Statistica I

Unità G: istogrammi e boxplot

Tommaso Rigon

Università Milano-Bicocca

Anno Accademico 2020-2021

Unità G

Argomenti affrontati

- Numero di intervalli in un istogramma
- Intervalli di ampiezza diversa, concetto di densità
- Diagrammi a bastoncini
- Una variante dei diagrammi a scatola con baffi (boxplot)

Riferimenti al libro di testo

- §3.4
- §6.4 — §6.5 (escluse pagine 169–170)
- **Nota**. Il concetto di istogramma perequato non è materia d'esame.

Istogrammi

- Nella costruzione di un istogramma esiste un elemento di arbitrarietà: la scelta di **quanti** e **quali** intervalli utilizzare.

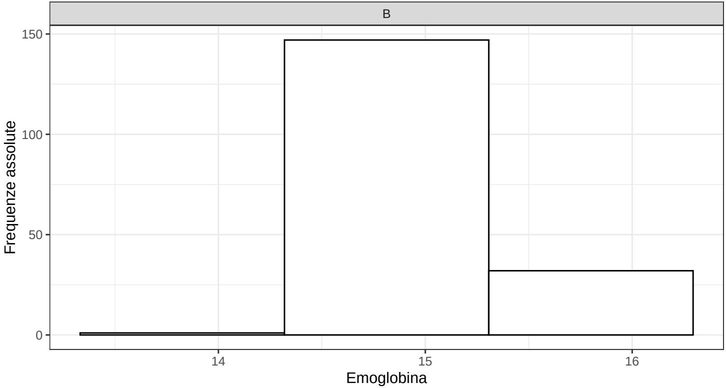
- Finora, abbiamo presentato una **versione semplificata** dell'istogramma, anche se corretta.

- Ad esempio, abbiamo finora assunto che l'ampiezza delle classi fosse uguale.

- Inoltre, vorremmo anche considerare istogrammi basati su "frequenze relative".

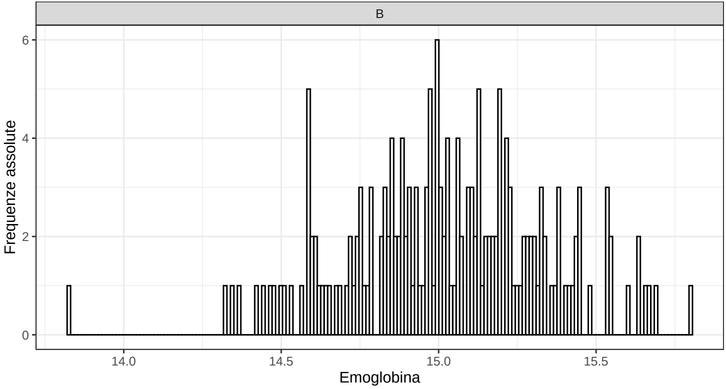
Istogrammi: numero di intervalli

- Consideriamo anzitutto il primo problema: la scelta del **numero di intervalli**.
- Consideriamo le misurazioni dell'emoglobina, metodologia B (si veda l'unità F).
- Un numero **troppo basso** di intervalli comporta una **perdita di informazione**.



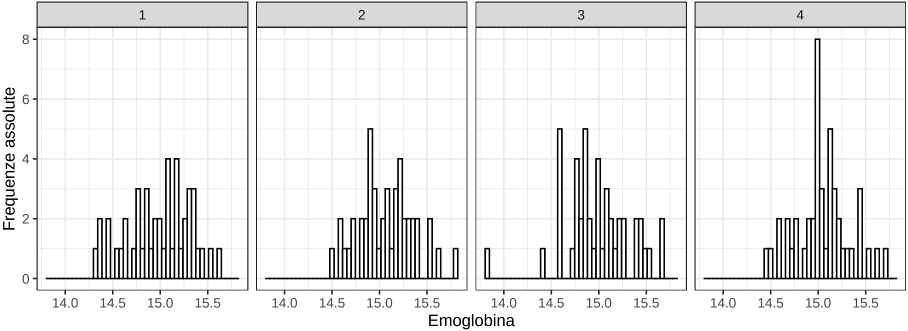
Istogrammi: numero di intervalli

- Viceversa, un numero **troppo alto** di intervalli comporta una **perdita di sintesi**.
- Le oscillazioni che osserviamo sono probabilmente **rumore**, caratteristiche particolari dei dati disponibili più che della metodica usata per il dosaggio.



