Statistica I

Unità F: indici di variabilità

Tommaso Rigon

Università Milano-Bicocca

Anno Accademico 2020-2021

Unità F

Argomenti affrontati

- Concetto di variabilità
- Varianza e scarto quadratico medio
- Altre misure di variabilità (campo di variazione, scarto interquartile, MAD)
- Standardizzazione dei dati
- Il coefficiente di variazione

Riferimenti al libro di testo

- §5.1 — §5.3
- **Nota**. Nel libro di testo sono presenti proprietà della varianza aggiuntive.

Descrizione del problema

- Siamo interessati a confrontare l'efficacia di due diverse metodologie, chiamate A e B, per la **misurazione dell'emoglobina** nel sangue.
- Si è creato in laboratorio del **sangue artificiale** contenente 15 grammi di emoglobina ogni 100cm^3.
- Dal composito sono stati estratti in totale $n = 360$ campioni.
- Di questi, in $n_A = 180$ campioni l'emoglobina è stata misurata utilizzando la metodologia A mentre per i restanti $n_B = 180$ campioni è stata usata la metodologia B.
- Alcuni dati sono riportati nella prossima slide. Le differenze tra le diverse misurazioni sono da attribuire in larga parte agli errori di misura delle due diverse metodologie.

I dati grezzi

Grammi di emoglobina ogni 100cm^3, metodologia A. $n_A = 180$

```
 [1] 14.98654 15.14828 15.15741 14.78573 15.00364 15.06475
 [7] 14.99282 14.92189 14.93331 14.94189 15.22719 14.64697
[13] 14.85369 15.35937 15.09510 14.70682 14.88053 15.11486
[19] 15.01449 15.10965 14.72711 14.98090 14.75410 14.90115
[25] 15.21905 14.97432 15.04769 14.90602 15.11311 14.78668
...
```

Grammi di emoglobina ogni 100cm^3, metodologia B. $n_B = 180$

```
 [1] 14.62067 15.26097 14.87602 15.45027 15.09104 15.31831
 [7] 15.06252 15.19373 14.31944 15.36786 15.48341 15.01780
[13] 14.34351 14.58493 14.97563 15.29785 15.28055 15.53123
[19] 13.82846 15.12360 14.83586 15.60325 14.85619 15.01115
[25] 14.64376 14.95360 15.53356 15.69041 15.10458 14.56744
...
```

Indici di posizione

- Per cominciare, descriviamo i dati utilizzando i concetti che abbiamo appreso finora.

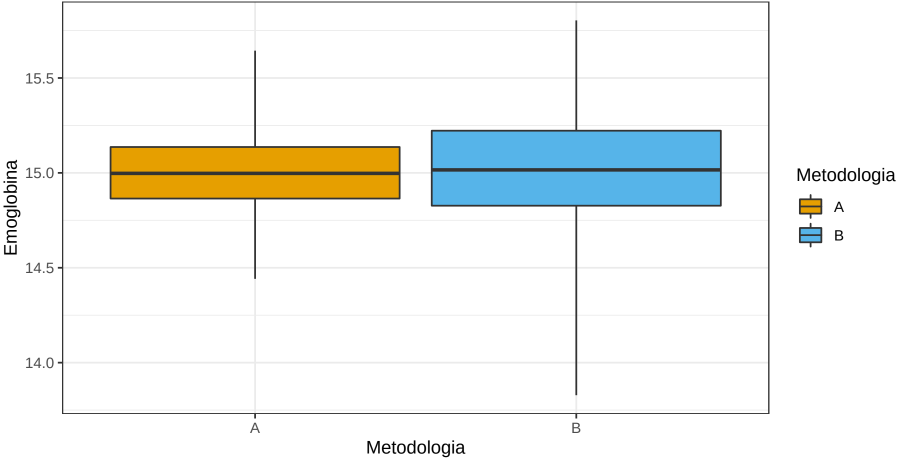
	Metodologia **A**	Metodologia **B**
Minimo di Emoglobina	14.44	13.83
Primo quartile di Emoglobina	14.86	14.83
Media di Emoglobina	15.00	15.02
Mediana di Emoglobina	15.00	15.02
Terzo quartile di Emoglobina	15.14	15.22
Massimo di Emoglobina	15.64	15.80

- Dalla tabella seguente, è abbastanza chiaro che entrambe le metodologie, in media, registrano circa 15g di emoglobina nel sangue artificiale.

