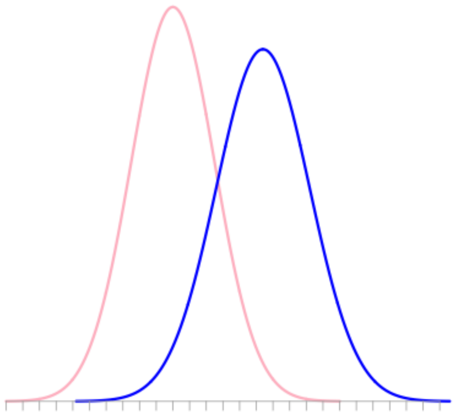
studio roccato
Data Science Training & Consulting

Introduzione ai Test di Ipotesi

Concetti base e applicazioni pratiche



Contenuti

- Introduzione ai Test di Ipotesi
- Formulazione delle Ipotesi
- Errori di Tipo I e di Tipo II
- Test unilaterali e bilaterali
- Assunzioni fondamentali
- Tipologie di Test Statistici
- Statistica di test e valore critico
- Intervallo di Confidenza
- Il p-value
- Utilizzo del software statistico
- Dimensione del Campione e Potenza del Test
- Conduzione del Test
- Esempi
- Appendice

Introduzione ai Test di Ipotesi

Spesso è necessario **prendere decisioni** in situazioni di incertezza basate su un'informazioni parziali, ottenute da **campioni**.

Per questo si **formulano ipotesi** che vengono verificate attraverso **test statistici**.

Esempi di ipotesi verificabili con i test statistici:

- Un nuovo farmaco porta miglioramenti significativi per una determinata patologia rispetto al placebo

- Il peso medio di un prodotto è inferiore rispetto a quello dichiarato dal fornitore

- La percentuale di bagagli persi da una compagnia aerea non supera il 6%

- Una nuova tecnica di comunicazione è più efficace della precedente per catturare l'attenzione dei clienti

- la percentuale di adulti residenti in una certa zona che possiede un cellulare è diversa da quella nazionale

Formulazione delle ipotesi

Definizione dell'enunciato

La **definizione dell'enunciato** parte da **ipotesi** complementari relative a una caratteristica della popolazione in studio.

- **L'ipotesi nulla (H_0)** *è l'ipotesi predefinita che viene assunta come vera finché non ci sono prove sufficienti per rifiutarla (per es. non c'è differenza tra un nuovo farmaco e il placebo)*
- **L'ipotesi alternativa (H_a)** *è quella che viene accettata se l'ipotesi nulla viene rifiutata (per es. il nuovo farmaco è più efficace del placebo)*

Di norma[1] si parte **assumendo vera l'ipotesi nulla**, fino a quando le prove raccolte non siano sufficienti per rifiutarla.

Il **mancato rifiuto dell'ipotesi nulla non implica che sia vera**: significa soltanto che il campione non fornisce prove sufficienti per rifiutarla.

In campo penale, l'**ipotesi nulla** che afferma che "***l'imputato è innocente***" contrapposta all'**ipotesi alternativa** "***l'imputato è colpevole***". L'ipotesi nulla viene rifiutata solo se le prove a favore dell'ipotesi alternativa superano ogni "*ragionevole dubbio*". Tuttavia, **il mancato rifiuto di H_0 non implica l'innocenza dell'imputato** (la corretta dizione è "*non colpevole*"): significa semplicemente che le prove disponibili non sono sufficienti per una condanna.

[1] "*Quando ci si confronta con una teoria non è bene cercare prove scientifiche che la sostengano, ma è più saggio cercare prove che la contraddicano, quindi che la falsifichino. È possibile dire, pertanto, che una proposizione è scientifica solo quando può essere confutata dall'esperienza*." Carl R. Popper, 1970, "Logica della scoperta scientifica".

Formulazione delle ipotesi

Il campione di sperimentale e il campione di controllo

L'**ipotesi nulla** deve essere necessariamente formulata in **termini quantitativi**, riferendosi a un **parametro** di interesse, come la **proporzione** o il **valor medio**.

I test di verifica di ipotesi si applicano spesso[1] **a due campioni**, confrontando il parametro di interesse per **disuguaglianza** (differenza).

- Il **campione di controllo** usato per calcolare il parametro secondo H_0
- Il **campione sperimentale** usato per calcolare il parametro secondo H_a

Questo metodo verifica se la **differenza** tra i due parametri è **significativa**.

Per valutare l'efficacia di un nuovo farmaco, ad esempio, si confronta la proporzione di pazienti migliorati nei due gruppi di trattamento. La percentuale di pazienti migliorati con il placebo è indicata con $\mathbf{p_0}$, mentre quella relativa al nuovo farmaco è indicata con $\mathbf{p_a}$. L'ipotesi nulla ($\mathbf{H_0}$) afferma che la percentuale dei pazienti migliorati con il nuovo farmaco è uguale o inferiore a quella del placebo ($\mathbf{p_a} \leq \mathbf{p_0}$), mentre l'ipotesi alternativa ($\mathbf{H_a}$) sostiene che il nuovo farmaco è più efficace ($\mathbf{p_a} > \mathbf{p_0}$).

L'obiettivo dello studio è verificare se la differenza tra $\mathbf{p_a}$ e $\mathbf{p_0}$ è statisticamente significativa, indicando un miglioramento reale con il nuovo farmaco.

[1] Possono essere applicati a **un solo campione** ma sono, in realtà, di scarsa utilità perché spesso **non si conosce il valore della media μ_0 o della proporzione p_0 della popolazione di riferimento**.

Errori di Tipo I e Tipo II

Definizione e significato degli errori

La decisione di accettare o rifiutare un'ipotesi dipende da quanto si è disposti a tollerare l'errore. Questo livello di tolleranza, chiamato "**livello di significatività**", determina un valore limite, chiamato "**valore critico**", al di sopra o al di sotto del quale l'ipotesi viene rifiutata.

Errore di Tipo I (α), detto anche livello di significatività

Si verifica quando **si rifiuta erroneamente l'ipotesi nulla quando questa è vera**. Il valore di α si pone di solito a **0,05** *(per es. si accetta un errore del 5**%** che un allarme antincendio suoni inutilmente).*

Il livello di confidenza ($1-\alpha$)

È la probabilità di prendere la decisione corretta di **non rifiutare l'ipotesi nulla quando questa è vera** *(per es. un livello di confidenza del 95% significa che su 100 incendi, ci si aspetta che l'allarme suoni correttamente in circa 95 casi).* Un livello di confidenza del 95% corrisponde a una tolleranza del 5% per errori di Tipo I.

Errore di Tipo II (β)

Si verifica quando **non si rifiuta l'ipotesi nulla, anche se è falsa**.
Il valore β si pone di solito a **0,2**[1] *(per es. **si accetta un errore del 20%** che l'allarme antincendio non suoni quando c'è un incendio).*

La potenza del test (1-β)

È la probabilità di **rifiutare correttamente l'ipotesi nulla quando è falsa** *(per es. se **1-**β è **0,8**, si vuole avere l' **80% di probabilità che un allarme antincendio suoni** quando c'è effettivamente un incendio).*

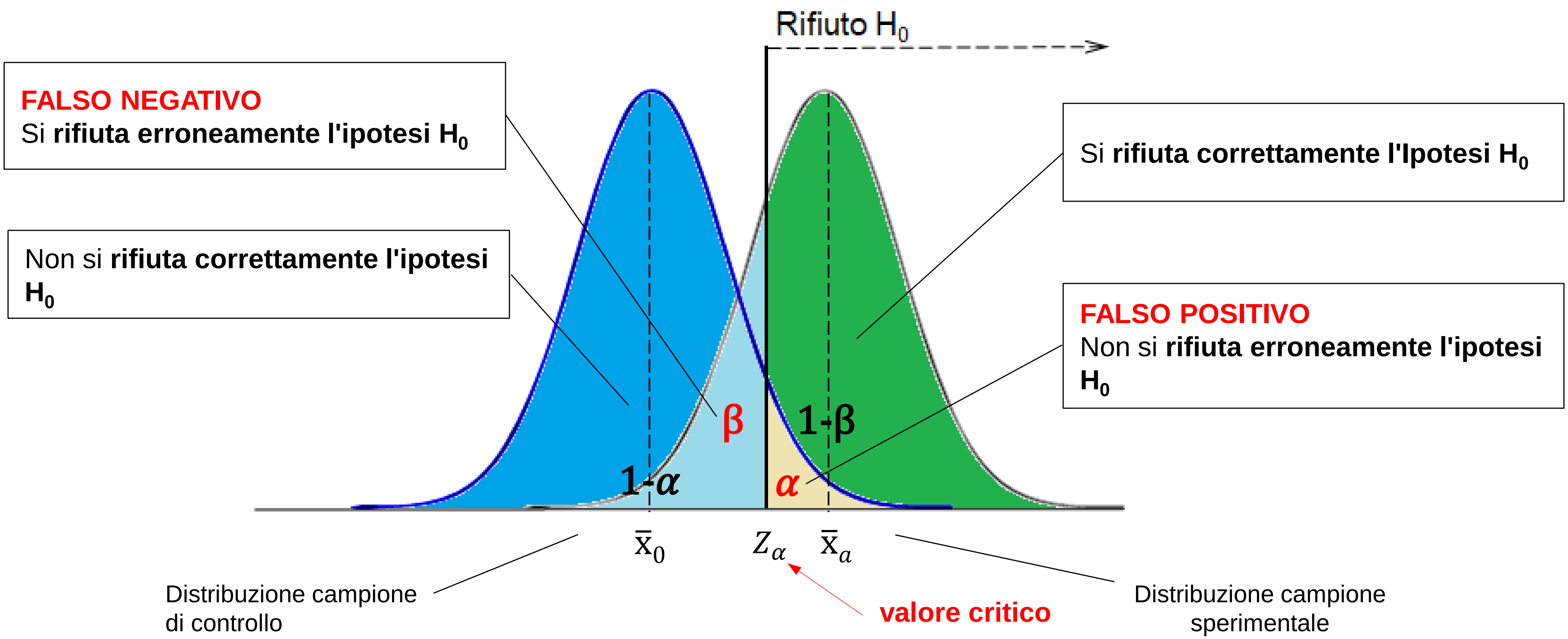
		Decisione	
		H_0 Non rifiutata	H_0 Rifiutata
Ipotesi	H_0 Vera	Decisione corretta $1-\alpha$	Errore Tipo I α
	H_0 Falsa	Errore Tipo II β	Decisione corretta $1-\beta$

[1] In molti contesti, un errore di Tipo II del 20% potrebbe avere conseguenze significative (per es., in ambito medico, non riuscire a rilevare una malattia potrebbe ritardare la diagnosi e compromettere le possibilità di cura).

Errori di Tipo I e Tipo II

Definizione e significato degli errori

Tra α e β sussiste una relazione inversa: **minore è il valore di α, maggiore è il valore di β**. Le **probabilità di commettere errori** corrispondono alle **aree sotto le code delle distribuzioni** dei due campioni.



Test unilaterali e bilaterali

Direzioni del test

Se si verificare se una **differenza significativa** esiste in una direzione specifica (**solo maggiore o solo minore rispetto al valore atteso**), si utilizza un **test a una coda**. Questo include **solo i valori estremi** nella coda sinistra o nella coda destra della distribuzione.

	Medie		Proporzioni		Test
Ipotesi Nulla H_0	$\bar{x}_a \leq/\geq \bar{x}_0$	la media $\bar{x}_a$ è ≤ o ≥ della media $\bar{x}_0$	$p_a \leq/\geq p_0$	la proporzione p_a è ≤ o ≥ della proporzione p_0	$\bar{x}_a - \bar{x}_0 \leq/\geq 0$ $p_a - p_0 \leq/\geq 0$
Ipotesi Alternativa H_a	$\bar{x}_a >/< \bar{x}_0$	la media $\bar{x}_a$ è > o < della media $\bar{x}_0$	$p_a >/< p_0$	la proporzione p_a è > o < della proporzione p_0	$\bar{x}_a - \bar{x}_0 >/< 0$ $p_a - p_0 >/< 0$

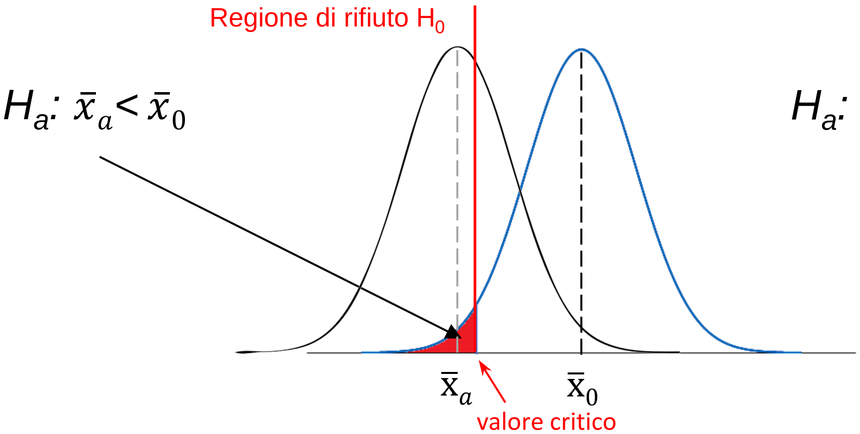
Se si vuole verificare se una **differenza significativa** esiste in **entrambe le direzioni** (positiva o negativa), si utilizza un test a due code. Questo include i **valori estremi sia nella coda sinistra sia nella coda destra** della distribuzione.

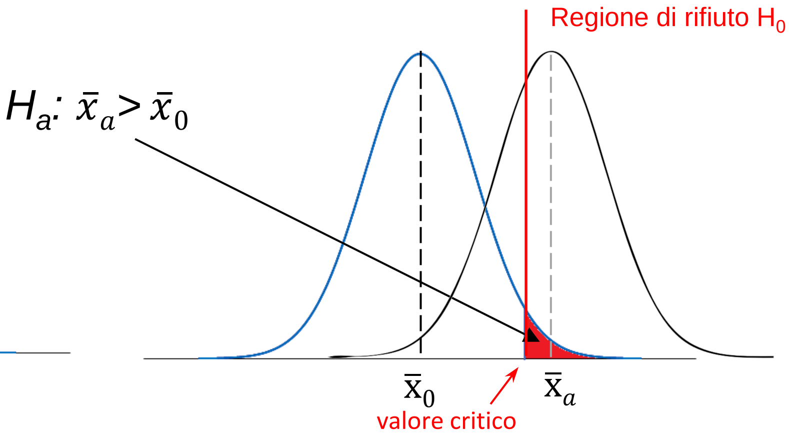
	Medie		Proporzioni		Test
Ipotesi Nulla H_0	$\bar{x}_a = \bar{x}_0$	la media $\bar{x}_a$ è **uguale** alla media $\bar{x}_0$	$p_a = p_0$	la proporzione p_a è **uguale** alla proporzione p_0	$\bar{x}_a - \bar{x}_0 = 0$ $p_a - p_0 = 0$
Ipotesi Alternativa H_a	$\bar{x}_a \neq \bar{x}_0$	la media $\bar{x}_a$ è **diversa** dalla media $\bar{x}_0$	$p_a \neq p_0$	la proporzione p_a è **diversa** dalla proporzione p_0	$\bar{x}_a - \bar{x}_0 \neq 0$

Test unilaterali e bilaterali

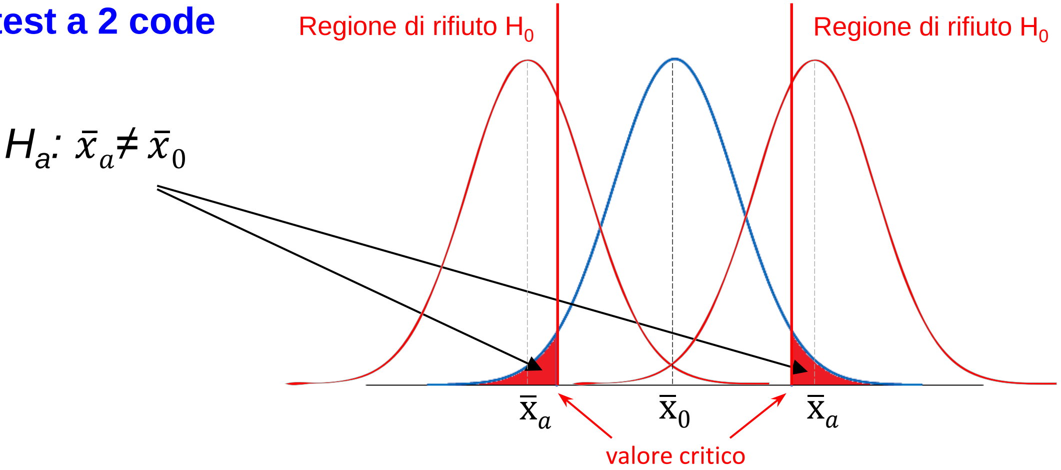
Regione di rifiuto

La regione di rifiuto nel caso di **test a 1 coda** (sinistra o destra)





La regione di rifiuto nel caso di **test a 2 code**



Assunzioni fondamentali

Normalità e indipendenza dei dati

Verificare le assunzioni è fondamentale per garantire la **validità dei risultati** dei test statistici e evitare di trarre conclusioni errate.

