UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Dipartimento di scienze economiche,
aziendali, matematiche e statistiche
“Bruno de Finetti”

Statistica

Inferenza: verifica d’ipotesi e stima intervallare

Francesco Pauli

A.A. 2016/2017

Indice

Logica della verifica d'ipotesi: facciamo un gioco

Partecipiamo a un gioco in cui si vince se esce testa al lancio di una moneta.

Lancio	Esito	
1	Croce	

Ho perso, pazienza, continuo

Logica della verifica d'ipotesi: facciamo un gioco

Partecipiamo a un gioco in cui si vince se esce testa al lancio di una moneta.

Lancio	Esito	
1	Croce	
2	Croce	

Ho perso, pazienza, continuo

Ho perso di nuovo, continuo

Logica della verifica d'ipotesi: facciamo un gioco

Partecipiamo a un gioco in cui si vince se esce testa al lancio di una moneta.

Lancio	Esito	
1	Croce	
2	Croce	
3	Croce	

Ho perso, pazienza, continuo
Ho perso di nuovo, continuo
E tre, prima o poi la fortuna gira

Logica della verifica d'ipotesi: facciamo un gioco

Partecipiamo a un gioco in cui si vince se esce testa al lancio di una moneta.

Lancio	Esito	
1	Croce	
2	Croce	
3	Croce	
4	Croce	

Ho perso, pazienza, continuo
Ho perso di nuovo, continuo
E tre, prima o poi la fortuna gira
Insisto

Logica della verifica d'ipotesi: facciamo un gioco

Partecipiamo a un gioco in cui si vince se esce testa al lancio di una moneta.

Lancio	Esito	
1	Croce	
2	Croce	
3	Croce	
4	Croce	
5	Croce	

Ho perso, pazienza, continuo
Ho perso di nuovo, continuo
E tre, prima o poi la fortuna gira
Insisto
Insisto

Logica della verifica d'ipotesi: facciamo un gioco

Partecipiamo a un gioco in cui si vince se esce testa al lancio di una moneta.

Lancio	Esito		
1	Croce		Ho perso, pazienza, continuo
2	Croce		Ho perso di nuovo, continuo
3	Croce		E tre, prima o poi la fortuna gira
4	Croce		Insisto
5	Croce		Insisto
6	Croce		
7	Croce		
8	Croce		
10	Croce		Mah!?!

Logica della verifica d'ipotesi: facciamo un gioco

Partecipiamo a un gioco in cui si vince se esce testa al lancio di una moneta.

Lancio	Esito	
1	Croce	Ho perso, pazienza, continuo
2	Croce	Ho perso di nuovo, continuo
3	Croce	E tre, prima o poi la fortuna gira
4	Croce	Insisto
5	Croce	Insisto
6	Croce	
7	Croce	
8	Croce	
10	Croce	Mah!?!
11	Croce	
12	Croce	
13	Croce	
14	Croce	

Logica della verifica d'ipotesi: facciamo un gioco

Partecipiamo a un gioco in cui si vince se esce testa al lancio di una moneta.

Lancio	Esito		
1	Croce		Ho perso, pazienza, continuo
2	Croce		Ho perso di nuovo, continuo
3	Croce		E tre, prima o poi la fortuna gira
4	Croce		Insisto
5	Croce		Insisto
6	Croce		
7	Croce		
8	Croce		
10	Croce		Mah!?!
11	Croce		
12	Croce		
13	Croce		
14	Croce		
15	Croce		Continuo?

Logica della verifica d'ipotesi: facciamo un gioco

Partecipiamo a un gioco in cui si vince se esce testa al lancio di una moneta.

Lancio	Esito		
1	Croce		Ho perso, pazienza, continuo
2	Croce		Ho perso di nuovo, continuo
3	Croce		E tre, prima o poi la fortuna gira
4	Croce		Insisto
5	Croce		Insisto
6	Croce		
7	Croce		
8	Croce		
10	Croce		Mah!?!
11	Croce		
12	Croce		
13	Croce		
14	Croce		
15	Croce		Continuo?

Logica della verifica d'ipotesi: facciamo un gioco

Partecipiamo a un gioco in cui si vince se esce testa al lancio di una moneta.

Lancio	Esito	Probabilità	
1	Croce	0.5	Ho perso, pazienza, continuo
2	Croce	0.25	Ho perso di nuovo, continuo
3	Croce	0.125	E tre, prima o poi la fortuna gira
4	Croce	0.062	Insisto
5	Croce	0.031	Insisto
6	Croce	...	
7	Croce	...	
8	Croce	...	
10	Croce	0.001	Mah!?!
11	Croce	...	
12	Croce	...	
13	Croce	...	
14	Croce	...	
15	Croce	3.0517578×10^{-5}	Continuo?

Logica della verifica d'ipotesi: facciamo un gioco

Partecipiamo a un gioco in cui si vince se esce testa al lancio di una moneta.

Lancio	Esito	Probabilità
1	Croce	0.5
2	Croce	0.25
3	Croce	0.125
4	Croce	0.062
5	Croce	0.031
6	Croce	...
7	Croce	...
8	Croce	...
10	Croce	0.001
11	Croce	...
12	Croce	...
13	Croce	...
14	Croce	...
15	Croce	3.0517578×10^{-5}

Ho perso, pazienza, continuo

Ho perso di nuovo, continuo

E tre, prima o poi la fortuna gira

Insisto

Insisto

Man mano che si va avanti, la faccenda comincia a puzzare: quello che accade è un po' troppo improbabile se la moneta è equilibrata.

A un certo punto non credo più che la moneta sia equilibrata.

Logica della verifica d'ipotesi: facciamo un gioco

Partecipiamo a un gioco in cui si vince se esce testa al lancio di una moneta.

Lancio	Esito	Probabilità
1	Croce	0.5
2	Croce	0.25
3	Croce	0.125
4	Croce	0.062
5	Croce	0.031
6	Croce	...
7	Croce	...
8	Croce	...
10	Croce	0.001
11	Croce	...
12	Croce	...
13	Croce	...
14	Croce	...
15	Croce	3.0517578×10^{-5}

Inizialmente partecipo al gioco perché, ritenendo che la moneta sia equilibrata, il gioco è equo.

Ai primi giri, non ho ragione di pensare altrimenti, perché l'esito a me sfavorevole non è strano.

Sono partito con un'ipotesi 'la moneta è equa', osservando un po' di esperimenti ad un certo punto rifiuto quell'ipotesi.

Ipotesi nulla

Abbiamo una popolazione alla quale è associato un parametro (cioè, una caratteristica della distribuzione è data da un parametro):

$$\theta : \text{parametro}; \quad (\text{probabilità, media})$$

Abbiamo un'ipotesi sul valore del parametro, che si indica con H_0 ed è detta **ipotesi nulla**, questa può essere di diversi tipi

- $H_0 : \theta = \theta_0$ (sopra era $\theta = 0.5$), ipotesi semplice;
- $H_0 : \theta \geq \theta_0$ ipotesi composta;
- $H_0 : \theta \leq \theta_0$ ipotesi composta;

Osserviamo un campione e il nostro obiettivo è valutare se l'ipotesi nulla (H_0) è compatibile o no col campione.
In altre parole, l'osservazione campionaria potrebbe smentire la mia ipotesi.

Esempio: probabilità di nascita di un maschio

Formuliamo l'ipotesi H_0: *nelle popolazioni umane, la probabilità che nasca un maschio è 1/2*.

Come campione, disponiamo dei dati sui nati a Muggia in un anno: 38 maschi e 47 femmine.

La proporzione di maschi nel campione, $\hat{\pi}$ ha, approssimativamente, distribuzione normale con media

$$\pi = \text{Probabilità che nasca un maschio}$$

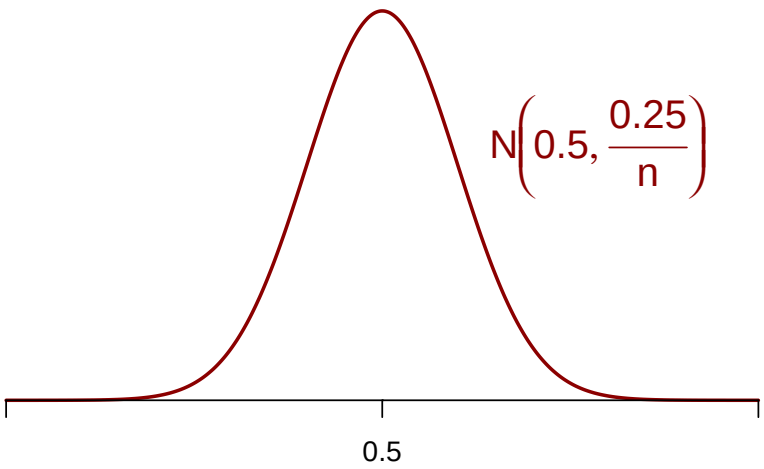
e varianza $\pi(1-\pi)/n$.

Se l'ipotesi nulla è vera,

$$\hat{\pi} \sim \mathcal{N}(0.5, 0.25/n)$$

Confrontiamo questa distribuzione nel'ipotesi nulla con quanto osservato.

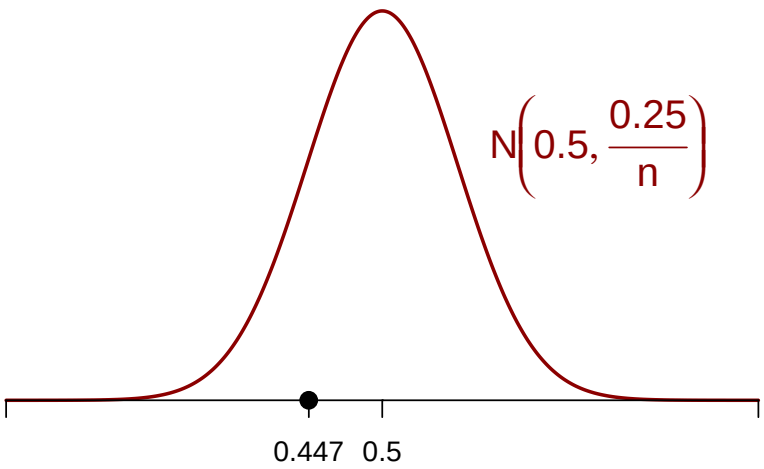
Probabilità di un maschio



Se l'ipotesi nulla è vera, la proporzione campionaria su n unità è distribuita secondo una

$$\mathcal{N}\left(0.5, \frac{0.25}{n}\right)$$

Probabilità di un maschio



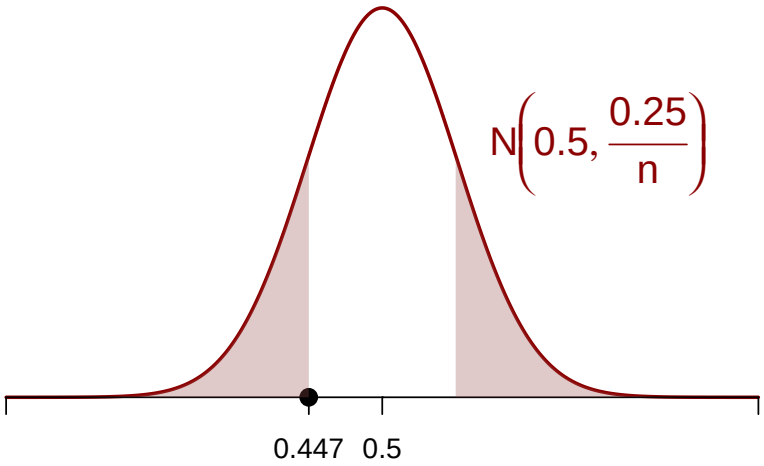
Se l'ipotesi nulla è vera, la proporzione campionaria su n unità è distribuita secondo una

$$\mathcal{N}\left(0.5, \frac{0.25}{n}\right)$$

Si è osservato

$$\hat{\pi} = \frac{38}{85} \approx 0.447$$

Probabilità di un maschio



Se l'ipotesi nulla è vera, la proporzione campionaria su n unità è distribuita secondo una

$$\mathcal{N}\left(0.5, \frac{0.25}{n}\right)$$

Si è osservato

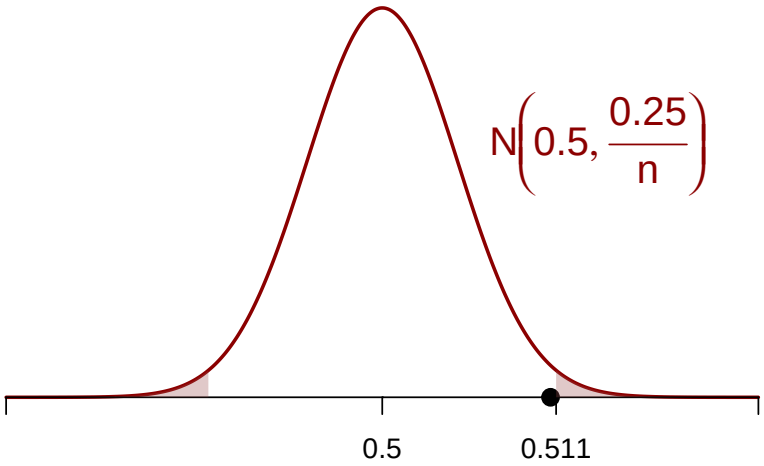
$$\hat{\pi} = \frac{38}{85} \approx 0.447$$

Misuriamo quanto questo è compatibile con l'ipotesi sulla base della probabilità di osservare uno scostamento altrettanto grande da 0.5 se l'ipotesi è vera

$$P(|\hat{\pi} - 0.5| \geq |0.5 - 0.447|) = P\left(\left|\frac{\hat{\pi} - 0.5}{0.5/\sqrt{n}}\right| \geq \frac{0.0529}{0.5/\sqrt{n}}\right) = 2\left(1 - \Phi\left(0.976\right)\right) = 0.329$$

Dato che la probabilità in questione non è per niente piccola, valutiamo che il campione è compatibile con l'ipotesi fatta. In altre parole, l'osservazione fatta non permette di escludere la validità dell'ipotesi.

Probabilità di un maschio, regione FVG



Consideriamo un campione più ampio, nella regione Friuli Venezia Giulia ci sono state 10337 nascite, di cui 5286 maschi, sicché

$$\hat{\pi} = \frac{5286}{10337} \approx 0.511$$

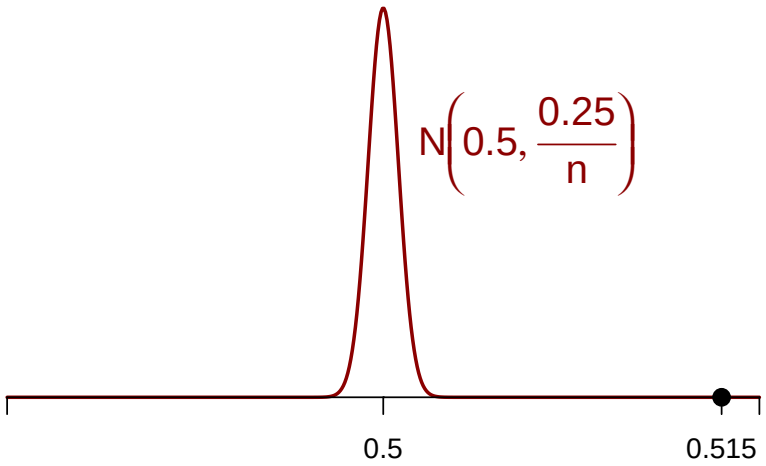
Procedendo come sopra calcoliamo

$$P(|\hat{\pi} - 0.5| \geq |0.5 - 0.511|) = P\left(\left|\frac{\hat{\pi} - 0.5}{0.5/\sqrt{n}}\right| \geq \frac{0.0114}{0.5/\sqrt{n}}\right) = 2\left(1 - \Phi\left(2.31\right)\right) = 0.02081$$

La probabilità è ora relativamente piccola, tale da mettere in dubbio l'ipotesi.

Potremmo decidere di raccogliere un campione più grande.

