Stat 27850/30850: Problem set 2

1. In class we considered a modified BH where we:

   (1) estimate the proportion of nulls π_0 with $\widehat{\pi}_0 = \frac{\#\text{ p-values} > \gamma}{n(1-\gamma)}$, then

   (2) run the BH with this estimate of π_0: find the largest k such that there are k many p-values $\leq \frac{\alpha k}{\widehat{\pi}_0 n}$ (but truncating at γ, i.e. never rejecting any p-value $P_i > \gamma$). Equivalently, we can first define $\widetilde{P}_i = P_i \cdot \widehat{\pi}_0$ and compare the $\widetilde{P}_i$'s against $\alpha \cdot \frac{k}{n}$.

   In this problem, consider a related setting: suppose that your n hypotheses are grouped into G many groups, and (before looking at the data) you have reason to believe that the proportion of nulls might be very different from one group to another. Suppose that the group sizes are $n_1, \ldots, n_G$ (with $n_1 + \cdots + n_G = n$). Let's try a modified procedure:

   (1) For each group $g = 1, \ldots, G$, estimate $\widehat{\pi}_0^g = \frac{\#\text{ p-values in group } g \text{ which are} > \gamma}{n_g(1-\gamma)}$

   (2) Now run a group-adaptive BH by defining $\widetilde{P}_i = P_i \cdot \widehat{\pi}_0^g$ for each p-value P_i in group g, for each group g; then compare the $\widetilde{P}_i$'s against $\alpha \cdot \frac{k}{n}$. Make sure to truncate at γ, i.e. to never reject any p-values which are $> \gamma$ (a simple way to do this is to replace $\widetilde{P}_i$ with $+\infty$ for any i with $P_i > \gamma$).

   In this problem you will run simulations to test the FDR control and power of this type of procedure.

   (a) You should start by writing a function

   ```
   group_adaptive_BH = function(P,group_sizes,alpha,gamma)
   ```

   which takes a vector of p-values `P` in $[0,1]^n$, a vector `group_sizes` which specifies $n_1, n_2, \ldots, n_K$, a level `alpha`, and runs your procedure (splitting the p-values into groups by placing the first n_1 of them into group 1, the next n_2 into group 2, etc).
   Note that if you set `group_sizes` to be a vector with a single entry n, then you are actually running the method described in class where we produce a single estimate $\widehat{\pi}_0$ and proceed with the modified BH.

   (b) As a base case, try the following: take $n = 1000$ and $n_1 = \cdots = n_{20} = 50$. Suppose that 10 groups are all null while 10 groups are 50% null. Set $\alpha = 0.1$. Generate the nulls as Unif$[0,1]$ and the signals as p-values from a z-test where the z-scores are $N(2,1)$. Run this for many iterations. What is the average FDP and the average power (proportion of signals which were discovered)? Compare the group-adaptive BH to the "single-group" BH (i.e. the modified BH method with a single $\widehat{\pi}_0$ estimate). You can use $\gamma = 0.5$ to estimate the π_0^g's / to estimate π_0. 20 groups

   (c) Now try different settings to examine the performance of this method across different scenarios, and how it compares to single-group BH. Some questions you may want to explore: What is the effect of non-uniform groups, i.e. groups where the proportion of signals and/or the signal strength is varied from one group to another? What is the effect of group size for both power and FDR control—in particular, for smaller groups, are the $\widehat{\pi}_0^g$'s reliable enough for FDR control? What about changing γ?
   You can also try conservative estimates of the $\widehat{\pi}_0^g$'s,

   $$\widehat{\pi}_0^g = \frac{1 + (\#\text{ p-values in group } g \text{ which are} > \gamma)}{n_g(1-\gamma)}.$$

   How does this affect power and FDR, as compared to the usual estimates without adding 1 in the numerator, and how does the difference between these two estimates change as you change the number of groups / group size?
   You do not need to address every single question and idea listed here, but should develop a thoughtful empirical exploration of at least 2-3 questions (either those suggested here or other ideas you may have). Write a report of your findings explaining the different questions and directions you explored, showing results (probably with plots, e.g. if you are looking at the effect of changing the group size on the resulting FDR then plot FDR against group size), and your conclusions.

2. **Mixture model.** For this problem, suppose that your p-values come from a "two-groups" mixture model: each hypothesis has a chance π_0 of being null, in which case it's drawn from the null distribution Unif$[0,1]$ distribution, and a $(1-\pi_0)$ chance of being a non-null, in which case it's drawn from the signal distribution $\mathcal{D}_\epsilon$, which has a CDF given by $\mathbb{P}\{P_i \leq t\} = t^\epsilon$, where $\epsilon \in (0,1)$ is a signal strength parameter. (Lower ϵ = stronger signal.) Let F be the overall distribution of each p-value (i.e. without knowing ahead of time whether it's null or not). That is,

$$P_i \overset{\text{iid}}{\sim} F = \pi_0 \cdot \text{Unif}[0,1] + (1-\pi_0) \cdot \mathcal{D}_\epsilon.$$

   (a) Write a function for the Bayesian FDR, i.e. the expected FDR for the mixture model, when you set your rejection threshold at t, i.e. you reject any p-value that is $\leq t$, which we will write as

$$\text{BayesFDR}(t) = \mathbb{P}\{P \text{ came from the null distribution} | P \leq t\}.$$

   (Here the probability is taken for P drawn from the mixture model F.) This is a continuous and increasing function of t. To control FDR at the level α we would just find the value of t for which BayesFDR$(t) = \alpha$; call this value $\widehat{t}_{\text{Bayes}}$. Find the value of $\widehat{t}_{\text{Bayes}}$, as a function of π_0, α, and ϵ.

