

Statistica inferenziale- IV

STATISTICA NUMERICA

A.Y. 2023-2024

Outline

1) Intervalli di confidenza

STATISTICA NUMERICA, CAP. 6.2.5

9

2) Test di ipotesi

STATISTICA NUMERICA CAP. 6.3

11

1) Intervalli di confidenza

STATISTICA NUMERICA, CAP. 6.2.5

Intervalli di confidenza per la stima della media campionaria: popolazione normale

Siano X_1, X_2, \dots, X_n SRS(n) da una distribuzione normale con (media= μ , sd= σ).

Consideriamo come statistica campionaria la media \bar{X} .

Stimando con la media campionaria la media della distribuzione, si commette un errore non noto.

Fissato α in $[0,1]$ e $z_{\alpha/2}$ il quantile di indice $\alpha/2$ della distribuzione normale standard,

$$\mu=0 \quad \sigma=1$$

l'intervallo:

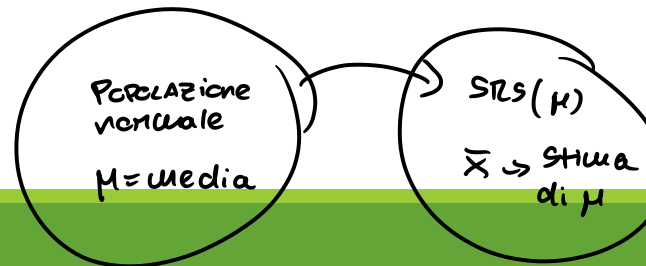
$$I_{\alpha} = \left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

È detto intervallo di confidenza $100(1-\alpha)\%$ di μ .

α è il livello di confidenza.

1) Conosco σ → deviazione standard della popolazione

2) Non conosco σ



Quantili → designano quanto "lontano" dalla stima puntuale possiamo andare per avere una certa possibilità che il parametro vero (es. μ) sia incluso nell'intervallo.

Un intervallo di confidenza (IC) fornisce un range di valori per un parametro della popolazione, con un certo livello di fiducia.

Intervalli di confidenza per la stima della media campionaria: popolazione normale

Possiamo anche dire che la probabilità che il valore esatto, μ sia nell'intervallo di confidenza è' del $100(1 - \alpha)\%$.

L'intervallo è casuale perché dipende dalla variabile aleatoria media campionaria

Possiamo fare le seguenti considerazioni.

- Per un livello di confidenza fissato $1 - \alpha$, se n aumenta, l'intervallo di confidenza diminuisce.
- Per n fissato, se $1 - \alpha$ aumenta, l'intervallo di confidenza aumenta.

Esempio 6.11 di Statistica Numerica.

$$\alpha = 0,05 \rightarrow 100(1 - \alpha)\% = 95\%$$

↳ Siamo confidenti al 95% che l'intervallo contenga la vera media della popolazione

Intervalli di confidenza per la stima della media campionaria: popolazione normale

Se la deviazione standard della popolazione non e' nota?

Si utilizza la deviazione standard campionaria:

$$\bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$$

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

deviazione
standard
campionaria

\bar{X}

Intervalli di confidenza per la stima della media campionaria: popolazione normale

Cosa posso dire di questa approssimazione?

- SE n e' grande S tende alla deviazione standard esatta
- Se n e' piccolo, sostituisco il valore del quantile della distribuzione normale con quello della distribuzione t di student (df=1):

Non conosco σ (caso b)

$n \geq 40$

stats.norm. pp%

$n < 40$

stats.t. pp%

$$\bar{X} \pm t_{\alpha/2}(\cancel{df=1}) \frac{S}{\sqrt{n}}.$$

$$df = n - 1$$

↓
gradi di libert 

\bar{X}

Intervalli di confidenza per la stima della media campionaria: popolazione **non** normale

Come caso b

Siano X_1, X_2, \dots, X_n SRS(n) da una distribuzione **qualunque** con (media= μ , sd= σ).

Consideriamo come statistica campionaria la media \bar{X} .

Per il Teorema del Limite Centrale, se n è sufficientemente grande la variabile aleatoria media campionaria si comporta come nel caso di campioni estratti da distribuzione normale.

