STATISTICA 1 - Dispersione

Riccardo Corradin, Andrea Gilardi

Introduzione

- Nelle scorse slides abbiamo studiato indici e formule per caratterizzare la **posizione** di una distribuzione di frequenze.
- Un'altra domanda naturale che possiamo chiederci è quanto sia **dispersa** tale distribuzione.
- Per rispondere a tale domanda, introdurremo nelle prossime slides degli indici atti a misurare la **variabilità di un carattere** o la sua **concentrazione**.

Per dispersione di un carattere si intende la **tendenza** dello stesso **ad assumere modalità differenti**.

Per esempio, consideriamo le due sequenze

$$A, A, B, A, A, A, A, C, A, A,$$

e

$$A, B, B, A, B, A, C, C, C, A.$$

Anche se assumono gli stessi valori, possiamo notare come la prima sequenza sia più concentrata sulla modalità A, mentre la seconda sia più dispersa su diverse modalità.

Variabilità

Introduzione

- Due concetti fondamentali sono la variabilità e l'assenza di variabilità in un carattere.

Per **variabilità** di un carattere si intende la **tendenza** dello stesso **ad assumere modalità differenti**.

Per **assenza di variabilità** intendiamo una situazione dove tutte le osservazioni assumono la medesima modalità.

Un esempio è dato dalla sequenza

$$1, 1, 1, 1, 1, 1, 1, 1, 1, 1$$

in cui tutte le osservazioni che abbiamo presentano la modalità 1.

- Il nostro obiettivo sarà individuare degli indici in grado di quantificare la **variabilità** di un certo carattere, misurando l'allontanamento dalla situazione di **assenza di variabilità**.

Introduzione

Requisiti per un indice di variabilità:

- L'indice deve essere pari a 0 se e solo se c'è assenza di variabilità.
- L'indice deve essere maggiore di 0 in presenza di variabilità.

Dividiamo gli indici di variabilità in due categorie fondamentali.

- **Assoluti**, che vengono espressi nella medesima unità di misura del carattere, e la cui interpretazione deve tenere conto della scala di riferimento.
  - → Intervalli di variazione.
  - → Scostamenti medi da un valore medio.
  - → Differenze medie.

- **Relativi**, numeri puri, che non dipendono dalla scala di misurazione del carattere di riferimento.
  - → Coefficiente di variazione.

Intervalli di variazione

Supponiamo di avere un insieme di osservazioni $\{x_1, \ldots, x_n\}$. Utilizziamo due quantità principalmente per vedere quanto ampiamente varia un carattere rispetto al suo supporto.

- **Campo di variazione** o **range**, ovvero

$$R = x_{(N)} - x_{(1)},$$

dove $x_{(1)} = \min\{x_1, \ldots, x_N\}$ è il più piccolo valore osservato e $x_{(N)} = \max\{x_1, \ldots, x_N\}$ è il più grande valore osservato.

- **Distanza interquartile**, ovvero

$$Q_3 - Q_1,$$

Dove Q_1 è il quantile $q_{0.25}$ di livello 0.25 e Q_3 è il quantile $q_{0.75}$ di livello 0.75.

$\rightarrow$ La distanza interquartile è più adeguata in presenza di **valori anomali**.

Intervalli di variazione - Esempio

Example

I seguenti dati rappresentano i punteggi ottenuti da due squadre, A e B, durante 6 manche di un quiz di cultura generale.

Manche	**1**	**2**	**3**	**4**	**5**	**6**	**Totale**
Squadra A	66	66	66	66	66	70	400
Squadra B	50	62	68	74	80	66	400

Si richiede di:

1. Calcolare il **punteggio medio** per ciascuna squadra;
2. Determinare il **range** dei punteggi per ciascuna squadra **commentando il risultato ottenuto**;

Intervalli di variazione - Esempio

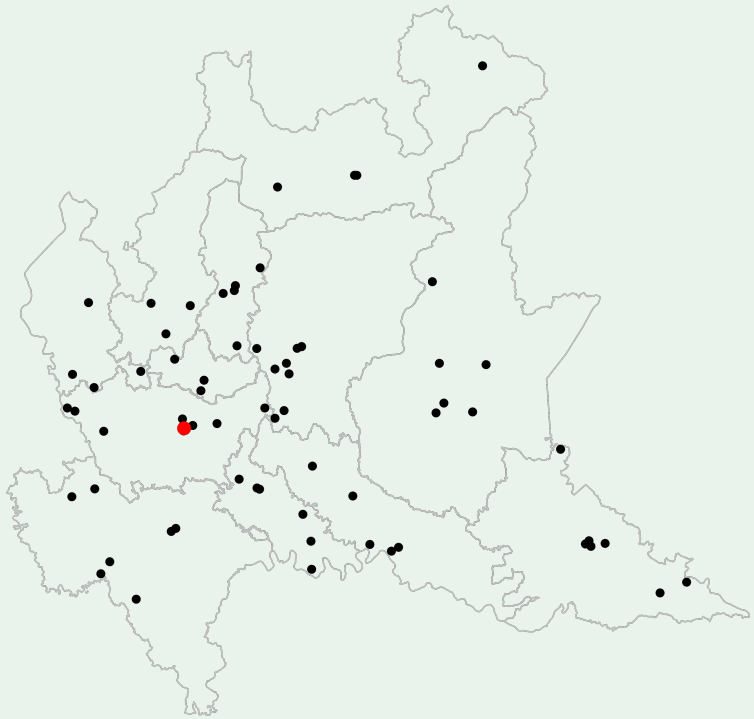
Intervalli di variazione - Esempio

Example

Dal sito di ARPA Lombardia è possibile scaricare la serie storica delle rilevazioni del PM_{10} per un panel di stazioni fisse sparse per tutta la regione.

La seguente tabella riassume le rilevazioni (medie giornaliere dal 2019-01-01 al 2020-12-31) di PM_{10} per una centraline posizionata a Milano in Via Senato.

