

Capitolo 6. Distribuzione binomiale. Deviazione standard di una proporzione. Distribuzione di Poisson. Assortimenti.

Distribuzione binomiale

La distribuzione binomiale consente di valutare la probabilità che una modalità di un evento con probabilità individuale a priori (p) si verifichi un determinato numero di volte (i) entro un numero totale (n) di eventi. Con esempi la faccenda può essere più chiara. Ad esempio, la distribuzione binomiale valuta la probabilità che:

- su 10 figli 7 siano maschi ($n=10$, $i=7$, modalità='maschio' con $p=0.5$)
- su 8 lanci di dado il due esca tre volte ($n=8$, $i=3$, modalità 'due' con $p=1/6$)
- sui 13 risultati della schedina compaiano 10 x ($n=13$, $i=10$, modalità 'x' con $p=1/3$)

La formula è la seguente:

$$P_{n,i,p} = \frac{n!}{i!(n-i)!} \cdot p^i \cdot q^{n-i}$$

Chiamiamo p (minuscolo) la probabilità a priori della modalità in questione e q la probabilità - complementare a p - di tutte le altre modalità:

modalità

modalità complementare

maschio, $p=0.5$

non-maschio (femmina), $q=0.5$

due ai dadi, $p=1/6$

non-due ai dadi, $q=5/6$

x della schedina, $p=1/3$

non-x della schedina $q=2/3$

Ovviamente, $p + q = 1$.

Attenzione: le modalità a cui si riferiscono p e q devono essere mutualmente esclusive, esaustive e sequenzialmente indipendenti!

La formula della distribuzione binomiale altro non è che il termine $(n+1-i)^{mo}$ dello sviluppo di $(p+q)^n$ di cui

$\frac{n!}{i!(n-i)!}$ rappresenta il cosiddetto **coefficiente**, mentre

$p^i \cdot q^{n-i}$ rappresenta la cosiddetta **parte letterale** del termine.

Ad esempio, la probabilità che su tre lanci di dado ($n=3$) il due ($p=1/6$) esca 1 volta ($i=1$) viene calcolata come:

$$P_{n,i,p} = \frac{3!}{1!(3-1)!} \cdot (1/6)^1 \cdot (5/6)^{3-1}$$

che corrisponde esattamente all' $(n+1-i)^{m_0} = (3+1-1)^{m_0} =$ terzo termine dello sviluppo di $(1/6 + 5/6)^3$:

$$(1/6)^3 + 3(1/6)^2 \cdot (5/6) + \mathbf{3(1/6) \cdot (5/6)^2} + (5/6)^3$$

Tutti conosciamo dalle scuole medie lo sviluppo del quadrato e del cubo di un binomio (a proposito, la distribuzione binomiale si chiama così appunto per tale corrispondenza). Ma nessuno potrebbe conoscere a mente lo sviluppo delle potenze di un binomio oltre un certo grado. In realtà, il famoso triangolo di Tartaglia consente di andare un pò più in là, in quanto ci dà la serie di coefficienti a cui applicare la parte letterale con esponenti decrescenti (da n a 0) di p e crescenti (da 0 a n) di q .

Triangolo di Tartaglia (fino al grado 7)

						1											$(p+q)^0$
						1		1									$(p+q)^1$
					1		2		1								$(p+q)^2$
			1		3		3		1								$(p+q)^3$
		1		4		6		4		1							$(p+q)^4$
	1		5		10		10		5		1						$(p+q)^5$
	1		6		15		20		15		6		1				$(p+q)^6$
1		7		21		35		35		21		7		1			$(p+q)^7$

Es. $(p+q)^7 = \mathbf{1}p^7 + \mathbf{7}p^6q + \mathbf{21}p^5q^2 + \mathbf{35}p^4q^3 + \mathbf{35}p^3q^4 + \mathbf{21}p^2q^5 + \mathbf{7}pq^6 + \mathbf{1}q^7$

