36-744 Fall 2019

(Some) Statistical Aspects of Reproducibility

The one and only homework (second mini)

Lecturer : Aaditya Ramdas

Question 1 (Online FWER control) *Let's say we have just four p-values: 0.01, 0.02, 0.1, 0.2 that appear one at a time (in the online multiple testing framework), and target error level* $\alpha = 0.1$*. How many rejections does online Bonferroni make? (since it needs a sequence of constants that sum to one, answer the question with constants of* $0.4, 0.3, 0.2, 0.1$*, and with* $0.1, 0.2, 0.3, 0.4$*. Obviously, this sequence cannot be picked in hindsight, so this question is just for the purpose of analysis.) In the same setting, how many rejections would the online fallback procedure have made, where a rejection results in passing all the current wealth to the next step?*

Question 2 (Graphical procedure) *In the offline FWER setting, let's say we have the same four p-values: 0.01, 0.02, 0.1, 0.2, and* $\alpha = 0.1$*. Recall the graphical procedure discussed in class: we start with a directed graph where the outgoing edges from each node sum to 1, and with levels on each node that sum to* α*; we then see if any node's p-value is below its level and if yes, it is rejected, and the graph is updated by deleting the node and passing on its level to other nodes (by an appropriate formula). The graph and initial levels are obviously not allowed to depend on the data, it needs to be fixed beforehand. However, for the purposes of intuition, please design one non-line graph (+levels) would have resulted in the maximum number of rejections, and one non-line shaped graphs (+levels) that would have resulted in the minimum number of rejections.*

Question 3 (Interactive testing) *In class, we saw the use of masked p-values* $g(p) = \min(p, 1-p)$ *and the missing bit* $h(p) = 2 \cdot I(p > 1/2)$*. (a) If the p-values are conservative, provide some reasoning for why the resulting procedure (like STAR) will be less powerful than if the p-values were uniform. Can you design a different masking function* $g(p)$ *that is robust to conservative p-values, and in fact will be more powerful than if the p-values were uniform? Recall that the condition needed for FDR control was weaker than independence, it was that for null p-values,* $E[h(p)|g(p)] \geq 1$*. (b) If the p-values are uniform, but we change* $h(p) = I(p > \lambda)/(1-\lambda)$*, then re-design* $g(p)$ *so that* $E[h(p)|g(p)] \geq 1$ *still holds.*

Question 4 (Naive permutations do not suffice) *In class, we saw that model-X knockoffs was one way to perform rigorous variable selection (conditional independence testing of* $H_j : X_j \perp Y | X_{-j}$*) when a model for* P_X *is known. Design a simulation experiment to* *demonstrate that the naive permutation method does not control FDR (or FWER) for variable selection. The "naive permutation method" is the following: given an $n \times 1$ vector y and an $n \times p$ matrix X, calculate a sensible test statistic (score) for each variable that would be high if the variable were important; then randomly permute y and recalculate the p scores, and repeat this process B times (for some large B); reject H_j if the original score for the j-th variable was larger than $(1-\alpha)$ fraction of the B permuted scores for the j-th variable. What null hypotheses is this method actually testing?*

Question 5 (Anytime p-values) *Design a simulation experiment to convince yourself that sequential tests with naive batch p-values can heavily inflate the type-1 error. For example, first pick the favorite batch hypothesis test of your choice (MMD two sample test, distance covariance independence test, t-test, chi-squared test, etc) — perhaps pick one with a simple computationally efficient rejection rule in the batch setting. Draw data under an appropriate null hypothesis, and apply that test's rule at time $t = T_{min} \ldots, T_{max}$ (where $T_{min} \geq 2$ and T_{max} has to be finite for a simulation). Repeat this B times. Plot time t on the x axis and the empirical type-1 error up to time t on the y-axis (number of false rejections before t divided by B), which should be an increasing curve. How high can it get? Do you know how to run this test correctly in a sequential manner? (is it an open problem?)*