

Campioni casuali e distribuzioni campionarie

7.1 | Introduzione

Nello studio di fenomeni variabili, le unità osservate costituiscono sovente una parte soltanto - un campione - dell'insieme che si intende indagare. Trarre dalle caratteristiche del campione le proprietà statistiche di un insieme di ordine superiore (popolazione) significa effettuare una inferenza *induttiva*. Per *campione* si intende l'insieme delle unità osservate; mentre la *popolazione* è l'insieme di ordine superiore da cui il campione è tratto. La popolazione può essere:

- *finita*: una classe chiusa, limitata nello spazio e nel tempo, composta da una pluralità reale di elementi tutti osservabili;
- *non finita*: una classe aperta, caratterizzata da un numero virtualmente illimitato (e pertanto indefinito) di costituenti.

Quando, in una grande popolazione, il fenomeno indagato si manifesta secondo una legge probabilistica, il problema induttivo viene risolto nel contesto teorico di un campionamento da popolazioni non finite.

Una induzione statistica presuppone un campione rappresentativo della popolazione da cui proviene. Per convenzione, si ritiene rappresentativo un campione formato con criterio casuale: l'eventuale mancata

conformità di tale insieme alla popolazione è effetto del solo errore di campionamento. La casualità può essere:

- *governata* tramite un'operazione di sorteggio, o una procedura equivalente a un sorteggio, degli elementi (campionamento probabilistico);
- *postulata*, sulla base di un criterio di formazione del campione indipendente dal carattere oggetto di indagine.

7.2 | Universo dei campioni

L'operazione di "campionamento probabilistico" su di un carattere variabile ha per caratteristica di tradursi in una virtuale pluralità di risultati: un insieme ipotetico detto *universo dei campioni*. In tale "universo" una qualunque costante campionaria diventa una variabile in quanto tende ad assumere, da campione a campione, valori non coincidenti.

Un campione probabilistico è dunque un risultato fra tanti possibili e perciò l'effettivo grado di rappresentatività (nel senso di conformità alla popolazione di riferimento) di un singolo campione è per se stesso indeterminabile. Si può tuttavia darne una teorizzazione formale attraverso le leggi distributive delle variabili in cui si traducono le costanti campionarie.

Universo dei campioni o spazio campionario -
 è l'insieme virtuale costituito da tutti i campioni casuali di uguale numerosità estraibili da una popolazione data, compatibilmente con la procedura di campionamento prescelta.

- *Campionamento bernoulliano*: le unità vengono estratte dalla popolazione una dopo l'altra *con reimmissione*: reinserita prima di ogni successiva estrazione, l'unità appena entrata nel campione può essere nuovamente estratta. È lo schema di prove ripetute di Bernoulli: prove indipendenti a probabilità costante. Data una popolazione di numerosità finita N , l'universo bernoulliano dei campioni di ordine n è costituito da N^n campioni (è il numero di disposizioni con ripetizione

di N elementi presi a gruppi di n) : sono compresi anche campioni composti da una stessa unità ripetuta, al limite, n volte.

- *Campionamento in blocco*: le unità vengono estratte dalla popolazione una dopo l'altra *senza reimmissione* (prescindendo in tale caso, dall'*ordine* con cui le unità entrano nel campione): ogni unità può entrare nel campione una e una sola volta. Data una popolazione di numerosità finita N , l'universo dei campioni in blocco di dimensione n ha cardinalità $\binom{N}{n}$ (è il numero delle combinazioni senza ripetizione di N elementi presi a gruppi di n) e costituisce un sottoinsieme dell'universo bernoulliano del medesimo ordine: il sottoinsieme composto dai soli campioni in cui nessuna unità è ripetuta.

□—————

◇ **Osservazione 7.1.** Se, estraendo le unità una dopo l'altra senza reimmissione, non si prescinde dall'ordine con cui le unità entrano nel campione, la cardinalità dell'universo campionario è pari a $N(N-1)(N-2) \dots (N-n+1)$, cioè al numero delle disposizioni semplici di N elementi presi a gruppi di n .

◇ **Osservazione 7.2.** La divergenza tra campionamento bernoulliano e campionamento in blocco - nonché la differenza tra numerosità dei rispettivi universi campionari: N^n e $\binom{N}{n}$ - si riduce al crescere della dimensione N della popolazione: diventa trascurabile quando la frazione di campionamento n/N è molto piccola; è pressoché inesistente quando la popolazione è una classe aperta.

—————□

Esempio 7.2.1.

Si immagini, per semplicità, un carattere quantitativo che assuma, in una "popolazione" di numerosità $N=7$, i seguenti valori: 1, 2, 3, 4, 5, 6, 7.

È come immaginare un'urna contenente sette palline numerate da 1 a 7.

Nella popolazione, il carattere ha media aritmetica $\mu = (1+2+3+4+5+6+7)/7 = 4$ e varianza $\sigma^2 = [(1-4)^2 + (2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2]/7 = 28/7 = 4$.

Se si estrae un campione bernoulliano di $n = 2$ elementi, questo è uno degli $N^n = 7^2 = 49$ campioni bernoulliani: tanti quante sono le disposizioni con ripetizione di 7 elementi presi a due a due.

1 1	2 1	3 1	4 1	5 1	6 1	7 1
1 2	2 2	3 2	4 2	5 2	6 2	7 2
1 3	2 3	3 3	4 3	5 3	6 3	7 3
1 4	2 4	3 4	4 4	5 4	6 4	7 4
1 5	2 5	3 5	4 5	5 5	6 5	7 5
1 6	2 6	3 6	4 6	5 6	6 6	7 6
1 7	2 7	3 7	4 7	5 7	6 7	7 7

L'universo dei campioni in blocco è invece costituito da $\binom{N}{n} = \binom{7}{2} = 7!/2!5! = 21$ campioni: le combinazioni di 7 elementi a due a due. Sono i campioni scritti al di sotto (o al di sopra) della diagonale tracciata attraverso i campioni 11, 22, 33, 44, 55, 66, 77, che divide idealmente in due parti uguali lo schema dell'universo bernoulliano.

Per la teoria dei campioni, tutti i 49 campioni probabilistici componenti l'universo bernoulliano sono da ritenersi "rappresentativi" in quanto casuali (e, come tali, non "distorti" per fatto sistematico), ma soltanto 7 tra essi (il 14,28%) riproducono esattamente il parametro della popolazione $\mu = 4$. Fissando un intervallo ± 1 intorno alla media aritmetica, ben 29 campioni (pari al 59,16%) hanno media compresa tra 3 e 5. Nell'universo dei 49 campioni bernoulliani (tabella 7.1), la media aritmetica ha distribuzione simmetrica.

La variabile descritta da tutte le medie aritmetiche dei possibili campioni sia \bar{X} e le sue proprietà statistiche, nello spazio campionario, siano espresse dalla $M(\bar{X})$ e dalla $V(\bar{X})$.

Nell'esempio, la media aritmetica delle medie campionarie risulta

Tabella 7.1

medie aritmetiche	numero campioni	frequenze percentuali
1	1	2,04
1,5	2	4,08
2	3	6,12
2,5	4	8,16
3	5	10,20
3,5	6	12,24
4	7	14,28
4,5	6	12,24
5	5	10,20
5,5	4	8,16
6	3	6,12
6,5	2	4,08
7	1	2,04
	49	100

$$M(\bar{X}) = \frac{1}{49}(1 \times 1 + 1,5 \times 2 + 2 \times 3 + \dots + 6,5 \times 2 + 7 \times 1) = 4$$

e la varianza:

$$V(\bar{X}) = \frac{1}{49}[(1-4)^2 \times 1 + (1,5-4)^2 \times 2 + \dots + (6,5-4)^2 \times 2 + (7-4)^2 \times 1] = 2.$$

La media aritmetica di tutte le medie aritmetiche campionarie è dunque uguale alla media aritmetica della popolazione. La varianza delle medie aritmetiche campionarie non è uguale, bensì inferiore, alla varianza della popolazione.

Per $n = 3$, l'universo bernoulliano dei campioni è composto da $7^3 = 343$ campioni. La distribuzione campionaria della media aritmetica in tale universo è riportata nella tabella 7.2.

Per $n = 4$, l'universo bernoulliano è composto da $7^4 = 2401$ campioni.

La distribuzione campionaria della media aritmetica cambia dunque al cambiare della numerosità n del campione: è triangolare per $n = 2$, e già per $n = 3$ e 4 accenna a conformarsi a un profilo campaniforme.

Tabella 7.2

medie aritmetiche	numero campioni	frequenze percentuali
1	1	0,3
1,33	3	0,9
1,67	6	1,7
2	10	2,9
2,33	15	4,4
2,67	21	6,1
3	28	8,2
3,33	33	9,6
3,67	36	10,5
4	37	11,0
4,33	36	10,5
4,67	33	9,6
5	28	8,2
5,33	21	6,1
5,67	15	4,4
6	10	2,9
6,33	6	1,7
6,67	3	0,9
7	1	0,3
	343	100,0

Negli universi di campioni figurano dunque campioni con media assai lontana dalla media della popolazione. Tuttavia, un campione che la riproduca è individualmente più probabile di un campione che non la riproduca: quest'ultimo è tanto più improbabile quanto più la sua media si allontana dalla media della popolazione.

L'operazione di campionamento probabilistico dà quindi luogo a campioni "conformi" e a campioni "difformi", ma con probabilità diverse.

L'asserto vale per la distribuzione campionaria di qualunque proprietà statistica (*parametro*) della popolazione. In particolare, per le costanti cui contribuiscono tutti i valori individuali, all'aumentare di n cresce la frequenza di costanti campionarie (*statistiche*) comprese in un intorno prefissato del parametro della popolazione.

Nell'esempio, l'universo dei campioni di dimensione $n = 2$, comprende il 59,16% dei campioni con media da 3 a 5, mentre nell'universo dei campioni di dimensione $n = 3$, tale frazione sale al 67,6% e in quello di dimensione $n = 4$, al 70%: all'aumentare di n aumenta dunque la probabilità di ottenere campioni con media prossima al parametro da stimare.

7.3 | Distribuzioni campionarie

Si chiama statistica campionaria S una qualsiasi funzione $h(\cdot)$, indipendente da parametri incogniti, che riferita alle variabili aleatorie campionarie $X_1, \dots, X_j, \dots, X_n$ genera una variabile aleatoria: $S = h(X_1, \dots, X_j, \dots, X_n)$. Applicata ai valori $(x_1, \dots, x_j, \dots, x_n)$ del campione, la funzione $h(\cdot)$ assume un valore numerico: $s = h(x_1, \dots, x_j, \dots, x_n)$.

La distribuzione di probabilità della statistica S , coinvolgendo il campione, prende il nome di *distribuzione campionaria*. Essa dipende dalla distribuzione di probabilità del campione e quindi, come quest'ultima, dal modello descrittivo della popolazione.

Da quanto detto, emerge che, in questo artefatto virtuale che è l'universo dei campioni, esistono numerosissime variabili aleatorie, tra queste verranno prese in esame quelle più frequentemente usate nell'inferenza statistica.

7.3.1 | Media aritmetica campionaria

In generale, dato un carattere descritto da una variabile continua X , osservato in campioni via via più numerosi, tratti da una popolazione indefinitamente grande, la distribuzione campionaria della variabile media aritmetica tende ad assumere un andamento gaussiano con media $M(\bar{X}) = \mu$ e varianza $V(\bar{X}) = \sigma^2/n$, dove μ , e σ^2 sono, rispettivamente, la media e la varianza della popolazione. È, questa, una conseguenza del "teorema centrale del limite".

Il teorema centrale del limite implica che:

- data una popolazione in cui il carattere X ha media μ e varianza finita σ^2 ;
- estratto da questa un campione casuale di numerosità n ;

Media aritmetica campionaria - è la variabile descritta dalle medie aritmetiche calcolate nei campioni appartenenti allo spazio campionario. È così definita: $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$

la distribuzione *asintotica* della v.a. media aritmetica campionaria è normale, con:

- media μ pari alla media di popolazione;
- varianza pari a $1/n$ -esimo della varianza σ^2 di popolazione (e scarto quadratico medio pari a $1/\sqrt{n}$ dello scarto σ):

$$\bar{X} \underset{n \rightarrow \infty}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Se nella popolazione il carattere X è normodistribuito, allora la variabile \bar{X} - combinazione lineare di normali - è normodistribuita come proprietà finita, valida per qualunque numerosità campionaria n .

□—————

◇ **Osservazione 7.3.** Se il campionamento viene effettuato in blocco da popolazione finita, la varianza di \bar{X} risente del fattore di esaustività $\frac{N-n}{N-1}$.

	Campionamento bernoulliano	Campionamento in blocco
Popolazione non finita	$E(\bar{X}) = \mu$ $\sqrt{V(\bar{X})} = \sigma/\sqrt{n}$	
Popolazione finita	$E(\bar{X}) = \mu$ $\sqrt{V(\bar{X})} = \sigma/\sqrt{n}$	$E(\bar{X}) = \mu$ $\sqrt{V(\bar{X})} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

◇ **Lineamenti teorici generali.**

Data una popolazione P definita rispetto a un carattere variabile X , con parametri $M(X)=\mu$ e $V(X)=\sigma^2$, un campione bernoulliano di n elementi:

$$x_1, \dots, x_i, \dots, x_n$$

può essere inteso come una determinazione empirica della n -pla di variabili aleatorie:

$$X_1, \dots, X_i, \dots, X_n$$

reciprocamente indipendenti e ugualmente distribuite. La generica variabile X_i descrive infatti l'insieme dei valori che possono avverarsi quando si assume l' i -esimo elemento campionario, mentre x_i è il valore effettivamente avvertosi. Ogni variabile X_i coincide quindi con la variabile X nella popolazione. E dunque:

$$X_1 = \dots = X_i = \dots = X_n = X.$$

Ed è pure (per ogni i):

$$\begin{aligned} M(X_i) &= M(X) = \mu \\ V(X_i) &= V(X) = \sigma^2. \end{aligned}$$

La somma

$$S_n = X_1 + \dots + X_i + \dots + X_n$$

è una variabile aleatoria con media uguale alla somma delle medie:

$$M(S_n) = \sum_i M(X_i) = n\mu$$

e varianza uguale alla somma delle varianze:

$$V(S_n) = \sum_i V(X_i) = n\sigma^2.$$

Per il “teorema centrale del limite”, la variabile aleatoria standardizzata

$$Z_n = \frac{S_n - M(S_n)}{\sqrt{V(S_n)}} = \frac{S_n - n\mu}{\sigma/\sqrt{n}} \quad (7.1)$$

tende asintoticamente a distribuirsi, per $n \rightarrow \infty$, come la normale standardizzata $N(0, 1)$. Si consideri ora la variabile “media aritmetica campionaria”:

$$\frac{S_n}{n} = \frac{X_1 + \dots + X_i + \dots + X_n}{n} = \bar{X} \quad (7.2)$$

Essa ha media

$$M(\bar{X}) = M\left(\frac{S_n}{n}\right) = \frac{1}{n}M(S_n) = \frac{1}{n}n\mu = \mu \quad (7.3)$$

e varianza:

$$V(\bar{X}) = V\left(\frac{S_n}{n}\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n} \quad (7.4)$$

La media aritmetica $M(\bar{X})$ di tutte le possibili medie campionarie è dunque uguale alla media della popolazione.

La varianza $V(\bar{X})$ è l'ennesima parte della varianza $V(X)$: la dispersione delle medie aritmetiche campionarie attorno al loro baricentro $M(\bar{X}) = \mu$, è dunque tanto più ridotta rispetto a quella della popolazione quanto maggiore è la dimensione del campione.

La \bar{X} standardizzata è:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{(S_n/n) - \mu}{\sigma/\sqrt{n}} = Z_n. \quad (7.5)$$

Poiché la Z_n tende alla $N(0, 1)$, la variabile “media aritmetica campionaria” \bar{X} è asintoticamente, per $n \rightarrow \infty$, una distribuzione normale con media $M(\bar{X}) = \mu$, varianza $V(\bar{X}) = \sigma^2/n$. (Si conviene che lo scarto quadratico medio sia $\sqrt{V(\bar{X})} = \sigma/\sqrt{n}$).

La rapidità con cui la distribuzione campionaria delle medie tende alla “gaussiana” al crescere di n dipende dalla distribuzione del carattere X nella popolazione. Quanto più è asimmetrica, tanto più n deve essere grande perchè si realizzi la convergenza. Per questo motivo molti metodi di stima dei parametri incogniti di una popolazione o di controllo di ipotesi sui parametri stessi impongono la condizione di normalità distributiva del carattere nella popolazione.

◇ **Osservazione 7.4.**

Data una popolazione Bernoulliana, in cui $X \sim Ber(p)$, avremmo che il campione ottenuto è in genere una sequenza di 1 e 0.

La v.a. Bernoulliana assume infatti valore 1 “successo”, con probabilità p e valore 0 “insuccesso”, con probabilità $1 - p$, e la sua funzione di probabilità è $f(x) = p^x(1 - p)^{1-x}$, $x = 0, 1$.

In tal caso la media di popolazione è

$$\mu = 1 \cdot p + 0 \cdot (1 - p) = p$$

e la varianza è

$$\sigma^2 = 1^2 \cdot p + 0^2 \cdot (1 - p) - p^2 = p(1 - p)$$

di conseguenza la v.a. media aritmetica campionaria avrà: $E(\bar{X}) = p$ e $Var(\bar{X}) = p(1 - p)/n$.

Si può inoltre affermare che la distribuzione della media campionaria nel caso di una popolazione Bernoulliana è proporzionale alla binomiale,

$$\bar{X} \sim \frac{1}{n} Bin(n, p)$$

infatti:

$$P(\bar{X} = \bar{x}) = P(n\bar{X} = n\bar{x}) = \binom{n}{n\bar{x}} p^{n\bar{x}} (1 - p)^{n - n\bar{x}}$$

dove $n\bar{X}$ è la variabile aleatoria “numero di successi”.

Teoricamente, è sempre possibile determinare la distribuzione di probabilità di \bar{X} a partire da un dato modello di popolazione; tuttavia è interessante osservare che la distribuzione campionaria della media tende ad assumere una forma sempre più stabile e indipendente dal modello descrittivo della popolazione, via via che l'ampiezza del campione aumenta, convergendo alle soglie della distribuzione normale.

Tale risultato trova la sua giustificazione nel teorema centrale del limite.

—————□

7.3.2 | Frequenza relativa campionaria

Le relazioni addotte per i caratteri misurabili si possono estendere alle distribuzioni per attributi. Indicate con:

- f la frequenza di un certo attributo in un generico campione bernoulliano di ampiezza n ,
- p la frequenza dello stesso attributo nella popolazione,
- Fr la v.a. frequenza relativa campionaria,

per il teorema centrale del limite la v.a. Fr è *asintoticamente* normo-distribuita, con media uguale alla frequenza p in popolazione e varianza uguale a $p(1 - p)/n$:

$$Fr \underset{n \rightarrow \infty}{\sim} N \left(p, \sqrt{\frac{p(1 - p)}{n}} \right). \tag{7.6}$$

Frequenza relativa campionaria - è la variabile aleatoria descritta dalle frequenze relative nello spazio dei campioni.

□—————

◇ **Osservazione 7.5.** Se il campionamento viene effettuato in blocco da popolazione finita, la variabilità di Fr è modificata dal fattore di esaustività $\frac{N-n}{N-1}$:

	Campionamento bernoulliano	Campionamento in blocco
Popolazione non finita	$E(Fr) = p$ $\sqrt{V(Fr)} = \sqrt{p(1 - p)/n}$	
Popolazione finita	$E(\bar{X}) = \mu$ $\sqrt{V(FR)} = \sqrt{p(1 - p)/n}$	$E(\bar{X}) = \mu$ $\sqrt{V(Fr)} = \sqrt{\frac{p(1-p)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

—————□

7.3.3 | Varianza campionaria

Indicando con \hat{s}^2 le singole varianze campionarie, la distribuzione della variabile \hat{S}^2 , descritta da tutte le varianze dei campioni, ha media non coincidente con quella della popolazione. Infatti è $M(\hat{S}^2) < \sigma^2$.

Moltiplicando la varianza per il fattore correttivo $n/(n-1)$, così da porre a denominatore il numero $n-1$, si ottiene:

$$s^2 = \hat{s}^2 \frac{n}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n} \cdot \frac{n}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad (7.7)$$

che genera una nuova variabile aleatoria S^2 : la “varianza campionaria corretta”. Essa ha media uguale a σ^2 :

$$M(S^2) = M\left(\hat{S}^2 \frac{n}{n-1}\right) = M(\hat{S}^2) \frac{n}{n-1} = \sigma^2. \quad (7.8)$$

Varianza campionaria - è la variabile aleatoria descritta dalle varianze nell'universo dei campioni. Indicata con S^2 la varianza campionaria “corretta”: $S^2 = \frac{\sum_{j=1}^n (X_j - \bar{X})^2}{n-1}$

Il fattore di correzione $n/(n-1)$ - dovuto a Bessel - vale per la media della variabile descritta da tutte le varianze campionarie. Non di meno, la “correzione” viene apportata alla varianza del singolo campione, ignorando se veramente abbia per effetto di renderla più prossima o meno prossima al parametro da stimare.

□—————

◇ **Osservazione 7.6.** Come per qualsiasi altra statistica, la distribuzione della v.a. varianza campionaria corretta S^2 dipende dal modello distributivo della popolazione, tuttavia è possibile determinare il valore atteso e la varianza in termini generali. Si dimostra infatti che:

$$E(S^2) = \sigma^2, \quad V(S^2) = \frac{\sigma^4}{n} \left(\beta_2 + 2 \frac{n}{n-1} \right)$$

dove $\beta_2 = \frac{\bar{\mu}_4}{\sigma^4} - 3$ è l'indice di disnormalità della popolazione, essendo $\bar{\mu}_4$ il quarto momento centrale.

