

MATHESIS

Sezioni di Napoli e C.mare di Stabia

Scuola Estiva di Formazione - Secondo Ciclo di Istruzione

La Matematica dell'incerto

Ferdinando Casolaro

ferdinando.casolaro@unina.it

Insegnamento della Probabilità e della Statistica

Quale percorso nella scuola secondaria di secondo grado?

Castellammare di Stabia, 15-18 luglio 2018

Motivazioni

Resistenza a considerare il Calcolo delle Probabilità e la Statistica come pilastri centrali su cui impostare la programmazione dell'insegnamento della Matematica.

Mancanza di Formazione dei docenti nei pregressi corsi universitari e nei loro passati percorsi di studenti.

L'insegnamento dell'Aritmetica, dell'Algebra, della Geometria, ecc.... viene proposta sull'assiomatica elementare già dalla Scuola del Primo Ciclo.

La *Matematica dell'incerto* viene catapultata direttamente agli esami di Stato o attraverso le prove Invalsi.

Programma da distribuire nel corso di 5 anni, come avviene per l'aritmetica, l'algebra e le geometrie.

1. L'aritmetica dell'incerto

Le medie: media aritmetica, media aritmetica ponderata; media geometrica; media armonica.

Medie di posizione: mediana; moda.

Indici di variabilità: scarto semplice, varianza, scarto quadratico medio.

Elementi di analisi combinatoria: Principi fondamentali del calcolo combinatorio. Fattoriale di un numero intero positivo, coefficiente binomiale, binomio di Newton.

Disposizioni, Permutazioni, Combinazioni semplici e con ripetizioni.

Introduzione alla Probabilità

Cenni storici; *la definizione classica; la definizione frequentista; la definizione soggettiva; la definizione assiomatica.*

Eventi incompatibili; eventi indipendenti. La probabilità condizionata; Principio di probabilità totale e teorema di Bayes.

Le variabili casuali

Definizione di variabile casuale. Distribuzione di probabilità. Variabili casuali discrete e continue: il valore medio, la varianza, la deviazione standard.

Alcuni esempi di distribuzioni

Distribuzione binomiale; distribuzione normale; distribuzione di Poisson.

La probabilità

Calcolo delle probabilità: i vari aspetti della probabilità e le questioni riguardanti gli insiemi discreti (finiti e numerabili).

1) *Definizione classica di probabilità.*

E' il rapporto tra il numero n_f dei casi favorevoli ed il numero n dei casi possibili:

$$p(a) = \frac{n_f}{n}$$

Tale definizione è valida solo se i casi sono equiprobabili.

2) *Definizione frequentista di probabilità*

E' una definizione che nasce dall'esperienza, cioè dall'osservazione di una ripetizione di prove e dal numero di volte in cui si verifica l'evento richiesto.

La probabilità frequentista è, dunque, il rapporto tra la frequenza f con cui si è verificato l'evento richiesto in n osservazioni precedenti ed il numero n stesso.

$$p(a) = \frac{f}{n}$$

quando il numero delle osservazioni è “abbastanza grande”.

L'espressione “*abbastanza grande*” ha un significato relativo allo specifico evento che si sta analizzando.

Guido Castelnuovo definisce la probabilità *frequentista* mediante la seguente affermazione (legge empirica del caso):

“In una serie di prove ripetute un gran numero di volte, nelle stesse condizioni, ciascuno degli eventi possibili si manifesta con una frequenza relativa (probabilità frequentista) che è presso a poco uguale alla sua probabilità, l'approssimazione cresce al crescere del numero delle prove”.

L'affermazione di Castelnuovo esprime la cosiddetta *legge dei grandi numeri*:

Con il crescere del numero delle prove, è sempre più probabile che la frequenza relativa di un evento si avvicini alla sua probabilità.

3) Definizione soggettiva di probabilità.

Bruno de Finetti (1906-1985), alla domanda: “che cos’è la probabilità?” era solito rispondere: “la probabilità non esiste!”.
Quale significato attribuire a tale risposta? Il significato si evince dal concetto di *probabilità soggettiva*.

La probabilità soggettiva p di un evento E è la misura del grado di fiducia espresso dal numero reale p , tale che una scommessa di quota p su E sia coerente, cioè tenga conto delle condizioni reali.

*La probabilità soggettiva è utilizzata nel caso in cui non abbia senso considerare ciò che è avvenuto per una successione di eventi analoghi o si deve assegnare una probabilità anche agli eventi in cui i casi possibili sono *infiniti*.*

Dato un numero reale p ($0 < p < 1$) ed una somma di danaro Q , diciamo che si effettua una scommessa di quota p su un evento E se, versando la somma pQ si riceve l’importo Q solo se si verifica l’evento E .

4) Definizione assiomatica di probabilità.

E' una definizione che si basa su un'assiomatica che presenta analogie alla struttura della geometria euclidea ed alla costruzione della teoria della misura.

Precisamente, si fissano degli assiomi su cui viene costruita una serie di operazioni che permettono l'analisi della previsione di eventi.

Ad ogni evento A dello spazio campione Ω ($A \in \Omega$) associamo un numero reale p , tale che:

$$0 \leq p(A) \leq 1$$

cioè, la probabilità è una funzione che ad ogni elemento dello spazio campione Ω associa un numero reale compreso tra 0 e 1.

1. Se S è l'evento certo, si ha $p(S) = 1$.
2. Se $\emptyset$ è l'evento impossibile, si ha $p(\emptyset) = 0$.
3. Se $A_1, A_2, \dots, A_n$ sono n eventi che si escludono a vicenda, cioè se:

$$A_i \cap A_j = \emptyset, \quad \forall i \neq j$$

si ha:

$$p\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n p(A_i)$$

Estensione della probabilità assiomatica ad un'infinità numerabile

La definizione assiomatica si estende ad insiemi infiniti numerabili.

In una σ -algebra (cioè per un insieme di infinità numerabile), si ha:

$$p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i)$$

La concezione assiomatica della probabilità permette di concepire l'insieme di tutti i possibili esiti che si possono verificare come uno spazio (spazio di probabilità o spazio dei campioni).

Un esito (o un evento) è detto punto campione.

