Unit 6: Further Topics

Eugene Katsevich

December 1, 2021

Units 1-5 focused on estimation and inference in linear models and generalized linear models. In Unit 6, we explore further topics (the exact topics are still TBD).

1 Multiple testing

In this class, we have talked a lot about hypothesis testing, e.g. testing the significance of a coefficient in a (generalized) linear model. But frequently, there are multiple hypotheses we care about testing; let us denote these null hypotheses by $H_1, \dots, H_m$. After obtaining p-values for each null hypothesis—denote these by $p_1, \dots, p_m$—we may want to answer questions about this entire collection of hypotheses. In particular:

- Global testing: Test the *global null hypothesis* $H_0 : H_1 \cap \cdots \cap H_m$.
- Multiple testing: Find a subset $S \subseteq \{1, \dots, m\}$ of null hypotheses to reject so that the set S satisfies some notion of Type-I error.

We discuss global testing in Section 1.1 and multiple testing in Section 1.2.

1.1 Global testing

Global testing problem setup. Here we want to test whether *any* of the null hypotheses $H_1, \dots, H_m$ is false. For example, suppose that $H_j : \beta_j = 0$, where β_j are the coefficients in a GLM. Then, $H_0 : \beta_1 = \cdots = \beta_m = 0$. We recognize this hypothesis as something we would test using an F-test or, more generally, a likelihood ratio test. Here we are concerned with the more general problem of aggregating m p-values for individual hypotheses (whatever these hypotheses may be) into one p-value (i.e. one test) for the global null. A level-α test $\phi(p_1, \dots, p_m)$ of the global null must satisfy

$$\mathbb{E}_{H_0}[\phi(p_1, \dots, p_m)] \leq \alpha. \tag{1}$$

The multiplicity problem. A naive test would separately test the m hypotheses, and then reject if any are significant:

$$\phi_{\text{naive}}(p_1, \dots, p_m) = \mathbb{1}\left(p_j \leq \alpha \text{ for some } j = 1, \dots, m\right). \tag{2}$$

This test does not control the Type-I error. In fact, assuming the input p-values are independent, we have

$$\mathbb{E}_{H_0}[\phi_{\text{naive}}(p_1, \dots, p_m)] = 1 - (1 - \alpha)^m \to 1 \quad \text{as} \quad m \to \infty. \tag{3}$$

This is an illustration of *the multiplicity problem*: The more hypotheses we test, the more likely one of them is going to appear significant just by chance. This is related to data-snooping and

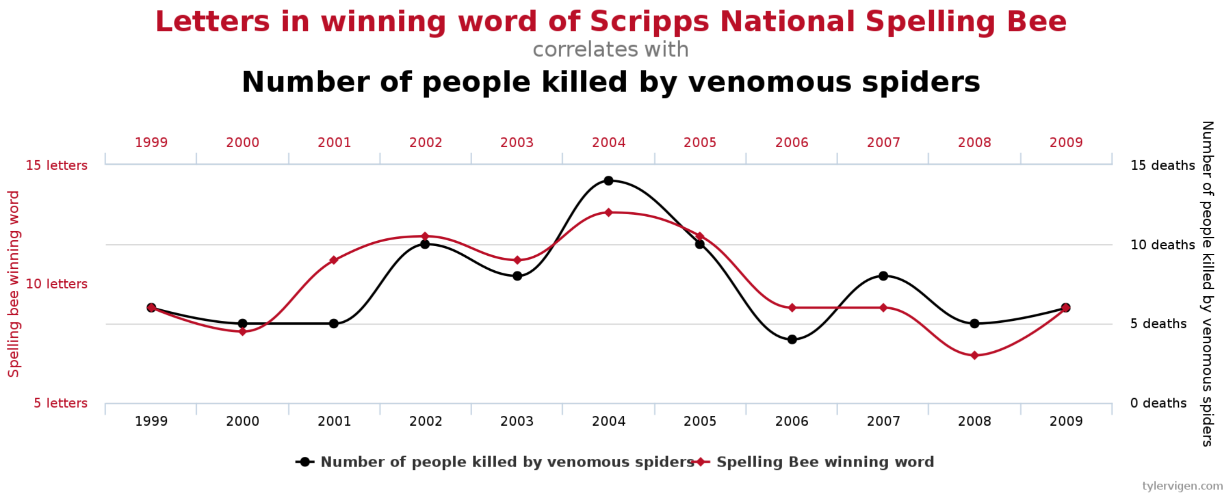

Figure 1: A spurious correlation resulting from data snooping.

the issue of selection bias. If we had chosen just one hypothesis a priori, then we can compare its p-value to the nominal level of α. If we chose the hypothesis by looking ("snooping") at the p-values of m hypotheses and choosing the most significant, we have incurred selection bias that must be corrected for. See Figure 1. There are several ways of properly correcting for this selection bias, i.e. several valid global tests in the sense of definition (1). Here we highlight two:

- Fisher combination test: Powerful against many weak signals.
- Bonferroni test: Powerful against few strong signals.

