Stat 359 Assignment 1

2023-01-29

Question 2

Step one: create a data frame that represents the given data

```
plant_plot <- c(1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2)
plant_pot <- c(1,1,2,2,3,3,1,1,2,2,3,3,1,1,2,2,3,3,1,1,2,2,3,3)
plant_treatment <- c(1,1,1,1,1,1,2,2,2,2,2,2,1,1,1,1,1,1,2,2,2,2,2,2)
plant_growth <- c(14.6,15.2,13.2,12.9,16.4,12.2,
                  7.1,7.7,6.8,6.0,10.0,8.3,
                  18.5,16.7,22.2,18.8,24.7,20.3,
                  9.7,8.8,6.8,9.0,10.4,11.3)
df <- data.frame(plant_plot, plant_pot, plant_treatment, plant_growth)
df
```

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

Step two: Sort the data by plant growth

```
df[order(plant_growth),]
```

```
##    plant_plot plant_pot plant_treatment plant_growth
## 10          1         2               2          6.0
## 9           1         2               2          6.8
```

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

Step three: Calculate the mean and standard deviation of plant growth

```r
mean(plant_growth)
```

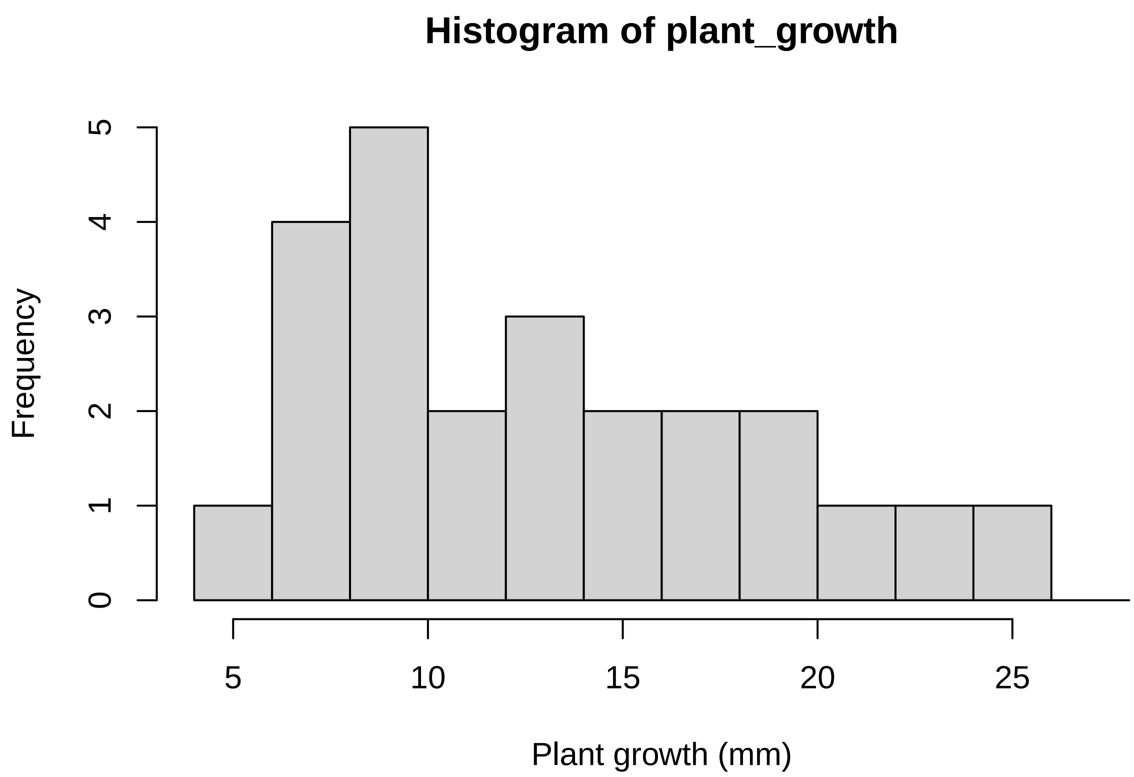
```
## [1] 12.81667
```

```r
sd(plant_growth)
```

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

Step four: plot the data as a histogram

```r
hist(plant_growth, xlab = "Plant growth (mm)", breaks = seq(from=4, to=28, by=2))
```



Question 3

```
calculate_variance <- function(y) {
  n <- length(y)
  mean_y <- mean(y)
  variance <- sum((y - mean_y)^2) / (n - 1)
  return(variance)
}

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

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

Question 4

First, read in the data from the provided tv.txt

```
tv_data<-read.table(file ='~/Desktop/school/assignments/Stat359/A1/tv.txt', sep="",header=TRUE, na.stri
```

Let's calculate some stats

