Giacomo Bulgarelli
Ufficio Servizi Statistici

SERVIZIO DAF: FONTI STATISTICHE

Mercoledì 3 ottobre 2012

4. La Statistica (III)

Indici di posizione

Nella ricerca scientifica e tecnologica, così come nelle scienze economiche, sociali e politiche, è importante misurare la reale efficacia di interventi e modifiche sul sistema oggetto di studio. Si cerca, cioè, nella mutevolezza ed instabilità dei risultati individuali, di valutare gli effetti complessivi indotti da una causa nota. Per questo sono necessarie misure sintetiche che posizionino la distribuzione di frequenza di un certo fenomeno e consentano il passaggio da una pluralità di informazioni (modalità e rispettive frequenza) ad un solo numero.

Obiettivo di una misura di posizione è quello di **sintetizzare in un singolo valore numerico l'intera distribuzione di frequenza** per effettuare **confronti** nel tempo, nello spazio o tra circostanze differenti. Talvolta, ciò è rilevante per verificare se le conseguenze di un'azione nota abbiano prodotto un risultato desiderato, in quale direzione e con quale intensità.

Indici di posizione

Media aritmetica

È il valore di posizione per eccellenza, spesso indicata senza altre aggettivazioni.
Disponendo di n osservazioni distinte $\{x_1, x_2, \ldots, x_n\}$ la media aritmetica è definita da:

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

Indici di posizione

Media aritmetica

Alcune proprietà

È sempre compresa tra il minimo e il massimo delle modalità della variabile *(criterio di internalità di Cauchy)*

La somma degli scarti dalla media aritmetica è nulla, per cui la media è il *baricentro* di una distribuzione di frequenza

Se la variabile X ha media μ, allora la variabile $\alpha+\beta X$ possiede media aritmetica pari a: $\alpha+\beta\mu$ *(linearità della media aritmetica)*. Pertanto aggiungendo o sottraendo una costante α alla variabile X, la rispettiva media sarà modificata dello stesso ammontare, mentre se la variabile X è moltiplicata per una costante β, la media risulterà moltiplicata dello stesso ammontare.

Esempio 5

Voto agli esami dello studente A

30, 18, 18, 24, 28, 30, 30, 30, 28, 27, 30, 24, 28, 27, 30, 30, 30, 26, 30, 28, 30

$$\mu = \frac{1}{21}(30+18+18+\ldots+30+28+30) = \frac{576}{21} = 27{,}43$$

Voto	Frequenze assolute
18	2
24	2
26	1
27	2
28	4
30	10
TOTALE	**21**

$$\mu = \sum_{i=1}^{k} x_i n_i \Big/ \sum_{i=1}^{k} n_i$$

Esempio 6

Voto agli esami dello studente A

Voto	Frequenze assolute
18	2
24	2
26	1
27	2
28	4
30	10
TOTALE	**21**

$$\mu = \sum_{i=1}^{k} x_i n_i \Big/ \sum_{i=1}^{k} n_i$$

$$\mu = \frac{(18*2+24*2+26*1+27*2+28*4+30*10)}{2+2+1+2+4+10} = \frac{576}{21} = 27{,}43$$

Esempio 7

E se lo studente A fa un altro esame e prende 29, come cambia la media aritmetica?

$$\mu_{(n+1)} = \frac{n\mu_{(n)} + x_{(n+1)}}{n+1} = \frac{n}{n+1}\mu_{(n)} + \frac{1}{n+1}x_{(n+1)} = \mu_{(n)} + \frac{x_{(n+1)} - \mu_{(n)}}{n+1}$$

$$\mu_{(22)} = \frac{21*27{,}43+29}{22} = \frac{21}{22}*27{,}43+\frac{1}{22}*29 = 27{,}43+\frac{29-27{,}43}{22} = 27{,}5$$

Indici di posizione

Mediana

È il valore della variabile che bipartisce la distribuzione ordinata delle modalità, cioè tale che metà delle osservazioni sia inferiore alla mediana e metà sia ad essa superiore. In altre parole, la mediana è la modalità dell'unità statistica che occupa il posto centrale nella distribuzione ordinata delle osservazioni

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

Indici di posizione

Mediana

Alcune proprietà

Il numero degli scarti (x_i-Me) positivi è uguale al numero degli scarti negativi.

La mediana è quel valore che minimizza la somma degli scarti assoluti.

Dove collocare un deposito (di merci, carburante, pezzi di ricambio, …) lungo un'autostrada con i punti di vendita ai km x_1, x_2, …, x_n in modo da minimizzare i costi di rifornimento dei punti di vendita? Si tratta di individuare un punto x in $[x_1, x_n]$ tale che sia minima la quantità $\Sigma c|\ x_i - x_n|$, in cui c è il costo unitario per rifornire il punto di vendita sito in x_i partendo dal deposito collocato in x. Il valore che minimizza tale costo complessivo è proprio x=Me.

A differenza della media aritmetica, la mediana non risente della presenza di valori anomali, in quanto tiene conto solo dell'ordinamento delle osservazioni, limitandosi a considerare la modalità dell'elemento centrale (resistenza).

