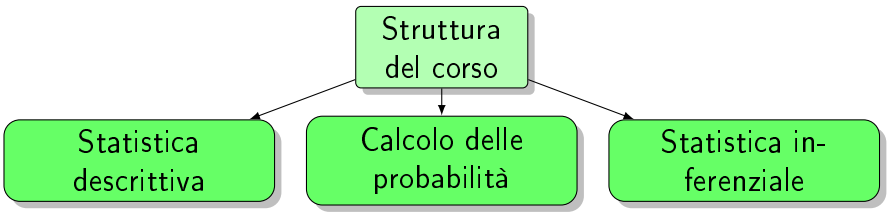
Appunti di Probabilità e Statistica

a.a. 2014/2015 C.d.L. Informatica – Bioinformatica

I. Oliva

Lezione 1

1 Introduzione



Statistica descrittiva: metodi per organizzare, riassumere, presentare dati in modo informativo

Esempio 1.1. *Il 49% degli elettori in un Comune conosce il primo libro della Bibbia*

Probabilità: strumento matematico per la *misura dell'incertezza*

Statistica inferenziale: tecniche matematiche per avere una predizione su una popolazione, basate sull'analisi di una porzione (*stime*)

Esempio 1.2. *indice di share televisivo*

STATISTICA: disciplina che elabora i principi e le metodolgie che presiedono al processo di rilevazione e di raccolta dei *dati,* alla loro rappresentazione sintetica ed alla loro interpretazione e, laddove ve ne siano le condizioni, alla generalizzazione delle evidenze osservate.

Dati (statistici): informazioni espresse numericamente, riferite ad un insieme di entità omogenee, rispetto ad un determinato punto di vista (*insieme di riferimento*).

dato	**insieme di riferimento**
Variazione media prezzi ISTAT	prezzi di beni e servizi sul mercato al consumo
Num. occupati/disoccupati	popolazione attiva
Num. incidenti mortali nel secondo trimestre 2014	sinistri verificatisi
orientamento di voto (sondaggi)	cittadini aventi diritto di voto

Cenni storici

- 550 a.C. (Libro di Confucio) $\longrightarrow$ Informazioni su statistica agraria, artigiana, commerciale in Cina
- $\approx$ 1550 $\longrightarrow$ Statistiche demografiche (Concilio di Trento)
- XVII secolo $\longrightarrow$ Political Arithmetic (uso del metodo empirico induttivo nelle scienze sociali) (J. Graunt, W. Patty – Inghilterra) & introduzione corso universitario di Scienze Politiche (H. Conring – Germania)
- XVIII-XIX secolo $\longrightarrow$ Calcolo delle Probabilità (Pascal, Fermat, De Moivre, Bernoulli, Bayes, Laplace, Legendre, Gauss)
- XIX-XX secolo $\longrightarrow$ Statistica nelle scienze empiriche (scienze naturali, scienze economiche), applicata a fenomeni sociali

2 Concetti preliminari

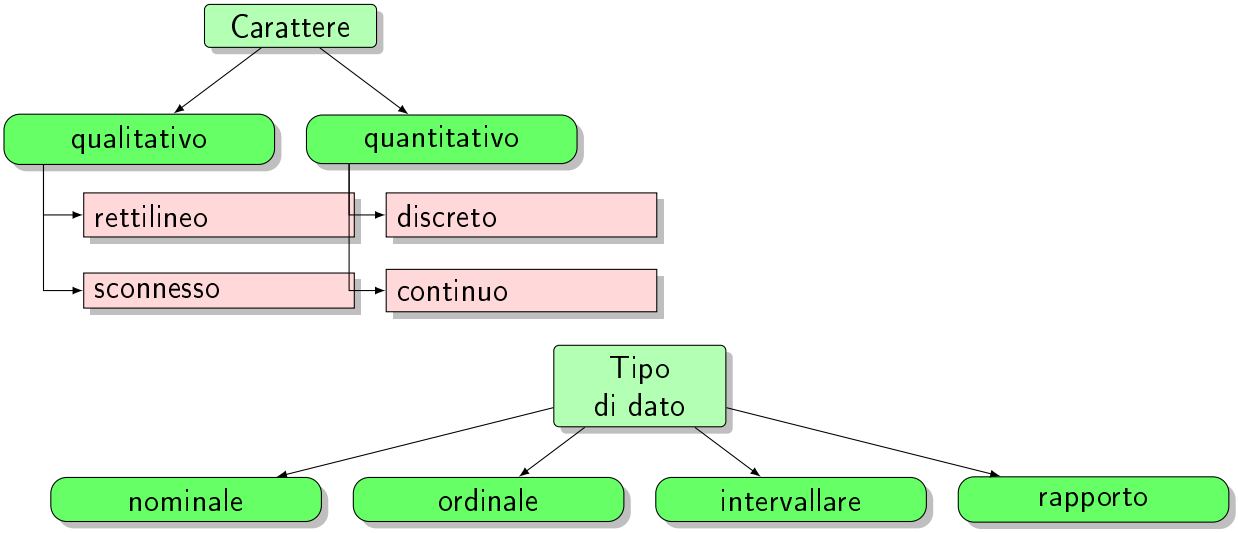
Unità statistica: caso individuale, oggetto di osservazione

Collettivo statistico: insieme di più unità statistiche, omogenee rispetto ad uno o più aspetti

Carattere: aspetto elementare che è oggetto di rilevazione tra le unità statistiche

Modalità: modi con i quali il carattere di presenta nelle unità statistiche del collettivo

2.1 Classificazione dei caratteri



Esempio 2.1. *Indagine statistica relativa al mezzo di trasporto utilizzato per raggiungere l'Università dagli studenti di questo Corso.*

collettivo: *tutti gli studenti universitari che frequentano il corso di Probabilità e Statistica di Univr.*

unità: *singolo studente*

3 Rappresentazione dei dati

I dati codificati di una rilevazione statistica effettuata su n unità statistiche, con riferimento a p caratteri (variabili), vengono raccolti in una tabella che viene chiamata *matrice dei dati*.