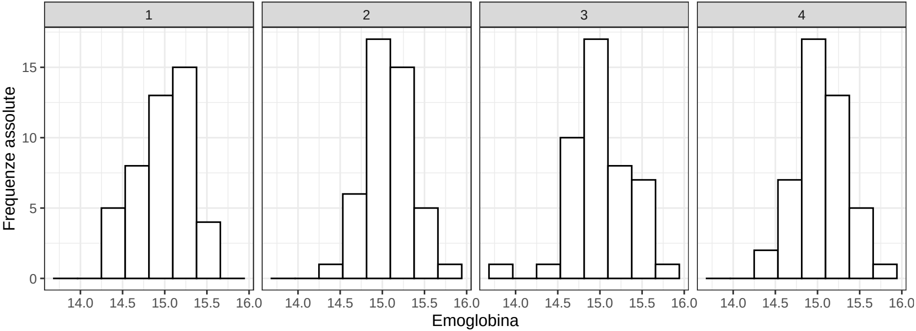
Istogrammi: troppi intervalli e rumore

- Perché troppi intervalli sono un problema?
- Se **dividiamo a caso i dati** in 4 gruppi ci aspettiamo che i relativi istogrammi siano simili, dato che provengono dalla stessa distribuzione.
- Intervalli troppo piccoli enfatizzano il **rumore**.



Istogrammi: troppi intervalli e rumore

- Usando meno intervalli, si riduce anche il **rumore**. I sottogruppi sono gli stessi della slide precedente.
- Gli istogrammi presentati qui di seguito sono più coerenti tra loro, anche se diversi.



Istogrammi: scelta del numero di intervalli

- In pratica, conviene fare più di un grafico e determinare il numero "ottimale" sulla base dei risultati.
- Si tenga presente che il **numero degli intervalli** deve dipendere dal **numero dei dati**: ripartire 1000 osservazioni in 40 intervalli può anche dare risultati sensati, usare gli stessi 40 intervalli per 20 dati non può che dare un risultato erratico.
- Sono state proposte varie formule per identificare il "numero ottimale" di intervalli. Vanno però prese come dei suggerimenti e non usate in maniera automatica.
- **Sturges**. Il numero di intervalli, approssimato all'intero più vicino, è

$$(\text{numero di intervalli}) = 1 + \log_2 n.$$

- **Freedman & Diaconis**. Il numero di intervalli, approssimato all'intero più vicino, è

$$(\text{numero di intervalli}) = \frac{x_{(n)} - x_{(1)}}{2(\mathcal{Q}_{0.75} - \mathcal{Q}_{0.25})} n^{1/3}.$$

Istogrammi: intervalli di differenti lunghezze

- Gli istogrammi che abbiamo considerato finora sono stati costruiti ponendo

$$\text{(base rettangoli)} = \text{(lunghezza intervalli)}$$
$$\text{(altezza rettangoli)} = \text{(frequenze assolute)}$$

- Questa definizione **non è appropriata** se gli **intervalli** hanno **dimensioni diverse**.

- Infatti, sia per scelta che per necessità, può capitare di dover rappresentare un istogramma con un intervalli di ampiezze diverse.

- In tal caso, è importante rendersi conto che le altezze dei rettangoli non devono essere pari alle frequenze osservate ma **proporzionali** alla **densità** delle osservazioni nelle singole classi. La densità è definita come

$$d_j = \text{(densità di un intervallo)} = \frac{\text{(frequenza assoluta)}}{\text{(lunghezza intervallo)}} = \frac{n_j}{a_j - a_{j-1}}, \quad j = 1, \dots, k.$$

Istogrammi: intervalli di differenti lunghezze

- Per capire la definizione si pensi alla popolazione. È la densità della popolazione e non il numero totale di abitanti che ci dice quanto questi sono addensati in una certa zona.

