HW2

Yue Zhang

2024-08-29

R Markdown

A study was conducted to demonstrate that soybeans inoculated with nitrogen-fixing bacteria yield more and grow adequately without expensive environmentally deleterious synthesized fertilizers. The trial was conducted under controlled conditions with uniform amounts of soil. The initial hypothesis was that inoculated plants would outperform their uninoculated counterparts. This assumption is based on the facts that plants need nitrogen to manufacture vital proteins and amino acids and that nitrogen-fixing bacteria would make more of this substance available to plants, increasing their size and yield. There were 8 inoculated plants (I) and 8 uninoculated plants (U). The plant yield as measured by pod weight for each plant is given in Table 2.20.

2.36 Use graphic methods to compare the two groups.

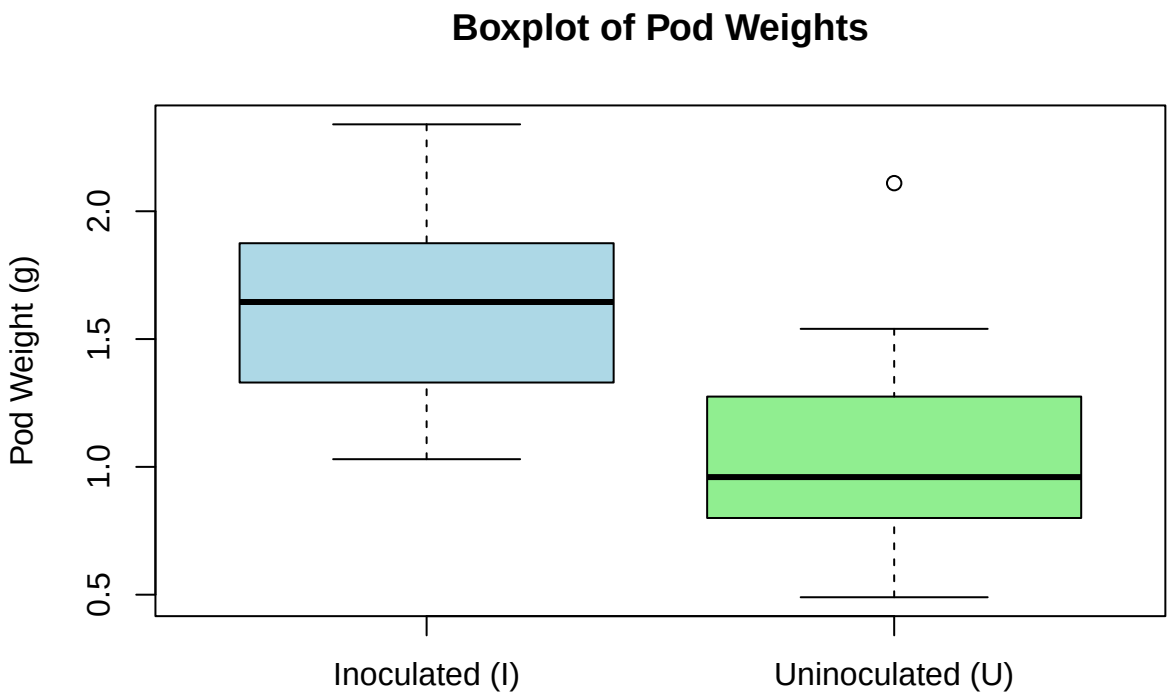
```
PodWeight = data.frame(I = c(1.76, 1.45, 1.03, 1.53, 2.34, 1.96,
    1.79, 1.21), U = c(0.49, 0.85, 1, 1.54, 1.01, 0.75, 2.11,
    0.92))

mean_I = mean(PodWeight$I)
sd_I = sd(PodWeight$I)
summary_I = summary(PodWeight$I)
print(summary_I)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.030   1.390   1.645   1.634   1.833   2.340
```