Nell'ipotesi che il modello descrittivo della popolazione sia *normale*, è possibile determinare la distribuzione di probabilità di S^2 , pertanto:

- data una popolazione in cui il carattere $X \sim N$ con varianza σ^2 ;
- estratto da questa un campione casuale di numerosità n ;
- la varianza campionaria S^2 ha distribuzione di probabilità proporzionale a quella di una χ^2 con $n - 1$ gradi di libertà; sia infatti

$$S^2 \sim \frac{\sigma^2}{n-1} \chi_{n-1}^2$$

la v.a. campionaria $(n-1)S^2/\sigma^2$ segue una distribuzione χ^2 con $\nu = n - 1$ gradi di libertà:

$$\frac{(n-1)}{\sigma^2} S^2 \sim \chi_{n-1}^2.$$

Essendo $E[\chi_{n-1}^2] = n - 1$, anche

$$E \left[\frac{(n-1)}{\sigma^2} S^2 \right] = \frac{(n-1)}{\sigma^2} E[S^2] = n - 1$$

da questa eguaglianza si ricava che

$$E(S^2) = \sigma^2$$

Inoltre, essendo $Var[\chi_{n-1}^2] = 2(n-1)$, risulta

$$Var \left[\frac{(n-1)}{\sigma^2} S^2 \right] = \frac{(n-1)^2}{\sigma^4} Var[S^2] = 2(n-1)$$

da cui si ricava:

$$V(S^2) = \frac{2\sigma^4}{n-1}.$$

—————□

7.4 | Funzione di verosimiglianza

Dato un modello descrittivo di riferimento della popolazione $f(x, \theta)$ con $\theta \in \Theta$, osservata una realizzazione casuale $(x_1, \dots, x_j, \dots, x_n)$ della n -pla di variabili aleatorie $(X_1, \dots, X_j, \dots, X_n)$ distribuite secondo tale modello; se $f(x_j, \theta)$, misura la probabilità che l'elemento x_j entri nel campione, allora la funzione di probabilità congiunta (se la variabile X è discreta) o la densità congiunta (se la variabile X è continua) della n -pla di variabili campionarie che dipende dall'incognito parametro, si dice “funzione di verosimiglianza”:

$$\mathcal{L}(\theta) = \mathcal{L}(\theta; x_1, \dots, x_j, \dots, x_n). \quad (7.9)$$

La funzione di verosimiglianza è indicata con \mathcal{L} , che è l'iniziale di Likelihood, ed è la funzione di (densità di) probabilità congiunta del campione assunta come funzione del parametro θ . Nella (7.9) il parametro θ , che può essere sia uno scalare che un vettore, è la variabile e le x_j , sono le costanti. Si può quindi dire che la (7.9) trovi il suo insieme di definizione nello spazio parametrico Θ ed associ ad ogni valore $\theta \in \Theta$ un livello di (densità di) probabilità al campione $(x_1, \dots, x_j, \dots, x_n)$. Indica pertanto come variano le probabilità di uno stesso campione al variare dei valori del parametro (avendo come dominio lo spazio parametrico).¹

La funzione di verosimiglianza $\mathcal{L}(\theta)$, assume forme differenti a seconda del criterio di campionamento adottato:

- nel caso più semplice in cui le variabili $(X_1, \dots, X_j, \dots, X_n)$, siano i.i.d. (campionamento probabilistico a probabilità costante)²:

$$\mathcal{L}(\theta) = L(\theta; x_1, \dots, x_j, \dots, x_n) = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \dots \cdot f(x_n, \theta) \propto \prod_{j=1}^n f(x_j; \theta)$$

- se le variabili $(X_1, \dots, X_j, \dots, X_n)$, sono reciprocamente indipendenti, ma non identicamente distribuite,

¹Ad esempio per una Bernoulli lo spazio parametrico è $0 < p < 1$.

²La relazione di proporzionalità può essere sostituita da una relazione *operativa* di uguaglianza, poiché, ai fini induttivi, funzioni di verosimiglianza tra loro proporzionali possono considerarsi equivalenti

$$\mathcal{L}(\theta) = \prod_{j=1}^n f_j(x_j; \theta),$$

- mentre se le variabili $(X_1, \dots, X_j, \dots, X_n)$, non sono né reciprocamente indipendenti, né identicamente distribuite

$$\mathcal{L}(\theta) = f_1(x_1; \theta) \cdot \prod_{j=1}^n f_{j|1,2,\dots,j-1}(x_j|x_1, \dots, x_{j-1}; \theta).$$

La funzione di verosimiglianza riveste un ruolo particolarmente importante nell'ambito della stima, essendo alla base del metodo della “*massima verosimiglianza*” introdotto da R.A Fisher.

7.5 | Esercizi

- 7.1** Quanti campioni bernoulliani e quanti campioni in blocco di numerosità $n=2$ è possibile estrarre da una popolazione finita di numerosità:
- $N=7$;
 - $N=21$;
 - $N=43$?
- 7.2** Quanti campioni bernoulliani e quanti campioni in blocco di numerosità $n=3$ è possibile estrarre da una popolazione finita di numerosità:
- $N=4$;
 - $N=24$;
 - $N=46$?
- 7.3** Si consideri una popolazione costituita da $N=4$ ipermercati A, B, C, D. Si determinino e si confrontino tra loro:
- l'universo dei campioni bernoulliani di dimensione $n=2$;
 - l'universo dei campioni in blocco del medesimo ordine.

- 7.4** I 5 dipendenti di una piccola azienda manifatturiera - Elio, Angelo, Dario, Stefano e Mauro - decidono di estrarre a sorte i 2 di loro che dovranno restare al lavoro nella settimana di ferragosto. Fra quante e quali coppie avverrà il sorteggio?
- 7.5** Uno studio legale composto da 5 soci - A, B, C, D, E - vince un soggiorno-vacanza per 3 persone. I soci decidono, di comune accordo, di estrarre a sorte i tre di loro che potranno godere del premio. Quante e quali sono le terne che parteciperanno al sorteggio?
- 7.6** Si consideri la popolazione costituita dai quattro ipermercati A, B, C, D. Le vendite effettuate da ciascuno di essi nel periodo 1/1/06 - 31/12/06 sono riportate nella seguente tabella:

Ipermercato	A	B	C	D
Vendite (in milioni di euro)	4	1	3	2

- a) Si calcolino la media e lo scarto quadratico medio di popolazione;
- b) effettuando un campionamento bernoulliano si calcolino:
- 1) la media aritmetica in ciascuno dei campioni di numerosità $n=2$ estraibili dalla popolazione;
 - 2) la media aritmetica (valore atteso) e lo scarto quadratico medio della v.a. media aritmetica campionaria confrontandoli, rispettivamente, con la media e con lo scarto quadratico medio di popolazione;
- c) effettuando un campionamento in blocco si calcolino:
- 1) la media aritmetica in ciascuno dei campioni di numerosità $n=2$ estraibili dalla popolazione;
 - 2) la media aritmetica (valore atteso) e lo scarto quadratico medio della v.a. media aritmetica campionaria confrontandoli, rispettivamente, con la media e con lo scarto quadratico medio di popolazione.

- 7.7 Presso l'ufficio studi di una grande banca è noto che l'80% dei prelievi effettuati tramite bancomat riguarda importi inferiori alle trecentomila lire. Estrahendo campioni casuali di 200 prelievi, in quale proporzione di campioni la percentuale di prelievi inferiori alle trecentomila lire è compresa fra il 70% e l'80% ?
- 7.8 Il 9% delle lenti da vista prodotte da una certa azienda presenta delle imperfezioni. Estrahendo dalla produzione campioni casuali di 300 lenti, in quale proporzione i campioni presenteranno una frequenza di pezzi difettosi:
- compresa fra l'8% e il 9%;
 - minore o uguale al 7%?
- 7.9 Un negozio di articoli sportivi riceve da un fornitore una partita di $n = 20$ boomerang. Secondo le dichiarazioni dell'azienda costruttrice, per quel tipo di boomerang i valori della precisione di ritorno (intesa come distanza del punto di arrivo dal punto di lancio dopo un giro completo) si distribuiscono normalmente con uno scarto pari a 1 metro. Si calcoli:
- la probabilità che la varianza della precisione nella partita ricevuta sia inferiore a $0,55 \text{ m}^2$;
 - la media e lo scarto quadratico medio della v.a. varianza campionaria corretta.
- 7.10 Il tempo (in minuti) necessario a un certo impiegato dell'anagrafe per disbrigare le pratiche ha una media aritmetica $\mu = 20$ e uno scarto quadratico medio $\sigma = 5$. Se si estrae un campione casuale di 50 pratiche d'ufficio:
- qual la probabilità che il tempo impiegato per ciascuna pratica sia, in media, non superiore ai 18 minuti?
 - qual la probabilità che il tempo impiegato per ciascuna pratica sia, in media, superiore ai 22 minuti?
 - qual il numero di minuti al di sotto del quale la media aritmetica del campione si colloca con una probabilità pari al 95%?

Il problema della stima

8.1 | Introduzione

Dei campioni ci si può avvalere per stimare le proprietà statistiche di popolazioni di cui costituiscono sottoinsiemi casuali, o per saggiare l'attendibilità di ipotesi circa le popolazioni stesse.

Il riferimento teorico è sempre all'“universo dei campioni”, vale a dire alle *distribuzioni campionarie* dei parametri della popolazione.

In una popolazione finita P di numerosità N , sia $\mathcal{L}(X; \theta)$ la legge distributiva di un carattere quantitativo rappresentato dalla variabile X , con parametro θ .

Sia C un campione casuale di n unità tratto dalla popolazione P . Tale campione è un elemento dello spazio composto da tutti i campioni di ampiezza n compatibili con la procedura di campionamento adottata.

Essendo P composta da N unità statistiche, l'universo dei campioni è anch'esso finito.

L'induzione circa le proprietà statistiche di un insieme – “popolazione” - sulla base di un suo sottoinsieme – “campione” - può attenere alla stima del valore incognito di un parametro: *stima puntuale*; o degli stati di grandezza fra i quali il parametro può, con assegnata probabilità, essere compreso: *stima intervallare*.

Può altresì riguardare il grado di attendibilità di una *ipotesi* su qualche proprietà statistica della popolazione.

8.2 | Stimatori e stima puntuale

Sia X la variabile che descrive gli stati di grandezza di un carattere quantitativo in una popolazione P . Sia θ un parametro di tale variabile in P . Sia (x_1, x_2, \dots, x_n) una n -pla di valori costituenti un sottoinsieme-campione di P . Se la popolazione è finita, si ammetta che il campione è bernoulliano. (Se la popolazione può ritenersi non finita, questa condizione è irrilevante).

La n -pla campione può allora essere riguardata come la determinazione empirica di una n -pla di variabili aleatorie campionarie eguali e indipendenti X_1, X_2, \dots, X_n , con funzione di densità $f(X, \theta)$. La stima puntuale di θ impone di definire un algoritmo sulle variabili campionarie, $g(X_1, X_2, \dots, X_n) = T$, da determinare sui dati del campione: $g(x_1, x_2, \dots, x_n) = t$.

La $g(X_1, X_2, \dots, X_n) = T$ si dice *stimatore puntuale di θ* ; T è la variabile aleatoria che descrive la “statistica” t nell’insieme virtuale di tutti i possibili campioni compatibili col criterio di campionamento. Un singolo valore t è una “stima puntuale” del “parametro” θ .

Assunto un campione (x_1, x_2, \dots, x_n) , la stima di θ tratta dalla n -pla dei valori è uno stato di grandezza dalla variabile descritta, nello spazio campionario, dalla t .

8.2.1 | Proprietà degli estimatori

Le proprietà che si richiedono a uno stimatore sono essenzialmente tre: la correttezza, l’efficienza, la consistenza.

Correttezza. Uno stimatore si dice *corretto* (nel senso di “centrato”) se il suo valore medio coincide col parametro θ della popolazione:

$$M(t) = M[g(X_1, X_2, \dots, X_n)] = \theta \quad (8.1)$$

ossia

$$M(t - \theta) = 0.$$

Si dice allora che il valore atteso¹ della statistica nel campione è $E(t) = \theta$.

Ad esempio, lo “stimatore media aritmetica” è corretto in quanto la sua distribuzione campionaria (tendenzialmente gaussiana) ha per baricentro μ . Infatti, per la (8.1):

$$M(\bar{X}) = \frac{1}{n}[M(X_1) + \dots + M(X_n)] = \frac{1}{n}n\mu = \mu \quad (8.2)$$

La media aritmetica delle medie campionarie è uguale alla media aritmetica della popolazione.

In ragione di questa identificazione tra media della popolazione e media delle medie di tutti i possibili campioni, si dice che il valore atteso della media \bar{x} di un campione è la media della popolazione.

In simboli: $E(\bar{x}) = \mu$.

Se tale uguaglianza non si avvera, lo stimatore si dice *distorto*.

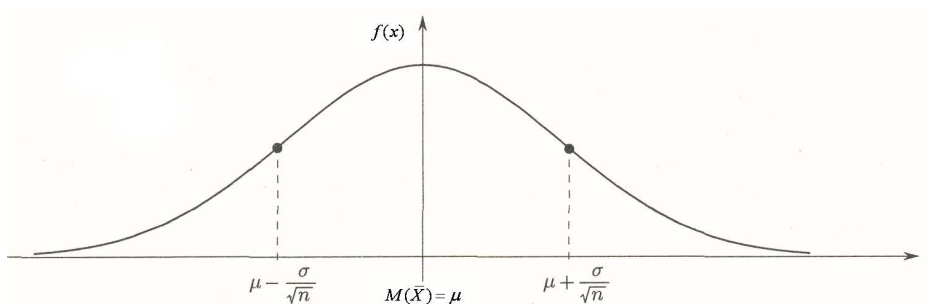


fig.8.1

Si debba stimare il parametro σ^2 , ossia la varianza della popolazione.

Si consideri allora la variabile “varianza campionaria”, vale a dire la variabile descritta, nello spazio dei campioni, da tutte le varianze $s_1^2, s_2^2, \dots, s_i^2, \dots$ dei possibili campioni. La media di tale variabile è:

¹Il simbolo E , introdotto da Pascal per indicare l'*Expérance mathématique*, è pure l'iniziale della parola *Expectation*.

$$\begin{aligned} M(\hat{S}^2) &= M(X_i - \bar{X})^2 = M(X_i - \mu + \mu - \bar{X})^2 = \\ &= M(X_i - \mu)^2 - M(\bar{X} - \mu)^2 = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2 \frac{n-1}{n} \end{aligned} \quad (8.3)$$

Il valore atteso $E[(x_i - \bar{x})^2/n]$ non è dunque uguale a σ^2 , bensì inferiore. È infatti $M(\hat{S}^2) < \sigma^2$. La varianza campionaria calcolata come $\sum (x_i - \bar{x})^2/n$ si definisce allora stimatore distorto di σ^2 . Stimatore corretto di σ^2 è, per quanto premesso, il prodotto della media delle varianze campionarie per il fattore di correzione $(n-1)/n$. In conseguenza di ciò, si dice “varianza corretta” la varianza di un campione calcolata ponendo a denominatore il numero $n-1$:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n}. \quad (8.4)$$

Il numero $n-1$ è il “numero dei gradi di libertà”, e identifica il numero dei termini indipendenti: quelli, tra gli n del campione, che sono liberi da vincoli.

Tale correzione vale per la media delle varianze campionarie.

Si adotta così, per la varianza di un campione, s^2 , il fattore di correzione che vale per la varianza campionaria media: \bar{s}^2 .

Efficienza. Di due stimatori corretti T_1 e T_2 si dice più *efficiente* quello con varianza minore.

Ad esempio: media aritmetica campionaria e mediana campionaria sono entrambe stime corrette di μ . ma la varianza della distribuzione campionaria della media (che è σ^2/n) è minore della varianza della distribuzione campionaria della mediana (che è $\pi\sigma^2/n$).

Consistenza. Uno stimatore si dice *consistente* (nel senso di “convergente”) quando vale il limite:

$$\lim_{n \rightarrow \infty} \Pr \{|T - \theta| < \varepsilon\} = 1$$

ossia quando, all’aumentare indefinito del numero n dei casi osservati, tende alla certezza la probabilità che la differenza tra lo stimatore e il parametro stimato si mantenga inferiore a un numero ε piccolo quanto si vuole.

Ad esempio: la varianza campionaria e la varianza campionaria corretta sono stimatori consistenti di σ^2 .

8.3 | Metodi di stima

Per individuare stimatori con il maggior numero di requisiti richiesti, soccorrono diversi principi induttivi. I più noti, e più usati, sono: il “metodo dei mo-menti” , il “metodo della massima verosimiglianza” , il “metodo bayesiano” (e, per particolari costanti statistiche, il “metodo dei minimi quadrati”).

Metodo dei momenti (K. Pearson). Consiste nel sostituire ai parametri incogniti da stimare le corrispondenti statistiche calcolate nel campione. È un primo modo di rispondere al problema teorico di una metodologia statistica induttiva tesa a generalizzare le proprietà statistiche accertate su una n -pla di casi.

La media aritmetica μ , stimata con la media del campione \bar{x} , la varianza σ^2 con la varianza del campione s^2 , e così via. Questo metodo, nei limiti delle proprietà che lo distinguono, può offrire efficaci e rapide soluzioni.

Metodo della massima verosimiglianza (R.A. Fisher). Consiste nel massimizzare l’informazione campionaria assumendo come stima del parametro di un dato modello distributivo la costante del campione che ha la massima probabilità di avverarsi.

Formalmente, dato un modello distributivo $f(x;\theta)$ e una ennu-
pla campionaria x_1, x_2, \dots, x_n , se $f(x_j;\theta)$ misura la probabi-
lità dell’elemento x_j , di entrare nel campione, la funzione di densità
congiunta

$$L(X; \theta) = \mathcal{L}(x_1, \dots, x_n; \theta)$$

è detta funzione di verosimiglianza.

In un campionamento probabilistico a probabilità costante, si ha, per l’indipendenza degli eventi (principio del prodotto):

$$L(X; \theta) = L(x_1, \dots, x_n; \theta) = f(x_1; \theta) f(x_2; \theta) \dots f(x_n; \theta) = \prod f(x_i; \theta). \quad (8.5)$$

Il valore di θ che massimizza la funzione di verosimiglianza $\mathcal{L}(x_1, \dots, x_n; \theta)$ si dice “stimatore di massima verosimiglianza”. Se il modello distribu-
tivo è funzione di più parametri

$$L(X; \theta_1, \dots, \theta_k) = \mathcal{L}(x_1, \dots, x_n; \theta_1, \dots, \theta_k) = \prod f(x_i; \theta_1, \dots, \theta_k),$$

massimizzare tale funzione significa risolvere un sistema di k equazioni del tipo

$$\left\{ \begin{array}{l} \frac{\partial L(X; \theta_1, \dots, \theta_k)}{\partial \theta_1} = 0 \\ \dots\dots\dots \\ \frac{\partial L(X; \theta_1, \dots, \theta_k)}{\partial \theta_j} = 0 \\ \dots\dots\dots \\ \frac{\partial L(X; \theta_1, \dots, \theta_k)}{\partial \theta_k} = 0 \end{array} \right.$$

Gli stimatori di massima verosimiglianza sono asintoticamente normali, efficienti e consistenti, e pertanto sono spesso preferiti quando il campione non sia molto piccolo.

Ad esempio, dato un campione casuale semplice x_1, \dots, x_n estratto da una popolazione in cui il carattere X si distribuisce secondo una $\mathcal{N}(\mu, \sigma)$:

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

il metodo consiste nel massimizzare la funzione di verosimiglianza ottenuta dal prodotto

$$\begin{aligned} f(X; \mu, \sigma) &= f(x_1; \mu, \sigma) f(x_2; \mu, \sigma) \dots f(x_n; \mu, \sigma) = \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\frac{\sum (x_i - \mu)^2}{2\sigma^2}} \end{aligned}$$

Sulla trasformata logaritmica, si cercano i valori dei parametri che rendono massima la funzione:

$$L(X; \mu, \sigma) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \sigma^2 - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$$

Si deve quindi risolvere, per μ , e per σ , il sistema:

$$\begin{cases} \frac{\partial L(X; \mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum (x_i - \mu) = 0 \\ \frac{\partial L(X; \mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (x_i - \mu)^2 = 0 \end{cases}$$

La soluzione è:

$$\mu = \frac{1}{n} \sum x_i = \bar{x} \text{ e } \sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = \hat{s}^2$$

Quando la distribuzione è normale, dunque, gli stimatori di massima verosimiglianza di μ e σ coincidono con gli stimatori ottenuti con il metodo dei momenti.

Dato un campione casuale semplice x_1, \dots, x_n estratto da una popolazione in cui il carattere X si distribuisce secondo un modello bernoulliano $f(X, p) = p^x(1-p)^{1-x}$, dove ogni x_i può assumere solo i valori 0 e 1, indicatori dell'assenza o della presenza di un'assegnata modalità A (il campione esprime la frequenza relativa f_i dei casi portatori della modalità A), la "funzione di verosimiglianza" è il prodotto

$$\begin{aligned} L(x, p) &= f(x_1; p)f(x_2; p)\dots f(x_n; p) = \prod_i f(x_i; p) \\ &= \prod p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i} \end{aligned}$$

Posta la trasformata logaritmica

$$l(X, p) = \sum x_i \log p + \left(n - \sum x_i\right) \log(1-p),$$

uguagliando a zero la derivata rispetto a p (con $0 < p < 1$):

$$\frac{\partial l(X, p)}{\partial p} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = 0$$

si ottiene:

$$p = \frac{\sum x_i}{n}.$$

La frazione $\sum x_i/n$ è la frequenza relativa f_A dei portatori di A nel campione osservato. Pertanto, la frequenza relativa f_A della modalità di un carattere distinto in due sole modalità alternative, A e $non-A$, è stimatore di massima verosimiglianza della probabilità $Pr\{A\}$.

Criterio bayesiano. Si propone come alternativa al criterio del trarre tutta l'informazione dal solo contesto empirico campionario, vale a dire dalla sola probabilità di accadimento accidentale dei particolari risultati (la “verosimiglianza”), e si ispira al principio dell'integrare tale informazione assegnando gradi di attendibilità a priori ai possibili valori del parametro incognito da stimare.

