Practical 2.5

CHEN

2024-03-12

1. Simulation of the probability of goodness-of-fit test

In the lecture you learned how to use the chi-square value to measure the discrepancy between observed data and expected data. Based on Season preferences data, use a simulation to get the distribution of chi2 and calculate the p-value (the probability that the discrepancy is larger than the observed data).

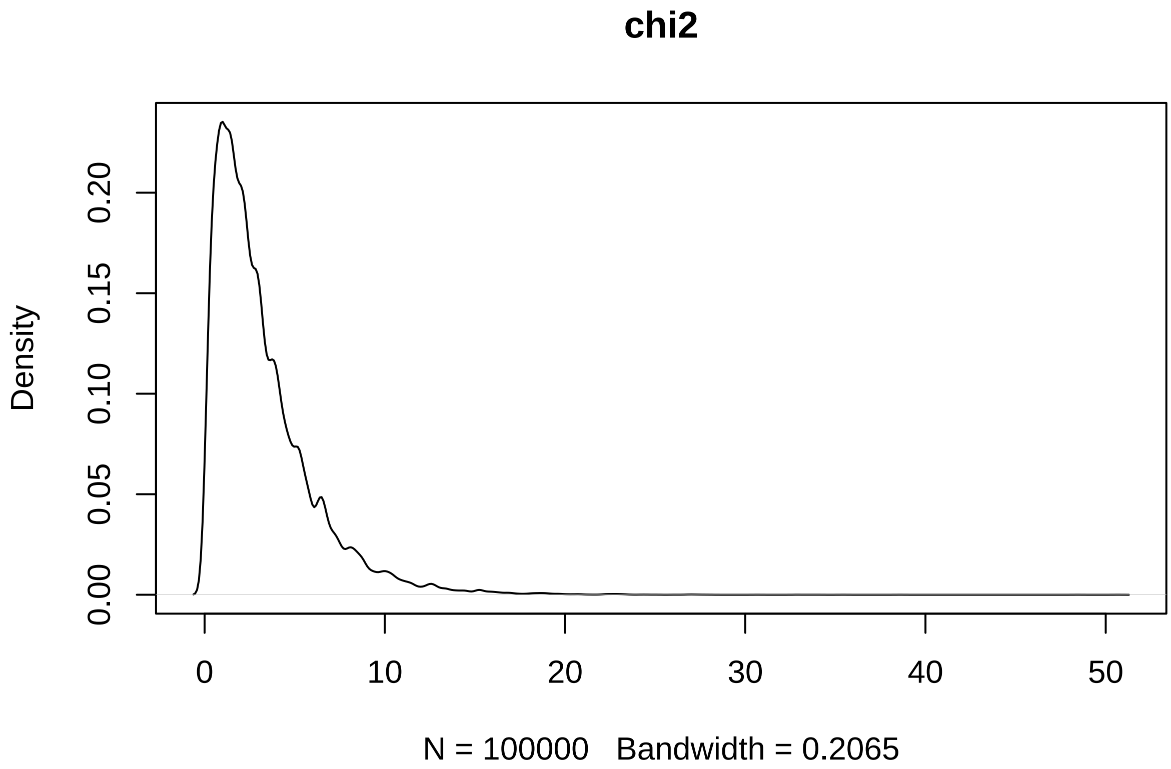
```
Poll_seasons <- data.frame(Spring = 40, Summer = 30, Autumn = 18, Winter = 28)
```

Hint: If there is no preference for a particular season, the frequencies should follow a 0.25 in each category. Generate a (large) population with the expected proportions, sample 116 values (same as the values in the poll) and calculate the chi2, replicate this a large number of times to get an approximate chi2 distribution.

```
# Calculate expected proportions
equal_preferences <- sum(Poll_seasons) * 0.25
population <- rep(c("Spring", "Summer", "Autumn", "Winter"), equal_preferences)

# Initialize simulation variables
num_of_simulation <- 100000
chi_sq_value <- numeric(length = num_of_simulation)

# Generate simulation to get distribution of chi2
for (i in seq_along(chi_sq_value)) {

  # Sample from the population
  sample <- sample(population, 116, replace = TRUE)

  # Calculate chi square value with the two dataset
  observed_frequency <- table(sample)
  chi_sq_value[i] <- sum((observed_frequency - equal_preferences)^2/observed_frequency)
}
```

Use plot(density(. . .)) to visualise the distribution of simulated chi2 values.

