Observaciones y medición

Antonio Falcó

Seminario 4

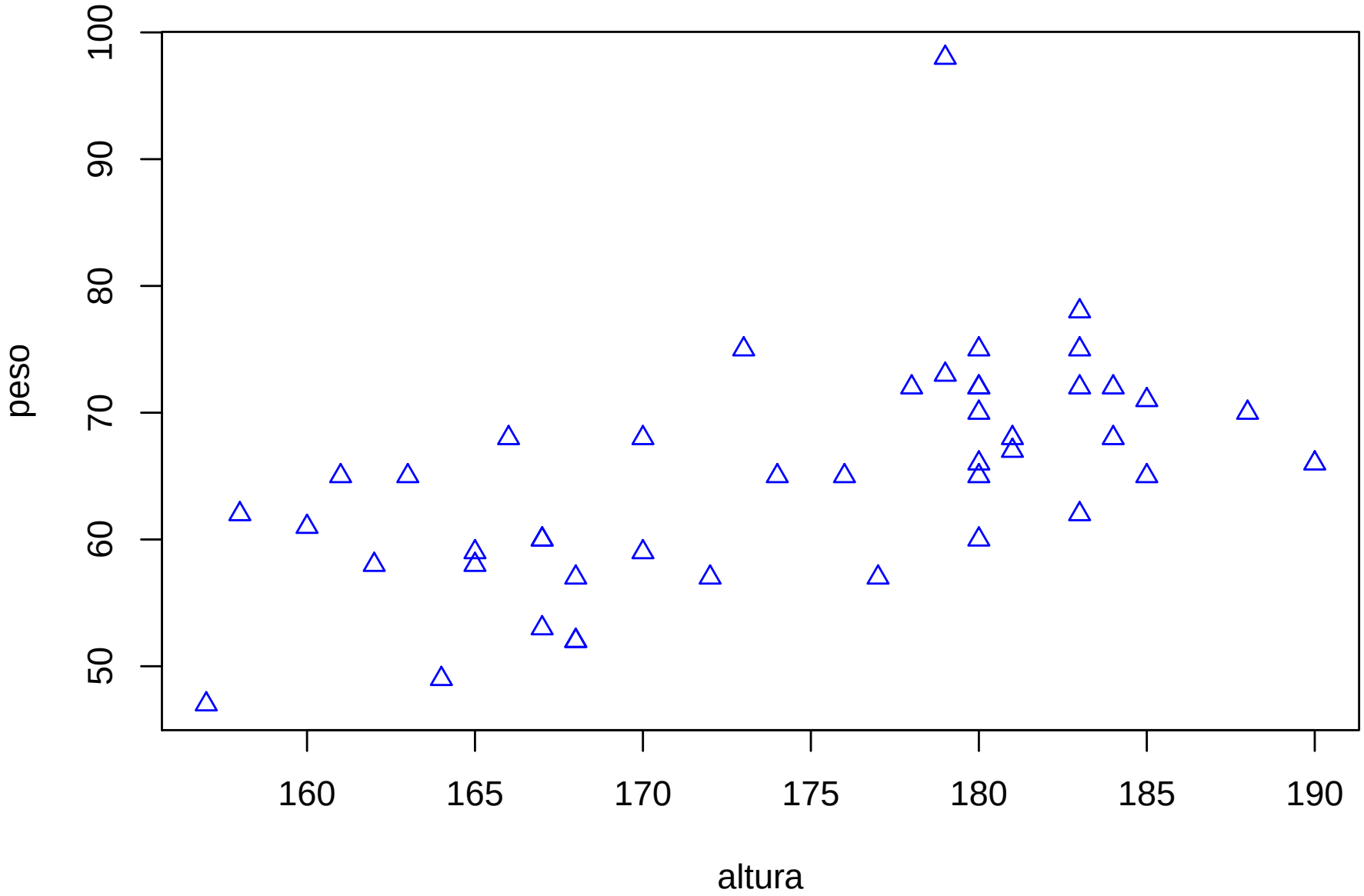
Consideremos las observaciones siguientes obtenidas de una población Ω que representan una muestra de la talla y el peso de 45 individuos, que podemos representar de forma conjunta en la tabla:

	altura	peso
1	180	70
2	177	57
3	180	60
4	180	66
5	183	62
6	184	68
7	185	65
8	184	72
9	174	65
10	180	72
11	168	52
12	180	75
13	183	75
14	181	68
15	180	65

	altura	peso
16	190	66
17	183	78
18	167	60
19	181	67
20	179	98
21	173	75
22	170	68
23	170	59
24	183	72
25	179	73
26	180	72
27	188	70
28	176	65
29	178	72
30	185	71

	altura	peso
31	168	52
32	157	47
33	167	53
34	168	57
35	163	65
36	167	60
37	166	68
38	164	49
39	172	57
40	165	59
41	158	62
42	161	65
43	160	61
44	162	58
45	165	58

Graficamente se puede representar como:



