Lab #4: Statistical Models with Inappropriate Assumptions

Send what you have to [afodor@uncc.edu](mailto:afodor@uncc.edu) by March 3rd

Schedule:

Thursday Feb 24th (this lab)

March 3rd (you will have the study guide for the mid-term; lab will be for review)

March 8th, March 10th: Spring break, no classes.

Midterm: Tuesday, March 15th class time (1:00 pm)

-----------

Download this dataset:

<http://afodor.github.io/classes/stats2015/longitdunalRNASeqData.zip>

(right-click in the browser and say “save as…”)

Each row in the spreadsheet represents a different gene in an RNA-seq experiment. The samples are E. Coli derived from two different mice under identical conditions (i.e. biological replicates).

(1) Read the dataset into R using commands something like...

setwd("C:\\classes\\Advanced\_Stats\_Spring2022\\Lab4\_HW1")

myT <- read.table("nc101\_scaff\_dataCounts.txt",header=TRUE,row.names=1)

(2) On a log10-log10 scale, show a plot of the counts for the two samples “D2\_01” and “D2\_02”.

Qualitatively, do the biological replicates appear to have similar patterns of gene expression?

(3) On a log10-log10 scale, plot the variance of all genes (across all genes) vs. the mean (across all genes) with a red line on your graph representing the identity line. Does the mean equal the variance for these samples?

(The commands apply(myT, 1, mean) and apply(myT, 1, var) will be of some use).

(4) Consider the first gene in the spreadsheet (e.g. NC101\_00003). Make a two by two contingency table:

|  |  |  |
| --- | --- | --- |
|  | Sequences in D2\_01 | Sequences in D2\_02 |
| Assigned to NC101\_00003 |  |  |
| Not assigned to NC101\_00003 |  |  |

use the two sided fisher.test to generate a p-value for the null hypothesis that the columns and rows of the contingency table are independent.

(5) Now generate a p-value for all the genes in the spreadsheet from the Fisher test. Plot out those p-values in a histogram. Are they uniformly distributed? Would you expect them to be? Are the p-values more significant, less significant or what we would expect under a uniform distribution? How does the p-value distribution change if you remove low abundance genes (with for example myT <- myT[ (myT$D2\_01 + myT$D2\_02 > 50),]

(6) Add 1 to every value in the table ( with something like myT = myT + 1 ). This is called adding a pseudo-count. Now consider the first gene (NC101\_00003 ) again. From the first experiment, calculate

expected frequency = p =

(# Assigned to NC101\_00003 in D2\_01)/total # of sequences in D2\_01)

Now use poisson.test to assign a p-value for the null hypothesis that value of p derived from D2\_01 could have produced the number of reads observed for this gene in D2\_02 .

(7) Repeat the calculation in (6) for every gene in the spreadsheet. Graph these p-values against the p-values produced in (5) on a log10-log10 plot. How well do they agree?

(Note: since the mean does not equal the variance, the models make assumptions that are not supported by the data! Also, with n=1 in both conditions, our sample size is small and we would want to be very careful about putting too much weight in any conclusions we draw from such a small sample size...)