Statistics of Big Data, Fall 2023-4

Class notes 11

The QPD schema (Aharoni et al. 2011), (Rosset et al. 2014)

This is a schema to build and maintain a database for scientific research, assuring:

- Statistical validity of all the results being generated on the database
- Usefulness (=maintaining power) for later users

The main idea is to increase the data size as the database is used. In this way we can maintain power even if the levels decrease in α spending!

Assume at time t we have data of size n_{t-1} and remaining α pool of $\alpha \cdot q^{n_{t-1}}$ for given some fixed parameter $q < 1$, say $q = 0.999$. Then the tth test arrives, meaning a scientist has in mind:

- A pair of hypotheses about some parameter θ:

$$H_{0t} : \theta = \theta_0 \ \ , \ \ H_{At} : \theta = \theta_A.$$

  Note this encodes what the test statistic is (through Neyman-Pearson or monotone likelihood), and also what the effect size they think they will find is through θ_A.

- A desired power π_t

At this point we find c_t such that:

$$\alpha_t = \alpha \cdot q^{n_{t-1}}(1 - q^{c_t}),$$

is an appropriate level for getting power π_t for the desired test.

Then after we perform the test and get c_t more samples we have $n_t = n_{t-1} + c_t$ samples and remaining pool of:

$$\alpha \cdot q^{n_{t-1}} - \alpha \cdot q^{n_{t-1}}(1 - q^{c_t}) = \alpha \cdot q^{n_{t-1}+c_t}.$$

Hence by definition our pool will never run out.

Example of power calculation with n_t samples and required effect size θ_t, in the simple normal means case, with known variance σ^2:

$$\begin{aligned} \pi(\alpha_t) &= \mathbb{P}_{\theta_t}\left(\bar{X} \geq Z_{1-\alpha_t}\frac{\sigma}{\sqrt{n_t}}\right) = \mathbb{P}_{\theta_t}\left(\frac{\bar{X} - \theta_t}{\frac{\sigma}{\sqrt{n_t}}} \geq \frac{Z_{1-\alpha_t}\frac{\sigma}{\sqrt{n_t}} - \theta_t}{\frac{\sigma}{\sqrt{n_t}}}\right) = \\ &= 1 - \Phi\left(\frac{Z_{1-\alpha_t}\frac{\sigma}{\sqrt{n_t}} - \theta_t}{\frac{\sigma}{\sqrt{n_t}}}\right) \end{aligned}$$

Theorem (Aharoni et al. 2011):
For many families of testing problems, including:

1. Any string of simple tests that use Neyman-Pearson
2. Tests of normal means

and many others, the simple recipe above guarantees that $c_t \leq c_0$ is bounded in the following sense: A test of a specific effect size $\tilde{\theta}$ at a specific required power $\tilde{\pi}$ will never cost more than $c_0(\tilde{\theta}, \tilde{\pi})$ samples at any time t.

In practice, this leads to diminishing and not only bounded costs (this can in fact be proven rigorously based on the same ideas from the original proof).

Important conclusion: If you come later, you will gain power and/or money.

False discoveries vs false non-coverage

Many times the interesting statistical inference is reporting confidence intervals in addition or instead of results of hypothesis testing. Then a similar quantity to FDR is FCR. Assume we choose to declare a (random) group of findings $S \in \{1, ..., K\}$, and for each one report a confidence interval I_i for the parameter θ_i (for example, the effect of the ith SNP in GWAS). Then:

$$FCR = \mathbb{E}\left(\frac{\sum_{i \in S} \mathbb{I}\{\theta_i \notin I_i\}}{|S|}\right).$$

Note the similarity to FDR, since we are seeking to control the % out of reported results that are not accurate: for FDR, falsely rejected, for FCR non-covering.

By analogy to the FDR result, we can show that for FCR, if we choose S in a "reasonable" way (specifically, by employing the BH procedure), and build all confidence intervals at confidence level $1 - \frac{\alpha|S|}{K}$, then we guarantee $FCR \leq \alpha$. Note that taking $S = \{1, \ldots, K\}$ gives in this case regular confidence intervals at level $1 - \alpha$, with the known standard interpretation, that for each parameter θ_i there is $1 - \alpha$ chance that the (random) CI I_i does not cover θ_i.

Selective inference: statistical inference after selection

The FDR/FCR way of thinking gets us closer to explicitly thinking about selective inference. When we publish we usually choose and report a group S of *interesting findings* out of the potential hypotheses we examined. Selective inference explicitly acknowledges that we are not going to report all results, and aims to control the false discovery / non-coverage over the set of reported results. Both FDR and FCR are already solutions to this problem and they control errors **on average in the selected (AoS)**.

