# Analysing Categorical Data

sn

2024-05-22

## Chi-Square

## Categorical Data

### Want to see if there is a difference between an observed variable and expected data?

- Goodness-of-fit tests: The chi-square goodness-of-fit test is used to determine whether a sample of data comes from a population with a specific distribution. For example, it can test whether observed frequencies differ significantly from expected frequencies.
- Fisher's Exact Test: alternative for contingency tables when sample sizes are small.

### Want to see if two or more variables have different distributions of a categorical variable?

- Chi-square homogeneity test/ Fisher's exact test

### Want to see if two or more response variables are independent on an explanatory variable?

- Test for independence: In a contingency table, the chi-square test for independence can determine whether two categorical variables are independent of each other.

pchisq(), chisq.test(), fisher.test()

## Ordinal Variable

### Statistic analysis: compare the median values across samples.

- Wilcoxon test: 1- or 2-sample test, especially useful for paired samples
- Kruskal-Wallis 1-way test: 3- or more sample test, non-parametric alternative to the 1-way ANOVA. kruskal.test(data~group)

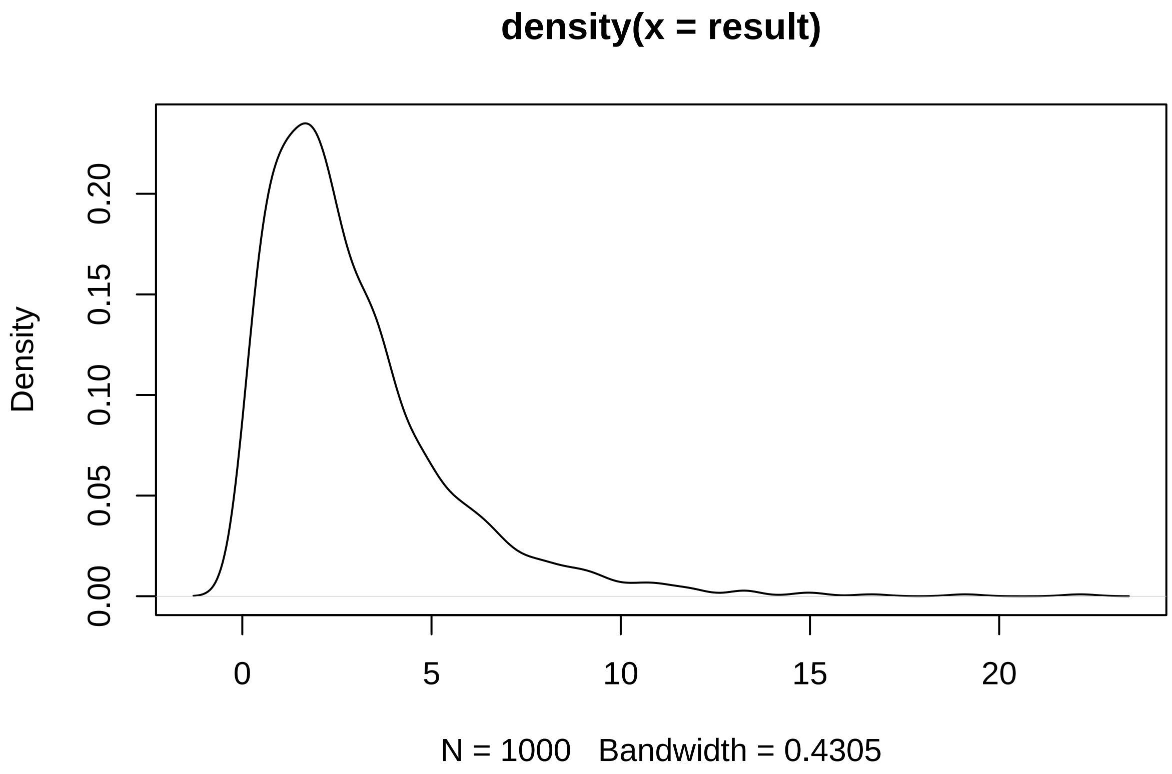
## 1. Simulation of the probability of goodness-of-fit test

