ELEMENTI

DI

STATISTICA METODOLOGICA

*

DISTRIBUZIONI DI PROBABILITA’

*

STATISTICA INFERENZIALE

SABO

STATISTICA

Il punto di partenza per la statistica è il:

Fenomeno: fatto che si verifica e che viene osservato; può essere

tipico: soggetto a leggi di tipo generale; si presenta sempre con le stesse caratteristiche (modello matematico).

atipico: non traducibile in modello matematico se preso singolarmente; acquista una certa regolarità quando si effettuano molte osservazioni.

La statistica è la scienza che studia i fenomeni atipici, con il metodo induttivo (dall'osservazione di un singolo fenomeno si traggono leggi di validità generale), per costruire, in base ai principi dell' analisi matematica e del calcolo delle probabilità, i modelli statistici (statistica teorica).

Lo studio della statistica si suddivide in due branche:

Statistica Metodologica: detta anche descrittiva, è l' insieme dei criteri e dei metodi utilizzati per approfondire la conoscenza dei fenomeni collettivi in modo da pervenire ad una visione il più possibile vicino alla realtà.

Statistica Inferenziale: detta anche applicata, si interessa dei fenomeni collettivi specifici di ogni ramo scientifico (Stat. Economica, Stat. Demografica, Stat. Sociale, etc.).

STATISTICA METODOLOGICA

Si basa essenzialmente sulla

INDAGINE STATISTICA	- raccolta dati -spoglio e trascrizione dati - elaborazione dati

è necessario inquadrare bene il fenomeno, formulando in modo corretto l'ipotesi che si vuole sottoporre a verifica e scegliendo i dati in modo da consentire lo studio del fenomeno; bisogna, poi, definire univocamente l'unità statistica, cioè la più piccola unità della quale si raccolgono i dati e che può avere:

- **Carattere qualitativo**: modalità del fenomeno (fenomeno designato da aggettivi o nomi);
- **Carattere quantitativo**: intensità del fenomeno (fenomeno soggetto ad unità di misura);

nonché il tempo dell'indagine:

- Rilevazione continua (senza limiti di tempo; es. nascite, morti, etc.);
- Rilevazione periodica (si ripete ad intervalli regolari; es. censimento);
- Rilevazione occasionale (nel caso si presentino problemi particolari).

Definito ciò si passa alla vera indagine:

-Raccolta dati:	di tipo globale (es. censimento) di tipo campionario (una parte campione)		
-Spoglio dati:	enumerazione dati classificazione in gruppi trascrizione in tabelle	$\longrightarrow$	semplici (due colonne) composte (più colonne) doppia entrata (lettura righe colonne)
-Elaborazione dati:	è di tipo matematico ed ha lo scopo di individuare alcuni indicatori sintetici, significativi del fenomeno oggetto di studio.		

preventivamente è opportuno, nel caso, effettuare la <u>ponderazione dei dati</u> (raggruppare insieme dati aventi le stesse caratteristiche), in modo da esplicitare il peso o frequenza dei termini costituenti il fenomeno oggetto di studio.

Quale utile strumento visivo, i risultati possono essere formulati sotto forma di:

Rappresentazione Grafica: utile per capire l' andamento del fenomeno (utilizzabile in alcuni

casi anche prima di iniziare l' elaborazione); si concretizza in:

-coordinate cartesiane
-istogrammi
-cartogrammi
-ideogrammi
-diagrammi di composizione (torta)
-diagrammi a nastro (bande orizzontali)

FONDAMENTALI FORME DI ELABORAZIONE

- Rapporti statistici;
- Medie;
- Variabilità e Concentrazione;
- Numeri indici.

Rapporti Statistici

Rapporto di Composizione: è un numero puro; può essere:

assoluto: indica quale peso ha ciascuna frequenza sul totale delle frequenze $\left[\frac{f_i}{\sum f_i}\right]$

relativo o percentuale: si ottiene dall' assoluto moltiplicandolo per 100

$$\left[\frac{f_i}{\sum f_i} \cdot 100 = x\%\right]$$

n. b. i rapporti percentuali rendono più facile il raffronto tra i dati, svincolandoli dal particolare e riferendoli tutti ad una stessa base (100).

esempio: nella quinta "A", formata da 16 alunni, vi sono 9 femmine e 7 maschi. Calcolare i rapporti di composizione.

assoluti: femmine = 9/16 = 0,5225 maschi = 7/16 = 0,4375

relativi: femmine = (9/16)∗100 = 52,25% maschi = (7/16)∗100 = 43,75%

Rapporto di Coesistenza: evidenzia la relazione tra due fenomeni diversi in uno stesso luogo o di uno stesso fenomeno in luoghi diversi.

esempio: calcolare il rapporto di coesistenza in quinta "A"

9/7 ≈ 1,29 ——> per ogni maschio poco più di una femmina

Rapporto di Derivazione: evidenzia la relazione tra fenomeni tra loro dipendenti.

esempio: nella scuola, su 494 alunni, 182 risultano provenienti dalla zona a Nord dell' autostrada; calcolare il rapporto di derivazione.

(182/494)∗100 = 36,84% ——> tale percentuale abita a monte della autostrada.

Rapporto di Densità e di Frequenza: (non sono numeri puri) evidenziano la relazione tra l'intensità o la frequenza di un fenomeno e la misura di una grandezza riferita ad un altro fenomeno che può essere messo in relazione con il primo.

esempio: nella sessione estiva dell'a. s. 93/94 su 87 alunni delle terze classi, 51 sono risultati promossi alla classe quarta; calcolare i rapporti di densità e di frequenza.

densità: (51/87)∗100 = 58,62% ——> 6 promossi ogni 10;
frequenza: 87/51 = 1,70 ——> circa 3 promossi ogni 5 alunni.

Rapporto di Durata: (numero non puro) evidenzia come varia in un certo intervallo di tempo l'intensità di un fenomeno a carattere dinamico (che evolve nel tempo); è utilizzato, in genere, in campo demografico, finanziario, aziendale.

esempio: al mare, nel mese di luglio sono presenti 5.000 villeggianti; durante tutta l'estate ne sono arrivati 42.300 e ne sono partiti 41.500; alla fine di settembre ce ne sono ancora 3.000; calcolare il rapporto di durata considerando l' estate formata di tre mesi.

$$\frac{\frac{5000+3000}{2}}{\frac{42300+41500}{2}} = 0{,}0955 \rightarrow 0{,}0955 \cdot 3\text{mesi} = 0{,}864 \rightarrow 0{,}2864 \cdot 30\text{gg} \cong 9$$

il tempo medio di permanenza di un villeggiante al mare e di circa 9 giorni.

Rapporto di Ripetizione: è l' inverso del precedente; indica ogni quanto tempo si rinnova un certo fenomeno.

esempio: calcolare il rapporto di ripetizione relativamente all'esempio precedente

$$\frac{\frac{42300+41500}{2}}{\frac{5000+3000}{2}} = 10{,}475 \cong 10{,}50$$

i villeggianti si rinnovano completamente circa 11 volte in tutta l' estate.

Rapporto Incrementale: esprime la variazione, positiva o negativa, subita da un certo fenomeno nel tempo; si può ottenere da:

- differenza tra ogni dato della serie e il precedente;
- differenza tra ogni dato della serie ed il primo dato;
- differenza tra due dati qualunque della serie.

può essere assoluto o relativo; quello relativo indica la variazione percentuale subita dal fenomeno rispetto ad un dato prefissato.

esempio: calcolare i rapporti incrementali relativamente al numero degli alunni della quinta "A" dall'a. s. 1990/91 all'a. s. 1994/95, rispetto all'a. s. 1990/91.

a. s.	N° Alunni	Incrementi Assoluti	Incrementi Relativi (%)
90/91	21	/	/
91/92	15	-6	-28,57
92/93	16	-5	-23,81
93/94	16	-5	-23,81
94/95	16	-5	-23,81

gli incrementi relativi si ottengono dividendo l'incremento assoluto per il dato prefissato e moltiplicando il risultato per 100:

incremento relativo del 91/92 rispetto al 90/91: $\frac{15-21}{21}\cdot 100 = -28{,}57\%$

incremento relativo del 92/93 rispetto al 90/91: $\frac{16-21}{21}\cdot 100 = -23{,}81\%$

VALORI MEDI

Media: valore riassuntivo di un insieme di dati costituenti un fenomeno; è compreso tra il valore minimo e il valore massimo.

n.b. si possono avere diversi tipi di medie, in relazione alla natura del fenomeno indagato.

Media Aritmetica: indica come l' intensità totale di un fenomeno si distribuisce su tutti i dati, nell' ipotesi in cui ciascun dato abbia la stessa intensità (è il valore che sostituito ad ognuno dei dati non altera la loro somma). $M = \sum(x_i/n)$

Media Geometrica: è utilizzata quando assume significato, ai fini dell' elaborazione, il prodotto dei termini da mediare; in pratica si ricorre alla media geometrica quando i termini della distribuzione tendono a variare in progressione geometrica. $M_g = \sqrt[n]{\prod x_i}$

esempio: un grossista acquista della merce venduta con lo sconto del 6%, poiché è un cliente abituale riceve un ulteriore sconto del 4%; infine pagando in contanti riceve ancora uno sconto del 2%; calcolare il tasso medio di sconto.

