Chapter 5 Hypothesis Testing

Angelo LaCommare-Soto

2025-02-28

Research Question 1: Does the presence of L. latifolium affect the abundance of soil invertebrates?

Section 1 - Importing Data

```r
# set working directory for all chunks in this file (default working directory is wherever Rmd file is)
getwd()
```

```
## [1] "C:/Users/Angelo L/Documents/GitHub/BIOL710/RCode710/RCode/working_directory"
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
# Importing data
marsh <- read.csv("marsh.csv",header=TRUE)

# viewing the structure of the data
str(marsh)
```

```
## 'data.frame':    119 obs. of  9 variables:
##  $ sample.name : chr  "2 high 80 Lep" "2 high 60 Lep" "2 mid 60 Lep" "2 mid 0 Lep" ...
##  $ n           : int  1976 377 513 1330 345 1130 133 232 264 699 ...
##  $ s           : int  28 28 34 42 35 36 30 42 38 41 ...
##  $ H           : num  1.07 2 2.22 1.76 2.29 ...
##  $ season      : chr  "Fall" "Fall" "Fall" "Fall" ...
##  $ elevation   : chr  "high" "high" "mid" "mid" ...
##  $ plant       : chr  "Lep" "Lep" "Lep" "Lep" ...
##  $ stage_lep   : chr  "early_senescence" "early_senescence" "early_senescence" "early_senescence" ..
##  $ invertebrate: chr  "canopy" "canopy" "canopy" "canopy" ...
```

```r
head(marsh)
```

```
##     sample.name    n  s        H season elevation plant        stage_lep
## 1 2 high 80 Lep 1976 28 1.073908   Fall      high   Lep early_senescence
## 2 2 high 60 Lep  377 28 2.000021   Fall      high   Lep early_senescence
## 3  2 mid 60 Lep  513 34 2.215133   Fall       mid   Lep early_senescence
## 4   2 mid 0 Lep 1330 42 1.761914   Fall       mid   Lep early_senescence
## 5  2 low 40 Lep  345 35 2.286133   Fall       low   Lep early_senescence
## 6  2 low 20 Lep 1130 36 1.814826   Fall       low   Lep early_senescence
##   invertebrate
## 1       canopy
## 2       canopy
## 3       canopy
## 4       canopy
## 5       canopy
## 6       canopy
```

Question Answers

a. The 'marsh' dataset has 119 observations of 9 variables.
b. Given the question 'Does the presence of L. latifolium affect the abundance of soil invertebrates?', we are interested in the 'plant' and 'n' variables.
c. Considering the experiment and research question, the treatment is 'plant' species/type.

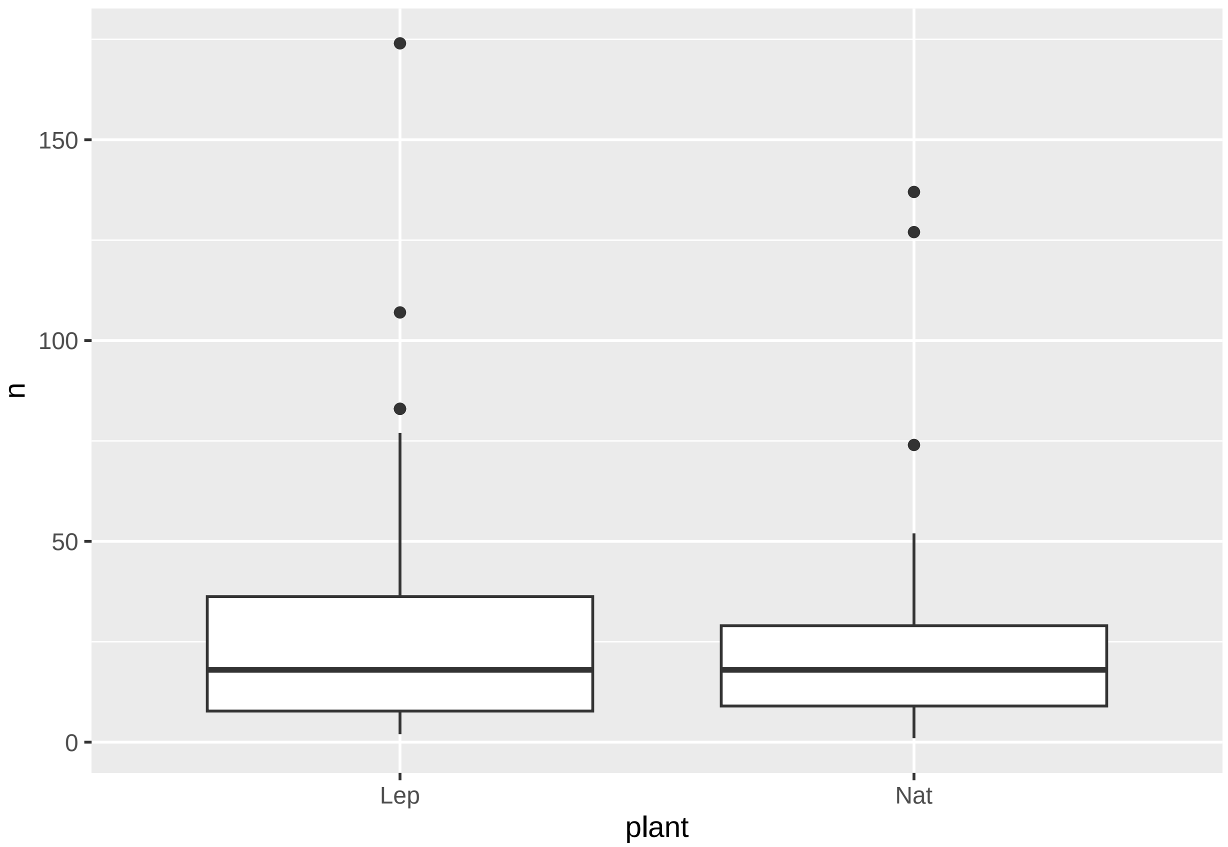
Section 2 - Stating the Null and Alternative Hypothesis