- Ricapitolando, costruiremo gli istogrammi ponendo

$$\text{(base rettangoli)} = \text{(lunghezza intervalli)}$$

$$\text{(altezza rettangoli)} = \lambda \times \text{(densità)} = \lambda \times \frac{\text{(frequenze assolute)}}{\text{(lunghezza intervalli)}}$$

  dove $\lambda > 0$ è un numero qualsiasi.

- Quando gli intervalli sono tutti uguali, allora si può porre $\lambda =$ (lunghezza intervalli) e si ritorna alla definizione originaria.

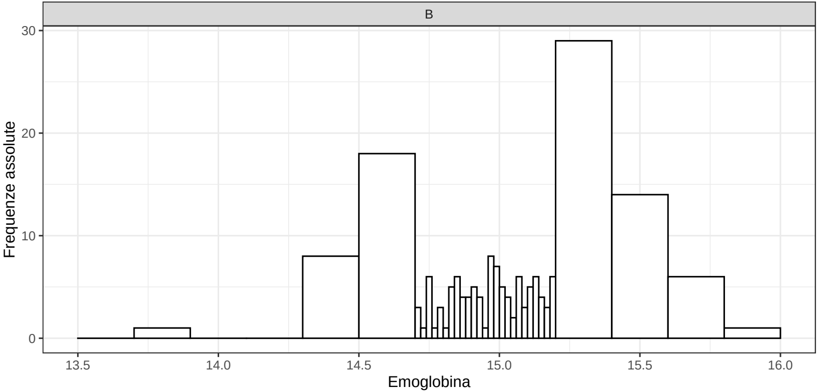
- Sebbene in teoria λ possa essere scelto a piacere, tipicamente si pone $\lambda = 1/n$, ovvero λ è quel numero che rende l'**area complessiva dei rettangoli pari a 1**.

- L'uso della densità è anche legato alla nostra **percezione**. Visivamente infatti "alto" è implicitamente associato a "tanto", come illustrato nei seguenti esempi.

Istogramma sbagliato

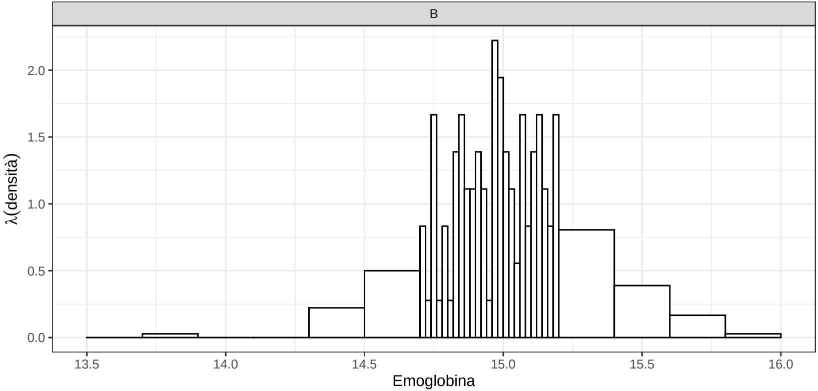
- L'istogramma della figura sottostante **è errato**. Viene mostrato per illustrare la necessità della densità.

- L'utilizzato errato della definizione originaria in presenza di intervalli diversi crea un "buco".



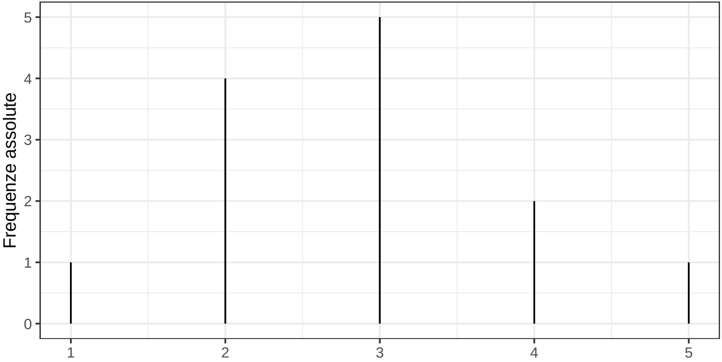
Istogramma corretto

- Il buco al centro è sparito. Il grafico correttamente ci dice che le osservazioni sono addensate attorno ai 15g per 100cm^3.



