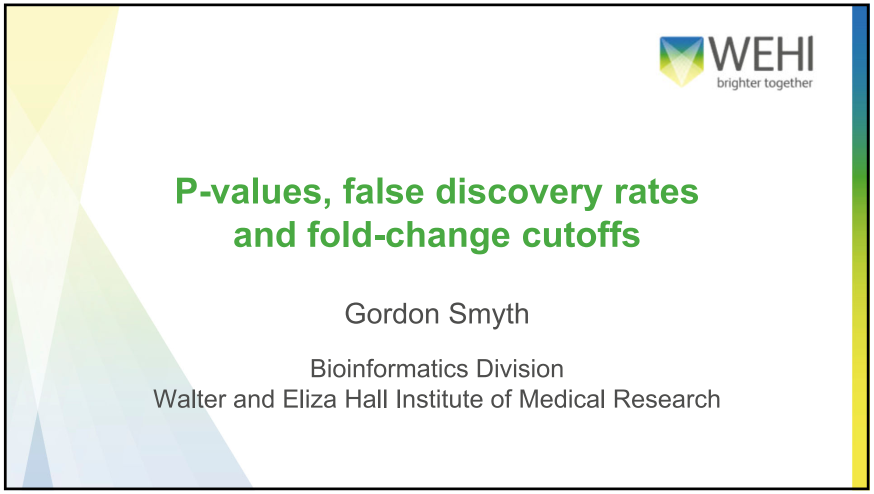


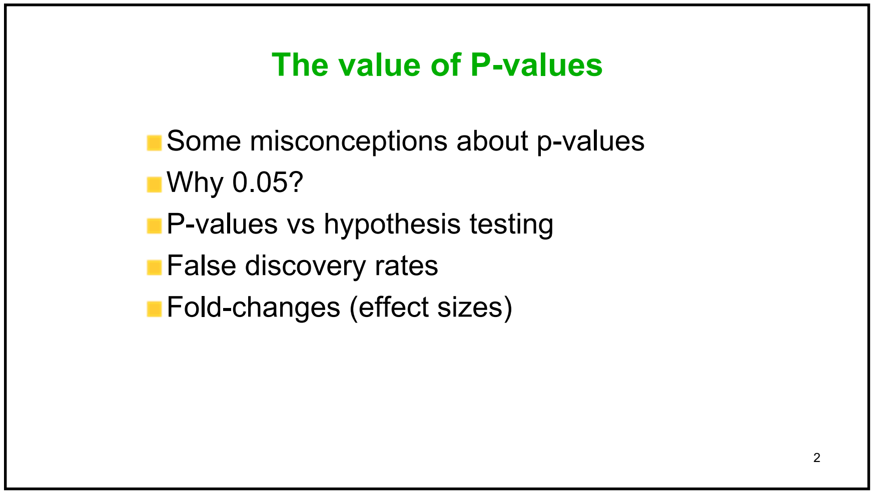
WEHI
brighter together

P-values, false discovery rates and fold-change cutoffs

Gordon Smyth

Bioinformatics Division
Walter and Eliza Hall Institute of Medical Research

1

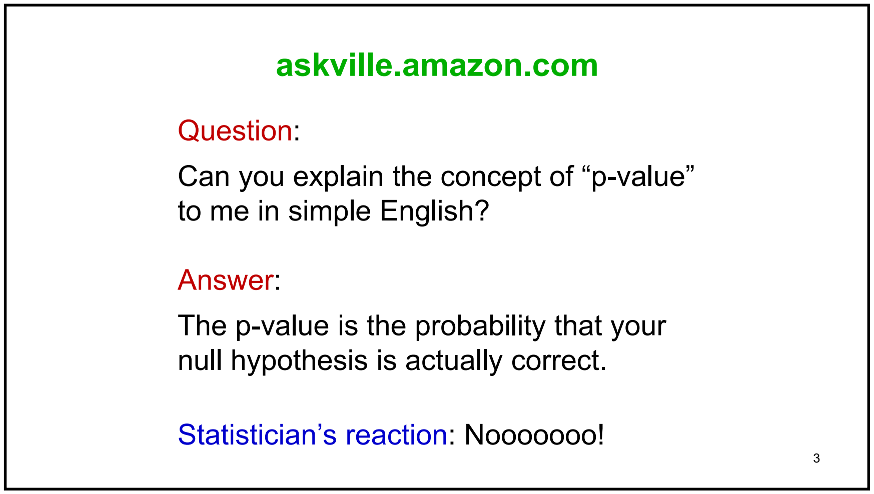


The value of P-values

- Some misconceptions about p-values
- Why 0.05?
- P-values vs hypothesis testing
- False discovery rates
- Fold-changes (effect sizes)

2

2



askville.amazon.com

Question:

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

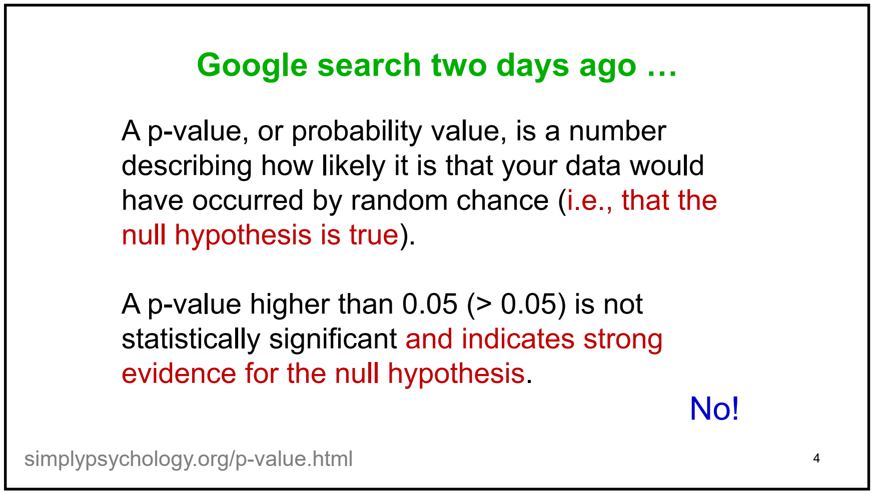
Answer:

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

Statistician’s reaction: Nooooooo!

3

3



Google search two days ago …

A p-value, or probability value, is a number describing how likely it is that your data would have occurred by random chance (i.e., that the null hypothesis is true).