$\sqrt[3]{(1-0{,}06)\cdot(1-0{,}04)\cdot(1-0{,}02)} \cong 0{,}95986 \longrightarrow$ 1-0,95986 = 0,04014 = 4,014%

Media Armonica: si utilizza quando tra i termini da mediare esiste un rapporto di proporzionalità inverso, cioè quando i termini tendono a variare in progressione armonica (i loro reciproci variano in progressione aritmetica). Risulta particolarmente utile per valutare il potere di acquisto della moneta. $M_a = \frac{n}{\sum(1/\ x_i)}$

esempio: il ricavo della vendita di un prodotto, pari a Lit. 1.000, è stato ottenuto vendendone 10 pezzi a Lit. 50 con un ricavo di Lit. 500, 25 pezzi a Lit. 20 con un ricavo di Lit. 500; determinare il prezzo medio.

$$\frac{1000}{P_m} = \frac{500}{50} + \frac{500}{20} \rightarrow \frac{1000}{P_m} = \frac{500(50+20)}{50\cdot 20} \rightarrow P_m = \frac{1000}{500}\cdot\frac{50\cdot 20}{50+20} = \frac{1}{\frac{1}{50}+\frac{1}{20}} = 28{,}5$$

Media Quadratica: è utilizzata quando, presentando i termini della distribuzione valori positivi e negativi, si vuole eliminare l' influenza del segno $M_q = \sqrt{\sum(x_i)^2 / \ n}$

Il calcolo di queste medie, dette medie semplici, si effettua quando ogni dato si presenta una sola volta nella distribuzione.

Quando, invece, più dati presentano delle ripetizioni (peso) si utilizza la media ponderata che prevede esplicitamente il diverso peso di ciascun dato. Indicando con f_i il numero di volte che si presenta il dato x_i, risulta: $\sum f_i = n$; il che indica come, in pratica, le formule prima descritte, pur assumendo forma diversa, non cambiano significato. Considerando, infatti, la media aritmetica ponderata si ha:

$$M_p = \frac{\sum(x_i \cdot f_i)}{\sum f_i} = \frac{(x_1 + x_1 + \ldots) + (x_2 + x_2 + \ldots) + \ldots + (x_n + x_n + \ldots)}{n} = \frac{\sum x_i}{n}$$

Relazione tra le Medie: $\mathbf{M_q > M > M_g > M_q}$

Tenendo presente, inoltre, che manipolando opportunamente i termini di una distribuzione, tutte le medie considerate possono essere riportate alla media aritmetica, risulta evidente che, tra tutte, è quest' ultima ad assumere un ruolo preminente ed è, pertanto, la più utilizzata.

MEDIE DI POSIZIONE

sono valori che si ottengono senza dover fare calcoli come è, invece, per le medie precedenti.

Mediana: valore corrispondente al termine posto a metà di tutti gli altri. Rappresenta il termine che occupa il posto centrale quando i dati sono disposti in ordine crescente. Se si è in presenza di termini che si ripetono si ricorre alle frequenze cumulate (somma delle singole frequenze), rappresentando la mediana il termine in corrispondenza del quale la frequenza cumulata supera la semisomma delle frequenze (cioè il 50%).

Moda: termine al quale corrisponde la massima frequenza. Risulta da ciò che la moda è un valore calcolabile solo nel caso di una distribuzione ponderata di dati.

VARIABILITÀ

Parametro statistico che dà informazioni sulla distribuzione dei dati di un fenomeno rispetto alla loro media. Si possono avere indici assoluti (riferiti ad uno stesso fenomeno ed espressi nella stessa unità di misura dei dati), ed indici relativi che, rapportati al valor medio, permettono di confrontare la variabilità di distribuzioni riferite a grandezze diverse.

Campo di Variazione: differenza tra il valore massimo ed il valore minimo di una distribuzione. E' un dato molto grezzo che non sempre riesce a fornire informazioni utili sulla variabilità del fenomeno.

Deviazione Quartile: connesso al concetto di mediana, divide il campo di variazione in quattro parti:

- 1° quartile: valore che supera il 25% della distribuzione (mediana della prima metà del campo di variazione);
- 2° quartile: valore che supera il 50% della distribuzione (coincide con la mediana);
- 3° quartile: valore che supera il 75% della distribuzione (mediana della seconda metà del campo di variazione);
- 4° quartile: comprende il 100% della distribuzione.

La differenza tra il terzo ed il primo quartile dà la deviazione quartile assoluta.

Scarto Semplice Medio: valore dipendente da tutti i dati del fenomeno; indica di quanto i valori stessi si discostano dal valore medio. E' dato dalla media aritmetica dei valori assoluti degli scarti dal valor medio.

$$S = \frac{\left|x_i - \mu\right|}{n}$$

Viene utilizzato, soprattutto, quando si considera come valore medio la mediana, perché la somma dei valori assoluti degli scarti dalla mediana è minima.

esempio: calcolare lo scarto semplice medio tra le classi del corso “A”:

Classe Corso “A”	numero Alunni	Frequenza Relativa	Frequenza Cumulata	Scarto da μ in valore assoluto	Scarto da M in valore assoluto
1°	26	26,53%	26,53%	6,4	7
2°	20	20,41%	46,94%	0,4	1
3°	17	17,35%	64,29%	2,6	2
4°	19	19,38%	83,67%	0,6	0
5°	16	16,33%	100,00%	3,6	3
Totale	98	100,00%	///	13,6	13

media aritmetica = M = 19,6 mediana = M_e = 19

S(mediana) < S(media) —> Scarto(media) = 13,6/5 = 2,72

Scarto(mediana) = 13,0/5 = 2,60

S(mediana) indica che mediamente le classi variano tre di loro di 2,6 alunni.

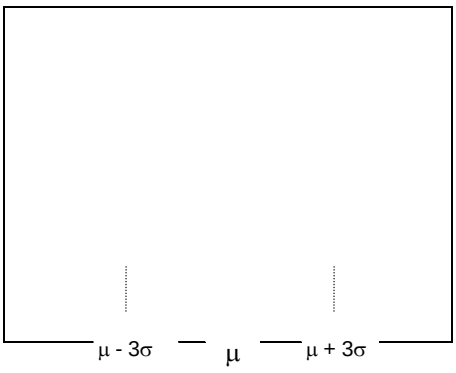
Scarto Quadratico Medio: è il più utilizzato tra gli stimatori della variabilità, poiché rappresenta con buona sensibilità anche le più piccole variazioni tra i valori di un fenomeno. E’ dato dalla radice quadrata della media degli scarti al quadrato. E’ utilizzato, soprattutto, quando come valore medio si utilizza la media aritmetica. Più alto è σ più è alta la variabilità del fenomeno.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

n. b. il valore di σ è fondamentale per la definizione della curva di Gauss che rappresenta graficamente l’ intensità dei fenomeni aventi distribuzione normale (distribuzione simmetrica rispetto alla media), cioè di una distribuzione la cui quasi totalità dei dati è interna all’ intervallo $\mu \pm 3\sigma$.

Analiticamente la curva di Gauss è espressa da:

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2} \right]$$



CONCENTRAZIONE

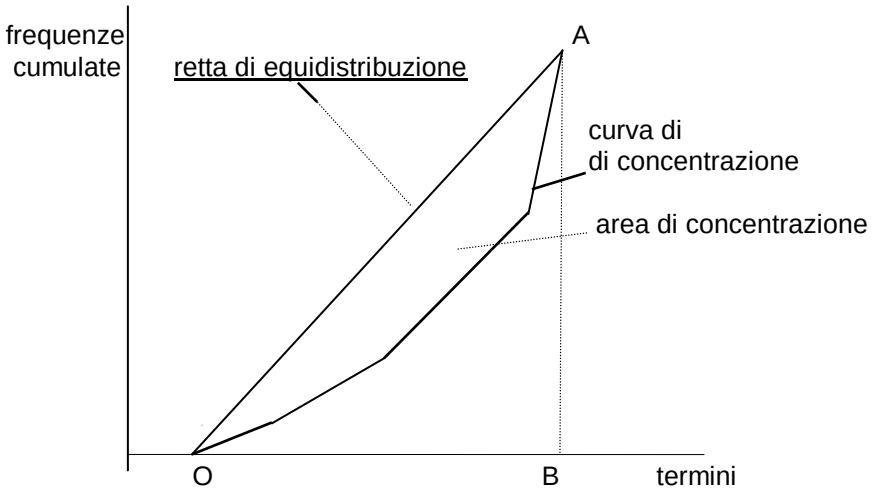
Aspetto particolare della variabilità, la concentrazione pone in evidenza come i termini di una distribuzione possono distribuirsi in modo non uniforme e dà, quindi, la possibilità di verificare se e come è presente l'addensamento dell'intensità di un fenomeno.

