Homework #2
Stat 212A, Fall 2015: Topics in Selective Inference
Instructor: Will Fithian
Assigned Oct. 19, 2015. Due 11:59pm Nov. 5, 2015

You are welcome to work with each other or consult articles or textbooks online, but you should then go away and write up the problem by yourself. If you collaborate or use other resources, please list your collaborators and cite the resources you used. Please show your work and include code where appropriate.

You can turn in the problem set in class Nov. 5 or under my door (Evans 301) Thursday night.

1. FDR Control for Arbitrary Dependence Consider the usual setup with m p-values $p_1, \dots, p_m$ and null hypotheses $H_i : p_i \sim U[0,1]$, true for $i \in I_0$.

In class we stated, but did not prove, the theorem from Benjamini and Yekutieli (2001) that under *any* dependendency structure among the p-values, the BH procedure at level α controls FDR at

$$\text{FDR} \leq \alpha \sum_{k=1}^{m} \frac{1}{k}$$

In this problem we will prove the theorem. The key to the proof is the quantity $\pi_{i,j,r}$, defined as

$$\pi_{i,j,r} = \mathbb{P}\left(R^i = r,\ \frac{(j-1)\alpha}{m} \leq p_i \leq \frac{j\alpha}{m}\right).$$

Recall the definition of R^i as the number of hypotheses BH would reject if we replaced p_i with 0, so that $R^i \perp\!\!\!\perp y_i$. Also recall our lemma that

$$R = r,\ i \in S \iff R^i = r,\ p_i \leq \frac{\alpha r}{m}$$

(a) Show that

$$\text{FDR} = \sum_{i \in I_0} \sum_{r=1}^{m} r^{-1} \sum_{j=1}^{r} \pi_{i,j,r}$$

Solution: The result follows from our usual decomposition:

$$\begin{aligned}\text{FDR} &= \sum_{i \in I_0} \sum_{r=1}^{m} r^{-1} \mathbb{P}(i \in S, R = r) \\ &= \sum_{i \in I_0} \sum_{r=1}^{m} r^{-1} \mathbb{P}(R^i = r, p_i \leq \alpha r/m),\end{aligned}$$

with the additional observation that

$$\left\{R^i = r, p_i \leq \frac{\alpha r}{m}\right\} = \bigcup_{j=1}^{r} \left\{R^i = r, \frac{\alpha(j-1)}{m} < p_i \leq \frac{\alpha j}{m}\right\},$$

a disjoint union.

(b) Continuing, show that we can exchange the indices j and r to obtain

$$\text{FDR} \leq \sum_{i \in I_0} \sum_{j=1}^{m} j^{-1} \sum_{r=j}^{m} \pi_{i,j,r}$$

(**Note:** be sure to justify replacing r^{-1} with j^{-1}.)

Solution: We can rewrite the sum from part (a) as a sum over $m(m+1)/2$ terms, with $j^{-1} \geq r^{-1}$ in each term:

$$\begin{aligned}\text{FDR} &= \sum_{i \in I_0} \sum_{1 \leq j \leq r \leq m} r^{-1} \pi_{i,j,r} \\ &\leq \sum_{i \in I_0} \sum_{1 \leq j \leq r \leq m} j^{-1} \pi_{i,j,r} \\ &= \sum_{i \in I_0} \sum_{j=1}^{m} j^{-1} \sum_{r=j}^{m} \pi_{i,j,r}.\end{aligned}$$

(c) Show that $\sum_{r=1}^{m} \pi_{i,j,r} \leq \alpha/m$ and use this to complete the proof.

Solution: The events

$$A_{i,j,r} = \left\{ R^i = r, \frac{\alpha(j-1)}{m} < p_i \leq \frac{\alpha j}{m} \right\}$$

are disjoint for different r or different j, and

$$\begin{aligned}\bigcup_{r=1}^{m} A_{i,j,r} &= \left\{ R^i > 0, \frac{\alpha(j-1)}{m} < p_i \leq \frac{\alpha j}{m} \right\} \\ &= \left\{ \frac{\alpha(j-1)}{m} < p_i \leq \frac{\alpha j}{m} \right\},\end{aligned}$$

which has probability α/m. The second equality comes from the fact that R^i must always be at least 1 (after replacing p_i with 0, BH would at least reject H_i).