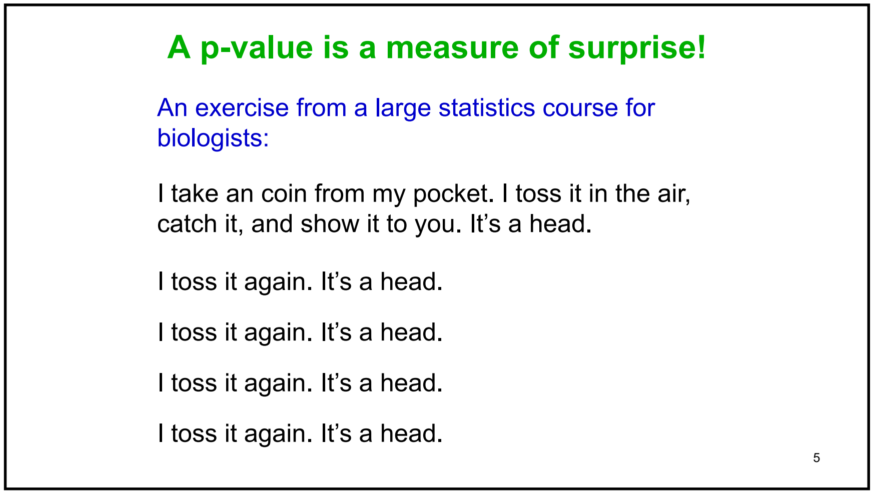
A p-value higher than 0.05 (> 0.05) is not statistically significant and indicates strong evidence for the null hypothesis.

No!

simplypsychology.org/p-value.html

4

4



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.

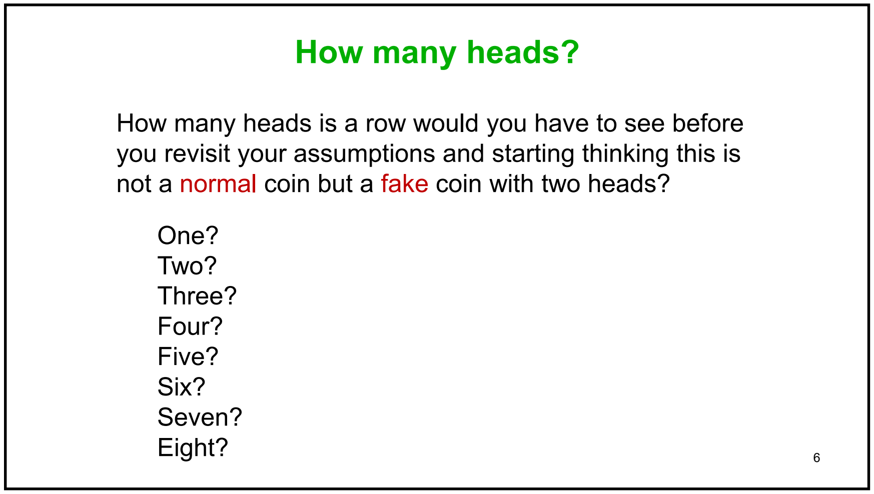
I toss it again. It’s a head.

I toss it again. It’s a head.

I toss it again. It’s a head.

I toss it again. It’s a head.

5

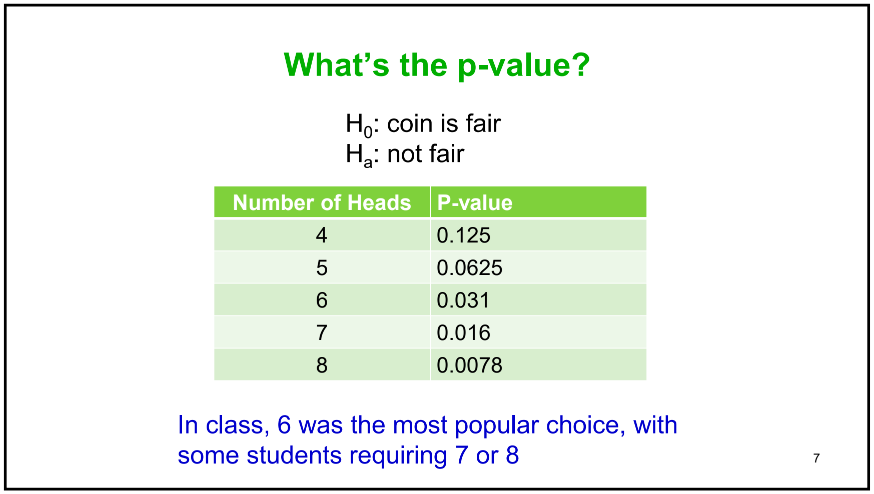


How many heads?

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

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

6



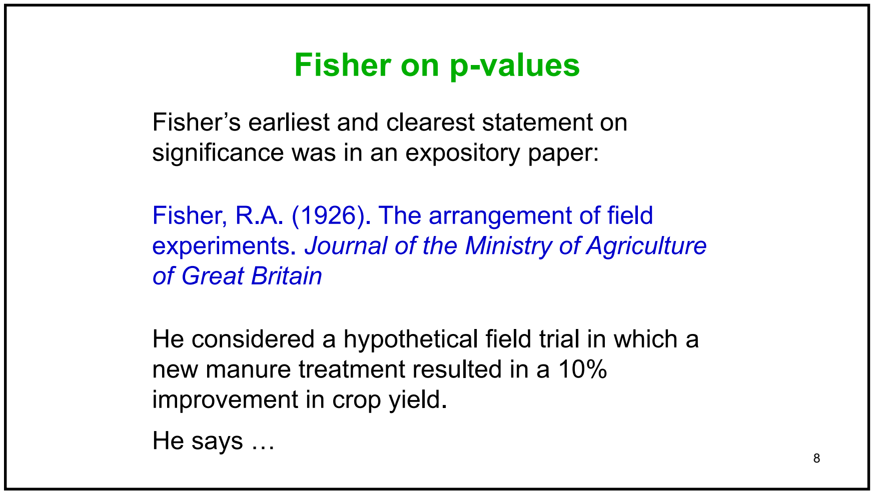
What’s the p-value?

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

7



Fisher on p-values

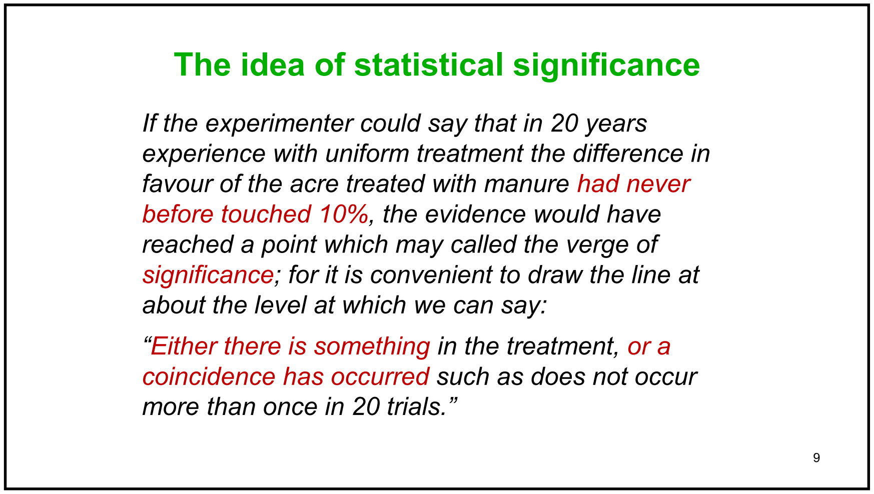
Fisher’s earliest and clearest statement on significance was in an expository paper:

