Statistica - Introduzione a R

a cura di Antonio Iovanella

iovanella@disp.uniroma2.it
http://www.disp.uniroma2.it/Users/iovanella

Verifica di ipotesi

Introduzione

Idea di base

Supponiamo di avere un'idea del valore (incognito) di una media di un campione, magari attraverso la media campionaria.

Vogliamo trovare una strategia che, sulla base dei dati osservati, ci permetta di confermare o smentire la supposizione iniziale.

L'obbiettivo è quindi quello di sottoporre a verifica una ipotesi posta.

Introduzione

Verifica di ipotesi sulla media

Sia $X_1, X_2, \ldots, X_n$ un campione i.i.d. di variabili casuali gaussiane di media μ (incognita) e varianza (nota) σ^2.

Ci proponiamo di sottoporre a verifica l'ipotesi statistica che il vero valore della media sia μ_0.

Chiamiamo ***ipotesi nulla*** l'ipotesi da verificare e la indichiamo con $\boldsymbol{H_0}$.

In pratica, il nostro test di verifica deve affermare che la probabilità che la distanza $|\, \overline{X}_n - \mu_0 \,|$ non sia troppo elevata.

Introduzione

Alternative

Alla fine del test abbiamo due alternative:

- o rifiutiamo l'ipotesi nulla H_0
- o non rifiutiamo l'ipotesi nulla H_0

Se rifiutiamo l'ipotesi nulla quando questa è falsa, allora non commettiamo errori.

Se non rifiutiamo l'ipotesi nulla quando questa è vera, allora non commettiamo errori.

Ci sono però altre due situazioni:

Introduzione

Alternative

Se rifiutiamo l'ipotesi nulla quando questa è vera, allora commettiamo un errore:

Errore di I^0 tipo

Se non rifiutiamo l'ipotesi nulla quando questa è falsa, allora commettiamo errore.

Errore di II^0 tipo

La probabilità di commettere errori del I^0 tipo viene indicata con α, mentre la probabilità di commettere errori del II^0 tipo viene indicata con β.

Introduzione

Riassumendo

	Rifiuto H_0	**Non rifiuto H_0**
H_0 è vera	Errore di I^0 tipo, α	Nessun errore, $1 - \alpha$
H_0 è falsa	Nessun errore, $1 - \beta$	Errore di II^0 tipo, β

