STT465 HW2

2018/9/24

Question 1

```
Gout <- read.table("https://raw.githubusercontent.com/gdlc/STT465/master/gout.txt",
                   header = T, sep = " ")
```

a)

Denote the gout variable by X_i, where $X_i \in \{0, 1\}$, $i = 1, ..., n$ with No $= 0$ and Yes $= 1$.

Denote the sex-by-race group by G_i, where $G_i \in \{1, 2, 3, 4\}$ with White Male $= 1$, White Female $= 2$, African American Male $= 3$, African American Female $= 4$.

The likelihood function for the entire sample is

$$\begin{aligned} L(\theta) &= P(X_1, ..., X_n | \theta_{mle}, G_{i's}) \\ &= P(X_1|\theta_{G_1})P(X_2|\theta_{G_2})...P(X_n|\theta_{G_n}) \\ &= \theta_1^{\sum X_i}(1-\theta_1)^{\sum 1-X_i}\theta_2^{\sum X_i}(1-\theta_2)^{\sum 1-X_i}\theta_3^{\sum X_i}(1-\theta_3)^{\sum 1-X_i}\theta_4^{\sum X_i}(1-\theta_4)^{\sum 1-X_i} \end{aligned}$$

Since $\sum X_i = n * \bar{X}$ and $\sum 1 - X_i = n - n * \bar{X} = n(1 - \bar{X})$,

$$L(\theta) = \theta_1^{n_1 * \bar{X_1}}(1-\theta_1)^{n_1(1-\bar{X_1})}\theta_2^{n_2*\bar{X_2}}(1-\theta_2)^{n_2(1-\bar{X_2})}\theta_3^{n_3*\bar{X_3}}(1-\theta_3)^{n_3(1-\bar{X_3})}\theta_4^{n_4*\bar{X_4}}(1-\theta_4)^{n_4(1-\bar{X_4})}$$

b)

In order to obtain the maximum likelihood estimator for each of the success probabilitites,

$$\begin{aligned} l(\theta) &= log(\theta_1^{n_1*\bar{X_1}}(1-\theta_1)^{n_1(1-\bar{X_1})}\theta_2^{n_2*\bar{X_2}}(1-\theta_2)^{n_2(1-\bar{X_2})}\theta_3^{n_3*\bar{X_3}}(1-\theta_3)^{n_3(1-\bar{X_3})}\theta_4^{n_4*\bar{X_4}}(1-\theta_4)^{n_4(1-\bar{X_4})}) \\ &= log(\theta_1^{n_1*\bar{X_1}}) + log((1-\theta_1)^{n_1(1-\bar{X_1})}) + log(\theta_2^{n_2*\bar{X_2}}) + log((1-\theta_2)^{n_2(1-\bar{X_2})}) \\ &+ log(\theta_3^{n_3*\bar{X_3}}) + log((1-\theta_3)^{n_3(1-\bar{X_3})}) + log(\theta_4^{n_4*\bar{X_4}}) + log((1-\theta_4)^{n_4(1-\bar{X_4})}) \\ &= n_1 * \bar{X_1} * log(\theta_1) + n_1(1-\bar{X_1})log(1-\theta_1) + n_2 * \bar{X_2} * log(\theta_2) + n_2(1-\bar{X_2})log(1-\theta_2) \\ &+ n_3 * \bar{X_3} * log(\theta_3) + n_3(1-\bar{X_3})log(1-\theta_3) + n_4 * \bar{X_4} * log(\theta_4) + n_4(1-\bar{X_4})log(1-\theta_4) \end{aligned}$$

$$l'(\theta_1) = \frac{\partial l(\theta)}{\partial \theta_1} = \frac{n_1 * \bar{X_1}}{\theta_1} - \frac{n_1(1-\bar{X_1})}{(1-\theta_1)} = 0$$

$$\frac{n_1 * \bar{X_1}}{\theta_1} = \frac{n_1(1-\bar{X_1})}{(1-\theta_1)}$$

$$\theta_1 * n_1(1-\bar{X_1}) = n_1 * \bar{X_1}(1-\theta_1)$$

$$\hat{\theta_1} = \bar{X_1}$$

Similarly,

$$l'(\theta_2) = \frac{\partial l(\theta)}{\partial \theta_2} = \frac{n_2 * \bar{X_2}}{\theta_2} - \frac{n_2(1-\bar{X_2})}{(1-\theta_2)} = 0$$

