Riassunto e presentazione dei dati (statistica descrittiva)

2 Data Summary and Presentation

CHAPTER OUTLINE

La media campionaria

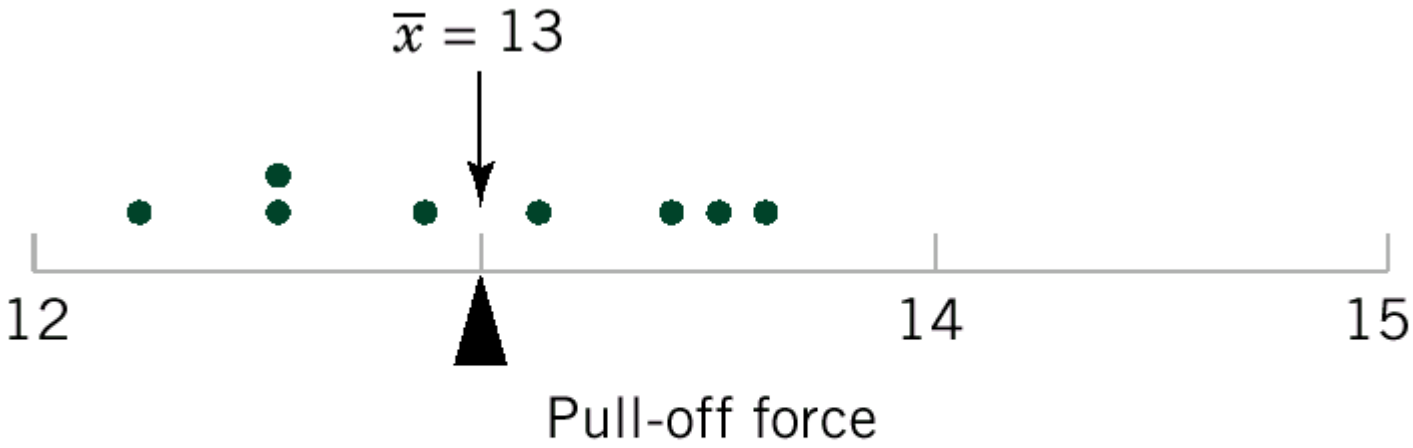
Definizione:

Se n osservazioni (dati) in un campione sono $x_1, x_2, \ldots, x_n$, allora la **media campionaria** vale:

$$\bar{x} \stackrel{\Delta}{=} \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Coincide con la media aritmetica degli n dati nel campione

Interpretazione della media campionaria



Il significato fisico della media campionaria è il centro di massa (baricentro) di un insieme di più masse unitarie collocate sull'asse della grandezza considerata nelle posizioni corrispondenti ai valori x_i dei dati

Calcolo della media campionaria

i	x_i
1	12.6
2	12.9
3	13.4
4	12.3
5	13.6
6	13.5
7	12.6
8	13.1
	$\sum_{i=1}^{n} x_i = 104.0$

$n = 8$ dati

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{104.0}{8} = 13.0$$

(valori di forza di rottura)

La varianza campionaria

Definizione:

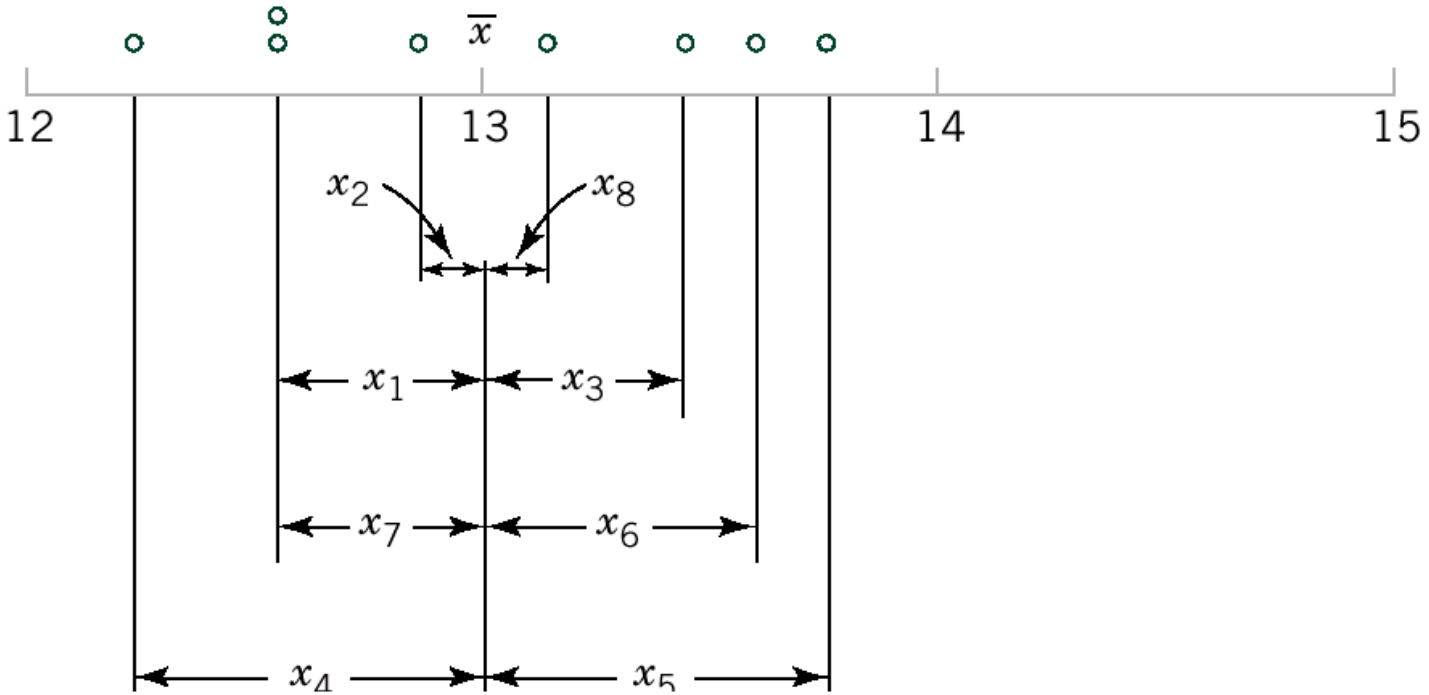
Se n osservazioni in un campione sono denotate $x_1, x_2, \ldots, x_n$, allora la **varianza campionaria** vale:

$$s^2 \stackrel{\Delta}{=} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

s^2 è la "media" degli scarti quadratici (sommatoria diviso n-1)

La **deviazione standard campionaria** s è la radice quadrata positiva della varianza campionaria: $s = \sqrt{s^2}$

Interpretazione della varianza campionaria

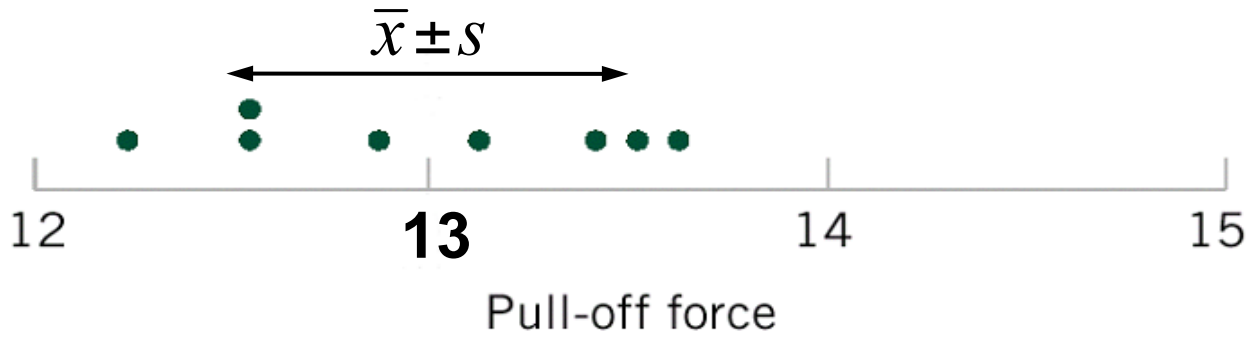


Si calcola la la media, con (n-1) gradi di libertà, dei quadrati delle distanze dal centro (media campionaria): è una **misura della dispersione dei dati**.

Calcolo della varianza campionaria

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	12.6	−0.4	0.16
2	12.9	−0.1	0.01
3	13.4	0.4	0.16
4	12.3	−0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	−0.4	0.16
8	13.1	0.1	0.01
	104.0	0.0	1.60

$\bar{x} = 13$



$$s^2 = \frac{1.60}{8-1} = 0.228$$

$$s = \sqrt{0.228} = 0.48$$

Calcolo rapido/semplificato della varianza campionaria

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2
1	12.6	−0.4	0.16	158.8
2	12.9	−0.1	0.01	166.4
3	13.4	0.4	0.16	179.6
4	12.3	−0.7	0.49	151.3
5	13.6	0.6	0.36	185.0
6	13.5	0.5	0.25	182.3
7	12.6	−0.4	0.16	158.8
8	13.1	0.1	0.01	171.6
	104.0	0.0	1.60	1353.6

La varianza campionaria si può anche calcolare come la **somma dei singoli valori elevati al quadrato meno *n* volte il valor medio al quadrato**, il tutto diviso per **(*n*-1) gradi di libertà**.

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

$$s^2 = \frac{1353.6 - 8 \cdot (104.0)^2}{8-1} = 0.228$$

$$s = \sqrt{0.228} = 0.48$$

Dimostrazione

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

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

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

$$=\sum_{i=1}^{n}x_i^2-n\bar{x}^2$$

Media e varianza di una popolazione

Se l'intera popolazione del fenomeno studiato contiene un numero finito N di dati:

MEDIA

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

VARIANZA

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

Se il numero di dati dell'intera popolazione è infinito non è possibile calcolare media e varianza della popolazione. Infatti non disporremo mai di tutti i dati e non potremo certo effettuare le sommatorie di infiniti termini. In questo caso cercheremo di stimare i parametri della popolazione a partire dai corrispondenti valori ($\bar{x}$ e s^2) calcolati per un campione particolare. Se nel campione considerato per i calcoli si riuscisse ad avere $n \to \infty$, potremmo ritenere che $\bar{x} \to \mu$ e $s^2 \to \sigma^2$.

Media campionaria e varianza campionaria sono degli stimatori per media e varianza dell'intera popolazione.

Diagramma a ramo-e-foglia (stem-and-leaf)

Il diagramma a punti è utile per campioni non troppo ampi (fino a circa 20 dati). Con un maggiore numero di dati è preferibile un altro tipo di diagramma: un diagramma **ramo-e-foglia** è un valido strumento per ottenere immediate informazioni visuali su un campione X, dove ogni dato x_i abbia almeno due cifre.

Per costruire il diagramma si divide ciascun valore numerico x_i del dato i-esimo in due parti:

- un **ramo**, consistente in una o due cifre (le più significative)
- una **foglia**, formata dalle rimanenti cifre (le meno significative)

Si riportano quindi i dati separati su due colonne (ramo e foglia).

Esempio di campione da rappresentare

80 misure di forza di compressione (pressione in “psi” ovvero *pounds per square inch*) per una lega di litio e alluminio

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

minimo valore: 76

massimo valore: 245

3 cifre: centinaia, decine, unità

possibili valori [0-999]

Diagramma stem-and-leaf

cifre delle centinaia e delle decine

numero di occorrenza

Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

cifre delle unità

Diagramma stem-and-leaf considerazioni

Quando i rami sono pochi si perde in leggibilità (*a*), è quindi consigliabile dividere la prima cifra in sottoparti (*b*), senza esagerare per non ottenere l'effetto opposto (*c*) di un eccessivo sparpagliamento.

