STATISTICA NOZIONI BASE

SAPIENZA
UNIVERSITÀ DI ROMA

annarita.vestri@uniroma1.it

Statistica

Descrittiva

Ho un insieme di dati e li voglio descrivere, sintetizzare e commentare

Inferenziale

Ho un insieme di dati e li utilizzo per fare induzione e previsione

STATISTICA

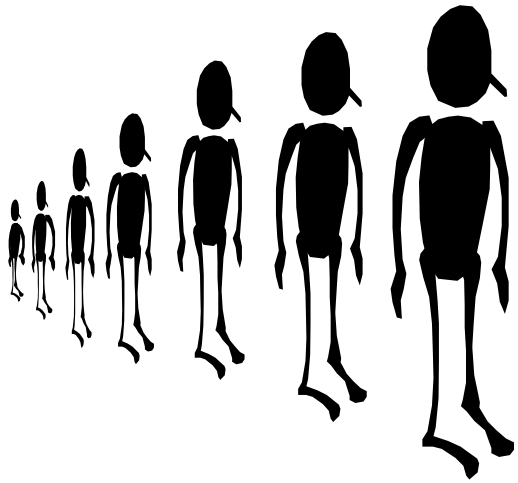
insieme di procedure finalizzate al trattamento di informazioni relative a fenomeni collettivi, che si manifestano con determinazioni tipicamente non costanti

Oggetto della STATISTICA sono quei fenomeni che variano all'interno di un collettivo di riferimento, la POPOLAZIONE STATISTICA, costituito da UNITA' STATISTICHE o elementari.

POPOLAZIONE STATISTICA

qualsiasi insieme di persone, animali, piante o cose da cui possono essere raccolte le informazioni

oggetto di interesse dell'indagine: insieme di entità sulle cui caratteristiche vogliamo trarre conclusioni



UNITA' STATISTICA

elemento di base della popolazione sul quale viene effettuata la rilevazione o la misurazione di uno o più fenomeni oggetto di studio

oggetto della raccolta dei dati, detentore dell'informazione che vogliamo rilevare e analizzare

POPOLAZIONI DI INTERESSE PER LA STATISTICA APPLICATA ALLA MEDICINA:

- **insieme di esseri umani** (residenti in una certa area; soggetti sani, malati oppure deceduti);
- **insieme di unità amministrative** (reparti, ospedali, comuni);

- **ematocrito dei ricoverati presso il reparto di ematologia del policlinico Umberto I nell'anno 2019**
- **tempi di sopravvivenza dopo il trapianto di cuore…**
- **Numero dei nati nell'ospedale di una provincia nel periodo 2010-2020**

POPOLAZIONI DI INTERESSE PER LA STATISTICA APPLICATA ALLA MEDICINA:

esempio (popolazione di residenti - sani o malati):

indagine ISAYA → adulti di età 20-44 anni residenti in Italia nel 1998-2000 selezionati indipendentemente dallo stato di salute

esempio (popolazione di soggetti sani):

sperimentazione sull'effetto del fluoro nel prevenire l'insorgenza di carie nei bambini → bambini sani (senza carie)

esempio (popolazione di soggetti malati):

sperimentazione sull'effetto di un chemioterapico per la cura di una particolare patologia tumorale → soggetti che presentano la patologia

Molte ricerche vengono programmate con lo scopo di pervenire a **conclusioni generali**, valide per tutte le unità statistiche della popolazione, sfruttando i risultati ottenuti da un numero ridotto di osservazioni

CAMPIONE STATISTICO:

sottoinsieme di unità statistiche appartenenti alla popolazione che vengono selezionate per l'analisi (sono quelle realmente studiate)

GENERALIZZAZIONE DELLE CONCLUSIONI

NB: il campione deve essere **rappresentativo** (stesse caratteristiche della popolazione dalla quale è stato estratto)

CAMPIONAMENTO CASUALE

SCHEMA LOGICO DELLA STATISTICA

CAMPIONAMENTO
teoria delle probabilità

CAMPIONE

STATISTICA DESCRITTIVA

Sintesi e presentazione dei dati raccolti sul campione

STATISTICA INFERENZIALE
generalizzazione delle informazioni raccolte sul campione

Studio delle caratteristiche della popolazione

Nozioni di base

Caratteri

aspetti del fenomeno oggetto di studio

Modalità (x_i)

modo di manifestarsi del carattere

La classificazione dei caratteri

I caratteri possono essere classificati in:

-Caratteri **qualitativi** distinti in:

- ***ordinabili*:** è possibile *ordinare* le modalità del carattere in senso crescente o decrescente (es: titolo di studio, livello di gravità della diagnosi...);
- ***sconnessi*:** non c'è alcun ordinamento intrinseco tra le modalità (es: colore degli occhi, sesso,stato civile, religione...);

- Caratteri **quantitativi** distinti in:

- ***discreti*:** le modalità del carattere sono numeri interi (es: numero di medici, numero di figli per donna, numero dei soggetti guariti,….)
- ***continui***: le modalità del carattere sono misurate su una scala continua (es: peso, altezza...).

