BRSM Results Analysis

George Paul

2024-02-19

```r
rm(list = setdiff(ls(), lsf.str()))
```

```r
# install.packages('readxl')
library(readxl)
```

```r
excel_path <- 'D:\\FILES\\BRSM_Results Visualization.xlsx'
data <- read_excel(excel_path)
data
```

```
## # A tibble: 43 x 2
##    Group `No. of Mosquitoes`
##    <chr>                <dbl>
##  1 Beer                    27
##  2 Beer                    19
##  3 Beer                    20
##  4 Beer                    20
##  5 Beer                    23
##  6 Beer                    17
##  7 Beer                    21
##  8 Beer                    24
##  9 Beer                    31
## 10 Beer                    26
## # i 33 more rows
```

```r
nums <- data[["No. of Mosquitoes"]]
nums
```

```
##  [1] 27 19 20 20 23 17 21 24 31 26 28 20 27 19 25 31 24 28 24 29 21 21 18 27 20
## [26] 21 19 13 22 15 22 15 22 20 12 24 24 21 19 18 16 23 20
```

```r
grps <- data[["Group"]]
grps
```

```
##  [1] "Beer"  "Beer"  "Beer"  "Beer"  "Beer"  "Beer"  "Beer"  "Beer"  "Beer"
## [10] "Beer"  "Beer"  "Beer"  "Beer"  "Beer"  "Beer"  "Beer"  "Beer"  "Beer"
## [19] "Beer"  "Beer"  "Beer"  "Beer"  "Beer"  "Beer"  "Beer"  "Water" "Water"
## [28] "Water" "Water" "Water" "Water" "Water" "Water" "Water" "Water" "Water"
## [37] "Water" "Water" "Water" "Water" "Water" "Water" "Water"
```

```r
count_b <- sum(grps == "Beer")
count_b
```

```
## [1] 25
```

```r
count_w <- sum(grps == "Water")
count_w
```

```
## [1] 18
```

```r
beer_list <- subset(data, Group == "Beer")
beer_list
```

```
## # A tibble: 25 x 2
##    Group `No. of Mosquitoes`
##    <chr>                <dbl>
##  1 Beer                    27
##  2 Beer                    19
##  3 Beer                    20
##  4 Beer                    20
##  5 Beer                    23
##  6 Beer                    17
##  7 Beer                    21
##  8 Beer                    24
##  9 Beer                    31
## 10 Beer                    26
## # i 15 more rows
```

```r
watr_list <- subset(data, Group == "Water")
obs_stat_med <- median(beer_list$`No. of Mosquitoes`) - median(watr_list$`No. of Mosquitoes`)
obs_stat_med
```

```
## [1] 4
```

```r
obs_stat_t <- t.test(watr_list$`No. of Mosquitoes`, beer_list$`No. of Mosquitoes`)
as.double(obs_stat_t$statistic)
```

