Probabilidad

Antonio Falcó

Seminario 2

Definición

La epistemología, del griego episteme (conocimiento) y lógos (estudio), es la rama de la filosofía que estudia el conocimiento científico, su naturaleza, posibilidad, alcance y fundamentos.

- La epistemología estudia las circunstancias históricas, psicológicas y sociológicas que llevan a la obtención del conocimiento científico y los criterios por los cuales se lo justifica o invalida, así como la definición clara y precisa de los conceptos epistémicos más usuales, tales como verdad, objetividad, realidad o justificación.

Cuestión

¿Cuál es el objeto de la estadística dentro de la ciencia? o bien ¿qué papel juega la estadística dentro del conocimiento científico?

La validación científica a través de la experimentación

Pilares del método científico

- La falsabilidad o refutabilidad es la capacidad de una teoría o hipótesis de ser sometida a potenciales pruebas que la contradigan.
- La reproducibilidad de es la capacidad de una teoría científica de ser refutada experimentalmente.

Caso a estudio

Teoría científica: Identificación y cuantificación de factores de riesgo de una enfermedad.

Definición

Se llama **prevalencia** a la cantidad de individuos de una población susceptibles de padecer una determinada enfermedad en un instante de tiempo dado. Se llama **prevalencia relativa** a la proporción de indivíduos de la población susceptibles de padecer una determinada enfermedad en un instante de tiempo dado.

Cuestión

¿Cómo podemos conocer la prevalencia de una determinada enfermedad?

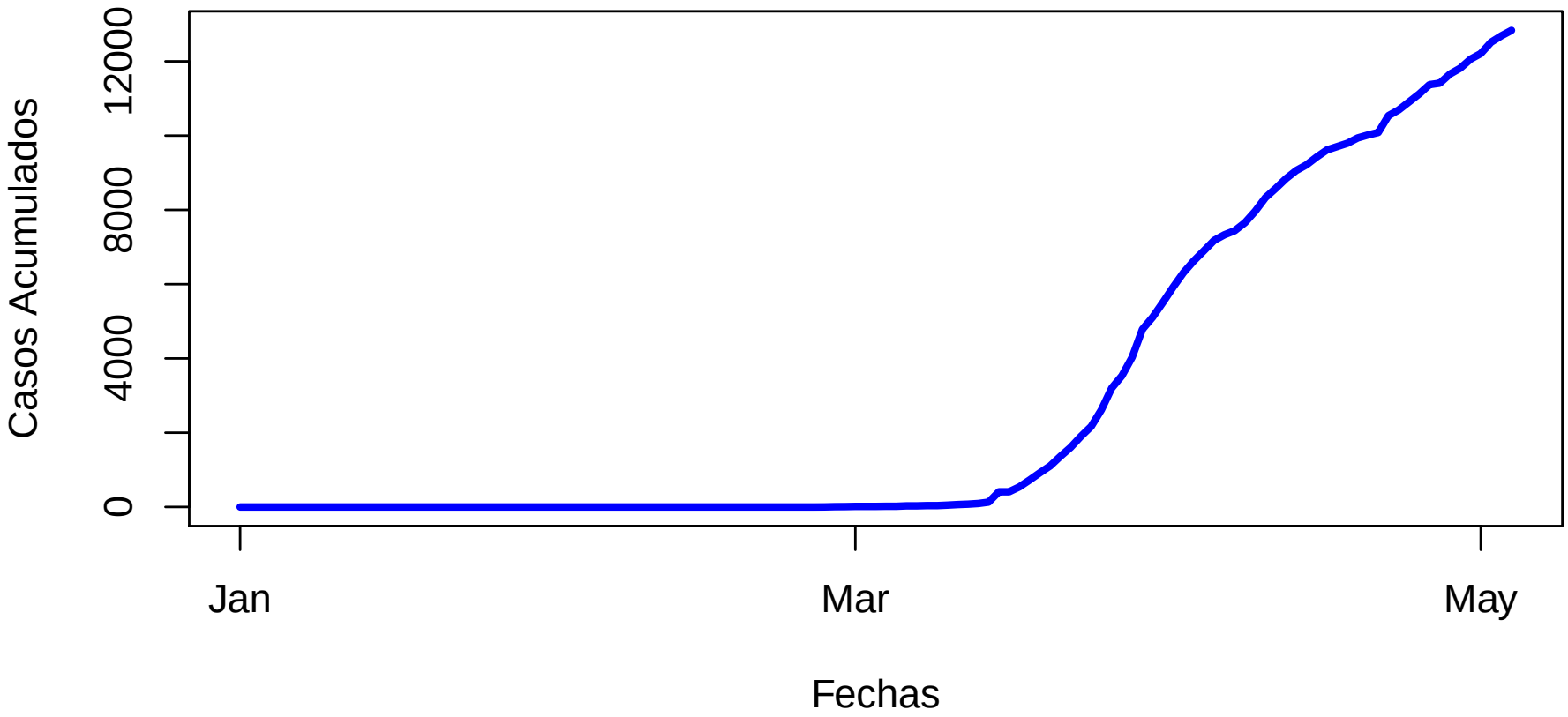
Respuesta

Mediante la experimentación, en particular, recolectando datos de enfermos declarados de dicha enfermedad.

