Test statistici di verifica di ipotesi

Test e verifica di ipotesi

Il test delle ipotesi consente di verificare se, e quanto, una determinata ipotesi (di carattere biologico, medico, economico,...) è supportata dall'evidenza empirica.

Il fenomeno studiato deve essere rappresentato mediante una distribuzione di probabilità e l'**ipotesi sulle caratteristiche del fenomeno** studiato è tradotta in **ipotesi su uno o più parametri** della distribuzione (**test parametrico**).

Esempio - Una moneta truccata?

Nel lancio di una moneta si vince se esce testa e si perde se esce croce. Il lanciatore garantisce che la moneta non è truccata.

Prima di giocare stiamo un po' a vedere e osserviamo che su 20 lanci esce testa solo 6 volte, un numero un po' basso rispetto al valore atteso di 10.

Ci chiediamo se il lanciatore ci sta ingannando o se il valore osservato sia un ragionevole frutto del caso. Al di là di risposte soggettive ed opinabili, è possibile attuare un test statistico per decidere se denunciare o meno il lanciatore.

Scopo del test è verificare se il dato osservato sia **probabilisticamente credibile**, assumendo (ipotesi) la moneta non sia truccata.

La probabilità dell'evento è (usando ad esempio la distribuzione $B(20, 1/2)$)

$$P(X = 6) = \frac{\binom{20}{6}}{2^{20}} \simeq 0.037,$$

quindi il risultato è decisamente poco probabile (meno del 4%).

Questo però non ci autorizza a concludere nulla

Infatti tutti i valori di $P(X = k)$ con $k = 0, ..., 20$ sono piuttosto piccoli. Ad esempio

$$P(X = 10) = \frac{\binom{20}{10}}{2^{20}} \simeq 0.18$$

ed in effetti, su 20 lanci ci aspettiamo che testa esca "circa" 10 volte, e non "esattamente" 10.

Calcolo dell'intervallo di confidenza

Se la moneta non è truccata allora $p(T) = 0.5$. Determiniamo un intervallo di confidenza per $p(T)$ a livello del 95%.
$N = 20$, $m_{20} = \frac{6}{20}$, $\alpha/2 = 0.025$ e $z_{\alpha/2} = 1.96$. Dunque

$$p(T) \in \left[m_N - z_{\alpha/2}\sqrt{\frac{m_N(1-m_N)}{N}}, m_N + z_{\alpha/2}\sqrt{\frac{m_N(1-m_N)}{N}}\right] = [0.1, 0.5]$$

Sulla base del risultato non possiamo escludere con forza l'ipotesi che la moneta non sia truccata.

Test e verifica di ipotesi

Pensandoci bene, per rispondere al nostro dubbio dobbiamo calcolare qual'è la **probabilità di osservare un risultato "sospetto" o "estremo" quanto e anche più di quello osservato**

In questo senso i possibili risultati estremi sono $X = 6, 5, 4, 3, 2, 1, 0$ e anche $X = 14, 15, 16, 17, 18, 19, 20$ (6 e 14 sono alla stessa distanza dal valore atteso 10).

La probabilità dell'unione di questi eventi è

$$\begin{aligned} p = \quad & P(X = 0) + \cdots + P(X = 6) + \\ & + P(X = 14) + \cdots + P(X = 20) \simeq 0.12 \end{aligned}$$

Quindi, assumendo la moneta non truccata, la probabilità di ottenere un risultato estremo è del 12%, una percentuale abbastanza alta per non avere forti dubbi che la moneta sia truccata.

Lo stesso risultato su 30 lanci avrebbe dato

$$p = \quad P(X = 0) + \cdots + \cdots + P(X = 6) + \\ + P(X = 24) + \cdots + P(X = 30) \simeq 0.0014$$

In questo caso il sospetto che la moneta sia truccata sarebbe seriamente fondato

definizione

Si chiama test statistico ogni procedura atta a verificare se un dato è in accordo con una teoria e si articola nelle seguenti fasi:

- **formulazione dell'ipotesi da verificare**, detta ipotesi nulla e indicata con H_0;
- **calcolo della probabilità p di ottenere un risultato estremo come e più di quello osservato**, nell'ipotesi che H_0 sia vera; p è detta valore p del test o p-value;
- **valutazione di p**; se p è troppo piccolo si rifiuta l'ipotesi H_0, se è grande la si accetta.

