STAT 135 CONCEPTS OF STATISTICS
HOMEWORK 6

Assigned July 27, 2021, due August 3, 2021

This homework pertains to materials covered in Lecture 15, 16 and 17. The assignment can be typed or handwritten, with your name on the document, and with properly labeled input code and computer output for those problems that require it. If not specified, **please try to perform the hypothesis testings from scratch** without using the built-in `anova` function in `R` to obtain full credit. If you choose to collaborate, the write-up should be your own. Please show your work! Upload the file to the Week 6 Assignment on bCourses.

Note in this homework, we use the following abbreviations: Analysis of Variance (ANOVA), confidence intervals (CIs).

Problem 1. For any numbers y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, k$,

(1) prove that

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{k} n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2,$$

where $\bar{y}_{i\cdot} = n_i^{-1} \sum_j y_{ij}$, $\bar{y}_{\cdot\cdot} = \sum_i n_i \bar{y}_{i\cdot} / \sum_i n_i$;

(2) prove that for any $\mu \in \mathbb{R}$,

$$\sum_{i=1}^{k} n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{k} n_i (\bar{y}_{i\cdot} - \mu)^2 - n(\bar{y}_{\cdot\cdot} - \mu)^2.$$

Problem 2. Prove the sums of squares identity for the two-way layout:

$$SS_{\text{Tot}} = SS_A + SS_B + SS_{AB} + SS_E.$$

Also show that the sums of squares on the right-hand side of the above equation are mutually independent under the two-way ANOVA model assumption in Lecture 16.

Problem 3. We have seen the Bonferroni corrections for the simultaneous confidence intervals a couple of times but we never explained why that works. Now let's take a look.

(1) Consider two events A and B. Prove that $P(A \cap B) \geq P(A) + P(B) - 1$. Demonstrate using a Venn diagram will suffice.

(2) Consider m events $A_1, \ldots, A_m$. Show that the probability that they happen simultaneously can be bounded as follows:

$$P\left(\bigcap_{i=1}^{m} A_i\right) \geq \sum_{i=1}^{m} P(A_i) - (m-1).$$

(3) Suppose we have m confidence intervals for m different parameters. Each confidence interval statement can be construed as an event. Show that to have an overall confidence level of $1-\alpha$, each individual confidence interval can simply be adjusted to the level of $1-\alpha/m$.
(4) Compared to Tukey's method, does the Bonferroni procedure produce more conservative or more confident CIs?

Problem 4. One researcher collected data to see whether there exists difference in the energy use of four gas ranges for seven menu days. He was in the process of performing the one-way ANOVA analysis when his laptop crashed and all his data were erased. Help him restore the lost information and complete the following ANOVA table for him:

Source	**Sum Sq**	**Df**	**Mean Sq**	F **statistic**	p-**value**
Treatment	64.42	?	?	8.98	?
Residual	?	?	2.39		
Total	?	20			

Problem 5. To determine diet quality, male weanling rats were fed diets with various protein levels. Each of 18 rats was randomly assigned to one of three diets, and their weight gain in grams was recorded in Table 1.

Diet protein level		
Low	Medium	high
3.89	8.54	20.39
3.87	9.32	24.22
3.26	8.76	39.91
2.70	9.30	22.78
3.82	10.45	26.33
3.23	8.94	
	10.37	

Table 1. Weanling rat diet data set

(1) Calculate the sums of squares and fill out a one-way ANOVA table;
(2) Denote the unique effects of the protein levels by α_1, α_2 and α_3. Test

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$$

with significance level $\alpha = 0.05$.
(3) Derive the Bonferroni simultaneous 95% CIs for α_1, α_2 and α_3. Do these CIs overlap with each other? What information can we learn if they don't overlap?

Problem 6. In Lecture 15, we derived the confidence intervals for the paired difference $\alpha_i - \alpha_r$ using the fact that

$$\frac{(\bar{Y}_{i\cdot} - \bar{Y}_{r\cdot}) - (\alpha_i - \alpha_j)}{\sqrt{S_{ir}^2 \left(\frac{1}{n_i} + \frac{1}{n_r}\right)}} \sim t_{n_i+n_r-2}, \text{ for any pair } i \neq r.$$

Now let's generalize the above CIs and look at a *contrast* which is defined to be $\sum_{i=1}^k t_i \alpha_i$ in which $\sum_{i=1}^k t_i = 0$. You can see that $\alpha_i - \alpha_r$ is an specific example of a contrast.

(1) Prove that

$$\frac{\sum_{i=1}^k t_i \bar{Y}_{i\cdot} - \sum_{i=1}^k t_i \alpha_i}{\sqrt{MS_W \sum_{i=1}^k t_i^2 / n_i}} \sim t_{n-k},$$

in which $MS_W = (n-k)^{-1} \sum_{i=1}^k \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$.

