

PSI Cheat Sheet 2C

Andrea Falbo - Quack

Cheat Sheet per il Secondo Compitino di Probabilità e Statistica per l'Informatica.

Lavoro diviso in due parti:

- Quiz a Risposta Multipla: Risoluzione degli esercizi pubblicati su E-Learning nell'anno accademico 2022/2023.
- Domande Aperte: Un esercizio per ogni argomento principale (basato su formulario), le soluzioni sono quelle fornite dai Prof, il mio è stato solo un partizionamento.

Osservazione: Nel formulario è presente un test sulla regressione verso la media ma non ho trovato nessun esercizio aperto svolto. Se ne trovate, oppure trovate errori, potete comunicarlo con Issues. Grazie della collaborazione.

Anno Accademico 2021/2022

1 Appello 2021/2022: Giugno

Domanda 5. Sia X una variabile aleatoria con legge binomiale $B(50, \frac{1}{2})$. Qual è, tra le seguenti, la migliore approssimazione di $\mathbb{P}(X > 30)$? (nel seguito indichiamo con Φ la f.d.r. di una variabile aleatoria Gaussiana standard.)

$1 - \Phi(30)$

$1 - \Phi(\sqrt{2})$

$\Phi(5)$

$1 - \Phi(\frac{5}{\sqrt{2}})$

$$n = 50 \quad p = \frac{1}{2}$$

$$\mu = E[X] = np = 50 \cdot \frac{1}{2} = 25$$

$$\sigma^2 = \text{Var}[X] = np(1-p) = 50 \cdot \frac{1}{2} \cdot \frac{1}{2} = 12.5$$

$$\mathbb{P}(X > 30) = \mathbb{P}(X - 25) / \sqrt{12.5} > (30 - 25) / \sqrt{12.5} = 1 - \mathbb{P}(Z \leq 5 / \sqrt{12.5}) = 1 - \phi(5 / \sqrt{2})$$

Domanda 6. Per un campione casuale estratto da una popolazione normale con varianza nota, l'intervallo bilatero per la media μ a livello di confidenza del 95% è $(-0.61, 0.78)$. Consideriamo il test per la verifica dell'ipotesi nulla $H_0 : \mu = 1$ vs $H_1 : \mu \neq 1$, cosa possiamo dire del p-value $\bar{\alpha}$ del test?

$\bar{\alpha} < 0.05$

$\bar{\alpha} > 0.025$

$\bar{\alpha} > 0.95$

$\bar{\alpha} < 0.025$

$IC = (-0.61, 0.78)$. $H_0: \mu = 1$, ma $1 \notin IC$. Allora al livello 0.05 rifiuto H_0 . Disegno p-value
accetto H_0 $\overline{\text{e rifiuto } H_0}$ Di conseguenza, so per certo che $\alpha < 0.05$

Domanda 7. In un modello di regressione lineare semplice

$$y_i = \alpha + \beta x_i + e_i \quad i = 1, \dots, n$$

si esegue un test di ipotesi con ipotesi nulla $H_0 : \beta \geq 1$ vs $H_1 : \beta < 1$ e si rifiuta H_0 a livello di significatività del 3%. Allora

- si rifiuta H_0 anche a livello di significatività del 2%
- c'è evidenza empirica di regressione verso la media
- Il p-value del test è $\bar{\alpha} = 1\%$
- I dati non sono in contraddizione significativa con H_0 .

Si può lavorare in due modi:

- a. Conosco il test (ultima pag. formulario) e so che se rifiuto H_0 c'è regressione verso la media.
- b. Non conosco il test, vado ad esclusione:
 3. Non posso confermarlo perché
 3. Non ho abbastanza dati per calcolare p-value
 4. Falso, rifiuto H_0 quindi i dati sono in contraddizione con H_0 .

Domanda 8. Si vuole verificare l'ipotesi che un certo medicinale abbassi il livello di colesterolo nel sangue. Si misura il livello di colesterolo a 100 persone prima e dopo la cura con il medicinale. Quale test usereste per verificare l'efficacia del farmaco?

- test z sulla differenza tra le medie di due campioni normali indipendenti
- test chi-quadrato di indipendenza
- test t sulla differenza tra le medie di due campioni normali accoppiati
- test z approssimato sulla proporzione

Tips: se il test è del tipo: "misuro su 1 stessa popolazione prima e dopo x" allora è test t dati accoppiati. Se invece è "confronto 2 popolazioni: una usai x e una non usai x / usai y" allora è test z indipendenti. In entrambi i casi è necessario controllare che la popolazione sia normale o che n sia sufficientemente grande.

2 Appello 2021/2022: Luglio

Domanda 5. Siano X_1, X_2, \dots, X_{100} variabili aleatorie indipendenti e identicamente distribuite, con legge di Poisson di parametro 1. Quale legge, tra le seguenti, meglio approssima la distribuzione della somma $X_1 + X_2 + \dots + X_{100}$?

- Gaussiana di media 100 e varianza 100
- Gaussiana di media 0 e varianza 1
- Gaussiana di media 50 e varianza 50
- Gaussiana di media 100 e varianza 10000

Quindi $\mu = E[X] = 100 \cdot 1 = 100$

$\sigma^2 = \text{Var}[X] = 100 \cdot 1 = 100$

Domanda 6. Per un campione casuale estratto da una popolazione normale con media μ e varianza σ^2 entrambi incognite, l'intervallo bilatero per la media μ a livello di confidenza del 99% è $(-0.12, 0.77)$. Con gli stessi dati, calcoliamo un intervallo di confidenza bilatero al livello del 95%. Quale tra questi è un possibile risultato?

- $(-0.08, 0.79)$
- $(-0.12, +\infty)$
- $(-0.08, 0.73)$
- $(-0.15, 0.79)$

dimensione intervallo aumenta all'aumentare della confidenza.

99 % ha $(-0.12, 0.77)$, a 95% sarà più piccolo, quindi $(-0.08, 0.73)$

Domanda 7. In un test chi-quadrato di buon adattamento alla distribuzione esponenziale di parametro $\frac{1}{2}$, si rifiuta l'ipotesi nulla a livello di significatività del 2.5%. Allora possiamo concludere che

- c'è evidenza empirica contro la distribuzione esponenziale
- Il p-value del test è $\bar{\alpha} = 1\%$
- c'è evidenza empirica contro la distribuzione esponenziale di parametro $\frac{1}{2}$
- i dati non sono in contraddizione significativa con l'ipotesi nulla

Possiamo ragionare in due modi:

- a. Rifiuto H_0 e quindi so che c'è evidenza empirica contro il test.
- b. Ad esclusione:
 1. Se rifiuto non ho evidenza contro la distribuzione, posso trovare un α non troppo grande (in questo caso anche 2,6%) tale per cui avrei accettato.
 2. Non posso dirlo con certezza
 4. Falso, lo sono, ho rifiutato.

Domanda 8. Si vuole verificare l'ipotesi che un certo medicinale regoli il livello di colesterolo nel sangue. Si misura il livello di colesterolo a 100 persone ipercolestolemiche, divise in due gruppi A e B , si somministra il medicinale alle persone del gruppo A e un placebo alle persone del gruppo B . Dopo il trattamento, si misura nuovamente il livello di colesterolo a tutti i pazienti. Tra quelli proposti, quale test usereste per verificare l'efficacia del farmaco?

- test z approssimato sulla proporzione
- test t sulla differenza tra le medie di due campioni normali indipendenti
- test chi-quadrato di indipendenza
- test t sulla differenza tra le medie di due campioni normali accoppiati

Come spiegato alla domanda 8 della pagina precedente, ho 2 popolazioni con una che varia x ed una y . n è suff. grande ($x_n = y_n = 50 \geq 30$) quindi t-indipendenti.

3 Appello 2021/2022: Settembre

Domanda 5. Sia $X \sim Bin(n, p)$ una variabile aleatoria Binomiale di parametri $n = 100$ e $p = 0.5$. Quale legge, tra le seguenti, meglio approssima la distribuzione della variabile aleatoria $Y = X - 50$?

