Controlling False Discoveries Rates

Agus Salim, Ph.D

Big Data Analytics: Application to Modern Genetics
Universitas Jember, 23 Dec 2015

A lot of variables, few samples

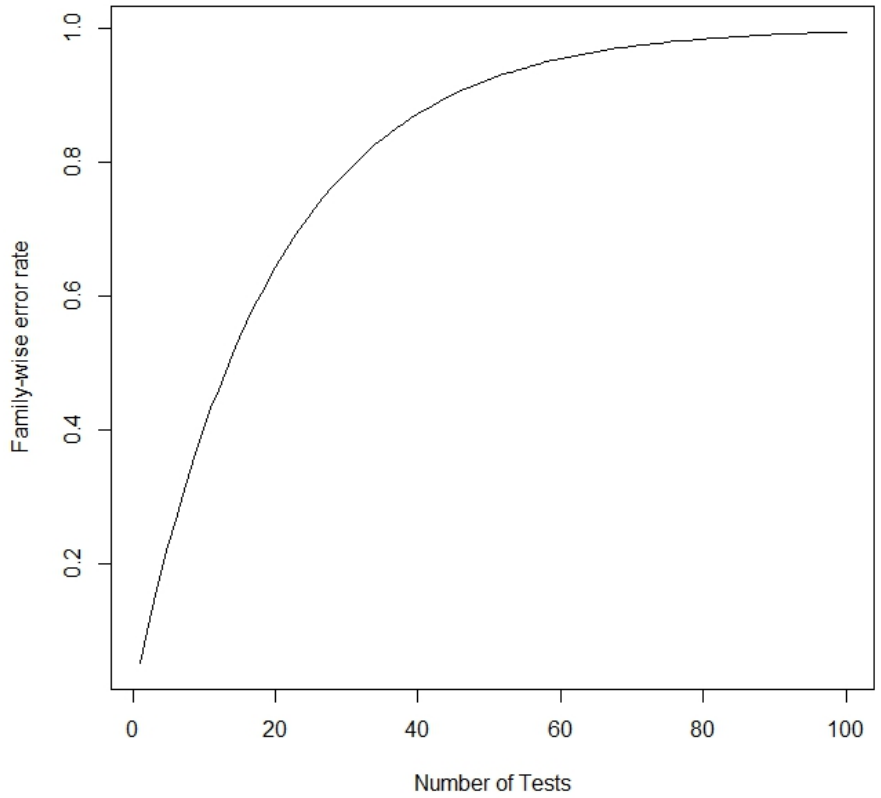
- With the advance in omics (e.g. genomics, proteomics) technology, it is increasingly common to measure thousands of variables (e.g., protein levels) simultaneously on relatively few subjects (less than 10 is common!).
- This is $n \ll p$ problem, for which a lot of statistical methods will not work
- Usually, the subjects were allocated one of the two treatments on offer and statistical test is performed to compare the two groups.
- There will be m separate tests and each test is performed using set Type I error probability $= \alpha$.

Family-wise error rate

- Type I error = Falsely rejecting the null hypothesis
- α = P(Reject H_0 | H_0 is true)
- Let us define V as the number of times we reject H_0 when it is actually true.
- We are interested in the overall (family-wise) error rate, $P(V \geq 1)$.
- If we perform only one test, $P(V \geq 1) = P(V = 1) = \alpha$
- If we perform two tests $P(V \geq 1) = P(V = 1) + P(V = 2) = 1 - P(V = 0) = 1 - (1 - \alpha)^2$
- If we perform m tests, $P(V \geq 1) = 1 - P(V = 0) = 1 - (1 - \alpha)^m$

Family-wise error rate, $\alpha = 0.05$

- When $m = 20$, the error-rate is approximately 0.6.
- At $m = 100$, the error-rate is approximately 1 (we are guaranteed an error!).



Bonferroni Correction

- To achieve family-wise error rate $= \alpha$, significance levels for individual test needs to be set at α/m.
- $P(V \geq 1) = 1 - (1 - \frac{\alpha}{m})^m \approx \sum_{i=1}^{m} \frac{\alpha}{m} = \alpha$
- Essentially, for decision-making purpose (reject/accept), the original p-value is adjusted by multiplying it with the number of hypothesis tested (m) and the adjusted p-value is compared to α.

$$p_{adj}^{bonferroni} = m \times (pvalue)$$

- This controls the family-wise error well, but it will be extremely difficult to have positive discovery (i.e., to reject H_0). Do you know why?

