Cenni di Statistica Inferenziale

Teorema del limite centrale

Data una variabile, qualsiasi sia la sua distribuzione, la media di tutti i suoi campioni di ampiezza n ha una distribuzione normale:

$$\bar{x} = \mathrm{N}\left(\mu, \sigma/\sqrt{\mathrm{n}}\right)$$

dove: $\bar{x}$

- *è la media campionaria*
- *μ è la media della popolazione*
- *σ è la deviazione standard della popolazione*
- *n è la dimensione del campione*

Uso del teorema del limite centrale

Dall'equazione: $\bar{x} = \mathrm{N}\,(\mu, \sigma/\sqrt{\mathrm{n}}\,)$

e dal fatto che circa il 95% dei valori in una distribuzione normale cade al più a distanza di 2 deviazioni standard dalla media

emerge che

dato un singolo campione di n elementi con media $\bar{x}$

con confidenza 95% $\bar{x}$ cade nell'intervallo

$$[\mu - 2 \times {}^{\sigma}/_{\sqrt{n}}\,,\ \mu\ +\ 2 \times {}^{\sigma}/_{\sqrt{n}}\,]$$

ovvero

possiamo ritenere che la stima μ della media di una popolazione sia accettabile al 95% se la media $\bar{x}$ di un campione di dimensione n cade nell'intervallo:

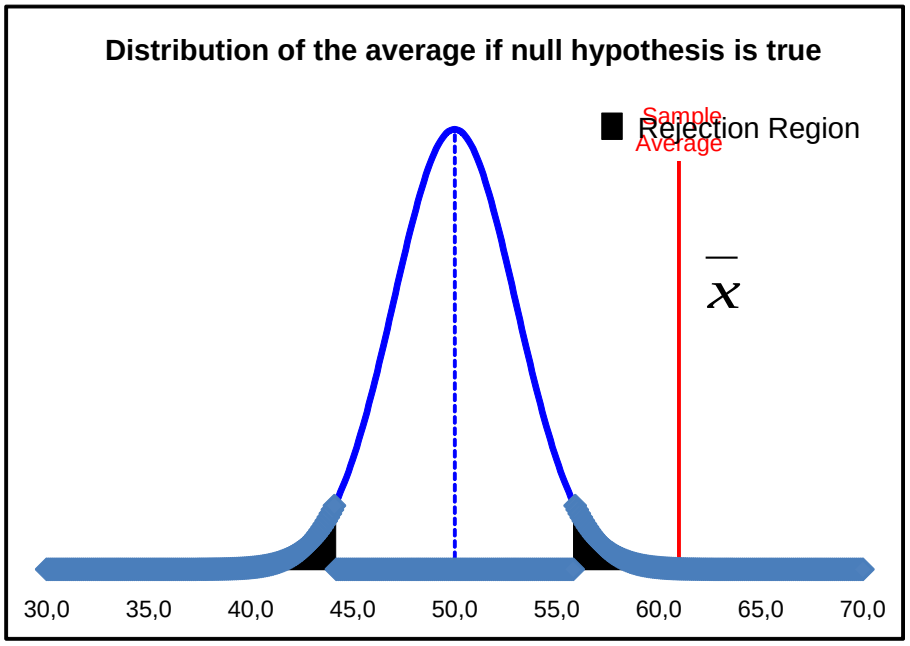
$$[\mu - 2 \times {}^{\sigma}/_{\sqrt{n}}\,,\ \mu\ +\ 2 \times {}^{\sigma}/_{\sqrt{n}}\,]$$

Esempio

- Ipotesi: la media della popolazione è $\mu = 50$
- Sappiamo che:
  - Deviazione standard della popolazione $\sigma = 15$
  - dimensione del campione N = 25
  - Media del campione $\bar{x} = 61$
- Intervallo di accettazione con confidenza al 95%