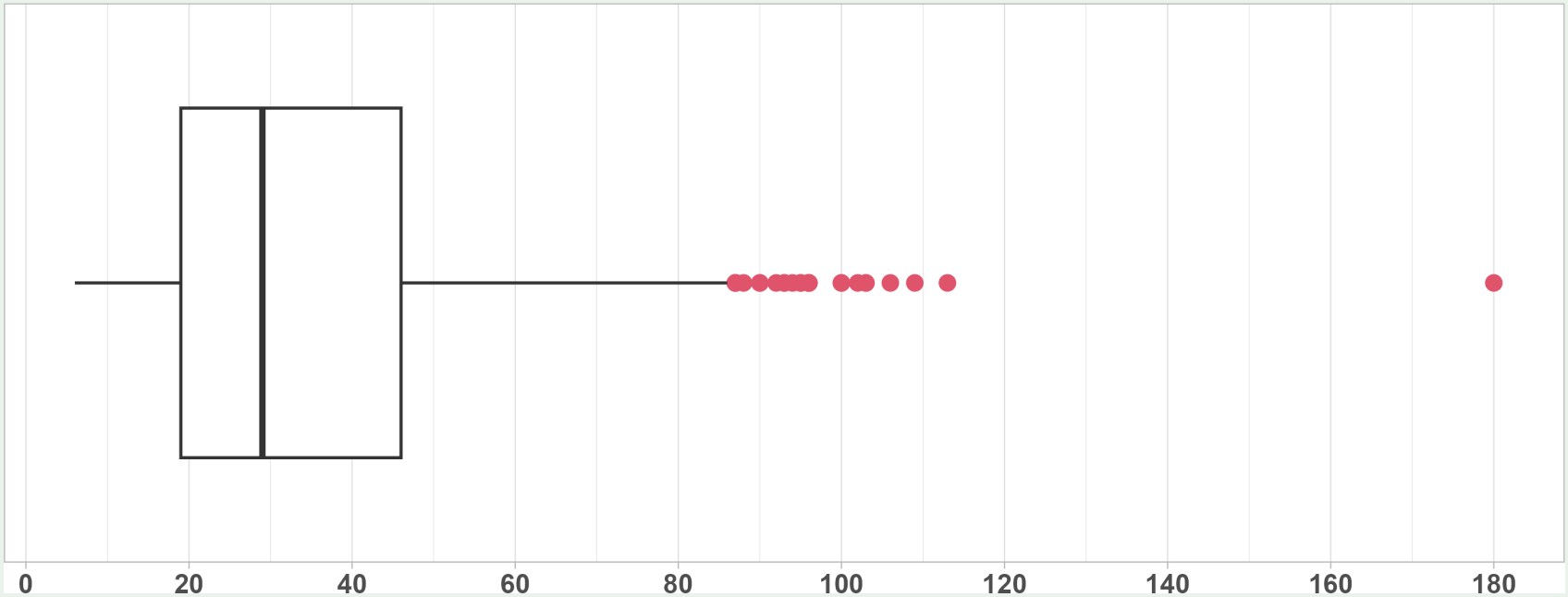
Tempo	Valore PM_{10} $[\mu g/m^3]$
2019-01-01	95
2019-01-02	24
2019-01-03	23
$\vdots$	$\vdots$
t	x_t
$\vdots$	$\vdots$
2020-12-31	55



Intervalli di variazione - Esempio

Example

Il seguente boxplot riassume la distribuzione del PM_{10} nel periodo sotto esame.



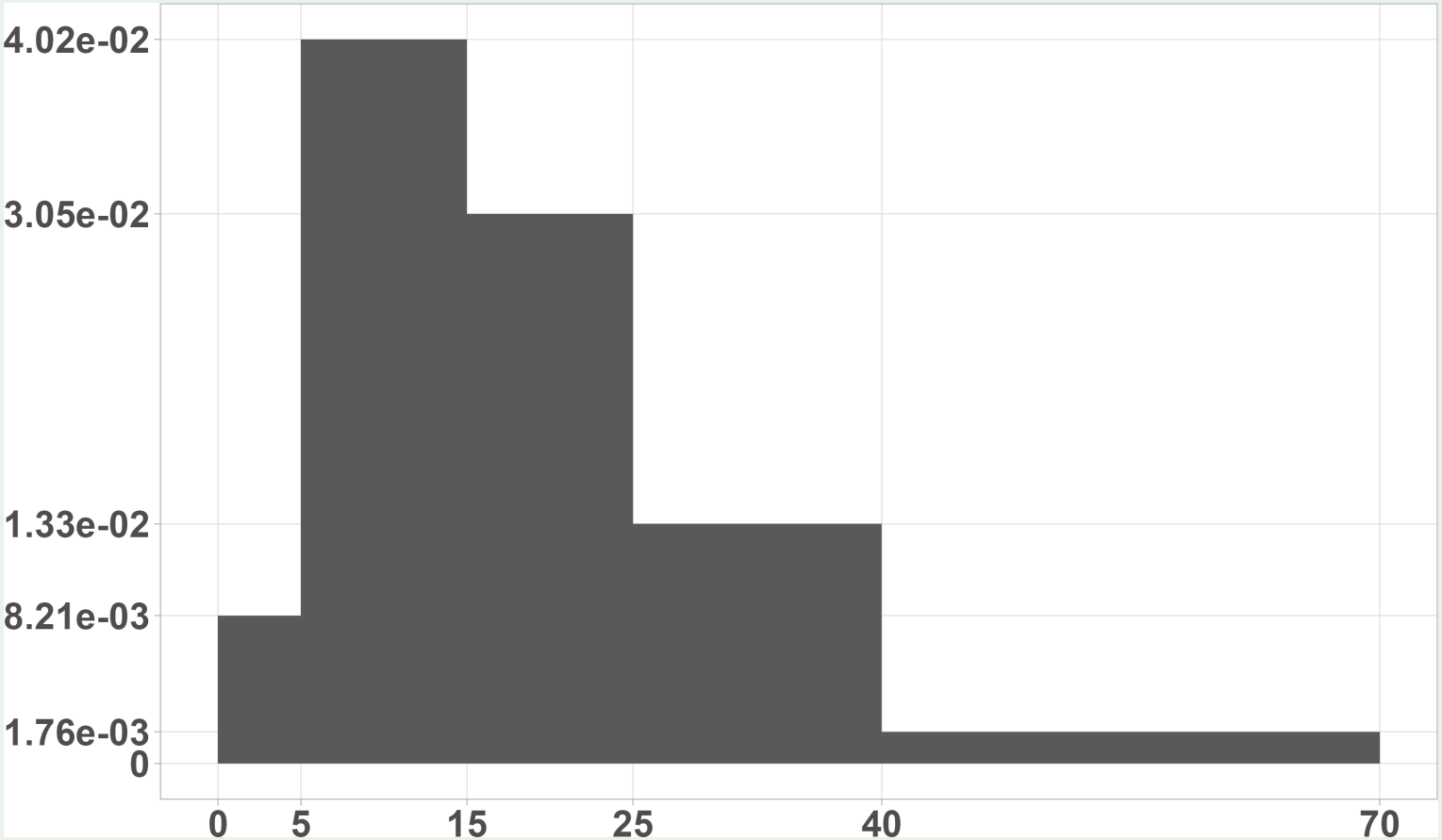
- Si calcoli (approssimativamente) il **range** e la **distanza interquartile** di tale distribuzione **commentando il risultato ottenuto**.
- Quale fra le due statistiche di sintesi vi sembra più appropriata? Perchè?

