

Hypothesis Testing



Wolfgang Huber, Bernd Klaus, 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

Hypothesis Testing

General idea: Set up a “null” hypotheses

H_0 : a model of reality which lets us make specific predictions of how the data should look like.

If we can show that the probability of getting the actually observed data (if H_0 is true) is small, then we can ‘reject’ H_0 and conclude that something else is likely to be true.

Examples of null hypotheses:

- The coin is fair
- The new drug is no better (or worse) than a placebo
- The observed CellTitreGlo signal is no different from that of negative controls

Example Hypothesis Testing

Toss a coin a given number of times \Rightarrow

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

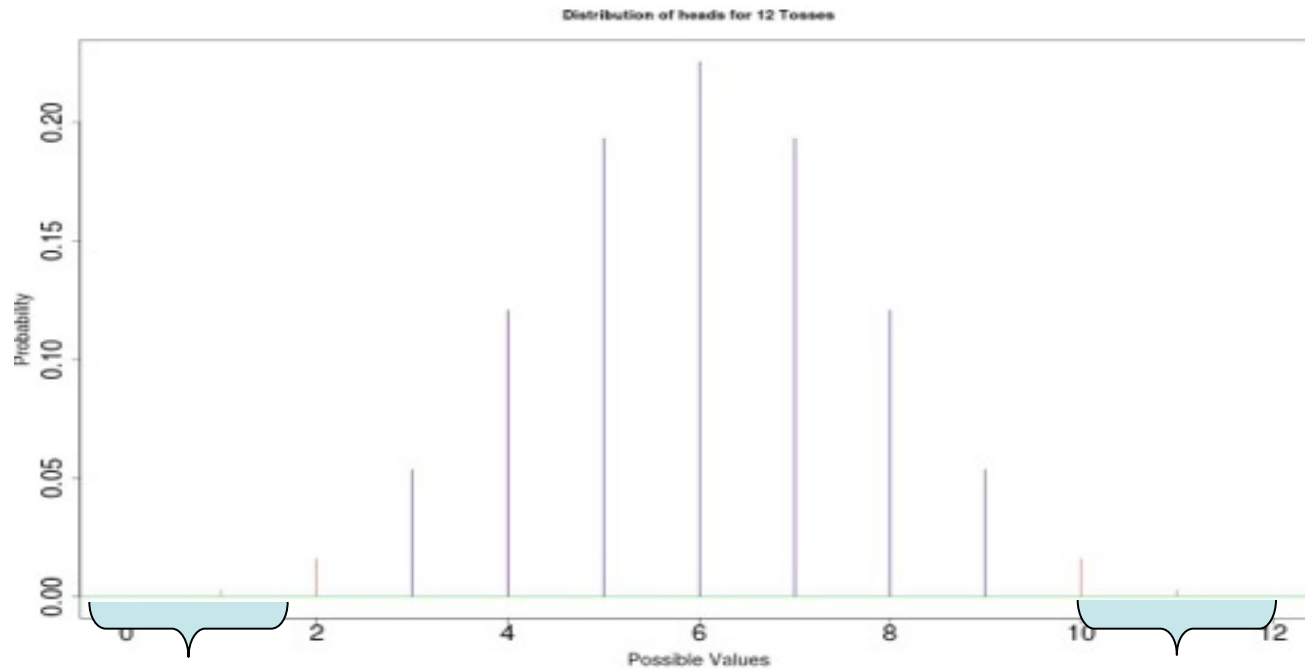
But what is “roughly”? We use combinatorics / probability theory to quantify this.

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 unprobable to observe extreme events such as more than 9 heads

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

If we observe 10 heads in a trial the null hypotheses 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		Decision	
		not rejected ('negative')	rejected ('positive')
Truth	H ₀ true	True negative (specificity)	False Positive Type I error α
	H ₀ false	False Negative Type II error β	True Positive (sensitivity)

False positive rate and false discovery rate

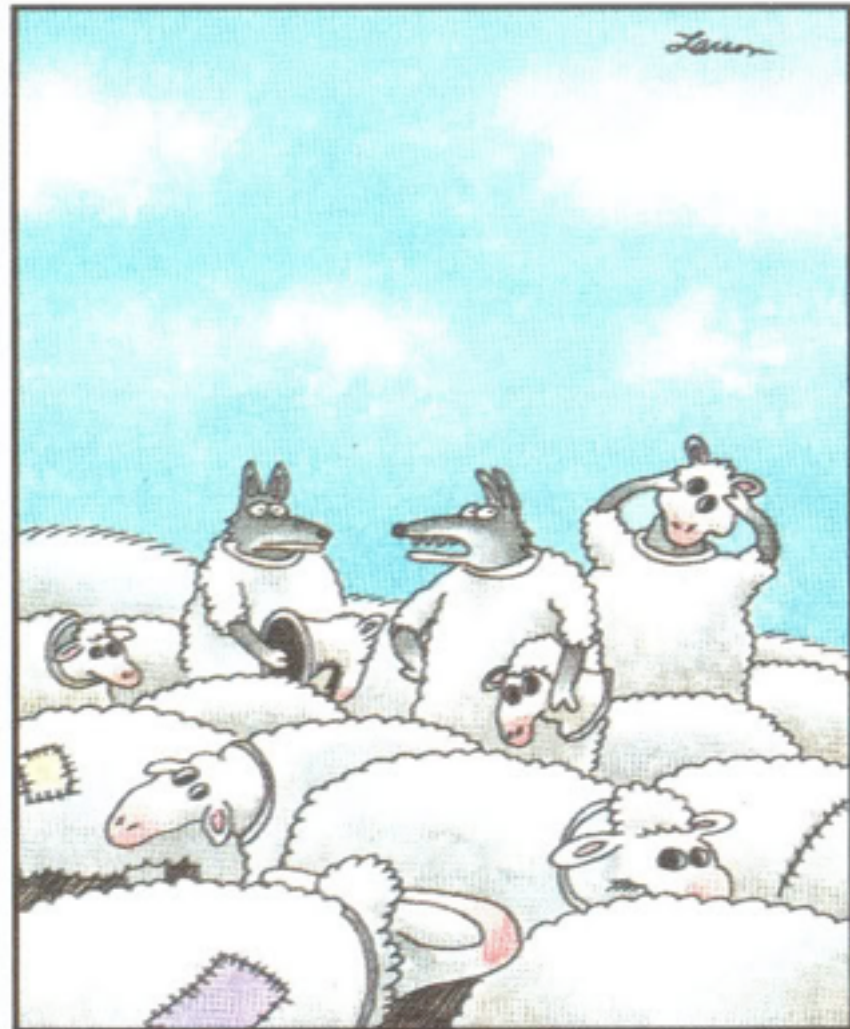
FPR: fraction of FP among all genes (etc.) tested

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?"

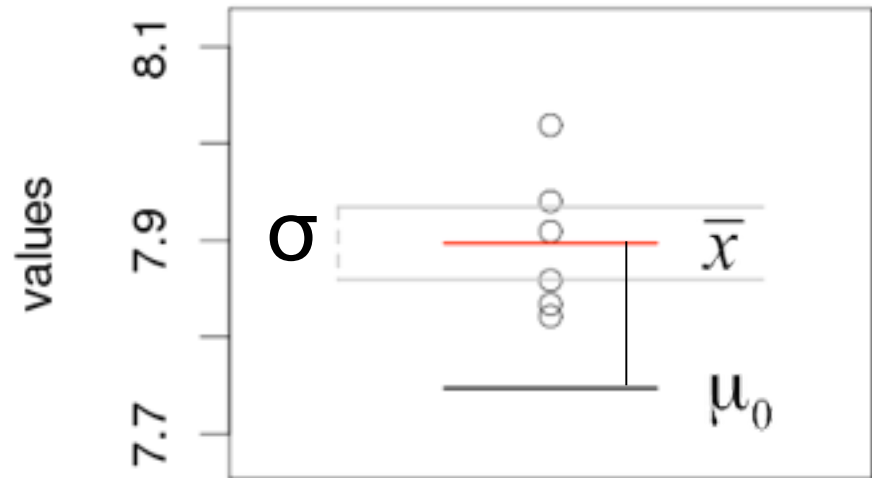
One sample t-test

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

One sample t-test: compare to a fixed value μ_0

Without n: z-score

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



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

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:

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

t-test

If H_0 is correct, t follows a known distribution: t-distribution

The shape of the t-distribution depends on the number of observations: if the average is made of n observations, it follows the t-distribution with $n-1$ degrees of freedom (T_{n-1}).

