

DISTRIBUCIÓN MULTINOMIAL

HÉCTOR OLIVERO Q. - VÍCTOR RIQUELME F.

1. INTRODUCCIÓN

Considere una población con artículos pertenecientes a k categorías distintas. Supongase que se extrae un artículo de dicha población, y se quiere ver de que tipo es. Podemos modelar lo anterior por una variable aleatoria X , que indica a que categoría pertenece el artículo. Llamemos y_1, \dots, y_k a las distintas categorías. Entonces X toma valores en el conjunto $\{y_1, \dots, y_k\}$, y definimos las probabilidades $p_i = \mathbb{P}(X = y_i)$. Es claro que $\sum_{i=1}^k p_i = 1$.

Supongase ahora que se toma una MAS de tamaño n con reposición (o si el tamaño de la población es “grande” se puede considerar sin reposición). Definamos el vector aleatorio $\vec{N} = (N_1, \dots, N_k)$ que indica en cada componente i -ésima la frecuencia de ocurrencia del tipo y_i en la MAS. Entonces la distribución de \vec{N} es una multinomial de parámetros n y $\vec{p} = (p_1, \dots, p_k)$:

$$\mathbb{P}(N_1 = n_1, \dots, N_k = n_k) = \binom{n}{n_1, \dots, n_k} p_1^{n_1} \dots p_k^{n_k} \mathbf{1}_{\{\sum_{i=1}^k n_i = n\}}(n_1, \dots, n_k)$$

con $\binom{n}{n_1, \dots, n_k} = \frac{n!}{n_1! n_2! \dots n_k!}$

2. ¿DE DONDE VIENE?

La deducción tiene dos patas:

- La parte de las probabilidades
- El coeficiente que acompaña, asociado a (n_1, \dots, n_k) , que llamaremos α_{n_1, \dots, n_k} .

Primero, la parte de las probabilidades es relativamente fácil. Es convencerse que la probabilidad de obtener una configuración de n_1 objetos de tipo y_1 , n_2 objetos de tipo y_2, \dots, n_k objetos de tipo y_k (si no importara el orden en que salen con respecto al total) es $p_1^{n_1} \dots p_k^{n_k}$.

La segunda parte tiene que ver con la cantidad de las configuraciones anteriores posibles. Para ello, definamos:

- $\alpha_1 = \{\text{numero de formas de elegir } n_1 \text{ art. de tipo } y_1 \text{ entre los } n \text{ disponibles}\}$

$$= \frac{n!}{(n - n_1)! n_1!}$$
- $\alpha_2 = \{\text{numero de formas de elegir } n_2 \text{ art. de tipo } y_2 \text{ entre los } n - n_1 \text{ restantes disponibles}\}$

$$= \frac{(n - n_1)!}{(n - n_1 - n_2)! n_2!}$$
- ⋮
- $\alpha_{k-1} = \{\text{numero de formas de elegir } n_{k-1} \text{ art. de tipo } y_{k-1} \text{ entre los } n - n_1 - \dots - n_{k-2} \text{ restantes disponibles}\}$

$$= \frac{(n - n_1 - \dots - n_{k-2})!}{(n - n_1 - \dots - n_{k-2} - n_{k-1})! n_{k-1}!}$$
- $\alpha_k = \{\text{numero de formas de elegir } n_k \text{ art. de tipo } y_k \text{ entre los } n - n_1 - \dots - n_{k-2} - n_{k-1} = n_k \text{ disponibles}\}$

$$= 1 = \frac{(n - n_1 - \dots - n_{k-1})!}{(n - n_1 - \dots - n_{k-1} - n_k)! n_k!}$$

Es relativamente fácil convencerse de que $\alpha_{n_1, \dots, n_k} = \alpha_1 \dots \alpha_k$, y desarrollando un poco la expresión de la derecha se obtiene el coeficiente multinomial.

3. RELACIONES ENTRE MULTINOMIAL Y BINOMIAL

Para el caso en que $k = 2$, uno puede ver que la distribución multinomial coincide con la binomial: interpretando que si no se está en la categoría y_1 , se está fuera de la categoría y_1 (llamamos y_2 a esa otra categoría). Como $p_1 + p_2 = 1$, $q \equiv p_2 = 1 - p_1$ y definimos $p = p_1$. De igual forma, $n_2 = n - n_1$. Reemplazando en la distribución multinomial los valores anteriores, se obtiene que $\mathbb{P}(N_1 = n_1, N_2 = n_2) = \mathbb{P}(N_1 = n_1)$, donde N_1 se distribuye como una binomial de parámetros n y $p = p_1$.

Para el caso en que se tienen k categorías nos interesará la distribución marginal de N_i . Primero, si seguimos el razonamiento anterior, el hecho que no se seleccione un elemento de la categoría y_i significa que se selecciona un elemento de el resto de categorías. Esto se hace con probabilidad $1 - p_i$, por lo que la distribución marginal de N_i debiera ser una binomial de parámetros n y p_i .