Introduzione

Livello di significatività

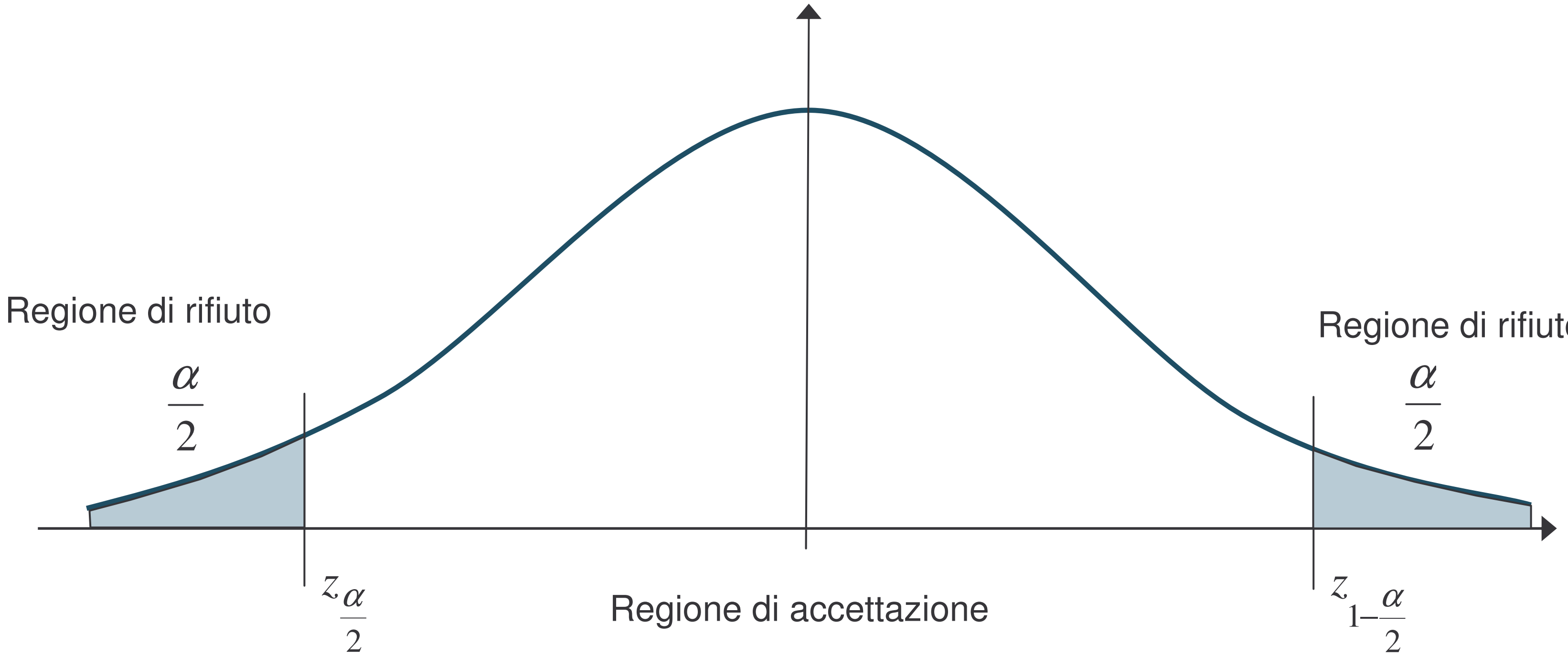
Si definisce *livello di significatività* α quel valore tale per cui l'ipotesi nulla H_0 viene rifiutata quando questa è vera con probabilità α.

Se la regola di rifiuto è $|\overline{X}_n - \mu_0| > k$, allora:

$$P(|\overline{X}_n - \mu_0| > k \mid H_0) = \alpha$$

Il valore $k = k_\alpha$ è chiamato *valore soglia* del test.

Introduzione



$$\alpha = P\left(Z < -\frac{k}{\frac{\sigma}{\sqrt{n}}} \quad e \quad Z > \frac{k}{\frac{\sigma}{\sqrt{n}}} \right) = P\left(Z < z_{\frac{\alpha}{2}} \quad e \quad Z > z_{1-\frac{\alpha}{2}} \right)$$

Introduzione

Statistica test e Test a due code

Indicando con *Z* la statistica test

$$Z = \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Se per un dato campione calcoliamo un valore *z* di *Z*:

$$z = \frac{\overline{x}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

il test ci dice di rifiutare l'ipotesi nulla $H_0 : \mu = \mu_0$ in favore dell'ipotesi $H_1 : \mu \neq \mu_0$ se z cade all'esterno della regione di accettazione.

Introduzione

Caso di varianza non nota

Se siamo in presenza del caso più realistico di varianza non nota, sostituiamo a σ^2 la sua stima e la statistica test diventa una *t* di Student:

$$t = \frac{\overline{X}_n - \mu_0}{\sqrt{\frac{S_n^2}{n}}} \sim t^{n-1}$$

e rifiuto H_0 se:

$$|\, t \,| > t_{1-\frac{\alpha}{2}}^{n-1}$$

Introduzione

Caso di varianza non nota con R

In R abbiamo gia visto **`t.test`** che effettua il test t di Student ad un campione, a due campioni e per dati appaiati.
La sintassi completa di questa funzione è:

```
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95, ...)
```

dove:
`x` , **`y`** : sono vettori numerici di dati;

`alternative`: specifica l'ipotesi alternativa, a seconda che si tratti di una verifica di ipotesi bilaterale o unilaterale;

Introduzione

Caso di varianza non nota con R

e, ancora:

`mu`: un numero che indica il valore reale della media (o la differenza tra le medie se si sta effettuando un test a due campioni);

`paired`: una variabile logica che indica se si vuole effettuare un test *t* per dati appaiati;

`var.equal`: una variabile logica che indica se porre le varianze dei due campioni uguali fra loro;

`conf.level`: livello di confidenza dell'intervallo.

