Problem Set 2.5: Categorical variables - Solutions

ADS2

Semester 2, 2023/24

1. A candy company produces a new type ice cream with 5 flavours: “Mint”, “Vanilla”,“Chocolate”,“Lemon” and “Orange”. The company sent out samples to 200 people and let them choose their favourite flavour. Here is the result:

Mint	Vanilla	Chocolate	Lemon	Orange
40	32	48	57	23

Based on the result, do you think the company should produce the same amount of ice cream with different flavours?

1.1 What is your null hypothesis. Use a simulation to generate the curve of chi-square values in this situation and get the p-value. Compare the result with *chisq.test()* result.

```
# simulate the population
popu2 <- c(rep("A", 500000), rep("B", 500000), rep("C", 500000),
           rep("D", 500000), rep("E", 500000))
chi2 <-function(){
  sam <- sample(popu2, 200, replace = FALSE)
  n_A <- length(which(sam == "A"))
  n_B <- length(which(sam == "B"))
  n_C <- length(which(sam == "C"))
  n_D <- length(which(sam == "D"))
  n_E <- 200 - n_A - n_B - n_C - n_D
  # 200 / 5 = 40
  X2 <- (n_A-40)^2/40 + (n_B-40)^2/40 + (n_C-40)^2/40 + (n_D-40)^2/40 + (n_E-40)^2/40
  return(X2)
}
chi_sti<- replicate(10000, chi2())
#calculate the p-value
Obs <- c(40, 32, 48, 57, 23)
chi_survey <- sum((Obs-40)^2/40)
length(which(chi_sti >= chi_survey)) / 10000
```

```
## [1] 9e-04
```

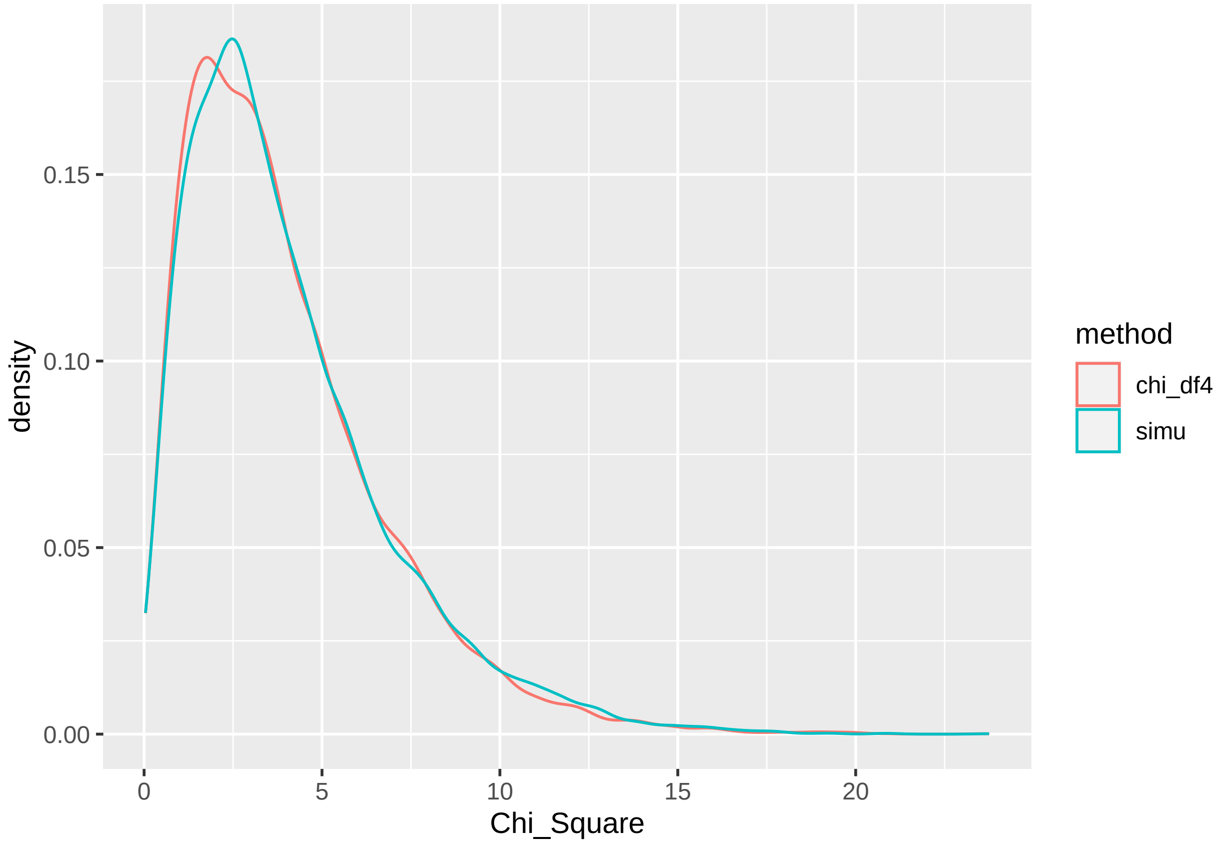
```
## [1] 0.0012
chisq.test(Obs, p = rep(0.2, 5))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  Obs
## X-squared = 17.65, df = 4, p-value = 0.001444
```

```
##
## Chi-squared test for given probabilities
##
## data: Obs
## X-squared = 17.65, df = 4, p-value = 0.001444
```

