**14 Distribuciones discretas**

En este capítulo se mostrarán las funciones de R para distribuciones discretas.

**14.1 Funciones disponibles para distribuciones discretas**

Para cada distribución discreta se tienen 4 funciones, a continuación el listado de funciones y su utilidad.

dxxx(x, ...) *# Función de masa de probabilidad, f(x)*

pxxx(q, ...) *# Función de distribución acumulada hasta q, F(x)*

qxxx(p, ...) *# Cuantil para el cual P(X <= q) = p*

rxxx(n, ...) *# Generador de números aleatorios.*

En el lugar de las letras xxx se de debe colocar el nombre de la distribución en R, a continuación el listado de nombres disponibles para las 6 distribuciones discretas básicas.

binom *# Binomial*

geo *# Geométrica*

nbinom *# Binomial negativa*

hyper *# Hipergeométrica*

pois *# Poisson*

multinom *# Multinomial*

Combinando las funciones y los nombres se tiene un total de 24 funciones, por ejemplo, para obtener la función de masa de probabilidad f(x)f(x) de una binomial se usa la función dbinom( ) y para obtener la función acumulada F(x)F(x) de una Poisson se usa la función ppois( ).

**Ejemplo binomial**

Suponga que un grupo de agentes de tránsito sale a una vía principal para revisar el estado de los buses de transporte intermunicipal. De datos históricos se sabe que un 10% de los buses generan una mayor cantidad de humo de la permitida. En cada jornada los agentes revisan siempre 18 buses, asuma que el estado de un bus es independiente del estado de los otros buses.

1. Calcular la probabilidad de que se encuentren exactamente 2 buses que generan una mayor cantidad de humo de la permitida.

Aquí se tiene una distribucion Binomial(n=18,p=0.1)Binomial(n=18,p=0.1) y se desea calcular P(X=2)P(X=2). Para obtener esta probabilidad se usa la siguiente instrucción.

dbinom(x=2, size=18, prob=0.10)

## [1] 0.2835121

Así P(X=2)=0.2835P(X=2)=0.2835.

1. Calcular la probabilidad de que el número de buses que sobrepasan el límite de generación de gases sea al menos 4.

En este caso interesa calcular P(X≥4)P(X≥4), para obtener esta probabilidad se usa la siguiente instrucción.

sum(dbinom(x=4:18, size=18, prob=0.10))

## [1] 0.09819684

Así P(X≥4)=0.0982P(X≥4)=0.0982

1. Calcular la probabilidad de que tres o menos buses emitan gases por encima de lo permitido en la norma.

En este caso interesa P(X≤3)P(X≤3) lo cual es F(x=3)F(x=3), por lo tanto, la instrucción para obtener esta probabilidad es

pbinom(q=3, size=18, prob=0.10)

## [1] 0.9018032

Así P(X≤3)=F(x=3)=0.9018P(X≤3)=F(x=3)=0.9018

1. Dibujar la función de masa de probabilidad.

Para dibujar la función de masa de probabilidad para una Binomial(n=18,p=0.1)Binomial(n=18,p=0.1) se usa el siguiente código.

x <- 0:18 *# Soporte (dominio) de la variable*

Probabilidad <- dbinom(x=x, size=18, prob=0.1)

plot(x=x, y=Probabilidad,

type='h', las=1, lwd=6)

