

## Exam 1 Review

A medical center urology group is studying the association between prostate-specific antigen (PSA) levels and a number of clinical measurements on 97 men with cancer about to undergo prostatectomies as treatment. We focus here on the association of PSA levels (this is a non-invasive blood test result) and Gleason scores (cancer severity score ranging from 6 to 8 with higher scores indicating worse prognosis). Note that PSA levels and Gleason scores are higher for this group than they would be for a random sample of men because all of these men have advanced cancers. The data are available as prostate.txt, prostate.xls, prostate.csv on the course website. There are 9 variables for each man:

id = identification number

PSA = prostate-specific antigen level (mg/ml)

cancvol = estimate of prostate cancer volume (cc)

weight = prostate weight (gm)

age = age of patient (yrs)

bph = amount of benign prostatic hyperplasia (cm<sup>2</sup>)

sem = presence/absence of seminal vesicle invasion

capspen = degree of capsular penetration (cm) gleason = grade of disease (6, 7, or 8 for these men)

- (a) What type of study is this, and explain how you know.
- (b) Identify the following as specifically as possible: population, sample, explanatory variable, response variable, confounding variables
- (c) Before examining the results of an ANOVA we should check our model assumptions. This can be done in two ways: by examining the data in each group for normality and constant variance or by running an ANOVA and examining the residuals. Here it is sufficient to take the first approach. Comment on the assumptions for ANOVA.
- (d) A logarithmic or square root transformation often helps with continuous positive measurements such as the PSA measurements. Reconsider the assumptions using log transformed data and square-root transformed data (create new variables by including LOGY=LOG(Y); and SQRTY=SQRT(Y); in the DATA step of your SAS program). Do the transformed data satisfy the model assumptions? Which transformation works best?
- (e) For the remainder of the question use the log transformed data. Is there evidence that median PSA levels vary across groups? Explain.
- (f) Identify the pairs of groups which differ significantly using the Bonferroni approach.
- (g) Define and analyze contrasts to see if there is a:
  - a. linear trend of PSA vs Gleason scores
  - b. quadratic trend of PSA levels vs Gleason scores.

What does it mean if both contrasts are significant?

- (h) Describe your findings on these data in a paragraph.

SAS hints: Use SAS PROC UNIVARIATE and a normal probability plot for each group defined by the Gleason score to assess the assumptions in (a); a scatterplot of PSA vs Gleason score may also be useful. PROC GLM automatically produces residual plots for ANOVA and regression, but these may not be drawn by group.