2075

2075

2024-01-10

1. Benefits of swimming for long-distance runners

a. What would be a suitable statistical test for these data and why? Note you may need to tidy these data before deciding on which test to use.

Import and tidy the data, choose the suitable statistical test.

```
#import the data
swimming <- read.csv("swimming.csv")
head(swimming)
```

```
##   names.before_minutes.before_seconds.after_minutes.after_seconds
## 1                                        Hannah\t121\t51\t120\t31
## 2                                       Crystal\t121\t28\t119\t53
## 3                                           Ian\t126\t52\t128\t39
## 4                                       Mutee'a\t125\t46\t126\t53
## 5                                    Karthikeyan\t131\t29\t136\t10
## 6                                       Mercedes\t118\t30\t115\t4
```

```
#separate the data
new_swimming <- separate(data = swimming,
        col = "names.before_minutes.before_seconds.after_minutes.after_seconds",
        into = c("names", "before_m", "before_s", "after_m", "after_s"), sep = "\t" )
head(new_swimming)
```

```
##         names before_m before_s after_m after_s
## 1      Hannah      121       51     120      31
## 2     Crystal      121       28     119      53
## 3         Ian      126       52     128      39
## 4     Mutee'a      125       46     126      53
## 5 Karthikeyan      131       29     136      10
## 6    Mercedes      118       30     115       4
```

```
#check the variable type
summary(new_swimming)
```

```
##     names              before_m           before_s           after_m
##  Length:50          Length:50          Length:50          Length:50
##  Class :character   Class :character   Class :character   Class :character
```

```
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##    after_s
##  Length:50
##  Class :character
##  Mode  :character
```

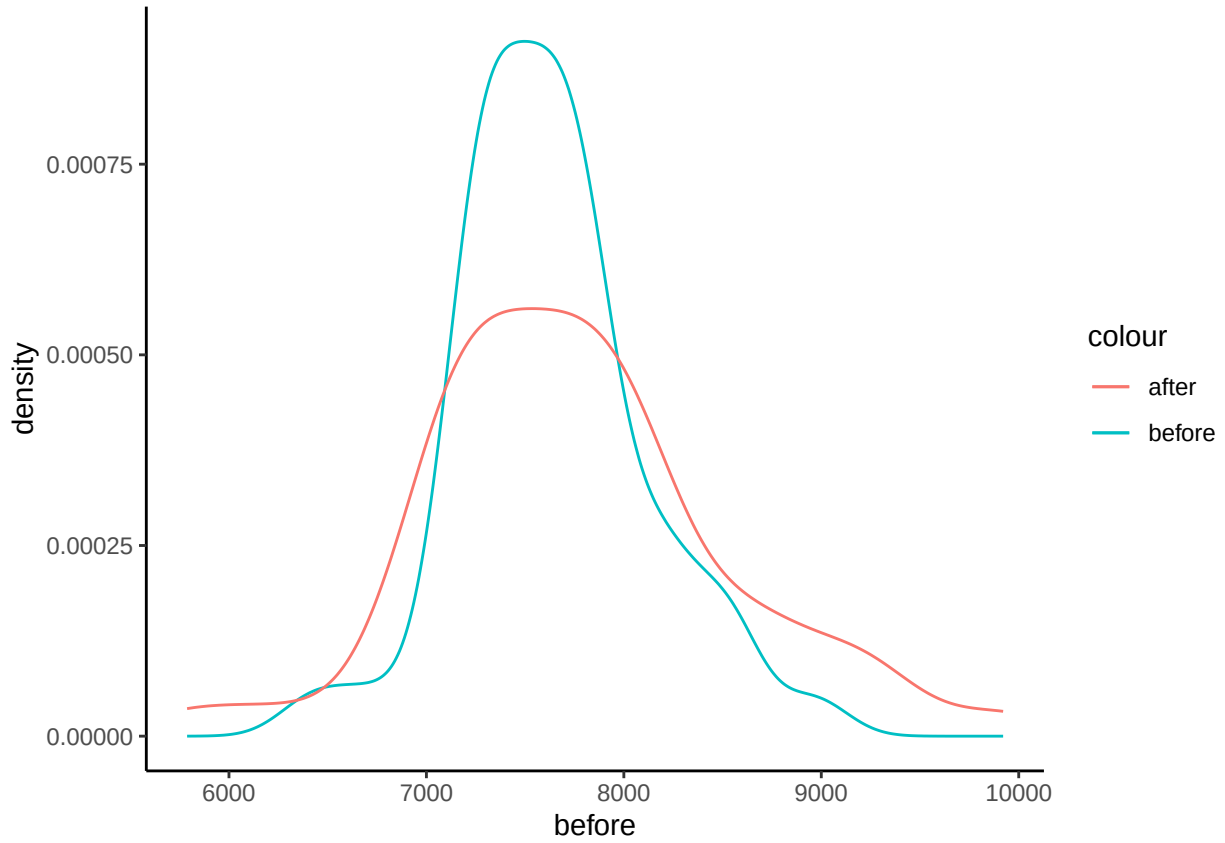
```
#change the type of variable
add_swim <- new_swimming %>%
  mutate(before_m = as.numeric(before_m), before_s = as.numeric(before_s),
         after_m = as.numeric(after_m), after_s = as.numeric(after_s))
#add the minutes and seconds
add_swim$before = add_swim$before_m*60 + add_swim$before_s
add_swim$after = add_swim$after_m*60 + add_swim$after_s
head(add_swim)
```

```
##         names before_m before_s after_m after_s before after
## 1      Hannah      121       51     120      31   7311  7231
## 2     Crystal      121       28     119      53   7288  7193
## 3         Ian      126       52     128      39   7612  7719
## 4     Mutee'a      125       46     126      53   7546  7613
## 5 Karthikeyan      131       29     136      10   7889  8170
## 6    Mercedes      118       30     115       4   7110  6904
```

