ESERCIZIO 1

Uno studente vuole indagare, in un suo gruppetto di 6 amici, se c'è o meno una relazione fra giorni passati senza studiare nemmeno un'ora e voto preso all'ultimo esame. Nello specifico, va a vedere in quanti dei 15 giorni precedenti all'esame i suoi amici non hanno studiato nemmeno un'ora (variabile X) e il voto preso a tale esame (variabile Y). Ottiene i valori presenti nella tabella:

Amico	Giorni	Voto
1	3	29
2	8	25
3	6	22
4	4	27
5	10	24
6	5	29

a) Trova l'equazione di previsione e interpretala. Trova qual è il voto atteso per una persona che non studia per 15 giorni.

b) Verifica se c'è un'associazione negativa significativa fra le due variabili, con $\alpha = 0.05$ e interpreta i risultati.

c) Trova la correlazione fra le due variabili e interpretala.

ESERCIZIO 2

Un'indagine su studenti e ideologia politica è andata a vedere se c'erano differenze nell'ideologia politica fra studenti vegetariani e non vegetariani. Le risposte sull'ideologia politica (da 1, molto liberale, a 7, molto conservatrice) avevano una media di 2.22 e una dev. stand. di 0.67 per i 9 studenti vegetariani e una media di 3.18 e una dev. stand. di 1.72 per i 51 studenti non-vegetariani.

a) Verifica se esiste una differenza significativa nell'ideologia politica fra gli studenti non vegetariani e vegetariani. Svolgi una verifica di ipotesi con $\alpha = 0.05$ e interpreta i risultati.

b) Se costruissi un intervallo di confidenza al 95%, questo intervallo conterrebbe lo zero? Rispondi a questo punto spiegando il perché e senza calcolare l'intervallo di confidenza.

QUESITO TEORICO

Nella regressione lineare cos'è il "Metodo dei Minimi Quadrati"?

DOMANDA BREVE 1

Un esperimento su dati appaiati riguardante la rilevazione del rumore sotto due condizioni ha utilizzato un campione di dodici bambini di 9 mesi e ha riportato una stima della differenza delle medie di 70.1 con relativa deviazione standard di 49.4.

Verifica, con $\alpha = 0.05$, se c'è una differenza significativa fra le due condizioni e spiega, brevemente, qual è l'aspetto teorico che permette di fare inferenza in questo modo.

DOMANDA BREVE 2

<u>Più di una risposta può essere corretta:</u>

Sia β la probabilità di commettere un errore del 2° tipo. Con un livello $\alpha = 0.05$ si verifica H_0: $\mu \leq 0$ contro H_1: $\mu > 0$ con $n = 30$ osservazioni, $\beta = 0.36$ per $\mu = 4$.

a) Se $\mu = 5$, allora per $n = 30$ e $\alpha = 0.05$, $\beta > 0.36$

b) se $\alpha = 0.01$, allora per $n = 30$ e $\mu = 4$, $\beta > 0.36$

c) Se $n = 50$, allora per $\mu = 4$ e $\alpha = 0.05$, $\beta > 0.36$

d) La potenza del test è 0.64 per $\mu = 4$, $n = 30$ e $\alpha = 0.05$

e) L'ipotesi deve essere falsa, perché necessariamente $\alpha + \beta = 1$.

RISOLUZIONE:

esercizio 1:

n = 6. x = giorni precedenti all'esame in cui non si ha studiato nemmeno un'ora. y = voto preso.

a)

Per trovare l'equazione di previsione $\hat{y} = a + bx$ utilizzo le formule: $\mathbf{a = \bar{y} - b\bar{x}}$ e $\boldsymbol{b = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})*(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$**.**

Prima di tutto calcolo media e deviazione standard sia di x che di y. $\mathbf{\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}}$ e $\mathbf{s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}}$ e $\mathbf{\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}}$ e $\mathbf{s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$**.**

Ottengo che: $\mathbf{\bar{x} = 6,}$ $\mathbf{s_x = 2.61,}$ $\mathbf{\bar{y} = 26,}$ $\mathbf{s_y = 2.83.}$

A questo punto trovo il coefficiente angolare **b**. Per semplificarmi i calcoli (qui non riportati) osservo che il denominatore della formula per trovare **b** corrisponde al numeratore della formula per trovare la deviazione standard di x.