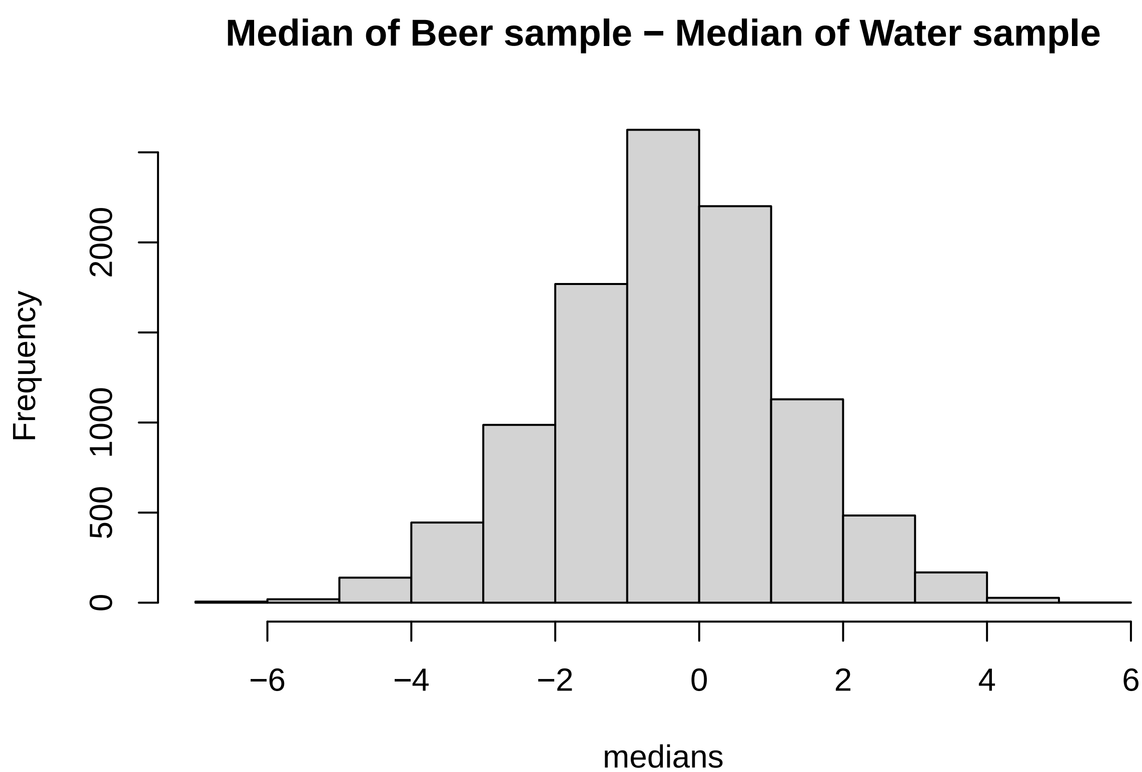
```
## [1] -3.658245
```

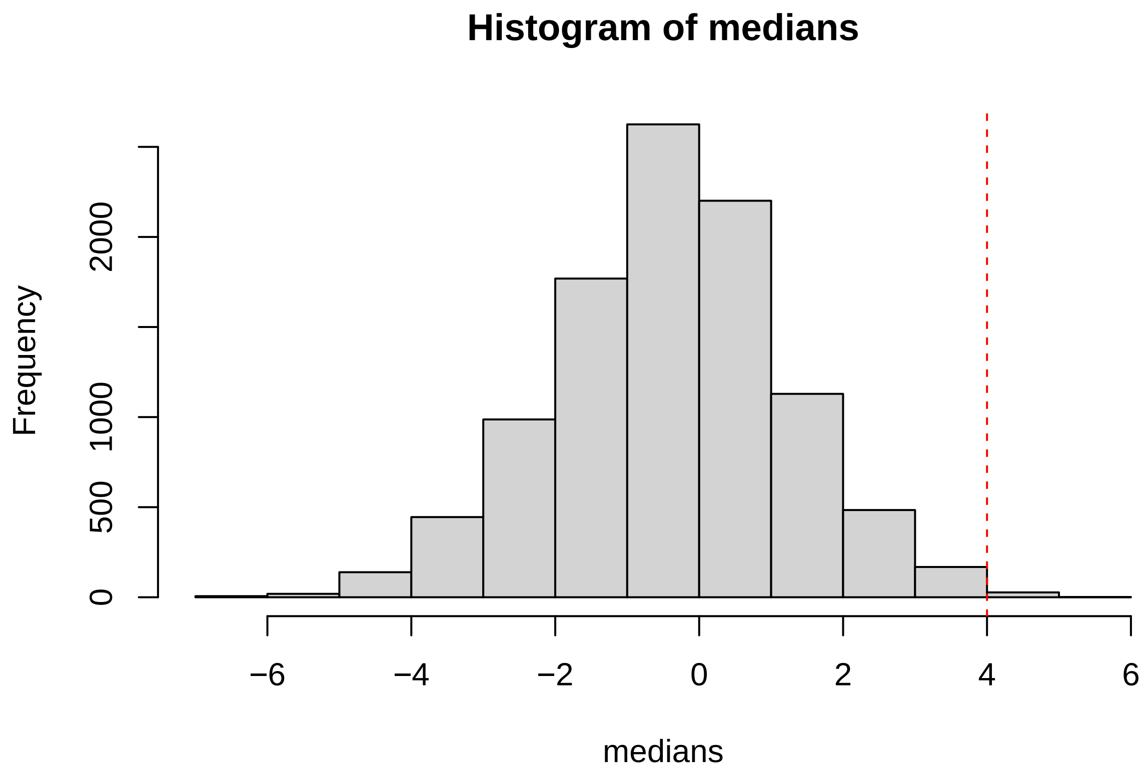
Question 2.a

```r
get_group_median_diff <- function() {
  beer_sample <- sample(nums, count_b, replace = TRUE)
  watr_sample <- sample(nums, count_w, replace =TRUE)
  median(beer_sample) - median(watr_sample)
}
get_group_median_diff()
```

```
## [1] -1
```

```r
counter <- 0
iter_count <-10000

medians <- c()

repeat {
  medians <- c(medians, get_group_median_diff())

  counter <- counter + 1
  if (counter >= iter_count) {
    break
  }
}

plot(hist(medians, main = "Median of Beer sample - Median of Water sample"))+
abline(v = obs_stat_med, col = "red", lty = 2)
```






```
## integer(0)
```

```r
p_val_qa <- sum(medians >= obs_stat_med) / length(medians)
p_val_qa
```

```
## [1] 0.0114
```

