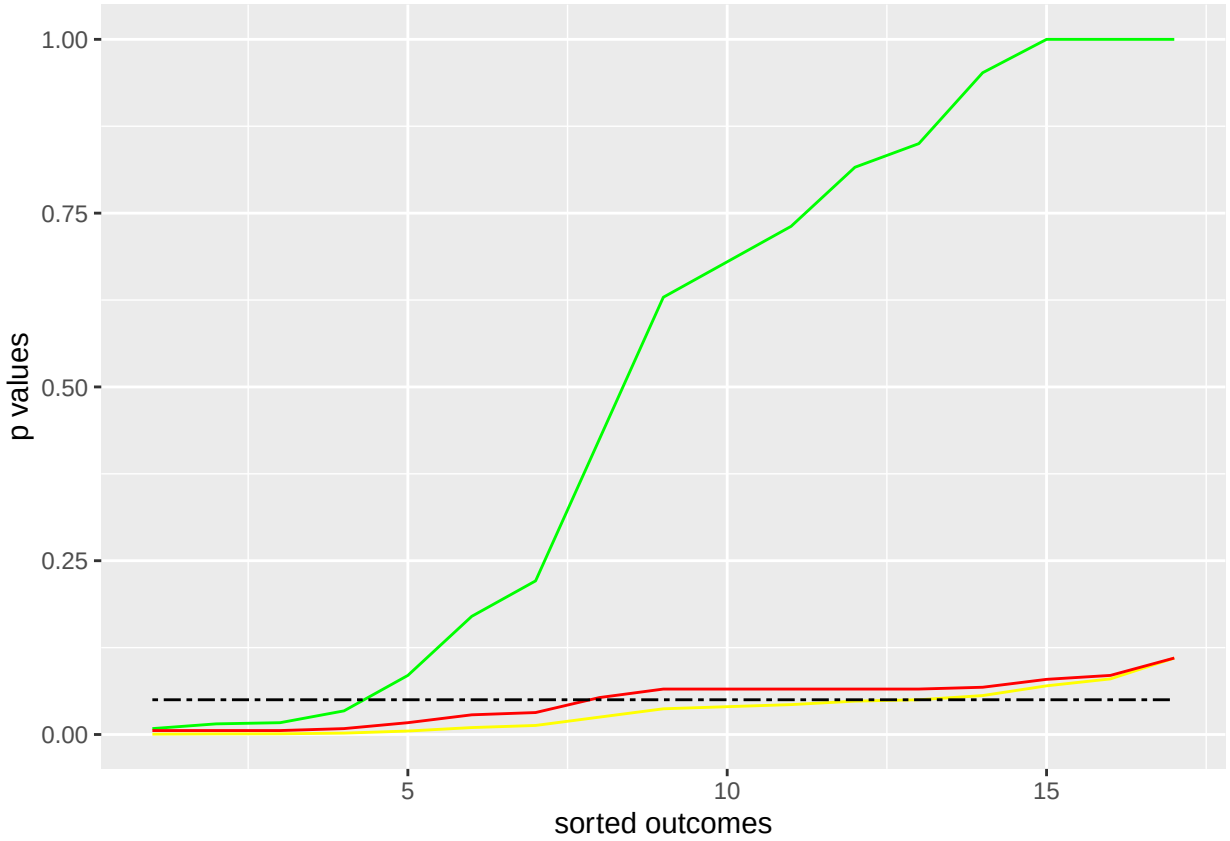
Multiple testing & Permutation Test Practicals

Akash C R

2023-02-20

Question 1 : Correction to p-values.

```
library(ggplot2)

p_values = c(0.0050, 0.0010, 0.0100, 0.0005, 0.0009, 0.0400, 0.0560, 0.0500,
0.0480, 0.0130, 0.0370, 0.0430, 0.0020, 0.0250, 0.1100, 0.0700, 0.0800)

p_values = sort(p_values)

bonferroni = p.adjust(p_values, method = "bonferroni")
bh = p.adjust(p_values, method = "BH")
alpha = c(rep(0.05,17))

x_values = c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17)

mydf = data.frame(x_values, p_values, bonferroni, bh, alpha)


ggplot(mydf, aes(x = x_values)) +
  geom_line(aes(y = p_values), color = "yellow") +
  geom_line(aes(y = bonferroni), color = "green") +
  geom_line(aes(y = bh), color = "red") +
  geom_line(aes(y = alpha), color = "black", linetype = "twodash") +
  labs( y = "p values", x = "sorted outcomes")
```



Legend for the line graph :

- Yellow line = Unadjusted p-values
- Red line = Benjamini-Hochberg corrected p-values
- Green line = Bonferroni corrected p-values
- Black dotted line = alpha (0.05)

Explanation as per the line plot.