Eventi incompatibili (esempio pag. 4)

Due eventi si dicono *incompatibili* se si escludono a vicenda. In tal caso, se E_1 ed E_2 sono due eventi incompatibili, risulta:

$$E_1 \cap E_2 = \emptyset$$

Eventi indipendenti (esempio pag. 5)

Due eventi si dicono *indipendenti* quando il verificarsi del primo non altera la probabilità del verificarsi dell'altro.

Evento totale - Evento composto (esempio pag. 6)

Dati due o più eventi (parziali):

- si dice *evento totale* (*unione degli eventi*) di essi, l'evento che consiste nel verificarsi dell'uno o dell'altro dei vari eventi parziali.
- si dice *evento composto* (*intersezione degli eventi*) da essi, l'evento che consiste nel verificarsi di tutti gli eventi parziali.

Principio della probabilità totale (esempio pag. 6)

Dati due eventi parziali, la probabilità del loro evento totale è uguale alla somma delle probabilità dei due eventi parziali diminuita della probabilità del loro evento composto. Nel linguaggio degli insiemi si ha:

$$p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2)$$

Nel caso che gli eventi siano incompatibili, risulta: $E_1 \cap E_2 = \emptyset$, per cui si ha:

$$p(E_1 \cup E_2) = p(E_1) + p(E_2)$$

che è il principio della probabilità totale per eventi incompatibili.

Principio della probabilità composta (esempio pag. 6)

Se un evento è composto di due o più eventi indipendenti, la sua probabilità è il prodotto delle probabilità dei vari eventi componenti.

Nel linguaggio degli insiemi si ha:

$$p(E_1 \cap E_2) = p(E_1) \cdot p(E_2)$$

Probabilità condizionata

La valutazione della probabilità di un evento dipende anche dallo stato di informazione di cui si è in possesso.

Didatticamente, riteniamo che sia opportuno introdurre l'argomento attraverso esempi (*insegnamento per problemi*).

I due esempi che seguono rappresentano due situazioni diverse:

1. *Calcolare la probabilità che il getto di due dadi dia 10 (evento A).*
2. *Calcolare la probabilità che il getto di due dadi dia 10, sapendo che un dado è uscito 6 (evento A/B).*

Nel primo caso si deve considerare l'intero universo U degli eventi elementari derivanti dal getto dei due dadi; nel secondo si deve considerare solamente un suo sottoinsieme, e precisamente quello i cui elementi sono coppie di numeri di cui almeno uno sia 6.

Nel primo caso, lo **Spazio Campione** U è costituito dai seguenti 36 elementi:

$$U \equiv \{(1,1); (1,2); (1,3); (1,4); (1,5); (1,6); (2,1); (2,2); (2,3); (2,4); (2,5); (2,6); (3,1); (3,2); (3,3); (3,4); (3,5); (3,6); (4,1); (4,2); (4,3); (4,4); (4,5); (4,6); (5,1); (5,2); (5,3); (5,4); (5,5); (5,6); (6,1); (6,2); (6,3); (6,4); (6,5); (6,6)\}$$

Nel secondo caso si deve considerare solamente un suo sottoinsieme, che indichiamo con U_6 . Precisamente lo **Spazio Campione** è costituito dalle coppie di numeri di cui almeno uno sia 6:

$$U_6 \equiv \{(1,6); (2,6); (3,6); (4,6); (5,6); (6,6); (6,1); (6,2); (6,3); (6,4); (6,5)\}.$$

Nel primo caso, il sottoinsieme U_3 dei punti di U costituiti da coppie che danno per somma 10 è:

$$U_3 \equiv \{(4,6); (6,4); (5,5)\}$$

e, quindi, essendo gli eventi equiprobabili, ci basta applicare la definizione classica; la probabilità

$$p(A) = \frac{3}{36} = \frac{1}{12}$$

Nel secondo caso, l'insieme U_6 è costituito da 11 elementi, quindi il sottoinsieme dei punti di U_6 costituiti da coppie che danno per somma 10 è:

$$U_2 \equiv \{(4,6); (6,4)\}$$

per cui risulta che la probabilità $p(B)$ è $\frac{2}{11}$.

Questo secondo esempio ci introduce il concetto di probabilità condizionata.

Probabilità condizionata

Calcolare la probabilità che si verifichi l'evento $A \equiv \{\text{il getto di due dadi dia } 10\}$, a condizione che sia dato l'evento $B \equiv \{\text{un dado ha dato } 6\}$, che è individuato dai punti di U_6 .

In tal caso diciamo che si calcola la probabilità che si verifichi l'evento A condizionato a B che si esprime con $p(A/B)$.

Osserviamo che l'insieme $U_2 \equiv \{(4,6); (6,4)\}$ è costituito dai punti comuni sia ad A che a B , cioè ad $A \cap B$ la cui probabilità nello spazio campione U , è data da:

$$p(A \cap B) = \frac{2}{36} = \frac{1}{18}$$

mentre l'insieme U_6 , costituito dai punti che hanno una coordinata uguale a 6, è costituito da 11 elementi, per cui la probabilità che si verifichi l'evento B è:

$$p(B) = \frac{11}{36}$$

Allora risulta:

$$p(A/B) = \frac{p(A \cap B)}{p(B)} = \frac{\frac{2}{36}}{\frac{11}{36}} = \frac{2}{11}$$

Generalizzando quanto detto, si hanno dunque le due relazioni:

$$p(A/B) = \frac{p(A \cap B)}{p(B)}$$

$$p(A \cap B) = p(A/B) \cdot p(B)$$

che individuano ***le leggi della probabilità condizionata.***

N.B. Il risultato precedentemente ottenuto:

$$p(A/B) = \frac{2}{11}$$

si può considerare anche di tipo classico, perché la probabilità del verificarsi di una qualsiasi coppia è $\frac{1}{36}$, cioè gli eventi sono equiprobabili:

- i casi favorevoli sono 2 su 36: $\{(4,6); (6,4)\}$;

- i casi possibili (*un dado abbia dato 6*) sono 11 su 36:

$\{(1,6); (2,6); (3,6); (4,6); (5,6); (6,6); (6,1); (6,2); (6,3); (6,4); (6,5)\}$

per cui risulta:

$$p(A/B) = \frac{n_f}{n} = \frac{2}{11}$$

Teorema di Bayes (esempio pag. 11)

Nelle applicazioni (in particolare in ambito ambientale e geologico), due qualsiasi eventi H_1 e H_2 , si possono immaginare come *cause possibili per un evento A osservato*.

