2118

2118

2024-01-10

1. Benefits of swimming for long-distance runners

```
swim = read.table("swimming.txt", sep = '\t', header = T)
# head(swimming)
# summary(swim)
# str(swim)
```

1.1 Tidy the data and decide on suitable statistical test.

Any NA?

```
anyNA(swim)
```

```
## [1] FALSE
```

```
which(!complete.cases(swim))
```

```
## integer(0)
```

```
length(rownames(swim[swim$names=="", ]))
```

```
## [1] 0
```

```
length(rownames(swim[swim$before_minutes=="", ]))
```

```
## [1] 0
```

```
length(rownames(swim[swim$before_seconds=="", ]))
```

```
## [1] 0
```

```
length(rownames(swim[swim$after_minutes=="", ]))
```

```
## [1] 0
```

```
length(rownames(swim[swim$after_seconds=="", ]))
```

```
## [1] 0
```

- There is no NA.

Any duplicated?

```
which(duplicated(swim))
```

```
## integer(0)
```

- There is no duplicated.

Anything strange in data types?

```
str(swim)
```

```
## 'data.frame':    45 obs. of  5 variables:
##  $ names         : chr  "Mercedes" "Xavier" "Britney" "Warren" ...
##  $ before_minutes: int  118 137 127 126 125 138 124 128 125 142 ...
##  $ before_seconds: int  30 32 51 0 5 13 14 34 52 45 ...
##  $ after_minutes : int  115 145 130 127 125 147 124 131 127 154 ...
##  $ after_seconds : int  4 59 15 16 46 6 24 25 2 29 ...
```

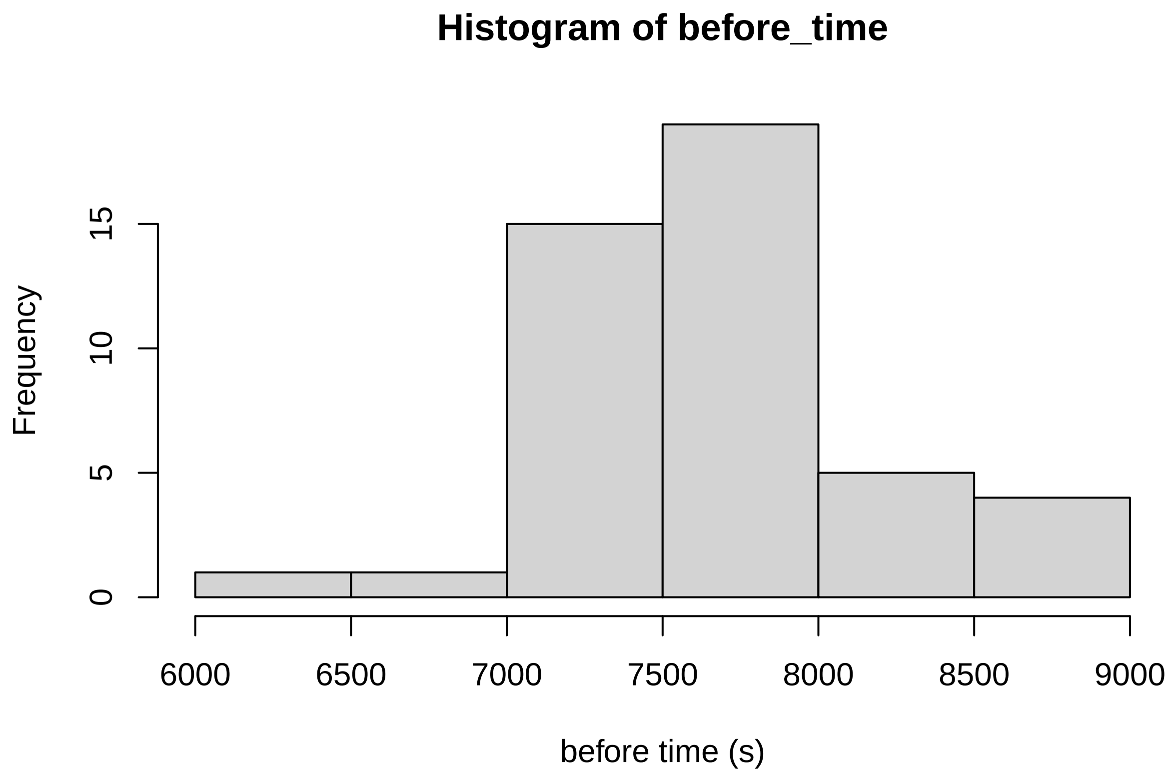
- Nothing strange in datatypes.