Hacendo el cálculo:

$$\begin{aligned}
\mathbb{P}(N_i = n_i) &= \sum_{\substack{\{n_1, \dots, n_k\} \setminus \{n_i\} \\ \sum_{j=1}^k n_j = n}} \mathbb{P}(N_1 = n_1, \dots, N_i = n_i, \dots, N_k = n_k) \\
&= \sum_{\substack{\{n_1, \dots, n_k\} \setminus \{n_i\} \\ \sum_{j \neq i} n_j = n - n_i}} \binom{n}{n_1, \dots, n_k} p_1^{n_1} \dots p_{i-1}^{n_{i-1}} p_i^{n_i} p_{i+1}^{n_{i+1}} \dots p_k^{n_k} \\
&= p_i^{n_i} \sum_{\substack{\{n_1, \dots, n_k\} \setminus \{n_i\} \\ \sum_{j \neq i} n_j = n - n_i}} \frac{n!}{n_1! \dots n_{i-1}! n_{i+1}! \dots n_k!} p_1^{n_1} \dots p_{i-1}^{n_{i-1}} p_{i+1}^{n_{i+1}} \dots p_k^{n_k} \\
&= \frac{n!}{(n - n_i)! n_i!} p_i^{n_i} \sum_{\substack{\{n_1, \dots, n_k\} \setminus \{n_i\} \\ \sum_{j \neq i} n_j = n - n_i}} \frac{(n - n_i)!}{n_1! \dots n_{i-1}! n_{i+1}! \dots n_k!} p_1^{n_1} \dots p_{i-1}^{n_{i-1}} p_{i+1}^{n_{i+1}} \dots p_k^{n_k} \\
&= \frac{n!}{(n - n_i)! n_i!} p_i^{n_i} (p_1 + \dots + p_{i-1} + p_{i+1} + \dots + p_k)^{n - n_i} \quad \text{de la fórmula multinomial} \\
&= \frac{n!}{(n - n_i)! n_i!} p_i^{n_i} (1 - p_i)^{n - n_i} \\
&= \binom{n}{n_i} p_i^{n_i} (1 - p_i)^{n - n_i}
\end{aligned}$$

Observación: Como se dijo, la interpretación es como si hubieran dos clases (categorías): $x_1 = y_i$, y $x_2 = \bigcup_{j \neq i} y_j$. Entonces $(M_1, M_2) = (N_i, \sum_{j \neq i} N_j)$ es un vector de frecuencias para las categorías x_1 y x_2 , y la distribución del vector es Multinomial de parámetros n y $q_1 = p_i$, $q_2 = \sum_{j \neq i} p_j$. La distribución marginal de M_2 es una binomial de parámetros n y q_2 .

4. ESPERANZA, VARIANZA, COVARIANZA

Como se dijo anteriormente, la distribución de N_i es Binom(n, p_i). Por lo tanto, se tiene que

$$\begin{aligned}
\mathbb{E}(N_i) &= np_i \\
\text{Var}(N_i) &= np_i(1 - p_i)
\end{aligned}$$

Como ya se vio, $N_i + N_j \sim \text{Binom}(n, p_i + p_j)$, por lo que

$$\begin{aligned}
\text{Var}(N_i + N_j) &= \text{Var}(N_i) + \text{Var}(N_j) + 2\text{Cov}(N_i, N_j) \\
n(p_i + p_j)(1 - p_i - p_j) &= np_i(1 - p_i) + np_j(1 - p_j) + 2\text{Cov}(N_i, N_j) \\
\text{Cov}(N_i, N_j) &= -np_i p_j
\end{aligned}$$

5. EJEMPLO SENCILLO

Supongase que el 23% de las personas que asisten a cierto partido de "baseball" viven a menos de 10 millas del estadio, el 59% de ellas viven a entre 10 y 50 millas del estadio, y el 18% vive a mas de 50 millas. Se seleccionan

al azar 20 personas entre los asistentes al partido (que son miles). Calcular la probabilidad de que siete de los seleccionados vivan a menos de 10 millas, ocho vivan entre 10 y 50 millas, y cinco vivan a más de 50 millas del estadio.

