Metodi matematici e informatici per la biologia

canale A-E, a.a. 2019/20

Docente: Giada Basile

Informazioni generali

Questo corso è l'ultimo modulo (3CFU) del corso di Calcolo, Biostatistica e Metodi Matematici per la Biologia.

Il corso comprende 26 ore di lezioni frontali + 10 ore di esercitazioni nel laboratorio di informatica sotto la supervisione di un tutor.

L'esame consiste in una prova scritta in laboratorio. Il voto finale dell'esame di CBMMIB è la media ponderata del voto dell'esame di Calcolo e di quello di MMIB (pesi: 2/3, 1/3)

Informazioni generali

Pagina web del corso su

http://elearning2.uniroma1.it/

Nome del corso

Metodi Matematici e Informatici per la Biologia (canale A-E)

a.a. 2019/2020

Docente

Giada Basile

Informazioni generali

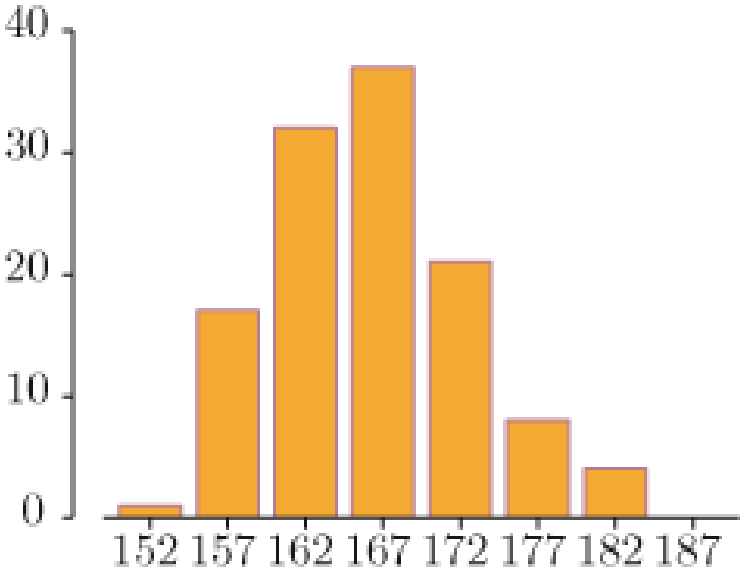
Sulla pagina troverete (tra le altre cose…)

- Diario delle lezioni
- Slides delle lezioni
- Link alla pagina elearning di MMIB (tutti i canali) utile per consultare il calendario dei prossimi appelli e prenotarsi all'esame

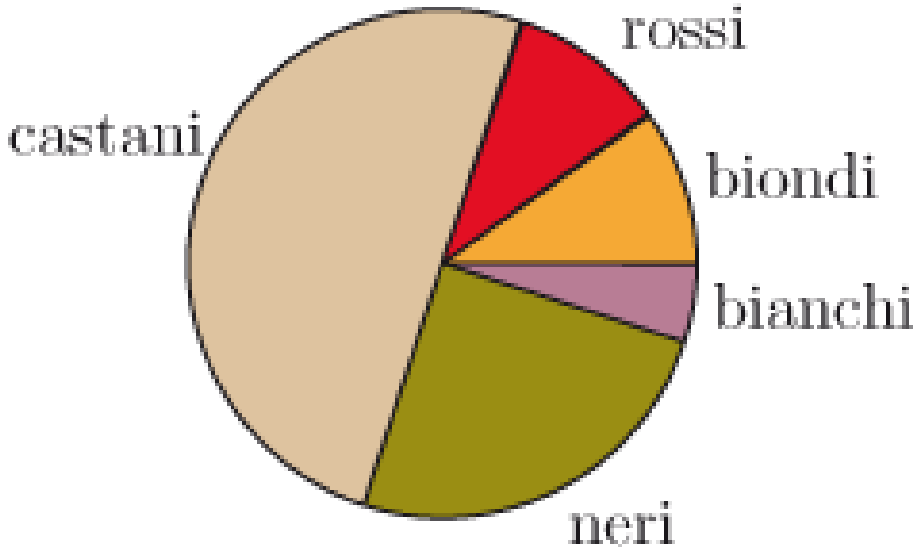
Contenuti del corso

Elementi di statistica con esempi di applicazioni alla biologia.

Iparerete a ....



Organizzare e presentare i dati



- Riassumere i dati in pochi numeri (media, mediana, varianza…)

- Determinare eventuali correlazioni tra i dati

  Esempio: relazione tra quantità d'acqua fornita alle piantine di basilico e l' altezza delle piantine

- Testare delle ipotesi statistiche.

  Esempio: testare l'efficacia di un farmaco

Strumenti

- Probabilità
- Statistica (elementare...)
- Software informatici (Excel, LibreOffice)

Statistica

“La statistica è l’arte di apprendere dai dati”

Si occupa di:

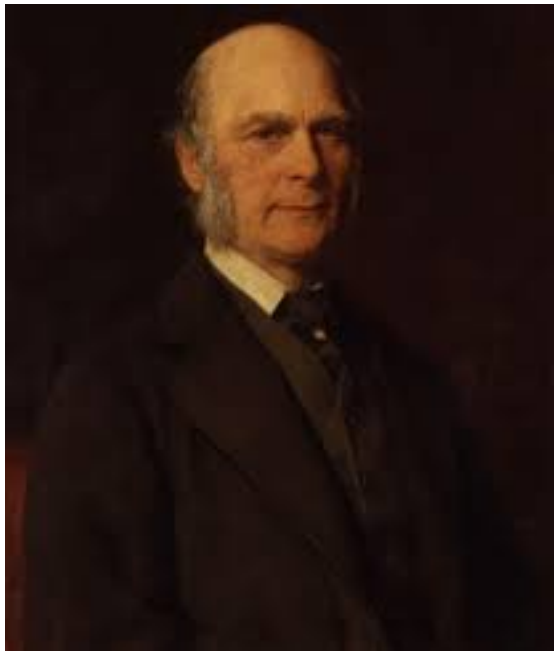
- Raccogliere i dati: ideare un procediment ottimale per la raccolta dei dati
- Elaborare i dati: descrivere e sintetizzare i dati raccolti
- Analizzare i dati: trarre informazioni dai dati raccolti

Cenni storici

- Il termina è collegato a *“stato”* e si usava per indicare una raccolta di informazioni che interessavano allo stato (censimenti, redditi)
- l’origine della statistica nella sua accezione moderna rivale al XVI secolo, quando gli stati europei chiedevano alle parrocchie di registrare nascite, matrimoni e decessi (demografia)
- Nel 1662 il commerciante inglese John Graunt usò i bollettini di mortalità di Londra per costruire una prima e rudimentale "tavola della mortalità". In particolare riuscì a stimare il numero di abitanti a partire dalle tavole.