- Gaussiana di media 50 e varianza 25
- Gaussiana di media 0 e varianza 1
- Gaussiana di media 0 e varianza 25
- Gaussiana di media 0 e varianza 50

$$\text{Var}[X] = n \cdot p \cdot (1-p) = 25$$

$$E[Y] = E[X - 50] = E[X] - 50 = 50 - 50 = 0$$

$$\text{Var}[Y] = \text{Var}[X - 50] = \text{Var}[X] = 25$$

Domanda 6. In un test per la verifica dell'ipotesi $H_0 : \mu \geq 2$ la regione critica è data da

$$C := \{(x_1, \dots, x_n) : T(x_1, \dots, x_n) < 1\}$$

dove T è un'opportuna statistica. Supponiamo che il campione dei dati osservati sia tale che $T(x_1, \dots, x_n) = 0.8$ e che il vero valore di μ sia 2.5. Allora

- si commette un errore di prima specie
- si commette un errore di seconda specie
- l'ipotesi H_0 è accettata
- non si commette alcun errore

In questi esercizi mi comporto così:

1. guardo se lo $T(x_1, \dots, x_n)$ osservato è C , cioè soddisfa condizione. Se si mi segno "rifiuto H_0 " mentalmente, "accetto H_0 " altrimenti. **Ricorda:** C è reg. critica, non vogliamo i nostri dati lì n!

In questo caso osservo 0.8 che appartiene a $T(x_1, \dots, x_n) < 1$. Quindi rifiuto.

2. Guardo se lo μ osservato rispetta H_0 . Se si mi segno " H_0 è vera" altrimenti " H_0 è falsa"

In questo caso $\mu=2.5$ rispetta $H_0: \mu \geq 2$.

3. Arrivo alle conclusioni:

· rifiuto H_0 e H_0 è vera \rightarrow errore 1° specie (il nostro caso)

· accetto H_0 e H_0 è falsa \rightarrow errore 2° specie

· accetto H_0 e H_0 è vera / rifiuto H_0 e H_0 è falsa \rightarrow no errore.

Domanda 7. In un modello di regressione lineare

$$Y = \alpha + \beta X + e$$

si esegue un test di ipotesi per verificare la bontà del modello, si sceglie quindi $H_0 : \beta = 0$.

A livello di significatività $\gamma = 5\%$ il campione delle risposte porta a concludere che il modello di regressione lineare è buono. Con gli stessi dati campionari si esegue il test al livello di significatività $\gamma = 10\%$. Cosa possiamo affermare con certezza?

- Il test a livello $\gamma = 10\%$ porterà a concludere che il modello non è buono
- Il test a livello $\gamma = 10\%$ porterà a concludere che il modello è buono
- Le informazioni non sono sufficienti a determinare l'esito del test a livello $\gamma = 10\%$
- il p-value del test è $\alpha = 4\%$

Modello buono \rightarrow rifiuto H_0 . In questo caso:

accetto H_0 $\xrightarrow{0.05}$ rifiuto H_0 Se rifiuto per 0.5, sicuramente rifiuto per 0.1
 \nearrow
 0.1

Domanda 8. Due campioni normali indipendenti X_1, \dots, X_{n_x} e Y_1, \dots, Y_{n_y} hanno uguale varianza campionaria $S_x^2 = S_y^2$. Cosa possiamo affermare con certezza?

- le numerosità sono diverse: $n_x \neq n_y$
- la varianza campionaria combinata è diversa da S_x^2
- le medie campionarie sono uguali: $\bar{x} = \bar{y}$
- la varianza campionaria combinata è uguale a S_x^2

Se $S_x^2 = S_y^2$, la loro combinata sarà anch'essa uguale.

Si dimostra in alternativa con la formula. Del formulario:

$$S_p^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2} = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_x^2}{n_x + n_y - 2} = \frac{S_x^2(n_x + n_y - 2)}{n_x + n_y - 2} = S_x^2$$

4 Appello 2021/2022: Gennaio

Domanda 5. Siano X_1, \dots, X_{100} v.a. i.i.d. uniformi continue sull'intervallo $(0, 1)$. Quale tra le seguenti è la migliore approssimazione di $P(X_1 + \dots + X_{100} \geq 50)$?

- $\frac{1}{2}$
- $e^{-\frac{1}{2}}$
- 1
- $\frac{1}{50}$

$$\begin{aligned} E[X] &= \frac{\alpha + \beta}{2} = \frac{1}{2} \quad , \quad E[X_1 + \dots + X_{100}] = 100 \cdot \frac{1}{2} = 50 \\ \text{Var}[X] &= \frac{(\beta - \alpha)^2}{12} = \frac{1}{12} \quad \text{Var}[X_1 + \dots + X_{100}] = \frac{100}{12} = \frac{25}{3} \\ &\text{C è continua, prob in 1 punto è } 0 \\ P\left(\frac{X_1 + \dots + X_{100} - 50}{\sqrt{25/3}} \geq \frac{50 - 50}{\sqrt{25/3}}\right) &= 1 - P(Z \leq 0) = 1 - 0.5 = 0.5 \end{aligned}$$

Oss: risolvibile anche graficamente

Domanda 6. In un test per la verifica dell'ipotesi $H_0 : \mu \geq 2$ la regione critica è data da

$$C := \{(x_1, \dots, x_n) : T(x_1, \dots, x_n) \leq 1\}$$

dove T è un'opportuna statistica. Supponiamo che il campione dei dati osservati sia tale che $T(x_1, \dots, x_n) = 0.7$. Allora

- rifiuto l'ipotesi H_0
- l'ipotesi H_0 non può essere rifiutata
- la probabilità dell'errore di prima specie è 0.05
- il p -value del test è 0.05

Procedo come spiegato alla pag. precedente:

1. $0.7 \leq 1 \rightarrow$ rifiuto H_0 (posso già concludere guardando le opzioni)
2. Non so l'osservazione di μ quindi non so se commetto errore, ma sicuramente rifiuto H_0

Domanda 7. L'ampiezza di un intervallo bilatero di confidenza al 98% per la media μ di un campione x_1, x_2, \dots, x_{100} estratto da una popolazione normale con varianza nota pari a 1 è (ricorda che $P(Z > z_\beta) = \beta$ con Z v.a. Gaussiana standard)

- $\frac{1}{10} z_{0.02}$
- $\frac{1}{5}$
- $\frac{1}{5} z_{0.01}$
- $z_{0.01}$

Ricorda: intervallo bilatero = 2 · scarto

In questo caso ha varianza nota pari a 1, $\alpha=0.02$ e $n=100$ quindi dal formulario

$$2 \cdot z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 2 \cdot z_{0.01} \cdot \frac{1}{\sqrt{100}} = 2 \cdot \frac{1}{10} \cdot z_{0.01} = \frac{1}{5} z_{0.01}$$

Domanda 8. In un test chi-quadrato di adattamento alla distribuzione ϱ_0 l'ipotesi nulla $H_0 : \varrho = \varrho_0$ viene rifiutata a livello di significatività $\alpha = 3\%$. Cosa possiamo dire con certezza?

- la probabilità dell'errore di prima specie è 0.05
- c'è una forte evidenza statistica di buon adattamento dei dati alla distribuzione ϱ_0
- il p -value del test è minore di 0.03
- la probabilità dell'errore di seconda specie è 0.03

Solita disegno accetto H_0 \overline{x} rifiuto H_0
0.03

5 Appello 2021/2022: Febbraio

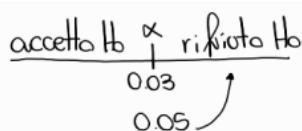
Domanda 5. Siano X_1, \dots, X_{100} v.a. i.i.d. con media 0 e varianza 1. Quale tra le seguenti è la migliore approssimazione di $P(X_1 + \dots + X_{100} \geq 0)$?

