

LAVORARE CON I FOGLI DI CALCOLO

- ▶ La Statistica di Laboratorio nei fogli di calcolo
- ▶ 1 e 4 dicembre 2017
- ▶ Istituto Zooprofilattico Sperimentale del Lazio e della Toscana M. Aleandri

Che cos'è la statistica

Scienza che ha come scopo la conoscenza quantitativa dei fenomeni collettivi.

E' un insieme di tecniche e metodologie sviluppate per aiutarci a rispondere a domande precise relative a fenomeni sociali, economici, di salute pubblica.....

La STATISTICA si occupa di raccogliere ed elaborare informazioni su un fenomeno che si vuole studiare.

La raccolta e l'elaborazione dei dati costituiscono l'INDAGINE STATISTICA vera e propria



Di che cosa si occupa la statistica?

- ▀ La statistica è la scienza che permette di trarre conclusioni generali relative ad un insieme numeroso di dati (popolazione, campione)
- ▀ La popolazione è la totalità degli oggetti o individui a cui si riferisce l'indagine statistica
- ▀ Il campione è una parte della popolazione che deve rappresentare in modo significativo l'intera popolazione

La Statistica descrittiva

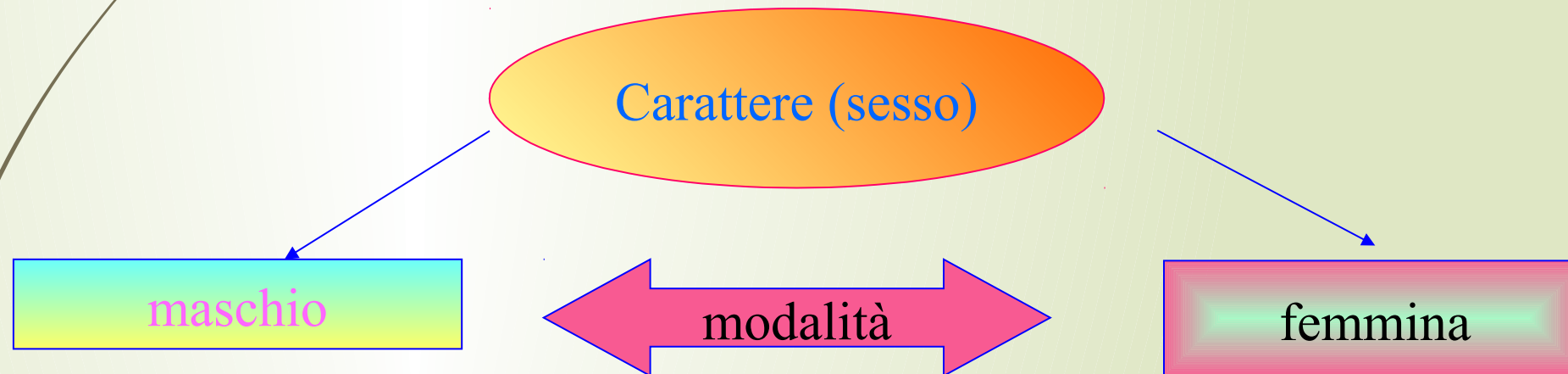
- Lo scopo della statistica descrittiva è quello di descrivere efficacemente una grande massa di dati mediante tabelle e grafici e di sintetizzare le informazioni in indici matematici in modo da individuare le caratteristiche fondamentali del campione

Il linguaggio della Statistica descrittiva

- Popolazione statistica: è l'insieme di tutti i possibili oggetti dell'indagine statistica
- Individuo (o unità statistica): è un qualsiasi elemento della popolazione
- Variabile: è una qualsiasi caratteristica di ogni individuo della popolazione, soggetta a variazioni di valore da un individuo all'altro

Carattere

- Su ogni unità statistica vengono rilevati diversi aspetti ciascuno dei quali è detto carattere. Il carattere è un attributo qualunque posseduto da una unità statistica. Esso può manifestarsi in diverse maniere ognuna delle quali è detta modalità



Le caratteristiche

- ▶ I caratteri o variabili che rappresentano l'oggetto di un'analisi statistica possono esprimere delle qualità (caratteri qualitativi) o delle quantità (caratteri quantitativi)
- ▶ I caratteri quantitativi a loro volta possono essere discreti, quando sono espressi da un numero intero naturale N , o continui quando sono espressi da un numero reale R
- ▶ I caratteri qualitativi possono essere ordinati, (es: laureato, diplomato...), oppure sconnessi, quando non seguono un ordine.