25 dati con valori tra 61 e 95

(*a*)

Stem	Leaf
6	1 3 4 5 5 6
7	0 1 1 3 5 7 8 8 9
8	1 3 4 4 7 8 8
9	2 3 5

(*b*)

Stem	Leaf
6L	1 3 4
6U	5 5 6
7L	0 1 1 3
7U	5 7 8 8 9
8L	1 3 4 4
8U	7 8 8
9L	2 3
9U	5

Lower (0-4)

Upper (5-9)

(*c*)

Stem	Leaf
6z	1
6t	3
6f	4 5 5
6s	6
6e	
7z	0 1 1
7t	3
7f	5
7s	7
7e	8 8 9
8z	1
8t	3
8f	4 4
8s	7
8e	8 8
9z	
9t	2 3
9f	5
9s	
9e	

Diagramma stem-and-leaf ordinato

In ogni diagramma i dati sono sempre ordinati in modo crescente nella colonna dei rami.
Il diagramma può essere reso più leggibile se si ordinano i dati anche nella colonna delle foglie.
Di norma ciò è consigliabile solo per procedure automatiche al computer, in quanto l'operazione di riordino può richiedere parecchio tempo e non aggiunge molta informazione.

7	6
8	7
9	7
10	1 5
11	0 5 8
12	0 1 3
13	1 3 3 4 5 5
14	1 2 3 5 6 8 9 9
15	0 0 1 3 4 4 6 7 8 8 8 8
16	0 0 0 3 3 5 7 7 8 9
17	0 1 1 2 4 4 5 6 6 8
18	0 0 1 1 3 4 6
19	0 3 4 6 9 9
20	0 1 7 8
21	8
22	1 8 9
23	7
24	5

Valori caratteristici di un insieme di dati

Oltre a **media**, $\bar{x}$ o μ, e **deviazione standard**, s o σ, esistono altri **valori** (puntuali) **caratteristici** di un insieme di dati e che forniscono ulteriori informazioni sulla distribuzione dei dati (tendenza, centro, dispersione, asimmetrie).

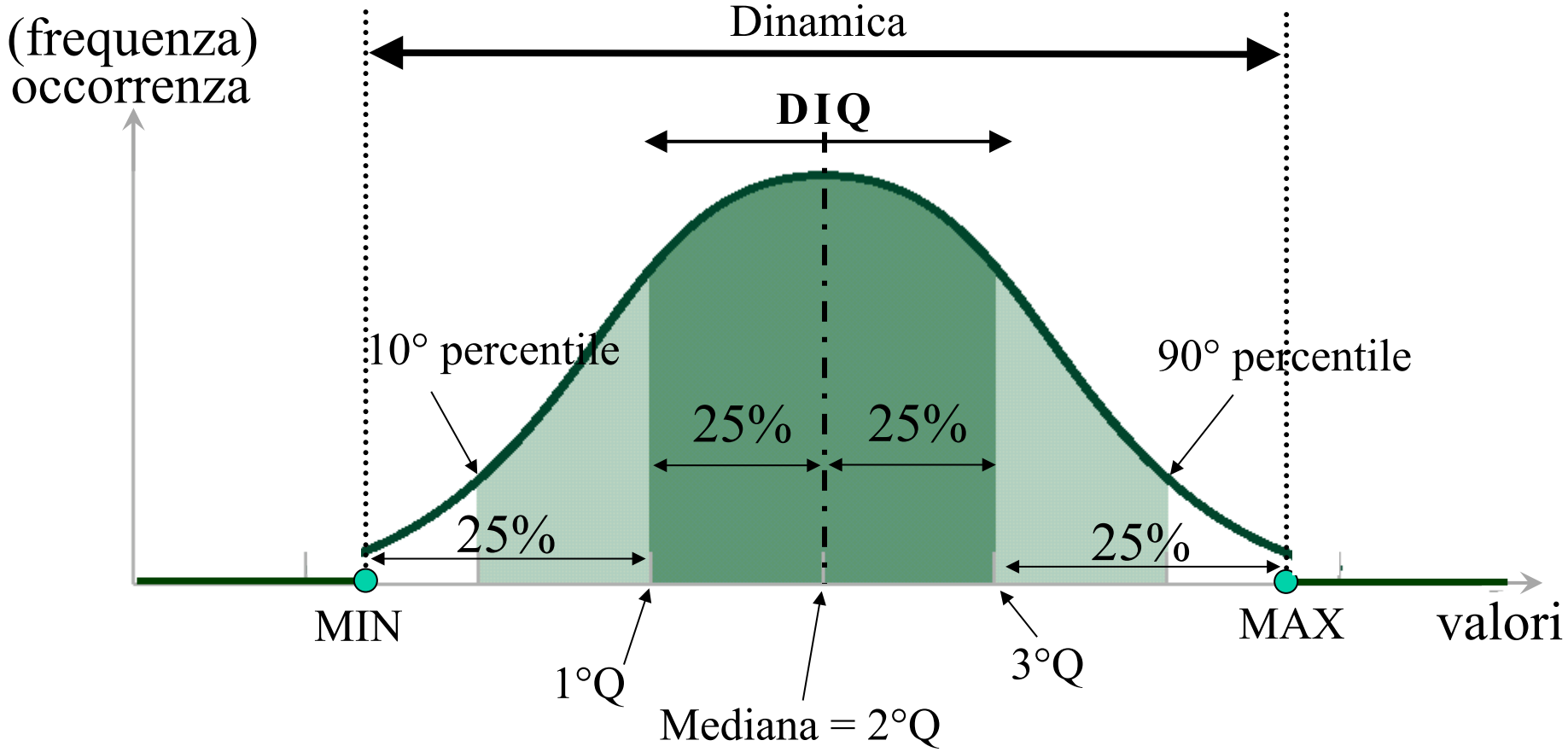
- **Moda (M_0):** valore che ha la massima occorrenza (o frequenza) nell'insieme di dati o valore più probabile. [è il picco della distribuzione di probabilità (ci sono anche distribuzioni bi- o multi-modali)]
- **Mediana (M_e)**: è una misura del centro del campione che divide i dati in due parti uguali (tanti dati al di sopra quanti al di sotto).
  Se il numero di dati è pari, la mediana è a metà tra i due valori centrali; oppure, se dispari, la mediana coincide con il valore centrale.
  [è il centro, tra le aree di sx e dx sotto la distribuzione di probabilità – punto a valore 0.5 nella cumulativa – (solo per distribuzioni simmetriche sta nel mezzo della dinamica e coincide con la media)]
- **Dinamica (*range*)**: differenza tra il valore massimo e il valore minimo
  (è anche detta escursione da minimo a massimo o picco-picco).

Valori caratteristici (percentili e quartili)

- ***k*-esimo percentile**: valore superiore al k% delle osservazioni e inferiore al $(100-k)$%.
- **Primo quartile** (quartile basso o di sx): valore tale che un quarto delle osservazioni abbia un valore inferiore; corrisponde al 25-esimo percentile.
- **Secondo quartile**: valore tale che due quarti delle osservazioni abbiano un valore inferiore; corrisponde al 50-esimo percentile e alla mediana.
- **Terzo quartile** (quartile alto o di dx): valore tale che tre quarti delle osservazioni abbiano un valore inferiore; corrisponde al 75-esimo percentile.
- **Dinamica interquartile (DIQ)**: differenza tra il quartile superiore (terzo) e il quartile inferiore (primo); è la distanza tra i due quartili (1°e 3°).

Interpretazione dei valori caratteristici

N.B. Il grafico è solo qualitativo e non in scala (a percentuali di popolazione uguali dovrebbero corrispondere aree uguali).



Le percentuali sono riferite all'area sottesa alla curva (probabilità) o anche si possono vedere come frazione del totale dei valori.

Metodo di calcolo del k-esimo percentile

Consideriamo un campione di *n* dati ordinati in maniera crescente.
L'indice, di posizione, del *k*-esimo percentile è : $\boldsymbol{I_{k\%} = (n+1)\times k /100}$
Dall'indice si ricava quindi il valore esatto con un'interpolazione lineare tra i due dati con indici pari all'intero prima e dopo di $I_{k\%}$

Esempio 1: n=14 dati x_i. Calcoliamo il 23-esimo precentile.
$I_{23\%} = (14+1)\times 23 /100 = 3.45$
Il valore del 23-esimo percentile sarà compreso tra il 3° ed il 4° dato (x_3 e x_4).
Numericamente poi $X_{23\%} = x_3 + (x_4 - x_3) \times 0.45$ (interpolazione lineare)

Esempio 2: n=78 dati x_i. Calcoliamo il 25-esimo precentile (1° quartile).
$I_{25\%} = (78+1)\times 25 /100 = 19.75$
Il valore del 25-esimo percentile sarà compreso tra il 19° ed il 20° dato (x_{19} e x_{20}).
Numericamente poi $X_{25\%} = x_{19} + (x_{20} - x_{19}) \times 0.75$ (interpolazione lineare)

Alcuni programmi di calcolo (Matlab) si limitano ad effettuare la media tra i due dati adiacenti che comprendono il percentile considerato. Quando n è grande l'imprecisione commessa è comunque trascurabile.

Esempio di calcolo dei quartili

Consideriamo ancora il campione di 80 dati (es. forza di compressione), in un diagramma a ramo-e-foglia ordinato.

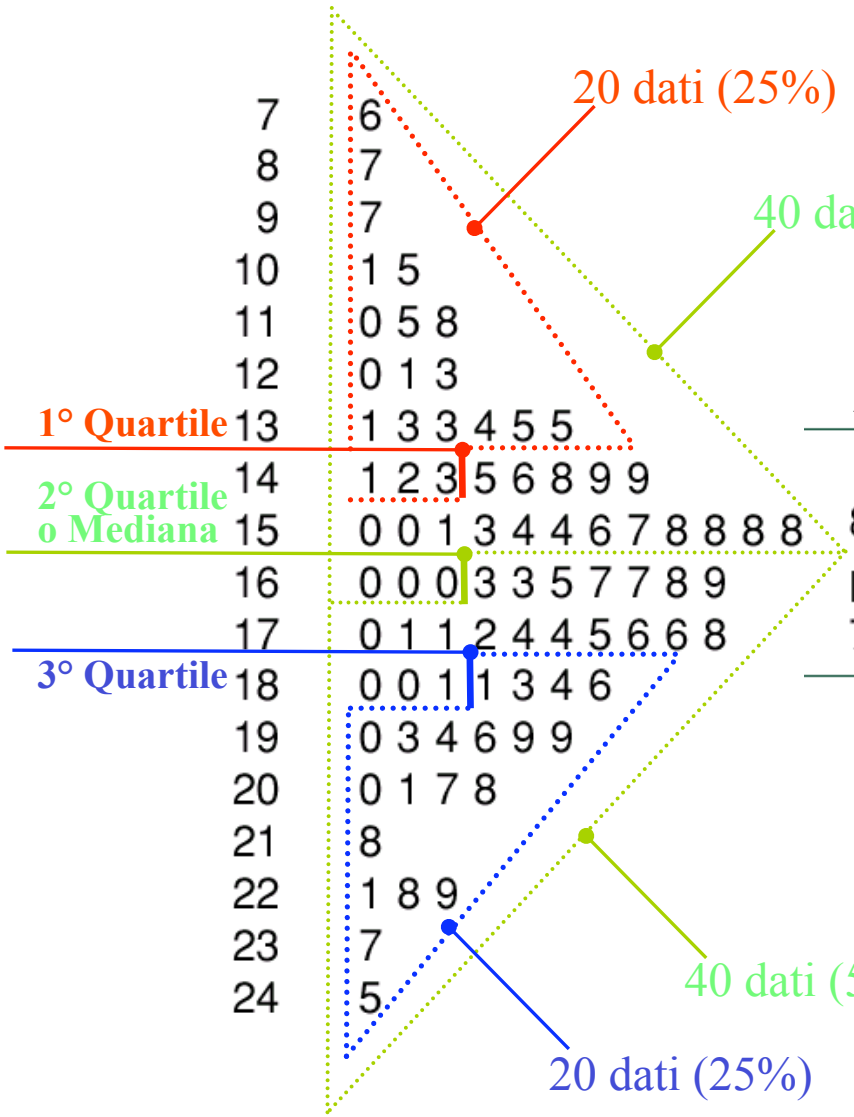
L'indice del primo quartile Q_1 vale:
$I_{25\%} = (80+1)\times 25 /100 = 20.25$
Dunque $\boldsymbol{Q_1} = x_{20} + (x_{21} - x_{20}) \times 0.25 =$
$= 143+(145-143) \times 0.25 =$ **143.5**

