Hypothesis Testing

Wolfgang Huber, EMBL

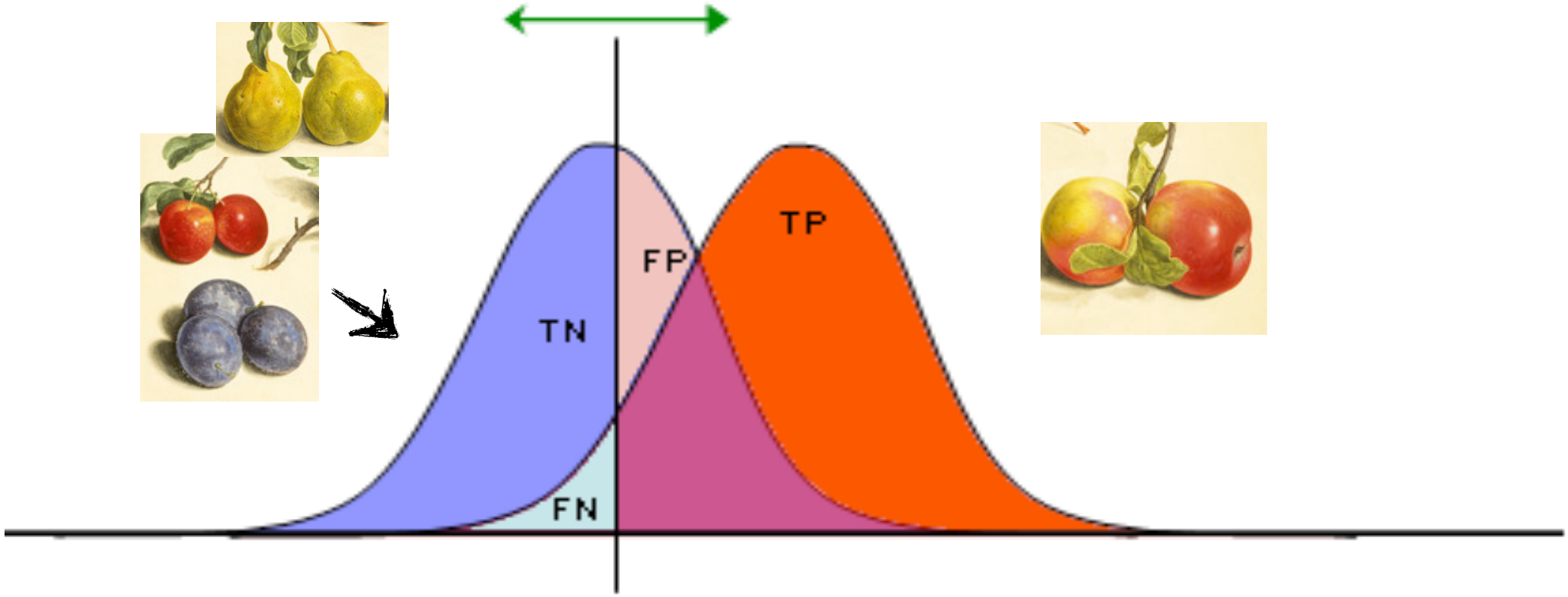
Aims for this lecture

Understand the basic principles of hypothesis testing, and its pitfalls

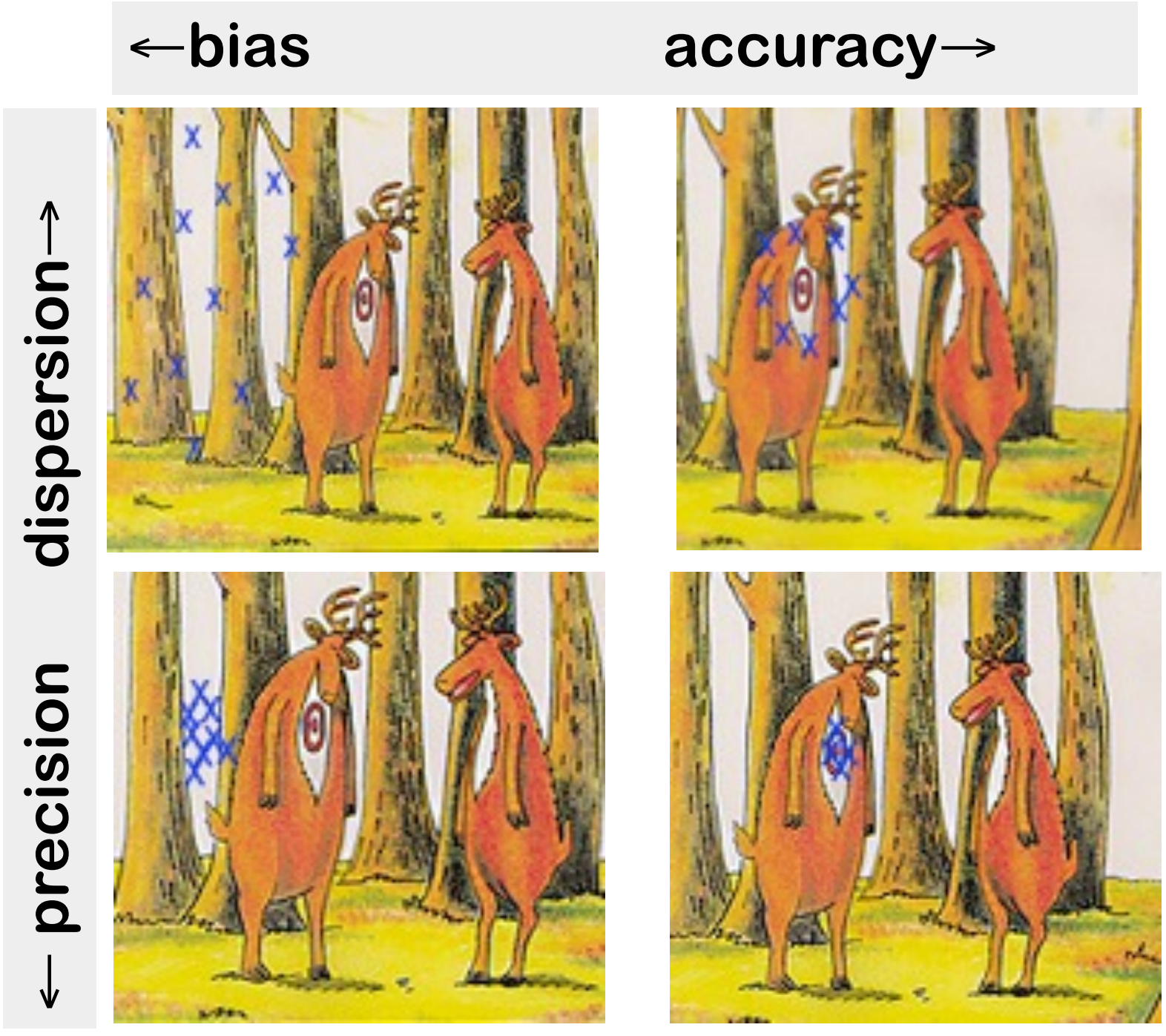
What changes when we go from single to multiple testing?

Understand the benefits and pre-conditions of independent filtering

Testing vs classification



Accuracy vs Precision - Bias vs Variance



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, H0

Step 2: Do an experiment, collect data

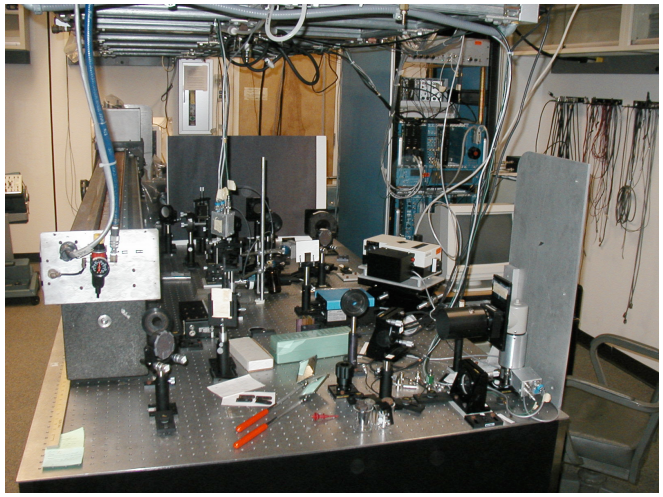
Step 3: Compute the probability of the data in this model

Step 4: Make a decision: reject the model if the computed probability is deemed too 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 effect of that RNAi-treatment on my cells is no different than that of a negative control treatment



The four steps of hypothesis testing

Step 1: Set up a model of reality: null hypothesis, H0

Step

Step

Step **d**

prob

H_0: a **of**

how **e**

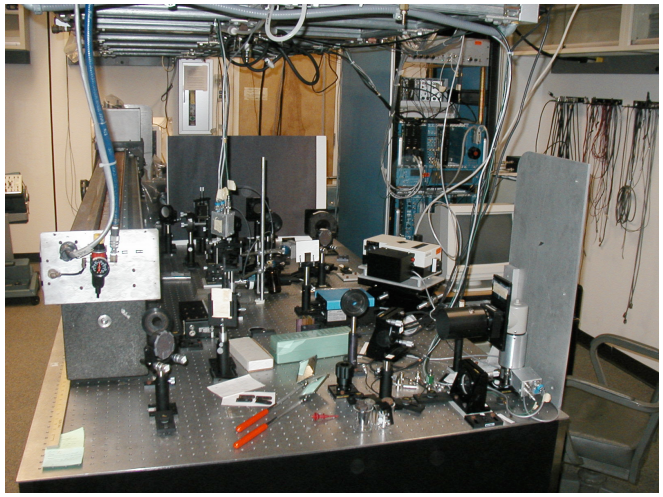
mathematical theory of probability.

These are **not** null hypotheses:

- The number of heads and tails is the same
- The outcomes in my patient cohort are exactly the same
- The measured Cell-Titro signal from the cells is the same between RNAi-treatment and negative control

Examples of null hypotheses:

- The coin is fair
- The new drug is no better or worse than a placebo
- The effect of that RNAi-treatment on my cells is no different than that of a negative control treatment

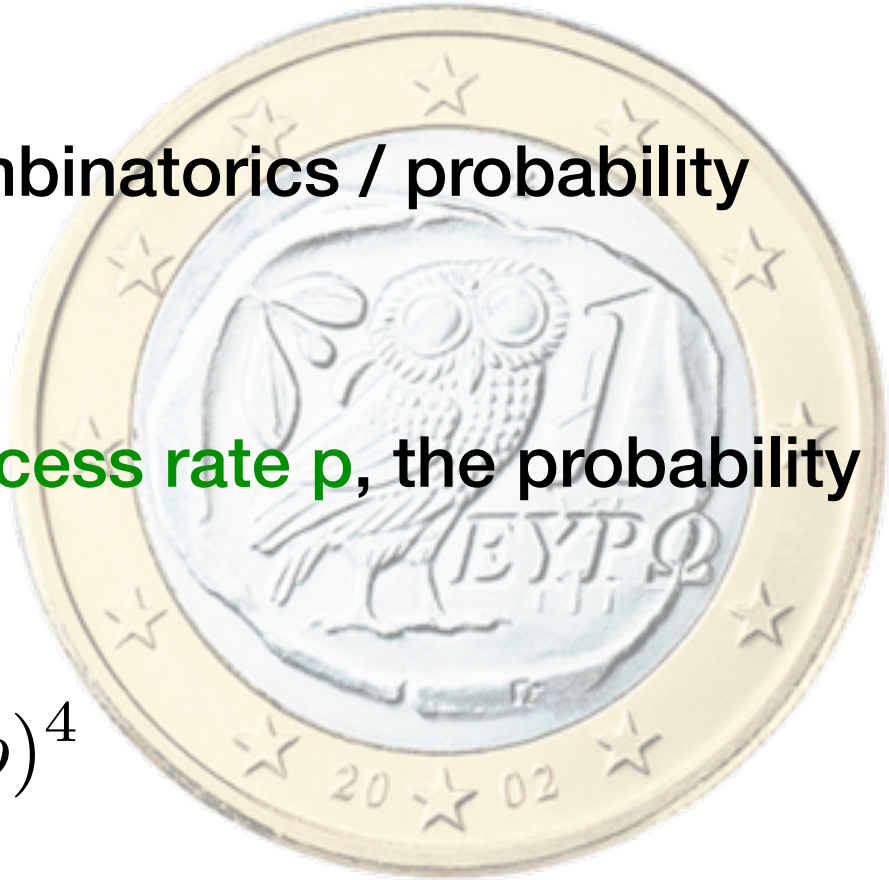


Example

Toss a coin a number of times ⇒

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.

