STA138
Final Project

Ryan Yang
Zheng Wang
Michelle Chen
Instructor: Andrew Farris

Report on Investigation of Byssinosis

Abstract:

Cotton is the main agricultural export product in North Carolina. Cotton dust is thus present in the air during the handling and processing of cotton. This dust contains a mixture of substances including group up plant matter, fibers, bacteria, fungi, soil, pesticides, non-cotton matter, and other contaminants. While cotton processing is decreasing in industrialized countries, it is increasing in developing countries. Cotton processing, particularly in the early processes of spinning, can cause byssinosis.

1.Introduction:

In the 1700s, occupational medicine has grown to encompass an array of respiratory conditions, and one such condition is Byssinosis, a collection of respiratory symptoms elicited by exposure to raw non-synthetic textiles during their manufacturing process. Over the years, Byssinosis has been referred to as a cotton worker's lung, brown lung disease, Monday fever, and mill fever. In 1973, a large cotton textile company in North Carolina participated in a study to investigate the prevalence of byssinosis, a form of pneumoconiosis to which workers exposed to cotton dust are subject. The report data is collected from 5419 workers, with their workplace, employment years, whether or not to smoke, sex, and race defined, to find out the relationship or pattern between the byssinosis and the categories mentioned above. In this analysis report, we are going to investigate relationships between this disease on the one hand and smoking status, sex, race, length of employment, smoking, and dustiness of workplace on the other. In doing such, we will apply Chi-Square Null Distribution, Maximum Likelihood Estimation, and Best Logistic Model methods to explore the data.

2.Summarizing the dataset:

The Byssinosis data is collected from 5419 workers. Data is stored in .csv file format. Here is the statistical summary, including the count, means, standard deviation, the minimum, and the maximum values as well as many percentile values. It is the numerical table to show the correlation between each category.

```
Employment Smoking  Sex    Race     Workspace   Byssinosis
<10  :24   No :36   F:36   O:36   Min.   :1   Min.   : 0.000
>=20 :24   Yes:36   M:36   W:36   1st Qu.:1   1st Qu.: 0.000
10-19:24                         Median :2   Median : 1.000
                                 Mean   :2   Mean   : 2.292
                                 3rd Qu.:3   3rd Qu.: 3.000
                                 Max.   :3   Max.   :31.000

Non.Byssinosis
Min.   :  0.00
1st Qu.:  4.00
Median : 33.00
Mean   : 72.97
3rd Qu.:101.00
Max.   :495.00
```

Figure 1: Summary Table

In order to have a better understanding of the frequency distribution of the number of Byssinosis patients from different types of workgroups, we apply histogram and boxplot to analyze the Byssinosis workers distribution. From the histogram graph below, we can see that the majority of the investigated workgroups have 0 to 5 workers get Byssinosis, and 10 to 35 workers found out to have Byssinosis is less common to see in the investigated workgroups. In general, it is common to find out 0 to 5 people who get Byssinosis are in a textile workgroup.

