CATEGORICAL

2024-05-25

Import and organize the data

```
genotype <- read.csv("genotype.csv")
head(genotype)
```

```
##   ID    sex genotype        BD
## 1  1 female      het 2024-2-16
## 2  2   male      het 2024-2-16
## 3  3   male      het 2024-2-16
## 4  5 female      het  2024-3-2
## 5  6 female      mut  2024-2-3
## 6  7   male      het  2024-2-3
```

```
str(genotype)
```

```
## 'data.frame':    80 obs. of  4 variables:
##  $ ID      : int  1 2 3 5 6 7 8 9 10 11 ...
##  $ sex     : chr  "female" "male" "male" "female" ...
##  $ genotype: chr  "het" "het" "het" "het" ...
##  $ BD      : chr  "2024-2-16" "2024-2-16" "2024-2-16" "2024-3-2" ...
```

```
cross_table <- table(genotype$sex, genotype$genotype)
print(cross_table)
```

```
##
##          het mut WT
##   female  26   5  7
##   male    30   2 10
```

Describe the data in a useful way

```
ggplot(data = genotype, aes(x = genotype, fill = sex)) +
  geom_bar(position = "dodge") +
  labs(x = "Genotype", y = "Count", fill = "Sex") +
  ggtitle("Genotype Distribution by Sex")
```

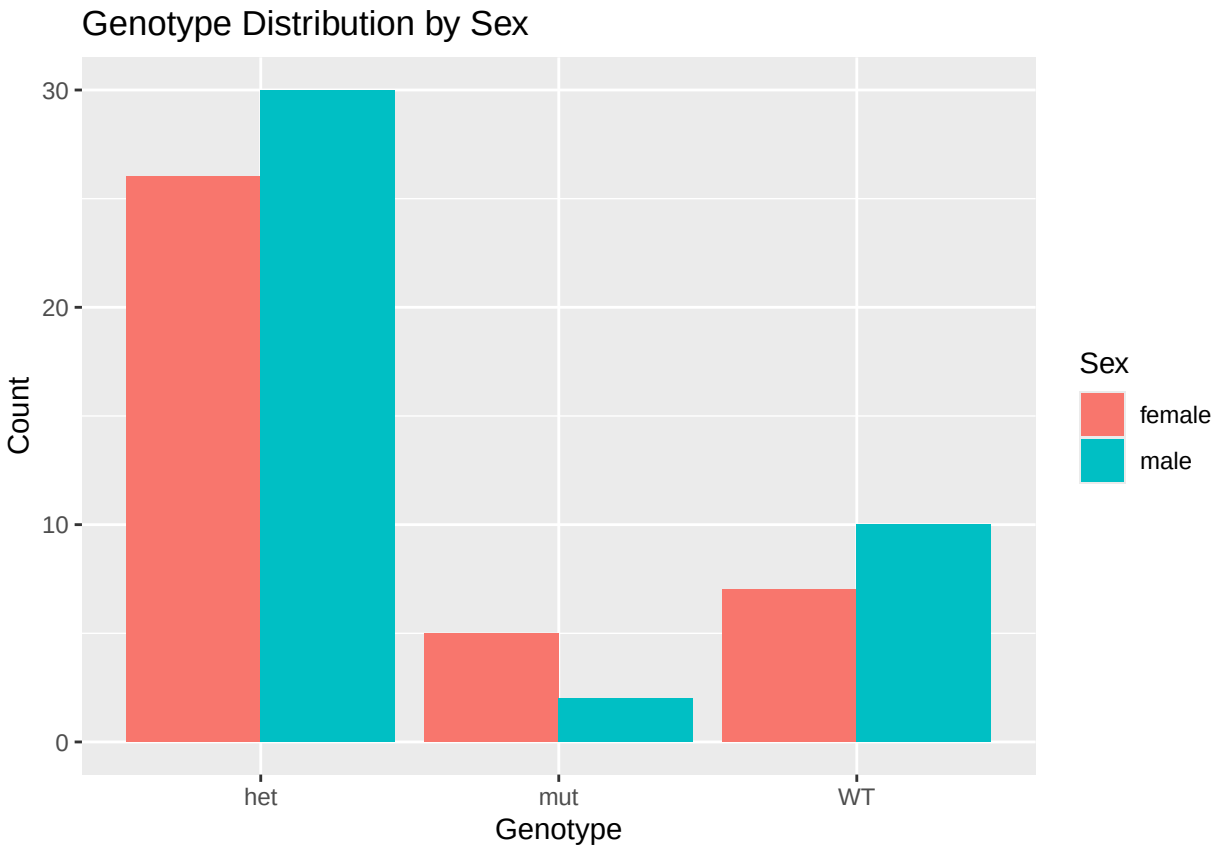


Table 1

	WT	*het*	*mut*
Females	7	26	5
Males	10	30	2

What would you expect under Mendelian inheritance?

