Metodi matematici e informatici per la biologia

Flavia Mascioli

A.A. 2015-2016

Informazioni generali

- Trovate le notizie riguardanti il corso sul sito:

  http://elearning2.uniroma1.it/

→Biologia→laurea triennale→I anno, II sem

→Corso Metodi Mat. e Inf. per la Biologia (informazioni per tutti i canali)

Informazioni generali

- **Laboratorio: Via Tiburtina 205, aula 17**

- **Open Office**

  www1.mat.uniroma1.it/people/giacomel/mmib/index.html

ESERCITATORE: Emanuela Giacomelli

Date Esami

Le Date d'esame RELATIVE AL SOLO MODULO DI MMIB (a. a. 2015-2016) le trovate sul sito di

elearning2.uniroma1.it

Metodi Matematici e Informatici per la Biologia
(informazioni per tutti i canali)

Obiettivi del corso

- Far capire l'importanza della statistica nel trattamento dei dati biologici
- Far capire la logica del ragionamento statistico
- Introdurre i principali elementi di:
  - disegno del metodo di raccolta dei dati
  - analisi esplorativa dei dati
- Familiarizzare lo studente con il particolare vocabolario della statistica
- Mostrare come l'uso di un software possa aiutare nei calcoli in modo semplice ed efficace

Perché la statistica in biologia?

- A causa della grande variabilità intrinseca al materiale biologico:

--variabilità genetica tra gli individui,

--variabilità dovuta alla crescita e allo sviluppo degli individui,

--variabilità delle reazioni di uno stesso individuo in momenti diversi, o delle misurazioni eseguite sullo stesso individuo in tempi diversi.

- Il concetto di variabilità è fondamentale nella sperimentazione scientifica.

- **La statistica aiuta nello studio quantitativo della variabilità, permettendo di capire, gestire e ridurre la variabilità.**

La variabilità

- I biologi considerano la **variabilità** come “la materia prima” dell’evoluzione: senza variabilità non esisterebbe nemmeno l’uomo.

Fattori casuali e variabilità individuale

- I motivi (fattori) che rendono ogni individuo **diverso** da ogni altro sono praticamente infiniti. La genetica, l'età il sesso, le condizioni di vita o di allevamento, l'alimentazione, il clima e un'infinità di altre variabili esercitano tutte sull'individuo un effetto grande o piccolo.
- Ovviamente, alcuni di questi fattori sono più importanti di altri; tuttavia, è sempre la **somma degli effetti di molti fattori diversi** che rende ogni individuo diverso dall'altro. L'insieme di tutti questi fattori (o cause) interferisce **imprevedibilmente** su un individuo. Si parla di fattori casuali, ossia **dovuti al caso.**

La (Bio)Statistica

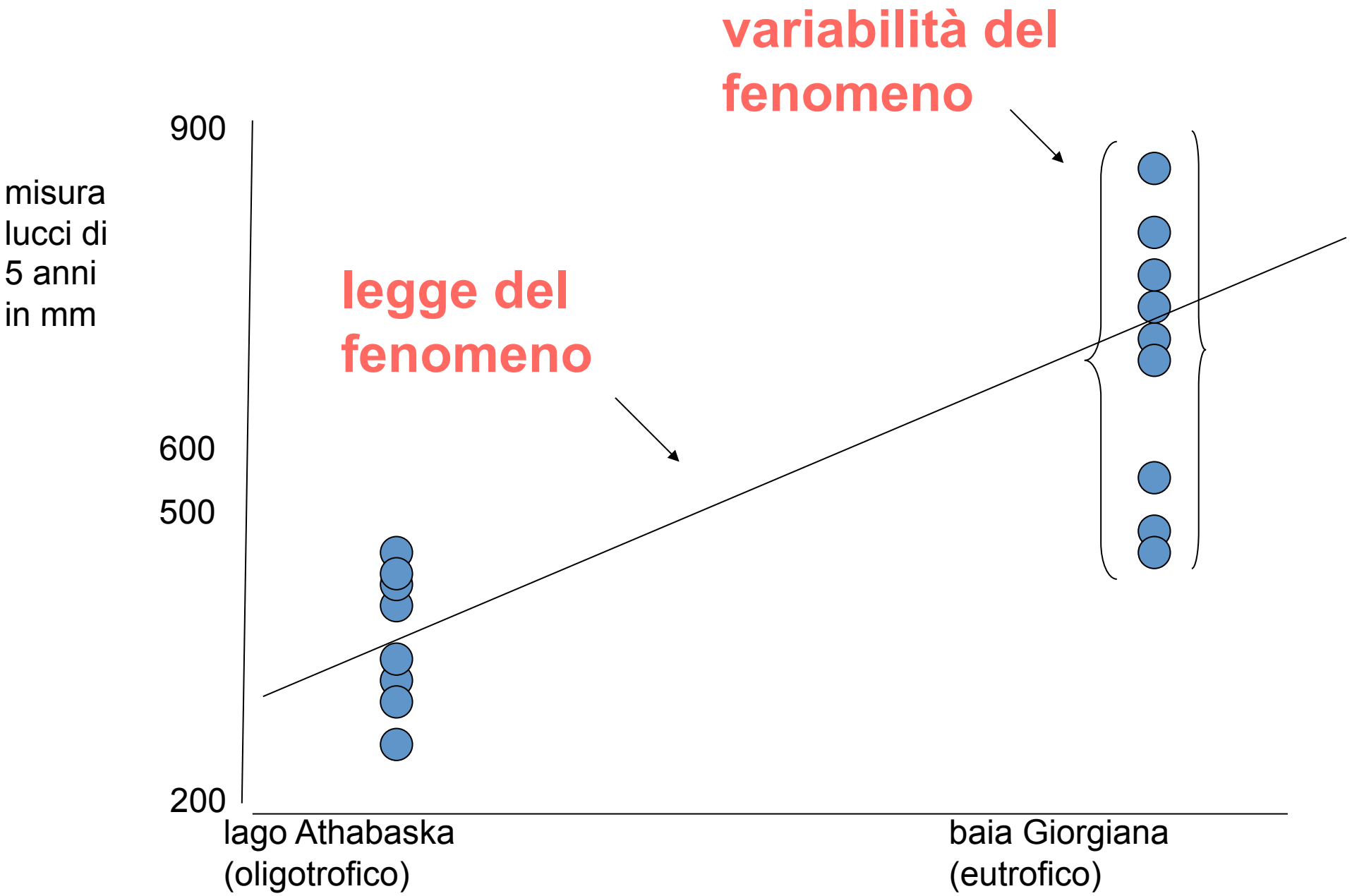
- La Biostatistica è l'applicazione **di metodi statistici** alla soluzione di problemi biologici

- La Statistica è lo studio scientifico dei **dati** che descrivono **la variabilità presente in natura**

- I dati possono essere i risultati di una misura sperimentale (la lunghezza di un organismo, o la quantità di una sostanza chimica nel sangue) o di un conteggio (il n° di ciglia in un organismo)

La variabilità: esempio

- Nell'esempio che segue si fa riferimento ad un articolo di Scott e Crossman, "Poisson d'eau douce du Canada", (1974), dove si confrontano le dimensioni dei lucci in 2 laghi canadesi.
- Si può studiare la **variabilità del fenomeno** ripetendo un esperimento o una misura nelle stesse condizioni.
- Si può studiare la **legge del fenomeno** facendo **variare** le condizioni dell'esperimento o dell'osservazione.


variabilità del fenomeno
legge del fenomeno
misura lucci di 5 anni in mm
900
600
500
200
lago Athabaska (oligotrofico)
baia Giorgiana (eutrofico)

La variabilità

Ad es. si supponga di effettuare un esperimento per controllare come funziona una dieta sui topi.

1) Osserveremo una variabilità "intrinseca", ossia i topi sono diversi geneticamente.
2) Osserveremo una variabilità "estrinseca", dovuta a una mancanza di uniformità nel condurre l'esperimento. Ossia, i topi di gabbie diverse sono soggetti a condizioni diverse (calore, luce, altri fattori).

2 cause principali di variabilità. La statistica aiuta a controllarle

La necessità dei metodi statistici

- L'esempio che segue illustra la necessità dei metodi statistici per l'analisi di dati quantitativi.

- Nella tabella appaiono i risultati di un esperimento per studiare gli effetti dell'irrigazione sulla crescita di piante di cavolo piantate a quattro diverse distanze (Mead, Curnow, Hasted, Technical Report, 1986).
- I valori che appaiono sono i pesi dei cavoli in kg.

Peso (kg) del raccolto di cavoli in 24 appezzamenti di terreno

irrigazione	distanza	Campo A	Campo B	Campo C
frequente	1 (45cm)	1.11	1.03	0.94
frequente	2 (40cm)	1.00	0.82	1.00
frequente	3 (35cm)	0.89	0.80	0.95
frequente	4 (25cm)	0.87	0.65	0.85
raramente	1 (45cm)	0.97	0.86	0.92
raramente	2 (40cm)	0.80	0.91	0.68
raramente	3 (35cm)	0.57	0.72	0.77
raramente	4 (25cm)	0.60	0.69	0.51

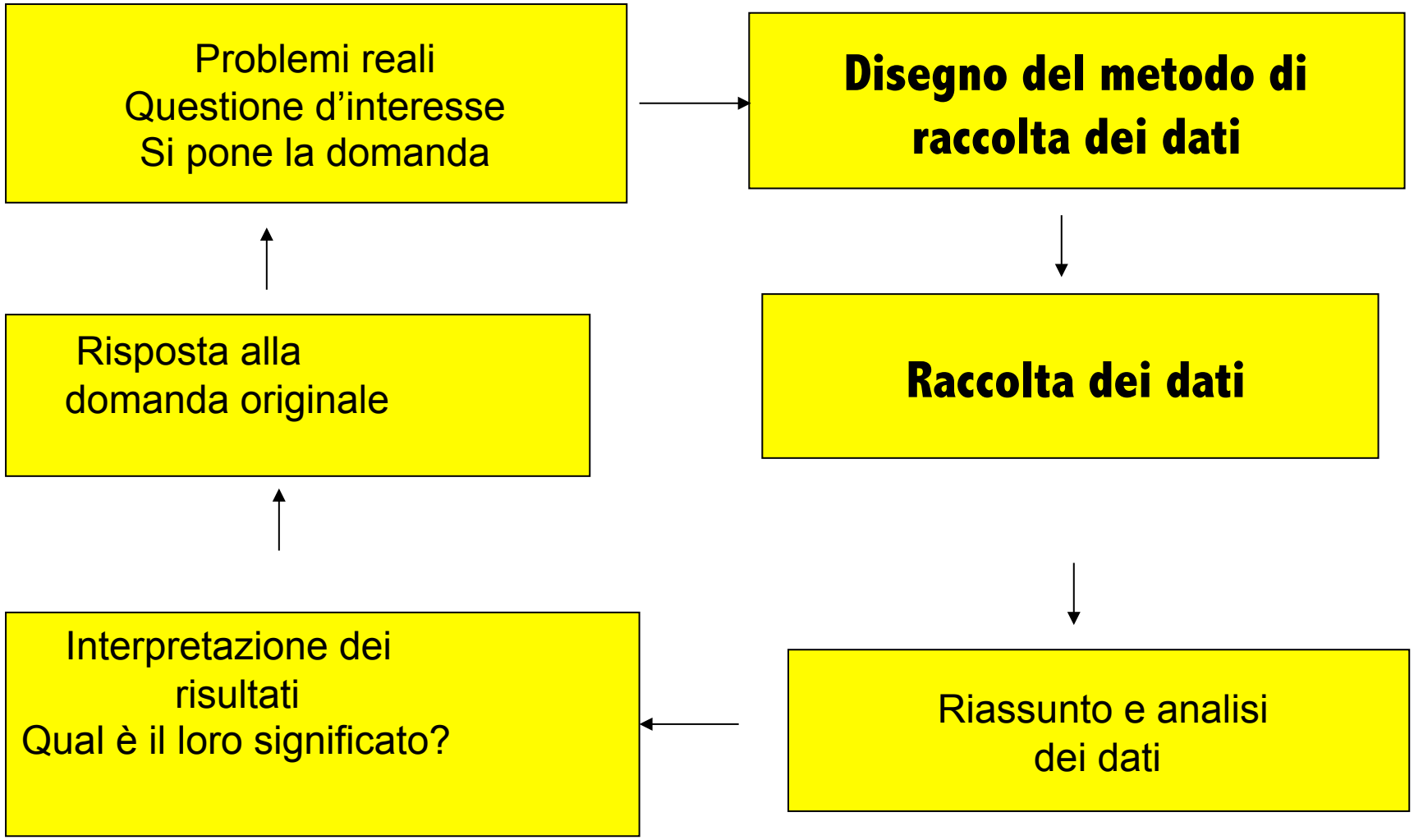
- Sono state provate tutte le 8 combinazioni di irrigazione e distanza fra piante, perché l'effetto dell'irrigazione può essere diverso per differenti distanze. Inoltre le 8 combinazioni sono state provate su 3 campi diversi. In questo modo si avranno informazioni sulla variabilità del materiale sperimentale, e si otterranno stime più precise sugli effetti dovuti ai trattamenti.
- Alcuni effetti importanti possono essere determinati osservando la tabella: l'irrigazione frequente produce cavoli più grandi e una distanza minore ne riduce la dimensione.

- Tuttavia, rimangono diverse domande a cui si vorrebbe rispondere e che necessitano di un'analisi statistica più elaborata.
- Di quanto aumenta il peso con irrigazioni frequenti e con distanze maggiori?
- C'è un relazione tra peso e distanza?
- E' possibile prevedere il peso per una distanza diversa da quelle considerate nell'esperimento?
- A parità di irrigazione e distanza c'è differenza tra i raccolti dei 3 campi?

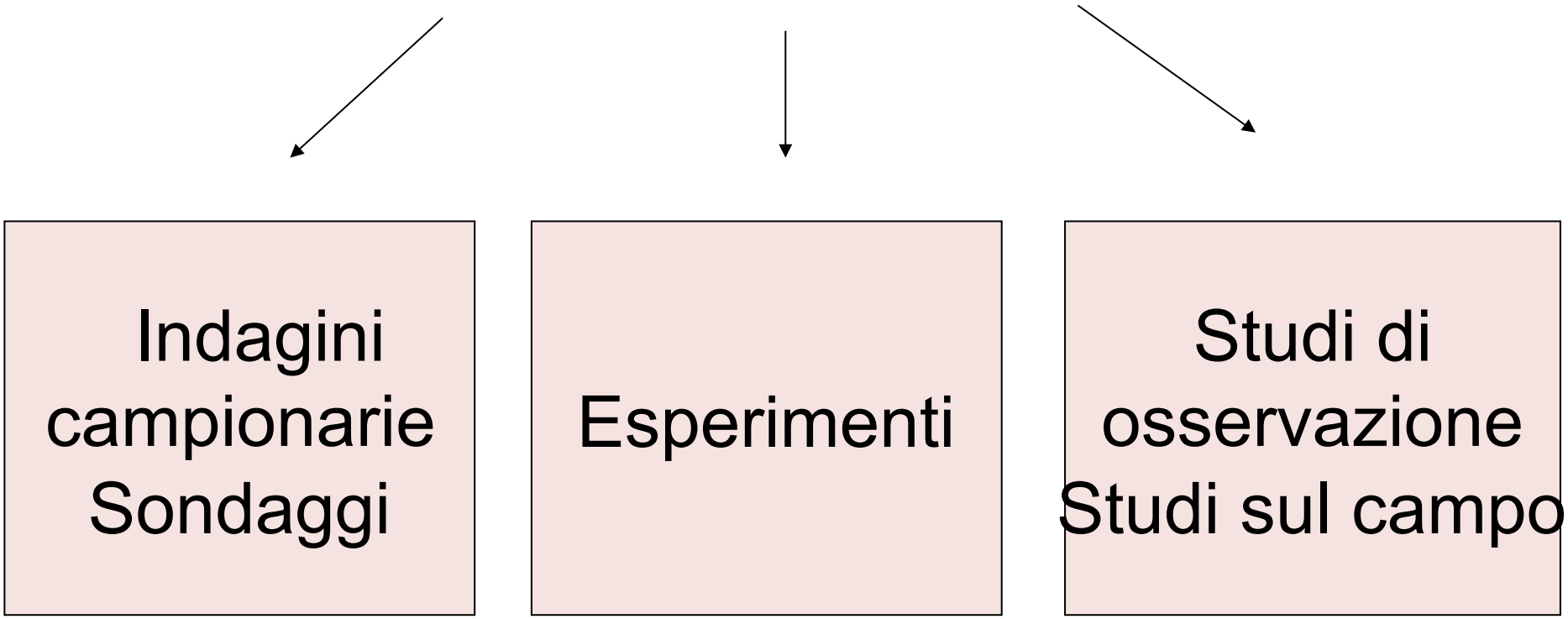
Passi principali del disegno di una ricerca in biologia

- Identificare gli scopi della ricerca.
- Pianificare la ricerca al fine di studiare il problema per il quale si cerca una risposta.
- Come ottenere i dati di cui si ha bisogno?
- Quale metodo statistico usare per analizzarli?
- Come interpretare i risultati?
- Come presentare i risultati?

Il ciclo di una ricerca scientifica e statistica

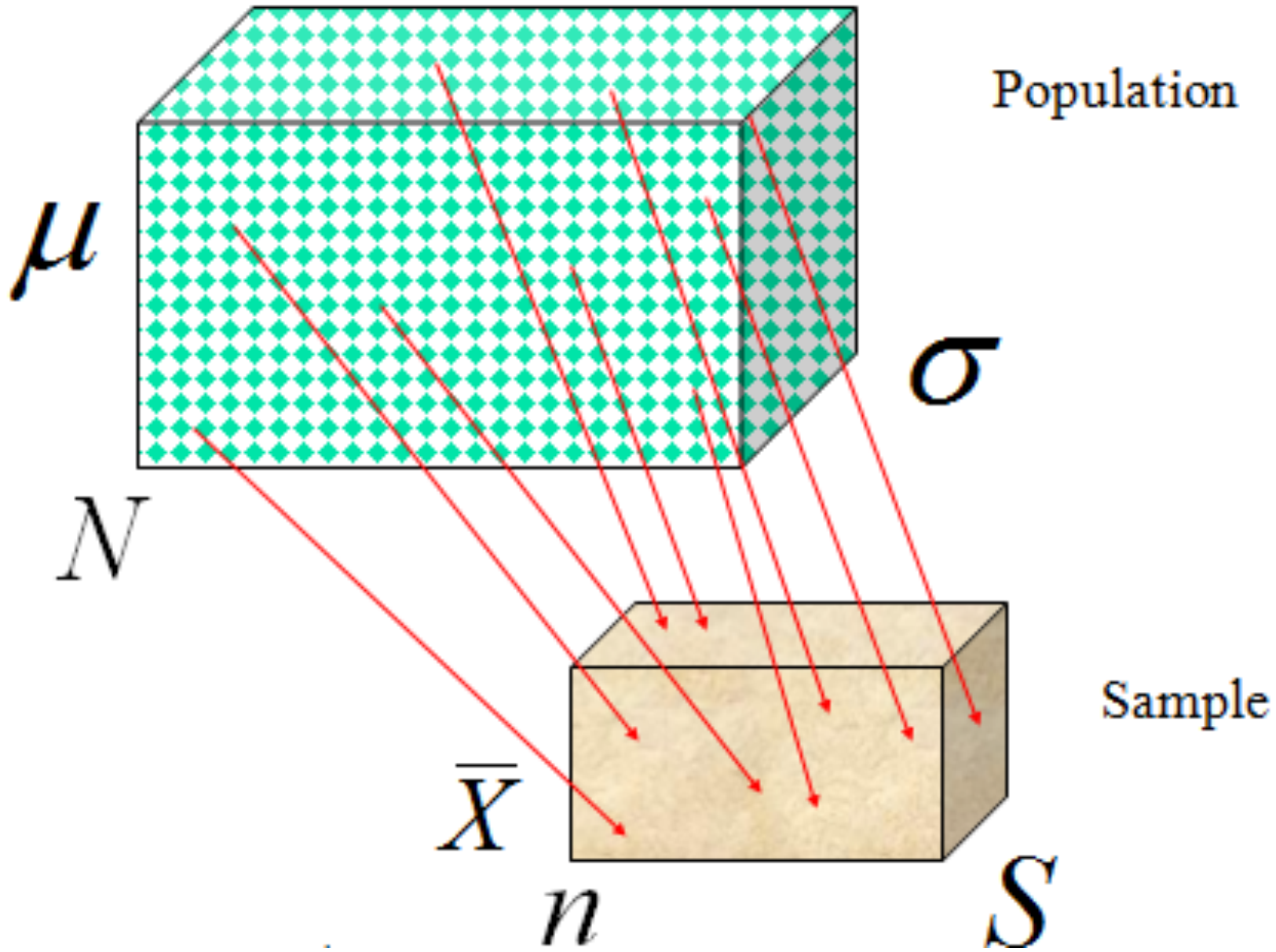


Quale metodo di raccolta dei dati?



Indagini campionarie, Sondaggi

- In un'indagine campionaria, o in un sondaggio si studiano i dati di **un campione** dalla popolazione per ottenere informazioni sull'intera popolazione.
- **La popolazione** è l'intero gruppo di unità sul quale vogliamo ottenere informazioni.
- **Un campione** è il sottoinsieme della popolazione che viene esaminato per ottenere le informazioni che interessano.


Population
μ
σ
N
Sample
$\bar{X}$
n
S

Campioni -Popolazioni

Supponiamo ad esempio di voler misurare la concentrazione di polveri sottili (PM10) nel comune di Padova, in un certo istante.

In questo caso la “popolazione” `e costituita dai valori del PM10 in tutti i punti della superficie comunale.

Scegliere un “campione” significa selezionare un certo numero di punti sulla superficie comunale, in cui piazzare dei rilevatori.

Campioni -Popolazioni

Una **popolazione statistica** può essere un insieme di persone o animali, o un insieme di misure o di osservazioni.

Popolazioni ipotetiche: es. tutte le possibili estrazioni di due carte da un mazzo;

Tutti i possibili valori della pressione sistolica letti in un paziente dallo stesso osservatore

– Popolazioni fisiche: bambini di V C di una certa scuola

– Popolazioni finite : tutti i caprioli che vivono nelle Alpi

– Popolazioni infinite: tutti i multipli di 13

Campioni -Popolazioni

- Campione:

un insieme di individui misurati o testati (10 cavie sulle quali si effettua un esperimento)

un insieme di diversi siti (per i quali si misura la concentrazione di nitrati, o la numerosità di fenicotteri)

- Popolazione:

infinita è formata da un n° **potenzialmente infinito** di unità statistiche che, in un certo istante, possono anche non esistere fisicamente.

La popolazione **infinita**, per sua natura, può essere osservata solo parzialmente.

infinita (l'insieme delle concentrazioni di nitrati misurabili in tempi diversi in un lago).

finita (tutti i geni del genoma umano)

Popolazione statistica e popolazione biologica

- Attenzione a non confondere la **popolazione statistica con la popolazione biologica.**

- La popolazione biologica si riferisce a tutti gli individui di una determinata specie che si trovano in un'area specifica ad un determinato tempo→ la stessa specie in uno stesso luogo

- A volte una popolazione statistica può non avere alcuna relazione con quella biologica

Popolazione statistica e popolazione biologica

- La pianta *Frittillaria n*el passato era molto diffusa nelle zone umide in Inghilterra. A causa della distruzione del suo habitat ora è confinata in poche località. In questo caso le popolazioni biologiche sono definite chiaramente e separate tra loro.

- Se si vogliono fare ricerche sulla *Frittillaria* le popolazioni statistiche coinvolte, che possono essere studiate, dipendono dalla domanda che ci si pone.

Popolazione statistica e popolazione biologica

Domanda	Unità di campionamento	Pop. Statistica
Che proporzione di piante sta fiorendo? Quanti semi per pianta?	Una singola pianta	Tutte le piante nella pop. biologica
Quante piante/m^2 nel campo		
Quanto tempo impiegano le api a visitare un fiore in un'area fissata?	????	
Quante api visitano in un periodo di osservazione di 5 min un certo fiore in un giorno?	????	

Popolazione statistica e popolazione biologica

Domanda	Unità di campionamento	Pop. Statistica
Che proporzione di piante sta fiorendo?	Una singola pianta	Tutte le piante nella pop. biologica
Quanti semi per pianta?	Una singola pianta in fiore	Tutte le piante in in fiore
Quante piante/m^2 nel campo	Un'area di $1m^2$	Tutte le aree di $1m^2$ nel campo
Quanto tempo impiegano le api a visitare un fiore in un'area fissata?	**????**	
Quante api visitano in un periodo di osservazione di 5 min un certo fiore in un giorno?		

Popolazione statistica e popolazione biologica

Quanto tempo impiegano le api a visitare un fiore in un'area fissata?	Una visita di un'ape a un fiore	Tutte le visite delle api ai fiori
Quante api visitano in un periodo di osservazione di 5 min?	Un periodo di osservazione di 5 min	Tutti i periodi di osservazione di 5 min che possono esser osservati

Popolazione statistica e popolazione biologica

- ESEMPIO:

Vogliamo descrivere e contare il n° di specie diverse di invertebrati in aree quadrate di una certa zona→

una sola popolazione statistica: la popolazione di tutti i quadrati

diverse popolazioni biologiche

- ESEMPIO:

Si vuole analizzare il n° di globuli rossi e bianchi in diversi settori di una piastra

Popolazioni statistiche in laboratorio

- Se si fa un incrocio genetico e si determina il fenotipo di ogni discendente l'unità campionaria è un individuo, ma la popolazione è quella di tutti i possibili individui che potrebbero derivare da un incrocio di quel tipo.

- Si vuole misurare una concentrazione sconosciuta per conoscerne il valore.
- Quale campione?
- Quale unità campionaria?
- Quale popolazione?

Popolazioni statistiche in laboratorio

- Il campione potrebbe, ad es., consistere in 3 volumi di 10ml ottenuti dalla concentrazione sconosciuta.
- L'unità campionaria è un volume di 10ml
- La popolazione comprende tutti i valori della concentrazione che potrebbero essere ottenuti.

- Ci riferiremo sempre a popolazioni statistiche.

Popolazione statistica

La popolazione può corrispondere ad un insieme finito di elementi come ad es. tutti i lupi che vivono nel parco nazionale d'Abruzzo.

In generale, si preferisce definire la popolazione statistica come un insieme infinito di unità:

1) perché in alcuni casi l'insieme è realmente infinito come ad es. l'insieme delle lunghezze delle tigri del Bengala che si sono misurate, si misurano e si misureranno in Birmania, o l'insieme di tutti i risultati di un esperimento che può essere ripetuto un numero infinito di volte (almeno in teoria)

2) perché la dimensione della popolazione è sempre molto maggiore di quella del campione.

POPOLAZIONE: ESEMPI

- Indagine sull'età media dei lupi in Abruzzo.
  Quale popolazione statistica?

- Indagine sulla produzione di tabacco in Italia.
  Quale popolazione statistica?

Disegno campionario

- Il **disegno campionario o piano di campionamento** è il metodo usato per selezionare il campione.
- Occorre che il campione venga scelto **in modo casuale** in modo da non favorire l'inserimento di alcuni elementi rispetto ad altri, o l'autoselezione tra chi deve rispondere (il campione deve essere rappresentativo della popolazione).

N. B. Scegliere in modo "casuale" non vuol dire "a casaccio"

La randomizzazione nelle indagini campionarie

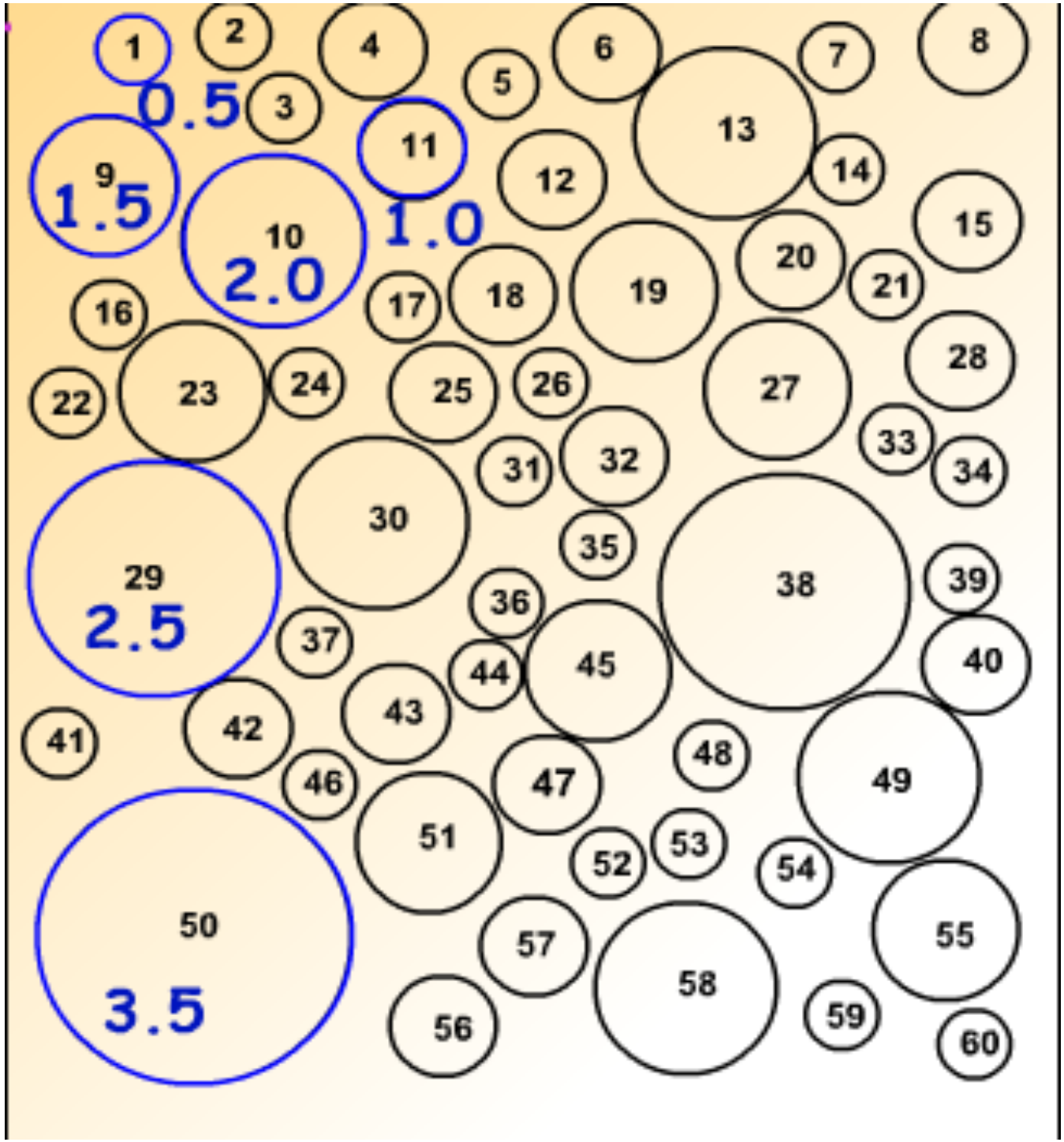
- Scegliendo un campione in modo casuale (random) viene data, a tutti gli elementi, la stessa probabilità di essere scelti.
- Nel **campionamento casuale semplice** (**CCS**) si estrae un campione in cui ogni unità della popolazione ha la stessa probabilità di essere selezionata. Inoltre, campioni della stessa dimensione hanno tutti la stessa probabilità di essere selezionati.

60 cerchi (popolazione):

si vuole stimare il diametro medio.

Occorre estrarre un campione

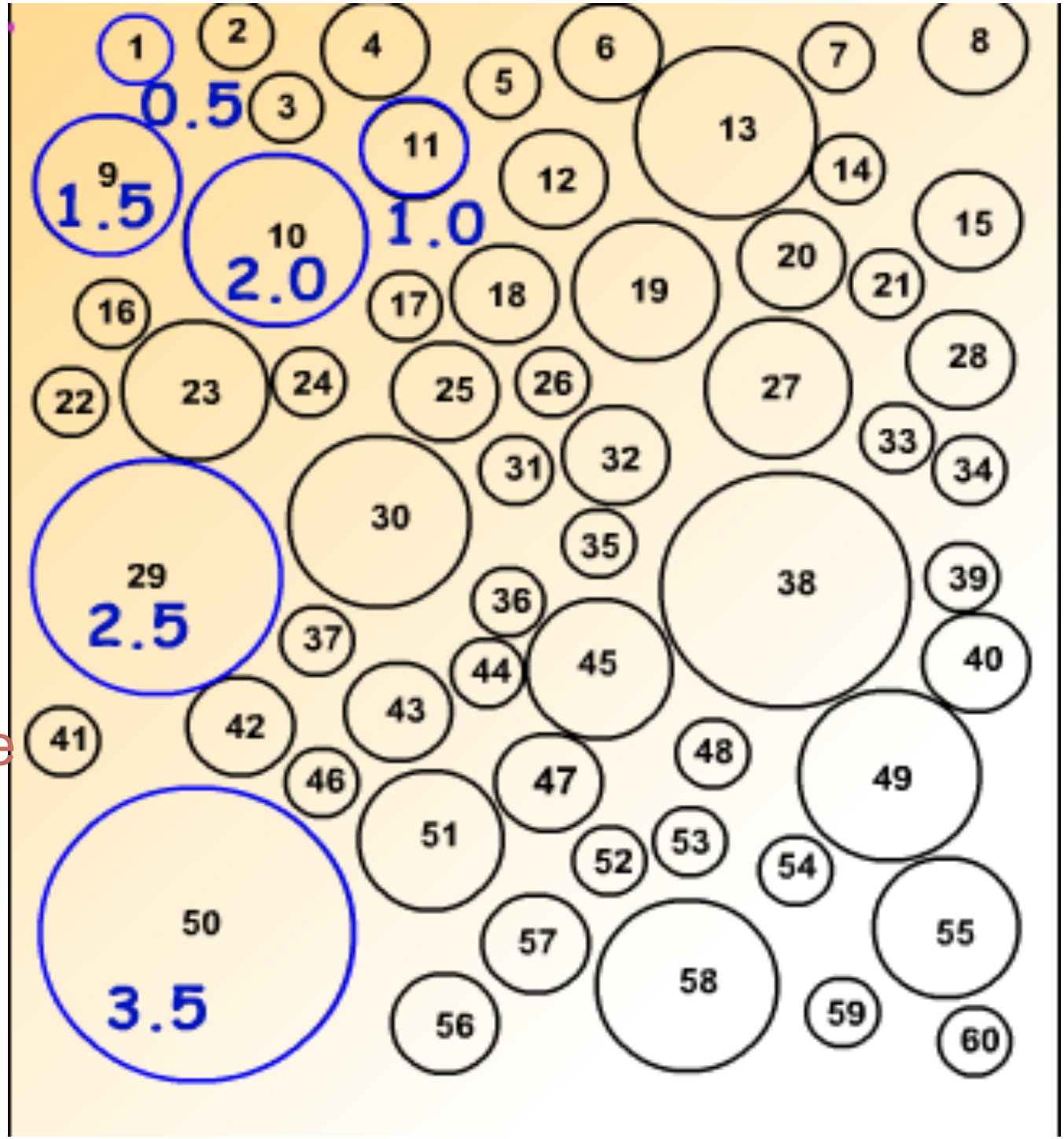
Che tipo di campione?
Quale dimensione?



60 cerchi:
si vuole
stimare il
diametro medio.
(μ = 1cm)

Si estrae un
campione
casuale semplice
di dimensione 5.

Come?



Come si sceglie un campione casuale semplice?

- Il CCS: analogia con le estrazioni senza reimmissione delle unità da un'urna (ad es. gioco del lotto)

- a) etichettare le unità
- b) usare il computer
- c) oppure, usare le tavole di numeri casuali

Tavole di numeri casuali

….5965 2913 5612 6361 7075 5490 9626 4307 0840 7945 5801 9383
6173 8358 9236 5543 5811 5520 5814 7864 1223 5344 3649 6397
1678 4400 7715 7614 1209 7729 0220 2108 0784 8837 3916 0282
4490 3442 6471 6593 4131 9772 7594 8863 0874 1864 8117 6411
7012 2682 3074 5746 2723 5681 0989 8015 0818 5380 9981 3758
2939 6585 6658 7756 7916 9770 2868 2128 2665 2386 6003 5982
8829 2833 8160 2101 3365 4121 4522 8216 2039 2993 4362 6363
2914 4955 6364 5237 6456 5561 0176 2425 2968 3834 6077 4302
3499 9938 7231 2136 2161 1365 2764 7836 1584…….

