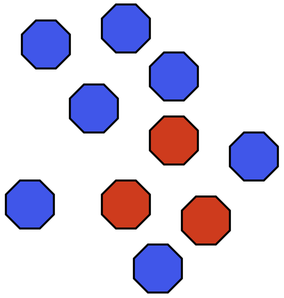
Classical statistics
Issues in their application to ma data

Overview of parametric statistics

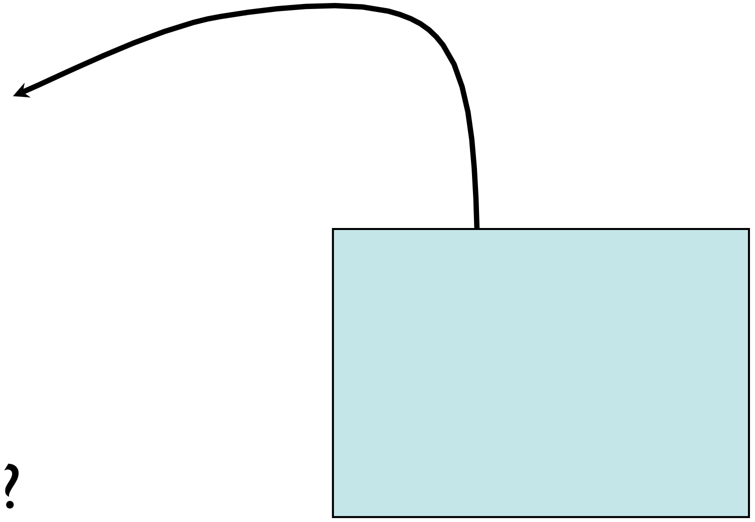
Sampling and testing

Discrete observations

#red = 3



Random sample of 10 balls from the box



10% red balls and 90% blue balls

When do I think that I am not sampling from this box anymore?

How many reds could I expect to get just by chance alone!

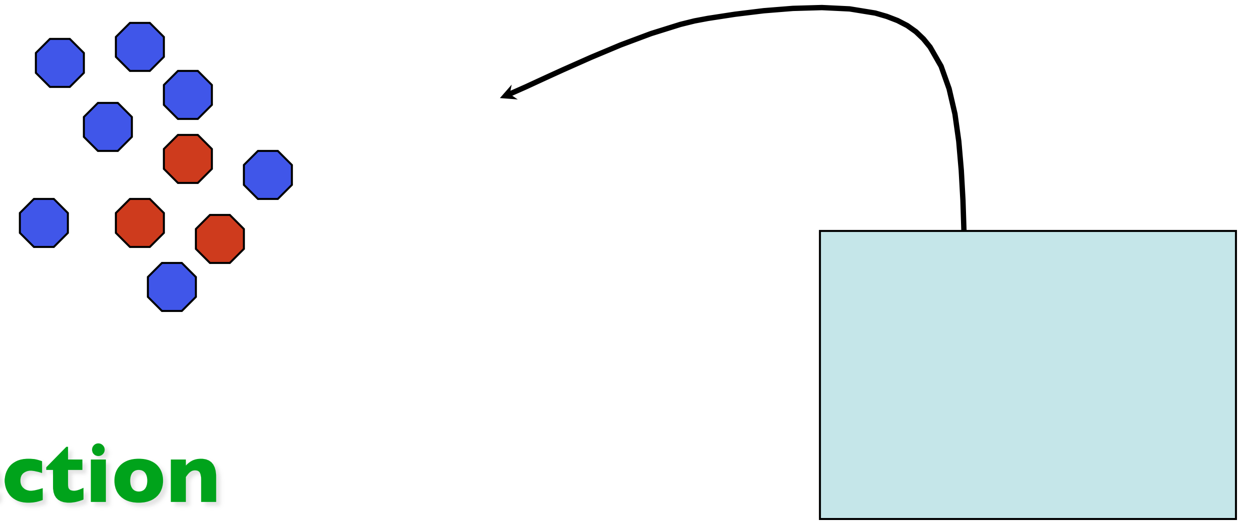
Sample

Random sample of 10 balls from the box

Discrete observations

#red = 3

Test statistic



Rejection criteria (based on your observed sample, do you have evidence to reject the hypothesis that you sampled from the null population)