Intervalli di variazione - Esempio

Intervalli di variazione - Esempio

Example

Il seguente istogramma riassume le rilevazioni orarie del PM_{10} nel periodo 2019-01-01 / 2019-12-31 di una centralina ARPA localizzata nel comune di Sondrio.



Intervalli di variazione - Esempio

Example

Richieste:

1. Partendo dalle informazioni racchiuse nell'istogramma, si ricostruisca la tabella di **frequenze (divise in classi)**;
2. Si calcoli la **media** e la **mediana** dei valori di PM10 **commentando i risultati ottenuti**;
3. Si confronti il **campo di variazione** e la **distanza interquartile**.

Intervalli di variazione - Esempio

Example

Soluzioni:

1. Dall'istogramma possiamo ricavare (approssimativamente) sia i valori estremi di ciascuna classe che le frequenze relative specifiche:

Classi	Val. Centrale	Ampiezza	Freq. Rel. Spec	Freq. Rel.
$00 \dashv 05$	2.5	5	0.00821	0.0411
$05 \dashv 15$	10	10	0.0402	0.402
$15 \dashv 25$	20	10	0.0305	0.305
$25 \dashv 40$	32.5	15	0.0133	0.1995
$40 \dashv 70$	55	30	0.00176	0.0528

2. La **media** è pari a

$$2.5 \cdot 0.0411 + 10 \cdot 0.402 + 20 \cdot 0.305 + 32.5 \cdot 0.1995 + 55 \cdot 0.0528 = 19.6105$$

La frequenze relative cumulate sono

$$[0.0411; 0.4431; 0.7481; 0.9476; 1]$$

Intervalli di variazione - Esempio

Example

Le frequenze relative retrocumulate sono pari a

$$[1; 0.9593; 0.5573; 0.2523; 0.0528]$$

La **classe mediana** è quindi $15 \dashv 25$ e la **mediana** interpolata è pari a

$$\text{Me} = 15 + (0.5 - 0.4431) \cdot \frac{1}{0.0305} \simeq 16.866$$

Osserviamo che la mediana è leggermente minore della media e tale fenomeno è legato alla presenza di alcuni valori molto maggiori dal blocco centrale dei dati.

3. Il **range** è pari a $70 - 0 = 70$. Per calcolare la **distanza interquartile** dobbiamo ricavare Q1 e Q3. Q1 giace nella classe $5 \dashv 15$ e il suo valore interpolato è pari a

$$\text{Q1} = 5 + (0.25 - 0.0411) \cdot \frac{1}{0.0402} \simeq 10.197.$$

Intervalli di variazione - Esempio

Example

Q3 giace nella classe $25 \dashv 40$ e il valore interpolato è pari a

$$\text{Q3} = 25 + (0.75 - 0.7481) \cdot \frac{1}{0.0133} \simeq 25.143.$$

La **distanza interquartile** è pari a $25.143 - 10.197 = 14.946$. Troviamo nuovamente indicazione che la distribuzione presenta dei valori "anomali" (nel senso di lontani dalla massa centrale dei dati) poiché il range è molto maggiore della distanza interquartile.

Scostamenti medi da un valore medio

Gli scostamenti vengono calcolati **solo per caratteri quantitativi**, discreti o continui.

Rappresentano una funzione di **quanto differiscono** le singole **osservazioni** da un **valore medio** osservato.

Non ci interessa il segno, o la direzione, dello scostamento. Invece, ci interessa **l'intensità** di ogni scostamento, il suo valore assoluto.

Supponiamo di avere un insieme di osservazioni $\{x_1, \ldots, x_N\}$. Gli scostamenti delle osservazioni da un valore medio M sono definiti come

$$|x_i - M| \geq 0, \qquad i = 1, \ldots, N.$$

- $\rightarrow$ $|x_i - M| = 0$ per ogni $i = 1, \ldots, N$ se e solo se siamo in una situazione di assenza di variabilità.
- $\rightarrow$ $|x_i - M| > 0$ per almeno un valore $i = 1, \ldots, N$ se siamo in presenza di variabilità.