Any outlier? How are the before and after time distributed?

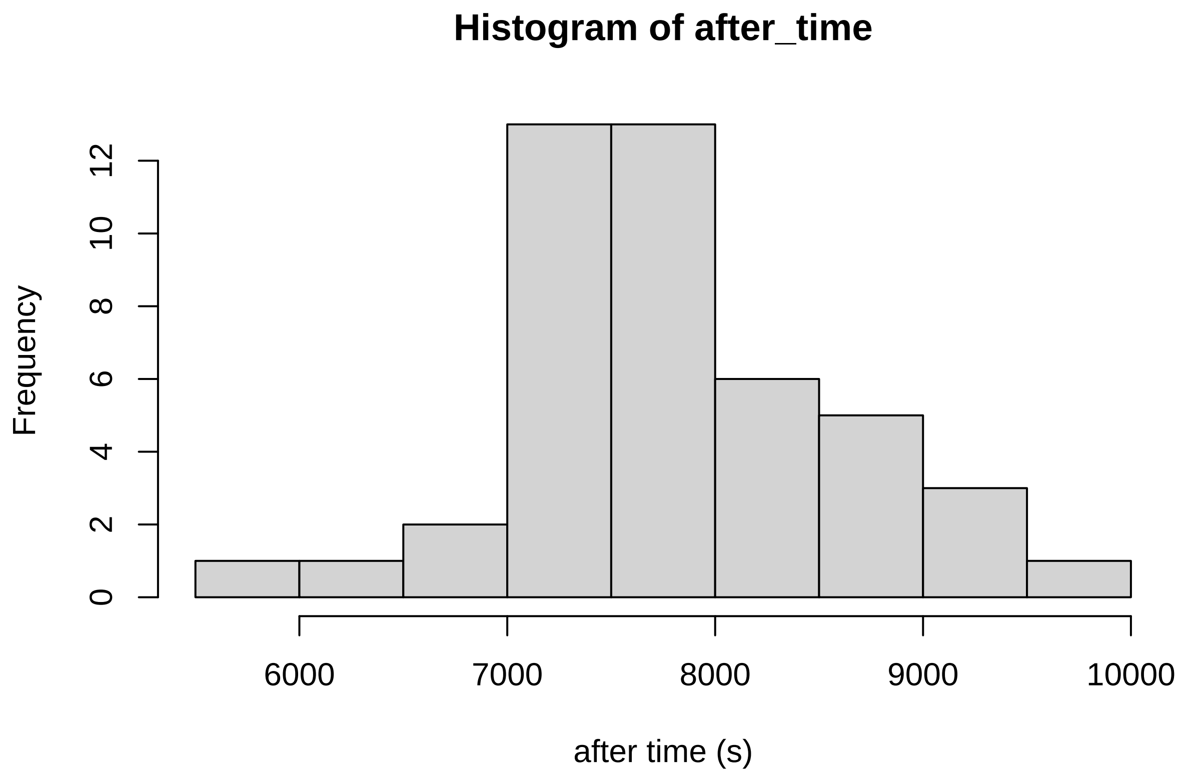
```
swim[swim$before_minutes < 0 | swim$after_minutes < 0 |
     swim$before_seconds < 0 | swim$after_seconds < 0 , ]
```

```
## [1] names          before_minutes before_seconds after_minutes  after_seconds
## <0 rows> (or 0-length row.names)
```

