CORSO DI STATISTICA (parte 1) - ESERCITAZIONE 2

Dott.ssa Antonella Costanzo

a.costanzo@unicas.it

Indici di posizione variabilità e forma per caratteri qualitativi

Il seguente data set riporta la rilevazione di alcuni caratteri su un collettivo di 20 studenti.

Studente	Sesso	Età	Red	Istituto di provenienza	Voto al diploma	Statura (cm)	Colore occhi	Voto esame	Giud. sul corso
1	M	22	0,7	ITC	96	173	Nero	26	Pessimo
2	F	20	0,2	Liceo Classico	92	168	Marrone	26	Ottimo
3	F	30	1,6	Liceo Classico	90	165	Marrone	30	Buono
4	M	22	2,5	Liceo Scient	85	180	Nero	25	Buono
5	F	26	3,2	ITI	100	163	Azzurro	30	Pessimo
6	F	20	0,5	ITC	74	160	Nero	24	Pessimo
7	M	26	4,2	Liceo Scient	60	177	Marrone	20	Suff
8	M	30	1,3	ITC	76	164	Verde	18	Ottimo
9	F	27	1,2	Liceo Scient	100	158	Azzurro	29	Ottimo
10	F	25	1,7	ITI	95	170	Nero	25	Pessimo
11	F	25	1,9	ITI	85	167	Nero	25	Buono
12	M	22	0,7	ITC	97	159	Marrone	27	Buono
13	F	21	0,4	Liceo Classico	65	174	Azzurro	21	Ottimo
14	F	24	1,8	Liceo Scient	70	164	Verde	30	Suff
15	M	20	1,9	Liceo Scient	80	177	Nero	28	Suff
16	F	21	3,2	Liceo Classico	93	172	Nero	27	Pessimo
17	F	27	2,1	ITC	100	166	Marrone	26	Suff
18	F	22	0,1	ITI	84	160	Marrone	24	Buono
19	M	23	1,6	Liceo Scient	92	170	Azzurro	27	Ottimo
20	F	23	2,2	Liceo Scient	73	184	Verde	23	Buono

Promemoria

Caratteri qualitativi nominali	Posizione: Moda Variabilità: indice di eterogeneità di Gini
Caratteri qualitativi ordinali	Posizione: Moda, mediana, quantili Variabilità: indice di eterogeneità di Gini, indice di dispersione D Forma: indice di simmetria

Caratteri qualitativi nominali: Sesso, Istituto di provenienza, Colore degli occhi. Caratteri qualitativi ordinali: Giudizio sul corso

Quesiti:

1. Calcolare la moda, l'indice di eterogeneità di Gini per le variabili Colore degli Occhi e Istituto di Provenienza.
2. Calcolare moda, mediana, primo quartile e terzo quartile, 15-esimo percentile per la variabile Giudizio sul corso. Calcolare inoltre l'indice di eterogeneità di Gini, l'indice di Dispersione D, e l'indice di forma F

Soluzione Q.1

La **moda** di una distribuzione di frequenza è la modalità a cui è associata la massima frequenza assoluta o relativa. Corrisponde quindi al valore "più rappresentativo" della distribuzione, quello che si è verificato più spesso. Si calcola per caratteri qualitativi sconnessi, caratteri qualitativi ordinali e per caratteri quantitativi discreti e continui organizzati in classi (classe modale). In tal caso se le classi sono di ampiezza diversa si fa riferimento alla densità di frequenza e non alla frequenza di ciascuna classe.

Indici di posizione: moda

Colore degli occhi	n_i	f_i	p_i
Azzurro	4	0.20	20%
Marrone	6	0.30	30%
Nero	**7**	**0.35**	35%
Verde	3	0.15	15%
Totale (n)	20	1	100%

Moda: Nero

Istituto di Provenienza	n_i	f_i	p_i
ITC	5	0.25	25%
ITI	4	0.20	20%
Liceo Classico	4	0.20	20%
Liceo Scientifico	**7**	**0.35**	35%
Totale (n)	20	1	100%

Moda: Liceo Scientifico

Indici di variabilità (mutabilità): Indice di eterogeneità di Gini

La variabilità per caratteri qualitativi esprime essenzialmente il concetto di "diversità" tra le unità (mutabilità) ossia l'attitudine di un carattere ad assumere differenti modalità qualitative. Si possono avere due casi:

- **Assenza di variabilità** (massima omogeneità/minima eterogeneità): tutte le unità statistiche assumono la stessa modalità. Nella distribuzione di frequenza appare una sola modalità a cui è associata la massima frequenza.
- **Massima variabilità** (minima omogeneità/massima eterogeneità): nella distribuzione di frequenza appaiono tutte le k modalità, e ad ognuna di esse è associata la medesima frequenza (assoluta o relativa)

Ovviamente tutti i casi intermedi sono caratterizzati da un dato livello di **eterogeneità**: misura della variabilità delle frequenze relative delle k modalità associate al carattere.