Probabilità di un maschio, Italia



Consideriamo un campione ancora più ampio, in Italia ci sono state 561944 nascite, di cui 289185 maschi, sicché

$$\hat{\pi} = \frac{289185}{561944} \approx 0.515$$

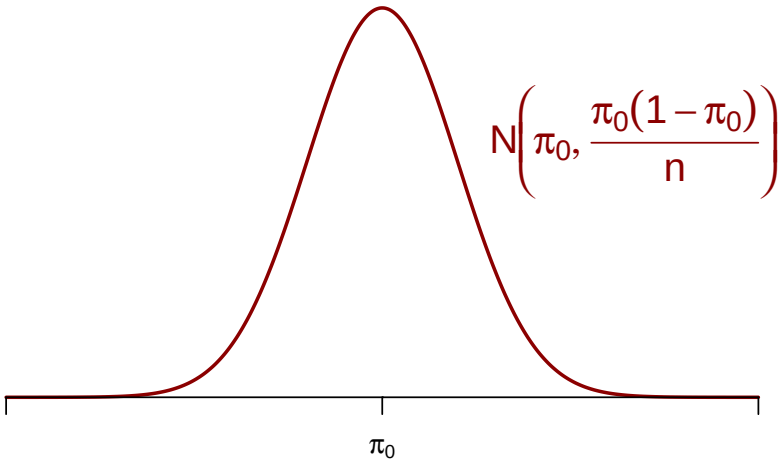
Procedendo come sopra calcoliamo

$$P(|\hat{\pi}-0.5| \geq |0.5-0.515|) = P\left(\left|\frac{\hat{\pi}-0.5}{0.5/\sqrt{n}}\right| \geq \frac{0.0146}{0.5/\sqrt{n}}\right) = 2\left(1-\Phi\left(21.9\right)\right) = 1.9889688\times 10^{-106}$$

La probabilità è ora estremamente bassa, l'ipotesi è ragionevolmente smentita.

Per approfondire sul rapporto maschi/femmine nelle popolazioni umane si veda, ad esempio, wikipedia.

Test di significatività bilaterale per una proporzione



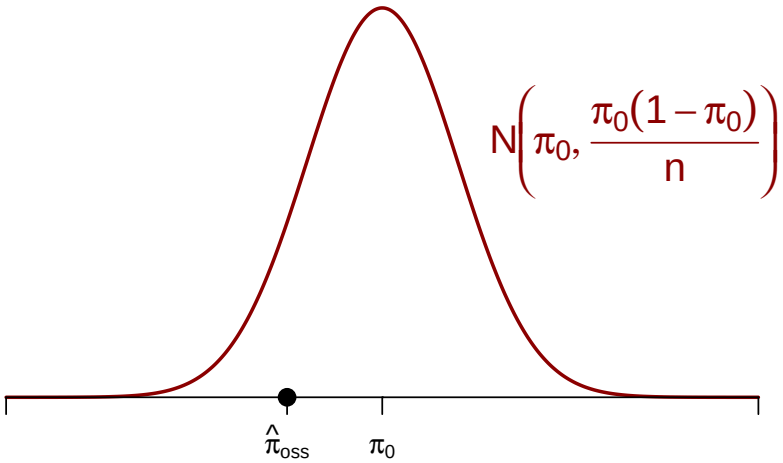
In generale, se l'ipotesi nulla è

$$H_0 : \pi = \pi_0$$

Se l'ipotesi nulla è vera, la proporzione campionaria su n unità è distribuita approssimativamente secondo una normale

$$D_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \mathcal{N}(0,1)$$

Test di significatività bilaterale per una proporzione



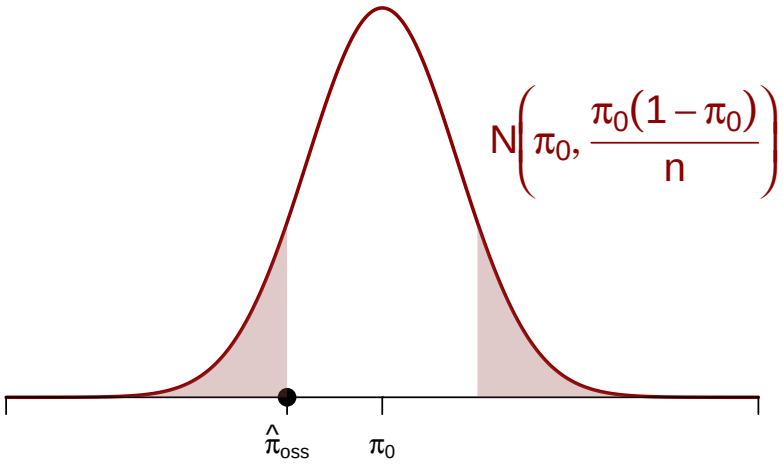
In generale, se l'ipotesi nulla è

$$H_0 : \pi = \pi_0$$

Se l'ipotesi nulla è vera, la proporzione campionaria su n unità è distribuita approssimativamente secondo una normale

$$D_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \mathcal{N}(0,1)$$

Test di significatività bilaterale per una proporzione



In generale, se l'ipotesi nulla è

$$H_0 : \pi = \pi_0$$

Se l'ipotesi nulla è vera, la proporzione campionaria su n unità è distribuita approssimativamente secondo una normale

$$D_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \mathcal{N}(0,1)$$

Si osserva $\hat{\pi} = \hat{\pi}_{oss}$: misuriamo quanto questo è compatibile con l'ipotesi sulla base della probabilità di osservare uno scostamento altrettanto grande da π_0 se l'ipotesi è vera, detto

valore p

$$P(|D_0| > |d_0|) = P\left(\left|\frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right| > \left|\frac{\hat{\pi}_{oss} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right|\right) = 2\left(1 - \Phi\left(\left|\frac{\hat{\pi}_{oss} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right|\right)\right)$$

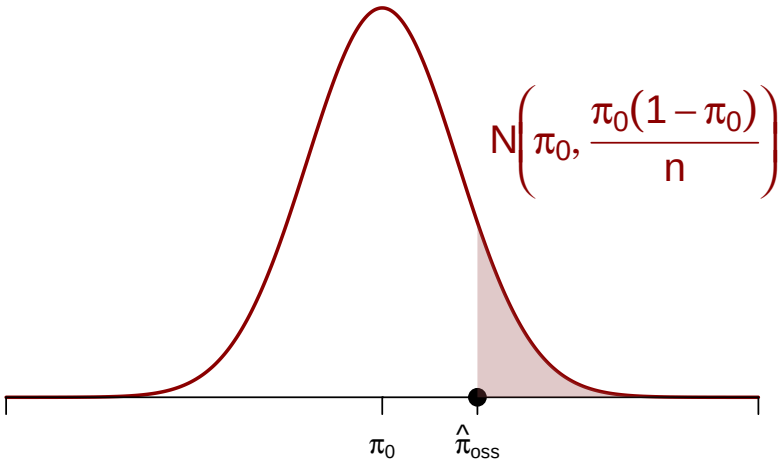
Se la probabilità in questione è molto piccola, valutiamo che il campione non è compatibile con l'ipotesi fatta.

Generalizzazioni

Dobbiamo generalizzare quanto sopra al caso di

- Altri tipi d'ipotesi:
  - bilaterale $H_0 : \pi = \pi_0$
  - unilaterale $H_0 : \pi \leq \pi_0$
  - unilaterale $H_0 : \pi \geq \pi_0$
- Altri modelli
  - la proporzione
  - la media di una normale, con varianza nota
  - la media di una normale, con varianza non nota
- Poi generalizzeremo anche al caso di due popolazioni

Test di significatività unilaterale per una proporzione



Se l'ipotesi nulla è

$$H_0 : \pi \leq \pi_0$$

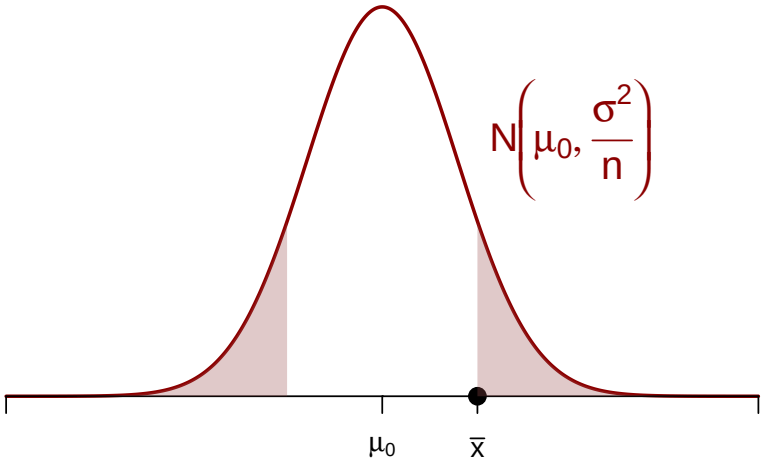
e si osserva $\hat{\pi} = \hat{\pi}_{oss}$, questo è suscettibile di smentire l'ipotesi nulla se è **maggiore** di π_0.

Misuriamo quanto questo è compatibile con l'ipotesi sulla base della probabilità di osservare uno scostamento altrettanto grande da π_0 se l'ipotesi è vera

$$P(D_0 > d_0) = P\left(\frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} > \frac{\hat{\pi}_{oss} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right) = 1 - \Phi\left(\frac{\hat{\pi}_{oss} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right)$$

Come si imposta il problema se $H_0 : \pi \geq \pi_0$?

Test di significatività per la normale con varianza nota



Se l'ipotesi nulla è

$$H_0 : \mu = \mu_0$$

la distribuzione della media campionaria nell'ipotesi nulla è una normale e si ha

$$D_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

La misura dello scostamento tra ipotesi nulla e osservazione $\bar{x}$ è

$$P(|D_0| > |d_0|) = P\left(\left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| > \left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right|\right) = 2\left(1 - \Phi\left(\left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right|\right)\right)$$

E se $H_0 : \mu \leq \mu_0$ o $H_0 : \mu \geq \mu_0$?

Test di significatività per la normale con varianza non nota

Se l'ipotesi nulla è

$$H_0 : \mu = \mu_0$$

la quantità di riferimento è

$$D_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}.$$

Se indichiamo con $\bar{x}$ il valore osservato, la misura dello scostamento tra ipotesi nulla e osservazione è

$$P(|D_0| > |d_0|) = P\left(\left|\frac{\bar{X} - \mu_0}{S/\sqrt{n}}\right| > \left|\frac{\bar{x} - \mu_0}{S/\sqrt{n}}\right|\right)$$

e si calcola con riferimento alla funzione di ripartizione della t di Student con $n - 1$ gradi di libertà.

Test di significatività: riassunto

La procedura dei test di significatività, in breve, è

a. formulare un'ipotesi
b. osservare un campione
c. misurare lo scostamento tra campione e ipotesi con la probabilità di osservare un campione altrettanto distante dall'ipotesi quando questa è vera.

La misura in questione, detta anche **valore p** (p-value) è stata anche chiamata fattore sorpresa, risponde alla domanda su quanto è sorprendente il campione rispetto all'ipotesi nulla.

Valore p: quadro

	Sistema d'ipotesi		
	$H_0 : \mu = \mu_0$	$H_0 : \mu \leq \mu_0$	$H_0 : \mu \geq \mu_0$
Normale, σ noto	$2\left(1 - \Phi\left(\left\lvert\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right\rvert\right)\right)$	$1 - \Phi\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)$	$\Phi\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)$
Normale, σ non noto	$2\left(1 - F_{t_{n-1}}\left(\left\lvert\frac{\bar{x}-\mu_0}{S/\sqrt{n}}\right\rvert\right)\right)$	$1 - F_{t_{n-1}}\left(\frac{\bar{x}-\mu_0}{S/\sqrt{n}}\right)$	$F_{t_{n-1}}\left(\frac{\bar{x}-\mu_0}{S/\sqrt{n}}\right)$
	$H_0 : \pi = \pi_0$	$H_0 : \pi \leq \pi_0$	$H_0 : \pi \geq \pi_0$
Proporzione	$2\left(1 - \Phi\left(\left\lvert\frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right\rvert\right)\right)$	$1 - \Phi\left(\frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right)$	$\Phi\left(\frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\right)$

Dove $F_{t_{n-1}}$ indica la funzione di ripartizione della t_{n-1}.

Valore p: quadro

			Sistema d'ipotesi		
		d	$H_0 : \mu = \mu_0$	$H_0 : \mu \leq \mu_0$	$H_0 : \mu \geq \mu_0$
Normale	σ noto	$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	$2\left(1-\Phi\left(\lvert d\rvert\right)\right)$	$1-\Phi\left(d\right)$	$\Phi\left(d\right)$
	σ non noto	$\frac{\bar{x}-\mu_0}{S/\sqrt{n}}$	$2\left(1-F_{t_{n-1}}\left(\lvert d\rvert\right)\right)$	$1-F_{t_{n-1}}\left(d\right)$	$F_{t_{n-1}}\left(d\right)$
			$H_0 : \pi = \pi_0$	$H_0 : \pi \leq \pi_0$	$H_0 : \pi \geq \pi_0$
Proporzione		$\frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	$2\left(1-\Phi\left(\lvert d\rvert\right)\right)$	$1-\Phi\left(d\right)$	$\Phi\left(d\right)$

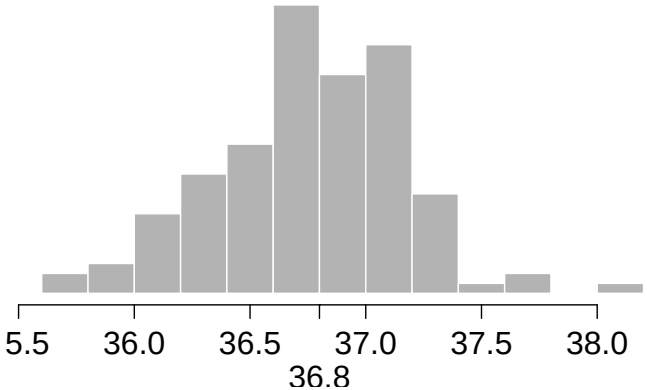
Dove $F_{t_{n-1}}$ indica la funzione di ripartizione della t_{n-1}.

Esempio: quando si ha la febbre?

Chi decide qual è la temperatura corporea 'normale'?

- Carl Wunderlich nel XIX secolo, basandosi su un gran numero di misure, ottenne una temperatura normale di $37.0°C$
- (Era il valore medio di milioni di misure.)
- Nel 1992 viene fatto un secondo studio, volto a controllare quello di Wunderlich del quale si dubita per via
  - della precisione dei termometri dell'epoca;
  - delle modalità di misurazione.
- (Shoemaker & College (1996) *What's Normal? – Temperature, Gender, and Heart Rate*, Journal of Statistics Education v.4, n.2)
- (Mackowiak, Wasserman, Levine (1992) *A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich*, Journal of the American Medical Association, 268, 1578-1580.)

Temperatura corporea

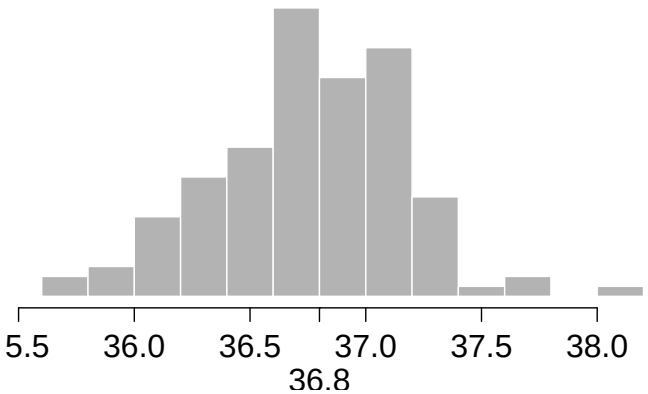


$$n = 130$$

$$\bar{x} = 36.8$$

$$S^2 = 0.166$$

Temperatura corporea



$$n = 130$$

$$\bar{x} = 36.8$$

$$S^2 = 0.166$$

L'ipotesi nulla è

$$H_0 : \mu = 37$$

la varianza è stimata, si calcola la quantità

$$D_0 = \frac{\bar{x} - 37}{S/\sqrt{n}} = \frac{36.8 - 37}{0.407/\sqrt{130}} = \frac{36.8 - 37}{0.036} = -5.556$$

Il riferimento è alla t di Student con 129 gradi di libertà, che di fatto è quasi uguale alla normale, quindi calcoliamo il valore p usando la FdR della normale standard

$$2\left(1 - \Phi\left(\left|\frac{\bar{x} - 37}{S/\sqrt{n}}\right|\right)\right) = 2(1 - \Phi(5.556)) \approx 0$$

Indice

Sistemi di ipotesi

L'approccio di Neyman-Pearson fa riferimento a un sistema d'ipotesi

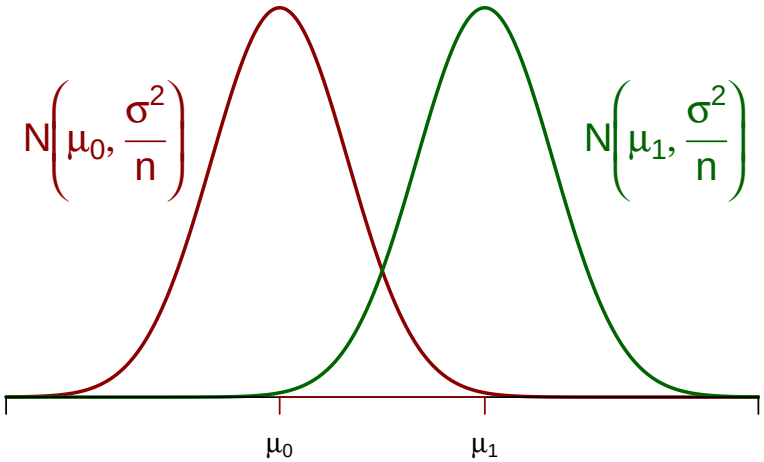
- ipotesi nulla, H_0
- ipotesi alternativa, H_1

e si vuole, idealmente prima di osservare il campione, formulare una regola per cui, a seconda del valore assunto dal campione si accetta o rifiuta l'ipotesi nulla, cioè si definisce un insieme di valori campionari $\mathcal{R}$, detta **regione di rifiuto** per cui

- se $X \in \mathcal{R}$ si rifiuta l'ipotesi nulla;
- se $X \notin \mathcal{R}$ si accetta l'ipotesi nulla;

il complementare della regione di rifiuto è anche detto regione di accettazione.