Nascita della statistica inferenziale

- Verso la fine dll’800 gli statistici iniziarono a trarre conclusioni dai dati numerici. Il primo lavoro fu quello di Francis Galton, che studiando la relazione tra l'altezza dei figli e l'altezza dei padri, constatò che lo scarto dalla media regredisce.
- Seguì il lavoro di Karl Pearson, che sviluppò il test del chi-quadro.



Oggi la statistica si utilizza in diversi campi (demografia, economia, medicina, genetica, fisica, big data) per trarre informazioni da un grande numero di dati.

Perchè la statistica in biologia?

Perché la statistica in biologia?

Per la 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.

La variabilità

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

Esempio

- Si vuole determinare se con una maggiore irrigazione l'altezza delle piante di basilico aumenta
- Ciascuna pianta è geneticamente diversa dalle altre e risponde in maniera diversa alle sollecitazioni esterne (irrigazione)

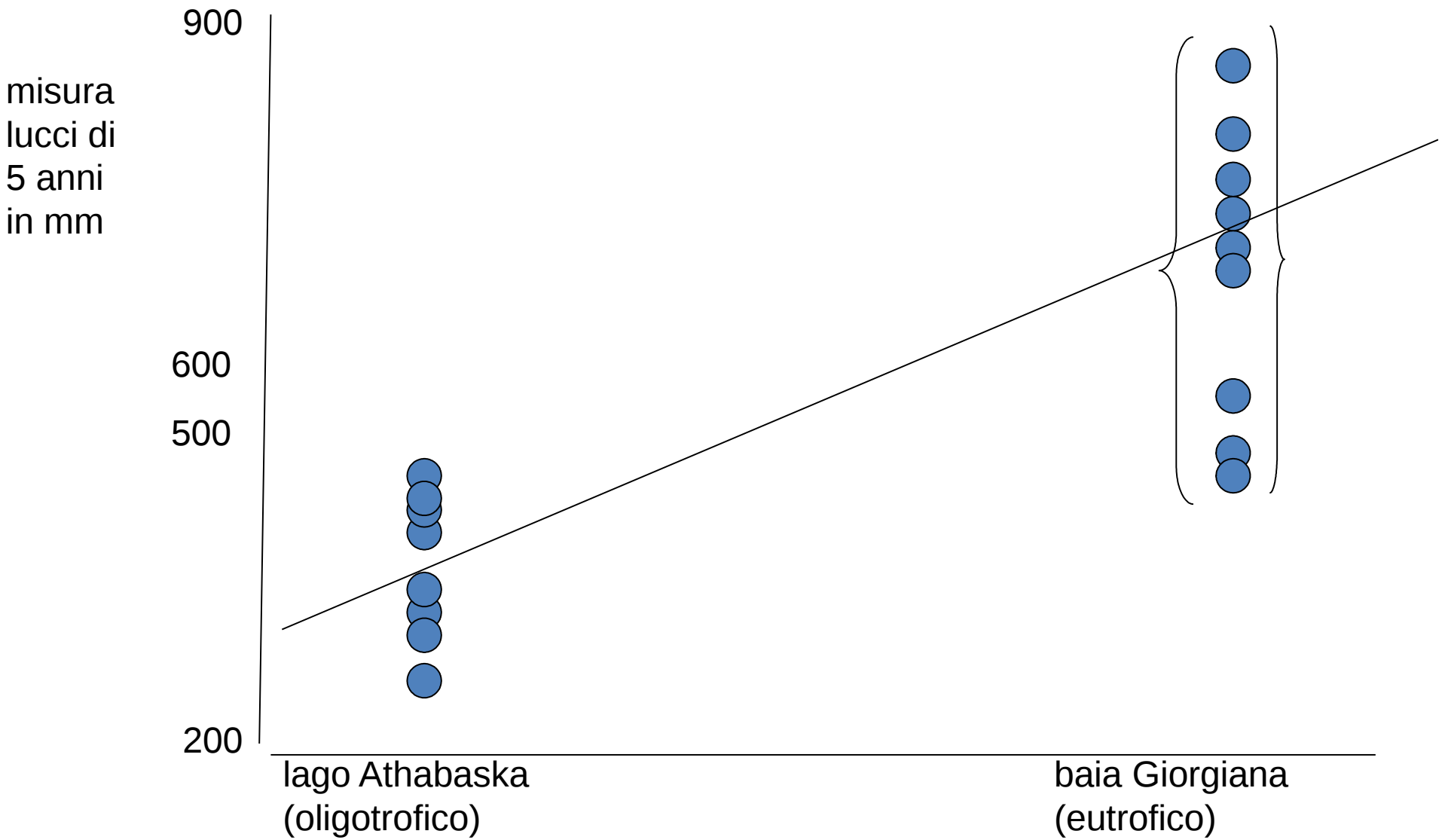
- Confrontando i risultati ottenuti su due piante non si ottiene nessuna informazione
- Per trarre delle conclusioni bisogna effettuare l'esperimento su un numero ragionevolmente grande di piante

Esempio

Rif. 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 misurano diversi lucci dello stesso lago

-si confrontano le misurazioni nei due laghi


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

Esempio

Si supponga di effettuare un esperimento per controllare come funziona una dieta sui topi.

1. variabilità “intrinseca”: i topi sono diversi geneticamente.
2. variabilità “estrinseca”: i topi di gabbie diverse sono soggetti a condizioni diverse (calore, luce, altri fattori).

Esempio III

- 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

Conclusione: l'irrigazione frequente produce cavoli più grandi e una distanza minore ne riduce la dimensione.

Domande:

- 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?

Popolazioni e campioni

Popolazioni

La popolazione (o popolazione satistica) è l'intero gruppo di elementi (unità) sul quale vogliamo ottenere informazioni.

