STA 712 Challenge Assignment 6: Fun with multiple testing!

Due: Thursday, November 10, 12:00pm (noon) on Canvas.

Instructions:

- Submit your work as a single PDF. You may include scanned handwritten work.
- 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.

Introduction

In class, we have briefly discussed some properties of p-values, and problems with simultaneously testing multiple hypotheses. One situation in which multiple testing occurs is when we want to do simultaneous pairwise comparisons between many different groups. For example, on HW you worked with the SFN data, which had several different forums. We can test for *any* difference between forums using a nested test, but knowing there is *some* difference between forums doesn't tell us *which* forums are different. So, we need to compare each pair of forums, but this requires many hypothesis tests. As you will see, testing many hypotheses leads to problems with type I errors.

On this challenge assignment, you will further explore p-values, type I errors, and the family-wise error rate; show that the Bonferroni correction controls the family-wise error rate, but is conservative; and learn about Tukey's range test as an alternative method for simultaneous pairwise comparisons.

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 increasing 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 $X_{1,1}^{n_1}, ..., X_{m,1}^{n_m}$, 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? (Your answer should be a function of α and m).

   (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 $p < \frac{\alpha}{m}$.

   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. To compare all means simultaneously, we could use an ANOVA F-test to test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$. But what if we want to know *which* means are different?

   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 $\overline{Y}_i = \frac{1}{n}\sum_{j=1}^{n} Y_{i,j}$ be the sample mean for group i, and let $s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_{i,j} - \overline{Y}_i)^2$ 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

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

is large.

(a) Suppose we use a two-sample t-test for each pairwise test, rejecting when $t_{ij} > t_{2n-2,1-\frac{\alpha}{2}}$ (the $1-\alpha/2$ quantile of a t distribution with $2n-2$ degrees of freedom). Conduct a simulation with $k = 10$, in which $H_0 : \mu_i = \mu_j$ is true for all i, j, to show that we fail to control the FWER error rate at level α.

(b) With k groups, we perform $\binom{k}{2}$ pairwise tests. One option for controlling the FWER is to use the Bonferroni method with our two-sample t-tests from part (a), rejecting when

$$t_{ij} > t_{2n-2,1-\frac{\alpha}{2\binom{k}{2}}}.$$

Using your simulation from (a), show that the Bonferroni method controls the FWER at level α.

4. While the Bonferroni correction works, it is *conservative* when not all null hypotheses are true. This means that we lose power to detect real differences. An alternative to the Bonferroni correction when comparing multiple group means is *Tukey's honest significance difference (HSD)* test, also called *Tukey's range test.*

Rather than use a t-distribution to test the pairwise hypotheses $H_0 : \mu_i = \mu_j$, Tukey's HSD uses the *studentized range distribution.* In particular, we reject H_0 when

$$q_{ij} = \frac{|\overline{Y}_i - \overline{Y}_j|}{s_p/\sqrt{n}} > q_{1-\alpha;k;k(n-1)},$$

where $q_{1-\alpha;k;k(n-1)}$ denotes the $1-\alpha$ quantile of the studentized range distribution with k groups and $k(n-1)$ degrees of freedom.

In this question, we will derive the studentized range distribution, then show its use for multiple pairwise comparisons.

(a) Let $X_1, ..., X_k$ be k iid draws from some continuous distribution with density f and cdf F. Let $X_{(1)}, ..., X_{(k)}$ be the order statistics, so $X_{(1)} = \min\{X_1, ..., X_k\}$ and $X_{(k)} = \max\{X_1, ..., X_k\}$. Let $R = X_{(k)} - X_{(1)}$. Argue that the density of R is given by

$$f_R(r; k) = k(k-1) \int_{-\infty}^{\infty} f(u+r) f(u) [F(u+r) - F(u)]^{k-2} du.$$

(b) Now suppose that $X_1, ..., X_k \overset{iid}{\sim} N(0, 1)$, and let $S \sim \chi_\nu$ be independent of the X_i's. (Note that S follows a χ_ν distribution, *not* a χ^2_ν distribution! This just means that $S^2 \sim \chi^2_\nu$.) Let $R = X_{(k)} - X_{(1)}$ and let $Q = R/S$.

i. Using independence, show that the density f_Q of Q is given by

$$f_Q(q; k, \nu) = \int_{-\infty}^{\infty} f_S(s; \nu) \cdot s f_R(sq; k) ds,$$

where f_S is the density of S.

ii. The density of a χ_v distribution is

$$f_S(s;\nu) = \begin{cases} 0 & s < 0 \\ \dfrac{s^{\nu-1}e^{-s^2/2}}{2^{(\nu/2-1)}\Gamma(\nu/2)} & \nu \geq 0. \end{cases}$$

Conclude that

$$f_Q(q;k,\nu) = \frac{k(k-1)}{2^{(\nu/2-1)}\Gamma(\nu/2)} \int_0^\infty s^\nu e^{-s^2/2} \int_{-\infty}^\infty \varphi(u+sq)\varphi(u)\left[\Phi(u+sq)-\Phi(u)\right]^{k-2} du\, ds,$$

where φ is the standard normal density, and Φ is the standard normal CDF.

(c) The distribution with density $f_Q(q;k,\nu)$ is called the *studentized range distribution*, and is parameterized by the number of groups k and the degrees of freedom ν. In R, use the `ptukey` function to plot the cdf $F_Q(q;k,\nu)$ for several different combinations of k and ν.

(d) Now suppose we observe k sample means $\overline{Y}_1, ..., \overline{Y}_k$. Assuming each of the k groups has the same variance σ^2 and the same sample size n, then each $\overline{Y}_i \approx N(\mu_i, \sigma^2/n)$. Let s_p be the pooled sample variance, as above. Then $\dfrac{s_p}{\sigma} \approx \chi_{k(n-1)}$. Using the fact that s_p is independent of the means $\overline{Y}_i$, explain why the density of

$$\frac{\overline{Y}_{(k)} - \overline{Y}_{(1)}}{s_p/\sqrt{n}}$$

should be approximately $f_Q(q;k,k(n-1))$.

(e) Consider the pairwise tests $H_0 : \mu_i = \mu_j$. Argue that rejecting H_0 for each test when

$$q_{ij} = \frac{|\overline{Y}_i - \overline{Y}_j|}{s_p/\sqrt{n}} > q_{1-\alpha;k;k(n-1)}$$

will control the FWER at level α.

(f) Repeat your simulations from 3(b), but this time reject when

$$q_{ij} = \frac{|\overline{Y}_i - \overline{Y}_j|}{s_p/\sqrt{n}} > q_{1-\alpha;k;k(n-1)}.$$

(In R, the `qtukey` function will calculate quantiles of the studentized range distribution). What is the FWER using Tukey's range test?