10% red balls and
90% blue balls

Null hypothesis (about the population that is being sampled)

Sample

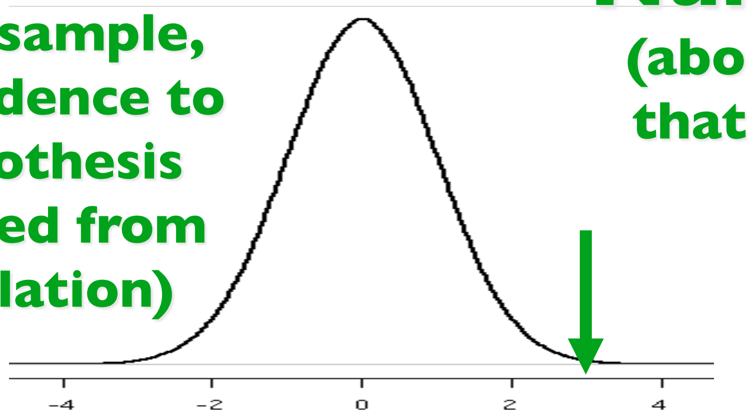
4, 2.3, 5.2, 4.7, 2.1, 3.5, ……..

Continuous observations

mean = 3, sd = 0.6

Test statistic

Rejection criteria (based on your observed sample, do you have evidence to reject the hypothesis that you sampled from the null population)



Null hypothesis

(about the population that is being sampled)

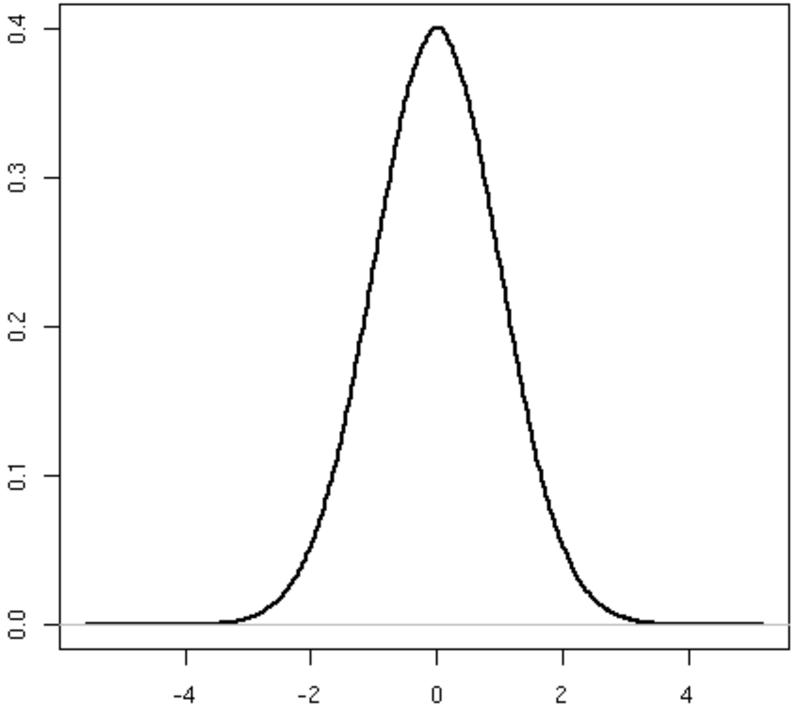
Lets formalise the test!

