Semana 4
Bolas, urnas y celdas

4-1. Urnas y muestras

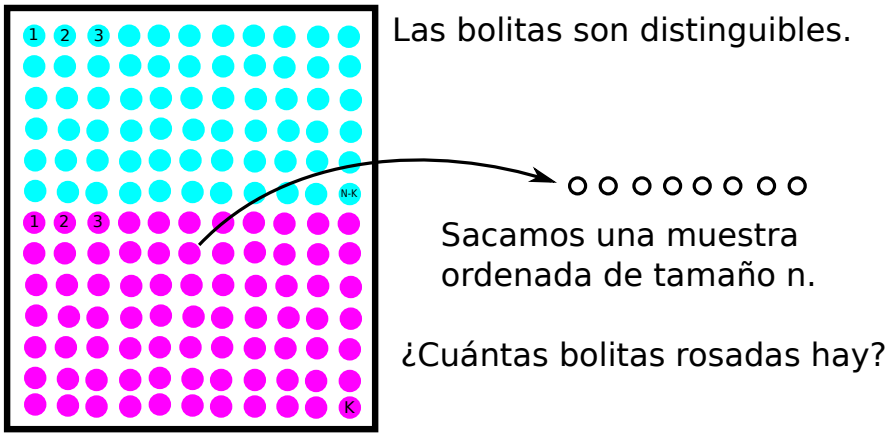
La gran mayoría de los experimentos aleatorios cuyos espacios muestrales son finitos se pueden modelar usando urnas y/o celdas. En este capítulo vamos a discutir algunas propiedades de estos modelos. Comenzaremos por los modelos de urnas que son más naturales.

Experimento: tenemos una urna con N bolas, numeradas de 1 a N:

$$B = \{b_1, b_2, \ldots, b_N\},$$

de las cuales K son rosadas y $N - K$ son celestes.

Extraemos n bolas para formar una muestra *ordenada*.



¿Cuál es el espacio muestral Ω? Para empezar, debemos distinguir dos formas de extraer las bolas:

- Sin reposición

$$\Omega_{SR} = \{\omega = (\omega_1, \ldots, \omega_n) : \omega_i \in B \text{ para todo } i, \text{ y } \omega_i \neq \omega_j \text{ si } i \neq j\}.$$

  En este caso, al extraer una bola no la volvemos a poner en la urna, por lo que en la muestra todas la bolas deben ser diferentes entre sí.

- Con reposición

$$\Omega_{CR} = \{\omega = (\omega_1, \ldots, \omega_n) : \omega_i \in B \text{ para todo } i\}.$$

  En este caso, al extraer una bola la volvemos a poner en la urna, por lo que puede aparecer más de una vez en la lista.

Caso sin reposición: la distribución hipergeométrica

En este caso el espacio muestral es

$$\Omega_{SR} = \{\omega = (\omega_1, \ldots, \omega_n) : \omega_i \in B \text{ para todo } i, \text{ y } \omega_i \neq \omega_j \text{ si } i \neq j\}.$$

Notar que en este caso se debe imponer $0 \leq n \leq N$.

Como en todo espacio muestral finito, los eventos son todos los subconjuntos del espacio muestral. Así que lo más importante es determinar cuáles son las probabilidades de los eventos elementales, es decir, de cada una de las listas ordenadas de Ω_{SR}.

Cada una de las listas tiene la misma chance de ocurrir que cualquier otra, ya que no hay ninguna razón por la cual suponer que alguna de ellas tiene una preferencia de ser elegida. Entonces, la probabilidad de cada lista ω es simplemente $1/|\Omega_{SR}|$.

Para calcular el número total de muestras posibles razonamos de la siguiente manera: para elegir la primer bola tenemos N posibilidades, para la segunda $N-1$ (ya que no volvemos a poner la bola que recién extrajimos), y así sucesivamente, hasta que cuando vayamos a elegir la n-ésima bola quedan $N-(n-1) = N-n+1$ bolas en la urna. Luego el número total de muestras es

$$|\Omega_{SR}| = N \times (N-1) \times \cdots \times (N-n+1).$$

Podemos escribir este número de forma más compacta usando factoriales

$$|\Omega_{SR}| = \frac{N!}{(N-n)!}.$$

La fórmula de la derecha se conoce como arreglos de N en n y se suele escribir como $(N)_n$ o A_n^N.

Lo que nos interesa en este momento es contar cuántas bolas de determinado color hay en la muestra. Para esto introducimos la variable aleatoria

$$X : \Omega_{SR} \to \mathbb{R}$$

que para cada $\omega \in \Omega_{SR}$ devuelve el número de bolas rosadas en la muestra. Más precisamente,

$$X(\omega_1, \ldots, \omega_n) = |\{i : \omega_i \text{ es rosada}\}|\,.$$

Comencemos por estudiar qué valores puede tomar X. Los valores posibles que una variable puede tomar se llama el recorrido de la variable. Así, nos preguntamos ¿Cuál es el recorrido de X?

El valor más chico que X puede tomar en general es 0, pero eso depende de cuántas bolas celestes haya en la urna. Por ejemplo, si hay una sola bola celeste en la urna ($N - K = 1$) y extraemos $n = 5$ bolas, entonces al menos 4 serán rosadas, por lo que X será mayor o igual a 4 siempre.

En general, si extraemos más bolas que la cantidad de bolas celestes, es decir si $n > N - K$, entonces como mínimo habrán $n - (N - K)$ bolas rosadas. En cambio, si la cantidad de bolas celestes es mayor o igual que la cantidad de bolas que extraemos, entonces X podría tomar el valor cero.

Juntando ambos casos, vemos que el valor más chico que X puede tomar es $m_n = \text{máx}\{n - (N - K), 0\}$.

Razonando del mismo modo podemos ver cuál es el valor más grande que puede tomar X. Si la cantidad K de bolas rosadas es mayor o igual que n, entonces X podría tomar el valor n. De lo contrario, el valor más grande que podría tomar X es K. Juntando ambos casos, vemos que el valor más grande que X puede tomar es $M_n = \text{mín}\{K, n\}$.

En resumen, X toma valores enteros y verifica

$$m_n = \text{máx}\{n - (N - K), 0\} \leq X \leq \text{mín}\{K, n\} = M_n.$$

Para cualquier entero k entre m_n y M_n la función de probabilidad puntual de X nos dice con que probabilidad X toma el valor k.

Para calcular esta probabilidad, vemos primero que el evento que nos interesa es

$$\{X = k\} = \{\omega \in \Omega_{SR} : k \text{ de los } \omega_i \text{ son rosadas y } n - k \text{ son celestes}\}.$$

Como todas las muestras son igualmente probables, la probabilidad de $\{X = k\}$ es

$$\mathbf{P}\{X = k\} = \frac{\text{casos favorables}}{\text{casos posibles}} = \frac{|\{X = k\}|}{|\Omega_{SR}|}.$$

Contando:

$$|\{X = k\}| = \overbrace{\binom{n}{k}}^{\substack{\text{elegimos los lugares} \\ \text{de las } k \text{ rosadas}}} \cdot \underbrace{(K)_k}_{\substack{\text{elegimos} \\ \text{las } k \text{ rosadas}}} \cdot \overbrace{(N-K)_{n-k}}^{\substack{\text{elegimos las} \\ n-k \text{ celestes}}}$$

Juntando todo nos queda

$$\mathbf{P}\{X = k\} = \frac{\binom{n}{k}(K)_k(N-K)_{n-k}}{(N)_n}.$$

Esta fórmula no parece muy cómoda de usar. Sin embargo, podemos re-agrupar términos

$$\begin{aligned}\mathbf{P}\{X=k\} &= \frac{\binom{n}{k}(K)_k(N-K)_{n-k}}{(N)_n} \\ &= \frac{n!}{k!(n-k)!}\cdot\frac{K!}{(K-k)!}\cdot\frac{(N-K)!}{(N-K-(n-k))!}\cdot\frac{(N-n)!}{N!} \\ &= \frac{K!}{k!(K-k)!}\cdot\frac{(N-K)!}{(n-k)!(N-K-(n-k))!}\cdot\frac{n!(N-n)!}{N!} = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.\end{aligned}$$

Es decir, hemos obtenido

$$\mathbf{P}\{X=k\} = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Llama la atención las combinaciones que aparecen en el denominador. Esto simplemente refleja el hecho que X no depende del orden en el que aparecen las bolas.

Podríamos haber modelado el experimento con muestras no ordenadas, con el espacio muestral

$$\Omega^*_{SR} = \Big\{\{\omega_1,\ldots,\omega_n\} \text{ subconjuntos de tamaño } n\Big\}.$$

De este modo, obtendríamos directamente que para cada k entre m_n y M_n, la probabilidad de que X sea igual a k está dada por

$$\mathbf{P}\{X=k\} = \overbrace{\binom{K}{k}}^{\substack{\text{elegimos}\\ \text{las } k \text{ rosadas}}} \cdot \overbrace{\binom{N-K}{n-k}}^{\substack{\text{elegimos las}\\ n-k \text{ celestes}}} \Big/ \underbrace{\binom{N}{n}}_{\substack{\text{total de}\\ \text{muestras}}}$$

Sin embargo, el modelo con muestras ordenadas nos será útil más adelante para escribir a X como una suma de n variables aleatorias.

Definición. *La distribución obtenida en el caso sin reposición se llama* Hipergeométrica *de parámetros N, K y n. La función de probabilidad puntual está dada por*

$$\mathbf{P}\{X=k\} = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

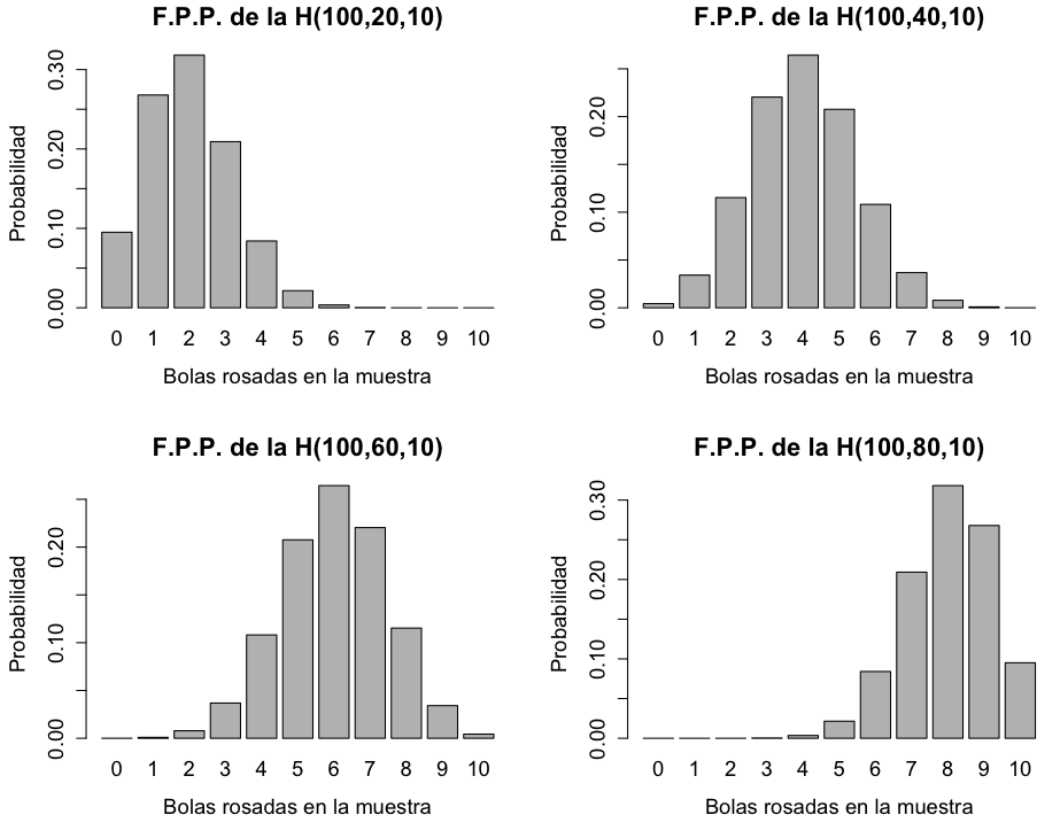
para todo $m_n \le k \le M_n$. Escribimos $X \sim \mathcal{H}(N,K,n)$ para indicar que X tiene distribución hipergeométrica.

