

Le principali procedure inferenziali: nozioni, schemi di procedimento ed esempi di applicazione



IC al livello $(1-\alpha)$ % per la media μ

Ipotesi: nella popolazione il fenomeno si distribuisce secondo una Normale con media μ (incognita) e varianza σ^2 nota

Formula valida anche per il caso di una popolazione non Normale SE il campione è grande ($n \geq 30$)

Formula valida anche se non si conosce la varianza della popolazione σ^2 SE il campione è grande ($n \geq 30$) (al posto di σ^2 si usa la varianza calcolata nel campione)

Formula:
$$\left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

Quantile: $N(0,1)$

$1-\alpha$	$z_{\frac{\alpha}{2}}$
90%	1.64
95%	1.96

Nota: quando non si conosce la varianza σ^2 e il campione è piccolo si usano i quantili della distribuzione T di Student con $(n-1)$ gradi di libertà

T-test su una media al livello di significatività $\alpha=5\%$

$$H_0: \mu = \mu_0$$

Ipotesi: nella popolazione il fenomeno si distribuisce secondo una Normale con media μ (incognita) e varianza σ^2 nota

Formula valida anche per il caso di una popolazione non Normale SE il campione è grande ($n \geq 30$)

Formula valida anche se non si conosce la varianza della popolazione σ^2 SE il campione è grande ($n \geq 30$) (al posto di σ^2 si usa la varianza calcolata nel campione)

Statistica test: la media calcolata nel campione, standardizzata:

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

Calcolare il p-value sulla tavola della $N(0,1)$
oppure
Confrontare t con il limite della regione di rifiuto z:

H_1	z
$\mu > \mu_0$ ($\mu < \mu_0$)	+1.64 (-1.64)
$\mu \neq \mu_0$	± 1.96

Nota: quando non si conosce la varianza σ^2 e il campione è piccolo si usa la distribuzione T di Student con (n-1) gradi di libertà

T-test per confrontare 2 medie da campioni indipendenti al livello di significatività $\alpha=5\%$

$$H_0: \mu_1 = \mu_2 \text{ vs } H_1: \mu_1 \neq \mu_2$$

Ipotesi: i due gruppi provengono da due popolazioni rispettivamente con media μ_1 e μ_2 (incognite) e uguale varianza σ^2 incognita; entrambi i campioni sono grandi ($n_1, n_2 > 30$)

Calcolare:

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

Statistica test: la differenza delle medie calcolate nel campione, standardizzata:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Calcolare il p-value sulla tavola della $N(0,1)$
oppure
Confrontare t con il limite della regione di rifiuto z = 1.96

Confronto di gruppi indipendenti o appaiati

- I 2 gruppi sono **indipendenti** quando le unità statistiche del primo gruppo non coincidono con quelle del secondo gruppo, o comunque non condividono con esse alcun possibile fattore esplicativo con quelle; i due gruppi possono avere numerosità diverse
- I 2 gruppi sono **dipendenti o appaiati** quando si tratta degli stessi soggetti, oppure di soggetti (o unità statistiche, o osservazioni) che condividono un importante fattore esplicativo; i due gruppi hanno la stessa numerosità.
Esempi:
 - Misurazione del peso (dei WBC; della pressione; ...) prima e dopo la terapia
 - Confronto fra due gruppi di fratelli (condividono fattori genetici / ambientali)
 - Confronto fra organi o parti di organi dello stesso paziente (es i 2 occhi, 2 lembi di pelle, etc)
 - Osservazione dell'outcome in studi sperimentali con cross-over, o in studi osservazionali con matching individuale (vd. Epidemiologia)
- Esistono test specifici per dati appaiati

Metodi parametrici e non-parametrici

- I test "**parametrici**" valgono sotto certe ipotesi sulla distribuzione dei dati (ad es: Normalità) o nei grandi campioni.
- I test "**non-parametrici**" non hanno bisogno di ipotesi restrittive sulla distribuzione da cui provengono i dati. Essi sono preferibili quando le ipotesi non sono verificate o verificabili, e con piccoli campioni.
- Tuttavia, un test parametrico usato quando valgono le ipotesi ha maggiore potenza di un test non-parametrico. Ossia, i non-parametrici sono test più "conservativi": solo una forte evidenza porta al rifiuto dell'ipotesi nulla, anche se essa è falsa.

