Global Testing and Multiple Testing

Andrew Nobel

October, 2020

Review of Hypothesis Testing

Basic Ingredients

- Family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of distributions on sample space $\mathcal{X}$
- Distinguished subset $\Theta_0 \subseteq \Theta$, and $\Theta_1 = \Theta_0^c$.

Goal: Use observation $X \sim P_\theta \in \mathcal{P}$ to distinguish between hypotheses

- $\mathrm{H}_0 : \theta \in \Theta_0$ (Null)
- $\mathrm{H}_1 : \theta \in \Theta_1$ (Alternative)

Review of P-Values

Recall: A valid *p-value* is a function $p : \mathcal{X} \to [0,1]$ such that

$$p(X) \overset{d}{\geq} \mathsf{U}(0,1) \text{ whenever } X \sim P_\theta \text{ with } \theta \in \Theta_0$$

- If $S : \mathcal{X} \to \mathbb{R}$ be a test statistic with large values favoring H_1 then

$$p(x) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(S(X) \geq S(x)) \text{ is a valid p-value}$$

- If $p(x)$ is a valid p-value then the test $(p(\cdot), [0, \alpha])$ has level α

Multiple Hypothesis Tests

Setting: Interested in multiple, related, hypothesis tests

$$\mathrm{H}_{0,i} \text{ vs. } \mathrm{H}_{1,i} \text{ for } 1 \leq i \leq n$$

- For each test obtain p-value p_i from appropriate test statistic and rejection region
- Standard assumption: if $\mathrm{H}_{0,i}$ is true then $p_i \sim \mathrm{U}(0,1)$

Fundamental Problems: Given p-values $p_1, \ldots, p_n$

- Global testing: Decide whether all nulls $\mathrm{H}_{0,i}$ are true
- Multiple testing: Identify subset of tests for which $\mathrm{H}_{0,i}$ is false.

Issue: When testing many hypotheses, some p-values will be small by chance

Example: Testing Coins

Setting: Have n coins with $P(\text{coin } i \text{ is heads}) = \theta_i$. Key questions

- Are some of the coins biased, that is $\theta_i \neq 1/2$?
- If so, which ones?

Experimental Procedure

- Flip each coin 20 times
- Use outcomes for coin i to test hypothesis $\mathrm{H}_{0,i} : \theta_i = 1/2$
- Obtain p-values $p_1, \ldots, p_n$ from Binomial hypothesis test

Key Issue: If n is large, some p-values will be small *just by chance*

- How to draw rigorous conclusions from the p-values $p_1, \ldots, p_n$?

Testing for Differential Gene Expression

Basic design: Two groups of individuals

- n_a with disease (treatment)
- n_b without disease (control)

For each individual, measure expression of $m \approx 20K$ genes in biopsied tissue

Data: Two matrices $X^a \in \mathbb{R}^{n_a \times m}$ (treatment) and $X^b \in \mathbb{R}^{n_b \times m}$ (control) where

$$X^c_{i,j} = \text{ expression level of gene } j \text{ in individual } i \text{ of group } c$$

Scientific Questions:

- Do any genes show differential expression between T and C groups?
- If so, which genes?

Testing for Differential Gene Expression

Model Assumption: Suppose that for each gene $j = 1, \ldots, m$

- n_a expression measurements $X^a_{\cdot j}$ are iid $\sim \mathcal{N}(\mu_j, \sigma_j^2)$
- n_b expression measurements $X^b_{\cdot j}$ are iid $\sim \mathcal{N}(\nu_j, \tau_j^2)$

Approach

- Test hypotheses $\mathrm{H}_{0,j} : \mu_j = \nu_j$ for genes $j = 1, \ldots, m$
- For each j calculate a two-sample t-statistic T_j from $X^a_{\cdot j}$ and $X^b_{\cdot j}$
- Compute p-values p_j from T_j using percentile of t-distribution

Issue: If m is large, some p-values will be small *just by chance*

- How to draw rigorous conclusions from p-values $p_1, \ldots, p_m$?

Global Testing

Definition

- *Global null* $\mathrm{H}_0 = \wedge_{i=1}^{n} \mathrm{H}_{0,i}$ asserts that all nulls $H_{0,i}$ are true
- *Global test* uses p-values $p = p_1, \ldots, p_n$ to test H_0 against variety of alternatives

Underlying Issues

- Control of Type I error
- Power against different classes of alternatives

Fisher Combination Test

Fisher Combination Test: Fix desired size $\alpha \in (0, 1)$

- Combine p-values using statistic $T(p) = \sum_{i=1}^{n} -2 \log p_i$
- Reject global null H_0 if $T(p) \geq \chi^2_{2n}(1 - \alpha)$

Fact: If $p_1, \ldots, p_n$ are independent then FCT has size α.