L'indice del secondo quartile Q_2 vale:
$I_{50\%} = (80+1)\times 50 /100 = 40.5$
Dunque Q_2 è la media tra i due dati centrali (coincide con la mediana
$\boldsymbol{Q_2} =$ **161.5**)

L'indice del terzo quartile Q_3 vale:
$I_{75\%} = (80+1)\times 75 /100 = 60.75$
Dunque $\boldsymbol{Q_3} = x_{60} + (x_{61} - x_{60}) \times 0.75 =$
$= 181+(181-181) \times 0.75 =$ **181**

7	6
8	7
9	7
10	1 5
11	0 5 8
12	0 1 3
13	1 3 3 4 5 5
14	1 2 3 5 6 8 9 9
15	0 0 1 3 4 4 6 7 8 8 8 8
16	0 0 0 3 3 5 7 7 8 9
17	0 1 1 2 4 4 5 6 6 8
18	0 0 1 1 3 4 6
19	0 3 4 6 9 9
20	0 1 7 8
21	8
22	1 8 9
23	7
24	5

Esempio di valori caratteristici



7	6
8	7
9	7
10	1 5
11	0 5 8
12	0 1 3
13	1 3 3 4 5 5
14	1 2 3 5 6 8 9 9
15	0 0 1 3 4 4 6 7 8 8 8 8
16	0 0 0 3 3 5 7 7 8 9
17	0 1 1 2 4 4 5 6 6 8
18	0 0 1 1 3 4 6
19	0 3 4 6 9 9
20	0 1 7 8
21	8
22	1 8 9
23	7
24	5

N	$\bar{x}$	M_e	s
N	Mean	Median	StDev
80	162.66	161.50	33.77
Min	Max	Q1	Q3
76.00	245.00	143.50	181.00

Q_1=143+(145-143)×0.25=143.50

Q_2=160+(163-160)×0.50=161.50

Q_3=181+(181-181)×0.75=181.00

Versione "MINITAB"

Distribuzione di frequenza

La **distribuzione di frequenza** è una sommarizzazione dei dati più compatta del diagramma ramo-e-foglia

Per costruire una distribuzione di frequenza è necessario dividere la dinamica dei dati in intervalli, chiamati **classi**, tipicamente ma non necessariamente di uguale ampiezza

La scelta del numero di intervalli dipende dal numero complessivo di acquisizioni e dalla dispersione dei dati

Spesso un numero sensato di intervalli da usare è pari alla radice quadrata del numero di dati $n_{\text{int}}=\sqrt{n}$ (è semplicemente una scelta per rendere significativo il diagramma). Meglio ancora $n_{\text{int.}}=1+\log_2 n$.

La distribuzione di frequenza è legata, come si vedrà nel Capitolo 3, al concetto di probabilità (vedremo come, nota la distribuzione di frequenza, è possibile calcolare la "probabilità")

Distribuzione di frequenza

La **frequenza (assoluta)** è il numero di dati nella rispettiva classe

La **frequenza relativa** è pari alla frequenza divisa per il numero complessivo di dati (n) del campione

La **frequenza relativa cumulativa** è la somma delle frequenze relative, a partire dalla classe minima sino alla classe corrente

La frequenza assoluta è un numero intero positivo (solitamente >1)
la frequenza relativa è un numero razionale sempre <1

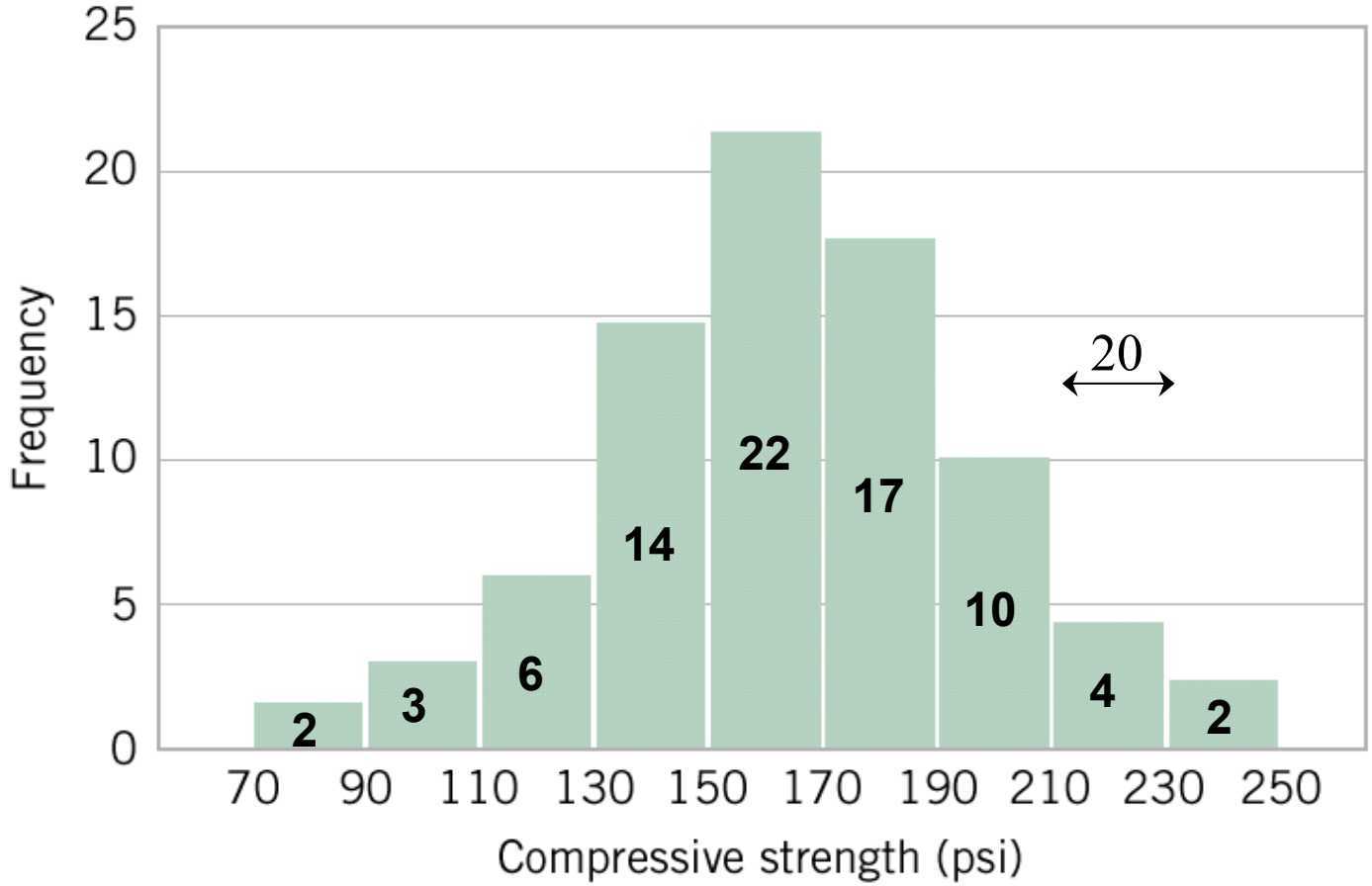
Class Interval (psi)	Tally	Frequency	Cum. Freq.	Relative Frequency	Cumulative Relative Frequency
$70 \leq x < 90$	||	2	2	0.0250	0.0250
$90 \leq x < 110$	|||	3	5	0.0375	0.0625
$110 \leq x < 130$	𝍸 |	6	11	0.0750	0.1375
$130 \leq x < 150$	𝍸 𝍸 ||||	14	25	0.1750	0.3125
$150 \leq x < 170$	𝍸 𝍸 𝍸 𝍸 ||	22	47	0.2750	0.5875
$170 \leq x < 190$	𝍸 𝍸 𝍸 ||	17	64	0.2125	0.8000
$190 \leq x < 210$	𝍸 𝍸	10	74	0.1250	0.9250
$210 \leq x < 230$	||||	4	78	0.0500	0.9750
$230 \leq x < 250$	||	2	80	0.0250	100% → 1.0000

5 dati (conteggi)

(Frequenza assoluta e relativa differiscono solo per un fattore costante: il numero *n* di dati)

Istogramma (di frequenza)

La rappresentazione grafica di una distribuzione di frequenza è detta **istogramma** (diagramma a barre).

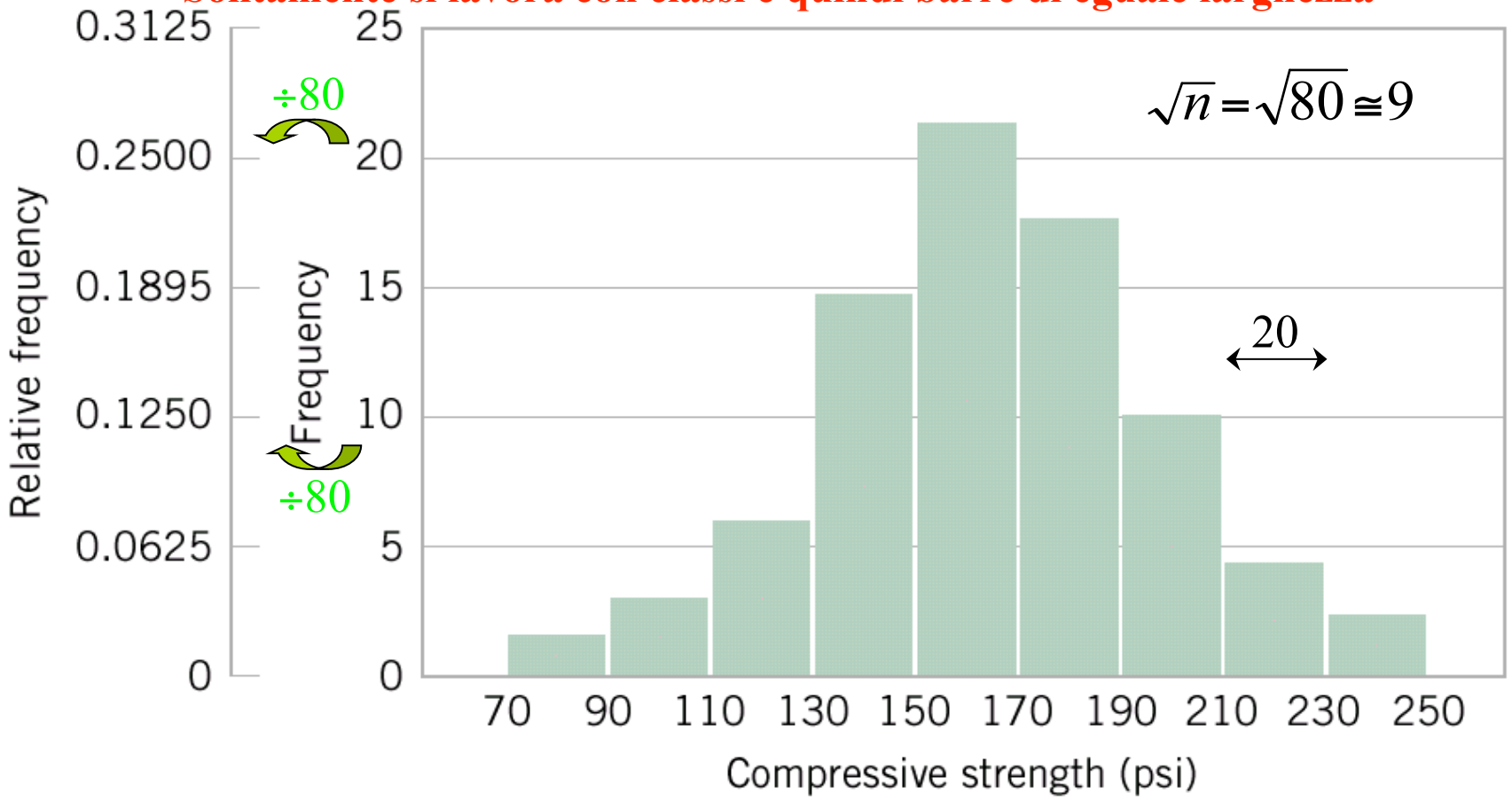


Sono state utilizzate 9 classi (infatti $\sqrt{n} = \sqrt{80} \cong 9$)