Le unità (o unità statistiche) possono essere individui, ma anche oggetti o misurazioni.

Esempi:

gli abitanti del Lazio; i pazienti di un certo ospedale; le temperature massime registrate quotidianamente a Roma nel 2019; le piante di girasole presenti in un certo campo.

Popolazioni statistiche

(altri esempi, con alcune distinzioni)

- L' insieme dei caprioli che vivono nelle Alpi (popolazioni fisiche)
- L'insieme degli abitanti di Roma
- L'insieme dei valori di temperatura rilevati a Roma il 6 marzo alle 14:00 degli anni dal 1980 al 2018 (popolazioni finite)
- L'insieme dei geni del genoma umano
- L'insieme di tutte le possibili estrazioni di due carte da un mazzo (popolazioni ipotetiche)
- L'insieme dei multipli di 13 (popolazioni infinite)

Osservabili

Data una certa popolazione, vogliamo trarne una informazione

Esempi:

- l'altezza media dei caprioli che vivono sulle Alpi;
- la pressione sistolica dei pazienti di un certo reparto;
- il reddito pro-capite annuo dei cittadini del Lazio;
- la media dei voti dell'esame di MMIB degli studenti del corso

Campioni

In alcuni casi possiamo trarre l'informazione guardando ciascun individuo della popolazione (censimento)

Se non è possibile (es. popolazione troppo numerosa) cerchiamo di ottenere l'informazione da un sottoinsieme dei suoi elementi

Un **campione** è il sottoinsieme della popolazione che viene esaminato per ottenere le informazioni che interessano.

Popolazioni e campioni. Esempio.

Vogliamo misurare la concentrazione di polveri sottili (PM10) presente nel comune di Padova martedì 6 marzo alle 17:10

Popolazione: i valori del PM10 in tutti i punti della superficie comunale alle 17:10 del 6 marzo

Campione: i valori del PM10 in certo numero di punti sulla superficie comunale, in cui abbiamo piazzato dei rilevatori.

Popolazioni e campioni. Esempio.

Vogliamo stimare l'incidenza del covid-19 sugli abitanti della Lombardia.

Popolazione: gli abitanti della Lombardia (circa 10 milioni)

Campione: il sottoinsieme degli abitanti della Lombardia a cui è stato effettuato il tampone (5723 individui)

Campione e disegno campionario

- Il disegno campionario o piano di campionamento è il metodo usato per selezionare il campione.

- il campione deve essere rappresentativo della popolazione.

- Campioni scelti in modo assolutamente casuale in una popolazione sono rappresentativi.

Campione casuale

Definizione. Un campione di k membri di una popolazione (un campione di dimensione k) si dice campione casuale se gli elementi sono scelti in modo che tutte le possibili scelte dei k elementi siano ugualmente probabili

Osservazione. Ogni unità della popolazione ha la stessa probabilità di essere selezionata e campioni della stessa dimensione hanno la stessa probabilità di essere selezionati.

Esercizio

Si vuole selezionare un campione casuale di dieci studenti da una classe di 122 studenti.

Probabilità che venga selezionato il più giovane?

Probabilità che venga selezionato il più anziano?

Numero di possibili campioni?

Probabilità che un certo campione di dimensione 10 venga selezionato?

Scegliere un campione casuale semplice

Il campionamento casuale (o campionameno casuale semplice CCS) è analogo alle estrazioni senza reimmissione delle unità da un'urna.

a) etichettare le unità della popolazione

b) usare il generatore di numeri casuali

c) 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

Esempio

Sorteggiare 10 studenti da una classe di 122 usando le tavole dei numeri casuali

- Numero gli studenti da 1 a 122
- Sulla tavola dei numeri casuali guardo le triplette a partire dalla prima riga; se il numero è minore o uguale a 122 lo tengo, altrimenti lo scarto e vado avanti, fino a collezionarne 10.
- Scelgo gli studenti che hanno I numeri corrispondenti

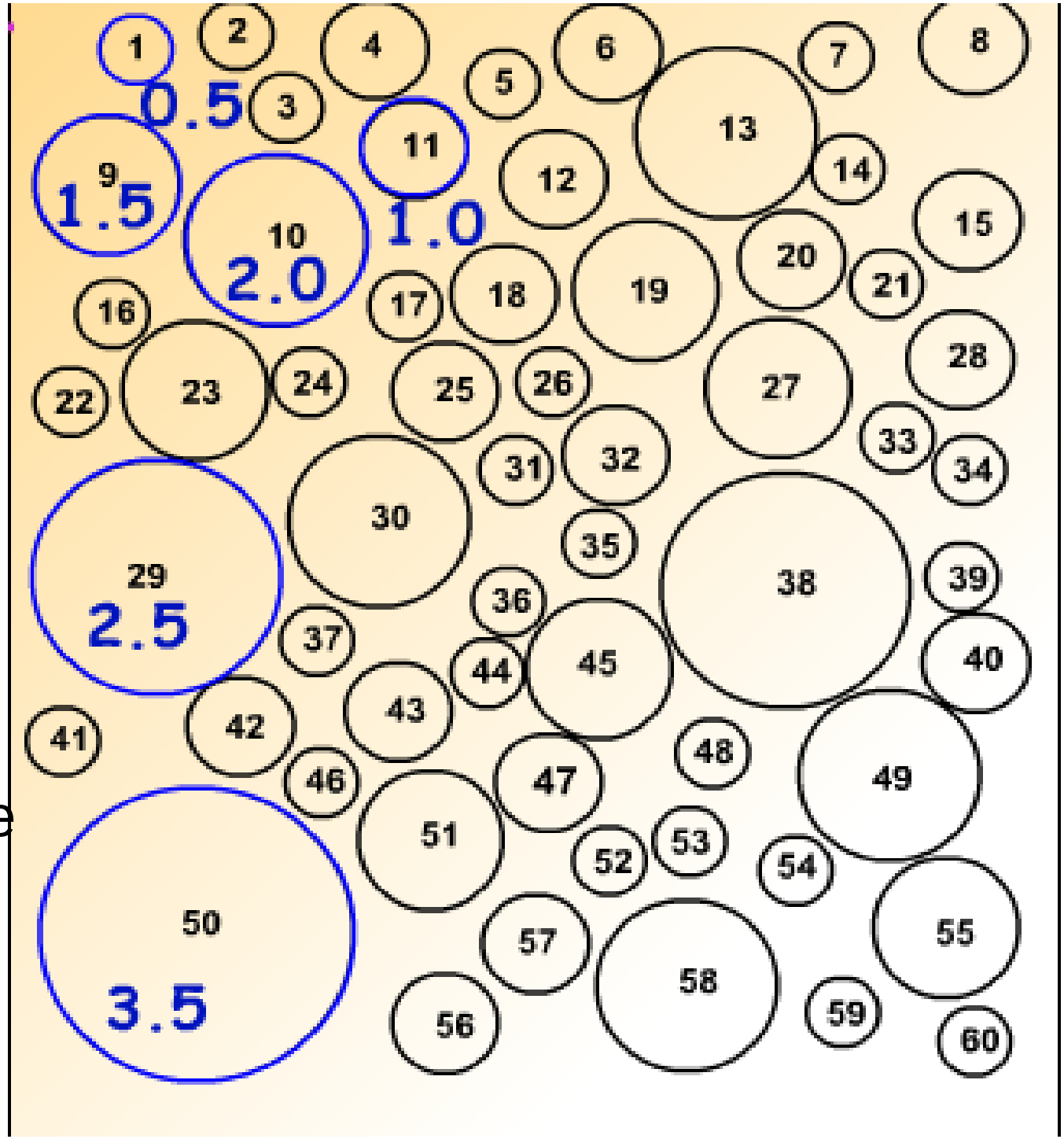
(084, 122, 120, 022, 078, 028, 098, 087, 098, 081)