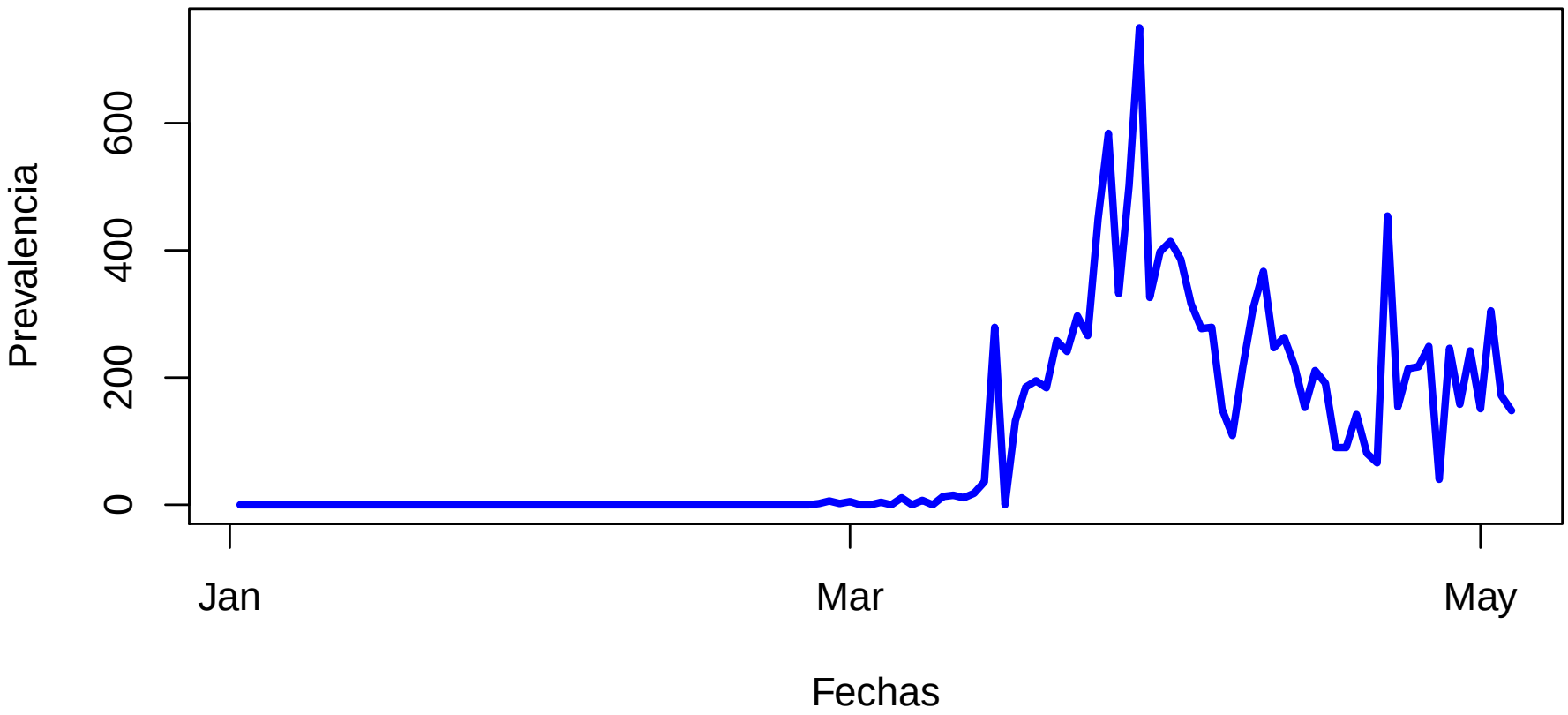
Cuestión

Imaginemos que calculamos esa cifra, ¿podemos afirmar es una cantidad objetiva y no subjetiva?

