GENOME 560, Spring 2018
Problem Set #5

(Due June 7th 11:59pm)

---

1. **[30 points] Partitioning Total Variation**

   In class, we discussed that the total variation (SST) is partitioned into variation between groups (SST_G) and variation within groups (SST_E).

   Prove that the sum of squared deviations about the grand mean across all N ($= \sum_{i=1}^{K} n_i$) observations (SST) is equal to the sum of SST_G and SST_E. SST, SST_G and SST_E are defined as:

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2, \tag{1}$$

$$SST_G = \sum_{i=1}^{K} n_i (\bar{x}_{i.} - \bar{x}_{..})^2, \tag{2}$$

$$SST_E = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2. \tag{3}$$

2. **[30 points] One-Way ANOVA Table**

   Complete the ANOVA table above from the values given. The sample size is $N = 20$. (Hint: See slide 34 of lecture note 18.)

Source of Variation	**Sum of Squares**	**df**	**Mean Square**	**F**
Group	56.7			
Error		14	13.5	
Total				

3. **[40 points] Two-Way ANOVA Table**

   Complete the two-way ANOVA table for the example problem we discussed in class today. The data table is on slide 18 of lecture note 19.

Source of Variation	Sum of Squares	df	Mean Square	F
Gender				
Genotype				
Gender-Genotype				
Error				
Total				

4. **[70 points] Two-Way ANOVA**

   In this exercise, we will examine the influence of the stress reduction method and age on the stress level. The dataset contains a hypothetical sample of 27 participants who are divided into three stress reduction treatment groups (mental, physical, and medical) and three age groups (young, mid, and old). The stress reduction values are represented on a scale that ranges from 0 to 10. This dataset can be conceptualized as a comparison between three stress treatment programs – one using mental methods, one using physical training, and one using medication – across three age groups. The stress reduction values represent how effective the treatment programs were at reducing participant's stress levels, with higher numbers indicating higher effectiveness. Note that the numbers in this dataset are not from real studies. The dataset can be loaded with the code:

   ```
   > dat <- read.table(file="http://www.cs.washington.edu/homes/suinlee/
   genome560/data/stress_reduction.txt")
   ```

   Two-way ANOVA determines how the stress reduction value is affected by two factors – treatment and age. Perform a two-way ANOVA on this dataset using the R.

   (a) **[10 points]** We learned that we need to first test whether there is a significant interaction between the factors. Is there an interaction between the treatment and age (i.e., p-value$<$ 0.05)? What is the p-value of the interaction term? If you think that there is an interaction, solve part (b) and (c). Otherwise, solve part (d) and (e).

   (b) **[15 points]** If an interaction is significant, we conclude that both effects are important. Does the effect of one factor seem to be "masked" by the interaction? Which factor is more masked?

   (c) **[15 points]** To better understand how the effect of each factor is masked by the interaction, we learned that we need to perform a one-way ANOVA for each level of the other factor. Perform the one-way ANOVA for the factor whose effect seems to be less masked for each level of the other factor. Copy and paste the output of ANOVA. Provide your R script.

(d) **[10 points]** If an interaction is not significant, we can then test for significance of the main effects separately, again using an one-way ANOVA. Report the result of an one-way ANOVA by copying and pasting the relevant parts of the results. What are the p-values?

(e) **[20 points]** There are 3 levels in each factor. We learned that when there are more than two levels, we can better understand the mean differences among the levels, through pairwise comparisons of the levels. For the factor(s) that has a significant main effect (i.e., p-value < 0.05), perform three separate one-way ANOVA for three pairs of levels, for example, a 'young' and 'mid' pair, a 'mid' and 'old' pair, and a 'young' and 'old' pair.

5. **[30 points] Comparison among Different Correction Methods**

We are given the genotype data from 334 mouse individuals produced by the backcross experiment. The genotype data measure the genotype values of 1333 genetic markers for each mouse, and the phenotype data measure the normalized blood cholesterol levels. The genotype data have binary values, because the mice were produced from the backcross experiment. Given these data, we want to find the quantitative trait loci (QTLs) that contribute to elevated cholesterol level.

For each marker, we will perform the t-test to measure the significance of the the differences in the phenotype between the mice that have genotype '0' and those that have genotype '1'. (Hint: You need to define a function to perform the t-test that takes both the genotype data for each marker and phenotype as input.)

The genotype and phenotype data can be downloaded from:
`https://sites.google.com/a/cs.washington.edu/genome560-spr18/homework`.

(a) **[10 points]** Generate a set of p-values each of which means the significance of the association between each marker and the phenotype. Plot the histogram of the 1333 p-values. Provide the code. How many markers have p-value smaller than 0.05?

(b) **[10 points]** Apply the Bonferroni correction method, Holm method, and B-H FDR correction method to the set of p-values you obtained in part (a). With the significance level of 0.05 and 0.01, how many markers are considered signifiant after correction in each of the three methods?

(c) **[10 points]** Plot the histograms of the corrected p-values in each of the three methods. Show the four histograms in one figure.