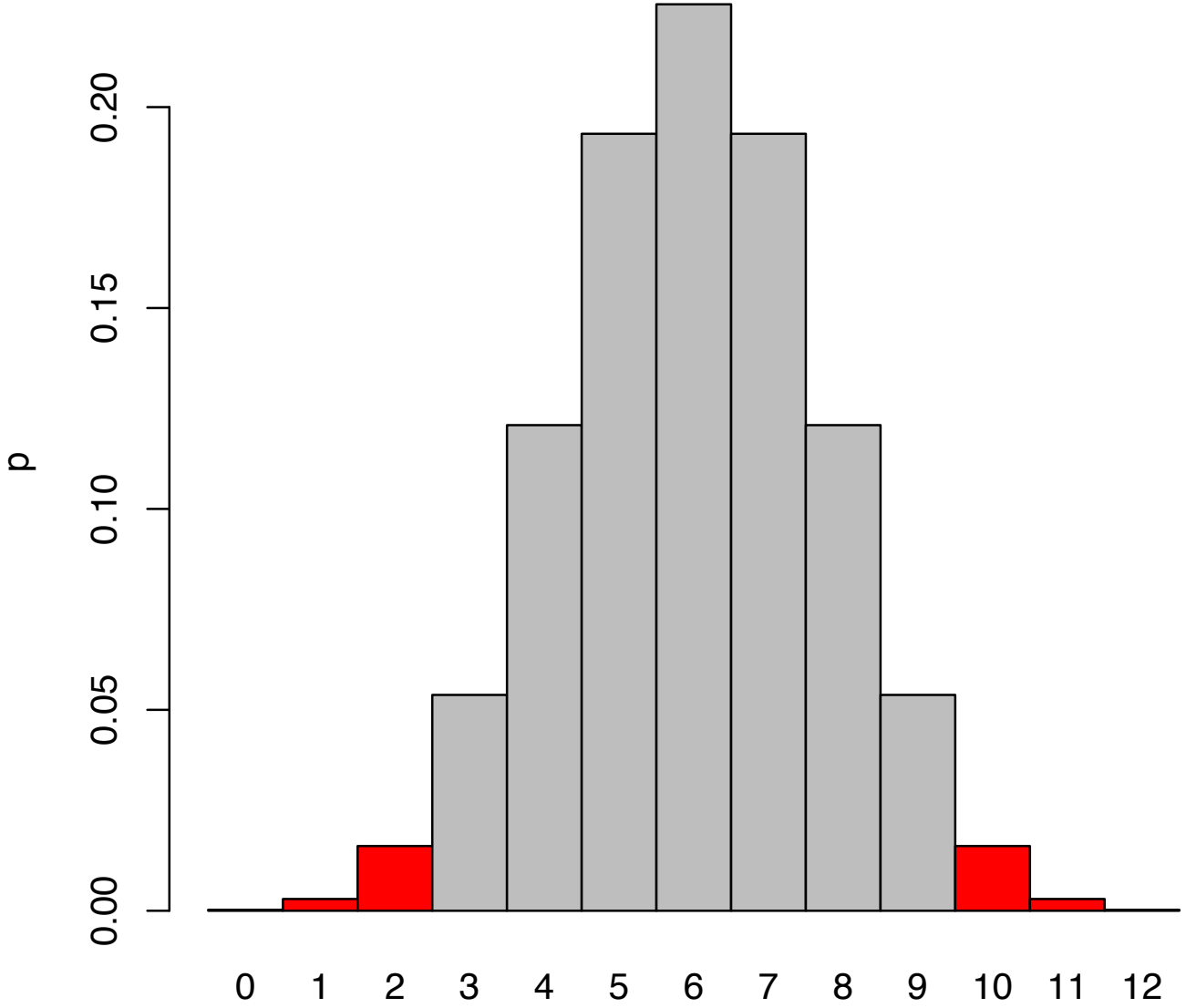


For example, in 12 tosses with success rate p, the probability of seeing exactly 8 heads is

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

Binomial Distribution

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



P(Heads ≤ 2) = 0.0193

P(Heads ≥ 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(heads ≥ 10 | H_0) = “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<α, 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. Make a decision: reject H_0 - yes / no ?

Errors in hypothesis testing

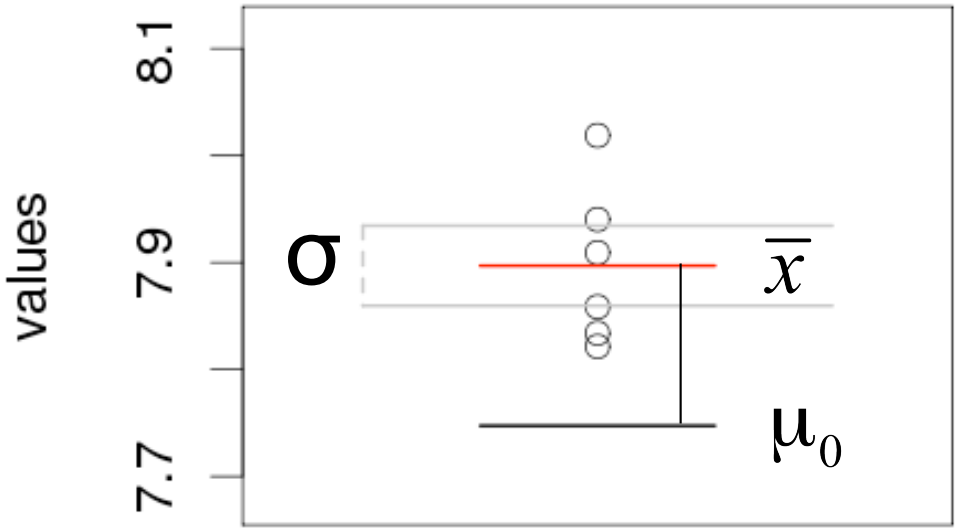
Truth \ Decision	not rejected ('negative')	rejected ('positive')
H_0 true	True negative (specificity)	False Positive Type I error α
H_0 false	False Negative Type II error β	True Positive (sensitivity)

One sample *t*-test

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

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

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

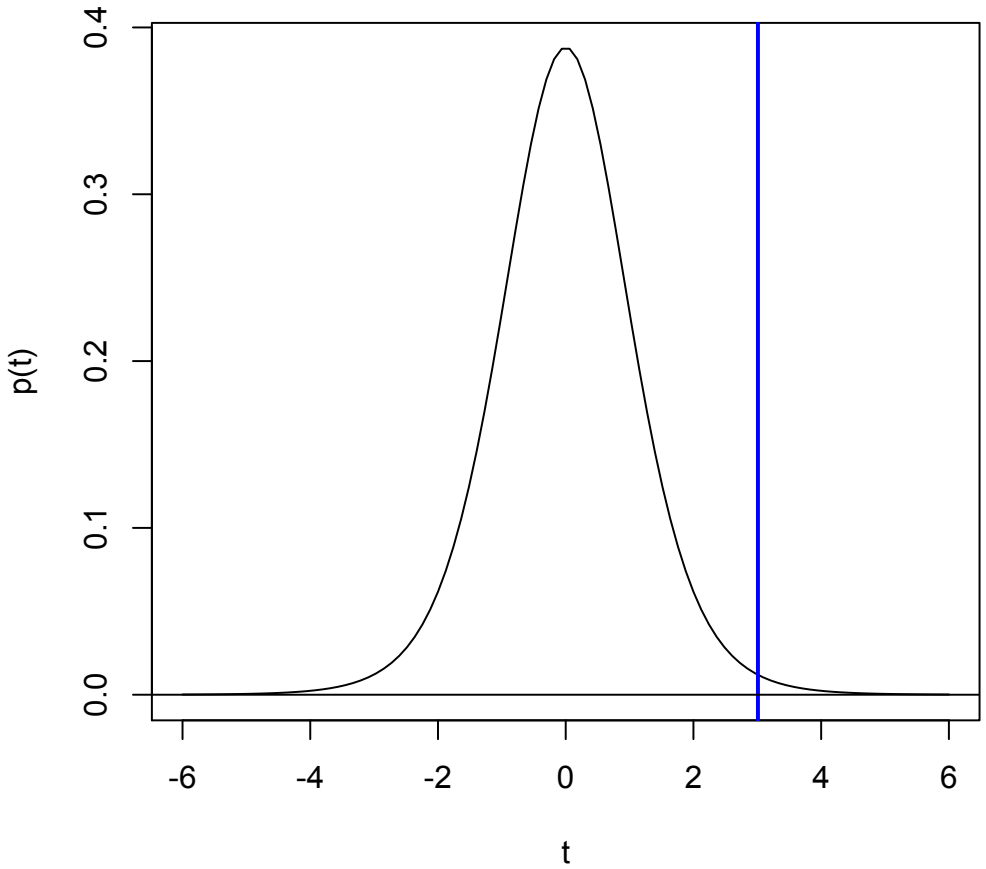
The 10 data points have a mean of 0.60 and a standard deviation of 0.62.

From that, we calculate the *t*-statistic:

