Corso di Studio SM13 – CHIMICA

Introduzione alla chemiometria ed al disegno sperimentale (018CM)

27/10/2022 (lezione 2)

Pierluigi Barbieri

UNIVERSITÀ DEGLI STUDI DI TRIESTE

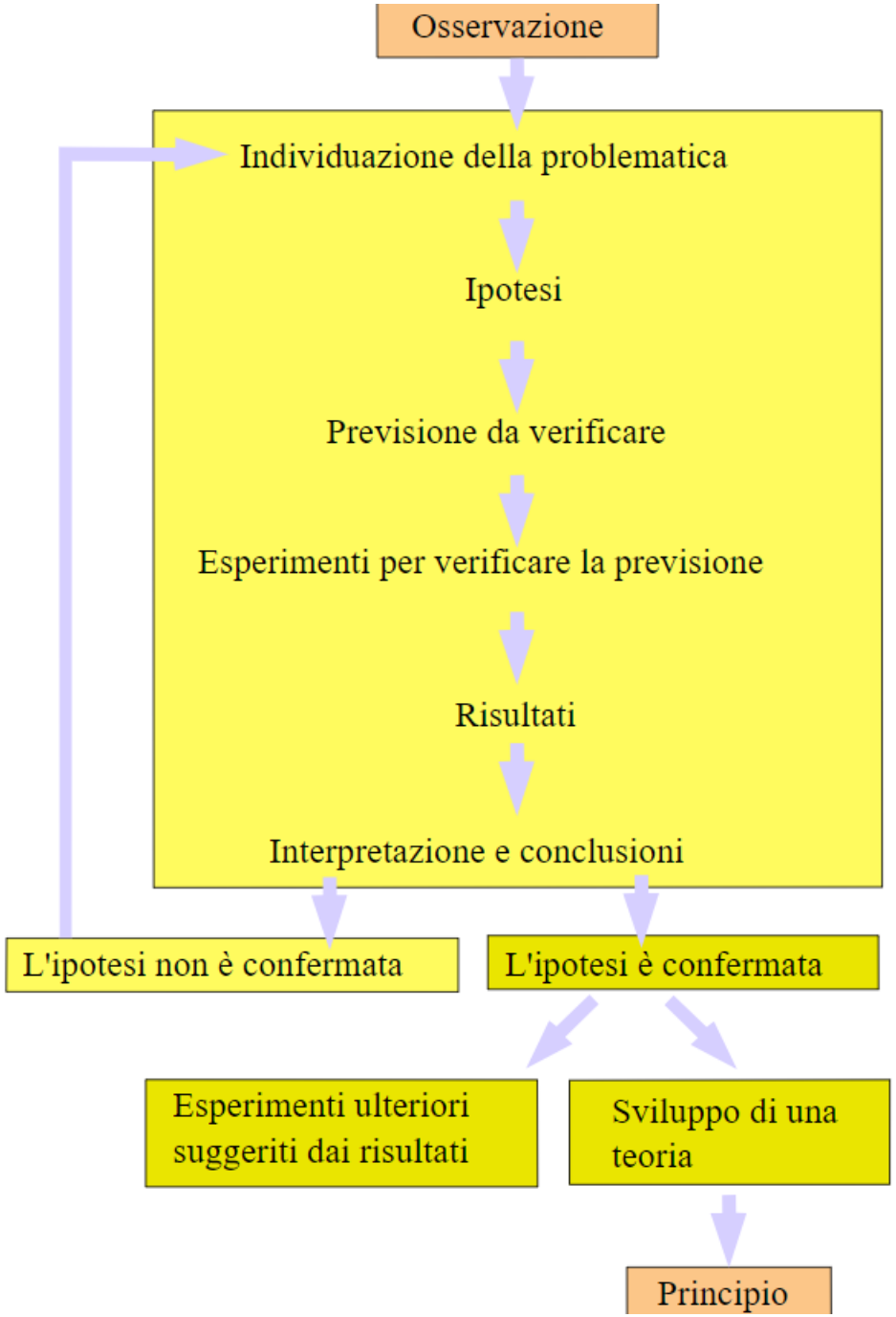
Dipartimento di Scienze Chimiche e Farmaceutiche

Gruppo di ricerca in Scienze Analitiche applicate alle InterazionI Uomo-Ambiente (ASHEI)

INTRODUZIONE ALLA CHEMIOMETRIA E DISEGNO SPERIMENTALE
018CM - Corso di StudioSM13 - CHIMICA

- ***1. Approcci statistici classici: indicatori e statistiche univariate. Descrizione di descrittori numerici semplici delle raccolte di dati: statistiche descrittive parametriche e non parametriche. Visualizzazione di statistiche uni- e bi-variate. ~~Esempi pratici in ambiente R software.~~***
- 2. Visualizzazione di set di dati reali con diverse tecniche di rappresentazione grafica, analisi dei risultati tramite esplorazione visiva, tecniche di individuazione di dati anomali. Illustrazione del concetto di carta di controllo dei dati. Esempi pratici in ambiente R software.

Metodo Scientifico induttivo



CONOSCENZA INDUTTIVA

Quantificazione:
Misure,
Produzione di dati/valori
Su sistemi chimici

Es. l'assorbimento di una radiazione elettromagnetica ad una certa λ da parte di una soluzione dipende dal cammino ottico e dalla concentrazione degli analiti che assorbono alla λ considerata, per soluzioni diluite.

Osservazione L'*osservazione* è il punto di partenza (e di arrivo) del ciclo di acquisizione della conoscenza nel senso che costituisce lo stimolo per la ricerca di una legge che governa il fenomeno osservato ed anche la verifica che la legge trovata sia effettivamente sempre rispettata. Si tratta di identificare le caratteristiche del fenomeno osservato, effettuando delle misurazioni adeguate, con metodi esattamente riproducibili. In fisica, infatti, tale parola è spesso usata come sinonimo di misura.

Esperimento L'*esperimento*, dove possibile, è programmato dall'osservatore che perturba il sistema e misura le risposte alle perturbazioni. Esistono tecniche di programmazione sperimentale, che consentono di porsi nelle condizioni migliori per perturbare in maniera minimale, ma significativa, al fine di osservare le risposte nel migliore dei modi.

Correlazione fra le misure L'analisi della correlazione fra le misure, che si colloca nel ciclo immediatamente dopo la fase di osservazione, costituisce la parte iniziale del patrimonio tecnico-scientifico utilizzabile per la costruzione del modello. Il dato grezzo, che è costituito in genere da tabelle di misure, può venire *manipolato* in vari modi, dalla costruzione di un grafico alla trasformazione logaritmica, dal calcolo della media alla interpolazione tra i punti sperimentali, utilizzando i metodi della statistica descrittiva.

Bisogna prestare attenzione nella scelta del tipo di funzione che correla i dati perché, citando Rescigno[24], le modulazioni dei dati ne cambiano il contenuto informativo. Infatti, se le manipolazioni mettono in evidenza alcune informazioni contenute nei dati, possono eliminarne altre. Quindi il contenuto informativo può diventare *inferiore* a quello dei dati originali.

Modello fisico Per facilitare il compito di scrivere la legge che esprime l'andamento di un certo fenomeno, si costruisce mentalmente un modello fisico, con elementi di cui si conosce il funzionamento, e che si suppone possa rappresentare il comportamento complessivo del fenomeno studiato.

Va notato che spesso un medesimo fenomeno può venire descritto con modelli fisici, e quindi anche con modelli matematici, diversi. Ad esempio i gas possono essere considerati come fluidi comprimibili oppure come un insieme di molecole. Le molecole possono essere pensate come puntiformi oppure dotate di una struttura; fra di loro interagenti oppure non interagenti: tutti modelli diversi. Ancora, la luce può venire considerata un fenomeno ondulatorio oppure un flusso di particelle e così via.

L'empirismo radicale sostiene che non è possibile avanzare oltre la conoscenza contenuta nei dati grezzi e quindi rifiuta il fatto che la conoscenza induttiva, sulla quale si fondano leggi empiriche e modelli, costituisca nuova conoscenza. Viceversa, la posizione realista è molto più flessibile e consente di parlare anche di concetti non direttamente osservabili, come la forza di attrazione gravitazionale o il campo elettromagnetico, la cui conoscenza è resa possibile adattando opportuni modelli all'osservazione degli effetti di tali entità e utilizzando a fondo le possibilità dell'induzione.

Chimica analitica

= la ***scienza dell'informazione*** chimica

Chemiometria

= una ***data science*** "chimica''

Dati chimici = misure, rapporto tra grandezza molecolare o atomica rilevata e grandezza di riferimento (un «metro» -> metrologia)

1.10 Misurazione

processo volto a ottenere sperimentalmente uno o più **valori** che possono essere ragionevolmente attribuiti a una **grandezza** (VIM 2.1). Una **misurazione** è costituita da una serie di azioni (fasi, stadi) che avvengono in maniera definita. Alcune **misurazioni** sono costituite da una singola fase, altre hanno più stadi. Questo concetto è in contrasto con l'opinione di alcuni che la **misurazione** sia quanto indicato da uno strumento, ad esempio per un'aliquota di un estratto di un campione. Invece, è chiaro che il termine **misurazione** si riferisce **all'*intero processo* necessario per ottenere** il **valore di una grandezza** e non dovrebbe essere utilizzato per indicare il valore numerico ottenuto.

https://www.eurachem.org/index.php/publications/guides/terminology-in-analytical-measurement VIM vocabulary in metrology

https://www.eurachem.org/images/stories/Guides/pdf/TAM_2011_IT.pdf Rapporti ISTISAN

È bene avere ***il Senso della misura*** (*e ‘‘della misurazione’’*)

Senso:

«Capire il significato di…»

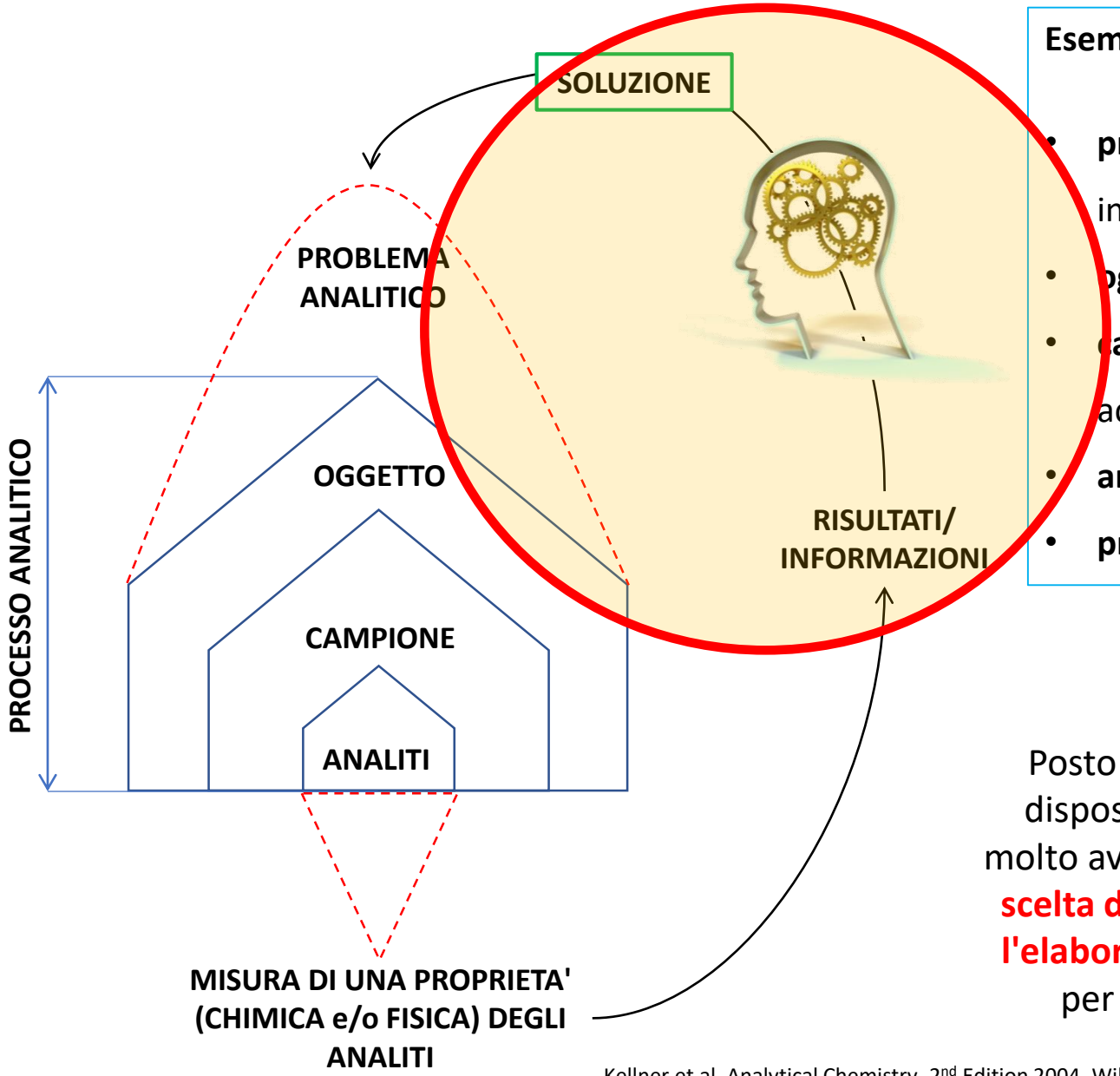
«dare valore a…»

«avere sensibilità per…»

Misura

Risultato di processo analitico / di una misurazione (definizione del rapporto/ relazione tra la proprietà di un *campione* la cui entità/grandezza è incognita ed un’altra omogenea, nota, presa a riferimento unitario)

Un **problema** scientifico, tecnico, sociale o economico origina un **problema analitico** che a sua volta richiede la messa a punto di un **processo analitico** per ottenere **risultati/informazioni** che vanno interpretati rispetto al problema posto.



