Sintesi della distribuzione di un carattere: la variabilità

Obiettivi dell'unità didattica

- Calcolare i principali indici di variabilità
- Giudicare la rappresentatività della media aritmetica
- Confrontare la variabilità di differenti distribuzioni
- Introdurre il concetto di concentrazione
- Standardizzare le informazioni
- Leggere la variabilità attraverso i grafici

In questa unità didattica verranno trattati i seguenti argomenti:

- **La variabilità di una distribuzione**
- **La varianza**
- **Gli indici relativi di variabilità**
- **La concentrazione**
- **La standardizzazione**
- **Il grafico box-plot**

Premessa e contenuti

Nella Unit precedente abbiamo visto come sia possibile riassumere tutte le informazioni disponibili attraverso l'utilizzo di una misura sintetica (come la media aritmetica o la mediana).

Generalmente, tuttavia, i fenomeni (sociali, economici, ecc...) che vengono studiati tendono a variare (nel tempo, nello spazio, tra individui, ecc...), elemento che giustifica l'esistenza stessa delle metodologie quantitative (se il reddito disponibile fosse lo stesso per tutte le famiglie, che senso avrebbe studiare la distribuzione dei redditi all'interno di una comunità?!); pertanto, sarà necessario comprendere in che modo tale variabilità si manifesta all'interno dell'analisi che stiamo facendo.

Questo assume un significato ancor maggiore alla luce del fatto che, come abbiamo visto in precedenza, quando sintetizziamo una distribuzione con una sola informazione quantitativa guadagniamo qualcosa in termini di capacità informativa, ma perdiamo qualcos'altro rispetto all'informazione disaggregata. Sarà dunque necessario capire se la sintesi effettuata è "buona", in altri termini se l'informazione ottenuta riassume bene i dati di partenza, altrimenti possono risultare errate le conclusioni alle quali perveniamo (proprio perché la misura utilizzata non rappresenta bene i dati disponibili).

In questa Unità didattica, inoltre, approfondiremo un concetto, quello della concentrazione, che è, in un certo senso, complementare a quello della variabilità. Infatti, studiare come varia un fenomeno significa anche cercare di comprendere se questo è sostanzialmente equidistribuito tra le unità statistiche considerate o se, al contrario, è concentrato nelle mani di pochi soggetti (o imprese).

Quindi, sulla base di alcuni degli indici che abbiamo incontrato nelle precedenti lezioni, impareremo in che modo trasformare i dati di partenza così da avere delle informazioni "standardizzate", e, dunque, confrontabili tra loro.

Infine, cercheremo di valutare la variabilità utilizzando una forma grafica particolare, il grafico box-plot, che ci consente di confrontare le caratteristiche di base di differenti distribuzioni da un punto di vista "visivo" e, quindi, come già abbiamo detto in precedenza, più intuibile ed immediato.

CAMPO DI VARIAZIONE

rappresenta la differenza tra la più grande e la più piccola modalità osservate all'interno della distribuzione; è una misura assoluta di variabilità; risente molto della presenza di "valori anomali" all'interno della distribuzione.

DIFFERENZA INTERQUARTILE

è data dalla differenza tra il terzo e il primo quartile della distribuzione; è una misura di variabilità assoluta; è preferibile rispetto al campo di variazione in quanto risente meno della presenza di "valori anomali".

VARIABILE SCARTO

è data dalla differenza tra una singola modalità e la media aritmetica: $(x_i - \bar{x})$.

DEVIANZA

è data dalla somma degli scarti elevati al quadrato: $\sum_{i=1}^{n} (x_i - \bar{x})^2$; in sostanza, si tratta del numeratore della varianza.

VARIANZA

è la principale misura di variabilità che abbiamo studiato. Viene calcolata come rapporto tra la sommatoria di tutti gli scarti elevati al quadrato (ossia, la devianza) e la numerosità totale:

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n}$$

SCARTO QUADRATICO MEDIO

(o DEVIAZIONE STANDARD) è dato dalla radice quadrata della varianza: $\sigma = \sqrt{\sigma^2}$; viene calcolato per riportare i dati all'unità di misura originaria.