- Distribution from null population is normal, defined by a mean and variance
  - $N(\mu_0,\sigma_0)$
- Take a sample size n, observe
  - $N(\mu_1,\sigma_1)$
- Null hypothesis : random sample comes from population with $N(\mu_0,\sigma_0)$

- If $x=c(x_1,\ldots x_n) \in N(\mu_0,\sigma_0)$

$$\frac{\text{mean}(x) - \mu_0}{\sqrt{\sigma_0 / n}} \sim N(0, 1)$$

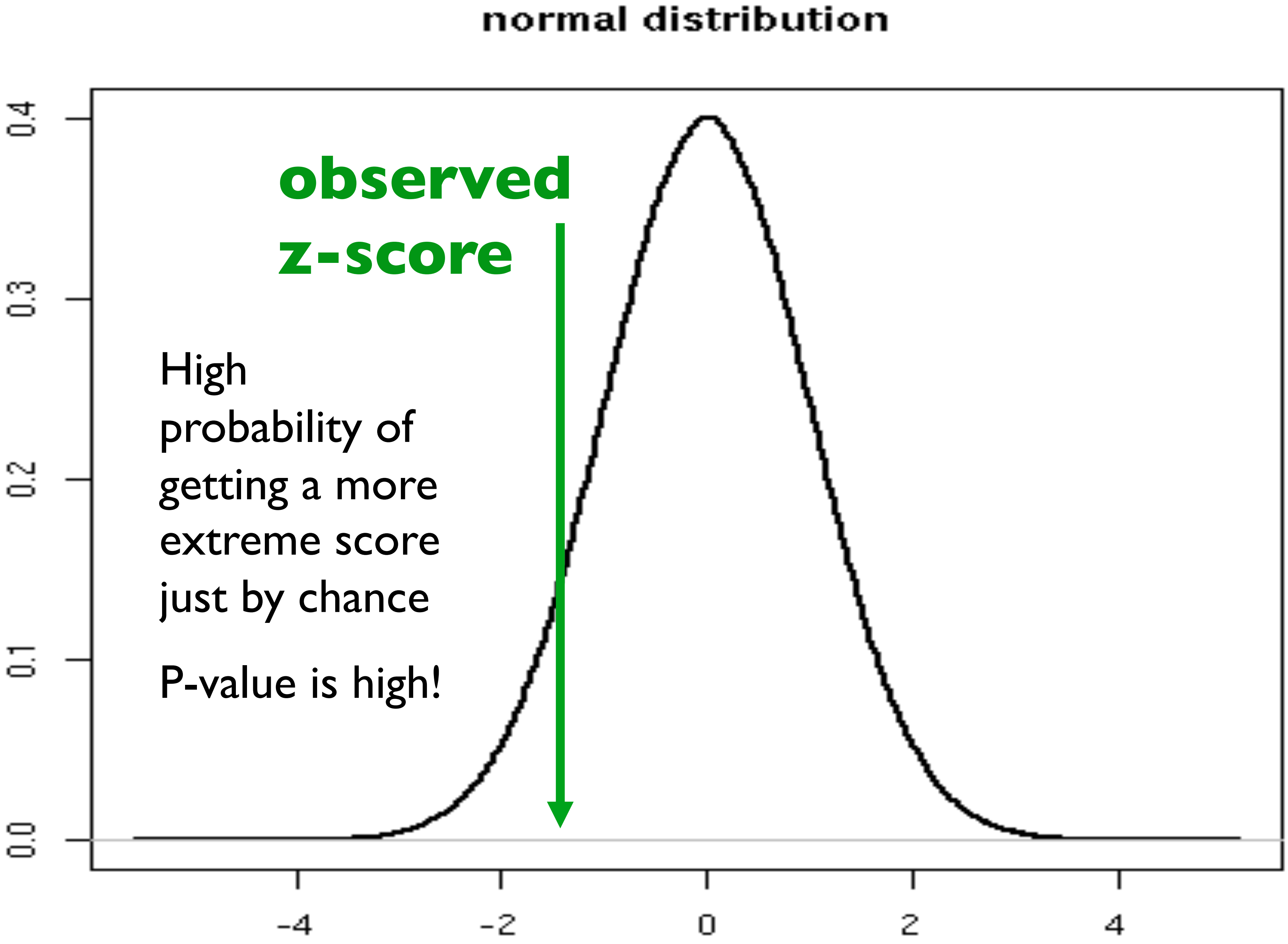