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 …

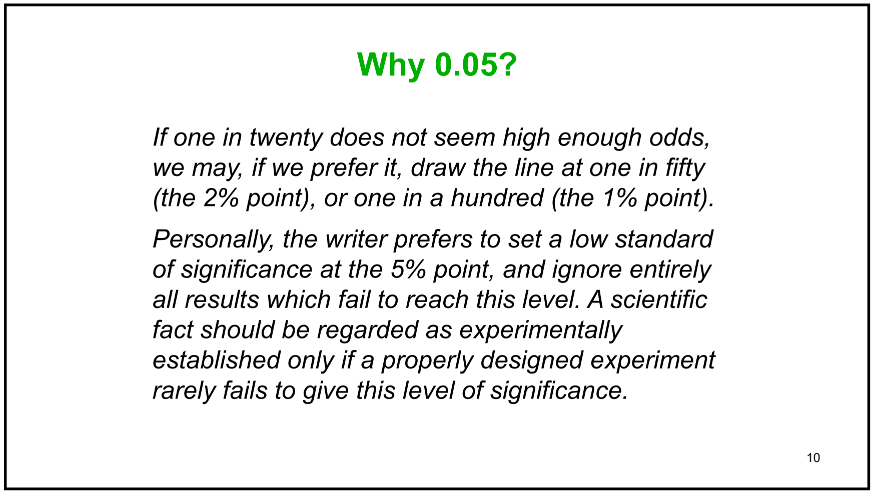
8



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:

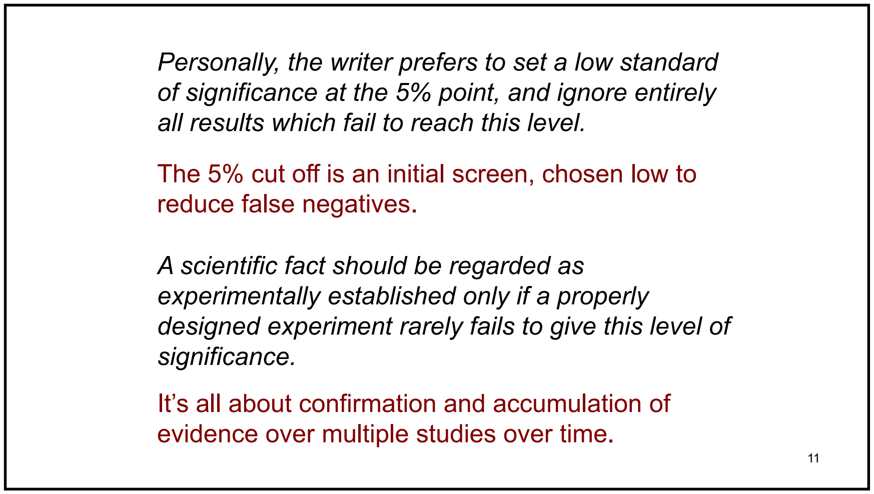
"Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in 20 trials."



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.

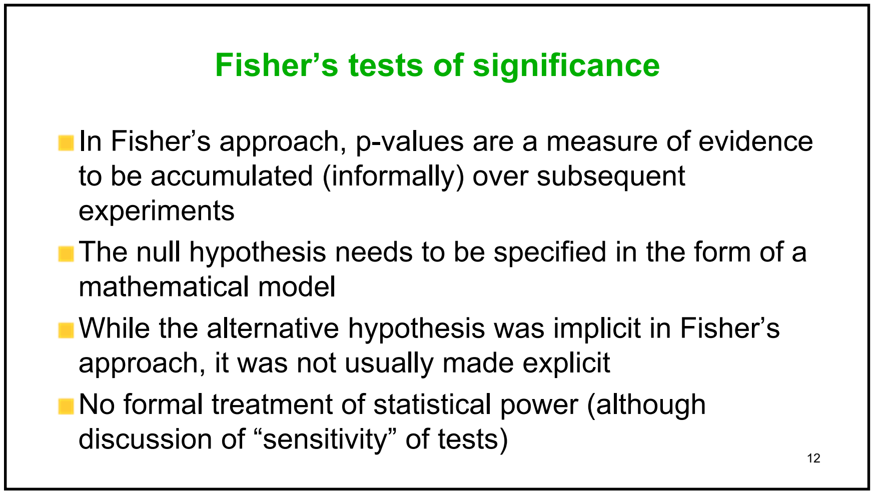


Personally, the writer prefers to set a low standard of significance at the 5% point, and ignore entirely all results which fail to reach this level.

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

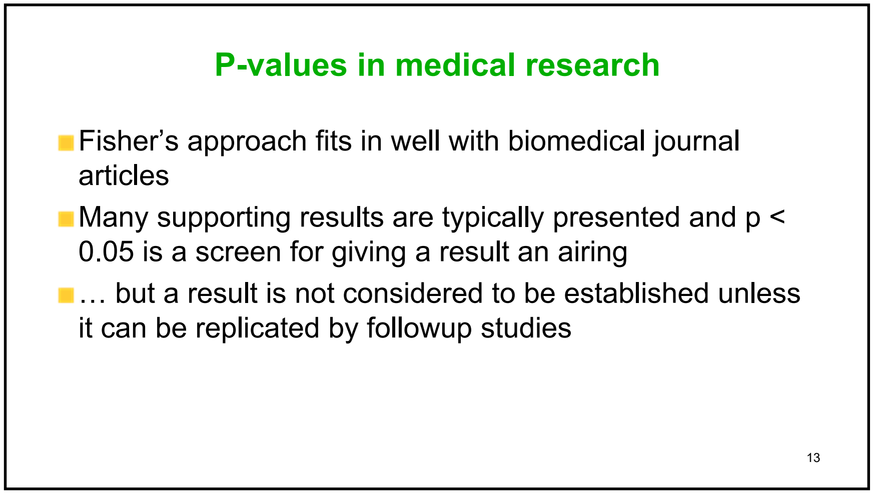
A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.

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



Fisher's tests of significance

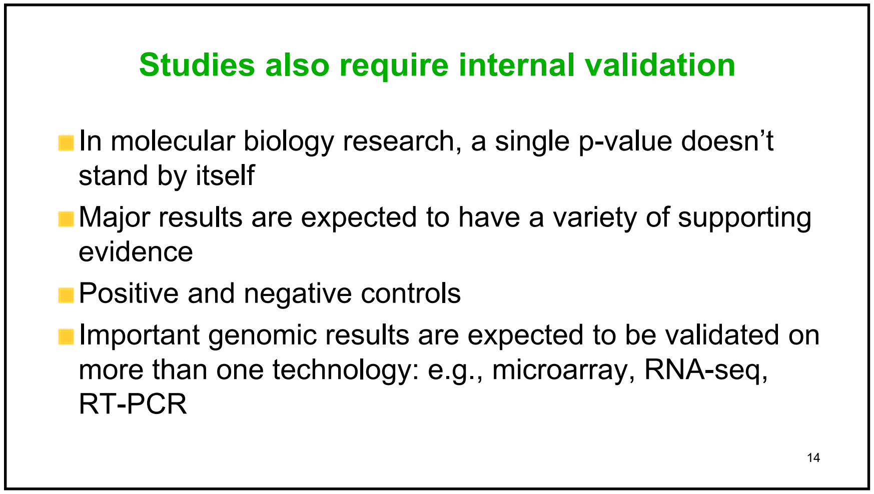
- In Fisher's approach, p-values are a measure of evidence to be accumulated (informally) over subsequent experiments
- The null hypothesis needs to be specified in the form of a mathematical model
- While the alternative hypothesis was implicit in Fisher's approach, it was not usually made explicit
- No formal treatment of statistical power (although discussion of "sensitivity" of tests)



