STATISTICA DESCRITTIVA MONOVARIATA: ESERCIZI SVOLTI

ESERCIZIO 1

La matrice Caso x Variabile presentata di seguito contiene i dati relativi alle proprietà "genere, istruzione ed età" rilevate in un gruppo di soggetti che hanno fatto domanda di iscrizione a un corso di formazione per intervistatori.

ID	GENERE	TS	ETA'
1	1	2	29
2	1	2	21
3	2	4	27
4	1	3	26
5	2	3	22
6	2	3	20
7	2	2	30
8	1	2	22
9	1	4	37
10	1	3	25
11	2	3	32
12	2	3	24
13	2	3	31
14	1	2	23
15	1	4	35

Dati previsti nelle schede di iscrizione.

Legenda delle etichette e dei codici assegnati alle variabili ed alle rispettive modalità.

ID: identificativo dei soggetti

GENERE
1: maschio
2: femmina

TS: titolo di studio conseguito
0: nessun titolo
1: licenza elementare
2: licenza media inferiore
3: licenza media superiore
4: diploma di laurea

ETA': età in anni compiuti

Utilizzando i dati in matrice, calcolare:
1) la distribuzione di frequenze assolute di ciascuna delle variabili;
2) la distribuzione di frequenze assolute dell'età che si ottiene raggruppando i soggetti in tre classi aventi la stessa ampiezza;
3) la tendenza centrale della distribuzione relativa al genere.

❐ Considerazioni preliminari

Il primo vettore colonna della matrice indica semplicemente l'identificativo dei soggetti ed è costituito da un numero progressivo che varia da 1 a N, dove N rappresenta la numerosità del collettivo in esame. Tale vettore non è oggetto di analisi e pertanto l'attenzione deve concentrarsi sui restanti, che rappresentano rispettivamente una proprietà categoriale (il genere), una ordinale (l'istruzione, rilevata in termini di titolo di studio conseguito dai soggetti) e una cardinale (l'età).

& Svolgimento

Una distribuzione di frequenze assolute non è altro che il conteggio del numero di volte con cui ciascuna modalità di una variabile si presenta nel collettivo considerato.

Iniziamo dal genere. Per ottenere la distribuzione di frequenze, è sufficiente contare il numero di maschi e di femmine presenti sulle righe della matrice: il risultato di questa operazione consiste nella frequenza assoluta di ciascuna modalità della variabile, che riportiamo nella colonna contrassegnata dal simbolo n.
La distribuzione così generata prende il nome di serie sconnessa di frequenze.

GENERE	**n**
1: maschio	8
2: femmina	7
Totale	15

La medesima operazione va effettuata anche per il terzo vettore colonna della matrice, che rappresenta il titolo di studio conseguito dai soggetti iscritti al corso di formazione; in questo caso, la distribuzione che si origina è una serie ordinata di frequenze, trattandosi di una variabile per cui è possibile stabilire un ordine fra le modalità.

TS: titolo di studio conseguito	**n**
0: nessun titolo	0
1: licenza elementare	0
2: licenza media inferiore	5
3: licenza media superiore	7
4: diploma di laurea	3
Totale	15

Come si può notare, contrariamente a quanto previsto prima dell'effettiva rilevazione, tra coloro che hanno presentato domanda d'iscrizione al corso non figurano soggetti con la licenza elementare o privi di un titolo. Si tratta di un elemento informativo che può essere evidenziato soltanto completando la serie con tutti i codici prestabiliti nelle schede di iscrizione[1].
La terza variabile della matrice è l'età che, data la natura cardinale, possiamo descrivere attraverso una seriazione di frequenze.

ETA'	**n**
20	1
21	1
22	2
23	1
24	1
25	1
26	1
27	1
29	1
30	1
31	1
32	1
35	1
37	1
Totale	15

[1] Una simile distribuzione non potrebbe essere generata con un software statistico, perché i programmi informatici che eseguono automaticamente queste operazioni lavorano esclusivamente con i codici presenti nelle celle della matrice dei dati.