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

p-value and test decision

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



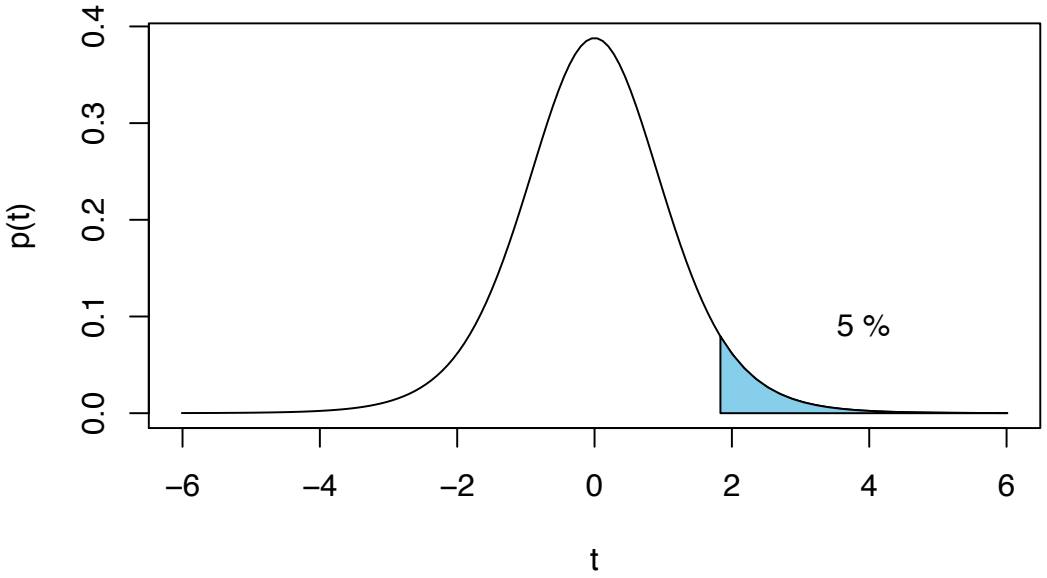
p-value: $P(|T_9| \geqq 3.0) = 0.015$

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

One-sided vs two-sided test

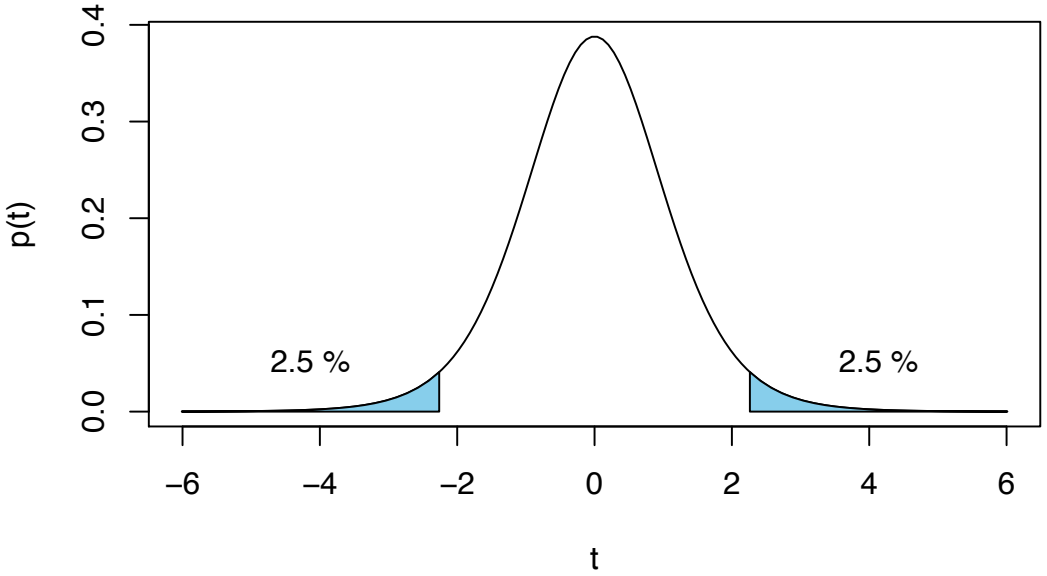
One-sided

e.g. H_0: $\mu<0$



Two-sided

e.g. H_0: $\mu=0$



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 = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, …)
```

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?

Avoid fallacy

The p-value is the probability that the 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 ≠ evidence of absence

Comments and pitfalls

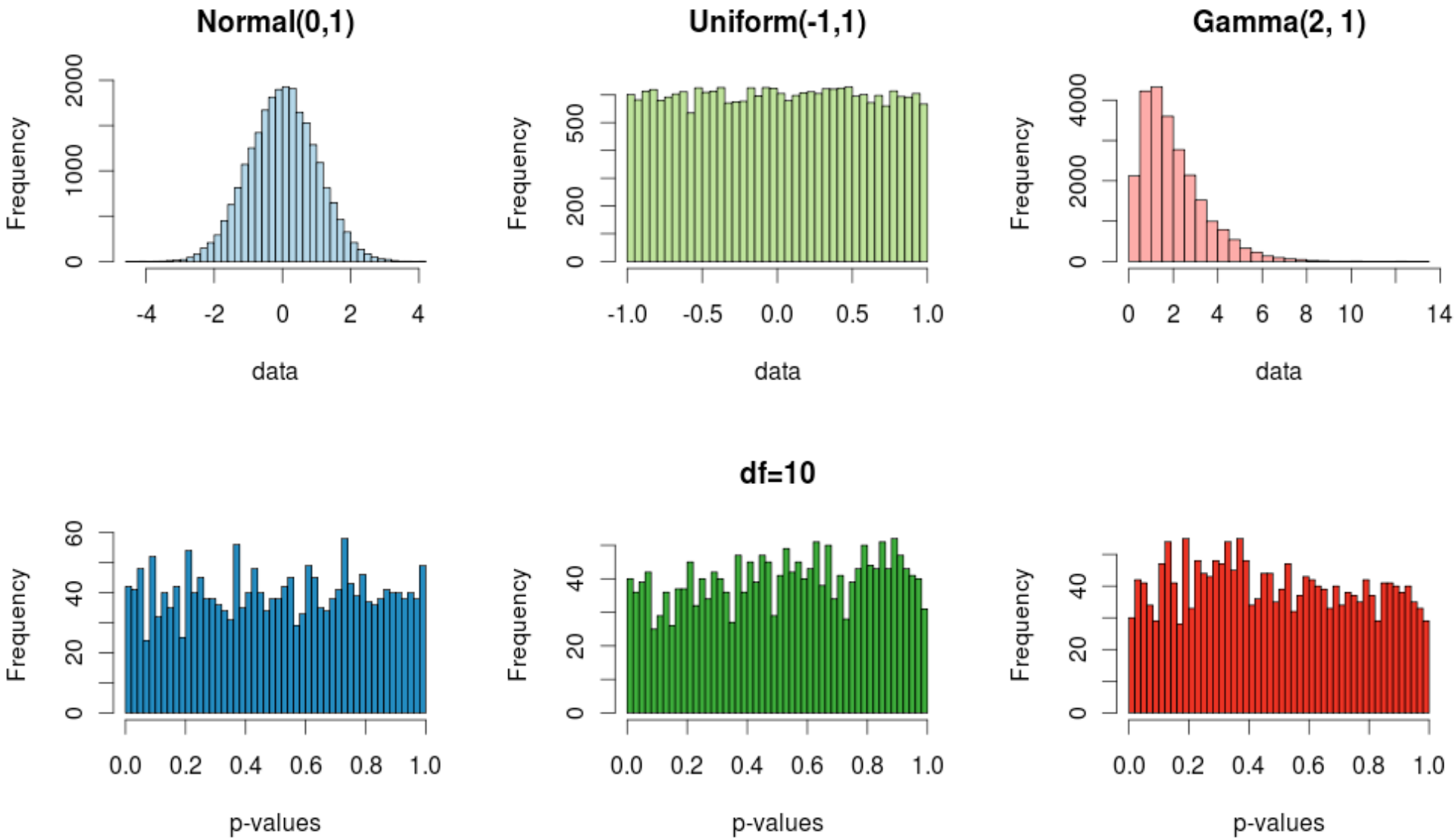
The derivation of the *t*-distribution assumes that the observations are independent and that they follow a Normal distribution.

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

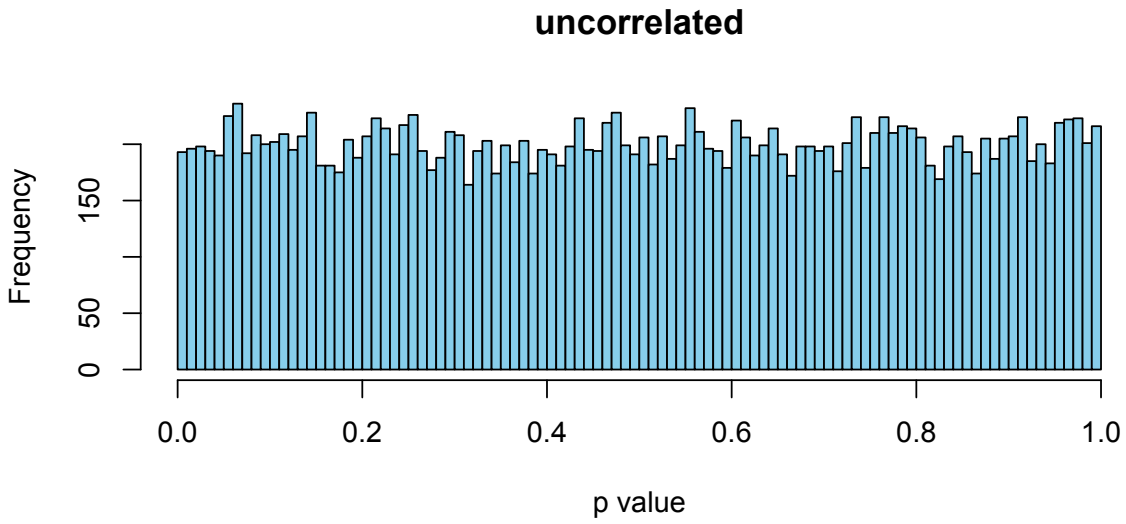
Options: Wilcoxon test, permutation tests

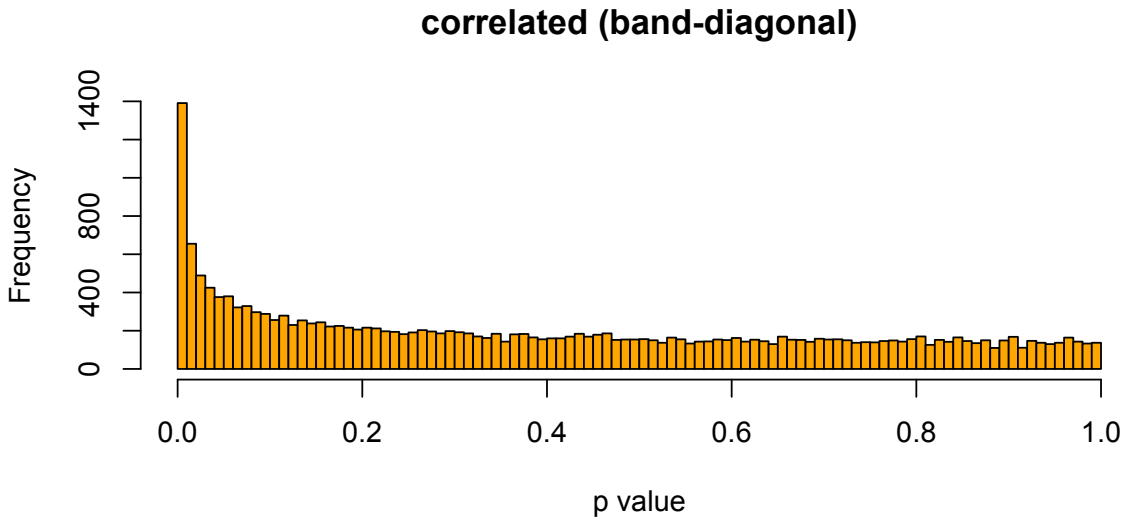
If the data are dependent, then p-values will likely be totally wrong (e.g., for positive correlation, too optimistic).

different data distributions – independent case



t-test can loose error control if independence assumption does not hold






```r
library("mvtnorm")
library("genefilter")

p = 30     ## number of samples
n = 20000  ## number of genes

mu = rep(0, p)
dp = diag(p)

sigma = list(
 'uncorrelated' = dp,              ## unity matrix
 'correlated (band-diagonal)' =    ## band diagonal
      (row(dp)==col(dp)) + 0.5 * (abs(row(dp)-col(dp))==1))

lapply(sigma, print)