##ESEMPIO

Esempio

Un gruppo di 22 volontari presso un centro di ricerca medica viene esposto a vari tipi di virus influenzali e tenuto sotto controllo medico.
Ad un campione casuale di 10 volontari viene somministrato un grammo di vitamina C quattro volte al giorno. Agli altri 12 volontari viene somministrato un placebo non distinguibile dal farmaco.
I volontari vengono poi visitati spesso da un medico che non conosce la divisione in gruppi e non appena uno di essi viene trovato guarito si registra la durata della malattia. Alla fine dell'esperimento si possiedono i seguenti dati:

Vitamina C	5.5, 6.0, 7.0, 6.0, 7.5, 6.0, 7.5, 5.5, 7.0, 6.5
Placebo	6.5, 6.0, 8.5, 7.0, 6.5, 8.0, 7.5, 6.5, 7.5, 6.0, 8.5, 7.0

##ESEMPIO

Esempio (continua)

Si può concludere che l'assunzione di 4 grammi di vitamina C al giorno abbia accorciato il decorso medio della malattia? A che livello di significatività?

Per prima cosa inseriamo i valori in R:

```
vitamina_C <- c(5.5, 6.0, 7.0, 6.0, 7.5, 6.0, 7.5, 5.5,
7.0, 6.5)
placebo <- c(6.5, 6.0, 8.5, 7.0, 6.5, 8.0, 7.5, 6.5,
7.5, 6.0, 8.5, 7.0)
```

##ESEMPIO

Esempio (soluzione)

Per provare l'ipotesi fatta, la assumiamo come ipotesi alternativa e vediamo se rifiutiamo l'ipotesi nulla corrispondente al livello di significatività desiderato.

Eseguiamo quindi il seguente test:

$$H_0 : \mu_{vitamina\_C} \geq \mu_{placebo} \quad \text{contro} \quad H_1 : \mu_{vitamina\_C} < \mu_{placebo}$$

supponendo che le varianze della durata della malattia nei due casi siano uguali.

##ESEMPIO

Esempio (soluzione)

In R:

```
t.test(vitamina_C, placebo, alternative = "less",
paired = FALSE, var.equal = TRUE, conf.level = 0.95)

        Two Sample t-test

data:  vitamina_C and placebo
t = -1.8987, df = 20, p-value = 0.03606
alternative hypothesis: true difference in means is
less than 0
95 percent confidence interval:
        -Inf -0.06185013
sample estimates:
mean of x mean of y
    6.450     7.125
```

##ESEMPIO

Esempio (considerazioni)

Se la statistica test t si trova nella regione di rifiuto, allora $p > \alpha$, se invece t si trova nella regione di accettazione avremo $p > \alpha$.

La regola di decisione è quindi:

rifiutare H_o : se $p > \alpha$
non rifiutare H_o : se $p < \alpha$

Tale regola resta valida anche per il test a due code.

##ESEMPIO

Esempio (conclusioni)

L'ipotesi nulla viene accettata solo per livelli di significatività $\alpha < 0.036$ = p-value;

l'ipotesi nulla H_0 viene dunque rifiutata ad un livello di significatività del 5%:

quindi, a questo livello di significatività, i dati raccolti evidenziano un accorciamento del decorso dell'influenza somministrando vitamina C.

##ESERCIZIO

Esercizio

Considerando l'esempio precedente, risolvere il problema senza considerare la funzione **`t.test`**.

Vitamina C	5.5, 6.0, 7.0, 6.0, 7.5, 6.0, 7.5, 5.5, 7.0, 6.5
Placebo	6.5, 6.0, 8.5, 7.0, 6.5, 8.0, 7.5, 6.5, 7.5, 6.0, 8.5, 7.0