Osserviamo che, detto x il risultato osservato, il valore p è dato dalla formula

$$\boxed{p = P(|X - E| \geq |x - E|)}$$

Livelli di significatività del test

Nella pratica statistica i valori critici di p, detti livelli di significatività del test sono fissati dalla seguente convenzione.

convenzione

- Se $p \geq 0.05$, la discrepanza tra dato osservato e valore atteso non è statisticamente significativa (cioè può trattarsi di un effetto casuale del campionamento) e H_0 viene accettata.
- Se $p < 0.05$, H_0 viene, in genere, rifiutata e la discrepanza viene detta
  - statisticamente significativa se $0.01 \leq p < 0.05$;
  - molto significativa se $0.001 \leq p < 0.01$;
  - estremamente significativa se $p < 0.001$.

Attenzione: Il p-value non è la probabilità che H_0 sia vera (cosa che non ha senso), ma la probabilità del verificarsi di eventi estremi assumendo H_0 vera, cioè rappresenta un **livello di confidenza** del test.

Una moneta un po' truccata

Il test effettuato nell'esempio precedente non ci permette di rifiutare l'ipotesi che il lanciatore sia onesto. Vediamo però nel seguente esempio che non possiamo nemmeno escludere che la moneta sia **un po' truccata**, addirittura a nostro favore.

Esempio - Una moneta molto truccata

Assumiamo come H_0 che $p(T) = 0.3$.

Siccome il fenomeno è descritto dalla binomiale $B(N, p(T))$ con $N = 20$ allora si ha

$$E = N \cdot p(T) = 20 \cdot 0.3 = 6$$

Siccome tutti i valori di k da 0 a 20 sono estremi come o più di 6, si avrà $p = 1$.

Siccome $p > 0.05$ non possiamo escludere quindi nemmeno che la moneta sia molto truccata a favore del lanciatore.

Esempio - Una moneta un po' truccata

Supponiamo ora che H_0 sia $p(T) = 0.52$

Usando la distribuzione $B(20, 0.52)$ il valore atteso è

$$E = 20 \cdot 0.52 = 10.4$$

Siccome $10.4 - 6 = 4.4$, sono da considerare estremi i casi

$$k = 6, 5, 4, 3, 2, 1, 0 \text{ e } k = 20, 19, 18, 17, 16, 15$$

Una moneta un po' truccata

Si ha allora

$$
\begin{aligned}
p &= \sum_{k=0}^{6} \binom{20}{k} 0.52^k 0.48^{20-k} + \sum_{k=15}^{20} \binom{20}{k} 0.52^k 0.48^{20-k} \\
&= 0.48^{20} + 20 \cdot 0.52 \cdot 0.48^{19} + 190 \cdot 0.52^2 \cdot 0.48^{18} \\
&\quad + 1140 \cdot 0.52^3 \cdot 0.48^{17} + 4845 \cdot 0.52^4 \cdot 0.48^{16} \\
&\quad + 15504 \cdot 0.52^5 \cdot 0.48^{15} + 38760 \cdot 0.52^6 \cdot 0.48^{14} \\
&\quad + 0.52^{20} + 20 \cdot 0.48 \cdot 0.52^{19} + 190 \cdot 0.48^2 \cdot 0.52^{18} \\
&\quad + 1140 \cdot 0.48^3 \cdot 0.52^{17} + 4845 \cdot 0.48^4 \cdot 0.52^{16} \\
&\quad + 15504 \cdot 0.48^5 \cdot 0.52^{15} \simeq 0.07
\end{aligned}
$$

Disponendo di un calcolatore, il calcolo può essere eseguito nel modo seguente

$$
\begin{aligned}
p &= P(X \leq 6) + P(X \geq 15) = P(X \leq 6) + [1 - P(X \leq 14)] \\
&= DISTRIB.BINOM(6; 20; 0,52; 1) + \\
&\quad + [1 - DISTRIB.BINOM(14; 20; 0,52; 1)] \\
&\simeq 0.04 + [1 - 0.97] = 0.04 + 0.03 = 0.07
\end{aligned}
$$

Una moneta un po' truccata

Poichè $p > 0.05$ dobbiamo accettare anche che la moneta possa essere un po' truccata addirittura a nostro favore.