## generate data
x  = lapply(sigma, function(s) rmvnorm(n = n, mean = mu, sigma = s))

## tests
tt = lapply(x, rowttests)

par(mfrow=c(length(tt), 1))
for(i in seq(along=tt))
    hist(tt[[i]]$p.value, breaks=100, col=c("skyblue", "orange")[i],
         main=names(tt)[i], xlab="p value")
```

Summary single hypothesis testing

We ‘prove’ something by rejecting the opposite (the null hypothesis)

Not rejecting does not prove the null hypothesis

All this reasoning is probabilistic

p-values are intended to be used for rational decision making

In genomics, they’re often also used for data integration

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Routledge
Taylor & Francis Group

Editorial

David Trafimow and Michael Marks
New Mexico State University

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

With the banning of the NHSTP from BASP, what are the implications for authors? The following are anticipated questions and their corresponding answers.

Question 1. *Will manuscripts with p-values be desk rejected automatically?*

Answer to Question 1. No. If manuscripts pass the preliminary inspection, they will be sent out for review. But prior to publication, authors will have to remove all vestiges of the NHSTP (*p*-values, *t*-values, *F*-values, statements about "significant" differences or lack thereof, and so on).

Question 2. *What about other types of inferential statistics such as confidence intervals or Bayesian methods?*

Answer to Question 2. Confidence intervals suffer from an inverse inference problem that is not very different from that suffered by the NHSTP. In the NHSTP, the problem is in traversing the distance from the probability of the finding, given the null hypothesis, to the probability of the null hypothesis, given the finding. Regarding confidence