Dongyang Wang
Professor Taeb
Stat 528
2/21/2023

HW4

1. The following table contains result for 1.1. It turns out that the coverages are similar but the average confidence interval gets narrower as sample size increases.

n	cover	C_lower	C_upper
50	0.90	-65.54	53.25
100	0.86	-117.72	154.24
200	0.86	-4.11	4.29
400	0.91	-3.30	3.50

The following table contains result for 1.2. It turns out that the coverages are similar but the average confidence interval gets narrower as sample size increases.

n	cover	C_lower	C_upper
50	0.91	-4.32	4.01
100	0.90	-4.08	4.10
200	0.87	-3.95	3.98
400	0.94	-3.82	3.89

For random forest, the average interval gets narrower but not as much in the previous case. Coverage is slightly better than the previous case.

2.

2.1. See the graph attached in the following page.

2.2. As will be detailed in the next page.

2.2a Significant genes under Bonferroni correction: 332 610 1720

2.2b Significant genes under Holm's procedure: 332 610 1720

2.2c Significant genes under FDR control: 2 11 332 364 377 579 610 637 694 698 702 721 735 739 805 905 914 921 1068 1077 1089 1113 1130 1314 1346 1557 1588 1589 1720 2370 2856 2897 2945 3017 3260 3269 3282 3292 3375 3505 3600 3647 3665 3930 3940 3991 4000 4040 4073 4088 4104 4154 4316 4331 4396 4518 4546 4549 4552 4981

2.2e Significant genes under pFDR control: 2 11 292 298 332 364 377 452 579 610 637 694 698 702 721 735 739 805 905 914 921 1068 1077 1089 1113 1130 1314 1346 1491 1557 1588 1589 1647 1659 1720 1966 2370 2856 2897 2945 3017 3208 3260 3269 3282 3292 3375 3505 3600 3647 3665 3930 3940 3991 4000 4040 4073 4088 4104 4154 4316 4331 4396 4492 4496 4515 4518 4546 4549 4552 4981

2.3 The methods are for hypothesis testing, especially multiple testing. The idea is to test whether we should reject our initial assumption, which is called null hypothesis, when compared with an alternative idea we have proposed. In the multiple testing setting, we seek to find out whether we have sufficient evidence in the data to make inference about a group of hypotheses. Basically, we are treating the group as a whole and pick and study each individual hypothesis based on the trend in the entire group.

Based on the evaluations from the two sections, it turns out that the p-values make sense, since they all fall into the range of (0,1). Also, with single tests, we simply need to look at each individual p value and make decisions respectively. However, collectively, we might reject more or less based on which method we choose. If we prefer lower Type I error, we might seek to use the Bonferroni or Holm methods. On the contrary, if we can tolerate some Type I error we gain power by FDR-controlling procedures.

