Statistica

per Tossicologia dell'Ambiente

AA2007/08

F.-L. Navarria

Obiettivi, crediti

- **Il corso si propone di fornire i principali elementi di statistica descrittiva, calcolo delle probabilità, stima campionaria, teoria delle decisioni ed analisi di tendenza, essenziali per le applicazioni pratiche in campo ambientale ed analitico-strumentale.**
- **2 CFU, 14h lezione, 5 lezioni (14/11, 20/11, 22/11, 27/11, 29/11)**

Programma

- **Introduzione. Analisi dei dati, campionamento, classificazioni, distribuzioni di frequenza, istogrammi. Misure di tendenza, media, mediana. Misure di dispersione, scarto quadratico medio,quantili.**
- **Probabilità: assiomatica, oggettiva, empirica. Regole di calcolo delle probabilita`. Probabilità soggettiva, teorema di Bayes. Inferenza statistica, uso della probabilità.Variabile stocastica discreta e continua. Valore di aspettazione. Analisi combinatoria, disposizioni, permutazioni, combinazioni.**
- **Distribuzione binomiale. Distribuzione normale o di Gauss. Variabile casuale normalizzata. Approssimazione gaussiana della binomiale. Integrali della gaussiana. Statistica di Poisson. Conteggi, errore statistico. Approssimazione gaussiana.**
- **Teoria statistica della stima. Tests di ipotesi statistiche. Test sulla media (test t di Student), test sulla varianza (test F), test sulla frequenza (test del chi-quadrato). Livelli di confidenza. Errori di I e II specie. Fit di dati con leggi note. Metodo dei minimi quadrati con dipendenza lineare o linearizzabile. Regressione lineare, coefficiente di correlazione, stima dei parametri.**

Citazioni

There are three kinds of lies: lies, bloody (oppure, in altre versioni, ‘damned’) lies, and statistics.

[frase attribuita a Benjamin Disraeli da Mark Twain]

La statistica addolcisce la vita.

[Universum Science Center, Bremen]

Bibliografia

- Copie dei lucidi delle lezioni
- F.R. Cavallo e F.-L. Navarria, Appunti di probabilità e statistica …, CLUEB, 2000
- (H.T. Hayslett, Statistics made simple, Heinemann, 1981)
- Qualche esercizio di probabilità e statistica si trova alla pagina Web http://www.bo.infn.it/ctf/eser
- …

Modalità didattiche/esame

- **Metodi didattici:**
- **Lezioni con qualche esercitazione/esercizio utile ai fini della valutazione dell'apprendimento**

- **Modalità di verifica dell'apprendimento:**
- **Esame scritto, integrabile se sufficiente con una prova orale (facoltativa)**

- **Primi appelli utili: 25/01/08 e 25/02/08**

Dove mi trovo

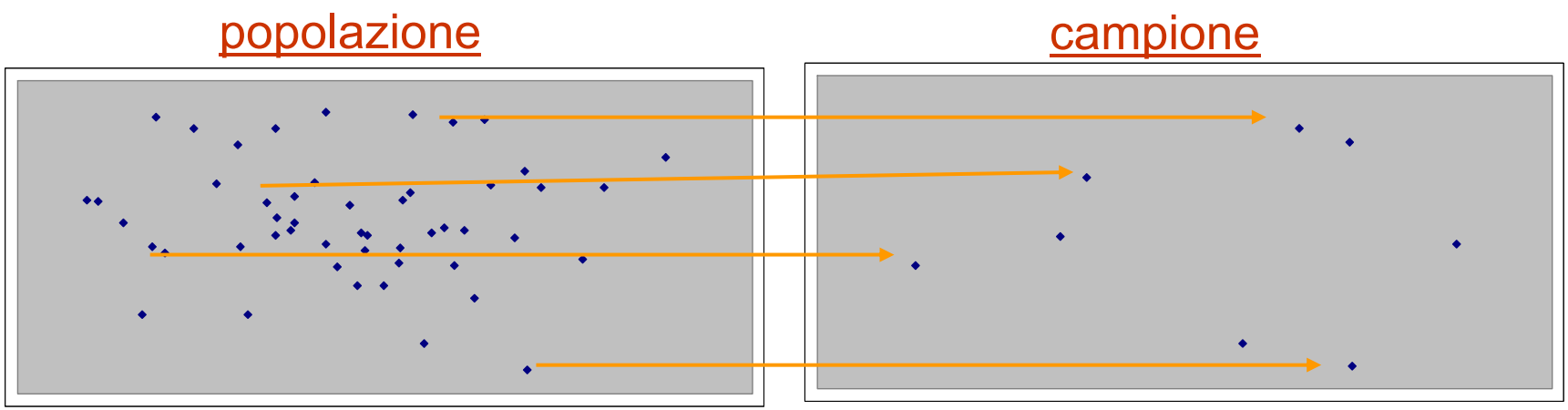
- Dip. di Fisica, V.le Berti-Pichat 6/2, p. II, 40127 Bologna
- Ricevimento, vedi http://www.bo.infn.it/ctf/eser - Lun 13-14, Mar 13-14, Mer 12-13 (possono cambiare in funzione degli orari di lezione)
- e-mail: navarria@bo.infn.it (funziona dovunque, o quasi, mi trovi)