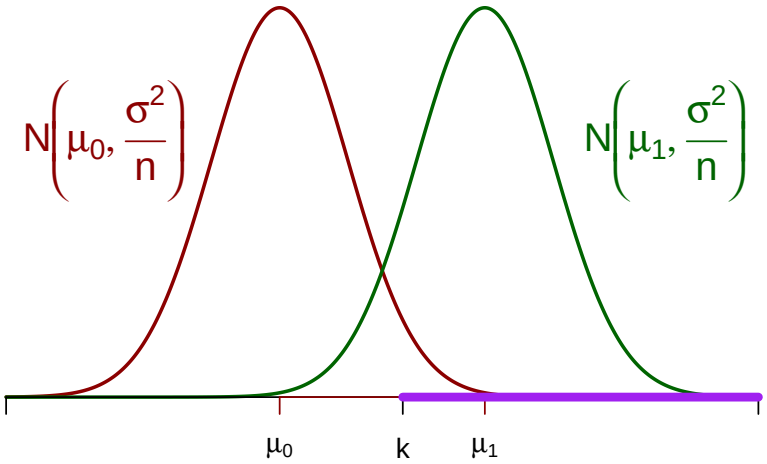
Regioni di rifiuto, media della normale con varianza nota



Consideriamo il sistema di ipotesi

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 (> \mu_0) \end{cases}$$

Regioni di rifiuto, media della normale con varianza nota



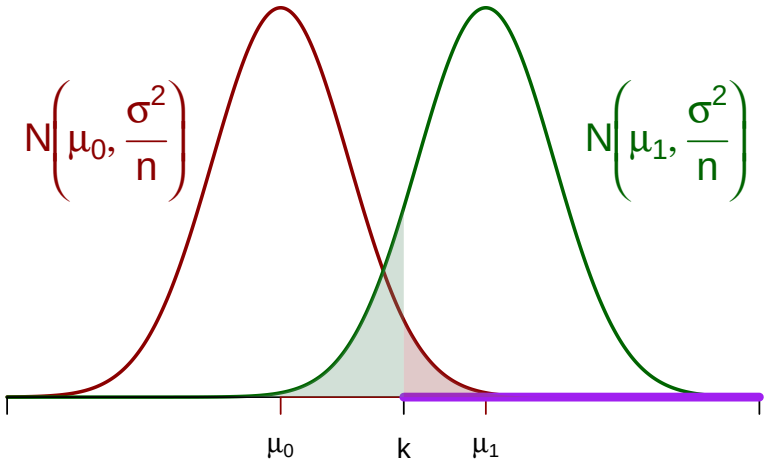
Consideriamo il sistema di ipotesi

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 (> \mu_0) \end{cases}$$

e consideriamo la regione di rifiuto (in viola in figura)

$$\mathcal{R} = \{\bar{X} > k\}$$

Regioni di rifiuto, media della normale con varianza nota



Consideriamo il sistema di ipotesi

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 (> \mu_0) \end{cases}$$

e consideriamo la regione di rifiuto (in viola in figura)

$$\mathcal{R} = \{\bar{X} > k\}$$

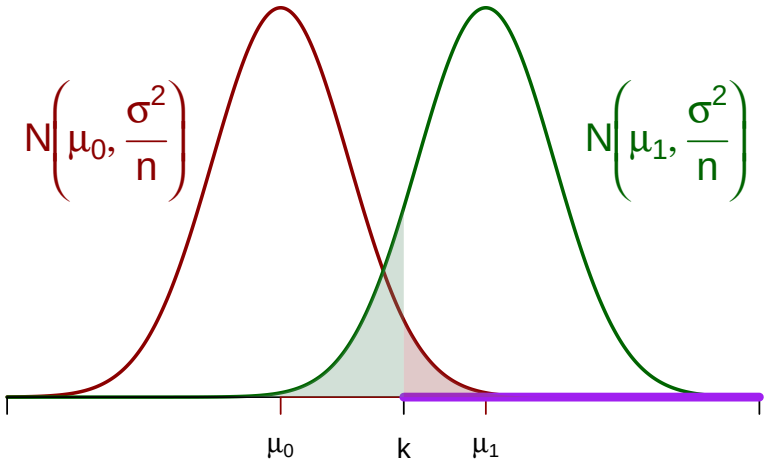
Per scegliere k facciamo riferimento alle probabilità di errore

		Stato reale	
		H_0 vera	H_0 falsa
Decisione	Non rifiuto: $X \notin \mathcal{R}$	Corretto	Errore II tipo
	Rifiuto: $X \in \mathcal{R}$	Errore I tipo	Corretto

In figura si rappresenta

- in rosso $P(\bar{X} \in \mathcal{R}; H_0) = P(\bar{X} > k; H_0) = P\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{k-\mu_0}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{k-\mu_0}{\sigma/\sqrt{n}}\right)$
- in verde $P(\bar{X} \notin \mathcal{R}; H_1) P(\bar{X} \leq k; H_1) = P\left(\frac{\bar{X}-\mu_1}{\sigma/\sqrt{n}} \leq \frac{k-\mu_1}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{k-\mu_1}{\sigma/\sqrt{n}}\right)$

Regioni di rifiuto, media della normale con varianza nota



Consideriamo il sistema di ipotesi

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 (> \mu_0) \end{cases}$$

e consideriamo la regione di rifiuto (in viola in figura)

$$\mathcal{R} = \{\bar{X} > k\}$$

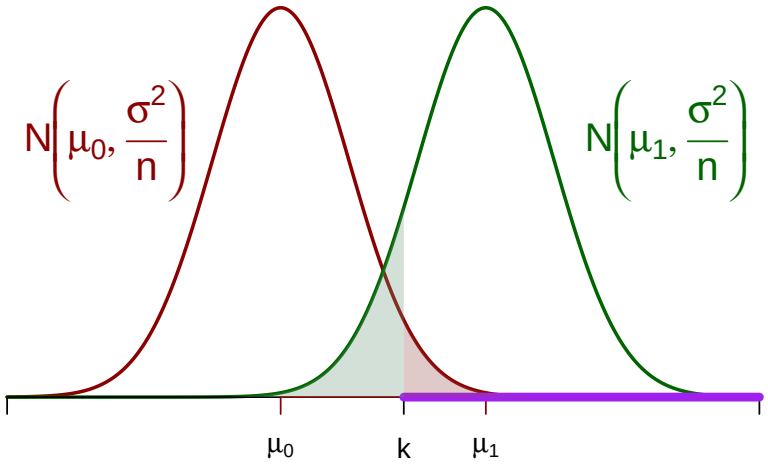
Per scegliere k facciamo riferimento alle probabilità di errore

		Stato reale	
		H_0 vera	H_0 falsa
Decisione	Non rifiuto: $X \notin \mathcal{R}$	Corretto	$\Phi\left(\frac{k-\mu_1}{\sigma/\sqrt{n}}\right)$
	Rifiuto: $X \in \mathcal{R}$	$1 - \Phi\left(\frac{k-\mu_0}{\sigma/\sqrt{n}}\right)$	Corretto

Com'è evidente, al crescere di k

- diminuisce la probabilità di errore di I tipo
- aumenta la probabilità di errore di II tipo

Regioni di rifiuto, media della normale con varianza nota



Consideriamo il sistema di ipotesi

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 (> \mu_0) \end{cases}$$

e consideriamo la regione di rifiuto (in viola in figura)

$$\mathcal{R} = \{\bar{X} > k\}$$

Per determinare un k, si fissa la probabilità di errore di I tipo

$$\alpha = 1 - \Phi\left(\frac{k - \mu_0}{\sigma/\sqrt{n}}\right)$$

da cui

$$k = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

La regione di rifiuto con probabilità di errore di I tipo al più pari a α è detta regione di livello α.

Il complemento a 1 della probabilità di errore di II tipo è detto potenza

$$1 - P(X \in \mathcal{R}; H_1) = 1 - \Phi\left(\frac{k - \mu_1}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{1-\alpha}\right)$$

Verifica d'ipotesi, sintesi

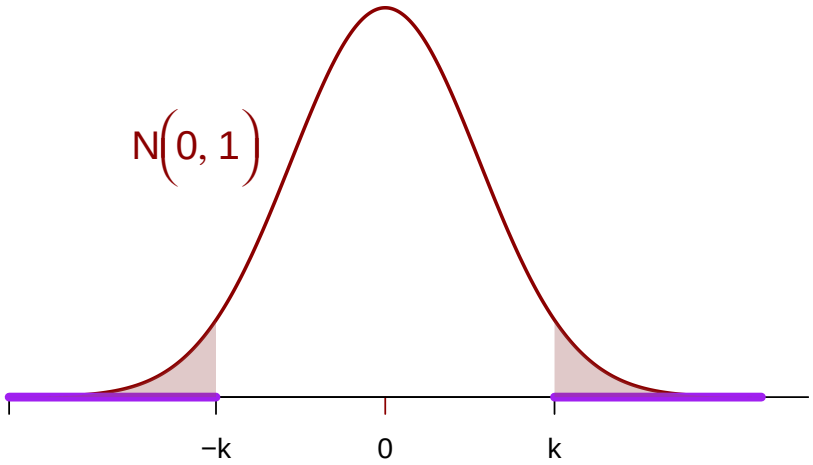
Per verificare un sistema d'ipotesi col paradigma di Neyman-Pearson

(1) Si formulano le due ipotesi.

(2) Si determina la forma della regione di rifiuto.
   - Nei casi che ci interessano sarà naturale.

(3) Si sceglie una particolare regione di rifiuto in modo che la probabilità di errore di I tipo sia α.
   - La probabilità di errore di II tipo viene di conseguenza, si fa in sostanza la scelta di tenere sotto controllo l'errore di I tipo (l'idea è che esso sia "più grave" del II tipo).
   - Si noti che per ottenere la regine di rifiuto si fa riferimento alla distribuzione campionaria nel'ipotesi nulla (l'altra non è rilevante).

Regioni di rifiuto: caso bilaterale



Consideriamo il sistema di ipotesi bilaterale

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

conviene qui fare riferimento alla quantità

$$D_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

che nell'ipotesi H_0 è distribuita come una $N(0,1)$, consideriamo la regione di rifiuto (in viola in figura)

$$\mathcal{R} = \{|D_0| > k\} = \left\{ \left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| > k > 0 \right\}$$

La probabilità di errore di I tipo è $2(1 - \Phi(k))$, si ha

$$\alpha = 2(1 - \Phi(k)) \Leftrightarrow k = z_{1-\alpha/2}$$

Altri sistemi di ipotesi

Per il modello normale con varianza nota, si consideri il sistema di'ipotesi

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0$$

e si mostri che

$$\mathcal{R} = \{D_0 > z_{1-\alpha}\} = \left\{\bar{X} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right\}$$

è una regione di rifiuto di livello α.

Per il modello normale con varianza nota, si consideri il sistema di'ipotesi

$$H_0 : \mu \geq \mu_0, \quad H_1 : \mu < \mu_0$$

e si mostri che

$$\mathcal{R} = \{D_0 < -z_{1-\alpha}\} = \left\{\bar{X} < \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right\}$$

è una regione di rifiuto di livello α.

Regioni di rifiuto: proporzione

Nel caso di una proporzione, si usa il fatto già ricordato, che se $\pi = \pi_0$ allora è approssimativamente vero che

$$D_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim \mathcal{N}(0,1)$$

da cui si ricavano le regioni di rifiuto:

Sistema d'ipotesi		
$H_0 : \pi = \pi_0$ $H_1 : \pi \neq \pi_0$	$H_0 : \pi \leq \pi_0$ $H_1 : \pi > \pi_0$	$H_0 : \pi \geq \pi_0$ $H_1 : \pi < \pi_0$
$\left\lvert \frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \right\rvert > z_{1-\alpha/2}$	$\frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} > z_{1-\alpha}$	$\frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} < -z_{1-\alpha}$

Regioni di rifiuto: media con varianza non nota

Nel caso di una verifica su una media con varianza non nota, si usa il fatto già ricordato, che se $\mu = \mu_0$ allora è approssimativamente vero che

$$D_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}.$$

da cui le regioni di rifiuto

Sistema d'ipotesi		
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$
$\left\lvert \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \right\rvert > t_{n-1,1-\alpha/2}$	$\frac{\bar{X}-\mu_0}{S/\sqrt{n}} > t_{n-1,1-\alpha}$	$\frac{\bar{X}-\mu_0}{S/\sqrt{n}} < -t_{n-1,1-\alpha}$

Regioni di rifiuto: quadro complessivo

	Sistema d'ipotesi		
	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$
Normale, σ noto	$\left\lvert \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \right\rvert > z_{1-\alpha/2}$	$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}$	$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < -z_{1-\alpha}$
Normale, σ non noto	$\left\lvert \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \right\rvert > t_{n-1,1-\alpha/2}$	$\frac{\bar{X}-\mu_0}{S/\sqrt{n}} > t_{n-1,1-\alpha}$	$\frac{\bar{X}-\mu_0}{S/\sqrt{n}} < -t_{n-1,1-\alpha}$
	$H_0 : \pi = \pi_0$ $H_1 : \pi \neq \pi_0$	$H_0 : \pi \leq \pi_0$ $H_1 : \pi > \pi_0$	$H_0 : \pi \geq \pi_0$ $H_1 : \pi < \pi_0$
Proporzione	$\left\lvert \frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \right\rvert > z_{1-\alpha/2}$	$\frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} > z_{1-\alpha}$	$\frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} < -z_{1-\alpha}$

Esempio: controllo di processo, varianza nota

Un macchinario per l'inscatolamento di fagioli produce confezioni il cui peso è distribuito secondo una normale di media μ e varianza $\sigma^2 = 9$ nota.

Le confezioni hanno peso dichiarato $100g$, e così viene impostata la macchina, tuttavia con l'uso la macchina potrebbe produrre con media diversa, nel qual caso è necessario sospendere la produzione e registrare la macchina.

Per decidere quando è necessario registrare la macchina si prende, periodicamente, un campione di n confezioni e le si pesa, se il peso medio si discosta troppo da 100 si sospende la produzione.

Per decidere quando il peso medio è troppo diverso da 100 si impiega una regione di rifiuto al 5%.

Controllo di processo

Si determini la regione di rifiuto supponendo di osservare campioni di $n = 10$ confezioni.

Il sistema d'ipotesi è

$$H_0 : \mu = 100, \quad H_1 : \mu \neq 100$$

quindi la regione di rifiuto è

$$\left| \frac{\bar{x} - 100}{3/\sqrt{10}} \right| > z_{0.975} = 1.96$$

in altre parole si sospende la produzione se il peso medio delle 10 confezioni cade fuori dall'intervallo

$$100 \pm 1.96 \times 0.948 \rightarrow [98.14, 101.86]$$

Controllo di processo

Se la macchina funzionasse perfettamente (la media è sempre 100), quante volte ci si aspetta di sospendere (inutilmente) la produzione?