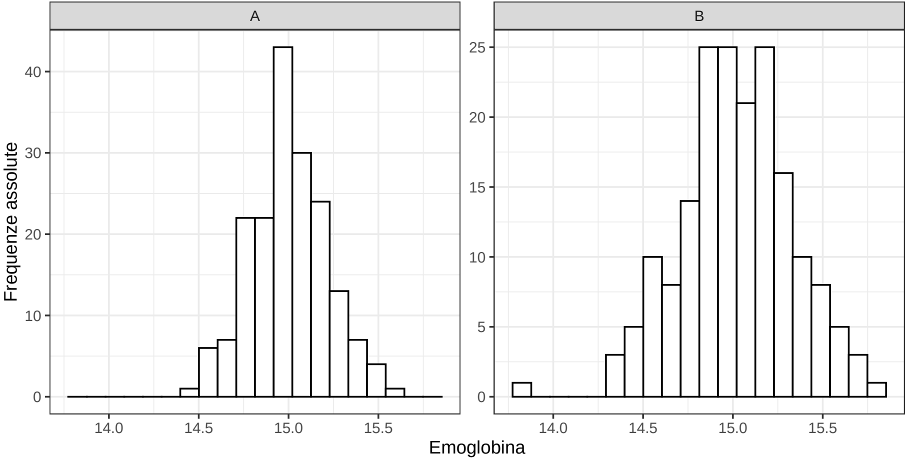
- Pertanto, entrambe le metodologie sono **ben calibrate**.

- Possiamo quindi concludere che le due metodologie sono equivalenti, oppure una delle due è preferibile?

Boxplot

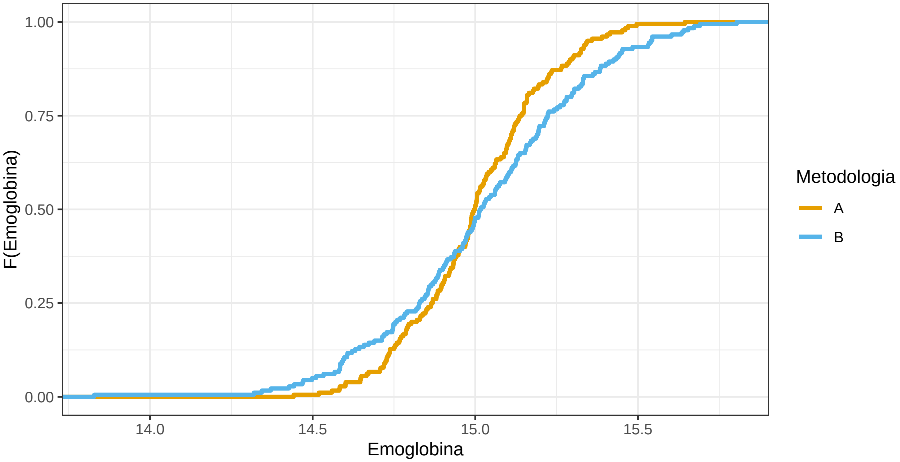


Istogrammi



A
B
Frequenze assolute
0
10
20
30
40
0
5
10
15
20
25
14.0
14.5
15.0
15.5
Emoglobina

Funzioni di ripartizione



- **Nota**. È importante che lo studente cerchi di capire che l'incrocio delle due funzioni di ripartizione empiriche è dovuto alla differente variabilità dei due insiemi di dati.

Commento al problema

- Ambedue le metodiche sembrano essere state **tarate accuratamente** visto che i due insiemi di dati si distribuiscono intorno al valore nominale, ovvero 15g.
- Tuttavia, gli errori di misura della metodica B sembrano essere più grandi. Infatti in questo caso i dati sono più **dispersi** intorno al valore nominale.
- In termini più appropriati, si dice che sono caratterizzati da una **variabilità** maggiore.
- Come possiamo **misurare** la variabilità di una distribuzione? Un possibile indice è la varianza.

