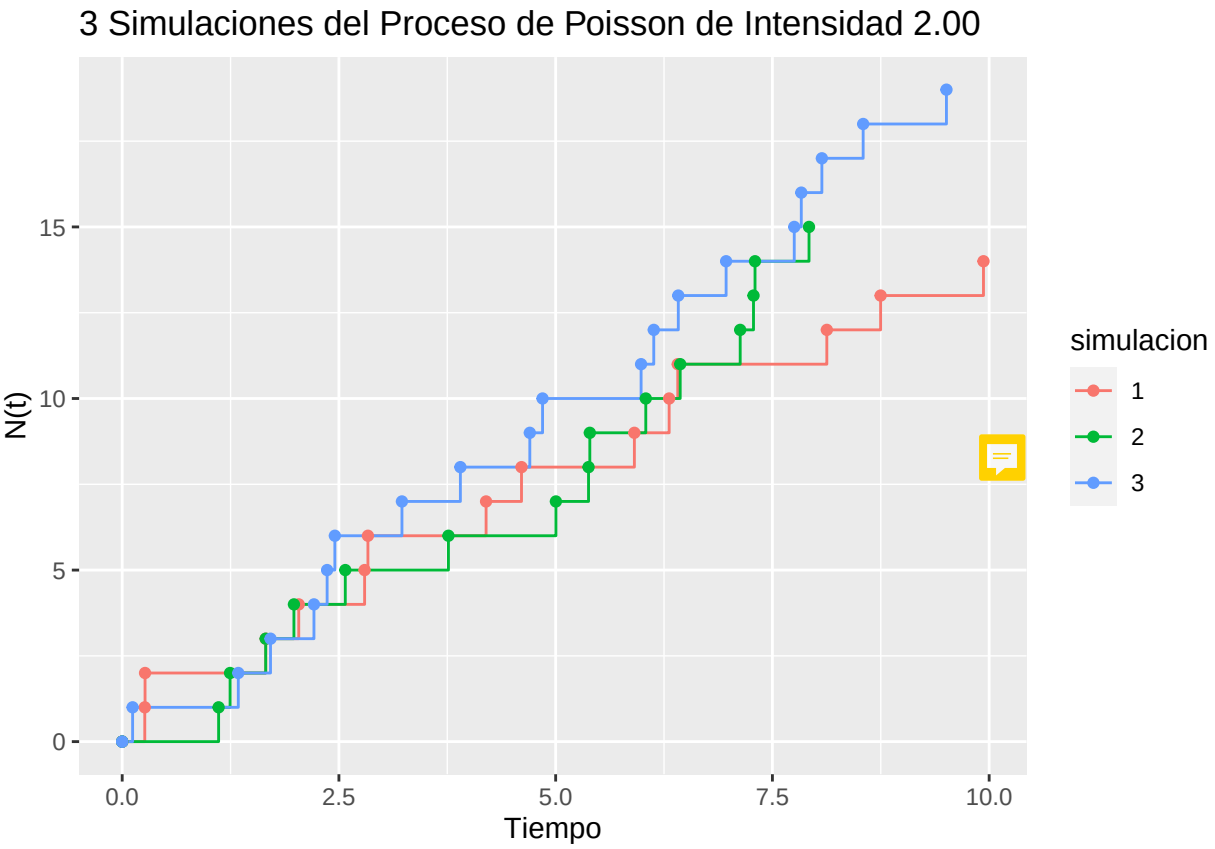
```
ProcesoPois<- function(t,lambda){
  N<- rpois(1,t*lambda) #Paso 1
  C<- sort(runif(N,0,t)) #Paso 2 y 3
  data.frame(x=c(0,0,C),y=c(0,0:N))
}
```

```
library(plyr)
NPois<-function(n,t,rate){
  C<- lapply(1:n, function(n)
    #Genera N dataframes con los procesos
    data.frame(ProcesoPois(t,rate),simulacion=n))
```

```
  C<-ldply(C, data.frame) # Une en una sola dataframe
  C$simulacion<-factor(C$simulacion) # Convierte en factores
  C
}
```

Gráficamos 3 simulaciones

```
simulacion_process_a <- NPois(3,10,2)

# Graficamos.
qplot(x,y,data=simulacion_process_a,geom=c("step","point"),color=simulacion,
      xlab="Tiempo",ylab="N(t)",main=sprintf("%d Simulaciones del Proceso de Poisson de Intensida
```