$$\hat{\theta_2} = \bar{X_2}$$

$$l'(\theta_3) = \frac{\partial l(\theta)}{\partial \theta_3} = \frac{n_3 * \bar{X_3}}{\theta_3} - \frac{n_3(1-\bar{X_3})}{(1-\theta_3)} = 0$$

$$\hat{\theta_3} = \bar{X_3}$$

$$l'(\theta_4) = \frac{\partial l(\theta)}{\partial \theta_4} = \frac{n_4 * \bar{X_4}}{\theta_4} - \frac{n_4(1-\bar{X_4})}{(1-\theta_4)} = 0$$

$$\hat{\theta_4} = \bar{X_4}$$

c)

Since the 95% confidence interval refers to the middle 95% of the observations, the remaining 5% is equally divided between the two tails and each tail gets 2.5%. Thus, the corresponding z value is 1.96.

Similarly, the corresponding z value for the 99% confidence interval is 2.576.

```
MW <- subset(Gout, sex == "M" & race == "W")
FW <- subset(Gout, sex == "F" & race == "W")
MB <- subset(Gout, sex == "M" & race == "B")
FB <- subset(Gout, sex == "F" & race == "B")
```

```
# White Male
# the MLE
n1 <- nrow(MW)
xBar1 <- nrow(subset(MW, gout == "Y"))/n1
xBar1
```

```
## [1] 0.08450704
```

```
# Approximate 95% CI
samplingVariance <- xBar1*(1-xBar1)/n1
xBar1+c(-1,1)*sqrt(samplingVariance)*1.96
```

```
## [1] 0.03875759 0.13025649
```

```
# Approximate 99% CI
xBar1+c(-1,1)*sqrt(samplingVariance)*2.576
```

```
## [1] 0.0243792 0.1446349
```

```
# White Female
# the MLE
n2 <- nrow(FW)
xBar2 <- nrow(subset(FW, gout == "Y"))/n2
xBar2
```

```
## [1] 0.04819277
```

```
# Approximate 95% CI
samplingVariance <- xBar2*(1-xBar2)/n2
xBar2+c(-1,1)*sqrt(samplingVariance)*1.96
```

```
## [1] 0.01561154 0.08077400
```

```
# Approximate 99% CI
xBar2+c(-1,1)*sqrt(samplingVariance)*2.576
```

```
## [1] 0.005371726 0.091013816
```

```r
# African American Male
# the MLE
n3 <- nrow(MB)
xBar3 <- nrow(subset(MB, gout == "Y"))/n3
xBar3
```

```
## [1] 0.1212121
```

```r
# Approximate 95% CI
samplingVariance <- xBar3*(1-xBar3)/n3
xBar3+c(-1,1)*sqrt(samplingVariance)*1.96
```

```
## [1] 0.009855984 0.232568258
```

```r
# Approximate 99% CI
xBar3+c(-1,1)*sqrt(samplingVariance)*2.576
```

```
## [1] -0.02514166  0.26756590
```

```r
# African American Female
# the MLE
n4 <- nrow(FB)
xBar4 <- nrow(subset(FB, gout == "Y"))/n4
xBar4
```

```
## [1] 0.1016949
```

```r
# Approximate 95% CI
samplingVariance <- xBar4*(1-xBar4)/n4
xBar4+c(-1,1)*sqrt(samplingVariance)*1.96
```

```
## [1] 0.02457055 0.17881928
```

```r
# Approximate 99% CI
xBar4+c(-1,1)*sqrt(samplingVariance)*2.576
```

```
## [1] 0.0003314612 0.2030583693
```

Can also use

qnorm(mean=xBar(.),
sd=sqrt(samplingVariance),
p=c(.025,.0975))

for the 95% CI and p=c(.005,.995) for the 99%CI.

Note, in some cases because the normal approximation is poor the lower bounds are negative. In those cases you may use 0 as lower bound beause the MLE in this model won't ever be negatvie. [see 99% CI for African American Male]

d)

No. Since $\hat{\theta_1} \in [0.039, 0.130]$, $\hat{\theta_2} \in [0.016, 0.081]$, $\hat{\theta_3} \in [0.010, 0.233]$, and $\hat{\theta_4} \in [0.025, 0.179]$ by using the 95% confidence interval of the prevalence of gout in each of the four sex-by-race groups, the confidence intervals overlap each other, which indicates that the differences betwen groups are not statistically significant.

