SM2 - Esame 2022/23 - Sessione 1

Dr Giorgio Pioda

2023-02-09

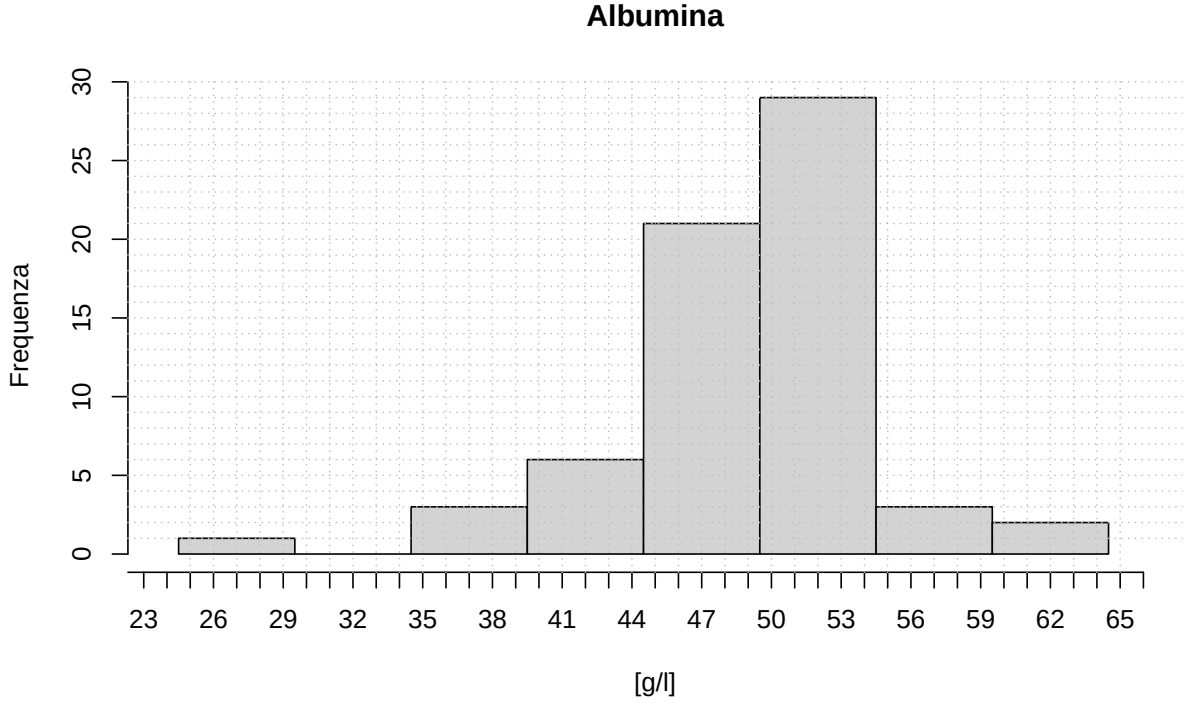
Condizioni generali

Il test è di tipo open book e open web. Si ha accesso alla rete. Si possono consultare pagine web informative di qualsiasi tipo. E invece proibito l'uso di chat, social media e qualsiasi altra forma di **comunicazione attiva**. Qualsiasi violazione ha come conseguenza l'esclusione dall'esame. Tempo 1h 30min max.

La restituzione dell'esame viene fatta su Moodle nell'apposito spazio. Si accettano script .R, file .Rmd, file .zip compressi con l'intero progetto di R, ecc. I comandi devono essere efficaci e generare l'output richiesto. Per le domande teoriche può essere consegnata una versione manoscritta. Il massimo del risultato lo si ottiene raggiungendo **40 punti**.

Esercizio 1 (esercizio su carta) [18p]

E' dato un istogramma di dati con la misurazione dell'albumina. Si consideri che è un istogramma fatto a regola d'arte e che i limiti delle classi tengono conto della precisione dello strumento, che non è particolarmente elevata.

