fondamenti di Matematica e biostatistica biotecnologie

Biotecnologie (Università degli Studi di Siena)

Variabili e misure

DEFINIZIONE DI STATISTICA: la raccolta, la presentazione e la elaborazione numerica delle informazioni, per agevolare l'analisi dei dati ed i processi decisionali.... ma anche insieme di metodi e norme atti a cogliere, analizzare ed interpretare dati numerici o numerabili

Il problema della misurazione

Il dato statistico è rappresentato dall'assegnazione di numeri a variabili o più in generale a concetti. Esistono diversi modi per rilevare variabili o concetti e anche la loro misurazione può presentare diversi gradi di difficoltà

- Quando si compie una rilevazione di dati, che in seguito saranno elaborati mediante tecniche descrittive, in realtà si rilevano i valori o le modalità con le quali si presentano determinate VARIABILI.
- Sesso, età, altezza, colore degli occhi ed altre ancora non sono che variabili le cui determinazioni possono essere oggetto di una rilevazione statistica.
- Ognuna di queste presenta particolari caratteristiche che permettono di classificarla in una ben precisa categoria.

- Tipo di variabili

- Qualitativa: categorie, attributi, non numerica

-Nominale: le categorie rappresentano le modalità con le quali la variabile si presenta
-Ordinale: come sopra, ma le categorie sono ordinabili secondo una scala logica

- Quantitativa: numerica, conteggi, misurazioni

-Discreta: numeri che mostrano solo determinati valori puntuali
-Continua: numeri che possono assumere tutti i valori infiniti all'interno di un dato intervallo

A quale categoria appartengono le seguenti variabili?

VARIABILE	DESCRIZIONE	DETERMINAZIONI	CATEGORIA
X	Numero di prestazioni radiologiche dell'U.O. nel mese	(103; 250; 125; ...)	Quantitativa discreta
Y	Tipo di prestazione ricevuta	(RX; TAC; RSM; ...)	Quantitativa nominale
Z	Costo mensile delle prestazioni (migliaia €)	(53,1; 125,7; 87,9; ...)	Quantitativa continua
K	Classe d'età degli studenti	(Meno di 30 anni; tra 30 e 55 anni; ...)	Qualitativa ordinale
W	Grado di soddisfazione del servizio ricevuto	(Scarso; buono; ...)	Qualitativa ordinale

Le misure che vengono effettuate su più soggetti in relazione alla stessa variabile sono rappresentate in forma vettoriale.
-n=numero di soggetti su cui viene effettuata la misurazione (N se il gruppo coincide con la popolazione oggetto dello studio)
-X=è la variabile misurata
-xi=sono le misurazioni della variabile realizzate sui soggetti con i=1,..,n

$$X = (x1, x2, \dots , xn)$$

Se per esempio fosse stata realizzata un'indagine su n=504 soggetti e fossero state misurate le seguenti variabili ...
X = genere
Y = sesso
K = pressione sistolica

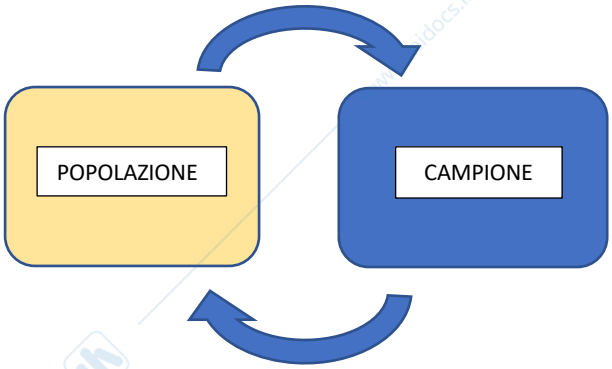
X = (x1, x2, ... , x504)=(32, 64, ... , 56)
Y = (y1, y2, ... , y504)=(maschio, femmina, ... , femmina)
K = (k1, k2, ... , k504)=(135, 109, ... , 151)