Per come è costruita la regione di rifiuto, il 5% delle volte.

Se la macchina produce confezioni con peso medio 98, qual è la probabilità che la produzione venga interrotta?

Si tratta della probabilità di rifiutare supponendo pari a 98 la media del processo, che si ottiene come

$$\Phi\left(\frac{98.14-98}{3/\sqrt{10}}\right)+\left(1-\Phi\left(\frac{101.86-98}{3/\sqrt{10}}\right)\right)\approx 0.559$$

Controllo di processo

Come cambiano le risposte alle domande precedenti se si osservano $n = 30$ confezioni?

- la regione di rifiuto al 5% è

$$100 \pm 1.96 \times 3/\sqrt{30} = 100 \pm 1.96 \times 0.548 \rightarrow [98.93, 101.07]$$

- la frequenza con cui si ferma la macchina se questa funziona perfettamente è sempre il 5%
- la probabilità cercata è

$$\Phi\left(\frac{98.93 - 98}{3/\sqrt{30}}\right) + \left(1 - \Phi\left(\frac{101.07 - 98}{3/\sqrt{30}}\right)\right) \approx 0.955$$

Si noti che, fissata la probabilità di errore di I tipo, aumentando la numerosità campionaria diminuisce la probabilità di errore di II tipo.

Controllo di processo, varianza non nota

Si determini la regione di rifiuto supponendo di osservare campioni di $n = 10$ confezioni ma dove la varianza non è nota.

Il sistema d'ipotesi è sempre il medesimo

$$H_0 : \mu = 100, \quad H_1 : \mu \neq 100$$

la regione di rifiuto è però

$$\left| \frac{\bar{x} - 100}{S/\sqrt{10}} \right| > t_{9,0.975} = 2.26$$

Se, ad esempio, si osservasse un campione con $S^2 = 9$ (pari alla varianza nota nella prima trattazione), si sospenderebbe la produzione se il peso medio delle 10 confezioni cadesse fuori dall'intervallo

$$100 \pm 2.26 \times 0.948 \rightarrow [97.86, 102.14]$$

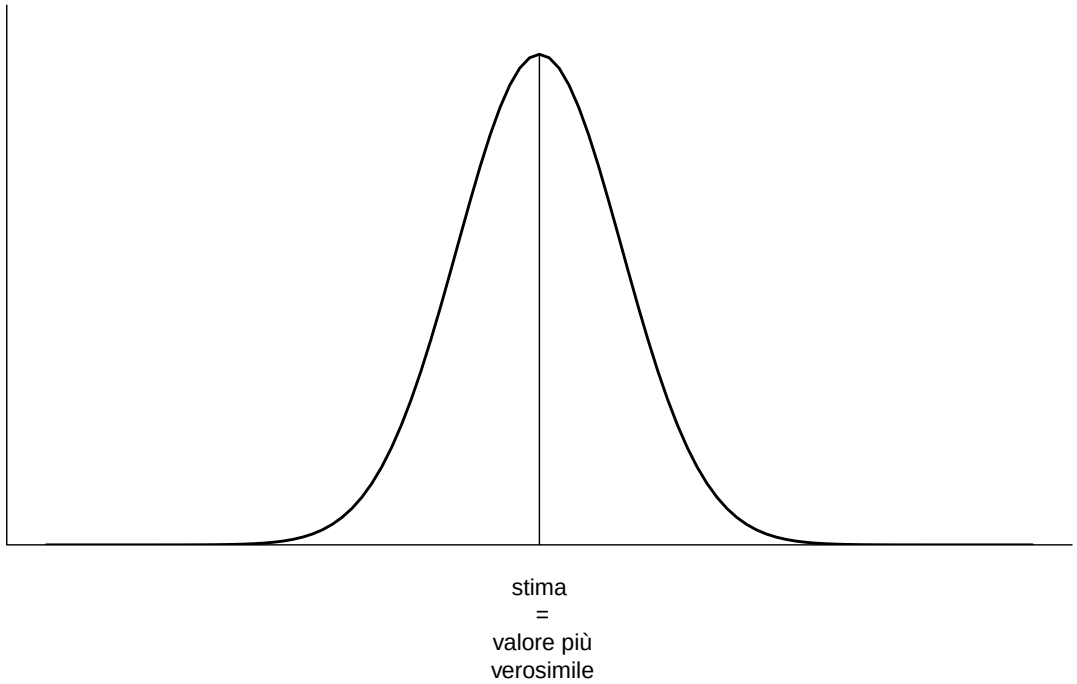
che è, si noti, più ampio di quello ottenuto supponendo $\sigma^2 = 9$ nota.

Indice

Stima puntuale e intervallare

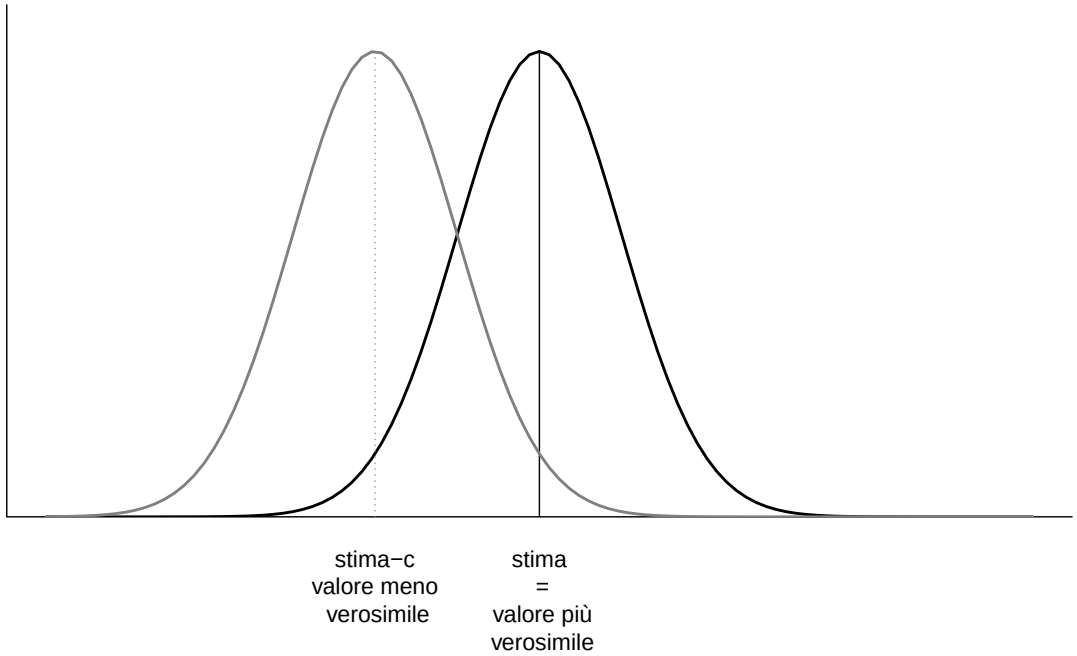
- Nella stima puntuale, calcoliamo dal campione un valore plausibile per il parametro.
- Il c.d.p. ci dice che quel valore è probabilmente vicino al vero valore del parametro.
- Ci dice qualcosa di più, perché ci dice, attraverso la distribuzione campionaria, quanto distante.
- Possiamo usare questa informazione per ottenere una stima intervallare.
- Un intervallo è un modo di esprimere stima e incertezza insieme.

Idea generale



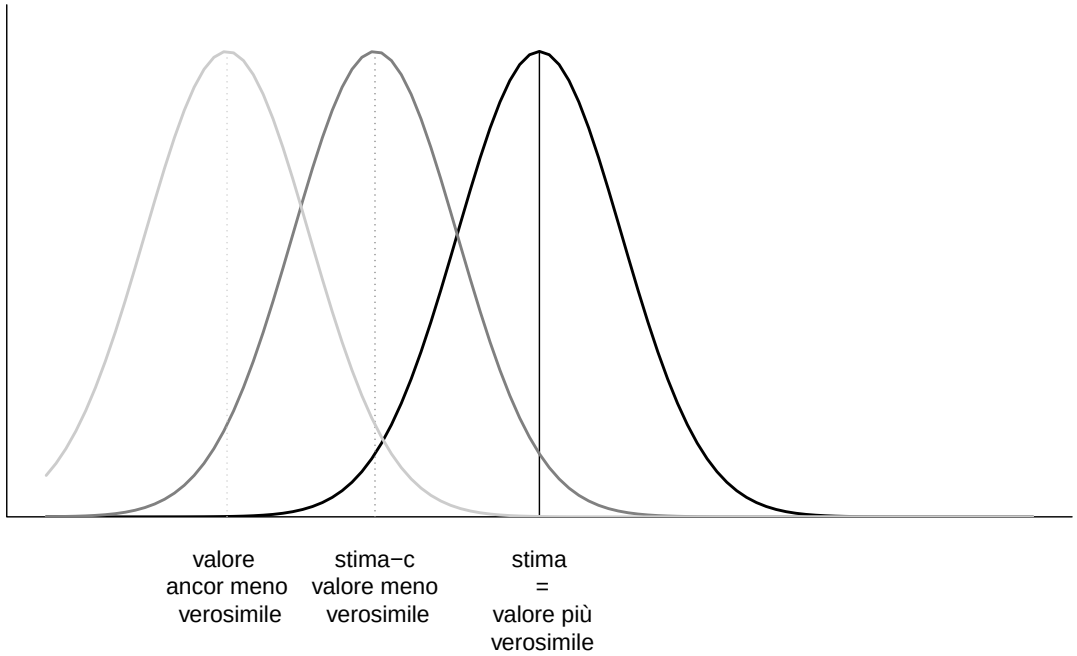
- Dal campione si calcola la stima, che è il valore più verosimile alla luce del campione.
- Sappiamo, un po' vagamente, che il parametro "non è tanto distante" dalla stima, probabilmente.

Idea generale (continua)



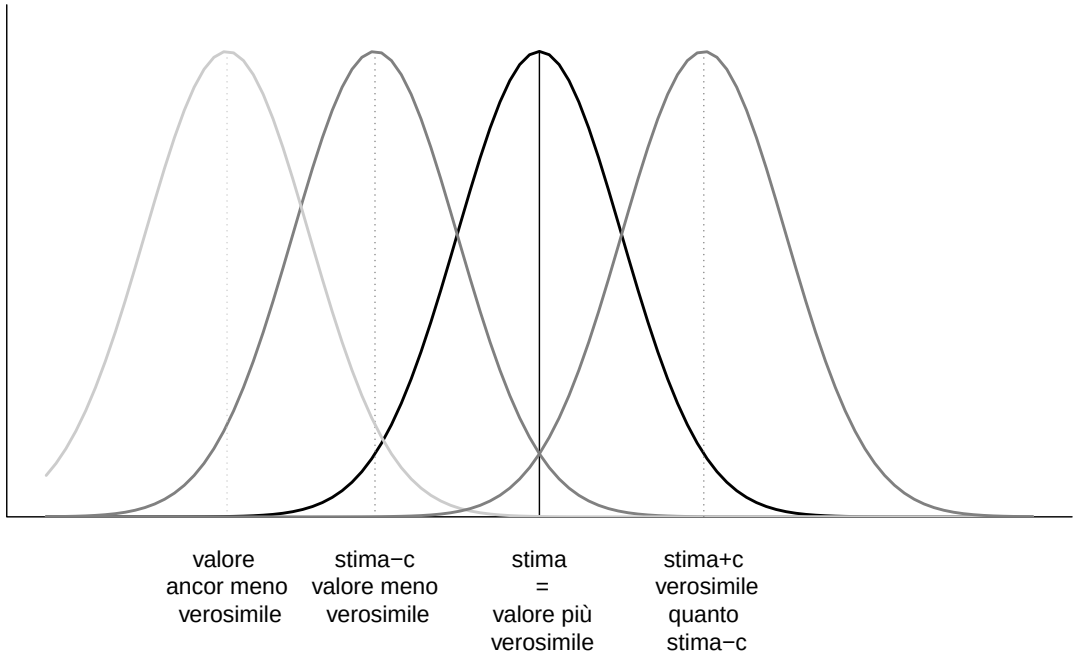
- Se ci discostiamo dalla stima, in stima $-$ c per esempio, il valore è meno verosimile.

Idea generale (continua)



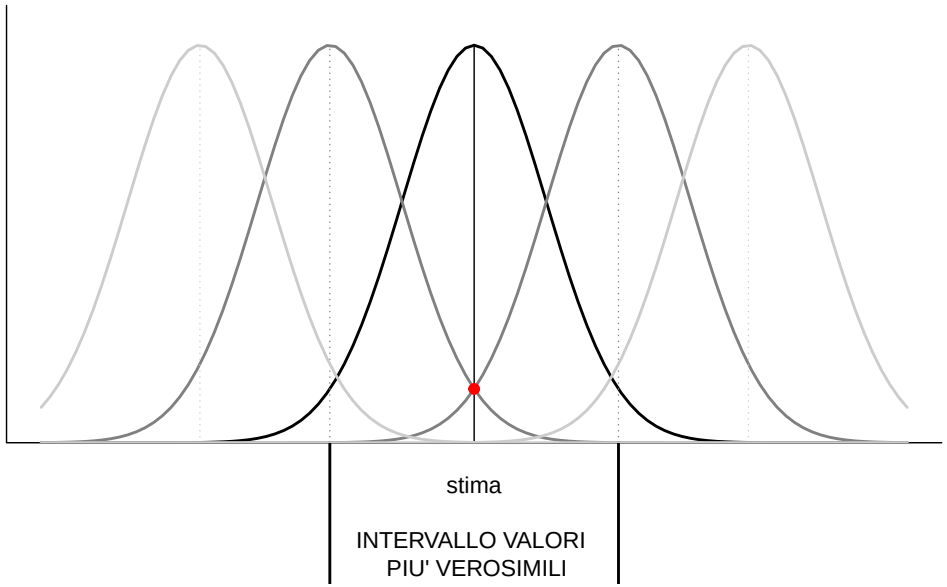
- Se ci discostiamo dalla stima, in stima $-$ c per esempio, il valore è meno verosimile.
- Più ci spostiamo meno verosimile è il valore.

Idea generale (continua)



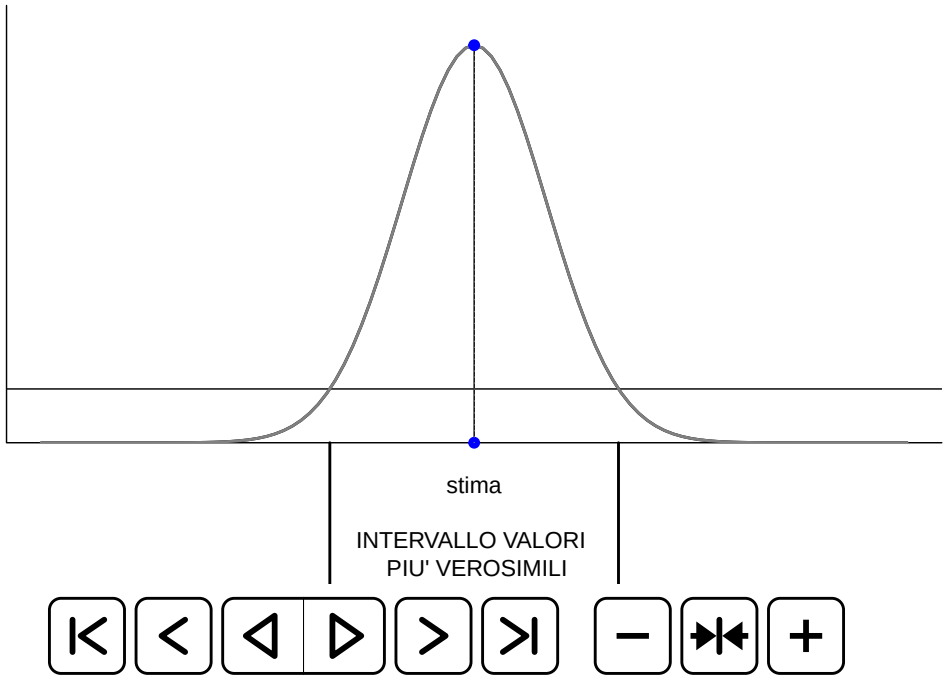
- Se ci discostiamo dalla stima, in stima $-$ c per esempio, il valore è meno verosimile.
- Più ci spostiamo meno verosimile è il valore.
- Se ci spostiamo dall'altra parte, accade lo stesso, simmetricamente.