The calculated p-value is: $\frac{\text{Number of values where val} \geq \text{Observed Statistic}}{\text{Total Number of values}}$

This was calculated to be p-value $\approx 0.01 < \alpha$. Hence we can conclude that it is statistically significant.

```r
p_val_qa_nondir <- (sum(medians >= obs_stat_med) + sum(medians <= -obs_stat_med)) / length(medians)
p_val_qa_nondir
```

2.c for (a) step statistic

```
## [1] 0.0278
```

The calculated p-value for non-directional hypothesis is: $\frac{\text{Number of values where val} \geq \text{Observed Statistic+Number of values where val} \leq (-}{\text{Total Number of values}}$

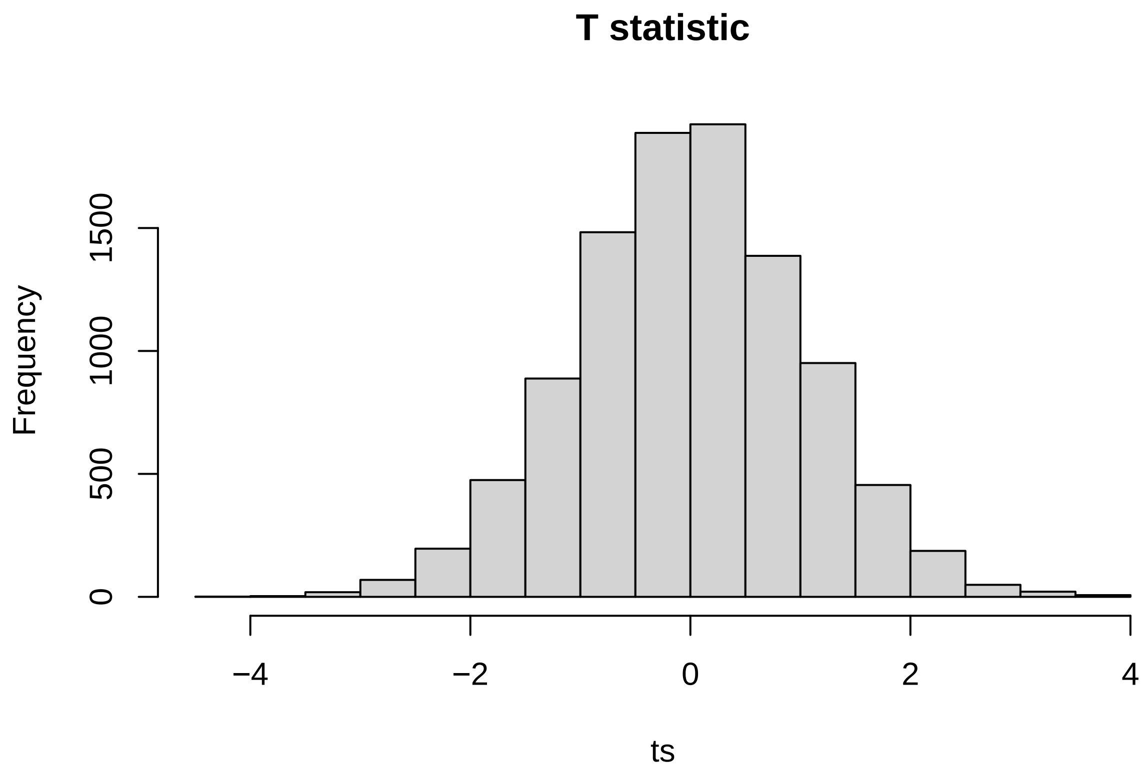
This was calculated to be p-value $\approx 0.02 < \alpha$. Hence we can conclude that it is statistically significant.