Challenge 1

H0: The presence of L. latifolium does not affect the abundance of soil invertebrates in a wetland ecosystem

HA: The presence of L. latifolium does affect the abundance of soil invertebrates in a wetland ecosystem.

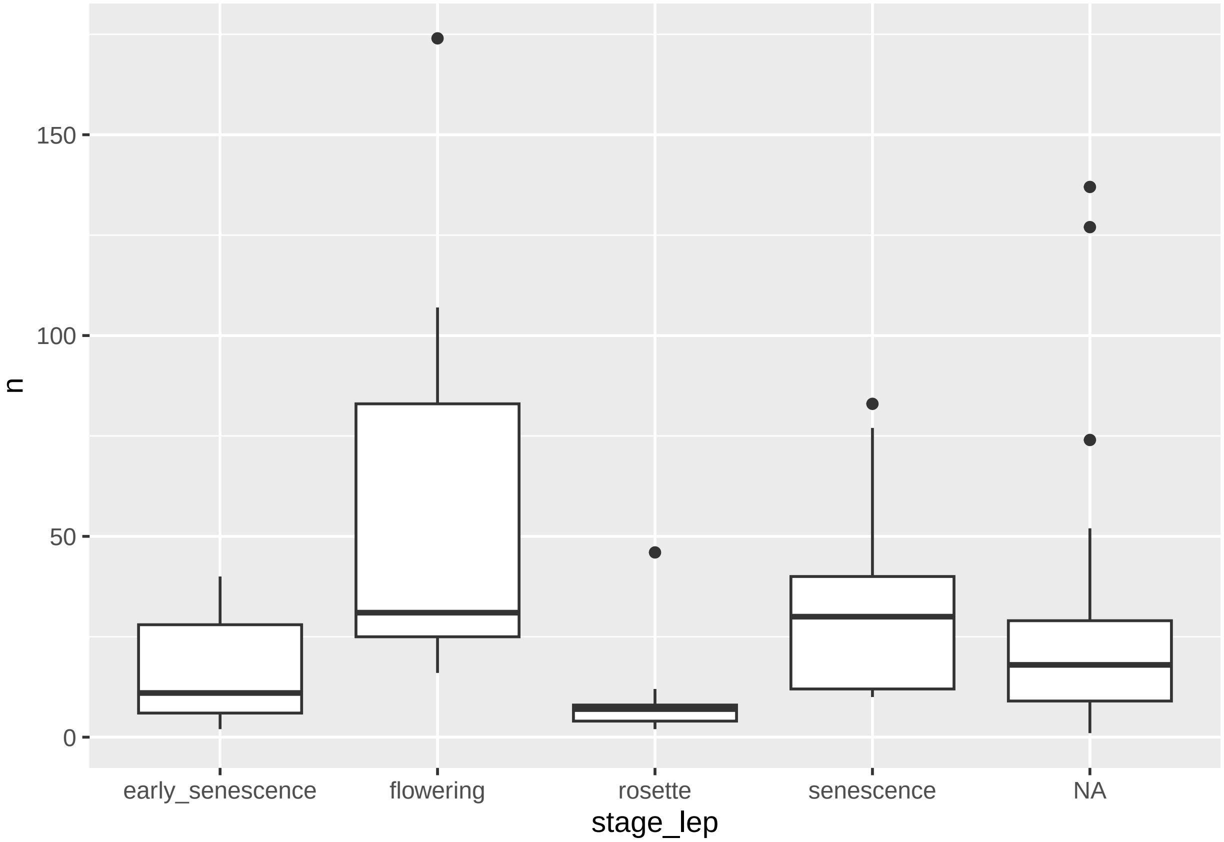
```r
# visualizing the variables of interest
# filtering by invertebrate type
soil <- filter(marsh,invertebrate=="soil")

# boxplot for abundance across treatment
p1 <- ggplot(soil,aes(x=plant,y=n)) +
  geom_boxplot()
p1
```