Quindi l'intervallo di confidenza si calcola sempre come:

$$\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

1
 $n \geq 40$
Stats. norm. pp8

$n < 40$
Stats. t. pp8

→

intervallo di confidenza

$$\bar{X} + t_{\alpha/2} \frac{S}{\sqrt{n}}$$

↓
 $t_{\alpha/2} \rightarrow$ quantile di ordine $\alpha/2$ della distribuzione t-di student ($df = n-1$)

Intervalli di confidenza per la stima della media campionaria: popolazione **non** normale

Purtroppo è molto difficile nel caso di distribuzioni **non normali** avere la deviazione standard esatta quindi la sostituisco con quella campionaria S .

I

$$\bar{X} \pm z_{\alpha/2} \frac{S}{\sqrt{n}}$$

Funzioni python

Funzione Python per calcolare il quantile di ordine p della distribuzione normale standard:

```
from scipy import stats  
Q=stats.norm.ppf(p)
```

Nel caso di distribuzione normale con media= m e deviazione standard= s :

```
Q=stats.norm.ppf(p, loc=m,scale=s)  
Q=stats.norm.ppf(0.95, loc=0,scale=1.5)
```

Nel caso di distribuzione t di student con gradi di libertà n :

```
Q=stats.t.ppf(p, df)  
Q=stats.t.ppf(0.99, 4)
```

Zuccheri ≥ 37 . $(0,3)$

Test di ipotesi

- IPOTESI nulla H_0
- IPOTESI alternativa $H_a \rightarrow$ Contraddice l'ipotesi nulla

$H_0: \mu = \mu_0$ $(0,3)$ ipotesi nulla

$H_a: \mu > \mu_0$ ipotesi alternativa

Risultato $\rightarrow H_0$ rigettata (in favore di H_a)

$\rightarrow H_0$ non è rigettata

Se H_0 non è rigettata non significa che $\mu = \mu_0$ ma significa che non c'è sufficientemente evidenza che H_a sia vera, cioè $\mu > \mu_0$

2) Test di ipotesi

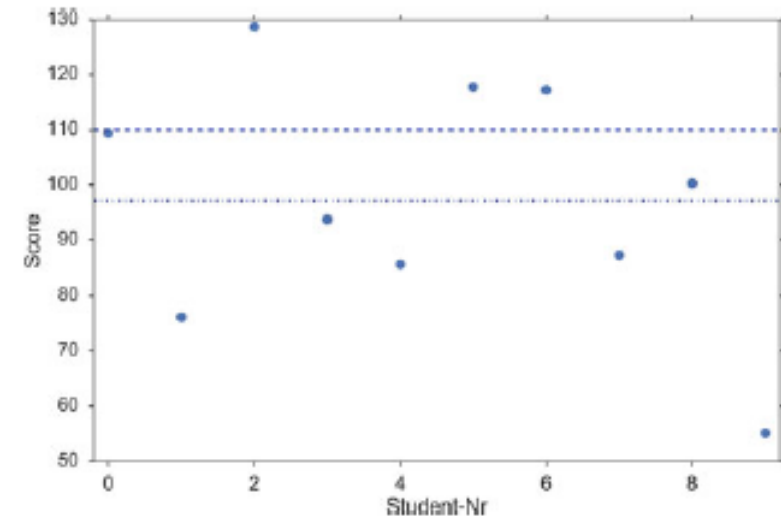
STATISTICA NUMERICA CAP. 6.3

Test di ipotesi

Supponiamo che tu sia il direttore di una scuola. Se gli studenti ottengono un punteggio di ¹⁰⁵~~110~~ nell'esame finale, visto che la media nazionale è di 100, ottieni un incentivo.

Se il voto è significativamente minore non lo ottieni (devi assumere più insegnanti), ma anche se è significativamente maggiore non lo ottieni (hai speso troppo e quindi devi licenziare delle insegnanti).

Come fare a decidere?



Test di ipotesi: procedura generale

Step:

1. Considero un SRS(n) (*il campione dei voti considerato*) 109

2. Formulo una ipotesi nulla H_0 (*La media dei voti è 110*)

3. Formulo una ipotesi alternativa H_1 (*la media dei voti è diversa da 110*)

4. Calcolo una statistica (*la media campionaria*)

5. Confronto il valore della statistica calcolata con quello della ipotesi nulla e calcolo un valore detto **p-value**

6. Interpreto il p-value e decido se l'ipotesi nulla è da rigettare oppure no

$$H_0 = \mu = 105 (\mu_0)$$

$$H_a = \mu \neq 105$$

funzione
python

→ Stimula la media μ della popolazione
(metodo MLE, oppure con la media campionaria)

Test di ipotesi: interpretazione del p-value

Il p-value è un valore p nell'intervallo $[0,1]$.