Esempio 8

Voto agli esami dello studente A

Dopo aver ordinato le modalità in senso non decrescente

18, 18, 24, 24, 26, 27, 27, 28, 28, 28, 28, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30

poiché n=21 è dispari, la **MEDIANA** sarà:

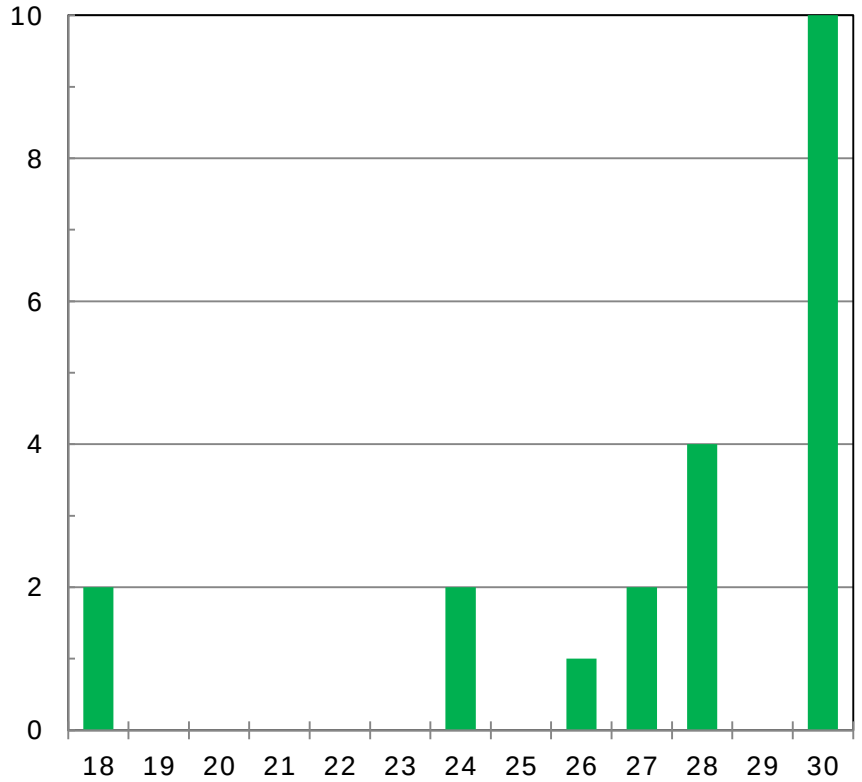
$$Me = x_{(n+1)/2} = x_{(21+1)/2} = x_{11} = 28$$

Indici di posizione

Moda

La moda di una distribuzione di frequenza è la **modalità** (o la classe di modalità) a cui corrisponde la massima frequenza (o la massima densità di frequenza, nel caso in cui le classi non siano equi-ampie); in altre parole, è il valore che compare più di frequente. Sintetizzare una variabile X tramite la sua moda significa, quindi, assumere come valore più rappresentativo della distribuzione quello che si è verificato più spesso di tutti gli altri.

