

Hypothesis Testing



Das Orakel zu Delphi.

Wolfgang Huber, EMBL

Karl Popper (1902-1994)

Logical asymmetry between verification and falsifiability.

No number of positive outcomes at the level of experimental testing can confirm a scientific theory, but a single counterexample is logically decisive: it shows the theory is false.



The four steps of hypothesis testing

Step 1: Set up a model of reality: null hypothesis, H_0

Step 2: Do an experiment, collect data

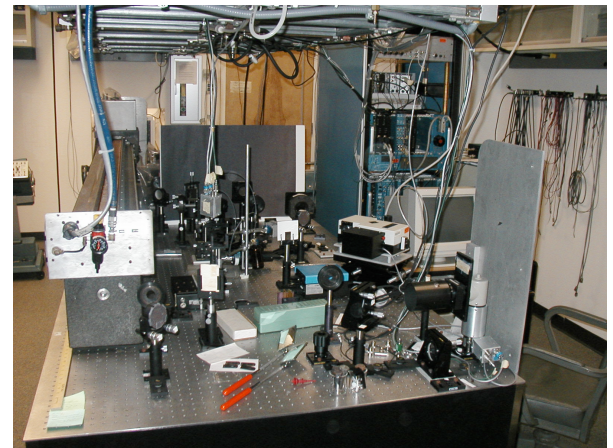
Step 3: Compute the probability of the data in this model

Step 4: Make a decision: reject model if the computed probability is deemed to small

H_0 : a model of reality that lets us make specific predictions of how the data should look like. The model is stated using the mathematical theory of probability.

Examples of null hypotheses:

- The coin is fair
- The new drug is no better or worse than a placebo
- The observed CellTitreGlo signal for my RNAi-treated cells is no different from that of the negative controls



Example

Toss a coin a certain number of times \Rightarrow

If the coin is fair, then heads should appear half of the time (roughly).

But what is “roughly”? We use combinatorics / probability theory to quantify this.

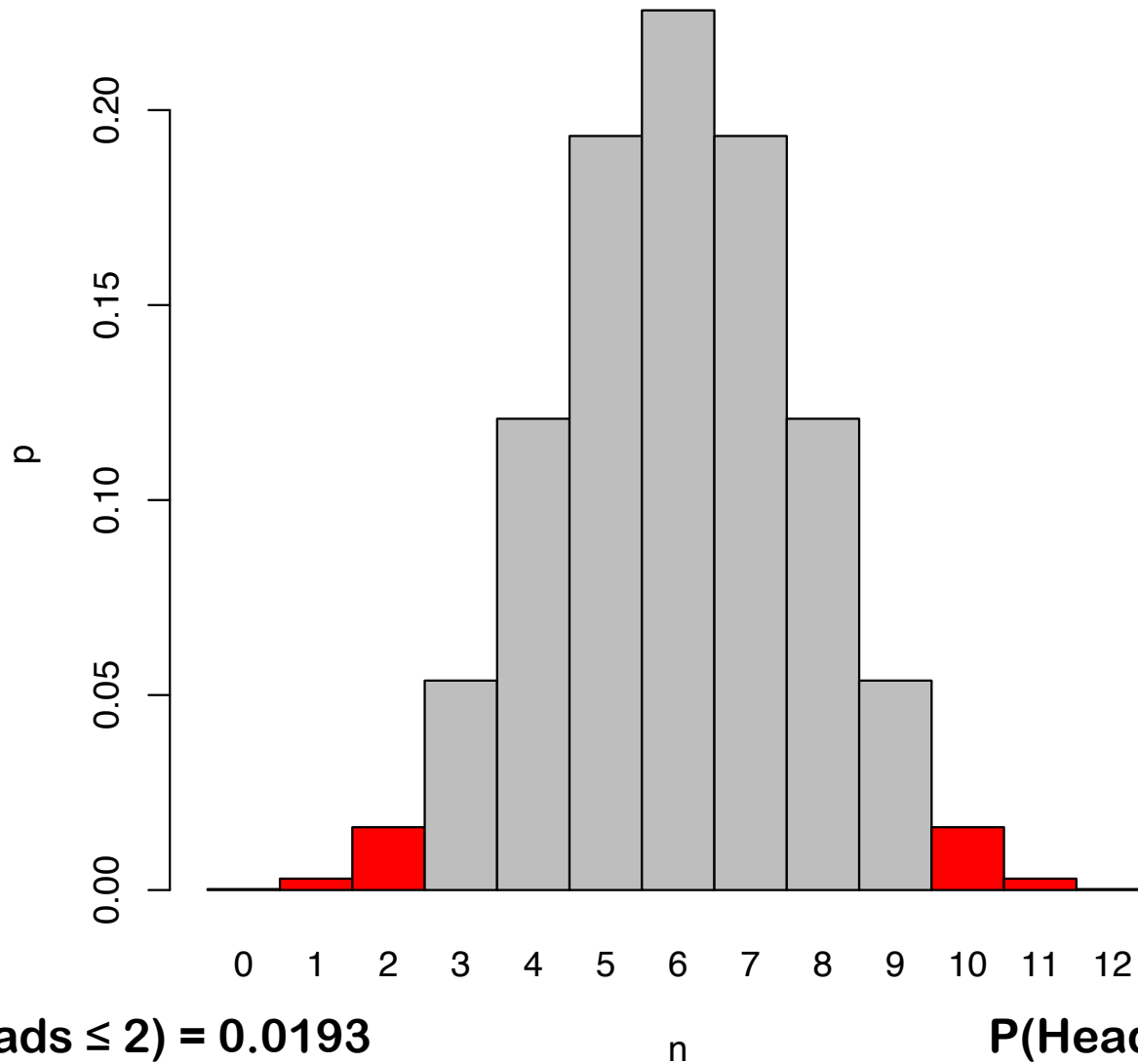
For example, in **12** tosses with **success rate p** , the probability of seeing exactly **8** heads is

$$\binom{12}{8} p^8 \cdot (1 - p)^4$$



Binomial Distribution

H_0 here: $p = 0.5$. Distribution of number of heads:



$P(\text{Heads} \leq 2) = 0.0193$

$P(\text{Heads} \geq 10) = 0.0193$

Significance Level

If H_0 is true and the coin is fair ($p=0.5$), it is improbable to observe extreme events such as more than 9 heads

$$0.0193 = P(\text{Heads} \geq 10 \mid H_0) = \text{“p-value”}$$

If we observe 10 heads in a trial, the null hypothesis is likely to be false.

An often used (but entirely arbitrary) cutoff is 0.05 (“significance level α ”): if $p < \alpha$, we reject H_0

Two views:

Strength of evidence for a certain (negative) statement

Rational decision support

Statistical Testing Workflow

1. Set up hypothesis H_0 (that you want to reject)
2. Find a test statistic T that should be sensitive to (interesting) deviations from H_0
3. Figure out the null distribution of T , if H_0 holds
4. Compute the actual value of T for the data at hand
5. Compute p-value = the probability of seeing that value, or more extreme, in the null distribution.
6. Test Decision: Rejection of H_0 - yes / no ?

Errors in hypothesis testing

Decision		Truth	
		not rejected ('negative')	rejected ('positive')
H ₀	H ₀	True negative (specificity)	False Positive Type I error α
	H ₁	False Negative Type II error β	True Positive (sensitivity)

One sample t-test

t-statistic (1908, William Sealy Gosset, pen-name “Student”)

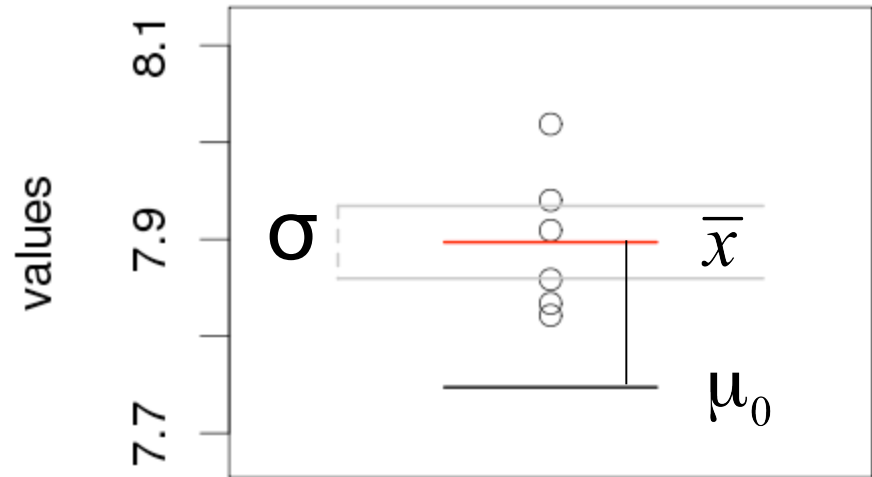
$$t = \sqrt{n} \frac{\bar{x} - \mu_0}{\hat{\sigma}}$$

compare to a fixed value μ_0

Without n: z-score

With n: t-statistic

If data are normal, null distribution can be computed: “t-distribution”, with a parameter called “degrees of freedom”, equal to n-1



One sample t-test example

Consider the following 10 data points:

-0.01, 0.65, -0.17, 1.77, 0.76, -0.16, 0.88, 1.09, 0.96, 0.25

We are wondering if these values come from a distribution with a true mean of 0: one sample t-test