Summary of Test Outcomes

	Null True	Alt True	Total
Not called significant	U	T	$m - R$
Called Significant	V	S	R
	m_0	$m - m_0$	m

- Family-wise error rate concerns with at least making one false discovery ($V \geq 1$)
- Instead of being concerned with making at least one false discovery, we can accept that we will make false discoveries anyway, but we will be happy if the *false discovery rates (FDR)* can be controlled.
- $FDR = E\left(\frac{V}{R}\right)$, the expected proportion of false discoveries among the rejected hypothesis.

Estimating False Discovery Rates (Storey and Tibshirani, *PNAS* 2003: 9440-5)

- Suppose that U is the random variable representing the p-value and we reject all hypothesis where the observed p-value is $\leq u$

$$\begin{aligned} FDR(u) &= P(H_0 \mid U \leq u) \\ &= \frac{P(U \leq u, H_0)}{P(U \leq u)} \\ &= \frac{P(H_0)P(U \leq u \mid H_0)}{P(U \leq u)} \end{aligned}$$

- Note: under H_0 (equality), the pvalue is distributed as $U(0,1)$ r.v (perform simulation to prove it yourself!), so $P(U \leq u \mid H_0) = u$.
- $P(H_0)$ is the probability of null hypothesis being true $= \pi_0$.

Estimating False Discovery Rates

- Hence, we can write

$$FDR(u) = \frac{\pi_0 \times u}{P(U \leq u)}$$

$P(U \leq u)$ is the CDF of observed p-values at u

Estimating False Discovery Rates

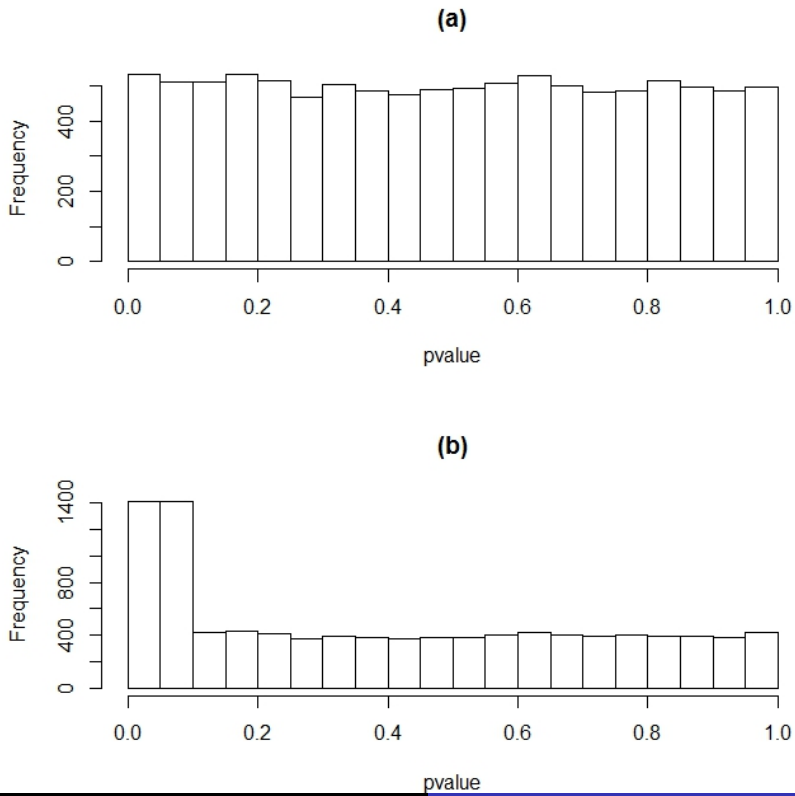
- Let $p_1, p_2, \ldots, p_m$ be the observed p-values and $p_{(1)}, p_{(2)}, \ldots p_{(m)}$ be the ordered p-values, so that $p_{(i)}$ is the ith most significant p-value.
- What is $FDR(p_{(i)})$?
- To compute FDR, we need $P(U \leq p_{(i)})$, using Glivenko-Cantelli Lemma, we can estimate this probability using its empirical version, $P^{emp}(U \leq p_{(i)}) = \frac{i}{m}$.
- Hence,

$$FDR(p_{(i)}) = \frac{m\pi_0 \times p_{(i)}}{i}$$

- We can now estimate FDR, if we know π_0.

