# 528HW4

Coco_Luo

2023-02-13

## Question 1

Let X be a 50-dimensional random vector that is distributed according to $N(0, I)$, and $y = 3tanh\beta^T X + \epsilon$, we sample coordinates of $\beta$ from a normal distribition $N(0, 1)$ and fix this for the entire analysis.

### 1.1

For 100 independent trials, I generated the pair $(X_{n+1}, y_{n+1})$. Then, I use the training data $\{(X_i, Y_i)\}$ and the test data $X_{n+1}$, and apply the split conformal prediction procedure (with a linear pre- diction model) to obtain a prediction interval for $y_{n+1}$. The empirical probabilities of $y_{n+1}$ belong in this prediction interval over the 100 independent trials are shown below ($\alpha = 0.05$).

Table 1: linear model simulation

| | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| empirical probrobability | 0.95 | 0.94 | 0.96 | 0.97 |
| mean upper bound | -224.02 | -144.99 | -4.73 | -4.09 |
| mean lower bound | 196.18 | 235.77 | 5.07 | 3.89 |

The table shows us that, with linear regression stimulation, the probability that true value in the interval is greater when we have greater sample sizes, and the interval gets smaller when we have smaller sample sizes.

### 1.2

Table 2: Random forest simulation

| | 50 | 100 | 200 | 400 |
|---|---|---|---|---|
| empirical probrobability | 0.93 | 0.92 | 0.91 | 0.96 |
| mean upper bound | -4.85 | -4.65 | -4.57 | -4.20 |
| mean lower bound | 4.55 | 4.56 | 4.33 | 4.46 |

The table shows us that, with random forest model stimulation, the empirical probability of true value is higher than the previous linear regression model. For the linear model, the conformal intervals in linear model are crazy because of overfitting (we are using half of the total sample to fit the model). However, random forest does not have the same problem and it generates more stable intervals.

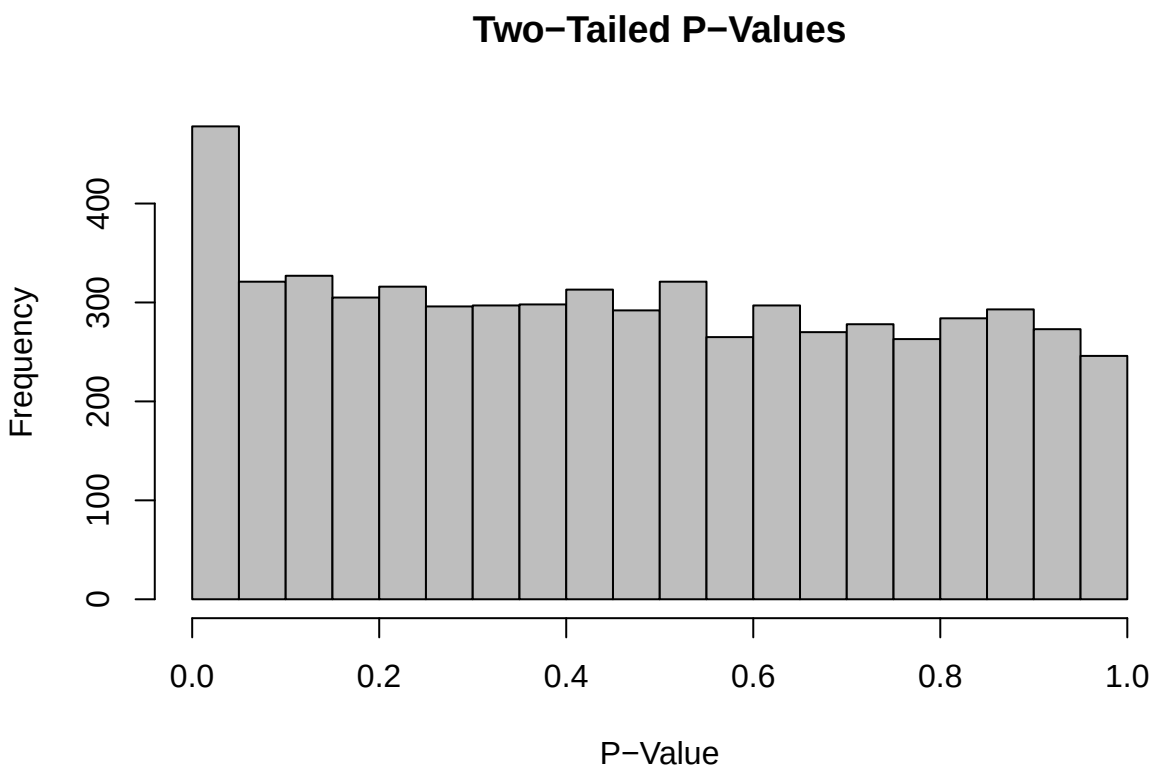
## Question 2

In this question I will analyze data described in Efron and Hastie (2016). In a prostate cancer study, data were collected on 50 controls and 52 patients with cancer, with the genetic activity being measured at m = 6033 genes. The investigators were hoping to spot non-null genes, ones for which patients and controls would

respond differently. The gene.txt file contains 6033 x-scores. Under the null, these statistics follow an N(0, 1) distribution.