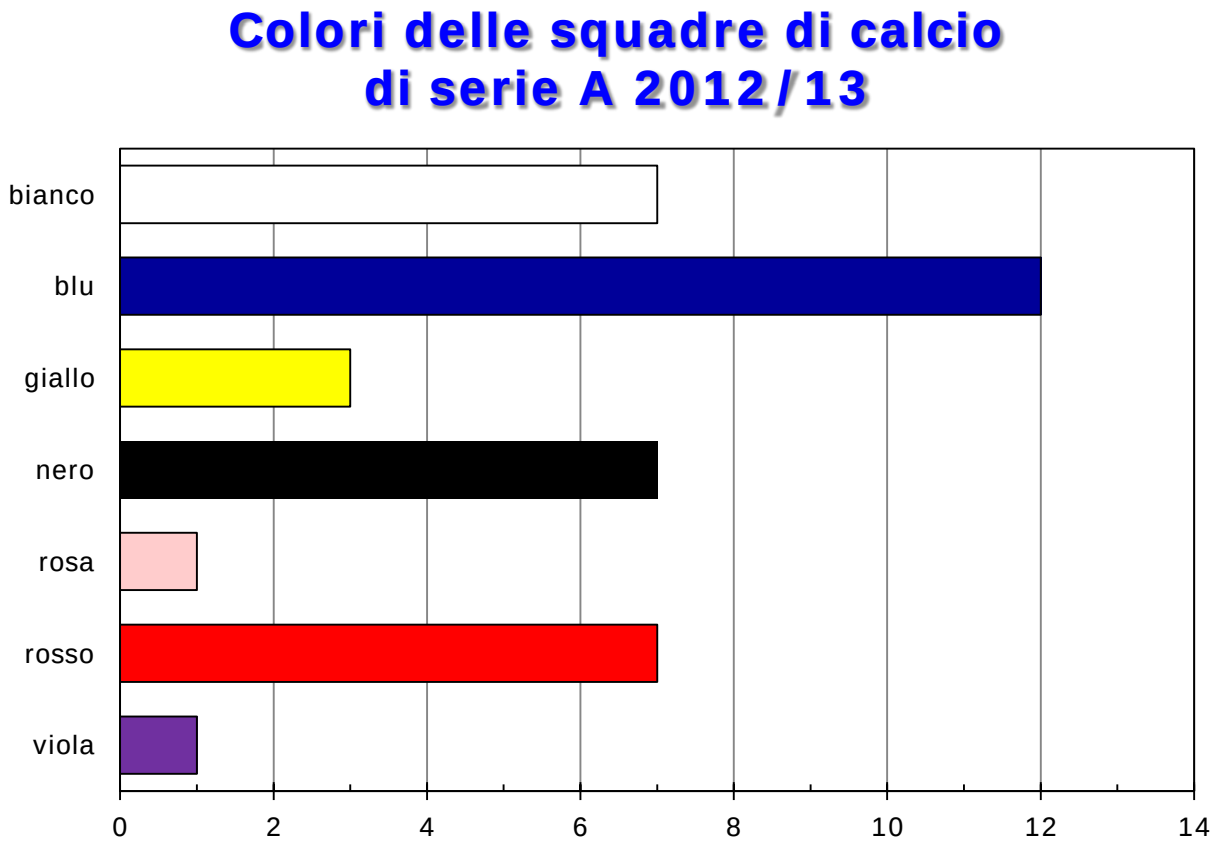
Voto agli esami dello studente A



Indici di posizione

Moda

L'utilità della moda risiede nell'essere l'unico degli indici di posizione a poter descrivere caratteri qualitativi nominali (o sconnessi).



Indici di posizione

Quantili

In statistica il **quantile** di ordine *a* è un valore *qa* che divide la popolazione in due parti, proporzionali ad *a* e (*1-a*) e caratterizzate da valori rispettivamente minori e maggiori di *qa*.

La **mediana** è il quantile di ordine 1/2.

I **quartili** sono i quantili di ordini 1/4, 2/4 e 3/4.

I **decili**, di ordine *m/10*, dividono la popolazione in 10 parti uguali.

I **centili**, di ordine *m/100*, dividono la popolazione in 100 parti uguali. Vengono anche chiamati **percentili**, esprimendo l'ordine in percentuale: *m/100=m%*.

Indici di posizione

Quantili

ECTS

Il sistema europeo di accumulazione e trasferimento dei crediti (ECTS) è basato sul carico di lavoro richiesto ad uno studente per raggiungere gli obiettivi di un corso di studio, obiettivi espressi preferibilmente in termini di risultati dell'apprendimento e di competenze acquisite.
La prestazione dello studente è documentata localmente da un voto che dipende dal sistema in uso. Nel caso italiano le votazioni sono espresse in trentesimi, con distribuzione delle votazioni dipendenti dall'insegnamento. Per favorire la trasferibilità dei crediti è allora buona pratica aggiungere alla votazione locale il voto ECTS.
Il sistema è basato sull'individuazione di 5 livelli di votazione basati sull'effettiva distribuzione dei voti nel corso frequentato nel periodo di riferimento, secondo lo schema che segue:

A al migliore 10%, **B** al successivo 25%, **C** al successivo 30%, **D** al successivo 25%, **E** al successivo 10%.

Esempio 9

Quantili - ECTS

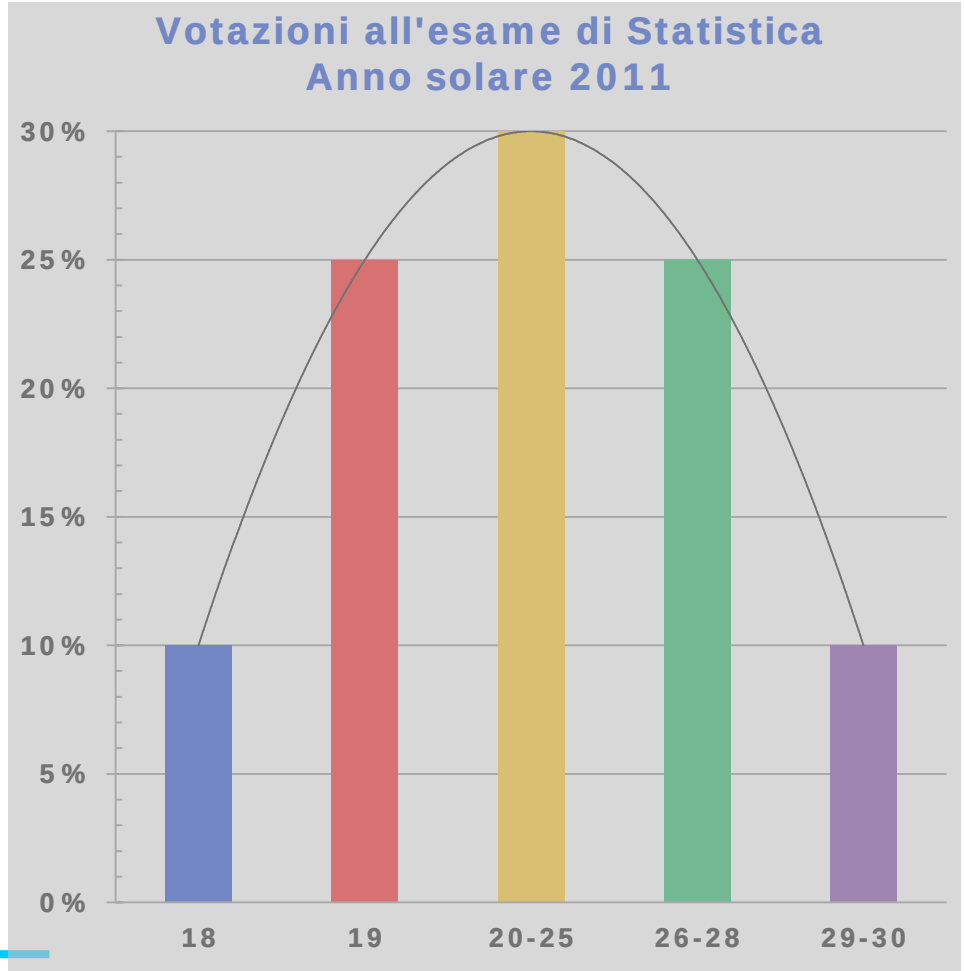
All'insegnamento di Statistica del corso di laurea in Scienze Politiche corrisponde nell'archivio degli esami 2011 la seguente sequenza: A=30, B=28, C=25, D=19, E=18. L'informazione che deve essere impiegata per il passaggio alla votazione ECTS è la seguente:

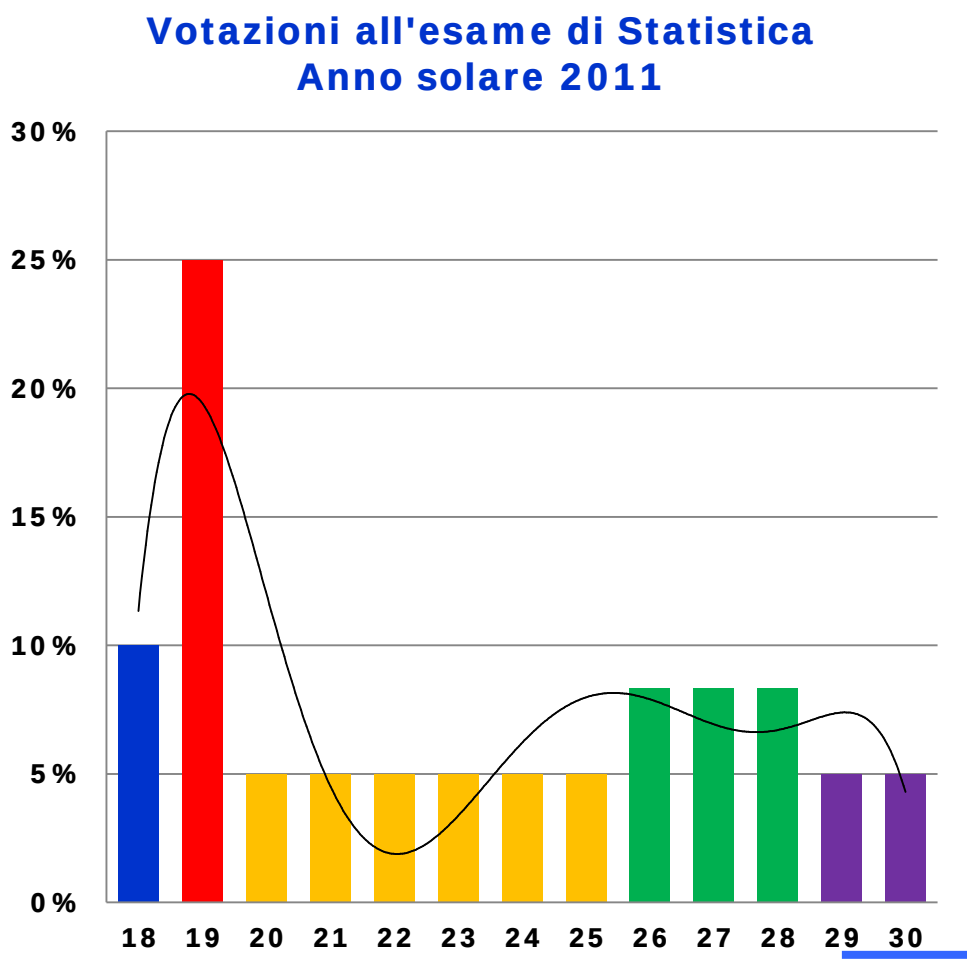
Votazione ECTS	Votazioni in trentesimi
A (30)	29, 30
B (28)	26, 27, 28
C (25)	20, 21, 22, 23, 24, 25
D (19)	19
E (18)	18

Nel caso in cui lo stesso voto espresso in trentesimi ricada in più classi, è opportuno riferirsi alla specifica documentazione ECTS o, in alternativa, adottare convenzioni locali.