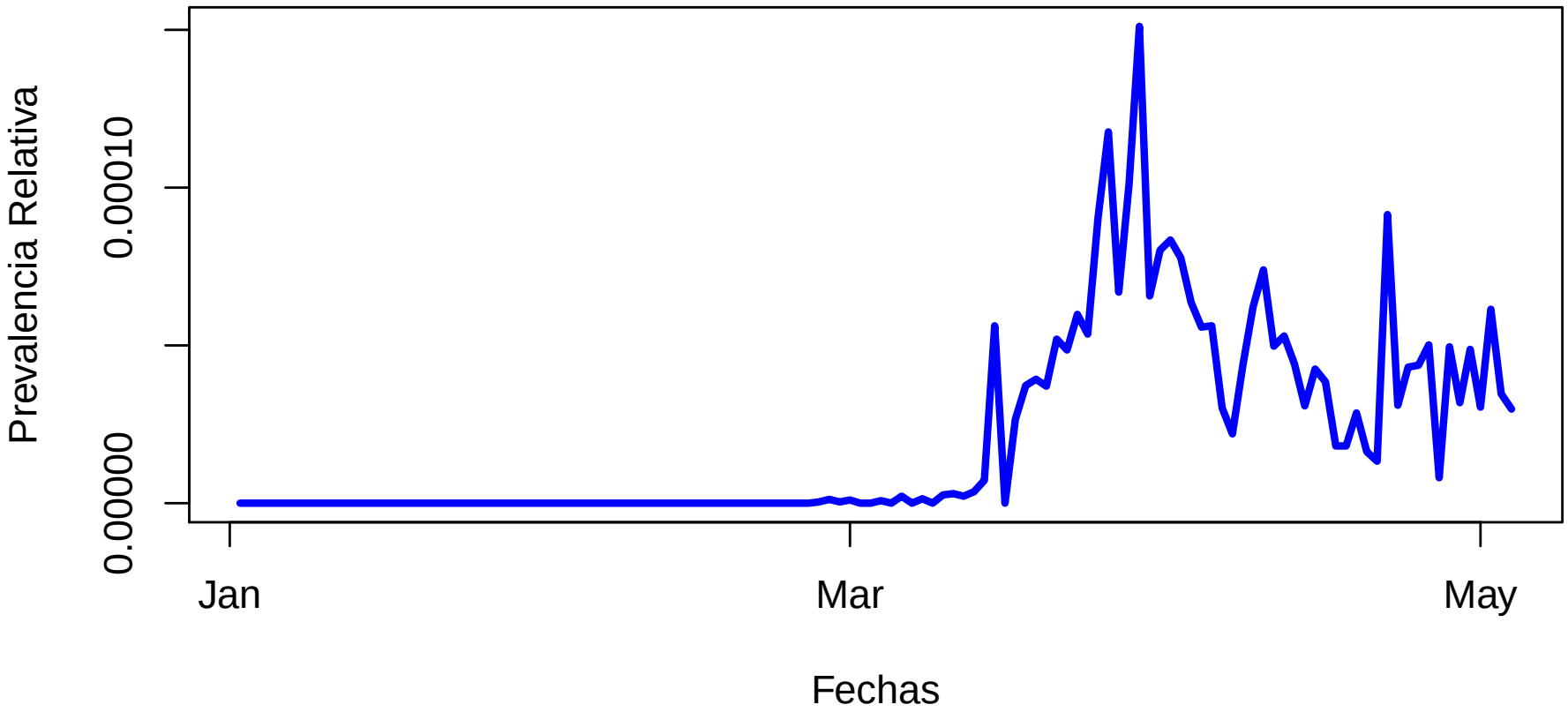
Consideremos las cifras de casos acumulados del COVID19 en la Comunidad Valenciana que tiene 4963703 habitantes.



Para calcular la prevalencia diaria calculamos los casos diarios, y construimos la curva de prevalencia diaria



Ahora dividimos los casos por el número de habitantes de la Comunidad Valenciana:



El valor máximo de la prevalencia relativa (diaria) es de 1.5109687×10^{-4}, lo que equivale a unos 15 casos cada 100000 indivíduos y se alcanzo en la fecha 2020-03-29

Definición

Se denomina **incidencia** anual (mensual, diaría) al número de nuevos casos de la enfermedad aparecidos a lo largo de un año (mes, día). Se llama **incidencia relativa** anual (mensual, diaría) a la proporción de nuevos casos de la enfermedad aparecidos a lo largo de un año (mes, día)

Ejemplo

La incidencia mensual en Enero fué de 0, en Febrero de 8, en Marzo de 5500 y en Abril de 6550 casos. La incidencia relativa fue en Enero de 0, en Febrero de 1.6117×10^{-6}, en Marzo de 0.001108 y en Abril de 0.0013196.

Regla habitual

La prevalencia y la incidencia se expresan habitualmente en **número de casos por cada 100000 habitantes por unidad de tiempo**. Por ejemplo, en Abril el COVID19 en la Comunidad Valenciana tuvo una incidencia (mensual) de 132 casos por cada 100000 habitantes. Este número se obtiene multiplicando la incidencia relativa del mes de Abril $0.0013196 \times 100000 = 131.9579354$ y redondeando la cifra.

Tasa de ataque

Se denomina tasa de ataque de una enfermedad a la proporción de la población que contrae la enfermedad durante una epidemia:

$$\text{TA} = \frac{\text{numero de nuevos casos en un periodo de tiempo}}{\text{número de personas en riesgo durante el mismo periodo de tiempo}}$$

Ejemplo

En los meses de Enero-Abril el número de casos en la Comunidad Valenciana ha sido de 12058 casos y el número de personas expuestas ha sido de 4.963703×10^6, en consecuencia la tasa de ataque del primer cuatrimestre de 2020 ha sido de

$$\frac{12058}{4.963703 \times 10^6} = 0.0024292,$$

es decir del 2.4292348 por 1000 de la población a contraido el COVID19.