Introduzione

- Statistica: due significati
  - numeri scelti per fornire indicazioni su un campione (età media degli studenti, valutazione massima ottenuta in un test d'ìngresso …); la mediana, la moda etc. sono una statistica
  - tecniche e metodi usati per raccogliere, analizzare, interpretare e presentare I dati
- La statistica è usata per prendere decisioni sulla base di dati/informazioni incompleti (ossia sempre).
- 1) Raccolta dati. Misura o campionamento di una popolazione. Esperimento ripetibile o meno.
- 2) Statistica descrittiva. Descrizione pittorica/grafica dei dati, istogrammi. Classificazione numerica, tabelle, stime di tendenza, dispersione, asimmetria del campione.
- 3) Statistica inferenziale – la scienza di prendere decisioni (test di ipotesi, stime probabilistiche in conseguenza di misura o campionamento).

Popolazione e campione

- Popolazione: gli abitanti di Imola
- Campioni:
  - le donne di I. di età superiore a 18 anni — biased
  - gli abitanti di I. biondi — "
  - gli abitanti di I. con gli occhi verdi — "
  - un campione casuale di abitanti di Imola — unbiased, random



Raccolta dati/campionamento

Esito di un TAS. Insieme di dati: studenti e punteggio

classificazione

ordinati 20 a 20 in ordine crescente

elementi

osservazioni

Campione casuale di 100 TAS

591	550	568	566	624
557	629	579	585	569
603	555	618	611	618
592	605	491	549	553
612	608	557	603	618
507	604	695	606	576
573	633	502	554	597
575	590	592	511	607
589	580	561	591	579
579	556	521	514	565
504	524	554	616	512
515	621	605	593	594
534	601	533	589	549
630	542	646	638	639
558	636	569	550	535
558	624	511	528	581
571	541	553	580	573
553	569	560	600	634
580	620	556	639	563
570	577	599	587	541

Dati TAS - ordinati su 5 colonne

504	52[illegible]	491	511	512
507	541	502	514	535
515	542	511	528	541
534	550	521	549	549
553	555	533	550	553
557	556	553	554	563
558	569	554	566	565
558	577	556	580	569
570	580	557	585	573
571	590	560	587	576
573	601	561	589	579
575	604	568	591	581
579	605	569	593	594
580	608	579	600	597
589	620	592	603	607
591	621	599	606	618
592	624	605	611	618
603	629	618	616	624
612	633	646	638	634
630	63[illegible]	695	639	639

estremi

Ad esempio

Campione di una popolazione – 100 studenti - Test Attitudine Scolastica

Raccolta dati/campionamento-2

- I dati sperimentali grezzi, ad es. i 100 valori del TAS, contengono tutta l'informazione possibile sul campione e di conseguenza sulla popolazione. Sono però spesso scomodi da visualizzare e da utilizzare.
- → Si passa ad una classificazione (istogramma, tabella etc.) che renda i dati più intellegibili e immediati ed alla riduzione ad un piccolo numero di parametri che caratterizzano il campione.
- In questo passaggio si perde l'informazione sul singolo dato. → Bisogna usare indicatori rappresentativi.
- Occorre anche evitare tutti quei raggruppamenti che 'falsano' i dati (ad es. istogrammi con canali di ampiezza diversa etc.)

Tabelle & istogrammi (raggruppamento dei dati)

Distribuzioni di frequenza

Tabella 1

Classe	Limiti	Frequenza
1	475-499	1
2	500-524	9
3	525-549	10
4	550-574	27
5	575-599	23
6	600-624	20
7	625-649	9
8	650-674	0
9	675-699	1

Fig. 1

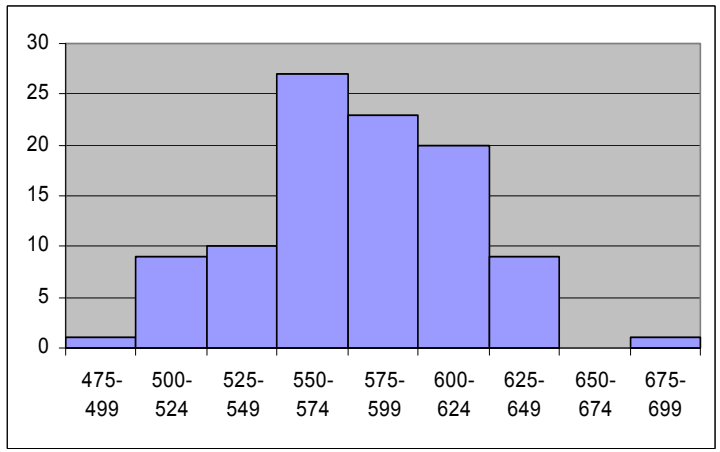
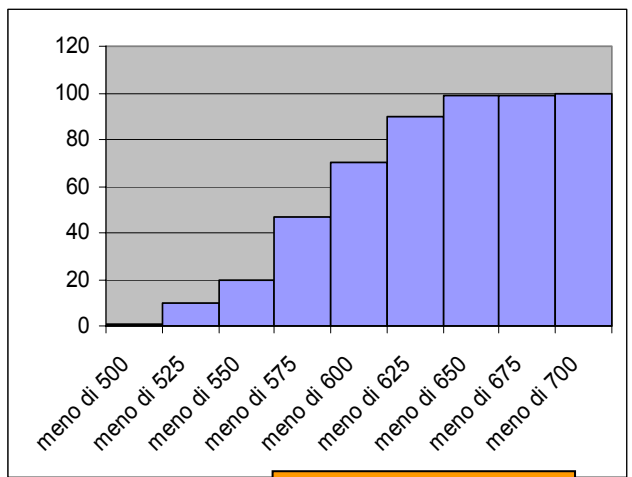