Esempio 9

Quantili - ECTS





Indici di variabilità

La variabilità di un fenomeno è la sua attitudine ad assumere differenti modalità. Per misurarla occorre controllare se le singole unità statistiche presentano modalità più o meno stabili rispetto ad un indice di posizione, che è rappresentativo dell'intera distribuzione di frequenza.

L'indice più importante per misurare la variabilità di una distribuzione è espresso dalla media degli scarti dalla media μ al quadrato. Tale quantità si chiama **varianza** (Pearson, 1918).

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

Indici di variabilità

Da un punto di vista computazionale si può affermare che la varianza è pari alla media aritmetica dei quadrati meno il quadrato della media aritmetica. In pratica, per ottenere σ^2 basta sommare i valori delle modalità ed i corrispondenti quadrati, facendone poi le rispettive medie.

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

$$\sigma^2 = \frac{1}{21} * 16070 - 27{,}43^2 = 12{,}91$$

Indici di variabilità

Una difficoltà nell'interpretazione della varianza deriva dal fatto che è espressa nell'unità di misura del fenomeno al quadrato. Pearson pertanto propose lo **scarto quadratico medio** (o deviazione standard), che rappresenta la media quadratica degli scarti dalla media μ

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

$$\sigma = \sqrt{\frac{1}{21} * 16070 - 27{,}43^2} = \sqrt{12{,}91} = 3{,}59$$

Indici di variabilità

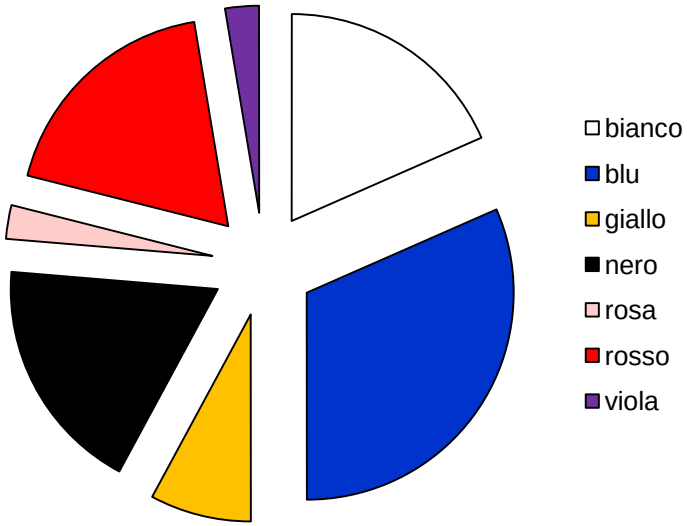
Poiché varianza e scarto quadratico medio sono indici assoluti, è opportuno introdurre indici relativi o normalizzati. Un indice relativo molto usato, purché μ>0, è il rapporto tra lo scarto quadratico medio σ e la media aritmetica μ: il **coefficiente di variazione** Cv

$$Cv = \frac{\sigma}{\mu} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}}{\mu} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\mu}\right)^2}$$

Il coefficiente di variazione Cv è un indice di dispersione che permette di confrontare misure di fenomeni riferite a unità di misura differenti, in quanto si tratta di un numero puro. Esso misura la variazione media del fenomeno in rapporto alla sua media aritmetica.

Rappresentazioni grafiche

Carattere qualitativo sconnesso
Colori delle squadre di Serie A 2012/13



Torta

Il **diagramma a torta** è un grafico che si ottiene dividendo un cerchio in spicchi (settori circolari), le cui ampiezze angolari sono proporzionali alle classi di frequenza, mentre le aree sono proporzionali alle frequenze. Tale grafico evidenzia la composizione di un fenomeno, è utile quando si è interessati a valutare una parte sul tutto, ma va usato con attenzione, specialmente quando il carattere è composto da molte modalità e quando queste assumono valori simili.

