DS 102 Data, Inference, and Decisions **Fall 2019**

Lecture 3: False Discovery Rate Control

Lecturer: Michael I. Jordan

1 Recap of Decision Theory

Recall the decision-theoretic concepts introduced in Lecture 2.

Let X be a data set generated from some distribution P_θ which is indexed by a ground-truth parameter θ. We define a procedure $\delta(X)$ that operates on the data to make a decision. Typically, $\delta(X)$ is trying to "guess" θ from X. To measure how good the guess of our procedure is, we take a loss function $\ell(\theta, \delta(X))$, which takes as input the ground-truth parameter, as well as our prediction.

We'd like to design a procedure δ that minimizes the loss. However, we can't evaluate $\ell(\theta, \delta(X))$ directly, because θ is unknown, as is possibly X at the time of defining the procedure δ. Instead, we can take expectations, resulting in a risk function, which we will also seek to minimize. Now we will derive the frequentist risk function, defined as

$$R(\theta) = \mathbb{E}[\ell(\theta, \delta(X))],$$

for two important loss functions ℓ.

1.1 Example 1: 0/1 Loss

Suppose $\theta \in \{0, 1\}$, and similarly $\delta(X) \in \{0, 1\}$. The 0/1 loss is defined as:

$$\ell(\theta, \delta(X)) = \begin{cases} 1, \text{ if } \theta \neq \delta(X) \\ 0, \text{ if } \theta = \delta(X). \end{cases}$$

Compactly we can write this as $\ell(\theta, \delta(X)) = \mathbf{1}\{\delta(X) \neq \theta\}$. Then, the frequentist risk is equal to:

$$R(\theta) = \begin{cases} \mathbb{E}[\mathbf{1}\{\delta(X) = 1\}], \text{ if } \theta = 0 \\ \mathbb{E}[\mathbf{1}\{\delta(X) = 0\}], \text{ if } \theta = 1 \end{cases} = \begin{cases} \mathbb{P}(\delta(X) = 1), \text{ if } \theta = 0 \\ \mathbb{P}(\delta(X) = 0), \text{ if } \theta = 1 \end{cases},$$

where we use the fact that $\mathbb{E}[\mathbf{1}\{E\}] = \mathbb{P}(E)$, for any event E.

Now the question is how to create a good decision rule $\delta(X)$. At the same time we want both $\mathbb{P}(\delta(X) = 1 \mid \theta = 0)$ and $\mathbb{P}(\delta(X) = 0 \mid \theta = 1)$ to be small. There is no single right answer to this question. One natural solution would be to minimize $\mathbb{P}(\delta(X) = 1 \mid \theta = 0) + \mathbb{P}(\delta(X) = 0 \mid \theta = 1)$.

Another solution is given by the Neyman-Pearson hypothesis testing framework discussed in Lecture 2. We could minimize $\mathbb{P}(\delta(X) = 0 \mid \theta = 1)$, while constraining $\mathbb{P}(\delta(X) = 1 \mid \theta = 0)$ to be under a pre-specified level α (e.g. 0.05).

We could also minimize the weighted average of $\mathbb{P}(\delta(X) = 1 \mid \theta = 0)$ and $\mathbb{P}(\delta(X) = 0 \mid \theta = 1)$. For example, we would take the weights to be equal to π_0, which is our *prior belief* of the probability that $\theta = 0$, with $1 - \pi_0$ the probability that $\theta = 1$. In that case, we would be minimizing

$$\pi_0 \cdot \mathbb{P}(\delta(X) = 1 \mid \theta = 0) + (1 - \pi_0) \cdot \mathbb{P}(\delta(X) = 0 \mid \theta = 1).$$

This quantity is equal to the Bayes risk from Lecture 2.

1.2 Example 2: ℓ_2 Loss

Now we allow $\theta \in \mathbb{R}$, and likewise $\delta(X) \in \mathbb{R}$. We define the ℓ_2 loss to be equal to:

$$\ell(\theta, \delta(X)) = (\theta - \delta(X))^2.$$

We compute the frequentist risk under the ℓ_2 loss as:

$$\begin{aligned} R(\theta) &= \mathbb{E}[(\theta - \delta(X))^2] = \mathbb{E}[((\theta - \mathbb{E}[\delta(X)]) - (\delta(X) - \mathbb{E}[\delta(X)]))^2] \\ &= (\theta - \mathbb{E}[\delta(X)])^2 + \mathbb{E}[(\delta(X) - \mathbb{E}[\delta(X)])^2] - 2(\theta - \mathbb{E}[\delta(X)])\mathbb{E}[(\delta(X) - \mathbb{E}[\delta(X)])] \\ &= (\theta - \mathbb{E}[\delta(X)])^2 + \mathbb{E}[(\delta(X) - \mathbb{E}[\delta(X)])^2] \\ &:= \text{bias}^2 + \text{variance} \end{aligned}$$

Therefore, the frequentist risk can be decomposed into a bias term and a variance term. We want our decision rule to have small bias, meaning that on average it gives the right value, but we also want it to have low variance, meaning its output is "consistent" and does not exhibit a lot of variation. Typically, there is a trade-off between bias and variance; small bias implies big variance, and vice versa. Notice that this is similar to the trade-off between false positives and false negatives we saw earlier in Example 1.