Scostamenti medi da un valore medio

- **Scostamento medio dalla media aritmetica**, ovvero quanto differiscono mediamente le osservazioni dalla loro media aritmetica. In formule

$$S_{M_1} = \frac{1}{N}\sum_{i=1}^{N}|x_i - M_1|.$$

  - S_{M_1} è sulla stessa scala di misura delle osservazioni.

- **Deviazione standard** o **scarto quadratico medio**, rappresenta quanto si discostano le osservazioni dalla loro media aritmetica, in media quadratica. In formule

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|x_i - M_1|^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - M_1)^2}$$

  - σ è sulla stessa scala di misura delle osservazioni.

- **Scostamento medio dalla mediana**. In formule

$$S_{Me} = \frac{1}{N}\sum_{i=1}^{N}|x_i - Me|.$$

  - S_{Me} è sulla stessa scala di misura delle osservazioni.

Scostamenti medi da un valore medio - Esercizio

Example

Il sabato mattina Andrea va al mercato di Sesto San Giovanni per acquistare frutta e verdura dalla sua commerciante preferita. Egli, durante l'ultima spedizione, ha acquistato 7 mele il cui peso (in g) è dato da

$$\{450, 550, 350, 375, 220, 650, 510\}.$$

Calcolare S_{M_1}, σ, e S_{Me} per il carattere *"Peso delle mele acquistate da Andrea"*.

Scostamenti medi da un valore medio - Esercizio

Varianza

- Lo scostamento più comunemente utilizzato è la **varianza**. Rappresenta la media aritmetica degli scostamenti quadratici. In formule

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - M_1)^2.$$

Corrisponde al quadrato della deviazione standard.

Procedimento indiretto per il calcolo della varianza

La varianza può essere scritta come la media aritmetica del quadrato delle osservazioni meno il quadrato della media aritmetica delle osservazioni,

$$\sigma^2(X) = M_1\big((X - M_1(X))^2\big) = M_1(X^2) - \big(M_1(X)\big)^2.$$

In formule operative abbiamo che

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - M_1)^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - M_1^2.$$

Dimostrazione del procedimento indiretto per il calcolo di σ^2

Abbiamo

$$\begin{aligned}
\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i - M_1)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i^2 - 2x_i M_1 + M_1^2) \\
&= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\sum_{i=1}^{N} 2x_i M_1 + \frac{1}{N}\sum_{i=1}^{N} M_1^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - 2M_1\frac{1}{N}\sum_{i=1}^{N} x_i + \frac{1}{N}\sum_{i=1}^{N} M_1^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - 2M_1^2 + M_1^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - M_1^2,
\end{aligned}$$

che conclude la dimostrazione.

L'espressione precedente rende molto più immediato il calcolo della varianza.

Proprietà della varianza

1) La varianza di una trasformazione lineare di un carattere X

$$Y = a + bX, \qquad a, b \in \mathbb{R},$$

può essere espressa come

$$\sigma^2(Y) = b^2\sigma^2(X).$$

Dimostrazione

$$\begin{aligned}
\sigma^2(Y) &= \frac{1}{N}\sum_{i=1}^{N}(y_i - M_1(Y))^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}(a + bx_i - a - bM_1(X))^2 = \frac{1}{N}\sum_{i=1}^{N}(bx_i - bM_1(X))^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} b^2(x_i - M_1(X))^2 = b^2\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - M_1(X))^2\right] \\
&= b^2\sigma^2(X),
\end{aligned}$$

che conclude la dimostrazione.

Varianza

2) Supponiamo di avere un insieme di N unità statistiche $\{x_1, \ldots, x_N\}$ suddivise in k gruppi, con numerosità $N_1, \ldots, N_k$, dove

$$\sum_{j=1}^{k} N_j = N.$$

Definiamo

- $x_{i,j}$ la i-esima osservazione del j-esimo gruppo.
- $\bar{x}_j$ come media aritmetica di X nel j-esimo gruppo.
- σ_j^2 la varianza di X nel j-esimo gruppo.
- $\bar{x}$ la media aritmetica di X nell'intera popolazione.

Possiamo scrivere la varianza di X calcolata sull'intera popolazione come

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{k} N_j\sigma_j^2 + \frac{1}{N}\sum_{j=1}^{k} N_j(\bar{x}_j - \bar{x})^2,$$

dove

- Il primo termine è la media delle varianze parziali, chiamato varianza nei gruppi.
- Il secondo termine è la varianza delle medie parziali, chiamato varianza fra gruppi.

Varianza

Dimostrazione: partendo dalla definizione di varianza, abbiamo

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{N_j}(x_{i,j} - \bar{x})^2 = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{N_j}(x_{i,j} - \bar{x}_j + \bar{x}_j - \bar{x})^2$$

$$= \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{N_j}\left[(x_{i,j} - \bar{x}_j)^2 + 2(x_{i,j} - \bar{x}_j)(\bar{x}_j - \bar{x}) + (\bar{x}_j - \bar{x})^2\right]$$

$$= \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{N_j}(x_{i,j} - \bar{x}_j)^2 + \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{N_j}2(x_{i,j} - \bar{x}_j)(\bar{x}_j - \bar{x}) + \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{N_j}(\bar{x}_j - \bar{x})^2$$

$$= \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{N_j}(x_{i,j} - \bar{x}_j)^2 + 2\frac{1}{N}\sum_{j=1}^{k}(\bar{x}_j - \bar{x})\sum_{i=1}^{N_j}(x_{i,j} - \bar{x}_j) + \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{N_j}(\bar{x}_j - \bar{x})^2$$

$$= \frac{1}{N}\sum_{j=1}^{k}\frac{N_j}{N_j}\sum_{i=1}^{N_j}(x_{i,j} - \bar{x}_j)^2 + \frac{1}{N}\sum_{j=1}^{k}N_j(\bar{x}_j - \bar{x})^2$$

$$= \frac{1}{N}\sum_{j=1}^{k}N_j\sigma_j^2 + \frac{1}{N}\sum_{j=1}^{k}N_j(\bar{x}_j - \bar{x})^2,$$

che conclude la dimostrazione.

