False Discovery Rate: A Novel Approach for Multiple Hypothesis Testing

Ansu Chatterjee
Subho Majumdar

School of Statistics, University of Minnesota

Table of contents

1. Hypothesis testing: introduction
2. The Family Wise Error approach
3. False Discovery Rate
4. Applications
5. Some other methods

Outline

1. Hypothesis testing: introduction
2. The Family Wise Error approach
3. False Discovery Rate
4. Applications
5. Some other methods

Hypothesis testing: the basics

- Statistical hypothesis testing is about evaluation of evidence to decide which of two competing "hypothesis" is more plausible.
- Both the hypothesis are statements about population parameters, only one of which can possibly be true.
- One of these hypotheses is called the **null hypothesis** H_0, the alternative to which is called the **alternative hypothesis** H_1 (or H_a).

A simple example

Example

Suppose $X_1, \ldots X_n$ are a random sample from a *Normal* distribution with mean μ and variance 1.
We want to test

$$\begin{aligned} H_0 : \mu &= 0 \text{ against} \\ H_1 : \mu &\neq 0 \end{aligned}$$

The basics steps

1. Make adequate and suitable assumptions.
2. Use relevant and testable hypotheses.
3. **Test statistic**: In order to test any pair of hypothesis, we need a *test statistic*. This is usually some transformation based on the point estimate of the parameter about which we hypothesize.

Example

A typical problem (continued) $X_1, \ldots X_n$ are a random sample from a *Normal* distribution with mean μ and variance 1. We want to test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$.
An estimator for μ is $\bar{X} = \sum X_i/n$.
We may use $\bar{X}$, or $\sqrt{n}\bar{X}$ as a *test statistic*.

The basics steps

1. Make adequate and suitable assumptions.
2. Use relevant and testable hypotheses.
3. **Test statistic**: In order to test any pair of hypothesis, we need a *test statistic*. This is usually some transformation based on the point estimate of the parameter about which we hypothesize.

Example

A typical problem (continued) $X_1, \ldots X_n$ are a random sample from a *Normal* distribution with mean μ and variance 1. We want to test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$.

An estimator for μ is $\bar{X} = \sum X_i/n$.

We may use $\bar{X}$, or $\sqrt{n}\bar{X}$ as a *test statistic*.

The basics steps

1. Make adequate and suitable assumptions.
2. Use relevant and testable hypotheses.
3. **Test statistic**: In order to test any pair of hypothesis, we need a *test statistic*. This is usually some transformation based on the point estimate of the parameter about which we hypothesize.

Example

A typical problem (continued) $X_1, \ldots X_n$ are a random sample from a *Normal* distribution with mean μ and variance 1. We want to test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$.
An estimator for μ is $\bar{X} = \sum X_i / n$.
We may use $\bar{X}$, or $\sqrt{n}\bar{X}$ as a *test statistic*.

The basics steps: p-values

1. Make adequate and suitable assumptions.
2. Use relevant and testable hypotheses.
3. **Test statistic**: In order to test any pair of hypothesis, we need a *test statistic*. This is usually some transformation based on the point estimate of the parameter about which we hypothesize.
4. ***P*-value**: This is a tool for evaluating whether the data favors the null or the alternative hypothesis.
   The *test statistic* takes some numeric value derived from the data. A *p-value* is usually computed as the probability of the test statistic taking what we see or more extreme values, assuming the null hypothesis is true.
5. Reject the null hypothesis is the *p*-value is too small, else don't reject it.

The basics steps: p-values

1. Make adequate and suitable assumptions.
2. Use relevant and testable hypotheses.
3. **Test statistic**: In order to test any pair of hypothesis, we need a *test statistic*. This is usually some transformation based on the point estimate of the parameter about which we hypothesize.
4. ***P*-value**: This is a tool for evaluating whether the data favors the null or the alternative hypothesis.
   The *test statistic* takes some numeric value derived from the data. A *p-value* is usually computed as the probability of the test statistic taking what we see or more extreme values, assuming the null hypothesis is true.
5. Reject the null hypothesis is the *p*-value is too small, else don't reject it.