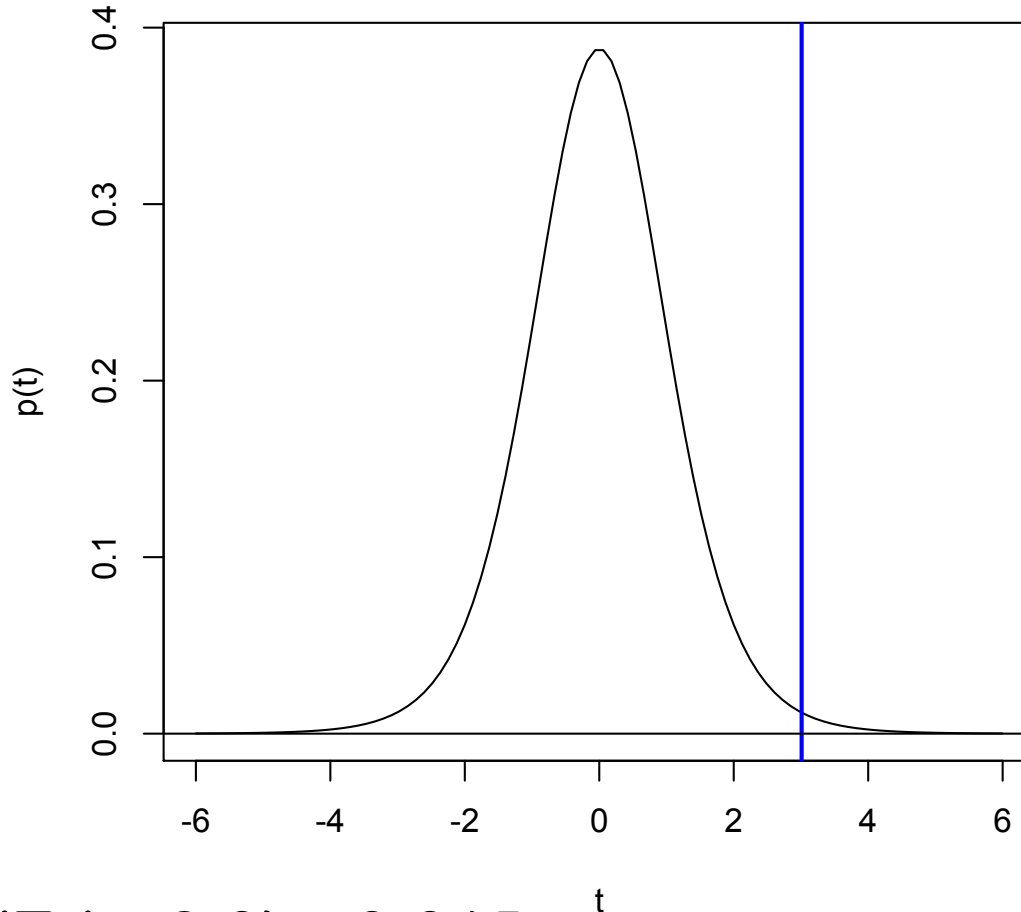
The 10 data points have a mean of 0.60 and a standard deviation of 0.62.

From that, we calculate the t-statistic:

$$t = 0.60 / 0.62 * 10^{1/2} = 3.0$$

p-value and test decision

10 observations → compare observed t-statistic to the t-distribution with 9 degrees of freedom

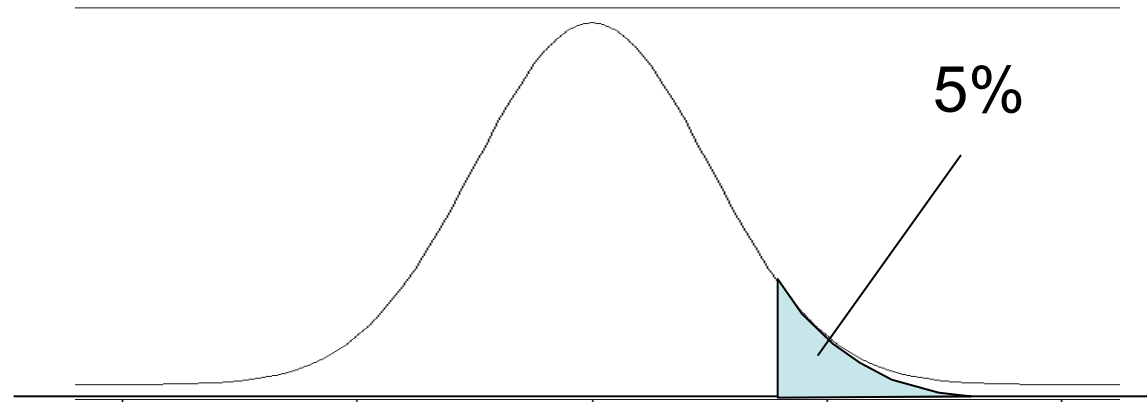


p-value: $P(|T_9| \geq 3.0) = 0.015$

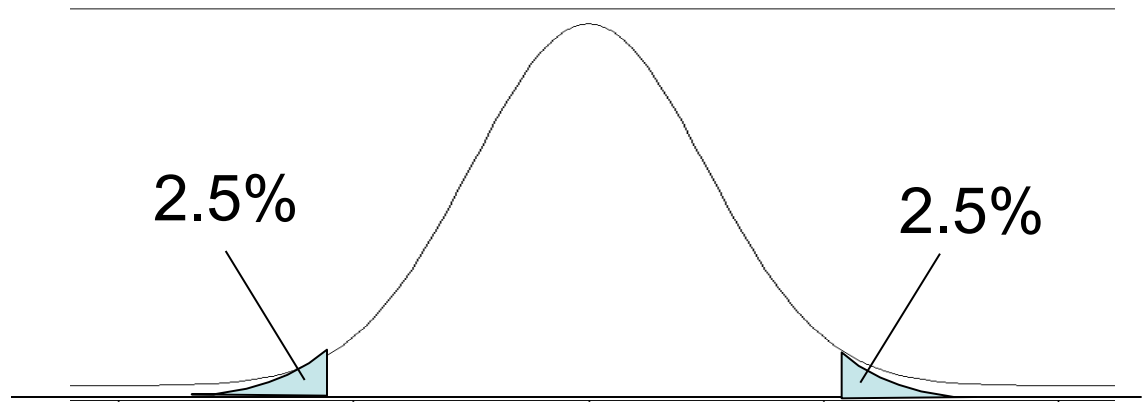
In R: `pt(3.0, df=9, lower.tail=FALSE)`

One-sided vs two-sided test

One-sided
e.g. $H_0: \mu < 0$



Two-sided
e.g. $H_0: \mu = 0$



Avoid fallacy

The p-value is the probability that the observed data could happen, under the condition that the null hypothesis is true.

It is not the probability that the null hypothesis is true.

Absence of evidence \neq evidence of absence

Two samples t-test

Do two different samples have the same mean ?

$$t = \frac{\bar{y} - \bar{x}}{SE}$$

\bar{y} and \bar{x} are the average of the observations in the two populations

SE is the standard error for the difference

If H_0 is correct, test statistic follows a t-distribution with $n+m-2$ degrees of freedom

(n, m : number of observations in each sample)

t-test in R

```
t.test(x, y, alternative, paired, var.equal)
```

x,y: Data (only x needs to be specified for one-group test, specify target mu instead)

paired: paired (e.g. repeated measurements on the same subjects) or unpaired

var.equal: Can the variances in the two groups assumed to be equal?

alternative: one- or two-sided test?

Comments and pitfalls

The derivation of the t-distribution assumes that the observations are independent and that they follow a Normal distribution.

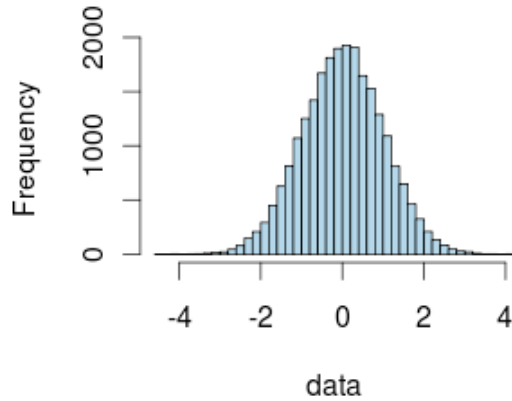
Deviation from Normality - heavier tails: test still maintains type-I error control, but may no longer have optimal power.