In realtà, com'è possibile osservare, questo tipo di rappresentazione non è efficace: le singole modalità della variabile hanno una frequenza assoluta talmente bassa che la seriazione finisce per essere una riproduzione pressoché identica della colonna presente nella matrice dei dati.
Una descrizione più sintetica può essere ottenuta riorganizzando i dati in classi di uguale ampiezza, come richiesto dal secondo quesito dell'esercizio.
A tal fine occorre anzitutto determinare l'intervallo di valori che la proprietà ha assunto nel collettivo: poiché il soggetto più giovane ha 20 anni e quello con l'età maggiore ne ha 37, l'intervallo di variazione dell'età è pari a 37-20, ovvero 18 anni. In secondo luogo, per calcolare l'ampiezza delle classi è sufficiente dividere tale intervallo per il numero di classi desiderato, che in questo caso è pari a 3; pertanto l'ampiezza della generica classe è: $a_k=18/3=6$
A questo punto si calcola la frequenza assoluta per ognuna delle tre classi.

Classi di età	**n**
20 - 25	7
26 - 31	5
32 - 37	3
Totale	15

Infine, possiamo indicare la tendenza centrale della distribuzione del genere individuando la moda, che coincide con la modalità più ricorrente, ovvero "maschio".

ESERCIZIO 2

In una ricerca di mercato è stato chiesto ad un gruppo di consumatori di esprimere il proprio grado di interesse per il prototipo di un nuovo prodotto.
Con quale operatore è possibile individuare la tendenza centrale della distribuzione dei giudizi?

Livello di interesse	Frequenze assolute
Basso	132
Medio	44
Alto	138
Non risponde	47
Totale	361

❐ Considerazioni preliminari

La distribuzione riportata in tabella riguarda una variabile ordinale, e pertanto l'operatore adeguato per individuare la tendenza centrale dei giudizi è la mediana. Nello svolgimento dell'esercizio è necessario prestare attenzione alla modalità "non risponde", che evidentemente non è ordinabile, ed alla numerosità (N) del collettivo.

& Svolgimento

Occorre considerare soltanto il totale dei casi validi (soggetti che hanno espresso un giudizio), e ciò significa eliminare dalla distribuzione la modalità "non risponde" e calcolare le frequenze relative sul nuovo totale.
A questo punto è possibile procedere al calcolo della mediana, ricordando che essa coincide con la modalità della variabile a cui appartiene il caso (detto *caso mediano*) che divide esattamente in due la distribuzione.
Per agevolare l'individuazione del caso mediano, dopo aver controllato che le modalità della variabile siano poste in ordine crescente, è utile affiancare alle colonne delle frequenze semplici quelle cumulate.
Iniziamo i calcoli ricordando che la frequenza percentuale della generica classe k è uguale a $q_k = f_k \cdot 100 = \frac{n_k}{N} \cdot 100$, mentre la frequenza cumulata e quella percentuale cumulata si ottengono nel seguente modo: $n_k^{'} = n_1 + ... + n_k$ e analogamente $q_k^{'} = q_1 + ... + q_k$.

Livello di interesse	n	n'	q	q'
Basso	132	132	42 %	42 %
Medio	44	176	14 %	56 %
Alto	138	314	44 %	100 %
Totale	314		100 %	

Osservando il vettore delle frequenze percentuali cumulate (q'), notiamo che oltre il 50% dei casi ha espresso almeno un interesse medio per il prodotto e ciò significa che la modalità di risposta "medio" individua la mediana della distribuzione.
Procedendo in modo più analitico (e preciso), possiamo stabilire la mediana individuando il caso, o i casi, che dividono a metà la distribuzione. Il totale dei soggetti esaminati è pari e perciò i casi mediani sono due ${}^{1}C_{Mdn}=N/2$ e ${}^{2}C_{Mdn}=(N/2)+1$, ovvero:
${}^{1}C_{Mdn=}=314/2=157$ e ${}^{2}C_{Mdn}=(314/2)+1=158$
Dalla colonna delle frequenze assolute cumulate (n'), si constata che entrambi appartengono alla modalità di risposta "medio", che dunque rappresenta la mediana della distribuzione relativa al livello di interesse per l'eventuale nuovo prodotto.