	Metodologia **A**	Metodologia **B**
Varianza di Emoglobina	0.046	0.099

La varianza

- Così come per la posizione, siamo interessati a indici che ci permettano di valutare in maniera sintetica la variabilità di un insieme di dati.

- **Varianza**. La varianza dei dati $x_1, \dots, x_n$ è

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

- A volte indicheremo la varianza della variabile x con il simbolo $\text{var}(x)$.

- La **varianza** è quindi una misura di quanto i **dati** siano **distanti dalla media aritmetica**.

- Tale distanza è valutata usando i quadrati delle differenze.

La varianza: definizione alternativa

- La varianza può anche essere definita in una **maniera alternativa**, ovvero come una media delle differenze al quadrato tra tutte le possibili coppie di dati.

- **Varianza**. La varianza dei dati $x_1, \dots, x_n$ è

$$\sigma^2 = \frac{1}{2n^2}\sum_{i=1}^n\sum_{j=1}^n (x_i - x_j)^2 = \frac{1}{n^2}\sum_{i<j}(x_i - x_j)^2.$$

- Questo definizione è **meno utilizzata** in pratica perché più scomoda da calcolare.

- Tuttavia, tale definizione chiarisce che la varianza si può interpretare come la media delle distanze tra tutte le coppie di valori.

Dimostrazione

- L'equivalenza tra le due definizione si dimostra come segue

$$
\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n (x_i - x_j)^2 &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n [(x_i - \bar{x}) - (x_j - \bar{x})]^2 \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n (x_i - \bar{x})^2 + \frac{1}{n^2}\sum_{j=1}^n\sum_{j=1}^n (x_j - \bar{x})^2 + \\
&\qquad - \frac{2}{n^2}\sum_{i=1}^n\sum_{j=1}^n (x_i - \bar{x})(x_j - \bar{x}) \\
&= \frac{2n}{n^2}\sum_{i=1}^n (x_i - \bar{x})^2 - 2\left[\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})\right]^2 \\
&= 2 \times \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 = 2\sigma^2.
\end{aligned}
$$

La varianza: formula per il calcolo

- La varianza ammette una ulteriore definizione, **spesso utilizzata** in pratica perchè semplice da calcolare.

- **Varianza**. La varianza dei dati $x_1, \dots, x_n$ è

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

- La dimostrazione si ottiene facilmente come segue:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 + \frac{1}{n}\sum_{i=1}^n \bar{x}^2 - 2\bar{x} \times \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{n}\sum_{i=1}^n x_i^2 + \bar{x}^2 - 2\bar{x}^2$$
$$= \frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2.$$

Esempio di calcolo della varianza

- **Dati** $x_1, \dots, x_4$: `1, 3, 2, 5.`

- La media aritmetica (**momento primo**) dei dati è $\bar{x} = (1 + 3 + 2 + 5)/4 = 2.75$.

- La media dei quadrati (**momento secondo**) è pari a

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1^2 + 3^2 + 2^2 + 5^2}{4} = 9.75.$$

- Pertanto la varianza è pari a

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = 9.75 - 2.75^2 = 2.19.$$

- **Esercizio**. Si ottenga la varianza dei dati $x_1, \dots, x_4$ utilizzando le altre definizioni.

Esempio di calcolo della varianza

- Supponiamo di avere i seguenti dati **discreti**, le cui $k = 3$ modalità sono presentate nel seguito.

Modalità c_j	4	6	7
Frequenze assolute n_j	2	8	3

- La media aritmetica vale $\bar{x} = (4 \times 2 + 6 \times 8 + 7 \times 3)/13 = 5.9231$.