Varianza - Esercizi

Example

Un veterinario sta studiando la variabilità del peso dei cani appartenenti a due razze diverse: Labrador e Beagle. Ha raccolto i seguenti dati (in kg) per un campione di cani:

$$\textbf{Labrador} : 30, 32, 35, 31$$

$$\textbf{Beagle} : 12, 14, 11, 15$$

1. Calcolare la **varianza totale** del peso dei cani;
2. Calcolare la **varianza delle medie parziali dei pesi fra i gruppi**;
3. Calcolare la **varianza del peso nei gruppi**;
4. Dimostrare empiricamente la formula appena vista.

Varianza - Esercizi

Varianza - Esercizi

Varianza - Esercizi

Varianza - Esercizi

Varianza - Esercizi

Example

Un medico sta studiando la **pressione sanguigna sistolica** in tre gruppi di pazienti affetti da diabete di tipo 2. I tre gruppi sono: A) pazienti sotto trattamento farmacologico; B) pazienti che seguono una dieta controllata; C) pazienti senza nessun trattamento.

Le misurazioni della pressione dei tre gruppi sono le seguenti:

A - Farmaci : $120, 125, 130$

B - Dieta : $135, 140, 138$

C - Nessun trattamento : $150, 155, 160$

Si mostri empiricamente la validità della formula di scomposizione della varianza per il problema in esame.

Varianza - Esercizi

Differenze medie

Un'altra quantità frequentemente utilizzata sono le **differenze medie**. In pratica, guardiamo quanto **differiscono mediamente** tra loro le osservazioni che abbiamo, considerando tutte le **coppie possibili** che possiamo formare. In formule abbiamo

$$\Delta = \frac{1}{N(N-1)} \sum_{\substack{i=1 \\ i \neq j}}^{N} \sum_{j=1}^{N} |x_i - x_j|.$$

La precedente è chiamata **differenza media semplice**.

Possiamo definire la formula precedente come segue.

1) Costruiamo tutte le coppie possibili di osservazioni tra gli N valori che abbiamo
$$(x_i, x_j), \qquad i = 1, \ldots, N, \quad j = 1, \ldots, N,$$
dove $i \neq j$, quindi abbiamo $N(N-1)$ coppie totali.
2) Associo a ciascuna coppia il valore corrispondente della differenza assoluta
$$|x_i - x_j|, \qquad i = 1, \ldots, N, \quad j = 1, \ldots, N,$$
che quantifica la distanza tra i valori osservati di x_i e x_j. Abbiamo quindi $N(N-1)$ distanze calcolate tra tutte le possibili coppie di valori.

Differenze medie

3) Calcoliamo la media aritmetica delle $N(N-1)$ distanze osservate, come

$$\Delta = \frac{\text{somma delle } N(N-1) \text{ distanze}}{\text{numero di distanze}} = \frac{1}{N(N-1)} \sum_{\substack{i=1 \\ i \neq j}}^{N} \sum_{j=1}^{N} |x_i - x_j|.$$

Esiste una versione alternativa della precedente, dove consideriamo anche la presenza delle coppie formate dalla stessa osservazione ripetuta, chiamata **Δ di Gini con ripetizione**.

$$\Delta_R = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j| = \frac{N(N-1)}{N^2} \Delta.$$

Differenze medie

Operativamente, esistono metodi semplificati per il calcolo di Δ e Δ_R.
Possiamo infatti notare che

$$\Delta = \frac{S}{N(N-1)}, \qquad \Delta_R = \frac{S}{N^2},$$

dove S, ossia la somma di tutte le distanze di coppia, può essere riscritta come

$$\begin{aligned} S &= \sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j| \\ &= 2 \sum_{i=1}^{N} x_{(i)}(2i - N - 1) \end{aligned}$$

Con l'espressione precedente semplifichiamo notevolmente il procedimento necessario per il calcolo di Δ e Δ_R.

NB nell'ultimo termine dell'equazione precedente le osservazioni sono ordinate in senso crescente.

Differenze medie - Esercizi

Example

In una competizione sportiva, 5 atleti hanno corso i 100 metri ottenendo i seguenti tempi:

$$\{10.4; 10.8; 10.6; 10.9; 10.5\}.$$

Calcolare Δ e Δ_R usando la formula semplificata per il calcolo di S.

Variabilità con distribuzioni di frequenze

Nel caso di **distribuzioni di frequenze**, abbiamo

$$S_{M_1} = \frac{1}{N}\sum_{i=1}^{k} n_i |x_i - M_1|,$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{k} n_i (x_i - M_1)^2},$$

$$S_{Me} = \frac{1}{N}\sum_{i=1}^{k} n_i |x_i - Me|,$$

$$\Delta = \frac{1}{N(N-1)}\sum_{i=1}^{k}\sum_{j=1}^{k} n_i n_j |x_i - x_j|.$$

Nel caso di **distribuzione di frequenze in classi**, nelle precedenti rimpiazziamo le modalità x_i con i valori centrali delle classi x_i^c, $i = 1, \ldots, k$.

Indici relativi di variabilità

Gli indici proposti finora sono denominati **indici assoluti** in quanto espressi nella stessa unità di misura con cui si rilevano i valori del carattere.

Non è possibile confrontare la variabilità di caratteri espressi in unità di misura differenti usando indici assoluti. Dobbiamo ricorrere agli **indici relativi**.

