Inferenza Statistica

Lorenzo Camponovo
SUPSI DTI

Contenuti

1. Teorema del Limite Centrale

- Consideriamo una successione di variabili aleatorie X_n, $n \in \mathbb{N}$, con funzione di distribuzione F_n.

- Si dice che X_n converge in distribuzione verso la variabile aleatoria Z, cioè,
$$X_n \rightarrow_d Z,$$
se il seguente limite esiste,
$$\lim_{n \rightarrow \infty} F_n(x) = F(x),$$
per ogni $x \in \mathbb{R}$, dove F è la funzione di distribuzione di Z.

1. Teorema del Limite Centrale

- Valgono i seguenti teoremi.

- $X_n \to_d Z$ se e solo se per ogni funzione continua e limitata $g(x)$ vale,

$$\lim_{n\to\infty} E[g(X_n)] = E[g(Z)].$$

- Se $X_n \to_d Z$ e $h(x)$ è una funzione continua, allora,

$$h(X_n) \to_d h(Z).$$

1. Teorema del Limite Centrale

- Siano $X_1, \ldots, X_n$ delle variabili aleatorie i.i.d, con,

$$E[X_i] = \mu, \qquad Var(X_i) = \sigma^2.$$

- Sia $S_n = \sum_{i=1}^n X_i$ e $\bar{X}_n = \frac{1}{n} S_n$. Allora la distribuzione F_n della variabile aleatoria,

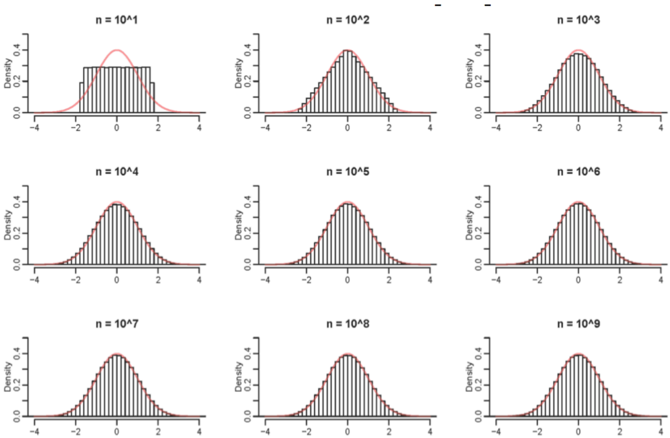
$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}},$$

converge per $n \to \infty$ verso la distribuzione normale standard,

$$F_n(Z_n) \to_d F_Z(Z).$$

1. Teorema del Limite Centrale

- Siano $X_1, \ldots, X_n$ distribuite secondo la distribuzione Uniforme $[0, 1]$,



1. Teorema del Limite Centrale

- Teorema di De Moivre e Laplace: Sia S_n una variabile aleatoria con distribuzione binomiale di parametri n e p.

- Allora la sua distribuzione converge per $n \to \infty$ verso la distribuzione normale con i seguenti momenti,

$$F_B(S_n; n, p) \to^d F_N(S_n; np, np(1-p)).$$

Analogamente la distribuzione di,

$$Z_n = \frac{S_n - np}{\sqrt{np(1-p)}} = \frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}},$$

converge per $n \to \infty$ verso la distribuzione normale standard,

$$F_n(Z_n) \to_d F_Z(Z).$$

1. Teorema del Limite Centrale

- Esempio: siano $X_1, \ldots, X_{12}$ delle variabili aleatorie i.i.d. con distribuzione uniforme nell'intervallo $[-0.5, 0.5]$.

- Sia $S_{12} = \sum_{i=1}^{12} X_i$.

- Utilizzando il teorema del limite centrale approssimare la seguente propabilità,
$$P(S_{12} > 1).$$

1. Teorema del Limite Centrale

- Esempio: lanciamo una moneta 100 volte, ottenendo testa esattamente 60 volte.

- Utilizzando il teorema del limite centrale approssimare la probabilità di ottenere almeno 60 volte testa.

- Pensate che la moneta sia truccata?

2. Introduzione: Esempio

- Sia $\mu = E[X]$ lo stipendio medio annuale (in migliaia di franchi) di una persona in CH.

- Dato un campione casuale $(X_1, \ldots, X_n)$ di dimensione n, possiamo utilizzare lo stimatore

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

  per stimare il parametro sconosciuto μ.

- Supponiamo che basandoci su dati passati crediamo che $\mu = 50$.

- Tuttavia, supponiamo di aver osservato nel nostro campione aleatorio $\hat{\mu}_n = 55$. Allora dovremmo rifiutare l'ipotesi $\mu = 50$? E se avessimo osservato $\hat{\mu}_n = 60$?