Factores de riesgo de una enfermedad

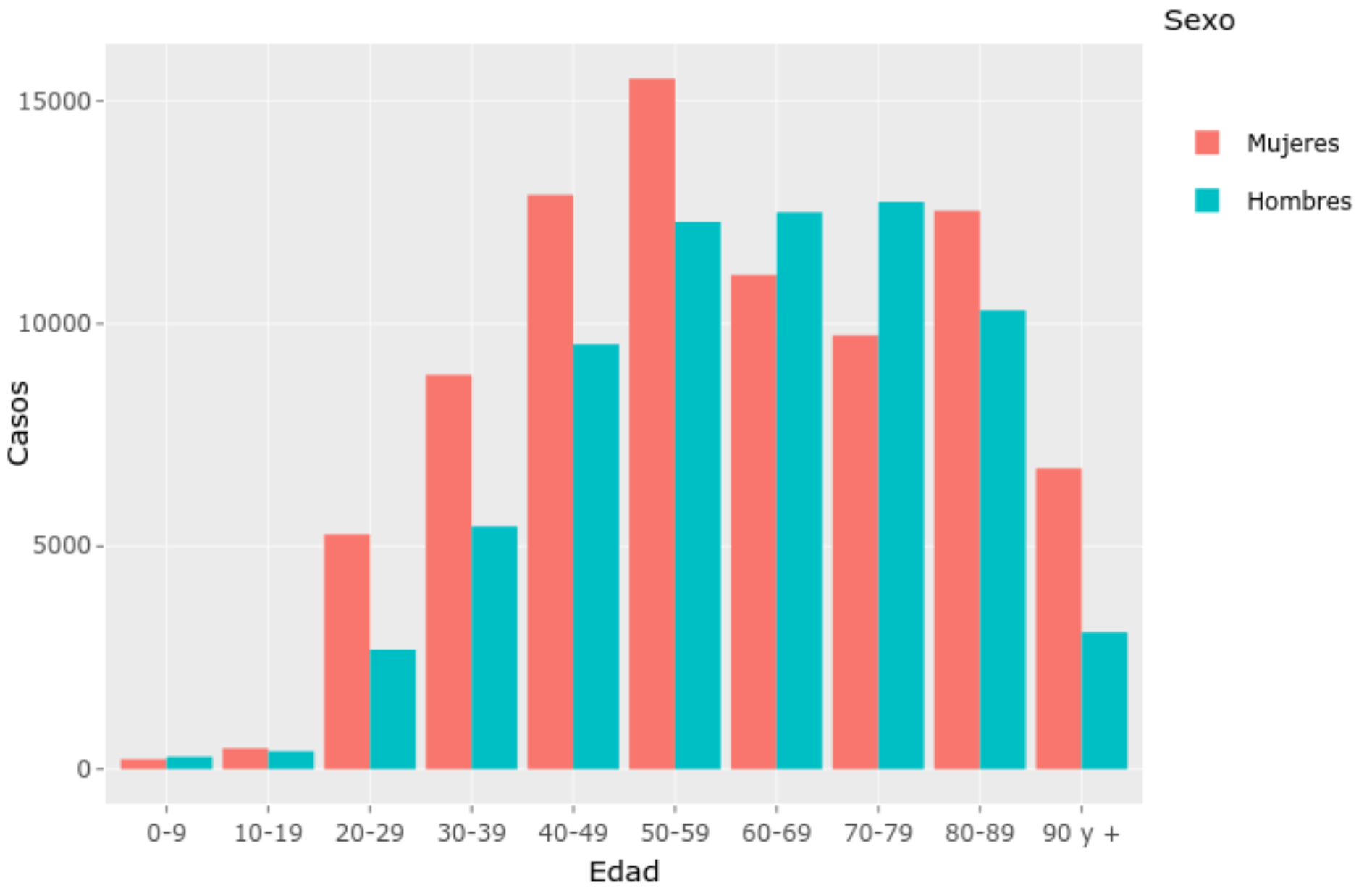
Ejemplo

Durante la pandemia producida por el COVID19 hemos oido hablar de los factores de riesgo de la enfermedad. Por ejemplo un factor de riesgo que hemos oido nombrar es que el **tener una edad avanzada,** por ejemplo superior a 70 años de edad, es un factor de riesgo para el COVID19.

