1. Suppose we wish to test $n$ normal means $\mu_i$ from $X_i \sim N(\mu_i, 1)$. One way to do this might be as follows: simulate $n$ i.i.d. random variables $\tilde{X}_i$ from $N(0, 1)$ and select $\tau$ as

$$\tau = \min \left\{ t : \frac{1 + |\tilde{S}(t)|}{1 + |S(t)|} \leq q \right\}, \quad (1)$$

where

$$\tilde{S}(t) = \left\{ i : |\tilde{X}_i| \geq t \text{ and } |\tilde{X}_i| > |X_i| \right\}$$

$$S(t) = \left\{ i : |X_i| \geq t \text{ and } |X_i| > |\tilde{X}_i| \right\}$$

then reject those hypotheses in $S(\tau)$.

(a) Do you expect this to control the FDR at level $q$? Explain why or why not.

Answer

Definitions and Reformulations

Define the following quantities:

$$Y_i := |X_i| \lor |\tilde{X}_i|; \quad Z_i := \begin{cases} 1 & |X_i| \geq |\tilde{X}_i| \\ -1 & \text{Otherwise} \end{cases}$$

$$W_i = Y_i Z_i; \quad B_i = 1 \{ Z_i = 1 \}$$

Then, we see that the following are equivalent definitions of $S(t)$ and $\tilde{S}(t)$:

$$S(t) = \left\{ i : W_i \geq t \right\}; \quad \tilde{S}(t) := \left\{ i : W_i \leq -t \right\}$$

For the order statistics $Y_{(1)} \leq \ldots \leq Y_{(n)}$, let $Z_{[1]}, \ldots, Z_{[n]}$ be the corresponding values of $Z_i$. Correspondingly, $B_{[i]} = 1 \{ Z_{[i]} = 1 \}$

Consider the expression $\frac{1 + |\tilde{S}(t)|}{1 + |S(t)|}$. We see that it is a piecewise-constant function that changes values at $t = Y_{(i)}$ for each $i$. Therefore, $\tau$ must be equal to the smallest $Y_{(j)}$ such that $\frac{1 + |\tilde{S}(Y_{(j)})|}{1 + |S(Y_{(j)})|} \leq q$. In addition, by the same piecewise-nature:

$$|\tilde{S}(Y_{(j)})| = \sum_{k=j}^{n} (1 - B_{[k]}) \text{ and } |S(Y_{(j)})| = \sum_{k=j}^{n} B_{[k]}$$
Making this insight precise, define the following:

\[
J = \min \left\{ j : \frac{1 + \sum_{k=j}^{n} (1 - B_k)}{1 \lor \sum_{k=j}^{n} B_k} \leq q \right\}
\]

Then, \( \tau = Y_{(J)} \).

**Main Proof**

Now, consider the FDP,

\[
FDP = \frac{\# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 1 \}}{1 \lor \# \{ j : Y_j \geq Y_{(J)}, B_j = 1 \}}
\]

\[
= \frac{\# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 1 \}}{1 + \# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 0 \} + \frac{1}{1 \lor \# \{ j : Y_j \geq Y_{(J)}, B_j = 0 \}} \leq \frac{\# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 1 \}}{1 + \# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 0 \} + \frac{1 + \sum_{k=j}^{n} (1 - B_k)}{1 \lor \sum_{k=j}^{n} B_k} \leq \frac{\# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 1 \}}{1 + \# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 0 \} \times q}
\]

where the last inequality follows by the construction of \( J \).

We now examine the expectation of the first term,

\[
E \left[ \frac{\# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 1 \}}{1 \lor \# \{ j : Y_j \geq Y_{(J)}, B_j = 1 \}} \right] = E \left[ \frac{1 + \# \{ j \text{ null} : Y_j \geq Y_{(J)} \}}{1 + \# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 0 \} - 1 \right]
\]

Define \( K := \# \{ j \text{ null} : Y_j \geq Y_{(J)} \} \). Then, we have:

\[
E \left[ \frac{\# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 1 \}}{1 \lor \# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 0 \} \left| K \right] \right] = E \left[ \frac{1 + K}{1 + \# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 0 \} \left| K \right] \right] - 1
\]