Popolazione:
60 cerchi.

Si vuole stimare il diam. medio.
(**μ = 1cm**)

Si estrae un campione casuale semplice di dimensione 5.

Come?



Campione di 5 elementi da una popolazione di 60 cerchi.

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)

Esempio

Osservazione:

Posso anche guardare le coppie della prima riga e scartarle se maggiori di 60

5965 2913 5612 6361 7075 5490 9626 4307 0840 7945 5801 9383 6173 8358

Risultato (59, 29, 56, 54, 43)

Oppure: (13, 12, 26, 07, 40)

Calcolando la media campionaria nei tre casi, ottengo risultati diversi (variabilità campionaria)

Esempio

- 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

- Quanti sono i campioni possibili?

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

Popolazione di 6 vermi di lunghezza (cm): 8, 9, 10, 10, 11, 12
Campione casuale di dimensione 2

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 (insieme delle medie campionarie ottenute)**

Media della popolazione = 10cm

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

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.

- 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)

Un esempio storico

Durante le elezioni presidenziali USA del 1936, il “Literary Digest” aveva previsto la vittoria del candidato Alfred Landon contro Franklin Roosvelt (che invece vinse pesantemente).

Il giornale si era basato su un sondaggio compiuto su un campione di elettori scelti dalle liste dei proprietari di automobili e di apparecchi telefonici.

Qual’è stato il motivo dell’errore?
Si è trattato di un campione distorto?

Campionamento casuale a strati

Consideriamo una scuola superiore che abbia 300 studenti del primo anno, 500 del secondo, 600del terzo, 600 del quarto e 500 del quinto.

Si vogliono conoscere le opinioni degli studenti sulla scelta universitaria.

Invece di estrarre a caso 100 studenti, vogliamo dividere la popolazione in 5 strati (le 5 classi) e scegliere casualmente un certo numero di studenti in ogni classe.

Visto che la frazione di studenti del primo anno è 300/2500=0,12 la proporzione nel campione deve rimanere invariata, dobbiamo scegliere a caso 0,12*100=12 studenti del primo anno nel campione

Campionamento casuale a strati

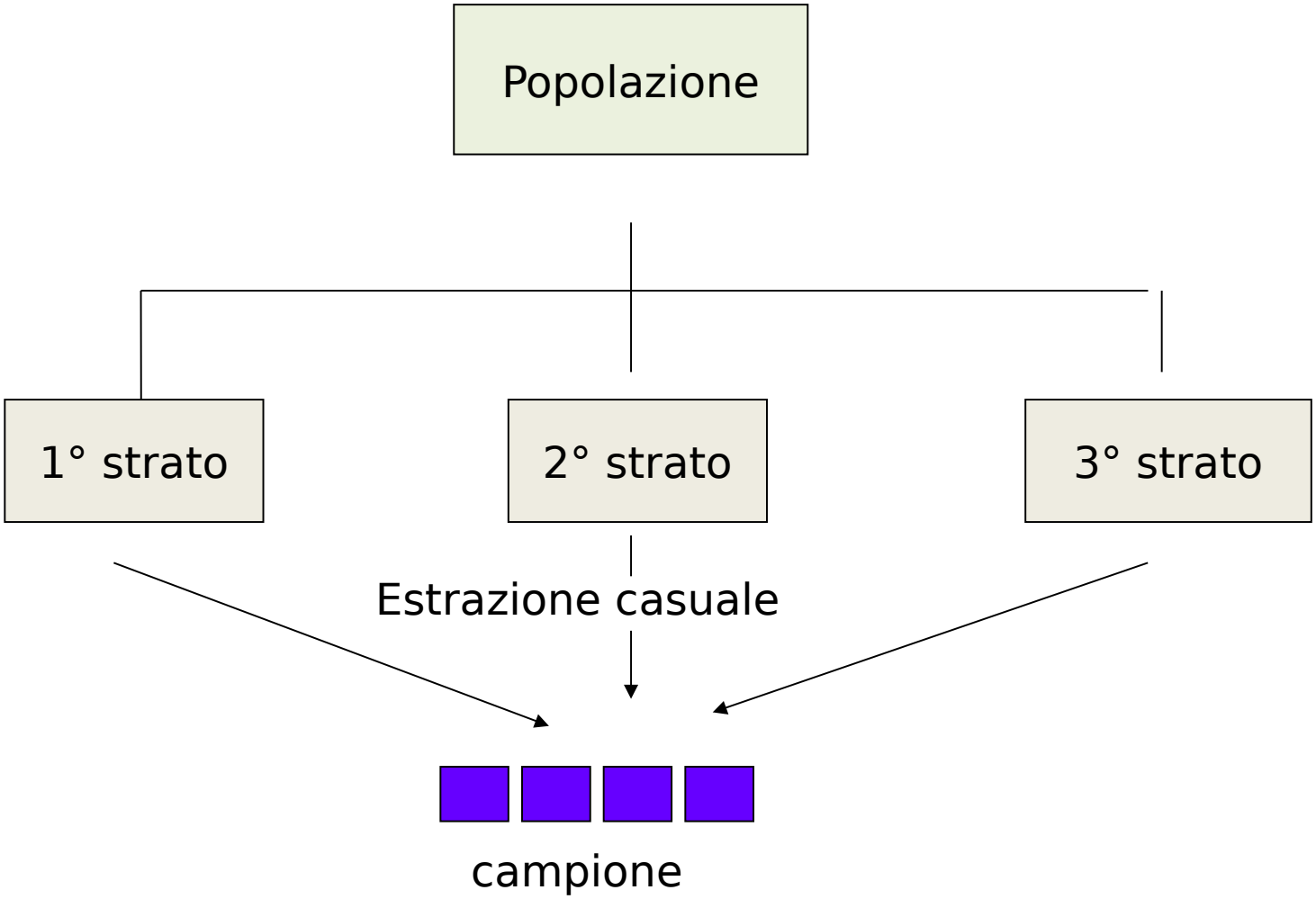
Nel campionamento casuale a strati si suddivide la popolazione in gruppi di unità simili, o strati.