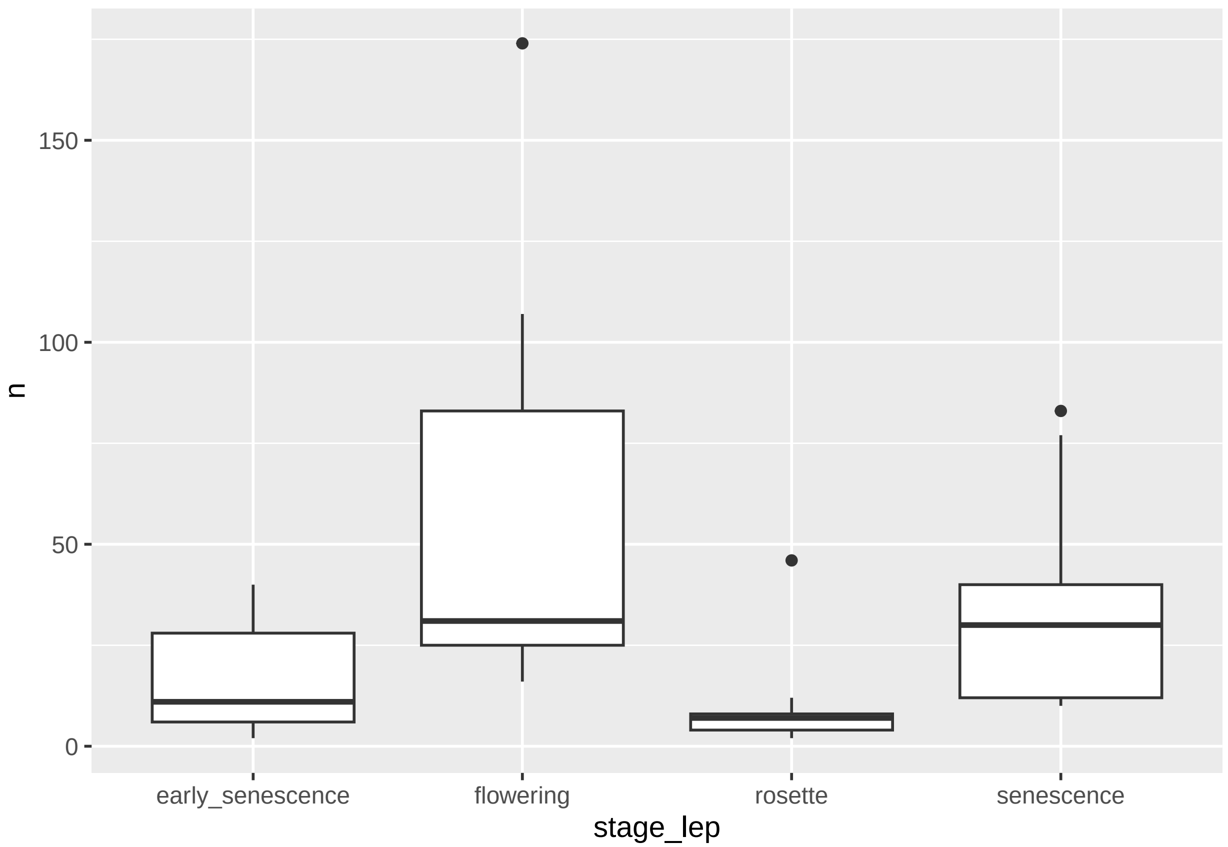
Question Answer

a. From the boxplots, which exhibit very similar distributions among treatments, it appears as though plant type/species does not have an effect on soil invertebrate abundance.

```
# Digging deeper and investigating other potential features of L. latifolium that could be contributing
# boxplot for abundance across L. latifolium stages
p2 <- ggplot(soil,aes(x=stage_lep,y=n)) +
  geom_boxplot()
p2
```



