Multiple testing and the false discovery rate

Śaunak Sen

2019-05-21

Seminar series on statistical reasoning in biomedical research

- Apr 30: P-values: What they are and what they are not (Fridtjof Thomas, PhD)
- May 07: Should We eliminate P-Values or Use More of Them: A Discussion on the P-Value Controversy (Saunak Sen, PhD)
- May 14: The Bayesian Approach to Data Analysis (Fridtjof Thomas, PhD)
- May 21: Multiple Testing and the False Discovery Rate (Saunak Sen, PhD)
- May 28: The Perfect Doctor: An introduction to Causal Inference (Fridtjof Thomas, PhD)
- Jun 04: Enhancing Statistical Methods in Grants and Papers (Saunak Sen, PhD)

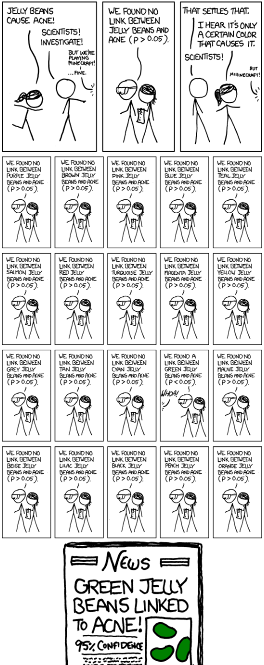
Outline

Should we adjust for multiple comparisons?

Family-wise error rate (FWER)

False discovery rate (FDR)

XKCD


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE

To adjust or not?

Consider a thought experiment.

Scenario 1: Postdoc A tests association between gene A and phenotype; publishes paper. Postdoc B tests association between gene B and phenotype; publishes paper. Postdoc C tests association between gene C and phenotype; publishes paper. And so on, for Postdoc Y. Each paper stands on own.

Scenario 2: Postdoc Z uses high-throughput technology to test association of genes A-Y and same phenotype. Does postdoc Z have to adjust for multiple comparisons? Should Z be penalized for using fancier technology?

Examples

- Deviant search from mutagenesis
- Genome-wide association studies
- Differential gene expression using RNAseq
- Compare lipid profile of subjects with and without statins
- Compare microbiome of obese and healthy individuals

Family-wise error rate (FWER)

Appropriate when multiple tests are used to test a single hypothesis.

- Bonferroni procedure
- Holm procedure
- Fisher combination procedure

Bonferroni procedure

If we have m ordered p-values $p_1 \leq p_2 \leq ... \leq p_m$, then the FWER corrected p-value is $m \times p_1$.

For example, if we perform three tests and get p-values 0.1, 0.04, and 0.01, and we want to control FWER to be 5% or less, then we will reject the null hypothesis only if the smallest p-value is 0.05/3 or less. In this case, one p-value (0.01) is smaller, and therefore we can reject the null hypothesis controlling the FWER at 5%. The Bonferroni-corrected p-value of the *family* of tests is $0.01{\times}3{=}0.03$.

- Simple and easy to conduct
- No assumption on dependence between tests
- Low power if m is large

Fisher combination procedure

Suppose we have m p-values $p_1, p_2, \ldots, p_m$ from *independent* tests. Let $T = -2\sum \ln(p_i)$. This is called the combination statistic.

Compare this to a χ^2 distribution with $2m$ degrees of freedom to get the combination p-value for the family of tests.

In the example considered above the combination statistic is $T = 2(2.30 + 3.22 + 4.61) = 20.26$ which gives a p-value of 0.0025 when compared to a χ^2 distribution with 6 degrees of freedom.

- More complex than Bonferroni
- Assumes independence of tests
- More power than Bonferroni
- Weights p-values equally regardless of information content

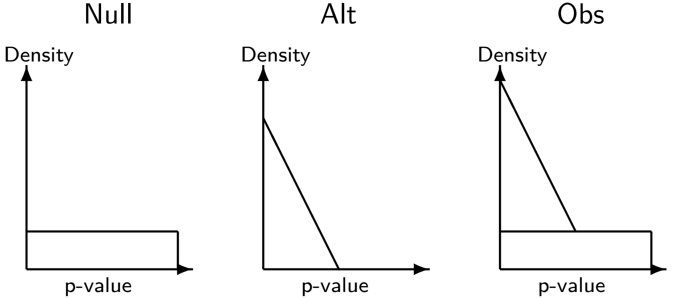
False discovery rate (FDR)

Appropriate when we want to evaluate a large number of similar hypotheses individually.

