CORSO DI STATISTICA (parte 1) - ESERCITAZIONE 3

Dott.ssa Antonella Costanzo

a.costanzo@unicas.it

Esercizio 1. Sintesi a cinque e misure di variabilità rispetto ad un centro

Una catena di fast-food ha selezionato un campione di 9 ristoranti al fine di valutare la possibilità di aprire un nuovo ristorante. Sono state raccolte sul campione esaminato le seguenti informazioni:

POSTI numero di posti a sedere

INCASSI incassi giornalieri (Euro)

LOC location non centrale? (SI =1 / NO = 0)

	POSTI	INCASSI	LOC
1	75	1.83	1
2	60	2.17	1
3	120	4.35	0
4	150	6.15	0
5	65	1.96	1
6	90	5.85	0
7	70	1.75	1
8	80	2.12	0
9	55	1.27	1
Total	**765**	**27.45**	

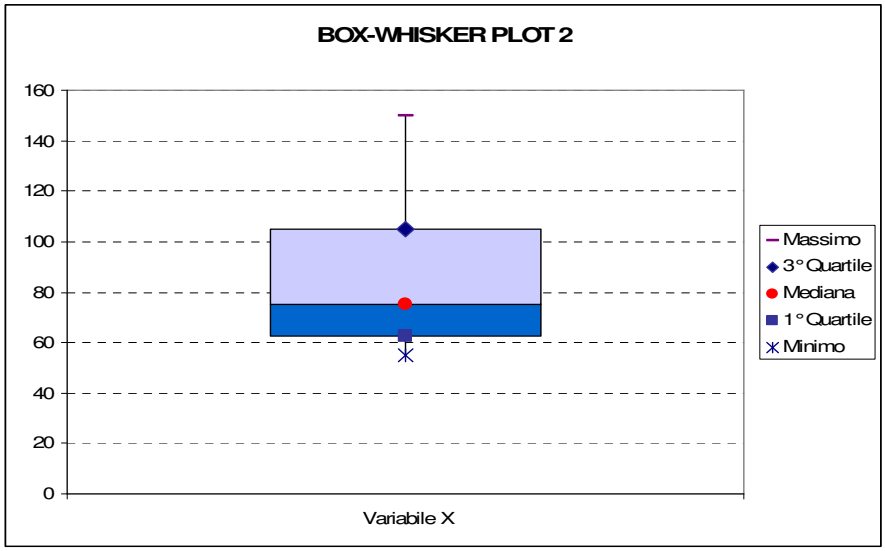
a) Costruire il box plot per la variabile POSTI A SEDERE.

b) Con riferimento alla variabile INCASSI GIORNALIERI organizzata in classi (1.26, 2], (2.1, 4.5], [4.5, 7] calcolare: lo scostamento semplice medio dalla media e lo scostamento semplice medio dalla mediana, lo scostamento quadratico medio dalla media (scarto quadratico medio), la devianza e la varianza.

Sol.

a)Il grafico a scatola (box-plot) è una particolare rappresentazione di una distribuzione. E'ottenuto a partire da 5 numeri di sintesi: minimo, 1° quartile (Q1), mediana, 3° quartile (Q3), massimo.

Il box plot o diagramma a scatola e baffi si ottiene riportando su un asse verticale (oppure orizzontale) i 5 numeri di sintesi. La scatola del box plot ha come estremi inferiore e superiore rispettivamente Q1 e Q3. La differenza tra Q3 e Q1 costituisce il campo di variazione interquartile, indicato con CVI=Q3-Q1. La mediana divide la scatola in due parti. I baffi si ottengono congiungendo Q1 al minimo e Q3 al massimo.



Variable X	Index Values
Minimum (minimo)	55
1° Quartile	62.5
Median (mediana)	75
3° Quartile	105
Maximum (massimo)	150

Soglie e valori anomali:

I valori anomali (distanti rispetto a tutti gli altri valori che caratterizzano la distribuzione) vengono determinati dal confronto con il campo di variazione interquartile. In particolare vengono considerate due soglie:

- il valore al di sotto del quale una modalità viene considerata outlier:

  $$Q_1 - 1.5(Q_3 - Q_1) = 62.5 - 1.5(42.5) = 62.5 - 63.75 = -1.25$$

- il valore al di sopra del quale una modalità viene considerata outlier:

  $$Q_3 + 1.5(Q_3 - Q_1) = 105 + 1.5(42.5) = 168.75$$

Nel nostro caso, non si riscontra la presenza di outliers.