L' indice di concentrazione, R, si ottiene dal rapporto tra la situazione reale e la combinazione delle situazioni teoriche che prevedono un'equidistribuzione tra i termini e la massima concentrazione di un solo termine.

$$R = \frac{\text{area concentrazione}}{\text{area triangolo OAB}} \qquad 0 \leq R \leq 1$$

R = 0 —> equidistribuzione

R = 1 —> max concentrazione



n. b. R = 1 indica, in pratica, che tutti i dati sono nulli tranne uno che possiede la massima intensità.

NUMERI INDICI

Rapporti statistici utilizzati per misurare l' intensità di uno stesso fenomeno o di fenomeni diversi, in riferimento ad epoche successive. Si suddividono in:

- numeri indici semplici: riferiti ad uno stesso fenomeno;
- numeri indici composti: riferiti a fenomeni diversi.

in entrambi i casi possono aversi:

* numeri indici a base fissa;
* numeri indici a base mobile

base fissa: si assume uno dei dati del fenomeno, posto uguale a 100, come riferimento e si rapportano ad esso tutti gli altri dati.

base mobile: si ottengono dividendo ogni dato per il precedente.

esempio: calcolare i numeri indici relativi alla popolazione scolastica dall'a. s. 1989/90 all'a. s. 1993/94, prendendo come base l' anno scolastico 1989/90..

Anno Scolastico	Numero Alunni	Numeri Indici (%) a base fissa	Numeri Indici (%) a base mobile
89/90	480	100,00	///
90/91	514	107,08	107,08
91/92	470	97,92	91,44
92/93	414	86,25	88,08
93/94	351	73,12	84,78

I numeri indici composti rappresentano la variazione complessiva di due o più fenomeni. Possono calcolarsi o come media dei numeri indici semplici dei singoli fenomeni o come numeri indici delle medie dei singoli fenomeni.

RELAZIONI STATISTICHE

Spesso nello studiare un fenomeno statistico ci si trova di fronte al problema dell' interpolazione dei dati, sia al fine della descrizione dei fenomeni, sia al fine del sostegno alla decisione razionale. Si ricorre, in tal caso ad alcune tecniche che permettono di superare il problema; le più importanti sono:

INTERPOLAZIONE

tecnica utilizzata per determinare i valori mancanti di una serie lacunosa di dati (interpolazione per punti); oppure per correggere, in una serie di dati, valori affetti da errori accidentali che possono alterare lo studio del fenomeno (interpolazione tra punti).

Interpolazione per Punti: utilizzata per quei fenomeni retti da leggi del tipo y = f(x). Il caso più frequente è l' interpolazione lineare, espressa analiticamente dall' equazione della retta passante per due punti:

$$\frac{y - y_1}{y_2 - y_1} = \frac{x - x_1}{x_2 - x_1}$$

tale metodo permette di ricercare la funzione il cui grafico passa per tutti i punti assegnati, attraverso una spezzata.

Interpolazione tra Punti: utilizzata per ricercare la retta che meglio si adatta a rappresentare un fenomeno ad andamento lineare (i valori si addensano intorno ad una retta). Il calcolo che meglio permette di ricercare la retta interpolante è espresso dal metodo dei minimi quadrati.

Si tratta, dunque, di ricercare la retta y=mx+p determinando i parametri m e p in modo che risulti minima la somma dei quadrati degli scostamenti dei dati teorici (ottenuti dall' interpolazione) dai rispettivi dati empirici, rispetto alla somma dei quadrati degli scarti che si otterrebbero con qualsiasi altra retta interpolante.

$$p = \bar{y} - m\,\bar{x} \qquad\qquad m = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

x_i , y_i = coppie di valori di cui si vuole determinare la retta dei minimi quadrati;

$\bar{x}$ = media aritmetica dei valori x_i;

$\bar{y}$ = media aritmetica dei valori y_i.

Estrapolazione: viene utilizzata per determinare valori esterni all' intervallo di osservazione; permette, in pratica, di prevedere l' evolversi di un fenomeno descritto dalla retta interpolante.

n. b. tale tecnica può dare risultati non attendibili quando si ricercano valori molto lontani dall' intervallo dei valori tabulati.

Perequazione: utilizzata per sostituire valori empirici di un fenomeno (quelli rilevati) con valori teorici più rispondenti all' andamento del fenomeno stesso.

Un metodo è quello della funzione interpolante; un altro, meno analitico e più meccanico, è quello per medie mobili che consiste, essenzialmente, nel sostituire ciascun termine con la media aritmetica considerando, oltre il termine da sostituire, un uguale numero di termini a sinistra e a destra di esso (in genere si scelgono medie di tre o cinque termini).

CONNESSIONE - REGRESSIONE - CORRELAZIONE

Vengono utilizzate quando è necessario stabilire se esiste un legame tra i dati di due fenomeni o tra più dati di uno stesso fenomeno. In particolare, con riferimento a legami di tipo lineare, si ricorre alla:

Connessione: per studiare il grado di dipendenza tra due caratteri qualitativi;

Regressione: per studiare il grado di dipendenza tra due caratteri quantitativi;

Correlazione: per stabilire il grado di interdipendenza di caratteri quantitativi di un dato fenomeno.

n. b. mentre con la connessione e con la correlazione si ottengono degli indicatori del grado di dipendenza tra le due grandezze in esame, con la regressione si ha una funzione che, per legami di tipo lineare, viene detta retta di regressione.

TEORIA DELLA CONNESSIONE

si ordinano i dati rilevati in una tabella a doppia entrata (tabella di connessione) in cui le righe contengono i valori qualitativi del carattere y e le colonne i valori qualitativi del carattere x; i valori n_{ij}, risultanti dall' incrocio delle righe e colonne, rappresentano il numero dei dati che hanno contemporaneamente modalità x_j e y_i; per essi valgono le relazioni:

Carattere X							
Carattere Y	x_1	x_2	...	x_j	...	x_k	Totale
y_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1k}	n_{10}
y_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2k}	n_{20}
...	...	...	...	...	...	...	...
y_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ik}	n_{i0}
...	...	...	...	...	...	...	...
y_h	n_{h1}	n_{h2}	...	n_{hj}	...	n_{hk}	n_{h0}
Totale	n_{01}	n_{02}	...	n_{0j}	...	n_{0k}	n

$$n_{i0} = n_{i1} + n_{i2} + ... + n_{ik} = \sum n_{ij} \quad (i = 1,...,h)$$

$$n_{0j} = n_{1j} + n_{2j} + ... + n_{hj} = \sum n_{ij} \quad (j = 1,...,k)$$

$$\sum n_{i0} = \sum n_{0j} = \sum \sum n_{ij} = n$$

Partendo da questi dati è possibile costruire una nuova tabella (tabella delle frequenze teoriche), simile nella struttura alla precedente, il cui generico elemento N_{ij} si ottiene dalla proporzione:

$$N_{ij} : n_{i0} = n_{0j} : n \quad \longrightarrow \quad N_{ij} = (n_{i0} * n_{0j})/n$$

supponendo una perfetta indipendenza tra i due caratteri. Se per tutte le coppie di indice i e j risulta $n_{ij} = N_{ij}$, i due caratteri sono indipendenti; se, invece, per qualche coppia risulta $n_{ij} \neq N_{ij}$, vuol dire che esiste una connessione tra i due caratteri. In tal caso si costruisce un' ulteriore tabella, sempre simile nella struttura alla prima (tabella di contingenza), i cui elementi si ottengono dalla relazione:

$$c_{ij} = n_{ij} - N_{ij}$$

noti i valori c_{ij}, è possibile calcolare l' indice di contingenza χ^2 (chi-quadro) o di Pearson attraverso la relazione:

$$\chi^2 = \sum_{ij} \frac{c_{ij}^2}{N_{ij}}$$

e quindi il coefficiente di contingenza di Pearson:

$$c = \sqrt{\frac{\chi^2}{(\chi^2 + n)}} \qquad (0 \leq c \leq 1)$$

c = 0 ——> perfetta indipendenza c = 1 ——> perfetta dipendenza

esempio: determinare se esiste un qualche legame ed in che misura tra i risultati finali dell' a. s. 1993/94, avendo suddiviso gli alunni in biennio e triennio.