En la figura de abajo se muestran cuatro gráficos de la función de probabilidad puntual de la distribución hipergeométrica. En todos los casos $N = 100$, es decir hay 100 bolas en la urna, y $n = 10$, es decir extraemos 10. La cantidad de bolas rosadas varía de $K = 20$ a $K = 80$.

Notar que en cada uno de los casos, la forma de la distribución es acampanada. Sin embargo no es simétrica. En el primer caso en que sólo hay $K = 20$ bolas

rosadas en la urna, la variable X se concentra en valores chicos de k, teniendo un máximo para $k = 2$.

Cuando $K = 40$ la distribución es bastante más simétrica, aunque no del todo. En este caso el máximo se da en $k = 4$. A medida que K aumenta, la distribución se va corriendo hacia la derecha, en donde para $K = 60$ el máximo se da en $k = 6$, y para $K = 80$ el máximo se da en $k = 8$.



En la segunda figura se muestran tres gráficos más. En este caso $N = 1000$, es decir hay 1000 bolas en la urna y se extrae una muestra de tamaño $n = 100$. En el primer caso la cantidad de bolas rosadas es $K = 250$, por lo que la distribución está concentrada en valores pequeños de k, con un máximo en $k = 25$. Lo opuesto ocurre en el tercer caso, en donde $K = 750$.

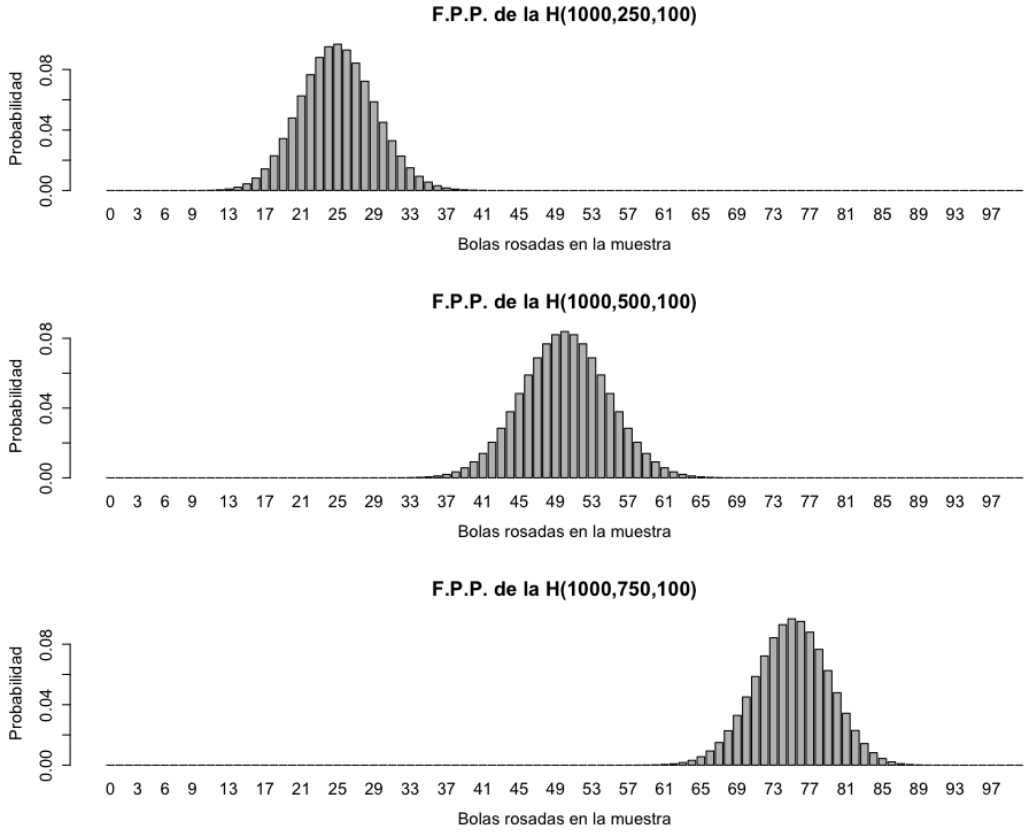
Notar que el segundo gráfico es perfectamente simétrico respecto de $k = 50$. Esto refleja el hecho de que hay la misma cantidad de bolas rosadas $K = 500$ que celestes $N - K = 500$ en la urna. La probabilidad de que X tome los valores $k = 50 \pm x$ son iguales, para todo x entre -50 y 50.

Ejemplo 1

Tal vez pueda parecer sorprendente, pero ya hemos visto la distribución hipergeométrica antes en el curso. De hecho, en el ejemplo de La Catadora de Té la variable X que cuenta el número de aciertos que tiene la Sra. verifica

$$\mathbf{P}\{X = k\} = \frac{\binom{4}{k}\binom{4}{4-k}}{\binom{8}{4}},$$

para k entre 0 y 4. Esta es exactamente la fórmula de la distribución hipergeométrica de parámetros $N = 8$, $K = 4$ y $n = 4$. Es decir $X \sim \mathcal{H}(8, 4, 4)$.



En este caso las bolas en las urnas son las 8 tazas de té, y las bolas rozadas son las tazas en las cuales el té se ha servido primero. El tamaño de la muestra es $n = 4$ y corresponde a la lista de tazas elegida por la Sra.

Ejemplo 2

Supongamos que una lotería funciona de la siguiente manera: de una urna que contiene 44 bolas numeradas del 1 al 44, se extraen al azar 5 de ellas y sin reposición. Los participantes compran tickets en los cuales indican una lista de 5 números. El premio mayor se otorga a aquellos participantes que acierten los 5 números, pero existen premios menores para aquellos que acierten 3 o más.

Este juego lo podemos modelar con urnas y bolas. Imaginemos que decidimos comprar el ticket que contiene los números $\{26, 9, 27, 28, 2\}$. Estas serán las bolas rosadas, por lo que $N = 44$, $K = 5$ y $n = 5$. Denotemos por X la cantidad de coincidencias entre nuestra lista y aquella que sale sorteada. Entonces X es la cantidad de bolas rosadas en la muestra.

La probabilidad de ganar el premio mayor es

$$\mathbf{P}\{X = 5\} = \frac{\binom{5}{5}\binom{39}{0}}{\binom{44}{5}} = \frac{1}{1\,086\,008} \approx 9{,}2 \times 10^{-7}.$$

Sin embargo, la probabilidad de ganar algún premio es

$$\begin{aligned}\mathbf{P}\{X \geq 3\} &= \mathbf{P}\{X = 3\} + \mathbf{P}\{X = 4\} + \mathbf{P}\{X = 5\} \\ &= \frac{\binom{5}{3}\binom{39}{2} + \binom{5}{4}\binom{39}{1} + \binom{5}{5}\binom{39}{0}}{\binom{44}{5}} = \frac{7\,606}{1\,086\,008} \approx 0{,}007.\end{aligned}$$

Aunque sigue siendo una probabilidad muy chica, notar que es 7606 veces mayor que la anterior.

Caso con reposición: la distribución binomial

Recordemos que en este caso el espacio muestral es

$$\Omega_{CR} = \{\omega = (\omega_1, \ldots, \omega_n) : \omega_i \in B \text{ para todo } i\}.$$

Lo primero que observamos es que en este caso no hay restricción sobre el tamaño n de la muestra. Esto es así porque cada vez que retiramos una bola la volvemos a poner para sacar la siguiente. Por tanto se puede tener $n > N$.

El número total de muestras posibles es $|\Omega_{CR}| = N^n$. Esto se ve fácilmente ya que para cada una de las n coordenadas tenemos N posibilidades distintas. Como ninguna de las secuencias ω tiene preferencia para ser elegida, la probabilidad de cada una de ellas es $1/|\Omega_{CR}|$.

Como en el caso sin reposición, consideremos la variable aleatoria $X : \Omega_{CR} \to \mathbb{R}$ que a cada $\omega \in \Omega_{CR}$ asigna el número de bolas rosadas en la muestra. Esto es

$$X(\omega_1, \ldots, \omega_n) = |\{i : \omega_i \text{ es rosada}\}|.$$

¿Cuál es el recorrido de X? En el caso con reposición el análisis es más fácil, ya que independientemente de la cantidad relativa de bolas rosadas y celestes en la urna, la variable X toma valores enteros y verifica $0 \leq X \leq n$.

Para determinar la función de probabilidad puntual de X, debemos calcular la probabilidad de X sea igual a k, para todo valor de k entre 0 y n. El evento que nos interesa es por lo tanto

$$\{X = k\} = \{\omega \in \Omega_{CR} : k \text{ de las } \omega_i \text{ son rosadas y } n - k \text{ son celestes}\}.$$

Por un lado, como todas las secuencias son equiprobables tenemos

$$\mathbf{P}\{X = k\} = \frac{\text{casos favorables}}{\text{casos posibles}} = \frac{|\{X = k\}|}{|\Omega_{CR}|}.$$

Contando:

$$|\{X = k\}| = \overbrace{\binom{n}{k}}^{\substack{\text{elegimos los} \\ \text{lugares de las} \\ k \text{ rosadas}}} \cdot \underbrace{K^k}_{\substack{\text{elegimos las} \\ k \text{ rosadas}}} \cdot \overbrace{(N-K)^{n-k}}^{\substack{\text{elegimos las} \\ n-k \text{ celestes}}}$$

Juntando todo nos queda

$$\mathbf{P}\{X = k\} = \binom{n}{k} \frac{K^k (N-K)^{n-k}}{N^n}.$$

Podemos re-agrupar los términos para que la fórmula sea más fácil de interpretar. Si escribimos N^n como $N^k N^{n-k}$, obtenemos

$$\begin{aligned}\mathbf{P}\{X=k\} &= \binom{n}{k}\frac{K^k(N-K)^{n-k}}{N^n} = \binom{n}{k}\frac{K^k(N-K)^{n-k}}{N^k N^{n-k}} \\ &= \binom{n}{k}\left(\frac{K}{N}\right)^k\left(1-\frac{K}{N}\right)^{n-k}\end{aligned}$$

Es decir, denotando por $p = K/N$ la fracción de bolas rosadas en la urna, la fórmula queda

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

La expresión a la que hemos llegado para la función de probabilidad de X tiene una interpretación muy útil. Notemos primero que $p = K/N$ es la probabilidad de sacar una bola rosada en un sólo intento.

