Probablity and Statistics - BB503/BB602 - Homework II

Dec 2, 2021

Note: Deadline for submission is 9th December 2021, 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/BB503_BB602_21_22/main/data/prostate_cancer.csv"
prca_df <- read.csv(URL)

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

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 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 prostate-specific antigen 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**. Follow the steps of hypothesis testing discussed in class: Please state the hypothesis clearly, check the assumptions of the chosen test and apply transformation if necessary. Using the appropriate hypothesis tests (take $\alpha = 0.5$), answer the following questions:

1. [20 pts] Is there any difference between patients who had seminal vesicle invasion (`invasion = 1`) and who had not (`invasion = 0`) with regards the serum prostate-specific antigen levels (`logPSA`)?
2. [25 pts] Are there differences between the groups of patients determined by the pathologically determined grade of disease (`Gleason`) with regards to the serum prostate-specific antigen levels (`logPSA`)?
3. [30 pts] Assess any differences in age of patients (`age`) between each pair of pathologically determined grades of disease. Adjust the resulting p values according via the Bonferroni method. Interpret the results.
4. [25 pts] Is there an association between seminal vesicle invasion (`invasion`) and pathologically determined grade of disease (`Gleason`)?