Probabilidad: una introducción (2)

Angelo Santana

Ejemplo 2

Imagen extraida de Wikipedia, De Teknad - Trabajo propio, CC BY-SA 4.0

Se ha realizado un estudio para evaluar si es admisible la hipótesis de que la sex ratio de la especie de cangrejos *Plagusia depressa* es 1 (esto es, que machos y hembras se encuentran en una proporción 50%-50%). Para ello se ha capturado una muestra aleatoria de 60 cangrejos de esta especie, de los cuáles 21 han resultado ser machos y 39 hembras. **¿Constituye esta muestra evidencia suficiente de que en esa especie hay más hembras que machos?**

Hipótesis inicial

- El problema es idéntico al que ya hemos visto en el ejemplo 1 (determinar si es admisible la hipótesis de que hay el mismo número de alumnos y alumnas en una facultad universitaria); sin embargo en aquel ejemplo la población era **finita** de **tamaño conocido y relativamente pequeño** (268 estudiantes en total), mientras que en este caso, el tamaño de la población de cangrejos es **desconocido**, aunque presumiblemente **muy grande**.

- Si la hipótesis de sex ratio igual a 1 fuese cierta, en una muestra aleatoria cabría esperar un número similar, **aunque no necesariamente igual**, de machos y hembras.

- El resultado observado (21 machos y 39 hembras), *¿es esperable (tiene una probabilidad alta de ocurrencia) bajo la hipótesis inicial, o no?*

Regla de decisión

- De nuevo utilizaremos la misma regla para decidir sobre la validez de la hipótesis de partida (que la sex ratio es igual a 1): suponiendo que dicha hipótesis es cierta, calcularemos cuál es la probabilidad de observar 21 machos y 39 hembras en una muestra de 60 ejemplares.
- Si esta probabilidad tiene un valor alto, concluiremos que la muestra observada es perfectamente compatible con dicha hipótesis, y por tanto no hay evidencia suficiente que indique que es falsa.
- Por el contrario, si esta probabilidad tiene un valor muy bajo, ello indicaría que la ocurrencia de este suceso es muy difícil bajo esa hipótesis, por lo que tendríamos cierta evidencia para rechazarla.
- **¿Cómo calcular la probabilidad de 21 machos y 39 hembras en este caso?**

Probabilidad

[Pierre Simon de Laplace en Wikipedia]

En este caso no parece haber una manera evidente de aplicar la regla de Laplace,

$$\Pr(A) = \frac{\#\ \text{Casos Favorables a A}}{\#\ \text{Casos Posibles}}$$

ya que en una población de tamaño desconocido no se puede calcular el número de casos favorables y de casos posibles del mismo modo que hicimos con los alumnos de la facultad del ejemplo anterior.

¡Es preciso generalizar la definición de probabilidad!

Definición axiomática de probabilidad

Sea Ω el conjunto de los posibles sucesos asociados a un **fenómeno aleatorio** (experimento u observación cuyo resultado exacto no puede conocerse *a priori*).

Una **medida de probabilidad** es una función definida sobre los sucesos de Ω que satisface los siguientes axiomas:

1. $0 \leq \Pr(A) \leq 1$ para todo suceso $A \subseteq \Omega$
2. $\Pr(\Omega) = 1$
3. Dados dos sucesos A y B tales que $A \cap B = \emptyset$, entonces:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$

Algunas propiedades de la Probabilidad

1. $\Pr(\emptyset) = 0$
2. $\Pr\left(A^C\right) = 1 - \Pr(A)$
3. Si $A \subset B$, entonces $\Pr(A) \leq \Pr(B)$
4. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

Probabilidad condicionada

Un suceso A contiene información de otro suceso B, cuando la ocurrencia o no de A modifica la probabilidad de B.

Ejemplos:

- Se lanza un dado dos veces: si la primera vez sale un 2 ¿ello nos da alguna información sobre lo que saldrá la segunda vez?
- En una urna hay dos bolas blancas y una negra. Se sacan dos bolas, sucesivamente y sin reemplazamiento. ¿Informa el color de la primera bola sobre el color de la segunda?.
- Si se saca una tercera bola, también sin reemplazamiento. ¿Informa el color de las dos primeras sobre el color de la tercera?