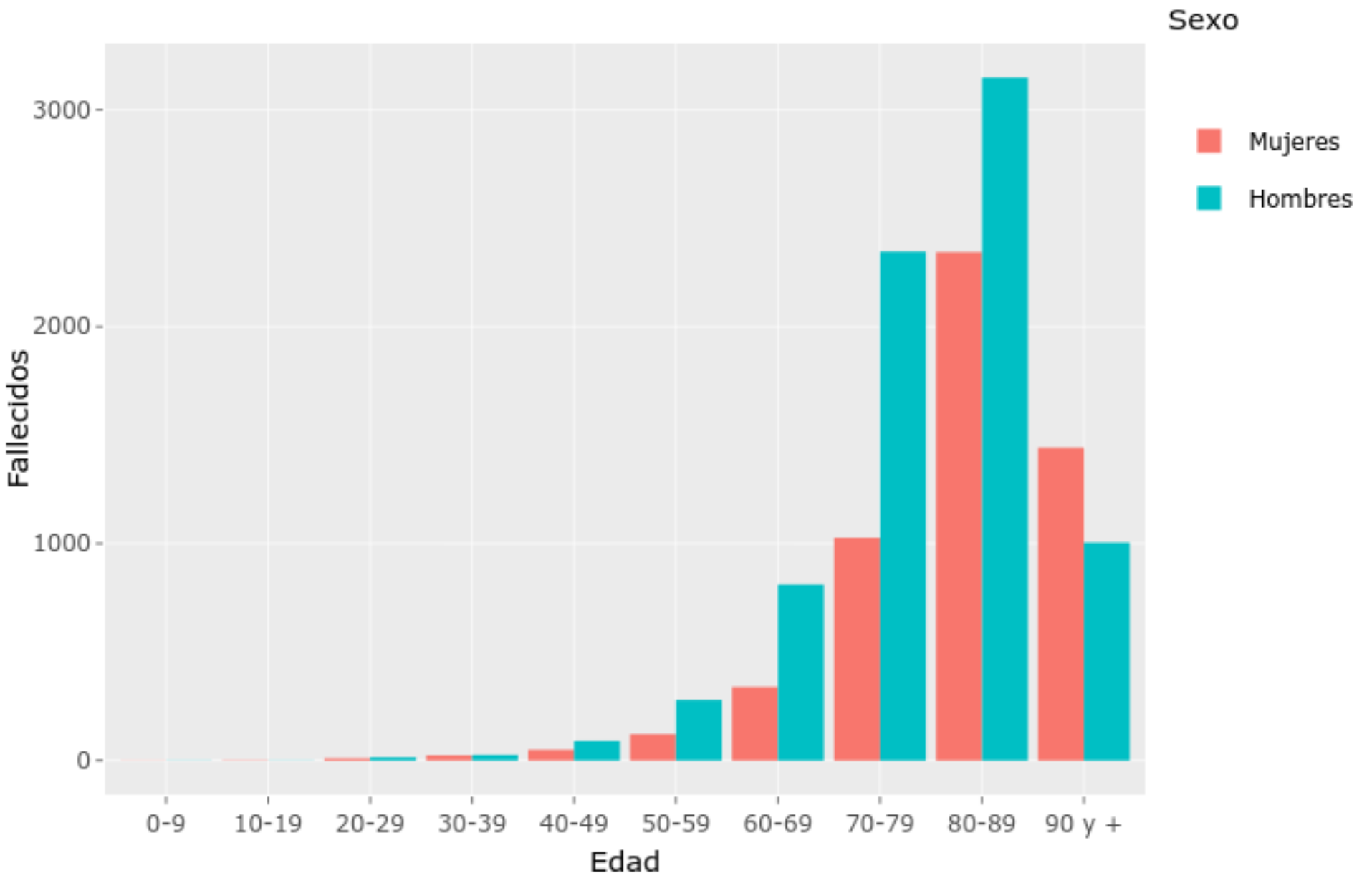
Cuestión

¿Podemos demostrar la validez científica de esta afirmación?

Datos España



Datos España



Conclusiones

1. En el primer gráfico no vemos que el número de casos de COVID19 entre las personas de más de 70 años sea significativo (en valores absolutos).
2. Lo que si podemos deducir del segundo gráfico es que la tasa de mortalidad dentro de este grupo de personas es (en términos absolutos) significativamente superior.
3. En consecuencia, podemos afirmar que es un grupo de riesgo (y no un factor de riesgo) debido a que si una persona de este grupo (mayor de 70 años) contrae la enfermedad, podemos intuir visualizando la gráfica anterior, que las **probabilidades** de fallecer son muy altas.

Primer ejemplo de causalidad

¿El contraer el COVID19 con una edad superior a 70 años es causa de un aumento en el riesgo de fallecimiento (aumento de la probabilidad de fallecer)? Volveremos sobre esta cuestión más adelante.

Cuestiones

- ¿Cómo se calculan estas probabilidades en casos reales?, pero antes de nada:
- ¿Qué es una probabilidad?

Definición de probabilidad frecuentista

Tenemos una población de individuos que denotamos por Ω y para cada subconjunto de individuos A de la población susceptibles de compartir una propiedad $\mathcal{P}$, definimos una asignación

$$\mathbb{P}(A) := \begin{array}{l}\text{proporción de individuos susceptibles} \\ \text{de tener la propiedad } \mathcal{P} \text{ en la población } \Omega.\end{array}$$

es decir,

$$\mathbb{P}(A) = \frac{\textbf{número de individuos en } \Omega \textbf{ que tienen la propiedad } \mathcal{P}}{\text{número total de individuos de } \Omega}$$

Problema experimental

No podemos conocer en la mayoría de casos la cantidad de personas de la población que comparten una determinada propiedad.

Ejemplo 1

- Consideremos Ω la población de la Comunidad Valenciana,
- Ω contiene 4.963703×10^6-individuos (en el lenguaje probabilista se les llama sucesos elementales ya que son las unidades más pequeñas en las que podemos dividir la población) .
- Consideremos el conjunto de individuos A de la Comunidad Valenciana que cumplen la propiedad $\mathcal{P}$ siguiente: **estar reportado oficialmente como contagiado por el COVID19 en la fecha 2020-03-25**.
- Conocemos que habían 2616 individuos reportados oficialmente como contagiados en esa fecha, en consecuencia la probabilidad de estar reportado oficialmente como contagiado por el virus en esa fecha es de

$$\mathbb{P}(A) = \frac{2616}{4.963703 \times 10^6} = 5.2702589 \times 10^{-4}.$$

Ejemplo 2

- Consideremos Ω la población de la Comunidad Valenciana,
- Ω contiene 4.963703×10^6-individuos (en el lenguaje probabilista se les llama sucesos elementales ya que son las unidades más pequeñas en las que podemos dividir la población) .
- Consideremos el conjunto de individuos A' de la Comunidad Valenciana que cumplen la propiedad $\mathcal{P}'$ siguiente: **estar contagiado con el COVID19 en la fecha 2020-03-25**.
- En general **no conocemos cuantos individuos contagiados no reportados oficialmente existían en esa fecha (asintomáticos, no detectados, ect)**. En consecuencia, no podemos obtener expresamente

$$\mathbb{P}(A') = \frac{\textbf{numero de individuos de la C.V. que cumplen } \mathcal{P}'}{4.963703 \times 10^6}.$$

Como todo individuo en A también está en A', entonces se tiene que cumplir

$$\mathbb{P}(A') \geq \mathbb{P}(A) = 5.2702589 \times 10^{-4}.$$

Aproximación experimental

En la vida real extraemos una muestra (pequeña) Ω_1 que sea representativa de la población Ω y calculamos

$$\mathbb{P}(A|\Omega_1) = \frac{\textbf{número de individuos en } \Omega_1 \textbf{ que tienen la propiedad } \mathcal{P}}{\text{número total de individuos de } \Omega_1}$$

y empleamos

$$\mathbb{P}(A|\Omega_1) \text{ como aproximación de } \mathbb{P}(A).$$

Al conjunto de individuos de Ω_1 se les llama **grupo experimental**.