- $e^{-\frac{1}{2}}$
- $\frac{1}{2}$
- 1
- $\frac{1}{10}$

$$n=100, \mu=0, \sigma^2=1, P\left(\frac{X_1+\dots+X_{100}-0}{\sqrt{100}} \geq \frac{0-0}{\sqrt{100}}\right) = 1 - P(Z \leq 0) = 1 - 1/2 = 1/2$$

Domanda 6. Il p -value di un test per la verifica dell'ipotesi nulla H_0 è pari a 0.03. Allora

- rifiuto l'ipotesi H_0 al livello di significatività del 5%
- l'ipotesi H_0 non può essere rifiutata al livello di significatività del 7%
- rifiuto l'ipotesi H_0 quando l'errore di prima specie è pari a 0.01
- rifiuto l'ipotesi H_0 quando l'errore di prima specie è 0.02



Domanda 7. Quanto vale l'ampiezza di un intervallo bilatero di confidenza al 99% per la media μ di un campione x_1, x_2, \dots, x_{100} estratto da una popolazione normale con varianza nota pari a 1? (Si ricordi che $P(Z > z_\beta) = \beta$ con Z v.a. normale standard.)

- $\frac{1}{5} z_{0.01}$
- $\frac{1}{10}$
- $\frac{1}{5} z_{0.005}$
- $z_{0.005}$

Come prima: $2 \cdot z_{0.005} \frac{1}{10} = \frac{1}{5} z_{0.005}$

Domanda 8. In un modello di regressione lineare

$$Y = \alpha + \beta X + e$$

si esegue un test di ipotesi per verificare la bontà del modello, si sceglie quindi $H_0 : \beta = 0$.

A livello di significatività 3% il campione delle risposte porta a rifiutare H_0 . Con gli stessi dati campionari si esegue il test al livello di significatività 5%. Cosa possiamo affermare?

- Il test a livello 5% porterà a concludere che il modello è buono
- Il test a livello 5% non permetterà di concludere che il modello è buono
- Le informazioni non sono sufficienti a determinare l'esito del test a livello 5%
- il p-value del test è esattamente 0.05

Anche questo già visto:

modello regressione buono
modello regressione non buono

Anno Accademico 2020/2021

1 Appello 2020/2021: Giugno

Domanda 5. Si lancia 100 volte una moneta equilibrata. Qual è la probabilità di ottenere almeno 45 volte testa? Scegliere l'approssimazione migliore, dove Φ denota la funzione di ripartizione di una normale standard.

- $\Phi(1)$
- $1 - \Phi(1)$
- $\Phi(45)$
- $1 - \Phi(45)$

Lancio 100 volte una moneta, probabilità che escano almeno 45 teste si traduce in $X_1 + \dots + X_{100} \sim \sum X_i \sim \text{Bin}(100, \frac{1}{2})$, $P(X \geq 45)$?

Allora risolu:

$$\begin{aligned}\mu &= E[X] = np = 100 \cdot \frac{1}{2} = 50 \\ \sigma^2 &= \text{Var}[X] = np(1-p) = 100 \cdot \frac{1}{2} \cdot \frac{1}{2} = 25 \\ P\left(\frac{X-50}{\sqrt{25}} \geq \frac{45-50}{\sqrt{25}}\right) &= 1 - P(Z \leq -1) = \cancel{1} - \cancel{P(Z \leq 1)} = \phi(1)\end{aligned}$$

Domanda 6. Si consideri un campione casuale di ampiezza $n = 100$ estratto da una popolazione normale con varianza nota $\sigma^2 = 1$ e si costruisca l'intervallo di confidenza per la media al 95%. Successivamente, estraendo un campione dalla stessa popolazione, si vuole costruire un intervallo di confidenza per la media al 95% la cui ampiezza sia la metà di quella del primo intervallo considerato. Quante osservazioni occorrerà fare?

- 400
- 25
- 200
- Non si può rispondere perché non si conoscono i dati campionari

L'ampiezza dell'intervallo è inversamente proporzionale alla radice della numerosità.

In questo caso voglio metà ampiezza e quindi n quadruplicherà: $\left(\frac{1}{2}\right)^{-1} = 4$



Domanda 7. Sia X_1, \dots, X_n un campione casuale con distribuzione $\mathcal{N}(\mu, \sigma^2)$. Viene effettuato un test per la verifica dell'ipotesi $H_0 : \mu < 5$ contro $H_1 : \mu \geq 5$. Si ottiene un p-value pari a 0.06. Scegliere l'alternativa corretta.

- A un livello $\alpha = 7\%$ non vi è evidenza empirica che sia $\mu \geq 5$.
- A un livello $\alpha = 5\%$ non vi è evidenza empirica che sia $\mu \geq 5$.
- A un livello $\alpha = 10\%$ non si rifiuta l'ipotesi nulla
- Con le informazioni che abbiamo non si può rispondere perché non si conoscono i dati campionari

Riscrivo i dati come $\frac{\text{acc. } H_0}{0.05} \quad \frac{\bar{x}}{0.06} \quad \frac{\text{rif. } H_0}{0.07 \quad 0.01}$. Quindi la risposta è la b.

Domanda 8. Sia X_1, \dots, X_n un campione casuale con distribuzione $\mathcal{N}(\mu, \sigma^2)$. Viene effettuato un test per la verifica dell'ipotesi $H_0 : \mu < 5$ contro $H_1 : \mu \geq 5$. Si ottiene un p-value pari a 0.06. Scegliere l'alternativa corretta.

- A un livello $\alpha = 7\%$ non vi è evidenza empirica che sia $\mu \geq 5$.
- A un livello $\alpha = 10\%$ non si rifiuta l'ipotesi nulla
- A un livello $\alpha = 5\%$ non vi è evidenza empirica che sia $\mu \geq 5$.
- Con le informazioni che abbiamo non si può rispondere perché non si conoscono i dati campionari

Uguale alla precedente

2 Appello 2020/2021: Luglio

Domanda 5. Siano X_1, \dots, X_{100} variabili aleatorie i.i.d. Poisson di parametro $\lambda = \frac{1}{2}$. Allora la somma $X_1 + \dots + X_{100}$:

- è ben approssimata da una variabile aleatoria gaussiana di media 50 e varianza 50
- è ben approssimata da una variabile aleatoria gaussiana di media 50 e varianza 100
- è ben approssimata da una variabile aleatoria gaussiana di media 0 e varianza 1
- è ben approssimata da una variabile aleatoria gaussiana ma le informazioni non sono sufficienti per indicarne la media e la varianza.

$$\text{Pois}\left(\frac{1}{2}\right) \text{ per } n=100, \text{ allora } E[X] = 100 \cdot \frac{1}{2} = 50, \text{ Var}[X] = 100 \cdot \frac{1}{2} = 50$$

Domanda 6. Siano X_1, \dots, X_n variabili aleatorie i.i.d. con distribuzione uniforme sull'intervallo $(0, \lambda)$. Uno stimatore non distorto per λ è: (Sugg: può essere utile calcolare la media di una variabile aleatoria con distribuzione uniforme sull'intervallo $(0, \lambda)$.)

- \bar{X}_n
- $2\bar{X}_n$
- $\frac{\bar{X}_n}{2}$
- X_1

Ho una $U(0, \lambda)$ e devo trovare stimatore non distorto di λ . Allora dal formulario:

$$\bar{X}_n = E[X] = \frac{\lambda}{2} \text{ quindi tiro fuori } \lambda \text{ ed ottengo } \lambda = 2\bar{X}_n$$

Domanda 7. Si vuole verificare l'uguaglianza delle medie di due campioni di numerosità n , con varianze ignote. Allora

- si sceglie sicuramente un test t sulla differenza delle medie di due campioni normali accoppiati
- si sceglie sicuramente un test t sulla differenza delle medie di due campioni normali indipendenti
- si sceglie sicuramente un test t sulla differenza delle medie di due campioni normali accoppiati se $n < 30$, indipendenti altrimenti.
- le informazioni non sono sufficienti per determinare il tipo di test

Non possiamo dire nulla, non sappiamo se il campione è normale e nel caso non lo fosse non sappiamo la numerosità per approssimarla ad una normale.