Probabilidad condicionada

Se define la probabilidad del suceso B condicionado por el suceso A como:

$$\Pr\left(B \mid A\right) = \frac{\Pr\left(A \cap B\right)}{\Pr\left(A\right)}$$

Regla multiplicativa

De la definición de probabilidad condicionada se sigue que:

$$\Pr\left(A \cap B\right) = \Pr\left(A\right)\Pr\left(B \mid A\right)$$

Independencia de sucesos

Un suceso **B es independiente de A** cuando A no contiene información sobre B, esto es:

$$\Pr\left(B \mid A\right) = \Pr\left(B\right)$$

Si B es independiente de A, entonces A es independiente de B:

$$\Pr\left(A \,|B\right) = \frac{\Pr\left(A \cap B\right)}{\Pr\left(B\right)} = \frac{\Pr\left(A\right)\Pr\left(B \,|A\right)}{\Pr\left(B\right)} = \frac{\Pr\left(A\right)\Pr\left(B\right)}{\Pr\left(B\right)} = \Pr\left(A\right)$$

Por tanto A y B se dicen *mutuamente independientes* o simplemente **independientes**.

Independencia de sucesos

De acuerdo con la definición de probabilidad condicionada:

$$\Pr\left(B \mid A\right) = \frac{\Pr\left(A \cap B\right)}{\Pr\left(A\right)}$$

Si A y B son independientes, entonces $\Pr\left(B \mid A\right) = \Pr\left(B\right)$ y por tanto de la definición anterior se sigue que:

$$\Pr\left(A \cap B\right) = \Pr\left(A\right)\Pr\left(B\right)$$

Esta última probabilidad se puede generalizar a n sucesos independientes:

$$\Pr\left(A_1 \cap A_2 \cap \cdots \cap A_n\right) = \Pr\left(A_1\right)\Pr\left(A_2\right)\cdots\Pr\left(A_n\right)$$

Volvamos a nuestro ejemplo.

- Llamemos $P\left(H\right)$ la probabilidad de elegir un cangrejo hembra. La probabilidad de que un cangrejo sea macho será $P\left(M\right) = 1 - P\left(H\right)$.

- Como la población es infinita, podemos asumir que las probabilidades $P\left(H\right)$ y $P\left(M\right)$ se mantienen constantes durante el muestreo *aunque no haya reemplazamiento*. Nótese que ésto no ocurriría en una población finita: si solo hubiese 100 machos y 100 hembras en la población y ya tenemos una muestra con 0 machos y 50 hembras, cuando capturemos el siguiente ejemplar, la probabilidad de que sea macho es mayor que la probabilidad de que sea hembra.

- Supondremos además que hay **independencia** entre el sexo de los sucesivos cangrejos que se incorporan a la muestra, esto es, conocer el sexo de los cangrejos que ya forman parte de la muestra no nos da información sobre cuál será el sexo del siguiente cangrejo que capturemos.

Volvamos a nuestro ejemplo.

- Dado que el sexo de cada cangrejo es independiente del resto, la probabilidad de que para formar una muestra de 60 cangrejos se capturen 21 machos y 39 hembras (**en ese orden**) sería:

$$\overbrace{P(M)\,P(M)\ldots P(M)}^{21}\overbrace{P(H)\,P(H)\ldots P(H)}^{39} = P(M)^{21}P(H)^{39}$$

- Ahora bien, podrían capturarse primero las 39 hembras y después los 21 machos. La probabilidad de que la muestra se obtenga de esta forma es entonces:

$$\overbrace{P(H)\,P(H)\ldots P(H)}^{39}\overbrace{P(M)\,P(M)\ldots P(M)}^{21} = P(M)^{21}P(H)^{39}$$

Volvamos a nuestro ejemplo.

- Cualquier otro orden en que se capturen tendría la misma probabilidad:
  - Una hembra primero, 21 machos y otras 38 hembras:

$$P(H)\overbrace{P(M)\,P(M)\ldots P(M)}^{21}\overbrace{P(H)\,P(H)\ldots P(H)}^{38} = P(M)^{21}P(H)^{39}$$

  - Dos hembras, 20 machos, 37 hembras a continuación y por último otro macho:

$$P(H)\,P(H)\overbrace{P(M)\,P(M)\ldots P(M)}^{20}\overbrace{P(H)\,P(H)\ldots P(H)}^{37}P(M) =$$

$$= P(M)^{21}P(H)^{39}$$

Volvamos a nuestro ejemplo.

- Por tanto, la probabilidad total de capturar 21 machos y 39 hembras puede obtenerse sumando el valor $P(M)^{21} P(H)^{39}$ tantas veces como maneras distintas haya de realizar la captura.

- El número de maneras en que puede realizarse la captura es igual al número de formas en que se pueden colocar los 21 machos en las 60 posiciones disponibles en la muestra (o 39 hembras en las 60 posiciones):

$$\binom{60}{21} = \binom{60}{39} = \frac{60!}{21! \cdot 39!} = 7.98446 \cdot 10^{15}$$

- Por tanto, la **probabilidad total** de capturar 21 machos (y 39 hembras) en una muestra de 60 ejemplares es:

$$P\left(21 \text{ machos}\right) = \binom{60}{21} P(H)^{21} P(M)^{39}$$

Volvamos a nuestro ejemplo.