intervals, the problem is that, for example, a 95% confidence interval does not indicate that the parameter of interest has a 95% probability of being within the interval. Rather, it means merely that if an infinite number of samples were taken and confidence intervals computed, 95% of the confidence intervals would capture the population parameter. Analogous to how the NHSTP fails to provide the probability of the null hypothesis, which is needed to provide a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that when in a state of ignorance, the researcher should assign an equal probability to each possibility. The problems are well documented (Chihara, 1994; Fisher, 1973; Glymour, 1980; Popper, 1983; Suppes, 1994; Trafimow, 2003, 2005, 2006). However, there have been Bayesian proposals that at least somewhat circumvent the Laplacian assumption, and there might even be cases where there are strong grounds for assuming that the numbers really are there (see Fisher, 1973, for an example). Consequently, with respect to Bayesian procedures, we reserve the right to make case-by-case judgments, and thus Bayesian procedures are neither required nor banned from BASP.

Question 3. *Are any inferential statistical procedures required?*

Answer to Question 3. No, because the state of the art remains uncertain. However, BASP will require strong descriptive statistics, including effect sizes. We also encourage the presentation of frequency or distributional data when this is feasible. Finally, we encourage the use of larger sample sizes than is typical in much psychology research, because as the sample size increases, descriptive statistics become increasingly stable and sampling error is less of a problem. However, we will stop short of requiring particular sample sizes, because it is possible to imagine circumstances where more typical sample sizes might be justifiable.