In this case, you would expect mice to have males and females as 50/50 and WT, heterozygotes, and mutant mice as 25/50/25.

Table 2

	WT	*het*	*mut*
Females	10	20	10
Males	10	20	10

Choose and justify the appropriate statistical test, state the statistical hypotheses, and carry the test out an appropriate test on whether the mutation affects the survival of mice. Clearly,

it is an expected vs observed distribution. Thus, we need to use the χ^2 goodness-of-fit test.

Assumptions include:

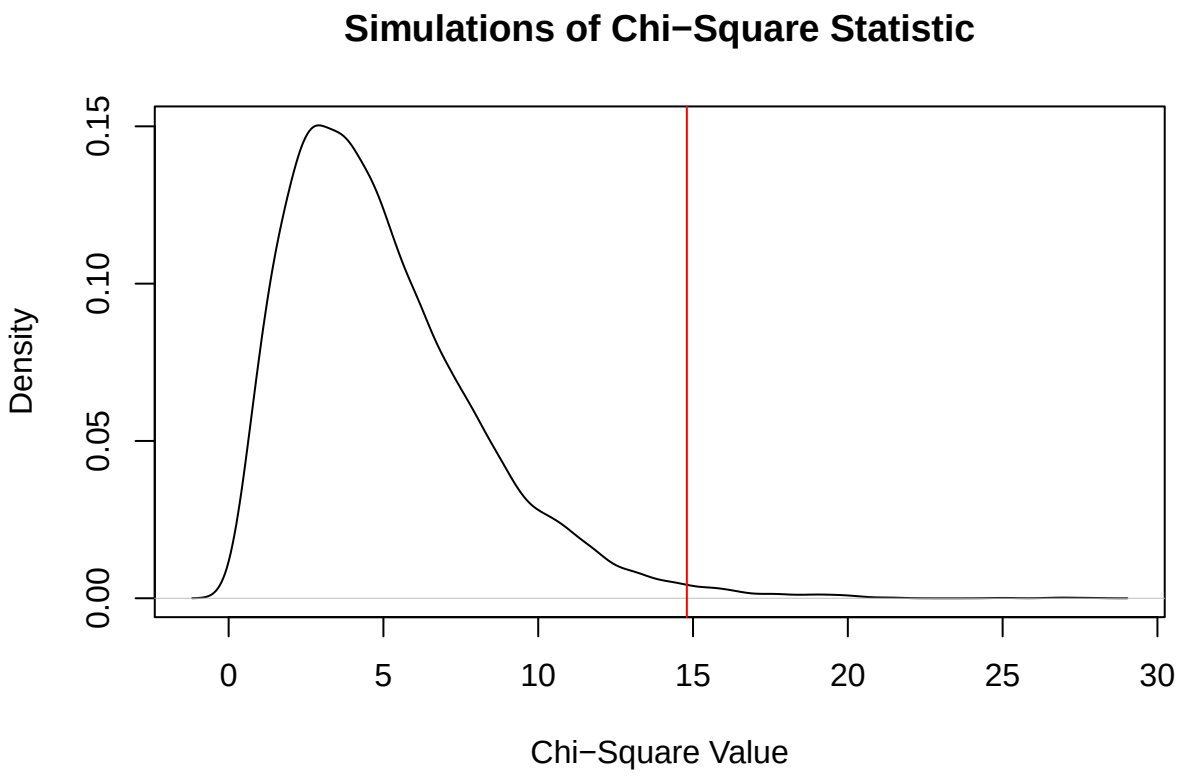
- The variables must be categorical. – Fits
- Observations must be independent. – Can assume from the task. Fits.
- Cells in the contingency table are mutually exclusive.– Fits
- The expected value of cells should be 5 or greater in at least 80% of cells.– See Table 2. Fits. (may not see Fisher's exact test)

State the statistical hypotheses H0: The data follow the expected distribution (Table 2).
HA: The data does not follow the expected distribution (Table 2).