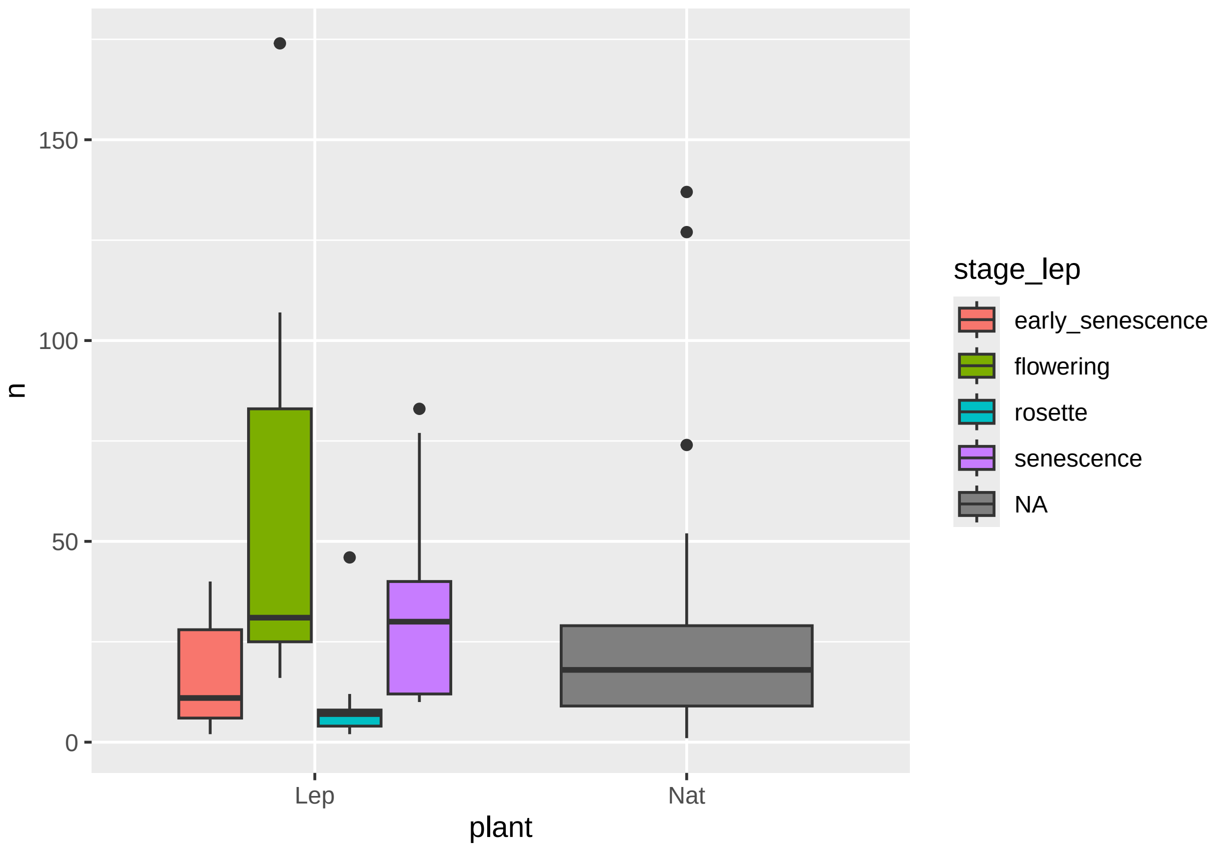
```
# removing NAs from the ggplot
p3 <- ggplot(drop_na(soil),aes(x=stage_lep,y=n)) +
  geom_boxplot()
p3
```