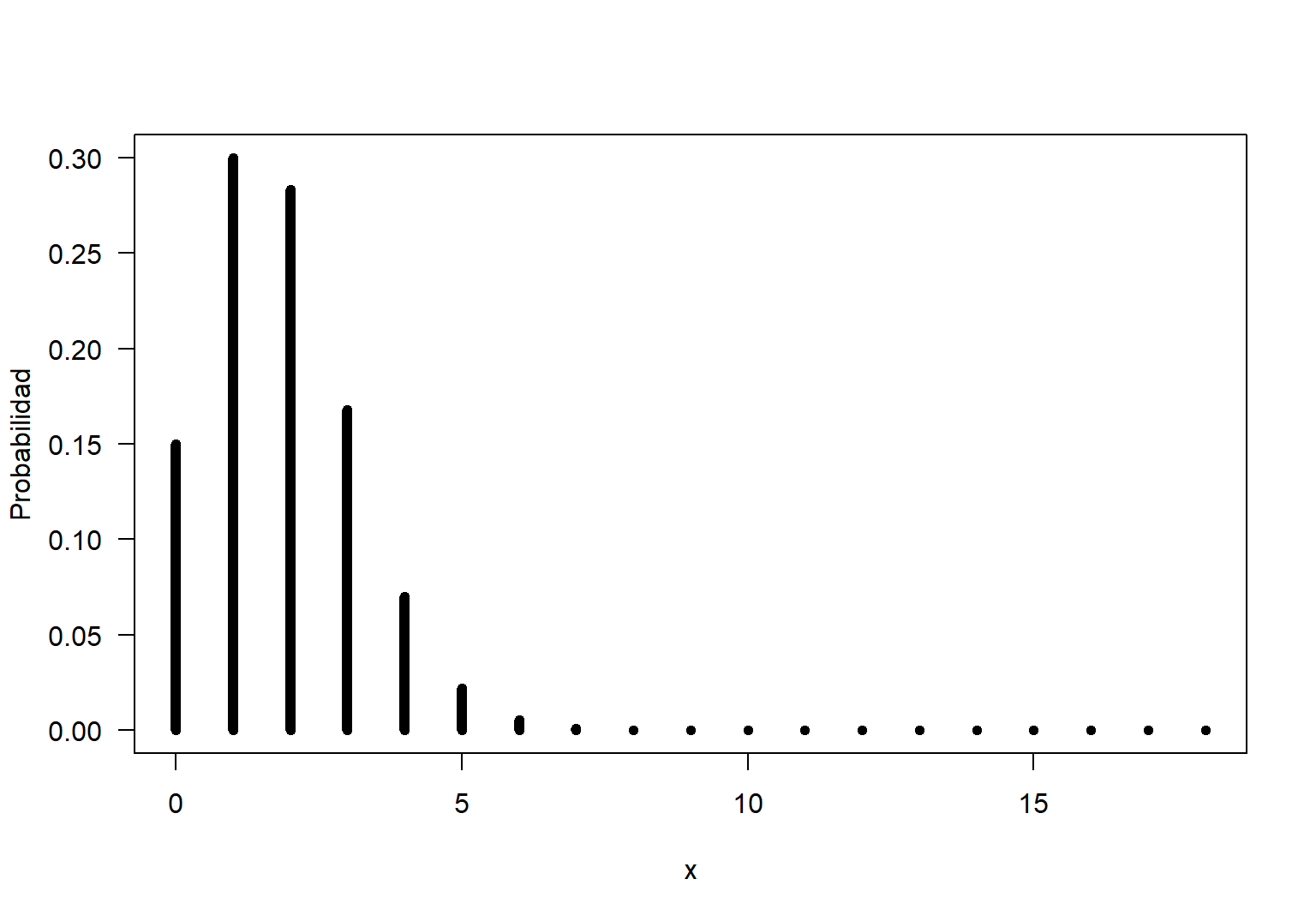


Figure 14.1: Función de masa de probabilidad para una Binomial(n=18,p=0.1)Binomial(n=18,p=0.1).

En la Figura [14.1](https://fhernanb.github.io/Manual-de-R/discretas.html#fig:binom1) se muestra la función de masa de probabilidad para la Binomial(n=18,p=0.1)Binomial(n=18,p=0.1), de esta figura se observa claramente que la mayor parte de la probabilidad está concentrada para valores pequeños de XX debido a que la probabilidad de éxito individual es p=0.10p=0.10. Valores de X≥7X≥7 tienen una probabilidad muy pequeña y es por eso que las longitudes de sus barras son muy cortas.

