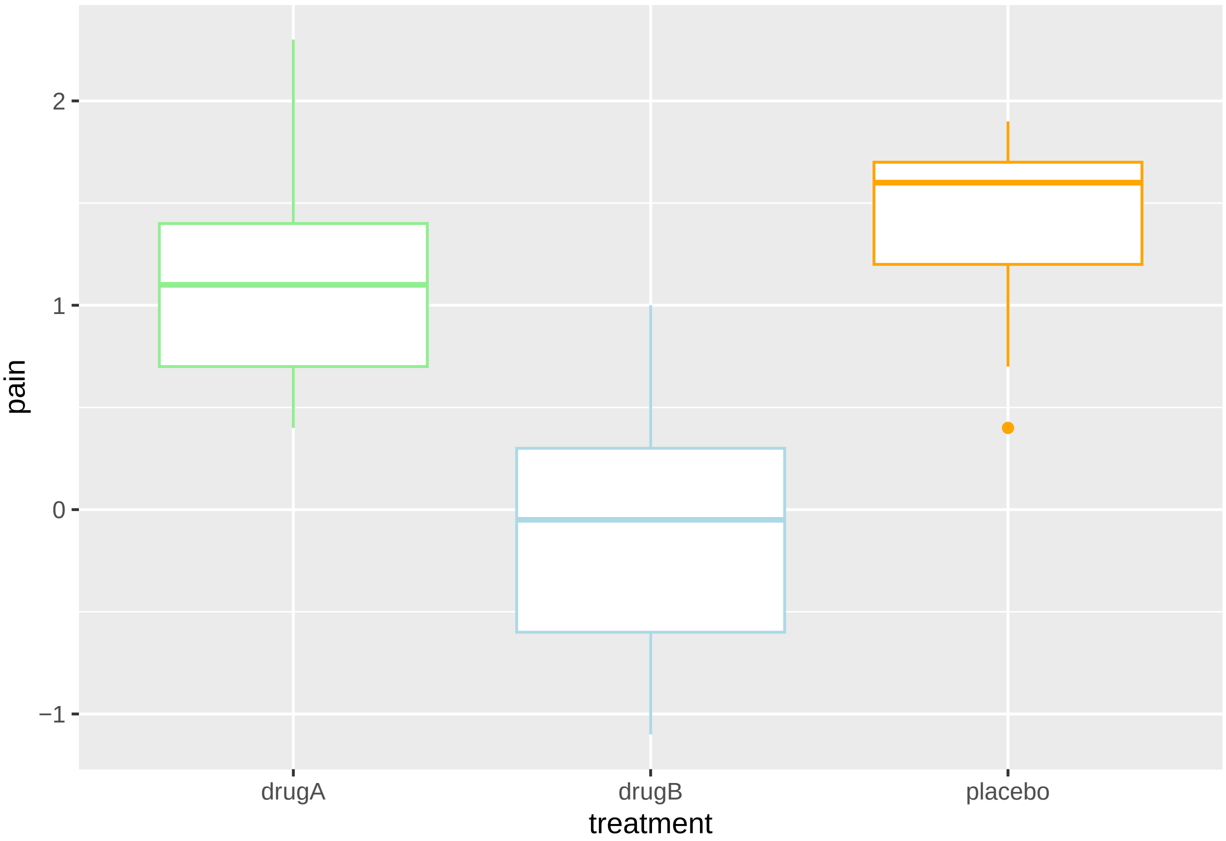
Practical 2.1

CHEN

2024-01-16

Looking at the dataset

```
drug <- read.csv("drug_trial.csv")
library(ggplot2)
library(tidyr)
# Plot the data in a boxplot
g1 <- ggplot(data = drug, aes(x = treatment, y = pain))
g1_1 <- g1 +geom_boxplot(color = c("lightgreen","lightblue","orange"))
g1_1
```



Reminding of H0 and HA

H0: The groups are the same in their relative pain levels. HA: There is differences between these groups in their relative pain levels. Or: H0: Any random two data points of relative pain levels from different groups will not be different from that from a same group. HA: Two random data points of relative pain levels from different groups will be different from that from the same group.

Getting a sense of variability

Draw two data points from the "trail" dataset.

```r
# randomly two data points
sample_index <- sample(1:nrow(drug),2)
random <- drug[sample_index,]
random
```

```
##    treatment pain
## 4    placebo  1.1
## 36     drugB  0.3
```

For this sample:

```r
# Are they in the same group or in different groups?
random$treatment[1] == random$treatment[2]
```

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

```r
# absolute difference
random$pain[1]-random$pain[2]
```