Cuestión

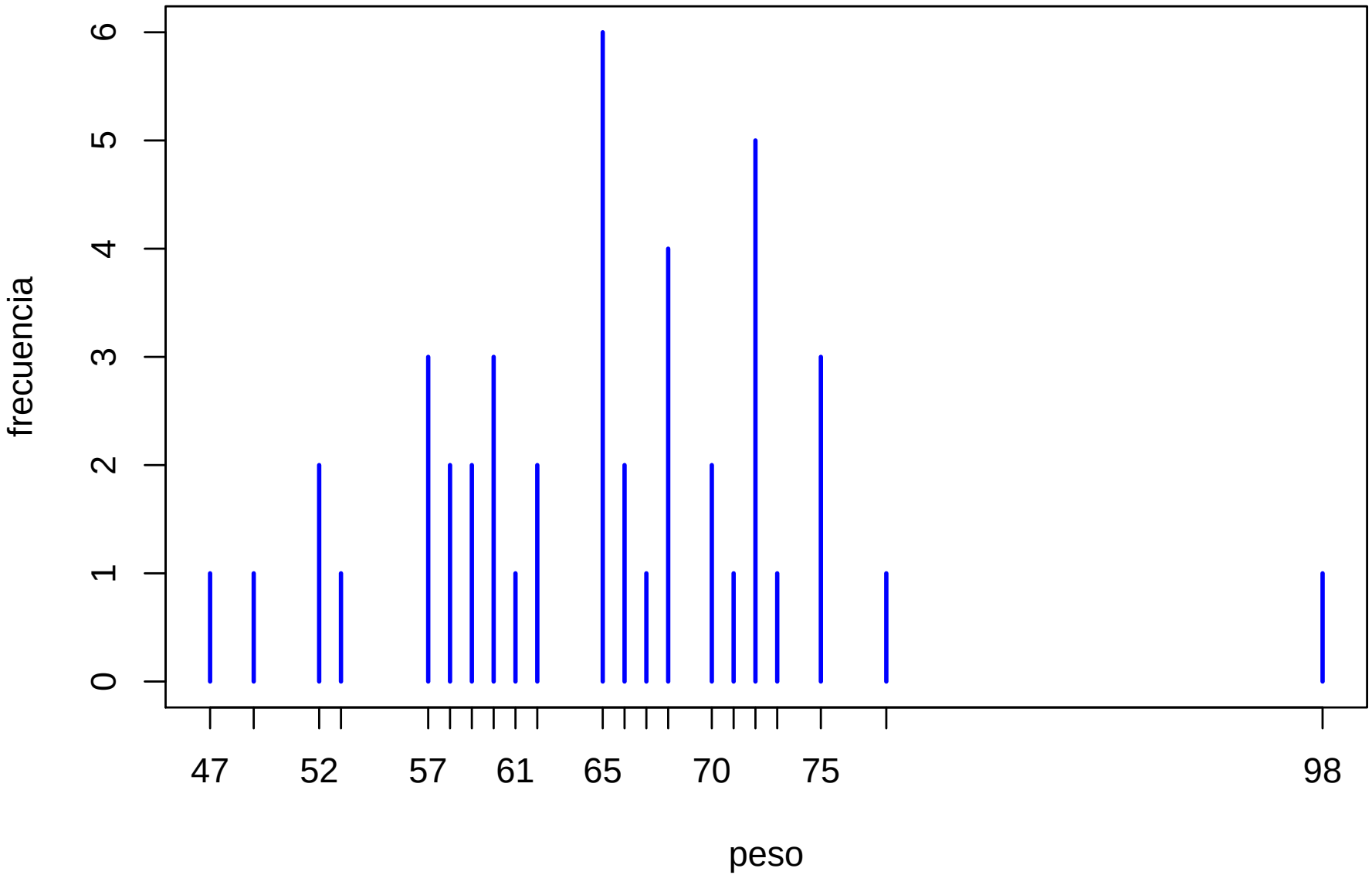
¿Podemos afirmar que en nuestra población (de la que se ha extraído esta muestra) las personas de mayor peso presentan una mayor altura con una **alta probabilidad**, y viceversa?

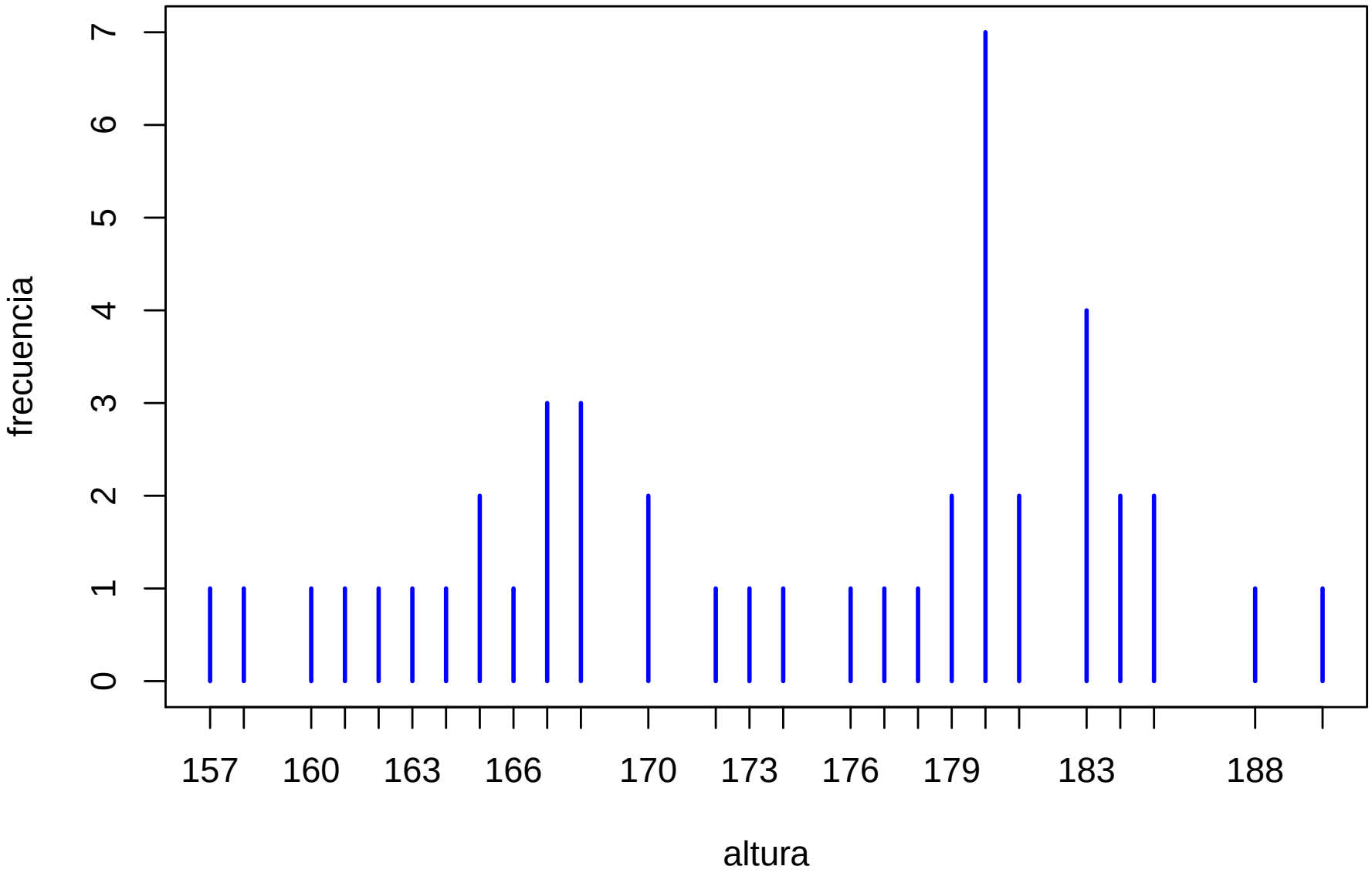
Cuestión científica

¿Cómo podemos definir el concepto de alta probabilidad y cómo lo podemos calcular en la práctica?