  $[50 - 2\times15/\sqrt{25},\ 50 + 2\times15/\sqrt{25}] = [44,56]$
- L'ipotesi sarebbe confermata se la media stimata cadesse nell'intervallo: in questo caso viene rifiutata

Graficamente



Conclusion: Reject the Null Hypothesis

Hypothesis Test

Ho: Mean = 50

- (●) Ha: Mean <> 50 (2-tailed)
- ( ) Ha: Mean > 50 (1-tailed)
- ( ) Ha: Mean < 50 (1-tailed)

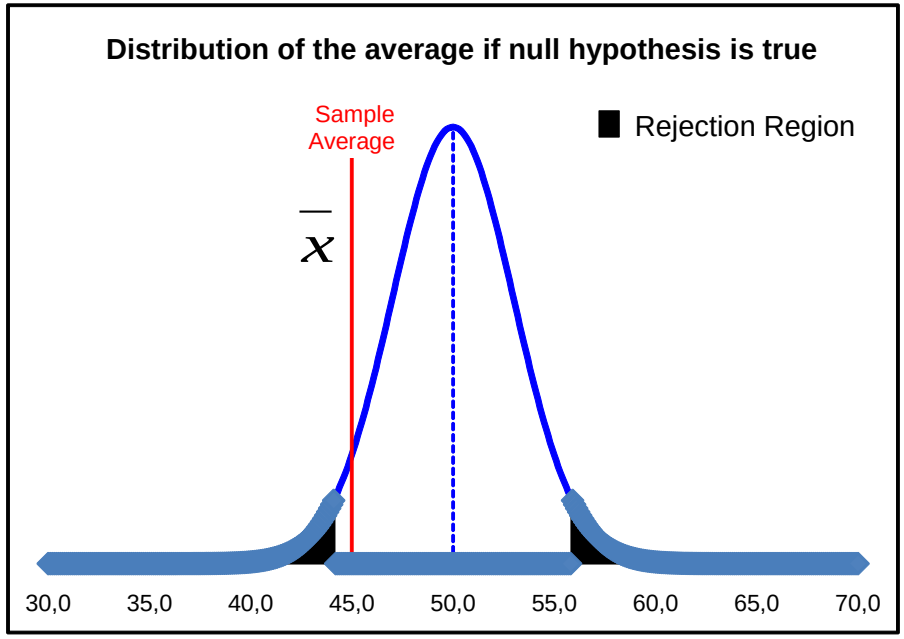
Population Mean:	50
Population Sigma:	15
Sample Average:	61
Sample Size:	25
Standard Error:	3,000
Alpha:	0,05
Lower Critical Value:	**44,120**
Upper Critical Value:	**55,880**
Z-Score:	3,667
P-value (2-sided):	**0,02%**

Esempio

- Ipotesi: la media della popolazione è $\mu = 50$
- Sappiamo che:
  - Deviazione standard della popolazione $\sigma = 15$
  - dimensione del campione N = 25
  - Media del campione $\bar{x} = 45$
- Intervallo di accettazione con confidenza al 95%

  $[50 - 2\times15/\sqrt{25},\ 50 + 2\times15/\sqrt{25}] = [44,56]$
- L'ipotesi sarebbe confermata se la media stimata cadesse nell'intervallo: in questo caso viene accettata

Graficamente



Conclusion: Do Not Reject the Null Hypothesis

Hypothesis Test

Ho: Mean = 50

- (●) Ha: Mean <> 50 (2-tailed)
- ( ) Ha: Mean > 50 (1-tailed)
- ( ) Ha: Mean < 50 (1-tailed)

Population Mean:	50
Population Sigma:	15
Sample Average:	45
Sample Size:	25
Standard Error:	3,000
Alpha:	0,05
Lower Critical Value:	**44,120**
Upper Critical Value:	**55,880**
Z-Score:	-1,667
P-value (2-sided):	**9,56%**

Riassumendo:

- Dati
  - una ipotesi μ sulla media di una popolazione
  - La deviazione standard σ della popolazione
  - un campione di dimensione n della popolazione
- Si può accettare l'ipotesi con confidenza 95% se la media del campione cade nell'intervallo

$$[\mu - 2 \times {}^{\sigma}/_{\sqrt{n}} \, , \, \mu \, + \, 2 \times {}^{\sigma}/_{\sqrt{n}} \,]$$

- o equivalentemente se μ cade nell'intervallo

$$[\bar{x} - 2 \times {}^{\sigma}/_{\sqrt{n}} \, , \, \bar{x} \, + \, 2 \times {}^{\sigma}/_{\sqrt{n}} \,]$$

Ovvero

- Dati
  - un campione di dimensione n con media $\bar{x}$ di una popolazione
  - La deviazione standard σ della popolazione

- Possiamo stimare la media della popolazione con confidenza al 95% come
  - $\bar{x} \pm 2 \times {}^{\sigma}/_{\sqrt{n}}$
- E' evidente che la stima è tanto più precisa quanto maggiore è n, che determina l'ampiezza dell'intervallo (la famosa forchetta delle previsioni)

Z-values

- Dall'equazione $\bar{x} = N(\mu, \sigma\sqrt{n})$, sottraendo μ e dividendo per $\sigma / \sqrt{n}$
- Otteniamo $\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = N(0,1)$
- Il valore $\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ è quindi distribuito normalmente con media 0 e deviazione standard 1
- Uno z-value, scritto z_p, è il punto z tale che, rispetto alla gaussiana standard la probabilità che che un valore sia minore o uguale a z_p è proprio p
- Per esempio $z_{0,95}$ è 1,96 poichè il 95% dei valori su una gaussiana standard sono minori di 1,96

Intervalli di confidenza

- La definizione di z-values, considerando anche il fatto che siamo interessati a escludere i valori ai due estremi ci porta a questa equazione, in cui α è il livello di significatività:

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

- Per esempio se α=0,05 allora $z_{1-0.05/2} = 1.96$ (circa 2 come assunto in precedenza) e quindi i limiti dell'intervallo di confidenza al 95% sono:

$$\left[\bar{x} - 1{,}96 \times {}^{\sigma}/_{\sqrt{n}}\ ,\ \bar{x} + 1{,}96 \times {}^{\sigma}/_{\sqrt{n}}\right]$$

Tabella degli intervalli di confidenza

1-α	Intervallo di confidenza
0,800	$\bar{x} \pm 1.282 \times \sigma / \sqrt{n}$
0,900	$\bar{x} \pm 1.645 \times \sigma / \sqrt{n}$
0,950	$\bar{x} \pm 1.960 \times \sigma / \sqrt{n}$
0,990	$\bar{x} \pm 2.576 \times \sigma / \sqrt{n}$
0,999	$\bar{x} \pm 3.290 \times \sigma / \sqrt{n}$

Testing di ipotesi

1. osserva il fenomeno
2. **ripeti**
    1. formula una teoria
    2. raccogli i dati
    3. analizza i dati

    finche' l'analisi conferma la teoria

Gli ingredienti del test

1. una ipotesi nulla H_0: rappresenta la teoria attuale del fenomeno sotto analisi

2. una ipotesi alternativa H_a: è una teoria alternativa che viene accettata nel caso che la ipotesi nulla venga respinta.

   Spesso l'ipotesi alternativa è l'ipotesi che si vuole accettare.

   In un esempio medico, nello studio degli effetti di un nuovo farmaco, l'ipotesi nulla può essere che il nuovo farmaco non abbia miglior comportamento rispetto alla terapia standard

3. Una statistica di controllo (test statistic): è calcolata sui dati per decidere se accettare o respingere l'ipotesi nulla

4. una regione di rifiuto: specifica l'insieme dei valori del test statistico per cui rifiutare l'ipotesi nulla.

Un esempio

In una fabbrica di componenti elettronici l'attuale standard di produzione fa rilevare che in ogni partita prodotta il numero di componenti difettosi ha una media di 50 e una deviazione standard di 15.