ESERCIZIO 3

Al fine di misurare l'abilità nel riconoscimento di figure ad un gruppo di individui sono stati presentati 50 stimoli visivi, la cui forma da individuare è stata parzialmente occultata.
Nella tabella seguente è riportato il numero di figure riconosciute da ciascuno dei 16 soggetti che hanno partecipato alla sessione sperimentale.
Determinare il numero medio di figure riconosciute da questo gruppo di individui.

Identificativo	Numero di figure riconosciute X
1	31
2	15
3	44
4	37
5	25
6	23
7	14
8	46
9	29
10	15
11	17
12	42
13	30
14	38
15	34
16	45
Totale	485

❐ Considerazioni preliminari.

La variabile "numero di figure riconosciute" origina da una operazione di conteggio e pertanto si riferisce ad una proprietà misurata attraverso una scala assoluta, ovvero ad una variabile cardinale per cui è possibile calcolare una media aritmetica.

& Svolgimento

La media della distribuzione si ottiene sommando tutti i valori della variabile X, da x_1 a x_N, e dividendo tale somma per il numero di casi.
La formula per il calcolo della media aritmetica è: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Il gruppo analizzato ha una numerosità di 16 individui, che complessivamente ricordano 485 figure; di conseguenza, rispetto al compito di riconoscimento, l'abilità media del gruppo è:

$\bar{x} = \frac{1}{16} \cdot 485 = 30{,}3125$ ovvero 30 figure.

E' opportuno notare che questo operatore di tendenza centrale risulta informativo soltanto se nella distribuzione non compaiono valori estremi che possono alterare il valore della sommatoria posta a numeratore della formula.
Ad esempio, se nel gruppo esaminato vi fossero due individui estremamente inabili, che riconoscessero soltanto una figura sulle cinquanta presentate, la prestazione media del gruppo risulterebbe fortemente condizionata da questi due valori anomali.

Identificativo	Numero di figure riconosciute X'
1	31
2	15
3	1
4	37
5	25
6	23
7	14
8	1
9	29
10	15
11	17
12	42
13	30
14	38
15	34
16	45
Totale	397

$$\overline{x'} = \frac{1}{16} \cdot 397 = 24{,}8125$$

In questo caso è preferibile ricorrere alla mediana che, utilizzando soltanto l'informazione relativa all'ordine fra le modalità, risulta meno distorta dai valori assunti dalle osservazioni estreme.
A titolo esemplificativo, riportiamo i valori medi e mediani per le due distribuzioni analizzate nell'esercizio:

	media	mediana
X	30,3125	30,5
X'	24,8125	27,0

ESERCIZIO 4

In un grande Comune deve essere progettato il nuovo piano relativo ai percorsi urbani dei mezzi pubblici, che implica altresì la determinazione del numero di mezzi da impiegare per ciascuna linea. A tal proposito è stata effettuata una indagine presso 8000 cittadini, ai quali è stato chiesto di dichiarare con quale frequenza, nell'ambito di una settimana, si servono dei trasporti pubblici per recarsi al lavoro. I dati sono presentati nella successiva tabella.
Calcolare la variabilità della distribuzione servendosi dell'operatore più opportuno.

Utilizzo settimanale dei mezzi pubblici	Frequenze assolute
Mai	2800
Uno o due giorni a settimana	1608
Tre o quattro giorni a settimana	1024
Tutti i giorni	2568
Totale	8000

❒ Considerazioni preliminari.

La distribuzione presentata nell'esercizio è una serie ordinata di frequenze, perciò si tratta di quantificare una variabilità non metrica.
Quando le proprietà in esame sono rilevate a livello di scala ordinale, non è sufficiente limitarsi ad analizzate la mutabilità della distribuzione, perché ciò significa ignorare l'informazione contenuta nell'ordine delle categorie della variabile. Infatti, riteniamo meno dispersi i casi che si concentrano nelle modalità contigue piuttosto che, ad esempio, quelli appartenenti alle due categorie poste agli estremi della distribuzione.
Per risolvere l'esercizio è dunque necessario impiegare l'operatore di dispersione adatto a misurare la variabilità non metrica di una distribuzione, ovvero il D* di Gini.