1.3 Properties of Expectation

We recall several useful properties of mathematical expectation. Denote by X and Y random variables, and by a and b real-valued constants.:

(a) $\mathbb{E}[aX] = a\mathbb{E}[X]$ (homogeneity)

(b) $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ (additivity)

(c) $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$ (tower property, a.k.a. total law of expectation)

Homogeneity and additivity are together called linearity. Keep in mind that a conditional expectation $\mathbb{E}[X \mid Y]$ is a *random variable*, one possibly different than X, whose expectation is the same as that of X.

2 Multiple Hypothesis Testing

Multiple hypothesis testing is a problem that arises when we want to test many hypotheses while controlling an appropriate error rate. The Neyman-Pearson testing framework does not give satisfactory guarantees when applied to many hypotheses. In particular, controlling the probability of (falsely) discovering a null under α is vacuous if there are many tested nulls. For example, if there are m of them, the probability of falsely discovering at least one null might be as high as $m\alpha$! This is a useless guarantee if m is large.

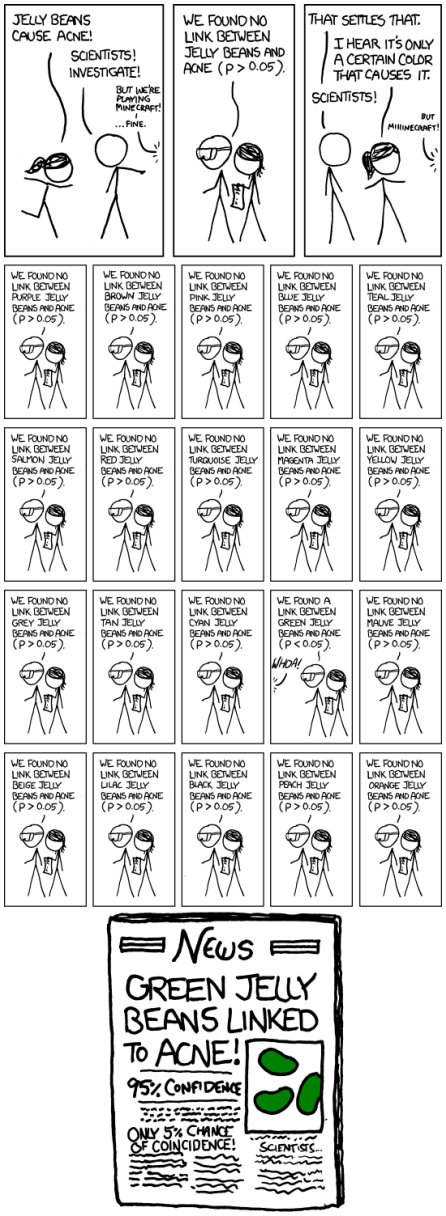

Figure 1.1: The multiple testing problem.

3 Bonferroni Correction

The initial approach to multiple testing was to control the probability of making at least one false discovery, also known as the **family-wise error rate** FWER:

$$\text{FWER} := \mathbb{P}(\text{at least one false discovery is made}).$$

A simple procedure that ensures FWER control is the **Bonferroni correction**. Instead of testing each hypothesis under level α, the idea is to test each hypothesis under level α/N, where N is the total number of hypotheses. Denote by $\{E_i = 1\}$ the event that a false discovery is made on the i-th test. By a union-bound argument, it follows that the FWER is controlled under α:

$$\text{FWER} = \mathbb{P}(\cup_{i=1}^{N}\{E_i = 1\}) \leq \sum_{i=1}^{N} \mathbb{P}(E_i = 1) \leq \sum_{i=1}^{N} \alpha/N = \alpha,$$

where the last inequality uses the fact that hypothesis tests by design have probability of false discovery controlled under the chosen significance level, which in this case is α/N.

The Bonferroni correction is unfortunately too stringent, and often does not make any discoveries; as N grows large, α/N is too small. And indeed, this makes intuitive sense: if we test a million hypotheses, preventing the possibility of making a *single* false discovery in a million tests necessarily implies that we have to be extremely conservative and only proclaim a discovery if there is overwhelmingly strong signal of something interesting.