Viene proposto un nuovo processo di produzione. Il test del processo consiste nel produrre un campione di 25 partite.

Il numero medio di difetti per partita risultante è 45.

Vale la pena passare al nuovo processo di produzione?

Il test dell'ipotesi

- H_0: il nuovo processo non introduce miglioramenti

- H_a: il nuovo processo introduce miglioramenti, ovvero la diminuzione di difetti non è dovuta a una fluttuazione random

ovvero

- H_0 : il numero medio di componenti difettosi con il nuovo processo è ancora 50
- H_a : il numero medio di componenti difettosi con il nuovo processo non è 50

Statistica di controllo (test statistic) e regione di rifiuto

- In questo caso la statistica di controllo è la media $\bar{X}$ del campione, ovvero 45
- Se vogliamo una confidenza del 95%, l'intervallo di confidenza, ovvero la regione di rifiuto dell'ipotesi alternativa è:

$$\left[50 - 1.96 \times 15/\sqrt{25}, 50 + 1.96 \times 15/\sqrt{25}\right]$$

=

[44.12,55.88]

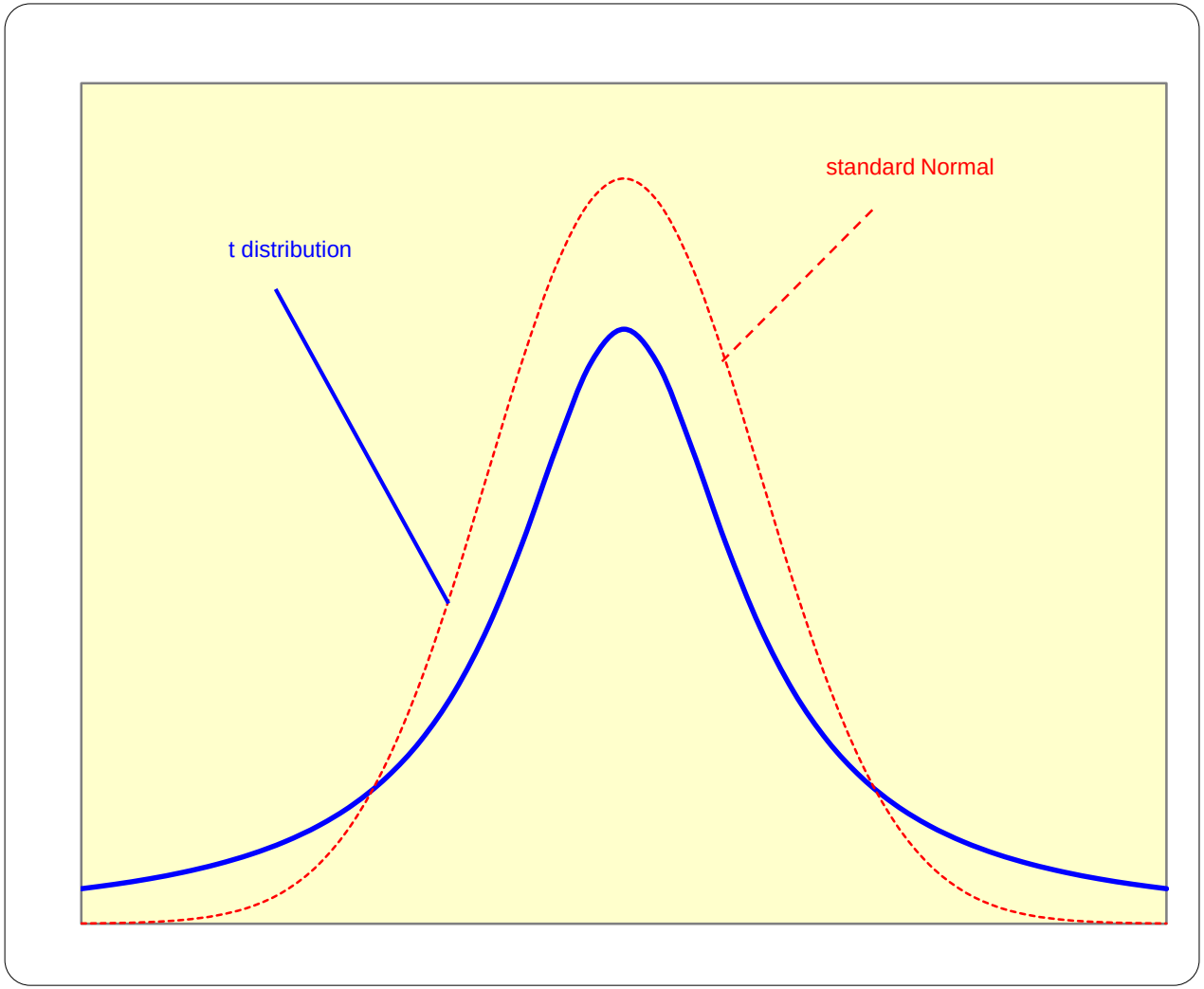
- Poiché 45 è incluso nella regione di rifiuto, l'ipotesi alternativa viene rifiutata, ovvero il nuovo metodo è rifiutato e viene accettata l'ipotesi nulla.

