Primi concetti di statistica descrittiva

a cura di

Francesco Fabi

ce3s

CENTRO STUDI STATISTICI E SOCIALI

Statistica impostazioni

- Esistono metodologie utilizzate per sintetizzare e comunicare i risultati osservati:
- Statistica descrittiva o esplorativa.
- Per l'interpretazione dei risultati e l'estensione dall'osservato all'osservabile esistono altre metodologie appropriate:
- Statistica inferenziale o inferenza statistica.

Dall'analisi esplorativa dei dati all'inferenza

Analisi esplorativa dei dati		*Inferenza statistica*
L'obiettivo consiste nell'esplorare i dati senza precise motivazioni, cercando di trovare dei modelli interessanti.	⟷	L'obiettivo è rispondere a delle domande specifiche poste prima di osservare i dati.
E' possibile arrivare a delle conclusioni solo per le unità di cui si hanno i dati.	⟷	E' possibile estendere le conclusioni anche a un gruppo più grande di unità.
Le conclusioni sono informali basate su ciò che si osserva nei dati.	⟷	Le conclusioni sono formali e sono accompagnate da un'affermazione sulla confidenza che abbiamo in esse.

STATISTICA ESPLORATIVA

Come si procede

Fenomeni collettivi

- I metodi statistici permettono di passare da considerazioni ***qualitative*** a considerazioni ***quantitative nello studio di fenomeni collettivi***:
- I fenomeni collettivi sono quei fenomeni riferibili ad una moltitudine di oggetti in cui interessi studiare l'insieme degli oggetti nel suo complesso e non i singoli individui nei quali il fenomeno si manifesta secondo caratteristiche individuali (*fenomeno individuale*).

Il fumo della madre e la salute del neonato

- Una delle raccomandazioni mediche che appaiono sui pacchetti di sigarette negli Stati Uniti dice che il fumo in gravidanza può provocare danni al feto, nascita prematura, e peso basso alla nascita.

- Su che cosa si basano tali raccomandazioni?

Gli studi statistici

- Alla base c'è uno studio statistico in cui sono state osservate nei neonati le caratteristiche su cui si voleva indagare e messe in relazione (statistica) con il fumo in gravidanza. I dati raccolti sono stati analizzati, sintetizzati e presentati per trarre conclusioni (con un certo margine di errore statistico).

  (J.Yerushalmy. The relationship of parent's cigarette smoking to outcome of pregnancy-implications as to the problem of inferring causation from observed associations. *Am. J. Epidemiol., 93, 1971).*

Alcuni dati dallo studio sul fumo

- Consideriamo un sottoinsieme dei dati dello studio condotto sulle donne in gravidanza tra il 1960 e il 1967 a San Francisco. Allo studio hanno partecipato 15000 famiglie con un livello di studio e di reddito medio-alto.
- Diverse caratteristiche del bambino venivano registrate alla nascita, insieme all'informazione sulle abitudini al fumo della madre.
- Consideriamo il peso alla nascita per 1236 maschi, nati tra il 1960 e il 1961, e che sono sopravissuti almeno 28 giorni.

Caratteristica	Descrizione
Peso alla nascita	Peso alla nascita in once (0,035 once=1gr)
Abitudine al fumo	Indicatore dell'abitudine al fumo in gravidanza. Fumo si (1), no (0)

Lo studio - Come procedere?

- Come possiamo registrare i dati raccolti?
- Come possiamo analizzare i dati e sintetizzare i risultati?
- Come possiamo presentarli?
- Come possiamo interpretarli?
- Le eventuali differenze osservate possono dirsi sistematiche?

Le matrice codificata dei dati

unit	smoke	birth weight
1	0	120
2	0	113
3	1	128
4	0	123
5	1	108
6	0	136
7	0	138
8	0	132

Unità statistiche e variabili

Le **unità statistiche** sono gli oggetti descritti tramite un insieme di dati. Le unità statistiche possono essere persone, ma anche animali o cose.

Una **variabile** è qualsiasi caratteristica associata a un'unità. Una variabile in generale assume valori diversi su unità statistiche diverse.

