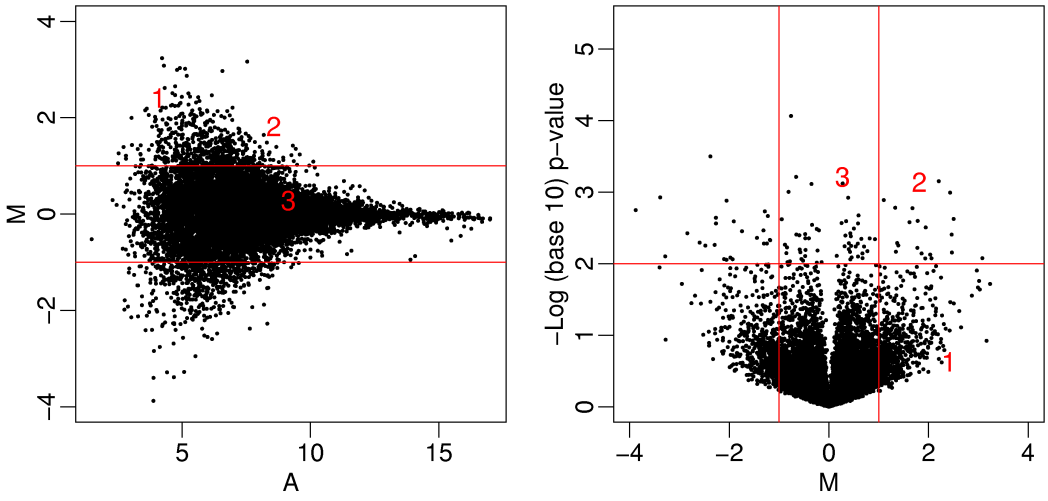
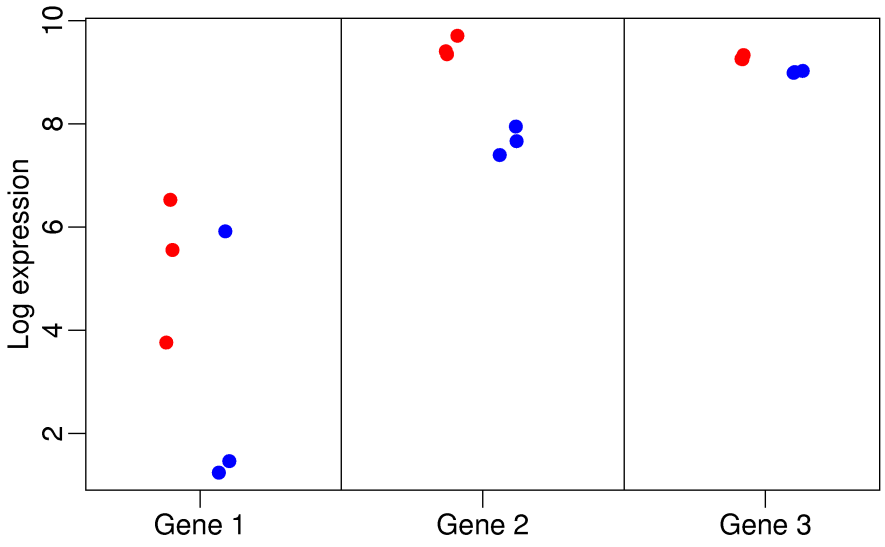
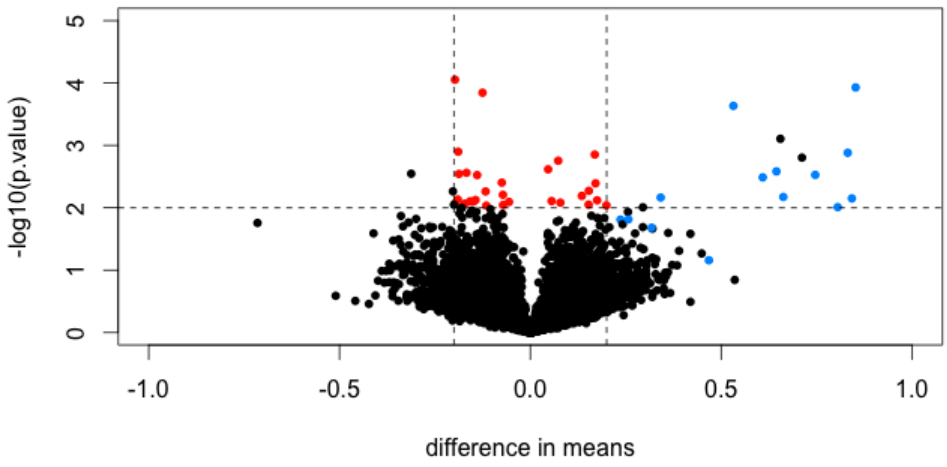
Differential Expression

