Lezione 11/10/2022

La probabilità condizionata (esercizi)

La probabilità e l'affidabilità dei tamponi

I teoremi delle probabilità totali e di Bayes permettono di conoscere l'affidabilità di un tampone di una malattia infettiva X. Consideriamo i seguenti eventi:

- M = una persona ha la malattia infettiva X;
- NM = una persona non ha la malattia infettiva X;
- $+$ = una persona è positiva al tampone di X;

$P(M|+)$ indica quanto è affidabile il tampone, perchè rappresenta la probabilità che una perona positiva sia davvero infetta. Per il teorema di Bayes e il teorema delle probabilità totali:

$$P(M|+) = \frac{P(+|M)P(M)}{P(+)} = \frac{P(+|M)P(M)}{P(+|M)P(M) + P(+|NM)P(NM)}.$$

Le probabilità $P(+|M)$, $P(+|NM)$, $P(M)$ e $P(NM)$ si calcolano utilizzando la definizione di probabilità frequentista:

1. Si considerano i seguenti insiemi
   - Ω = un insieme di persone;
   - Ω_M = tutte le persone di Ω infette;
   - Ω_{NM} = tutte le persone di Ω che non sono infette;
   - Ω_M^+ = tutte le persone di Ω_M che sono positive;
   - Ω_{NM}^+ = tutte le persone di Ω_{NM} che sono positive.
2. Le probabilità sono le seguenti:

$$P(M) = \frac{\#\Omega_M}{\#\Omega}, P(NM) = \frac{\#\Omega_{NM}}{\#\Omega}, P(+|M) = \frac{\#\Omega_M^+}{\#\Omega_M} \text{ e } P(+|NM) = \frac{\#\Omega_{NM}^+}{\#\Omega_{NM}}.$$

La probabilità per risalire alla causa di un effetto

Si considerano i seguenti eventi:

- $B =$ si osserva una macchia sulla radiografia;
- $A_1 =$ tumore maligno;
- $A_2 =$ tumore benigno;
- $A_3 =$ altra causa.

Si vuole stabilire quale sia la causa (la malattia) che con maggiore probabilità ha generato l'evento B (la macchia).

Si calcola $P(A_1|B)$, $P(A_2|B)$ e $P(A_3|B)$ con il teorema di Bayes.

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)}.$$

Le probabilità $P(B|A_1)$, $P(A_1)$, $P(B|A_2)$, $P(A_2)$, $P(B|A_3)$ e $P(A_3)$ si calcolano con la definizione frequentista di probabilità.

1. Si considerano i seguenti insiemi:
   - $\Omega =$ un insieme di persone;
   - $\Omega_{A_1} =$ tutte le persone di Ω che hanno un tumore maligno;
   - $\Omega_{A_2} =$ tutte le persone di Ω che hanno un tumore benigno;
   - $\Omega_{A_3} =$ tutte le persone di Ω che non hanno un tumore;
   - $\Omega^B_{A_1} =$ tutte le persone di Ω_{A_1} che hanno la macchia;
   - $\Omega^B_{A_2} =$ tutte le persone di Ω_{A_2} che hanno la macchia;
   - $\Omega^B_{A_3} =$ tutte le persone di Ω_{A_3} che hanno la macchia.
2. Le probabilità sono le seguenti:

$$P(A_i) = \frac{\#\Omega_{A_i}}{\#\Omega} \quad \text{e} \quad P(B|A_i) = \frac{\#\Omega^B_{A_i}}{\#\Omega_{A_i}} \quad \text{con } i \in \{1, 2, 3\}.$$

Se ad esempio $P(A_1|B) = 0.2$, $P(A_2|B) = 0.8$ e $P(A_3|B) = 0.3$ allora a B viene attribuita la causa A_2.

Estrazioni da un'urna, eventi dipendenti e dipendenti.

