Unit 6: Further Topics

Eugene Katsevich

December 6, 2021

Units 1-5 focused on estimation and inference in linear models and generalized linear models. In Unit 6, we explore further topics: multiple testing (Section 1) and high-dimensional inference under the model-X assumption (Section 2).

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

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

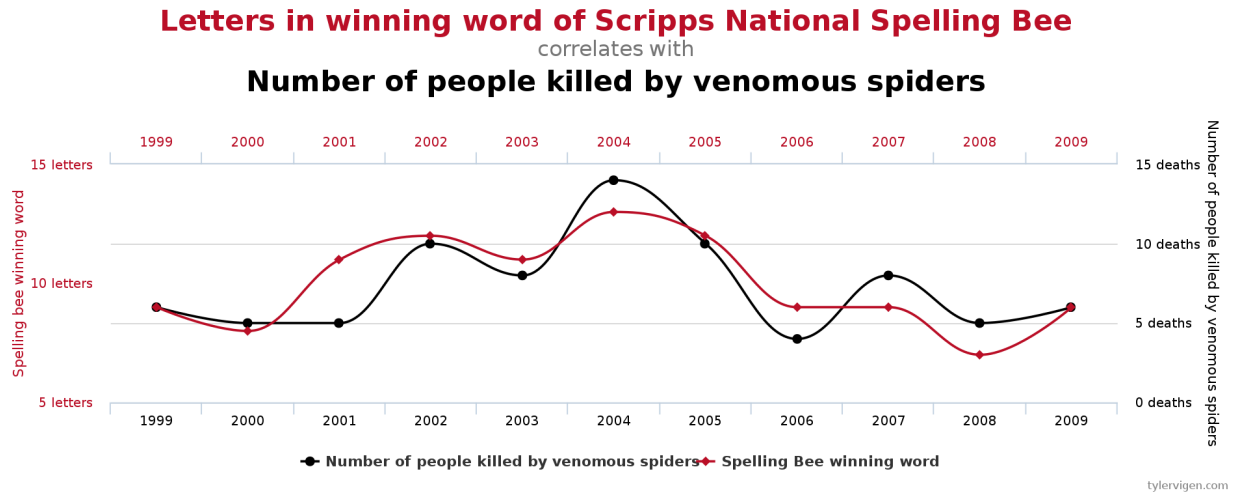

Figure 1: A spurious correlation resulting from data snooping.

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 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

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

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 $\{1, \ldots, m\} = \mathcal{H}_0 \cup \mathcal{H}_1$, 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

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

The BH procedure is then defined via

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

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: $\widehat{\mathrm{FDP}}(0) = \frac{0}{0} \equiv 0$.

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

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

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.

2 High-dimensional inference under Model-X

All of the statistical inference done so far in this class was *low-dimensional*: we assumed that the number of predictors p was fixed and at most equal to the sample size n. However, some modern applications fall outside of this regime and therefore require new statistical methodology. We discuss here a line of work initiated by Candès, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(3), 551–577.

2.1 Motivation and problem statement

All statistical inference requires assumptions, and inherently difficult problems like high-dimensional inference require strong assumptions. One such assumption is the

$$\textit{Model-X assumption:} \text{ That the joint distribution of } (x_1, \ldots, x_p) \text{ is known.} \tag{21}$$

This assumption is in some sense the opposite of what we have been considering in this class so far: Usually we assume nothing about the joint distribution of covariates (we treat these as fixed anyways), and assume instead that $y|(x_1, \ldots, x_p)$ follows a generalized linear model. Notably, this assumption is stronger than correct specification of a parametric model for $(x_1, \ldots, x_p)$; it states that we know not only a model for this distribution but all of its parameters as well. Below we discuss the motivation for this assumption, and the inference problem that grows out of it.

