

Principio delle probabilità totali e formula di Bayes

Esempio

Indichiamo con $+$ l'evento "il test da esito positivo", con I l'evento "individuo infetto"

$$P(I) = 0.001, \quad P(+|I) = 0.99, \quad P(+|\bar{I}) = 0.02$$

Calcolare la $P(+)$ e $P(I|+)$

$$\begin{aligned} P(+) &= P(+ \cap I) + P(+ \cap \bar{I}) \\ &= P(+|I) \cdot P(I) + P(+|\bar{I}) \cdot P(\bar{I}) \\ &= 0.99 \cdot 0.001 + 0.02 \cdot (1 - 0.001) \\ &= 0.02097 \end{aligned}$$

$$\begin{aligned} P(I|+) &= \frac{P(I \cap +)}{P(+)} = \frac{P(+|I) \cdot P(I)}{P(+)} \\ &= \frac{P(+|I) \cdot P(I)}{P(+|I) \cdot P(I) + P(+|\bar{I}) \cdot P(\bar{I})} \\ &= \frac{0.99 \cdot 0.001}{0.02097} = 0.0472 \end{aligned}$$

Possiamo rileggere il problema in questo modo. Siano rispettivamente

$$P(D) = 0.001, \quad P(S|D) = 0.99, \quad P(S|\bar{D}) = 0.02$$

la probabilità che un pezzo sia difettoso, la probabilità che sia segnalato difettoso dato che lo sia e la probabilità che sia segnalato difettoso dato che non lo sia. La probabilità che sia segnalato un pezzo difettoso è

$$P(D) = 0.02097$$

Mentre la probabilità che un pezzo segnalato difettoso lo sia realmente è

$$P(D|S) = 0.0472$$

Se noi scartiamo tutti i pezzi segnalati difettosi di quelli scartati solo il 5% lo sono realmente!!!!

Principio delle probabilità totali e formula di Bayes

Partizione di Ω

k eventi A_1, A_2, \dots, A_k sono una partizione se e solo se

$$\bigcup_{i=1}^k A_i = \Omega \quad \text{e}$$

scelti due qualsiasi A_i e A_j allora $A_i \cap A_j = \emptyset$

Principio delle probabilità totali

se A_i sono una partizione di Ω e E è un qualsiasi altro evento

$$P(E) = \sum_{i=1}^k P(E|A_i) \cdot P(A_i)$$

Formula di Bayes

se A_i sono una partizione di Ω e E è un qualsiasi altro evento

$$P(A_j|E) = \frac{P(E|A_j) \cdot P(A_j)}{\sum_{i=1}^k P(E|A_i) \cdot P(A_i)}, \quad j = 1, \dots, k$$

Esercizio

Cinque scatole di dischi magnetici sono numerate con le cifre 1, 2, ..., 5. Di 10 dischi, la i -esima scatola ne contiene “ i ” danneggiati e “ $10-i$ ” intatti. Si sceglie una scatola a caso e, senza poterne leggere il numero che la identifica, se ne estrae un disco. Con che probabilità il disco è difettoso? Constatando che il disco estratto è difettoso, con che probabilità proviene dalla i -esima scatola?

S_i = “si estrae la scatola numero i ”

D = “si estrae dischetto difettoso”

$$P(S_1) = P(S_2) = P(S_3) = P(S_4) = P(S_5) = \frac{1}{5}$$

$$P(D|S_i) = \frac{i}{10}, \quad i = 1, \dots, 5$$

$$P(D) = \sum_{i=1}^5 P(D|S_i) \cdot P(S_i) = \sum_{i=1}^5 \frac{i}{10} \cdot \frac{1}{5} = \frac{1}{50} \sum_{i=1}^5 i = \frac{3}{10}$$

$$P(S_i|D) = \frac{P(D|S_i) \cdot P(S_i)}{P(D)} = \frac{\frac{i}{10} \cdot \frac{1}{5}}{\frac{3}{10}} = \frac{i}{15}$$

Classificazione e distribuzioni di frequenza

Problema: studiare l'andamento del primo semestre di un corso di laurea che prevede 3 esami per semestre

Rileviamo:

- numero di esami superati E (0, 1, 2 o 3)
- voti in trentesimi V negli esami superati
- valutazione del rendimento R nel singolo esame (*sufficiente, buono, ottimo*)
- condizione C di studente (*lavoratore o non lavoratore*)
- frequenza ai corsi F (*si o no oppure assidua, saltuaria*).

