Stats305a Etude 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
\[ \mathbb{P}(p \leq u) \leq u \]
for all $u \in [0, 1]$. (So it is typically larger than a uniform random variable $U \sim \text{Uni}[0, 1]$.) Relatedly, we call a nonnegative random variable $E \geq 0$ an $e$-value (for expected-value) if
\[ \mathbb{E}[E] \leq 1. \]

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) $e = \frac{1}{\sqrt{p}}$.

You may use that if $Z$ is a nonnegative random variable, then $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \geq t)dt$.\footnote{This follows by a change of variables and Fubini’s theorem, as $\mathbb{E}[Z] = \int_0^\infty zdP(z) = \int_0^\infty \int_0^\infty 1(z \geq t)dtdP(z) = \int_0^\infty \int_0^\infty 1(z \geq t)dP(z)dt = \int_0^\infty \mathbb{P}(Z \geq t)dt.$}

(b) 2 pts. Let the typical linear model hold, that is, $Y = X\beta + \varepsilon$ where $\varepsilon \sim \text{N}(0, \sigma^2 I)$ and $X \in \mathbb{R}^{n \times d}$ has rank $d$. Let $\hat{\beta} = (X^TX)^{-1}X^TY$ be the usual estimator of $\beta$ and $\hat{\varepsilon} = Y - X\hat{\beta} = (I - H)\varepsilon$ for $H = X(X^TX)^{-1}X^T$. For $j = 1, \ldots, d$, define the statistics
\[ T_j := \frac{\hat{\beta}_j}{s_n \sqrt{[(X^TX)^{-1}]*_{jj}}}, \quad s_n^2 := \frac{1}{n-d}\|\hat{\varepsilon}\|^2_2. \]

For $m \leq \frac{n-d}{4}$, 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
\[ \mathbb{E}[A^m] = \left(\frac{d_2}{d_1}\right)^m \frac{\Gamma(d_1/2 + m)\Gamma(d_2/2 - m)}{\Gamma(d_1/2)\Gamma(d_2/2)} \quad \text{for } m < \frac{d_2}{2}. \]

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 $E \geq \frac{1}{\alpha}$. Show the test has level at most $\alpha$.

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 $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$ and finds the largest $k$ satisfying
\[ p_{(k)} \leq \frac{k\alpha}{N}, \quad \text{(BH)} \]
then rejects all associated nulls $H_{(i)}$ for $i \leq k$. The Benjamini-Yekutieli procedure is a bit more conservative (to allow for dependence among the $p$-values) and finds the largest $k$ satisfying
\[ p_{(k)} \leq \frac{k\alpha}{c(N)N} \quad \text{where } c(N) = \sum_{i=1}^N \frac{1}{i} \approx \log N + \frac{1}{2N} + 0.5772156649. \quad \text{(BY)} \]
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 \( E_1 \geq E_2 \geq \cdots \geq E_N \) (note the flipped order of sorting), and find the largest \( k \) satisfying

\[
E_k \geq \frac{N}{k\alpha}
\]

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

\[
E_j \geq \frac{N}{R\alpha}.
\]

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

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

\[
\frac{\text{card}(\mathcal{N} \cap R)}{\max\{R,1\}} \overset{(i)}{=} \sum_{j \in \mathcal{N}} \frac{1\{j \in R\}}{\max\{R,1\}} \overset{(ii)}{\leq} \sum_{j \in \mathcal{N}} \frac{1\{j \in R\}}{\max\{R,1\}} \cdot \frac{R\alpha E_j}{N} \overset{(iii)}{\leq} \frac{\alpha}{N} \sum_{j \in \mathcal{N}} E_j.
\]

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

(f) 10 pts. We come to a numerical comparison between the testing procedures: (i) the Bonferroni correction (union bound) that rejects \( p \) values when \( p_j \leq \frac{\alpha}{N} \), (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. \( \mathcal{N}(0,1) \) entries, and set \( \beta \in \mathbb{R}^d \) to have its first 10 entries \( \mathcal{N}(0,.01) \) and the last 20 to be zero.

ii. Sample \( Y = X\beta + \varepsilon \), where \( \varepsilon \sim \mathcal{N}(0,I_n) \).

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.