1.1.1 Fisher combination test

Suppose that $p_1, \ldots, p_m$ are independent (though this is a strong assumption that is often violated). Then, the Fisher combination test is

$$\phi(p_1, \ldots, p_m) \equiv \mathbb{1}\left(-2\sum_{j=1}^{m} \log p_j \geq Q_{1-\alpha}[\chi^2_{2m}]\right). \tag{4}$$

Type-I error control (1) is based on the fact that

$$\text{if } p_1, \ldots, p_m \overset{\text{i.i.d.}}{\sim} U[0,1], \text{ then } -2\sum_{j=1}^{m} \log p_j \sim \chi^2_{2m}. \tag{5}$$

If we have $X_j \sim N(\mu_j, 1)$ and the p-values are defined via $p_j = 2\Phi(-|X_j|)$, then

$$-2\log p_j \approx X_j^2. \tag{6}$$

Therefore,

$$-2\sum_{j=1}^{m} \log p_j \approx \sum_{j=1}^{m} X_j^2. \tag{7}$$

This helps us build intuition for what the Fisher combination test is doing. It's averaging the strengths of the signal across hypotheses.

1.1.2 Bonferroni test

Instead of averaging the signal across p-values, we might want to find the *strongest* signal among the p-values. It makes sense that such a strategy would be powerful against sparse alternatives. We define the Bonferroni test via

$$\phi(p_1, \dots, p_m) \equiv \mathbb{1}\left(\min_{1 \le j \le m} p_j \le \alpha/m\right). \tag{8}$$

The Bonferroni global test rejects if any of the p-values crosses the *multiplicity-adjusted* or *Bonferroni-adjusted* significance threshold of α/m. The more hypotheses we test, the more stringent the significance threshold must be. We can verify the Type-I error control of the Bonferroni test via a union bound:

$$\mathbb{P}_{H_0}\left[\min_{1 \le j \le m} p_j \le \alpha/m\right] \le \sum_{j=1}^{m} \mathbb{P}_{H_0}\left[p_j \le \alpha/m\right] = m \cdot \alpha/m = \alpha. \tag{9}$$

Importantly, while the Fisher combination test is valid only for independent p-values, *the Bonferroni test is valid for arbitrary p-value dependency structures.* However, the Bonferroni bound derived above is tightest for independent p-values. For example, if the p-values are perfectly dependent, then no multiplicity correction is required at all.

1.2 Multiple testing

While global testing seeks to detect the presence of *any* signals, multiple testing seeks to *localize* these signals, i.e. find a subset S of the null hypotheses that are false. Let $\{1, \dots, m\} = \mathcal{H}_0 \cup \mathcal{H}_1$, where $\mathcal{H}_0, \mathcal{H}_1$ are the sets of null hypotheses that are true and false, respectively. Ideally, we would like to have $S = \mathcal{H}_1$, but of course we typically cannot do this. We design methods such outputting sets S satisfying satisfying some Type-I error control criterion, and compare their performance based on their power, e.g. as quantified by $\mathbb{E}[|S \cap \mathcal{H}_1|/|\mathcal{H}_1|]$. There are several Type-I error control criteria of interest, but we highlight the two most important ones:

- Family-wise error rate (FWER), defined

$$\text{FWER} \equiv \mathbb{P}[S \cap \mathcal{H}_0 \neq \varnothing]. \tag{10}$$

- False discovery rate (FDR), defined

$$\text{FDR} \equiv \mathbb{E}\left[\frac{|S \cap \mathcal{H}_0|}{|S|}\right], \quad \text{where} \quad \frac{0}{0} \equiv 0. \tag{11}$$

The random quantity $\frac{|S \cap \mathcal{H}_0|}{|S|}$ is called the *false discovery proportion* (FDP). Note that the FWER is a stricter error rate than the FDR. Controlling the FWER at level α implies that, with probability $1 - \alpha$, the set S contains no false discoveries at all. Controlling the FDR at level q means that, on average, at most a proportion q of the set S can be false discoveries. Many methods have been proposed to control each of these error rates, but we highlight one each.

1.2.1 The Bonferroni procedure for FWER control

We discussed the Bonferroni test for the global null. This test can be extended to an FWER-controlling procedure:

$$S \equiv \{j : p_j \le \alpha/m\}. \tag{12}$$

Note that not all global tests can be extended to FWER-controlling procedures in this way. For example, the Fisher combination test does not single out any of the hypotheses, as it only aggregates the p-values. By contrast, the Bonferroni test searches for p-values that are individually very small, allowing for it to double as an FWER-controlling procedure. It is easy to verify that the Bonferroni procedure controls the FWER:

$$\mathbb{P}[S \cap \mathcal{H}_0 \neq \varnothing] = \mathbb{P}\left[\min_{j \in \mathcal{H}_0} p_j \leq \alpha/m\right] \leq \sum_{j \in \mathcal{H}_0} \mathbb{P}[p_j \leq \alpha/m] = \frac{|\mathcal{H}_0|}{m}\alpha \leq \alpha. \tag{13}$$

Note that the FWER is actually controlled at the level $\frac{|\mathcal{H}_0|}{m}\alpha \leq \alpha$, making the Bonferroni test conservative to the extent that $|\mathcal{H}_0| < m$. The null proportion $\frac{|\mathcal{H}_0|}{m}$ has such an effect on the performance of many multiple testing procedures.