	Biennio	Triennio	Totale
Promossi	59	193	252
non Promossi	113	94	207
Totale	172	287	459

Si costruisce la tabella delle frequenze teoriche in base alla relazione: $N_{ij} = \frac{n_{i0} \cdot n_{0j}}{n}$

ad esempio per calcolare N_{11} si ha: $N_{11} = \frac{252 \cdot 172}{459} = 94{,}43$

	Biennio	Triennio	totale
Promossi	94,43	157,57	252
non Promossi	77,57	129,43	207
Totale	172	287	459

da questa si costruisce, poi, la tabella di contingenza in base alla relazione: $c_{ij} = n_{ij} - N_{ij}$ da cui è possibile ricavare l' indice di contingenza χ^2 o di Pearson:

$$\chi^2 = \frac{(-35{,}43)^2}{94{,}43} + \frac{(35{,}43)^2}{157{,}57} + \frac{(35{,}43)^2}{77{,}57} + \frac{(-35{,}43)^2}{129{,}43} = 47{,}141$$

e quindi il coefficiente di contingenza di Pearson:

$$c = \sqrt{\frac{47{,}141}{(47{,}141 + 459)}} = 0{,}305$$

essendo c molto prossimo a zero si può affermare che i due fenomeni non sono tra loro influenzati.

TEORIA DELLA REGRESSIONE

Con l' interpolazione si è visto che, data una serie di valori di x, è possibile stimare attraverso la retta interpolante i corrispondenti valori di y.

E' possibile, inoltre, con il metodo dei minimi quadrati descrivere l' equazione della retta che regola la dipendenza tra le due variabili x e y e, quindi, di rappresentare graficamente sia i valori effettivi sia i valori teorici; in particolare la retta che rappresenta i valori teorici prende il nome di Retta di Regressione

$$y = mx + p$$

Nell' interpretare tale retta assume notevole importanza il coefficiente m (coefficiente di regressione) che rappresenta la variazione media di y per un incremento unitario di x; in particolare si ha:

m = 0 ——> non esiste alcuna relazione tra le due variabili;
m > 0 ——> al crescere di x cresce y;
m < 0 ——> al crescere di x decresce y.

Rappresentando, d' altronde, tale retta dei valori teorici, è opportuno misurare la sua attendibilità tramite una misura della dispersione dei dati intorno ad essa; ciò può essere fatto calcolando l' Errore Standard di Stima

$$S_{y,x} = \sqrt{\frac{(y - y^*)^2}{N}}$$

essendo y i valori osservati ed y* i valori calcolati; più quest' indice è piccolo più è precisa la relazione tra x e y, al limite se esso è nullo tutti i punti giacciono sulla retta di regressione.

esempio: determinare la relazione esistente tra i promossi e i rimandati del corso "A" nell' a. s. 1993/94

Classi corso "A"	Promossi x_i	Rimandati y_i	Scarto $(x_i - \bar{x})$	Scarto $(y_i - \bar{y})$	Prodotto $(x_i - \bar{x})(y_i - \bar{y})$	Scarto quad $(x_i - \bar{x})^2$	Scarto quad $(y_i - \bar{y})^2$
1° A	8	8	-2,75	2,00	-5,50	7,56	4,00
2° A	8	9	-2,75	3,00	-8,25	7,56	9,00

3° A	14	4	3,25	-2,00	-6,50	10,56	4,00
4° A	13	3	2,25	-3,00	-6,75	5,06	9,00
Somma	43	24	///	///	-27,00	30,75	26,00

$\bar{x} = 10,75$; $\bar{y} = 6,00$; $\sigma_x = 2,77$; $\sigma_y = 2,55$

costruzione retta di regressione: $y = mx + p$

$$m = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum(x_i - \bar{x})} = -\frac{27,00}{30,75} \cong -0,88 \quad ; \quad p = \bar{y} - m \cdot \bar{x} = 6,00 + 0,88 \cdot 10,75 \cong 15,44$$

$$y = -0,88x + 15,44$$

Calcolo dell' errore di stima:

x_i	y_i	y^*_i	$(y_i - y^*_i)$
8	8	8,40	0,16
8	9	8,40	0,36
14	4	3,12	0,77
13	3	4,00	1,00
		Somma	2,29

$S_{x,y} = \sqrt{\Sigma(y_i - y^*_i)^2/N} = \sqrt{2,29/4} \approx 15,44$

TEORIA DELLA CORRELAZIONE

E' utilizzata per misurare il grado di influenza esistente tra due variabili, supposto che tra di esse esista un certo legame. Per poter stimare il grado di correlazione si utilizza il coefficiente di correlazione lineare o Indice di Bravais:

$$r = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n \cdot \sigma_x \cdot \sigma_y}$$

essendo n il numero dei dati osservati (stesso valore per x e per y), σ_x lo scarto quadratico medio della distribuzione x, σ_y lo scarto quadratico medio della distribuzione y.

Il coefficiente r è interno all' intervallo $-1 \leq r \leq 1$

- $r > 0$ ——> correlazione diretta (al crescere di x cresce y);
- $r < 0$ ——> correlazione inversa (al crescere di x decresce y);
- $r = 1$ ——> perfetta correlazione positiva;
- $r = -1$ ——> perfetta correlazione negativa;
- $r = 0$ ——> variabili perfettamente indipendenti tra loro.

esempio: calcolare la correlazione esistente tra i promossi e i rimandati del corso "A" nell' a. s. 1993/94.

Corso "A"	Promossi	Rimandati	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x}) * (y - \bar{y})$

1° A	8	8	-2,75	2,00	-5,50
2° A	8	9	-2,75	3,00	-8,25
3° A	14	4	3,25	-2,00	-6,50
4° A	13	3	2,25	-3,00	-6,75
somma	43	24	///	///	-27,00

$M_p = 10,75$; $M_R = 6,00$; $\sigma_p = 2,77$; $\sigma_R = 2,55$

$$r = \frac{-27,00}{(4 \cdot 2,77 \cdot 2,55)} = -0,96$$

esiste una correlazione inversa tra i due fenomeni molto marcata, come d'altronde c'era da aspettarsi, dal momento che all' aumentare dei promossi diminuiscono i rimandati e viceversa.

DISTRIBUZIONI DI PROBABILITA’

Data una distribuzione statistica di un fenomeno caratterizzata da frequenze assolute, è possibile trasformare queste ultime in frequenze relative potendole, così, interpretare in termini frequentistici e quindi, per la legge empirica del caso, come probabilità. E' possibile, in tal modo, leggere una distribuzione statistica come una distribuzione di probabilità di variabile casuale; ciò risulta di grande importanza pratica perché è possibile, in tal modo assegnare valori numerici, trattabili matematicamente, agli eventi di uno spazio campione relativo ad una prova e di quantificarne i suoi risultati in forma tabellare, in forma grafica e mediante formule. Questo assunto dà la possibilità di sostituire ad un modello statistico concreto un modello probabilistico teorico, scegliendo tra le distribuzioni di probabilità quella che meglio si adatta alla distribuzione statistica considerata sulla base del confronto tra la rappresentazione grafica dei dati empirici e la forma della funzione di distribuzione di probabilità.

E' opportuno, dunque, studiare alcune tra le più importanti e più utilizzate tra le distribuzioni di probabilità, sia nel discreto che nel continuo.

Distribuzione Binomiale

Si consideri il seguente problema: calcolare la probabilità che, lanciando 15 volte due dadi, la somma delle facce dia per 8 volte come risultato 6. I dati in possesso sono:

Evento richiesto: somma pari a 6 ⟶ p = 5/36

Evento contrario: somma non pari a 6 ⟶ q = 1 - p = 31/36

(i due eventi risultano tra loro indipendenti)

si richiede la probabilità che tale somma esca 8 volte, si tratta, pertanto, di una probabilità composta (prodotto delle probabilità dei singoli lanci).

La probabilità richiesta sarà data dal prodotto di 8 volte la probabilità che esca tale somma per 7 volte la probabilità che non esca:

$$p = \left(\frac{5}{36}\right)^8 \cdot \left(\frac{31}{36}\right)^7$$

tale espressione, però, ipotizza che si abbiano prima i successi e poi gli insuccessi; mentre, generalmente essi possono presentarsi distribuiti secondo le combinazioni di 15 elementi presi a gruppi di 8, l' espressione, pertanto, assume la forma:

$$p = \binom{15}{8} \cdot \left(\frac{5}{36}\right)^8 \cdot \left(\frac{31}{36}\right)^7 = 3{,}13 \cdot 10^{-4}$$

Generalizzando il discorso, la probabilità che, effettuando n prove, un evento avente probabilità costante di verificarsi pari a p e probabilità contraria q = 1 - p, si presenti x volte ($x \leq n$) è data da:

$$p = \binom{n}{x} \cdot p^x \cdot q^{(n-x)} \qquad \text{(prove ripetute)}$$

Indicando, poi, con X il numero di volte che si presenta l' evento richiesto nelle n prove, tale numero è una variabile casuale definita da:

$$X \begin{cases} 0 & 1 & 2 & \dots & x & \dots & n \\ p_{n,0} & p_{n,1} & p_{n,2} & \dots & p_{n,x} & \dots & p_{n,n} \end{cases}$$

detta Distribuzione Binomiale o di Bernoulli, avente rispettivamente media e varianza:

$$\mu = n*p \quad ; \quad Var(X) = n*p*q \quad \left[\sigma(x) = \sqrt{n \cdot p \cdot q}\right]$$

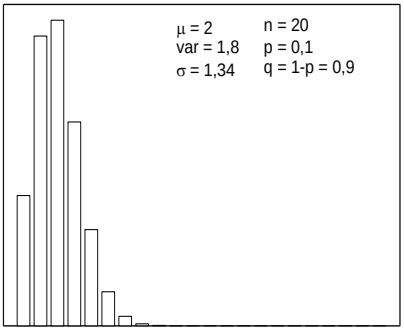
Di questa distribuzione è, in pratica, importante conoscere il numero n delle prove e la probabilità p dell' evento, giacché è proprio al variare di tali valori che varia l' andamento della distribuzione.