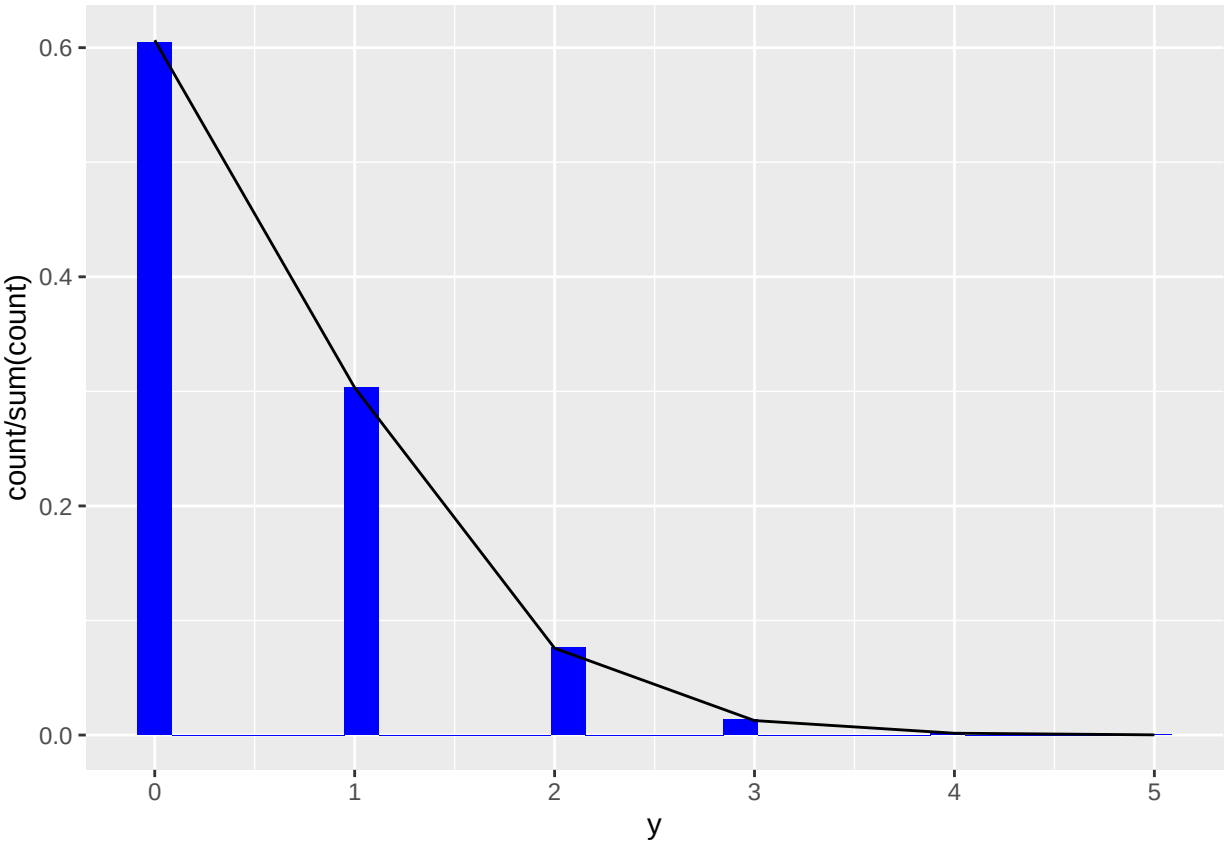


Graficamos la distribución Poission y el histograma de las simulaciones:

```
set.seed(13)
prueba <- NPois(10^4, 1,0.5)

prueba %>% group_by(simulacion) %>%
    top_n(1, y) %>% distinct(y, simulacion) %>%
  ggplot(aes(y))+geom_histogram(fill="blue", aes(y=stat(count)/sum(count)))+
  geom_line(data=data_frame(x=seq(0,5),y=dpois(x=seq(0,5),lambda = 0.5)), aes(x,y))
```

```
## Warning: `data_frame()` is deprecated, use `tibble()`.
## This warning is displayed once per session.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Observamos que las simulaciones realizadas es una forma de realizar un experimento con distribución Poisson.

8. En una oficina de correo los paquetes llegan según un proceso de Poisson de intensidad λ. Hay un costo de almacenamiento de c pesos por paquete y por unidad de tiempo. Los paquetes se acumulan en el local y se despachan en grupos cada T unidades de tiempo (es decir, se despachan en $T, 2T, 3T, \cdots$). Hay un costo por despacho fijo de K pesos (es decir, el costo es independiente del número de paquetes que se despachen). (a) ¿Cuál es el costo promedio por paquete por almacenamiento en el primer ciclo $[0, T]$? (b) ¿Cuál es el costo promedio por paquete por almacenamiento y despacho en el primer ciclo? (c) ¿Cuál es el valor de T que minimiza este costo promedio?

RESPUESTA

Sea X el número de paquetes que llegan al correo en un intervalo de tiempo T, este se distribuye como un proceso Poisson con intensidad λ. Entonces el costo total promedio por almacenamiento es:

$$\mathbb{E}[C] = \mathbb{E}[X \cdot c \cdot T] = cT\mathbb{E}[X] = cT \cdot (\lambda T) = cT^2\lambda.$$

Y ahora el número esperado de paquetes en el primer ciclo es:

$$\mathbb{E}[X] = \lambda T.$$

Por lo que, **(a) el costo promedio por paquete por almacenamiento es:**

$$\frac{cT^2\lambda}{\lambda T} = cT.$$

Ahora sea G el costo total de almacenamiento y despacho para el primer ciclo $[0, T]$ definido como

$$G = cXT + K.$$

Entonces el costo promedio total por almacenamiento y despacho es

$$\mathbb{E}[G] = \mathbb{E}[cXT + K] = cT^2\lambda + K.$$

Lo anterior implica que **el costo promedio por paquete por almacenamiento y despacho en el primer ciclo** es:

$$\bar{\mathbb{E}[G]} = \frac{cT^2\lambda + K}{\lambda T}.$$

Utilizando el resultado anterior, diferenciamos e igualamos a cero para encontrar el mínimo.

$$\bar{\mathbb{E}'[G]} = c - \frac{K}{\lambda T^2}$$

Igualamos a cero:

$$\begin{aligned} c - \tfrac{K}{\lambda T^2} &= 0 \\ T^2 &= \tfrac{K}{c\lambda} \\ T &= \sqrt{\tfrac{K}{c\lambda}}. \end{aligned}$$

Usando el criterio de segunda derivada para determinar si es un máximo o minino:

$$\bar{\mathbb{E}''[G]} = 2\frac{K}{\lambda T^3}.$$

Evaluando la segunda derivada en $T = \sqrt{\frac{K}{c\lambda}}$, observamos que $\bar{\mathbb{E}''[G]} > 0$, por lo que podemos concluir que es un mínimo. En conclusión, **el valor de T para el cuál minimiza el costo promedio por paquete por almacenamiento y despacho en el primer ciclo es** $\sqrt{\frac{K}{c\lambda}}$ ■.

9. Considere la siguiente función

$$F(x) = \begin{cases} 0 & \text{para } x < 0 \\ 0.1 & \text{para } x = 0 \\ 0.1 + 0.8x & \text{para } 0 < x < 3/4 \\ 1 & \text{para } 3/4 \leq x \end{cases}$$

¿Es una función de distribución? Si es una función de distribución, ¿corresponde a una variable aleatoria discreta o continua?

RESPUESTA

Observemos por como esta definida la función tenemos que $0 \leq F(x) \leq 1$. Y además $\lim_{x\to\infty^-} F(x) = 0$ y $\lim_{x\to\infty^+} F(x) = 1$. Por lo que podemos concluir que $F(x)$ si es una función de distribución. Ahora, observemos que la función esta definida para $x = 0$ y $x = 3/4$, si X fuera una fuera una variable continua, por definición $F(x = a) = 0$, por lo que X es discreta en $x = 0$ y $x = 3/4$. Y como $F(x)$ es continua en $0 < x < 3/4$ podemos decir que X es continua en ese intervalo. Entonces como X es continua y discreta para ciertos valores, decimos que X es "mixta". Esto igual se puede mostrar observando la grafica de la función $F(X)$ ■.

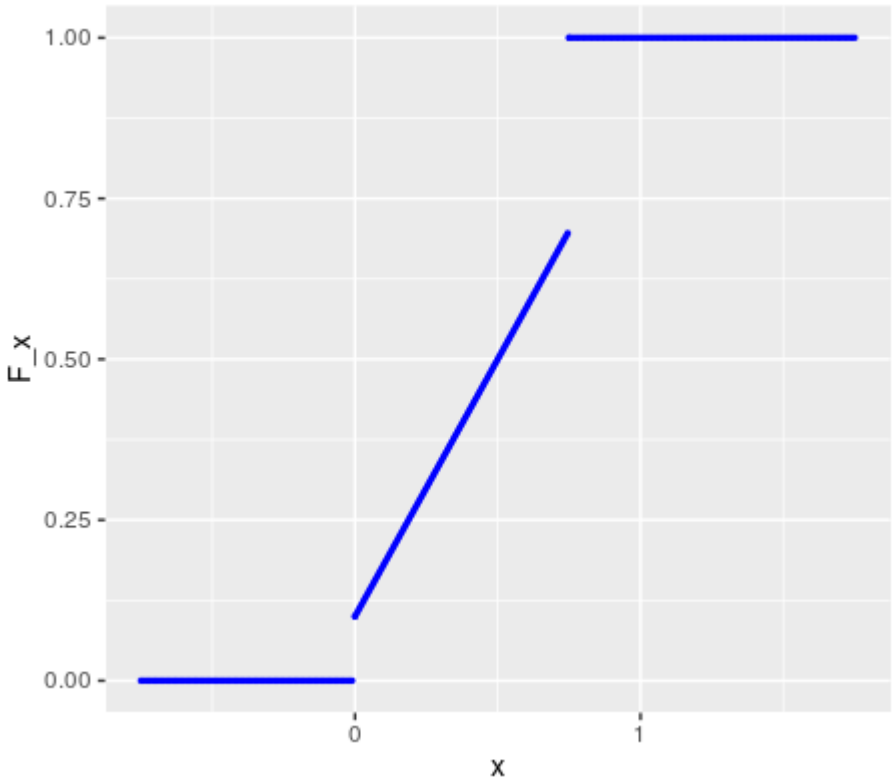


Figura 1: Función de densidad mixta.

10. **Este es un problema al que se recurrirá en el futuro,** su intención es que empiecen a jugar con datos reales. El archivo Delitos.csv contiene información sobre los delitos denunciados en la ciudad de Aguascalientes, para el período comprendido entre enero de 2011 a junio del 2016. Dicho archivo contiene 5 columnas: la primera columna contiene la fecha de denuncia del delito; la columna TIPO muestra una descripción del tipo de delito; la columna CONCATENAD presenta un descripción más amplia del delito; la columna SEMANA contiene la semana del año a la que corresponde la fecha de denuncia; y la columna SEMANA_COMPLETAS indica la semana a lo largo del estudio en la cual se presentó la denuncia. A través de métodos gráficos (e.g. boxplots) traten de determinar el comportamiento semanal de los delitos y discutan alternativas de modelos para describir los delitos cometidos en forma relativamente apropiada.

RESPUESTA

Realicemos un analisis exploratorio. Cargamos los datos:

```
# Cargamos las librerias a ocupar.
library(tidyverse)

# Leamos los datos.
df_delitos <- read_csv(file = "Delitos.csv")
```

Conozcamos un poco los datos. Nombre de las columnas y mostramos los primeros 5 registros.

```
names(df_delitos)
```

```
## [1] "FECHA"            "TIPO"             "CONCATENAD"       "SEMANA"
## [5] "SEMANA_COMPLETAS"
```

```r
head(df_delitos,3)
```

```
## # A tibble: 3 x 5
##   FECHA      TIPO      CONCATENAD                          SEMANA SEMANA_COMPLETAS
##   <date>     <chr>     <chr>                                <dbl>            <dbl>
## 1 2011-01-01 COMERCIAL COMERCIAL/EMPRESA/INDUSTRIA/FARD~      1                1
## 2 2011-01-04 COMERCIAL COMERCIAL/EMPRESA/INDUSTRIA/FARD~      1                1
## 3 2011-01-16 COMERCIAL COMERCIAL/EMPRESA/INDUSTRIA/FARD~      3                3
```

Exploramos el tipo de cada variable.

```r
str(df_delitos)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 44212 obs. of  5 variables:
##  $ FECHA           : Date, format: "2011-01-01" "2011-01-04" ...
##  $ TIPO            : chr  "COMERCIAL" "COMERCIAL" "COMERCIAL" "COMERCIAL" ...
##  $ CONCATENAD      : chr  "COMERCIAL/EMPRESA/INDUSTRIA/FARDERO" "COMERCIAL/EMPRESA/INDUSTRIA/F
##  $ SEMANA          : num  1 1 3 3 3 4 4 4 5 6 ...
##  $ SEMANA_COMPLETAS: num  1 1 3 3 3 4 4 4 5 6 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   FECHA = col_date(format = ""),
##   ..   TIPO = col_character(),
##   ..   CONCATENAD = col_character(),
##   ..   SEMANA = col_double(),
##   ..   SEMANA_COMPLETAS = col_double()
##   .. )
```

Imprimimos los valores unicos, observamos que existen 23 diferentes.