Domanda 8. Si testa il buon adattamento del campione X_1, \dots, X_{100} alla distribuzione esponenziale di parametro $\lambda = \frac{1}{3}$, osservati i dati x_1, \dots, x_{100} , tramite un test chi-quadrato di buon adattamento. Il p-value del test è pari all' 1%. Allora a livello di significatività del 3% si deduce che

- il campione non segue una distribuzione esponenziale
- il campione segue una distribuzione esponenziale di di parametro $\frac{1}{3}$
- l'errore di seconda specie è pari al 3%
- il campione non segue una distribuzione esponenziale di di parametro $\frac{1}{3}$

Come specificato dal formulario nel test χ^2 buon adattamento H_0 indica che la popolazione ha una distribuzione. Dunque tornando all'esercizio acc. H_0 $\bar{\chi}^2$ 0.01 rif. H_0 0.03 quindi è la d.

3 Appello 2020/2021: Settembre

Domanda 5. In un test per la verifica dell'ipotesi $H_0 : \mu \leq 2$ la regione critica è data da

$$C := \{(x_1, \dots, x_n) : T(x_1, \dots, x_n) > 1\}$$

dove T è un'opportuna statistica. Supponiamo che il campione dei dati osservati sia tale che $T(x_1, \dots, x_n) = 0.6$ e che il vero valore di μ sia 1.5. Allora

- si commette un errore di prima specie;
- si commette un errore di seconda specie;
- l'ipotesi H_0 è rifiutata;
- non si commette alcun errore.

Procedo con:

- $0.6 < 1$ quindi accetto H_0
- $1.5 \leq 2$ quindi H_0 è vera

Essendo H_0 accettata e vera non commetto errori.

Domanda 6. Si lancia 900 volte un dado non truccato, sia X la variabile aleatoria che conta il numero di volte in cui è uscito 6 in questi 900 lanci. Qual è la migliore approssimazione di $\mathbf{P}(X > 150)$? (nel seguito indichiamo con Z una variabile aleatoria Gaussiana standard)

- $\mathbf{P}(Z > 150);$
- $\mathbf{P}(Z > 0);$
- $\mathbf{P}(Z \leq 150);$
- $\mathbf{P}(Z > 1).$

H_0 una Binomiale con $n=900$, $p=\frac{1}{6}$ e $x=150$. Allora

$$\mathbb{E}[X] = np = 150, \quad \text{Var}[X] = np(1-p) = 125 \quad \text{e quindi}$$

$$\mathbf{P}(X > 150) = 1 - \mathbf{P}\left(\frac{X - 150}{\sqrt{125}} \leq \frac{150 - 150}{\sqrt{125}}\right) = 1 - \mathbf{P}(Z \leq 0) = \mathbf{P}(z > 0)$$

Domanda 7. In un modello di regressione lineare

$$Y = \alpha + \beta X + e$$

si esegue un test di ipotesi per verificare se c'è regressione verso la media, si sceglie quindi $H_1 : \beta < 1$. A livello di significatività $\gamma = 10\%$ il campione delle risposte porta a concludere che c'è regressione verso la media. Con gli stessi dati campionari si esegue il test al livello di significatività $\gamma = 5\%$.

- Il test a livello $\gamma = 5\%$ porterà sempre a concludere che c'è regressione verso la media;
- Il test a livello $\gamma = 5\%$ porterà sempre a concludere che non c'è regressione verso la media;
- Le informazioni non sono sufficienti a determinare l'esito del test a livello $\gamma = 5\%$;
- se il p-value è 7% , allora il test a livello $\gamma = 5\%$ porterà a concludere che c'è regressione verso la media.

Ottengo regressione verso la media se accetto H_0 quindi: $\frac{\text{acc } H_0}{?} \xrightarrow{?} \text{rif. } H_0$
Non posso concludere dato che si trova a livello 5% , se \bar{x} fosse $?0.05?$
stato 0.07 non avrei avuto regressione in quanto accettavo H_0 a livello 0.5

Domanda 8. Due campioni normali indipendenti X_1, \dots, X_{n_x} e Y_1, \dots, Y_{n_y} hanno uguale varianza campionaria $S_x^2 = S_y^2$. Allora

- hanno sempre la stessa media campionaria $\bar{x} = \bar{y}$;
- la varianza campionaria combinata è sempre uguale a S_x^2 ;
- non si può usare il test t sulla differenza delle medie;
- nessuna delle precedenti è corretta.

Già svolta [qui](#)

4 Appello 2020/2021: Gennaio

Domanda 5. Sia X una variabile aleatoria con legge binomiale $B(100, \frac{1}{2})$. Qual è la migliore approssimazione di $\mathbb{P}(X > 60)$? (nel seguito indichiamo con φ la f.d.r. di una variabile aleatoria Gaussiana standard.)

- $\Phi(1)$
- $1 - \Phi(60)$
- $\Phi(2)$
- $1 - \Phi(2)$

Binomiale con $n=100$, $p=\frac{1}{2}$ e $x=60$ dunque $E[x]=50$, $\text{Var}[x]=25$ e
 $\mathbb{P}(X > 60) = 1 - \mathbb{P}(X \leq 60) = 1 - \mathbb{P}\left(Z \leq \frac{60-50}{\sqrt{25}}\right) = 1 - \varphi(2)$

Domanda 6. Si consideri una popolazione Bernoulliana con parametro p incognito. Per stimare la media si estrae il campione X_1, \dots, X_{3n} e si considerano le due statistiche

$$T_1 = \frac{X_1 + \dots + X_n}{n}, \quad T_2 = \frac{X_1 + \dots + X_{3n}}{3n}.$$

Dire se T_1 e se T_2 sono entrambi estimatori non distorti per la media, e, nel caso lo siano, dire qual è preferibile, volendo minimizzare l'errore standard.

- Sono entrambi estimatori non distorti e T_1 è preferibile a T_2 .
- Sono entrambi estimatori non distorti e T_2 è preferibile a T_1 .
- T_1 è non distorto, mentre T_2 lo è.
- T_2 è non distorto, mentre T_1 lo è.

Sono entrambi estimatori in quanto sono medie campionarie per una Bernoulli. T_2 è preferibile in quanto più numeroso.

Domanda 7. In un test di ipotesi, l'ipotesi nulla NON viene rifiutata a livello di significatività del 4%. Quale delle seguenti affermazioni è sicuramente vera?

- Il p-value del test è maggiore di 0.04
- Il p-value del test è compreso tra 0.04 e 0.05.
- L'ipotesi nulla NON viene rifiutata a livello di significatività del 5%.
- L'ipotesi nulla viene rifiutata a livello di significatività del 5%.

acc. $H_0 \neq$ rif. H_0
0.04

Domanda 8. In un modello di regressione lineare semplice, il coefficiente di determinazione è pari a 0.64. Il coefficiente di correlazione

- può essere uguale a 0.64
- deve essere uguale a 0.8
- può essere uguale a -0.8
- nessuna delle precedenti affermazioni è corretta

Dal formulario $R^2 = r_{x,y}^2$ quindi $x = \pm \sqrt{64} \rightarrow x = \pm 8$, allora scelgo l'opzione con "può".

5 Appello 2020/2021: Febbraio

Domanda 5. Consideriamo variabili aleatorie indipendenti X_1, X_2, \dots, X_{50} , tutte con la stessa distribuzione di Poisson di parametro $\lambda = 2$. Che cosa possiamo dire sulla distribuzione della somma $X_1 + X_2 + \dots + X_{50}$?