- Ogni numero in tabella è, con uguale probabilità, una della 10 cifre comprese tra 0 e 9.
- Scegliendo una coppia a caso è ugualmente probabile che sia una delle 100 coppie 00, 01,…,99
- Scegliendo una tripletta a caso è ugualmente probabile che sia una delle 1000 triplette 000, 001,…,999

- Le cifre sono indipendenti l'una dall'altra

Tavole di numeri casuali: esempio

5 elementi da una popolazione di 60.

Si osservano ad es. le righe 75 e 76 :

2868 2128 2665 2386 6003 5982 8829 2833 8160
2101 3365 4121 4522 8216 2039 2993 4362 6363

Si osservano le coppie di numeri
28 68 **21** 28 **26** 65 **23** 86 **60** 03 59 82 88 29 28 33
81 60 21 01 33 65 41 21 45 22 82 16 20 39 29 93
43 62 63 63

(28 21 26 23 60)

Variabilità campionaria: esempio dei cerchi

Cosa può succedere estraendo un campione?

- Variabilità campionaria:

Estraggo 3 CCS dalla pop. e calcolo la media $\bar{x}$ del campione

1) 59 29 13 56 12

1 2,5 2 1 1 → $\bar{x}$ = **1,5**

2) 55 43 58 11 20

1,5 1 2 1 1 → $\bar{x}$ = **1,3**

3) 2 20 21 8 7

0,5 1 0,5 1 0,5 → $\bar{x}$ = **0,7**

Distorsione: esempio dei cerchi

Cosa può succedere estraendo un campione?

- **Distorsione**

Estraggo 2 CCS dalla pop. e calcolo la media $\bar{x}$ del campione

(solo cerchi piccoli)

1) 1 2 3 5 7

0,5 0,5 0,5 0,5 0,5 → $\bar{x}$ = **0,5**

(solo cerchi grandi)

2) 29 38 50 30 13

2,5 2,5 3,5 2 2 → $\bar{x}$ = **2,5**

Possibili problemi nel campionamento casuale semplice

Il campionamento casuale semplice spesso non è effettuabile come ad es. nel caso si voglia misurare la lunghezza media di una popolazione di pesci in un certo fiume.

Oppure, come nel caso di misurazioni di tipo ecologico dove, ad es., si voglia stimare la concentrazione media di anidride carbonica in un lago.

Quanti campioni? Popolazione finita

- Data una popolazione di 6 vermi di lunghezza (cm):

  8, 9, 10, 10, 11, 12 vogliamo stimarne la lunghezza media osservando un campione casuale semplice, di dimensione 2, estratto senza reimmissione.

Media della popolazione = ?

La media della popolazione è un **parametro incognito**

Campionamento casuale semplice senza ripetizione

La probabilità di estrarre un particolare campione di numerosità **n** da una popolazione di dimensione finita **N** è pari a:

$$1 \Big/ \binom{N}{n}$$

Ci sono $\binom{N}{n}$ campioni che possono essere generati in questo modo e quindi tutti con la stessa probabilità, ossia per mezzo di un campionamento casuale semplice. Questo tipo di campionamento viene anche detto "**in blocco**", poiché non si è interessati all'ordine di estrazione delle n unità campionarie.

Campioni casuali semplici da una **popolazione finita-** Variabilità tra campionj

- Data una popolazione di 6 vermi di lunghezza (cm):

  8, 9, 10, 10, 11, 12 vogliamo stimarne la lunghezza media osservando un campione casuale (CCS), di dimensione 2, estratto senza reimmissione.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
8	8	8	8	8	9	9	9	9	10	10	10	10	10	11
9	10	10	11	12	10	10	11	12	10	11	12	11	12	12

$\overline{x}$ = **8.5 9.0 9.5 10.0 10.5 11.0 11.5** (valori ottenuti da ciascun campione)

Media della popolazione = 10cm (che supponiamo nota in quest'es.)

15 possibili campioni di dimensione 2

Variabilità tra campioni e all'interno dei campioni

- La variabilità tra campioni dipende in parte dalla variabilità della popolazione. Ad es. una serie di campioni sulla temperatura corporea mostrerà poca variabilità tra l'uno e l'altro, mentre gli stessi campioni mostreranno grande variabilità se le misure saranno riferite alla pressione del sangue.

- La variabilità tra campioni dipende in parte anche dalla dimensione dei campioni. Più sono grandi più si somigliano

- La variabilità all'interno dei campioni dipende dal fatto che gli individui non sono tutti uguali.

Parametri e stime

- Nell'esempio precedente si vuole stimare la media incognita della popolazione , ossia un **parametro incognito** della popolazione.
- A tale scopo si fa uso della media campionaria $\overline{x}$ calcolata su un solo campione, ossia di una **statistica campionaria.**
- Si noti come il valore di $\overline{x}$ **vari al variare** del campione → possibile errore di campionamento

Errori di campionamento

- **L'errore di campionamento** è la differenza tra il risultato relativo al campione e quello relativo alla popolazione effettiva; tale errore è dovuto alle fluttuazioni casuali nei campioni.
- Nell'es dei vermi gli **errori di campionamento** sono dati dalla differenza tra il valore 10 (media della pop) e i valori delle medie campionarie $\bar{x}$
- Un **errore non dovuto** al campionamento si verifica quando i dati del campione sono raccolti o analizzati in modo sbagliato (ad es. campione distorto, strumento di misura impreciso, o dati registrati erroneamente)

Errori di misura

- Errore di misura- difficile da evitare, può rappresentare una componente importante della variabilità.
- Per es. è noto che la caratteristiche comportamentali hanno una bassa ripetibilità: la misurazione di un comportamento in un individuo potrebbe dare un risultato diverso se ripetuta in un secondo momento.
- Si può ridurre l'errore di misura eseguendo misurazioni precise. Se impossibile si misura più volte ogni individuo e si usa la misura media

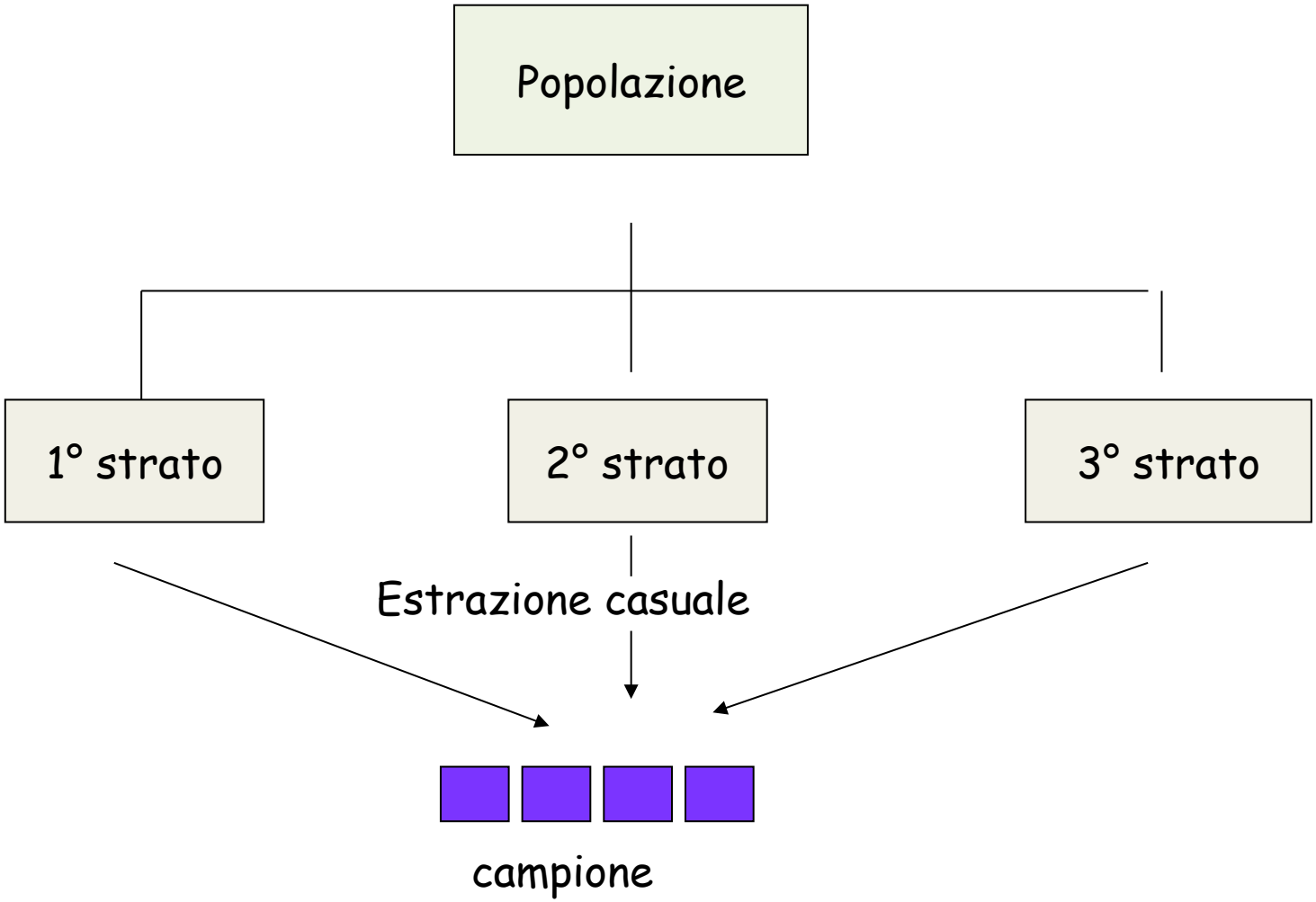
Il campionamento casuale a strati

Ci sono altri disegni campionari che costruiscono **campioni probabilistici** come il CCS.

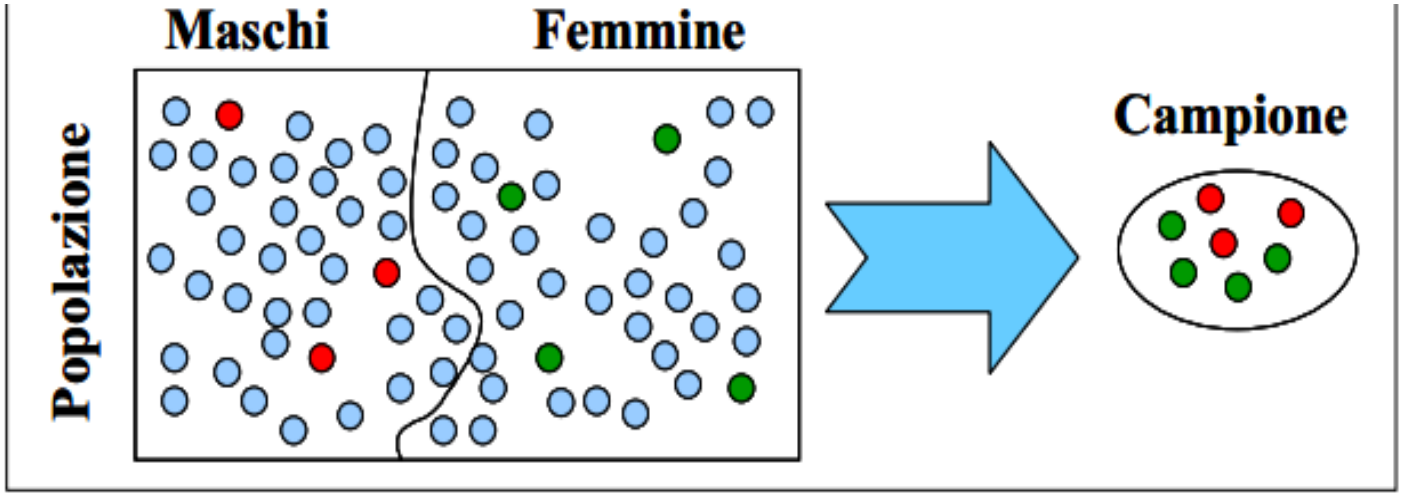
Nel campionamento casuale a strati si suddivide la popolazione in gruppi di unità simili, o strati. Si sceglie, poi, un **CCS** per ogni strato (non necessariamente della stessa dimensione) e si uniscono alla fine tutti i **CCS** formando un unico campione.

- Il campionamento a strati ha senso solo se la caratteristica da analizzare mostra differenze consistenti tra i vari strati.
- Se gli strati sono ben scelti, si ha una maggior precisione delle stime dei parametri della popolazione rispetto al campionamento casuale semplice

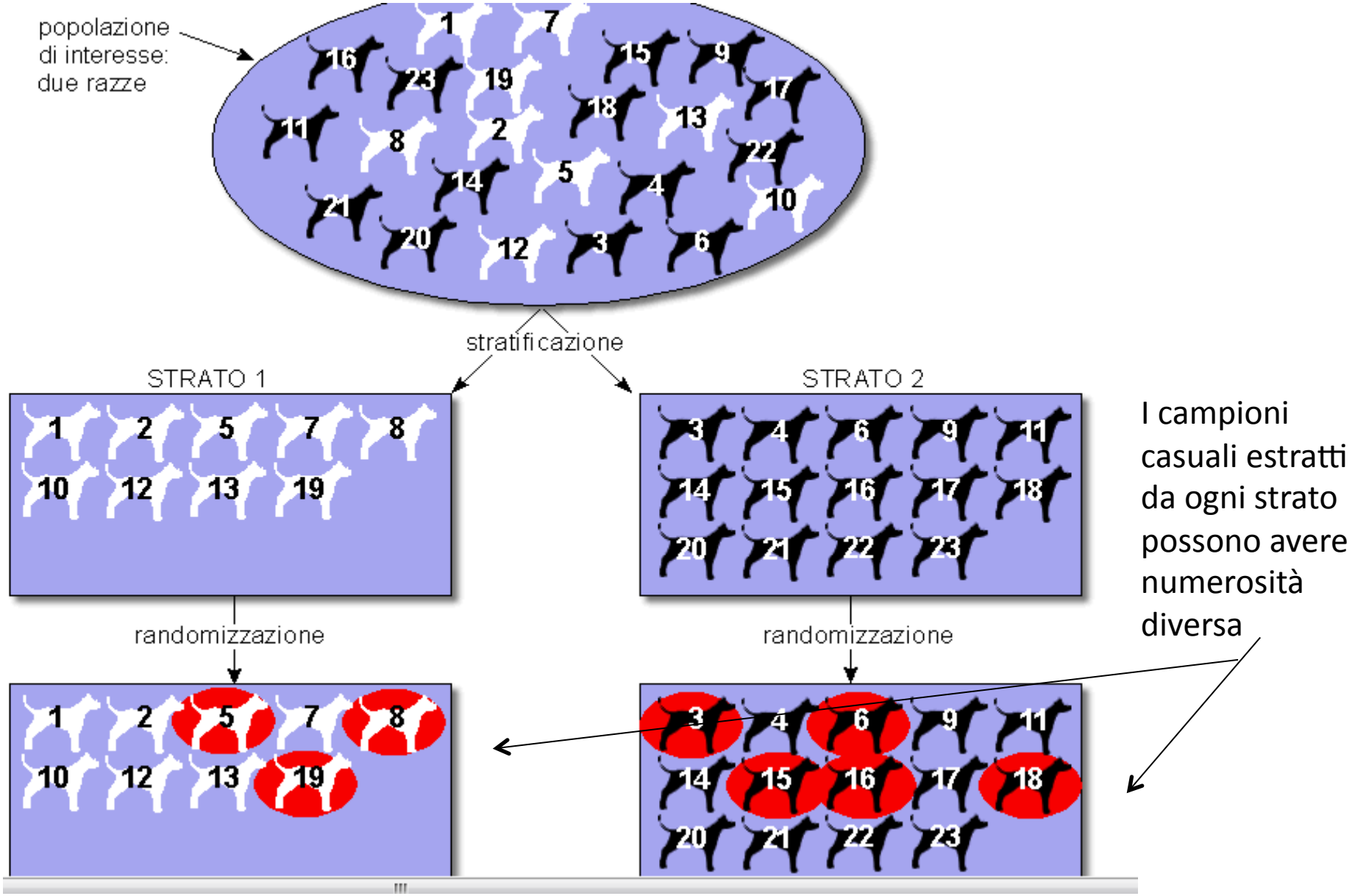
Campionamento a strati



Campionamento a strati



I campioni casuali estratti da ogni strato possono avere numerosità diversa


popolazione
di interesse:
due razze
stratificazione
STRATO 1
STRATO 2
randomizzazione
randomizzazione
I campioni
casuali estratti
da ogni strato
possono avere
numerosità
diversa

ESEMPIO. Supponi di voler studiare la produzione di latte delle bovine in una regione ove vengono allevate vacche di due diverse razze: la Bianca Val Padana e la Frisona. È noto che la produzione di quest'ultima è superiore (per motivi genetici) rispetto alla Bianca; perciò, converrà suddividere la popolazione-oggetto in due strati (strato «Bianca» e strato «Frisona») e poi campionare all'interno di ciascuno di essi per randomizzazione semplice o sistematica. Supponi inoltre di conoscere che, nella regione considerata, il rapporto numerico Frisona/Bianca sia 9/1; allora, potrai scegliere un campione proporzionale che rispetti la proporzione esistente nella popolazione: un campione di 100 vacche sarà composto da 90 Frisone e 10 Bianche Val Padana.

Frisona Bianca Val Padana

- Esempio. Si vuole estrarre un campione di piante da un grande appezzamento sperimentale in cui le piante vengono cresciute in 3 file. Lo scopo è di stimare il raccolto totale dell'appezzamento. Un'ispezione ha mostrato che ci sono differenze tra i raccolti delle piante nelle diverse file. Le **file** sono usate come **strati**.
- **Quante piante dobbiamo campionare da ogni fila?**
- Più sono numerose le piante in una fila, o più sono variabili i raccolti delle piante in una fila, più deve essere grande la dimensione del campione.
- Supponiamo che per ragioni pratiche la numerosità del campione non possa superare 140.

Continuazione esempio

i	N_i	s_i
1	100	2
2	200	4
3	400	1

Noti il numero di piante **Ni** e la deviazione standard **si** per ogni strato, è possibile procedere in 3 diversi modi per scegliere la dimensione del campione per ogni strato.

Cosa fareste?

i	N_i	s_i
1	100	2
2	200	4
3	400	1

Noti il numero di piante **Ni** e la deviazione standard **si** per ogni strato, è possibile procedere in 3 diversi modi per scegliere la dimensione del campione per ogni strato.

i	campioni uguali	campioni proporzionali (20% di ogni strato)	campioni ottimali (proporzionali a $N_i s_i$)
1	47	20	20
2	47	40	80
3	47	80	40
	141	140	140
Er. st. della stima del raccolto totale	118	126	100

- Nello schema di campionamento ottimale, allo strato 2 corrisponde la dimensione campionaria maggiore, perché è lo strato più variabile ed è il secondo più grande.

- Perché non un CCS?

Campionamento a strati

i	N_i	s_i
1	100	2
2	200	4
3	400	1

Esempio. Si vuole estrarre un campione di piante da un grande appezzamento sperimentale in cui le piante vengono cresciute in 3 file.

Lo scopo è di stimare il raccolto totale dell'appezzamento.

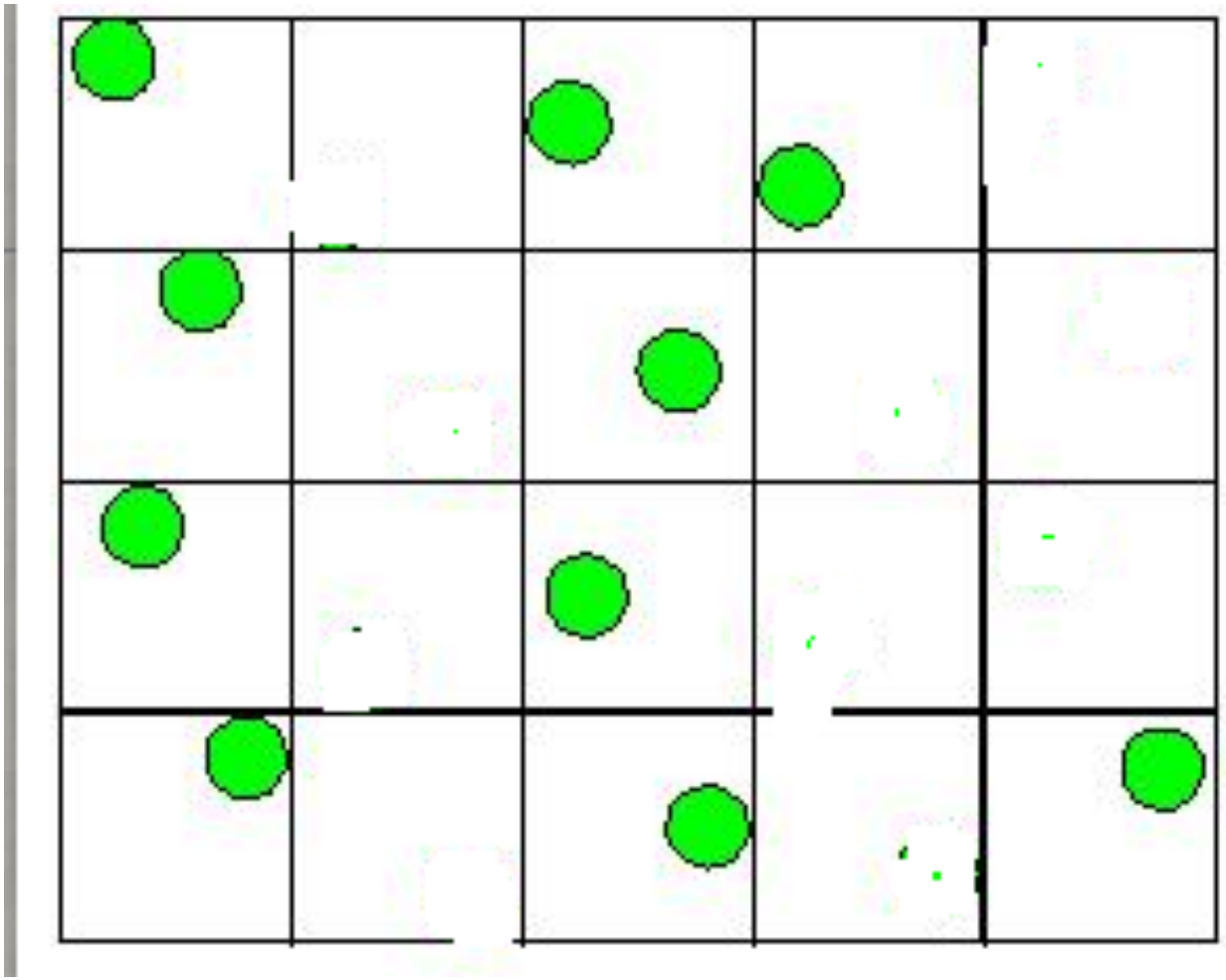
Un'ispezione ha mostrato che ci sono differenze tra i raccolti delle piante nelle diverse file.

Campionamento a strati

Le file sono usate come strati.

- Perché non un CCS?

Campionamento: come?



La copertura non è equilibrata e alcune zone sono scoperte.
Un unico CCS per tutta la superficie?

Dividiamo la popolazione di unità campionarie in strati orizzontali di unità in qualche modo simili tra loro.

Poi estraiamo CCS da ogni strato

Campionamento a due stadi

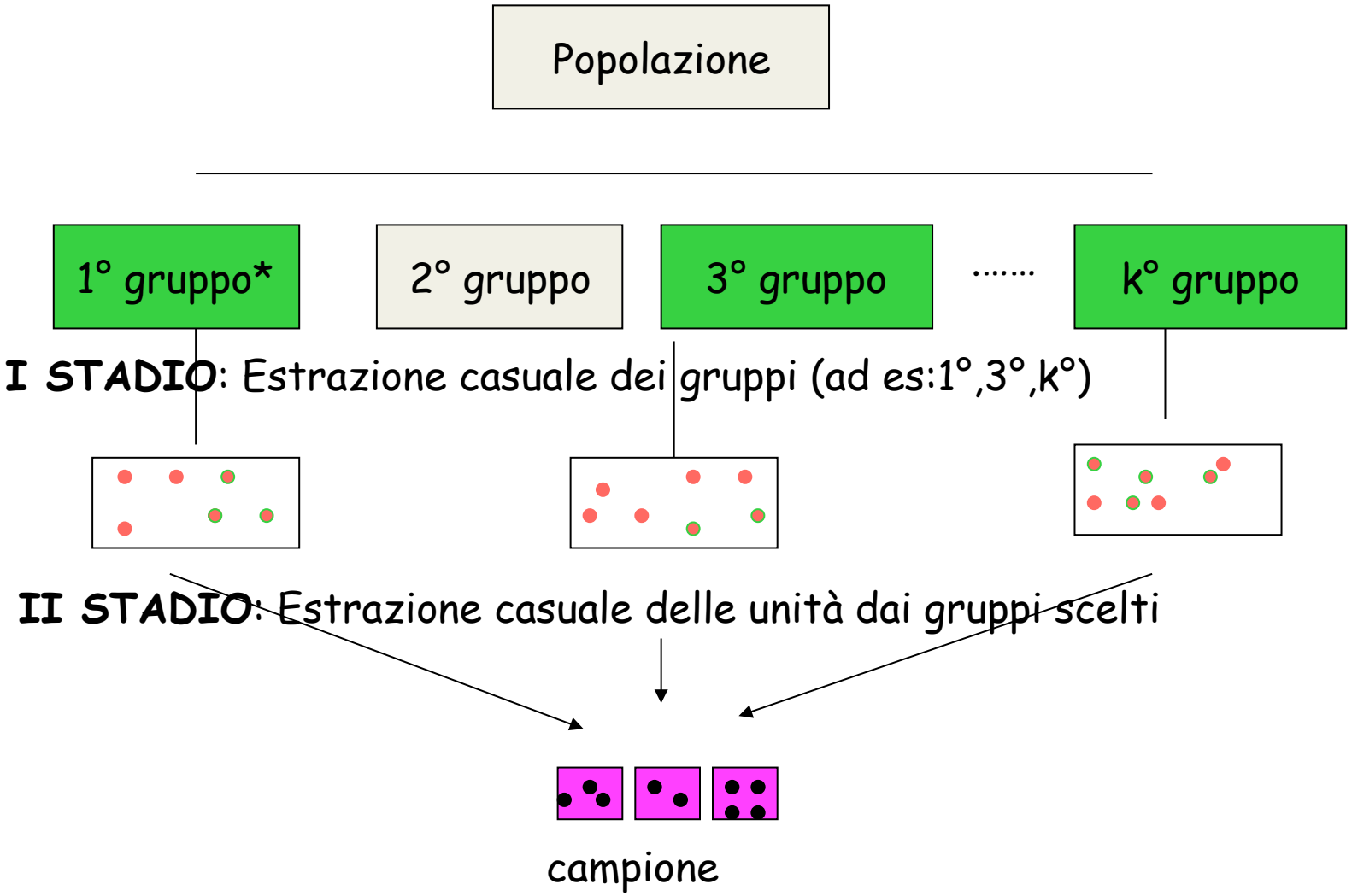
Questo tipo di campionamento viene usato se si deve campionare da una popolazione molto numerosa (ad es. di piante, microrganismi, ecc.).

Gli elementi della popolazione sono divisi in gruppi possibilmente della stessa numerosità. Si estrae poi un campione casuale di n gruppi (1° stadio) e si sceglie un elemento (o più elementi) a caso in ciascuno di questi n gruppi (2°stadio). I gruppi possono ad es. essere nidiate o gabbie che sono le unità su cui effettuare il campionamento.

Il campione risultante sarà non distorto.

Attenzione La scelta di un appropriato piano di campionamento è importante nel piano di uno studio e può condizionare fortemente l'esito dello stesso.

Campionamento a grappoli (a 2 stadi)



***si parla di grappolo o gruppo di unità della popolazione**

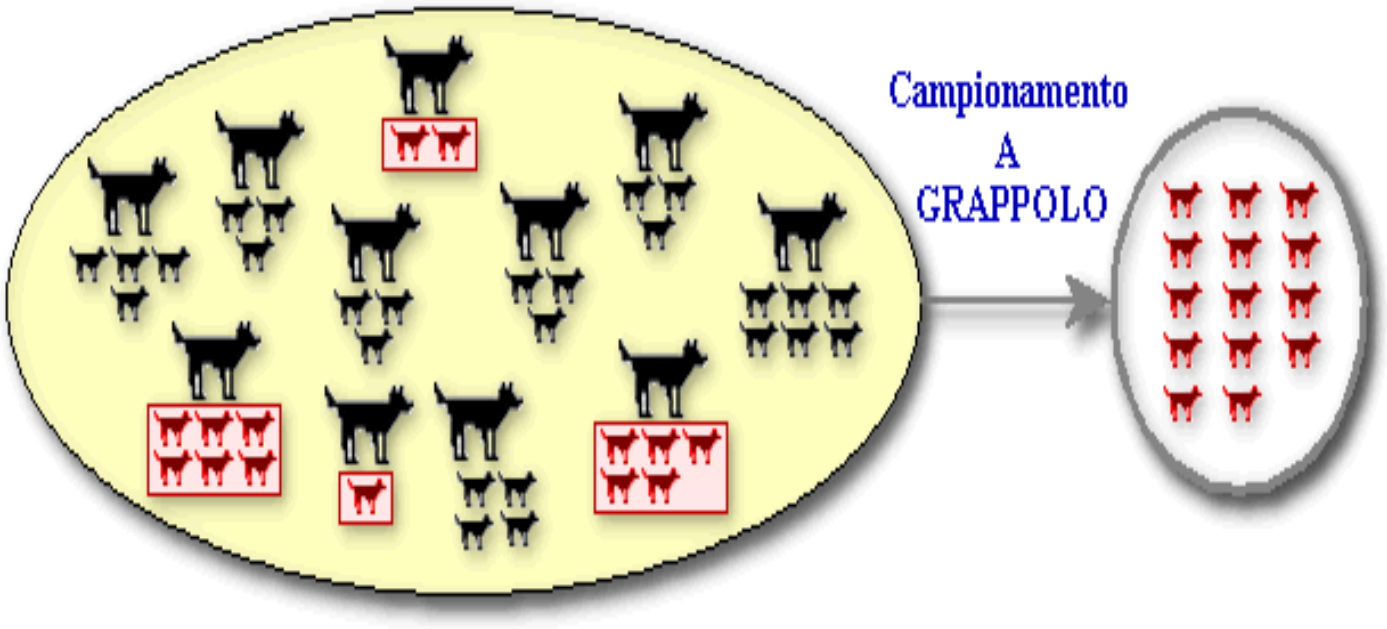
Campionamento a grappolo

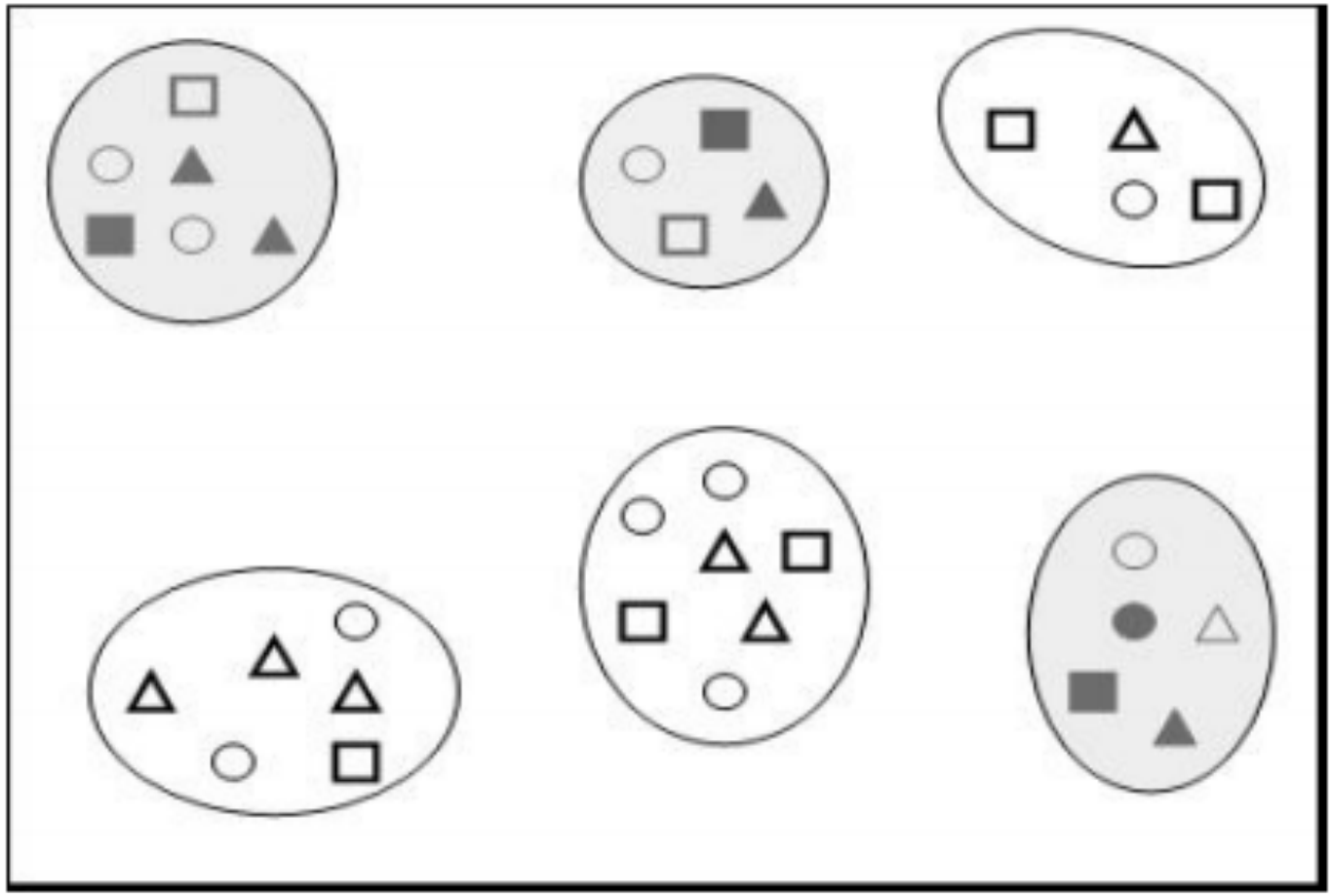
- Il campionamento a grappolo è meno preciso dei campionamenti visti prima, pertanto il suo uso è giustificato se la pop. è molto numerosa o se ci sono ragioni economiche. Se il budget è limitato questo tipo di campionamento permette di ottenere campioni maggiori rispetto agli altri metodi.
- Il grappolo è di fatto una popolazione in miniatura, che ne rispetta tutte le caratteristiche fondamentali.
- La condizioni per cui abbia senso effettuare **un piano di campionamento a grappoli è che ci sia molta eterogeneità all'interno dei grappoli e molta omogeneità tra loro.** Basta prenderne quindi solo alcuni per rappresentare le caratteristiche dell'intera popolazione

Campionamento a 2 stadi (a grappolo)
Esempio

empio esempio esempio

ESEMPIO. Si deve stimare la presenza di una malattia che colpisce i cuccioli di cane poco dopo la nascita. L'unità di indagine è rappresentata dal «cucciolo». Si procede ad effettuare un campionamento a grappolo, selezionando, mediante randomizzazione semplice o sistematica, un certo numero di *nidiate*.





molta eterogeneità all’interno dei grappoli (o gruppi) e molta omogeneità tra loro

ESEMPIO

Una scuola media superiore è formata da 78 classi (gruppi), ciascuna di 24 studenti.

Si desidera estrarre un campione per stimare il numero medio di ore settimanali dedicate allo sport.

Per motivi organizzativi e di costo si decide di selezionare 10 classi (**I stadio**)e di porre la domanda a tutti gli studenti delle classi estratte (**II stadio**), durante una delle ore di lezione.

- Esempio. Si vuole stimare la superficie totale delle foglie degli alberi in una determinata area.
- Si può usare il campionamento a 2 stadi: prima si effettua un campionamento di piante, poi un campionamento di foglie.