Cuando extraemos n veces con reposición, las condiciones en cada nueva extracción son las mismas que en la primer extracción. Es decir, el procedimiento se puede pensar como la repetición de n veces de una sola extracción. Además las extracciones son independientes entre sí.

Si definimos como "éxito" sacar una bola rosada, X cuenta el número de éxitos en la repetición del experimento n veces. Más aún, X solo distingue si las coordenadas son rosadas o no, por lo que podríamos haber modelado el experimento con el espacio muestral

$$\Omega^*_{CR} = \{(\omega_1,\ldots,\omega_n) : \text{ para todo } i,\ \omega_i = 0 \text{ o } 1\}.$$

En este caso, una coordenada igual a 1 significa que sale una bola rosada, y una igual a 0 que sale una celeste. Entonces X cuenta el número de unos en dicha secuencia.

Para que X sea igual a k deben haber k unos en la secuencia y $n-k$ ceros. Si nos olvidamos por el momento de cuáles son los unos y cuáles son los ceros, la probabilidad de que esto ocurra es $p^k(1-p)^{n-k}$. Lo que falta es tener en cuenta de cuántas formas posibles podemos elegir los lugares para los k unos (pues al elegirlos, los lugares de los $n-k$ ceros quedan automáticamente determinados). Esto se puede hacer precisamente de $\binom{n}{k}$ formas distintas. Esto explica la fórmula que obtuvimos.

Definición. *La distribución obtenida en el caso con reposición se llama* Binomial *de parámetros n y p. La función de probabilidad puntual está dada por*

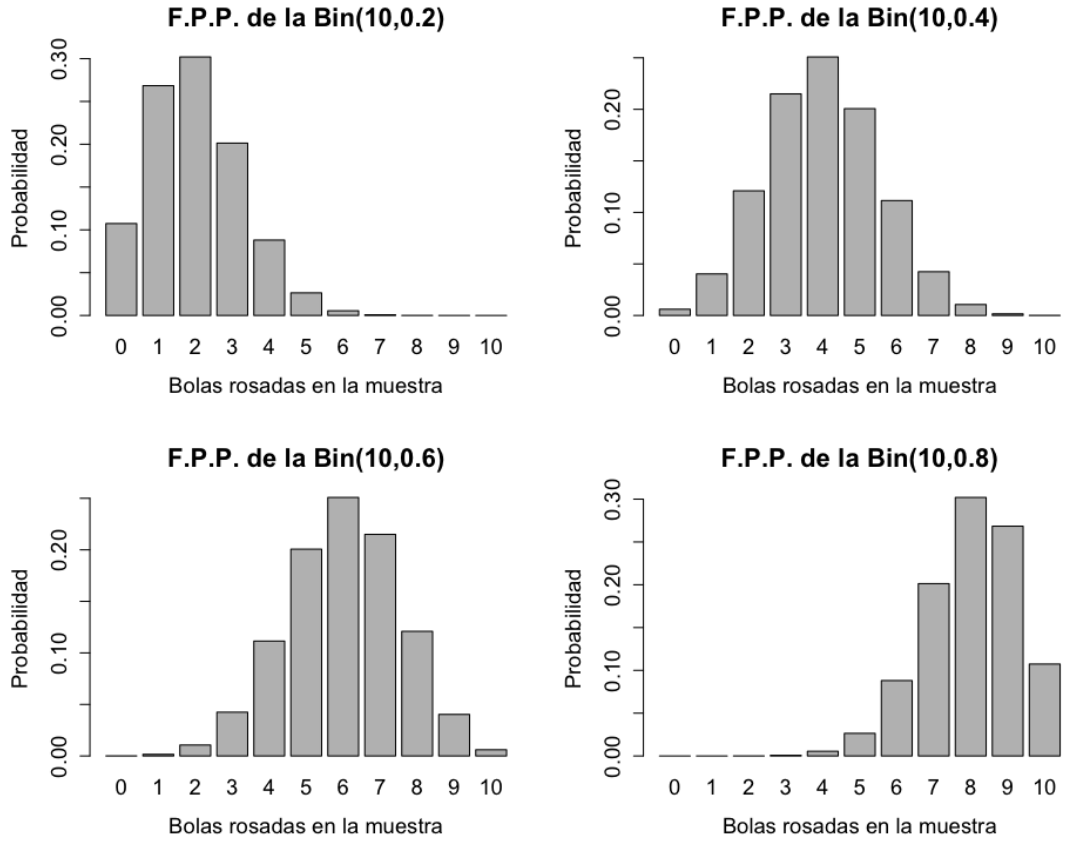
$$\mathbf{P}\{X=k\} = \binom{n}{k}p^k(1-p)^{n-k},$$

para todo k entre 0 y n. Escribimos $X \sim \mathrm{Bin}(n,p)$ para indicar que X tiene distribución binomial.

En la figura de abajo se muestran cuatro gráficos de la función de probabilidad puntual de la distribución binomial. En todos los casos $n = 10$, es decir extraemos (o repetimos) 10 veces de la urna. La proporción de bolas rosadas varía de $p = 0{,}2$ a $p = 0{,}8$.

Notar que en cada uno de los casos, al igual que en el caso de la distribución hipergeométrica, la forma de la distribución es acampanada. Sin embargo no es simétrica. En el primer caso en que la proporción es sólo de $p = 0{,}2$ bolas rosadas en la urna, la variable X se concentra en valores chicos de k, teniendo un máximo para $k = 2$.

Cuando $p = 0{,}4$ la distribución es bastante más simétrica, aunque no del todo. En este caso el máximo se da en $k = 4$. A medida que p aumenta, la distribución se va corriendo hacia la derecha, en donde para $p = 0{,}6$ el máximo se da en $k = 6$, y para $p = 0{,}8$ el máximo se da en $k = 8$.



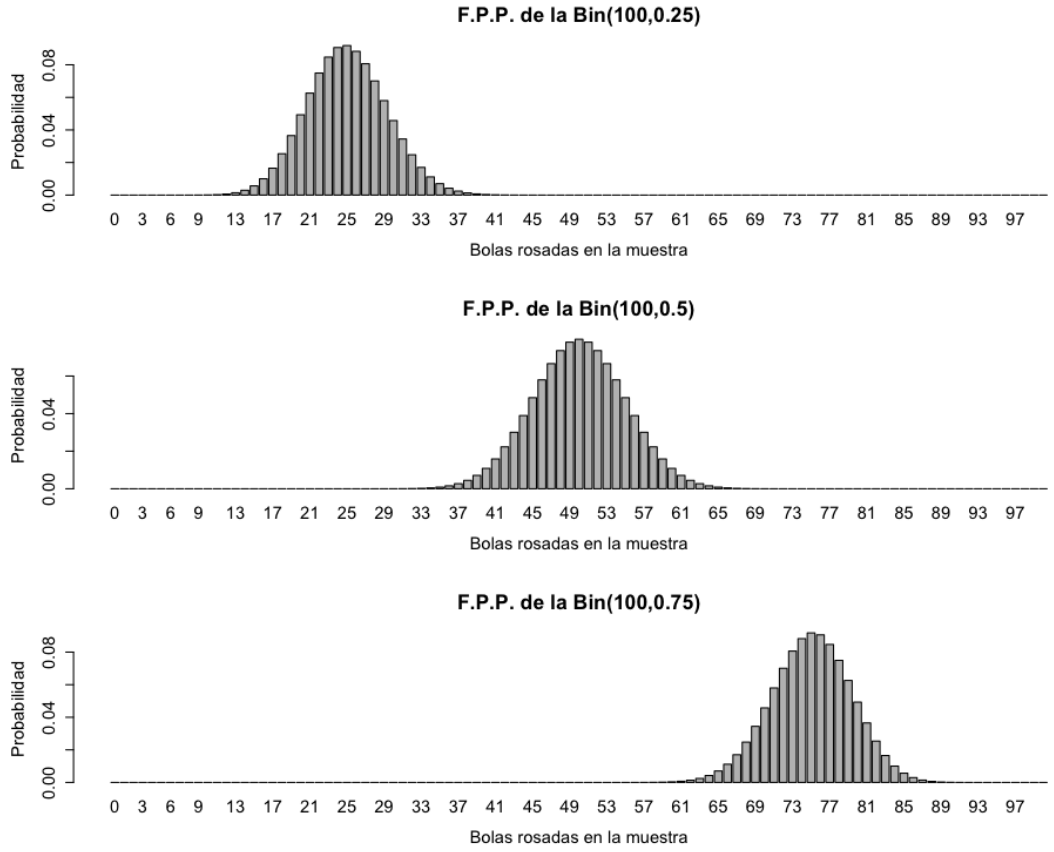
En la segunda figura se muestran tres gráficos más. En este caso $n = 100$, es decir extraemos 100 bolas de la urna. En el primer caso la proporción de bolas rosadas es $p = 0{,}25$, por lo que la distribución está concentrada en valores pequeños de k, con un máximo en $k = 25$. Lo opuesto ocurre en el tercer caso, en donde $p = 0{,}75$.

Notar que el segundo gráfico es perfectamente simétrico respecto de $k = 50$. Esto refleja el hecho de que hay la misma cantidad de bolas rosadas que celestes $p = 0{,}5$ en la urna. La probabilidad de que X tome los valores $k = 50 \pm x$ son iguales, para todo x entre -50 y 50.

Ejemplo 1

También hemos encontrado esta distribución antes. En el ejemplo del estudio del psicólogo sobre la forma de sentarse de los estudiantes en una cantina estudiantil. En este ejemplo, la urna contiene $N = 6$ bolas, de las cuales 4 son rosadas, y

corresponden a las configuraciones en las cuales los estudiantes se sientan en lados adyacentes de la mesa. Las 2 bolas celestes, corresponden por lo tanto a las configuraciones en las cuales los estudiantes se sientan en lados opuestos.



El psicólogo observa a 197 parejas, y en nuestro modelo esto equivale a extraer $n = 197$ bolas de la urna con reposición. Notar que en este caso claramente $n > N$. Vimos que la variable S que cuenta el número de parejas que se sientan en lados adyacentes verificaba

$$\mathbf{P}\{S = k\} = \binom{197}{k} \left(\frac{2}{3}\right)^k \left(\frac{1}{3}\right)^{n-k}$$

para todo k entre 0 y 197. Esta fórmula corresponde a la función de probabilidad puntual de la distribución Bin $\left(197, \frac{2}{3}\right)$.

Ejemplo 2

Un estudio intenta determinar si las personas son capaces de distinguir el género a partir de la escritura. Para esto, a un participante del estudio se le presentaron 20 tarjetas escritas por personas distintas, entre ellas mujeres y hombres.