Istogramma (di freq. relativa)

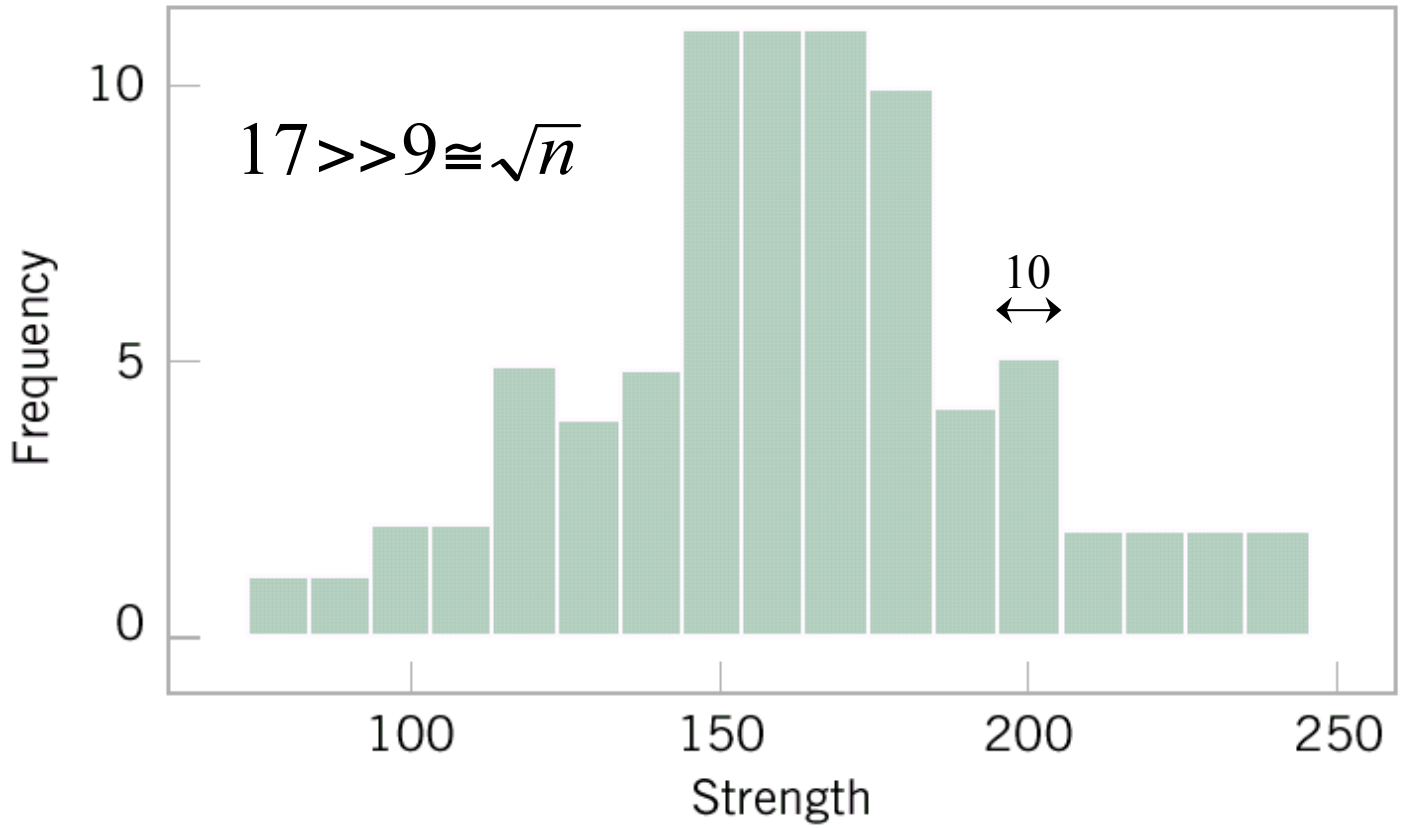
L'altezza di una barra indica la frequenza relativa della classe considerata e la larghezza di una barra è proporzionale alla larghezza dell'intervallo che definisce la classe ⇒ con "basi unitarie" l'**area delle barre è la probabilità** della classe (la somma di tutte le aree deve dare 1=100%).

Solitamente si lavora con classi e quindi barre di eguale larghezza



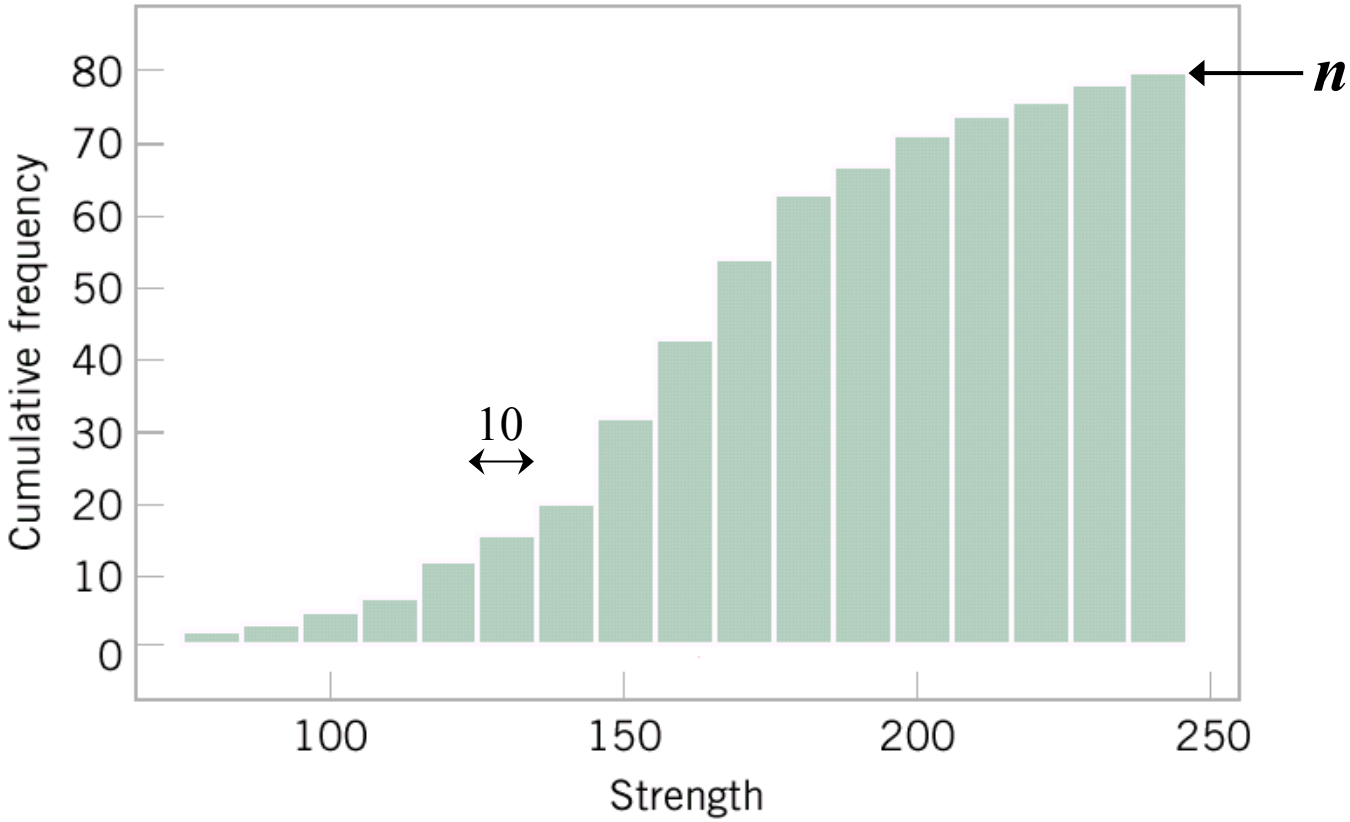
Istogramma con "troppe" classi

Aumentando eccessivamente il numero di classi l'istogramma non fornisce una informazione chiara:
nell'esempio considerato **17 classi sono troppe**



Distribuzione cumulativa (istogramma)

In questo grafico, l'altezza di ogni barra è data dal numero di osservazioni con valore inferiore al limite superiore della classe corrispondente (ovvero è la somma delle frequenze di tutte le classi precedenti più quella corrente).

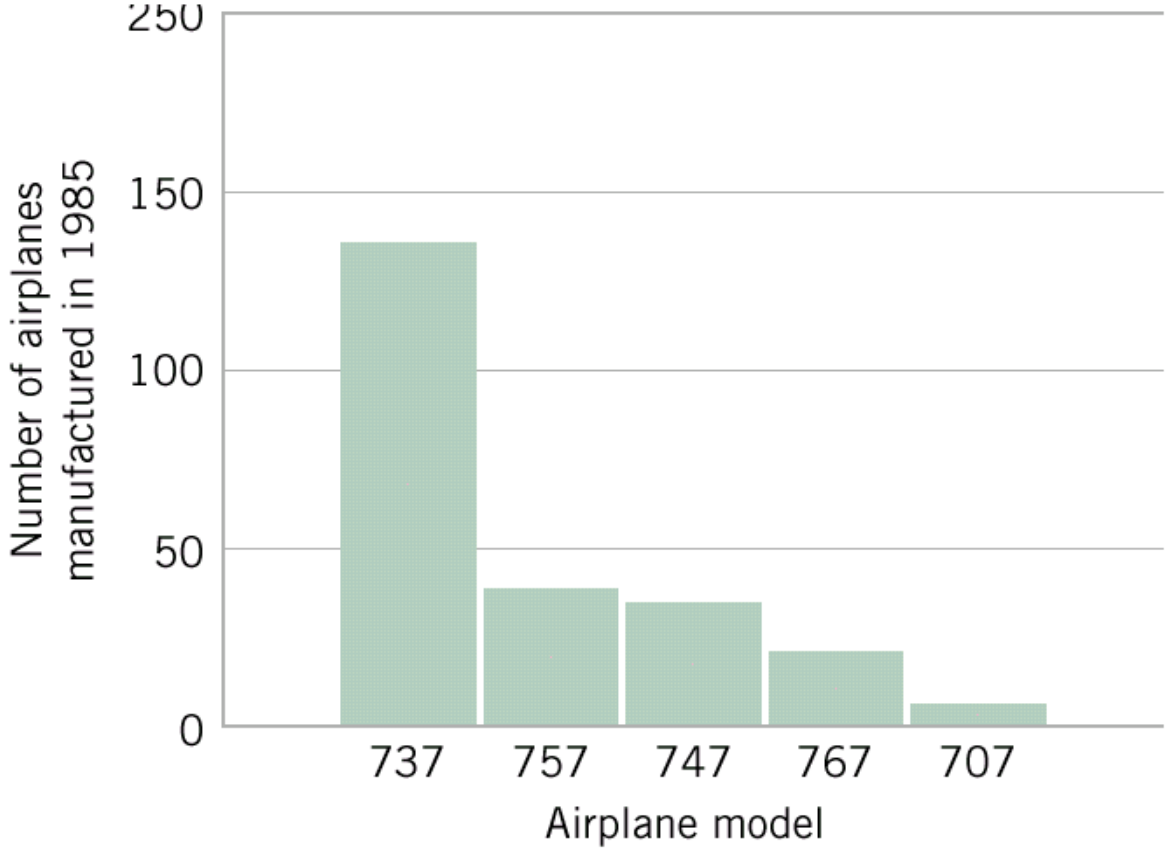


Naturalmente esiste anche una distribuzione **cumulativa relativa**, che si ottiene dal grafico/valori della distribuzione cumulativa dividendo per il numero *n* di dati. Il **picco** della cumulativa relativa è sempre **uguale a 1, o al 100 %**.