Ulteriori considerazioni:

Confrontando tra loro le lunghezze dei due baffi (che rappresentano le distanze tra Q1 e il minimo e tra Q3 e il massimo) e le altezze dei due rettangoli che costituiscono la scatola (che rappresentano le distanze tra Q1 e mediana e tra mediana e Q3) si ottengono informazioni sulla simmetria della distribuzione: questa è tanto più simmetrica quanto le lunghezze dei baffi risultano simili tra loro e le altezze dei due rettangoli risultano simili tra loro. Nel nostro caso la variabile POSTI A SEDERE presenta una coda più allungata a destra per cui risulta lievemente asimmetrica positiva: la distanza tra la mediana e Q3 è maggiore della distanza della mediana rispetto a Q1).

b) Misure di variabilità (rispetto ad un "centro")

X=Incassi	c_i	n_i	fi	Fi	$c_i - \overline{x}$	$\lvert ci - \overline{x} \rvert * n_i$	$(c_i - \overline{x})^2$	$(c_i - \overline{x})^2 * n_i$	$\lvert c_i - Me \rvert * n_i$
(1.26, 2]	1.63	4	0.45	0.45	-1.47	5.88	2.1609	8.6436	3.32
(2.1, 4.5]	3.3	3	0.33	0.78	0.20	0.6	0.04	0.12	2.52
[4.5, 7]	5.75	2	0.22	1	2.65	5.3	7.0225	14.045	6.58
Totale n		9	1			11.78		22.8086	12.42

$\bar{x} = 3.10$

Me=2.46

Scostamento semplice medio dalla media: $S_{\bar{x}} = \frac{\sum_{i=1}^{n} |c_i - \overline{x}| * n_i}{n} = \frac{11.78}{9} = 1.308$

Tanto più piccolo è lo scarto semplice medio tanto più i valori si addensano attorno alla media aritmetica.

Scostamento semplice medio dalla mediana: $S_{Me} = \frac{\sum_{i=1}^{n} |c_i - Me| * n_i}{n} = \frac{12.42}{9} = 1.38$

Scostamento quadratico medio dalla media (o scarto quadratico medio):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (c_i - \bar{x})^2 n_i}{n}} = \sqrt{\frac{22.8086}{9}} = 1.59$$

Varianza: $\sigma^2 = \frac{22.8086}{9} = 2.53$ la varianza è un indice assoluto che risente dell'unità di misura della variabile.

Devianza: numeratore della varianza=22.8086

Si può dimostrare inoltre che la varianza può essere calcolata alternativamente come: $\sigma_x^2 = \frac{\sum_{i=1}^{n} x_i^2 n_i}{n} - \bar{x}^2$ ossia la varianza è pari alla media aritmetica dei quadrati meno il quadrato della media aritmetica. Nel nostro caso siccome si hanno dati in classi: $\sigma_x^2 = \frac{\sum_{i=1}^{n} c_i^2 n_i}{n} - \bar{x}^2$

La varianza assume valore minimo (zero) quando tutte le modalità sono uguali tra loro e aumenta all'aumentare della differenza tra i valori osservati. Il massimo può essere infinito perché gli scarti possono essere infinitamente grandi (ovvero le modalità infinitamente lontane dalla media aritmetica).

Esercizio 2. Confronti in termini di variabilità

In un piccolo paesino, ci sono soltanto due banche. Dei depositi dei clienti di due banche si conosce la distribuzione per classi:

Depositi bancari (migliaia di euro)	**Banca 1**	**Banca 2**
[0, 20)	47	91
[20, 50)	79	137
[50, 100)	111	205
[100, 200)	64	129
[200, 500)	31	73

a) Si calcoli la varianza della variabile depositi dei clienti nelle due banche.

b) Si confronti la variabilità dei depositi bancari nelle due banche.