Bonferroni Global Test

BGT: Fix desired level α. Given p-values $p_1, \ldots, p_n$ reject global null H_0 if

$$T(p) = \min_{1 \le i \le n} p_i \ \le \ \frac{\alpha}{n}$$

Equivalently: Reject H_0 if one or more $\mathrm{H}_{0,i}$ is rejected at level α/n

Fact

(a) BGT has level α regardless of the dependence between the p-values

(b) If p-values are independent then level $\approx 1 - e^{-\alpha}$

Bonferroni vs Fisher

Ordered p-values

- If p-values $p_1, \ldots, p_n$ iid $\sim$ U$(0,1)$ then $\mathbb{E}p_{(r)} = r/(n+1)$
- Under H$_0$ expect $(1, p_{(1)}), \ldots, (n, p_{(n)})$ close to line with slope $(n+1)^{-1}$
- Points lying below the line provide evidence against H$_0$

Note

- Fisher test aggregates information across p-values via sum. Powerful when many p-values are slightly below the line
- Bonferroni test considers only the smallest p-value. Powerful when smallest p-value(s) far below the line

Gaussian Sequence Model (GSM)

General Setting

- Observe $Y_1, \ldots, Y_n$ independent with $Y_i \sim \mathcal{N}(\theta_i, 1)$
- In vector form, $Y \sim \mathcal{N}_n(\theta, I)$ with $\theta = (\theta_1, \ldots, \theta_n)^t$
- Of interest: inference about mean vector θ
- Individual tests $\mathrm{H}_{0,i} : \theta_i = 0$ vs. $\mathrm{H}_{1,i} : \theta_i \neq 0$
- Global null H_0 asserts that mean vector $\theta = 0$

Attraction: Useful setting in which to study optimality properties of testing procedures under different alternatives

Gaussian Sequence Model Under Sparse Alternative

Observe: $Y_1, \ldots, Y_n$ independent with $Y_i \sim \mathcal{N}(\theta_i, 1)$

- Let $\theta = (\theta_1, \ldots, \theta_n)^t$ be the mean vector of the observations
- Global null $\mathsf{H}_0 : \theta = 0$
- Sparse alternative, sometimes called "needle in a haystack",

$$\mathsf{H}_1 : \theta \in \{s_n e_1, \ldots, s_n e_n\}$$

  where $e_i = i$th canonical basis vector and $s_n =$ signal strength

Fact: In GSM

- $p_i = \overline{\Phi}(Y_i)$ is a valid one-sided p-value for $1 \leq i \leq n$
- BGT rejects H_0 if $\max_i Y_i \geq z(\alpha/n) = \Phi^{-1}(1 - \alpha/n)$

Power Analysis of BGT in Gaussian Sequence Model

Fact: Let $\alpha \in (0,1)$ and $\epsilon > 0$ be fixed.

(a) If signal strength $s_n \;\geq\; (1+\epsilon)\sqrt{2\log n}$ then

$$\lim_{n\to\infty} \mathbb{P}_1(\text{BGT rejects H}_0) \;=\; 1$$

(b) If signal strength $s_n \;\leq\; (1-\epsilon)\sqrt{2\log n}$ then

$$\limsup_{n\to\infty} \mathbb{P}_1(\text{BGT rejects H}_0) \;\leq\; 1 - e^{-\alpha}$$

Limits to Testing Under Sparse Alternatives

Q: How well can *any* testing procedure do against sparse alternatives?

A: Essentially, no better than Bonferroni.

Basic Idea

- Create simple alternative $\overline{\mathrm{H}}_1$ by averaging over vectors $s_n e_i$
- Obtain densities f_0 and f_1 for H_0 and $\overline{\mathrm{H}}_1$
- Assume signal strength $s_n = (1-\epsilon)\sqrt{2\log n}$
- Argue that size α test based on $f_1/f_0 \geq \tau$ must have Type II error $\approx 1-\alpha$
- Neyman-Pearson shows that no size α test has higher power

Limits to Testing Under Sparse Alternatives

Let $Y \sim \mathcal{N}_n(\theta, I)$. Consider testing

- $\mathsf{H}_0 : \theta = 0$
- $\overline{\mathsf{H}}_1 : \theta \sim \mathsf{Unif}(s_n e_1, \ldots, s_n e_n)$ with $s_n = (1-\epsilon)\sqrt{2 \log n}$

Proposition: Let $\alpha \in (0,1)$ and $\delta > 0$ be fixed. When n is sufficiently large any level α test for H_0 vs $\overline{\mathsf{H}}_1$ will have Type II error probability $\geq 1 - \alpha - \delta$.

