```
In [1]: library(tidyverse)
```

```
Warning message in system("timedatectl", intern = TRUE):
“running command 'timedatectl' had status 1”
── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
✔ ggplot2 3.4.0     ✔ purrr   1.0.0
✔ tibble  3.1.8     ✔ dplyr   1.0.10
✔ tidyr   1.2.1     ✔ stringr 1.5.0
✔ readr   2.1.3     ✔ forcats 0.5.2
── Conflicts ────────────────────────────────────────── tidyverse_conflic
ts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
```

Question 1

Part a. State null and alternate hypotheses

H_0: There is no relationship between gender and the decision to consider energy efficiency labeling when purchasing large home appliances

H_1: There is a relationship between gender and the decision to consider energy efficiency labeling when purchasing large home appliances

Part b. Compute test statistic

```
In [2]: # create a table from data
h_app <- matrix(c(115, 32, 8, 135, 16, 8), ncol=3, byrow=TRUE)
colnames(h_app) <- c("Yes","No", "Undecided")
rownames(h_app) <- c("Men","Women")
h_app <- as.table(h_app)
h_app
# perform chi-squared test
chisq.test(h_app)
```

```
      Yes  No Undecided
Men   115  32         8
Women 135  16         8


	Pearson's Chi-squared test

data:  h_app
X-squared = 6.8835, df = 2, p-value = 0.03201
```

Test statistic: $X^2 = 6.8835$

Part c. Compute p-value

p-value = 0.03201

Part d. Conclusion using significance level of α = 0.05

Since the p-value is less than α = 0.05, we reject the null hypothesis and conclude that the data give evidence that there is a relationship between gender and the decision to consider energy efficiency labeling when purchasing large home appliances.

Question 2

Let's denote the proportion of Canadian workers using a certain type of transportation by p, and the proportion of American workers using that same type of transportation by p_0. Then, our hypotheses are:

H_0: The proportion of Canadian workers using a certain type of transportation is equal to the proportion of American workers using that same type of transportation, given by:

$$p_1 = p_{10}, \ldots, p_6 = p_{60}$$

H_1: The proportion of Canadian workers using a certain type of transportation is different from the proportion of American workers using that same type of transportation in at least one transportation category, given by:

$$\text{At least one } p_1 \neq p_{10}, \ldots, p_6 \neq p_{60}$$

In [3]:

```
# create two vectors for data, length is 6
ca_counts = c(320, 100, 30, 20, 10, 20)
usa_proportions = c(.757, .122, .047, .029, .012, .033)
chisq.test(ca_counts, p=usa_proportions)
```

```
	Chi-squared test for given probabilities

data:  ca_counts
X-squared = 41.269, df = 5, p-value = 8.278e-08
```

Test statistic: $X^2 = 41.269$

p-value = $8.278 \text{ x } 10^{-8}$

Conclusion: Since the p-value is less than α = 0.01, we reject the null hypothesis and conclude that the data give evidence that the proportion of Canadian workers using a certain type of transportation is different from the proportion of American workers using that same type of transportation in at least one transportation category.

Question 3

In [4]:
```
# set counter
counter = 0
# set iteration
iter = 1000
# read the pre-made csv
gh = read.csv("q3.csv", header=TRUE, sep=",")
gh
```

A data.frame: 16 × 4

Observation	Greenhouse	Humidity.Level	Plant
<int>	<int>	<chr>	<chr>
1	1	H	Northern Light
2	1	H	Northern Light
3	1	H	B52
4	1	H	B52
5	2	H	Northern Light
6	2	H	Northern Light
7	2	H	B52
8	2	H	B52
9	3	L	Northern Light
10	3	L	Northern Light
11	3	L	B52
12	3	L	B52
13	4	L	Northern Light
14	4	L	Northern Light
15	4	L	B52
16	4	L	B52

In [5]:
```
# count the number of unique of greenhouses
num.gh = unique(gh$Greenhouse) # 1,2,3,4
num.gh = length(num.gh) # 4
num.gh
```

4

```
In [6]: # generate the error terms for between greenhouses (1 kg)
        error1 = rnorm(num.gh, mean=0, sd=1)
        gh$error1 = rep(error1, each=num.gh)
        gh
```