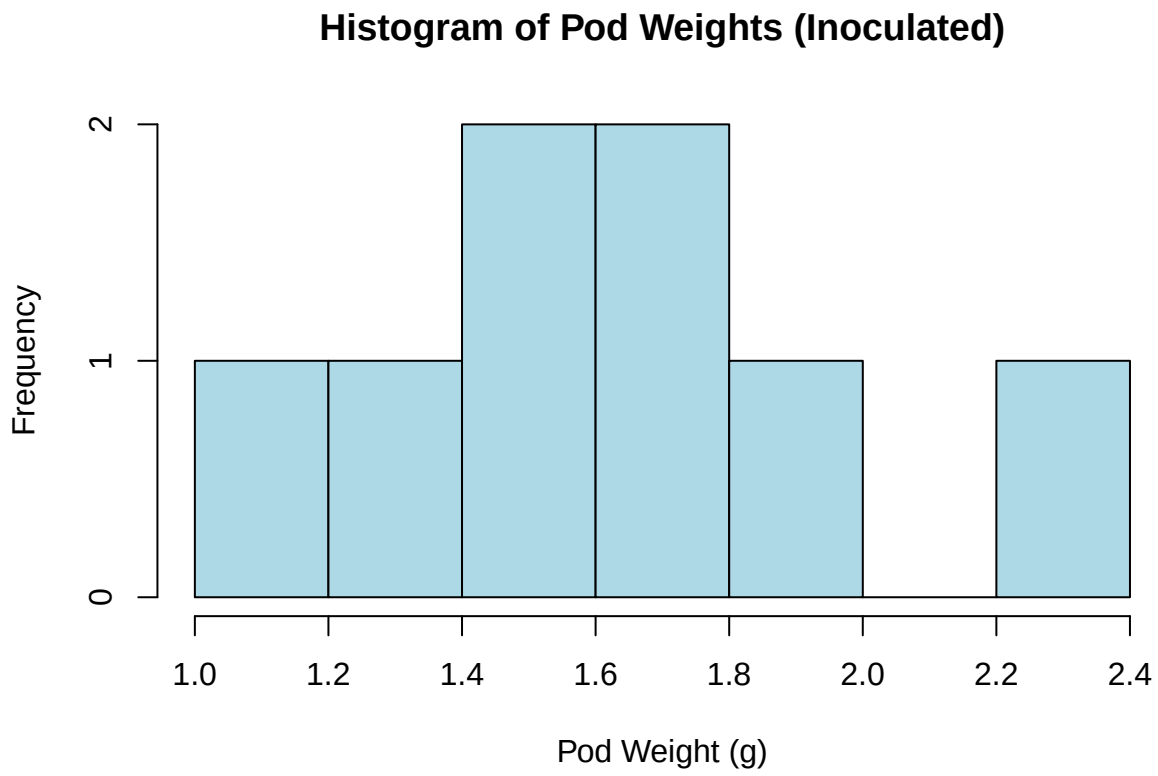
```
mean_U = mean(PodWeight$U)
sd_U = sd(PodWeight$U)
summary_U = summary(PodWeight$U)
print(summary_U)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.490   0.825   0.960   1.084   1.143   2.110
```