Before choose the suitable test, check if normality

judge by eyes

```
ggplot(data = add_swim) +
  geom_line(aes(x=before, col='before'), stat='density') +
  geom_line(aes(x=after,col='after'), stat='density') + theme_classic()
```



The data seems approximately normally distributed ### Use Shapiro test

```
shapiro.test(add_swim$before)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  add_swim$before
## W = 0.96728, p-value = 0.179
```

```
shapiro.test(add_swim$after)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  add_swim$after
## W = 0.96733, p-value = 0.1799
```

P-value is lager than 0.05. Therefore, we could not claim that the data distribution is NOT normal with 5% significance. So, we choose t-test.

b. What are your null and alternative hypotheses?

HA: Half-marathon time is improved after swimming than before swimming H0: Half-marathon time is not improved after swimming than before swimming

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

Therefore, we should perform paired t-test:

```
# Are the variance same or different?
var.test(add_swim$before, add_swim$after)#p-value<0.05, not equal
```

```
##
##  F test to compare two variances
##
## data:  add_swim$before and add_swim$after
## F = 0.37871, num df = 49, denom df = 49, p-value = 0.0009053
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2149065 0.6673509
## sample estimates:
## ratio of variances
##          0.3787057
```

```
# t-test
t.test(add_swim$before, add_swim$after,
paired = T,
var.equal=F)
```

```
##
##  Paired t-test
##
## data:  add_swim$before and add_swim$after
## t = -2.9199, df = 49, p-value = 0.005278
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -205.72727  -37.99273
## sample estimates:
## mean difference
##         -121.86
```