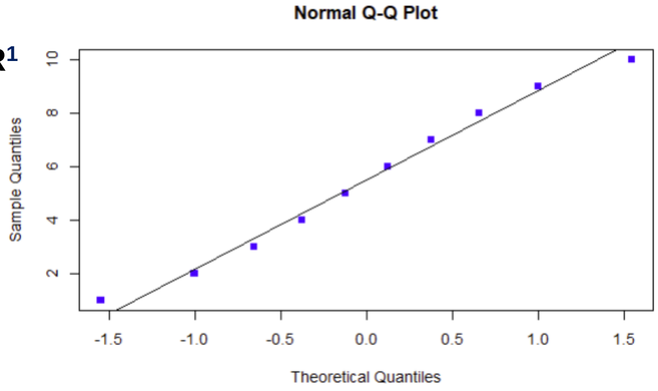
- **Normalità: Medie**

  Normalità: La **distribuzione** dei dati deve essere **approssimativamente normale**.

  Questa si può verificare con un esame visivo dei dati (per es. un Box plot o un Q-Q plot) o con dei test formali come quello di Shapiro-Wilk.

Esempio di verifica di normalità dei dati utilizzando il **software R**[1]

```
dati <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
qqnorm(dati, pch=15, col="blue"); qqline(dati)
shapiro.test(dati)

Shapiro-Wilk normality test
W = 0.97016, p-value = 0.8924
```



Quanto più il valore `W` è vicino a 1, tanto più i dati tendono a conformarsi a una distribuzione normale. Il risultato va però interpretato insieme al `p-value`: se $p \geq 0{,}05$, non ci sono evidenze sufficienti per rifiutare l'ipotesi di normalità.

[1] Vedi Appendice

Assunzioni fondamentali

Normalità e indipendenza dei dati

- **Normalità: Proporzioni**

  Una proporzione segue una distribuzione binomiale. Tuttavia, quando la dimensione del campione è sufficientemente grande, può essere approssimata da una distribuzione normale.
  Regola pratica: $np \geq 10$ e $n(1-p) \geq 10$, dove n rappresenta la dimensione del campione e p è la proporzione attesa (sotto l'ipotesi nulla) o osservata.

- **Omogeneità della varianza**

  L'assunzione di omogeneità delle varianze è fondamentale per alcuni test statistici, come il classico test t di Student.

  Tuttavia, questa assunzione può essere **difficile da verificare** e spesso **non è soddisfatta nei dati reali**. Per ovviare a questa limitazione, si utilizzano **test t che correggono i gradi di libertà**, rendendoli più robusti anche in caso di varianze diverse.

  In questo corso si privilegiano i test robusti come metodo standard. Gli esempi pratici utilizzano funzioni del software R, che tengono conto di eventuali disuguaglianze tra varianze, garantendo risultati affidabili e flessibili per diverse tipologie di dati.

- **Indipendenza dei dati**

  Le osservazioni all'interno di ciascun campione **devono essere indipendenti tra loro**.

Tipologie di Test Statistici

Test non parametrici

Sebbene questo corso sia incentrato sui test parametrici, in alcune situazioni i **test non parametrici**, che **non richiedono assunzioni sulla distribuzione dei dati o sull'omogeneità della varianza**, potrebbero risultare più **appropriati**.

In questa tabella vengono riportati i più comuni.

Test	Quando si usa	Descrizione
Test di Wilcoxon per un campione	Dati ordinali o non normali per confrontare con un valore noto.	Verifica se la mediana di un campione differisce da un valore specifico.
Test di Mann-Whitney	Due gruppi indipendenti con dati ordinali o non normali.	Confronta se i due gruppi provengono dalla stessa distribuzione.
Test di Kolmogorov-Smirnov	Due campioni indipendenti, confronto delle distribuzioni.	Verifica differenze nella distribuzione cumulativa tra due gruppi.
Test di Wilcoxon Signed-Rank	Due osservazioni appaiate con dati ordinali o non normali.	Analizza i cambiamenti tra due misurazioni sugli stessi soggetti (pre-post).
Test di McNemar	Dati categorici appaiati (pre-post, casi/controlli).	Valuta i cambiamenti tra due osservazioni dello stesso gruppo per dati dicotomici.
Test di Kruskal-Wallis	Più di due gruppi indipendenti con dati ordinali o non normali.	Estensione del test di Mann-Whitney per più campioni.
Test esatto di Fisher	Variabili categoriche, piccoli campioni (es. < 5 osservazioni per cella).	Alternativa al chi-quadro per piccoli numeri, valuta la dipendenza tra categorie.
Correlazione di Spearman	Relazione tra due variabili ordinali o continue non lineari.	Misura la relazione monotona (non lineare) tra due variabili.

Poiché questi test **non vengono trattati in questo corso**, si consiglia di consultare risorse aggiuntive per un approfondimento

Tipologie di Test Statistici

Test parametrici su due popolazioni

Per verificare se il valore medio di una distribuzione si discosta significativamente da un certo valore di riferimento, si usano la **statistiche di test Z** e **t** definite come:

$$Z, t = \frac{\bar{x} - \mu}{SE} \text{ e, per le proporzioni, } Z = \frac{p - \pi}{SE}$$

$SE = s / \sqrt{n}$ è chiamato **errore standard**, dove s è la deviazione standard del campione e n la sua numerosità. Nel caso delle proporzioni $SE = \sqrt{\pi(1-\pi)} / \sqrt{n}$, dove $\sqrt{\pi(1-\pi)}$ è la deviazione standard delle proporzioni.

- Si utilizza il test Z quando la **deviazione standard** della **popolazione è nota,** indipendentemente dalla dimensione del campione.
- Si utilizza il test t quando la **deviazione standard** della **popolazione non** è **nota**, indipendentemente dalla dimensione del campione; tuttavia, per n>30, i risultati del test t si avvicinano a quelli del test Z.

Test di ipotesi	Scopo	Statistica del test standardizzata[1]
Per una media	**Testare una media osservata con una teorica**	$Z,t = (\bar{x} - \mu) / (s/\sqrt{n})$
Per differenza tra due medie	**Testare la differenza tra le medie di 2 campioni**	$Z,t = (\bar{x}_a - \bar{x}_0) / \sqrt{s_a^2/n_a + s_0^2/n_0}$
Per una proporzione	**Testare una proporzione osservata con una teorica**	$Z = (p - \pi) / \sqrt{\pi(1-\pi)/n}$
Per differenza tra due proporzioni	**Testare la differenza tra le proporzioni di 2 campioni**	$Z = (p_a - p_0) / \sqrt{p_a(1-pa)/n_a + p_0(1-p_0)/n_0}$

[1] μ_0 e π_0 rappresentano rispettivamente la media e la proporzione della popolazione di riferimento (o nel campione di controllo, in caso di confronto tra due campioni).

Tipologie di Test Statistici

Test parametrici per dati appaiati

Un **errore frequente** nell'analisi statistica è quello di **applicare test per campioni indipendenti** **a dati** che in realtà sono **appaiati.**

Si consideri uno studio sull'efficacia di un nuovo farmaco antidolorifico, misurando il livello di dolore di un gruppo di pazienti prima e dopo il trattamento. In questo caso, **le osservazioni "prima" e "dopo" sono legate allo stesso individuo** e **non possono essere considerate indipendent**i. Applicare un test per campioni indipendenti a questi dati porterebbe a una sovrastima della significatività dell'effetto del farmaco, portando potenzialmente a decisioni cliniche errate.

I test in questo caso per le **medie** usano la statistica test $\boldsymbol{t}$ (o $\boldsymbol{Z}$) definita come[1] $\boldsymbol{Z, t = \bar{d}/SE}$ dove $\boldsymbol{\bar{d}}$ **è la media delle differenze** e $\boldsymbol{SE = s_d/\sqrt{n}}$ è l'errore standard, con $\boldsymbol{s_d}$ che è la deviazione standard campionaria delle differenze e $\boldsymbol{n}$ è la dimensione del campione (vedi esempio).

Nel caso delle **proporzioni**, si utilizza la **statistica del chi-quadrato** in quanto più preciso per confrontare proporzioni in una tabella di contingenza (vedi esempio).

Test di ipotesi	Scopo	Statistica del test standardizzata
Per le medie	Testare la media delle differenze tra due campioni ($\boldsymbol{\bar{d}}$)	$Z, t = \frac{\bar{d}}{s_d}\sqrt{n}$
Per le proporzioni	Testare se la proporzione di "successi" in un campione è diversa da un valore teorico o da un'altra proporzione osservata.	Statistica del chi-quadrato $\mathbf{X^2}$

[1] Contrariamente alla formula precedente, non si è sottratto a $\boldsymbol{\bar{d}}$ il valore atteso della media delle differenze sotto l'ipotesi nulla $\boldsymbol{\mu}$ in quanto **è solitamente uguale a 0**.

Statistica di test e valore critico

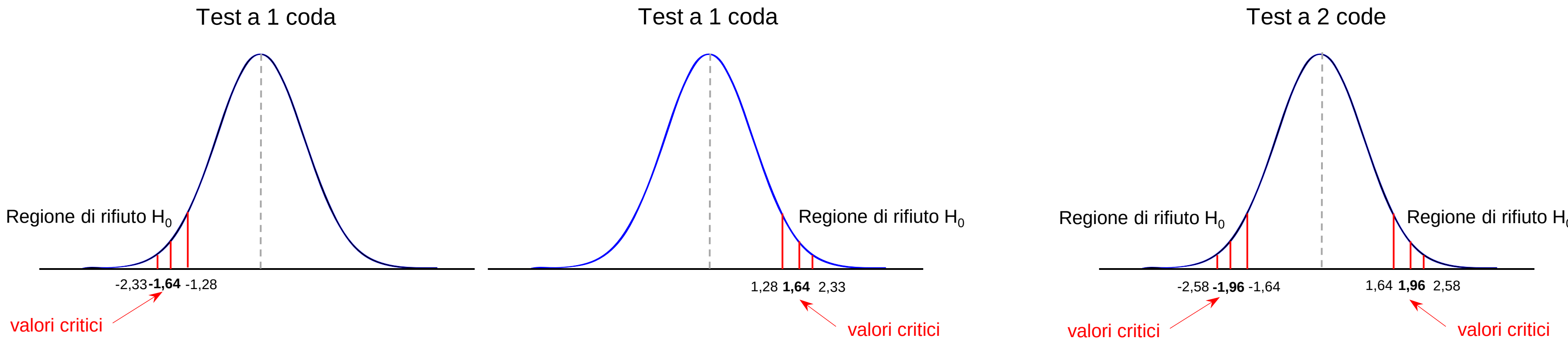
Statistica di test Z

La statistica di test **Z**, per $n \geq 30$, segue una **distribuzione normale standard**[1] con media $\mu = 0$ e deviazione standard $\sigma = 1$.

A ogni **livello di significatività** (α) corrisponde un **valore critico**, che delimita l'area di rifiuto dell'ipotesi nulla.

Valori critici più comuni per un test a due code:

Livello di significatività α	**Valore critico Z_α (test a 1 coda)**	**Valore critico Z_α (test a 2 code)**[2]
1%	-2,326 oppure 2,326	-2,576 e 2,576
5%	**-1,645 oppure 1,645**	**-1,960 e 1,960**
10%	-1,282 oppure 1,282	-1,645 e 1,645



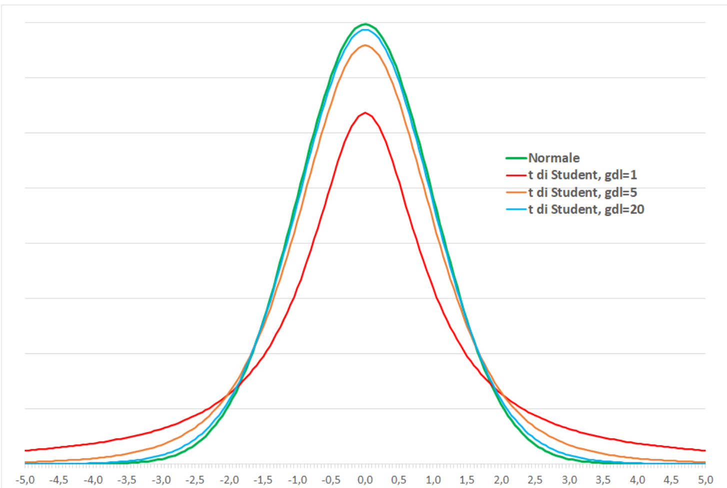
[1] Nel caso di proporzioni, la condizione è valida se $np_0 \geq 10$ e $n(1-p_0) \geq 10$. In caso contrario, è opportuno utilizzare la **distribuzione binomiale**.
[2] Nel **test a 2 code** il valore di α **viene diviso per 2.**

Statistica di test e valore critico

Statistica di test t

Quando la **deviazione standard della popolazione σ è sconosciuta**, si utilizza la **distribuzione t di Student** per il calcolo del valore critico, indipendentemente dalla dimensione del campione. Tuttavia, per campioni molto grandi ($n \geq 30$), i **risultati** del **test t e del test Z tendono a essere simili**. La distribuzione $\boldsymbol{t}$ dipende non solo dal livello di significatività α, ma anche dai gradi di libertà ($n-1$) del campione.

All'aumentare di n, la distribuzione t si avvicina sempre di più a una distribuzione normale.



Confronto tra la distribuzione Z e t

	α				
GdL	0,005	0,01	0,025	0,05	0,10
1	63,6567	31,8205	12,7062	6,3138	3,0777
2	9,9248	6,9646	4,3027	2,9200	1,8856
3	5,8409	4,5407	3,1824	2,3534	1,6377
4	4,6041	3,7469	2,7764	2,1318	1,5332
5	4,0321	3,3649	2,5706	2,0150	1,4759
6	3,7074	3,1427	2,4469	1,9432	1,4398
7	3,4995	2,9980	2,3646	1,8946	1,4149
8	3,3554	2,8965	2,3060	1,8595	1,3968
9	3,2498	2,8214	2,2622	1,8331	1,3830
10	3,1693	2,7638	2,2281	1,8125	1,3722
11	3,1058	2,7181	2,2010	1,7959	1,3634
12	3,0545	2,6810	2,1788	1,7823	1,3562
13	3,0123	2,6503	2,1604	1,7709	1,3502
14	2,9768	2,6245	2,1448	1,7613	1,3450
15	2,9467	2,6025	2,1314	1,7531	1,3406
16	2,9208	2,5835	2,1199	1,7459	1,3368
17	2,8982	2,5669	2,1098	1,7396	1,3334
18	2,8784	2,5524	2,1009	1,7341	1,3304
19	2,8609	2,5395	2,0930	1,7291	1,3277
20	2,8453	2,5280	2,0860	1,7247	1,3253
21	2,8314	2,5176	2,0796	1,7207	1,3232
22	2,8188	2,5083	2,0739	1,7171	1,3212
23	2,8073	2,4999	2,0687	1,7139	1,3195
24	2,7969	2,4922	2,0639	1,7109	1,3178
25	2,7874	2,4851	2,0595	1,7081	1,3163
26	2,7787	2,4786	2,0555	1,7056	1,3150
27	2,7707	2,4727	2,0518	1,7033	1,3137
28	2,7633	2,4671	2,0484	0,0000	1,3125
29	2,7564	2,4620	2,0452	1,6991	1,3114
30	**2,7500**	**2,4573**	**2,0423**	**1,6973**	**1,3104**

Livello di significatività α	Valore critico $t_{\alpha,29}$ (test a 1 coda)	Valore critico $t_{\alpha,29}$ (test a 2 code)
1%	-2,462 oppure 2,462	-2,756 e 2,756
5%	**-1,699 oppure 1,699**	-2,045 e 2,045
10%	-1,311 oppure 1,311	-1,699 e 1,699

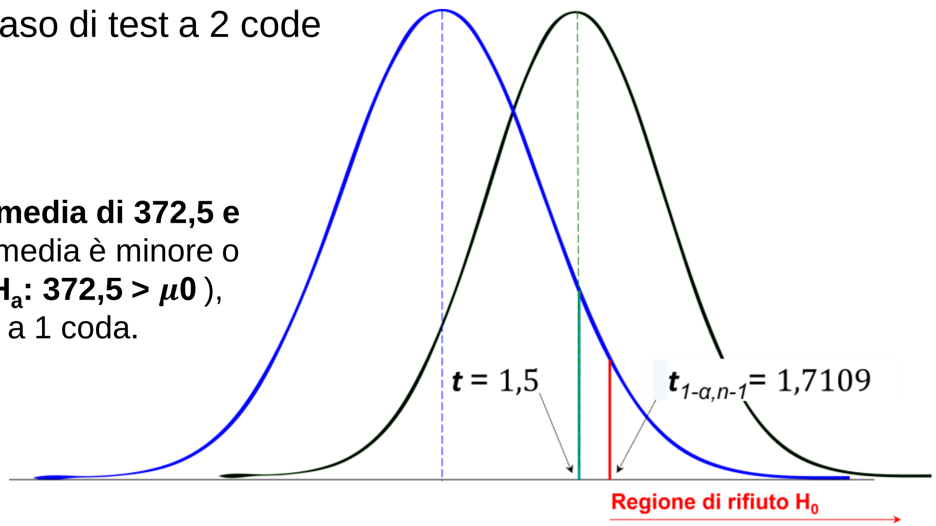
[1] I gradi di libertà corrispondono al numero di informazioni indipendenti che sono libere di variare nel calcolo di una determinata stima di un parametro. Si ottengono sottraendo uno dal numero totale di osservazioni in un campione statistico.