   Note that in practice, the parameters π_0, ϵ would not be known; however for a very large n it would be easy to estimate them with high accuracy by fitting a mixture model (or we could use a Bayesian approach instead), so this Bayesian procedure is something that we could carry out in practice, assuming that we believe the mixture model is a good approximation to the real data.

   (b) Next suppose that you are going to use the BH procedure instead. Recall that, when you're working with p-values, the BH procedure estimates the FDP for a cutoff t as

$$\widehat{\text{FDP}}(t) = \frac{n \cdot t}{\#\text{ p-values} \leq t},$$

   and then takes the largest t such that $\widehat{\text{FDP}}(t) \leq \alpha$. We will look at the "expected" behavior of BH on our mixture model. Assuming that n is very large, the denominator in the expression of $\widehat{\text{FDP}}(t)$ should be very predictable, i.e. very close to its expectation. Assuming that this is the case, what is $\widehat{t}_{\text{BH}}$, the t value for which $\widehat{\text{FDP}}(t) = \alpha$?

   (c) Now suppose that you are going to use Storey's modification of the BH procedure instead. First, we'll have to estimate π_0 using the method from class. Use a cutoff p-value of 0.5. First, write down the expected value of $\widehat{\pi}_0$ (which will differ from the true value π_0).

   (d) Now, Storey's modification of BH will do the following: compute

$$\widehat{\text{FDP}}(t) = \frac{\widehat{\pi}_0 \cdot n \cdot t}{\#\text{ p-values} \leq t},$$

   and then find the largest t such that $\widehat{\text{FDP}}(t) \leq \alpha$. Call this cutoff value $\widehat{t}_{\text{Storey}}$, and reject all p-values $\leq \widehat{t}_{\text{Storey}}$. Now, assuming that n is very large, the denominator in our estimate of $\widehat{\text{FDP}}(t)$ will be extremely close to its expectation, and $\widehat{\pi}_0$ will be close to its expectation. Assuming that this is the case, what is $\widehat{t}_{\text{Storey}}$?

   (e) Now let's compare the three methods numerically. Set $\alpha = 0.2$ and $\pi_0 = 0.7$ and take ϵ over the range $[0.05, 0.5]$. Compute $\widehat{t}_{\text{Bayes}}$, $\widehat{t}_{\text{BH}}$, and $\widehat{t}_{\text{Storey}}$ as a function of these parameters, across the range of ϵ values. In a single plot, show the trajectory of the $\widehat{t}$'s over the range of ϵ values and compare the behavior of the three methods.

   (f) Continuing with the numerical study, compute FDR and power (recall power is the expected proportion of signals which will be discovered) as a function of the rejection threshold t. In a single plot, show the trajectory of FDR for the three methods over the range of ϵ values. In a single plot, show the trajectory of power over the range of ϵ values. Discuss your findings. In particular, how does the achieved FDR compare with the target level α, and how is this affected by ϵ, for the three methods?

3. Write code to simulate the following scenario:

   - $N = 1000$ people each vote for one of $n = 20$ candidates. For any given voter, each candidate $i = 1, \ldots, n$ has probability p_i of being preferred, where $p = (p_1, \ldots, p_n)$ is a probability vector (adds up to 1). Use `rmultinom` to simulate a draw of the vote outcome.
   - For any i, we could build a confidence interval for p_i with the normal approximation to the binomial distribution—the formula is: $\text{CI}_i = \widehat{p}_i \pm z \cdot \sqrt{\frac{\widehat{p}_i(1-\widehat{p}_i)}{N}}$, where z is the critical z-value at threshold α, and $\widehat{p}_i$ is the proportion of voters who chose candidate i.
   - Let $\widehat{i}$ be the winning candidate (in case of ties, just pick any one), and consider the CI for the winner, $\text{CI}_{\widehat{i}}$. Does it actually contain the true value, in this case $p_{\widehat{i}}$? We'd like the CI to be wrong only α of the time.

   Set $\alpha = 0.1$ for this problem.

   (a) What coverage rate do you get when you build the CI at level α? This is the "naive" CI, which does not correct for selection. (Coverage rate = how often the CI actually contains the true value, in this case $p_{\widehat{i}}$, across many repetitions of this procedure—you'll want to run a large number of trials so that we can reliably compare to the target 90% coverage rate).

   Test this question with two settings for the true probability vector:

   $$p = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$$

   and

   $$p = (1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n})/(\text{normalizing constant})$$

   You should get fairly different coverage rates for the naive CI, for these two settings. Describe what you see and if possible, explain why you see this difference.

   (b) What coverage rate do you get when you do a corrected CI as described in class, and build your CI at level α/n? (Note that the number of selected i's is just 1, in this case.) Is this correction too conservative? Try it for both vectors p above.

.