Variabili categoriche e quantitative

Una **variabile categorica** colloca un'unità in una tra diverse categorie (o modalità).

Una **variabile quantitativa** assume valori numerici che misurano, in opportune unità di misura, le caratteristiche per ogni unità.

La distribuzione dell'abitudine al fumo

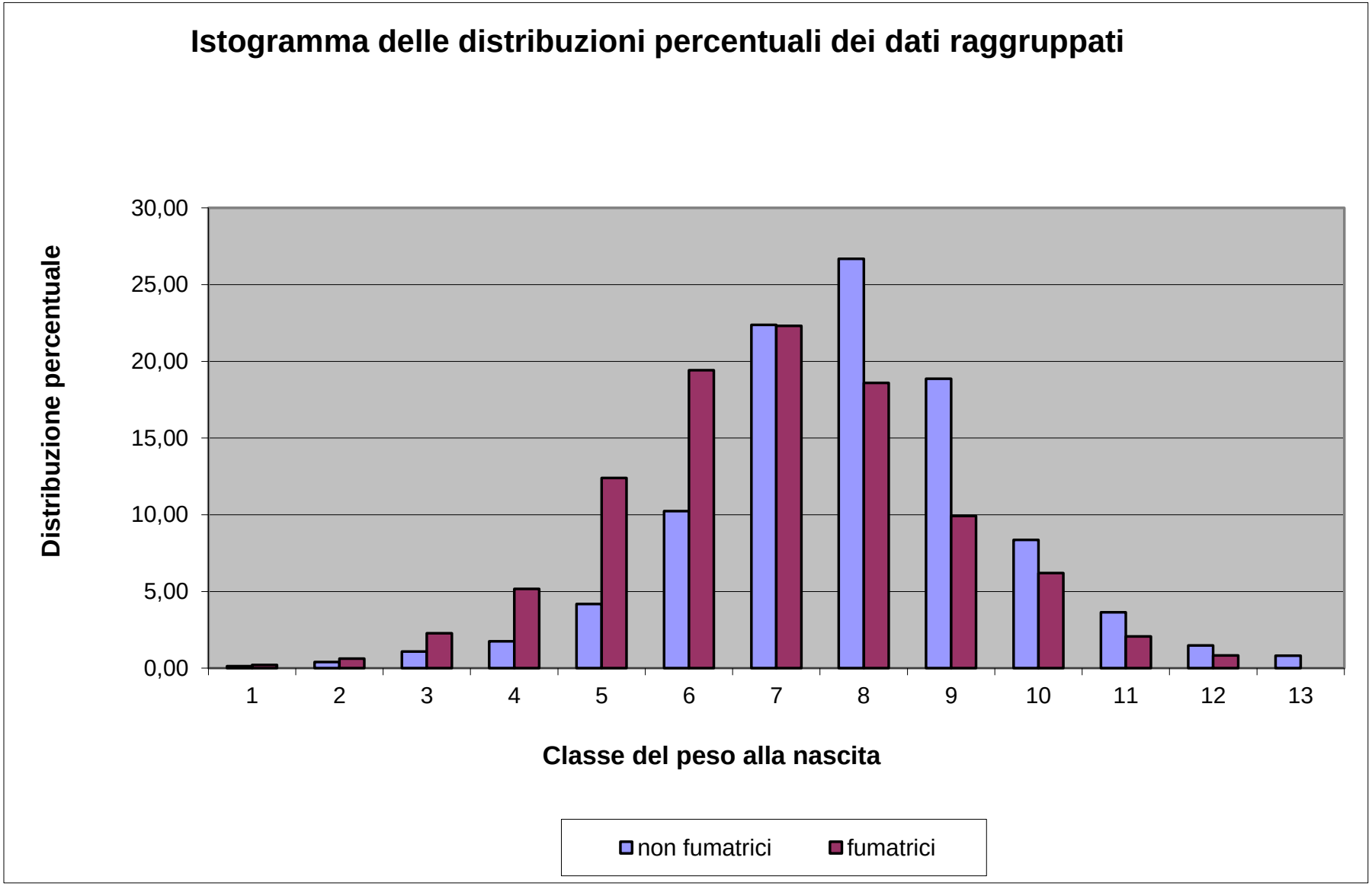
Possiamo sintetizzare i dati qualitativi attraverso semplici conteggi delle unità che presentano una stessa modalità purché l'ordine in cui sono registrate le unità stesse non sia influente sul fenomeno in studio. Si ottiene così una tabella sintetica di frequenze.