If n is large, T_{n-1} is close to a normal distribution

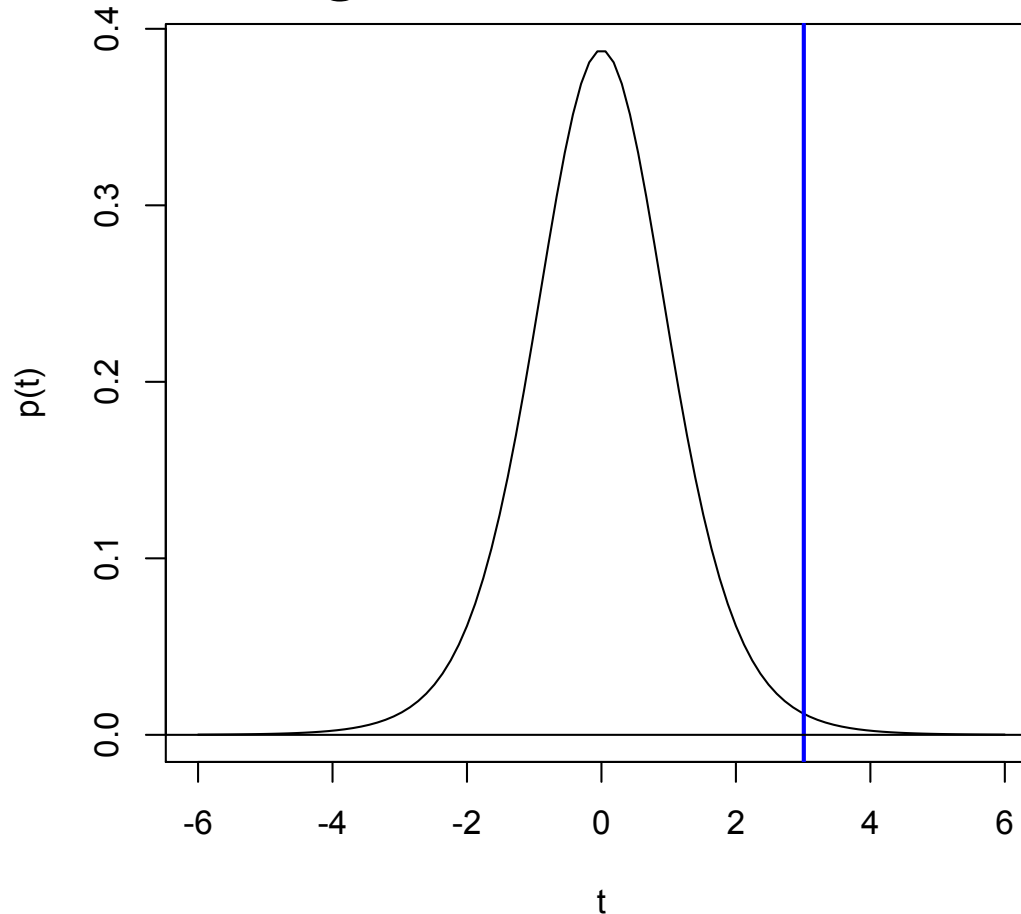
If n is small, T_{n-1} is more spread out than a normal distribution.

This penalty takes into account that the data-based estimate of the standard deviation can underestimate* the true value.

(*in principle: also overestimate)

p-value and test decision

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



p-value: $P(|t| \geq 3.01) = 0.014$

Avoid fallacy

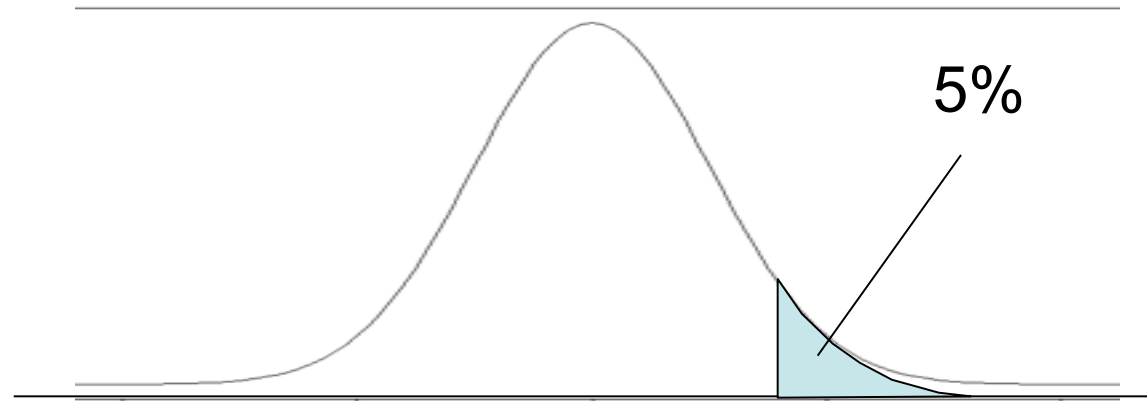
The p-value is the probability that the observed data could happen, under the condition that the null hypothesis is true.

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

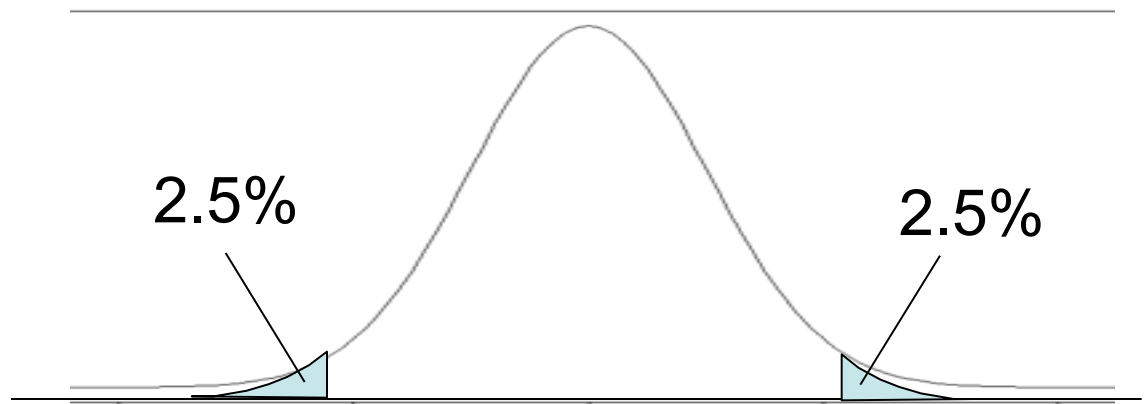
Absence of evidence \neq evidence of absence

One-sided vs two-sided test

One-sided
e.g. $H_A: \mu > 0$



Two-sided
e.g. $H_A: \mu \neq 0$



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 both 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 the number of observations in each sample).

Comments and pitfalls

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

Some deviations from Normality, e.g. heavier tails, are actually rarely a problem for the t-test, unsymmetric (skewed) distributions are \Rightarrow use Wilcoxon tests based on ranks!

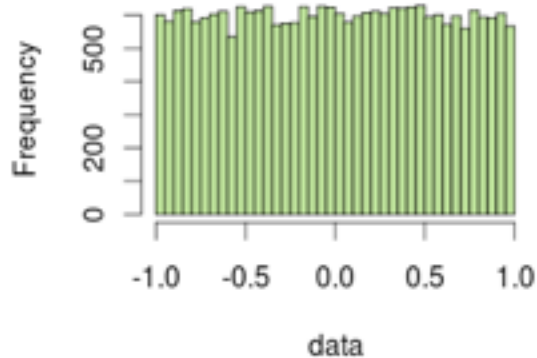
If the data are dependent, then p-values will likely be totally wrong (e.g., for positive correlation, too optimistic).

different data distributions – independent case

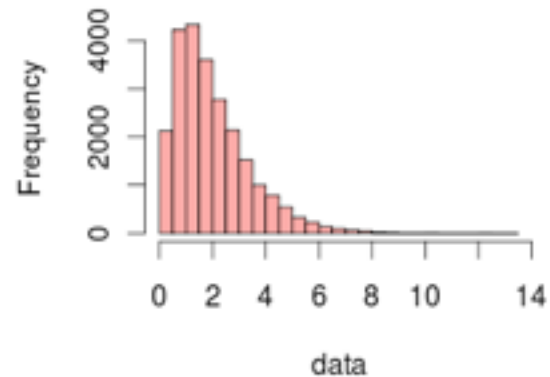
Normal(0,1)



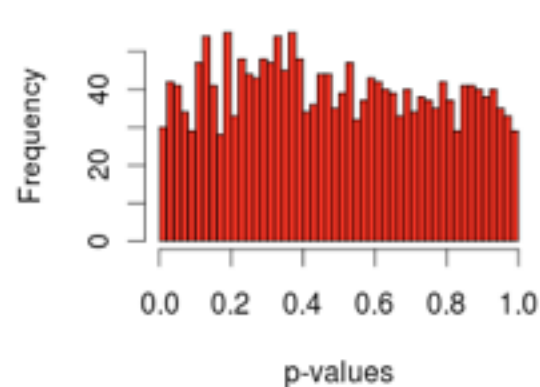
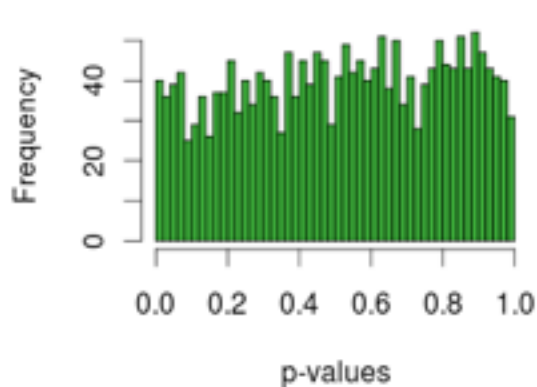
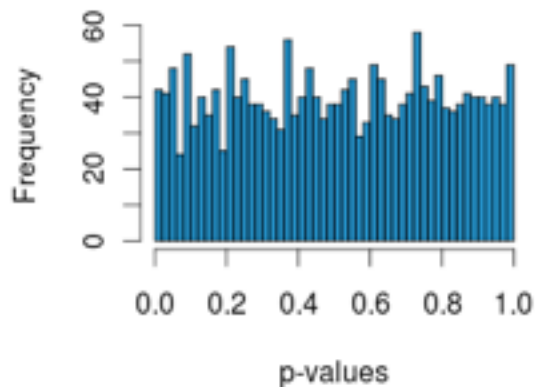
Uniform(-1,1)