Tipologie di fenomeni

- E è il risultato di un'operazione di conteggio (numero di esami): è un fenomeno *quantitativo discreto*
- V è il risultato di una *misura* della performance dello studente: è un fenomeno *quantitativo continuo* (qui è *discretizzato*)
- R è una rilettura della variabile V . Può assumere i valori S = “sufficiente” (18-22), B = “buono” (23-26) e O = “ottimo” (27-30): è un fenomeno *qualitativo rilevato su scala ordinale*
- C : è un fenomeno *qualitativo rilevato su scala nominale* (*si o no*)
- F : è un fenomeno qualitativo. Su scala nominale se *si* o *no*, su scala ordinale se invece avessimo utilizzato la classificazione *assidua* e *saltuaria*

Tipologie di fenomeni

qualitativo	$\left\{ \begin{array}{ll} \text{su scala nominale,} & =, \neq \\ \text{su scala ordinale,} & =, \neq, < \end{array} \right.$
quantitativo	$\left\{ \begin{array}{ll} \text{discreto,} & =, \neq, < \text{ (conteggio)} \\ \text{continuo,} & =, \neq, < \text{ (misurazione)} \end{array} \right.$

Per i fenomeni quantitativi esiste un'ulteriore classificazione

Tipologie di scale

scale di intervalli:	$=, \neq, <, +, -$
scale di rapporti:	$=, \neq, <, +, -, *, \div$

Grado	Denominazione	Effetti
1	Strumentale	È percepita solo dai sismografi.
2	Leggerissima	È avvertita solo dalle persone ipersensibili in momenti di quiete e ai piani più elevati.
3	Leggera	Viene avvertita da un numero maggiore di persone, le quali non si allarmano perché generalmente non si rendono conto che si tratta effettivamente di scosse telluriche.
4	Mediocre	Le persone che sono in casa l'avvertono e qualcuna anche tra quelle che si trovano all'aperto. I lampadari oscillano, i pavimenti possono dare degli scricchiolii.
5	Forte	Sentita tanto dalle persone che si trovano in casa quanto da quelle fuori casa. Gli oggetti sospesi oscillano ampiamente, gli orologi a pendolo si fermano, si hanno tremiti dei vetri e delle stoviglie. Si ha risveglio brusco dal sonno e può generare panico senza danni alle persone.
6	Molto forte	Gli oggetti cadono e così i calcinacci dei muri in cui si possono formare lievi lesioni. La popolazione, presa dal panico, abbandona le case.
7	Fortissima	Possono cadere comignoli e tegole, mentre i muri presentano lesioni non molto gravi. Suono di campane.
8	Rovinoso	Lesioni gravi ai fabbricati, crollo di qualche muro interno. Qualche ferito, raramente vittime.
9	Disastrosa	Alcuni crolli di case, altri edifici gravemente lesionati. Molti i feriti, non numerose le vittime.
10	Distruttrice	Crolli di molti fabbricati. Paresche le vittime, moltissimi i feriti.
11	Catastrofe	Numerose vittime. Quasi tutti gli edifici crollati.
12	Grande catastrofe	Formazione di crepacci e frane. Distruzione di qualsiasi opera umana.

Che tipo di informazione vogliamo ricavare da questi dati?

unità statistica	E	V	R	F	C
27	2	28	O	si	no
27	2	25	B	si	no
131	1	29	O	si	no
271	3	28	O	no	si
271	3	18	S	no	si
271	3	18	S	no	si
311	1	18	S	no	no
321	2	26	B	si	si
321	2	25	B	si	si

- Qual'è il rendimento *medio*?
- Qual'è il comportamento più frequentemente riscontrato?
- Qual'è il voto medio?
- Esiste un'effetto positivo dovuto alla frequenza dei corsi?