Options: Wilcoxon test, permutation tests

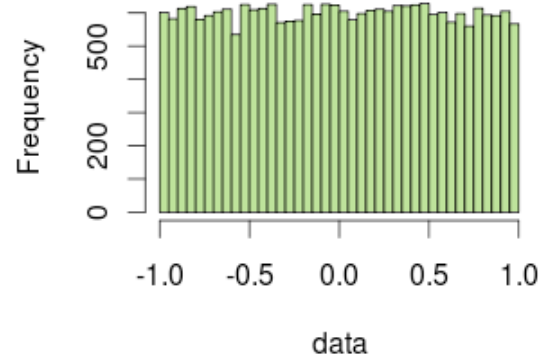
If the data are **dependent**, then p-values will likely be totally wrong (e.g., for positive correlation, too optimistic).

different data distributions – independent case

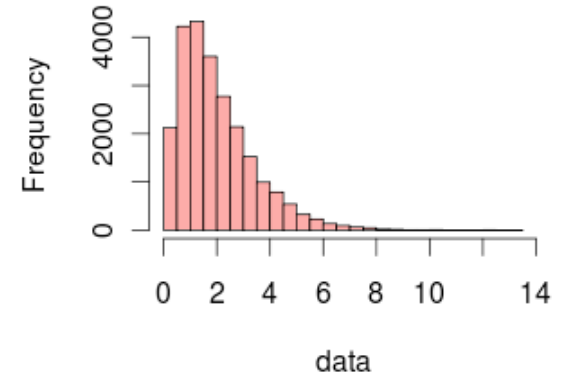
Normal(0,1)



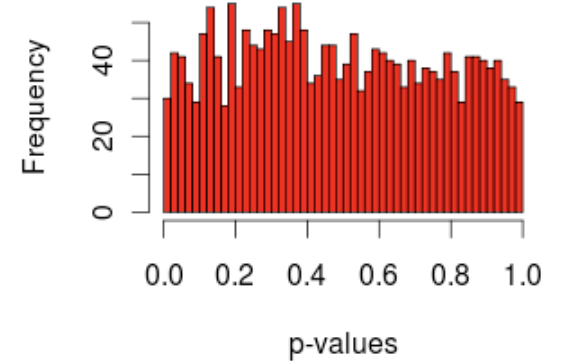
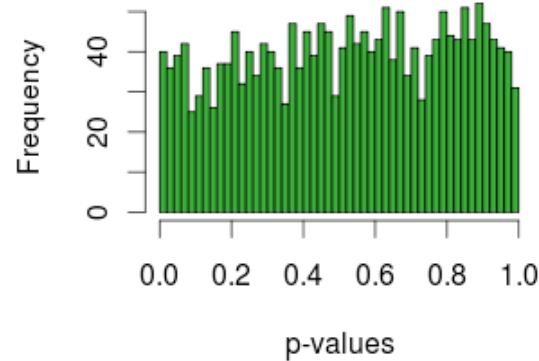
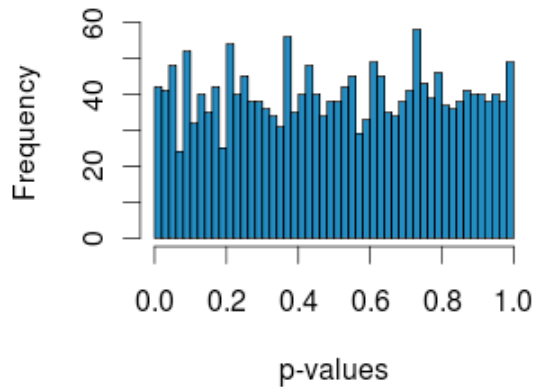
Uniform(-1,1)



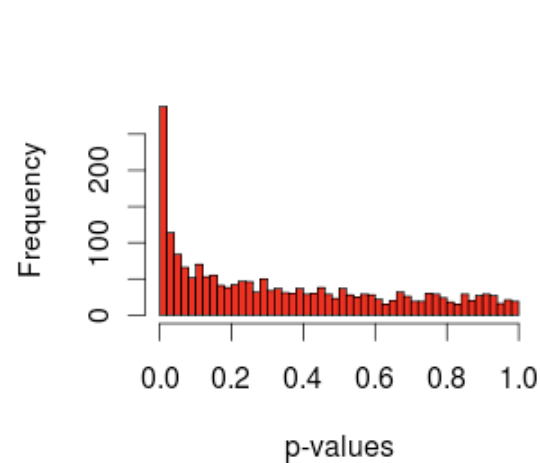
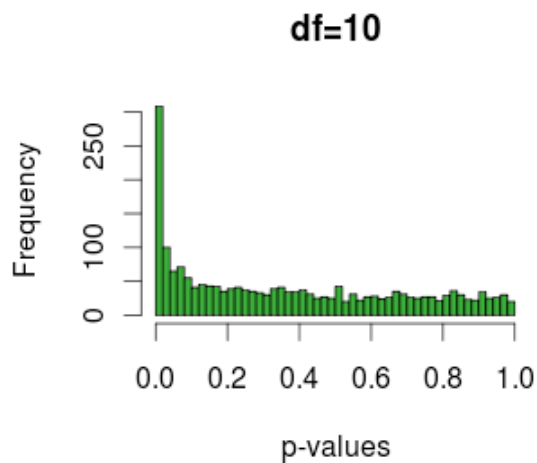
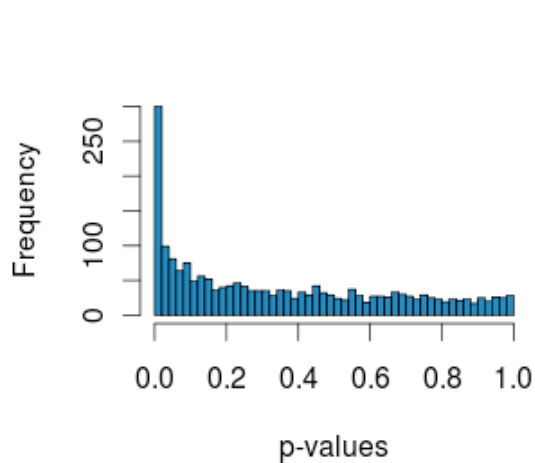
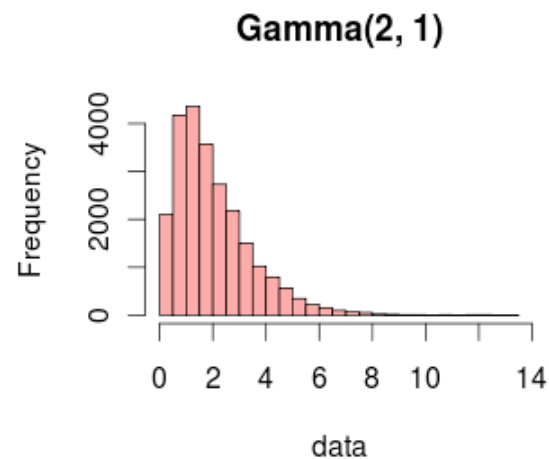
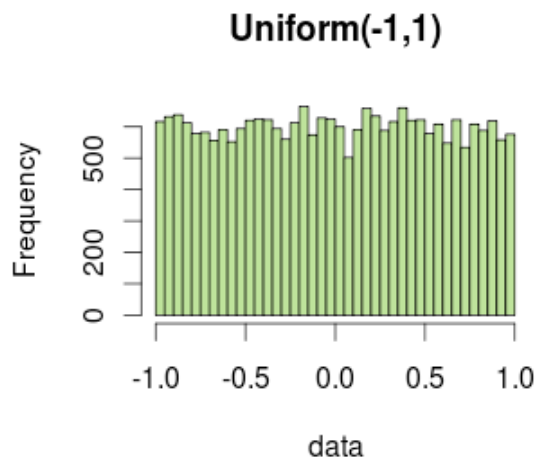
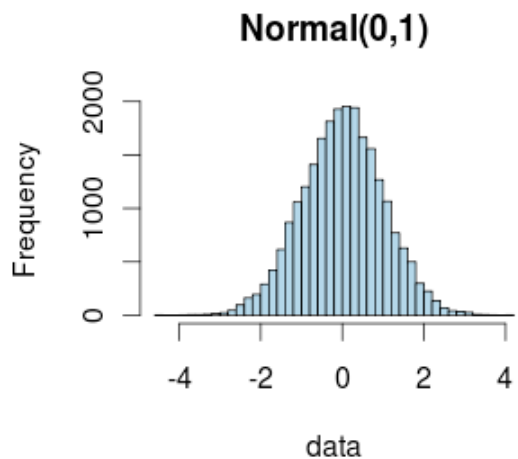
Gamma(2, 1)



df=10



different data distributions – correlated case



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 MINECRAFT!

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 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 ($P > 0.05$).

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 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 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