- È approssimativamente normale con media 2 e varianza 2.
- È approssimativamente normale con media 100 e varianza 100.
- È approssimativamente normale con media 100 e varianza 200.
- Non è approssimativamente normale.

$$\text{Pois}(2), n=50, E[X] = 2, \text{Var}[X] = 2$$

$$\mu = E[X_1 + \dots + X_{50}] = 50 \cdot 2 = 100$$

$$\sigma^2 = \text{Var}[X_1 + \dots + X_{50}] = 50 \cdot 2 = 100$$

Domanda 6. In un test per la verifica dell'ipotesi $H_0 : \mu \leq 2$ la regione critica è data da

$$C := \{(x_1, \dots, x_n) : T(x_1, \dots, x_n) > 1\}$$

dove T è un'opportuna statistica. Supponiamo che il campione dei dati osservati sia tale che $T(x_1, \dots, x_n) = 0.85$. Allora

- l'ipotesi H_0 non può essere rifiutata
- rifiuto l'ipotesi H_0
- il p -value del test è 0.05
- nessuna delle precedenti è corretta

Domanda 7. Due campioni normali indipendenti X_1, \dots, X_{n_x} e Y_1, \dots, Y_{n_y} hanno uguale varianza campionaria $S_x^2 = S_y^2$. Allora

- hanno sempre la stessa media campionaria $\bar{x} = \bar{y}$
- la varianza campionaria combinata è sempre uguale a S_x^2
- non si può usare il test t sulla differenza delle medie
- nessuna delle precedenti è corretta

Già svolta [qui](#)

Domanda 8. In un test chi-quadrato di adattamento alla distribuzione π_0 l'ipotesi nulla $H_0 : \pi = \pi_0$ viene rifiutata a livello di significatività $\alpha = 3\%$. Quale delle seguenti affermazioni è corretta?

- il p -value del test è maggiore di 0.03
- c'è una forte evidenza statistica di buon adattamento dei dati alla distribuzione π_0
- la probabilità dell'errore di prima specie è 0.05
- c'è una forte evidenza statistica che i dati non seguano la distribuzione π_0

Come già visto, se rifiuto H_0 non segue dist. π_0

Stima media: Test z - varianza nota

5

Esercizio 3. Tra i pasticcini prodotti artigianalmente in una pasticceria se ne prelevano $n = 100$; risulta che il loro peso medio è pari a 35 g. Si sa che la deviazione standard del peso di tutti i pasticcini prodotti dalla pasticceria è pari a $\sigma = 4$ g.

- (a) Si trovi l'intervallo di confidenza per il peso medio di tutti i pasticcini prodotti a livello di confidenza del 98%.
- (b) Quanto deve essere numeroso il campione se si vuole che l'ampiezza dell'intervallo a livello di confidenza del 98% si dimezzi?
- (c) Si determini quanti pasticcini occorre ancora estrarre se si vuole che lo stimatore del peso medio si discosti dal vero peso medio per meno di un grammo con probabilità del 96%.

Soluzione 3. Il campione $\{x_1, \dots, x_n\}$ di numerosità $n = 100$ è estratto da una popolazione normale di media incognita μ e deviazione standard nota $\sigma = 4$.

- (a) Un intervallo bilatero di confidenza per μ al livello del $100(1 - \alpha)\%$ è

$$\left(\bar{x}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right), \quad (1)$$

dove $z_{\alpha/2}$ denota il $100(1 - \alpha/2)$ -esimo percentile della distribuzione normale standard. Nel nostro caso $n = 100$, $\alpha = 1 - 0.98 = 0.02$, $\bar{x}_n = 35$, $\sigma = 4$, $\alpha/2 = 0.01$ e $z_{\alpha/2} = z_{0.01}$. Dalle tavole della distribuzione Gaussiana si ha che

$$\Phi(2.32) = 0.9898 \quad \text{e} \quad \Phi(2.33) = 0.99010$$

per cui interpolando troviamo $z_{0.01} \approx 2.325$.

Per sostituzione in (1) troviamo che un intervallo di confidenza bilatero per μ al livello del 98% è

$$\left(35 - 2.325 \frac{4}{\sqrt{100}}, 35 + 2.325 \frac{4}{\sqrt{100}} \right) = (34.07, 35.93).$$

- (b) L'ampiezza di un intervallo bilatero di confidenza per μ al livello del $100(1 - \alpha)\%$ della forma (1) è

$$2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}; \quad (2)$$

fissato α l'ampiezza diminuisce al crescere di n , e diminuisce come \sqrt{n} , quindi per dimezzare l'ampiezza bisogna prendere un campione che sia numeroso il quadruplo: per un campione di numerosità $4n$ l'ampiezza è infatti data da

$$2z_{\alpha/2} \frac{\sigma}{\sqrt{4n}} = \frac{1}{2} \left(2z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right),$$

che è la metà di (2).

- (c) Una stima per μ è \bar{x}_n ; tramite l'approssimazione normale per rispondere alla richiesta si può chiedere che la semiampiezza di un intervallo bilatero di confidenza per μ al livello del $96\% = 100(1 - \alpha)\%$ (formula (1)) sia minore a 1. La semiampiezza è data da

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}};$$

nel nostro caso $\alpha = 100 - 0.96 = 0.04$, $z_{\alpha/2} = z_{0.02} \approx 2.055$ (infatti dalle tavole $\Phi(2.05) = 0.9798$ e $\Phi(2.06) = 0.9803$) e, ricordando che $\sigma = 4$ si deve imporre

$$2.055 \cdot \frac{4}{\sqrt{n}} < 1 \text{ ossia } \sqrt{n} > 8.22 \text{ da cui } n > (8.22)^2 = 67.5684$$

e passando all'intero successivo si ottiene $n \geq 68$.

Stima media: Test t - varianza incognita

5

Esercizio 3. Nella ditta XXX l'ammontare delle fatture segue approssimativamente una distribuzione gaussiana con media e varianza incognite. Si vuole valutare se l'ammontare delle fatture sia uguale a 120 euro. Dai valori delle 12 fatture campionate si ha

$$\bar{x}_n = 112.85 \text{ euro}, \quad s_n = 20.80 \text{ euro} .$$

- (a) Si costruisca l'intervallo di confidenza al livello del 95% per l'ammontare medio delle fatture.
- (b) Tramite un opportuno test, verificare, a livello di significatività del 5%, l'ipotesi che l'ammontare medio delle fatture sia uguale a 120 euro.
- (c) Stimare il p-value del test al punto b) e commentare il risultato.

Soluzione 3. Il campione $\{x_1, \dots, x_n\}$ di numerosità $n = 12$ è estratto da una popolazione normale di media incognita μ e deviazione standard incognita σ .

- (a) Un intervallo bilatero di confidenza per μ al livello del $100(1 - \alpha)\%$ è

$$\left(\bar{x}_n - t_{n-1,\alpha/2} \frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{n-1,\alpha/2} \frac{s_n}{\sqrt{n}} \right),$$

dove $t_{n-1,\alpha/2}$ denota il $100(1 - \alpha/2)$ -esimo percentile della distribuzione t di Student. Nel nostro caso $n = 12$, $\alpha = 0.05$, $\bar{x}_n = 112.85$ e $s_n = 20.80$ e $t_{11,0.025} = 2.201$. Per sostituzione troviamo che un intervallo di confidenza bilatero per μ al livello del 95% è

$$\left(112.85 - 2.201 \frac{20.80}{\sqrt{12}}, 112.85 + 2.201 \frac{20.80}{\sqrt{12}} \right) = (99.63, 126.07).$$

- (b) Eseguiamo un test di verifica delle ipotesi

$$H_0 : \mu = 120 \qquad \qquad H_1 : \mu \neq 120$$

a livello di significatività $\alpha = 0.05$. Dal punto (a) notiamo che il valore 120 appartiene all'intervallo di confidenza per μ al livello del 95% dunque i dati non consentono di rifiutare l'ipotesi nulla a livello del 5%.

Soluzione alternativa. Eseguiamo il test di verifica delle ipotesi

$$H_0 : \mu = 120 \qquad \qquad H_1 : \mu \neq 120$$

a livello $\alpha = 0.05$ la cui regione critica è

$$\left| \frac{\bar{x}_n - \mu_0}{s_n} \sqrt{n} \right| > t_{n-1,\alpha/2}.$$

Nel nostro caso, $\mu_0 = 120$ e

$$\left| \frac{112.85 - 120}{20.80} \sqrt{12} \right| = 1.19 < 2.201$$

quindi l'ipotesi nulla non viene rifiutata a livello del 5% (e quindi a tutti i livelli di significatività inferiori).