Statistica di test e valore critico

Rifiuto dell'ipotesi nulla

A questo punto si può **rifiutare l'ipotesi nulla** se la **statistica di test *t*** (o ***Z***, nel caso delle proporzioni) **cade nella regione di rifiuto H_0**:

- $t < t_{\alpha,n-1}$ nel caso di test a 1 coda (sinistra)
- $t > t_{1-\alpha,n-1}$ nel caso di test a 1 coda (destra)
- $t < t_{\frac{\alpha}{2},n-1}$ oppure $t > t_{1-\frac{\alpha}{2},n-1}$ nel caso di test a 2 code



Esempio. Avendo un campione di **25 misure,** con una **media di 372,5 e** una **deviazione standard di 15**, si vuole testare se la media è minore o uguale alla **media ipotizzata di 368** (**H_0: 372,5 ≤ μ0 ; H_a: 372,5 > μ0**), con un livello di significatività **α = 5%**. Il test sarà quindi a 1 coda.

Il valore della **statistica di test *t*** = $(\bar{x} - \mu_0) / (s/\sqrt{n})$ è:
t=(372,5 – 368) / (15/√25) = **1,5**.
Il **valore critico** per α = 5% è t_0 $_{(1-\alpha=\mathbf{0,95},\ \mathrm{gdl}=\mathbf{24})}$ = **1,7109**

Essendo il **valore *t* inferiore al valore critico t_0** (fuori dalla regione di rifiuto), **non si rifiuta l'ipotesi nulla H_0** al livello di significatività α = 5%.

Intervallo di Confidenza

Definizione, calcolo e interpretazione

Quando si **stima una caratteristica** di una popolazione (media o proporzione), si calcola un **Intervallo di Confidenza (IC)**, ovvero un intervallo di valori plausibili che, con un livello di confidenza scelto, indica dove potrebbe trovarsi il vero valore del parametro.

IC = ***statistica di test*** ± (***valore critico t***) * ***errore standard (SE)***

Per le medie: $IC = \bar{x} \pm t_{1-\alpha/2\ ,n-1}\ {}^{s}/_{\sqrt{n}}$

Se il valore ipotizzato sotto H_0 (es. μ_0 o π_0) **rientra nell'IC**, **non si rifiuta l'ipotesi nulla** perché non vi è evidenza sufficiente per supportare l'alternativa.

Un livello di confidenza del 95% non rappresenta la probabilità che il parametro cada nell'IC. Significa che, ripetendo l'esperimento molte volte, circa il 95% degli intervalli calcolati includerà il vero valore del parametro.

Utilizzando l'esempio precedente, si vuole verificare se la media è uguale a μ_0 **= 368** con un livello di significatività α = 5%. L'**IC** al 95% (372,5 - (2,064)*15/√25, 372,5 + (2,064)*15/√25) è compreso tra [**366,3 , 378,7**].
Poiché il **valore ipotizzato μ_0 = 368 è compreso nell'IC, non si rifiuta l'ipotesi nulla**.

L'ampiezza dell'IC riflette l'incertezza della stima: maggiore è l'ampiezza, maggiore è l'incertezza, dovuta alla variabilità nei dati o a un livello di confidenza elevato. Per es., un IC al 99% [155, 195] cm è troppo ampio per trarre conclusioni utili.

Intervallo di Confidenza

Intervallo di confidenza per un campione

test a 1 coda (sinistra), media: $-\infty, \bar{x} + t_{1-\alpha,n-1}\frac{s}{\sqrt{n}}$, proporzione: $0,\ p_a + Z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}$

test a 1 coda (destra), media: $\bar{x} - t_{1-\alpha,n-1}\frac{s}{\sqrt{n}}, +\infty$, proporzione: $p_a - Z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}, 1$

test a 2 code, media: $(\bar{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}})$

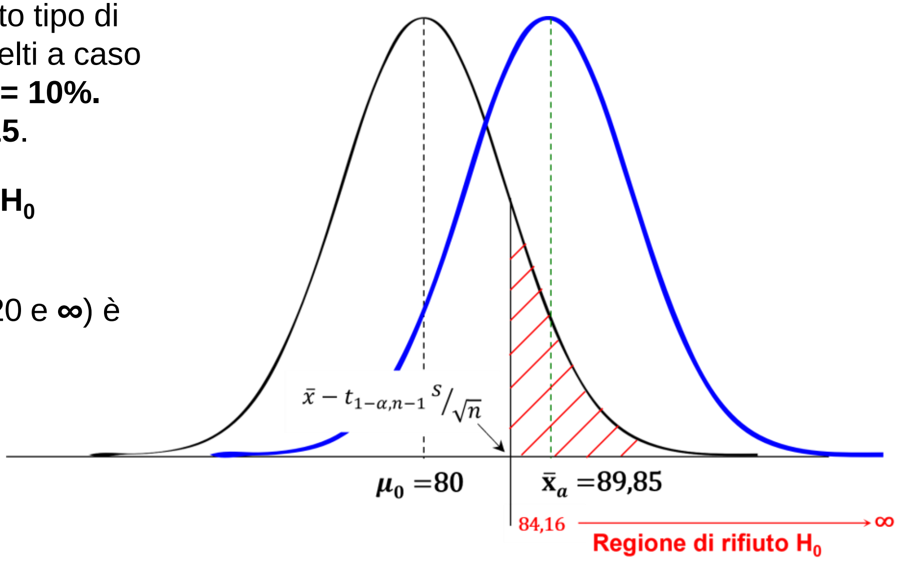
proporzione: $(p_a - Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p_a + Z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}})$

Esempio. Si vuole verificare se la durata media di un certo tipo di causa giudiziaria è più di **80** giorni utilizzando **20 casi** scelti a caso (H_0: $\mu_0 \leq 80$; H_a: $\mu_0 > 80$), con un livello di significatività **α = 10%.** La media del campione è **89,85** con una dev. std. di **19,15**.

La statistica di test ***t* = 2,3003** è **nella regione di rifiuto H_0** essendo maggiore di $t_{1-\alpha,n-1}$ = **1,328**.

L'**IC al 90%** ($\bar{x}$ - $t_{1-\alpha,n-1}$*s/√n cioè 89,85-(1,3277)*19,15/√20 e **∞**) è compreso tra [**84,16, +∞**].

Poiché μ_0 è **fuori dall'IC al 90%, si rifiuta l'ipotesi nulla H_0** per cui la durata media delle cause giudiziarie **è significativamente più grande** di 80 giorni.



Intervallo di Confidenza

Intervallo di confidenza per due campioni

test a 1 coda (sinistra), medie: $(-\infty, \bar{x}_a - \bar{x}_0 + t_{1-\alpha,n_a+n_0-2}\sqrt{\frac{s_a^2}{n_a} + \frac{s_0^2}{n_0}})$,

proporzioni: $(0, p_a - p_0 + Z_{1-\alpha}\sqrt{\frac{p_a(1-p_a)}{n_a} + \frac{p_0(1-p_0)}{n_0}})$

test a 1 coda (destra), medie: $(\bar{x}_a - \bar{x}_0 - t_{1-\alpha,n_a+n_0-2}\sqrt{\frac{s_a^2}{n_a} + \frac{s_0^2}{n_0}}, +\infty)$,

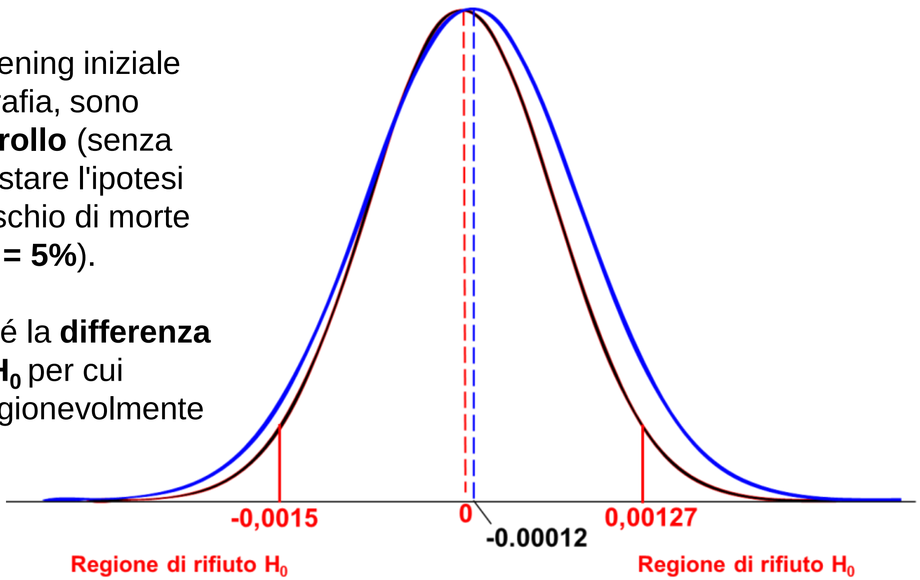
proporzioni: $(p_a - p_0 - Z_{1-\alpha}\sqrt{\frac{p_a(1-p_a)}{n_a} + \frac{p_0(1-p_0)}{n_0}}, 1)$

test a 2 code, medie: $\bar{x}_a - \bar{x}_0 \pm t_{1-\alpha/2,n_a+n_0-2}\sqrt{\frac{s_a^2}{n_a} + \frac{s_0^2}{n_0}}$,

proporzioni: $p_a - p_0 \pm Z_{1-\alpha/2}\sqrt{\frac{p_a(1-p_a)}{n_a} + \frac{p_0(1-p_0)}{n_0}}$

Esempio, Durante uno studio trentennale, dopo uno screening iniziale di 5 anni, **500 donne su 44.425**, sottoposte a mammografia, sono decedute per cancro al seno mentre **nel gruppo di controllo** (senza mammografia) ne sono morte **505 su 44.405**. Si vuole testare l'ipotesi che le mammografie possano diminuire o aumentare il rischio di morte rispetto ai normali esami (H_0: p_a-p_0=0; H_a: p_a-p_0≠0, con α **= 5%**).

L'IC al 95% sarà compreso tra [**-0,151%, 0,127%**]**;** poiché la **differenza zero è compresa nell'IC, non si rifiuta l'ipotesi nulla H_0** per cui la differenza nei tassi di mortalità per cancro al seno è ragionevolmente spiegata dal caso.



Il p-value

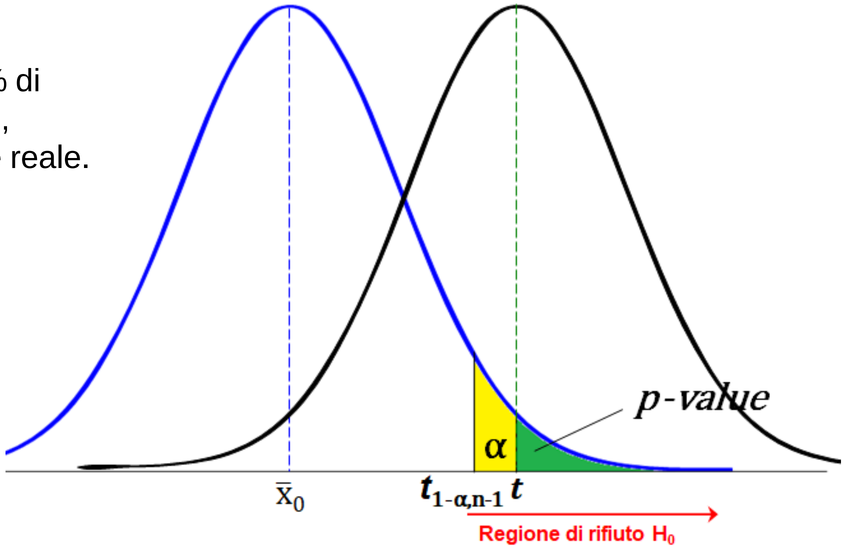
Definizione

Man mano che aumenta il valore assoluto della statistica di test ***z*** (o ***t***), **l'area sotto la curva oltre il valore di *t*** si riduce. Questa area, nota come "**p-value**", rappresenta la **probabilità di osservare un risultato uguale o più estremo** rispetto a quello osservato, assumendo che l'ipotesi nulla sia vera.

È importante sottolineare che il *p-value* non rappresenta la probabilità che l'ipotesi nulla sia vera o falsa. Piuttosto, è una **misura dell'evidenza statistica contro l'ipotesi nulla**: **più il *p-value* è piccolo, maggiore è l'evidenza** che il risultato non sia dovuto al caso**.**

Il *p-value* serve quindi a **valutare se la differenza** tra il risultato osservato e quello ipotizzato sia **statisticamente significativa**.

Per es. un *p-value* di 0,0062 indica che c'è solo lo 0,62% di probabilità di ottenere un risultato così estremo per caso, suggerendo che la differenza osservata è probabilmente reale.



Il p-value

Calcolo

Per **calcolare il *p-value*** si utilizza la **funzione di distribuzione cumulata (CDF)**, che misura **l'area sotto la curva della distribuzione a partire da un valore specifico**.

Distribuzione normale standardizzata (Z), per un livello di significatività α = 5%:

$Z_{1-\alpha} = Z_{0,95}$=**1,6448** da cui: *p-value* = 1 − $CDF_Z(1{,}6448) \approx$ **0,05**

Distribuzione t di Student (t), per un campione di n=25 e α = 5%:

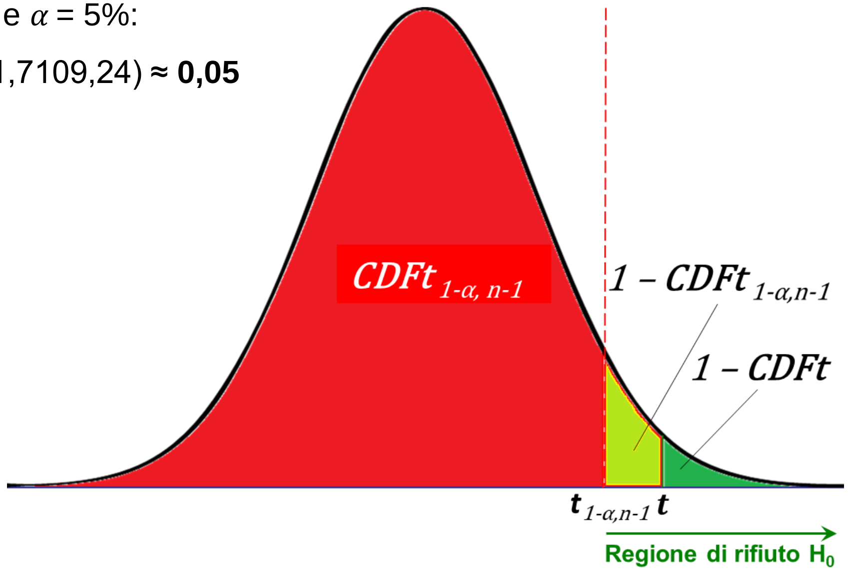
$t_{1-\alpha,n-1} = t_{0,95,24}$ =**1,7109** da cui: *p-value* = 1 − $CDF_t(1{,}7109{,}24) \approx$ **0,05**

Esempio. Si supponga di calcolare la statistica t e ottenere un valore di t **= 3,3696** con α **= 5% e** n **= 25**

In questo caso, il *p-value* è:
p-value = 1 − $CDF_t(3{,}3696{,}24)$ **= 0,00127**.