Esempio:

- **problema analitico**: possibile inquinamento acqua di falda;
- **oggetto**: pozzo;
- **campione**: prelievo di aliquote di acqua;
- **analiti**: POPs;
- **proprietà**: spettro di massa

Posto che al giorno d'oggi si hanno a disposizione strumenti tecnicamente molto avanzati per effettuare le misure, **la scelta dell'oggetto/i e del campione/i** e **l'elaborazione dei risultati** sono **cruciali** per ottenere **soluzioni adeguate**

Kellner et al. Analytical Chemistry, 2nd Edition 2004, Wiley

Raccolta di informazione da sistemi chimici

Approccio Univariato: *interesse per singolo aspetto di un sistema chimico, misure ripetute di una caratteristica selezionata (approccio semplice e in alcuni casi efficace)*

Ma esistono anche

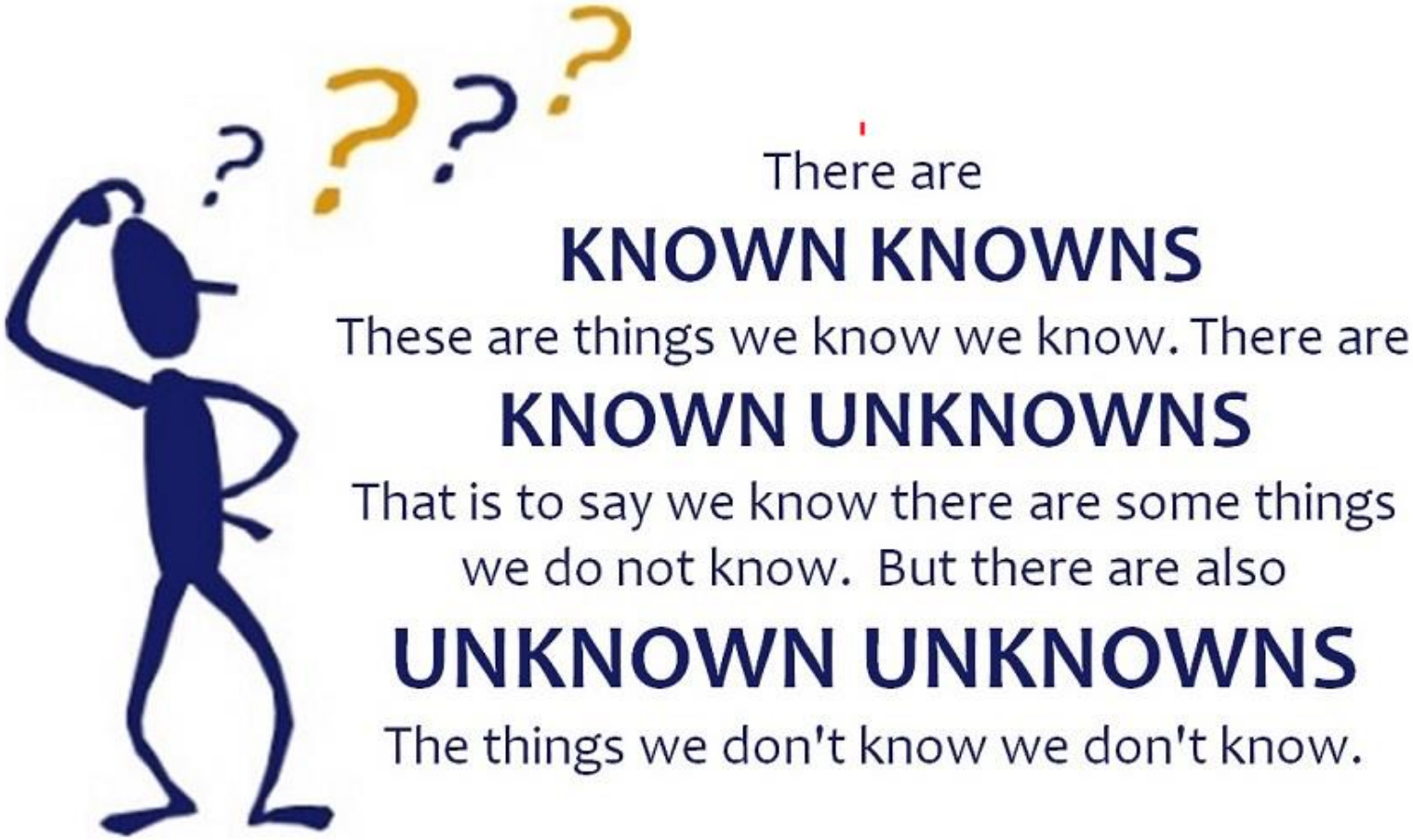
Sistemi multi- analita

multi- sensore

spettri associati a assorbimento ed emissione di radiazione elettromagnetica e spettri di massa (anche da nuove tecnologie es. trappola ionica orbitale)

Molta informazione – vedremo…

Ad abundantiam…



For example, we refer to a compound suspected to be present in a mixture whose identity is to be confirmed by mass spectrometric analyses as a “known known.”
A “known unknown” is a compound that is unknown to the investigator, but is cited in the chemical literature or mass spectrometry reference databases.
Lastly, an “unknown unknown” is a compound that is not previously cited.

https://doi.org/10.1007/s13361-010-0034-3

Statistiche univariate di una raccolta di dati

Approccio Univariato: *interesse per singolo aspetto di un sistema chimico, misure ripetute di una caratteristica selezionata (approccio semplice e in alcuni casi efficace)*

Statistical Methods In analytical Chemistry

Peter Meier
Richard Zund

Wiley, 2000

«Una (singola) misurazione non è una misurazione»: fare repliche della misurazione

Riproducibilità sperimentale indica misure affidabili

Misure hanno una componente determinata e una componente stocastica/casuale/random (rumore)

La componente determinata è il valore atteso che è identificato da valore mediato di misure replicate, ad esempio come media aritmetica

https://www.pdfdrive.com/statistical-methods-in-analytical-chemistry-chemical-analysis-a-series-of-monographs-on-analytical-chemistry-and-its-applications-e156630510.html

A seguito di una raccolta di misure - *assumendo che ci sia una distribuzione normale dei valori del misurando* - possiamo definire modi semplici per identificare qual è il valore tipico che rappresenta al meglio questi valori.

Metodi semplici per trovare il valore più probabile , e l'intervallo di confidenza per comparare i risultati.

Quando iniziamo uno studio, impostiamo la sperimentazione (raccolta dati), avendo un modello in mente, ad esempio

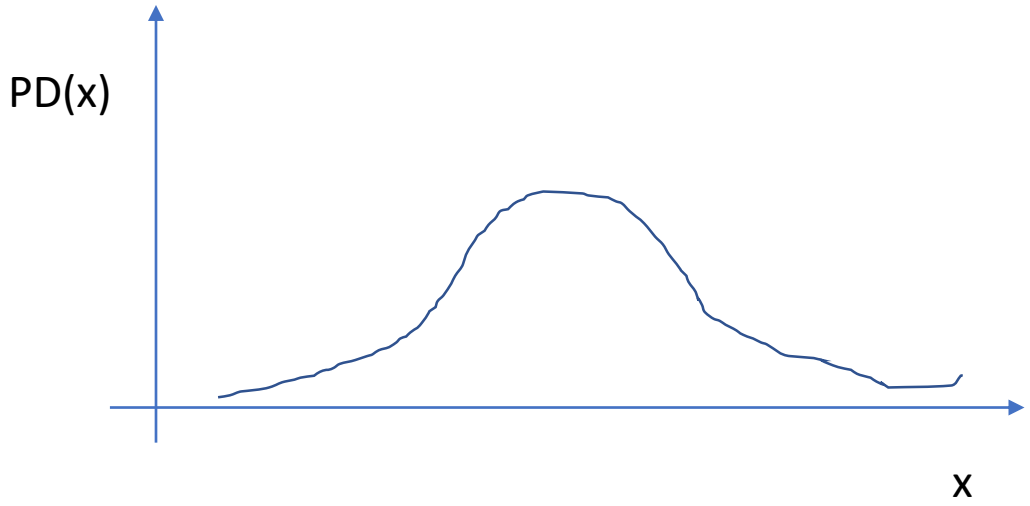
Che la grandezza da misurare abbia una *distribuzione normale*, descritta da una funzione di densità di probabilità a campana, espressa analiticamente come

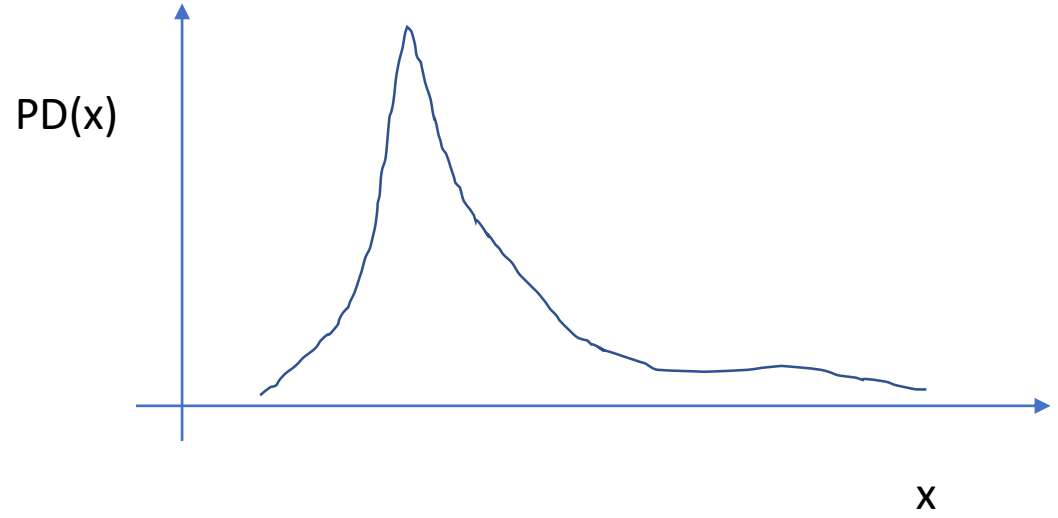
$$PD = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot \exp\left(\frac{-1}{2} \cdot \left(\frac{x - \mu}{\sigma}\right)^2\right)$$

Dove x è il valore osservato, μ è il valore vero dedotto dalla teoria o da un gran numero di osservazioni, σ la deviazione standard vera dedotta dalla teoria o da un gran numero di osservazioni, PD la densità di probabilità in funzione di x

Distribuzione di valori («modello» della variabilità del valore vero) non è sempre identica

Es distribuzione gaussiana , d. gamma





Misura tratta da un insieme di possibili realizzazioni

(Campionamento?)

Caratterizziamo un’entità finita

Per poche misure e in casi di asimmetria

Per determinare il valore più probabile

la mediana (valore che biseca l'insieme delle osservazioni ordinate) è statistica più appropriata della media .

La mediana è poco influenzata da dati anomali. E' una statistica robusta

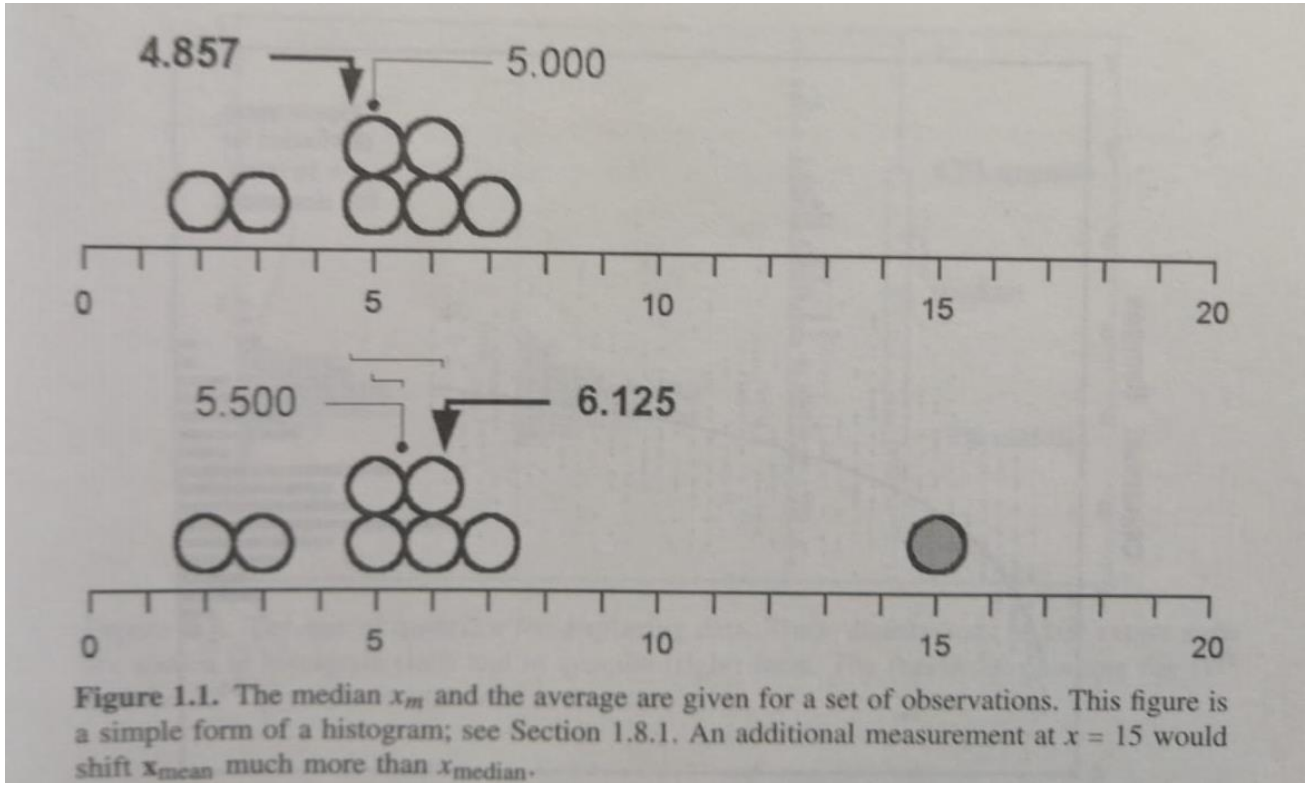


Figure 1.1. The median x_m and the average are given for a set of observations. This figure is a simple form of a histogram; see Section 1.8.1. An additional measurement at $x = 15$ would shift $\mathbf{x}_{\text{mean}}$ much more than x_{median}.

Data set x()=2,3,5,5, 6,6,7, a cui aggiungiamo una misura di 15. Come cambiano media e mediana?

Dispersione

Affidabilità delle media è valutata da distribuzioni delle misure individuali attorno alla media .

