Activities - Solution

Activity 3

a) Hypothesis tests for two populations:

- a.1) Choose the best way to import file "feeders_select_exercise.csv" and save it to an object named feeders.

This file contains measurements (OD_avg and pH) taken from a sucrose solution present in 10 different bird feeders. Measurements were made at day 1 and at day 7 after changing the solution. The purpose will be to see differences between both days

```
feeders <- read.csv("data/feeders_select_exercise.csv")
```

- a.2) Check the str() to confirm that, al least, both numeric variables (OD_avg and pH) are in fact numeric and not factors. DurGrowth should be a factor.

```
# change DurGROWTH to a factor
feeders$DurGROWTH <- factor(feeders$DurGROWTH)

levels(feeders$DurGROWTH)
```

```
## [1] "1" "7"
```

- a.3) Calculate the mean, sd and median of variables pH and OD_avg for day 1 and day 7.

```
by(feeders$OD_avg, feeders$DurGROWTH, mean, na.rm=TRUE)
```

```
## feeders$DurGROWTH: 1
## [1] 0.038025
## --------------------------------------------------------
## feeders$DurGROWTH: 7
## [1] 0.042325
```

```
by(feeders$OD_avg, feeders$DurGROWTH, median, na.rm=TRUE)
```

```
## feeders$DurGROWTH: 1
## [1] 0.03865
## --------------------------------------------------------
## feeders$DurGROWTH: 7
## [1] 0.042
```

```
by(feeders$OD_avg, feeders$DurGROWTH, sd, na.rm=TRUE)
```

```
## feeders$DurGROWTH: 1
## [1] 0.002459251
## --------------------------------------------------------
## feeders$DurGROWTH: 7
## [1] 0.002273427
```

```
by(feeders$pH, feeders$DurGROWTH, mean, na.rm=TRUE)
```

```
## feeders$DurGROWTH: 1
## [1] 4.4401
## --------------------------------------------------------
## feeders$DurGROWTH: 7
## [1] 3.6765
```

```r
by(feeders$pH, feeders$DurGROWTH, median, na.rm=TRUE)
```

```
## feeders$DurGROWTH: 1
## [1] 4.2535
## --------------------------------------------------------
## feeders$DurGROWTH: 7
## [1] 3.671
```

```r
by(feeders$pH, feeders$DurGROWTH, sd, na.rm=TRUE)
```

```
## feeders$DurGROWTH: 1
## [1] 0.4024783
## --------------------------------------------------------
## feeders$DurGROWTH: 7
## [1] 0.3874203
```

- a.4) How many measurement (n) were taken in day 1 and in day 7?

```r
summary(feeders)
```

```
##        X              TRIAL             DATE      TIME      SITE
##  Min.   :  1.00   Min.   :1.00   03/jul/17:4   06:00:4   Glide:20
##  1st Qu.:  5.75   1st Qu.:1.75   11/jul/17:3   06:15:7
##  Median : 10.50   Median :2.00   13/jun/17:3   06:40:2
##  Mean   : 22.25   Mean   :2.10   17/jul/17:4   07:10:4
##  3rd Qu.: 15.25   3rd Qu.:3.00   19/jun/17:2   07:20:3
##  Max.   :110.00   Max.   :3.00   27/jun/17:4
##  DurGROWTH     FEED.N     TREATMENT OBSERVER    FEED.WT
##  1:10      Min.   : 1.0   cage:20   CGL :5   Min.   :1305
##  7:10      1st Qu.: 3.0             TH  :8   1st Qu.:1460
##            Median : 5.5             NA's:7   Median :1534
##            Mean   : 5.5                      Mean   :1517
##            3rd Qu.: 8.0                      3rd Qu.:1605
##            Max.   :10.0                      Max.   :1633
##    FEED.TEMP           pH             OD_1              OD_2
##  Min.   :52.70   Min.   :3.011   Min.   :0.03360   Min.   :0.03410
##  1st Qu.:57.50   1st Qu.:3.712   1st Qu.:0.03875   1st Qu.:0.03760
##  Median :60.35   Median :4.251   Median :0.04035   Median :0.03960
##  Mean   :60.62   Mean   :4.058   Mean   :0.04038   Mean   :0.03997
##  3rd Qu.:64.10   3rd Qu.:4.253   3rd Qu.:0.04202   3rd Qu.:0.04280
##  Max.   :71.10   Max.   :5.530   Max.   :0.04600   Max.   :0.04660
##      OD_avg
##  Min.   :0.03385
##  1st Qu.:0.03848
##  Median :0.03997
##  Mean   :0.04018
##  3rd Qu.:0.04229
##  Max.   :0.04620
```

```r
# 10 measurements for each day
```