Dal punto di vista teoretico-induttivo, l'impostazione bayesiana della teoria della stima è la più completa e la più coerente, soprattutto nella fattispecie dei piccoli campioni. Tuttavia, al crescere del numero dei casi su cui è costruita la verosimiglianza, s'attenua l'importanza delle probabilità *a priori*. (Queste, per altro, non sono sempre facilmente determinabili nelle esperienze concrete delle procedure di stima e non sempre riducibili a canoni intersoggettivi) .

È vero che in termini di attesa soggettiva può sempre essere adottata una distribuzione *a priori* dei possibili valori (la quale può avvicinare al “bersaglio” così come può peggiorare la “mira”); ma è anche vero che in tal modo si introducono quelle componenti di giudizio non codificabili che la teoria dei campioni, nel pensiero dei fondatori (di Neyman, soprattutto), aveva espressamente inteso escludere dai principi del campionamento.

8.4 | Stima per intervalli

8.4.1 | Intervalli di confidenza della media

Le medie aritmetiche di tutti i campioni appartenenti a un medesimo universo sono stati di grandezza della variabile aleatoria “media aritmetica campionaria”: essa esprime una variabilità dovuta all'errore *di campionamento* . Per quanto premesso, la distribuzione campionaria di tale variabile è asintoticamente “normale”, con valor medio aritmetico coincidente con la media aritmetica della popolazione di riferimento $M(X) = \mu$, e pertanto il “valore atteso” di una media campionaria è $E(\bar{x}) = \mu$. La varianza della variabile “media aritmetica campionaria”, in un campione di n elementi, è uguale a un ennesimo della varianza della popolazione (σ^2)

$$V(\bar{X}) = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \quad (8.6)$$

Nella radice quadrata di tale varianza viene identificato lo scarto quadratico medio della variabile “media aritmetica campionaria”:

$$\sqrt{V(\bar{X})} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Esso dipende, quindi, dalla variabilità del carattere nella popolazione e dalla dimensione del campione.

Assumendo la media di un campione come stima della media incognita della popolazione, la precisione di tale stima tende dunque a migliorare all'aumentare del numero delle unità che costituiscono il campione. Quando la varianza σ^2 della popolazione è, come avviene in generale, ignota, la si deve a sua volta stimare. La varianza del campione s^2 è in genere l'unica informazione di cui si dispone. Quando tale varianza sia calcolata riferendo la devianza al numero dei gradi di libertà, essa può assumere, in ragione dell'accennata relazione valida in media, il ruolo di stima corretta dell'incognita σ^2 .

La stima dell'errore della media campionaria diventa allora $s_{\bar{x}} = s/\sqrt{n}$. Ciò introduce un ulteriore elemento di incertezza; al variare dei campioni, infatti, anche lo scarto quadratico medio assume valori diversi e configura, a sua volta, una variabile aleatoria S con una propria distribuzione campionaria.

Poichè la “variabile campionaria media aritmetica” \bar{X} tende a distribuirsi come una normale con $M(\bar{X}) = \mu$, e $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, dove μ e σ sono i parametri della popolazione di riferimento, la variabile standardizzata:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (8.7)$$

tende a distribuirsi come una normale con $M(\bar{X}) = 0$ e $V(\bar{X}) = 1$.

Noti i parametri μ e σ , i campioni del rispettivo universo campionario aventi media aritmetica compresa nell'intervallo definite dagli estremi $\mu - \frac{\sigma}{\sqrt{n}}$ e $\mu + \frac{\sigma}{\sqrt{n}}$ sono il 68,27 per cento; quelli compresi tra $\mu - 1,96\frac{\sigma}{\sqrt{n}}$ e $\mu + 1,96\frac{\sigma}{\sqrt{n}}$ sono il 95 per cento; quelli compresi tra $\mu - 2\frac{\sigma}{\sqrt{n}}$ e $\mu + 2\frac{\sigma}{\sqrt{n}}$ sono il 95,45 per cento; quelli compresi tra $\mu - 3\frac{\sigma}{\sqrt{n}}$ e $\mu + 3\frac{\sigma}{\sqrt{n}}$ sono il 99,73 per cento.

Si può allora definire la probabilità dell'avverarsi di un campione con media m compresa entro assegnati multipli dello scarto quadratico medio della variabile \bar{X} .

$$\Pr\left(\mu - k\frac{\sigma}{\sqrt{n}} \leq m \leq \mu + k\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad (8.8)$$

In particolare:

per $k = 1$ la probabilità è $1 - \alpha = 0,6827$

per $k = 1,96$ la probabilità è $1 - \alpha = 0,95$

per $k = 2$ la probabilità è $1 - \alpha = 0,9545$

per $k = 3$ la probabilità è $1 - \alpha = 0,9973$.

Questa è la soluzione del “problema diretto”: prevedere una proprietà statistica del campione, nota quella della popolazione. Ma l'induzione statistica attiene soprattutto al “problema inverso”: inferire una proprietà statistica della popolazione, nota quella di un campione. Nella stima per intervallo, ciò significa attribuire una probabilità all'ipotesi che l'incognito parametro sia compreso in un assegnato intorno della costante osservata nel campione. Così, se il parametro è la media aritmetica della popolazione ed è nota la media m di un campione di tale popolazione, la *stima intervallare* del parametro si traduce nell'attribuire una probabilità all'ipotesi che la media incognita μ , sia compresa tra gli estremi $m - k\frac{\sigma}{\sqrt{n}}$ e $m + k\frac{\sigma}{\sqrt{n}}$: una approssimazione tanto più attendibile quanto più il campione è numeroso.

Nella fattispecie inferenziale più comune anche il parametro σ è sconosciuto. Lo si sostituisce allora con lo scarto quadratico medio s del campione: un valore anch'esso non invariante da campione a campione e tuttavia tanto più atto a sostituire σ quanto più il campione è numeroso.

La variabile di riferimento non è più la (8.7), bensì la standardizzata

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \quad (8.9)$$

che, in generale, non si distribuisce normalmente. È unimodale simmetrica, ma con una variabilità più accentuata, rispetto alla distribuzione normale, dovuta alla sostituzione di σ con s .

La variabile standardizzata (8.9) è la variabile t di Student: essa descrive più duttilmente della normale l'andamento campionario delle medie calcolate in piccoli campioni e ha distribuzione di probabilità

$$f(t) = K \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad (8.10)$$

ove la costante $K = \Gamma\left(\frac{\nu+1}{2}\right) / \left[\Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi\nu}\right]$ è il fattore di normalizzazione, ossia il fattore per cui è

$$\int_{-\infty}^{+\infty} f(t) dt = 1$$

La $f(t)$ dipende dal numero $\nu = n - 1$ dei “gradi di libertà”.

L'integrale tra gli estremi $\pm t_\alpha$

$$\int_{-t_\alpha}^{+t_\alpha} f(t) dt$$

misura la probabilità di ottenere un valore di t compreso tra $-t_\alpha$ e $+t_\alpha$, mentre il complemento a 1

$$1 - \int_{-t_\alpha}^{+t_\alpha} f(t) dt = \int_{-\infty}^{-t_\alpha} f(t) dt + \int_{+t_\alpha}^{+\infty} f(t) dt \quad (8.11)$$

misura la probabilità di ottenere valori di t esterni alle soglie $-t_\alpha$ e $+t_\alpha$:

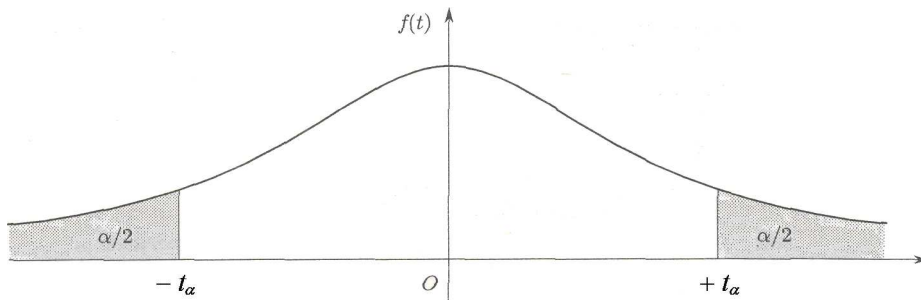


fig. 8.2

a questi valori di t , collocati nelle “code” della distribuzione, corrisponde una probabilità inferiore a un certo livello α : il grado di avverabilità dei valori estremi della variabile \bar{X} . I valori integrali della t , calcolati per diversi gradi di libertà, sono riportati nella tavola in Appendice.

Ponendo uguale ad α (numero reale da 0 a 1) la frazione di probabilità attribuita agli eventi estremi, sull'uno e sull'altro lato della distribuzione, restano indicati i “valori soglia” della variabile e precisamente: $-t_{\alpha/2}$ per gli eventi “minusvarianti” estremi e $+t_{\alpha/2}$ per gli eventi “plusvarianti” estremi (fig. 8.2). Il valore della probabilità α è scelto convenzionalmente. Nella pratica, si usa dividere la probabilità unitaria in due frazioni, solitamente fissate in: $1 - \alpha = 0,95$ e $\alpha = 0,05$. In particolari ricerche, la partizione può essere: $1 - \alpha = 0,99$ e $\alpha = 0,01$, o addirittura: $1 - \alpha = 0,999$ e $\alpha = 0,001$.

In generale, la probabilità di ottenere un valore di t compreso fra gli estremi $-t_\alpha$ e $+t_\alpha$ è:

$$\Pr \left(-t_\alpha \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq +t_\alpha \right) = 1 - \alpha \quad (8.12)$$

da cui:

$$\Pr \left(\bar{X} - t_\alpha \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_\alpha \frac{s}{\sqrt{n}} \right) = 1 - \alpha \quad (8.13)$$

La doppia disequazione

$$m - t_\alpha \frac{s}{\sqrt{n}} \leq \mu \leq m + t_\alpha \frac{s}{\sqrt{n}} \quad (8.14)$$

definisce un *intervallo di confidenza* della media.

All'ammissibilità di questa ipotesi circa μ viene così attribuito un “grado di fiducia” di valore $1 - \alpha$.

All'aumentare dei gradi di libertà, ossia all'aumentare della dimensione del campione, la $f(t)$ tende asintoticamente a coincidere con la normale standardizzata. Infatti, essendo stimatore consistente, al crescere di n la “varianza campionaria” s^2 tende a stabilizzarsi e a convergere al parametro della popolazione di riferimento, riconducendo così la t alla

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (8.15)$$

Si possono definire due livelli $-z_\alpha$ e $+z_\alpha$ così da escludere simmetricamente, nelle code sottese alla normale, i valori estremi, con probabilità inferiore a un livello convenzionale. Naturalmente questa procedura è adottata quando non si possa o non si voglia definire una probabilità a priori e non si possa o non si voglia, quindi, avvalersi dell'algoritmo bayesiano.

8.4.2 | Intervalli di confidenza della frequenza

Se la proprietà della popolazione da stimare è la frequenza relativa (probabilità statistica) p_A di una certa caratteristica ed è nota la frequenza relativa $f_A = n_A/n$ di tale caratteristica in un sottoinsieme campione (casuale), la stima intervallo si risolve nell'assegnare probabilità alla condizione

$$f_A - k\sqrt{\frac{f_A(1-f_A)}{n}} \leq p_A \leq f_A + k\sqrt{\frac{f_A(1-f_A)}{n}} \quad (8.16)$$

essendo la varianza $f_A(1-f_A)/n$ assunta a stima della $p_A(1-p_A)/n$, in genere sconosciuta.

All'aumentare di n , la frequenza relativa f_A tende a distribuirsi come una normale

$$\mathcal{N}\left(p_A, \sqrt{\frac{p_A(1-p_A)}{n}}\right)$$

e la variabile standardizzata

$$Z_A = \frac{f_A - p_A}{\sqrt{\frac{p_A(1-p_A)}{n}}} \quad (8.17)$$

tende a una $\mathcal{N}(0, 1)$.

Si può allora porre l'intervallo di confidenza, ad un assegnato livello α :

$$f_A - z_\alpha\sqrt{\frac{f_A(1-f_A)}{n}} \leq p_A \leq f_A + z_\alpha\sqrt{\frac{f_A(1-f_A)}{n}}. \quad (8.18)$$

8.4.3 | Intervalli di confidenza della varianza

Se in ogni campione si calcola la varianza corretta

$$s^2 = \frac{\sum (x_j - \bar{x})^2}{n - 1} \quad (8.19)$$

genera una variabile aleatoria

$$(n - 1) S^2 = \sum_j (x_j - \bar{x})^2$$

che tende a comporsi, se il carattere X nella popolazione è normalmente distribuito, come un particolare modello teorico: la funzione della variabile χ^2 . Infatti, la somma dei quadrati di n variabili aleatorie normali standardizzate:

$$\sum (X_i - \mu)^2 = \sigma^2 \quad (8.20)$$

converge a una variabile aleatoria “chi quadrato con $n - 1$ gradi di libertà.

In contesto campionario, il numeratore della (8.20) può essere sostituito con la variabile $(n - 1) S^2$.

La variabile

$$\frac{(n - 1) S^2}{\sigma^2}$$

si distribuisce come una $f(\chi_{n-1}^2)$ con $\nu = n - 1$ gradi di libertà:

$$f(\chi_\nu^2) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right) 2^{\nu/2}} e^{-x^2/2} (x^2)^{(\nu/2)-1} \quad (8.21)$$

ove la costante $K = 1/\Gamma\left(\frac{\nu}{2}\right) 2^{\nu/2}$ è il fattore di normalizzazione, ossia la quantità per cui è

$$\int_0^\infty f(\chi_\nu^2) d\chi^2 = 1 \quad (8.22)$$

La (8.21) è la funzione di densità di probabilità asimmetrica di una variabile che varia da zero all'infinito positivo e dipende dal numero dei

gradi di libertà. All'aumentare di questi, la distribuzione si fa sempre più simmetrica e sempre più conforme alla "normale".

La funzione integrale

$$\int_0^{\chi_\alpha^2} f(\chi_\nu^2) d\chi^2 = 1 - \alpha \quad (8.23)$$

misura la probabilità di un valore di χ^2 compreso tra 0 e χ_α^2 ; mentre il complemento a 1 della (8.23)

$$1 - \int_0^{\chi_\alpha^2} f(\chi_\nu^2) d\chi^2 = \int_{\chi_\alpha^2}^{+\infty} f(\chi_\nu^2) d\chi^2 = \alpha \quad (8.24)$$

misura la probabilità di un valore χ^2 esterno alla soglia χ_α^2 , (fig.8.3)

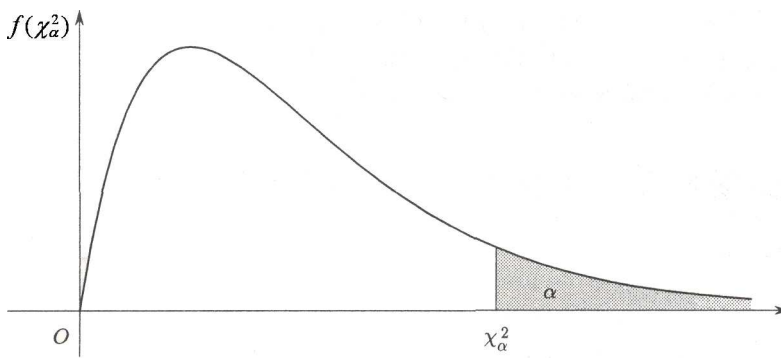


fig. 8.3

È pertanto possibile definire intervalli di confidenza della varianza, con assegnate probabilità.

$$\Pr \left\{ \chi_{\nu, 1-\alpha/2}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\nu, \alpha/2}^2 \right\} = 1 - \alpha \quad (8.25)$$

e quindi

$$\Pr \left\{ \frac{(n-1)S^2}{\chi_{\nu, \alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{\nu, 1-\alpha/2}^2} \right\} = 1 - \alpha \quad (8.26)$$

Sostituendo a S^2 la varianza corretta s^2 calcolata nel campione e alle espressioni $\chi_{\nu, \alpha/2}^2$ e $\chi_{\nu, 1-\alpha/2}^2$ i corrispondenti valori di $f(\chi_\nu^2)$, si ottengono gli estremi di un intervallo che si conviene comprenda l'incognito valore σ^2 con un livello di confidenza pari a $1 - \alpha$.

L'intervallo di confidenza della varianza, per un assegnato livello α , è allora:

$$\frac{(n-1)s^2}{\chi_{\nu, \alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{\nu, 1-\alpha/2}^2} \quad (8.27)$$

8.5 | La determinazione della numerosità campionaria

8.6 | Esercizi

Capitolo 9

Teoria dei test

9.1 | Introduzione

In generale per controllo di ipotesi si intende la proposizione di ipotesi circa l'incognita popolazione (relativamente ad una sua legge distributiva, un suo parametro ecc...) e il conseguente controllo di tale ipotesi.

Le procedure per il controllo statistico in campo parametrico, attono al controllo di ipotesi relative ad uno o più parametri della popolazione presupponendo la conoscenza della legge distributiva (modello) del carattere nella popolazione da cui il campione è estratto. Il controllo statistico delle ipotesi ammette una pluralità di procedure che differiscono principalmente sotto il profilo logico-metodologico. Ci si riferirà, in questa sede, a quelle che trovano le loro premesse nella teoria della significatività e nella teoria dei test. Queste teorie sono molto diverse quanto alle impostazioni teoriche e logiche ma, sotto particolari condizioni, convergono tecnicamente.

9.2 | Ipotesi statistica e test statistico

Dato un carattere statistico X osservato in una popolazione definita. Si immagini di conoscere la forma distributiva $f(x, \theta)$ dove il parametro

θ è però incognito.

Si dice *ipotesi statistica* una congettura circa l'incognito parametro θ della popolazione.

Si dice *statistica-test* ogni procedimento inteso a trarre dalla n -pla campionaria indicazioni oggettive in merito all'ipotesi formulata su θ .

L'ipotesi può, ad esempio, attribuire al parametro θ un valore numerico θ_0 :

$$H_0 : \theta = \theta_0 \quad (9.1)$$

osservato un campione tratto dalla popolazione di riferimento con criterio casuale (probabilistico) in cui la stima del parametro di riferimento è t , ci si chiede se il valore t è compatibile col valore di θ_0 , ovvero se la differenza tra t e θ_0 può essere attribuita esclusivamente alla procedura di campionamento adottata.

L'ipotesi H_0 viene detta correntemente *ipotesi nulla*¹.

***Ipotesi nulla o ipotesi H_0** - è l'ipotesi di completa accidentalità dei risultati campionari osservati rispetto alla struttura della popolazione, qualunque differenza tra la costante del campione e il valore θ_0 attribuito al parametro è dovuta al solo errore di campionamento.*

Sull'ipotesi H_0 si fonda una metodologia inferenziale detta *teoria della significatività*. Nel contesto logico di questa teoria, tale ipotesi può essere rifiutata o non rifiutata, in ragione della probabilità dell'avverarsi del risultato campionario, se H_0 è vera. Quando questa probabilità risulta molto piccola o, comunque, inferiore a un prestabilito livello, l'ipotesi viene rifiutata con la seguente motivazione: o si è avverato un'evento raro (campione estremo) o l'ipotesi H_0 è falsa. Poichè gli eventi rari sono "rari", si decide di ritenere falsa l'ipotesi H_0 .

Sorta nell'ambito dell'indagine naturalistica e divenuta canone della ricerca nel contesto della variabilità biologica, la "teoria della significatività" (dovuta principalmente a R.A.Fisher) è intesa a saggiare anzitutto gli effetti della accidentalità campionaria sui risultati sperimentali.

¹Traduzione (banale) di *null hypothesis*.

Teoria della significatività (R.A Fisher) $\rightarrow H_0$
 pone un'unica ipotesi, H_0 , che può essere rifiutata o non rifiutata. Quando non è rifiutata si ritiene provvisoriamente vera.

Dalla metodologia della significatività, suscitata dai problemi investigativi della ricerca scientifica nel contesto della variabilità naturale, ha tratto sviluppo la *teoria dei tests d'ipotesi* (dovuta principalmente a J. Neyman), che ha trovato utile impiego in svariati settori operativi.

In questa metodologia, si contrappone ad H_0 un'ipotesi parametrica alternativa H_1 : e si conviene che se una delle due ipotesi è rifiutata, l'altra risulta accettata. Il criterio per decidere è affidato alla valutazione e al confronto delle probabilità associate al risultato campionario.

L'ipotesi H_1 può essere “*semplice*”:

$$H_1 : \theta = \theta_1$$

oppure “*complessa*”:

$$H_1 : \theta < \theta_0 \quad \text{ovvero} \quad H_1 : \theta > \theta_0.$$

Queste ipotesi vengono dette “unidirezionali”. Ipotesi “bidirezionale” sarebbe la seguente:

$$H_1 : \theta \neq \theta_0$$

ossia l'ovvia alternativa all'ipotesi H_0 , implicita nella teoria della significatività². Le tecniche statistiche per la scelta tra le due ipotesi si risolvono appunto nel rifiutare (o non rifiutare) l'ipotesi H_0 e, contestualmente, nell'accettare (o non accettare) l'ipotesi H_1 .

9.2.1 | La statistica-test

Nella metodologia della “significatività” per confutare l'ipotesi H_0 vengono utilizzate delle grandezze variabili denominate *statistiche-test*,

²Nella teoria neymaniana si assume il “non rifiuto” di H_0 come “accettazione” di H_0 (conseguente al rifiuto di H_1). Tra non rifiutare H_0 e accettare H_0 corre tuttavia una sottile ma profonda differenza concettuale. Nell'analisi della significatività, non rifiutare H_0 significa non ritenere l'esito campionario sufficientemente probatorio della presenza di fattori estranei all'errore di campionamento. E il rifiuto di H_0 è soltanto uno degli elementi di prova, fra i tanti di cui la ricerca si avvale.

Teoria dei test d'ipotesi (Neyman-Pearson) →
 H_0 e H_1
 pone due ipotesi a confronto, l'ipotesi nulla H_0 e l'ipotesi alternativa H_1 ; il rifiuto di un'ipotesi implica la necessaria accettazione dell'altra, (si conviene di rifiutare H_0 , e quindi accettare H_1 quando l'esito campionario risulta, data H_0 , meno verosimile rispetto ad H_1).

esse saranno diverse a seconda del tipo di ipotesi da controllare, pertanto si avranno “statistiche test” che faranno riferimento a distribuzioni campionarie di tipo Normale, t di Student, χ^2 , F di Fisher-Snedecor, ecc...