- We know that Bonferroni method is more conservative one. The results of the correction is indeed support that. Since we just multiply the number of observations to correct the p-values in Bonferroni its really high values compared to unadjusted p-values.

- In case Benjamin-Hochberg, adjusting is done keeping the constraints in mind, so intuitevely it has to be higher than un-adjusted p-values and lower than Bonferroni which is indeed true as per the graph plotted.

Question 2 : Attractiveness of humans post consumption of beer to mosquitoes.

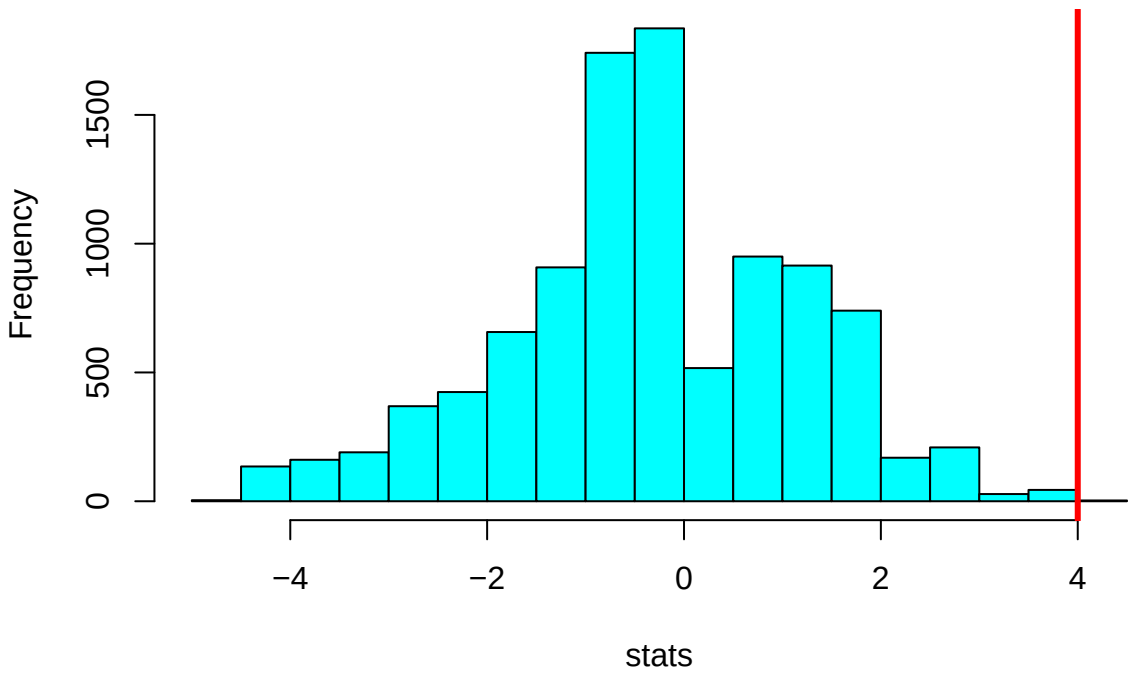
Part a : Permutation test using difference between medians

```
library("xlsx")

qn2 = read.xlsx("BRSM_Results.xlsx", sheetIndex = 1)
beer = head(qn2, 25)
water = tail(qn2, 18)


beer_median = median(beer$No..of.Mosquitoes)
water_median = median(water$No..of.Mosquitoes)

diff = beer_median - water_median
print(paste("Difference between medians is : ", diff))
```

```
## [1] "Difference between medians is :  4"
```

```
p <- 10000
variable = qn2$No..of.Mosquitoes
n = length(qn2$No..of.Mosquitoes)
permsample = matrix(0, nrow = n, ncol = p)

stats<-rep(0, p)

for ( i in 1:p)
{
  permsample[,i] <- sample(variable, size = n, replace = FALSE)
}

for (i in 1:p)
{
  stats[i] <- median(permsample[qn2$Group == "Beer", i]) -
                              median(permsample[qn2$Group == "Water", i])
}
```

```
hist(stats, col = 'cyan', main = "Distribution of median differences from 10000 permutations")
abline(v=diff, lwd = 3, col = "red")
```




```
p_value = sum(stats >= diff)/p
print(paste("Observed p values is : ",p_value))
```

```
## [1] "Observed p values is :  0.0047"
```