Carattere	Tipologia	Modalità
mezzo di trasporto	qualitativo connesso	Auto, scooter, autobus, treno
tempo di percorrenza	quantitativo continuo	$x \in \mathbb{R}$
costo	quantitativo continuo	$x \in \mathbb{R}$
Sesso	qualitativo	maschile, femminile
Residenza	qualitativo sconnesso	città, regione, indirizzo
Status lavorativo	qualitativo	lavoratore, non lavoratore

IND.	SESSO	ANASC.	SCUOLA	MAT.	ORDR	STAT	ECON	DIR	RMAT
3	2	1963	07	58	06	30	30	31	3
1	1	1964	07	50	06	23	24	26	3
2	2	1964	02	51	20	26	27	30	1
3	2	1963	07	48	06	23	24	27	-2
3	2	1968	02	48	08	24	29	26	1
1	1	1966	02	60	11	31	30	30	2
1	1	1968	02	42	07	31	30	30	2
1	2	1967	07	53	11	30	28	28	2
3	1	1968	02	40	10	31	19	21	0
3	1	1968	01	60	10	30	24	30	0
2	1	1968	02	40	07	21	21	24	-1
1	2	1968	07	50	04	29	24	26	2
2	1	1964	02	44	22	28	24	30	3
2	2	1967	02	50	13	19	30	30	2
1	1	1967	07	46	09	24	30	28	2
1	1	1966	02	36	11	25	23	22	0
3	2	1968	01	60	11	24	26	30	-3
3	2	1965	05	50	07	18	24	27	-3

Legenda:
IND: Indirizzo 1 - Economico; 2 - Sociologico; 3 - Amministrativo; 4 - Altro
SESSO: 1 - Maschio; 2 - Femminina
ANASC.: Anno di nascita
SCUOLA: 01 - Liceo Classico; 02 - Liceo Scientifico; 03 - Liceo Linguistico; 05 - Ist. Magistrale; 06 - ITIS; 07 - Ist. Tec. Commerciale; 08 - Ist. Tec. Geometra; 12 - Ist. Tec. Aziendale; 14 - Ist. Tec. non specificato; 21 - Ist. Profess. non specificato;
MAT.: Voto conseguito alla maturità
ORDR: Ordine di registrazione sul libretto dell'esame di statistica
STAT: Voto conseguito in statistica (31 per lode)
ECON: Voto conseguito in economia
DIR: Voto conseguito in diritto
RMAT: Opinione dello studente circa l'influenza (positiva o negativa) delle conoscenze precedenti di matematica sull'esito dell'esame di statistica (da -3 a $+3$)

Osserviamo che:

- la matrice dei dati contiene tutte le informazioni analitiche di ciascuna

unità statistica

- la riga $i-$sima rappresenta l'$i-$sima unità statistica, la colonna $j-$sima rappresenta il $j-$simo carattere
- Quando i dati sono molti, l'analisi diretta della matrice non consente di cogliere in via immediata gli aspetti salienti del fenomeno
- Occorre una sintesi attraverso un'elaborazione statistica dei dati (*indici statistici*)

Cosa succede quando il numero di modalità che il carattere può assumere è molto elevato? Si ricorre al *raggruppamento dei dati statistici.*

- Se il carattere è *qualitativo* $\longrightarrow$ accorpamento delle modalità
- Se il carattere è *quantitativo* $\longrightarrow$ suddivisione in classi

In entrambi i casi, si parla di *classi di modalità.* I criteri per la costruzione sono:

1. il numero di classi deve essere abbastanza piccolo da fornire una adeguata sintesi, ma abbastanza grande da mantenere un livello accettabile di dettaglio dell'informazione
2. le classi devono essere *disgiunte* (mutua esclusività)
3. le classi devono comprendere tutte le possibili modalità del carattere (esaustività)
4. le classi devono avere la stessa ampiezza (criterio facoltativo)

Dato un carattere quantitativo, è possibile determinare classi *equiampie* o *equifrequenti.*

Classi equiampie:

- X_{max} e X_{min} sono il più alto ed il più basso valore del carattere X
- ampiezza delle classi A
- il numero di classi K
- La relazione cui si fa riferimento è $A = \frac{(X_{max} - X_{min})}{K}$

Classi equifrequenti:

- ordinamento crescente dei valori della modalità del carattere
- frequenza associata a ciascuna classe (quante volte la modalità x si presenta nel collettivo?)

PROBLEMA: una volta raccolti i dati, come li rappresentiamo?

SOLUZIONE: *distribuzione di frequenza* $\longrightarrow$ numero di unità statistiche che presentano una determinata modalità

modalità	x_1	$x_2 \cdots$	x_k	TOT
frequenza	n_1	$n_2 \cdots$	n_k	N

dove n_i numero di unità che presentano la modalità x_i (frequenza assoluta) e N totale unità statistiche osservate.

$$f_i = \frac{n_i}{N},\ \forall\, i = 1, \ldots, k, \quad \text{frequenza relativa}$$
$$p_i = f_i \cdot 100,\ \forall\, i = 1, \ldots, k, \quad \text{frequenza percentuale}$$
$$N_i = \sum_{j=1}^{i} n_j,\ \forall\, i = 1, \ldots, k, \quad \text{frequenza assoluta cumulata}$$
$$F_i = \sum_{j=1}^{i} f_j,\ \forall\, i = 1, \ldots, k, \quad \text{frequenza relativa cumulata}$$

Esempio 3.1. *Due esempi di distribuzioni di frequenza (caratteri qualitativi e quantitativi)*

età	n_i	f_i	p_i
10–29	5	0.25	25%
30–49	9	0.45	45%
50–69	4	0.2	20%
70–89	2	0.1	10%
tot	20	1	100%

colore occhi	n_i	f_i	p_i
nero	8	0.4	40%
marrone	4	0.2	20%
azzurro	6	0.3	30%
verde	2	0.1	10%
tot	20	1	100%

3.1 Rappresentazione grafica dei dati

Una volta che i dati statistici siano stati raccolti e raggruppati, occorre illustrarli, in modo che la successiva analisi risulti il più facile possibile.