Probabilidad y observación experimental

- Una probabilidad $\mathbb{P}$ se define sobre los subconjuntos de una población experimental de individuos Ω.
- ¿Qué ocurre si deseamos estudiar características como el `peso` y la `altura` que afectan a cada individuo de la población de forma diferente?
- Por ejemplo el número de individuos que tienen un peso de 60 kg es de 3 sobre un total de 45, si intentamos representar los pesos o las alturas en una gráfica nos encontraremos la situación siguiente.


frecuencia
0
1
2
3
4
5
6
47
52
57
61
65
70
75
98
peso


frecuencia
0
1
2
3
4
5
6
7
157
160
163
166
170
173
176
179
183
188
altura

Cuestión científica

¿Cómo podemos relacionar características medibles (como el `peso` y la `altura`) de una población experimental Ω con la probabilidad $\mathbb{P}$ definida sobre ella?

Observaciones sobre una población

Supongamos que tenemos una población experimental de 5 individuos

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\},$$

entonces puedo obtener experimentalmente para cada individuo su peso y su altura escribiendo para cada elemento individual ω de la población experimental Ω su peso y altura como:

$$\texttt{peso}(\omega) = x \text{ kg}, \ \texttt{altura}(\omega) = y \text{ cm}.$$

Definición (variable aleatoria)

Sea Ω una población experimental, se llama **variable aleatoria** a cualquier asignación X definida sobre cada individuo ω de la población de forma que $X(\omega)$ **es el valor numérico de esta observación experimental**. Además, para cada individuo ω solo podemos obtener una única observación experimental $X(\omega)$.

Ejemplos

- $X = \texttt{peso}$.
- $X = \texttt{altura}$.
- $X = \texttt{temperatura}$.
- $X = \texttt{presion\_arterial}$.
- $X = \texttt{glucosa}$.

Variables aleatorias y probabilidades

- Sea Ω una población experimental y $\mathbb{P}$ una probabilidad definida sobre la misma.
- Sea X una variable aleatoria definida sobre Ω.

Entonces definimos la función de distribución F que relaciona la medida observacional X y P del modo siguiente:

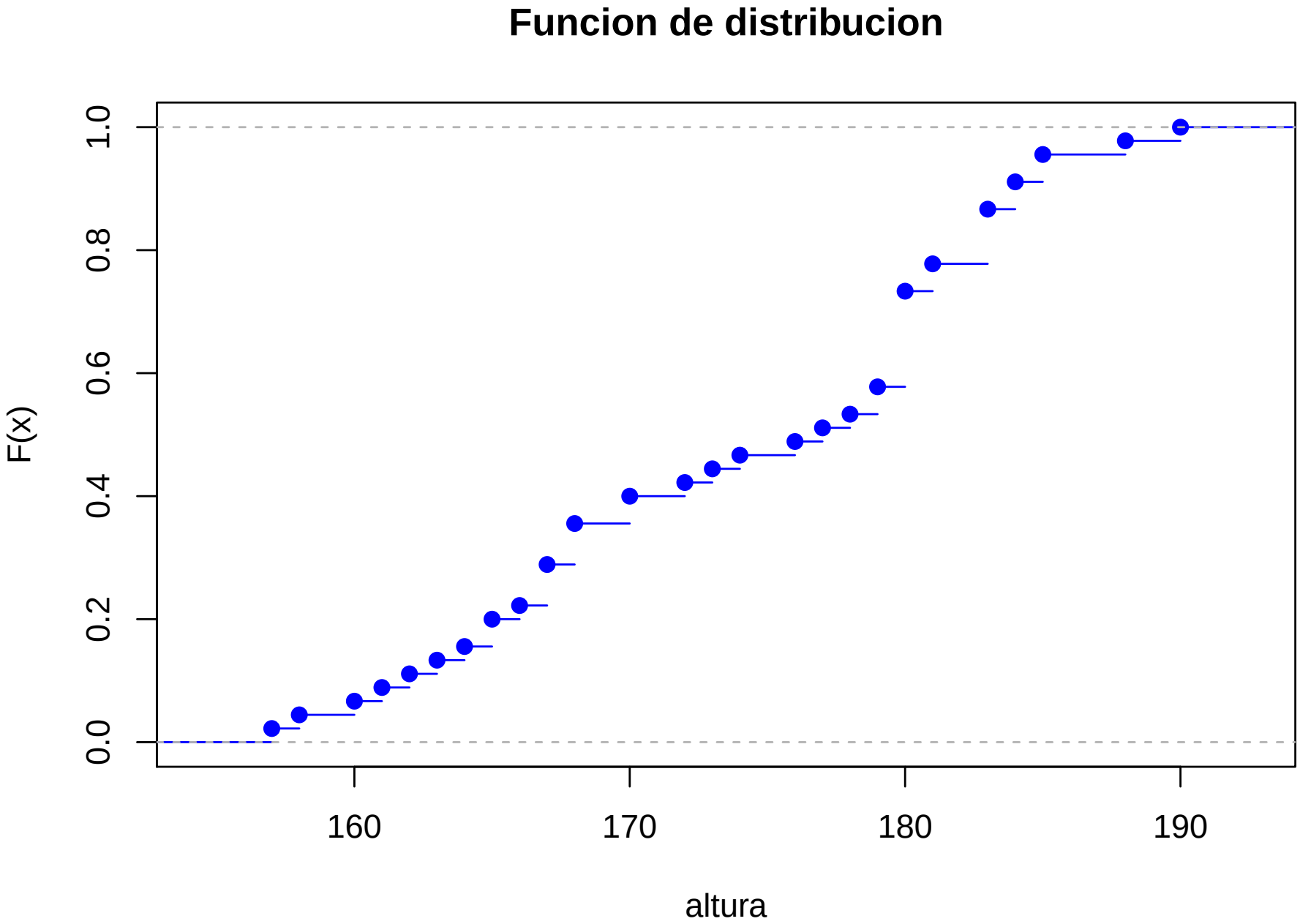
$$F(x) = \mathbb{P}(A_x)$$

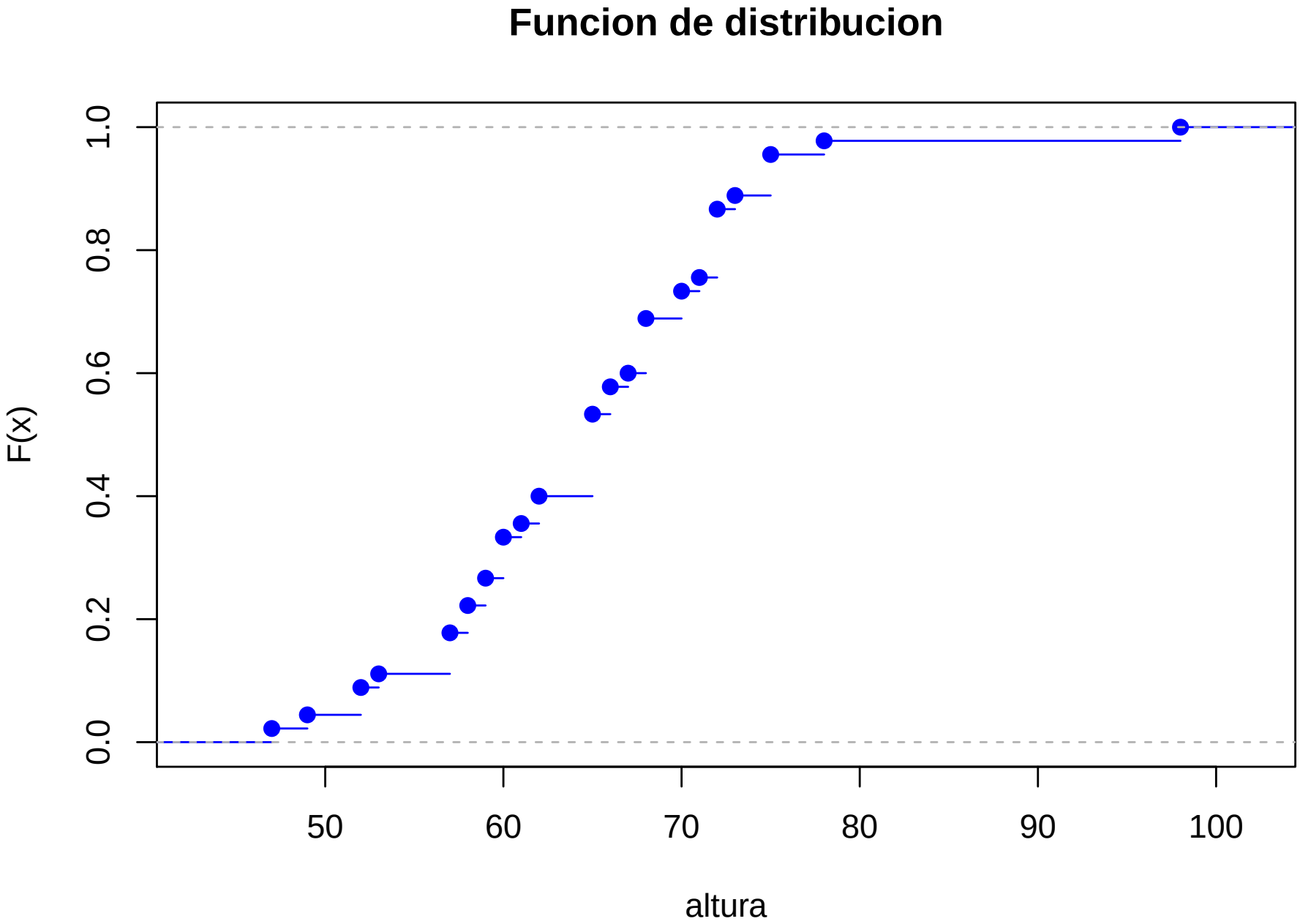
donde

$$A_x := \{\omega : X(\omega) \leq x\},$$

son los individuos de la población experimental susceptibles de tener una observación con un valor no superior a x, con los que lo podemos expresar como

$$F(x) = \mathbb{P}(\{\omega : X(\omega) \leq x\}) = \mathbb{P}(X \leq x)$$


Funcion de distribucion
F(x)
0.0
0.2
0.4
0.6
0.8
1.0
160
170
180
190
altura


Funcion de distribucion
F(x)
0.0
0.2
0.4
0.6
0.8
1.0
50
60
70
80
90
100
altura

Propiedad 1

- Sea Ω una población experimental y $\mathbb{P}$ una probabilidad definida sobre la misma.
- Sea X una variable aleatoria definida sobre Ω.
- Sea F la función de distribución para X en $\mathbb{P}$.

Entonces por construcción se cumple que

1. F es una función no decreciente, es decir, si $x < y$ entonces $F(x) \leq F(y)$
2. $F(-\infty) = 0$ y $F(\infty) = 1$ y
3. Se cumple
$$F(x-) \leq F(x) = F(x+),$$
(*fonction cadlag:* continue à droite, limite à gauche)

Definición

1. Si la función F **no tiene saltos** se dice que la variable X es **continua**.
2. Si la función F **tiene forma de escalera**, entonces se dice que la variable X es **discreta**.

Cuestión

En las clases de teoría se ha visto que la variable `altura` es **continua**, pero ahora según esta definición diríamos que la variable `altura` es **discreta**, ya que la función $F(x)$ que obtenemos es una función escalera. ¿Qué es lo que está ocurriendo?