Motivation: Genome-wide association studies (GWAS). In GWAS, $x_1, \ldots, x_p \in \{0, 1, 2\}$ represent *genotypes* of an individual at p genomic locations. Suppose that humans typically have either an A or a T at genomic location j, where A is more common. Since we have two sets of chromosomes (one maternal and one paternal), the *genotype* at this location can either be AA, AT, or TT. The allele T is called the *minor allele* because it is less common, and x_j is defined as the number of minor alleles an individual has at location j: AA implies $x_j = 0$, AT implies $x_j = 1$, TT implies $x_j = 2$. We collect this genotype information at p genomic locations from each individual, as well as a response variable y, like disease status. The goal is to find the genomic locations whose genotypes are associated with the response. The nice thing in this application is that the joint distribution $(x_1, \ldots, x_p)$ has been studied extensively in the field of population genetics, and is well-approximated by a *hidden Markov model*. This motivates the model-X assumption.

Problem statement. It turns out that if we have a model for the joint distribution of the predictors, we need not make any assumptions on the distribution of the response given the predictors. But this leaves us with the following awkward question: If we have no parametric model for the response, then what even are the hypotheses we are testing? Well, for each genomic location j, we are trying to test whether the genotype at that location is associated with the response, controlling for the genotypes at other genomic locations. Probabilistically, this may be written as:

$$H_{0j} : x_j \perp\!\!\!\perp y \mid \boldsymbol{x}_{\text{-}j}. \tag{22}$$

Under mild assumptions, this hypothesis turns out to coincide with the usual $H_{0j} : \beta_j = 0$ in the case when y does follow a GLM. The problem statement, then, is to test the hypotheses H_{0j} based on data

$$(x_{i1}, \ldots, x_{ip}, y_i) \overset{\text{i.i.d.}}{\sim} F_{\boldsymbol{x},y}, \quad i = 1, \ldots, n, \tag{23}$$

given knowledge of the distribution $F_{\boldsymbol{x}}$. Note that *regularized regression* methods such as the LASSO have been developed to get estimates of regression coefficients in high dimensions. However, the issue with these shrinkage-based estimation methodologies is that they do not come with inferential guarantees and therefore cannot provide valid tests of the conditional independence hypothesis (22). Under the model-X assumption, we can get around this roadblock.

2.2 Conditional randomization test

One idea is to view x_j as a treatment (though not necessarily binary) and $\boldsymbol{x}_{\text{-}j}$ as a set of covariates. The model-X assumption gives us knowledge of the *propensity function* $p(x_j|\boldsymbol{x}_{\text{-}j})$, i.e. the distribution of treatment assignments given the covariates. In the spirit of Fisher's randomization test (see Homework 5 Problem 1), we can build a null distribution for any test statistic $T(\boldsymbol{X}, \boldsymbol{y})$—e.g. a lasso coefficient—by *randomly reassigning the treatment x_j to each individual based on its covariates $\boldsymbol{x}_{\text{-}j}$*. More explicitly, let

$$\widetilde{x}_{ij}|\boldsymbol{X}, \boldsymbol{y} \overset{\text{ind}}{\sim} F_{x_j|\boldsymbol{x}_{\text{-}j}=\boldsymbol{x}_{i,\text{-}j}}. \tag{24}$$

Let $\widetilde{\boldsymbol{X}}$ be the matrix obtained by replacing the jth column in $\boldsymbol{X}$ with $\widetilde{\widetilde{x}}_{*j}$ as defined above. For a test statistic T, we then define the CRT p-value by comparing the test statistic's value on the original data with its distribution under resampling:

$$p_j^{\text{CRT}} \equiv \mathbb{P}[T(\widetilde{\boldsymbol{X}}, \boldsymbol{y}) \geq T(\boldsymbol{X}, \boldsymbol{y})|\boldsymbol{X}, \boldsymbol{y}]. \tag{25}$$

In practice, we approximate this p-value by resampling a finite number B of instances $\widetilde{\boldsymbol{X}}^b$ and setting

$$\widehat{p}_j^{\text{CRT}} \equiv \frac{1}{B+1}\sum_{b=1}^{B} \mathbb{1}(T(\widetilde{\boldsymbol{X}}^b, \boldsymbol{y}) \geq T(\boldsymbol{X}, \boldsymbol{y})). \tag{26}$$

The CRT is a simple and elegant inferential framework that gives finite-sample valid p-values for high-dimensional inference. However, its adoption has been slowed by the computational cost of resampling.