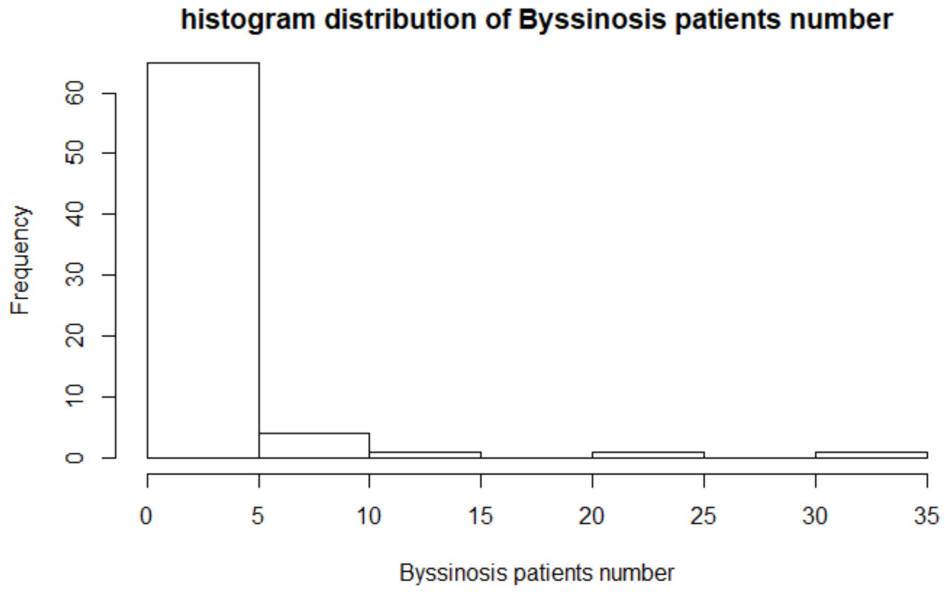

Figure 2: Histogram of the Byssinosis Workers Distribution

Furthermore, we apply a box plot method to display the distribution of data based on a five-number summary. From the boxplot graphic, we can see that the data is tightly clustered to the 0 - 5 number of Byssinosis patient groups, which the data is skewed to the bottom area. From the boxplot diagram, it is easy to see that there are five points drawn outside of the boxplot which

are the outliers. Based on that, we can simply conclude that it is not a normal case to find out that more than five people have Byssinosis from a workgroup since the maximum number of Byssinosis patients of the quantile is about five people.

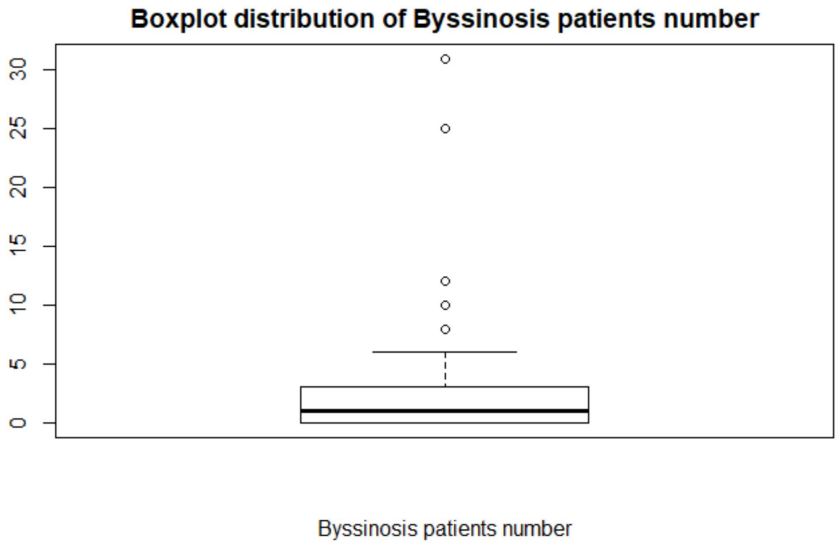

Figure 3: Boxplot Distribution of the Byssinosis Workers

With the same analysis purpose, we would like to define the distribution of the non-Byssinosis people from different workgroups. From the histogram distribution of non-Byssinosis patients, it tells that more than forty working groups have 0 to 50 workers that didn't have byssinosis; it is normal for a workgroup to have 50 to 100, 100 to 150, or 150 to 200 workers that didn't have Byssinosis; and it is infrequently for a group to have 200 to 500 range of worker that didn't have Byssinosis.