Il teorema di Bayes permette di calcolare la probabilità affinché si possa verificare l'evento H_1 (l'evento H_2) una volta che si è osservato l'evento A già verificato.

Con semplici passaggi algebrici, dalla relazione di *probabilità condizionata*, si deduce la formula di Bayes:

$$p(H_1 / A) = \frac{p(H_1) \cdot p(A / H_1)}{p(H_1) \cdot p(A / H_1) + p(H_2) \cdot p(A / H_2)}$$

Tale relazione va estesa ad un numero n di eventi (cause per l'evento A) :

$$p(H_1 / A) = \frac{p(H_1) \cdot p(A / H_1)}{p(H_1) \cdot p(A / H_1) + p(H_2) \cdot p(A / H_2) + \dots + p(H_n) \cdot p(A / H_n)}$$

Esempio

Il montaggio di un'apparecchiatura è effettuato con componenti di buona qualità o con componenti di qualità mediocre.

Nel primo caso, la probabilità di funzionamento corretto per una durata di tempo T è di 0,95, mentre nel secondo caso è di 0,70.

Il 40% delle apparecchiature contiene componenti di buona qualità.

Supponendo che al collaudo un'apparecchiatura funzioni correttamente per la durata T , si calcoli la probabilità che essa sia costituita da componenti di buona qualità (si applichi il teorema di Bayes).

Evento A : “l'apparecchiatura funziona al collaudo per un tempo T ”.

Evento H_1 : “l'apparecchiatura è montata con elementi di buona qualità”.

Evento H_2 : “l'apparecchiatura è montata con elementi di mediocre qualità”.

La probabilità $p(H_1) = 0.40$;

La probabilità $p(H_2) = 0.60$;

La probabilità $p(A/H_1) = 0.95$;

La probabilità $p(A/H_2) = 0.70$;

Pertanto, risulta:

$$p(H_1/A) = \frac{p(A/H_1) \cdot p(H_1)}{p(A/H_1) \cdot p(H_1) + p(A/H_2) \cdot p(H_2)} =$$

$$= \frac{0,95 \cdot 0,40}{0,95 \cdot 0,40 + 0,70 \cdot 0,60} = 0,475$$

Distribuzione di probabilità - Variabili casuali discrete e continue

Si chiama variabile casuale una funzione che associa ad ogni evento elementare E dello Spazio Campionario S uno ed un solo numero reale.

Una *variabile casuale* si dice *discreta* quando può assumere solo particolari valori in punti isolati di un intervallo.

Una *variabile casuale* si dice *continua* se i suoi valori possono variare con continuità in un intervallo, che può essere limitato o illimitato.

Problema 1. - *Consideriamo lo spazio degli eventi (spazio campionario) associato al lancio di tre monete. Qual è il numero totale delle volte in cui si presenta croce?*

Osserviamo che, lanciando contemporaneamente le tre monete, croce si può presentare:

1. zero volte (evento E_0 - tre volte testa). 2. una volta (evento E_1 - due volte testa). 3. due volte (evento E_2 - una volta testa). 4. tre volte (evento E_3 - zero volte testa).

E' opportuno far rappresentare agli allievi la seguente tabella, da cui si evince immediatamente la variabile indipendente che indichiamo con X:

Punti dello spazio campionario	Numero di volte che esce croce:
CCC	3
CCT	2
CTC	2
CTT	1
TCC	2
TCT	1
TTC	1
TTT	0

Attribuendo ad ogni valore della variabile casuale X la corrispondente probabilità, si ottiene la *distribuzione di probabilità di X*.

Nel nostro caso lo spazio campionario S è costituito da otto elementi:

$$S \equiv \{CCC, CCT, CTC, CTT, TCC, TCT, TTC, TTT\}.$$

Poiché la variabile X può assumere: {una sola volta il valore zero, tre volte il valore uno, tre volte il valore due, una volta il valore tre}, l'insieme

$$\{X_0=0, X_1=1, X_2=2, X_3=3\}$$

individua il dominio **D** della nostra funzione ai cui valori sono associate, rispettivamente, le seguenti probabilità:

$$\mathbf{D} \equiv \{p_0=1/8, p_1=3/8, p_2=3/8, p_3=1/8\}.$$

La legge che associa ad ogni valore della variabile casuale X_k , (con $k = 0 \dots 3$), il corrispondente valore della probabilità p_k è detta ***distribuzione di probabilità della variabile casuale discreta X*** .

Problema 2. - *Determinare la distribuzione di probabilità della variabile X che esprime il punteggio ottenuto lanciando due dadi*

I valori di X ai quali corrispondono le probabilità $p(X)$ sono:

X:	2	3	4	5	6	7	8	9	10	11	12
$p(X)$:	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36.

Pertanto, l'insieme

$\{1/36 \ 2/36 \ 3/36 \ 4/36 \ 5/36 \ 6/36 \ 5/36 \ 4/36 \ 3/36 \ 2/36 \ 1/36\}$

rappresenta la ***distribuzione di probabilità*** della variabile casuale X.

E' opportuno far rappresentare agli allievi le coppie $[X, p(X)]$ sul piano cartesiano perché uno degli aspetti più significativi dei problemi proposti è l'osservazione che, con i valori di X sulle ascisse e i valori di $p(X)$ sulle ordinate, ***i punti ad ordinata massima si hanno nella parte centrale della rappresentazione e i punti ad ordinata minima agli estremi della rappresentazione***, come gli stessi allievi avranno modo di constatare negli anni successivi con la rappresentazione della curva di Gauss.

Spazio di Probabilità

Indicato con x il numero reale che rappresenta il valore della variabile casuale X , la terna $\{S, D, p(x)\}$ è detta *Spazio di Probabilità*.