- a.5) Is pH data for day 7 and day 1 normally distributed? What about OD_avg? Which type of tests would you use for each variable.

```r
# for OD_avg
shapiro.test(feeders[feeders$DurGROWTH == "7", "OD_avg"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  feeders[feeders$DurGROWTH == "7", "OD_avg"]
## W = 0.95379, p-value = 0.7134
```

```
shapiro.test(feeders[feeders$DurGROWTH == "1", "OD_avg"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  feeders[feeders$DurGROWTH == "1", "OD_avg"]
## W = 0.91631, p-value = 0.3272
```

Don't reject H0, OD_avg data is normally distributed follow parametric hypothesis tests

```
# for pH
shapiro.test(feeders[feeders$DurGROWTH == "1", "pH"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  feeders[feeders$DurGROWTH == "1", "pH"]
## W = 0.55361, p-value = 1.556e-05
```

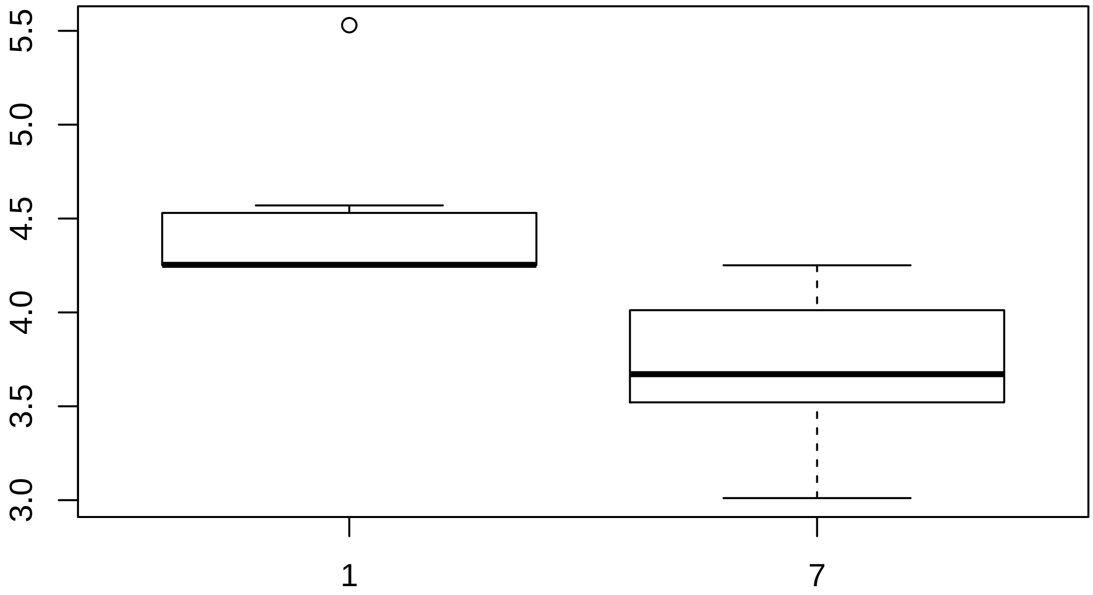
```
shapiro.test(feeders[feeders$DurGROWTH == "7", "pH"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  feeders[feeders$DurGROWTH == "7", "pH"]
## W = 0.97058, p-value = 0.8962
```

```
# reject HO for day1, pH data is in this day is not normally distributed
# Unless you have any evidence for the fact that this type of data is,
# in fact, normally distributed (other samplings from other trials, problem during the
# measurements, ...), go for non-parametric tests, even if normality is assumed for day7
```

- a.6) Draw a boxplot for pH as function of DurGrowth. Is there something you could to the data to solve the normality issue?

```
boxplot(formula = pH ~ DurGROWTH,data = feeders)
```



```
# maybe remove that one outlier and repeat shapiro-test, but population "day1"
# doesn't look very simetrical, so sampling could be repeated, eventually
# increasing sample n
```

- a.7) measurements in day 1 and day 7 were made always on the same feeder. What kind of strategy should be used for to compare these two days?

```
# This can be viewed as a "before and after" situation, so it's paired data
```

- a.8) Perform the Bartlet test and the t.test/wilcoxon (depending on the normality tests results) to investigate if pH and OD_avg are different between the two days.