Consideriamo l'esperimento casuale che cosniste nell'estrarre palline da un'urna con 3 palline rosse e 2 bianche e gli eventi $A =$ *la prima pallina estratta è rossa* e $B =$ *la seconda pallina estratta è bianca.*

- Se le estrazioni sono *con rimpiazzo* (o *reimbussolamento* o *reimmissione*, cioè la pallina estratta si reinserisce nell'urna prima della di una nuova estrazione), gli eventi A e B sono indipendenti.
- Se le estrazioni sono *senza rimpiazzo* (o *senza reimbussolamento* o *senza reimmissione*, cioè la pallina estratta non si reinserisce nell'urna prima della di una nuova estrazione), gli eventi A e B sono dipendenti.

Eventi indipendenti ed eventi contrari

Se A e B sono indipendenti allora

- $\bar{A}$ e $\bar{B}$ sono indipendenti;
- $\bar{A}$ e B sono indipendenti;
- A e $\bar{B}$ sono indipendenti.

Esercizi sulla probabilità condizionata

Esercizio 1. *Siano A e B due eventi indipendenti tali che $P(A) = 0.4$ e $P(B) = 0.3$.*

(a) Calcola la probabilità che si verifichi A oppure B.

(b) Calcola la probabilità che si verifichi A e contemporaneamente B non si verifichi.

(c) Calcola la probabilità di A dato B.

(d) Calcola la probabilità che nè A e nè B si verifichino.

Svolgimento. *(a) Osserviamo che*

- *non si sa se A e B siano incompatibili, per cui $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;*
- *dato che A e B sono indipendenti $P(A \cap B) = P(A)P(B)$.*

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(A)P(B) = 0.4 + 0.3 - 0.4 \cdot 0.3 = 0.53.$

(b) Osserviamo che se A e B sono indipendenti allora anche A e $\bar{B}$ sono indipendenti, quindi $P(A \cap \bar{B}) = P(A)P(\bar{B}) = P(A)(1 - P(B))$ (per la proprietà 1)$=0.4 \cdot (1 - 0.3) = 0.28$.

(c) Dato che A e B sono indipendenti, $P(A|B) = P(A)$.

(d) $P(\bar{A}\cap\bar{B}) = P(\overline{A\cup B}) = 1 - P(A\cup B)$ *(per la proprietà 1)=1-0.58=0.42.*

Osservazione In generale, $P(\bar{A}\cap\bar{B}) = P(\overline{A\cup B})$ e $P(\bar{A}\cup\bar{B}) = P(\overline{A\cap B})$ per le leggi di De Morgan ($\bar{A}\cap\bar{B} = \overline{A\cup B}$ e $\bar{A}\cup\bar{B} = \overline{A\cap B}$).

Esercizio 2. *Abbiamo una moneta sbilanciata e due urne che contengono palline bianche e nere. Lancio la moneta e se esce testa estraiamo una pallina dalla prima urna, altrimenti la estraiamo dalla seconda. Tenendo conto della tabella:*

Moneta	Urna 1	Urna 2
P(testa)= $\frac{4}{10}$	P(pallina nera)=$\frac{3}{5}$	P(pallina nera)=$\frac{4}{5}$
P(croce)=$\frac{6}{10}$	P(pallina bianca)=$\frac{2}{5}$	P(pallina bianca)= $\frac{1}{5}$

(a) *calcola la probabilità di estrarre una pallina nera;*

(b) *calcola la probabilità che dal lancio della moneta è uscito testa sapendo che è stata estratta una pallina nera.*

Svolgimento. *(a) Consideriamo gli eventi*
$N =$ esce una pallina nera,
$U_1 =$ ho estratto la pallina dall'urna 1 *e*
$U_2 =$ ho estratto la pallina dall'urna 2.

