Statistica I

Unità C: Indici di posizione

Tommaso Rigon

Università Milano-Bicocca

Unità C

Argomenti affrontati

- Media aritmetica
- Mediana, quartili e percentili
- Diagramma a scatola con baffi (boxplot)
- Medie di Bonferroni, di Chisini e di Wald

Riferimenti al libro di testo

- §4.1 — §4.4
- §4.6
- §4.9
- **Nota**. La moda verrà presentata nell'unità I, i boxplot verranno nuovamente affrontati nell'unità G.

Misure (indici) di posizione

- Nell'unità precedente abbiamo visto che la distribuzioni parto prematuro / parto non prematuro differiscono soprattutto per la diversa **posizione**.
- Possiamo quantificare di quanto è più basso il DDE tra le donne con parto non prematuro?
- Vogliamo quindi **sintetizzare** le singole distribuzioni in un **unico numero** che indichi dove la distribuzione stessa è posizionata.
- Il confronto di questi indici consente di rispondere alla domanda del punto precedente
- Gli indici di posizione più "famosi" sono la **media aritmetica**, la **mediana** ed i **quantili**.
- In questo corso ne presenteremo anche altri: la **moda** (discussa nelle unità successive), la media geometrica, la media armonica, etc.

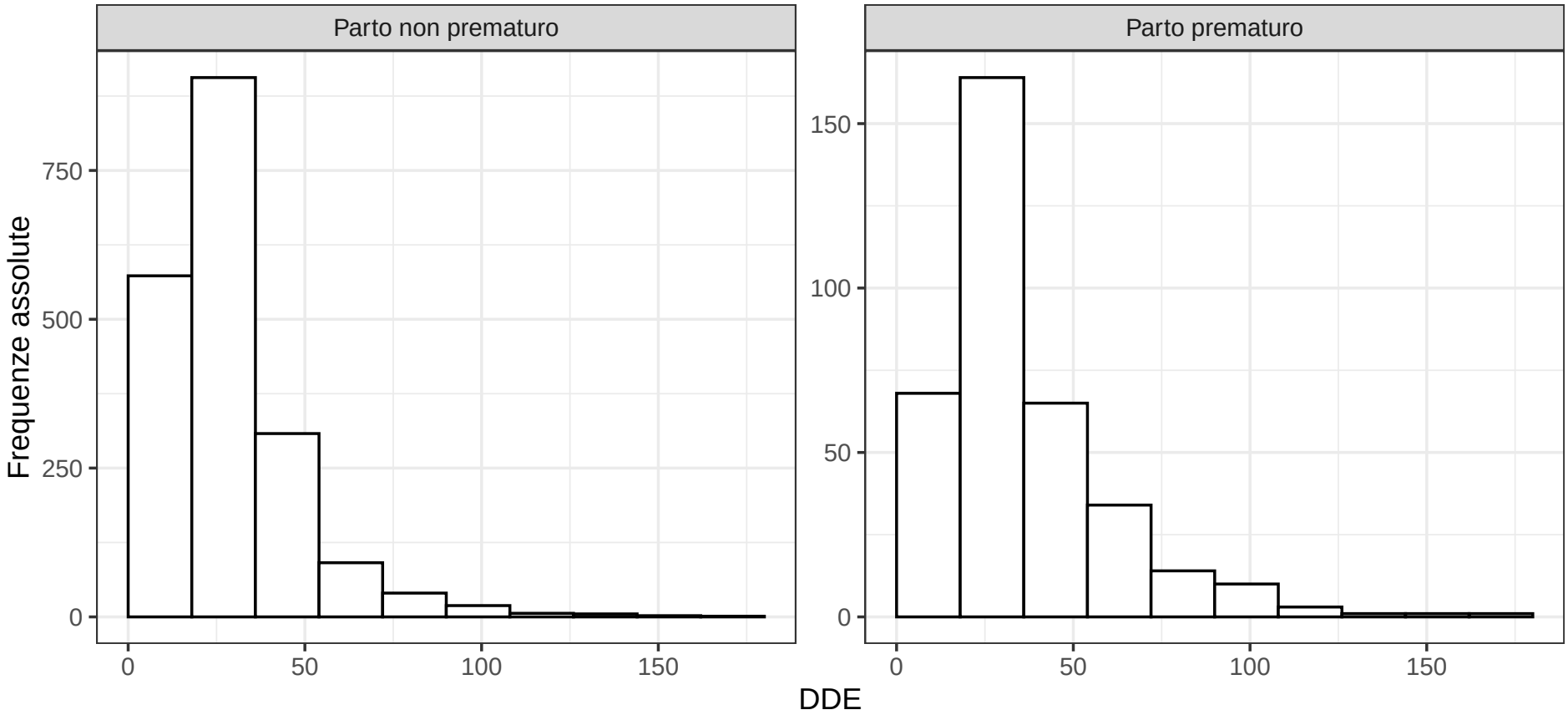
La media aritmetica

- Supponiamo di aver rilevato un certo fenomeno (esprimibile numericamente) su n unità statistiche diverse.
- Come fatto in precedenza, indichiamo con $x_1, \ldots, x_n$ i valori osservati (ovvero i dati).
- **Media aritmetica**. La media aritmetica dei dati $x_1, \ldots, x_n$ è

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

- Nonostante esistano altri tipi di "medie", quella aritmetica è senza dubbio quella più utilizzata.
- Per questa ragione, viene comunemente indicata come **la media**, senza ulteriore aggettivazione.

La media aritmetica



	Parto non prematuro	Parto prematuro
Media di DDE (mg/L)	29.142	36.203

Proprietà della media: rappresentatività

- Se i **dati** sono **tutti uguali** ad un valore a allora, anche la media è uguale ad a.
- Infatti, se

$$x_1 = x_2 = \cdots = x_n = a,$$

allora

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(a + \cdots + a) = \frac{na}{n} = a,$$

dato che lo stesso numero a viene sommato n volte.

- Questa proprietà della media viene chiamata, a volte, **rappresentatività**.
- La quasi totalità degli indici di posizione possiede questa proprietà.

Proprietà della media: internalità

- La media è sempre **compresa** tra il più **piccolo** e il più **grande** dei valori osservati.
- In simboli, si ha che

$$x_{(1)} \leq \bar{x} \leq x_{(n)},$$

dove

$$x_{(1)} = \min\{x_1, \ldots, x_n\}, \qquad x_{(n)} = \max\{x_1, \ldots, x_n\}.$$

- Infatti, per quanto riguarda la prima disuguaglianza, si ha che

$$x_{(1)} = \frac{x_{(1)} + \cdots + x_{(1)}}{n} \leq \frac{x_1 + \cdots + x_n}{n} = \bar{x}.$$

- Questa proprietà della media viene chiamata, a volte, **internalità**.
- Anche in questo caso, la maggior parte degli indici di posizione possiede questa proprietà.

Proprietà della media: associatività

- La media rimane invariata se un **sotto-insieme** di dati viene rimpiazzato con la loro **media parziale**.
- In simboli, si ha che la media $\bar{x}$ dei dati

$$x_1, \ldots, x_k, x_{k+1}, \ldots, x_n$$

coincide con la media di

$$m, \ldots, m, x_{k+1}, \ldots, x_n,$$

dove m è la media del sotto-insieme $x_1, \ldots, x_k$.

- Infatti, basta notare che

$$\frac{m + \cdots + m + x_{k+1} + \cdots + x_n}{n} = \frac{1}{n}\left(k \times \frac{1}{k}\sum_{i=1}^{k} x_i + \sum_{i=k+1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- Questa proprietà della media viene chiamata, a volte, **associatività**.

Proprietà della media: trasformazione lineare

- La media di una **trasformazione lineare** dei dati coincide con la trasformazione lineare della media.