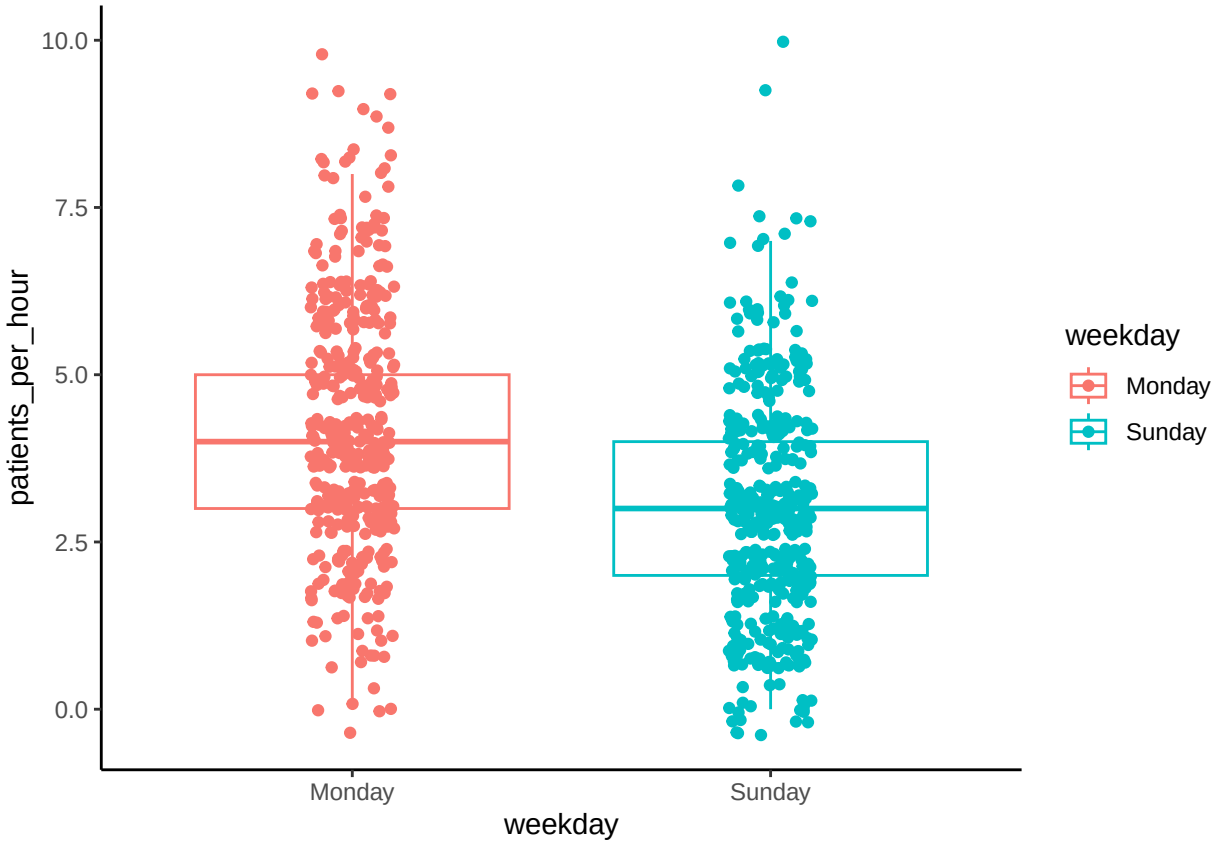
2. Number of emergency room admissions

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

```
#import the data
hospital <- read.csv("hospital_admissions.csv")
head(hospital)
```

```
##   week weekday hour patients_per_hour
## 1    1  Monday    1                 2
## 2    1  Monday    2                 4
## 3    1  Monday    3                 7
## 4    1  Monday    4                 3
## 5    1  Monday    5                 3
## 6    1  Monday    6                 2
```

```
#plot the data
g1 <- ggplot(data = hospital,
  aes(x = weekday, y = patients_per_hour, col = weekday)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_classic()
print(g1)
```