- Per rispondere a queste domande dobbiamo effettuare un test statistico.

2. Introduzione: Ipotesi

- Ipotesi nulla (l'ipotesi che vogliamo testare, denotata con H_0),

$$H_0 : \mu = 50.$$

- Ipotesi alternativa (l'ipotesi che contrasta quella nulla, denotata con H_1),

$$H_1 : \mu \neq 50.$$

2. Introduzione: Scenari

- 4 scenari possibili,

	H_0 is true	H_1 is true
decision: H_0	$1-\alpha$	β
decision: H_1	α	$1-\beta$

2. Introduzione: Errori del I e II tipo

- α è la significatività del test. È anche l'errore di primo tipo.
- β è l'errore di secondo tipo.
- $1 - \beta$ è la potenza del test.

2. Introduzione: Test Statistico

- Determinare il parametro di interesse θ e un suo stimatore $\hat{\theta}_n$.
- Determinare H_0, H_1 e la significatività α.
- Determinare p-values / valori critici.
- Decidere se rifiutare o meno H_0.

3. Test per la media

- Supponiamo che il parametro di interesse (e sconosciuto) sia il valore atteso di una data popolazione $\mu = E[X]$.

- Supponiamo invece che la varianza della popolazione sia conosciuta $\sigma^2 = Var(X) = 9$.

- Sulla base di dati precedenti, pensiamo che $\mu = 30$. Tuttavia, stimando il parametro μ attraverso un campione di 25 elementi otteniamo $\hat{\mu}_n = 29$.

- I dati ottenuti ci possono far pensare che $\mu < 30$. Un test della media ci permette di sciogliere questo dubbio.

3. Test per la media: p-value

- Ipotesi nulla: $H_0 : \mu = 30$.
- Ipotesi alternative: $H_1 : \mu < 30$.
- Livello di significatività: $\alpha = 0.05$.
- Test Statistica:

$$T = \frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}} = \frac{\hat{\mu}_n - 30}{3/5},$$

  Sotto l'ipotesi nulla, grazie al TLC, sappiamo che $T \sim N(0,1)$.

- Calcoliamo il p-value:

$$p - value = P\left(T < \frac{29 - 30}{3/5}\right) = 0.0475.$$

- Il p-value è inferiore al livello di significatività $\alpha = 0.05$, dunque rifiutiamo H_0.

3. Test per la media: valore critico

- Ipotesi nulla: $H_0 : \mu = 30$.
- Ipotesi alternative: $H_1 : \mu < 30$.
- Livello di significatività: $\alpha = 0.05$.
- Test Statistica:

$$T = \frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}} = \frac{\hat{\mu}_n - 30}{3/5},$$

  Sotto l'ipotesi nulla, grazie al TLC, sappiamo che $T \sim N(0,1)$.

- Dalla tabella $N(0,1)$ sappiamo che,

$$P(T < -1.64) = 0.05.$$

  Otteniamo il valore critico k risolvendo $\frac{k-30}{3/5} = -1.64$, da cui otteniamo: $k = 29.016$.

- Il valore osservato è minore al valore critico dunque rifiutiamo H_0.

3. Test per la media

- In questo esempio abbiamo osservato $\hat{\mu}_n = 29$.

- Chiaramente se avessimo osservato $\hat{\mu}_n = 30.5$ l'ipotesi alternativa sarebbe cambiata.

- Più precisamente avremmo ottenuto i seguenti test statistici.

3. Test per la media: p-value

- Ipotesi nulla: $H_0 : \mu = 30$.
- Ipotesi alternative: $H_1 : \mu > 30$.
- Livello di significatività: $\alpha = 0.05$.
- Test Statistica:
$$T = \frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}} = \frac{\hat{\mu}_n - 30}{3/5},$$
Sotto l'ipotesi nulla, grazie al TLC, sappiamo che $T \sim N(0, 1)$.
- Calcoliamo il p-value:
$$p - value = P\left(T > \frac{30.5 - 30}{3/5}\right) = 0.2033.$$
- Il p-value è superiore al livello di significatività $\alpha = 0.05$, dunque non rifiutiamo H_0.

3. Test per la media: valore critico

- Ipotesi nulla: $H_0 : \mu = 30$.
- Ipotesi alternative: $H_1 : \mu > 30$.
- Livello di significatività: $\alpha = 0.05$.
- Test Statistica:

$$T = \frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}} = \frac{\hat{\mu}_n - 30}{3/5},$$

  Sotto l'ipotesi nulla, grazie al TLC, sappiamo che $T \sim N(0, 1)$.

- Dalla tabella $N(0, 1)$ sappiamo che,

$$P(T > 1.64) = 0.05.$$

  Otteniamo il valore critico k risolvendo $\frac{k-30}{3/5} = 1.64$, da cui otteniamo: $k = 30.984$.