Per il teorem della probabilità totali,

$$P(N) = P(N|U_1)P(U_1) + P(N|U_2)P(U_2).$$

Osserviamo che $P(U_1)$*=P(testa)=*$\frac{4}{10}$, $P(U_2)$ *=P(croce)=*$\frac{6}{10}$, $P(N|U_1) = \frac{3}{5}$ *e* $P(N|U_2) = \frac{4}{5}$, *quindi*

$$P(N) = \frac{3}{5}\cdot\frac{4}{10} + \frac{4}{5}\cdot\frac{6}{10} = \frac{18}{25}.$$

(b) *Ricordo che scelgo è uscito testa se e solo se ho scelto l'urna 1, quindi devo calcolare* $P(U_1|N)$. *Per il teorema di Bayes,*

$$P(U_1|N) = \frac{P(N|U_1)P(U_1)}{= P(N|U_1)P(U_1) + P(N|U_2)P(U_2)} = \frac{\frac{3}{5}\cdot\frac{4}{10}}{\frac{18}{25}} = \frac{1}{3}.$$

La statistica descrittiva

La statistica

- è la branca della matematica che si occupa dello studio dei dati,
- si divide in *statistica descrittiva* e *statistica inferenziale*.

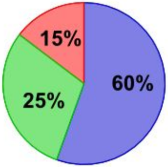
La statistica descrittiva si occupa di raccogliere, elaborare e rappresentare i dati di una popolazione (l'insieme delle entità d'interesse). La statistica inferenziale invece studia i dati di un campione, cioè un sottoinsieme, della popolazione; successivamente, le informazioni ottenute sul campione si cercano di estendere a tutta la popolazione utilizzando la teoria delle probabilità.

Statistica Descrittiva e Statistica Inferenziale

Esempio: Vogliamo sapere come i cittadini di un paese si ripartiscono tra tre partiti politici P1, P2 e P3.

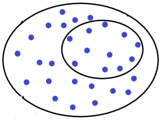
Statistica Descrittiva

1. Si chiede ad ogni cittadino la sua preferenza
2. Si elaborano i dati ottenuti (ad esempio calcolando le percentuali: P1=15%, P2=25% e P3=60%)
3. Si rappresentano graficamente



Statistica Inferenziale

1. Si interroga un numero limitato di cittadini (sondaggio)



2. Si ci chiede se a partire dal sondaggio possiamo inferire il voto di tutti i cittadini
3. Cosa succede se consideriamo due sondaggi diversi? Otteniamo gli stessi risultati?

Statistica descrittiva

Fenomeno studiato	Il colore degli occhi degli italiani	Altezza (misurata in metri) degli studenti di una classe	L'anno di nascita degli iscritti a una palestra
Popolazione	Tutti gli italiani	Gli studenti della classe	Tutti gli iscritti alla palestra
Carattere	Il colore degli occhi	La misura dell'altezza	L'anno di nascita
Modalità	Verdi, azzurri, marroni ecc.	Per esempio: 1,72 m; 1,85 m; 1,78 m...	..., 1970, ..., 1981, 1965, ...
Tipo di carattere	Qualitativo	Quantitativo continuo	Quantitativo discreto

Frequenza assoluta e distribuzione di frequenza

Si definisce frequenza assoluta di una modalità il numero che esprime quante volte si è manifestata quella modalità.

Si definisce funzione di distribuzione delle frequenze la funzione che associa a ogni modalità di un carattere la rispettiva frequenza.

Studente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Colore degli occhi	N	M	A	N	N	M	A	V	V	M	M	N	N	V	V	N	M	N

Una distribuzione di frequenze si rappresenta con una tabella a due colonne (o a due righe): nella prima sono riportate le modalità, nella seconda le rispettive frequenze.

Colore degli occhi	Frequenza (= numero di studenti)
Nero	7
Marrone	5
Azzurro	2
Verde	4

Distribuzione di frequenze

Se indichiamo con X un generico carattere, con $x_1, \ldots, x_k$ le differenti modalità osservate su n individui, e con $f_1, \ldots, f_k$ le frequenze delle varie modalità, la distribuzione di frequenze del carattere si può rappresentare tramite una tabella che può essere di due tipi

X	Frequenza
x_1	f_1
x_2	f_2
.....	.....
x_k	f_k
Totale	n

X	x_1	x_2	.....	x_k	**Totale**
Frequenza	f_1	f_2	.....	f_k	n

Nell'ultima riga (o colonna), per comodità, si è soliti riportare la somma delle frequenze, che è uguale al numero di individui del collettivo.