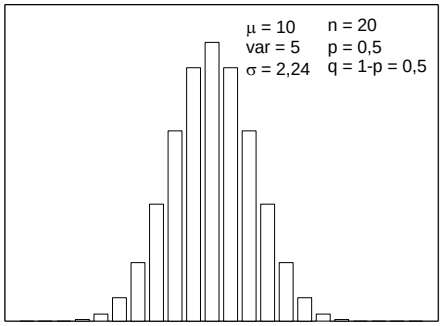
E’ interessante notare, a questo proposito ciò che accade tenendo costante uno dei due valori e facendo variare l’ altro:

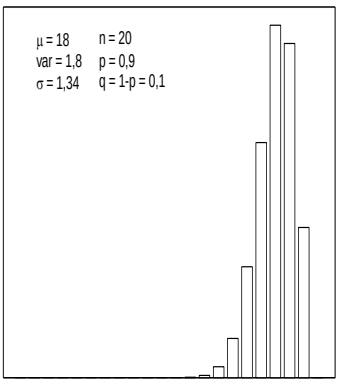
1° caso:

n = cost. ; p = variabile

in questo caso si nota che per valori di p lontani da 0,5 la concentrazione si avvicina all’ origine per p<0,5 e se ne allontana per p>0,5



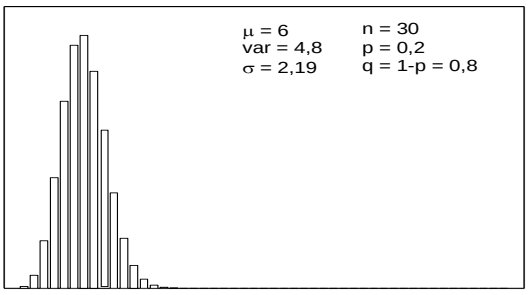


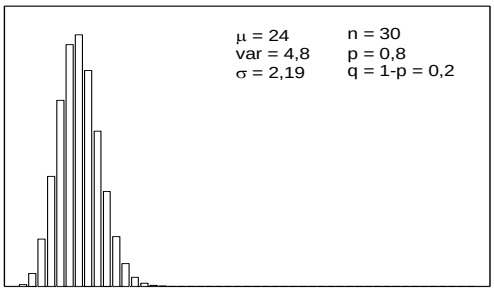


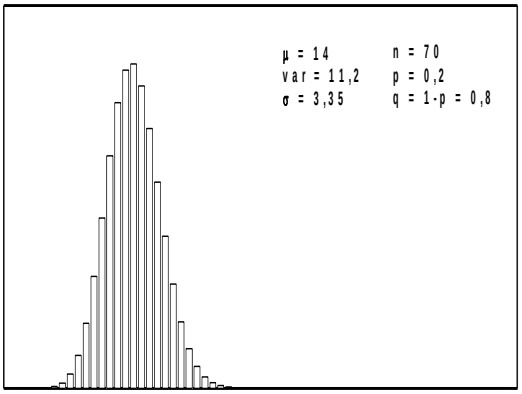
(le due concentrazioni sono perfettamente simmetriche); nel caso di p = 0,5 la concentrazione, invece, si ha in prossimità del valore centrale di n (punto per cui passa l’ asse di simmetria della distribuzione).

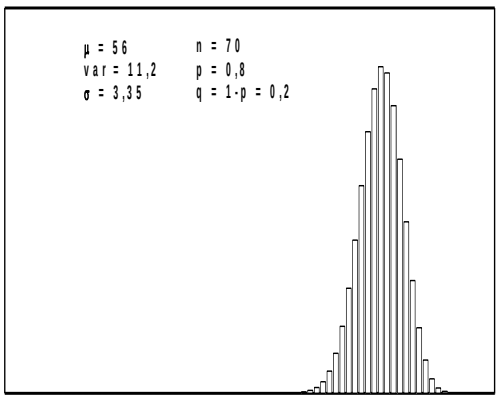
2° caso: n = variabile ; p = cost.

in questo caso si nota che pur variando n la concentrazione si mantiene abbastanza vicino all’ asse delle ordinate quando p < 0,5; tende, invece, ad allontanarsi quando p > 0,5 ed n è molto alto.







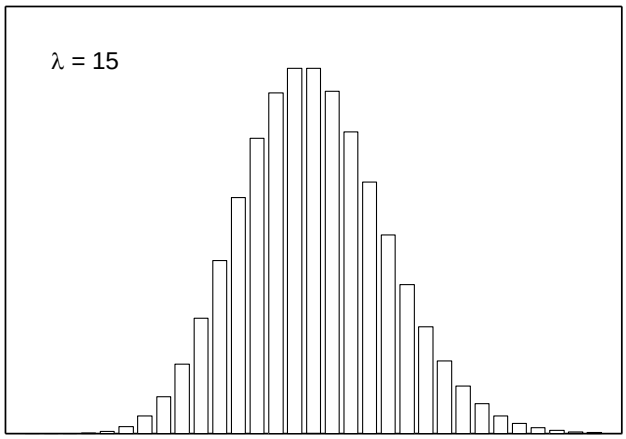


Distribuzione di Poisson

Quando in una distribuzione binomiale il numero dei successi x è molto piccolo rispetto al numero delle prove e la probabilità che essi si verifichino è molto bassa (n —>∞ e p ≤ 0,05) è possibile fare delle semplificazioni sulla distribuzione binomiale trasformandola in una nuova distribuzione detta Distribuzione di Poisson, la cui funzione di probabilità è:

$$P(X) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}$$

essendo λ una costante positiva (nell' approssimazione della binomiale alla poissoniana risulta $\lambda = n*p$). In pratica la distribuzione di Poisson viene utilizzata quando si considerano problemi nei quali non si conosce il numero delle prove ma di cui, tuttavia, si conosce la frequenza λ, costante, di un certo numero di successi



che si presentano secondo una certa legge esponenziale.

esempio: si è accertato che nella segreteria telefonica di un ufficio pervengono nel corso di un'ora, mediamente, 4 telefonate; calcolare la probabilità che in un'ora non arrivino telefonate, che arrivi una telefonata, che in due ore arrivino due telefonate.

Dai dati risulta $\lambda = 4$, per cui l'arrivo delle telefonate in un'ora ha probabilità pari a:

$$P(X) = \frac{4^x}{x!} \cdot e^{-4} \qquad (x = 1, 2, 3, ...)$$

a) la probabilità che in un'ora non arrivino telefonate è:

$$x = 0 \longrightarrow \qquad P(x = 0) = \frac{4^0}{0!} \cdot e^{-4}$$

b) la probabilità che in un'ora arrivi una telefonata è:

$$x = 1 \longrightarrow \qquad P(x = 1) = \frac{4^1}{1!} \cdot e^{-4}$$

c) per calcolare la probabilità che in due ore possano arrivare due telefonate, bisogna tener presente che possono arrivare due telefonate durante la prima ora e nessuna nella seconda; oppure nessuna nella prima ora e due nella seconda; o, ancora, una telefonata durante la prima ora e una durante la seconda ora. Ognuno di questi casi comporta una probabilità composta e tutti insieme una probabilità totale, per cui la probabilità richiesta è:

$$P = \left(\frac{4^2}{2!}\cdot e^{-4}\cdot\frac{4^0}{0!}\cdot e^{-4}\right)+\left(\frac{4^0}{0!}\cdot e^{-4}\cdot\frac{4^2}{2!}\cdot e^{-4}\right)+\left(\frac{4^1}{1!}\cdot e^{-4}\cdot\frac{4^1}{1!}\cdot e^{-4}\right) = 0{,}0107$$

n. b. la distribuzione di Poisson è caratterizzata dall'avere il valore medio uguale alla varianza:

$$\mu = \text{var} = \lambda \quad \longrightarrow \sigma = \sqrt{\lambda}$$

Distribuzioni di Variabili Continue

Le distribuzioni finora viste sono relative a variabili casuali che assumono valori interi. Molto spesso, però, si presentano problemi in cui i casi possibili si presentano con continuità, dando luogo a variabili casuali continue.

variabile casuale continua: variabile che può assumere qualunque valore all'interno di un intervallo limitato o illimitato (a, b) con una probabilità $P(a < x < b)$.