Question 2

a)

Since $\theta \sim Beta(1, 1)$,

$$P(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

For each of the sex-by-race groups, the joint posterior distribution for the probability of developing gout is

$$\begin{aligned} P(\theta_{G_i}|X_1, ..., X_n) &= P(X_1, ..., X_n|\theta_{G_i})P(\theta_{G_i}) \\ &\propto {\theta_{G_i}}^{n*\bar{X}}(1-\theta_{G_i})^{n(1-\bar{X})}{\theta_{G_i}}^{\alpha-1}(1-\theta_{G_i})^{\beta-1} \\ &\propto {\theta_{G_i}}^{n*\bar{X}+\alpha-1}(1-\theta_{G_i})^{n(1-\bar{X})+\beta-1} \end{aligned}$$

Let $\tilde{\alpha} = n * \bar{X} + \alpha$ and $\tilde{\beta} = n(1 - \bar{X}) + \beta$, then $\theta|X \sim Beta(\tilde{\alpha}, \tilde{\beta})$.

Therefore,

$$P(\theta_{G_i}|X_1, ..., X_n) = \frac{\Gamma(\tilde{\alpha} + \tilde{\beta})}{\Gamma(\tilde{\alpha})\Gamma(\tilde{\beta})} {\theta_{G_i}}^{\tilde{\alpha}-1}(1 - \theta)_{G_i}{}^{\tilde{\beta}-1}$$

b)

```
alpha0 <- 1
beta0 <- 1
# White Male
alpha1 <- alpha0+n1*xBar1
beta1 <- alpha0+n1*(1-xBar1)
# Approximate 95% Posterior Credibility Regions
qbeta(shape1=alpha1,shape2=beta1,p=c(.025,.975))
```

```
## [1] 0.04929683 0.14199641
```

```
# White Female
alpha2 <- alpha0+n2*xBar2
beta2 <- alpha0+n2*(1-xBar2)
# Approximate 95% Posterior Credibility Regions
qbeta(shape1=alpha2,shape2=beta2,p=c(.025,.975))
```

```
## [1] 0.02493599 0.09220614
```

```
# African American Male
alpha3 <- alpha0+n3*xBar3
beta3 <- alpha0+n3*(1-xBar3)
# Approximate 95% Posterior Credibility Regions
qbeta(shape1=alpha3,shape2=beta3,p=c(.025,.975))
```

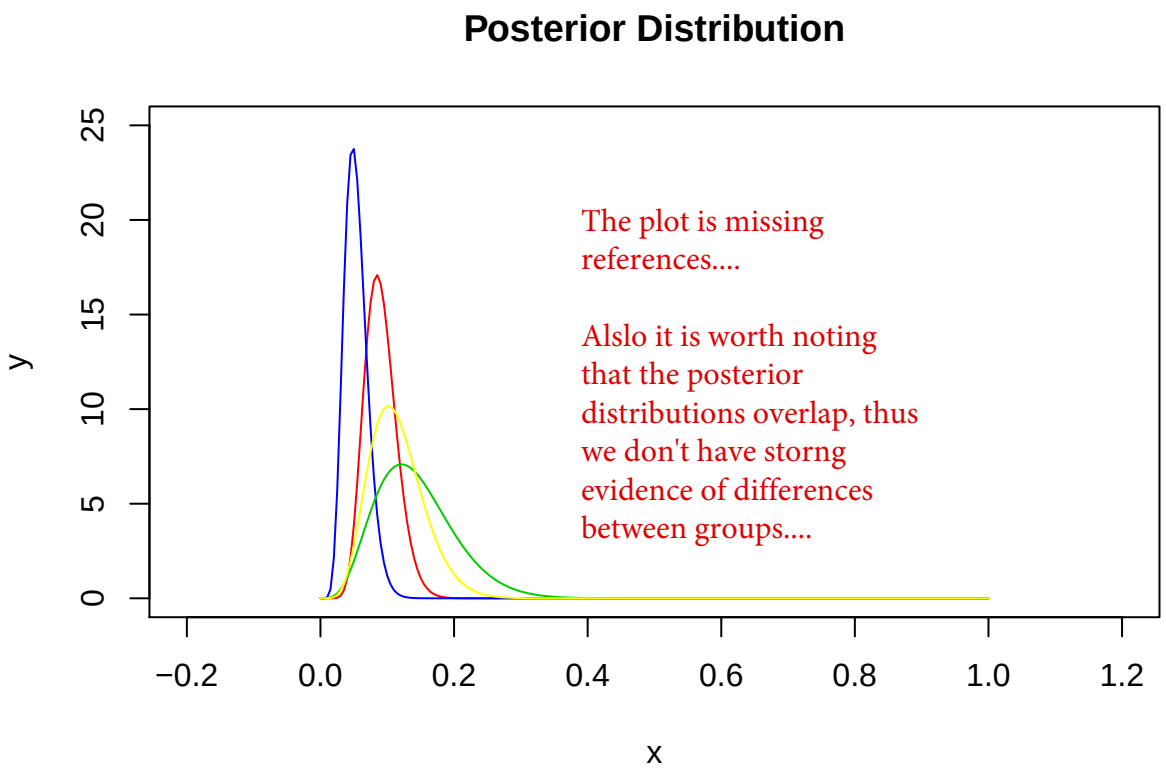
```
## [1] 0.04952846 0.27450349
```

```
# African American Female
alpha4 <- alpha0+n4*xBar4
beta4 <- alpha0+n4*(1-xBar4)
# Approximate 95% Posterior Credibility Regions
qbeta(shape1=alpha4,shape2=beta4,p=c(.025,.975))
```

```
## [1] 0.04821484 0.20505774
```