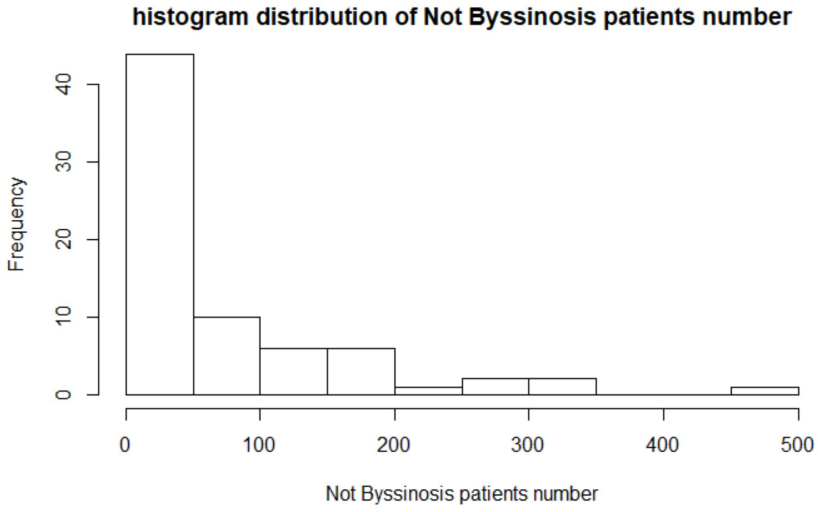

Figure 4: Histogram of the Not Byssinosis Workers Distribution

Continuing with the boxplot analysis, we can define that the data is tightly clustered to the bottom area which is in the range of 0 to 100 workers, and the largest data point excluding any outliers (the maximum point) is about 250 workers. Thus, it is clearer for us to draw a conclusion that it is common to see 0 to 250 workers don't have byssinosis from a workgroup and it is not common to have more than 250 workers don't have byssinosis from a work group.

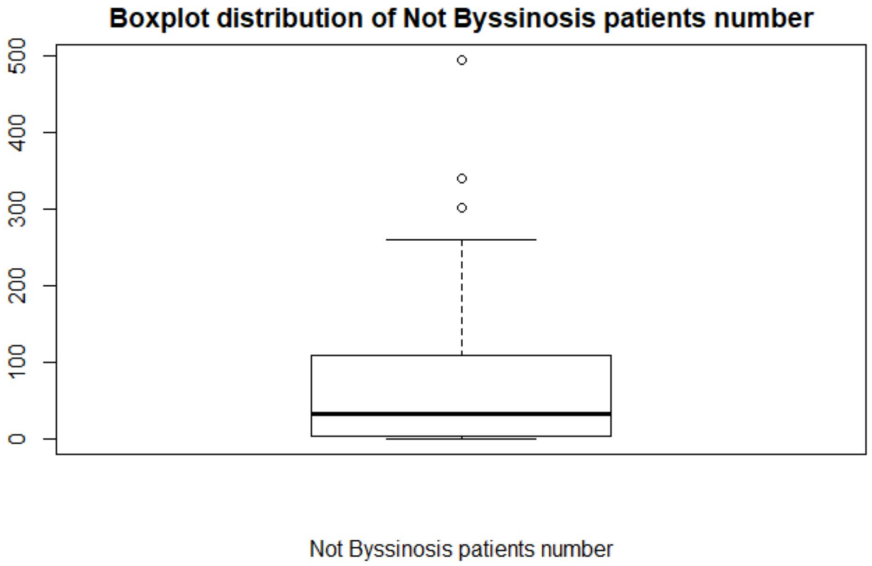

Figure 5: Boxplot Distribution of the Not Byssinosis Workers

3.Analysis and Interpretation:

3.1. Chi-Square Null Distribution:

To show the relationship between Bysssinosis disease and the researching categories: employment, smoking, sex, race, and workspace has an effect on the prevalence of Byssinosis, we applied the chi-square null method to distribute the dataset, which the P-value will be calculated by squaring the result from the observation value minus the expected value, divided by the mean, then find out the appropriate p-value to the result.

$$c^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

Firstly, we initially predicted that the Employment feature has no effect on the prevalence of Byssinosis. We collected the total number of workers who have Byssinosis is 165, and workers who don't have Byssinosis is 5254. After doing the Chi-square distribution, we received the P-

value of 0.0062. The expected significance is 0.01 at this point. Then, we fail to reject the null hypothesis, which can conclude that there is no sufficient evidence to conclude that then Employment has an effect on the prevalence of the Byssinosis disease.