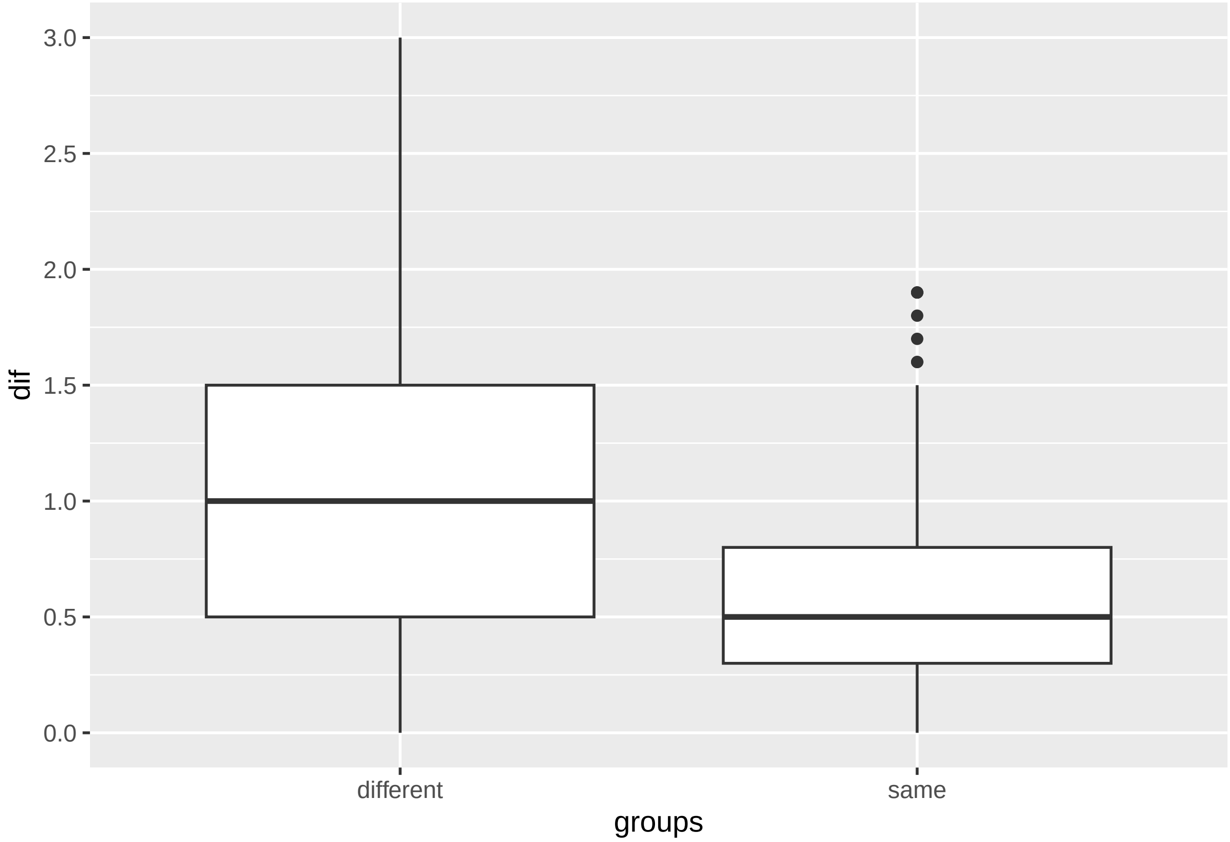
```
## [1] 0.8
```

Do this a few times to see a few examples of differneces within groups and between groups.

```r
# create empty dataframe for data filling
differences <- data.frame(matrix(nrow = 1000, ncol = 2))

for (i in 1:1000) {
# choose 2 (different) rows from the total number of rows
  sample_index <- sample(1:nrow(drug),2)
# read out those two rows from trial. This is your sample
# (maybe save it as a separate object, but it's not necessary)
  sample <- drug[sample_index,]
# For the two points in your sample, read out the pain indices
# and determine their absolute difference
  difference <- abs(sample$pain[1]-sample$pain[2])
# For the two points in your sample, decide whether they belong
# to the same or to different treatment groups
  if (sample$treatment[1]==sample$treatment[2]) {
    differences[i,] <- c('same',difference)
  } else {
    differences[i,] <- c('different',difference)
  }
}

# set column names
colnames(differences) <- c('groups','dif')

# change data types
differences$dif <- as.numeric(differences$dif)
```

Plot the result of differences

```
g2 <- ggplot(differences, mapping = aes(x = groups, y = dif, group = groups))
g2_1 <- g2 + geom_boxplot() +
  scale_y_continuous(breaks=seq(0,3.5,0.5))
g2_1
```



Do they look the same, do they look different? If we compare the means, we get

```
same <- c(differences[differences$groups=="same",]$dif)
sm <- mean(same)
sm
```

```
## [1] 0.5770186
```

for the distances between data points within the same groups and

```r
different <- c(differences[differences$groups=="different",]$dif)
dm <- mean(different)
dm
```

```
## [1] 1.071534
```

for the distances between data points in different groups. (The exact values may be a bit different for you, since we obtained them using random sampling!) The difference between those means is

```r
abs(dm-sm)
```

```
## [1] 0.4945153
```

This is a significant difference.

Run a test to find out!

We now have only two groups that need comparing: absolute differences between participants belonging to the same group (placebo, drug A or drug B), and absolute differences between participants belonging to different groups.

How do we compare two groups? A t-test is not a good idea here (why not?), but you could run a non-parametric alternative, such as the Wilcoxon rank sum test.What p-value do you get, and how do you interpret this?

```r
wilcox.test(dif~groups, data = differences, alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dif by groups
## W = 154606, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

This agrees with an alternative hypothesis that there are differences between different groups.

Also, what information do you think is still missing?