Con $p(T) = 0.6$ si ha $E = 20 \cdot 0.6 = 12$ e dunque sono estremi i valori $k = 0 - 6$ e $k = 18, 19, 20$. Si ha in tal caso

$$\begin{aligned} p &= P(X \leq 6) + P(X \geq 18) = P(X \leq 6) + [1 - P(X \leq 17)] \\ &= DISTRIB.BINOM(6; 20; 0,6; 1) + \\ &\quad + [1 - DISTRIB.BINOM(17; 20; 0,6; 1)] \\ &\simeq 0.01 \end{aligned}$$

che indica un notevole scostamento dal valore atteso. In tal caso la discrepanza è statisticamente significativa e l'ipotesi $p(T) = 0.6$ va rifiutata.

Z-test

Consente di effettuare il calcolo del p-value in maniera **approssimata** ma molto **più veloce**.

Idea: approssimare la distribuzione discreta con una Normale.

Siccome si usa il TLC, il risultato sarà tanto più accurato quanto più è alto il numero di prove N.

Z-test

N=numero di prove
k=numero di successi
q=valore ipotizzato del parametro (H_0)

$$\boxed{p = P\left(|Z| \geq \frac{|k - Nq|}{\sqrt{Nq(1-q)}}\right)}$$

dove $Z \sim \mathcal{N}(0, 1)$.

Il valore $s = \frac{|k - Nq|}{\sqrt{Nq(1-q)}}$ con cui va confrontata la normale è detto statistica del test

Infatti, siccome

$$X \sim \mathcal{N}\big(Nq, Nq(1-q)\big)$$

allora, standardizzando, si ha

$$Z = \frac{X - Nq}{\sqrt{Nq(1-q)}} \sim \mathcal{N}(0,1)$$

quindi

$$\begin{aligned} p &= P(|X - E| \geq |k - E|) && (1)\\ &= P(|X - Nq| \geq |k - Nq|) && (2)\\ &= P(|Z| \geq \frac{|k - Nq|}{\sqrt{Nq(1-q)}}) && (3) \end{aligned}$$

Esempio: Z-test su H_0 = moneta non truccata ($q = 0.5$)

$N = 20$
$k = 6$
$q = 0.5$ (H_0)

Si ha

$$s = \frac{|k - Nq|}{\sqrt{Nq(1-q)}} = \frac{|6 - 10|}{\sqrt{10(1-0.5)}} = \frac{4}{\sqrt{5}} \simeq 1.79$$

quindi

$$\begin{aligned} p &= P\big(|Z| \geq s)\big) = P\big(|Z| \geq 1.79\big) = 2P\big(Z \geq 1.79)\big) \\ &= 2 \cdot \big(1 - P(Z \leq 1.79)\big) = 2 \cdot (1 - 0.9633) = 0.0734 > 0.05 \end{aligned}$$

quindi l'ipotesi H_0: $q = 1/2$ non si può rigettare.

Esempio: Z-test su H_0 = moneta un po' truccata ($q = 0.52$)

$N = 20$
$k = 6$
$q = 0.52$ (H_0)

Si ha

$$s \simeq 1.96$$

e

$$p = 0.05$$

quindi l'ipotesi H_0: $q = 0.52$ non si può rigettare.