- **NOTA** Il campionamento a strati e quello a 2 stadi possono sembrare simili, ma sono completamente diversi. Nel campionamento a strati si formano dei gruppi all'interno dei quali le unità siano il più possibile omogenee, e **si campiona da tutti gli strati**. Nel campionamento a 2 stadi si formano dei gruppi, sulla base di raggruppamenti naturali o artificiali, tali che siano il più possibile disomogenei al loro interno (in modo da evitare di scegliere un campione di gruppi "particolari"). **Si campiona solo da pochi gruppi estratti.**

Esempi di campionamento

In uno studio (*Determining Reef Fish Abundance in Marine Protected Areas in the Northern Mariane Islands-2008*), si voleva stimare l'abbondanza dei pesci di alcuni atolli. Campionamento a strati.

Gli strati sono basati sull'habitat qualitativo o sulla profondità marina.

No. of strata in which species occur (u_j)	Frequency (No. of species) (f_j)	Percentage of species $\left(\frac{f_j}{n} \times 100\right)$	Cumulative percentage
1	117	35.5	35.5
2	61	18.5	53.9
3	37	11.2	65.2
4	24	7.3	72.4
5	23	7.0	79.4
6	12	3.6	83.0
7	14	4.2	87.3
8	10	3.0	90.3
9	9	2.7	93.0
10+	23	7.0	100.0
	n = 330	100.0	

G2, L4, S4

Campionamento a strati e a stadi

- Sia gli strati, sia i gruppi sono sottoinsiemi della popolazione che non si sovrappongono.
- Nel campione finale sono rappresentati tutti gli strati, ma solo un sottoinsieme di gruppi.
- Col campionamento a strati i migliori risultati si ottengono quando gli elementi negli strati sono internamente omogenei.
- Col campionamento a stadi i migliori risultati si ottengono quando gli elementi all'interno dei gruppi sono internamente eterogenei.

Quale campionamento?

1) Si deve stimare la quantità media del raccolto in una certa piantagione di tabacco. L'unità di indagine è rappresentata da…..?

Che tipo di campionamento?

2) Si vuole estrarre un campione di studenti del II anno in una certa università per studiarne le conoscenze informatiche. Come?

Stratificazione per facoltà, nell'ambito della facoltà per corso di laurea, considero il II anno e stratifico per sesso.

3)Un ricercatore della J.Hopkins University intervista tutti i pazienti cardiopatici di 30 ospedali che ha scelto a caso nel suo stato. Quale campionamento ha usato?

Campioni probabilistici

La struttura generale per il campionamento statistico è il campione probabilistico. I disegni campionari che abbiamo studiato producono campioni probabilistici

Campione probabilistico

Un **campione probabilistico** è un campione scelto a caso. Per ottenere un campione probabilistico è necessario conoscere quali campioni sono possibili e che probabilità ha ogni campione di essere estratto.

Attenzione agli errori di copertura, alle mancate risposte, alle distorsioni nelle risposte.
Si noti che più è grande il campione, più è efficace la randomizzazione nel produrre un campione rappresentativo.

La variabilità

- Come descrivere gli "individui che variano"?
- L'unico modo per risolvere il problema sarebbe di considerarli tutti.
- Esempio

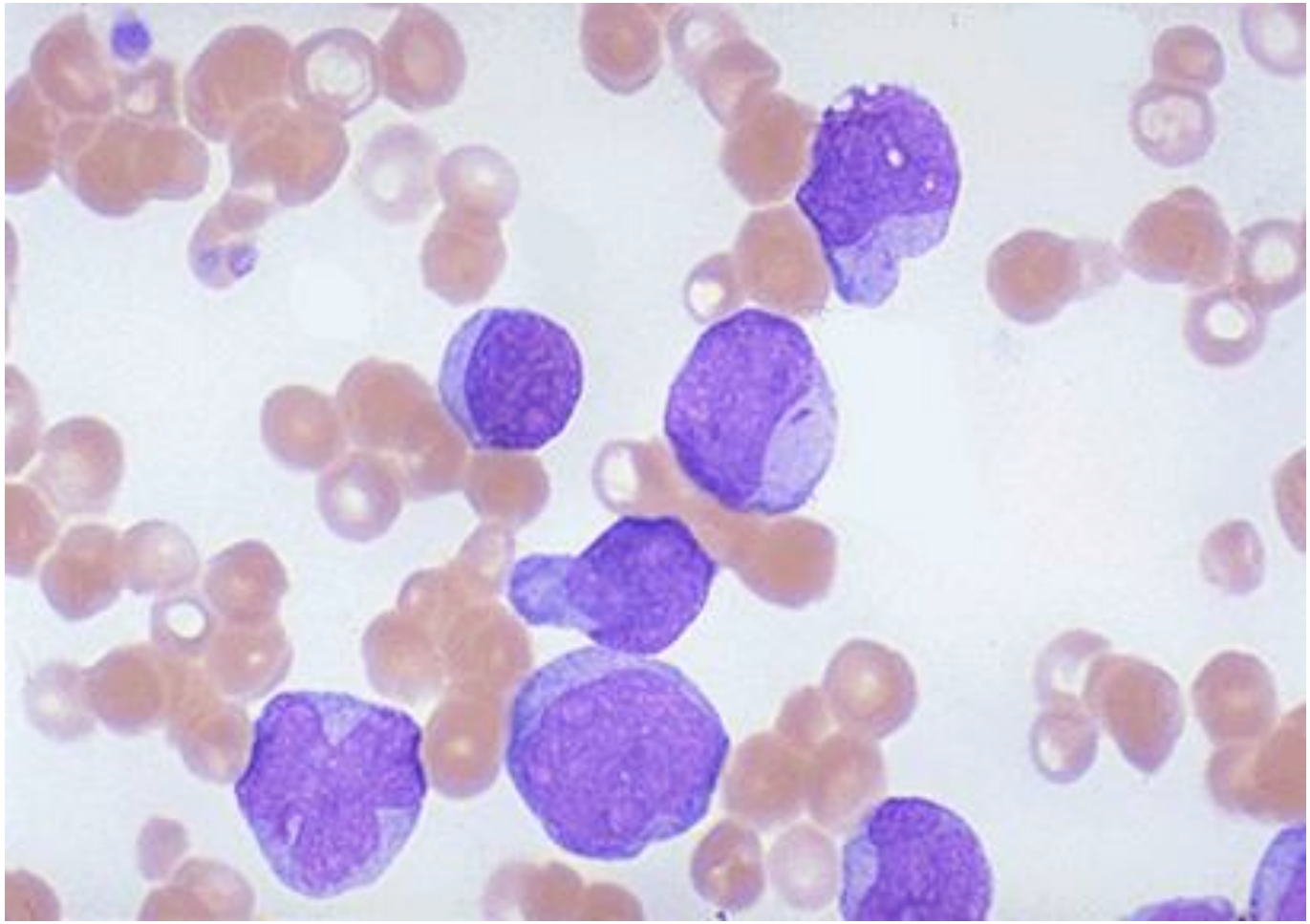
Diversi tipi di cellule nel sangue dei vertebrati (globuli rossi e diverse forme di globuli bianchi).

Se volessimo descrivere in quale proporzione questi globuli si presentano nelle rane sane, dovremmo trovare qual è la percentuale per ciascun tipo.

Se esaminassimo al microscopio una piccola quantità di sangue, potremmo avere una situazione come in figura 1.

La risposta varierebbe a seconda del punto di osservazione nella slide, ma si potrebbero esaminare molte diverse slides, e molte rane sane!! **Enorme numero di osservazioni possibili**.

Fig 1



Distorsione

- Un problema è legato all'uso di un campione
- Come scegliamo il settore in cui contiamo le cellule?
- Si potrebbe incorrere in una sottostima o sovrastima sistematica del parametro della popolazione se il campionamento non è effettuato correttamente o lo strumento non è adeguato.
- Questo tipo di errore è detto **distorsione**
- **La distorsione (bias)** è una discrepanza **sistematica** tra le stime e il valore vero della caratteristica di interesse della popolazione.

Campione non rappresentativo-Distorsione

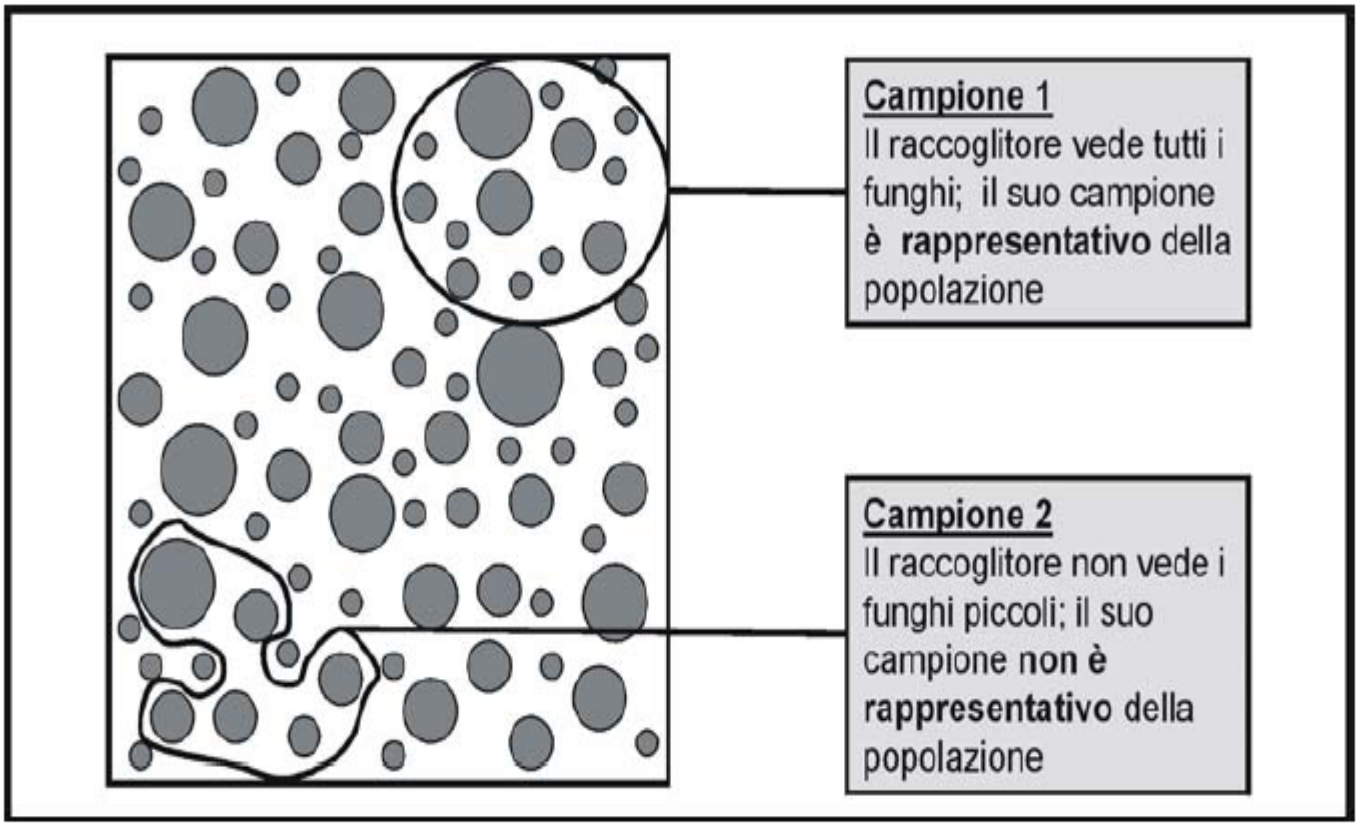


Figura 1.3. Due campioni di funghi raccolti da due sperimentatori, il secondo dei quali molto miope. Il primo campione è rappresentativo della popolazione dei funghi, il secondo non lo è perché il raccoglitore miope non vedrà i funghi più piccoli.

Peso dei topi campagnoli attirati in una trappola

E' nota la media dei pesi della popolazione dei topi:

M = 134,5 g

a) molta distorsione, poca variabilità

110, 112, 111, 112, 114, 107, 105 ($\bar{x}$=110,1, s=3,1)

b) poca distorsione, molta variabilità

111, 135, 140, 120, 130, 150,132 ($\bar{x}$=131,1 s=12,8)

c) molta distorsione, molta variabilità

110, 105, 135, 165, 149, 130, 135 ($\bar{x}$=147, s=42,1)

d) poca distorsione, poca variabilità

139, 132, 140, 143, 133, 135, 137 ($\bar{x}$=137 s=4)

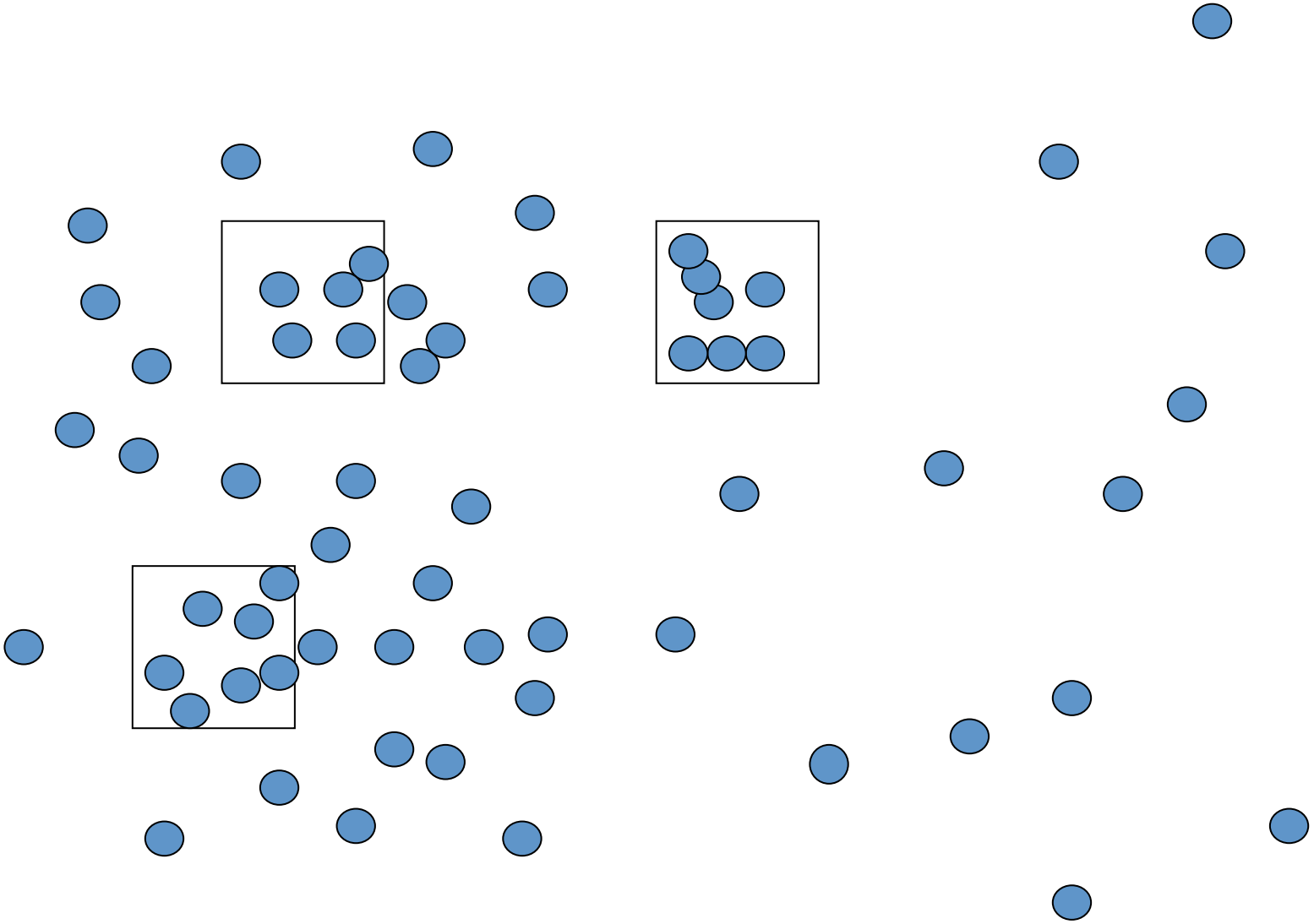
Campione non rappresentativo Distorsione

Esempio: per stimare la densità di animali in una determinata zona, è stato scelto un campione di superfici quadrate nella zona stessa (v. figura).

Questo campione sovrastima la densità media e sottostima la varianza. Infatti, per caso, tutti i quadrati sono in zone di grande densità.

Tale campione è distorto e non è rappresentativo della popolazione. Se si possedesse una mappa delle densità della zona si rifiuterebbe un simile campione che non è rappresentativo della popolazione.

Nella pratica può essere difficile realizzare un campionamento del tutto casuale.



Campioni non probabilistici: esempi

Si ha un campionamento non probabilistico quando le unità della popolazione non hanno una probabilità predefinita e non nulla di entrare a far parte del campione.

Campionamento accidentale:

si ha un campionamento accidentale quando il ricercatore sceglie come rispondenti alla sua indagine le prime persone che capitano, senza criteri definiti.

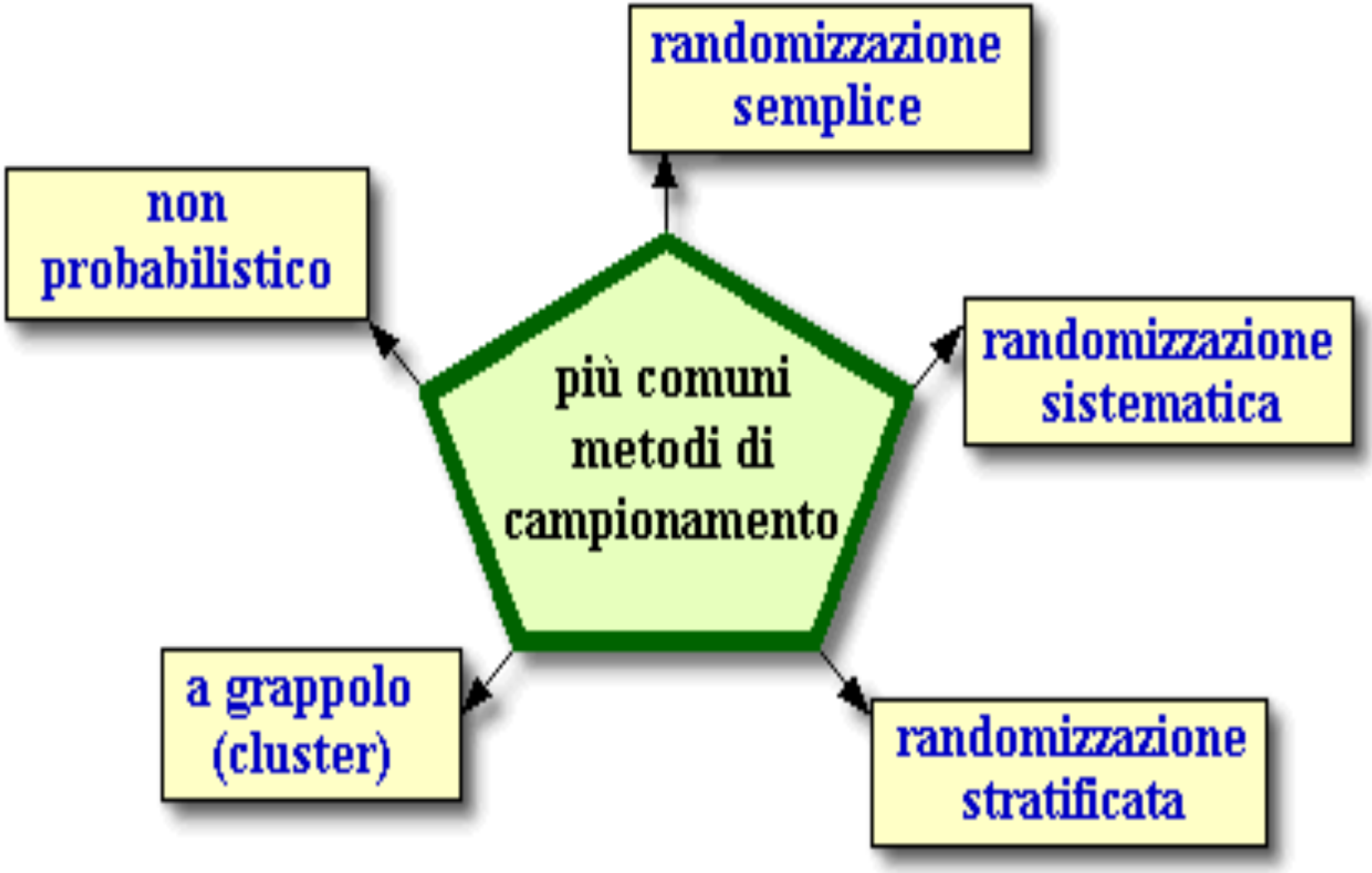
Campionamento a valanga:

composto da più fasi, dopo aver intervistato alcune persone dotate delle caratteristiche richieste, queste persone servono per identificare altri soggetti che possono essere intervistati in una fase successiva e che a loro volta producono informazioni per identificare altri soggetti con le caratteristiche per essere inclusi nel campione, creando così un effetto a valanga.

Campionamento a elementi rappresentativi:

si ha un campionamento a elementi rappresentativi quando si selezionano all'interno della popolazione gli elementi che il ricercatore ritiene rappresentativi per gli obiettivi della ricerca.

I metodi di campionamento più comuni



Campionamento non probabilistico

Gli esperimenti

- Negli studi sperimentali i dati vengono raccolti tramite un **esperimento**.
- In **un esperimento** si sottopongono le unità sperimentali (**soggetti**) ad alcuni trattamenti per osservarne le reazioni (**risposte**).
- **Un trattamento** è una condizione sperimentale applicata ai soggetti.

Gli esperimenti

Esperimento

Unità sperimentale	Trattamento	Risposta
pomodori	fertilizzante	raccolto
topo	radiazioni	mortalità
paziente	farmaco	pressione

Si parla anche di **variabile esplicativa** (trattamento e di **variabile di risposta** (risposta)

Disegno degli esperimenti--Disegni completamente randomizzati

Si vuole confrontare una nuova varietà di granturco, con contenuto alterato di amminoacidi, con una varietà normale.

Si hanno a disposizione 30 pulcini. Si assegna, a caso, ciascuna varietà a 15 pulcini maschi di 1 giorno. Dopo 21 giorni si misura l'aumento di peso nei 2 gruppi.

Come avviene l'assegnazione casuale?

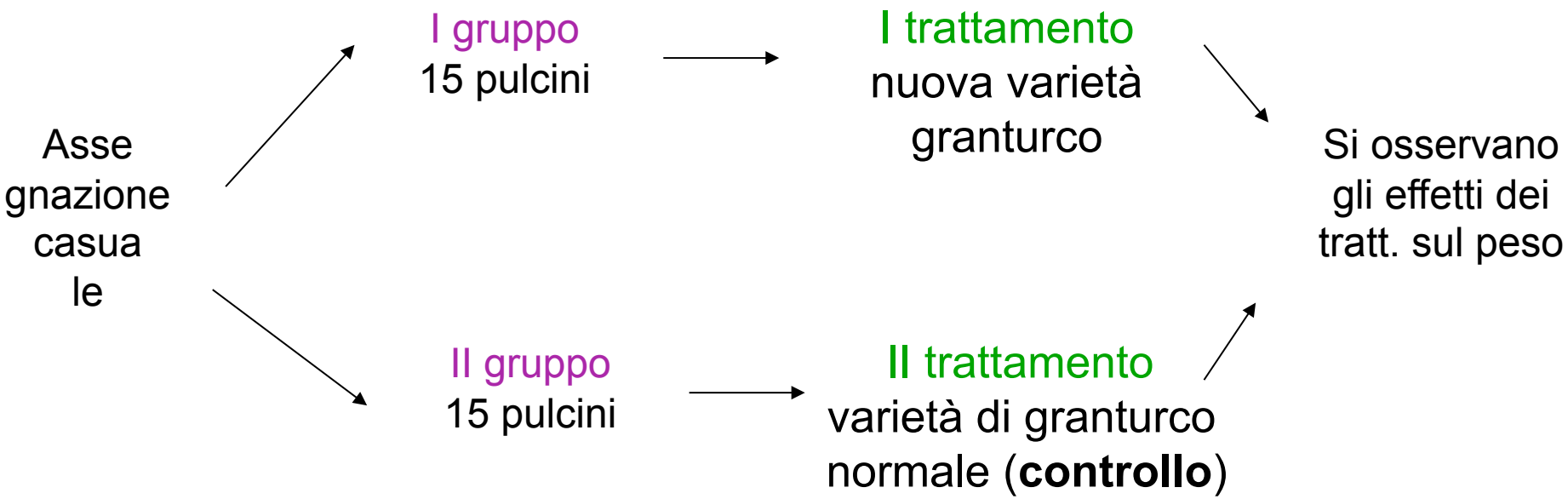
Disegni completamente randomizzati

Basta assegnare a caso i primi 15 pulcini al I trattamento

(occorre generare 30 numeri casuali da 1 a 30)

- 1) numerare i soggetti (30)
- 2) usare le tavole dei numeri casuali per assegnare i soggetti ai trattamenti
  - 11309 22069 26239 42564 19623 92280 07246…….!

Disegno degli esperimenti--Disegni completamente randomizzati



1) numerare i 30 pulcini
2) usare i numeri casuali generati per assegnare i pulcini ai trattamenti

Disegno degli esperimenti--Disegno completamente randomizzato

1) numerare i soggetti (40)
2) usare le tavole dei numeri casuali per assegnare i soggetti ai trattamenti

11369 23569 26339 42564 39623 92280 17246…….!

La randomizzazione

- **In un disegno completamente randomizzato tutti i soggetti sono assegnati in modo casuale ai trattamenti.**

Fattori di confondimento

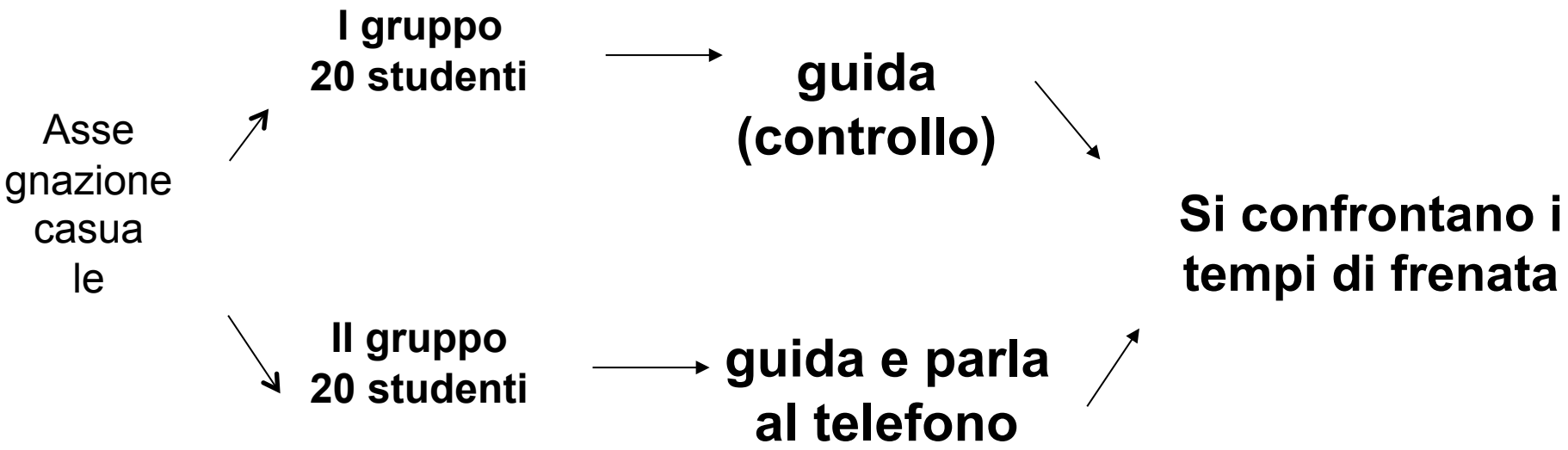
- Ad esempio, supponiamo che un **esperimento** voglia verificare l'efficacia di un farmaco (causa) per ridurre la pressione (effetto).
- Soggetti con caratteristiche simili vengono assegnati a caso ai due gruppi (trattati e controllo).

- Se, invece, il farmaco fosse assegnato a un solo gruppo di soggetti, l'effetto dell'assunzione del farmaco potrebbe **confondersi** con le caratteristiche dei soggetti quali, ad esempio, peso, tipo di dieta, attività fisica svolta……
- Perciò non sarebbe possibile stabilire una **relazione di causa-effetto**.

Fattori di confondimento

- In statistica, con il termine **fattori di confondimento** si indica la presenza di altri fattori o variabili che interferiscono con la possibilità di identificare un particolare fattore come agente causale di un dato evento.

Disegno completamente randomizzato



Il controllo

- **Il confronto** con un **gruppo di controllo** permette di ridurre il più possibile l'influenza di altri fattori che potrebbero **confondere** l'effetto vero del trattamento.

- In generale, **il disegno sperimentale** tende a controllare il più possibile i fattori esterni ai quali non si è interessati.

La Replica

- E' importante ripetere le osservazioni o i trattamenti per avere un'indicazione della variabilità dei risultati e quindi un'idea della precisione delle nostre stime.
- Come ripetere ogni trattamento in un esperimento pianificato?
- Esempio: confronto tra 3 diversi fertilizzanti (A, B, C) da applicare a un campo.

C	B
A	A
B	A
C	C
B	A
B	C

Possiamo confrontare la variabilità dei raccolti dei 4 plots a cui è stato somministrato il fertilizzante *A*.
Più repliche abbiamo, meglio stimeremo i raccolti medi e le differenze tra questi

I principi base di un esperimento

- Sono 3 i principi base di un esperimento:

1. **Randomizzazione**
2. **Controllo**
3. **Repliche**

Studi di osservazione

In **uno studio di osservazione** si studiano e si confrontano le unità a cui è capitato di "ricevere un trattamento". Il trattamento non viene imposto e non si cerca di influenzare le risposte.

Nota: Le indagini campionarie viste prima, che selezionano una parte della popolazione d'interesse per studiarne tutto l'insieme, sono un esempio di studio di osservazione

Studi di osservazione

Studio di osservazione

Unità	Trattamento	Risposta
patate	condizioni meteo	raccolto
individuo	radiazioni	mortalità
paziente	fumo	tumore(polmone)

Studi sul campo

- Anche lo **studio sul campo** è uno studio di osservazione.
- In questo caso, si osserva direttamente il comportamento che interessa studiare, in genere in un ambito naturale.
- Molte delle conoscenze sul comportamento animale sono state ottenute con questo tipo di studi.
- Questo vale per tutte quelle aree di ricerca dove **indagini campionarie e esperimenti sono impossibili.**

Studi di osservazione e Esperimenti

- Gli studi di osservazione **non** possono essere usati per valutare gli effetti di un qualche intervento sulle unità sottoposte a studio, possono solo identificare possibili cause di effetti.

- Solo un esperimento ben disegnato e ben eseguito può stabilire un **rapporto di causa ed effetto** (tra il trattamento e la risposta).

- Torniamo all'esempio dell' **esperimento** che vuole verificare l'efficacia di una nuova varietà di granturco per pulcini (causa) per aumentare il peso (effetto).
- Soggetti con caratteristiche simili vengono assegnati a caso ai due gruppi: trattati e controllo (granturco normale).
- Se, invece, un gruppo di soggetti fosse semplicemente **osservato (studio di osservazione**), l'effetto della nuova varietà di granturco potrebbe confondersi con le caratteristiche dei soggetti quali, ad esempio, peso, sesso……
- Perciò non sarebbe possibile stabilire una **relazione di causa-effetto.**

Studi di osservazione
Un esempio

- Uno studio di osservazione, durato 11 anni, su un gruppo di fumatori e non fumatori, ha mostrato che ci sono state 7 morti per tumore al polmone su 100000, nel campione di non fumatori, mentre ce ne sono state 166 su 100000, nel campione di fumatori.
- Tuttavia questo non prova che fumare causa il tumore ai polmoni, perché i fumatori, ad esempio, potrebbero fumare a causa dello stress o dello stile di vita e questo potrebbe provocare il tumore polmonare.

Un esperimento

- Torniamo all'esempio dello studio di osservazione sul fumo.
- Per controllare il fattore (stress) che si **confonde** col fattore fumo, si possono dividere i due campioni in diverse categorie di stress (esperimento).
- Quindi si confrontano i fumatori e non fumatori che sono nella stessa categoria di stress.
- Solo così si può stabilire una relazione di **causa (fumo) – effetto (tumore).**

Riassunto

- Gli studi di osservazione e gli esperimenti producono dati che servono a rispondere a domande specifiche.
- Le indagini campionarie, che selezionano una parte della popolazione d'interesse per studiarne tutto l'insieme, sono un esempio di studio di osservazione.
- Negli esperimenti, a differenza degli studi di osservazione, i soggetti vengono sottoposti a trattamenti.
- Gli studi di osservazione spesso non riescono a mostrare l'influenza di una variabile su un'altra, perché possono esserci effetti di confondimento.
- Il disegno campionario è il metodo usato per scegliere il campione. (Attenzione alla distorsione e variabilità)

- Il campione probabilistico più importante è il campione casuale semplice.
- Altri campioni probabilistici sono: il campione a strati e il campione a due stadi
- Per scegliere un CCS si possono usare le tavole dei numeri casuali o un software statistico.
- In un esperimento si somministrano uno o più trattamenti ai soggetti.
- Il disegno di un esperimento descrive la scelta dei trattamenti e il modo in cui i soggetti sono assegnati ai trattamenti.
- Con gli esperimenti è possibile provare l'esistenza di relazioni causa-effetto.
- Controllo, randomizzazione e repliche alla base di un disegno degli esperimenti

ESERCIZI

- 1. Quali tra le seguenti fonti di dati possono essere considerate fonti attendibili di dati statistici?
- (a) Dati ISTAT
- (b) Un esperimento di laboratorio
- (c) L'esperienza personale
- (d) Un sondaggio aperto sul sito internet de "La Repubblica"
- (e) Il censimento

- ESERCIZI
- In un'indagine statistica del 1998 si sono confrontati due trattamenti per la cura del tumore al seno: la mastectomia e la radioterapia.
- Lo studio è stato effettuato nel seguente modo:
- Si sono confrontati i tempi di sopravvivenza delle pazienti in 25 ospedali dal momento della prima diagnosi e si sono riportati i dati nella seguente tabella:

	sopravvivenza > 3 anni	sopravvivenza < 3 anni
Mastectomia	60	40
Radioterapia	210	90

Da questa indagine statistica possiamo dedurre che la radioterapia è più efficace della mastectomia nel prolungare l'aspettativa di vita? V F

- Per ciascuna delle seguenti variabili:

Tempo di sopravvivenza

Tipo di intervento: mastectomia/radioterapia

indicare se sono:

(a) variabile esplicativa

(b) variabile di risposta

- In uno studio sull'efficacia di due diversi farmaci A e B per ridurre la pressione, è stato confrontato un gruppo di 50 maschi di 60 anni che aveva fatto uso del farmaco A, con un altro gruppo di 50 maschi della stessa età che aveva usato il farmaco B.
- Si è notato che, mediamente, il livello di pressione è inferiore nei soggetti che assumono il farmaco A.

1. Si tratta di un esperimento o di uno studio di osservazione?
2. Qual è la variabile di risposta?
3. E' possibile concludere che il farmaco A ha effetto nel ridurre la pressione.

Esempio

1) quale delle seguenti affermazioni è vera?

I. Un'indagine campionaria è un esempio di disegno sperimentale

II. Il miglior metodo per studiare relazioni casuali è uno studio di osservazione

(A) I
(B) II
(C) Tutte
(D) Nessuna

Esempio

Viene testato un nuovo farmaco contro il raffreddore su 200 individui:
100 uomini e 100 donne.
Gli uomini ricevono il farmaco e le donne no.

Alla fine dello studio si osserva che gli uomini hanno avuto meno raffreddori.

E' corretto? Come pianifichereste questo esperimento?
Non ci sono controlli
Il sesso si confonde con l'uso del farmaco →
Uomini e donne potrebbero essere assegnati a caso al farmaco
Un gruppo potrebbe ricevere un placebo

Nella tabella seguente sono riportate le informazioni relative a un campione di sette piantine di pomodori di una data piantagione Lo scopo dello studio è di confrontare la crescita media delle piantine di pomodoro trattate con i 3 fertilizzanti A, B, C.