Idea generale (continua)



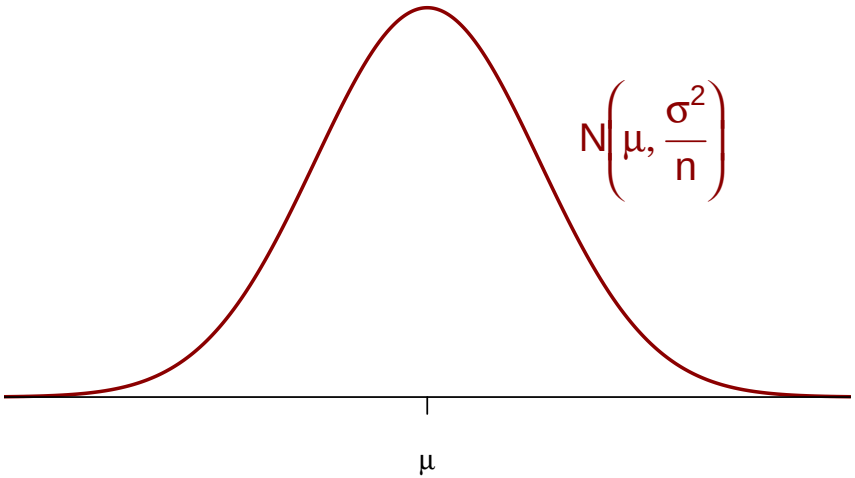
Ha senso quindi determinare un intervallo di valori, più verosimili degli altri.

Idea generale (continua)



I.C. per la media di una normale, varianza nota

Consideriamo lo stimatore media campionaria, $\bar{X}$ per la media di una normale

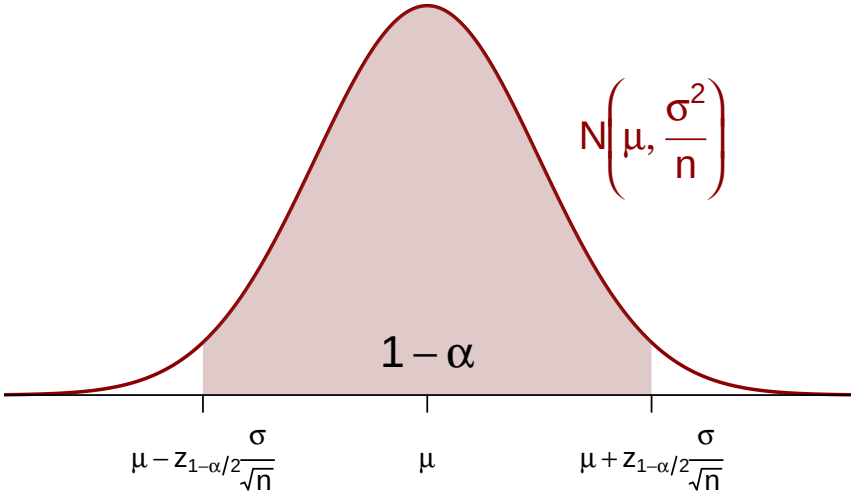


Il punto di partenza è

$$D = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

I.C. per la media di una normale, varianza nota

Consideriamo lo stimatore media campionaria, $\bar{X}$ per la media di una normale



Il punto di partenza è

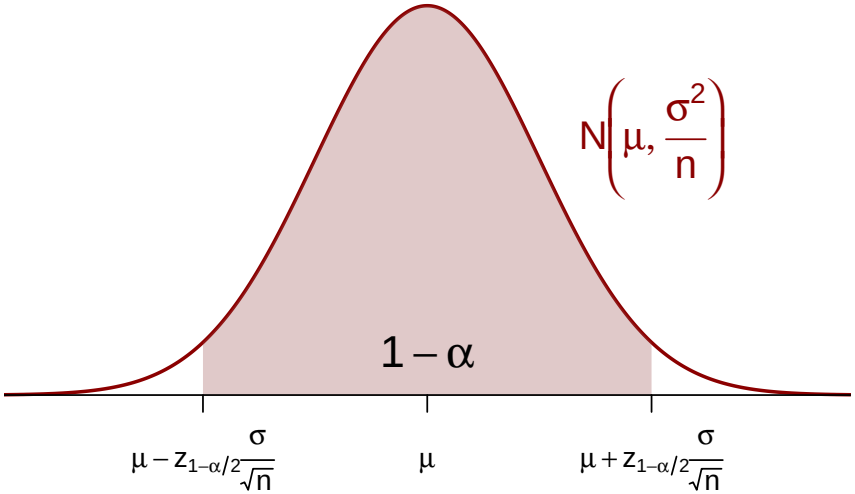
$$D = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

da cui si ha

$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

I.C. per la media di una normale, varianza nota

Consideriamo lo stimatore media campionaria, $\bar{X}$ per la media di una normale



Il punto di partenza è

$$D = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

da cui si ha

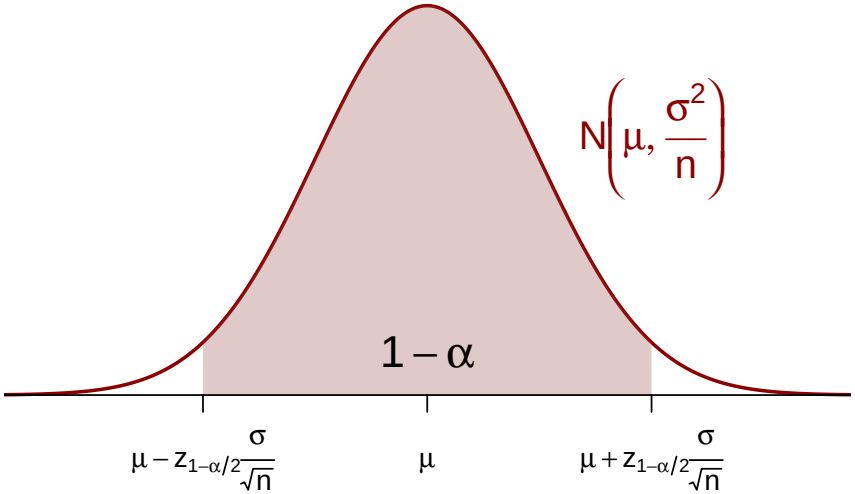
$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

equivalentemente possiamo scrivere

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

I.C. per la media di una normale, varianza nota

Consideriamo lo stimatore media campionaria, $\bar{X}$ per la media di una normale



Il punto di partenza è

$$D = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

da cui si ha

$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

equivalentemente possiamo scrivere

$$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

L'intervallo aleatorio di estremi

$$\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

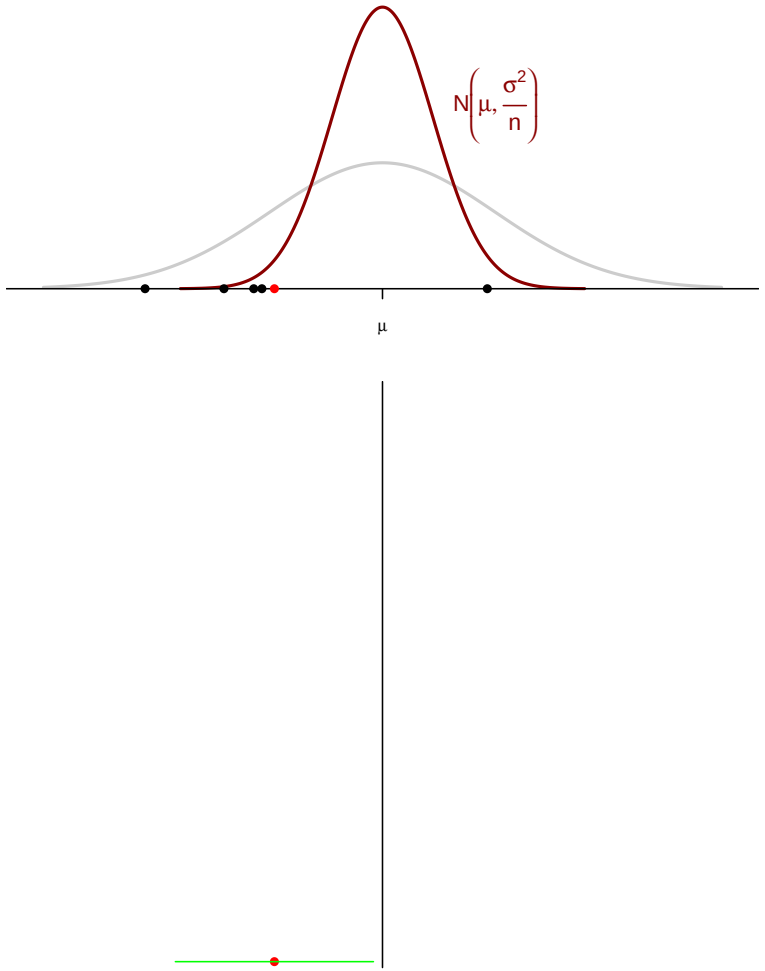
contiene la media della popolazione μ con probabilità $1 - \alpha$.

Intervallo di confidenza

Un intervallo di confidenza è un intervallo i cui estremi sono funzione del campione (esso è dunque aleatorio).

Gli estremi sono costruiti in modo che sia pari a un livello prefissato, detto livello di confidenza, la probabilità che l'intervallo contenga il vero valore del parametro.

Intervallo di confidenza



Esempio

Il macchinario che inscatola i fagioli in un'industria alimentare produce confezioni il cui peso effettivo si distribuisce secondo una normale con una media μ e deviazione standard pari a 3 grammi, si vuole stimare il peso medio μ.

Si pesano 10 confezioni e la media campionaria è pari a 101.

Si vuole ottenere un intervallo di confidenza al 95% per μ.

Se l'intervallo desiderato è al 95% il quantile rilevante della normale è

$$z_{1-0.05/2} = z_{0.975} = 1.96$$

L'intervallo ha dunque estremi

$$101 \pm 1.96 \times \frac{3}{\sqrt{10}} = 101 \pm 1.86$$

Si vuole ottenere un intervallo di confidenza al 99% per μ.

Esempio

Il macchinario che inscatola i fagioli in un'industria alimentare produce confezioni il cui peso effettivo si distribuisce secondo una normale con una media μ e deviazione standard pari a 3 grammi, si vuole stimare il peso medio μ.

Si vuole ottenere un intervallo di confidenza al 95% che sia però non più ampio di 1 grammo, quanto dev'essere grande il campione.

L'ampiezza dell'intervallo è

$$2z_{0.975}\frac{3}{\sqrt{n}}$$

se si vuole che sia al più 1 dev'essere

$$2z_{0.975}\frac{3}{\sqrt{n}} \leq 1 \Rightarrow n \geq (6z_{0.975})^2 = 138.29$$

quindi il campione deve avere almeno 139 unità.

Ampiezza dell'intervallo e dimensione campionaria

L'ampiezza dell'intervallo è

$$2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

per cui se si vuole che l'intervallo di livello $1-\alpha$ sia al più lungo c dev'essere

$$2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq c$$

da cui

$$n \geq 4\left(z_{1-\alpha/2}\frac{\sigma}{c}\right)^2$$

Intervallo di confidenza per la media della normale, varianza non nota

Nel caso in cui la varianza non sia nota, si usa lo stimatore S^2 e il punto di partenza è

$$D = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

da cui si ha

$$P\left(-t_{n-1,1-\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1,1-\alpha/2}\right) = 1 - \alpha$$

che si riscrive

$$P\left(\bar{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

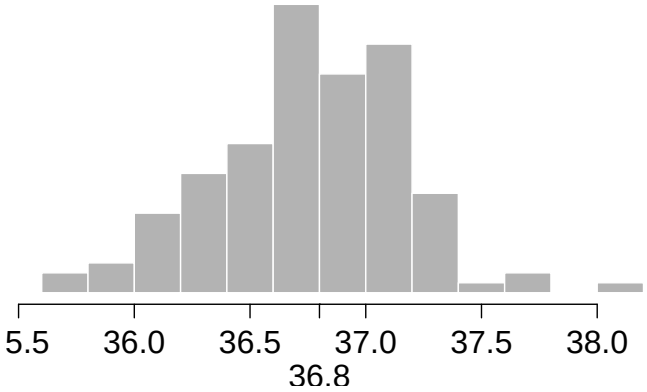
L'intervallo aleatorio di estremi

$$\left[\bar{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

contiene la media della popolazione μ con probabilità $1 - \alpha$.

Esempio: i.c. per la temperatura corporea

Riconsideriamo i dati sulla temperatura corporea



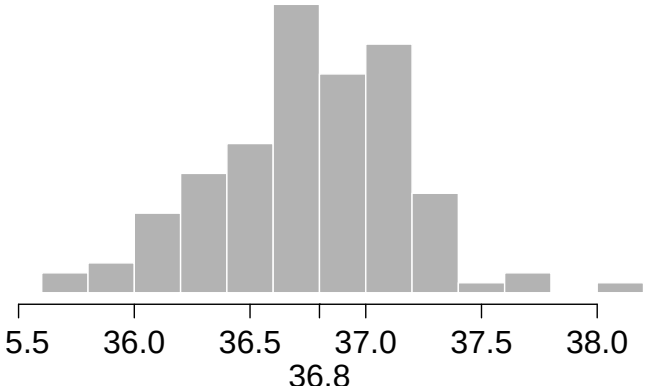
$$n = 130$$

$$\bar{x} = 36.8$$

$$S^2 = 0.166$$

Esempio: i.c. per la temperatura corporea

Riconsideriamo i dati sulla temperatura corporea



$$n = 130$$

$$\bar{x} = 36.8$$

$$S^2 = 0.166$$

Il riferimento è alla t di Student con 129 gradi di libertà (che poi di fatto è quasi uguale alla normale standard)

$$\left[\bar{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

$$\left[36.8 - t_{129,0.975}\frac{0.407}{\sqrt{130}}, 36.8 + t_{129,0.975}\frac{0.407}{\sqrt{130}}\right]$$

$$[36.8 - t_{129,0.975} \times 0.036, 36.8 + t_{129,0.975} \times 0.036]$$

$$[36.73, 36.87]$$

Intervallo di confidenza per la proporzione

Per la proporzione si ha, approssimativamente

$$D = \frac{\hat{\pi} - \pi}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} \sim \mathcal{N}(0,1)$$

da cui

$$P\left(\pi - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq \hat{\pi} \leq \pi + -z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) = 1-\alpha$$

che si riscrive

$$P\left(\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq \pi \leq \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) = 1-\alpha$$

L'intervallo aleatorio di estremi

$$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

contiene la proporzione della popolazione π con probabilità $1-\alpha$.

Stima intervallare della probabilità che nasca un maschio

Usiamo i dati relativi alla città di Muggia, dove su $n = 85$ nati $x = 38$ sono maschi, si ha allora

$$\hat{\pi} = \frac{38}{85} = 0.447$$

un intervallo di confidenza al 95% è dunque

$$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

$$\left[0.447 - z_{0.975}\sqrt{\frac{0.447(1-0.447)}{85}}, 0.447 + z_{0.975}\sqrt{\frac{0.447(1-0.447)}{85}}\right]$$

$$[0.447 - z_{0.975}0.0539, 0.447 + z_{0.975}0.0539]$$

$$[0.341, 0.553]$$

Stima intervallare della probabilità che nasca un maschio

Usando i dati relativi alla regione FVG, dove su $n = 10337$ nati $x = 5286$ sono maschi, si ha allora

$$\hat{\pi} = \frac{5286}{10337} = 0.511$$

un intervallo di confidenza al 95% è dunque

$$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

$$\left[0.511 - z_{0.975}\sqrt{\frac{0.511(1-0.511)}{10337}}, 0.511 + z_{0.975}\sqrt{\frac{0.511(1-0.511)}{10337}}\right]$$

$$[0.511 - z_{0.975}0.00492, 0.511 + z_{0.975}0.00492]$$

$$[0.501, 0.521]$$

Stima intervallare della probabilità che nasca un maschio

Nella tabella sotto si riportano gli i.c. per la probabilità che nasca un maschio valutati usando come campione i nati di Muggia, quelli del FVG e quelli dell'Italia.

x	n	$\hat{\pi}$	$\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$	i.c. 95%		i.c. 98%	
38	85	0.447	0.053900	0.341	0.553	0.308	0.586
5286	10337	0.511	0.004920	0.501	0.521	0.498	0.524
289185	561944	0.515	0.000667	0.514	0.516	0.513	0.517

Intervalli di confidenza: quadro

Normale, σ noto	$\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$
Normale, σ non noto	$\left[\bar{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right]$
Proporzione	$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$

Indice

Due popolazioni

Sin qui si è considerata l'inferenza per una singola popolazione (cioè interessavano le proprietà di degli individui omogenei).