Esempio: il tirocinante e le infezioni ospedaliere
Un tirocinante viene incaricato dal suo tutor di analizzare un gruppo di record clinici riguardanti alcuni ricoverati (n=15) degli ultimi mesi per studiare le infezioni ospedaliere che si verificano in reparto. Il tutor chiede al tirocinante di estrarre le seguenti variabili:
X=Età in anni={0, ... , 120}
Y=Residenza={Siena/Altra provincia/Estero}
Z=Sesso={F/M}
K=Infezione ospedaliera={No/Si}
J=Soddisfazione del servizio={Molto/Abbastanza/Poco/Per niente}
W=Leucociti all'ingresso in reparto={0,00 milioni/L, ... , 20,00 milioni/L}
V=Altre infezioni pregresse al ricovero={Spesso/Qualche volta/Raramente/Mai}

Esempio: il tirocinante e le infezioni ospedaliere
Il tirocinante, al termine dell'analisi delle cartelle cliniche costruisce questo archivio:

PAZIENTE	X	Y	Z	K	J	W	V
1	35	Siena	F	No	Molto	7,45	Raramente
2	61	Altra prov	M	No	Abbastanza	7,45	Qualche volta
3	56	Altra prov	F	No	Molto	8,33	Raramente
4	44	Siena	M	No	Abbastanza	5,23	Raramente
5	73	Siena	M	Si	Poco	15,67	Spesso
6	85	Siena	F	Si	Molto	13,58	Qualche volta
7	49	Altra prov	F	No	Per niente	9,31	Mai
8	57	Estero	M	No	Abbastanza	7,66	Raramente
9	61	Siena	F	No	Per niente	4,89	Mai
10	72	Siena	M	Si	Molto	21,65	Qualche volta
11	68	Altra prov	F	Si	Abbastanza	11,25	Spesso
12	51	Altra prov	F	No	Per niente	6,73	Mai
13	87	Altra prov	F	Si	Abbastanza	9,94	Spesso
14	65	Siena	M	Si	Per niente	17,15	Qualche volta
15	58	Estero	M	No	Poco	8,32	Raramente

X è la variabile
N la numerosità



x è la variabile
n la numerosità

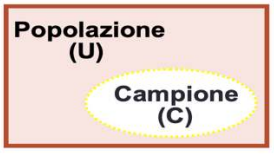
• Il campione: è semplicemente l'insieme degli elementi (detti anche unità campionarie o sperimentali) sui quali effettuiamo misure o osservazioni (per esempio, 20 marmotte catturate con trappole). Costituisce una frazione della popolazione statistica, un gruppo più grande di elementi che potenzialmente potremmo osservare e misurare.
• La popolazione: può corrispondere ad un insieme finito di individui che hanno alcune caratteristiche in comune (per esempio, tutte le marmotte che vivono nelle Alpi). In generale, comunque, si preferisce definire la popolazione statistica come un insieme infinito di elementi

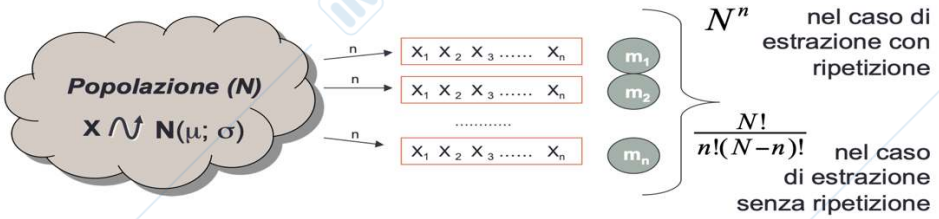
- Esempi di popolazioni:
• tutti i gatti caduti dagli edifici di New York
• tutti i geni del genoma umano
• tutti gli individui maggiorenni in Australia
• tutti i serpenti volanti del paradiso nel Borneo
• tutti i bambini asmatici di Milano.
- Esempi di campioni:
• i gatti caduti portati in un singolo ambulatorio in un certo intervallo di tempo
• 20 geni umani
• i frequentatori maggiorenni di un pub in Australia
• otto serpenti volanti del Borneo
• 50 bambini asmatici a Milano

Popolazioni e campioni

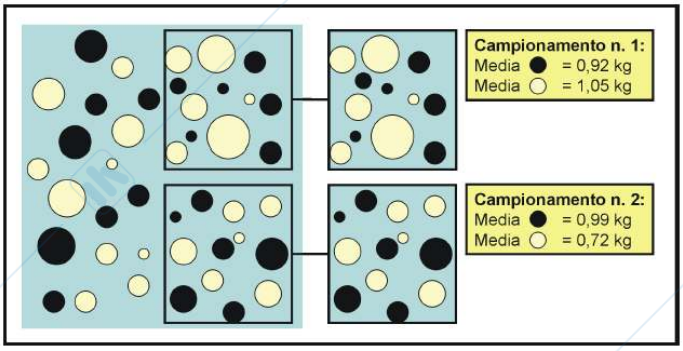
Come detto, il campione è sempre un sottoinsieme della popolazione

