Assignment1_stats359

Koki Itagaki

2023-01-30

#Question 2: Suppose the following data comes from a study on plant growth (mm) #where 2 plants are in each pot, 3 pots are within each plot and 2 plots are #given one of two fertilizer treatments.

```
#(a) Arrange the data into a dataframe so that it can be analysed. Print out
this dataframe.

Dataframe <- data.frame(growth = c(14.6,15.2, 18.5, 16.7,13.2,
12.9,22.2,18.8,
16.4,12.2,24.7,20.3,7.1,7.7, 9.7,8.8,6.8,6.0,6.8,9.0,10.0,8.3,10.4,11.3),
plot = c(1,1,2,2,1,1,2,2,1,1,2,2,1,1,2,2,1,1,2,2,1,1,2,2), Pot =
c(1,1,1,1,2,2,2,2,3,3,3,3,1,1,1,1,2,2,2,2,3,3,3,3), treatment =
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))
Dataframe
```

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

```r
attach(Dataframe)
#(b) Sort the data by plant growth.
sort(growth)
```

```
##  [1]  6.0  6.8  6.8  7.1  7.7  8.3  8.8  9.0  9.7 10.0 10.4 11.3 12.2 12.9
13.2
## [16] 14.6 15.2 16.4 16.7 18.5 18.8 20.3 22.2 24.7
```

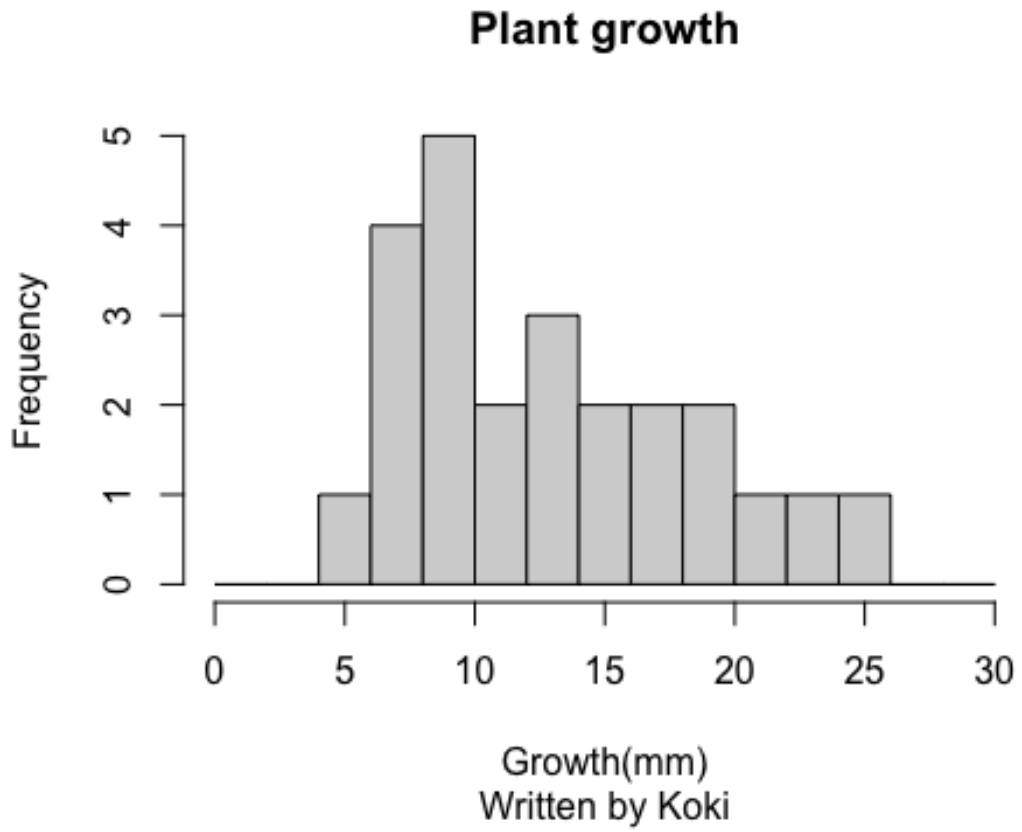
```r
#(c) Calculate the mean, and standard deviation of the data.
mean(growth)
```

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

```r
sd(growth)
```

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

```r
#(d) Plot the data using a histogram (R function hist()).
#Clearly label the axis, title and use bin sizes of 2 mm.
hist(growth, xlab = "Growth(mm)", main = "Plant growth",sub = "Written by
Koki",
     breaks = seq(0, 30, by = 2))
```



#Question 3: Write a function that uses the short cut formula to calculate #the sample variance of a data vector. #Use the vector y=(11,11,10,8,11,3,15,11,7,6) as your test vector.

```
sample.variance<-function(y){
 return.variance <-  (1/(length(y)-1))*sum((y - sum(y)/length(y))^2)
 return.variance
}
y <-c(11,11,10,8,11,3,15,11,7,6)
sample.variance(y)
```