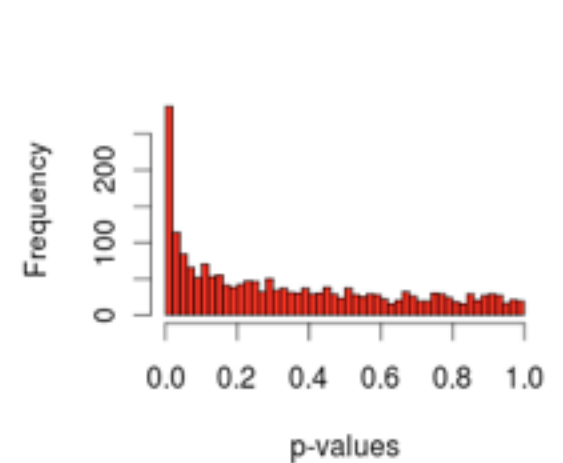
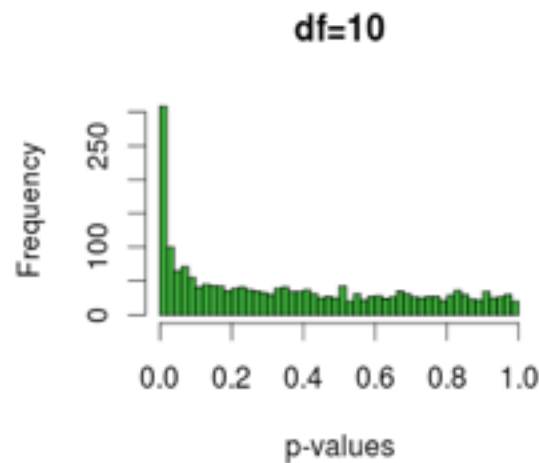
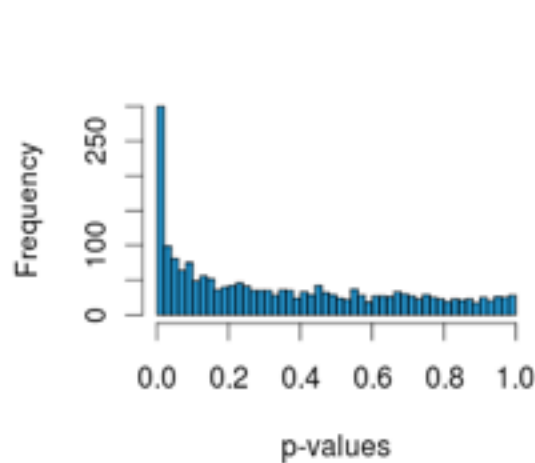
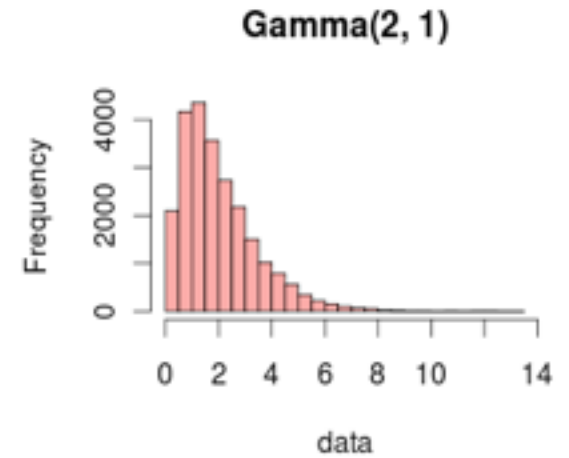
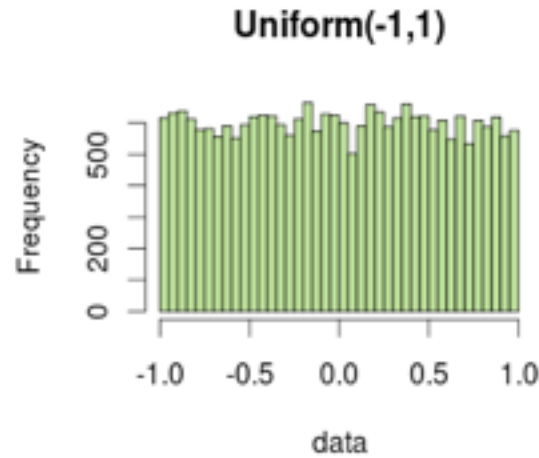
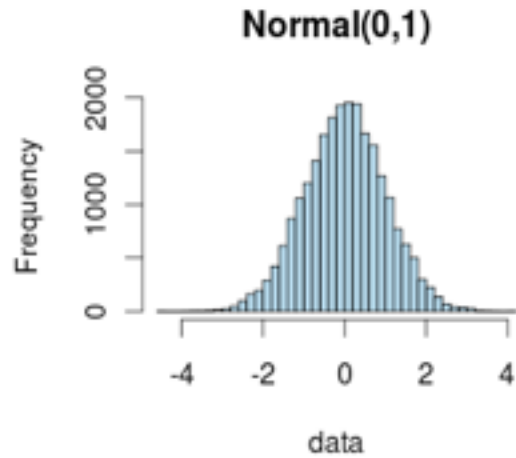
Gamma(2, 1)



df=10

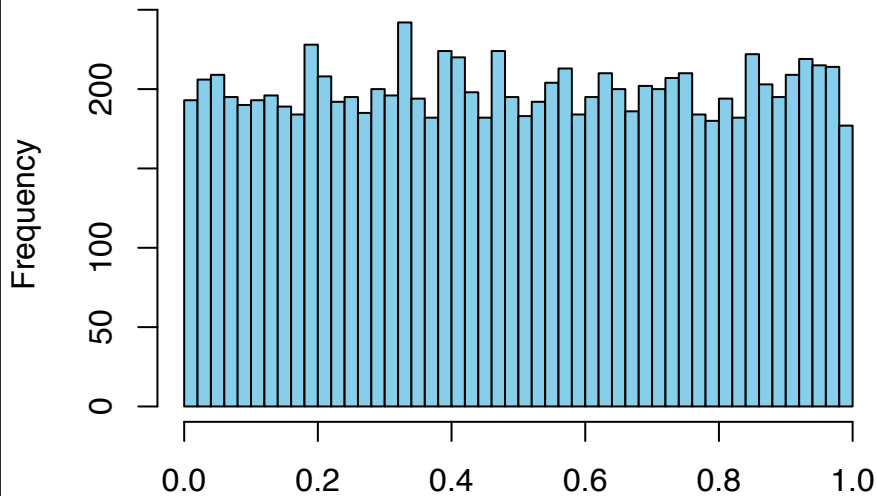


different data distributions – correlated case

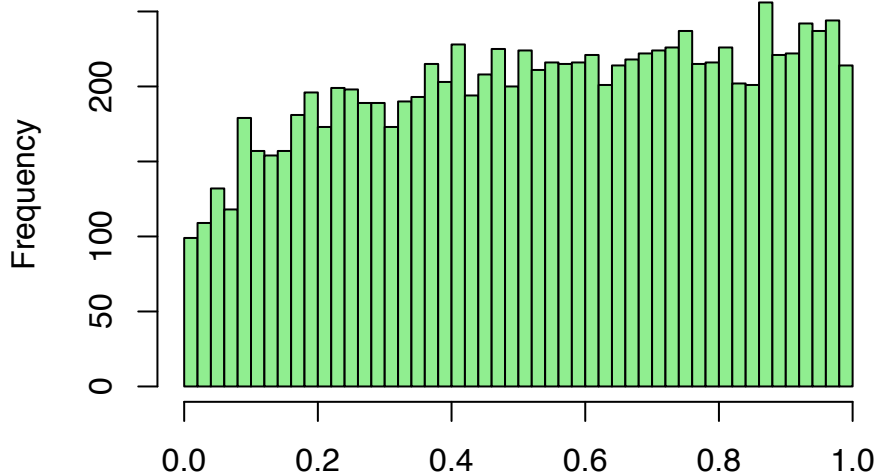


Batch effects or “latent variables”

Histogram of `rt1$p.value`



Histogram of `rt2$p.value`



n = 10000

m = 20

```
x = matrix(rnorm(n*m), nrow=n, ncol=m)
```

```
fac = factor(c(rep(0, 10), rep(1, 10)))
```

```
rt1 = rowttests(x, fac)
```

```
x[, 6:15] = x[, 6:15]+1
```

```
rt2 = rowttests(x, fac)
```

sva package; Leek JT, Storey JD.
Capturing heterogeneity in gene
expression studies by surrogate
variable analysis. PLoS Genet. 2007