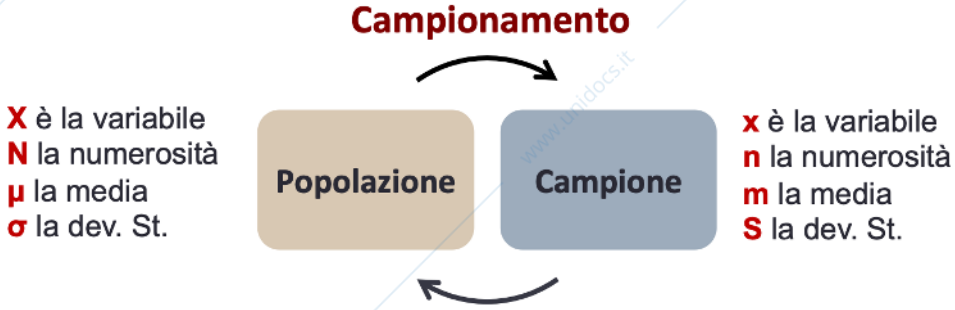
$C \subset U$





La variabilità sperimentale può portare per puro effetto del caso a risultati diversi in diversi campionamenti.





- La rappresentatività del campione rispetto alla popolazione è influenzata in maniera determinante anche dal criterio di campionamento che viene utilizzato.
- Alcune tecniche di campionamento: • Casuale semplice,
- Casuale semplice senza ripetizione, • Sistematico,
- Stratificato,
- Grappolo.

Il campionamento casuale semplice

- Data una popolazione P composta da N elementi, si tratta di selezionare in modo casuale un campione di ordine n.
- È necessario...
- conoscere la numerosità della popolazione,
- numerare in una lista i suoi componenti,
- estrarre il numero da un'urna, usare opportune tavole dei numeri casuali, oppure un qualsiasi software dotate di un algoritmo per la generazione dei numeri casuali.
- Ogni elemento ha la stessa probabilità di essere estratto, in quanto l'eventuale pallina ogni volta viene reinserita nell'urna.

Le tavole dei numeri casuali

- Per esempio, se la popolazione da cui campionare ha numerosità fosse N=1000.
- È necessario prendere in considerazione colonne con 3 cifre.
- I primi elementi estratti saranno il 152, il 486, l'860, ecc.. Esaurita la prima colonna si può passare alla seconda fino che ho estratto tutti gli n elementi del campione.

1528	7066	3672
4864	3988	5464
8609	1770	3760
9091	1233	1559
7044	0734	9303
9624	9617	8414
5874	6322	0229
4449	5050	5690
9871	4191	6591
5668	3565	4693
0862	3582	4553
8539	4972	3808

Il campionamento casuale semplice non ripetitivo

• Data una popolazione P composta da N elementi, anche in questo caso si tratta di selezionare in modo casuale un campione di ordine n.
• È necessario...
- conoscere la numerosità della popolazione,
- numerare in una lista i suoi componenti,
- estrarre il numero da un'urna, usare opportune tavole dei numeri casuali, oppure un qualsiasi software dotate di un algoritmo per la generazione dei numeri casuali..
• Ogni elemento ha la stessa probabilità di essere estratto, ma questa aumenta in quanto non viene reinseritol'elemento estratto.
• Selezionare da una popolazione P composta da N elementi, un campione casuale di ordine n.
• È necessario...
- conoscere la numerosità della popolazione,
- ordinare la lista degli elementi della popolazione secondo una caratteristica correlata con l'oggetto della ricerca,
- individuare il valore del parametro K (passo di campionamento) ottenuto da rapporto tra N e n.
- estrarre casualmente il primo elemento tra i primi k della lista e successivamente estrarre ogni k elementi.
• Per esempio:
- la popolazione P è composta da N=1.000 elementi,
- la numerosità del campione è n=100,
- il passo di campionamento è k=10 (N/n),
- estraggo un numero casuale compreso tra 1 e 10,
- se per ipotesi il numero estratto fosse 6, il sesto elemento della lista sarà il primo elemento del campione; i successivi saranno il 16° (6+k), il 26° (16+k), e così via ...