Pertanto, uno *Spazio di probabilità* è una terna $\{S, D, p(x)\}$, dove S è un insieme qualunque (in genere pensato come l'insieme dei risultati possibili di un esperimento casuale), D è detta σ -algebra, ovvero un insieme (gli eventi) per i quali si può calcolare una probabilità, e $p(x)$ è una misura di probabilità su S ; precisamente:

$$p(x) : S \rightarrow [0, 1].$$

Distribuzione di variabili casuali continue

Se i valori di una *variabile casuale* variano con continuità in un intervallo, essa si dice *continua*.

Distribuzione di Poisson

Consideriamo il caso delle cosiddette “*n prove ripetute*”: ad esempio, si debba calcolare la probabilità p che un evento si verifichi un numero k di volte su n prove ripetute, con n relativamente piccolo. In generale, in questi casi la probabilità p assume valori abbastanza piccoli e si può determinare solo con un grande numero di prove.

Si utilizzano allora delle formule approssimate di cui la più significativa è la distribuzione di Poisson che è espressa dalla relazione:

$$p_k = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

dove λ è il valore medio della variabile casuale binomiale x e k è il numero di volte in cui si verifica l'evento richiesto.

Esempio.

Ad un centralino di una ditta pervengono mediamente 10 telefonate al minuto, la probabilità che in un minuto pervengano 6 telefonate è data da:

$$p_6 = \frac{10^6}{6!} \cdot e^{-10} \cong 6.3\%$$

Costruzione di semplici grafici relativi a problemi naturali

- 1) Con un campione di mille studenti, costruire un istogramma che presenta
- sull'asse delle ascisse le seguenti fasce di altezza (in centimetri):
<150 150-160 160-170 170-180 180-190 190-200 >200;
 - sull'asse delle ordinate, il numero di studenti per ogni fascia.

Si vede immediatamente che la fascia centrale presenta i valori massimi, mentre gli studenti di altezza inferiore a 150 cm e gli studenti di altezza superiore a 200 centimetri presentano i valori più bassi.

- 2) Costruire l'istogramma che presenta sull'asse delle ascisse i mesi dell'anno e sulle ordinate la misura della fascia oraria compresa tra l'alba e il tramonto.

Orientativamente, si hanno i seguenti dati:

gennaio e dicembre: 8.00 → 17.00 - giugno-luglio: 5.00 → 21.00.

*I valori massimi sono concentrati nei mesi centrali, **giugno-luglio**, i valori minimi agli estremi rappresentati dai mesi di **gennaio e dicembre**.*

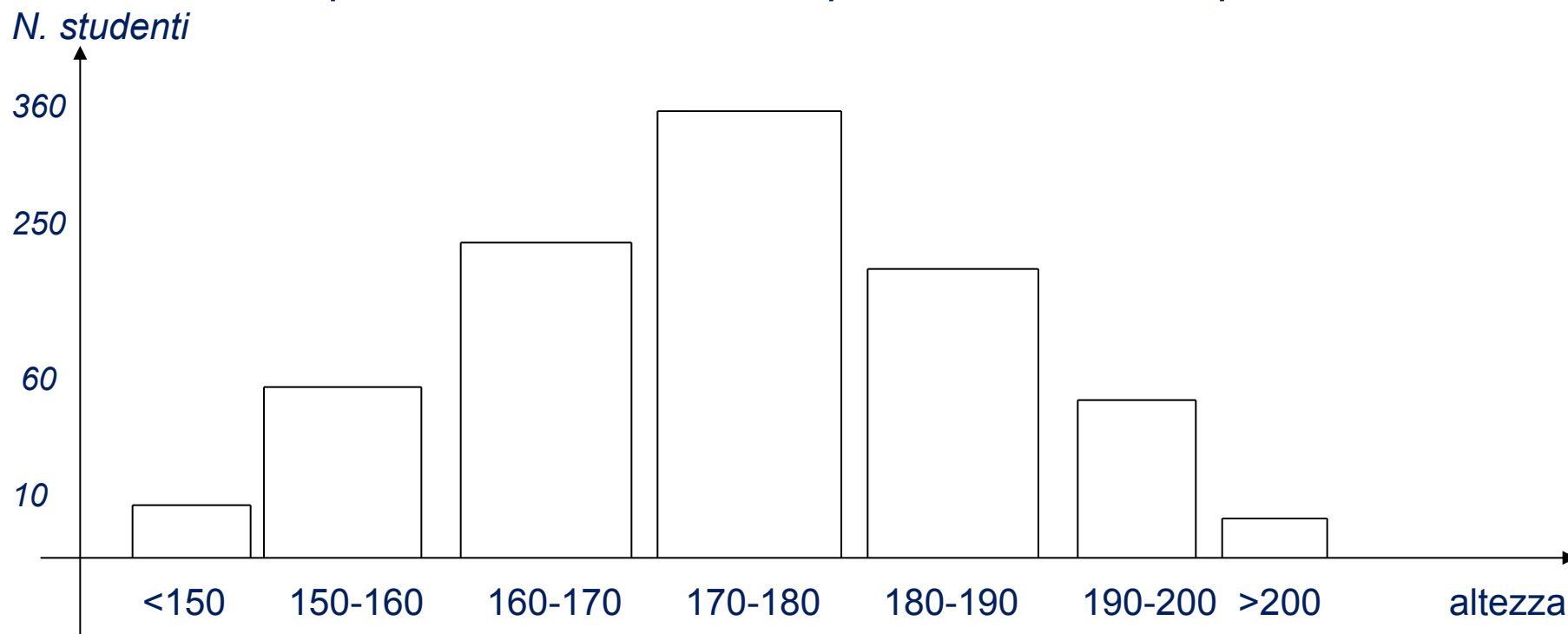
- 3) Disegnare un vulcano ed indicare sulla figura i punti in cui si sviluppa la massima energia.

*E' evidente che la figura del vulcano ha l'andamento della curva di Gauss; è interessante far notare agli allievi come i valori di massima energia si hanno nella parte centrale in cui è concentrata una massa maggiore, coerentemente al principio di equivalenza **massa-energia**.*

Distribuzione degli studenti in base all'altezza (campione di 1000 studenti di età superiore ai 18 anni)

- 1) Con un campione di mille studenti, costruire un istogramma che presenta
- sull'asse delle ascisse le seguenti fasce di altezza (in centimetri):
<150 150-160 160-170 170-180 180-190 190-200 >200;
 - sull'asse delle ordinate, il numero di studenti per ogni fascia.