Comunque, anche con il triangolo di Tartaglia è molto scomodo calcolare potenze di grado superiore come $(p+q)^{500}$: occorrerebbe sviluppare centinaia di righe con centinaia di coefficienti! Pertanto è bene usare la formula della distribuzione con i fattoriali. Attenzione: i fattoriali sono capaci di produrre numeri enormi, non rappresentabili in molti calcolatori a 8-10 cifre o rappresentabili in forma esponenziale con solo 8-10 cifre significative e conseguente perdita di informazione. Per questo spesso si ricorre alla trasformazione logaritmica, sapendo che il logaritmo di un prodotto corrisponde alla somma dei logaritmi dei fattori, e quindi il logaritmo di $k!$ corrisponde alla sommatoria dei logaritmi primi k interi: $\log(1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot \dots \cdot k) = \log 1 + \log 2 + \log 3 + \log 4 + \log 5 + \dots + \log k$, e rimettendo poi le cose a posto elevando la base del logaritmo adottato (decimale o naturale, è indifferente) per il risultato.

Ci si chiederà a che serve tutta questa dissertazione sulla distribuzione binomiale. A noi serve per conoscere e rappresentare il modello di variabilità con cui valutare, ad esempio, le frequenze attese da confrontare con quelle osservate nel nostro campione. Se qualcuno ci chiedesse quale è la probabilità di estrarre dal sacchetto della tombola (con dentro tutti i numeri da 1 a 90) un numero compreso tra 21 e

30, potremmo rispondere facilmente: $10/90 = 0.111...$ Infatti l'intervallo tra 21 e 30 (inclusi) comprende dieci numeri, su un totale di 90. Questo lo possiamo dire solo perché conosciamo esattamente quali sono i numeri della tombola. Qualsiasi distribuzione, conosciuta in dettaglio, ci consente di rispondere ad ogni quesito concernente la probabilità dei suoi eventi.

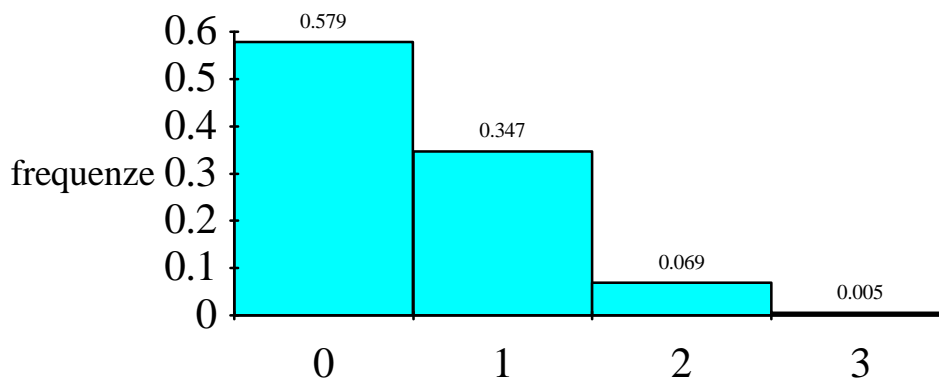
Il seguente è uno schema che riassume tutti gli aspetti del calcolo della distribuzione binomiale.

Probabilità che il numero due del dado con

- probabilità di evento a priori pari a $1/6$ ($p=1/6$, $q=5/6$)
- esca 0, 1, 2, 3 volte ($i=0, 1, 2, 3$)
- su 3 lanci ($n=3$)

			espressione generalizzata		espressione semplificata	valori
possibilità			coefficiente dato dal triangolo di Tartaglia o dalla formula $\frac{n!}{i!(n-i)!}$	parte letterale p con esponenti decrescenti q con esponenti crescenti $p^i \cdot q^{n-i}$		
i = 0	0 volte su 3	no no no	$\frac{3!}{0!(3-0)!}$	$(1/6)^0 \cdot (5/6)^{3-0}$	q^3	0.579 (57.9%)
i = 1	1 volta su 3	sì no no no no sì no sì no	$\frac{3!}{1!(3-1)!}$	$(1/6)^1 \cdot (5/6)^{3-1}$	$3pq^2$	0.347 (34.7%)
i = 2	2 volte su 3	sì sì no sì no sì no sì sì	$\frac{3!}{2!(3-2)!}$	$(1/6)^2 \cdot (5/6)^{3-2}$	$3p^2q$	0.069 (6.9%)
i = 3	3 volte su 3	sì sì sì	$\frac{3!}{3!(3-3)!}$	$(1/6)^3 \cdot (5/6)^{3-3}$	p^3	0.005 (0.5%)
					totale $(p+q)^3$	totale 1 (100%)