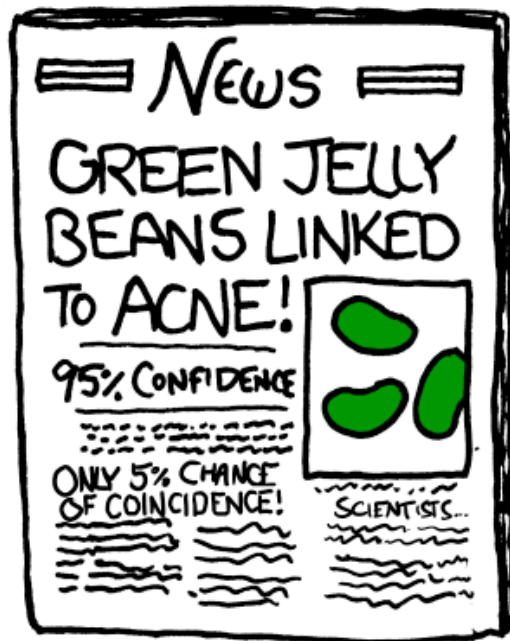
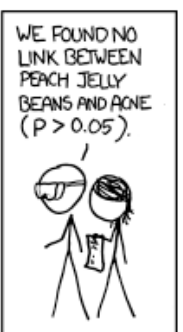
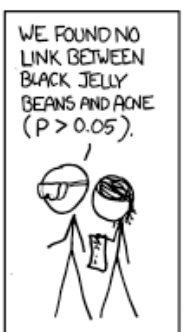
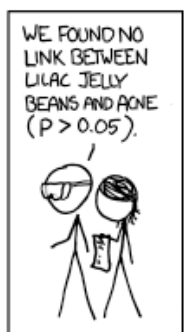
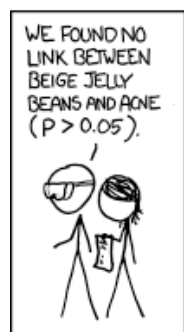
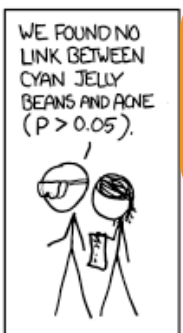
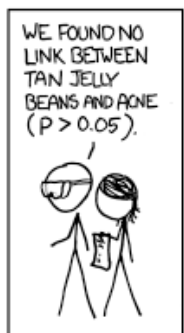
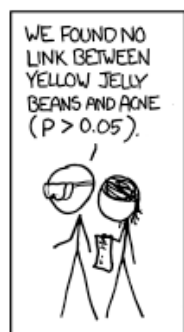
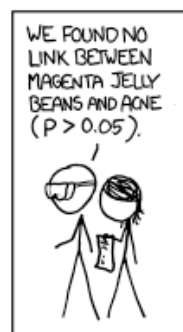
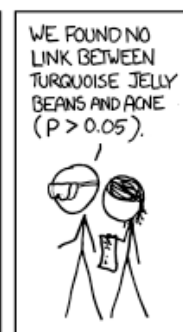
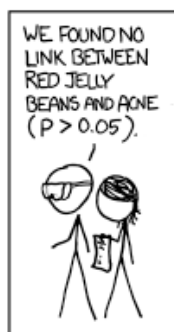
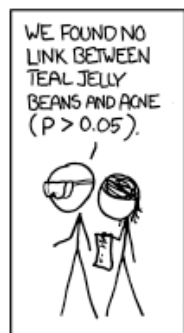
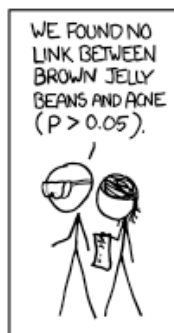
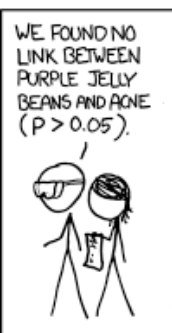
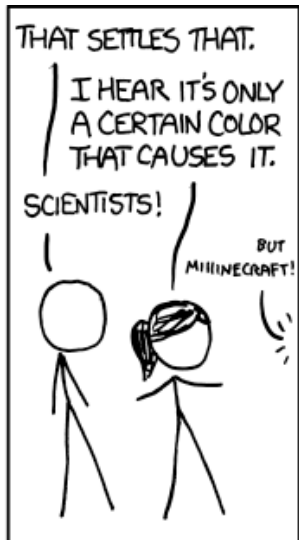
GREEN JELLY BEANS LINKED TO ACNE!

95% CONFIDENCE

ONLY 5% CHANCE OF COINCIDENCE!

SCIENTISTS...





The Multiple Testing Problem

When performing a large number of tests, the type I error goes up: for $\alpha=0.05$ and performing n tests, the probability of no false positive result is:

$$\underbrace{0.95 \cdot 0.95 \cdot \dots \cdot 0.95}_{n\text{-times}} \lll 0.95$$

⇒ The larger the number of tests performed, the higher the probability of a false rejection!

Multiple Testing Examples

Many data analysis approaches in genomics rely on item-by-item (i.e. multiple) testing:

Microarray or RNA-Seq expression profiles of “normal” vs “perturbed” samples: gene-by-gene

ChIP-chip: locus-by-locus

RNAi and chemical compound screens

Genome-wide association studies: marker-by-marker

QTL analysis: marker-by-marker and trait-by-trait

False positive rate and false discovery rate

FPR: fraction of FP among all genes (etc.) tested

FDR: fraction of FP among hits called

**Example:
20,000 genes, 100 hits, 10 of them wrong.**

FPR: 0.05%

FDR: 10%



"Wait a minute! Isn't anyone here a real sheep?"

Experiment-wide type I error rates

	Not rejected	Rejected	Total
True null hypotheses	U	V	m
False null hypotheses	T	S	m
Total	$m - R$	R	m

Family-wise error rate (FWER): $P(V > 0)$, the probability of one or more false positives. For large m_0 , this is difficult to keep small.

False discovery rate (FDR): $E[V / \max\{R, 1\}]$, the expected fraction of false positives among all discoveries.

FWER: The Bonferroni correction

Suppose we conduct a hypothesis test for each gene $g = 1, \dots, m$, producing

an observed test statistic: T_g

an unadjusted p -value: p_g .

Bonferroni adjusted p -values:

$$\tilde{p}_g = \min(mp_g, 1).$$

Selecting all genes with $\tilde{p}_g \leq \alpha$ controls the FWER at level α , that is, $Pr(V > 0) \leq \alpha$.

Controlling the FDR (Benjamini/Hochberg)

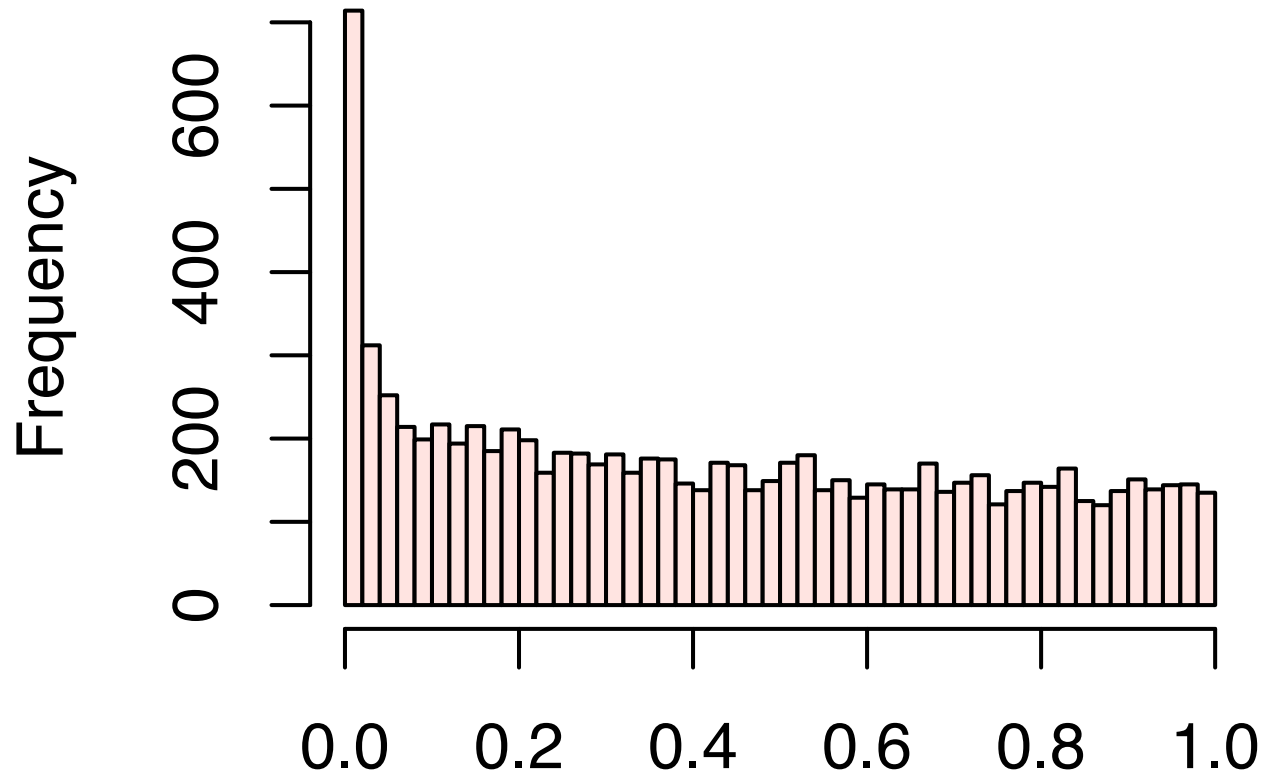
- FDR: the expected proportion of false positives among the significant genes.
- Ordered unadjusted p -values: $p_{r_1} \leq p_{r_2} \leq \dots \leq p_{r_m}$.
- To control $FDR = E(V/R)$ at level α , let

$$j^* = \max\{j : p_{r_j} \leq (j/m)\alpha\}.$$

Reject the hypotheses H_{r_j} for $j = 1, \dots, j^*$.

- Is valid for independent test statistics and for some types of dependence.