MASSIMA VARIABILITA' (σmax)

è il massimo valore teorico raggiungibile da σ in una determinata distribuzione. È dato dal prodotto tra la media aritmetica e la radice quadrata di (n-1). Ci è utile per giudicare la rappresentatività della media aritmetica.

COEFFICIENTE DI VARIAZIONE

è dato dal rapporto tra lo s.q.m. e la media aritmetica, moltiplicato a 100: $CV = \frac{\sigma}{x} 100$; viene calcolato per confrontare la variabilità di differenti distribuzioni. Più è elevato, maggiore è la variabilità della distribuzione.

CARATTERE TRASFERIBILE un carattere è detto trasferibile quando può essere ceduto, in tutto o in parte, da un'unità statistica ad un'altra

CARATTERE EQUIDISTRIBUITO quando tutte le unità statistiche possiedono lo stesso ammontare del carattere

CONCENTRAZIONE un carattere (trasferibile) si dice concentrato quando una gran parte dell'ammontare complessivo è nelle "mani" di pochi soggetti (al limite, di uno solo)

MINIMA CONCENTRAZIONE quando tutte le unità statistiche possiedono lo stesso ammontare del carattere (ossia, il carattere è EQUIDISTRIBUITO)

MASSIMA CONCENTRAZIONE quando l'ammontare complessivo del carattere è detenuto da una sola unità statistica

RAPPORTO DI CONCENTRAZIONE DI GINI (R)

È l'indice che misura la concentrazione di un carattere; è dato dalla formula:

$$R = \frac{\sum_{i=1}^{n-1} (F_i - Q_i)}{\sum_{i=1}^{n-1} F_i} ;$$

tale indice può variare tra 0 (nel caso di minima concentrazione) ed 1 (nel caso di massima concentrazione).

STANDARDIZZAZIONE

È una trasformazione lineare dei dati che rende le informazioni tratte da diverse distribuzioni tutte confrontabili. La formula è la seguente:

$$Z = \frac{x_i - \bar{x}}{\sigma}$$

e l'elemento fondamentale è che una distribuzione (qualsiasi essa sia) una volta che i dati sono stati standardizzati si trova ad avere media pari a zero e varianza pari a uno.

GRAFICO BOX PLOT

È una rappresentazione grafica che mostra da un punto di vista visivo tutte le caratteristiche fondamentali di una distribuzione (minimo, massimo, valore centrale e variabilità)

Sintesi della distribuzione di un carattere:

la variabilità (prima parte)

Introduzione alla variabilità

- Nelle lezioni precedenti abbiamo visto come sia possibile riassumere tutte le informazioni disponibili attraverso l'utilizzo di una misura sintetica (come la media aritmetica o la mediana).

- I fenomeni (sociali, economici, ecc...) che vengono studiati tendono a variare. Tale elemento giustifica l'esistenza stessa delle metodologie quantitative; pertanto, sarà necessario comprendere in che modo tale variabilità si manifesta all'interno dell'analisi che stiamo effettuando.

Quando sintetizziamo una distribuzione con una sola informazione quantitativa guadagniamo qualcosa in termini di capacità informativa, ma perdiamo qualcos'altro rispetto all'informazione disaggregata. Sarà dunque necessario capire se la sintesi effettuata è "buona", in altri termini se l'informazione ottenuta riassume bene i dati di partenza, altrimenti possono risultare errate le conclusioni alle quali perveniamo (proprio perché la misura utilizzata non rappresenta bene i dati disponibili)

A)	40	50	60	$\overline{X}=50$
B)	50	50	50	$\overline{X}=50$
C)	0	0	150	$\overline{X}=50$

Distribuzioni con media aritmetica uguale ma differente variabilità

$$R = x_n - x_1$$

A)	40	50	60
B)	50	50	50
C)	0	0	150

A) $R_A = X_n - X_1 = 60 - 40 = 20$
B) $R_B = X_n - X_1 = 50 - 50 = 0$
C) $R_C = X_n - X_1 = 150 - 0 = 150$

Difetti: Ha un minimo, ma non ha un massimo definito; è difficilmente interpretabile nel caso di *outliers*

Alcune considerazioni

Le misure appena presentate:

- Prendono in considerazione solo due soggetti
- Non hanno estremi definiti
- Sono spesso di difficile interpretazione
- Sono misure assolute rendono difficile il confronto

La varianza

La **VARIANZA** è data dalla sommatoria di tutti gli scarti tra le singole modalità e la media aritmetica, elevate al quadrato e rapportata alla numerosità totale:

$$Var = \sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n}$$

X	X^2
1	1
2	4
3	9
5	25

$$s.q.m. = \sigma = \sqrt{\sigma^2}$$

Esempio 1 – Varianza e S.Q.M.