We want to estimate the probability that a “discovery” is false, given that we have a discovery (low p-value).

Note similarity with diagnostic tests: we want to estimate the chance that a discovery (postitive test result) is false (from a normal sample).

P-value distributions under null, alternative, and observed



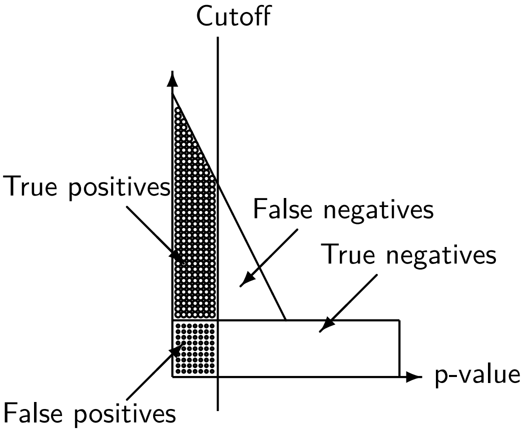
$F_0(t)$ $\quad$ $F_1(t)$ $\quad$ $\pi_0 F_0(t) + (1-\pi_0) F_1(t)$

Analogy with diagnostic tests: The diagnostic test is based on the p-value. The characteristics of the diagnostic test under null (normal sample, F_0) is known, and the alternative (diseased sample, F_1) is partially known. The disease prevalence $(1-\pi_0)$ is not known.

Actual and declared true/false hypotheses

		Declared		
		Null	Alt	Total
Actual	Null	U_t	V_t	m_0
	Alt	T_t	S_t	m_1
	Total	W_t	R_t	m

Cartoon


Cutoff
True positives
False negatives
True negatives
p-value
False positives

Two approaches to FDR

There are two main approaches. The first, due to Benjamini and Hochberg seeks to find a cutoff, t, given a target proportion α, such that

$$E\left(\frac{V_t}{R_t}, R_t > 0\right) \leq \alpha.$$

The second, due to Storey, and closely related to that of Efron and Tibshirani, seeks to find the *q-value,* for a fixed cutoff t such that

$$q_t = P\left(\frac{V_t}{R_t} | R_t > 0\right).$$

Thus, the first approach estimates a cutoff given a target *FDR;* the second estimates the *FDR* given a cutoff.

False discovery rate (FDR) given cutoff (q-value)

The q-value, or FDR corresponding to a cutoff of t is the probability of a false discovery given that there has been a discovery (a p-value under the cutoff).

We can use Bayes theorem for that.

$$P(H_0|T\leq t) = \frac{P(H_0)p(T\leq t|H_0)}{P(T\leq t)}$$

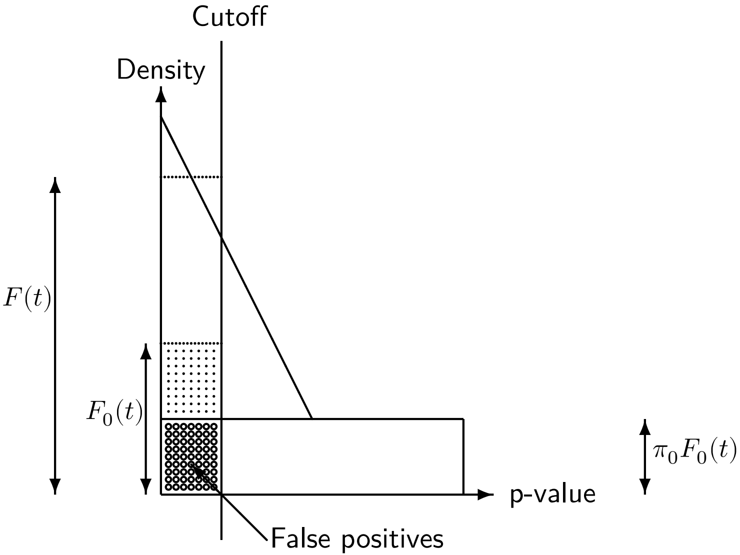
$$q(t) = \pi_0 F_0(t)/F(t),$$

where $F = \pi_0 F_0 + \pi_1 F_1$. We can estimate $F(t)$ from the empirical distribution of p-values, and $F_0(t) = t$ since F_0 is uniform. Thus