1. Generar con 100 de una distribución Binomial(n=18,p=0.1)Binomial(n=18,p=0.1) y luego calcular las frecuencias muestrales y compararlas con las probabilidades teóricas.

La muestra aleatoria se obtiene con la función rbinom y los resultados se almacenan en el objeto m, por último se construye la tabla de frecuencias relativas, a continuación el código usado.

m <- rbinom(n=100, size=18, prob=0.1)

m *# Para ver lo que hay dentro de m*

## [1] 0 1 4 1 3 4 2 3 1 0 3 2 1 4 1 1 1 3 0 5 0 3 5 3 2 3 1 1 2 4 1 3 2 2 3 2 2

## [38] 3 2 3 0 2 2 2 1 1 3 1 1 1 1 1 1 1 1 3 1 2 4 4 7 2 4 0 1 2 1 1 2 3 2 1 3 1

## [75] 1 2 3 5 1 5 0 1 2 1 4 1 1 2 1 2 0 2 2 3 3 0 1 1 3 2

prop.table(table(m)) *# Tabla de frecuencia relativa*

## m

## 0 1 2 3 4 5 7

## 0.09 0.35 0.24 0.19 0.08 0.04 0.01

A pesar de ser una muestra aleatoria de sólo 100 observaciones, se observa que las frecuencias relativas obtenidas son muy cercanas a las mostradas en la Figura [14.1](https://fhernanb.github.io/Manual-de-R/discretas.html#fig:binom1).

**Ejemplo geométrica**

En una línea de producción de bombillos se sabe que sólo el 1% de los bombillos son defectuosos. Una máquina automática toma un bombillo y lo prueba, si el bombillo enciende, se siguen probando los bombillos hasta que se encuentre **un** bombillo defectuoso, ahí se para la línea de producción y se toman los correctivos necesarios para mejorar el proceso.

1. Calcular la probabilidad de que se necesiten probar 125 bombillos para encontrar el primer bombillo defectuoso.

En la distribución geométrica, la variable XX representa el número de fracasos antes de encontrar el único éxito, por lo tanto, en este caso el interés es calcular P(X=124)P(X=124). La instrucción para obtener esta probabiliad es la siguiente.

dgeom(x=124, prob=0.01)

## [1] 0.002875836

1. Calcular P(X≤8)P(X≤8).

En este caso interesa P(X≤50)P(X≤50) lo que equivale a F(8)F(8), la instrucción para obtener la probabilidad es la siguiente.

pgeom(q=50, prob=0.01)

## [1] 0.401044

1. Encontrar el cuantil qq tal que P(X≤q)=0.40P(X≤q)=0.40.

En este caso interesa encontrar el cuantil qq que cumpla la condición de que hasta qq esté el 40% de las observaciones, por esa razón se usa la función qgeom como se muestra a continuación.

qgeom(p=0.4, prob=0.01)

## [1] 50

Note que las funciones pxxx y qxxx están relacionadas, pxxx entrega la probabilidad hasta el cuantil qq mientras qxxx entrega el cuantil en el que se acumula pp probabilidad.

**Ejemplo binomial negativa**

Una familia desea tener hijos hasta conseguir **2 niñas**, la probabilidad individual de obtener una niña es 0.5 y se supone que todos los nacimientos son individuales, es decir, un sólo bebé.

1. Calcular la probabilidad de que se necesiten 4 hijos, es decir, 4 nacimientos para consguir las dos niñas.

En este problema se tiene una distribución binomial negativa con r=2r=2 niñas, los éxitos deseados por la familia. La variable XX representa los fracasos, es decir los niños, hasta que se obtienen los éxitos r=2r=2 deseados.