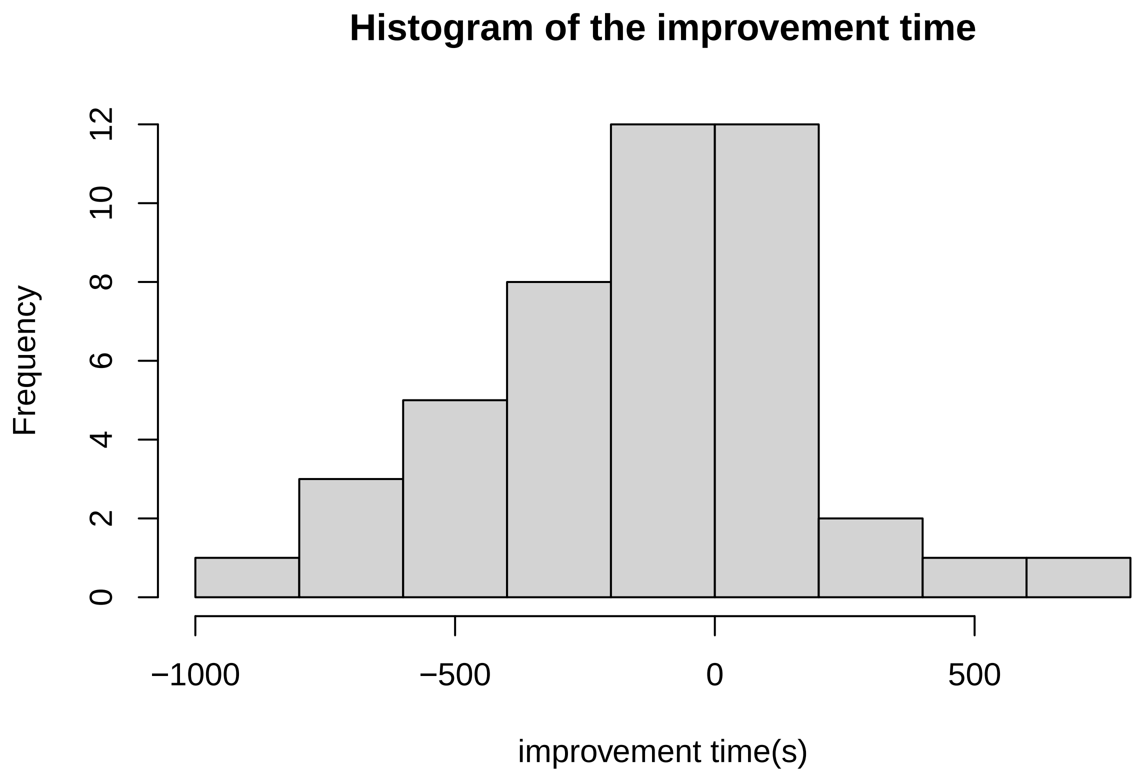
```
before_time = swim$before_minutes * 60 + swim$before_seconds
before_time = as.integer(before_time)
after_time = swim$after_minutes * 60 + swim$after_seconds
after_time = as.integer(after_time)
hist(before_time, xlab = "before time (s)")
```




```r
hist(after_time, xlab = "after time (s)")
```




```r
hist(before_time - after_time, xlab = "improvement time(s)",
     main = "Histogram of the improvement time")
```




```
swim2 = data.frame(swim$names, before_time, after_time)
# head(swim2)
```

- There is no outlier.
- The before_time, the after_time and the improvement time (the difference between the after_time and the before_time) are all normally distributed.
- Since every person in the data has a before_time and an after_time, the 2 values are paired.
- We decided to use the paired 2-sample t-test.

1.2 The null and alternative hypotheses

- The null hypothesis (H0): The time used for the half-marathon after the swimming training is no shorter than that before the swimming training.
- The alternative hypothesis (HA) : The time used for the half-marathon after the swimming training is shorter than that before the swimming training.