Impiego degli istogrammi

Un istogramma può essere utile anche per rappresentare differenti tipi di dati (**classi tipo**, contraddistinte da una particolare proprietà/tipologia).

Nell'esempio è rappresentato il numero di aeroplani prodotti nel 1985, al variare del **modello**. In questo caso le barre devono avere tutte uguale larghezza.



Misure della forma di distribuzione dei dati

Nella statistica descrittiva si definiscono alcune **misure/indici di forma di una distribuzione** e che vanno sotto il nome di **asimmetria e curtosi**.

DEFINIZIONI:

Si dice **simmetrica** una distribuzione che, rispetto alla posizione centrale, assume uguale struttura delle frequenze sia nella parte destra che nella parte sinistra. (Media=Mediana)

Per avere una distribuzione simmetrica la perfetta coincidenza delle tre misure di tendenza è solo condizione necessaria ma non sufficiente.

Si dice **normale** una distribuzione che:

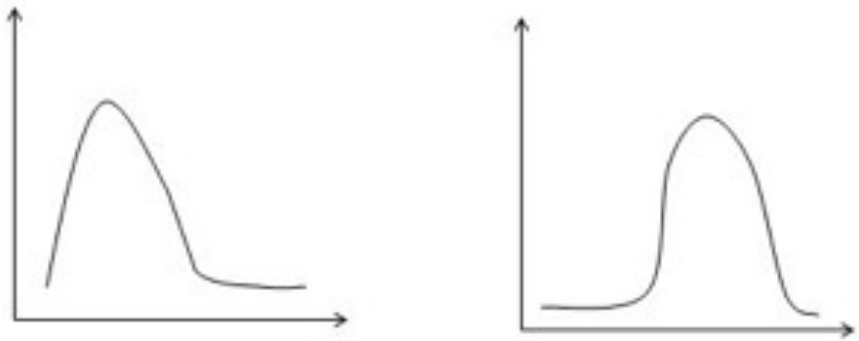
- presenta con una forma a campana e simmetrica rispetto alla posizione centrale
- tutti gli indici di posizione centrale assumono uguale valore (Media=Mediana=Moda)

Asimmetria

• **Asimmetria (*skewness*):** misura la mancanza di simmetria della distribuzione. Si dice asimmetrica una distribuzione la cui forma non si presenta speculare rispetto alla posizione centrale.
In particolare, si ha:

Asimmetria **positiva** o **destra** quando la distribuzione ha una coda allungata verso **destra**

Asimmetria **negativa** o **sinistra** quando la distribuzione ha una coda allungata verso **sinistra**



Asimmetria positiva o destra

Asimmetria negativa o sinistra

$M_0 < M_e < \mu \Rightarrow$ *asimmetria positiva* (destra)

$\mu < M_e < M_0 \Rightarrow$ *asimmetria negativa* (sinistra)

Indice normalizzato di asimmetria

Un semplice indice di asimmetria si ottiene mediante la **differenza tra media e mediana rapportata allo SQM**, scarto (dev. st.)che si dimostra essere pari al massimo della differenza al numeratore.

$$A = \frac{\mu - M_e}{\sigma}$$

$$poichè \left| \mu - M_e \right| \leq \sigma \Rightarrow -1 \leq A \leq 1$$

Tale indicatore è un indice normalizzato e quindi varia tra -1 e +1 a prescindere dall'unità di misura della variabile originaria.

$$A < 0 \Rightarrow asimmetria\ negativa$$

$$A = 0 \Rightarrow simmetria$$

$$A > 0 \Rightarrow asimmetria\ positiva$$

Sostituendo M_e con M_0 si ottiene l'indice di asimmetria di Pearson

Un indice più evoluto di asimmetria, **indice di asimmetria di Fisher**, si ottiene dalla media dei rapporti tra gli scarti e la deviazione standard elevati alla terza potenza

$$a = Sk = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(x_i - \bar{x})^3}{s^3}$$

Curtosi

• **Curtosi (*curtosis*):** misura del maggiore o minore **peso delle code** rispetto alla parte centrale, e conseguentemente, il minore o maggiore **appuntimento** della distribuzione:

(nella misura β si utilizza la variabile standardizzata Z – scarto tra un dato e la media, normalizzato alla deviazione standard – che vedremo in maggiore dettaglio nel seguito)

Indice di curtosi di Pearson

L'indice di curtosi di Pearson misura la curtosi come media aritmetica delle quarte potenze della variabile standardizzata $Z=(x-\mu)/\sigma$:

$$K = \beta = \frac{1}{N-1}\sum_{i=1}^{N}\frac{(x_i - \bar{x})^4}{s^4}$$

Questo indice β assume valore **pari a 3 nel caso** in cui la distribuzione abbia una forma **normale**.

Leptocurtosi e platicurtosi

Quando la distribuzione ha una forma **maggiormente appuntita rispetto alla normale** si parla di forma leptocurtica (poche-code) e l'indice sarà $\beta > 3$.

Quando la distribuzione ha una forma **meno appuntita rispetto alla normale** si parla di forma platicurtica (molte-code) e l'indice sarà $\beta < 3$.

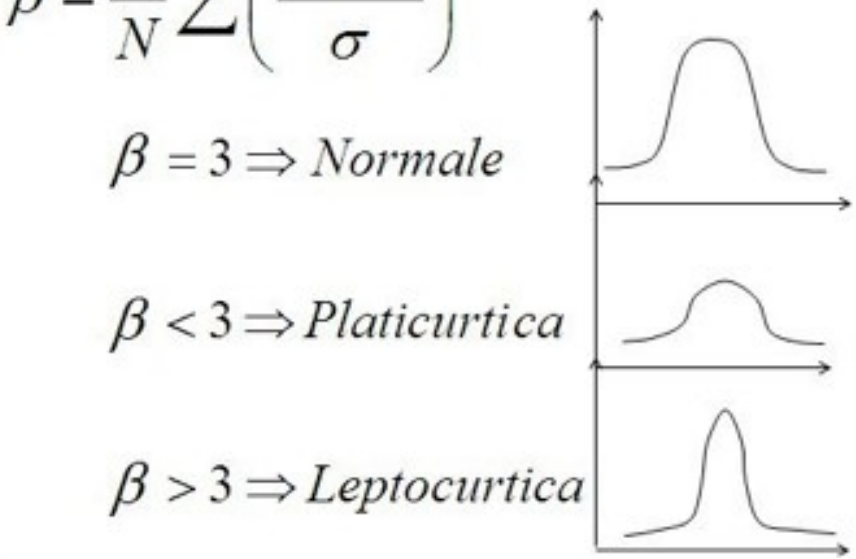
Sottraendo la costante 3 all'indice di Pearson si ottiene una versione centrata rispetto alla distribuzione normale: $\beta^*=0$ forma normale; $\beta^*>0$ forma leptocurtica; $\beta^*<0$ forma platicurtica

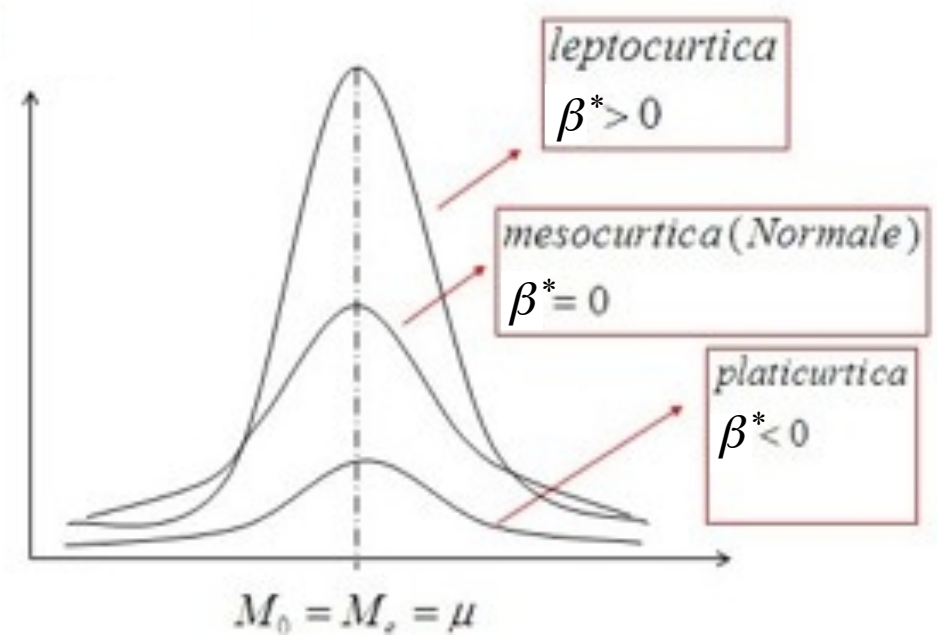
$$\beta = \frac{1}{N}\sum\left(\frac{x_i - \mu}{\sigma}\right)^4$$