Stegle O, Parts L, Durbin R, Winn J. A
Bayesian framework to account for
complex non-genetic factors in gene
expression levels greatly increases
power in eQTL studies. PLoS Comput
Biol. 2010.

t-test and wilcoxon 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

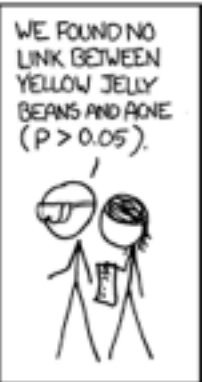
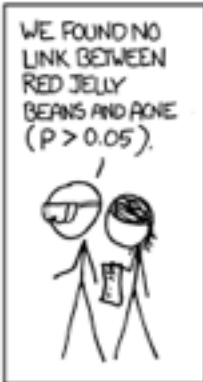
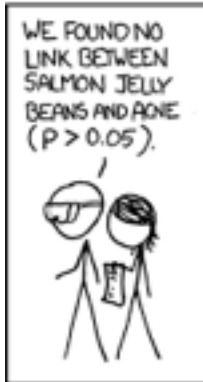
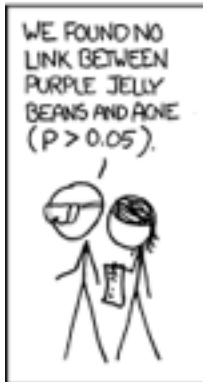
var.equal: Can the variances in the two groups assumed to be equal?

alternative: one- or two-sided test?

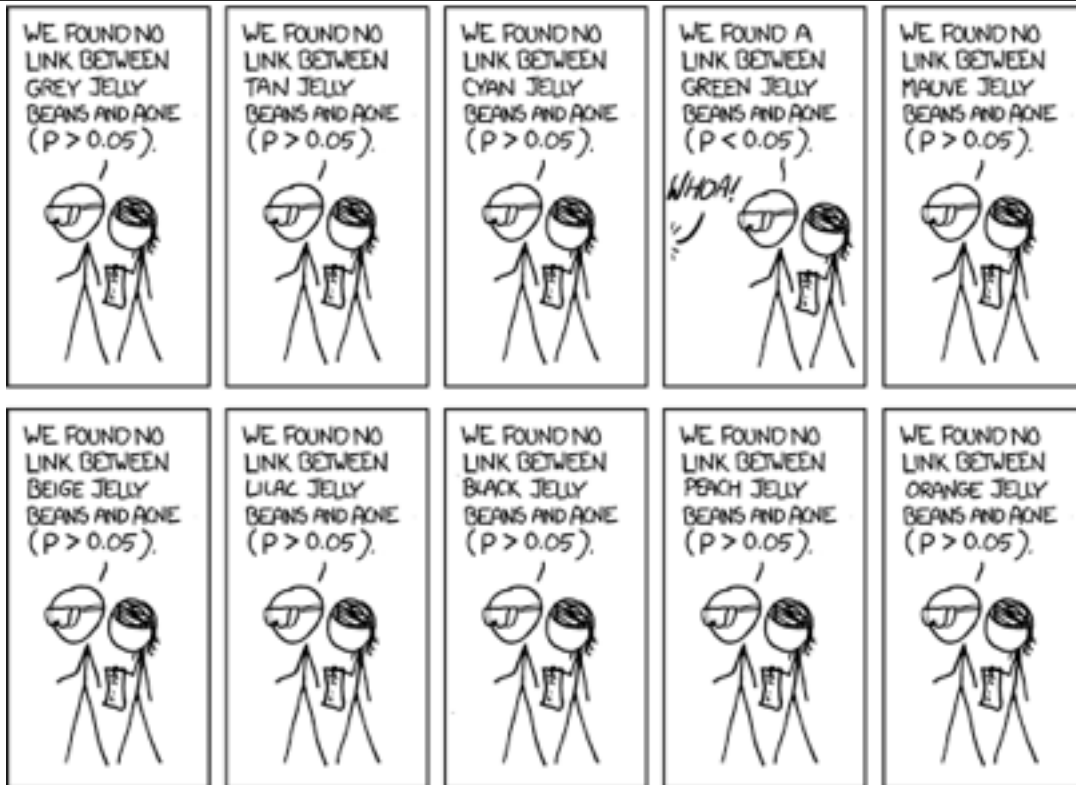
```
wilcox.test(x, y, alternative, paired, exact)
```

... just like the t-test,

exact: shall computations be performed using permutations? (slow for large samples)



xkcd




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 is inflated: 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

Experiment-wide type I error rates

	Not rejected	Rejected	Total
True null hypotheses	U	V	m_0
False null hypotheses	T	S	m_1
Total	$m - R$	R	m

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

False discovery rate: $E[V / \max\{R, 1\}]$, 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)

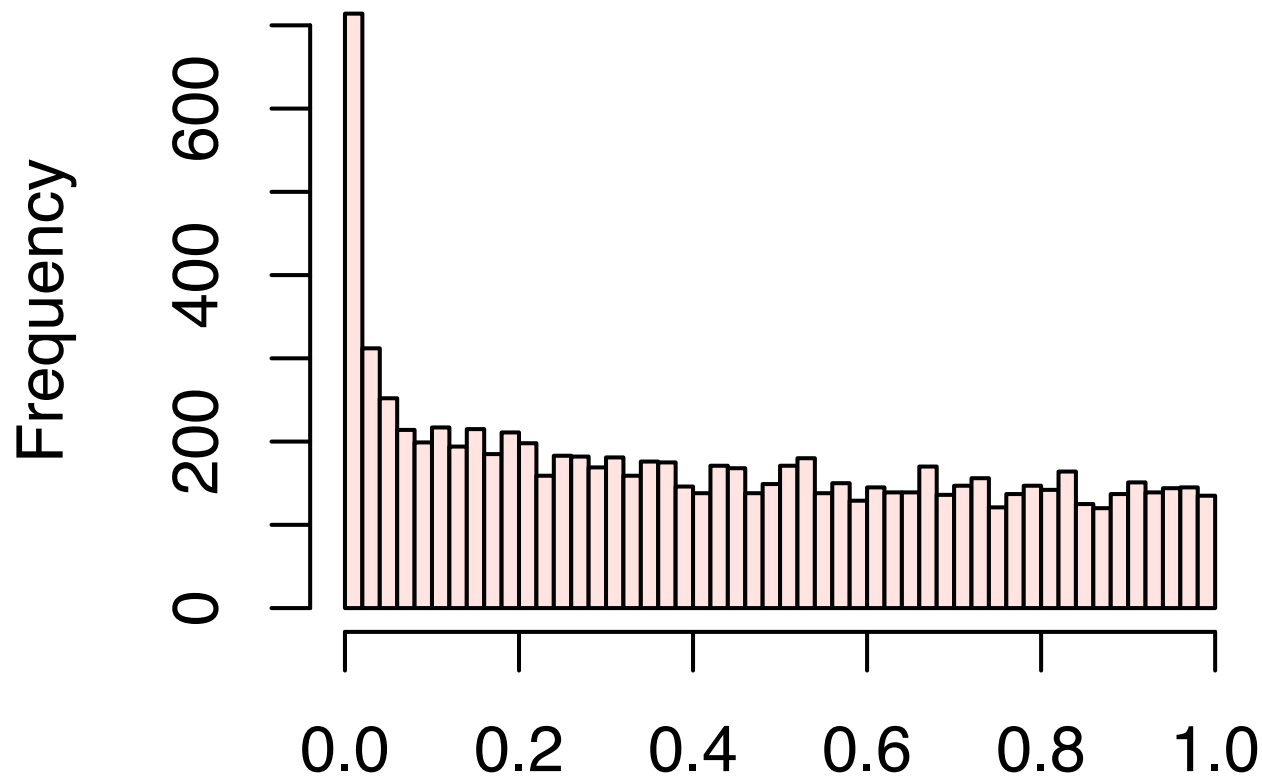
- FDR: the expected proportion of false positives among the significant genes.
- Ordered unadjusted p -values: $p_{r_1} \leq p_{r_2} \leq \dots \leq p_{r_m}$.
- To control $FDR = E(V/R)$ at level α , let