Empirical Bayes and shrinkage





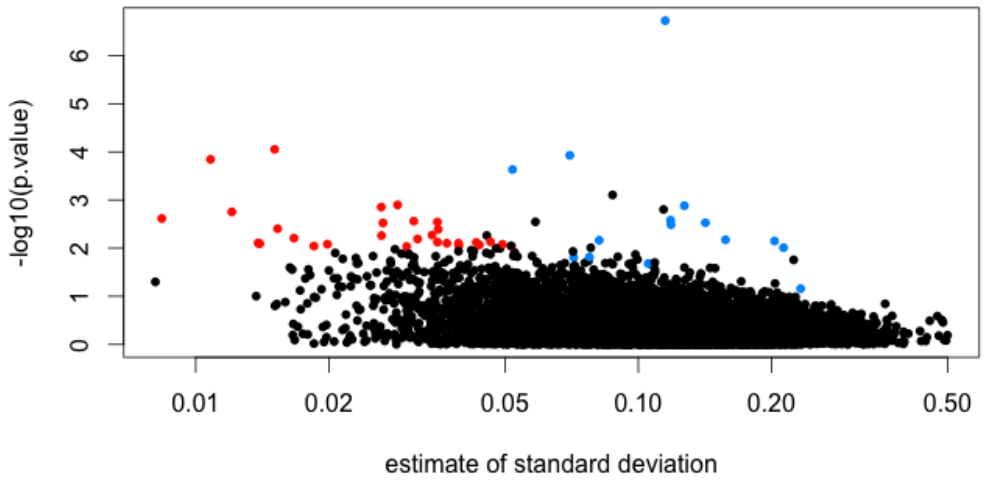
A spike-in experiment



Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[140.668]

A spike-in experiment



[140.668]

ANOVA models for microarray data

A microarray experiment may involve multiple arrays to compare multiple samples. Every measurement in a microarray experiment is associated with a particular combination of an array in the experiment, a dye (red or green), a variety, and a gene. Let y_{ijkg} denote the measurement from the i^{th} array, j^{th} dye, k^{th} variety, and g^{th} gene. To account for the multiple sources of variation in a microarray experiment, consider the model

$$\log(y_{ijkg}) = \mu + A_i + D_j + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \epsilon_{ijkg}, \qquad (1)$$

where μ is the overall average signal, A_i represents the effect of the i^{th} array, D_j represents the effect of the j^{th} dye, V_k represents the effect of the k^{th} variety, G_g represents the effect of the g^{th} gene, $(AG)_{ig}$ represents a combination of array i and gene g (i.e., a particular spot on a particular array), and $(VG)_{kg}$ represents the interaction between the k^{th} variety and the g^{th} gene. The error terms ϵ_{ijkg} are assumed to be independent and identically distributed with mean 0. The array effects A_i account for differences

[PMID 11382364]

Gordon K Smyth

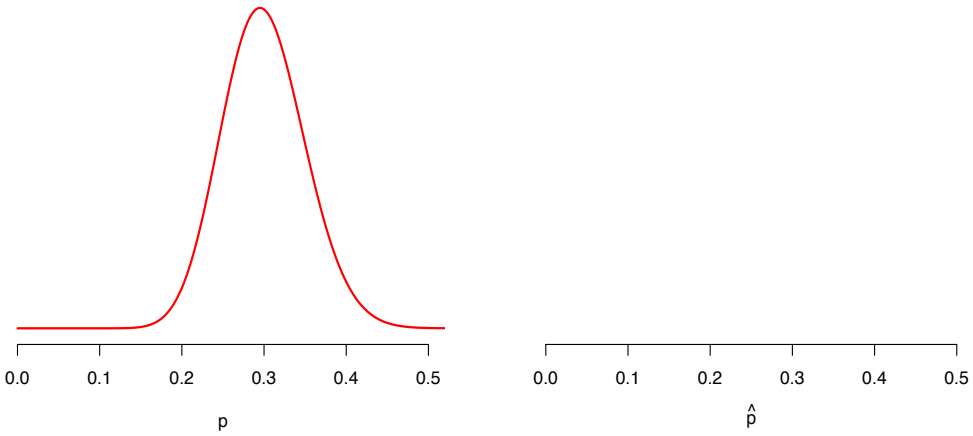
Follow

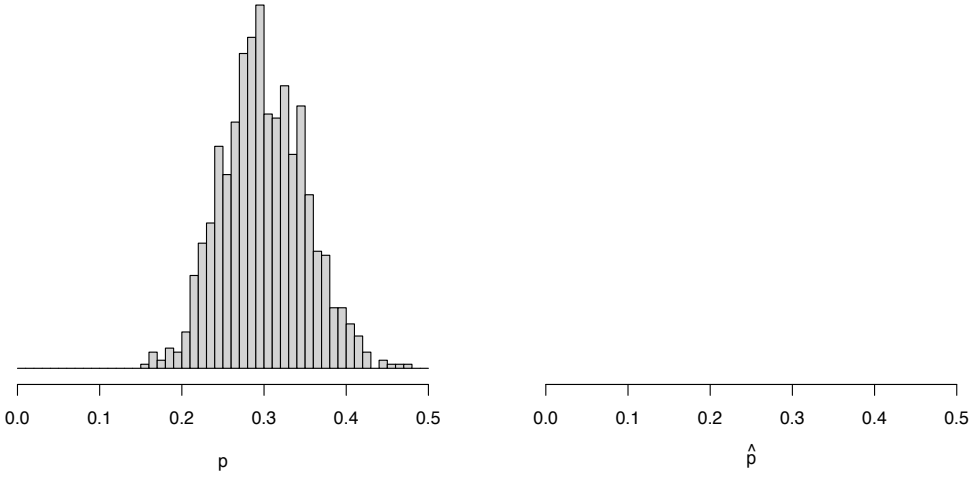
Division Head, Bioinformatics, Walter and Eliza Hall Institute of Medical Research
Bioinformatics, statistics, biostatistics, genomics, statistical computing
Verified email at wehi.edu.au - Homepage