La valutazione dei dati campionari passa quindi attraverso una sintesi, la *statistica-test* $T|H_0$, costruita per il particolare problema inferenziale. La variabile $T|H_0$ descritta dai valori possibili di tale statistica è suddivisa in due regioni: la “regione di rifiuto” e la “regione di non rifiuto” dell'ipotesi H_0 .

Statistica test $T|H_0$ - è la variabile descritta dai possibili valori della statistica campionaria sotto H_0 nello spazio dei campioni.

Sull'intero campo di variabilità di $T|H_0$ è ripartita una probabilità unitaria $p = 1$ (è la probabilità dell' “evento certo”, essendo certo che la statistica-test assuma uno dei valori compresi nel campo di variabilità).

Una quota di tale probabilità, pari ad α , è attribuita alla regione di rifiuto di H_0 e la restante quota $(1 - \alpha)$ a quella di non rifiuto.

Si conviene di rifiutare H_0 quando il valore della statistica-test calcolato nel campione, si situa entro la regione di rifiuto, essendo piccola la probabilità di ottenere un risultato siffatto per il solo errore di campionamento. La regola è dunque di dubitare della completa causalità di un risultato che abbia una probabilità molto piccola di avverarsi sotto H_0 .

Rifiutare H_0 equivale a ritenere il risultato campionario “significativo”, nel senso di “significativo della presenza di fattori estranei all'errore di campionamento” (tanto più significativo quanto più α è piccolo).

La probabilità α , detta *livello di significatività*, è dunque la probabilità di un rifiuto improprio dell'ipotesi H_0 (fig. 9.1): errore di I tipo.

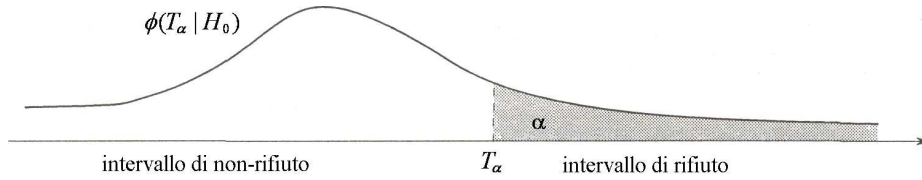


fig. 9.1

Fin qui l'analisi della significatività. La teoria dei test d'ipotesi prende in considerazione anche la funzione di distribuzione $\phi(T|H_1)$. L'intervallo di tutti i possibili valori della statistica test è ancora diviso in due intervalli: la regione di rifiuto di H_0 , che diventa accettazione di H_1 e la regione di accettazione di H_0 , che diventa rifiuto di H_1 .

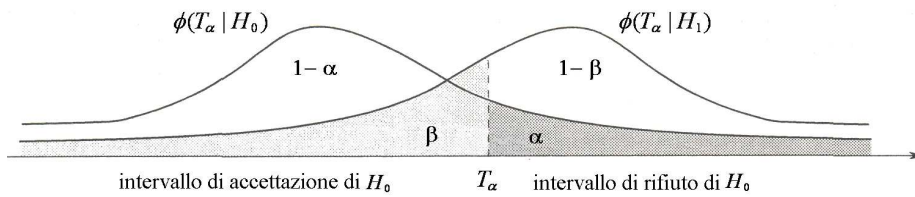


fig. 9.2

La figura 9.2 rende evidente la variabile descritta dalla “statistica-test” e la distribuzione delle probabilità, subordinatamente alle due ipotesi rivali.

La figura specifica inoltre le probabilità: α , $1 - \alpha$, β , $1 - \beta$. L'area α misura la probabilità di rifiutare come falsa una ipotesi H_0 vera: errore di I tipo. Mentre l'area β misura la probabilità di accettare come vera un'ipotesi H_0 falsa: errore di II tipo.

9.3 | Considerazioni generali sulle teorie

9.3.1 | La teoria della significatività

Nella teoria della significatività, l'ipotesi H_0 può essere soltanto “rifiutata” o “non rifiutata”. Se è rifiutata, è aperta la via alla ricerca o alla definizione del fattore o dei fattori differenzianti. Naturalmente, trattandosi di fenomeni soggetti alla variabilità statistica, il rifiuto o il non rifiuto dell'ipotesi H_0 non ha la conclusività della falsificazione di un'ipotesi nell'ambito della “logica del certo”, dove un'ipotesi è confutata dall'avverarsi di un evento incompatibile con lo spazio degli eventi possibili.

Posta H_0 come “antecedente”, il “conseguente” è una variabile $T|H_0$: la variabile definita dalla pluralità disgiuntiva ed esaustiva dei valori della statistica-test compatibili con H_0 ; su tale variabile è distribuita la probabilità dell'“evento certo”.

Invece, nella logica del probabile, in cui si colloca il controllo di H_0 , non è possibile una confutazione definitiva di tale ipotesi perchè tutti i risultati campionari sono virtualmente compatibili con l'ipotesi parametrica H_0 , seppur con diversa probabilità. La confutazione dell'ipotesi H_0 , infatti, non è determinata dall'avverarsi di eventi incompatibili con l'ipotesi stessa, bensì di eventi che - sotto tale ipotesi - hanno una piccola probabilità di avverarsi. Un criterio giustificato dal “principio di semplice disgiunzione” (Fisher): “o si è avverato un evento raro, o l'ipotesi H_0 non è vera”.

Nel contesto della ricerca scientifica, il rifiuto improprio di H_0 può indurre alla vana ricerca di un fattore inesistente; più grave è il non rifiuto improprio di H_0 , in quanto può distogliere dalla ricerca intorno a circostanze che potrebbero essere rilevanti.

Le tecniche per la confutazione dell'ipotesi H_0 variano a seconda del parametro oggetto di inferenza, del problema investigativo specifico, della legge distributiva del carattere in esame, ma si risolvono tutte nel ripartire una probabilità unitaria $p = 1$ sulla variabile descritta dai valori campionari possibili della statistica-test.

L'“analisi della significatività”, nel senso di Fisher, attiene dunque al momento preliminare di ogni investigazione intorno alle proprietà statistiche di caratteri variabili in un contesto empirico. Una sola ipo-

tesi (l'ipotesi H_0) è messa alla prova dei fatti (campione) ed è fissata nello spazio virtuale dei risultati (in ragione della probabilità di accadimento accidentale: la *verosimiglianza*) una "soglia critica" di rifiuto, così da delimitare la sottoclasse degli eventi con minor probabilità: un processo logico riconducibile all'"asserzione ipotetica": "Se H , allora E ", (ove H è l'"antecedente" e E il "conseguente"). Ma quando l'antecedente è H_0 , il conseguente è una pluralità di esiti campionari non certi, dotati ciascuno di una propria probabilità.

L'asserto è pertanto: "Se H_0 , allora $T|H_0$, essendo

$$\int_{-\infty}^{+\infty} \phi(T|H_0) dT|H_0 = 1.$$

A stretto rigore logico - giova ripeterlo - nessuno dei risultati possibili descritti dalla variabile $T|H_0$ può falsificare l'ipotesi, e pertanto per rendere operativo il criterio si deve convenire, con Fisher, di rifiutare H_0 quando all'esito campionario osservato si associa una probabilità inferiore a un valore-soglia prescelto.

La variabile $T|H_0$ descrive i possibili valori della statistica-test, subordinatamente all'ipotesi nulla H_0 , e α è la probabilità attribuita ai valori "estremi": graficamente è l'area sottesa alla curva descritta dalla $\phi(X|H_0)$ sul tratto di ascissa attribuito all'"intervallo di rifiuto". Tale intervallo attiene ai valori della variabile $T|H_0$ superiori a un limite T_α , e la probabilità α è espressa allora dall'integrale

$$\int_{T_\alpha}^{+\infty} \phi(T|H_0) dT|H_0 = \alpha$$

il complemento a 1 dà la probabilità $1 - \alpha$:

$$1 - \int_{T_\alpha}^{+\infty} \phi(T|H_0) dT|H_0 = 1 - \alpha.$$

□—————

◇ **Osservazione 9.1.** Per intendere come nella "logica del probabile" l'ipotesi venga rifiutata per l'avverarsi di un evento non

incompatibile con essa, ma soltanto poco probabile sotto il suo dominio, si immagini di aver lanciato dieci volte una moneta e di aver ottenuto due volte “testa”.

Cosa si può affermare circa l'accidentalità di tale esito?

Posta l'ipotesi H_0 (che qui sta per ipotesi di moneta non asimmetrica e di lanci non tendenziosi), lo spazio dei risultati comprende undici esiti fra i quali è ripartita una probabilità unitaria (tabella 9.1).

Tabella 9.1

$T H_0$	$\Pr\{T H_0\}$
0	1/1024
1	10/1024
2	45/1024
3	120/1024
4	210/1024
5	252/1024
6	210/1024
7	120/1024
8	45/1024
9	10/1024
10	1/1024
	1024/1024

La somma delle undici probabilità vale $(1024/1024) = 1$, ed è la probabilità dell'evento certo: lanciando dieci volte una moneta e contando le uscite di testa, è certo l'avverarsi di uno degli undici esiti possibili.

La probabilità del particolare risultato ottenuto (due “teste” in dieci lanci) è:

$$\Pr\{2; 10|H_0\} = \binom{10}{2} \frac{1}{2^{10}} = 45/1024.$$

E corrisponde al terzo addendo dello sviluppo del binomio

$$\left(\frac{1}{2} + \frac{1}{2}\right)^{10}.$$

È lecito affermare la non completa casualità del risultato?

Due sono le risposte possibili:

(a) il risultato è uno dei possibili eventi dedotti da H_0 ; non essendo incompatibile con tale ipotesi, questa non può essere confutata (logica del certo);

(b) il risultato non è incompatibile con l'ipotesi H_0 , ma è compreso in un intervallo di esiti cui è attribuita una piccola probabilità; quindi può essere lecito sospettarne la non completa accidentalità: l'ipotesi può essere confutata (logica del probabile).

Ovviamente, se la moneta fosse truccata a svantaggio dell'uscita di "testa", l'avverabilità di due "teste" in dieci lanci sarebbe inferiore a quella determinata sotto l'ipotesi H_0 . Il contrario, se fosse truccata a vantaggio di tale "faccia".

Eventuali informazioni sullo stato della moneta, tali da consentire di fissare una probabilità *a priori* dell'uscita di testa diversa da $p = 1/2$, in concorso con la verosimiglianza $\Pr\{2; 10|H_0\} = 45/1024$, consentirebbero di valutare (bayesianamente) la probabilità dell'ipotesi H_0 dopo il risultato "2 teste in 10 lanci": $\Pr\{H_0|2; 10\}$. Ma, in una situazione di indifferenza *a priori* delle informazioni e delle opinioni, rispetto alle due ipotesi alternative:

$$\Pr\{H_0\} = \Pr\{\neg H_0\},$$

ci si deve affidare alla sola verosimiglianza $\Pr\{2; 10|H_0\}$.

L'analisi della significatività impone di fissare una soglia critica nella variabile descritta dai possibili esiti, che distingua lo spazio degli eventi, in ragione di un livello prescelto α di significatività. Per esempio, la probabilità (unitaria) dell'evento certo può essere ripartita in modo da attribuire un valore $1 - \alpha = 0,95$ all'intervallo di non rifiuto e il residuo $\alpha = 0,05$ all'intervallo di rifiuto.

Ove non sussistano motivi per sospettare che l'eventuale non casualità dell'esito vada a vantaggio di una soltanto delle due facce (in tal caso l'area di rifiuto insisterà tutta sui risultati estremi a favore di tale faccia), l'intervallo di rifiuto di H_0 attiene ad entrambi gli estremi: il criterio è a "due code" e si bipartisce la probabilità α , attribuendo $\alpha/2$ a ciascuna coda.

Se il test è a una sola coda, il risultato "due teste in dieci lanci" ($p = 45/1024 = 0,044$) cade nell'intervallo di rifiuto di

H_0 al livello di significatività $\alpha \leq 0,05$. Se il risultato fosse “una testa in dieci lanci” ($p = 10/1024 = 0,009$) l’ipotesi H_0 sarebbe confutata al livello di significatività $\alpha \leq 0,01$.

—————□

9.3.2 | La teoria dei test d’ipotesi

Altro è il contesto, altra la logica, della teoria neymaniana, ove il rifiuto dell’ipotesi H_0 implica l’accettazione simultanea di una prefissata ipotesi alternativa H_1 . E’ quasi superfluo osservare che, se le due ipotesi a confronto sono esaustive, si rientra, in sostanza, nella teoria della significatività; se invece non sono esaustive, possono essere entrambe false. Tuttavia, una delle due, questa è la regola, deve essere accettata come vera. Nella ricerca scientifica, il cui intento è la conoscenza, un siffatto criterio è insostenibile. In un contesto strategico, tale criterio risponde a principi di opportunità operativa.

La regola di scelta tra le due ipotesi a confronto consiste ancora nel suddividere la variabile descritta da tutti gli esiti possibili in intervalli: l’intervallo di rifiuto dell’ipotesi H_0 e quello di non rifiuto, inteso come intervallo di accettazione di H_0 . La linea di confine tra i predetti intervalli è determinata in ragione della distribuzione dei possibili valori della *statistica-test*.

Quando il valore della statistica-test calcolata nel campione cade nell’intervallo cui è attribuita la probabilità α , l’ipotesi H_0 viene rifiutata ed è quindi accettata l’ipotesi H_1 . La decisione di respingere H_0 può essere sbagliata (α misura appunto la probabilità del rifiuto errato di H_0) e così l’accettazione, contestuale, di H_1 . Tale decisione impropria su H_0 viene chiamata “errore del primo tipo” (respingere come falsa un’ipotesi vera), mentre l’accettazione di un’ipotesi che dovrebbe invece essere rifiutata perchè “non vera” viene chiamato “errore del secondo tipo” (assumere come vera un’ipotesi falsa).

Indicando con R_{H_0} la regione di rifiuto dell’ipotesi H_0 , con A_{H_0} la regione di accettazione di H_0 e con T_c il valore concreto assunto dalla statistica-test, si distinguono le seguenti probabilità:

$$\begin{aligned}\Pr\{T_c \in R_{H_0} | H_0\} &= \alpha \\ \Pr\{T_c \in A_{H_0} | H_0\} &= 1 - \alpha \\ \Pr\{T_c \in A_{H_0} | H_1\} &= \beta \\ \Pr\{T_c \in R_{H_0} | H_1\} &= 1 - \beta.\end{aligned}$$

Se $T|H_1$ è la variabile costituita da tutti i possibili esiti della statistica-test subordinatamente all'ipotesi alternativa H_1 , e $\phi(T|H_1)$ è la sua distribuzione di probabilità, la probabilità β è espressa dall'integrale

$$\int_{-\infty}^{T_\alpha} \phi(T|H_1) dT|H_1 = \beta$$

e il complemento a 1 è la probabilità $1 - \beta$

$$1 - \int_{-\infty}^{T_\alpha} \phi(T|H_1) dT|H_1 = 1 - \beta.$$

9.3.3 | Inferenza bayesiana

La metodologia neymaniana dei tests d'ipotesi si risolve dunque nella scelta di una tra due ipotesi rivali, affidata alle “verosimiglianze” delle medesime, ossia al rispettivo grado di probabilità nei confronti dei dati campionari.

Una scelta tra due ipotesi alternative dovrebbe - a dir vero - tenere conto anche del grado di ammissibilità *a priori* delle ipotesi stesse, che è come dire della loro credibilità, indipendentemente dal particolare contesto campionario sul quale vengono confrontate. (Tuttavia, la teoria dei tests d'ipotesi trova una sua giustificazione nel “lungo andare” e ha quindi un ruolo nell'ambito delle scelte strategiche standard.)

Se si conviene di misurare il grado di ammissibilità *a priori* delle due ipotesi con un numero reale da 0 a 1, da intendersi come probabilità - $\Pr\{H_0\}$ e $\Pr\{H_1\}$ - l'algoritmo di Bayes-Laplace, che formalizza le inferenze induttive, consente di trarre le “probabilità a posteriori” dal concorso di tali probabilità con le verosimiglianze $\Pr\{T_c|H_0\}$ e $\Pr\{T_c|H_1\}$:

$$\Pr \{H_0|T_c\} = \frac{\Pr \{H_0\} \Pr \{T_c|H_0\}}{\Pr \{T_c\}} \quad (9.2)$$

$$\Pr \{H_1|T_c\} = \frac{\Pr \{H_1\} \Pr \{T_c|H_1\}}{\Pr \{T_c\}}.$$

Questo è il più coerente criterio di scelta razionale tra due ipotesi a confronto in un contesto campionario. Ed è anche il più intuitivo. Non si dovrebbe scegliere tra ipotesi rivali senza considerarne, preliminarmente, l'ammissibilità. L'algoritmo bayesiano,³ tuttavia, consente di dimostrare il convergere asintotico della probabilità inversa e della verosimiglianza all'ingrandirsi del contesto empirico sul quale le verosimiglianze vengono commisurate. Svela quindi la crescente irrilevanza delle probabilità a priori, $\Pr\{H_0\}$ e $\Pr\{H_1\}$, all'aumentare del numero dei costituenti il campione. (Ovviamente, se nessuna delle probabilità a priori è posta uguale a 0 o 1).

9.4 | Errore di I e II tipo e funzione potenza

Nell'ambito della teoria neymaniana è quindi possibile incorrere in due tipi di errori: "errore del primo tipo" (respingere come falsa un'ipotesi vera) o "errore del secondo tipo" (assumere come vera un'ipotesi falsa).

Riassumendo: α misura la probabilità di respingere l'ipotesi H_0 quando essa è vera; β misura la probabilità di accettare l'ipotesi H_0 quando essa è falsa; $1 - \beta$ misura la probabilità di rifiutare H_0 quando è falsa: è la cosiddetta "potenza del test" (un test è tanto più "potente" quando più è atto a respingere ipotesi false).

Al concetto di potenza è ispirato un criterio di scelta del livello α inteso a rendere minima la probabilità di un errore di probabilità β , e massima la probabilità $1 - \beta$ di rifiutare H_0 quando è falsa: è detto "lemma di Neyman-Pearson jr."

³L'impostazione bayesiana dell'inferenza statistica si rifà - almeno nel nome - a un saggio di Thomas Bayes apparso (postumo) nel 1763: *An essay toward solving a problem in the doctrine of chances*, "Philosophical Transactions", London, oltre che a un saggio di Pierre Simon de Laplace, apparso nel 1774: *Mmoire sur la probabilitis des causes par les vnements*, Acadmic des Sciences", Paris.

□—————

◇ Data una n -pla campionaria (x_1, \dots, x_n) , tratta da una popolazione in cui una variabile X ha distribuzione $\psi(X; \theta)$ e data l'alternativa tra le ipotesi:

$$H_0 : \theta = \theta_0; \quad H_1 : \theta = \theta_1,$$

il valore α che massimizza, a date condizioni, la potenza del test è identificato confrontando, in termini di rapporto, le rispettive funzioni di verosimiglianza:

$$\prod_{i=1}^n \psi(X; \theta_0) / \prod_{i=1}^n \psi(X; \theta_1).$$

—————□

È evidente (fig. 9.2) come elevando la soglia T_α diminuisca α e aumenti β ; e come, abbassando tale soglia, aumenti α e diminuisca β . All'aumentare di α aumenta pure $1 - \beta$; al diminuire di α , diminuisce $1 - \beta$. E' altresì evidente l'impossibilità di ridurre entrambe le probabilità α e β . Una riduzione simultanea di α e di β può avvenire soltanto riducendo la dispersione delle due distribuzioni; e ciò può ottenersi aumentando il numero n delle unità campionarie.

Quando la n -pla campionaria è fissata, si deve trovare una soluzione di compromesso tra i due tipi di errore, in ragione del contesto, e scegliere le probabilità α e β soppesando il "rischio" conseguente all'uno e all'altro tipo di errore.

In sintesi lo schema delle scelte e delle rispettive probabilità è il seguente:

9.5 | I test del rapporto di verosimiglianza

La funzione di verosimiglianza trova un'importante applicazione anche nell'ambito della verifica delle ipotesi, specie nel contesto dell'impostazione di J. Neyman e E. Pearson. Essa consente di superare le restrittive

Realtà	Decisione	
	<i>non rifiuto di H_0</i>	<i>rifiuto di H_0</i>
H_0 è vera	<i>decisione corretta</i> (probabilità $1 - \alpha$)	<i>errore di I tipo</i> (probabilità α)
H_0 è falsa	<i>errore di II tipo</i> (probabilità β)	<i>decisione, corretta</i> (probabilità $1 - \beta$)

condizioni del lemma di Neyman-Pearson, riferito ad ipotesi semplici e limitate a parametri scalari.

Dato lo spazio parametrico Θ suddiviso in due regioni ω e il suo complemento $\bar{\omega}$. Si considerino le seguenti ipotesi:

$$\begin{aligned} H_0 &: \theta \in \omega \subset \Theta \\ H_1 &: \theta \in \bar{\omega} \subset \Theta \end{aligned}$$

dove θ può anche essere un vettore. Sia λ il seguente rapporto:

$$\lambda = \frac{\max_{\theta \in \omega} L(\theta; x_1, \dots, x_j, \dots, x_n)}{\max_{\theta \in \bar{\omega}} L(\theta; x_1, \dots, x_j, \dots, x_n)} \quad (9.3)$$

la zona di rifiuto del test sarà quella in cui $\lambda \leq k$, dove k è una soglia definita a priori.

Il criterio può essere descritto nel seguente modo: *l'ipotesi zero* viene rifiutata se la spiegazione “al meglio” del campione osservato sotto H_0 (espressa dal numeratore di λ) è sensibilmente inferiore alla spiegazione “al meglio” sotto l'ipotesi alternativa (espressa dal denominatore di λ).