We can take a step back and build a hierarchy of control types under selection:

A **Simultaneous over all possible (SoP)**. In this setting, we do not consider our specific selection and want to make sure we control false discovery simultaneously over all possible selection policies for this data. For example:

$$\forall S \in \{1, \ldots, K\}, \; ; \mathbb{P}(V(S) > 0) \leq \alpha.$$

So, we find ourselves back in the familiar FWER territory (since V is maximized when $S = \{1, \ldots, K\}$), and of course Bonferroni is a solution.

B **Simultaneous over all selected (SoS)**. Here we want FWER-type control, but only for the (random) set we select to report:

$$\mathbb{P}(V(S) > 0) \leq \alpha.$$

Can we use Bonferroni with $|S|$ hypotheses? No, because we may have selected the smallest p values to report. For example, assume all nulls are true and we decide in advance to report only the smallest p-value. Then if we do Bonferroni for $|S| = 1$ which is no correction:

$$\mathbb{P}(V(1) = 1) = \mathbb{P}(p_{(1)} \leq \alpha) = 1 - ((1 - \alpha)^K).$$

For $K = 100$ this is 0.99. In fact, if we choose the smallest, we know that we need SoP to control false discovery. On the other hand, if we choose at random, then Bonferroni with $|S|$ will work fine. So designing SoS policies depends on the nature of the selection policy. We will not expand on methods for SoS control, but this is an active area of research.

C **Conditional on selected (CoS)**. This is a delicate definition, that challenges us to think carefully about probabilistic notions. The definition here for hypothesis testing is:

$$\mathbb{P}(V_i = 1 | i \in S) \leq \alpha \ , \ \forall i.$$

This acknowledges that the selection set S is random, and requires that for any hypothesis, conditional on its selection to report as "discovery", the probability it is false is at most α. By integrating this gives us:

$$\mathbb{E}(V|R) \leq \alpha R \ \Rightarrow \ \mathbb{E}(V/R|R) \leq \alpha,$$

and by integrating again over R:

$$\Rightarrow \ \mathbb{E}(V/R) \leq \alpha.$$

So CoS guarantees FDR control on the reported set (but not the other way around!).

D **On Average on selected (AoS)**. This is explicitly what FDR /FCR do. For hypothesis testing:

$$\mathbb{E}(V(S) \mid R(= |S|)) \leq \alpha.$$

As we have just seen CoS $\Rightarrow$ AoS, it is also easy to see that SoS $\Rightarrow$ AoS as long as we keep the $0/0 = 0$ convention.

So overall we have a clear hierarchy of strictness where each definition is more strict and implies the next: $SoP \ \Rightarrow \ SoS \ \Rightarrow \ CoS \ \Rightarrow \ AoS$.

Selective inference and multiple testing problem

Our general setup: we are either testing m hypotheses or building confidence intervals for m parameters. We may select a subset $S \in \{1, ..., m\}$ of them as "interesting".

1. State whether each of these claims is true or false and explain **briefly and clearly**:

(a) Building confidence intervals at the Bonferroni level $1 - \alpha/m$ guarantees FCR control at level α for any subset S.

(b) If we choose a set of rejected hypotheses by the BH procedure at level α, obtaining R rejections, and then build confidence intervals at level $1 - \alpha \cdot R/m$, then the FCR is also controlled at level α.

(c) If we decide to select all m hypotheses as "interesting", then selective inference (i.e., controlling FCR) is reduced to inference "on average", meaning we are controlling the expected percentage of errors of our m hypotheses.

2. Consider Table 1 in the Science paper by Zeggini et al. (link to the original paper can be found the class homepage).

(a) Assume we calculate FCR-corrected confidence intervals for the second to last column of the table. Considering the results from the first part of this problem, and the p values in the table, explain why these FCR-corrected intervals are not expected to cross below 1.

(b) Assume now that we were to take a different approach, collect all the SNPs that were significant in *any* of the participating studies, and declare all of them "selected" (their number can be much bigger than the ten on slide 11), and then build FCR-corrected CI's for them at FCR level α, based on the entire meta analysis (like the last two columns of the table). Do you expect that some of these intervals will cross 1? Explain. What percentage of non-coverage do you expect over these selected? Specifically, do you expect this percentage to be about α, smaller than that, or larger than that? Explain.

3. Assume now that instead of selecting interesting results, we can order our hypotheses a-priori from "the most important" to the least important. We care most about the first hypothesis and least about the last. The following procedure is known as hierarchical testing:

- Test the most important null hypothesis at level α. If not rejected, stop and don't consider the other hypotheses.
- If rejected, continue to the second null hypothesis and test it at level α. If this second hypothesis not rejected, stop.
- Continue until a non-rejected null. Then stop.

(a) Does this procedure control FWER at level 0.05? Prove your answer (a full formal proof is not required, but a clear and correct argument is required).

(b) Given a large number m of hypotheses, consider Bonferroni, α-spending, and this hierarchical approach, all at level α. Can you predict which one would make more rejections? Why yes or why not?