Poiché il *p-value* calcolato (0,00127) è inferiore al livello di significatività α = 5%, **la statistica t si trova nella regione di rifiuto della distribuzione**.

Questo significa che è possibile rifiutare l'ipotesi nulla e concludere che **il risultato osservato è statisticamente significativo**.



Il p-value

Interpretazione

La tabella seguente fornisce delle linee guida per valutare l'evidenza empirica contro l'ipotesi nulla.

p-value	Evidenza contro H_0
$p > 0,10$	Debole o nessuna
$0,05 < p \le 0,10$	Moderata
$0,01 < p \le 0,05$	Forte
$p \le 0,01$	Fortissima

Esempio. Secondo la Sorveglianza PASSI[1], nel biennio 2020-2021, la percentuale di adulti di età dai 18 ai 69 in Italia che sono in sovrappeso è del 42,9%. Il sindaco di una città teme che molti suoi cittadini siano in sovrappeso, sopra la percentuale nazionale. Da un campione casuale di 150 adulti di questa città si scopre che 76 sono classificati come sovrappeso (cioè il 50,67%). Si vuole verificare l'ipotesi che la percentuale in questa città sia maggiore della percentuale nazionale al livello di significatività del 5%.

Si vuole quindi testare

H_0: $\mu_0 \le 0,429$; H_a: $\mu_0 > 0,429$ per cui $Z = (p - \pi_0)/\sqrt{\pi_0(1-\pi_0)/n}$; Z = (0,5067-0,429)/√((0,429*0,571)/150); Z **= 1,922**.

- ✓ $Z_{1-\alpha}$ **= 1,645**. La statistica di test Z**= 1,923 è nella regione di rifiuto H_0** essendo maggiore di 1,645.
- ✓ L'**IC** al 95% (p_a - Z_α*s/√(n) cioè 0,5067-(1,645)*√((0,429*0,571)/150) e 1) è compreso tra [**0,44, 1**].
  Il parametro π_0 (**0,429**) **essendo fuori dall'IC**, porta a **rifiutare l'ipotesi nulla H_0**.
- ✓ Il ***p-value*** = 1- CDF_Z = **0,0273**.

Essendo **0,0273 ≤ 0,05**, **ci sono prove sufficienti** per affermare che la **proporzione è inferiore al 42,9%**.

[1] hptts://www.epicentro.iss.it/passi/infoPassi/infoGen

Utilizzo del software statistico

- Con il foglio di calcolo Microsoft Excel

Livello di Confidenza	Livello di significatività		Statistica di test	
$1-\alpha$	α	Valore Critico Z	Valore di test	p-value
0,95	0,05	-1,6449	-1,5	0,06681

=INV.NORM.S(D4)

=DISTRIB.NORM.ST.N(F4;VERO)

Numerosità campione	Livello di Confidenza	Livello di significatività		Statistica di test	
n	$1-\alpha$	α	Valore Critico t	Valore di test	p-value
30	0,95	0,05	-1,6991	-1,5	0,07221

=INVT(D11;B11-1)

=DISTRIB.T.N(F11;B11-1;VERO)

Utilizzo del software statistico

Con il software R[1]

1. Calcolo del valore critico

```
Liv_Confidenza = 0.95; alfa = 1-Liv_Confidenza; Coda_inferiore = alfa; n = 30

Val_Critico_Z = qnorm(Coda_inferiore); Val_Critico_Z
[1] -1.644854

Val_Critico_t = qt(Coda_inferiore, n-1); Val_Critico_t
[1] -1.699127
```

2. Calcolo del *p-value*

```
Statistica_test_Z = 1.5

# CDF Z
p_value_Z = pnorm(Statistica_test_Z, lower.tail²=FALSE); p_value_Z
[1] 0.0668072

# CDF t
Statistica_test_t = 1.5
p_value_t = pt(Statistica_test_t, n-1, lower.tail=FALSE); p_value_t
[1] 0.07221185
```

[1] Vedi Appendice
[2] FALSE considera la coda superiore (l'area alla destra del valore di test). Calcola la probabilità di ottenere un valore maggiore di quello specificato.
TRUE considera la coda inferiore (l'area alla sinistra del valore di test). Calcola la probabilità di ottenere un valore minore o uguale a quello specificato.

Dimensione del campione e Potenza del Test

Minimizzare il rischio di errore

Poiché il risultato del test porta a rifiutare o non rifiutare l'ipotesi nulla, senza sapere con certezza quale sia l'ipotesi vera, è fondamentale **minimizzare il rischio di errore**.

L'**errore di tipo α** (probabilità di rifiutare l'ipotesi nulla quando è vera) può essere **ridotto abbassando il livello di significatività**, per es. da 0,05 a 0,01. Questo consente di avere una fiducia del 99% (1-α) nel non rifiutare l'ipotesi nulla quando è vera.

D'altra parte, se si ritiene probabile che l'**ipotesi alternativa sia vera**, è importante **ridurre l'errore di tipo β**. Ridurre β aumenta la potenza del test (1-β), cioè la capacità di rilevare una differenza quando questa esiste realmente.

Un modo efficace per **aumentare la potenza del test**, senza sacrificare il livello di significatività α, è **incrementare la numerosità del campione**. Questo permette di ridurre sia α sia β, migliorando complessivamente la qualità del test.

Dimensione del campione e Potenza del Test

Dimensione minima del campione (differenze tra medie)

Fissati **il livello di significatività** (α), la potenza del test **desiderata** ($1-\beta$), una **stima dell'effetto atteso** e della variabilità dei dati, è possibile calcolare la dimensione minima del campione necessaria per il test.

Per le **differenze tra medie**: $$n = \frac{2\,\sigma^2 (Z_{1-\alpha/2} + Z_{1-\beta})^2}{d^2}$$

Esempio.

d	= 10	differenza ipotetica tra le medie (per calcolare il campione sufficiente a rilevare una diff. significativa)
sd	= 16	deviazione standard ipotetica
1-β	**= 0,80**	potenza del test
1-α	= 0,05	livello di significatività con un test t a 2 code

Per soddisfare questi requisiti, sono necessari almeno **42 soggetti per gruppo** (ossia **84 soggetti totali**).

Per il calcolo si può usare la funzione ***power.t.test*** del **software R**[1].

```
power.t.test (d=10, sd=16, sig.level=0.05, power=0.80, type="two",alt="two")
Two-sample t test power calculation
          n = 41.16902
      delta = 10
         sd = 16
  sig.level = 0.05
      power = 0.8
alternative = two.sided
NOTE: n is number in *each* group
```

[1] Vedi Appendice

Dimensione del campione e Potenza del Test

Dimensione minima del campione (proporzioni)

Fissati il **livello di significatività** (α), la **potenza del test desiderata** (1−β), **una stima dell'effetto atteso** e della variabilità dei dati, è possibile calcolare la dimensione minima del campione necessaria per il test.

Per le **differenze tra proporzioni**: $n = \dfrac{(p_a q_a + p_0 q_0)(Z_{1-\alpha/2} + Z_{1-\beta})^2}{(p_a - p_0)^2}$

Esempio.
p1 = 0,20 proporzione nel gruppo sperimentale
p2 = 0,15 proporzione nel gruppo di controllo
1-β = 0,90 potenza del test
1-α = 0,05 livello di significatività con un test t a 2 code

In questo caso servono almeno **1212 soggetti per gruppo**, per un **totale di 2024 soggetti**.

Per il calcolo si può usare la funzione ***power.prop.test*** del software **R**:

```
power.prop.test (p1=0.2, p2=0.15, sig.level= 0.05, power=0.90, alt="two")
Two-sample comparison of proportions power calculation

              n = 1211.529
             p1 = 0.2
             p2 = 0.15
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

Dimensione del campione e Potenza del Test

Rapporto di numerosità (differenze tra medie)

Se richiesta una **diversa numerosità dei campioni**, bisognerà tener conto dell'"***allocation ratio***" che determina come il campione totale viene suddiviso tra i gruppi. In alcune situazioni (ad es., negli studi clinici, quando il trattamento è costoso, possono essere giustificati rapporti diversi dall'ottimale 1:1).

Per le **differenze tra medie**:

Esempio.
delta = 10 differenza ipotetica tra le medie (per calcolare il campione sufficiente a rilevare una diff. significativa)
sd = 20 deviazione standard ipotetica
$1-\beta$ = 0,80 potenza del test
$1-\alpha$ = 0,05 livello di significatività con un test t a 2 code
rapp = 2/3 il secondo campione più piccolo del 66,67% rispetto al primo

Sono quindi necessari **80 soggetti** per il **1° gruppo** e **54 soggetti** per il **2° gruppo** (ossia **134 soggetti totali**).

Per il calcolo si può usare la funzione ***ttest.2samp*** del **software R**.

```
library(powertools)
ttest.2samp(delta = 10, sd1 = 20, power = 0.80, alpha = 0.05, sides = 2, n.ratio = 2/3, v=T)
Two-sample t test power calculation

    n1, n2 = 79.85539, 53.23692
     delta = 10
  sd1, sd2 = 20, 20
     alpha = 0.05
     power = 0.8
     sides = 2
```

Dimensione del campione e Potenza del Test

Rapporto di numerosità (differenze tra proporzioni)

Se richiesta una **diversa numerosità dei campioni**, bisognerà tener conto dell'"***allocation ratio***" che determina come il campione totale viene suddiviso tra i gruppi. In alcune situazioni (ad es., negli studi clinici, quando il trattamento è costoso, possono essere giustificati rapporti diversi dall'ottimale 1:1).

Per le **differenze tra proporzioni**:

Esempio.
p1 = 0,20 proporzione nel gruppo sperimentale
p2 = 0,15 proporzione nel gruppo di controllo
1-β = 0,80 potenza del test
1-α = 0,05 livello di significatività con un test t a 2 code
marg = 0 margine di non-inferiorità o superiorità (0 per difetto).
rapp = 2/3 il secondo campione più piccolo del 66,67% rispetto al primo

Sono quindi necessari **1103 soggetti** per il **1° gruppo** e **736 soggetti** per il **2° gruppo** (ossia **1839 soggetti totali**).

Per il calcolo si può usare la funzione ***prop.2samp*** del **software R**.

```
library(powertools)
prop.2samp(p1=0.2, p2=0.15, margin=0, power=0.8, alpha=0.05, sides = 2, n.ratio = 2/3, v=T)
Two sample comparison of proportions power calculation

    n1, n2 = 1102.7676, 735.1784
    p1, p2 = 0.20, 0.15
    margin = 0
     alpha = 0.05
     power = 0.8
     sides = 2
```

Dimensione del campione e Potenza del Test

Calcolo della Potenza del Test (differenze tra medie)

Fissato un **livello di significatività** α, la **numerosità dei campioni** n e una **differenza attesa** tra le ipotesi, è possibile **calcolare la potenza del test** $(1-\beta)$. Questa rappresenta la probabilità di rifiutare correttamente l'ipotesi nulla quando questa è falsa.

Esempio.

n	**= 35**	numero di campioni in ciascun gruppo
d	= 10	differenza significativa ipotetica tra le due medie
sd	= 16	deviazione standard ipotetica
1-α	= 0,05	livello di significatività con un test t a 2 code

Si ottiene una **potenza del test 1-β** = **0,732**, che è la probabilità di rifiutare correttamente l'ipotesi nulla e **β = 0,268**, che è probabilità di commettere un errore del 26,8% nel non rifiutare l'ipotesi nulla quando questa è falsa.

Per il calcolo si può usare la funzione ***power.t.test*** del **software R**.

```
power.t.test (n=35, d=10, sd=16, sig.level= 0.05, type="two.sample", alternative="two")
Two-sample t test power calculation

          n = 35
      delta = 10
         sd = 16
  sig.level = 0.05
      power = 0.7315533
alternative = two.sided

NOTE: n is number in *each* group
```

Dimensione del campione e Potenza del Test

Calcolo della Potenza del Test (proporzioni)

Fissato un **livello di significatività** α, la **numerosità dei campioni** n e una **differenza attesa** tra le ipotesi, è possibile **calcolare la potenza del test** $(1-\beta)$. Questa rappresenta la probabilità di rifiutare correttamente l'ipotesi nulla quando questa è falsa.

Esempio.
n **= 1000** numero di campioni in ciascun gruppo
p1 = 0,20 proporzione nel gruppo sperimentale
p2 = 0,15 proporzione nel gruppo di controllo
1-α = 0,05 livello di significatività con un test t a 2 code

Si ottiene una **potenza del test 1-β** = **0,837**, che è la probabilità di rifiutare correttamente l'ipotesi nulla. Con un valore di **β** = **0,163**, c'è una probabilità del 16,3% di non rifiutare l'ipotesi nulla anche quando questa è effettivamente falsa.

Per il calcolo si può usare la funzione ***power.t.test*** del **software R**