En este caso lo que interesa es P(familia tenga 4)P(familia tenga 4), en otras palabras interesa P(X=2)P(X=2), la instrucción para calcular la probabilidad es la siguiente.

dnbinom(x=2, size=2, prob=0.5)

## [1] 0.1875

1. Calcular P(familia tenga al menos 4 hijos)P(familia tenga al menos 4 hijos).

Aquí interesa calcular P(X≥2)=P(X=2)+P(X=3)+…P(X≥2)=P(X=2)+P(X=3)+…, como esta probabilidad va hasta infinito, se debe usar el complemento así:

P(X≥2)=1−[P(X=0)+P(X=1)]P(X≥2)=1−[P(X=0)+P(X=1)]y para obtener la probabilidad solicitada se puede usar la función dnbinom de la siguiente manera.

1 - sum(dnbinom(x=0:1, size=2, prob=0.5))

## [1] 0.5

Otra forma para obtener la probabilidad solicitada es por medio de la función pnbinom de la siguiente manera.

1 - pnbinom(q=1, size=2, prob=0.5)

## [1] 0.5

**Ejemplo hipergeométrica**

Un lote de partes para ensamblar en una empresa está formado por 100 elementos del proveedor A y 200 elementos del proveedor B. Se selecciona una muestra de 4 partes al azar sin reemplazo de las 300 para una revisión de calidad.

1. Calcular la probabilidad de que todas las 4 partes de la muestra sean del proveedor A.

Aquí se tiene una situación que se puede modelar por medio de una distribución hipergeométrica con m=100m=100 éxitos en la población, n=200n=200 fracasos en la población y k=4k=4 el tamaño de la muestra. El objetivo es calcular P(X=4)P(X=4), para obtener esta probabilidad se usa la siguiente instrucción.

dhyper(x=4, m=100, k=4, n=200)

## [1] 0.01185408

1. Calcular la probabilidad de que dos o más de las partes sean del proveedor A.

Aquí interesa P(X≥2)P(X≥2), la instrucción para obtener esta probabilidad es.

sum(dhyper(x=2:4, m=100, k=4, n=200))

## [1] 0.4074057

**Ejemplo Poisson**

En una editorial se asume que todo libro de 250 páginas tiene en promedio 50 errores.

1. Encuentre la probabilidad de que en una página cualquiera no se encuentren errores.

Este es un problema de distribución Poisson con tasa promedio de éxitos dada por:

λ=50erroreslibro=0.2errorespaginaλ=50erroreslibro=0.2errorespaginaEl objetivo es calcular P(X=0)P(X=0), para obtener esta probabilidad de usa la siguiente instrucción.

dpois(x=0, lambda=0.2)

## [1] 0.8187308

Así P(X=0)=0.8187P(X=0)=0.8187.

**14.2 Distribuciones discretas generales**

En la práctica nos podemos encontramos con variables aleatorias discretas que no se ajustan a una de las distribuciones mostradas anteriormente, en esos casos, es posible manejar ese tipo de variables por medio de unas funciones básicas de R como se muestra en el siguiente ejemplo.

**Ejemplo**

El cangrejo de herradura hembra se caracteriza porque su caparazón se adhieren los machos de la misma especie, en la Figura [14.2](https://fhernanb.github.io/Manual-de-R/discretas.html#fig:crab) se muestra una fotografía de este cangrejo. Los investigadores están interesado en determinar cual es el patrón de variación del número de machos sobre cada hembra, para esto, se recolectó una muestra de hembras a las cuales se les observó el color, la condición de la espina, el peso en kilogramos, el ancho del caparazón en centímetros y el número de satélites o machos sobre el caparazón, la base de datos está disponible en el siguiente [enlace](https://raw.githubusercontent.com/fhernanb/datos/master/crab).



Figure 14.2: Fotografía del cangrejo de herradura, tomada de <http://sccoastalresources.com/home/2016/6/21/a-night-of-horseshoe-crab-tagging>

1. Encontrar la distribución de probabilidad para la variable Sa que corresponde al número de machos sobre el caparazón de cada hembra.