For any null \( j \), by symmetry, it is equally likely that \( B_j = 1 \) or \( 0 \). Therefore, we see that:

\[
\left\{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 0 \right\} \sim Binom \left( K, \frac{1}{2} \right)
\]

We bound the value of the conditional expectation:
\[ E \left[ \begin{array}{c} \frac{1 + K}{1 + \# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 0 \}} \\ 1 + \# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 1 \} \end{array} \right] K \]

\[ = \sum_{k=0}^{K} \binom{K}{k} \frac{1}{2^K} \frac{1 + K}{1 + k} = \frac{1}{2^K} \sum_{k=0}^{K} \binom{K + 1}{k + 1} \]

\[ = \frac{1}{2^K} \sum_{k=1}^{K+1} \binom{K + 1}{k} \leq \frac{1}{2^K} \sum_{k=0}^{K+1} \binom{K + 1}{k} = \frac{2^{K+1}}{2^K} = 2 \]

Therefore, we have:

\[ E \left[ \begin{array}{c} \# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 1 \} \\ 1 + \# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 0 \} \end{array} \right] \leq E [2] - 1 \leq 1 \]

Taking the expectation of the FDP equation, we then obtain:

\[ FDR = E [FDP] \leq E \left[ \frac{\# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 1 \}}{1 + \# \{ j \text{ null} : Y_j \geq Y_{(J)}, B_j = 0 \}} \times q \right] \leq q \]

Therefore, we see that this procedure controls the FDR at level \( q \).

(b) Take \( n = 1000 \) and simulate the FDR and power of the above method in the following two settings: (1) 80% of the \( X_i \)'s are N(0,1) and 20% are N(5,1). (2) 80% of the \( X_i \)'s are N(0,1) and 20% are N(2,1). Compare FDR and power with BHq and comment on your findings.

**Answer** Table 1 and 2 shows average FDP and power for different \( q \) over \( N = 1000 \) simulations. Procedure for knockoff is at end. Both methods control FDR in all situations. Knockoff has less power than BHq at all levels. Figure 1 plots histograms of FDP and power when \( q = 0.2 \), and weak signal (\( X_i \sim i.i.d. \text{N}(2,1) \) for alternatives). Knockoff method has longer tail and larger standard deviation, which can be because of randomness when constructing knockoffs.

**Table 1: Comparing power and FDR of knockoff and BHq** Average FDP and power over \( N = 1000 \) simulations. Total number of hypothesis is \( n = 1000 \), 80% are nulls, from N(0,1), alternatives are from N(2,1) (weak signal). Both methods control FDR, knockoff is more powerful at medium level.

<table>
<thead>
<tr>
<th>( q )</th>
<th>0.01</th>
<th>0.05</th>
<th>0.1</th>
<th>0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FDR</strong></td>
<td>BHq</td>
<td>0.007</td>
<td>0.04</td>
<td>0.079</td>
</tr>
<tr>
<td></td>
<td>knockoff</td>
<td>0</td>
<td>0.031</td>
<td>0.084</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>BHq</td>
<td>0.02</td>
<td>0.106</td>
<td>0.199</td>
</tr>
<tr>
<td></td>
<td>knockoff</td>
<td>0</td>
<td>0.06</td>
<td>0.167</td>
</tr>
</tbody>
</table>
Table 2: Comparing power and FDR of knockoff and BHq

Average FDP and power over \( N = 1000 \) simulations. Total number of hypothesis is \( n = 1000 \), \( 80\% \) are nulls, from \( N(0, 1) \), alternatives are from \( N(5, 1) \) (strong signal). Both methods control FDR, knockoff is less powerful than BHq at all levels.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>0.01</th>
<th>0.05</th>
<th>0.1</th>
<th>0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDR</td>
<td>BHq</td>
<td>0.008</td>
<td>0.040</td>
<td>0.080</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>knockoff</td>
<td>0.007</td>
<td>0.048</td>
<td>0.099</td>
<td>0.199</td>
</tr>
<tr>
<td>Power</td>
<td>BHq</td>
<td>0.971</td>
<td>0.992</td>
<td>0.996</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>knockoff</td>
<td>0.912</td>
<td>0.985</td>
<td>0.992</td>
<td>0.996</td>
</tr>
</tbody>
</table>

(e)  Same as (b) but with 95% of nulls instead.