Esercizio 12.11 del testo

*Viene analizzato un campione di 1235 semi importati. Di essi 22 risultano transgenici. La ditta produttrice garantisce che la percentuale di semi transgenici tra i suoi prodotti è dell'*1%. *Si testi l'ipotesi nulla che la ditta affermi il vero.*

Esercizio 12.11 del testo

*Viene analizzato un campione di 1235 semi importati. Di essi 22 risultano transgenici. La ditta produttrice garantisce che la percentuale di semi transgenici tra i suoi prodotti è dell'*1%. *Si testi l'ipotesi nulla che la ditta affermi il vero.*

Indicata con $p(T)$ la percentuale di semi transgenici si ha

$$H_0) \qquad p(T) = 0.01$$

Possiamo procedere in 3 modi

1. calcolando il p-value con la formula binomiale
2. calcolando il p-value con lo Z-test
3. determinando un intervallo di confidenza al 95%

1. Calcolo del p-value con formula binomiale

Si ha

$$E = N \cdot p(T) = 1235 \cdot 0.01 = 12.35, \qquad x - E = 22 - 12.35 = 9.65$$

Sono quindi da considerare estremi tutti i valori di k che distano almeno 9.65 dal valore atteso 12.35, cioè

$$k \in \{0, 2, 22, 23, \ldots, 1235\}.$$

Dunque

$$\begin{aligned} p &= P(X \leq 2) + P(X \geq 23) = P(X \leq 2) + 1 - P(X \leq 22) \\ &= DISTRIB.BINOM(22; 1235; 0,01; 1) \\ &\quad +1 - DISTRIB.BINOM(22; 1235; 0,01; 1) \\ &= 0.0004 + 1 - 0.9959 = 0.0045 \end{aligned}$$

Siccome $p < 0.05$ l'ipotesi nulla è da rifiutare. In effetti $0.001 < p < 0.01$, quindi lo scostamento dal valore medio è statisticamente **molto significativo**.

1. Calcolo del p-value con Z-test

$N = 1235$
$k = 22$
$q = 0.01$
Si ha $s \simeq 2.957$ e $p = 0.0102$
Siccome $p < 0.05$ l'ipotesi nulla è da rifiutare.

3. Intervallo di confidenza al 95%

Si ha

$$m_N = x/N = 22/1235 \simeq 0.018, \quad z_{\alpha/2} = z_{0.025} = 1.96,$$

$$\begin{aligned} m_N \pm z_{\alpha/2}\sqrt{\frac{m_N(1-m_N)}{N}} &= 0.0178 \pm 1.96\sqrt{\frac{0.0178(1-0.0178)}{1235}} \\ &= 0.018 \pm 0.006 \end{aligned}$$

quindi

$$p(T) \in [0.012, 0.024]$$

Poiché l'estremo inferiore dell'intervallo è maggiore di 0.01 l'ipotesi nulla è da rifiutare.

Vantaggi e svantaggi dello Z-test

+ i conti sono più semplici
− è meno preciso
− si applica a modelli che coinvolgono una sola variabile

Test di adattamento del χ^2 (Pearson)

È utile quando il modello coinvolge più di una variabile aleatoria.

In questo test la **statistica** viene confrontata con la v.a. χ^2.

Test di adattamento del χ^2 (Pearson)

Esempio (esperimento di Mendel)

Incrociando tra loro piante di piselli di due linee pure, a **fiore rosso** e a **fiore bianco**, Mendel osservò

705 piante a fiore rosso
224 piante a fiore bianco

Prima legge di Mendel

La proporzione tra piante di fenotipo dominante (rosso) e fenotipo recessivo (bianco) è $3:1$ (i.e. $p(R) = 3/4$ e $p(B) = 1/4$).

Testiamo l'ipotesi

H_0) i dati sono in accordo con la legge di Mendel

Esempio (esperimento di Mendel)

Potremmo procedere in 3 modi:

1. test binomiale (esercizio)
2. Z - test (esercizio)
3. test del χ^2

	R	B	*totale*
ferquenze reali	$F_R = 705$	$F_B = 224$	929
valori attesi teorici	$E_R = 929\frac{3}{4} = 696.75$	$E_B = 929\frac{1}{4} = 232.5$	929
scarti	$F_R - E_R = 8.25$	$F_B - E_B = -8.25$	0

La statistica del test è

$$s = \frac{|F_R - E_R|^2}{E_R} + \frac{|F_B - E_B|^2}{E_B} = \frac{8.25^2}{696.75} + \frac{(-8.25)^2}{232.5} \simeq 0.39$$

Test di adattamento del χ^2 (Pearson)

Esempio (esperimento di Mendel)

$$\text{statistica del test } = s \simeq 0.39$$

È da ritenere che se s è piccolo allora vi sia un buon accordo con l'ipotesi nulla. Se è grande l'ipotesi va rifiutata.

