STA 712 Challenge Assignment 6: Fun with multiple testing!

Due: Wednesday, November 9, 12:00pm (noon) on Canvas.

Instructions:

- Submit your work as a single typed PDF (you should not need to type much, if any, math on this assignment).
- You are welcome to work with others on this assignment, but you must submit your own work.
- You can probably find the answers to many of these questions online. It is ok to use online resources! And using online documentation and examples is a very important part of coding.

Distribution of p-values

Let $X_1^n = X_1, ..., X_n$ be a sample from a continuous distribution, with density function $f(x; \theta)$. Consider testing the null hypothesis $H_0 : \theta = \theta_0$, with test statistic $T(X_1, ..., X_n)$, and rejecting when T is large. The *p-value* for this hypothesis test is given by

$$p = P_{\theta_0}(T(X_1^*, ..., X_n^*) > T(X_1, ..., X_n)),$$

where $X_1^*, ..., X_n^* \sim f(x; \theta_0)$ is a sample under H_0, and P_{θ_0} denotes the probability when $\theta = \theta_0$. In other words, the p-value is the "probability of our data or more extreme", if the null hypothesis were true.

1. Under these conditions, the p-value has a very nice distribution: $p \sim Uniform(0, 1)$ when H_0 is true.
   (a) Argue that $p = 1 - F_T(T)$, where F_T is the cumulative distribution function (cdf) of T under H_0.
   (b) Using the fact that F_T is a continuous, monotonic increase function under our assumptions, show that $P_{\theta_0}(p < s) = s$ for any $s \in (0, 1)$. Conclude that $p \sim Uniform(0, 1)$.
   (c) Show that if we reject when $p < \alpha$, then the type I error of our test is α.

Multiple hypothesis testing

2. Suppose we now have m samples $X_1^{n_1}, ..., X_1^{n_m}$, from distributions with parameters $\theta_1, ..., \theta_m$ respectively. For each sample i, we test the hypothesis $H_0 : \theta_i = \theta_{i,0}$.
   (a) The *family-wise error rate* (FWER) is the probability of making at least one type I error in our m tests. Suppose all our tests are independent, H_0 is true for all the tests, and for each test we reject H_0 when $p < \alpha$. What is the family-wise error rate?
   (b) Clearly, rejecting each test when $p < \alpha$ does *not* control the FWER at level α. The *Bonferroni method* is a simple and popular method for controlling the FWER by changing the p-value threshold. When testing m hypotheses, the Bonferroni method rejects for each test when $p < \dfrac{\alpha}{m}$.

Using the union bound,

$$P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i),$$

show that the Bonferroni method controls the FWER at level α.

(c) Simulate $m = 100$ samples from some continuous distribution, and test some null hypothesis H_0 for each sample. Simulate your data so that H_0 is true for every sample. Using the Bonferroni correction to control the FWER at level $\alpha = 0.05$, do you reject H_0 for any of the tests?

(d) Repeat part (c) 1000 times; for each repetition, record whether you rejected H_0 for any of the tests. In what fraction of your 1000 repetitions do you reject H_0 for at least one test?

Multiple pairwise comparisons

3. Now suppose we have k different groups we want to compare, and let $\mu_1, ..., \mu_k$ denote the means of each group. We are interested in all pairwise comparisons of these means: that is, we test $H_0 : \mu_i = \mu_j$ for every $i \neq j$. We want to control the FWER across all our pairwise comparisons.

   We observe a sample $Y_{i,1}, ..., Y_{i,n}$ of size n from each group $i = 1, ..., k$ (note we are assuming the same sample size for every group). Let $\overline{Y}_i = \frac{1}{n} \sum\limits_{j=1}^{n} Y_{i,j}$ be the sample mean for group i, and let $s_i^2 = \frac{1}{n-1} \sum\limits_{j=1}^{n} (Y_{i,j} - \overline{Y}_i)^2$ be the sample variance for group i. Assuming the true variance for each group is the same, the *pooled sample variance* is then

   $$s_p^2 = \frac{1}{k} \sum_{i=1}^{k} s_i^2.$$

   We reject $H_0 : \mu_i = \mu_j$ when

   $$q_{ij} = \frac{|\overline{Y}_i - \overline{Y}_j|}{s_p \sqrt{2/n}}$$

   is large.

   With k groups, we perform $\binom{k}{2}$ pairwise tests. One option for controlling the FWER is to test each hypothesis with a two-sample t-test, and