Tema 3: PROBABILIDAD - V

Variables aleatorias discretas v

Biología sanitaria 2017/18. Universidad de Alcalá

M. Marvá. Actualizado: 2017-10-22

Ejemplo: otro de variable binomial. Para ello usaremos el genoma del bacteriófago ΦX174 que fue el primer genoma basado en ADN que se secuenció en 1977. Usaremos Bioconductor para descargar el genoma completo. Sólo necesitamos el *número de acceso* en la base de datos GenBank del NCBI

```
# Instalamos el software de BioConductor
source("https://bioconductor.org/biocLite.R")
if(!require(annotate))biocLite("annotate")
# Descargamos el genoma
phiX174 = getSEQ("NC_001422.1")
phiX174
```

```
## [1] "GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTCGGATATTTCTGATGAGTCGAAAAATTAT
```

```
# ¿Número de bases?
nchar(phiX174)
```

```
## [1] 5386
```

```
# Separamos la secuencia en caracteres (nucleótidos)
genoma = strsplit(phiX174, split = "")[[1]]
head(genoma)
```

```
## [1] "G" "A" "G" "T" "T" "T"
```

Ejemplo: (continuación)

```
# Y hacemos su tabla de frecuencias
(frecNucltd = table(genoma))
```

```
## genoma
##    A    C    G    T
## 1291 1157 1254 1684
```

Interesados en

$$X = \textbf{número de citosinas entre los 25 bases elegidas al azar}$$

- Se fija $n = 25$ nucleótidos del genoma.
- Para cada uno de los 25, es un éxito si el nucleótido elegido es citosina (C) y un fracaso en otro caso (A, T, G).

```
n = 25
(p =  unname(frecNucltd["C"] / length(genoma)))
```

```
## [1] 0.2148162
```

(pregunta a tu profesor por `unname`) para definir la variable:

Entonces $X \sim B(25, 0.21)$.

Ejemplo: (continuación)

- ¿Cuál es la probabilidad de que 5 de los 25 nucleótidos elegidos al azar sean citosinas? Con la fórmula de la binomial

$$P(X = 5) = \binom{25}{5} (0.2148)^5 (0.7852)^{20}$$

Usando R:

```
## [1] 0.2148162
```

```
k = 5
(q = 1 - p)
```

```
## [1] 0.7851838
```

```
choose(n, k) * p^k * q^(n - k)
```

```
## [1] 0.1927976
```

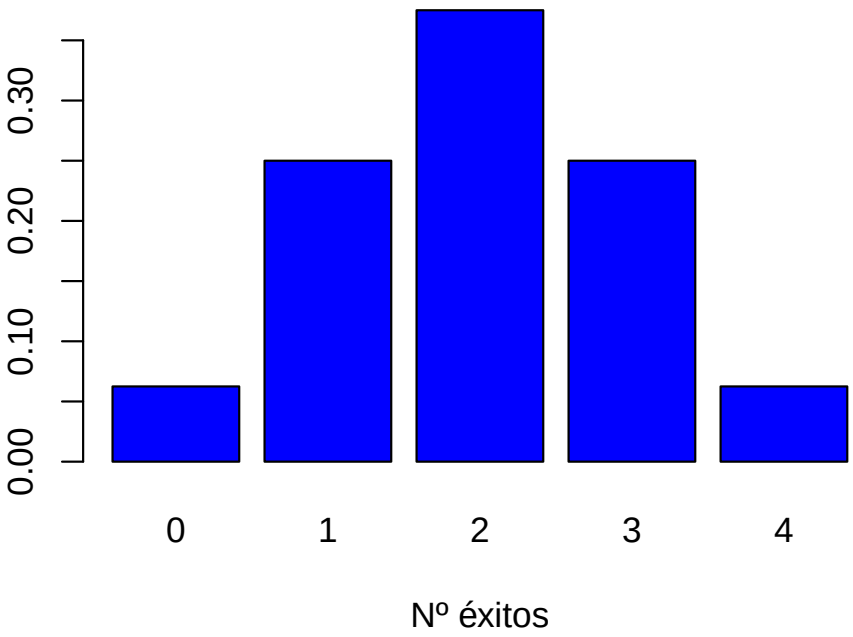
- Directamente con `dbinom`:

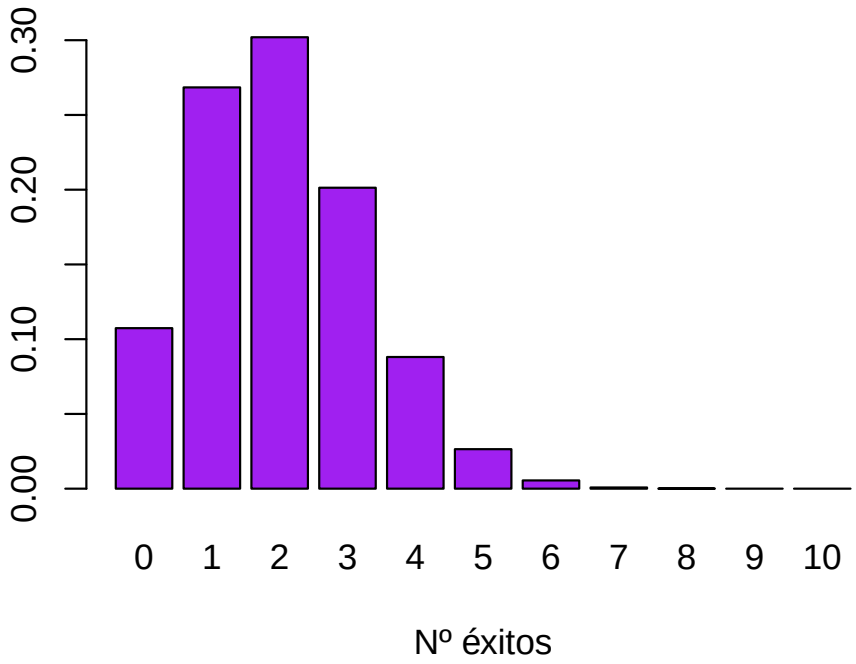
```
dbinom(x = k, size = n, prob = p)
```

```
## [1] 0.1927976
```

Ejemplo: dispersión de una variable aleatoria discreta.

Sean $X \sim B(4, 0.5)$ e $Y \sim B(10, 0.2)$,





Las probabilidades P(Y=7), P(Y=8), P(Y=9), P(Y=10) son muy pequeñas.

Ambas cumplen

$$\mu_X = 2, \qquad \mu_Y = 2$$

pero en una los valores están más agrupados que la otra

La varianza de una variable aleatoria discreta

Inspirado en la media de una variable estadística discreta que toma valores x_1, x_2,..., x_k con frecuencias absolutas f_1, f_2,..., f_k:

$$Var(x) = \frac{\sum_{i=1}^{k} f_i \cdot (x_i - \bar{x})^2}{n} = \sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot \frac{f_i}{n}$$

aquí $\frac{f_i}{n}$ es la frecuencia relativa número i, que se identifica con su probabilidad.

Si X es una variable aleatoria discreta, que toma los valores $x_1, x_2, \ldots, x_k$, con las probabilidades $p_1, p_2, \ldots, p_k$ (donde $p_i = P(X = x_i)$) y valor esperado μ, entonces la **varianza** de X es:

$$\sigma_X^2 = \sum_{i=1}^{k} (x_i - \mu)^2 \cdot P(X = x_i) = (x_1 - \mu)^2 p_1 + \cdots + (x_k - \mu)^2 p_k.$$

y su **desviación típica**

$$\sigma_X = \sqrt{\sum_{i=1}^{k} (x_i - \mu)^2 \cdot P(X = x_i)}.$$

Ejemplo: $X =$ "Número de descendientes con ojos verdes de un total 5, con $P(V) = 0.2$".

```
valores_X = 0:5
probabilidades_X = dbinom(0:5, size = 5, prob = 0.2)
(mu_X= sum(valores_X*probabilidades_X))
```

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

```
(sigma2_X=sum((valores_X-mu_X)^2*probabilidades_X))
```

```
## [1] 0.8
```