1.2 What is the degree of freedom (df) in the test? Use *rchisq()* to get the distribution of chi-square with the specific df. Does this curve match the curve in 1.1?

```
#simulate the chi-square values
chi_df4 <- rchisq(10000, df = 4)
#compare the curves
library(tidyr)
library(ggplot2)
chidat <- data.frame(simu = chi_sti, chi_df4)
chidat <- gather(chidat, key = "method", value = "Chi_Square")
ggplot(data = chidat, aes(x = Chi_Square, color = method)) +
  geom_density()
```



2. Analyze the 3-way mouse data from the lecture.

	WT	KO
	Male	Female
Alive	40	34

	WT	KO
——	——	———
Dead	9	7

2.1 Input the data into an array in R. Print it out.

```
# input the data into array
mouse_data <- array(c(40, 9, 34, 7, 20, 15, 25, 20), dim = c(2, 2, 2))
dname <- list(status = c("Alive", "Dead"),
              sex = c("Male", "Female"),
              Genotype = c("WT", "KO"))
dimnames(mouse_data) <- dname
mouse_data
```

```
## , , Genotype = WT
##
##        sex
## status  Male Female
##   Alive   40     34
##   Dead     9      7
##
## , , Genotype = KO
##
##        sex
## status  Male Female
##   Alive   20     25
##   Dead    15     20
```

```
## , , Genotype = WT
##
##        sex
## status  Male Female
##  Alive    40     34
##   Dead     9      7
##
## , , Genotype = KO
##
##        sex
## status  Male Female
##  Alive    20     25
##   Dead    15     20
```