La definizione di un **indice relativo** parte dall'intensità degli **scostamenti relativi rispetto alla media aritmetica**, cioè

$$\frac{|x_i - M_1(X)|}{M_1(X)}; \qquad M_1(X) > 0.$$

L'indice relativo più impiegato è il **coefficiente di variazione (CV)** definito come la **media quadratica** degli scostamenti relativi rispetto alla media aritmetica:

$$CV(X) = M_2\left(\frac{|X - M_1(X)|}{M_1(X)}\right) = \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - M_1(X))^2}{M_1^2(X)}} = \frac{\sigma(X)}{M_1(X)}$$

Indici relativi di variabilità - Esercizi

Example

In una competizione sportiva, 5 atleti alti rispettivamente

$$\{1.86\text{m}; 1.65\text{m}; 1.88\text{m}; 1.70\text{m}; 1.75\text{m}\}$$

hanno corso i 100 metri ottenendo i seguenti tempi:

$$\{10.4\text{s}; 10.8\text{s}; 10.6\text{s}; 10.9\text{s}; 10.5\text{s}\}.$$

Calcolare il **coefficiente di variazione** di entrambi i caratteri commentando i risultati ottenuti. Inoltre, senza fare ulteriori calcoli, rispondere alla seguente domanda: come cambierebbe $CV(\text{Altezza})$ se le altezze fossero misurate in cm?

Concentrazione

Concentrazione

Un'altra metodologia spesso utilizzata per studiare quanto una distribuzione sia diffusa è valutare la sua **concentrazione**. La concentrazione vuole valutare quanto ciò che osserviamo sia concentrato in poche o molte modalità del carattere in studio.

Si pensi per esempio alla ricchezza prodotta in un paese da una determinata popolazione. Tale ricchezza è **distribuita in modo omogeneo** tra gli individui, oppure **concentrata** nelle mani di pochi?

Idealmente vogliamo costruire delle misure che ci fornisca un'indicazione di quanto sia equamente distribuito o meno tra gli individui il totale di quanto osserviamo.

Concentrazione

Supponiamo di avere quindi un **insieme di osservazioni** $\{x_1, \ldots, x_N\}$ e definiamo il totale nuovamente come

$$T = \sum_{i=1}^{N} x_i.$$

Indichiamo inoltre con $\{x_{(1)}, \ldots, x_{(N)}\}$ lo stesso insieme ordinato in senso crescente, dove $x_{(i)} \leq x_{(i+1)}$. Le due situazioni estreme che possiamo osservare sono le seguenti.

1) **Equidistribuzione**, o equiripartizione, dove tutte le osservazioni assumono il medesimo valore

$$x_{(1)} = x_{(2)} = \cdots = x_{(N)} = M_1(X),$$

e siamo dunque in una situazione di **assenza di variabilità**.

Modalità	frequenza
M_1	N
Totale	N

La tabella di sinistra mostra la distribuzione di frequenze in caso di equiripartizione.

Concentrazione

2) **Massima concentrazione**, in cui una sola osservazione manifesta l'intero totale osservato, ovvero

$$\begin{cases} x_{(1)} & = x_{(2)} = \cdots = x_{(N-1)} = 0, \\ x_{(N)} & = T. \end{cases}$$

Modalità	frequenza
0	$N-1$
T	1
Totale	N

La tabella di sinistra mostra la distribuzione di frequenze in caso di massima concentrazione.

Le situazioni usualmente osservate nella concentrazione dei fenomeni studiati si collocano tra le due situazioni estreme precedenti.

La concentrazione presente in un fenomeno in studio può quindi essere valutato misurando **quanto si discosta** tra le situazioni estreme precedentemente descritte.

Concentrazione - Esempio

Example

ABC

Diagramma di Lorenz

Il **diagramma di Lorenz** rappresenta uno strumento grafico per valutare la concentrazione di un fenomeno. Consideriamo la distribuzione di frequenze descritta nella seguente tabella

valori di x_j	freq. n_j	tot. della classe $n_j \times x_j$	freq. cum. relative $p_j = \frac{C_j}{N}$	tot. cum. relativi $q_j = \frac{Q_j}{T} = \frac{\sum_{i=1}^{j} x_i n_i}{T}$
x_1	n_1	$n_1 \times x_1$	p_1	q_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_j	n_j	$n_j \times x_j$	p_j	q_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	$n_k \times x_k$	p_k	q_k
totale	N	T		

dove

$$N = \sum_{j=1}^{k} n_j, \qquad T = \sum_{j=1}^{N} n_j x_j.$$

Diagramma di Lorenz

Nella tabella precedente abbiamo due quantità fondamentali, ovvero

- La quota della popolazione o del campione che mostra una modalità del carattere minore o uguale a x_j, ovvero

$$p_j = \frac{C_j}{N} = \frac{\sum_{i=1}^{j} n_i}{N}, \qquad j = 1, \ldots, k.$$

- La quota del totale accumulato dalla popolazione o dal campione fino alla $j$esima modalità,

$$q_j = \frac{Q_j}{T} = \frac{\sum_{i=1}^{j} x_i n_i}{\sum_{i=1}^{k} x_i n_i}.$$

Supponiamo che $x_1, \ldots, x_k$ siano k redditi distinti misurati all'interno di un campione di ampiezza $N > k$. Supponiamo che per una generica classe $j$esima abbiamo

$$p_j = 0.5, \qquad q_j = 0.3.$$

Vuol dire che, il 50% più povero della popolazione detiene il 30% complessivo del reddito totale della popolazione.

Diagramma di Lorenz

Per convenzione, poniamo $q_0 = p_0 = 0$. Il diagramma di Lorenz o spezzata di Lorenz consiste nella rappresentazione grafica in un piano cartesiano dei punti (p_j, q_j), $j = 0, \ldots, k$.