1.3 Is there a statistically significant improvement on runners' times after swimming?

```
t.test(after_time, before_time, paired = T, alternative = "less")
```

```
##
```

```
##  Paired t-test
##
## data:  after_time and before_time
## t = 2.8221, df = 44, p-value = 0.9964
## alternative hypothesis: true mean difference is less than 0
## 95 percent confidence interval:
##      -Inf 206.0872
## sample estimates:
## mean difference
##        129.1778
```

- $p > 0.05$
- We cannot reject the null hypothesis.
- There is insufficient evidence to conclude that the time used for the half-marathon after the swimming training is shorter than that before the swimming training.
- Therefore, there is not a statistically significant improvement on runners' times after swimming.

2. Number of emergency room admissions

2.1 Import the dataset and plot the data in a useful way

```
hosp = read.csv("hospital_admissions.csv")
# head(hosp)
# str(hosp)
hosp$week = as.factor(hosp$week)
hosp$weekday = as.factor(hosp$weekday)

hosp1 = aggregate(hosp$patients_per_hour,
                  by = list(hosp$week, hosp$weekday),
                  FUN = sum)
names(hosp1)[1] = "week"
names(hosp1)[2] = "weekday"
names(hosp1)[3] = "patients"
# str(hosp1)
# summary(hosp1)
g2.1 = ggplot(data = hosp1[hosp1$weekday == "Monday",],
            mapping = aes(x = week, y = patients)
            )
g2.1 = g2.1 + geom_bar(stat="identity", fill = "orange")
g2.1
```