& Svolgimento

I dati di cui disponiamo riguardano le frequenze assolute; quindi dobbiamo passare alle frequenze relative ($f_k = \frac{n_k}{N}$) perché la formula del D* di Gini è: $D^* = 2 \cdot \sum_{k=1}^{K-1} \left[f_k^{'} \cdot \left(1 - f_k^{'}\right) \right]$

dove $f_k^{'}$ e $\left(1 - f_k^{'}\right)$ sono rispettivamente una frequenza relativa cumulata ed il corrispondente complemento ad uno, k è una generica categoria, mentre K è il numero totale di categorie della variabile.
Per visualizzare i passaggi, aggiungiamo alla tabella dell'esercizio quattro nuove colonne, che andremo a riempire con i fattori utili a calcolare il valore dell'operatore.

Utilizzo settimanale dei mezzi pubblici	n	f	f'	$1 - f'$	$f' \cdot (1 - f')$
Mai	2800	0,350	0,350	0,650	0,2275
Uno o due giorni a settimana	1608	0,201	0,551	0,449	0,2474
Tre o quattro giorni a settimana	1024	0,128	0,679	0,321	0,2180
Tutti i giorni	2568	0,321	1	0	0
Totale	8000	1			0,6929

Completando la tabella abbiamo già la sommatoria delle frequenze relative cumulate moltiplicate per il corrispondente complemento a uno, quindi possiamo calcolare direttamente il D*, che è uguale a:

$$D^* = 2 \cdot \sum_{k=1}^{K-1} \left[f_k^{'} \cdot \left(1 - f_k^{'}\right) \right] = 2 \cdot 0{,}6929 = 1{,}3858$$

Per trarre delle conclusioni circa la variabilità della distribuzione è utile relativizzare il D* rispetto al range di valori che esso può assumere nel caso in esame. Il valore minimo del D* è zero, che significa assenza di variabilità, mentre il valore massimo è pari a $\frac{K-1}{2}$; la variabile dell'esercizio ha quattro modalità e quindi il valore massimo del D* è: $\frac{4-1}{2} = 1{,}5.$

Possiamo già notare che la variabilità della distribuzione osservata è notevole, poiché il valore dell'operatore è prossimo al suo massimo; procediamo a relativizzare il D*:

$d^* = D^* \cdot \frac{1}{\frac{k-1}{2}} = 1{,}3858 \cdot \frac{1}{1{,}5} = 1{,}3858 \cdot 0{,}6667 = 0{,}9239$ che possiamo approssimare a 0,924.

Poiché il d* varia fra zero e uno ($0 \leq d^* \geq 1$), ed il valore relativo osservato è 0,924, concludiamo affermando che la variabilità interna alla distribuzione è quasi massima.
In effetti, la maggioranza dei cittadini contattati si colloca nelle due categorie estreme della variabile ("Mai" e "Tutti i giorni"), e ciò aumenta la dispersione della distribuzione.

ESERCIZIO 5

In una ricerca svolta allo scopo di verificare gli effetti di un nuovo farmaco sulla memoria dei soggetti in età senile, si sono suddivisi 20 volontari in due gruppi, ciascuno di 10 individui scelti a caso, che chiameremo gruppo A e gruppo B. Nella sessione sperimentale, ai soggetti del gruppo A (gruppo sperimentale) è stato somministrato il farmaco, mentre quelli del gruppo B (gruppo di controllo) hanno ricevuto un placebo. In seguito, entrambi i gruppi sono stati sottoposti ad una prova di memoria, il cui punteggio è ricavato a partire dal numero di sillabe senza senso ricordate correttamente. I risultati dell'esperimento sono riportati nella tabella.
Determinare quale dei due gruppi ha reso la migliore performance, e quale ha avuto una prestazione più omogenea.

Gruppo A		**Gruppo B**	
Identificativo	Punteggio X_A	Identificativo	Punteggio X_B
1	10	1	10
2	7	2	5
3	9	3	3
4	8	4	6
5	6	5	7
6	10	6	8
7	7	7	5
8	6	8	9
9	5	9	4
10	8	10	7

❒ Considerazioni preliminari.