The basics steps: p-values

1. Make adequate and suitable assumptions.
2. Use relevant and testable hypotheses.
3. **Test statistic**: In order to test any pair of hypothesis, we need a *test statistic*. This is usually some transformation based on the point estimate of the parameter about which we hypothesize.
4. ***P*-value**: This is a tool for evaluating whether the data favors the null or the alternative hypothesis.

   The *test statistic* takes some numeric value derived from the data. A *p-value* is usually computed as the probability of the test statistic taking what we see or more extreme values, assuming the null hypothesis is true.
5. Reject the null hypothesis is the *p*-value is too small, else don't reject it.

Why use *"reject/(not reject) the null"* terminology?

- The null hypothesis H_0 is the "protected" hypothesis, you do not want to reject it unless there is *strong evidence* in support of the alternative.
- The null hypothesis is like "*not guilty*" hypothesis. This hypothesis stands unless proven otherwise.
- Just evidence against the null will not do, you need evidence in favor of the alternative, otherwise the null hypothesis stands.
- This is like saying "*guilty as charged*". The "as charged" part of it is important.

The more classical approach

Example

A typical problem (continued) $X_1, \ldots X_n$ are a random sample from a *Normal* distribution with mean μ and variance 1. We want to test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$.
An estimator for μ is $\bar{X} = \sum X_i/n$.
We may use $\bar{X}$, or $\sqrt{n}\bar{X}$ as a *test statistic*.
The null hypothesis H_0 is rejected if $\sqrt{n}|\bar{X}| > C_\alpha$ for some pre-specified C_α.

The more classical approach

A typical problem (continued)

The null hypothesis H_0 is rejected if $\sqrt{n}|\bar{X}| > C_\alpha$ for some pre-specified C_α.
(*i.e.* $\sqrt{n}\bar{X} > C_\alpha$ *or* $\sqrt{n}\bar{X} < -C_\alpha$.)

- The values $\pm C_\alpha$ are known as **critical values.**
- The interval $(-C_\alpha, C_\alpha)$ is called the **acceptance region**, the region outside it is called **rejection region or critical region.**

The more classical approach

A typical problem (continued)

The null hypothesis H_0 is rejected if $\sqrt{n}|\bar{X}| > C_\alpha$ for some pre-specified C_α.
(*i.e.* $\sqrt{n}\bar{X} > C_\alpha$ *or* $\sqrt{n}\bar{X} < -C_\alpha$.)

- Two types of error may be committed in hypothesis tests:
  1. A **Type 1 error** occurs when H_0 is rejected when it is true.
  2. A **Type 2 error** occurs when H_0 is accepted when it is false.

The more classical approach

A typical problem (continued)

The null hypothesis H_0 is rejected if $\sqrt{n}|\bar{X}| > C_\alpha$ for some pre-specified C_α.

Two kinds of errors are possible: *Type 1: reject H_0 when it is true*, *Type 2: don't reject H_0 when it is false*.

- Since H_0 is to be protected, *Type 1* error is the more serious one.
- Making *Type 1* error is like convicting an innocent person.
- That's serious; that's why you've got those "beyond reasonable doubt" stuff in the courts.
- In statistical hypothesis tests, we ensure that the probability that a *Type 1 error* occurs is not more than a pre-specified α, called the **level of significance**.

The more classical approach

A typical problem (continued)

Two kinds of errors are possible: *Type 1: reject H_0 when it is true*, *Type 2: don't reject H_0 when it is false*.
Type 1 is more serious.
Probability of Type 1 error is not allowed to be more than pre-specified level of significance α.