1.2.2 The Benjamini-Hochberg procedure for FDR control

Designing procedures with FDR control, as well as verifying the latter property, is typically harder than for FWER control. It is harder to decouple the effects of the individual hypotheses, as the denominator $|S|$ in the FDR definition (11) couples them together. Both the FDR criterion and the most popular FDR-controlling procedure were proposed by Benjamini and Hochberg in 1995.

Procedure. To define the BH procedure, consider thresholding the p-values at $t \in [0, 1]$. We would expect $\mathbb{E}[|\{j : p_j \leq t\} \cap \mathcal{H}_0|] = |\mathcal{H}_0|t$ false discoveries among $\{j : p_j \leq t\}$. Since $|\mathcal{H}_0|$ is unknown, we can bound it from above by mt. This leads to the FDP estimate

$$\widehat{\text{FDP}}(t) \equiv \frac{mt}{|\{j : p_j \leq t\}|}. \tag{14}$$

The BH procedure is then defined via

$$S \equiv \{j : p_j \leq \widehat{t}\}, \quad \text{where} \quad \widehat{t} = \max\{t \in [0, 1] : \widehat{\text{FDP}}(t) \leq q\}. \tag{15}$$

In words, we choose the most liberal p-value threshold for which the estimated FDP is below the nominal level q. Note that the set over which the above maximum is taken is always nonempty because it at least contains 0: $\widehat{\text{FDP}}(0) = \frac{0}{0} \equiv 0$.

FDR control under independence. Benjamini and Hochberg established that their procedure controls the FDR if the p-values are independent. Here we present an alternative argument due to Storey, Taylor, and Siegmund (2004).

Proof. We have

$$\begin{aligned} \text{FDR} = \mathbb{E}\left[\text{FDP}(\widehat{t})\right] &= \mathbb{E}\left[\frac{|\{j \in \mathcal{H}_0 : p_j \leq \widehat{t}\}|}{|\{j : p_j \leq \widehat{t}\}|}\right] \\ &= \mathbb{E}\left[\frac{|\{j \in \mathcal{H}_0 : p_j \leq \widehat{t}\}|}{m\widehat{t}}\widehat{\text{FDP}}(\widehat{t})\right] \leq q \cdot \mathbb{E}\left[\frac{|\{j \in \mathcal{H}_0 : p_j \leq \widehat{t}\}|}{m\widehat{t}}\right]. \end{aligned} \tag{16}$$

To prove that the last expectation is bounded above by 1, note that

$$M(t) \equiv \frac{|\{j \in \mathcal{H}_0 : p_j \leq t\}|}{mt} \tag{17}$$

is a backwards martingale with respect to the filtration

$$\mathcal{F}_t = \sigma(\{p_j : j \in \mathcal{H}_1\}, |\{j \in \mathcal{H}_0 : p_j \leq t'\}| \text{ for } t' \geq t), \tag{18}$$

with t running backwards from 1 to 0. Indeed, for $s < t$ we have

$$\mathbb{E}[M(s)|\mathcal{F}_t] = \mathbb{E}\left[\frac{|\{j \in \mathcal{H}_0 : p_j \leq s\}|}{ms}\middle|\mathcal{F}_t\right] = \frac{\frac{s}{t}|\{j \in \mathcal{H}_0 : p_j \leq t\}|}{ms} = \frac{|\{j \in \mathcal{H}_0 : p_j \leq t\}|}{mt} = M(t). \tag{19}$$

The threshold $\widehat{t}$ is a stopping time with respect to this filtration, so by the optional stopping theorem, we have

$$\mathbb{E}\left[\frac{|\{j \in \mathcal{H}_0 : p_j \leq \widehat{t}\}|}{m\widehat{t}}\right] = \mathbb{E}[M(\widehat{t})] \leq \mathbb{E}[M(1)] = \frac{|\mathcal{H}_0|}{m} \leq 1. \tag{20}$$

This completes the proof. □

FDR control under dependence. The BH procedure has empirically been shown to control the FDR for a wide variety of dependency structures besides independence. However, theoretical FDR control results for the BH procedure are available only for a few dependency structures. A notable example is a type of positive dependency called *positive regression dependence on a subset*, or PRDS. Benjamini and Yekutieli proved FDR control for BH under PRDS in 2001. This theoretical condition is somewhat hard to verify in practice, however. The simplest example of a set of PRDS p-values is when $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^m$ where $\boldsymbol{\Sigma}$ has all positive entries and the p-values are derived based on one-sided tests. Outside of this special case, there are few known instances of PRDS p-values.