```
power.prop.test (n=1000, p1=0.2, p2=0.15, sig.level= 0.05, alternative="two")
Two-sample t test power calculation
            n = 1000
           p1 = 0.2
           p2 = 0.15
    sig.level = 0.05
        power = 0.8375944
  alternative = two.sided
NOTE: n is number in *each* group
```

Conduzione del test

Passaggi operativi per condurre un Test

Per condurre un test di ipotesi bisogna definire i passaggi essenziali:

- ✓ ***Raccogliere un campione casuale dei dati***
- ✓ ***Verificare le assunzioni*** *sulla distribuzione e specificare i parametri (per es.* $\bar{x}$ o μ*).*
- ✓ ***Formulare le ipotesi****:* H_0 *(ipotesi nulla),* H_a *(ipotesi alternativa); test a 1 o 2 code*
- ✓ ***Stabilire il livello di significatività*** (α)
- ✓ ***Calcolare la statistica di test****, l'intervallo di confidenza e il p-value.*
- ✓ ***Prendere una decisione*** *basata sui risultati.*

Esempi

Questi esempi utilizzano la funzione ***h.test***[1] (vedi Appendice)

Esempio 1: livello colesterolo tra popolazioni (medie, 1 campione, 1 coda)

Esempio 2: sondaggio popolarità (prop., 1 campione, 1 coda)

Esempio 3: efficacia nuova pubblicità (prop., 1 campione, 1 coda)

Esempio 4: controllo qualità (medie, 1 campione, 2 code)

Esempio 5: efficacia nuovo farmaco (medie, 2 campioni, 2 code)

Esempio 6: scelta tra due soluzioni di web design (prop., 2 campioni)
A/B Test

Esempio 7: differenza IMC tra generi (medie, 2 campioni, 2 code)

Esempio 8: differenza diabete tra i due generi (prop., 2 campioni, 2 code)

Esempio 9: miglioramento processo (prop., 1 campione, 1 coda)
distribuzione binomiale

Esempio 10: differenza di peso prima e dopo una dieta (medie, stesso campione, 2 code)
campioni appaiati

Esempio 11: valutazione del premier dopo sei mesi (prop., stesso campione, 2 code)
campioni appaiati

Esempio 12: efficacia farmaco rispetto al precedente (medie, 2 campioni, 1 coda)
dimensione campioni data potenza del test con diverso rapporto di numerosità
test di non inferiorità

Esempio 13: efficacia trattamento oncologico rispetto al precedente (prop., 2 campioni, 1 coda)
test di non inferiorità

[1] Sviluppata dall'autore delle presenti note, questa funzione offre un approccio facilitato al calcolo e l'interpretazione di test di ipotesi semplificando il processo decisionale.

Esempi

Esempio 1: test livello di colesterolo (1 campione su popolazione)

1. Raccolta dei dati e obiettivo del test

   Data una popolazione maschile che va dai 20 ai 74 anni si ha un livello medio di colesterolo di 211 mg/100 ml con una deviazione standard di 46 mg/100 ml. Si vuole testare l'ipotesi alternativa che, preso un campione maschile di 25 soggetti di età compresa tra i 20 e i 24 anni, il livello medio di colesterolo sia più basso e che quindi le due distribuzioni di probabilità siano diverse.

Campione	n.: 25	media: 180 mg/100 ml	dev.std. : **-**
Popolazione	n: **-**	media: 211 mg/100 ml	dev. std.: 46 mg/100 ml

2. Ipotesi

   H_0: $\overline{x}_a \geq \overline{x}_0$; H_a: $\overline{x}_a < \overline{x}_0$ il test è a 1 coda (sinistra)

3. Livello di significatività $\boldsymbol{\alpha}$ **= 5%** con **24** (n-1) **gradi di libertà** a cui corrisponde il valore critico $\boldsymbol{t_\alpha}$ **= -1,7109**

4. Test statistici

   $t = (\overline{x} - \mu) / (\sigma/\sqrt{n})$, ***t*** **= (180 - 211) / (46/√24) = -3,3696** a cui corrisponde un ***p-value*** **= 0,00127**
   e un '**IC** al 95% (-∞, $\bar{x}_0 + Z_{1-\alpha}$*s/√n ,180+1,7109*46/√25 = 195,74) compreso tra [**-∞**, **195,74**].

5. Decisione

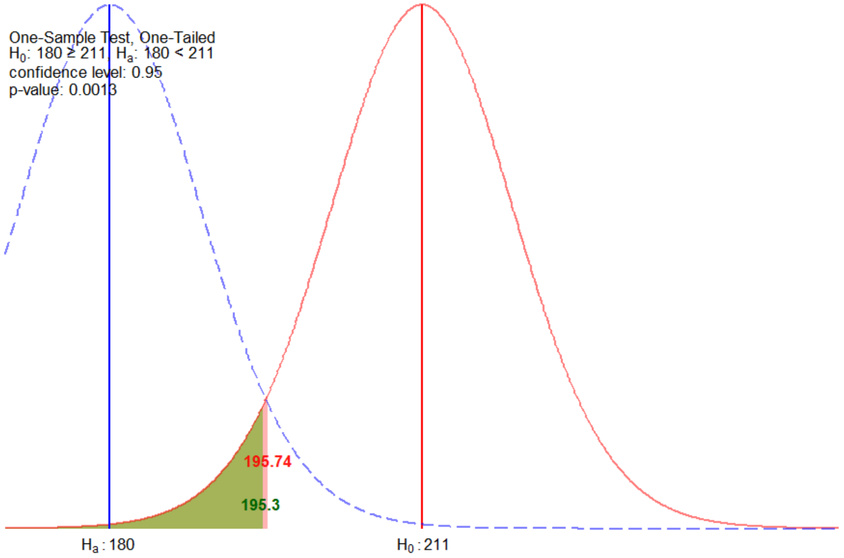
   Questo risultato suggerisce che i giovani adulti potrebbero avere un profilo lipidico più favorevole rispetto alla popolazione generale. L'analisi statistica ha evidenziato una **differenza significativa** tra il livello medio di colesterolo nei **giovani adulti** (20-24 anni) e quello della **popolazione maschile generale**. Il campione di giovani adulti mostra un livello medio di colesterolo significativamente inferiore, come confermato dal valore del *p-value* molto basso (0,00127).

   Questo risultato suggerisce che i **giovani adulti possiedano un profilo lipidico più favorevole** rispetto alla popolazione generale.

Esempi

Esempio 1: test livello di colesterolo (1 campione su popolazione)

```
h.test(n_a=25, x_a=180, sd_0=46, x_0=211, alt= "less", conf_level=0.95, graph=TRUE)

One-Sample Test, One-Tailed. Student's t distribution
H0: 180 ≥ 211, Ha: 180 < 211, confidence level = 0.95, alpha = 0.05

test statistic = -3.36957, critical value (0.05,24) = -1.7109, Std. Err. = 9.2
95% confidence interval: from -Inf to 195.7401
p-value = 0.00127

the null hypothesis H0 is rejected at the 5% significance level
```



Esempi

Esempio 2: test popolarità candidato sindaco (1 campione su popolazione)

1. Raccolta dei dati e obiettivo del test

Viene effettuato un sondaggio per prevedere se un candidato alla carica di sindaco di una città vincerà il ballottaggio. Vengono fatte 200 interviste, nelle quali il candidato riceve 105 preferenze. Durante le precedenti elezioni il candidato aveva vinto con il 55% delle preferenze. Si può affermare ad un livello di significatività del 5% che il candidato ha perso popolarità?

Campione	proporzione: 0,525 (105 / 200)	
Popolazione	proporzione: 0,55	**- / -**

2. Ipotesi

$H_0: p_a \geq p_0$; $H_a: p_a < p_0$ il test è a 1 coda (sinistra)

3. Livello di significatività $\boldsymbol{\alpha}$ **= 5%** a cui corrisponde il valore critico Z_α **= -1,6448**

4. Test statistici

$Z = (p_a - p_0) / \sqrt{p_0(1-p_0)/n}$, Z = (0,525-0,55)/√(0,55*0,45/200) **= -0,7107** a cui corrisponde un ***p-value* = 0,23864**. L'IC al 95% (*0,* $p_a - p_0 + Z_{1-\alpha}\sqrt{p(1-p)/n}$) = *0, 0,525+1,6448*√(0,55*0,45/200)*) è compreso tra [**0, 0,5828**].

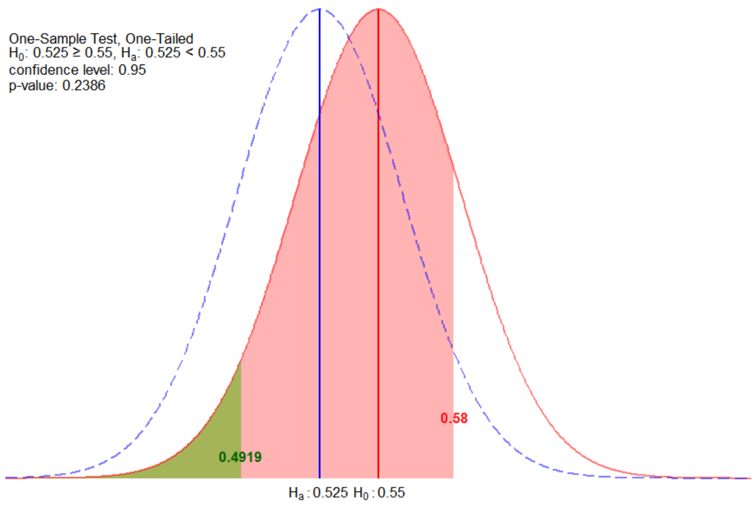
5. Decisione

Poiché il valore della statistica test Z (-0,7107) è maggiore del valore critico (-1,6448) e il *p-value* è superiore a 0,05, **non si può rifiutare l'ipotesi nulla**. Questo significa che, al livello di significatività del 5%, **non ci sono prove sufficienti per affermare che il candidato abbia perso popolarità** rispetto alle elezioni precedenti. Inoltre, l'IC al 95% per la proporzione di voti ottenuti dal candidato include il valore ipotizzato del 55%, confermando tale conclusione.

Esempi

Esempio 2: test popolarità candidato sindaco (1 campione su popolazione)

```
h.test(p_a=0.525, N_a=200, p_0=0.55, alt= "less", correct=FALSE, conf_level=0.95, graph=T)

For a more precise estimate, the binomial test is recommended.
Use the argument binom = TRUE.

One-Sample Test, One-Tailed. Chi-square distribution
H0: 0.525 ≥ 0.55, Ha: 0.525 < 0.55, confidence level = 0.95, alpha = 0.05

test statistic = -0.71067, critical value (0.05) = -1.6449, Std. Err. = 0.035
95% confidence interval: from 0 to 0.5824
p-value = 0.23864

the null hypothesis H0 is not rejected at the 5% significance level
```



[1] Con `correct=FALSE` non viene applicata la correzione di continuità di Yates per compatibilità con i calcoli manuali.

Esempi

Esempio 3: test nuova tecnica di comunicazione (1 campione su popolazione)

1. Raccolta dei dati e obiettivo del test

   L'ufficio marketing di una azienda telefonica afferma che la nuova pubblicità online è più efficace di quella usata in precedenza, che dava un tasso di visite al sito del 25%. È stato scelto un campione di 200 clienti sul quale è stata testata questa pubblicità in anteprima e 63 di loro sono andati a visitare il sito. Al livello di significatività α = 5%, si può provare quanto affermano le persone del marketing?

Campione	proporzione: 0,315 (63 / 200)
Popolazione	proporzione: 0,25 **- / -**

2. Ipotesi

   H_0: $p_a \leq p_0$, H_a: $p_a > p_0$ il test è a 1 coda (destra)

3. Livello di significatività $\boldsymbol{\alpha}$ = 5% a cui corrisponde il valore critico $Z_{1-\alpha}$ = 1,6449

4. Test statistici

   $Z > Z_{1-\alpha}$ (2,12289 > 1,6449); p_0 (0,25) è fuori dall'IC al 95% [0,26375, 1] e il *p-value* (0,01688) < 0,05

5. Decisione

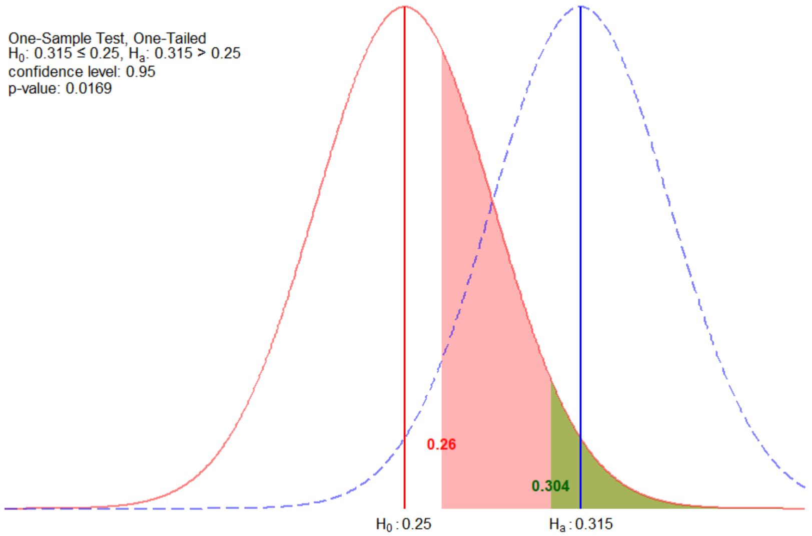
   I risultati dell'analisi statistica indicano che la **nuova campagna pubblicitaria online ha superato le aspettative**, generando un **tasso di visite al sito significativamente superiore** rispetto alla campagna precedente. Con un livello di confidenza del 95%, si può affermare che la nuova pubblicità è più efficace nell'attirare l'attenzione dei consumatori e nel guidarli verso il sito web.

   Questi risultati positivi suggeriscono che **l'investimento nella nuova campagna è stato fruttuoso** e che potrebbe portare a un aumento delle conversioni e del fatturato.

Esempi

Esempio 3: test nuova tecnica di comunicazione (1 campione su popolazione)

```
h.test (n_a=63, N_a=200, p_0=0.25, alt="greater", correct=F, conf_level=0.95, graph=T)

For a more precise estimate, the binomial test is recommended.
Use the argument binom = TRUE.

One-Sample Test, One-Tailed. Chi-square distribution
H0: 0.315 ≤ 0.25, Ha: 0.315 > 0.25, confidence level = 0.95, alpha = 0.05

test statistic = 2.12289, critical value (0.95) = 1.6449, Std. Err. = 0.03285
95% confidence interval: from 0.2637 to 1
p-value = 0.016881

the null hypothesis H0 is rejected at the 5% significance level
```



Esempi

Esempio 4: test controllo qualità (1 campione su popolazione)

1. Raccolta dei dati e obiettivo del test

   Un'azienda che produce bottiglie d'acqua da 500 ml. Si vuole verificare se la media del contenuto delle bottiglie prodotte è effettivamente 500 ml o se c'è una differenza significativa (in eccesso o in difetto).

Campione	n.: 30	media: 495 ml	dev.std. : 5 ml
Popolazione	n: **-**	media: 500 ml	dev. std.: -

2. Ipotesi

   H_0: $\overline{x}_a - \overline{x}_0 = 0$; H_a: $\overline{x}_a - \overline{x}_0 \neq 0$ il test è a 2 code

3. Livello di significatività $\boldsymbol{\alpha}$ = 5% a cui corrisponde il valore critico $t_{1-\alpha/2,\, n-1}$ = -2,045

4. Test statistici

   $t = (\overline{x} - \mu) / (\sigma/\sqrt{n})$, t = (495-500) / (5/√30) = -5,477 ; $\boldsymbol{t < t_{1-\alpha/n-1}}$ (-5,48 > -2,045)

   $t < t_{1-\alpha/2,\, n-1}$ (-5,477 < -2,045); il valore teorico atteso, 500, è fuori dall'IC al 95% [493,13 , 496,87] e il *p-value* (0,00000674) < 0,05

5. Decisione

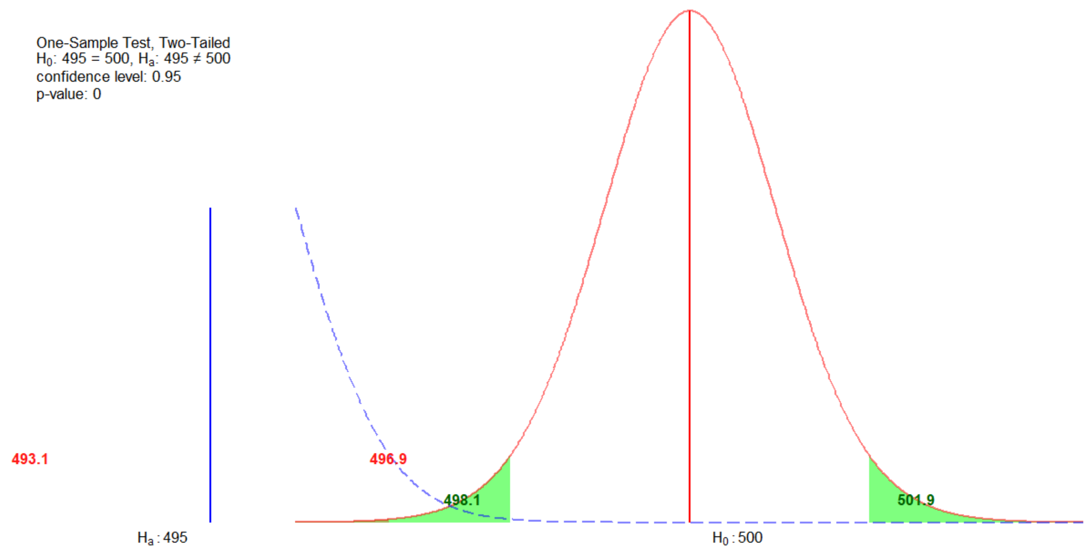
   Poiché il valore di t cade nella regione di rifiuto, si conclude che c'è una **differenza significativa** tra il volume medio delle bottiglie e il valore teorico di 500 ml. Si può anche affermare che **il volume medio delle bottiglie tende a essere inferiore rispetto a quello dichiarato**. Di conseguenza, l'azienda dovrebbe indagare sul processo produttivo per capire la causa di questa discrepanza.

Esempi

Esempio 4: test controllo qualità (1 campione su popolazione)

```
h.test (x_a=495, sd_a=5, n_a=30, x_0=500, alt="two", conf_level=0.95, graph=T)
```