2.3 Model-X knockoffs

An alternative to the CRT for model-X inference is *model-X knockoffs*. This methodology requires constructing a set of p new *knockoff* variables $(\widetilde{x}_1, \ldots, \widetilde{x}_p)$, whose joint distribution with the original variables satisfies the following exchangeability criterion:

$$\text{for each } j, \quad (x_j, \widetilde{x}_j) \stackrel{d}{=} (\widetilde{x}_j, x_j) \mid \boldsymbol{x}_{\text{-}j}, \widetilde{\boldsymbol{x}}_{\text{-}j}. \tag{27}$$

Knockoff variables are meant to serve as valid *negative controls* for the original variables: they have the same dependency structure but they have no association with the response y. Constructing such knockoff variables is a nontrivial endeavor that depends on the joint distribution of the original variables. If this can be done, then we can sample an entire knockoff matrix $\widetilde{\boldsymbol{X}}$, row by row. We then assess the significant of all $2p$ variables using test statistics $Z_1(\boldsymbol{X}, \widetilde{\boldsymbol{X}}, \boldsymbol{y}), \ldots, Z_p(\boldsymbol{X}, \widetilde{\boldsymbol{X}}, \boldsymbol{y}), \widetilde{Z}_1(\boldsymbol{X}, \widetilde{\boldsymbol{X}}, \boldsymbol{y}), \ldots, \widetilde{Z}_p(\boldsymbol{X}, \widetilde{\boldsymbol{X}}, \boldsymbol{y})$, constructed to ensure the following swap-equivariance property: swapping $\boldsymbol{X}_{*j}$ with $\widetilde{\boldsymbol{X}}_{*j}$ results in $Z_j(\boldsymbol{X}, \widetilde{\boldsymbol{X}}, \boldsymbol{y})$ swapping with $\widetilde{Z}_j(\boldsymbol{X}, \widetilde{\boldsymbol{X}}, \boldsymbol{y})$, while all the other test statistics stay the same. For example, consider running the LASSO of $\boldsymbol{y}$ on the *augmented design matrix* $[\boldsymbol{X}, \widetilde{\boldsymbol{X}}]$, and defining the Z_j's as the fitted coefficients for the corresponding variables. With these Z_j's in hand, the idea is to define the significance of the jth original variable by comparing the test statistics for itself and for its knockoff:

$$T_j(\boldsymbol{X}, \widetilde{\boldsymbol{X}}, \boldsymbol{y}) \equiv Z_j(\boldsymbol{X}, \widetilde{\boldsymbol{X}}, \boldsymbol{y}) - \widetilde{Z}_j(\boldsymbol{X}, \widetilde{\boldsymbol{X}}, \boldsymbol{y}). \tag{28}$$

Large values of T_j are evidence against H_{0j}. If the knockoffs are constructed correctly, then the test statistics T_j for null j can be shown to have symmetric distributions around zero. In other words, the original variable and its knockoff are equally likely to be more significant. Using this observation, a clever multiple testing algorithm called *Selective SeqStep* can be used to choose a cutoff $\widehat{t}$ for the test statistics in a way that provably controls the FDR at a pre-specified level q. Remarkably, this entire construction bypasses the construction of p-values!

2.4 Comparing CRT to MX knockoffs

There are pros and cons to both the CRT and MX knockoffs. Both procedures offer valid, finite-sample inference in high dimensions, which sets them apart from many other inferential methodologies. Both procedures require the model-X assumption, however, which may or may not be reasonable in a given application. MX knockoffs is the more popular methodology at this time, due to its computational speed. It can be used to carry out inference for all p hypotheses in "one shot", by running one big regularized regression on the augmented design matrix. It has been applied successfully to genome-wide association studies, using a hidden Markov model as the model for X. On the other hand, MX knockoffs is a randomized procedure, giving different results for different realizations of $\widetilde{\boldsymbol{X}}$. Furthermore, it does not provide p-values quantifying the significance of individual

predictors, which hinders the interpretability of its results. On the other hand, the CRT requires more computation than knockoffs, so it has been slower to be adopted in practice. But this procedure is not randomized in the same way that knockoffs is; with more computation its results can be arbitrarily "de-randomized." Furthermore, the CRT does have a p-value output, which facilitates easy interpretation and more flexibility for downstream multiple testing.