Tipo di fertilizzante	Altezza finale (cm)	Altezza iniziale (cm)
Tipo C	87	3
Tipo A	68	4
Tipo B	76	2
Tipo A	77	5
Tipo B	80	4
Tipo A	74	3
Tipo C	91	7

Qual è la popolazione di riferimento? Qual è l'unità statistica?

Si tratta di un esperimento o di uno studio di osservazione?

Qual è il parametro che vogliamo stimare? Come lo stimiamo?

Esercizio

- Un campione di pazienti ipertesi viene suddiviso in modo casuale in due gruppi. Ad un gruppo viene somministrato un farmaco contro l'ipertensione, mentre all'altro gruppo viene somministrato un placebo.
- Dopo sei mesi dall'inizio della terapia, sono stati confrontati i valori delle pressioni del sangue nei due gruppi.

Dire se si tratta di:

- (a) un esperimento
- (b) uno studio di osservazione

ESERCIZIO

Supponendo di voler condurre un'indagine, a Roma, sull'uso di cannabis da parte degli studenti che frequentano l'ultimo anno delle scuole superiori

i) poiché le scuole superiori sono molte, che tipo di campionamento effettuereste?

ii) descrivete, in modo sintetico, come lo effettuereste

Quali domande sui dati per un'indagine statistica?

- **Perché?**
  - **_ Qual è lo scopo dell'indagine**
- **Chi?**
  - **_ Quali unità statistiche, quante**
- **Quali variabili?**
  - **_ Quali variabili, quante, quali unità di misura**

L'analisi esplorativa dei dati

- Nelle prossime lezioni studieremo i metodi per a) **esplorare** e b) **descrivere** i dati.
- A tale scopo faremo uso di
  a) **grafici**
     istogrammi, grafici ramo-foglia, box-plot,...
  b) **riassunti numerici**
     centro, dispersione, percentili,...
- Attraverso l'analisi esplorativa dei dati cerchiamo di capire cosa i dati "vogliono dire".

Popolazione, unità statistiche, caratteri

- In statistica, il termine popolazione indica qualunque insieme di elementi o unità statistiche che sono l'oggetto della ricerca.
- Una variabile è un qualunque carattere o caratteristica misurabile o osservabile su un'unità statistica.
- I caratteri possono assumere modalità o valori differenti sulle diverse unità statistiche.

Popolazione

Campione

Unità statistica o di campionamento

Variabile o carattere statistico
(es. altezza, colore del vestito)

Valore o Modalità
(es. altezza = 1.65, colore vestito = rosso)

Esempi:

Ex1. Se misuriamo il peso di 100 topi: l'**unitá di campionamento** é il singolo topo (= unitá biologica); la **variabile** studiata é il peso; un'**osservazione individuale** (o **item**) é il peso di ciascun topo; il **campione statistico** é la collezione dei pesi di tutti e 100 i topi selezionati; se da tale osservazione traiamo conclusioni riguardanti il peso di tutti i topi, la **popolazione** é costituita dai pesi di tutti i topi esistenti.

ESEMPIO

Ex2. Se contiamo il numero di donne e bimbi in 100 famiglie: l'**unitá di campionamento** é una famiglia ($\neq$ individuo biologico); le **variabili** studiate sono due: il numero di donne ed il numero di bimbi; un **item** é il numero di donne e di bimbi in una tra le famiglie selezionate; il **campione** é la collezione di tutti i numeri di donne e di bimbi nelle 100 famiglie; se si vogliono trarre conclusioni sulle famiglie italiane, la **popolazione** sará la collezione dei numeri di donne e di bimbi in tutte le famiglie italiane.

Tipi di caratteri o variabili

I caratteri o variabili possono essere qualitativi o quantitativi.

- Una variabile qualitativa definisce il gruppo (categoria) di appartenenza.
  Ad es. il gruppo sanguigno, il sesso
- Una variabile quantitativa misura o conta qualcosa.
  Ad es. l'altezza di un individuo, il numero di figli in una famiglia

Variabili qualitative

2 tipi:

- **Variabili categoriche** (con scala nominale) che non possono essere ordinate

  Ad es. la specie, il sesso, il tipo di habitat

  Variabili categoriche di uso frequente sono le variabili binarie con 2 sole categorie possibili.

- **Variabili ordinate** (con scala ordinale) che possono essere ordinate

  Ad es. la scala di abbondanza per la classificazione dell'abbondanza di diverse specie di piante (dominante, abbondante, frequente, non comune, rara)

Variabili qualitative

- Molti metodi dell'inferenza statistica **non** possono essere applicati a **v. qualitative**

- Ma anche un semplice calcolo come la media non può essere effettuato.

- Es: gruppo sanguigno medio per 20 individui???????

Variabili quantitative

- Variabili quantitative discrete
- Le variabili discrete derivano in genere da **conteggi**

  Ad es. il numero di uova deposte da un uccello,

  Il n° di figli per famiglia

- Variabili quantitative continue

  Ad es. la temperatura, il peso di un individuo,

  la pressione sanguigna

Che tipo di variabile? Esempio

- Se un biologo che studia il comportamento di animali di una certa specie vuole codificare, in una serie di esperimenti, le reazioni degli animali come:

a) molto aggressivo b) aggressivo c) neutrale d) sottomesso e) molto sottomesso

- Qual è la variabile che si studia?
- Di che tipo di variabile si tratta?

Popolazioni - campioni - osservazioni

Se le osservazioni sono misurazioni

Si vuole misurare la lunghezza del becco di corvi catturati su un posatoio

Unità statistica:
Variabile:
Osservazione o valore:
Campione:
Popolazione statistica:

Popolazione biologica:

Popolazioni - campioni - osservazioni

Se le osservazioni sono misurazioni

Unità statistica: un corvo di un posatoio
Variabile: lunghezza del becco
Osservazione o valore: 8 mm
Campione: i corvi catturati sul posatoio e misurati
Popolazione statistica: i corvi del posatoio che sono **c**atturabili e misurabili
Popolazione biologica: la popolazione biologica può comprendere uccelli non catturabili (ad es. maschi e femmine di altri posatoi).

Oppure la popolazione statistica può includere individui di più popolazioni biologiche (ad es. se ci sono migratori invernali).

Se le osservazioni sono conteggi

Quante conchiglie su una certa spiaggia?

Unità statistica o campionaria:
Variabile:
Osservazione:
Campione:
Popolazione statistica:
Popolazione biologica:

Se le osservazioni sono conteggi

Osservazione: 23

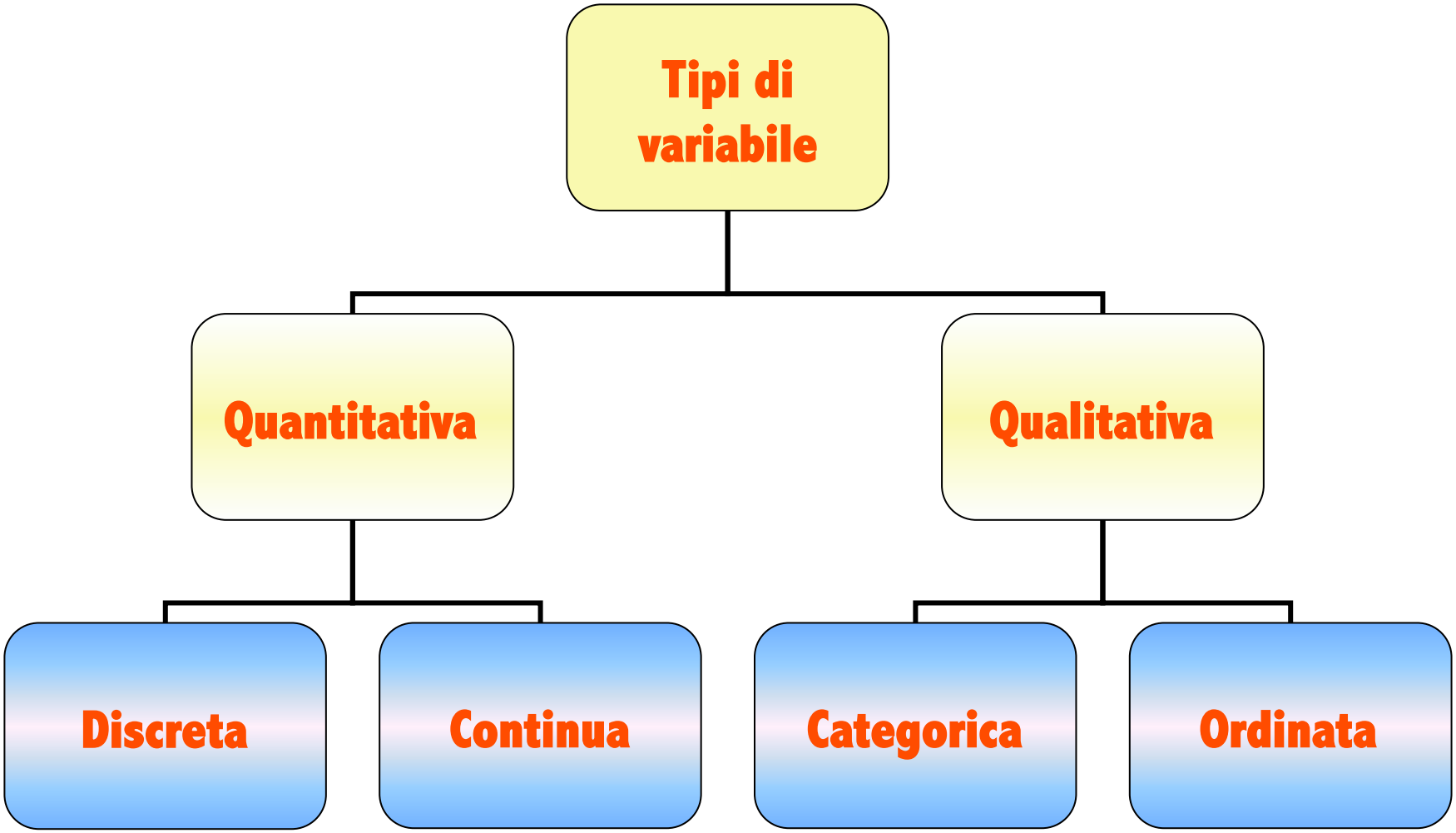
Variabile: numero di conchiglie

Unità statistica: un quadrato di area determinata in cui si setaccia la sabbia e si contano le conchiglie

Campione: il numero di quadrati (unità campionarie) esaminati

Popolazione statistica: il numero totale di quadrati che si possono individuare in tutta l'area di studio

Popolazione biologica: nessuna relazione tra popolazione statistica e biologica



E’ importante precisare il livello di misura dei dati osservati per determinare la procedura statistica da usare per analizzarli.

Proporzioni, percentuali, rapporti, tassi

- A volte, in biologia, si opera sui dati osservati per ottenere dei numeri derivati.

- Esempi importanti di variabili derivate sono le proporzioni, le percentuali, i rapporti e i tassi.

Proporzione-- Percentuale

- Una **proporzione** è il rapporto di una parte sul tutto.

► Ad es. se la lunghezza totale del corpo (testa+torace +addome) di un insetto è 7.2mm e se la testa è lunga 2.7mm, la proporzione della testa rispetto al corpo è 2.7/7.2=0.37.

► Ad es. Mortalità = N° morti/ N° abitanti

- Una **percentuale** è una proporzione moltiplicata per 100.

► Ad es. 2.7/7.2=0.37, se si moltiplica per 100 →37%

Rapporti

- Un **rapporto** è una parte divisa per un'altra parte (il numeratore non è compreso nel denominatore).

►Se la larghezza della capsula della testa di un insetto è 1.31mm e la lunghezza è 2.7mm il rapporto larghezza/lunghezza è pari a 1.31/2.7.

►Se in un campione ci sono 25 femmine e 32 maschi, il rapporto femmine/maschi è pari a 25/32=0.78 o 1:32/25 = 1/1.28, ossia, il rapporto femmine/maschi è 1/1.28.

Tassi

- In biologia, spesso, si fa riferimento ai **tassi** considerati come rapporti tra un'osservazione e un periodo di tempo.
- I tassi sono utili per esprimere variabili quali la crescita e i cambiamenti di una popolazione.

► Ad es. un germoglio cresce 15cm in 5 giorni

Il rapporto è 15:5 = 3:1

Il tasso è 3 cm/giorno

ESERCIZIO

Per ciascuna delle seguenti variabili:

Tempo di sopravvivenza

Tipo di intervento: mastectomia/radioterapia

indicare se sono:

(a) quantitative continue

(b) quantitative discrete

(c) qualitative ordinate

(d) qualitative categoriche

ESERCIZIO

Pressione del sangue (in millimetri di mercurio)

Livello di calcio nel sangue (microgr./ml)

- indicare se sono:
- (a) quantitative continue
- (b) quantitative discrete
- (c) qualitative ordinate
- (d) qualitative categoriche

ESERCIZIO

Si vuole condurre uno studio ecologico sullo stato di un lago e, come primo passo, si individuano 4 variabili che possano descriverlo:

la temperatura T dell'acqua (gradi centigradi),

il ph (moli per decimetro cubo),

il volume V di acqua contenuta nell'invaso (decimetri cubi),

la torbidita' dell'acqua (numero di particelle solide non solubili contenute in un decimetro cubo)

Per ogni variabile, indicare se si tratta di una variabile:

- (a) quantitativa continua,
- (b) qualitativa ordinata,
- (c) quantitativa discreta,
- (d) qualitativa categorica.

ESERCIZIO

- Per ciascuna delle variabili che seguono:
- tipo di habitat
- intensità di un terremoto
- posizione nella graduatoria di una gara: 1°,…10°, ecc…
- mese di nascita di un individuo
-
- indicare se sono:
- (a) Variabile qualitativa ordinata
- (b) Variabile qualitativa categorica
- (c) Variabile quantitativa continua
- (d) Variabile quantitativa discreta
-

I Grafici

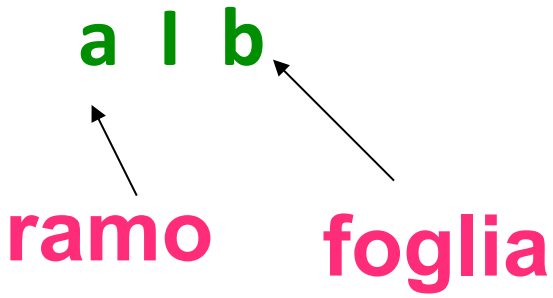
- Per costruire il grafico di una variabile occorre conoscere la sua **distribuzione**, ossia i valori (o modalità) che assume la variabile e quante volte li assume (frequenza).

Grafici per variabili quantitative

- Grafici ramo-foglia (stem and leaf)
- Istogrammi
- Diagrammi a segmenti

Grafici ramo-foglia: variabili

- Ogni numero è diviso in due parti:



Esempio:
lunghezza dell'ala di 10 passeri (mm)
59 64 68 71 73 75 75 77 80 80 (dati ordinati)

ramo	foglie
5	9
6	4 8
7	1 3 5 5 7
8	0 0

Grafici ramo-foglia

5	9
6	4 8
7	1 3 5 5 7
8	0 0

Il grafico precedente si può espandere dividendo ogni ramo a metà.

5	9	
6	4	← Comprende le foglie 0, 1, 2, 3, 4
6	8	← Comprende le foglie 5, 6, 7, 8, 9
7	1 3	
7	5 5 7	
8	0 0	

Grafici ramo-foglia

In genere, la foglia contiene l'ultima cifra del numero e il gambo le altre cifre.

Ad es. per i dati (in euro):

121.69, 281.06, 393.36

si arrotonda all'euro più vicino:

12 | 2 = 122 euro

28 | 1 = 281 euro

39 | 3 = 393 euro

Grafici ramo-foglia

Suggerimenti:

- Questi grafici si usano quando la dimensione n del campione

$$15 \leq n \leq 150$$

- Un grafico appropriato dovrebbe avere tra 5 – 20 gambi.
- Usare foglie di una sola unità. Se necessario, arrotondare i numeri.
- Specificare sempre l'unità di misura.

ESEMPIO

26
13
9
11
57

0 9
1 13
2 6
3
4
5 7

Media **Mediana** **Dispersione**

??????

Mean	StDev	Minimum	Median	Maximum
23,20	20,03	9,00	13,00	57,00

Istogrammi: variabili continue

Esempio 1 Le lunghezze (cm) dei coyote (CCS con n = 40)

- **Femmine n= 40**

93.0 97.0 92.0 101.5 93.0 84.5 102.5 97.8 91.0 98.0 93.5 91.7
90.2 91.5 80.0 86.4 91.4 83.5 88.0 71.0 81.3 88.5 86.5 90.0
84.0 89.5 84.0 85.0 87.0 88.0 86.5 96.0 87.0 93.5 93.5 90.0
85.0 97.0 86.0 73.7

- **Maschi**

97.0 95.0 96.0 91.0 95.0 84.5 88.0 96.0 96.0 87.0 95.0 100.0
101.0 96.0 93.0 92.5 95.0 98.5 88.0 81.3 91.4 88.9 86.4 101.6
104.1 88.9 92.0 91.0 90.0 85.0 93.5 78.0 91.0 83.8 103.0 100.5
105.0 86.0 95.5 86.5 90.5 80.0 80.0

Istogrammi: variabili continue

- Come si costruisce un istogramma?

1) Si divide il campo di variazione (range) delle osservazioni in classi possibilmente di uguale ampiezza
2) Si conta il numero di osservazioni in ogni classe
3) Si costruisce la tabella di frequenze e frequenze relative
4) Si disegna l'istogramma

NOTA: date n osservazioni ordinate in senso crescente, il campo di variazione è la differenza tra la più grande e la più piccola delle osservazioni

Istogrammi: variabili continue

- Un istogramma appropriato dovrebbe avere tra **7 – 20 classi** possibilmente di ugual ampiezza

N° valori (osservazioni)	N° classi (intervalli)
≤ 100	7--10
101—200	11—15
> 200	13--20

Istogrammi

Distribuzione delle frequenze e delle frequenze relative delle lunghezze dei coyote femmina

7 Classi	Frequenza (n_i)	Frequenza relativa (n_i/n)	Freq. relativa percentuale %
70- 75	2	0.05	5
75- 79	0	0	0
80- 84	6	0.15	15
85- 89	12	0.3	30
90- 94	13	0.325	32.5
95-99	5	0.125	12.5
100-105	2	0.05	5
Totale	n=40	1.00	100

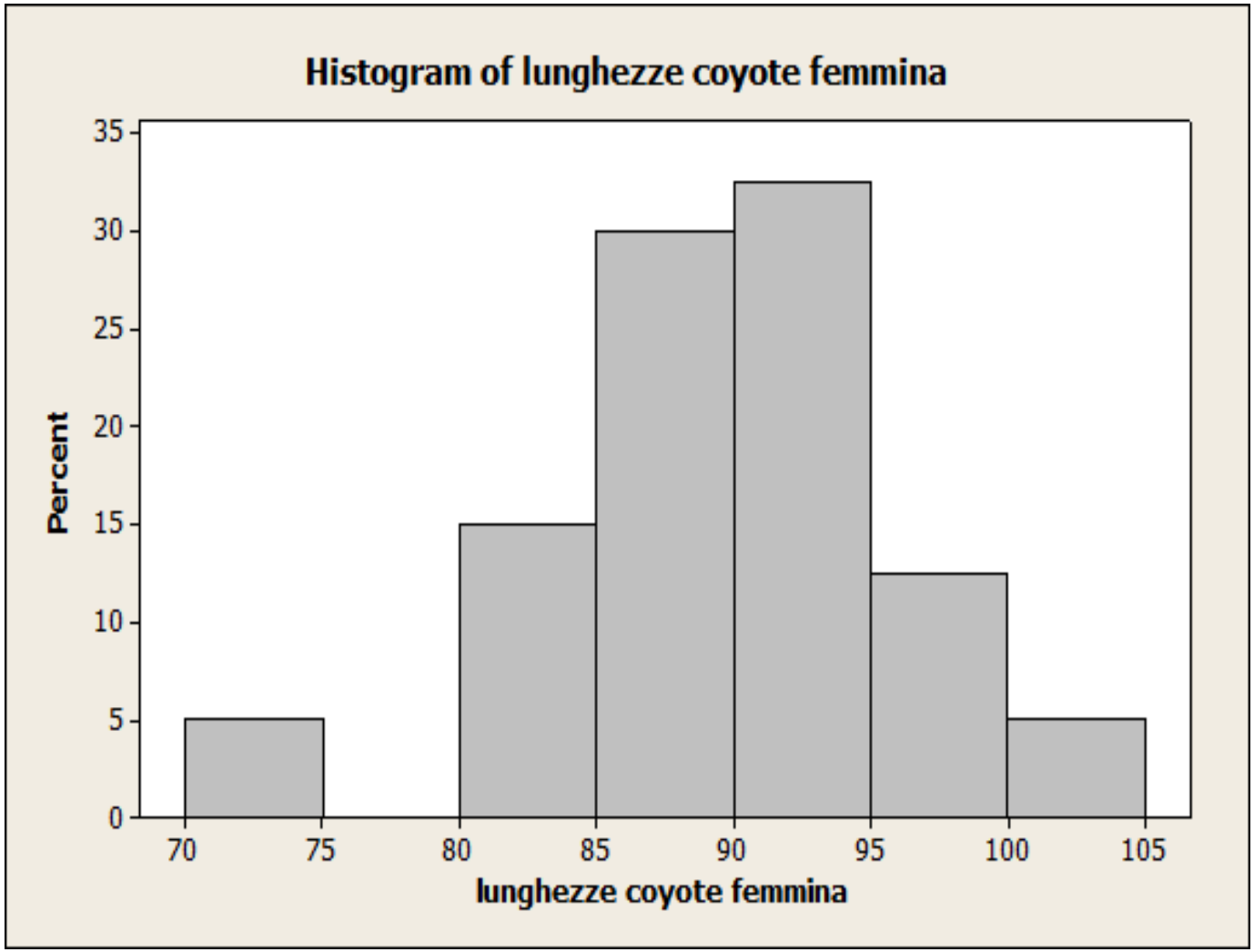
Istogramma delle lunghezze (cm) dei coyote femmina

Ogni ramo è diviso a metà

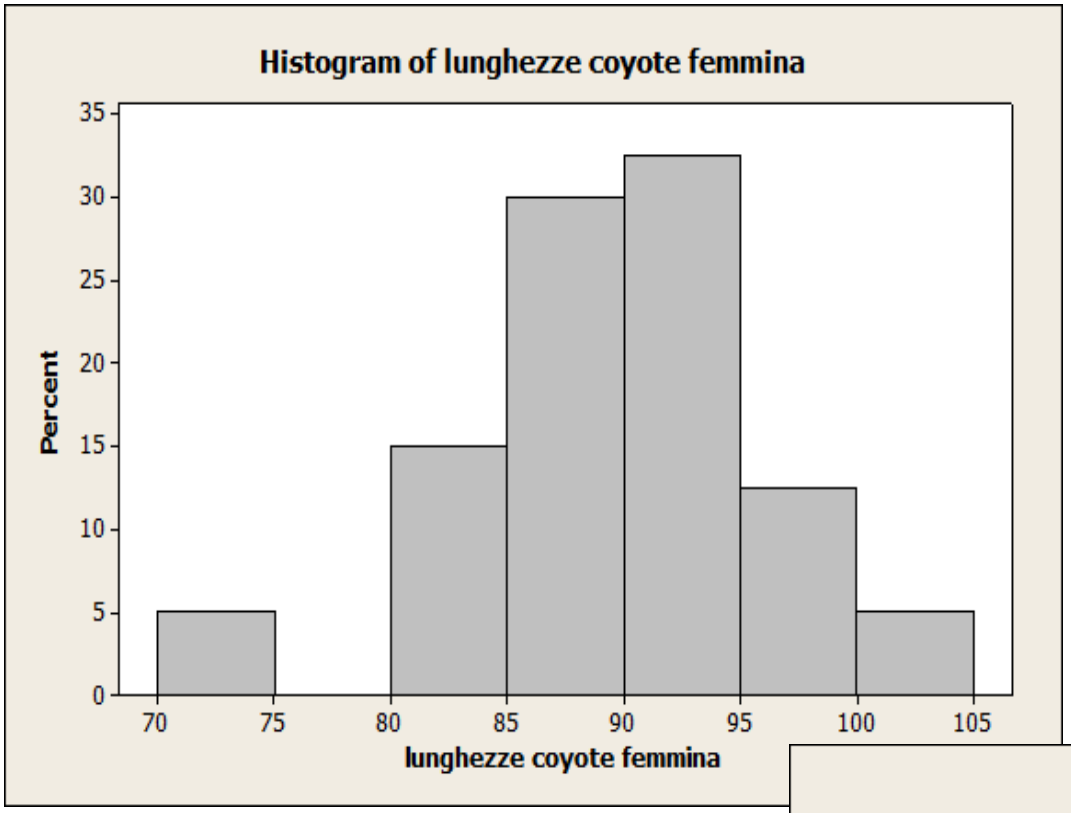
Ramo-foglia

7	13
7	
8	013444
8	556666778889
9	0001111233333
9	67778
10	12

L'altezza di ogni barra è uguale (o proporzionale) alla freq relativa percentuale della classe corrispondente



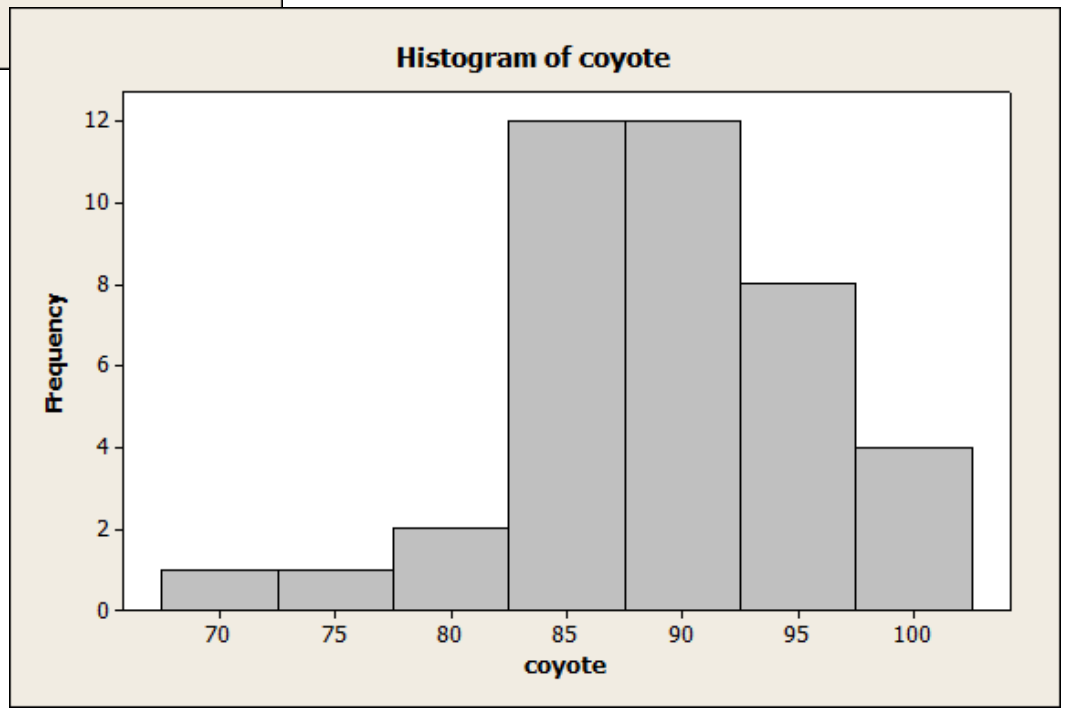
7 classi ampiezza 5 misure in cm



Frequenze relative percentuali (7 classi)

Frequenze assolute (7 classi)

Notare **la differenza** tra i 2 grafici



Istogrammi- Divisione in classi (coyote femmine)

7 Classi chiuse a sinistra Ampiezza classi = 5

70-75→70.0-74.9→ 70.0 ≤ lunghezze < 75.0

75-80→ 75.0-79.9→ 75.0 ≤ lunghezze < 80.0

……

100-105→100-104.9→100 ≤ lunghezze < 80.0

OPEN OFFICE→ Classi chiuse a destra (7 classi) Ampiezza classi = 5

70-75→70.1-75.0→ 70.0 <lunghezze ≤ 75.0

75-80→ 75.1-80.0→ 75.0< lunghezze ≤ 80.0

……

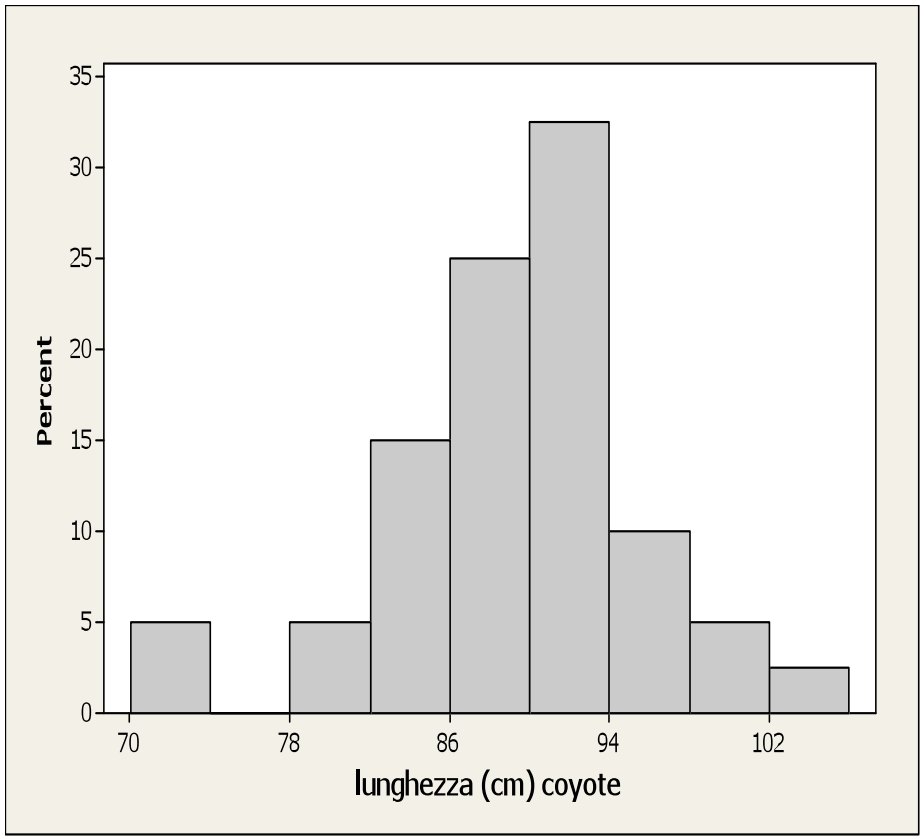
100-105→100-104.9→100 < lunghezze ≤ 105

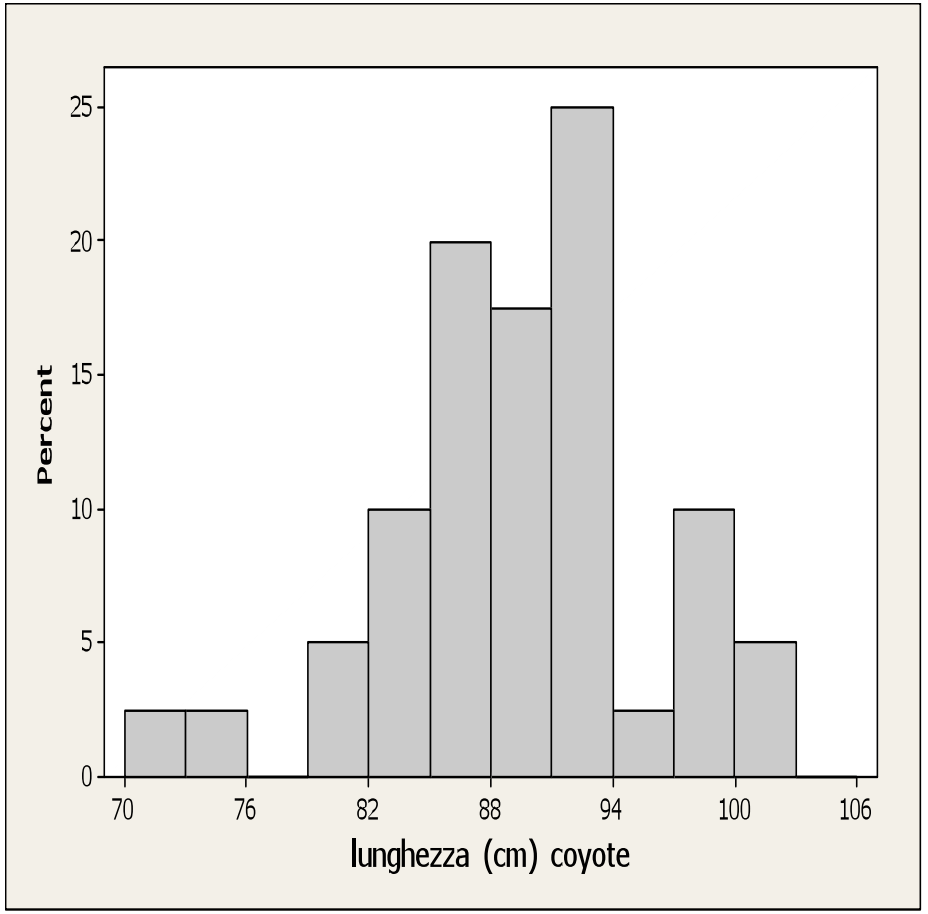
Istogrammi

- **Attenzione alla scelta delle classi**
- Scelte diverse delle classi possono portare a istogrammi di aspetto diverso.

- Usare le frequenze relative soprattutto se si confrontano 2 o più grafici

9 classi di ampiezza 4:
70-74-78-…-102-106





12 classi di ampiezza 3:
70-73-76-…..-103-106

L’istogramma

- Indicate la **scala e l’etichetta** per l’asse delle y
- Indicate la **scala e l’etichetta** per l’asse delle x
- Date un titolo all’istogramma

ESEMPIO

- Considerate le seguenti collezioni di dati (campioni) a due cifre:

1) 56 46 53 11 65 41 91 13 62 84 09 43 82 27 64 52 89 25 68 53 26 46 70

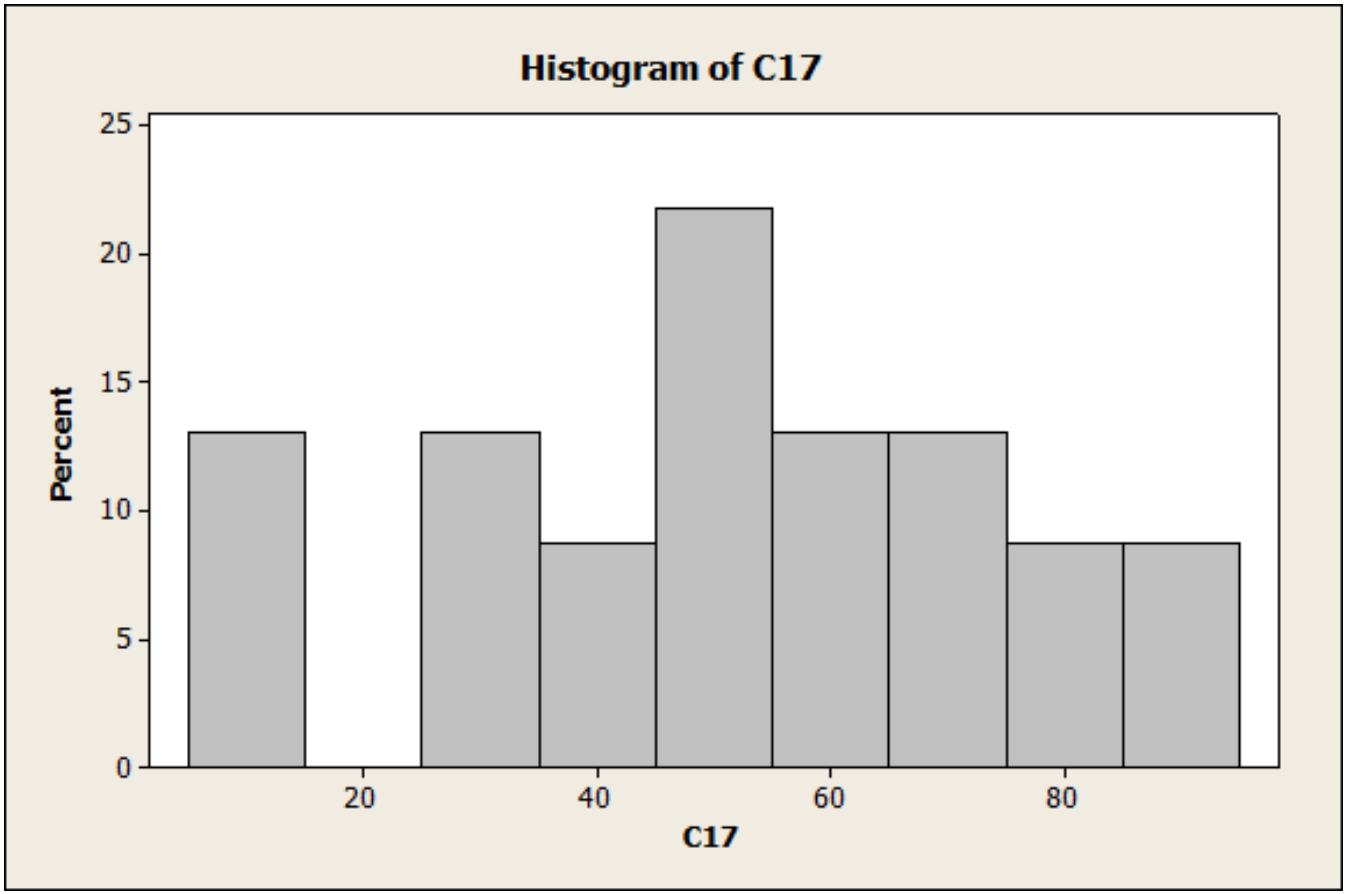
2) 94 83 68 17 40 59 35 26 41 86 84 12 36 78 93 40 29 64 94 78 65 66 67 20.