Justificación metodológica

Ω_1 es una **muestra representativa** si cumple que

$$\frac{\mathbb{P}(A|\Omega_1)}{\mathbb{P}(A)} = 1 - \varepsilon \approx 1$$

donde $\varepsilon > 0$ representa una cantidad pequeña (digamos por ejemplo $\varepsilon = 0.00001$).

Cuestión

Deseamos conocer el alcance de una determinada epidemia en una población de 100000 individuos. Sabemos que podemos emplear una muestra representativa Ω_1 de 1500 indivíduos. En este caso representativa quiere decir que el error se estima en 0.05 (5%). Si obtenemos que 2 de los 1500 individuos de la muestra tienen la enfermedad, ¿Cuaĺ es la probabilidad de padecer la enfermedad en la población aproximadamente?

Respuesta

Conocemos que $\mathbb{P}(\Omega_1) = \frac{1500}{100000} = \frac{3}{200}$ y que $\varepsilon = 0.05$. Además, $\mathbb{P}(E|\Omega_1) = \frac{2}{1500}$, en consecuencia

$$\frac{\mathbb{P}(E|\Omega_1)}{\mathbb{P}(E)} = \frac{3/200}{\mathbb{P}(E)} = 1 - \varepsilon = 1 - 0.05 = 0.95,$$

luego

$$\mathbb{P}(E) = \frac{3/200}{0.95} = 0.01425 \approx 14 \text{ de cada } 1000 \text{ individuos.}$$

Definición (probabilidad)

Una probabilidad es una asignación $\mathbb{P}$ definida sobre una población experimental Ω de forma que para cualquier subcojunto A de la población Ω se cumple

$$0 \leq \mathbb{P}(A) \leq 1,$$

y $\mathbb{P}(\Omega) = 1$. Además, si A y B son dos subconjuntos de la población sin elementos en común ($A \cap B = \emptyset$) entonces la probabilidad que ocurra A o B es igual a la suma de las probabilidades individuales:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

Propiedad 1

Una notación que se suele emplear es la siguiente: Supongamos que

$$A := \text{ son los individuos de } \Omega \text{ que comparten la propiedad } \mathcal{P}$$

entonces denotamos por

$$\overline{A} := \text{ son los individuos de } \Omega \text{ que } \textbf{NO} \text{ comparten la propiedad } \mathcal{P}.$$

Como $\Omega = A \cup \overline{A}$ y ambos no tienen individuos en común ($A \cap \overline{A} = \emptyset$), entonces

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(\overline{A}),$$

luego

$$\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A).$$

Ejemplo 1

Supongamos que una enfermedad genética E en una determinada población se da con una frecuencia de 1 por cada 100000 indivíduos ¿cuál será la probabilidad de NO tener esa enfermedad genética $\overline{E}$?

$$\mathbb{P}(E) = \frac{1}{100000} \text{ luego } \mathbb{P}(\overline{E}) = 1 - \frac{1}{100000} = \frac{99999}{100000}.$$

Ejemplo 2

Supongamos que una enfermedad E se da en una determinada población y que F denota los individuos susceptibles de tener un determinado factor de riesgo que suponemos permite detectar la enfermedad. **Buscamos calcular la probabilidad del factor de riesgo en la población en relación a los individuos enfermos**. Observemos que

$$E \cap F = \text{ son los individuos enfermos que presentan ese factor de riesgo y,}$$

$$\overline{E} \cap F = \text{ son los individuos NO enfermos que presentan ese factor de riesgo,}$$

y ambos grupos no tienen individuos en común. Como los individuos que presentan el factor de riesgo los podemos clasificar en dos grupos sin individuos en común:

$$F = (E \cap F) \cup (\overline{E} \cap F)$$

entonces

$$\mathbb{P}(F) = \mathbb{P}(E \cap F) + \mathbb{P}(\overline{E} \cap F)$$

Ejemplo 3

Supongamos que una enfermedad E se da en una determinada población y que Ω_1 denota los individuos de una muestra elegidos para detectar la enfermedad. **Buscamos calcular la probabilidad de tener la enfermedad en la población en relación a los individuos de la muestra**. Observemos que

$$\Omega_1 \cap E = \text{ son los individuos de la muestra que están enfermos y,}$$

$$\overline{\Omega}_1 \cap E = \text{ son los individuos que NO están en la muestra y están enfermos,}$$

y ambos grupos no tienen individuos en común. Como los individuos que presentan la enfermedad $E = (\Omega_1 \cap E) \cup (\overline{\Omega}_1 \cap E)$ entonces

$$\mathbb{P}(E) = \mathbb{P}(\Omega_1 \cap E) + \mathbb{P}(\overline{\Omega}_1 \cap E)$$

Cuestión

¿Cómo podemos calcular en la práctica $\mathbb{P}(\Omega_1 \cap E)$ o $\mathbb{P}(E \cap F)$?