Questione: come determinare se s è grande o piccolo?

Ciò è determinato dal p-value

$$p = P(\chi^2_1 \geq s) = P(\chi^2_1 \geq 0.39) > P(\chi^2_1 \geq 1.64) = 0.2$$

dove il numero di gradi di libertà è scelto pari a 1 perché in effetti c'è realmente una sola variabile indipendente

Il valore p del test è dunque molto alto e l'ipotesi confermata

Test di adattamento del χ^2 (Pearson)

In generale

Il test del χ^2 confronta l'accordo (o adattamento) tra frequenza osservata e frequenza attesa di dati organizzati in n categorie qualitative.

Supponiamo di estrarre da una popolazione un campione di dimensione N e di osservare nel campione le frequenze F_1, F_2, ..., F_n.

Se le frequenze relative delle diverse categorie sono q_1, q_2, ..., q_n (ipotesi nulla), i valori attesi di tali frequenze sono $E_i = Nq_i$.

La statistica del test è il numero

$$s = \sum_{i=1}^{n} \frac{|F_i - E_i|^2}{E_i}$$

e l'ipotesi nulla va valutata in relazione al p-value

$$p = P(\chi^2_{n-1} \geq s)$$

Test di adattamento del χ^2 (Pearson)

Esempio

Per testare l'efficacia di un principio attivo si preparano 3 farmaci V_1, V_2 e V_3 dove

V_1 non contiene il principio attivo
V_2 contiene il principio in quantità q
V_3 contiene il principio in quantità $2q$

Si osservano i seguenti risultati nella sperimentazione

	V_1	V_2	V_3	*totale*
pazienti migliorati	12	5	29	46
pazienti non migliorati	114	80	90	284
totale	126	85	119	330

Test di efficacia di un farmaco

Frequenze reali

	V_1	V_2	V_3	*totale*
pazienti migliorati	12	5	29	46
pazienti non migliorati	114	80	90	284
totale	126	85	119	330

H_0 = il farmaco è inefficace

= gli eventi "miglioramento" e "assunzione del farmaco" sono indipendenti

$H_0 \implies$ valore atteso teorico di pazienti migliorati che hanno assunto V_3

= probabilità che un paziente migliorato abbia assunto V_3 x 330

$$= \frac{46}{330} \cdot \frac{119}{330} \cdot 330 = \frac{46}{330} 119 \simeq 16.6$$

Test di efficacia di un farmaco

Frequenze reali

	V_1	V_2	V_3	*totale*
pazienti migliorati	12	5	29	46
pazienti non migliorati	114	80	90	284
totale	126	85	119	330

Valori attesi teorici

	V_1	V_2	V_3
pazienti migliorati	17.6	11.8	16.6
pazienti non migliorati	108.4	73.2	102.4

Scarti

	V_1	V_2	V_3
pazienti migliorati	-5.6	-6.8	12.4
pazienti non migliorati	5.6	6.8	-12.4

Test di efficacia di un farmaco

Valori attesi teorici

	V_1	V_2	V_3
pazienti migliorati	17.6	11.8	16.6
pazienti non migliorati	108.4	73.2	102.4

Scarti

	V_1	V_2	V_3
pazienti migliorati	-5.6	-6.8	12.4
pazienti non migliorati	5.6	6.8	-12.4

$$s = \frac{(-5.6)^2}{17.6} + \frac{(-6.8)^2}{11.8} + \frac{(12.4)^2}{16.6} + \frac{(5.6)^2}{108.4} + \frac{(6.8)^2}{73.2} + \frac{(-12.4)^2}{102.4} \simeq 17.4$$

$$0.001 = P(\chi^2_5 \geq 20.5) < p = P(\chi^2_5 \geq 17.4) < P(\chi^2_5 \geq 16.7) = 0.005$$

L'ipotesi H_0 va rigettata. I risultati osservati sono statisticamente molto significativi e inducono a ritenere che il farmaco abbia effetto.