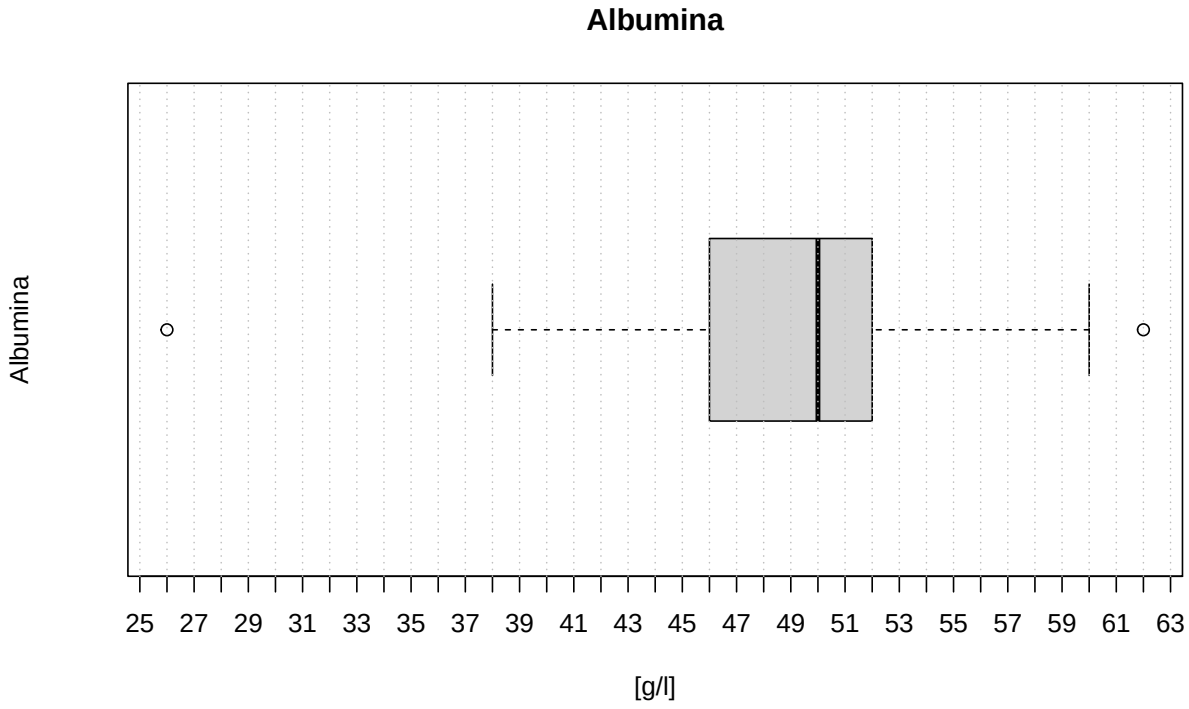


a) Costruire una tabella delle frequenze che contenga minimo e massimo delle classi, valore centrale, frequenza assoluta, frequenza relativa, e frequenza cumulata percentuale. [6p]

b) Calcolare la media ponderata dei valori di albumina [4p]

c) Calcolare la mediana per classi [4p]

Degli stessi dati è anche offerto un boxplot qui sotto



d) Determinare lo scarto interquartile [2p]

e) Determinare per entrambi gli outlier se si tratta di un outlier interno oppure esterno. [2p]

Esercizio 2 (esercizio principalmente su R) [18p + 4p bonus]

