Statistics of Big Data, Fall 2023-24

Class notes 10

Multiple testing review

Reminder: F_θ an unknown distribution with parameter(s) θ, in *hypothesis testing* we choose between:

$$H_0 \ : \ \theta \in \Theta_0 \text{ (usually } \Theta_0 = \{\theta_0\} \text{ a simple null)} \quad \text{vs.} \qquad H_A \ : \ \theta \in \Theta_A.$$

A *statistical test* at level α for this problem given data $\mathbb{X} \sim F_\theta$ (usually i.i.d sample) is defined by a test statistics $S(\mathbb{X})$ and a rejection region C_α such that

$$\mathbb{P}_\theta(S(\mathbb{X}) \in C_\alpha) \leq \alpha \ , \ \ \forall \theta \in \Theta_0.$$

Defining a *p-value*: intuitively a p-value is the strength of evidence ("surprise") against the null hypothesis from the data. In case the null is simple and there is monotone likelihood ratio:

$$\frac{\mathbb{P}_\theta(\mathbb{X})}{\mathbb{P}_{\theta_0}(\mathbb{X})} \searrow \text{ in } S(\mathbb{X}) \ \forall \theta \in \Theta_A,$$

then we can define a p-value as the probability under the null of being above the observed value in the data $S(\mathbb{X})$.

$$p - val = \mathbb{P}_{\theta_0}(S \geq S(\mathbb{X})).$$

The rejection region at level α is simply the rule $p - val \leq \alpha$. Note that under the null the p-value has a $U(0,1)$ distribution (by definition).

Power for a specific alternative is the probability of rejection if the alternative is true

$$\Pi_\theta = \mathbb{P}_\theta(S \in C_\alpha) = \mathbb{P}_\theta(p - val \leq \alpha).$$

Hence α the level of the test is a bound on the type-I error (probability of rejection if null is true).

Hypothesis testing is a major tool in science, and basically the goal is to define a null that is "we discovered nothing" and if we succeed in rejecting the null, we have made a discovery and it usually leads to a scientific publication (and maybe a Nobel prize...). So the goal of the scientist is to reject nulls, but this has to be done while preserving validity and avoiding *false discoveries*.

The setting of interest is when we have not a single hypothesis testing problem, but many testing problems at once:

$$H_{0k} : \theta_k = \theta_{0k} \quad H_{Ak} : \theta_k \in \Theta_{Ak} \quad k = 1, \ldots, K.$$

K can be in the hundreds, thousands or even millions (for example in GWAS).

The critical point in multiple testing scenarios is that we now want to control some measure of overall type-I errors. For example in GWAS, if we have a million hypotheses and we naively control each test at level $\alpha = 0.05$, it means that if all nulls are true (non-genetic phenotype) we expect $0.05 \times 10^6 = 5 \times 10^4$ false discoveries — not an acceptable number!

Denote the total number of rejections out of the K tests by R and the total number of false rejections (type-I errors) by V, then we can think of various notions of overall false discovery:

- **Family-wise error rate** $FWER = \mathbb{P}(V > 0)$ **(closely related:** $\mathbb{E}(V)$ **).** Note $FWER \leq \mathbb{E}(V)$ (why?). FWER is the probability of making even one false discovery in the entire corpus. FWER is controlled at level α by the following simple idea: divide α into K pieces $\alpha_1, \ldots, \alpha_K$ such that $\sum_k \alpha_k \leq \alpha$, then test H_{0k} at level α_k. Then it is trivial to see

$$FWER \leq \mathbb{E}(V) = \sum_k \alpha_k \leq \alpha.$$

  Simplest implementation is *Bonferroni's correction* where $\alpha_k = \frac{\alpha}{K}$.
  Critically this result does not depend on any assumption on the nature of the data and the tests (for example, the test statistics for the different tests can be dependent in any form of dependence), as long as all tests are marginally valid at the desired level α_k

- **False Discovery Rate (FDR).** Controlling FWER might be too conservative in the following sense: If we test $K = 10^6$ hypotheses and find 100 true discoveries, we can perhaps agree to also have a few false ones, because most of our discoveries will be correct. This approach would allow us to be less conservative and make more discoveries $\Rightarrow$ publish more papers and get more Nobel prizes, at the cost of a small number of errors (small = compared to the number of discoveries made).

  How can we be less conservative? The FDR idea is to control $\mathbb{E}(\frac{V}{R}) \leq \alpha$. A main problem is what happens when $R = 0$ (no discoveries). The controversial solution is to assume $0/0 = 0$ and so when we make no discoveries we do not contribute to this error. Why is this problematic? Since the following policy controls FDR at level α: flip a coin with probability α, if it comes out heads reject all K hypotheses, if tails reject none. Then with probability $1 - \alpha$ we have $R = V = 0$ and with probability α we have $V/R = 1$.