Si vede immediatamente che la fascia centrale presenta i valori massimi, mentre gli studenti di altezza inferiore a 150 cm e gli studenti di altezza superiore a 200 centimetri presentano i valori più bassi.



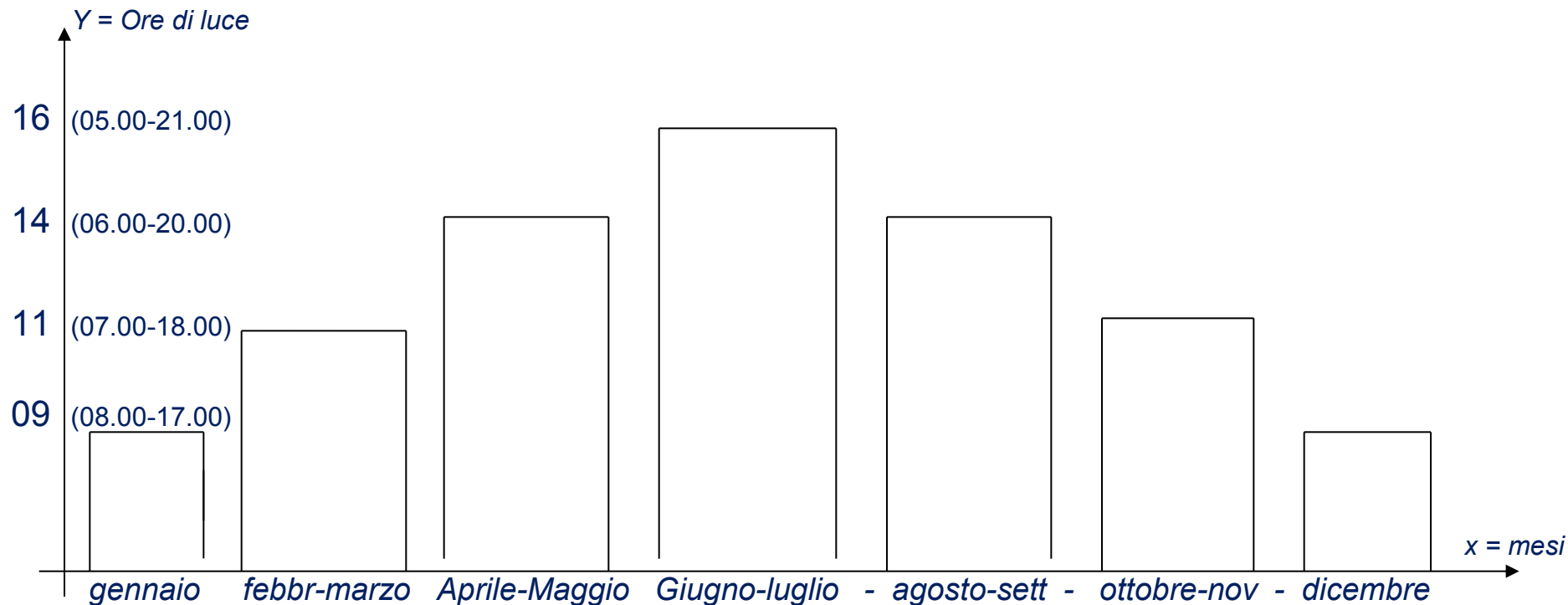
Distribuzione delle ore di luce durante la giornata

Costruire l'istogramma che presenta sull'asse delle ascisse i mesi dell'anno e sulle ordinate la misura della fascia oraria compresa tra l'alba e il tramonto.

Orientativamente, si hanno i seguenti dati:

gennaio e dicembre: 8.00 → 17.00 - giugno-luglio: 5.00 → 21.00.

*I valori massimi sono concentrati nei mesi centrali, **giugno-luglio**, i valori minimi agli estremi rappresentati dai mesi di **gennaio e dicembre**.*



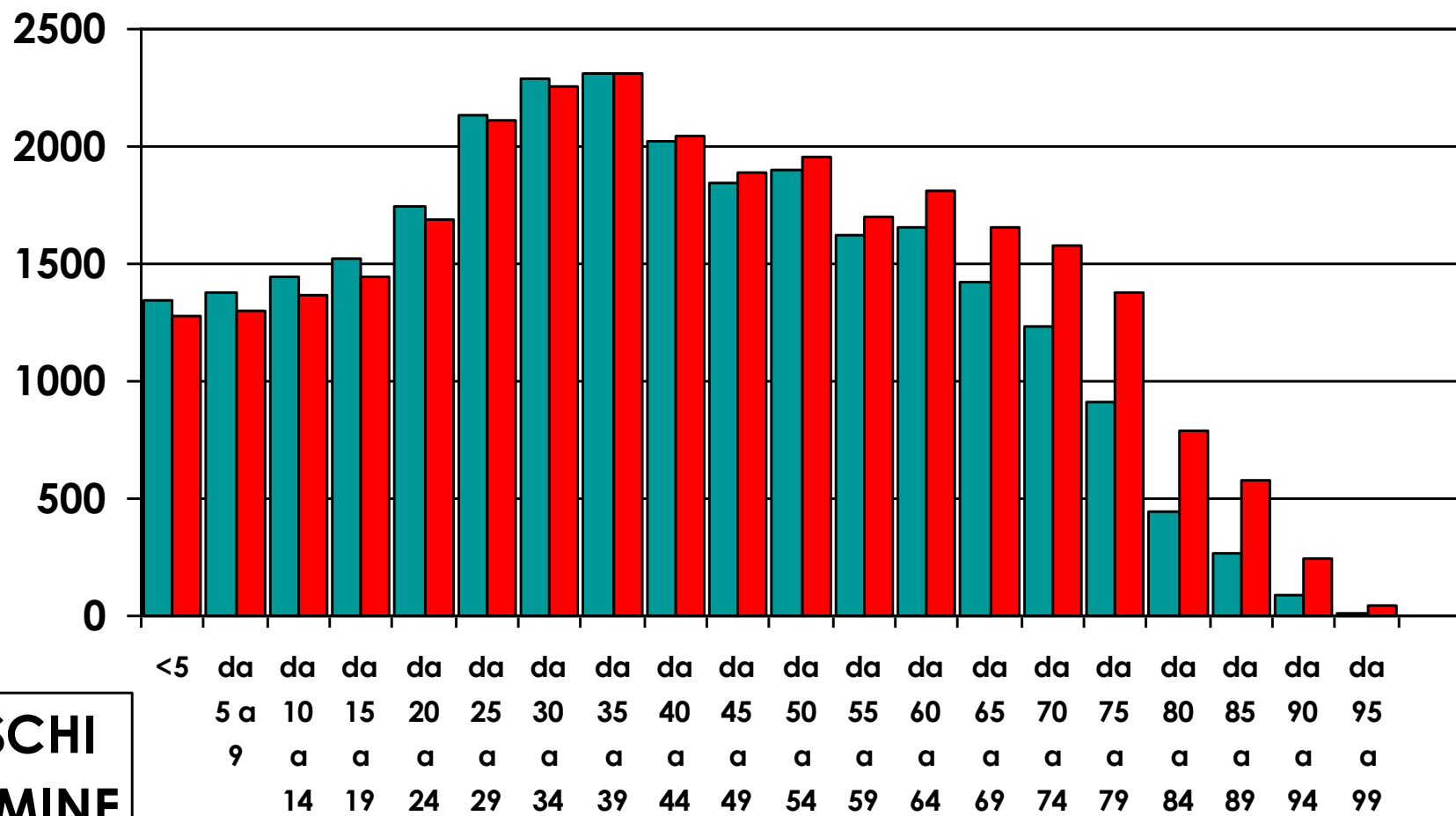
Costruzione di semplici grafici relativi a problemi naturali