Secondly, we initially predicted that Smoking has no effect on the prevalence of Byssinosis. After doing the Chi-square distribution, we receive the P-value of 7.378918e-06. The expected significance is 0.01 at this point. Then, we reject the null hypothesis and conclude that there is sufficient evidence to conclude that Smoking has an effect on the prevalence of Byssinosis.

Thirdly, we initially predicted that Sex has no effect on the prevalence of Byssinosis. After doing the Chi-square distribution, we received the P-value of 5.016859e-10. The expected significance is 0.01 at this point. Then, we reject the null hypothesis and conclude that there is sufficient evidence to conclude that Sex has an effect on the prevalence of Byssinosis.

Fourthly, we initially predicted that Race has no effect on the prevalence of Byssinosis. After doing the Chi-square distribution, we received the P-value of 0.01263537. The expected significance is 0.01 at this point. Then, we fail to reject the null hypothesis, and conclude that there is sufficient evidence to conclude that Race has an effect on the prevalence of Byssinosis.

Fifty, we initially predicted that Workspace has no effect on the prevalence of Byssinosis. After doing the Chi-square, we received the P-value of 0. The expected significance is 0.01. Then, we fail to reject the null hypothesis, and conclude that there is sufficient evidence to conclude that Workspace has an effect on the prevalence of Byssinosis.

Despite investigating the relationship between each categorical feature and Byssinosis. We are interested in defining the prevalence of the Byssinosis rate. By defining the percentage value, we followed the formula below:

Percent = byssinosis$Byssinosis/(byssinosis$Byssinosis + byssinosis$Non.Byssinosis)

Based on that, if the percentage is more than 0.05, we will assign 1 to its end, otherwise, we will assign 0.

```
##
## Call:  glm(formula = Y ~ Smoking + Sex, family = binomial, data = byssinosis)
##
## Coefficients:
## (Intercept)     Smoking1         Sex1
##     -3.6729       0.2233       2.8660
##
## Degrees of Freedom: 71 Total (i.e. Null);  69 Residual
## Null Deviance:       68
## Residual Deviance: 54.86     AIC: 60.86
```

Considering a fitted model, as alpha, beta1, and beta2 information shown above. The estimated log-odds for such Smoking features according to the model are alpha.

The estimated log-odds ratio of the number of workers who have Byssinosis in Smoking feature is represented as beta1. Thus, the estimated odds of workers who have Byssinosis under this model are 1.2501956 (e^beta1) times the Sexual rate.

The estimated log-odds ratio of workers who have Byssinosis in Sexual features is represented as beta2. Thus, the estimated odds of workers who have Byssinosis with a Smoking feature that is one higher than of another one under this model are 1.82212 (e^beta2) times those of the other. Therefore, the model suggests that Sexual features are more likely to affect the number of workers who have Byssinosis compared with Smoking features.

3.2.Analysis: Maximum Likelihood Estimation:

To further confirm the conclusions that we made previously, we applied the maximum likelihood estimation (MLE) method to estimate the parameters of probability distribution by maximizing the likelihood function. By doing this, we can assume the statistical model of the observed data is most probable.

To find the MLE, we use calculus and maximize l(pi) with respect to pi. In practice, mathematically, L(pi) = ln(l(pi)) which is similar to the log-likelihood function. It is easier to work with, and has the same maximum as l(pi). By doing this, we got the maximum likelihood estimate of the relative proportions of the employment period (less than 10, between 10 to 19, and greater than 20) in order are 0.3818182, 0.4606061, and 0.1575758. To model the probability of counts for each employment period, and also to pick the value of pi that maximizes the probability of getting Byssinosis with the employment feature. We received a maximum likelihood value of 0.0057647 for the Employment feature, which estimates the "TRUE" probability of having Byssinosis.