- In altri termini, se consideriamo i dati trasformati $y_1, \ldots, y_n$, tali che

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

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

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = a + b\bar{x}.$$

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

- La dimostrazione è anche in questo caso immediata

$$\bar{y} = \frac{(a + bx_1) + \cdots + (a + bx_n)}{n} = \frac{a + \cdots + a}{n} + b\frac{x_1 + \cdots + x_n}{n} = a + b\bar{x}$$

Proprietà della media: baricentro

- La somma, e dunque la media, delle differenze dei dati dalla loro media, detti **scarti**, è sempre pari a 0.

- In altri termini,

$$\sum_{i=1}^{n}(x_i - \bar{x}) = (x_1 - \bar{x}) + \cdots + (x_n - \bar{x}) = 0.$$

- Si tratta di una conseguenza delle proprietà precedente, con $a = -\bar{x}, b = 1$. Oppure, basta notare che

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \left(\sum_{i=1}^{n} x_i\right) - n\bar{x} = n\bar{x} - n\bar{x} = 0.$$

- Questo risultato mostra che la media costituisce il **baricentro** della distribuzione di frequenza.

- Infatti, alcuni scarti $(x_i - \bar{x})$ saranno positivi, altri negativi, ed alcuni (a volte) nulli. Tali scarti si **compensano esattamente**.

Proprietà della media: scarti quadratici

Lemma A

- Sia $a \in \mathbb{R}$ un numero qualsiasi, allora

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

Infatti:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - a)^2 &= \sum_{i=1}^{n}(x_i - a + \bar{x} - \bar{x})^2 = \sum_{i=1}^{n}[(x_i - \bar{x}) + (\bar{x} - a)]^2 \\
&= \sum_{i=1}^{n}\left[(x_i - \bar{x})^2 + (\bar{x} - a)^2 + 2(x_i - \bar{x})(\bar{x} - a)\right] \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - a)^2 + 2(\bar{x} - a)\overbrace{\sum_{i=1}^{n}(x_i - \bar{x})}^{=0} \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - a)^2.
\end{aligned}$$

Proprietà della media: scarti quadratici

- Il Lemma appena descritto ha una importante conseguenza.
- La somma degli **scarti quadratici** da una costante è **minima** se e solo se la costante è posta uguale alla media.
- In simboli, si ha che

$$\bar{x} = \arg\min_{a \in \mathbb{R}} \sum_{i=1}^{n} (x_i - a)^2.$$

- Infatti

$$\sum_{i=1}^{n} (x_i - a)^2 > \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad \text{se } a \neq \bar{x},$$

poiché il secondo termine che compare nel Lemma A è strettamente positivo.

- **Esercizio**. Si dimostri questo risultato studiando la funzione $\ell(a) = \sum_{i=1}^{n} (x_i - a)^2$, ovvero considerando derivata prima e seconda di $\ell(a)$.

Proprietà della media: calcolo ricorsivo

- Si indichi con $\bar{x}_n$ la media aritmetica dei valori $x_1, \ldots, x_n$.
- Si supponga che un **nuovo dato** x_{n+1} diventi disponibile e si indichi con $\bar{x}_{n+1}$ la media aritmetica dei dati $x_1, \ldots, x_{n+1}$. Allora

$$\bar{x}_{n+1} = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}x_{n+1}.$$

Infatti, si noti che

$$\bar{x}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} x_i = \frac{n}{n+1}\frac{1}{n}\left(\sum_{i=1}^{n} x_i + x_{n+1}\right) = \frac{n}{n+1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \frac{1}{n+1}x_{n+1}$$

- Tale relazione è detta **ricorsiva** perché permette di aggiornare la media aritmetica $\bar{x}_{n+1}$ in termini di $\bar{x}_n$ e del nuovo dato x_{n+1}, senza dover rifare tutti i calcoli.
- **Esercizio**. Sapendo che $\bar{x}_9 = 26$ e che $x_{10} = 30$, si calcoli la media $\bar{x}_{10}$.

Una non-proprietà della media

- La media di una **trasformazione non lineare** dei dati **non** coincide con la trasformazione della media.

- In altri termini, se consideriamo i dati trasformati $y_1, \ldots, y_n$, tali che

$$y_i = f(x_i), \qquad i = 1, \ldots, n,$$

dove $f(x)$ è una funzione non lineare qualsiasi, allora **in generale**

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} f(x_i) \neq f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = f(\bar{x}).$$

- **Esempio**. Se $f(x) = x^2$, allora **in generale** vale che

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \neq \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2,$$

ovvero la media dei quadrati non è pari al quadrato della media.

- **Esercizio**. Si verifichi la precedente affermazione ponendo $x_1 = 0$, $x_2 = 1$ e $x_3 = 2$.

Funzioni concave e convesse

- La precedente non-proprietà può essere resa più precisa quando $f(x)$ è una funzione **convessa** oppure **concava**.

- **Funzione convessa**. Una funzione $f(x) : (a, b) \to \mathbb{R}$ si dice convessa se vale che

$$f\left(\lambda x + (1-\lambda)y\right) \leq \lambda f(x) + (1-\lambda) f(y),$$

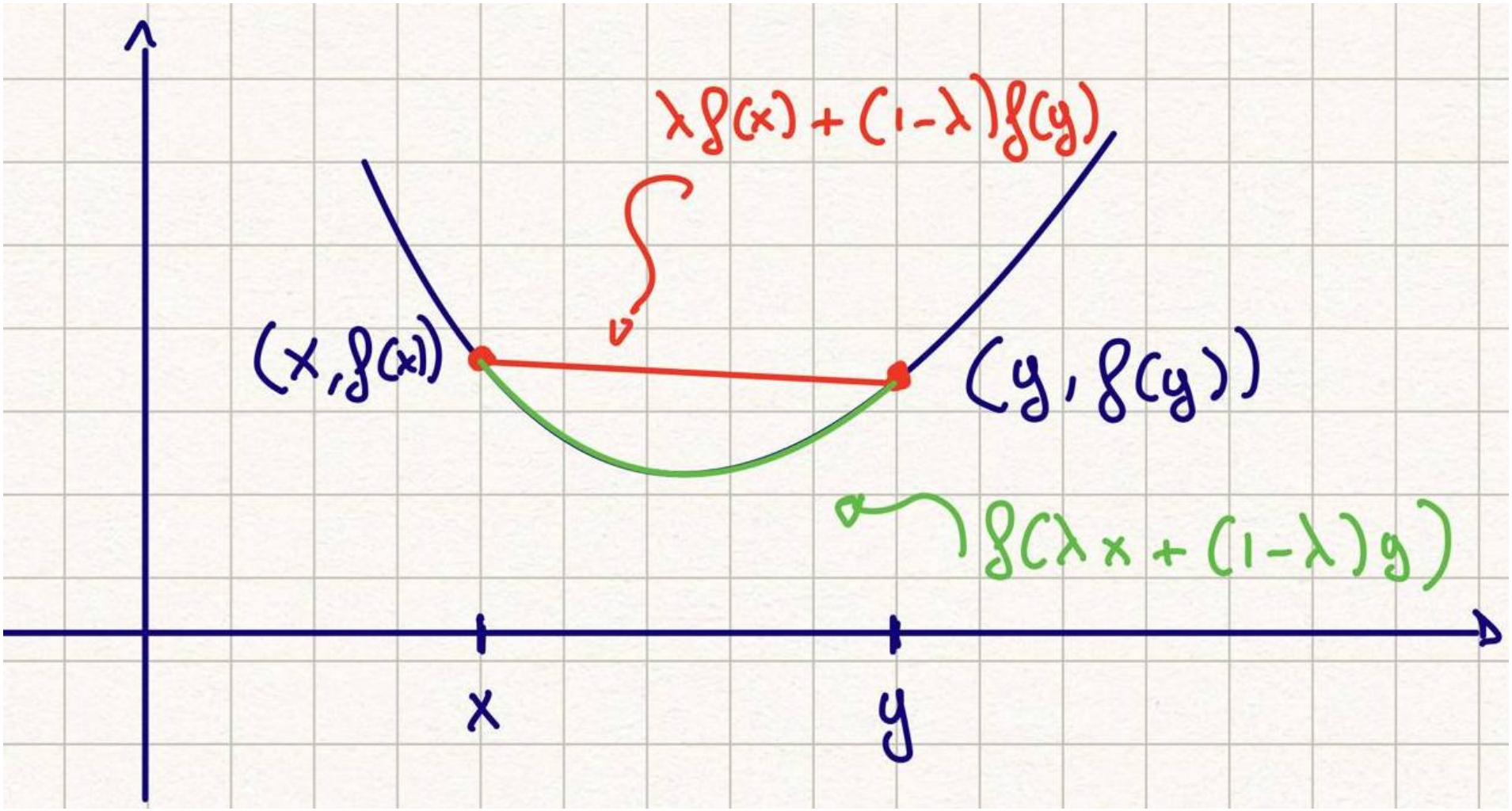
per ogni $x, y \in (a, b)$ e $0 < \lambda < 1$.