Question Answers

a. In contrast to the previous boxplots, these are more numerous and do not all demonstrate similar distributions or overlapping median values.
b. Biologically, this could mean that life history stage of L. latifolium dictates the abundance of soil invertebrates. Statistically, this could mean that there is an effect of life history stage of L. latifolium on soil invertebrate abundance.

```r
# plotting all variables together
p4 <- ggplot(soil,aes(x=plant,y=n,fill=stage_lep)) +
  geom_boxplot()
p4
```



Section 3 - Estimating the Test Statistic and its Precision

```
# Estimate the mean soil invertebrate abundance per plot per treatment
# filtering data by invertebrate type and treatment
s_lep <- filter(marsh,
               invertebrate=="soil" & plant=="Lep")

# mean abundance per plot
m_s_lep <- mean(s_lep$n)

m_s_lep
```

```
## [1] 30.61111
```

```
# standard error
se_s_lep <- sd(s_lep$n)/sqrt(36)

se_s_lep
```

```
## [1] 5.974719
```

Question Answer

a. The test statistic is a mean of about 31 soil invertebrates per plot with a precision of about 6 invertebrates (mean = 30.61, n=36, SEM = 5.97).

Section 4 - Estimating the Null Distribution and the p-value

```
# creating a dataframe to calculate p-value based on figure 3 in lab manual
Nat_df <- data.frame(MeanSoilInvert=c(11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 2
                     p=c(0.0001, 0.0007, 0.0008, 0.0014, 0.0056, 0.0098, 0.0162, 0.0265, 0.0375, 0.0444
#sum(Nat_df$p)

#filter out values equal to or more extreme than the test statistic of 30.6, rounded up to 31
Nat_df0 <- filter(Nat_df, MeanSoilInvert >= 31)

# because of two tail distribution with a center at 25
Nat_df1 <- filter(Nat_df, MeanSoilInvert <= 19)

#combine dataframes with rbind()
Nat_df01 <- rbind(Nat_df0, Nat_df1)

#calculate p-value from sums of probabilities
pvalue1 <- sum(Nat_df01$p)
pvalue1
```

```
## [1] 0.2938
```

Question Answers

a. The p-value is 0.294.
b. Given the p-value of 0.294 and our cutoff of p=<0.05, there is insufficient evidence to reject the null hypothesis because the probability of observing a mean soil invertebrate abundance of 31 individuals is relatively high under the null hypothesis of no affect of plant type on soil invertebrate abundance.

Research Question 2: Does the presence of L. latifolium in the rosette stage affect the abundance of soil invertebrates?

Stop, Think, Do: Test the null hypothesis against the alternative hypothesis that the presence of L. latifolium in rosette stage (circular arrangement of leaves before flowering) affects the abundance of soil invertebrates.

```
#filter to create a data frame with the data needed
rosette <- filter(soil, stage_lep == "rosette")

# calculate mean invertebrate abundance per plot
m_rosette <- mean(rosette$n)
m_rosette
```

```
## [1] 10.55556
```

```
sem_rosette <- sd(rosette$n)/sqrt(9)
sem_rosette
```

```
## [1] 4.540286
```

```
#filter out values equal to or more extreme than the test statistic of 10.5, rounded up to 11
Nat_df2 <- filter(Nat_df, MeanSoilInvert <= 11)

# because of two tail distribution with a center at 25
Nat_df3 <- filter(Nat_df, MeanSoilInvert >= 49)

Nat_df23 <- rbind(Nat_df2, Nat_df3)

#calculate p-value from sums of probabilities
pvalue2 <- sum(Nat_df23$p)
pvalue2
```

```
## [1] 3e-04
```

Question Answers

a. The p-value is 0.0003.
b. Given the p-value of 0.0003 and our cutoff of p=<0.05, there is sufficient evidence to reject the null hypothesis because the probability of observing a mean soil invertebrate abundance of 11 individuals is extremely low under the null hypothesis of no effect of plant type/life stage on soil invertebrate abundance.

Discussion Question Answers

a. The null hypothesis is a claim that implies that there is no effect, association, or difference between treatments in a study. The alternative hypothesis is a statement that implies an effect, association, or difference between treatments in a study.
b. We need a null distribution to test our hypotheses so that we can compare the test statistic to an array of possible outcomes from our study to ultimately determine the probability of observing the test statistic.
c. The p-value is a number that represents the probability of observing the value of a test statistic or something more extreme. These p-values are extremely important in hypothesis testing because they determine if we are able to reject or fail to reject the null hypothesis based on our accepted p-value cutoff.