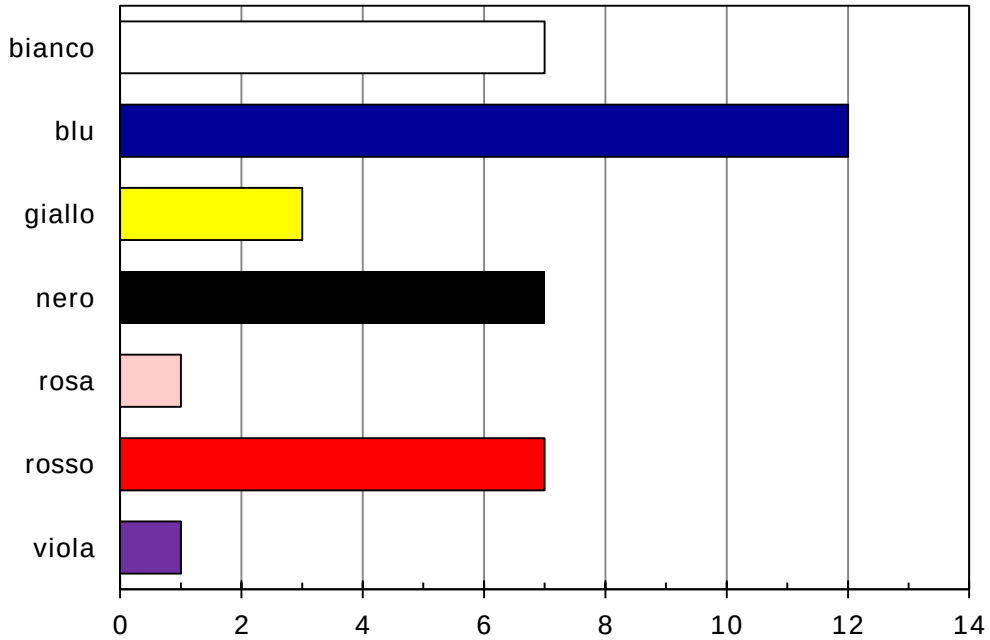


Diagramma a nastri

Il **diagramma a nastri** è un grafico che si ottiene costruendo tanti rettangoli quante sono le modalità del carattere da rappresentare. Tali rettangoli sono paralleli all'asse delle ascisse: la loro altezza è fissa, mentre la base è proporzionale alla frequenza (assoluta o relativa) che rappresentano.

Rappresentazioni grafiche

Carattere qualitativo ordinale
Studenti per titolo di studio

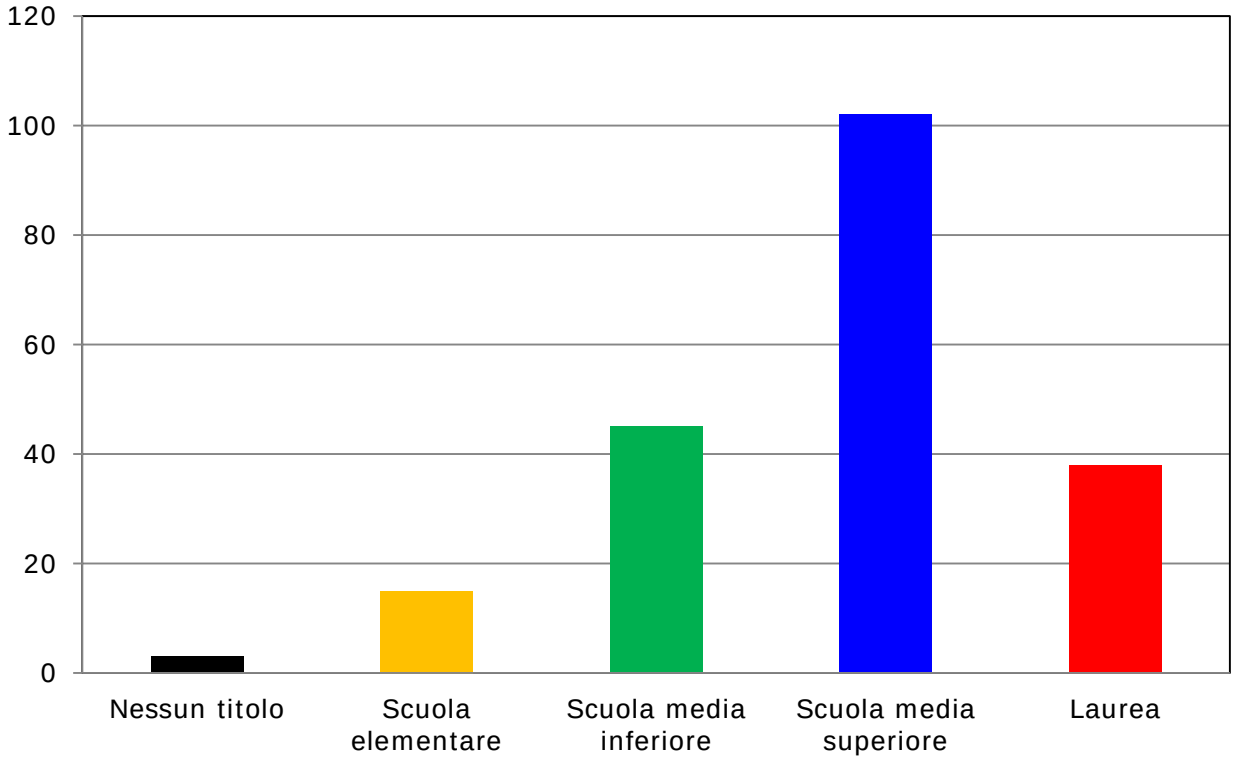


Diagramma a colonne

Il **diagramma a colonne** è un grafico che si ottiene costruendo tanti rettangoli quante sono le modalità del carattere da rappresentare. Tali rettangoli sono paralleli all'asse delle ordinate: la loro base è fissa, mentre l'altezza è proporzionale alla frequenza (assoluta o relativa) che rappresentano.