Solución Comenzamos por identificar todos los elementos del problema:

$n = 20$ (número de personas seleccionadas), $k = 3$ (cantidad de grupos de clasificación de las personas); $y_1 = \{\text{Personas que viven a menos de 10 millas del estadio}\}$, $y_2 = \{\text{Personas que viven a entre 10 y 50 millas del estadio}\}$, $y_3 = \{\text{Personas que viven a más de 50 millas del estadio}\}$;

$p_1 = 0.23, p_2 = 0.59, p_3 = 0.18$

Definiendo (N_1, N_2, N_3) el vector correspondiente a las frecuencias, se pide calcular

$$\begin{aligned} \mathbb{P}(N_1 = 7, N_2 = 8, N_3 = 5) &= \binom{20}{7, 8, 5} 0.23^7 0.59^8 0.18^5 \\ &= \frac{20!}{7!8!5!} 0.23^7 0.59^8 0.18^5 \\ &= 0.0094 \end{aligned}$$

6. PROBLEMAS

Problema 1: (a) Suponga que las variables aleatorias $(X_i)_{i=1}^k$ son independientes y que $X_i \sim \text{Poisson}(\lambda_i) \forall i \in \{1, \dots, k\}$. Demuestre que para todo $n \in \mathbb{N}$ la distribución del vector aleatorio $(X_i)_{i=1}^k$ condicional a que $\sum_{i=1}^k X_i = n$, es una multinomial de parámetros n y $\vec{p} = (p_i)_{i=1}^k$, con

$$p_i = \frac{\lambda_i}{\sum_{j=1}^k \lambda_j}$$

(b) A una heladería llegan clientes de tres tipos distintos: *normal*, *golozo* y *premium*, y el número de clientes de cada tipo (que llega en una hora) son Poisson con tasas $\lambda_n = 100$, $\lambda_g = 50$ y $\lambda_p = 20$ respect. Si se sabe que el número total de clientes que llegó en una hora es de 500 personas, calcular la probabilidad de que hayan llegado al menos 200 clientes premium.

Problema 2:

Se lanzan cinco dados equilibrados. ¿Cuál es la probabilidad de que el número 1 y el número 4 aparezcan el mismo número de veces?

Problema 3: (a) Supongase que el 16% de los estudiantes de un colegio son alumnos de primero medio, el 14% de segundo, el 38% de tercero, y el 32% de cuarto. Si se seleccionan al azar 15 estudiantes, cual es la probabilidad de que al menos 8 estudiantes sean de primero o segundo?

(b) Sea X_3 el número de estudiantes de tercero y X_4 el número de estudiantes de cuarto. Calcule el número esperado de alumnos de cada curso (en la muestra), y la esperanza y varianza de $X_3 - X_4$

7. RESOLUCIÓN DE PROBLEMAS

Solución (Problema 1): (a) Sea $n \in \mathbb{N}$. Definamos $\lambda = \sum_{j=1}^k \lambda_j$, $\vec{n} = (n_1, \dots, n_k)$. Una primera observación es que $Y = \sum_{j=1}^k X_j$ es una variable aleatoria que se distribuye *Poisson*(λ), pues las X_j son Poissons independientes (*).

$$\begin{aligned}
& \mathbb{P} \left(X_1 = n_1, \dots, X_k = n_k \mid \sum_{j=1}^k X_j = n \right) \\
&= \frac{\mathbb{P} \left(X_1 = n_1, \dots, X_k = n_k, \sum_{j=1}^k X_j = n \right)}{\mathbb{P} \left(\sum_{j=1}^k X_j = n \right)} \\
&= \frac{\mathbb{P}(X_1 = n_1, \dots, X_k = n_k) \mathbf{1}_{\{\sum_{j=1}^k n_j = n\}}(n_1, \dots, n_k)}{\mathbb{P} \left(\sum_{j=1}^k X_j = n \right)} \\
&= \frac{\mathbb{P}(X_1 = n_1) \dots \mathbb{P}(X_k = n_k)}{\mathbb{P}(Y = n)} \underbrace{\mathbf{1}_{\{\sum_{j=1}^k n_j = n\}}(\vec{n})}_{\alpha(\vec{n})} \quad (*) \\
&= \frac{e^{-\lambda_1} \lambda_1^{n_1} / n_1! \dots e^{-\lambda_k} \lambda_k^{n_k} / n_k!}{e^{-\lambda} \lambda^n / n!} \alpha(\vec{n}) \\
&= \frac{e^{-\sum_{j=1}^k \lambda_j} \lambda_1^{n_1} \dots \lambda_k^{n_k}}{e^{-\lambda} \lambda^n} \frac{n!}{n_1! \dots n_k!} \alpha(\vec{n}) \\
&= \frac{n!}{n_1! \dots n_k!} \left(\frac{\lambda_1}{\lambda} \right)^{n_1} \dots \left(\frac{\lambda_k}{\lambda} \right)^{n_k} \alpha(\vec{n}) \\
&= \binom{n}{n_1, \dots, n_k} p_1^{n_1} \dots p_k^{n_k} \mathbf{1}_{\{\sum_{j=1}^k n_j = n\}}(n_1, \dots, n_k)
\end{aligned}$$