Para minimizar posibles efectos diferentes al tipo de escritura, todas las tarjetas contenían el mismo texto:

Facultad de Ingeniería
Julio Herrera y Reissig 565 CP11300
Montevideo, Uruguay

A medida que se le presentaban las tarjetas al participante, éste debía indicar si la tarjeta había sido escrita por una mujer o por un hombre.

Suponiendo que el participante indica correctamente el género de la persona en 14 de las tarjetas. ¿Te parece que la performance del participante es significativamente mejor que la que se obtendría al elegir las respuestas al azar?

Este es un claro ejemplo de decisión que debemos tomar razonando por improbable. Para esto, pongamos a prueba la afirmación siguiente: el participante está tirando a embocar. Supongamos que la afirmación es verdadera y calculemos la probabilidad de observar algo tanto o más extremo que lo observado.

Si el participante está indicando el género al azar, podemos modelar el experimento usando la distribución binomial. De hecho, el modelo es equivalente a una urna con $N = 2$ bolas, una rosada (representando a la mujer M) y otra celeste (representando al hombre H), de la cual el participante extrae con reposición $n = 20$ bolas.

Supongamos que la secuencia correcta de géneros es

$$\omega^* = (M, H, H, H, M, M, H, M, M, M, H, M, H, M, M, H, H, M, M, H)\,.$$

Notar que hay 11 tarjetas escritas por mujeres (M) y 9 por hombres (H).

Para una secuencia posible ω de $M's$ y $H's$ denotamos por $X(\omega)$ el número de aciertos, esto es de coincidencias entre ω y ω^*. En símbolos

$$X(\omega) = |\{i : \omega_i = \omega_i^*\}|\,.$$

La variable X puede tomar cualquier valor entre 0 y 20. Además, X es igual a k si la cantidad de aciertos es exactamente k. La cantidad de secuencias que tienen exactamente k aciertos es $\binom{20}{k}$. Por tanto

$$\mathbf{P}\{X = k\} = \frac{1}{2^{20}}\binom{20}{k}.$$

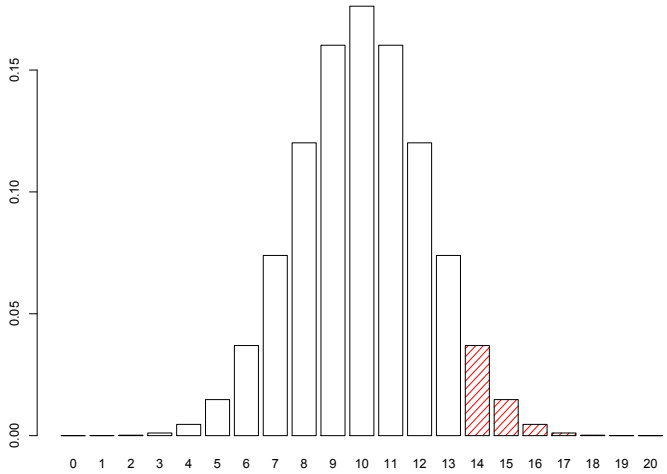
Esto significa que la variable X tiene distribución binomial de parámetros $n = 20$ y $p = 1/2$. En símbolos $X \sim \text{Bin}\,(20, 1/2)$.

Es importante notar que el parámetro $p = 1/2$ es consecuencia de que el participante está eligiendo al azar el género, y no depende de cuántas tarjetas escritas por mujeres o hombres haya en la secuencia correcta. Pusimos una secuencia con 11 mujeres y 9 hombres para resaltar este hecho.

Ahora que tenemos el modelo completo y hemos hecho uso de la afirmación cuya veracidad estamos discutiendo, es el momento de utilizar los datos observados. El valor observado de la variable X es $X_{\text{obs}} = 14$, y la probabilidad de observar algo tanto o más extremo que lo observado es $\mathbf{P}\{X \geq X_{\text{obs}}\} = \mathbf{P}\{X \geq 14\}$. Esta probabilidad la podemos calcular sumando

$$\mathbf{P}\{X \geq 14\} = \sum_{k=14}^{20} \mathbf{P}\{X = k\} = \frac{1}{2^{20}} \sum_{k=14}^{20} \binom{20}{k} \approx 0{,}13.$$

Esta probabilidad está representada por la suma de las alturas de las barras rayadas en rojo en la figura de abajo.



Hasta aquí llegan las cuentas. La conclusión es que si estuviera tirando a embocar, embocaría a 14 o más tarjetas con 13 % de chances. No es una probabilidad demasiado baja como para concluir que la afirmación es falsa sin dudarlo, pero sugiere que el resultado no es puramente causa del azar.

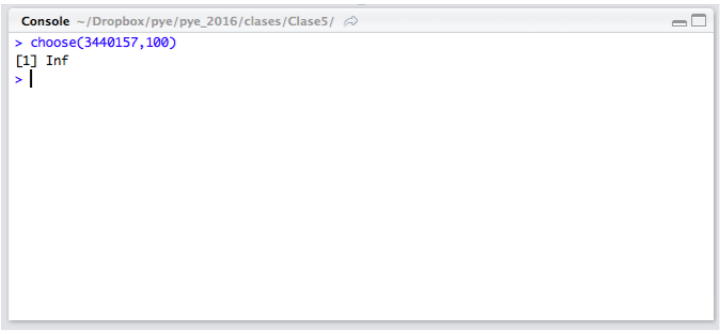
4-2. ¿Con o sin reposición?

Muchas veces lo natural es modelar un experimento usando el muestreo sin reposición. Por ejemplo, si queremos estimar la cantidad de hombres y mujeres en la población uruguaya, y nos disponemos a hacer esto a través de una encuesta, lo natural es tomar una muestra de la población sin incluir dos veces a la misma persona.

Supongamos que tomamos una muestra de $n = 100$ uruguayos, y nos interesa calcular la probabilidad de que haya k mujeres en la muestra, entonces debemos calcular

$$\mathbf{P}\{X = k\} = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

en donde N es el total de la población uruguaya, y K es el total de mujeres uruguayas en el momento de la muestra. Asumiendo que $N = 3440157$, al intentar calcular las combinaciones en el denominador con una computadora personal, el resultado es el que se muestra en la siguiente figura:



En otras palabras, el denominador en la fórmula de la probabilidad es un número extremadamente grande que la computadora no puede manejar.

Sin embargo, como la cantidad N de bolas en la urna es muy grande, es poco probable que si tomamos la muestra con reposición, se elija dos veces la misma bola. Esto sugiere que para una urna con muchas bolas, ambos modelos son en la práctica equivalentes. Vamos a probar que efectivamente la hipergeométrica se puede aproximar por la binomial cuando la urna es suficientemente grande.

Supongamos que tenemos una urna con N bolas de las cuales K son rosadas y extraemos con reposición una muestra de tamaño n. La cantidad n de bolas que extraemos será fija y veremos que sucede cuando N es grande. Supondremos ademas que la proporción de bolas rosadas en la urna es aproximadamente constante e igual a $p = K/N$. Dicho de otro modo, la cantidad de bolas rosadas es $K = pN$.

Consideremos el evento

$$D = \{(\omega_1, \ldots, \omega_n) \in \Omega : \omega_i \neq \omega_j \text{ si } i \neq j\},$$

que consiste en aquellas muestras en las cuales las bolas extraídas son todas distintas. Vamos a probar primero que la probabilidad de D tiende a 1 cuando N tiende a infinito. El argumento es muy similar al del ejercicio de los cumpleaños.

Por un lado, como todas las secuencias son igualmente probables, tenemos que

$$\mathbf{P}\{D\} = \frac{|\{(\omega_1, \ldots, \omega_n) \in \Omega : \omega_i \neq \omega_j \text{ si } i \neq j\}|}{|\Omega|}.$$

El cardinal de D es igual a

$$|\{(\omega_1, \ldots, \omega_n) \in \Omega : \omega_i \neq \omega_j \text{ si } i \neq j\}| = N(N-1)\cdots(N-(n-1)),$$

y el cardinal de Ω es igual a N^n. Por tanto

$$\mathbf{P}\{D\} = \frac{N(N-1)\cdots(N-(n-1))}{N\cdots N} = \prod_{i=0}^{n-1}\left(1 - \frac{i}{N}\right).$$

Notar que la productoria que aparece en el lado derecho tiene un número fijo, igual a n, de factores. Al hacer N tender a infinito, todos los factores tienden a uno, y por lo tanto también la productoria. En conclusión $\mathbf{P}\{D\} \to 1$ cuando $N \to \infty$.[1]

Llamemos X a la variable aleatoria que cuenta el número de bolas rosadas en la muestra. Sabemos que X tiene distribución binomial de parámetros n y p, y por lo tanto, para cada k entre 0 y n tenemos que

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

¿Cuál es la distribución de X condicionada a que D ha ocurrido? En otras palabras, cuál es la probabilidad $\mathbf{P}\{X = k \big| D\}$.

Por definición

$$\mathbf{P}\{X = k \big| D\} = \frac{\mathbf{P}\{\{X = k\} \cap D\}}{\mathbf{P}\{D\}}.$$

[1] Aunque no lo hemos escrito explícitamente, para no cargar la notación, el conjunto D depende de N.

La probabilidad de D ya la hemos calculado. Debemos contar ahora cuántas secuencias pertenecen al evento $\{X = k\} \cap D$, esto es, cuántas secuencias tienen exactamente k bolas rosadas y además todas las bolas son distintas. Es el mismo conteo que hicimos para deducir la función de probabilidad puntual de la hipergeométrica. Entonces

$$|\{X = k\} \cap D| = \binom{n}{k}(K)_k(N-K)_{n-k}.$$

Para calcular la probabilidad $\mathbf{P}\{\{X = k\} \cap D\}$ basta dividir el cardinal anterior por N^n, y esto da

$$\mathbf{P}\{\{X = k\} \cap D\} = \frac{\binom{n}{k}(K)_k(N-K)_{n-k}}{N^n}.$$

Finalmente basta dividir entre la probabilidad de $\mathbf{P}\{D\} = (N)_n/N^n$ para obtener

$$\mathbf{P}\left\{X = k \middle| D\right\} = \frac{\binom{n}{k}(K)_k(N-K)_{n-k}}{(N)_n}.$$

Re-agrupando términos se llega a la fórmula de la distribución hipergeométrica.

Conclusión: La distribución hipergeométrica coincide con la distribución binomial condicionada a que todas las bolas sean distintas.

Usando los dos hechos que probamos hasta ahora podemos probar el siguiente teorema.

Teorema (Aproximación de la hipergeométrica por la binomial)**.** *Sea X_N una variable con distribución hipergeométrica de parámetros N, K, y n. Suponemos que n está fijo y que $K/N = p \in [0,1]$. Entonces, para todo k entre 0 y n, tenemos*

$$\lim_{N\to\infty} \mathbb{P}(X_N = k) = \binom{n}{k}p^k(1-p)^{n-k}$$

Es decir, la distribución de X_N tiende a la distribución de una variable binomial de parámetros n y p.