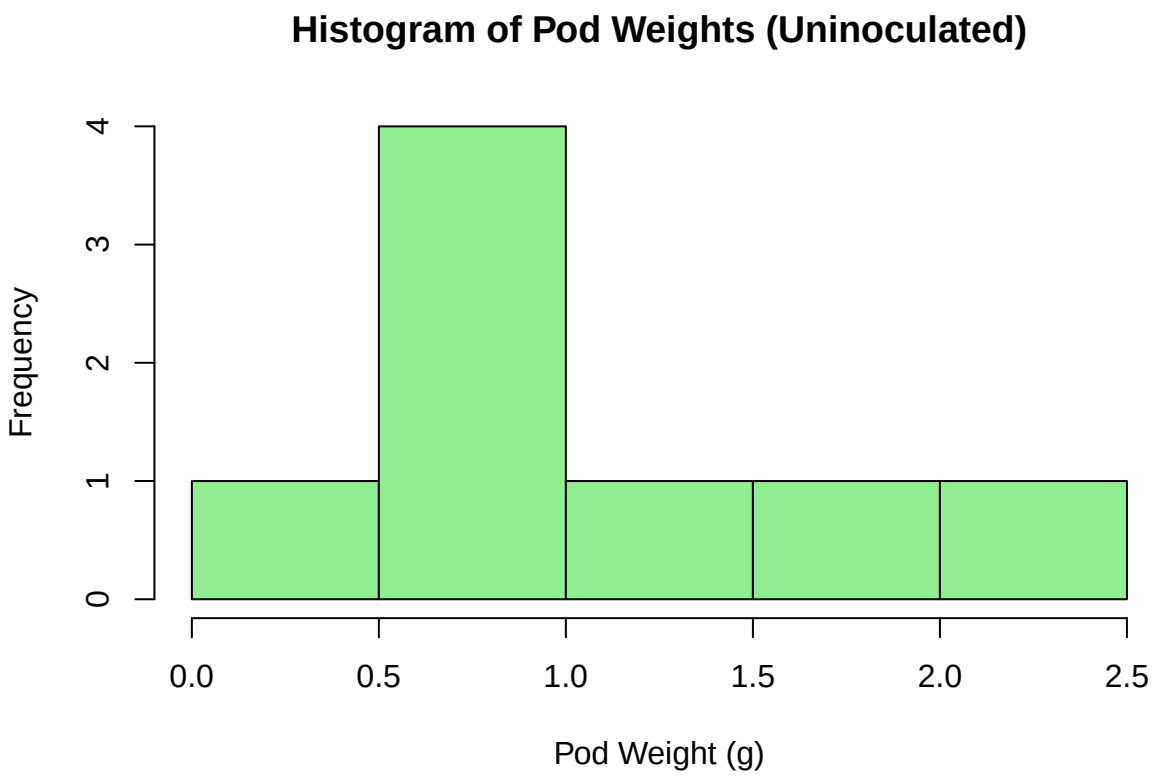
```
bp = boxplot(PodWeight$I, PodWeight$U, names = c("Inoculated (I)",
    "Uninoculated (U)"), main = "Boxplot of Pod Weights", ylab = "Pod Weight (g)",
    col = c("lightblue", "lightgreen"))
```




```r
h_I = hist(PodWeight$I, main = "Histogram of Pod Weights (Inoculated)",
    xlab = "Pod Weight (g)", ylab = "Frequency", col = "lightblue",
    border = "black", breaks = 5)
```




```r
h_U = hist(PodWeight$U, main = "Histogram of Pod Weights (Uninoculated)",
    xlab = "Pod Weight (g)", ylab = "Frequency", col = "lightgreen",
    border = "black", breaks = 5)
```



2.37 What is your overall impression concerning the pod weight in the two groups?

Answer:As the descriptive statistics show, the mean pod weight for I is higher than U, and the variance of I is lower than U, meaning that the U group will be more spread out than the I group. This means that inoculated with nitrogen-fixing bacteria will increase pod weight in this case, but the sample is too small. As the boxplot shows that, group I has a higher median and less spread compared to group U. Also, the distribution of pod weights for group I is higher than group U. Histograms of different groups show that most inoculated plants will cluster around higher values while uninoculated plants have a wider spread and lower overall weights.

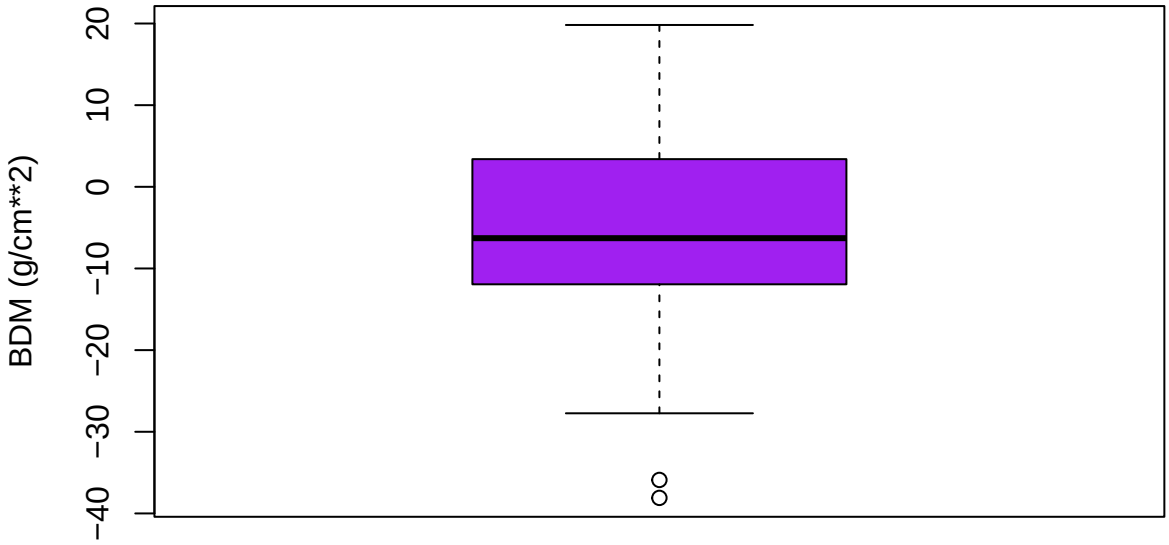
In Section 2.10, we described Data Set BONEDEN.DAT(at www.cengagebrain.com) concerning the effect of tobacco use on BMD.