L'istogramma della distribuzione così calcolata è:



Infine, una precisazione importante: la probabilità è sempre espressa in ragione di 1 (un valore compreso tra 0 ed 1). Ma se consideriamo una certa frequenza osservata (es., in 7 lanci, il due è uscito 6 volte; frequenza osservata = 6) e vogliamo confrontare questa con la frequenza attesa dalla distribuzione binomiale, questa non è la semplice probabilità $7 \cdot (1/6)^6 \cdot (5/6)^1 = 0.000125$, bensì la probabilità moltiplicata per il numero di lanci, cioè, nel nostro caso $0.000125 \cdot 7 = 0.000875$.

Deviazione standard di una proporzione

La proporzione rappresenta nel campo delle frequenze ciò che è la media nel campo delle variabili ponderali. Come abbiamo visto nel paragrafo precedente, la distribuzione binomiale non è necessariamente simmetrica, anzi, quando **p** è piccolo si presenta fortemente asimmetrica. Al crescere di **n** ed al tendere di **p** a 0.5 la distribuzione binomiale tende a diventare simmetrica. In pratica, quando $n \cdot p > 5$ la distribuzione può essere considerata sufficientemente simmetrica da poter valutare una deviazione standard da associare alla proporzione. Ricordiamo che la deviazione standard di un parametro statistico, sinonimo di errore standard, è l'espressione della variabilità o affidabilità della stima del parametro. La deviazione standard di una proporzione è data dalla formula:

$$s_p = \sqrt{pq / n}$$

Se, ad esempio, la proporzione riguarda 15 osservazioni su un totale di 70, diremo che

$$n = 70$$

$$r = 15$$

$$p = 15 / 70 = 0.21$$

$$n - r = 55$$

$$q = 1 - p = 0.79$$

$$s_p = \sqrt{0.21 \cdot 0.79 / 70} = 0.049$$

0.049 è la deviazione standard associata alle frequenze relative, sia **p** (0.21) che **q** (0.79).

Attenzione: al posto delle frequenze relative di **p** e **q** possiamo anche impiegare nella formula le loro frequenze percentuali, lasciando tuttavia **n** in frequenza assoluta

$$s_p = \sqrt{21 \cdot 79 / 70} = 4.9$$

Allo stesso modo possiamo calcolare la deviazione standard da associare alle frequenze assolute **r** (15) e **n-r** (55) calcoleremo

$$s_r = \sqrt{15 \cdot 55 / 70} = 3.4$$

Ovviamente 3.4 sta a 15 come 0.049 sta a 0.21 (a parte differenze dovute all'arrotondamento).

Quindi, esprimendo una proporzione, ed assumendo che la distribuzione binomiale definita dai valori **p** e **n** sia simmetrica, noi possiamo calcolare la deviazione standard da associare alla

- frequenza relativa
- frequenza percentuale
- frequenza assoluta

In analogia con quanto fatto per i limiti fiduciali della media, la deviazione standard della proporzione consente la stima dei limiti fiduciali della proporzione.

$$LF = p \pm t \cdot s_p$$

Considerando un livello di probabilità $\alpha=0.05$ per cui $t=1.96$, i limiti fiduciali di 0.21 sono:

$$LF = 0.21 \pm 1.96 \cdot 0.049$$

per cui riteniamo che la vera proporzione della popolazione sia compresa, con probabilità del 95%, tra 0.10 e 0.32.

Quando la distribuzione non sia approssimabile a quella normale (**np**≤5) bisogna ricorrere ad un calcolo più complesso o a delle tabelle che riportino i limiti fiduciali asimmetrici di ogni proporzione per diversi valori di **n**.

Sempre dalla stessa deviazione standard della proporzione possiamo ricavare il numero di osservazioni sufficiente ad ottenere una proporzione di una certa rappresentatività. Il problema ricalca anche in questo caso quello della grandezza del campione affrontato per la prima volta nel 2° capitolo riguardante le medie e la loro variabilità.