Di solito, la zona di rifiuto del test del rapporto di verosimiglianza viene indicata come segue:

$$\lambda = \frac{\max_{\theta \in \omega} L(\theta; x_1, \dots, x_j, \dots, x_n)}{\max_{\theta \in \Theta} L(\theta; x_1, \dots, x_j, \dots, x_n)} \leq k, \quad k < 1 \quad (9.4)$$

Siano $H_0 : \theta = \theta_0$ e $H_1 : \theta \neq \theta_0$, ipotesi nulla e alternativa sul parametro θ . Allora il test (9.4) diventa:

$$\lambda = \frac{L(\theta_0; x_1, \dots, x_j, \dots, x_n)}{\max_{\theta \in \Theta} L(\theta; x_1, \dots, x_j, \dots, x_n)} = \frac{L(\theta_0)}{L(\hat{\theta})} \leq k \quad (9.5)$$

dove $\hat{\theta}$ è la stima di massima verosimiglianza di θ e k è tale che $P(\lambda \leq k | H_0) = \alpha$. Chiaramente la costruzione del test implica la conoscenza della distribuzione di probabilità di λ .

□—————

◇ **Osservazione 9.2.** Si consideri la verifica dell'ipotesi sulla media di una popolazione normale con varianza nota. Siano $H_0 : \mu = \mu_0$ e $H_1 : \mu \neq \mu_0$. Si ha:

$$L(\mu_0) = \frac{1}{(\sigma 2\pi)^n} e^{-\sum_{i=1}^n (x_i - \mu_0)^2 / 2\sigma^2} \quad \text{e} \quad L(\hat{\mu}) = \frac{1}{(\sigma 2\pi)^n} e^{-\sum_{i=1}^n (x_i - \bar{x})^2 / 2\sigma^2},$$

essendo \bar{x} la stima di massima verosimiglianza di μ . Il rapporto λ assume perciò la forma

$$\lambda = \frac{e^{-\sum_{i=1}^n (x_i - \mu_0)^2 / 2\sigma^2}}{e^{-\sum_{i=1}^n (x_i - \bar{x})^2 / 2\sigma^2}} = e^{-n(\bar{x} - \mu_0)^2 / 2\sigma^2} \quad (9.6)$$

La zona di rifiuto è allora definita dalla relazione

$$\lambda = e^{-\sum_{i=1}^n (x_i - \mu_0)^2 / 2\sigma^2} \leq k.$$

◇ **Osservazione 9.3.** Data una n -pla campionaria (x_1, \dots, x_n) , tratta da una popolazione in cui una variabile X ha distribuzione $\psi(X; \theta)$ e data l'alternativa tra le ipotesi:

$$H_0 : \theta = \theta_0; \quad H_1 : \theta = \theta_1,$$

il valore α che massimizza, a date condizioni, la potenza del test è identificato confrontando, in termini di rapporto, le rispettive funzioni di verosimiglianza:

$$\prod_{i=1}^n \psi(X; \theta_0) \Big/ \prod_{i=1}^n \psi(X; \theta_1).$$

Nel caso quindi di due ipotesi semplici $H_0 : \theta = \theta_0$ e $H_1 : \theta = \theta_1$, il test (9.5) coincide con il test più potente individuato dal Lemma di Neyman-Pearson. Al di fuori di questo caso particolare, il test gode di apprezzabili proprietà. Nell'osservazione (9.2), ad esempio, il test ottenuto è uniformemente più potente e non distorto.

◇ **Osservazione 9.4.** Si dimostra inoltre che il *test del rapporto di verosimiglianza*, gode di due interessanti proprietà:

1. è asintoticamente consistente;

$$\lim_{n \rightarrow \infty} P(\lambda_n \leq k | H_1) = 1, \quad \forall \theta \neq \theta_0$$

2. sotto H_0 la statistica $-2 \ln \lambda_n \xrightarrow[n \rightarrow \infty]{} \chi_r^2$,

dove “r” rappresenta il numero delle componenti del vettore θ che devono essere stimate. Se il parametro è uno scalare allora si avrà ovviamente un solo grado di libertà.

◇ **Osservazione 9.5.** Nel caso vettoriale sia $f(x, \theta)$, il modello descrittivo della popolazione con $\theta = (\theta_1, \theta_2, \dots, \theta_r)$. Le ipotesi da saggiare vertono congiuntamente su tutte le r componenti. Data la verifica dell'ipotesi nulla semplice $H_0 : \theta = \theta^*$, contro l'alternativa composta che almeno una delle r uguaglianze non valga. In queste situazioni il rapporto di verosimiglianza diventa:

$$\lambda = \frac{L(\theta^*; x_1, \dots, x_j, \dots, x_n)}{\max_{\theta \in \Theta} L(\theta; x_1, \dots, x_j, \dots, x_n)} = \frac{L(\theta^*)}{L(\hat{\theta})} \leq k \quad (9.7)$$

dove $\hat{\theta}$ è il vettore delle stime di massima verosimiglianza delle r componenti di $\theta = (\theta_1, \theta_2, \dots, \theta_r)$.

Anche in questo caso valgono le proprietà asintotiche.

◇ **Osservazione 9.6.** Nel caso di ipotesi nulla composta $H_0 : \theta_1 = \theta_1^*, \theta_2 = \theta_2^*, \dots, \theta_m = \theta_m^*, \theta_{m+1}, \dots, \theta_r$ che specifica i valori di m (con $m < r$) componenti del vettore $\theta = (\theta_1, \theta_2, \dots, \theta_r)$, lasciando indeterminate le altre, il rapporto di verosimiglianza risulta

$$\lambda = \frac{L(\theta_1^*, \theta_2^*, \dots, \theta_m^*, \theta_{m+1}, \dots, \theta_r; x_1, \dots, x_j, \dots, x_n)}{L(\hat{\theta}; x_1, \dots, x_j, \dots, x_n)} \quad (9.8)$$

Anche in questo caso valgono le proprietà asintotiche, e sotto H_0 la statistica $-2 \ln \lambda_n \xrightarrow[n \rightarrow \infty]{} \chi_m^2$.

—————□

9.6 | Esercizi

9.1 Si vuole studiare la durata di un processo produttivo che dal materiale grezzo porta al prodotto finito. Il venditore del meccanismo di produzione sostiene che la durata del processo si distribuisce normalmente con media pari a 11 ore e scarto quadratico medio pari a 4 ore. L'acquirente, sulla base delle valutazioni di un esperto, sospetta invece che, pur distribuendosi normalmente e con scarto quadratico medio pari a 4, la durata media del processo sia di 14 ore. Si mettono allora in produzione 16 pezzi e si decide che il meccanismo di produzione verrà acquistato soltanto se la durata media della produzione nel campione è inferiore a 13 ore. Si calcolino:

- a) la probabilità dell'errore di I tipo (α),
- b) la probabilità dell'errore di II tipo (β)

associati al criterio di decisione sopra riportato.

9.2 Si consideri il processo produttivo descritto nell'esercizio precedente, la cui durata si distribuisce normalmente con scarto quadratico medio pari a 4. Volendo stabilire se la durata sia uguale a 11 (H_0) o 14 (H_a) ore.

- a) Quali valori deve assumere la durata media della produzione di un campione di 5 pezzi affinché si possa accettare $\mu = 11$ ad un livello di significatività del 5%?
- b) Se nel campione di 5 pezzi si fosse ottenuto un tempo medio di 12 ore, che cosa si concluderebbe circa la sorte di H_0 e H_a ?

c) Quanto vale, nel problema in esame, β , cioè la probabilità dell'errore di II tipo?

d) Quanto varrebbe β se anziché fissare $\alpha=0,05$ si fosse scelto $\alpha=0,01$?

9.3 Si hanno due tipi di lampadine, apparentemente identiche, ma tali che l'una ha una durata media di 1.000 ore, le altre durata media di 1.100 ore. Si ipotizzi che la durata delle lampadine si distribuisca normalmente, con scarto quadratico medio $\sigma=150$ ore. Si abbia poi a disposizione una partita di lampadine di cui non si conosce il tipo, ma si sa che è lo stesso per tutte. Si consideri quindi un campione casuale di lampadine per decidere, se la partita appartenga al tipo di durata 1.000 ore o a quello di durata 1.100 ore.

a) Di quante lampadine deve essere costituito il campione perché si verifichi $\alpha = \beta=0,01$?

b) Si scelga $\alpha=0,05$ e si selezioni un campione di 100 unità statistiche. Quanto vale ora β ?

c) A parità di α , quanto varrebbe β se il campione contenesse soltanto 49 lampadine?

Test su medie, proporzioni e varianze

10.1 | Introduzione

Sulla base della teoria generale illustrata nel capitolo precedente, risulta evidente che il rifiuto o l'accettazione dell'ipotesi nulla, avvenga sulla base del campione osservato.

Tale procedimento teso a trarre dalla n -pla campionaria indicazioni oggettive in merito all'ipotesi formulata, é comunemente noto come *test statistico*.

Fissato α , (il livello di significativit del test), e dedotta sotto H_0 la distribuzione della statistica-test, si suddivide tale variabile in due regioni: la *regione di rifiuto* di H_0 , cui é attribuita la probabilitá α , e la *regione di accettazione* di H_0 , cui é attribuita la probabilitá $(1 - \alpha)$.

Se il test statistico (valore assunto nel campione dalla statistica-test) é compreso nell'intervallo di rifiuto, l'ipotesi H_0 viene rifiutata e viene accettata l'ipotesi H_1 .

Convenzionalmente si pone $\alpha = 0,10$, $\alpha = 0,05$, $\alpha = 0,01$ oppure $\alpha = 0,001$.

10.2 | Test per la media aritmetica di una popolazione.

In una logica di test di ipotesi in cui sia ammissibile identificare un'ipotesi H_1 alternativa all'ipotesi H_0 , è possibile procedere all'accettazione provvisoria di H_1 conseguentemente al rifiuto dell'ipotesi H_0 che viene messa sotto controllo.

Sia m la media di un campione (x_1, \dots, x_n) e sia μ , l'incognita media della popolazione di riferimento. La variabile \bar{X} descritta dalle medie aritmetiche di tutti i campioni possibili ha media μ e scarto quadratico medio σ/\sqrt{n} , e tende a distribuirsi come una normale:

$$L(\bar{X}) \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Osservata la media m in un campione, porre l'ipotesi

$$H_0 : \mu = \mu_0$$

è come postulare che, se sussiste una differenza $|\mu - \mu_0| > 0$, essa è dovuta al solo errore casuale di campionamento.

Per saggiare l'ipotesi H_0 , ci si avvale, come si è anticipato, della variabile standardizzata

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}. \quad (10.1)$$

Scelto un "livello di significatività" α , restano fissate, in ascissa, le soglie $-z_{\alpha/2}$ e $+z_{\alpha/2}$ tali che

$$1 - \alpha = \Pr(-z_{\alpha/2} < |z| < +z_{\alpha/2})$$

sia la probabilità di un esito non esterno all'intervallo $[-z_{\alpha/2}; +z_{\alpha/2}]$.

Se, ove ammissibile, l'alternativa all'ipotesi H_0 è bidirezionale

$$H_0 : \mu \neq \mu_0,$$

al di fuori dell'intervallo di probabilità $1 - \alpha$ esistono due intervalli (due "code") che costituiscono la regione di rifiuto, a ciascuno dei quali è attribuita una probabilità $\alpha/2$.

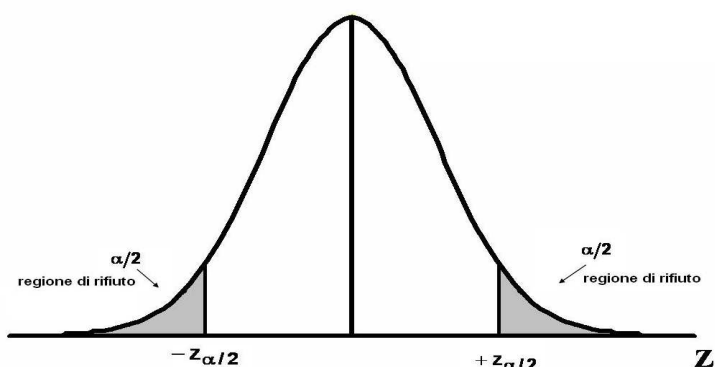


fig. 10.1 - Verifica d'ipotesi bidirezionale

Se invece l'ipotesi alternativa è:

$$H_1 : \mu < \mu_0, \quad \text{oppure: } H_1 : \mu > \mu_0$$

la regione di rifiuto consta di un solo intervallo (una "coda"), al quale è attribuita l'intera probabilità α

$$\alpha = \Pr \left(\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} > z_\alpha \right).$$

Se l'alternativa è bidirezionale, l'ipotesi H_0 è rifiutata quando il valore della statistica-test commisurata alla variabile Z cade nella regione di rifiuto, ossia entro uno degli intervalli $[-\infty; -z_{\alpha/2}]$; $[+z_{\alpha/2}; +\infty]$ cui corrisponde la probabilità di un valore della statistica-test $|z| > z_{\alpha/2}$.

Se l'alternativa è unidirezionale nella forma:

$$H_1 : \mu > \mu_0,$$

allora l'ipotesi H_0 è rifiutata (e perciò è accettata l'ipotesi H_1) quando la statistica-test assume un valore $z > z_\alpha$.

Se l'alternativa è unidirezionale nella forma:

$$H_1 : \mu < \mu_0,$$

allora l'ipotesi H_0 è rifiutata (e perciò viene accettata H_1 quando la statistica-test assume un valore $z < -z_\alpha$.

In generale, la varianza della popolazione, σ^2 , non è nota e la si deve stimare induttivamente calcolando la varianza campionaria corretta:

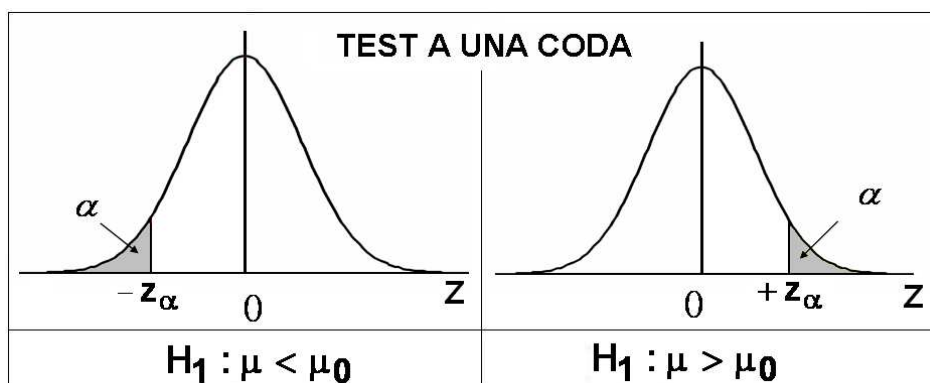


fig. 10.2 - Verifica d'ipotesi unidirezionale

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}.$$

Se H_0 è vera, la variabile

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \quad (10.2)$$

segue la distribuzione di Student con $n-1$ gradi di libertà. La statistica-test è allora:

$$t = \frac{m - \mu}{s/\sqrt{n}}. \quad (10.3)$$

Se l'alternativa è $H_1 : \mu \neq m$, si rifiuta H_0 quando il valore concreto della statistica-test t_c cade in una qualsiasi delle due code; ovvero nell'unica coda, se l'alternativa è unidirezionale.

Nella logica dei tests d'ipotesi il rifiuto di H_0 implica sempre l'accettazione dell'ipotesi alternativa.

	<i>Ipotesi su una media</i> $H_0 : \mu = \mu_0$		
	<i>(σ noto)</i>	<i>(σ non noto)</i>	
		campioni $n > 30$	campioni $n \leq 30$
Variabile statistica-test	$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$	$Z = \frac{\bar{X} - \mu_0}{\hat{S}/\sqrt{n}} \sim N(0, 1)$	$T = \frac{\bar{X} - \mu_0}{\hat{S}/\sqrt{n}} \sim t_{n-1}$
Test statistico	$z_c = \frac{m - \mu_0}{\sigma/\sqrt{n}\sqrt{n}}$	$z_c = \frac{m - \mu_0}{\hat{s}/\sqrt{n}}$	$t_c = \frac{m - \mu_0}{\hat{s}/\sqrt{n}}$
test a 2 code $H_1 : \mu \neq \mu_0$ RIFIUTO H_0 SE	$ z_c \geq z_{\alpha/2}$	$ z_c \geq z_{\alpha/2}$	$ t_c \geq t_{\alpha/2; n-1}$
test 1 coda $H_1 : \mu < \mu_0$ RIFIUTO H_0 SE	$z_c \leq -z_{\alpha}$	$z_c \leq -z_{\alpha}$	$t_c \leq -t_{\alpha; n-1}$
test 1 coda $H_1 : \mu > \mu_0$ RIFIUTO H_0 SE	$z_c \geq z_{\alpha}$	$z_c \geq z_{\alpha}$	$t_c \geq t_{\alpha; n-1}$

Esempio 10.2.1.

Si immagini, ad esempio, che una casa produttrice di automobili, per stimare il numero medio di chilometri percorribili con un litro di benzina “super”, accerti su un campione casuale di 15 autovetture collaudate nelle identiche condizioni, il numero medio aritmetico di chilometri per litro. Sia tale media $m = 17,4 \text{ km/l}$ e lo scarto quadratico medio sia $s = 2,6$. Questo risultato è compatibile con l'ipotesi di una prestazione media di 18 km/l , a parità di scarto quadratico medio?

L'ipotesi da sottoporre a controllo è del tipo $H_0 : \mu = \mu_0$, ossia $H_0 : \mu = 18 \text{ km/l}$; una sua alternativa può essere $H_1 : \mu < \mu_0$, ossia: $H_1 : \mu < 18 \text{ km/l}$.

Il valore concreto della (10.3) è

$$t_c = \frac{|17,4 - 18|}{2,6/\sqrt{15}} = \frac{0,6}{0,67} = 0,89.$$

Tale valore della statistica-test $t_c = 0,89$ va confrontato con il valore teorico t_t , calcolato sotto l'ipotesi H_0 , per un assegnato livello di significatività α e per un numero ν di gradi di libertà.

Posto un livello di significatività $\alpha = 0,05$, il valore teorico della t , riportato nella tavola II dell'Appendice, per $(15 - 1) = 14$ gradi di libertà, risulta $t_{0,05(14)} = 2,145$. Pertanto, essendo $t_c < t_t$ non si rifiuta l'ipotesi che il campione provenga da una catena produttiva standardizzata su un consumo medio di 18 *km/l*. (Naturalmente, se l'ipotesi H_0 non è rifiutata per un assegnato α , essa resta implicitamente non rifiutata per ogni altro livello inferiore ad α).

10.3 | Test per una frequenza

Nell'analisi dei dati campionari, il confronto tra due proporzioni circa la presenza di un'assegnata caratteristica A nella popolazione di riferimento si avvale della distribuzione binomiale.

Sia $Pr(A) = p$ la probabilità dell'evento A e $Pr(\neg A) = 1 - p = q$ la probabilità dell'evento *non-A*. La probabilità che A si avveri x volte, in n prove indipendenti, è

$$Pr(x; n) = \binom{n}{x} p^x q^{n-x}.$$

La distribuzione cumulata della binomiale offre la probabilità di ottenere l'evento A un numero di volte inferiore o uguale a un assegnato livello k .

$$\sum_{x=0}^k Pr(x; n) = \sum_{x=0}^k \binom{n}{x} p^x q^{n-x}.$$

Se tale probabilità risulta inferiore alla soglia prefissata α , si può rifiutare l'ipotesi H_0 di casualità dei risultati nella popolazione di riferimento.

Esempio 10.3.1.

Per valutare l'effetto della stanchezza sull'apprendimento, uno psicologo ha sottoposto un campione casuale di 8 studenti alla seguente prova: a ciascuno studente sono stati affidati due brani di 30 parole,

uno da imparare a memoria la mattina, prima dell'inizio delle lezioni, l'altro il pomeriggio, al termine delle lezioni. Il giorno successivo ogni studente è stato invitato a ripetere perfettamente uno dei due brani, a piacere. Solo due studenti hanno scelto il brano memorizzato nel pomeriggio; i restanti sei hanno preferito il brano studiato al mattino. Questo risultato è compatibile con l'ipotesi di scelta casuale dei brani da parte degli studenti $H_0 : p_1 = p_2 = 0,5$, oppure configura una diversa capacità di memoria in condizione di minore affaticamento mentale? L'ipotesi alternativa è quindi $H_1 : p_1 > 0,5$.

Per rispondere, si può adottare il modello binomiale e calcolare, per $x = 0, 1, 2$, la distribuzione di ripartizione nei valori estremi:

$$\begin{aligned} \sum_{x=0}^2 \Pr(x; n=8, p=0,5) &= \binom{8}{0} 0,5^0 \times 0,5^8 + \binom{8}{1} 0,5^1 \times 0,5^7 + \\ &+ \binom{8}{2} 0,5^2 \times 0,5^6 = 0,1445. \end{aligned}$$

La probabilità di ottenere non più di 2 esiti di un certo tipo in 8 prove vale 0,1445, ed è maggiore del valore critico $\alpha = 0,05$: si conviene allora di non respingere l'ipotesi H_0 che gli studenti abbiano scelto a caso il brano da ripetere.

Quando il numero delle prove sia appena elevato, il calcolo delle probabilità cumulate della distribuzione binomiale può essere non sempre agevole; conviene allora adottare come approssimazione calcolatoria la distribuzione normale nella forma standardizzata:

$$z = \frac{f - p}{\sqrt{pq/n}}, \quad (10.4)$$

dove f è la frequenza incognita di una data modalità nel campione che può essere assunta come stima della corrispondente probabilità nella popolazione, p è la probabilità incognita della modalità nella popolazione di riferimento e n è il numero delle unità del campione.

Esempio 10.3.2.

Si è osservato il comportamento di 35 clienti, tutte donne, nella scelta tra due casse all'uscita da un supermercato; nella prima l'operatore era una femmina, nella seconda, un maschio: 23 clienti hanno preferito l'operatore femmina. Il risultato è compatibile con l'ipotesi di una scelta del tutto casuale?