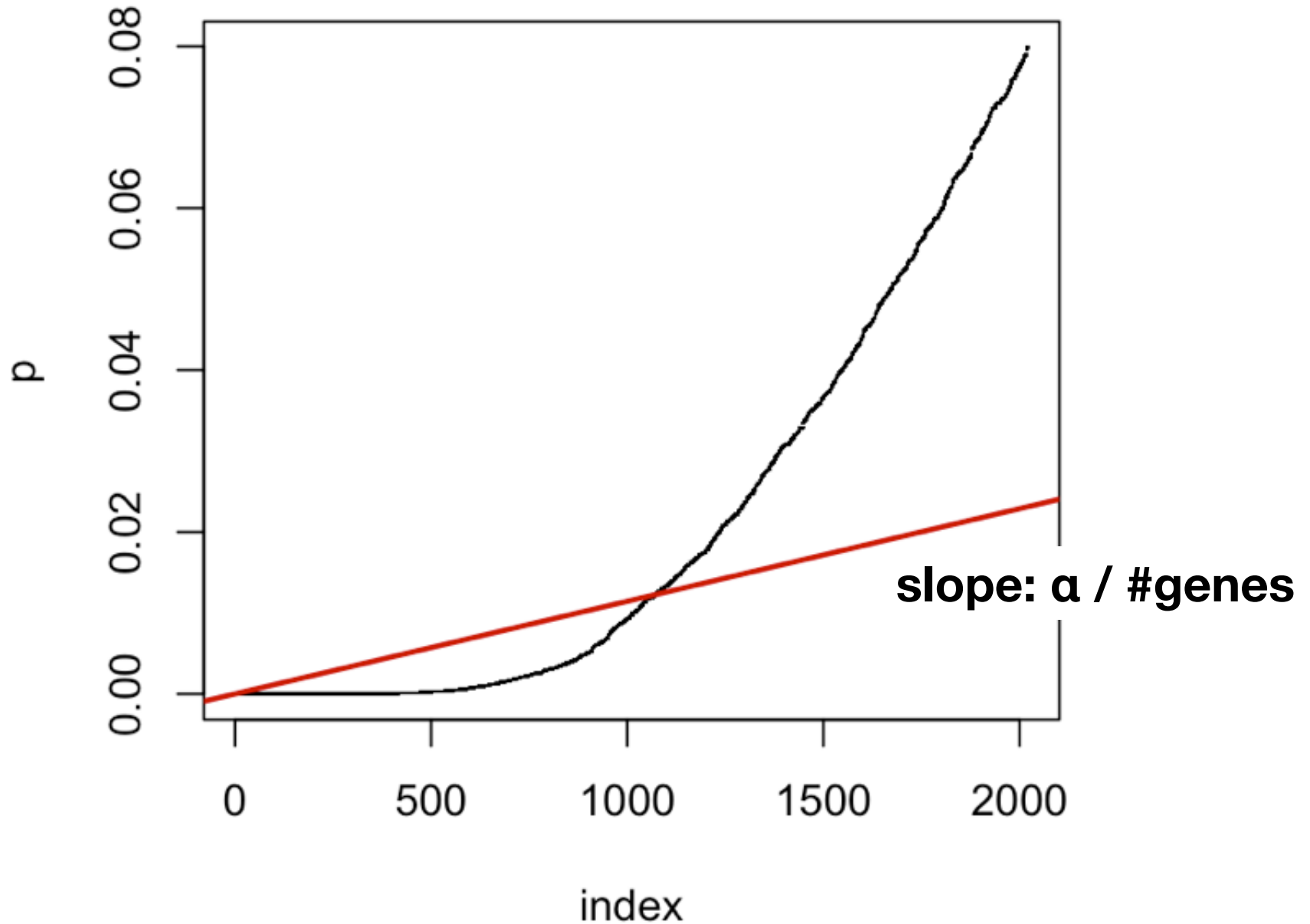
Diagnostic plot: the histogram of p-values



Observed p-values are a mix of samples from

- a uniform distribution (from true nulls) and
- from distributions concentrated at 0 (from true alternatives)

Benjamini Hochberg multiple testing adjustment



Benjamini Hochberg multiple testing adjustment

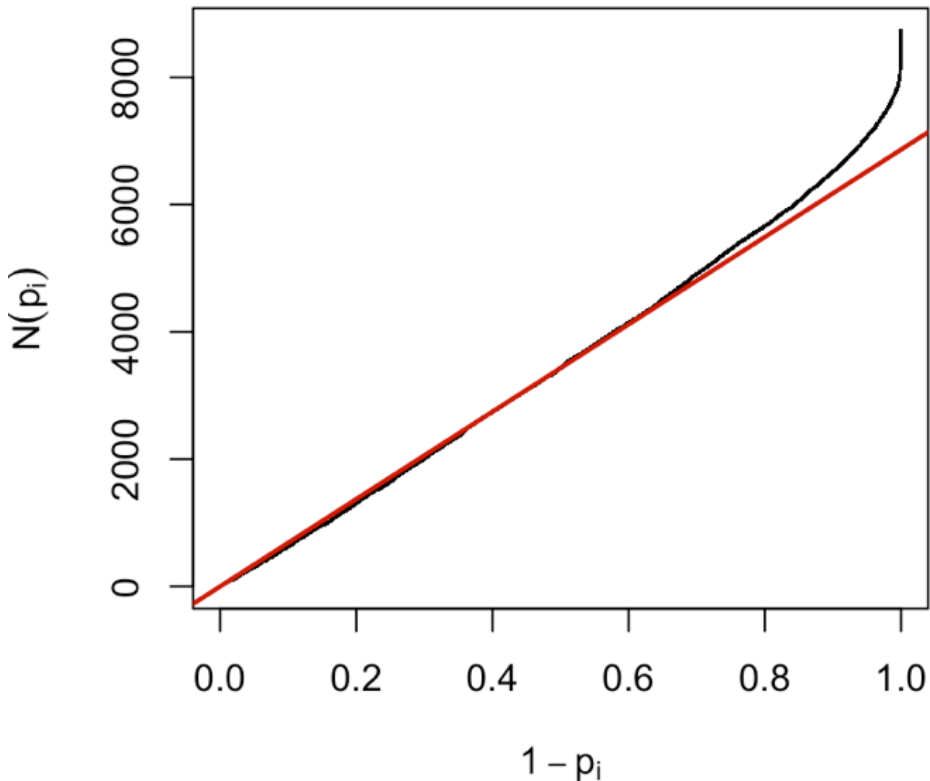


```
p  BH = {  
    i <- length(p) : 1  
    o <- order(p, decreasing = TRUE)  
    ro <- order(o)  
    pmin(1, cummin(n/i * p[o]))[ro]  
}
```

0 500 1000 1500 2000

index

How to estimate the number (not: the identity) of differentially expressed genes



For a series of hypothesis tests $H_1 \dots H_m$ with p-values p_i , plot

$$(1 - p_i, N(p_i)) \quad \text{for all } i$$

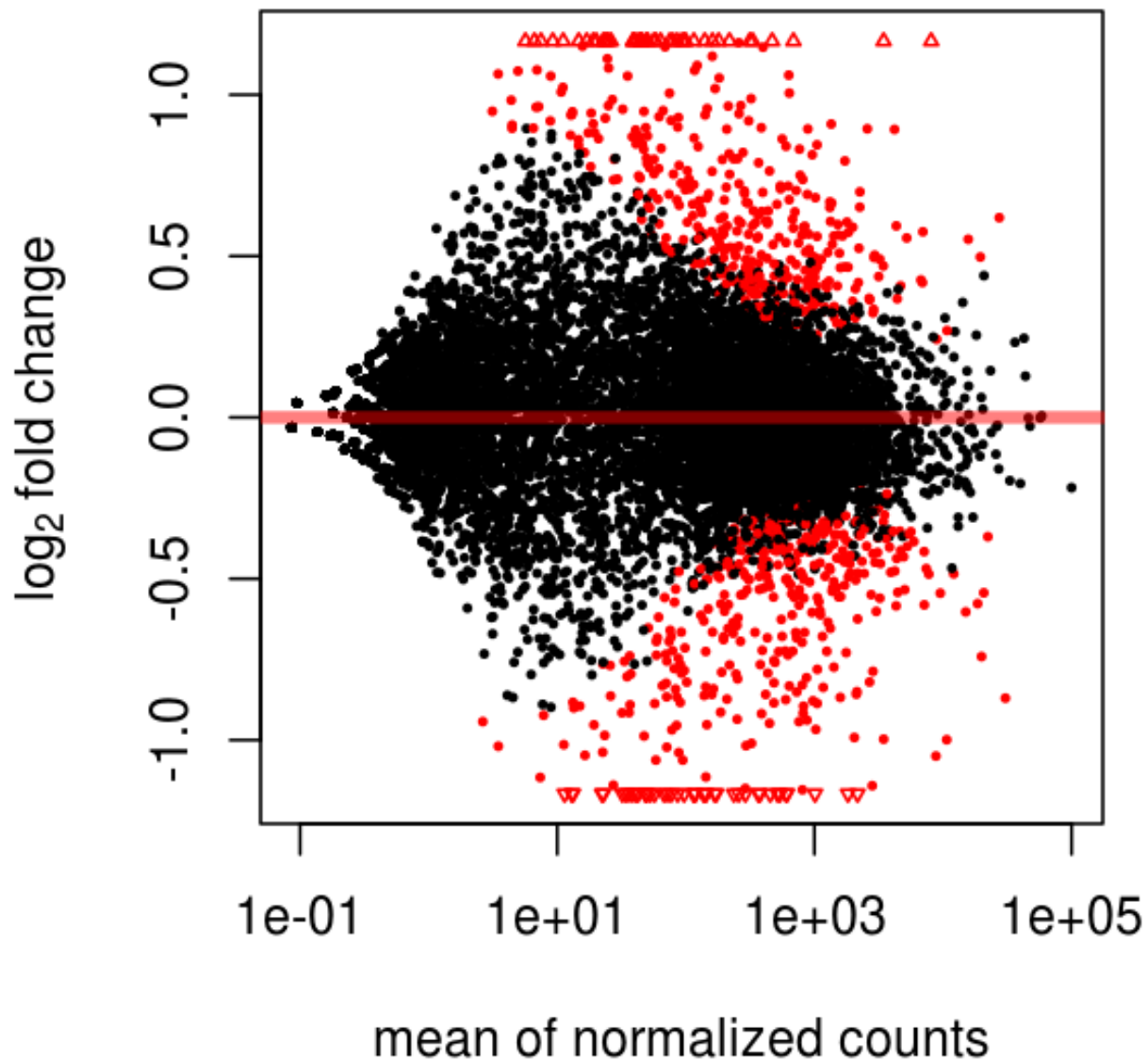
where $N(p)$ is the number of p-values greater than p .