- Viceversa, si dice **concava** se la precedente disuguaglianza ha verso invertito.

- **Proprietà** (di Analisi Matematica). Se la derivata seconda $f''(x)$ esiste, allora una funzione è convessa in (a, b) se e solo se $f''(x) \geq 0$ per ogni $x \in (a, b)$.

- Viceversa è la funzione è concava se $f''(x) \leq 0$.

Funzione convessa



Disuguaglianza di Jensen

Teorema (Disuguaglianza di Jensen)

- Sia $f(x)$ una funzione **convessa** nell'intervallo (a, b) e siano $x_1, \ldots, x_n$ dei dati contenuti in tale intervallo. Allora

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \geq f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

- Il verso della disuguaglianza è invertito quando $f(x)$ è una funzione **concava**.

- **Esempio**. Se $f(x) = \log x$, allora $f(x)$ è **concava** per ogni $x > 0$ e vale che

$$\frac{1}{n}\sum_{i=1}^{n} \log x_i \leq \log\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right).$$

- **Esercizio (difficile)**. Si dimostri la disuguaglianza di Jensen tramite il principio di induzione. Si noti che se $n = 2$ la disuguaglianza è vera per definizione di convessità.

Un difetto della media

- Alcuni insiemi di dati possono contenere una frazione di osservazioni **anomale** o **atipiche**.
- Si tratta di osservazioni che sembrano provenire da una popolazione diversa o essere state generate da un meccanismo differente.
- Nel caso più banale, potrebbe perfino trattarsi di errori di trascrizione.
- In una situazione del tipo descritto, bisogna tenere presente che la media aritmetica è **molto sensibile** alla presenza di **osservazioni anomale** e può fornire risultati fuorvianti.
- Come è facile capire dalla definizione stessa, **una sola osservazione** molto grande o molto piccola **può dominare** il valore assunto dalla media.

Un difetto della media

- **Esercizio**. Si supponga di avere $n = 10.000$ osservazioni $x_1, \ldots, x_n$ tali che $x_i \in (0, 1)$ per ogni $i = 2, \ldots, 10.000$.
- In altri termini, tutte le osservazioni eccetto la prima sono comprese tra 0 e 1.
- Si mostri che

$$\lim_{x_1 \to -\infty} \frac{1}{n} \sum_{i=1}^{n} x_i = -\infty.$$

- Si commenti il risultato.

La media aritmetica ponderata

- Nella definizione di media aritmetica, le unità statistiche concorrono "alla pari" nella determinazione della media.
- Esistono tuttavia delle situazioni in cui tale approccio non è adeguato.
- Ad esempio, la media dei voti universitari è **pesata** tramite i crediti formativi (CFU).
- Se a ciascuna unità statistica x_i è associato un **peso numerico** w_i, allora si può usare la media aritmetica ponderata.
- **Media aritmetica ponderata**. La media aritmetica dei dati $x_1, \ldots, x_n$ con pesi $w_1, \ldots, w_n$ è

$$\bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i'=1}^n w_{i'}} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n} = \sum_{i=1}^n \tilde{w}_i x_i,$$

dove $\tilde{w}_i = w_i / \sum_{i'=1}^n w_{i'}$ sono i **pesi standardizzati**.

Dati raggruppati: approssimazione della media

Classi	$(a_0, a_1]$	$(a_1, a_2]$	$(a_2, a_3]$	$\ldots$	$(a_{k-1}, a_k]$
Frequenze assolute	n_1	n_2	n_3	$\ldots$	n_k

- Supponiamo di non conoscere i dati individuali ma solo una distribuzione di frequenza per intervalli, come nell'esempio sopra.
- La media **non si può calcolare** esattamente.
- Un'**approssimazione** che viene spesso usata in questi casi è

$$\bar{x} \approx \frac{1}{n} \sum_{j=1}^{k} n_j m_j = \sum_{j=1}^{k} f_j m_j, \qquad f_j = n_j / n,$$

dove m_j è il punto centrale dell'intervallo j-esimo, ovvero

$$m_j = \frac{a_{j-1} + a_j}{2}, \qquad j = 1, \ldots, k.$$

- Questa approssimazione è una **media aritmetica ponderata**, con pesi $w_j = n_j$.

Dati raggruppati: proprietà della media

- **Esercizio**. Si dica quale delle due seguenti affermazioni è ragionevolmente corretta
  - Più gli intervalli sono grandi (lunghi) più l'approssimazione è accurata.
  - Più gli intervalli sono piccoli (corti) più l'approssimazione è accurata.

- **Esercizio - proprietà**. Si dimostri che l'approssimazione del punto precedente coincide con la media aritmetica, ovvero

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j m_j = \sum_{j=1}^{k} f_j m_j, \qquad f_j = n_j/n,$$

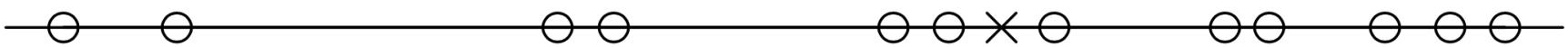
quando tutte le osservazioni nell'intervallo j-esimo sono uguali a m_j (variabili discrete).

- **Esercizio - proprietà**. Più in generale, si dimostri che l'identità $\bar{x} = 1/n \sum_{j=1}^{k} n_j m_j$ vale anche nel caso in cui m_j sia pari alla media aritmetica delle osservazioni contenute nell'intervallo j-esimo, ovvero se

$$m_j = \frac{1}{n_j} \sum_{i\,:\,x_i \in (a_{j-1}, a_j]} x_i, \qquad j = 1, \ldots, k.$$

La mediana

- L'idea alla base della **mediana** è trovare quel numero che sia più grande di **circa** il 50% delle osservazioni e più piccolo della restante parte.
- Nel grafico seguente, le osservazioni $x_1, \ldots, x_{13}$ corrispondono ai punti disegnati con un cerchio. La mediana invece è stata contrassegnata con una croce.
- La mediana lascia sia a sinistra che a destra 6 osservazioni.