3) Disegnare un vulcano ed indicare sulla figura i punti in cui si sviluppa la massima energia.

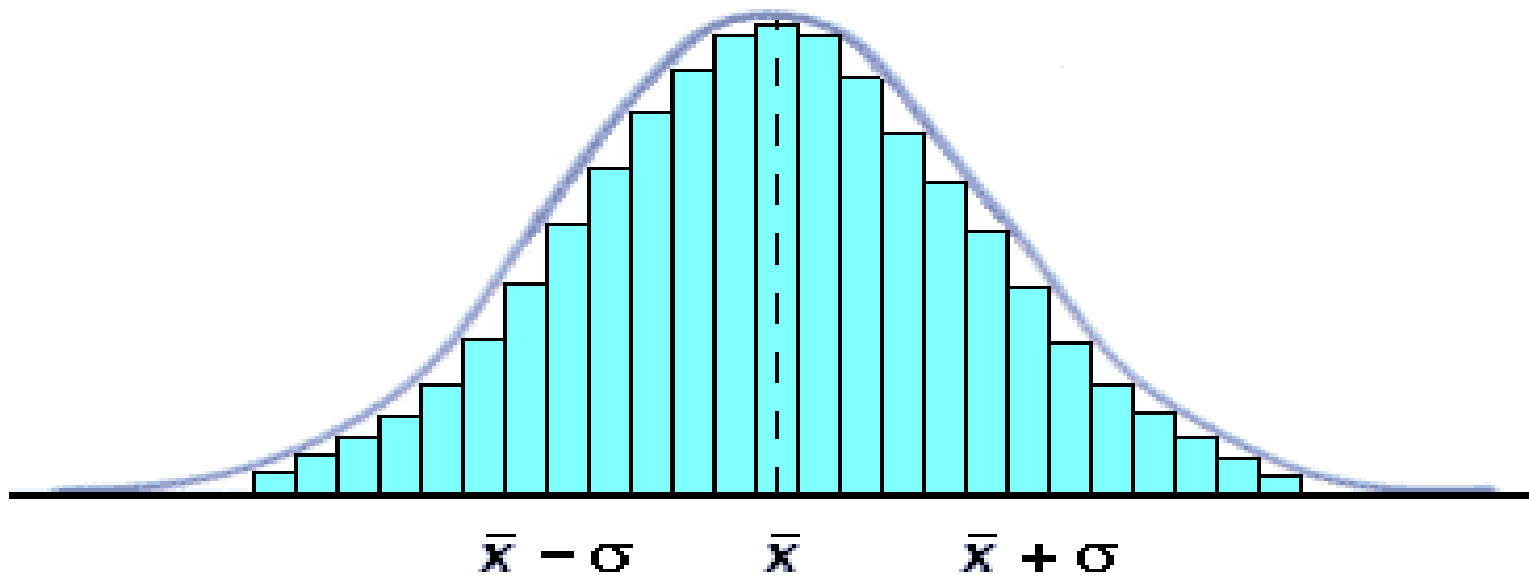
*E' evidente che la figura del vulcano ha l'andamento della curva di Gauss; è interessante far notare agli studenti come i valori di massima energia si hanno nella parte centrale in cui è concentrata una massa maggiore, coerentemente al principio di equivalenza **massa-energia**.*



Distribuzione popolazione residente in Italia per sesso e classi di età - Censimento 2001 (campione di 50000)