Per ogni strato si sceglie un CCS e si uniscono alla fine tutti i **CCS** formando un unico campione.

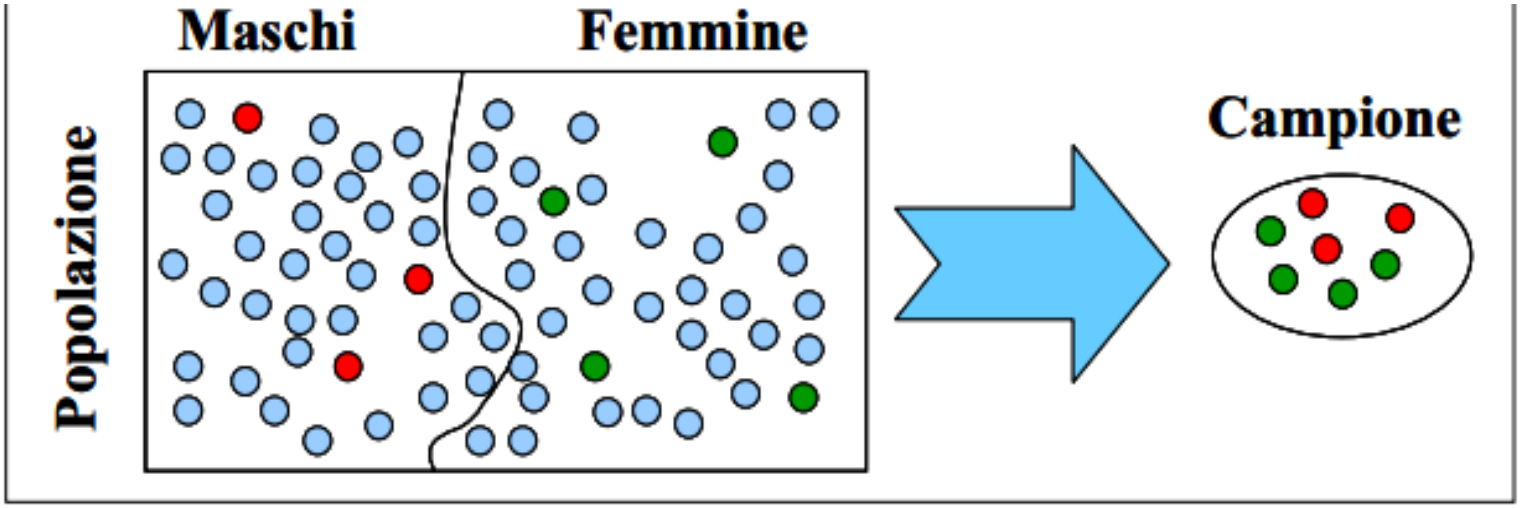
IL n. dei soggetti del campione di ciascuno strato è proporzionale alla dimensione dello strato nella popolazione

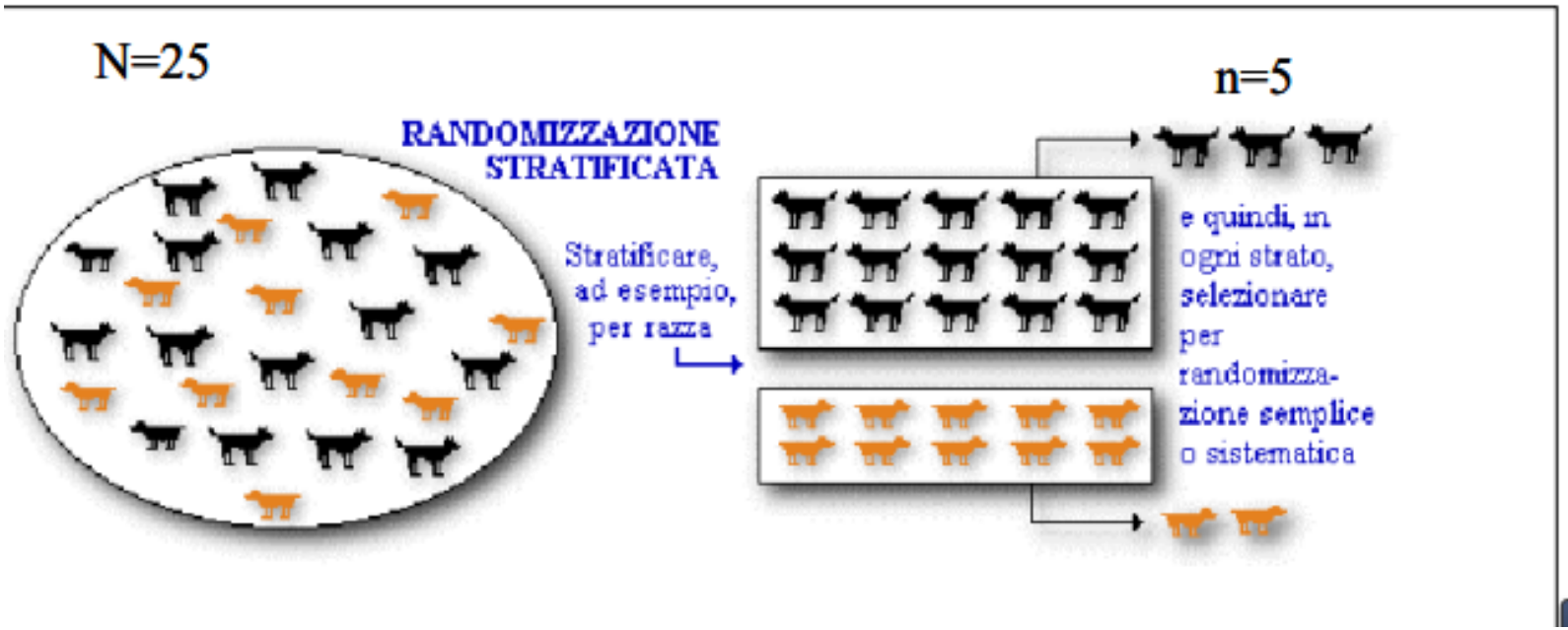
Il campionamento a strati ha senso solo se la caratteristica da analizzare mostra differenze consistenti tra i vari strati.