- *hint: look at column FEED.N (unique ID of the feeder) to understand the order of the measurements in the table*

```
# variance Homogeneity

bartlett.test(pH ~ DurGROWTH, data = feeders)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  pH by DurGROWTH
## Bartlett's K-squared = 0.012394, df = 1, p-value = 0.9114
```

```
#variance not homogeneous
```

```
bartlett.test(OD_avg ~ DurGROWTH, data = feeders)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  OD_avg by DurGROWTH
## Bartlett's K-squared = 0.052579, df = 1, p-value = 0.8186
```

```
#variance is homogeneous (for a sign. level of 0.05)
```

```
## Hypothesis tests for OD_avg
```

```r
# the order of the values for day 1 and 7 allows matching of the pairs,
# therefore we may use the formulaformula = OD_avg ~ DurGROWTH directly

t.test(formula = OD_avg ~ DurGROWTH, data = feeders, paired = TRUE, var.equal = TRUE)
```

```
##
##  Paired t-test
##
## data:  OD_avg by DurGROWTH
## t = -2.9611, df = 9, p-value = 0.01593
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.007585021 -0.001014979
## sample estimates:
## mean of the differences
##                 -0.0043
```

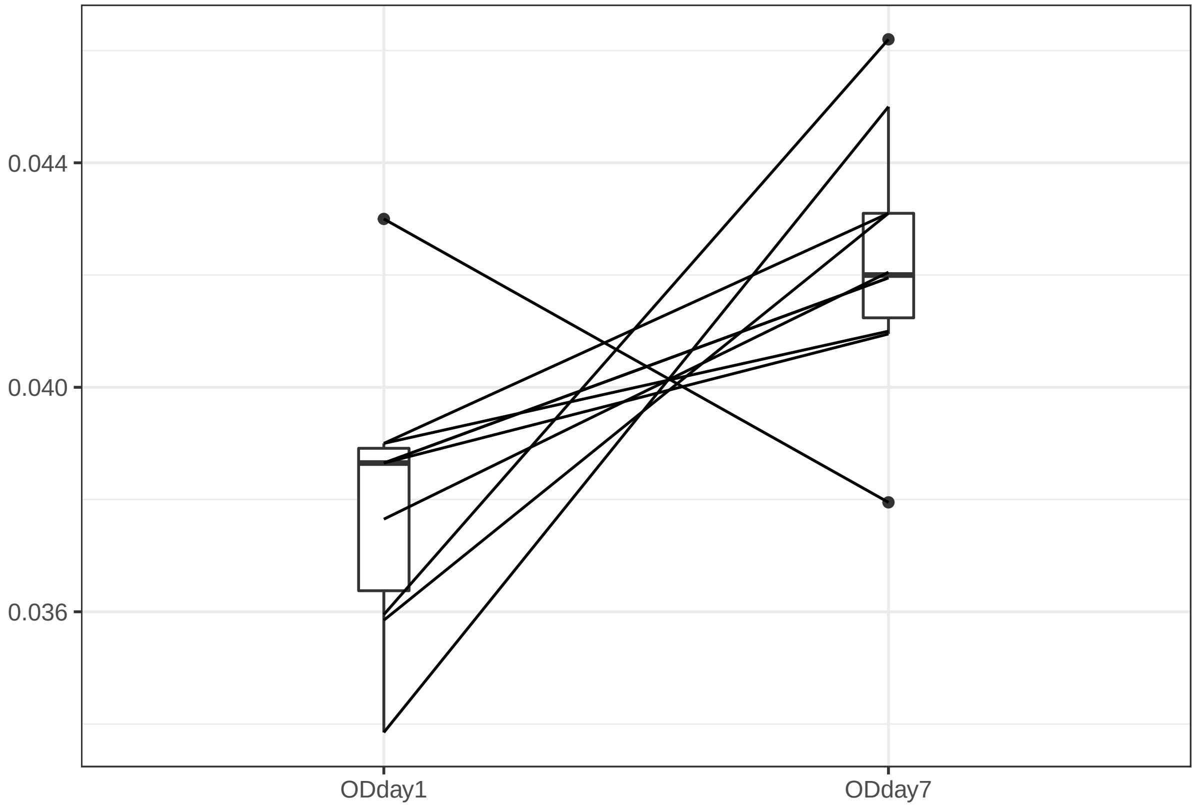
```r
#or get values in invidivual vectors (confirm the order is correct)

ODday1 <- feeders[feeders$DurGROWTH == 1,"OD_avg"]
ODday7 <- feeders[feeders$DurGROWTH == 7,"OD_avg"]

t.test(ODday1, ODday7, paired = TRUE, var.equal = TRUE)
```

```
##
##  Paired t-test
##
## data:  ODday1 and ODday7
## t = -2.9611, df = 9, p-value = 0.01593
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.007585021 -0.001014979
## sample estimates:
## mean of the differences
##                 -0.0043
```

```r
#plot for pairs
library(PairedData)
```

```
## Loading required package: MASS

## Loading required package: gld

## Warning: package 'gld' was built under R version 3.5.2

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 3.5.2

## Loading required package: lattice

## Loading required package: ggplot2

##
## Attaching package: 'PairedData'

## The following object is masked from 'package:base':
##
##     summary
```

```
pd = paired(ODday1, ODday7)
plot(pd, type = "profile") + theme_bw()
```



```
## Hypothesis tests for pH

wilcox.test(formula = pH ~ DurGROWTH, data = feeders, paired = TRUE)
```