Ejemplo: (continuación)

total_descendientes	Valor_esperado	varianza
1	0.2	0.16
2	0.4	0.32
3	0.6	0.48
4	0.8	0.64
5	1.0	0.80

Ejemplo: Y_1 = "Resultado de lanzar 1 dado".

```
valores_Y1 = 1:6
probabilidades_Y1 = rep(1,6)/6
(mu_Y1= sum(valores_Y1*probabilidades_Y1))
```

```
## [1] 3.5
```

```
(sigma2_Y1=sum((valores_Y1-mu_Y1)^2*probabilidades_Y1))
```

```
## [1] 2.916667
```

Ejemplo: Y = "Suma del resultado de lanzar 2 dados".

```
valores_Y2 = 2:12
probabilidades_Y2 = c(1:6,5:1)/36
(mu_Y2= sum(valores_Y2*probabilidades_Y2))
```

```
## [1] 7
```

```
(sigma2_Y2=sum((valores_Y2-mu_Y2)^2*probabilidades_Y2))
```

```
## [1] 5.833333
```

Propiedades de la media y la varianza

Y si X_1, X_2 son dos variables aleatorias, se tiene:

$$E(X_1 + X_2) = E(X_1) + E(X_2).$$

Si además X_1 y X_2 son *independientes*, entonces

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2).$$

Si X es una variable aleatoria, y a, b son números cualesquiera, entonces

$$E(a \cdot X + b) = a \cdot E(X) + b, \quad \text{Var}(a \cdot X + b) = a^2 \cdot \text{Var}(X).$$

Además, para cualquier variable aleatoria X con media μ y varianza σ^2

la variable aleatoria

$$Z = \frac{X - \mu}{\sigma}$$

es una nueva variable aleatoria que cumple:

$$\mu_Z = 0 \qquad \text{y} \qquad \sigma_Z^2 = 1.$$

Se dice que Z es una **variable aleatoria estandarizada**.

Ejemplo: Sea C="Temperatura en Alcalá, en Celsius" y suponer que la temperatura esperada (temperatura media) es

$$E[C] = 13.8, \qquad Var[C] = 3.7$$

Si queremos la temperatura esperada en Fahrenheit, como

$$F = 32 + \frac{9}{5}C$$

entonces

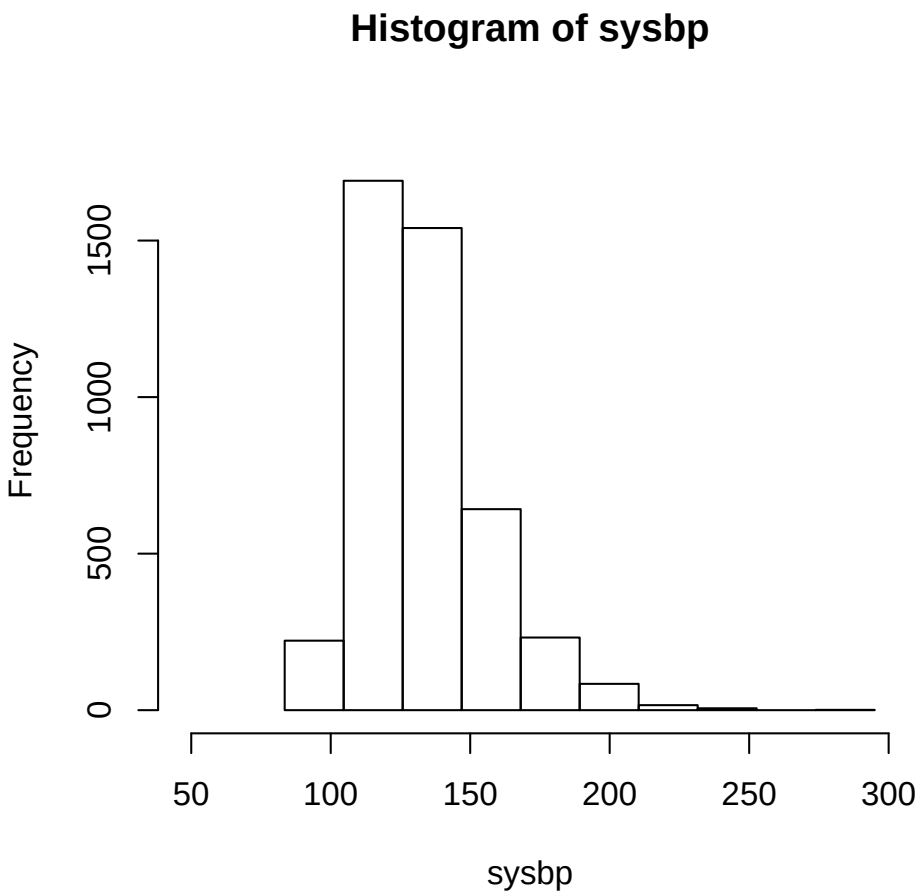
$$\begin{aligned} E[F] &= E\left[32 + \frac{9}{5}C\right] \\ &= 32 + \frac{9}{5}E[C] \\ &= 32 + \frac{9}{5}13.8 = 56.84F \end{aligned}$$