FASI DELLA RICERCA STATISTICA

L'indagine statistica si realizza in cinque fasi:

- Pianificazione
- Rilevazione
- Elaborazione
- Presentazione
- Interpretazione

Pianificazione

La pianificazione consiste nella definizione del fenomeno da studiare e nell'individuazione della popolazione o del campione a cui si riferisce, nella scelta dei caratteri del collettivo che interessano lo studio e nella definizione delle relative modalità o dei processi di misura.

(Esempio: preparazione test da somministrare al campione)

Rilevazione

La rilevazione è quel complesso di operazioni attraverso i quali si acquisiscono le informazioni sulle caratteristiche di interesse.

Da questa fase si acquisiscono i dati statistici elementari o dati grezzi, che saranno catalogati in tabelle.

La rilevazione si dice totale se interessa tutta la popolazione

La rilevazione si dice parziale se si limita ad esaminare una parte soltanto delle unità statistiche, cioè un campione.

(Esempio: somministrare il test al campione scelto)



Rilevazione dei dati.

Ci permette di pervenire alla conoscenza delle caratteristiche delle singole unità statistiche. Essa può dividersi in tre sotto fasi:

- raccolta;
- classificazione;
- rappresentazione grafica dei dati.



Elaborazione

L'elaborazione è quel complesso di operazioni attraverso le quali vengono codificati i dati e sintetizzati in tabelle o grafici, più facilmente interpretabili.

Presentazione

La presentazione è l'esposizione dei dati, attraverso le tabelle o i grafici preparati.

Interpretazione

L'interpretazione è la spiegazione delle tabelle e dei grafici e dei risultati finali ottenuti, con osservazioni ed eventuali collegamenti con altre indagini simili.

Unità statistica, variabili e modalità

L unità statistica è il possessore del fenomeno individuale che costituisce il fenomeno collettivo (impresa-allevamento-ospedale)

Variabile o carattere è la caratteristica o l'attributo che si osserva sulle unità statistiche (età- sesso- data di nascita)

Le variabili sono presenti nell'unità statistica con una determinata modalità (30 anni-maschio-2750 gr)

Tipologie di variabili

- ✓ Quantitative (o numeriche)

Specificano le unità con un numero positivo o negativo,
Intero o decimale, definito relativamente ad una scala di misura

- ✓ Qualitative (o categoriche)

Specificano le unità con termini descrittivi, con un attributo, un aggettivo o una qualità

Variabili qualitative

Dicotomiche :

(assumono due modalità mutuamente esclusive)

ES: Sesso: maschio/femmina

Stato: Vivo/morto

Presenza infezione (o malattia):si/no

Politomiche

(assumono più di due modalità mutuamente esclusive e non ordinabili)

ES: Stato civile: celibe/coniugato/separato/vedovo

Gruppo sanguigno: A,B, AB,0

Tipo di cavallo: da sella /da galoppo/da trotto

Ordinali

(assumono due o più modalità ordinabili)

ES: livello d'istruzione: elementare, media ,superiore ,diploma, laurea

Gravità di un sintomo: lieve/media/forte

Variabili quantitative

Discrete

Possono essere messe in corrispondenza di numeri interi(conteggi).

Assumono solo determinati valori nella scala numerica.

ES: numero decessi giornalieri (4,5,6)

Numero figli (1,2,3,4)

Titolo prova sierologica

Continue

Possono assumere un numero infinito di livelli (le misurazioni)

ES: peso /altezza/ pressione sanguigna/livello di glicemia



Indici statistici

Misure di posizione

Quando ci troviamo di fronte a valori di una variabile che esprimono quantità (discrete o continue), compresi i punteggi, è possibile sintetizzare i dati attraverso misure di posizione e di variabilità

Le distribuzioni di frequenza delle stesse variabili raggruppate possono non essere sufficientemente precise o essere poco sintetiche

FREQUENZA

La frequenza di un valore è il numero di individui della popolazione per i quali la variabile assume tale valore

La frequenza relativa è il rapporto tra la frequenza del valore e il numero di individui della popolazione:

$$\text{freq. relat.} = \text{freq. ass.} / \text{totale individui}$$

La frequenza percentuale si ottiene normalizzando a 100 il totale della popolazione:

$$\text{freq. percentuale} = \text{freq. relativa} * 100$$

Distribuzione

La distribuzione di frequenza descrive come si distribuisce un carattere rispetto alla sua modalità. Le distribuzioni di frequenza possono essere assolute, relative, percentuali e cumulate. Utilizzando le distribuzioni di frequenza è possibile, in genere, rappresentare graficamente i caratteri oggetto di studio. Ogni rappresentazione grafica deve essere appropriata rispetto al tipo di carattere