Dispersione può essere valutata come intervallo o come deviazione standard σ_x

Intervallo R definito come $x_{max} - x_{min}$, valore influenzato da estremi (dati anomali)

Nel caso di n (numero di osservazioni) > 9 è possibile ricorrere ai quantili , tagliando i dati estremi ottenendo una stima robusta della dispersione (es box con 2/3 dei dati (scatola tra 17° e 83° percentile, corrisponde a intervallo $\mu \pm 1\sigma$ in una distribuzione normale, che contiene 68,3% delle osservazioni)

$$S_{xx} = \Sigma(x_i - \mathbf{x}_{\text{mean}})^2 = \Sigma(r_i)^2$$
$$S_{xx} = \Sigma(x_i^2) - (\Sigma x_i)^2/n$$
$$V_x = S_{xx}/(n-1)$$
$$s_x = \sqrt{V_x}$$

S_{xx} = *Somma dei quadrati dei residui*

V_x = *Varianza di x*
(n-1) gradi di libertà (fissata la media e (n-1) x_i
l'n° è determinato
s_x = *deviazione standard di x*

RSD o coefficiente di variazione = 100 $S_{x\,/}\,X_{mean}$ usato per comparare riproducibilità di serie di misure

Indipendenza delle misure

Quando 2 campioni sono indipendenti? dipende

A campioni con trattamento e analisi offline

B sensore immerso in unmezzo di reazione che acquisce dati in continuo

Caso A campioni devono essere preparati individualmente

Ma attenzione per valutare la dispersione, dipende dalla tecnica

Per analisi UV : pesata e diluizione portano errore maggiore rispetto a lettura delle repliche (campione va ripreparato)

Per analisi HPLC : separazione per mobilità nella colonna hanno errore comparabile a quello della preparazione (ha senso reiniettare stesso campione)

I fattori e le operazioni che contribuiscono maggiormente alla varianza totale devono essere ripetuti

B : se registrazione continua, misure indipendenti si mediano su intervallo temporale che deve essere distante almeno 5 T, costante temporale dello strumento, dopo l'ultimo valore della mediazione precedente

Serve che il sensore si sia adattato alle nuove condizioni

La costante temporale dello stremento che rileva, viene determinata provocando una risposta con scalino di segnale (step response) vedi fig. successiva.

Si devono distinguere condizioni di campione statico e dinamico

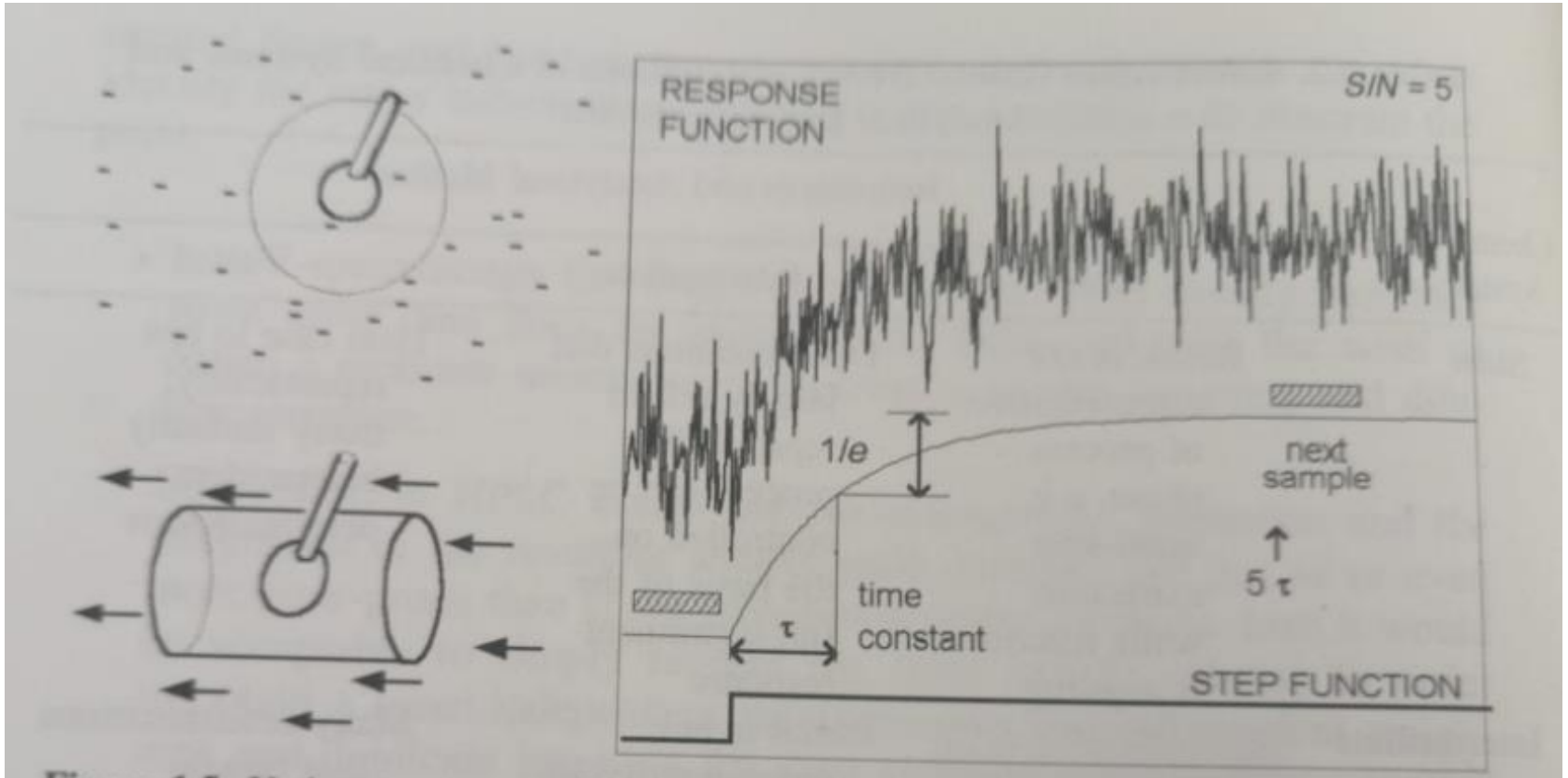


Figure 1.5. Under stagnant conditions a sensor will sample a volume, that is, the average response for that volume is obtained. A sensor in a current yields an average reading over time and cross-section. The observed signal S over time t is the convolute of the local concentration with the sensor's sampling volume and time constant. Two measurements are only then independent when they are separated at least by five time constants and/or a multiple of the sampling volume's diameter. At the left, the sampled volumes are depicted. At the right, a typical signal-*versus*-time record (e.g., strip-chart recorder trace) and the system response to a step change in concentration are shown. Tau (τ) is the time constant defined by an approximately 63.2% change $(1 - 1/e) = 0.63212$, with $e = 2.71828 \ldots$. The hatched bars indicate valid averages taken at least $5 \cdot \tau$ after the last disturbance.

Statistical
Methods
In analytical
Chemistry

Peter Meier
Richard Zund

Wiley, 2000

Ripetibilità e riproducibilità per valutare errori casuali

Ripetibilità σ_x considerando in un tempo breve misure di uno stesso laboratorio, strumento, operatore,

Riproducibilità considera additività delle varianze e si riferisce a stima della dispersione di misure in un tempo lungo

$$V_{\text{reprod}} = V_{\text{repeat}} + V_{\text{temp}} + V_{\text{operator}} + V_{\text{chemicals}} + V_{\text{work-up}} + V_{\text{population}} + \ldots$$

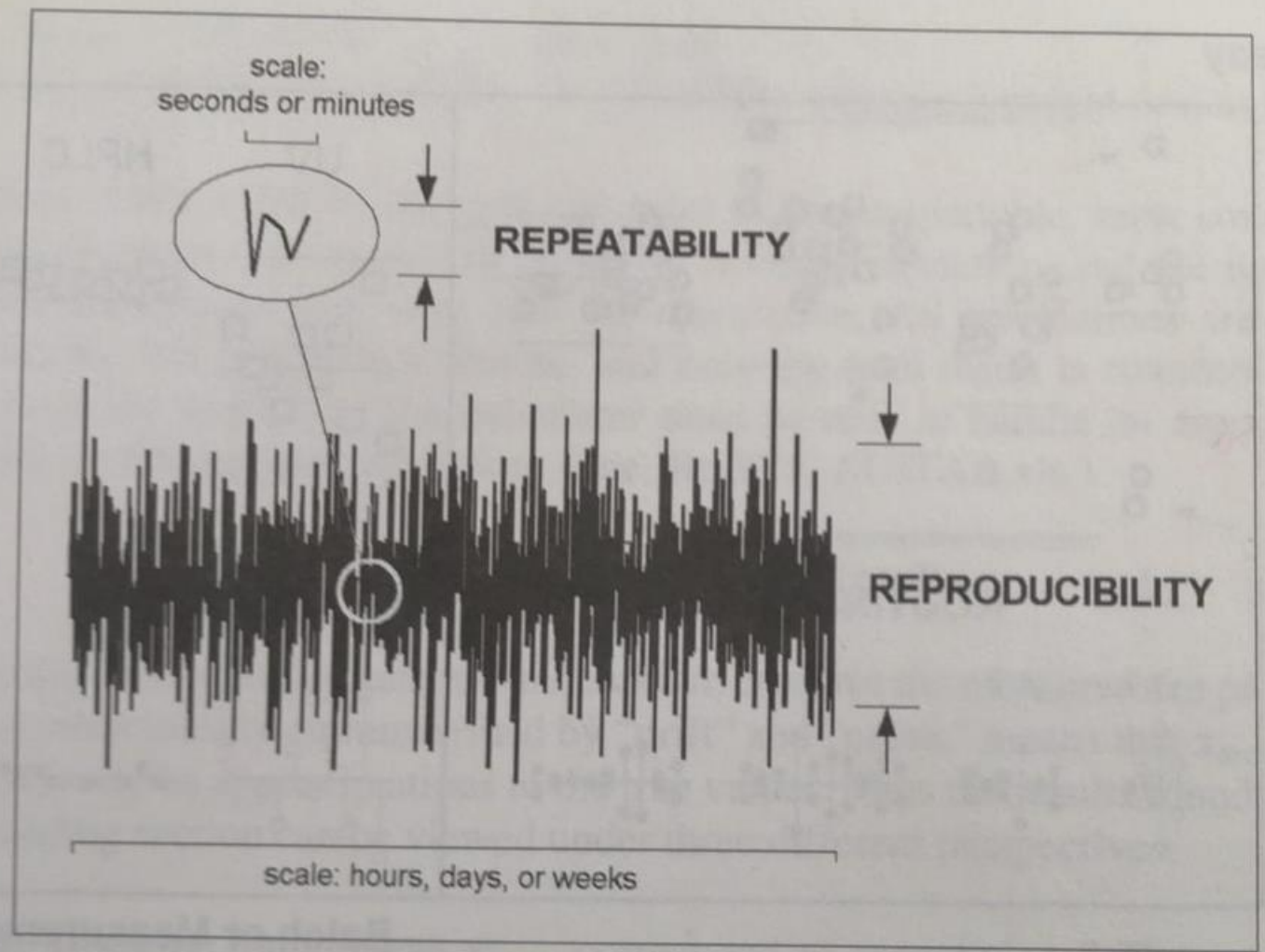


Figure 1.6. Repeatability and reproducibility are defined using historical data. The length of the time interval over which the parameter is reviewed is critical: the shorter it is, the better defined the experimental boundary conditions tend to be; the repeatability sets the limit on what could potentially be attained, the reproducibility defines what is attained in practice using a given set of instrumentation and SOPs.

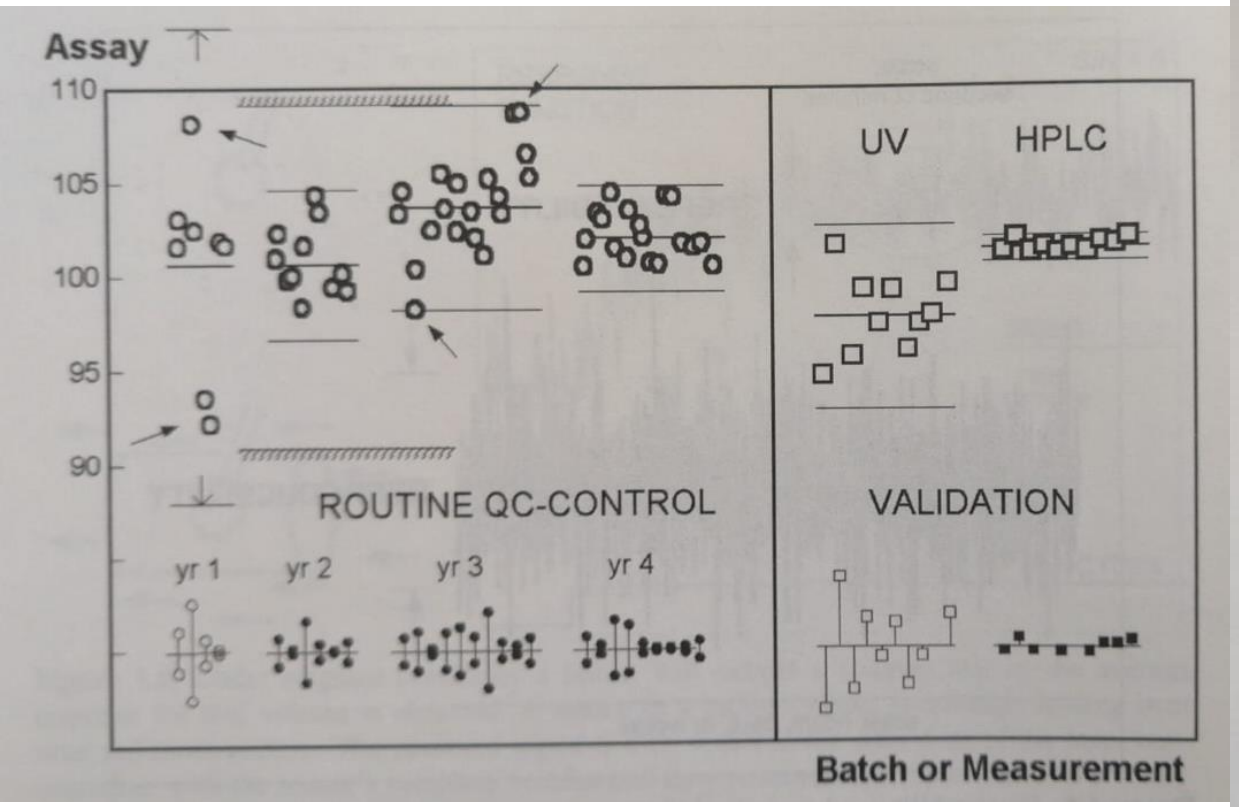


Figure 1.7. Reproducibility and repeatability. For a cream the assay data for the active principle is shown for retrospective surveys (left) and validation runs (right). This particular product is produced in about 20 batches a year. At the end of every year, product release analysis data for a number of randomly picked batches is reviewed for an overall picture of the performance of the laboratory. In four successive years, 30 batches (circles) were investigated; the repeat determinations are given by simple (UV) and bold (HPLC) and the respective mean $\mathbf{x}_{mean}$ and CL(x) are indicated by horizontal lines; the CL($\mathbf{x}_{mean}$) are given by the symbols bars. For definitions of CL see Section 1.3. The residuals for the double determinations are shown below (dots). The following conclusions can be drawn: (a) All data are within the ±9.1% specifications (hatched bars), because otherwise the releases would not have been granted; (b) The mean of the third group is higher than the others ($p < 0.025$, 95% CL being shown); (c) Four pairs of data points are marked with arrows; because the individual points within a pair give typical residuals, either one of three artifact-causing mechanisms must be investigated: (1) over- or under-dosing during production, (2) inhomogeneity, and (3) errors of calibration. Points 1 and 2 can be cleared up by taking more samples and checking the production records; point 3 is a typical problem found in routine testing laboratories (deadlines, motivation). This is a reason why Good Manufacturing Practices (GMP) regulations mandate that reagent or calibration solutions be marked with the date of production, the shelf life, and the signature of the technician, in order that gross mistakes are avoided and such questions can retrospectively be cleared. In the right panel, validation data for an outdated photometrical method (squares) and the HPLC method (bold squares) are compared. HPLC is obviously much more reliable. The HPLC-residuals in the righthand panel (repeatability, same technician, day, and batch) should be compared with those in the lefthand panel (reproducibility: several technicians, different days and batches) to gain a feeling for the difference between a research and a routine lab.