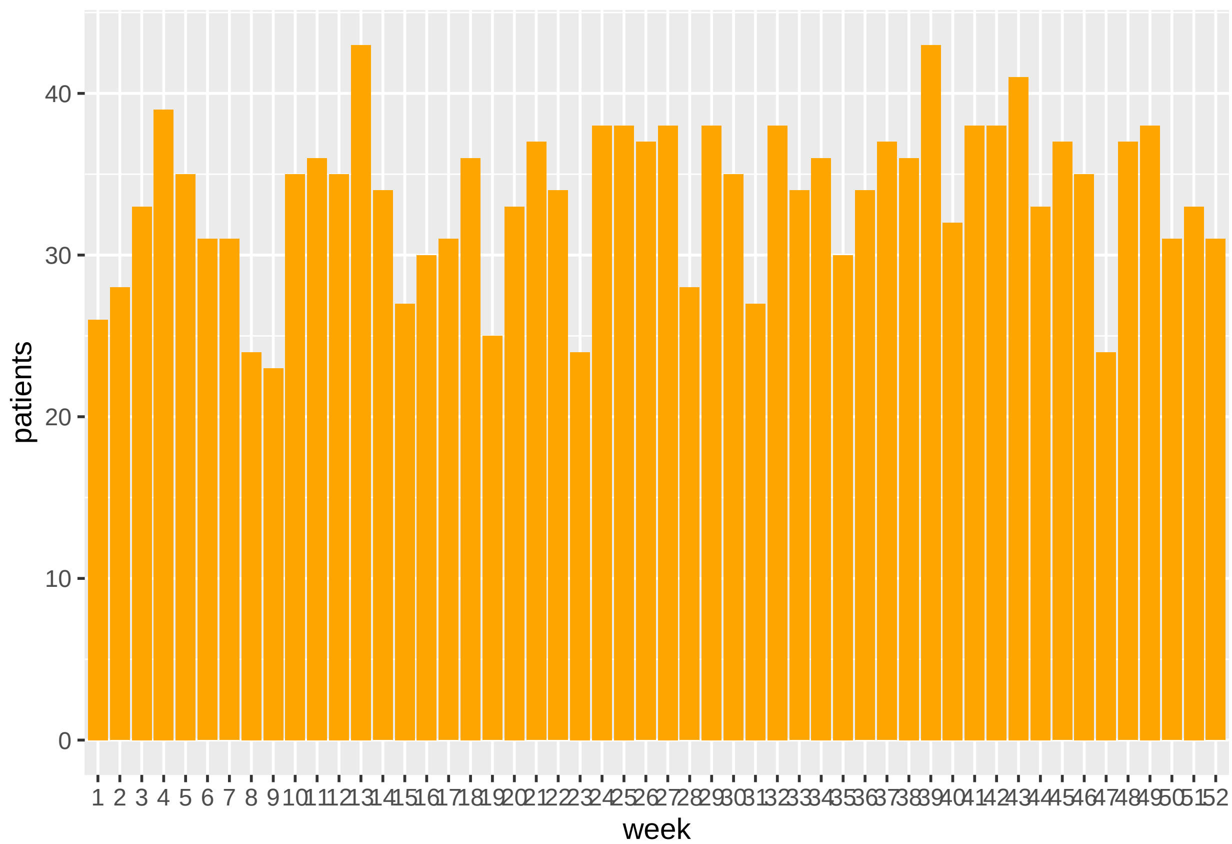


Figure 2.1: Patients on Monday during the year.

```
g2.2 = ggplot(data = hosp1[hosp1$weekday == "Sunday",],
              mapping = aes(x = week, y = patients)
              )
g2.2 = g2.2 + geom_bar(stat="identity", fill = "purple")
g2.2
```

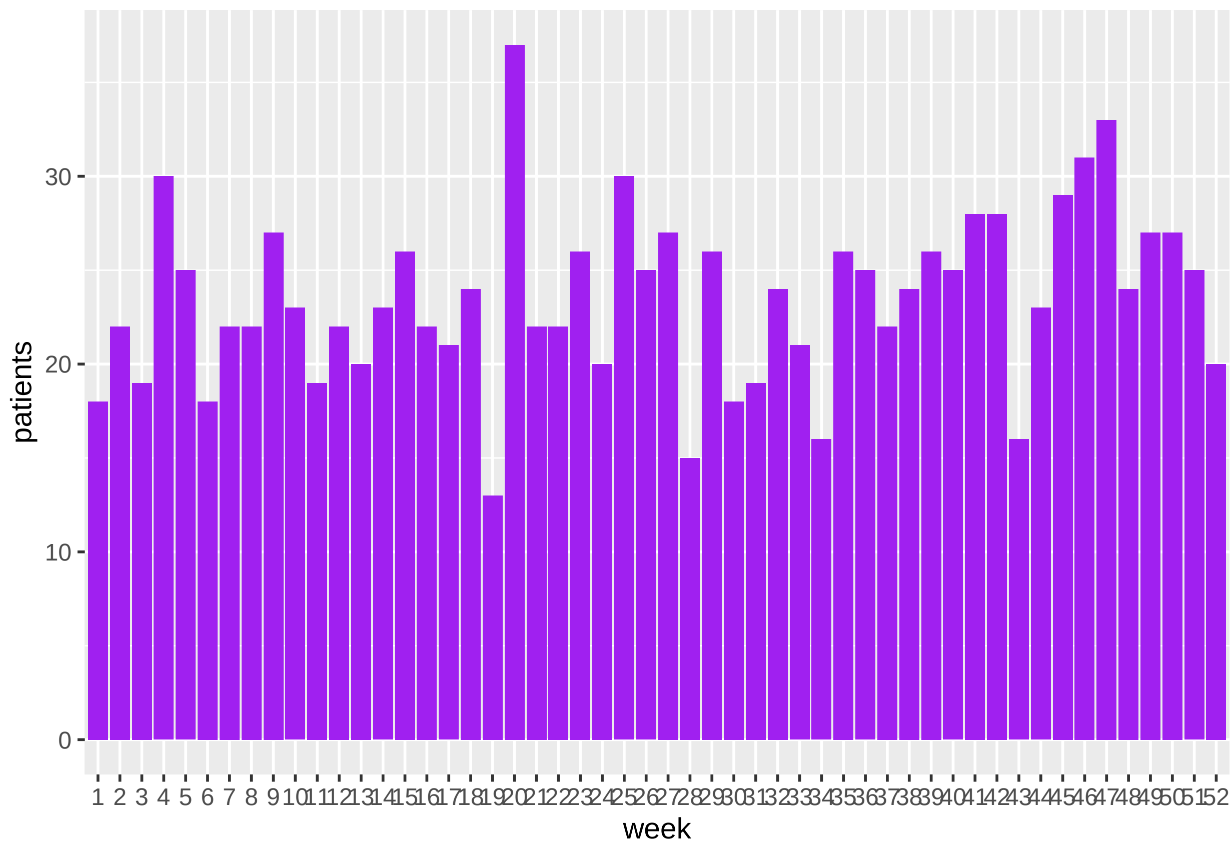


Figure 2.2: Patients on Sunday during the year.