Tabella 2

Classe	Limiti	Frequenza cumulata	Frequenza cumulata precentuale
1	meno di 500	1	1%
2	meno di 525	10	10%
3	meno di 550	20	20%
4	meno di 575	47	47%
5	meno di 600	70	70%
6	meno di 625	90	90%
7	meno di 650	99	99%
8	meno di 675	99	99%
9	meno di 700	100	100%

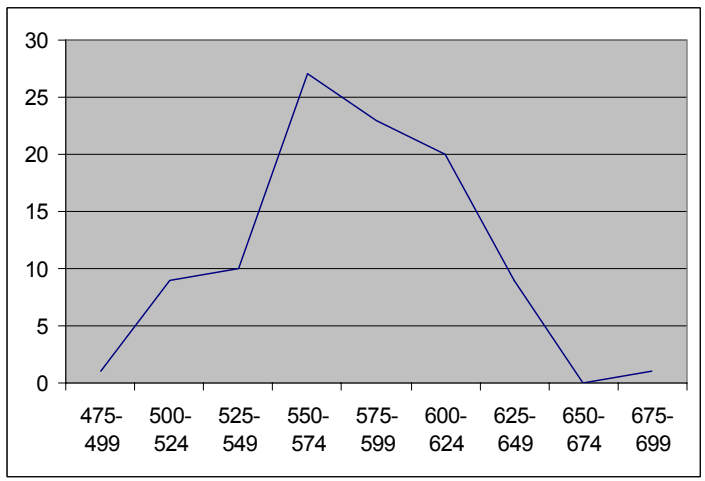
Fig. 2



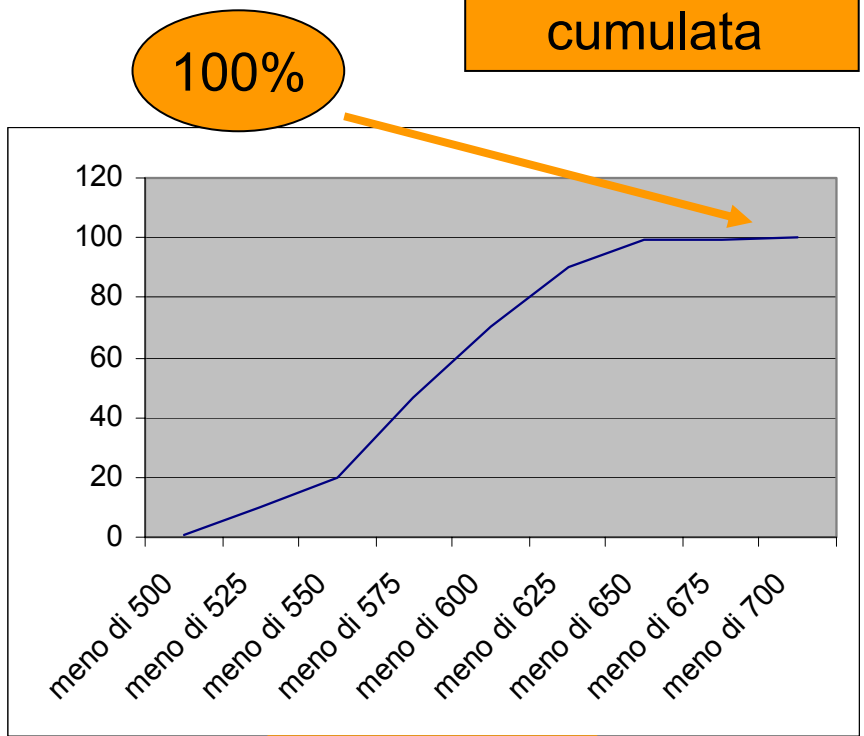
Distribuzione cumulativa

Altre rappresentazioni grafiche dei dati raggruppati

Poligono della frequenza



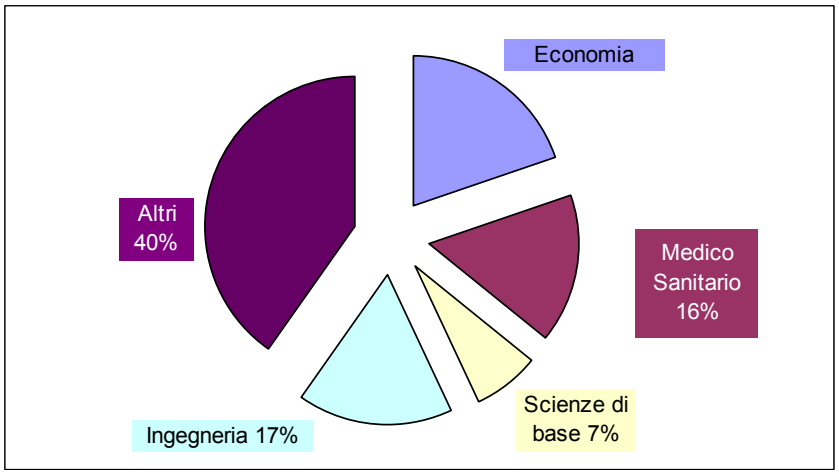
Poligono della frequenza cumulata

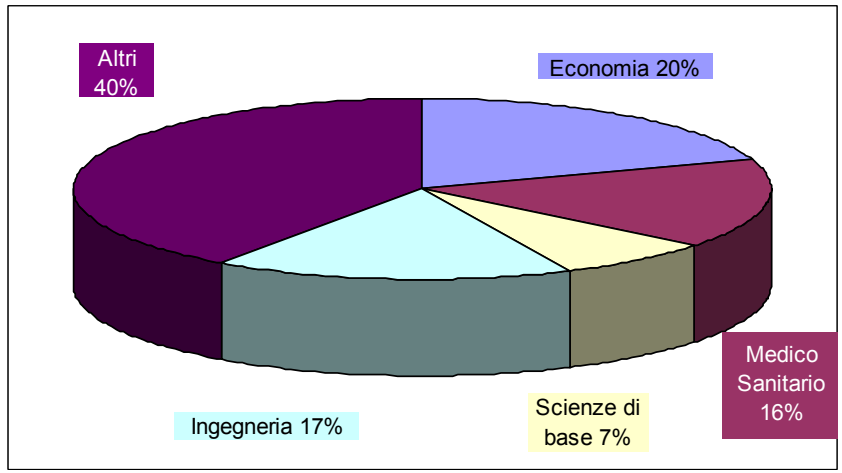


tipicamente a forma di S

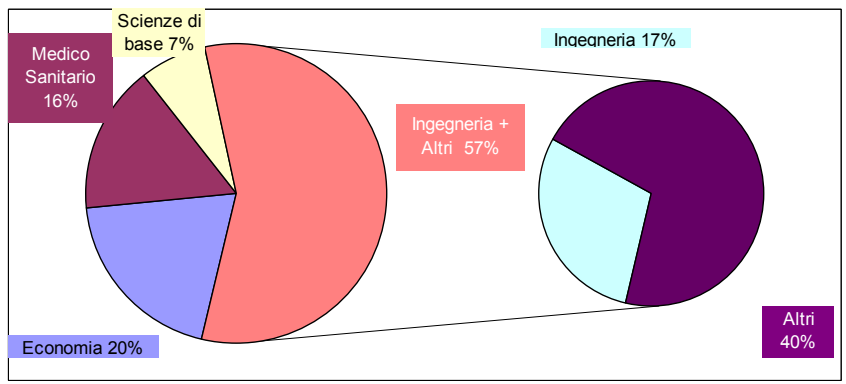
(Tabelle e grafici sono stati prodotti con Excel)

Altre rappresentazioni (torte)





Indirizzo di studio	Frequenza	Freq. relat.	Angolo in gradi
Economia	2234	0.197	71
Medico Sanitario	1829	0.161	58
Scienze di base	807	0.071	26