Respuesta

- Si conocemos $\mathbb{P}(\Omega_1)$ o $\mathbb{P}(F)$,
- Si conocemos $\mathbb{P}(E|\Omega_1)$ o $\mathbb{P}(E|F)$, donde

$$\mathbb{P}(E|\Omega_1) := \frac{\text{número de individuos enfermos en la muestra}}{\text{número de individuos de la muestra}}$$

y

$$\mathbb{P}(E|F) := \frac{\text{número de individuos enfermos que tienen el factor de riesgo}}{\text{número de individuos que tienen el factor de riesgo}}$$

Entonces

$$\mathbb{P}(\Omega_1 \cap E) = \mathbb{P}(E|\Omega_1)\mathbb{P}(\Omega_1) \text{ y } \mathbb{P}(E \cap F) = \mathbb{P}(E|F)\mathbb{P}(F).$$

Cuestión

Tenemos una población 100000 individuos donde sabemos que 1 de cada 10000 padece una enfermedad E. Además, sabemos que 5 de cada 10 enfermos que llegaron a un hospital presentaron un factor de riesgo F. ¿Cuántos individuos estan enfermos y presentan el facfor de riesgo aproximadamente?

Respuesta

Si **conocemos** (información a priori) que una persona está enferma, entonces tiene una probabilidad de $\frac{5}{10} = \frac{1}{2}$ de tener el factor de riesgo, luego

$$\mathbb{P}(F|E) = \frac{1}{2}$$ (con la notación $F|E$ indicamos que conocemos que está enfermo).

Además, la probabilidad de padecer la enfermedad es $\mathbb{P}(E) = \frac{1}{10000}$, entonces

$$\mathbb{P}(E \cap F) = \mathbb{P}(F|E)\mathbb{P}(E) = \frac{1}{2} \times \frac{1}{10000} = \frac{1}{20000}.$$

Si hay 100000 individuos en la población tenemos que unos $\frac{1}{20000} \times 100000 = 5$ indivíduos estaran enfermos y presentaran el factor de riesgo aproximadamente.

Definición (Probabilidad Condicional)

Sean

- Ω una población y $\mathbb{P}$ una probabilidad definida sobre la población.
- A es un subconjunto de individuos de la población que comparten una propiedad $\mathcal{P}$, (información a priori).

Entonces dado cualquier otro subconjunto B de individuos la población que comparten una propiedad $\mathcal{Q}$, se define la probabilidad de que un individuo que cumple la propiedad $\mathcal{P}$, y en consecuencia está en A, sea susceptible de cumplir la propiedad $\mathcal{Q}$, es decir que esté en B, como

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\text{número de individuos de } A \text{ que están a su vez en } B}{\text{número de individuos de } A}.$$

$\mathbb{P}(B|A)$ es la **probabilidad de que ocurra** B **condicionada a que ocurra** A.

Independencia probabilista

Sean

- Ω una población y $\mathbb{P}$ una probabilidad definida sobre la población.
- Los individuos A cumplen la propiedad $\mathcal{P}$,
- Los individuos B cumplen la propiedad $\mathcal{Q}$.

Se dice que las propiedades $\mathcal{P}$ y $\mathcal{Q}$ relativas a los conjuntos A y B son independientes si

$$\mathbb{P}(B|A) = \mathbb{P}(B),$$

es equivalente a decir que

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Cuestión

En un grupo de 25 pacientes 5 presentan vómitos, 10 padecen migrañas, y 2 de los pacientes presentan a su vez vómitos y migrañas. ¿Podemos afirmar que presentar vómitos es independiente de padecer migrañas?

Respuesta

Conocemos que

$$\begin{aligned} A &= \text{ser susceptible de presentar vómitos y} \\ B &= \text{ser susceptible de padecer migrañas} \end{aligned}$$

entonces de los datos facilitados se tiene que

$$\mathbb{P}(A) = \frac{5}{25} = \frac{1}{5}, \mathbb{P}(B) = \frac{10}{25} = \frac{2}{5} \text{ y } \mathbb{P}(A \cap B) = \frac{2}{25},$$

como

$$\mathbb{P}(A \cap B) = \frac{2}{25} = \frac{1}{5} \times \frac{2}{5} = \mathbb{P}(A)\,\mathbb{P}(B),$$

podemos afirmar que en nuestra población de pacientes presentar vómitos es independiente de padecer migrañas, es decir $\mathbb{P}(A|B) = \mathbb{P}(A)$ y $\mathbb{P}(B|A) = \mathbb{P}(B)$.

Cuestión

En un grupo de 25 pacientes 6 presentan vómitos, 10 padecen migrañas, y 2 de los pacientes presentan a su vez vómitos y migrañas. ¿Podemos afirmar que presentar vómitos es independiente de padecer migrañas?

Respuesta

Conocemos que

$$\begin{aligned} A &= \text{ser susceptible de presentar vómitos y} \\ B &= \text{ser susceptible de padecer migrañas} \end{aligned}$$

entonces de los datos facilitados se tiene que

$$\mathbb{P}(A) = \frac{6}{25}, \mathbb{P}(B) = \frac{10}{25} = \frac{2}{5} \text{ y } \mathbb{P}(A \cap B) = \frac{2}{25},$$

como

$$\mathbb{P}(A \cap B) = \frac{2}{25} \neq \frac{12}{125} = \frac{6}{25} \times \frac{2}{5} = \mathbb{P}(A)\,\mathbb{P}(B),$$

podemos afirmar que en nuestra población de pacientes presentar vómitos no es independiente de padecer migrañas.