(b) Si suponemos que las llegadas de los distintos tipos de clientes son independientes, tendríamos aplicando la parte anterior $n = 500$, $p_n = \frac{100}{170}$, $p_p = \frac{20}{170}$, $p_g = \frac{50}{170}$. Definiendo (X_n, X_p, X_g) el vector de llegadas, entonces este vector, dado que su suma es 500, se distribuye *Multin*(500, (p_n, p_p, p_g)). Luego $X_p | X_n + X_p + X_g = 500 \sim \text{Binom}(n, p_p)$

$$\mathbb{P}(X_p \geq 200 | X_n + X_p + X_g = 500) = \sum_{j=200}^{500} \binom{500}{j} \left(\frac{20}{170} \right)^j \left(\frac{100}{170} + \frac{50}{170} \right)^{500-j}$$

Solución (Problema 2):

En este caso $n = 5$, $k = 6$; $y_i = \{i\}$, $p_i = \frac{1}{6} \forall i \in \{1, \dots, 6\}$. Lo que se pide es $\mathbb{P}(N_1 = N_4)$

$$\begin{aligned}
\mathbb{P}(N_1 = N_4) &= \sum_{j=0}^{\infty} \mathbb{P}(N_1 = j, N_4 = j) \\
&= \sum_{j=0}^2 \mathbb{P}(N_1 = j, N_4 = j)
\end{aligned}$$

Para resolver lo anterior habría que condicionar sobre los posibles valores de las otras frecuencias, y quedaría algo “monstruoso”. La forma “inteligente” de resolver el problema es “colapsar” ciertas clases. definamos entonces las clases $x_1 = y_1$, $x_4 = y_4$, $x_{resto} = y_2 \cup y_3 \cup y_5 \cup y_6$, donde se tiene que $\bar{p}_1 = 1/6 = \bar{p}_4$, $p_{resto} = 4/6$ (\bar{p}_j es la probabilidad asociada a x_j); y sea (M_1, M_4, M_{resto}) el vector de frecuencias asociado a (x_1, x_4, x_{resto}) . Así, y con la observación

de que este vector también tiene una distribución $Multinomial(5, (1/6, 1/6, 4/6))$, calculemos $\mathbb{P}(N_1 = j, N_4 = j)$.

$$\begin{aligned}\mathbb{P}(N_1 = j, N_4 = j) &= \mathbb{P}(M_1 = j, M_4 = j, M_{resto} = 5 - 2j) \\ &= \binom{5}{j, j, 5 - 2j} \left(\frac{1}{6}\right)^j \left(\frac{1}{6}\right)^j \left(\frac{4}{6}\right)^{5-2j}\end{aligned}$$

Así,

$$\begin{aligned}\mathbb{P}(N_1 = N_4) &= \sum_{j=0}^2 \binom{5}{j, j, 5 - 2j} \left(\frac{1}{6}\right)^j \left(\frac{1}{6}\right)^j \left(\frac{4}{6}\right)^{5-2j} \\ &= \frac{5!}{0!0!5!} \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^0 \left(\frac{4}{6}\right)^5 + \frac{5!}{1!1!3!} \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 \left(\frac{4}{6}\right)^3 + \frac{5!}{2!2!1!} \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^2 \left(\frac{4}{6}\right)^1 \\ &= \dots \\ &= \frac{2424}{6^5}\end{aligned}$$

Solución (Problema 3): (a) Definamos las clases $x_1 = \{\text{Estudiantes de primero o segundo}\}$, $x_2 = \{\text{Estudiantes de tercero o cuarto}\}$, con probabilidades $q_1 = 0.3$, $q_2 = 0.7$. Sea (M_1, M_2) el vector de frecuencias de dichos grupos, con $n = 15$. Entonces $N_1 \sim Binomial(15, 0.3)$.

$$\begin{aligned}\mathbb{P}(N_1 \geq 8) &= \sum_{j=8}^{15} \mathbb{P}(N_1 = j) \\ &= \sum_{j=8}^{15} \binom{15}{j} (0.3)^j (0.7)^{15-j} \\ &= 0.05\end{aligned}$$

(b) Recordemos que $N_3 \sim Binom(15, 0.38)$, $N_4 \sim Binom(15, 0.32)$. Entonces

$$\begin{aligned}\mathbb{E}(N_3 - N_4) &= \mathbb{E}(N_3) - \mathbb{E}(N_4) \\ &= 15 \times 0.38 - 15 \times 0.32 \\ &= 15 \times 0.06 \\ &= 0.9\end{aligned}$$

La varianza la calculamos

$$\begin{aligned}\text{Var}(N_3 - N_4) &= \text{Var}(N_3) + \text{Var}(N_4) - 2\text{Cov}(N_3, N_4) \\ &= 15 \times 0.38 \times 0.62 + 15 \times 0.32 \times 0.68 - 2 \times 15 \times 0.38 \times 0.32 \\ &= 3.534 + 3.264 - 3.648 \\ &= 3.15\end{aligned}$$