Nel diagramma ramo-foglia associato a queste collezioni di dati, le foglie corrispondenti al ramo 2 sono **??**

Con riferimento al 1° campione dell'es . precedente

Leaf Unit = 1.0

0	9
1	13
2	567
3	
4	1366
5	2336
6	2458
7	0
8	249
9	1



ISTOGRAMMI

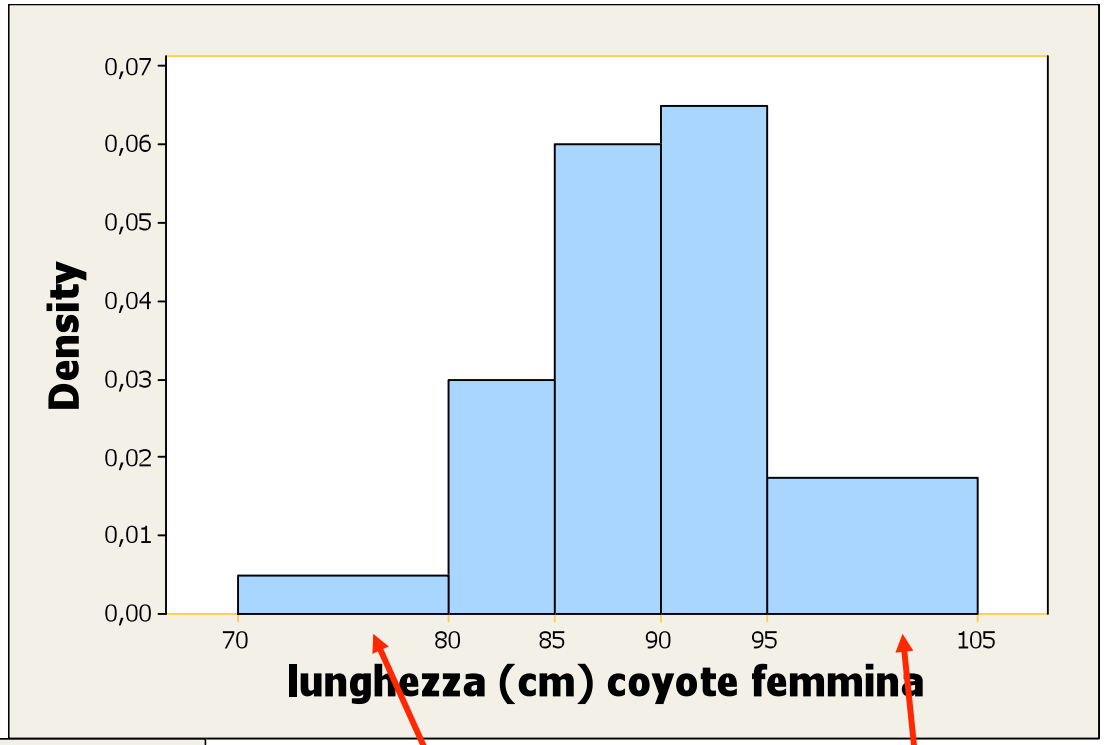
- A volte può essere opportuno considerare classi di ampiezze diverse tra loro.
- Si supponga di voler ripartire le lunghezze dei coyote in 5 classi raggruppando le lunghezze delle prime due classi e delle ultime due.

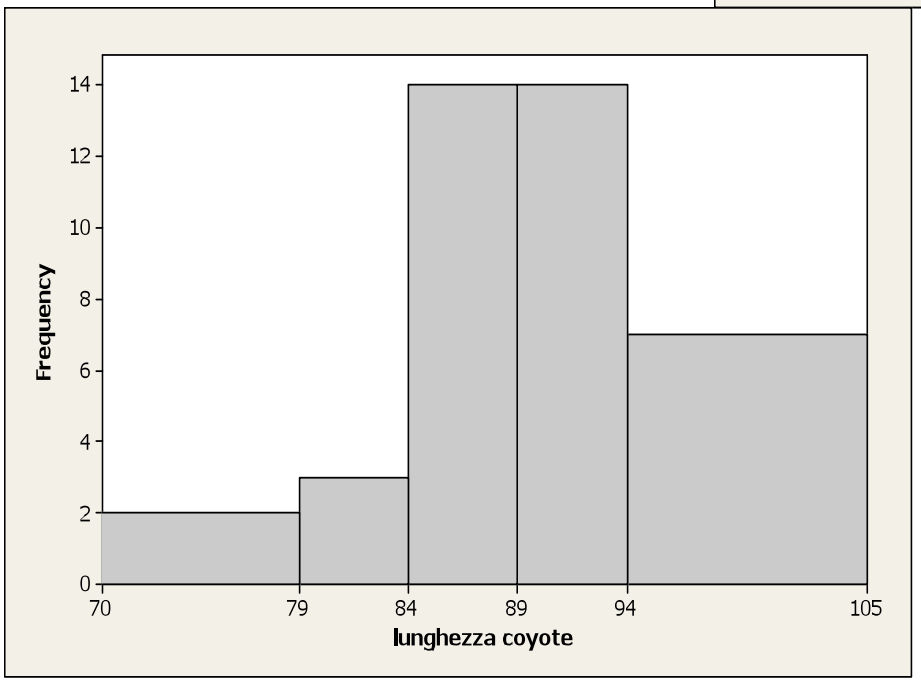
Classi	Frequenza	Frequenza relativa (nj/n)
70- 79	2	0.05
80- 84	6	0.15
85- 89	12	0.3
90- 94	13	0.325
95-105	7	0.175
Totale	40	1.00

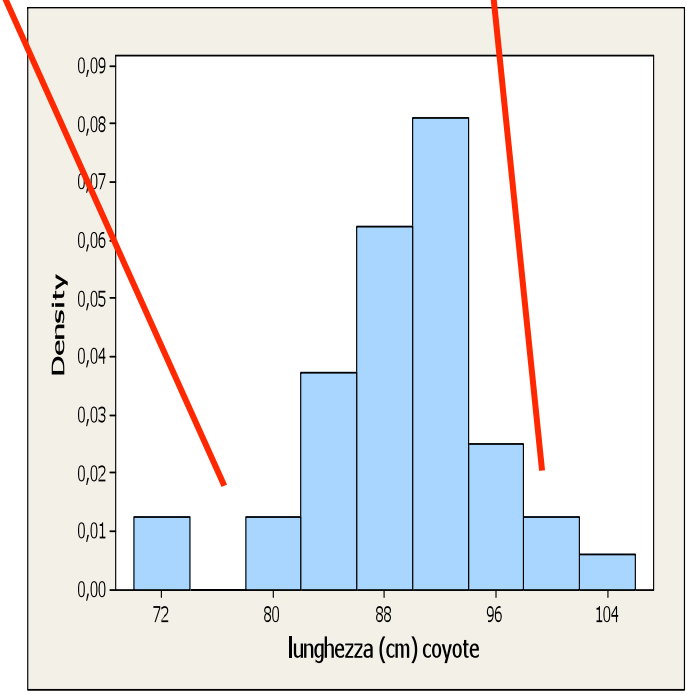
Istogrammi

- Quando le ampiezze delle classi sono diverse, per evitare distorsioni visive, attribuendo un valore sproporzionato alle barre con base più ampia, saranno **le aree** e non più le altezze delle barre ad **essere proporzionali alle corrispondenti frequenze relative.**
- In tal caso **l'altezza del rettangolo** corrispondente a una classe viene chiamata **densità.**
- Densità = $\frac{\text{frequenza relativa}}{\text{ampiezza classe}}$

La somma delle aree di tutti i rettangoli è pari ad 1

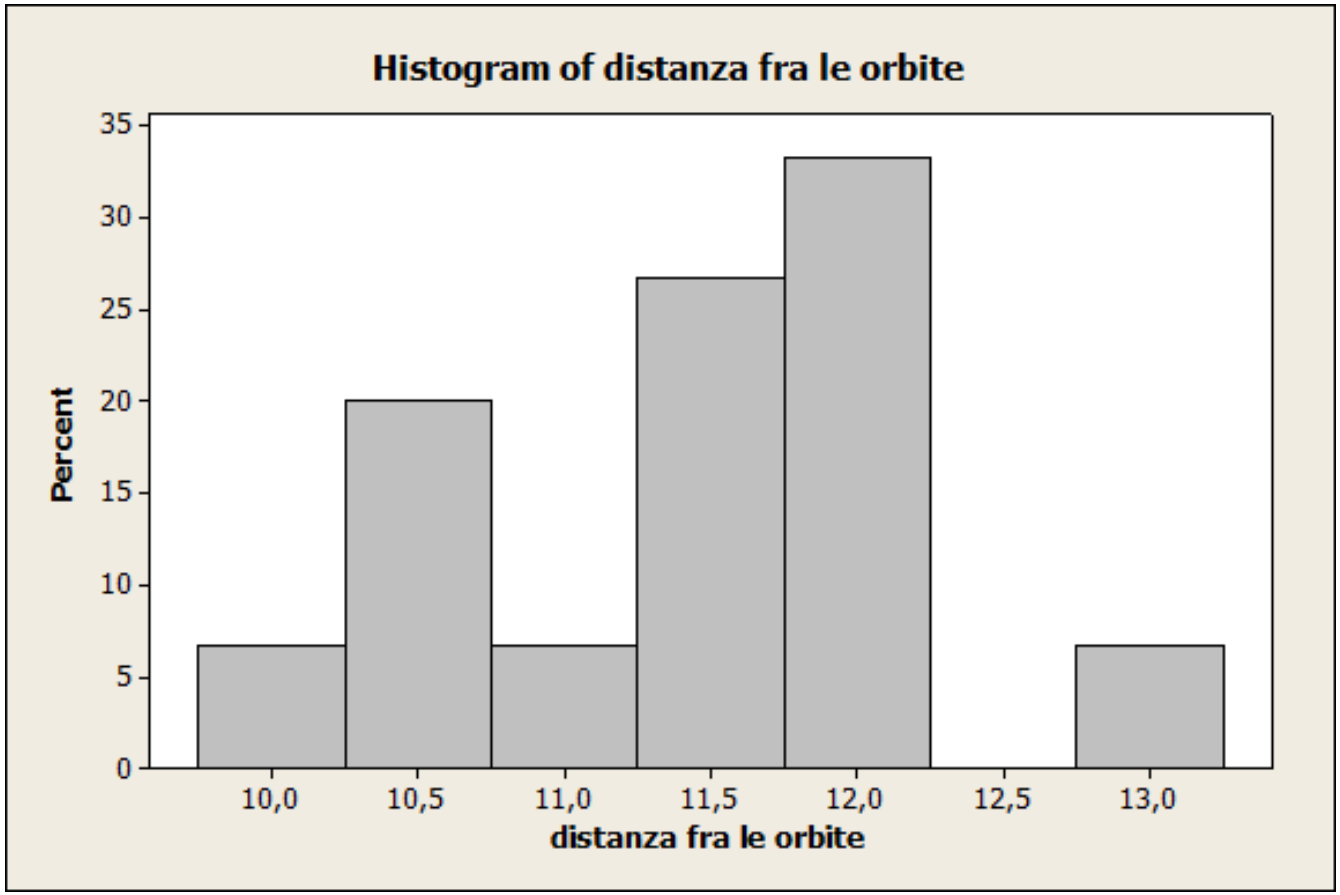






Misure (mm) della distanza tra le orbite di 15 piccioni domestici

	Valori ordinati
10,2	10,2
12,2	10,4
10,7	10,4
12,1	10,7
10,8	10,8
12,9	11,3
11,5	11,3
11,9	11,5
11,6	11,6
11,8	11,8
11,3	11,9
10,4	11,9
10,4	12,1
11,9	12,2
11,3	12,9

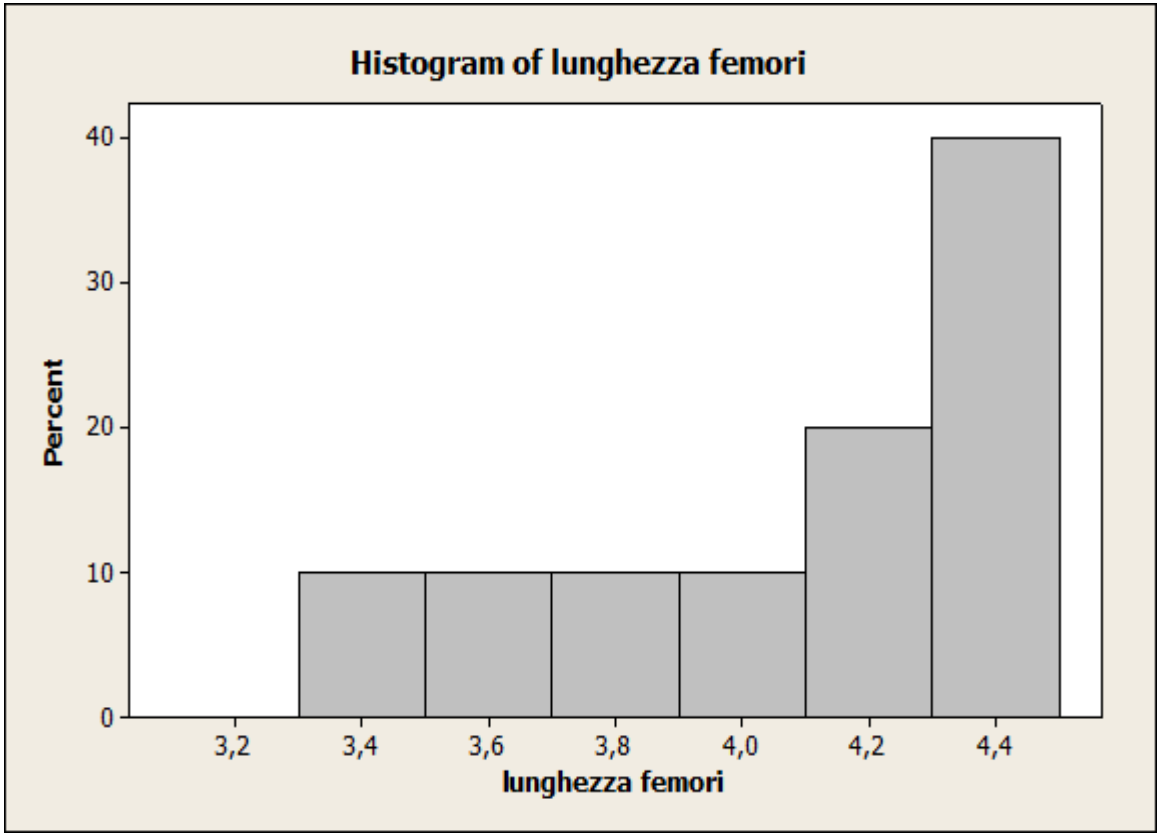


Minitab
Classi chiuse a sinistra

Classi: 9,75—10,25
10,25--- 10,75
…….
12,75---13,25

Lunghezza di **10 femori** dell'afide *Pemphigus* *(mm x 10^{-1})*

3,8
3,3
3,9
4,1
3,6
4,3
4,4
4,4
4,1
4,3



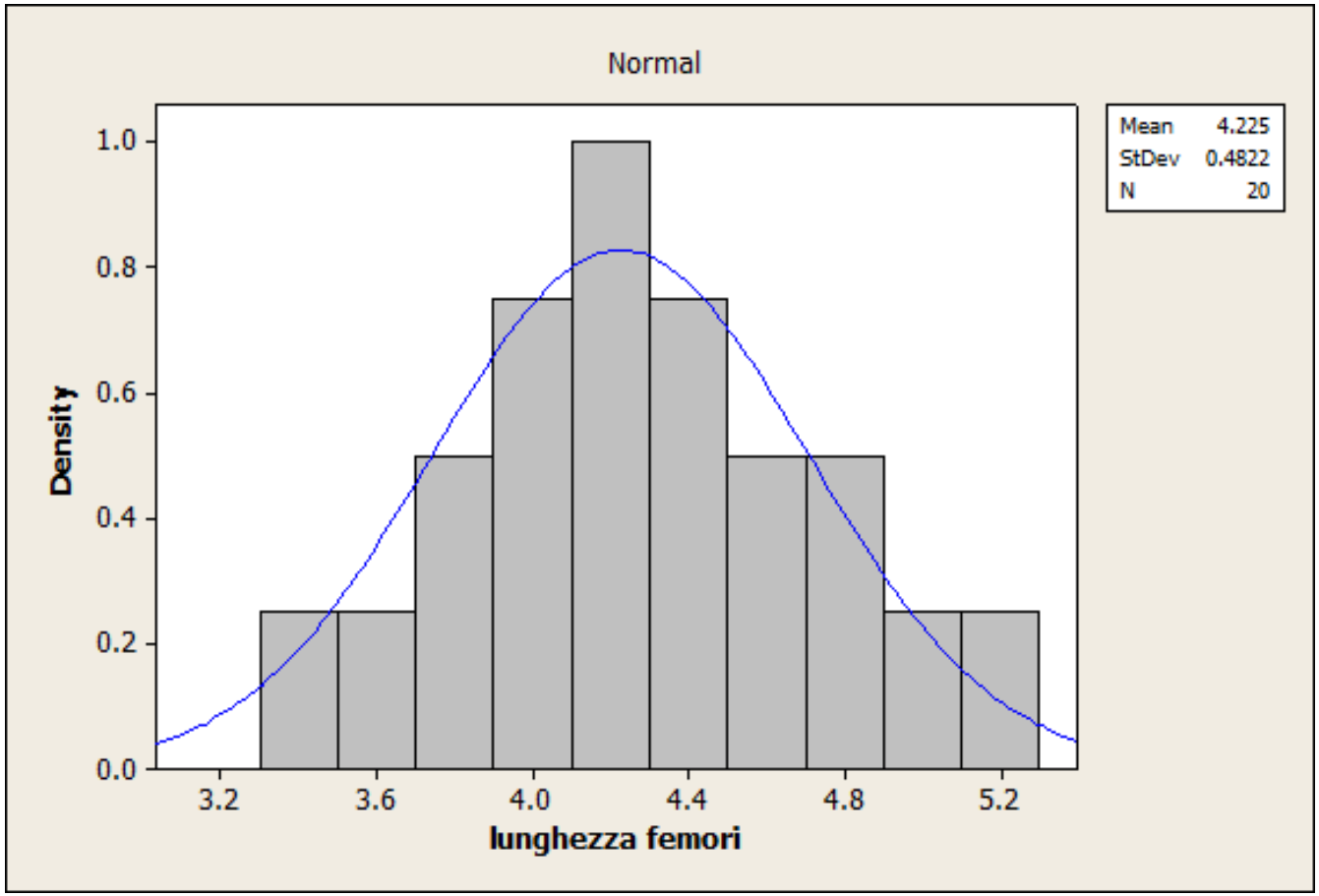
Mean	StDev	Min	Q1	Median	Q3	Max
4,020	0,368	3,300	3,750	4,100	4,325	4,400

Skewness	Kurtosis
-0,85	-0,08

3.8
3.3
3.9
4.1
3.6
4.3
4.4
4.6
4.1
4.3
3.9
3.7
4.0
5.2
4.9
4.8
4.8
4.6
4.1
4.1

Lunghezza di **20 femori** dell'afide *Pemphigus* *(mm x 10^{-1})*

Mean	StDev	Minimum	Q1	Median	Q3	Maximum
4.225	0.482	3.300	3.900	4.100	4.600	5.200



In questo es. si noti come all'aumentare del campione (da 10 a 20), aumentando le osservazioni, la forma del grafico diventa più regolare

Istogrammi

Suggerimenti:

- Un istogramma appropriato dovrebbe avere tra

  7 – 20 classi

- Gli istogrammi si usano quando la dimensione n del campione è abbastanza numerosa
- Gli istogrammi mostrano:

  la forma della distribuzione,

  outlier, valori raggruppati, interruzioni nei valori.

Diagrammi a segmenti: variabili discrete

Esempio . Numero di protozoi contati in un CCS di 33 unità campionarie prelevate da uno stagno.

163 165 165 165 166 166 166 166 168 168
168 168 169 169 169 169 169 169 169 169
171 171 171 171 171 171 172 172 172 174
174 175 175

Diagrammi a segmenti

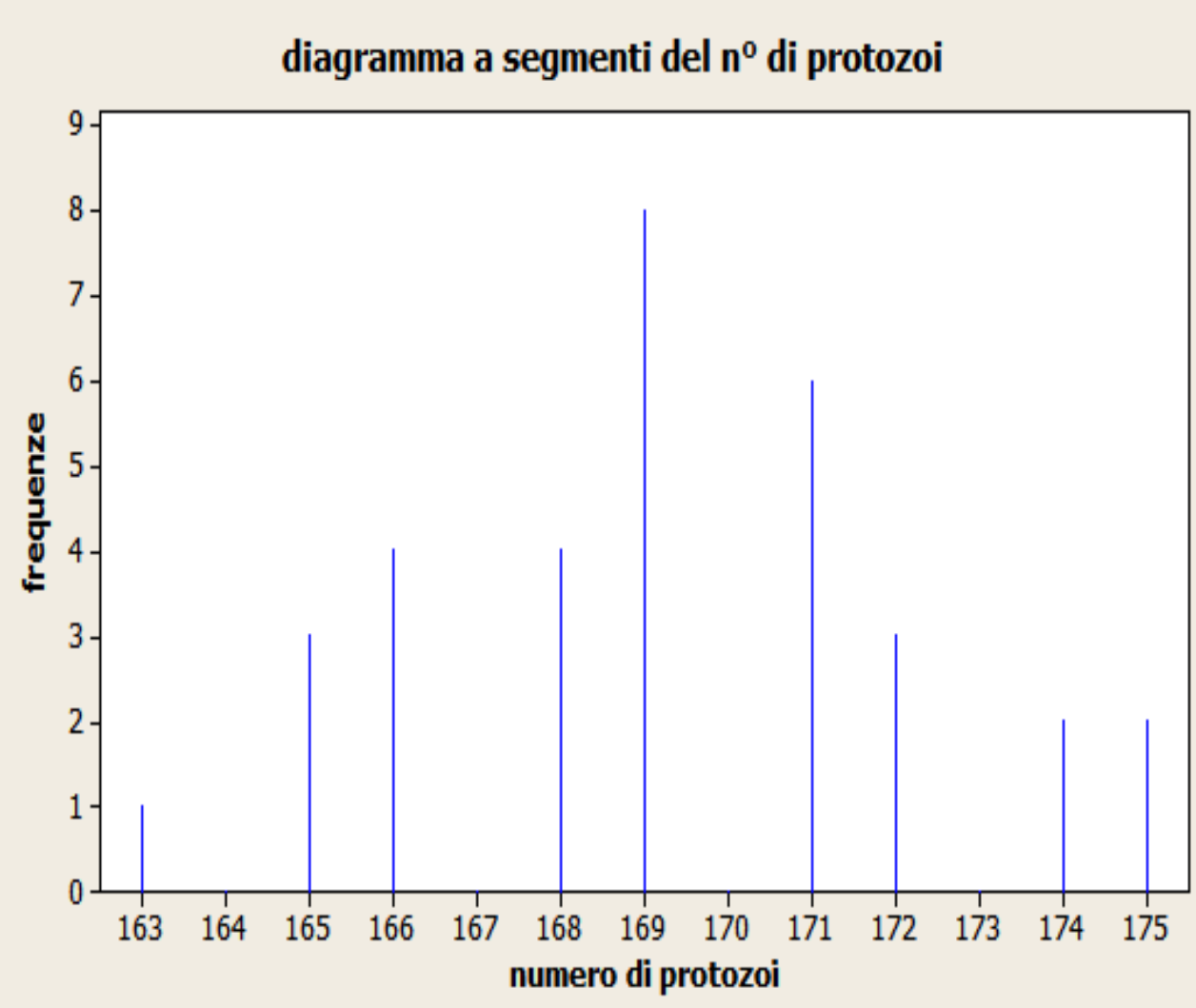
Distribuzione delle frequenze

Numero di protozoi	Frequenza
163	1
165	3
166	4
168	4
169	8
171	6
172	3
174	2
175	2
	Tot. 33

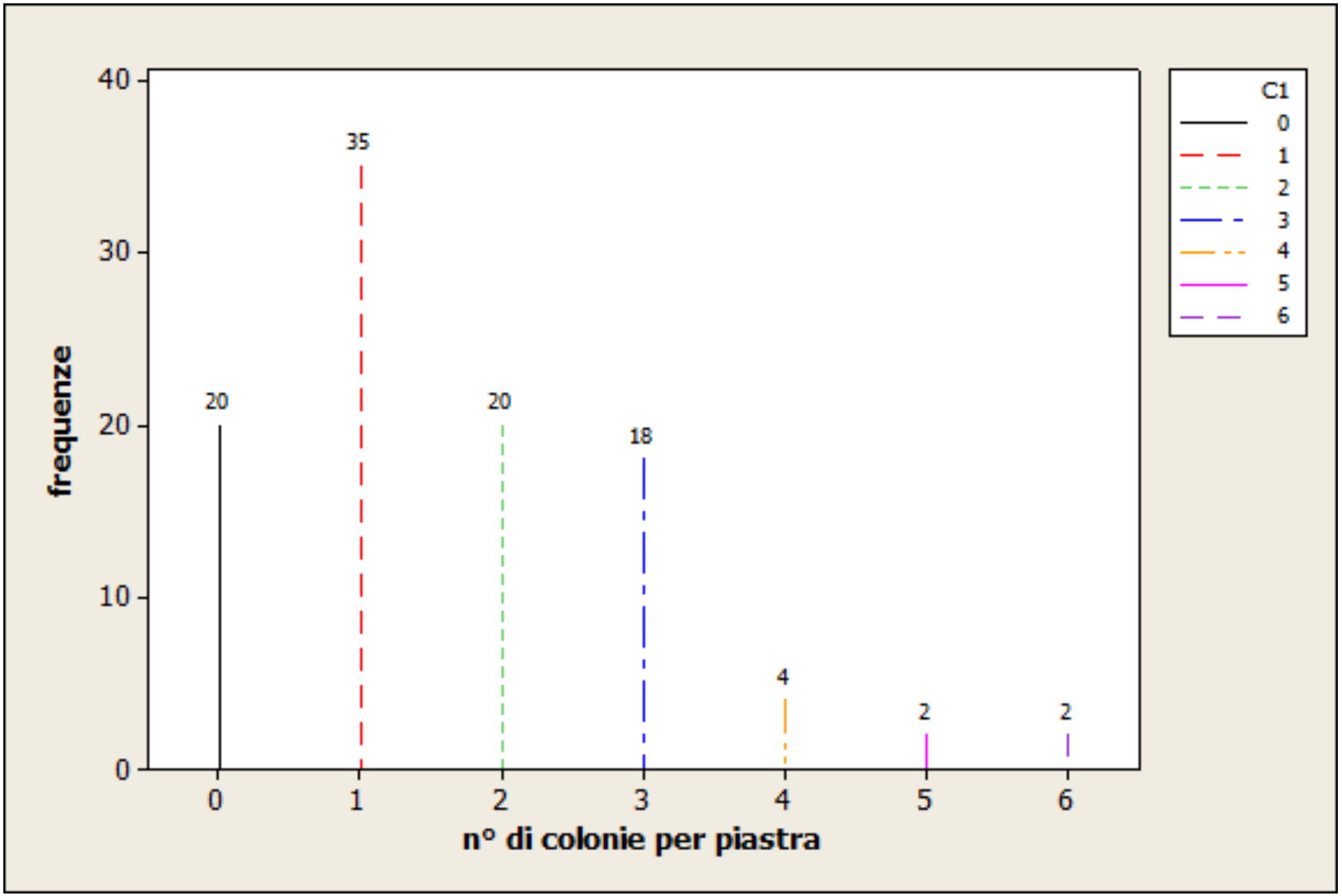
Diagramma ramo-foglia

Diagramma a segmenti

163 0
164
165 000
166 0000
167
168 0000
169 00000000
170
171 000000
172 000
173
174 00
175 00



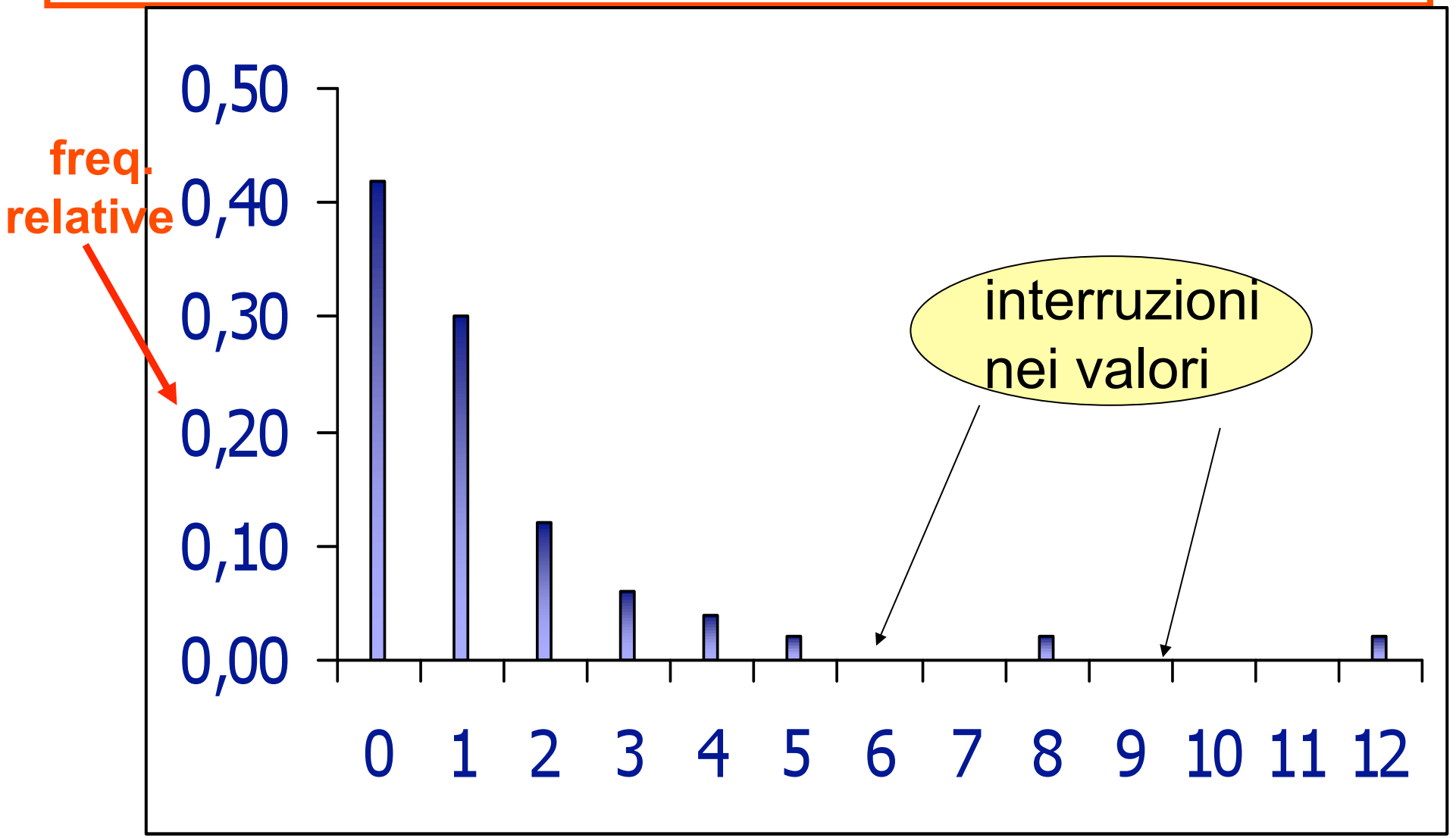
Frequenza del n° di colonie di microrganismi su 90 piastre: diagramma a segmenti (v. quantitativa discreta)



Variabile discreta: numero di orchidee in un CCS di 50 quadrati scelti a caso (Tab. frequenze)

n° orchidee	frequenza assoluta	frequenza relativa	frequenza rel.cumulata
0	21	0.42	0.42
1	15	0.30	0.72
2	6	0.12	0.84
3	3	0.06	0.90
4	2	0.04	0.94
5	1	0.02	0.96
8	1	0.02	0.98
12	1	0.02	1.00
totali	50	1.00	

diagramma a segmenti del n° di orchidee



Come si interpretano i grafici ramo-foglia, gli istogrammi e i diagrammi a segmenti?