La mediana

- **Mediana**. Siano $x_1, \ldots, x_n$ un insieme di dati e siano $x_{(1)}, \ldots, x_{(n)}$ le **osservazioni ordinate**. La mediana è quindi pari a

$$\text{Me} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ è dispari,} \\ \left(x_{(n/2)} + x_{(n/2+1)}\right)/2, & \text{se } n \text{ è pari.} \end{cases}$$

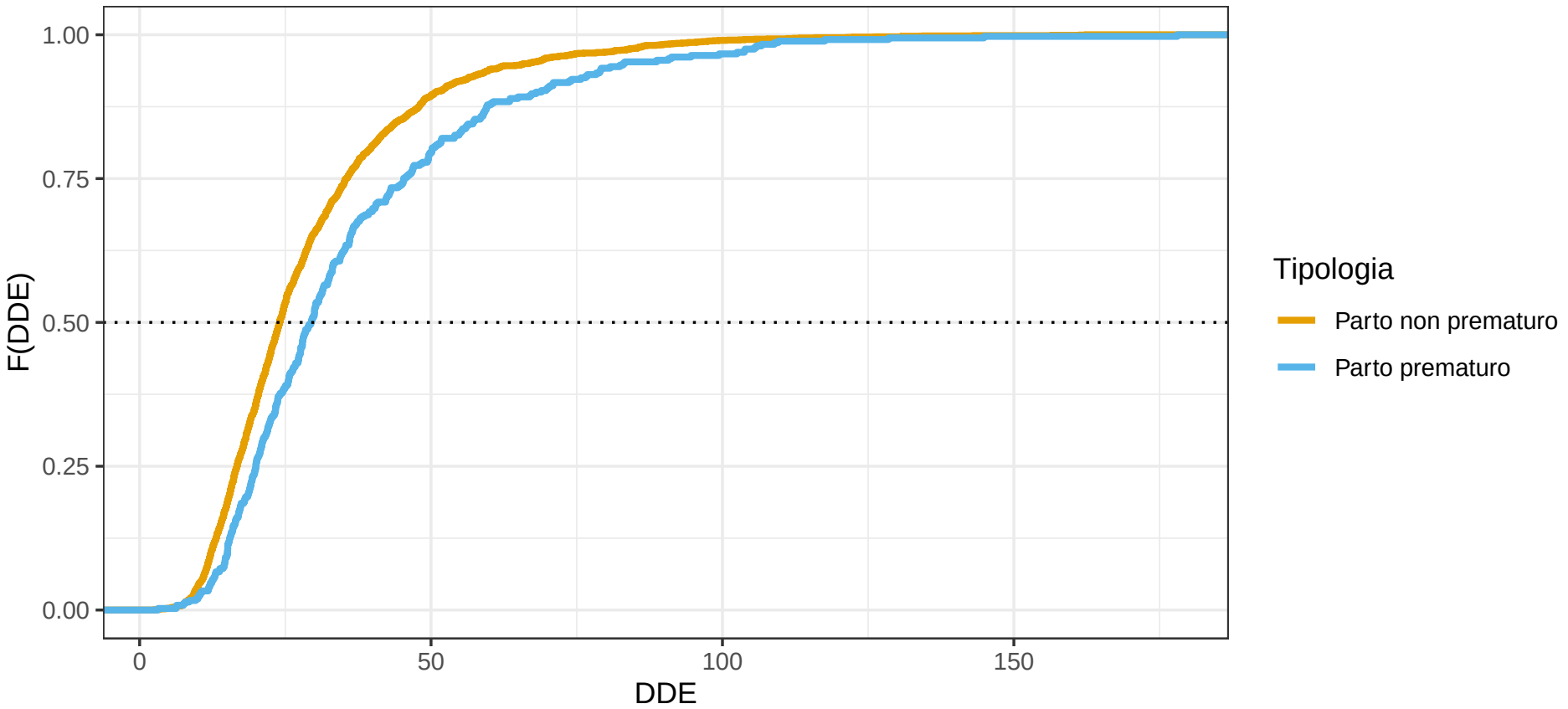
- La mediana è quindi il valore centrale dei dati ordinati se questi sono in numero dispari, mentre è la media dei due valori centrali quando i dati sono in numero pari.

- Ricordando la definizione della funzione di ripartizione empirica, si nota che la mediana è un valore tale per cui

$$F(\text{Me}) \approx \frac{1}{2}.$$

- **Nota**. Quando n è pari, la media dei due valori centrali è una **scelta convenzionale**, anche se largamente condivisa. In realtà, qualsiasi valore compreso nell'intervallo $[x_{(n/2)}, x_{(n/2+1)}]$ sarebbe accettabile (ci torniamo poi!).

La mediana



	Parto non prematuro	Parto prematuro
Media di DDE (mg/L)	29.14	36.20
Mediana di DDE (mg/L)	24.04	29.46

Esempi di calcolo della mediana

- **Dati** $x_1, \ldots, x_5$: `1, 4, 2, 9, 3`.
- **Dati ordinati** $x_{(1)}, \ldots, x_{(5)}$: `1, 2, 3, 4, 9`.

- In questo caso $n = 5$ è dispari. Secondo la definizione di mediana data in precedenza si ottiene:

$$\text{Me} = x_{(3)} = 3.$$

- Tuttavia, non esiste un numero che lascia esattamente il 50% delle osservazioni alla sua sinistra. Infatti, in questo caso si ha che

$$F(3) = \frac{3}{5} = 0.6,$$

ricordando che $F(x)$ è la funzione di ripartizione empirica.

Esempi di calcolo della mediana

- **Dati** $x_1, \ldots, x_4$: `1, 2, 1, 5`.
- **Dati ordinati** $x_{(1)}, \ldots, x_{(4)}$: `1, 1, 2, 5`.

- Poiché $n = 4$ è pari, otteniamo

$$\text{Me} = \frac{1}{2}\left(x_{(2)} + x_{(3)}\right) = (1+2)/2 = 1.5.$$

- In questo caso, la mediana lascia esattamente il 50% delle osservazioni alla sua sinistra. Infatti

$$F(1.5) = \frac{2}{4} = 0.5.$$

- Tuttavia, qualsiasi numero compreso tra 1 e 2 godrebbe della stessa proprietà, ad esempio $F(1.7) = 0.5$.

- Il valore 1.7 è una "mediana alternativa" rispetto a quella convenzionalmente usata. Ad ogni modo, le differenze tra la mediana "canonica" e quelle alternative sono spesso trascurabili in pratica.

Esempi di calcolo della mediana

- **Dati** $x_1, \ldots, x_{10}$: 4, 3, 2, 2, 5, 2, 6, 5, 1, 3.
- **Dati ordinati** $x_{(1)}, \ldots, x_{(10)}$: 1, 2, 2, 2, 3, 3, 4, 5, 5, 6.

- Poiché $n = 10$ è pari, otteniamo

$$\text{Me} = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = (3+3)/2 = 3.$$

- In questo caso non ci sono ambiguità: il valore 3 è l'unico ammissibile.
- Tuttavia, il valore 3 **non** lascia esattamente il 50% delle osservazioni alla sua sinistra. Infatti

$$F(3) = \frac{6}{10} = 0.6.$$

Esempi di calcolo della mediana

- Supponiamo ora di avere i seguenti dati **raggruppati**

Classi	$(0, 1]$	$(1, 2]$	$(2, 3]$	$(3, 4]$	$(4, 5]$
Frequenze assolute	1	4	4	2	1

- In totale abbiamo $n = 12$ osservazioni, pertanto la mediana dovrebbe essere pari a

$$\text{Me} = \frac{x_{(6)} + x_{(7)}}{2}.$$

- Dalla tabella è chiaro che entrambi i valori $x_{(6)}$ e $x_{(7)}$ appartengono all'intervallo $(2, 3]$, quindi necessariamente anche la mediana appartiene a questo intervallo.

- Tuttavia, il calcolo preciso della mediana è **impossibile**.

- In pratica, potrebbe essere conveniente scegliere un valore preciso compreso tra 2 e 3, ma questo sarebbe necessariamente una **scelta arbitraria**.