Rappresentazioni grafiche

Carattere quantitativo discreto
Famiglie per numero di componenti

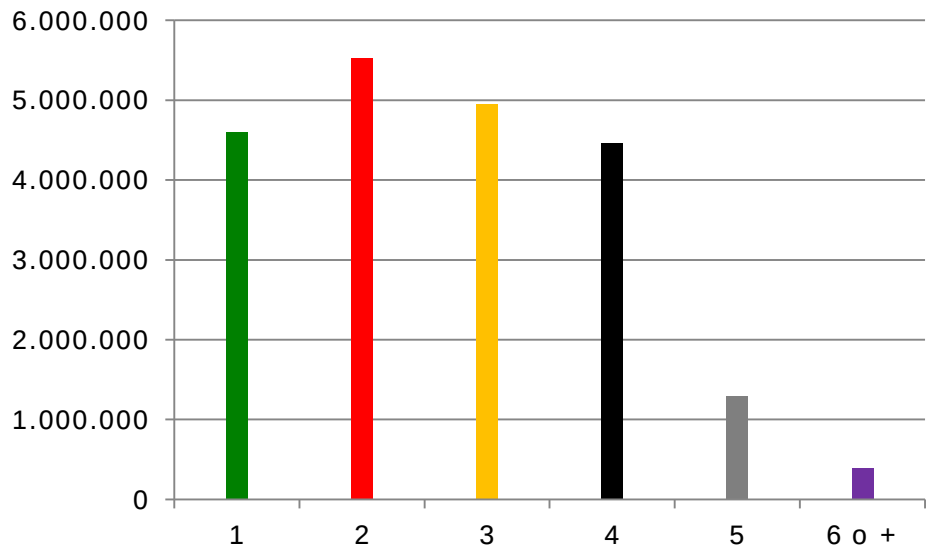
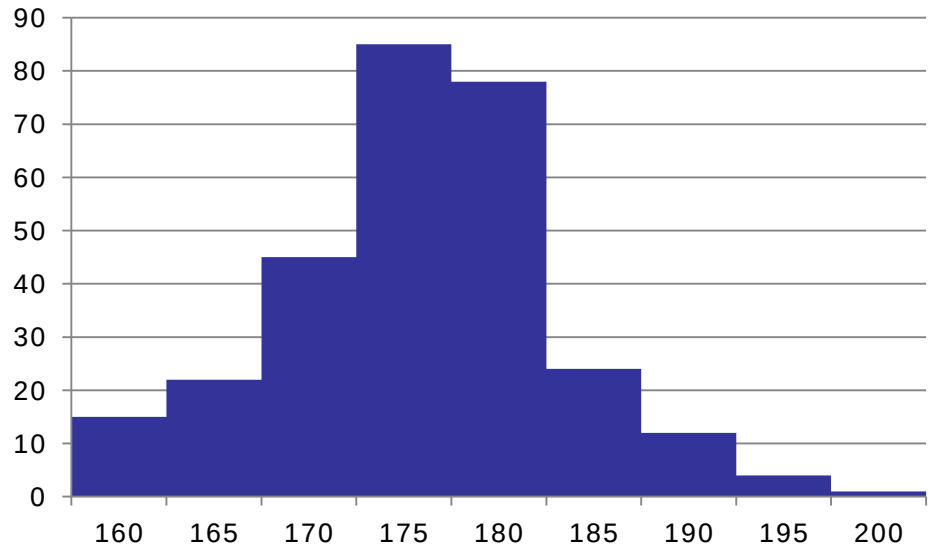


Diagramma a bastoncini

Il **diagramma a bastoncini** è un grafico che si costruisce disegnando, in corrispondenza di ogni valore osservato, un bastoncino (perpendicolare all'asse delle ascisse) di lunghezza uguale alla frequenza assoluta con cui quel valore è stato osservato.

Rappresentazioni grafiche

Carattere quantitativo continuo
Altezza di un gruppo di studenti



Istogramma

L'**istogramma** è costituito da rettangoli adiacenti, le cui basi sono allineate su un asse orientato. Ogni rettangolo ha la base di lunghezza pari all'ampiezza della corrispondente classe, mentre l'altezza è calcolata come densità di frequenza, pari al rapporto fra la frequenza associata alla classe e l'ampiezza della classe. L'area della superficie di ogni rettangolo coincide con la frequenza associata alla classe cui il rettangolo si riferisce.