- Frequenze assolute n_i : numero di volte che osservo la modalità del carattere; totale frequenze N
- Frequenze relative: $f_i = n_i/n$ frequenza assoluta associata alla i -esima modalità diviso il totale delle frequenze osservate
- La frequenza cumulata assoluta corrispondente ad una certa modalità di un carattere, indica il numero di unità della popolazione considerata che presentano un valore del carattere minore o uguale a quella modalità. Analogamente le frequenze cumulate relative (e percentuali) si riferiscono a frazioni del collettivo considerato. Le frequenze cumulate si ottengono “cumulando” progressivamente le frequenze assolute o relative associate a ciascuna modalità del carattere



INDICATORI SINTETICI

Misure di tendenza centrale

Rappresentano i valori attorno a cui i dati tendono ad aggregarsi. Le più diffuse sono:

Moda(caratteri qualitativi non ordinabili)

Mediana (caratteri qualitativi ordinabili)

Media(carattere quantitativi)

La moda o modalità prevalente

La moda (M_o) di un insieme di misure è quella osservazione che si presenta con maggior frequenza nella distribuzione dei dati

E' l'unico indice sintetico per dati qualitativi

Se, però, il numero dei casi preso in considerazione è esiguo, la moda può non essere unica, ovvero ce ne possono essere più di una

Esempio: moda

Razza	n.Capi (freq.ssolute)	Freq. Relative
Bruna	72	10,64
Frisona	443	65,44
Meticcia	70	10,34
Pezzata rossa	92	13,59
Totale	677	100

qualifica funzionale	N.impiegati
II	58,038
III	308,249
IV	287,707
V	71,974
VI	52,232
VII	28,081
VIII	12,259
totale	818,54

Moda:

E' la modalità più frequente nel collettivo di unità considerato

Vantaggi: può essere calcolata per qualunque tipo di distribuzione di frequenza, anche per variabili qualitative non ordinali

Limiti: non è molto utile perché poco indicativa

E' possibile avere distribuzioni multimodali

La mediana

Considerando un carattere le cui modalità sono ordinabili, ad esempio lo stadio di una malattia, possiamo calcolare un altro indice sintetico per la nostra distribuzione di frequenza: la mediana

La mediana è quel valore che in un insieme ordinato lo divide in due parti che contengono lo stesso numero di osservazioni

Corrisponde al valore centrale quando i valori sono ordinati in modo crescente

Corrisponde al valore al di sotto e al di sopra del quale si trovano il 50% delle unità:

Esempio: $n=15$

3 5 5 5 5 7 7 9 9 10 10 12 15 40 41

Essendo $n=15$, ossia n è dispari, la mediana è data dal valore $n+1/2$ quindi

$15+1/2=8$

Quindi la nostra mediana sarà pari a 9 che è il valore che ritroviamo in ottava posizione

Mediana=esempio

Esempio: $n=14$

3 5 5 5 5 5 7 7 9 9 10 10 12 15

La mediana nel caso in cui n è pari è il valore corrispondente alla media aritmetica dei valori della 7ma e dell'ottava unità.

Mediana=

$$14/2=7$$

$$14/2+1=8$$

$$7+8/2=15/2=7$$

La mediana è quel valore della variabile quantitativa ordinabile che, nella successione di valori osservati, disposti in ordine crescente o decrescente, occupa la posizione centrale: ovvero il numero di quelle che possiedono il carattere in quantità superiore alla mediana

Mediana.....

Se n è dispari si ha una sola mediana, ed la modalità corrispondente all'unità $n/2+1$ (nella distribuzione ordinata)

Se n è pari si hanno 2 valori mediani in corrispondenza delle osservazioni $n/2$ ed $(n/2+1)$. Se il carattere è quantitativo allora la mediana è la media aritmetica dei due valori mediani.

Mediana....

- Vantaggi: rispetto alla media aritmetica non risente di valori estremi ed anomali, per questa ragione è uno stimatore più «robusto» della media, ed in caso di distribuzione molto asimmetrica è preferibile alla media
- Limiti: usa un numero limitato di informazioni e non gode di alcune proprietà matematiche e statistiche di grande importanza come la media aritmetica
- Risulta poco adatta per alcuni test statistici

La media aritmetica

- Se disponiamo di una serie di osservazioni su di un carattere quantitativo, cioè abbiamo a disposizione un insieme di valori, uno degli indici sintetici più utilizzati per dare informazioni sulla serie è la media aritmetica. Questa si ottiene sommando i valori osservati e dividendoli per il loro numero.