Gli aspetti da tenere in considerazione per una efficace rappresentazione grafica sono:

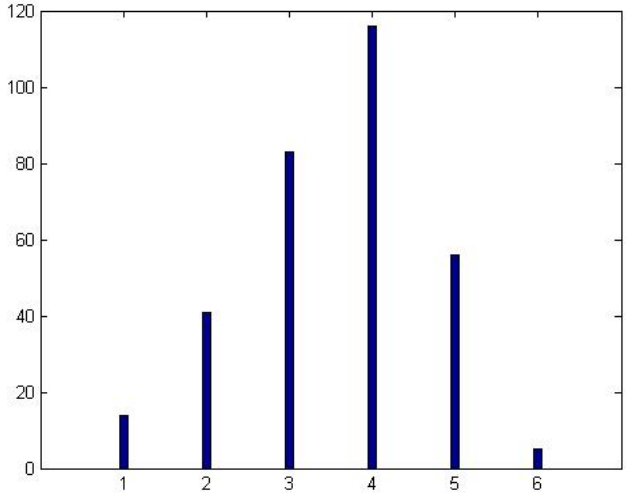
- accuratezza
- semplicità
- chiarezza
- aspetto
- struttura

Grafici a barre: ciascuna barra è associata ad una modalità del carattere considerato, inoltre tutte le barre hanno la stessa larghezza, mentre l'altezza delle barre è proporzionale alle frequenze delle modalità.

Molto utili per rappresentare distribuzioni di frequenze per caratteri qualitativi.

Per esempio, si consideri la distribuzione di frequenza del numero di esami sostenuti alla fine del primo anno:

num. esami	0	1	2	3	4	5	TOT
frequenza	14	41	83	116	56	5	315



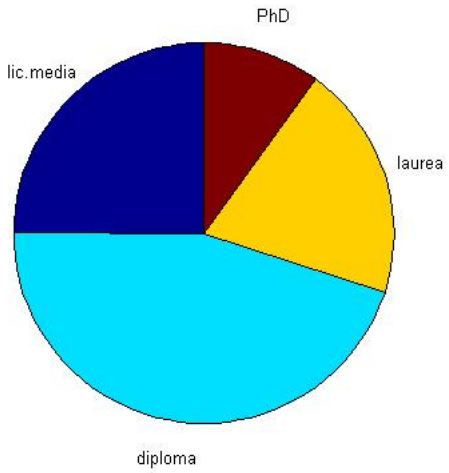
Grafici a torta: utili per rappresentare la composizione di un aggregato. Ciascuna *fetta di torta* (settore circolare) è associata ad una modalità del carattere. L'ampiezza di ciascuna fetta è proporzionale alla frequenza della modalità.

L'angolo al centro g_i associato all'$i-$sima modalità con percentuale p_i è dato da

$$p_i \ : \ 100 = 360 \ : \ g_i, \quad \text{da cui } g_i = \frac{p_i \cdot 360}{100} \ .$$

Per esempio, si consideri la distribuzione di frequenze percentuali relativa al titolo di studio dei padri dei 200 studenti iscritti al primo anno (C.d.L. Informatica):

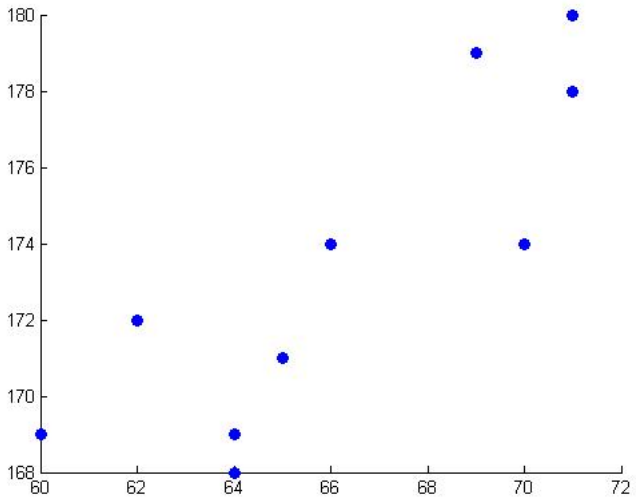
titolo di studio	p_i
licenza media	50
diploma	90
laurea	40
PhD	20



Grafici a punti: utili per rappresentare il valore assunto da due variabili su una stessa unità statistica, in modo da verificare se esista connessione tra le variabili. Ogni unità statistica è rappresentata da un punto nel piano cartesiano.

Per esempio, si consideri la distribuzione di peso e altezza di 10 atleti:

atleta	peso (kg)	altezza (cm)
M	66	174
P	64	168
L	65	171
G	71	178
S	64	169
F	70	174
A	71	180
O	62	172
B	60	169
E	69	179



Istogrammi: grafico costituito da barre non distanziate, con basi non necessariamente uguali. L'area di ogni barra è proporzionale alla frequenza della modalità cui si riferisce.

Se il carattere è quantitativo, discreto o continuo, la distribuzione di frequenza può essere ottenuta a partire da classi di stessa ampiezza o ampiezze diverse; nel primo caso, si avrà un istogramma a basi regolari.

L'area di ciascun rettangolo deve essere proporzionale alla frequenza, l'altezza h deve pertanto essere proporzionale al rapporto tra la frequenza da rappresentare e l'ampiezza della $i-$sima classe.