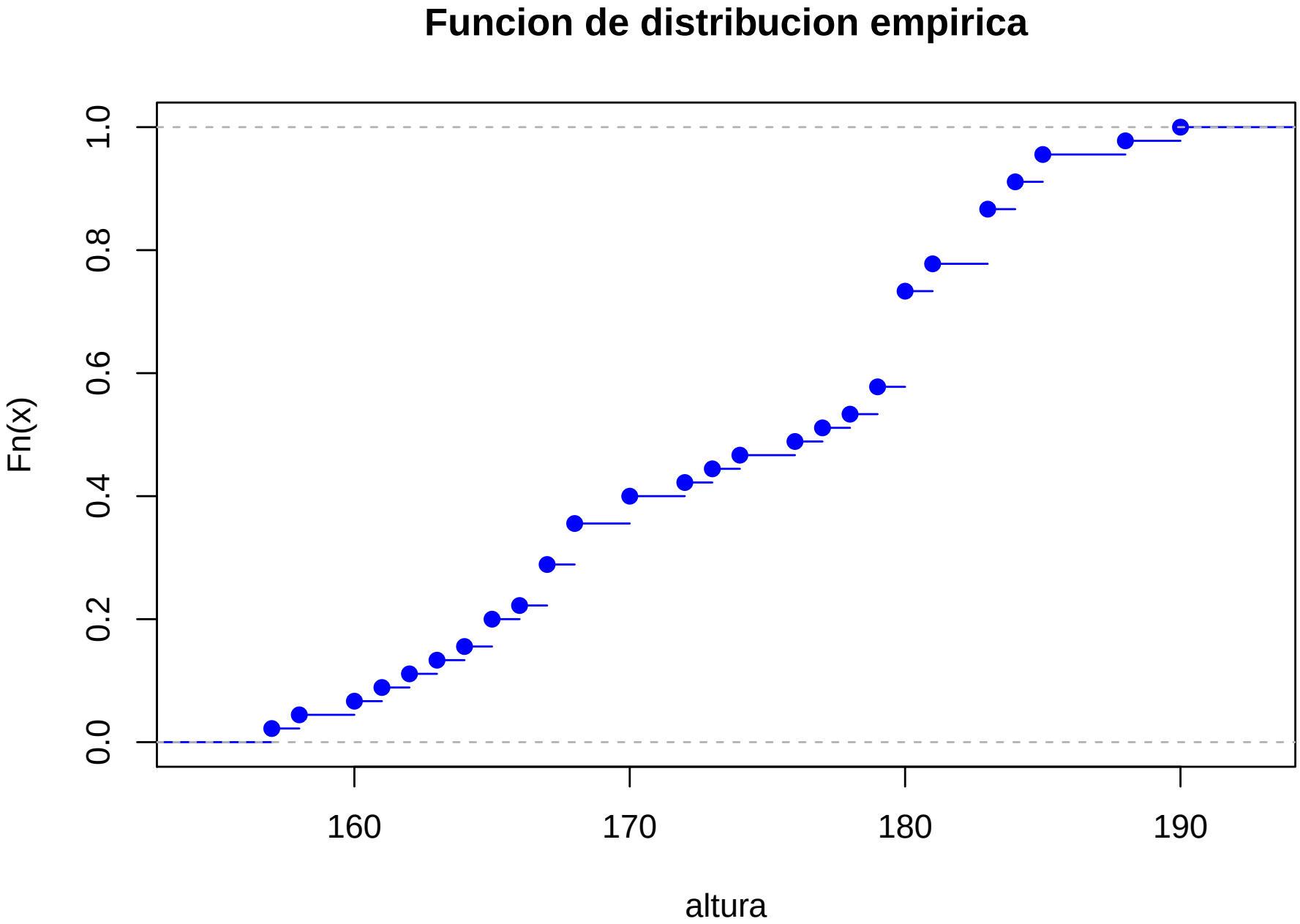
Respuesta

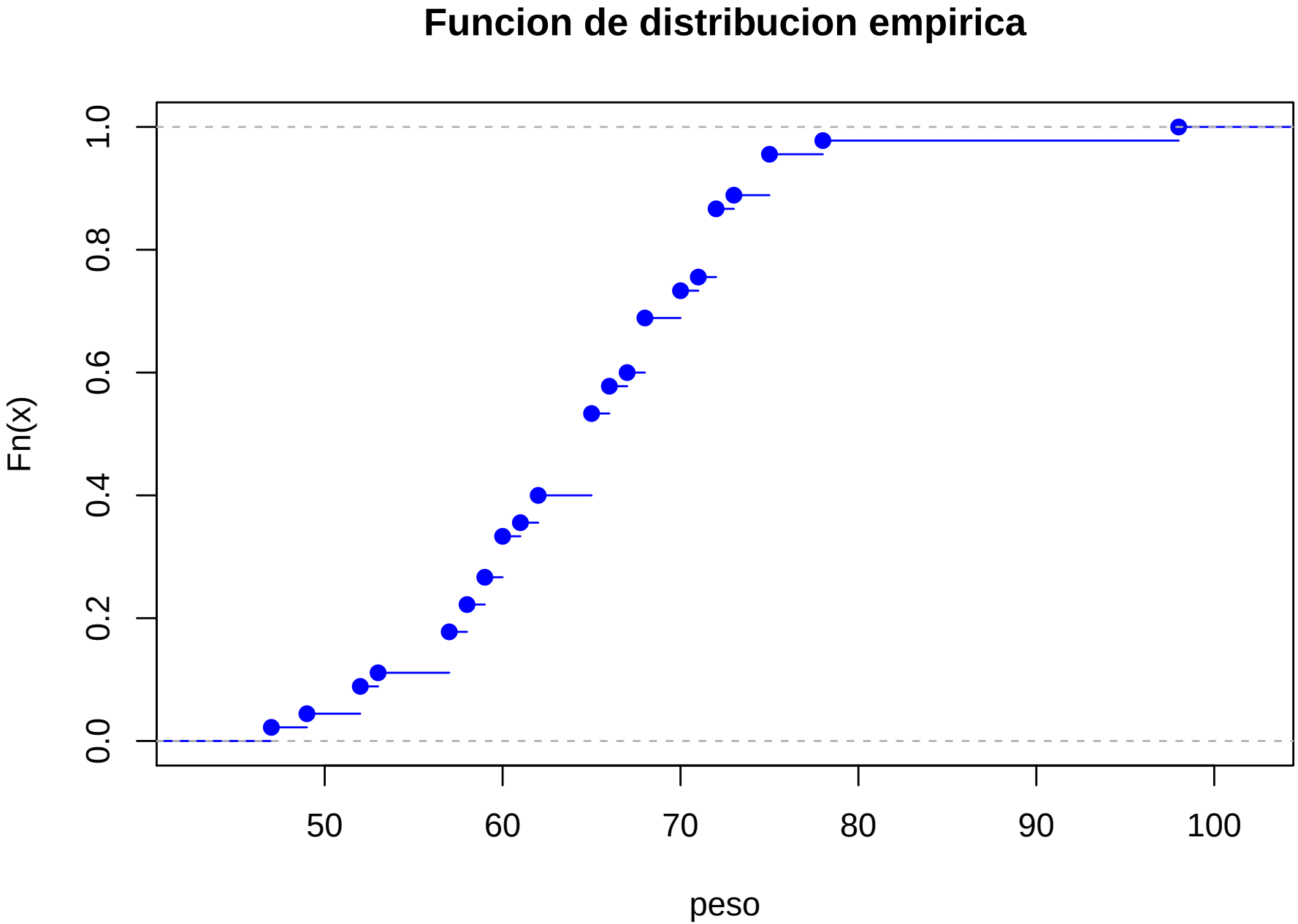
La función F la calculamos empleando solo una muestra Ω_n de una población mayor Ω, en consecuencia, podemos denotar a esta función como F_n donde el subíndice indica que proviene de una muestra y la llamamos **Función de distribución empírica.** En realidad, tendríamos

