Probablity and Statistics - BB503/BB602 - Homework I

Oct 27, 2022

1. [25 pts] Write an R function that detects outliers, and provide a working example.
2. [25 pts] Please calculate **by hand** the 10th, 25th, 50th, 75th, and 90th percentiles for blood cholesterol data given below (Using the "type 2" quantile algorithm):

```
##  [1] 158 171 174 180 181 183 183 184 187 188 191 191 191 192 193 193 193 194 194
## [20] 195 196 196 199 200 200 200 202 204 204 205 206 211 212 213 213 216 218 220
## [39] 221
```

3. The ELISA test is used to screen blood for HIV.

- When the blood contains HIV, it gives a positive result 99% of the time.
- When the blood does not contain HIV, it gives a negative result 95% of the time.

a) [10 pts] Assume the prevalence of HIV is about 0.5% in the adult male population. For an adult male patient who has just tested positive, what is the probability that he has HIV?
b) [10 pts] Assume the prevalence of HIV is about 0.1% in the adult male population. For an adult male patient who has just tested positive, what is the probability that he has HIV?

4. `data/fev data.txt` includes data for a number of patients on:

- age (in years)
- forced expiratory volume, a measure used to measure lung function (in liters)
- height (in inches)
- sex (0 for females, 1 for males)
- smoking status (0 for non-smokers, 1 for smokers)

You can import the dataset directly from the GitHub repo by:

```
data_URL <- "https://raw.githubusercontent.com/egeulgen/BB503_BB602_22_23/main/data/fev_data.txt"
fev_df <- read.delim(data_URL)
```

Import this data set into R and use R to provide answers to the following items:

a) [3 pts] List the variable types (subtypes of discrete/continuous) for age, forced expiratory volume, height, sex, and smoking status.
b) [3 pts] How many patients are there?
c) [3 pts] How many of them were females and how many males? Also provide percentages.
d) [3 pts] How many of them were smokers and how many non-smokers? Also provide percentages.
e) [3 pts] Create a frequency table for age (You can use the R function that you provide for the bonus question). Provide interpretations of the results.
f) [10 pts] Calculate 10th, 25th, 50th, 75th, and 90th quantiles, mean, variance, standard deviation of forced expiratory volume for males and females separately, and interpret the results.
g) [5 pts] Plot a scatter-plot for age in the x-axis and forced expiratory volume in the y-axis and interpret the graph.

Bonus question [15 pts]: Write an R function to create a frequency table for a continuous variable (using regular intervals). Using your function, obtain the frequency tables given below (please pay attention to the square bracket and parenthesis in the tables) for the blood cholesterol data above.

Table 1: right-closed

Class	Frequency	Relative_Frequency	Percentage
[160-170]	1	0.02632	2.6316
(170-180]	3	0.07895	7.8947
(180-190]	6	0.15789	15.7895
(190-200]	16	0.42105	42.1053
(200-210]	5	0.13158	13.1579
(210-220]	7	0.18421	18.4211

Table 2: left-closed

Class	Frequency	Relative_Frequency	Percentage
[160-170)	0	0.00000	0.0000
[170-180)	2	0.05263	5.2632
[180-190)	7	0.18421	18.4211
[190-200)	13	0.34211	34.2105
[200-210)	8	0.21053	21.0526
[210-220]	8	0.21053	21.0526