Simulation

```
popu <- c(rep("WT_M", 5000), rep("het_M", 10000),
          rep("mut_M", 5000), rep("WT_FM", 5000),
          rep("het_FM", 10000), rep("mut_FM", 5000))

chi <- function(){
  sam <- sample(popu, 80, replace = FALSE)
  n_WT <- length(which(sam == "WT_M"))
  n_het <- length(which(sam == "het_M"))
  n_mut <- length(which(sam == "mut_M"))
  n_WT_F <- length(which(sam == "WT_FM"))
  n_het_F <- length(which(sam == "het_FM"))
  n_mut_F <- length(which(sam == "mut_FM"))
  X2 <- (n_WT-10)^2/10 + (n_het-20)^2/20 +
    (n_mut-10)^2/10 + (n_WT_F-10)^2/10 +
    (n_het_F-20)^2/20 + (n_mut_F-10)^2/10
  return(X2)
}
chi_sti <- replicate(10000, chi())
Obs <- c(7,20,5,10,30,2)
chi_survey <- sum((7-10)^2/10 + (20-20)^2/20 + (5-10)^2/10 +
                  (10-10)^2/10 + (30-20)^2/20 + (2-10)^2/10)
p.val <- length(which(chi_sti >= chi_survey)) / 10000
plot(density(chi_sti), main = "Simulations of Chi-Square Statistic", xlab = "Chi-Square Value")
abline(v = chi_survey, col = "red")
```




```
print(p.val)
```

```
## [1] 0.0106
```

```
#mock
chisq.test(Obs, simulate.p.value = TRUE, p = c(0.125, 0.25, 0.125, 0.125, 0.25, 0.125))
```

```
##
##  Chi-squared test for given probabilities with simulated p-value (based
##  on 2000 replicates)
##
## data:  Obs
## X-squared = 15.514, df = NA, p-value = 0.007996
```

Carry test out an appropriate test

As we need to run χ^2 for goodness-of-fit, we need to provide the expected values. In chisq.test(), we can do that by providing a matrix of expected probabilities under the p argument.

Table 3: The expected probabilities

	WT	*het*	*mut*
Females	0.125	0.25	0.125
Males	0.125	0.25	0.125

Use all the data for chi goodness fit

```
observed <- c(26, 5, 7, 30, 2, 10)  #directly 1-d OK!!
expected <- c(0.25, 0.125, 0.125, 0.25, 0.125, 0.125)
test_result <- chisq.test(observed, p = expected)
print(test_result)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 16.6, df = 5, p-value = 0.005324
```

Most easy version

```
tulip <- c(81, 50, 27)
res <- chisq.test(tulip, p = c(1/2, 1/3, 1/6))
res
```

```
##
##  Chi-squared test for given probabilities
##
## data:  tulip
## X-squared = 0.20253, df = 2, p-value = 0.9037
```

```
res$p.value
```

```
## [1] 0.9036928
```

p-value = 0.005324 < 0.05, we can reject H0 now.
Thus, the distribution of sex and genotype deviates from the one under Mendelian inheritance law. The mutation did affects the survival of mice.

Chi-square test of independence:the null hypothesis is that two variables (criteria of classifications defined for the same population) are independent, so the distribution of one of them in no way depends on the distribution of the other.
Chi-square test of homogeneity:this test is a generalization of the Z/ t test for the difference between two population proportions. It can be used to test the null hypothesis that several population proportions. It can be used to test the null hypothesis that several populations are homogeneous in the sense that categories of a single qualitative variable, or class intervals of a single quantitative variable, have the same distribution in these populations.
Similarly to the chi-square test of independence, the chi-square test of homogeneity is also based on a contingency table, and the two tests have the same test statistic. However, the two tests differ in terms of the underlying rationale. in the test of independence the expected (joint) frequencies are based on the multiplication rule for independent events, while in the test of homogeneity they are calculated from the pooled sample.

Chi-square test of homogeneity

Question: Is there a difference between the distribution of the allergic reaction and the preferred seasons? Assumptions include:

- The variables must be categorical. – Fits
- Observations must be independent. – Can assume from the task. Fits.
- Cells in the contingency table are mutually exclusive.– Fits
- The expected value of cells should be 5 or greater in at least 80% of cells.– See Table 2. Fits.(may not see Fisher's exact test)

State the statistical hypotheses H0: The distribution of allergic reactions is the same for the people who preferred different seasons.
HA: The distribution of allergic reactions is *not* the same for the people who preferred different seasons.