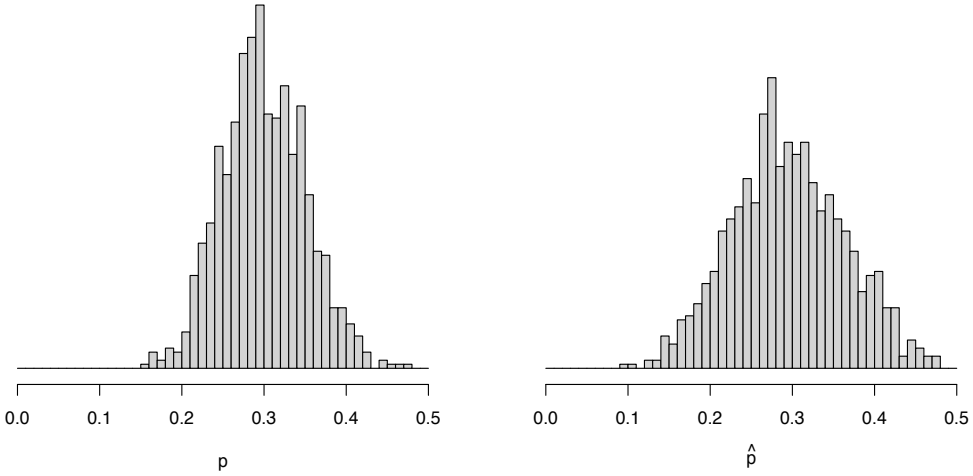
Title 1–20	Cited by	Year
Bioconductor: open software development for computational biology and bioinformatics RC Gentleman, VJ Carey, DM Bates, B Bolstad, M Dettling, S Dudoit, ... Genome biology 5 (10), R80	9510	2004
Linear models and empirical Bayes methods for assessing differential expression in microarray experiments GK Smyth Statistical applications in genetics and molecular biology 3 (1), Article 3	9116	2004

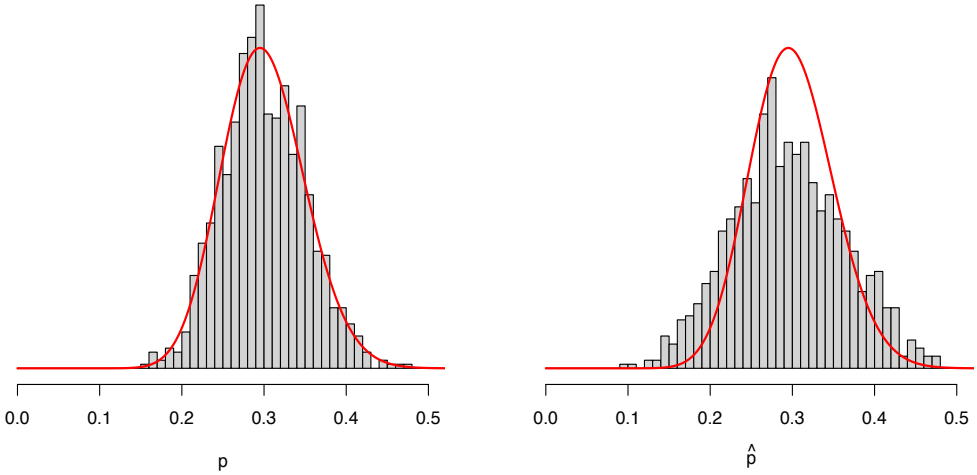
DATE	OPP	RESULT	AB	R	H	2B	3B	HR	RBI	BB	SO	SB	CS	OBP	SLG	OPS	AVG
Apr 1	@ NYY	W 8-2	5	1	3	0	0	0	1	0	1	0	0	.600	.600	1.200	.600
Apr 3	@ NYY	W 7-4	4	1	2	1	0	0	0	0	1	0	0	.556	.667	1.223	.556
Apr 4	@ NYY	L 4-2	3	0	2	0	0	0	0	0	0	0	0	.583	.667	1.250	.583
Apr 5	@ TOR	W 6-4	0	0	0	0	0	0	0	0	0	0	0	.615	.667	1.282	.583
Apr 6	@ TOR	L 5-0	Did not play														
Apr 7	@ TOR	W 13-0	5	1	2	1	0	0	0	0	1	0	0	.556	.647	1.203	.529
Apr 8	vs BAL	W 3-1	3	0	0	0	0	0	0	0	0	0	0	.476	.550	1.026	.450
Monthly Totals			**20**	**3**	**9**	**2**	**0**	**0**	**1**	**0**	**3**	**0**	**0**	**.476**	**.550**	**1.026**	**.450**