$\boldsymbol{b = \frac{[(3-6)*(29-26)+(8-6)*(25-26)....]}{34}}$ **= - 0.706.**

Adesso che ho **b** trovo il valore dell' intercetta **a**.

a = 26 – (-0.706 * 6) = 26 + 4.236 = 30.236.

Quindi l'equazione di previsione è $\mathbf{\hat{y} = 30.236 - 0.706x.}$

L'interpretazione dell'intercetta è: Il voto atteso ($\hat{y}$) per una persona con x = 0, ovvero per una persona che non ha passato nemmeno un giorno precedente all'esame senza studiare, è circa 30.

L'interpretazione del coefficiente angolare è: all'aumentare unitario di x, ovvero per ogni giorno in più passato senza studiare, il voto atteso ($\hat{y}$) diminuisce di circa 0.7.

Per trovare qual è il voto atteso per una persona che non studia per 15 giorni sostituisco x = 15 nell'equazione di previsione trovata:

$\hat{y}$ = 30.236 – 0.706*15 = 30.236 – 10.59 = 19.95.

Quindi il voto atteso è circa 20.

b)

Il punto mi chiede di verificare se c'è un' associazione negativa. Quindi l'ipotesi è unidirezionale:

H_0: $\beta \geq 0$ e H_1: $\beta < 0$

Per trovare la statistica test t: $\boldsymbol{t = \frac{b - b_0}{se_b}}$ ho bisogno di: $\boldsymbol{se_b = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$**;** $\boldsymbol{s_e^2 = \frac{SQerr}{n-2}}$**;** $\boldsymbol{SQerr = \sum_{i=1}^{n}(y_i - \hat{y_i})^2}$**.**

SQerr è la sommatoria dei residui elevati al quadrato, dove il residuo è la differenza fra il valore "vero" y di un'osservazione e il valore previsto dall'equazione di previsione per tale osservazione. Devo quindi trovare i valori previsti per le 6 osservazioni campionarie.

$\hat{y}_1$ = 30.236 – 0.706*3 = 28.118; $\hat{y}_2$ = 30.236 – 0.706*8 = 24.588;

$\hat{y}_3 = 30.236 - 0.706*6 = 26$; $\hat{y}_4 = 30.236 - 0.706*4 = 27.412$;

$\hat{y}_5 = 30.236 - 0.706*10 = 23.176$; $\hat{y}_2 = 30.236 - 0.706*5 = 26.706$;

Una volta trovati i valori previsti trovo quindi le differenze al quadrato.

$(y_1 - \hat{y}_1)^2 = (29 - 28.118)^2 = 0.882^2$; $(y_2 - \hat{y}_2)^2 = (25 - 24.588)^2 = 0.412^2$;

$(y_3 - \hat{y}_3)^2 = (22 - 26)^2 = (-4)^2$; $(y_4 - \hat{y}_4)^2 = (27 - 27.412)^2 = (-0.412)^2$;

$(y_5 - \hat{y}_5)^2 = (24 - 23.176)^2 = 0.82^2$; $(y_6 - \hat{y}_6)^2 = (29 - 26.706)^2 = 2.23^2$;

Sommo quindi queste differenze fra di loro è ottengo il risultato ricercato:

SQerr = 23.056. (Per motivi di approssimazione potrebbe non venire esattamente uguale).