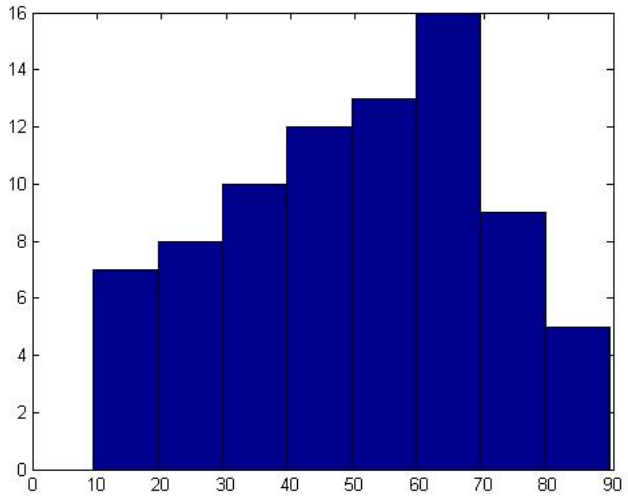
L'altezza dei rettangoli si chiama *densità di frequenza*

$$n_i = a_i \times h \;\Rightarrow\; h = \frac{n_i}{a_i}$$

È possibile sostituire la frequenza assoluta n_i con la frequenza relativa f_i.

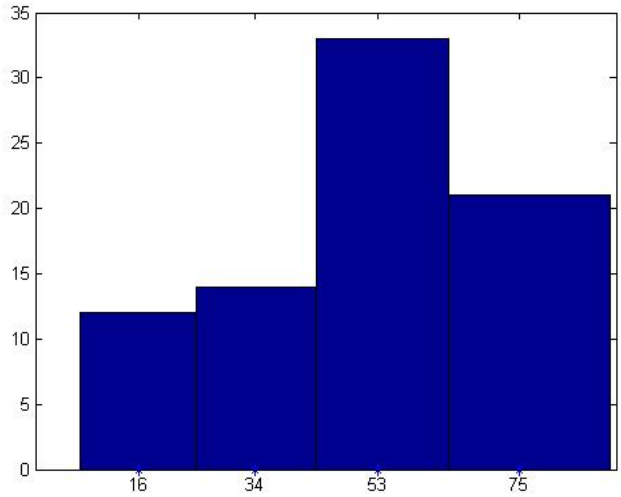
Per esempio, si consideri la distribuzione dell'età (in anni compiuti) in un condominio. Le classi, in questo caso, sono già definite ed hanno tutte la stessa ampiezza:

età	n_i	f_i
10–19	7	0.087
20–29	8	0.100
30–39	10	0.125
40–49	12	0.150
50–59	13	0.163
60–69	16	0.200
70–79	9	0.113
80–90	5	0.062
tot	80	1

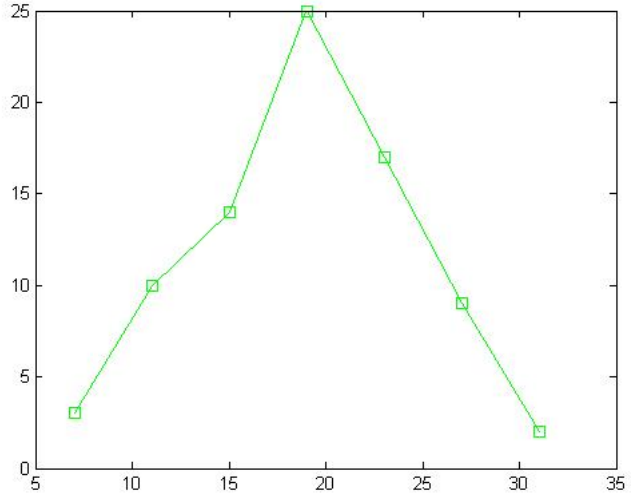


Altro esempio: stesso collettivo statistico, raggruppato in classi non equispaziate. In tal caso, occorre determinare anche la densità di ciascuna classe.

età	a_i	n_i	f_i	d_i
10–22	12	10	0.15	0.0125
23–45	22	19	0.45	0.02
46–60	14	23	0.25	0.018
61–90	29	28	0.15	0.005
tot		80	1	20



Poligono di frequenza: linea poligonale che unisce i punti centrali delle basi superiori dei rettangoli dell'istogramma. Vantaggio: agevola il confronto tra distribuzioni diverse, utilizzando lo stesso grafico.



Funzione di ripartizione: permette di rappresentare la distribuzione delle frequenze relative cumulate.

Si consideri il carattere X quantitativo discreto con $K+1$ modalità t.c. $x_0 \leq x_1 \leq \ldots \leq x_K$, oppure il carattere X quantitativo continuo, suddiviso in K classi $[x_0, x_1], (x_1, x_2], \ldots, (x_{K-1}, x_K]$. Allora, si ha:

$$F(x) = \begin{cases} 0, & \text{se } x \leq x_0 \\ F_1, & \text{se } x_0 < x \leq x_1 \\ F_2, & \text{se } x_1 < x \leq x_2 \\ \ldots & \ldots \\ F_K, & \text{se } x_{K-1} < x \leq x_K \\ 1, & \text{se } x \geq x_K \end{cases}$$

dove F_i è l'$i-$sima frequenza relativa cumulata.

La funzione di ripartizione di X con campo di variazione $[x_0, x_K]$ gode delle seguenti proprietà (che dimostreremo in seguito):

1. $F(X) = 0$ per $x < x_0$
2. $F(X) = 1$ per $x > x_K$
3. $F(X)$ è una funzione non decrescente

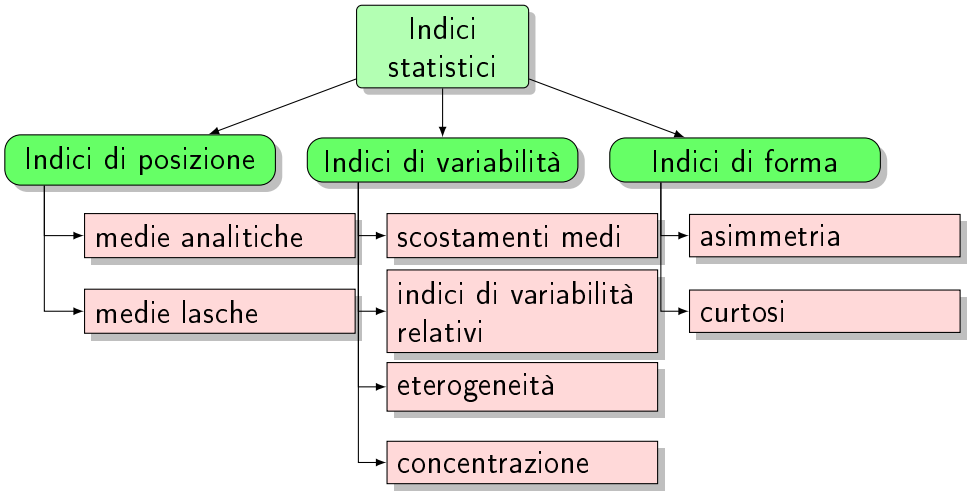
Si rappresenta attraverso una *step function,* nel caso di carattere discreto, o attraverso una *polinomiale a tratti,* nel caso di un carattere continuo.

