Statistics 416: Statistical Design and Analysis of Gene Expression Experiments
Homework #1, Spring 2017
Due: in class on 02/07/17

1. Suppose researchers were interested in studying the effects of different fertilizer amount (Low Nitrogen and High Nitrogen) on gene expression in two different genotypes (one energy line and one grain line) of sorghum. For each genotype, there were six pots of one-week-old seedlings available, and each pot held two seedlings. For each genotype, the researchers randomly assigned three pots to high nitrogen treatment (H) and the remaining three pots to low nitrogen (L) treatment. After 2 weeks, the sorghum roots of each seedling were obtained and one RNA sample was prepared for each seedling. Hence, there were a total of 24 RNA samples. One library was prepared for each RNA sample and sequenced using Illumina HiSeq platform.
   i. Name the treatment factors considered in this experiment.
   ii. Name the levels of each treatment factor.
   iii. Do we have a full factorial treatment design for this experiment?
   iv. What are the experimental units in this experiment? How many experimental units are used in this experiment? (If there are two or more types of experimental units, specify the number of experimental units for each type.)
   v. What are the observational units in this experiment?
   vi. Does this experiment involve blocking? If so, name the blocks.
   vii. Is this experiment best described as a completely randomized design, randomized complete block design, split-plot design, incomplete block design, or Latin square design? Justify your answer.

2. For the experiment described in question 1, the sequencing was done using multiplexing platform with 8 samples per lane. The 8 samples for each lane include one pot (both seedlings) from each genotype and each nitrogen condition. Three lanes on one flowcell was used for this RNA-seq experiment. The sequence data were aligned to the sorghum reference genome, and a count table with one row for each gene and one column for each of the 24 samples was generated.

   i. Identify the potential sources of variability in this RNAseq count dataset for this nitrogen-genotype experiment, i.e., what contributes to the differences across the 24 counts for a given gene?
   ii. For each source of variability, categorize whether it is a technical variability, biological variability, treatment effect under study, or some other kind of variability such as block effect.
   iii. Are biological replicates involved in this experiment? Are technical replicates involved in this experiment?

3. There was one experiment to examine gene expression levels in liver of mouse. Suppose we got a RNA sample from one mouse, and we sequenced half of the sample on lane 1, and the other half on lane 2 of the same flow cell. From lane 1, we obtained 10 million total mapped reads, with 50 reads mapped to gene *g*. For lane 2, we have 8 million total mapped reads with 25 reads mapped to gene *g*. Use Fisher's exact test to check whether there is any significant

lane effect in the mapping rate for this gene. You may use the code I posted to start. Include the following in your answer:

i. define your parameter(s),

ii. write down the hypotheses in terms of the parameter(s) you defined,

iii. present your R output together with your R code, and

iv. state clearly your conclusion based on the result from iii. For simplicity, you do not need to consider the control of multiple testing error here.

4. Similar to question 3, one experiment examined gene expression levels in liver of mice. Suppose we got an RNA sample from one mouse, and we sequenced the same sample on four lanes of a flow cell. The count of reads mapped to a gene (gene 4 in the dataset hwk1_4.csv) and the total count for each lane is given below:

	lane.1	lane.2	lane.3	lane.4
gene 4	54	67	56	74
total	615878	615867	617439	739028

Use Goodness-of-Fit test to check whether the model, Poisson($C_j\mu$), fits the data for this gene, where j denotes lane j and C_j denotes the total count for lane j and μ denotes the mapping rate of the gene. Include the following in your answer:

i. Give the null hypothesis and the alternative hypothesis.

ii. Use hand calculation (give the formula and then plug in the numbers) to obtain the maximum likelihood estimate (MLE) for the mapping rate (μ) of this gene while assuming the mapping rates are the same across all four lanes. Verify your answer with R.

iii. Use hand calculation (provide the formula and then plug in the numbers) to compute the value of the χ^2 test statistic. Verify your answer with R.

iv. Give the degrees of freedom (df) for the χ^2 goodness-of-fit test.

v. Calculate the p-value based on your test statistic and the χ^2 distribution with appropriate df.

vi. State clearly your conclusion based on the result from (v). For simplicity, you do not need to consider the control of multiple testing error for this test.

5. A dataset (hwk1_4.csv) stores counts for all 10,000 genes obtained from the experiment described in question 4. Apply the goodness-of-fit test for each gene, and provide a qqplot to check whether Poisson models fit the dataset well. In your homework, (i) specify your null hypothesis and alternative hypothesis for gene g, (ii) attach your qq-plot, and (iii) answer the question whether the Poisson models fit well or not.

Hint1: you may either check whether the test statistics follow a χ^2 distribution with appropriate df or whether the p-values follow a Uniform(0,1) distribution.

Hint2: Using functions on matrices directly (such as the function `apply`) will be fast. But if that seems difficult to program, you may use a `for` loop to get the test statistics and/or the p-value for each gene. The following example of `for` loop calculates the mean for each row of a data matrix **`X`** and stores the means in the vector `avg`.

```
avg = rep(0,length(X[,1]))
for (i in 1: length(X[,1]))
   {
   Y = X[i,]
   avg[i] = mean(Y)  # or use only a single line in { }: avg[i] = mean(data[i,])
   }
```