$$F_n(x) = \mathbb{P}(X \leq x|\Omega_n) \approx \mathbb{P}(X \leq x) = F(x).$$

Recordemos que (empleando el Teorema de Bayes)

$$\begin{aligned} F(x) = \mathbb{P}(X \leq x) &= \mathbb{P}(X \leq x|\Omega_n)\mathbb{P}(\Omega_n) + \mathbb{P}(X \leq x|\overline{\Omega}_n)\mathbb{P}(\overline{\Omega}_n) \\ &= F_n(x)\mathbb{P}(\Omega_n) + \mathbb{P}(X \leq x|\overline{\Omega}_n)(1 - \mathbb{P}(\Omega_n)). \end{aligned}$$


Funcion de distribucion empirica
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
160
170
180
190
altura


Funcion de distribucion empirica
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
50
60
70
80
90
100
peso

Empleamos para construirla en la práctica la función de R `ecdf()` *empirical cumulative distribution function*

```
Fn <- ecdf(datos$altura)
```

Ejemplo

Si consideramos $X = \texttt{altura}$ en cm entonces podemos calcular

$$F_n(165) = 0.2 = \mathbb{P}\left(\texttt{altura} \leq 165\right),$$

$$F_n(175) = 0.4666667 = \mathbb{P}\left(\texttt{altura} \leq 175\right),$$

$$F_n(185) = 0.9555556 = \mathbb{P}\left(\texttt{altura} \leq 185\right),$$

$$F_n(195) = 1 = \mathbb{P}\left(\texttt{altura} \leq 195\right),$$

lo que nos indica que el 100% de los individuos de nuestra población experimental tienen una altura inferior o igual al 195 cm.

Cuestión

¿Cuál es la probabilidad de encontrar en esta población experimental un indivíduo con una altura de entre 165 y 185 cm?

Respuesta

$$
\begin{aligned}
\mathbb{P}(165 < X \leq 185) &= \mathbb{P}(\{X > 165\} \cap \{X \leq 185\}) \\
&= \mathbb{P}(\overline{\{X \leq 165\}} \cap \overline{\{X > 185\}}) \\
&= \mathbb{P}(\overline{\{X \leq 165\} \cup \{X > 185\}}) \\
&= 1 - \mathbb{P}(\{X \leq 165\} \cup \{X > 185\}) \\
&(\{X \leq 165\} \cap \{X > 185\} = \emptyset) \\
&= 1 - \mathbb{P}(\{X \leq 165\}) - \mathbb{P}(\{X > 185\}) \\
&= 1 - \mathbb{P}(\{X > 185\}) - F_n(165) \\
&= \mathbb{P}(\overline{\{X > 185\}}) - F_n(165) \\
&= \mathbb{P}(\{X \leq 185\}) - F_n(165) = F_n(185) - F_n(165) \\
&= 0.7555556 = 75.5555556\%.
\end{aligned}
$$

Propiedad 2

- Sea Ω una población experimental y $\mathbb{P}$ una probabilidad definida sobre la misma.
- Sea X una variable aleatoria definida sobre Ω.
- Sea F la función de distribución para X en $\mathbb{P}$.

Entonces para cada par de números $a \leq b$ se cumple

$$\mathbb{P}(a < X \leq b) = F(b) - F(a).$$

Cuestión

¿Que relación tiene la función de distribución empírica F_n de la variable `altura` con la **verdadera** función de distribución F?