P-values in medical research

- Fisher’s approach fits in well with biomedical journal articles
- Many supporting results are typically presented and p < 0.05 is a screen for giving a result an airing
- … but a result is not considered to be established unless it can be replicated by followup studies

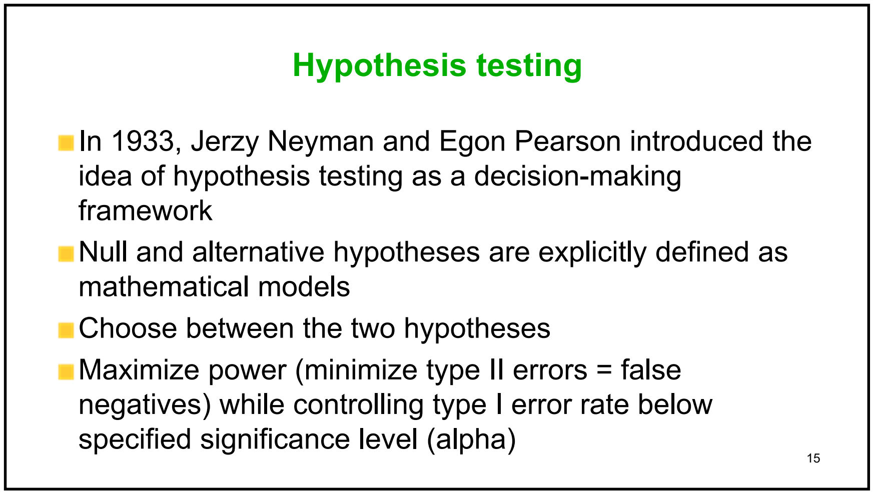
13



Studies also require internal validation

- In molecular biology research, a single p-value doesn’t stand by itself
- Major 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

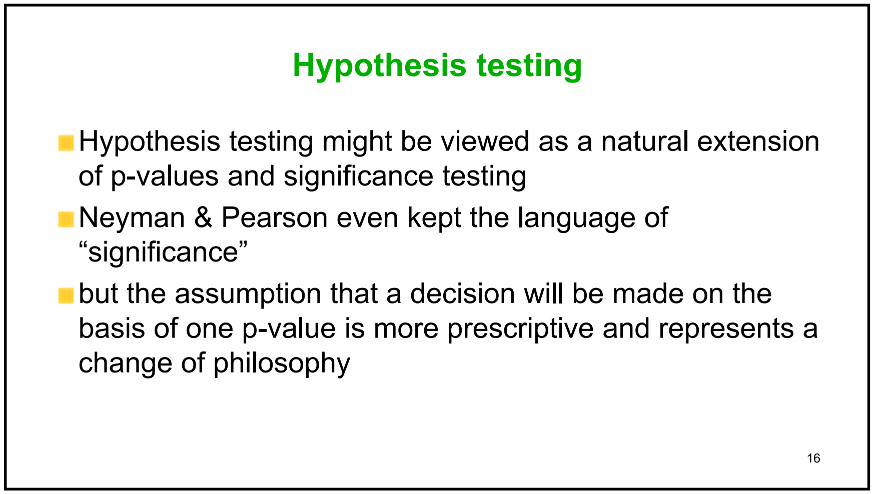
14



Hypothesis testing

- In 1933, Jerzy Neyman and Egon Pearson introduced the idea of hypothesis testing as a decision-making framework
- Null and alternative hypotheses are explicitly defined as mathematical models
- Choose between the two hypotheses
- Maximize power (minimize type II errors = false negatives) while controlling type I error rate below specified significance level (alpha)

15



Hypothesis testing

- Hypothesis testing might be viewed as a natural extension of p-values and significance testing
- Neyman & Pearson even kept the language of “significance”
- but the assumption that a decision will be made on the basis of one p-value is more prescriptive and represents a change of philosophy

16

Likelihood ratio

The Neyman-Pearson lemma proved that the likelihood ratio was the most powerful statistic for comparing simple hypotheses

> *If we show that the frequency of accepting a false hypothesis is minimum when we use (likelihood) tests, I think it will be quite a thing*
>
> (Neyman in a letter to Pearson)

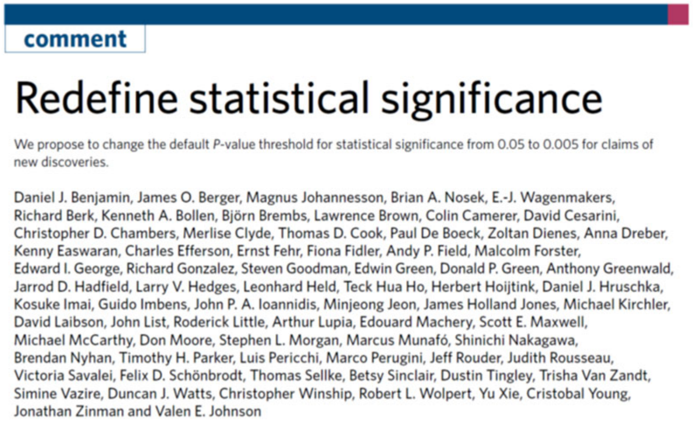
17

Fisher's reaction

Fisher 1955 on hypothesis testing:

> *[The unfounded presumptions] would scarcely have been possible without that insulation from all living contact with the natural sciences, which is a disconcerting feature of many mathematical departments*

18



comment

Redefine statistical significance

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

19

Historical figures

Francis Galton. Bivariate normal, correlation (r), regression to mean, standard deviation.

Karl Pearson. PCA, chi-squared test.

William Gosset. t-test.

Ronald Fisher. ANOVA, variance, design of experiments, significance testing, models, maximum likelihood, sufficiency, efficiency, information etc.

Jerzy Neyman, **Egon Pearson**. Hypothesis testing, confidence interval, C(alpha) tests etc.

John Tukey. “bit”, “software”, FFT, exploratory vs confirmatory data analysis, box plots, Tukey range etc.

George Box. “robust”.

20

Neyman came to London to do a postdoc with Karl Pearson in 1925, prompting Gosset to write privately to Fisher:

> *He is fonder of algebra than correlation tables and is the only person except yourself I have heard talk about maximum likelyhood as if he enjoyed it.*

Adjusting for 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 …

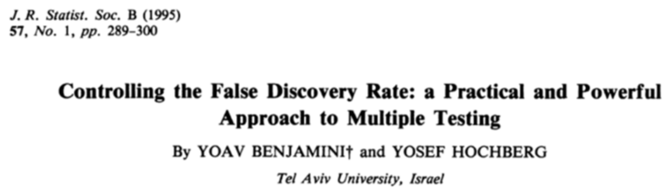
Family-wise error rate

Control the probability of any false positive amongst *n* tests. OK for a modest number of tests.

