1. **Problem 1: Read the dataset.**

setwd("D:\\2022\_spring\\8310\_bioinformatics\_advanced\_stats\\R\_working\_dir\\advanced\_statistics\\lab\_4")

myT <- read.table(

"longitdunalRNASeqData\\nc101\_scaff\_dataCounts.txt",

header=TRUE,

row.names=1

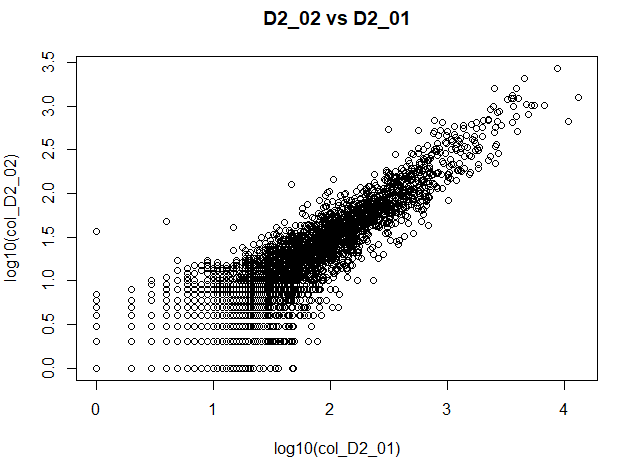
)

nrows = dim(myT)[1]

ncols = dim(myT)[2]

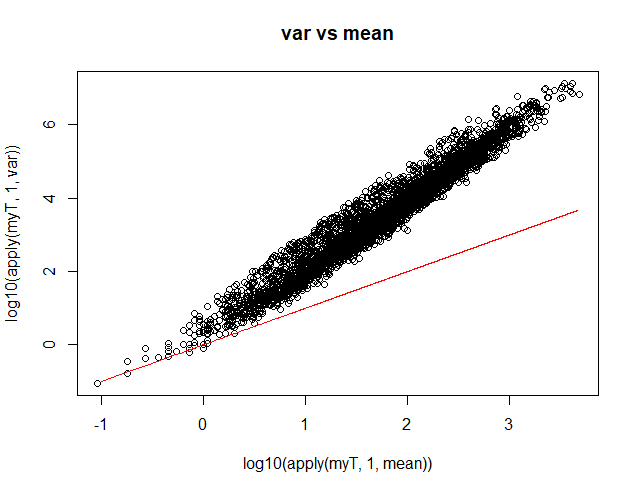
paste('nrows: ', nrows, ', ncols: ', ncols)

1. **Problem 2: Qualitatively, do the biological replicates appear to have similar patterns of gene expression?**



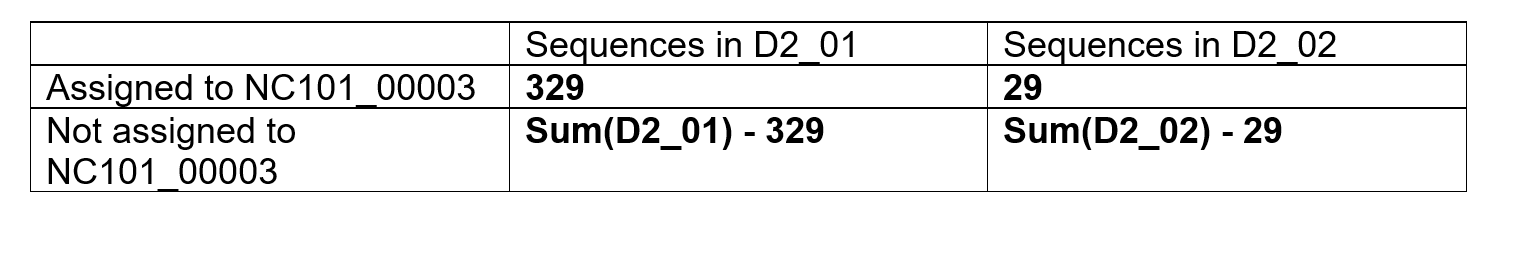
It’s not perfectly linear, but it seems like there are some similarities in the patterns of their gene expressions (even though slight difference with the magnitudes in a log10 scale).

1. **Problem 3: Does the mean equal the variance for these samples?**



No, the mean does not seem to equal the variance (for most cases, the variance is higher than the mean).

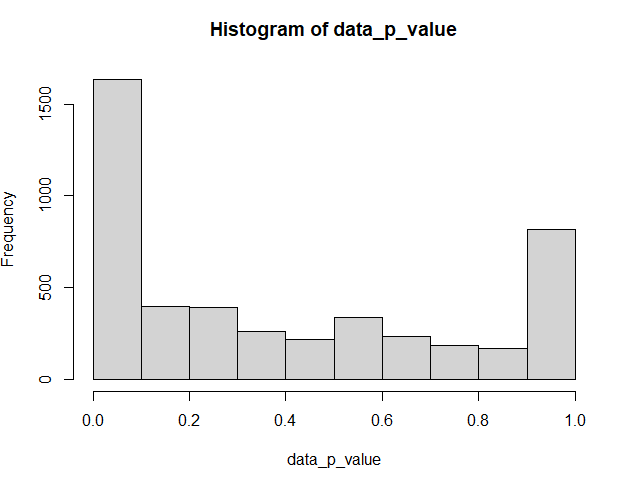
1. **Problem 4:**



"p\_value using first gene: 1.67001714123218e-11". Based on the p-value, It seems like we can not accept the null hypothesis, therefore they does not seem to be independent.