2.38 For each pair of twins, compute the following for the lumbar spine: A = BMD for the heavier-smoking twin − BMD for the lighter-smoking twin = x1 − x2 B = mean BMD for the twinship = (x1 + x2)/2 C = 100% × (A/B) Derive appropriate descriptive statistics for C over the entire study population.

```
Boneden = read_xls("E:/Biostat/Biostatistics/PHL_1700/Data/Raw/boneden-1.xls")
Boneden$A = Boneden$ls2 - Boneden$ls1
Boneden$B = (Boneden$ls1 + Boneden$ls2)/2
Boneden$C = 100 * (Boneden$A/Boneden$B)
summary(Boneden$C)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -38.095 -11.940  -6.289  -4.950   3.390  19.820
```

```
bp2 = boxplot(Boneden$C, xlab = "Boxplot of C", ylab = "BDM (g/cm**2)",
    col = "purple")
```



The median and distribution of the boxplot shows that smoking seems to have a slightly negative effect on BMD since the median is below zero and most of the data points lie within the range of approximately -30% to 20%. Two points are outliers which means that the difference in BMD is more negative than other points, in this case, the heavier-smoking twin had significantly lower BMD.

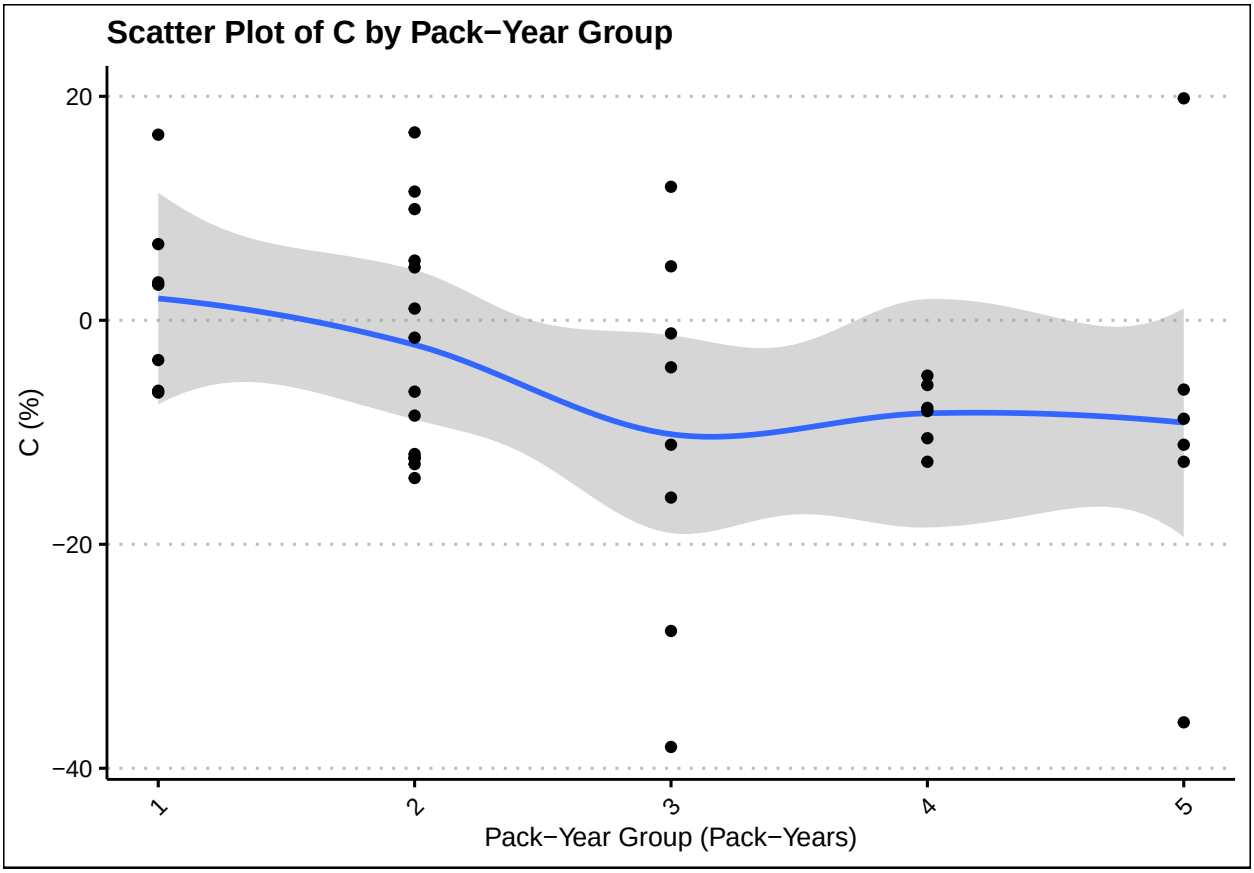
2.39 Suppose we group the twin pairs according to the difference in tobacco use expressed in 10 pack-year groups (0–9.9 pack-years/10–19.9 pack-years/20–29.9 pack-years/30–39.9 pack-years/40+ pack-years). Compute appropriate descriptive statistics, and provide a scatter plot for C grouped by the difference in tobacco use in pack-years.

```
packyears = abs(Boneden$pyr2 - Boneden$pyr1)
Boneden = Boneden %>%
    mutate(packyear_group = case_when(packyears < 10 ~ "0-9.9",