Respuesta

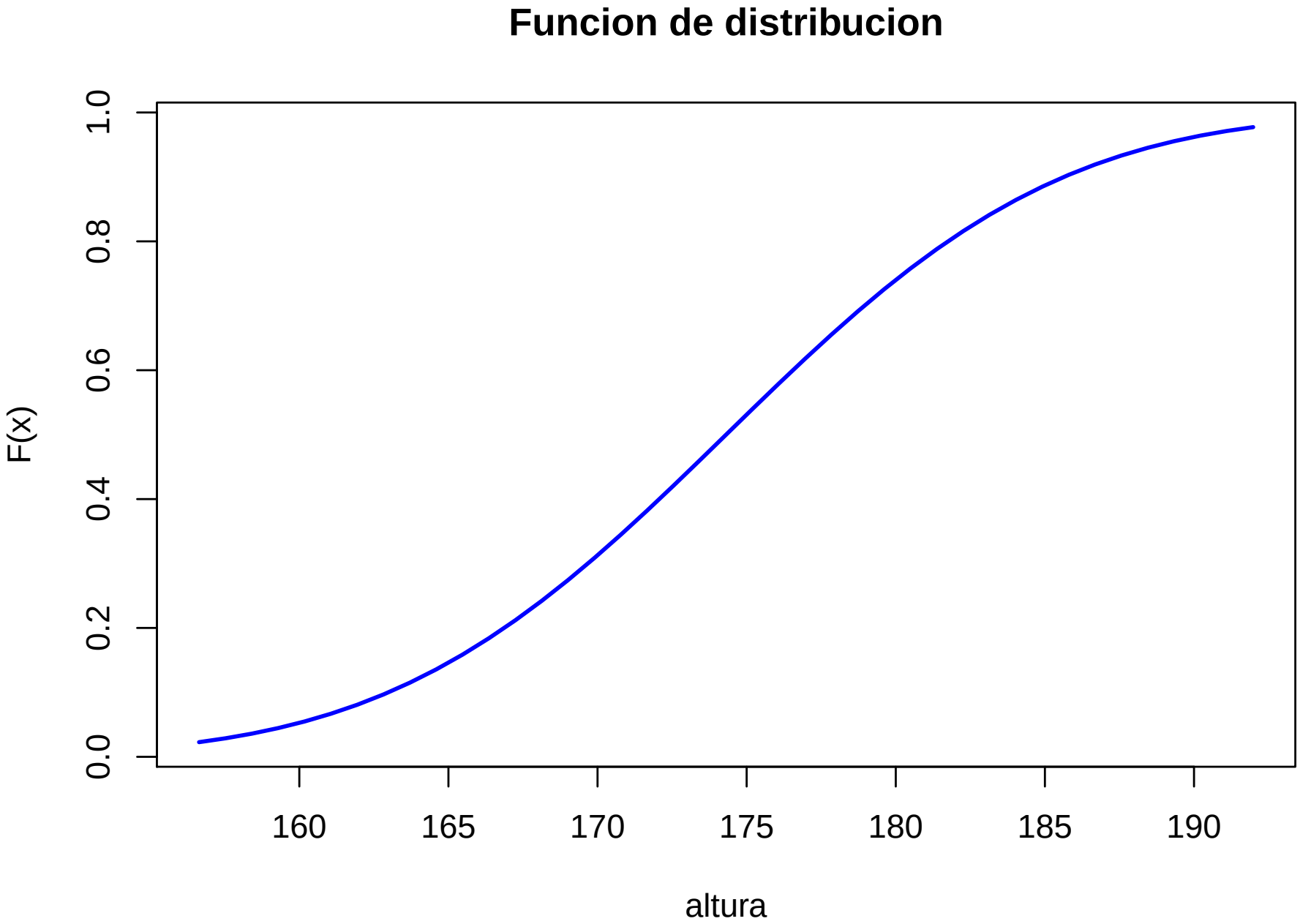
Calculemos la media

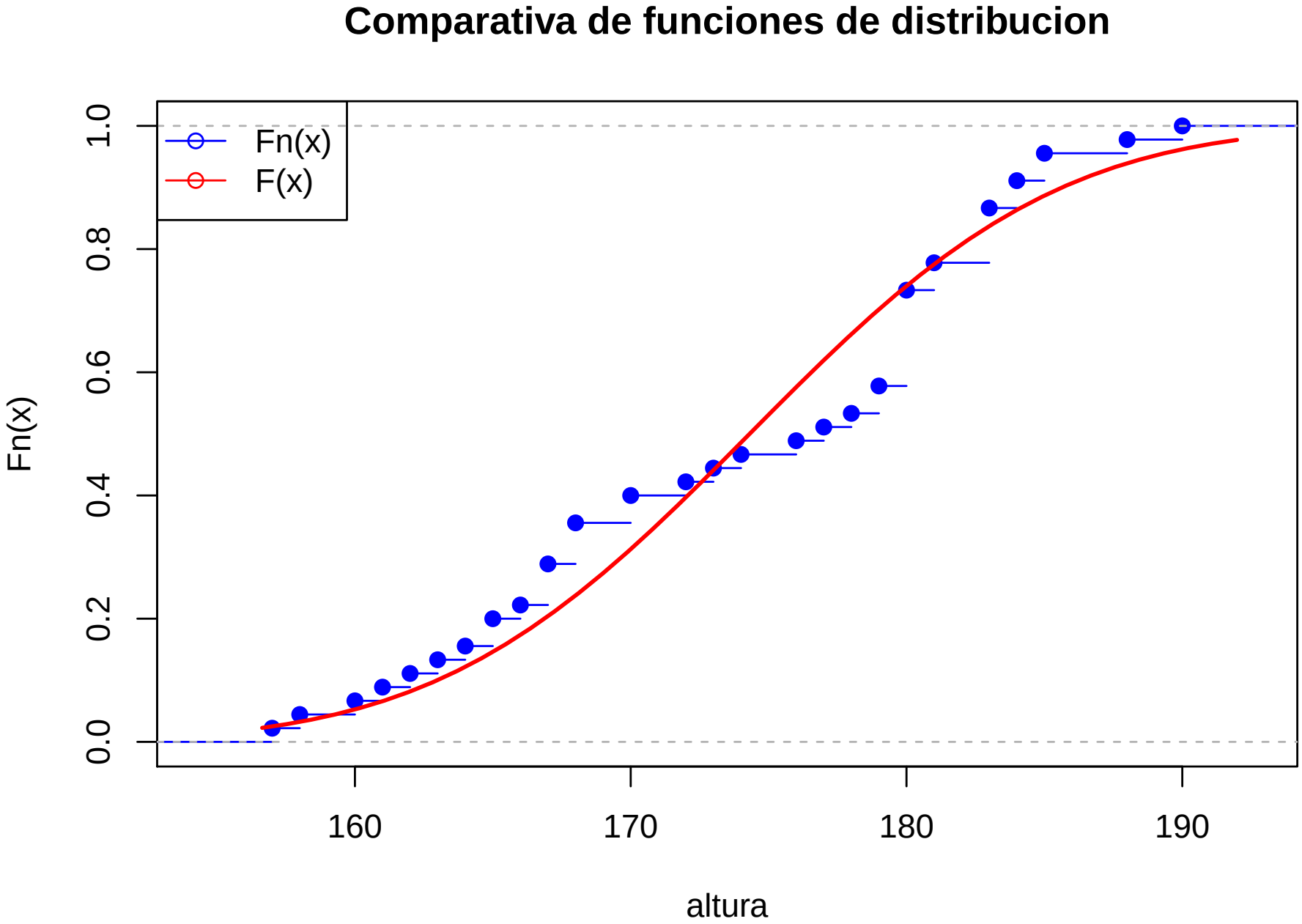
$$\mu = \mathbb{E}(\texttt{altura}) = 174.3111111 \text{ cm},$$

y la desviación típica

$$\sigma = \sqrt{\text{Var}(\texttt{altura})} = 8.8364489 \text{ cm},$$

de la variable altura y supongamos que la variable `altura` sigue una distribución normal $\mathcal{N}(174.3111111, 8.8364489)$, podemos entonces visualizar la función de distribución de esta variable aleatoria normal.


Funcion de distribucion
F(x)
0.0
0.2
0.4
0.6
0.8
1.0
160
165
170
175
180
185
190
altura


Comparativa de funciones de distribucion
Fn(x)
F(x)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
160
170
180
190
altura

Discusión

Podemos considerar la función de distribución empírica de la variable `altura` $F_n(x)$ como una **aproximación** de la función de distribución (verdadera) $F(x)$ de la variable `altura`.