4 Indici statistici

Gli indici statistici consentono di esprimere con un'unica misura numerica l'intera distribuzione di un carattere su un collettivo.

VANTAGGI:

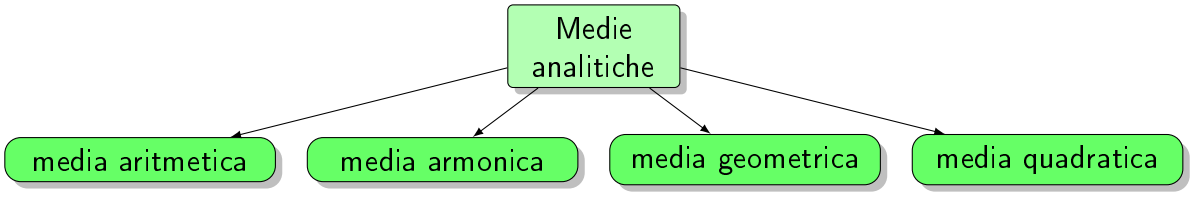
- si possono confrontare distribuzioni di un carattere nel tempo e/o nello spazio, in circostanze diverse
- è possibile verificare gli effetti (in termini di variazione, direzione e intensità) di una determinata azione sulla distribuzione del carattere considerato



4.1 Indici di posizione

Gli indici di posizione sono rappresentati dalle *medie,* i.e., indicatori statistici che permettono di rappresentare *l'ordine di grandezza* del fenomeno osservato.

Distinguiamo le *medie analitiche* e le *medie lasche.*



Le medie analitiche tengono conto di tutti i valori e vengono calcolate attraverso operazioni algebriche su modalità di caratteri quantitativi. Si parla anche di *medie di potenze.*

Media aritmetica: Si indica con la lettera greca μ. Indichiamo con N il numero totale di modalità e con x_j la $j-$sima modalità del carattere, allora:

$$\mu = \frac{1}{N}\sum_{j=1}^{N} x_j, \quad \text{dati disaggregati}$$

$$\mu = \frac{\sum_{j=1}^{N} x_j n_j}{\sum_{j=1}^{N} n_j}, \quad \text{dati organizzati in frequenze}$$

$$\mu = \sum_{j=1}^{N} x_j f_j, \quad \text{media per frequenze relative}$$

$$\mu = \frac{1}{N}\sum_{j=1}^{N} \bar{x}_j n_j, \quad \text{media per dati organizzati in classi, dove}$$

$$\bar{x}_j = \frac{c_j + c_{j-1}}{2},\ c_{j-1}, c_j \text{ estremi della classe } \forall\, j = 1, \ldots, N.$$

Proprietà:

1. Criterio di internalità: se $m = \min\{x_1, \ldots, x_N\}$ e $M = \max\{x_1, \ldots, x_N\}$, allora $m \leq \mu \leq M$.

   Proof. Consideriamo i dati della distribuzione ed ordiniamoli: otterremo $x_{(1)}, \ldots, x_{(i)}, \ldots, x_{(N)}$, con $x_{(1)} \leq x_{(i)} \leq x_{(N)}$, per ogni $i = 1, \ldots, N$. Sommiamo i tre membri della precedente disuguaglianza:

   $$\begin{aligned} &\sum_{i=1^N} x_{(1)} \leq \sum_{i=1^N} x_{(i)} \leq \sum_{i=1^N} x_{(N)} \\ &\Rightarrow\ N x_{(1)} \leq N\mu \leq N x_{(N)} \\ &\Rightarrow\ x_{(1)} \leq \mu \leq x_{(N)} \end{aligned}$$

   dove $x_{(1)} = \min\{x_1, \ldots, x_N\}$ e $x_{(N)} = \max\{x_1, \ldots, x_N\}$. □

2. Baricentro: la somma degli scarti dalla media è nulla, in simboli $\sum_{j=1}^{N}(x_j - \mu) = 0$.

Proof. Avremo:

$$\sum_{j=1}^{N}(x_j - \mu) = \sum_{j=1}^{N} x_j - \sum_{j=1}^{N} \mu = N\mu - N\mu = 0 \, .$$

□

3. Linearità: se $Y = aX + b$, allora $\mu(Y) = a\mu(X) + b$.

   Proof. Esercizio. □

4. Associatività: sia X una variabile osservata su più gruppi. La media può essere ottenuta come media delle medie calcolate in ciascun gruppo, tenendo conto della differente numerosità dei singoli gruppi. Il collettivo è suddiviso in K gruppi di numerosità $n_1, n_2, \ldots, n_K$. La media del carattere X sul collettivo è μ. Per la proprietà associativa, si avrà

   $$\mu = \mu_1 \cdot \frac{n_1}{N} + \ldots + \mu_K \cdot \frac{n_K}{N} \, .$$

5. Minimizzazione dei quadrati degli scarti: la media aritmetica rende minima la somma dei quadrati degli scarti, in simboli $\sum_{j=1}^{N}(x_j - \mu)^2 = \min$.
6. Non robustezza
7. Rappresentatività nei confronti di distribuzioni simmetriche

Media armonica: costruita come il reciproco della media aritmetica dei reciproci delle modalità, riferite alle N unità di un carattere quantitativo.