To estimate the Smoking feature, we got the maximum likelihood estimates of the relative proportions of whether smoke or not is 0.2424242 for non-smokers and 0.7575758 for smokers. For the maximum likelihood value of the Smoking feature, we received a value of 0.07230896. To estimate the Sex feature, we got the maximum likelihood estimates of the relative proportions of male and female are 0.7757576 and 0.2242424. For the maximum likelihood value of the Sex feature, we received a value of 0.07428561.

To estimate the Race features, we got the maximum likelihood estimates of the relative proportions of white people and other people are 0.4424255 and 0.5575745. For the maximum likelihood value of the Race feature, we receive a value of 0.06243482.

3.3.Best Logistic Model:

We want to calculate the prevalence of byssinosis rate, so we define: Percent = byssinosis$Byssinosis/(byssinosis$Byssinosis + byssinosis$Non.Byssinosis). If the Percent is > 0.05, we assign 1 to its end, otherwise, we assign 0. In other words, if the percentage of byssinosis is larger than 5% among all the people(byssinosis+non byssinosis), we assign 1 to its end, otherwise, we assign 0.

```
Step:  AIC=34.93
Y ~ isWorkspace1 + Sex

                   Df Deviance    AIC     LRT Pr(>Chi)
<none>                  22.099 34.929
+ isWorkspace2      1   20.670 37.777 1.42903   0.2319
+ isWorkspace3      1   20.670 37.777 1.42903   0.2319
+ isEmloyee10to19   1   21.385 38.492 0.71394   0.3981
+ Smoking           1   21.733 38.839 0.36643   0.5450
+ Race              1   21.733 38.839 0.36643   0.5450
+ isEmloyeege20     1   21.914 39.021 0.18489   0.6672
+ isEmloyeel10      1   21.914 39.021 0.18489   0.6672
+ Sex:isWorkspace1  1   22.082 39.189 0.01701   0.8962
```

The above is the result of model-select by R, since workspace and employee numbers have three levels, I use as.integer() to separate them into three estimators. The model suggests that workspace has an effect on the response. The best model is :

$$log(\frac{\pi}{1-\pi}) = \alpha + \beta_1 \cdot isWorkspace1 + \beta_2 \cdot Sex$$

Not working the dustiest workspace and Female is the ‘baseline’ case. The estimated log-odds for baseline case according to this model are α = -7.958.

All other things are equal, the estimated log-odds ratio of getting Byssinosis with working in the dustiest workspace vs. not working in the dustiest space is $\beta_1 = 5.551$.
All other things are equal, the estimated log-odds ratio of the patient whose sex is male Vs. whose sex is female is 4.814.

```
Call:  glm(formula = Y ~ isWorkspace1 + Sex, family = binomial, data = byssinosis)

Coefficients:
 (Intercept)  isWorkspace1           Sex
      -7.958         5.551         4.814

Degrees of Freedom: 71 Total (i.e. Null);  69 Residual
Null Deviance:      68
Residual Deviance: 22.1         AIC: 28.1
```

4. Conclusion:

From the Chi-Square Null Distribution method, the model suggests that Sexual features are more likely to affect the number of workers who have Byssinosis compared with Smoking features. Using the maximum likelihood estimation (MLE) method, we estimate the different parameters of probability distribution by maximizing the likelihood function and we see the different population estimate of sex, race, smoking and year of employment. Also, we got our best-fit model using the dustiest workspace and sex to predict Byssinosis from Forward AIC method.