Consideriamo per esempio la tabella seguente.

j	p_j	q_j
0	0	0
1	0.3	0.1
2	0.6	0.22
3	0.7	0.3
4	1	1

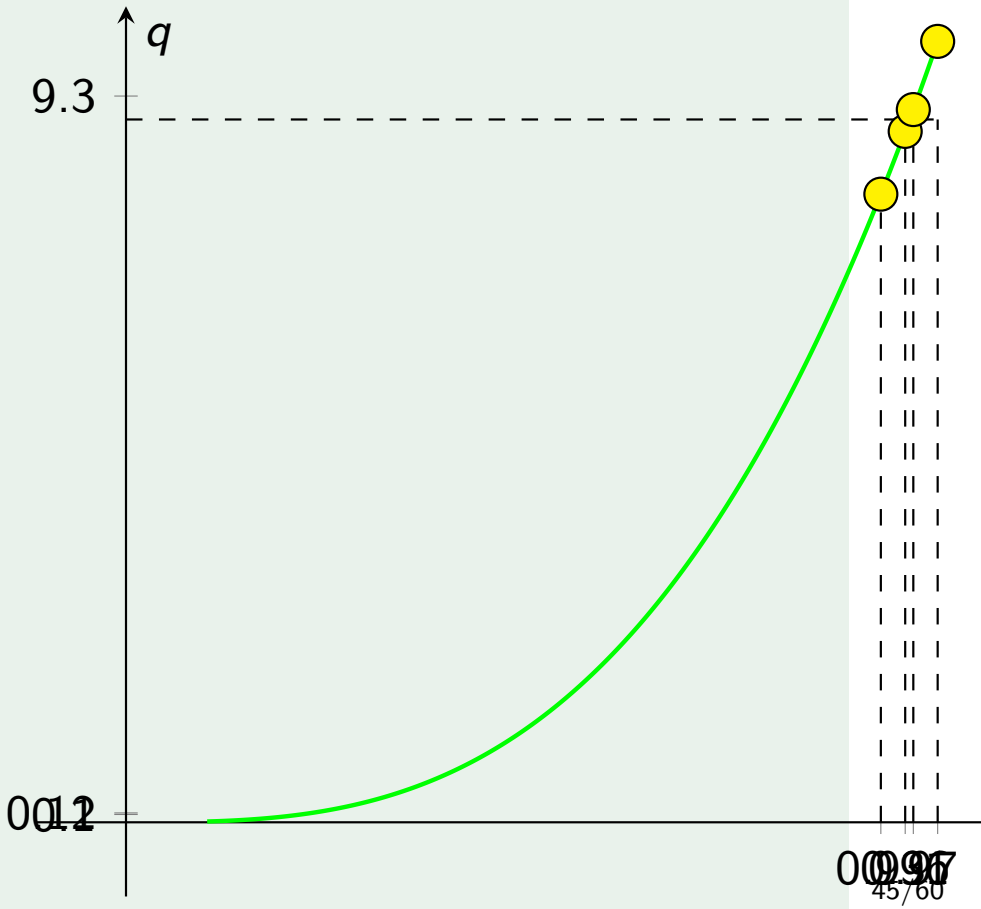


Diagramma di Lorenz

Nei casi estremi abbiamo le seguenti situazioni.

- Quand siamo in **equiripartizione**, abbiamo $k = 1$ e $p_1 = q_1 = 1$, ottenendo come tabella di frequenze e diagramma di Lorenz

j	p_j	q_j
0	0	0
1	1	1

La precedente viene detta **spezzata di equiripartizione**.

Diagramma di Lorenz

- Quand siamo in **equiripartizione**, abbiamo $k = 2$, ottenendo come tabella di frequenze e diagramma di Lorenz

j	p_j	q_j
0	0	0
1	$\frac{N-1}{N}$	0
2	1	1

La precedente viene detta **spezzata di massima concentrazione**.

Diagramma di Lorenz

Proprietà del diagramma di Lorenz

- Nel caso di distribuzioni di unità, dove osserviamo valori distinti $x_1, \ldots, x_N$, abbiamo

$$k = N,$$
$$p_j = \frac{j}{N}, \qquad j = 1, \ldots, N,$$
$$q_j = \frac{1}{T} \sum_{i=1}^{j} x_{(i)}, \qquad j = 1, \ldots, N,$$

  dove $x_{(1)}, \ldots, x_{(N)}$ rappresentano i valori $x_1, \ldots, x_N$ ordinati in senso crescente.

- I segmenti che compongono la spezzata del diagramma di Lorenz hanno inclinazioni **crescenti o costanti**, ovvero

$$p_j \geq q_j, \qquad j = 1, \ldots, k.$$

  La precedente implica che la spezzata di Lorenz giace sotto la retta di equiripartizione.

Diagramma di Lorenz

Example

Esempio curva di Lorenz con distribuzioni di unità

Diagramma di Lorenz

Partendo dal diagramma di Lorenz, possiamo **costruire una misura** per valutare la concentrazione di una distribuzione.

In particolare, possiamo valutare di quanto si allontana la spezzata che osserviamo dalla retta di equiripartizione.

Tanto più l'area compresa tra la retta di equiripartizione e la spezzata osservata è piccola, tanto più siamo **prossimi ad una situazione di equiripartizione**.

Un indice comunemente utilizzato è l'indice di concentrazione di Gini R.

Indice di concentrazione di Gini

L'**indice di concentrazione di Gini** può essere definito in diversi modi, qui consideriamo il seguente rapporto

$$R = \frac{\text{area di concentrazione osservata}}{\text{area di massima concentrazione}}.$$

Di seguito vediamo come calcolare entrambe le quantità.

- **Area di massima concentrazione**, ovvero l'area compresa tra la bisettrice e la spezzata di massima concentrazione.