```
One-Sample Test, Two-Tailed. Student's t distribution
H0: 495 = 500, Ha: 495 ≠ 500, confidence level = 0.95, alpha/2 = 0.025

test statistic = -5.47723, critical value (0.025,29) = -2.0452, Std. Err. = 0.913
95% confidence interval: from 493.133 to 496.867
p-value = 6.74e-06

the null hypothesis H0 is rejected at the 5% significance level
```



Esempi

Esempio 5: test efficacia statine (2 campioni)

1. Raccolta dei dati e obiettivo del test

   Da uno studio teso a confrontare i valori di colesterolemia in due gruppi di soggetti, trattati con statine o con placebo, dopo un anno di trattamento, si vuole testare l'efficacia del farmaco.

Gruppo sperimentale (statina)	N.: 40	Media, mg/dL: 194,4	Dev. Std.: 15,6
Gruppo di controllo (placebo)	N.: 40	Media, mg/dL: 202,1	Dev. Std.: 14,8

2. Ipotesi

   H_0: $\bar{x}_a - \bar{x}_0 = 0$; H_a: $\bar{x}_a - \bar{x}_0 \neq 0$ il test è a 2 code

3. Livello di significatività $\alpha = 5\%$ a cui corrisponde il valore critico $t_{1\text{-}\alpha/2,\, n1+n2\text{-}2}$ **= -1,990847**

4. Test statistici

   $t = (\bar{x}_a - \bar{x}_0) \,/\, \sqrt{s_a^2/n_a + s_0^2/n_0}$, t= **-2,265** con l'IC: $(\bar{x}_a$ - $\bar{x}_0) \pm |t|_{\alpha/2}{}^* \sqrt{s_a^2/n_a + s_0^2/n_0}$, -7,7 ± 1,991* √ (15,6²/40+14,8²/40) = -7,7 ± 1,991 *3,4 compreso tra [**-14,47, -0,93**] e un *p-value* = 0,026.

5. Decisione

   Lo studio ha dimostrato che il **trattamento con statine è efficace** nel ridurre significativamente i livelli di colesterolo nel sangue. I pazienti trattati con statine hanno mostrato una **riduzione media di 7,7 mg/dL** rispetto al gruppo placebo, con un errore di ± 6,77, entro un IC al 95% di [-14,47, -0,93] e un *p-value* di 0,0263, inferiore a 0,05.

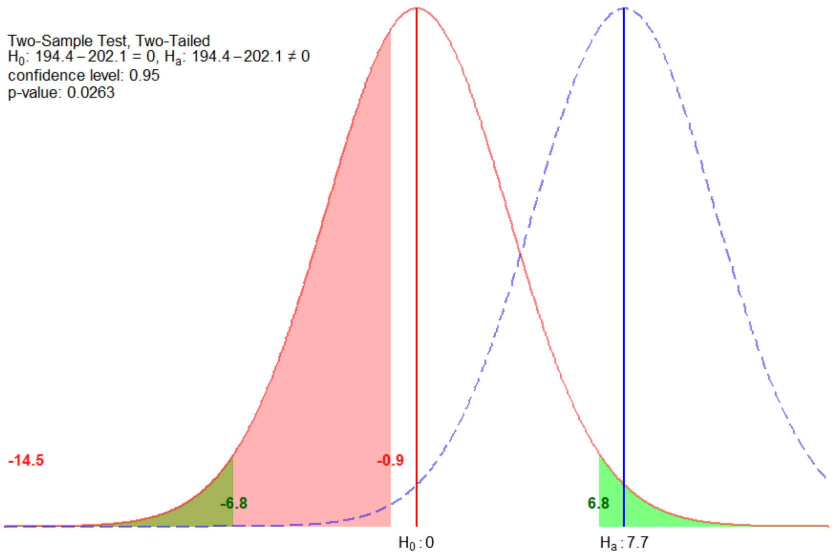
   Questi risultati suggeriscono **che le statine costituiscono un'opzione terapeutica valida** per la gestione dell'ipercolesterolemia.

Esempi

Esempio 5: test efficacia statine (2 campioni)

```
h.test(x_a=194.4, sd_a=15.6, n_a=40, x_0=202.1, sd_0=14.8, n_0=40, alt="two",
conf_level=0.95, manual=TRUE, graph=T)
```

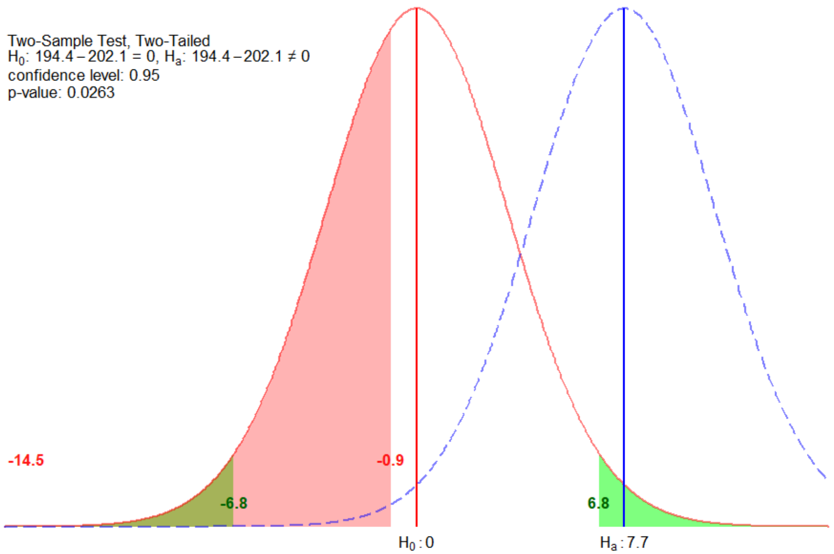
```
Two-Sample Test, Two-Tailed. Student's t distribution
H0: 194.4-202.1 = 0, Ha: 194.4-202.1 ≠ 0, confidence level = 0.95, alpha/2 = 0.025

test statistic = -2.26471, critical value (0.025,78) = -1.9909, Std. Err. = 3.4
95% confidence interval: from -14.4692 to -0.9308
p-value = 0.026307

the null hypothesis H0 is rejected at the 5% significance level
```



Esempi

Esempio 5: test efficacia statine (2 campioni)

```
h.test(x_a=194.4, sd_a=15.6, n_a=40, x_0=202.1, sd_0=14.8, n_0=40, alt="two",
conf_level=0.95, manual=FALSE¹, graph=TRUE)

Two-Sample Test, Two-Tailed. Student's t distribution
H0: 194.4-202.1 = 0, Ha: 194.4-202.1 ≠ 0, confidence level = 0.95, alpha/2 = 0.025

test statistic = -2.26471, critical value (0.025,77.78²) = -1.9909, Std. Err. = 3.4
95% confidence interval: from -14.4692 to -0.9308
p-value = 0.026315

the null hypothesis H0 is rejected at the 5% significance level
```



[1] I calcoli della funzione *t.test()* differiscono da quelli ottenuti manualmente (ad es., per piccoli campioni applica una correzione della varianza).
[2] I gradi di libertà sono calcolati con il metodo **Satterthwaite** (per varianze diverse tra gruppi).

Esempi

Esempio 6: A/B Test (2 campioni)

1. Raccolta dei dati e obiettivo del test

   Un'azienda di Digital Marketing vuol sapere se le due nuove soluzioni (A e B) per le campagne di e-mail sono più efficaci della soluzione finora adottata (Winner) attraverso due A/B test[1] in termini di rapporto *click-through* (numero di clic sul numero di consegne di messaggi).

Versione	Open	Click	Click/Open
A	205	24	0,11707
B	194	23	**0,11856**
Winner	582	51	0,08763

   Si potrebbe pensare che la versione B sia abbastanza buona per sceglierla.
   Si fanno quindi 2 test per "quantificare" questa fiducia.

2. Ipotesi

   H_0: p_a (A) **-** p_0 (Winner) = 0 , p_a (A) **-** p_0 (Winner) ≠ 0   il test è a 2 code
   H_0: p_a (B) **-** p_0 (Winner) = 0 , p_a (B) **-** p_0 (Winner) ≠ 0   il test è a 2 code

3. Livello di significatività α = 5% a cui corrisponde il valore critico $Z_{1-\alpha/2}$ = 1,96

4. Test statistici (vedi risultati nelle pagine seguenti)

Test	*p-value*
A / Winner	**0,2451**
B / Winner	**0,2342**

5. Decisione

   Sebbene le nuove soluzioni per le campagne di e-mail abbiano mostrato un leggero miglioramento rispetto alla soluzione Winner, con **i risultati dei test A/B non è possibile affermare con certezza che le nuove soluzioni siano più efficaci della soluzione già in uso** piuttosto che al caso.

   Per trarre conclusioni più definitive, potrebbero essere necessarie ulteriori indagini con campioni più grandi o test di durata più lunga

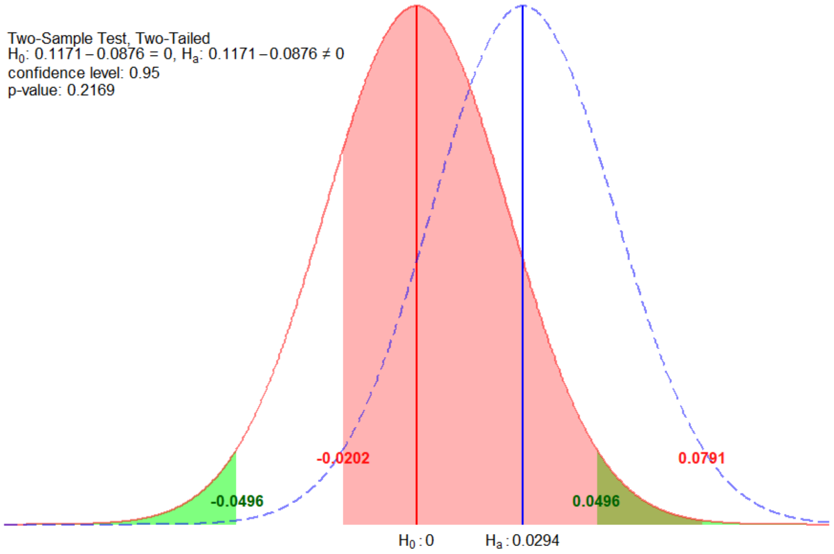
[1] https://en.wikipedia.org/wiki/A/B_testing#Email_marketing

Esempi

Esempio 6: A/B Test (2 campioni)

```
h.test(n_a=24, N_a=205, n_0=51, N_0=582, alt="two", correct=F, conf_level=0.95, graph=T)
```

```
Two-Sample Test, Two-Tailed. Chi-square distribution
H0: 0.1171-0.0876 = 0, Ha: 0.1171-0.0876 ≠ 0, confidence level = 0.95, alpha/2 = 0.025

test statistic = 1.16244, critical value (0.975) = 1.96, Std. Err. = 0.02533
95% confidence interval: from -0.0202 to 0.0791
p-value = 0.24506

the null hypothesis H0 is not rejected at the 5% significance level
```



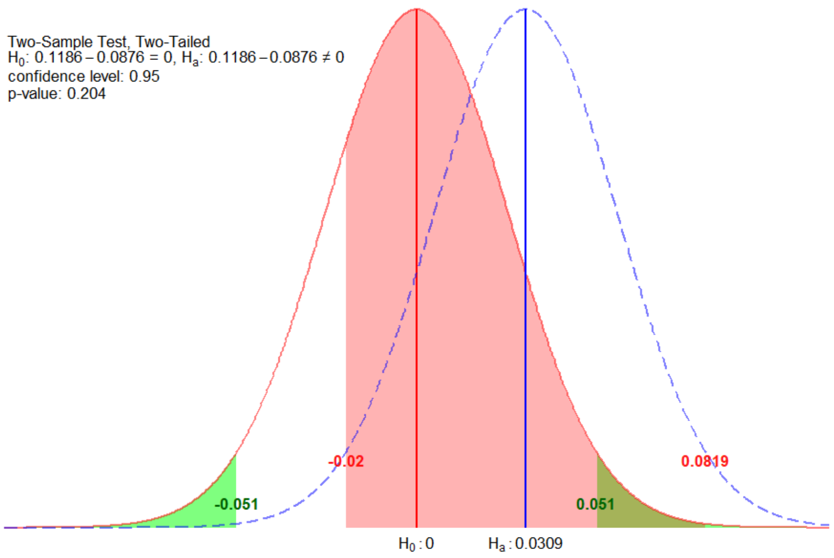
Segue

Esempi

Esempio 6: A/B Test (2 campioni)

```
h.test(n_a=23, N_a=194, n_0=51, N_0=582, alt="two", correct=F, conf_level=0.95, graph=T)
```

```
Two-Sample Test, Two-Tailed. Chi-square distribution
H0: 0.1186-0.0876 = 0, Ha: 0.1186-0.0876 ≠ 0, confidence level = 0.95, alpha/2 = 0.026

test statistic = 1.1895, critical value (0.975) = 1.96, Std. Err.= 0.02321
95% confidence interval: from -0.02 to 0.0819
p-value = 0.23424

the null hypothesis H0 is not rejected at the 5% significance level
```



Esempi

Esempio 7: test differenza di grasso corporeo tra uomini e donne (2 campioni)

1. Raccolta dei dati e obiettivo del test

Un modo per valutare lo stato di forma di una persona è quello di misurarne la percentuale di grasso corporeo. Le percentuali medie del grasso corporeo di un individuo variano con l'età, ma, secondo alcune linee guida, dovrebbero rientrare nel 15-20 % per gli uomini e nel 20-25 % per le donne. I dati campione vengono da un gruppo di uomini e donne che si sono allenati in palestra tre volte a settimana per un anno. Le medie sono “abbastanza simili” per concludere che il grasso corporeo medio è lo stesso anche per la popolazione più ampia di uomini e donne che frequentano le palestre?

Gruppo uomini	**13,3**	**6,0**	**20,0**	**8,0**	**14,0**	**19,0**	**18,0**	**25,0**	**16,0**	**24,0**	**15,0**	**1,0**	**15,0**
Gruppo donne	**22,0**	**16,0**	**21,7**	**21,0**	**30,0**	**26,0**	**12,0**	**23,0**	**28,0**	**23,0**			

2. Ipotesi

H_0: p_a (Gr. Donne) - p_0 (Gr. Uomini) $= 0$, p_a (Gr. Donne) - p_0 (Gr. Uomini) $\neq 0$ il test è a 2 code

3. Livello di significatività α = 5% a cui corrisponde il valore critico $t_{1-\alpha/2,\, n1+n2-2}$ = 2,0797

4. Test statistici

$t = (\bar{x}_a - \bar{x}_0) \,/\, \sqrt{s_a^2/n_a + s_0^2/n_0}$ = (22,29-14,94) / $\sqrt{}$(5,31²/10 + 6,84²/ 13) = 2,8958

$t > t_{1-\alpha}$ (2,8958 > 2,0797); 0 è fuori dall'IC al 95% [-12,618, -2,067] e il *p-value* (0,0086) < 0,05

5. Decisione

L'analisi statistica ha evidenziato una **differenza significativa nella percentuale media di grasso corporeo tra gli uomini e le donne che frequentano la palestra**. Questo risultato suggerisce che, contrariamente all'ipotesi nulla, **gli uomini e le donne non presentano la stessa composizione corporea**, anche dopo un periodo di allenamento regolare.

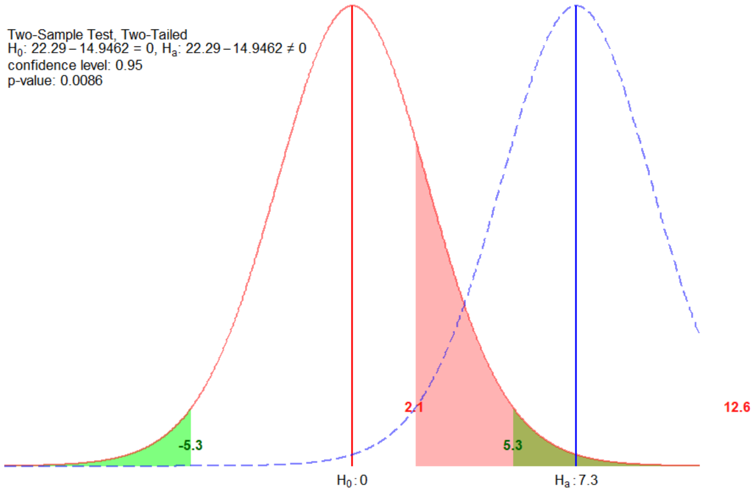
Le differenze osservate sono in linea con le linee guida esistenti e sottolineano l'importanza di considerare il genere come un fattore determinante nella valutazione dello stato di forma fisica.

Esempi

Esempio 7: test differenza di grasso corporeo tra uomini e donne (2 campioni)

```
gm = c(13.3 ,  6.0 , 20.0 ,  8.0 , 14.0 , 19.0 , 18.0 , 25.0 ,16.0 , 24.0 , 15.0 , 1.0 , 15.0)
media_gm = mean(gm); sd_gm = sd(gm); n_gm = length(gm)
gf = c(22.0 , 16.0 , 21.7 , 21.0 , 30.0 , 26.0 , 12.0 , 23.2 , 28.0 , 23.0)
media_gf = mean(gf); sd_gf = sd(gf); n_gf = length(gf)
h.test(x_a=media_gm, n_a=n_gm, sd_a=sd_gm, x_0=media_gf, n_0=n_gf, sd_0=sd_gf, alt="two", conf_level=0.95,
manual=F, graph=T)
```