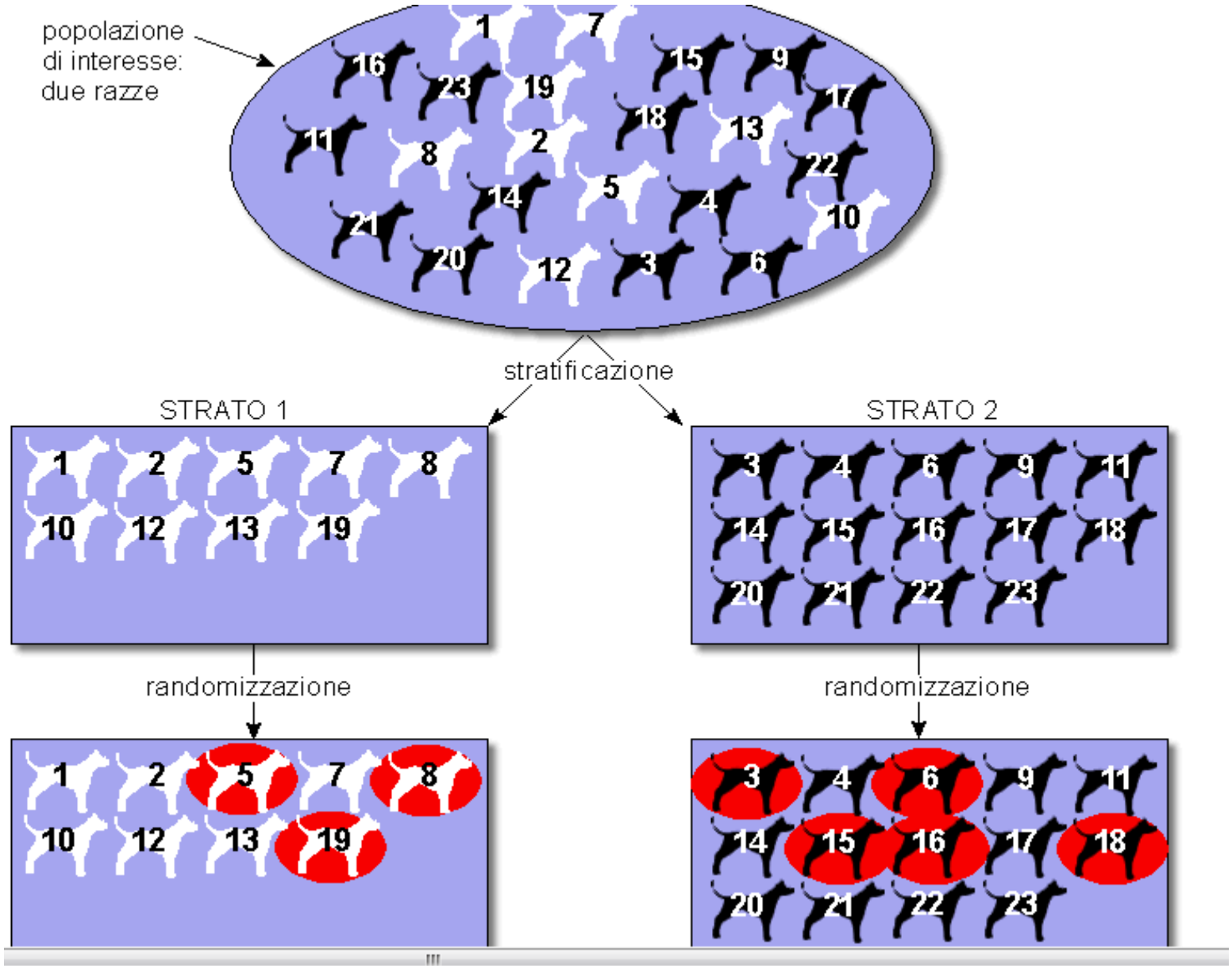
Campionamento a strati



Campionamento a strati






popolazione
di interesse:
due razze
stratificazione
STRATO 1
STRATO 2
randomizzazione
randomizzazione

Campionamento a due stadi o a grappolo

Si vuole determinare il tempo medio di degenza dei pazienti negli ospedali italiani nel 2019.

Poichè prendere in considerazione tutti gli ospedali può essere troppo dipendioso, si sceglie a caso un certo numero di ospedali e in ciascuno di essi si estrae un campione (CCS).