	Distribuzione statistica dell'abitudine al fumo		
Valore	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
1	742	0,60	60,03
2	484	0,39	39,16
3	10	0,01	0,81
Totale	1236	1,00	100,00
	1=non fumatrice, 2=fumatrice, 3=missing		

Frequenza

- frequenza assoluta corrispondente ad una certa modalità (valore) il numero di unità statistiche che presenta tale modalità,
- frequenza relativa corrispondente ad una certa modalità il rapporto tra il numero di unità statistiche che presenta tale modalità e il totale delle unità statistiche considerate,
- frequenza percentuale è la frequenza relativa moltiplicata per 100.

Rappresentazione grafica



Distribuzione di frequenza

- La distribuzione statistica fornisce un modo compatto di rappresentazione dei dati che così risultano più organizzati e dunque più leggibili. Nella tabella ad ogni modalità della variabile è associata la sua frequenza.
- Nella distribuzione relativa o percentuale non compare il numero di unità statistiche considerate, occorre allora fornire tale dato nella descrizione della rilevazione.
- L'informazione data dalle frequenze percentuali calcolate su un campione di 10000 unità, infatti, è ben diversa da quella ottenuta su un campione di 50 unità.

Analizzare una distribuzione

In qualsiasi distribuzione è importante individuare il miglior **modello interpretativo** e le **deviazioni** evidenti rispetto a tale modello.

È possibile individuare il modello di un istogramma attraverso la **forma**, il **centro** e la **dispersione**.

Gli Indici Statistici

Per confrontare tra loro due o più popolazioni, o le variazioni della stessa popolazione in periodi diversi, non basta l'aver raccolto i dati e averli sintetizzati in tabelle di frequenza, è spesso necessario sintetizzare ulteriormente la distribuzione in un solo valore attorno a cui i dati si "addensano" e sapere in che misura ciò accade, ovvero studiare la variabilità dei dati.

Se non ci fosse variabilità all'interno di una popolazione, non ci sarebbe bisogno della statistica. Una singola unità sarebbe sufficiente a descrivere l'intera popolazione.

Misure di tendenza centrale

- Rappresentano ***i valori attorno a cui i dati tendono ad aggregarsi (indici di posizione)***. Le più diffuse sono:
  - **moda** (anche per caratteri qualitativi non ordinabili);
  - **mediana** e **quantili** (almeno caratteri qualitativi ordinabili);
  - **media** (solo caratteri quantitativi).

La moda o modalità prevalente

- **Dal dizionario**: ***"Usanza più o meno mutevole secondo il gusto prevalente, che si impone nelle abitudini, nel modo di vivere e specialmente nelle forme del vestire".*** Dire ad esempio che ora è di moda, per gli uomini, portare i capelli lunghi, significa nient'altro che nella distribuzione di frequenza degli uomini secondo la lunghezza dei capelli, alla modalità *lunghi* corrisponde la massima frequenza.
- La ***moda di un collettivo***, distribuito secondo un carattere, è ***la modalità prevalente del carattere*** ossia ***quella a cui è associata la massima frequenza***.
- ***La moda è utilizzabile per tutti i tipi di caratteri organizzati in distribuzioni di frequenza.***

La mediana

La ***mediana*** (**Me**) è quel valore della **variabile ordinabile** o quantitativa che, nella successione di valori osservati, disposti in ordine crescente o decrescente, **occupa la posizione centrale**; ovvero il numero delle unità che possiedono il carattere in quantità inferiore alla **mediana** è uguale al numero di quelle che possiedono il carattere in quantità superiore alla **mediana**.

