Primer on multiple testing

Joshua Loftus

July 23, 2015

One hypothesis, many kinds of errors

We have a null hypothesis H_0 which seems reasonable *a priori*. After observing some data, we decide to accept or reject H_0.

- Type 1 (**false positive**) H_0 is actually true but we rejected it.
- Type 2 (**false negative**) H_0 is actually false but we accepted it.
- Type 3? Asking the wrong question, making the right decision for the wrong reason, etc.

Classical statistical decision theory has two goals

- Guarantee that the probability of a Type 1 error is below a pre-specified level α (usually 5%)
- Maximize the *power*, i.e. minimize the probability of Type 2 error, subject to the previous constraint

Many hypotheses, even more kinds of errors

- Type 1 (or 2) errors for each individual hypothesis
- The number of Type 1 errors
- Proportions or rates of Type 1 errors

The **family-wise error rate** (FWER) is the probability of making *any* Type 1 errors at all.

The **false discovery rate** (FDR) is the expected proportion of false rejections out of all rejections.

A simulation example

Consider n normal random variables. Test $H_{0,i} : \mu_i = 0$ vs. $\mu_i > 0$. Truth: first k of them have mean $\mu > 0$, the rest have mean 0.

```r
bunch_of_tests <- function(n, k, mu) {
  stats <- rnorm(n, mean = 0)
  stats[1:k] <- stats[1:k] + mu
  rejections <- which(stats > qnorm(.95))
  # family-wise error
  FWE <- any(rejections > k)
  # false discovery proportion
  FDP <- sum(rejections > k)/max(1,length(rejections))
  # true discovery proportion
  TPP <- sum(rejections <= k)/max(1,k)
  return(c(FWE, FDP, TPP))
}
```

Simulation results $n = 100$, $k = 10$, $\mu = 1$

Perform the testing procedure 1000 times to estimate FDR, etc.

```
results <- replicate(1000, bunch_of_tests(100, 10, 1))
row.names(results) <- c("FWER", "FDR", "TPR")
rowMeans(results)
```

```
##      FWER       FDR       TPR
## 0.9930000 0.6443149 0.2551000
```

This example shows that using many individual tests at level 5% does **not** control FWER or FDR at level 5%.

Simulation results $n = 20$, $k = 10$, $\mu = 2$

```r
results <- replicate(1000, bunch_of_tests(20, 10, 2))
row.names(results) <- c("FWER", "FDR", "TPR")
rowMeans(results)
```

```
##       FWER        FDR        TPR
## 0.39000000 0.06503925 0.63710000
```

If the truth is more favorable, we make fewer errors.

But can we **control** these error rates, making them lower than 5% regardless of whether the truth is favorable?

Bonferroni controls FWER

The **Bonferroni correction** (credit: Olive Jean Dunn in 1959, Carlo Emilio Bonferroni) guarantees FWER $\leq \alpha$ by decreasing the level for all the individual tests to α/n.

$$\mathbb{P}(\text{any Type 1 error}) \leq \sum_{i=1}^{n} \mathbb{P}(\text{Type 1 error for test } i) \leq \sum_{i=1}^{n} \frac{\alpha}{n} = \alpha$$

- Works even if the test statistics are not independent
- Very conservative if n is large
- Can find one very big needle-in-a-haystack, but not many small effects
- The Holm-Bonferroni method has better power

Interlude on p-values

A p-value is...

- a random variable on the interval [0,1]
- distributed like $U[0,1]$ if the null hypothesis is true
- usually smaller if the null hypothesis is false
- i.e. reject if $p < \alpha$
- often transformed from $T \sim F(\cdot)$ to get $p = F(T)$

Many multiple testing procedures begin by sorting all the p-values, since the smallest ones provide the strongest evidence for rejecting their corresponding null hypothesis. Usually we reject the hypotheses with the smallest p-values up to some point, and we just need to decide that stopping point (e.g. Holm-Bonferroni).

Benjamini-Hochberg controls FDR...

The **Benjamini-Hochberg** procedure (1995, initially rejected...)

- Sort the p-values $p_1, \dots, p_n$ to get $p_{(1)} \leq \cdots \leq p_{(n)}$.
- Find the largest k such that $p_{(k)} \leq k \cdot \alpha / n$
- Reject the hypotheses corresponding to $p_{(1)}, \dots, p_{(k)}$

If the p-values are independent then FDR $\leq \alpha$.

If they are not independent, then FDR $\lessapprox \log(n)\alpha$, so we still improve from Bonferroni by using $\alpha / log(n)$ instead of α / n.

Special topic: selective inference

- Motivated by performing inference *after* model selection, e.g. with the Lasso
- Fithian, Sun, Taylor: `http://arxiv.org/abs/1410.2597`
- Suppose we look at the data first and then choose which hypotheses to test
- The *selective* Type 1 error rate is $\mathbb{P}(H_0 \text{ rejected } | H_0 \text{ chosen})$

Conditional probability

Do we need this?

Selection breaks traditional methods

Suppose we begin with *n potential* tests, e.g. we have normal random variables $X_1, \ldots, X_n$ and for each one we could ask if its mean is positive.

Before we perform any tests, we first *select* only the ones that look interesting. For example, suppose that $m < n$ of the X_i have $X_i > 1$. These are the cases that look promising. Call them $Z_1, \ldots Z_m$.

Now do Bonferroni with level α/m instead of α/n. Bonferroni is usually conservative, but will this control anything?

Breaking Bonferroni

```
selected_tests <- function(n) {
  X <- rnorm(n)
  Z <- X[X > 1]
  m <- length(Z)
  rejections <- sum(Z > qnorm(1-.05/m))
  FWE <- as.integer(rejections > 0)
  FDP <- rejections/max(1, m)
  return(c(FWE, FDP))
}
results <- replicate(1000, selected_tests(100))
row.names(results) <- c("FWER", "FDR")
rowMeans(results)
```

```
##       FWER        FDR
## 0.27100000 0.02117014
```

How we fix it

To adjust our tests for selection we use the conditional probability distribution to determine the significance threshold. I.e. instead of *qnorm* we need quantiles of the truncated normal distribution: $Z|Z > 1$.

In general, the kind of truncated distribution depends on the kind of selection method being used. My advisor and his students (including me) have done a lot of work solving various cases, e.g. forward stepwise.

Consultation considerations

- Discuss goals/constraints (e.g. journal standards)
- Caution about multiple testing
- Researchers *need* positive results, be empathic and learn how to be persuasive or they may ignore you
- Remember some convincing examples and explanations
- If they are fooled by randomness it could be embarassing in the long run even if they get published in the short run