Modalità	1	2	3	4	5
Frequenze assolute	1	4	5	2	1

- Se la variabile in esame è **discreta**, possiamo anche evitare del tutto la scelta sul numero di intervalli ed usare un **diagramma a bastoncini**.



Boxplot e valori estremi

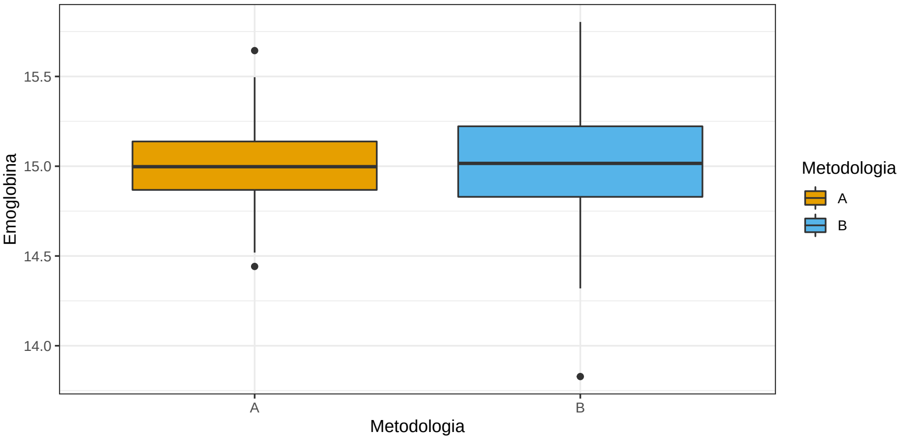
- Spesso con un boxplot si vuole: (i) descrivere in maniera stilizzata la distribuzione dei dati ma anche anche (ii) evidenziare eventuali **valori estremi** (outlier).
- Una **variante** del diagramma usata a questo scopo può essere costruita come segue:
- La scatola è costruita come descritto nell'unità C a partire dai tre quartili.
- I baffi si estendono fino ai dati più lontani che siano però non più distanti di

$$\lambda \times (\mathcal{Q}_{0.75} - \mathcal{Q}_{0.25})$$

dalla scatola.

- λ è una costante arbitraria tipicamente scelta uguale a 1.5.
- Le osservazioni che sono oltre i baffi sono disegnate opportunamente sul grafico, ad esempio utilizzando un pallino.

Boxplot e outlier



Boxplot e valori estremi: esempio di calcolo

- **Dati ordinati** $x_{(1)}, \dots, x_{(12)}$: `1.1, 1.3, 1.4, 1.6, 1.8, 1.9, 2.0, 2.5, 2.9, 3.2, 4.1, 5.6`.

- Pertanto, abbiamo che $\mathcal{Q}_{0.25} = 1.4$, $\mathrm{Me} = 1.95$ e $\mathcal{Q}_{0.75} = 2.9$. Quindi si ottiene che
$$1.5(\mathcal{Q}_{0.75} - \mathcal{Q}_{0.25}) = 1.5 \times 1.5 = 2.25.$$

- Allora, la scatola si estende da 1.4 a 2.9 con mediana indicata da una linea a 1.95.

- Il baffo inferiore si estende fino all'osservazione più bassa tra quelle maggiori di $1.4 - 2.25 = -0.85$. Ovvero, in questo caso, fino a al minimo $x_{(1)} = 1.1$.

- Il baffo superiore si estende fino all'osservazione più alta tra quelle minori di $2.9 + 2.25 = 5.15$, ovvero fino a $x_{(11)} = 4.1$.

- Il punto $x_{(12)} = 5.6$ viene disegnato esplicitamente con un punto.

- **Esercizio**. Disegnare il boxplot.