Per variabili quantitative:

- Se **n** è dispari si ha **una sola mediana**, ed è il valore corrispondente all'**unità $(n+1)_{/2}$** (nella distr. ordinata);
- Se **n** è pari si hanno **2 valori mediani** in corrispondenza delle osservazioni **$n_{/2}$** e **$(n_{/2}+1)$**. In questo caso, la **mediana** è per convenzione la media aritmetica dei due valori.

Troviamo la mediana

- Per le madri non fumatrici, che sono 742, il peso alla nascita ha due valori mediani, in corrispondenza delle posizioni: 371 e 372, entrambi i valori sono di 123 once.
- Per le madri fumatrici, che sono 484, i due valori corrispondono alle posizioni: 242 e 243 e entrambi sono di 115 once.

La media aritmetica

- **La media** di un insieme di n misure è data da

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

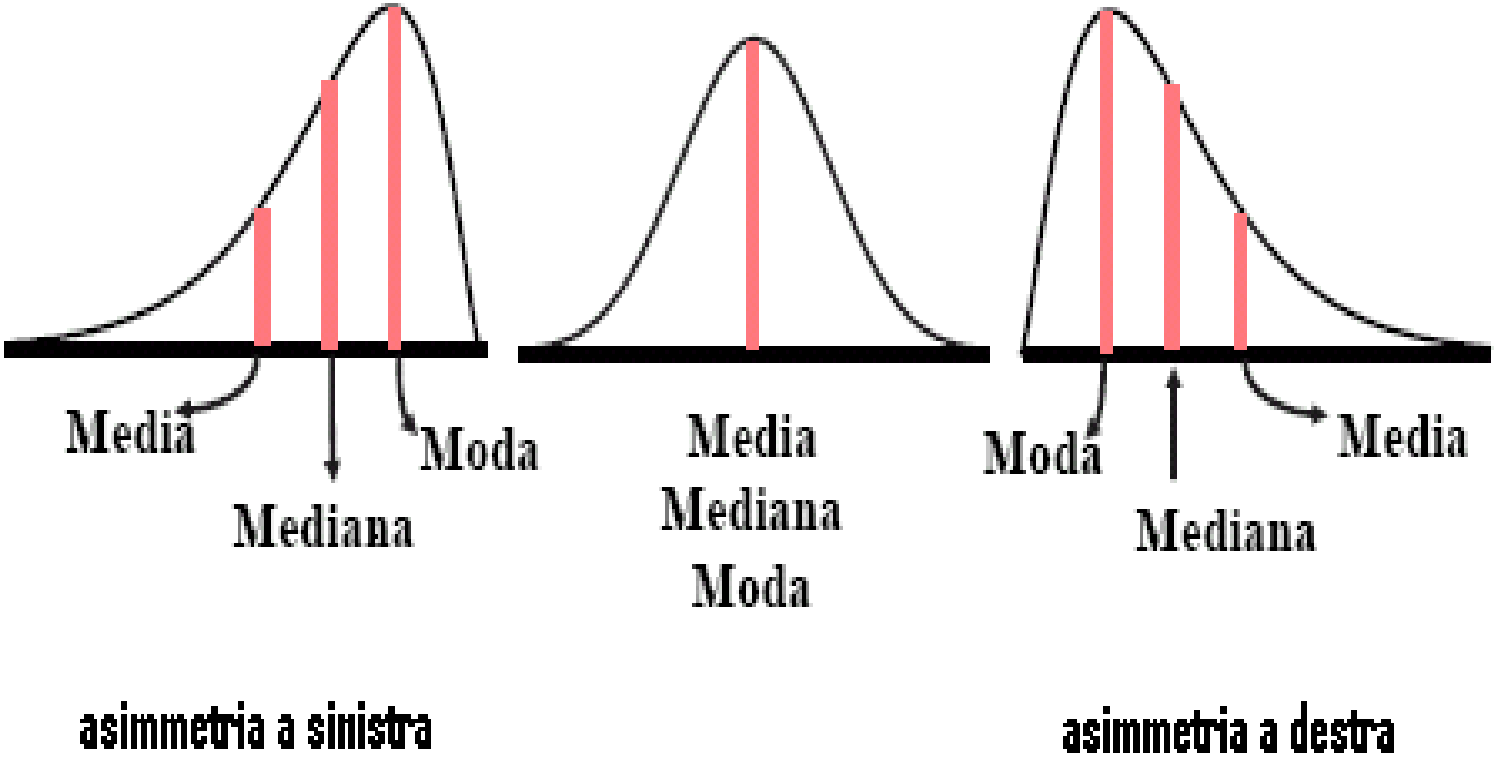
se i dati sono sintetizzati in una distribuzione di frequenze, cioè il valore x_j compare con la frequenza assoluta f_j (j = 1, 2, ..., s) si può usare la formula sotto (proprietà associativa dell'addizione):

