Stats305a Étude 2

Due: Monday, November at 5:00pm on Gradescope.

Note: All data files available at https://web.stanford.edu/class/stats305a/Data/.

Question 2.1: A random variable p is a *p-value* if it is super-uniform, meaning that

$$\mathbb{P}(p \le u) \le u$$

for all $u \in [0, 1]$. (So it is typically larger than a uniform random variable $U \sim \mathsf{Uni}[0, 1]$.) Relatedly, we call a nonnegative random variable $E \ge 0$ an *e-value* (for expected-value) if

$$\mathbb{E}[E] \le 1.$$

We develop analogues of the Benjamini-Hochberg multiple hypothesis testing procedure with e-values, which allow us to provide false discovery rate control with arbitrary dependence.

We begin by first developing a few e-values.

(a) **2 pts.** Let p be a p-value. Show that the following are e-values: (i) $e = \log \frac{1}{p}$, and (ii) $e = \frac{1}{2\sqrt{p}}$. You may use that if Z is a nonnegative random variable, then $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \ge t)dt$.[1]

(b) **2 pts.** Let the typical linear model hold, that is, $Y = X\beta + \varepsilon$ where $\varepsilon \sim \mathsf{N}(0, \sigma^2 I)$ and $X \in \mathbb{R}^{n\times d}$ has rank d. Let $\widehat{\beta} = (X^T X)^{-1} X^T Y$ be the usual estimator of β and $\widehat{\varepsilon} = Y - X\widehat{\beta} = (I - H)\varepsilon$ for $H = X(X^T X)^{-1}X^T$. For $j = 1, \ldots, d$, define the statistics

$$T_j := \frac{\widehat{\beta}_j}{s_n\sqrt{[(X^T X)^{-1}]_{jj}}}, \quad s_n^2 := \frac{1}{n-d}\|\widehat{\varepsilon}\|_2^2.$$

For $m \le \frac{n-d}{4}$, define

$$M_j(m) := T_j^{2m}.$$

Give the largest scalar $c > 0$ you can such that $cM_j(m)$ is an e-value. *Hint.* If A follows an F-distribution with d_1 d.o.f. in the numerator and d_2 in the denominator, then it has moments

$$\mathbb{E}[A^m] = \left(\frac{d_2}{d_1}\right)^m \frac{\Gamma(\frac{d_1}{2} + m)\Gamma(\frac{d_2}{2} - m)}{\Gamma(\frac{d_1}{2})\Gamma(\frac{d_2}{2})} \quad \text{for } m < \frac{d_2}{2}.$$

Now, we start elucidating properties of e-values. First, a simple argument by Markov's inequality shows that they can function as a test statistic for a hypothesis test:

(c) **2 pts.** Consider a test that rejects if an e-value $E \ge \frac{1}{\alpha}$. Show the test has level at most α.

Given a collection $\{p_j\}_{j=1}^N$ of p-values for nulls $\{H_j\}_{j=1}^N$, the Benjamini-Hochberg procedure sorts the p-values into their order statistics $p_{(1)} \le p_{(2)} \le \cdots \le p_{(N)}$ and finds the largest k satisfying

$$p_{(k)} \le \frac{k\alpha}{N}, \tag{BH}$$

then rejects all associated nulls $H_{(i)}$ for $i \le k$. The Benjamini-Yekutieli procedure is a bit more conservative (to allow for dependence among the p-values) and finds the largest k satisfying

$$p_{(k)} \le \frac{k\alpha}{c(N)N} \quad \text{where} \quad c(N) = \sum_{i=1}^N \frac{1}{i} \approx \log N + \frac{1}{2N} + .5772156649. \tag{BY}$$

[1] This follows by a change of variables and Fubini's theorem, as $\mathbb{E}[Z] = \int_0^\infty z dP(z) = \int_0^\infty \int_0^\infty 1\{z \ge t\} dt dP(z) = \int_0^\infty (\int_0^\infty 1\{z \ge t\} dP(z)) dt = \int_0^\infty \mathbb{P}(Z \ge t) dt$.

The analogue of these procedures for the e-value case is the following: given a collection $\{E_j\}_{j=1}^N$ of e-values and associated null hypotheses $\{H_j\}_{j=1}^N$, sort the e-values so that $E_{(1)} \ge E_{(2)} \ge \cdots \ge E_{(N)}$ (note the flipped order of sorting), and find the largest k satisfying

$$E_{(k)} \ge \frac{N}{k\alpha} \tag{EV}$$

then reject the associated nulls $H_{(j)}$ for $j \le k$. A key property of the procedure (EV) is that if $\mathcal{R}$ denotes the set of rejected hypotheses and $R = \mathrm{card}(\mathcal{R})$, then any rejected hypothesis j satisfies

$$E_j \ge \frac{N}{R\alpha}.$$

Let $\mathcal{N}$ denote the collection of true nulls in a multiple hypothesis test, and define the False Discovery Proportion by $\mathrm{FDP} := \frac{\mathrm{card}(\mathcal{N}\cap\mathcal{R})}{\max\{R,1\}}$ and the False Discovery Rate $\mathrm{FDR} := \mathbb{E}[\mathrm{FDP}]$.

(d) **2 pts.** Justify each of the following string of equalities and inequalities:

$$\frac{\mathrm{card}(\mathcal{N}\cap\mathcal{R})}{\max\{R,1\}} \stackrel{(i)}{=} \sum_{j\in\mathcal{N}} \frac{1\{j\in\mathcal{R}\}}{\max\{R,1\}} \stackrel{(ii)}{\le} \sum_{j\in\mathcal{N}} \frac{1\{j\in\mathcal{R}\}}{\max\{R,1\}} \cdot \frac{R\alpha E_j}{N} \stackrel{(iii)}{\le} \frac{\alpha}{N} \sum_{j\in\mathcal{N}} E_j.$$

(e) **2 pts.** Show that the procedure with rejection threshold (EV) satisfies $\mathrm{FDR} \le \frac{\mathrm{card}(\mathcal{N})}{N}\alpha$.

(f) **10 pts.** We come to a numerical comparison between the testing procedures: (i) the Bonferroni correction (union bound) that rejects p values when $p_j \le \frac{\alpha}{N}$, (ii) the Benjamini-Yekutieli corrected procedure (BY), and (iii) the e-value procedure with rejections (EV) using the e-values from part (b) for $m = 1, 2, 4, 8, 16$. (A log-gamma function may be useful.)

Perform the following experiment 1000 times with sample size $n = 900$ and dimension $d = 30$:

i. Construct a design $X \in \mathbb{R}^{n\times d}$ with i.i.d. $\mathsf{N}(0,1)$ entries, and set $\beta \in \mathbb{R}^d$ to have its first 10 entries $\mathsf{N}(0, .01)$ and the last 20 to be zero.
ii. Sample $Y = X\beta + \varepsilon$, where $\varepsilon \sim \mathsf{N}(0, I_n)$.
iii. Run each of the procedures enumerated above for nulls $\{H_j : \beta_j = 0\}$, $j = 1, \ldots, d$.
iv. For each procedure, record the FDP and the number of rejected hypotheses.

For each procedure, report a histogram (across the 1000 experiments) of the FDPs and the number of rejected hypotheses at level $\alpha = .1$. Explain (in a few sentences) your results.