# STA 138 Project

*Dandi Peng 915553480 Yuhan Ning 915486450*

*2/1/2019*

## Project 1

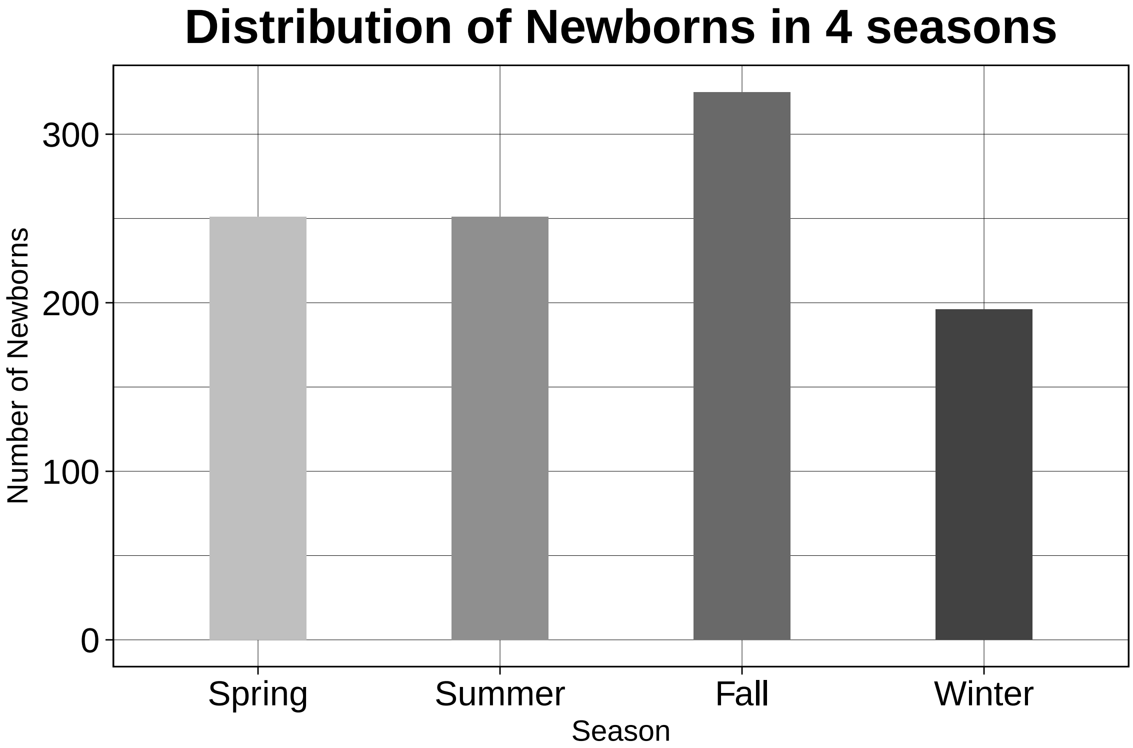
### 1. Introduction

With a hypothesis that newborns are uniformly distributed among 4 seasons, here are the data we own to have the hypothesis test $H_0: \ \pi_1 = 0.25, \ \pi_2 = 0.25, \ \pi_3 = 0.25, \ \pi_4 = 0.25$ and $H_a: \ at\ least\ two\ \pi_i's\ are\ not\ the\ null\ hypothesized\ values.$

### 2. Summary

```
## [1] 1023
```

```
##
## Spring Summer   Fall Winter
##    251    251    325    196
```



Our data has 1023 subjects, where the numbers of newborns in Spring, Summer, Fall and Winter are 251, 251, 325 and 196.

From the above barplot, comparing the heights for number of newborns in four seasons, it is apparently shown that there are the highest number of newborns in Fall and the lowest number of newborns in Winter with a visually comparable difference.

## 3. Analysis

With the null hypothesis that a subject is equally likely to be born in any season, it brings us to the multinomial case, where the Pearson's Chi-Square Test and Wilson-Adjusted $(1 - \alpha/2g) \times 100\%$ Bonferroni Corrected Confidence Intervals can be applied.

**Pearson's Chi-Square Test**

```
## X-squared
##  32.88661
```

```
## [1] 3.403013e-07
```

**Wilson-Adjusted Bonferroni Corrected Confidence Intervals**

```
##            Lower     Upper
## Spring 0.2143639 0.2813963
## Summer 0.2143639 0.2813963
## Fall   0.2832842 0.3557149
## Winter 0.1639398 0.2253590
```

## 4. Interpretation

From the outputs in part **3**, the p value of Pearson's Test is $3.403013 \times 10^{-7}$. This p value tells us that we would observe our data or more extreme with probability $3.403013 \times 10^{-7}$, if the null is true.