To control the FWER, sort p-values from smallest to largest and adjust:

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

False discovery rate



J. R. Statist. Soc. B (1995)
57, No. 1, pp. 289–300

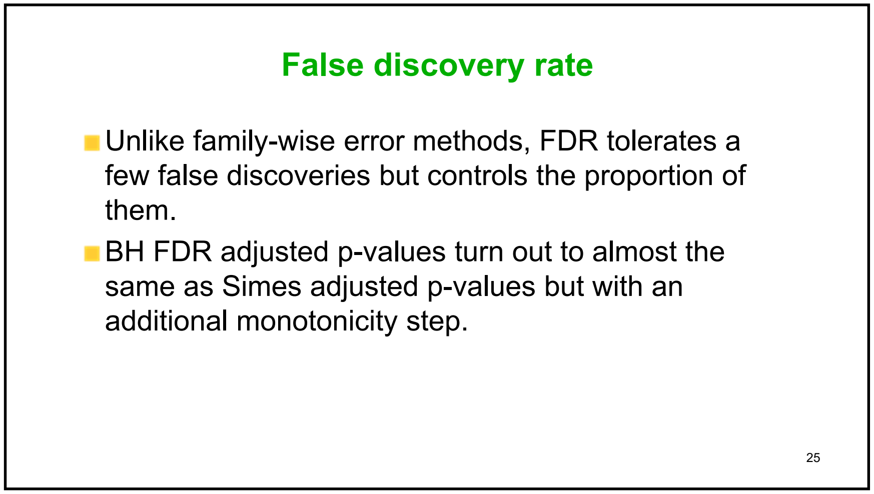
Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing

By YOAV BENJAMINI† and YOSEF HOCHBERG

Tel Aviv University, Israel

No p-values in BH's paper, they took an entirely hypothesis testing approach.

So I re-interpreted BH's method in terms of adjusted p-values and contributed code to p.adjust() in R.

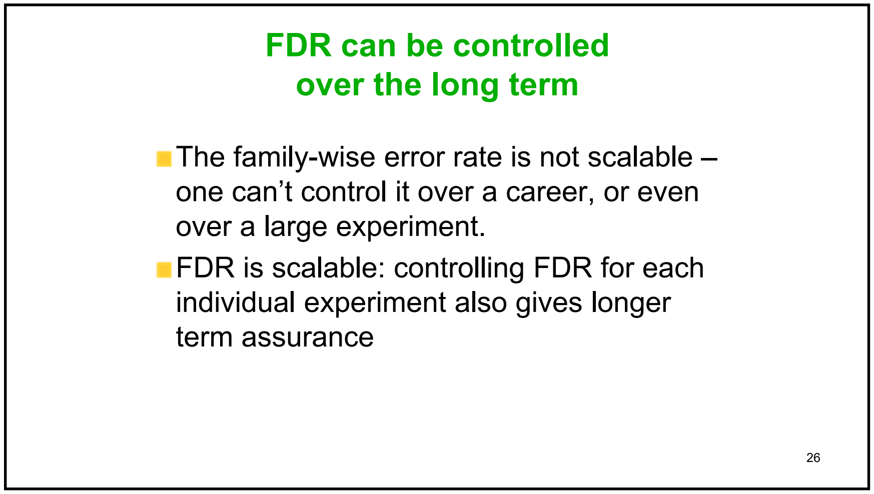


False discovery rate

- Unlike family-wise error methods, FDR tolerates a few false discoveries but controls the proportion of them.
- BH FDR adjusted p-values turn out to almost the same as Simes adjusted p-values but with an additional monotonicity step.

25

25

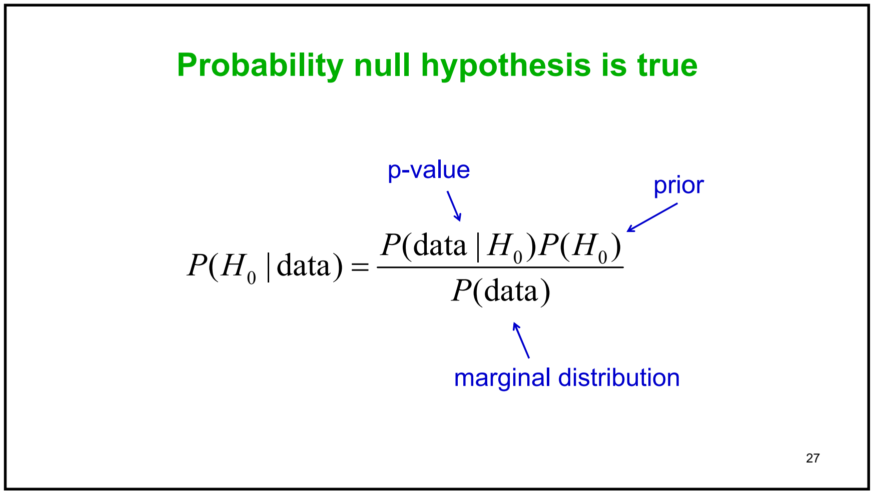


FDR can be controlled over the long term

- 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

26

26



Probability null hypothesis is true

p-value

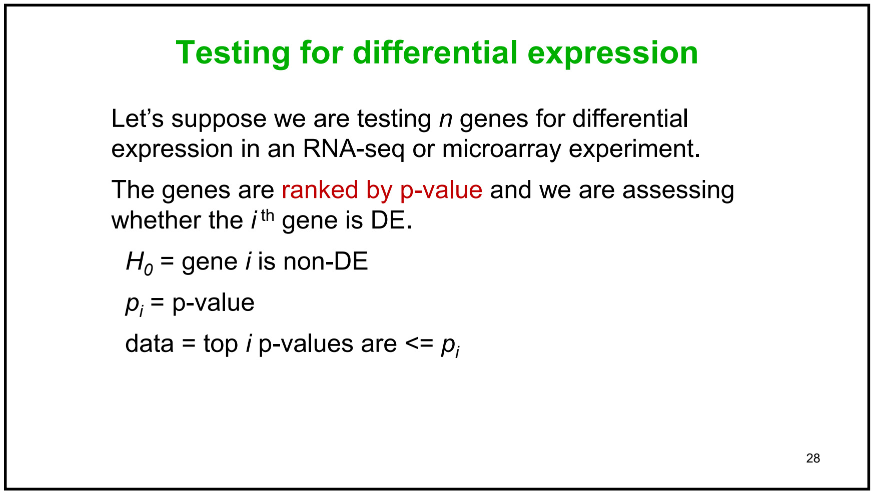
prior

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

marginal distribution

27

27



Testing for differential expression

Let’s suppose we are testing *n* genes for differential expression in an RNA-seq or microarray experiment.

The genes are ranked by p-value and we are assessing whether the i^{th} gene is DE.