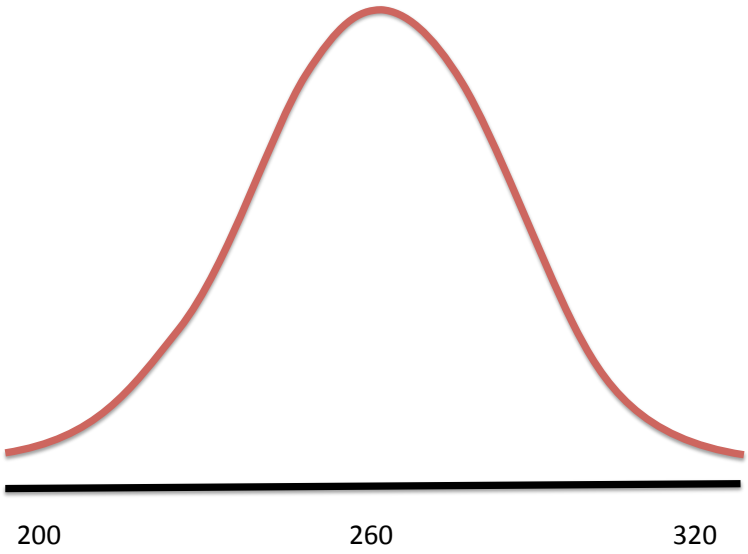
espn.go.com


0.0
0.1
0.2
0.3
0.4
0.5
p
0.0
0.1
0.2
0.3
0.4
0.5
p̂


0.0
0.1
0.2
0.3
0.4
0.5
p
0.0
0.1
0.2
0.3
0.4
0.5
p̂


0.0
0.1
0.2
0.3
0.4
0.5
p
0.0
0.1
0.2
0.3
0.4
0.5
p̂


0.0
0.1
0.2
0.3
0.4
0.5
p
0.0
0.1
0.2
0.3
0.4
0.5
p̂



A (rough) sketch of the MLB batting average distribution.

[RI]

A hierarchical model

$$\theta \sim N(\mu, \tau^2)$$

$$Y|\theta \sim N(\theta, \sigma^2)$$

Here, θ denotes *any* batting average among the MLB players, and *Y* denotes the player's batting average. The parameter τ quantifies the prior standard deviation, and σ describes the sampling standard deviation. Specifically:

$$\theta \sim N(260, 34^2)$$

$$Y|\theta \sim N(\theta, 110^2)$$

A hierarchical model

Best guess for the players batting average, given the observed data:

$$\begin{aligned} E(\theta|Y) &= B\mu + (1-B)Y \\ &= \mu + (1-B)(Y-\mu) \end{aligned}$$

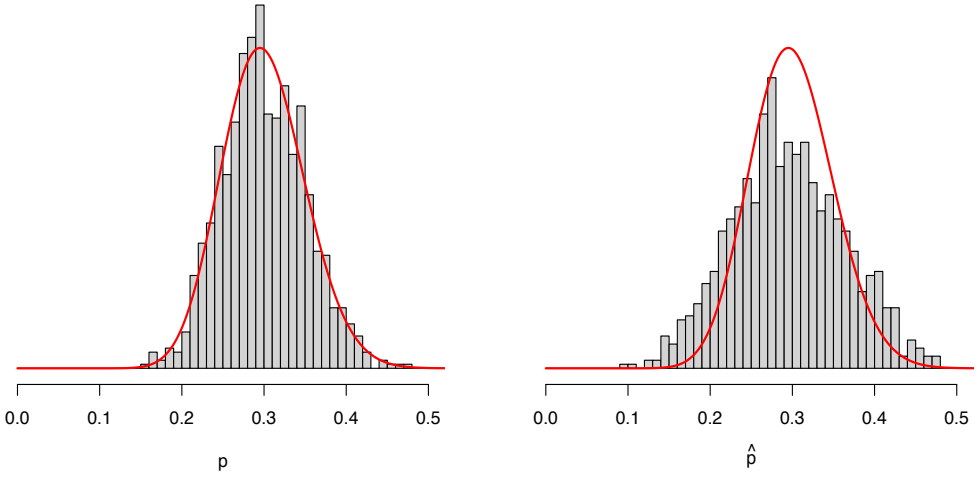
$$B = \frac{\sigma^2}{\sigma^2 + \tau^2}$$

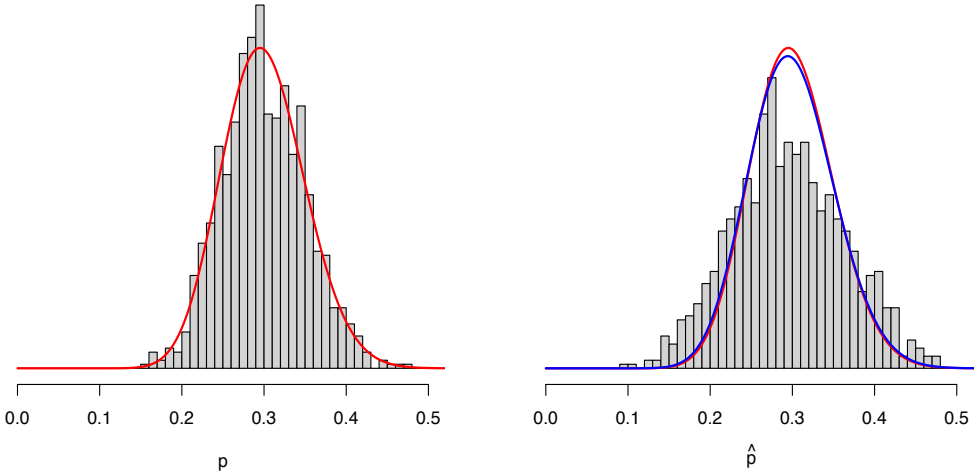
Specifically:

$$\begin{aligned} E(\theta|Y=450) &= B \times 260 + (1-B) \times 450 \\ &= 260 + (1-B)(450-260) \end{aligned}$$