Table 4: The allergic reaction and the preferred seasons

Category	Spring	Summer	Fall	Winter
Severe	5	1	1	9
Mild	8	5	2	5
Sporadic	9	8	3	9
Never	18	16	12	5

```
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)
chisq.test(Two_categories)# NOTHING IS IN THE LEFTEST COLOMN
```

```
## 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
```

```
#try fisher.test for better convincing
fisher.test(Two_categories,workspace = 2e7)
```

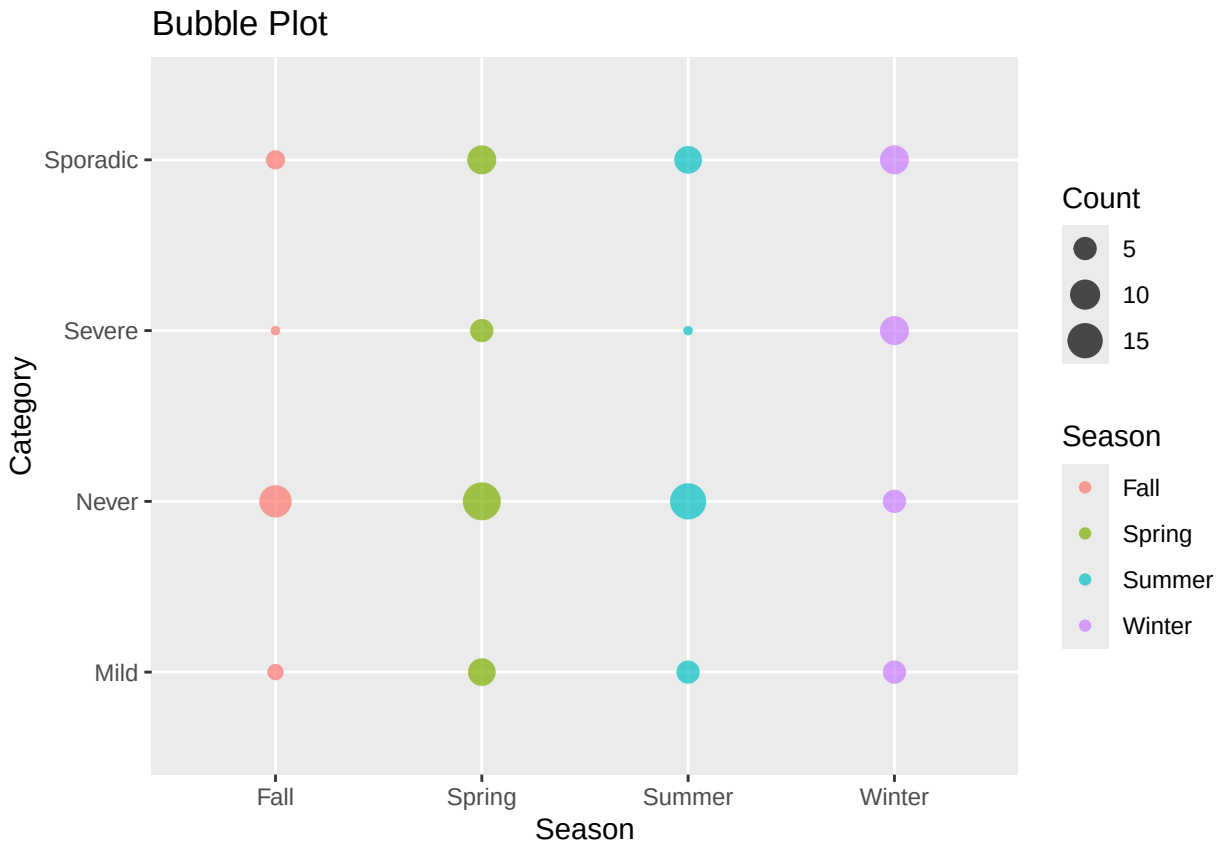
```
##
##  Fisher's Exact Test for Count Data
##
## data:  Two_categories
## p-value = 0.02984
## alternative hypothesis: two.sided
```

```r
Two_categories <- Two_categories %>%
  mutate(Category = c("Severe","Mild","Sporadic","Never"))


Two_categories_long <- Two_categories %>%
  pivot_longer(cols = c(Spring, Summer, Fall, Winter),
               names_to = "Season", values_to = "Count")
print(Two_categories_long)##Format for bubble
```

```
## # A tibble: 16 x 3
##    Category Season Count
##    <chr>    <chr>  <dbl>
##  1 Severe   Spring     5
##  2 Severe   Summer     1
##  3 Severe   Fall       1
##  4 Severe   Winter     9
##  5 Mild     Spring     8
##  6 Mild     Summer     5
##  7 Mild     Fall       2
##  8 Mild     Winter     5
##  9 Sporadic Spring     9
## 10 Sporadic Summer     8
## 11 Sporadic Fall       3
## 12 Sporadic Winter     9
## 13 Never    Spring    18
## 14 Never    Summer    16
## 15 Never    Fall      12
## 16 Never    Winter     5
```

```r
# Bubble Plot
ggplot(Two_categories_long, aes(x = Season, y = Category, size = Count, color = Season)) +
  geom_point(alpha = 0.7) +
  labs(title = "Bubble Plot", x = "Season", y = "Category",
       size = "Count")
```