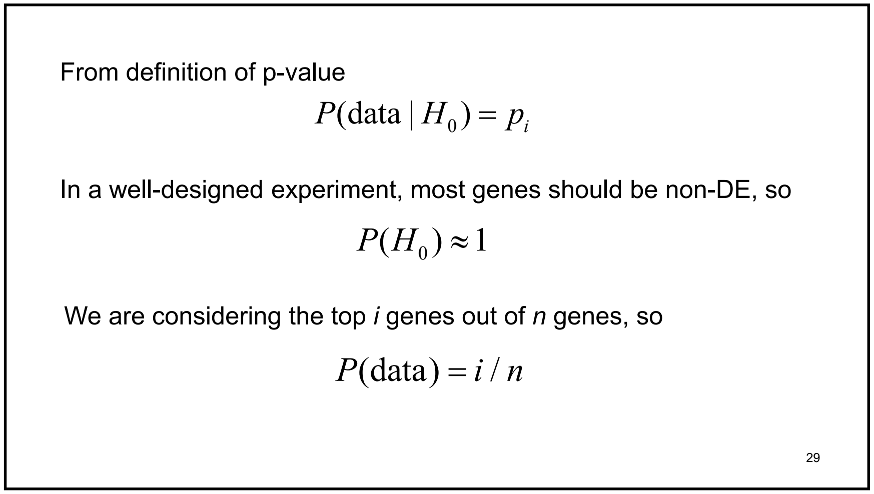
H_0 = gene *i* is non-DE

p_i = p-value

data = top *i* p-values are <= p_i

28

28



From definition of p-value

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

In a well-designed experiment, most genes should be non-DE, so

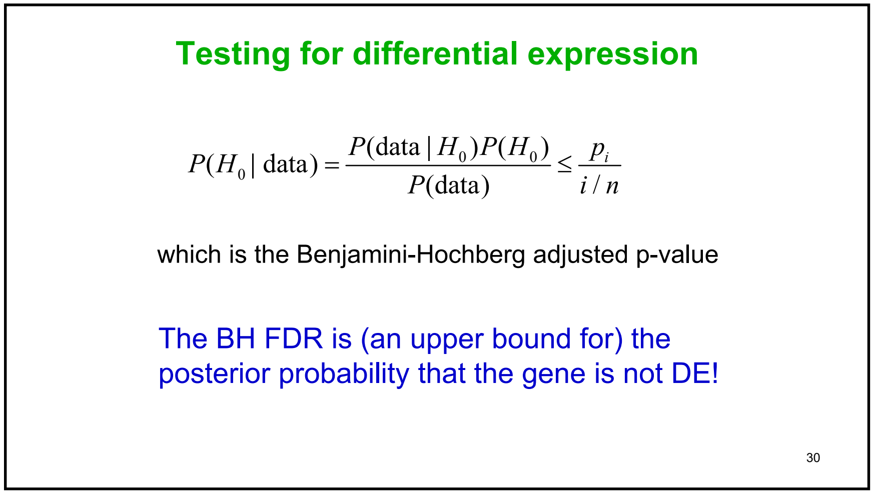
$$P(H_0) \approx 1$$

We are considering the top *i* genes out of *n* genes, so

$$P(\text{data}) = i / n$$

29

29



Testing for differential expression

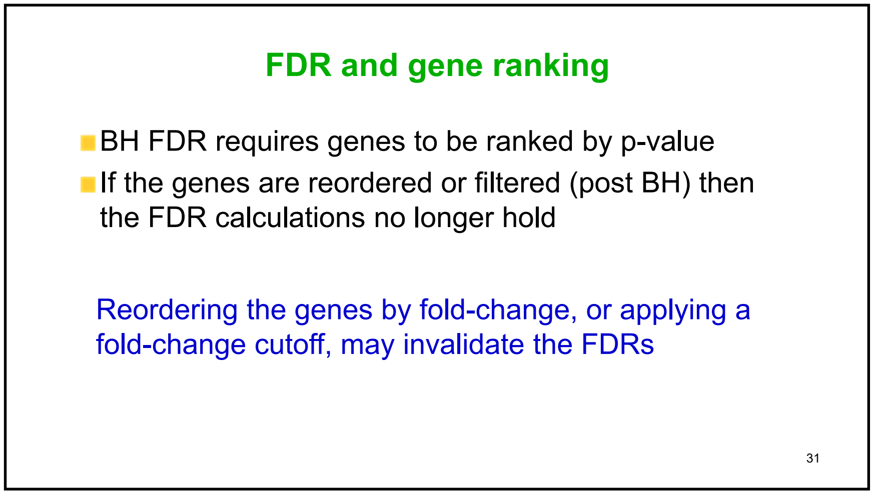
$$P(H_0 \mid \text{data}) = \frac{P(\text{data} \mid H_0) P(H_0)}{P(\text{data})} \le \frac{p_i}{i / n}$$

which is the Benjamini-Hochberg adjusted p-value

The BH FDR is (an upper bound for) the posterior probability that the gene is not DE!

30

30



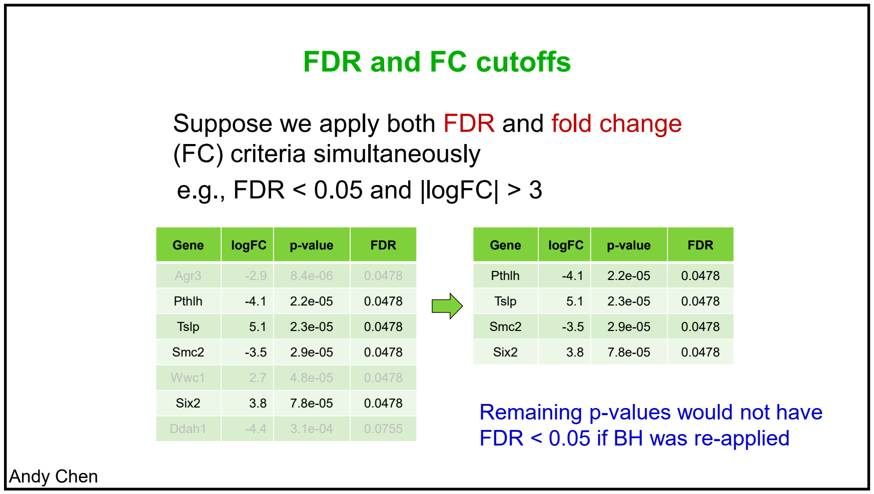
FDR and gene ranking

- BH FDR requires genes to be ranked by p-value
- If the genes are reordered or filtered (post BH) then the FDR calculations no longer hold

Reordering the genes by fold-change, or applying a fold-change cutoff, may invalidate the FDRs

31

31



FDR and FC cutoffs

Suppose we apply both FDR and fold change (FC) criteria simultaneously

e.g., FDR < 0.05 and |logFC| > 3

Gene	logFC	p-value	FDR
Agr3	-2.9	8.4e-06	0.0478
Pthlh	-4.1	2.2e-05	0.0478
Tslp	5.1	2.3e-05	0.0478
Smc2	-3.5	2.9e-05	0.0478
Wwc1	2.7	4.8e-05	0.0478
Six2	3.8	7.8e-05	0.0478
Ddah1	-4.4	3.1e-04	0.0755