Alla base di tale classificazione dei caratteri vi è la **'scala di misura'** con cui sono espresse le modalità: se attraverso dei **numeri** o delle '**etichette**'.

Classificazione dei caratteri e scala di misura

CARATTERE		SCALA
qualitativo	Sconnesso	Nominale
	Ordinabile	Ordinale
quantitativo		Ad intervalli (scala numerica discreta o continua)

Operazioni che è possibile fare sui caratteri in base alla loro classificazione

Operazioni sulle modalità del carattere	Carattere		
	qualitativi		Quantitativi *(discreti/continui)*
	sconnessi	*ordinabili*	
= ; ≠	si	si	si
> ; <	no	si	si
+ ; -	no	no	si

Nozioni di base

In ogni collettivo ogni modalità può presentarsi più volte

Il numero delle volte che una modalità si presenta prende il nome di frequenza assoluta (n_i) (o semplicemente frequenza)

Nozioni di base

L’insieme delle modalità e delle frequenze costituisce la distribuzione statistica

Nel caso di un carattere qualitativo (mutabile) la distribuzione si chiamerà serie

Nel caso di un carattere quantitativo (variabile) la distribuzione viene detta seriazione

Tabulazione dei dati

I dati raccolti vengono riportati in apposite tabelle

Si distinguono:

Tabelle semplici (singole): relative ad un unico carattere

Tabelle doppie (a doppia entrata): relative a 2 caratteri "incrociati"

Tabelle multiple: relative a 3 o più caratteri

Distribuzioni di frequenza

frequenze assolute

Una distribuzione di frequenza è la determinazione della frequenza con cui compare, in una certa popolazione, ciascun valore di una data variabile.

3	2	4	5
7	7	4	6
6	2	3	6
7	5	5	4
1	6	6	5
9	4	2	6
5	8	7	4
4	6	8	5
5	5	5	5
6	3	4	5

Periodi di incubazione di una malattia in 40 pazienti

Giorni (x)	Frequenza (ni)
1	1
2	3
3	3
4	7
5	11
6	8
7	4
8	2
9	1
Totale	40

Distribuzione di frequenze dei periodi di incubazione della malattia in 40 pazienti

Distribuzione di frequenze dei periodi di incubazione della malattia in 40 pazienti

Giorni (xi)	ni	fi	fi%
1	1	0,025	2,5
2	3	0,075	7,5
3	3	0,075	7,5
4	7	0,175	18
5	11	0,275	28
6	8	0,2	20
7	4	0,1	10
8	2	0,05	5
9	1	0,025	2,5
totale	40	1	100

FREQUENZA RELATIVA: $(f_i = n_i / n)$

rapporto tra il numero di osservazioni corrispondente ai diversi valori (modalità/intervalli di classe) della variabile e la dimensione campionaria

$\Rightarrow \quad 0 \leq f_i \leq 1$

$\Rightarrow \quad \Sigma^{K}_{i=1} f_i = f_1 + f_2 + .. + f_K = 1$

FREQUENZA RELATIVA PERCENTUALE: $(f_i^{\%} = n_i / n * 100)$

indica quanto volte un fenomeno si manifesta su una casistica di 100 osservazioni

$\Rightarrow \quad 0\% \leq f_i^{\%} \leq 100\%$

$\Rightarrow \quad \Sigma^{K}_{i=1} f_i{}^{\%} = f_1{}^{\%} + f_2{}^{\%} + .. + f_K{}^{\%} = 100\%$

TABELLA 1 - DISTRIBUZIONE DEL MOTIVO PER CUI I I CASI DI COVID-19 DIAGNOSTICATI IN ITALIA SONO STATI TESTATI
PERIODO: 21 SETTEMBRE - 4 OTTOBRE 2020