As we can see from the above p-value is less than 0.05, there is a very high chance that the alternate directional hypothesis is true. Thus the observed statistic is significant

Part b : Permutation test using t score

```
original_t = t.test(beer$No..of.Mosquitoes, water$No..of.Mosquitoes, var.equal =  TRUE)
original_t = original_t$statistic[['t']]
print(paste("Original t score is :  ", original_t))
```

```
## [1] "Original t score is :  3.58698438321434"
```

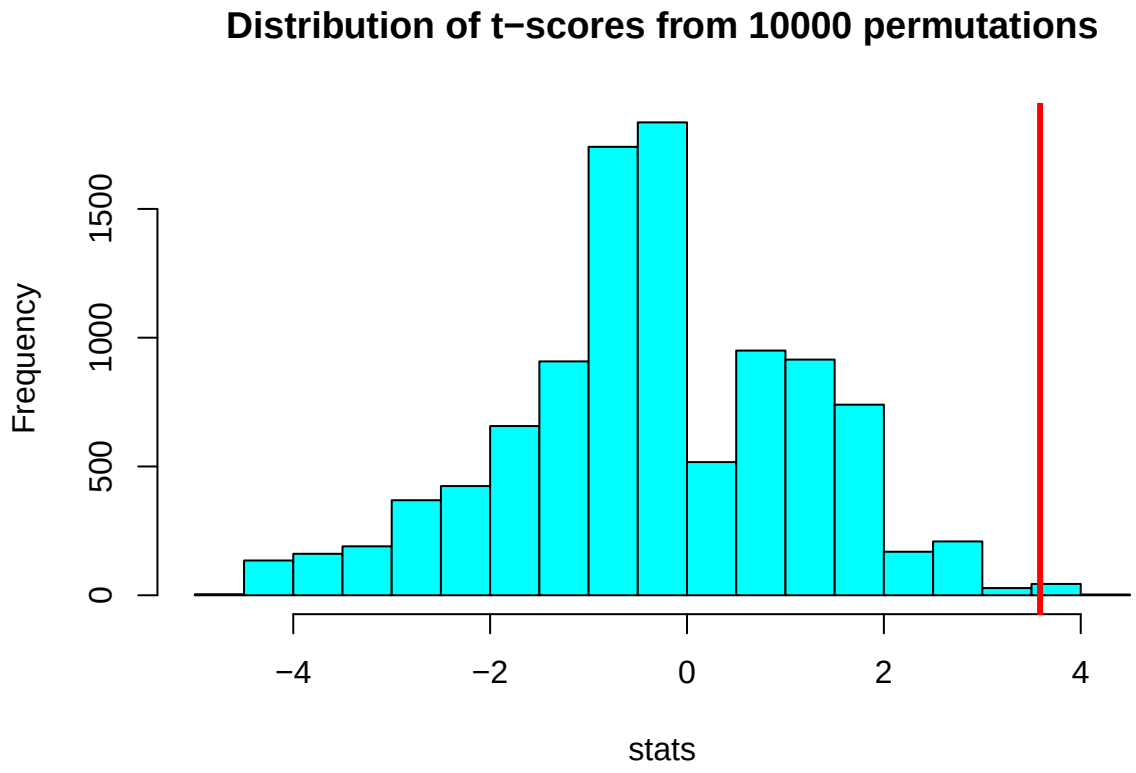
```
stats2 = rep(0,p)
for (i in 1:p)
{
  tbeers = head(permsample[,i], 25)
  twater = tail(permsample[,i], 18)
  tstat = t.test(tbeers, twater, var.equal = TRUE)
```

```
  stats2[i] <- tstat$statistic[['t']]
}

p_value = sum(stats2 >= original_t)/p
print(paste("observed p_value is : ", p_value))
```

```
## [1] "observed p_value is :  5e-04"
```