Esempio:

Ore di lavoro straordinarie (totale 96 ore) effettuate nel mese di giugno 2016 dai 12 dipendenti di una ipotetica impresa

0 5 5 7 7 7 10 10 10 11 12 12


Media= $\text{somma}(0+5+5+7+7+7+10+10+10+11+12+12)/12=96/12=8,45$

La media aritmetica

- E' la misura di tendenza centrale più usata
- Si definisce media aritmetica semplice di n osservazioni il numero che si ottiene dividendo la loro somma per il numero n
- Si ripartisce l'ammontare totale tra tutte le unità ottenendo un valore medio che se fosse assegnato a ciascuna unità darebbe lo stesso totale.
- Vantaggi: sfrutta tutte le informazioni a disposizione
- Limite: risente dei valori estremi e anormali

La media aritmetica

- Se i dati sono rappresentati con una distribuzione di frequenza, cioè la modalità X_j compare con la frequenza F_j ($j=1,2,\dots,k$) si può usare questa formula:
- $\text{Media} = (x_1f_1 + x_2f_2 + \dots + x_nf_n) / (f_1 + f_2 + f_3 + \dots + f_n) = \text{somma}(x_i f_i / n)$
- $[(0 \cdot 1) + (5 \cdot 2) + (7 \cdot 3) + (10 \cdot 3) + (11 \cdot 1) + (12 \cdot 2)] / 12 = 96 / 12 = 8.45$



ore lavoro straordinarie	N impiegati(frequenze)
0	1
5	2
7	3
10	3
11	1
12	2
totale	12

La media aritmetica

- La media aritmetica gode delle seguenti proprietà:
- È sempre compresa tra il minimo ed il massimo degli n valori
- Dati i valori x_1, x_2, \dots, x_n e la loro media aritmetica, si definiscono scarti dalla media le differenze $x_1 - m_x, x_2 - m_x, \dots, x_n - m_x$. La somma di tali differenze è nulla, cioè
- $$\text{somma}(x_i - m_x) = 0$$

Le misure di posizione

- Moda: in una serie di dati, la moda è il valore che presenta la frequenza più elevata
- Mediana: in un insieme ordinato di valori, la mediana è quel valore che divide l'insieme in due parti che contengono lo stesso numero di osservazioni
- Media aritmetica: rapporto tra la somma dei valori che costituiscono una serie di osservazioni ed il numero di osservazioni

Le misure di posizione

- Tra moda, media e mediana, quale scegliere per rappresentare la nostra popolazione?
- 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 valori estremi e/o anomali.
- Media: è calcolabile solo per i caratteri quantitativi, è la più informativa, ma è influenzata dai valori estremi e/o anomali

Le misure di posizione

Tipo di variabile	Misura di posizione
Qualitativo non ordinabile	Moda
qualitativo ordinabile	Moda+Mediana
Quantitativo	Moda+Mediana+Media

La media geometrica

Spesso usata per variabili quantitative come alternativa alla mediana quando le osservazioni sono distribuite in forma asimmetrica a destra

Dato un certo numero di osservazioni X_1, X_2, \dots, X_n la media geometrica è:

$$X_g = \text{radq}(n) \text{ Produttoria } X_i^n$$

La media geometrica di una distribuzione è uguale all'antilogaritmo della media aritmetica dei logaritmi.

Es: data una serie di 7 valori

5 5 5 7 10 20 102 la media geometrica è:

$$\text{Radq}(7) [5 \cdot 5 \cdot 5 \cdot 7 \cdot 10 \cdot 20 \cdot 102] = 10,86$$

La media geometrica

Alternativamentesi calcola la media aritmetica dei logaritimi

5 5 5 7 10 20 102

$\text{Log}(5)+\text{Log}(5)+\text{Log}(5)+\text{Log}(7)+\text{Log}(10)+\text{Log}(20)+\text{Log}(102)=$

$1,609+1,609+1,609+1,945+2,302+2,995+4,624=2,38$

Media geometrica= $\exp(2,38) = 10,86$



Ricapitolando

Valori caratteristici di posizione:

Valori calcolati a partire dalle osservazioni fatte, per individuare un valore tipico di una regione della distribuzione (es. la mediana)

Valori di tendenza centrale:

Valori calcolati a partire dalle osservazioni fatte, per esprimere sinteticamente il valore centrale attorno al quale sono disposte le osservazioni (es. la media)

Misure di dispersione

Gli indici di posizione dicono attorno a quale valore le osservazioni sono centrate e sono tanto più significative quanto più i dati sono concentrati vicino ad essi. E' interessante notare misurare il grado di dispersione dei dati intorno a tali indici.