A data.frame: 16 × 5

Observation	Greenhouse	Humidity.Level	Plant	error1
<int>	<int>	<chr>	<chr>	<dbl>
1	1	H	Northern Light	-0.9985371
2	1	H	Northern Light	-0.9985371
3	1	H	B52	-0.9985371
4	1	H	B52	-0.9985371
5	2	H	Northern Light	-1.0684437
6	2	H	Northern Light	-1.0684437
7	2	H	B52	-1.0684437
8	2	H	B52	-1.0684437
9	3	L	Northern Light	1.8086484
10	3	L	Northern Light	1.8086484
11	3	L	B52	1.8086484
12	3	L	B52	1.8086484
13	4	L	Northern Light	0.7632905
14	4	L	Northern Light	0.7632905
15	4	L	B52	0.7632905
16	4	L	B52	0.7632905

In [7]:
```
# generate the error terms for within greenhouses (0.5 kg)
num.plants = length(gh$Observation)
gh$error2 = rnorm(num.plants, mean = 0, sd=0.5)
gh
```

A data.frame: 16 × 6

Observation	Greenhouse	Humidity.Level	Plant	error1	error2
<int>	<int>	<chr>	<chr>	<dbl>	<dbl>
1	1	H	Northern Light	-0.9985371	0.205707113
2	1	H	Northern Light	-0.9985371	-0.263541164
3	1	H	B52	-0.9985371	-0.547475391
4	1	H	B52	-0.9985371	0.044597098
5	2	H	Northern Light	-1.0684437	-0.368651355
6	2	H	Northern Light	-1.0684437	1.185826310
7	2	H	B52	-1.0684437	-0.203617268
8	2	H	B52	-1.0684437	-0.472470849
9	3	L	Northern Light	1.8086484	0.300097391
10	3	L	Northern Light	1.8086484	0.412981977
11	3	L	B52	1.8086484	-0.022067605
12	3	L	B52	1.8086484	-0.235413058
13	4	L	Northern Light	0.7632905	0.066038654
14	4	L	Northern Light	0.7632905	0.594986745
15	4	L	B52	0.7632905	0.211189843
16	4	L	B52	0.7632905	0.001598197

In [8]:
```
# generate the effects for humidity,
# difference can by anything but considered 1 kg
# the df has 8 Hs followed by 8 Ls
gh$humidity.effect[1:8] = 2
gh$humidity.effect[9:16] = 1
# not showing LHS df becaues of larger-size printing issue
gh[, 5:7]
```

A data.frame: 16 × 3

error1	error2	humidity.effect
<dbl>	<dbl>	<dbl>
-0.9985371	0.205707113	2
-0.9985371	-0.263541164	2
-0.9985371	-0.547475391	2
-0.9985371	0.044597098	2
-1.0684437	-0.368651355	2
-1.0684437	1.185826310	2
-1.0684437	-0.203617268	2
-1.0684437	-0.472470849	2
1.8086484	0.300097391	1
1.8086484	0.412981977	1
1.8086484	-0.022067605	1
1.8086484	-0.235413058	1
0.7632905	0.066038654	1
0.7632905	0.594986745	1
0.7632905	0.211189843	1
0.7632905	0.001598197	1

In [9]:
```r
# generate the effects for plant variety, difference is 1 kg
# the df has 2 Northern Lights followed by 2 B52s and is repeated 4 times
plant.effect = c(2,2,1,1)
gh$plant.effect = rep(plant.effect, times=4)
gh[, 5:8]
```

A data.frame: 16 × 4

error1	error2	humidity.effect	plant.effect
<dbl>	<dbl>	<dbl>	<dbl>
-0.9985371	0.205707113	2	2
-0.9985371	-0.263541164	2	2
-0.9985371	-0.547475391	2	1
-0.9985371	0.044597098	2	1
-1.0684437	-0.368651355	2	2
-1.0684437	1.185826310	2	2
-1.0684437	-0.203617268	2	1
-1.0684437	-0.472470849	2	1
1.8086484	0.300097391	1	2
1.8086484	0.412981977	1	2
1.8086484	-0.022067605	1	1
1.8086484	-0.235413058	1	1
0.7632905	0.066038654	1	2
0.7632905	0.594986745	1	2
0.7632905	0.211189843	1	1
0.7632905	0.001598197	1	1