Per entrambi i gruppi il punteggio è ottenuto tramite una operazione di conteggio e pertanto la variabile dell'esercizio è cardinale; per confrontare le prestazioni dei due gruppi è possibile utilizzare gli operatori di tendenza centrale e di dispersione appropriati per tale tipo di variabili, ovvero media, varianza e deviazione standard. Inoltre, poiché nella scala utilizzata per il punteggio lo zero non è convenzionale, è possibile aggiungere ai precedenti operatori il coefficiente di variazione.
L'esercizio richiede sostanzialmente di descrivere le due distribuzioni, di effettuare un confronto ed argomentarne l'esito mediante delle considerazioni inerenti gli effetti che il farmaco ha avuto sul gruppo sperimentale.

& Svolgimento

Per iniziare riportiamo le formule che utilizzeremo per risolvere l'esercizio.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad S^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 \qquad S = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} \qquad CV = \frac{S}{\bar{x}}$$

Ricordiamo che la varianza (S^2) è una misura quadratica che non può dunque essere direttamente confrontata con la media. Per tale motivo si preferisce utilizzare la deviazione standard (S) che viceversa è espressa nella stessa unità di misura del suddetto operatore di tendenza centrale.
Per semplificare i passaggi calcoliamo i singoli costituenti delle formule.

Gruppo A			
Identificativo	Punteggio X_A	$x_i - \bar{x}$	$\left(x_i - \bar{x}\right)^2$
1	10	2,40	5,76
2	7	-0,60	0,36
3	9	1,40	1,96
4	8	0,40	0,16
5	6	-1,60	2,56
6	10	2,40	5,76
7	7	-0,60	0,36
8	6	-1,60	2,56
9	5	-2,60	6,76
10	8	0,40	0,16
Totale		0,00	26,40

$$\overline{x_A} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{10}\cdot 76 = 7{,}6$$

$$S_A^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 = \frac{1}{10}\cdot 26{,}40 = 2{,}64$$

$$S_A = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} = \sqrt{2{,}64} = 1{,}625$$

$$CV_A = \frac{S}{\bar{x}} = \frac{1{,}625}{7{,}6} = 0{,}214$$

Gruppo B			
Identificativo	Punteggio X_B	$x_i - \bar{x}$	$\left(x_i - \bar{x}\right)^2$
1	10	3,60	12,960
2	5	-1,40	1,960
3	3	-3,40	11,560
4	6	-0,40	0,160
5	7	0,60	0,360
6	8	1,60	2,560
7	5	-1,40	1,960
8	9	2,60	6,760
9	4	-2,40	5,760
10	7	0,60	0,360
Totale		0,00	44,400

$$\overline{x_B} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{10}\cdot 64 = 6{,}4$$

$$S_B^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 = \frac{1}{10}\cdot 44{,}40 = 4{,}44$$

$$S_B = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2} = \sqrt{4{,}44} = 2{,}107$$

$$CV_B = \frac{S}{\bar{x}} = \frac{2{,}107}{6{,}4} = 0{,}392$$

Osservando i risultati possiamo concludere che il gruppo sperimentale (A) non solo ha fornito una miglior prestazione, perché in media ha ricordato circa 8 sillabe contro le 6 ricordate dal gruppo B, ma è risultato anche più omogeneo: confrontando le deviazioni standard si osserva che la minor dispersione dei punteggi si realizza proprio nel gruppo sperimentale.
Il coefficiente di variazione offre risultati analoghi e, in questo caso, non porta informazioni aggiuntive rispetto alla deviazione standard: data la relazione di grandezza tra la media e la deviazione standard non pare che i due gruppi si discostino per un fattore di scala. Detto in altro modo, il gruppo caratterizzato dalla maggior dispersione è quello con la media più piccola.
Queste misure descrittive non permettono un giudizio sull'efficacia del farmaco generalizzabile a tutta la popolazione in età senile, tuttavia consentono almeno di congetturare che esso sembra aver moderatamente aumentato e uniformato l'abilità dei soggetti a cui è stato somministrato.