For the 95% Wilson_Adjusted Confidence Intervals of the numbers of newborns in four seasons, it demonstrates that:

We are 95% confident that the proportion of newborns in Spring and Summer are both between 0.2143639 and 0.2813963, the proportion of newborns in Fall is between 0.2832842 and 0.3557149, and the proportion of newborns in Winter is between 0.1639398 and 0.2253590.

## 5. Conclusion

The p value is $3.403013 \times 10^{-7}$, which is significantly less than $\alpha = 0.05$. Meanwhile, the 95% Wilson_Adjusted Confidence Intervals shows that the proportions of newborns in Fall and Winter don't cover 0.25.

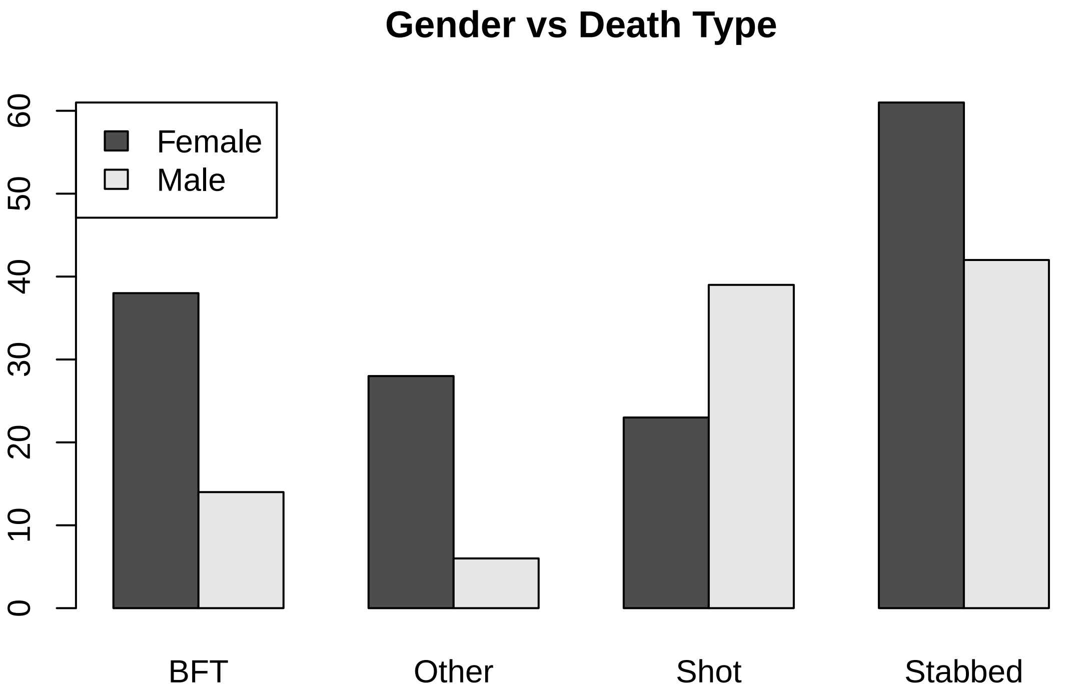
Both the Pearson's Test and the confidence intervals state that we reject the null hypothesis, and we are not able to conclude that a subject is equally likely to be born in any season.

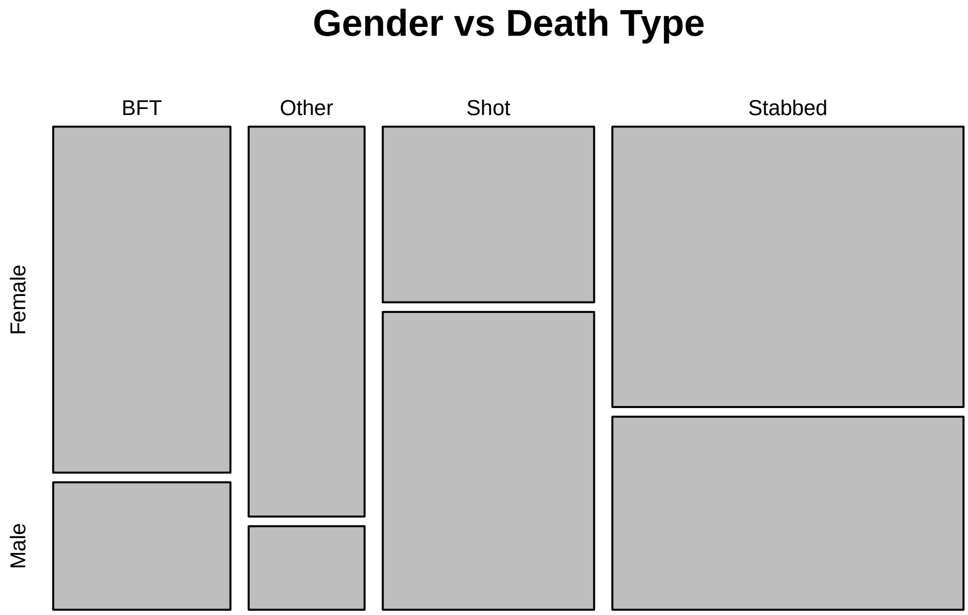
# Project 2

## 1. Introduction

We are interested in how gender and death type are related in dataset 'horror'. In this project, we aim to check dependence of two variables with assumption that all samples are random.