$$\mu_a = \frac{N}{\sum_{j=1}^{N} \frac{1}{x_j}}, \quad \text{dati disaggregati}$$

$$\mu_a = \frac{\sum_{j=1}^{N} n_j}{\sum_{j=1}^{N} \frac{n_j}{x_j}}, \quad \text{dati organizzati in frequenze}$$

$$\mu_a = \frac{\sum_{j=1}^{N} n_j}{\sum_{j=1}^{N} \frac{n_j}{\bar{x}_j}}, \quad \text{media per dati organizzati in classi, dove}$$

$$\bar{x}_j = \frac{c_j + c_{j-1}}{2}, \; c_{j-1}, c_j \text{ estremi della classe } \forall\, j = 1, \ldots, N.$$

Proprietà:

1. la media armonica è principalmente usata nei problemi in cui vi siano legami inversi del fenomeno considerato con altri fenomeni (e.g., velocità e tempo)
2. la media armonica è principalmente usata quando i dati si presentano sottoforma di progressione armonica ($x_j = x_{j-1} + d$)
3. se $x_j = 0$, per qualche $j = 1, \ldots, K$, non si può calcolare la media armonica.

Media geometrica: definita come la radice $N-$sima del prodotto dei valori assunti dal carattere quantitativo.

$$\mu_g = \sqrt[N]{\prod_{j=1}^{N} x_j}, \quad \text{dati disaggregati}$$

$$\mu_g = \sqrt[N]{\prod_{j=1}^{N} x_j^{n_j}}, \quad \text{dati organizzati in frequenze}$$

$$\mu_g = \sqrt[N]{\prod_{j=1}^{N} \bar{x}_j^{n_j}}, \quad \text{media per dati organizzati in classi, dove}$$

$$\bar{x}_j = \frac{c_j + c_{j-1}}{2},\ c_{j-1}, c_j \text{ estremi della classe } \forall\, j = 1, \ldots, N.$$

Proprietà:

1. La media geometrica si esprime anche in modo diverso, ma del tutto equivalente, in termini di funzione esponenziale. Questa forma risulta essere più comoda da applicare:

$$\mu_g = \exp\left\{\frac{1}{N}\sum_{j=1}^{N} \ln(x_j)\right\}, \quad \text{dati disaggregati}$$

$$\mu_g = \exp\left\{\frac{1}{N}\sum_{j=1}^{N} n_j \ln(x_j)\right\}, \quad \text{dati organizzati in frequenze}$$

$$\mu_g = \exp\left\{\frac{1}{N}\sum_{j=1}^{N} \ln(\bar{x}_j)\right\}, \quad \text{media per dati organizzati in classi, dove}$$

$$\bar{x}_j = \frac{c_j + c_{j-1}}{2},\ c_{j-1}, c_j \text{ estremi della classe } \forall\, j = 1, \ldots, N$$

Proof. Si applicano le proprietà della funzione logaritmo e della funzione esponenziale. □

2. la media geometrica non può essere calcolata se esiste almeno un indice $j = 1, \ldots, N$ tale che $x_j \leq 0$

   Proof. Banale. □

3. La media geometrica è usata nel caso in cui i dati sipresentano sottoforma di progressione geometrica ($x_j = x_{j-1} \cdot r$)
4. La media geometrica è usata quando i dati variano nel tempo, secondo un certo tasso di incremento/decremento, o per calcolare l'incremento/decremento medio per u.d.t.

Media quadratica: definita come la radice quadratica della media aritmetica dei quadrati delle modalità di un carattere quantitativo.

$$\mu_q = \sqrt{\frac{\sum_{j=1}^{N} x_j^2}{N}}, \quad \text{dati disaggregati}$$

$$\mu_q = \sqrt{\frac{\sum_{j=1}^{N} x_j^2 n_j}{\sum_{j=1}^{N} n_j}}, \quad \text{dati organizzati in frequenze}$$

$$\mu_q = \sqrt{\frac{\sum_{j=1}^{N} \bar{x}_j^2}{N}}, \quad \text{media per dati organizzati in classi, dove}$$

$$\bar{x}_j = \frac{c_j + c_{j-1}}{2},\ c_{j-1}, c_j \text{ estremi della classe } \forall\, j = 1, \ldots, N$$

Si parla di *medie di potenze,* in quanto tutti i tipi di media analitica visti finora posso essere scritti nella forma seguente:

$$\mu_t = \sqrt[t]{\frac{\sum_{j=1}^{N} x_j^t}{N}},$$

dove:

$$\mu_t = \begin{cases} \mu, & \text{se } t = 1 \\ \mu_a, & \text{se } t = -1 \\ \mu_g, & \text{se } t \to 0 \\ \mu_q, & \text{se } t = 2 \end{cases}.$$

Proposizione 4.1. *Le medie analitiche soddisfano la seguente catena di diseguaglianze:*

$$\mu_a \leq \mu_g \leq \mu \leq \mu_q \ .$$

Proof. 1. Iniziamo col verificare che $\mu_g \leq \mu$, ossia $\sqrt[n]{x_1 \dots x_n} \leq \frac{1}{n}\sum_{i=1}^{n} x_i$. Supponiamo per iniziare $n = 2$, allora:

$$\begin{aligned} 0 \leq (x_1 - x_2)^2 &= x_1^2 + x_2^2 - 2x_1x_2 \\ &\Rightarrow\ 4x_1x_2 \leq x_1^2 + x_2^2 + 2x_1x_2 = (x_1 + x_2)^2 \\ &\Rightarrow\ x_1x_2 \leq \left(\frac{x_1 + x_2}{2}\right)^2 \ . \end{aligned}$$

Analogamente, se $n = 4$:

$$\begin{aligned} x_1x_2 \leq \left(\frac{x_1 + x_2}{2}\right)^2 \ , \quad x_3x_4 \leq \left(\frac{x_3 + x_4}{2}\right)^2 \\ \Rightarrow x_1x_2x_3x_4 \leq \left(\frac{x_1 + x_2}{2}\,\frac{x_3 + x_4}{2}\right)^2 \ . \end{aligned}$$