Entrambe le soluzioni portano alla stessa conclusione: *a livello di significatività del 5% i dati non ci permettono di concludere che l'ammontare delle fatture sia diverso da 120.*

- (c) Per stimare il p-value $\bar{\alpha}$ eseguiamo il test a livello di significatività $\alpha = 20\%$ e $\alpha = 40\%$: dato che $t_{11,0.2} = 0.876$ e $t_{11,0.1} = 1.363$ e dal punto precedente

$$\left| \frac{112.85 - 120}{20.80} \sqrt{12} \right| = 1.19,$$

i dati consentono di rifiutare H_0 a livello di significatività del 40% ma non lo consentono a livello di significatività del 20%, da cui si ottiene che $0.2 < \bar{\alpha} < 0.4$. *Quindi dati non permettono di concludere che l'ammontare medio delle fatture sia diverso da 120 euro.*

Stima media, Test z - popolazione Bernoulliana 5

Esercizio 3. Un vivaista controlla un campione di 300 piantine scelte in modo casuale da una coltura e osserva che di queste 20 presentano una determinata malattia.

- (a) Fornire una stima puntuale della proporzione di piantine malate e un intervallo di confidenza al livello del 95% per la proporzione di piantine malate.
- (b) Quante piantine avrebbe dovuto controllare il vivaista per essere sicuro al livello di confidenza del 95% che la precisione della stima puntuale di p fosse inferiore a 0.02?
- (c) Eseguire un opportuno test statistico, con ipotesi nulla $p_0 = 0.05$.

Soluzione 3. Il campione $\{x_1, \dots, x_n\}$ di numerosità $n = 300$ è estratto da una popolazione Bernoulliana di parametro p incognito, dove p rappresenta la proporzione di piantine malate.

- (a) La stima puntuale di p è data da $\bar{x}_n = \frac{20}{300} = 0.067$, da cui è immediato verificare che $300\bar{x}_n = 20 > 5$, $300(1 - \bar{x}_n) = 280 > 5$, quindi sappiamo che l'intervallo di confidenza per la proporzione al livello del $100(1 - \alpha) = 95\%$ è dato da

$$\left(\bar{x}_n - z_{\alpha/2} \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}}, \bar{x}_n + z_{\alpha/2} \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}} \right).$$

Nel nostro caso $n = 300$, $\bar{x}_n = 0.067$, $\alpha = 0.05$, e sulle tavole della distribuzione normale cerchiamo $z_{0.025}$ tale che $\Phi(z_{0.025}) = 0.975$ trovando

$$z_{0.025} = 1.96$$

Per sostituzione si ottiene che un intervallo di confidenza bilatero per la proporzione di piantine malate al livello del 95% è

$$\begin{aligned} & \left(0.067 - 1.96 \sqrt{\frac{0.067(1 - 0.067)}{300}}, 0.067 + 1.96 \sqrt{\frac{0.067(1 - 0.067)}{300}} \right) \\ & = (0.067 - 0.028, 0.067 + 0.028) = (0.039, 0.095) \end{aligned}$$

- (b) La semiampiezza dell'intervallo di confidenza a livello $100(1 - \alpha)\%$ per la proporzione di piantine malate è data da

$$z_{\alpha/2} \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}} \% \leq 1.96 \frac{1}{2\sqrt{n}}$$

dove nell'ultimo passaggio si è tenuto conto del fatto che $\alpha = 0.05$ e $z_{\alpha/2} = 1.96$ e $\bar{x}_n(1 - \bar{x}_n) \leq \frac{1}{4}$. Ne segue quindi che

$$0.98 \frac{1}{\sqrt{n}} \leq 0.02$$

da cui

$$n \geq (49^2) = 2401.$$

- (c) Impostiamo un test sulla proporzione, a patto che sia $np_0 \geq 5$, $n(1 - p_0) \geq 5$: nel nostro caso abbiamo $n = 300$, $p_0 = 0.05$ e le condizioni sono verificate. Eseguiremo un test con ipotesi nulla $H_0 : p = p_0$ e ipotesi alternativa $H_1 : p \neq p_0$ e i dati consentono di rifiutare H_0 a livello di significatività $\alpha = 5\%$ se $|z| = \left| \frac{\bar{x}_n - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n} \right| > z_{\alpha/2}$. Senza fare ulteriori calcoli per l'esecuzione del test, dal punto (a) notiamo che il valore 0.05 appartiene all'intervallo di confidenza per μ al livello del 95% dunque accettiamo l'ipotesi nulla a livello del 5%.

Test t - Dati Indipendenti

7

Esercizio 4. La presenza di zinco nell'acqua potabile ne influenza il sapore e può risultare nociva. Viene condotta un'analisi in sei punti diversi di un fiume e in ciascun punto viene misurata la concentrazione in superficie e in profondità. I dati raccolti sono i seguenti:

zone	1	2	3	4	5	6
concentrazione in profondità	0.430	0.266	0.567	0.531	0.707	0.716
concentrazione in superficie	0.415	0.238	0.390	0.410	0.605	0.609

Si può concludere che la (vera) concentrazione media di zinco in profondità eccede quella in superficie? Per rispondere alla domanda si effettua un test di ipotesi per “dati accoppiati”.

- (a) Quale ipotesi aggiuntiva è necessario fare? Specificare l'ipotesi nulla H_0 , l'ipotesi alternativa H_1 e la statistica del test
- (b) Effettuare il test a livello di significatività del 10% .
- (c) Dare una stima per il p -value del test: cosa si può concludere in relazione alla domanda posta?

Soluzione 4. Eseguiremo un test t sulla differenza delle medie di due campioni normali

- (a) Dobbiamo fare l'ipotesi che i due campioni siano estratti da popolazioni normali. Eseguiremo un test sui due campioni accoppiati X_1, \dots, X_6 (di media μ_x) che rappresenta la concentrazione di zinco in profondità nei 6 diversi punti, e il campione Y_1, \dots, Y_6 (di media μ_Y) che rappresenta la concentrazione di zinco in superficie nei 6 diversi punti.

Eseguiremo il test t sul campione delle differenze $D_1 = X_1 - Y_1 = 0.015$, $D_2 = X_2 - Y_2 = 0.028$, $D_3 = X_3 - Y_3 = 0.177$, $D_4 = X_4 - Y_4 = 0.121$, $D_5 = X_5 - Y_5 = 0.102$, $D_6 = X_6 - Y_6 = 0.107$, denotiamo con μ_d la media, con $\bar{D}_n = \frac{1}{n}(D_1 + \dots + D_n)$ la media campionaria e con $S_d^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D}_n)^2$ la varianza campionaria del campione D_1, \dots, D_n con $n = 6$,

Eseguiremo un test con

H_0	H_1	Statistica
$\mu_X \leq \mu_Y \rightsquigarrow \mu_d \leq 0$	$\mu_X > \mu_Y \rightsquigarrow \mu_d > 0$	$T = \frac{\bar{D}_n}{S_d} \sqrt{n}$

- (b) La regione critica del test di cui al punto precedente è data da $\frac{\bar{D}_n}{S_d} \sqrt{n} > t_{n-1,\alpha}$, con $\alpha = 0.10$. Dalle tavole ricaviamo che $t_{n-1,\alpha/2} = t_{5,0.10} = 1.476$ mentre dai dati campionari $\bar{D}_n = \frac{0.55}{6} = 0.09167$ e $S_d^2 = 0.003069$; si ha quindi $t = 3.6998$: i dati consentono di rifiutare H_0 a livello di significatività del 10% e di concludere che la percentuale di zinco in profondità è maggiore alla percentuale di zinco in superficie.
- (c) Dalle tavole si vede che $t_{5,0.01} = 3.365$, a livello $\alpha = 1\%$ i dati consentono di rifiutare H_0 e $t_{5,0.004} = 4.032$, a livello $\alpha = 0.5\%$ i dati non consentono di rifiutare H_0 : denotato con p il p -value si ha $0.005 < p < 0.01$ c'è una forte evidenza statistica che la concentrazione media di zinco in profondità ecceda quella in superficie.