Red line: $(1 - p_i, (1 - p)^*m)$

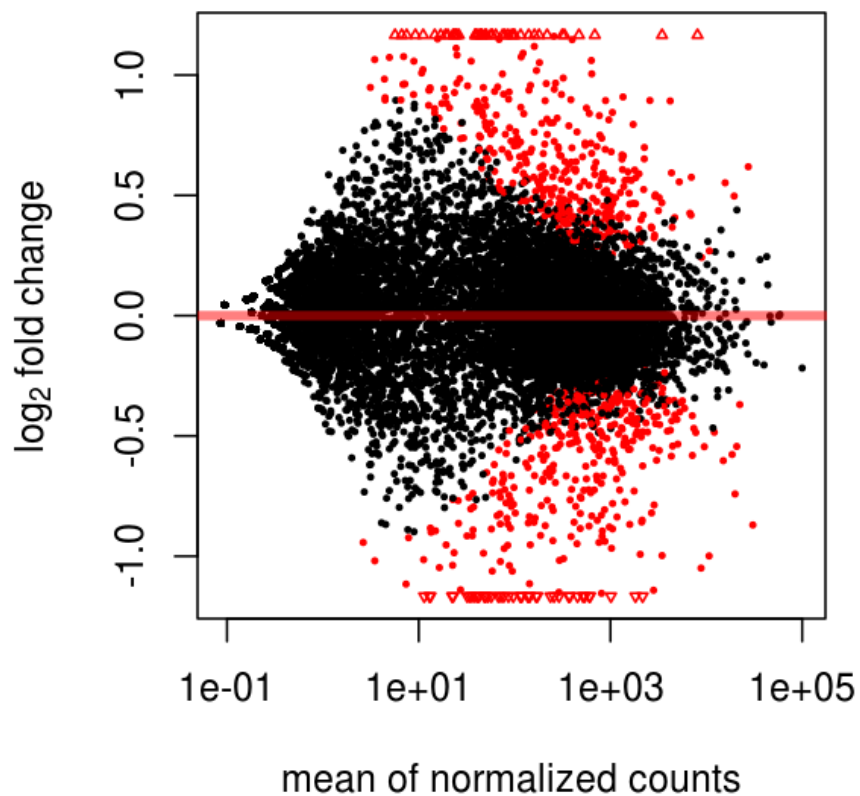
$(1 - p)^*m =$ expected number of p-values greater than p

Schweder T, Spjøtvoll E (1982) Plots of P-values to evaluate many tests simultaneously. *Biometrika* 69:493–502.
See 'genefilter' vignette for an example.

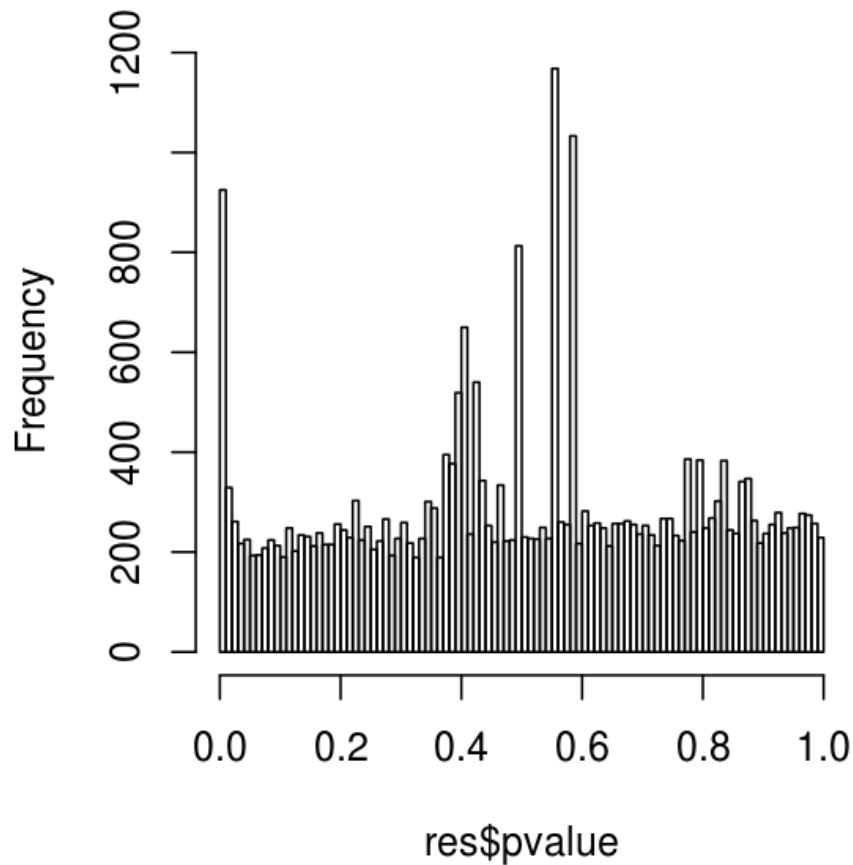
parathyroid dataset



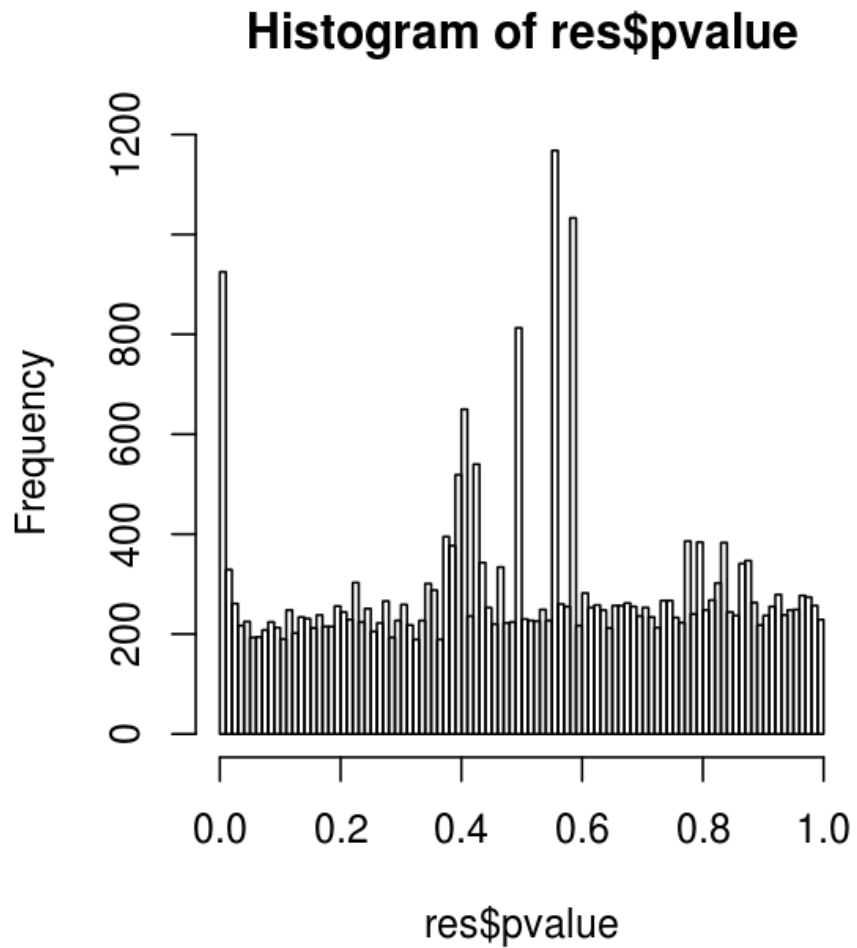
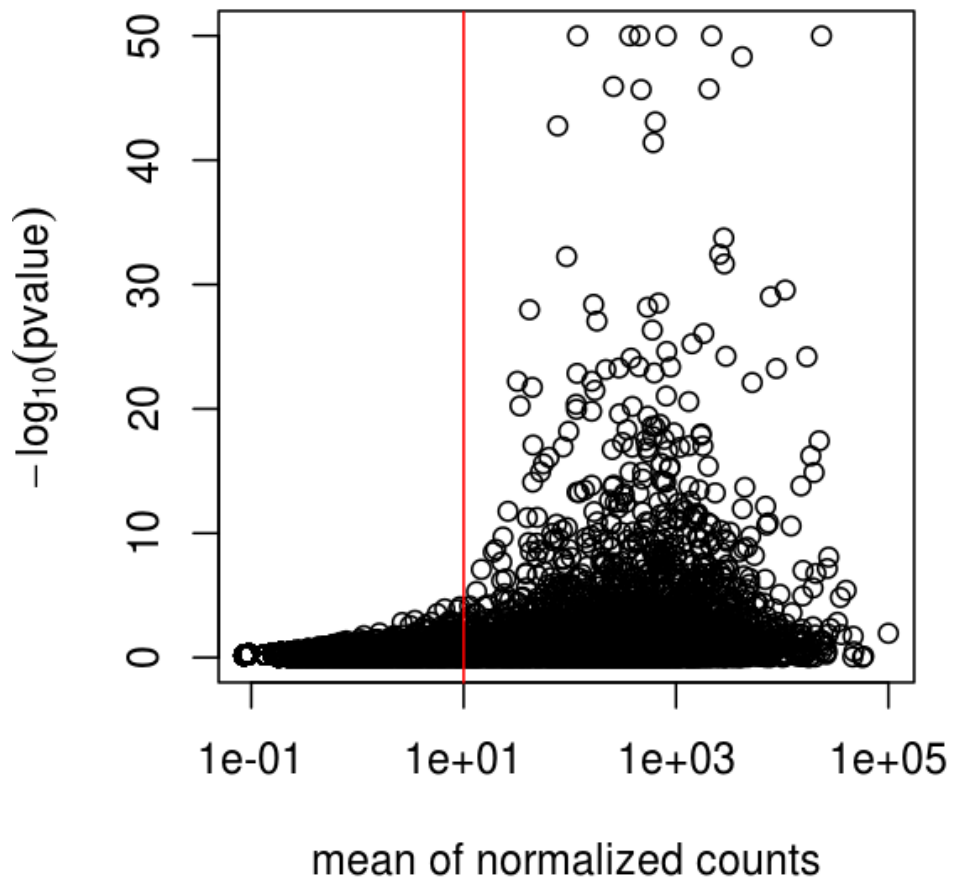
parathyroid dataset