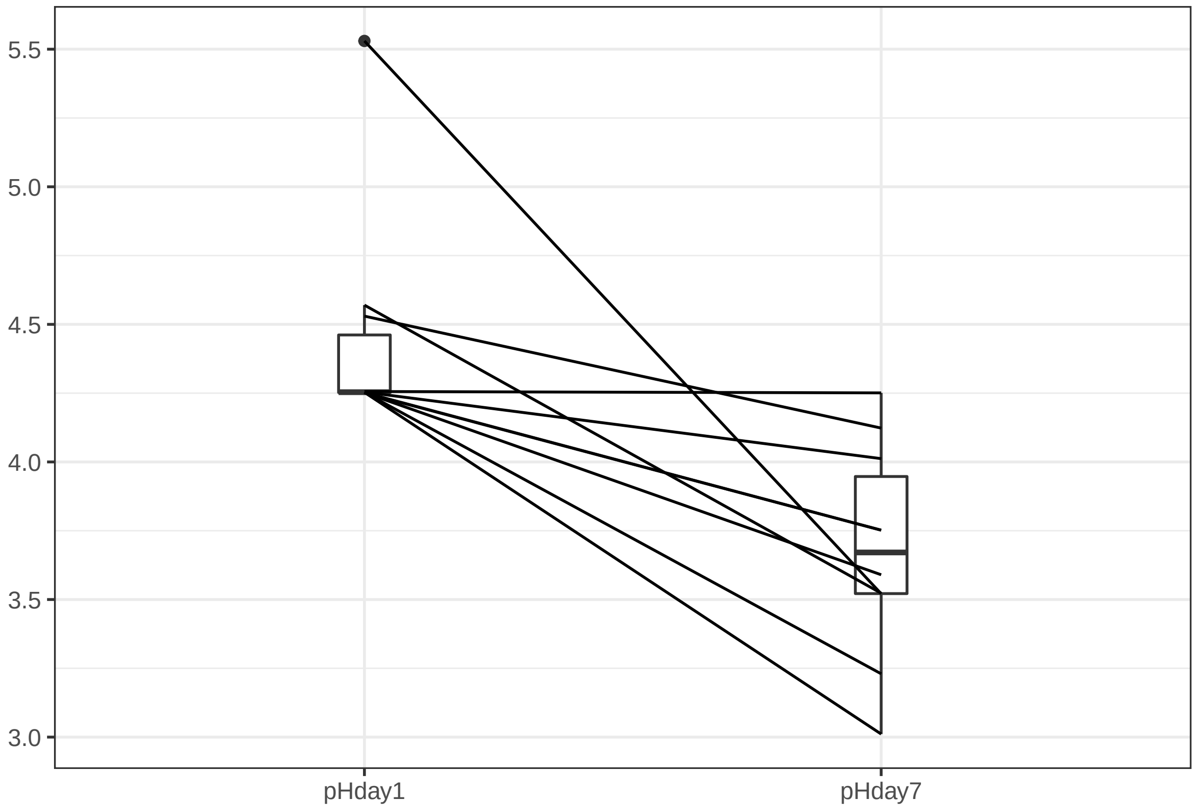
```
##
##  Wilcoxon signed rank test
##
## data:  pH by DurGROWTH
## V = 55, p-value = 0.001953
## alternative hypothesis: true location shift is not equal to 0
```

```
# or
pHday1 <- feeders[feeders$DurGROWTH == 1,"pH"]
pHday7 <- feeders[feeders$DurGROWTH == 7,"pH"]

wilcox.test(pHday1, pHday7, paired = TRUE)
```

```
##
##  Wilcoxon signed rank test
##
## data:  pHday1 and pHday7
## V = 55, p-value = 0.001953
## alternative hypothesis: true location shift is not equal to 0
```

```
# plot for pairs

pd = paired(pHday1, pHday7)
```

```
plot(pd, type = "profile") + theme_bw()
```



```
feeders$pH
```

```
##  [1] 4.251 4.254 5.530 4.252 4.530 4.256 4.252 4.253 4.570 4.253 3.752
## [12] 4.012 3.521 3.752 4.123 4.251 3.590 3.230 3.523 3.011
```

Sampling for day1 yielded very similar values (e.g. 4.25), and the distribution is not very simetrical, which can be a problem for hypothesis tests, even non-parametric

Wilcox.test (aka, sum of rank test) works by ordering all values of both samples together, and calculate significance values based on the sum of the ranks of these values, for each sample.

In case of many ties when ranking the values, (e.g. many equal values) a warning message may appear calling the attention for uncertainty in the p-value

What to do in that case… maybe repeat experiment, increase the n, …