4 Recap of P-values

In hypothesis testing we often use **p-values** to guide our decisions. The p-value is the probability that the **hypothetical data from the null hypothesis** would be equal to, or more extreme than, the actual observed samples. This concept should be familiar to you from previous classes, but we review it here for convenience.

Let $\mathcal{S} \sim \mathcal{P}_\theta$ be the observed data set, and let $T(\mathcal{S})$ be some real-valued summary statistic obtained from $\mathcal{S}$. For example, $\mathcal{S}$ could be a set of n i.i.d. samples $X_1, \ldots, X_n$, and $T(\mathcal{S})$ could be their average $\frac{1}{n}\sum_{i=1}^{n} X_i$.

Let the null hypothesis be $\theta \in \Theta_0$, and let the alternative be $\theta \in \Theta_1$. Suppose that for every $\theta_0 \in \Theta_0$, we can compute the distribution of $T(\mathcal{S}_0)$, where $\mathcal{S}_0 \sim \mathcal{P}_{\theta_0}$. Then, the p-value is defined as:

$$P := \sup_{\theta_0 \in \Theta_0} \mathbb{P}(T(\mathcal{S}_0) \geq T(\mathcal{S}) \mid T(\mathcal{S})),$$

where the "hallucinated" sample $\mathcal{S}_0$ is independent from $\mathcal{S}$. We proclaim a discovery if P is less than or equal to the significance level α.

Going forward, we will use a few basic properties of p-values. First, they take values in $[0, 1]$, which is not surprising given that they are a probability. Second, if the null is true, they are *uniformly*

distributed:

$$\text{If the null is true, } \mathbb{P}(P \leq u) = u, \text{for all } u \in [0, 1].$$

If the alternative is true, on the other hand, we do not know anything about the distribution of P, but it is reasonable to expect P to be more "biased" toward smaller values.

5 False Discovery Proportion

Recall that in Lecture 2 we defined the **false discovery proportion** (FDP) as:

$$\text{FDP} = \frac{n_{01}}{n_{01} + n_{11}}.$$

		decision		
		null (0)	*non-null (1)*	
reality	*null*	n_{00}	n_{01}	$n_{00} + n_{01}$
	non-null	n_{10}	n_{11}	$n_{10} + n_{11}$
		$n_{00} + n_{10}$	$n_{01} + n_{11}$	N

Table 1.1: Different ground truth and decision relationships in multiple testing.

We also mentioned that this quantity can be thought of as being an estimate of a conditional probability $\mathbb{P}(H = 0 \mid D = 1)$, where D is a random variable denoting the decision, and H is a random variable denoting the ground truth about the hypothesis. Unlike sensitivity and specificity, this quantity is dependent on the prevalence (prior probability of null, i.e. $\mathbb{P}(H = 0)$).

By Bayes' theorem, we can rewrite this conditional probability as

$$\mathbb{P}(H = 0 \mid D = 1) = \frac{\mathbb{P}(D = 1 \mid H = 0)\, \mathbb{P}(H = 0)}{\mathbb{P}(D = 1)} := \frac{\mathbb{P}(\text{type I error})\, \pi_0}{\mathbb{P}(D = 1)}.$$

We can simply bound π_0 by 1; this even makes a reasonable assumption, because in practice most things we test are truly null. In both frequentist and Bayesian thinking, $\mathbb{P}(\text{type I error})$ is assumed known, because we assume we know how the data behave under the null. The denominator $\mathbb{P}(D = 1)$ is equal to

$$\mathbb{P}(D = 1) = \pi_0 \mathbb{P}(D = 1 | H = 0) + (1 - \pi_0)\mathbb{P}(D = 1 | H = 1),$$

so this term does indeed depend on the prevalence. However, a fortunate circumstance is that it can easily be estimated from data. An "obvious" estimate of this quantity is

$$\widehat{\mathbb{P}}(D = 1) = \frac{n_{01} + n_{11}}{N}.$$

Therefore, it seems reasonable to find a procedure that ensures

$$\mathbb{P}(H = 0 \mid D = 1) \approx \frac{\mathbb{P}(\text{type I error})}{\widehat{\mathbb{P}}(D = 1)} \leq \alpha.$$

We will return to this idea in the next section.

6 False Discovery Rate Control

The **false discovery rate** is defined as the expectation of the false discovery proportion:

$$\text{FDR} = \mathbb{E}[\text{FDP}] = \mathbb{E}\left[\frac{n_{01}}{n_{01}+n_{11}}\right].$$

Since data are random, we cannot hope to control the FDP under a target level α with probability one. For example, if all of our hypotheses are null, there is a possibility that all corresponding p-values are extremely small, small enough to be rejected (maybe even under the Bonferroni correction). This event indeed happens with very small probability, however it prevents us from controlling FDP across *all* events. For this reason, we are happy if we can control the FDP *on average*; in other words, we want to design decision rules which will guarantee that the FDR is kept under a target level α (e.g. 0.05). Because the FDR is not random but is simply a number, this will be manageable.

