Biostatistics - MED131/MBG211 - Homework II

Nov 17, 2022

Note: Deadline for submission is December 1st, 2022, 16:00

You'll be working on the prostate cancer data under `data/prostate_cancer.csv`. You may read the directly from the GitHub repository as follows:

```
URL <- "https://raw.githubusercontent.com/egeulgen/MED131_22_23/main/data/prostate_cancer.csv"
prca_df <- read.csv(URL)
```

Because the PSA level is not normally distributed, you'll use the log-transformed PSA levels:

```
# instead of PSA, use logPSA
prca_df$logPSA <- log(prca_df$PSA)

# turn necessary columns into factor
prca_df$Gleason <- as.factor(prca_df$Gleason)
prca_df$invasion <- as.factor(prca_df$invasion)
```

The main aim of collecting this data set was to inspect the association between prostate-specific antigen (PSA) and prognostic clinical measurements in men with advanced prostate cancer. Data were collected on 97 men who were about to undergo radical prostectomies.

The data contains the following variables:

Column	Variable	Description
1	PSA	Serum prostate-specific antigen (PSA) level (mg/mL)
2	vol	Estimate of prostate cancer volume (cc)
3	wt	Prostate weight (g)
4	age	Age of patient (years)
5	BPH	Amount of benign prostate hyperplasia (cm2)
6	invasion	Presence or absence of seminal vesicle invasion: 1 if yes; 0 otherwise
7	penetration	Degree of capsular penetration (cm)
8	Gleason	Pathologically determined grade of disease using total score of two patterns (6, 7, or 8; higher scores indicating worse prognosis)
9	logPSA	log-transformed Serum PSA level (mg/mL)

Please answer the following questions using R. **Notice that in the above code, I've created a new variable `logPSA` by log-transforming `PSA` values. Please use this variable instead of PSA**. For question 2-4, follow the steps of hypothesis testing discussed in class: Please state the hypotheses clearly, state your conclusion. Using the appropriate hypothesis tests (take $\alpha = 0.01$), answer the following questions:

1. [15 pts] Generate a box plot to visually compare the log(serum PSA levels) (`logPSA`) between patients who had seminal vesicle invasion (`invasion = 1`) and who had not (`invasion = 0`). Briefly interpret the plot.
2. [25 pts] Is there any difference between patients who had seminal vesicle invasion (`invasion = 1`) and who had not (`invasion = 0`) with regards the log(serum PSA levels) (`logPSA`)?
3. [25 pts] Is the the log(serum PSA levels) (`logPSA`) higher in patients who had seminal vesicle invasion (`invasion = 1`) compared to who had not (`invasion = 0`)?
4. [35 pts] Are there differences between the groups of patients determined by the pathologically determined grade of disease (`Gleason`) with regards to the log(serum PSA levels) (`logPSA`)? If necessary, also perform post-hoc testing (Tukey test) to determine which grade(s) is significantly different from the others.