L'ipotesi da sottoporre a controllo è $H_0 : p = 1 - p = 0,50$.

Si calcola $f_x = 23/35 = 0,66$. E allora, per la (10.4),

$$z = \frac{0,66 - 0,50}{\sqrt{0,50 \times 0,50/35}} = \frac{0,16}{0,084} = 1,90.$$

Poiché, ad un livello di significatività $\alpha = 0,05$, il risultato non è esterno ai valori soglia $z = \pm 1,96$, il test suggerisce di non respingere (o, se si vuole, di conservare) l'ipotesi $H_0 : p = 0,50$: ciò significa che le clienti del supermercato sono indifferenti alla scelta dell'operatore di cassa, sia esso maschio o femmina.

10.4 | Test per il confronto fra due medie aritmetiche.

10.4.1 | Campioni appaiati

Si hanno due campioni appaiati quando gli stessi individui vengono osservati più di una volta, generalmente in circostanze diverse, o quando si esaminano due diversi gruppi i cui costituenti sono il più possibile simili a due a due tranne che per il fattore sperimentale oggetto d'interesse.

Nell'esame di dati appaiati l'attenzione si rivolge alla media delle differenze tra le coppie e alla variabilità di queste differenze. Interessa quindi la variabilità entro i soggetti piuttosto che tra i soggetti. Il vantaggio del *disegno sperimentale per campioni appaiati* sta nel fatto che la variabilità tra i soggetti può essere rimossa osservando solo le differenze entro i soggetti: è come se si avesse a che fare con un solo campione, e quindi si possono adottare metodi di analisi analoghi a quelli precedentemente illustrati.

Un prerequisito essenziale al controllo della significatività delle differenze fra due campioni appaiati è che la distribuzione delle differenze fra i valori del carattere osservato sia di tipo normale.

Indicata con δ la media aritmetica delle differenze fra i dati appaiati nelle popolazioni da cui i campioni provengono, l'ipotesi nulla $H_0 : \delta = 0$ configura l'irrelevanza dei fattori sperimentali differenzianti. La statistica test sarà allora, in analogia con la (10.3):

$$t_c = \frac{\bar{d} - 0}{se(\bar{d})} \quad (10.5)$$

dove \bar{d} è la media delle differenze calcolate sui due campioni e $se(\bar{d})$ è l'errore standard di \bar{d} ottenuto come rapporto fra lo scarto quadratico medio corretto di d calcolato sui dati campionari e la radice del numero delle coppie, s_d/\sqrt{n} . Tale statistica si distribuisce come una t di Student con $n - 1$ gradi di libertà. Fissato α e individuato il valore $|t_{\alpha/2; n-1}|$, si procede come illustrato nel paragrafo (10.1). Per n sufficientemente grande la (10.4) converge a una $N(0, 1)$ e quindi si può ricorrere al test normale.

10.4.2 | Campioni indipendenti

Siano due campioni con medie m_1 e m_2 tra loro differenti. L'analisi della significatività pone sotto controllo l'ipotesi (H_0) che i due campioni provengano da due popolazioni con uguale valor medio.

$$H_0 : \mu_1 = \mu_2 = \mu.$$

In quanto postula l'uguaglianza tra le medie aritmetiche delle popolazioni da cui i due campioni provengono, l'ipotesi attribuisce la disuguaglianza $m_1 \neq m_2$ all'errore casuale di campionamento. Un'eventuale ipotesi alternativa può essere "bilaterale":

$$H_1 : \mu_1 \neq \mu_2$$

e può essere "unilaterale", nella forma:

$$H_1 : \mu_1 > \mu_2$$

oppure nella forma:

$$H_1 : \mu_1 < \mu_2.$$

La variabile descritta dalle differenze tra le medie di tutti i possibili campioni casuali provenienti da una medesima popolazione, ovvero da due popolazioni identiche per media e varianza, è distribuita simmetricamente attorno allo zero e tanto più dispersa quanto più sono differenti tra loro le medie campionarie.

Essendo la distribuzione campionaria della media aritmetica \bar{X} asintoticamente una “gaussiana”, anche la variabile aleatoria differenza $(\bar{X}_1 - \bar{X}_2)$ tra le medie di due campioni appartenenti a due popolazioni con medie uguali $\mu_1 = \mu_2 = \mu$ ha andamento gaussiano, con media zero:

$$M(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 = 0$$

e varianza uguale alla somma delle varianze:

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2).$$

Standardizzando, si ottiene la variabile

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{V(\bar{X}_1) + V(\bar{X}_2)}}. \quad (10.6)$$

Il modello di riferimento è ancora “normale” e la statistica-test viene modellata sulla variabile standardizzata

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

Sotto l'ipotesi $H_0 : \mu_1 = \mu_2 = \mu$, la differenza $\mu_1 - \mu_2$ è nulla; e pertanto:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}. \quad (10.7)$$

Poichè le varianze delle popolazioni sono in genere sconosciute, ammessa l'uguaglianza $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (le due popolazioni sono “omoschedastiche”), si stima σ^2 con la media aritmetica ponderata delle varianze dei due campioni (i pesi sono i rispettivi gradi di libertà):

$$s^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

La statistica test è dunque

$$z = \frac{m_1 - m_2}{s\sqrt{1/n_1 + 1/n_2}} \quad (10.8)$$

e per grandi campioni si distribuisce come una normale standardizzata.

Quando il divario tra le medie (standardizzate) di due campioni rimane entro i valori soglia $-z_{\alpha/2}$ e $+z_{\alpha/2}$, l'ipotesi H_0 non è rifiutata; quando invece è minore di $-z_{\alpha/2}$ o maggiore di $+z_{\alpha/2}$ (ossia si colloca in una delle "code" della distribuzione), l'ipotesi H_0 è rifiutata.

Esempio 10.4.2.1

Si voglia, ad esempio, saggiare la propensione al risparmio, a seconda del sesso, in un gruppo omogeneo di persone: dirigenti di un ente pubblico appartenenti ad una stessa fascia di reddito. Estratti due campioni casuali, C_1 e C_2 , rispettivamente di 82 femmine e 36 maschi (onde tener conto della proporzione tra maschi e femmine nell'insieme di riferimento), in C_1 risulta un risparmio medio annuo (in migliaia di lire) $m_1 = 5.600$ e, in C_2 , $m_2 = 4.200$, con rispettivi scarti quadratici medi $s_1 = 650$ e $s_2 = 710$.

Che cosa si può dire della differenza tra le medie dei due campioni? Essa può dipendere dall'errore di campionamento (in tale ipotesi i campioni si ritengono appartenenti ad una stessa popolazione), oppure può esprimere l'intervento di un fattore sistematico da identificarsi, in prima approssimazione, in un comportamento differente dei due sessi nei confronti del risparmio.

Posta l'ipotesi $H_0 : \mu_1 = \mu_2 = \mu$ e adottato il modello normale, si deve calcolare la quantità z secondo la (10.8)

$$z = \frac{|5.600 - 4.200|}{668,67\sqrt{\frac{1}{82} + \frac{1}{36}}} = \frac{1.400}{134} = 10,45$$

Se si sceglie il livello $\alpha = 0,05$, i valori soglia corrispondenti a $|z_{0,05}|$ sono $-1,96$ e $+1,96$. Poichè il valore $z = 10,12$ risulta esterno a tale intervallo, l'ipotesi $H_0 : \mu_1 = \mu_2 = \mu$, viene respinta. La probabilità di ottenere, in questa ipotesi, un risultato $z = 10,12$ è infatti inferiore al

5 per cento; essendo 10,12 esterno anche alle soglie $z = \pm 3$, la probabilità è addirittura inferiore al 2 per cento.

Nei piccoli campioni, la variabile aleatoria standardizzata “differenza tra due medie” campionarie non si distribuisce come una gaussiana, bensì come una “studentiana” con gradi di libertà: $\nu = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$. Infatti la ridotta dimensione dei campioni non consente di stimare asintoticamente la varianza. Pertanto il denominatore della T è dato dalla variabile aleatoria varianza campionaria, e quindi nella:

$$T = \frac{|\bar{X}_1 - \bar{X}_2|}{S\sqrt{1/n_1 + 1/n_2}} \quad (10.9)$$

Se stimiamo S^2 con:

$$s^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

La statistica-test è allora

$$t = \frac{m_1 - m_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (10.10)$$

Il test t vale nel confronto tra le medie di due campioni provenienti da popolazioni con varianze uguali. Se non si può ammettere l’omoschedasticità nelle popolazioni, il test non garantisce una risposta coerente rispetto alla soglia α di probabilità.

Esempio 10.4.2.2

Si immagini che l’impresa produttrice di automobili di cui all’esempio 10.2.1, voglia valutare le eventuali differenze nel consumo di benzina “senza piombo” sullo stesso modello di utilitaria quando sia dotato di marmitta catalitica. In un campione di 10 autovetture con tale marmitta risultano una media $m = 16,9 \text{ km/l}$ e uno scarto quadratico medio $s = 2,9 \text{ km/l}$. La differenza rispetto al precedente campione può ritenersi significativa?

L’ipotesi da sottoporre a controllo è ancora $H_0 : \mu_1 = \mu_2 = \mu$. L’ipotesi alternativa potrebbe essere $H_1 : \mu_1 > \mu_2$.

La statistica-test risulta:

$$t_c = \frac{|17,4 - 16,9|}{2,72\sqrt{\frac{1}{15} + \frac{1}{10}}} = \frac{0,5}{1,12} = 0,45$$

essendo:

$$s = \sqrt{\frac{(2,6)^2 \times 14 + (2,9)^2 \times 9}{23}} = 2,72.$$

Posto il livello di significatività $\alpha = 0,05$, il valore teorico t_t , riportato nella tavola II dell'Appendice, per $(15 - 1) + (10 - 1) = 23$ gradi di libertà, risulta $t_{0,05(23)} = 2,069$. Non essendo $t_c > t_t$ non c'è motivo di rifiutare l'ipotesi che i campioni provengano da una produzione "standardizzata" su un identico consumo medio. Leggendo il risultato sulla falsariga dei test d'ipotesi, H_0 viene accettata ed è perciò rifiutata l'eventuale ipotesi alternativa. Ovviamente, se l'ipotesi H_0 non è rifiutata al livello $\alpha = 0,05$, resta implicitamente non rifiutata a qualunque livello α inferiore a tale soglia.

L'uso del test è corretto, essendo le due popolazioni omoschedastiche (l'ipotesi $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ non è infatti da rifiutare).

Esempio 10.4.2.3

Si intende valutare il rendimento medio in quintali per ettaro di due qualità diverse, A e B , di grano duro. Un terreno agricolo sufficientemente omogeneo è stato suddiviso in 50 zone sperimentali; ciascuna qualità di grano è poi stata seminata in 25 di queste zone individuate secondo un piano di campionamento casuale. La qualità A di grano duro, in media, ha reso 19,1 q/ha , con uno scarto quadratico medio $s = 2,4$, mentre la qualità B ha reso 22,8 q/ha con uno scarto quadratico medio $s = 2,7$.

Calcolato lo scarto quadratico medio:

$$s = \sqrt{\frac{(2,4)^2 \times 24 + (2,7)^2 \times 24}{48}} = 2,55$$

si ottiene

$$t_c = \frac{|19,1 - 22,8|}{2,55\sqrt{1/25 + 1/25}} = \frac{3,7}{0,72} = 5,13.$$

Scelto, ad esempio, un livello di significatività $\alpha = 0,001$, il valore teorico della t , riportato nella tavola dell'Appendice, per $(25-1)+(25-1) = 48$ gradi di libertà, risulta $t_{0,001(48)} = 3,5$ ed è superato dal valore t_c . Pertanto, essendo $t_c > t_t$, si può rifiutare l'ipotesi $H_0 : \mu_A = \mu_B = \mu$, che i campioni provengano da popolazioni con uguale rendimento medio. L'ipotesi risulta ovviamente confutata per tutti i valori di α maggiori di 0,001. Se, coerentemente con la teoria dei test d'ipotesi, è stata formulata una ipotesi alternativa H_1 , questa, qualunque essa sia, resta accettata.

10.5 | Test per il confronto fra due frequenze.

Date due frequenze relative campionarie tra loro diverse, ci si può chiedere se la differenza osservata dipende dal solo errore di campionamento o può essere attribuita ad almeno una circostanza differenziante.

Per rispondere, la teoria della significatività pone sotto controllo l'ipotesi $H_0 : p_1 = p_2 = p$ e ne deduce gli esiti possibili, ossia le proporzioni da attendersi in assenza di differenze tra le due popolazioni; è come postulare che la differenza tra le frequenze campionarie osservate, f_1 e f_2 , sia tutta attribuibile all'“errore di campionamento”.

Per il calcolo del test di significatività occorre standardizzare la differenza tra le frequenze relative rispetto allo scarto quadratico medio della variabile “differenza tra due frequenze”. Esso è uguale alla radice quadrata della somma delle due varianze $f_1(1 - f_1)/n_1$ e $f_2(1 - f_2)/n_2$

$$s_f = \sqrt{s_{f_1}^2 + s_{f_2}^2} = \sqrt{\frac{f_1(1-f)}{n_1} + \frac{f_2(1-f_2)}{n_2}}.$$

Il test è quindi

$$z = \frac{|f_1 - f_2|}{\sqrt{\frac{f_1(1-f)}{n_1} + \frac{f_2(1-f_2)}{n_2}}}. \quad (10.11)$$

Esempio 10.5.1

In un campione casuale C_1 di 32 punti di vendita di una grande catena commerciale, 17 accettano i pagamenti con tessera “bancomat”;

in un altro campione casuale C_2 di 47 punti di vendita di una azienda concorrente, 29 accettano tale tessera. I risultati sono indicativi di differenti strategie aziendali tese a favorire l'uso di questa forma di pagamento?

Essendo $f_1 = 17/32 = 0,53$ e $f_2 = 29/47 = 0,62$, si calcola il valore della radice

$$s = \sqrt{\frac{0,53 \times 0,47}{32} + \frac{0,62 \times 0,38}{47}} = 0,006.$$

E quindi il test è

$$z = \frac{|0,53 - 0,62|}{0,006} = \frac{0,09}{0,006} = 15.$$

Il valore calcolato $z_c = 15$ è esterno all'intervallo ± 3 , corrispondente a $\mu \pm 3\sigma$ in una normale non standardizzata. Pertanto la probabilità che per il solo effetto della variabilità campionaria si avveri una differenza tra le frequenze pari a quella osservata è minore del 2,7 per mille ($1 - 0,9973 = 0,0027$). Nella logica del test si può pertanto rifiutare, a un livello di significatività $\alpha < 0,002$, l'ipotesi $H_0 : p_1 = p_2 = p$ e accettare l'ipotesi alternativa H_1 che le differenze osservate siano conseguenti a diverse scelte aziendali.

10.6 | Test per il confronto fra due varianze.

L'ipotesi $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$, postula l'omoschedasticità di due popolazioni rispetto a un dato carattere X e può essere saggiata, su due campioni, mediante la *statistica-test* F di Fisher-Snedecor, espressa dal rapporto tra le due varianze campionarie:

$$F = \frac{s_1^2}{s_2^2} \quad (10.12)$$

(essendo $s_1^2 > s_2^2$).

La regola è di rifiutare l'ipotesi H_0 se il quoziente tra le due varianze supera, nella distribuzione teorica, la soglia F_{α, ν_1, ν_2} scelta in ragione di un livello α di significatività e della particolare combinazione dei gradi

di libertà (ν_1 e ν_2). (I valori teorici della F sono riportati nelle tavole dell'Appendice).

Poiché la somma dei quadrati di n variabili normali standardizzate tra loro indipendenti si distribuisce secondo la $f(\chi^2)$ con $(n-1)$ gradi di libertà, dividendo numeratore e denominatore del rapporto per σ^2 si ha:

$$F = \frac{\sum_j ({}_1X_j - \mu_1)^2}{\sigma_1^2(n_1 - 1)} \bigg/ \frac{\sum_j ({}_2x_j - m_2)^2}{\sigma_2^2(n_2 - 1)} = \frac{\chi^2(n_1 - 1)}{\chi^2(n_2 - 1)}.$$

La distribuzione di F cambia in ragione dei gradi di libertà. Ammessa l'ipotesi H_0 , il quoziente del rapporto

$$\frac{\sum_j ({}_1X_j - \mu_1)^2 / \sigma_1^2(n_1 - 1)}{\sum_j ({}_2x_j - m_2)^2 / \sigma_2^2(n_2 - 1)}$$

stimato dal rapporto

$$F = \frac{\sum_{j=1}^{n_1} ({}_1x_j - m_1)^2 / s_1^2(n_1 - 1)}{\sum_{j=1}^{n_2} ({}_2x_j - m_2)^2 / s_2^2(n_2 - 1)}$$

si approssima a 1.

La condizione per poter usare il test F è che il carattere X si distribuisca come una normale con media μ e varianza σ^2 .

Esempio 10.6.1

Nell'esempio 10.4.2.2, circa la variabilità dei consumi di benzina osservati nei due campioni di automobili dotate, rispettivamente, di marmitte normali e catalitiche, sinteticamente espressi dalle varianze $s_1^2 = 2,6$ e $s_2^2 = 2,9$, è opportuno accertare se la differenza tra le medie possa in qualche modo risentire della diversa variabilità interna ai due gruppi (eteroschedasticità).

Si deve preliminarmente saggiare l'ipotesi di omoschedasticità $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ confrontando il quoziente:

$$F_c = \frac{(2,9)^2}{(2,6)^2} = \frac{8,41}{6,76} = 1,24$$

con il valore teorico F_t riportato nella Tavola IV dell'Appendice. Scelta una soglia di significatività $\alpha = 0,05$ ed essendo $\nu_2 = (10 - 1) = 9$ i gradi di libertà della varianza maggiore (a numeratore) e $\nu_1 = (15 - 1) = 14$ quelli della varianza minore (a denominatore), il valore teorico si trova nella tavola della F per $\alpha = 0,05$, all'incrocio tra $\nu_2 = 9$ e $\nu_1 = 14$ gradi di libertà. Esso risulta $F_{0,05(9,14)} = 3,03$. Il valore calcolato F_c non supera il valore teorico F_t e pertanto non sussiste motivo di rifiutare l'ipotesi di omoschedasticità. La differenza di variabilità interna ai due gruppi non sembra esorbitare dall'errore di campionamento.

Anche nel confronto tra i rendimenti medi in quintali per ettaro di due diverse qualità di grano duro (sempre nel paragrafo precedente), è opportuno saggiare l'ipotesi di omoschedasticità delle varianze nelle popolazioni. Si calcola il test F , ponendo a numeratore la varianza maggiore $(2,7)^2$ e a denominatore quella minore $(2,4)^2$:

$$F_c = \frac{(2,7)^2}{(2,4)^2} = \frac{7,29}{5,76} = 1,27$$

Il valore teorico F_t , letto nella Tavola IV dell'Appendice in corrispondenza di 24 e 24 gradi di libertà, per $\alpha = 0,05$, risulta $F_{0,05(24,24)} = 1,98$. Poiché il valore concreto non supera il valore teorico, non si rifiuta l'ipotesi $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ e si può assumere che i due campioni provengano da popolazioni con uguale variabilità.

10.7 | Test per il confronto fra più medie aritmetiche: l'analisi della varianza.

Un campione di n unità statistiche viene suddiviso in k sottocampioni sulla base delle modalità di un carattere assunto a criterio di classificazione; si individuano in tal modo k sottocampioni indipendenti, determinazioni empiriche di altrettante sottopopolazioni.

Si ipotizza che, con riferimento alla distribuzione del carattere Y , le k sottopopolazioni, da cui i sottocampioni sono stati estratti, siano:

- a) normali,
- b) omoschedastiche,
- c) tra loro indipendenti.

Calcolata in ciascun sottocampione la media aritmetica del carattere Y si avrà, in generale:

$$m_1 \neq m_2 \neq \dots \neq m_i \neq \dots \neq m_k.$$

Ci si chiede se la differenza fra le medie dei campioni sia attribuibile al solo errore di campionamento o se invece l'aver suddiviso il campione iniziale in k sottocampioni, in ragione di un assegnato criterio di classificazione, abbia introdotto una differenza sistematica fra i valori medi.

L'ipotesi nulla sotto controllo può essere posta nel modo seguente:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_i = \dots = \mu_k = \mu$$

I metodi per il controllo di questa ipotesi nulla sono dovuti a R.A. Fisher e sono noti con il nome di Analisi della Varianza (in inglese Analysis of Variance, da cui la sigla AN.O.VA).

In un insieme suddiviso in k sottocampioni si consideri la scomposizione della devianza di un carattere Y in due quantità, usualmente definite come devianza entro i sottogruppi e devianza tra i sottogruppi:

$$Dev_{tot} = Dev_{entro} + Dev_{tra}$$

cioè

$$\sum_i^k \sum_j^{n_i} (y_{ij} - m)^2 = \sum_i^k \sum_j^{n_i} (y_{ij} - m_i)^2 + \sum_i^k (m_i - m)^2 n_i$$

cui competono rispettivamente $(n-1)$, $(n-k)$ e $(k-1)$ gradi di libertà con:

$$(n - 1) = (n - k) + (k - 1)$$

La varianza entro:

$$\frac{\sum_i^k \sum_j^{n_i} (y_{ij} - m_i)^2}{n - k}$$

è, per costruzione – vere le condizioni (a), (b) e (c) – stima corretta di σ_y^2 , la varianza del carattere Y nella popolazione.

La varianza tra:

$$\frac{\sum_i^k (m_i - m)^2 n_i}{k - 1}$$

è stima corretta di σ_y^2 , solo se H_0 è vera.

In regime di ipotesi H_0 , varianza entro e varianza tra possono ritenersi entrambe stime corrette e indipendenti dell'incognita varianza σ_y^2 . La valutazione della significatività della differenza fra i k valori medi si traduce allora nella determinazione della significatività della differenza fra due varianze, mediante un test F costruito come rapporto della varianza “tra” alla varianza “entro”:

$$F_c = \frac{\sum_i^k (m_i - m)^2 n_i / k - 1}{\sum_i^k \sum_j^{n_i} (y_{ij} - m_i)^2 / n - k} = \frac{s_{ytra}^2}{s_{yentro}^2} \quad (10.13)$$

Il valore concreto F_c ottenuto sui dati campionari deve essere confrontato con il valore di F_α tabulato per $(k - 1)$ e $(n - k)$ gradi di libertà al livello di significatività prescelto.