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_s f_s}{f_1 + f_2 + \ldots + f_s} = \frac{\sum_{i=1}^{s} x_i f_i}{n}$$

Calcoliamo la media

- La media del peso alla nascita per le madri non fumatrici è 123,05 once
- La media del peso alla nascita per le madri fumatrici è 114,11 once
- La media non è uno dei valori osservati.

Relazioni tra media mediana e moda



Media, mediana e moda dei pesi in grammi

	NF	F
moda	3685.714	3285.714
mediana	3514.286	3285.714
media	3515.714	3260.286

Indici di dispersione o variabilità

- Gli **indici di posizione** (misure di tendenza centrale) dicono *attorno a quale valore le osservazioni sono centrate* e *sono tanto più significativi quanto più* ***i dati sono concentrati intorno ad essi***.
- Per ottenere un'informazione più accurata, è quindi necessario **misurare il grado di dispersione dei dati intorno a tali indici**. Questo è possibile, soltanto per i caratteri quantitativi, associando alle misure di tendenza centrale delle **misure di dispersione o variabilità.**

Misure di variabilità

- Il range o intervallo di variazione, che rappresenta l'intervallo tra il minimo e il massimo valore osservato.
- I percentili, sono quei valori che dividono la distribuzione in 100 parti di uguale numerosità.

Il 25-esimo, 50-esimo e 75-esimo percentile (ossia, primo quartile Q_1, mediana Q_2 e terzo quartile Q_3) dividono la distribuzione in 4 parti uguali.