Se $p < 0.05$ possiamo interpretare così:

↗ SE l'ipotesi nulla è vera, la probabilità di trovare un valore della media campionaria più estremo (maggiore o minore) di quello osservato è del 5%.

N.B. Non significa che l'ipotesi nulla sia falsa e nemmeno che l'ipotesi alternativa sia vera!

Se H_0 fosse vera, la probabilità di trovare $\bar{x} \neq 105$ è $< 5\%$.

Test di ipotesi: interpretazione del p-value

In pratica però di solito il p-value si utilizza così per decidere il test di ipotesi:

Se $p < 0.05$ si rigetta l'ipotesi nulla a favore di quella alternativa, per una differenza statisticamente significativa del valore osservato rispetto all'ipotesi nulla.

Ricordiamo però che il p-value rappresenta solo una **verosimiglianza che l'ipotesi nulla sia vera, niente di più!**

RESULTATO TEST
DI IPOTESI

→ $p < 0,05 \Rightarrow H_0$ rigettata

→ $p \geq 0,05 \Rightarrow H_0$ non rigettata

Test di ipotesi: tipi di errore

producer risk errors

H_0 è vera \rightarrow ma la statistica dice no (H_0 sarebbe vera)

Errori del primo tipo (Type I errors):

Sono gli errori che si commettono quando l'ipotesi nulla è vera anche se la statistica misurata differisce significativamente.

Sono detti producer risk errors perché si rigettano dati che in realtà rispondono ai requisiti richiesti.

Errori del secondo tipo (Type II errors):

Sono gli errori che si commettono quando si accetta l'ipotesi nulla perché la statistica non differisce da essa significativamente anche se l'ipotesi nulla è falsa.

Sono detti consumer risk errors perché si accettano misure anche se non conformi ai requisiti richiesti.

Consumer risk errors

H_0 accettata \rightarrow la statistica converge

ma H_0 in realtà è falsa

problema \rightarrow si accettano misure non corrette.

Test di ipotesi: sensitività e specificità

Sopponiamo di fare un test di ipotesi utilizzando una grandezza che identifica l'esistenza o meno di un tumore. Quindi:

Ipotesi nulla: il paziente ha il tumore

Ipotesi alternativa: il paziente NON ha il tumore.

Sensitività: anche detta *potenza*. Proporzione di **positivi correttamente identificati da un test**.

Specificità: Proporzione di **negativi correttamente identificati da un test**.

Positive Predicted Value (PPV): proporzione di pazienti con test **positivo** che sono stati **correttamente diagnosticati**

Negative Predicted Value (NPV): proporzione di pazienti con test **negativo** che sono stati **correttamente diagnosticati**

Test di ipotesi: sensibilità e specificità

		Condition		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	True Positive	False Positive (Type I error)	Positive predictive value = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	False Negative (Type II error)	True Negative	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$ Sensitivita'	Specificity = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$ Specificita'	

PPV → tra tutti i positivi che ha detto quanti erano veri

Sensitivita' → Tra tutti i positivi della condizione quanti ne ha azzeccati

Test di ipotesi: sensitività e specificità

Mentre la sensitività e la specificità caratterizzano un test, non indicano quale sensitività e specificità caratterizza un test, non indicano quale porzione di pazienti con test “anormale” sono veramente “non normali”. Questa informazione è data dai valori PPV e NPV.

