The value of p-values

Gordon Smyth
Walter+Eliza Hall Institute of Medical Research

FOAM
28 March 2014

1

askville.amazon.com

Question:

Can you explain the concept of "p-value" to me in simple English?

Answer:

The p-value is the probability that your null hypothesis is actually correct.

Statistician's reaction: Noooooooo!

2

A p-value is a measure of surprise!

An exercise from a large statistics course for biologists:

I take an coin from my pocket. I toss it in the air, catch it, and show it to you. It's a head.

I toss it again. It's a head.

I toss it again. It's a head.

I toss it again. It's a head.

I toss it again. It's a head.

3

How many heads?

How many heads is a row would you have to see before you starting thinking this is not a normal coin but a fake coin with two heads?

One?
Two?
Three?
Four?
Five?
Six?
Seven?
Eight?

4

What's the p-value?

H_0: coin is fair
H_a: not fair

Number of Heads	P-value
4	0.125
5	0.0625
6	0.031
7	0.016
8	0.0078

In class, 6 was the most popular choice, with some students requiring 7 or 8

5

Fisher on p-values

Fisher's earliest and clearest statement on significance was in an expository paper:

Fisher, R.A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*

He considered a hypothetical field trial in which a new manure treatment resulted in a 10% improvement in crop yield.

He says …

6

The idea of statistical significance

If the experimenter could say that in 20 years experience with uniform treatment the difference in favour of the acre treated with manure had never before touched 10%, the evidence would have reached a point which may called the verge of significance; for it is convenient to draw the line at about the level at which we can say:

"Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in 20 trials."

7

Why 0.05?

If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2% point), or one in a hundred (the 1% point).

Personally, the writer prefers to set a low standard of significance at the 5% point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance

8

Personally, the writer prefers to set a low standard of significance at the 5% point, and ignore entirely all results which fail to reach this level.

The 5% cut off is an initial screen, chosen low to reduce false negatives.

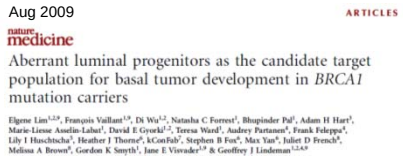
A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.

It's about confirmation and accumulation of evidence over multiple studies over time.

9

Confirmation in biomed practice

Proposed cell of origin for breast cancer


Aug 2009 ARTICLES
nature medicine
Aberrant luminal progenitors as the candidate target population for basal tumor development in *BRCA1* mutation carriers

Physical confirmation


Sep 2010
Cell Stem Cell
Article
***BRCA1* Basal-like Breast Cancers Originate from Luminal Epithelial Progenitors and Not from Basal Stem Cells**

10

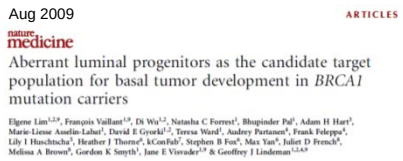
Studies require internal validation

- In molecular biology research, a single p-value never stands by itself
- Results are expected to have a variety of supporting evidence
- Positive and negative controls
- Important genomic results are expected to be validated on more than one technology: e.g., microarray, RNA-seq, RT-PCR

11

Example of internal validation

Proposed cell of origin for breast cancer


Aug 2009 ARTICLES
nature medicine
Aberrant luminal progenitors as the candidate target population for basal tumor development in *BRCA1* mutation carriers

- Genomic analysis linked gene expression patterns in luminal progenitor cells with that of basal breast tumours
- Also showed that women with BRCA1 mutations produce abnormally many luminal progenitor cells even before they get cancer

12

Multiple testing

In genomic research, we report on tens of thousands of tests (at very least) for each paper.

Doing thousands of tests at 5% significance would lead to an unacceptable number of false positives.

There are two main approaches …

13

Family-wise error rate

For modest sized experiments (up to a few dozen tests), can control the probability of any false positive.

Sort p-values from smallest to largest:

	p_1	p_2	$\cdots$	p_n
Bonferroni:	$\times n$	$\times n$	$\cdots$	$\times n$
Holm:	$\times n$	$\times (n-1)$	$\cdots$	$\times 1$
Simes:	$\times n$	$\times \frac{n}{2}$	$\cdots$	$\times 1$

14

The second approach is one more suited to genomic research, and called is the

"false discovery rate"

This brings us back to …

the probability that the null hypothesis is actually correct.

15

Probability null hypothesis is true

p-value

prior

$$P(H_0 \mid \text{data}) = \frac{P(\text{data} \mid H_0) P(H_0)}{P(\text{data})}$$

marginal distribution

16

Probability the hypothesis is true

In genomic research, we want to find genes that have specific behaviour, so $P(H_0)$ is typically large, and we can drop it:

$$P(H_0 \mid \text{data}) = \frac{P(\text{data} \mid H_0) P(H_0)}{P(\text{data})} \approx 1$$

$$\leq \frac{P(\text{data} \mid H_0)}{P(\text{data})}$$

17

False discovery rate

In genomic research, we have thousands of p-values, so can use empirical marginal distribution:

$$P(H_{0i} \mid \text{data}) \leq \frac{P(\text{data} \mid H_{0i})}{P(\text{data})}$$

$$= \frac{p_i}{P(\text{data}_i)}$$ ordered p-value

$$= \frac{p_i}{i/n}$$

which is equivalent to the Benjamini-Hochberg procedure for controlling the false discovery rate.

18

FDR can be controlled over the long term

- The family-wise error rate is not scalable – one can't control it over a career, or even over a large experiment.
- FDR is scalable: controlling FDR for each individual experiment also gives longer term assurance

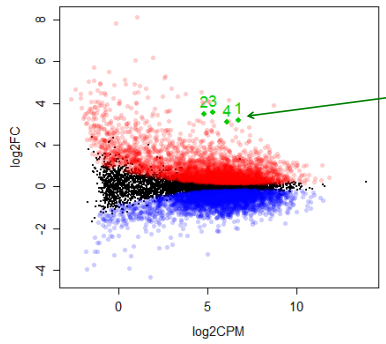
19

Significance vs effect size

- Consider the problem of identifying differentially expressed genes between two groups (using microarray or RNA-seq data)
- Ranking genes by p-value doesn't necessarily rank by biological importance
- Ranking by fold-change tends to be more reproducible between experiments

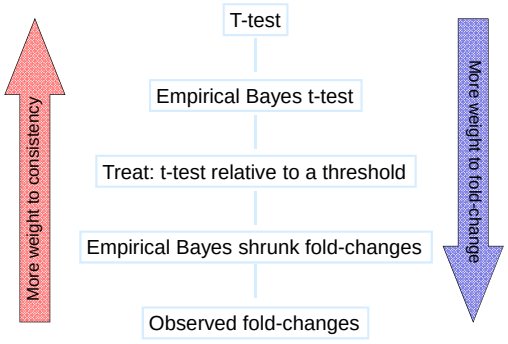
20

Most significant genes don't have largest fold-changes



21

Strategies for ranking genes



22

Summary

- P-values are a tremendously useful practical tool
- Never make a decision on one p-value alone
- FDR better than p-value for genomic problems
- P-values are a building block in the larger scientific puzzle

23

Acknowledgements

- Yifang Hu
- Ross Dickins
- Grace Liu
- Geoff Lindeman
- Jane Visvader

24