$$\widehat{q(t)} = \pi_0 t/\widehat{F(t)} \leq t/\widehat{F(t)}.$$

Sharper estimates possible by estimating π_0.

Cartoon



Get cutoff given desired False discovery rate (FDR)

Let $p_1, p_2, \ldots, p_m$ be the ordered p-values. If π_0 is the proportion of null hypotheses, then the cutoff t such that

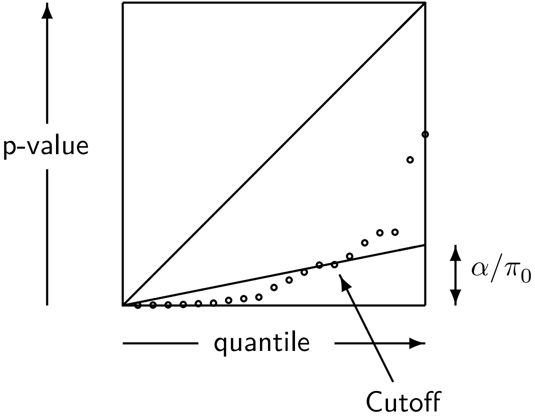
$$E\left(\frac{V_t}{R_t}, R_t > 0\right) \leq \alpha$$

is

$$t = \max\left\{p_{(i)} : p_{(i)} \leq \left(\frac{i}{m}\right)\left(\frac{\alpha}{\pi_0}\right)\right\}.$$

In practice, we do not know π_0, so a conservative choice is $\pi_0 = 1$.

Benjamini-Hochberg procedure



Deviant search example

Battery of 6 tests (vertical movement, ambulatory movement, RER; in dark and light phases) per mouse. A FDR of 10% was desired.

	Number		FDR		P-value	Bonferroni	
	Mice	Tests	Mice	Tests	Cutoff	Mice	Tests
Control training	24	144	0	0	0	0	0
Control Test	24	144	2	6	0.0016	2	6
Mutant Test	22	132	17	49	0.0371	12	28

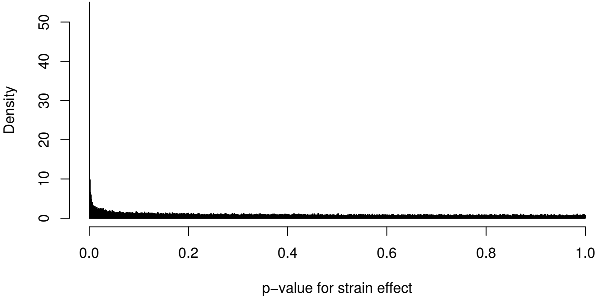
Gene expression example

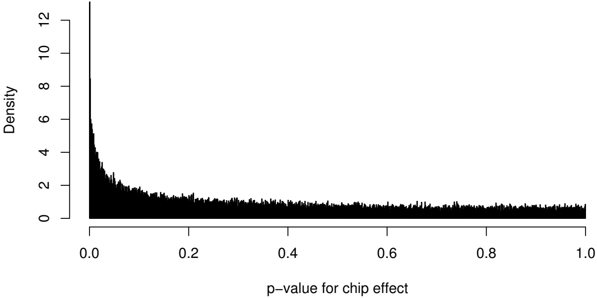
Compare genomewide gene expression between two mouse strains in spinal cord tissue samples. The Illumina mouse arrays that we used can hybridize up to six samples on the same "chip". Spinal cord mRNA from 6 B6 mice, and 6 LP/J mice. Two chips were hybridized. In each chip mRMA hybridized to three B6 and three LP/J samples. A total of 48358 probes. Our goal is to find out which genes (probes) are differentially expressed between strains.

f we perform a Bonferroni correction to the p-values, then only genes with p-values smaller than $0.05/48358 \simeq 10^{-6}$ would be considered significant. Only about 654 genes fit this bill.