As a result,

$$\sum_{r=j}^{m} \pi_{i,j,r} \leq \sum_{r=1}^{m} \pi_{i,j,r} = \alpha/m.$$

Plugging this into the last expression in part (b) and summing over i gives

$$\text{FDR} \leq \frac{\alpha m_0}{m} \sum_{j=1}^{m} j^{-1}$$

2. BY Intervals for Marginal Screening For the construction of the Benjamini-Yekutieli (BY) intervals in class, we defined

$$R^i_{\min} = \min \left\{ |S(y_1, \ldots, y_{i-1}, x, y_{i+1}, \ldots, y_m)| : \; x \in \mathbb{R}, i \in S(y_1, \ldots, y_{i-1}, x, y_{i+1}, \ldots, y_m) \right\}.$$

We showed in particular that for the Benjamini-Hochberg selection procedure, we always have $R^i_{\min} = R$ for every $i \in S(y)$.

A simpler selection algorithm we could use is *marginal screening*: selecting all indices for which $|y_i|$ surpasses some *fixed* threshold t. That is,

$$S(y_1, \ldots, y_m) = \big\{ i \in \{1, \ldots, m\} : \; |y_i| > t \big\}.$$

(a) Show that for this selection procedure, we also have $R^i_{\min} = R$ for every $i \in S(y)$.

Solution: The value of y_i has no effect on the decision to select or not select any other index but i, so we can write

$$\begin{aligned} R^i_{\min} &= \min\left\{ \#\{j \neq i : \ |y_j| > t\} + 1\{|x| > t\} : \ |x| > t \right\} \\ &= 1 + \#\{j \neq i : \ |y_j| > t\}. \end{aligned}$$

If $i \in S$, then $|y_i| > t$, so this last sum is just R.

(b) What is $R^i_{\min} - R$ for the unselected indices $i \notin S(y)$?

Solution: If $i \notin S$ then $|y_i| \leq t$ so the last sum is $R + 1$. Thus, the difference is 1.

3. Convergence of Conditional to Nominal Intervals In class we defined the equal-tailed conditional intervals for the truncated $N(\mu, 1)$ distribution conditional on $|y| > t$ for a fixed value of $t > 0$. If $k_1(\mu)$ and $k_2(\mu)$ are the lower and upper $\alpha/2$ quantiles of the conditional distribution with density

$$f^t_\mu(y) = \frac{e^{-(y-\mu)^2/2}}{\sqrt{2\pi}} \cdot \frac{1\{|y| > t\}}{\mathbb{P}_\mu(|y| > t)},$$

we defined the conditional intervals as

$$C(y; t) = \left[k_2^{-1}(y), k_1^{-1}(y)\right].$$

Consider taking a limit with fixed $t > 0$ and $y \to \infty$. Show that the conditional interval tends to the nominal interval $y \pm z_{\alpha/2}$. That is, prove that

$$\lim_{y\to\infty} \left|k_2^{-1}(y) - (y - z_{\alpha/2})\right| \ = \ \lim_{y\to\infty} \left|k_1^{-1}(y) - (y + z_{\alpha/2})\right| \ = \ 0.$$

Solution: Define the conditional CDF

$$F^t_\mu(x) = \mathbb{P}_\mu(y \leq x \mid |y| > t)$$

$$\begin{aligned} F^t_\mu(x) &= \mathbb{P}_\mu(y \leq x \mid |y| > t) \\ &= 1 - \mathbb{P}_\mu(y > x, |y| > t)/\mathbb{P}_\mu(|y| > t). \end{aligned}$$

Let $\pi^t_\mu = \mathbb{P}_\mu(|y| > t)$, and note that for any fixed t, $\pi^t_\mu \to 1$ as $\mu \to \infty$.

We start by establishing a lemma:

Lemma 1. *As $\mu \to \infty$, $F^t_\mu(x)$ converges uniformly to $\Phi(x - \mu)$:*

$$\lim_{\mu\to\infty} \sup_{x\in\mathbb{R}} |F^t_\mu(x) - \Phi(x - \mu)| = 0$$

Proof. For $x \leq t$, $F^t_\mu(x)$ and $\Phi(x - \mu)$ are both very small:

$$\begin{aligned} \sup_{x\leq t} |F^t_\mu(x) - \Phi(x - \mu)| &\leq \sup_{x\leq t} F^t_\mu(x) + \Phi(x - \mu) \\ &\leq F^t_\mu(t) + \Phi(t - \mu) \\ &\to 0 \quad \text{as } \mu \to \infty. \end{aligned}$$