In particolare, la disuguaglianza per $n = 2$ vale per $(x_1 + x_2)/2$ e $(x_3 + x_4)/2$, dunque

$$\begin{aligned} &\frac{(x_1 + x_2)}{2}\frac{(x_3 + x_4)}{2} \leq \left(\frac{x_1 + x_2 + x_3 + x_4}{4}\right)^2 \\ &\Rightarrow\ x_1x_2x_3x_4 \leq \left(\frac{(x_1 + x_2)}{2}\frac{(x_3 + x_4)}{2}\right)^2 \leq \left(\frac{x_1 + x_2 + x_3 + x_4}{4}\right)^4 \\ &\Rightarrow\ \sqrt[4]{x_1x_2x_3x_4} \leq \frac{x_1 + x_2 + x_3 + x_4}{4} \ . \end{aligned}$$

Questo ragionamento si applica a tutti gli $n = 2^k$, $k \geq 1$. Infine, applicando il principio di induzione a ritroso, si ottiene:

$$\begin{aligned} &A := \frac{x_1 + \dots + x_{n-1}}{n-1} \ \Rightarrow\ x_1 + \dots + x_{n-1} = (n-1)A \\ &x_1 \cdot x_2 \dots x_{n-1} \cdot A \leq \left(\frac{x_1 + \dots + x_{n-1} + A}{n}\right)^n \\ &= \left(\frac{(n-1)A + A}{n}\right)^n = A^n \\ &\Rightarrow\ x_1 \cdot x_2 \dots x_{n-1} \leq A^{n-1} \ . \end{aligned}$$

2. $\mu_a \leq \mu_g \Leftrightarrow \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \leq \sqrt[n]{\prod_{i=1}^{n} x_i}$.

   La disuguaglianza del punto precedente vale per ogni x_i, dunque anche per $1/x_i$, allora:

$$\frac{1}{\sqrt[n]{x_1 \cdots x_n}} = \sqrt[n]{\frac{1}{x_1} \cdots \frac{1}{x_n}} \leq \frac{\frac{1}{x_1} + \cdots \frac{1}{x_n}}{n}$$
$$\Rightarrow \frac{n}{\frac{1}{x_1} + \cdots \frac{1}{x_n}} \leq \sqrt[n]{x_1 \cdots x_n} \, .$$

3. $\mu \leq \mu_q \Leftrightarrow \frac{1}{n} \sum_{i=1}^{n} x_i \leq \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2}$. Consideriamo il caso $n = 2$, la proprietà si generalizza facilmente al caso $n > 2$:

$$\frac{x_1^2 + x_2^2}{2} = \frac{(x_1 + x_2)^2 - 2x_1x_2}{2} = \frac{(x_1 + x_2)^2}{2} - x_1x_2$$
$$\geq \frac{(x_1 + x_2)^2}{2} \geq \frac{(x_1 + x_2)^2}{4} = \left(\frac{x_1 + x_2}{2}\right)^2$$
$$\Rightarrow \frac{x_1 + x_2}{2} \leq \sqrt{\frac{x_1^2 + x_2^2}{2}} \, .$$

□

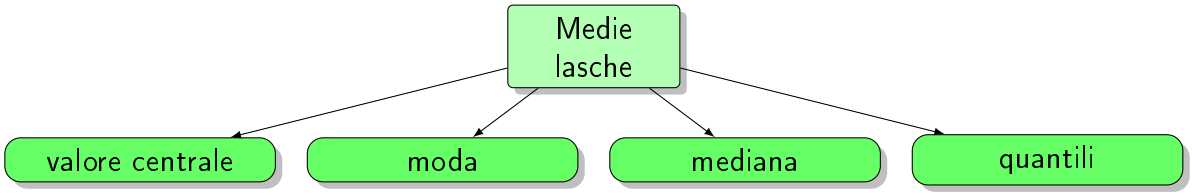
Infine, osserviamo che, nel calcolo delle medie analitiche semplici (dati disaggregati), tutte le modalità e le unità statistiche hanno la stessa importanza, o , equivalentemente, ciascuna modalità ha un *peso* pari a $1/n$, nel determinare il valore μ_t.

Le modalità di un carattere possono avere una diversa importanza: in questi casi, ciascuna di esse viene moltiplicata per una quantità (peso) che ne misura l'importanza. Le medie analitiche ottenute grazie a questi valori pesati sono dette *medie analitiche ponderate*

$$\mu_t^{\omega} = \sqrt[t]{\frac{\sum_{j=1}^{N} x_j^t \omega_j}{\sum_{j=1}^{N} \omega_j}} \, .$$

Quando $\omega_j = n_j$, per ogni j, otteniamo le espressioni precedenti, nel caso di dati organizzati in frequenze.

Medie lasche: tengono conto solo di alcuni valori della distribuzione.



Valore centrale: semisomma dei valori estremi, i.e., del valore più piccolo e del valore più grande osservati, ottenuti previo ordinamento dei dati. Consideriamo le modalità $x_1, \ldots, x_N$ ed ordiniamole in modo crescente: $x_{(1)}, \ldots, x_{(N)}$, allora

$$VC = \frac{x_{(1)} + x_{(N)}}{2} .$$

Moda: corrisponde alla modalità con la frequenza assoluta (relativa) più alta.

Per esempio, dato un collettivo di 10 unità statistiche, si consideri la seguente serie di osservazioni: $\{1, 2, 3, 4, 4, 4, 4, 3, 4, 1\}$.

La moda, indicata con Mo, risulta pari a 4, dal momento che la modalità 4 è presente cinque volte nel collettivo.

Cosa succede se i dati sono raggruppati in classi? Se le classi sono equi-ampie, si fa riferimento alla frequenza relativa di ciascuna classe. Se le classi hanno ampiezze diverse, si fa riferimento alla densità di frequenza di ciascuna classe. In quest'ultimo caso, la moda si definisce come la classe di modalità con massima densità di frequenza.