## 2. Summary





There are 251 samples in the dataset ‘horror’ and the two-way table of counts are shown below:

```
##          death
## gen        BFT Other Shot Stabbed
##   Female  38    28   23      61
##   Male    14     6   39      42
```

We plot barplot and mosaicplot to catch a better glimpse of the counts in different categories. Clearly, for different death types, group female has highest number of deaths in level ‘Other’ and lowest number of deaths in ‘Shot’. The uneven distribution of counts implies the two factors may not be independent. Therefore, we use hypothesis tests and confidence intervals to check further.

## 3. Analysis

### Pearson’s Chi-Square Test

$H_0 : P(female|BFT) = P(female|Other) = P(female|Shot) = P(female|Stabbed)$ $H_a$ : At least one probability in H0 is not the same.

Under the null hypothesis, $X^2$ follows chi-square distribution with d.f = (2-1)*(4-1) = 3

Test statistic is 24.30671 and p-value is $2.155483 \times 10^5$.

We got table of expected value, table of standard residuals and table of $X^2$:

```
##          death
## gen            BFT    Other     Shot  Stabbed
##   Female 31.0757 20.31873 37.05179 61.55378
##   Male   20.9243 13.68127 24.94821 41.44622
```

```
##          death
## gen               BFT      Other        Shot    Stabbed
##   Female  2.1991361  2.8891377 -4.1938149 -0.1449092
##   Male   -2.1991361 -2.8891377  4.1938149  0.1449092
```

```
##          death
## gen                BFT       Other        Shot     Stabbed
##   Female 1.542876698 2.903823139 5.329104657 0.004982272
##   Male   2.291401037 4.312608622 7.914511866 0.007399413
```

Stabbed differed the least from what was expected under the null. Shot tended to report less than expected under H0 and it contributed the most to the rejection of H0.

### Wilson-Adjusted Bonferroni Corrected Confidence Intervals

To know how gender depends on death types, we’d like to compare the probability of being female for different death types using Bonferroni corrected confidence interval.

Since we have four groups (BFT, Other, Shot, Stabbed), we compare 6 differences in total (4 chooses 2 equals 6).

$g = 6$; $alpha = 0.05$

```
##                                            Lower       Upper
## P(female|BFT)-P(female|Other)     -0.3202795270  0.15361286
## P(female|BFT)-P(female|Shot)       0.1206199680  0.57382448
## P(female|BFT)-P(female|Stabbed)   -0.0729209474  0.33641301
## P(female|Other)-P(female|Shot)     0.1943894702  0.66672164
## P(female|Other)-P(female|Stabbed) -0.0001284661  0.43028720
## P(female|Shot)-P(female|Stabbed)  -0.4192395177 -0.01171286
```

**Partitioning Tables**

Continually, we partition the table to determine what sub categories are independent/dependent and check with the conclusion we got using confidence interval. In total, we get 3 tables, denoted as 'table_p1', 'table_p2', 'table_p3'.

## 4. Interpretation

From Pearson's Test, p-value is $2.155483 \times 10^5$. If gender and death type are independent, we would observe our data or more extreme with probability $2.155483 \times 10^5$. Since p value is much smaller than alpha, we reject H0 and conclude that gender and death types are dependent.

Continually, we compare 6 pairs of difference. Among the results:

1. Confidence Intervals for P(female|BFT)-P(female|Other), P(female|BFT)-P(female|Stabbed), P(female|Other)-P(female|Stabbed) contain 0;
2. P(female|BFT)-P(female|Shot), P(female|Other)-P(female|Shot) are larger than 0;
3. P(female|Shot)-P(female|Stabbed) is smaller than 0.

We are overall 95% confident that the probabilities for female to die in BFT, Other and Stabbed are the same, which is larger than the probability of Shot. In other words, it is least probable for women to die of shot.

By partition tables, we got

- p-value from Pearson's Test for table_p1 as 0.3195, which is larger than alpha. So we fail to reject $H_0 : P(female|BFT) = P(female|Other)$;
- P-value for table_p2 is $1.1716 * 10^{-6}$, which is smaller than alpha and we reject $H_0 : P(female|Shot) = P(female|BFT + Other)$;
- P-value for table_p3 is 0.8848, which is larger than alpha and we fail to reject $H_0 : P(female|Stabbed) = P(female|BFT + Other + Shot)$

The result is consistent with findings in confidence interval.