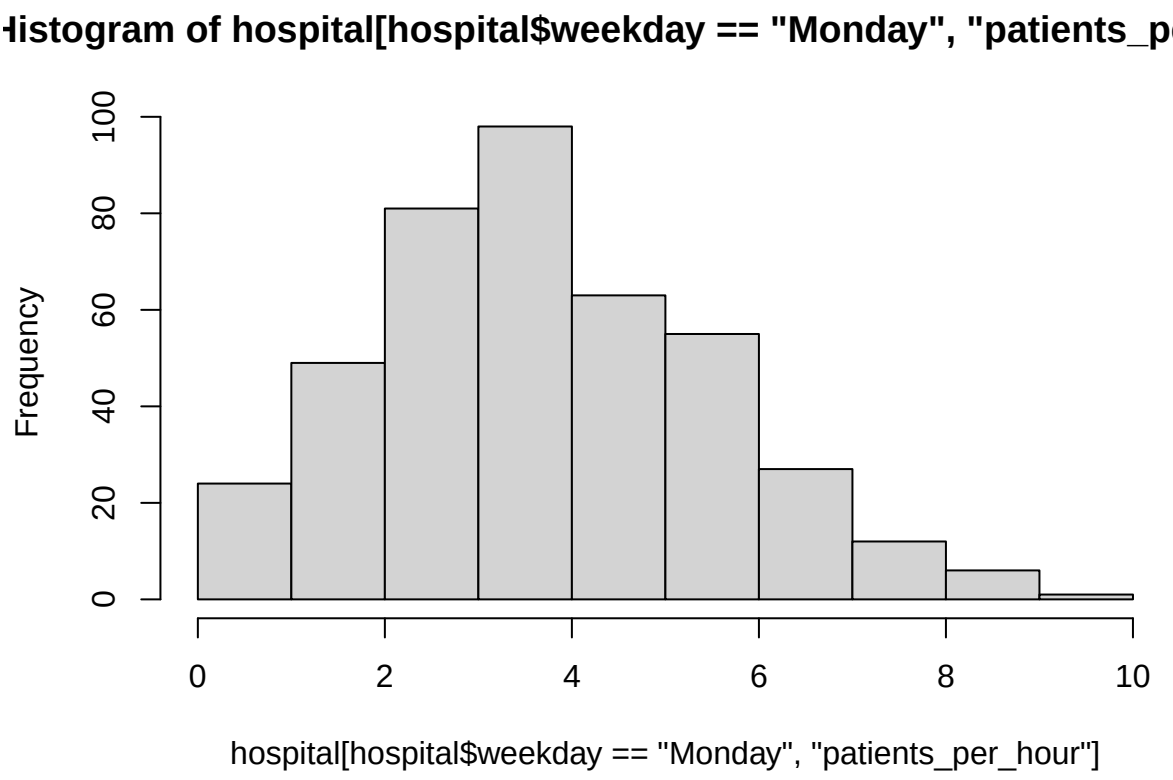


b. Is there a difference in patient admission rates between Mondays and Sundays?

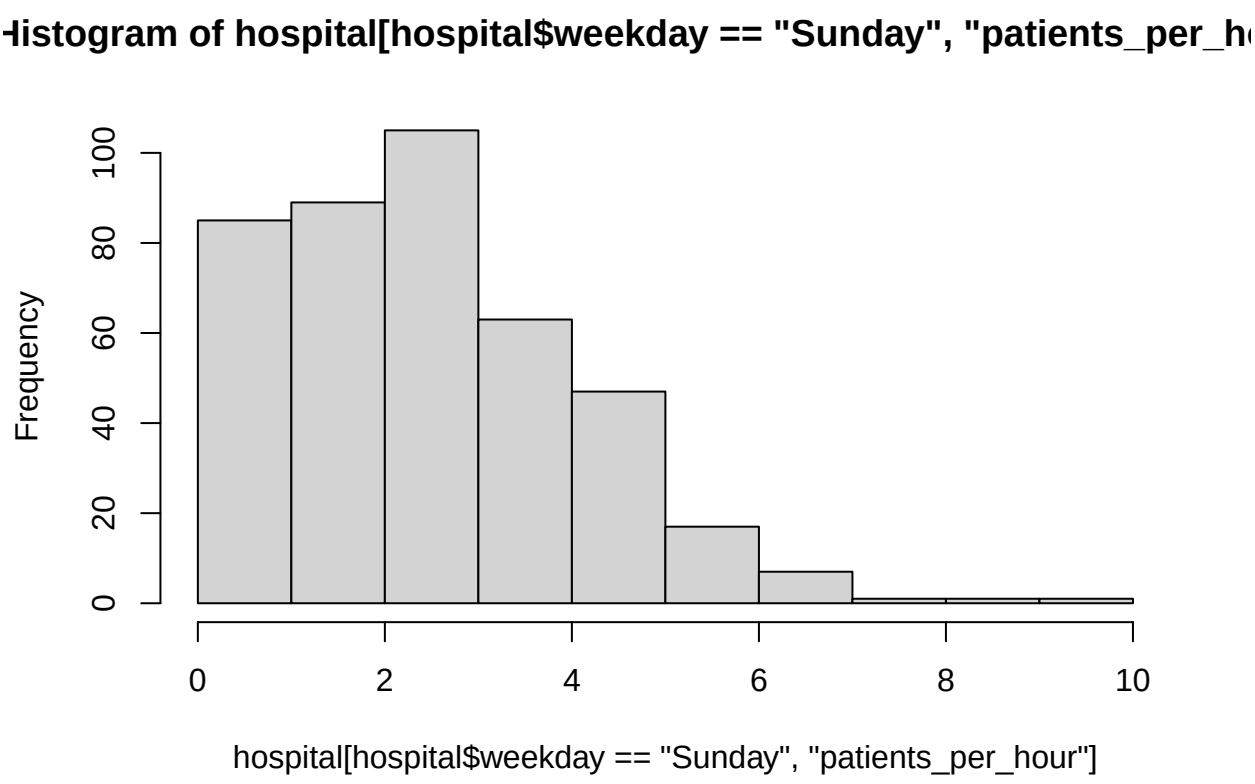
H0: Patient admission rates in monday is smaller than Sunday. HA: Patient admission rates in monday is larger than Sunday.