Indice di eterogeneità di Gini:

$$E = 1 - \sum_i {f_i}^2$$

L'indice di eterogeneità di Gini per la variabile *Colore degli occhi* è definito come segue:

$$E = 1 - \sum_i {f_i}^2 = 1 - (0.2^2 + 0.30^2 + 0.35^2 + 0.15^2) = 1 - 0.275 = 0.725$$

In caso di minima variabilità l'indice di Gini assumerà valore E = 0

In caso di massima variabilità l'indice assume valore E = $1 - \frac{1}{\mathrm{k}}$ dove k = numero di modalità

Valore massimo dell'indice di Gini nel nostro caso sarà pari a: $E_{max} = 1 - \frac{1}{k} = 1 - \frac{1}{4} = 0.75$

Avendo definito il valore di E_{max} è possibile ricavare la versione normalizzata dell'indice di eterogeneità di Gini. In particolare:

$$E_{norm} = \frac{E}{E_{max}}$$

Nel nostro caso

$$E_{norm} = \frac{E}{E_{max}} = \frac{0.725}{0.75} = 0.97$$

La normalizzazione consente di interpretare più agevolmente l'indice poiché si definiscono un massimo e un minimo convenzionali entro cui l'indice può variare. In tal caso l'indice di Gini normalizzato varierà tra 0 (min) e 1 (max). In questo caso, un valore di E_{norm} prossimo a 1 indica che la distribuzione è molto eterogenea (variabile): tutte le modalità sono presente e con frequenze simili tra loro.

Per il carattere istituto di provenienza:

Istituto di Provenienza	n_i	f_i	p_i
ITC	5	0.25	25%
ITI	4	0.20	20%
Liceo Classico	4	0.20	20%
Liceo Scientifico	**7**	**0.35**	35%
Totale (n)	20	1	100%

Moda: Liceo Scientifico

L'indice di eterogeneità di Gini è il seguente:

$E = 1 - \sum_i f_i^2 = 1 - (0.25^2 + 0.20^2 + 0.25^2 + 0.35^2) = 1 - 0.288 = 0.712$

La versione normalizzata dell'indice di eterogeneità di Gini sarà data da: $E_{norm} = \frac{E}{E_{max}}$

Nel nostro caso:

$$E_{norm} = \frac{E}{E_{max}} = \frac{0.712}{0.75} = 0.95$$

Anche qui il valore di E_{norm} è prossimo a 1 per cui la distribuzione è molto eterogenea (frequenze equilibrate tra loro)

Soluzione Q.2

La **mediana** è la modalità dell' unità statistica che occupa il posto centrale nella distribuzione ordinata delle osservazioni. Si calcola per variabili qualitative ordinali e per variabili quantitative discrete e continue organizzate in classi.

Per dati grezzi bisogna anzitutto distinguere i due casi n pari ed n dispari.

- Nel caso di n pari la mediana è la modalità che occupa la posizione $(\frac{n}{2}, \frac{n}{2} + 1)$
- Nel caso di n dispari la mediana è la modalità che occupa la posizione $\frac{n+1}{2}$

Per dati organizzati in tabelle di frequenza (come nel nostro caso), la mediana è quel valore x_i tale che:

- $N(x_i) = \frac{n+1}{2}$ se uso le frequenze assolute cumulate
- $F(x_i) = 0.50$ se uso le frequenze relative cumulate
- $P(x_i) = 50$ se uso le frequenze cumulate percentuali

Giudizio sul corso	n_i	f_i	N_i	F_i	P_i	RF_i
Pessimo	5	0.25	5	0.25	25	1
Sufficiente	4	0.2	9	0.45	45	0.75
Buono	6	0.3	15	0.75	75	0.55
Ottimo	5	0.25	20	1	100	0.25
Totale (n)	20	1				

Mediana: utilizzando le frequenze cumulate relative la mediana è quella modalità tale che: $F(x_i) = 0.5$ quindi in questo caso Me=Buono

Allo stesso modo, per dati organizzati in tabelle di frequenza, si procede per il calcolo dei quartili. I quartili sono particolari percentili della distribuzione e la dividono in 4 parti uguali; sono indici di posizione calcolabili per caratteri qualitativi ordinali, per caratteri quantitativi discreti e quantitativi continui organizzati in classi.