- Cercare di individuare un andamento generale per i dati osservati
- Individuare eventuali scostamenti da tale andamento
- Individuare le caratteristiche più significative della distribuzione dei dati quali la forma, il centro, e la dispersione
- Individuare eventuali outlier (osservazioni estreme). Possono essere **errori o osservazioni interessanti/inusuali**

Variabili quantitative-Le caratteristiche principali di una distribuzione

1. **Forma**

unimodale bimodale multimodale

simmetrica coda a destra coda a sinistra

2. **Centro**
3. **Dispersione**

Numero di piccoli sopravissuti in 22 nidi
es. di distribuzione bimodale

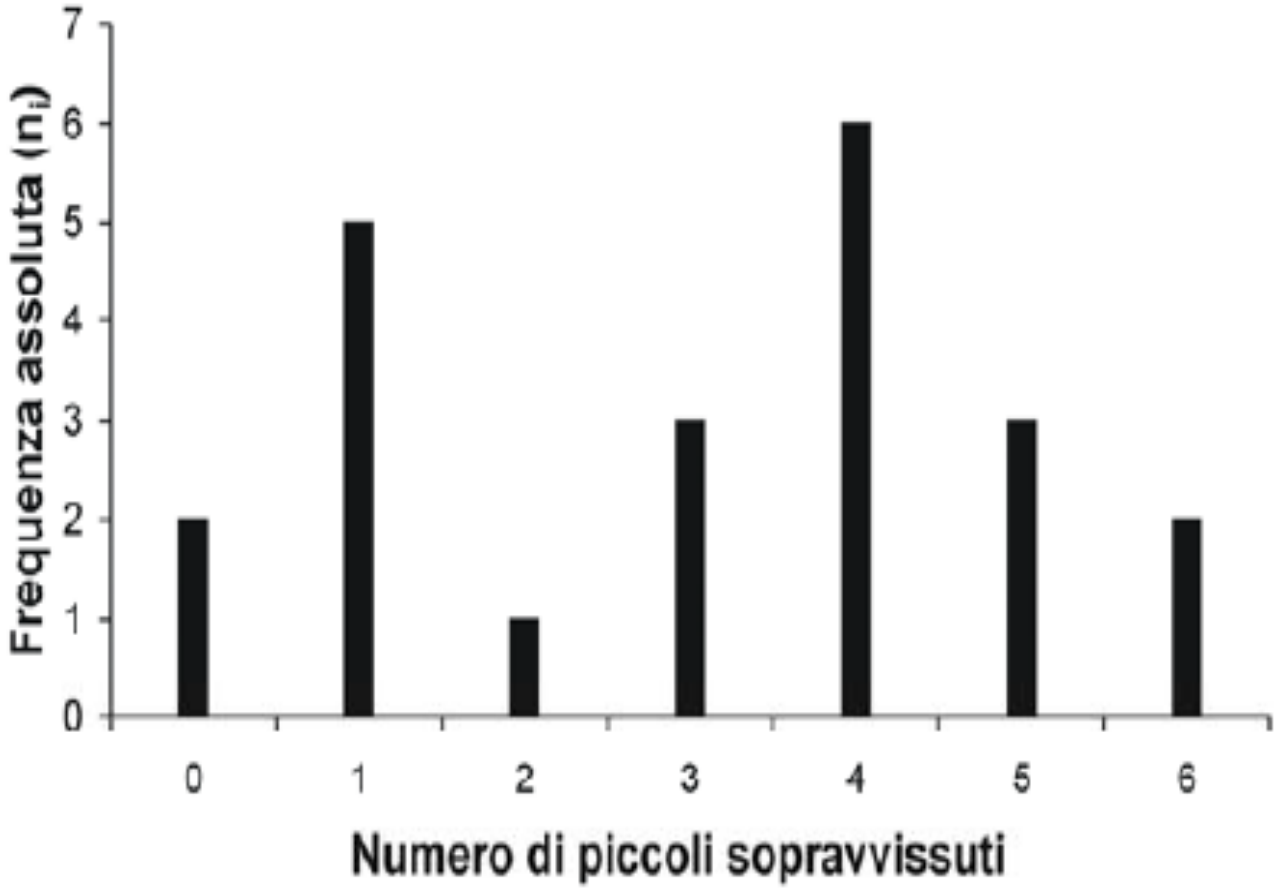


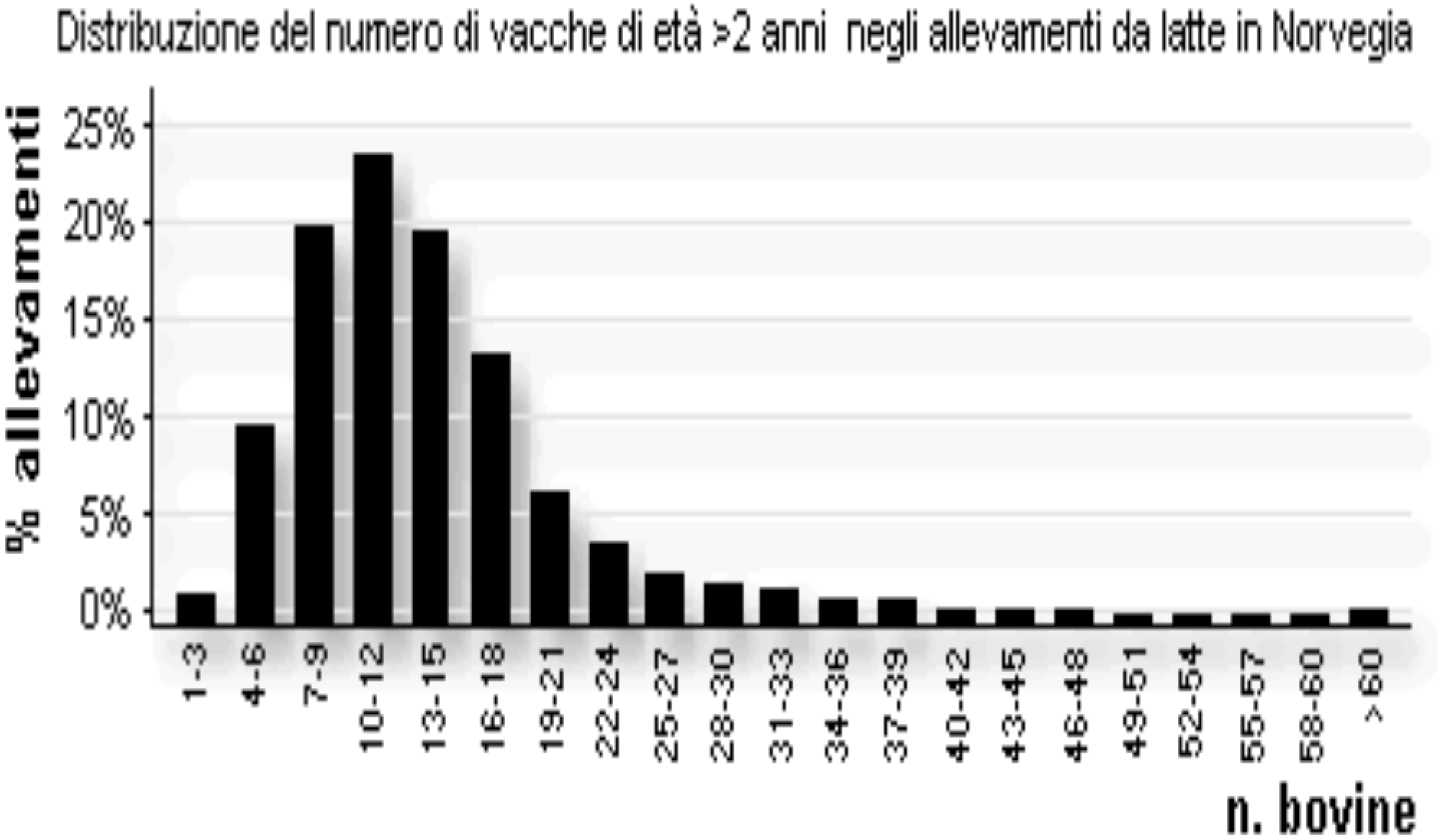
Figura 3.2. Esempio di una distribuzione bimodale.

Numero di piccoli sopravissuti in 22 nidi: es. di distribuzione bimodale

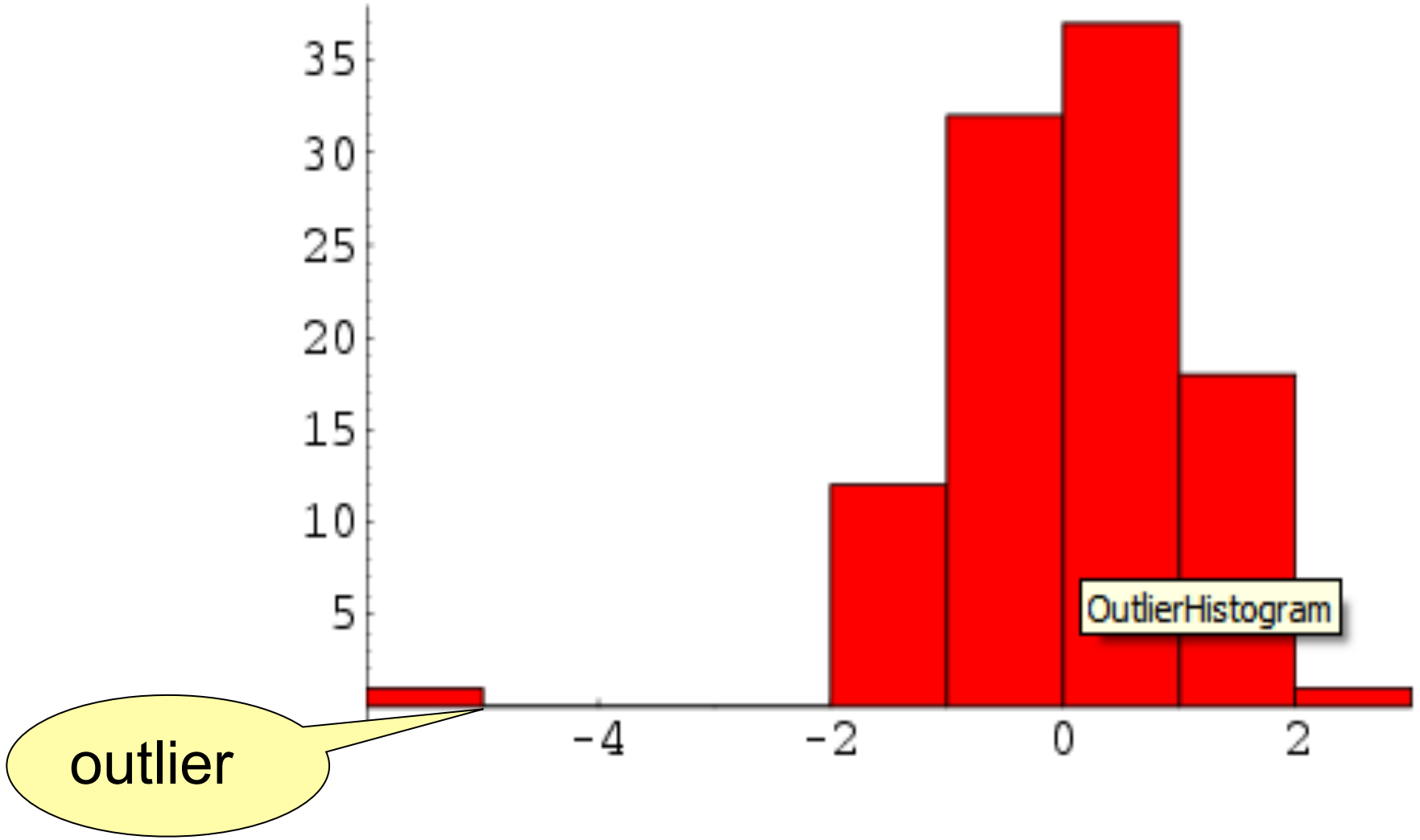
- Di solito in biologia le distribuzioni sono unimodali, ma, se per esempio, la distribuzione del numero di piccoli sopravvissuti per nido fosse di tipo bimodale, potremmo pensare che il campione analizzato si riferisca a due gruppi ben distinti di genitori che si differenziano per il numero medio di uova deposte o per l'abilità di allevare i piccoli.

- Tali gruppi potrebbero corrispondere a due specie diverse difficilmente distinguibili, oppure a due classi di età.

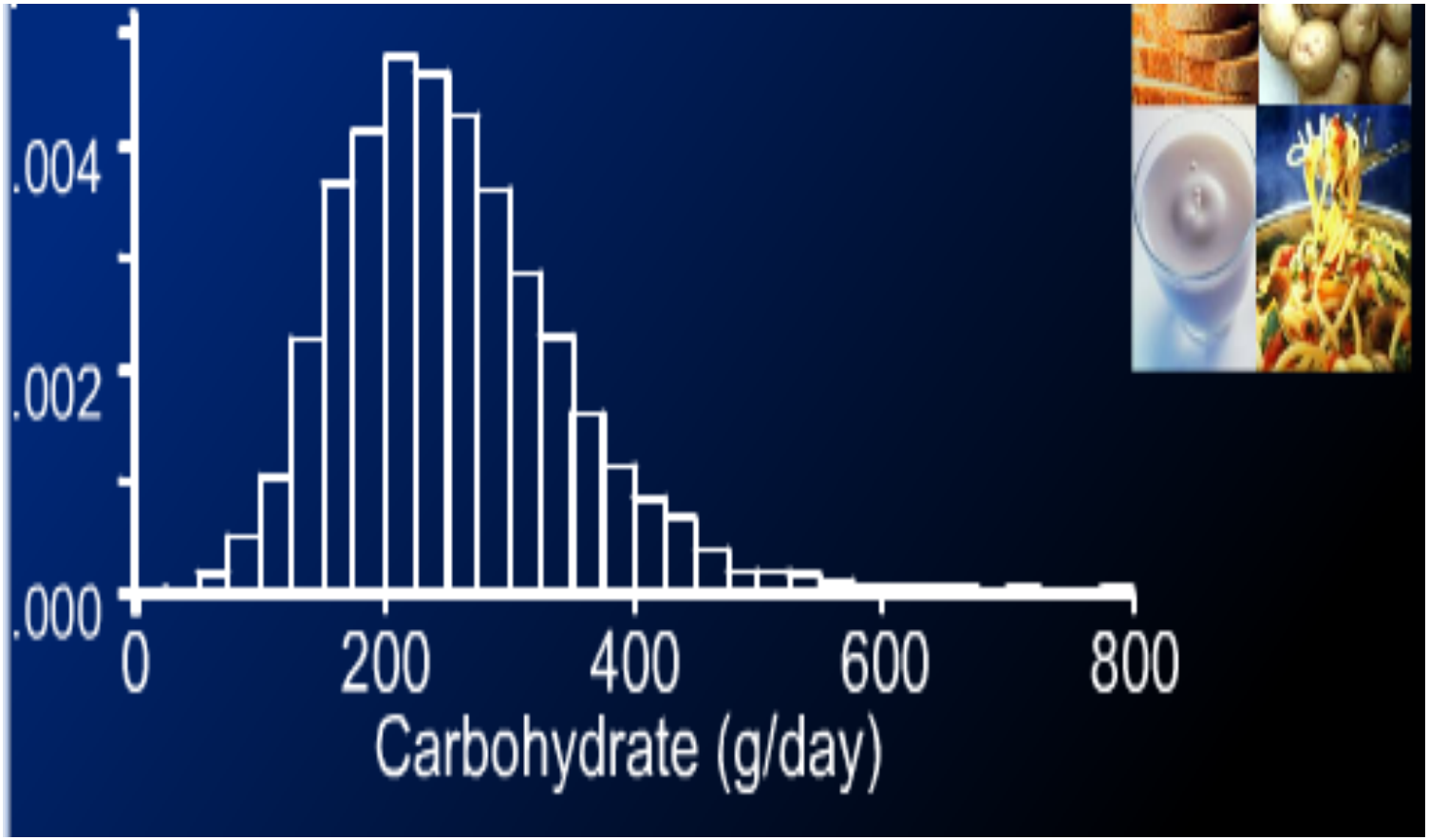
Distribuzione asimmetrica a destra



Distribuzione con outlier (valori estremi)



Assunzione media giornaliera di carboidrati nella dieta di 5929 individui



Asimmetria a destra.

Grande variabilità: il consumo più grande è più di 10 volte il consumo più piccolo

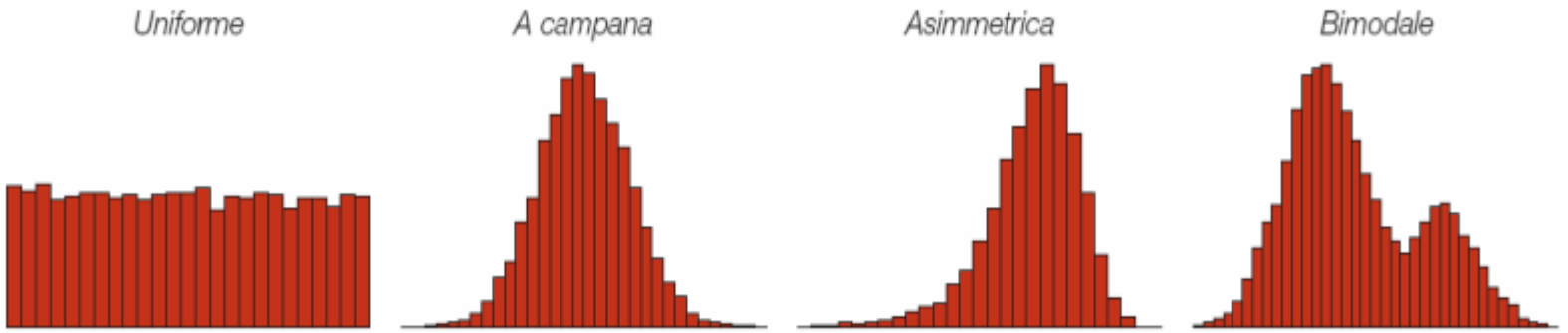


Figura 2.1-3
Alcune possibili forme di distribuzioni di frequenza.

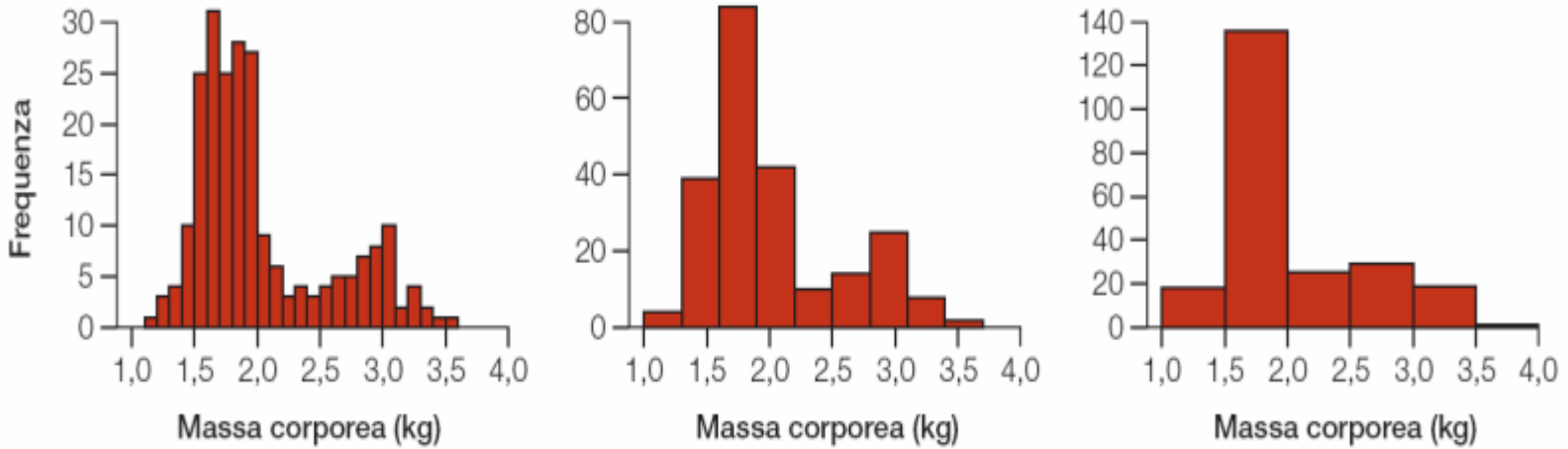


Figura 2.1-4
Massa corporea di 228 femmine di salmone rosso campionate dal Pick Creek in Alaska (Hendry *et al.*, 1999). In ciascun caso sono presentati gli stessi dati, ma le ampiezze degli intervalli sono diverse: 0,1 kg (a sinistra), 0,3 kg (in centro) e 0,5 kg (a destra).

Grafici per variabili qualitative

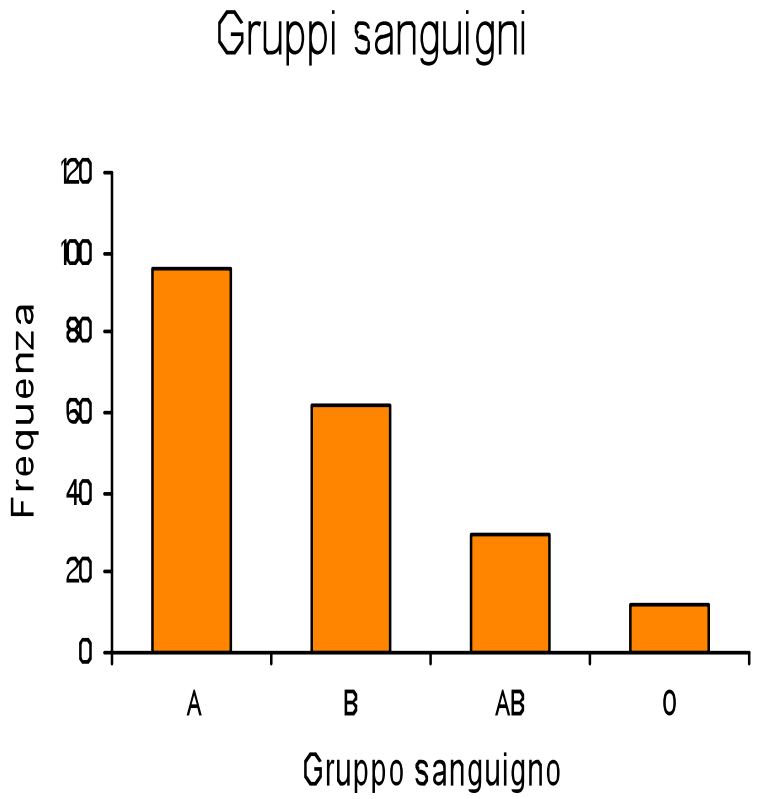
- Grafici a barre
- Grafici a torta

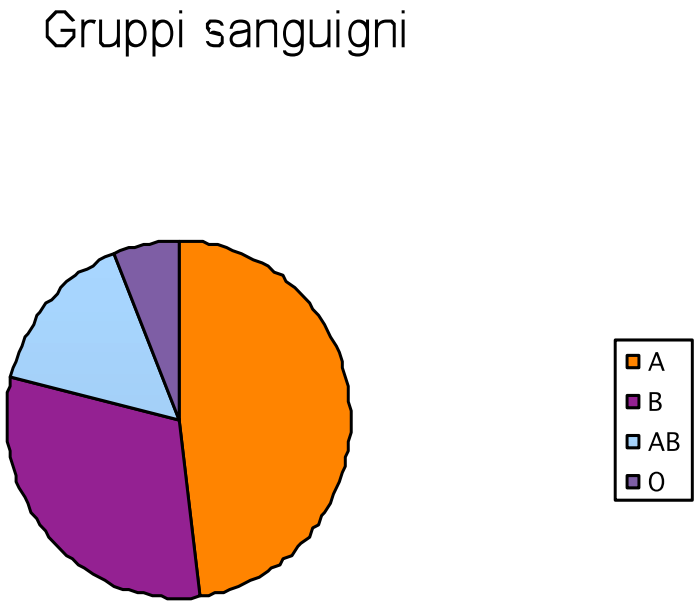
Grafici a barre

Esempio . In un laboratorio sono state eseguite 200 analisi e sono stati osservati i gruppi sanguigni

Gruppo	Frequenza	Frequenza relativa %
0	96	48%
A	62	31%
B	30	15%
AB	12	6%

- grafici a barre e a torta per l'esempio





Grafici a barre e grafici a torta

- Nei grafici a barre ogni frequenza è rappresentata da una barra (rettangolo).
- I rettangoli hanno la stessa base e l'altezza è proporzionale alla frequenza.
- I grafici a barre sono **diversi dagli istogrammi**.
- Nel grafico a torta si visualizzano le diverse parti in cui è stato diviso un tutto.
- Le ampiezze dei settori circolari sono proporzionali alle corrispondenti frequenze percentuali.

Riassunto

- Un insieme di dati contiene informazioni su un certo numero di unità.
- Per ogni unità i dati riportano valori riferiti a una o più variabili.
- Variabili qualitative o quantitative.
- L'analisi esplorativa dei dati si serve di grafici e indici numerici per descrivere il comportamento delle variabili in un insieme di dati.
- La distribuzione di una variabile descrive i valori che questa assume e la frequenza con cui li assume.

- Distribuzione delle frequenze, delle frequenze relative e delle frequenze relative cumulate.
- Per descrivere una distribuzione è utile cominciare con un grafico.
- Nell'analisi di un grafico o distribuzione cercare l'andamento generale (forma, centro, dispersione) e le eventuali deviazioni degne di nota.
- Gli outlier sono osservazioni che si discostano molto dal modello generale della distribuzione.

ESERCIZIO Nella seguente distribuzione di frequenze, il valore nella casella mancante è :

a) 34

b) 40

c) 32

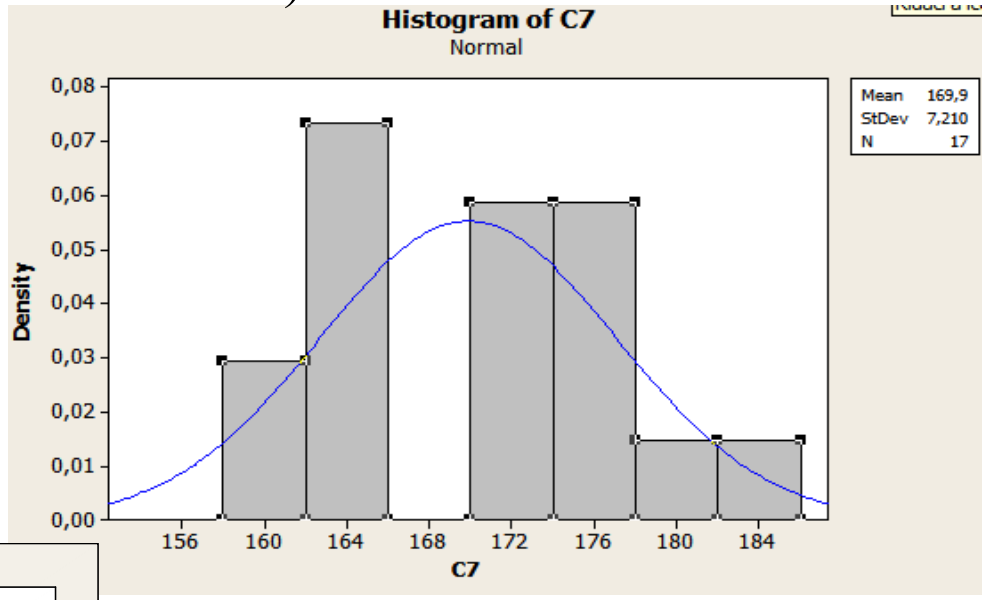
d) nessuno dei precedenti

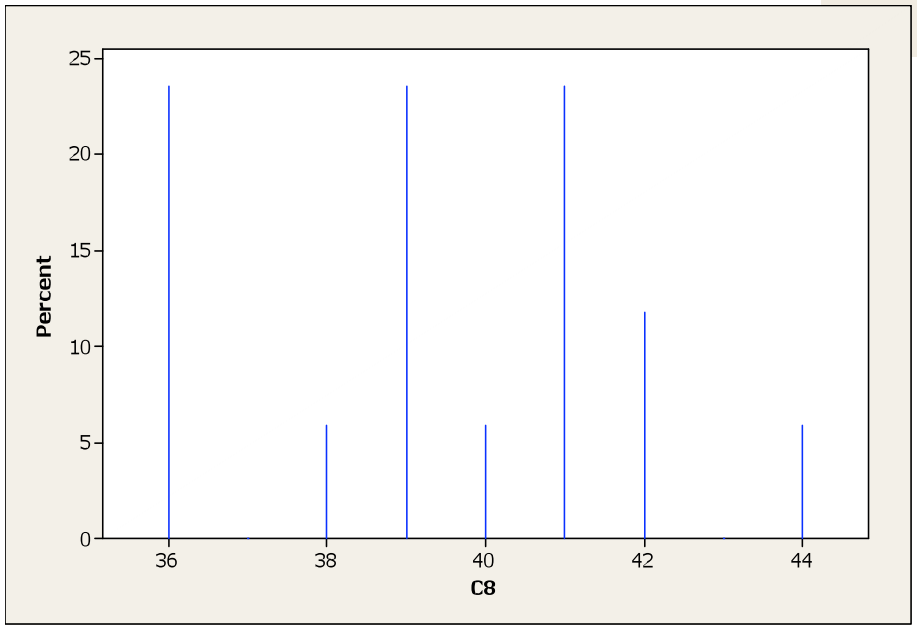
Modalità	**Frequenze assolute**	**Frequenze percentuali**
A	**13**	**10,8%**
B	**27**	**22,5%**
C		**26,7%**
D	**48**	**40%**

Nella tabella che segue sono riportati alcuni dati biometrici di 17 individui.

Genere1	Altezza cm	Num scarpe	Colore occhi
M	185	39	marrone
F	165	39	marrone
F	165	40	marrone
F	170	36	marrone
F	175	42	verde
M	176	42	marrone
F	171	38	marrone
M	170	41	marrone
F	163	39	verde
F	170	39	marrone
F	160	36	marrone
F	174	41	blu
F	164	36	blu
M	177	41	marrone
M	180	44	verde
F	161	36	marrone
M	162	41	marrone

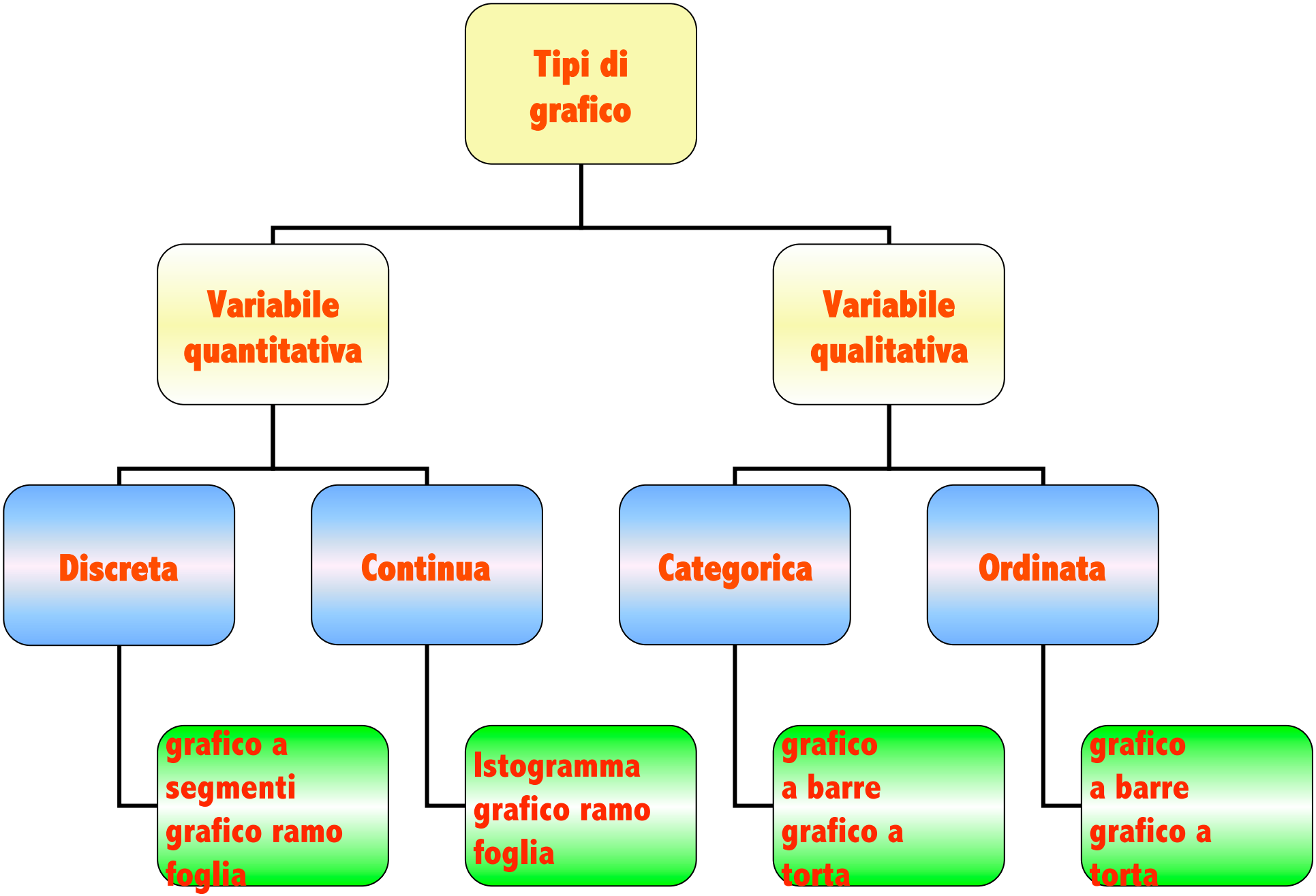
Disegnate i grafici appropriati per ciascuna delle seguenti variabili altezza, numero di scarpa, e colore degli occhi (per l'altezza considerate classi di ampiezza pari a 4 cm a partire dal valore 158)





ESERCIZIO: Un insieme è costituito da 200 osservazioni. Se 50 osservazioni sono comprese nella classe 10-15, qual è la densità associata alla classe?

(a) 0,05
(b) 0,25
(c) 5
(d) 50
(e) nessuna delle precedenti


Tipi di grafico
Variabile quantitativa
Variabile qualitativa
Discreta
Continua
Categorica
Ordinata
grafico a segmenti
grafico ramo foglia
Istogramma
grafico ramo foglia
grafico a barre
grafico a torta
grafico a barre
grafico a torta

Come rappresentare numericamente una distribuzione?
Indici riassuntivi

- **Indici di posizione**

  media, mediana, moda, percentili

- **Indici di variabilità o dispersione**

  intervallo di variazione, varianza (o deviazione standard),

  coefficiente di variazione, box-plot

- **Indici di forma**

  curtosi, indice di simmetria

Indici di posizione centrale

La media (aritmetica)

Esempio 7

I diametri delle cappelle di un fungo commestibile in un campione di 6 sono:

9.3 cm 7.8 cm 6.2 cm 7.0 cm 8.3 cm 9.9 cm

la **media (aritmetica)** dei diametri, ossia il diametro medio, pari a **8.08 cm**, è dato da

$$media = \frac{somma\ \ dei\ \ valori\ \ delle\ \ osservazioni}{numero\ \ delle\ \ osservazioni}$$

La media

- Se i dati sono rappresentati con una distribuzione di frequenze, ossia la modalità (il valore) $\mathbf{x_j}$ compare con la frequenza $\mathbf{f_j}$ (j = 1, 2, …, k) si può usare la formula:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_k f_k}{f_1 + f_2 + \ldots + f_k} = \frac{\sum_{i=1}^{k} x_i f_i}{n}$$

La media

Esempio 8. Il numero di formiche del legno catturate in 7 trappole, poste di notte in un bosco, è:

25 4 12 9 15 8 202

Qual è il numero medio di formiche per ogni trappola?

$\bar{x}$ = **39.3**. Tale valore è più grande di 6 delle 7 osservazioni ed è molto più piccolo dell'ultima. La media utilizza il valore effettivo di ogni osservazione, perciò potrà essere distorta da un singolo valore eccezionale (**non è robusta**).

La mediana

- ✓Nell'esempio precedente la **mediana** può essere un indice di centralità più appropriato.
- ✓La **mediana** è il valore di mezzo in un insieme di osservazioni che sono state ordinate in ordine crescente.
- ✓ Quindi metà delle osservazioni sono più piccole e metà più grandi della mediana.

La mediana

- ✓ Ordiniamo i dati sulle formiche:

  4 8 9 12 15 25 202

  mediana

- ✓ La **mediana** è più **robusta** della media, ossia non è influenzata da singoli valori estremi.
- ✓ Nell'esempio la mediana vale 12 qualunque sia il valore della settima osservazione (20, 202 o 2002).

La mediana

Come si calcola la mediana?