## 5. Conclusion

Based on above tests and calculations, we find that gender and death type are dependent. The probability for female to die of shot is less than other death types. The probability of other 3 death types for female is approximately the same.

## R Appendix

```r
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(MASS)
born <- read.csv('born.csv', header = T)
born$Season <- factor(born$Season,levels(born$Season)[c(2,3,1,4)])
nrow(born)
born.table = table(born$Season)
born.table
ggplot(born,aes(x=Season,fill=Season))+geom_bar(stat='count',width = 0.4)+
  ylab("Number of Newborns") + ggtitle("Distribution of Newborns in 4 seasons")+
  theme_linedraw() + scale_fill_manual(values=c('gray75','gray56','grey41','grey26'))+
  theme(axis.text=element_text(size=13),
          plot.title = element_text(size = 18, hjust=0.5, face = 'bold'),
          legend.position="none")
# Pearson's Chi-Square Test
the.test <- chisq.test(born.table,p = c(1/4,1/4,1/4,1/4),correct = FALSE)
the.test$statistic
the.test$p.value
## Wilson-Adjusted Bonferroni Corrected Confidence Intervals
alpha <- .05  ## at the 95% confidence level
g <- 4
n <- sum(born.table)
y1 <- born.table[1]
y2 <- born.table[2]
y3 <- born.table[3]
y4 <- born.table[4]

CI1 = prop.test(y1+2,n+4,conf.level = 1-alpha/(g),correct = FALSE)$conf.int
CI2 = prop.test(y2+2,n+4,conf.level = 1-alpha/(g),correct = FALSE)$conf.int
CI3 = prop.test(y3+2,n+4,conf.level = 1-alpha/(g),correct = FALSE)$conf.int
CI4 = prop.test(y4+2,n+4,conf.level = 1-alpha/(g),correct = FALSE)$conf.int

results = rbind(CI1,CI2,CI3,CI4)
colnames(results) = c("Lower","Upper")
rownames(results) = names(born.table)
results
horror = read.csv('horror.csv')
n_sum = nrow(horror)
table_data = table(horror)
y1 = table_data[1,1]
y2 = table_data[1,2]
y3 = table_data[1,3]
```

```
y4 = table_data[1,4]
n1 = sum(table_data[,1])
n2 = sum(table_data[,2])
n3 = sum(table_data[,3])
n4 = sum(table_data[,4])
barplot(table_data, beside = TRUE, main = 'Gender vs Death Type', legend.text = rownames
table2 = table(horror$death, horror$gen)
mosaicplot(table2, main = 'Gender vs Death Type')
table_data
pearson.test = chisq.test(table_data, correct = FALSE)
test.stat = pearson.test$statistic
p.val = pearson.test$p.value
expt = pearson.test$expected
stdres = pearson.test$stdres
x_sqr = (pearson.test$observed - pearson.test$expected)^2/pearson.test$expected
expt
stdres
x_sqr
g = 6
alpha = 0.05
ci1 = prop.test(c(y1+1, y2+1), c(n1+2, n2+2), correct = FALSE, conf.level = 1 - alpha/g
ci2 = prop.test(c(y1+1, y3+1), c(n1+2, n3+2), correct = FALSE, conf.level = 1 - alpha/g
ci3 = prop.test(c(y1+1, y4+1), c(n1+2, n4+2), correct = FALSE, conf.level = 1 - alpha/g
ci4 = prop.test(c(y2+1, y3+1), c(n2+2, n3+2), correct = FALSE, conf.level = 1 - alpha/g
ci5 = prop.test(c(y2+1, y4+1), c(n2+2, n4+2), correct = FALSE, conf.level = 1 - alpha/g
ci6 = prop.test(c(y3+1, y4+1), c(n3+2, n4+2), correct = FALSE, conf.level = 1 - alpha/g
results = rbind(ci1, ci2, ci3, ci4, ci5, ci6)
colnames(results) = c('Lower','Upper')
rownames(results) = c('P(female|BFT)-P(female|Other)','P(female|BFT)-P(female|Shot)','P(
results
table_p1 = table_data[,c(1,2)]
table_p2 = cbind(rowSums(table_p1), table_data[,3])
colnames(table_p2) = c('B+O','Shot')
table_p3 = cbind(rowSums(table_p2), table_data[,4])
colnames(table_p3) = c('B+O+Shot','Stabbed')

pearson.test1 = chisq.test(table_p1, correct = FALSE)
test.stat1 = pearson.test1$statistic
p.val1 = pearson.test1$p.value

pearson.test2 = chisq.test(table_p2, correct = FALSE)
test.stat2 = pearson.test2$statistic
p.val2 = pearson.test2$p.value
```

```r
pearson.test3 = chisq.test(table_p3, correct = FALSE)
test.stat3 = pearson.test3$statistic
p.val3 = pearson.test3$p.value
```