QA/QC: Quali fattori intervengono nel determinare il risultato dell'analisi olfattometrica?

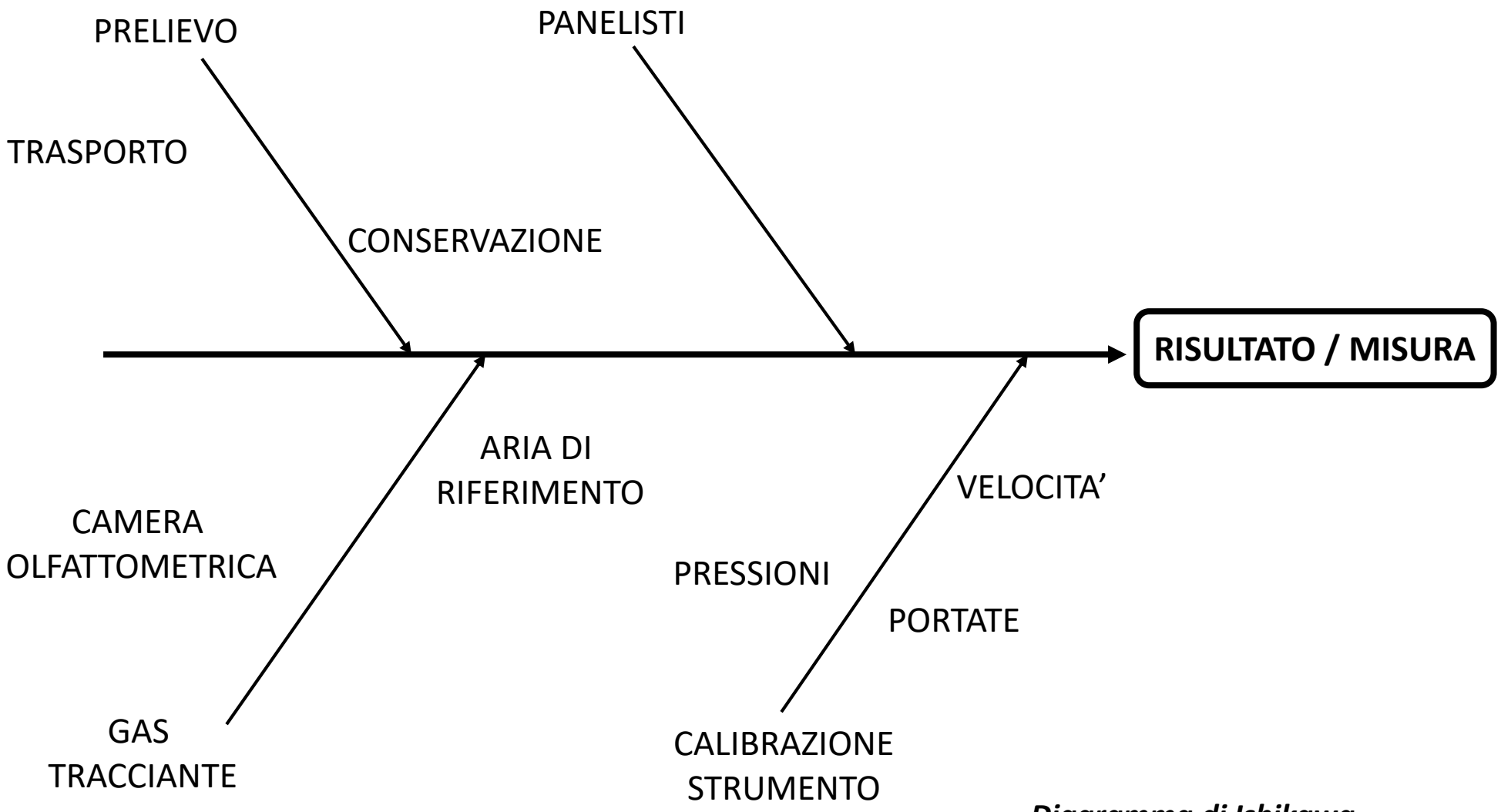


Diagramma di Ishikawa

Modulo 2:

Approcci statistici classici: indicatori e statistiche univariate

Dati ottenuti da studi/esperimenti

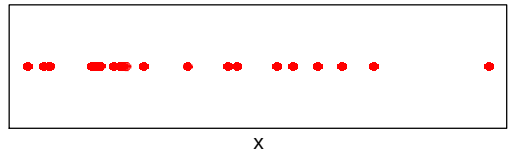
In genere l'"aspetto" di un set di dati ricavati da uno studio o un esperimento, se correttamente organizzato dovrebbe essere rappresentato in questo modo:

Campione/ osservazione	Variabile/ Proprietà $_1$	Variabile/ Proprietà $_2$	...	...	Variabile/ Proprietà $_m$
C_1					
C_2					
...					
...					
C_n					

Come esplorare i dati per estrarre informazione?

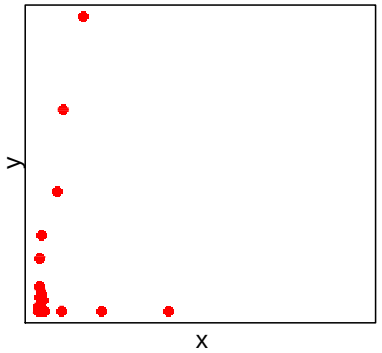
Osservare il comportamento di **una** variabile alla volta

Campione/ osservazione	Variabile/ Proprietà $_1$	Variabile/ Proprietà $_2$	...	...	Variabile/ Proprietà $_m$
C_1					
C_2					
...					
...					
C_n					



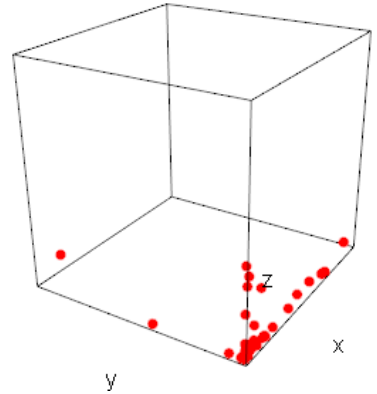
Osservare il comportamento di **coppie** di variabili

Campione/ osservazione	Variabile/ Proprietà $_1$	Variabile/ Proprietà $_2$	...	...	Variabile/ Proprietà $_m$
C_1					
C_2					
...					
...					
C_n					



Osservare il comportamento di variabili a **gruppi di 3**

Campione/ osservazione	Variabile/ Proprietà $_1$	Variabile/ Proprietà $_2$	...	...	Variabile/ Proprietà $_m$
C_1					
C_2					
...					
...					
C_n					



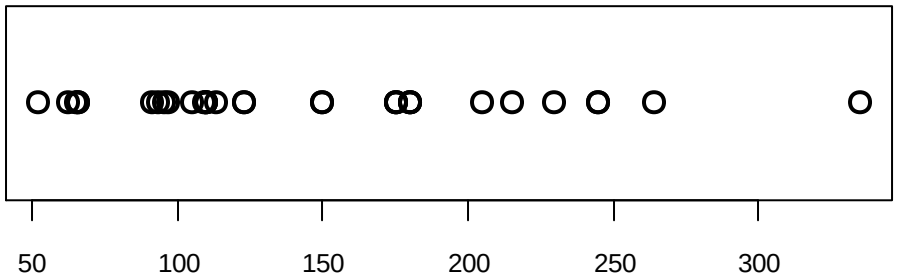
Osservare il comportamento di **nuove variabili** (di solito 2-3) che "rappresentano" la variabilità delle variabili sperimentali

Campione/ osservazione	Variabile/ Proprietà $_1$	Variabile/ Proprietà $_2$	...	...	Variabile/ Proprietà $_m$
C_1					
C_2					
...					
...					
C_n					

ANALISI MULTIVARIATA

Analisi statistica descrittiva univariata

La statistica descrittiva è un insieme di tecniche usate per descrivere le caratteristiche di base dei dati raccolti in un esperimento/studio.



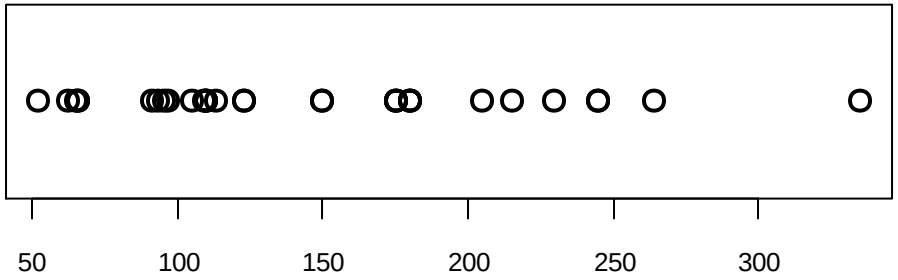
Per descrivere una distribuzione univariata si utilizzano:

- **Misure di tendenza centrale**
- **Misure di dispersione**
- **Misure di forma**

Fonti: http://docenti.unimc.it/cristina.davino/teaching/2014/12821/files/statistiche-univariate/l2-indici-di-tendenza-centrale

Misure di tendenza centrale (o di posizione)

In genere una serie di dati omogenea (cioè più misure di una unica variabile) tende a raggrupparsi intorno ad alcuni valori caratteristici (di solito centrali) che consentono di descrivere la serie stessa.

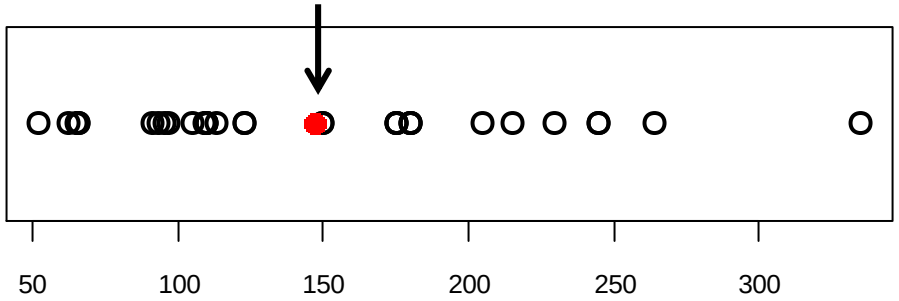


Tipi di misure di tendenza centrale sono:

- **Media**
- **Mediana (e quartili)**
- **Moda**

Fonti: http://docenti.unimc.it/cristina.davino/teaching/2014/12821/files/statistiche-univariate/l2-indici-di-tendenza-centrale

Misure di tendenza centrale: la media



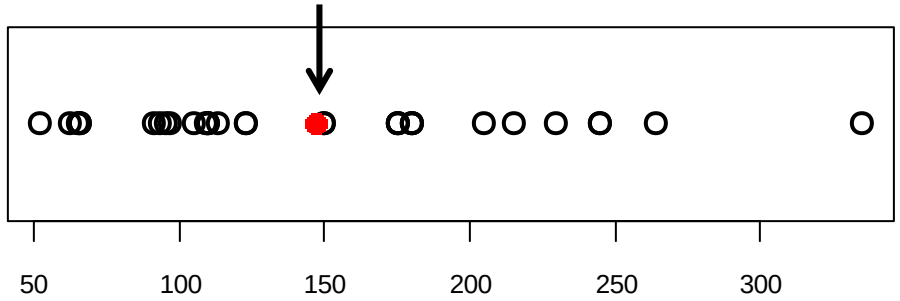
$$\underbrace{\bar{x} = \mu}_{\text{Media}} = \frac{\sum_{i=1}^{n} x_i}{n}$$

x_i → valore assegnato all' i-esima osservazione

n → numero di osservazioni

Fonti: http://docenti.unimc.it/cristina.davino/teaching/2014/12821/files/statistiche-univariate/l2-indici-di-tendenza-centrale

Proprietà della media



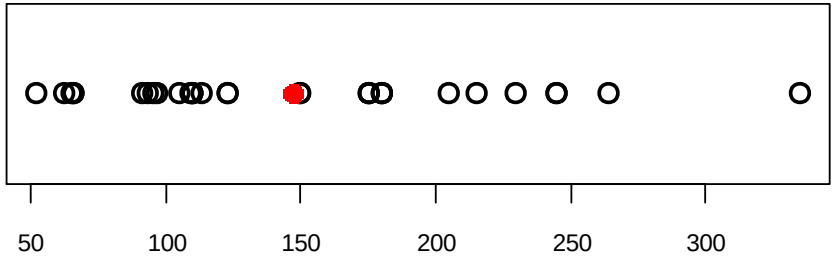
Alcune proprietà della media sono:

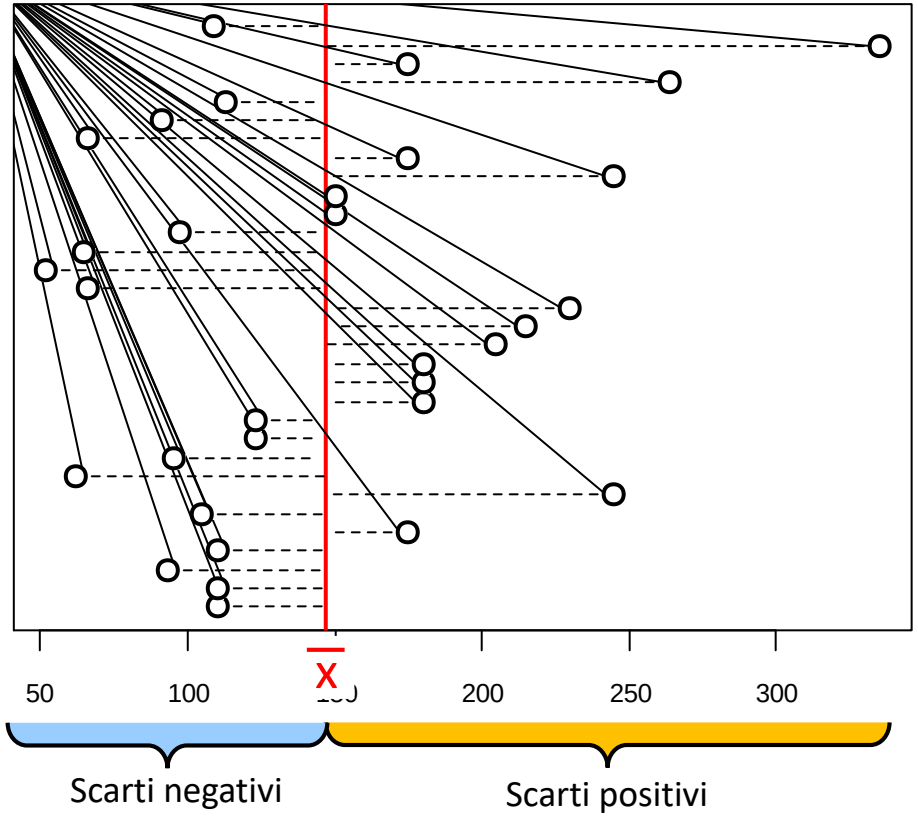
- **Internalità:** la media si trova sempre tra il minimo e il massimo della distribuzione di dati considerata;
- **Linearità:** una nuova variabile B data da B=aX+b avrà media uguale a $\overline{B}=a\overline{X}+b$ (ad es. se si converte in un'altra unità di misura);

SEGUE

Fonti: http://docenti.unimc.it/cristina.davino/teaching/2014/12821/files/statistiche-univariate/l2-indici-di-tendenza-centrale

Proprietà della media (2)

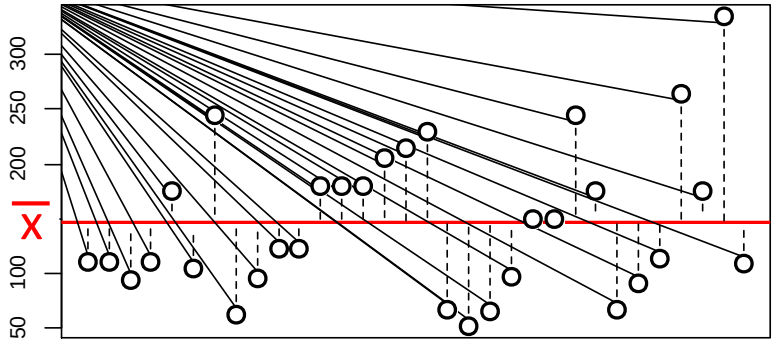




- **Baricentro:** la somma degli scarti dalla media è nulla:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$

Oppure, girando gli assi, si può "vedere" anche così:



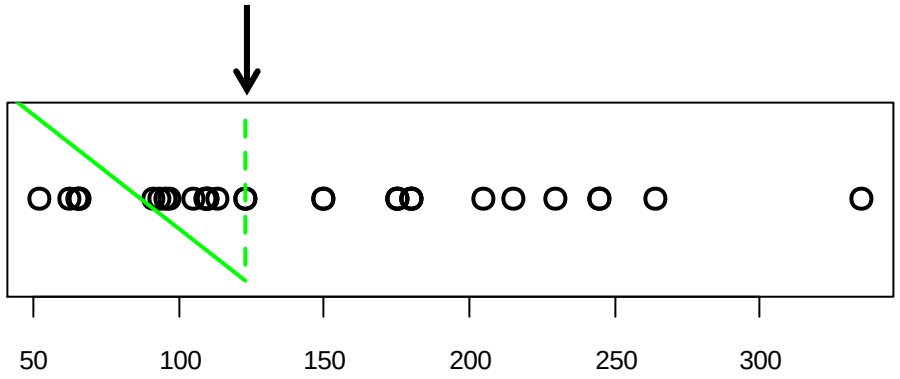
Fonti: http://docenti.unimc.it/cristina.davino/teaching/2014/12821/files/statistiche-univariate/l2-indici-di-tendenza-centrale

Misure di tendenza centrale: la mediana

La mediana è data dal valore dell'osservazione che si trova in posizione centrale nell'elenco dei valori assunti dalla variabile, dopo averli ordinati in modo crescente.

Come si effettua il calcolo della mediana:

1) I valori della variabile vengono ordinati dal più piccolo al più grande;
2) Si individua la posizione centrale a seconda che il numero **n** di osservazioni sia **pari** o **dispari**;
3) Si individua il valore corrispondente alla posizione centrale.



$$Med = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

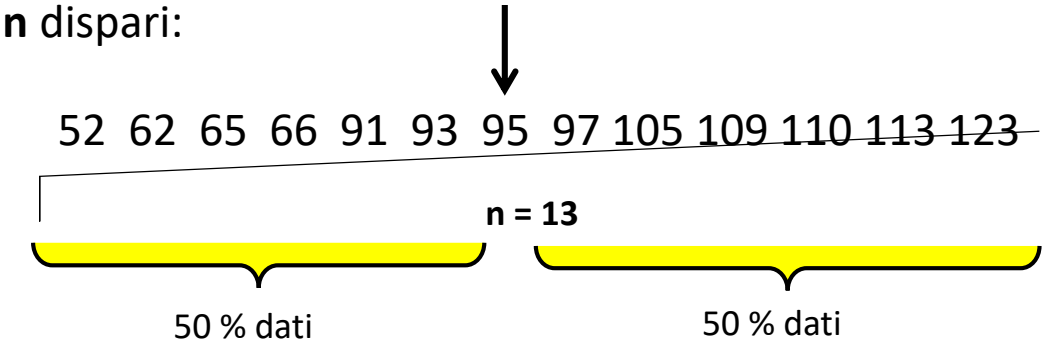
$$Med = X_{\frac{n+1}{2}}$$

SEGUE

Fonti: http://docenti.unimc.it/cristina.davino/teaching/2014/12821/files/statistiche-univariate/l2-indici-di-tendenza-centrale

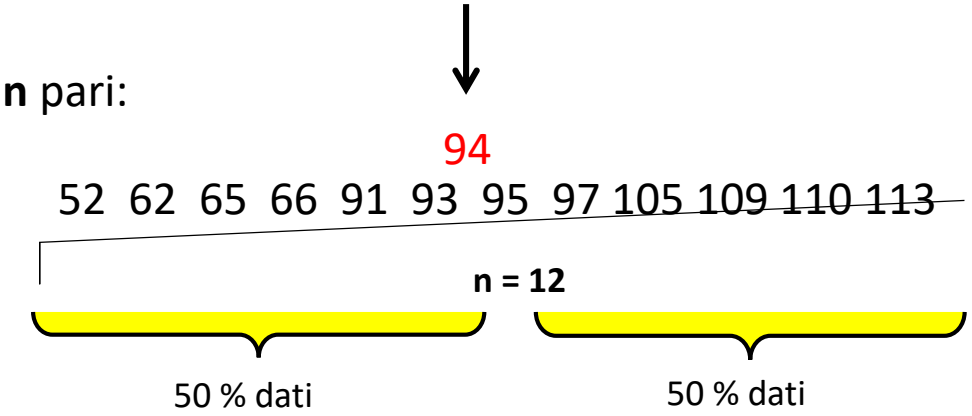
La mediana (2)

Esempio con **n** dispari:



$$Med = X_{\frac{n+1}{2}}$$

Esempio con **n** pari:



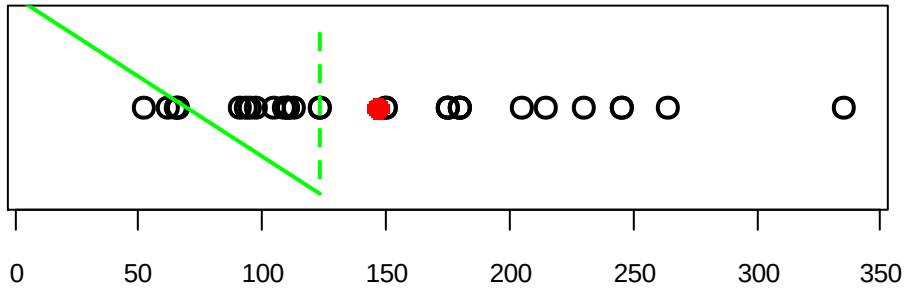
$$Med = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

La mediana (3)

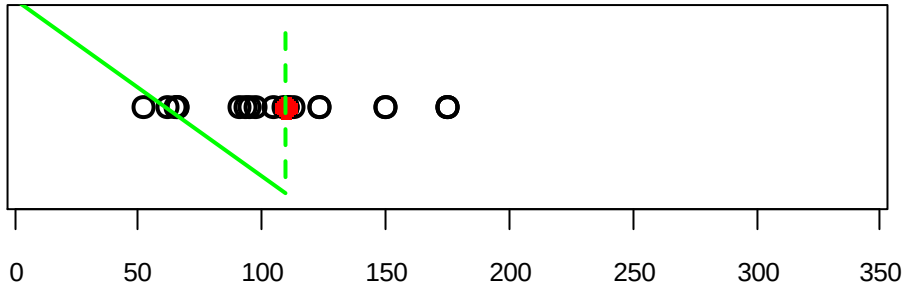
La mediana è anche detta "media robusta" perché il suo valore è meno influenzato da valori estremi rispetto alla media.

Tre situazioni possibili:

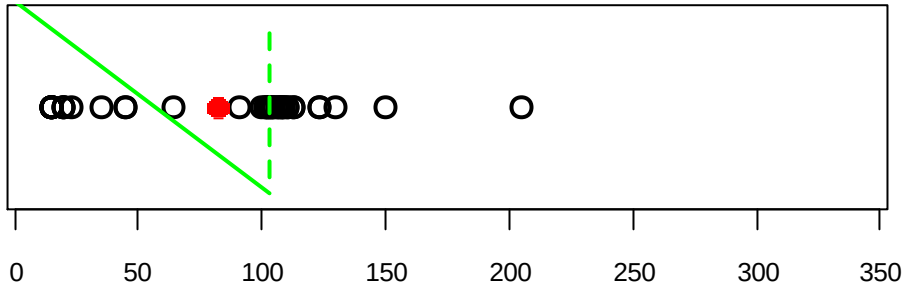
$\mu > Med$



$\mu = Med$



$\mu < Med$



Misure di tendenza centrale: i quartili

I quartili sono 3 valori che dividono i valori della variabile (già ordinati in ordine crescente) in quattro parti:

Q_1 = primo quartile

Q_2 = secondo quartile = mediana

Q_3 = terzo quartile

Q_1 = valore tale che il 25% dei valori della variabile è minore di esso e il 75% è maggiore

$$Q_1 = X_{\frac{n+1}{4}}$$

Q_3 = valore tale che il 75% dei valori della variabile è minore di esso e il 25% è maggiore

$$Q_3 = X_{\frac{3\cdot(n+1)}{4}}$$

SEGUE

Fonti: http://www.unife.it/economia/economia/insegnamenti/statistica/materiale-corso-a-a-2017-2018-gruppo-c/capitolo-03

Misure di tendenza centrale: i quartili (2)

Come si effettua il calcolo dei quartili:

1) I valori della variabile vengono ordinati dal più piccolo al più grande;

2) Si calcola il risultato di $\frac{n+1}{4}$ (oppure $\frac{3\cdot(n+1)}{4}$) ; → POSIZIONE

3)

a) Se il risultato è un numero intero allora i quartili sono i valori presenti alla posizione corrispondente;

b) Se il risultato è a metà tra due numeri interi allora il valore del quartile è la media dei valori corrispondenti ai due numeri interi;

c) Se il risultato è un numero reale che non si trova a metà tra due numeri interi, per individuare la posizione si approssima per eccesso o difetto al primo numero intero vicino, e il quartile sarà il valore corrispondente a quella posizione.