- Si es cierta la hipótesis de que la sex-ratio es 1, entonces $P(H) = P(M) = \frac{1}{2}$ y la probabilidad anterior vale:

$$P(21\text{machos}) = \binom{60}{21}\left(\frac{1}{2}\right)^{21}\left(\frac{1}{2}\right)^{39} = 0.00692542$$

- Nuevamente: **¿esta probabilidad es alta o baja?**

- Para responder a esta pregunta debemos evaluar las probabilidades de los otros sucesos que pueden ocurrir bajo esta hipótesis: que se capturen 0 machos, solo 1 macho, 2 machos, 3, ... o que todos los ejemplares que se capturen sean machos

Distribución binomial

Si llamamos X al número de machos en la muestra, y $p = P(M)$ la probabilidad de que haya k machos en una muestra de 60 cangrejos sería:

$$P(X = k) = \binom{60}{k} p^k (1-p)^{60-k}$$

Y, en general, si la muestra estuviese compuesta por n cangrejos:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Este reparto o distribución de probabilidades entre los distintos valores de k recibe el nombre de **distribución binomial de parámetros n y p**. Suele denotarse mediante la expresión $X \approx B(n, p)$.

Distribución binomial

- Las probabilidades correspondientes a la distribución binomial pueden calcularse en R mediante la función `dbinom(k,n,p)`.

- Así, en nuestro caso, el número de machos es $X \approx B(60, \frac{1}{2})$ y la probabilidad de que ocurran 21 machos se calcularía mediante:

```
dbinom(21,60,0.5)
```

```
## [1] 0.00692542
```

En nuestro problema:

Para decidir si la probabilidad anterior es grande o pequeña, debemos compararla con la probabilidad del resto de sucesos posibles:

```r
library(tidyverse)
data.frame(k=0:60) %>%
  mutate(p=dbinom(k,60,0.5))
```

En nuestro problema:

k	p(X=k)
0	8.674e-19
1	5.204e-17
2	1.535e-15
3	2.968e-14
4	4.230e-13
5	4.737e-12
6	4.342e-11
7	3.350e-10
8	2.219e-09
9	1.282e-08
10	6.539e-08
11	2.972e-07
12	0.000001

k	p(X=k)
13	0.000004
14	0.000015
15	0.000046
16	0.000130
17	0.000336
18	0.000802
19	0.001774
20	0.003636
21	0.006925
22	0.012277
23	0.020284
24	0.031270

k	p(X=k)
25	0.045029
26	0.060617
27	0.076332
28	0.089963
29	0.099269
30	0.102578
31	0.099269
32	0.089963
33	0.076332
34	0.060617
35	0.045029
36	0.031270

k	p(X=k)
37	0.020284
38	0.012277
39	0.006925
40	0.003636
41	0.001774
42	0.000802
43	0.000336
44	0.000130
45	0.000046
46	0.000015
47	0.000004
48	0.000001

k	p(X=k)
49	2.972e-07
50	6.539e-08
51	1.282e-08
52	2.219e-09
53	3.350e-10
54	4.342e-11
55	4.737e-12
56	4.230e-13
57	2.968e-14
58	1.535e-15
59	5.204e-17
60	8.674e-19

En nuestro problema:

- Por tanto, si la hipótesis de sex-ratio = 1 fuera cierta, la probabilidad de observar un suceso **tanto o más improbable que el que hemos observado** (21 machos y 39 hembras, con probabilidad 0.006925) se obtendría sumando todas las probabilidades que se muestran en rojo en la tabla anterior (las menores o iguales que 0.006925).

- Dicha suma vale:

```
data.frame(k=0:60) %>%
  mutate(p=dbinom(k,60,0.5)) %>%
  filter(p<=dbinom(21,60,0.5)) %>%
  summarize(sum(p))
```

```
##       sum(p)
## 1 0.02734013
```

En nuestro problema:

- Por tanto, lo que hemos observado **es muy improbable** si fuera cierta la hipótesis de sex-ratio=1.
- En la práctica científica habitual, cuando la probabilidad total de ocurrencia de sucesos tanto o más improbables que el observado es menor que **0.05**, se considera que **hay evidencia suficiente para rechazar la hipótesis de partida**.
- Por tanto en este caso decidiríamos que hay evidencia suficiente de que la sex-ratio en esta especie de cangrejos, al menos durante el periodo en que se realizó el muestreo, es distinta de uno.