In un'azienda lavorano 5 dipendenti, per i quali l'azienda stessa sostiene un costo annuo pari, rispettivamente, a:

25.000€ 37.000€ 19.000€ 21.000€ 28.000€

Si calcoli:

A) il costo medio annuo per dipendente

B) la varianza

C) lo scarto quadratico medio

A) (25.000+37.000+19.000+21.000+28.000)/5=**26.000**

B)

$$\text{Var} = \sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n}$$

[(25.000-26.000)² + (37.000-26.000)² + (19.000-26.000)² + + (21.000-26.000)² + (28.000-26.000)²]/5 =

= [(-1.000)² + (+11.000)² + (-7.000)² + (-5.000)² + (+2.000)²]/5 =

= (1.000.000 + 121.000.000 + 49.000.000 + 25.000.000 + + 4.000.000)/5 = (200.000.000 / 5) = 40.000.000

C)

$$\sigma = \sqrt{\sigma^2} = \sqrt{40.000.000} = 6.324,5$$

Esempio 2 –Varianza e S.Q.M.

x_i	n_i	$x_i \cdot n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 * n_i$
1	11	11	-2,33	5,43	59,73
2	21	42	-1,33	1,77	37,17
3	32	96	-0,33	0,11	3,52
4	28	112	0,67	0,45	12,60
5	16	80	1,67	2,79	44,64
6	7	42	2,67	7,13	49,91
	115	383			207,57

$\bar{X}$= 383/115 = 3,33

MAI NEGATIVA

$$\text{Var} = \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot n_i}{n} = \frac{207,57}{115} = 1,80$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{1,80} = 1,34$$

Esempio 3 –Varianza e S.Q.M.

x_i	n_i	c_i	$c_i \cdot n_i$
1-5	45	3	135,0
6-9	35	7,5	262,5
10-19	26	14,5	377,0
20-29	20	24,5	490,0
30-39	15	34,5	517,5
40-49	10	44,5	445,0
	151		2.227,0

a) $\bar{X}$ = 2.227 / 151 = 14,75

b) $R = X_n - X_1 = 49 - 1 = 48$

Calcolare:

a) La media aritmetica

b) Il campo di variazione

c) La varianza e lo scarto quadratico medio

x_i	n_i	c_i	$c_i \cdot n_i$	$(c_i - \bar{x})$	$(c_i - \bar{x})^2$	$(c_i - \bar{x})^2 * n_i$
1-5	45	3	135,0	-11,75	138,06	6.212,70
6-9	35	7,5	262,5	-7,25	52,56	1.839,60
10-19	26	14,5	377,0	-0,25	0,06	1,56
20-29	20	24,5	490,0	9,75	95,06	1.901,20
30-39	15	34,5	517,5	19,75	390,06	5.850,90
40-49	10	44,5	445,0	29,75	885,06	8.850,60
	151		2.227,0			24.656,56

c)

$$\text{Var} = \sigma^2 = \frac{\sum_{i=1}^{n} (c_i - \bar{x})^2 \cdot n_i}{n} = \frac{24.656,56}{151} = 163,29$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{163,29} = 12,78$$

Indici di variabilità relativi

- Giudicare la rappresentatività della media aritmetica
  Devo capire se la media che ho calcolato sintetizza bene le informazioni

- Confrontare la variabilità di due distribuzioni
  Varianza e scarto quadratico medio dipendono dall'unità di misura e dal "livello";
  non posso fare confronti