Quindi trovo in sequenza tutte gli altri valori di cui ho bisogno:

$\boldsymbol{s_e^2 = \frac{23.056}{6-2}}$ **= 5.765.**

$\boldsymbol{se_b = \sqrt{\frac{5.765}{34}}}$ **= 0.412.**

$\boldsymbol{t = \frac{-0.706 - 0}{0.412}}$ **= - 1.714.**

Ora che ho la statistica test devo confrontarla con la statistica test t "critica" che lascia alla sinistra della distribuzione il 5% della distribuzione per vedere se il risultato è statisticamente significativo. Per la regressione lineare i gradi di libertà sono n – 2, quindi 4.

Sulla tavola t a pag. 287 del libro di Luccio, incrociando 4 gdl con la probabilità 0.05 trovo che il valore t critico è 2.13 (visto che stiamo guardando la coda sinistra sarebbe -2.13).

Confronto in valore assoluto (per evitare confusione) i due valori di t trovati e osservo che:

|- 1.714| < |-2.13|. Essendo inferiore non ho evidenze per rifiutare l'ipotesi nulla.

Quindi, concludo che non ci sono evidenze a sostegno dell'ipotesi alternativa. Ovvero concludo che non sembra esserci un'associazione negativa significativa fra giorni precedenti all'esame senza studiare e voto preso all'esame. In altre parole, sembrerebbe non esserci associazione significativa fra non studiare nei 15 giorni precedenti all'esame e il peggioramento del voto preso a tale esame.

c)

Per trovare la correlazione utilizzo la formula che fa uso del coefficiente angolare **b**. $\boldsymbol{r_{xy} = b * \frac{s_x}{s_y}}$

Ho a disposizione il necessario quindi ricavo che: . $\boldsymbol{r_{xy} = -0.706 * \frac{2.61}{2.83}}$ **= - 0.65.**

L'interpretazione è: Una correlazione di -0.65 mi dice che, all'aumentare di una deviazione standard delle x, il valore atteso di y diminuisce di 0.65 deviazioni standard. Questa correlazione è un valore medio-alto, il che mi suggerirebbe una buona intensità di associazione fra le due variabili, ma la piccola dimensione campionaria non mi permette di trarre molte conclusioni.

ESERCIZIO 2

$n_1 = 9$, $\bar{x}_1 = 2.22$, $s_1 = 0.67$; $n_2 = 51$ $\bar{x}_2 = 3.18$, $s_2 = 1.72$

a)

Questo punto mi chiede se c'è una differenza significativa, quindi è un'ipotesi bilaterale.

H_0: $\mu_2 - \mu_1 \neq 0$ H_1: $\mu_2 - \mu_1 = 0$.

Per verificare se la differenza è statisticamente significativa devo calcolare la statistica test $\boldsymbol{t = \frac{(\bar{x}_2 - \bar{x}_1) - (\mu_2 - \mu_1)_0}{se}}$**.** In questo problema i due campioni sono indipendenti, quindi la formula dell'errore standard è **se =** $\boldsymbol{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$**.**

Trovo che **se =** $\boldsymbol{\sqrt{\frac{0.67^2}{9} + \frac{1.72^2}{51}}}$ **= 0.328.**

La statistica test t è quindi: $\boldsymbol{t = \frac{(3.18 - 2.22) - 0}{0.328}}$ **= 2.93.**

Per verificare se questo risultato è statisticamente significativo devo confrontarlo con l'appropriato t "critico". Per farlo devo sapere i gradi di libertà e per campioni indipendenti i gradi di libertà appropriati li trovo con la formula di Welch-Satterthwaite: **v =** $\boldsymbol{\frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}}$ Sostituendo i valori: **v = 30.77** (≈**31**).

Quindi sulla tavola t a pag. 287 del libro di Luccio, incrociando 30 gdl (approssimo a 30 visto che a infinito sarebbe esagerato) con la probabilità 0.025 trovo che il valore t critico è 2.04. Osservo che **2.93 > 2.04**.

Essendo la statistica test t trovata superiore ho evidenze per rifiutare l'ipotesi nulla. Quindi concludo che ci sono evidenze a favore dell'ipotesi alternativa. Ovvero, sembrerebbe che ci siano differenze nell' ideologia politica fra studenti vegetariani e studenti non vegetariani.

b)

Se costruissi un intervallo di confidenza quest'ultimo NON conterrebbe lo zero. Perché dal punto **a)** posso concludere che 0 è un valore implausibile.