c)

```
# White Male
x <- seq(0,1,0.005)
y <- dbeta(x=x,shape1=alpha1,shape2=beta1)
plot(y~x,type='l',col=2,xlim=c(-.2,1.2),ylim=c(0,25),main='Posterior Distribution')
# White Female
y <- dbeta(x=x,shape1=alpha2,shape2=beta2)
lines(y~x,type='l',col=4)
# African American Male
y <- dbeta(x=x,shape1=alpha3,shape2=beta3)
lines(y~x,type='l',col=3)
# African American Female
y <- dbeta(x=x,shape1=alpha4,shape2=beta4)
lines(y~x,type='l',col=7)
```



d)

The above posterior distribution plot indicates that the highest point of each posterior distribution is centered around $\hat{\theta}_{G_i}$. Also, as the sample size grows, the posterior distribution becomes narrower and vice versa.

Question 3

a)

```
Fish <- read.csv("fish.csv", header = T)
summary(Fish$count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   3.296   2.000 149.000
```

```
table(Fish$count)
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  11  13  14  15  16  21  22
## 142  31  20  12   6  10   4   3   2   2   1   1   1   1   2   1   2   1
##  29  30  31  32  38  65 149
##   1   1   1   2   1   1   1
```

b)

Denote the fish count variable by X_i, where $X_i \in \{0, 1, ..., n\}$.

The Poisson model is

$$\begin{aligned} P(X_i|\lambda) &= P(X_1|\lambda)P(X_2|\lambda)...P(X_n|\lambda) \\ &= \frac{\lambda^{X_1}e^{-\lambda}}{X_1!} * \frac{\lambda^{X_2}e^{-\lambda}}{X_2!} * ... * \frac{\lambda^{X_n}e^{-\lambda}}{X_n!} \\ &= \frac{\lambda^{\sum X_i}e^{-n\lambda}}{X_1!X_2!...X_n!} \\ &= \frac{\lambda^{n*\bar{X}}e^{-n\lambda}}{\prod X_i!} \\ &\propto \lambda^{n*\bar{X}}e^{-n\lambda} \end{aligned}$$

The left-hand side should be P(X1,X2,...,Xn| lambda).

In order to obtain the maximum likelihood estimator of the Poisson parameter,

$$l(\lambda) = log(\lambda^{n*\bar{X}}e^{-n\lambda}) = log(\lambda^{n*\bar{X}}) + log(e^{-n\lambda}) = n * \bar{X} * log(\lambda) - n\lambda$$

$$l'(\lambda) = \frac{n * \bar{X}}{\lambda} - n = 0$$

$$\hat{\lambda} = \bar{X}$$

```
# the MLE
n <- nrow(Fish)
xBar <- summary(Fish$count)[[4]]
xBar
```