**2.1**

First, I calculate the p-values corresponding to the z-scores and generate a histogram of these values.




**2.2**

Next, let's see which genes are flagged as significant under 4 different procedures. For the first one

a). Bonferroni controlling the FWER at 5%

I reject the genes id that has a p value smaller than 0.05/6033.

```
## [1] "Number of significant genes: 3"
```

We reject: 332, 610, 1720

(b) Holm's procedure controlling the FWER at 5%

Following the Holm's procedure, I reject the ids where the sorted p-value is smaller or equal to 0.05/(6033+1-i), where i is the rank number of the sorted p-value. The ids we want to rejected here the same as before, but Holm's procedure usually rejects more than Bonferroni.

```
## We accept when i equals 4
```

We reject: 610, 1720, 332.

(c) Benjamini and Hochberg's FDR control, apply with control at the 10% level

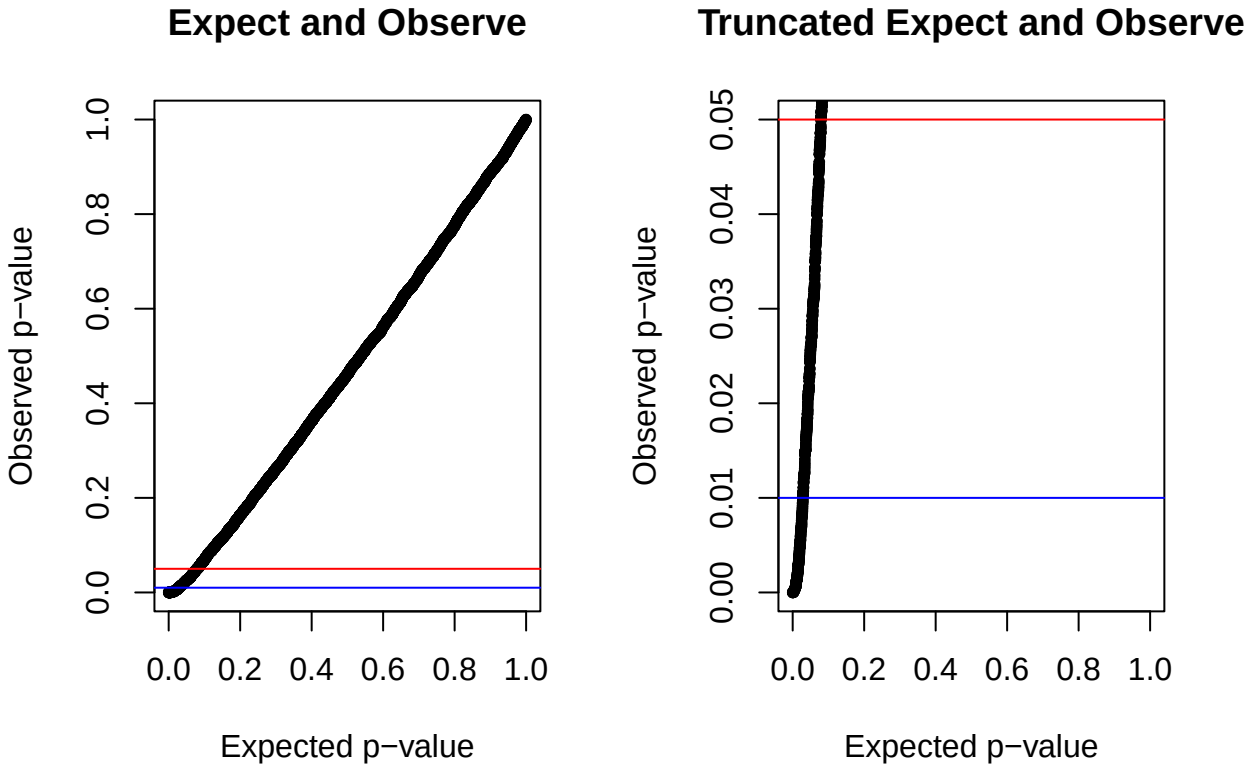
I will find the smallest index of sorted p-value that bigger than (i/6033)*0.1. Thus, we should reject the first 60 p-value of the sorted p-value since 61 is my smallest index.

```
## We accept when i equals 61
```

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

(d) Control the expected number of false discoveries (EFD). Find the genes that correspond to EFD of 1, and then also with 5