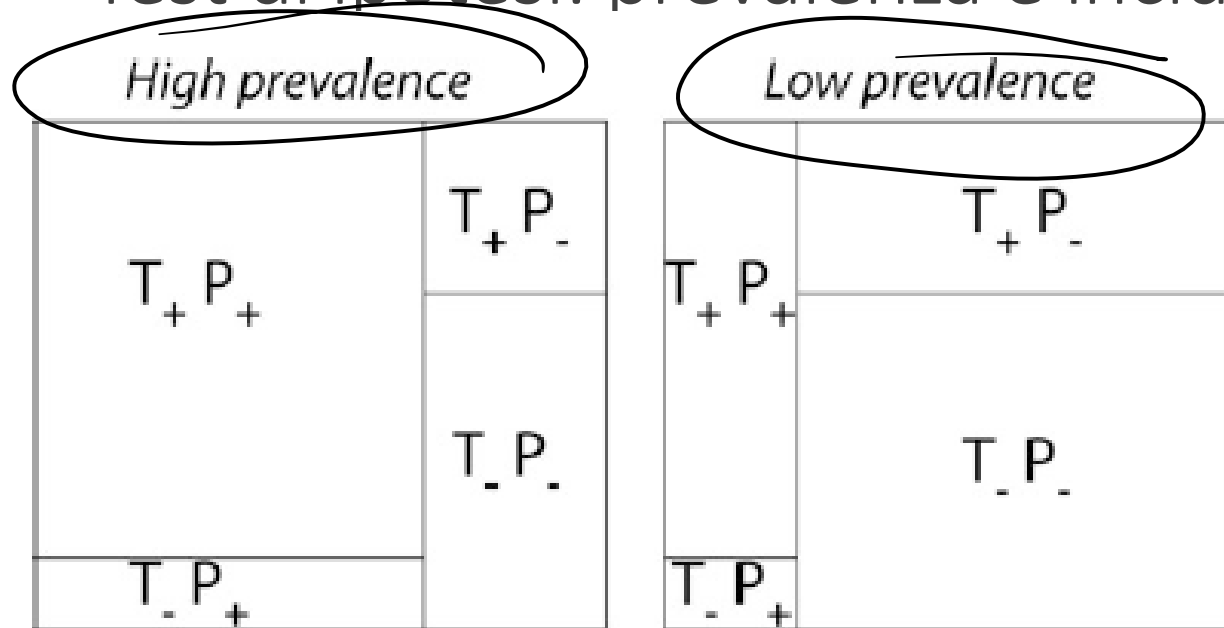
Tuttavia, questi valori da soli non sono sufficienti per una valutazione affidabile.

Per rispondere a questa domanda:

Se un paziente ha un test positivo, qual è la probabilità che sia veramente ammalato?

Dobbiamo considerare altri due fattori: la prevalenza e l'incidenza.

Test di ipotesi: prevalenza e incidenza



Prevalenza: quante persone su 100000(numero fissato) è ammalata

incidenza: il numero di nuovi casi diagnosticati su 100000 persone.

Fig. 7.8 Effect of prevalence on PPV and NPV. “T” stands for “test,” and “P” for “patient.” (For comparison with below: T+P+ = TP, T-P- = TN, T+P- = FP, and T-P+ = FN.)

Test di ipotesi: risultati di un esempio

1.5b

/ hypothesis tests

		Condition			
		Condition Positive	Condition Negative		
Test Outcome	Test Outcome Positive	True Positive (TP) = 25	False Positive (FP) = 175	Positive predictive value = = TP / (TP+FP) = 25 / (25+175) = 12.5%	$\frac{TP}{TP+FP}$
	Test Outcome Negative	False Negative (FN) = 10	True Negative (TN) = 2000	Negative predictive value = = TN / (FN+TN) = 2000 / (10+2000) = 99.5%	$\frac{TN}{TN+FN}$
		Sensitivity = = TP / (TP+FN) = 25 / (25+10) = 71%	Specificity = = TN / (FP+TN) = 2000 / (175+2000) = 92%		
		$\frac{TP}{TP+FN}$	$\frac{TN}{TN+FP}$		

Test di ipotesi: risultati di un esempio

Errore di tipo 1 \rightarrow Nel nostro test, Rifiuta l'ipotesi nulla (H_0) quando in realtà H_0 è vera

Errore di tipo 2 \rightarrow Nel nostro test, Non rifiuta l'ipotesi nulla (H_0) quando in realtà è falsa (H_1 è vera)

- False positive rate (α) = type I error = $1 - \text{specificity} = \frac{FP}{FP+TN} = \frac{175}{175+2000} = 8\%$
- False negative rate (β) = type II error = $1 - \text{sensitivity} = \frac{FN}{TP+FN} = \frac{10}{25+10} = 29\%$
- Power = sensitivity = $1 - \beta$
- positive likelihood ratio = $\frac{\text{sensitivity}}{1 - \text{specificity}} = \frac{71\%}{1 - 92\%} = 8.9 \rightarrow$ Un LR+ di 8.9 significa che un test positivo è quasi 9 volte più probabile in una persona malata che in una persona sana.
- negative likelihood ratio = $\frac{1 - \text{sensitivity}}{\text{specificity}} = \frac{1 - 71\%}{92\%} = 0.32 \rightarrow$ Un LR- di 0.32 significa che un test negativo è circa $1/3$ meno probabile in una persona malata che in una persona sana.