Supponiamo che i nostri dati con **n**=70 derivino da uno studio preliminare. Vogliamo ora sapere quanto grandi debbano essere i futuri campioni in modo tale che i limiti fiduciali delle loro proporzioni non si discostino dalla proporzione vera

della popolazione più - diciamo - del 5% del loro valore (questo è il grado di rappresentatività voluto). Scriviamo quindi il sistema:

$$\begin{aligned} LF &= p \pm t \cdot s_p \\ LF &= p \pm 0.05 p \end{aligned}$$

da cui ricaviamo:

$$t \cdot s_p = 0.05 p$$

sostituendo s_p con $\sqrt{p \cdot q / n}$ otteniamo:

$$t \cdot \sqrt{p \cdot q / n} = 0.05 p$$

da cui infine ricaviamo n:

$$n = (t^2 \cdot q) / (0.05^2 \cdot p)$$

Nel nostro caso

$$n = (1.96^2 \cdot 0.79) / (0.05^2 \cdot 0.21) = 5781$$

Per soddisfare il nostro desiderio dovremmo quindi esaminare un campione piuttosto grande.

Per semplicità non si è inclusa la correzione per la continuità, che comunque si dovrebbe usare per ottenere stime più accurate.

Infine, in questi calcoli abbiamo usato i valori di t con GDL = ∞ per i quali t è distribuito normalmente. Questo in quanto abbiamo assunto che la proporzione sia distribuita normalmente.

Distribuzione di Poisson

Quando l'oggetto non presenta modalità o tipi - o non si è interessati ai tipi - si può fare solo un conteggio degli oggetti. Ad esempio possiamo essere interessati al numero di auto che percorrono una certa strada senza considerare le varie marche di auto, o al numero di impulsi elettrici prodotti da un neurone senza considerare le caratteristiche di tali impulsi, o al numero di eruzioni di un vulcano, o al numero di cellule presenti in un certo tessuto, ecc. In tali casi dobbiamo definire esattamente il contenitore di tali fenomeni, che può essere un ambito fisico di tempo e/o di spazio o anche un ambito logico entro cui rientra il fenomeno. In genere si parla di

- **frequenza** quando ci si riferisce a numero di...per intervallo di **tempo** e di
- **densità** quando ci si riferisce a numero di...per intervallo di **spazio**

Per i nostri scopi, frequenza e densità sono equivalenti. Per semplicità, parleremo di frequenza per tutti i casi. E' chiaro che la frequenza di un determinato fenomeno non è mai perfettamente costante nel tempo o nello spazio, ma può variare. Noi ora ci occupiamo, o meglio ci preoccupiamo di valutare tale variabilità secondo il modello della distribuzione di Poisson. Secondo tale distribuzione, la probabilità di

trovare **i** oggetti in un determinato intervallo di spazio o tempo nel quale, in condizioni di omogeneità, dovrebbero trovarsene **m** (numero medio atteso) è data da:

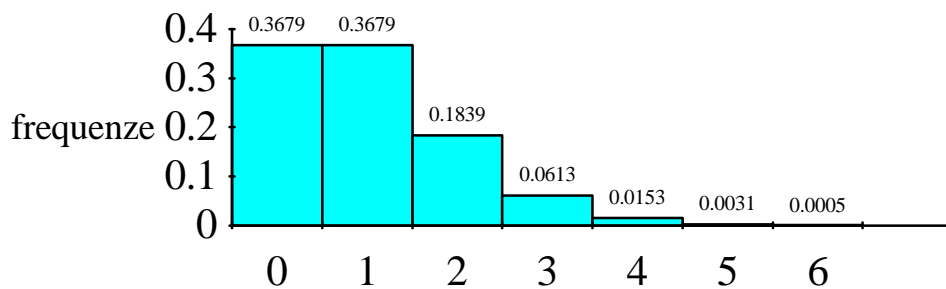
$$P_{i,m} = \frac{m^i}{i!} e^{-m}$$

ricordando che **e** è la base dei logaritmi naturali ed è circa **2.7182**.