Come si è già detto nel contesto della statistica descrittiva, le situazioni di maggiore interesse prevedono dei confronti tra gruppi diversi, ad esempio

- vogliamo verificare se i redditi medi di due regioni siano diversi e in che misura;
- vogliamo verificare se un farmaco di nuova concezione è più efficace dell'esistente (o di un placebo) e in che misura lo è.

Occorrono a tali fini degli strumenti inferenziali per costruire intervalli di confidenza e fare test (ottenere valori p o regioni di rifiuto) nel contesto del confronto di popolazioni.

Esempio: efficacia di un farmaco

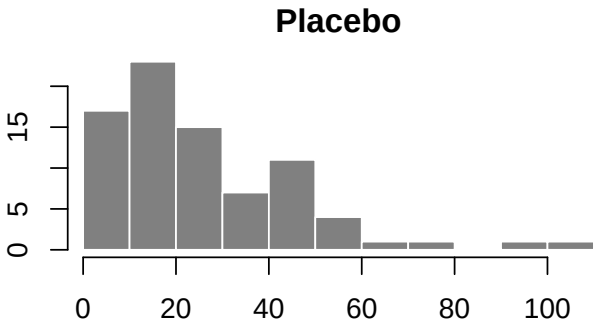
La verifica dell'efficacia di un farmaco è una delle applicazioni tipiche della statistica oggi.
Una casa farmaceutica vuole verificare l'efficacia di un farmaco contro l'insonnia

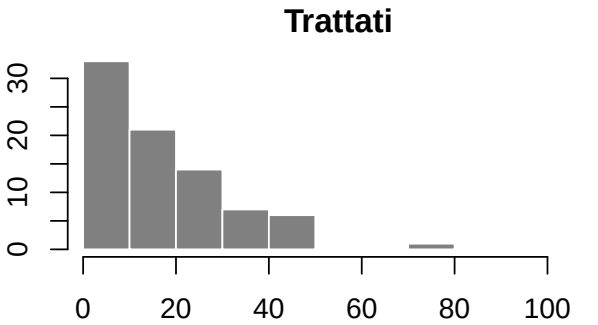
- Consideriamo n pazienti che soffrono di insonnia,
- li dividiamo casualmente in due gruppi,
- a uno somministriamo il farmaco, all'altro il placebo.

Misuriamo l'efficacia con

- Latenza (minuti) prima del sonno prolungato dopo un risveglio notturno (misurato oggettivamente).

Dati sulla latenza nei due gruppi





	pl	tr
n	81.00	82.00
media	26.05	17.26
S2	414.05	196.12
S	20.35	14.00
min	1.00	1.00
max	109.00	73.00

Naturalmente, il fenomeno "sonno" ha una certa variabilità, cioè, se anche il sonnifero è efficace, non è che tutti quelli a cui è stato somministrato si riaddormentano in meno tempo.

Confronto i tempi medi nei due gruppi.

Il fatto che il tempo medio nel gruppo dei trattati sia minore

- è quello che ci si aspetta se il sonnifero funziona
- potrebbe essere però effetto del caso (se il sonnifero è inefficace mi aspetto lo stesso tempo nei due gruppi, ma non sarà mai esattamente uguale).

Confronto di popolazioni

Ragioniamo pensando di avere due popolazioni,

- quelli del placebo
- i trattati

Osservando due campioni, osservo due medie campionarie $\bar{X}_1$ e $\bar{X}_2$ che, se le due popolazioni sono normali, sono distribuite secondo normali

$$\bar{X}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \bar{X}_2 \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Confronto di popolazioni

Ragioniamo pensando di avere due popolazioni, a ciascuna popolazione è associato un tempo medio di latenza

- quelli del placebo tempo medio latenza: μ_1
- i trattati tempo medio latenza: μ_2

Osservando due campioni, osservo due medie campionarie $\bar{X}_1$ e $\bar{X}_2$ che, se le due popolazioni sono normali, sono distribuite secondo normali

$$\bar{X}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \bar{X}_2 \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Quantità di interesse e sua stima

L'oggetto di interesse è ora la differenza tra le medie

$$\mu_1 - \mu_2$$

e in particolare interessa **verificare l'ipotesi** che il farmaco sia inefficace, cioè

$$\mu_1 = \mu_2 \quad \text{ovvero } \mu_1 - \mu_2 = 0$$

La differenza $\mu_1 - \mu_2$ si può ragionevolmente stimare con

$$\bar{X}_1 - \bar{X}_2$$

dove notiamo che, supponendo siano **indipendenti** i due campioni,

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Quantità di riferimento per il confronto

Allora come prima usavo il fatto che

$$\bar{X}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

per verificare un'ipotesi su μ_1 ma anche per ricavare un intervallo di confidenza per μ_1, ora potrei fare l'analogo per $\mu_1 - \mu_2$ usando $\bar{X}_1 - \bar{X}_2$, usiamo cioè il fatto che

$$D = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$$

per ricavare intervalli di confidenza, test di significatività e regioni di rifiuto per $\mu_1 - \mu_2$.

Intervallo di confidenza per la differenza tra medie

Si parta da

$$P\left(-z_{1-\alpha/2} \leq \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

che può essere riscritta

$$P\left(\bar{X}_1 - \bar{X}_2 - z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

che mostra che l'intervallo di estremi

$$(\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

è un intervallo di confidenza di livello $1 - \alpha$ per $\mu_1 - \mu_2$.

Test di significatività bilaterale per la differenza tra medie

Consideriamo l'ipotesi nulla

$$H_0 : \mu_1 - \mu_2 = 0$$

che, si noti, equivale a $H_0 : \mu_1 = \mu_2$, e definniamo

$$D_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (\sim \mathcal{N}(0,1) \text{ se } H_0 \text{ vera})$$

che è tanto più grande in valore assoluto quanto più il campione si discosta dall'ipotesi, il valore p è allora, indicando con d_0 il valore osservato di D_0,

$$\begin{aligned} P(|D_0| > |d_0|) &= P\left(\left|\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right| > \left|\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right|\right) \\ &= 2\left(1 - \Phi\left(\left|\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right|\right)\right) \end{aligned}$$

Test di significatività unilaterale per la differenza tra medie

Se l'ipotesi nulla è

$$H_0 : \mu_1 - \mu_2 \leq 0$$

(che equivale a $H_0 : \mu_1 \leq \mu_2$), ci riferiamo sempre alla quantità D_0 ma lo scostamento è indicato da valori grandi (positivi), quindi il valore p è

$$\begin{aligned} P(D_0 > d_0) &= P\left(\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right) \\ &= 1 - \Phi\left(\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right) \end{aligned}$$

(Si ricavi il valore p per il caso di $H_0 : \mu_1 - \mu_2 \geq 0$.)

Regioni di rifiuto bilaterale per la differenza tra medie

Consideriamo il sistema d'ipotesi

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 \neq 0$$

che, si noti, equivale a $H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$.

La regione

$$\mathcal{R} = \{|D_0| > z_{1-\alpha/2}\} = \left\{ \left| \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| > z_{1-\alpha/2} \right\}$$

è una regione di rifiuto di livello α, infatti

$$P_{H_0}(|D_0| > z_{1-\alpha/2}) = \alpha$$

Regioni di rifiuto unilaterale per la differenza tra medie

Consideriamo il sistema d'ipotesi

$$H_0 : \mu_1 - \mu_2 \leq 0, \quad H_1 : \mu_1 - \mu_2 > 0$$

che, si noti, equivale a $H_0 : \mu_1 \leq \mu_2, \quad H_1 : \mu_1 > \mu_2$.

La regione

$$\mathcal{R} = \{D_0 > z_{1-\alpha}\} = \left\{ \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{1-\alpha} \right\}$$

è una regione di rifiuto di livello α.

si ottenga la regione di rifiuto per il caso
$H_0 : \mu_1 - \mu_2 \geq 0, \quad H_1 : \mu_1 - \mu_2 < 0$

Applicazione al sonnifero?

I metodi sopra non possono essere applicati ai dati sul sonnifero in quanto non conosciamo le varianze (come tipicamente avviene).

Occorre quindi estendere quanto visto sopra al caso in cui la varianza debba essere stimata.

Le cose procedono in maniera analoga a quanto fatto nel caso di un campione, salvo un'ipotesi aggiuntiva: l'eguaglianza delle varianze nei due gruppi.

Indice

Caso della varianza non nota

Nel caso la varianza debba essere stimata, bisogna supporre che sia la medesima nelle due popolazioni,

$$\mathcal{N}(\mu_1, \sigma^2), \;\; \mathcal{N}(\mu_2, \sigma^2)$$

La varianza verrà stimata con

$$\begin{aligned} S_p^2 &= \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (X_{1,i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2} (X_{2,i} - \bar{X}_2)^2 \right) \\ &= \frac{1}{n_1 + n_2 - 2} \left((n_1 - 1) S_1^2 + (n_2 - 1) S_2^2 \right) \end{aligned}$$

e si ha dunque

$$D = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}$$

Intervallo di confidenza per la differenza tra medie con varianza non nota

Dal fatto che

$$D = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

e quindi $P(-t_{n_1+n_2-2,1-\alpha/2} \leq D \leq t_{n_1+n_2-2,1-\alpha/2}) = 1 - \alpha$, cioè

$$P\left((\mu_1 - \mu_2) - t_{n_1+n_2-2,1-\alpha/2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \bar{X}_1 - \bar{X}_2 \leq (\mu_1 - \mu_2) + t_{n_1+n_2-2,1-\alpha/2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha$$

si ricava che l'intervallo di estremi

$$\bar{X}_1 - \bar{X}_2 \pm t_{n_1+n_2-2,1-\alpha/2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

è un i.c. di livello $1 - \alpha$ per $\mu_1 - \mu_2$.

Intervallo di confidenza al 95% per la differenza dei tempi di latenza

Nell'esperimento sul sonnifero si ha

$$\bar{X}_1 = 26.05; \quad \bar{X}_2 = 17.26$$

inoltre $S_1^2 = 414.05$ e $S_2^2 = 196.12$ e $n_1 = 81$, $n_2 = 82$, quindi

$$S_p^2 = \frac{1}{81+82}(81 \times 414.05 + 82 \times 196.12 = 304.42)$$

e $S_p^2 = \sqrt{304.42} = 17.45$, infine $t_{161,0.975} \approx 1.97$, sicché l'intervallo cercato ha estremi

$$(26.05 - 17.26) \pm 1.97 \times 17.45\sqrt{\frac{1}{81} + \frac{1}{82}} = (26.05 - 17.26) \pm 5.38$$

cioè è $[3.41, 14.17]$.

Test per la differenza tra medie di popolazioni normali con varianza non nota

Nell'ipotesi

$$H_0 : \mu_1 - \mu_2 = 0$$

si ha

$$D_0 = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

da cui si ottiene il valore p

$$2(1 - F_{t_{n_1+n_2-2}}(|d_0|))$$

Mentre la regione di rifiuto di livello α è

$$|d_0| > t_{n_1+n_2-2,1-\alpha/2}$$

Valore p per l'ipotesi di eguaglianza delle medie dei tempi di latenza

Si ha $n_1 = 81$, $n_2 = 82$ e

$$\bar{X}_1 = 26.05; \quad \bar{X}_2 = 17.26; \quad S_p = 17.45$$

quindi

$$D = \frac{26.05 - 17.26}{17.45 \times \sqrt{\frac{1}{81} + \frac{1}{82}}} = \frac{8.79}{2.73} = 3.22$$

per calcolare il valore p dovrei conoscere la FdR della t_{161}, ma questa è approssimativamente uguale alla normale standard, quindi

$$2(1 - F_{t_{161}}(3.22)) \approx 2(1 - \Phi(3.22)) \approx 0.0013$$

Regione di rifiuto per l'ipotesi di eguaglianza delle medie dei tempi di latenza

Si ha

$$t_{161,0.975} \approx 1.97$$

e quindi si rifiuta l'ipotesi nulla al livello di significatività del 5% se $|d_0| > 1.97$, avendo osservato $d_0 = 3.22$ l'ipotesi nulla è rifiutata.

Indice

Inferenza sulla media per grandi campioni

Se l'obiettivo è l'inferenza sulla media o sulla differenza tra medie e il/i campione/i sono molto grandi, si hanno due semplificazioni

- si possono usare i test ricavati sopra con l'assunzione di normalità anche se la popolazione non è normale
  - questo perché le medie campionarie sono distribuite comunque approssimativamente secondo una normale in virtù del teorema del limite centrale
- si possono usare i test sviluppati per varianze note anche se le varianze sono stimate
  - se l'inferenza è su una singola popolazione, questo secondo punto non cambia sostanzialmente nulla
  - se l'inferenza è sul confronto tra medie, questo consente di evitare l'assunzione di eguaglianza tra le medie.

Differenza tra medie, grandi campioni

Si hanno due popolazioni X_1 e X_2 dalle quali si osservano due campioni di numerosità n_1 e n_2, entrambi elevati, dai quali si ottengono le medie campionarie $\bar{X}_1$ e $\bar{X}_2$, per esse si ha, approssimativamente

$$\bar{X}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \bar{X}_2 \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

supponendo siano indipendenti i due campioni, si ha, approssimativamente,

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \mathcal{N}(0, 1)$$

da cui si possono ricavere i.c. e test.

Intervallo di confidenza per la differenza tra medie, grandi campioni

Si ha dunque l'intervallo di livello $1 - \alpha$ di estremi

$$(\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Test di significatività per la differenza tra medie, grandi campioni

Consideriamo l'ipotesi nulla

$$H_0 : \mu_1 - \mu_2 = 0,$$

il valore p corrispondente è

$$P\left(\left|\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}\right| > \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}\right)$$

che è pari a

$$2\left(1 - \Phi\left(\left|\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}\right|\right)\right)$$

Regioni di rifiuto per la differenza tra medie, grandi campioni

Consideriamo il sistema d'ipotesi

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 \neq 0$$

Si ha la regione di rifiuto di livello α

$$\mathcal{R} = \left\{ \left| \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \right| > z_{1-\alpha/2} \right\}$$

Indice

La poliomielite

La poliomielite è una malattia virale riconosciuta dal 1840, il suo agente eziologico, il poliovirus, è stato identificato nel 1908

Nel 1880, in Europa, iniziarono grandi epidemie che, poco dopo, si diffusero anche negli Stati Uniti, rendendo la poliomielite una delle malattie infantili più temute del XX secolo.

Nel 1910 gran parte del mondo ha sperimentato un drammatico aumento di casi di polio, e le epidemie sono diventate eventi regolari, soprattutto nelle grandi città e durante i mesi estivi.

Le epidemie hanno fornito l'impulso per una corsa verso lo sviluppo di un vaccino.

(Tratto da Wikipedia)

Vaccino per la poliomielite

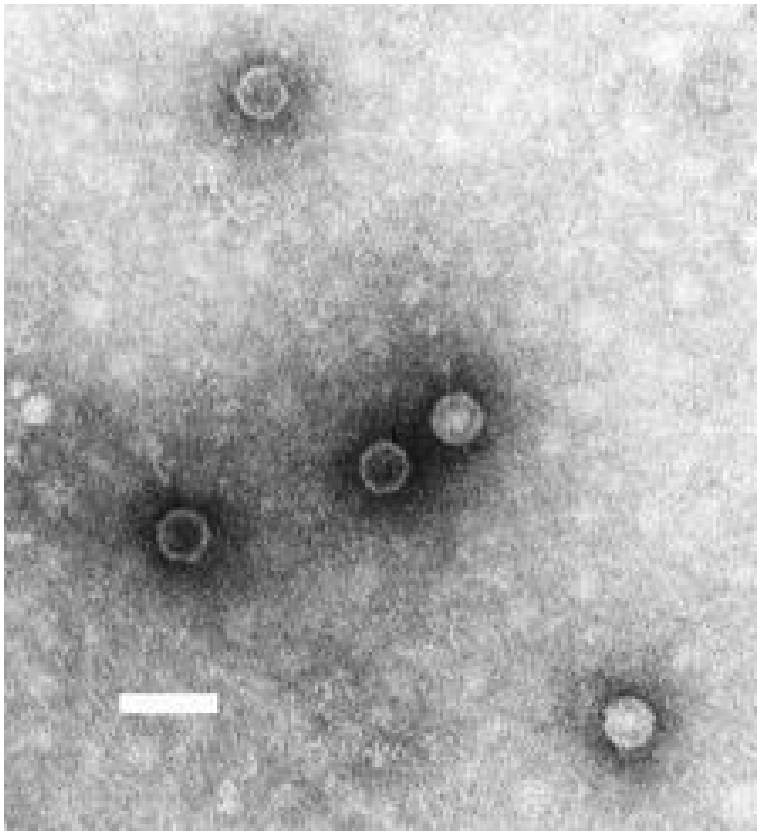
Nel 1950 Jonas Salk sviluppa un vaccino.