Histogram of res\$pvalue



parathyroid dataset



Independent filtering

**From the set of all tests to be done,
first filter out those that seem to have insufficient power
anyway,
then formally test for differential expression on the rest.**

Literature

von Heydebreck, Huber, Gentleman (2004)

Chiaretti et al., Clinical Cancer Research (2005)

McClintick and Edenberg (BMC Bioinf. 2006) and references therein

Hackstadt and Hess (BMC Bioinf. 2009)

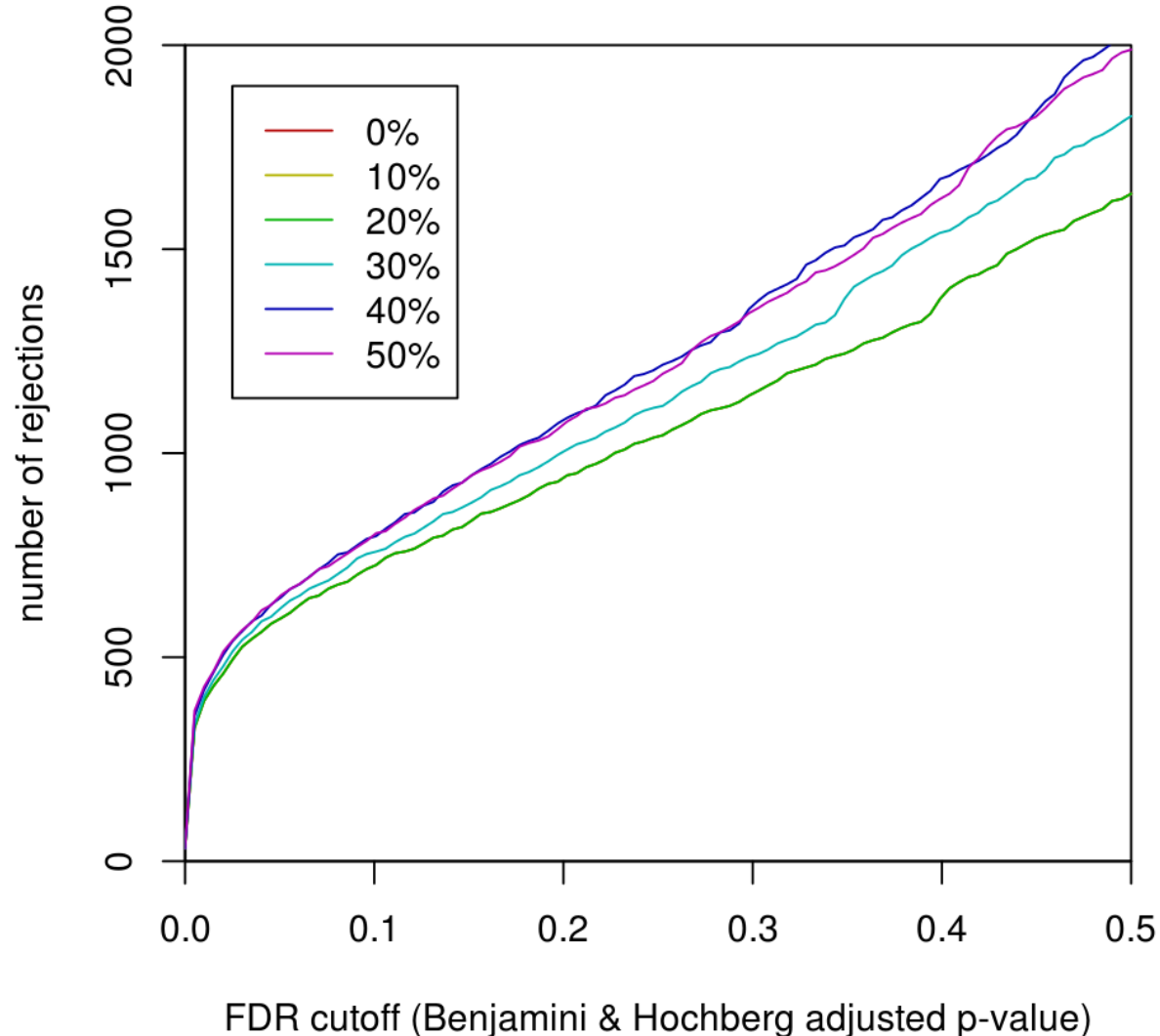
Bourgon, Gentleman and Huber (PNAS 2010)

Many others.

Increased detection rates

Stage 1 filter: sum of counts, across samples, for each gene, and remove the fraction (10%, 20%, ...) of genes where that is smallest

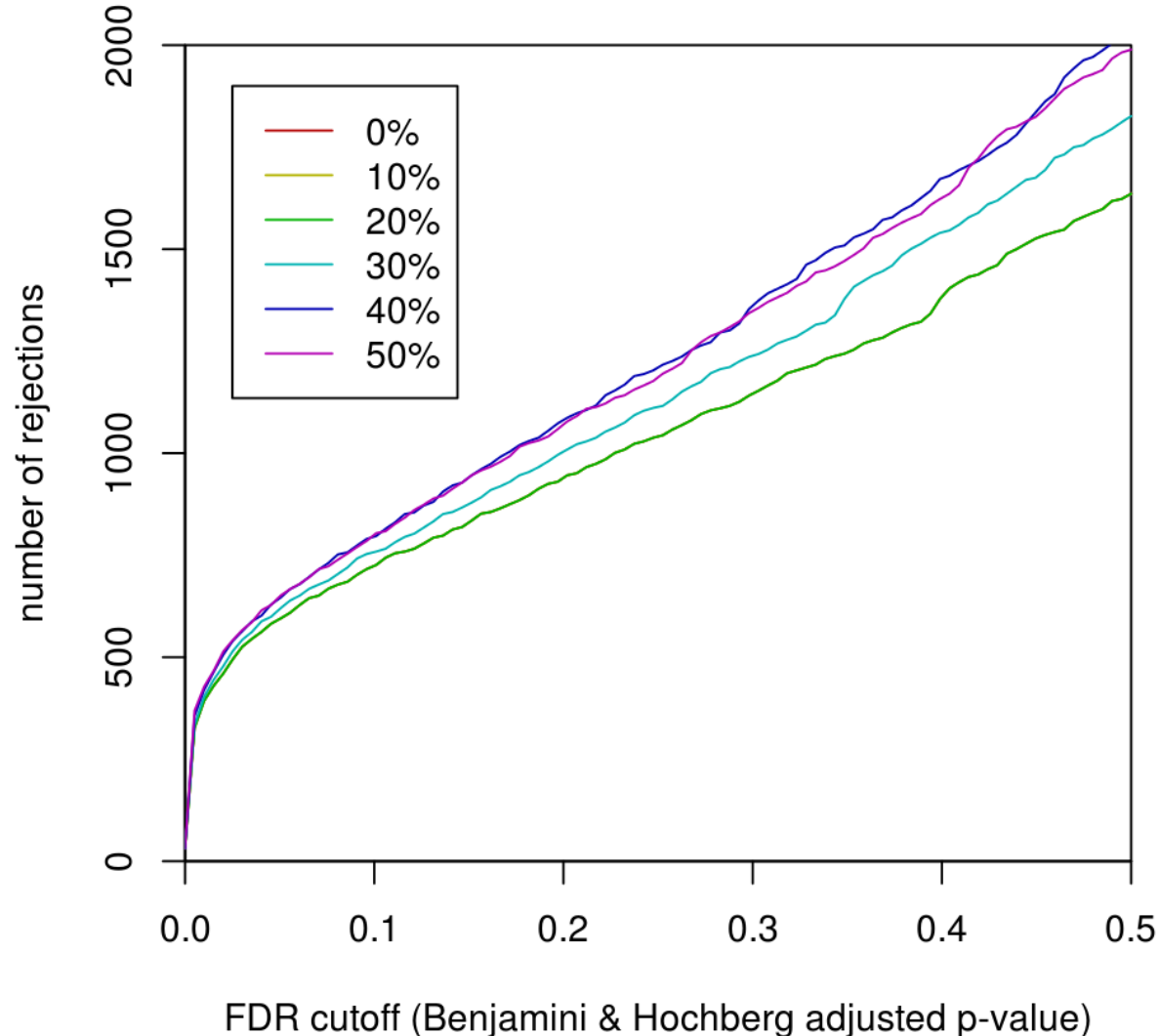
Stage 2: standard NB-GLM test



Increased power?