Demostración. Consideremos como antes una urna con N bolas, de las cuales K son rosadas. Extraemos n bolas de la urna con reposición, y denotamos por X la cantidad de bolas rosadas en la muestra. Por lo que vimos antes, sabemos que

$$\mathbf{P}\{X_N = k\} = \mathbf{P}\{X = k|D\} = \frac{\mathbf{P}\{\{X = k\} \cap D\}}{\mathbf{P}\{D\}}.$$

El denominador $\mathbf{P}\{D\} \to 1$ cuando N tiende a infinito. Para el numerador, observar que

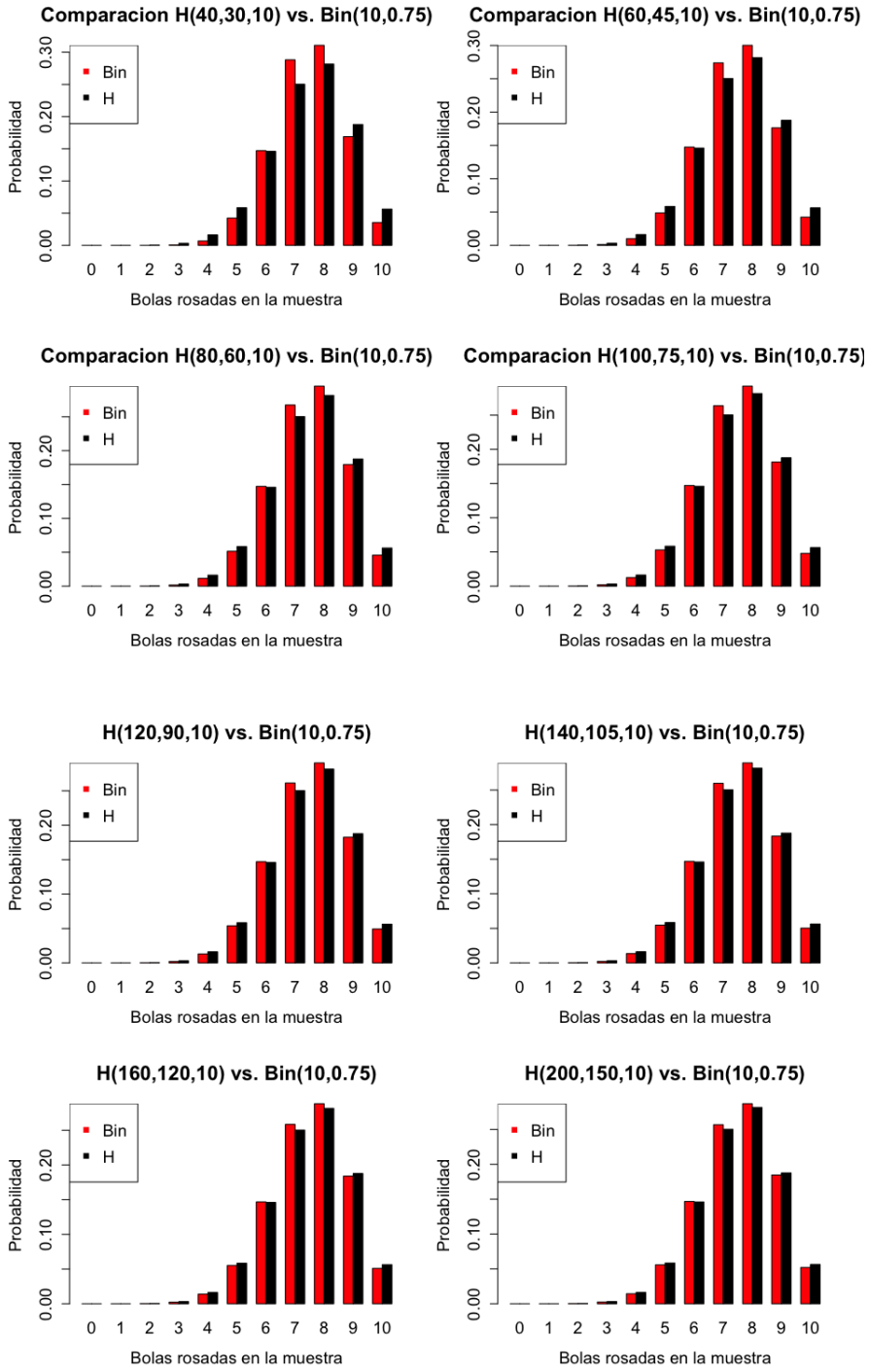
$$\mathbf{P}\{\{X = k\} \cap D\} \leq \mathbf{P}\{X = k\} \leq \mathbf{P}\{\{X = k\} \cap D\} + \mathbf{P}\{D^c\}.$$

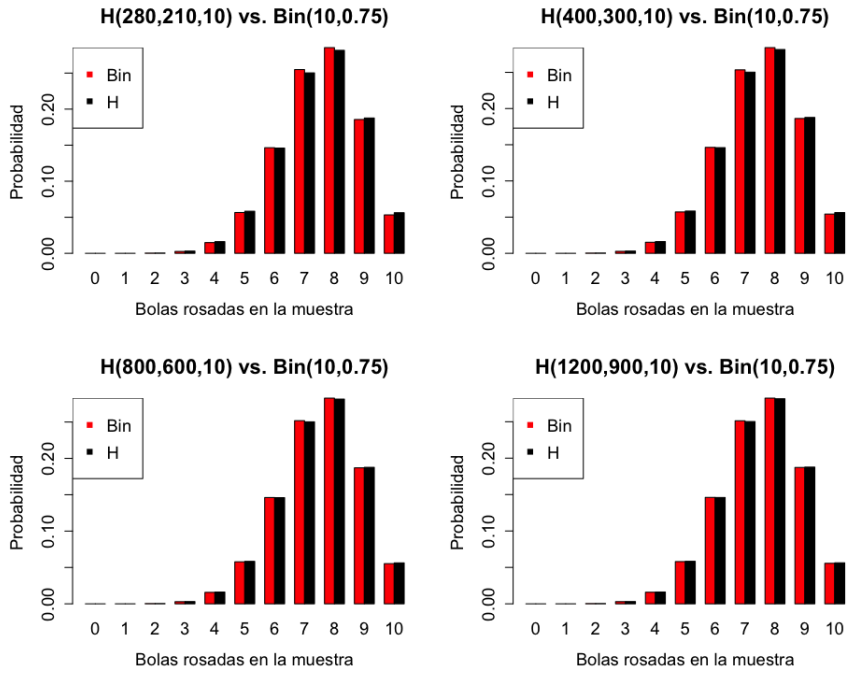
Como $\mathbf{P}\{D^c\} \to 0$, vemos que

$$\mathbf{P}\{\{X = k\} \cap D\} \to \mathbf{P}\{X = k\} = \binom{n}{k}p^k(1-p)^{n-k},$$

que es lo que queríamos probar. □

En las figuras que siguen se muestra en un ejemplo como la distribución hipergeométrica converge a la binomial. En este caso $p = 0{,}75$, $n = 10$, y N varía de 40 a 1200. Notar que K varía de forma tal que $K = pN$ en todos los casos.





Comentario sobre cómo usar la aproximación

Usar la aproximación es muy simple. Si queremos calcular alguna probabilidad en la que esté involucrada una variable X con distribución hipergeométrica de parámetros N, K, y n, debemos hacer lo siguiente:

1. Calculamos el valor de $p = K/N$.
2. Suponemos que la distribución de X es binomial de parámetros n y p.

Sin embargo, el teorema de que probamos más arriba no nos dice cuál es el error que cometemos al usar la aproximación. El error no es difícil de calcular, pero las cuentas son un poco tediosas. Se puede probar que el error *relativo* al aproximar la hipergeométrica por la binomial es a lo sumo

$$\epsilon_{\text{rel}} \leq \frac{2n^2}{p(1-p)N}.$$

Viene bien recordar la distinción entre el error relativo y el error absoluto cuando hacemos una aproximación. Supongamos que queremos aproximar una probabilidad p usando un valor aproximado p_{aprox}. Entonces

- El *error absoluto* es $\epsilon_{\text{abs}} = |p - p_{\text{aprox}}|$;
- El *error relativo* es $\epsilon_{\text{rel}} = \left|\frac{p}{p_{\text{aprox}}} - 1\right|$.

Notar que el error relativo refiere a la diferencia entre 1 y el cociente entre p y p_{rel}. Como las probabilidades pueden ser números muy pequeños, no tiene mucho sentido calcular errores absolutos. Por eso, cuando estamos trabajando con probabilidades, es mejor considerar los errores relativos en las aproximaciones.

Por ejemplo, en ejemplo con el cual empezamos esta sección (tomar una muestra de tamaño $n = 100$ de la probalción uruguaya), tenemos que $N = 3440157$ y $K = 1777273$. De aquí resulta que $p = 0{,}5166$. Si usamos la aproximación binomial, la fórmula del error relativo nos asegura que cometemos un error de a lo sumo 0,0233. Es decir, un error menor al 2 %.

4-3. Descomposición en sumas

Tanto la distribución hipergeométrica como la binomial pueden descomponerse como una suma de variables que valen cero o uno. Las variables que solo toman los valores cero o una son importantes y se llaman variables *Bernoulli*. Son las variables más simples, y su distribución queda determinada por la probabilidad con la cual toman el valor uno. Si X puede valer 0 o 1 y $\mathbf{P}\{X = 1\} = p$ decimos que X tiene distribución Bernoulli de parámetro p. Esto lo escribimos $X \sim \text{Ber}(p)$.

Veamos entonces como podemos escribir la distribución hipergeométrica como suma de variables Bernoulli. Supongamos que la urna tiene N bolas, de las cuales K son rosadas y extraemos sin reposición una muestra de tamaño n.

Consideremos las n variables $X_1, \ldots, X_n$ que indican si la i-ésima bola es rosada o no. Esto es

$$X_i(\omega) = \begin{cases} 1 & \text{si } \omega_i \text{ es rosada,} \\ 0 & \text{si } \omega_i \text{ es celeste.} \end{cases}$$

Entonces el número total de bolas rosadas está dado por $X = X_1 + \cdots + X_n$.

Comencemos por determinar la distribución de cada X_i. Para esto basta con calcular la probabilidad de que X_i sea igual a 1. Entre todas las secuencias de Ω debemos contar cuántas tienen la i-ésima coordenada rosada. Debemos elegir cuál es la bola rosada entre las K posibles, y el resto de las bolas las elegimos con la única restricción de que la muestra sea sin reposición. Entonces

$$\mathbf{P}\{X_i = 1\} = \frac{|\{\omega \in \Omega : \omega_i \text{ es rosada}\}|}{|\Omega|} = \frac{K\Big[(N-1)\cdots(N-(n-1))\Big]}{N(N-1)\cdots(N-(n-1))} = \frac{K}{N} = p.$$

Es decir, la variable X_i vale uno con probabilidad p en donde p es la proporción K/N de bolas rosadas en la urna. Notar que todas las variables $X_1, \ldots, X_n$ tienen la misma distribución $\text{Ber}(p)$.