In entrambi i casi, non parleremo di moda, ma di *classe modale.*

Vale la pena di sottolineare che la moda di una distribuzione non è unica. Distingueremo tra *distribuzioni unimodali,* intese come distribuzioni di frequenza che hanno un solo punto di massimo (che rappresenta sia il massimo relativo che il massimo assoluto della distribuzione) e *distribuzioni bimodali o $k-$modali,* ossia, distribuzioni di frequenza che presentano due o k mode, che hanno due o k massimi relativi.

Se tutte le modalità hanno la stessa frequenza, allora si parla di classe zeromodale.

Infine, la moda viene utilizzata solamente a scopi descrittivi, perchè è meno stabile e meno oggettiva delle altre medie lasche.

Mediana: corrisponde alla modalità osservata sulla unità statistica centrale nella distribuzione ordinata delle osservazioni.

Se il carattere è quantitativo discreto, allora

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

Esempio 4.1. *Per un collettivo di* 15 *unità, si consideri*

$$\{29, 7, 18, 15, 27, 23, 14, 1, 25, 13, 18, 24, 28, 22, 5\}.$$

Le osservazioni ordinate sono

$$\{1, 5, 7, 13, 14, 15, 18, 18, 22, 23, 24, 25, 27, 28, 29\}.$$

Dato che $N = 15$ *è dispari, la mediana sarà* $Me = x_{\left(\frac{15+1}{2}\right)} = x_{(8)} = 18$.

Esempio 4.2. *Per un collettivo di* 12 *unità, si consideri*

$$\{34, 42, 1, 34, 19, 42, 25, 35, 21, 15, 9, 10\}.$$

Le osservazioni ordinate sono

$$\{1, 9, 10, 15, 19, 21, 25, 34, 34, 35, 42, 42\}.$$

Dato che $N = 12$ *è pari, la mediana sarà* $Me = \frac{x_{\left(\frac{12}{2}\right)} + x_{\left(\frac{12}{2}+1\right)}}{2} = \frac{x_{(6)} + x_{(7)}}{2} = \frac{21+25}{2} = 23$.

Se i dati del carattere X discreto sono noti mediante una distribuzione di frequenze, allora l'individuazione della mediana avviene attraverso l'uso della funzione di ripartizione $F(x)$ (supponendo di aver ordinato le modalità di X in modo crescente).

Più, precisamente, la mediana sarà la modalità x_j tale che

$$\begin{aligned} F(x_{(j-1)}) &< 0.5 \\ F(x_{(j)}) &\geq 0.5 \end{aligned}$$

Nel caso di dati raggruppati in classi:

1. si individua la *classe mediana,* ossia la classe (c_{j-1}, c_j) che ha funzione di ripartizione $F(x_j) \geq 0.5$
2. si calcola la mediana all'interno di tale classe:

$$Me = c_{j-1} + \frac{0.5 - F_{j-1}}{f_j}(c_j - c_{j-1}) \ .$$

Quantili: costituiscono una famiglia di misure che si distinguono a seconda del numero di parti uguali in cui suddividono una distribuzione.

Si definisce *quantile di ordine* $\alpha \in (0,1)$ quel numero che divide l'insieme delle osservazioni in due gruppi, lasciando a sinistra l'$\alpha \times 100$ delle osservazioni più piccole del quantile e a destra l'$(1-\alpha) \times 100$ delle osservazioni più grandi.

Per esempio, il *primo quartile* Q_1 corrisponde alla modalità assunta dall'unità statistica, il 25% delle quali presenta valori ad essa inferiori. Il *secondo quartile* Q_2 coincide con la mediana, mentre il *terzo quartile* Q_3 corrisponde alla modalità assunta dall'unità statistica, il 75% delle quali presenta valori ad essa inferiori.

I *decili* ripartiscono la graduatoria non decrescente in dieci gruppi, dunque le soglie saranno $10\%, 20\%, 30\%, 40\%, \ldots$

I *percentili* son generalizzazione dell'indice di posizione a qualunque percentuale della distribuzione.

Come si calcolano i quantili di una distribuzione?

- ordinare le modalità in modo crescente
- calcolare $i = \left(\frac{\alpha}{100}\right) \cdot N$, dove α è il *percentile di interesse* e N il numero di modalità
- se i è un intero, il valore corrispondente ad α è la media tra la posizione i e la posizione $i+1$
- se i non è un intero, arrotondare per eccesso ottenendo i^*. Il valore di interesse è quello corrispondente alla posizione i^*.

Esempio 4.3. *Consideriamo un collettivo di* 15 *unità statistiche con le seguenti osservazioni (non ordinate):*

$$\{29, 7, 18, 15, 27, 23, 14, 1, 25, 13, 18, 24, 28, 22, 5\} \ .$$

Ordiniamo le osservazioni e otteniamo:

$$\{1, 5, 7, 13, 14, 15, 18, 18, 22, 23, 24, 25, 27, 28, 29\} .$$

Vogliamo calcolare i quartili, il 68*-simo ed il* 20*-simo percentile.*

I quartili si ottengono dividendo la distribuzione in quattro parti. Così, il primo, secondo e terzo quartile si calcolano ponendo $\alpha = 25$, $\alpha = 50$ *e* $\alpha = 75$, *rispettivamente, con* $n = 15$.

In particolare, Q_1 *coincide con l'elemento* i*-simo della serie di osservazioni ordinate, dove*

$$i = \frac{25}{100} \cdot 15 = 3.75 \Rightarrow i^* = 4 ,$$

dunque $Q_1 = X_{(4)} = 13$. *Analogamente per gli altri quartili,*

$$Q_2 = X_{(8)} = 18, \ \textit{essendo } i = \frac{50}{100} \cdot 15 = 7.5 \Rightarrow i^* = 8$$
$$Q_3 = X_{(12)} = 25, \ \textit{essendo } i = \frac{75}{100} \cdot 15 = 11.25 \Rightarrow i^* = 12$$
$$P_{68} = X_{(11)} = 24, \ \textit{essendo } i = \frac{68}{100} \cdot 15 = 10.2 \Rightarrow i^* = 11$$
$$P_{20} = (X_{(3)} + X_{(4)})/2 = 10, \ \textit{essendo } i = \frac{20}{100} \cdot 15 = 3 .$$