stat359_A1_wducharme

Wesley Ducharme

2024-09-22

Question 2:

```
plant_growth <- read.csv(file ='C:/Users/wesch/uvic/stat359 Data
↪  Analysis/Assignments/A1/a1_plantgrowth_dataframe.csv')
```

a)

```
plant_growth
```

```
##    rowId treatment growth pot plot
## 1      1         1   14.6   1    1
## 2      2         1   15.2   1    1
## 3      3         1   13.2   2    1
## 4      4         1   12.9   2    1
## 5      5         1   16.4   3    1
## 6      6         1   12.2   3    1
## 7      7         1   18.5   1    2
## 8      8         1   16.7   1    2
## 9      9         1   22.2   2    2
## 10    10         1   18.8   2    2
## 11    11         1   24.7   3    2
## 12    12         1   20.3   3    2
## 13    13         2    7.1   1    1
## 14    14         2    7.7   1    1
## 15    15         2    6.8   2    1
## 16    16         2    6.0   2    1
## 17    17         2   10.0   3    1
## 18    18         2    8.3   3    1
## 19    19         2    9.7   1    2
## 20    20         2    8.8   1    2
## 21    21         2    6.8   2    2
## 22    22         2    9.0   2    2
## 23    23         2   10.4   3    2
## 24    24         2   11.3   3    2
```

b)

```r
attach(plant_growth)
plant_growth[order(growth),]
```

```
##    rowId treatment growth pot plot
## 16    16         2    6.0   2    1
## 15    15         2    6.8   2    1
## 21    21         2    6.8   2    2
## 13    13         2    7.1   1    1
## 14    14         2    7.7   1    1
## 18    18         2    8.3   3    1
## 20    20         2    8.8   1    2
## 22    22         2    9.0   2    2
## 19    19         2    9.7   1    2
## 17    17         2   10.0   3    1
## 23    23         2   10.4   3    2
## 24    24         2   11.3   3    2
## 6      6         1   12.2   3    1
## 4      4         1   12.9   2    1
## 3      3         1   13.2   2    1
## 1      1         1   14.6   1    1
## 2      2         1   15.2   1    1
## 5      5         1   16.4   3    1
## 8      8         1   16.7   1    2
## 7      7         1   18.5   1    2
## 10    10         1   18.8   2    2
## 12    12         1   20.3   3    2
## 9      9         1   22.2   2    2
## 11    11         1   24.7   3    2
```

c)

```r
growth_mean <- mean(growth)
growth_mean
```

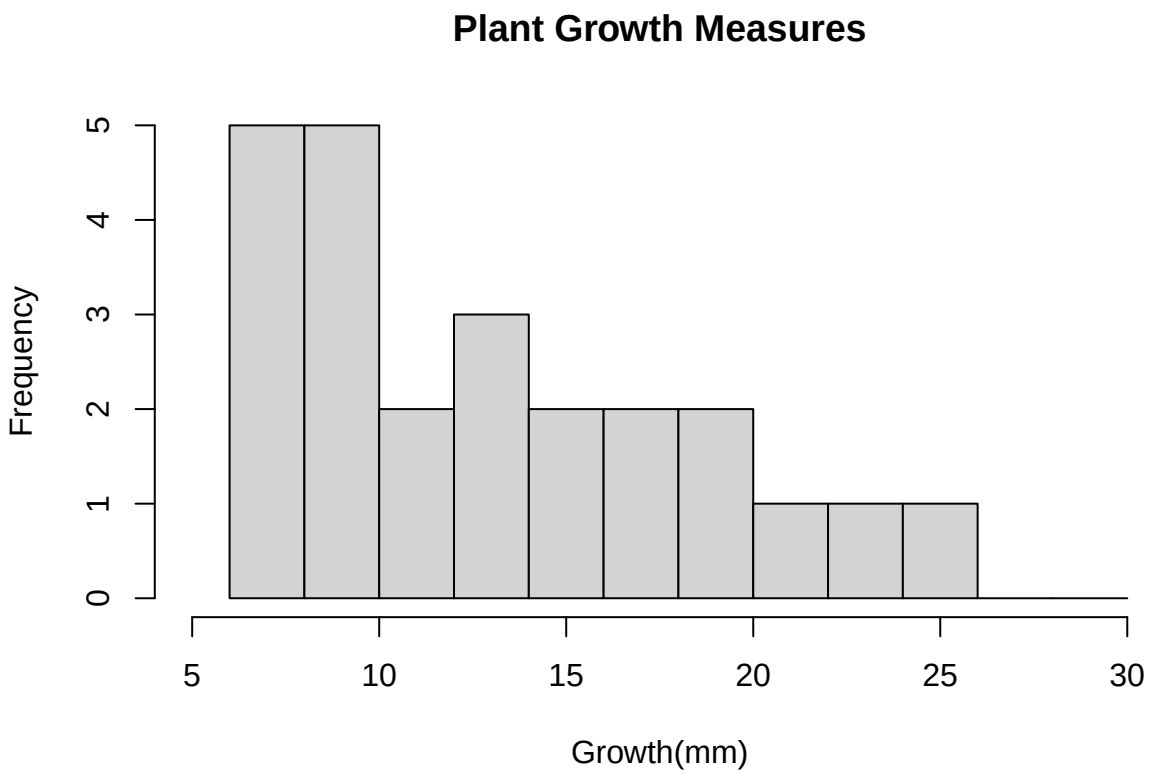
```
## [1] 12.81667
```

```r
growth_sd <- sqrt(var(growth))
growth_sd
```

```
## [1] 5.296813
```

d)

```r
hist(growth, breaks=seq(6,30,by=2), main="Plant Growth Measures",
     xlab="Growth(mm)", ylab="Frequency", xlim=range(5,30))
```