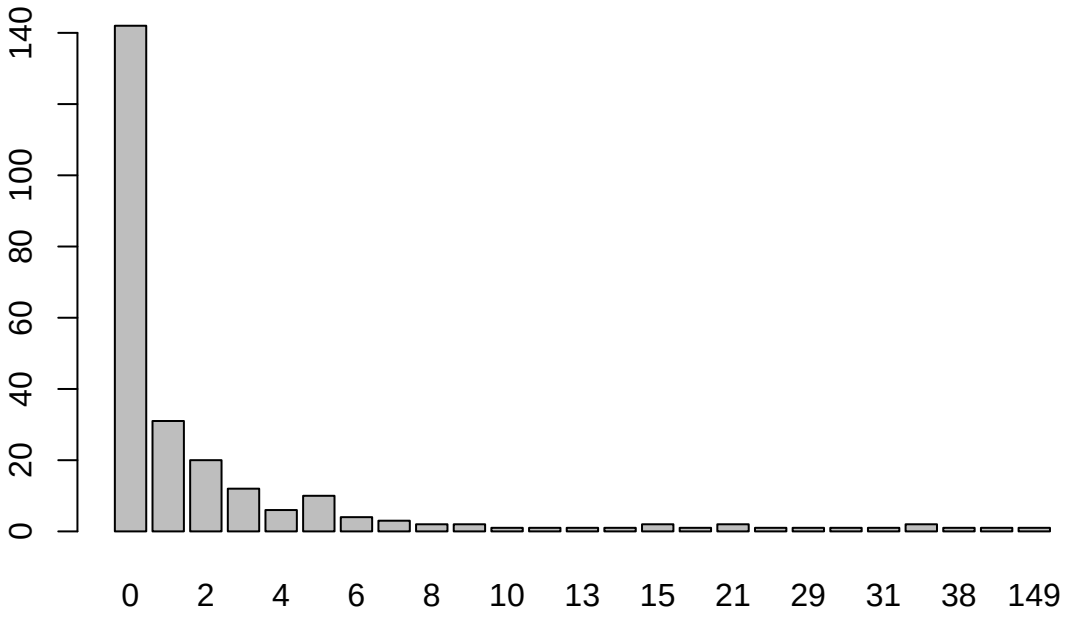
```
## [1] 3.296
```

```
# Approximate 95% CI
samplingVariance <- xBar/n
xBar+c(-1,1)*sqrt(samplingVariance)*1.96
```

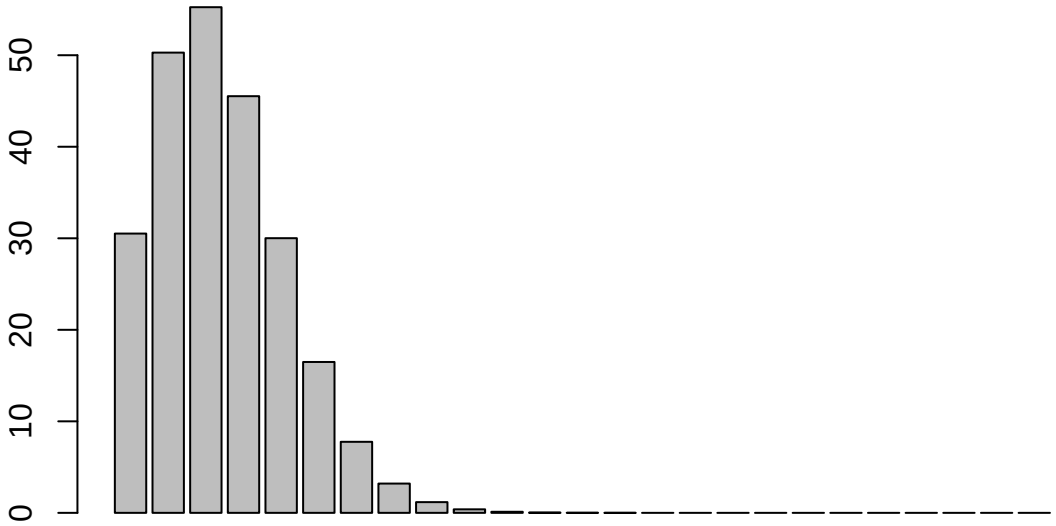
```
## [1] 3.07095 3.52105
```

c)

```
# the Observed Frequencies
barplot(table(Fish$count))
```



```r
# the Predicted Frequencies
Count <- as.data.frame(table(Fish$count))
Count$Var1 <- as.integer(Count$Var1)
barplot(dpois(Count$Var1,lambda=xBar)*sum(Count$Freq))
```



d)

The Poisson model does not fit the data well. Although they have the same parameter, the observed data is more right-skewed and concentrated than the predicted value set.

Data do not have parameters,... models do....the emprical distribituion does not have any parameters.