Posso concluderlo perché esiste una relazione fra test bidirezionale e intervallo di confidenza. Se errore standard e α coincidono, essi mi danno risultati coerenti. Se il test afferma che un valore è plausibile (o implausibile) allora l'intervallo di confidenza mi porterà a concludere la stessa cosa.

Infatti, se nel test bidirezionale la regione di rifiuto è α, l'ampiezza della regione di non-rifiuto 1-α corrisponde all'intervallo di confidenza costruito con α centrato sul valore assunto sotto H_0. Se il mio risultato cade nella regione α, allora un intervallo di confidenza con ampiezza 1-α non conterrebbe il valore assunto sotto H_0.

Essendo la media assunta sotto H_0 uguale a 0, allora la differenza fra le medie campionarie 3.18 e 2.22 sarebbe caratterizzata da un intervallo che non conterrebbe questo 0.

QUESITO TEORICO:

Nella regressione lineare il metodo dei minimi quadrati è un metodo utilizzato nella computazione dell'equazione di previsione.

Per un certo set di osservazioni ci sono infinite possibili equazioni di previsione per prevedere i dati. Ogni osservazione del set ha un residuo/errore a seconda della retta utilizzata (residuo = $y_i - \hat{y}_i$). Per calcolare la variazione dei dati intorno alla retta di previsione si fa la somma di tutti i residui al quadrato (ovvero SQerr).

Il metodo dei minimi quadrati mi permette di estrapolare quella retta di previsione che minimizza il valore di SQerr. Ovvero, mi permette di calcolare la retta che riduce al minimo gli errori dalle osservazioni a disposizione.

DOMANDA BREVE 1:

L'esperimento descritto è su dati appaiati, ovvero su campioni dipendenti. n = 12, $\overline{d}$ = 70.1 e sd = 49.4.

Per verificare se c'è una differenza significativa applico il test t per campioni dipendenti. Le ipotesi sono:

H_0: $\delta \neq 0$ H_1: $\delta = 0$.

La statistica test t si calcola con: $\boldsymbol{t = \frac{\overline{d} - \delta_0}{se}}$**.** Dove $\boldsymbol{se = \frac{s_d}{\sqrt{n}}}$**.**

Ricavo che: $\boldsymbol{se = \frac{49.4}{\sqrt{12}}}$ **= 14.26.**

Ricavo che: $\boldsymbol{t = \frac{70.1 - 0}{14.26}}$ **= 4.92.**

A questo punto per vedere se il risultato è statisticamente significativo devo confrontare la statistica test t trovata con il t "critico". Per confronto fra campioni dipendenti i gradi di libertà sono uguali a n-1.

Quindi sulla tavola t a pag. 287 del libro di Luccio, incrociando 11 gdl con la probabilità 0.025 trovo che il valore t critico è 2.20. Osservo che **4.92 > 2.20.** Quindi rifiuto l'ipotesi nulla.

Per confronto fra campioni dipendenti il test statistico si può considerare come un test t per una singola media. Il motivo per cui è possibile fare ciò e perché per questo confronto si può utilizzare la media delle differenze fra le osservazioni appaiate dei due campioni invece che la differenza fra le medie dei due campioni. Infatti, queste due coincidono ("La media delle differenze tra le osservazioni dei due gruppi è uguale alla differenza tra le medie dei due gruppi").

Così facendo non ci sono più due campioni di osservazioni ma solo uno.