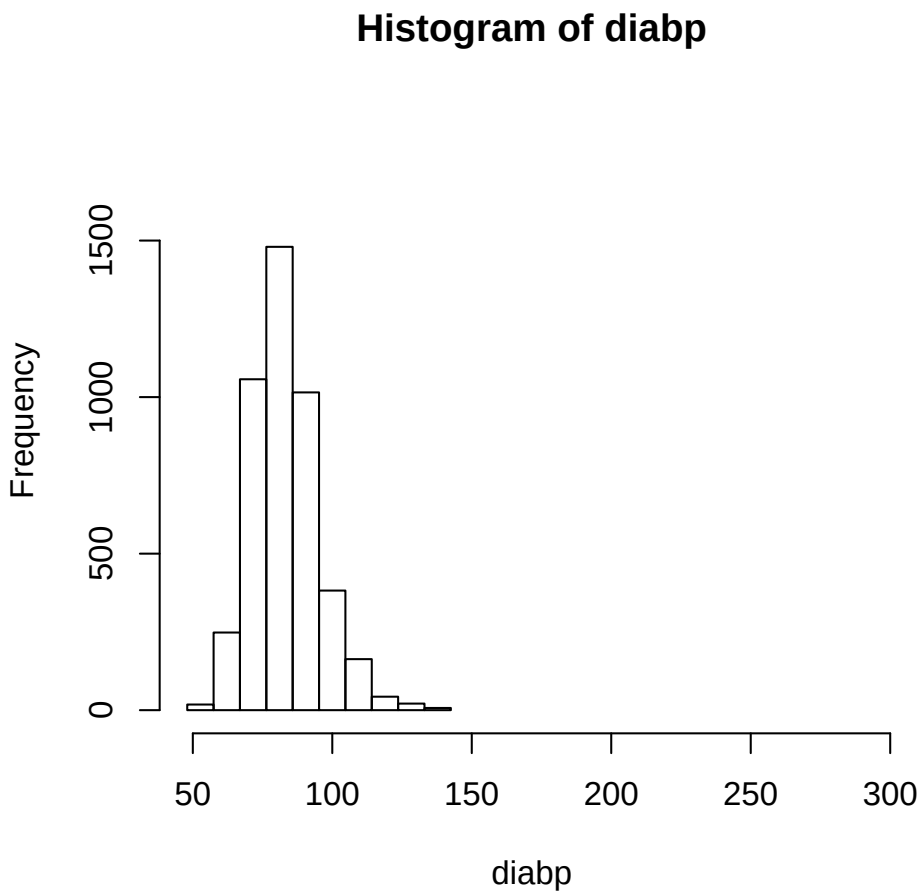
y, adem'as

$$\begin{aligned} Var[F] &= Var\left[32 + \frac{9}{5}C\right] \\ &= \left(\frac{9}{5}\right)^2 Var[C] = 11.988F^2 \end{aligned}$$

Ejemplo: Presiones sistólica y diastóolica de los datos del experimento de Framhinham