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

#Question4: On the course webpage you will find a dataset with filename 'tv.txt' #The data arise from a study examining the time teenagers spend watching tv. #A random sample of n = 100 eighth grade American high school students was #obtained, and the number of minutes spent watching TV during the first week #of October was recorded. A similar sample of m = 90 Canadian students was also #obtained. In this study it is of interest to compare the TV watching habits of #the teenagers from the two different countries, specifically to determine #if Canadian students watch less TV than their American counterparts.

```
#(a)Compare the two samples using appropriate descriptive statistics,
including side-by-side boxplots
tv<-read.table(file ='~/Desktop/stat359/data/tv.txt', sep="",header=TRUE)
tv
```

```
##         Canada       US
## 1    76.972030 65.71819
## 2    81.930050 66.84723
## 3    10.287570 72.77606
## 4    46.531230 73.58473
## 5    84.947600 69.39871
## 6    91.493270 67.56986
## 7    45.420800 71.57007
## 8    48.490550 65.10214
## 9    74.460080 63.34076
## 10   89.445170 71.82348
## 11   76.351960 62.37042
## 12   78.265030 71.44686
## 13   43.435870 62.32819
## 14   54.291900 71.52563
## 15   49.196370 64.38735
## 16   77.869930 66.11336
## 17   68.456650 75.35511
## 18   85.307040 67.26076
## 19  107.262700 72.18155
## 20   87.499820 70.73481
## 21   59.375800 74.22746
## 22  104.794000 64.42946
## 23   77.580930 70.44878
```

```
## 24   83.481240 65.90180
## 25   57.427760 66.22887
## 26   85.615570 61.76221
## 27   54.605680 73.59683
## 28   32.062650 68.53207
## 29   37.905940 64.52083
## 30   93.184760 74.59507
## 31   47.369310 79.51543
## 32   90.605810 77.31957
## 33   80.776960 62.89537
## 34   82.678450 77.64112
## 35   53.531920 72.38140
## 36   51.586710 59.25923
## 37   28.954730 65.63279
## 38   77.009840 73.83070
## 39   35.947980 71.76344
## 40   35.962380 72.22720
## 41   70.216140 70.38777
## 42   63.404160 64.79948
## 43   90.616150 68.61543
## 44  109.433200 65.26773
## 45   38.803220 77.55310
## 46   57.180420 61.87569
## 47   44.269680 72.74478
## 48    6.780787 75.87815
## 49   84.085740 68.30550
## 50   72.804490 68.63948
## 51   59.398720 71.63996
## 52   67.532420 67.19000
## 53   82.476580 72.01224
## 54   81.046340 72.10090
## 55   50.568320 75.21945
## 56   15.471450 62.60472
## 57   94.195660 67.48466
## 58   39.907960 73.76562
## 59   93.990450 77.86371
## 60   69.309160 69.54561
## 61   91.233800 75.83243
## 62   56.742160 61.08572
## 63   87.393340 66.01869
## 64   55.501590 59.36347
## 65   69.728300 66.80645
## 66   67.182820 57.13115
## 67   58.493330 60.71688
## 68   64.017000 76.74675
## 69   61.051710 71.23923
## 70   69.191360 66.35301
## 71   22.838230 72.38831
## 72   71.295020 63.04627
## 73   58.090690 75.19135
```

```
## 74   74.525230 75.51897
## 75   44.924740 65.63971
## 76   84.234600 66.29252
## 77   54.469070 63.66657
## 78   74.041100 65.40320
## 79   69.108830 74.43947
## 80   88.787660 63.27804
## 81   53.340000 72.99844
## 82   61.522300 66.51039
## 83    7.260062 69.39116
## 84  105.424900 77.31849
## 85   36.410640 73.69178
## 86   21.535670 65.93512
## 87   42.868120 68.92336
## 88   48.353670 53.18675
## 89   73.210680 63.68953
## 90   75.492600 79.67048
## 91          NA 71.48828
## 92          NA 71.18937
## 93          NA 63.76111
## 94          NA 69.72101
## 95          NA 79.73981
## 96          NA 75.97058
## 97          NA 68.51689
## 98          NA 71.24320
## 99          NA 79.45313
## 100         NA 73.08269
```

```
summary(tv)
```

```
##      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
```

```
min_canada<-(tv$Canada[!is.na(tv$Canada)])
min_US<-(tv$US[!is.na(tv$US)])
c(mean(min_canada),mean(min_US))
```

```
## [1] 64.31260 69.33279
```