**Answer**  See table 3 and 4 and figure 2 for results. Both methods control FDR at all levels, BHq has higher power at all levels. From figure 2, it occurs more often for knockoffs to make no discovery. When it does make a discovery, it has higher power than BHq procedure.

**Knockoff procedure:**

1. Construct knockoffs: \( X_i \overset{i.i.d.}{\sim} N(0, 1), i = 1, \ldots, n, n = 1000 \).
2. Compute \( W_j = \max(|X_j|, |\tilde{X}_j|) \times \text{sign}(|X_j| - |\tilde{X}_j|) \).
3. For all \( t = |W_j| \), compute \( \tilde{S}(t) \) and \( S(t) \).
   \[
   S(t) = \{ j : W_j \geq t \} \\
   \tilde{S}(t) = \{ j : W_j < -t \}
   \]
4. Find the smallest \( |W_j| \) such that \( \frac{1+|\tilde{S}(t)|}{1+|S(t)|} \leq q \), set it as threshold \( \tau \).
5. Reject all the hypothesis in \( S(\tau) \).
6. Compute number of nulls and non-nulls that are rejected, calculate
   \[
   \text{FDP} = \frac{\text{number of nulls rejected}}{\text{total number of rejection}},
   \]
   \[
   \text{power} = \frac{\text{number of non nulls rejected}}{\text{total number of non-nulls}}.
   \]

Code for this problem is attached at end.
Figure 1: Histogram of power and FDR for knockoff and BHq

Histogram of $N = 1000$ simulations. Number of total hypothesis is $n = 1000$, $80\%$ are nulls. Nulls are $X_i \overset{i.i.d.}{\sim} N(0, 1)$ and alternative are $N(2, 1)$ (weak signal). $q = 0.2$. Both methods control FDR, knockoff method has higher power than BHq, but its power has larger variation, so is its empirical FDP, which can be due to randomness when generating knockoffs.

2.

Suppose I wish to test whether a given regression coefficient in a logistic regression is null or not.

(a) Explain you would compute p-values classically (for instance, explain how R computes p-values) Are these classical p-values truely uniform under the null? If not, when are they approximately uniform?

Answer One method to compute p-values is to use normal approximation. Assume all estimated parameters are jointly normal: $\hat{\beta} \sim N(\beta, \Sigma)$, where we estimate $\hat{\Sigma}$ by $\hat{\Sigma} = E\left[-\frac{\partial^2 l}{\partial \beta^2}\right]_{\hat{\beta}}$, compute p-value assuming each estimated parameter is normally distributed. They may not be truely uniform under the null if $p$ is large, for fixed $p$, as $n \to \infty$, these
Table 3: Comparing power and FDR of knockoff and BHq
Average FDP and power over N = 1000 simulations. Total number of hypothesis is n = 1000, 95% are nulls, from N(0, 1), alternatives are from N(2, 1) (weak signal). Both methods control FDR, knockoff is more powerful at medium level.

<table>
<thead>
<tr>
<th>q</th>
<th>0.01</th>
<th>0.05</th>
<th>0.1</th>
<th>0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDR</td>
<td>BHq</td>
<td>0.009</td>
<td>0.054</td>
<td>0.093</td>
</tr>
<tr>
<td></td>
<td>knockoff</td>
<td>0</td>
<td>0.009</td>
<td>0.033</td>
</tr>
<tr>
<td>Power</td>
<td>BHq</td>
<td>0.01</td>
<td>0.004</td>
<td>0.071</td>
</tr>
<tr>
<td></td>
<td>knockoff</td>
<td>0</td>
<td>0.0009</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Table 4: Comparing power and FDR of knockoff and BHq
Average FDP and power over N = 1000 simulations. Total number of hypothesis is n = 1000, 95% are nulls, from N(0, 1), alternatives are from N(5, 1) (strong signal). Both methods control FDR, knockoff is less powerful than BHq at all levels.