En el caso con reposición podemos definir las variables $X_1, \ldots, X_n$ de la misma forma. En este caso, la probabilidad de que X_i sea igual a uno también es p. El razonamiento es un poco distinto: tenemos N posibilidades para cada una de las $n-1$ coordenadas distintas de la i-ésima, y K posibles bolas rosadas para elegir la i-ésima coordenada. Por lo tanto

$$\mathbf{P}\{X_i = 1\} = \frac{|\{\omega \in \Omega : \omega_i \text{ es rosada}\}|}{|\Omega|} = \frac{KN^{n-1}}{N^n} = \frac{K}{N} = p.$$

En conclusión, en cualquiera de los dos casos las variables X_i tienen distribución Bernoulli de parámetro p.

Sin embargo, para disponer de toda la información relevante a estas variables debemos saber cuál es la dependencia entre ellas. Veamos primero el caso con reposición. Consideremos una secuencia arbitraria de $0's$ y $1's$, de largo n, que denotaremos por $(x_1, \ldots, x_n)$. Queremos calcular

$$\mathbf{P}\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}.$$

Las condiciones sobre las coordenadas son: si $x_i = 1$ queremos que la bola sea rosada, y si $x_i = 0$ que la bola sea celeste. En el primer caso tenemos K posibilidades, y en el segundo $N - K$. Si r es la cantidad de $i's$ tales que $x_i = 1$, entonces

$$\mathbf{P}\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} = \frac{K^r(N-K)^{n-r}}{N^n} = p^r(1-p)^{n-r}.$$

Por otro lado, para cada una de las coordenadas tenemos que

$$\mathbf{P}\{X_i = x_i\} = \begin{cases} p & \text{si } x_i = 1 \\ 1-p & \text{si } x_i = 0 \end{cases}$$

por lo que

$$\mathbf{P}\{X_1 = x_1\}\,\mathbf{P}\{X_2 = x_2\} \cdots \mathbf{P}\{X_n = x_n\} = p^r(1-p)^{n-r}.$$

Juntando ambas igualdades deducimos que

$$\mathbf{P}\{X_1 = x_1, \ldots, X_n = x_n\} = \mathbf{P}\{X_1 = x_1\} \cdots \mathbf{P}\{X_n = x_n\}.$$

Esto significa que las variables $X_1, \ldots, X_n$ son independientes. Esto es completamente claro desde un punto de vista intuitivo: si extraemos la muestra con reposición, el resultado de la i-ésima extracción no influye sobre el resultado de las otras extracciones.

Claramente este no es el caso si la muestra la obtenemos sin reposición. Veamos por ejemplo qué ocurre con X_1 y X_2. Por un lado, la probabilidad

$$\mathbf{P}\{X_1 = 1, X_2 = 1\} = \frac{K(K-1)\Big[(N-2)\cdots(N-(n-1))\Big]}{N(N-1)\cdots(N-(n-1))} = \frac{K(K-1)}{N(N-1)}.$$

Sin embargo, las probabilidades por separado son iguales a p, de donde el producto

$$\mathbf{P}\{X_1 = 1\}\,\mathbf{P}\{X_2 = 1\} = p^2 = \frac{K^2}{N^2} \neq \frac{K(K-1)}{N(N-1)} = \mathbf{P}\{X_1 = 1, X_2 = 1\}.$$

Esto muestra que X_1 y X_2 no son independientes.

En resumen: si X denota la cantidad de bolas rosadas en la muestra, entonces

$$X = X_1 + \cdots + X_n$$

con

- $X_1, \ldots, X_n$ independientes si la muestra es con reposición;
- $X_1, \ldots, X_n$ no independientes si la muestra es sin resposición.

En ambos casos $X_i \sim \text{Ber}(p)$ para todo $i = 1, \ldots, n$.

4-4. Valor esperado de variables discretas

El objetivo de esta sección es responder a la pregunta ¿Si repitiéramos varias veces el experimento de extraer muestras de una urna, cuántas bolas rosadas veríamos en promedio? Para esto debemos introducir el concepto de valor esperado de una variable aleatoria.

Motivación

Comencemos por un ejemplo simple. Supongamos que somos gerentes de un pequeño emprendimiento en el cual cada dos semanas debemos decidir entre dos opciones:

1. cerramos un negocio seguro que nos provee una ganancia neta de \$1 500;
2. o realizamos una inversión que de salir bien nos aportaría una ganancia neta de \$3 000, pero de salir mal conllevaría una perdida neta de \$1 500. Además, en este caso estimamos que la probabilidad de que la inversión sea exitosa es 0,75.

Imaginemos que tomamos la decisión de invertir en n semanas consecutivas. Denotemos por n_+ el número de veces que la inversión ha resultado exitosa, y $n_- = n - n_+$ el número de veces que dio pérdidas. Entonces, las ganancias totales $G(n)$ en esas n semanas son

$$G(n) = 3\,000n_+ - 1\,500n_-.$$

Si queremos calcular las ganancias por semana de nuestro negocio, debemos dividir por el número n de semanas, de donde

$$g(n) = \frac{G(n)}{n} = 3\,000\frac{n_+}{n} - 1\,500\frac{n_-}{n}.$$

¿Qué ocurre a la larga con las ganancias por semana $g(n)$? De la interpretación frecuentista de la probabilidad asumimos que las frecuencias relativas n_+/n y n_-/n convergen, cuando n tiende a infinito, a las probabilidades de que la inversión sea exitosa o fracase respectivamente. Entonces

$$\begin{aligned}\lim_{n\to\infty} g(n) &= 3\,000 \lim_{n\to\infty} \frac{n_+}{n} - 1\,500 \lim_{n\to\infty} \frac{n_-}{n} \\ &= 3\,000\,(\text{proba. de éxito}) - 1\,500\,(\text{proba. de fracaso}) \\ &= 3\,000 \cdot 0{,}75 - 1\,500 \cdot 0{,}25 = 1\,875.\end{aligned}$$

Es decir, a medida que n crece, las ganancias por semana se aproximan más y más al valor \$1 875.

Si hubiéramos optado por la opción segura, las ganancias por semana serían iguals a $g(n) = 1\,500$. Como las ganancias por semana son mayores para la opción 2 que para la opción 1, es mejor arriesgar invirtiendo el dinero, siempre y cuando seamos capaces de invertir durante una cantidad grande de semanas.

Si pensamos a las ganancias semanales como una variable aleatoria G, que toma los valores $3\,000$ y $-1\,500$ con probabilidades 0,75 y 0,25 respectivamente, entonces la cantidad $1\,875$ que calculamos más arriba se llama *el valor esperado* de G. Esto lo escribimos $\mathbf{E}\,(G) = 1\,875$.

Definición

La misma idea nos sirve como motivación para definir el valor esperado de una variable discreta en general. Supongamos que X es una variable discreta cuyo recorrido es $R_X = \{x_1, x_2, \ldots\}$. Imaginemos que realizamos el experimento n veces y para cada una de estas registramos el valor de X. Llamemos a estos valores por $X(1), X(2), \ldots, X(n)$. Cada uno de los $X(i)$ puede ser igual a cualquiera de los valores posibles de X (los valores del recorrido de X).

El promedio de las n realizaciones de X es

$$\text{Prom}(X(1), \ldots, X(n)) = \frac{X(1) + \cdots X(n)}{n}.$$

Podemos reordenar los valores $X(1), \ldots, X(n)$ y agruparlos de acuerdo a su valor de modo que la suma

$$X(1) + \cdots + X(n) = n_1 x_1 + n_2 x_2 + \cdots ,$$

en donde n_j es el número de veces que ha ocurrido el valor x_j. En símbolos esto lo podemos escribir como

$$n_j = |\{i : X(i) = x_j\}| .$$

Con esta forma de escribir la suma, tenemos que

$$\text{Prom}(X(1), \ldots, X(n)) = x_1 \frac{n_1}{n} + x_2 \frac{n_2}{n} + \cdots .$$

Al realizar más veces el experimento, y hacer n tender a infinito, las frecuencias relativas convergen a

$$\frac{n_j}{n} \to \mathbf{P}\{X = x_j\} .$$

De este modo, el valor "por ensayo" de X, para n tendiendo a infinito queda

$$\sum_{j=1}^{\infty} x_j \mathbf{P}\{X = x_j\} .$$

Esto es lo que definiremos como el valor esperado de X.

Definición. *Sea X una variable aleatoria discreta cuyo recorrido es $R_X = \{x_1, x_2, \ldots\}$. Definimos el valor esperado de X (o la esperanza de X) como*

$$\mathbf{E}(X) = \sum_{j=1}^{\infty} x_j \mathbf{P}\{X = x_j\} .$$

Observar que la serie anterior en la definición de valor esperado se puede escribir también como

$$\mathbf{E}(X) = \sum_{x \in R_X} x \mathbf{P}\{X = x\} ,$$

que puede resultar útil, sobre todo cuando no tenemos una preferencia natural para ordenar los valores del recorrido de X.

Linealidad del valor esperado

Calcular el valor esperado en la práctica puede ser bastante engorroso. Veremos una propiedad, la linealidad, que nos va a permitir calcular el valor esperado de sumas de variables de forma más simple.

Proposición. *Sean X e Y dos variables discretas. Entonces la suma $X+Y$ es una variable aleatoria discreta y su valor esperado es*

$$\mathbf{E}\,(X+Y) = \mathbf{E}\,(X) + \mathbf{E}\,(Y)\,.$$

Demostración. El recorrido de $X+Y$ es

$$R_{X+Y} = \{x+y : x \in R_X, y \in R_Y\},$$

que es numerable, por lo que $X+Y$ es discreta.

Notar que estos valores de la suma pueden repetirse, ya que $x+y$ puede ser igual a $x'+y'$ aunque $x \neq x'$ e $y \neq y'$. Sin embargo, para cada valor posible z de la suma, tenemos que el evento $\{X+Y=z\}$ se descompone como unión disjunta

$$\{X+Y=z\} = \bigcup_{x+y=z} \{X=x, Y=y\},$$

en todas las formas distintas de escribir z como suma de algún x y algún y.