Is there a difference in patient admission rates between Mondays and Sundays?

- We first form the null (H0) and alternative (HA) hypothesis for this question.
- H0: The patient admission rate on Mondays is not higher then that on Sundays.
- HA: The patient admission rate on Mondays is higher then that on Sundays.
- We can see that the patients on both Monday and Sunday during the year are not normally distribution.
- Therefore, We use paired Wilcoxon test.

```
# hosp1[hosp1$weekday == "Monday",]$week ==
# hosp1[hosp1$weekday == "Sunday",]$week
wilcox.test(hosp1[hosp1$weekday == "Monday",]$patients,
            hosp1[hosp1$weekday == "Sunday",]$patients,
            alternative = 'greater',paired = T)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  hosp1[hosp1$weekday == "Monday", ]$patients and hosp1[hosp1$weekday == "Sunday", ]$patients
## V = 1345, p-value = 1.15e-09
## alternative hypothesis: true location shift is greater than 0
```

- p-value < 0.05

- We reject H0.
- There is sufficient evidence to conclude that the patient admission rate on Mondays is higher than that on Sundays.
- Therefore, there is a significant difference in patient admission rates between Mondays and Sundays.

2.3 Based on your findings, what advice would you give Dr. Horsey?

- We should arrange more staff on Mondays than on Sundays.

3. Spinal cord injury and novel biomaterials

3.1 Import, arrange the data (merge both pieces of data and make the data possible to analyse), and makeit suitable for analysis.

```
data1 = read.csv("SCI_before.csv")
data2 = read.csv("SCI_after.csv")
# head(data1)
# head(data2)
data1$patient_ID = as.factor(data1$patient_ID)
# levels(data1$patient_ID)
data2$patient_ID = as.factor(data2$patient_ID)
# summary(data1)
# summary(data2)
# library(dplyr)
# levels(data2$patient_ID)
data1 = arrange(data1,data1$patient_ID)
data2 = arrange(data2,data2$patient_ID)
# data1$patient_ID == data2$patient_ID
data = cbind(data1, data2$AIS_after)
names(data)[3] = "AIS_after"
# head(data)
```

Any NA?

```
anyNA(data)
```

```
## [1] FALSE
```

- No NA.

Any duplicated?

```
idx1 = which(duplicated(data))
idx2 = which(duplicated(data, fromLast = T))
idx1
```

```
## [1]  3  6  8 10 18 31 35 37
```

```r
# data[c(idx1, idx2),  ]
data = data[-idx1, ]
```

Data type?

```r
# summary(data)
data$AIS_before = as.factor(data$AIS_before)
data$AIS_after = as.factor(data$AIS_after)
```

Documentation: Remove 8 duplicated rows.

3.2 Check your data carefully. Identify features of the data and discuss your conclusions. Make illustrative plots.

```r
g3.1 = ggplot(data = data,
            mapping = aes(x = AIS_before))
g3.1 = g3.1 + geom_bar(fill = "orange")
g3.1
```