—

Esperimenti di laboratorio mostrarono che il vaccino era efficace e non pericoloso.

—

Nel 1954 si passa alla sperimentazione umana, gestita dal *Public Health Service*.



Disegno sperimentale /1

I soggetti dell'esperimento sono bambini tra i 6 e gli 8 anni (I,II e III classe), che sono le età di maggiore vulnerabilità.

---

tra i partecipanti (volontari) all'esperimento

- ad alcuni, scelti a caso, viene somministrato il vaccino $\rightarrow$TRATTATI;
- agli altri viene somministrato un placebo $\rightarrow$CONTROLLI.

si osserva dopo un tempo congruo quanti si ammalano nei due gruppi.

Il vaccino per la poliomielite, risultati

I risultati sperimentali sono i seguenti

	#	Casi	Tasso
Vaccinati	200745	33	0.00016
Placebo	201289	115	0.00057

Si noti che

- la percentuale di ammalati è inferiore nel gruppo dei vaccinati
- come al solito, questo non basta, il fatto di ammalarsi è casuale, la differenza potrebbe essere dovuta al caso e non all'efficacia del vaccino.

Confronto delle proporzioni di due popolazioni

Osservo due campioni di numerosità n_1 e n_2 dai quali si ottengono le proporzioni campionarie di successo $\hat{\pi}_1$ e $\hat{\pi}_2$, per esse si ha

$$\hat{\pi}_1 \sim \mathcal{N}\left(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1}\right), \quad \hat{\pi}_2 \sim \mathcal{N}\left(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

dove π_1 e π_2 sono le probabilità di successo nelle due popolazioni. Supponendo siano indipendenti i due campioni

$$\hat{\pi}_1 - \hat{\pi}_2 \sim \mathcal{N}\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

e si ha dunque

$$\frac{\hat{\pi}_1 - \hat{\pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \sim \mathcal{N}(0,1)$$

Intervallo di confidenza per la differenza tra proporzioni

Per costruire un intervallo di confidenza si fa riferimento alla quantità, approssimatvamente normale,

$$D = \frac{\hat{\pi}_1 - \hat{\pi}_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} \sim \mathcal{N}(0,1)$$

dove $\hat{\pi}_i$ è la proporzione campionaria per il campione i.

L'intervallo di estremi

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

è dunque un intervallo di confidenza di livello $1-\alpha$ per $\pi_1 - \pi_2$.

Test per la differenza tra proporzioni

Per i test sulla differenza tra proporzioni si impiega la quantità

$$D_0 = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}_p(1-\hat{\pi}_p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

dove

$$\hat{\pi}_p = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2}$$

è la proporzione campionaria complessiva per i due campioni.

Nell'ipotesi $H_0 : \pi_1 = \pi_2$ la quantità D_0 è approssimativamente distribuita secondo una $\mathcal{N}(0,1)$, da cui i valori p e le regioni di rifiuto.

Test per il vaccino contro la poliomielite

Calcoliamo

$$\hat{\pi}_p = \frac{33 + 115}{200745 + 201289} \approx 0.0004$$

si ha quindi

$$D_0 = \frac{0.00016 - 0.00057}{\sqrt{0.0004 \times 0.9996 \left(\frac{1}{200745} + \frac{1}{201289}\right)}} = \frac{0.00016 - 0.00057}{0.000063} \approx -6.5$$

il valore p è perciò

$$2(1 - \Phi(6.5)) \approx 3.8 \times 10^{-11}$$

Indice

I vari test

Le verifiche d'ipotesi relative a tutti i sistemi seguenti

- media μ;

$$H_0 : \mu \leq / \geq / = \mu_0; \quad H_1 : \mu > / < / \neq \mu_0$$

- proporzione π;

$$H_0 : \pi \leq / \geq / = \pi_0; \quad H_1 : \pi > / < / \neq \pi_0$$

- differenza tra medie $\mu_1 - \mu_2$;

$$H_0 : \mu_1 - \mu_2 \leq / \geq / = 0; \quad H_1 : \mu_1 - \mu_2 > / < / \neq 0$$

- differenza tra proporzioni $\pi_1 - \pi_2$;

$$H_0 : \pi_1 - \pi_2 \leq / \geq / = 0; \quad H_1 : \pi_1 - \pi_2 > / < / \neq 0$$

N.B.: μ_0, π_0 sono noti!
Sono riconducibili a uno schema comune.

Quadro generale

		Quantità di interesse: Proporzione	Media: σ noto	Media: σ non noto
Singolo, osservo un campione, l'obiettivo è la caratteristica di un unico gruppo	Parametro/i:	π	μ	μ, σ^2
	Stimatore/i:	$\hat{\pi}$	$\bar{X}$	$\bar{X}, S^2$
	Ipotesi:	$\pi \leq / \geq / = \pi_0$	$\mu \leq / \geq / = \mu_0$	
Differenza, osservo due campioni e l'obiettivo è il confronto tra i due	Parametri:	π_1, π_2	μ_1, μ_2	μ_1, μ_2, σ^2
	Stimatori:	$\hat{\pi}_1, \hat{\pi}_2$	$\bar{X}_1, \bar{X}_2$	$\bar{X}_1, \bar{X}_2$
		$\hat{\pi}_p = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1+n_2}$		$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$
	Ipotesi:	$\pi_1 - \pi_2 \leq / \geq / = 0$	$\mu_1 - \mu_2 \leq / \geq / = 0$	

Statistica D_0 in generale

Ci si basa su una statistica che ha la struttura

$$D_0 = \frac{\left[\begin{array}{c}\text{STIMATORE}\\ \text{(dal campione)}\end{array}\right] - \left[\begin{array}{c}\text{IPOTESI}\\ \text{(su popolazione)}\end{array}\right]}{\left[\begin{array}{c}\sqrt{\text{var stim}}\\ \text{secondo}\\ \text{l'ipotesi}\end{array}\right]} = \begin{cases} \frac{\bar{X}-\mu_0}{\text{sd}(\bar{X})} \\ \frac{\hat{\pi}-\pi_0}{\text{sd}(\hat{\pi})} \\ \frac{\bar{X}_1-\bar{X}_2}{\text{sd}(\bar{X}_1-\bar{X}_2)} \\ \frac{\hat{\pi}_1-\hat{\pi}_2}{\text{sd}(\hat{\pi}_1-\hat{\pi}_2)} \end{cases}$$

	Proporzione	Media	
		σ noto	σ non noto
Singolo	$\frac{\hat{\pi}-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$	$\frac{\bar{X}-\mu_0}{S/\sqrt{n}}$
Differenza	$\frac{\hat{\pi}_1-\hat{\pi}_2}{\sqrt{\hat{\pi}_p(1-\hat{\pi}_p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ $\hat{\pi}_p = \frac{n_1\hat{\pi}_1+n_2\hat{\pi}_2}{n_1+n_2}$	$\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$\frac{\bar{X}_1-\bar{X}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$

Valore p

Sulla base del valore d_0 della statistica D_0 calcolo il valore p

		Bilaterale	Unilaterale	
			Destro	Sinistro
	$D_0 \sim F_D$	$2(1 - F_D(\lvert d_0 \rvert))$	$1 - F_D(d_0)$	$F_D(d_0)$
Proporzione Diff. proporzioni Media (σ^2 noto) Diff. medie (σ^2 noto)	$D_0 \sim \mathcal{N}(0,1)$	$2(1 - \Phi(\lvert d_0 \rvert))$	$1 - \Phi(d_0)$	$\Phi(d_0)$
Media (σ^2 non noto)	$D_0 \sim t_{n-1}$	$2(1 - F_{t_{n-1}}(\lvert d_0 \rvert))$	$1 - F_{t_{n-1}}(d_0)$	$F_{t_{n-1}}(d_0)$
Diff. medie (σ^2 non noto)	$D_0 \sim t_{n_1+n_2-2}$	$2(1 - F_{t_{n_1+n_2-2}}(\lvert d_0 \rvert))$	$1 - F_{t_{n_1+n_2-2}}(d_0)$	$F_{t_{n_1+n_2-2}}(d_0)$

Regioni di rifiuto

Sulla base della statistica D_0, si definisce la regione di rifiuto

		Bilaterale	Unilaterale	
			Destro	Sinistro
	$D_0 \sim F_D$	$\lvert d_0 \rvert > F_D^{-1}(1-\alpha/2)$	$d_0 > F_D^{-1}(1-\alpha)$	$d_0 < F_D^{-1}(\alpha)$
Proporzione Diff. proporzioni Media (σ^2 noto) Diff. medie (σ^2 noto)	$D_0 \sim \mathcal{N}(0,1)$	$\lvert d_0 \rvert > z_{1-\alpha/2}$	$d_0 > z_{1-\alpha}$	$d_0 < z_{\alpha}$
Media (σ^2 non noto)	$D_0 \sim t_{n-1}$	$\lvert d_0 \rvert > t_{n-1,1-\alpha/2}$	$d_0 > t_{n-1,1-\alpha/2}$	$d_0 < t_{n-1,\alpha/2}$
Diff. medie (σ^2 non noto)	$D_0 \sim t_{n_1+n_2-2}$	$\lvert d_0 \rvert > t_{n_1+n_2-2,1-\alpha/2}$	$d_0 > t_{n_1+n_2-2,1-\alpha/2}$	$d_0 < t_{n_1+n_2-2,\alpha/2}$

Intervalli di confidenza

Anche gli intervalli di confidenza hanno una struttura simile, che si basa sulla statistica

$$D = \frac{\left[\begin{array}{c}\text{STIMATORE}\\ \text{(dal campione)}\end{array}\right] - \left[\begin{array}{c}\text{PARAMETRO}\\ \text{(su popolazione)}\end{array}\right]}{\left[\sqrt{\begin{array}{c}\text{varianza}\\ \text{stimatore}\end{array}}\right]}$$

La struttura dell'intervallo è

$$\left[\begin{array}{c}\text{STIMATORE}\\ \text{(dal campione)}\end{array}\right] \pm \left[\begin{array}{c}\text{quantile } 1-\alpha/2\\ \text{della distr. di } D\end{array}\right] \times \left[\sqrt{\begin{array}{c}\text{varianza}\\ \text{stimatore}\end{array}}\right]$$

Intervalli di confidenza

	Singolo	Differenza
Proporzione	$\hat{\pi} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$	$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n}}$ $\hat{\pi}_p = \frac{n_1\hat{\pi}_1+n_2\hat{\pi}_2}{n_1+n_2}$
Media σ^2 noto	$\bar{X} \pm z_{1-\alpha/2}\sigma/\sqrt{n}$	$\bar{X}_1 - \bar{X}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Media σ^2 non noto	$\bar{X} \pm t_{n-1,1-\alpha/2}S/\sqrt{n}$	$\bar{X}_1 - \bar{X}_2 \pm t_{n_1+n_2-2,1-\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$

Indice

In generale: verifica d'ipotesi

Indicando con θ il parametro e con $\hat{\theta}$ il suo stimatore, si calcola

$$D_0 = \frac{\hat{\theta} - \theta_0}{\sqrt{V_0(\hat{\theta})}} \sim F_0$$

dove

- $V_0(\hat{\theta})$ è la varianza di $\hat{\theta}$ nell'ipotesi nulla.
- F_0 è la FdR di D_0 nell'ipotesi nulla,

Valore p e regioni di rifiuto sono allora

	Sistema d'ipotesi		
	$H_0 : \theta = \theta_0$ $H_1 : \theta \neq \theta_0$	$H_0 : \theta \leq \theta_0$ $H_1 : \theta > \theta_0$	$H_0 : \theta \geq \theta_0$ $H_1 : \theta < \theta_0$
valore p	$2(1 - F_0(d_0))$	$1 - F_0(d_0)$	$F_0(d_0)$
regione di rifiuto	$\lvert D_0 \rvert > F_0^{-1}(1 - \alpha/2)$	$D_0 > F_0^{-1}(1 - \alpha)$	$D_0 < -F_0^{-1}(1 - \alpha)$

In generale: intervalli di confidenza

Anche gli intervalli di confidenza seguono uno schema comune, che si basa sulla quantità

$$D = \frac{\hat{\theta} - \theta}{\sqrt{V(\hat{\theta})}} \sim F$$

dove

- $V(\hat{\theta})$ è la varianza di $\hat{\theta}$
- F è, secondo i casi, una normale o una t di Student

l'intervallo di livello $1 - \alpha$ è

$$\hat{\theta} \pm F^{-1}(1 - \alpha/2)\sqrt{V(\hat{\theta})}$$

Modello	θ	θ_0	$\hat{\theta}$	V_0	V	F_0, F
$X \sim \mathcal{N}(\mu, \sigma^2)$ σ^2 noto	μ	μ_0	$\bar{X}$	$\frac{\sigma^2}{n}$		N
$X \sim \mathcal{N}(\mu, \sigma^2)$ σ^2non noto	μ	μ_0	$\bar{X}$	$\frac{S^2}{n}$		t_{n-1}
$X \sim \text{Ber}(\pi)$	π	π_0	$\hat{\pi}$	$\frac{\pi_0(1-\pi_0)}{n}$	$\frac{\hat{\pi}(1-\hat{\pi})}{n}$	N
$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ σ_1^2, σ_2^2 note	$\mu_1 - \mu_2$	0	$\bar{X}_1 - \bar{X}_2$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$		N
$X_1 \sim \mathcal{N}(\mu_1, \sigma^2)$ $X_2 \sim \mathcal{N}(\mu_2, \sigma^2)$ σ^2non nota	$\mu_1 - \mu_2$	0	$\bar{X}_1 - \bar{X}_2$	$S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$		$t_{n_1+n_2-1}$
$X_1 \sim \text{Ber}(\pi_1)$ $X_2 \sim \text{Ber}(\pi_2)$	$\pi_1 - \pi_2$	0	$\hat{\pi}_1 - \hat{\pi}_2$	$\hat{\pi}_p(1-\hat{\pi}_p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$	$\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}$	N

Indice

Il primo esempio

Austin Bradford Hill (1897-1991) nel 1948 verifica l'efficacia della streptomicina per la TBC su un campione di 107 tubercolitici, assegnando a caso i trattamenti (l'idea viene dall'agricoltura dove ha sostituito i confronti storici)

- 55, scelti a caso, vengono trattati con streptomicina e riposo
- gli altri 52 vengono trattati col solo riposo

Al termine dell'esperimento

	Morti	Sopravvissuti
Trattati (streptomicina)	4	51
Controlli	14	38

Essendo stati scelti a caso i malati a cui dare la streptomicina, è legittimo attribuire la differenza o al caso o alla streptomicina (tertium non datur).

Valutazione del rischio di credito

L'ufficio prestiti di una banca vuole valutare la probabilità che dei potenziali clienti restituiscano i soldi prestati sulla base degli esiti dei prestiti degli anni passati.

	clienti	insolventi
autonomi	528	45
dipendenti	1021	75
pensionati	754	44
Totale	2303	164

1) Si fornisca una stima della probabilità che il prestito non venga restituito per i lavoratori dipendenti.
2) Si ottenga un i.c. al 98% per la probabilità che il prestito non venga restituito da un lavoratore dipendente.
3) Si dica se, e in che misura, i dati suggeriscano che la probabilità di restituzione è diversa per dipendenti e pensionati.