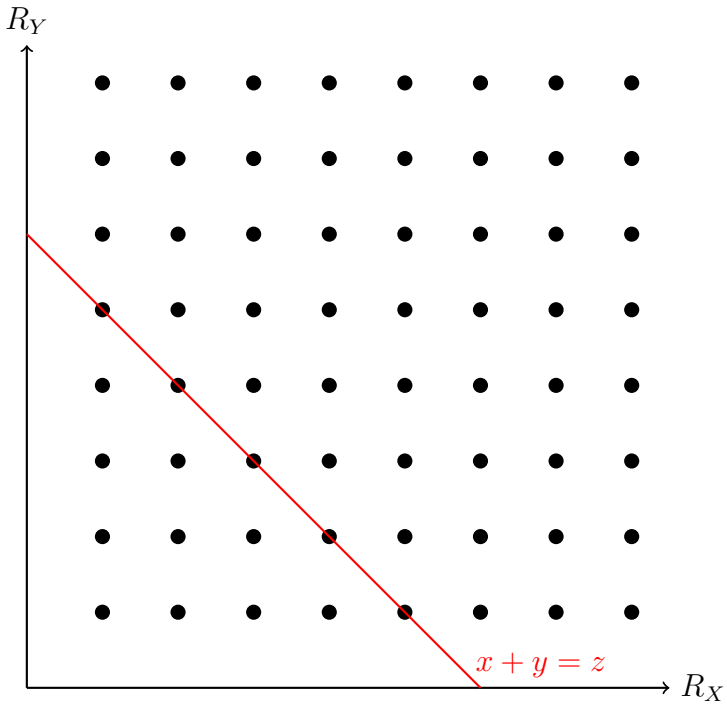
Al tomar probabilidades, obtenemos

$$\mathbf{P}\,\{X+Y=z\} = \sum_{x+y=z} \mathbf{P}\,\{X=x, Y=y\}\,.$$

Nos resta ahora sumar en z:

$$\begin{aligned}
\mathbf{E}\,(X+Y) &= \sum_{z\in R_{X+Y}} z\mathbf{P}\,\{X+Y=z\} = \sum_{z\in R_{X+Y}} z \sum_{x+y=z} \mathbf{P}\,\{X=x, Y=y\} \\
&= \sum_{z\in R_{X+Y}} \sum_{x+y=z} z\mathbf{P}\,\{X=x, Y=y\} \\
&= \sum_{z\in R_{X+Y}} \sum_{x+y=z} (x+y)\,\mathbf{P}\,\{X=x, Y=y\}
\end{aligned}$$

Aquí viene un punto ligeramente sutil de la demostración. Este consiste en notar que sumar en aquellos $x's$ e $y's$ que suman z y luego sumar en todos los valores posibles de z, es lo mismo que sumar en todos los valores posibles de x e y. Esto se puede ver mejor con un dibujo, como el que se muestra en la página siguiente. En el mismo vemos la diagonal roja que corresponde a todos los valores de x e y que suman un cierto valor de z. Claramente, al variar z, las diagonales cubren todo el cuadrante. El cuadrante corresponde a todos los pares posibles de x e y.



De esta forma podemos re-escribir la suma anterior como

$$\sum_z \sum_{x+y=z} (x+y)\,\mathbf{P}\{X=x, Y=y\} = \sum_{x,y} x\mathbf{P}\{X=x, Y=y\} + \sum_{x,y} y\mathbf{P}\{X=x, Y=y\}.$$

Notar que para cada x fijo, la suma

$$\sum_{y\in R_Y} \mathbf{P}\{X=x, Y=y\} = \mathbf{P}\{X=x\}.$$

De aquí resulta que el primer término en la suma anterior es igual a

$$\sum_{x\in R_X}\sum_{y\in R_Y} x\mathbf{P}\{X=x, Y=y\} = \sum_{x\in R_X} x\mathbf{P}\{X=x\} = \mathbf{E}(X).$$

Un razonamiento análogo muestra que el segundo término es igual a $\mathbf{E}(Y)$. Esto termina la demostración. □

Otra propiedad útil es la siguiente.

Proposición. *Sean X una variable aleatoria discreta y c una constante. Entonces*

$$\mathbf{E}(cX) = c\mathbf{E}(X).$$

Demostración. Directamente de la definición tenemos que

$$\mathbf{E}(cX) = \sum_{x\in R_X} cx\mathbf{P}\{X=x\} = c\sum_{x\in R_X} x\mathbf{P}\{X=x\} = c\mathbf{E}(X),$$

que es lo que queríamos demostrar. □

Valor esperado de bolas rosadas en la muestra

Volviendo a las urnas, queremos calcular el número esperado de bolas rosadas que aparecen en la muestra. La mejor forma de hacer esto es usando la propiedad de linealidad. Notar que en ambos casos, en el muestreo con o sin reposición, podemos escribir la cantidad X de bolas rosadas en la muestra como una suma

$$X = X_1 + \cdots X_n.$$

Entonces, el valor esperado de X es $\mathbf{E}(X) = \mathbf{E}(X_1) + \cdots + \mathbf{E}(X_n)$.

Además, vimos que las variables X_i (que indican si la i-ésima bola es rosada o no) tienen todas distribución Bernoulli de parámetro $p = K/N$. Entonces

$$\mathbf{E}(X_i) = 1 \cdot \mathbf{P}\{X_i = 1\} + 0 \cdot \mathbf{P}\{X_i = 0\} = p.$$

Finalmente, de aquí resulta que $\mathbf{E}(X) = np$.

Recordar que esto lo interpretamos de la siguiente manera: si repitiéramos n veces el experimento, a la larga a medida que n crece, observaríamos que en promedio np de las bolas de la muestra son rosadas.

En particular tenemos que:

- Si $X \sim \text{Bin}(n, p)$, entonces $\mathbf{E}(X) = np$.
- Si $X \sim \mathcal{H}(N, K, n)$, entonces $\mathbf{E}(X) = np$ con $p = K/N$.

4-5. Poisson, Student y la producción de cerveza

En la producción de cerveza es importante conocer con precisión la cantidad de levadura utilizada para la fermentación. Demasiada levadura produce una cerveza amarga, y poca hace que la cerveza no fermente bien. En 1904, la Guiness contrató al matemático Student[2] (William Gosset) para mejorar la calidad de la cerveza. Student ideó un modelo probabilístico para controlar la cantidad de levadura usada en la fermentación.

El modelo de Student era el siguiente: supongamos que un líquido (un cultivo diluido de células de levadura) es vertido y extendido sobre una placa, formando una capa delgada de 0,01 mm de espesor. Queremos estimar la cantidad promedio μ de células de levadura por unidad de área en la placa. La unidad de área es $1/400$ mm^2 y la placa tiene en total N unidades de área. Debemos imaginarnos la placa dividida en N cuadrados de área $1/400$ mm^2.

Asumiendo que el líquido ha sido bien mezclado, una célula dada tendrá la misma probabilidad de caer en cualquier unidad de área. En total hay μN células.

Fijemos C una unidad de área cualquiera de la placa, que llamaremos una celda, y sea X el número de células de levadura que caen en C. Nuestro primer objetivo es determinar la distribución de X.

[2]William Gosset usaba el seudónimo Student para publicar sus investigaciones debido a que la compañía Guiness no autorizaba a sus empleados a revelar información importante sobre el proceso de producción de cerveza.

La variable X puede tomar cualquier valor entre 0 y μN. Podemos pensar el experimento como un experimento con urnas. Las distintas celdas son las N bolas numeradas en la urna, y extraemos una muestra de tamaño μN y con reposición. De este modo la i-ésima coordenada de la muestra obtenida representa la celda en la cual cayó la i-ésima célula. Estamos asumiendo implícitamente que las células son distinguibles entre ellas, podemos imaginarnos que cada célula tienen un nombre própio que la identifica (ahí va la célula llamada Carlos... por ejemplo).

Con este modelo, la variable X cuenta cuántas coordenadas tienen la letra C. Por lo tanto X tiene distribución binomial de parámetros $n = \mu N$ y $p = 1/N$. Recordar que N es la cantidad total de celdas.

Entonces, para cada k entre 0 y μN, tenemos que

$$\mathbf{P}\{X = k\} = \binom{\mu N}{k} \left(\frac{1}{N}\right)^k \left(1 - \frac{1}{N}\right)^{\mu N - k}.$$

Notar que los parámetros de la distribución binomial de X satisfacen la relación $np = \mu$. Además, el número de unidades N es cercano al millón. Usaremos estas dos propiedades para aproximar la distribución de X por otra más adecuada a los cálculos. Vamos a tomar el límite cuando N tiende a infinito en la fórmula de la función de probabilidad puntual de X.

Teorema (Aproximación de Poisson a la binomial)**.** *Para cada $n \in \mathbb{N}$, sea X_n una variable con distribución binomial de parámetros n y p_n. Supongamos que*

$$np_n = \mu$$

para un cierto parámetro fijo $\mu > 0$. Entonces, para todo $k \in \mathbb{N}$, tenemos que

$$\lim_{n\to\infty} \mathbf{P}\{X_n = k\} = \frac{\mu^k}{k!} e^{-\mu}.$$

Demostración. Fijemos k un natural cualquiera. Como X_n tiene distribución binomial $\mathrm{Bin}(n, p_n)$, y $np_n = \mu$, entonces

$$\begin{aligned}
\mathbf{P}\{X_n = k\} &= \binom{n}{k}(1-p_n)^{n-k}p_n^k = \binom{n}{k}\left(1 - \frac{\mu}{n}\right)^{n-k}\left(\frac{\mu}{n}\right)^k \\
&= \frac{n!}{k!(n-x)!}\left(1 - \frac{\mu}{n}\right)^{n-k}\left(\frac{\mu}{n}\right)^k \\
&= \frac{\mu^k}{k!}\Big[n(n-1)\cdots(n-k+1)\Big]\left(1 - \frac{\mu}{n}\right)^n \frac{1}{\left(1-\frac{\mu}{n}\right)^k}\left(\frac{1}{n}\right)^k
\end{aligned}$$

Estudiemos la última ecuación por separado. Primero, observar que cuando $n \to \infty$ la productoria

$$n(n-1)\cdots(n-k+1) \sim n^k,$$

pues estamos multiplicando un número fijo, igual a k, de factores. A su vez, deben recordar de los cursos de Cálculo que

$$\left(1 - \frac{\mu}{n}\right)^n \to e^{-\mu},$$

cuando $n \to \infty$. Por último, el factor

$$\left(1 - \frac{\mu}{n}\right)^k \to 1,$$

pues k está fijo y n es cada vez más grande.

Luego

$$\mathbf{P}\{X_n = k\} \sim \frac{\mu^k}{k!} n^k \left(1 - \frac{\mu}{n}\right)^n \frac{1}{n^k} \to \frac{\mu^k e^{-\mu}}{k!}.$$

Esto es exactamente lo que queríamos probar. □

El nombre del teorema resulta de que la distribución límite a la que hemos llegado es bien conocida: se llama distribución de *Poisson* de parámetro μ.

Definición. *Una variable aleatoria X tiene distribución de Poisson de parámetro μ si toma valores naturales y su función de probabilidad está dada por*

$$\mathbf{P}\{X = k\} = \frac{\mu^k e^{-\mu}}{k!},$$

para todo $k \geq 0$ natural. Esto lo escribimos $X \sim \text{Pois}(\mu)$.

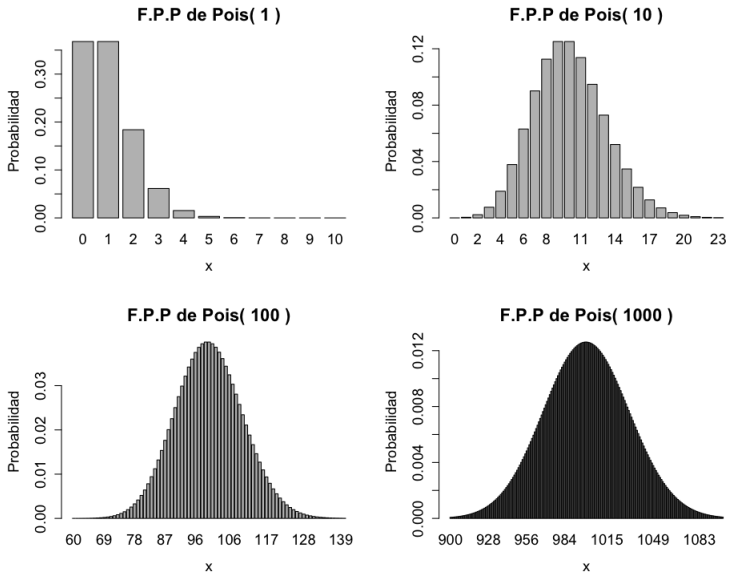
Notar que la suma de las probabilidades es igual a uno (¡como debe ser!). De hecho, la función exponencial se puede escribir como la serie

$$e^{\mu} = \sum_{k=0}^{\infty} \frac{\mu^k}{k!},$$

por lo que

$$\sum_{k=0}^{\infty} \frac{\mu^k e^{-\mu}}{k!} = e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!} = e^{-\mu} e^{\mu} = 1.$$

En la figura mostramos la función de probabilidad puntual de una variable X con distribución de Poisson para varios valores del parámetro μ. El valor de μ varía de 1 a 1 000. Notar como a medida que μ crece, la distribución de X se corre hacia la derecha y se concentra en valores cada vez mayores de k.