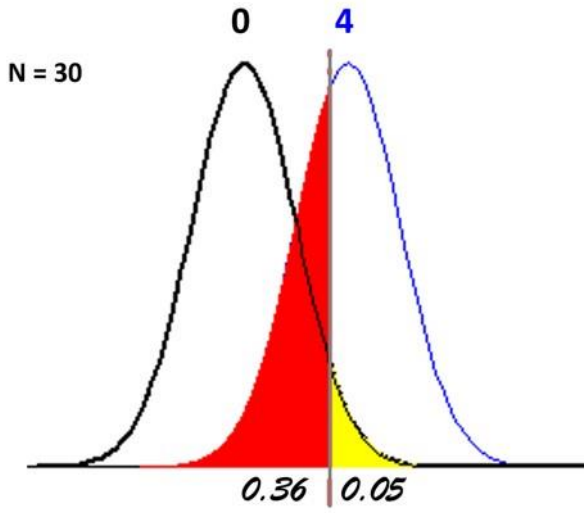
DOMANDA BREVE 2:

Le risposte corrette sono **b)** e **d)**.

Spiegazione:

α è **l'errore del primo tipo**. Ovvero la probabilità di rifiutare H_0 quando invece NON bisognerebbe rifiutarla.

β è invece **l'errore del secondo tipo**. Ovvero la probabilità di NON rifiutare H_0 quando invece bisognerebbe rifiutarla.

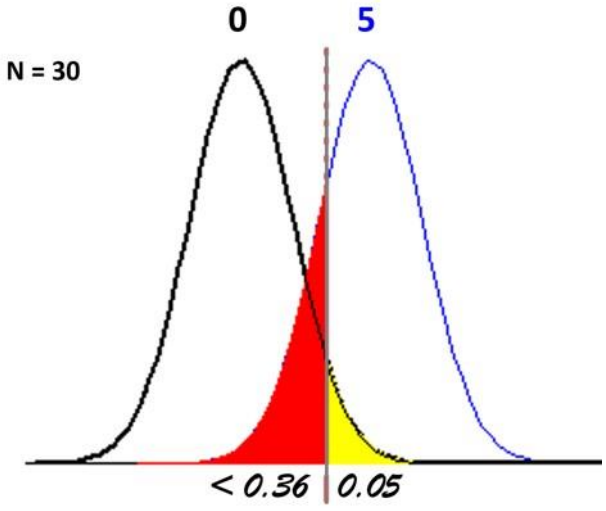


La potenza di un test si calcola con **(1 – β)**. Rappresenta la probabilità di rifiutare H_0 quando essa è effettivamente falsa (ovvero è la capacità del test di non commettere un **errore del secondo tipo**).

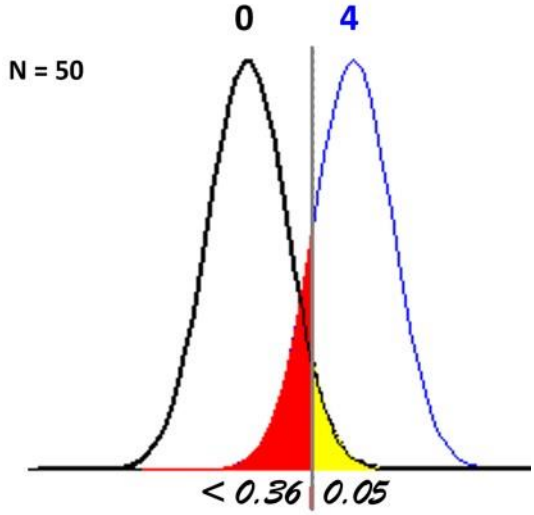
β è complessa da calcolare ma ha delle proprietà. In particolare, dipende da **α**, dalla dimensione campionaria **n** e dalla distanza del parametro VERO della popolazione dal parametro assunto sotto H_0.

β diminuisce quando:

1) Il valore del parametro VERO si <u>allontana</u> da quello ipotizzato in H_0 (c'è meno sovrapposizione).



2) La dimensione campionaria **n** aumenta. (Le distribuzioni si “snelliscono”).



3) La probabilità di commettere un errore del primo tipo, **α**, aumenta. Aumentando α, si riduce l’area di protezione dell’ipotesi nulla (1- α), con conseguente diminuzione della probabilità di un errore di secondo tipo e quindi aumento della potenza del test (1- β).