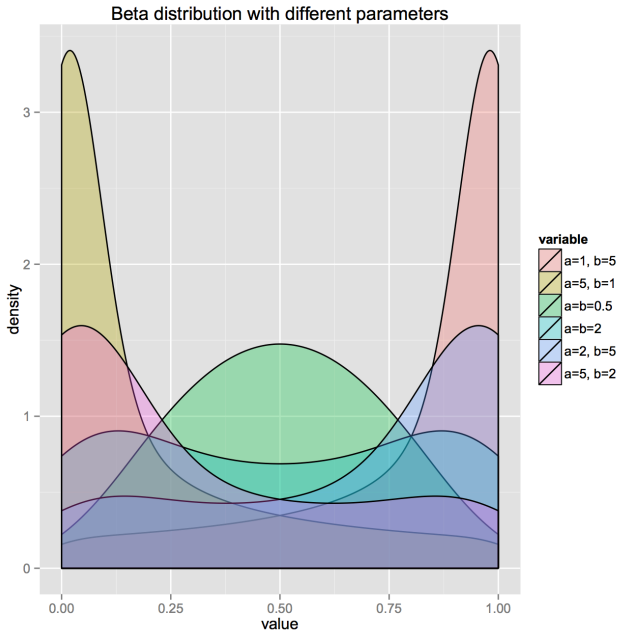
$$B = \frac{110^2}{110^2 + 34^2}$$

$$E(\theta|Y=450) \approx 270$$





A better solution



A better solution

In this context, we can express our model as:

$$k_i \sim Binomial(n_i, p_i)$$

$$p_i \sim Beta(a, b), i = 1...N$$

where N is total number of observations and a and b are parameters to be estimated. Such model is also called Empirical Bayes. Unlike traditional Bayes, in which we pull prior distribution and it's parameters out of the thin air, Empirical Bayes estimates prior parameters from the data.

In order to estimate parameters of the prior, we calculate marginal distribution as

$$m(k|a, b) = \int \prod_{i=1}^{N} f(k_i|p)\pi(p|a, b)dp = \prod_{i=1}^{N} \binom{n_i}{k_i} \frac{\Gamma(a+b)\Gamma(a+k_i)\Gamma(n_i-k_i+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+n_i)}$$

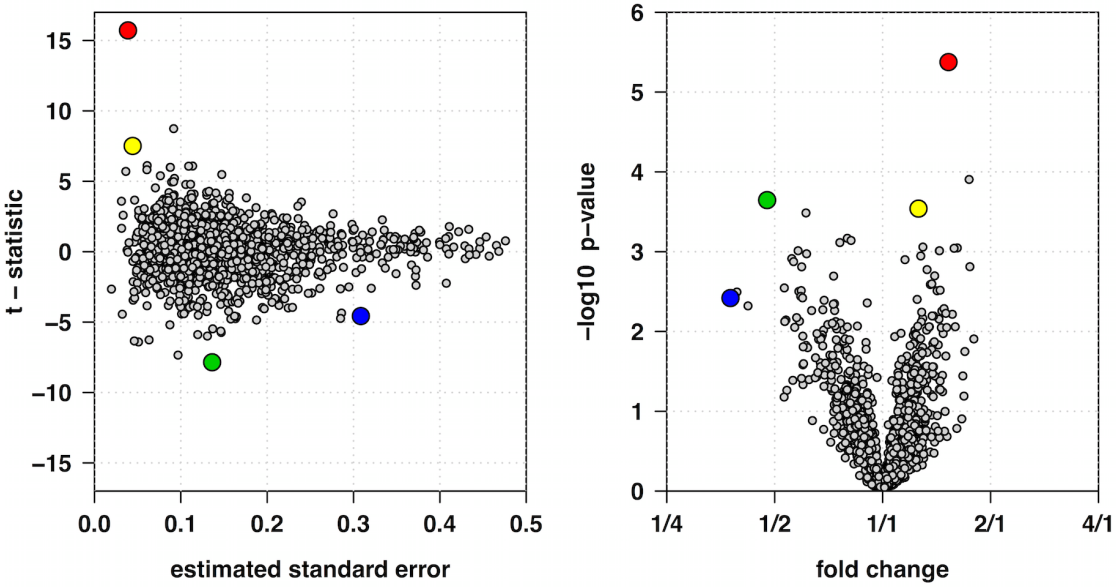
where f and π are density functions of binomial and beta distributions, respectively. Parameter estimates $\hat{a}$ and $\hat{b}$ can be obtained by maximizing the log likelihood of the marginal distribution.