Code Appendix

```
library(rvest)
library(dplyr)
library(plyr)
knitr::opts_chunk$set(message = FALSE)

byssinosis <- read.csv("Byssinosis.csv")
# Summary of Statistics also here using the table

# Employment
byssinosis %>%
   group_by(Employment) %>%
   summarise(
      Sumb =sum(Byssinosis),
      SumNb = sum(Non.Byssinosis)
      )
#Smoking
byssinosis %>%
   group_by(Smoking) %>%
   summarise(
      Sumb =sum(Byssinosis),
      SumNb = sum(Non.Byssinosis)
      )
#Sex
byssinosis %>%
   group_by(Sex) %>%
   summarise(
      Sumb =sum(Byssinosis),
      SumNb = sum(Non.Byssinosis)
      )
#race
byssinosis %>%
   group_by(Race) %>%
   summarise(
      Sumb =sum(Byssinosis),
      SumNb = sum(Non.Byssinosis)
      )
#Workspace
```

```
byssinosis %>%
   group_by(Workspace) %>%
   summarise(
      Sumb =sum(Byssinosis),
      SumNb = sum(Non.Byssinosis)
      )

# Employment
counts_employ <- matrix(c(63,76,26,2666,1902,686),nrow=3)
colnames(counts_employ) <- c(as.roman(1:2))
rownames(counts_employ) <- c("<10",">=20","10-19")
O <- counts_employ  # observed
E <- rowSums(O)%*%t(colSums(O))/sum(O) # expected
pearsonStatistic_employ <- sum((O-E)^2/E)
pearsonpVal_employ <- 1-pchisq(pearsonStatistic_employ,2) # df = 3 !
pearsonpVal_employ
#Smoking
counts_smoking <- matrix(c(40,125,2190,3064),nrow=2)
colnames(counts_smoking) <- c(as.roman(1:2))
rownames(counts_smoking) <- c("No","Yes")
O <- counts_smoking  # observed
E <- rowSums(O)%*%t(colSums(O))/sum(O) # expected
pearsonStatistic_smoking <- sum((O-E)^2/E)
pearsonpVal_smoking <- 1-pchisq(pearsonStatistic_smoking,1) # df = 3 !
pearsonpVal_smoking
#Sex
counts_sex <- matrix(c(37,128,2466,2788),nrow=2)
colnames(counts_sex) <- c(as.roman(1:2))
rownames(counts_sex) <- c("Female","Male")
O <- counts_sex  # observed
E <- rowSums(O)%*%t(colSums(O))/sum(O) # expected
pearsonStatistic_sex <- sum((O-E)^2/E)
pearsonpVal_sex <- 1-pchisq(pearsonStatistic_sex,1) # df = 3 !
pearsonpVal_sex
#race
counts_race <- matrix(c(73,92,1830,3424),nrow=2)
colnames(counts_race) <- c(as.roman(1:2))
rownames(counts_race) <- c("Others","Whie")
O <- counts_race  # observed
E <- rowSums(O)%*%t(colSums(O))/sum(O) # expected
```

```
pearsonStatistic_race <- sum((O-E)^2/E)
pearsonpVal_race <- 1-pchisq(pearsonStatistic_race,1) # df = 3 !
pearsonpVal_race
#Workspace
counts_workspace <- matrix(c(105,18,42,564,1282,3408),nrow=3)
colnames(counts_workspace) <- c(as.roman(1:2))
rownames(counts_workspace) <- c("1","2","3")
O <- counts_workspace  # observed
E <- rowSums(O)%*%t(colSums(O))/sum(O) # expected
pearsonStatistic_workspace <- sum((O-E)^2/E)
pearsonpVal_workspace = 1-pchisq(pearsonStatistic_workspace,2) # df = 3 !
pearsonpVal_workspace
```

