STA305/1004 - Homework #3

Due Date: March 20, 2017 by 22:00

- E-mail submissions will NOT be accepted.
- You will be required to submit the assignment via Crowdmark.
- A Crowdmark submission link will be sent to you via your UofT email. Please check your UofT e-mail account for this link. The link will be ready during the week of March 6, 2017.

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. Suppose that twenty independent two-sample t tests are conducted at the 0.05 level, and that all the null hypotheses are true.

(a) What is the distribution of the total number of tests that will be significant? How many tests do you expect would produce a significant result? Briefly explain.

(b) The following R code simulates two independent t-tests based on sample sizes of 100 when H_0 is true and calculates the probaility at least one H_0 is rejected.

```
y1 <-replicate(1000,t.test(rnorm(100,0,1),rnorm(100,0,1))$p.value)
y2 <-replicate(1000,t.test(rnorm(100,0,1),rnorm(100,0,1))$p.value)
sum(y1<=0.05|y2<=0.05)/1000
```

NB: The R code `sum(y1<=0.05 | y2 <=0.05)` means add up all the elements of `y1` or `y2` that are less than or equal to 0.05. So for example the code below checks if each element of `y1` or `y2` is less than or equal to 0.05. If at least one of the elements is less than or equal to 0.05 then it is counted in the sum. Here is a simple example:

```
y1 <- c(0.01,0.06,0.03)
y2 <- c(0.05,0.09,0.07)
y1<=0.05|y2<=0.05
```

```
## [1]  TRUE FALSE  TRUE
```

```
sum(y1<=0.05|y2<=0.05)
```

```
## [1] 2
```

Extend the simulation above to calculate the probability that at least one null hypothesis is rejected among the twenty studies. Does this probability correpond to a type I or type II error? Briefly explain.

(c) Suppose that k independent t-tests are conducted at level α. Show that

$$P\left(\text{at least one } H_0 \text{ rejected}\right) = 1 - (1-\alpha)^k.$$

This is called the experimentwise error rate. Verify the results of part (b) by choosing the appropriate values of α and k.

(d) Does the experimentwise error rate increase or decrease as k increases for a fixed α? Interprete this probability in terms of conducting multiple hypothesis tests.

(e) Consider a study of the effects of a person's sex (e.g., male or female) on salary at a university. In such a case, it would be impossible to assign subjects to be men or women, but a nonrandomized study might be conducted by carefully matching males to females on factors such as age and years of experience. In one such study, fifty variables were recorded for each subject such as promotion, external awards, articles cited, and number of children born. After termination of the study, the male and female salaries were compared on each of these fifty variables, and it was found that there was a "significant difference (at the 1% level)" in the salary between men and women that had a large number of highly cited articles during their career. Explain why there might be a problem with this "significant finding".

2. Suppose that we would like to compare sales from four different versions of the landing page of a company's website. The company would like to test if there is evidence of a difference in sales between different langing pages. Assume that sales on each landing page are normally distributed with means of $\mu_1 = 55, \mu_2 = 58, \mu_3 = 50,$ and $\mu_4 = 60$ (μ_i is the mean sales of landing page $i = 1, \ldots, 4$), $\alpha = 0.05$ and that a reasonable estimate of the error variance is $\sigma^2 = 50$.

(a) What are the null and alternative hypotheses?

(b) Suppose that each page can generate ten sales. Calculate the power of the test directly using the non-central F distribution. Make sure to specify how you calculated the non-centrality parameter and degrees of freedom. Hand in your R output.

(c) What is the effect size that this study is trying to detect? How many observations in each group would be required to detect an effect size that is twice as large, as this effect size, to acheive the same power that you calculated in part (b) (for this calculation you may use one of the power functions built into R). Explain why the required sample size (per group) increases or decreases.

3. (Adapted from Box, Hunter and Hunter) A study was conducted to determine the effects of individual bathers on the fecal and total coliform bacterial populations in water. The variables of interest were the time since the subject's last bath, the vigor of the subject's activity in the water, and the subject's sex. The experiments were perfomed in a 100-gallon polyethylene tub using dechlorinated tap water at 38°C. The bacterial contribution of each bather was determined by subtracting the bacterial concentration measured at 15 and 30 minutes from that measured initially. A replicated 2^3 factorial design was used for this experiment. The data obtained are presented below. (Note: Because the measurement of bacterial populations in water involves a dilution technique, the experimental errors do not have constant variance. Rather, the variation increases with the value of the mean.) Perform analysis using a logarithmic transfomation of the data (the natural logarithm in R is obtained from the function `log()`).

The data can be read into R using the code:

```
prb0506 <- read.table(file = "prb0506.dat",header = T)
```

run	x1	x2	x3	y1	y2	y3	y4
1	-1	-1	-1	1	1	3	7
2	1	-1	-1	12	15	57	80
3	-1	1	-1	16	10	323	360
4	1	1	-1	4	6	183	193
5	-1	-1	1	153	170	426	590
6	1	-1	1	129	148	250	243
7	-1	1	1	143	170	580	450
8	1	1	1	113	217	650	735
9	-1	-1	-1	2	4	10	27
10	1	-1	-1	37	39	280	250
11	-1	1	-1	21	21	33	53
12	1	1	-1	2	5	10	87
13	-1	-1	1	96	67	147	193
14	1	-1	1	390	360	1470	1560
15	-1	1	1	300	377	665	810
16	1	1	1	280	250	675	795

Code	Name	Low Level	High Level
x_1	Time since last bath	1 hour	24 hours
x_2	Vigor of bathing activity	Lethargic	Vigorous
x_3	Sex of bather	Female	Male

Code	Name
y_1	Fecal coliform contribution after 15 minutes (organisms/100mL)
y_2	Fecal coliform contribution after 30 minutes (organisms/100mL)
y_3	Total coliform contribution after 15 minutes (organisms/100mL)
y_4	Total coliform contribution after 30 minutes (organisms/100mL)

(a) Briefly explain why this is a 2^3 factorial design.

(b) Calculate the factorial effects (main and interaction effects) on total coliform populations after 15 minutes. Hand in your R output. Interpret the main effect of x_1, and the interaction between x_1 and x_3.

(c) Let w_1, w_2 be the results from two individual runs under the same conditions from this 2^3 factorial design. For example, the responses for Fecal coliform contribution after 15 minutes (y1) under the condition x_1, x_2, x_3 set at -1 -1 -1 are 1 and 2 respectively. (runs 1 and 9).

Show that the estimated variance $s_i^2, i = 1, ..., 8$ at the i^{th} conditions is:

$$s_i^2 = \frac{(w_1 - w_2)^2}{2}.$$

(d) Use the result in part (c) to show that the variance of any of the factorial effects in a 2^3 factorial design with replicated runs is

$$V(effect) = \left(\frac{1}{8} + \frac{1}{8}\right) s^2,$$

where $s^2 = \frac{\sum_{i=1}^{8} s_i^2}{8}$.