$\beta = 3 \Rightarrow Normale$

$\beta < 3 \Rightarrow Platicurtica$

$\beta > 3 \Rightarrow Leptocurtica$



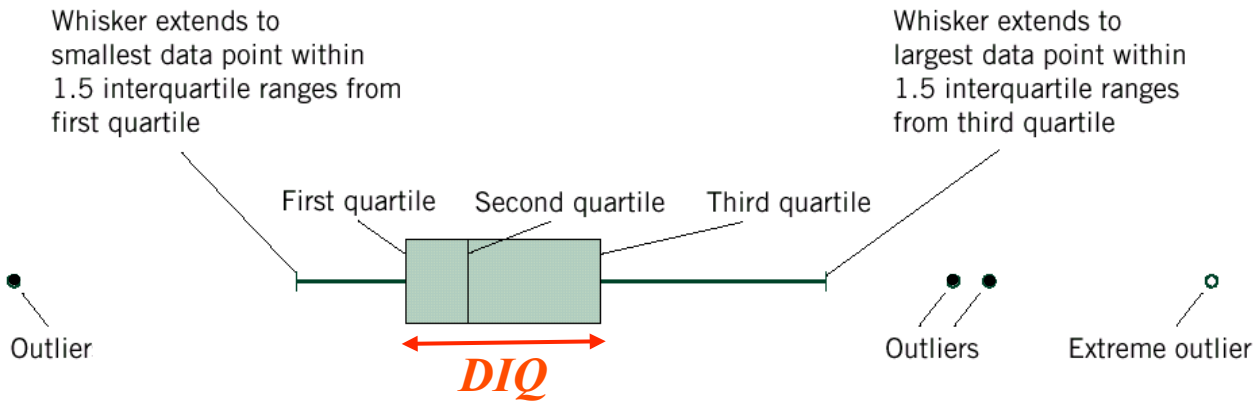


Box plot

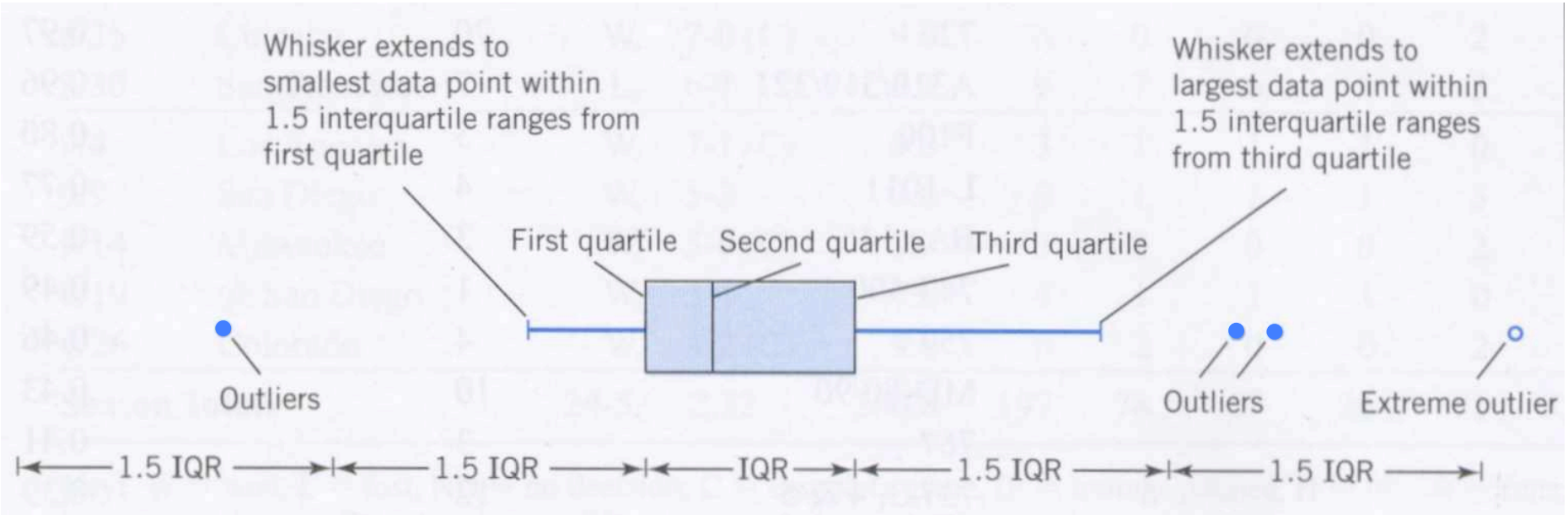
Il diagramma a ramo-e-foglia e l'istogramma danno una **visione generale** (qualitativa) di un insieme di dati. Singoli valori numerici come la media, la varianza o i quartili forniscono **informazioni puntuali** (quantitative) su uno specifico aspetto del campione.

Un indicatore grafico (visione d'insieme) che descrive anche e contemporaneamente diverse importanti caratteristiche quantitative del campione è il ***box plot***: è una **scatola delimitata dai quartili**, e "tagliata" dalla mediana, che riporta anche **due baffi** estesi fino a "circa" 1.5 volte la **dinamica interquartile (*DIQ*=Q_3-Q_1)**.

Eventuali punti esterni ai baffi vengono riportati singolarmente (*outliers*).
Un punto più lontano di 3 dinamiche interquartili, dal quartile corrispondente, è detto *outlier* estremo.



Box plot con punti esterni



Inter Quartile Range IQR ≡ DIQ

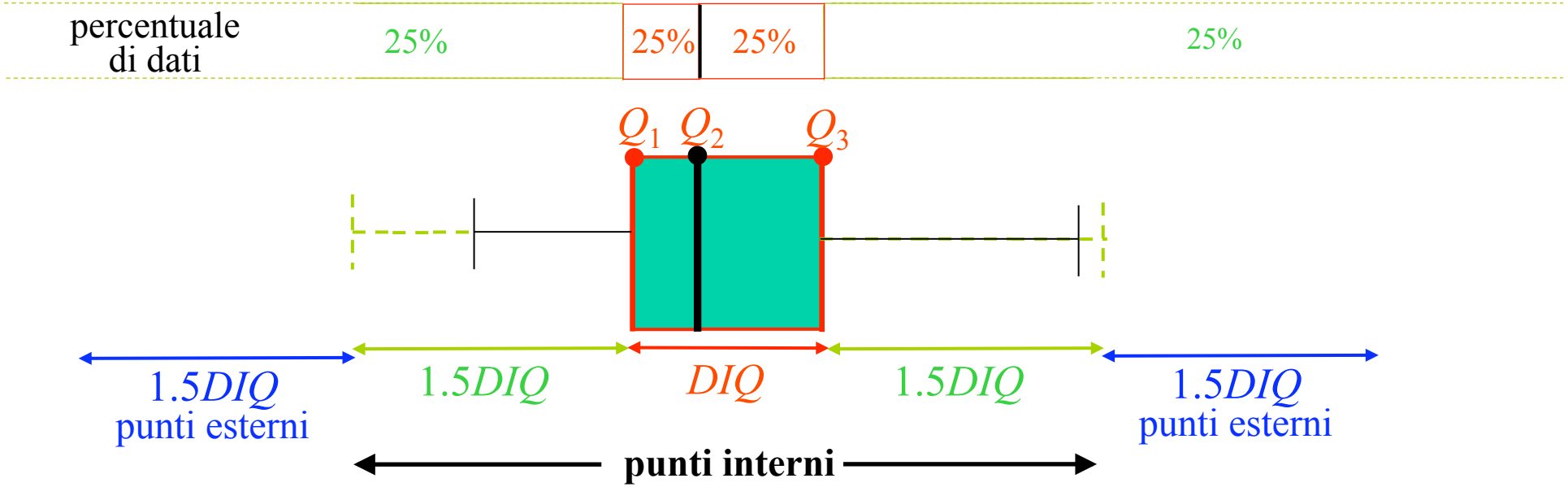
delimita la "scatola" (box) che contiene il 50 %, centrale, dei dati

Parametri del box plot

$DIQ=Q_3-Q_1$ dinamica interquartile

$W_{L,lim}=Q_1-1.5DIQ$ limite inferiore per il baffo basso

$W_{H,lim}=Q_3+1.5DIQ$ limite superiore per il baffo alto



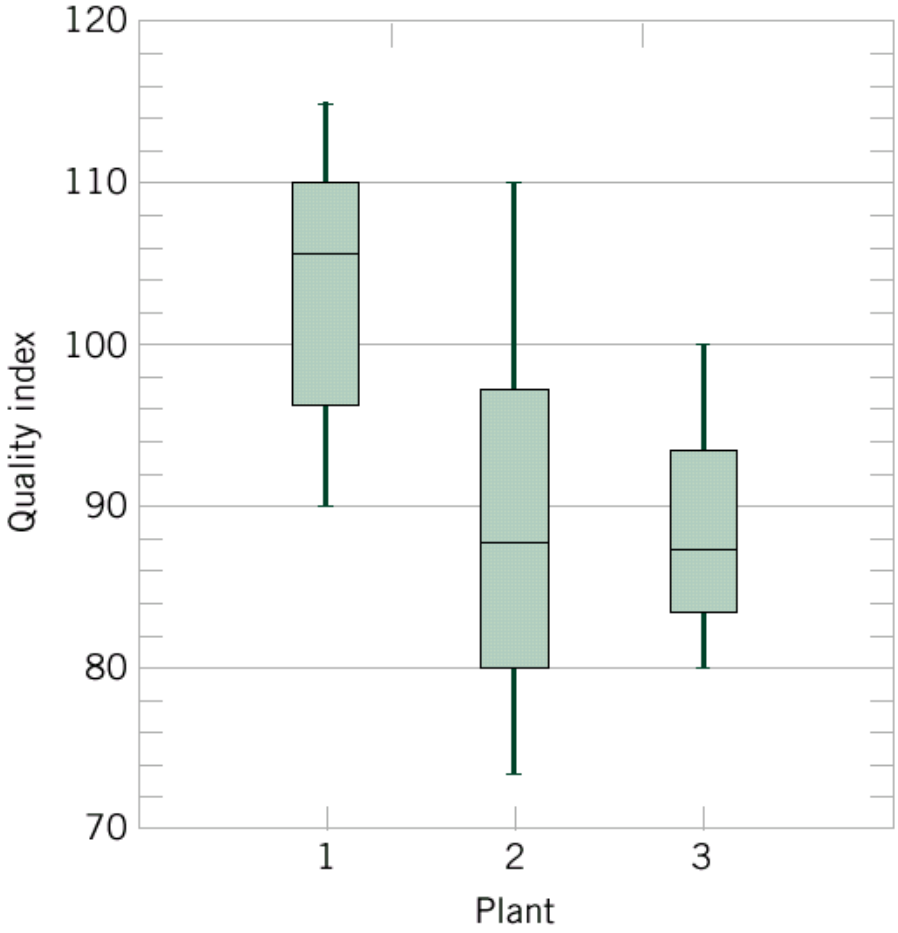
Dati nell'intervallo [Q_1-1.5DIQ , Q_3+1.5DIQ] sono **punti interni**.
Dati entro gli intervalli [Q_1-3DIQ , Q_1-1.5DIQ] oppure [Q_3+1.5DIQ , Q_3+3DIQ] sono detti **punti esterni (*outliers*)**.
Dati con valore inferiore a Q_1-3DIQ o superiore a Q_3+3DIQ sono **punti esterni estremi**.

Utilità dei box plot

I *box plot* sono molto utili per il confronto diretto (visivo) di dati provenienti da campioni diversi: in figura si riportano gli indici di qualità riferiti a tre impianti di produzione.

Il ***box plot*** è un indicatore grafico che fornisce importanti informazioni quantitative su un insieme di dati:

– Posizione e tendenza centrale
– Variabilità e dispersione
– Simmetria o asimmetria
– Identificazione dei punti esterni

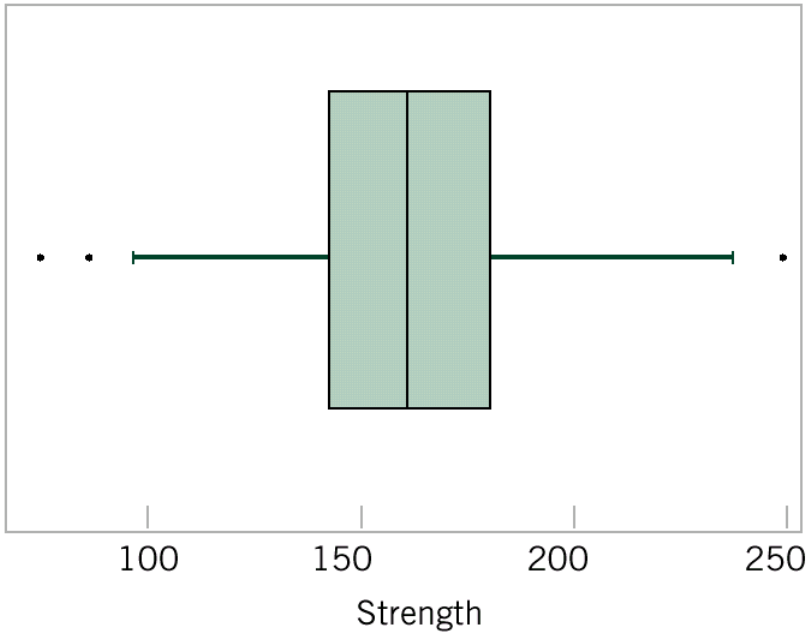


Esempio di box plot

Riportiamo il *box plot* relativo alle misure di pressione, a destra del diagramma ramo-e-foglia corrispondente.

