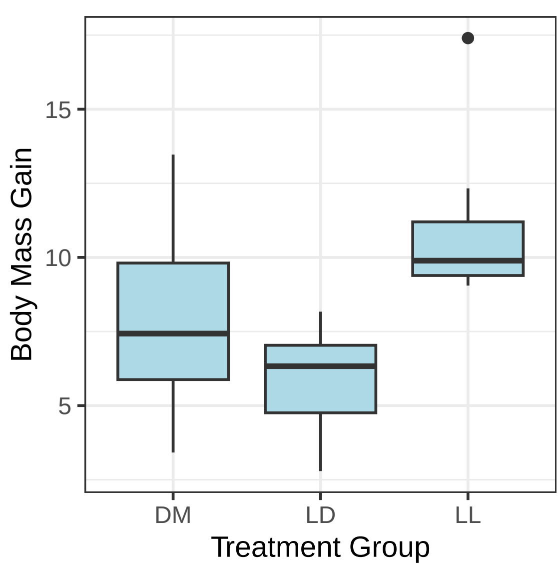
Homework2

Yuki Joyama

```
# libraries
library(tidyverse)
library(ggplot2)
library(perm)

# setup plot theme
theme_set(
  theme_bw() +
    theme(legend.position = "top")
  )

# import data
df = read_csv('./data/light.csv')
```

1. I will use boxplot to show the outcome by treatment group.

```
# boxplot
p1 = ggplot(df, aes(x = Light, y = BMGain)) +
  geom_boxplot(fill = "lightblue") +
  labs(x = "Treatment Group", y = "Body Mass Gain")
p1
```



2. Here I will subset the data to only consider LD (dark light) and LL (bright light) groups.

```r
# filter by these two groups
df2 = df |>
  filter(Light == 'LL' | Light == 'LD')
summary(df2)
```

```
##     Light               BMGain       Corticosterone       DayPct
##  Length:17          Min.   : 2.790   Min.   :  3.00   Min.   :21.85
##  Class :character   1st Qu.: 6.340   1st Qu.: 23.40   1st Qu.:40.50
##  Mode  :character   Median : 9.050   Median : 52.00   Median :61.45
##                     Mean   : 8.618   Mean   : 59.86   Mean   :57.49
##                     3rd Qu.: 9.890   3rd Qu.: 70.47   3rd Qu.:81.60
##                     Max.   :17.400   Max.   :191.22   Max.   :87.26
##   Consumption     GlucoseInt            GTT15           GTT120
##  Min.   :3.387   Length:17          Min.   :226.6   Min.   :118.3
##  1st Qu.:3.791   Class :character   1st Qu.:280.0   1st Qu.:153.7
##  Median :4.240   Mode  :character   Median :348.8   Median :227.3
##  Mean   :4.427                      Mean   :347.8   Mean   :251.8
##  3rd Qu.:4.873                      3rd Qu.:392.4   3rd Qu.:328.7
##  Max.   :7.177                      Max.   :500.0   Max.   :470.2
##     Activity
##  Min.   : 153
##  1st Qu.: 877
##  Median :1649
##  Mean   :2660
##  3rd Qu.:4482
##  Max.   :6702
```

3. I will redefine the variables using generic names as follows:

- LL group: $A = 1$
- LD group: $A = 0$
- BMGain: Y_{obs}

```r
# edit df2 accordingly
df2 = df2 |>
  mutate(A = ifelse(Light == "LL", 1, 0), # add variable A and input 1 for LL, 0 for LD group
         Y_obs = BMGain) |> # add new outcome column
  select(-Light, -BMGain)
```

To evaluate the causal effect of light at night on weight gain, I will need the following quantities:

```r
# define/calculate the quantities
N1 = sum(df2$A == 1)
N0 = sum(df2$A == 0)
N = N1 + N0
Yb_obs1 = df2 |>
  filter(A == 1) |>
  summarize(mean_Y_obs = mean(Y_obs)) |>
  pull(mean_Y_obs)
Yb_obs0 = df2 |>
  filter(A == 0) |>
  summarize(mean_Y_obs = mean(Y_obs)) |>
  pull(mean_Y_obs)
```

- Number of mice in LL group: $N_1 = 9$
- Number of mice in LD group: $N_0 = 8$
- Total number of mice in LL and LD group: $N = 17$
- Mean of the outcome variable for LL group: $\bar{Y}_1^{obs} = 11.01$
- Mean of the outcome variable for LD group: $\bar{Y}_0^{obs} = 5.93$

4.

```r
# calculate t_obs
T_obs = Yb_obs1 - Yb_obs0
```

$T_{obs} = \bar{Y}_1^{\text{obs}} - \bar{Y}_0^{\text{obs}} = 5.08$

5. Under the completely randomized experiment where $N1$ and $N0$ are fixed, there are $\binom{N}{N_1} = 24310$ possibilities for A.

```r
# enumerate them in a matrix
A = chooseMatrix(N, N1)
```

6. The sharp null hypothesis:
   $H_0 : Y_i^1 = Y_i^0 \quad \text{for all } i$
   where Y_i^1 is the potential outcome for mouse i if they are assigned to $A = 1$, and Y_i^0 is the potential outcome for mouse i if they are assigned to $A = 0$.

```r
# create df that has the group assignment based on the first row of matrix A
df3 = df2
df3$A = A[1,]

# calculate t under the first possibility of A, under the sharp null hypothesis
T_stat = mean(df3$Y_obs[df3$A == 1]) - mean(df3$Y_obs[df3$A == 0])
```

Under the sharp null hypothesis, the test statistic under the first row of matrix A is 1.55.