We conclude with one last thought. Some might view the NHSTP ban as indicating that it will be easier to publish in BASP, or that less rigorous manuscripts will be acceptable. This is not so. On the contrary, we believe

Correspondence should be sent to David Trafimow, Department of Psychology, MSC 3452, New Mexico State University, P.O. Box 30001, Las Cruces, NM 88003-8001. E-mail: dtrafimo@nmsu.edu

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015

PDF | Rights & Permissions

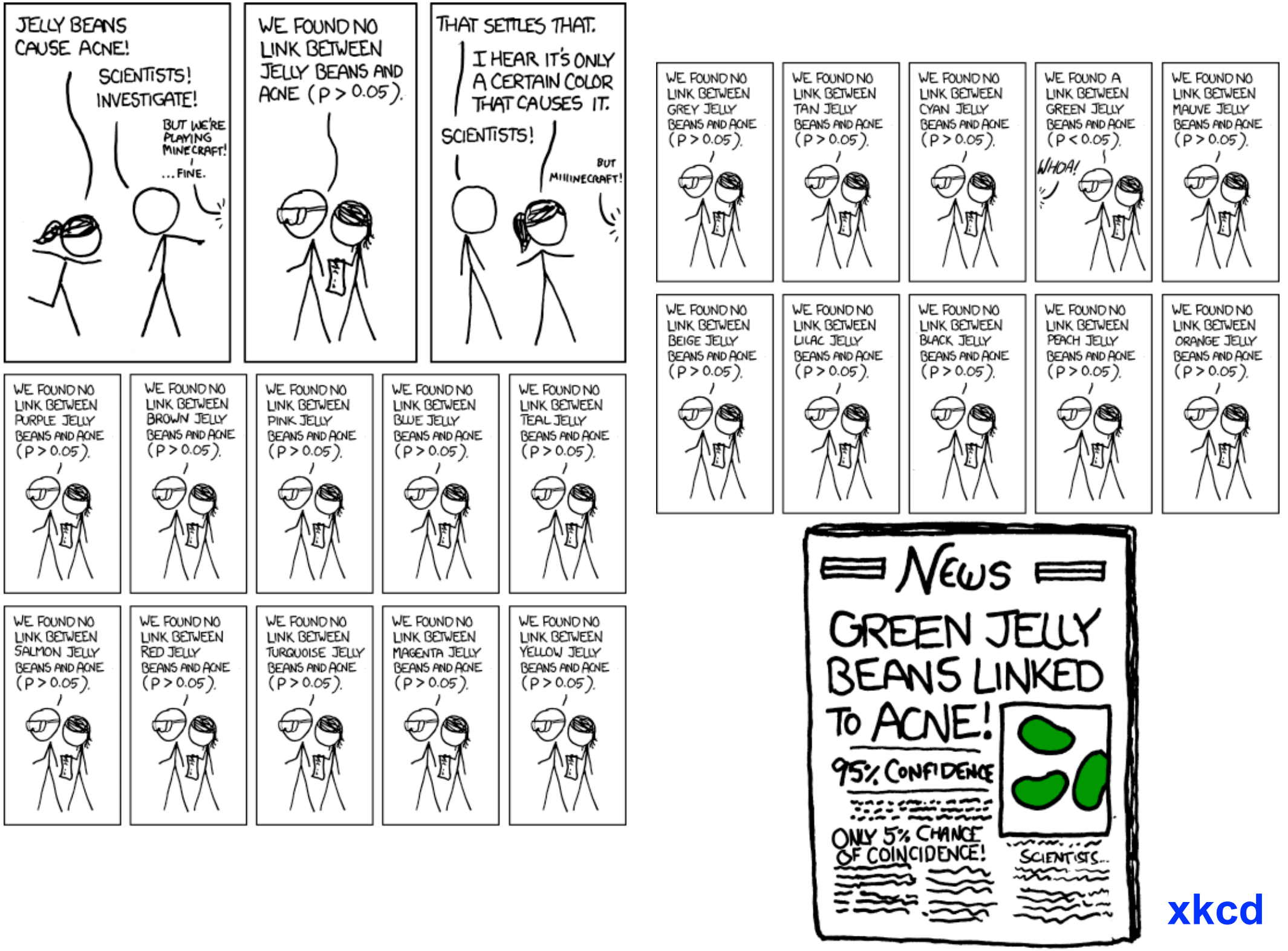
A controversial statistical test has finally met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values because the statistics were too often used to support lower-quality research[1].

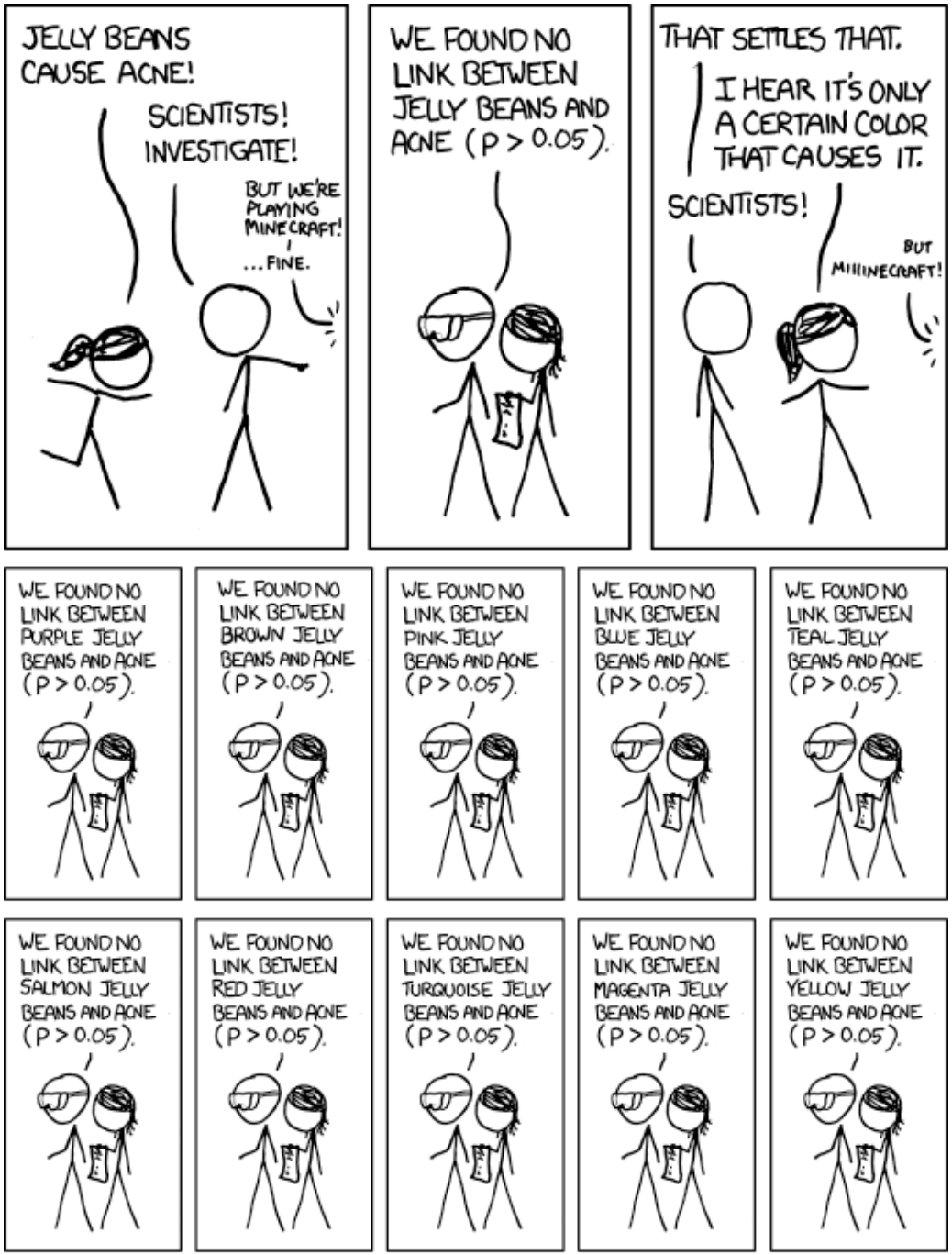
Authors are still free to submit papers to *BASP* with *P* values and other statistical measures that form part of 'null hypothesis significance testing' (NHST), but the numbers will be removed before publication. Nerisa Dozo, a PhD student in psychology at the University of Queensland in Brisbane, Australia, tweeted:

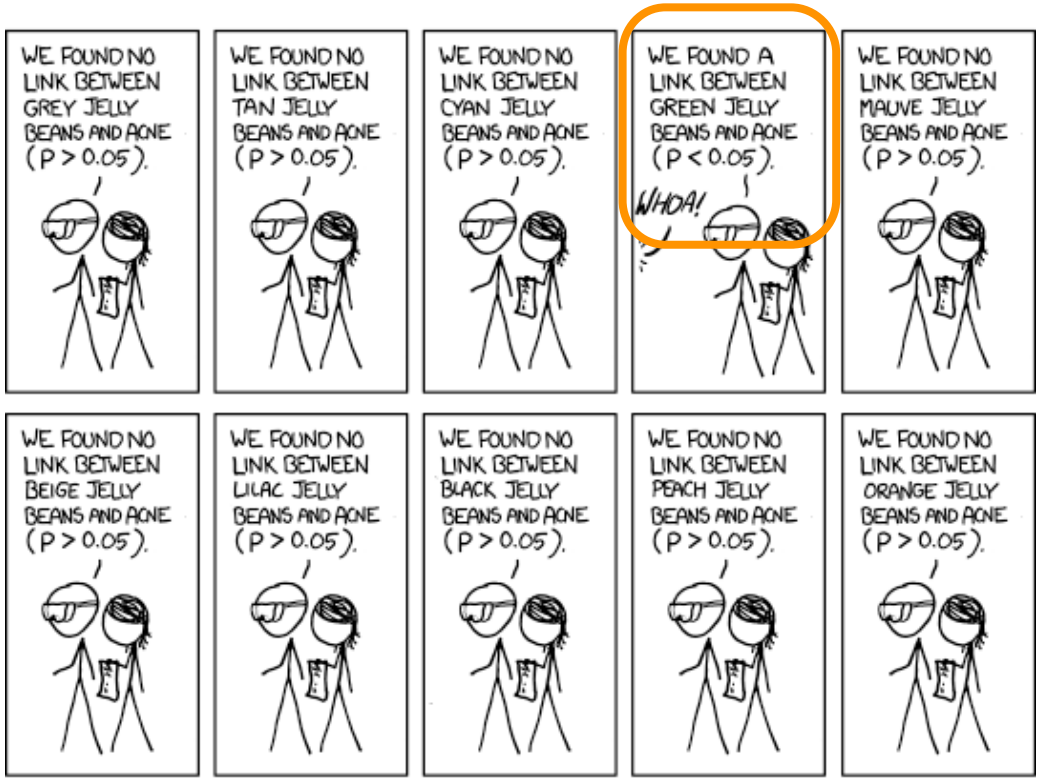
Good discussion thread:

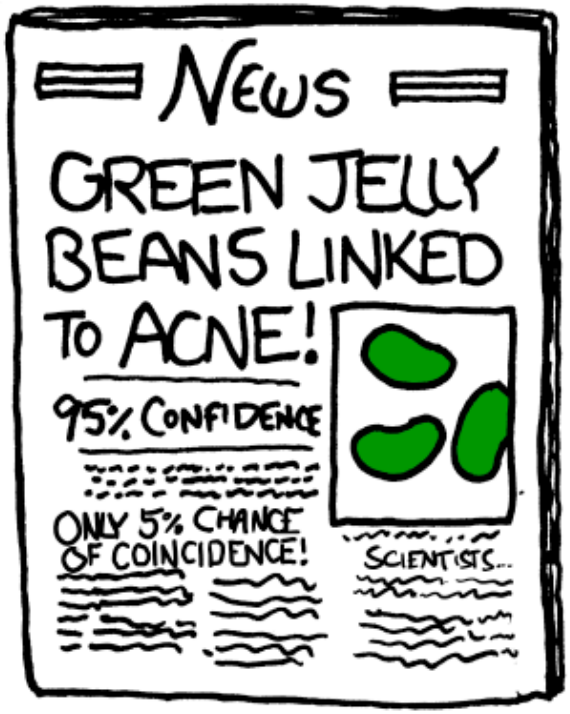
http://stats.stackexchange.com/questions/139290/a-psychology-journal-banned-p-values-and-confidence-intervals-is-it-indeed-wise

"A bad workman blames his tools."


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 MIIINECRAFT!
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 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 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 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).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...
xkcd