Ingegneria	1912	0.169	61
Altri	4563	0.402	145
Totale	11345	1.000	360



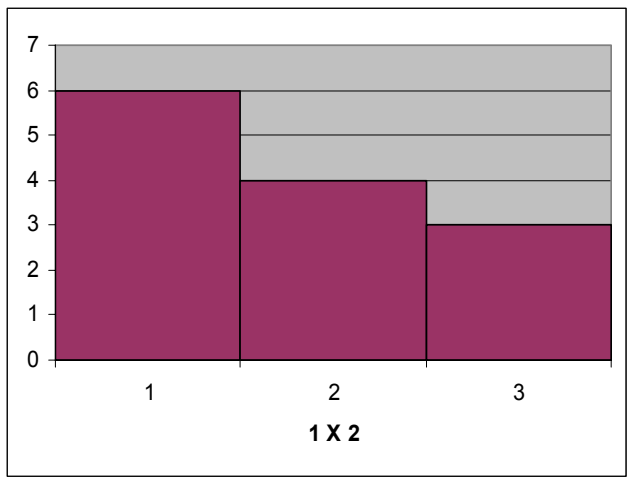
Note

- Campione di una popolazione (finita, ad es. nel caso degli studenti – infinita, ad es. tutte le possibili uscite T/C nei successivi lanci di una moneta).
- Campionamento casuale (random): ogni elemento ha la stessa possibilità di essere scelto e la scelta di un dato el. non influenza la scelta di un successivo el. [pop. infin. o pop. finita con rimpiazzamento; se non si rimpiazza, tutti i campioni della stessa dimensione devono avere uguale possibilità]
- Variabile aleatoria (discreta, ad es. il risultato del TAS – continua, ad es. l'altezza h o la massa m di una persona [il numero di cifre significative dipende dalla precisione della misura])
- Distribuzioni di frequenza – riga 1: tabella 1, figura 1
- Distribuzioni cumulative di frequenza – riga 2: tabella 2, figura 2