Confronto di medie fra gruppi indipendenti: altre situazioni

- Osservando k gruppi indipendenti di unità, si vuole verificare l'ipotesi che le medie delle popolazioni da cui provengono siano fra loro tutte uguali, contro l'ipotesi che siano diverse (ossia, che almeno 2 fra k medie siano diverse)
- I principali test di riferimento sono:

	<i>Parametrici</i>	<i>Non-parametrici</i>
<i>k=2 gruppi</i>	T-test per campioni indipendenti	Wilcoxon; Mann-Whitney
<i>k>2 gruppi</i>	ANOVA	Kruskal-Wallis

- T-test e ANOVA fanno **ipotesi di Normalità e Omoschedasticità** (uguale varianza nei gruppi). Esistono test e altri metodi per verificarle:
 - Normalità: Metodi grafici (es. Normal Probability Plot); Test di Kolmogorov-Smirnov o di Shapiro-Wilks
 - Omoschedasticità: Test F; test di Levene
- Lavorare su una trasformata dei dati può rendere le ipotesi valide (Es. se Y ha distribuzione asimmetrica a destra, logY può avere distribuzione simmetrica)

IC al livello $(1-\alpha)$ % e test al livello 5% per la prob. π

Ipotesi: il campione è grande ($n \geq 30$)

Formula IC95%:

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

Quantile: N(0,1)	$1-\alpha$	$z_{\frac{\alpha}{2}}$
	90%	1.64
	95%	1.96

Statistica test per $H_0: \pi = \pi_0$: la proporzione calcolata nel campione, standardizzata:

$$t = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}}$$

Calcolare il p-value sulla tavola della N(0,1)
oppure
Confrontare t con il limite della regione di rifiuto z:

H_1	z
$\pi > \pi_0$ ($\pi < \pi_0$)	+1.64 (-1.64)
$\pi \neq \pi_0$	± 1.96

Nota: in sostanza, si tratta la probabilità come la media di un campione, SE il campione è grande. Si può usare anche il T-test per confrontare 2 probabilità da campioni indipendenti e GRANDI, l'alternativa non-parametrica è il test Chi-Quadrato (e, per campioni piccoli, il test F di Fisher)

Test Chi-Quadrato per l'associazione fra due caratteri qualitativi (tabella doppia rxc)

H_0 : X (r modalità) e Y (c modalità) indipendenti vs H_1 : X,Y associati

Ipotesi: nella tabella doppia, tutte le frequenze attese sono >5

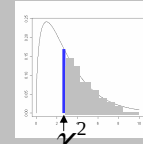
Calcolare le frequenze attese sotto l'ipotesi H_0 di indipendenza:

$$\tilde{n}_{ij} = \frac{n_{i.} \cdot n_{.j}}{n_{..}} \quad \text{ovvero, per ogni cella: } \text{Attesa} = \frac{\text{tot riga} \cdot \text{tot colonna}}{\text{tot generale}}$$

Statistica test:

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}} = \sum \frac{(\text{Osservate} - \text{Attese})^2}{\text{Attese}}$$

Calcolare il p-value sulla tavola del Chi-quadrato con gradi di libertà $(r-1)(c-1)$



oppure

Confrontare χ^2 con il limite della regione di rifiuto z : solo per tabella 2x2:

α	z
0.05	3.841
0.01	6.635

Studio dell'associazione fra 2 caratteri continui

- I principali metodi di riferimento sono:

	misura del grado di relazione e test associati	
relazione lineare	(Covarianza) Coefficiente di Correlazione Lineare di Pearson (r) Coefficiente di regressione (pendenza della retta di regressione)	$\text{cov}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x} \cdot \bar{y}$ $r_{xy} = \frac{\text{cov}_{xy}}{\text{std}_x \cdot \text{std}_y}$ $\hat{b} = r_{xy} \frac{\text{std}_y}{\text{std}_x} \quad \text{Intercetta: } \hat{a} = \bar{y} - \hat{b}\bar{x}$
relazione non-lineare	Indici Rho di Spearman o Tau di Kendall	

Es: IC 95% per la media

Si vuole stimare il valore medio della pressione sanguigna fra i soggetti sottoposti ad un certo trattamento farmacologico. Si dispone di un campione di 130 soggetti, con media campionaria pari a 160 e deviazione standard pari a 25.

Data l'ampiezza del campione, è possibile applicare la formula dell'intervallo di confidenza usando le tavole della Normale; scegliamo il livello 95%

$$\bar{x} = 160$$

$$s = 25, n = 130 \Rightarrow s.e.(\bar{X}) = 25 / \sqrt{130} = 2.19$$

$$z_{\frac{\alpha}{2}} = 1.96$$

$$\begin{aligned} 95\%CI &= (\bar{x} - 1.96 \cdot \sigma / \sqrt{n}, \bar{x} + 1.96 \cdot \sigma / \sqrt{n}) \\ &= (160 - 1.96 \cdot 2.19, 160 + 1.96 \cdot 2.19) \\ &= (155.7, 164.3) \end{aligned}$$

Es: test su una media. i) p-value

Assumendo che in una popolazione di soggetti sani il valore della pressione sanguigna si distribuisca secondo una distribuzione Normale con media 150 e deviazione standard pari a 15, e osservando un campione di 20 soggetti con media campionaria pari a 160, verificare l'ipotesi che tali soggetti siano stati estratti da quella popolazione, contro l'ipotesi che siano soggetti ipertesi.

Ipotesi e tipo di Test: $H_0 : \mu = 150$ vs $H_1 : \mu > 150$; T-test per la media

Ipotesi del test: la distribuzione è Normale e la varianza è nota: soddisfatte.