Primero se debe leer la base de datos usando la url suministrada y luego se construye la tabla de frecuencia relativa y se almacena en el objeto t1.

url <- 'https://raw.githubusercontent.com/fhernanb/datos/master/crab'

crab <- read.table(file=url, header=T)

t1 <- prop.table(table(crab$Sa))

t1

##

## 0 1 2 3 4 5

## 0.358381503 0.092485549 0.052023121 0.109826590 0.109826590 0.086705202

## 6 7 8 9 10 11

## 0.075144509 0.023121387 0.034682081 0.017341040 0.017341040 0.005780347

## 12 14 15

## 0.005780347 0.005780347 0.005780347

La anterior tabla de frecuencias relativas se puede representar gráficamente usando el siguiente código.

plot(t1, las=1, lwd=5, xlab='Número de satélites',

ylab='Proporción')

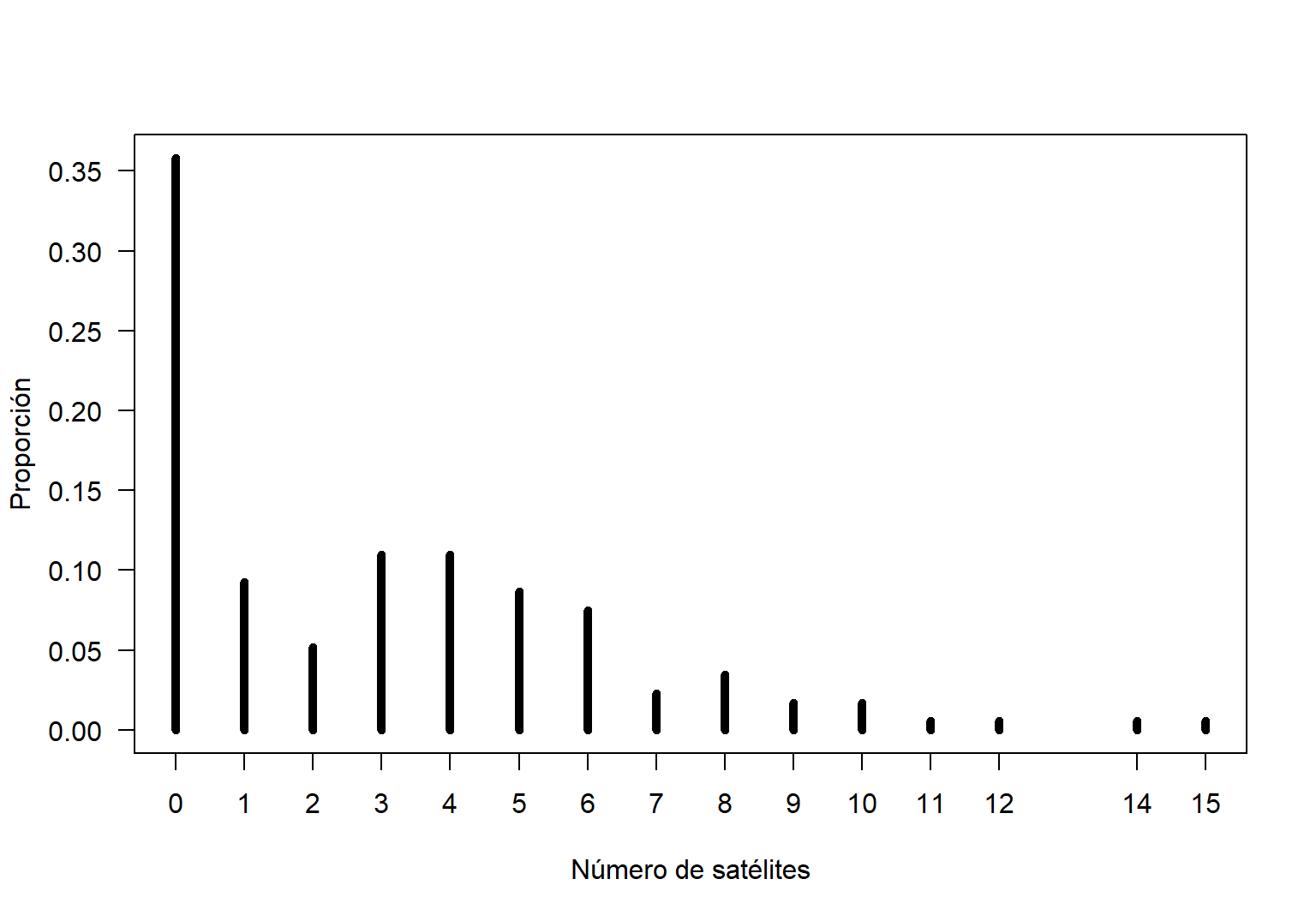


Figure 14.3: Función de masa de probabilidad para el número de satélites por hembra.

1. Sea XX la variable número de satélites por hembra, construir la función F(x)F(x).

Para construir F(x)F(x) se utiliza la función ecdf o *empirical cumulative density function*, a esta función le debe ingresar el vector con la información de la variable cuantitativa, a continuación del código usado. En la Figura [14.4](https://fhernanb.github.io/Manual-de-R/discretas.html#fig:Fcrab) se muestra la función de distribución acumulada para para el número de satélites por hembra.

F <- ecdf(crab$Sa)

plot(F, las=1, main='')