- The critical value C_α is determined according to this level of significance α.
- Among tests of equal level α, the one for which the probability of *Type 2 error* is small is obviously better.
- The **Power** of a test is the *probability of not making Type 2 error*, *i.e.* the probability of rejecting H_0 when it is false.
- **Power**$= 1 - P[$ Type 2 error$]$.

What happens with repeated hypothesis tests?

Example

- Consider the data from *Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer* **(Nature Medicine. 13, 361-366 (2007))**.
- About 200 control cases +patients, data on about 23000 gene expressions.
- Want to check which genes are expressed in lung cancer cases. So we test if for each gene, the expressions in the cancer and the no-cancer group have the same mean.

What happens with repeated hypothesis tests?

Example

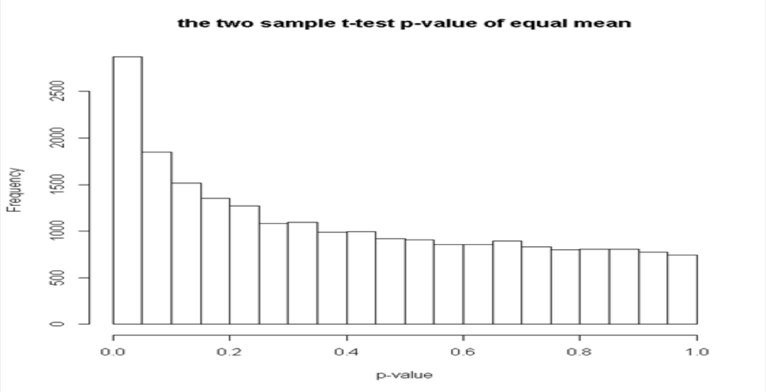

Figure : About 23K p-values, $\approx 3K$ are below 0.05, only $\approx$ 20 should be true discoveries

Outline

1. Hypothesis testing: introduction
2. **The Family Wise Error approach**
3. False Discovery Rate
4. Applications
5. Some other methods

Definition of Family Wise Error

- Need to test a *Family* of null hypotheses.
- The **Family Wise Error Rate** (FWER) is the probability of making one or more type-I error (i.e. false positive) in testing all the hypotheses.
- For example, If we test for 5 hypotheses simultaneously, each at 95% significance level, the the FWER is at most $0.05 \times 5 = 0.25$.
- There are different procedures of multiple hypotheses testing which utilize FWER.

Some FWER procedures

Bonferroni's procedure

Suppose we need to test *m* hypotheses. To keep the FWER less than α, simply do level-$\frac{\alpha}{m}$ tests for each of them.

- A conservative procedure, since the confidence level of each individual test is less than the overall confidence level. Probability of false positive is less.
- When the number of hypotheses is large, the level for an individual hypothesis test is very small and probability of false negatives (i.e. Type-II error) becomes very high.

Some FWER procedures

Tukey's procedure

Used for comparison of two populations.

Suppose the mean of the populations are m_1, m_2, then the test statistic is $\frac{m_1 - m_2}{SE}$, with SE being the standard error of the data.

A major limitation of this procedure is that to test for difference between a group of populations, one needs to test for each pair. This becomes cumbersome for a large number of populations (${}^{k}C_2$ comparisons for k populations).

Limitations of the FWER approach

- Several methods, each applicable for specific type of problem.
- Most of the methods do not work well for a large number of null hypotheses.

Outline

1. Hypothesis testing: introduction
2. The Family Wise Error approach
3. **False Discovery Rate**
4. Applications
5. Some other methods

The Setup

- We need to simultaneously test m hypotheses, of which m_0 are true.
- In the test procedure, suppose **R** hypotheses are declared significant.
- Also **U**, **T**, **V** and **S** are defined as in the table:

	Declared not significant	Declared significant	Total
True null	**U**	**V**	m_0
Non-true null	**T**	**S**	$m - m_0$
	$m -$ **R**	**R**	m

Defining the FDR

- Our goal is to determine that among all the hyoptheses deemed significant, how many are wrongly done, i.e. those hypotheses that are actually true but have been rejected by the test procedure.
- For this define another random variable $\mathbf{Q} = \mathbf{V}/(\mathbf{V} + \mathbf{S}) = \mathbf{V}/\mathbf{R}$. Also define $\mathbf{Q} = 0$ when $\mathbf{R} = 0$.