Upshot: In the GSM with sparse alternative, no test can significantly outperform BGT

Chi-Squared Test

Observations: $Y \sim \mathcal{N}_n(\theta_n, I)$

Chi-squared test: Fix desired level α. Given Y reject global null H_0 if

$$T(Y) \;=\; ||Y||^2 \;=\; \sum_{i=1}^{n} Y_i^2 \;\geq\; \chi_n^2(1-\alpha)$$

Fact: For the chi-squared test with size α, the asymptotic probability of Type II error is

$$\begin{cases} 0 & \text{if } ||\theta||^2/\sqrt{2n} \to \infty \\ 1-\alpha & \text{if } ||\theta||^2/\sqrt{2n} \to 0 \end{cases}$$

The chi-squared is essentially optimal against $\mathrm{H}_1 = \{\theta : ||\theta|| = \eta\}$ if $\eta^2/\sqrt{2n} \to 0$

Multiple Testing

Setting

- Testing hypotheses $\mathsf{H}_{0,i}$ for $1 \leq i \leq n$
- Associated p-values $p_1, \ldots, p_n$
- *Key assumption:* p-value $p_i \sim \mathsf{U}(0,1)$ if $\mathsf{H}_{0,i}$ is true

New ingredient: Configuration of true null-hypotheses

- Which $\mathsf{H}_{0,i}$ are true?
- Binary vector $c \in \{0,1\}^n$ with $c_i = \mathbb{I}(\mathsf{H}_{0,i}$ true$)$

Multiple Testing Procedure

Definition: A *multiple testing procedure* γ rejects or accepts each $\mathrm{H}_{0,i}$ based on p-values $p = (p_1, \ldots, p_n)$. Formally, $\gamma : [0,1]^n \to \{0,1\}^n$ where

$$\gamma(p)_i = \begin{cases} 1 & \text{reject } \mathrm{H}_{0,i} \\ 0 & \text{accept } \mathrm{H}_{0,i} \end{cases}$$

Rejected hypotheses are referred to as *discoveries*. Two kinds of errors

- *False discovery:* Reject $\mathrm{H}_{0,i}$ when $\mathrm{H}_{0,i}$ is true
- *Missed discovery:* Accept $\mathrm{H}_{0,i}$ when $\mathrm{H}_{1,i}$ is true

Outcome of Multiple Testing Procedure

	γ accepts null	γ rejects null	Total
null true	U	V	n_0
alternative true	T	S	$n - n_0$
	$n - R$	R	n

Focus: Controlling false discoveries

- $V =$ Number of false discoveries [depends on p, $\gamma(\cdot)$, and true nulls]
- $R =$ Total number of discoveries [depends on p and $\gamma(\cdot)$]
- Familywise error rate (FWER) $= \mathbb{P}(V \geq 1)$
- False discovery rate (FDR) $= \mathbb{E}[V/(R \vee 1)]$

Familywise Error Rate

Definition: Familywise error rate (FWER) of multiple testing procedure γ is the probability of one or more false discoveries, $\mathbb{P}(V \geq 1)$.

Given $\alpha \in (0, 1)$ we say that γ has

- *Weak control* of FWER if $\mathbb{P}(V \geq 1) \leq \alpha$ under global null H_0
- *Strong control* of FWER if $\mathbb{P}(V \geq 1) \leq \alpha$ for any configuration of true nulls

Example: Two-Step Procedure for GSM

Description: Observe $Y \sim \mathcal{N}_n(\theta, I)$. Define $p_i = \overline{\Phi}(Y_i)$ for $1 \leq i \leq n$

- Reject H_0 if $\min p_i \leq \alpha/n$
- If we reject H_0 then reject $\mathrm{H}_{0,i}$ if $p_i \leq \alpha$.

Fact: Two-step procedure gives only weak control of FWER.