- Il primo quartile Q1 è la modalità x_i tale che $N(x_i) = \frac{n}{4}$, $F(x_i) = 0.25$, $P(x_i) = 25$ ragionando rispettivamente con le frequenze assolute cumulate, relative cumulate e percentuali cumulate
- Il secondo quartile Q2=Me, è la mediana
- Il terzo quartile Q3 è la modalità tale che $N(x_i) = \frac{3n}{4}$, $F(x_i) = 0.75$, $P(x_i) = 75$ ragionando rispettivamente con le frequenze assolute cumulate, relative cumulate e percentuali cumulate

I decili dividono la distribuzione in 10 parti uguali, i percentili dividono la distribuzione in 100 parti uguali. Nel nostro caso sappiamo che il generico percentile $p_{x_i} = \frac{i\,n}{100} = \frac{15*1}{100} = 0.15$. Il 15-esimo percentile è quindi, utilizzando le F_i la modalità $x_i: F(x_i) = 0.15$ =Pessimo.

Giudizio sul corso	n_i	f_i	N_i	F_i	P_i	RF_i
Pessimo	5	0.25	5	0.25	25	1
Sufficiente	4	0.2	9	0.45	45	0.75
Buono	6	0.3	15	0.75	75	0.55
Ottimo	5	0.25	20	1	100	0.25
Totale (n)	20	1				

Primo quartile: Pessimo

15-esimo percentile: Pessimo

Terzo quartile= Buono

Indice di eterogeneità di Gini per la variabile *Giudizio sul corso*:

$E = 1 - \sum_i f_i^2 = 1 - (0.25^2 + 0.20^2 + 0.30^2 + 0.25^2) = 1 - 0.255 = 0.745$

$E_{norm} = \frac{E}{E_{max}} = \frac{0.745}{0.75} = 0.99$ prossimo alla massima eterogeneità

Nota: Per i caratteri qualitativi ordinali non solo posso verificare quanto sono "diverse" tra loro le unità (eterogeneità) ma in più, sfruttando l'ordinamento delle modalità, posso studiare la disposizione delle unità (dispersione).

Indice di dispersione D:

L'indice D per il calcolo della dispersione in variabili qualitative ordinali si basa sulle frequenze cumulate F_i e retro cumulate RF_i. A differenza dell'indice di eterogeneità di Gini consente di tener conto anche della relazione d'ordine che sussiste tra le modalità delle variabili.

$$D = \sum_{i=1}^{n} [F_i(1 - F_i) + RF_i(1 - RF_i)]$$

Per agevolare i calcoli costruiamo la seguente tabella:

Giudizio sul corso	n_i	f_i	F_i	$1-F_i$	RF_i	$1-RF_i$	$F_i(1-F_i)$	$RF_i(1-RF_i)$
Pessimo	5	0.25	0.25	0.75	1	0	0.1875	0
Sufficiente	4	0.2	0.45	0.55	0.75	0.25	0.2475	0.1875
Buono	6	0.3	0.75	0.25	0.55	0.45	0.1875	0.2475
Ottimo	5	0.25	1	0	0.25	0.75	0	0.1875
Totale (N)	20	**1**						

$$D = 0.1875 + 0.435 + 0.435 + 0.1875 = 1.245$$

NB. Se si cambia l'ordine delle modalità della distribuzione del carattere l'indice di eterogeneità di Gini resta invariato mentre l'indice di dispersione D cambierà in quanto tiene conto dell'ordine delle modalità.

Il valore minimo dell'indice D è 0 (assenza di dispersione), mentre il suo valore massimo (nel caso di distribuzioni caratterizzate ma massima dispersione) dipende dalla numerosità delle osservazioni. In particolare, si hanno due casi:

1. **n pari,** $D^{max} = \frac{k-1}{2}$ (per normalizzare l'indice è ok utilizzare questo valore, indipendentemente dalla numerosità di n!)
2. **n dispari,** $D^{max} = \frac{k-1}{2}(1 - \frac{1}{n^2})$

Nel nostro caso, n=20 k=4 quindi $D^{max} = \frac{k-1}{2} = \frac{4-1}{2} = 1.5$. La versione normalizzata dell'indice D sarà quindi:

$$D^{norm} = \frac{D}{D^{max}} = \frac{1.245}{1.5} = 0.83$$

Nota utile:

E' possibile calcolare l'indice D (non normalizzato) senza far ricorso al calcolo delle frequenze retrocumulate, sfruttando la seguente equivalenza:

$$D = \sum_{i=1}^{n} [F_i(1 - F_i) + RF_i(1 - RF_i)] = 2\sum_{i=1}^{k-1} F_i(1 - F_i)$$

Nel nostro caso avremo:

$$D = 2[0.25(1 - 0.25) + 0.45(1 - 0.45) + 0.75(1 - 0.75)] = 1.25$$

$$D^{norm} = \frac{D}{D^{max}} = \frac{1.245}{1.5} = 0.83$$

L'indice D può essere sfruttato per studiare la forma della distribuzione di un carattere qualitativo ordinale. Data la distribuzione ordinata del carattere Giudizio sul corso calcoliamo l'**indice di asimmetria F**:

1. divido la distribuzione in due parti; n=20 pari quindi prendo le prime $\frac{n}{2}$ e le ultime $\frac{n}{2}$ osservazioni

Giudizio sul corso	$\boldsymbol{n_i}$	$\boldsymbol{f_i}$	$\boldsymbol{F_i}$
Pessimo	5	0.25	0.25
Sufficiente	4	0.2	0.45
Buono	6	0.3	0.75
Ottimo	5	0.25	1
Totale (n)	20	**1**	

2. prima metà della distribuzione (n=10)

Giudizio sul corso	$\boldsymbol{n_i}$	$\boldsymbol{f_i}$	$\boldsymbol{F_i}$	$\boldsymbol{1 - F_i}$	$\boldsymbol{F_i(1 - F_i)}$
Pessimo	5	0.5	0.5	0.5	0.25
Sufficiente	4	0.4	0.9	0.1	0.09
Buono	1	0.1	1	0	0
Totale (n)	10	1			

l'indice D per le prime $\frac{n}{2}$ osservazioni, $D_s = 2[0.25 + 0.09] = 0.68$

3. seconda metà della distribuzione (n=10)

Giudizio sul corso	$\boldsymbol{n_i}$	$\boldsymbol{f_i}$	$\boldsymbol{F_i}$	$\boldsymbol{1 - F_i}$	$\boldsymbol{F_i(1 - F_i)}$
Buono	5	0.5	0.5	0.5	0.25
Ottimo	5	0.5	1	0	0
Totale (n)	10	1			

l'indice D per le seconde $\frac{n}{2}$ osservazioni, $D_d = 2[0.25] = 0.5$

4. Confronto: $D_s > D_d$ segnale di asimmetria negativa (sinistra) e calcolo indice F

$$F = D_d - D_s = 0.5 - 0.68 = -0.18$$

5. Indice F normalizzato $-1 \leq F \leq 1$

$$F_{norm} = \frac{D_d - D_s}{D_d + D_s} = \frac{-0.18}{1.18} = -0.15$$

Esercizio 2.

Per il carattere Voto all'esame calcolare la moda, la mediana, i quartili e il 12-esimo percentile

Voto all'esame	n_i	f_i	p_i	N_i	F_i	P_i
18	1	0.05	5%	1	0.05	5
20	1	0.05	5%	2	0.1	10
21	1	0.05	5%	3	0.15	15
23	1	0.05	5%	4	0.20	20
24	2	0.1	10%	6	0.30	30
25	4	0.2	20%	10	0.50	50
26	2	0.1	10%	12	0.60	60
27	3	0.15	15%	15	0.75	75
28	1	0.05	5%	16	0.80	80
29	1	0.05	5%	17	0.85	85
30	3	0.15	15%	20	1	100
Totale (n)	20	1	100%			

Moda=25

Mediana e quantili:

Essendo *Voto all'esame* un carattere quantitativo discreto si può ragionare in termini di dati grezzi o (più comodamente) siccome i dati sono organizzati in una tabella di frequenza calcolare gli indici di posizione sfruttando le informazioni sulle frequenze cumulate (assolute, relative o percentuali). Ricordiamo che per caratteri quantitativi discreti la mediana è quel valore x tale che:

- Se n è pari: $Me = \frac{\left(X_{\frac{n}{2}} + X_{\frac{n}{2}+1}\right)}{2}$

- Se n è dispari: $Me = X_{\frac{n+1}{2}}$

NOTA OPERATIVA: indipendentemente dalla numerosità delle osservazioni (pari o dispari) è sempre possibile individuare la posizione e il valore della mediana con la regola utilizzata nel caso di n dispari e cioè $Me = X_{\frac{n+1}{2}}$