1 4 7 **9** 10 12 14

med = 9

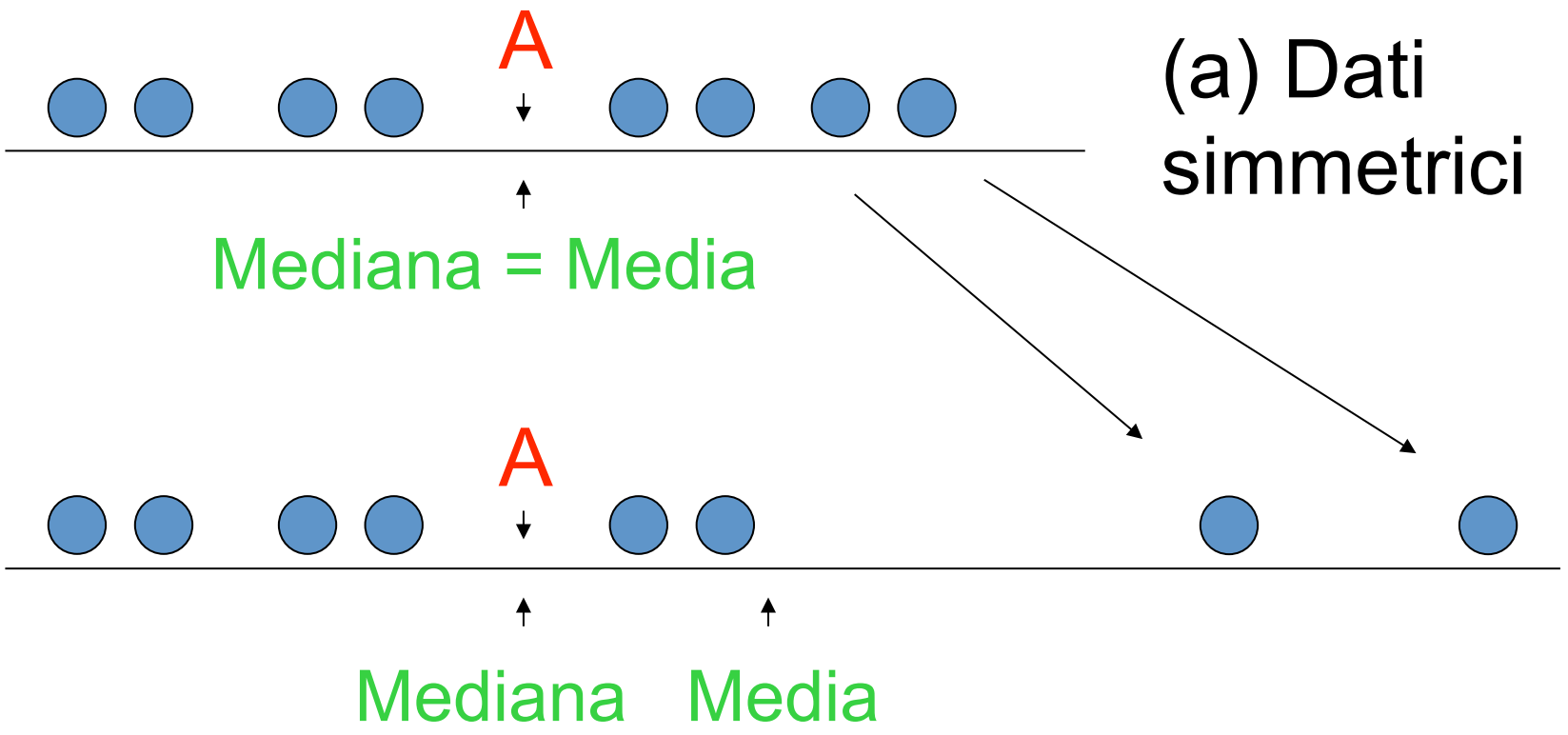
11 13 15 **16** | **19** 21 22 25

med = 17.5

Se le osservazioni sono in numero **dispari** la mediana è l'elemento che occupa il posto centrale

Se le osservazioni sono in numero **pari** la mediana è la semisomma dei due elementi di posto centrale

Media contro mediana

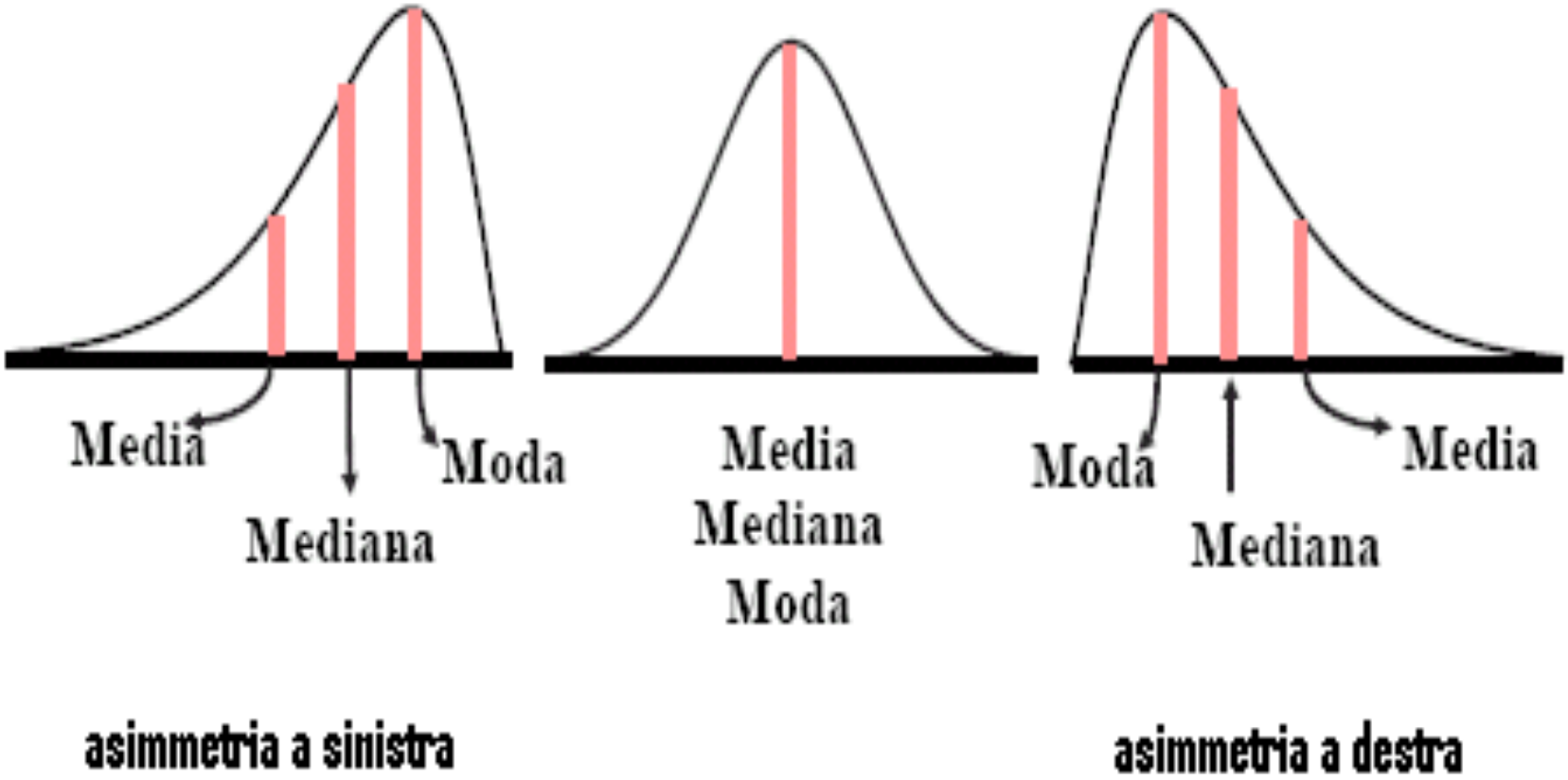


(b) Le due osservazioni con il valore più grande si sono spostate a destra (dati asimmetrici)

La moda

- In una distribuzione di dati la moda o (classe modale) è il valore che si verifica con maggior frequenza o (la classe che contiene il maggior numero di osservazioni).
- Se due (più di due) valori si verificano con una maggior frequenza la distribuzione è detta bimodale (multimodale).
- La moda è l'unica misura di centralità che può essere usata con dati qualitativi.

Relazioni tra media mediana e moda



Le misure di posizione centrale

Tra moda, mediana, e media quale scegliere per rappresentare la nostra distribuzione?

Moda: è sempre calcolabile, ma è poco potente dal punto di vista informativo.

Mediana: è calcolabile soltanto per caratteri almeno ordinabili e trascura l'informazione relativa alla grandezza quantitativa dei dati. Ha però il vantaggio di non essere influenzata dai dati estremi.

Media: è calcolabile soltanto per caratteri quantitativi, è la più informativa, ma è influenzata dai dati estremi.

Moda, mediana, media hanno le stesse unità di misura delle osservazioni individuali.

Misure di posizione "**non centrale**"
Quartili, Percentili

- I **quartili**, rispettivamente, primo quartile Q_1, mediana Q_2 e terzo quartile Q_3 dividono la distribuzione dei dati ordinati in 4 parti uguali.
- Il **primo quartile** è la mediana di tutte le osservazioni la cui posizione è inferiore alla posizione della mediana.
- Il **terzo quartile** è la mediana delle osservazioni con posizione superiore.
- Nota: nella letteratura vengono date diverse definizioni di quartili, ma l'idea è sempre la stessa. Anche Excel usa una definizione diversa da quella che è stata data qui.

I quartili

Esempio 10

11	13	15	16		19	21	22	25

$Med = Q_2 = 17.5$

$Q_1 = 14$

$Q_3 = 21.5$

Q_1 lascia alla propria destra il 75% dell'intera distribuzione.Q_3 lascia alla propria destra il 25% della distribuzione.

I percentili

- I percentili sono quei valori che dividono la distribuzione dei dati ordinati in 100 gruppi di uguale numerosità, dove ogni gruppo contiene circa l'1% di tutti i valori.

I percentili

- Il p-percentile è quel valore tale che il p percento delle osservazioni cade fino a quel valore compreso; il resto delle osservazioni è maggiore
- Es. livelli di cotinina nella saliva (nmol/l) dopo aver fumato

data	count	cumcnt	percent	cumperc
18	1	1	14,29	14,29
33	1	2	14,29	28,57
58	1	3	14,29	42,86
67	1	4	14,29	57,14
73	1	5	14,29	71,43
93	1	6	14,29	85,71
147	1	7	14,29	100,00

1/7 = 0,1429 → 14,29 è la frequenza rel. percentuale di ogni osservazione

Percentili

- n=7 osservazioni
- Si vuole calcolare il 50mo percentile -> si sceglie l'osservazione campionaria che occupa il posto:

  p *x* (n+1) = **0.50** x (7+1)=4 ->**67**
- Si vuole calcolare il 20mo percentile -> si sceglie l'osservazione campionaria che occupa il posto:

  p x (n+1) = **0.20** *x* (7+1)=1.6 -> **25.5** è l'osservazione a metà circa tra la prima e la seconda.

Qual è il 10mo percentile? Qual è il 1° percentile?

Percentili

- 10-mo percentile:
- 0.10*(7+1)=0.8 -> approssimativamente la **prima osservazione** del campione, ossia 18

- 1° percentile:
- 0.01*(7+1)= 0.08 →**nessuna osservazione** del campione

ci sono diversi metodi che, dato un percentile, calcolano il corrispondente valore campionario.

I percentili : operazione inversa

- Il p-percentile è quel valore tale che il p percento delle osservazioni cade fino a quel valore compreso.
- **Dato un particolare valore x di una distribuzione di osservazioni a quale percentile corrisponde?**

$$\text{percentile di } x = 100 * \frac{\text{n° dei valori} \leq x}{\text{n° totale dei valori}}$$

71,0
73,7
80,0
81,3
83,5
84,0
84,0
84,5
85,0
85,0
86,0
86,4
86,5
86,5
87,0
87,0
88,0
88,0
88,5
89,5
90,0
90,0
90,2
91,0
91,4
91,5
91,7
92,0
93,0
93,0
93,5
93,5
93,5
96,0
97,0
97,0
97,8
98,0
101,5
102,5

Lunghezza (cm) di 40 coyote femmina (v. esempio 1)

Vogliamo trovare il percentile corrispondente alla lunghezza 91cm.

24 : 40 = x : 100 

x = 60

La lunghezza 91cm è il 60-mo percentile, ossia il 60% dei valori cade alla sua sinistra.

Es. coyote

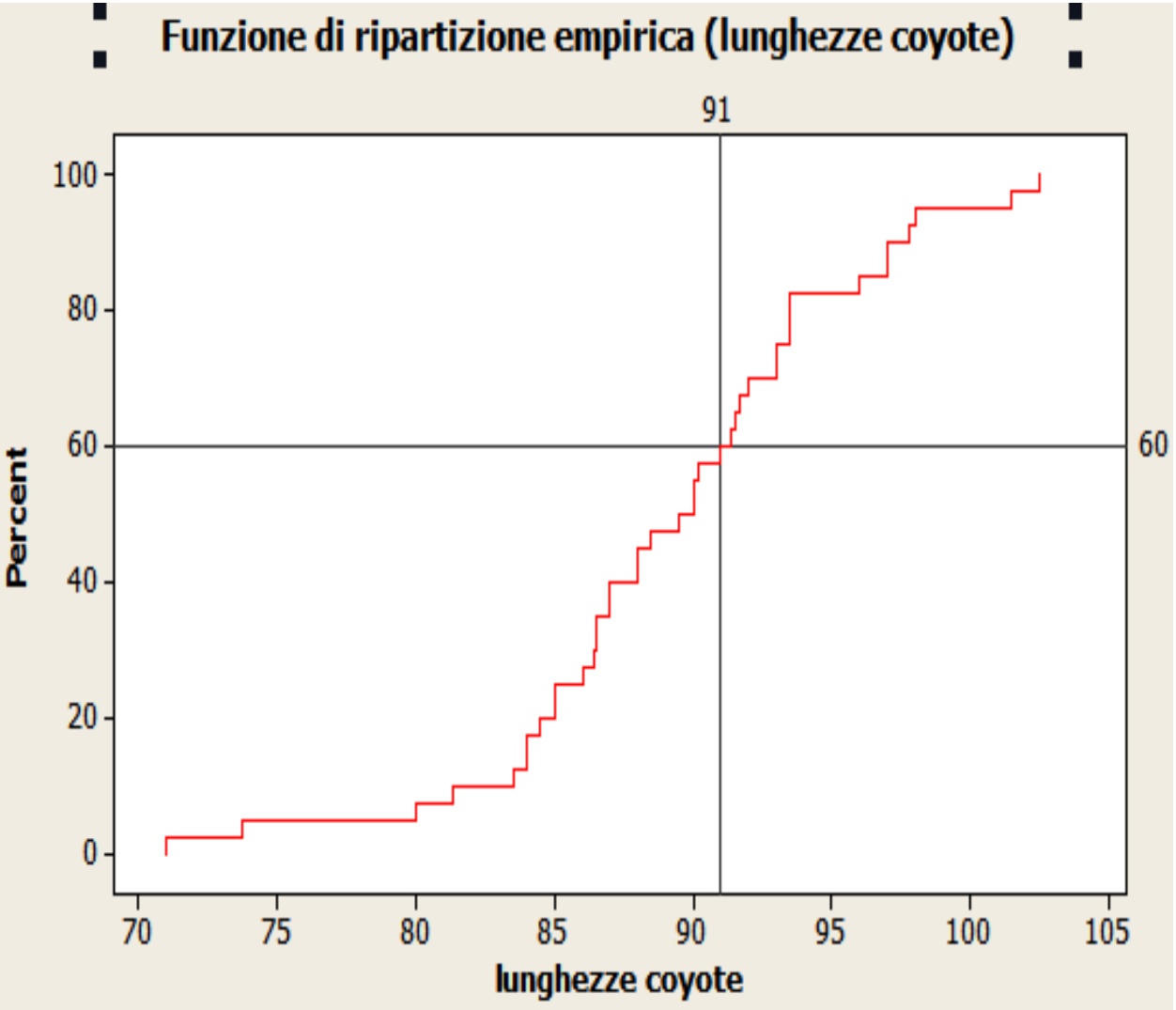
Distribuzione delle frequenze e delle frequenze relative delle lunghezze dei coyote femmina

7 Classi	Frequenza (n_i)	Frequenza relativa (n_i/n)
70- 75	2	0.05
75- 79	0	0
80- 84	6	0.15
85- 89	12	0.3
90- 94	13	0.325
95-99	5	0.125
100-105	2	0.05
Totale	n=40	1.00

Frequenza cumulata percentuale
Grafico della distribuzione cumulata

C1	Count	CumPct
71,0	1	2,50
73,7	1	5,00
80,0	1	7,50
81,3	1	10,00
83,5	1	12,50
84,0	2	17,50
84,5	1	20,00
85,0	2	25,00
86,0	1	27,50
86,4	1	30,00
86,5	2	35,00
87,0	2	40,00
88,0	2	45,00
88,5	1	47,50
89,5	1	50,00
90,0	2	55,00
90,2	1	57,50
91,0	1	60,00
91,4	1	62,50
91,5	1	65,00
91,7	1	67,50
92,0	1	70,00
93,0	2	75,00
93,5	3	82,50
96,0	1	85,00
97,0	2	90,00
97,8	1	92,50
98,0	1	95,00
101,5	1	97,50
102,5	1	100,00
N=	40	

1/40 = 0.025



Percentili

- Esempio: un bambino che superi il 90° percentile avrà un valore (es. di altezza) superiore al 90% di tutti i bambini considerati.
- Esempio: la più piccola osservazione in un insieme di 20 è il quinto percentile (5%), l'osservazione successiva è il 10-mo percentile (10%).

$1 : 20 = x : 100$

I percentili

il 68% circa dei bracchi
è più basso di 62 cm.

Altezza al garrese di 659 cani "Bracco italiano"

(dati fittizi)

classe cm	freq.	%	% cumulativa
53.0 - 53.9	4	0.6	0.6
54.0 - 54.9	7	1.1	1.7
55.0 - 55.9	13	2.0	3.6
56.0 - 56.9	25	3.8	7.4
57.0 - 57.9	41	6.2	13.7
58.0 - 58.9	56	8.5	22.2
59.0 - 59.9	69	10.5	32.6
60.0 - 60.9	74	11.2	43.9
61.0 - 61.9	85	12.9	56.8
62.0 - 62.9	76	11.5	68.3
63.0 - 63.9	67	10.2	78.5
64.0 - 64.9	55	8.3	86.8
65.0 - 65.9	41	6.2	93.0
66.0 - 66.9	26	3.9	97.0
67.0 - 67.9	12	1.8	98.8
68.0 - 68.9	5	0.8	99.5
69.0 - 69.9	3	0.5	100.0

moda

colonna della "distribuzione percentuale". E' in forma standardizzata e quindi facilita il confronto con altri dati

colonna della "percentuale cumulativa". E' utile per il calcolo dei percentili

La distribuzione dei dati segue un andamento particolare, in quanto le barre disegnano una sorta di 'campana'

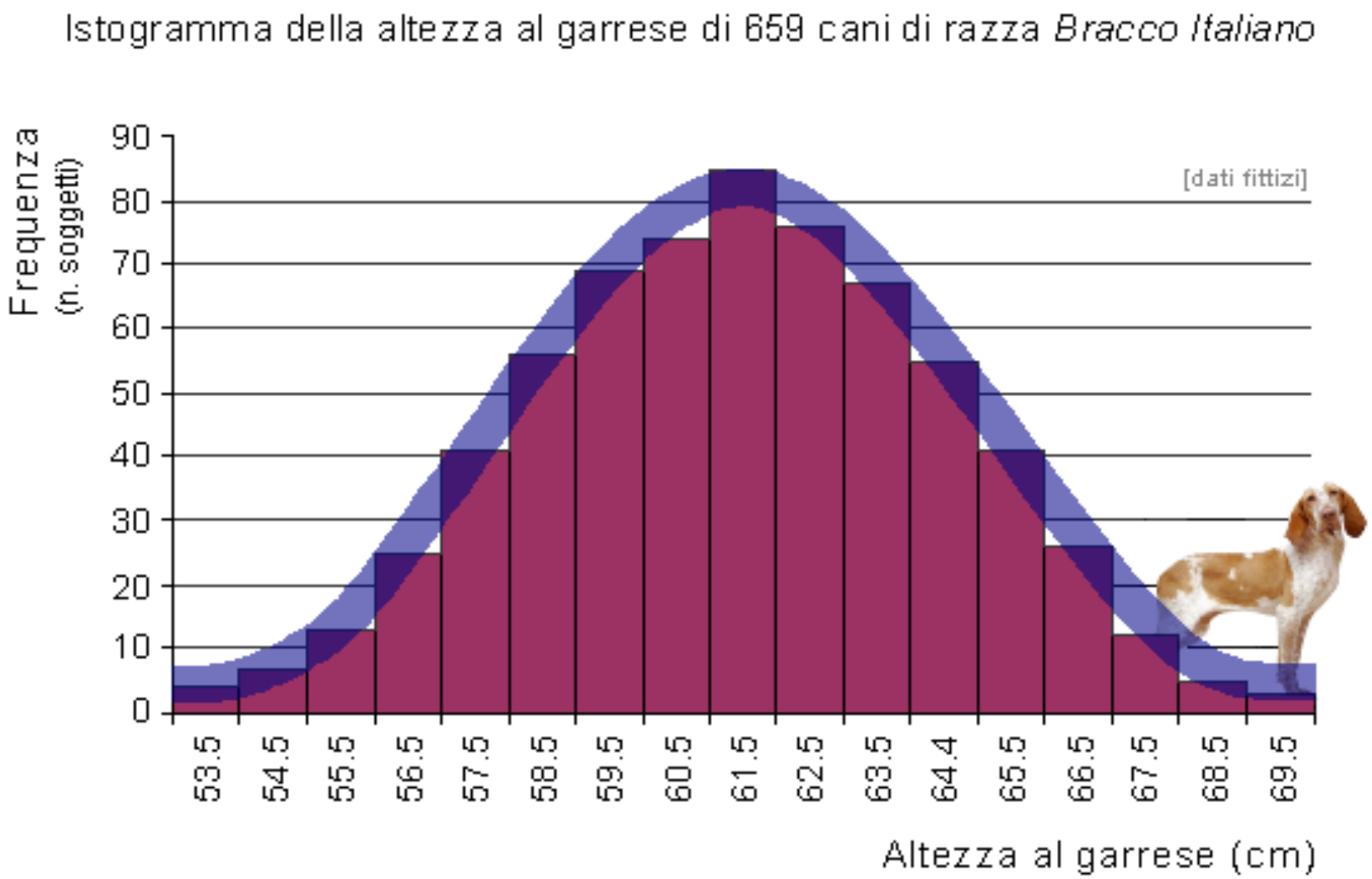
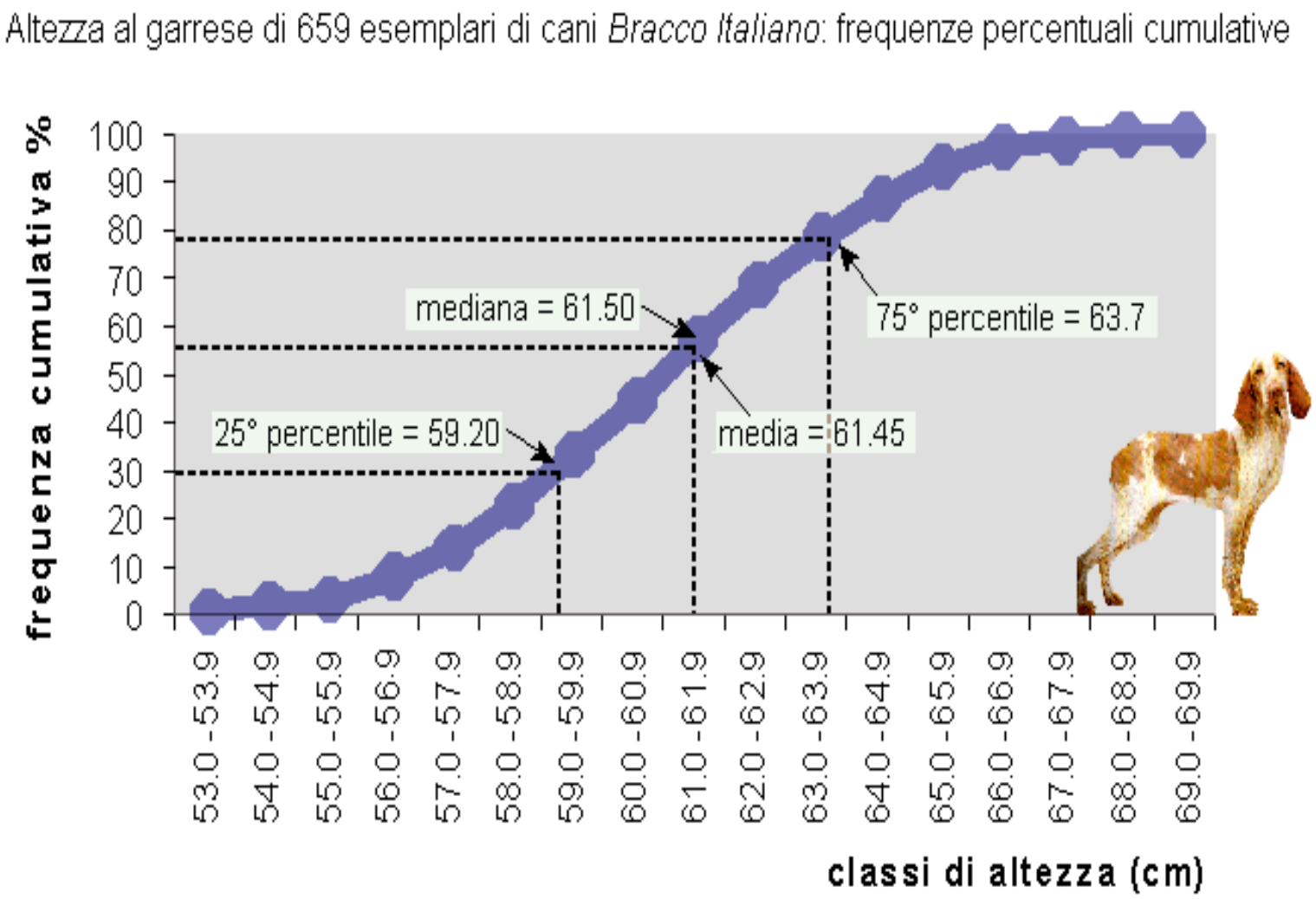


Grafico della distribuzione cumulativa delle frequenze dei dati



sull'asse delle ascisse sono state riportate le classi di frequenza e sull'asse delle ordinate le *percentuali cumulative*.

Indici di variabilità

- Se non ci fosse variabilità all'interno di una popolazione non ci sarebbe bisogno della statistica. Una singola unità campionaria sarebbe sufficiente a descrivere l'intera popolazione.
- Come si misura la variabilità o dispersione di una distribuzione di dati?

Misure di variabilità o dispersione: Intervallo di variazione

- **Intervallo di variazione (range)**

= osservazione più grande – osservazione più piccola

- E’ influenzato dalle osservazioni estreme.

Misure di variabilità: Differenza (o range) interquartile

- **Differenza interquartile**
  = terzo quartile Q_3 - primo quartile Q_1

- Misura la dispersione del 50% dei valori centrali della distribuzione dei dati.
- Esempio
  Per i dati dell'esempio 10 la differenza interquartile è data da

  11 13 15 16 19 21 22 25

  Q_1 = 14 **Q_3 = 21.5**

  Diff. interq. = 21.5 – 14 = **7.5**

Misure di variabilità: Differenza interquartile

Per l'esempio precedente si ha(valori calcolati con il software):

media=17,7

mediana=17,5

Q1=13,50

Q3=21,75

Diff. interq=8.25

Misure di variabilità

- Gli **indici di posizione centrale** dicono *attorno a quale valore le osservazioni sono centrate* e *sono tanto più significativi quanto più* ***i dati sono concentrati vicino ad essi***.

- Per ottenere un'informazione più accurata, è quindi necessario **misurare il grado di dispersione dei dati intorno a tali indici**. Ciò può farsi, soltanto per i caratteri quantitativi, associando alle misure di tendenza centrale delle **misure di variabilità.**

Misure di variabilità: la varianza e la deviazione standard

Varianza →

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

Deviazione standard (radice quadrata della varianza) →

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

La varianza campionaria

- La varianza campionaria s^2 è uno **stimatore non distorto** della varianza della popolazione σ^2

- Questo vuol dire che i valori di s^2 tendono a centrare il valore di σ^2 e non a sovrastimarlo o sottostimarlo in modo sistematico.

La lunghezza dei petali di 3 specie di fiori.
3 distribuzioni a confronto.

La lunghezza dei petali di 3 specie di fiori.
3 distribuzioni a confronto.

Variazioni di scala

I dati della tabella rappresentano i valori della temperatura corporea rilevati su 65 soggetti sani e misurati in gradi Fahrenheit.

La loro media è

$m_f = 98,0$

Temperature corporee in gradi Farenheit			
96,30	97,60	98,20	98,70
96,70	97,70	98,20	98,70
96,90	97,80	98,20	98,80
97,00	97,80	98,30	98,80
97,10	97,80	98,30	98,80
97,10	97,80	98,40	98,90
97,10	97,90	98,40	99,00
97,20	97,90	98,40	99,00
97,30	98,00	98,40	99,00
97,40	98,00	98,50	99,10
97,40	98,00	98,50	99,20
97,40	98,00	98,60	99,30
97,40	98,00	98,60	99,40
97,50	98,00	98,60	99,50
97,50	98,10	98,60	
97,60	98,10	98,60	**Media**
97,60	98,20	98,60	**98,10**

Come cambia la media se cambia l'unità di misura?

- I dati della tabella rappresentano i valori della temperatura corporea degli stessi 65 soggetti sani in gradi Celsius.
- La loro media è $m_c = 36,73$

Temperature corporee in gradi Celsius			
35,73	36,45	36,78	37,06
35,95	36,50	36,78	37,06
36,06	36,56	36,78	37,12
36,12	36,56	36,84	37,12
36,17	36,56	36,84	37,12
36,17	36,56	36,89	37,17
36,17	36,62	36,89	37,23
36,23	36,62	36,89	37,23
36,28	36,67	36,89	37,23
36,34	36,67	36,95	37,28
36,34	36,67	36,95	37,34
36,34	36,67	37,00	37,39
36,34	36,67	37,00	37,45
36,39	36,67	37,00	37,50
36,39	36,73	37,00	
36,45	36,73	37,00	**Media**
36,45	36,78	37,00	**36,73**

Trasformazione delle scale di misura

- Le due scale di misura sono legate dalla seguente trasformazione:

$$T_c = \frac{5}{9}(T_f - 32)$$

dove T_c= temp. in gradi Celsius

che applicata alle **medie** si scrive:

$$m_c = \frac{5}{9}(m_f - 32)$$

e fornisce lo stesso valore del calcolo diretto di m_c a partire dai dati trasformati, ossia

36.73 = 5/9 x (98.10 – 32)

- Con lo stesso insieme dei dati si può verificare che anche la **mediana** gode della stessa proprietà (per questi dati, inoltre, la mediana coincide con la media).

Variazioni di scala

- Cosa succede alla deviazione standard?
- La dev. standard delle temperature in gradi Fahrenheit è pari a 0,70, mentre per le temperature in gradi Celsius è pari a 0,39.
- Se si applica la trasformazione precedente alla dev. standard:

$$devstC = \frac{5}{9}(devstF - 32)$$

non si ottiene 0,39.

ATTENZIONE!! perchè

0,39 = 5/9 x 0,70

Proprietà della media e della deviazione standard

1. Se a tutti gli elementi di una serie di dati viene sommato o sottratto un numero, la media risulterà aumentata o diminuita dello stesso numero, mentre la deviazione standard non cambia.

2. Se tutti gli elementi di una serie di dati vengono moltiplicati per una costante, sia la media sia la deviazione standard risulteranno moltiplicati per la stessa costante.

Lo scarto standard: una regola empirica per dati con una distribuzione approssimativamente normale

- Lo scarto standard, insieme con la media, fornisce una indicazione utile circa l'intera distribuzione dei dati.

- Infatti se la distribuzione è approssimativamente simmetrica (v. normale), l'intervallo

$$\overline{x} \pm s$$

comprende circa il 68% di tutti i valori.

Misure di variabilità: Il coefficiente di variazione

- La deviazione standard risente dell'unità di misura e dell'ordine di grandezza dei dati.

Esempio 13

2 campioni di maschi

	Campione 1	Campione 2
Età	**25 anni**	**11 anni**
Peso medio	**70 kg**	**36 kg**
Dev.st.	**3 kg**	**3 kg**

I due campioni hanno la stessa variabilità?

Coefficiente di variazione

- Calcoliamo il coefficiente di variazione dato da

$$CV = \frac{s}{\bar{x}} \times 100\%$$

Campione 1: C. V. = 3/70 (100) = **4.2%**

Campione 2: C. V. = 3/36 (100) = **8%**

Il coefficiente di variazione esprime s come percentuale di $\bar{x}$ ed è indipendente dall'unità di misura.

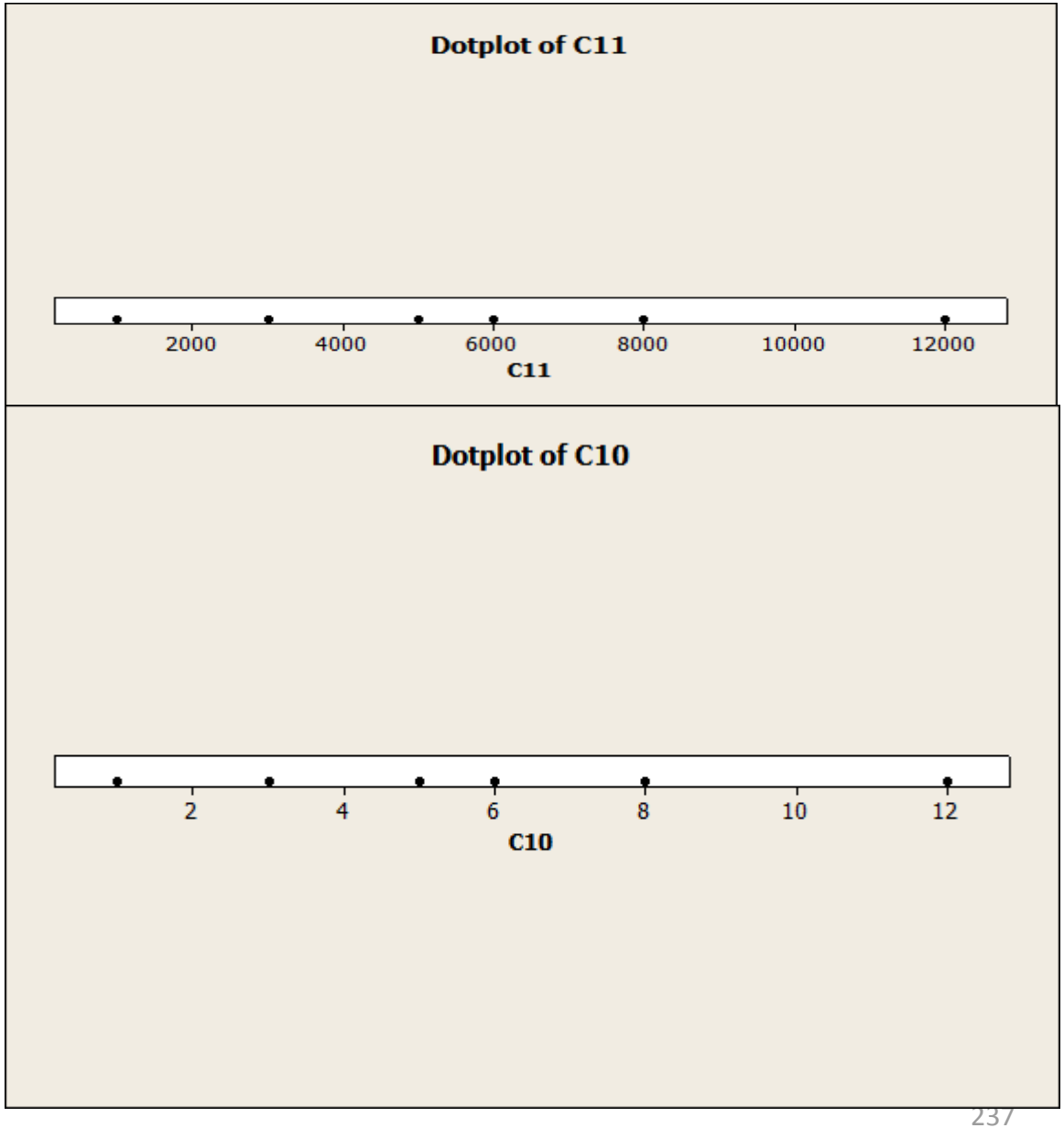
Il coefficiente di variazione

- Due distribuzioni con deviazioni standard **s** molto vicine non hanno necessariamente un'analoga dispersione; infatti **s** è "grande" o "piccolo" rispetto all'ordine di grandezza delle misure a cui si riferisce, ossia ad un indice di posizione come $\bar{x}$.

- Per confrontare le dispersioni di due diverse distribuzioni occorre confrontare indici **indipendenti dall'unità di misura**. L'indice più utilizzato è il Coefficiente di Variazione

C10	C11
1	1000
3	3000
6	6000
5	5000
8	8000
12	12000

Variable	Mean	StDev
C10	5,83	**3,87**
C11	5833	**3869**



Coefficiente di variazione

- Esempio

Gli elefanti hanno orecchie che probabilmente sono 100 volte più grandi di quelle dei topi.