```r
unique(df_delitos$TIPO)
```

```
##  [1] "COMERCIAL"                          "TRANSEUNTE"
##  [3] "CRISTAL"                            "MOTOCICLETA"
##  [5] "VEHICULO"                           "TRANSEUNTE EN VEHICULO"
##  [7] "BICICLETA"                          "TRANSPORTE DE PASAJEROS CIUDAD"
##  [9] "DOMICILIARIO"                       "INSTITUCIONES PUBLICAS"
## [11] "INSTITUCION POLITICA"               "REMOLQUE/PLATAFORMA"
## [13] "INSTITUCION FINANCIERA"             "OTRO"
## [15] "TARJETA BANCARIA/COMERCIAL"         "TRANSPORTE DE CARGA CIUDAD"
## [17] "MAQUINARIA PESADA"                  "TRANSPORTE DE CARGA CARRETERA"
## [19] "GANADO"                             "INSTITUCION BANCARIA"
## [21] "TRANSPORTE DE PASAJEROS CARRETERA" "No Capturado"
## [23] "TRACTOR AGRICOLA"
```

Contamos los

```r
library(dplyr)
df_delitos %>% group_by(TIPO) %>%
  tally() %>% arrange(desc(n)) %>% head()
```

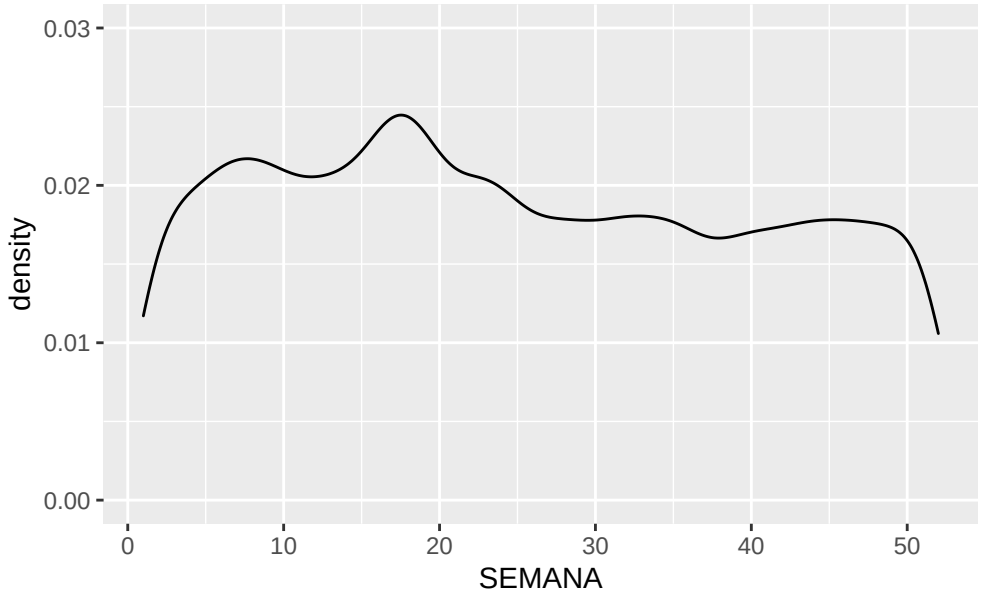
```
## # A tibble: 6 x 2
##   TIPO                      n
```

```
##   <chr>                  <int>
## 1 DOMICILIARIO           11295
## 2 VEHICULO                8336
## 3 TRANSEUNTE              8178
## 4 COMERCIAL               6179
## 5 CRISTAL                 6013
## 6 INSTITUCIONES PUBLICAS  1440
```

Esto puede deberse a que no todos los delitos se reportan, lo que puede indica que exista un sesgo en los delitos. Puede deberse a que en México no reportan cuando el delito no es muy grave como el robo de una bicicleta, o incluso a que algunos delitos son más complicados que otros como robar un tractor agricola.

La distribución de los delitos reportados durante todo el año se ve de la siguiente forma:

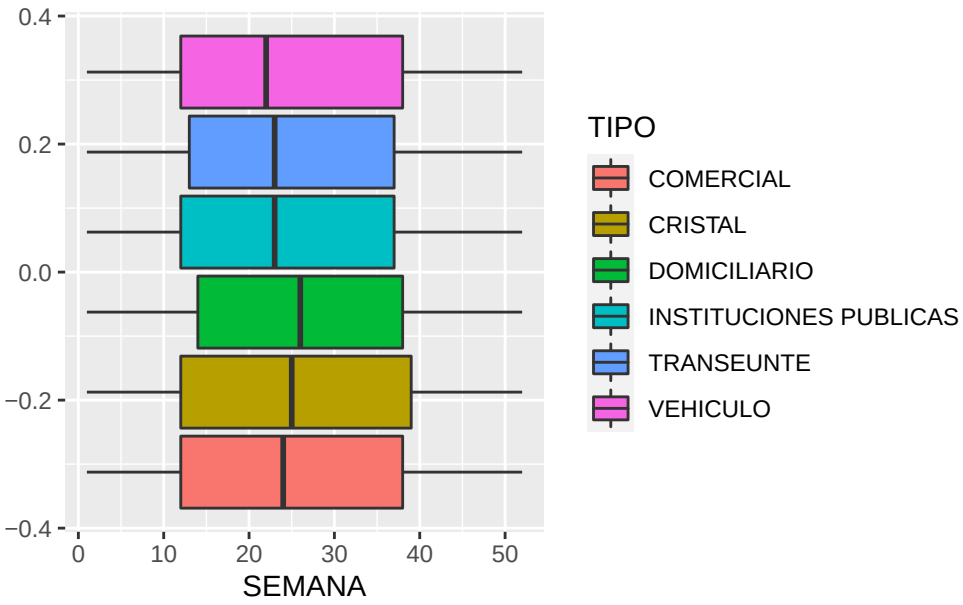
```
ggplot(data=df_delitos, aes(x=SEMANA))+
  geom_density()+
  ylim(c(0,0.030))
```



Se observamos que en las semanas 15-19 existe un pico notorio con respecto al resto de las semanas, esto se puede deber a que existe por esas fechas es semana santa, y que exista más vulnabilidad para cometer delitos.

Ahora como los no tienen la misma posiilidad de comenterse, observemos como se ven las semanas de los delitos top 6 con más registros esto para ser más entendible las gráficas:

```
top_5sup <- c(df_delitos %>% group_by(TIPO) %>%
  tally() %>% arrange(desc(n)) %>% head() %>%select(TIPO))

df_delitos_top <- df_delitos %>% filter(TIPO %in% top_5sup$TIPO)

ggplot(data=df_delitos_top , aes(x=SEMANA, fill=TIPO))+
         geom_boxplot()
```



Observamos que los delitos estan centrados en las semanas 20-30, lo que implica que puede estar relacionado con las vacacioens de verano.

Ahora una forma de modelar el número delitos que ocurrirán en una semana del semana es considerando un proceso poisson. Como observamos la intensidad no es la misma en cada semana del año, por lo que puede ser que un proceso poisson no homogeneo sería una buena alternativa. Aclaro que cada delito tiene que tener su proceso poisson debido a que existe una diferencia entre la intensidad de los delitos. ■.

Ejercicios de las notas.

- Distribución uniforme continua.

I) Crea una columna con 100 valores de una $Unif(0,1)$ en el software de tu preferencia.

RESPUESTA
Ocupando el software estadístico R:

```
set.seed(080801997)
y <- runif(100, 0, 1)
head(y) # Mostramos algunos valores.
```

```
## [1] 0.02946391 0.58192044 0.78700956 0.23089343 0.11966928 0.84819126
```

II) Construye otra columna con la fórmula $x = -2\log(1-y)$

RESPUESTA
Creamos la variable y con la formula:

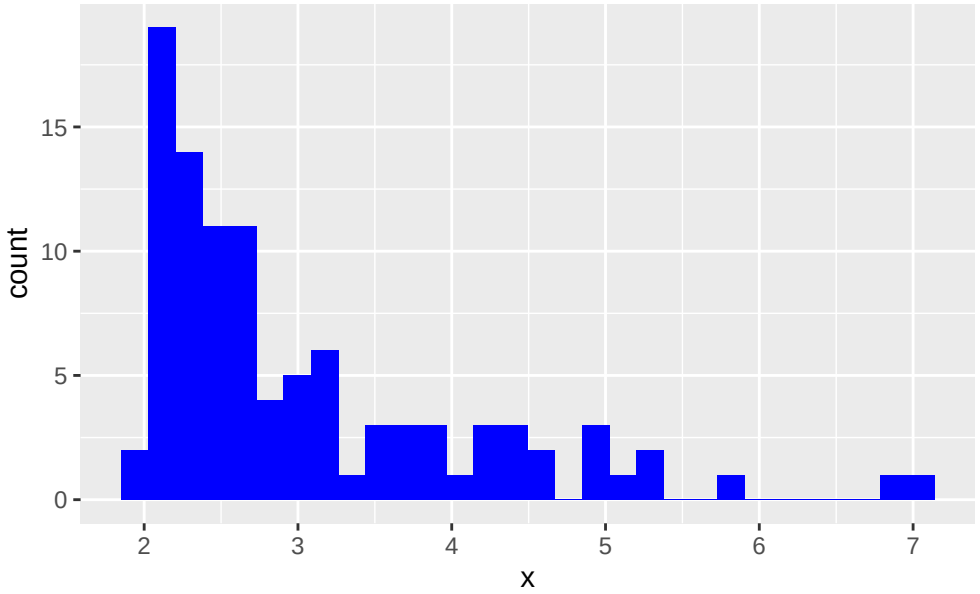
```
x <- 2-log(1-y)
head(x) # Mostramos algunos valores.
```

```
## [1] 2.029907 2.872084 3.546508 2.262526 2.127458 3.885134
```

III) Construye el histograma de esta nueva columna y concluye.

RESPUESTA
Utilizando la librería R:

```
library(tidyverse)

ggplot(data=data_frame(x=x), aes(x))+
  geom_histogram(fill="blue")
```



Por como se construyo la variable x si despejamos la expresión del inciso anterior, obtenemos que

$$y = 1 - e^{-1/2x}.$$

Por lo que podemos concluir que x se distribuyen como una distribución Exponencial y esto mismo se observa cuando se observa el histograma del inciso anterior. ■.

- El modelo Normal o Gaussiano. Se han realizado ciertas pruebas de resistencia en ladrillos obteniéndose las mediciones que a continuación se muestran, agrupadas en una tabla de frecuencias.

Interv.	De	Hasta	Frecuencia	Frec. relativa
1	28.70	32.65	5	5.56
2	32.65	36.60	6	6.67
3	36.60	40.55	11	12.22
4	40.55	44.50	17	18.89
5	44.50	48.45	19	21.11
6	48.45	52.40	19	21.11
7	52.40	56.35	7	7.78
8	56.35	60.30	2	2.22
9	60.30	64.25	3	3.33
10	64.25	68.20	1	1.11

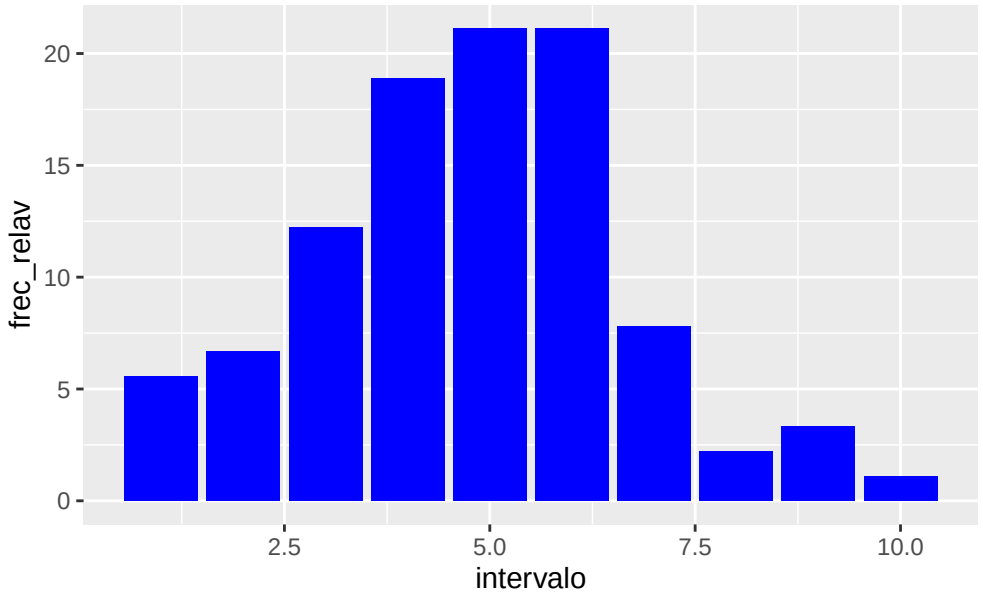
a) Traza el histograma. **RESPUESTA**
El histograma es:

```
df_ladrillo <- data_frame(intervalo=seq(1,10,1),
                          de=seq(28.70, 64.25, 3.95),
                          hasta=seq(32.65, 68.20, 3.95),
                          frec=c(5, 6, 11, 17, 19, 19, 7, 2, 3, 1))
```

```
df_ladrillo <- df_ladrillo %>%
  mutate(frec_relav=round(frec/sum(frec)*100,2))
```

Graficamos

```
ggplot(data=df_ladrillo, aes(intervalo, frec_relav))+
  geom_histogram(stat="identity",bins = 10,fill="blue")
```



b) Calcula la probabilidad de cada intervalo de clase de la tabla de frecuencias asumiendo que las resistencias siguen una distribución normal con media 45.47 y varianza 58.19.

RESPUESTA

Cómo no se especifica que metodología usar:

```
sd_prob <- sqrt(58.19)
media_prob <- 45.47

df_ladrillo <- df_ladrillo %>%
  mutate(prob_a=pnorm(de,media_prob, sd_prob),
         prob_b=pnorm(hasta,media_prob, sd_prob),
         prob=(prob_b-prob_a)*100)
```

c) Compara las frecuencias relativas con las probabilidades bajo normalidad ¿Qué se puede concluir? Esta es la idea base de una prueba de Bondad de Ajuste conocida como χ^2 de Pearson. Si la media y la varianza de la normal no se conocen de antemano, podemos usar los valores muestrales correspondientes como una aproximación de los mismos. A esto lo llamamos estimación de parámetros y ya hablaremos más adelante de las cualidades de los estimadores.

Veamos las frecuencias del inciso anterior:

```
df_ladrillo[c(1,5,8)]
```

```
## # A tibble: 10 x 3
##    intervalo frec_relav   prob
##        <dbl>      <dbl>  <dbl>
##  1         1       5.56  3.25
```

```
##  2          2       6.67  7.60
##  3          3      12.2  13.7
##  4          4      18.9  19.0
##  5          5      21.1  20.3
##  6          6      21.1  16.6
##  7          7       7.78 10.5
##  8          8       2.22  5.10
##  9          9       3.33  1.90
## 10         10       1.11  0.547
```

Observamos que la diferencias entre las probabilidades son muy pequeñas, podría decir que tienen la mismas forma. ■.

- El modelo Weibull.

1. Gráficar la función Weibull para los siguientes valores de los parámetros:

$$\alpha = 1,\ \beta = 2$$
$$\alpha = 2,\ \beta = 2$$
$$\alpha = 3,\ \beta = 4$$

Algebraicamente, o bien generando variables con ésta distribución y construyendo un histograma.

RESPUESTA
Recordemos que la función de distribución de una Weibull es:

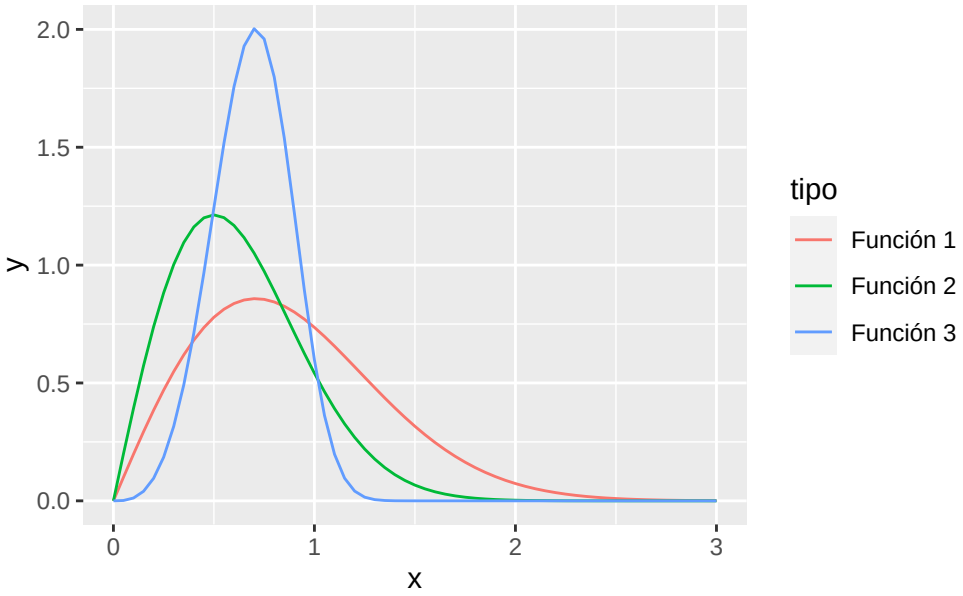
$$f(x) = \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}.$$

Para no tener problemas de versión de parametrización con R, definamos una función que calcule la expreción anterior

```
beta_clases <-function(x,alpha, beta){
  alpha*beta*(x)^{beta-1}*exp({-alpha*x^{beta}})
}
```

Ahora con la ayuda de la función anterior gráficamos la distribución de los distintos parámetros de la Weibull.

```
x <- seq(0,3, 0.05)
len_x<-length(x)
y_1 <- beta_clases(x, 1, 2)
y_2 <- beta_clases(x, 2, 2)
y_3 <- beta_clases(x, 3, 4)

df_weibull <- data.frame(x = rep(x,3), y =c(y_1, y_2, y_3),
                         tipo=c(rep("Función 1",len_x), rep( "Función 2",len_x),rep("Función 3",l