Distribuzione di Gauss

Per poter definire la distribuzione di probabilità di una tale variabile bisogna far riferimento ad una funzione $f(x) \geq 0$, detta Densità di Probabilità, il cui grafico è una curva continua (curva di densità) che dovendo, comunque, rispettare le caratteristiche proprie di una variabile casuale, possiede un asse di simmetria intorno al quale si concentra la maggior parte dei valori, con un massimo che giace su tale asse e con valori decrescenti via via che ci si allontana da esso.

Ciò porta a concludere che la curva assume un andamento a campana ed è descritta analiticamente dalla funzione:

$$P(X = x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{(x - \mu)^2}{2\sigma^2} \right]$$

detta Distribuzione di Probabilità Normale o Funzione di Gauss, in cui μ è il valore medio e σ^2 è la varianza della distribuzione.

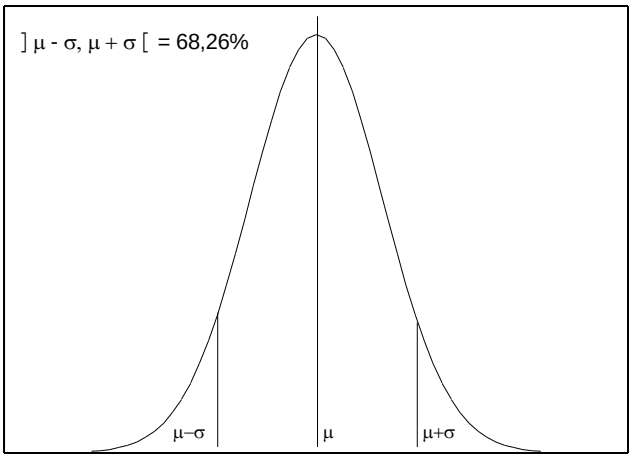
Questi due parametri caratterizzano fortemente l' andamento della curva, in particolare:

μ = valor medio: rappresenta il valore massimo assunto dalla curva; controlla la distanza dell' asse di simmetria dall' asse delle ordinate.

σ^2 = varianza: controlla la concentrazione della curva intorno al suo asse di simmetria (σ piccolo indica una curva stretta ed alta; σ grande indica una curva larga e bassa).

Da notare, infine, che il fattore $1/(\sigma\sqrt{2\pi})$ è un fattore di scala per le ordinate che controlla la grandezza della curva e rende pari ad 1 l' area ad essa sottesa.

Da quanto detto è facile intuire che la curva presenta la maggiore concentrazione nell' intervallo $]\mu-\sigma\ ,\ \mu+\sigma[$ il che porta a dire, anche, che i punti $x = \mu \pm \sigma$ (e quindi σ) rappresentano i punti di flesso della curva, i punti cioé, dove varia il suo andamento.



Altri intervalli tipici definiti da σ sono: $]\mu-2\sigma\ ,\ \mu+2\sigma[$ in cui cade il 95,44% dei punti della distribuzione e $]\mu-3\sigma\ ,\ \mu+3\sigma[$ in cui cade il 99,73%; quest' ultimo intervallo, anzi, permette di affermare che il valore assunto dalla variabile casuale risulta essere quasi certamente compreso tra tali estremi.

Distribuzione Normale Standard

Ponendo nella distribuzione normale

$$z = \frac{x - \mu}{\sigma} \qquad \text{con } \mu = 0 \quad \text{e } \sigma = 1$$

si ottiene la funzione:

$$p(z) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{z^2}{2}\right)$$

detta Distribuzione Normale Standard che, non avendo parametri, facilita i calcoli nella maggior parte dei casi che si possono presentare nella pratica (anche perché tale funzione risulta tabellata).

Ha, inoltre, il vantaggio che, essendo simmetrica rispetto all'asse delle ordinate, permette di calcolare la probabilità che la variabile assuma un valore negativo considerando l'opposto di tale valore, cioè $p(-x) = p(x)$.

Ma soprattutto, tenendo presente che anche nella normale standard l'area sottesa alla curva deve essere pari ad 1, la normale standard è importante, ed è stata tabellata (tav. 2), per il calcolo della funzione di ripartizione F(x) di una variabile casuale, cioè quando si vogliono calcolare distribuzioni di probabilità del tipo $F(x) = P(X \leq x)$ (probabilità che la variabile casuale X assuma valore non superiore a x) o del tipo $F(x) = P(x_1 \leq X \leq x_2)$.

I casi che si possono presentare per questi tipi di calcolo sono:

1° caso: calcolare la probabilità che la variabile casuale assuma un valore compreso tra $-\infty$ e 2 (valore non superiore a 2)

sulla tav. 2 in corrispondenza di z = 2 si legge f(z) = 0,9772 e tale valore è appunto la probabilità richiesta.

2° caso: calcolare la probabilità che la variabile casuale assuma un valore compreso tra $x_1 = 0$ e $x_2 = 1$

$$P(0 \leq x \leq 1) = (\text{area a sinistra di } 1) - (\text{area a sinistra di } 0) = F(1) - F(0) = 0{,}8413 - 0{,}5000 = 0{,}3413$$

3° caso: calcolare la probabilità che la variabile casuale assuma un valore compreso tra -1,85 e 0; per la simmetria rispetto all'asse delle ordinate risulta:

$$P(-1{,}85 \leq x \leq 0) = P(0 \leq x \leq 1{,}85) = F(1{,}85) - F(80) = 0.9678 - 0{,}5000 = 0{,}4678$$

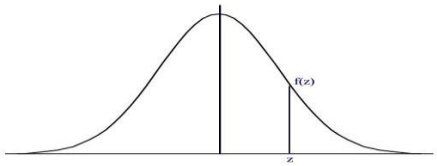
4° caso: calcolare la probabilità che la variabile casuale assuma un valore non superiore a -3: per la simmetria risulta:

$$P(X \leq -3) = P(X \geq 3) = (\text{area totale}) - (\text{area a sinistra di } 3) = 1 - F(3) = 1 - 0{,}9987 = 0{,}0013$$

5° caso: calcolare la probabilità che la variabile casuale assuma un valore compreso tra -2 e 1:

$$P(-2 \leq X \leq 1) = (\text{area a sinistra di } 1) - (\text{area a sinistra di } -2) =$$
$$= F(1) - (\text{area a destra di } 2) = F(1) - [(\text{area totale}) - (\text{area a sinistra di } 2)] =$$
$$= F(1) - [1 - F(2)] = F(1) + F(2) - 1 = 0{,}8413 + 0{,}9772 - 1 = 0{,}8185$$