Sol.

a)

Depositi Banca 1	c_i	n_i	$c_i n_i$	c_i^2	$c_i^2 n_i$
[0, 20)	10	47	470	100	4700
[20, 50)	35	79	2765	1225	96775
[50, 100)	75	111	8325	5625	624375
[100, 200)	150	64	9600	22500	1440000
[200, 500)	350	31	10850	122500	3797500
Totale		332	32010	151950	5963350

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} c_i n_i = \frac{1}{332}32010 = 96.42$$

$$\bar{x}_1^2 = 96.42^2 = 9295.98$$

$$\sigma_1^2 = \frac{1}{n}\sum_{i=1}^{n} c_i^2 n_i - \bar{x}_1^2 = \frac{1}{332}(5963350) - 9295.98 = 8665.92$$

Depositi Banca 2	c_i	n_i	$c_i n_i$	c_i^2	$c_i^2 n_i$
[0, 20)	10	91	910	100	9100
[20, 50)	35	137	4795	1225	167825
[50, 100)	75	205	15375	5625	1153125
[100, 200)	150	129	19350	22500	2902500
[200, 500)	350	73	25550	122500	8942500
Totale		635	65980	151950	13175050

$$\bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n} c_i n_i = \frac{1}{635}65980 = 103.91$$

$$\bar{x}_2^2 = 103.91^2 = 10796.36$$

$$\sigma_2^2 = \frac{1}{n}\sum_{i=1}^{n} c_i^2 n_i - \bar{x}_2^2 = \frac{1}{635}(13175050) - 10796.36 = 9951.76$$

b) Per confrontare la variabilità di un carattere in due gruppi di unità statistiche non è opportuno ricorrere alla varianza, il cui valore dipende, oltre che dalla variabilità, dall'ordine di grandezza del carattere X considerato (si dice, quindi, che la varianza è una misura dimensionata). Per questo motivo, è preferibile confrontare i coefficienti di variazione, che standardizzano gli scarti quadratici medi rispetto al diverso livello medio del fenomeno. Nel nostro caso, essendo i valori assunti dalla nostra variabile tutti positivi, possiamo calcolare i coefficienti di variazione:

Coefficiente di variazione Depositi bancari 1 :

$$CV_1 = \frac{\sigma_1}{\bar{x}_1} = \frac{\sqrt{8665.92}}{96.42} = 0.97$$

Coefficiente di variazione Depositi bancari 2:

$$CV_2 = \frac{\sigma_2}{\bar{x}_2} = \frac{\sqrt{9951.76}}{103.91} = 0.92$$

Il coefficiente di variazione è un indice relativo, indipendente dall'unità di misura e dall'ordine di grandezza della variabile. Ha il minimo uguale a zero e il massimo non definito, giacché varia al variare del tipo di distribuzione. *Note utili*: se la media, in valore assoluto, risulta prossima a zero (per effetto di parziali compensazioni fra valori positivi e negativi), il CV può segnalare, in maniera errata, una variabilità molto elevata del fenomeno.

Dal confronto tra i coefficienti di variazione relativi ai due gruppi di unità statistiche, concludiamo che la variabilità dei depositi è maggiore nella Banca 1, contrariamente a quello che avrebbe suggerito il confronto tra le due varianze.

Altri indici relativi di variabilità (utili per il confronto tra distribuzioni):

Indici percentuali di variabilità o di variabilità relativa alla media

Siccome per gli indici di variabilità non si conosce il massimo che possono assumere è conveniente dividere il valore dell'indice per il corrispondente indice di posizione scelto per misurare la dispersione. Ad esempio: la versione relativa dello scostamento semplice medio dalla media e dello scostamento semplice medio dalla mediana si ottiene dividendo tali indici per la corrispondente media o la corrispondente mediana rispettivamente. Per cui avremo:

DEPOSITI BANCA 1

Scostamento semplice medio dalla media relativo:

$$S_{xBanca1}^{rel} = \frac{S_{\bar{x}_1}}{\bar{x}_1} = \frac{75.311}{103.91} = 0.72$$

Scostamento semplice medio dalla mediana relativo:

Me=71.88

$$S_{MeBanca1}^{rel} = \frac{S_{Me}}{Me} = \frac{65.68}{71.88} = 0.91$$

DEPOSITI BANCA 2: il calcolo dello scostamento semplice medio dalla media relativo e lo scostamento semplice medio dalla mediana relativo è lasciato per esercizio. Si commentino inoltre i risultati ottenuti.