2. Model:

We want to calculate the prevalence of byssinosis rate, so we define: Percent = byssinosis$Byssinosis/(byssinosis$Byssinosis + byssinosis$Non.Byssinosis). If the Percent is > 0.05, we assgin 1 to its end, otherwise, we assgin 0.

```{r}
byssinosis$percent <- byssinosis$Byssinosis/(byssinosis$Byssinosis +
byssinosis$Non.Byssinosis)
byssinosis[is.na(byssinosis)] <- 0
for (variable in 1:72) {
   if(byssinosis$percent[variable] > 0.05){
      byssinosis$Y[variable] = 1
   }
   else if(byssinosis$percent[variable] <=  0.05){
      byssinosis$Y[variable] = 0
   }
   else{
      byssinosis$Y[variable] = 0
   }
}

byssinosis$Smoking <- ifelse(byssinosis$Smoking ==  "Yes", 1, 0)
byssinosis$Sex <- ifelse(byssinosis$Sex ==  "M", 1, 0)
byssinosis$Race <- ifelse(byssinosis$Race ==  "W", 1, 0)
byssinosis

isEmloyeel10 = as.integer(byssinosis$Employment == "<10")
isEmloyee10to19 = as.integer(byssinosis$Employment == "10-19")
```

```
isEmloyeege20 = as.integer(byssinosis$Employment == ">=20")
isWorkspace1 = as.integer(byssinosis$Workspace == 1)
isWorkspace2 = as.integer(byssinosis$Workspace == 2)
isWorkspace3 = as.integer(byssinosis$Workspace == 3)
result <- step(glm(Y ~ 1, binomial, byssinosis),
               scope = ~
Smoking*Sex*Race*isEmloyeel10*isEmloyee10to19*isEmloyeege20*isWorkspace1*isWorksp
ace2*isWorkspace3,
               direction = "forward",
               test = "LRT",
               k = log(72))
fittedModel4 <- glm(Y ~isWorkspace1 + Sex, family=binomial, data=byssinosis)
fittedModel4

BYS_Employeement = Byssinosis %>%
  group_by(Employment) %>%
  summarise(Sum_Byssinosis = sum(Byssinosis))
counts = BYS_Employeement$Sum_Byssinosis
multinomLik = function(pi, y){
  dmultinom(y,prob = pi)
}
allPi = function(other_pis){
  c(1-sum(other_pis), other_pis)
}
init = rep(1/3, 2)
optOutput = optim(init, fn = function(other_pis){ -multinomLik(allPi(other_pis), counts)})
MLE_Employeement = allPi(optOutput$par)
MLE_Employeement
ML_Employeement = multinomLik(allPi(optOutput$par), counts)
ML_Employeement

BYS_Somking = Byssinosis %>%
  group_by(Smoking) %>%
  summarise(Sum_Byssinosis = sum(Byssinosis))
counts = BYS_Somking$Sum_Byssinosis
multinomLik = function(pi, y){
  dmultinom(y,prob = pi)
}
allPi = function(other_pis){
  c(1-sum(other_pis), other_pis)
```

```
}
init = rep(1/2, 1)
optOutput = optim(init, fn = function(other_pis){ -multinomLik(allPi(other_pis), counts)})
MLE_Smoking = allPi(optOutput$par)
MLE_Smoking
ML_Smoking = multinomLik(allPi(optOutput$par), counts)
ML_Smoking

BYS_Sex = Byssinosis %>%
  group_by(Sex) %>%
  summarise(Sum_Byssinosis = sum(Byssinosis))
counts = BYS_Sex$Sum_Byssinosis
multinomLik = function(pi, y){
  dmultinom(y,prob = pi)
}
allPi = function(other_pis){
  c(1-sum(other_pis), other_pis)
}
init = rep(1/2, 1)
optOutput = optim(init, fn = function(other_pis){ -multinomLik(allPi(other_pis), counts)})
MLE_Sex = allPi(optOutput$par)
MLE_Sex
ML_Sex = multinomLik(allPi(optOutput$par), counts)
ML_Sex

BYS_Race = Byssinosis %>%
  group_by(Race) %>%
  summarise(Sum_Byssinosis = sum(Byssinosis))
counts = BYS_Race$Sum_Byssinosis
multinomLik = function(pi, y){
  dmultinom(y,prob = pi)
}
allPi = function(other_pis){
  c(1-sum(other_pis), other_pis)
}
init = rep(1/2, 1)
optOutput = optim(init, fn = function(other_pis){ -multinomLik(allPi(other_pis), counts)})
MLE_Race = allPi(optOutput$par)
MLE_Race
ML_Race = multinomLik(allPi(optOutput$par), counts)
```