(2) For the trout toxin example of Lecture 15, derive the Bonferroni simultaneous 95% CIs for the following contrasts:

$$\alpha_1 - \alpha_2,\ \alpha_1 - 2\alpha_2 + \alpha_4,\ 3\alpha_3 - \alpha_1 - 2\alpha_2,$$
$$5\alpha_2 - 4\alpha_3 - \alpha_4,\ 2\alpha_2 - \alpha_1 - \alpha_3.$$

(3) Scheffé (1959) coined a quite elegant approach to deriving simultaneous CIs on **all** contrasts. This procedure is valid for any number of contrast CIs concurrently via specifying a larger critical value $c = (k-1)F_{k-1,n-k}(\alpha)$; that is, the overall probability is $1 - \alpha$ that

$$\sum_{i=1}^k t_i \alpha_i \in \sum_{i=1}^k t_i \bar{Y}_{i\cdot} \pm c\sqrt{MS_W \sum_{i=1}^k t_i^2/n_i}$$

simultaneously for all $\mathbf{t} = (t_1, \ldots, t_k)$ such that $\sum t_i = 0$.

Calculate the Scheffé simultaneous 95% CIs for the contrasts in (2), and visualize the two sets of CIs side by side to compare.

Problem 7. In Section 12.2, there is an example data from Kirchhoefer (1979), who studied the measurement of chlorpheniramine maleate in tablets. Now we look at the measurement data from another manufacturer. (You saw this dataset in Lab 11. We are only considering the first three columns here.)

Derive three sets of simultaneous CIs for all pairwise differences using the Bonferroni method, Tukey's method and Scheffé's method respectively. Discuss the widths of the CIs calculated from different methods. Which method is the least accurate and why is that?

Problem 8. A researcher ran a two-factor experiment to compare 3 different species (Species A, B and C) under different fertilizer treatments in a greenhouse. He assigned combinations of fertilizer and species levels to 72

Lab1	Lab2	Lab3
4.15	3.93	4.1
4.08	3.92	4.1
4.09	4.08	4.05
4.08	4.09	4.07
4.01	4.06	4.06
4.01	4.06	4.03
4	4.02	4.04
4.09	4	4.03
4.08	4.01	4.03
4	4.01	4.06

pots to have 6 replications in the greenhouse. This would be a referred to as 3×4 factorial treatment design.

The data is in Table 2 (You can read this data set via copying and pasting in `R`):.

(1) Make a Treatment Mean Plot to visually examine whether there is interaction between the two factors;
(2) Calculate the sums of squares and fill out a two-way ANOVA table. Is your table the same as the table output from `anova()` in `R`?
(3) Denote the differential effects of species as α_1, α_2 and α_3. Test
$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$$
with significance level $\alpha = 0.05$.
(4) Denote the differential effects of fertilizer treatments as β_1, β_2, β_3 and β_4. Test
$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
with significance level $\alpha = 0.05$.
(5) Denote the interactive effects between species and fertilizer treatments as δ_{ij}, $i = 1, 2, 3,\ j = 1, 2, 3, 4$. Test
$$H_0 : \delta_{ij} = 0$$
with significance level $\alpha = 0.05$.

Problem 9. Try to develop a parametrization for a balanced three-way layout. Define main effects and two-factor and three-factor interactions, and discuss their interpretation. What linear constraints do the parameters satisfy?

Problem 10. Pottery shards are collected from four sites in the British Isles: Llanedyrn (L), Caldicot (C), Isle Thorns (I) and Ashley Rails (A).

	Fertilizer Treatment			
	Control	**F1**	**F2**	**F3**
Species A	21.0	32.0	22.5	28.0
	19.5	30.5	26.0	27.5
	22.5	25.0	28.0	31.0
	21.5	27.5	27.0	29.5
	20.5	28.0	26.5	30.0
	21.0	28.6	25.2	29.2
Species B	23.7	30.1	30.6	36.1
	23.8	28.9	31.1	36.6
	23.7	34.4	34.9	37.1
	22.8	32.7	30.1	36.8
	22.8	32.7	30.1	36.8
	24.4	32.7	25.5	37.1
Species C	25.1	28.4	22.8	32.8
	22.6	26.4	23.2	34.3
	24.5	27.8	26.4	33.3
	23.7	26.7	23.8	31.9
	22.6	25.3	25.4	32.6
	23.9	25.9	22.7	30.6

TABLE 2. Greenhouse data set

Each pottery sample was returned to the laboratory for chemical assay. In these assays the concentrations of five different chemicals were determined: Aluminum (Al), Iron (Fe), Magnesium (Mg), Calcium (Ca) and Sodium (Na). The data set can be downloaded from the *data_sets* directory under bCourses.

Try to follow the example from Lecture 17 to perform MANOVA in `R` to test

$$H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_3 = \boldsymbol{\alpha}_4 = \mathbf{0}$$

with $\alpha = 0.05$, in which $\boldsymbol{\alpha}_i$'s are the unique effects of different sites.