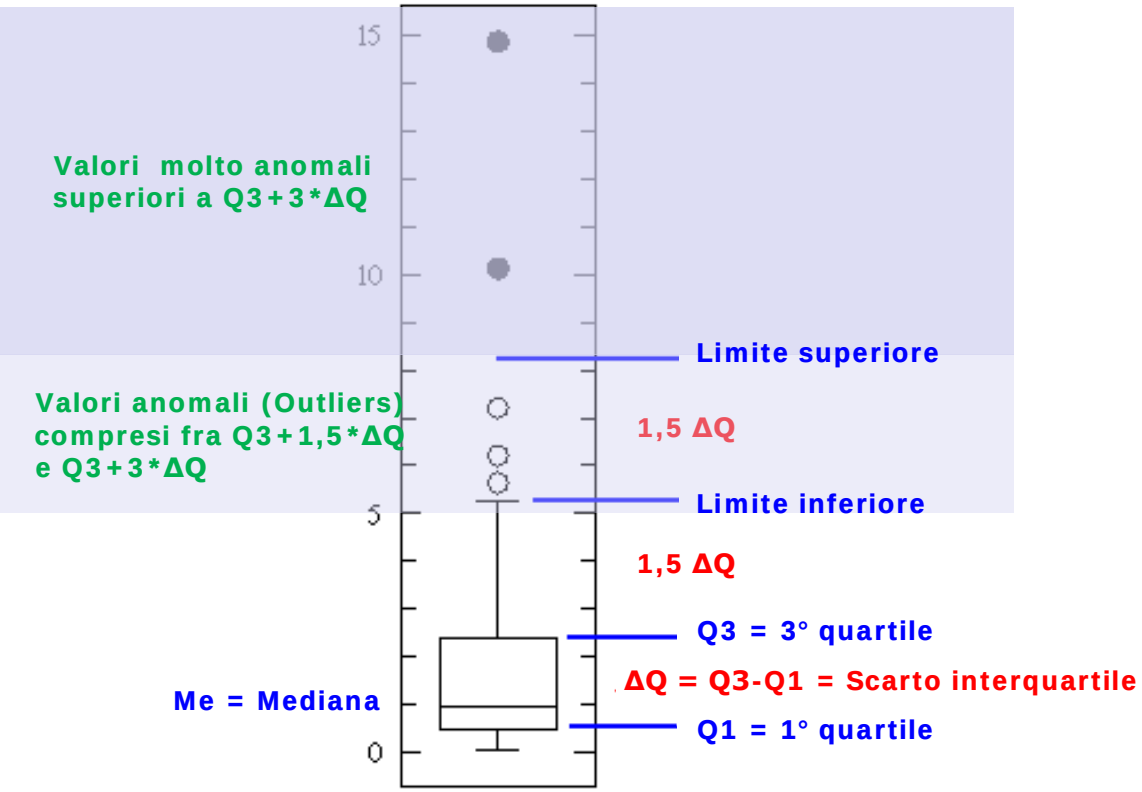
Rappresentazioni grafiche

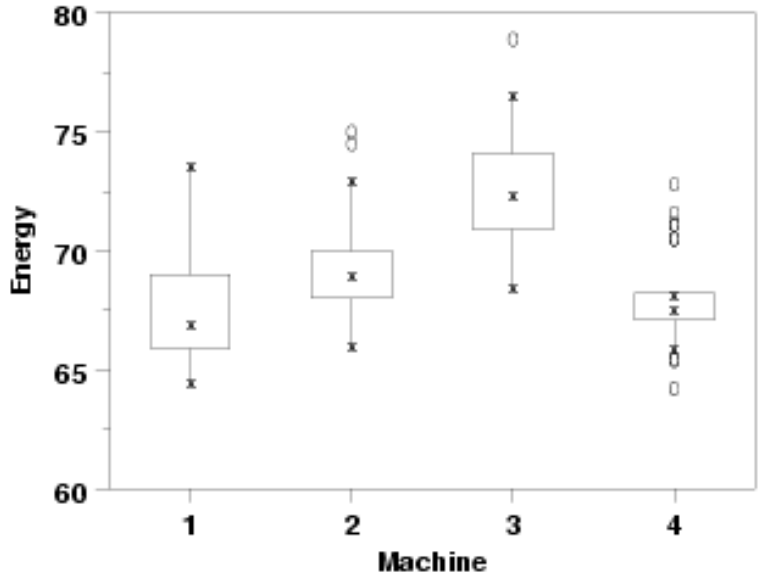
Il grafico MIGLIORE

Grafico per tipologia di carattere	Qualitativo		Quantitativo	
	Sconnesso	Ordinale	Discreto	Continuo / Discreto in classi
Torta	Sì	Sì	Sì	Sì
Nastri	**Sì**	Sì	Sì	Sì
Colonne	Sì	**Sì**	Sì	Sì
Bastoncini	No	No	**Sì**	No
Istogramma	No	No	No	**Sì**

Rappresentazioni grafiche

Carattere quantitativo





Box-plot

Il **box-plot** descrive la distribuzione di un carattere quantitativo attraverso indici di posizione e consente facili confronti fra distribuzioni diverse. È rappresentato da un rettangolo diviso in due parti, da cui escono due segmenti: il rettangolo (la "scatola") è delimitato dal 1° e 3° quartile (Q1 e Q3) e diviso al suo interno dalla mediana (Me); i segmenti (i "baffi") sono delimitati dal minimo e dal massimo dei valori. In alternativa, per evitare valori anomali, gli estremi sono calcolati come Q1-1,5*ΔQ e Q3+1,5*ΔQ.

Giacomo Bulgarelli
Ufficio Servizi Statistici

SERVIZIO DAF: FONTI STATISTICHE

Mercoledì 3 ottobre 2012

FINE PARTE 4