Stat 416 Solutions to Homework #1 Spring 2017

1. i. There are two treatment factors in this experiment, fertilizer amount and genotype.

   ii. The levels of fertilizer amount are Low Nitrogen and High Nitrogen; the levels of genotype are one energy line and one grain line.

   iii. We do have a full factorial treatment design for this experiment because all possible combinations of fertilizer amount and genotype were considered.

   iv. The pots are the experimental units, since fertilizer amount and genotypes were randomly assigned to the pots. A pot consisting of two seedlings is one experimental unit. The number of experimental units is 12 for pots.

   v. Gene expression was measured separately for each seedling, so the seedlings are the observational units. More specifically we could say that a sorghum root obtained from a single seedling is a single observational unit.

   vi. The experiment does not involve blocking.

   vii. The experiment is best described as a completely randomized design. The experimental units (pots) in this experiment are randomly allocated to the treatments.

2. i. Fertilizer amount, genotypes, pots, seedlings, lanes, sequencing depth.

   ii. **Fertilizer amount**: treatment effect under study.
   Genotypes: treatment effect under study.
   Pots: biological variability.
   Seedlings: biological variability.
   Lanes: block effect, technical variability.
   Sequencing depth: technical variability.

   iii. **Biological replicates**: 3 pots within the same fertilizer amount and the same genotype, 2 seedlings within each pot.
   There are **no technical replicates** in this experiment since we only measured each seedling once.

3. i. p_1 is the mapping rate for gene g in the first lane and p_2 is the mapping rate for gene g in the second lane.

   ii. $H_0 : p_1 = p_2$ vs. $H_a : p1 \neq p2$

   iii. **R code:**

```
FisherTestMatrix <-
  matrix(c(50, 10e6 - 50, 25,  8e6 - 25),
         nrow = 2,
         dimnames = list(gene = c("yes", "no"),
                         lane = c("1", "2")))
FisherTestMatrix
fisher.test(FisherTestMatrix)
```

   Output:

```
Fisher's Exact Test for Count Data

data:  FisherTestMatrix
p-value = 0.06245
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.971277 2.699292
sample estimates:
odds ratio
       1.6
```

iv. Since $p - value = 0.06245 > 0.05$, we don't have enough evidence to reject the null hypothesis H0: p1 = p2 at 0.05 significance level. Hence, based on the test for this gene, we lack of strong evidence for lane effect in the mapping rate for this gene.

4. i. H_0: The count of reads mapped to gene 4 across lanes follows $Poisson(C_j\mu)$.
H_a: The count of reads mapped to gene 4 across lanes does not follow $Poisson(C_j\mu)$.

ii. The maximum likelihood estimate for the mapping rate μ is

$$\hat{\mu} = \frac{\sum_{j=1}^{4} Y_j}{\sum_{j=1}^{4} C_j} = \frac{54 + 67 + 56 + 74}{615878 + 615867 + 617439 + 739028} = \frac{251}{2588212} = 9.698e - 5$$

iii. The χ^2 test statistic is

$$\begin{aligned} T.S. &= \sum_{j=1}^{4} \frac{(Y_j - C_j\hat{\mu})^2}{C_j\hat{\mu}} \\ &= \frac{(54 - 9.698 \times 6.159)^2}{9.698 \times 6.159} + \frac{(67 - 9.698 \times 6.159)^2}{9.698 \times 6.159} + \frac{(56 - 9.698 \times 6.174)^2}{9.698 \times 6.174} + \frac{(74 - 9.698 \times 7.390)^2}{9.698 \times 7.390} \\ &= 1.761 \end{aligned}$$

iv. The degrees of freedom for the χ^2 goodness-of-fit test is $4 - 1 = 3$.

v. The p-value for the χ^2 goodness-of-fit test is

$$P - value = Pr(\chi_3^2 > 1.761) = 0.623$$

vi. Because p-value is large (> 0.05), we don't have enough evidence to reject the null hypothesis H_0: The count of reads mapped to gene 4 across lanes follows $Poisson(C_j\mu)$.