Easy to see: if $\theta_j \geq (1+\epsilon)\sqrt{2\log n}$ for one j then $V(p) \approx \alpha\,|\{i : \theta_i = 0\}|$

Example: Bonferroni's Procedure

Short version: Reject null $\mathsf{H}_{0,i}$ iff $p_i \leq \alpha/n$

Bonferroni Procedure

1. Input: Error level α and p-values $p_1, \ldots, p_n$ from individual tests
2. Order p-values $p_{(1)} \leq \ldots \leq p_{(n)}$ and hypotheses $\mathsf{H}_{0,(1)}, \ldots, \mathsf{H}_{0,(n)}$
3. Let $\hat{k} =$ largest $k \geq 1$ such that $p_{(k)} \leq \alpha/n$
4. Reject $\mathsf{H}_{0,(1)}, \ldots, \mathsf{H}_{0,(\hat{k})}$ and accept all other null hypotheses

Fact: Bonferroni procedure gives strong control of FWER under arbitrary dependence. For any configuration of true null hypotheses

$$\text{FWER(Bonferroni)} \leq \left(\frac{n_0}{n}\right)\alpha$$

Holm's Step-Down Procedure

Holm's Procedure

1. Input: Error level α and p-values $p_1, \ldots, p_n$ from individual tests
2. Order p-values $p_{(1)} \leq \ldots \leq p_{(n)}$ and hypotheses $\mathrm{H}_{0,(1)}, \ldots, \mathrm{H}_{0,(n)}$
3. Let $\hat{k} =$ smallest $k \geq 1$ such that $p_{(k)} > \alpha/(n-k+1)$
4. Reject $\mathrm{H}_{0,(1)}, \ldots, \mathrm{H}_{0,(\hat{k}-1)}$ and accept all other hypotheses

Fact: Holm's procedure gives strong control of FWER under arbitrary dependence. For any configuration of true null hypotheses

$$\text{FWER(Holm)} \leq \alpha$$

False Discovery Rate

	γ accepts null	γ rejects null	total
null true	U	V	n_0
null false	T	S	$n - n_0$
	$n - R$	R	n

Definition: The *false discovery proportion* of a multiple testing procedure γ is

$$\text{FDP}(\gamma) = \frac{V}{R \vee 1} = \begin{cases} V/R & \text{if } R \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

The *false discovery rate* of the procedure is $\text{FDR}(\gamma) = \mathbb{E}[\text{FDP}(\gamma)]$.

Familywise error rate vs. False Discovery Rate

Recall: A multiple testing procedure γ gives strong control of FWER if, for any configuration of null hypotheses, the probability of a false discovery is at most α

Definition: A multiple testing procedure γ controls FDR if, for any configuration of null hypotheses, the expected proportion of false discoveries is at most α

Fact: For any multiple testing procedure γ

(a) Under global null FWER$(\gamma) =$ FDR(γ)

(b) Control of FDR gives weak control of FWER

(c) For any configuration of null hypotheses, FWER$(\gamma) \geq$ FDR(γ)

Benjamini-Hochberg Step-Up Procedure

BH Step-Up Procedure

(0) Input: Error level α and p-values $p_1, \ldots, p_n$ from individual tests

(1) Order p-values $p_{(1)} \leq \ldots \leq p_{(n)}$ and hypotheses $\mathrm{H}_{0,(1)}, \ldots, \mathrm{H}_{0,(n)}$

(2) Let $\hat{k} =$ largest $k \geq 1$ such that $p_{(k)} \leq \alpha k/n$

(3) Reject $\mathrm{H}_{0,(1)}, \ldots, \mathrm{H}_{0,(\hat{k})}$ and accept all other hypotheses

Note: Index $\hat{k}$ is the *last time* that $p_{(k)} \leq \alpha k/n$

FDR Control of BH Procedure: Independent Case

Theorem: If p-values $p_1, \ldots, p_n$ are independent then the BH procedure satisfies

$$\text{FDR} \;\leq\; \left(\frac{n_0}{n}\right)\alpha \;\leq\; \alpha$$

- Proportion $\pi_0 = n_0/n$ of true nulls $\approx$ prior probability that $\text{H}_{0,i}$ is true
- π_0 may be much less than 1, yielding improved error bounds
- Substantial literature on how to estimate π_0 from observed p-values

FDR Control of BH Procedure: General Case

Theorem: For p-values $p_1, \ldots, p_n$ with arbitrary dependence, the BH procedure satisfies

$$\text{FDR} \;\leq\; \left(\frac{n_0}{n}\right) s(n)\,\alpha \;\leq\; s(n)\,\alpha$$

where $s(n) = 1 + 1/2 + \cdots + 1/n$

- Proportion $\pi_0 = n_0/n$ of true nulls still present in the bound
- $s(n) \approx \log n + 0.577$, so worst case loss of performance logarithmic in n
- If p-values are positive regression dependent the iid bound continues to hold