* tratto da *“Matematica per la scuola superiore, vol.1”* – A. Giambò,



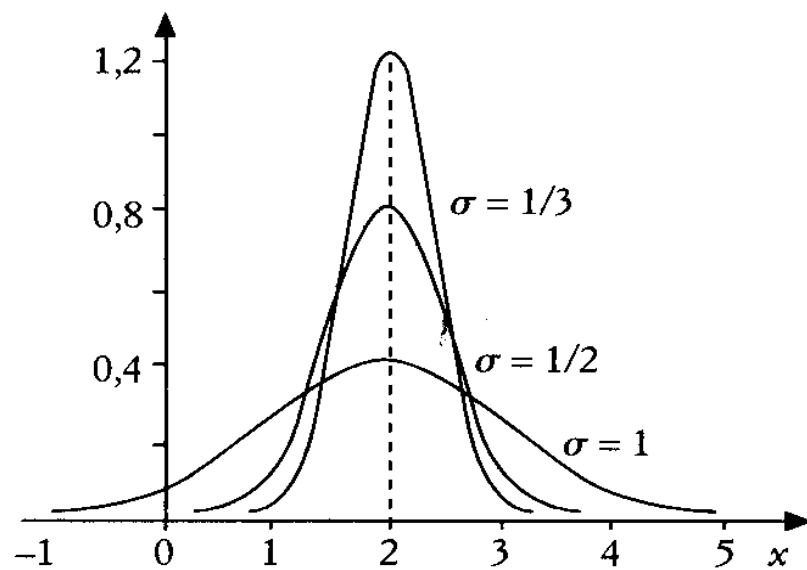
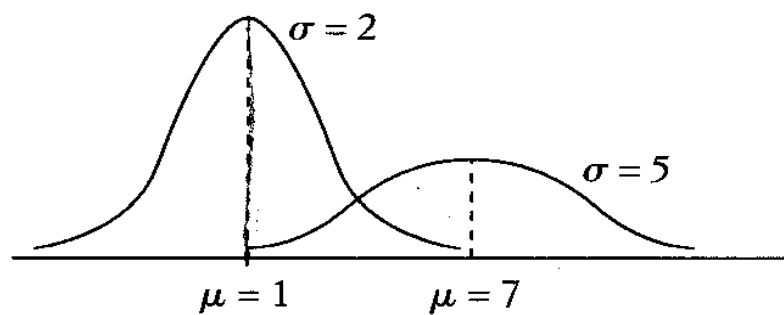
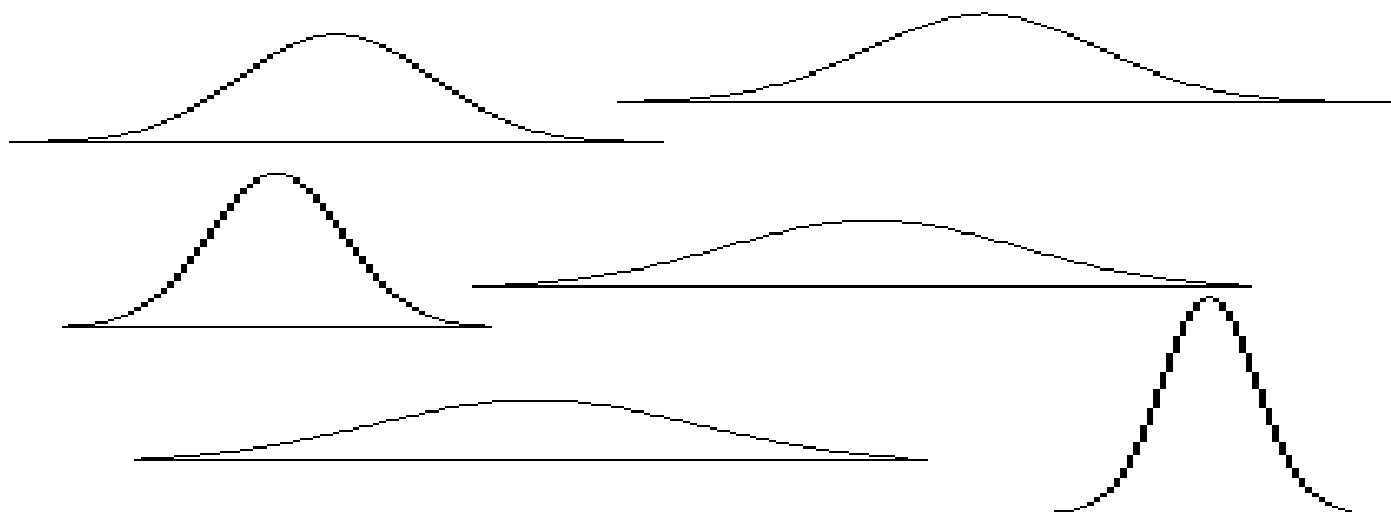
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Distribuzione di variabili casuali continue

I parametri essenziali nella rappresentazione analitica di una variabile casuale continua sono:

1. La ***media aritmetica*** μ (che nella curva normale, essendo simmetrica, coincide con la *moda* e la *mediana*) *corrisponde all'asse di simmetria della curva e definisce la posizione, sull'asse delle ascisse, della curva*. Cambiando la media della curva questa trasla lungo l'asse x .
2. La ***deviazione standard*** σ (o scarto quadratico medio) *corrisponde alla distanza tra la media e il punto di flesso della curva (dove la curva attraversa la sua tangente) e determina l'ampiezza della curva stessa*.
3. La ***varianza*** σ^2 fornisce una misura di quanto i valori assunti dalla variabile si discostino dalla *media*, per cui la ***varianza*** è un indice di variabilità.

Data una ***distribuzione*** di una variabile su una popolazione di n elementi, la ***varianza*** è la *media aritmetica del quadrato delle distanze dei valori dalla loro media*.



INFERENZA STATISTICA

L'inferenza statistica tratta lo studio delle ipotesi formulate su un'intera popolazione, attraverso l'analisi delle caratteristiche esaminate su un campione di n elementi estratti dagli N ($n \ll N$) componenti dell'intera popolazione.

L'ipotesi che si sottopone a verifica è detta *ipotesi nulla* e si indica con H_0 .

Precisamente, “*l'ipotesi nulla*” è quella che non modifica lo stato attuale dell'informazione; essa afferma che è “*poco significativa*” (nulla) la differenza tra due o più parametri statistici.

L'ipotesi che conduce a risultati diversi tra due parametri statistici è detta *ipotesi alternativa* e si indica con H_1 .

La verifica delle ipotesi avviene attraverso un *test statistico*.

VERIFICA DELLE IPOTESI SULLA BASE DI DUE O PIU' CAMPIONI

(caso qualitativo): il test χ (chi quadrato)

I risultati ottenuti nei campioni non sempre concordano esattamente con i risultati teorici attesi secondo le regole della probabilità.

Supponiamo che in un particolare campione si sia osservato che un insieme di possibili eventi: $E_1, E_2, \dots, E_n$, si presenti con frequenze rispettivamente, $O_1, O_2, \dots, O_n$, dette *frequenze osservate*, e che, secondo le regole della probabilità, ci si attende che si presenti con frequenze $e_1, e_2, \dots, e_n$, dette *frequenze attese*.