```
Two-Sample Test, Two-Tailed. Student's t distribution
H0: 22.29-14.9462 = 0, Ha: 22.29-14.9462 ≠ 0, confidence level = 0.95, alpha/2 = 0.025
test statistic = 2.89579, critical value (0.975,20.99[1]) = 2.0797, Std. Err. = 2.536
95% confidence interval: from 2.0697 to 12.618
p-value = 0.00865
the null hypothesis H0 is rejected at the 5% significance level
```



[1] I gradi di libertà sono calcolati con il metodo **Satterthwaite** (per varianze diverse tra gruppi).

Esempi

Esempio 8: test su differenza diabete tra uomini e donne anziani (2 campioni)

1. Raccolta dei dati e obiettivo del test

   È stato condotto un sondaggio casuale su 705 uomini anziani, di età pari o superiore a 70 anni, e il 14,7% è risultato affetto da diabete di tipo 2. Allo stesso tempo, è stato condotto un sondaggio casuale su 688 donne anziane, di età pari o superiore a 70 anni, e il 9,1% delle donne è risultato affetto da diabete di tipo 2.
   Si vuole verificare se c'è una differenza significativa tra i due campioni.

2. Ipotesi

   L'ipotesi nulla è che non ci sia una differenza significativa.
   H_0: 0,147 (104/705) - 0,0916 = 0, H_a: 0,147 - 0,091 ≠ 0   il test è a 2 code

3. Livello di significatività con α = 5% il valore critico $Z_{1-\alpha/2}$ = 1,960

4. Test statistici

   $Z = (p_a - p_0) / \sqrt{p_a(1-p_a)/n_a + p_0(1-p_0)/n_0}$ = (0,147-0,091)/ √(0,147*0,853/705+ 0,091*0,909/688) = 3,2435.

   L'IC al 95% è compreso tra [0,02216, 0,08984] e il *p-value* (0,00127) < 0,05.

   La **differenza tra le proporzioni dei due campioni è** 0,056 (**5,6%**), con un margine di errore del **3,38%.**

5. Decisione

   I risultati di questo studio suggeriscono che gli **uomini anziani hanno una probabilità significativamente più alta di sviluppare il diabete di tipo 2** rispetto alle donne della stessa fascia d'età. Questa differenza potrebbe essere dovuta a una combinazione di fattori biologici, comportamentali e sociali.

   Tuttavia, è importante sottolineare che questo studio si basa su un campione specifico e potrebbe non essere generalizzabile a tutta la popolazione anziana.

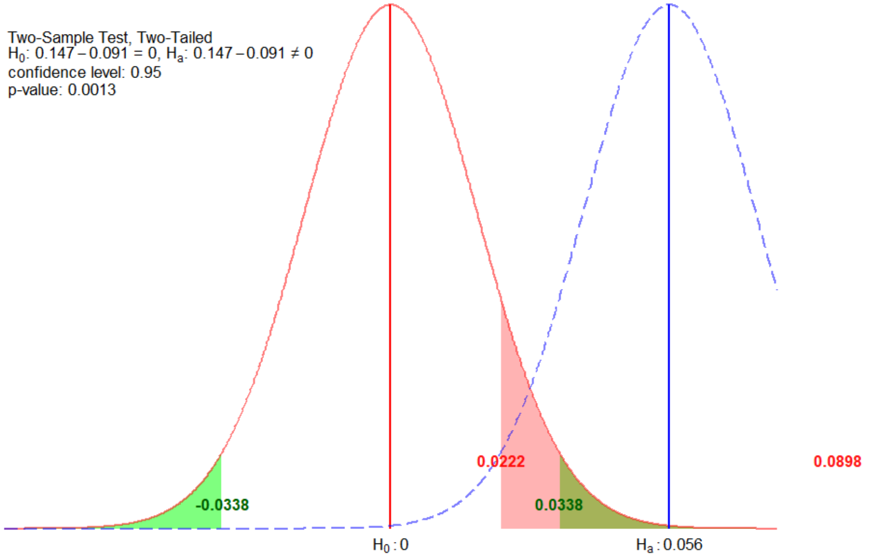
[1] Journal of Clinical Endocrinology & Metabolism, october 2016

Esempi

Esempio 8: test differenza diabete tra uomini e donne anziani (2 campioni)

```
h.test(p_a=0.147, N_a=705, p_0=0.091, N_0=688, alt="two", conf_level=0.95, correct=F,
manual=F, graph=T)
```

```
Two-Sample Test, Two-Tailed. Chi-square distribution
H0: 0.147-0.091 = 0, Ha: 0.147-0.091 ≠ 0, confidence level = 0.95, alpha/2 = 0.025

test statistic = 3.22331, critical value (0.975) = 1.96, Std. Err. = 0.01334
95% confidence interval: from 0.0222 to 0.0898
p-value = 0.0012672

the null hypothesis H0 is rejected at the 5% significance level
```



Esempi

Esempio 9: test su miglioramento processo (1 campione su popolazione)

1. Raccolta dei dati e obiettivo del test

   Storicamente, una fabbrica ha prodotto un componente nanotecnologico con un'affidabilità del 35%, cioè il 35% dei componenti superava i requisiti di garanzia della qualità. Dopo aver cambiato il processo di produzione, i tecnici sperano di aver migliorato l'affidabilità. Per testarlo, hanno selezionato un campione di 24 componenti prodotti con il nuovo processo e hanno riscontrato che 13 componenti hanno superato il controllo di qualità.

2. Ipotesi

   L'ipotesi nulla è che non ci sia stato un miglioramento significativo nel processo di produzione.

   H_0: 0,5417 (13/24) ≤ 035, H_a: 0,5417 > 0,35   il test è a 1 coda (destra)

3. Livello di significatività

   Con α = 5%, il valore critico per $\mathbf{Z}_{1-\alpha}$ = 1,6449.

4. Test statistici (*risultati dalla pagina seguente*)

   $Z > Z_{1-\alpha}$ (3,60555 > 1,6449), l'IC è compreso tra [0,3788, **1**] e il *p-value* = 0,04225.

   In questo caso, poiché la dimensione del campione è piccola (***n* = 24**) e $\boldsymbol{np_0}$ **< 10**, si è utilizzata la **distribuzione binomiale** per il test, invece della distribuzione normale.

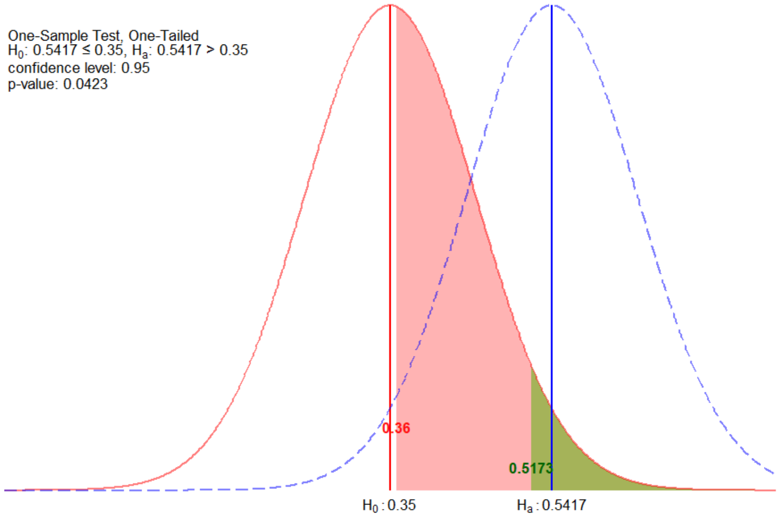
5. Decisione

   Poiché il valore calcolato di Z (3,60555) è maggiore del valore critico $Z_{1-\alpha}$ (1,6449) e il *p-value* (0,04225) è inferiore a 0,05, **si rifiuta l'ipotesi nulla**. Pertanto, **il nuovo processo mostra un miglioramento significativo** rispetto al vecchio.

Esempi

Esempio 9: test su miglioramento processo (1 campione su popolazione)

```
h.test(n_a=13, N_a=24, p_0=0.35, alt="gr", conf_level=0.95, manual=F, graph=T)

For a more precise estimate, the binomial test is recommended.
Use the argument binom = TRUE.
...

h.test(n_a=13, N_a=24, p_0=0.35, alt="gr", binom=TRUE, conf_level=0.95, manual=F, graph=T))

One-Sample Test, One-Tailed. Binomial distribution
H0: 0.5417 ≤ 0.35, Ha: 0.5417 > 0.35, confidence level = 0.95, alpha = 0.05

test statistic = 3.60555, critical value (0.95) = 1.6449, Std. Err. = 0.10171
95% confidence interval: from 0.3576  to 1
p-value = 0.042253

the null hypothesis H0 is rejected at the 5% significance level
```



Esempi

Esempio 10: test sulla differenza di peso prima e dopo una dieta (1 campione)

1. Raccolta dei dati e obiettivo del test

   Un gruppo di persone ha partecipato a uno studio per valutare l'efficacia di una nuova dieta. Per ogni individuo è stato registrato il peso corporeo prima e dopo la dieta. Si vuole verificare se la dieta ha determinato una variazione significativa nel peso medio dei partecipanti. Poiché **le osservazioni prima e dopo la dieta non sono indipendenti** (provengono dagli stessi individui), e quindi i **campioni sono appaiati**, si effettuerà **un test *t* per un solo campione** per verificare se la **media delle differenze** tra le due misurazioni è statisticamente significativa rispetto a zero. Per procedere con il calcolo bisogna determinare prima le differenze tra i pesi (in kg):

Prima dieta	**80**	**95**	**70**	**82**	**71**	**70**	**120**	**105**	**111**	**90**
Dopo dieta	**78**	**94**	**69**	**83**	**65**	**69**	**118**	**103**	**112**	**88**
Differenza	**2**	**1**	**1**	**-1**	**6**	**1**	**2**	**2**	**-1**	**2**

Differenza media = 1,5 kg

2. Ipotesi

   H_0: media delle differenze = 0 , H_a: media delle differenze ≠ 0 il test è a 2 code

3. Livello di significatività $\boldsymbol{\alpha}$ = 5% a cui corrisponde il valore critico $\boldsymbol{t}_{1-\alpha/2,\,n-1}$ = 2,2622

4. Test statistici (vedi risultati pagina seguente)

   $t = \frac{\bar{d}}{s_d}\sqrt{n}$ = 1,5/1,958 √10 = 2,42272; $t > t_{1-\alpha/2,9}$ (2,42272 > 2,2622); 0 è fuori dall'IC al 95% [0,09941, 2,90059] e il *p-value* (0,03844) < 0,05.

5. Decisione

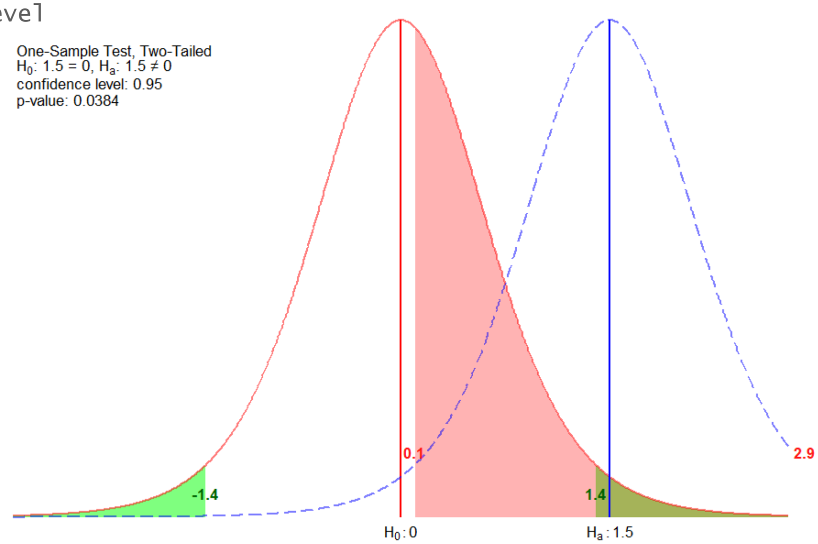
   Poiché l'IC non include zero, si rifiuta l'ipotesi nulla**. Ci sono prove sufficienti per affermare che la dieta abbia causato una riduzione media del peso diversa da zero**. Il test ha mostrato una **riduzione media di 1,5 kg**, che è statisticamente significativa al livello del 5%. Tuttavia, l'importanza pratica di questa riduzione dipende dalle aspettative individuali.

Esempi

Esempio 10: test sulla differenza di peso prima e dopo una dieta (1 campione)

```
prima <- c(80, 95, 70, 82, 71, 70, 120, 105, 111, 90)
dopo <- c(78, 94, 69, 83, 65, 69, 118, 103, 112, 88)
differenze <- prima - dopo
m_differenze <- mean(differenze); sd_differenze=sd(differenze); n_differenze=length(differenze)
h.test(x_a=m_differenze, sd_a=sd_differenze, n_a=n_differenze, x_0=0, alt="two", conf_level=0.95, graph=T)
```

```
One-Sample Test, Two-Tailed. Student's t distribution
H0: 1.5 = 0, Ha: 1.5 ≠ 0, confidence level = 0.95, alpha/2 = 0.025

test statistic = 2.42272, critical value (0.975,9) = 2.2622, Std. Err. = 0.619
95% confidence interval: from 0.0994 to 2.9006
p-value = 0.03844

the null hypothesis H0 is rejected at the 5% significance level
```



Esempi

Esempio 11: valutazione del premier dopo sei mesi (1 campione)

1. Raccolta dei dati e obiettivo del test

   La tabella seguente mostra i risultati di due sondaggi condotti a sei mesi di distanza, in cui gli stessi elettori hanno espresso un giudizio sull'operato del primo ministro inglese.

Secondo sondaggio			
Primo sondaggio	**Approva**	**Disapprova**	Totale
Approva	794	**150**	944
Disapprova	**86**	570	656
Totale	880	720	1600

   Si vuole verificare se c'è stato un **significativo aumento del numero di elettori che hanno cambiato opinione** passando **da un giudizio positivo a negativo**. Su un totale di **236 individui che hanno modificato la propria valutazione**, **150 sono passati da "*approva*" a "*disapprova*"**. Poiché i **campioni sono appaiati**[2] (gli stessi individui sono stati intervistati due volte), si procede con un **test Z per un solo campione** per valutare se questo aumento è statisticamente significativo.

2. Ipotesi

   L'ipotesi nulla è che non c'è stato alcun cambiamento significativo (**proporzione uguale a 50**).
   H_0: 0,6356 (**150/(86+150)**) **= 0,5**, H_a: 0,6356 (**150/(86+150)**) **≠ 0,5** per cui test è a 2 code

3. Livello di significatività con α = 5% il valore critico $Z_{1-\alpha/2}$ = 1,960

4. Test statistici

   $Z = (p_a - p_0) / \sqrt{p_0(1-p_0)/n} = (0{,}6356-0{,}5) / \sqrt{(0{,}5*(1-0{,}5)/236)}$ = 4,16605, l'IC al 95%: $0{,}636 \pm 1{,}96*\sqrt{(0{,}636*0{,}364/236)}$ è compreso tra [0,57, 0,69] con il *p-value* = 0,000030993

5. Decisione

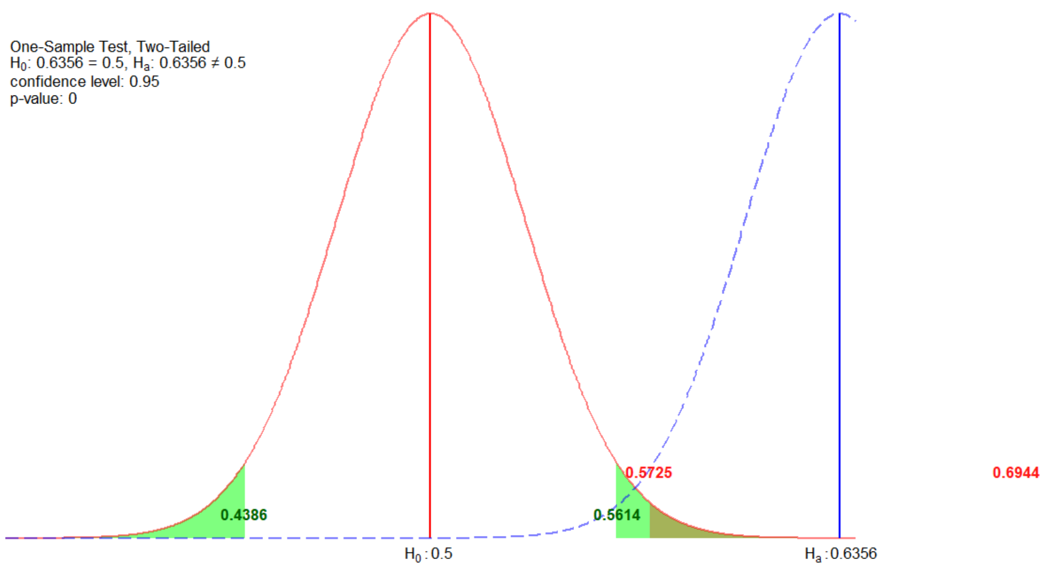
   Dopo sei mesi, circa il 63,6% di chi ha cambiato idea è passato da "*Approva*" a "*Disapprova*". Questo **cambiamento è statisticamente significativo**, indicando **un calo di popolarità del primo ministro** nel periodo analizzato. Tra gli elettori che hanno modificato la loro opinione, tra il 57,2% e il 69,4% ha smesso di approvarlo, mentre solo tra il 30,6% e il 42,8% ha cambiato idea in suo favore.

[1] Alan Agresti, Categorical Data Analysis, II Edition, 2002, John Wiley & Sons, p. 411
[2] In questo caso è più indicato il test di **McNemar**.

Esempi

Esempio 11: valutazione del premier dopo sei mesi (1 campione)

```
h.test (n_a=150, N_a=(86+150), p_0=0.5, alt="two", conf_level=0.95, manual=F, correct=F, graph=T)

For a more precise estimate, the binomial test is recommended.
Use the argument binom = TRUE.