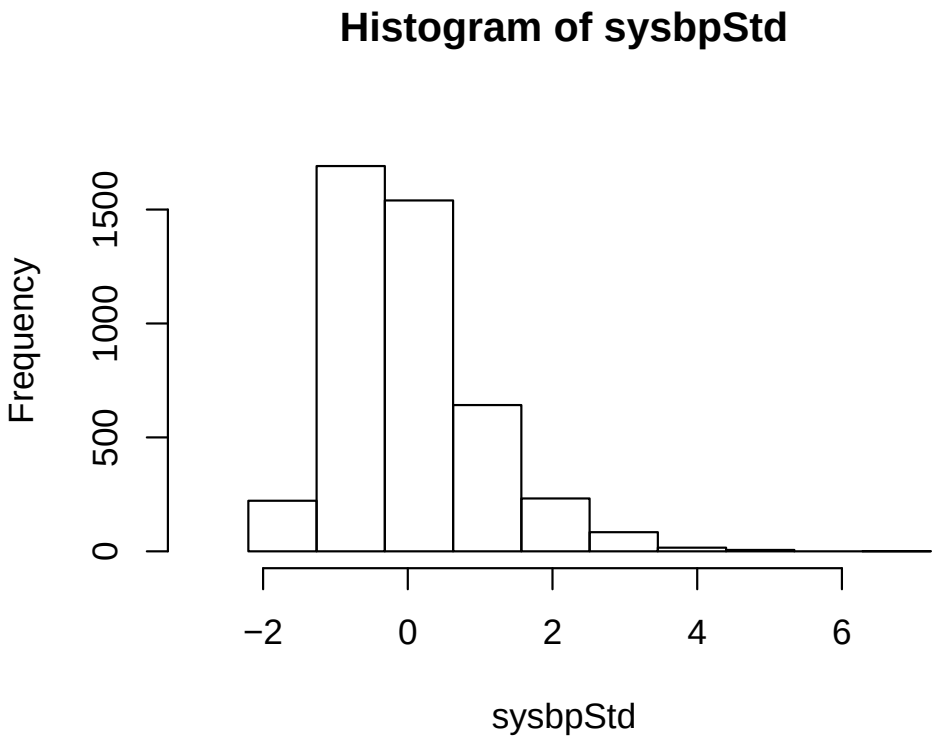
```
framhinham = read.table(file = "FraminghamDataSet.csv",
                        sep = ";", dec = ".", header = TRUE)

sysbp = framhinham$sysbp1; diabp = framhinham$diabp1
brks = function(x){seq(from=min(x), to = max(x), length.out = 11)}
par(mfrow = c(1,2))
hist(sysbp, breaks = brks(sysbp), xlim=c(min(sysbp, diabp), max(sysbp, diabp)), ylim = c(0, 1850))
hist(diabp, breaks = brks(diabp), xlim=c(min(sysbp, diabp), max(sysbp, diabp)), ylim = c(0, 1850))
```

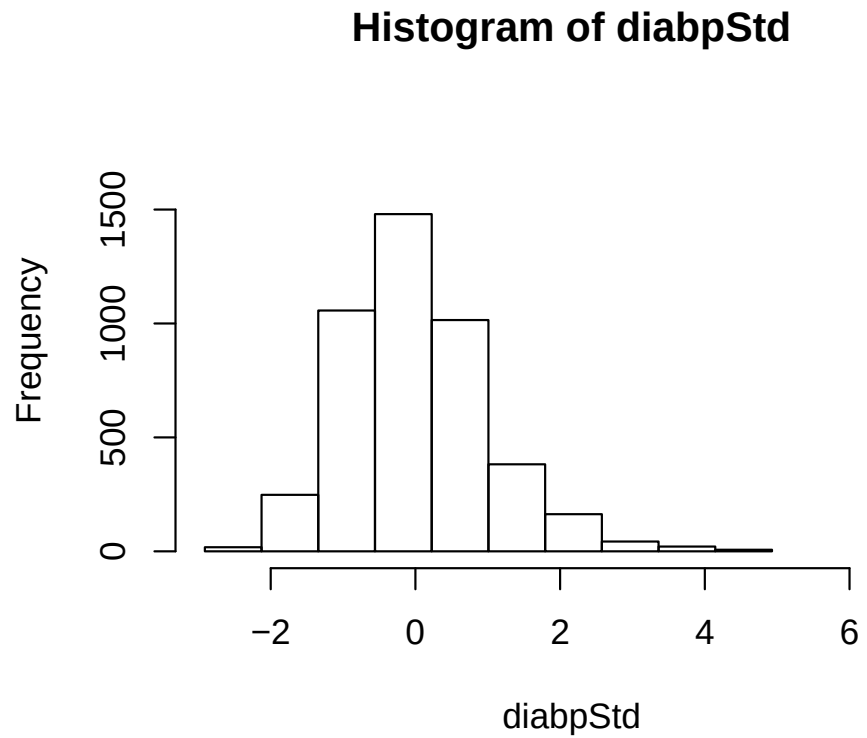





```
par(mfrow = c(1,1))
```

```r
par(mfrow = c(1,2))
# ESTANDARIZAR VARIABLES
sysbpStd = (sysbp-mean(sysbp))/sd(sysbp)
diabpStd = (diabp-mean(diabp))/sd(diabp)
hist(sysbpStd, breaks = brks(sysbpStd), xlim=c(min(sysbpStd, diabpStd), max(sysbpStd, diabpStd)), ylim = c(0, 1850))
hist(diabpStd, breaks = brks(diabpStd), xlim=c(min(sysbpStd, diabpStd), max(sysbpStd, diabpStd)), ylim = c(0, 1850))
```