For $x > t$, we can write

$$\begin{aligned} F^t_\mu(x) - \Phi(x - \mu) &= 1 - \frac{\mathbb{P}_\mu(y > x)}{\pi^t_\mu} - \Phi(x - \mu) \\ &= 1 - \Phi(x - \mu) - \frac{1 - \Phi(x - \mu)}{\pi^t_\mu} \\ &= (1 - \Phi(x - \mu))\left(1 - 1/\pi^t_\mu\right). \end{aligned}$$

As a result,

$$\begin{aligned}\sup_{x>t}|F_\mu^t(x)-\Phi(x-\mu)| &= \sup_{x>t}|1-\Phi(x-\mu)|\cdot|1-1/\pi_\mu^t| \\ &\le |1-1/\pi_\mu^t| \\ &\to 0.\end{aligned}$$

□

As a result, for any fixed $\varepsilon > 0$, we have

$$F^t_{y-z_{\alpha/2}+\varepsilon}(y) \to \Phi(z_{\alpha/2}-\varepsilon) < 1-\alpha/2,$$

implying that $k_2^{-1}(y) < y - z_{\alpha/2} + \varepsilon$ for sufficiently large y. An analogous argument shows that, for sufficiently large y, $k_2^{-1}(y) > y - z_{\alpha/2} - \varepsilon$, and that

$$y + z_{\alpha/2} - \varepsilon < k_1^{-1}(y) < y + z_{\alpha/2} + \varepsilon$$

4. Computing Conditional Intervals For the threshold $t = 4$, compute the equal-tailed conditional interval for $y = 4.001, 4.1, 5$, and 10 for $y \sim N(\mu, 1)$ given $|y| > t$. Show your code.

Hint: if you can implement $k_1(\mu)$ and $k_2(\mu)$ as functions then you can invert them via numerical root-finding.

Solution: See R code below:

```
> qtruncnorm <- function(p,mu,threshold) {
+   p.left <- pnorm(-threshold-mu)                    #P_mu(y   < -t)
+   p.right <- pnorm(threshold-mu,lower.tail=FALSE) #P_mu(y   >  t)
+   p.tot <- p.left + p.right                         #P_mu(|y| >  t)
+
+   ifelse(p <= p.left / p.tot,                       #if quantile in left lobe
+          qnorm(p*p.tot)+mu,                         # use left lobe formula
+          qnorm((1-p)*p.tot,lower.tail=FALSE)+mu)  # use right lobe formula
+ }
>
> alpha <- .05
> threshold <- 4
> mu.grid <- seq(-5,15,by=0.001)
> k1.grid <- qtruncnorm(p=alpha/2, mu=mu.grid, threshold=4)
> k2.grid <- qtruncnorm(p=1-alpha/2, mu=mu.grid, threshold=4)
>
> interval <- function(y) {
+     c(mu.grid[which.min((k2.grid-y)^2)],
+       mu.grid[which.min((k1.grid-y)^2)])
+ }
>
> interval(4.001)
[1] -0.433  0.453
> interval(4.1)
[1] -0.377  5.147
> interval(5)
[1] 1.068 6.933
> interval(10)
[1]  8.04 11.96
```

5. A Bit of Philosophy (**Note:** Graded for completion only; write as little or as much as you want, but write something.)

In class we have discussed the contrast between the "full model" viewpoint and the "submodel viewpoint" of PoSI. Come up with a concrete applied example (other than the "coefficient of IQ on salary" example I gave in class) where you think the submodel viewpoint seems more appropriate, and another concrete applied example where you think the full model viewpoint seems more appropriate. Here, "appropriate" really means "aligned with the actual scientific goals of the analysis."

Solution: For the submodel viewpoint, I would point to a "messy" biological problem such as a microarray experiment: say we observe some phenotype y_i for patient $i = 1, \ldots, n$, as well as expression levels x_{ij} of gene $j = 1, \ldots, 20,000$. Perhaps n is on the order of 1000. In this case, the full model is almost certainly badly misspecified, and even if it were not, we would not expect the "true" coefficients to actually be sparse. Thus, it is not really clear what it would mean to say that gene j's expression level to be correlated with the response y "adjusting for" linear effects for all 20K other genes, even if we had enough statistical power to estimate the partial correlation.

The full model viewpoint might be more appropriate in analyzing a large experiment in which the treatment was assigned at random, independently of all other predictor variables, and regression adjustment is simply used as a means of improving the precision with which we can estimate the treatment effect. In that case, it seems justifiable to believe that the causal treatment effect is best approximated by the regression that adjusts for as many confounders as possible.

References

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.