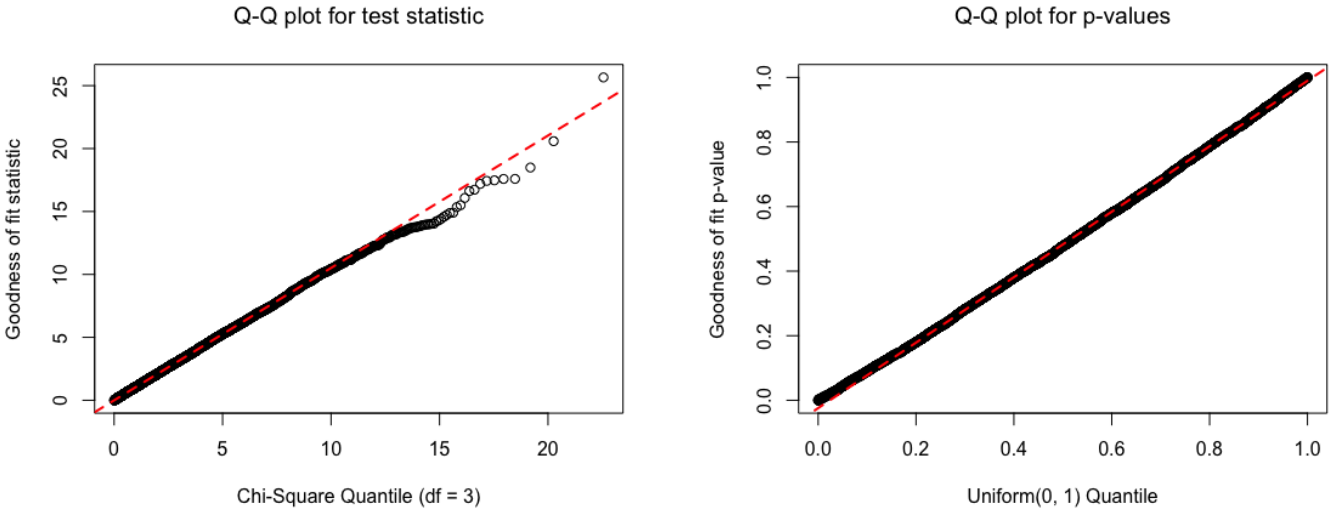
The R Code is given below:

```
Y = c(54, 67, 56, 74) # read counts for gene 4
C = c(615878, 615867, 617439, 739028) # total read counts for the 4 lanes
# calcualte the MLE for mu
mu.hat = sum(Y) / sum(C)
# calcualte the test statistic
```

```
t.s = sum(((Y - mu.hat * C)^2) / (mu.hat * C))
# get the p-value with a chisq distribution with df
df = 3
p.value = 1 - pchisq(t.s, df)
```

5. i. For each gene g, H_0: The count of reads mapped to gene g across lanes follows a Poisson distribution. H_a: The count of reads mapped to gene g across lanes does not follow a Poisson distribution.

   ii. The qq-plot on the left is to check whether the test statistics follow a χ^2_3 distribution with x-axis as theoretical quantiles for the χ^2_3 distribution and y-axis as quantiles for the goodness-of-fit test statistics. The qq-plot on the right is to check whether the p-values follow a $Uniform(0, 1)$ distribution with x-axis as theoretical quantiles for $Uniform(0, 1)$ and y-axis as quantiles of p-values resulting from the goodness-of-fit test.



   iii. Both qq-plots show that the points almost lie in a straight line, indicating that the Poisson models fit well for this count data.

The R Code is given below:

```
data = read.csv("hwk1_4.csv", head = T)
df = 3

# use a for loop
t.s = pvalue = rep(0, length(data[, 1]))
for(i in 1:length(data[,1])){
  Y = data[i,]
  mu.hat = sum(Y) / sum(C)
  t.s[i] = sum(((Y - mu.hat * C)^2) / (mu.hat * C))
  pvalue[i] = 1-pchisq(t.s[i], df)
```

```
}

# work on matrice directly
t.s = apply(data, 1, function(x){sum((x - sum(x)/sum(C) * C)^2 / (sum(x)/sum(C) * C))})
pvalue = 1-pchisq(t.s, df)

# qqplot for test statistic
qqplot(qchisq(ppoints(10000), df = 3), t.s,
       xlab = 'Chi-Square Quantile (df = 3)',
       ylab = 'Goodness of fit statistic',
       main = expression("Q-Q plot for test statistics"))
qqline(t.s, distribution = function(p){qchisq(p, df = 3)},
        col = "red", lwd = 2, lty = 2)

# qqplot for p-value
qqplot(qunif(ppoints(10000), 0, 1), pvalue,
       xlab = 'Uniform(0, 1) Quantile', ylab = 'Goodness of fit p-values',
       main = expression("Q-Q plot for p-values"))
qqline(pvalue, distribution = function(p){qunif(p, 0, 1)},
       col = "red", lwd = 2, lty = 2)
```