Calcolo della statistica test:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{160 - 150}{3.35} = 2.98$$

Metodo del p-value: $\Phi(2.98) = 0.999$ (sulle tavole)
 $p\text{-value} = (1 - 0.999) = 0.001$

Conclusione: Il valore osservato è significativamente diverso da 150: possiamo concludere che i soggetti osservati sono ipertesi rispetto alla norma.

Es: test su una media. ii) regione di rifiuto

Assumendo che in una popolazione di soggetti sani il valore della pressione sanguigna si distribuisca secondo una distribuzione Normale con media 150 e deviazione standard pari a 15, e osservando un campione di 20 soggetti con media campionaria pari a 160, verificare l'ipotesi che tali soggetti siano stati estratti da quella popolazione, contro l'ipotesi che siano soggetti ipertesi.

Ipotesi e tipo di Test: $H_0: \mu = 150$ vs $H_1: \mu > 150$; T-test per la media

Ipotesi del test: la distribuzione è Normale e la varianza è nota: soddisfatte.

Calcolo della statistica test:
$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{160 - 150}{3.35} = 2.98$$

Metodo delle regioni di rifiuto:

Fissiamo $\alpha=0.05$; trattandosi di un test a una coda, la zona è la coda destra, ed è delimitata dal valore 1.64 (proprietà della Normale: $\Phi(1.64) = 0.95$)

La statistica test vale 2.98: cade nella zona di rifiuto

Conclusione: ... (come prima)

Esempio: test su due medie

Due gruppi di soggetti assumono rispettivamente il trattamento A o B. Si misura il valore della pressione sanguigna, che si assume si distribuisca secondo una distribuzione Normale con stessa varianza. I risultati nei campioni sono in tabella. Verificare l'ipotesi che non ci sia differenza fra i due gruppi.

	gruppo	
	A	B
n	32	36
$\bar{y}$	94	92
s^2	18	16

Ipotesi e tipo di Test: $H_0: \mu_A = \mu_B$ vs $H_1: \mu_A \neq \mu_B$; T-test per 2 medie

Ipotesi del test: valgono Normalità e omoschedasticità (secondo il testo)

Calcolo della statistica test:

$$s = \sqrt{\frac{31 \cdot 18 + 35 \cdot 16}{32 + 36 - 2}} = 4.16 \quad t = \frac{94 - 92}{4.16 \sqrt{\frac{1}{32} + \frac{1}{36}}} = 1.98$$

Metodo delle regioni di rifiuto: Fissiamo $\alpha=0.05$; il test è a due code; il valore soglia è 1.96: La statistica test cade nella zona di rifiuto.

Conclusione: rigettiamo l'ipotesi di uguaglianza dei gruppi rispetto alla pressione.

Esempio: test su una probabilità

Si osserva un gruppo di 160 pazienti con leucemia acuta mieloide; 110 di essi ottengono la remissione completa dopo un ciclo di una chemioterapia sperimentale. Dunque la percentuale di successi osservata è pari al 69%. Si può concludere che il trattamento è più efficace della chemioterapia standard, che ha una probabilità di successo del 60%?

Ipotesi e tipo di Test: $H_0: \pi = 0.6$ vs $H_1: \pi > 0.6$; test di tipo T

Ipotesi del test: il test è valido: il campione è grande ($n=160$)

Calcolo della statistica test:

$$\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = \sqrt{\frac{0.69 \cdot 0.31}{160}} = \sqrt{0.0013} = 0.036 \quad z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} = \frac{0.69 - 0.6}{0.036} = 2.50$$

Metodo del p-value: $\Phi(2.50) = 0.994$ (sulle tavole)
p - value = $(1 - 0.994) = 0.006$

Conclusione: Il valore osservato è significativamente maggiore di 60%, possiamo concludere che il nuovo trattamento è superiore al trattamento standard.

Esempio: test Chi-Quadrato

I seguenti dati sono relativi all'utilizzo di antibiotici registrato in un anno in due gruppi di uomini e donne. Si vuole verificare se esiste una differenza significativa nei due gruppi.

	antibiotici		
sex	No	Si	tot
M	35	20	55
F	54	8	74
tot	89	28	129

Ipotesi e tipo di Test: H_0 : sesso e antibiotici sono indipendenti, H_1 : sono associati; Test Chi-Quadrato

freq.	37.9	11.9
attese	51	16

Ipotesi del test: tutte le frequenze attese sono >5 : il test è valido.

Calcolo della statistica test: $(\dots) \chi^2 = 9.8$

Metodo delle regioni di rifiuto: Fissiamo $\alpha=0.05$; il test è sempre a una coda, a destra; il valore soglia in corrispondenza di 1 grado di libertà (letto sulle tavole) è 3.841. La statistica test cade nella zona di rifiuto.

Conclusione: rifiutiamo l'ipotesi di indipendenza.