Sempre per uno studio sulle proteine seriche si analizza un data set di Gamma globuline. Il file è accessibile con vari meccanismi; è anche presente su Moodle.

```
gamma <- read.csv("http://web.ticino.com/gfwp/stat/dataset/gamma.csv") # Caricamento diretto dal web
# gamma <- read.csv("gamma.csv") # Caricamento dal file salvato in locale
str(gamma)
```

```
## 'data.frame':    1693 obs. of  1 variable:
##  $ gamma: num  9 10.2 8.6 12.3 10.4 8.7 10.6 11 10.8 9.4 ...
```

a) Si determini il numero ideale di classi secondo la formula di Sturges. [2p]

b) Si tracci un istogramma che **tenga in considerazione sia precisione sia la numerosità dei dati**. Va determinata l'ampiezza ideale di una singola classe. L'arrotondamento risulterà essere molto naturale. [4p]

c) Si commenti l'istogramma. [2p]

d) Si calcolino, mediana, primo e terzo quartile, e i quantili al 2.5% e al 97.5% sui dati grezzi. [2p]

e) Si calcoli la media e si commenti la differenza con il valore della mediana. [2p]

f) Ammesso che la distribuzione potesse essere considerata *normale*, si calcoli la soglia di concentrazione che fa in modo che la coda di destra abbia il 5% dei dati. Si può sia procedere a mano con le tabelle, sia con l'apposita funzione integrata di R. [2p]

g) Si calcoli il 95% percentile dei dati. Si confronti questo dato con quello del punto precedente commentando il risultato. [2p]

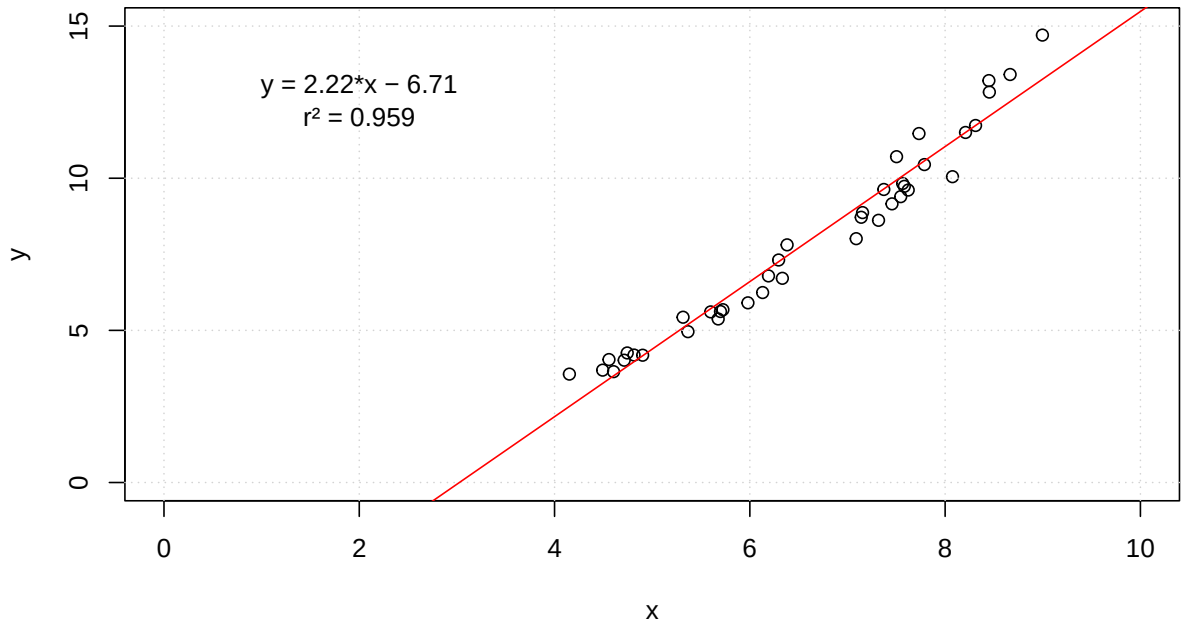
h) Si valuti l'aderenza dei dati al modello normale con una apposita tecnica e un commento sensato. [4p]

i) *(domanda bonus)* Si analizzino i dati con l'uso della kernel density. Si crei un plot della kernel density sovrapposto all'istogramma. Si calcoli infine il valore soglia che tiene alla sua destra il 5% dei dati della kernel density, confrontando il risultato con i valori ottenuti al punto g) e h) [4p]

Esercizio 3 (regressione) [4p + 4p bonus]

Si analizzano i dati bidimensionali qui messi a disposizione. E stampato qui sotto il plot dei dati e la retta di regressione lineare con l'equazione e il parametro r^2 già calcolati.