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 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 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...

xkcd

Multiple Testing Examples

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

- RNA-Seq (or μarray) expression profiles of “normal” vs “perturbed” samples: gene-by-gene
- ChIP-Seq: locus-by-locus
- RNAi and chemical compound screens
- Genome-wide association studies (GWAS): marker-by-marker
- QTL analysis: marker-by-marker and trait-by-trait

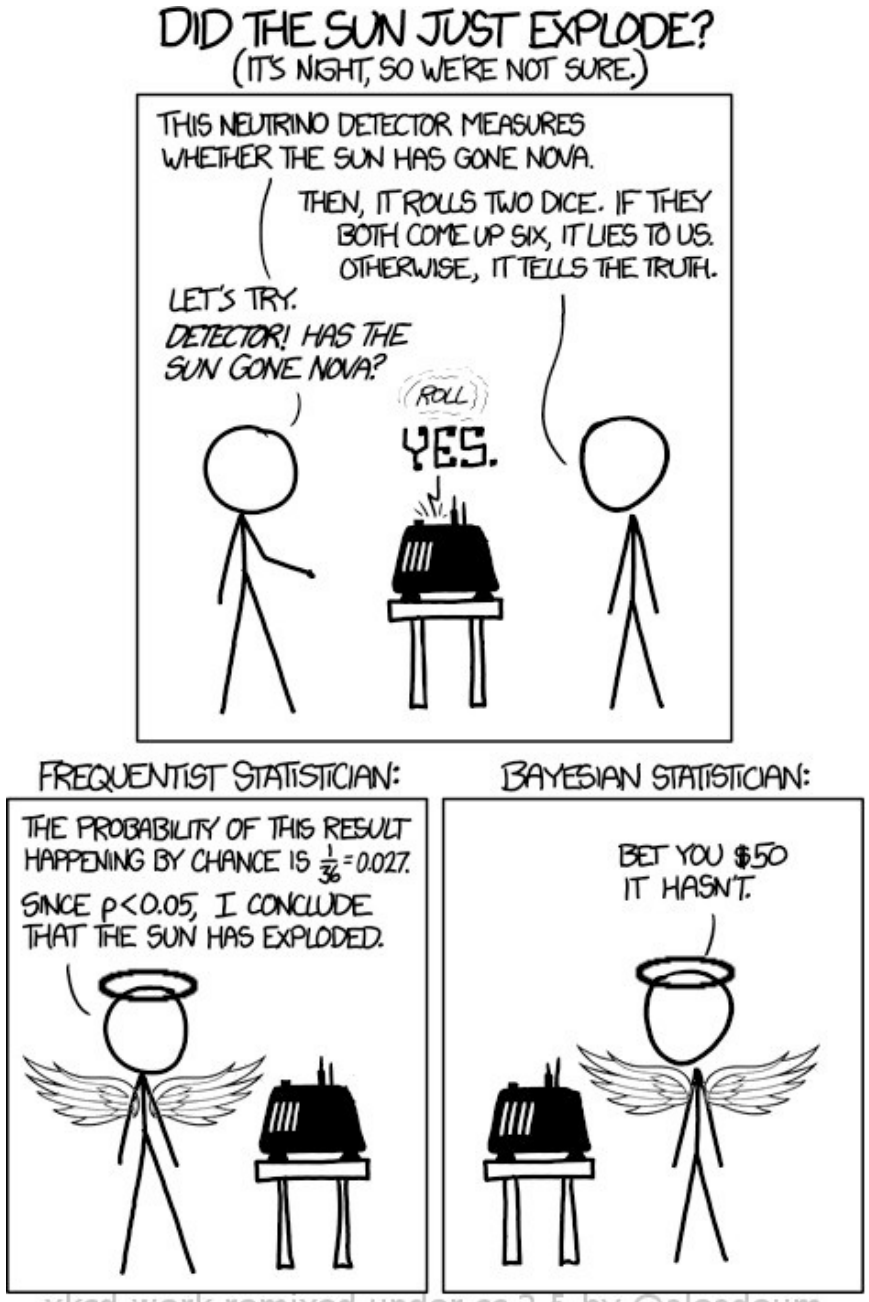
The Multiple Testing Problem

When performing many tests, the type I error goes up: for α = 0.05 and *n* tests, the probability of no false positive result is:

$$\underbrace{0.95 \cdot 0.95 \cdot \ldots \cdot 0.95}_{\text{n–times}} \quad \lll \quad 0.95$$

The larger the number of tests performed, the higher the probability of a false rejection (→1)

The Multiple Testing Opportunity



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_0
False null hypotheses	T	S	m_1
Total	m – R	R	m

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

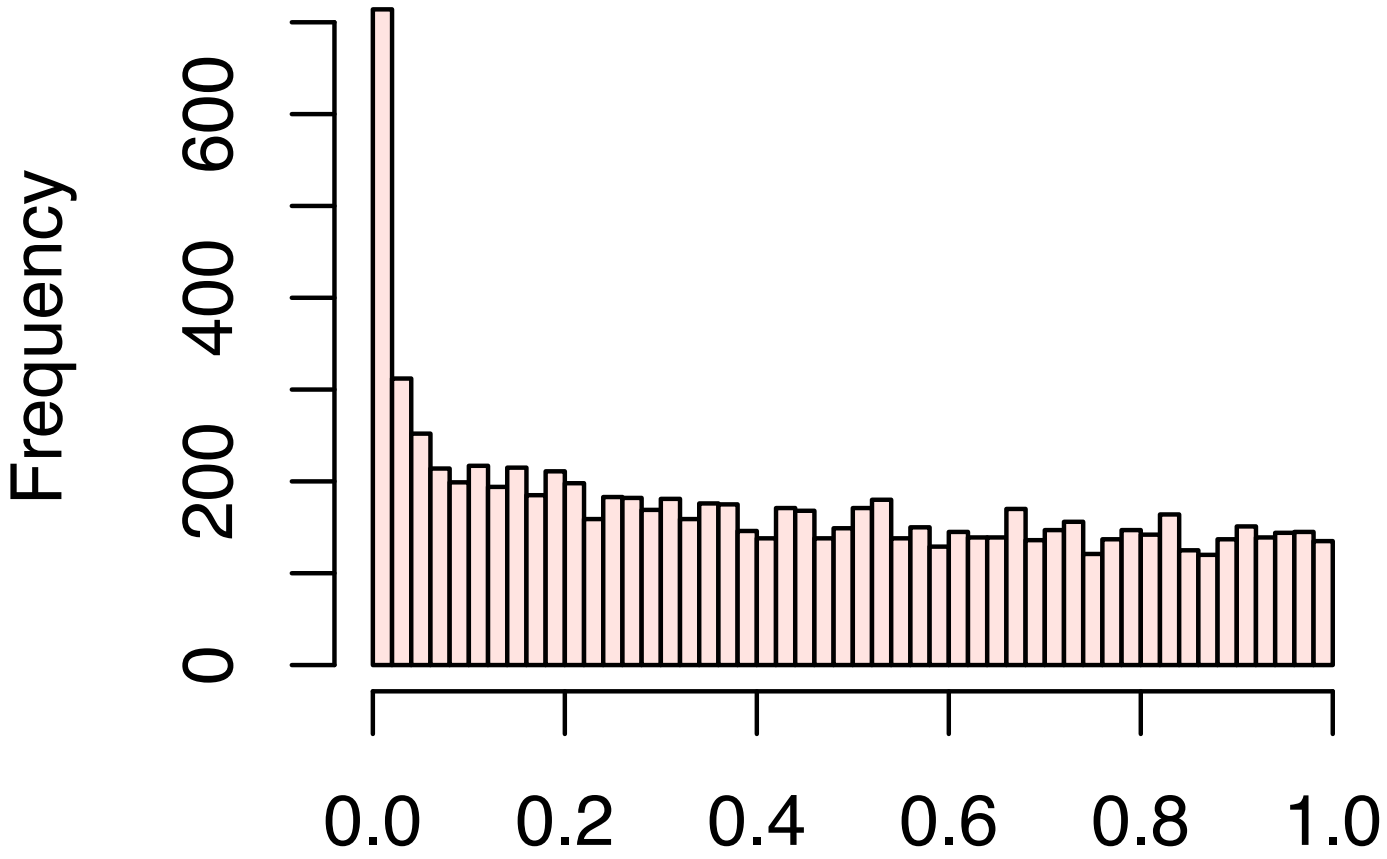
False discovery rate (FDR): The expected fraction of false positives among all discoveries, E[V / max {R, 1}].

Bonferroni correction

For m tests, multiply each p-value with m.

Then see if anyone still remains below α.

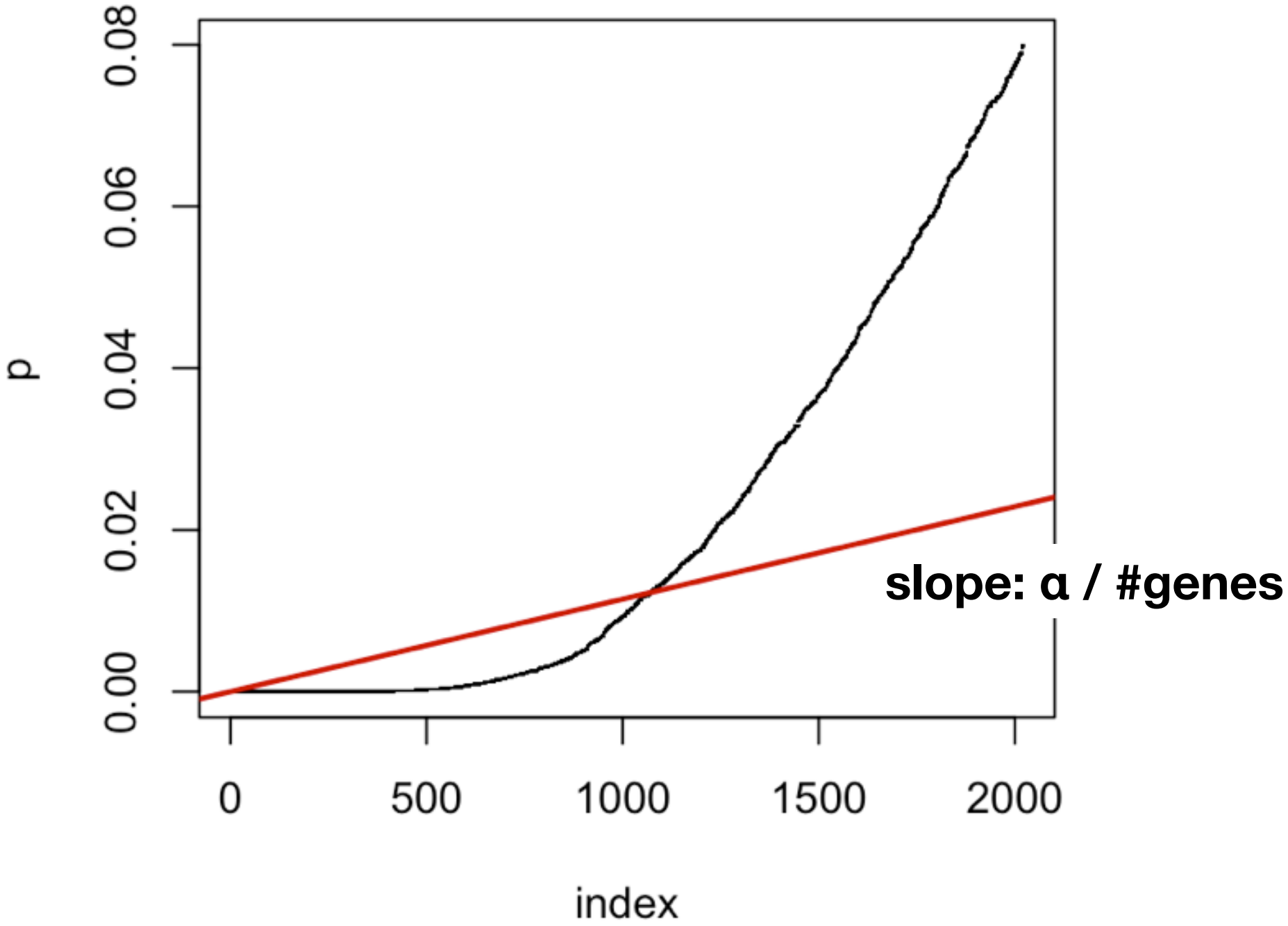
False discovery rate



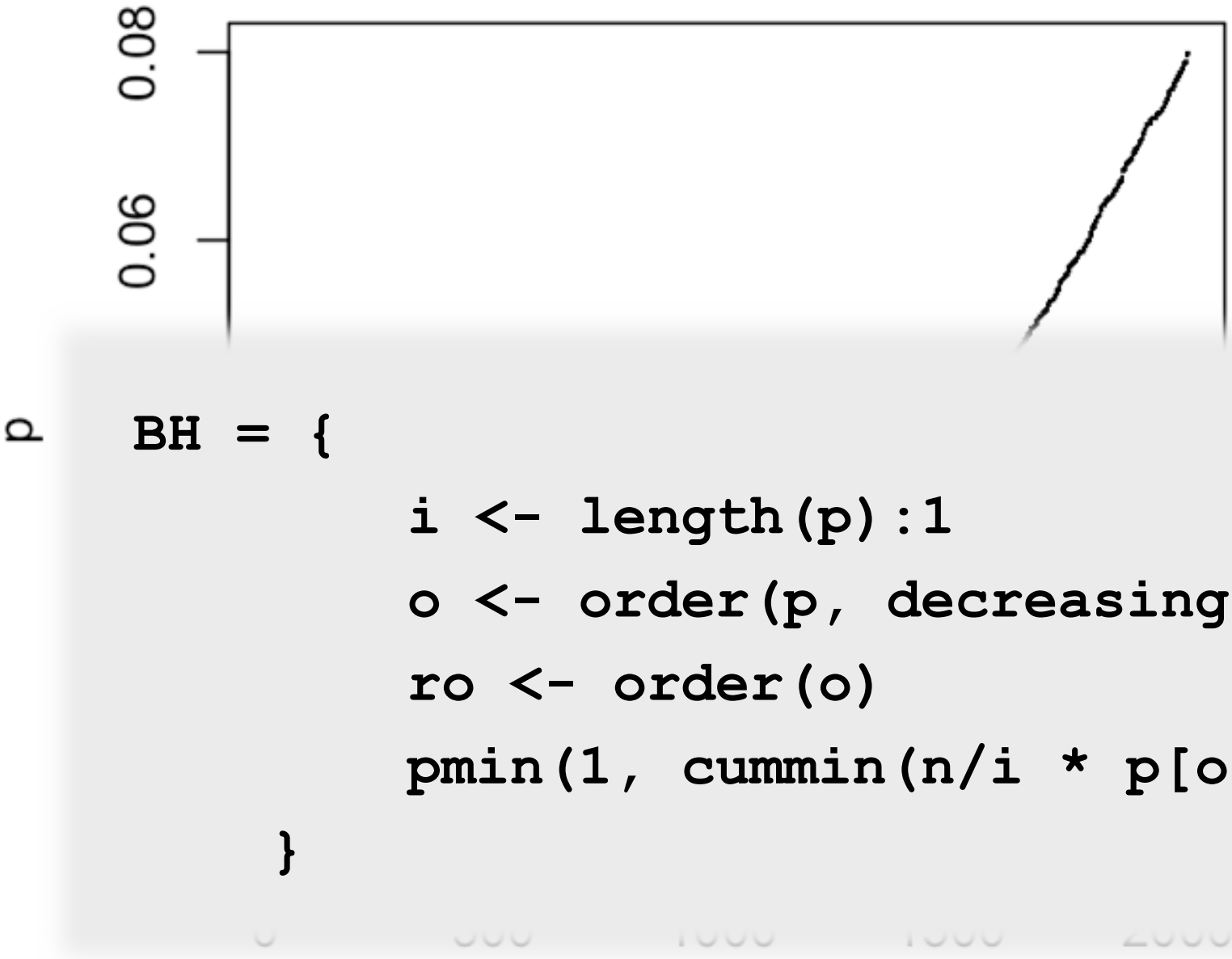
Observed p-values are a mix of samples from

- a uniform distribution (←nulls) and
- from one (more) distributions concentrated at 0 (←alternatives)

Benjamini Hochberg method

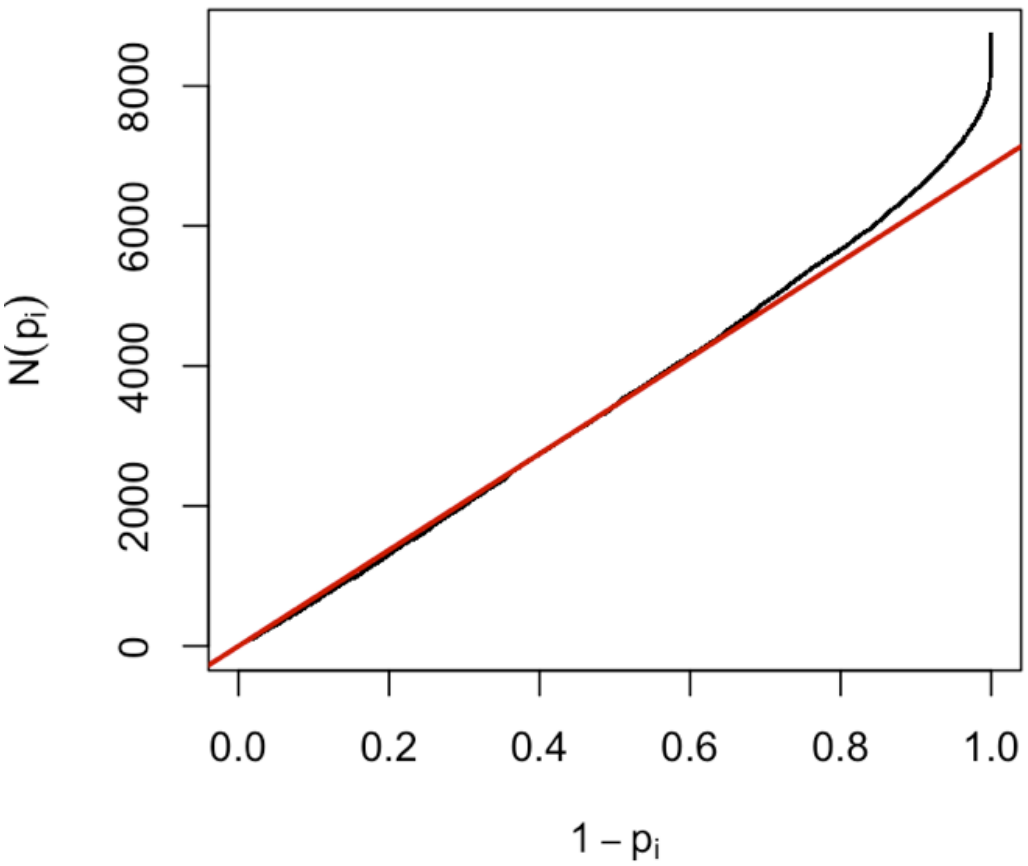


Benjamini Hochberg method




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

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