First, check if normality:

```
#judge by eyes
hist(hospital[hospital$weekday == "Monday","patients_per_hour"])
```




```
hist(hospital[hospital$weekday == "Sunday","patients_per_hour"])
```




```
#seems not normal

# Use Shapiro test
shapiro.test(hospital[hospital$weekday == "Monday","patients_per_hour"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  hospital[hospital$weekday == "Monday", "patients_per_hour"]
## W = 0.96785, p-value = 6.407e-08
```

```
shapiro.test(hospital[hospital$weekday == "Sunday","patients_per_hour"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  hospital[hospital$weekday == "Sunday", "patients_per_hour"]
## W = 0.94807, p-value = 6.617e-11
```

```
#p-value < 0.05, not normal
```

Therefore, we use non-parametric test: wilcox.test

```r
wilcox.test(hospital[hospital$weekday == "Monday","patients_per_hour"],
            hospital[hospital$weekday == "Sunday","patients_per_hour"],
            alternative="greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  hospital[hospital$weekday == "Monday", "patients_per_hour"] and hospital[hospital$weekday ==
## W = 119936, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

Conclusion: p-value < 0.05, patient admission rates in monday is larger than Sunday.

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

Arrange more staff on weekdays

3. Spinal cord injury and novel biomaterials

a. Import, arrange the data (merge both pieces of data and make the data possible to analyse), and make it suitable for analysis, e.g. the values. You should perform all the manipulations in R and provide the code.

Import the data:

```
before <- read.csv("SCI_before.csv")
after <- read.csv("SCI_after.csv")
head(before)
```

```
##   patient_ID AIS_before
## 1          6          A
## 2          1          A
## 3         10          B
## 4         12          A
## 5         21          A
## 6         18          A
```

```
head(after)
```

```
##   patient_ID AIS_after
## 1         13         A
## 2         10         C
## 3         24         A
## 4          4         A
## 5         11         B
## 6         16         A
```

Merge the data:

```
#Sort the data in descending order
new_before <- arrange(before, desc(patient_ID))
new_after <- arrange(after, desc(patient_ID))
#Merge the data
merge <- merge(new_before, new_after)
head(merge)
```

```
##   patient_ID AIS_before AIS_after
## 1          1          A         B
## 2          2          A         B
## 3          2          A         B
## 4          2          A         B
## 5          2          A         B
## 6          3          A         A
```

```
#remove the duplicated rows
no.merge = merge[!duplicated(merge),]
head(no.merge)
```

```
##    patient_ID AIS_before AIS_after
## 1           1          A         B
## 2           2          A         B
## 6           3          A         A
## 7           4          A         A
## 11          5          B         B
## 15          6          A         A
```

make it suitable for analysis:

```
no.merge$AIS_before[no.merge$AIS_before == "A"] <- 5
no.merge$AIS_before[no.merge$AIS_before == "B"] <- 4
no.merge$AIS_before[no.merge$AIS_before == "C"] <- 3

no.merge$AIS_after[no.merge$AIS_after == "A"] <- 5
no.merge$AIS_after[no.merge$AIS_after == "B"] <- 4
no.merge$AIS_after[no.merge$AIS_after == "C"] <- 3

merge11 <- no.merge %>%
  mutate(AIS_before = as.numeric(AIS_before),
         AIS_after = as.numeric(AIS_after))
head(no.merge)
```

```
##    patient_ID AIS_before AIS_after
## 1           1          5         4
## 2           2          5         4
## 6           3          5         5
## 7           4          5         5
## 11          5          4         4
## 15          6          5         5
```

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

```
ggplot
```

```
## function (data = NULL, mapping = aes(), ..., environment = parent.frame())
## {
##     UseMethod("ggplot")
## }
## <bytecode: 0x144d3d420>
## <environment: namespace:ggplot2>
```

Formulate the correct statistical hypothesis to compare the groups, choose the appropriate statistical test, and check assumptions for this test. Explain your choice briefly. Then, perform this test and identify whether the difference between the experimental groups is statistically significant.

HA: There is differences after a novel biomaterial, it may help to regenerate injured nervous tissue after SC
H0: There is no differences after a novel biomaterial, it may not help to regenerate injured nervous tissue after SC