Finally, Empirical Bayes estimator can be constructed as expectation of posterior distribution:

$$\hat{p_i} = E(p_i|k_i, \hat{a}, \hat{b}) = \frac{\hat{a}+k_i}{\hat{a}+\hat{b}+n_i}$$

blog.supplyframe.com/2013/09/10/empirical-bayes-estimation-of-p-using-r/

An iTRAQ experiment



- Assume that the true (unobservable) protein variances follow a scaled inverse χ^2 distribution:

$$\frac{1}{\sigma_p^2} \sim \frac{1}{d_0 \times s_0^2} \times \chi^2_{d_0}.$$

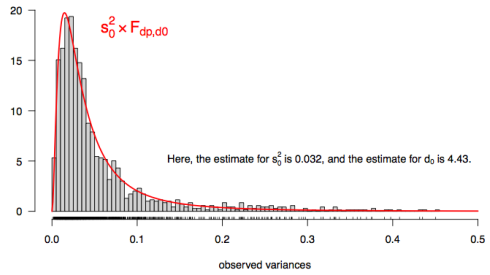
We estimate the parameters d_0 and s_0^2 later from the observed data.

- If the data for a particular protein are normally distributed with variance σ_p^2, it follows for the observed sample variance that

$$s_p^2 | \sigma_p^2 \sim \frac{\sigma_p^2}{d_p} \times \chi^2_{d_p},$$

where d_p are the degrees of freedom associated with the experiment.

- This implies that the observed sample variances s^2 follow a scaled F distribution: $s^2 \sim s_0^2 \times F_{d_p, d_0}$.



- For the test statistics the observed variances are shrunk towards the prior values with the degree of shrinkage depending on the relative sizes of the observed and prior degrees of freedom:

$$s^2_{p\ [\text{ moderated }]} = \frac{d_0 \times s_0^2 + d_p \times s_p^2}{d_0 + d_p}.$$

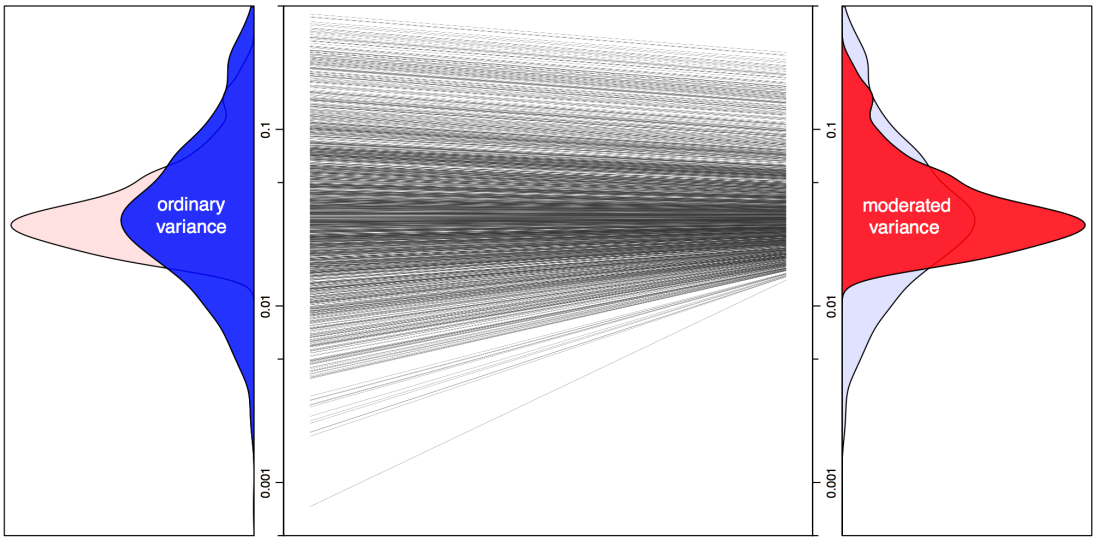
Ordinary t-statistic:

$$t_p = \frac{\text{estimated log fold change}}{\text{estimated standard error}} = \frac{\overline{X}_p - \overline{Y}_p}{s_p\sqrt{2/n}},$$

Moderated t-statistic:

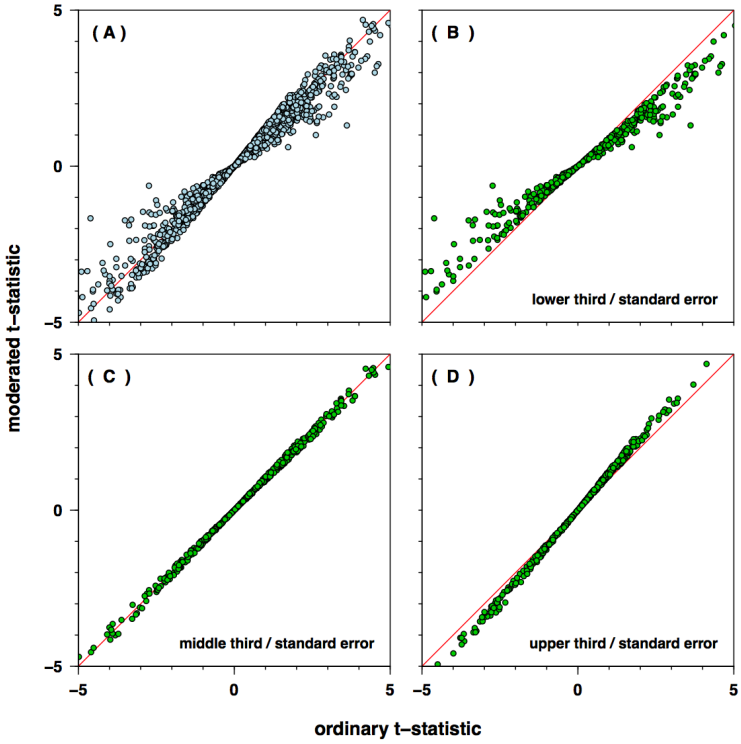
$$t_{p\,[\text{moderated}]} = \frac{\text{estimated log foldchange}}{\text{moderated standard error}} = \frac{\overline{X}_p - \overline{Y}_p}{s_{p\,[\text{moderated}]}\sqrt{2/n}}$$

[PMID 25821719]

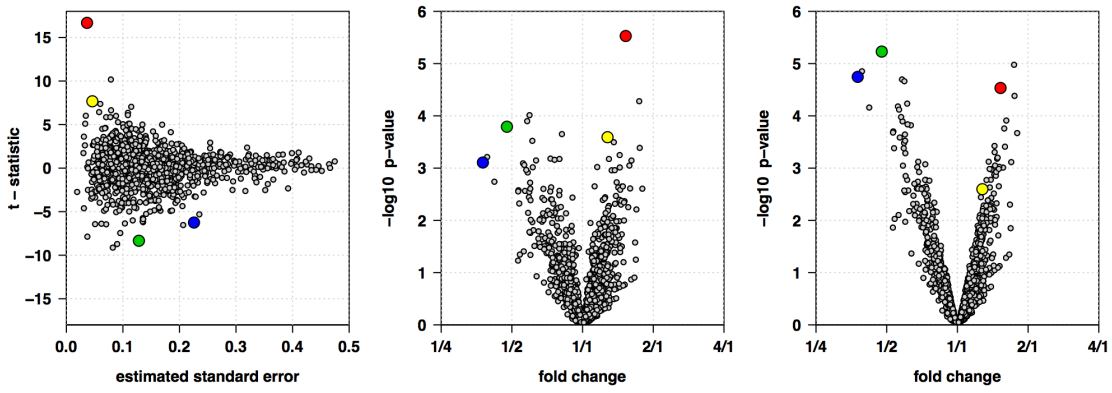


Ingo Ruczinski | Asian Institute in Statistical Genetics and Genomics | July 21-22, 2017

[PMID 25821719]



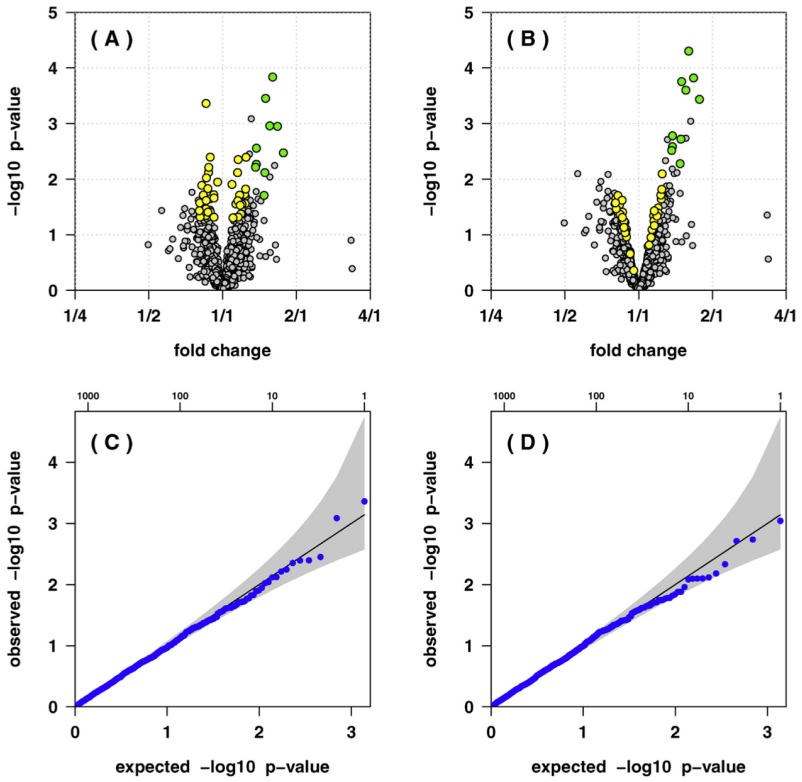
[PMID 25821719]