10.8 | Test per il confronto fra distribuzioni.

Le procedure per il controllo statistico in campo parametrico, attengono al controllo di ipotesi relative ad uno o più parametri della popolazione presupponendo la conoscenza della legge distributiva (modello) del carattere nella popolazione da cui il campione è estratto.

I test non parametrici servono al contrario dei precedenti a verificare ipotesi concernenti non più i valori dei parametri, ma il modello stesso che caratterizza una o più popolazioni, da cui il nome di test non parametrici.

Una proprietà importante di un test non parametrico è che la distribuzione della statistica test, sotto l'ipotesi nulla, non dipende dal modello distributivo della popolazione.

È facile quindi intuire che i seguenti test per il confronto fra distribuzioni rientrano tra quelli non parametrici.

10.8.1 | Confronto fra una distribuzione empirica e una teorica

Il χ^2 è sicuramente il primo test per l'analisi della significatività dovuto a K. Pearson, esso pone a confronto una distribuzione empirica e una distribuzione teorica relativa ad un medesimo carattere.

Dato un campione di n osservazioni in cui si sia rilevato un carattere X , ci si può chiedere se, nella popolazione da cui esso è stato estratto, il carattere X segua una particolare legge distributiva. Ci si chiede cioè se la distribuzione del carattere nel campione si conformi a un particolare modello distributivo. L'ipotesi nulla che si sottopone a controllo è quindi che, per ogni modalità i , nella popolazione le frequenze relative p_i siano uguali a quelle (p_i^*) desumibili dal modello distributivo ipotizzato:

$$H_0 : p_i = p_i^* \quad \forall i$$

Test χ^2 (un carattere)

Sia k il numero delle modalità del carattere X . Sia n_i la frequenza della generica modalità i -esima ed np_i^* la corrispondente frequenza assoluta attesa, in ragione del modello probabilistico ipotizzato, allora il controllo dell'ipotesi H_0 può essere effettuato mediante la seguente statistica test:

$$\chi_c^2 = \sum_{i=1}^k \frac{(n_i - np_i^*)^2}{np_i^*}. \quad (10.14)$$

Se il numero delle osservazioni è sufficientemente elevato da garantire l'assenza di frequenze teoriche inferiori a 5, questa statistica test

si distribuisce, nell'universo dei campioni, come una variabile χ^2 con $(k-1)$ gradi di libertà, se sono noti a priori i parametri del modello che si suppone si conformi ai dati, o con $(k-\nu-1)$ gradi di libertà - dove ν è il numero dei parametri da cui il modello dipende - se tali parametri vengono stimati con il metodo della massima verosimiglianza a partire dalle osservazioni campionarie.

Il valore χ_c^2 deve essere confrontato con quello desunto dalle tavole della variabile χ^2 in corrispondenza del livello di significatività prefissato e dei gradi di libertà. Se χ_c^2 supera il valore tabulato, l'ipotesi nulla di conformità viene rifiutata: il divario tra distribuzione empirica e distribuzione teorica è troppo elevato per essere attribuito al solo errore di campionamento.

È opportuno che in nessuna classe, la frequenza teorica sia $np_i^* > 5$, altrimenti il χ^2 , può risultare sovrastimato. Questo inconveniente può essere superato riunendo più classi, e sommando le frequenze (ovviamente, a condizione che tale sintesi non alteri il significato investigativo della classificazione). Inoltre, è già stato detto, che la dimensione del campione deve essere sufficientemente ampia, altrimenti la (10.14) non si distribuisce come una $f(\chi^2)$. Conviene, altrimenti adottare appositi test di conformità. Ad esempio il test di Kolmogorov-Smirnov.

□—————

◇ **Osservazione 10.1.** Se ad es. si vuole saggiare il grado di conformità del modello "Normale" ad una data distribuzione osservata, l'ipotesi H_0 sarà:

$$H_0 : L(X) = N(\mu, \sigma)$$

in questo caso np_i^* essendo il modello distributivo di riferimento una Normale, sarà:

$$np_i^* = \frac{n}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

◇ **Test di Kolmogorov-Smirnov.** Se è dato un campione di n unità statistiche in cui si è rilevato un carattere X ordinabile, ma non si vogliono raggruppare le osservazioni in classi o la ridotta numerosità campionaria non consente il ricorso al test χ^2 , per valutare la conformità della distribuzione del campione

a un modello teorico ci si può riferire al test di Kolmogorov-Smirnov.

L'ipotesi da controllare è

$$H_0 : L(X) = F(X)$$

dove $L(X)$ indica la legge distributiva del carattere X nella popolazione e $F(X)$ è il modello distributivo ipotizzato. Il test si basa sul confronto fra la funzione di ripartizione del campione, $S(X)$, e $F(X)$. Il massimo valore assoluto delle differenze fra tali funzioni di ripartizione, in corrispondenza della medesima modalità, costituisce la statistica test:

$${}_1D_c = \max(|S(X) - F(X)|) \quad (10.15)$$

che deve essere confrontata con il valore tabulato (vedi Tavola 6 in Appendice) in ragione del livello di significatività prescelto e della numerosità campionaria.

Se il valore empiricamente determinato supera la massima divergenza che ci si aspetterebbe di ottenere qualora il modello fosse adeguato per descrivere i dati, H_0 viene rifiutata.

Perché i risultati ottenuti in seguito all'uso del test di Kolmogorov-Smirnov siano attendibili è necessario che i parametri della distribuzione di riferimento siano noti a priori e non vengano perciò stimati a partire dai dati campionari.

—————□

Test χ^2 (due caratteri):

Si consideri una popolazione le cui unità siano raggruppate secondo le modalità di due caratteri X e Y di natura qualsiasi, dove $X_1, \dots, X_h, \dots, X_v$ sono le modalità di X ed $Y_1, \dots, Y_i, \dots, Y_u$ quelle di Y .

Come è noto, quando fra due caratteri esiste indipendenza stocastica le frequenze relative all'interno della tavola che contiene la distribuzione congiunta di tali caratteri sono tali che:

$$p_{ih} = p_{i0} \cdot p_{0h}$$

dove p_{i0} e p_{0h} sono le frequenze relative marginali rispettivamente della i -esima riga e della h -esima colonna.

Data una n -pla di osservazioni relative a due caratteri X e Y proveniente da tale popolazione, si vuole controllare se, nella popolazione, esiste indipendenza stocastica (o indipendenza distributiva) tra X e Y , si pone un problema di controllo della conformità di un particolare modello teorico ai dati: il modello dedotto appunto in ipotesi di indipendenza distributiva.

L'ipotesi di indipendenza che si vuole sottoporre a verifica è espressa pertanto da:

$$H_0 : p_{ih} = p_{i0} \cdot p_{0h}, \quad \forall i, h$$

Per quanto detto, quando fra due caratteri esiste indipendenza distributiva le frequenze all'interno della tavola che contiene la distribuzione congiunta di tali caratteri si deducono da:

$$n_{ih}^* = \frac{n_{i0} \cdot n_{0h}}{n} \quad (10.16)$$

dove n_{i0} e n_{0h} sono i totali marginali rispettivamente della i -esima riga e della h -esima colonna. L'allontanamento della distribuzione empirica dalla condizione di indipendenza viene quindi valutato, ancora una volta, attraverso la statistica χ^2 che, per il problema in esame, assume la forma:

$$\chi_c^2 = \sum_{i=1}^u \sum_{h=1}^v \frac{(n_{ih} - n_{ih}^*)^2}{n_{ih}^*} \quad (10.17)$$

dove n_{ih} , indica il numero di unità che portano simultaneamente la modalità i del carattere X e la modalità h del carattere Y e u e v sono, rispettivamente, il numero di righe e il numero di colonne della distribuzione doppia definita da X e Y .

Se il numero delle osservazioni è sufficientemente elevato da garantire l'assenza di frequenze teoriche inferiori a 5, questa statistica test si distribuisce, nell'universo dei campioni, come una variabile χ^2 con $(u - 1) \cdot (v - 1)$ gradi di libertà. Il valore χ_c^2 deve essere confrontato con quello desunto dalle tavole della variabile χ^2 in corrispondenza del livello di significatività prefissato e dei gradi di libertà. Se χ_c^2 supera il valore tabulato l'ipotesi nulla di indipendenza in distribuzione fra X e Y viene rifiutata: il divario tra distribuzione empirica e distribuzione di indipendenza è troppo elevato per essere attribuito al solo errore di

campionamento.

□—————

◇ *Osservazione 10.2.- Confronto fra due o più distribuzioni empiriche.* Dati due campioni indipendenti C_1 e C_2 , si può voler confrontare la distribuzione in essi di un carattere X per controllare se i due campioni provengano da una medesima popolazione, o meglio se la legge distributiva del carattere oggetto di interesse nelle due popolazioni da cui i campioni provengono sia la medesima. L'ipotesi nulla da controllare è allora che, per ogni modalità i del carattere osservato, valga:

$$H_0 : p_{i1} = p_{i2} = p_i \quad \forall i$$

Se le osservazioni sono raggruppate in classi e se la numerosità dei campioni è sufficientemente elevata, il controllo dell'ipotesi nulla può essere effettuato mediante il test χ^2 analogamente a quanto fatto in precedenza. Infatti, se nelle due popolazioni il carattere si distribuisce allo stesso modo, nei due campioni le distribuzioni del carattere devono essere simili - cioè avere frequenze relative uguali in corrispondenza di uguali modalità - o allontanarsi da tale condizione per il solo effetto dell'errore di campionamento. Se si riuniscono i due campioni in modo da formare una distribuzione doppia, in cui le righe sono definite dalle modalità del carattere X e le colonne identificano ciascun campione, la distribuzione per cui vale la condizione di similitudine delle distribuzioni parziali del carattere X coincide con quella di indipendenza stocastica. Il controllo dell'ipotesi di identità distributiva del carattere X nelle due popolazioni passa pertanto attraverso il controllo della significatività della divergenza fra la distribuzione doppia empirica e quella teorica costruita in base alla (10.18).

Il confronto avviene mediante la statistica (10.19) che si distribuisce come un χ^2 con $(u - 1) \cdot (v - 1)$ gradi di libertà (dove u è il numero delle modalità classificatorie del carattere X e v è il numero di gruppi). L'espressione della statistica test mostra come il suo impiego possa essere esteso al confronto fra un numero qualsiasi di distribuzioni. Se il carattere X presenta

due sole modalità e i campioni a confronto sono due, il test χ^2 perfettamente equivalente al test normale illustrato per il confronto fra due frequenze; tuttavia l'espressione generale della statistica test χ^2 mostra come il suo impiego possa essere esteso al confronto fra un numero qualsiasi di frequenze.

Se le osservazioni non sono raggruppate in classi, ma il carattere è ordinabile, o se il numero di unità attese risulta in qualche cella inferiore a 5 e quindi non sussistono le condizioni per la convergenza della statistica test a una variabile χ^2 , per il confronto fra due distribuzioni risulta opportuno ricorrere al test di Kolmogorov-Smirnov.

◇ *Osservazione 10.3.- Test di Kolmogorov-Smirnov per il confronto fra due o più distribuzioni empiriche.* L'ipotesi nulla che con tale test viene controllata è che nelle due popolazioni da cui i campioni sono stati estratti la funzione di ripartizione del carattere in esame sia la medesima. Indicata rispettivamente con $F_1(X)$ l'incognita funzione di ripartizione del carattere X nella popolazione da cui il campione C_1 è stato estratto e con $F_2(X)$ l'incognita funzione di ripartizione del medesimo carattere nella popolazione da cui proviene il campione C_2 , l'ipotesi nulla può essere posta come:

$$H_0 : F_1(X) = F_2(X) = F(X).$$

Il test per il controllo di questa ipotesi comporta inizialmente la determinazione, per i due campioni, delle funzioni di ripartizione del carattere in esame, indicate rispettivamente con $S_1(X)$ e $S_2(X)$ e, successivamente, la costruzione della statistica test:

$${}_2D_c = \max (|S_1(X) - S_2(X)|) \quad (10.18)$$

e il suo confronto con i valori che, in ragione dell'assegnato livello di significatività e delle numerosità campionarie, sono riportati nelle Tavole in Appendice.

Se infatti i due campioni provengono da popolazioni aventi la medesima funzione di ripartizione, le due funzioni di ripartizione campionarie sono poco discoste e quindi ci si attende che ${}_2D_c$ assuma valori bassi; se viceversa le funzioni di ripartizione nelle popolazioni sono diverse, ${}_2D_c$ potrà assumere anche valori

molto elevati. Se quindi la statistica ${}_2D_c$ assume valori maggiori di quelli che ci si potrebbe attendere qualora H_0 fosse vera, si decide per il rifiuto dell'ipotesi nulla.

—————□

10.9 | Test non parametrici.

Le procedure per il controllo statistico delle ipotesi trattate in questo paragrafo compaiono in letteratura con il nome di test non parametrici.

Esse attengono al controllo di ipotesi su quantità che non corrispondono ai parametri che indicizzano particolari famiglie distributive, e non presuppongono la conoscenza della legge distributiva del carattere nella popolazione da cui il campione è stato estratto. Appartengono a questo gruppo di procedure anche i test volti al controllo della conformità di un assegnato modello distributivo a un campione, o al confronto fra le leggi distributive di un medesimo carattere in due o più campioni indipendenti.

10.9.1 | Ipotesi su una frequenza relativa

In un campione casuale di numerosità n un certo evento si è avverato con frequenza relativa f . Ci si chiede se, sulla scorta dei risultati campionari, sia verosimile che la probabilità p del verificarsi di quell'evento nella popolazione sia uguale a un assegnato valore p_0

$$H_0 : p = p_0$$

Come è noto, il numero X di volte in cui un evento, la cui probabilità di accadimento è p_0 si verifica in n prove, si distribuisce come una binomiale con media np_0 e varianza $np_0(1 - p_0)$. La frequenza relativa dell'evento in n prove, cioè $Y = X/n$, si distribuisce, pertanto, con media p_0 e varianza $p_0(1 - p_0)/n$.

Test binomiale

L'ipotesi nulla $H_0 : p = p_0$ viene rifiutata se la probabilità che l'evento si presenti con frequenza relativa f o con frequenze a essa

più estreme è, vera H_0 , inferiore al livello di significatività α prefissato. La distribuzione cumulata della binomiale consente di ottenere la probabilità che l'evento oggetto di interesse si presenti, in n prove, con frequenza relativa inferiore o uguale a f o, equivalentemente, un numero di volte inferiore a $k = nf$:

$$P(X \leq k) = \sum_{x=0}^k \binom{n}{x} \cdot p_0^x (1 - p_0)^{n-x} \quad (10.19)$$

Si osservi che, su H_0 , si dovrà decidere in base al confronto tra la probabilità del verificarsi dell'esito campionario o di esiti a esso più estremi e il livello di significatività α e non, come accadeva nel capitolo precedente, tra il valore concreto della statistica test e il valore tabulato che stacca sulle code (o sulla coda) della distribuzione campionaria della statistica test stessa una probabilità pari ad α . Questo perché la binomiale è una variabile aleatoria discreta e quindi è possibile che non esista affatto un valore tale che la probabilità di eventi a esso uguali o più estremi sia esattamente uguale ad α .

Test normale

Il test appena illustrato si adotta solo per piccoli campioni. Infatti, quando la numerosità campionaria aumenta, la binomiale converge a una normale con la stessa media e la stessa varianza, consentendo così il ricorso al test normale. La statistica test diventa allora:

$$z_c = \frac{f - p_0}{\sqrt{p_0(1 - p_0)/n}} \quad (10.20)$$

10.9.2 | Confronto fra due frequenze

Quando si vogliono confrontare le frequenze relative con cui un certo evento si è manifestato in due campioni indipendenti C_1 e C_2 di numerosità n_1 e n_2 e ci si chiede se l'eventuale differenza tra esse sia imputabile al solo errore di campionamento, l'ipotesi nulla sotto controllo può essere formalizzata nel modo seguente:

$$H_0 : p_1 = p_2 = p,$$

dove p_1 e p_2 sono rispettivamente le frequenze relative con cui l'evento si manifesta nelle popolazioni da cui i due campioni sono stati estratti.

Per piccoli campioni la legge distributiva della variabile aleatoria differenza fra due variabili aleatorie frequenza relativa campionaria non è nota e quindi il controllo di ipotesi si deve avvalere di procedure fondate su criteri diversi da quelli fin qui descritti. Allo scopo risponde il test della probabilità esatta o test esatto di Fisher.

Test esatto di Fisher

Le informazioni che si desumono dai campioni possono essere riassunte in una distribuzione doppia 2×2 , in cui le righe identificano i campioni e le colonne la presenza (1) o l'assenza (0) dell'evento oggetto di interesse:

	1	0	
C_1	A	B	A+B
C_2	C	D	C+D
	A+C	B+D	

(A corrisponde alla quantità che nel capitolo 4 è stata indicata con n_{11} , B a n_{12} , ecc. La simbologia è stata qui modificata per adeguarsi a quella comunemente impiegata per il test esatto).

Una volta fissate le marginali, la probabilità di una data configurazione nella tavola si ottiene attraverso l'espressione dell'ipergeometrica (paragrafo 6.3), che consente di determinare la probabilità della configurazione osservata e di quelle a essa più estreme. Per ovviare alla complessità calcolatoria del test sono state predisposte apposite tavole. La statistica test è rappresentata da C (o D), il numero di volte in cui, nel campione C_2 , si è presentato (o non si è presentato) l'evento oggetto di interesse. Se il valore di C è minore o uguale a quello riportato nella Tavola 5 dell'Appendice, per assegnati valori di $A + B$, $C + D$ e A (o B) al livello di significatività prescelto, si rifiuta l'ipotesi nulla.

Test normale

Per grandi campioni, invece, in virtù della convergenza della binomiale alla normale, la distribuzione della variabile aleatoria differenza fra due variabili aleatorie indipendenti frequenza relativa campionaria è facilmente derivabile: essa è una normale avente per media la diffe-

renza fra le due medie e per varianza la somma delle varianze. Sotto $H_0 : p_1 = p_2 = p$ si ottiene quindi la statistica test:

$$z_c = \frac{f_1 - f_2}{\sqrt{f(1-f)(n_1 + n_2)/n_1 n_2}} \quad (10.21)$$

(dove f è una stima dell'incognita p ottenuta come $f = \frac{f_1 n_1 + f_2 n_2}{n_1 + n_2}$ e f_1 e f_2 sono le frequenze relative dell'evento oggetto di interesse, nei due campioni) il cui valore va confrontato con quello desunto dalla tavola della normale (in Appendice) al livello di significatività prescelto.

10.9.3 | Ipotesi sulla tendenza centrale di una popolazione

Nel capitolo precedente si è visto che, per piccoli campioni, il controllo di una ipotesi nulla su un valor medio (o sulla differenza fra due valori medi), quando non sussiste la condizione di normalità distributiva del carattere oggetto di interesse nella popolazione, non può essere effettuato tramite il test t . In questa situazione, o quando i dati da analizzare sono ranghi (posti d'ordine) e non variabili quantitative continue, si deve far ricorso a procedure alternative che richiedano assunzioni meno restrittive. A tal scopo può soccorrere il test dei segni, incentrato sulla mediana piuttosto che sulla media, il cui nome deriva dal fatto che i dati grezzi su cui si opera non sono valori numerici, ma soltanto segni $+$ e $-$.

L'ipotesi nulla da controllare è che la mediana della popolazione da cui il campione proviene, Med , sia uguale a un assegnato valore Med_0 . In simboli:

$$H_0 : Med = Med_0$$

Test dei segni

Se l'ipotesi nulla è vera, se cioè la mediana della popolazione da cui il campione è stato estratto è veramente Med_0 , ci si aspetta che, in un campione di n unità statistiche, il numero dei valori che si trovano al di sotto di Med_0 (contrassegnati con un segno $-$) eguagli approssimativamente il numero di quelli che si trovano al di sopra di essa (contrassegnati con un segno $+$). Ciò consente di porre l'ipotesi nulla anche come:

$$H_0 : P(+) = P(-) = 0,5,$$

cioè la probabilità di un + è uguale alla probabilità di un -, ed è quindi uguale a 0,5.

Già si è visto, nel paragrafo 10.1, che una ipotesi di questo tipo può essere controllata, per piccoli campioni, attraverso il test binomiale. Nel particolare problema in esame, indicato con k il numero dei segni meno frequenti, la distribuzione campionaria di k , in regime di ipotesi H_0 , è la distribuzione binomiale di parametro $p = 0,5$.

Si può perciò determinare la probabilità di avere un numero di segni di un certo tipo minore o uguale a k , in un campione di numerosità n , attraverso la Statistica test (10.1) e agire come indicato per il test binomiale.

Se $n \geq 25$ ci si può avvalere della convergenza della binomiale alla normale e quindi riferirsi alla (10.2).

Il test dei segni sfrutta solo una minima parte delle informazioni contenute nell'insieme dei dati: si limita infatti a valutare se ciascuna osservazione si trova al di sopra o al di sotto della mediana specificata in H_0 . Ha però il vantaggio di essere applicabile a campioni di osservazioni su caratteri ordinabili e provenienti da popolazioni di qualsiasi forma. Se si può ipotizzare che la distribuzione del carattere nella popolazione sia simmetrica e se il carattere è continuo, allora il centro di simmetria della popolazione è la mediana (o la media aritmetica, perché esse coincidono) e si può fare inferenza ricorrendo a strumenti che impieghino in modo più completo i dati.