Rappresentatività della media

Distribuzione dei clienti di un'agenzia di viaggi per numero di viaggi effettuati nell'ultimo anno

x_i	n_i
1	8
2	5
3	6
4	1
	20

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i \cdot n_i}{n} = \frac{40}{20} = 2$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \cdot n_i}{n}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{0,9} = 0,95$$

x_i	n_i	$x_i \cdot n_i$	$(x_i - \bar{x})^2 * n_i$
2	20	40	$(2 - 2)^2 \times 40 = 0$
	20	40	0

MINIMA VARIABILITA'

$$\sigma = \sqrt{\sigma^2} = \sqrt{0} = 0$$

Rappresentatività della media -2

x_i	n_i	$x_i \cdot n_i$	$(x_i - \bar{x})^2 * n_i$
0	19	0	$(0 - 2)^2$ x 19= 76
40	1	40	$(40 - 2)^2$ x 1= 1444
	20	40	1.520

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \cdot n_i}{n}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{76} = 8{,}72$$

MASSIMA VARIABILITA'

$$\sigma_{max} = \bar{x}\sqrt{(n-1)} = 2 \cdot \sqrt{(20-1)} = 2 \cdot \sqrt{19} = 2 \cdot 4{,}36 = 8{,}72$$

Rappresentatività della media -3

$$\frac{\sigma}{\sigma_{max}}$$

+ si avvicina a 0
+ Media rappresentativa

+ si avvicina a 1
+ Media NON rappresentativa

- **Nel nostro caso:**

$$\frac{\sigma}{\sigma_{max}} = \frac{0{,}95}{8{,}72} = 0{,}11$$

La Media rappresenta bene i dati

Confrontare la variabilità 1.

- **Confrontare la variabilità di due distribuzioni**
  Varianza e scarto quadratico medio dipendono dall'unità di misura e dal "livello"; non posso fare confronti

- **Esempio: confronto tra peso adulti e neonati**

78,4	65,8	72,2	85,6	75,2	58,7	69,9
3,875	2,954	3,458	4,512	2,722	3,158	

2.

Adulti

$$\bar{x} = \frac{(78{,}4 + 65{,}8 + 72{,}2 + 85{,}6 + 75{,}2 + 58{,}7 + 69{,}9)}{7} = 72{,}26$$

$$\sigma = \sqrt{\frac{\left[\begin{array}{l}(78{,}4 - 72{,}26)^2 + (65{,}8 - 72{,}26)^2 + (72{,}2 - 72{,}26)^2 + (85{,}6 - 72{,}26)^2 + \\ (75{,}2 - 72{,}26)^2 + (58{,}7 - 72{,}26)^2 + (69{,}9 - 72{,}26)^2\end{array}\right]}{7}} = 8{,}066$$

Neonati

$$\bar{x} = \frac{(3{,}875 + 2{,}954 + 3{,}458, +4{,}512 + 2{,}722 + 3{,}158)}{6} = 3{,}447$$

$$\sigma = \sqrt{\frac{\left[\begin{array}{l}(3{,}875 - 3{,}447)^2 + (2{,}954 - 3{,}447)^2 + (3{,}458 - 3{,}447)^2 + \\ (4{,}512 - 3{,}447)^2 + (2{,}722 - 3{,}447)^2 + (3{,}158 - 3{,}447)^2\end{array}\right]}{6}} = 0{,}601$$

3.

$$CV = \frac{\sigma}{\bar{x}} \cdot 100$$

Coefficiente di Variazione

Più è grande CV, maggiore è la variabilità

$$CV_{adu} = \frac{\sigma}{\bar{x}} \cdot 100 = \frac{8{,}066}{72{,}26} \cdot 100 = 0{,}112 \cdot 100 = 11{,}2$$

$$CV_{neo} = \frac{\sigma}{\bar{x}} \cdot 100 = \frac{0{,}601}{3{,}447} \cdot 100 = 0{,}174 \cdot 100 = 17{,}4$$

$CV_{neo} > CV_{adu}$ - La variabilità è maggiore nei neonati

Variabilità –competenze acquisite

Cosa abbiamo imparato?

- Cos'è la variabilità
- Misurare la variabilità
- Valutare la "bontà" della media aritmetica
- Confrontare la variabilità di differenti distribuzioni

Quali strumenti usare?

- Per misurare la variabilità → Varianza
- Per la rappresentatività della media → σ_{max}
- Per confrontare differenti distribuzioni → CV