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

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

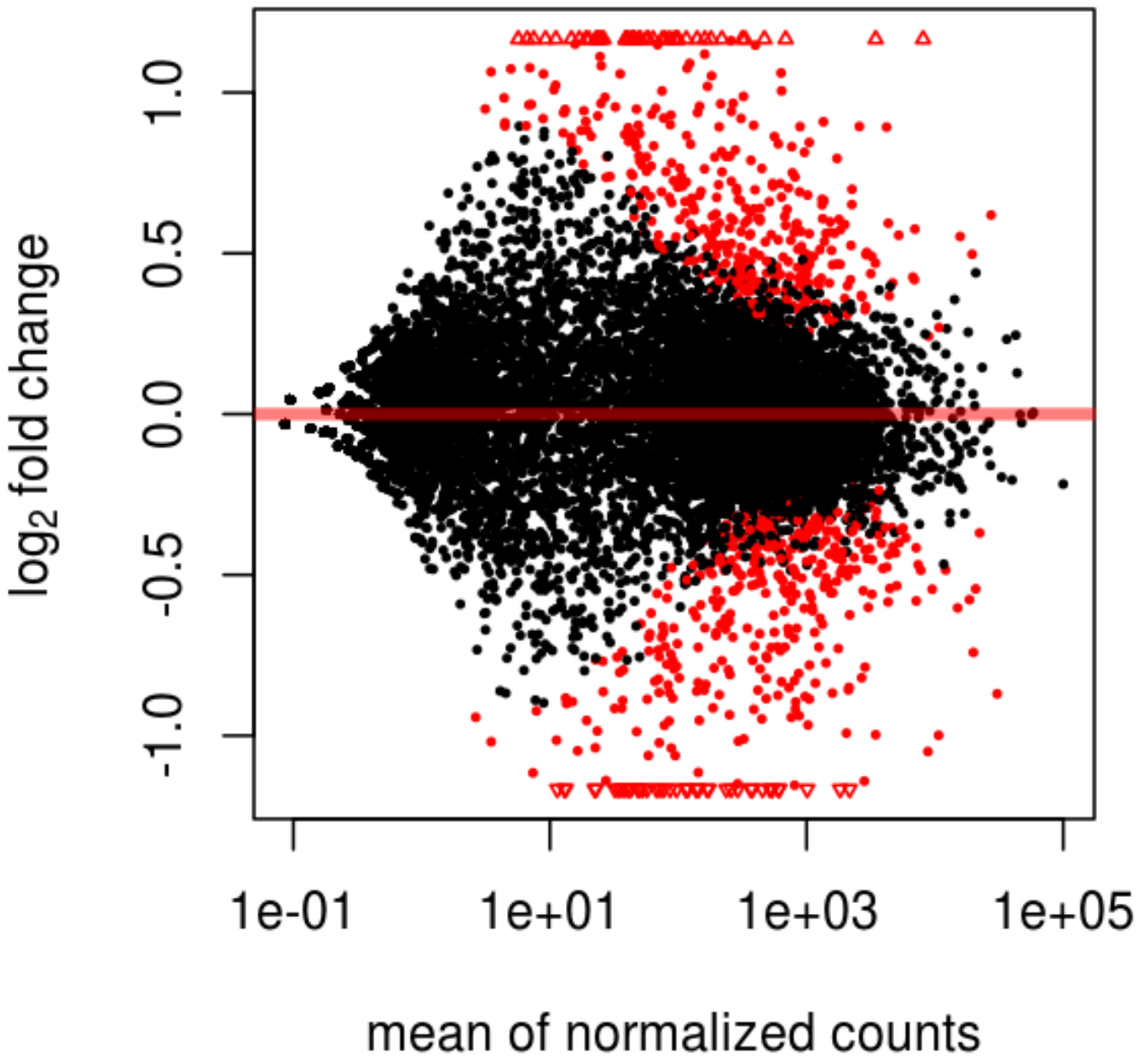
where $N(p)$ is the number of p-values greater than p.

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

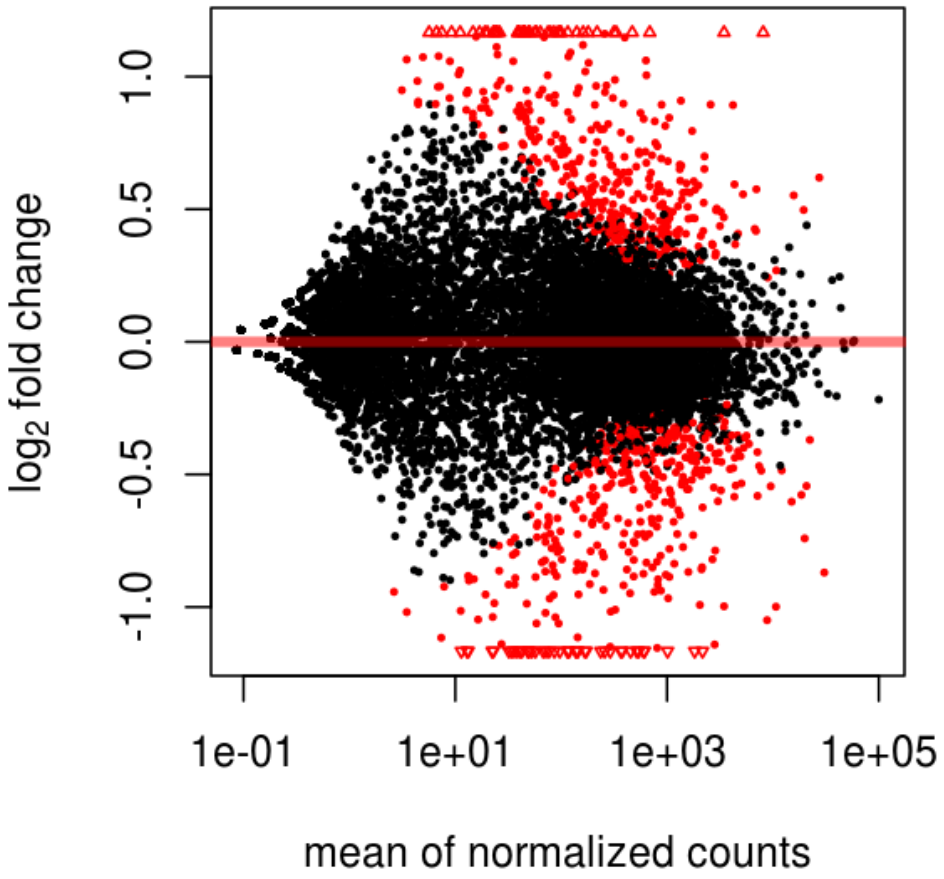
$(1 - p)\, m$ = expected number of p-values greater than p

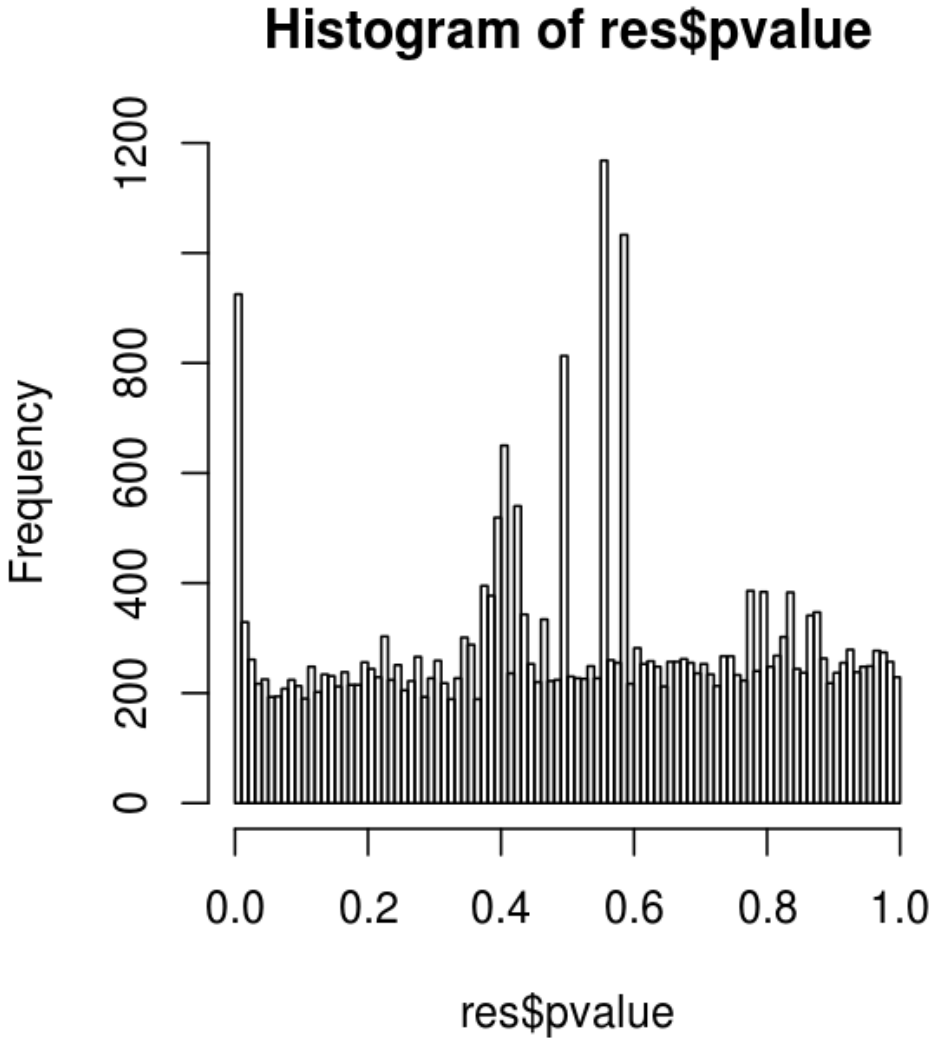
Schweder T, Spjøtvoll E (1982) Plots of P-values to evaluate many tests simultaneously. *Biometrika* 69:493–502. See 'genefilter' vignette for an example.

parathyroid dataset

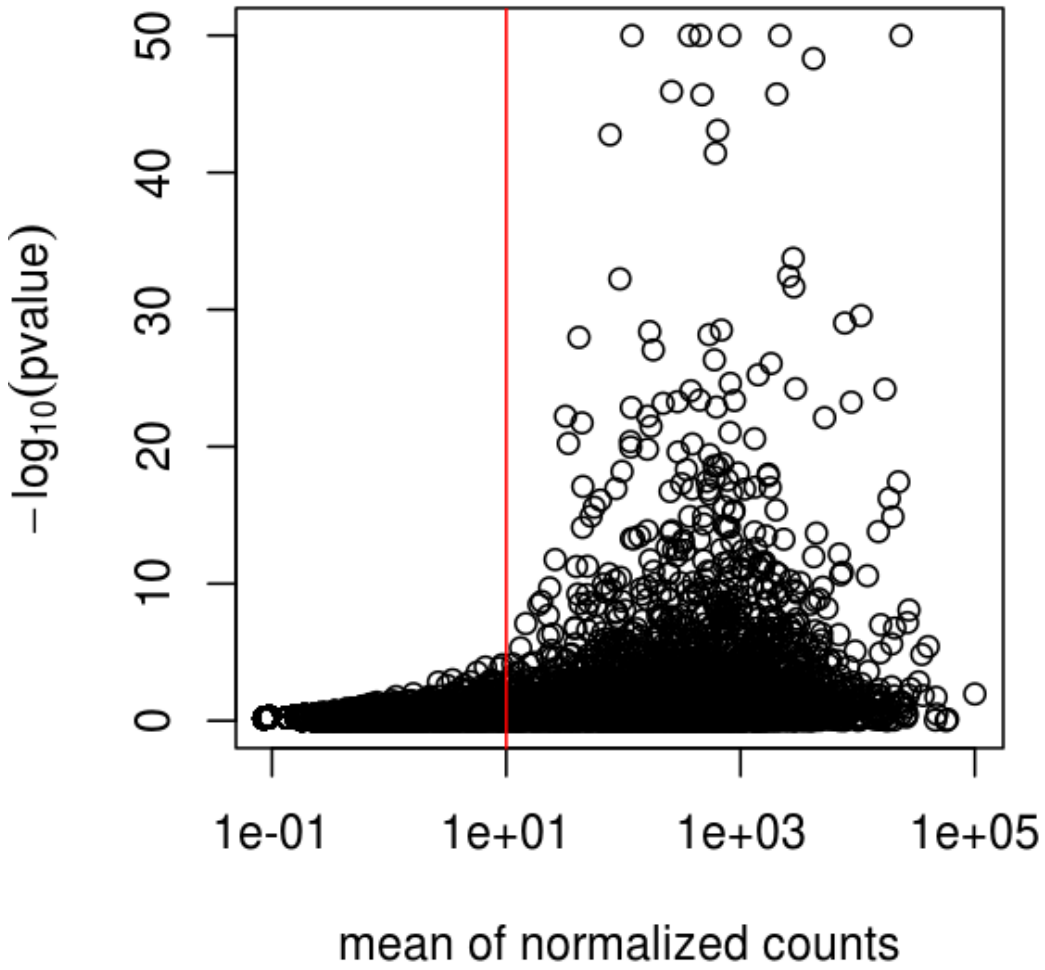


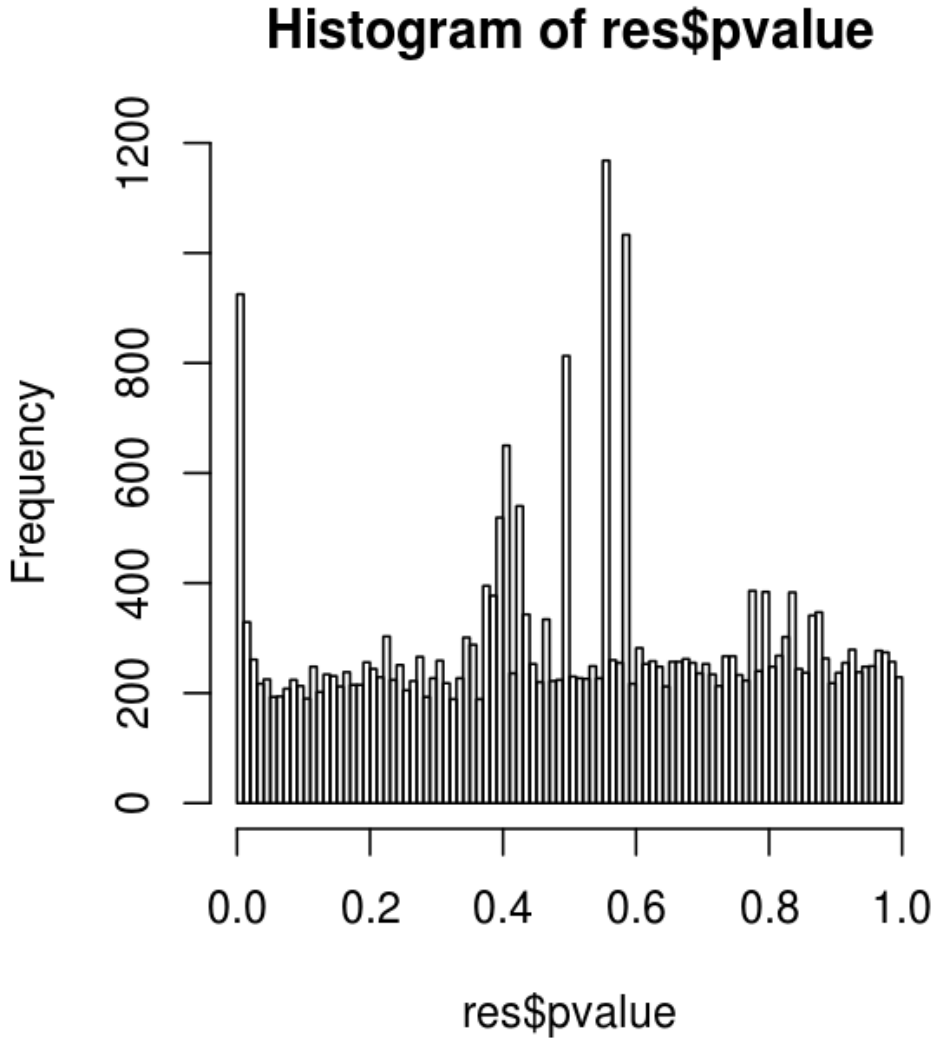
parathyroid dataset





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.

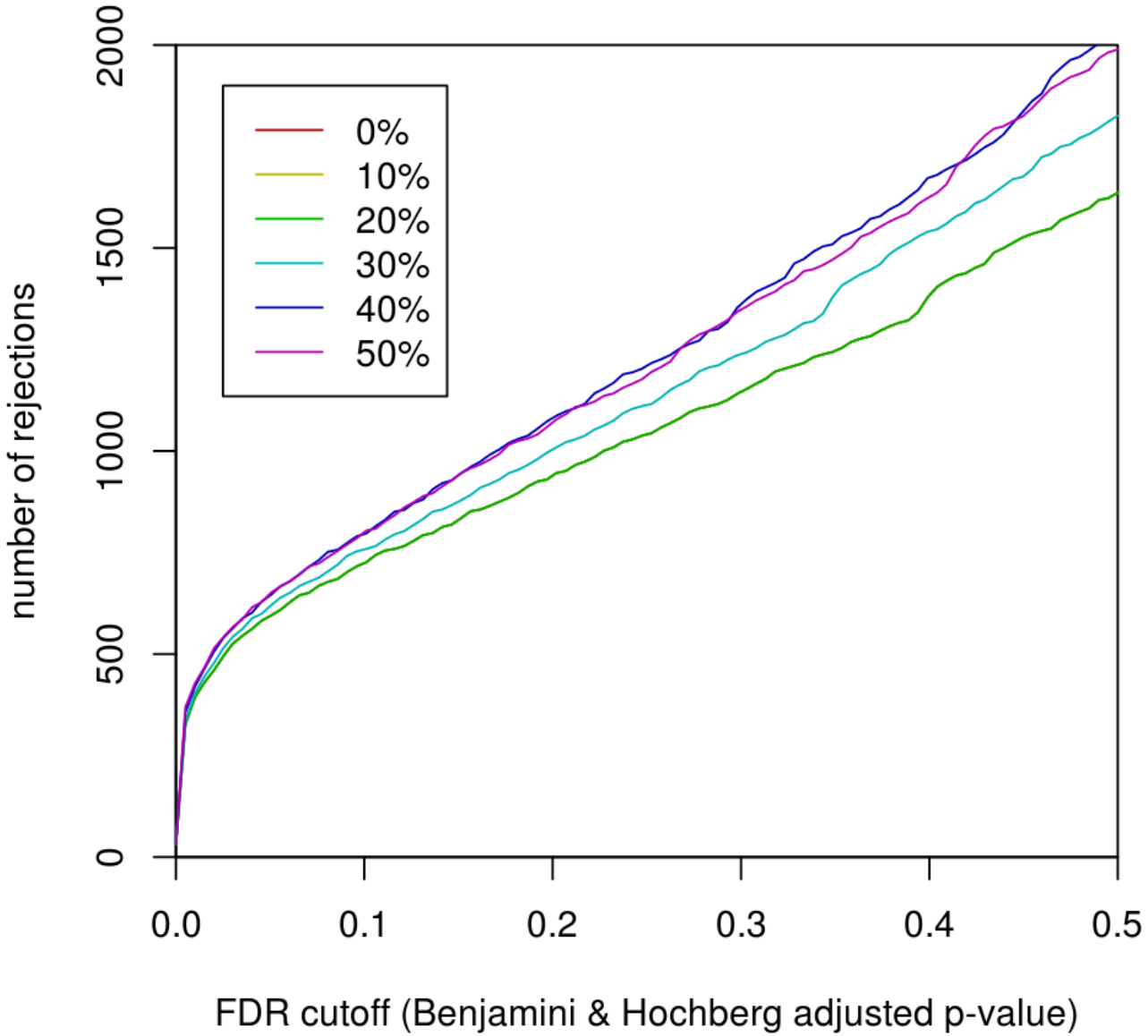
Literature

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.

Independent filtering can increase 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

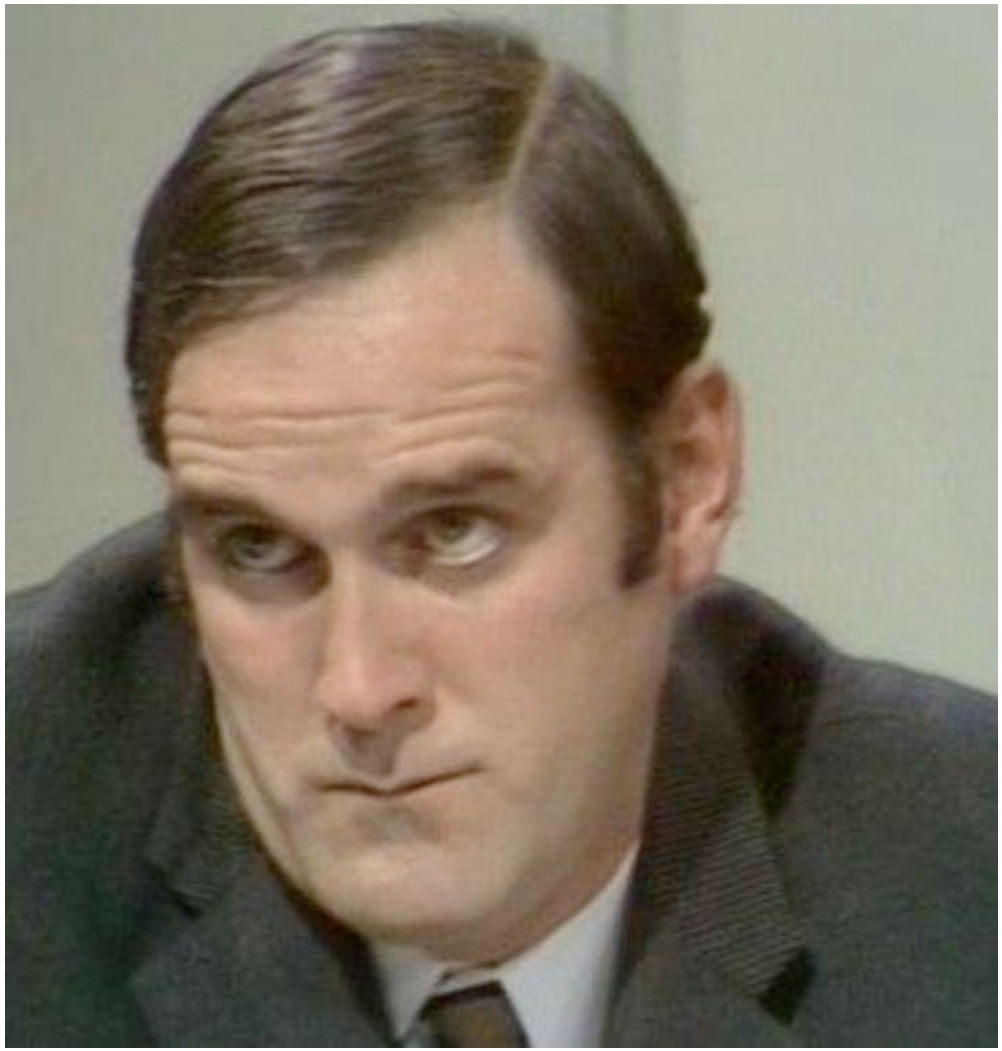
Stage 2: standard NB-GLM test



Really?

Increased detection rate implies increased power

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



Really?

Increase

only if

as bef

vel

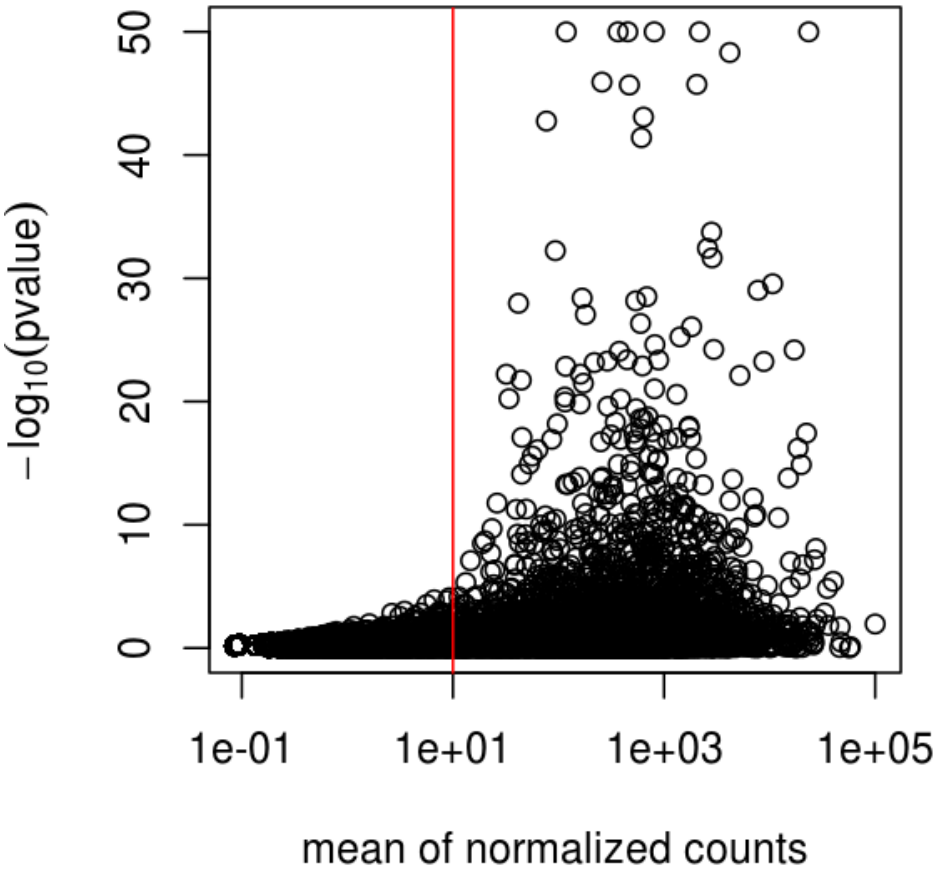
Concern:

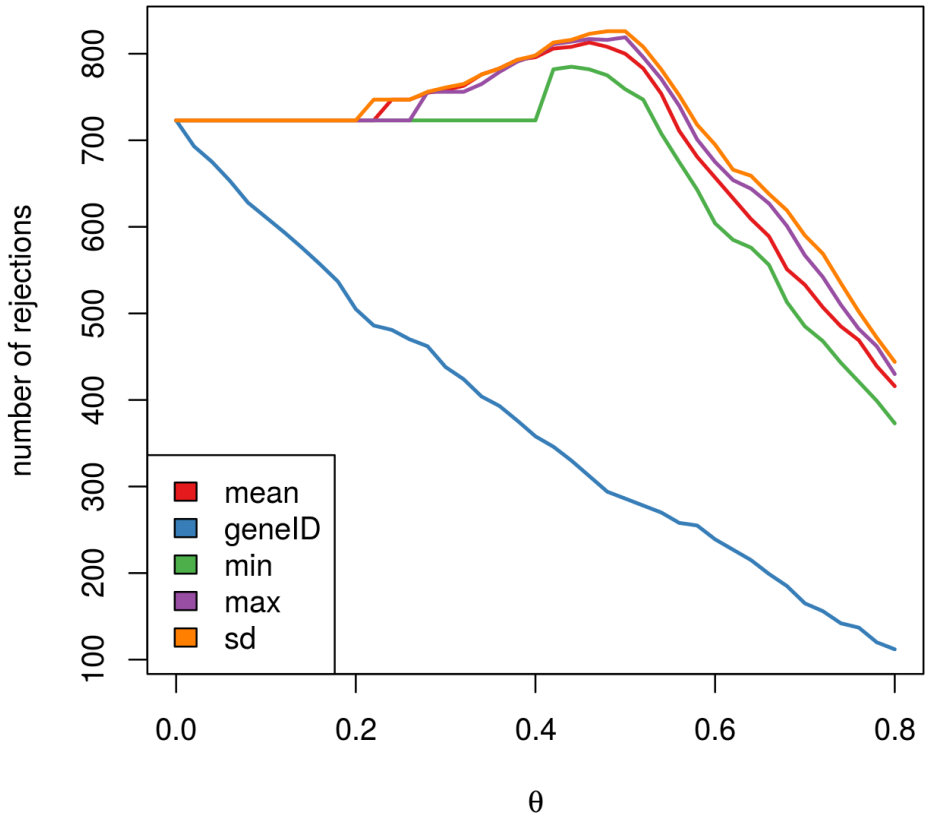
Since we use a data-driven criterion in stage 1, but do p-value and type-I error related computations only on the genes in stage 2, aren't we 'cheating'?

Informal justification:

Filter does not use covariate information

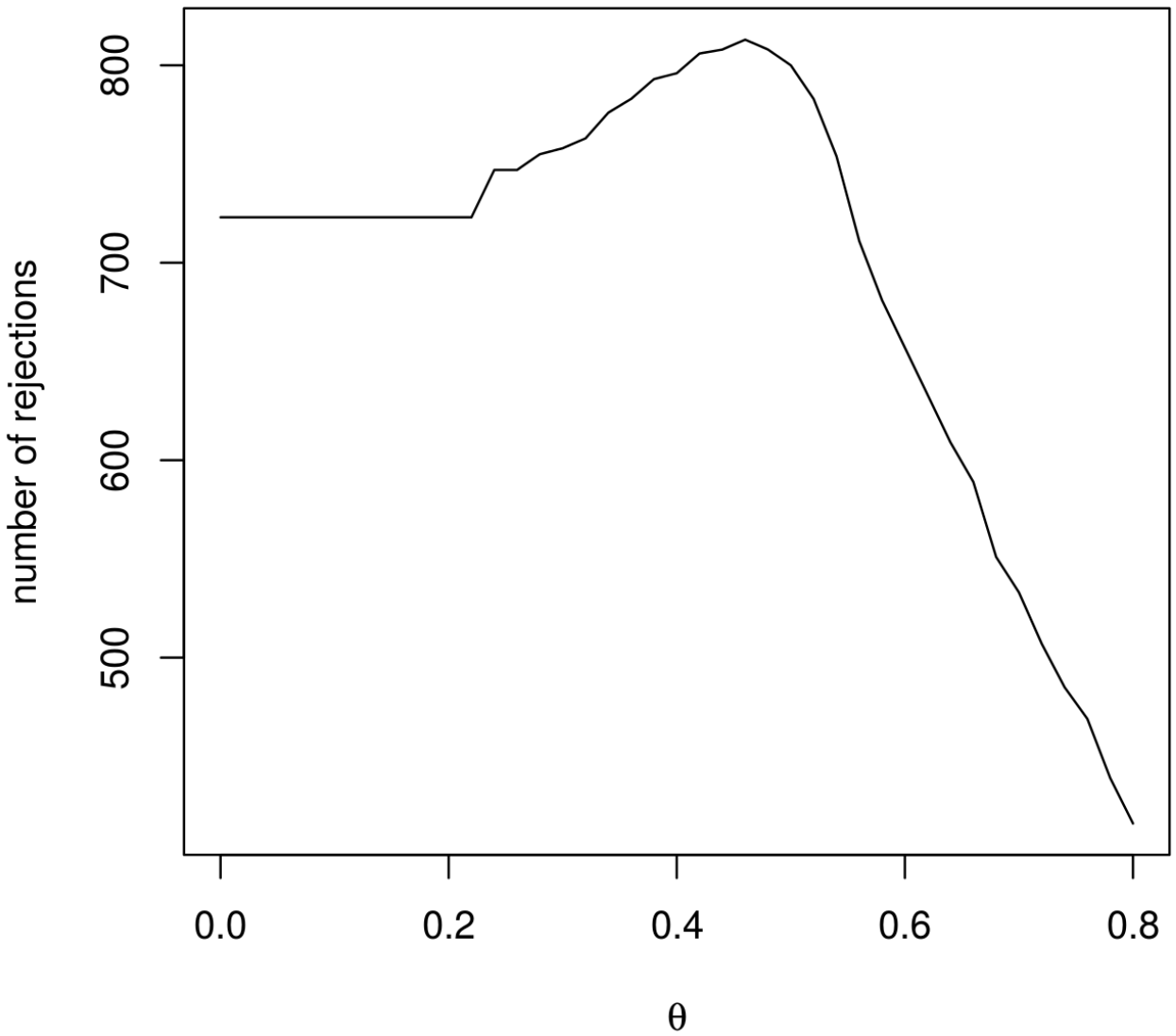
What makes a good filter?





genefilter vignette

How to choose the filter cutoff?

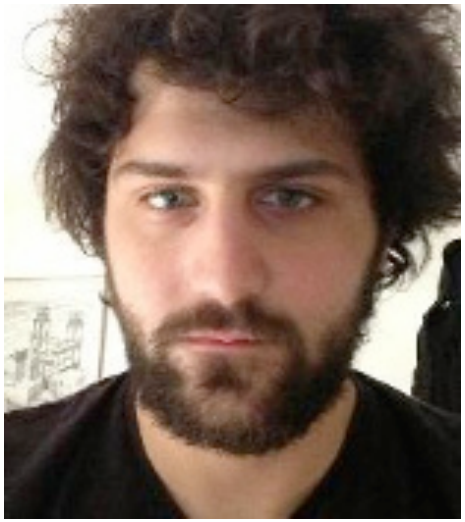


genefilter vignette
DESeq2 - inbuilt default
Caveat - overfitting, optimism of there are no alternatives

Acknowledgments

Robert Gentleman

Richard Bourgon



Nikos Ignatiadis