Dati raggruppati: approssimazione della mediana

Classi	$(a_0, a_1]$	$(a_1, a_2]$	$(a_2, a_3]$	$\dots$	$(a_{k-1}, a_k]$
Frequenze assolute	n_1	n_2	n_3	$\dots$	n_k

- Sebbene il calcolo preciso della mediana non sia possibile in caso di dati raggruppati, una **scelta ragionevole** è la seguente **approssimazione lineare**

$$\text{Me} \approx a_{j-1} + (a_j - a_{j-1}) \frac{1/2 - F(a_{j-1})}{F(a_j) - F(a_{j-1})},$$

  in cui a_{j-1} e a_j sono gli estremi dell'intervallo a cui la mediana appartiene ed $F(x)$ è la funzione di ripartizione.

- Tale formula si ottiene considerando la retta $y = ax + b$ passante per i punti $(a_{j-1}, F(a_{j-1}))$ e $(a_j, F(a_j))$. L'approssimazione della mediana sarà quel valore per cui

$$\frac{1}{2} \approx a \times \text{Me} + b, \qquad \text{ovvero} \qquad \text{Me} \approx \frac{1}{a}(1/2 - b).$$

- **Esercizio**. Calcolare esplicitamente i valori di a e di b e verificare la formula.

Esempi di calcolo della mediana

- Supponiamo di avere i seguenti dati **raggruppati**

Classi	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
Frequenze assolute	1	4	4	2	1

- Dalla tabella è chiaro che la mediana appartiene all'intervallo $(2,3]$, ovvero $a_{j-1} = 2$ e $a_j = 3$.

- Inoltre, si verifichi che

$$F(2) = \frac{5}{12} = 0.42 < 0.5, \qquad F(3) = \frac{9}{12} = 0.75 > 0.5.$$

- Utilizzando l'approssimazione lineare per la mediana, si ottiene che

$$\text{Me} \approx a_{j-1} + (a_j - a_{j-1}) \frac{1/2 - F(a_{j-1})}{F(a_j) - F(a_{j-1})} = 2 + (3-2) \frac{1/2 - 5/12}{9/12 - 5/12} = 2.25.$$

Proprietà della mediana

- È semplice mostrare che anche la mediana soddisfa il criterio di **rappresentatività**, ovvero se i dati sono tutti uguali a una costante a, allora $\text{Me} = a$.
- Inoltre, la mediana soddisfa il criterio di **internalità**, ovvero è sempre compresa tra il valore minimo ed il valore massimo. Questo fatto è piuttosto ovvio dalla definizione.

- La mediana è particolarmente apprezzata soprattutto perché è **poco sensibile** alla presenza di **valori anomali**.
- Per questa ragione, si dice che la mediana è **resistente**.
- **Esercizio**. Come mai questo succede? Supponendo che $x_1, \ldots, x_n$ siano tutti compresi tra 0 e 1, con $n = 10.000$. Cosa succederebbe alla mediana se al posto del valore di x_1 sostituissi $x_1 = 10^{2020}$? Cosa succederebbe invece alla media aritmetica?

Proprietà della mediana

- La somma degli **scarti in valore assoluto** da una costante è minima se (ma non solo se) tale costante è posta uguale alla mediana.
- In simboli, si ha che

$$\sum_{i=1}^{n} |x_i - \text{Me}| = \min_{a \in \mathbb{R}} \sum_{i=1}^{n} |x_i - a|$$

- Attenzione ai dettagli. Quando n è pari, infatti, esistono **infiniti valori** che rendono minimi gli scarti in valore assoluto. In altri termini, esistono delle "mediane alternative".
- In particolare, qualsiasi costante a compresa nell'intervallo $[x_{(n/2)}, x_{(n/2+1)}]$ minimizza $\sum_{i=1}^{n} |x_i - a|$. Infatti, la somma degli scarti assoluti è sempre la stessa.
- Quando n è dispari, invece, esiste un **unico valore** che minimizza $\sum_{i=1}^{n} |x_i - a|$, ovvero il valore in posizione centrale dei dati ordinati.

Dimostrazione per $n = 2$

- A fini illustrativi, dimostriamo questa proprietà nel caso $n = 2$. Supponiamo quindi di avere due osservazioni tali che $x_2 > x_1$. La dimostrazione nel caso $x_1 = x_2$ è ovvia.

- **Caso 1**. Si noti che per **qualsiasi valore** $a \in [x_1, x_2]$ allora

$$\sum_{i=1}^{2} |x_i - a| = (a - x_1) + (x_2 - a) = x_2 - x_1.$$

- **Caso 2**. Viceversa, se $a < x_1$ allora

$$\sum_{i=1}^{2} |x_i - a| = (x_1 - a) + (x_2 - a) = x_1 + x_2 - 2a > x_1 + x_2 - 2x_1 = x_2 - x_1.$$

- **Caso 3**. Infine, se $a > x_2$ allora

$$\sum_{i=1}^{2} |x_i - a| = (a - x_1) + (a - x_2) = 2a - x_1 - x_2 > 2x_2 - x_1 - x_2 = x_2 - x_1.$$

Proprietà della mediana: trasformazione monotona

- La mediana di una trasformazione che preserva l'ordinamento dei dati, detta trasformazione **monotona crescente**, coincide con la trasformazione della mediana.
- In simboli, se $f(x)$ è una trasformazione monotona e poniamo

$$y_i = f(x_i), \qquad i = 1, \ldots, n,$$

allora la mediana di $y_1, \ldots, y_n$ coincide con $f(\text{“mediana di } x_1, \ldots, x_n\text{”})$.

- **Esempio 1**. Alcune trasformazioni lineari $f(x) = a + bx$ sono trasformazioni monotone crescenti, ovvero quelle in cui $b \geq 0$.
- **Esempio 2**. Se $f(x) = x^2$ e i dati sono non-negativi, allora $f(x)$ è una trasformazione monotona crescente e la mediana di $y_1, \ldots, y_n$ coincide con $f(\text{“mediana di } x_1, \ldots, x_n\text{”})$.
- **Esercizio**. Si verifichi la precedente affermazione ponendo $x_1 = 0$, $x_2 = 3$ e $x_3 = 5$.

I quantili

- I **quantili** generalizzano la mediana.
- L'idea alla base di un **quantile-p**, dove $p \in (0, 1)$, è trovare quel numero che sia più grande di **circa** il $100 \times p\%$ delle osservazioni e più piccolo della restante parte.
- Ad esempio, un quantile-0.1 è il valore che lascia a sinistra circa il 10% delle osservazioni.
- I quantili con p uguale a 0.25, 0.5, 0.75 vengono spesso chiamati, rispettivamente, il primo, il secondo ed il terzo **quartile**. Dividono la popolazione in quattro parti uguali.
- **Nota**. Il secondo quartile coincide con la mediana.
- I quantili con $p = 0.1, \ldots, 0.9$ si chiamano **decili** mentre quelli con $p = 0.01, \ldots, 0.99$ si chiamano **percentili**.

I quantili

- **Quantile-p**. Siano $x_1, \ldots, x_n$ un insieme di dati, sia $p \in (0,1)$ e sia $F(x)$ la funzione di ripartizione empirica. Il quantile-p è quindi pari a

$$\mathcal{Q}_p = \inf\{x : F(x) \geq p\}.$$

- In altri termini, il quantile-p è il **più piccolo** valore che è **più grande** di **almeno** il $100 \times p\%$ dei dati.