Statistica non parametrica

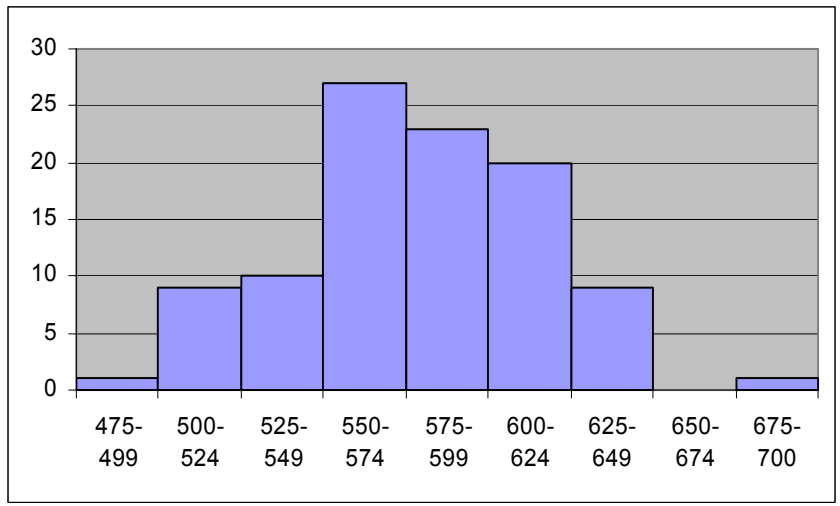
- Non tutti i risultati di un campionamento sono numerici e quindi suscettibili di essere ordinati numericamente (la probabilità, vedi più avanti, di fare ad es. X è comunque definita, se si ammette di giocare sempre le stesse partite). Altre osservazioni ad es. il colore dei capelli, degli occhi di una persona, cavalli in una corsa &tc.
- Istogramma del lunedì o dello scommettitore!

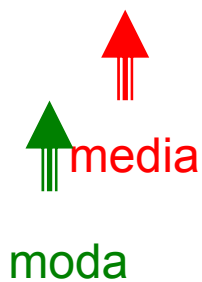
Classe	Segno	Frequenza
1	1	6
2	X	4
3	2	3



Indicatori di tendenza del campione

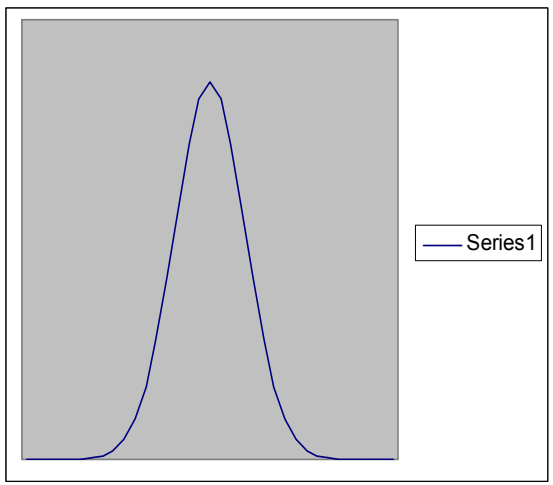
- **Media aritmetica**:
  $x_m = \underline{x} = \Sigma_{i=1,n}\, x_i/n$ (per una popolazione $\mu = \Sigma_{i=1,N}\, x_i/N$)
- **Mediana**: il valore che corrispondere a dividere a metà i dati (nell'es. 50 prima e 50 dopo la mediana)
- **Moda**: Il dato/la classe/il canale dell'istogramma con la massima frequenza
- **Centro dell'intervallo**: (dato più piccolo + dato più grande)/2