Chi-square test of independence

Question: Does geneX affect lifespan of mice? H0: The lifespan is independent on genX
H1: The lifespan is dependent on genX

Table 5:Observed data

	WT	KO	Total
Alive	7	2	9
Dead	3	7	10
Total	10	9	19

Table 6:Expected data

	WT	KO
Alive	4.7	4.3
Dead	5.3	4.7

```r
## Fisher (1962, 1970), Criminal convictions of like-sex twins
Convictions <- matrix(c(2, 10, 15, 3), nrow = 2,
                      dimnames =
              list(c("Dizygotic", "Monozygotic"),
               c("Convicted", "Not convicted")))
Convictions
```

```
##             Convicted Not convicted
## Dizygotic           2            15
## Monozygotic        10             3
```

```r
fisher.test(Convictions, alternative = "less")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  Convictions
## p-value = 0.0004652
## alternative hypothesis: true odds ratio is less than 1
## 95 percent confidence interval:
##  0.0000000 0.2849601
## sample estimates:
## odds ratio
## 0.04693661
```

```r
fisher.test(Convictions, conf.int = FALSE)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  Convictions
## p-value = 0.0005367
## alternative hypothesis: true odds ratio is not equal to 1
## sample estimates:
## odds ratio
## 0.04693661
```

```r
fisher.test(Convictions, conf.level = 0.95)$conf.int
```

```
## [1] 0.003325764 0.363182271
## attr(,"conf.level")
## [1] 0.95
```

```r
fisher.test(Convictions, conf.level = 0.99)$conf.int
```

```
## [1] 0.001386333 0.578851645
## attr(,"conf.level")
## [1] 0.99
```

```r
## A r x c table  Agresti (2002, p. 57) Job Satisfaction
Job <- matrix(c(1,2,1,0, 3,3,6,1, 10,10,14,9, 6,7,12,11), 4, 4,
           dimnames = list(income = c("< 15k", "15-25k", "25-40k", "> 40k"),
                           satisfaction = c("VeryD", "LittleD", "ModerateS", "VeryS")))
fisher.test(Job) # 0.7827
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  Job
## p-value = 0.7827
## alternative hypothesis: two.sided
```

```r
fisher.test(Job, simulate.p.value = TRUE, B = 1e5) # also close to 0.78
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  1e+05 replicates)
##
## data:  Job
## p-value = 0.7818
## alternative hypothesis: two.sided
```

Three way Chi-square test

input the data into array

Table 7

Hard to drad a three key table

	WT	WT	KO	KO
	Male	Female	Male	Female
Alive	40	34	20	25
Dead	9	7	15	20

```r
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#Divide into two smaller table
```