Distribuzioni di frequenza

n dati relativi a n individui

X il fenomeno (o variabile statistica)

x_i possibili valori che può essere assunto da X

n_i la frequenza con cui è assunto x_i

Esempio

5 studenti

E numero di esami sostenuti

$x_1 = 0, x_2 = 1, x_3 = 2, x_4 = 3$ e così via

$n_1 = 0, n_2 = 2, n_3 = 2, n_4 = 1$

x_i	n_i
1	2
2	2
3	1
Totale	5

Esempio Numero di esami sostenuti da 350 studenti

3, 1, 3, 1, 3, 1, 1, 3, 2, 2, 1, 3, 2, 1, 1, 2, 0, 2, 1,
 1, 1, 3, 2, 1, 1, 1, 1, 0, 2, 0, 0, 1, 3, 1, 2, 2, 2, 2,
 2, 3, 3, 2, 2, 1, 2, 0, 1, 2, 0, 0, 2, 3, 3, 0, 2, 2, 2,
 2, 1, 2, 2, 0, 2, 0, 2, 2, 1, 1, 1, 3, 1, 1, 2, 1, 1, 1,
 1, 1, 3, 2, 0, 2, 3, 1, 2, 2, 1, 1, 3, 0, 3, 3, 1, 0, 2,
 1, 0, 2, 2, 2, 0, 2, 2, 2, 2, 0, 2, 2, 0, 1, 11, 2, 0,
 2, 0, 1, 0, 1, 1, 3, 1, 2, 1, 0, 1, 2, 0, 2, 2, 2, 1, 1,
 0, 1, 1, 1, 0, 1, 2, 2, 1, 2, 2, 3, 0, 2, 10, 3, 1, 1,
 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 2, 2, 1, 0,
 2, 1, 2, 2, 1, 1, 2, 2, 2, 1, 3, 2, 1, 2 0, 1, 3, 1, 1,
 1, 2, 2, 0, 2, 1, 2, 3, 1, 1, 1, 3, 2, 2, 2, 1, 1, 1, 0,
 3, 1, 3, 0, 1, 2, 1, 0, 1, 3, 2, 2, 2, 2, 2, 0, 2, 2, 2,
 2, 1, 1, 2, 2, 2, 1, 2, 1, 2, 1, 0, 1, 1, 2, 2, 1, 2, 2,
 2, 3, 0, 2, 1, 2, 1, 0, 1, 1, 1, 1 2, 2, 2, 2, 2, 0, 1,
 1, 1, 2, 0, 1, 0, 1, 1, 2, 1, 1, 1, 2, 1, 2, 0, 1, 2, 2,
 1, 2, 1, 3, 1, 2, 2, 0, 2, 2, 3, 2, 2, 1, 1, 1, 2, 3, 3,
 1, 2, 1, 2, 1, 2, 2, 2, 1, 1, 1, 0, 1, 1, 2, 1, 2, 0, 2,
 1, 0, 1, 1, 2, 3, 1, 2, 3, 1, 0, 2, 1, 3, 3, 1, 2, 2, 2,
 2, 2, 2, 1, 2, 1, 1, 1

x_i	n_i
0	45
1	136
2	133
3	36
Totale	350

La frequenza relativa f_i è il rapporto tra n_i (la frequenza assoluta) e il numero totale delle informazioni disponibili n

x_i	n_i	$f_i = n_i/n$	$p_i = f_i \cdot 100\%$
0	45	0.13	13%
1	136	0.39	39%
2	133	0.38	38%
3	36	0.10	10%
Totale	$n = 350$	1.00	100%

Distribuzioni di frequenza

k : numero totale dei diversi valori assunti dal fenomeno statistico

n_i : frequenza con cui compare il valore x_i del fenomeno statistico

$n_1 + n_2 + \dots + n_k = n$, con n numero di osservazioni o ampiezza del campione

Le distribuzioni di frequenza sono l'insieme dei valori x_i e delle frequenze:

assolute: $\{x_i, n_i\}$

relative: $\left\{x_i, f_i = \frac{n_i}{n}\right\}$

percentuali: $\{x_i, p_i = f_i \cdot 100\%\}$

Esempio Rendimento (S = sufficiente, B = buono e O = ottimo) rilevato su 350 studenti:

S, S, O, S, O, O, B, B, B, O, O, B, O, B, B, O, NA, B, O, O, B, O,
 B, B, B, B, O, NA, S, NA, NA, O, B, O, O, B, O, O, B, S, O, B, O,
 B, O, NA, B, O, NA, NA, O, O, O, NA, B, S, B, B, B, O, B, NA, O,
 NA, B, O, O, O, O, B, O, B, O, B, B, B, O, O, B, S, NA, O, O, O,
 S, S, B, S, O, NA, O, B, B, NA, B, O, NA, O, O, S, NA, O, O, O, B,
 NA, O, O, NA, O, B, O, S, NA, B, NA, O, NA, B, B, B, O, O, B, NA,
 S, O, NA, O, O, O, B, B, NA, B, B, B, NA, S, O, O, B, B, O, S, NA,
 B, B, NA, O, B, S, O, B, B, B, B, O, S, O, O, O, O, O, O, B, B, O,
 O, O, NA, B, B, O, O, O, B, O, B, B, O, B, B, O, B, NA, O, O, O, B,
 O, O, O, NA, O, O, B, O, B, O, O, O, B, S, O, O, B, O, NA, O, O,
 S, NA, O, B, O, NA, B, O, B, B, O, O, O, NA, B, O, O, S, O, B, O,
 O, O, B, B, B, O, B, NA B, B, B, O, B, B, O, O, O, NA, O, O, O, B,
 NA, O, O, B, O, B, O, O, O, B, NA, B, B, B, O, NA, O, NA, B, B, O,
 O, O, B, O, S, O, NA, O, O, O, B, B, B, B, B, B, B, NA, O, O, B, O,
 B, O, O, B, O, O, O, B, B, O, B, O, O, B, O, S, B, B, NA, O, B, B,
 B, O, NA, B, B, NA, B, O, O, B, O, S, O, O, NA, O, O, B, O, O, O,
 O, B, O, B, B, O, O, B, O, B

x_i	n_i	f_i	p_i	f'_i	p'_i
S	23	0.07	7%	0.07	7%
B	124	0.41	41%	0.35	35%
O	158	0.52	52%	0.45	45%
NA	45	—	—	0.13	13%
Totale	350	1.00	100%	1.00	100%

Dati raccolti in classi

Se abbiamo un numero elevato di possibili valori x_i per il fenomeno X si possono raccogliere i dati in una tabella in cui nella prima colonna mettiamo le classi $x_i \vdash x_{i+1}$ e nella seconda le frequenze (assolute, relative o percentuali) di classe, cioè il numero totale di osservazioni che hanno valori del carattere X inclusi nella classe $x_i \vdash x_{i+1}$.

	Iscritti alla Società	Avviati al lavoro
Maschi	53.4	59.0
Femmine	46.6	41.0
	100.0	100.0
< 25	32.9	35.9
25-29	29.2	26.6
30-34	17.5	16.8
35-39	9.8	9.7
40-54	9.9	10.3
≥55	0.7	0.7
	100.0	100.0

Frequenze cumulate

Rispondono a domande del tipo:

- Quanti studenti hanno dato meno di 2 esami?
- Quanti studenti risultano avere un rendimento medio al più buono?

$45 + 136 = 181$, $181/350 \cdot 100\% = 51.7\%$, quindi oltre la metà degli studenti non ha dato più di un esame.

Frequenze cumulate

cumulate assolute: $N_i = \sum_{j=1}^i n_j = n_1 + n_2 + \cdots + n_i$

cumulate relative: $F_i = \sum_{j=1}^i f_j = f_1 + f_2 + \cdots + f_i$

cumulate percentuali : $P_i = \sum_{j=1}^i p_j = p_1 + p_2 + \cdots + p_i$

Esempio Numero esami sostenuti

x_i	n_i	f_i	p_i	N_i	F_i	P_i
0	45	0.13	13%	45	0.13	13%
1	136	0.39	39%	181	0.52	52%
2	133	0.38	38%	314	0.90	90%
3	36	0.10	10%	350	1.00	100%
Totale	$n = 350$	1.00	100%	—	—	—

Esempio Rendimento

x_i	n_i	f_i	p_i	N_i	P_i
S	23	0.07	7%	23	7%
B	124	0.41	41%	147	48%
O	158	0.52	52%	305	100%
NA	45	—	—	—	—
Totale	350	1.00	100%	305	100%

Si ricordi che valgono sempre le seguenti relazioni tra frequenze cumulate e frequenze assolute:

Relazioni da ricordare

Se abbiamo k frequenze assolute, allora

$$\begin{cases} N_1 & = n_1 \\ N_k & = n \\ N_i - N_{i-1} & = n_i \end{cases}$$

Analogamente per le F_i e le P_i .

Attenzione : *non ha alcun significato calcolare le frequenze cumulate se il fenomeno statistico non è di tipo ordinabile come i fenomeni qualitativi su scala nominale (fenomeno C nell'esempio degli studenti).*