Distribuzioni per classi

Rileviamo ora la statura di 18 studenti. Il risultato della rilevazione fornisce i dati grezzi riassunti in questa tabella

Studente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Altezza (in cm)	173	164	174	180	182	176	184	185	170	172	186	167	188	183	168	176	184	178

E' utile accorpare preliminarmente le modalità in intervalli tra loro disgiunti, detti classi, quindi costruire la distribuzione di frequenze delle classi. Per esempio, possiamo suddividere le possibili altezze degli studenti nei seguenti intervalli:

$$(160, 165](165, 170](170, 175](175, 180](180, 185](185, 190]$$

Altezza degli studenti (in cm)	(160,165]	(165,170]	(170,175]	(175,180]	(180,185]	(185,190]
Numero di studenti	1	3	3	4	5	2

Data una classe definita da un intervallo di estremi a e b, con $a < b$, si definisce ampiezza della classe il numero $b - a$ e valore centrale della classe il numero $\frac{a+b}{2}$.

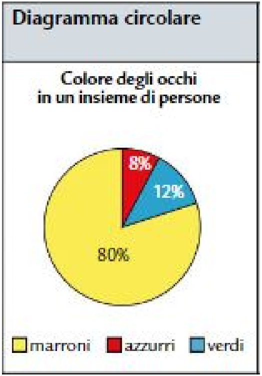
Per esempio, in riferimento al precedente, la classe $(160, 165]$ ha ampiezza 5 e il suo valore centrale è $\frac{160+165}{2} = 162,5$. Le classi in cui vengono suddivisi i dati possono avere tutte la stessa ampiezza (come nell'esempio precedente) o ampiezza diversa tra loro.

Distribuzioni di Frequenza relativa e percentuali

Si definisce frequenza relativa di una modalità il rapporto fra la sua frequenza assoluta e il numero complessivo di individui del collettivo.

Colore degli occhi	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
Nero	7	$\frac{7}{18} \simeq 0{,}389$	38,9%
Marrone	5	$\frac{5}{18} \simeq 0{,}278$	27,8%
Azzurro	2	$\frac{2}{18} = \frac{1}{9} \simeq 0{,}111$	11,1%
Verde	4	$\frac{4}{18} = \frac{2}{9} \simeq 0{,}222$	22,2%
Totale	18	**1**	**100%**

Diagramma circolare



I diagrammi circolari rappresentano la distribuzione del carattere studiato tramite settori circolari di ampiezza proporzionale alla frequenza con cui le modalità si presentano. Si utilizzano soprattutto per rappresentare le diverse parti in cui una totalità è suddivisa. Per determinare l'ampiezza del settore che rappresenta una data modalità basta moltiplicare la sua frequenza relativa per 360°.

Diagramma a Barre

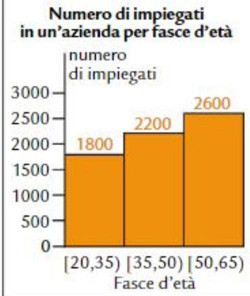
Diagramma a barre (rettangoli distanziati)

Voti

I diagrammi a barre rappresentano la distribuzione di frequenze del carattere studiato, tramite rettangoli, distanziati tra loro, aventi tutti la stessa base e altezza proporzionale alla frequenza (assoluta o relativa) della modalità osservata. L'asse orizzontale può essere graduato (se il carattere è quantitativo) oppure può servire soltanto come base di appoggio dei rettangoli (se il carattere è qualitativo).

Istogramma: classi di uguale ampiezza

Istogramma (rettangoli affiancati)



È un grafico costituito da rettangoli non distanziati, ciascuno dei quali ha un'*area* proporzionale alla frequenza della classe rappresentata.
Nei casi più comuni, come quello rappresentato (in cui le classi hanno tutte la stessa ampiezza), è sufficiente associare a ciascuna classe un rettangolo avente base di misura uguale all'ampiezza della classe e altezza di misura uguale alla frequenza.