7	6
8	7
9	7
10	1 5
11	0 5 8
12	0 1 3
13	1 3 3 4 5 5
14	1 2 3 5 6 8 9 9
15	0 0 1 3 4 4 6 7 8 8 8 8
16	0 0 0 3 3 5 7 7 8 9
17	0 1 1 2 4 4 5 6 6 8
18	0 0 1 1 3 4 6
19	0 3 4 6 9 9
20	0 1 7 8
21	8
22	1 8 9
23	7
24	5

pos. centrale ~160 psi
dispersione ~±20 psi

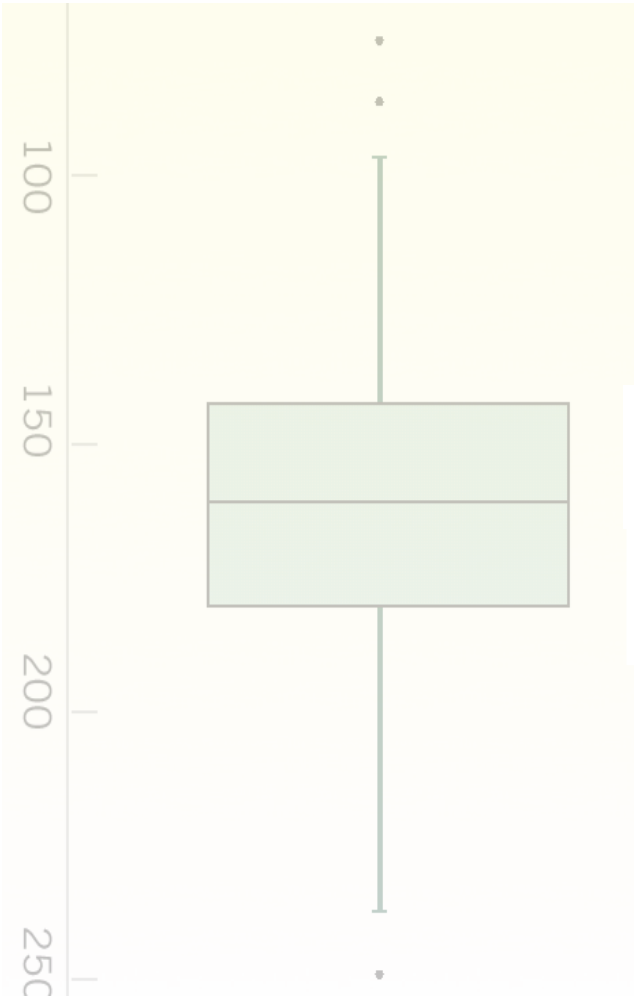


simmetria: SI
punti esterni: SI

Esempio di box plot

Riportiamo il *box plot* relativo alle misure di pressione, a destra del diagramma ramo-e-foglia corrispondente.

7	6
8	7
9	7
10	1 5
11	0 5 8
12	0 1 3
13	1 3 3 4 5 5
14	1 2 3 5 6 8 9 9
15	0 0 1 3 4 4 6 7 8 8 8 8
16	0 0 0 3 3 5 7 7 8 9
17	0 1 1 2 4 4 5 6 6 8
18	0 0 1 1 3 4 6
19	0 3 4 6 9 9
20	0 1 7 8
21	8
22	1 8 9
23	7
24	5



pos. centrale ~160 psi
dispersione ~±20 psi

simmetria: SI
punti esterni: SI

Grafici di serie temporali

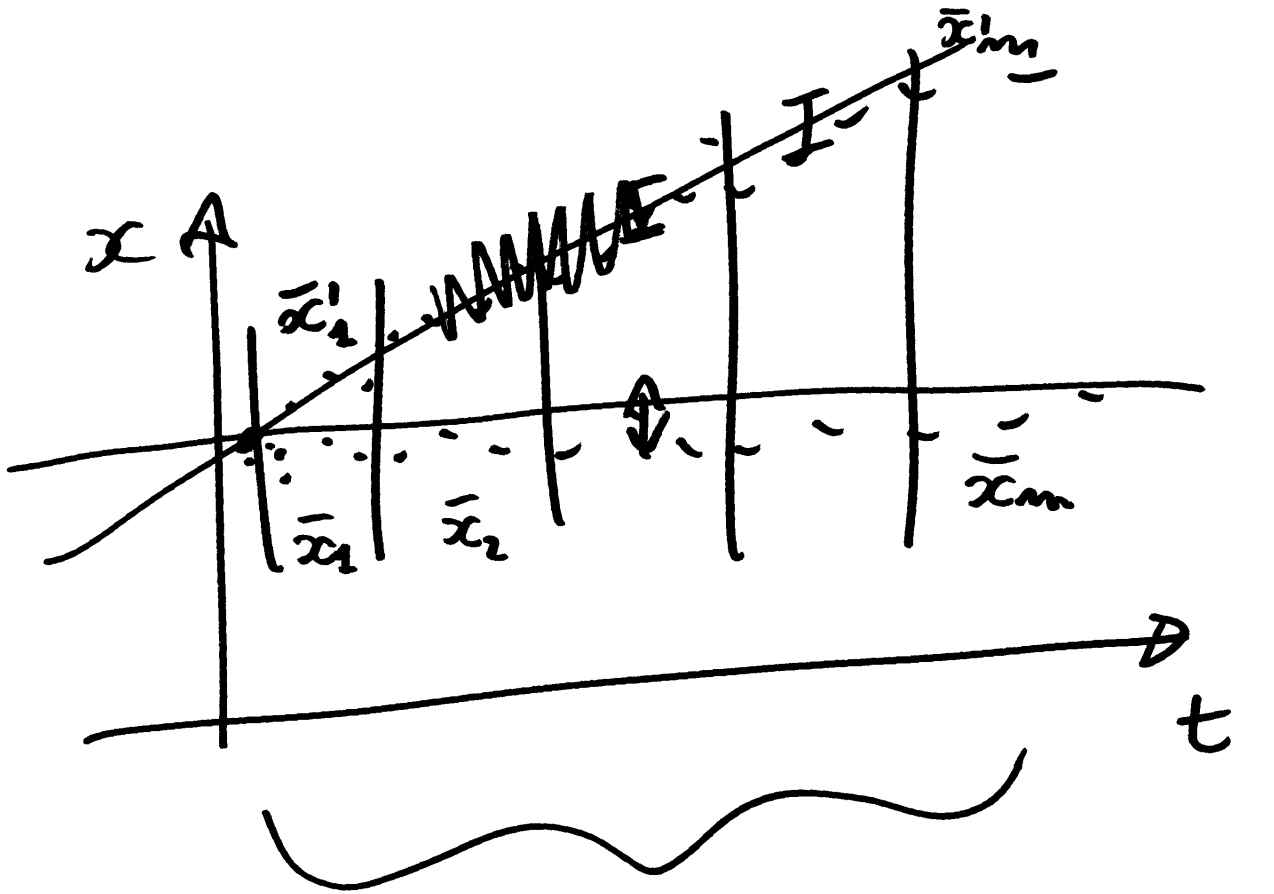
Le rappresentazioni grafiche considerate finora sono molto utili per stimare la variabilità statistica dei dati nel campione.

Spesso è necessario poter visualizzare anche l'**evoluzione nel tempo** del processo considerato.

Una serie nel tempo (**serie temporale**) è un insieme di dati in cui le osservazioni sono registrate nell'ordine cronologico in cui avvengono. Il grafico di una serie temporale riporta sull'**asse verticale** (ordinate) il valore del **dato** e sull'**asse orizzontale** (ascisse) il **tempo**.

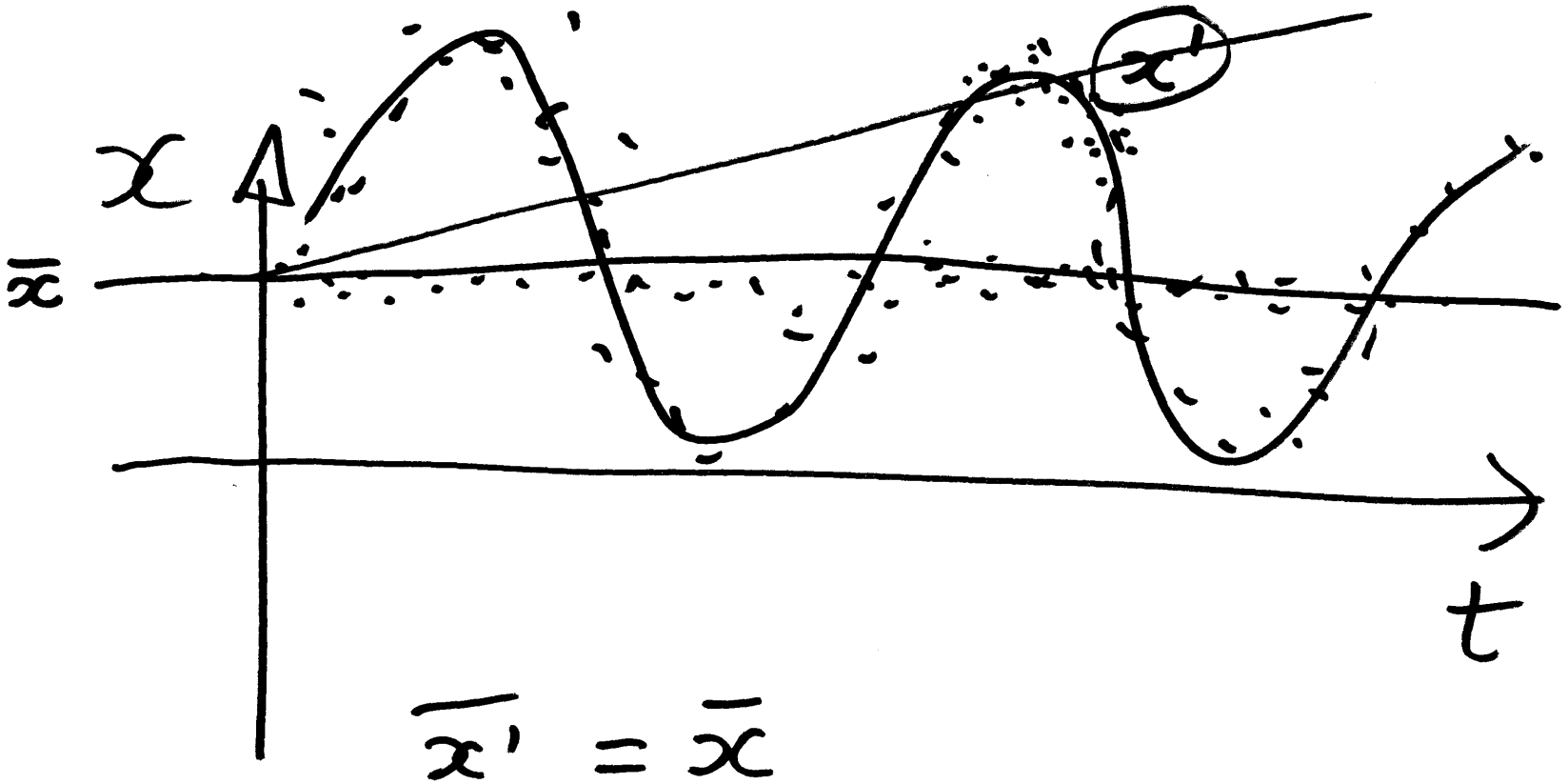
In questo tipo di grafici possiamo osservare degli **avvenimenti isolati**, le **derive**, i **cicli temporali**, e altri aspetti che non potrebbero essere visualizzati altrimenti.