FDR was proposed as an appropriate error metric in multiple testing by Benjamini and Hochberg, who argued that FWER is too stringent for modern testing practices. They also proposed the first procedure for FDR control, usually called the Benjamini-Hochberg (BH) procedure. It takes a target FDR level α, as well as a set of p-values, and outputs which of the p-values correspond to discoveries. It does so in a way that guarantees FDR $\leq \alpha$. The explicit procedure statement is given below.

Algorithm 1 The Benjamini-Hochberg Procedure

input: FDR level α, set of N p-values $P_1, \ldots, P_N$
Sort the p-values $P_1, \ldots, P_N$ in non-decreasing order $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(N)}$
Find $K = \max\{i \in \{1, \ldots, N\} : P_{(i)} \leq \frac{\alpha}{N} i\}$
Reject the null hypotheses (declare discoveries) corresponding to $P_{(1)}, \ldots, P_{(K)}$

Notice that it makes sense to sort p-values and start rejecting from the smallest one; small p-values are generally indicative of an interesting finding, and null p-values are very small with tiny probability (recall that they are uniformly distributed).

The picture below visualizes how the BH procedure works.

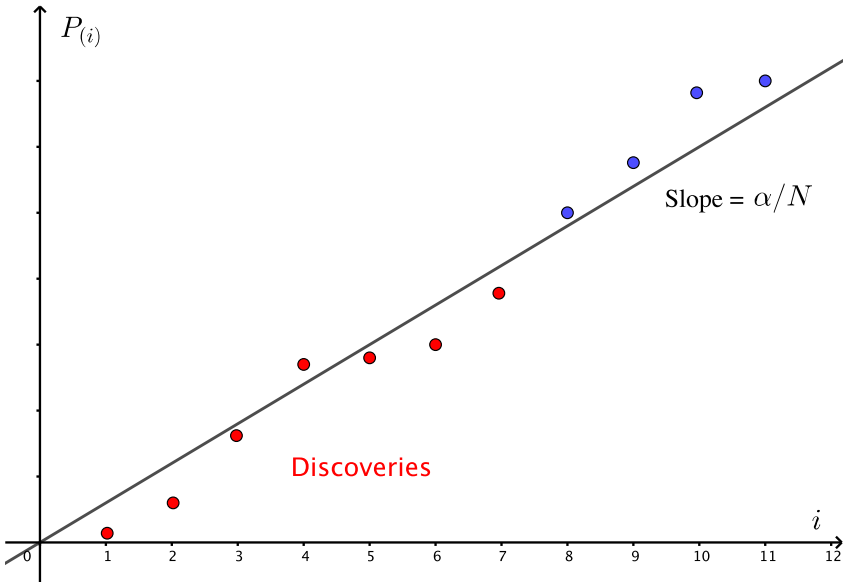

Figure 1.2: Illustration of the BH procedure.

Recall our discussion in Section 5, where we said that ensuring $\frac{\mathbb{P}(\text{type I error})}{\widehat{\mathbb{P}}(D=1)} \leq \alpha$ seems like a reasonable false discovery guarantee. In that case, we would like to pick the "least conservative" rule that will satisfy this. Suppose that we discover all p-values less than some threshold, and imagine that this threshold is equal to the K-th largest p-value, denoted $P_{(K)}$, for some fixed K. Under this decision rule, $\mathbb{P}(D = 1) = \frac{K}{N}$ by construction. On the other hand, the probability of a false positive $\mathbb{P}(\text{type I error})$ is equal to the probability that a null p-value is less than or equal to $P_{(K)}$. Since null p-values are uniform, this probability is exactly equal to $P_{(K)}$. Therefore, the condition $\frac{\mathbb{P}(\text{type I error})}{\widehat{\mathbb{P}}(D=1)} \leq \alpha$ can now be written as $\frac{P_{(K)}}{K/N} \leq \alpha$, i.e. $P_{(K)} \leq \frac{\alpha}{N}K$. Since we want our decision rule to be the least conservative one, we pick the maximum K such that $P_{(K)} \leq \frac{\alpha}{N}K$. Notice that this exactly corresponds to the BH procedure! To conclude, we have given a reinterpretation of the BH procedure, which says that it could actually be thought of as controlling an estimate of $\mathbb{P}(H = 0 \,|D = 1)$, the probability of a null given that we have made a discovery.