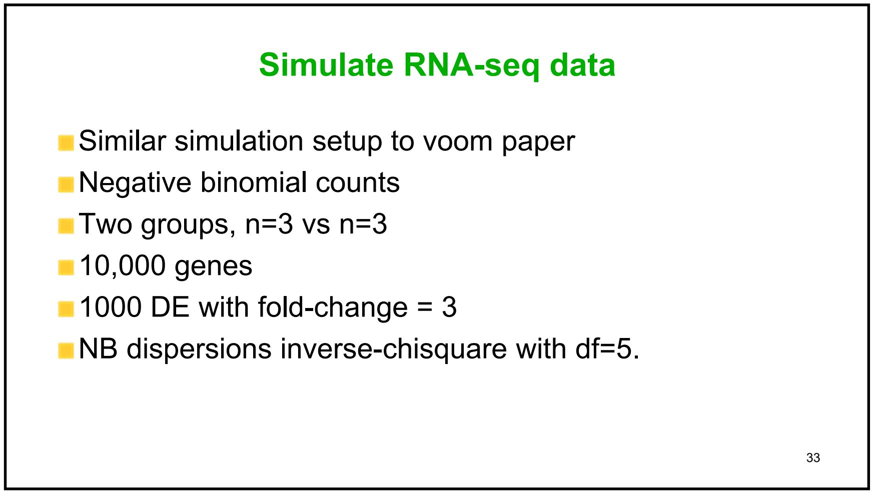
➡

Gene	logFC	p-value	FDR
Pthlh	-4.1	2.2e-05	0.0478
Tslp	5.1	2.3e-05	0.0478
Smc2	-3.5	2.9e-05	0.0478
Six2	3.8	7.8e-05	0.0478

Remaining p-values would not have FDR < 0.05 if BH was re-applied

Andy Chen

32

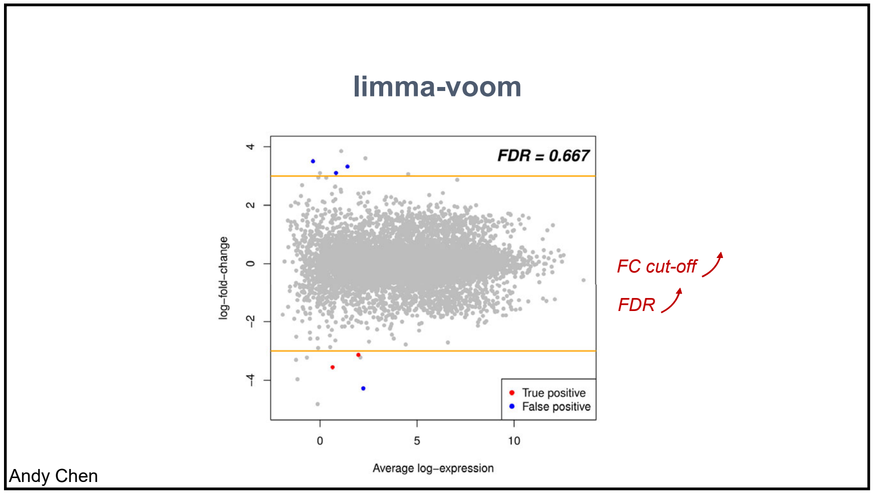


Simulate RNA-seq data

- Similar simulation setup to voom paper
- Negative binomial counts
- Two groups, n=3 vs n=3
- 10,000 genes
- 1000 DE with fold-change = 3
- NB dispersions inverse-chisquare with df=5.