Ovviamente, il numero medio atteso **m** è dato dal numero totale diviso il numero di frazioni di tempo o di spazio. Se ad esempio suddividiamo un litro di terreno di coltura, in cui si trovano 1000 cellule, in 1000 frazioni di 1 ml, in condizioni di perfetta omogeneità ci aspetteremo di trovare **m=1** cellula in ciascuna frazione. Ma così non avverrà sempre: ci saranno frazioni con 0, 1, 2, 3, 4, ecc. cellule le cui frequenze rispecchieranno la distribuzione di Poisson.

Applicando la formula all'esempio per **i** che va da 0 a 6 otteniamo le seguenti probabilità:

i	P
0	0.3679
1	0.3679
2	0.1839
3	0.0613
4	0.0153
5	0.0031
6	0.0005
...	...
	totale 1



Come già detto in altre circostanze, la definizione della distribuzione teorica è essenziale per fare previsioni e/o per valutare se esistano fattori che alterino la distribuzione naturale dei fenomeni. Se ad esempio ci fosse una tendenza delle cellule ad aggregare in gruppi di 4 o 5, potremmo trovare un aumento delle frazioni con 0, 4, e 5 cellule, parallelamente ad una diminuzione delle frequenze delle frazioni con 1, 2, e 3 cellule. L'ipotesi che la distribuzione osservata differisca

da quella attesa può essere valutata mediante test χ^2 . In quest'ultimo caso il test equivale a verificare se le cellule tendono significativamente ad aggregare.

Assortimenti

Il calcolo delle frequenze attese di certi eventi costituiti dall'assortimento di r elementi estratti da un insieme definito di n elementi in tutto, richiede di conoscere il numero totale dei possibili diversi assortimenti. Questo dipenderà dal fatto che:

1. l'ordine degli elementi faccia/non faccia differenza
2. gli elementi estratti possano/non possano essere ripetuti

Esaminiamo 4 diversi problemi:

quante sono le possibili diverse giocate (colonne) del totocalcio ?

=> *l'ordine fa differenza, ci possono essere ripetizioni*

quante diversi acquisti di 4 pezzi si possono fare da un campionario di 10 articoli ?

=> *l'ordine non fa differenza, ci possono essere ripetizioni*

quanti diversi ordini di arrivo di 5 atleti si possono avere da una gara con 20 atleti ?

=> *l'ordine fa differenza, non ci possono essere ripetizioni*

quante diverse mani di 5 carte si possono avere giocando a poker ?

=> *l'ordine non fa differenza, non ci possono essere ripetizioni*

	<i>l'ordine fa differenza</i> DISPOSIZIONI	<i>l'ordine non fa differenza</i> COMBINAZIONI
<i>elementi ripetuti</i>	$n^{\circ} \text{ totale} = n^r$ Esempio: quante possibili diverse colonne del totocalcio di $r=13$ segni presi da un insieme di $n=3$ elementi (1,2,x) ? $n^{\circ} \text{ totale} = 3^{13}$	$n^{\circ} \text{ totale} = \frac{(n+r-1)!}{r!(n-1)!}$ Esempio: quante diverse liste di $r=4$ oggetti si possono fare scegliendo da un campionario di 10 oggetti ? $n^{\circ} \text{ totale} = \frac{(10+4-1)!}{4!(10-1)!}$ Nota: si tratta di semplici liste, in cui la sequenza degli oggetti non ha importanza. Hanno scarsa applicazione.
<i>elementi non ripetuti</i>	$n^{\circ} \text{ totale} = \frac{n!}{(n-r)!}$ Esempio: quanti possibili diversi ordini di arrivo di $r=5$ atleti in una gara a cui partecipano $n=20$ atleti ? $n^{\circ} \text{ totale} = \frac{20!}{(20-5)!}$ Attenzione: se $r = n$ si parla di permutazioni . In tal caso, ovviamente: $n^{\circ} \text{ totale} = n!$	$n^{\circ} \text{ totale} = \binom{n}{r} = \frac{n!}{r!(n-r)!}$ Esempio: quante possibili diverse mani di $r=5$ carte qualsiasi si possono avere giocando a poker con un mazzo di $n=32$ carte ? $n^{\circ} \text{ totale} = \binom{32}{5} = \frac{32!}{5!(32-5)!}$ Attenzione: corrisponde al coefficiente binomiale del triangolo di Tartaglia (vedi distribuzione binomiale)