STA138-Final Project

Dec. 06 2021

Class: STA138 Fall 2021

Professor: Andrew Farris

Members: Shih-Chi Chen
(#ID:917995392)

I. Introduction

Data Introduction: 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.

This data were collected from 5,419 workers, including:

- Type of work place [1 (most dusty), 2 (less dusty), 3 (least dusty)]
- Years of employment [< 10, 10–19, >=20]
- Smoking [Smoker, or not in last 5 years]
- Sex [Male, Female]
- Race [White, Other]
- Byssinosis [Yes, No]

Project goal: To investigate relationships of variable Byssinosis vs five variables: Smoking Status, Sex, Race, Workplace (1,2,3), and Length of Employment, respectively.

II. Data Analysis

Using $\alpha = 0.05$

State the hypotheses:

H_0: There is no association between the two variables.

H_a: There is an association between the two variables.

1. Disease and Smoking status:

Using Pearson Test:

Pearson Statistic=20.0924

P-value=$7.378918e-06$.

Since p-value $< \alpha$, we would reject H_0 and conclude that there is an association between Disease and Smoking Status. (They are dependent).

Using Fisher's Exact Test:

Statistic value: 2190

p-value=$5.5633e-06$.

Since p-value $< \alpha$, we would reject H_0 and conclude that there is an association between Disease and Smoking Status. (They are dependent).

```
##           Byssinosis Non_Byssinosis
## Smoke            125           3064
## Non_Smoke         40           2190
```

Odd ratio :

Odd ratio= $\frac{125*2190}{3064*40)} = 2.2336$

From the odd ratio obtained from the sample, it shows that people who smoke are more likely (2.2336 times) to have Byssinosis than non-smoking people. In addition, since the odd ratio is not equal to 1, it means that two variables are not independent, which is same as the results obtained by Pearson Test and Fisher's Exact Test.

2. Disease and Sex:

Using Pearson Test:

Pearson Statistic=38.67069

P-value=$5.016859e - 10$.

Since p-value $< \alpha$, we would reject H_0 and conclude that there is an association between Disease and Sex. (They are dependent).

Using Fisher's Exact Test:

Statistic value: 2466

P-value=$1.768928e - 10$.

Since p-value $< \alpha$, we would reject H_0 and conclude that there is an association between Disease and Sex. (They are dependent).

```
##   Byssinosis Non_Byssinosis
## M        128           2788
## F         37           2466
```

Odd ratio:

Odd ratio= $\frac{128*2466}{2788*37} = 3.059909$

From the odd ratio obtained from the sample, it shows that Male are more likely (3.059909 times) to have Byssinosis than Female. In addition, since the odd ratio is not equal to 1, it means that these two variables are not independent, which is same as the results obtained by Pearson Test and Fisher's Exact Test.

3. Disease and Race:

Using Pearson Test:

Pearson Statistic=6.219459

P-value=0.01263537.

Since p-value $< \alpha$, we would reject H_0 and conclude that there is an association between Disease and Race. (They are dependent).

Using Fisher's Exact Test:

Statistic value: 1830

P-value=0.01604278.

Since p-value $< \alpha$, we would reject H_0 and conclude that there is an association between Disease and Race. (They are dependent).

```
##   Byssinosis Non_Byssinosis
## W         92           3424
## O         73           1830
```

Odd ratio:

Odd ratio=$\frac{92*1830}{3424*73} = 0.6735693$

From the odd ratio obtained from the sample, it shows that White people are less likely (0.6735693 times) to have Byssinosis than people of Other Races. In addition, since the odd ratio is not equal to 1, it means that these two variables are not independent, which is same as the results obtained by Pearson Test and Fisher's Exact Test.

4. Disease and Work Space:

Using Pearson Test:

Pearson Statistic=413.8151

P-value=$1.384185e - 9$.

Since p-value $< \alpha$, we would reject H_0 and conclude that there is an association between Disease and Work Space. (They are dependent).

Since this is not a 2*2 table, Fisher's Exact Test cannot be used.

```
##   Byssinosis Non_Byssinosis
## 1        105            564
## 2         18           1282
## 3         42           3408
```

Odd ratio for Work Space 1 (most dusty):

$$Odd_{Bys/1} = \frac{105/165}{1 - 105/165} = 1.75$$

$Odd_{NoBys/1} = \frac{564/5254}{1-564/5254} = 0.1202559$

Therefore, $Odd_{Bys/NoBys} = \frac{1.75}{0.1202559} = 14.5523$

From the odd ratio of Work Space 1obtained from the sample, it shows that people are more likely (14.5523 times) to have Byssinosis in the most dusty work place.

Odd ratio for Work Space 2 (less dusty):

$Odd_{Bys/2} = \frac{18/165}{1-18/165} = 0.122449$

$Odd_{NoBys/2} = \frac{1282/5254}{1-1282/5254} = 0.3227593$

Therefore, $Odd_{Bys/NoBys} = \frac{0.122449}{0.3227593} = 0.3793818$

From the odd ratio of Work Space 2 obtained from the sample, it shows that people are less likely (0.3793818 times) to have Byssinosis in the less dusty work place.