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

Ejemplo: Supongamos que $X \sim Bernoulli(p)$. Entonces

valores	X = 0	X = 1
Probabilidades	P(X = 0) = q = 1-p	P(X = 1) = p

La media es

$$\mu_X = 0 \cdot q + 1 \cdot p = p$$

La varianza es

$$\sigma_X^2 = (0 - p)^2 \cdot q + (1 - p)^2 \cdot p = p \cdot q$$

Ejemplo: Supongamos que $X \sim B(n, p)$. Podemos ver X como la suma de n variables aleatorias Bernoullí $Bernoulli(p)$ independientes

$$X = X_1 + X_2 + \cdots + X_n, \qquad X_i \sim Bernoulli(p)$$

por tanto, la media es

$$\begin{aligned} \mu_X = E[X] &= E[X_1 + \cdots + X_n] \\ &= E[X_1] + \cdots + E[X_n] \\ &= p + \cdots + p \\ &= n \cdot p \end{aligned}$$

La varianza es

$$\begin{aligned} \sigma_X^2 = \sigma^2[X] &= \sigma^2[X_1 + \cdots + X_n] \\ &= \sigma^2[X_1] + \cdots + \sigma^2[X_n] \\ &= p \cdot q + \cdots + p \cdot q \\ &= n \cdot p \cdot q \end{aligned}$$

Y la desviación típica

$$\sigma_X = \sqrt{n \cdot p \cdot q}$$

Probabilidades acumuladas de una variable discreta

Inspirados en las frecuencias relativas acumuladas

Ejemplo: $X =$ "Nº de descendientes con ojos verdes de un total de 5, con $P(V) = 0.2$".

Probabilidades de **que haya al menos** x descendientes con ojos verdes de un total de 5.

```
tabla = rbind(0:5, pbinom(0:5, size = 5, prob = 0.2))
rownames(tabla) = c("X", "P(X<=x)")
colnames(tabla) = rep(x = "", times = 6)
tabla
```

```
##
## X        0.00000 1.00000 2.00000 3.00000 4.00000 5
## P(X<=x) 0.32768 0.73728 0.94208 0.99328 0.99968 1
```