- Inoltre, la media dei valori al quadrato è pari a

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{n}\sum_{j=1}^{k} n_j c_j^2 = (2 \times 4^2 + 8 \times 6^2 + 3 \times 7^2)/13 = 35.9230.$$

- Quindi, la varianza si può calcolare come segue

$$\sigma^2 = 35.9230 - 5.9231^2 = 0.840.$$

Proprietà della varianza

- **Proprietà**. La varianza è per costruzione sempre **maggiore o uguale a zero**, ovvero
$$\sigma^2 \geq 0.$$
- Inoltre, la varianza è esattamente pari a zero solo se i dati sono uguali tra loro.
- Infatti ad esempio se
$$x_1 = x_2 = \cdots = x_n = a,$$
dove $a \in \mathbb{R}$ è una costante qualsiasi, allora
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left[(a-a)^2 + \cdots + (a-a)^2\right] = 0.$$
- Si può dimostrare anche il viceversa: se $\sigma^2 = 0$ allora necessariamente le osservazioni sono uguali tra loro (si veda la definizione alternativa di varianza per convincersene).

Proprietà della varianza: trasformazione lineare

- **Proprietà**. Se consideriamo i dati trasformati $y_1, \dots, y_n$, tali che

$$y_i = a + bx_i, \qquad i = 1, \dots, n,$$

dove $a, b \in \mathbb{R}$ sono due numeri qualsiasi, allora

$$\text{var}(y) = b^2 \text{var}(x).$$

- La relazione precedente permette di calcolare agevolmente la varianza delle y_i senza dover calcolare le y_i stesse.

- La dimostrazione segue dalle proprietà della media e delle sommatorie. Infatti:

$$\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^n (a + bx_i - a + b\bar{x})^2 = \frac{1}{n}\sum_{i=1}^n (bx_i - b\bar{x})^2 = b^2\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2.$$

- **Esercizio - proprietà**. La varianza delle y_i non dipende dalla costante a. Si spieghi come mai il contrario sarebbe stato per molti versi "preoccupante".

Lo scarto quadratico medio

- La radice quadrata della varianza è tipicamente chiamata **scarto quadratico medio** e scriveremo
$$(\text{scarto quadratico medio}) = \sigma = \sqrt{\sigma^2}.$$

- A volte useremo anche la notazione $\text{sqm}(x) = \sqrt{\text{var}(x)}$.

- L'unità di misura della varianza è uguale al quadrato dell'unità di misura dei dati.

- Invece, l'unità di misura dello scarto quadratico medio coincide con l'unità di misura dei dati.

Altre misure di variabilità

- In aggiunta alla varianza sono stati suggeriti una molteplicità di indici di variabilità. Ne presentiamo qui 3 tra i più **diffusi**.

- **Campo di variazione**. È la differenza tra il massimo ed il minimo, ovvero

$$(\text{Campo di variazione}) = x_{(n)} - x_{(1)}.$$

  È un indice molto semplice da calcolare, ma **estremamente sensibile** a **valori anomali**.

- **Scarto interquartile**. È la differenza tra il terzo ed il primo quartile, ovvero

$$(\text{Scarto interquartile}) = \mathcal{Q}_{0.75} - \mathcal{Q}_{0.25}.$$

  È molto più **resistente** della varianza in presenza di poche osservazioni estreme. È usato soprattutto nelle situazioni in cui si sospetta la presenza di osservazioni anomale.

- **MAD (Median Absolute Deviations)**. È definito come segue

$$\text{MAD} = \text{Mediana}(|x_1 - \text{Me}_x|, \dots, |x_n - \text{Me}_x|), \qquad \text{Me}_x = \text{Mediana}(x_1, \dots, x_n).$$

  Anche questo indice è poco sensibile alla presenza di dati anomali.

Misurazione dell'emoglobina: indici di variabilità

Emoglobina	Metodologia **A**	Metodologia **B**
Varianza	0.046	0.099
Scarto quadratico medio	0.214	0.315
Campo di variazione	1.202	1.975
Scarto interquartile	0.272	0.395
MAD	0.136	0.198

- Tutti gli indici considerati evidenziano una **maggiore variabilità** delle misure ottenute con la metodologia **B**.

Il coefficiente di variazione

- La varianza e lo scarto quadratico medio sono **indici assoluti** che guardano alle differenze tra unità statistiche.

- Tuttavia, il significato pratico di tali indici potrebbe dipendere dal livello del fenomeno considerato.

- Si pensi, ad esempio, al reddito. Una differenza di 30'000 euro nel reddito annuo è rilevante se confrontiamo lo stipendio ad esempio di due operai. La stessa differenza è praticamente irrilevante se confrontiamo i redditi di Jeff Bezos e Bill Gates.