Evento	$E_1, E_2, \dots, E_n$
Frequenze osservate	$O_1, O_2, \dots, O_n$
Frequenze attese	$e_1, e_2, \dots, e_n$

Vogliamo sapere se le frequenze osservate e le frequenze attese differiscano significativamente.

La statistica χ^2 (chi quadrato), espressa dalla relazione:

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \dots + \frac{(o_n - e_n)^2}{e_n} = \sum_{k=1}^n \frac{(o_k - e_k)^2}{e_k}$$

permette di valutare una misura della differenza esistente tra le frequenze osservate e le frequenze attese.

Se $\chi^2 = 0$, le frequenze osservate coincidono esattamente con quelle attese. Se invece $\chi^2 > 0$, esse differiscono.

Le frequenze teoriche (o frequenze attese) vengono calcolate sulla base di un'ipotesi H_0 .

Il test χ^2 dell'ipotesi consiste nel valutare se sotto l'ipotesi H_0 , il valore calcolato per χ^2 è più grande di un certo valore critico. In tal caso, dobbiamo concludere che le frequenze osservate differiscono significativamente dalle frequenze attese e dobbiamo rifiutare H_0 al corrispondente livello di significatività.

N.B. – Se χ^2 è troppo vicino allo zero, bisogna essere prudenti sull'accettazione del risultato ottenuto, perché è raro che le frequenze osservate concordino troppo bene con le frequenze teoriche.

In tali situazioni, possiamo determinare se il valore calcolato di χ^2 è minore di $\chi^2_{0.95}$ o di $\chi^2_{0.99}$ (valori critici ai livelli di significatività 0.05 e 0.01 rispettivamente) per decidere se l'accostamento è troppo buono ai livelli di significatività 0.05 o 0.01 rispettivamente.

Esempio

Si deve scegliere tra quattro detersivi per lavatrici di marche diverse, che indichiamo con A, B, C, D, quale sia meno inquinante per l'ambiente.

Utilizzando un campione di 1000 persone, scelto casualmente, è stato chiesto telefonicamente quale detersivo preferisse in base ai residui riscontrati all'interno della lavatrice. Il campione ha così risposto:

detersivo	A	B	C	D	altro
preferenze	221	193	204	195	187

Si può inferire, a partire da questi dati, che la maggioranza delle persone preferisce il detersivo A?

Seguendo il metodo della verifica delle ipotesi, l'ipotesi nulla è:

H_0 : non c'è differenza significativa (al 5%) nelle preferenze tra detersivi.

Se non ci fossero preferenze tra i detersivi, la distribuzione sarebbe (ipotesi attesa):

detersivo	A	B	C	D	altro
preferenze	200	200	200	200	200

Il valore di χ^2 è dato da:

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \dots + \frac{(o_n - e_n)^2}{e_n} = \sum_{k=1}^n \frac{(o_k - e_k)^2}{e_k}$$

in cui ognuno degli addendi è non negativo e quindi χ^2 non è mai negativo.

Inoltre, quanto più χ^2 è vicino allo zero, tanto minore è la differenza tra le frequenze osservate e quelle attese.

Per il calcolo di χ^2 è opportuno costruire la tabella dei dati:

detersivo	osservate	attese	χ^2
A	221	200	2.205
B	193	200	0.245
C	204	200	0.080
D	195	200	0.125
altro	187	200	0.845

da cui:

$$\chi^2 = 3.500$$

Il valore di χ^2 trovato si deve confrontare con quello costruito su una distribuzione normale, che si evince da opportune tavole già predisposte.

Il valore di χ^2 dipende anche da quanti sono gli addendi che abbiamo sommato, perché da essi si deduce il numero v di osservazioni indipendenti: v è detto grado di libertà.

Ipotizzando di accettare o rifiutare l'ipotesi nulla al 5% ed essendo $v = 4$ gradi di libertà, si ha: $\chi^2 = 9.49$.

Allora, la nostra decisione è la seguente: accettiamo l'ipotesi nulla se $\chi < 9.49$.

Poiché $\chi^2 = 3.5$, accettiamo l'ipotesi nulla e inferiamo che non c'è alcuna significativa preferenza tra le quattro marche di detersivo.

Seguendo il metodo della verifica delle ipotesi, l'ipotesi nulla è:

H : non c'è differenza significativa (al 5%) nelle preferenze tra detersivi.

Se non ci fossero preferenze tra i detersivi, la distribuzione sarebbe (ipotesi attesa):

detersivo	A	B	C	D	altro
preferenze	200	200	200	200	200

Il valore di χ è dato da:

$\chi = \dots$

in cui:

ognuno degli addendi è non negativo e quindi χ non è mai negativo.

Quanto più χ è vicino allo zero, tanto minore è la differenza tra le frequenze osservate e quelle attese.

Bibliografia

- F. (1990), *"Le funzioni elementari senza l'utilizzo dell'analisi matematica"* - Convegno nazionale Mathesis 1990, Ed. Luciani - Iseo 1990.

- F. (2004), Appunti estratti dal Corso di Statistica tenuto alla Facoltà di Scienze MM.FF.NN. dell'Università del Sannio nel periodo 2004-2007.

Giambò A., Giambò R. (2009), *Matematica Scuola Superiore*, Ed. Armando Scuola.

A., Squillante M., A.G. Ventre A.G.(2010). *Coherence for Fuzzy Measures and Applications to Decision Making*. Vol. 257, Springer-Verlag, Berlin Heidelberg 2010.

- Casolaro F., Prosperi R. (2011), “Atti Terni 2011: La Matematica per la Scuola di 2° secondo grado: un contributo per il docente di Matematica”. Ed. 2C Contact.

- Prosperi R. (2011), *La Matematica dell'incerto: l'insegnamento del “Calcolo delle Probabilità” e della “Statistica” nel I biennio della Scuola Secondaria di II grado*. Ed. 2C Contact 2011.

- Casolaro F. - Paladino L. (2012): *“Analisi sociale e rigore scientifico. Scelta di equilibrio per l'ottimizzazione dei risultati nell'insegnamento della matematica”* - Logica, linguaggio e didattica della Matematica, a cura della facoltà di Scienze della Formazione dell'Università di Salerno - Edizione Franco Angeli 2012.