Below is a truncated scatter plot for observed p-value in (0, 0.05) since we only need to check the p-value smaller than 0.01 and 0.05.




```
## For EFD of 1, we reject :

##   [1]  610 1720  332  364  914 3940 4546 1068  579 4331 1089 3647 1113 1557 1077
##  [16] 4518 4088 3991 3375 4316 4073 1130 3665  735 4549 1346  921 1589 1314 4981
##  [31] 4104 2897  739  702    2 4000 2370 3282 2856 3600 2945  905  694 3017 4396
##  [46] 4552  721 1588 3292 3930  698 3260 4154   11 3505 4040  377 3269  805  637
##  [61] 4492  292 4515 1659 1491 4496  452  298 1647 1966 3208 3879  718 3200 2968
##  [76]  684 4013 1572 1507 2912 2811 3585 3313  913 2852  478 3242 1329 4378 5159
##  [91] 4671  341 4500 1097  641  742 3961  493 3696  354 3343  995 3917 3922 5287
## [106] 1117 1476 2391 3835  381  987 1643 3746 1628 1090 4282 4499 2923 3712 4554
## [121] 4315 5305   73 1082   78  731 2764 3187  758 1185  348  813  489 4997 4550
## [136] 3793 5242 3265  594 1003 1620  692  476  724 2908 3557 1345  709 3761  680
## [151] 2868 4428 1073 4163 1702  611   98 3567 1019 1908 4134 1566 4538 3590  926
## [166] 2785 4364 3374 2872   44  918 4005

## --------------- total: 172 genes

## For EFD of 5, we reject :

##   [1]  610 1720  332  364  914 3940 4546 1068  579 4331 1089 3647 1113 1557 1077
##  [16] 4518 4088 3991 3375 4316 4073 1130 3665  735 4549 1346  921 1589 1314 4981
##  [31] 4104 2897  739  702    2 4000 2370 3282 2856 3600 2945  905  694 3017 4396
```

```
##  [46] 4552  721 1588 3292 3930  698 3260 4154   11 3505 4040  377 3269  805  637
##  [61] 4492  292 4515 1659 1491 4496  452  298 1647 1966 3208 3879  718 3200 2968
##  [76]  684 4013 1572 1507 2912 2811 3585 3313  913 2852  478 3242 1329 4378 5159
##  [91] 4671  341 4500 1097  641  742 3961  493 3696  354 3343  995 3917 3922 5287
## [106] 1117 1476 2391 3835  381  987 1643 3746 1628 1090 4282 4499 2923 3712 4554
## [121] 4315 5305   73 1082   78  731 2764 3187  758 1185  348  813  489 4997 4550
## [136] 3793 5242 3265  594 1003 1620  692  476  724 2908 3557 1345  709 3761  680
## [151] 2868 4428 1073 4163 1702  611   98 3567 1019 1908 4134 1566 4538 3590  926
## [166] 2785 4364 3374 2872   44  918 4005 1458 4186 2886  676 1751 1223 1193 1810
## [181]  448 4539  449 2941 2864 4530   90 5283 2789 5939 5731 4137  958 4143 4080
## [196] 1050 2993 3848  374 1853 1018  786  276 1918 1228 4405 1492 2967 4652 4371
## [211] 5017 2949 4367 1093 1362 4386   37  696 1508 1729 3804  485  729 2562 1177
## [226] 5257  327 4636  966 2702  490 1158  307 4007   68 2998 3913 3007 1907 1747
## [241]   26 1979 3520 2385 3378 1717  270 3789   35 1034 5784 1651 4965 4541 1226
## [256]  606 1215 1573 2421 4513 1692 5246  398 1356 1584 4905 5407 1736  929  305
## [271] 4508 2935 3308  670  627 2815 3370 1680  737 1504 5647 1995 1604 3366  152
## [286]  725 1975 1174 4112  390 3908 1524 4883 1857 3827 1916 2819 3842  733 5104
## [301] 2442 1822 1169 3826 2376 2801 4031 5336 2621 4023 1376 3432  134 1186 1674
## [316] 1230 2488  677  700 1669  996 4502 1219  460 3192   79  910 2866 5687 5250
## [331] 5158 4110 4996 2134  474 5142  144 4417 3189 4012 2883 5028   55 2211 1598
## [346] 1176 1694 2547 2302 1254 2409 1814 3397  614 1189 3958 5772 2793 3875 2983
## [361]  642 2429 3081 5533 1181 1762 3629 5950  393 2857 3782 3002 4556 2283 2971
## [376] 3498 4115  904 5898   72  423   38 3205 3556 1218 3797  470 1957 3838  212
## [391] 2900  596  126 1081  314   61 2879  285 5625 3048  962 1783   74 5434 4438
## [406]  940 1576 4255 5228 1843 3847  130 1142 1211  828 4412 1213 5686 1847 1227
## [421] 1782 4544  385  928 4295 1880 3671 5606 3801 2918 1637 2881 4511 4792 1275
## [436]  324 3227 2932 5697 3393 4124  667 4494 2026 4120 2833 5951  439   85 1209
## [451] 1012  716 4967  190 4155   81 4692  851 1495 1112  311 2693 4235  494  999
## [466] 1115 3274 4057 2840 1629 1700 1923 3401 1305 1841  577 2745 4169

## --------------- total: 478 genes
```

(e) Storey's q-values with a pFDR threshold of 10%. Also give the estimate of the proportion of null genes $\pi_0$.

