

Distribuzione Multinomiale e Distribuzione Ipergeometrica¹

Dobbiamo capire la differenza fra questi due esempi:

Esempio 1. 12 bambini vanno a una festa; 3 vengono dal paese A, 4 da B e 5 da C. Per 5 volte viene estratto un bambino fra i 12 e gli si dà un premio. Ogni volta, le probabilità di estrazione sono $3/12$ per A, $4/12$ per B e $5/12$ per C. Qual è la probabilità che i vincitori siano 1 di A, 2 di B e 2 di C.

Esempio 2. Una commissione di 12 persone è fatta da 3 rappresentanti del partito A, 4 del partito B e 5 del partito C. Si estrae a sorte una sottocommissione di 5 persone. Qual è la probabilità che i commissari scelti siano 1 di A, 2 di B e 2 di C.

In entrambi i casi possiamo pensare che ci sia un'urna con 12 palline numerate da 1 a 12; sulle prime 3 c'è scritto A, in quelle da 4 a 7 B e nelle restanti 5 C. Nel primo esempio si estrae una pallina (probabilità $3/12, 4/12, 5/12$ delle tre provenienze), *si rimette dentro* e si ripete l'estrazione per 5 volte. Nel secondo esempio si prendono 5 palline una dopo l'altra, *non rimettendo dentro* quelle via via estratte. In entrambi i casi siamo interessati alla composizione della cinquina estratta, ma i due casi sono chiaramente diversi. Osserviamo d'altra parte che se invece di 12 ci fossero 12 milioni di palline, estrarne 5 non cambierebbe sostanzialmente la composizione dell'urna quindi le probabilità cercate nei due casi dovrebbero essere simili. Scriveremo adesso un modello che ci permette di analizzare il problema.

Ci sono N palline numerate da 1 ad N . Ognuna è di un tipo $i \in T$, dove T è l'insieme dei tipi, finito (per esempio $T = \{\text{paese A, paese B, paese C}\}$, oppure $T = \{\text{bianco, rosso, giallo}\}$, $T = \{1, \dots, m\}$, eccetera). Ce ne sono N_i del tipo i , quindi per costruzione $\sum_{i \in T} N_i = N$. L'insieme delle palline di tipo i lo indichiamo con $T(i) \subseteq \{1, \dots, N\}$. Nell'esempio dei colori può essere che le palline rosse sono quelle numero 3, 28 e 144, nel qual caso $T(\text{rosso}) = \{3, 28, 144\}$ ed $N_{\text{rosso}} = 3$.

Con o senza reimmissione si estraggono $n \leq N$ palline. Siamo interessati alla composizione di questo “campione” di n elementi, precisamente a quanti elementi di ogni tipo contiene. Formalmente l'oggetto di interesse è il vettore $\vec{k} \equiv (k_i)_{i \in T}$, dove i k_i sono interi non negativi e $\sum_{i \in T} k_i = n$ (in entrambi gli esempi iniziali $\vec{k} = (1, 2, 2)$). Quindi in entrambi i casi lo spazio sul quale vogliamo assegnare una probabilità è

$$\hat{\Omega} = \left\{ \vec{k} = (k_i)_{i \in T} : \sum_{i \in T} k_i = n \right\}.$$

Caso 1: estrazioni con reimbussolamento (che sono prove ripetute indipendenti)

In questo caso il risultato ω dell'esperimento è una sequenza di n tipi, in altre parole un elemento dell'insieme $\Omega = T^n$, cioè $\Omega \ni (\omega_1, \dots, \omega_n)$ con $\omega_i \in T$, $i = 1, \dots, n$. Osserva che se

¹S Modica 2024 (in Italian, sorry about that). Ref. Georgii, §2.2-2.3.

$T = \{0, 1\}$ il modello è quello di n lanci di una moneta con probabilità di successo $p = N_1/N$ in ogni lancio (Baldi, esempio 1.20). In quel caso $\omega_i \in \{0, 1\}$ per ogni i , e in considerazione dell'indipendenza delle prove abbiamo dedotto che per ogni $\omega = (\omega_1, \dots, \omega_n)$ in cui gli 1 sono k e gli zeri $n - k$ debba essere $P(\{\omega\}) = p^k q^{n-k}$; e poiché di siffatti ω ce ne sono $\binom{n}{k}$ abbiamo concluso che $P(\{\omega: \sum_{i=1}^n \omega_i = k\}) = \binom{n}{k} p^k q^{n-k}$. Nel caso generale procediamo allo stesso modo. Ponendo $\rho(i) = N_i/N$, arriviamo alla conclusione che per ogni ω che contiene k_i elementi del tipo i per $i \in T$ deve essere $P(\{\omega\}) = \prod_{i \in T} \rho(i)^{k_i}$. Questi sono gli ω che risultano nel vettore $\vec{k}$ di interesse; e di questi come abbiamo visto (Baldi, Proposizione 1.31) ce n'è $\binom{n}{\vec{k}} \equiv n! / \prod_i (k_i!)$. Dunque la probabilità che dobbiamo assegnare al vettore $\vec{k}$ è $\binom{n}{\vec{k}} \prod_{i \in T} \rho(i)^{k_i}$. Questa si indica con $\mathcal{M}_{n,\rho}(\vec{k})$, e si chiama *distribuzione multinomiale* (il sottoscritto ricorda che dipende da n e dal vettore $(\rho(i))_{i \in T}$):

$$\mathcal{M}_{n,\rho}(\vec{k}) = \binom{n}{\vec{k}} \prod_{i \in T} \rho(i)^{k_i}, \quad \vec{k} \in \hat{\Omega}.$$