Valutazione del rischio di credito

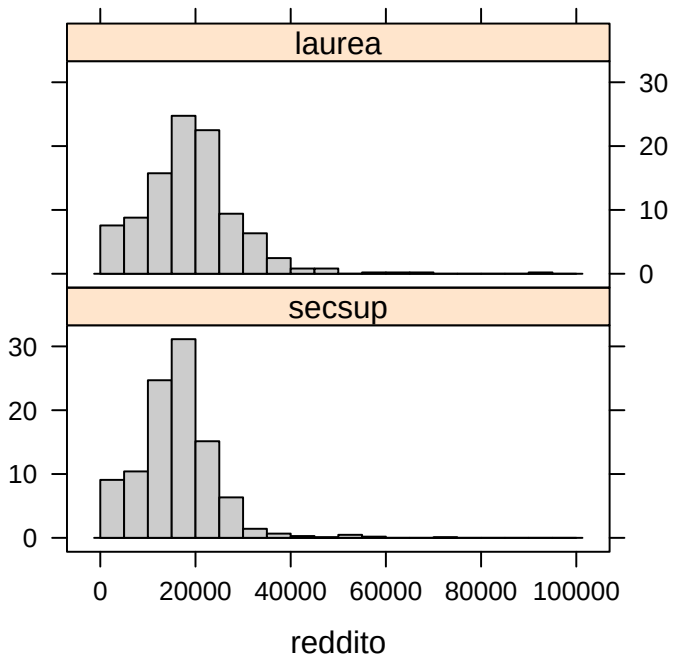
L'ufficio prestiti di una banca vuole valutare la probabilità che dei potenziali clienti restituiscano i soldi prestati sulla base degli esiti dei prestiti degli anni passati.

	clienti	ins	perc	se
autonomi	528	45	8.52	0.012
dipendenti	1021	75	7.35	0.008
pensionati	754	44	5.84	0.009
Totale	2303	164	7.12	0.005

1) Si fornisca una stima della probabilità che il prestito non venga restituito per i lavoratori dipendenti.
2) Si ottenga un i.c. al 98% per la probabilità che il prestito non venga restituito da un lavoratore dipendente.
3) Si dica se, e in che misura, i dati suggeriscano che la probabilità di restituzione è diversa per dipendenti e pensionati.

Dati sul reddito per titolo di studio

Al fine di valutare l'effetto del titolo di studio sul reddito dei lavoratori dipendenti si considera un campione costituito da lavoratori dipendenti con tiolo di scuola secondaria superiore e laureati ed età compresa tra 30 e 35 anni.



	secsup	laurea
n	1057.00	489.00
media	15844.74	19027.29
sqm	8000.36	10166.73
min	81.00	428.00
max	73292.00	94729.00

1) I dati suggeriscono una differenza?
2) Si stimi la differenza fornendo un i.c. al 95%.

Indice

Indice

Osservazioni e teoria

Modalità	Freq. ass.	Freq. rel.
y_1	n_1	f_1
$\vdots$	$\vdots$	$\vdots$
y_i	n_i	f_i
$\vdots$	$\vdots$	$\vdots$
y_k	n_k	f_k

Una distribuzione di frequenza è la descrizione di un insieme di dati.

A volte però abbiamo una teoria/modello su come potrebbe essere fatta la distribuzione, cioè una distribuzione di probabilità

$$\pi_1, \ldots, \pi_k$$

con $\pi_i \geq 0$ e $\sum_{i=1}^{k} = 1$.

Ad esempio, quando lanciamo un dado a 6 facce, abbiamo un modello per cui gli esiti sono equiprobabili.

Osservazioni e teoria

Modalità	Freq. ass.	Freq. rel.	Freq. teorica
y_1	n_1	f_1	π_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_i	n_i	f_i	π_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_k	n_k	f_k	π_k

Una distribuzione di frequenza è la descrizione di un insieme di dati.

A volte però abbiamo una teoria/modello su come potrebbe essere fatta la distribuzione, cioè una distribuzione di probabilità

$$\pi_1, \ldots, \pi_k$$

con $\pi_i \geq 0$ e $\sum_{i=1}^{k} = 1$.

Ad esempio, quando lanciamo un dado a 6 facce, abbiamo un modello per cui gli esiti sono equiprobabili.

Verifica dell'adattamento

Si pone il problema di confrontare la distribuzione di frequenza (osservata) con le frequenze (probabilità) teoriche (ipotizzate).

Ad esempio, lanciamo tante volte un dado e registriamo le facce uscite, poi vogliamo verificare se le frequenze osservate (che non saranno esattamente uguali) sono compatibili con le frequenze teoriche, uguali per tutte le modalità. (In sostanza, vogliamo vedere se il dado è equilibrato.)

Esiti di 96 lanci di un dado a sei facce

Esito	1	2	3	4	5	6
Freq	14	17	16	22	13	14

Confronto delle frequenze

Sulla base del modello, la frequenza attesa dell'i-esiam modalità è

$$n\pi_i$$

Una misura della differenza tra frequenze osservate n_i e frequenze attese $n\pi_i$ è

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i}$$

Nel caso degli esiti di 96 lanci del dado, essendo $\pi_i = 1/6$ per ogni i e quindi $n\pi_i = 16$, vale

$$\chi^2 = \frac{(14-16)^2}{16} + \frac{(17-16)^2}{16} + \frac{(16-16)^2}{16} + \frac{(22-16)^2}{16} + \frac{(13-16)^2}{16} + \frac{(14-16)^2}{16} = 3.375$$

Per stabilire se è grande o piccolo possiamo ragionare in termini di verifica d'ipotesi.

Distribuzione di riferimento

L'ipotesi nulla è

$$H_0 : P(Y = y_i) = \pi_i, \quad i = 1, \ldots, k$$

a fronte delle osservazioni

$$n_1, \ldots, n_k$$

Modalità	Freq. ass.	Freq. rel. teorica
y_1	n_1	π_1
$\vdots$	$\vdots$	$\vdots$
y_i	n_i	π_i
$\vdots$	$\vdots$	$\vdots$
y_k	n_k	π_k

Se

- le osservazioni provengono dalla distribuzione $\pi_1, \ldots, \pi_k$ ipotizzata
- n è grande

allora la statistica

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i}$$

è approssimativamente distribuita come un χ^2_{k-1}

Test di significatività e regione di rifiuto

Le osservazioni sono 'distanti' dall'ipotesi se la statistica χ^2 è grande, quindi il valore p è

$$P(\chi^2_{k-1} > \chi^2) = 1 - F_{\chi^2_{k-1}}(\chi^2)$$

dove $F_{\chi^2_{k-1}}$ indica la FdR di un χ^2 con $k-1$ gradi di libertà.

Inoltre, se $\chi^2_{k-1,1-\alpha}$ indica il quantile $1-\alpha$ di una distribuzione χ^2_{k-1}, la regione

$$\{\chi^2 > \chi^2_{k-1,1-\alpha}\}$$

è una regione di rifiuto di livello α.

Nel caso del dado, con $k = 6$ la regione di rifiuto al 5% è $\{\chi^2 > 11.1\}$ e quindi...

Esempio: Poisson

Si ipotizza che il numero di incidenti che si verifica ogni mese in un tratto di strada sia distribuito secondo una Poisson di parametro $\lambda = 1$.

Si osserva il numero di incidenti per 50 settimane, e si vuole verificare se tale numero è compatibile con l'ipotesi.

y	n
0	6
1	7
2	17
3+	20

Esempio: Poisson

Si ipotizza che il numero di incidenti che si verifica ogni mese in un tratto di strada sia distribuito secondo una Poisson di parametro $\lambda = 1$.

Si osserva il numero di incidenti per 50 settimane, e si vuole verificare se tale numero è compatibile con l'ipotesi.

y	n	π	$n\pi$
0	6	0.37	18.40
1	7	0.37	18.40
2	17	0.18	9.20
3+	20	0.08	4.00

Si calcolano le probabilità teoriche

$$\pi_i = \begin{cases} \frac{\lambda^i}{i!} e^{-\lambda} & i = 0, \ldots, 2 \\ 1 - \sum_{i=0}^{3-1} \pi_i & i = 3 \end{cases}$$

e le frequenze attese $n\pi_i$.

Esempio: Poisson

Si ipotizza che il numero di incidenti che si verifica ogni mese in un tratto di strada sia distribuito secondo una Poisson di parametro $\lambda = 1$.

Si osserva il numero di incidenti per 50 settimane, e si vuole verificare se tale numero è compatibile con l'ipotesi.

y	n	π	$n\pi$
0	6	0.37	18.40
1	7	0.37	18.40
2	17	0.18	9.20
3+	20	0.08	4.00

Si calcolano le probabilità teoriche

$$\pi_i = \begin{cases} \frac{\lambda^i}{i!} e^{-\lambda} & i = 0, \ldots, 2 \\ 1 - \sum_{i=0}^{3-1} \pi_i & i = 3 \end{cases}$$

e le frequenze attese $n\pi_i$.

quindi

$$\chi^2 = \frac{(6-18.4)^2}{18.4} + \frac{(7-18.4)^2}{18.4} + \frac{(17-9.2)^2}{9.2} + \frac{(20-4)^2}{4} = 86.033$$

mentre $\chi^2_{3,0.95} = 7.81$, quindi si rifiuta l'ipotesi.

Osservazione

Il test del χ^2 ci suggerisce di rifiutare l'ipotesi che la distribuzione da cui provengono le osservazioni sia una Poisson(1), nonci dice perché:

- la distribuzione potrebbe non essere Poisson
- la distribuzione potrebbe essere Poisson ma con $\lambda \neq 1$

In particolare, ricordando che se una variabile è distribuita secondo una Poisson(λ), allora ha media λ, si potrebbe stimare il valore di λ con la media campionaria, questa non è calcolabile con i dati a disposizione (per via del fatto che sono messe assieme le modalità maggiori di 3), però certamente è maggiore di

$$\frac{1}{50}\left(1 \times 7 + 2 \times 17 + 3 \times 20\right) = 2.02$$

che è comunque maggiore del valore ipotizzato.

Osservazione (continua)

Ha senso allora confrontare i valori osservati con una Poisson(2.02), usiamo sempre il χ^2

y	n	π	$n\pi$
0	6	0.13	6.65
1	7	0.27	13.40
2	17	0.27	13.55
3+	20	0.33	16.40

e si ottiene

$$\chi^2 = \frac{(6-6.65)^2}{6.65} + \frac{(7-13.4)^2}{13.4} + \frac{(17-13.55)^2}{13.55} + \frac{(20-16.4)^2}{16.4} = 4.7889$$

avendo stimato un parametro, il quantile a cui riferirsi non è quello di un χ^2_3 ma quello di un χ^2_2: $\chi^2_{2,0.95} = 5.99$

Indice

Riprendiamo il Titanic

Riprendiamo la tabella relativa all'esito del naufragio del Titanic per i diversi "tipi" di passeggeri

	1st	2nd	3rd	Sum
No	122	167	528	817
Yes	203	118	178	499
Sum	325	285	706	1316

Ci si era chiesti se le chance di sopravviere erano le stesse per i diversi "tipi" di passeggeri.

In altre parole, si vuole vedere se le distribuzioni del tipo condizionate al fatto di essere sopravvissuti o non sono le stesse (se i due gruppi sono omogenei).

Il problema può anche essere posto in termini di indipendenza: ci chiediamo se **la classe e l'esito sono indipendenti**.

Indipendenza in distribuzione

Ricordiamo che, in una tabella a doppia entrata

			X			
Y	x_1	$\cdots$	x_j	$\cdots$	x_t	totale
y_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1t}	n_{10}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{it}	n_{i0}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_s	n_{s1}	$\cdots$	n_{sj}	$\cdots$	n_{st}	n_{s0}
totale	n_{01}	$\cdots$	n_{0j}	$\cdots$	n_{0t}	N

Y è **indipendente in distribuzione** da X se, per qualsivoglia $i = 1, \ldots, s$,

$$\frac{n_{i1}}{n_{01}} = \frac{n_{i2}}{n_{02}} = \ldots = \frac{n_{ij}}{n_{0j}} = \ldots = \frac{n_{it}}{n_{0t}}$$

cioè se $n_{ij} = \hat{n}_{ij}$ con

$$\hat{n}_{ij} = \frac{n_{i0}\, n_{0j}}{N}.$$

Indice X^2 di Pearson: distribuzione asintotica

Se le due variabili sono indipendenti, la frequenza attesa in corrispondenza alle modalità i e j è

$$\hat{n}_{ij} = \frac{n_{i0} n_{0j}}{N}.$$

Si misura di quanto le frequenze osservate si scostano da quelle teoriche con

$$X^2 = \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}.$$

che abbiamo usato dal punto di vista descrittivo.

L'indice X^2 di Pearson può essere usato anche in termini inferenziali in quanto

- se vi è indipendenza,
- il campione è sufficientemente grande

l'indice X^2 è approssimativamente distribuito secondo un $\chi^2_{(s-1)(t-1)}$.

Test di significatività e regione di rifiuto

Le osservazioni sono 'distanti' dall'ipotesi se la statistica X^2 è grande, quindi il valore p è

$$P(\chi^2_{(s-1)(t-1)} > X^2) = 1 - F_{\chi^2_{(s-1)(t-1)}}(\chi^2)$$

dove $F_{\chi^2_{(s-1)(t-1)}}$ indica la FdR di un χ^2 con $(s-1)(t-1)$ gradi di libertà.

Inoltre, se $\chi^2_{(s-1)(t-1),1-\alpha}$ indica il quantile $1-\alpha$ di una distribuzione $\chi^2_{(s-1)(t-1)}$, la regione

$$\{X^2 > \chi^2_{(s-1)(t-1),1-\alpha}\}$$

è una regione di rifiuto di livello α.

Per il Titanic

Tabella osservata

	1st	2nd	3rd	Sum
No	122	167	528	817
Yes	203	118	178	499
Sum	325	285	706	1316

Tabella teorica

	1st	2nd	3rd	Sum
No	201.8	176.9	438.3	817.0
Yes	123.2	108.1	267.7	499.0
Sum	325.0	285.0	706.0	1316.0

Si calcolano dunque le differenze

$n_{ij} - \hat{n}_{ij}$

	1st	2nd	3rd
No	-79.8	-9.9	89.7
Yes	79.8	9.9	-89.7

Per il Titanic

Tabella osservata

	1st	2nd	3rd	Sum
No	122	167	528	817
Yes	203	118	178	499
Sum	325	285	706	1316

Tabella teorica

	1st	2nd	3rd	Sum
No	201.8	176.9	438.3	817.0
Yes	123.2	108.1	267.7	499.0
Sum	325.0	285.0	706.0	1316.0

Si calcolano dunque le differenze

$n_{ij} - \hat{n}_{ij}$

	1st	2nd	3rd
No	-79.8	-9.9	89.7
Yes	79.8	9.9	-89.7

le si elevano al quadrato

$(n_{ij} - \hat{n}_{ij})^2$

	1st	2nd	3rd
No	6362.7	98.7	8046.2
Yes	6362.7	98.7	8046.2

Per il Titanic

Tabella osservata

	1st	2nd	3rd	Sum
No	122	167	528	817
Yes	203	118	178	499
Sum	325	285	706	1316

Si calcolano dunque le differenze

$n_{ij} - \hat{n}_{ij}$

le si elevano al quadrato

$(n_{ij} - \hat{n}_{ij})^2$

e si dividono per le frequenze teoriche

$(n_{ij} - \hat{n}_{ij})^2 / \hat{n}_{ij}$

Tabella teorica

	1st	2nd	3rd	Sum
No	201.8	176.9	438.3	817.0
Yes	123.2	108.1	267.7	499.0
Sum	325.0	285.0	706.0	1316.0

	1st	2nd	3rd
No	-79.8	-9.9	89.7
Yes	79.8	9.9	-89.7

	1st	2nd	3rd
No	6362.7	98.7	8046.2
Yes	6362.7	98.7	8046.2

	1st	2nd	3rd
No	31.5	0.6	18.4
Yes	51.6	0.9	30.1

Per il Titanic

Tabella osservata

	1st	2nd	3rd	Sum
No	122	167	528	817
Yes	203	118	178	499
Sum	325	285	706	1316

Tabella teorica

	1st	2nd	3rd	Sum
No	201.8	176.9	438.3	817.0
Yes	123.2	108.1	267.7	499.0
Sum	325.0	285.0	706.0	1316.0

Si calcolano dunque le differenze

$n_{ij} - \hat{n}_{ij}$

	1st	2nd	3rd
No	-79.8	-9.9	89.7
Yes	79.8	9.9	-89.7

le si elevano al quadrato

$(n_{ij} - \hat{n}_{ij})^2$

	1st	2nd	3rd
No	6362.7	98.7	8046.2
Yes	6362.7	98.7	8046.2

e si dividono per le frequenze teoriche

$(n_{ij} - \hat{n}_{ij})^2 / \hat{n}_{ij}$

	1st	2nd	3rd
No	31.5	0.6	18.4
Yes	51.6	0.9	30.1

Il totale è pari a 133.05, si conclude che...