Interpretación

Conocemos mediante la función de distribución empírica que

$$0.7555556 = F_n(185) - F_n(165) \approx F(185) - F(165) = \mathbb{P}(165 < \texttt{altura} \leq 185).$$

Además podemos calcular la precisión de esta aproximación ya que empleando la distribución normal de la `altura` como $\mathcal{N}(174.3111111, 8.8364489)$ podemos calcular

$$\mathbb{P}(165 < \texttt{altura} \leq 185) = F(185) - F(165) = 0.7407844.$$

y obtener el error de aproximación

$$|0.7555556 - 0.7407844| = 0.0147712$$

Conclusión

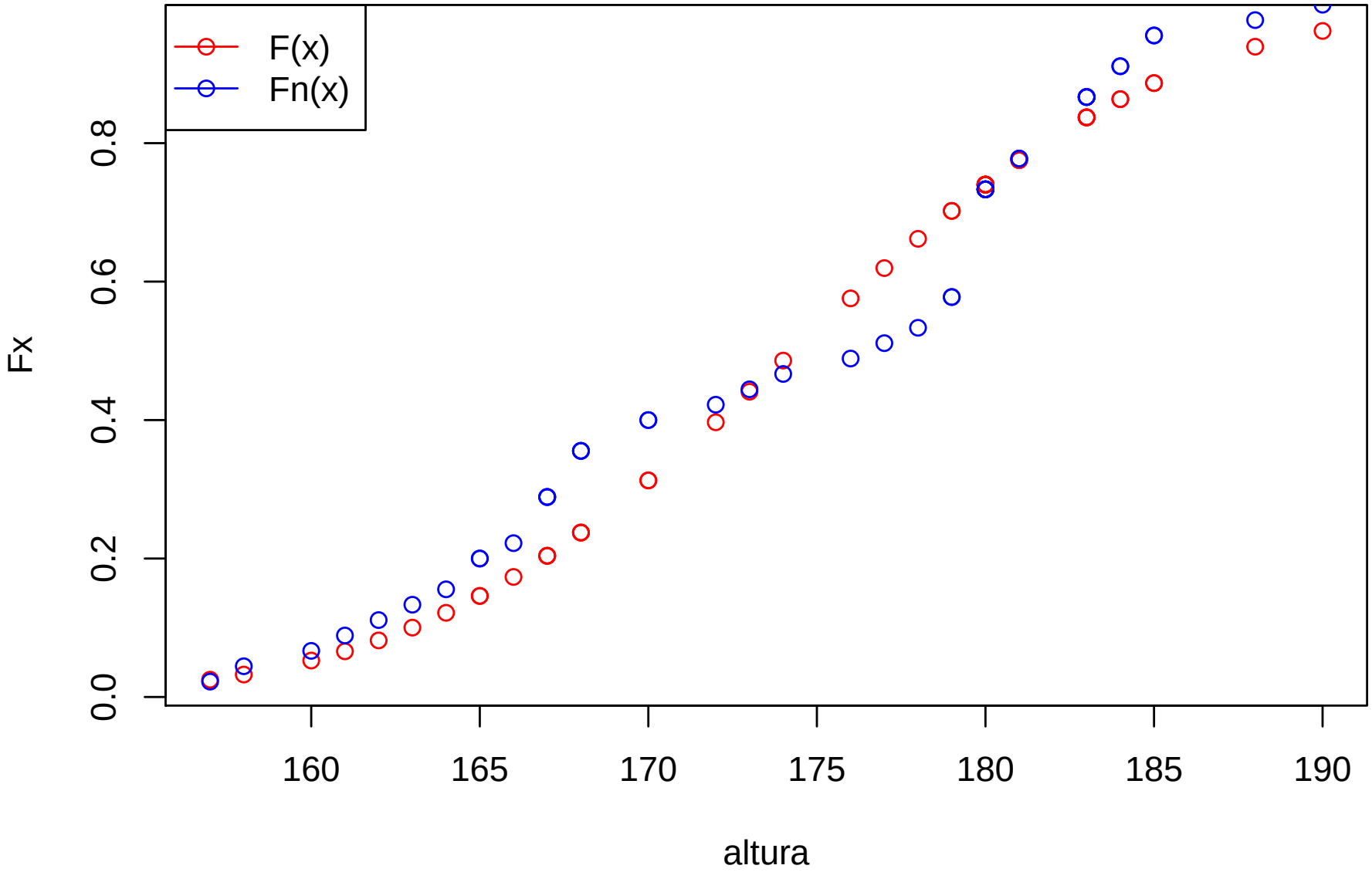
- Sea Ω una población experimental y $\mathbb{P}$ una probabilidad definida sobre la misma.
- Sea X una variable aleatoria definida sobre Ω.
- Sea F la función de distribución para X en $\mathbb{P}$.

Con los datos obtenidos mediante observaciones de una variable de interés X sobre una muestra Ω_n de la población experimental Ω construimos la función de distribución empírica

$$F_n(x) = \mathbb{P}(x \leq X|\Omega_n) \approx \mathbb{P}(x \leq X) = F(x),$$

como aproximación de la poblacional $F(x)$.

Evaluamos F y F_n sobre los 45-valores de la muestra:



altura	Fx	Fnx	diferencia	diferencia_cuadrado
180	0.0250530	0.0222222	0.0028308	0.0000080
177	0.0324541	0.0444444	-0.0119903	0.0001438
180	0.0526640	0.0666667	-0.0140027	0.0001961
180	0.0659839	0.0888889	-0.0229050	0.0005246
183	0.0817767	0.1111111	-0.0293344	0.0008605
184	0.1002635	0.1333333	-0.0330699	0.0010936
185	0.1216286	0.1555556	-0.0339270	0.0011510
184	0.1460064	0.2000000	-0.0539936	0.0029153
174	0.1460064	0.2000000	-0.0539936	0.0029153
180	0.1734681	0.2222222	-0.0487541	0.0023770
168	0.2040105	0.2888889	-0.0848784	0.0072043
180	0.2040105	0.2888889	-0.0848784	0.0072043
183	0.2040105	0.2888889	-0.0848784	0.0072043
181	0.2375476	0.3555556	-0.1180080	0.0139259
180	0.2375476	0.3555556	-0.1180080	0.0139259

Calculamos el error cuadrático medio entre la función empírica y la normal en cada uno de los valores de altura observados:

$$\frac{1}{45}\sum_{i=1}^{45}(F_n(x_i) - F(x_i))^2 = 0.0039835.$$

Esto nos permite afirmar que la variable `altura` se comporta aproximadamente como una distribución normal.