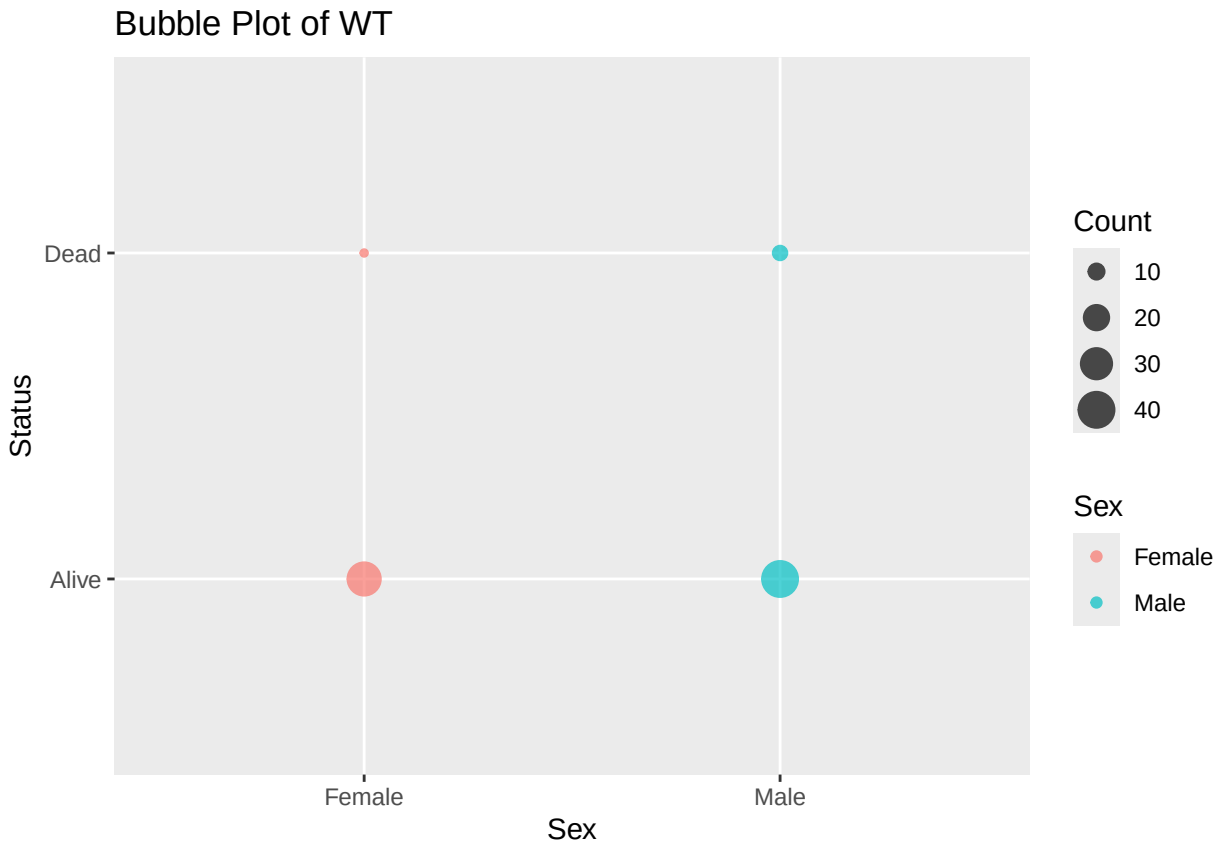
```
## , , Genotype = WT
##
##        sex
## status  Male Female
##   Alive   40     34
```

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

```
##Visulize two photos
mouse_data_WT <- data.frame(Status = c("Alive","Dead"),
                            Male = c(40, 9),
                            Female = c(34, 7))
mouse_data_WT_long <- mouse_data_WT %>%
  pivot_longer(cols = c(Male,Female),
               names_to = "Sex", values_to = "Count")
print(mouse_data_WT_long)
```

```
## # A tibble: 4 x 3
##   Status Sex    Count
##   <chr>  <chr>  <dbl>
## 1 Alive  Male      40
## 2 Alive  Female    34
## 3 Dead   Male       9
## 4 Dead   Female     7
```

```
ggplot(mouse_data_WT_long, aes(x = Sex, y = Status, size = Count, color = Sex)) +
  geom_point(alpha = 0.7) +
  labs(title = "Bubble Plot of WT", x = "Sex", y = "Status",
       size = "Count")
```