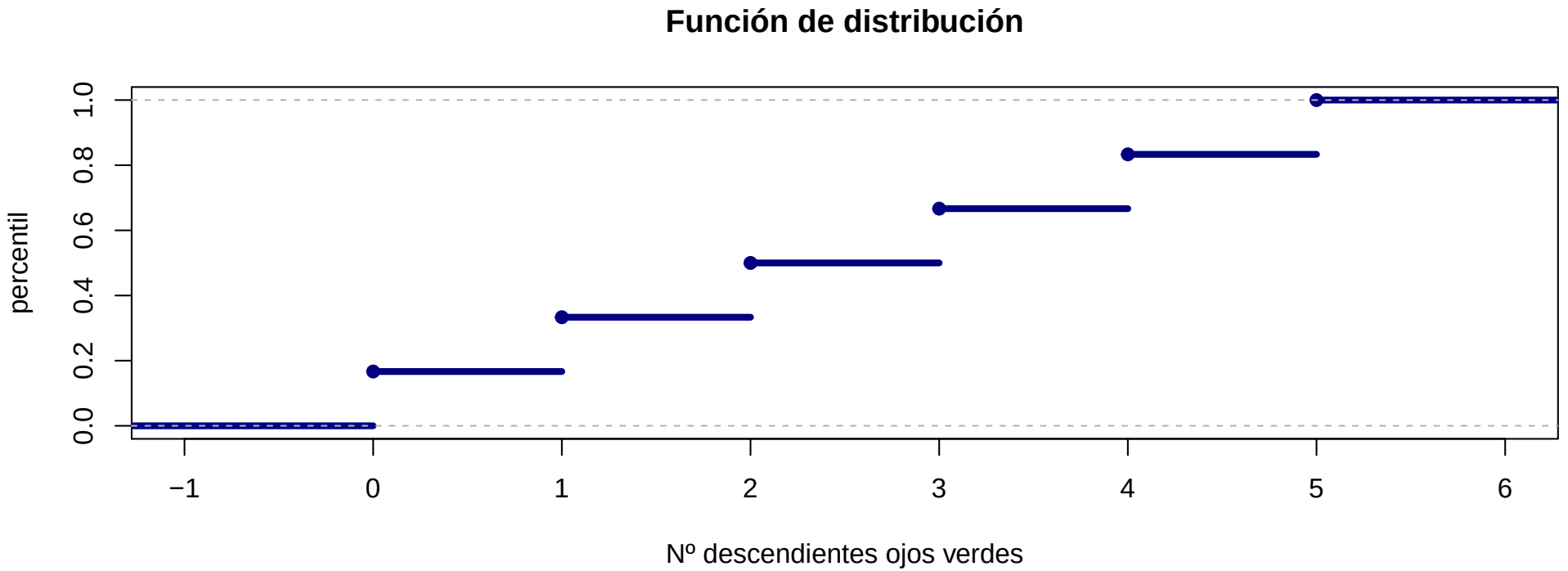
es la **función de distribución de probabilidades** de $X \sim B(5, 0.2)$.

Probabilidades de **que haya al menos** x descendientes con ojos verdes de un total de 5.

```
##
## X       0.00000 1.00000 2.00000 3.00000 4.00000 5
## P(X<=x) 0.32768 0.73728 0.94208 0.99328 0.99968 1
```

Visualización de la **función de distribución**

```
funcionDistribucion = ecdf(tabla[1, ])
plot(funcionDistribucion, lwd=4, col="navy",
     main="Función de distribución",
     xlab="Nº descendientes ojos verdes", ylab="percentil")
```



Formalicemos lo anterior

Si X es una variable aleatoria que sólo toma una cantidad finita de valores numéricos $x_1, x_2, x_3, \ldots, x_k$, con probabilidades $p_i = P(X = x_i)$, la tabla

Valor:	x_1	x_2	x_3	$\cdots$	x_k
Probabilidad:	p_1	$p_1 + p_2$	$p_1 + p_2 + p_3$	$\cdots$	$p_1 + \cdots + p_k$

es la **función de distribución de probabilidad** de la variable aleatoria X.

$$F(x) = P(X \leq x)$$

Unma variable aleatoria continua

Ejercicio: Sea $X =$ "Nº de horas de espera en el servicio de urgencias". A partir de una muestra grande se determina que

Minutos espera	Nº individuos
$[0, 60)$	25
$[60, 120)$	35
$[120, 180)$	20
$[180, 240)$	15
$[240, 300)$	3
Más de 300	2

1. Usa las frecuencias relativas para calcular la probabilidad de esperar entre 0 y 59 minutos, entre 60 y 119,...
2. Calcula la probabilidad de esperar entre 60 y 180 minutos.
3. Calcula la probabilidad de esperar entre 60 y 150 minutos.