Test t - dati indipendenti

7

Esercizio 4. Sono stati campionati 25 bambini i cui genitori sono affetti da diabete di tipo II e 25 bambini i cui genitori non sono affetti da diabete. I primi presentavano un livello medio di glicemia a digiuno pari a 86.1 mg/dl, mentre gli altri pari a 82.2 mg/dl. È noto che le deviazioni standard dei due campioni sono pari a 2.09 mg/dl per il campione dei bambini con genitori diabetici e a 2.49 mg/dl per il campione dei bambini con genitori non diabetici. Si vuole verificare con un opportuno test statistico se la malattia dei genitori modifica il livello medio di glicemia dei bambini, per questo si sottopone a test l'ipotesi nulla che il livello medio di glicemia nei due campioni sia uguale.

- (a) Calcolare la varianza campionaria combinata.
- (b) Scegliere un opportuno test per la verifica dell'ipotesi nulla che le medie dei due campioni coincidano ed eseguirlo a livello di significatività $\alpha = 5\%$ e a livello $\alpha = 1\%$.
- (c) Utilizzare i risultati ottenuti al punto precedente per dare una stima del p-value del test (basta dire se il p -value è maggiore/minore di un certo valore, o se è compreso tra due valori) e fornire un'interpretazione statistica.

Soluzione 4. (a) La varianza campionaria combinata è data da:

$$S_p^2 = \frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2} = \frac{1}{2}(S_x^2 + S_y^2)$$

dato che i campioni sono ugualmente numerosi. Quindi

$$s_p^2 = \frac{2.09^2 + 2.49^2}{2} \approx 5.28.$$

(b) Si tratta di due campioni indipendenti, X_1, \dots, X_{25} (glicemia nei bambini con genitori con diabete) e Y_1, \dots, Y_{25} (glicemia nei bambini con genitori senza diabete), entrambi di numerosità pari a 25, ossia $n_x = n_y = 25$, che possiamo supporre normali in quanto campioni di misurazioni (si misura la glicemia a digiuno). Possiamo applicare il test t sulla differenza tra le medie di due campioni indipendenti con la stessa varianza, infatti la condizione $\frac{1}{2} < \frac{s_x^2}{s_y^2} < 2$ è verificata dato che

$$\frac{s_x^2}{s_y^2} = \frac{2.09^2}{2.49^2} \approx 0.7045.$$

Siano μ_x e μ_y le medie (incognite) dei due campioni \bar{X} e \bar{Y} le medie campionarie e S_x^2 e S_y^2 le varianze campionarie. Dobbiamo eseguire il test per $H_0 : \mu_x = \mu_y$ vs $H_1 : \mu_x \neq \mu_y$. La statistica del test è data da

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

e rifiutiamo H_0 se

$$t = \left| \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \right| > t_{n_x+n_y-2,\alpha/2}.$$

Nel nostro caso $\bar{x} = 86.1$, $\bar{y} = 82.2$, $n_x = n_y = 25$, $\alpha/2 = 0.025$ e dalle tavole $t_{n_x+n_y-2,\alpha/2} = t_{48,0.025} = 2.011$; quindi

$$t = \left| \frac{5(86.1 - 82.2)}{\sqrt{2 * 5.28}} \right| \approx 6 > 2.011 :$$

i dati consentono di rifiutare H_0 al livello $\alpha = 5\%$.

Eseguiamo ora il test al livello $\alpha = 1\%$: $t_{n_x+n_y-2,\alpha/2} = t_{48,0.005} = 2.682$, poiché $t \approx 6$ i dati consentono di rifiutare H_0 anche al livello $\alpha = 1\%$.

- (c) Dal punto (b) deduciamo che il p -value del test è inferiore a 0.01, quindi è estremamente piccolo: c'è una forte evidenza statistica a favore di H_1 , ossia del fatto che le medie dei due campioni siano diverse, ossia del fatto che avere un genitore diabetico influisca sul livello medio di glicemia a digiuno dei figli.

ChiQuadro Discreta

7

Esercizio 4. Durante la seconda guerra mondiale, la parte meridionale di Londra fu colpita da 535 bombe volanti V1. Per analizzare la distribuzione geografica dei punti di impatto, tale area è stata suddivisa in 576 regioni di pari superficie, registrando quante bombe sono cadute in ciascuna regione. I dati sono riportati nella seguente tabella:

Bombe ricevute	0	1	2	3	4	5	6 o più
Numero di regioni (n_k)	229	211	93	35	7	1	0

L'obiettivo è vedere se si può affermare che il numero di bombe cadute in una regione segua una distribuzione di Poisson di parametro λ . Per far questo

- (a) Fornire una stima di λ dai dati.
- (b) Calcolare le frequenze attese delle classi assegnate.
- (c) Eseguendo un opportuno test, dire se a livello di significatività del 5% si può affermare che il numero di bombe cadute in una regione segua una distribuzione di Poisson.

Soluzione 4. Eseguiremo un test di adattamento ad una distribuzione di Poisson di parametro λ incognito.

- (a) Ricordiamo che la media di una distribuzione di Poisson di parametro λ è pari a λ , dunque scegliamo come stima di λ la media campionaria ossia ($n = 576$)

$$\bar{x}_n = \frac{0 \cdot 229 + 1 \cdot 211 + 2 \cdot 93 + 3 \cdot 35 + 4 \cdot 7 + 5 \cdot 1}{576} = \frac{535}{576} \approx 0.929.$$

- (b) Le frequenze attese sono date da

$$f_k = n \cdot \pi(k) = 576 \cdot e^{-0.929} \frac{(0.929)^k}{k!}, \quad k \in \{0, \dots, 5\},$$

mentre per l'ultima classe

$$f_6 = n \cdot (1 - (\pi(0) + \dots + \pi(5))) = n - (f_0 + \dots + f_5).$$

Essendo $n = 576$, si ottiene la seguente tabella:

Bombe ricevute	0	1	2	3	4	5	6 o più
Frequenze attese (f_k)	227.5	211.3	98.2	30.4	7.1	1.3	0.2

- (c) Vorremmo eseguire il test chi-quadrato di adattamento alla distribuzione di Poisson al livello di significatività $\alpha = 0.05$. le nostre ipotesi sono

H_0 : la popolazione segue una distribuzione di Poisson

H_1 : la popolazione non segue una distribuzione di Poisson

Notiamo però dal punto (b) che la regola empirica di applicabilità del test non è soddisfatta, infatti c'è una classe la cui frequenza attesa è < 1 . Raggruppando le ultime due classi otteniamo

Bombe ricevute	0	1	2	3	4	5 o più
Frequenze osservate (n_k)	229	211	93	35	7	1
Frequenze attese (f_k)	227.5	211.3	98.2	30.4	7.1	1.5

dove tutte le frequenze attese sono ≥ 1 e l'83.3% circa di esse è ≥ 5 , dunque le condizioni di applicabilità del test sono soddisfatte. Possiamo ora eseguire il test, la

sua regione critica è

$$q = \sum_{k=0}^5 \frac{(n_k - f_k)^2}{f_k} > \chi_{0.05,4}^2.$$

(Notare che i gradi di libertà della chi-quadrato sono 4 in quanto ci sono 6 classi e abbiamo stimato un parametro.) La statistica del test vale

$$q = \frac{(229 - 227.5)^2}{227.5} + \dots + \frac{(1 - 1.5)^2}{1.5} \approx 1.15.$$

Sulla tavola della distribuzione chi-quadrato troviamo che

$$\chi_{0.05,4}^2 = 9.49$$

dunque l'ipotesi nulla non può essere rifiutata: a livello di significatività 5% non si può escludere che i dati siano compatibili con una distribuzione di Poisson.

ChiQuadro assolutamente continua

7

Esercizio 4. Con un generatore di numeri casuali che dovrebbe simulare una distribuzione X con distribuzione uniforme in $[0, 1]$, sono stati generati 150 numeri ottenendo la seguente tabella che fornisce i dati raggruppati

classe	intervallo	frequenze osservate
1	$[0, 1/4)$	30
2	$[1/4, 1/2)$	37
3	$[1/2, 3/4)$	52
4	$[3/4, 1]$	31

Si vuole verificare, tramite un opportuno test, che X segua effettivamente una distribuzione uniforme in $[0, 1]$.