However, if we estimate the q-values, and find genes with a q-value smaller than 0.05, then we find 4163 genes. If we look for genes with a q-value smaller than 0.01, we find 2459 genes. An estimated 27% of the genes show differential expression.

P-value distributions



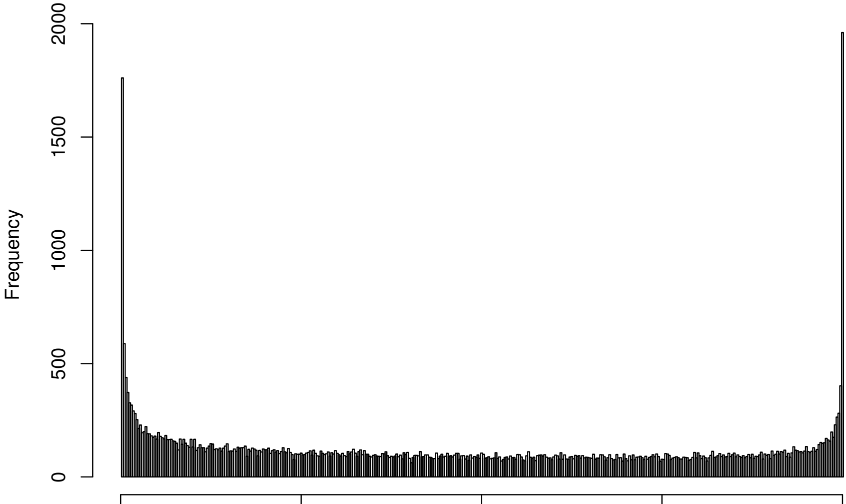


Using permutation

An alternative analysis would be to use a permutation test, by permuting the strain labels *within* each chip. The permutations allow us to obtain the distribution of the test statistic under the hypothesis of no strain effect, but in the presence of a possible chip effect. In this data, we have only $\binom{6}{3} = 20$ permutations possible. By combining the permutations in both chips, a total of 400 permutations are possible. We exhaustively enumerated each of them and calculated the average strain effect within chip. This was then compared to the observed average strain effect within chip and ranked.

For 1761 probes the observed effect was ranked 1, and for 1961 probes it was ranked 400. Thus the two possible extreme ranks were observed 3722 times compared to the expectation of $48358/200 \simeq 242$ under the null. Thus the genes with rank 1 or 400 have a q-value of at most $242/3722 \simeq 0.065$.

P-value distributions from permutations (exhaustive enumeration)



Notes

- Given a cutoff, t, the q-value is the proportion of null hypotheses among the discoveries. It is one minus the positive predictive value.
- The p-value under the null must be correctly calibrated, otherwise the FDR estimates are incorrect.
- The p-value distribution under the alternative is not needed, but it is assumed that small p-values are more likely under the alternative.
- Estimates of the proportion of null hypotheses (π_0) affects the q-value (and FDR), just like the disease prevalence affects the positive predictive value.
- Independence of the tests (p-values) is not required as long as we can get good estimates of the p-value distributions.
- There is an implicit understanding that the hypotheses/tests are exchangable (or similar).

Summary

- Consider research question carefully to decide if multiple comparisons adjustment is needed, and if so, what kind
- FWER appropriate when many tests are used to test a single hypothesis
- FDR appropriate when we want to evaluate a large number of similar hypotheses individually
- Be transparent regarding process and assumptions
- Report all p-values so that others can replicate or adjust assumptions

Further reading

- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561-576. URL.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65-70. URL
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, *57*, 289-300. URL
- Storey JD. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. Annals of Statistics, 31: 2013-2035. URL

R functions:

- `stats::p.adjust`
- `qvalue::qvalue`

Upcoming

- May 28: The Perfect Doctor: An introduction to Causal Inference (Fridtjof Thomas, PhD)
- Jun 04: Enhancing Statistical Methods in Grants and Papers (Saunak Sen, PhD)

Slides at https://tnctsi.uthsc.edu