Using the package provided above, I get the estimated proportion of null genes $\pi_0 = 0.8545$ and we should reject:

```
## [1] "Number of significant genes:  71"

## Significant genes:

##  [1]    2   11  292  298  332  364  377  452  579  610  637  694  698  702  721
## [16]  735  739  805  905  914  921 1068 1077 1089 1113 1130 1314 1346 1491 1557
## [31] 1588 1589 1647 1659 1720 1966 2370 2856 2897 2945 3017 3208 3260 3269 3282
## [46] 3292 3375 3505 3600 3647 3665 3930 3940 3991 4000 4040 4073 4088 4104 4154
## [61] 4316 4331 4396 4492 4496 4515 4518 4546 4549 4552 4981

## [1] "Estimated proportion of null genes:  0.854468261463924"
```

**2.3**

In this study, I examined the genetic activity data on 52 patients with cancer and 50 controls. Among 6033 genes information, my goals is to find genes that are expressed differently between the groups. The data I use includes z scores which quantify the level of expression of each of the 6033 genes. Using these information, I calculated my p values, which shows us how likely it is that my data would have occurred under my null hypothesis (no differential expression). While performing the multiple hypothesis tests, I used 5 different

methods to control the family-wise error rate and the false discovery rate. These methods help reduce false positive results, which may occur due to randomness.

From the result of the analysis, I found a great amount of genes with statistically significant differential expression between the patient group and the control group. The standard for significance I used differ by the probability of making a false positive error. The Bonferroni and the Holm's procedure control the family-wise error rate and make sure that the probability of making false discovery is very small. But they might also results in a high probability of false negatives, and meanwhile fail to identify true positives. On the other hand, the Benjamini and Hochberg's FDR control and the Storey's q-value method allow for a greater chance of false positives, but also do well on identifying the true positives. No matter which method to use, we should always make wise choice based on our scientific goals and evaluate the trade-off between false positives and false negatives.

## Question 3

Suppose we test m = 10 hypotheses

### 3.1

Suppose that we wish to control the Type I error for each null hypothesis at level $\alpha = 0.05$. I use the Bonferroni method. We should reject all the p-value smaller or equal to 0.005 (0.05/10 = 0.005), thus we should reject $H_{01}$, $H_{09}$, $H_{10}$.

### 3.2

Now suppose that we wish to control the FWER at level $\alpha = 0.05$. I use the Bonferroni-Holm correction. First, we sort the p-value in increasing order. Then, we compare the p-value with 0.05/(11 - i), where i is the rank index. For this question, we start to accept the hypothesis when i is 5, and reject $H_{10}$, $H_{01}$, $H_{09}$, $H_{08}$.

```
## Accept when i equals 5

## Therefore, we reject:

## integer(0)
## [1] 10
## [1] 1
## [1] 9
## [1] 8
```

### 3.3

Now suppose that we wish to control the FDR at level q = 0.05. I use the Benjamini and Hochberg's FDR control. I will now compare the p-value with (i/10)*0.05, where i is the rank of our p-value. We start to accept the hypothesis when i = 6, so we reject 5 hypothesis $H_{10}$, $H_{01}$, $H_{09}$, $H_{08}$, $H_{03}$.

```
## Accept when i equals 6

## Therefore, we reject :

## integer(0)
## [1] 10
## [1] 1
## [1] 9
## [1] 8
## [1] 3
```

### 3.4

Now suppose that we wish to control the FDR at level q = 0.2. Using the same method, I accept when i = 9, so I reject $H_{10}$, $H_{01}$, $H_{09}$, $H_{08}$, $H_{03}$, $H_{02}$, $H_{07}$, $H_{05}$.

```
## Accept when i equals 9

## Therefore, we reject :

## integer(0)
## [1] 10
## [1] 1
## [1] 9
## [1] 8
## [1] 3
## [1] 2
## [1] 7
## [1] 5
```

### 3.5

With q = 0.2, I reject 8 hypothesis, but from these hypothesis, we only have $H_{10}$, $H_{01}$, $H_{09}$, $H_{08}$, $H_{03}$, $H_{02}$ with a p-value $< 0.05$. Thus, our approximately false positives of the null hypothesis rejected at q = 0.2 is $6/8 = 0.75$.