- **Coefficiente di variazione**. Se $\bar{x} > 0$, il coefficiente di variazione è

$$\mathrm{CV} = \frac{\text{(Scarto quadratico medio)}}{\text{(Media aritmetica)}} = \frac{\sigma}{\bar{x}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\bar{x}}\right)^2}.$$

- Il CV è **indipendente** dall'unità di misura e aggiusta la variabilità tenendo conto anche del livello del fenomeno.

Standardizzazione dei dati

- A volte è utile trasformare un insieme di dati $x_1, \dots, x_n$ in maniera tale che i dati trasformati, indichiamoli $z_1, \dots, zn$, abbiano **media nulla** e **varianza unitaria**.

- È facile verificare che una trasformata appropriata consiste nel porre

$$z_i = \frac{y_i - (\text{media aritmetica})}{(\text{scarto quadratico medio})} = \frac{x_i - \bar{x}}{\sigma}, \qquad i = 1, \dots, n.$$

- I dati trasformati $z_1, \dots, z_n$ vengono usualmente chiamati **standardizzati**.

- **Esercizio - proprietà**. Si verifichi che

$$\bar{z} = 0, \qquad \text{var}(z) = 1.$$

Disuguaglianza di Chebyshev

Teorema (Disuguaglianza di Chebyshev)

Siano $x_1, \dots, x_n$ dei dati aventi media $\bar{x}$ e scarto quadratico medio σ. Per qualsiasi valore di $k \geq 1$ vale che

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(|x_i - \bar{x}| > k\sigma) \leq \frac{1}{k^2}.$$

- Equivalentemente, si ottiene che

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(|x_i - \bar{x}| \leq k\sigma) \geq 1 - \frac{1}{k^2}.$$

- In altri termini, la frequenza relativa dei dati distanti dal centro è limitata, ovvero

$$\frac{(\text{Numero di osservazioni che distano da } \bar{x} \text{ più di } k\sigma)}{(\text{Numero di osservazioni})} \leq \frac{1}{k^2}.$$

- **Esempio**. Se $\bar{x} = 0$ e $\sigma = 1$, allora la frazione di dati $x_1, \dots, x_n$ compresi tra -4 e 4 è almeno pari a $1 - 1/16 = 15/16 \approx 0.94$.

Dimostrazione

- La dimostrazione della disuguaglianza di Chebyshev segue dalla definizione di varianza.

$$
\begin{aligned}
n\sigma^2 = \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i : x_i - \bar{x} < -k\sigma} (x_i - \bar{x})^2 + \sum_{i : |x_i - \bar{x}| \leq k\sigma} (x_i - \bar{x})^2 + \sum_{i : x_i - \bar{x} > k\sigma} (x_i - \bar{x})^2 \\
&\geq \sum_{i : x_i - \bar{x} < -k\sigma} (x_i - \bar{x})^2 + \sum_{i : x_i - \bar{x} > k\sigma} (x_i - \bar{x})^2 \\
&\geq \sum_{i : x_i - \bar{x} < -k\sigma} (k\sigma)^2 + \sum_{i : x_i - \bar{x} > k\sigma} (k\sigma)^2 \\
&= \sum_{i : |x_i - \bar{x}| > k\sigma} (k\sigma)^2 = k^2\sigma^2 \sum_{i=1}^n \mathbb{1}(|x_i - \bar{x}| > k\sigma).
\end{aligned}
$$

- Il risultato segue dividendo entrambi i membri della disuguaglianza per $nk^2\sigma^2$.

Commenti alla disuguaglianza di Chebyshev

- La disuguaglianza di Chebyshev consente di determinare una limitazione della frequenza relativa delle **code** della distribuzione.
- È bene sottolineare che tale disuguaglianza vale per qualsiasi insieme di dati.
- In altri termini, tale disuguaglianza chiarisce che media e varianza sono **strumenti molto potenti**.
- Media e varianza consentono di "controllare" con un buon grado di approssimazione una qualsiasi distribuzione.
- **Esercizio**. Cosa succede alla disuguaglianza di Chebyshev se $k = 1$? Il risultato è corretto? È utile?