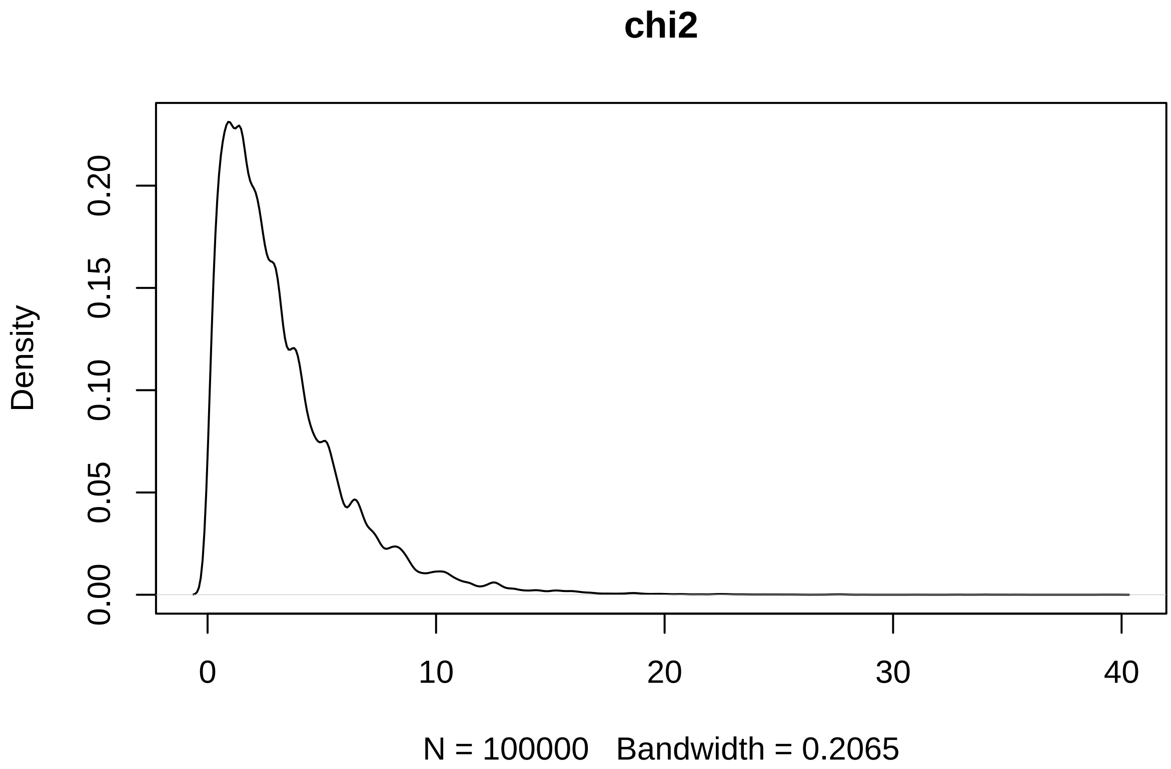
```
plot(density(chi_sq_value), main = "chi2")
```