Osserva che nel caso binario $T = \{0, 1\}$ il vettore $\vec{k}$ è del tipo $(k, n - k)$ dove k è il numero di successi, e ponendo $p = \rho(1) = N_1/N$ la $\mathcal{M}$ si riduce alla *distribuzione binomiale* che conosciamo:

$$\mathcal{B}_{n,p}(k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, \dots, n.$$

Caso 2: estrazioni senza reimbussolamento

In questo caso il risultato ω dell'esperimento è un sottoinsieme $\omega \subseteq \{1, 2, \dots, N\}$ di n elementi (le n palline estratte). Sicché $\Omega = \{\omega \subseteq \{1, 2, \dots, N\}: |\omega| = n\}$.² La sua cardinalità è $\binom{N}{n}$, ed è naturale assumere che tutte le possibili estrazioni hanno la stessa probabilità. Gli ω che risultano nel vettore $\vec{k}$ sono quelli tali che $|\omega \cap T(i)| = k_i, i \in T$. Quanti sono? Un siffatto ω si forma prendendo, per ogni i , k_i elementi dagli N_i di tipo i ; quindi sono in numero $\prod_{i \in T} \binom{N_i}{k_i}$.³ Dunque la probabilità che dobbiamo assegnare a $\vec{k}$ è in questo caso $\prod_{i \in T} \binom{N_i}{k_i} / \binom{N}{n}$; si indica con $\mathcal{H}_{n,\vec{N}}(\vec{k})$, dove $\vec{N} = (N_i)_{i \in T}$ è il vettore degli N_i , e si chiama *distribuzione ipergeometrica*:

$$\mathcal{H}_{n,\vec{N}}(\vec{k}) = \frac{\prod_{i \in T} \binom{N_i}{k_i}}{\binom{N}{n}}, \quad \vec{k} \in \hat{\Omega}.$$

Di nuovo, nel caso $T = \{0, 1\}$ il vettore $\vec{k}$ è del tipo $(k, n - k)$ quindi la $\mathcal{H}$ si riduce a

$$\mathcal{H}_{n,(N_0,N_1)}(k) = \frac{\binom{N_1}{k} \binom{N_0}{n-k}}{\binom{N}{n}}, \quad k = 0, \dots, n.$$

² $|A|$ indica la cardinalità dell'insieme finito A .

³In questo caso deve essere $k_i \leq N_i$, in caso contrario poniamo $\binom{N_i}{k_i} = 0$.

Probabilità di $\vec{k} = (1, 2, 2)$ nei due esempi di sopra

Nel primo caso la risposta è

$$\mathcal{M}_{5,(\frac{3}{12}, \frac{4}{12}, \frac{5}{12})}((1, 2, 2)) = \frac{5!}{1!2!2!} \cdot \frac{3}{12} \cdot \left(\frac{4}{12}\right)^2 \cdot \left(\frac{5}{12}\right)^2 = \frac{125}{864} \approx 0.145.$$

Nel secondo caso è

$$\mathcal{H}_{5,(3,4,5)}((1, 2, 2)) = \frac{\binom{3}{1}\binom{4}{2}\binom{5}{2}}{\binom{12}{5}} = \frac{5}{22} \approx 0.227.$$

Grandi popolazioni

In grandi popolazioni, come accennato, l'intuizione suggerisce che non dovrebbe fare differenza se le palline vengono rimesse nell'urna o meno. Per esempio se negli esempi di sopra prendiamo $N = 120$ anziché 12 mantenendo le stesse frazioni di tipi otteniamo multinomiale invariata, ma $\mathcal{H}_{5,(30,40,50)}((1, 2, 2)) = 0.150$ - molto più vicina alla multinomiale.

Per enunciare il risultato preciso usiamo la notazione di equivalenza asintotica di due funzioni: “ $a(N) \sim b(N)$ per $N \rightarrow \infty$ ” vuol dire $\lim_{N \rightarrow \infty} a(N)/b(N) = 1$. Vale il seguente

Teorema. Per $N \rightarrow \infty, N_\theta \rightarrow \infty$ ed $N_\theta/N \rightarrow \rho(\theta), \theta \in T$ si ha $\mathcal{H}_{n,\vec{N}}(\vec{k}) \sim \mathcal{M}_{n,\rho}(\vec{k}), \forall \vec{k} \in \hat{\Omega}$.

Dimostrazione. Ricorda che $\binom{n}{k} = n(n-1) \cdots (n-k+1)/k!$ (al numeratore k termini). Moltiplicando e dividendo per n^k otteniamo

$$\begin{aligned} \binom{n}{k} &= \frac{n^k n(n-1) \cdots (n-k+1)}{k! n^k} \\ &= \frac{n^k}{k!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \sim \frac{n^k}{k!} \quad \text{per } n \rightarrow \infty. \end{aligned}$$

Dunque per $N, N_\theta \rightarrow \infty$ ed $N_\theta/N \rightarrow \rho(\theta), \theta \in T$ abbiamo (usando $\sum_\theta k_\theta = n$)

$$\mathcal{H}_{n,\vec{N}}(\vec{k}) \sim \left[\prod_{\theta \in T} \frac{N_\theta^{k_\theta}}{k_\theta!} \right] \Big/ \frac{N^n}{n!} = \binom{n}{\vec{k}} \prod_{\theta \in T} \left(\frac{N_\theta}{N} \right)^{k_\theta} \rightarrow \mathcal{M}_{n,\rho}(\vec{k})$$

che è quanto volevamo. □