- Calcolare le frequenze attese delle classi assegnate.
- Calcolare il valore della statistica del test sui dati.
- Eseguire il test a livello di significatività del 5% e dell' 1% e commentare i risultati ottenuti.

Soluzione 4. Eseguiremo un test *chi*-quadrato di adattamento con ipotesi nulla $H_0 : X \sim \mathcal{U}([0, 1])$ con due parametri incogniti.

- Se vale H_0 tutte le classi sono equiprobabili con probabilità $1/4$ e la frequenza attesa di ciascuna classe è $f_1 = 150/4 = 37.5$, $i = 1, \dots, 4$.
- La statistica del test è

$$Q = \sum_{i=1}^4 \frac{(N_i - f_i)^2}{f_i}$$

e sui dati assume il valore

$$q = 206/25 = 8,24$$

- Si rifiuta H_0 a livello α se $q > \chi_{3,\alpha}^2$. Poichè

$$\chi_{3,0.05}^2 = 7,815, \quad \chi_{3,0.01}^2 = 11,345$$

i dati consentono di rifiutare H_0 a livello $\alpha = 5\%$ ma non a livello $\alpha = 1\%$, quindi il p-value $\bar{\alpha} \in (0.01, 0.05)$: c'è una buona evidenza statistica che il simulatore di numeri casuali non segua una distribuzione $\mathcal{U}([0, 1])$.

Test ChiQuadro Indipendenza

Esercizi aggiuntivi

Esercizio In un'indagine epidemiologica si sono classificate 100 persone secondo i seguenti caratteri

X = "di norma usa l'autobus"

Y = "influenzato durante l'inverno"

ottenendo la seguente tabella di contingenza :

	Influenzato	Non influenzato
Usa l'autobus	50	16
Non usa l'autobus	12	22

Verificare l'ipotesi di indipendenza tra X e Y a livello 5%.
Cosa concludete?

Svolgimento

Completiamo la tabella di contingenza con le frequenze marginali:

X \ Y	Influenzato	Non influenzato	TOT
Usa l'autobus	50 = m_{11}	16 = m_{12}	66 = m_1^x
Non usa l'autobus	12 = m_{21}	22 = m_{22}	34 = m_2^x
TOT	62 = m_1^y	38 = m_2^y	100 = m

Testiamo l'ipotesi:

H_0 : X e Y sono indipendenti

contro H_1 : X e Y non sono indipendenti

Dal formulare, la RC è uguale a livello di significatività $\alpha = 5\%$

$$q = \sum_{\substack{1 \leq i \leq 2 \\ 1 \leq j \leq 2}} \frac{\left(m_{ij} - \frac{m_i^x m_j^y}{m} \right)^2}{\frac{m_i^x m_j^y}{m}} > \chi^2_{1, \alpha}$$

Nel nostro caso, $\alpha = 0.05$ è

$$\begin{aligned} q = & \frac{\left(50 - \frac{66 \cdot 62}{100} \right)^2}{\frac{66 \cdot 62}{100}} + \frac{\left(16 - \frac{66 \cdot 38}{100} \right)^2}{\frac{66 \cdot 38}{100}} \\ & + \frac{\left(12 - \frac{34 \cdot 62}{100} \right)^2}{\frac{34 \cdot 62}{100}} + \frac{\left(22 - \frac{34 \cdot 38}{100} \right)^2}{\frac{34 \cdot 38}{100}} \approx 15.594 \end{aligned}$$

Dalle tabelle del χ^2 ,

$$\chi^2_{1, 0.05} = 3.841$$

dunque rifiuto H_0 : X e Y sono indip. a livello di significatività 5%. Quindi c'è evidenza empatica contro H_0 .

Regressione

6

Esercizio 4. In un esperimento diretto allo studio della relazione tra il numero di pulsazioni sotto sforzo per minuto (Y) e l'età in anni (X) sono stati rilevati i seguenti dati su $n = 5$ soggetti di sesso maschile:

X	20	25	29	31	45
Y	195	190	188	185	163

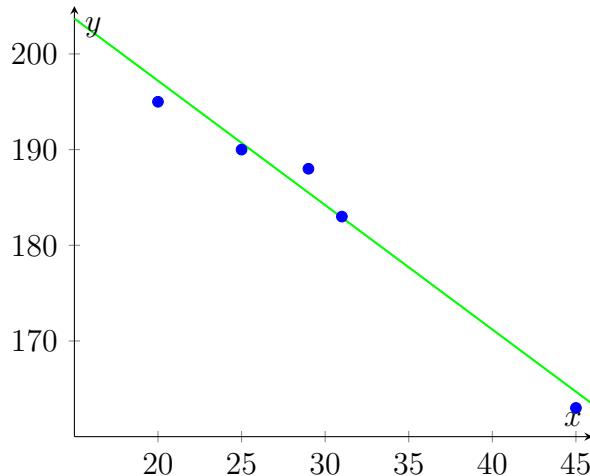
- (a) Rappresentare il diagramma di dispersione delle osservazioni dei due caratteri e calcolare il coefficiente di correlazione. Sulla base del grafico e del risultato ottenuto per il coefficiente di correlazione il modello di regressione lineare

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, \dots, 9$$

appare adeguato? Motivare la risposta.

- (b) Stimare i parametri α e β col metodo dei minimi quadrati e scrivere l'equazione della retta di regressione.
- (c) Si verifichi l'ipotesi nulla $H_0 : \beta = 0$ contro l'alternativa $H_1 : \beta \neq 0$ e si faccia un commento sulla bontà del modello, usando anche i risultati ottenuti ai punti precedenti.

Soluzione 4. (a) Il diagramma di dispersione (in verde è disegnata la retta di regressione che calcoleremo al punto successivo) è dato da



Il modello di regressione lineare appare adeguato in quanto i punti sul grafico sembrano disporsi approssimativamente lungo una retta (con coefficiente angolare < 0): al crescere dell'età tendono a diminuire le pulsazioni.

Calcoliamo ora il coefficiente di correlazione lineare ϱ : si ha

$$\bar{x} = 30, \quad \bar{y} = 184.2, \quad S_{xx} = \sum_{i=1}^5 (x_i - \bar{x})^2 = 352$$

$$S_{xY} = \sum_{i=1}^5 x_i y_i - 5 \bar{x} \bar{y} = -458, \quad S_{YY} = \sum_{i=1}^5 (y_i - \bar{y})^2 = 614.8$$

da cui

$$\varrho = \frac{S_{xY}}{\sqrt{S_{xx} S_{YY}}} = \frac{-458}{\sqrt{352 \times 614.8}} = -0.9845$$

e, pertanto, si può concludere che esiste una forte correlazione lineare negativa.

(b) Si ha

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}} = \frac{-458}{352} = -1.30, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 184.2 + 30 \cdot 1.30 = 223.2$$

dunque la retta di regressione è $y = 223.2 - 1.30 \cdot x$.

(c) Utilizzeremo un test per la verifica dell'ipotesi $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ con livello di significatività γ : rifiutiamo H_0 se

$$\left| \sqrt{\frac{S_{xx}(n-2)}{SS_R}} \hat{\beta} \right| > t_{n-2,\gamma/2}$$

dove SS_R è la somma dei quadrati residui ed è data da $SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}} = \frac{352 \cdot 614.8 - 458^2}{352} = 18.88$, quindi la statistica del test sui dati vale

$$\left| -\sqrt{\frac{352 \cdot 3}{18.88}} \cdot 1.3 \right| = 9.72.$$

Dalle tavole si ha che $t_{3,0.05} = 2.353$ quindi i dati consentono di rifiutare l'ipotesi nulla: viene confermata una forte evidenza statistica della validità del modello di regressione lineare semplice, come anticipato dal coefficiente di correlazione che è, in valore assoluto, molto vicino a 1.