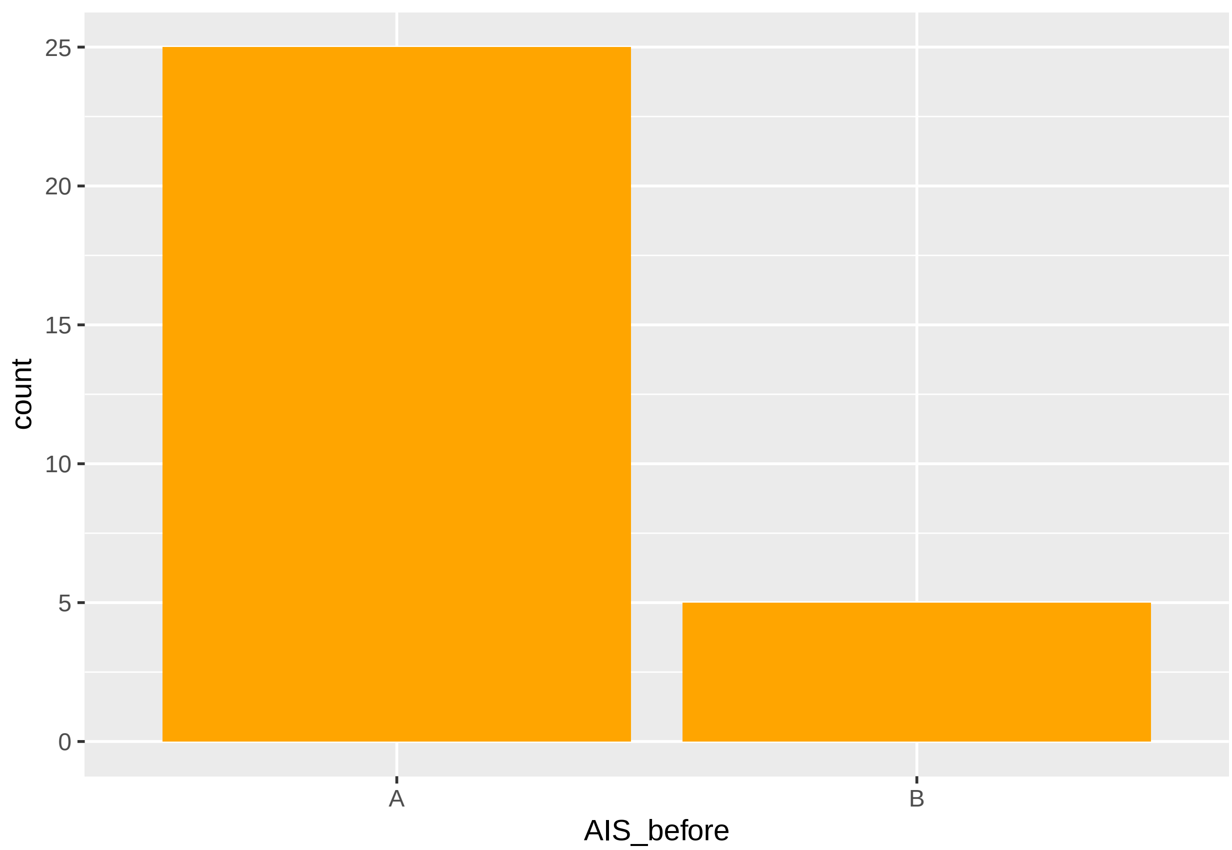


Figure 3.1: ALS level distribution before treatment.

```
g3.2 = ggplot(data = data,
             mapping = aes(x = AIS_after))
g3.2 = g3.2 + geom_bar(fill = "purple")
g3.2
```