Question 3:

```
y <- c(11,11,10,8,11,3,15,11,7,6)

shortcut_variance <- function(y){
  y_sqrd <- y^2
  n <- length(y)
  sum_of_squares_shortcut <- sum(y_sqrd) - ((sum(y)^2)/n)
  sample_variance <- sum_of_squares_shortcut/(n-1)
  sample_variance ##return the sample variance
}
shortcut_variance(y)
```

```
## [1] 11.34444
```

Question 4:

```
tv <- read.table(file ='C:/Users/wesch/uvic/stat359 Data Analysis/Assignments/A1/tv.txt',
↪  sep="",header=TRUE, na.strings="NA")

library(knitr)

attach(tv)
```

a)

```
## remove missing values from data vector Canada
Canada<-Canada[!is.na(Canada)]
## remove missing values from data vector US
US<-US[!is.na(US)]
```

```
mean(Canada)
```

```
## [1] 64.3126
```

```
mean(US)
```

```
## [1] 69.33279
```

```
## US has a slightly higher mean watch time than Canada.
```

```
median(Canada)
```

```
## [1] 67.99453
```

```
median(US)
```

```
## [1] 69.47216
```

```
## US also has a slightly higher median than Canada.
```

```
var(Canada)
```

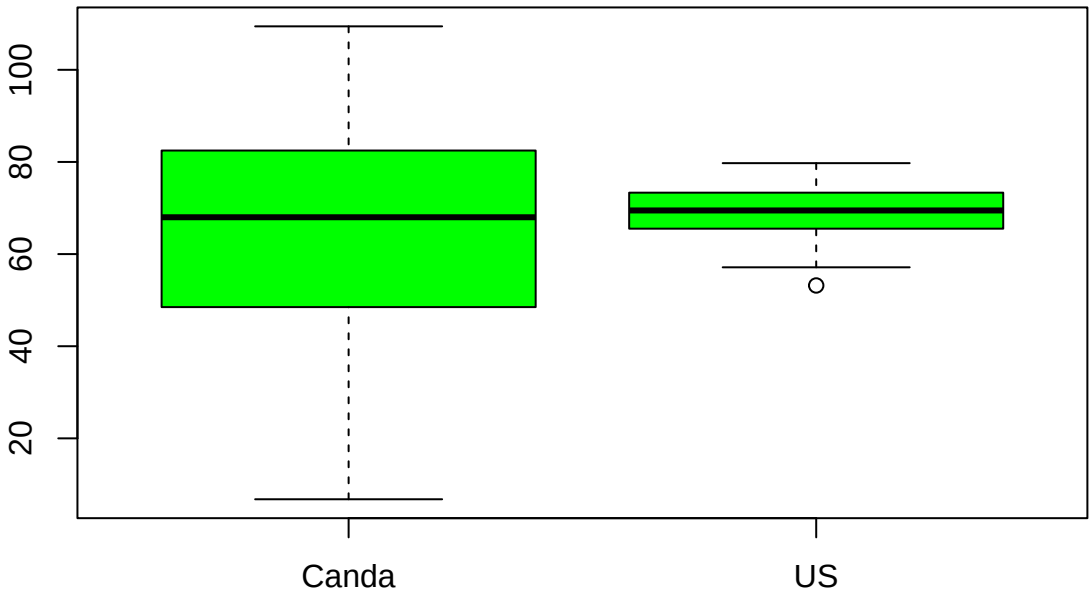
```
## [1] 533.1246
```

```
var(US)
```

```
## [1] 29.92508
```

```
## Canada's variance is very high in comparison to the US's Variance.
```

```
boxplot(tv$Canada, tv$US, col='green', names=c('Canda','US'))
title('Minutes watching tv')
```



From the boxpot we can see the confirmation of what was calculated earlier. With Canda and the US having very close medians but Canada's variance being much greater than the US's variance. Canada has more extreme outliers which contributes to it's greater variance.

b)

```
z.test <- function(y1, y2, H1){
  Z.obs <- (mean(y1) - mean(y2))/sqrt( (var(y1)/length(y1)) +
                                       (var(y2)/length(y2)))

  if(H1 == "two.sided"){
    p.value <- 2*(pnorm(Z.obs))
  }else if(H1 == "less"){
    p.value <- pnorm(Z.obs)
  }else if(H1 == "greater"){
    p.value <- 1 - pnorm(Z.obs)
  }
  p.value # return the p-value
}
```

c) mu1 = average for the Canadian population. mu2 = average for the US population.

I) for two sided H1: mu1 != mu2

```
p.two.sided <- z.test(Canada, US, "two.sided")
p.two.sided
```

```
## [1] 0.04417275
```

II) for one sided less than. H1: mu1 < mu2

```
p.less <- z.test(Canada, US, "less")
p.less
```

```
## [1] 0.02208637
```

III) for one sided greater than. H1: mu1 > mu2

```
p.greater <- z.test(Canada, US, "greater")
p.greater
```

```
## [1] 0.9779136
```

d) mu1 = average for the Canadian population. mu2 = average for the US population. The study outlines it's main question to be to determine if Canadian students watch less tv than American students. It would be best to use the alternative hypotheses H1: mu1 < mu2 which aligns best with the the question of "Do Canadian students watch less tv than American students.". The result of the test with this H1 is a p-value of 0.02208. Which means we have strong evidence against the null hypothesis H0: mu1 = mu2.