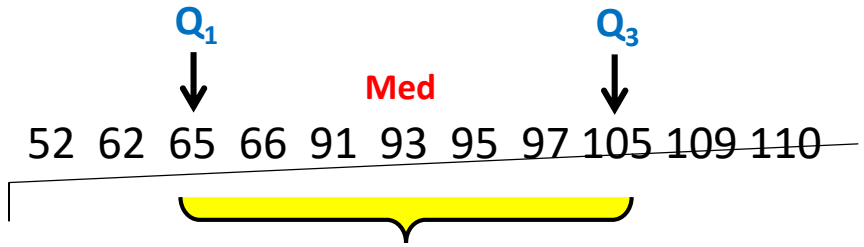
SEGUE

Fonti: http://www.unife.it/economia/economia/insegnamenti/statistica/materiale-corso-a-a-2017-2018-gruppo-c/capitolo-03

Misure di tendenza centrale: i quartili (3)

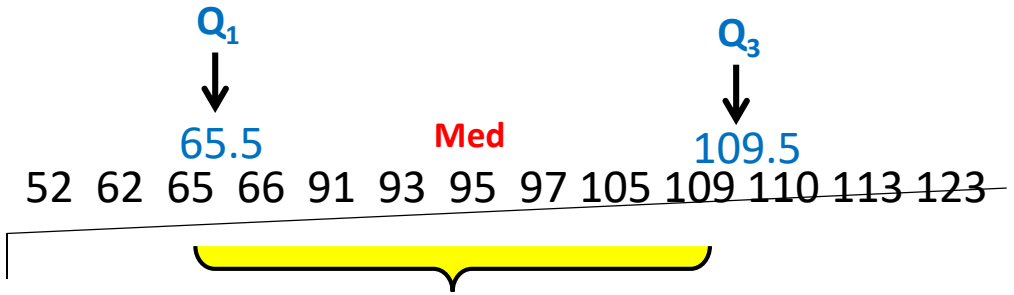
Esempi:

n = 11



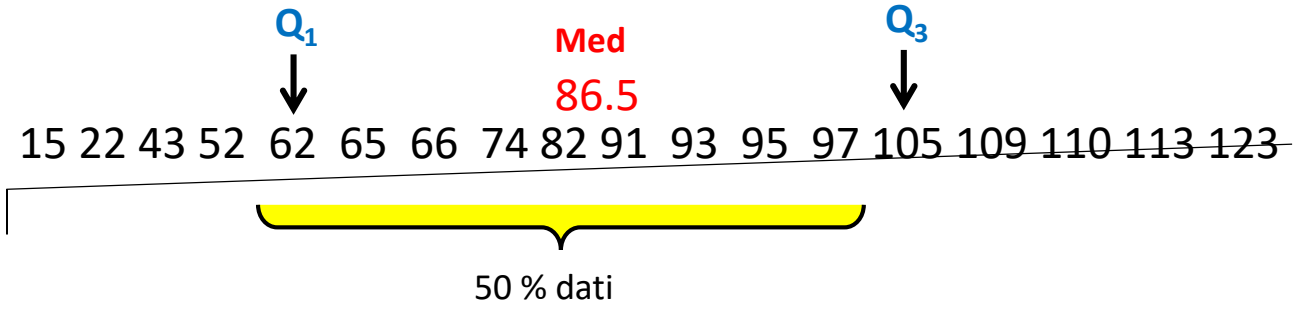
$\frac{n+1}{4} = 3$ $\frac{3\cdot(n+1)}{4} = 9$

n = 13



$\frac{n+1}{4} = 3.5$ $\frac{3\cdot(n+1)}{4} = 10.5$

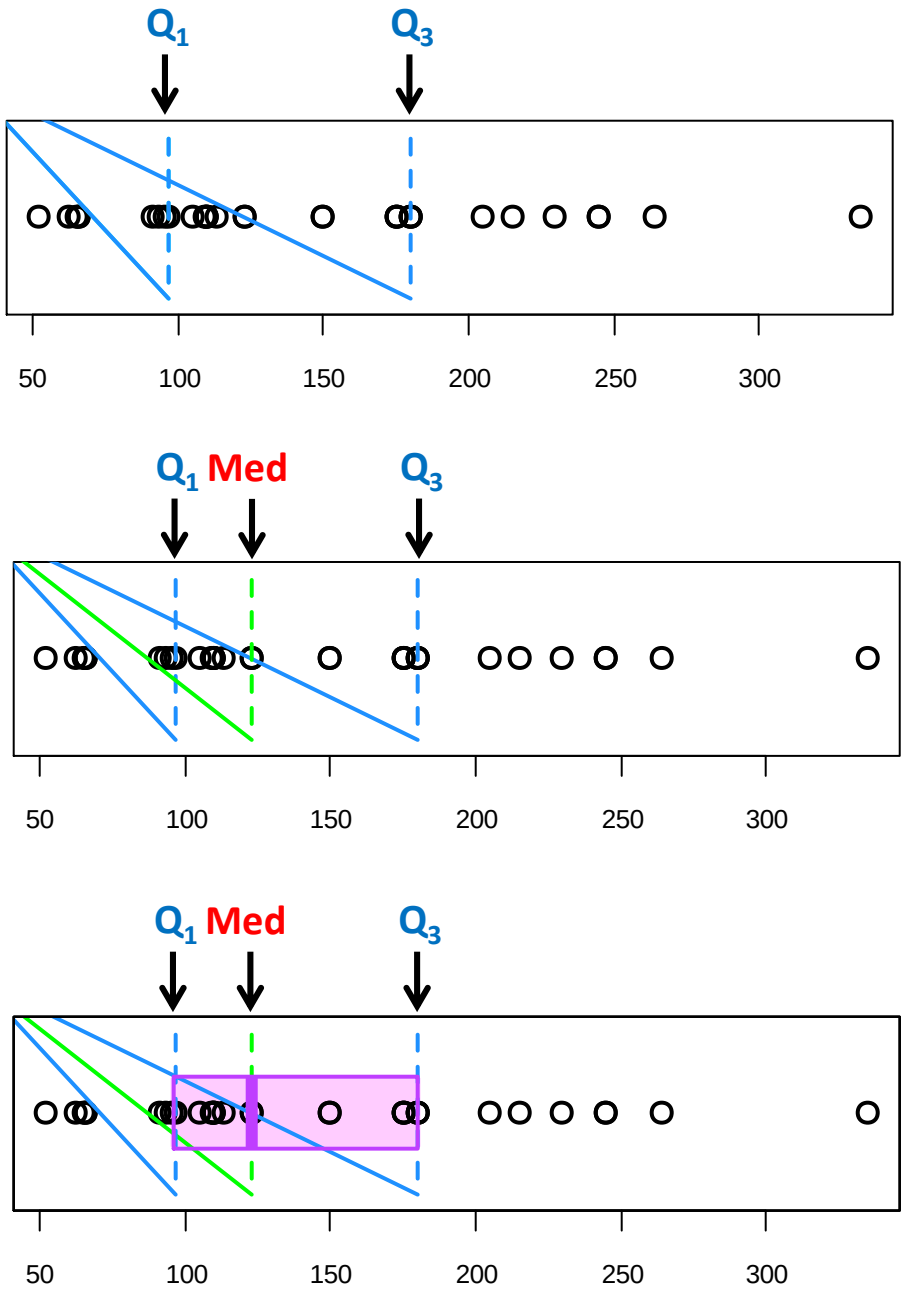
n = 18



$\frac{n+1}{4} = 4.75$ $\frac{3\cdot(n+1)}{4} = 14.25$

SEGUE

Misure di tendenza centrale: i quartili (4)



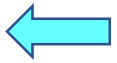
Misure di tendenza centrale: la moda

La moda è data dal valore rappresentato con maggiore frequenza nell'insieme di valori di una variabile, si può individuare in modo semplice quando si considerano variabili discrete

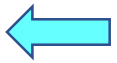
Una variabile può avere anche più di un valore rappresentato con alta frequenza, quindi il suo andamento può essere: unimodale (1 moda), bimodale (2 mode) o multimodale (più mode).

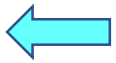
X	Freq
1	1
3	2
5	2
6	1
7	3
8	3
9	5
10	1
11	1

X	Freq
1	1
3	2
5	5
6	1
7	3
8	3
9	5
10	1
11	1



X	Freq
1	1
3	5
5	2
6	1
7	5
8	2
9	5
10	2
11	1





Misure di dispersione

Le misure di dispersione sono dei valori opportunamente calcolati che danno una indicazione sulla variabilità delle serie di valori registrati per una variabile.

Tipi di misure di dispersione sono:

- **L'intervallo di variazione (o *range*)**
- **L'intervallo di variazione interquartile**
- **Varianza**
- **Deviazione standard (o scarto quadratico medio)**
- **Coefficiente di variazione**

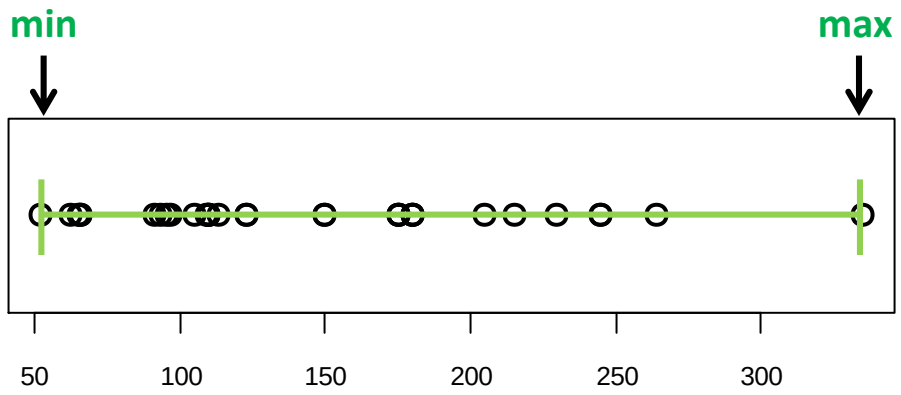
SEGUE

Fonti: http://www.unife.it/economia/economia/insegnamenti/statistica/materiale-corso-a-a-2017-2018-gruppo-c/capitolo-03

Misure di dispersione: gli intervalli di variazione

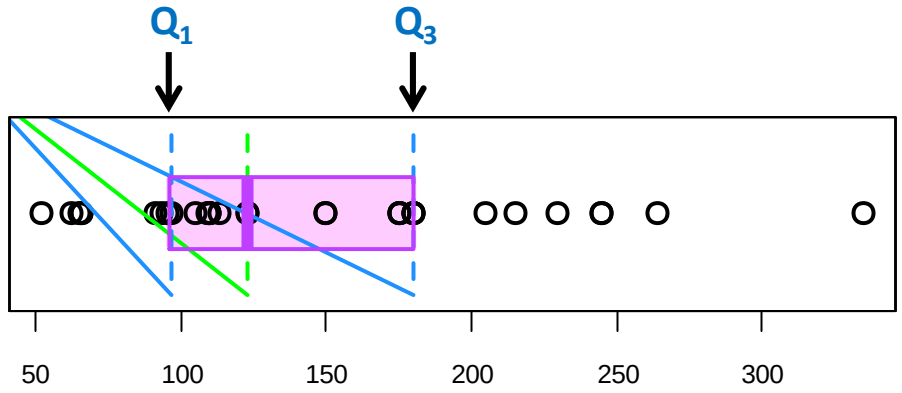
➢ L'intervallo di variazione (o *range*)

$$range = |\max - \min|$$



➢ L'intervallo di variazione interquartile

$$range\ interquartile = |Q_3 - Q_1|$$



Questa misura indica la variabilità del 50 % "centrale" dei dati

SEGUE

Fonti: http://www.unife.it/economia/economia/insegnamenti/statistica/materiale-corso-a-a-2017-2018-gruppo-c/capitolo-03

Misure di dispersione: la varianza e la deviazione standard

➢ La varianza

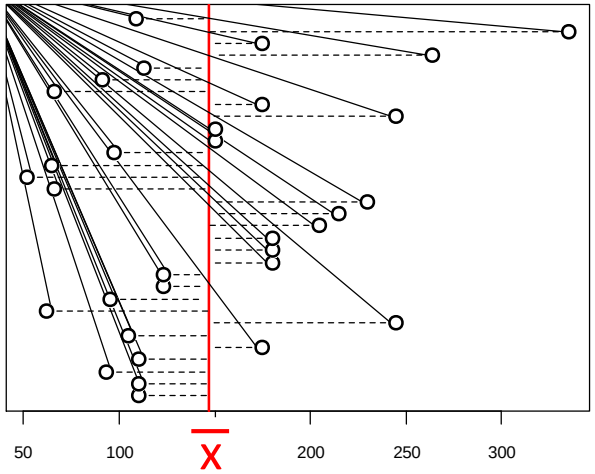
$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Queste misure sintetizzano la dispersione attorno alla media dei valori osservati

($\bar{x} = \mu$ = media;
n = numero di osservazioni)

➢ La deviazione standard

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



La deviazione standard (al contrario della varianza) mantiene la stessa unità di misura dei dati

SEGUE

Fonti: http://www.unife.it/economia/economia/insegnamenti/statistica/materiale-corso-a-a-2017-2018-gruppo-c/capitolo-03

Misure di dispersione: il coefficiente di variazione

$$CV = \frac{\sigma}{\mu}$$

σ = deviazione standard

μ = media

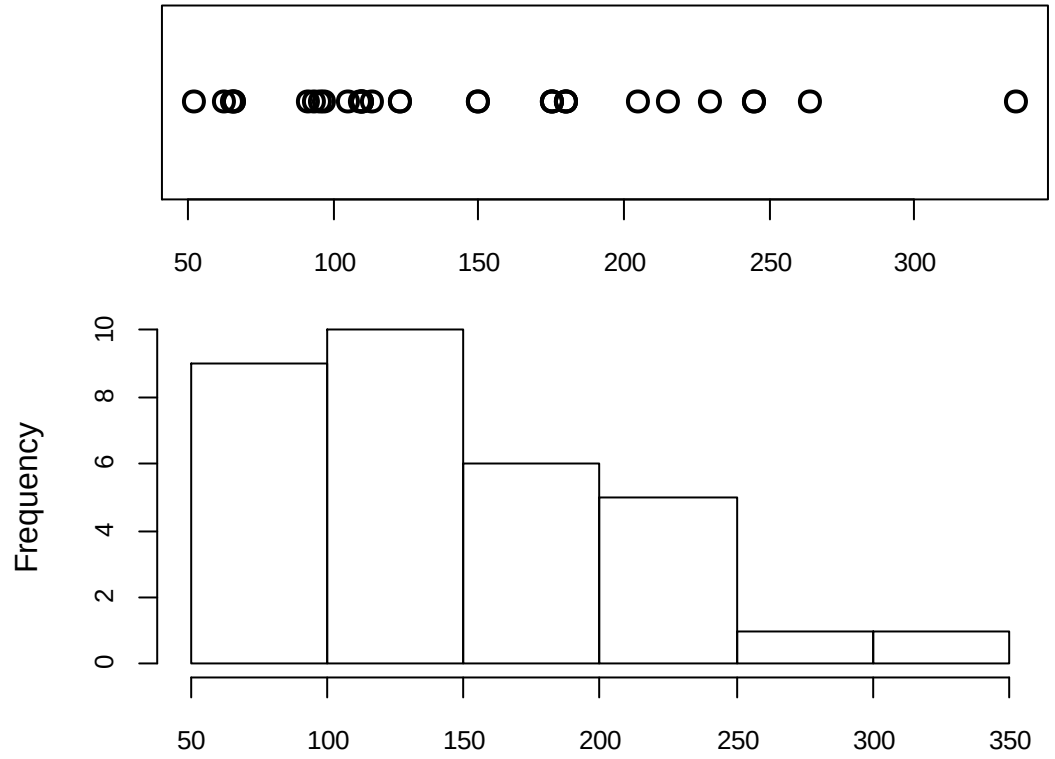
Essendo un numero puro (senza unità di misura) è utile per comparare distribuzioni di due variabili di unità di misura diversa. Può esser espresso anche come percentuale:

$$CV = \frac{\sigma}{\mu} \cdot 100$$

Fonti: http://www.unife.it/economia/economia/insegnamenti/statistica/materiale-corso-a-a-2017-2018-gruppo-c/capitolo-03