Odd ratio for Work Space 3 (least dusty):

$Odd_{Bys/3} = \frac{42/165}{1-42/165} = 0.3414634$

$Odd_{NoBys/3} = \frac{3408/5254}{1-3408/5254} = 1.846154$

Therefore, $Odd_{Bys/NoBys} = \frac{0.3414634}{1.846154} = 0.1849593$

From the odd ratio of Work Space 3 obtained from the sample, it shows that people are less likely (0.1849593 times) to have Byssinosis in the least dusty work place.

5. Disease and Length of Employment:

Using Pearson Test:

Pearson Statistic=10.16041

P-value=0.006218624.

Since p-value $< \alpha$, we would reject H_0 and conclude that there is an association between Disease and Length of Employment. (They are dependent).

Since this is not a 2*2 table, Fisher's Exact Test cannot be used.

```
##       Byssinosis Non_Byssinosis
## <10           63           2666
## 10-19         26            686
## >=20          76           1902
```

Odd ratio for Length of Employment (<10):

$Odd_{Bys/<10} = \frac{63/165}{1-63/165} = 0.6176471$

$$Odd_{NoBys/<10} = \frac{2666/5254}{1-2666/5254} = 1.030139$$

Therefore, $Odd_{Bys/NoBys} = \frac{0.6176471}{1.030139} = 0.5995765$

From the odd ratio of Length of Employment (<10) obtained from the sample, it shows that people are less likely (0.5995765 times) to have Byssinosis with Length of Employment less than 10 years.

Odd ratio for Length of Employment (10-19):

$$Odd_{Bys/10-19} = \frac{26/165}{1-26/165} = 0.1870504$$

$$Odd_{NoBys/10-19} = \frac{686/5254}{1-686/5254} = 0.1501751$$

Therefore, $Odd_{Bys/NoBys} = \frac{0.1870504}{0.1501751} = 1.245549$

From the odd ratio of Length of Employment (10-19) obtained from the sample, it shows that people are more likely (1.245549 times) to have Byssinosis with Length of Employment between 10 and 19 years.

Odd ratio for Length of Employment (>=20):

$$Odd_{Bys/>=20} = \frac{76/165}{1-76/165} = 0.8539326$$

$$Odd_{NoBys/>=20} = \frac{1902/5254}{1-1902/5254} = 0.5674224$$

Therefore, $Odd_{Bys/NoBys} = \frac{0.8539326}{0.5674224} = 1.504933$

From the odd ratio of Length of Employment (>=20) obtained from the sample, it shows that people are more likely (1.504933 times) to have Byssinosis with Length of Employment greater than 20 years.

III. Conclusion

	Disease vs Smoking status	**Disease vs Sex**	**Disease vs Race**	**Disease vs Work Space**	**Disease vs Length of Employment**
P-value (Person Test)	7.378918e-06	5.016859e-10 (smallest)	0.01263537 (largest)	1.384185e-9	0.006218624

Odd ratio	2.2336	3.059909	0.6735693	1. 14.5523 2. 0.3793818 3. 0.1849593	<10 0.5995765 10-19 1.245549 >=20 1.504933
Disease chance	Smoke more	Male more	White less	Type 1 more	>=20 more

From the table above, it shows that Work Space 1(most dusty) has the highest odd ratio, which is 14.5523. This means that people working in the most dusty place have the highest chance to get Byssinosis than other factors since the highest odd ratio is obtained in this group. In addition, the second highest odd ratio occurs on the Sex variable, which is 3.059909. Therefore, the males are more likely to get Byssinosis than others.

According to the P-values, it also shows that Sex and Work Space variables have the smallest p-values. Therefore, there is a strong evidence to conclude that these two variables are not independent from Byssinosis.

Moreover, the same results can be observed from the odd ratio. Since these two variables have large odd ratios, it means that there exists a strong relationship between Sex and Byssinosis, and work Space and Byssinosis, respectively.