Test di Wilcoxon

In questo contesto, dato un campione di n osservazioni indipendenti, si può considerare, oltre al segno degli scostamenti di ciascuna osservazione dalla mediana (o dalla media) specificata in H_0 : $Med = Med_0$, anche l'entità di ciascuno scostamento. Se H_0 specifica la vera mediana (o media) di una distribuzione simmetrica, allora gli scostamenti di una data intensità sono equiprobabili nei due sensi. Così, se si dispongono in ordine di grandezza tutti gli scostamenti dalla mediana esplicitata in H_0 e a essi si sostituiscono i rispettivi ranghi (1 per lo scostamento più piccolo e n per quello più grande), ponendo poi un segno meno davanti ai ranghi che corrispondono a valori inferiori alla

mediana, ci si aspetta, se H_0 è vera, una notevole dispersione dei segni più e meno. Inoltre, se si sommano rispettivamente i ranghi positivi (S_p) e quelli negativi (S_n) ci si aspetta che tali somme siano pressoché identiche. Una sproporzione fra i valori delle due somme indica che il valore specificato in H_0 non è la vera mediana della popolazione.

La statistica test, che porta al cosiddetto test di Wilcoxon, è rappresentata (per un test a due code) dalla quantità più piccola fra S_n e S_p :

$$S = \min(S_n, S_p) \quad (10.22)$$

(il caso di un test a una coda verrà illustrato negli esercizi).

La Tavola in Appendice riporta il massimo valore che, in un campione di n unità statistiche, tale quantità può assumere per portare al rifiuto di H_0 . Valori della statistica test inferiori al valore tabulato conducono pertanto al rifiuto di H_0 .

Se si presentano osservazioni uguali alla mediana ipotizzata, a esse viene attribuito rango zero e le si esclude quindi dall'analisi, riducendo, di conseguenza, la dimensione del campione. Se invece alcune osservazioni compaiono più di una volta, a ciascuna di esse viene attribuito un rango pari alla media aritmetica dei ranghi che a tali osservazioni sarebbero stati attribuiti se esse fossero state diverse.

Se $n \geq 25$, anche per il test di Wilcoxon vale una approssimazione alla normale standardizzata nella forma:

$$z_c = \frac{S_c - \frac{1}{4}n(n+1)}{\sqrt{n(n+1)(2n+1)/24}} \quad (10.23)$$

10.9.4 | Confronto tra le tendenze centrali di due campioni

Campioni appaiati

Si consideri la situazione sperimentale descritta nel paragrafo 9.3 *Confronto fra due valori medi. Campioni appaiati*. Se non si può postulare la normalità distributiva della variabile differenza si rende necessario ricorrere a un metodo di controllo di ipotesi alternativo al test t , per il quale la condizione di normalità è requisito essenziale. L'attenzione

si sposta sulla mediana della distribuzione delle differenze. Se il fattore sperimentale non ha alcun effetto, la differenza, per ogni coppia di unità (l'una appartenente a un campione e l'altra, a essa associata, appartenente all'altro campione), rispetto al carattere oggetto di interesse è solo accidentale. La mediana dell'ipotetica popolazione di differenze Med_D è dunque 0. L'ipotesi nulla può quindi porsi come:

$$H_0 : Med_D = 0.$$

Si ritorna quindi, anche in questo contesto, al problema del controllo di una ipotesi sulla tendenza centrale di una popolazione, e quindi si può fare ricorso al test dei segni o al test di Wilcoxon (qualora si possa ipotizzare la simmetria della distribuzione delle differenze) illustrati nel paragrafo precedente.

Campioni indipendenti

Quando si pone il problema del confronto fra le tendenze centrali di due campioni indipendenti e non sussistono i presupposti formali per il ricorso al test t , ci si può avvalere del test della mediana o del test di Wilcoxon-Mann-Whitney. I prerequisiti che devono essere soddisfatti affinché il ricorso al test della mediana risulti giustificato sono meno rigidi di quelli che vengono generalmente posti per la validità dei test parametrici. Si richiede infatti soltanto che i campioni siano casuali e indipendenti e che la variabile oggetto di interesse sia almeno ordinabile.

Test della mediana

Dati due campioni C_1 e C_2 di numerosità n_1 e n_2 , se essi provengono da popolazioni aventi la stessa mediana, le loro mediane sono entrambe stime dell'incognita mediana comune Med . In questo contesto, se si uniscono i due campioni e si determina la mediana dell'insieme di osservazioni così ottenuto, essa sarà ancora una stima di Med , sotto l'ipotesi nulla:

$$H_0 : Med_1 = Med_2 = Med.$$

Se è vera H_0 , allora approssimativamente metà delle osservazioni di ciascun campione sarà al di sotto della mediana comune e metà al di

sopra; al contrario, se i due campioni provengono da due popolazioni aventi diversa mediana, in un campione si avrà una preponderanza di valori al di sopra della mediana e nell'altro un prevalere delle osservazioni al di sotto.

Le informazioni che si traggono in seguito alle operazioni sopra descritte possono essere riassunte in una distribuzione doppia con due righe - che identificano la condizione di trovarsi sopra o sotto la mediana - e due colonne - che corrispondono ai due campioni.

Dire che, se H_0 è vera, ci si aspetta nei due campioni un ugual numero di valori sopra e sotto la mediana comune equivale a dire che le distribuzioni del carattere sopra o sotto la mediana sono, nei due campioni, tra loro simili e simili alla marginale e quindi che esiste indipendenza in distribuzione fra il carattere che definisce le righe e quello che definisce le colonne. Per controllare l'ipotesi che i due campioni provengano da due popolazioni aventi uguale mediana si può dunque adottare il test χ^2 (o il test esatto), confrontando la distribuzione empirica con quella teorica dedotta dallo schema di indipendenza. Il valore χ_c^2 andrà dunque confrontato con quello che si trova sulla tavola della variabile χ^2 (Tavola in Appendice) in corrispondenza di 1 grado di libertà e del livello di significatività prescelto.

Test di Wilcoxon-Mann-Whitney

Il test della mediana non utilizza in modo globale le informazioni contenute nei due campioni quando il carattere è una mutabile ordinabile o una variabile: esso riduce infatti il contenuto di informazione di una misura al solo fatto che essa sia al di sopra o al di sotto della comune mediana. Quando è possibile sostituire a ogni osservazione il rango da essa occupato e si può ipotizzare che, oltre alle condizioni poste per il test della mediana, le popolazioni da cui i campioni vengono estratti siano della stessa forma e differiscano al più per la tendenza centrale, può risultare opportuno confrontare le mediane dei due campioni indipendenti sulla base di un test proposto da Wilcoxon e da Mann e Whitney secondo due formulazioni diverse, che conducono però ai medesimi risultati.

La logica sottostante il test è molto simile a quella illustrata per il test di Wilcoxon relativo al controllo di ipotesi su una mediana. Invece di attribuire ai ranghi un segno uguale a quello della differenza fra il valore corrispondente e la mediana e sommare i ranghi di ugual segno,

si riuniscono le osservazioni dei due campioni in un unico campione e si attribuisce a ciascuna di esse un rango; si sommano poi i ranghi relativi a ciascun campione ottenendo le quantità S_1 e S_2 . Se i due campioni provengono da un'unica popolazione, ci si aspetta di trovare, in ciascuno di essi, un misto di osservazioni aventi rango basso, medio ed elevato. Se i due campioni provengono da popolazioni diverse rispetto alla mediana, ci si attende invece un prevalere di ranghi bassi in un campione e di ranghi elevati nell'altro.

In analogia con quanto visto per un solo campione, la statistica test è rappresentata dalla più piccola fra S_1 e S_2 .

Spesso le tavole sono riferite alle statistiche di Mann-Whitney U_1 e U_2 , che sono semplici trasformazioni monotone di S_1 e S_2 . È infatti $U_1 = S_1 - \frac{1}{2}n_1(n_1 + 1)$ e $U_2 = S_2 - \frac{1}{2}n_2(n_2 + 1)$. E quindi la statistica test da utilizzare, per un test a due code, è:

$$U_c = \min(U_1, U_2) \quad (10.24)$$

(il caso di un test a una coda verrà affrontato negli esercizi).

La Tavola in Appendice riporta, per piccoli campioni, la probabilità di valori di U minori o uguali a quello osservato. (Le probabilità di Tavola 9 sono a una coda. Per un test a due code si deve raddoppiare il valore di probabilità riportato nelle tavole.)

Probabilità inferiori al livello di significatività prescelto portano al rifiuto di H_0 . La Tavola 10 in Appendice riporta, invece, il massimo valore che U può assumere, in campioni di numerosità n_1 e n_2 più elevate, affinché il test risulti significativo. Valori di U minori o uguali al valore tabulato portano al rifiuto dell'ipotesi nulla.

Per valori di n_1 e n_2 superiori a 20 si può ricorrere alla espressione asintotica della statistica test:

$$z_c = \frac{U_c - n_1 n_2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1) / 12}} \quad (10.25)$$

che si distribuisce, vera l'ipotesi nulla, come una normale standardizzata.

10.9.5 | Confronto fra le tendenze centrali di $k > 2$ campioni

Qualora si vogliano controllare ipotesi sulle tendenze centrali di più di due popolazioni sulla base di $k \geq 2$ campioni indipendenti e non

si possa postulare la normalità distributiva del carattere oggetto di interesse nella popolazione, il controllo dell'ipotesi nulla:

$$H_0 : Med_1 = Med_2 = \dots = Med_k = Med.$$

non può avvenire in termini di Analisi della varianza.

Test della mediana

Se non si hanno informazioni sulla forma della distribuzione, può essere utile estendere semplicemente il test della mediana al caso di $k \geq 2$ campioni. Si riuniscono pertanto tutte le osservazioni in un unico campione di cui si determina la mediana. Si costruisce poi una distribuzione doppia con due righe - che identificano, rispettivamente, i valori al di sopra e al di sotto della mediana - e tante colonne quanti sono i campioni posti a confronto, contenente, per ciascun campione, il numero delle osservazioni maggiori o minori della mediana comune. (Le osservazioni coincidenti con la mediana vengono escluse dall'analisi.) Si confronta infine tale distribuzione con quella desunta in ipotesi di indipendenza in distribuzione fra la variabile che descrive il campione di appartenenza e quella che descrive la posizione delle unità nei confronti della mediana (secondo la (10.5)), mediante la statistica test χ_c^2 (10.6) che si distribuisce come un χ^2 con $(k - 1)$ gradi di libertà.

Test di Kruskal-Wallis

L'estensione del test di Wilcoxon-Mann-Whitney al caso di $k \geq 2$ campioni è rappresentata dal test di Kruskal-Wallis. Anche in questo contesto si determina il rango di ciascuna osservazione nell'insieme di tutte le osservazioni, prescindendo dal campione di appartenenza. Si ricostruiscono poi i campioni sostituendo a ciascuna osservazione il proprio rango.

In regime di ipotesi nulla ci si aspetta una equiripartizione di ranghi bassi, medi e alti in ciascun campione e quindi, dopo opportuni aggiustamenti volti a tener conto della eventuale diversa numerosità dei campioni, anche valori pressoché simili della somma dei ranghi nei diversi campioni. La statistica test H di Kruskal-Wallis, volta a valutare l'entità della differenza fra le somme dei ranghi dei vari campioni e a giudicare se tale differenza sia verosimile sotto H_0 , assume la forma:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(n+1) \quad (10.26)$$


dove R_i e n_i sono rispettivamente la somma dei ranghi e la numerosità del i -esimo campione e n è la numerosità del campione complessivo. Per piccoli campioni il valore così ottenuto deve essere confrontato con quello tabulato, per le assegnate numerosità campionarie e al livello di significatività prescelto (si veda la Tavola 11 in Appendice per tre campioni, tali che $n_1, n_2, n_3 \leq 5$). Per grandi campioni si può fare riferimento alle soglie della variabile χ^2 con $k - 1$ gradi di libertà. In entrambi i casi, valori empirici superiori a quelli tabulati portano al rifiuto dell'ipotesi nulla al livello di significatività prescelto.

Quando nel campione sono presenti valori ripetuti a essi si sostituisce il loro rango medio e si corregge la statistica test dividendo per $1 - \sum_j T_j / (n^3 - n)$, dove $T_j = t_j^3 - t_j$ e t_j è il numero delle replicazioni del j -esimo insieme di valori ripetuti.

10.10 | Esercizi

10.1 Si vuole controllare se il metodo di campionamento usato per un sondaggio di opinioni sia completamente casuale. Si ha infatti il sospetto che le unità della popolazione non avessero tutte la medesima probabilità di essere selezionate, ma che tale probabilità fossero legate al reddito posseduto. Dai dati del censimento si sa che nella comunità, il reddito familiare medio annuo ammonta a 45 milioni, con uno scarto quadratico medio di 3,5 milioni. Dalla ricerca, effettuata su un campione di 100 famiglie (la cui casualità è indubbia), risulta un reddito familiare medio di 49 milioni. Il sospetto del ricercatore è giustificato, posto $\alpha = 0,05$?

10.2 In un processo produttivo che funziona a regime il peso dei pezzi prodotti si distribuisce normalmente con media 50 g. Se in un campione di pezzi, in cui la media è 52 g e la varianza corretta 25 g², l'ipotesi nulla di un divario casuale fra la media attesa del processo e la media del campione è stata rifiutata al 5% (concludendo così che il processo non funziona a regime), quante unità conteneva almeno il campione?

- 10.3** Un gruppo ambientalista raccoglie campioni di 1 litro d'acqua per 15 zone diverse lungo il corso di un fiume e misura la quantità di ossigeno disciolto in ciascuno di essi. La media è 4,62 mg e lo scarto quadratico medio è $s=0,92$. Si può ragionevolmente affermare, a un livello di significatività del 5%, che, nell'ipotesi di normalità distributiva della quantità di ossigeno disciolto, il fiume ha un contenuto medio in ossigeno inferiore a 5 mg?
- 10.4** Una macchina impacchettatrice di funghi è stata programmata in modo da produrre pacchetti da 500 g. Il peso medio degli ultimi 60 pacchetti prodotti è risultato 487 g con uno scarto quadratico medio corretto di 15 g. Si può ipotizzare che il peso del prodotto sia diverso dai 500 grammi previsti a un livello di significatività $\alpha=0,01$?
-  **10.5** Dato un campione casuale di 9 batterie della marca A con durata media di 20 ore, sapendo che il carattere durata in popolazione è distribuito normalmente con scarto quadratico medio di 3 ore:
- saggiare l'ipotesi nulla che la media in popolazione sia di 22 ad un livello di significatività 0,05;
 - saggiare l'ipotesi al punto a) nel caso in cui sia noto solo lo scarto del campione pari sempre a 3 ore;
 - rappresentare il grafico di una distribuzione normale.
- 10.6** In un campione di 12 famiglie in cui il capofamiglia è laureato si è rilevato un reddito medio annuo m_1 , di 35 milioni di lire, mentre in un campione di 15 famiglie, il cui capofamiglia è diplomato, il reddito medio annuo, m_2 , è di 28 milioni. Ci si chiede se tali risultati consentano di affermare, a un livello di significatività $\alpha = 0,01$, che esiste una differenza nel reddito medio fra famiglie con capofamiglia laureato e famiglie con capofamiglia diplomato o se invece la differenza rilevata nei due campioni possa ragionevolmente essere attribuita al solo errore di campionamento, ipotizzando che, nelle categorie in esame, il reddito si distribuisca normalmente con scarto quadratico medio $\sigma_1 = 4,24$ e $\sigma_2 = 4,94$.
- 10.7** Un'industria dolciaria sta sperimentando due nuove ricette di gelati alla crema, per verificare se il tempo impiegato dai gelati

per sciogliersi differisca a seconda della ricetta utilizzata. A tal fine controlla otto gelati per ogni ricetta. I gelati prodotti con la prima ricetta si sciolgono mediamente dopo 4,13 minuti mentre i gelati prodotti con la seconda si sciolgono mediamente dopo 5,8 minuti, a parità di condizioni esterne. La varianza non corretta dei 16 tempi di scioglimento risulta pari a 0,79.

a) Quali condizioni devono valere affinché l'azienda possa confrontare i due valori medi con il test t di Student?

b) La differenza fra i due valori medi risulta significativa all'1% se le condizioni al punto a) sono soddisfatte?

10.8 In 12 uffici comunali si è valutato il numero di pratiche mediamente sbrigate in un anno prima e dopo aver introdotto una meccanizzazione nel servizio. Ci si chiede se la meccanizzazione abbia introdotto una variazione significativa nel numero medio delle pratiche o se invece sia stata ininfluente.

Prima	Dopo
4921	4800
3910	5260
4220	5470
3885	5640
5160	6180
5645	6390
4680	6515
5265	6805
5975	7515
6790	7515
6900	8230
7335	8770

a) Che condizioni devono valere affinché l'ipotesi possa essere controllata con i metodi finora visti?

b) Se tali condizioni sono soddisfatte il test risulta significativo all'1%?

- 10.9 Dati i seguenti campioni casuali, estratti al fine di analizzare il carattere Y =durata media dei pneumatici prodotti (in Km), da un'impresa nei due turni di lavoro:

	turno di giorno (C_1)	Turno di notte (C_2)
n	100	100
m	40.900 km	37.500 km
s	6.500	4.800

- a) controllare dato un livello di significatività $\alpha = 0,01$, se esiste una differenza significativa di durata tra i pneumatici prodotti dai 2 turni;
- b) supposto che Y si distribuisca normalmente, effettuare il medesimo controllo, dati due campioni di numerosità pari a 10.

- 10.10 Un produttore di cappelli impiega nella sua fabbrica, due diverse macchine, che producono, per ogni taglia, cappelli con la stessa circonferenza media. Egli tuttavia ha il sospetto che i prodotti delle due macchine differiscano quanto a variabilità media. Lo scarto quadratico medio corretto, in un campione di 20 cappelli prodotti con la macchina 1, è risultato $s_1 = 0,9$ cm, mentre quello relativo a un campione di 19 cappelli prodotti con la macchina 2 è $s_2 = 1,2$ cm. I dati confermano il sospetto del produttore, per $\alpha = 0,05$?

Quali condizioni devono sussistere affinché l'ipotesi possa essere controllata seguendo le indicazioni date in questo paragrafo?

- 10.11 Dati i seguenti tre campioni, estratti al fine di analizzare il carattere Y =durata media dei pneumatici prodotti (in Km), da un'impresa nei tre turni di lavoro :

saggiare, nell'ipotesi di normalità distributiva e di omoschedasticità della variabile Y nelle tre popolazioni da cui i campioni provengono, l'ipotesi di ininfluenza del tipo di turno sulla durata dei pneumatici, dato un livello di significatività $\alpha = 0,01$.

- 10.12 Il consumo medio, in litri di benzina, osservato su campioni casuali di veicoli di diversa cilindrata, per percorrere 100Km in ciclo urbano è riportato nella tabella seguente:

	turno 1	turno 2	turno 3
n	5	5	5
	40,87	37,65	32,21
	42,33	35,08	35,72
	38,78	37,81	31,78
	38,2	36,6	33,15
	40,39	34,75	32,82

Tipo di autoveicolo	n_i	Consumo medio
Utilitarie	45	7,8
Media cilindrata	102	10,1
Grossa cilindrata	18	11,7

Lo scarto quadratico medio corretto calcolato su tutti i 165 risultati campionari è $s=1,378$. Si ipotizzino la normalità e l'omoschedasticità delle distribuzioni dei consumi per le tre categorie di veicoli.

- a) Quale test statistico si deve utilizzare per controllare l'ipotesi nulla di uguaglianza dei tempi medi per le tre diverse categorie?
- b) Quale valore assume la statistica test? Il test è significativo all'1%?

10.13 In tre campioni omogenei di vigne coltivate con identici criteri, tranne che per il metodo di potatura, e costituiti rispettivamente da $n_1=20$, $n_2=12$, $n_3=7$ filari, si sono ottenuti i seguenti raccolti medi di uva (in quintali) per filare: $m_1=75$, $m_2=94$, $m_3=112$ con relativi scarti quadratici medi corretti pari a $s_1=35$, $s_2=28$, $s_3=47$.

Nell'ipotesi di normalità distributiva e di omoschedasticità della variabile "raccolto" nelle tre popolazioni da cui i campioni provengono, si è deciso di saggiare l'ipotesi nulla di ininfluenza del tipo di potatura sulla quantità di raccolto mediante una analisi della varianza. Si calcoli il valore della statistica test F e si dica se il test risulta significativo all'1%.

10.14 In un test attitudinale per la selezione del personale, si sottopongono ai candidati 50 domande a risposta binaria. Se il candi-

dato risponde bene a 34 domande è possibile che abbia risposto a caso? Si controlli al livello di significatività del 5%.

10.15 Su un quotidiano sportivo è apparsa la notizia che le persone che il lunedì sera guardano un famoso programma sportivo sono per il 70% uomini. Selezionato un campione casuale di 400 individui che seguono regolarmente tale programma, 220 sono uomini. Che test si dovrebbe usare per valutare se tale risultato contrasta con la notizia citata? Il test risulta significativo al livello di significatività del 5%?

10.16 In due campioni casuali di telespettatori, rispettivamente composti da 1750 maschi e 2300 femmine, le percentuali di coloro che hanno dichiarato di apprezzare una trasmissione sportiva in onda la domenica pomeriggio sono risultate pari all'81% dei maschi e al 77% delle femmine.

Ci si chiede se la preferenza per il programma sia diversa nei due sessi. Si controlli scegliendo $\alpha=0,05$.

10.17 Prima di decidere il lancio di una nuova marca di caffè una società di mercato realizza un sondaggio di opinioni a Milano e a Napoli su due campioni casuali di rispettive numerosità 150 e 100. Nel campione di Milano il 60% dei soggetti dichiara di preferire il nuovo tipo di caffè a quello abitualmente utilizzato, mentre nel campione di Napoli soltanto il 45% degli intervistati esprime la sua predilezione per il nuovo caffè.

Si può ipotizzare che la preferenza (e quindi il mercato) per il nuovo caffè sia diversa a Milano e a Napoli? Si scelga $\alpha=0,01$.

10.18 Un campione casuale di 400 telefonate urbane ha dato luogo alla seguente distribuzione di frequenza secondo la durata (in minuti) delle conversazioni:

Durata	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
Frequenza	28	64	142	118	3	18

Si stabilisca, al livello di significatività $\alpha=0,05$, se i dati confermano l'ipotesi che la durata delle conversazioni telefoniche segua una legge normale.