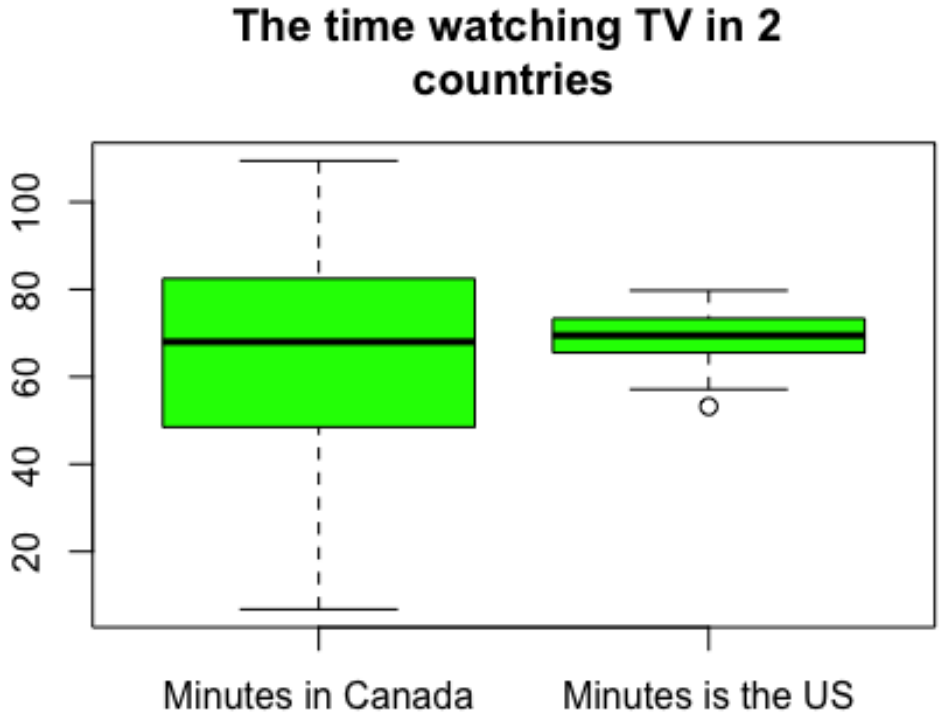
```
summary(min_canada)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   6.781  48.667  67.995  64.313  82.340 109.433
```

```
summary(min_US)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   53.19   65.58   69.47   69.33   73.21   79.74

boxplot(min_canada,min_US, col='green',main = "The time watching TV in 2
countries", sub = "Written by Koki Itagaki",
        names=c('Minutes in Canada','Minutes is the US'))
```




```
#According to the data above, we can see that the data of Canada was spreaded
#more than the data of the U.S. the both means are really close and,
#from the data above it is hard to find out the difference between 2 groups.



#determine if Canadian students watch less TV than their American
counterparts.

#(b)Write an R function z.test(y1,y2,H1) to compute the p-value for a large
sample z-test
#(discussed in lecture) for testing equality of two population means (H0 : μ1
= μ2).
#y1, a vector containing the sample measurements from the first population;
#y2, a vector containing the sample measurements from the second population;
```

```r
#and H1, a string variable, which takes one of three possible values:
#‘two.sided’, ‘less’ or ‘greater’ specifying the alternative hypothesis.
z.test<-function(y1,y2,H1){
  n<-length(y1)
  m<-length(y2)
  # compute the value of the test statistic
  Z.obs<-(mean(y1) - mean(y2))/sqrt( (var(y1)/length(y1)) +
(var(y2)/length(y2)))
  # compute the p-value
  if(n>=30 && m>=30){
    if(H1 == "less"){
      p.value.obs<-pnorm(Z.obs)
      p.value.obs
    }else if(H1 == "two.sided"){
      p.value.obs<-2*(1 - pnorm(abs(Z.obs)))
      p.value.obs
    }else if(H1 == "greater"){
      p.value.obs<-1 - pnorm(Z.obs)
      p.value.obs
  }
  }else{
    print("The sample size is not large enough for z-test")
  }
}

#(c) Apply your function to the TV data, computing the p-values for each of
the three possible
#alternative hypotheses.
z.test(min_canada,min_US,"two.sided")
```

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

```r
z.test(min_canada,min_US,"greater")
```

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

```r
z.test(min_canada,min_US,"less")
```

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

```r
#(d) Which of the three alternative hypotheses is relevant for the particular
#question being asked in this study? Comment on the results.

#My answer
#We would like to know if tenagers in Canada watch TV less time than
teenagers in the U.S.
#Let u1 = the average time teenagers in Canada watch TV and let u2 = the time
teenagers in the U.S
#"watch the TV. So the alternative hypotheses is Ha: u1 - u2 < 0 (or u1 < u2)
```

#(d) Which of the three alternative hypotheses is relevant for the particular #question being asked in this study? Comment on the results.

#My answer: We would like to know if tenagers in Canada watch TV less time than teenagers in the U.S.Let u1 = the average time teenagers in Canada watch TV and let u2 = the time teenagers in the U.S watch the TV. So the alternative hypotheses is Ha: u1 - u2 < 0 (or u1 < u2)