- Il valore critico è maggiore a quanto osservato dunque non rifiutiamo H_0.

3. Test per la media

- In questi esempi abbiamo considerato ipotesi alternative della forma,

$$H_1 : \mu < \mu_0, \text{ oppure } H_1 : \mu > \mu_0.$$

- Test di questa forma si dicono unilaterali.

- Tuttavia, avremmo potuto anche considerare come ipotesi alternativa,

$$H_1 : \mu \neq \mu_0.$$

- Test di questa forma si dicono bilaterali.

- Supponiamo di aver osservato $\hat{\mu}_n = 31$. Il test bilaterale si esegue nel seguente modo.

3. Test per la media: p-value

- Ipotesi nulla: $H_0 : \mu = 30$.
- Ipotesi alternative: $H_1 : \mu \neq 30$.
- Livello di significatività: $\alpha = 0.05$.
- Test Statistica:
$$T = \frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}} = \frac{\hat{\mu}_n - 30}{3/5},$$
Sotto l'ipotesi nulla, grazie al TLC, sappiamo che $T \sim N(0,1)$.
- Calcoliamo il p-value:
$$p - value = P\left(|T| > \left|\frac{31-30}{3/5}\right|\right) = 0.095.$$
- Il p-value è superiore al livello di significatività $\alpha = 0.05$, dunque non rifiutiamo H_0.

3. Test per la media: valore critico

- Ipotesi nulla: $H_0 : \mu = 30$.
- Ipotesi alternative: $H_1 : \mu \neq 30$.
- Livello di significatività: $\alpha = 0.05$.
- Test Statistica:
$$T = \frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}} = \frac{\hat{\mu}_n - 30}{3/5},$$
Sotto l'ipotesi nulla, grazie al TLC, sappiamo che $T \sim N(0,1)$.
- Dalla tabella $N(0,1)$ sappiamo che,
$$P(|T| > 1.96) = 0.05.$$
Otteniamo i valori critici k_1 e k_2 risolvendo $\frac{k_1 - 30}{3/5} = -1.96$ e $\frac{k_2 - 30}{3/5} = 1.96$ da cui otteniamo: $k_1 = 28.824$ e $k_2 = 31.176$.
- Il valore osservato è all'interno dei valori critici dunque non rifiutiamo H_0.

4. Test per differenza di medie

- Supponiamo ora di avere due popolazioni e due parametri sconosciuti $\mu_X = E[X]$ e $\mu_Y = E[Y]$.

- Supponiamo invece che le varianze delle due popolazioni siano conosciute $\sigma_X^2 = Var(X) = 9$ e $\sigma_Y^2 = Var(Y) = 25$.

- Sulla base di dati precedenti, pensiamo che $\mu_X = \mu_Y$. Tuttavia, stimando la differenza dei parametri sconosciuti su un campione di $n = 36$ elementi per la popolazione X e $m = 49$ per la popolazione Y otteniamo $\hat{\mu}_{X,n} - \hat{\mu}_{Y,m} = 1.5$.

- I dati ottenuti ci possono far pensare che $\mu_X > \mu_Y$. Un test per differenza di medie ci permette di sciogliere questo dubbio.

4. Test per differenza di medie: p-value

- Ipotesi nulla: $H_0 : \mu_X - \mu_Y = 0$.
- Ipotesi alternative: $H_1 : \mu_X - \mu_Y > 0$.
- Livello di significatività: $\alpha = 0.05$.
- Test Statistica:
$$T = \frac{\hat{\mu}_{X,n} - \hat{\mu}_{Y,m}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} = \frac{1.5}{\sqrt{\frac{9}{36} + \frac{25}{49}}},$$
Sotto l'ipotesi nulla, grazie al TLC, sappiamo che $T \sim N(0,1)$.
- Calcoliamo il p-value:
$$p - value = P\left(T > \frac{1.5}{\sqrt{\frac{9}{36} + \frac{25}{49}}}\right) = 0.0427.$$
- Il p-value è inferiore al livello di significatività $\alpha = 0.05$, dunque rifiutiamo H_0.

4. Test per differenza di medie: valore critico

- Ipotesi nulla: $H_0 : \mu_X - \mu_Y = 0$.
- Ipotesi alternative: $H_1 : \mu_X - \mu_Y > 0$.
- Livello di significatività: $\alpha = 0.05$.
- Test Statistica:

$$T = \frac{\hat{\mu}_{X,n} - \hat{\mu}_{Y,m}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}.$$

  Sotto l'ipotesi nulla, grazie al TLC, sappiamo che $T \sim N(0,1)$.