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

# expected
equal_preferences <- sum(Poll_seasons) * 0.25

# function
calculate <- function(observed, expected){
  x <- sum((observed-expected)^2/expected)
  return(x)
}

# simulation
result <- replicate(1000,
  {population <- sample(c('Spring', 'Summer', 'Autumn', 'Winter'), 1000, replace = TRUE, prob = c(0.25,
  # simulated
  sample <- sample(population, 116)
  simulated <- data.frame(Spring = sum(sample == 'Spring'), Summer = sum(sample == 'Summer'),
                       Autumn = sum(sample == 'Autumn'), Winter = sum(sample == 'Winter'))
  # calculate
  x <- calculate(simulated, equal_preferences)
  return(x)})
plot(density(result))
```

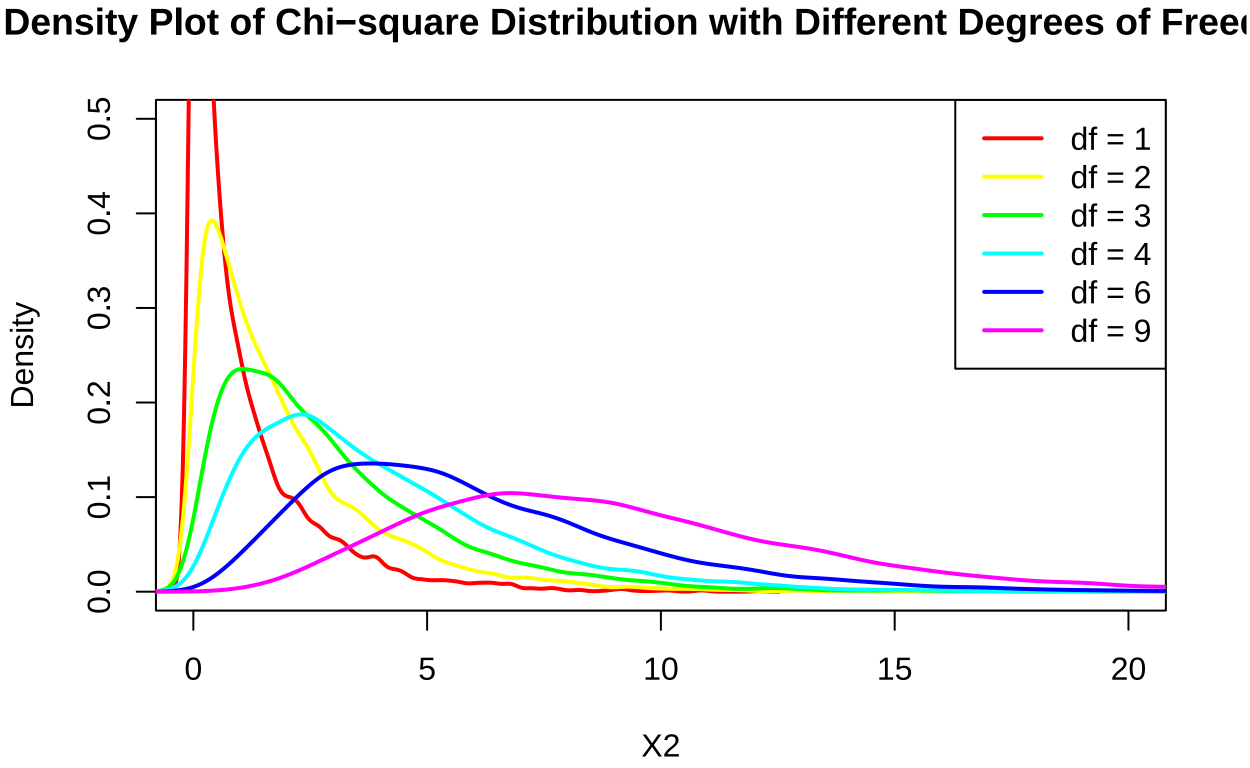


```r
# calculate p.value
observed <- calculate(Poll_seasons, equal_preferences)
p.value <- sum(result >= observed)/1000
```

## 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 x2 values for each degree of freedom.