ggplot(data=df_weibull, aes(x,y,fill=tipo,color=tipo))+
  geom_line()
```



2. Gráfica las funciones de confiabilidad y riesgo para cada uno de los casos anteriores.

RESPUESTA

Las funciones de confiabilidad y riesgo de una función de distribución Weibull son:

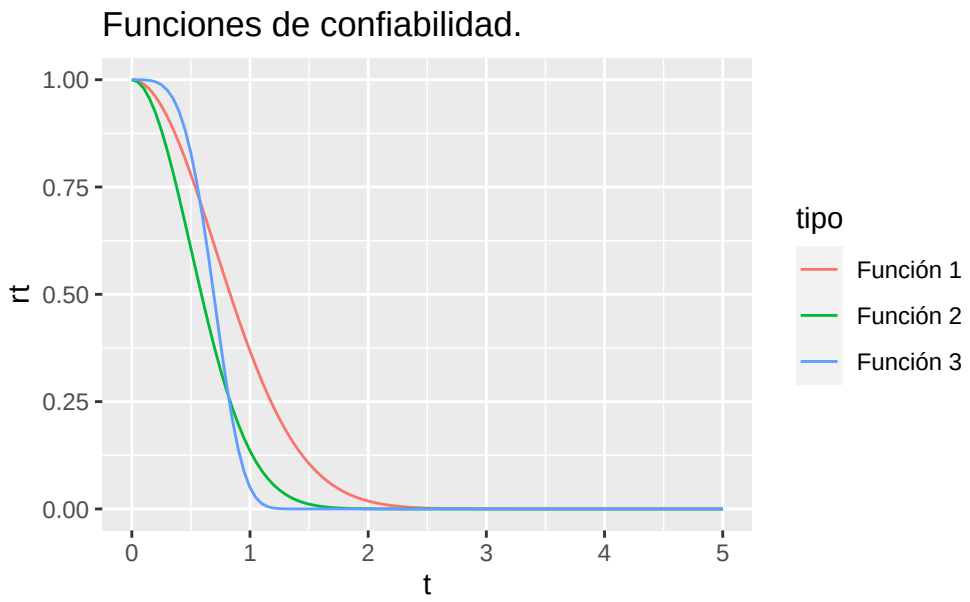
$$R(t) = e^{-\alpha t^{\beta}} \quad h(t) = \alpha\beta t^{\beta-1}$$

Creamos una función que avalué lo anterior:

```
confiabilidad_weibull <- function(t, alpha, beta){
  exp(-alpha*t^{beta})
}

riesgo_weibull <- function(t, alpha, beta){
  alpha*beta*t^{beta-1}
}
```

Ocupando las funciones anteriores generamos las gráficas:

```
t <- seq(0,5, 0.05)
len_t<-length(t)
rt_1 <- confiabilidad_weibull(t, 1, 2)
rt_2 <- confiabilidad_weibull(t, 2, 2)
rt_3 <- confiabilidad_weibull(t, 3, 4)

ht_1 <- riesgo_weibull(t, 1, 2)
ht_2 <- riesgo_weibull(t, 2, 2)
ht_3 <- riesgo_weibull(t, 3, 4)
df_weibull <- data.frame(t = rep(t,3), rt=c(rt_1, rt_2, rt_3),
                         ht=c(ht_1, ht_2, ht_3),
                         tipo=c(rep("Función 1",len_t), rep( "Función 2",len_t),rep("Función 3",l

ggplot(data=df_weibull, aes(t,rt,fill=tipo,color=tipo))+
  geom_line()+
  labs(title = "Funciones de confiabilidad.")
```




