STA 106 Winter 2018
Homework 2 - Due Friday, Jan 26^{th}

Book Homework (does not require R)

1. A rehabilitation center researcher was interested in examining the relationship between physical fitness prior to surgery and the time (in days) for successful completion of physical therapy. The fitness levels were measured as "Below Average" (group I), "Average" (group II), or "Above Average" (group III). Summary statistics follow:

	I	II	III	Overall
Sample mean	38	32	24	32
Sample Std. Dev	5.48	3.46	4.43	6.88
Sample Size	8	10	6	24

Data Source: "Applied Linear Statistical Models", Kutner, Nachtsheim, Neter, & Li.

   (a) Estimate all values of μ_i.
   (b) Estimate the expected value of Y_{ij} for every value of i, j.
   (c) Find SSTO, SSE, SSA
   (d) Find $d.f.\{SSTO\}$, $d.f.\{SSA\}$, $d.f.\{SSE\}$
   (e) Find MSTO, MSE, MSA.

2. Continue with problem 1.
   (a) Based on the sample values, would you conclude that the variance for each group is constant? Why or why not.
   (b) What group appears to have the fastest recovery time? Do you believe it is statistically significantly faster than the group that has the slowest recovery time? Without a confidence interval or hypothesis test, explain your answer.
   (c) State the null and alternative for testing if the population means are all equal.
   (d) Find the value of the test-statistic, and the approximate the p-value for the test in (c).
   (e) State your conclusion in terms of the problem if $\alpha = 0.01$.

3. A consumer organization studied the effect of age of automobile owner on cash offer for a used car (in hundreds of dollars). The age groups were "Young" (Y), "Middle"' (M), and "Old" (O). Summary statistics follow:

	M	O	Y	Overall
sample mean	27.75	21.42	21.50	23.56
sample std. dev	1.29	1.68	1.73	3.38
sample size	12	12	12	36

Data Source: "Applied Linear Statistical Models", Kutner, Nachtsheim, Neter, & Li.

   (a) Estimate the value of $\mu_i - \mu.$ for all values of i.
   (b) Describe what the values in (a) represent in words.
   (c) Find SSTO, SSE, SSA
   (d) Find $d.f.\{SSTO\}$, $d.f.\{SSA\}$, $d.f.\{SSE\}$
   (e) Find MSTO, MSE, MSA.

4. Continue with problem 2.
   (a) Without using a confidence interval or hypothesis test, does it appear that all groups differ significantly from each other? Explain.
   (b) State the distribution of $\bar{Y}_{2\cdot}$, being as specific as you can. Assume $i = 2$ corresponds to the "Young" group.
   (c) State the null and alternative hypothesis for testing if all population group averages are equal, and the corresponding test-statistic.
   (d) Estimate the p-value and interpret it in terms of the problem.
   (e) State your conclusion in terms of the problem, assuming $\alpha = 0.01$.
   (f) What two group means are most likely to differ from each other (if any)?

5. Prove the following identities. You may assume the Single Factor ANOVA model (and its assumptions) hold.
   (a) $\sum_j e_{ij} = 0$ for any i.
   (b) $\sum_i \sum_j e_{ij} = 0$
   (c) $\sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})(e_{ij}) = 0$
   (d) $E\{\bar{Y}_{\cdot\cdot}\} = \mu_\cdot$, where $\mu_\cdot = \sum_i \frac{n_i \mu_i}{n_T}$
   (e) $\sigma^2\{Y_{ij}\} = \sigma^2_\epsilon$

6. Answer the following questions with TRUE or FALSE. You should explain your answer as practice for an exam.
   (a) In general, for a particular population, the larger n_T, the larger SSTO is.
   (b) If we reject the null hypothesis, we may always conclude there is evidence to suggest that all population means are different from each other.
   (c) To reduce the probability of a Type I error, we would want to use the lowest value of α possible.
   (d) It is possible for SSE to equal 0.

R Homework (requires some use of R)

I. Online you will find the file "Cancer.csv". The csv file has the following columns:

Column 1. `Survival`: The survival time of the patient in days

Column 2. `Organ`: The organ where cancer was present - `Stomach`, `Bronchus`, `Colon`, `Ovary`, `Breast`

Data Source : From the article "Supplemental Ascorbate in the Supportive Treatment of Cancer: Reevaluation of Prolongation of Survival Times in Terminal Human Cancer" by Ewan Cameron and Linus Pauling, Proceedings of the National Academy of Sciences of the United States of America, Vol. 75, No. 9 (Sep., 1978), pp. 4538-4542.

(a) Create a group box plot of Survival by Organ type. Does there appear to be significant differences in the groups? Explain your answer.

(b) Find the sample averages of Survival by Organ type. Do you believe you would reject the null hypothesis of Single Factor ANOVA based on these values? Explain.

(c) Do you believe the standard deviations of each population are equal? Explain.

(d) What level of α would you suggest if concluding that the true average survival time was equal when in reality it was not, would be considered the most severe error?

II. Continue with problem I.

(a) Find the value of SSTO, SSA, SSE.

(b) Find the value of MSTO, MSA, MSE.

(c) Find the value of the test-statistic, and the corresponding p-value.

(d) State your conclusion in terms of the problem if $\alpha = 0.05$.

III. Online you will find the file "Hawks.csv". The csv file has the following columns:

Column 1. `Wing`: Length (in mm) of primary wing feather from tip to wrist it attaches to

Column 2. `Species`: CH=Cooper's, RT=Red-tailed, SS=Sharp-Shinned

The goal is to assess if these hawks can be easily distinguished by the length of their wing feathers.
Source: Professor Bob Black at Cornell College

(a) Find the standard deviation by groups.

(b) Make histogram of wing feather length by group.

(c) Do you believe that the assumption of equal variance by group is met? Explain your answer.

(d) What group appears to have the largest wing feather length (if any)?

(e) What group appears to have the smallest variance (if any)?

IV. Continue with problem II.

(a) Find the value of SSTO, SSA, SSE.

(b) Find the value of MSTO, MSA, MSE.

(c) Find the value of the test-statistic, and the corresponding p-value.

(d) State your conclusion in terms of the problem if $\alpha = 0.05$.

(e) Interpret your p-value in terms of the problem.