Definition

$\mathbf{V}, \mathbf{S}$ can not be observed, so we rather define the False Discovery Rate Q_e as the *expected* proportion of falsely rejected null hypotheses:

$$Q_e = E\left(\frac{\mathbf{V}}{\mathbf{R}}\right)$$

The FDR test procedure

A smaller proportion of wrongly rejected hypotheses means less false positives. So we want to ensure the expected proportion, i.e. the FDR in our test procedure does not go above an upper bound, say α. The algorithm below helps achieve that.

Suppose the m hypotheses to be tested are $H_1, ..., H_m$, and the p-values obtained from testing are $P_1, ..., P_m$, respectively. Now we follow these steps:

1. Order the p-values in increasing order: $P_{(1)} \leq P_{(2)} \leq ... \leq P_{(m)}$.
2. Let k be the largest i for which $P_{(i)} \leq \frac{i}{m}\alpha$.
3. Then reject the hypotheses $H_{(1)}, H_{(2)}, ..., H_{(k)}$. Accept other null hypotheses.

An example

- Suppose we have $m = 15$, and we want to test them so that FDR $\leq \alpha = 0.05$.
- The ordered p-values obtained from testing the individual hypotheses are 0.0001, 0.0004, 0.0019, 0.0095, 0.0201, 0.0278, 0.0298, 0.0344, 0.0459, 0.3240, 0.4262, 0.5719, 0.6528, 0.7590 and 1.000.
- Now, varying *i* between 1 and 15, the thresholds the respective p-value needs to be compared with are: 0.0033, 0.0067, 0.01, 0.0133, 0.0167, 0.02, 0.0233, 0.0267, 0.03, 0.0333, 0.0367, 0.04, 0.0433, 0.0467, 0.05.
- Upto $i = 4$ the p-values are less than the threshold.
- Hence we reject the null hypotheses corresponding to these p-values, accept others.

Outline

1. Hypothesis testing: introduction
2. The Family Wise Error approach
3. False Discovery Rate
4. **Applications**
5. Some other methods

Scope of FDR

- Making an overall decision based on testing multiple hypotheses. Example: comparing two treatments/drugs based on their different aspects.
- Making multiple separate decisions without making an overall decision. Example: comparing two drugs in different ethnic groups.
- Obtaining significant factors in an experimental design. By FDR one can control accuracy of the process by limiting the maximum permissible number of factors wrongly deemed significant.

Relevance in Bioinformatics

- Relevant case scenario: identifying differentially expressed genes in two different populations from microarray data.
- FDR provides a method to limit the number of falsely detected significant genes.
- The Family Wise Error approach becomes too conservative for a large number of null hypotheses. FDR has been shown to have more power (i.e. less probability of false negatives) than FWER as number of null hypotheses increases.
- FDR is robust against dependency among genes.

Outline

1. Hypothesis testing: introduction
2. The Family Wise Error approach
3. False Discovery Rate
4. Applications
5. **Some other methods**

Other methods and modifications

- Alternative procedures that perform better than FWER: Schweder and Spjotvol (1982), Hochberg and Benjamini (1990), Storey (2002).
- FDR for dependent data: Benjamini and Yekuteli (2001), Blanchard and Roquain (2009).
- Positive FDR (pFDR) and bayesian approach: Storey (2001), q-value: Storey (2003).
- 'Local FDR': Efron et al (2001).
- Extending FDR for Operating Characteristic curves: Genovese and Wasserman (2002).

THANK YOU!