(E' meno preciso dei campionamenti visti prima, pertanto il suo uso è giustificato da ragioni economiche)

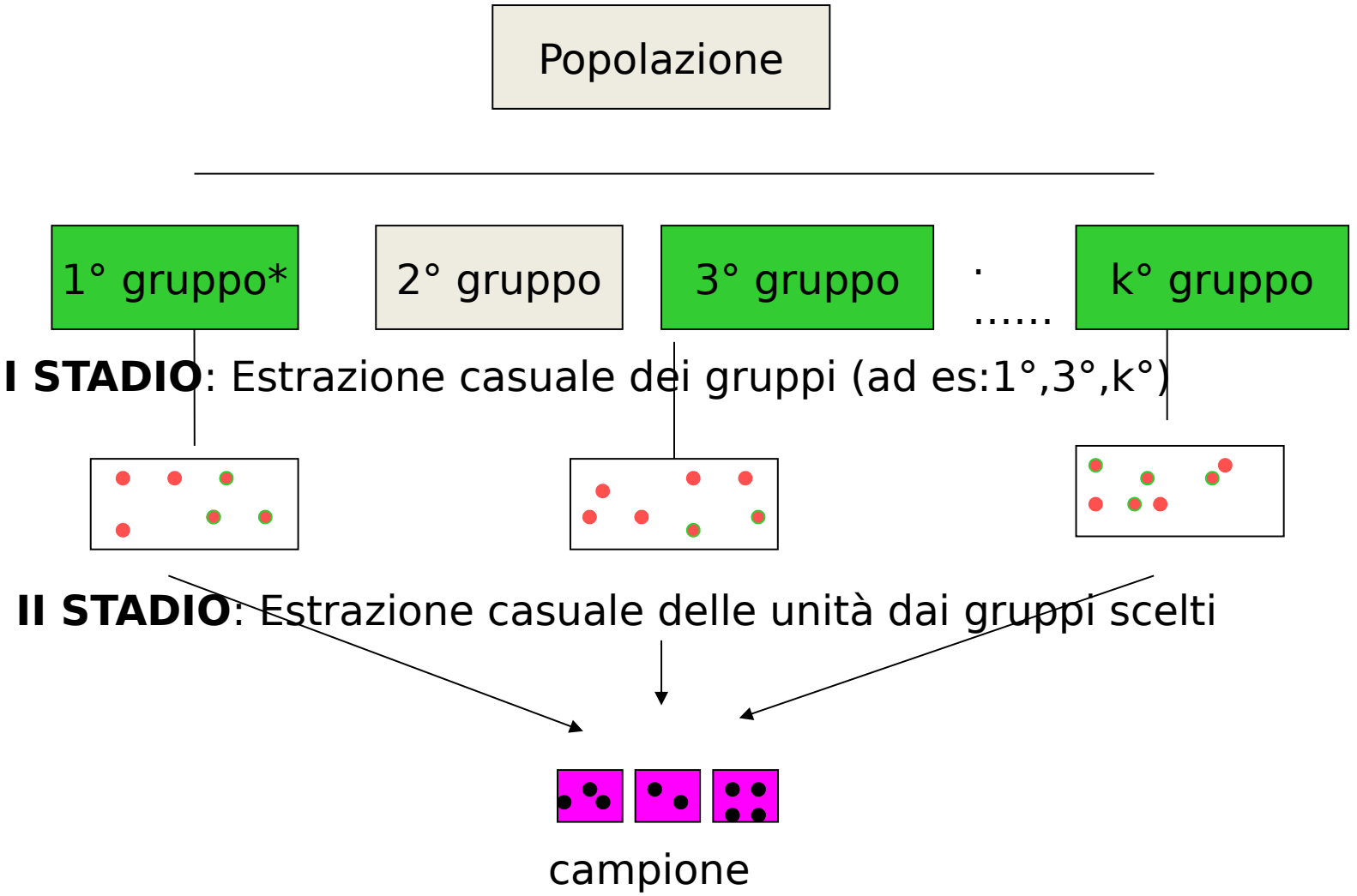
Campionamento a due stadi

Viene usato se si deve campionare da una popolazione molto numerosa.

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 campione casuale in ciascuno di questi n gruppi (2°stadio).

Il campione risultante sarà non distorto se i gruppi sono omogenei tra loro.

Campionamento a grappoli (a 2 stadi)

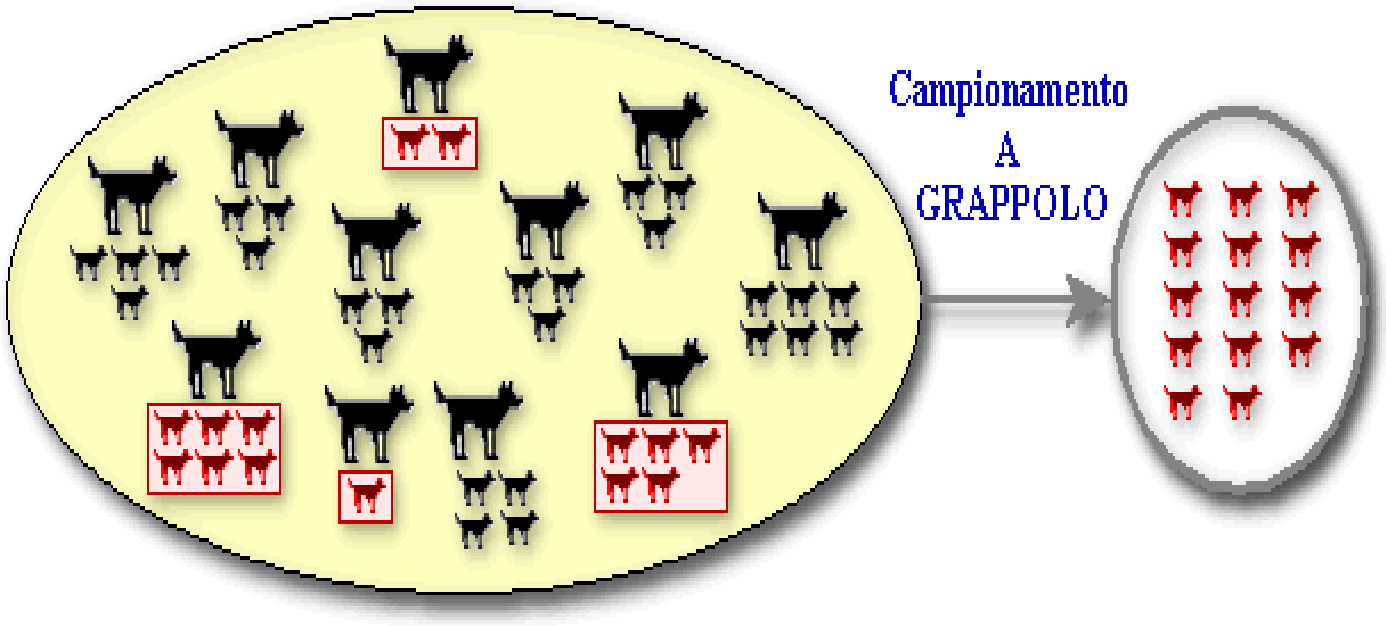


***si parla di grappolo o gruppo di unità della popolazione**

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*.