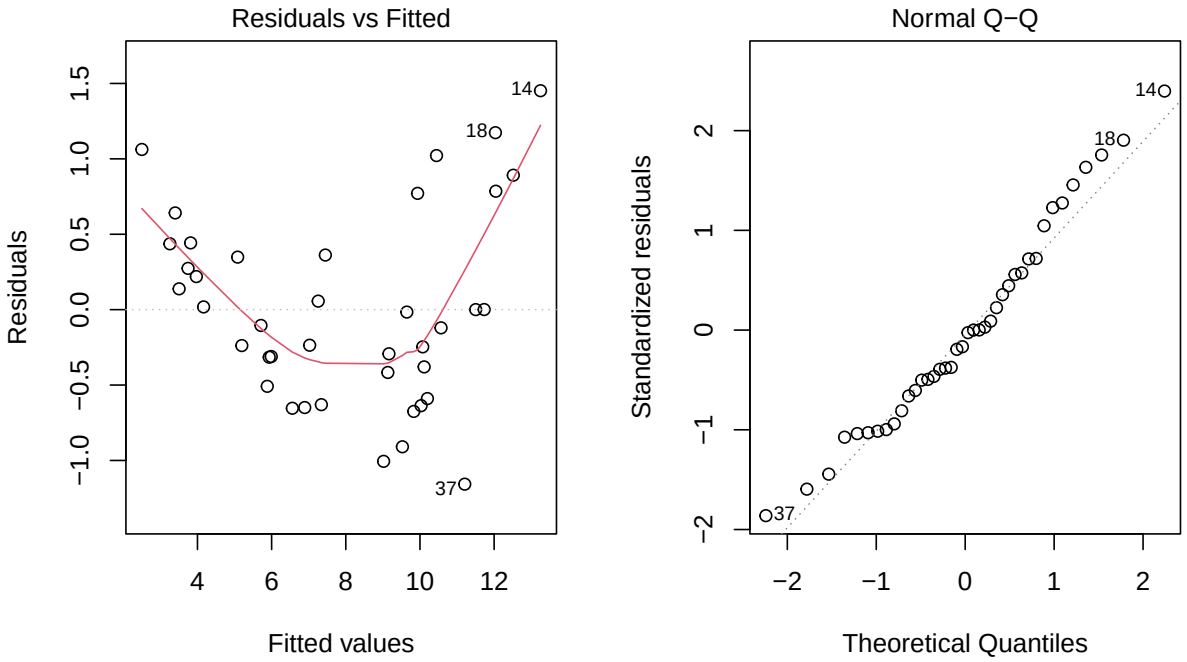
```
regr22 <- read.csv("http://web.ticino.com/gfwp/stat/dataset/regr22.csv")
# regr22 <- read.csv("regr22.csv")
plot(regr22$x,regr22$y,xlim = c(0,10),ylim = c(0,15),xlab="x",ylab="y")
grid()
fit <- lm(y ~ x, data = regr22 )
abline(fit,col="red")
text(2,13,paste0("y = ",signif( coef(fit)[2],3),"*x - ",signif(abs( coef(fit)[1]),3)))
text(2,12,paste0("r² = ",signif(summary(fit)$r.squared,3)))
```



a) Si determini il valore teorico di y per $x = 3.5$ [2p]

b) Si stimi il coefficiente di correlazione di Pearson con i dati presenti sul grafico. [2p]

Qui sotto sono stampati i primi due plot diagnostici della regressione precedente.



c) *(domanda bonus)* Si osservino i due grafici diagnostici; si nota qualcosa? Come si potrebbe migliorare la situazione? Ci si aiuti anche osservando nuovamente il plot precedente. [4p]