As we can see from the above p-value which is less than 0.05. We can confirm that the alternate hypothesis is true and the observed statistic is significant.

```
hist(stats, col = 'cyan', main = "Distribution of t-scores from 10000 permutations")
abline(v=original_t, lwd = 3, col = "red")
```



Part c : Calculating p-value assuming non-directional hypothesis

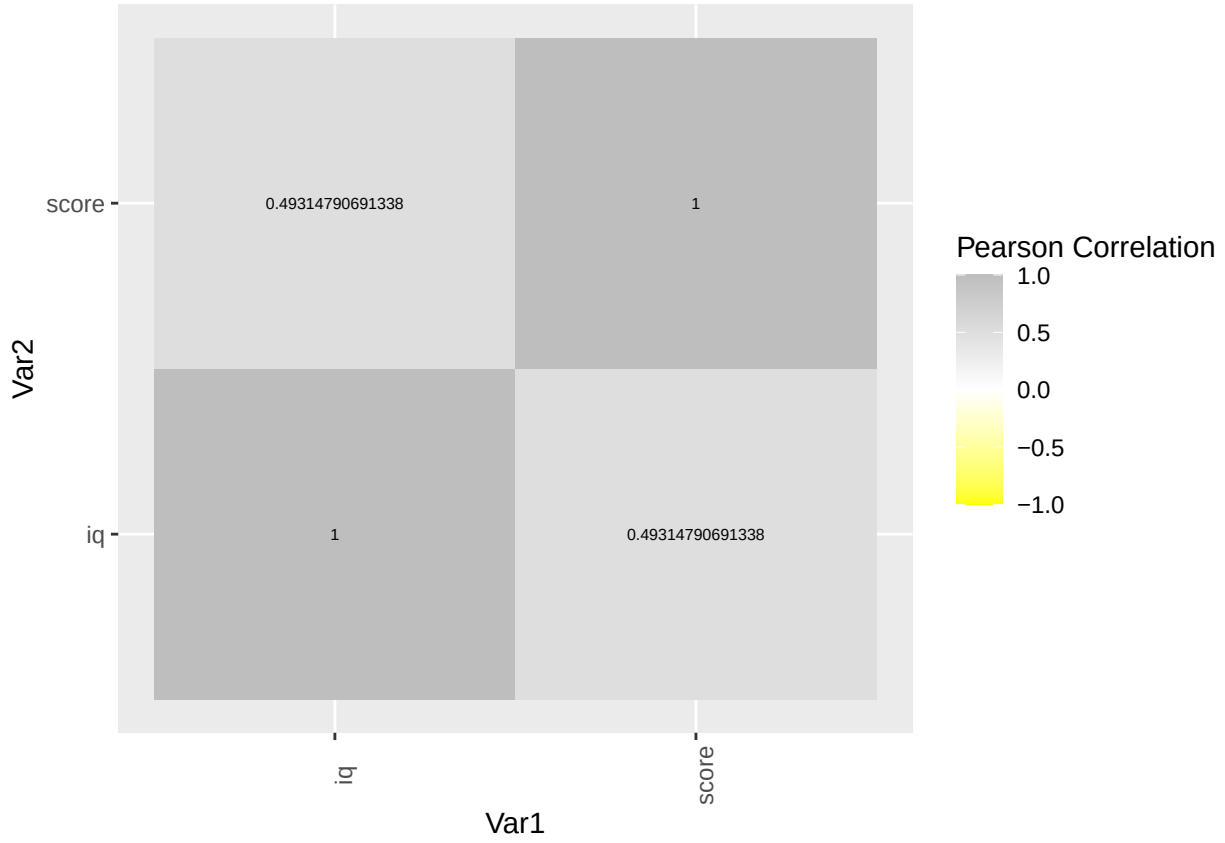
```
p_value = sum(abs(stats) >= diff)/p
print(p_value)
```

```
## [1] 0.0186
```