Estimating π_0

- If null hypothesis is true for all tests, then the p-value distribution is uniform(panel (a)).
- When there are some tests for which the alternative is true, the p-value distribution is skewed to the left and there is a 'deficit' on the right-hand side (panel (b)).



Estimating π_0

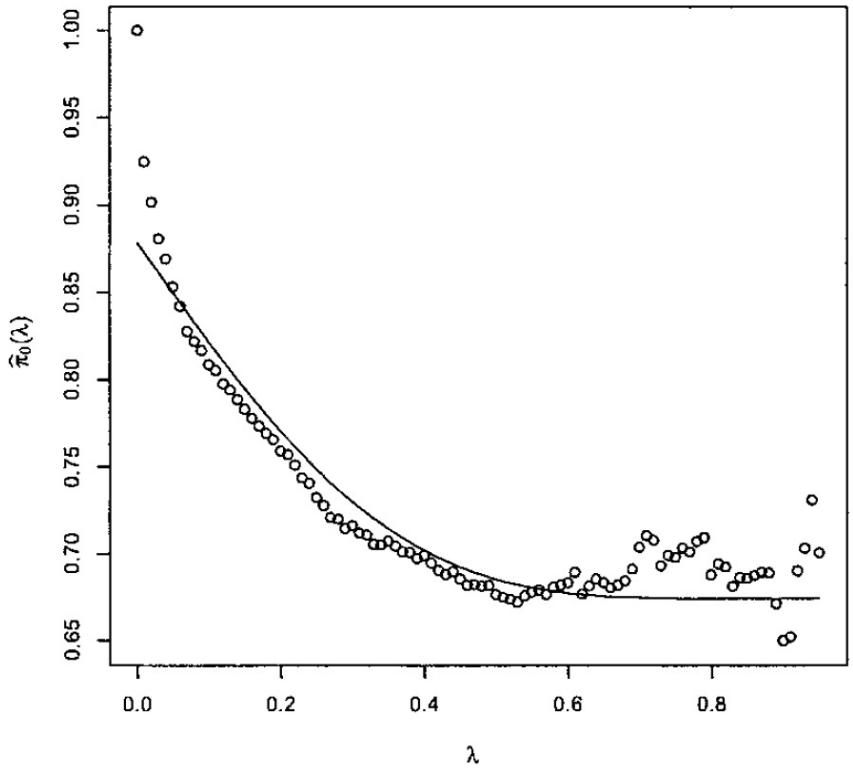
- Storey and Tibshirani (2003) proposed to estimate π_0 as the number of observed p-values $\geq \lambda$ divided by the expected number of p-values $\geq \lambda$ if the null hypothesis is always true.
- The expected number of p-values $\geq \lambda$ is null hypothesis is always true $= m(1-\lambda)$

$$\hat{\pi}_0(\lambda) = \frac{num(p_i \geq \lambda)}{m(1-\lambda)}$$

- The estimate depends on choice of λ. Ideally, we want to choose λ close to 1 as we are almost entirely sure that null hypothesis is true when p-value $=1$, but there will be very few tests with p-value ≈ 1, so the estimate will be unstable.

Estimating π_0

- Solution: estimate $\hat{\pi}_0(\lambda)$ for various values of λ and fit smoothing spline (or quadratic model) with $\hat{\pi}_0(\lambda)$ as response and λ as predictor (see Figure below from Storey and Tibshirani (2003)).
- Use the estimated model to calculate $\hat{\pi}_0 = \hat{\pi}_0(1)$!



Estimating False Discovery Rates

- We can estimate FDR for ith most significant p-value as

$$FDR(p_{(i)}) = \frac{m\hat{\pi}_0 \times p_{(i)}}{i}$$

- We are almost done, but the FDR estimates can be non-monotone. Storey fixed this by calculating the q-value (a monotone version of FDR).
- We can now declare all genes with FDR $\leq \alpha$ as statistically significant and only $100\alpha\%$ of these declared genes are expected to be 'false'.

Some Useful References

- Benjamini Y, Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSS B* **57:** 289-300.
- Pawitan Y, Murthy KR, Michiels S, Ploner A. (2005). Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics* **21:** 3865-72.
- Storey JD, Tibshirani R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **100:**9440-5.