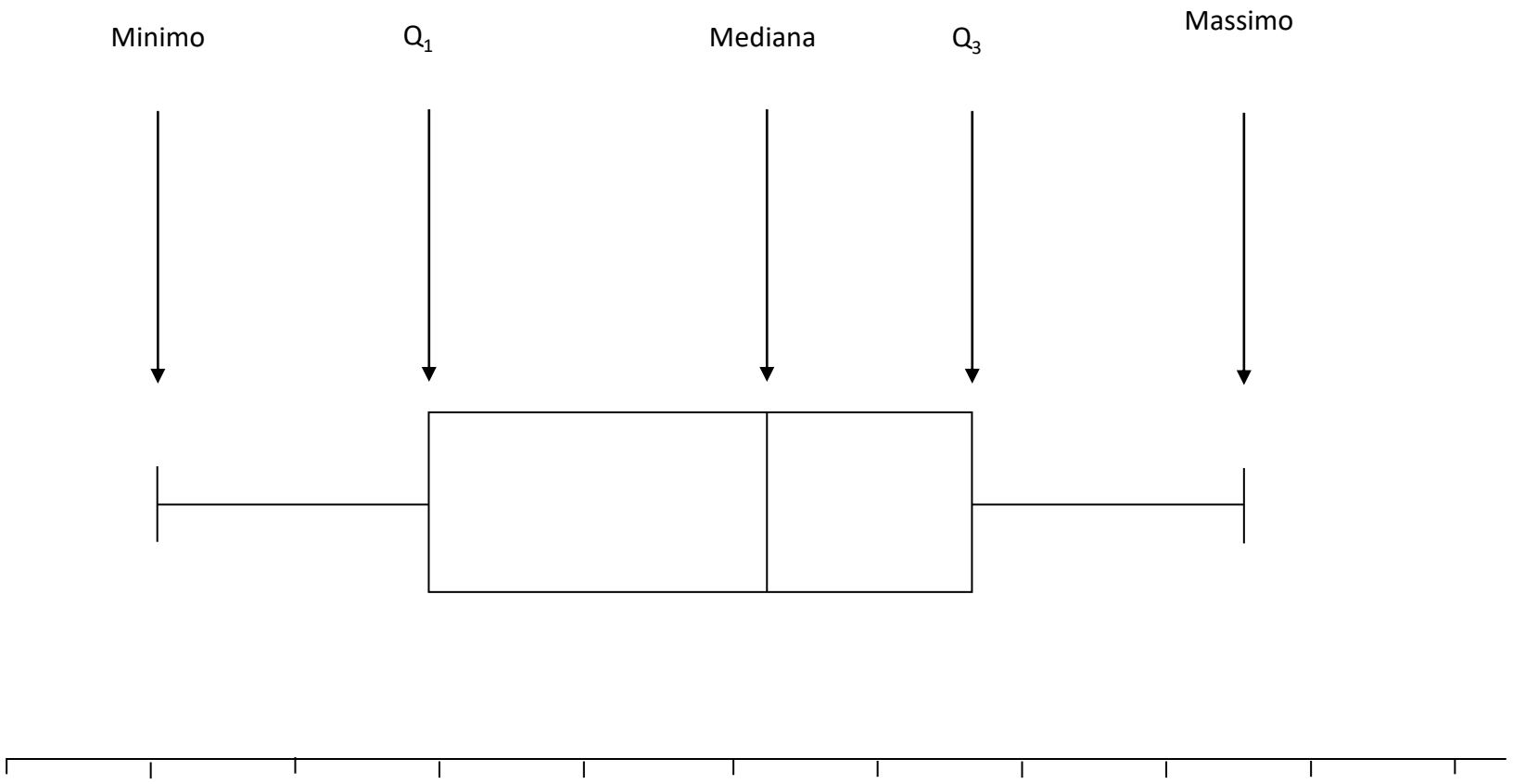
Per determinare i quartili possiamo considerare le due parti dei dati ottenute dal calcolo della mediana e trovare di ognuna la mediana.

Il primo e il terzo quartile individuano un intervallo centrale che contiene il 50% delle unità statistiche e che misura la dispersione dei valori centrali del collettivo osservato attorno alla mediana.