- Per definizione, si ha che $F(\mathcal{Q}_p) \geq p$, anche se spesso in pratica vale che

$$F(\mathcal{Q}_p) \approx p.$$

- **Esercizio: ambiguità dei quantili**. Si verifichi che

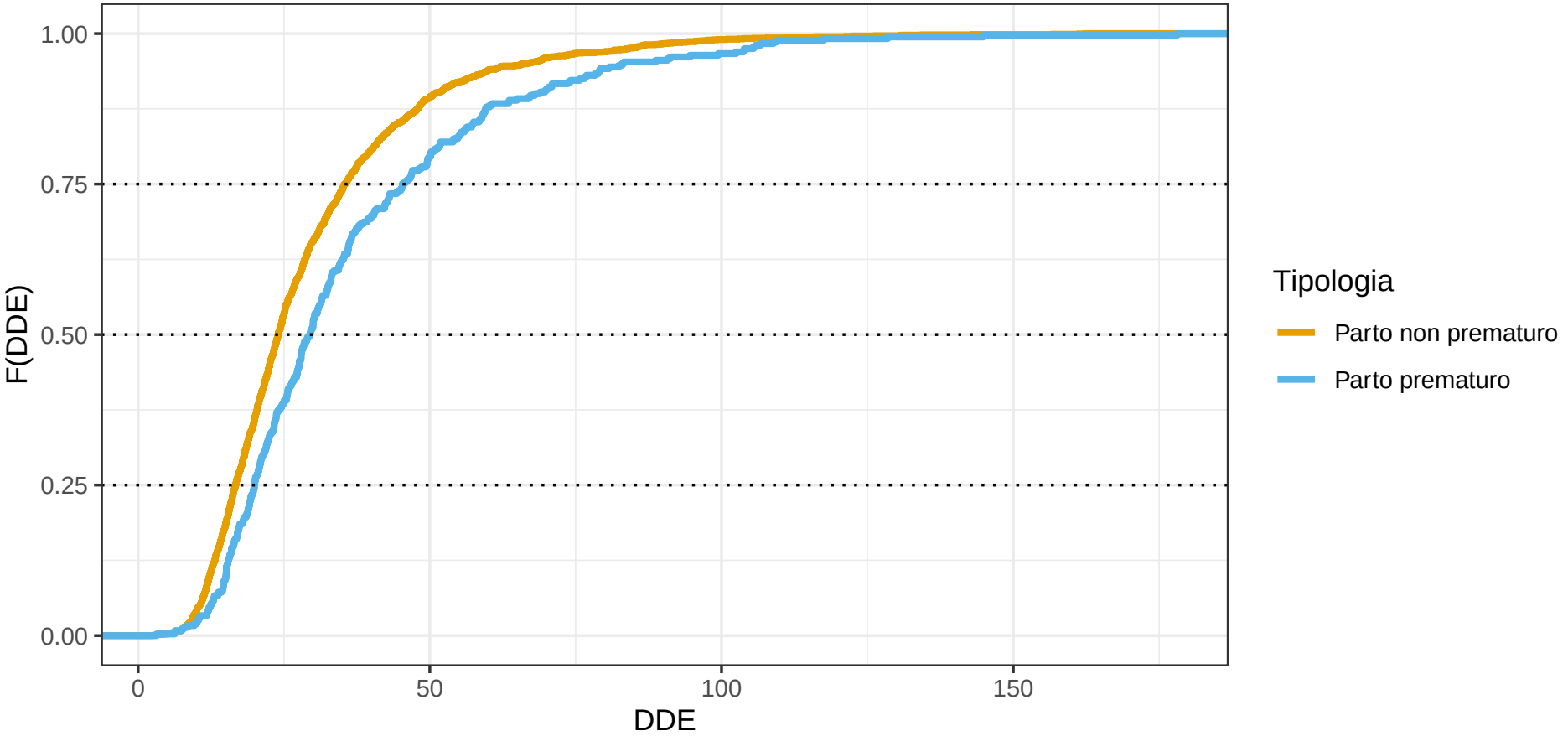
$$\mathcal{Q}_{0.5} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{se } n \text{ è dispari,} \\ x_{(n/2)}, & \text{se } n \text{ è pari.} \end{cases}$$

Si confronti $\mathcal{Q}_{0.5}$ con la mediana Me e si commenti.

Ambiguità nel calcolo dei quantili

- **Nota**. Esattamente come per la mediana, questa definizione di quantili-p è una **scelta convenzionale**.
- Differenti software e libri di testo usano diverse convenzioni! Ad esempio, il nostro libro di testo **non** usa la nostra stessa definizione.
- Ci sono infatti delle ambiguità nel calcolo dei quantili, dovute ad esempio al fatto che esistono infiniti valori più grandi del $100 \times p\%$ dei dati.
- Purtroppo, diversamente dalla mediana, non esiste una definizione di quantile-p "canonica" e largamente accettata da tutti.
- Esistono definizioni alternative per i quantili-p tali che la mediana è **sempre** pari al secondo quartile.
- Nonostante esistano numerose alternative, le differenze diventano sostanzialmente **trascurabili** quando n è sufficientemente grande, perché i dati diventano **più addensati**.

Quartili di DDE



	Parto non prematuro	Parto prematuro
Primo quartile di DDE (mg/L)	16.73	19.94
Secondo quartile di DDE (mg/L)	24.04	29.46
Terzo quartile di DDE (mg/L)	35.45	45.30

Quartili di DDE

Definizione alternativa I (default nel software R)

	Parto non prematuro	Parto prematuro
Primo quartile di DDE (mg/L)	16.735	19.94
Secondo quartile di DDE (mg/L)	24.040	29.46
Terzo quartile di DDE (mg/L)	35.400	45.30

Definizione alternativa II (default nel software SAS)

	Parto non prematuro	Parto prematuro
Primo quartile di DDE (mg/L)	16.73	19.94
Secondo quartile di DDE (mg/L)	24.04	29.31
Terzo quartile di DDE (mg/L)	35.35	45.30

Esempi di calcolo dei quantili

- **Dati ordinati**: `6.4, 6.7, 6.8, 7.0, 7.3, 7.5, 7.6, 7.7, 7.9, 8.1`.

- Il primo quartile è il più piccolo valore che sia più grande di almeno il 25% dei dati. Pertanto, abbiamo

$$\mathcal{Q}_{0.25} = x_{(3)} = 6.8,$$

dato che $F(6.8) = 0.3 > 0.25$. Questo è effettivamente il valore più piccolo possibile, infatti ad esempio $F(6.75) = 0.2 < 0.25$.

- **Ambiguità dei quantili**. Le mediana Me in questo caso è il valore medio tra 7.3 e 7.5, ovvero Me $= 7.4$. Tuttavia, il secondo quartile è leggermente diverso:

$$\mathcal{Q}_{0.5} = x_{(5)} = 7.3.$$

Questa differenza è tuttavia abbastanza trascurabile ai fini pratici.

Esempi di calcolo dei quantili

- **Dati** $x_1, \ldots, x_{10}$: `4, 3, 2, 2, 5, 2, 6, 5, 1, 3`.
- **Dati ordinati** $x_{(1)}, \ldots, x_{(10)}$: `1, 2, 2, 2, 3, 3, 4, 5, 5, 6`.

- Poiché $n = 10$, allora il terzo quartile, ovvero il quantile-0.75, è pari a

$$\mathcal{Q}_{0.75} = x_{(8)} = 5.$$

- Infatti, $F(5) = 0.9 > 0.75$. Tuttavia ad esempio $F(4.9) = 0.7 < 0.75$. In altri termini, il valore 5 è il più piccolo valore più grande di almeno il 75% dei dati.

- **Ambiguità dei quantili**. In questo caso particolare, il secondo quartile $\mathcal{Q}_{0.5} = 3 = \text{Me}$ coincide con la mediana Me . Questo accade perché i due valori centrali $x_{(5)}$ e $x_{(6)}$ sono uguali.