Es interesante notar la forma acampanada y simétrica de la distribución cuando μ es grande. Esto no es casualidad y lo estudiaremos más adelante.

Volviendo al ejemplo de Student y la producción de cerveza, el resultado del teorema nos dice que la variable X que cuenta el número de células de levadura que caen en determinada celda de la placa, debe tener distribución de Poisson de parámetro μ. Por lo menos en forma aproximada, pues N es del orden del millón.

Para comprobar su modelo, Student observó en el microscopio un área cuadrada de 1 mm^2 de la placa, que estaba dividida en 400 celdas cuadradas (esto se hace con un aparato llamado hemocitómetro), y contó el número de células en cada una de ellas.

El resultado está en la tabla siguiente[3]:

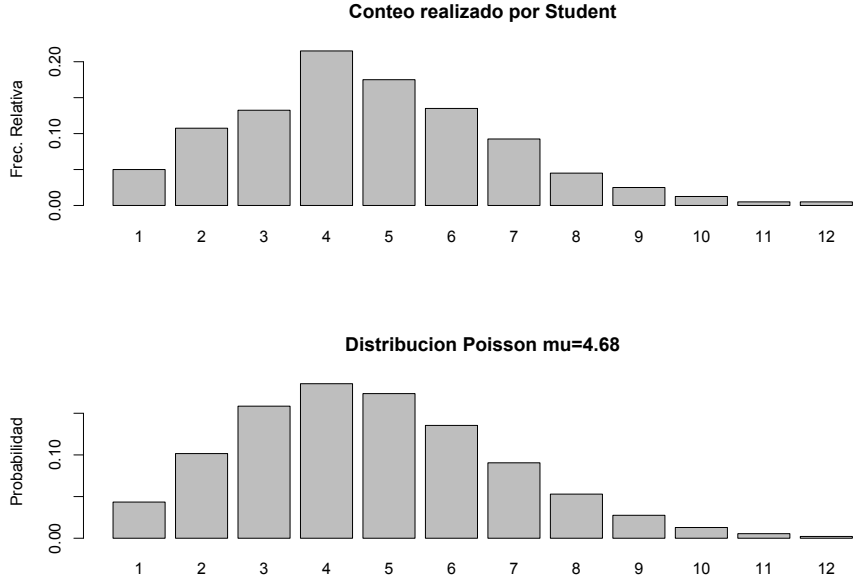
Distribución de células de levadura sobre 1 mm^2 *dividido en* 400 *cuadrados.*

2	2	4	4	4	5	2	4	7	7	4	7	5	2	8	6	7	4	3	4
3	3	2	4	2	5	4	2	8	6	3	6	6	10	8	3	5	6	4	4
7	9	5	2	7	4	4	2	4	4	4	3	5	6	5	4	1	4	2	6
4	1	4	7	3	2	3	5	8	2	9	5	3	9	5	5	2	4	3	4
4	1	5	9	3	4	4	6	6	5	4	6	5	5	4	3	5	9	6	4
4	4	5	10	4	4	3	8	3	2	1	4	1	5	6	4	2	3	3	3
3	7	4	5	1	8	5	7	9	5	8	9	5	6	6	4	3	7	4	4
7	5	6	3	6	7	4	5	8	6	3	3	4	3	7	4	4	4	5	3
8	10	6	3	3	6	5	2	5	3	11	3	7	4	7	3	5	5	3	4
1	3	7	2	5	5	5	3	3	4	6	5	6	1	6	4	4	4	6	4
4	2	5	4	8	6	3	4	6	5	2	6	6	1	2	2	2	5	2	2
5	9	3	5	6	4	6	5	7	1	3	6	5	4	2	8	9	5	4	3
2	2	11	4	6	6	4	6	2	5	3	5	7	2	6	5	5	1	2	7
5	12	5	8	2	4	2	1	6	4	5	1	2	9	1	3	4	7	3	6
5	6	5	4	4	5	2	7	6	2	7	3	5	4	4	5	4	7	5	4
8	4	6	6	5	3	3	5	7	4	5	5	5	6	10	2	3	8	3	5
6	6	4	2	6	6	7	5	4	5	8	6	7	6	4	2	6	1	1	4
7	2	5	7	4	6	4	5	1	5	10	8	7	5	4	6	4	4	7	5
4	3	1	6	2	5	3	3	3	7	4	3	7	8	4	7	3	1	4	4
7	6	7	2	4	5	1	3	12	4	2	2	8	7	6	7	6	3	5	4

Los valores de la tabla representan el número de células en cada una de las 400 celdas. Para comparar estos con la distribución de Poisson es mejor representar los valores de la tabla graficamente.

Lo que hacemos es observar cuántas celdas tienen, por ejemplo, 4 células. Al dividir este número por el total de células, es decir 400, tendremos la frecuencia relativa del valor 4, lo cual nos da una idea de la probabilidad de que caigan 4 células en una celda.

[3]Los datos se pueden bajar de la página del curso.



La figura de arriba muestra las frecuencias relativas del experimento realizado por Student. En la parte de abajo se compara con la función de probabilidad puntual de la distribución de Poisson de parámetro $\mu = 4{,}68$. Más adelante veremos métodos para cuantificar cuán buena es la aproximación de Poisson, respecto a la evidencia empírica aportada por los datos.

El parámetro μ puede interpretarse como la cantidad media de células por unidad de área. Esto queda justificado porque el valor esperado de X es igual a μ.

Proposición. *Si X es una variable aleatoria con distribución de Poisson de parámetro μ, entonces $\mathbf{E}(X) = \mu$.*

Demostración. De la definición de valor esperado:

$$\mathbf{E}(X) = \sum_{k=0}^{\infty} k \frac{\mu^k e^{-\mu}}{k!} = \sum_{k=1}^{\infty} k \frac{\mu^k e^{-\mu}}{k!},$$

ya que el primer término $(k = 0)$ de la serie se anula.

Entonces

$$\mathbf{E}(X) = \sum_{k=1}^{\infty} = \frac{\mu^k e^{-\mu}}{(k-1)!} = \mu \sum_{k=1}^{\infty} \frac{\mu^{k-1} e^{-\mu}}{(k-1)!}.$$

Poniendo $j = k - 1$ obtenemos

$$\mu \sum_{k=1}^{\infty} \frac{\mu^{k-1} e^{-\mu}}{(k-1)!} = \mu \sum_{j=0}^{\infty} \frac{\mu^j e^{-\mu}}{j!} = \mu.$$

Esto es lo que queríamos demostrar. □

4-6. Resumen

En este capítulo hemos visto dos formas distintas en las que se puede tomar una muestra de una urna. Este puede ser *con reposición* o *sin reposición*. También vimos varios ejemplos en los cuales estos modelos se aplican a diversas situaciones.

1. En el muestreo con reposición: el número X de bolas rosadas en la muestra tiene distribución binomial de parámetros n y p. Esto lo escribimos $X \sim \text{Bin}(n, p)$. En este caso la función de probabilidad puntual de X es

$$\mathbf{P}\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k} \quad \forall\, k \in \mathbb{N} : 0 \leq k \leq n.$$

2. En el muestreo csin reposición: el número X de bolas rosadas en la muestra tiene distribución hipergeométrica de parámetros N, K y n. Esto lo escribimos $X \sim \mathcal{H}(N, K, n)$. En este caso la función de probabilidad puntual de X es

$$\mathbf{P}\{X = k\} = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad \forall\, k \in \mathbb{N} : \text{máx}\{0, n-(N-K)\} \leq k \leq \text{mín}\{n, K\}.$$

3. Vimos que cuando la cantidad de bolas en la urna N es grande, y la proporción de bolas rosadas es $p = K/N$, podemos aproximar la distribución hipergeométrica por una binomial de parámetros n y p.

4. La distribución de una variable X que toma los valores 0 y 1 se llama *Bernoulli*. Si $\mathbf{P}\{X = 1\} = p$, esto lo escribimos $X \sim \text{Ber}(p)$.

5. Vimos que en cualquiera de los dos casos, la cantidad de bolas rosadas X se puede escribir como una suma de n variables Bernoulli de parámetro $p = K/N$:

$$X = X_1 + \cdots + X_n$$

   Además:

   - $X_1, \ldots, X_n$ son independientes si la muestra es con reposición;
   - $X_1, \ldots, X_n$ no son independientes si la muestra es sin resposición.

6. Una variable aleatoria tiene distribución de Poisson de parámetro μ si

$$\mathbf{P}\{X = k\} = \frac{\mu^k e^{-\mu}}{k!} \quad \forall\, k \in \mathbb{N} : k \geq 0.$$

   Esto lo escribimos $X \sim \text{Pois}(\mu)$.

7. Vimos que cuando n es grande y $np = \mu$, la distribución binomial de parámetros n y p se puede aproximar por una distribución de Poisson de parámetro μ.

8. Introdujimos la noción de valor esperado (o esperanza) de una variable aleatoria discreta. Si X es discreta con recorrido R_X entonces

$$\mathbf{E}(X) = \sum_{x \in R_X} x\mathbf{P}\{X = x\}.$$

Probamos que:

- Si $X \sim \text{Bin}(n, p)$, entonces $\mathbf{E}(X) = np$. El caso particular de $n = 1$ es el caso de $X \sim \text{Ber}(p)$, y por tanto $\mathbf{E}(X) = p$.
- Si $X \sim \mathcal{H}(N, K, n)$, entonces $\mathbf{E}(X) = np$ con $p = K/N$.
- Si $X \sim \text{Poiss}(\mu)$, entonces $\mathbf{E}(X) = \mu$.

Lecturas recomendadas

Recomendamos la siguientes lecturas de la página de Wikipedia. Como siempre, a aquellos que se sientan cómodos con el inglés les sugerimos revisen también la versión en inglés de las mismas.

1. Distribución binomial
   `https://es.wikipedia.org/wiki/Distribución_binomial`
2. Distribución hipergeométrica
   `https://es.wikipedia.org/wiki/Distribución_hipergeométrica`
3. Distribución de Poisson
   `https://es.wikipedia.org/wiki/Distribución_de_Poisson`
4. William Gosset
   `https://es.wikipedia.org/wiki/William_Sealy_Gosset`
5. Siméon Denis Poisson
   `https://es.wikipedia.org/wiki/Siméon_Denis_Poisson`