Distribuzione t (t-distribution)

- Abbiamo assunto che il valore di σ (la deviazione standard della popolazione) sia noto
- se non lo è (e in genere non lo è), si può usare invece il valore *s* della deviazione standard del campione (ovviamente noto)
- Gosset, impiegato alla birreria Guinness a Dublino, scoprì che il rapporto

$$\frac{\bar{x} - \mu}{s / \sqrt{n}}$$

in cui la deviazione standard del campione (s) sostituisce la deviazione standard della popolazione (σ) non ha una distribuzione normale, ma una leggermente diversa detta t-distribution.


standard Normal
t distribution

t-distribution

- La t-distribution è
  - *simmetrica,*
  - *centrata sullo 0*
  - Caratterizzata da un singolo parametro, detto *grado di libertà*, uguale alla dimensione del campione meno 1
- Per calcolare l'intervallo di confidenza usiamo ora i t-value, che dipendono sia dalla confidenza α richiesta, sia dal grado di libertà uguale a n-1, dove n è la dimensione del campione.
- L'intervallo di confidenza è:

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

- Definiamo t-value come $\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$

t-values

- In Excel il t-value è calcolato dalla funzione
  - TINV(p,df)
    - p è il p-value di una distribuzione a due code, per esempio 0,05
    - df è il grado di libertà
- Per esempio, con un campione di 25 elementi l'intervallo di rifiuto per una confidenza del 95% è:

$$\left[\bar{x} - 2.06\frac{s}{5}, \bar{x} + 2.06\frac{s}{5}\right]$$

Esempio

- L'amministrazione dell'università afferma che la spesa media di libri per anno a informatica umanistica è inferiore a €200.
- Un rappresentante degli studenti intervista 25 colleghi scelti casualmente e verifica che:
  - la spesa media del campione è di €220
  - La deviazione standard del campione è €50
- C'è sufficiente evidenza che l'amministrazione abbia sottostimato la spesa per libri?

Esempio (cont.)

- Ipotesi nulla H_0 : la spesa media per testi è €200
- Ipotesi alternativa H_a: la spesa media per testi è significativamente diversa da €200
- Costruiamo un intervallo di rifiuto al 95% ovvero:

$$\left[220 - t_{0.95/2,24}\frac{50}{\sqrt{25}}, 220 + t_{0.95/2,24}\frac{50}{\sqrt{25}}\right]$$

- In Excel possiamo calcolare $t_{0,95/2,24}$ con la funzione TINV(0.05,24)
- L'intervallo di rifiuto è quindi:

  [199.36,240,64]

- E quindi l'ipotesi alternativa è rifiutata e confermata l'ipotesi nulla.