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.

Campionamento a grappolo: commenti.

- E’ meno preciso dei campionamenti visti prima, pertanto il suo uso è giustificato se la popolazione è molto numerosa o se ci sono ragioni economiche.

- Il grappolo è di fatto pensato come 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.

Esempio

Si vuole stimare la lunghezza media delle foglie di ulivo presenti in una certa piantagione.

Per farlo si sceglie a caso un certo numero di alberi e da ciascuno di essi si preleva (a caso) una o più foglie (campionamento a due stadi).

Ipotesi: gli alberi sono “omogenei” tra di loro.

Esempio

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

Poichè l'abbondanza di pesci dipende fortemente dall'habitat qualitativo e in particolare dalla profondità marina, si è eseguito un campionamento a strati dove ogni strato è rappresentato da un particolare habitat.

No. of strata in which species occur (u_i)	Frequency (No. of species) (f_i)	Percentage of species $\left(\frac{f_i}{n} \times 100\right)$	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	

Campione non rappresentativo-Distorsione

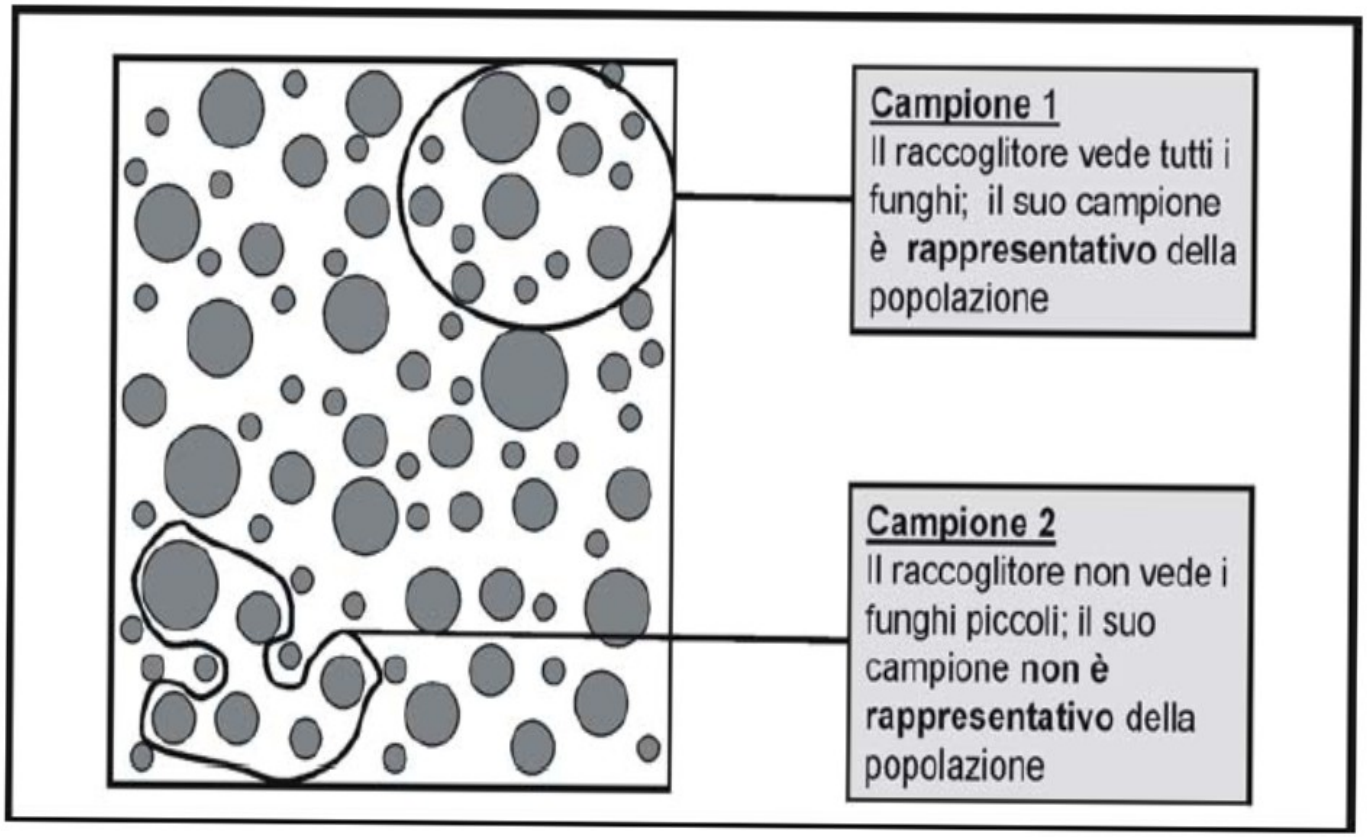


Figura 1.3. Due campioni di funghi raccolti da due sperimentatori, il secondo dei quali molto miope. Il primo campione è rappresentativo della popolazione dei funghi, il secondo non lo è perché il raccoglitore miope non vedrà i funghi più piccoli.

Esempi di campionamenti non probabilistici.

Campionamento accidentale:

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 selezionano all'interno della popolazione gli elementi che il ricercatore ritiene rappresentativi per gli obiettivi della ricerca.

Esercizio

In un grande ospedale si vuole stimare il valor medio μ della pressione sanguigna (mmHg) di tutti i pazienti maschi (di età 50-60 anni) subito dopo un particolare intervento chirurgico al cuore; lo studio prenderà in considerazione i dati raccolti negli ultimi tre anni. Per questo studio si vuole estrarre un campione di pazienti di dimensione 50 per calcolare la media campionaria.

- se si dispone dell'elenco in ordine alfabetico di tutti i pazienti maschi che hanno subito questo particolare intervento al cuore (di età 50-60 anni), che tipo di campione si può estrarre?
- Come effettuereste, in pratica, questo campionamento?
- Qual è la probabilità che il paziente più anziano faccia parte del campione?