<table>
<thead>
<tr>
<th>q</th>
<th>0.01</th>
<th>0.05</th>
<th>0.1</th>
<th>0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>FDR</td>
<td>BHq</td>
<td>0.0097</td>
<td>0.048</td>
<td>0.093</td>
</tr>
<tr>
<td></td>
<td>knockoff</td>
<td>0</td>
<td>0.041</td>
<td>0.09</td>
</tr>
<tr>
<td>Power</td>
<td>BHq</td>
<td>0.932</td>
<td>0.976</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>knockoff</td>
<td>0</td>
<td>0.942</td>
<td>0.97</td>
</tr>
</tbody>
</table>

p-values will be uniform. In this problem the X_i’s are independent, so the p-values will be asymptotically uniform distributed because the \( \hat{\beta} \)'s are independent.

(b) Imagine we have n samples of the form \((X_i, Y_i)\), where \(X_i \sim N(0, I_p)\) and \(Y_i = \pm 1\) with probability 1/2, independently from \(X\). That is, the covariates are i.i.d. \(N(0, 1)\) and we are under the global null. For \(n = 150\) and \(p = 500\), plot histograms of the empirical distribution of the p-values. Comment on your findings.

Answer Figure 3 plots histogram of empirical distribution of p-values. We have more p-values near 0. Figure 4 plots qqplot against \((0, 1)\) line which shows that the p-values are not uniform and tend to be smaller.

(c) Repeat (b) for various values of the ratio \(p/n\). What do you get when \(p/n > 1/2\)? Can you explain this phenomenon?

Answer Figure shows histogram of various \(p/n\) values keeping \(n = 1500\) fixed. If we use a qqplot, even when \(p/n = 0.1\), our p-values are smaller than they should be if uniform \((0, 1)\). When \(p/n > 0.5\), we sometimes can get perfect separation and algorithm to maximize likelihood does not converge.

(d) What would happen if you were to use classical p-values for multiple testing when \(p/n = 0.3\), say?
**Answer**  If we use these p-values, we would have higher false discovery rate than estimated, so we would not be able to control FDR at the desired level if we assume these p-values are uniform (0,1).

(e) We are in the setup of (b) but now have a fraction of non-zero regression coefficients. Does the distribution of a null p-value seem to change? Explain your answer.

**Answer**  Figure 6 shows qqplot of logistic regression when there are non-zero coefficients. Fix $n = 1500$ and $p = 500$, $n_1$ is number of non-nulls.

\[ Y_i \overset{ind}{\sim} \text{Ber}(p_i), \quad p_i = \text{logistic}(x_i^T \beta + z_i), \]

$\beta$ and $z_i$ are i.i.d. $N(0,1)$. p-values are all smaller than uniform distribution, more so compared to global null when $n_1 = 2$ or $n_1 = 10$. When $n_1 = 15$ there’s also perfect separation in the one simulation I did.
Consider the proof from Lecture 10, Section 3, and recall the ordering $x \geq y$ iff $x_i \geq y_i$ for all $i$. Show that if $f : \mathbb{R}^d \to \mathbb{R}$ is a non-decreasing function meaning that $x \geq y \implies f(x) \geq f(y)$, then for each $t_1 \leq t'_1$,

$$
E[f(X)|X_1 = t_1] \leq E[f(X)|X_1 = t'_1].
$$

Answer Recall result from lecture 13, write

$$
X = \begin{pmatrix} X_1 \\ X_{(-1)} \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_{(-1)} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{1,-1} \\ \Sigma_{-1,1} & \Sigma_{-1,-1} \end{pmatrix},
$$

$X \sim N(\mu, \Sigma)$, then conditional on $X_1 = x$,

$$
X_{(-1)}|X_1 = x \sim N(\mu_x, \Sigma'),
$$

$\mu_x = \mu_{(-1)} + \Sigma_{-1,1} \Sigma_{11}^{-1}(x - \mu_1), \quad \Sigma' = \Sigma_{(-1),(-1)} - \Sigma_{(-1),1} \Sigma_{11}^{-1} \Sigma_{1,(-1)}$