        packyears >= 10 & packyears < 20 ~ "10-19.9", packyears >=
            20 & packyears < 30 ~ "20-29.9", packyears >= 30 &
            packyears < 40 ~ "30-39.9", packyears >= 40 ~ "40+"))

desstat = Boneden %>%
    group_by(packyear_group) %>%
    summarize(count = n(), mean_C = mean(C), sd_C = sd(C), min_C = min(C),
        max_C = max(C))
print(desstat)
```

```
## # A tibble: 5 x 6
```

```
##   packyear_group count mean_C  sd_C  min_C max_C
##   <chr>          <int>  <dbl> <dbl>  <dbl> <dbl>
## 1 0-9.9              7   1.95  8.26  -6.45 16.6
## 2 10-19.9           14  -2.18 10.5  -14.1  16.8
## 3 20-29.9            8 -10.2  16.7  -38.1  11.9
## 4 30-39.9            6  -8.30  2.89 -12.6  -4.94
## 5 40+                6  -9.13 17.8  -35.9  19.8
```

```r
Boneden$packyear_group = as.numeric(as.factor(Boneden$packyear_group))
sp = ggplot(Boneden, aes(x = packyear_group, y = C)) + geom_smooth() +
    geom_point() + labs(title = "Scatter Plot of C by Pack-Year Group",
    x = "Pack-Year Group (Pack-Years)", y = "C (%)")
print(sp)
```



2.40 What impression do you have of the relationship between BMD and tobacco use based on Problem 2.39?

Answer:Descriptive statistics and the scatter plot show that tobacco use in pack-years tends to have a slightly negative effect on BMD. However, after group 3, the effect seems not to be consistent. The leveling off the trend line in the higher pack-year groups could indicate that after a certain level of smoking intensity, the impact on BMD does not continue to decrease. It could be the reason that the sample size or variability in these groups makes it difficult to detect a clear pattern. Also in group 5, there is an outlier.