Funzioni Python

- Test di ipotesi sulla media di un campione di dati distribuiti in modo normale: t-test

```
t, pVal = scipy.stats.ttest_1samp(a,  
popmean, axis=0, nan_policy='propagate', alternative='two-sided', *, keepdims=False)
```

[scipy.stats.ttest_1samp — SciPy v1.10.1 Manual](#)

Funzioni Python

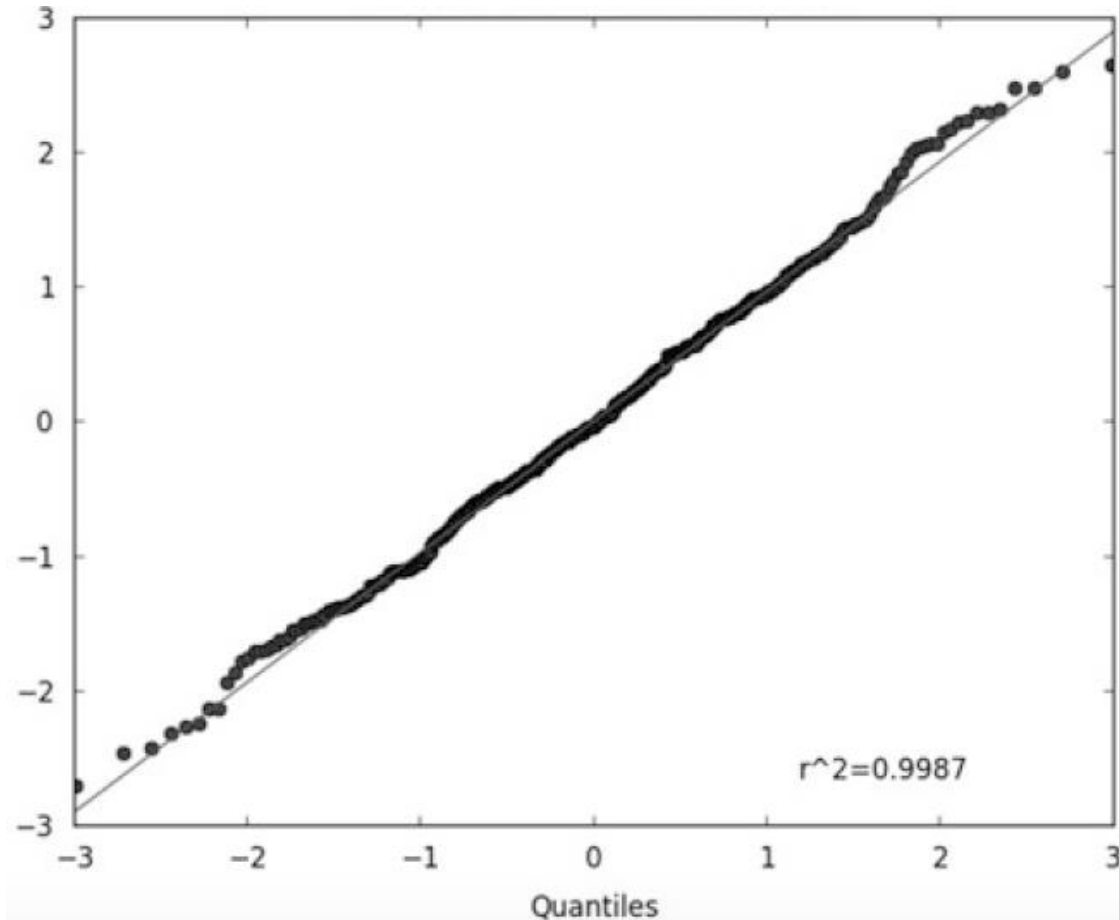
- Test di ipotesi per decidere se due campioni x e y provengono dalla stessa distribuzione:

```
Res=scipy.stats.wilcoxon(x, y=None, zero_method='wilcox', correction=False,  
alternative='two-sided', method='auto', *, axis=0, nan_policy='propagate', keepdims=False)
```

Res.statistic

Res.pvalue

[scipy.stats.wilcoxon — SciPy v1.10.1 Manual](#)



Test di normalità

QQ-plot: Il quantile del data set considerato è plottato rispetto al quantile di una distribuzione (normale in questo caso) di riferimento.

Se le due distribuzioni sono simili, i punti devono stare molto vicini alla retta.

Test di ipotesi di normalità

Ci sono diversi test di ipotesi di normalità basati sul confronto della distribuzione stimata dei dati rispetto alla distribuzione normale.

Uno dei più famosi è il **test di Shapiro-Wilk**, che si basa sulla matrice di covarianza delle statistiche ordinate delle osservazioni e può essere utilizzato anche con un numero ridotto (< 50) di osservazioni.

H_0 : residui normali

H_a : residui non normali.