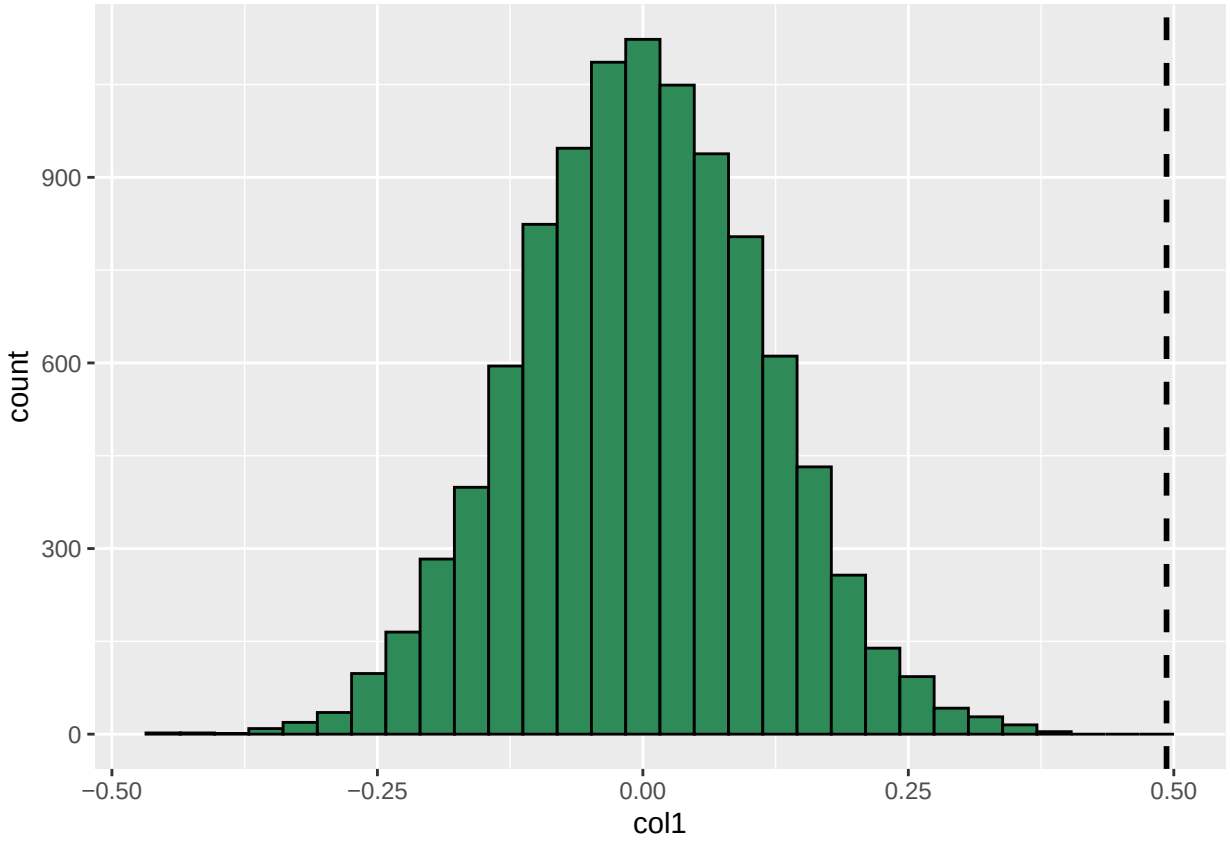
As we can see from the above result, its less than 0.05. So the non-directional hypothesis that, there will be a difference between the group is true and hence the obseved statistics is significant.

Q3. Correlation between IQ and test_scores

```
library(reshape2)
qn3 = read.xlsx("IQ.xlsx", sheetIndex = 1)
iqs = qn3$IQ
tscore = qn3$TESTSCORE

df = data.frame(col1 = c(iqs), col2 = c(tscore))
colnames(df) = c("iq", "score")
corr = cor(df)
corr = melt(corr)

ggplot(data = corr, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  geom_text(aes(label = value), size = 2) +
  scale_fill_gradient2(low = "yellow", high = "gray",
                       limit = c(-1,1), name="Pearson Correlation") +
  theme(axis.text.x = element_text(angle = 90))
```




```
ogcorr = cor(iqs, tscore)
p = 10000
finalstats = rep(0,p)

for ( i in 1:p )
{
  finalstats[i] = cor(iqs, sample(tscore))
}

newdf = data.frame(col1=c(finalstats))
ggplot(newdf, aes(x = col1)) +
  geom_histogram(color = "black", fill="seagreen") +
  geom_vline(xintercept = ogcorr, lwd = 1, lty = 2)
```




```
p_value = sum(finalstats >= ogcorr)/p
print(paste("p-value is : ", p_value))
```

```
## [1] "p-value is :  0"
```

Based on p-value which is almost 0, we can reject the null hypothesis.