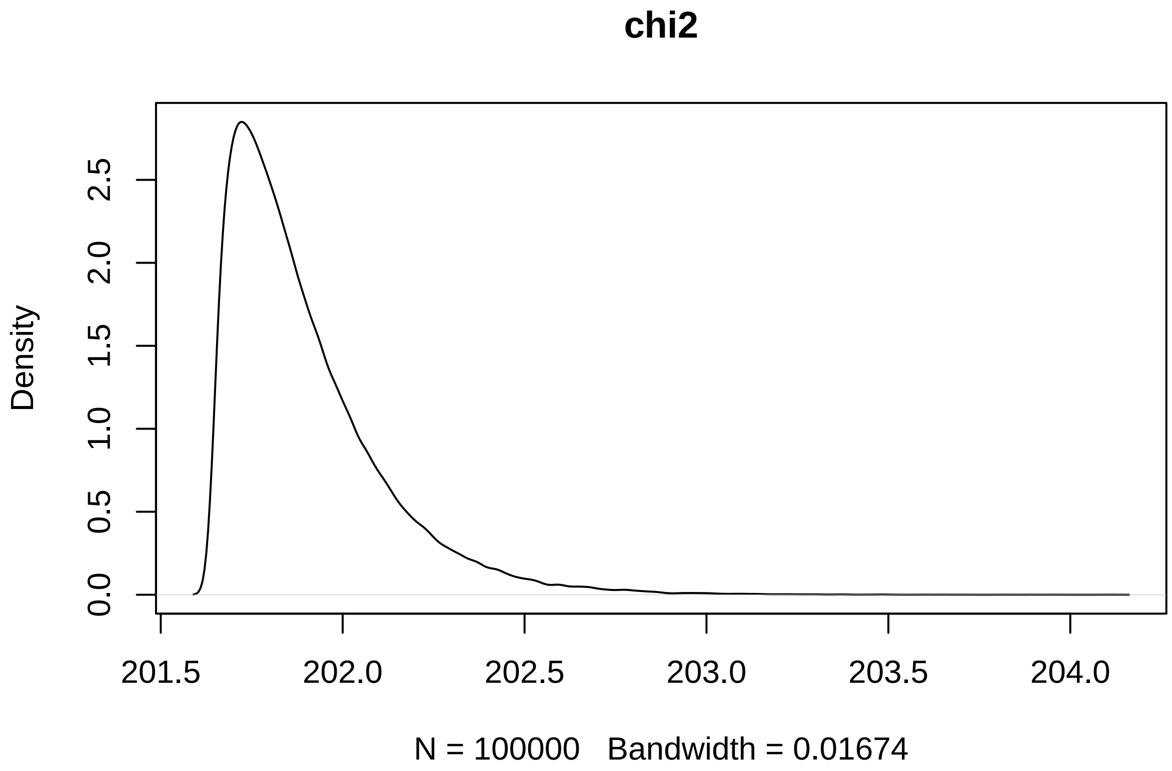
What is the curve like? Do you get the similar probability as in the lecture material?

How the curve changes when considering a higher population size? How about a larger sample size?

```
population1 <- rep(c("Spring", "Summer", "Autumn", "Winter"), 50)
chi_sq_value1 <- numeric(length = num_of_simulation)

for (i in seq_along(chi_sq_value1)) {

  # Sample from the population
  sample <- sample(population1, 116, replace = TRUE)

  # Calculate chi square value with the two dataset
  observed_frequency <- table(sample)
  chi_sq_value1[i] <- sum((observed_frequency - equal_preferences)^2/observed_frequency)

}

plot(density(chi_sq_value1), main = "chi2")
```



```r
chi_sq_value2 <- numeric(length = num_of_simulation)

for (i in seq_along(chi_sq_value2)) {

  # Sample from the population
  sample <- sample(population, 400, replace = TRUE)

  # Calculate chi square value with the two dataset
  observed_frequency <- table(sample)
  chi_sq_value2[i] <- sum((observed_frequency - equal_preferences)^2/observed_frequency)

}

plot(density(chi_sq_value2), main = "chi2")
```