Motivo del test	Casi	
	N	%
Screening	8.275	31,0
Contact tracing	8.997	33,6
Paziente con sintomi	7.803	29,2
Non noto	1.654	6,2
Totale	26.729	

Prodotto dall'Istituto Superiore di Sanità (ISS), Roma, 6 ottobre 2020

6

FREQUENZA CUMULATA

FREQUENZA ASSOLUTA CUMULATA (F_i)

numero di osservazioni il cui valore è **inferiore o uguale** ad una data modalità o a un dato valore x_i

$$F_i(-\infty)=0 \qquad F_i(+\infty)=n$$

FREQUENZA RELATIVA CUMULATA

($P_i = F_i / n$; $P_i^{\%} = F_i / n * 100\%$)

$$P_i(-\infty)=0 \qquad P_i(+\infty)=1$$

Distribuzioni di frequenza frequenze cumulate
si sommano le frequenze assolute iniziando dalla prima

giorni	ni	Fi	Fi%
1	1	1	2,5
2	3	4	10
3	3	7	17,5
4	7	14	35
5	11	25	62,5
6	8	33	82,5
7	4	37	92,5
8	2	39	97,5
9	1	40	100
totale	40		

Esempio: I dati seguenti si riferiscono al grado del trauma in 100 ricoverati al pronto soccorso:

0 2 1 1 1	2 0 0 1 0	1 1 0 0 0	3 1 2 0 1
1 0 0 1 0	1 1 0 2 0	0 0 1 0 1	0 2 1 2 0
0 2 0 1 0	1 0 1 0 3	1 2 0 0 0	0 1 0 0 0
1 0 1 0 1	0 2 0 1 2	1 2 0 1 0	2 2 1 0 1
0 0 0 0 4	0 1 1 2 0	0 2 1 0 2	0 0 2 1 0

modalità	conteggio	frequenza
	conteggio delle osservazioni	
assente	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 | | |	48
lieve	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 | |	32
grave	𝍸 𝍸 𝍸 | |	17
lesioni permanenti	| |	2
decesso	|	1
		100

modalità	tally	frequenza
assente	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |||	48
lieve	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ||	32
grave	𝍸 𝍸 𝍸 ||	17
lesioni permanenti	||	2
decesso	|	1
		100

MODALITA'	frequenza assoluta n_i	frequenza relativa n_i / n
assente	48	48/100 = 0,48
lieve	32	0,32
grave	17	0,17
lesioni permanenti	2	0,02
decesso	1	0,01
TOTALE	100	

Costruzione della tabella e calcolo di frequenze relative

esempio (grado del trauma):

distribuzione di frequenza assoluta, relativa e cumulativa della variabile "grado del trauma"

valore x_i	assoluta n_i	relativa p_i	relativa percentuale p_i (%)	assoluta cumulata N_i	relativa cumulata P_i	relativa cumulata percentuale P_i (%)
assente	48	0.48	48%	48	48 / 100 = 0.48	0.48 * 100 = 48%
lieve	32	0.32	32%	48 + 32 = 80	80 / 100 = 0.80	0.80 * 100 = 80%
grave	17	0.17	17%	80 + 17 = 97	97 / 100 = 0.97	0.97 * 100 = 97%
lesioni permanenti	2	0.02	2%	97 + 2 = 99	99 / 100 = 0.99	0.99 * 100 = 99%
decesso	1	0.01	1%	99 + 1 = 100	100 / 100 = 1	1 * 100 = 100%
TOTALE	100	1	100%			

La matrice dei dati

I dati codificati in una in una rilevazione statistica su n unità statistiche studiando x variabili sono raccolti in forma di tabella (matrice di dati)

N	sesso	Età (anni)	Peso (Kg)	Titolo di studio	n.° ricoveri
1	M	42	83	laurea	2
2	F	48	65	diploma	1
...	...	...	...	.....	.....
n	F	61	79	Licenza media inferiore	4

La matrice dei dati

Ogni riga corrisponde ad una unità statistica

N	sesso	Età (anni)	Peso (Kg)	Titolo di studio	n.° ricoveri
1	M	42	83	laurea	2
2	F	48	65	diploma	1
...	...	...	...	.....	.....
n	F	61	79	Licenza media inferiore	4

La matrice dei dati

Ogni colonna rappresenta una variabile

N	sesso	Età (anni)	Peso (Kg)	Titolo di studio	n.° ricoveri
1	M	42	83	laurea	2
2	F	48	65	diploma	1
...	...	...	...	.....	.....
n	F	61	79	Licenza media inferiore	4