Per gli altri quantili (e in particolare per Q1 e Q3) le formule "canoniche" sono:

$$Q_1 = \frac{X_{\frac{n}{4}} + X_{\frac{n}{4}+1}}{2}$$

$$Q_3 = \frac{X_{\frac{3n}{4}} + X_{\frac{3n}{4}+1}}{2}$$

Per il generico quantile q_i che divide la distribuzione in k parti uguali

$$q_i = x_{(\frac{i\,n}{K})}$$

Se si lavora su tabelle di frequenze il calcolo della mediana e dei quantili sfrutta le informazioni derivanti dalle frequenze cumulate assolute, relative o percentuali.

Per la variabile Voto all'esame procediamo dunque in entrambi i modi:

1. dati grezzi

Distribuzione ordinata dei voti all'esame:

1°	18
2°	20
3°	21
4°	23
5°	24
6°	24
7°	25
8°	25
9°	25
10°	25
11°	26
12°	26
13°	27
14°	27
15°	27
16°	28
17°	29
18°	30
19°	30
20°	30

Formule canoniche:

$$Q_1 = \frac{X_{\frac{n}{4}} + X_{\frac{n}{4}+1}}{2} = \frac{24+24}{2} = 24$$

$$Q_2 = Me = X_{\frac{n+1}{2}} = \frac{25+26}{2} = 25.5$$ (vedi nota operativa)

$$Q_3 = \frac{X_{\frac{3n}{4}} + X_{\frac{3n}{4}+1}}{2} = \frac{27+28}{2} = 27.5$$

Formule "alternative" (utili anche per i calcolo di altri percentili):

Data la distribuzione ordinata delle modalità del carattere X=$\{ x_{(1)}, x_{(2)} \dots x_{(k-1)}, x_{(k)} \}$:

- Individuo la posizione del quantile desiderato nella distribuzione ordinata di X, che corrisponderà sostanzialmente ad un *numero,* che indichiamo con Pos.

- Definisco una funzione $Dec(\text{Pos})$ che individua la parte decimale del *numero* che esprime la posizione del quantile desiderato risultante dall'operazione precedente.
- Il valore del quantile di una distribuzione può essere definito come la somma pesata dei valori corrispondenti alle posizioni del quantile desiderato. I pesi sono funzione della distanza del valore del quantile rispetto alla posizione individuata. Si può utilizzare la seguente regola:

$$[1 - Dec(\text{Pos})] * x_{(ARR.DIFETTO)} + Dec(\text{Pos}) * x_{(ARR.ECCESSO)}$$

Calcolo terzo quartile sfruttando questo metodo alternativo:

- $posizione\ Q_3 = \frac{3n+1}{4} = \frac{21}{4} = 15.25$

- Il valore del primo quartile si trova tra la quindicesima e la sedicesima posizione della distribuzione ordinata della variabile Voto all'esame, precisamente alla posizione 15.25 che risulta essere "più prossima" alla quinta posizione $x_{(15)} = 27$ che non alla sesta posizione $x_{(16)} = 28$

- Applico la funzione $Dec(\text{Pos})$, individuo quindi la parte decimale della posizione individuata al punto 1. In tal caso la parte decimale è 0.25 (la parte intera è 5).

- Valore del primo quartile $(1 - 0.25) * x_{(15)} + 0.25 * x_{(16)}$=(0.75)*27+(0.25)*28=27.25

Il calcolo del primo quartile con questo procedimento è lasciato per esercizio.

12-esimo percentile:

- $posizione\ Q_{0.12} = \frac{(12)*20}{100} = 2.4$
- Il valore del 12-esimo percentile si trova tra la seconda e la terza posizione della distribuzione ordinata della variabile Voto all'esame, precisamente alla posizione 2.4 che risulta essere "più prossima" alla seconda posizione $x_{(2)} = 20$ che non alla terza posizione $x_{(3)} = 21$
- Applico la funzione $Dec(\text{Pos})$, individuo quindi la parte decimale della posizione individuata al punto 1. In tal caso la parte decimale è 0.4 (la parte intera è 2).
- Valore del 12-esimo percentile $(1 - 0.4) * x_{(2)} + 0.4 * x_{(3)}$=(0.6)*20+(0.4)*21=20.4