In [10]:
```
# generate the yeilds
# yeild = humidity.effect + plant.effect + error1 + error2 (no interaction)
gh$yeild = gh$humidity.effect + gh$plant.effect + gh$error1 + gh$error2
gh[, 5:9]
```

A data.frame: 16 × 5

error1	error2	humidity.effect	plant.effect	yeild
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
-0.9985371	0.205707113	2	2	3.207170
-0.9985371	-0.263541164	2	2	2.737922
-0.9985371	-0.547475391	2	1	1.453988
-0.9985371	0.044597098	2	1	2.046060
-1.0684437	-0.368651355	2	2	2.562905
-1.0684437	1.185826310	2	2	4.117383
-1.0684437	-0.203617268	2	1	1.727939
-1.0684437	-0.472470849	2	1	1.459085
1.8086484	0.300097391	1	2	5.108746
1.8086484	0.412981977	1	2	5.221630
1.8086484	-0.022067605	1	1	3.786581
1.8086484	-0.235413058	1	1	3.573235
0.7632905	0.066038654	1	2	3.829329
0.7632905	0.594986745	1	2	4.358277
0.7632905	0.211189843	1	1	2.974480
0.7632905	0.001598197	1	1	2.764889

In [11]:
```
# perform ANOVA analysis on the model including the interation
result = aov(yeild~Humidity.Level * Plant +
             Error(Greenhouse:Humidity.Level), data=gh)
x=summary(result)
# view summary
x
# get p-value for plant effect
x = x$`Error: Within`
z = unlist(x)
pval = z[18]
# check if p-value < 0.05
counter = counter + (pval<0.05)
# check yes/no
counter
```

```
Error: Greenhouse:Humidity.Level
               Df Sum Sq Mean Sq F value Pr(>F)
Humidity.Level  1  7.787   7.787   52.01 0.0877 .
Residuals       1  0.150   0.150
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: Within
                     Df Sum Sq Mean Sq F value   Pr(>F)
Humidity.Level        1  3.319   3.319  17.541  0.00186 **
Plant                 1  8.061   8.061  42.605 6.66e-05 ***
Humidity.Level:Plant  1  0.017   0.017   0.089  0.77138
Residuals            10  1.892   0.189
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Pr(>F)2: 1

In [12]:
```
# repeat above procedure 1000 times
# and check the proportions of time we reject null hypothesis
for(i in 2:iter) {
    gh = read.csv("q3.csv", header=TRUE, sep=",")
    num.gh = unique(gh$Greenhouse)
    num.gh = length(num.gh)
    error1 = rnorm(num.gh, mean=0, sd=1)
    gh$error1 = rep(error1, each=num.gh)
    num.plants = length(gh$Observation)
    gh$error2 = rnorm(num.plants, mean = 0, sd=0.5)
    gh$humidity.effect[1:8] = 2
    gh$humidity.effect[9:16] = 1
    plant.effect = c(2,2,1,1)
    gh$plant.effect = rep(plant.effect, times=4)
    gh$yeild = gh$humidity.effect + gh$plant.effect + gh$error1 + gh$error2
    result = aov(yeild~Humidity.Level * Plant +
                 Error(Greenhouse:Humidity.Level), data=gh)
    x = summary(result)
    x = x$`Error: Within`
    z = unlist(x)
    pval = z[18]
    counter = counter + (pval<0.05)
}
```

In [13]:
```
# compute proportion
proportion = counter/iter
proportion
```

Pr(>F)2: 0.957

Based on the simulation with this design, we have a power of 0.95 for the experiment, which is higher than the proposed power of 0.80. This indicates that we can confidently detect the effects of the different plant varieties on the mean yield within 1 kg. Therefore, we will need 2 greenhouses for each level of humidity, resulting in a total of 4 greenhouses for the two levels of humidity (i.e., high and low).