Dati raggruppati: approssimazione dei quantili

Classi	$(a_0, a_1]$	$(a_1, a_2]$	$(a_2, a_3]$	$\ldots$	$(a_{k-1}, a_k]$
Frequenze assolute	n_1	n_2	n_3	$\ldots$	n_k

- Sebbene il calcolo preciso dei quantili non sia possibile in caso di dati raggruppati, una **scelta ragionevole** è la seguente **approssimazione lineare**

$$\mathcal{Q}_p \approx a_{j-1} + (a_j - a_{j-1}) \frac{p - F(a_{j-1})}{F(a_j) - F(a_{j-1})},$$

in cui a_{j-1} e a_j sono gli estremi dell'intervallo a cui il quantile-p appartiene ed $F(x)$ è la funzione di ripartizione.

- Tale formula si ottiene considerando la retta $y = ax + b$ passante per i punti $(a_{j-1}, F(a_{j-1}))$ e $(a_j, F(a_j))$. L'approssimazione del quantile-p sarà quel valore per cui

$$p \approx a \times \mathcal{Q}_p + b, \qquad \text{ovvero} \qquad \mathcal{Q}_p \approx \frac{1}{a}(p - b).$$

Esempi di calcolo dei quantili

- Supponiamo di avere i seguenti dati **raggruppati**

Classi	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
Frequenze assolute	1	4	4	2	1

- Dalla tabella è chiaro che il quantile-0.25, ovvero $\mathcal{Q}_{0.25}$, appartiene all'intervallo $(1,2]$, cioè $a_{j-1} = 1$ e $a_j = 2$.

- Inoltre, si verifichi che

$$F(1) = \frac{1}{12} = 0.09 < 0.25, \qquad F(2) = \frac{5}{12} = 0.45 > 0.25.$$

- Utilizzando l'approssimazione lineare per i quantili, si ottiene che

$$\mathcal{Q}_{0.25} \approx a_{j-1} + (a_j - a_{j-1}) \frac{1/4 - F(a_{j-1})}{F(a_j) - F(a_{j-1})} = 1 + (2-1) \frac{1/4 - 1/12}{5/12 - 1/12} = 1.5.$$

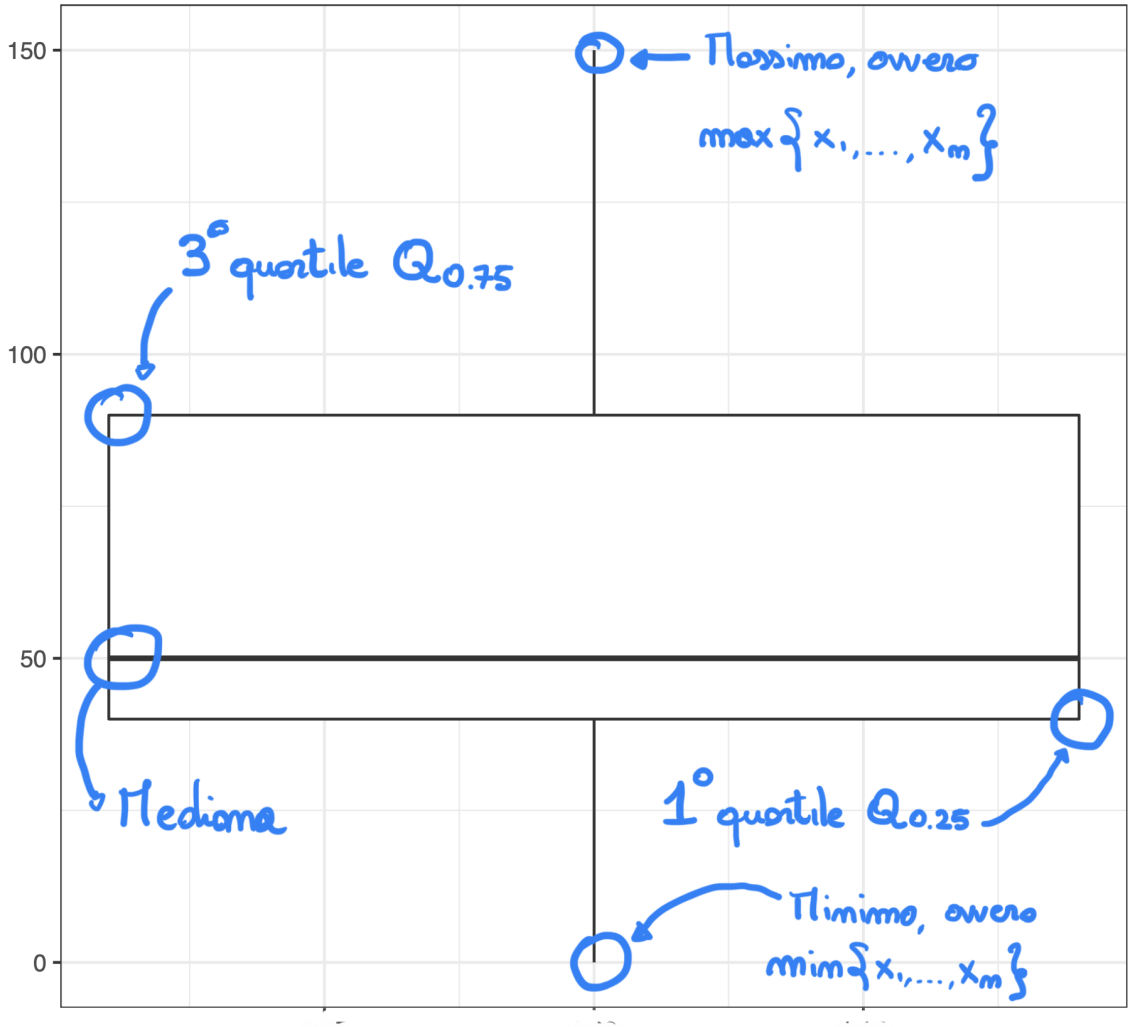
Un inciso: indici per dati raggruppati

Classi	$(a_0, a_1]$	$(a_1, a_2]$	$(a_2, a_3]$	$\ldots$	$(a_{k-1}, a_k]$
Frequenze assolute	n_1	n_2	n_3	$\ldots$	n_k

- Abbiamo fornito fino ad ora delle possibili **approssimazioni** per media, mediana e quantili in presenza di **dati raggruppati**, ovvero rappresentati come in tabella.
- Una **strategia generale** consiste nel **supporre** che i dati appartenenti alla classe $(a_{j-1}, a_j]$ siano pari al valore centrale $m_j = (a_{j-1} + a_j)/2$ e poi procedere normalmente.
- Per evitare di presentare ogni indice due volte (definizione & approssimazione), ci asterremo dal presentare le approssimazioni dei futuri indici che considereremo.
- Tuttavia, nelle applicazioni statistiche moderne, sono **rari** (anche se ne esistono!) i contesti in cui i dati sono forniti allo statistico già raggruppati in classi. Viceversa, è spesso lo statistico stesso a creare un raggruppamento per poterli meglio descrivere.
- Se i dati originali sono disponibili, gli indici possono essere calcolati esattamente, senza dover ricorrere ad approssimazioni.

I boxplot

- I diagrammi a **scatola con baffi**, chiamati **boxplot**, sono una sorta di istogramma semplificato basato sui quantili.