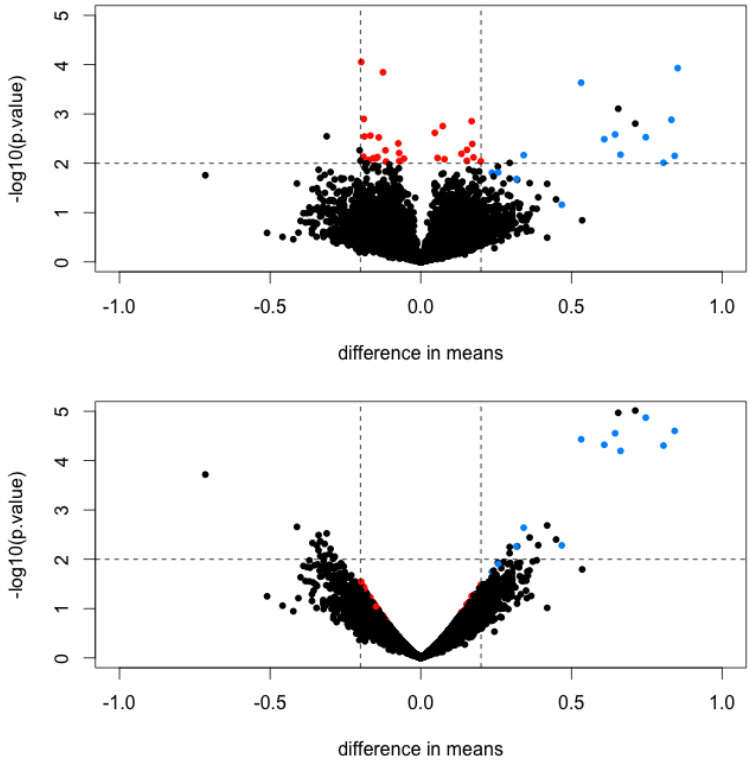
At a false discovery rate control of 1% only 1 protein is declared differentially expressed when using ordinary t-statistics, compared to 23 proteins when moderated t-statistics are used.

FDR of 5%: 30 and 98 proteins, respectively; FDR of 10%: 120 and 184 proteins, respectively.

[PMID 25821719]



[PMID 25821719]



[RI]

Limma

Observed gene sample variance $$s_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi^2_{d_g}$$

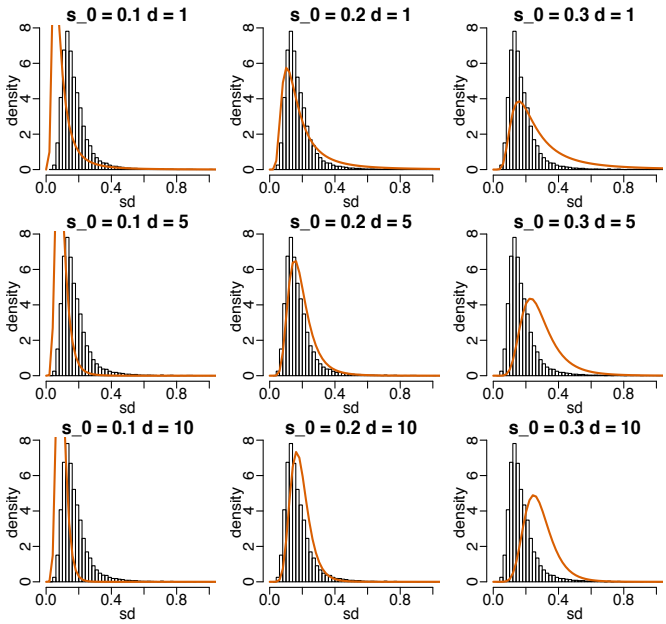
Variability between genes $$s^2 \sim s_0^2 F_{d,d_0}$$

Moderated gene variance $$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

[PMID 16646809]

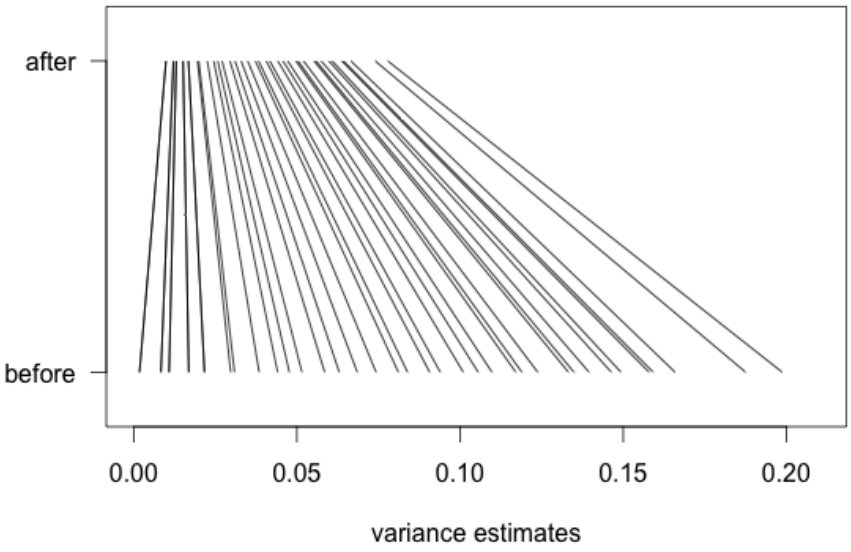
Variances as scaled F-distribution

$$s^2 \sim s_0^2 F_{d,d_0}$$



[RI]

Moderated sample variances



[RI]