Increased detection rate implies increased power

only if we are still controlling type I errors at the same level as before.



Increased power?

Increased detection rate implies increased power

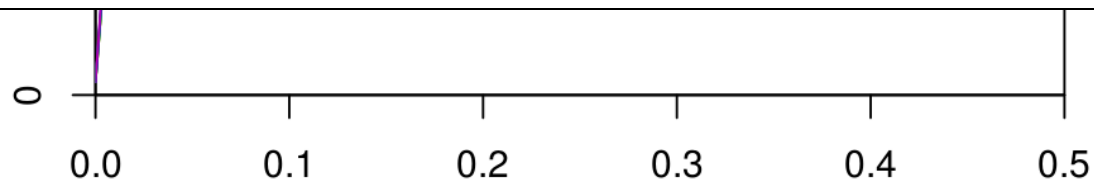
only if we are still controlling type I errors at the same level as before.

Concern:

Since we use a data-driven criterion in stage 1, but do p-value and type-I error related computations only on the genes in stage 2, aren't we 'cheating'?

Informal justification:

Filter does not use covariate information



FDR cutoff (Benjamini & Hochberg adjusted p-value)

What do we need for experiment-wide type I error (e.g.: FDR) control?

- I. Per gene p-values must be bona-fide p-values: for those genes for which H_0 holds, p must be Uniform distributed.
- II. Joint distribution of the p-values must comply with the assumptions of the multiple testing procedure (e.g. Benjamini-Hochberg)

What do we need for experiment-wide type I error (e.g.: FDR) control?

I. Per gene p-values must be bona-fide p-values: for those genes for which H_0 holds, p must be Uniform distributed.

II. Joint distribution of the p-values must comply with the assumptions of the multiple testing procedure (e.g. Benjamini-Hochberg)

If these conditions hold without filtering, and if the filtering is statistically independent from the test statistics under the null, they still hold with filtering.

(Bourgon, Gentleman, Huber, PNAS 2010)

Independence of filter and test statistics under the null hypothesis

For genes for which the null hypothesis is true (X_1, \dots, X_n exchangeable), f (filter) and g (test) are statistically independent in all of the following cases:

- **NB-test (DESeq2):**

f : overall count sum (or mean)

- **Normally distributed data (e.g. microarray data after `rma` or `vsN`):**

f : overall variance, overall mean

g : standard two-sample t-statistic, or any test statistic which is scale and location invariant.

- **Non-parametrically:**

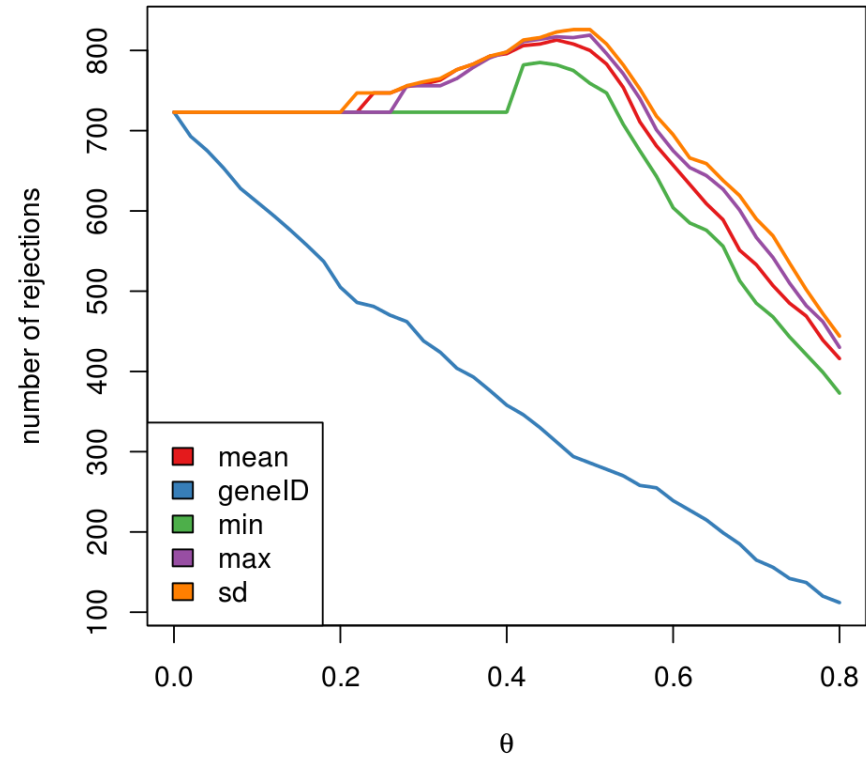
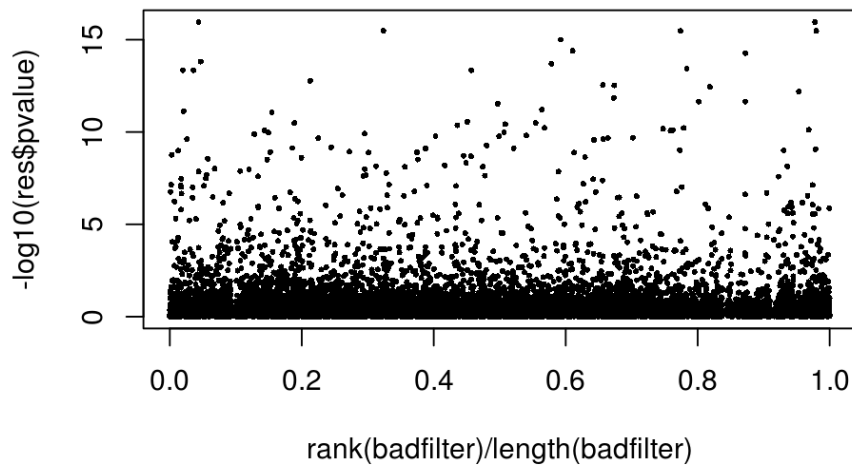
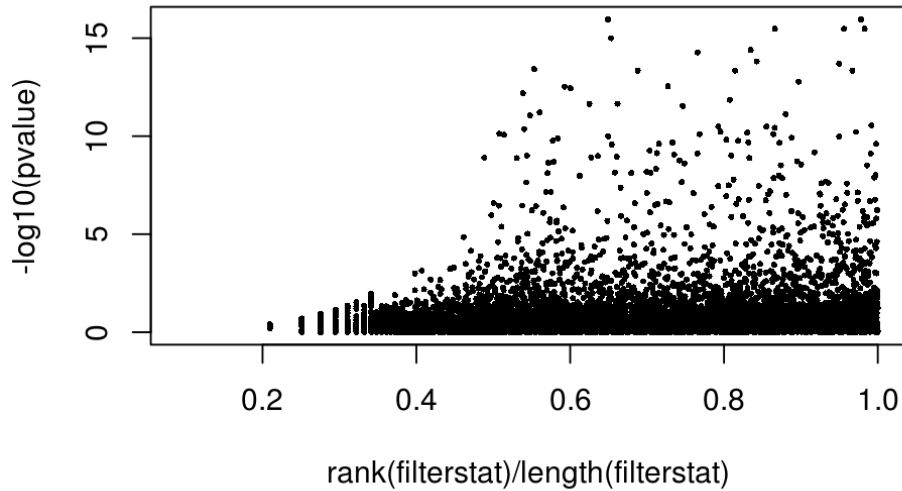
f : any function that does not depend on the order of the arguments. E.g. overall variance, IQR.

g : the Wilcoxon rank sum test statistic.

Also in the multi-class context: ANOVA, Kruskal-Wallis.

Diagnostics

(see: vignettes of genefilter, DESeq2 packages)



Conclusion

Independent filtering can substantially increase your power at same type I error.

Conclusion

Independent filtering can substantially increase your power at same type I error.



References

Bourgon R., Gentleman R. and Huber W. Independent filtering increases detection power for high-throughput experiments, PNAS (2010)

Bioconductor package `genefilter` vignette: Diagnostics for independent filtering

DESeq2 vignette

**Richard
Bourgon**

**Robert
Gentleman**

**Michael
Love**

Thank you



A G A G T T C T G C T C G
A G G G T T A T G C G C G
C G T T C G G G A A T C C
C G T T A G G A A A T C T
T C T T T G A C G A C T C

STUCK IN A DULL, LOW PAYING JOB?
WANT TO MAKE **BIG MONEY?**

**BE A
QUANTUM
MECHANIC!**

... EVEN IF YOU NEVER
FINISHED HIGH SCHOOL!

STUDY AT HOME!

THE COLUMBIA INSTITUTE OF QUANTUM MECHANICS, INC.

Not affiliated with the Columbia Broadcasting System, Columbia University, the District of Columbia, or Columbia, Gem of the Ocean.



CUT OUT AND SEND

Yes! I want to get in on the ground floor of this exciting new field. I understand no salesman will call.

NAME _____

ADDRESS _____

CITY, STATE, ZIP _____

COLUMBIA INSTITUTE OF QUANTUM MECHANICS

Suite 293, 1100 Back St., Providence, RI 02904