Aree delle superfici sotto la curva normale standard da -∞ al valore z



z	f(z)	z	f(z)	z	f(z)	z	f(z)	z	f(z)	z	f(z)	z	f(z)	z	f(z)
0,00	0,5000	0,50	0,6915	1,00	0,8413	1,50	0,9332	2,00	0,9772	2,50	0,9938	3,00	0,9987	3,50	0,9998
0,01	0,5040	0,51	0,6950	1,01	0,8438	1,51	0,9345	2,01	0,9778	2,51	0,9940	3,01	0,9987	3,51	0,9998
0,02	0,5080	0,52	0,6985	1,02	0,8461	1,52	0,9357	2,02	0,9783	2,52	0,9941	3,02	0,9987	3,52	0,9998
0,03	0,5120	0,53	0,7019	1,03	0,8485	1,53	0,9370	2,03	0,9788	2,53	0,9943	3,03	0,9988	3,53	0,9998
0,04	0,5160	0,54	0,7054	1,04	0,8508	1,54	0,9382	2,04	0,9793	2,54	0,9945	3,04	0,9988	3,54	0,9998
0,05	0,5199	0,55	0,7088	1,05	0,8531	1,55	0,9394	2,05	0,9798	2,55	0,9946	3,05	0,9989	3,55	0,9998
0,06	0,5239	0,56	0,7123	1,06	0,8554	1,56	0,9406	2,06	0,9803	2,56	0,9948	3,06	0,9989	3,56	0,9998
0,07	0,5279	0,57	0,7157	1,07	0,8577	1,57	0,9418	2,07	0,9808	2,57	0,9949	3,07	0,9989	3,57	0,9998
0,08	0,5319	0,58	0,7190	1,08	0,8599	1,58	0,9429	2,08	0,9812	2,58	0,9951	3,08	0,9990	3,58	0,9998
0,09	0,5359	0,59	0,7224	1,09	0,8621	1,59	0,9441	2,09	0,9817	2,59	0,9952	3,09	0,9990	3,59	0,9998
0,10	0,5398	0,60	0,7257	1,10	0,8643	1,60	0,9452	2,10	0,9821	2,60	0,9953	3,10	0,9990	3,60	0,9998
0,11	0,5438	0,61	0,7291	1,11	0,8665	1,61	0,9463	2,11	0,9826	2,61	0,9955	3,11	0,9991	3,61	0,9998
0,12	0,5478	0,62	0,7324	1,12	0,8686	1,62	0,9474	2,12	0,9830	2,62	0,9956	3,12	0,9991	3,62	0,9999
0,13	0,5517	0,63	0,7357	1,13	0,8708	1,63	0,9484	2,13	0,9834	2,63	0,9957	3,13	0,9991	3,63	0,9999
0,14	0,5557	0,64	0,7389	1,14	0,8729	1,64	0,9495	2,14	0,9838	2,64	0,9959	3,14	0,9992	3,64	0,9999
0,15	0,5596	0,65	0,7422	1,15	0,8749	1,65	0,9505	2,15	0,9842	2,65	0,9960	3,15	0,9992	3,65	0,9999
0,16	0,5636	0,66	0,7454	1,16	0,8770	1,66	0,9515	2,16	0,9846	2,66	0,9961	3,16	0,9992	3,66	0,9999
0,17	0,5675	0,67	0,7486	1,17	0,8790	1,67	0,9525	2,17	0,9850	2,67	0,9962	3,17	0,9992	3,67	0,9999
0,18	0,5714	0,68	0,7517	1,18	0,8810	1,68	0,9535	2,18	0,9854	2,68	0,9963	3,18	0,9993	3,68	0,9999
0,19	0,5753	0,69	0,7549	1,19	0,8830	1,69	0,9545	2,19	0,9857	2,69	0,9964	3,19	0,9993	3,69	0,9999
0,20	0,5793	0,70	0,7580	1,20	0,8849	1,70	0,9554	2,20	0,9861	2,70	0,9965	3,20	0,9993	3,70	0,9999
0,21	0,5832	0,71	0,7611	1,21	0,8869	1,71	0,9564	2,21	0,9864	2,71	0,9966	3,21	0,9993	3,71	0,9999
0,22	0,5871	0,72	0,7642	1,22	0,8888	1,72	0,9573	2,22	0,9868	2,72	0,9967	3,22	0,9994	3,72	0,9999
0,23	0,5910	0,73	0,7673	1,23	0,8907	1,73	0,9582	2,23	0,9871	2,73	0,9968	3,23	0,9994	3,73	0,9999
0,24	0,5948	0,74	0,7704	1,24	0,8925	1,74	0,9591	2,24	0,9875	2,74	0,9969	3,24	0,9994	3,74	0,9999
0,25	0,5987	0,75	0,7734	1,25	0,8944	1,75	0,9599	2,25	0,9878	2,75	0,9970	3,25	0,9994	3,75	0,9999
0,26	0,6026	0,76	0,7764	1,26	0,8962	1,76	0,9608	2,26	0,9881	2,76	0,9971	3,26	0,9994	3,76	0,9999
0,27	0,6064	0,77	0,7794	1,27	0,8980	1,77	0,9616	2,27	0,9884	2,77	0,9972	3,27	0,9995	3,77	0,9999
0,28	0,6103	0,78	0,7823	1,28	0,8997	1,78	0,9625	2,28	0,9887	2,78	0,9973	3,28	0,9995	3,78	0,9999
0,29	0,6141	0,79	0,7852	1,29	0,9015	1,79	0,9633	2,29	0,9890	2,79	0,9974	3,29	0,9995	3,79	0,9999
0,30	0,6179	0,80	0,7881	1,30	0,9032	1,80	0,9641	2,30	0,9893	2,80	0,9974	3,30	0,9995	3,80	0,9999
0,31	0,6217	0,81	0,7910	1,31	0,9049	1,81	0,9649	2,31	0,9896	2,81	0,9975	3,31	0,9995	3,81	0,9999
0,32	0,6255	0,82	0,7939	1,32	0,9066	1,82	0,9656	2,32	0,9898	2,82	0,9976	3,32	0,9995	3,82	0,9999
0,33	0,6293	0,83	0,7967	1,33	0,9082	1,83	0,9664	2,33	0,9901	2,83	0,9977	3,33	0,9996	3,83	0,9999
0,34	0,6331	0,84	0,7995	1,34	0,9099	1,84	0,9671	2,34	0,9904	2,84	0,9977	3,34	0,9996	3,84	0,9999
0,35	0,6368	0,85	0,8023	1,35	0,9115	1,85	0,9678	2,35	0,9906	2,85	0,9978	3,35	0,9996	3,85	0,9999
0,36	0,6406	0,86	0,8051	1,36	0,9131	1,86	0,9686	2,36	0,9909	2,86	0,9979	3,36	0,9996	3,86	0,9999
0,37	0,6443	0,87	0,8078	1,37	0,9147	1,87	0,9693	2,37	0,9911	2,87	0,9979	3,37	0,9996	3,87	0,9999
0,38	0,6480	0,88	0,8106	1,38	0,9162	1,88	0,9699	2,38	0,9913	2,88	0,9980	3,38	0,9996	3,88	0,9999
0,39	0,6517	0,89	0,8133	1,39	0,9177	1,89	0,9706	2,39	0,9916	2,89	0,9981	3,39	0,9997	3,89	0,9999
0,40	0,6554	0,90	0,8159	1,40	0,9192	1,90	0,9713	2,40	0,9918	2,90	0,9981	3,40	0,9997	3,90	1,0000
0,41	0,6591	0,91	0,8186	1,41	0,9207	1,91	0,9719	2,41	0,9920	2,91	0,9982	3,41	0,9997	3,91	1,0000
0,42	0,6628	0,92	0,8212	1,42	0,9222	1,92	0,9726	2,42	0,9922	2,92	0,9982	3,42	0,9997	3,92	1,0000
0,43	0,6664	0,93	0,8238	1,43	0,9236	1,93	0,9732	2,43	0,9925	2,93	0,9983	3,43	0,9997	3,93	1,0000
0,44	0,6700	0,94	0,8264	1,44	0,9251	1,94	0,9738	2,44	0,9927	2,94	0,9984	3,44	0,9997	3,94	1,0000
0,45	0,6736	0,95	0,8289	1,45	0,9265	1,95	0,9744	2,45	0,9929	2,95	0,9984	3,45	0,9997	3,95	1,0000
0,46	0,6772	0,96	0,8315	1,46	0,9279	1,96	0,9750	2,46	0,9931	2,96	0,9985	3,46	0,9997	3,96	1,0000
0,47	0,6808	0,97	0,8340	1,47	0,9292	1,97	0,9756	2,47	0,9932	2,97	0,9985	3,47	0,9997	3,97	1,0000
0,48	0,6844	0,98	0,8365	1,48	0,9306	1,98	0,9761	2,48	0,9934	2,98	0,9986	3,48	0,9997	3,98	1,0000
0,49	0,6879	0,99	0,8389	1,49	0,9319	1,99	0,9767	2,49	0,9936	2,99	0,9986	3,49	0,9998	3,99	1,0000
0,50	0,6915	1,00	0,8413	1,50	0,9332	2,00	0,9772	2,50	0,9938	3,00	0,9987	3,50	0,9998	4,00	1,0000

STATISTICA INFERENZIALE

Spesso nell'effettuare una indagine statistica ci si trova di fronte a fenomeni che non possono essere studiati nella loro completezza; ciò porta a non poter conoscere quei parametri che caratterizzano la loro distribuzione. Nasce, pertanto, la necessità di stimare tale parametri; ciò viene fatto tramite la Teoria del Campionamento, che concentra l'attenzione non su tutto l'universo statistico in esame ma solo su un suo campione, in modo che, applicando il metodo induttivo (dal particolare al generale), si possa giungere a formulare delle leggi di tipo probabilistico che descrivano il fenomeno.

Universo o Popolazione: insieme (N) di unità statistiche omogenee rispetto ad uno o più caratteri; può essere di ampiezza finita o infinita.

Campione: sottoinsieme dell'universo costituito da un numero finito di elementi (inferiore all'ampiezza dell' universo: $n < N$) e scelto in modo che siano rispettate alcune proprietà che consentano di trasferire le informazioni ottenute da esso a tutto l'universo.

Il problema, a questo punto, è di definire il grado di attendibilità del campione nei confronti dell'universo; nasce, cioè, il problema dell'Inferenza Statistica (con quale probabilità le informazioni fornite dal campione possono essere estese all' intero universo).

Chiaramente quanto più grande è l' ampiezza del campione (numerosità, n) tanto più le sue informazioni si avvicinano a quelle dell' universo; la scelta della numerosità del campione e, quindi, del campione stesso è un fattore molto importante poiché deve riprodurre, in modo non deformato, le caratteristiche dell' universo da cui proviene; il campione deve essere, cioè, rappresentativo dell'universo (riprodurre in piccolo le caratteristiche dell' universo); diventa, pertanto, fondamentale il modo in cui il campione viene scelto.

Criteri di Scelta di un Campione

si basano, essenzialmente, su due metodi:

scelta ragionata: il campione è scelto in base a criteri soggettivi che possono condurre anche a risultati lontani dalla realtà.

Scelta Casuale: il campione è scelto estraendo a caso le singole unità; può essere ottenuto utilizzando le tavole dei numeri casuali o, meglio ancora, effettuando delle estrazioni dall' universo.