```
ggplot(data=df_weibull, aes(t,ht,fill=tipo,color=tipo))+
  geom_line() +
  labs(title = "Funciones de riesgo.")
```

3. Si $h(t) = a + bt$, ¿cuál es la densidad asociada?

RESPUESTA

Cómo $f(t) = h(t) \cdot R(t)$, calculemos primero $R(t)$. Ocupando la siguente relación entre la función de riesgo y confiabilidad:

$$R(t) = Ce^{\int h(t)dt}$$

$$R(t) = Ce^{at+b/2t^2}$$

Evaluemos en $t = 0$ para encontrar C,

$$R(0) = Ce^0 = 1.$$

Por lo tanto, la densidad asociada es:

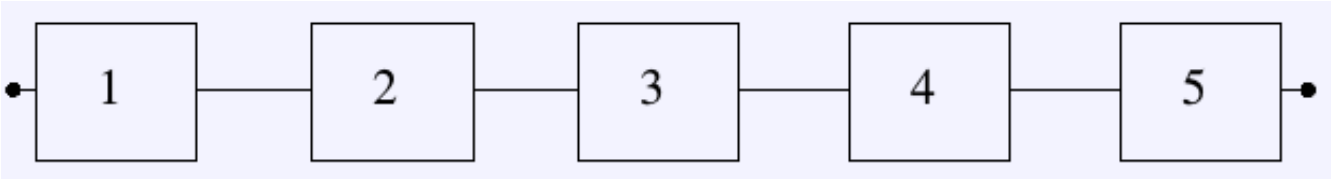
$$f(t) = (a + bt)e^{at+b/2t^2}.$$

4. Estudiar las funciones de riesgo para la distribución Normal, Gamma y Lognormal. ¿Cuál es la principal dificultad en estos casos?

RESPUESTA
Cada una de estas función si observamos su función de probabilidad acumulada, estas no poseen como tal una expresión para calcularlas para algún x. Por lo que, como la función de confiabilidad es $R(x) = 1 - F(x)$ para estas funciones no existe una "igualdad". Y por lo tanto como la función de riesgo es $h(t) = \frac{f(t)}{R(t)}$, como depende de $R(t)$ por lo que el manejo de $h(t)$ se complica debido a que no existe una forma sencilla de calcularla. Es decir, en la distribución Normal, la función acumulada es una integral que no se puede resolver si no esta definida de igual modo que la distribución LogNormal. Y para el caso de la distribución Gamma obtenemos una serie la cual es algo complicado de que converja a algo. ■.

- El modelo exponencial.

Consideremos un sistema formado por 5 componentes idénticos conectados en serie tal como se muestra a continuación:



Tan pronto como un componente falla, el sistema completa falla. Supongamos que cada componente sigue un modelo de tiempo a la falla exponencial con $\theta = 100$, y que los componentes fallan en forma independiente una de la otra. Definamos los eventos

$$A_i = \{i - \text{ésimo componente dura al menos t horas}\}\ i = 1, 2, 3, 4, 5.$$

Las $A_i's$ son independientes e identicamente distriuidas. Sea X el tiempo de la falla del sistema.

a) El evento $\{X \geq t\}$, ¿a cuál evento, en términos de las $A_i's$ es equivalente?

RESPUESTA
Por como se definieron los eventos de cada componente, tenemos que decir que el sistema fallé equivale a que algún componente fallé, lo que se puede escribir como:

$$\{X \geq t\} = \{A_1 \cap\ A_2 \cap A_3 \cap A_4 \cap A_5\}.$$

b) Usando la independencia de las $A_i's$ calcula $\mathbb{P}\{X \geq t\}$. Obtén $F(t) = \mathbb{P}(X \leq t)$. ¿Cuál es la distribución de X?

RESPUESTA
Ocupando la relación anterior y además como cada componente es independiente ($\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \cap \mathbb{P}(A_j)$):

$$\begin{aligned} \mathbb{P}\{X \geq t\} &= \mathbb{P}\{A_1 \cap\ A_2 \cap A_3 \cap A_4 \cap A_5\} \\ &= \mathbb{P}\{A_1\}\mathbb{P}\{A_2\}\mathbb{P}\{A_3\}\mathbb{P}\{A_4\}\mathbb{P}\{A_5\} \end{aligned}$$

Ahora considerando que cada componente sigue un modelo de tiempo a la falla exponencial con $\theta = 100$, podemos ver que $\mathbb{P}(A_i) = F(y) = \mathbb{P}(Y \leq y)$ donde $Y \sim Exp(1/100)$. Ahora sustituyendo en lo anterior:

$$= [1-(1-e^{-\frac{t}{100}})][1-(1-e^{-\frac{t}{100}})][1-(1-e^{-\frac{t}{100}})][1-(1-e^{-\frac{t}{100}})][1-(1-e^{-\frac{t}{100}})]$$

$$= [e^{-\frac{5t}{100}})]$$

Por lo que

$$F(t) = 1 - e^{-500t}.$$

Lo anterior implica que X tiene una distribución Exp(5/100).

c) Si en lugar de 5 componentes tenemos n, ¿cuál es la distribución de X?

RESPUESTA
Por inducción podemos probar que cuando existe n componentes

$$\mathbb{P}(X \geq t) = \prod_{i=1}^{n} \mathbb{P}\{A_i\}$$

Paso 1. Probamos para algún n. Considerando el inciso b) se cumple.
Paso 2. Suponemos que se cumple para n componentes.
Paso 3. Demostramos para $n+1$ componentes. Como existe independencia entre cualquier A_i, tenemos entonces que:

$$\mathbb{P}(X \geq t) = \bigcap_{i=1}^{n} \mathbb{P}\{A_i\} \cap A_{n+1} = \prod_{i=1}^{n} \mathbb{P}\{A_i\} \cdot A_{n+1} = \prod_{i=1}^{n+1} \mathbb{P}\{A_i\}.$$

Entonces utilizando lo mismo que el inciso anterior tendrías que

$$\mathbb{P}(X \geq t) = \prod_{i=1}^{n} \mathbb{P}\{A_i\} = [1-(1-e^{-\frac{t}{100}})]^n = e^{-\frac{nt}{100}}.$$

Entonces, podemos concluir que para n componentes $X \sim Exp(n/100)$ ■.

- El modelo Beta.

En el ejemplo anterior usa la relación con la Binomial para hacer el cálculo de la probabilidad en b).
RESPUESTA
Recordemos la relación entre una variable Beta y Binomial. Sea $Y \sim Beta(\alpha = k, \beta = n-(k-1))$ y $X \sim Bin(n,p)$

$$\mathbb{P}(Y > p) = \mathbb{P}(X \leq k-1).$$

Entonces la probabilidad de que al menos el .25 de los restaurantes fracasen considerando $Y \sim Beta(1,4)$ y $X \sim Bin(4, 0.25)$:

$$\mathbb{P}(Y > 0.25) = \mathbb{P}(X \leq 0) = \binom{4}{0}(0.25)^0(0.75)^4 = 0.3164.$$

Por lo que concluimos que **la probabilidad de que al menos el** .25 **de los restaurantes fracasen** es 0.3164062.

En muchos proyectos se emplea un método llamado PERT (Program Evaluation and Review Technique) para coordinar varias actividades en proyectos grandes y/o complejos (por ejemplo, fue empleado en la construcción de los Apolo). Un supuesto estándar de esta técnica es que el tiempo necesario para completar cualquier actividad una vez que ha comenzado, se puede modelar mediante una distribución Beta con

$$A = \text{tiempo optimista (si todo va bien)}$$

$$B = \text{tiempo pesimista (si todo va mal)}$$

Por ejemplo, supongamos que estamos construyendo casas habitación y que el tiempo necesario, en días, para poner los cimientos de una casa sigue una distribución Beta con $A = 2$ y $B = 5$. Además, se puede calcular que $\alpha = 2$ y $\beta = 3$. Calcula la probabilidad de que el tiempo para terminar los cimientos sea menor a 3 días.

RESPUESTA

Sea X una variable aleatoria $Beta(\alpha, \beta)$ con soporte en $[0, 1]$, con una transformación lineal podemos obtener una nueva variable $Y \sim Beta(\alpha, \beta)$ con soporte en $[A, B]$. Sea:

$$Y = X(B - A) + A$$

La función de probabilidad de Y sería:

$$f(y) = \frac{f(x)}{B-A} = \frac{\left(\frac{y-A}{B-A}\right)^{\alpha-1}\left(\frac{B-y}{B-A}\right)^{\beta-1}}{(B-A)B(\alpha,\beta)} = \frac{(y-A)^{\alpha-1}(B-y)^{\beta-1}}{(B-A)^{\alpha+\beta-1}B(\alpha,\beta)}.$$

Entonces para este problema sustituyen los valores de A y B tenemos que la distribución de probabilidad es:

$$f(y) = \frac{(y-2)^{2-1}(5-y)^{3-1}}{(5-2)^{2+3-1}Beta(2,3)} = \frac{(y-2)(5-y)^2}{81 \cdot B(2,3)},$$

donde $1/B(2,3)$ es la función Beta, la cual esta definida como $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$.

Por tanto la **probabilidad de que el tiempo para terminar los cimientos sea menor a 3 días** es:

$$\begin{aligned}
F(3) &= \int_2^3 \frac{(y-2)(5-y)^2}{81}\frac{\Gamma(5)}{\Gamma(2)\Gamma(3)}dy \\
&= \frac{4!}{81\cdot 2!}\int_2^3 y^3 - 12y^2 + 45y - 50dy \\
&= \frac{4}{27}\left(\frac{y^4}{4} - 4y^3 + \frac{45y^2}{2} - 50y\right)|_2^3 \\
&= \frac{4}{27} * 2.75 \\
&= 0.4074074. \blacksquare.
\end{aligned}$$