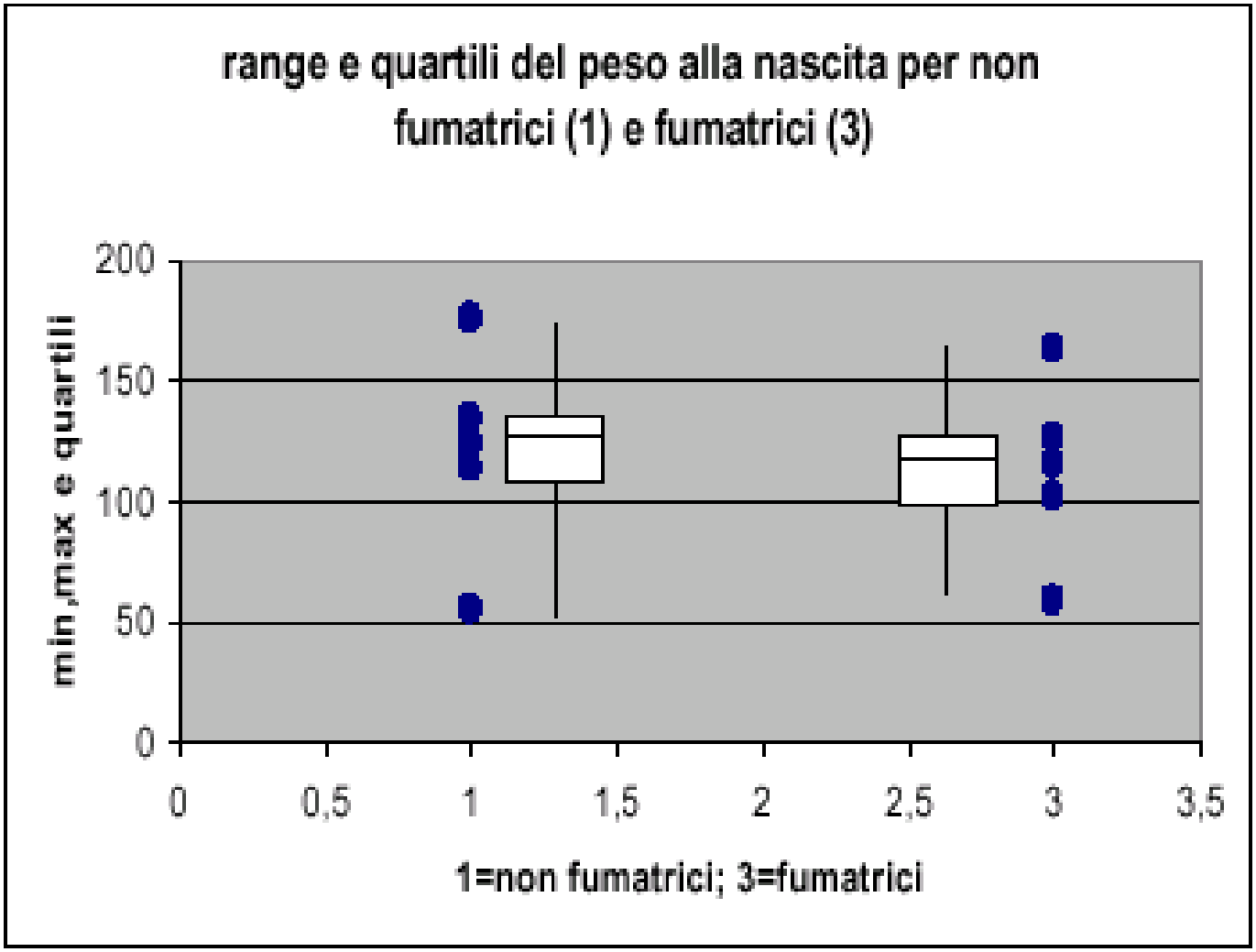
Rappresentazione grafica: il box plot

- Il ***Box-Plot*** rappresenta in modo compatto la distribuzione statistica attraverso alcuni indici sintetici: il ***range*** delle misure attraverso un *segmento verticale*, i ***3 quartili*** della distribuzione mediante un *rettangolo (box)*, tagliato internamente da un *segmento* in corrispondenza della ***mediana***, che contiene il 50% della distribuzione.
- La dimensione della base (o altezza) del rettangolo non rappresenta alcuna informazione, come pure la posizione del ***Box-Plot***, che può essere posto sia verticalmente che orizzontalmente.

Box plot



Rappresentazione grafica



La deviazione standard – il più importante indice di variabilità

- La Deviazione Standard è l'indice di dispersione più usato. E' importante, come lo è la media, perché è alla base di ulteriori analisi.
- Si calcola la differenza tra ogni valore e la media ($x_i - \bar{x}$), che si chiama variabile scarto. Si eleva al quadrato ogni differenza. Si sommano tali differenze al quadrato. Si divide la sommatoria ottenuta per il numero dei valori indipendenti nella variabile scarto. Il valore ottenuto si chiama varianza. Si estrae la radice quadrata della varianza per ottenere la Deviazione Standard.

La deviazione standard

La deviazione standard s

La **varianza** s^2 di un insieme di osservazioni è la media dei quadrati delle deviazioni delle osservazioni dalla loro media. In simboli, la varianza di n osservazioni $x_1, \ldots, x_n$ è

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

o, in forma più compatta,

$$s^2 = \frac{1}{n-1} \sum (x_i - \overline{x})^2$$

La **deviazione standard** s è la radice quadrata della varianza s^2:

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \overline{x})^2}$$