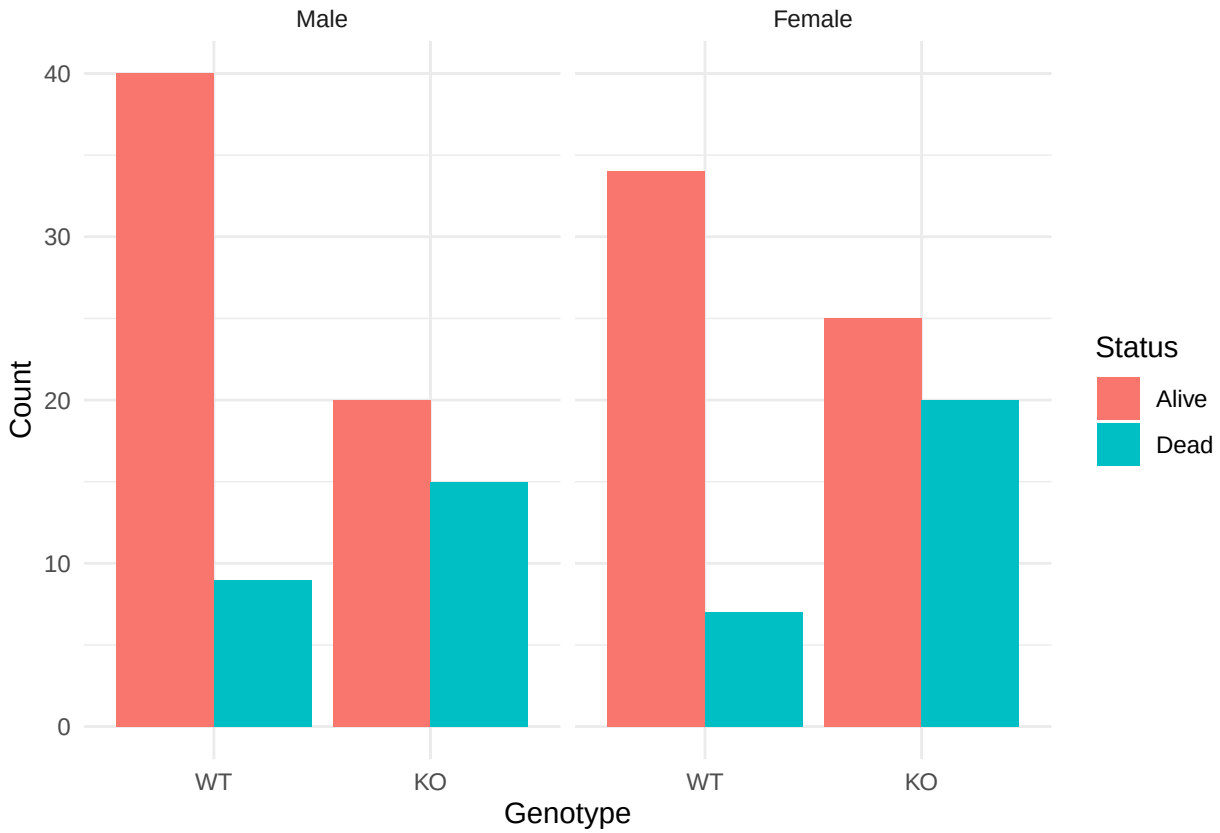



```
## 3 way how to convert visulization

mouse_table <- as.table(mouse_data)
mouse_df <- as.data.frame(mouse_table)
colnames(mouse_df) <- c("status", "sex", "Genotype", "value")
print(mouse_df)
```

```
##   status    sex Genotype value
## 1  Alive   Male       WT    40
## 2   Dead   Male       WT     9
## 3  Alive Female       WT    34
## 4   Dead Female       WT     7
## 5  Alive   Male       KO    20
## 6   Dead   Male       KO    15
## 7  Alive Female       KO    25
## 8   Dead Female       KO    20
```

```
ggplot(mouse_df, aes(x = Genotype, y = value, fill = status)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~sex) +
  labs(x = "Genotype", y = "Count", fill = "Status") +
  theme_minimal()
```



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
```

In this analysis, we utilized a three-way chi-square test to assess the relationship between genotype (WT and KO), sex (Male and Female), and status (Alive and Dead).

H0: There is no interaction among the three factors, implying that genotype, and sex do not affect survival status.
HA: There *is* no interaction among the three factors, implying that genotype, and sex do not affect survival status, indicating that at least one factor influences survival rate.

By organizing the data into a three-dimensional table, we conducted a three-way chi-square test to compare observed values with expected values, determining whether there is an interaction among genotype, sex, and status. The chi-square test results include the chi-square statistic and its corresponding p-value. If the p-value is less than a pre-set significance level (typically 0.05), we reject the null hypothesis, suggesting that

there is an interaction among the factors.

Through a statistical summary, we obtained the results of the three-way chi-square test, allowing us to determine whether we reject the null hypothesis and whether genotype, sex, and status have an interaction effect on survival rate.