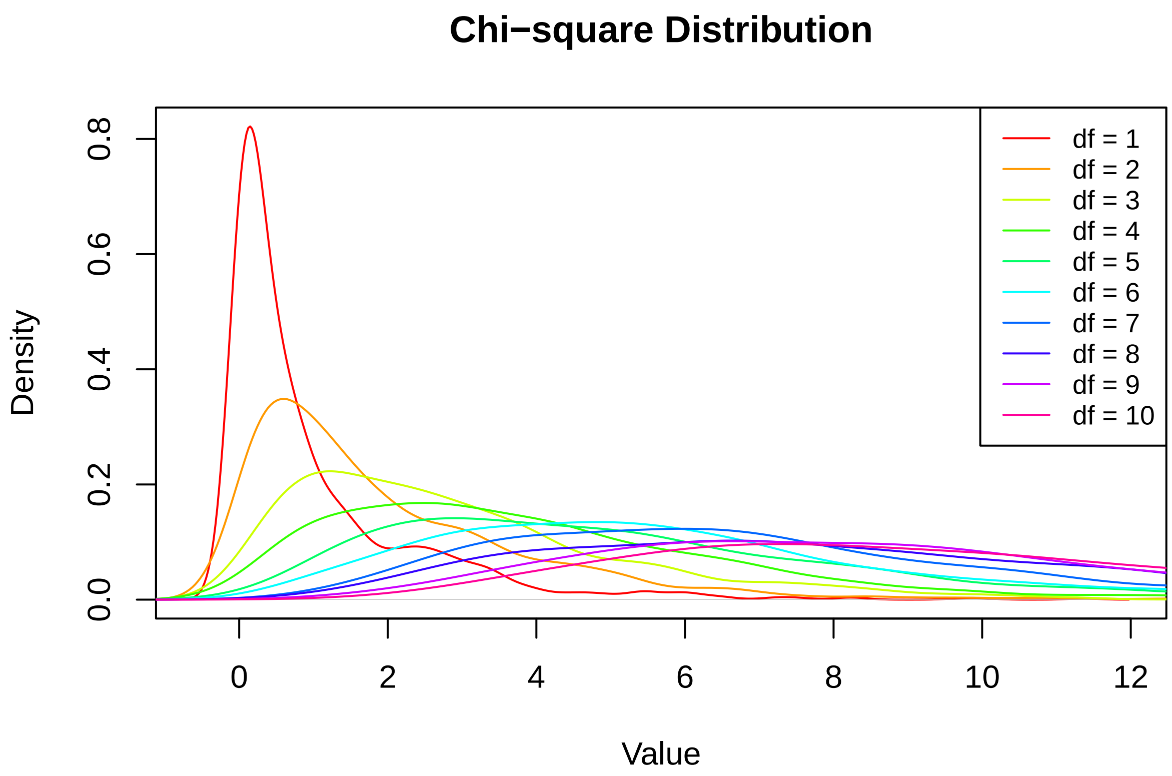
Compare the probability with the result from chisq.test().

```r
# Calculate the p-value
p_value_simulated <- mean(chi_sq_value >= chisq.test(Poll_seasons)$statistic)
```

2. Chi-square distribution and degree of freedom

Generate random chi-square values with different degrees of freedom. Use it as your simulation tool to get the curves as in the lecture. Hint: use rchisq() to directly obtain chi2 values for each degree of freedom.

```r
color <- rainbow(10)
plot(density(rchisq(1000, df = 1)), type = "l", lty = 1, col = color[1], main = "Chi-square Distributio

for (i in 2:10) {
  lines(density(rchisq(1000, df = i, ncp = 0)), col = color[i])
}

legend("topright", legend = c("df = 1", "df = 2", "df = 3", "df = 4", "df = 5", "df = 6", "df = 7", "df
```



3. Chi-square test of homogeneity

Input the data from the two categories (season preference and reported allergy) into a data frame. Visualize the data as bar, balloons and mosaics. Hint: Try mosaicplot()

```
season_preference <- c("Spring", "Summer", "Autumn", "Winter")
reported_allergy <- c("Severe", "Mild", "Sporadic", "Never")
Severe <- data.frame(Spring = 5, Summer = 1, Fall = 1, Winter = 9)
Mild <- data.frame(Spring = 8, Summer = 5, Fall = 2, Winter = 5)
Sporadic <- data.frame(Spring = 9, Summer = 8, Fall = 3, Winter = 9)
Never <- data.frame(Spring = 18, Summer = 16, Fall = 12, Winter = 5)
Two_categories <- rbind(Severe, Mild, Sporadic, Never)
```

Perform chi-square test on the data.