33

33



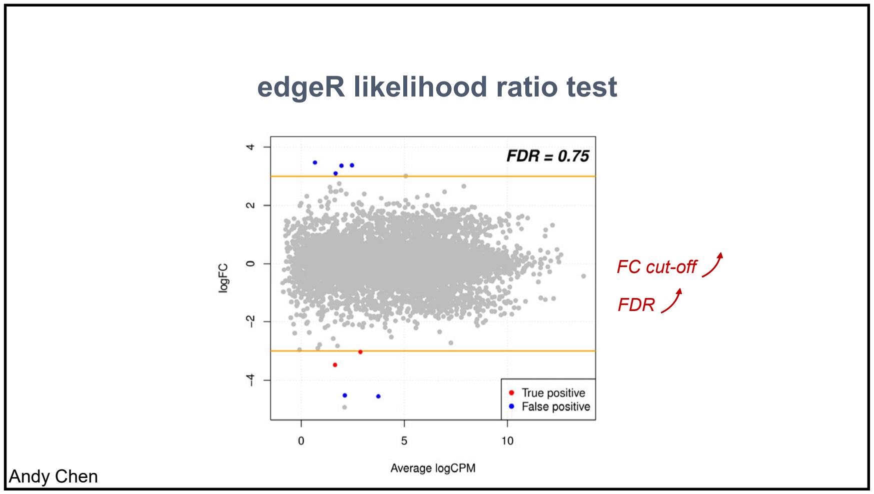
limma-voom

FC cut-off ↗

FDR ↗

Andy Chen

34



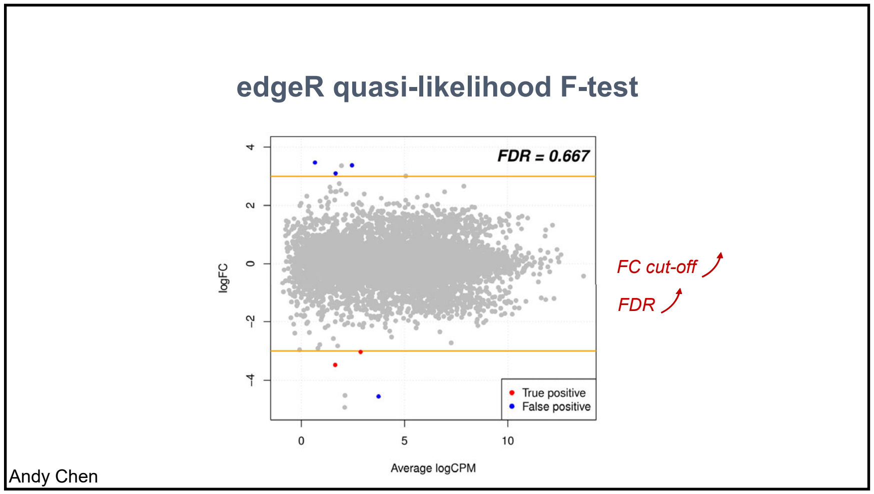
edgeR likelihood ratio test

FC cut-off ↗

FDR ↗

Andy Chen

35



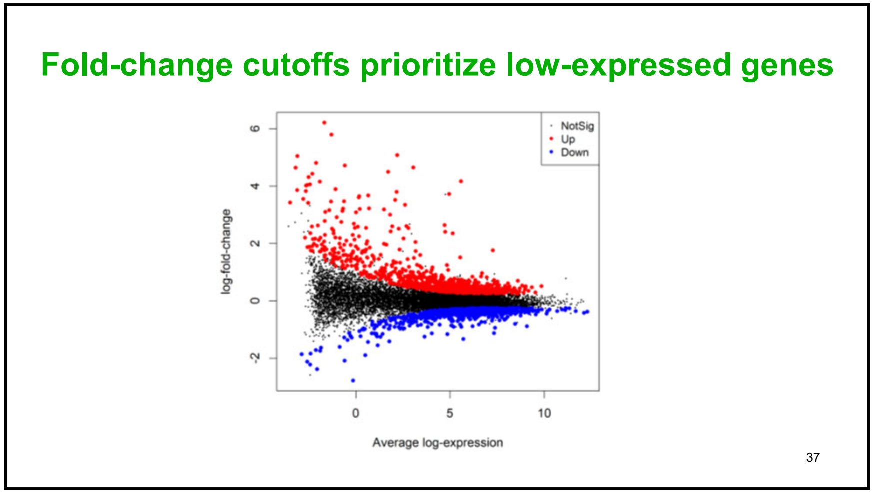
edgeR quasi-likelihood F-test

FC cut-off ↗

FDR ↗

Andy Chen

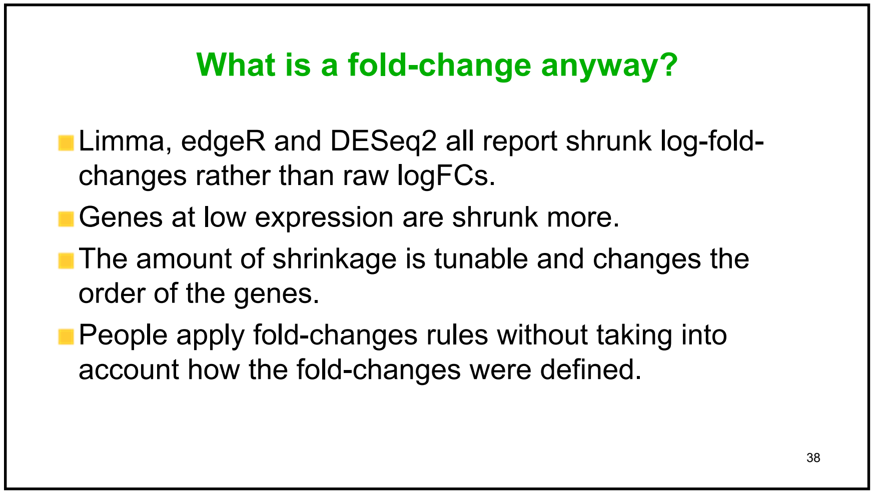
36



Fold-change cutoffs prioritize low-expressed genes

37

37

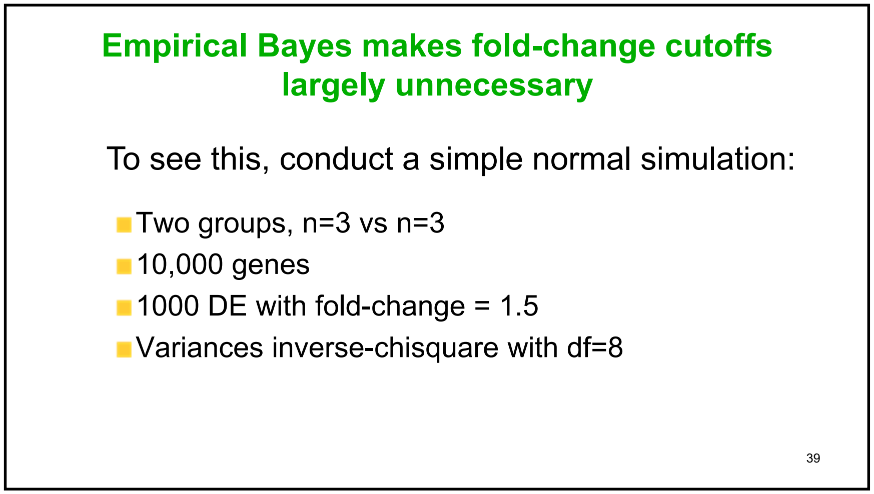


What is a fold-change anyway?

- Limma, edgeR and DESeq2 all report shrunk log-fold-changes rather than raw logFCs.
- Genes at low expression are shrunk more.
- The amount of shrinkage is tunable and changes the order of the genes.
- People apply fold-changes rules without taking into account how the fold-changes were defined.

38

38



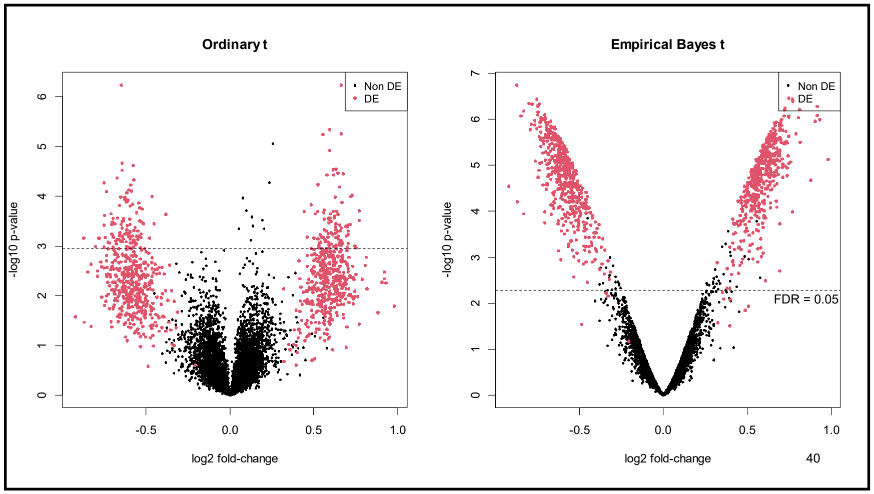
Empirical Bayes makes fold-change cutoffs largely unnecessary

To see this, conduct a simple normal simulation:

- Two groups, n=3 vs n=3
- 10,000 genes
- 1000 DE with fold-change = 1.5
- Variances inverse-chisquare with df=8

39

39



40

40

Testing DE relative to a fold-change threshold is a TREAT

- We created the TREAT methods in limma and edgeR to integrate fold-change thresholding and FDR control
- Works relative to true fold-changes rather than estimated
- P-values are redefined relative to the threshold so that BH is applied to properly ordered p-values
- limma::treat() and edgeR::glmTreat()

Summary

- P-values are a tremendously useful practical tool
- P-values are a building block in the larger scientific puzzle. Don’t make a decision on one p-value alone.
- The popular idea of p-value actually matches FDR much better
- Don’t change p-value orderings when using BH FDR
- Fold-change cut-offs are unnecessary in the empirical Bayes framework

Acknowledgements

- Andy Chen
- Terry Speed

References

- Fisher, RA (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*.
- Reid, C (1982). *Neyman from Life.* Springer.
- Lenhard, J (2006). Models and statistical inference: the controversy between Fisher and Neyman-Pearson. *British Journal for the Philosophy of Science* 57, 69-91.