```
summary(tv_data)
```

```
##      Canada              US
##  Min.   :  6.781   Min.   :53.19
##  1st Qu.: 48.667   1st Qu.:65.58
##  Median : 67.995   Median :69.47
##  Mean   : 64.313   Mean   :69.33
##  3rd Qu.: 82.340   3rd Qu.:73.21
```

```
##   Max.   :109.433   Max.   :79.74
##   NA's   :10
```

This gives us the mean, median and quartile values. Let's also calculate the sample variance for both the US and Canada

```r
cn_variance <- calculate_variance(tv_data[1:90,1])
cn_variance
```

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

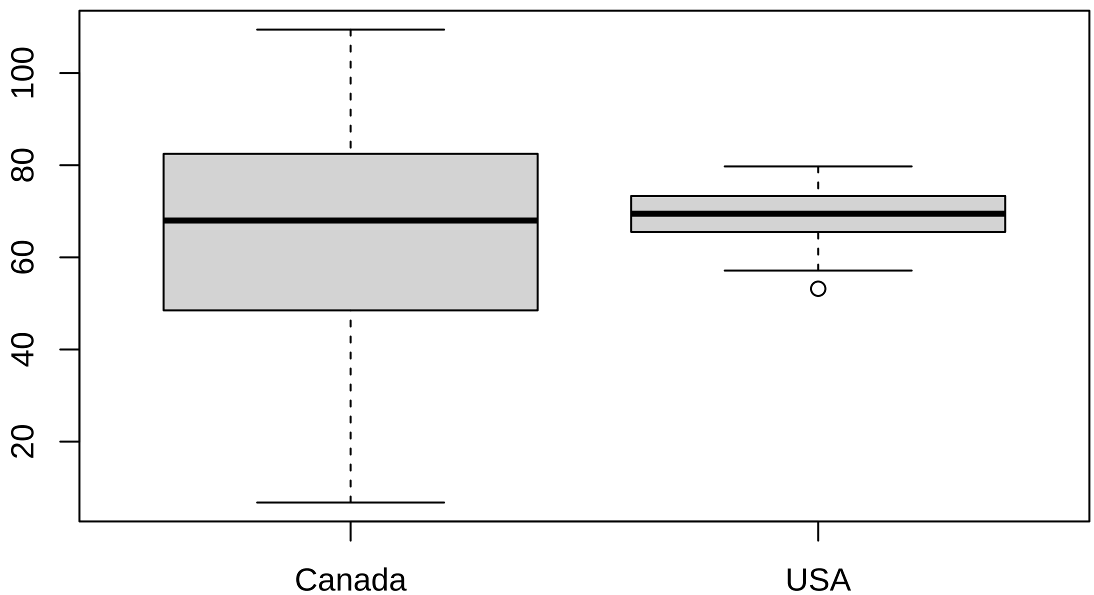
```r
us_variance <- calculate_variance(tv_data[,2])
us_variance
```

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

Whoa! Canada's variance is a lot higher! This makes sense, notice how the difference between Canada's max and min (6.7 vs 109.4) is significantly larger than that of the USA (53.2 vs 79.7)

Now let's create a box plot comparing the two

```r
boxplot(tv_data[1:90,1], tv_data[,2], names = c("Canada", "USA"))
```



Again, notice how much bigger Canada's range is.

Now let's create our Z test function

```r
z.test <- function(y1, y2, H1) {

  # Calculate mean of sample 1
  mu1 <- mean(y1)

  # Calculate mean of sample 2
  mu2 <- mean(y2)

  # Calculate standard deviation of sample 1
  s1 <- sd(y1)

  # Calculate standard deviation of sample 2
  s2 <- sd(y2)
```

```r
  # Calculate number of observations in sample 1
  n1 <- length(y1)

  # Calculate number of observations in sample 2
  n2 <- length(y2)

  # Calculate standard error of the difference in means
  se <- sqrt(s1^2/n1 + s2^2/n2)

  # Calculate the test statistic (z-score)
  z <- (mu1 - mu2) / se

  # Calculate p-value based on alternative hypothesis
  p <- if (H1 == "two.sided") {
    # two-sided test
    2 * (1 - pnorm(abs(z)))
  } else if (H1 == "less") {
    # less than alternative
    pnorm(z)
  } else if (H1 == "greater") {
    # greater than alternative
    1 - pnorm(z)
  }

  # Return p-value
  return(p)
}

two_sided_pval <- z.test(tv_data[1:90,1], tv_data[,2], "two.sided")
two_sided_pval
```

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

```r
less_pval <- z.test(tv_data[1:90,1], tv_data[,2], "less")
less_pval
```

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

```r
greater_pval <- z.test(tv_data[1:90,1], tv_data[,2], "greater")
greater_pval
```

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

The purpose of the study is to determine if Canadians watch less TV than Americans, the relevant alternative hypothesis is "less": $H_a : Ca - USA < 0$. The p value for this hypothesis is 0.022, suggesting it is likely true, since if it weren't we'd only have a approx 2% chance of seeing results this extreme.