Progetto di laborotorio - Calcolo delle Probabilità e Statistica Matematica 2

Žana Ilić - 898373

Settembre 2020

In questo progetto di Calcolo delle Probabilità e Statistica Matematica usiamo una database heart.csv dove ci sono 14 variabili che sono sintomi di malattie cardiache e 303 osservazioni cioè pazienti. Tutte le variabili sono espressi in numeri. Qualche sono in numeri interi positivi e qualche sono in numeri binari (0 e 1) che solitamente rapressentano che variabile è vera o falsa (oppure se sesso di paziente è maschio o femmina)...

Le variabili sono:
1. age - età (da 29 a 77)
2. sex - sesso (1 = maschio; 0 = femmina)
3. cp - tipo di dolore toracico (valori da 0 a 3)
4. trestbps - pressione sanguigna a riposo (da 94 a 200)
5. chol - colesterolo in mg/dl (da 126 a 564)
6. fbs - glicemia > 120 mg/dl (1 = vero; 0 = falso)
7. restecg - risultati elettrocardiografici a riposo (valori da 0 a 2)
8. thalach - battito cardiaco massimo raggiunto (da 71 a 202)
9. exang - angina indotta (1 = vero; 0 = falso)
10. oldpeak - depressione ST indotta (da 0 a 6.2)
11. slope - la pendenza del ST segmento (valori da 0 a 2)
12. ca - numero di navi principali (valori da 0 a 3)
13. thal (valori da 0 a 3)
14. target (1 = vero; 0 = falso)

Nel primo quesito vediamo la statistica descrittiva dei dati. Vogliamo rispondere alle domande:
- Guardando la variabile age e variabile sesso, cosa si può osservare?
- Trovare funzione di ripartizione empirica di pressione sanguigna quando il paziente è arrivato in ospedale.
- Abbiamo visto tante volte che il colesterolo è un killer silenzioso di pazienti affetti da malattie cardiache. Cosa si può osservare?

Prima osservo che, guardando la variabile age:
- range di anni dei nostri pazienti è da 29 a 77 anni
- mediana è 55 anni
- media è 54.37 anni
- media armonica è 52.73 anni
- media di pazienti maschi è 55.68 anni
- media di pazienti femmine è 53.76 anni
- deviazione standard è 9.082101
- quantili sono: 29 Min; 47.5 25%; 55 50%; 61 75%; 77 Max

Concludo che ci sono al più pazienti malatti tra 57.8 e 62.6 anni: 63 persone. In Figura 1 si può vedere il numero di pazienti in base alla loro età rappresentato con l'istogramma. Osserviamo che che le pazienti femmine si ammalano in giovane età rispetto ai maschi.

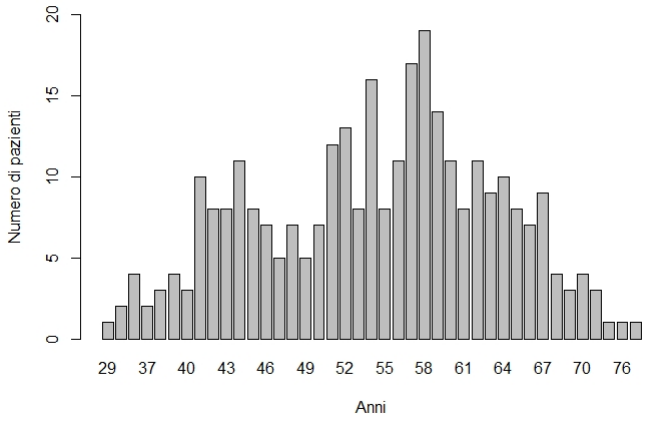


Figura 1: Numero di pazienti in base alla loro età

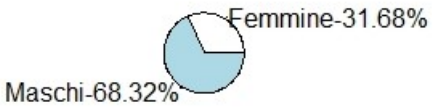


Figura 2: Relazione tra pazienti maschi e femmine

Per quanto riguarda il sesso, ci sono 96 pazienti femmine e 207 pazienti maschi. In percentuali ci sono 31.68% pazienti femmine e 68.32 % pazienti maschi, cioè ci sono più maschi che femmine con malattie cardiache. Si vede in pie chart nella Figura 2.

Ora troviamo pressione sanguigna a riposo (trestbps) quando il paziente è arrivato in ospedale. Minimo è 94 mmHg e il massimo è 200 mmHg. Rappresentiamolo con funzione di ripartizione empirica in Figura 3. Concludiamo che la maggior parte dei pazienti ha pressione sanguigna tra 125.8 e 136.4, 74 dei quali.