ESERCIZIO 6

I dati presentati in tabella si riferiscono a elezioni comunali e rappresentano il risultato dello spoglio delle schede complessivamente votate in una circoscrizione cittadina.
Rappresentare graficamente la distribuzione monovariata in esame servendosi delle frequenze assolute.

Schede	**n**
Schede bianche	19
Schede nulle	35
Voti validi	448
Totale	502

❐ Considerazioni preliminari.

La distribuzione presentata in tabella riguarda una variabile categoriale; poiché le modalità non sono ordinabili è opportuno rappresentarla tramite un grafico a torta oppure un grafico a barre.

& Svolgimento

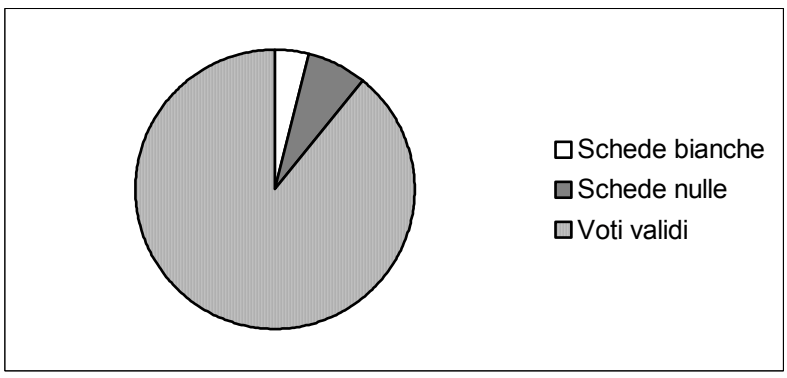
In un grafico a torta le "fette" corrispondono alle modalità ed hanno una ampiezza proporzionale alla frequenza della categoria a cui si riferiscono. Pertanto, partendo da una serie sconnessa di frequenze, è sufficiente determinare l'ampiezza degli angoli che definiscono i settori utilizzati per raffigurare la numerosità di ogni modalità della variabile.
Utilizzando le frequenze assolute, la formula che consente di calcolare l'ampiezza degli angoli per ciascuna delle K categorie è:

$\alpha_\kappa = \frac{360 \cdot n_k}{N}$

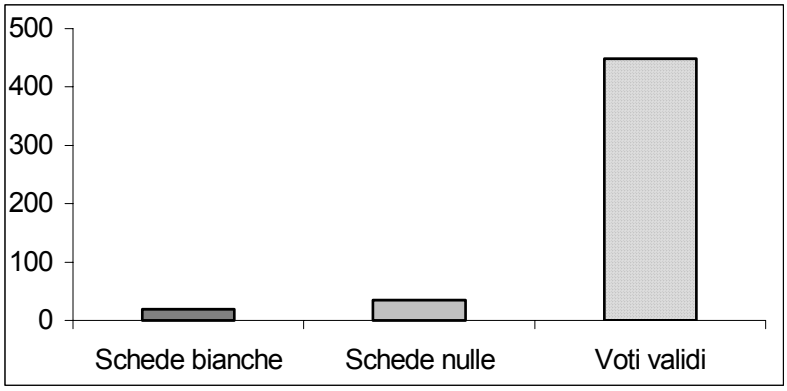
dove N ed n_k sono rispettivamente la numerosità dei casi rilevati (nel nostro caso, il totale delle schede spogliate) e la frequenza assoluta di una generica categoria appartenente alla variabile analizzata.
Svolgendo i calcoli per la distribuzione presentata in tabella abbiamo:

$\alpha_1 = \frac{360 \cdot 19}{502} = 13{,}625°$ $\alpha_2 = \frac{360 \cdot 35}{502} = 25{,}1°$ $\alpha_3 = \frac{360 \cdot 448}{502} = 321{,}275°$

A questo punto siamo in grado di disegnare il grafico a torta:



E' possibile rappresentare una serie sconnessa di frequenze anche tramite un grafico a barre.



Come si può notare, la base delle barre verticali è uguale per le tre categorie ed è convenzionale, cosi come l'ordine con cui vengono allineate sull'asse delle ascisse.