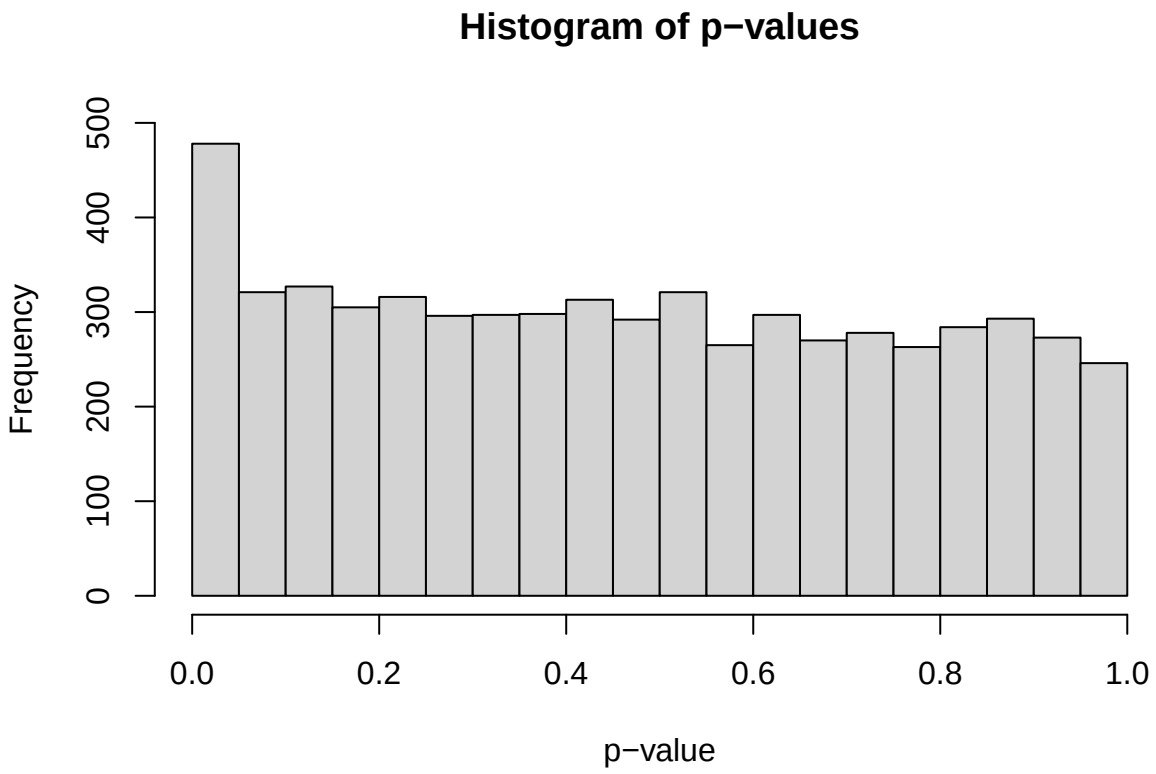
Question 2

Question 2.1

```
# Load data
url <- "https://web.stanford.edu/~hastie/CASI_files/DATA/prostz.txt"
df <- scan(url)
```

```
#calculate p-values
p_value=2*pnorm(abs(df),lower.tail=FALSE)

#plot histogram
hist(p_value,breaks=20,ylim=c(0,500),xlab='p-value',main = "Histogram of p-values",
)
```



Question 2.2

```
#Bonferroni
a=0.05
a_star=0.05/length(p_value)
cat("Significant genes under Bonferroni correction:",
    which(p_value<=a_star))
```

```
## Significant genes under Bonferroni correction: 332 610 1720
```

```
#Holm's procedure
p_value_sorted=sort(p_value,decreasing=F)
threshold=a/(length(p_value)-1:length(p_value)+1)
i0=min(which(p_value_sorted>threshold))
cat("Significant genes under Holm's procedure:",
    which(p_value <= max(p_value_sorted[1:(i0-1)])))
```

```
## Significant genes under Holm's procedure: 332 610 1720
```

```
#FDR control
a=0.1
I=(1:length(p_value))*a/length(p_value)
R=max(which(p_value_sorted<=I))
P_T=p_value_sorted[R]
cat("Significant genes under FDR control:",
    which(p_value<=P_T))
```

```
## Significant genes under FDR control: 2 11 332 364 377 579 610 637 694 698 702 721 735 739 805 905 914
```

```
#if (!require("BiocManager", quietly = TRUE))
#    install.packages("BiocManager")
#BiocManager::install("qvalue")
library(qvalue)
```

```
## Warning: package 'qvalue' was built under R version 4.2.1
```

```
#Storey's q-values
cat("Significant genes under FDR control:",
    which(qvalue(p_value, fdr.level = 0.1)$significant == TRUE))
```

```
## Significant genes under FDR control: 2 11 292 298 332 364 377 452 579 610 637 694 698 702 721 735 739
```

Question 2.3

3.

3.1. Since type I error is defined as the probability of falsely rejecting the true null hypothesis, we can simply reject the ones with p values smaller than 0.05, namely hypotheses 1-3, 8-10.

3.2. For Bonferroni Correction: a* = 0.05/10 = 0.005, so we reject hypotheses 1,9,10.

3.3. Based on the BH procedure, after some calculations in R by setting threshold as q times j divided by 10, 5 hypotheses meet the threshold criteria (are below the threshold). We reject hypotheses 1 3 8 9 10.

3.4. Repeat similar procedure, we can reject a bit more in this case. We reject hypotheses 1 2 3 5 7 8 9 10.

3.5. Since FDR = E(FD/K) <= q, and we have q = 0.2 and K = 8 according to question above. Multiplying gives us 1.6, meaning there can possibly be 2 False Positives.