

**Stat 401, Section F Homework 8**

**Due Date:** Wednesday, October 24

1. For an evaluation of diets used for routine maintenance of laboratory rats, researchers used a completely randomized design to allocate weanling male rats to five different diets (20 rats total). After four weeks, specimens of blood were collected and various biochemical variables were measured. We consider the results for blood urea concentration (mg/dl). The pooled standard deviation is 4.61 and the group means are as follows:

Diet	A	B	C	D	E
$\bar{Y}$	40.0	40.7	32.9	29.6	48.8

A partial ANOVA table looks like:

Source	df	RSS	MS	F
Between diets				10.507
Within diets				
Total		1214.15		

- Please complete the ANOVA table above. What is the  $p$ -value for this test? Explain in a few words what your conclusion is based on this ANOVA table.
  - Produce a family of confidence intervals for each of the five diets, such that the familywise confidence is at least 95%. Use the Bonferroni correction.
  - In part (a), you have probably rejected the null hypothesis that the means for all five groups are equal. In part (b) you have computed confidence intervals that may help you decide which of the means are different. However, the only meaningful way to decide which of the differences are significant is to either construct a confidence interval for each difference in means, or to conduct tests of hypotheses for every pair of group means. How many such tests would you need to perform for this experiment?
  - Use a method for determining significance that keeps the probability of rejecting one or more true null hypotheses no larger than 0.05. Provide either  $p$ -values for all comparisons of pairs of means or confidence intervals for the differences between all pairs of means to support your answer. Also, sketch a line plot that illustrates the significant and nonsignificant differences among the five means.
  - Suppose Tukey's minimum significant difference corresponding to 95% confidence is 8.55. Compute a confidence interval for all the differences in means (i.e. one for each pair). Based on these intervals, which of the differences appear to be significant? Use this HSD to determine which pairs of means differ from one another. Draw a line plot to better illustrate your results. What can you conclude based on this analysis? Which diet leads to the highest mean blood urea? Which leads to the smallest? Can you say anything about the order between the means?
2. Manatees (a.k.a. sea cows) live off the coast of Florida. Many manatees are killed or injured by powerboats. The program `manatee.sas` contains data on  $X$  =the number of Florida powerboat registrations (in 1000s) and  $Y$  =number of manatees killed near Florida. There is one point for each year from 1977 to 1990. Examine and run `manatee.sas`. Use the output to answer the following questions. (Most of your work will be in deciphering the SAS output. SAS will compute almost everything for you.)
- Examine the scatter plot produced by SAS. Is the relationship between  $Y$  and  $X$  roughly linear?
  - Is the correlation negative, zero, or positive?

- (c) Find the exact sample linear correlation between  $X$  and  $Y$  in the SAS output under the section for *The CORR Procedure*. Only one of the numbers in this section could possibly be the correlation between  $X$  and  $Y$ . You should be able to figure out which number is the correct one on your own.
  - (d) Give the equation of the least-squares regression line for predicting the number of manatee deaths as a function of powerboat registrations (use the values from SAS).
  - (e) Predict the number of manatee deaths for a year in which 600,000 powerboats are registered in Florida ( $X = 600$ ). You can do this using your answer above or you can get SAS to do the computation for you as follows:
    - (i) Add another observation to the data set that has a 600 value for  $X$  and a period in place of a  $Y$  value. (Periods indicate missing data in SAS.)
    - (ii) Replace *model*  $Y=X$ ; with *model*  $Y=X/p$ ; This will give you a predicted value for each observation in the data set, including the one you added with  $X = 600$ .
3. An exercise physiologist used skin-fold measurements to estimate the total body fat, expressed as a percentage of body weight, for 19 participants in a physical fitness program. The body fat percentages and the body weights are shown in the table:

Participant	Weight (kg)	Fat (%)
1	89	28
2	88	27
3	66	24
4	59	23
5	93	29
6	73	25
7	82	29
8	77	25
9	100	30
10	67	23
11	57	29
12	68	32
13	69	35
14	59	31
15	62	29
16	59	26
17	56	28
18	66	33
19	72	33

Actually, participants 1-10 are men, and participants 11-19 are women.

- (a) Use function `cor` in **R** to calculate the correlation coefficient between body weights and body fat for
  - i. men
  - ii. women
  - iii. all participants.

The answers may surprise you.

- (b) Draw a scatterplot with the points for men and women denoted by different symbols. After studying the scatterplot, try to sketch by eye the regression line of body fat on body weight. Do this part before you go to part (c), I want you to tell me what you think the line may be without doing the calculations.
- (c) Compute the regression line that you estimated in part (b) and draw it on the scatterplot.
- (d) Using the insight gained from parts (b) and (c), can you explain the discrepancy between the correlation coefficients computed in part (a)? Discuss.