IV. Appendix Code

```r
#read data
Byssinosis <- read.csv("~/Downloads/Byssinosis.csv")

#Disease and Smoking status
smok_bys=sum(subset(Byssinosis, Smoking == "Yes")$Byssinosis)
smok_nonbys=sum(subset(Byssinosis, Smoking == "Yes")$Non.Byssinosis)
nonsmok_bys=sum(subset(Byssinosis, Smoking == "No")$Byssinosis)
nonsmok_nonbys=sum(subset(Byssinosis, Smoking == "No")$Non.Byssinosis)
counts_smokbys <- matrix(c(smok_bys,nonsmok_bys,smok_nonbys,nonsmok_nonbys),
nrow=2)
rownames(counts_smokbys) <- c("Smoke","Non_Smoke")
colnames(counts_smokbys) <- c("Byssinosis","Non_Byssinosis")
counts_smokbys

#pearson test
pearsonStatistic <- chisq.test(counts_smokbys, correct=FALSE)$stat
pearsonpVal <- chisq.test(counts_smokbys, correct=FALSE)$p.val

#fisher exact test
fisherPval <- fisher.test(counts_smokbys)$p.val
odd_ratio_smok=(125*2190)/(3064*40)
```

```
#Disease and Sex
sexm_bys=sum(subset(Byssinosis, Sex == "M")$Byssinosis)
sexm_nonbys=sum(subset(Byssinosis, Sex == "M")$Non.Byssinosis)
sexf_bys=sum(subset(Byssinosis, Sex == "F")$Byssinosis)
sexf_nonbys=sum(subset(Byssinosis, Sex == "F")$Non.Byssinosis)
counts_sexbys <- matrix(c(sexm_bys,sexf_bys,sexm_nonbys,sexf_nonbys),
nrow=2)
rownames(counts_sexbys) <- c("M","F")
colnames(counts_sexbys) <- c("Byssinosis","Non_Byssinosis")
counts_sexbys

#pearson test
pearsonStatistic <- chisq.test(counts_sexbys, correct=FALSE)$stat
pearsonpVal <- chisq.test(counts_sexbys, correct=FALSE)$p.val

#fisher exact test
fisherPval <- fisher.test(counts_sexbys)$p.val
odd_ratio_sex=(128*2466)/(2788*37)
#Disease and Race
racew_bys=sum(subset(Byssinosis, Race == "W")$Byssinosis)
racew_nonbys=sum(subset(Byssinosis, Race == "W")$Non.Byssinosis)
raceo_bys=sum(subset(Byssinosis, Race == "O")$Byssinosis)
raceo_nonbys=sum(subset(Byssinosis, Race == "O")$Non.Byssinosis)
counts_racebys <- matrix(c(racew_bys,raceo_bys,racew_nonbys,raceo_nonbys),
nrow=2)
rownames(counts_racebys) <- c("W","O")
colnames(counts_racebys) <- c("Byssinosis","Non_Byssinosis")
counts_racebys

#pearson test
pearsonStatistic <- chisq.test(counts_racebys, correct=FALSE)$stat
pearsonpVal <- chisq.test(counts_racebys, correct=FALSE)$p.val

#fisher exact test
fisherPval <- fisher.test(counts_racebys)$p.val
odd_ratio_race=(92 *1830)/(3424*73)
#Disease and Work Space
workone_bys=sum(subset(Byssinosis, Workspace == "1")$Byssinosis)
workone_nonbys=sum(subset(Byssinosis, Workspace == "1")$Non.Byssinosis)
worktwo_bys=sum(subset(Byssinosis, Workspace == "2")$Byssinosis)
worktwo_nonbys=sum(subset(Byssinosis, Workspace == "2")$Non.Byssinosis)
workthr_bys=sum(subset(Byssinosis, Workspace == "3")$Byssinosis)
workthr_nonbys=sum(subset(Byssinosis, Workspace == "3")$Non.Byssinosis)

counts_workbys <- matrix(c(workone_bys,workone_nonbys,worktwo_bys,worktwo_non
bys,workthr_bys,workthr_nonbys),nrow = 3, byrow = TRUE)
rownames(counts_workbys) <- c("1","2","3")
colnames(counts_workbys) <- c("Byssinosis","Non_Byssinosis")
counts_workbys
```

```r
#pearson test
pearsonStatistic <- chisq.test(counts_workbys, correct=FALSE)$stat
pearsonpVal <- chisq.test(counts_workbys, correct=FALSE)$p.val
#Disease and length of employment
emp10_bys=sum(subset(Byssinosis, Employment == "<10")$Byssinosis)
emp10_nonbys=sum(subset(Byssinosis, Employment == "<10")$Non.Byssinosis)
emp19_bys=sum(subset(Byssinosis, Employment == "10-19")$Byssinosis)
emp19_nonbys=sum(subset(Byssinosis, Employment == "10-19")$Non.Byssinosis)
emp20_bys=sum(subset(Byssinosis, Employment == ">=20")$Byssinosis)
emp20_nonbys=sum(subset(Byssinosis, Employment == ">=20")$Non.Byssinosis)

counts_empbys <- matrix(c(emp10_bys,emp10_nonbys,emp19_bys,emp19_nonbys,emp20
_bys,emp20_nonbys),nrow = 3, byrow = TRUE)
rownames(counts_empbys) <- c("<10","10-19",">=20")
colnames(counts_empbys) <- c("Byssinosis","Non_Byssinosis")
counts_empbys

#pearson test
pearsonStatistic <- chisq.test(counts_empbys, correct=FALSE)$stat
pearsonpVal <- chisq.test(counts_empbys, correct=FALSE)$p.val
```