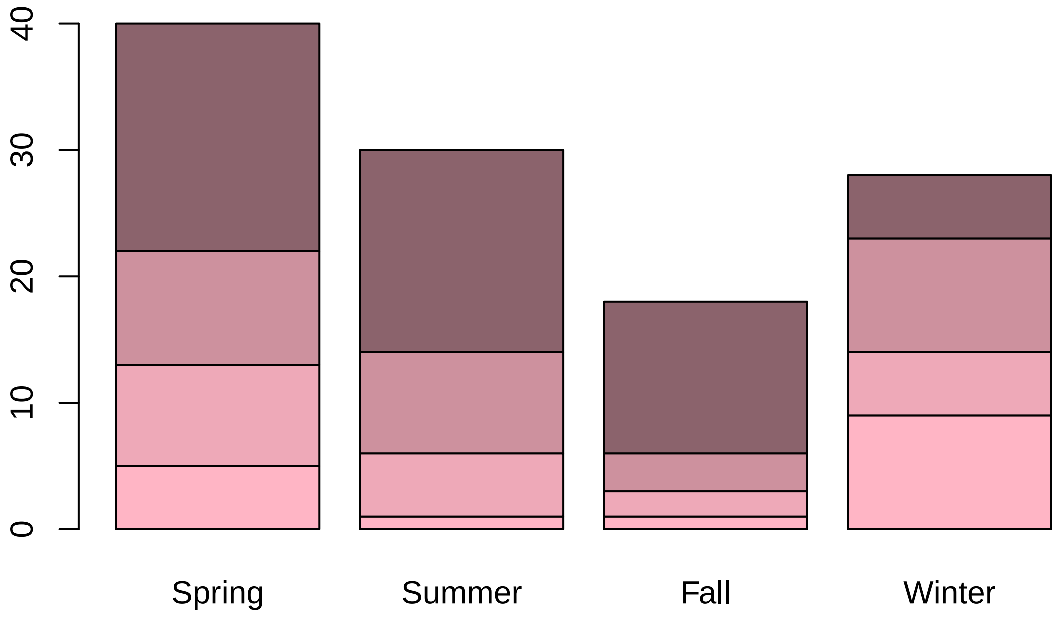
```
chisq.test(Two_categories)
```

```
## Warning in chisq.test(Two_categories): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
```

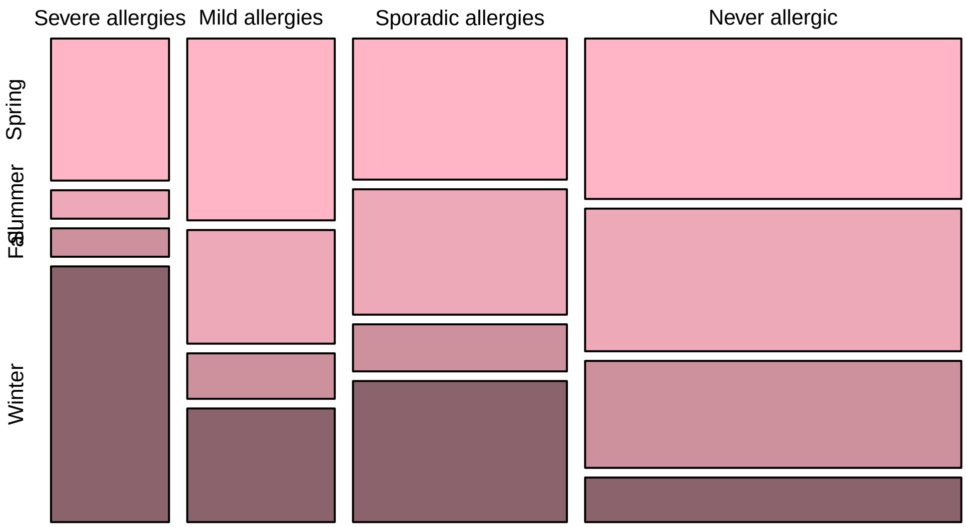
```
## data:  Two_categories
## X-squared = 18.994, df = 9, p-value = 0.02524
```

```
data<-data.frame(Spring=c(5,8,9,18),Summer=c(1,5,8,16),Fall=c(1,2,3,12),Winter=c(9,5,9,5),
                 row.names=c("Severe allergies","Mild allergies","Sporadic allergies","Never allergic")
barplot(as.matrix(data),col=c("pink1","pink2","pink3","pink4"))
```