Nell'utilizzare la scelta casuale si possono seguire due metodi:

estrazione bernoulliana: estrazione con ripetizione; si possono avere N^n campioni distinti (disposizione con ripetizione di N elementi di classe n).

estrazione in blocco: estrazione senza ripetizione; si possono avere $\binom{N}{n}$ campioni distinti (combinazione di N elementi di classe n).

Si capisce da ciò che da un universo si possono estrarre diversi campioni, ognuno dei quali risulta diverso dagli altri e con una sua probabilità di essere formato e, quindi, con una sua probabilità più o meno marcata di riflettere l' universo.

Supposto di estrarre tutti i possibili campioni di dimensione n da un universo di dimensione N e di avere calcolato media e varianza (s.q.m.) di ognuno dei campioni, è possibile considerare una distribuzione delle medie campionarie e una distribuzione delle varianze campionarie; d'altronde, in quanto distribuzioni, di esse si possono calcolare la media e la varianza (in genere si determina solo la media $\Rightarrow$ distribuzione campionaria delle medie o distribuzione della media campionaria).

Per ottenere, dunque, la distribuzione delle medie campionarie $\bar{X}$, bisogna calcolarne la media $M(\bar{X})$ [si dimostra che sia per estrazione bernoulliana sia per estrazione in blocco risulta $M(\bar{X})$ = media dell' universo] e lo scarto quadratico medio (varianza):

$$\sigma(\bar{X}) = \begin{cases} \dfrac{\sigma}{\sqrt{n}} & \longrightarrow \text{ estrazione bernoulliana} \\ \dfrac{\sigma}{\sqrt{n}} \cdot \sqrt{\dfrac{N-n}{N-1}} & \longrightarrow \text{ estrazione in blocco} \end{cases}$$

[σ = s.q.m. dell'universo = valore noto]

In pratica, essendo l'estrazione del campione del tutto casuale, la conoscenza di tali valori sintetici permette di determinare una misura della variabilità della distribuzione della media campionaria (interpretata come variabile casuale); questo significa che, pur non potendo sapere qual'è la media del campione che verrà estratto, se tuttavia n è molto grande ($n \geq 30$), si può affermare che tale media è interna ad una gaussiana di media $M(\bar{X})$ e varianza $\sigma^2(\bar{X})$ con una approssimazione tanto maggiore quanto più grande è n.

Si può, pertanto, considerare ogni campione come una variabile casuale che può assumere valori x_i con probabilità p_i, $[X_i(x_i,p_i)]$; si avrà in tal modo un insieme di variabili casuali tra

loro indipendenti, aventi tutte la stessa distribuzione di probabilità data dalla distribuzione di frequenza con cui un carattere si presenta nell'universo. D'altra parte il metodo di estrazione utilizzato concorre anch'esso a diversificare i campioni, tuttavia se dall'universo è possibile estrarre un numero elevato di campioni, tutti nelle stesse condizioni, se cioè il tasso di campionamento n/N è abbastanza piccolo ($\leq$ 5% $\longrightarrow$ n << N) i due metodi portano alla stessa conclusione e la distribuzione dei campioni tende ad una distribuzione normale, per cui è possibile trasformare una distribuzione statistica (modello concreto) in una distribuzione probabilistica (variabile casuale) (modello teorico); si possono, pertanto, introdurre i concetti propri al calcolo di una variabile casuale e, quindi, effettuare la stima campionaria sul valore medio e sulla varianza del campione (stimatori campionari).

valori sintetici per l' analisi del campione	media
	varianza o scarto qudratico medio

valori sintetici per l' universo $\Rightarrow$ parametri
valori sintetici per il campione $\Rightarrow$ statitiche camponarie o statistiche

stima campionaria: processo di analisi che, partendo dai dati di un campione, determina il corrispondente valore del parametro dell' universo.

parametro: valore caratteristico dell'universo (media, varianza)

stimatore: variabile casuale espressa come funzione delle variabili casuali che compongono il campione; ogni stimatore è dotato di una sua distribuzione, detta distribuzione campionaria, che può variare in base alla numerosità del campione e anche al tipo di stimatore considerato.

TEORIA DELLA STIMA

Viene utilizzata per ottenere da un campione risultati estendibili all'intero universo, determinandone la maggiore o minore esattezza di uno o più parametri; per fare ciò, tuttavia, è necessario che gli stimatori soddisfino a determinate condizioni:

correttezza: uno stimatore è corretto se il suo valore medio è uguale al parametro che deve stimare (in caso contrario si ha una stima distorta);

efficienza: uno stimatore è efficiente se la varianza delle stime è minore della varianza che si può ottenere con altri stimatori;

consistenza: uno stimatore è consistente se al crescere della numerosità del campione il suo valore si avvicina sempre più al parametro da stimare.

Uno stimatore può essere calcolato in due modi diversi:

stima puntuale: l'elaborazione dei dati del campione porta alla determinazione di un unico valore dello stimatore;

stima per intervallo di confidenza: l'elaborazione dei dati del campione porta all' individuazione di un intervallo che contiene, con una data probabilità, lo stimatore; si fissa, in pratica, un intervallo al quale si associa un'alta probabilità di contenere il valore reale del parametro (la precisione della stima è inversa all'ampiezza dell'intervallo, minore è tale ampiezza più precisa è la stima).

STIMA PUNTUALE DELLA MEDIA DELL' UNIVERSO

Si dimostra che il valore medio della media campionaria per un campionamento casuale, sia benoulliano sia in blocco, è uguale alla media dell'universo:

$$\text{media campione} = \text{media universo} \quad \Leftrightarrow \quad \mu_{\bar{x}} = \mu$$

cioè la media di tutte le medie campionarie è uguale a quella dell'universo.

Per quanto riguarda la sua distribuzione, invece, si ha un valore diverso della varianza in base al tipo di estrazione effettuata, in particolare per:

estrazione bernoulliana: $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$

estrazione in blocco: $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$

essendo σ^2 la varianza dell' universo, N l'ampiezza dell' universo, n la numerosità del campione.

n. b. se per il campione risulta n << N (in genere n inferiore al 20% di N) si ha:

$$\lim_{N\to\infty} \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} \cong \frac{\sigma^2}{n}$$

cioè le varianze, per i due tipi di estrazione, tendono a coincidere.

Questo risultato, per la legge dei grandi numeri, può essere esteso anche ad universi con distribuzioni non normali in base al:

Teorema del Limite Centrale: qualunque sia la distribuzione di un universo rispetto ad un carattere quantitativo avente media μ_x e varianza σ^2_x, le medie dei campioni, al crescere della numerosità del campione, tendono ad una distribuzione normale di media μ e varianza σ^2/n.

STIMA PUNTUALE DELLA VARIANZA DELL'UNIVERSO

detta s^2 la distribuzione campionaria della varianza, essa ha media:

$$M(s^2) = \frac{n-1}{n} \cdot \sigma^2 \quad \begin{cases} n = \text{dimensione campione} \\ \sigma^2 = \text{varianza universo} \end{cases}$$

essendo, però, $M(s^2) \neq \sigma^2$ la varianza campionaria è uno stimatore non corretto (distorto) della varianza dell'universo; affinché esso risulti corretto si deve definire la varianza come:

$$\hat{\sigma} = \frac{n}{n-1} \cdot \sigma^2$$

in tal modo si ottiene la distribuzione s^2 delle varianze campionarie corrette e, quindi, si ha:

$$M\left(\hat{s^2}\right) = M\left[\frac{n \cdot s^2}{n-1}\right] = \frac{n \cdot M(s^2)}{n-1} = \frac{n \cdot (n-1)}{(n-1) \cdot n} \cdot \sigma^2 = \sigma^2$$

In pratica, la varianza dell'universo viene assunta come media corretta della distribuzione della varianza campionaria.

STIMA PER INTERVALLO

Lo stimatore viene individuato da un intervallo (intervallo di fiducia o di confidenza) avente una prefissata probabilità di contenerlo (livello di fiducia); l'intervallo può essere rappresentato da:

$$\theta - \varepsilon \leq \bar{\theta} \leq \theta + \varepsilon$$

essendo ε l'errore che si commette nel sostituire il generico stimatore θ al corrispondente parametro $\bar{\theta}$ dell' universo; la relativa probabilità diventa:

$$p(\theta - \varepsilon \leq \bar{\theta} \leq \theta + \varepsilon) = 1 - \alpha$$

essendo α il livello di significatività (rischio di errore) e 1 - α il livello di fiducia che influenza la quantità ε.

Per distribuzioni campionarie di tipo normale la stima per intervallo diventa alquanto semplice e la probabilità che la media dell'universo sia interna all'intervallo di fiducia vale:

$$p\left[\bar{x} - z_C \cdot \frac{s}{\sqrt{n}} \leq m \leq \bar{x} + z_C \cdot \frac{s}{\sqrt{n}} = 1 - a\right]$$

essendo z_c il valore critico che, per un assegnato livello di fiducia (1-α) o rischio di errore α, si ricava dalla tabella della distribuzione normale standardizzata.