STAT 425 — Spring 2016 — Sections 1UG, 1GR

Homework 6

Please submit your assignment *on paper*, following the Formatting Guidelines for Homework Submission. (Even if correct, answers might not receive credit if they are too difficult to read.) Remember to include relevant computer output.

1. The file `BFHS.dat` contains data from the British Family Heart Study.[1] In each of thirteen different towns, two general medical practices were selected. In each pair of practices from the same town, one was chosen at random to receive an `Intervention`: a subset of males aged 40–59, randomly chosen from the practice's patient registry, received an initial screening and advice and support for leading a healthier lifestyle. The other practice (`ExternalComparison`) did not receive any such intervention. After one year, the males in the intervention groups and all males aged 40–59 in the comparison groups were screened, and their serum cholesterol levels (mmol/l) were recorded. The data set consists of, for each town, the average cholesterol level of the subset of males in the intervention practice, and the average cholesterol level of the males in the comparison practice.

   (a) [2 pts] Create an appropriate R data set for this data. Display a summary of it.

   (b) [2 pts] Perform an *appropriate* t-test for whether there is any mean difference in the cholesterol levels of the intervention and comparison groups.[2] Display a summary of the results. What do you conclude?

   (c) [2 pts] Perform an approximate *randomization test* (based on simulating the t-statistic under re-randomization) for whether there is any mean difference. What is your approximate p-value? (Show the R code you used to produce it.) Does your conclusion change?

   (d) [2 pts] Create a histogram of the *randomization distribution* of the t-statistic, based on your simulation of the previous part. (Show the R code you used to produce it.)

2. The file `Barley1928.csv` contains data from an experiment to examine the effect of five different soil treatments (1: no nitrogen, 2: cyanamide, single dressing, 3: sulphate of ammonia, single dressing, 4: cyanamide, double dressing, 5: sulphate of ammonia, double dressing) on the yield of straw (in quarter pounds). The experiment was conducted in a randomized complete block design (where blocks are designated by the variable `Block`). (Note: The R function `read.csv` might be useful for reading this data into R.)

   (a) [2 pts] Examine the data. How many blocks are there? How many experimental units in total?

   (b) [2 pts] Fit the linear model appropriate for analysis of this experiment. Display a summary of the results. (Remember: Make sure the treatment is a `factor` variable!)

   (c) [2 pts] Produce an ANOVA table for your model. Are treatment effects statistically significant? (Remember to state the p-value.)

[1] From Thompson, Simon G., Pyke, Stephen D., and Hardy, Rebecca J. (1997) "The design and analysis of paired cluster randomized trials: An application of meta-analysis techniques," *Statistics in Medicine*, 16, 2063–2079.

[2] Even though a one-sided test might be more appropriate, you should do a two-sided test.

(d) [2 pts] Produce Tukey simultaneous 95% confidence intervals for all mean differences between pairs of treatments.

(e) [2 pts] According to your Tukey intervals, which pairs of treatments have significantly different means (after adjusting for multiple comparisons)?

3. The file `Spelling1941.csv` contains data from an experiment to compare four different ways of testing the spelling abilities of children: Multiple Choice (MC), Second Dictation (SD), Wrongly Spelled (WS), and Skeleton Word (SW). In the experiment, there was an original dictation test of the words to be spelled, followed by a test using one of the four ways (MC, SD, WS, or SW). The response (`Number`) was "the number of words wrong in the original dictation but correct in the later test." The experiment was conducted in a Latin square design, with different lists of words as one blocking factor and different groups of children as the other blocking factor.

   (a) [2 pts] Examine the data. What are the names of the two blocking factors?

   (b) [2 pts] Display the particular Latin square that was used in this experiment. Identify which factor corresponds to the *rows* and which to the *columns*.

   (c) [2 pts] Fit the linear model appropriate for analysis of this experiment. Display a summary of the results. (Remember to convert variables to `factor` as needed.)

   (d) [2 pts] Produce an ANOVA table for your model. Are treatment effects statistically significant? (Remember to state the p-value.)

   (e) [2 pts] Produce Tukey simultaneous 95% confidence intervals for all mean differences between pairs of treatments.

   (f) [2 pts] According to your Tukey intervals, which pairs of treatments have significantly different means (after adjusting for multiple comparisons)?

4. [GRADUATE SECTION ONLY] Consider the following two possible assignments of treatments A, B, C, and D to a 4×4 square grid of experimental units:

Assignment I

A	B	C	D
C	A	D	B
B	D	A	C
A	D	C	B

Assignment II

A	B	C	D
D	A	B	C
C	D	A	B
B	C	D	A

For each of the assignments (I and II) separately, determine the *probability* of that particular assignment being chosen at random if the design is:

   (a) [2 pts] completely randomized (with 4 each of A, B, C, and D).

   (b) [2 pts] a randomized complete block design with *rows* as blocks.

   (c) [2 pts] a randomized complete block design with *columns* as blocks.

   (d) [2 pts] a Latin square design (with rows and columns as blocks), with the square chosen randomly from among the 576 total distinct 4×4 Latin squares.

Some reminders:

- Unless otherwise stated, all data sets are either automatically available or can be found in either the `alr4` package or the `faraway` package in R.
- Unless otherwise stated, use a 5% level ($\alpha = 0.05$) in all tests.