$\Sigma'$ does not depend on $x$. Write $f(x) = f(x_1, x_{(-1)})$ for simplicity,

$$
E[f(x, x_{(-1)})|X_1 = x] = \int f(x, x_{(-1)}) \frac{1}{(2\pi)^{d-1} |\Sigma'|^{1/2}} e^{-1/2(x_{(-1)}-\mu_x)^T\Sigma'^{-1}(x_{(-1)}-\mu_x)} dx_{(-1)}
$$

$$
= \int f(x, x_{(-1)}) + (\mu_x - \mu_{x'})(2\pi)^{d-1} |\Sigma'|^{1/2} e^{-1/2(x_{(-1)}-\mu_x)^T\Sigma'^{-1}(x_{(-1)}-\mu_x)} dx_{(-1)}
$$

$$
\leq \int f(x', x_{(-1)}) \frac{1}{(2\pi)^{d-1} |\Sigma'|^{1/2}} e^{-1/2(x_{(-1)}-\mu_x)^T\Sigma'^{-1}(x_{(-1)}-\mu_x)} dx_{(-1)}
$$

$$
= E[f(x', x_{(-1)})|X_1 = x'],
$$

the inequality is because $f(x)$ a non-decreasing function. Therefore

$$
E[f(X)|X_1 = t_1] \leq E[f(X)|X_1 = t'_1].
$$

Acknowledgement

Thanks to Xiaowei, Zhimei, Nikolas for helping me debug results in knockoff simulations and discussing results with me.

RCode for Question 1

```r
#sampled nulls
x.n<-rnorm(n0,0,1)
x.1<-rnorm(n1,5,1)
#my samples
x<-c(x.n,x.1)

#compute knockoffs
#for independent case, just need to sample i.i.d from normal N(0,1)
```

8
# knockoffs
x.k <- rnorm(n, 0, 1)

# compute test statistics W for each hypothesis
temp <- cbind(x, x.k)
W <- apply(abs(temp), 1, max) * (abs(x) - abs(x.k)) # test statistics as defined

# compute S, \tilde{S} for all thresholds i can use
# order all the statistics
W.sorted <- sort(abs(W), decreasing = FALSE)
# compute threshold value
thresh <- numeric(n)
for (i in 1:n)
  s.t <- sum(W <= (-1) * W.sorted[i])
  s <- sum(W >= W.sorted[i])
  thresh[i] <- (1 + s.t) / max(1, s)

# threshold
tt <- W.sorted[min(which(thresh <= q))]
# rejection set
rej.k <- which(W >= tt)

References

Figure 2: Histogram of power and FDR for knockoff and BHq. Histogram of \( N = 1000 \) simulations. Number of total hypothesis is \( n = 1000 \), 95% are nulls. Nulls are \( X_i \overset{i.i.d.}{\sim} N(0, 1) \) and alternative are \( N(2, 1) \) (weak signal). \( q = 0.2 \). Both methods control FDR, BHq has higher power. From histogram of power, it's more often for knockoffs to discover nothing. When knockoff does make discovery, it seems to have higher power on average than BHq.
Figure 3: Histogram logistic regression p-values. $Y_i$ are i.i.d. Ber(1/2), $p = 500$, $X_i$ are i.i.d. N(0, 1) and number of observations in $n = 1500$. The p-values are not uniform distributed, they tend to be smaller.

Figure 4: QQplot of logistic regression p-values. $Y_i$ are i.i.d. Ber(1/2), $p = 500$, $X_i$ are i.i.d. N(0, 1) and number of observations in $n = 1500$ (see figure 3). The line is (0,1) line. The p-values are smaller than uniformly distributed.
Figure 5: Histogram of logistic regression p-values. Setup is the same as in figure 3, fix $n = 1500$ and change proportion of $p/n$. Even when $p/n = 0.1$, the p-values are smaller than uniform distributed. When $p/n = 0.6$ maximum likelihood algorithm does not converge, indicating there can be a hyperplane that perfectly separates the two classes.
Figure 6: QQplot of logistic regression p-values for variables with zero coefficients. Line is (0,1) line. $p = 500$, $X_i \overset{i.i.d.}{\sim} N(0,1)$, $n_1$ is number of non-nulls. $Y_i \overset{ind}{\sim} \text{Ber}(p_i)$ where $p_i = \text{logistic}(x_i^T \beta + z_i)$, $\beta$ and $z_i$ are i.i.d. $N(0,1)$. p-values are all smaller than uniform distribution, more so compared to global null when $n_1 = 2$ or $n_1 = 10$. When $n_1 = 15$ there’s also perfect separation in the one simulation I did.