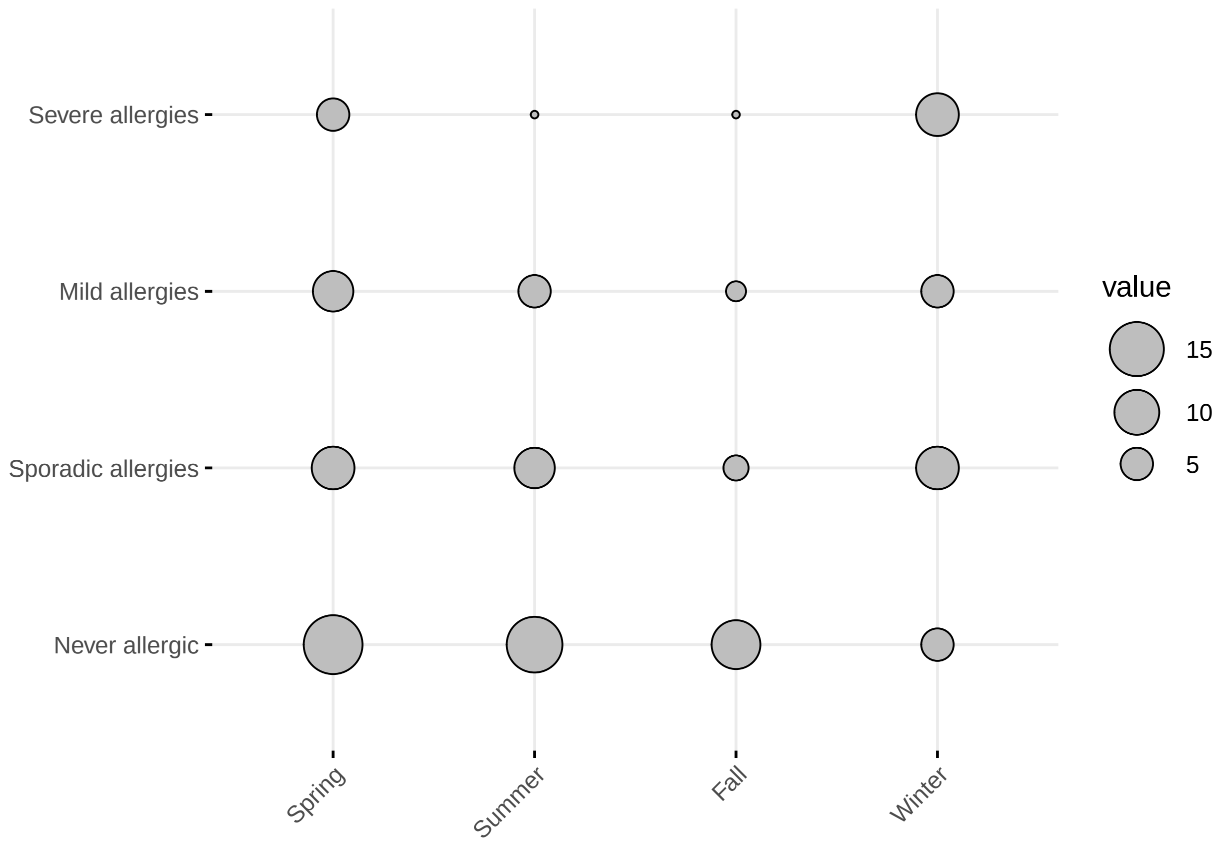


```
mosaicplot(data,col=c("pink1","pink2","pink3","pink4"))
library(ggplot2)
```

, 4L), dimnames = list(c("Severe allergies", "Mild allergies", "Sporadic a



```r
library(ggpubr)
ggballoonplot(data)
```



4. Chi-square test and Fisher's exact test

Input the data from the survival after geneX KO into a matrix.

Perform a chi-square test on the data. Turn off the Yates's continuity correct assigning the correct argument to FALSE. What warning message do you get? If you turn on the correction, what changes?

Perform a Fisher's exact test on the data. Hint: use fisher.test()

```r
gene_data <- data.frame(WT = c(7, 2), KO = c(3, 7), row.names = c("Alive", "Dead"))
fisher.test(gene_data)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  gene_data
## p-value = 0.06978
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    0.7520079 113.4668907
## sample estimates:
## odds ratio
##   7.166282
```