  For the case that the test statistics $S_1(\mathbb{X}), \ldots, S_K(\mathbb{X})$ are independent, the most famous policy for controlling FDR is the Benjamini-Hochberg (BH) approach, invented at TAU: given p-values $p_1, \ldots, p_K$, sort them in increasing order:

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(K)},$$

  and define:

$$i^* = \max \left\{ i \ : \ p_{(i)} \leq \frac{\alpha \cdot i}{K} \right\},$$

  and reject all hypotheses $1, \ldots, i^*$. This guarantees FDR control at level α.
  Comments:

  1. The Bonferroni policy is to reject if $p_k \leq \alpha/K$ which is the threshold for $\alpha_{(1)}$ in BH, so clearly BH always rejects more than Bonferroni

2. If all nulls H_{0k} are true, then BH still controls FWER, even though it is more liberal than Bonferroni
3. Dealing with dependence: Benjamini and Yekutieli proved that if all the p-values are positively correlated (makes sense for GWAS for example), then BH is still valid. If not, they showed that the following more conservative policy guarantees FDR control under general dependence: reject hypotheses $1, \ldots, i^*$ where:

$$i^* = \max\left\{ i \ : \ p_{(i)} \leq \frac{\alpha \cdot i}{K \cdot c(K)} \right\}, \ \ c(K) = \sum_{k=1}^{K} \frac{1}{k} \approx \log(K).$$

Because BH makes more discoveries than FWER control (and scientists care about discoveries and publications!) the BH approach has been extremely influential and widely used

In the context of GWAS, the Genetics journals typically require a p-value threshold of 5×10^{-8} for publication, which is essentially a Bonferroni correction for 10^6 tests. FDR based analysis is also very popular (since the positive dependence assumption is usually reasonable), but the top journals usually insist on the conservativeness of FWER control after many past "discoveries" turned out to be false and set the area back many years.

Publication bias

In a big data world, many different scientists have large data sets and they perform research.

Now assume the following simple model of how a scientific community works:

1. A number of different groups (say 20) work independently on the same or very similar problem (say GWAS on the same very important phenotype)
2. They each perform research that may include multiple testing on their data, and carefully correct for multiplicity (say control FWER at level 0.05 using Bonferroni on their K hypotheses)
3. Each group that got a significant result publishes, the ones that don't move on to the next project (do not publish failures)

Now we come next year and we see that one group published a paper with an important discovery. If we know about the other 19 groups, we can calculate the FWER for the entire corpus of research (assuming independence which is reasonable here):

$$\mathbb{P}(\text{any discovery if all nulls true}) = 1 - \mathbb{P}(\text{no discovery}) = 1 - (1 - 0.05)^{20} \approx 0.65,$$

so we conclude that there is not really a convincing discovery here. On the other hand in real life we typically do not know about the unpublished failures, and so we do not know that FWER is 0.65 and not 0.05!

In the modern scientific world, this publication bias issue turns out to be a very substantial concern and some claim that it threatens the validity of the entire paradigm of how research is done. The most famous paper that makes this (and other) points is *Why Most Published Research Is Wrong* by Ionides (2005).

Ideas for solution:

Publish every study, including ones that have negative results
$\Rightarrow$ we can calculate quantities like FWER across studies
However, this is not how scientific publishing works...

Share data: In a modern big data world, perhaps we can encourage joint research by the entire research community instead of 20 separate studies:

1. Collect data jointly to a central joint data repository
2. All research is done on the central repository, perhaps with a system where groups who give more data will be allowed to "use" the data more
3. Control FWER or FDR across all use of the joint resource. Important aspect: If we are serious about sharing data for scientific research, it may not be reasonable to assume that we know in advance how many hypotheses (or even what kind of hypotheses) will be tested on the joint resource.

A simple solution for point (3) is offered by α-*Spending* approaches that control FWER and $\mathbb{E}(V)$ by building an infinite series: $\alpha_1, \alpha_2, \ldots$ such that:

$$\sum_{k=1}^{\infty} \alpha_k \leq \alpha, .$$

and test the kth hypothesis at level α_k. This simple Bonferroni-like approach guarantees $FWER \leq \alpha$ for eternity.
An obvious problem: α_k are getting increasingly tiny, having to decrease quadratically so the series converges. Low level = low power of course.

So the question becomes: How can we implement the ideas in (1)-(3) above in a way that will be fair, useful and encourage the scientists to participate?

The Quality preserving database (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!

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. v

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.

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