Si possono calcolare solo per i caratteri quantitativi, associando alle misure di tendenza centrale delle misure di dispersione, quali

Campo di variazione (range)

Varianza

Deviazione standard

Coefficiente di variazione



Misurare la variabilità di una distribuzione

Le misure di variabilità forniscono una informazione sul grado di dispersione delle modalità di una distribuzione:

Fra due termini o due valori rappresentativi della distribuzione

Fra tutti i termini della distribuzione

Rispetto ad un valore centrale (es. media aritmetica)

Campo di variazione (o “range”)

Il campo di variazione di una distribuzione è la differenza tra il dato più grande e quello più piccolo della distribuzione:

$$C = x_{\max} - x_{\min}$$

Questo indice è abbastanza grossolano non dicendo nulla sulla variabilità dei dati intermedi.

Un limite di tale indice è che il suo valore dipende soltanto dai due valori estremi della serie ordinata dei numeri, e non tiene conto della variabilità interna degli altri valori

Esempio:

il campo di variazione della seguente distribuzione:

25 – 26 – 28 – 29 – 30 – 32 è

$$C = 32 - 25 = 7$$

Scarto

Lo scarto quadratico medio **campionario** misura quanto ciascun dato x_i si discosta dal valor medio, ovvero

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + \dots + (x_n - \mu)^2}{n - 1}}$$

Varianza della popolazione: misura che caratterizza molto bene la variabilità di una popolazione

Altro non è che il quadrato dello scarto quadratico medio

$$\sigma^2 = \frac{(x_1 - \mu)^2 + \dots + (x_n - \mu)^2}{n - 1}$$

Varianza

Vantaggi:

Usa le informazioni provenienti da tutte le osservazioni campionarie. E' facile da calcolare ed ha buone proprietà matematiche

Svantaggi

E' espressa in una unità di misura pari al quadrato dell'unità di misura delle osservazioni

Deviazione Standard

Misura lo scostamento medio delle osservazioni dalla media della distribuzione.....

Altro non è che la radice quadrata della varianza. Utilizzando la deviazione standard ci si riconduce ad un indice di variabilità omogeneo, cioè espresso nella stessa unità di misura della variabile considerata. Maggiore è la variabilità maggiore è la deviazione standard, che assume valore nullo solo nel caso in cui tutti i valori siano uguali

Associazione tra due variabili

$$\text{Concordanza} = (35 + 56) / 100$$

	malato	sano	
malato	35	7	42
sano	2	56	58
	37	63	100

$$\text{Concordanza} = (A + D) / N$$

	positivo	negativo	
positivo	A	B	A+B
negativo	C	D	C+D
	A+C	B+D	N

I test diagnostici

Sensibilità

E' la capacità do in test di identificare i casi malati

La proporzione $A/A+C$

veri positivi/malati è una misura della probabilità di identificare correttamente come positivi i soggetti malati

$$\underline{Se=A/A+C}$$

		verità		
		pos	neg	
test	pos	a(VP)	b(FP)	a+b
	neg	c(FN)	d(VN)	c+d
		a+c	b+d	a+b+c+d

Specificità

$$Sp = D / B + D$$

$$Sp = 80 / 100$$

E' la capacità di un test di individuare i non malati nella popolazione controllata

La proporzione $D/B+D$

Veri negativi/sani è una misura della probabilità di identificare correttamente come negativi i soggetti sani

	positivo	negativo	
positivo	A	B	A+B
negativo	C	D	C+D
	A+C	B+D	N



	malato	sano	
test+	89	20	109
test-	11	80	91
	100	100	200

Numero figli				
0	2	1	4	3
1	2	3	8	2
5	2	1	3	3
1	3	2	2	5
4	4	4	2	3
5	5	1	1	2
4	4	2	3	3
3	3	3	3	2

Esempio frequenza

N=40

Numero figli di coppie che abitano in un quartiere

Classe	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
0	1	0.025	2.5%
1	6	0.15	15.0%
2	10	0.25	25.0%
3	12	0.3	30.0%
4	6	0.15	15.0%
5	4	0.1	10.0%
6	0	0.	0.0%
7	0	0.	0.0%
8	1	0.025	2.5%
Totali	40	1.	100.