```r
# degree of freedom
df <- c(1, 2, 3, 4, 6, 9)
# Number of simulations
num_simulations <- 10000
# Generate random chi-square values for each degree of freedom
chi_square_values <- lapply(df, function(d) {
  rchisq(num_simulations, df = d)
})
# Plot density curves for each degree of freedom
plot(NULL, xlim = c(0, 20), ylim = c(0, 0.5), xlab = "X2", ylab = "Density", main = "Density Plot of Chi
colors <- rainbow(length(df))
for (i in 1:length(df)) {
  lines(density(chi_square_values[[i]]), col = colors[i], lwd = 2)
}
legend("topright", legend = paste("df =", df), col = colors, lwd = 2)
```

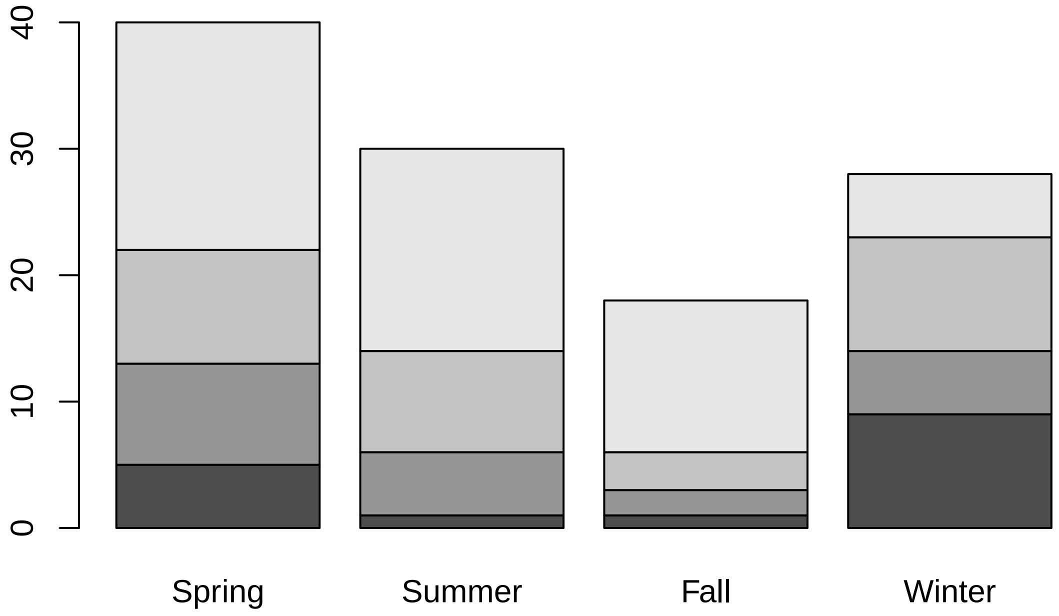


## 3. Chi-square test of homogeneity

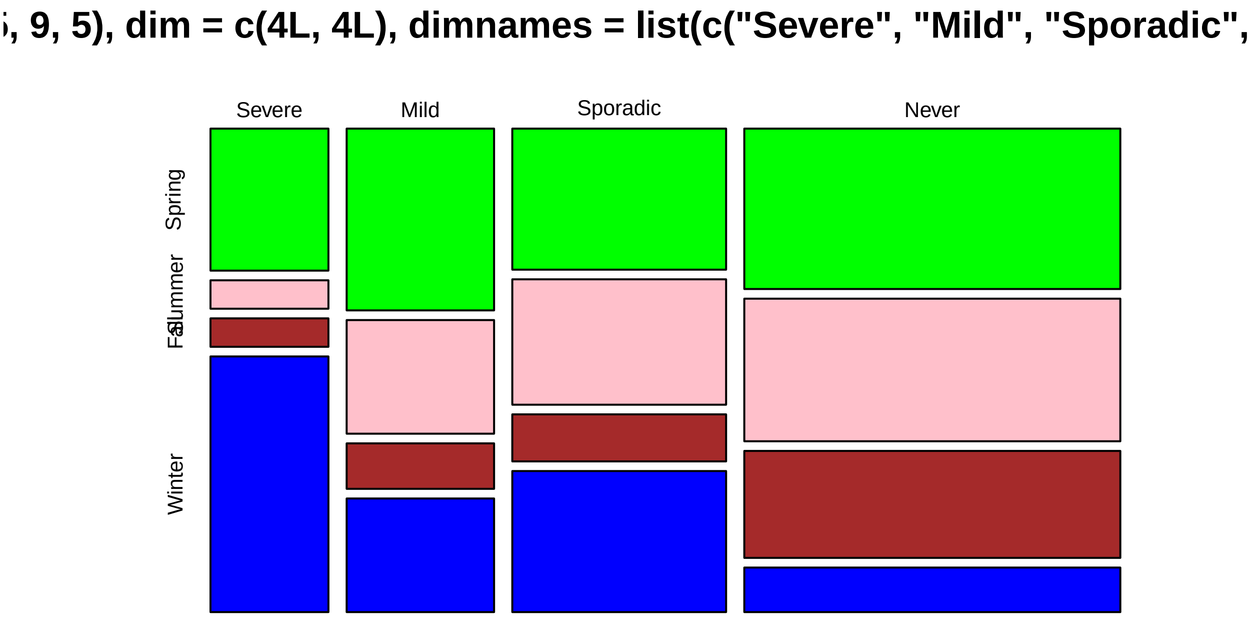
Input the data from the two categories (season preference and reported allergy) into a data frame. Visualise the data as bar, balloons and mosaics. Hint: Try mosaicplot() Perform chi-square test on the data.

```
Severe <- data.frame(Spring = 5, Summer =1, Fall = 1, Winter = 9, row.names = 'Severe' )
Mild <- data.frame(Spring = 8, Summer = 5, Fall = 2, Winter = 5, row.names = 'Mild' )
Sporadic <- data.frame(Spring = 9, Summer = 8, Fall = 3, Winter = 9, row.names = 'Sporadic' )
Never <- data.frame(Spring = 18, Summer = 16, Fall = 12, Winter = 5, row.names = 'Never' )
two_categories <- rbind(Severe, Mild, Sporadic, Never)

# bar
barplot(as.matrix(two_categories))
```



```
# mosaicplot
mosaicplot(two_categories, color = c('green', 'pink', 'brown', 'blue'))
```



```r
# chi-square
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
```

## 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
Alive <- data.frame(WT=7, KO=2, row.names= 'Alive')
Dead <- data.frame(WT=3, KO=7, row.names = 'Dead')
mouse <- rbind(Alive, Dead)
chisq.test(mouse, correct = FALSE)
```

```
## Warning in chisq.test(mouse, correct = FALSE): Chi-squared approximation may be
## incorrect
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  mouse
## X-squared = 4.3372, df = 1, p-value = 0.03729
```

```r
fisher.test(mouse)
```

```
## 
##  Fisher's Exact Test for Count Data
## 
## data:  mouse
## 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
```