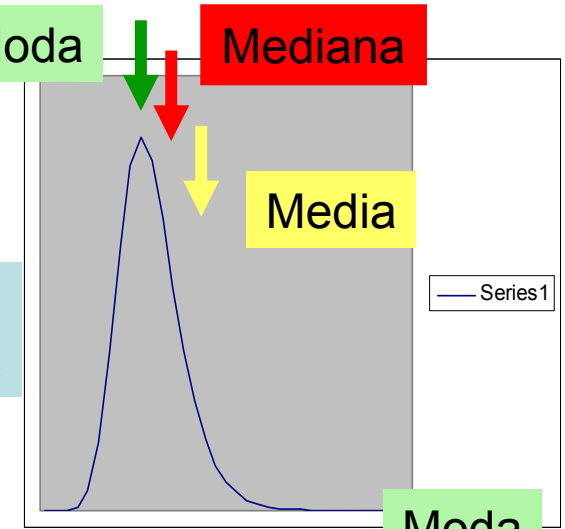
Media, Mediana, Moda

Moda=mediana=media

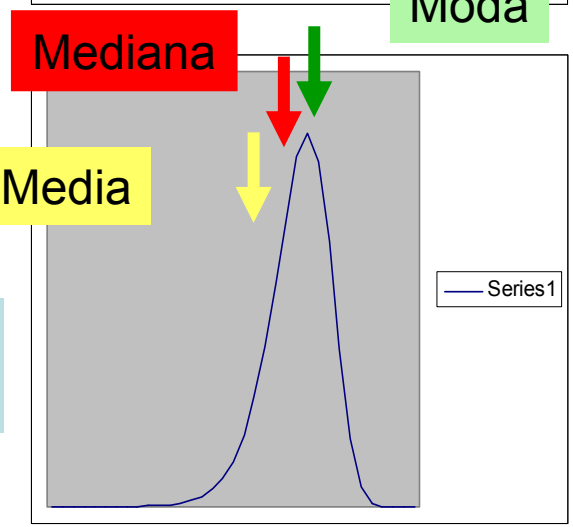


Distribuzione simmetrica (a campana)

Unimodale obliqua sinistra



Unimodale obliqua destra



Media aritmetica

- $x_m = \underline{x} = \Sigma_{i=1,n}\, x_i/n = (\Sigma_{i=1,n}\, x_i)/n$ n dati (con lo stesso peso)
- Media ponderata

  $x_m = \underline{x} = \Sigma_{i=1,n}\, w_i x_i/\Sigma_{i=1,n}\, w_i$ pesata con w_i diversi
- Ad es. media di un istogramma con frequenze f_i

  $x_m = \underline{x} = \Sigma_{i=1,n}\, f_i x_i/\Sigma_{i=1,n}\, f_i$ pesata con f_i

  (le probabilità sono proporzionali a f_j nel j-esimo canale)

  – dati raggruppati: n è in questo caso il numero di classi
- Se facciamo un cambiamento di origine si ha

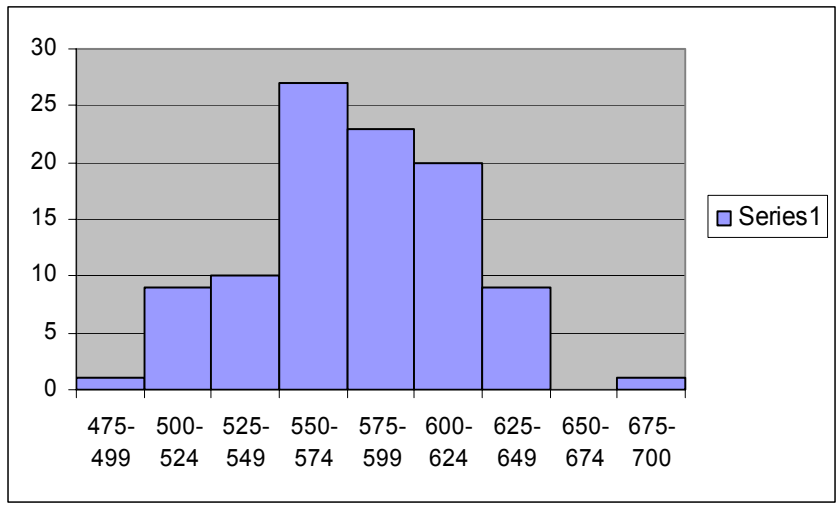
  $x_i \rightarrow y_i = x_i + A \Rightarrow \underline{x} \rightarrow \underline{y} = \underline{x} + A$

  $x_i \rightarrow y_i = x_i - \underline{x} \Rightarrow \underline{x} \rightarrow \underline{y} = \underline{x} - \underline{x} = 0$

  (def.: scarti dalla media) (→ la media degli scarti è nulla)

Indici di dispersione del campione

- Campo di variazione:
  massimo valore –
  minimo valore =
  $= x_{max} - x_{min}$
- scarto medio assoluto (rispetto alla media)
  $s.m.a. = \Sigma_{i=1,n} |x_i - \underline{x}|/n$
- scarto quadratico medio dalla media
- semi-differenza interquartile



± sqm

Campo di variazione

Scarto quadratico medio (deviazione standard)