Anche, se le orecchie degli elefanti non fossero più variabili delle orecchie dei topi, la deviazione standard delle lunghezze delle orecchie degli elefanti sarebbe 100 volte più grande della dev. standard delle lunghezze delle orecchie dei topi.

Coefficiente di variazione

I confronti fra deviazioni standard di diverse distribuzioni di dati, hanno senso se:

1) i caratteri sono della stessa natura e sono espressi nelle stesse unità di misura
2) le medie hanno grandezza simile

Nell'es. precedente (es. 13) i giovani del I gruppo hanno altezza media pari a 169cm con dev st pari a 4.1cm. Come si può confrontare 4.1cm (dev st delle altezze) con 3kg (dev st dei pesi)?

I cinque numeri di sintesi e il boxplot

- I cinque numeri di sintesi di una distribuzione sono:

il valore min. Q_1 mediana Q_3 il valore max.

- Questi cinque numeri danno una descrizione sintetica della distribuzione.
- Il boxplot fornisce una rappresentazione grafica dei dati sulla base dei cinque numeri.

I cinque numeri di sintesi

- Esempio 14

Date le osservazioni ordinate:

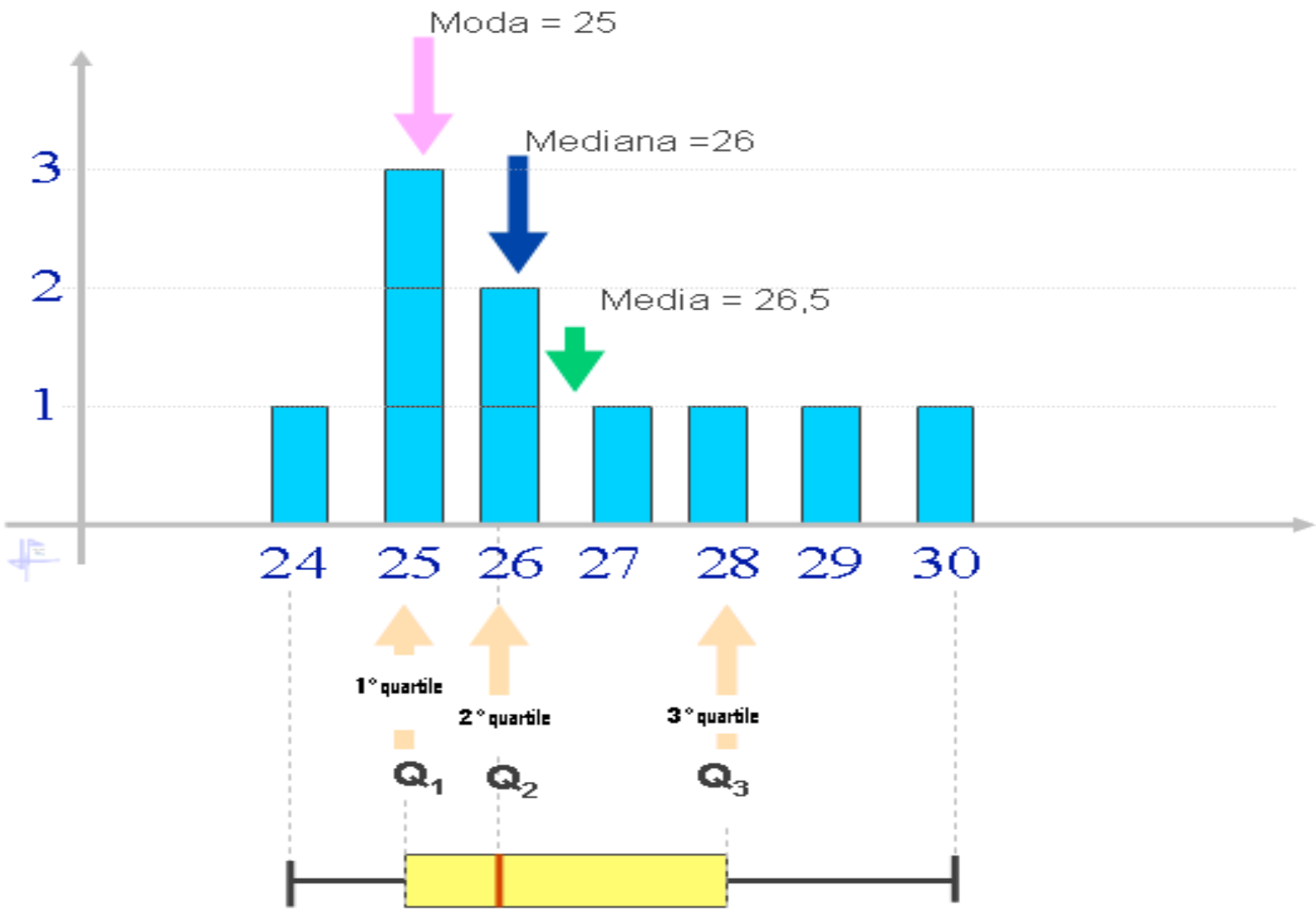
24 25 25 25 26 26 27 28 29 30

$Q_1 = 25$

$Q_2 = (26 + 26)/2 = 26$

$Q_3 = 28$

Boxplot


Moda = 25
Mediana =26
Media = 26,5
3
2
1
24
25
26
27
28
29
30
1° quartile
2° quartile
3° quartile
Q_1
Q_2
Q_3

Quali misure di centralità e dispersione?

- La mediana e i quartili sono misure resistenti, ai valori estremi, la media e la deviazione standard non lo sono.
- La media e la deviazione standard sono ottimi indici per le distribuzioni simmetriche.
- I cinque numeri di sintesi e il boxplot non sono sensibili ai valori estremi e quindi sono più adatti per le distribuzioni asimmetriche.

Boxplot e Istogramma

Età studenti del Corso	Frequenze assolute (n_i)
25	3
26	6
27	8
28	5
29	5
30	3
61	1
	31

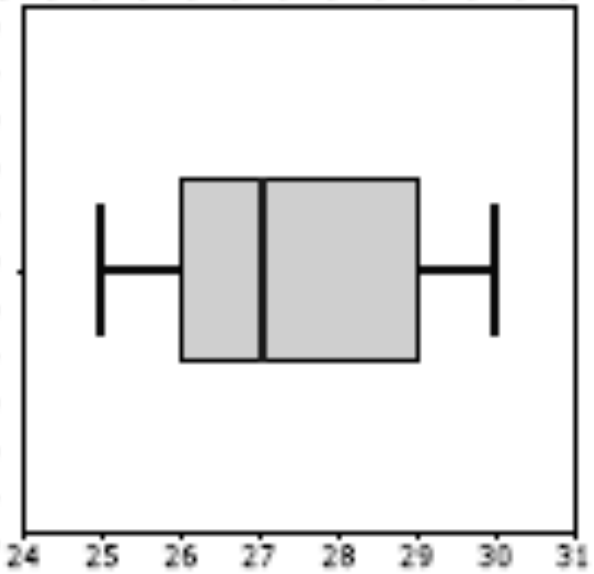
$Q_1 = 26$

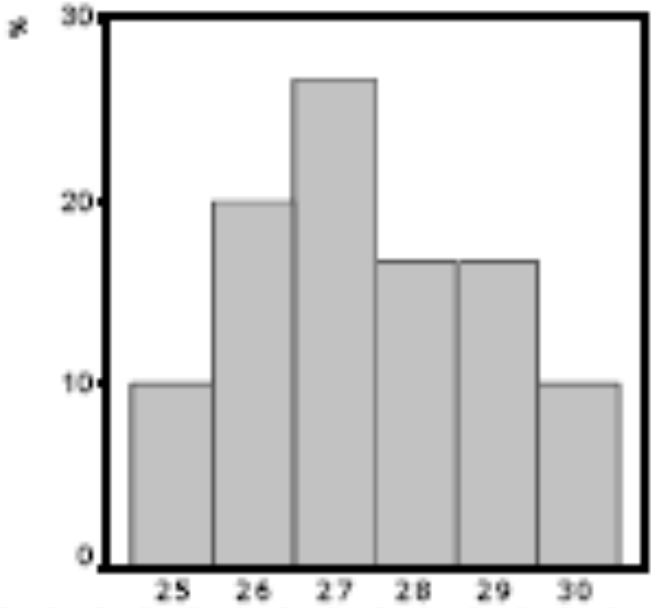
$Med = 27$

$Q_3 = 29$

$Q_3 - Q_1 = 29-26 = 3$

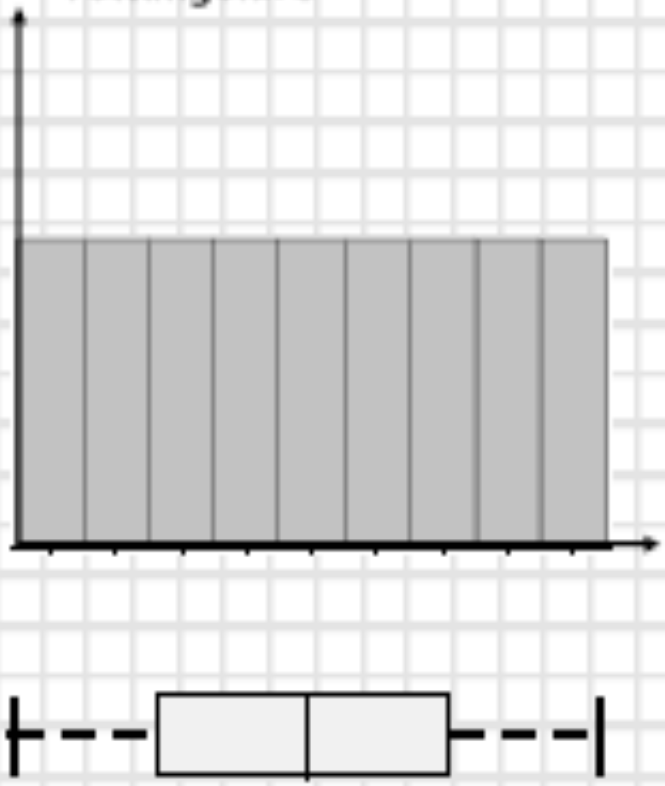
Quanto vale la media?



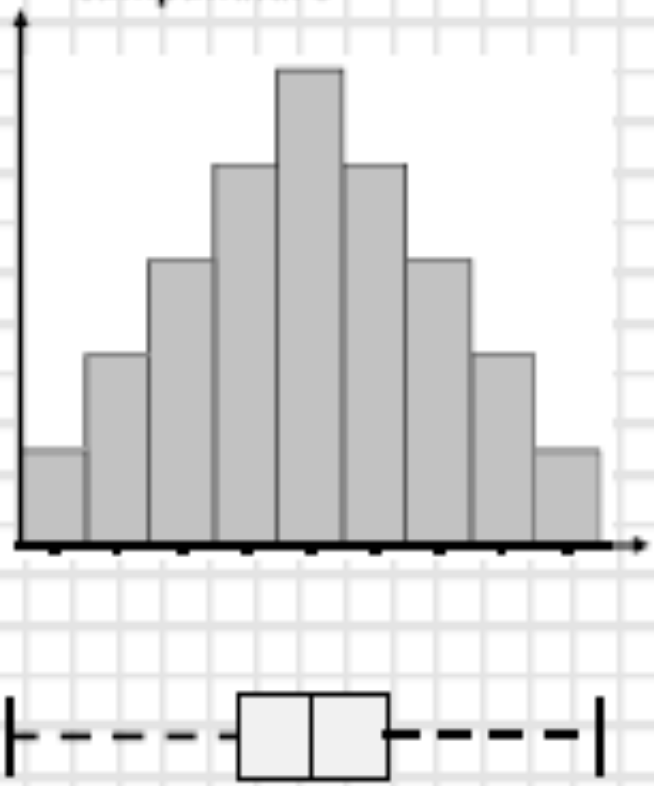


Boxplot e Istogramma

Distribuzione simmetrica rettangolare

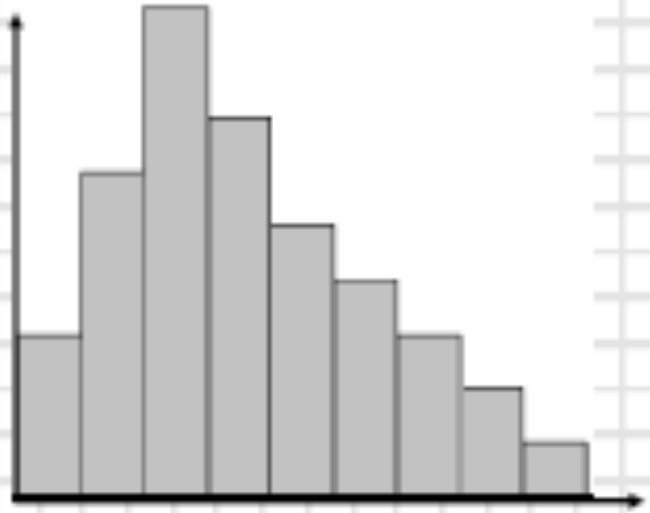


Distribuzione simmetrica campanulare



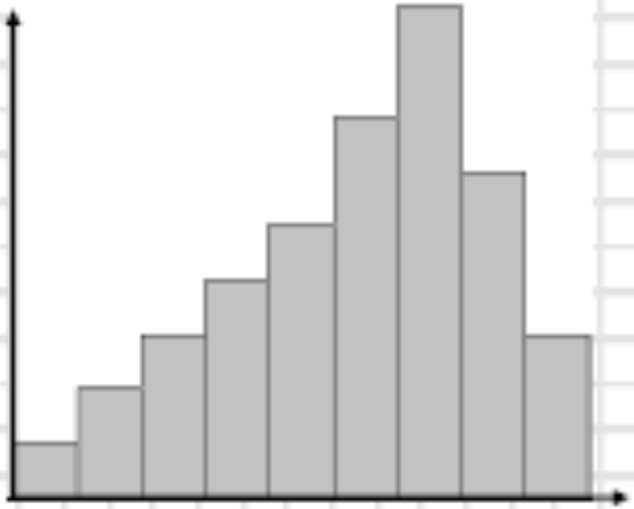
Boxplot e Istogramma

Distribuzione asimmetrica positiva



Me-Q1<Q3-Me

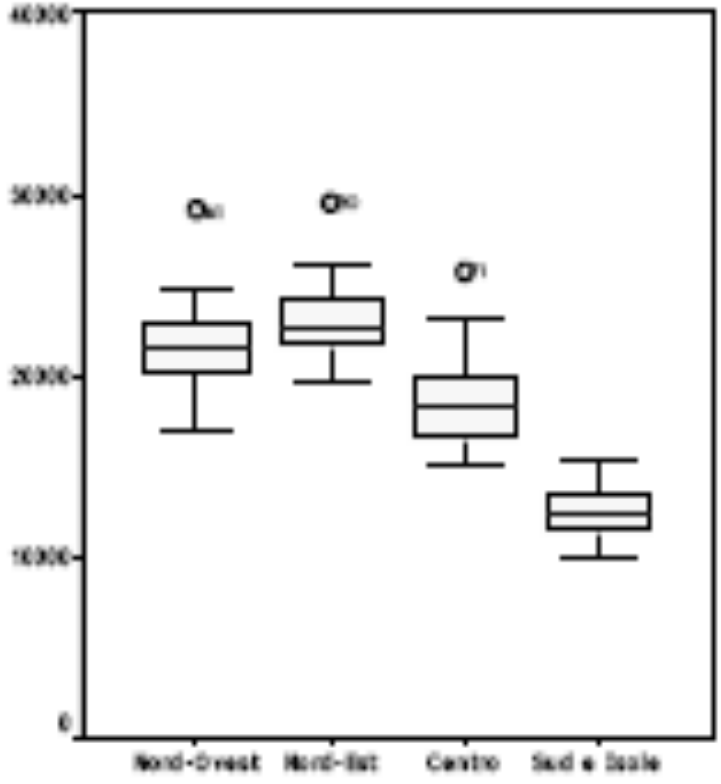
Distribuzione asimmetrica negativa



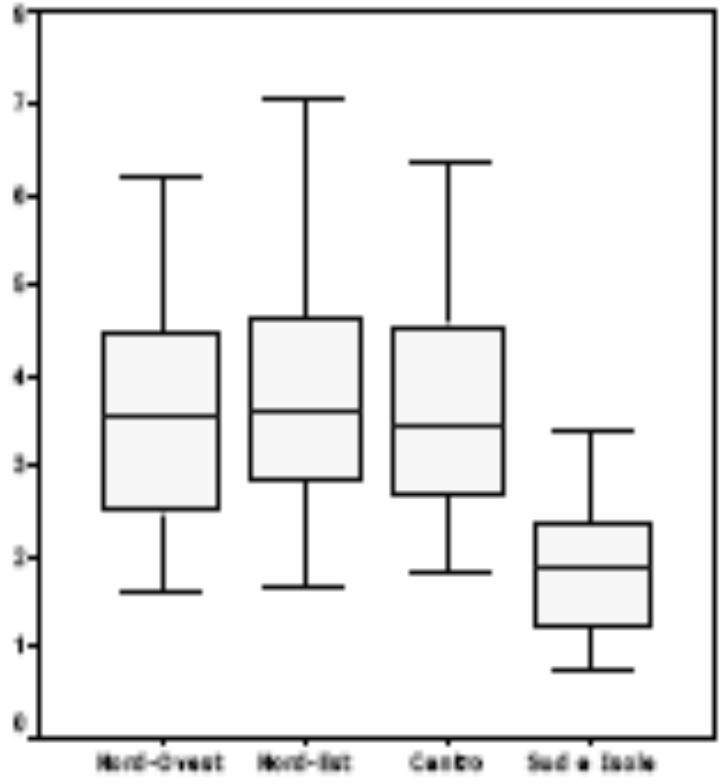
Me-Q1>Q3-Me

Boxplot multipli

Reddito p.c.
(in €)



Num. sale cinematografiche
(per 100mila ab.)



Come individuare gli outliers sospetti?

- Per individuare eventuali outlier la regola più comune è calcolare:

  1.5 x Δ

  dove **Δ** è la differenza interquartile data da

  $$\mathbf{Q_3 - Q_1}$$

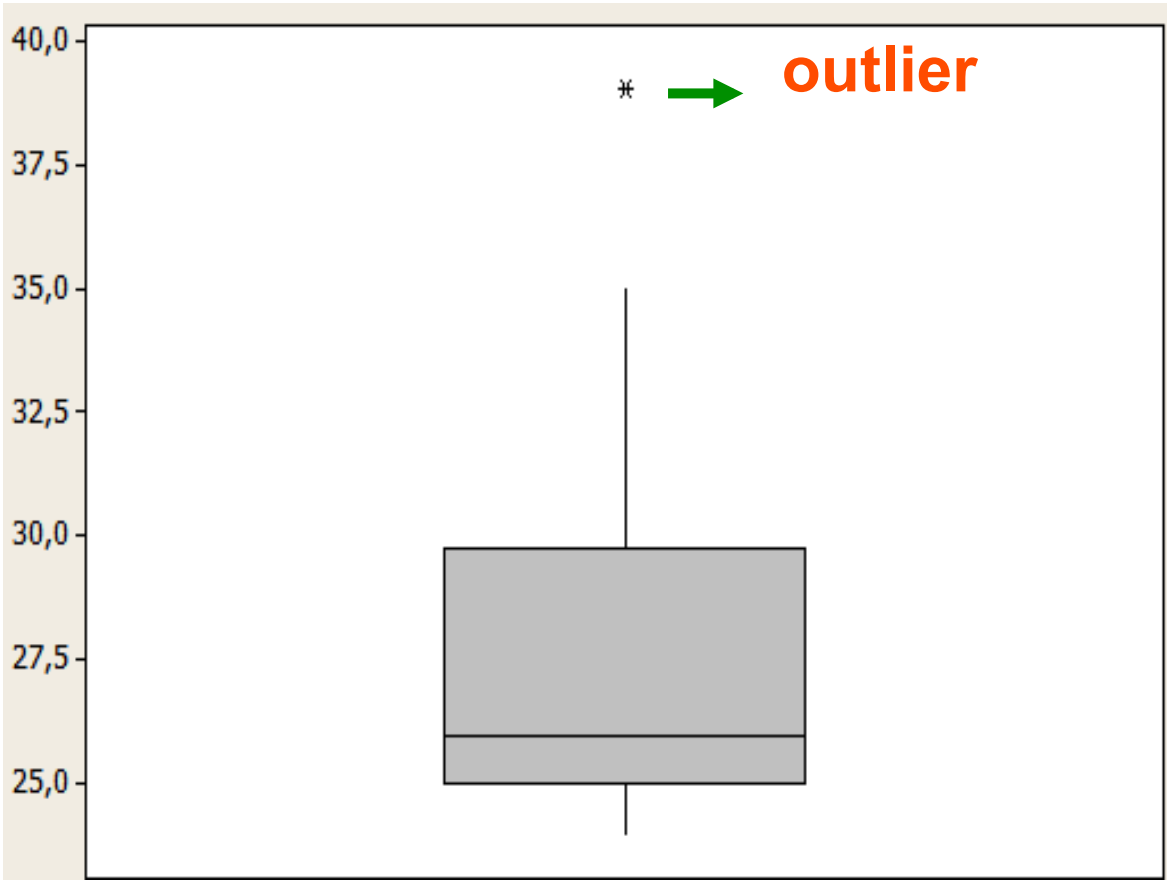
- Un punto viene considerato un possibile outlier se si trova più di **1.5 x Δ** al di sotto del primo quartile o al di sopra del terzo quartile.

Box Plot e outlier

Q1 = 25 Me = 26 Q3 = 29.75 Δ =(Q3-Q1) = 4.75

Q3+1.5(Q3-Q1) = 36.875

24
25
25
25
25
25
26
26
26
27
28
29
30
33
35
39 → outlier

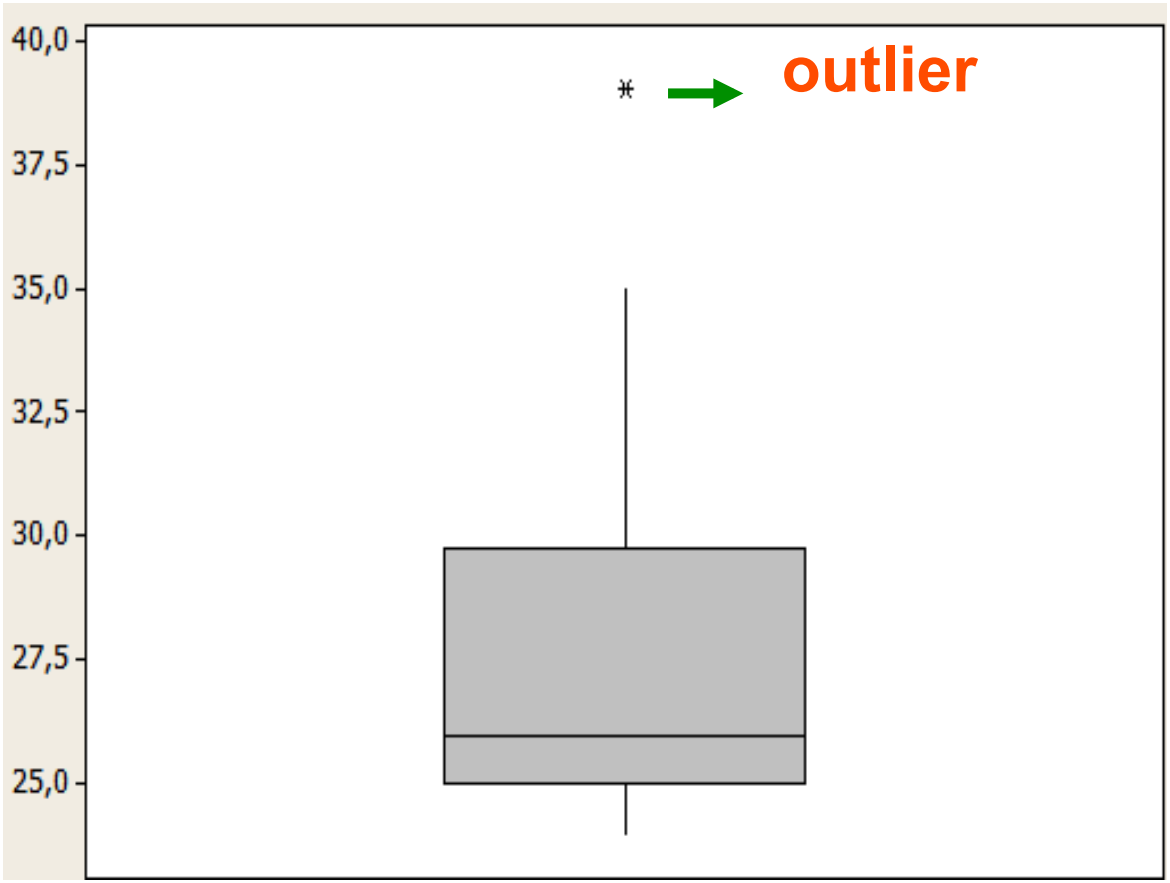


Box Plot e outlier

Q1 = 25 Me = 26 Q3 = 29.75 Δ =(Q3-Q1) = 4.75

Q3+1.5(Q3-Q1) = 36.875 Media=28

24
25
25
25
25
25
26
26
26
27
28
29
30
33
35
39 → outlier



Indici di forma e di simmetria

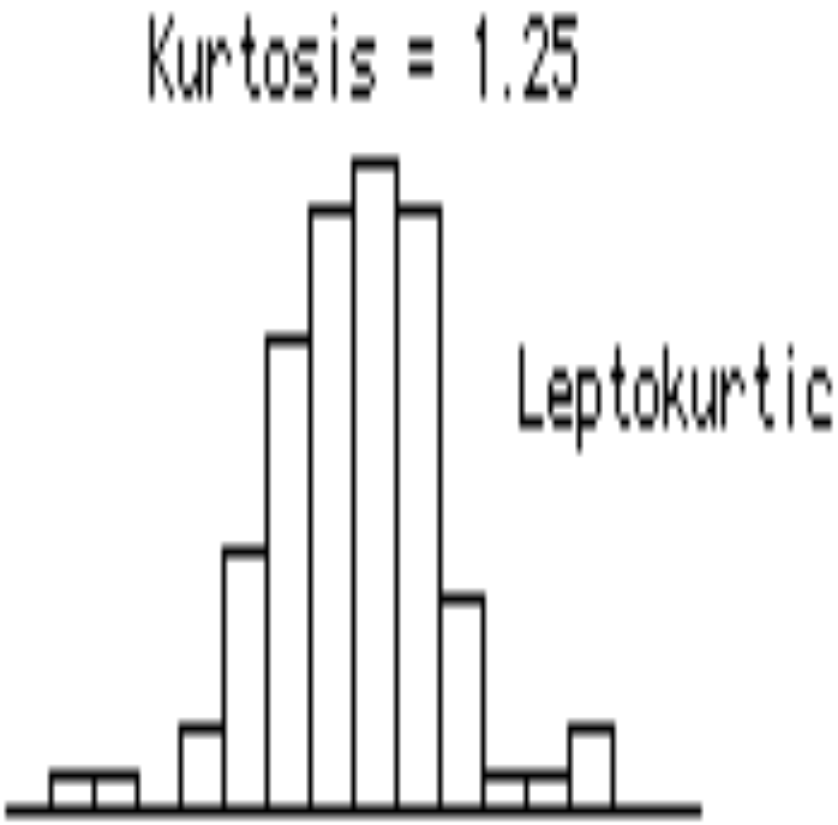
- In biologia, a volte, può essere utile confrontare un istogramma di frequenze osservate con una distribuzione normale con medesima media e deviazione standard. Un indice di "forma" utile, in tal caso, è l'indice di curtosi dato dalla

$$\gamma = \frac{1}{n}\sum_i \left(\frac{x_i - \bar{x}}{s}\right)^4 - 3$$

- L'indice di curtosi è una misura del peso relativo delle code della distribuzione rispetto alla parte centrale (è **tanto più grande quanto più grande è il peso delle code** rispetto alla parte centrale).

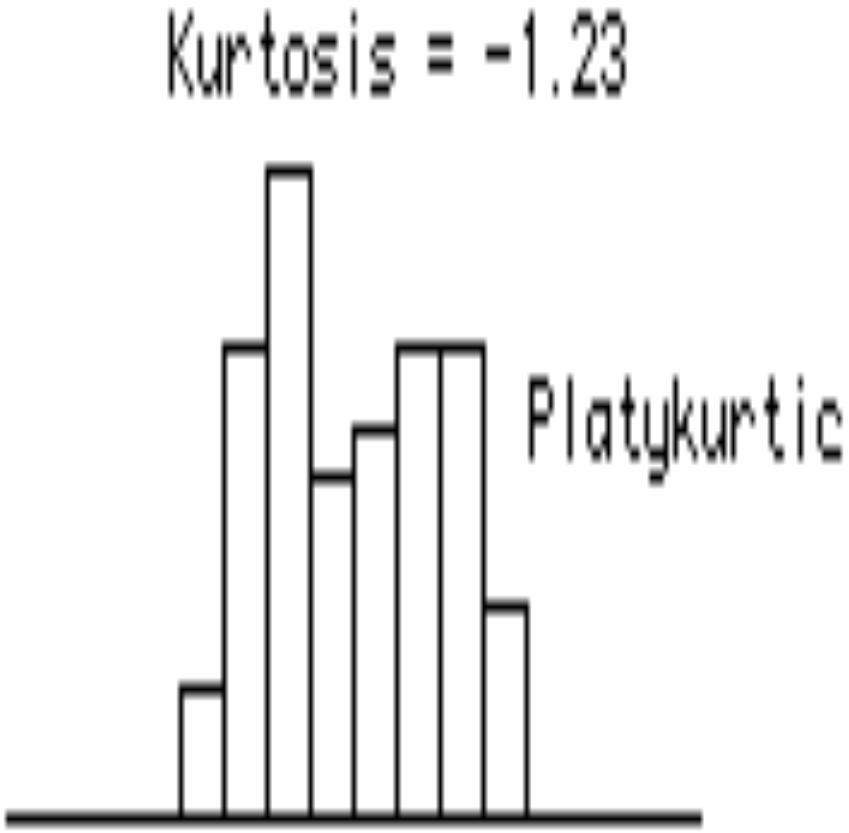
Indice di curtosi

Ipernormale $\gamma > 0$



maggior frequenza dei valori centrali ed estremi

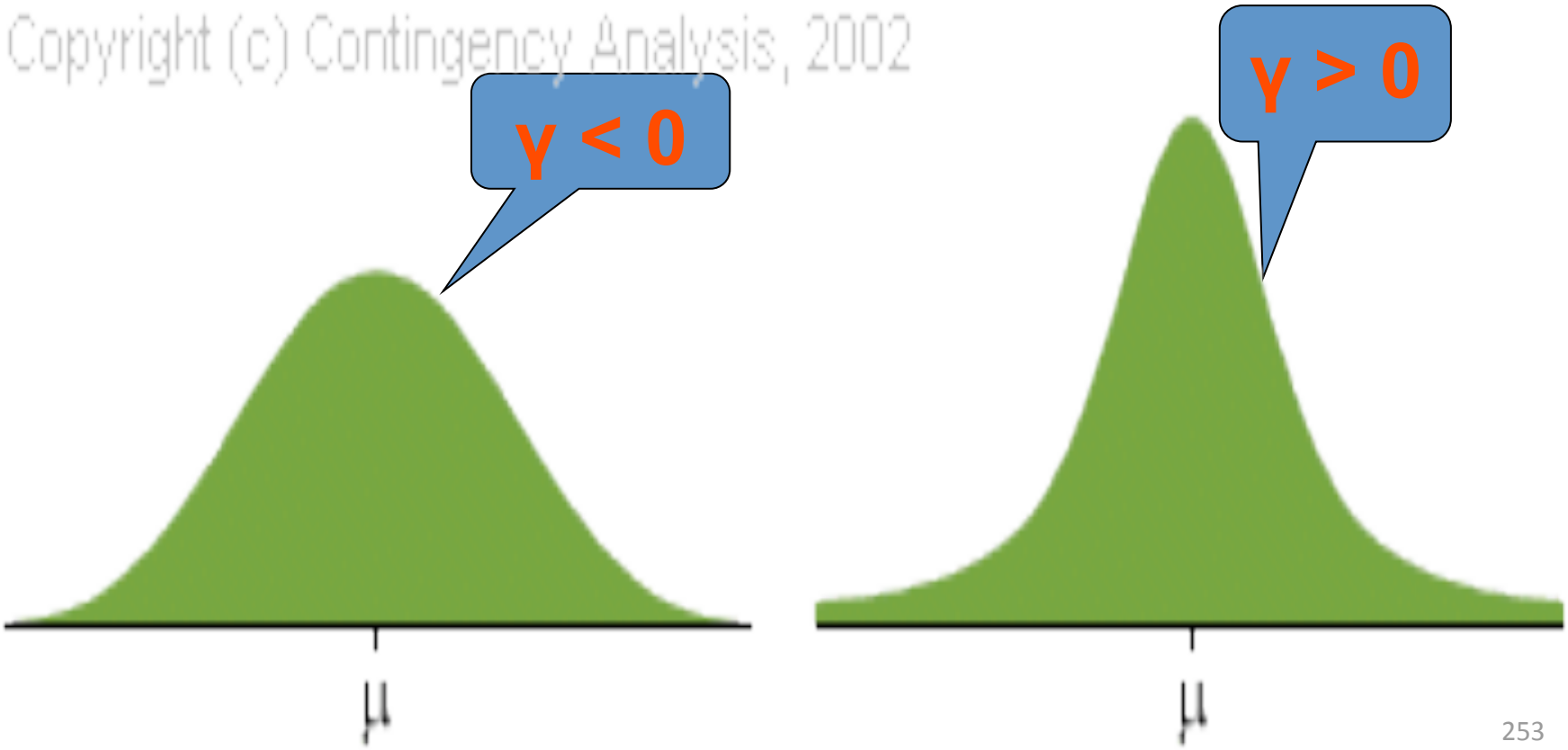
Iponormale $\gamma < 0$



minor frequenza dei valori centrali e estremi

Indice di Curtosi

L'**indice di curtosi** vale **0** (alcuni testi riportano il valore 3) se la popolazione è normale.



Indice di asimmetria

- L'**indice di asimmetria (skewness)** di Pearson è dato da

$$S = \frac{\bar{x} - \text{mod}\, a}{s}$$

- Se S = 0 si ha simmetria
- Se S ≥ 1.00 si ha una asimmetria a destra
- Se S ≤ -1.00 si ha una asimmetria a sinistra.

Riassunto

- Un riassunto numerico di una distribuzione deve riportare il centro e la dispersione.
- La media, la mediana e la moda descrivono il centro di una distribuzione in modi diversi.
- Se si usa la mediana per indicare il centro è opportuno rappresentare la dispersione con i quartili.
- I quartili e i percentili sono misure di posizione non centrale che dividono la distribuzione, rispettivamente, in 4 e 100 parti di uguale numerosità.

- La varianza e la sua radice quadrata, la deviazione standard, misurano la dispersione rispetto alla media come centro.
- Il coefficiente di variazione è una misura di dispersione che non risente dell'unità di misura e dell'ordine di grandezza dei dati.
- Il sommario a cinque numeri (mediana, quartili, minimo e massimo) fornisce una descrizione generale della distribuzione.
- Il box-plot è un grafico del sommario a cinque numeri.
- La mediana e i quartili sono misure resistenti.

- L’indice di curtosi e di asimmetria.
- Attenzione ai cambiamenti di scala e alle trasformazioni dei dati.

ESEMPIO

Se media, mediana e moda di una distribuzione sono 4,6,7 rispettivamente, allora la distribuzione risulta essere

(a)asimmetrica a sinistra

(b) simmetrica

(c) asimmetrica a destra

(d) bimodale

Analizzando il contentuto di 32 pacchi di biscotti, si `e contato il numero di biscotti rotti in ciascun pacco. La tabella di frequenze che segue riassume i dati del campione analizzato.

Num.bisc. rotti	Frequenza
2	5
4	6
6	7
9	2
11	1

Qual `e la mediana di questa distribuzione?

(a) 4 (b) 5 (c) 7 (d) 8 (e) nessuno dei valori precedenti

Qual `e la moda di questa distribuzione?

(a) 11 (b) 7 (c) 6 (d) 1 (e) nessuno dei valori precedenti