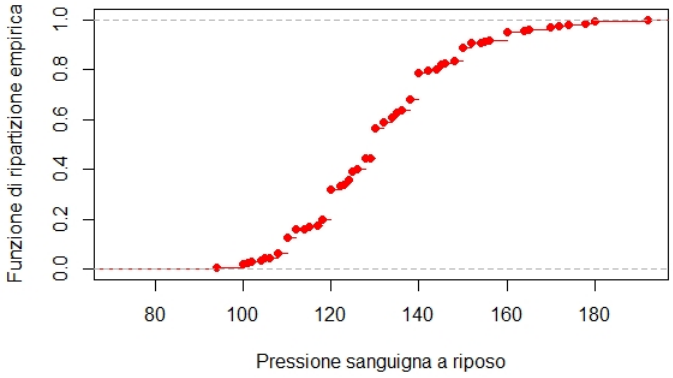


Figura 3: Funzione di ripartizione empirica di pressione sanguigna

La ricerca mostra che la maggior parte dei pazienti hanno la quantità di colesterolo nel sangue tra 213.6 e 257.4, loro 106. Rappresentiamo l'istogramma con una curva normale in Figura 4.

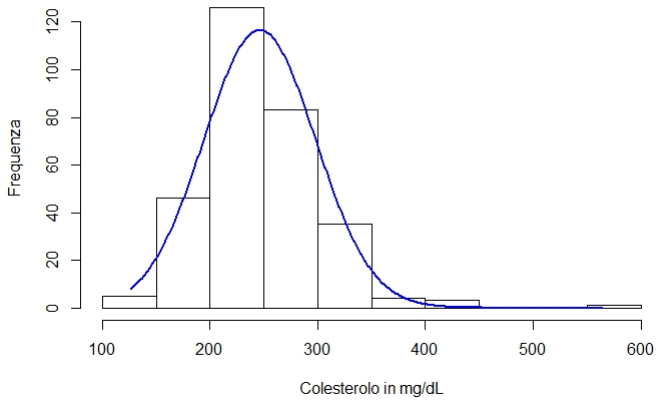


Figura 4: Histogramma con curva normale di quantità di colesterolo

Nel secondo quesito vogliamo fare un test di ipotesi per indicare battito cardiaco medio come una soglia per tachicardia. Poi facciamo un test di normalità per distribuzioni dei due campioni maschi e femmine rispetto a quantità di colesterolo. Facciamo anche due test per verificare se il sesso ha alcun effetto sulla quantità di colesterolo oppure sul depressione ST indotta.

Guardiamo variabile thalach, che significa battito cardiaco massimo raggiunto. Vogliamo indicare battito cardiaco medio per pazienti malatti, che può essere una indicazione per la tachicardia, con un errore 1% oppure 5%. Verifichiamo se il valore 150 è compatibile. Costruiamo un test di ipotesi per verificare l'ipotesi nulla $H_0 : \mu = \mu_0 = 150$ contro l'alternativa che è $H_1 : \mu < \mu_0$ cioè che il valore sia in realtà inferiore di parametro minimo richiesto 150. Abbiamo trovato che p-value è pari a 0.3943. Sia per $\alpha = 0.01$ che per $\alpha = 0.05$ abbiamo che $\alpha < p$, quindi il test non permette di rifiutare l'ipotesi nulla, ovvero possiamo dire che per un paziente la soglia per indicare tachicardia può essere battito cardiaco 150. Per calcolare questo abbiamo usato t-test per un campione (one sample t-test). Nella figura 5 si può vedere che per variabile thalach l'ipotesi di normalità può essere accettata. Verifichiamo questo con Kolmogorov-Smirnov test per un campione. Poichè p-value è pari a 0.09187 possiamo accetare l'ipotesi nulla - che distribuzione non è statisticamente diversa da una distribuzione normale.