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

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

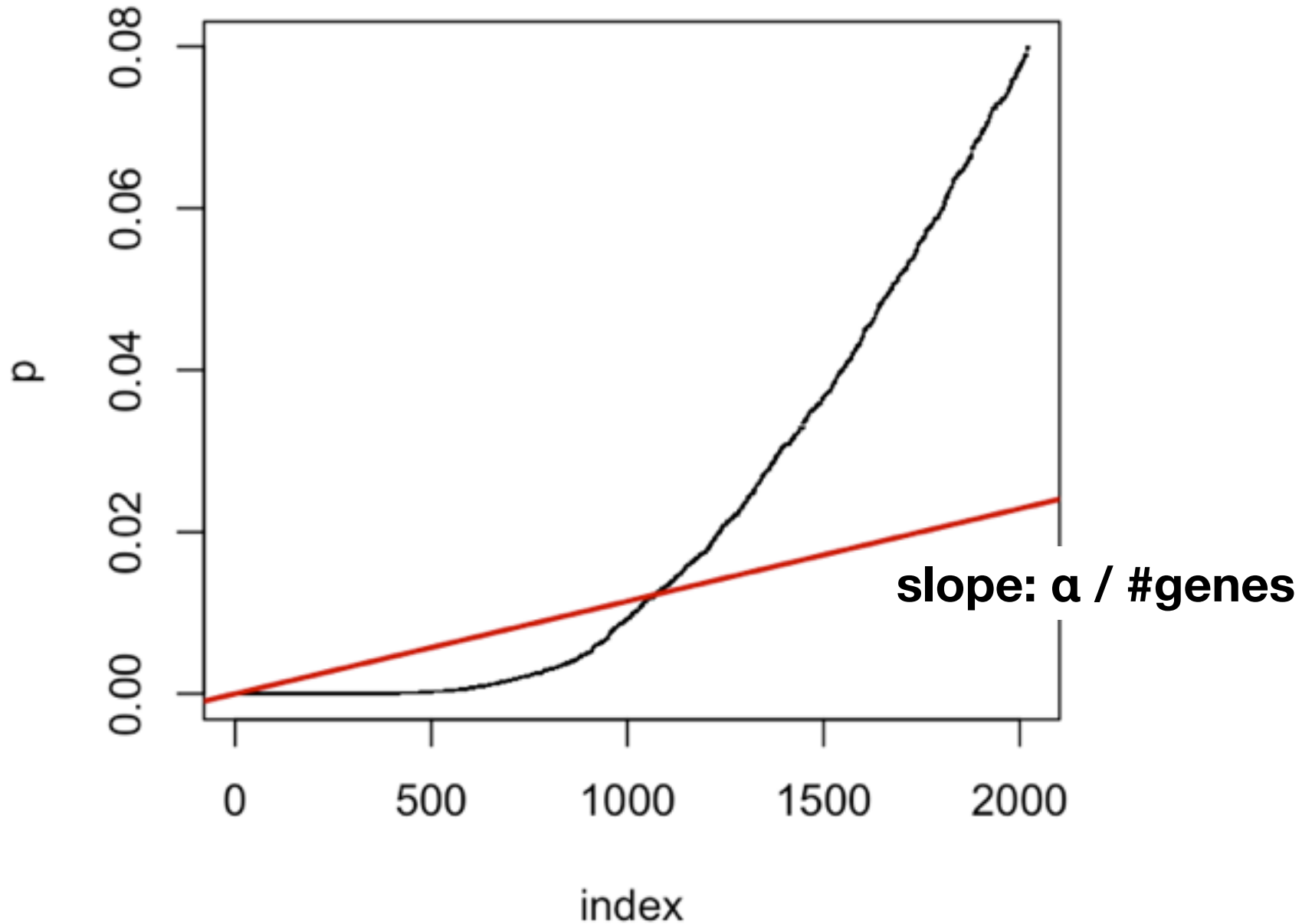
- Is valid for independent test statistics and for some types of dependence.

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

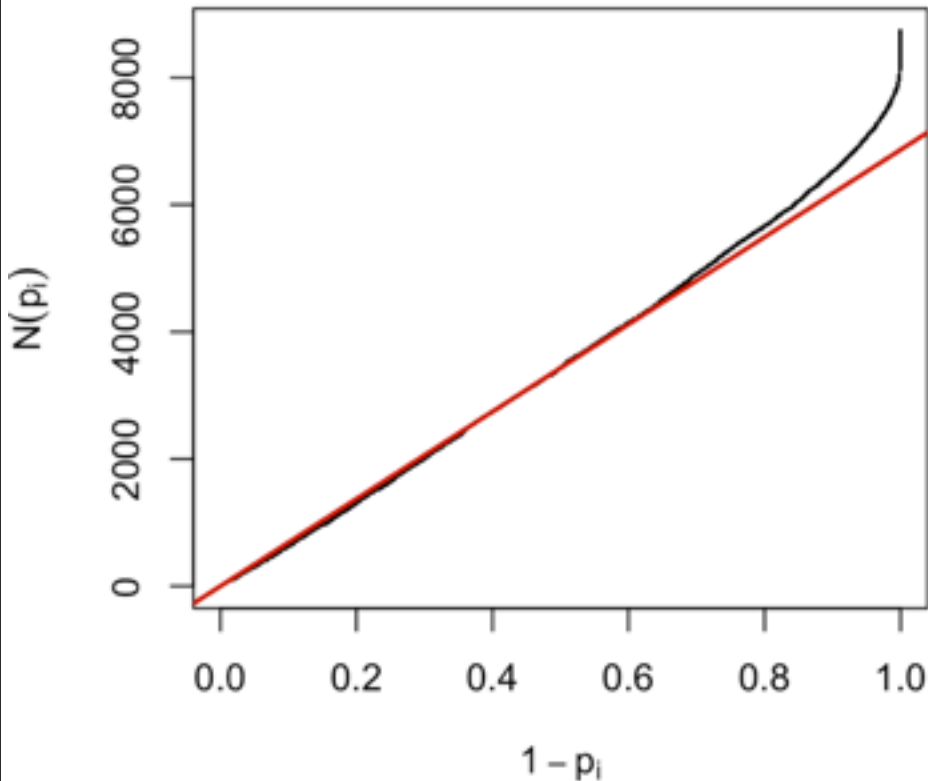


```
p BH = {  
  i <- length(p) : 1  
  o <- order(p, decreasing = TRUE)  
  ro <- order(o)  
  pmin(1, cummin(n/i * p[o]))[ro]  
}
```

0 500 1000 1500 2000

index

Schweder and Spjøtvoll p-value plot



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

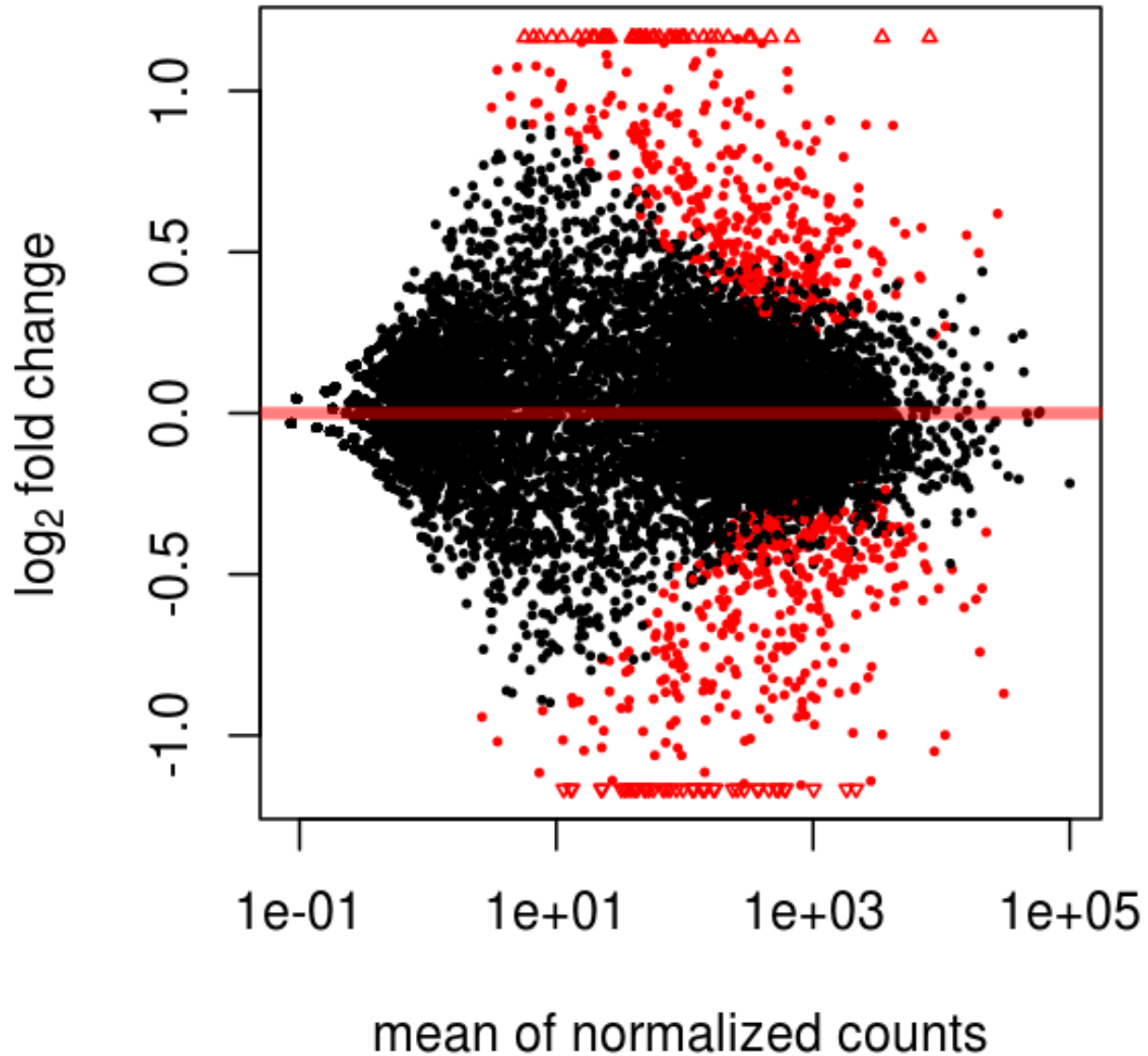
$$(1 - p_i, N(p_i)) \quad \text{for all } i$$