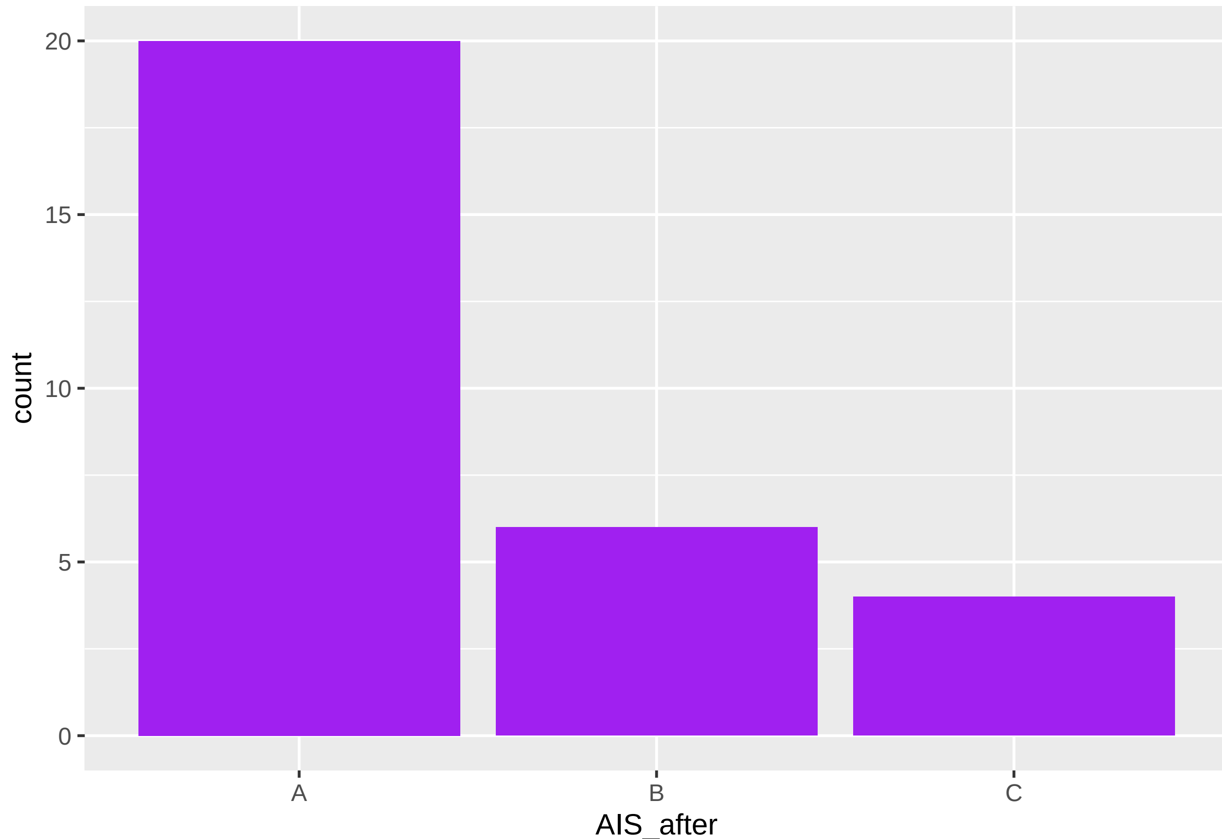


Figure 3.1: ALS level distribution after treatment.

3.3 Formulate the correct statistical hypothesis to compare the groups, choose the appropriate statistical test

- We first form the null (H0) and alternative (HA) hypothesis for this question.
- H0: The AIS score after treatment is no better than that before treatment.
- HA: The AIS score after treatment is better than that before treatment.
- Because the sample size is too small, we cannot decide whether it is normally distributed.
- Therefore, We use paired Wilcoxon test.
- We convert AIS score A,B,C,D,E into 0, 25, 50, 75 and 100.

```
a = c()
for (i in 1:nrow(data)) {
  x = data[i, "AIS_before"]
  if (x == "A")
    t = 0
```

```r
  if (x == "B")
    t = 25
  if (x == "C")
    t = 50
  if (x == "D")
    t = 75
  if (x == "E")
    t = 100
  a = c(a, t)
}
#print(a)
b = c()
for (i in 1:nrow(data)) {
  x = data[i, "AIS_after"]
  if (x == "A")
    t = 0
  if (x == "B")
    t = 25
  if (x == "C")
    t = 50
  if (x == "D")
    t = 75
  if (x == "E")
    t = 100
  b = c(b, t)
}
wilcox.test(a, b, alternative = 'less', paired = T)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  a and b
## V = 4, p-value = 0.01229
## alternative hypothesis: true location shift is less than 0
```

- p-value < 0.05
- We reject H0.
- There is sufficient evidence to conclude that the AIS score after treatment is better than that before treatment.

3.4 Discuss the results you got.

- The treatment has significant improvement effect on the spinal cord injury.
- The sample size is too small.
- The effect size.

```r
mean(b-a)/sd(b-a)
```

```
## [1] 0.4606464
```

- The effect size is so small.

- We should select more samples, and randomly sampled.
- We should consider more factors that influence the AIS score.
- We should consider to use a quantitative score to evaluate the health condition instead of AIS score.