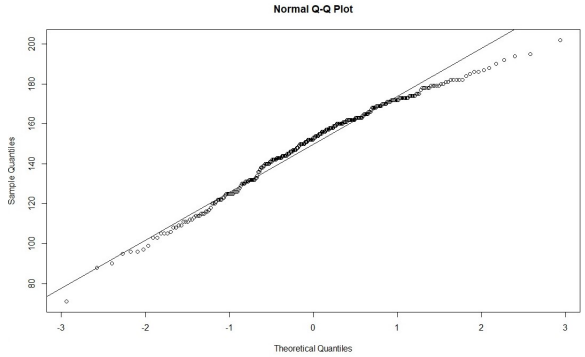


Figura 5: Q-Q plot e linea dei quantili teorici di insieme di dati normali per variabile thalach

Ora vogliamo dividere tutti i pazienti in due gruppi, maschi e femmine. Vogliamo fare test non parametrici di normalità di variabile chol che rappresenta colesterolo per due gruppi, cioè vogliamo vedere se si può rappresentare in distribuzione normale. Nella tabella si vedono i risultati di p-value per vari test.

normality test	maschi	femmine
Lilliefors	0.3154	0.02191
Pearson chi-square	0.4255	0.03896
Shapiro-Wilk	0.5273	3.017e-05
Anderson-Darling	0.3716	0.00346
Cramer-von Mises	0.3681	0.00967

Osserviamo che per maschi p-value $p > \alpha$ per ogni test e per i valori $\alpha = 0.01$ e $\alpha = 0.05$. Questo significa che variabile chol si può vedere come una variabile di densità normale. Per quanto riguarda femmine, si vede che per primi due test vale che $p > \alpha$ per $\alpha = 0.01$ ma non per $\alpha = 0.05$ cioè l'ipotesi di normalità dipende dal livello di significatività. Anche, per test di Shapiro-Wilk, Anderson-Darling e Cramer-von Mises questo non vale perchè $p < \alpha$ per $\alpha = 0.01$ e $\alpha = 0.05$. Nelle figure 6 e 7 si può vedere la curva normale tratteggiata in blu e densità di quantita di colesterolo per pazienti maschi e femmine.

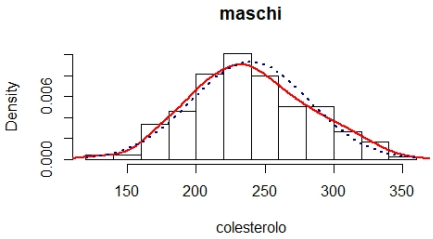


Figura 6

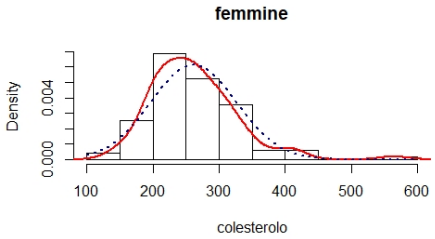


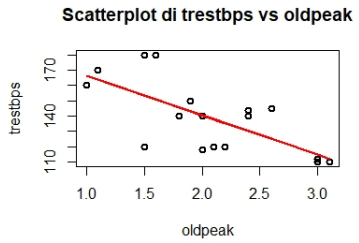
Figura 7

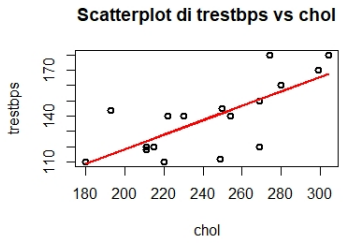
Con Kolmogorov-Smirnov test per un campione verifichiamo che variabile chol ha distribuzione normale poichè p-value è pari a 0.3097. Vogliamo vedere se esiste una differenza statisticamente significativa tra i sessi quando osserviamo il colesterolo. Variabile sex e chol non sono dipendenti. Abbiamo ipotesi H_0 che dire che maschi e femmine non sono significativamente diversi quando guardiamo variabile colesterolo e H_1 che dire che i valori di colesterolo differiscono statisticamente in modo significativo sui campioni di femmine e sui campioni di maschi. Confrontiamo le intere distribuzioni dei due campioni maschi e femmine, con un test non parametrico, il test di Kolmogorov-Smirnov per due campioni indipendenti. Otteniamo p-value $p = 0.01887$. Così si vede che le distribuzioni dei due campioni, maschi e femmine, non sono significativamente diverse in termini di colesterolo solo nel caso $\alpha = 0.01$, ma sono diverse nel caso $\alpha = 0.05$, quando rifiutiamo l'ipotesi H_0 - che due campioni hanno distribuzioni uguali di variabile chol.

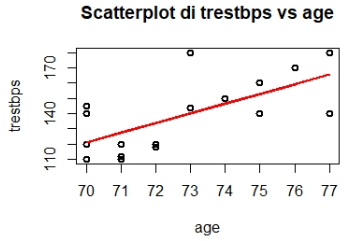
Se guardiamo la variabile oldpeak, possiamo concludere che non ha distribuzione normale per Kolmogorov-Smirnov test per un campione. Se abbiamo le stesse ipotesi nulla e alternativa come in esempio precedente (ma per oldpeak, non per chol), usiamo Wilcoxson sum rank test oppure Mann-Whitney U test e verifichiamo che p-value e 0.0802. Possiamo concludere che in questo caso per $\alpha = 0.01$ e $\alpha = 0.05$ non esiste alcuna differenza statisticamente significativa tra i sessi in termini di oldpeak cioè possiamo accettare l'ipotesi nulla. In altri termini, non è importante quale è sesso di paziente per avere depressione ST indotta.