Esempio di dati con deriva (trend) nel tempo



$$s^2(x') \gg s^2(x)$$

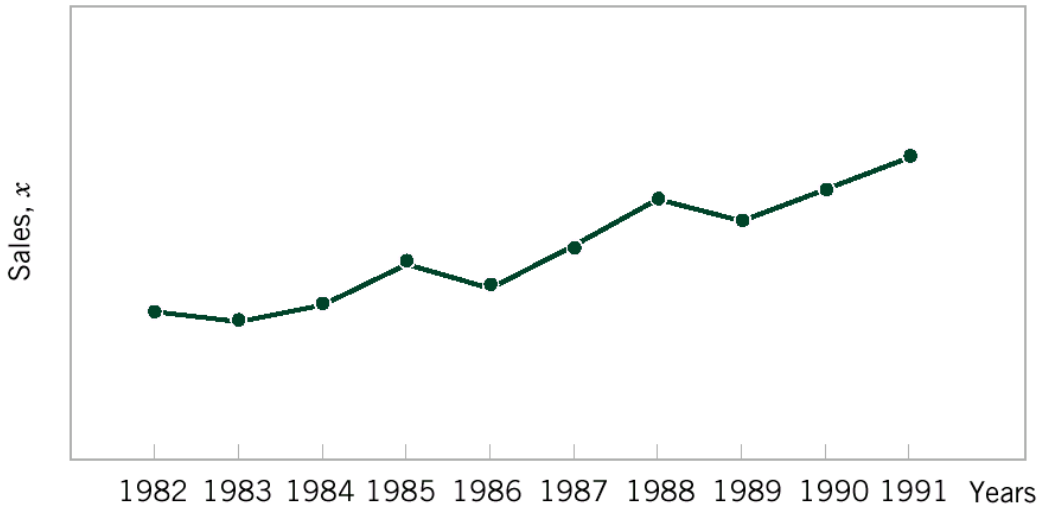
Esempio di dati con ciclo (oscillazione) nel tempo



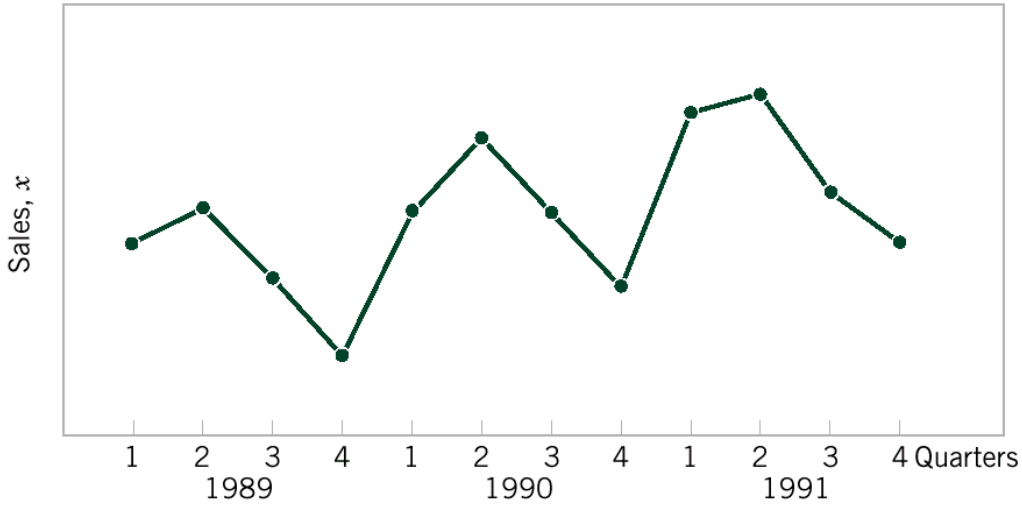
$$\overline{x'} = \bar{x}$$

$$s^2(x') \gg s^2(x)$$

Esempio di grafici nel tempo



Rappresentazione in serie temporale delle vendite di una azienda su 10 anni



Il lungo periodo mostra una tendenza alla crescita.
Nel breve periodo si evidenzia un'oscillazione annuale, sui 4 "quarti" dell'anno (ossia i trimestri)

Grafico combinato

A volte può essere utile combinare due tipi di rappresentazione grafica.

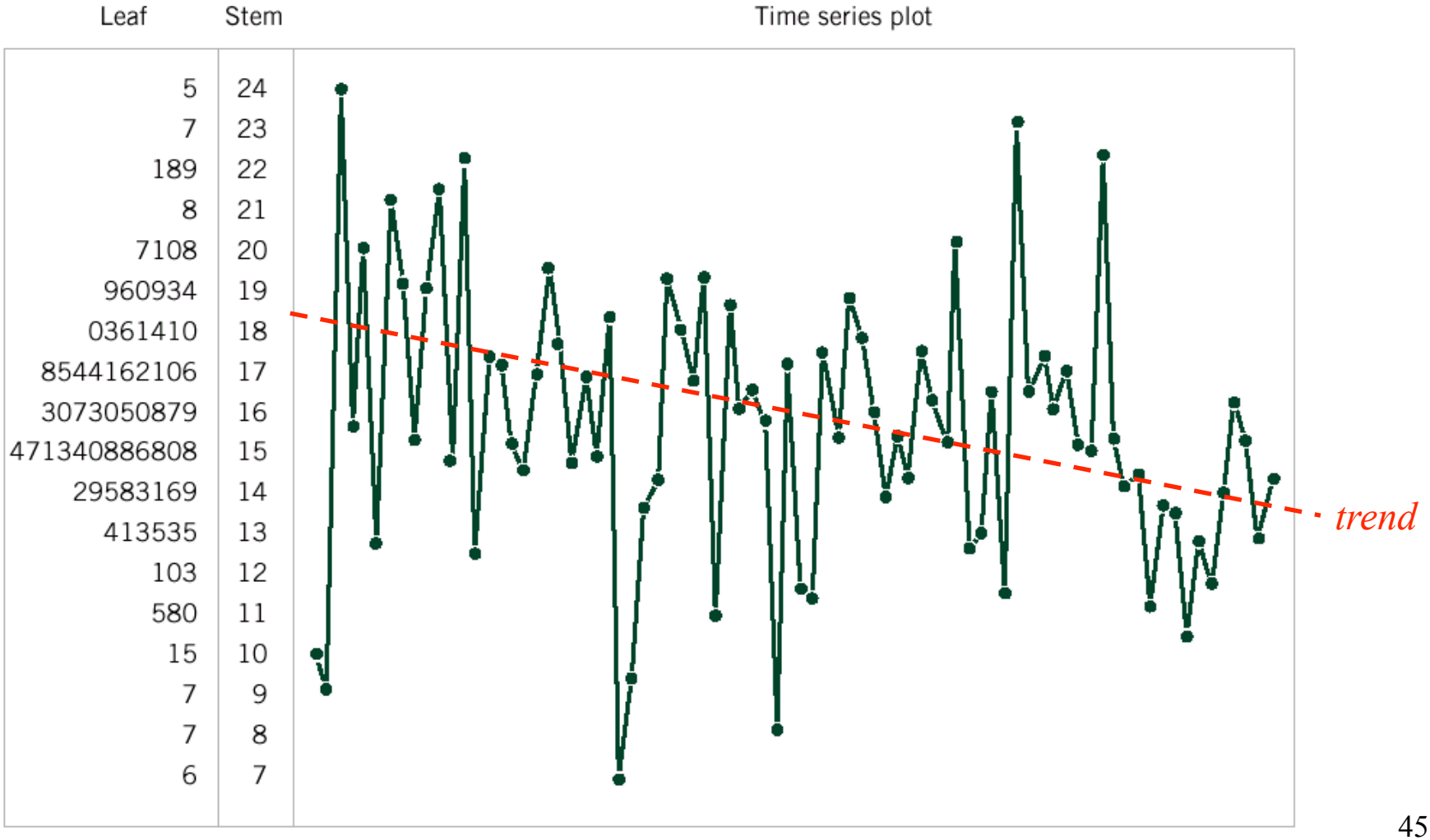
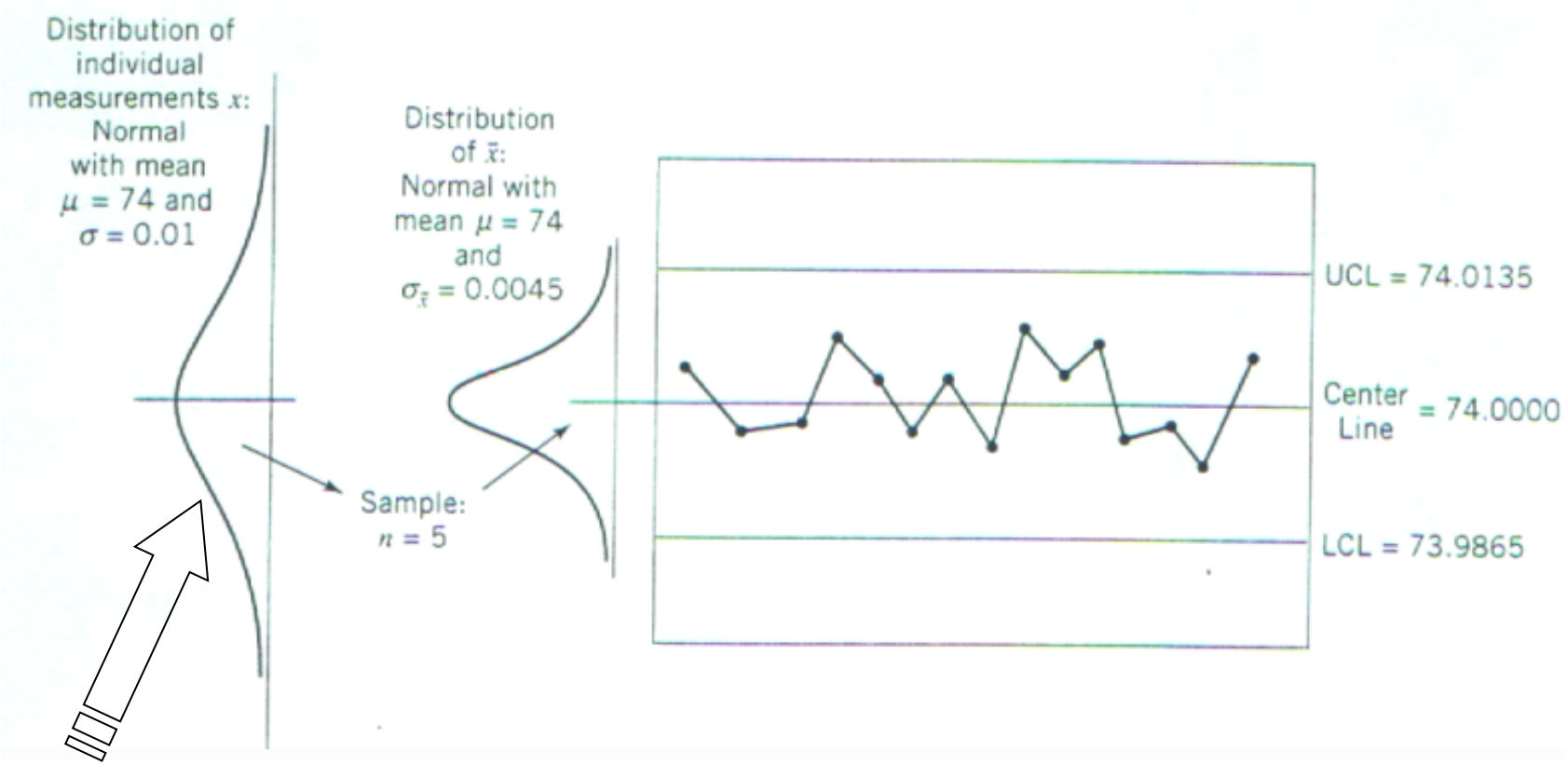


Grafico di controllo con PDF campionaria e del valor medio



PDF: limite dell'istogramma di frequenza per larghezza della classe → 0