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

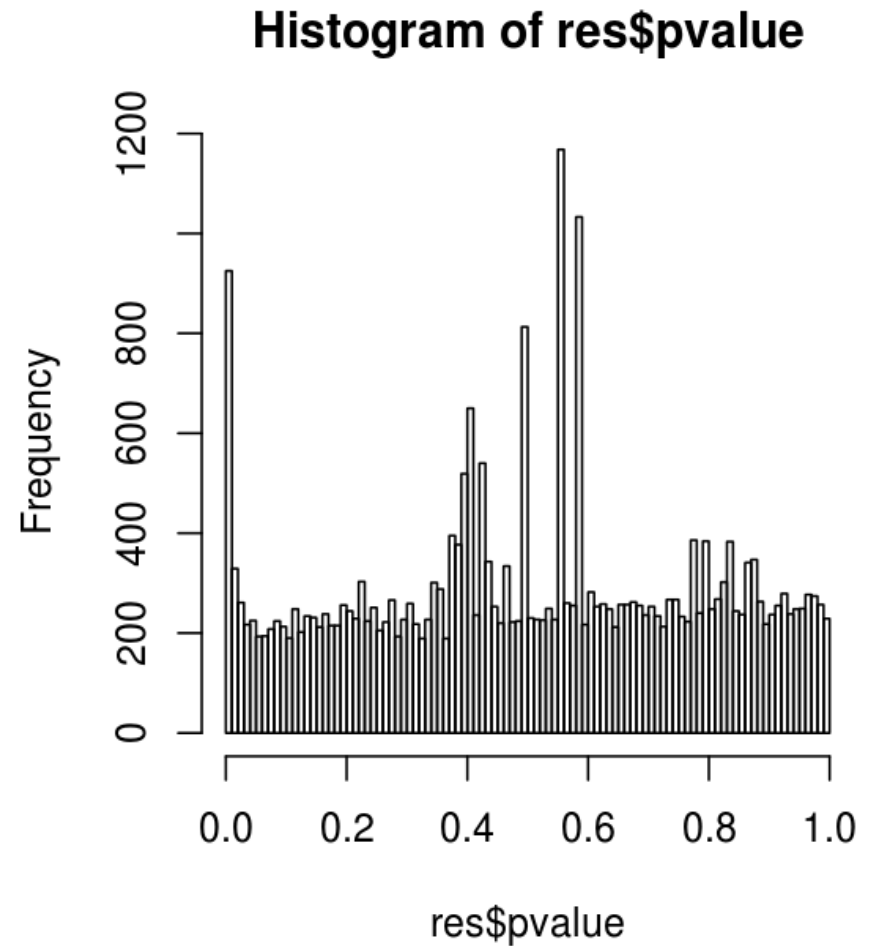
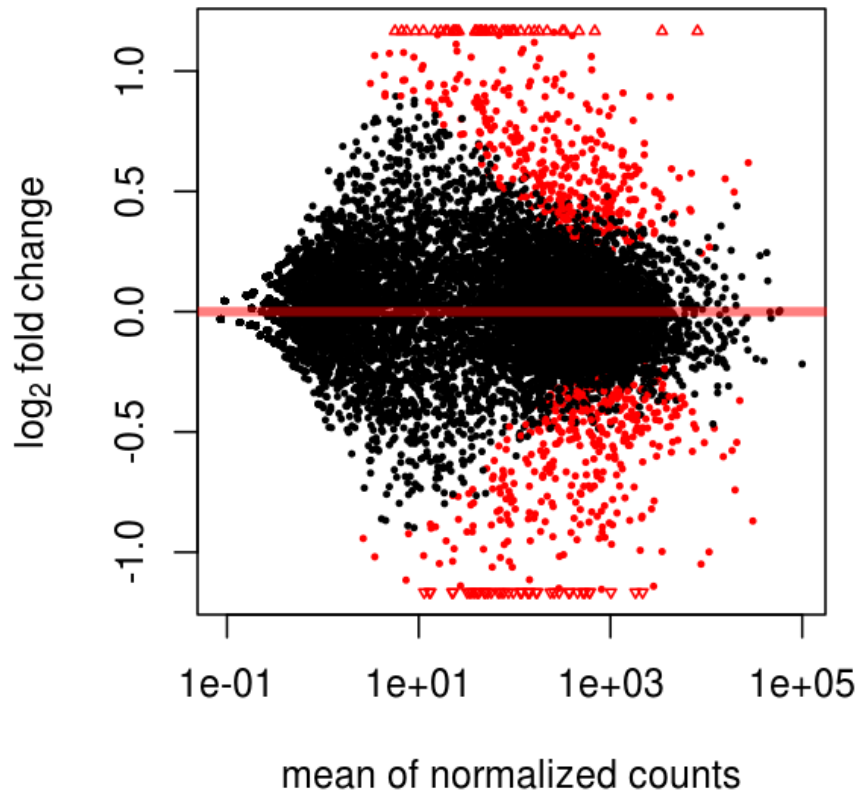
Red line: $(1 - p_i, (1 - p)^*m)$

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

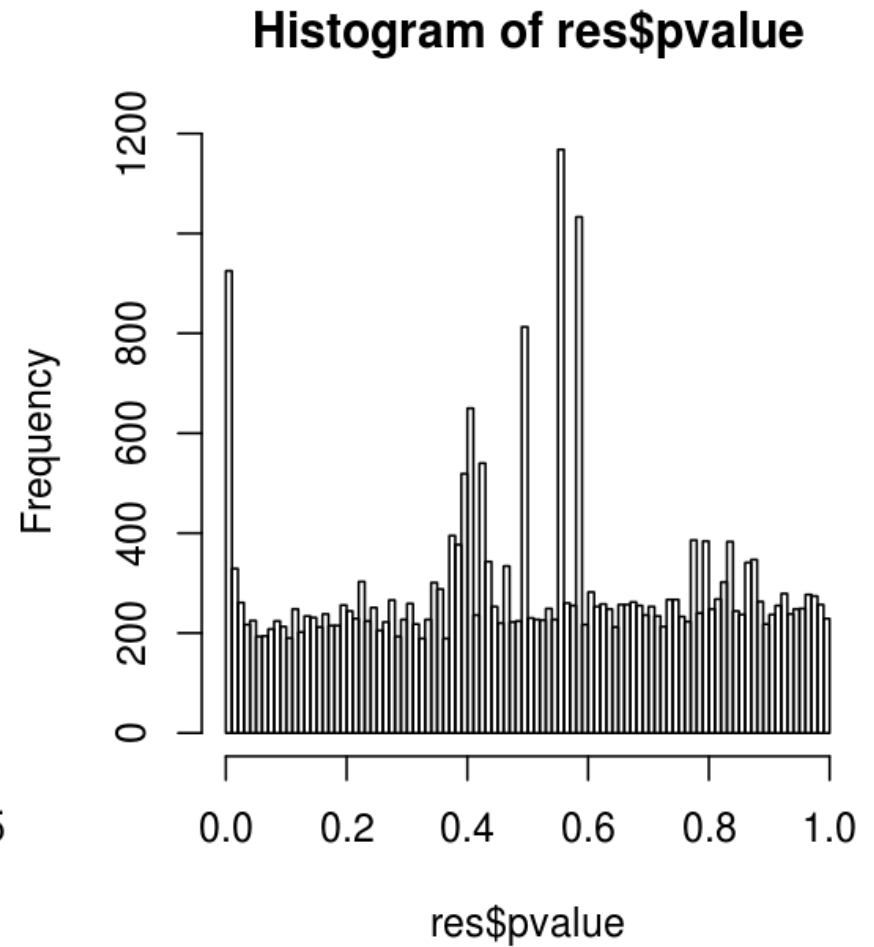
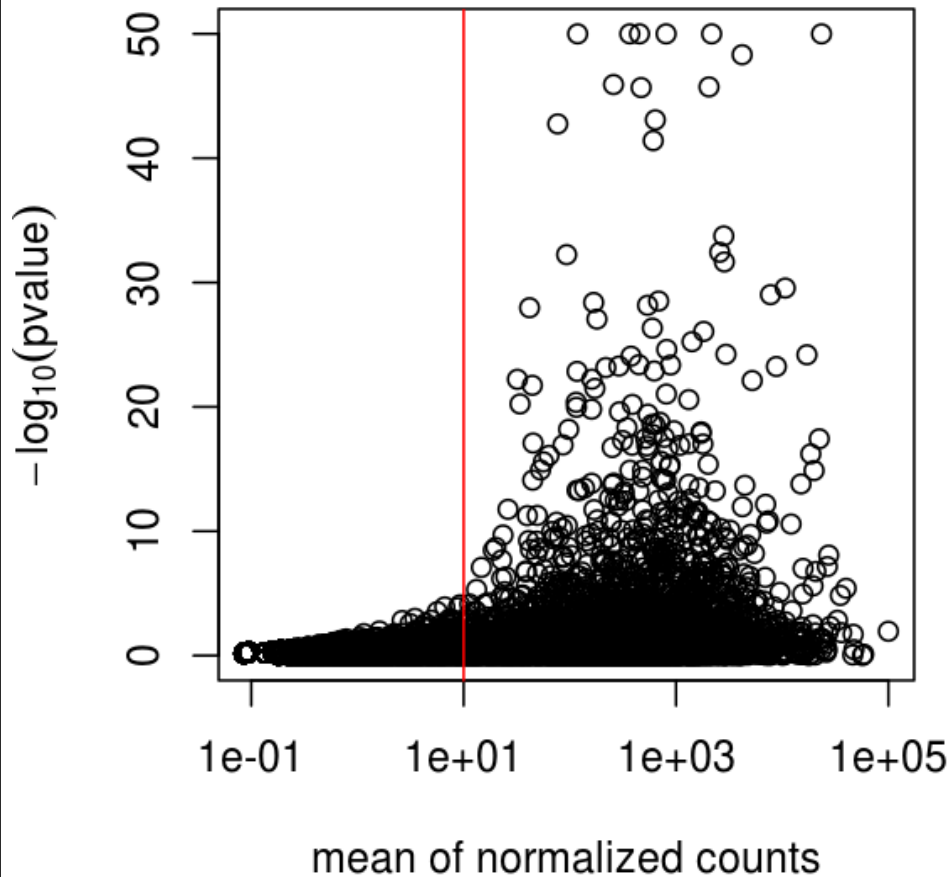
DESeq2 lab - parathyroid dataset