1. **Problem 5:**

**Plot out those p-values in a histogram. Are they uniformly distributed? Would you expect them to be?**

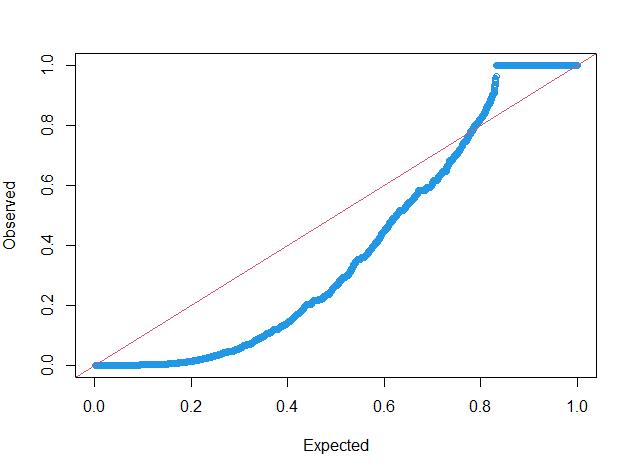


No, from the histogram of the p-values, it does not seem like they follow a uniform distribution.

The frequencies for the most and least significant values, have a large difference with the other frequencies.

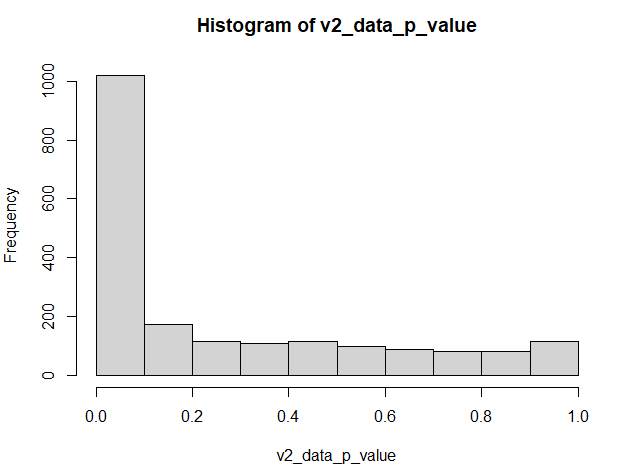
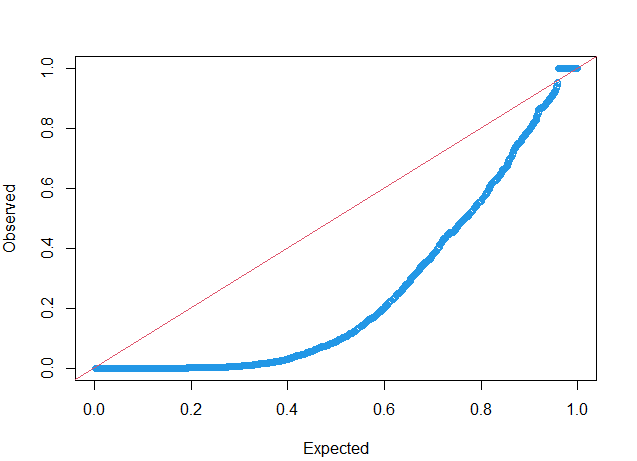
Yes, I was thinking they would show a bit more uniform-like trend considering these two samples are biological replicates. But I think its not rare to show dissimilarities even though they are biological replicates.

**Are the p-values more significant, less significant or what we would expect under a uniform distribution?**



In most cases, the p-values seem to be less significant to the expectation 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),]**

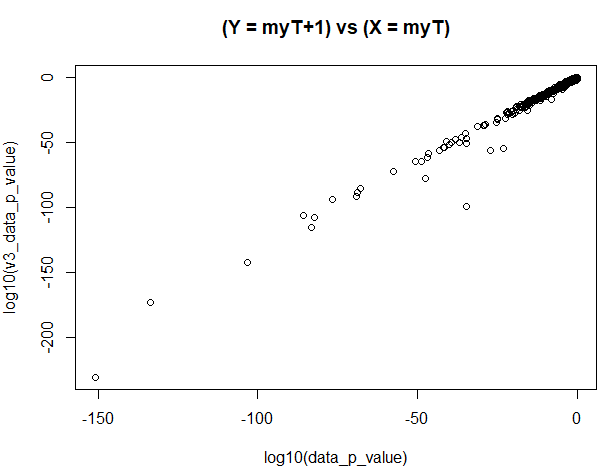
It seems like we got rid of some of the cases (genes) where the p-values were in the higher region. Now, almost for all the cases the observed p-values seem to be less significant than the expected ones.

1. **Problem 6: 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.**

"poisson p-value: 3.61898656558817e-14"

Considering the p-value, the poisson test seems to be rejecting the null hypothesis.

1. **Problem 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?**



They seem to agree very well on most of the cases, I am just seeing some slight disagreement with the magnitude of significance.