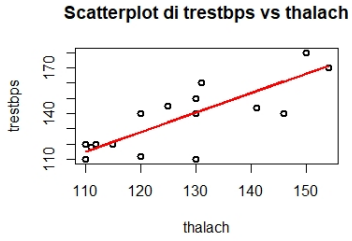
In terzo quesito vogliamo guardare un sottoinsieme di pazienti anziani che hanno più di 70 anni. Vogliamo vedere se, e in che modo, variabile trestbps - pressione sanguigna nei pazienti anziani dipende dalle altre quattro variabili: oldpeak - depressione ST, thalach - battito cardiaco, chol - colesterolo e età. Vogliamo trovare il miglior modello di regressione lineare multipla ridotto togliendo predittori non significativi. Poi troviamo un predittore che è il più significativo e le sue bande di confidenza e di previsione per retta di regressione.

Prima vediamo i modelli di regressione lineare semplice tra trestbps e altre quattro variabili. Così si vedono possibili relazioni tra variabili.









Sappiamo che alcune variabili possono essere non significative. Cerchiamo il miglior modello di regressione lineare multipla ridotto. Lo troviamo dai tre metodi automatici di selezione delle variabili (backward, forward e stepwise) tutti basati sull'AIC - Akaike information criterion. Concludiamo che dobbiamo togliere variabile chol per avere miglior modello ridotto. Nella figura 8 si vede scatterplot di residui vs valori stimati per la variabile trestbps e nella figura 9 normal probability plot dei residui.

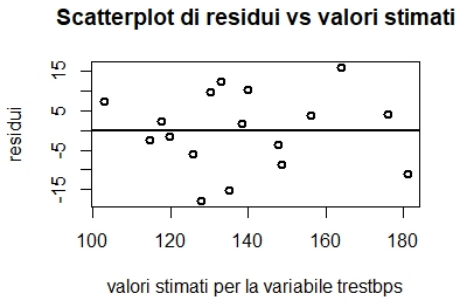


Figura 8

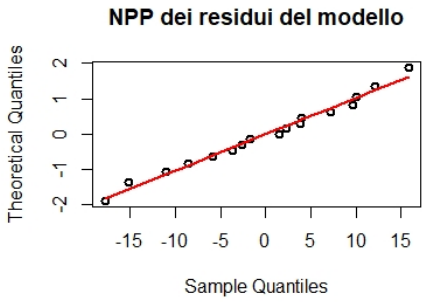


Figura 9

Poi troviamo un predittore che è il più significativo - in questo caso variabile oldpeak - e troviamo intervalli di confidenza per il valore medio di variabile trestbps e intervalli di previsione per il valore di variabile trestbps. In figura 10 si vede il grafico delle stime calcolate. La curva rossa tratteggiata rappresenta limiti di intervallo di confidenza e la curva blu tratteggiata limiti di intervallo di previsione per modello di regressione.

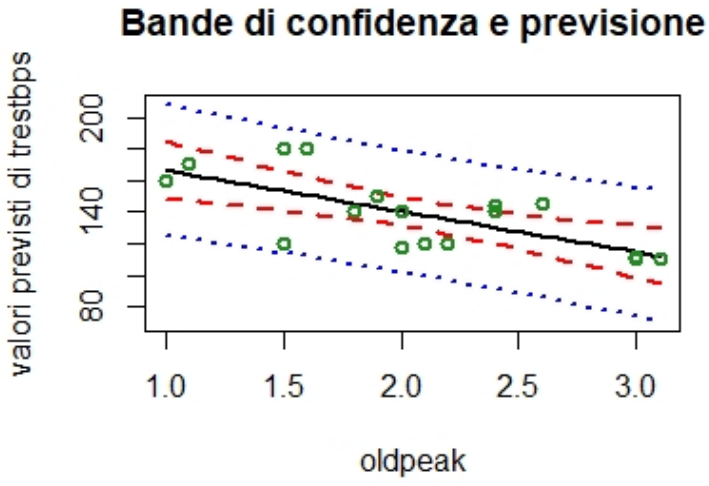


Figura 10