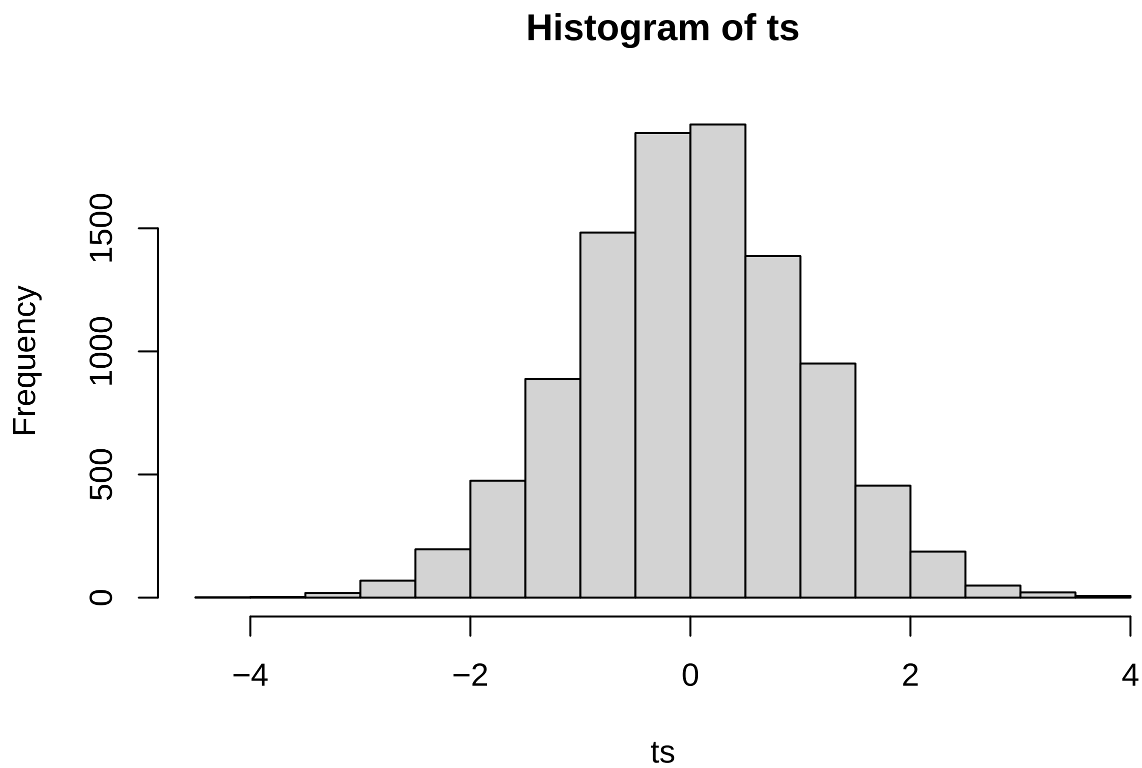
Question 2.b

```
get_group_t <- function() {
  beer_sample <- sample(nums, count_b, replace = TRUE)
  watr_sample <- sample(nums, count_w, replace =TRUE)
  as.double(t.test(beer_sample, watr_sample)$statistic)
}
get_group_t()
```

```
## [1] -0.2575364
```

```
counter <- 0
iter_count <-10000

ts <- c()

repeat {
  ts <- c(ts, get_group_t())

  counter <- counter + 1
  if (counter >= iter_count) {
    break
  }
}

plot(hist(ts, main = "T statistic"))
```


T statistic
Frequency
0
500
1000
1500
−4
−2
0
2
4
ts




```
# abline(v = obs_stat_t, col = "red", lty = 2)
```

Question 3

```
excel_path <- 'D:\\FILES\\iqdata.xlsx'
data <- read_excel(excel_path)
```

```
## New names:
## * `` -> `...1`
## * `GPA` -> `GPA...2`
## * `GPA` -> `GPA...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...15`
## * `` -> `...16`
## * `` -> `...18`
## * `` -> `...19`
## * `` -> `...20`
## * `` -> `...22`
## * `` -> `...23`
```