Conclusión

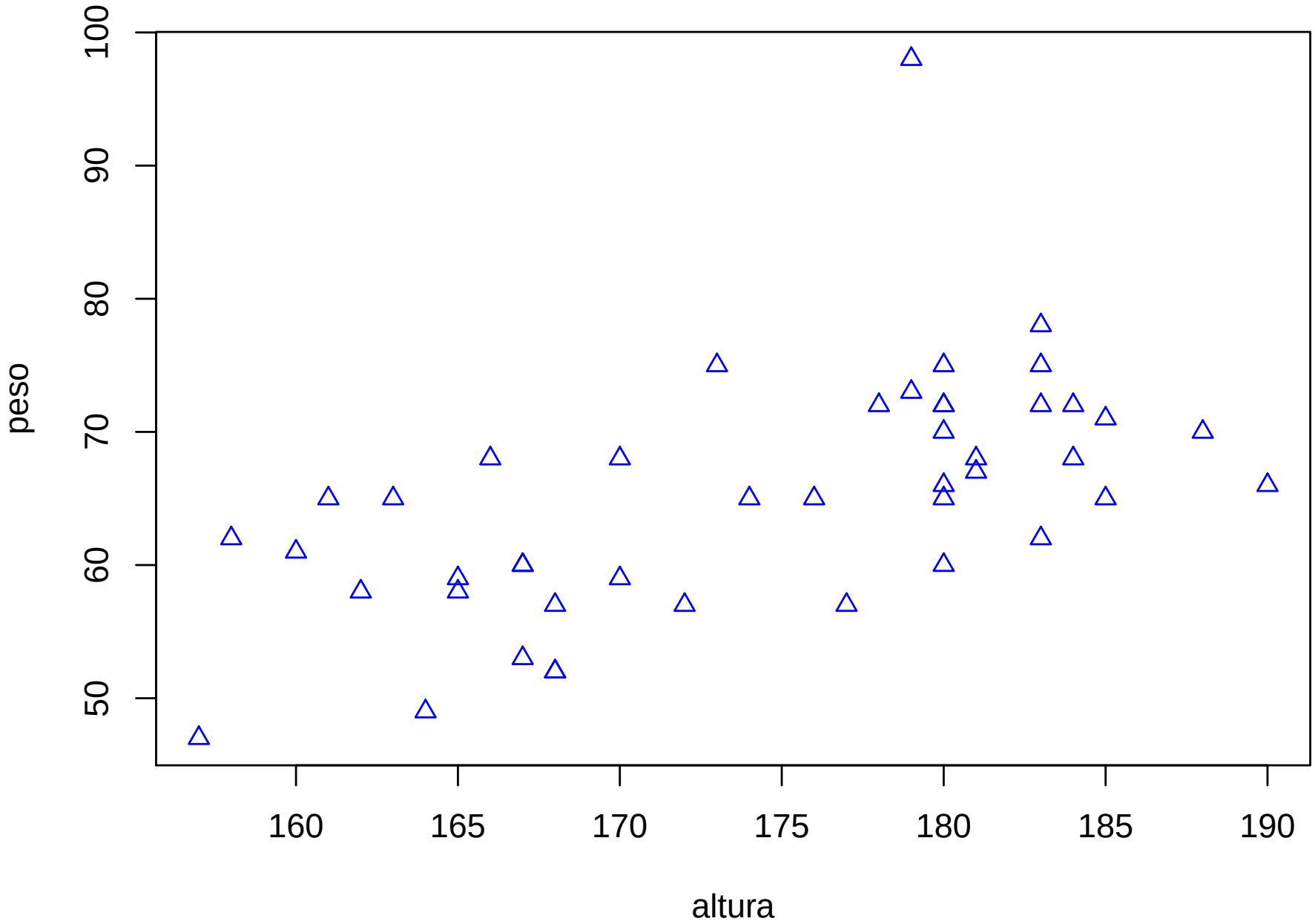
La noción de independencia probabilista es una noción que afecta a la relación numérica entre dos propiedades que caracterizan a los individuos de una población, y en consecuencia **no explica una relación causa-efecto entre estas dos propiedades.**

Cuestión científica

¿Qué clase de relación explica la noción de probabilidad condicional?

```
altura <- c(180, 177, 180, 180, 183, 184, 185, 184, 174,
180, 168,180, 183,181, 180, 190, 183, 167, 181, 179, 173,
170, 170, 183,179, 180, 188, 176,178, 185, 168, 157, 167,
168, 163, 167,166,164, 172, 165, 158, 161, 160,162, 165)
peso <- c(70, 57, 60, 66, 62, 68, 65, 72, 65, 72, 52, 75,
75, 68, 65,66, 78, 60, 67, 98, 75, 68, 59, 72, 73, 72, 70,
65, 72, 71, 52, 47,53,57, 65, 60, 68, 49, 57, 59, 62, 65,
61, 58, 58)
```

Consideremos una muestra de la talla y el peso de 45 individuos, que podemos representar en una gráfica.


peso
altura
50
60
70
80
90
100
160
165
170
175
180
185
190

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?