STA305/1004 - Homework #3

Due Date: March 22, 2016

Due date: Electronic submission on UofT Learning Portal course page by Tuesday, March 22, 2016 at 22:00. NB: e-mail submissions will NOT be accepted.

If you work with other students on this assignment then:

- indicate the names of the students on your solutions;
- your solutions must be written up independently (i.e., your solutions should not be the same as another students solutions).

1. A psychologist is designing an experiment to investigate the effects of four different learning methods on short term memory. Subjects will be shown a series of 20 words after undergoing some training in the learning method that they were assigned. The outcome of the experiment is the total number of words that a subject is able to recall after being trained in one of the learning methods. An equal number of subjects will be randomly assiged to each learning method.

Based on previous research the psychologist estimates that the mean and standard deviation for each method are:

Learning Method	Mean	Standard deviation
1	15	6
2	14.5	4
3	12.5	3.5
4	15.3	3

The psychologist would like to know how many subjects she will require so that her study has 80% power at the 5% significance level.

(a) Use `pwr.anova.test()` in R to calculate the effect sizes that the psychologist can detect if she uses the different variances of the different learning methods. The formula for effect size is

$$f = \sqrt{\frac{\sum_{i=1}^{k} \left(\mu_i - \bar{\mu}\right)^2 / k}{\sigma^2}}.$$

$\bar{\mu} = \sum_{i=1}^{k} \mu_i / k$, and σ^2 is the within group error variance.

For example, for the first effect size assume that the within group variance is 36, for the second effect size assume that the within group variance is 16, and so on. Now, assuming that she can enrol 15 subjects per group, what is the power to detect each effect size at the 5% level? (Hand in your R code and output)

(b) Use simulation to calculate the power of the study using 15 subjects per group assuming that the standard deviations for the four methods are not equal, but are as shown in the table above, and that the distribution of observations in each group is normal. A random sample of size n from a $N(\mu, \sigma^2)$ can be generated in R using the function `rnorm(n,mu,sigma)`. (Hand in your R output and R code)

(c) What does part (b) tell you about the assumption of a common within group variance in calculating power for an ANOVA experiment? Explain.

2. A clinical trial was conducted where patients were randomized to four different treatments. The data is available in the file `q2data.csv`. The outcome is a continuous response y_{ij} the response for the ith subject in the jth treatment group. There are three new treatments in this study and one control treatment. The control treatment is the third treatment ($j = 3$). The main objective of the study is to compare the three new treatments to the control treatment.

NB: The file can be read into R and put into a data.frame using the command

```
q2data <- read.csv("q2data.csv").
```

In this question use the 5% significance level.

(a) What are the averages and standard deviations of each treatment? Plot the distributions of the four treatment groups. Do the distributions look similar or different? (Hand in your R code and output)

(b) Use linear regression to calculate the ANOVA table. What do you conclude from the ANOVA table? (NB: when using linear regresssion to calculate the effects the treatment variable should be specified as a factor `as.factor(trt)`.) (Hand in your R code and output)

(c) Use the model you obtained in part (b) to obtain the appropriate parameter estimates using the treatment contrast (dummy coding) to answer the main objective. In R this can be done using the `contr.treatment()` function. Define the underlying statistical model in terms of dummy variables. Explictly state the dummy variables. Interpret the parameter estimates. Verify the paratemer estimates using the table of means that you obtained in part (a). (Hand in your R code and output)

(d) Obtain the parameter estimates using the Helmert contrast. In R this can be done using the `contr.helmert(4)` function. Explictly state the dummy variables. Define the underlying statistical model in terms of dummy variables. Interpret the parameter estimates. Verify the parameter estimates using the table of means that you obtained in part (a). (Hand in your R code and output)

(e) Which coding scheme do you think makes more sense for evaluating if there is a significant difference between any of the new treatments and placebo.

(f) Which pairs of treatments have a statistically significant difference? Do your results change if you adjust for multiple comparisons using either the Bonferroni or Tukey method? Compare all pairs of treatment means using no adjustement, Bonferroni, and Tukey. If the unadjusted, Bonferroni, and Tukey lead to different conclusions then explain why these methods give different results. Does it make sense to consider all pairs of treatment means given the main objective of this study? (Hand in your R code and output)

3. A chemical engineer studied three variables: temperature ($x1$), pH ($x2$), and agitation rate ($x3$). Each variable has two levels. The dependent variable is the yield of a chemical reaction. The data is shown in the table below and is available in the file `q3data.csv`.

run	x1	x2	x3	y
1	-1	-1	-1	59
2	1	-1	-1	60
3	-1	1	-1	63
4	1	1	-1	62
5	-1	-1	1	63

run	x1	x2	x3	y
6	1	-1	1	64
7	-1	1	1	54
8	1	1	1	59
9	-1	-1	-1	55
10	1	-1	-1	56
11	-1	1	-1	61
12	1	1	-1	58
13	-1	-1	1	63
14	1	-1	1	53
15	-1	1	1	56
16	1	1	1	56

(a) Explain why this a 2^k factorial experiment? What is the value of k. How many replications of each run did the engineer conduct? Explain.

(b) Estimate all the factorial effects and their standard errors by fitting a linear regression model in R. (Hand in your R code and output)

(c) Interpret all the main and interaction effects in the context of this study.

4. A food scientist is interested in studying the effect of two types of flour (call this factor A with levels a_1, a_2) and two amounts of flavouring (call this factor B with levels b_1, b_2) on taste (y). y is measured on scale of 0 to 10 with 0 meaning poor taste and 10 meaning excellent taste. The variance of y is σ^2.

(a) The scientist wants to compare the two types of flour so decides to keep factor B at the low level since he almost always uses B at the low level in most of the products he works on. So there are two treatments: $a_1 b_1$ and $a_2 b_1$. If the scientist obtains four observations for each treatment then what is the variance of the mean difference of the effect of flour (i.e., factor A)? Show your work.

(b) A 2^2 factorial design with factors A and B replicated twice is also conducted. What are the variances of the main effect of A? Show your work.

(c) Another scientist wants to know if the effect of flour is the same when the amount flavouring used is low and high? Which experiment can answer this question: the experiment in (a) or the experiment in (b)? Explain which estimate you would report to the scientist. Explain your reasoning.