Le medie di Bonferroni

- Un terzo indicatore di posizione molto utilizzato è la **moda**, che discuteremo più avanti.
- Sebbene media, mediana e quantili siano gli indicatori più diffusi, esistono molte altre definizioni di media.
- Presentiamo nel seguito una **generalizzazione** della **media aritmetica**, introdotta da Bonferroni, e in seguito studiate da Nagumo, Kolomogorov e de Finetti.
- **Medie di Bonferroni**. Sia $f(x)$ una funzione **continua** e **strettamente monotona** nell'intervallo $[a, b]$ e siano $x_1, \ldots, x_n$ dei dati contenuti in tale intervallo. Allora, una media $\mathbb{M}$ dei dati $x_1, \ldots, x_n$ è pari a

$$\mathbb{M} = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right),$$

  dove $f^{-1}(x)$ è la funzione inversa di $f(x)$.
- La media aritmetica è una media di Bonferroni, ponendo $f(x) = x$.

La media geometrica

- Uno esempio notevole di media di Bonferroni si ottiene ponendo $f(x) = \log x$.
- **Media geometrica**. La media geometrica dei dati strettamente positivi $x_1, \ldots, x_n$ è

$$\mathbb{G} = \sqrt[n]{x_1 x_2 \cdots x_n} = \left(\prod_{i=1}^{n} x_i \right)^{1/n} = \exp \left\{ \frac{1}{n} \sum_{i=1}^{n} \log x_i \right\}.$$

- Per dati strettamente positivi $x_1, \ldots, x_n$ vale che

$$\mathbb{G} \leq \bar{x},$$

ovvero la media geometrica è sempre minore o uguale della media aritmetica.

- **Esercizio**. Dimostrare la disuguaglianza precedente. Suggerimento: si veda la slide sulla disuguaglianza di Jensen.

La media armonica

- Uno secondo esempio di media di Bonferroni si ottiene ponendo $f(x) = 1/x$.
- **Media armonica**. La media armonica dei dati strettamente positivi $x_1, \ldots, x_n$ è

$$\mathbb{A} = \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}.$$

- Per dati strettamente positivi $x_1, \ldots, x_n$ vale che

$$\mathbb{A} \leq \mathbb{G} \leq \bar{x},$$

ovvero la media armonica è sempre minore o uguale della media geometrica.

Proprietà delle medie di Bonferroni: rappresentatività

- Se i **dati** sono **tutti uguali** ad un valore a, allora anche la media di Bonferroni è uguale ad a.

- Infatti, se

$$x_1 = x_2 = \cdots = x_n = a,$$

allora

$$\mathbb{M} = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right) = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(a)\right) = f^{-1}\left(\frac{nf(a)}{n}\right) = a,$$

dato che per definizione $f^{-1}(f(x)) = x$.

- Pertanto, tutte le medie di Bonferroni rispettano il criterio della **rappresentatività**.

Proprietà della medie di Bonferroni: internalità

- La media di Bonferroni è sempre **compresa** tra il più **piccolo** e il più **grande** dei valori osservati.

- In simboli, si ha che

$$x_{(1)} \leq \mathbb{M} \leq x_{(n)}.$$

- Infatti, per quanto riguarda la prima disuguaglianza, si noti che

$$\frac{f(x_{(1)}) + \cdots + f(x_{(1)})}{n} \leq \frac{f(x_1) + \cdots + f(x_n)}{n},$$

dato che $f(x)$ è una funzione monotona. Poiché anche $f^{-1}(x)$ è una funzione monotona, possiamo scrivere

$$x_{(1)} = f^{-1}\left(\frac{f(x_{(1)}) + \cdots + f(x_{(1)})}{n}\right) \leq f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right) = \mathbb{M}.$$

- Pertanto, tutte le medie di Bonferroni rispettano il criterio di **internalità**.

Proprietà delle medie di Bonferroni: associatività

- La media di Bonferroni rimane invariata se un **sotto-insieme** di dati viene rimpiazzato con la loro **media parziale**.
- In simboli, si ha che la media di Bonferroni $\mathbb{M}$ dei dati

$$x_1, \ldots, x_k, x_{k+1}, \ldots, x_n$$

coincide con la media di Bonferroni di

$$m, \ldots, m, x_{k+1}, \ldots, x_n,$$

dove m è la media di Bonferroni del sotto-insieme $x_1, \ldots, x_k$.

- La dimostrazione è sostanzialmente identica a quella della media aritmetica.
- Pertanto, tutte le medie di Bonferroni rispettano il criterio di **associatività**.

Le medie di Bonferroni ponderate

- Se alle osservazioni sono associate dei pesi $w_1, \ldots, w_n$, esiste una naturale estensione delle medie di Bonferroni.

- **Medie di Bonferroni ponderate**. Sia $f(x)$ una funzione **continua** e **strettamente monotona** nell'intervallo $[a, b]$ e siano $x_1, \ldots, x_n$ dei dati contenuti in tale intervallo. Allora, una media ponderata $\mathbb{M}_w$ dei dati $x_1, \ldots, x_n$ con pesi $w_1, \ldots, w_n$ è pari a

$$\mathbb{M}_w = f^{-1}\left(\frac{\sum_{i=1}^n w_i f(x_i)}{\sum_{i'=1}^n w_{i'}}\right) = f^{-1}\left(\sum_{i=1}^n \tilde{w}_i f(x_i)\right),$$

  dove $\tilde{w}_i = w_i / \sum_{i'=1}^n w_{i'}$ sono i **pesi standardizzati**

- La media aritmetica ponderata è una media di Bonferroni ponderata, ponendo $f(x) = x$.

Le medie di Chisini

- Esistono anche altri criteri per definire una media $\mathbb{M}$. Uno in particolare è descritto da Chisini, basato sull'idea di **trasferibilità**.

- **Medie di Chisini**. Sia $g(x_1, \ldots, x_n)$ una funzione dei dati $x_1, \ldots, x_n$. Una media $\mathbb{M}$ dei dati $x_1, \ldots, x_n$ secondo Chisini è pari a quel valore compreso tra $x_{(1)}$ e $x_{(n)}$ tale che

$$g(x_1, \ldots, x_n) = g(\mathbb{M}, \ldots, \mathbb{M}).$$

- La media secondo Chisini è quel valore che non altera il valore della **funzione di sintesi** $g(\cdot)$ quando si sostituisce alle osservazioni il valore costante $\mathbb{M}$.

- Le medie di Bonferroni sono medie secondo Chisini, in cui $g(x_1, \ldots, x_n) = \sum_{i=1}^{n} f(x_i)$.

- Nel caso della media aritmetica la funzione aggregatrice è $g(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$.

- In altri termini, la media aritmetica è quel singolo valore che sostituito alle osservazioni ne lascia inalterata la somma.

Le medie di Wald

- Una media secondo Wald è quel valore $\mathbb{M}$ che **minimizza una funzione di perdita** complessiva che si ottiene quando alle singole osservazioni $x_1, \ldots, x_n$ sostituiamo $\mathbb{M}$.

- **Medie di Wald**. Sia $\ell_i = \ell(x_i, a)$ la **perdita** o **costo** che subiamo nel sostituire a al valore x_i e sia $g(\ell_1, \ldots, \ell_n)$ una funzione che **sintetizza** tali perdite. Una media $\mathbb{M}$ dei dati $x_1, \ldots, x_n$ secondo Wald è pari a quel valore $\mathbb{M}$ tale che la perdita complessiva

$$g(\ell(x_1, \mathbb{M}), \ldots, \ell(x_n, \mathbb{M}))$$

  è **minima**.

- In altri termini, se tale valore è unico la media secondo Wald è pari a

$$\mathbb{M} = \arg\min_a g(\ell(x_1, a), \ldots, \ell(x_n, a)).$$

- Nel caso della media aritmetica si ha che $g(\ell_1, \ldots, \ell_n) = \sum_{i=1}^n \ell_i$ e che $\ell_i = (x_i - a)^2$. Quindi

$$\bar{x} = \arg\min_a g(\ell(x_1, a), \ldots, \ell(x_n, a)) = \arg\min_a \sum_{i=1}^n (x_i - a)^2.$$