Abbiamo un triangolo di base $\frac{N-1}{N}$, e altezza 1, quindi l'area possiamo ottenerla come

$$A_{max} = \frac{N-1N \times 1}{2} = \frac{N-1}{2N}.$$

Indice di concentrazione di Gini

- **L'area di concentrazione osservata** può essere decomposta nella somma di aree più semplici.

$$
\begin{aligned}
A_1 &= \frac{(p_1 - q_1)(p_1 - p_0)}{2} \\
&\quad + \frac{(p_1 - q_1)(p_2 - p_1)}{2} \\
&= \frac{(p_1 - q_1)(p_1 - p_0 + p_2 - p_1)}{2} \\
&= \frac{(p_1 - q_1)(p_2 + p_0)}{2} \\
A_2 &= \frac{(p_2 - q_2)(p_2 - p_1)}{2} \\
&\quad + \frac{(p_2 - q_2)(p_3 - p_2)}{2} \\
&= \frac{(p_2 - q_2)(p_3 + p_1)}{2} \\
&\vdots
\end{aligned}
$$

Indice di concentrazione di Gini

In generale, l'area di concentrazione osservata per k modalità o classi è uguale a

$$A_{oss} = \frac{1}{2}\sum_{j=1}^{k-1}(p_j - q_j)(p_{j+1} - p_{j-1}).$$

Ricordiamo che

$$p_j = \frac{C_j}{N}, \qquad j = 1, \ldots, k.$$

Di conseguenza,

$$p_{j+1} - p_{j-1} = \frac{C_{j+1} - C_{j-1}}{N} = \frac{\sum_{i=1}^{j+1} n_i - \sum_{i=1}^{j-1} n_i}{N} = \frac{n_{j+1} - n_j}{N}.$$

Utilizzando la precedente, possiamo semplificare l'area di concentrazione come

$$A_{oss} = \frac{1}{2N}\sum_{j=1}^{k-1}(p_j - q_j)(n_{j+1} - n_{j-1}).$$

Indice di concentrazione di Gini

- Concludendo, abbiamo che **l'indice di concentrazione di Gini** può essere scritto come

$$R = \frac{\text{area di concentrazione osservata}}{\text{area di massima concentrazione}} = \frac{\frac{1}{2N}\sum_{j=1}^{k-1}(p_j - q_j)(n_{j+1} - n_{j-1})}{\frac{N-1}{2N}}$$

$$= \frac{1}{N-1}\sum_{j=1}^{k-1}(p_j - q_j)(n_{j+1} - n_{j-1}).$$

- Nel caso di distribuzioni di unità, la precedente diventa

$$R = \frac{2}{N-1}\sum_{j=1}^{N-1}(p_j - q_j).$$

- Si può dimostrare che l'indice di concentrazione di Gini può essere scritto come

$$R = \frac{\Delta}{M_1},$$

dove Δ è la differenza media semplice (senza ripetizione) e M_1 la media aritmetica.

Proprietà dell'indice di concentrazione di Gini

- $R = 0$ nel caso di equiripartizione.
- $R = 1$ nel caso di massima concentrazione.
- L'indice aumenta se tutte le modalità vengono aumentate di una quantità generica $h > 0$.
- L'indice di concentrazione di Gini è sensibile a trasferimenti.
  - Se spostiamo una parte di una modalità tra le più alte osservate nelle modalità dal valore più basso, l'indice diminuisce.
  - Viceversa, se spostiamo una parte delle modalità più basse osservate nelle più alte, l'indice aumenta.
- L'indice è invariante rispetto a trasformazioni di scala, ovvero la concentrazione di un carattere X è uguale alla concentrazione di un carattere Y se

$$Y = aX, \qquad a \in \mathbb{R}.$$

Mutabilità

Mutabilità

Nel caso di variabili qualitative, l'analogo della variabilità viene detto **mutabilità** o **eterogeneità**.

Analogamente a quanto visto per la concentrazione, abbiamo due casi estremi.

- **Minima mutabilità**, ovvero quando le unità statistiche osservate assumono tutte la stessa modalità,

modalità	x_1	$\dots$	x_j	$\dots$	x_k
frequenza	0	$\dots$	N	$\dots$	0

- **Massima mutabilità**, ovvero quando le unità statistiche osservate assumono tutte modalità diverse tra loro,

modalità	x_1	$\dots$	x_j	$\dots$	x_k
frequenza	1	$\dots$	1	$\dots$	1

Idealmente, vogliamo costruire degli indici che misurano dove ci collochiamo all'interno di queste due situaizoni estreme.

Mutabilità

Un primo indice che consideriamo è **l'indice di mutabilità di Gini**, definito come

$$G = \sum_{j=1}^{k} f_j(1 - f_j) = 1 - \sum_{j=1}^{k} f_j^2,$$

dove $f_j = \frac{n_j}{N}$, $j = 1, \ldots, N$, sono le frequenze relative che osservaimo per ogni modalità.

- In condizioni di mutabilità minima, $G = 0$.
- In condizioni di mutabilità massima,

$$G = 1 - \sum_{j=1}^{k} f_j^2 = 1 - \frac{k}{k^2} = 1 - \frac{1}{k} = \frac{k-1}{k}.$$

- Possiamo normalizzare l'indice di mutabilità di Gini, ottenendo un indice che assume valori tra 0 e 1, dividendolo per il suo massimo, ovvero

$$G_{norm} = \frac{G}{\frac{k-1}{k}} = \frac{kG}{k-1}.$$

Mutabilità

Example

Esempio indice Gini

Mutabilità

Un secondo indice che consideriamo è **l'entropia di Shannon**, definito come

$$H = -\sum_{j=1}^{k} f_j \log(f_j),$$

dove $f_j = \frac{n_j}{N}$, $j = 1, \ldots, N$, sono le frequenze relative che osservaimo per ogni modalità.

- In condizioni di mutabilità minima, $H = 0$.
- In condizioni di mutabilità massima,

$$H = -\sum_{j=1}^{k} \frac{1}{k} \log\left(\frac{1}{k}\right) = -\log\left(\frac{1}{k}\right) = \log k.$$

- Possiamo normalizzare l'entropia di Shannon, ottenendo un indice che assume valori tra 0 e 1, dividendolo per il suo massimo, ovvero

$$H_{norm} = \frac{H}{\log k}.$$

Mutabilità

Example

Esempio entropia Shannon