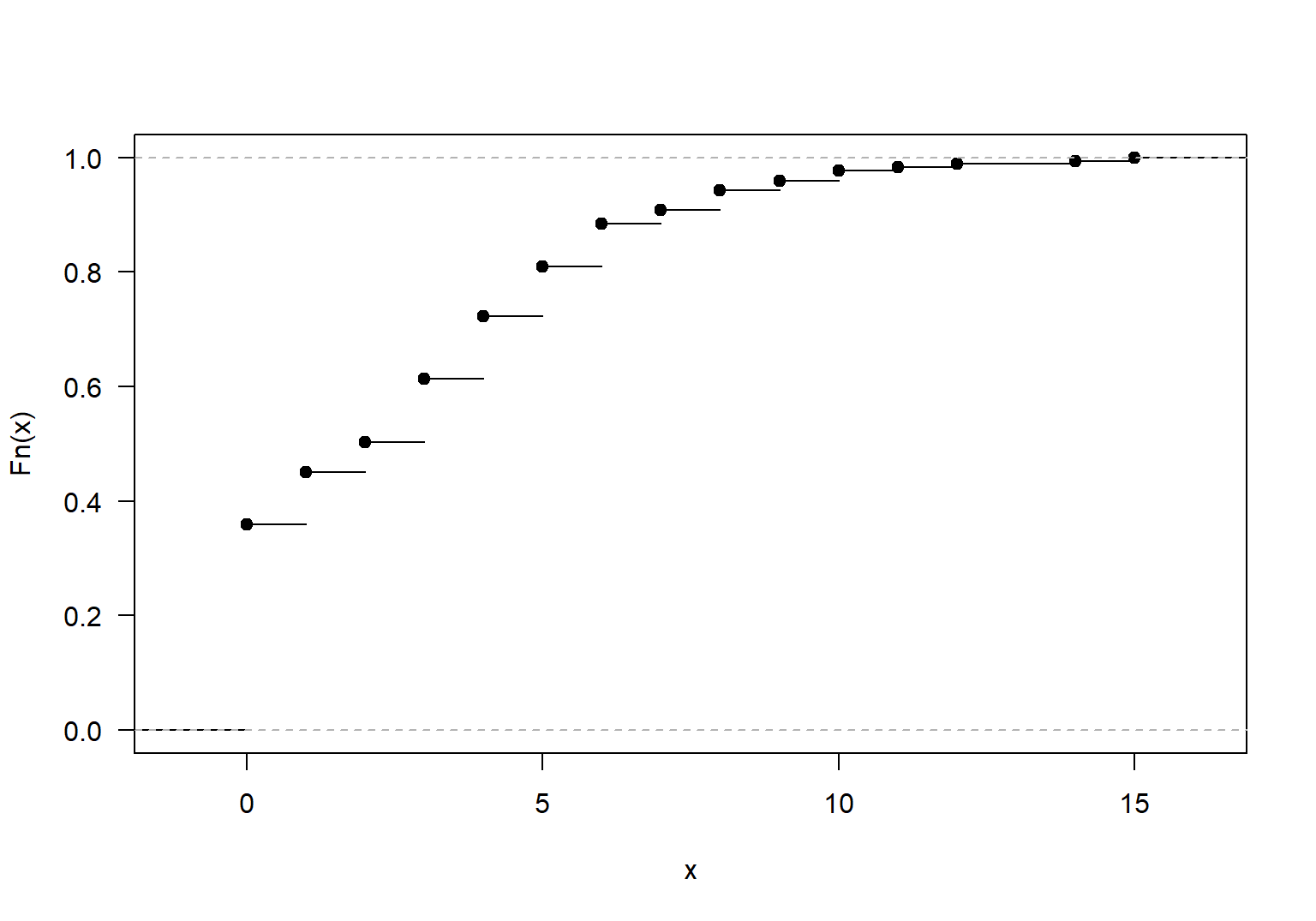


Figure 14.4: Función de distribución acumulada para el número de satélites por hembra.

1. Calcular P(X≤9)P(X≤9).

Para obtener esta probabilidad se usa el objeto F que es en realidad una función, a continuación la instrucción usada.

F(9)

## [1] 0.9595376

Así P(X≤9)=0.9595P(X≤9)=0.9595.

1. Calcular P(X>4)P(X>4).

Para obtener esta probabilidad se usa el hecho de que P(X>4)=1−P(X≤4)P(X>4)=1−P(X≤4), así la instrucción a usar es.

1 - F(4)

## [1] 0.2774566

Por lo tanto P(X>4)=0.2775P(X>4)=0.2775.

1. Suponga que el grupo 1 está formado por las hembras cuyo ancho de caparazón es menor o igual al ancho mediano, el grupo 2 está formado por las demás hembras. ¿Será F(x)F(x) diferente para los dos grupos?

Para realizar esto vamos a particionar el vector Sa en los dos grupos de acuerdo a la nueva variable grupo creada como se muestra a continuacion.

grupo <- ifelse(crab$Wt <= median(crab$Wt), 'Grupo 1', 'Grupo 2')

x <- split(x=crab$Sa, f=grupo)

El objeto x es una lista y para acceder a los vectores allí almacenados usamos dos corchetes [[]], uno dentro del otro. Luego para calcular F(x)F(x) para los dos grupos se procede así:

F1 <- ecdf(x[[1]])

F2 <- ecdf(x[[2]])

Para obtener las dos F(x)F(x) en la misma figura se usa el código siguiente.

plot(F1, col='blue', main='', las=1)

plot(F2, col='red', add=T)

legend('bottomright', legend=c('Grupo 1', 'Grupo 2'),

col=c('blue', 'red'), lwd=1)