DESeq2 lab - parathyroid dataset



DESeq2 lab - parathyroid dataset



Independent filtering

From the set of all rows in the table,
first filter out those that seem to report negligible signal,
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)

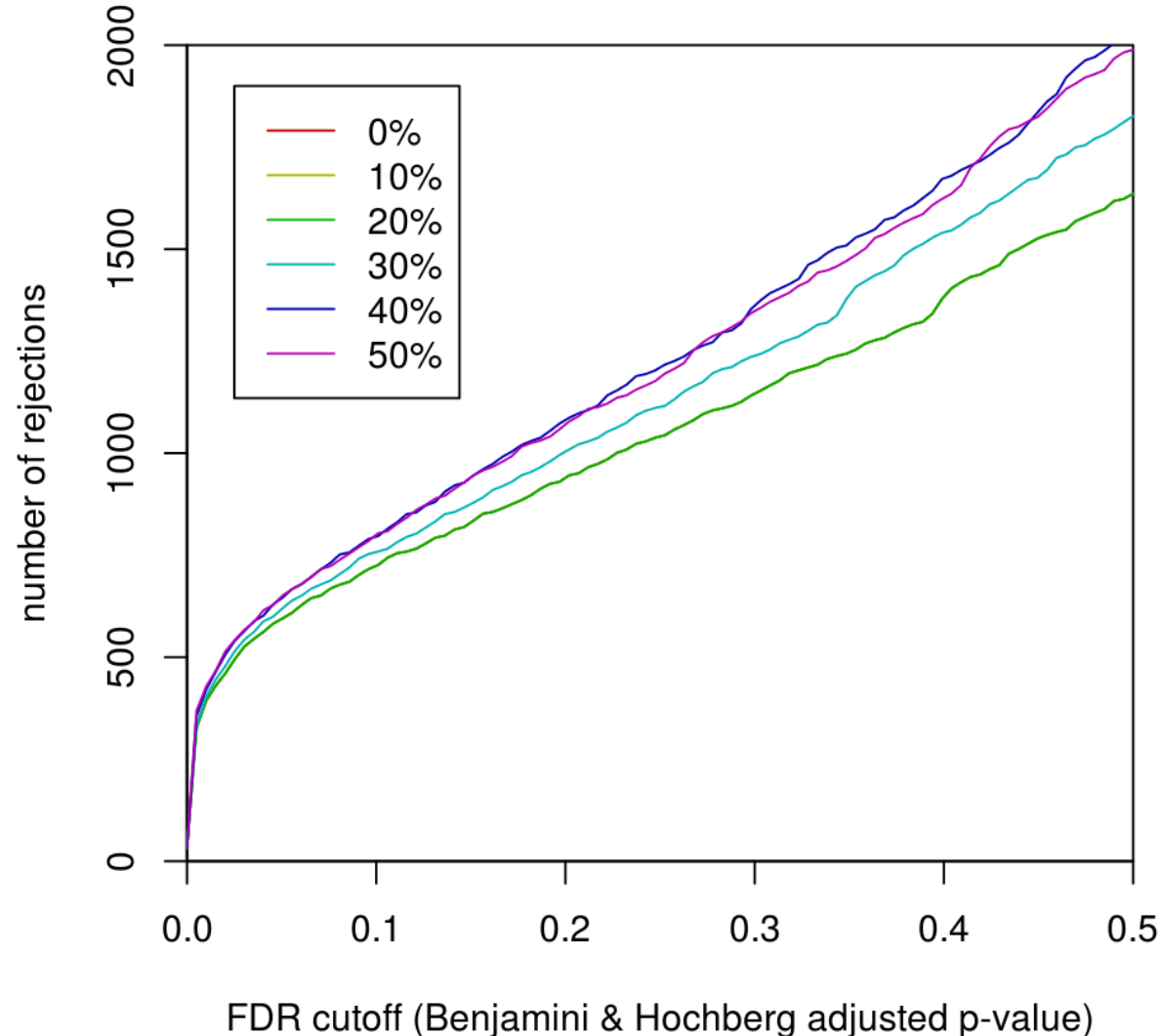
Bourgon et al. (PNAS 2010)

Many others.

Increased detection rates

Stage 1 filter: sum of counts, across samples, for each row, and remove the fraction θ that are smallest

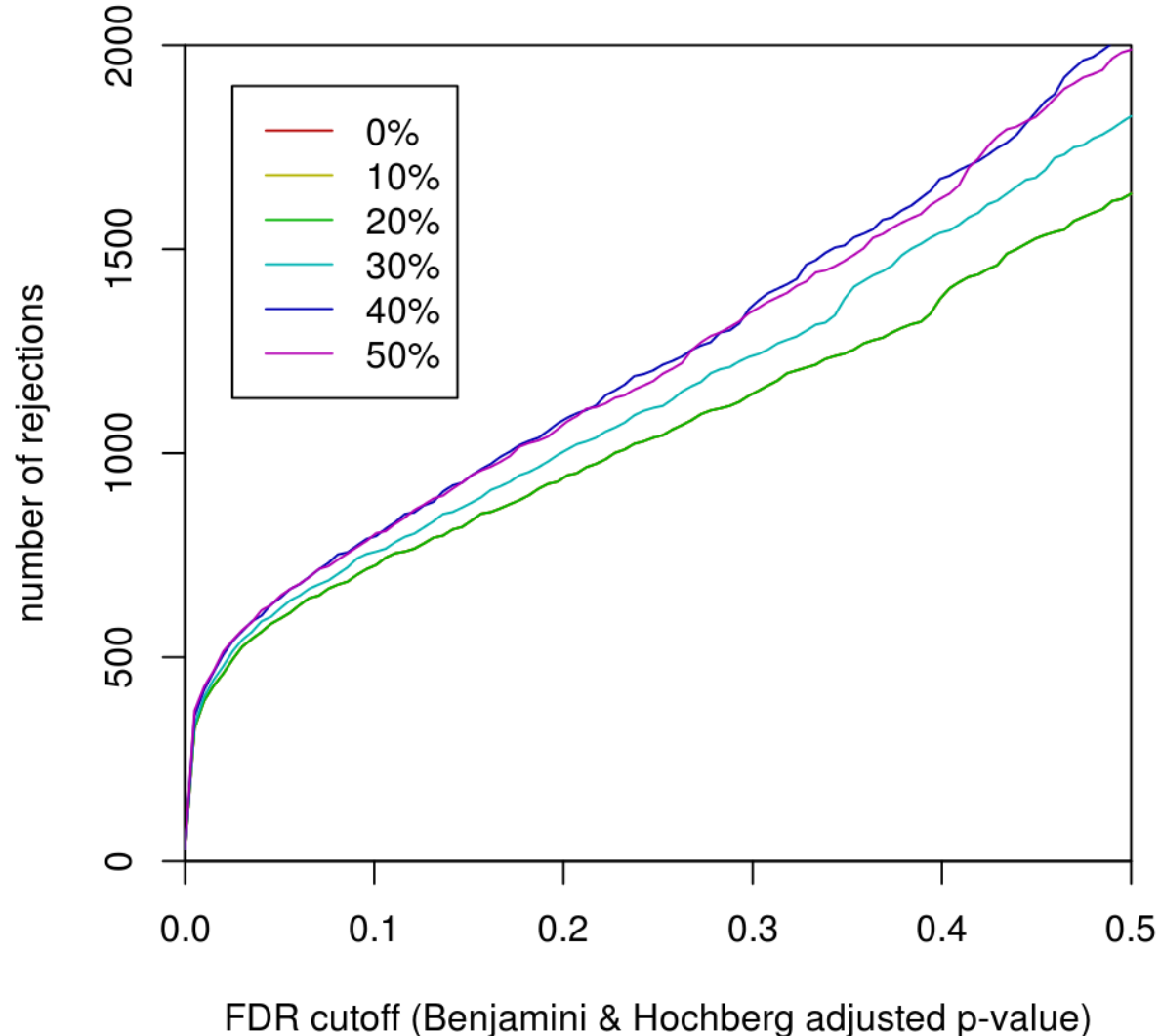
Stage 2: standard NB-GLM test



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

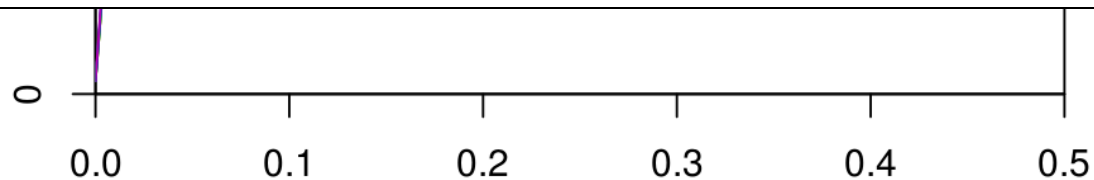
only if we are still controlling type I errors at the same level as before.