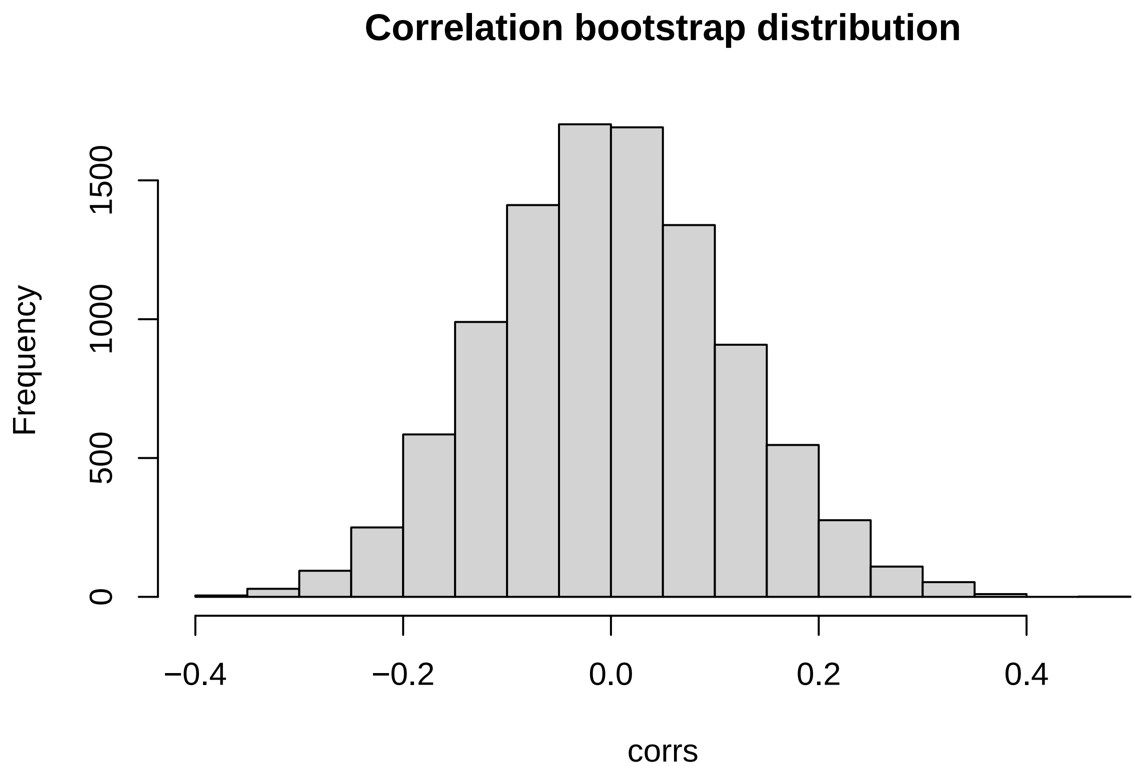
```r
data
```

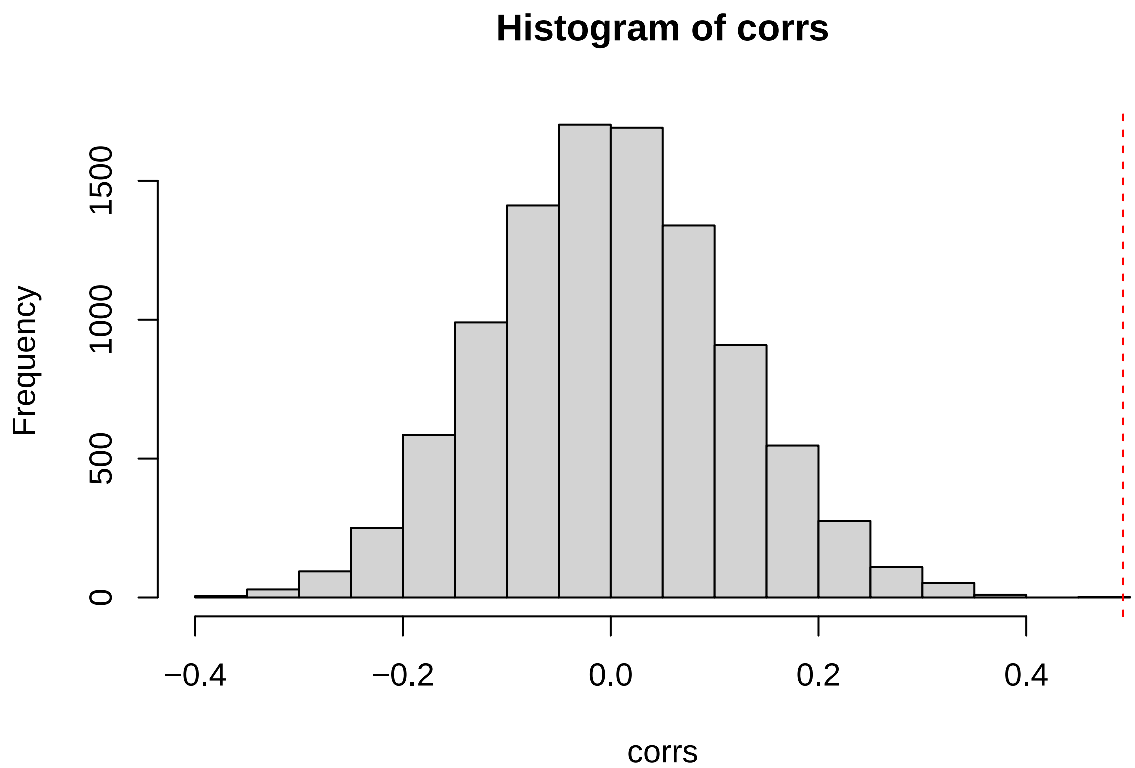
```
## # A tibble: 78 x 23
##     ...1 GPA...2    IQ GENDER `Placement \r\nTESTSCORE` GPA...6 ...7  ...8
##    <dbl>   <dbl> <dbl>  <dbl>                     <dbl>   <dbl> <lgl> <lgl>
##  1     1    7.94   111      2                        67    7.94 NA    NA
##  2     2    8.29   107      2                        43    8.29 NA    NA
##  3     3    4.64   100      2                        52    4.64 NA    NA
##  4     4    7.47   107      2                        66    7.47 NA    NA
##  5     5    8.88   114      1                        58    8.88 NA    NA
##  6     6    7.58   115      2                        51    7.58 NA    NA
##  7     7    7.65   111      2                        71    7.65 NA    NA
##  8     8    2.41    97      2                        51    2.41 NA    NA
##  9     9    6      100      1                        49    6    NA    NA
## 10    10    8.83   112      2                        51    8.83 NA    NA
## # i 68 more rows
## # i 15 more variables: Exerice_Times <dbl>, `Exercise code` <dbl>,
## #   Anxiety <dbl>, ...12 <lgl>, ...13 <lgl>, `anxiety scores` <chr>,
## #   ...15 <chr>, ...16 <lgl>, `t-Test: Paired Two Sample for Means` <chr>,
## #   ...18 <chr>, ...19 <chr>, ...20 <lgl>,
## #   `t-Test: Two-Sample Assuming Equal Variances` <chr>, ...22 <chr>,
## #   ...23 <chr>
```

```r
sco_list <- data$`Placement \r\nTESTSCORE`
iqs_list <- data$IQ
original_corr <- cor(sco_list, iqs_list)
original_corr
```

```
## [1] 0.4931479
```

```r
counter <- 0
iter_count <-10000

corrs <- c()

repeat {
  shuff_sco_list <- sco_list[sample(length(sco_list))]
  shuff_iqs_list <- iqs_list[sample(length(iqs_list))]
  cor(shuff_sco_list, shuff_iqs_list)
  corrs <- c(corrs, cor(shuff_sco_list, shuff_iqs_list))

  counter <- counter + 1
  if (counter >= iter_count) {
    break
  }
}

plot(hist(corrs, main = "Correlation bootstrap distribution"))+
abline(v = original_corr, col = "red", lty = 2)
```


Correlation bootstrap distribution
Frequency
0
500
1000
1500
−0.4
−0.2
0.0
0.2
0.4
corrs




```
## integer(0)
```

As it stands in the latest simulation, the originally calculated correlation is beyond the range of any of the randomly generated data. Hence the calculated p-value will be $\approx 0 \leq \alpha = 0.05$. We can conclude that the correlation that was found was statistically quite significant.