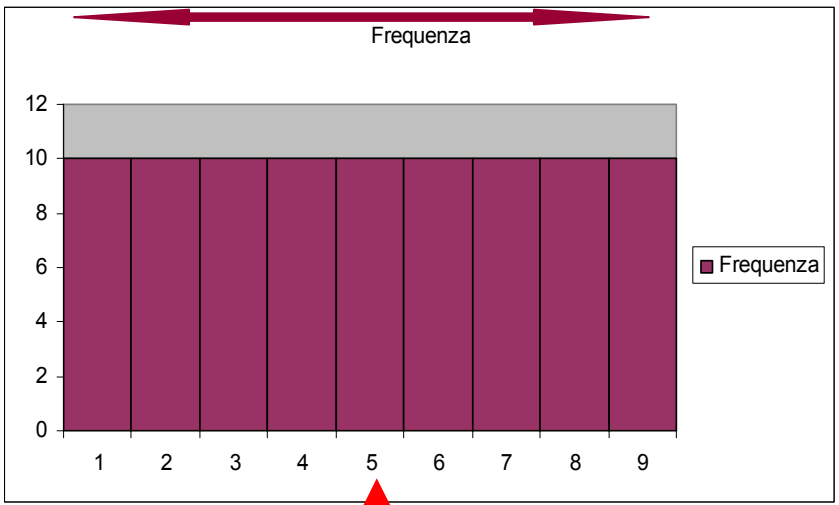
- $s = \sqrt{(\Sigma_{i=1,n} (x_i - \underline{x})^2/(n-1))}$ [campione]
  [N.B. per la popolazione -> σ, n]
- s^2 - varianza
- lo s.q.m. risulta minimo rispetto alla media aritmetica: supponiamo infatti di calcolarlo rispetto ad $a = \underline{x} + b$

$$\Sigma_{i=1,n} (x_i - a)^2 = \Sigma_{i=1,n} (x_i - \underline{x} - b)^2 =$$
$$= \Sigma_{i=1,n} (x_i - \underline{x})^2 - 2b\Sigma_{i=1,n} (x_i - \underline{x}) + nb^2$$
$$= \text{min se } b = 0 \qquad \underbrace{\text{----------}}_{= 0}$$