Istogrammi di frequenza

L'intervallo di valori assunti dalla variabile (min-max) viene diviso in intervalli più piccoli, detti anche **classi** o **bins**. Successivamente si calcola quanti valori sono presenti in ogni intervallo (**frequenza**) e si costruisce il grafico

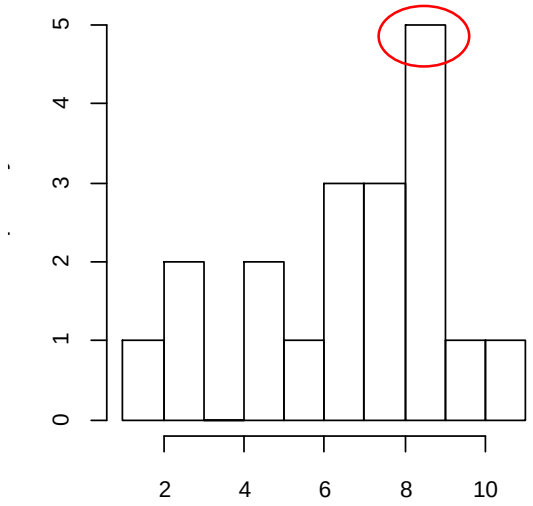


L'istogramma di frequenza è un modo di visualizzare la distribuzione di valori di una variabile (il numero di *bins* da utilizzare per la rappresentazione dipende dallo scopo)

SEGUE

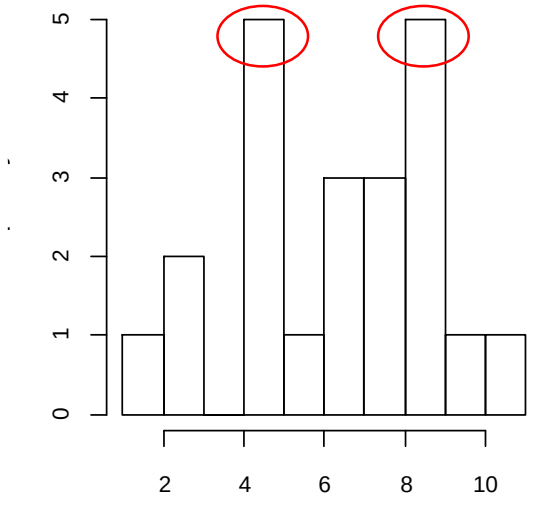
Istogrammi di frequenza assoluta

X	Freq
1	1
3	2
5	2
6	1
7	3
8	3
9	5
10	1
11	1



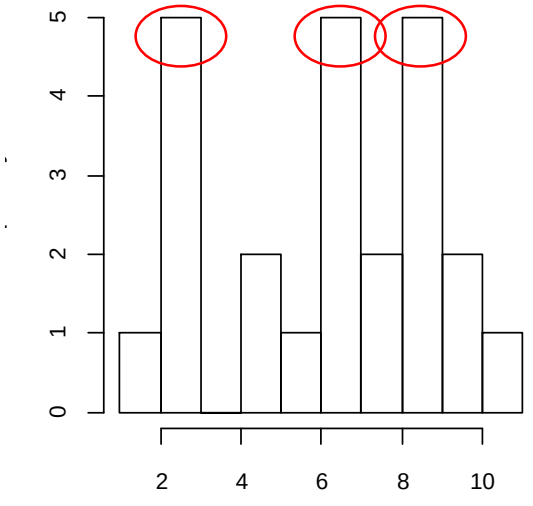
unimodale

X	Freq
1	1
3	2
5	5
6	1
7	3
8	3
9	5
10	1
11	1



bimodale

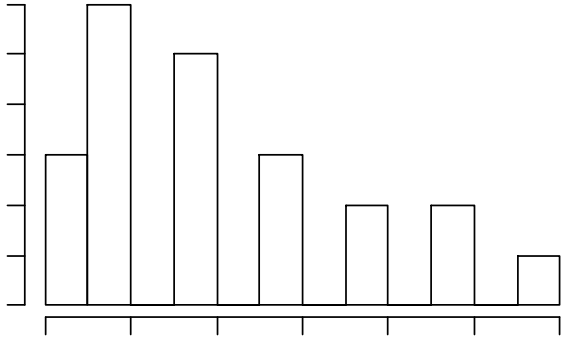
X	Freq
1	1
3	5
5	2
6	1
7	5
8	2
9	5
10	2
11	1



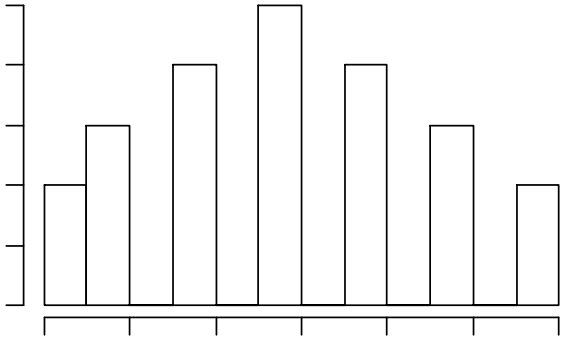
multimodale

Misure di forma: Asimmetria (o *skewness*)

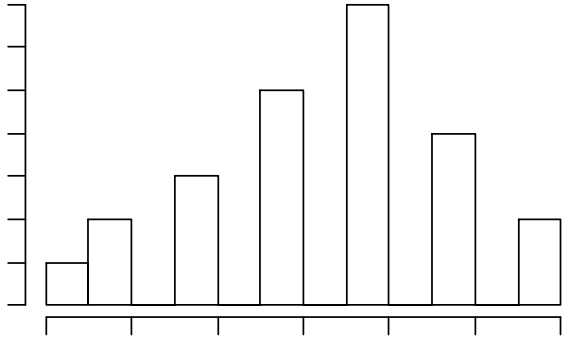
Le distribuzioni di dati unimodali possono essere simmetriche o asimmetriche (positive o negative).



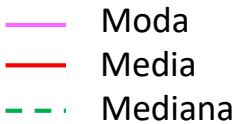
asimmetrica positiva

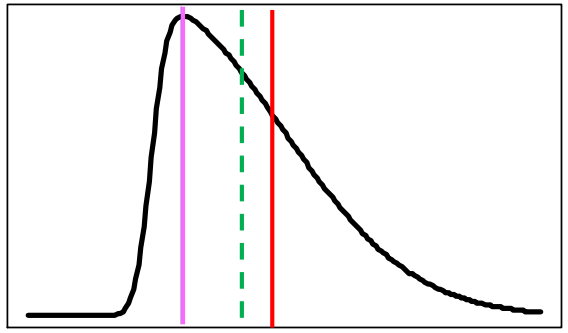


simmetrica

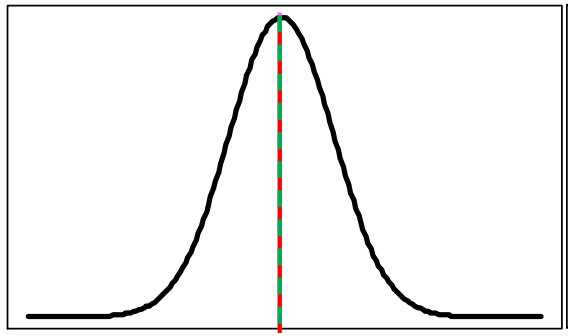


asimmetrica negativa

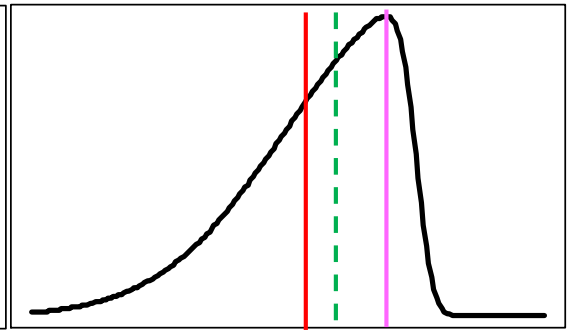




Media > Mediana

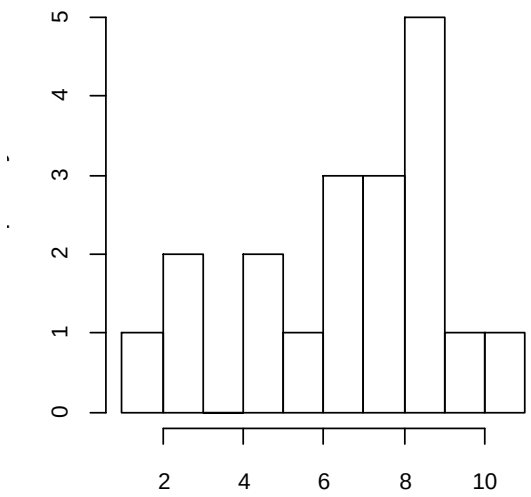


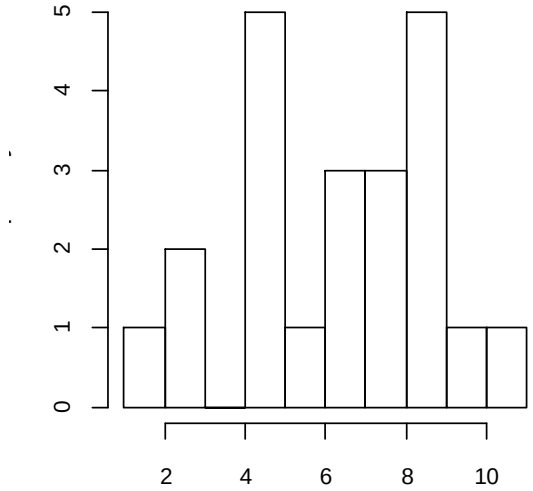
Media = Mediana

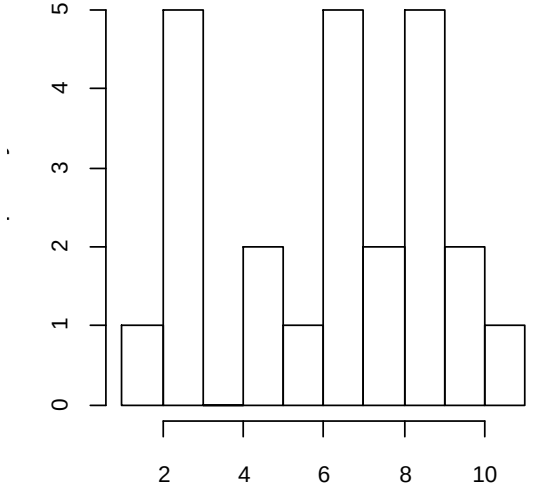


Media < Mediana

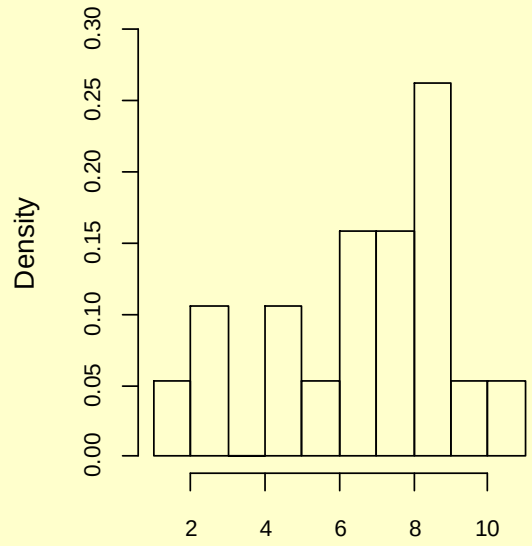
Densità di probabilità

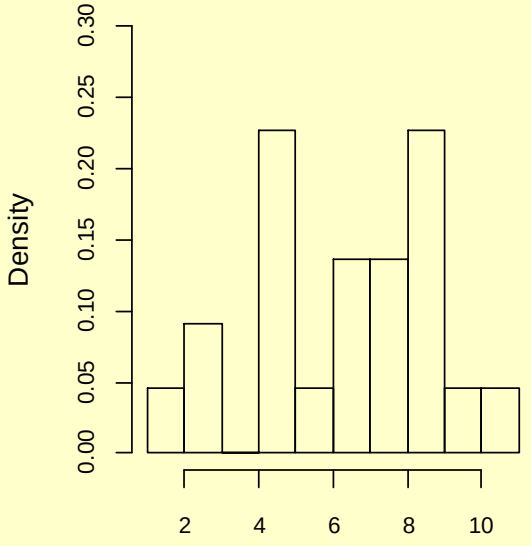


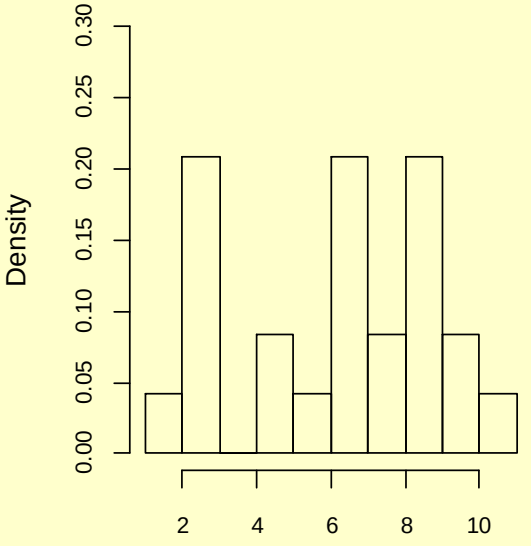




La densità di probabilità è rappresentata in modo che l'area totale dell'istogramma sia = 1







Permette di confrontare diverse serie di osservazioni della stessa variabile

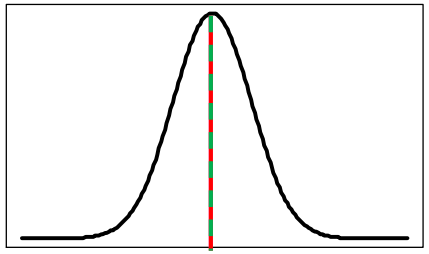
Densità di probabilità: approssimazione tramite curva gaussiana

Funzione di densità di una variabile distribuita in modo gaussiano o normale:

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

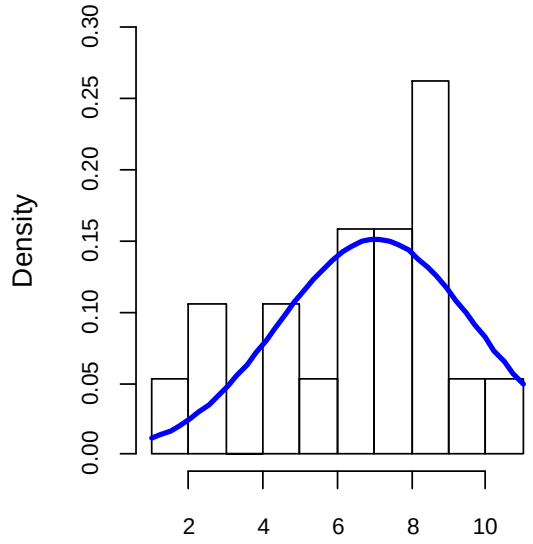
σ = deviazione standard

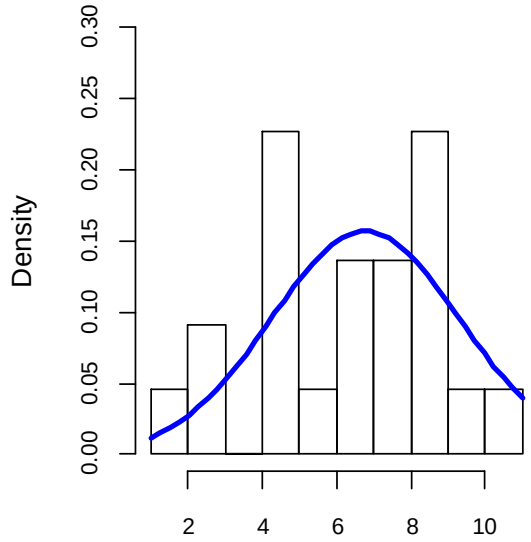
μ = media

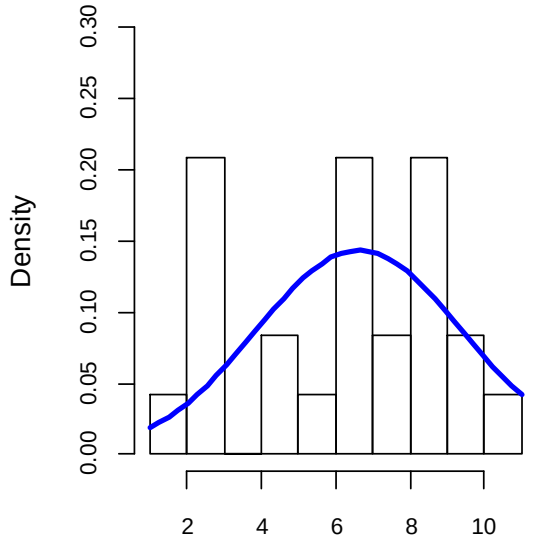


- ✓ La curva è simmetrica rispetto al valore centrale.
- ✓ I parametri che determinano la curva sono la media (μ) e la deviazione standard (σ)

E' importante perché alcuni test statistici (così detti parametrici) richiedono che i dati abbiano una distribuzione normale!



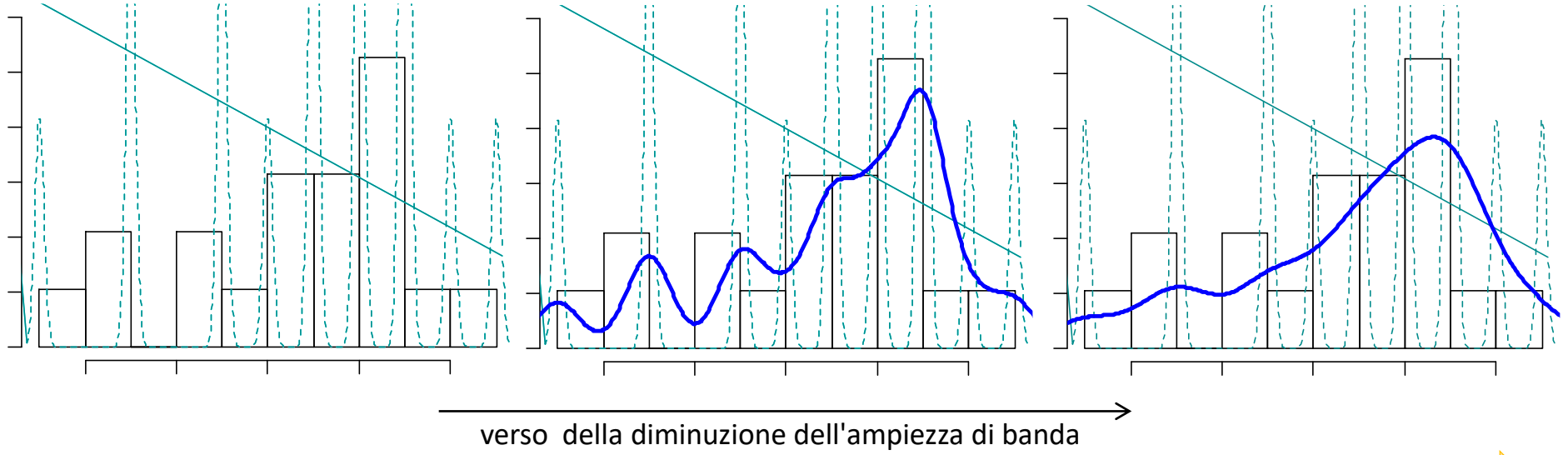




Densità di probabilità: approssimazione tramite curva di densità di kernel

La funzione di stima di densità di probabilità di kernel è un metodo non-parametrico (quindi non fondato principalmente sui valori di μ e σ^2 e distribuzione normale della variabile) che prevede la "modellazione" dell'istogramma di densità di probabilità utilizzando una curva che smussi (***smoothing***) il profilo dell'istogramma.

In pratica al posto delle barre vengono poste delle *piccole "gobbe"* che poi vengono unite assieme in una unica curva. La **forma delle gobbe** dipende dal tipo di funzione kernel utilizzata (ce ne sono diverse) e da un parametro di smussamento detto ampiezza di banda (***bandwidth***).

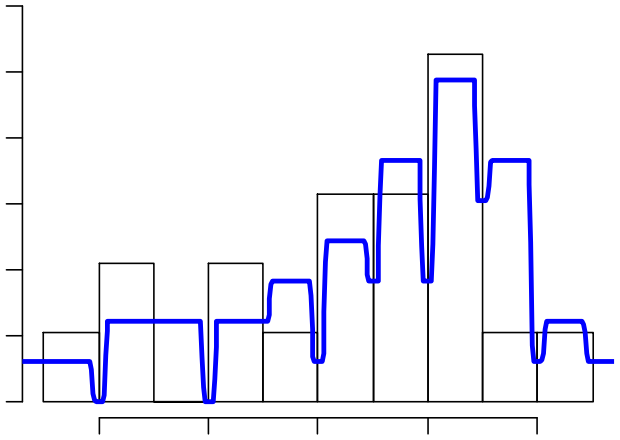


SEGUE

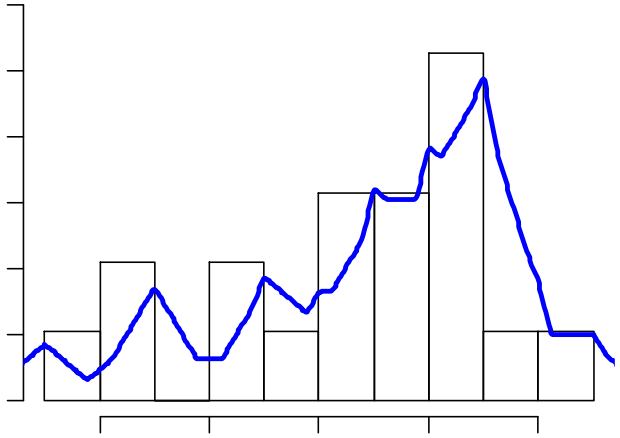
Fonti: https://impararfacendo.blogspot.com/search?q=kernel; https://mathisonian.github.io/kde/

Densità di probabilità: approssimazione tramite curva di densità di kernel (2)

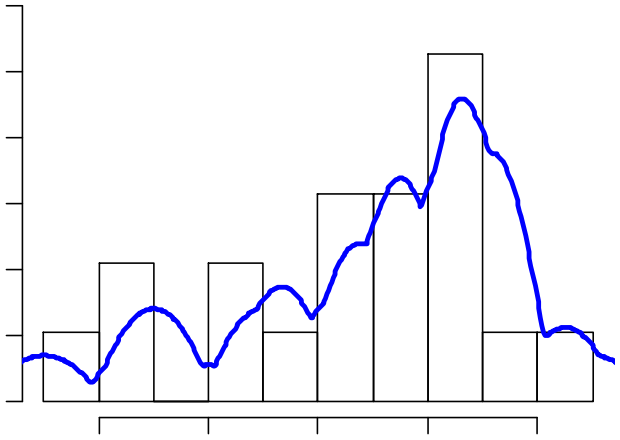
Stessa ampiezza di banda e diversa funzione di approssimazione della "gobba":



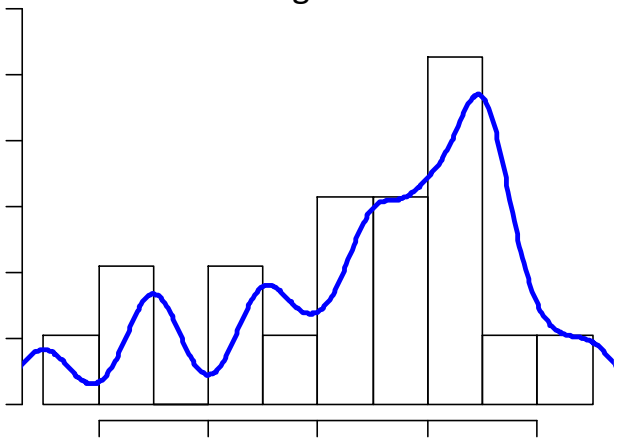
"rettangolare"



"triangolare"



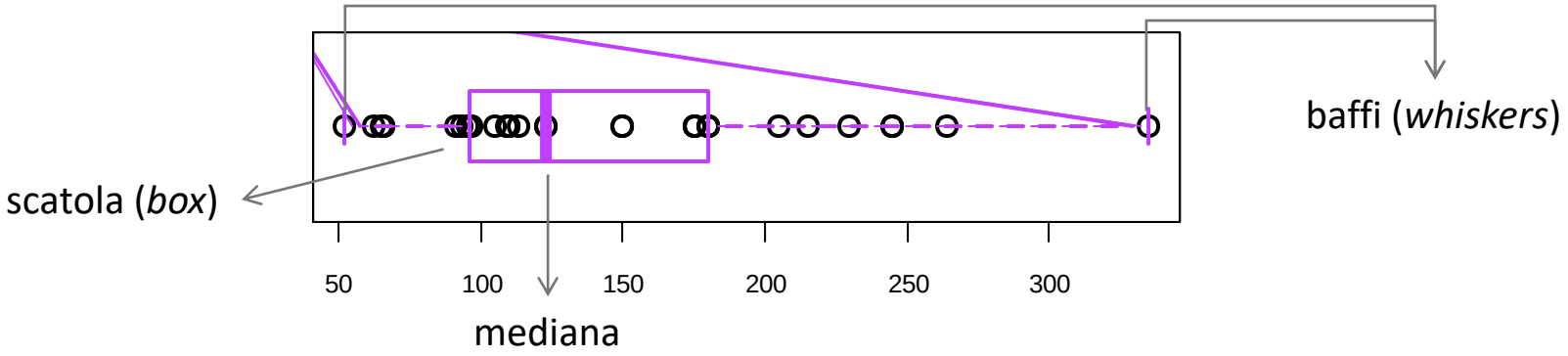
"Epanechnikov"



"Gaussiana"

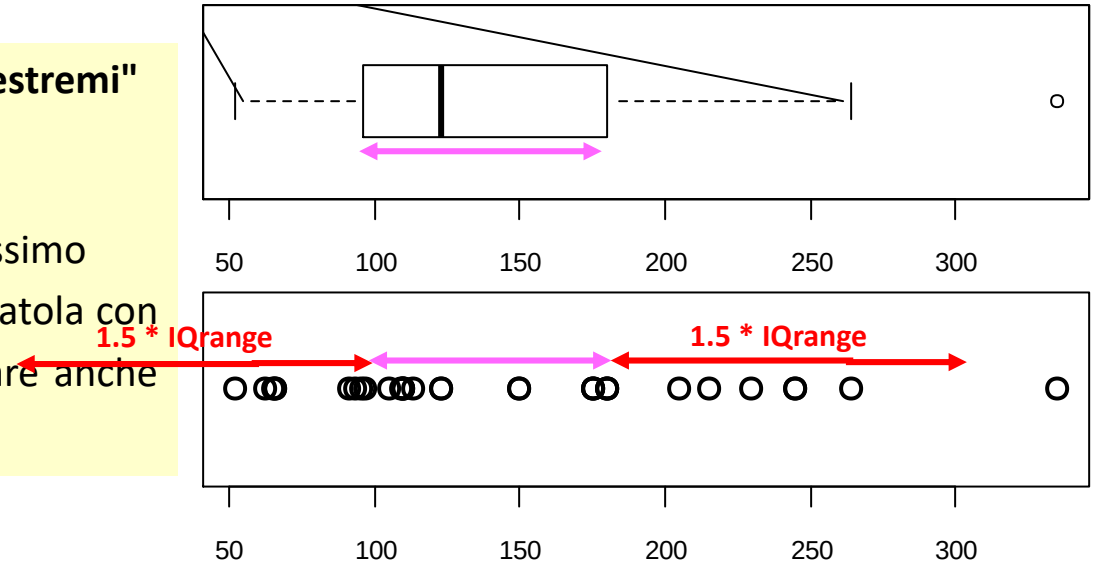
Grafici a scatola (o *boxplot*)

Un grafico a scatola viene costruito utilizzando i valori min, max e quartili



"Variante" per l'individuazione di "dati estremi" (*outlier*):

i baffi si estendono ai dati che sono al massimo = (N *range interquartile) lontani dalla scatola con N tipicamente 1.5, ma si possono utilizzare anche altri valori.



Fonti: http://www.unife.it/economia/economia/insegnamenti/statistica/materiale-corso-a-a-2017-2018-gruppo-c/capitolo-03

Riassumendo...