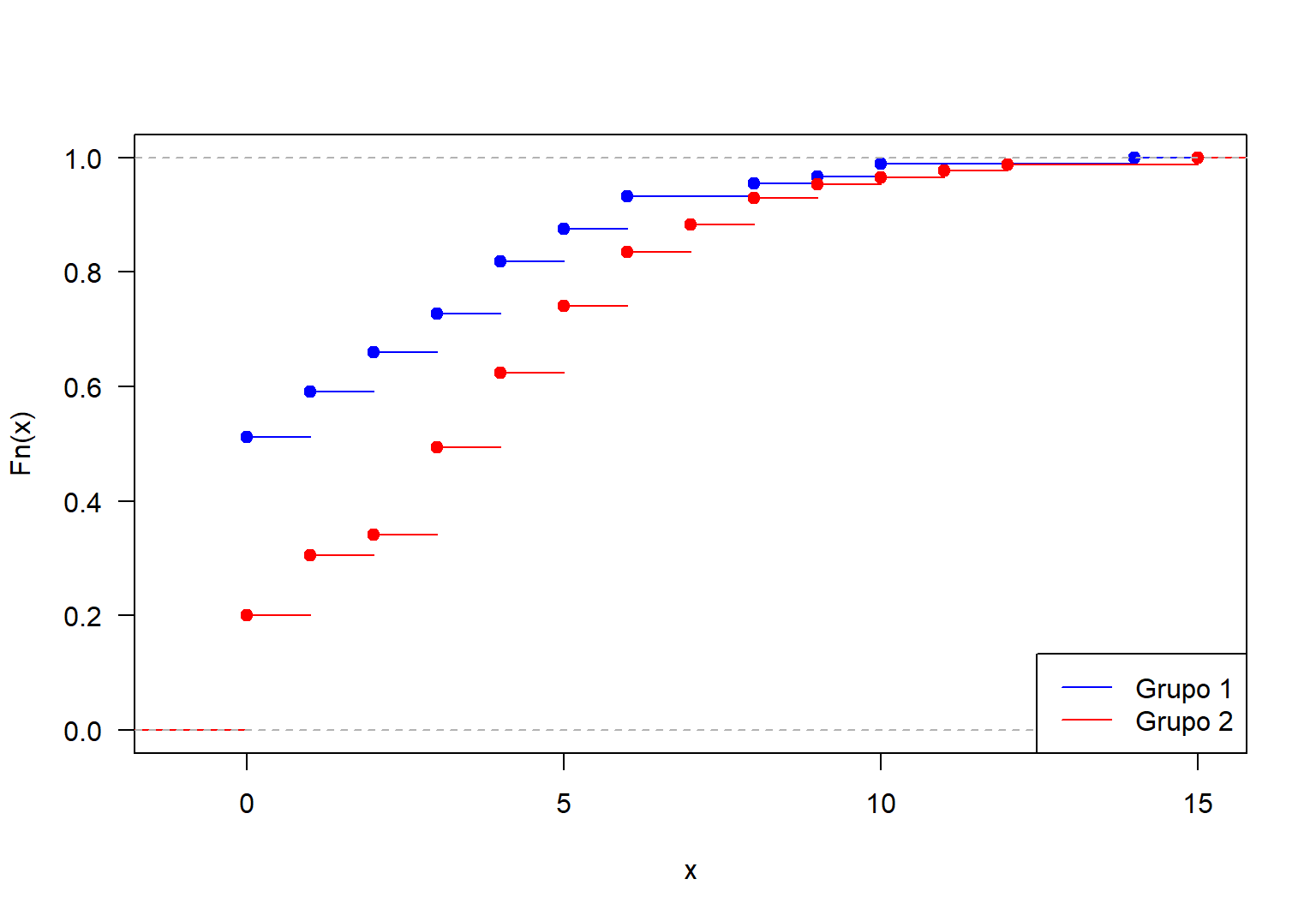


Figure 14.5: Función de distribución acumulada para el número de satélites por hembra diferenciando por grupo.

En la Figura [14.5](https://fhernanb.github.io/Manual-de-R/discretas.html#fig:Fcrabs) se muestran las dos F(x)F(x), en color azul para el grupo 1 y en color rojo para el grupo 2. Se observa claramente que las curvas son diferentes antes de x=9x=9. El hecho de que la curva azul esté por encima de la roja para valores menores de 9, es decir, F1(x)≥F2(x)F1(x)≥F2(x), indica que las hembras del grupo 1 tienden a tener menos satélites que las del grupo 2, esto es coherente ya que las del grupo 2 son más grandes en su caparazón.