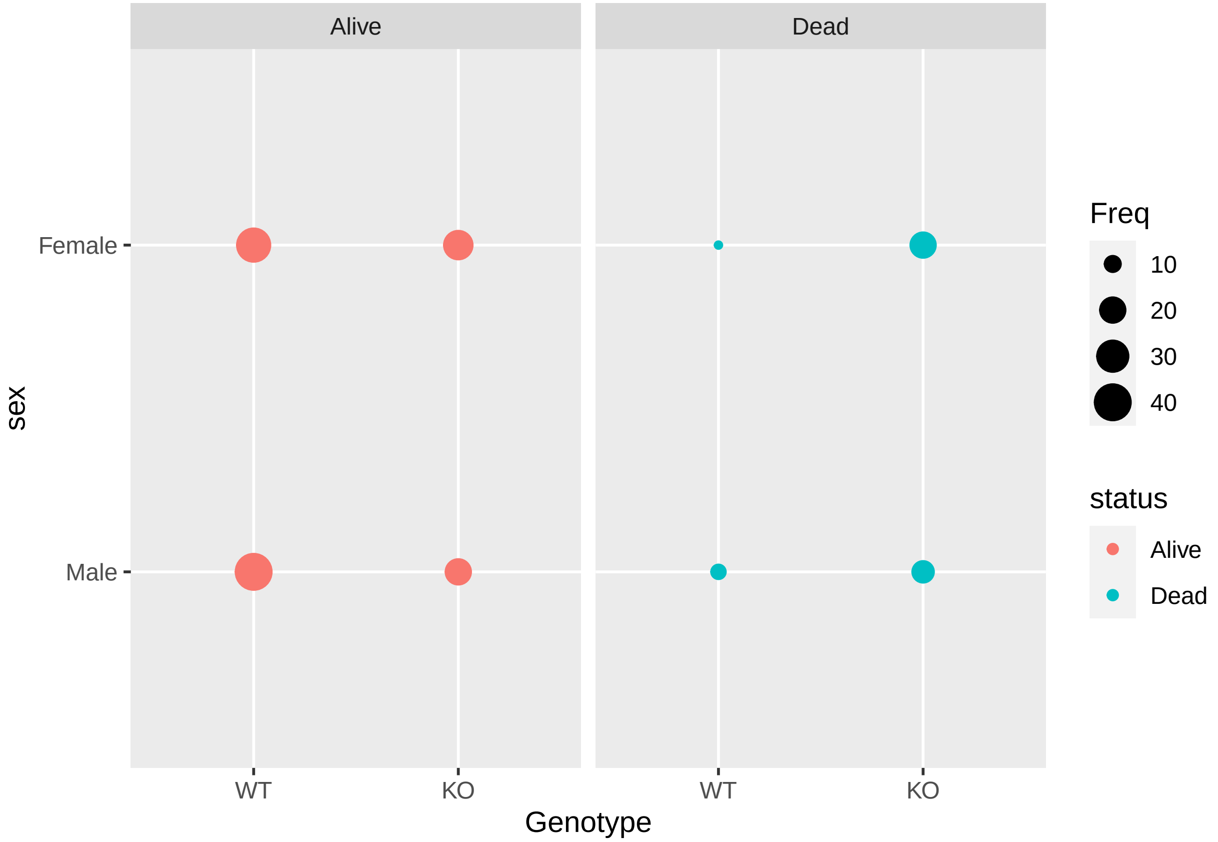
2.2 Apply the chi-square test. Note: You cannot use an 3-dimensional object in the chisq.test(). Convert the array to a table object and find the chi-square test result in the summary of the object. Your result should match the one in the lecture.

```
mouse_data <- as.table(mouse_data)
summary(mouse_data)
```

```
## Number of cases in table: 170
## Number of factors: 3
## Test for independence of all factors:
##  Chisq = 15.765, df = 4, p-value = 0.003351
## Number of cases in table: 170
## Number of factors: 3
## Test for independence of all factors:
## Chisq = 15.765, df = 4, p-value = 0.003351
```

2.3 Use ggplot2 to visualise the data in a dot plot.

```
mdf <- as.data.frame(mouse_data)
ggplot(data = mdf, aes(x = Genotype, y = sex, color = status)) +
  geom_point(aes(size = Freq)) +
  facet_grid(.~status)
```



Additional exercise. Change the question in the mouse experiment. Now we want to know if the survival of mice is dependent on geneX and sex (without assuming geneX and sex are independent on each other). What is the null hypothesis now? How do you reshape the date and perform the test?

```
# The null hypothesis is that the survival is independent on either geneX or sex.
mouse_data2 <- matrix(c(40, 9, 34, 7, 20, 15, 25, 20), nrow = 2)
row.names(mouse_data2) <- c("Alive", "Dead")
colnames(mouse_data2) <- c("WT_Male","WT_Female","KO_Male","KO_Female")
mouse_data2
```

```
##       WT_Male WT_Female KO_Male KO_Female
## Alive      40        34      20        25
## Dead        9         7      15        20
```

```
##       WT_Male WT_Female KO_Male KO_Female
## Alive      40        34      20        25
## Dead        9         7      15        20
chisq.test(mouse_data2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  mouse_data2
## X-squared = 13.646, df = 3, p-value = 0.003429
```

```
## Pearson's Chi-squared test
##
## data: mouse_data2
## X-squared = 13.646, df = 3, p-value = 0.003429
```

Previous versions by Chaochen Wang and Hugo Samano.

Last update by DJ MacGregor in 2024