Hypothesis tests on the mean

September 23, 2020

Answers for this lab must be submitted before **September 30th at 5pm on Moodle**. In your answer for each question, please include the R code used (if applicable) and the results obtained.

1. Ozone concentration in three gardens

For this exercise, we will use the `gardens.csv` dataset, which comes from Michael Crawley's book *Statistics: An Introduction Using R*. These data represent ozone concentrations (in parts per 100 million or pphm) measured in three gardens (A, B and C) on different days.

```
gardens <- read.csv("gardens.csv")
```

a) Observe the distribution of ozone measurements in each garden with a graph of type `geom_jitter` in *ggplot2*. What is the advantage of that type of graph for this dataset, compared with `geom_point` or `geom_boxplot`?

b) What is the mean and standard deviation of the ozone concentration in each garden? Is the mean a good indicator of the "typical" value in each garden?

c) From these data, test the null hypothesis that gardens A and B receive the same mean ozone concentration. What is your estimate of the difference between the mean and its 99% confidence interval? Does this test give a good idea of the difference between the two gardens? Explain your answer.

d) Repeat the previous exercise for the null hypothesis that gardens A and C receive the same mean ozone concentration. Comment on the difference between this result and the previous result.

2. Effect of dead leaves on the nitrogen supply of seedlings

The `nconc.csv` dataset present the results of a (fictitious) experiment to estimate the effect of dead leaves on the nitrogen supply of seedlings.

```
nconc <- read.csv("nconc.csv")
nconc
```

```
##   plot   litter no_litter
## 1    1 1.859543 1.8073724
## 2    2 1.461364 0.7367826
## 3    3 1.488136 1.6332546
## 4    4 1.325093 1.1615338
## 5    5 1.600666 0.9864743
## 6    6 2.038625 1.9011473
## 7    7 1.788214 1.3162220
## 8    8 1.994081 1.7849742
```

Eight plots were delimited in a forest and divided into two halves. In each plot, one half (chosen at random) was treated by systematically removing dead leaves from the ground. The last two columns of the data frame show the nitrogen concentration (expressed as a % of biomass) for the seedlings of the untreated (`litter`) and treated (`no_litter`) halves.

a) What is the advantage of conducting the experiment in this manner rather than completely removing the dead leaves from four of the eight plots? What type of t test is appropriate here to determine the effect of the treatment on the mean nitrogen concentration of the seedlings?

b) Perform the t test you selected and determine if the treatment has a significant effect at a significance threshold of $\alpha = 0.05$. What is the estimated mean effect (remember to interpret the sign of the difference) and its confidence interval?