Il campionamento stratificato

• Selezionare da una popolazione P composta da N elementi, un campione casuale di ordine n. Questo metodo è fortemente consigliato quando la caratteristica esaminata è presente in modo molto eterogeneo nella popolazione.
• È necessario...
- conoscere la numerosità della popolazione,
- procedere ad una divisione della popolazione di N elementi in L sottopopolazioni, dette strati.

• Ogni strato contiene Nh (h=1, ... ,L) elementi omogenei rispetto ad una determinata caratteristica e tali che:

$$\sum_{h=1}^{L} N_h = N$$

• L'efficacia del metodo aumenta se...
- gli strati hanno una grossa omogeneità all'interno (bassa varianza interna agli strati),
- gli strati hanno una grossa eterogeneità tra di loro (alta varianza tra gli strati).

• Per esempio: gli strati hanno una grossa omogeneità all'interno (bassa varianza interna agli strati),
• la popolazione P è composta da N=10.000 elementi,
• la numerosità del campione è n=1.000,
• la disposizione negli strati della numerosità della popolazione è la seguente:
N_1=6.000 N_2=3.000 N_3=1.000
$N_1 + N_2 + N_3$ =10.000
• Assegnata la popolazione negli strati, generalmente ho due modi per distribuire la numerosità campionaria:
- Assegnamento proporzionale: si assegnerà una numerosità campionaria ad ogni strato in modo proporzionale alla densità di popolazione
- Assegnamento uniforme: si assegnerà una quota costante di numerosità campionaria ad ogni strato.

• Per selezionare gli elementi che faranno parte del campione sarà necessario utilizzare le tecniche del campionamento casuale, semplice o non ripetitivo

Il campionamento a grappolo

• Tutte le tecniche di campionamento precedenti possono essere utilizzate quando la popolazione ha una numerosità relativamente bassa (numerabile).
• Se l'universo è composto da un numero molto elevato di soggetti (non numerabile), può risultare molto difficile se non impossibile estrarre un campione con le tecniche esaminate. È dunque necessario ricorrere al campionamento a grappolo.
• È necessario...
- suddividere la popolazione originaria N in un numero M di sottopopolazioni chiamate grappoli (Unità Primarie di Campionamento, UPC), tali che:

- estrarre un campione m di UPC tra gli M in cui ho suddiviso la popolazione;
- nelle m UPC si estraggono i campioni di numerosità ni<Ni ottenendo le cosiddette Unità Secondarie di Campionamento (USC). Ovviamente:

• Nel seguente esempio si ipotizza la divisione in M=10 grappoli della popolazione N, con l'estrazione di un campione di numerosità n da un numero di m=3 grappoli:

La rappresentatività del campione

Le osservazioni che vogliamo analizzare possono provenire da un campionamento (per esempio, il peso delle trote, ma anche le concentrazioni di un certo composto chimico in diversi terreni), oppure da un esperimento (per esempio, lo stato di salute dei pazienti trattati o meno con un farmaco). In entrambi i casi, il campione dei dati deve essere rappresentativo della popolazione.
Le fonti di distorsione del campione possono essere molteplici, ma sono campioni distorti quelli di convenienza o quelli composti da volontari (comunque non selezionati con metodi casuali), oppure:

- quando tutte le unità campionarie non hanno la stessa probabilità di essere estratte,
- quando la probabilità di estrarre una unità campionaria dipende dalla possibilità che ne sia stata estratta un'altra.

La statistica moderna può essere distinta in tre discipline:
La statistica descrittiva spiega come i dati raccolti devono essere riportati in tabella, rappresentati in grafici e sintetizzati in indici matematici, allo scopo di individuare le caratteristiche fondamentali del campione.
La statistica matematica presenta le distribuzioni teoriche sia per misure discrete sia per misure continue, allo scopo di illustrarne le caratteristiche fondamentali, le relazioni che esistono tra esse, gli usi possibili;
L'inferenza statistica serve per la verifica delle ipotesi e può essere distinta in vari capitoli, in rapporto alle caratteristiche dei dati (se permettono o meno il ricorso alla distribuzione normale: statistica parametrica e non parametrica) e al numero di variabili (se una, due o più: statistica univariata, bivariata, multivariata).