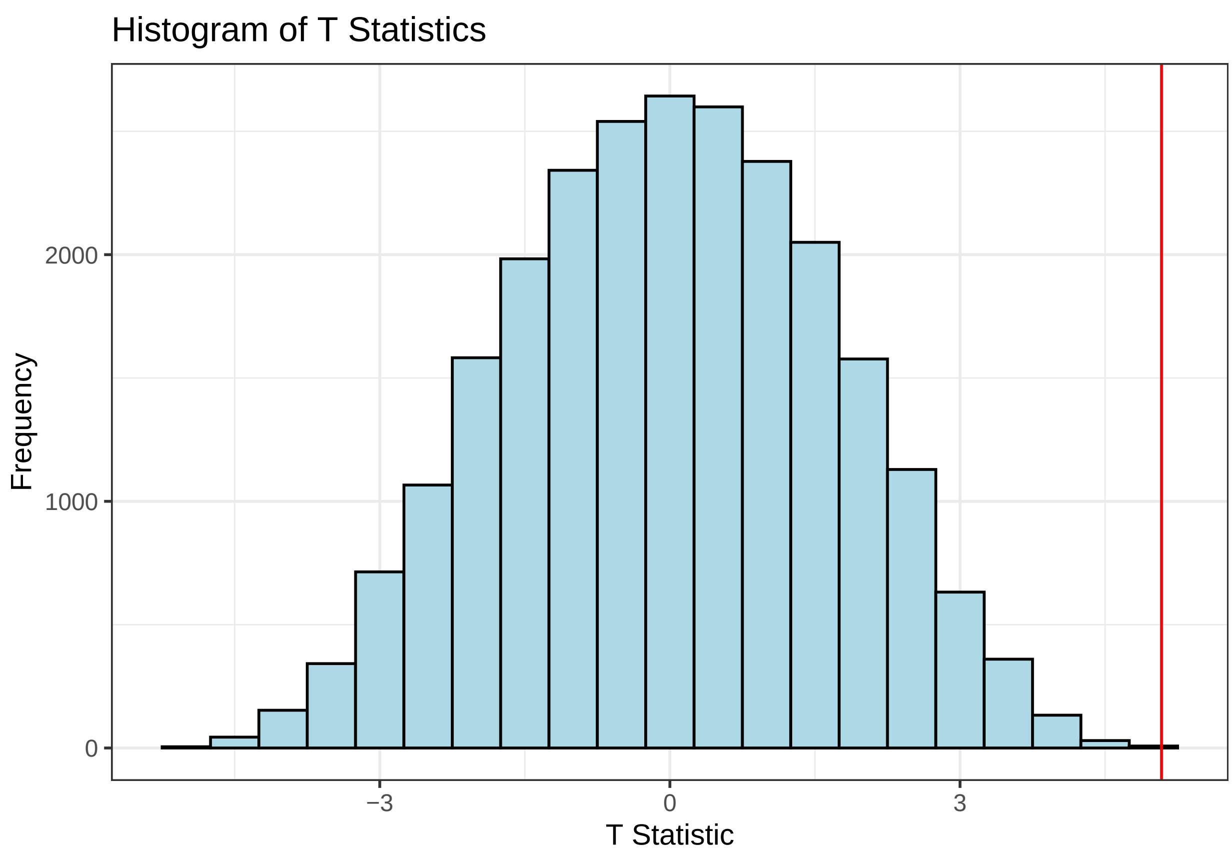
7. I will iterate the process in 6 for all the possibilities of matrix A to obtain the exact randomization distribution for T under the sharp null hypothesis.

```r
# set up df to store T statistic values
rdist = rep(NA, times = A_num)

# iteration
for (i in 1:A_num) {
  df_ite = df2
  df_ite$A = A[i,]
  rdist[i] = mean(df_ite$Y_obs[df_ite$A == 1]) - mean(df_ite$Y_obs[df_ite$A == 0])
}

# show the summary of the distribution
summary(rdist)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -4.98167 -1.21333  0.01444  0.00000  1.21153  5.08375
```

8. The T_{obs} is the red line in the plot.

```
# plot histogram
ggplot(data.frame(t_stat = rdist), aes(x = t_stat)) +
  geom_histogram(binwidth = 0.5, color = "black", fill = "lightblue") +
  geom_vline(aes(xintercept = T_obs), color = "red", size = 0.5) + # add T_obs line
  labs(title = "Histogram of T Statistics", x = "T Statistic", y = "Frequency")
```



9. Based on this distribution, we can obtain the exact p-value by the following formula: $P(T(A,Y) \geq T_{obs} | Y_i^1 - Y_i^0 = 0) = \frac{\sum I(T(A,Y) \geq T_{obs})}{K}$ where $K = \binom{N}{N_1}$.

```
# calculate the exact p-value
p_val = sum(rdist >= T_obs) / length(rdist)
```

The exact p-value is 0.00004113534.

10. Under $\alpha = 0.05$, the exact p-value < 0.05 suggests that the observed test statistic is unlikely to have occurred under the sharp null hypothesis. Therefore, we reject the null hypothesis. We can conclude that exposure to bright light at night compared to darkness at night has a causal effect on changes in body mass among mice over the 8 weeks of the experiment.