Stessa media/mediana, stesso range, diversa dispersione

unimodale

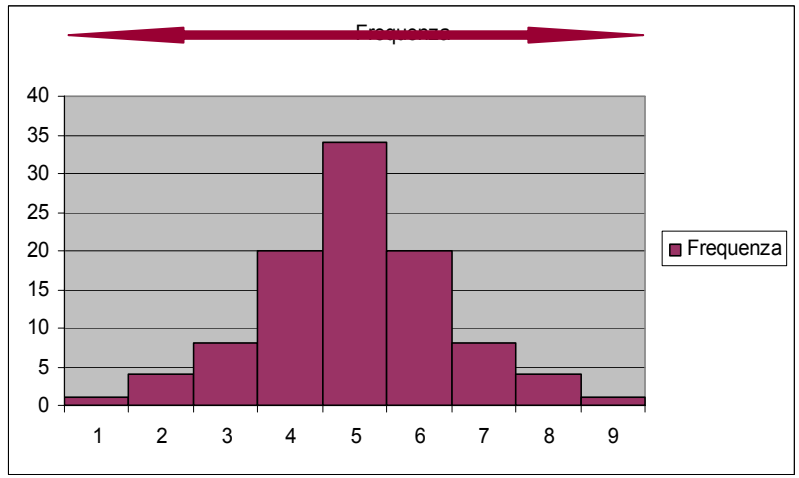
Range = x_{max}-x_{min}

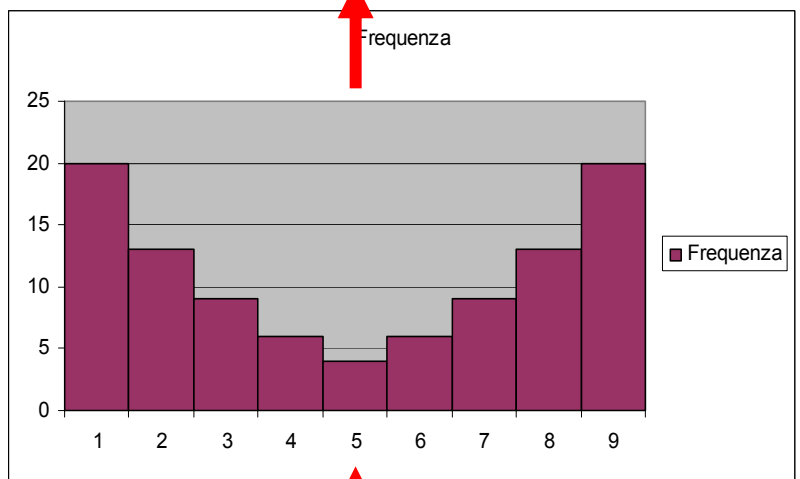


Media = (x_{min}+x_{max})/2

non ha moda

bimodale



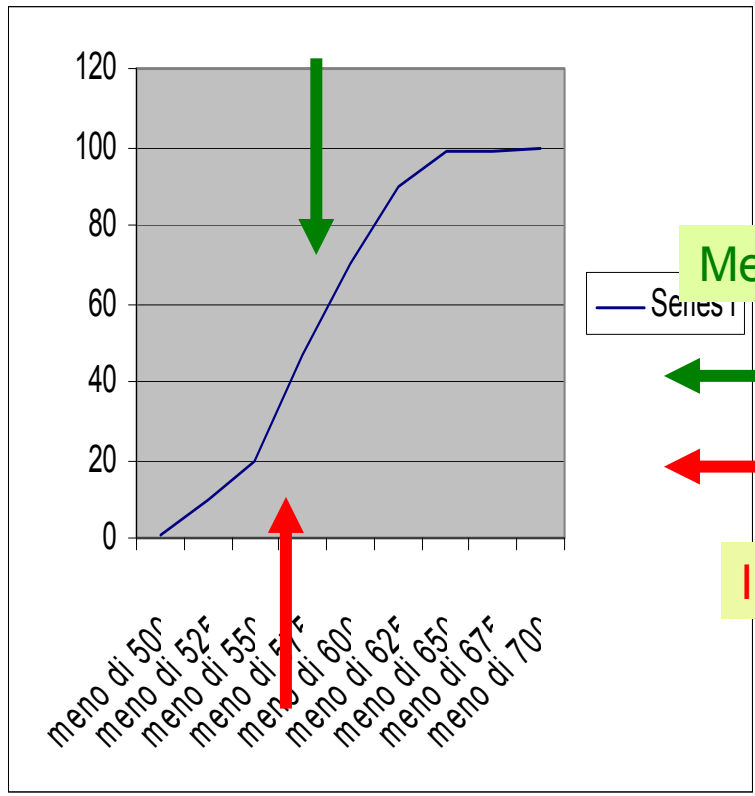


Dispersione dei dati raggruppati, dispersione relativa, variabile standardizzata etc.

- Campione, dati raggruppati, N classi con frequenza f_i
  $\mathbf{s^2 = (\Sigma f_i x_i^2 - n\underline{x}^2)/(n - 1)}$ n – numero di dati
- dispersione-relativa = dispersione-assoluta/media (numero puro, adimensionale) $s/\underline{x}$
- variabile standardizzata
  $z = (x - \underline{x})/s$ (adimensionale)
- { Momenti di ordine r dalla media
  $m_r = \Sigma_{j=1,n} (x_j - \underline{x})^r /n$ }
- { Asimmetria
  $a_3 = m_3/s^3$ che, essendo dispari, può essere +va/-va}
- { &tc. (la distribuzione di Gauss o normale ha $a_3 = 0$, $a_4 = 3$ …) }

Quantili

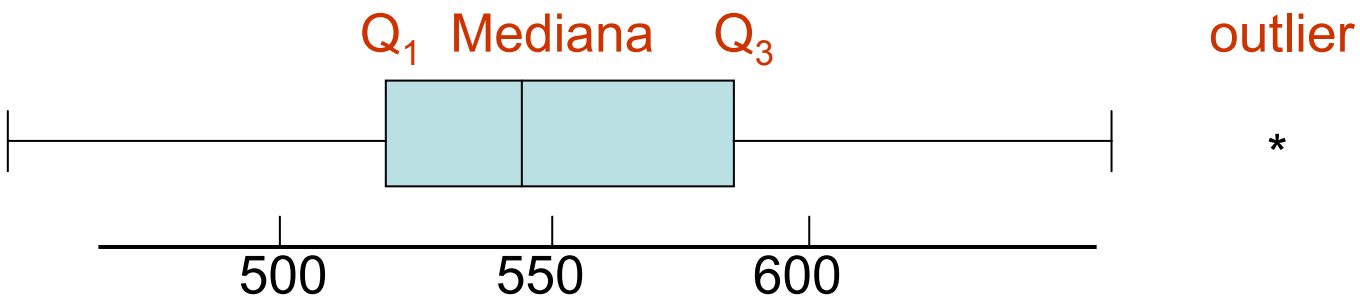
- Con dati ordinati in modo progressivo, il valore centrale è la mediana, che divide la serie di dati in due parti uguali.
- Analogamente si possono dividere I dati in 4 parti => quartili (I, II, III, IV quartile)
- Idem in 10 parti => decili
- Idem in 100 parti => percentili
- Semi-differenza interquartile: $(Q_3-Q_1)/2$



I quartile

Box-and-Wiskers plot

- Rappresentazione dei dati B&W (scatola e baffi): si costruisce una scatola con indicata la mediana e con estremi Q_1 e Q_3 sopra un intervallo graduato che contiene tutte le osservazioni sul campione, i baffi si ottengono con due segmenti lunghi 1.5xIQR=1.5x(Q_3-Q_1) a partire da Q_1 e Q_3 – i dati fuori dall'intervallo sono outlier



Note

- Media e deviazione standard per la popolazione sono indicate spesso con μ,σ rispettivamente, mentre $\underline{x}$,s sono le quantità campionarie - in Excel: StDevp/Varp e StDev/Var, rispettivamente
- Nel caso di variabili continue sarà necessario sostituire le sommatorie (Σ) con integrali (∫) ad es. la media di x su una distribuzione di frequenze continua f(x) sarà

  $\underline{x} = \int xf(x)dx / \int f(x)dx$ esteso a tutto il campo di variazione della x