One-Sample Test, Two-Tailed. Chi-square distribution
H0: 0.6356 = 0.5, Ha: 0.6356 ≠ 0.5, confidence level = 0.95, alpha/2 = 0.025

test statistic = 4.16605, critical value (0.975) = 1.96, Std. Err. = 0.03133
95% confidence interval: from 0.5725 to 0.6944
p-value = 3.0993e-05

the null hypothesis H0 is rejected at the 5% significance level
```



Esempi

Esempio 12: test sull'efficacia di un farmaco rispetto al precedente (2 campioni)

1. Raccolta dei dati e obiettivo del test

   Si vuole valutare l'efficacia del nuovo farmaco per l'artrite, *ArtiPlus*, rispetto al trattamento standard, *FlexiMax*.
   Non si vuole dimostrare che *ArtiPlus* sia superiore, ma **verificare che abbia un'efficacia comparabile, risultando al contempo più economico**. Per questo motivo, un classico test di ipotesi non è adatto: **invece di cercare una differenza significativa**, si deve **dimostrare la similarità**, rendendo il **test di non inferiorità** la scelta più appropriata.

   ArtiPlus sarà considerato efficace quanto *FlexiMax* se il suo punteggio medio (margine) non **risulterà inferiore di oltre 2 punti** rispetto alla media attesa di 10 per *FlexiMax*, ossia entro il 20%.

   Alla fine dello studio, l'**obiettivo è poter affermare**, con un livello di confidenza del 95**%** e una **probabilità dell'85%,** che ***ArtiPlus* rappresenta un'alternativa valida**.

2. Dimensione dei campioni

   Per ottenere maggiori informazioni sugli eventi avversi, il **campione del gruppo sperimentale** (*ArtiPlus*) sarà del **50% più grande** rispetto **a quello di controllo** (*FlexiMax*).

   Per garantire una **potenza del test dell'85%,** è necessario calcolare la **dimensione minima del campione per ciascun gruppo**, **tenendo conto del diverso rapporto** tra le due dimensioni.

   Supponendo che *ArtiPlus* abbia una **media di 9,5**, leggermente inferiore alla **media di 10** di *FlexiMax*, e che la **deviazione standard sia pari a 2,25** per entrambi i trattamenti, si ottiene una **numerosità** campionaria **di 28** per il **gruppo di controllo** e **42** per **quello sperimentale**.

   Con questi valori si è proceduto all'esperimento ottenendo:

ArtiPlus	N.: 42	Media, mg/dL: 9,405	Dev. Std.: 2,45
FlexiMax	N.: 28	Media, mg/dL: 10,07	Dev. Std.: 2,79

Esempi

Esempio 12: test sull'efficacia di un farmaco rispetto al precedente (2 campioni)

In questo esempio viene utilizzata la funzione ***ttest.2samp*** per il calcolo delle numerosità campionarie

```
library(powertools)
ttest.2samp ( n.ratio = 1.5,            # rapporto tra i campioni (50% in più per il secondo)
              delta = 9.5-10-(-2),      # differenza media (comprensiva del margine da sottrarre)
              sd1 = 2.25,
              sd.ratio = 1,             # rapporto tra le deviazioni standard
              df.method = "classical",
              alpha = 0.05,
              power = 0.85,             # potenza del test desiderata
              sides = 1,                # una coda
              v = T )
```

```
Two-sample t test power calculation

         n1, n2 = 27.51785, 41.27678
         delta = 1.5
         sd1, sd2 = 2.25, 2.25
         alpha = 0.05
         power = 0.85
         sides = 1
```

Esempi

Esempio 12: test sull'efficacia di un farmaco rispetto al precedente (2 campioni)

3. Ipotesi
   H_0: ArtiPlus **è inferiore** a FlexiMax di almeno 2 unità (**$\mu_1 - \mu_2 \leq -2$**)
   H_a: La differenza è superiore a -2 (**$\mu_1 - \mu_2 > -2$**) il test è a 1 coda (destra)

4. Livello di significatività
   α = 5%; valore critico $t_{1-\alpha,\, n1+n2-2}$ = 1,66757[1]

5. Test statistici
   $t = (\bar{x}_a - \bar{x}_0 - (-2)) \;/\; \sqrt{s_a^2/n_a + s_0^2/n_0}$ = (9,405-10,07-(-2)) / √ (2,45^2/42 + 2,79/28) = 2,058[2],
   con il *p-value* = 0,0022[2]

6. Decisione
   Il risultato, con un *p-value* di 0,022, **supporta l'ipotesi di non inferiorità**. Questo indica che **l'efficacia del farmaco ArtiPlus non è significativamente inferiore a quella di FlexiMax**, ossia il punteggio di sollievo per ArtiPlus è al massimo inferiore di 2 unità rispetto a FlexiMax.

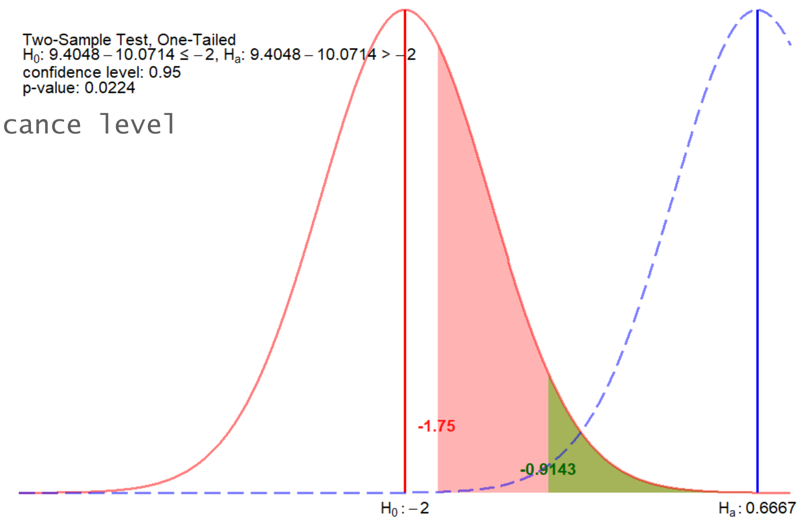
   Possiamo dunque concludere che ***ArtiPlus* rappresenta un'alternativa valida a *FlexiMax***, con un margine di errore del 5%.

[1] Per il valore corretto vedi pagina seguente.
[2] Valore approssimato.

Esempi

Esempio 12: test sull'efficacia di un farmaco rispetto al precedente (2 campioni)

```
FlexiMax <- c(9,14,13,8,10,5,11,9,12,10,9,11,8,11,4,8,11,16,12,10,9,10,13,12,11,13,9,4);
ArtiPlus <- c(7,14,8,4,10,11,7,7,13,8,8,13,10,9,12,9,11,10,12,7,8,5,10,7,13,12,13,11,7,12,10,11,10,8,6,
              9,11,8,5,11,10,8);
  media_prima = mean(FlexiMax ); sd_prima = sd(FlexiMax ); n_prima = length(FlexiMax )
  media_dopo  = mean(ArtiPlus); sd_dopo  = sd(ArtiPlus); n_dopo  = length(ArtiPlus)
  h.test (x_a=media_dopo, n_a=n_dopo, sd_a=sd_dopo, x_0=media_prima, n_0=n_prima, sd_0=sd_prima,
       alt="greater", delta=-2, conf_level=0.95, manual=F, graph=T);

  Two-Sample Test, One-Tailed. Student's t distribution
  H0: 9.4048-10.0714 ≤ -2, Ha: 9.4048-10.0714 > -2, confidence level = 0.95, alpha = 0.05

  test statistic = 2.05612, critical value (0.95,52.75¹) = 1.6743, Std. Err. = 0.648
  95% confidence interval: from -1.7524 to Inf
  p-value = 0.02237

  the null hypothesis H0 is rejected at the 5% significance level
```



[1] I gradi di libertà sono calcolati per difetto con il metodo Satterthwaite (per varianze diverse tra gruppi).

Esempi

Esempio 13: test sull'efficacia di un trattamento oncologico (2 campioni)

1. Raccolta dei dati e obiettivo del test

   In uno studio clinico randomizzato, sono stati inclusi bambini con un certo tipo di cancro al rene per cercare di dimostrare che la **chemioterapia (nuovo trattamento) non è inferiore alla radioterapia (trattamento standard).** Una risposta di successo è definita come una riduzione delle dimensioni del tumore. Il nuovo trattamento si considererà non inferiore al trattamento standard se non differisce di più di un margine di 0,1 (10%). I risultati del trattamento sono stati raccolti da 57 pazienti in ciascuno dei due gruppi.

Trattamento	Soggetti con riduzione
Chemio	53
Radio	51

2. Ipotesi

   H_0: p_a (Chemio) - p_0 (Radio) = -0,1; H_a: p_a (Chemio) - (Radio) $\neq$ -0,1 il test è a 2 code

3. Livello di significatività

   $\boldsymbol{\alpha}$ = 10%; valore critico $Z_{1-\alpha/2,}$= 1,64485

4. Test statistici

   Z = 2,2965; IC = [-0,0567, 0,1305], *p-value* = 0,010824

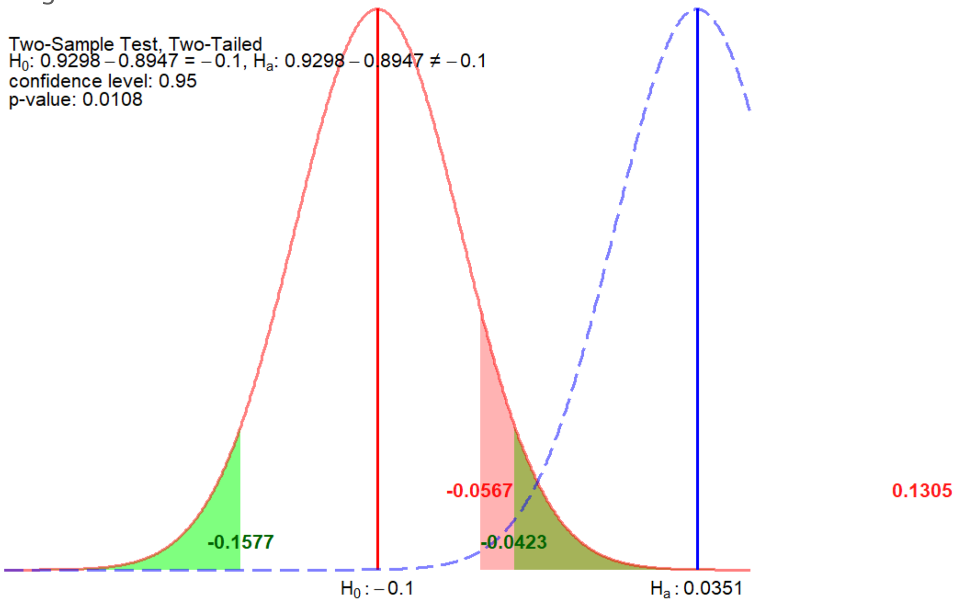
5. Decisione

   I risultati del test di non inferiorità indicano che **il nuovo gruppo di trattamento non è sostanzialmente inferiore al gruppo di controllo** (*p-value*=0,0108). La significatività del test è verificata dal limite inferiore dell'intervallo di confidenza al 90% (-0,0617) che è maggiore del limite di non inferiorità (-0,1).

Esempi

Esempio 13: test sull'efficacia di un trattamento oncologico (2 campioni)

```
h.test (p_a=53/57, N_a=57, p_0=51/57, N_0=57, alt="two", delta=-0.1, conf_level=0.9, graph=T)

Two-Sample Test, One-Tailed. Chi-square distribution
H0: 0.9298-0.8947 = -0.1, Ha: 0.9298-0.8947 ≠ -0.1, confidence level = 0.9, alpha = 0.1

test statistic = 2.29648  , critical value (0.9) = 1.64485  , Std. Err. = 0.03509
90% confidence interval: from -0.0567 to 0.1305
p-value = 0.010824

the null hypothesis H0 is rejected at the 10% significance level
```



Appendice

La funzione h.test

Per i calcoli utilizzati negli esempi l'**autore ha scritto la funzione h.test**[1] utilizzando le funzioni *t.test, prop.test*[2], *binom.test*[3], *farrington.manning*[4] e la funzione *ggplot* del **linguaggio R**[5].

```
source("C:/…/h.test.R")

h.test ( data_a       = NULL,   # data frame contenente i dati del campione sperimentale
         data_0       = NULL,   # data frame contenente i dati del campione di controllo
         x_a          = NULL,   # colonna del d.f. o valore media campione sperimentale
         x_0          = NULL,   # colonna del d.f. o valore media campione di controllo
         sd_a         = NULL,   # colonna del d.f. o dev. standard campione sperimentale
         sd_0         = NULL,   # colonna del d.f. o dev. standard campione di controllo
         p_a          = NULL,   # proporzione campione sperimentale
         p_0          = NULL,   # proporzione campione di controllo
         n_a          = NULL,   # colonna del d.f. o n. casi positivi campione sperimentale
         n_0          = NULL,   # colonna del d.f. o n. casi positivi campione di controllo
         N_a          = NULL,   # colonna del d.f. o n. casi campione sperimentale
         N_0          = NULL,   # colonna del d.f. o n. casi campione di controllo
         conf_level   = 0.95,   # livello di confidenza
         alt          = c("less","greater,"two.tailed"), # ipotesi alternativa
         delta        = 0,      # margine per il test di non-inferiorità4
         manual       = TRUE,   # calcolo manuale della statistica di test
         correct      = TRUE,   # Se TRUE apporta, se serve, la correzione di Yates
         binom        = FALSE,  # se TRUE utilizza, quando è il caso, il test binomiale
         graph        = FALSE)  # Se TRUE crea i grafici delle distribuzioni
```

[1] Contattare l'autore per maggiori dettagli e per richiedere il codice.
[2] L'Intervallo di Confidenza è calcolato con il metodo di Wilson, che può differire dalle formule classiche.
[3] La funzione *binom.test* è utilizzata quando $np < 10$ o $n(1-p) < 10$ e l'argomento `binom=TRUE`.
[5] Vedi Appendice

Appendice

Riferimenti

Bibliografia
Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. 3rd ed. Amsterdam; Boston, Academic Press.

Introduzione al software R
https://www.alfredoroccato.it/corso-di-introduzione-al-software-r/

E questa sperienza si faccia più volte, acciò che qualche accidente non impedissi o falsassi tal prova, che le sperienzia fussi falsa, e ch'ella ingannassi o no il suo speculatore.

Leonardo Da Vinci

There are people of courageous ideas, though very critical of their own ideas, they try to find out if their ideas are correct, trying to find out first whether they are wrong.
They operate with courageous conjectures and severe attempts to reject their own conjectures.

Karl Popper

Contatti

studio roccato
Data Science Training & Consulting

e-mail: alfredo.roccato(at)fastwebnet.it
www.alfredoroccato.it