- Calcoliamo il valore critico k definito come,

$$P(T > k) = 0.05.$$

  Usando le tabelle otteniamo: $k = 1.4299$.

- Il valore critico è minore a quanto osservato dunque rifiutiamo H_0.

4. Test per differenza di medie

- In questo esempio abbiamo osservato $\hat{\mu}_{X,n} - \hat{\mu}_{Y,m} = 1.5$.

- Chiaramente se avessimo osservato $\hat{\mu}_{X,n} - \hat{\mu}_{Y,m} = -0.5$, l'ipotesi alternativa sarebbe cambiata.

- Più precisamente avremmo ottenuto i seguenti test statistici.

4. Test per differenza di medie: p-value

- Ipotesi nulla: $H_0 : \mu_X - \mu_Y = 0$.
- Ipotesi alternative: $H_1 : \mu_X - \mu_Y < 0$.
- Livello di significatività: $\alpha = 0.05$.
- Test Statistica:

$$T = \frac{\hat{\mu}_{X,n} - \hat{\mu}_{Y,m}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} = \frac{-0.5}{\sqrt{\frac{9}{36} + \frac{25}{49}}},$$

  Sotto l'ipotesi nulla, grazie al TLC, sappiamo che $T \sim N(0,1)$.

- Calcoliamo il p-value:

$$p - value = P\left(T < \frac{-0.5}{\sqrt{\frac{9}{36} + \frac{25}{49}}}\right) = 0.2843.$$

- Il p-value è superiore al livello di significatività $\alpha = 0.05$, dunque non rifiutiamo H_0.

4. Test per differenza di medie: valore critico

- Ipotesi nulla: $H_0 : \mu_X - \mu_Y = 0$.
- Ipotesi alternative: $H_1 : \mu_X - \mu_Y < 0$.
- Livello di significatività: $\alpha = 0.05$.
- Test Statistica:

$$T = \frac{\hat{\mu}_{X,n} - \hat{\mu}_{Y,m}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}.$$

  Sotto l'ipotesi nulla, grazie al TLC, sappiamo che $T \sim N(0,1)$.

- Calcoliamo il valore critico k definito come,

$$P(T < k) = 0.05.$$

  Usando le tabelle otteniamo: $k = -1.4299$.

- Il valore critico è minore a quanto osservato dunque non rifiutiamo H_0.

4. Test per differenza di medie

- In questi esempi abbiamo considerato ipotesi alternative della forma,

$$H_1 : \mu_X - \mu_Y < 0, \text{ oppure } H_1 : \mu_X - \mu_Y > 0.$$

- Test di questa forma si dicono unilaterali.

- Tuttavia, avremmo potuto anche considerare come ipotesi alternativa,

$$H_1 : |\mu_X - \mu_Y| \neq 0.$$

- Test di questa forma si dicono bilaterali.

- Supponiamo di aver osservato $\hat{\mu}_{X,n} - \hat{\mu}_{Y,m} = 1$. Il test bilaterale si esegue nel seguente modo.

4. Test per differenza di medie: p-value

- Ipotesi nulla: $H_0 : \mu_X - \mu_Y = 0$.
- Ipotesi alternative: $H_1 : |\mu_X - \mu_Y| \neq 0$.
- Livello di significatività: $\alpha = 0.05$.
- Test Statistica:

$$T = \frac{|\hat{\mu}_{X,n} - \hat{\mu}_{Y,m}|}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} = \frac{|1|}{\sqrt{\frac{9}{36} + \frac{25}{49}}},$$

  Sotto l'ipotesi nulla, grazie al TLC, sappiamo che $T \sim N(0,1)$.

- Calcoliamo il p-value:

$$p - value = P\left(|T| > \frac{|1|}{\sqrt{\frac{9}{36} + \frac{25}{49}}}\right) = 0.2542.$$

- Il p-value è superiore al livello di significatività $\alpha = 0.05$, dunque non rifiutiamo H_0.

4. Test per differenza di medie: valore critico

- Ipotesi nulla: $H_0 : \mu_X - \mu_Y = 0$.
- Ipotesi alternative: $H_1 : |\mu_X - \mu_Y| < 0$.
- Livello di significatività: $\alpha = 0.05$.
- Test Statistica:
$$T = \frac{|\hat{\mu}_{X,n} - \hat{\mu}_{Y,m}|}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}.$$
Sotto l'ipotesi nulla, grazie al TLC, sappiamo che $T \sim N(0,1)$.
- Calcoliamo i valori critici k definito come,
$$P(|T| < k) = 0.05.$$
Usando le tabelle otteniamo: $k = -1.7089$ e $k = 1.7089$.
- Il valore osservato è all'interno dei valori critici dunque non rifiutiamo H_0.