Concern:

- Since we use a data-driven criterion in stage 1, but do type I error consideration only on number of 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 type I error control?

I. For each individual (per gene) test statistic, we need to know its correct null distribution

II. To the extent that the multiple testing procedure relies on a certain (in)dependence structure between the different test statistics, our test statistics need to comply.

I.: one (though not the only) solution is to make sure that by filtering, the null distribution is not affected - that it is the same before and after filtering

II.: See later

Result: 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 (DESeq(2)):**

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.

Derivation

Non-parametric case:

Straightforward decomposition of the joint probability into product of probabilities using the assumptions.

Normal case:

Use the spherical symmetry of the joint distribution, p -dimensional $N(0, 1\sigma^2)$, and of the overall variance; and the scale and location invariance of t .

This case is also implied by Basu's theorem

(V complete sufficient for family of probability measures P , T ancillary $\Rightarrow T, V$ independent)

What do we need for type I error control?

The distribution of the test statistic under the null.

- I. **Marginal**: for each individual (per gene) test statistic
- II. **Joint**: some multiple testing procedures relies on certain independence properties of the joint distribution

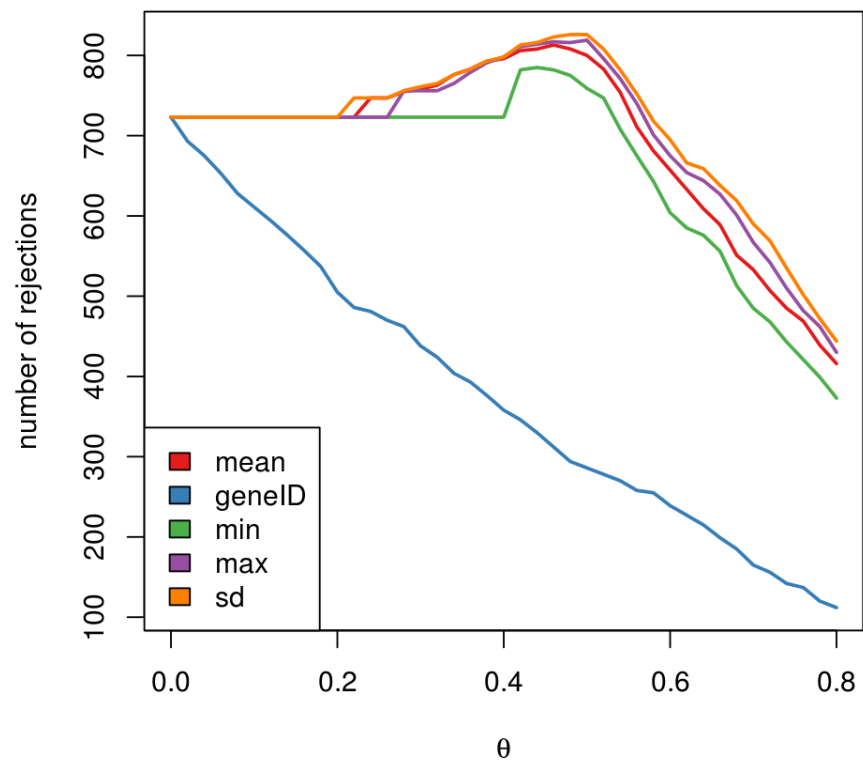
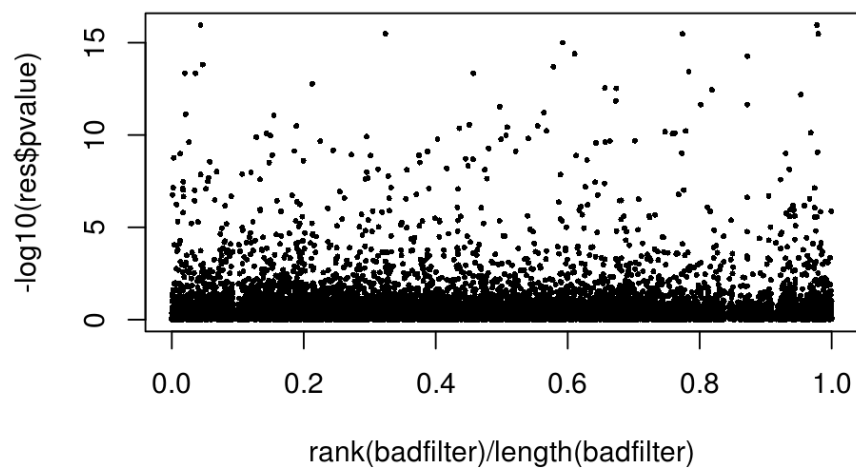
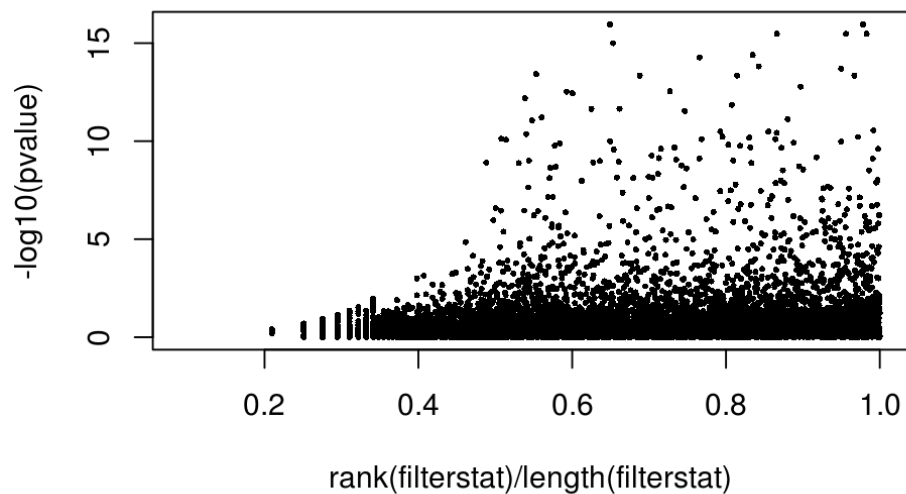
I.: one solution is to make sure that by filtering, the marginal null distribution is not affected - that it is the same before and after filtering (possible alternative: empirical nulls)



Multiple testing procedures and dependence

1. **Methods that work on the p-values only and allow general dependence structure: Bonferroni, Bonferroni-Holm (FWER), Benjamini-Yekutieli (FDR)**
2. **Those that work on the data matrix itself, and use permutations to estimate null distributions of relevant quantities (using the empirical correlation structure): Westfall-Young (FWER)**
3. **Those that work on the p-values only, and make dependence-related assumptions: Benjamini-Hochberg (FDR), q-value (FDR)**

Diagnostics



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**

Thank you

A photograph of a very crowded city street, likely a pedestrian walkway, with many people walking. The image has a greenish tint. At the bottom, a DNA sequence is overlaid in white text, with some letters enclosed in white boxes.

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

Derivation (non-parametric case)

$$P(f \in A, g \in B)$$

A, B: measurable sets
f: stage 1, g: stage 2

$$= \int_{i^n} \delta_A(f(X)) \delta_B(g(X)) dP_X$$

exchangeability

$$= \frac{1}{n!} \sum_{\pi \in \Pi_n} \int_{i^n} \delta_A(f \circ \pi(X)) \delta_B(g \circ (X)) dP_X$$

f's permutation invariance

$$= \int_{i^n} \delta_A(f(X)) \left(\frac{1}{n!} \sum_{\pi \in \Pi_n} \delta_B(g \circ (X)) \right) dP_X$$

distribution of g generated
by permutations

$$= \int_{i^n} \delta_A(f(X)) P(g \in B) dP_X$$

$$= P(f \in A) \cdot P(g \in B)$$

#

Positive Regression Dependency

On the subset of true null hypotheses:

If the test statistics are $X = (X_1, X_2, \dots, X_m)$:

For any increasing set D (the product of rays, each infinite on the right), and H_{0i} true, require that

$\text{Prob}(X \text{ in } D \mid X_i = s)$ is increasing in s , for all i .

Important Examples

Multivariate Normal with positive correlation

Absolute Studentized independent normal