normal distribution



This standardised statistic is called the z-statistic

Most of the data are between +/- 3 sd's

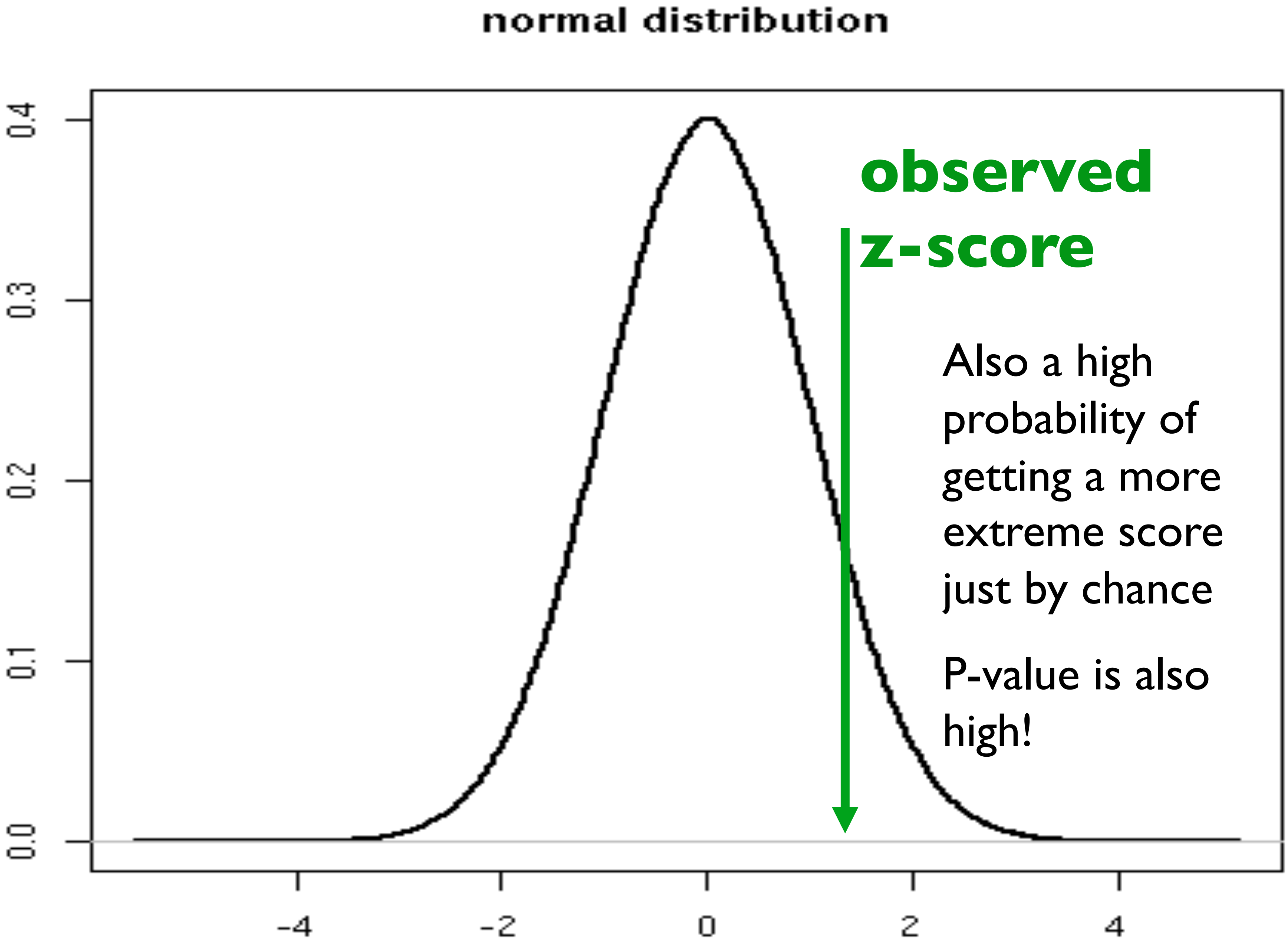
If you observed a z-score greater than +/- 3 this is evidence that you did not sample from the hypothesized distribution.



observed z-score

High probability of getting a more extreme score just by chance

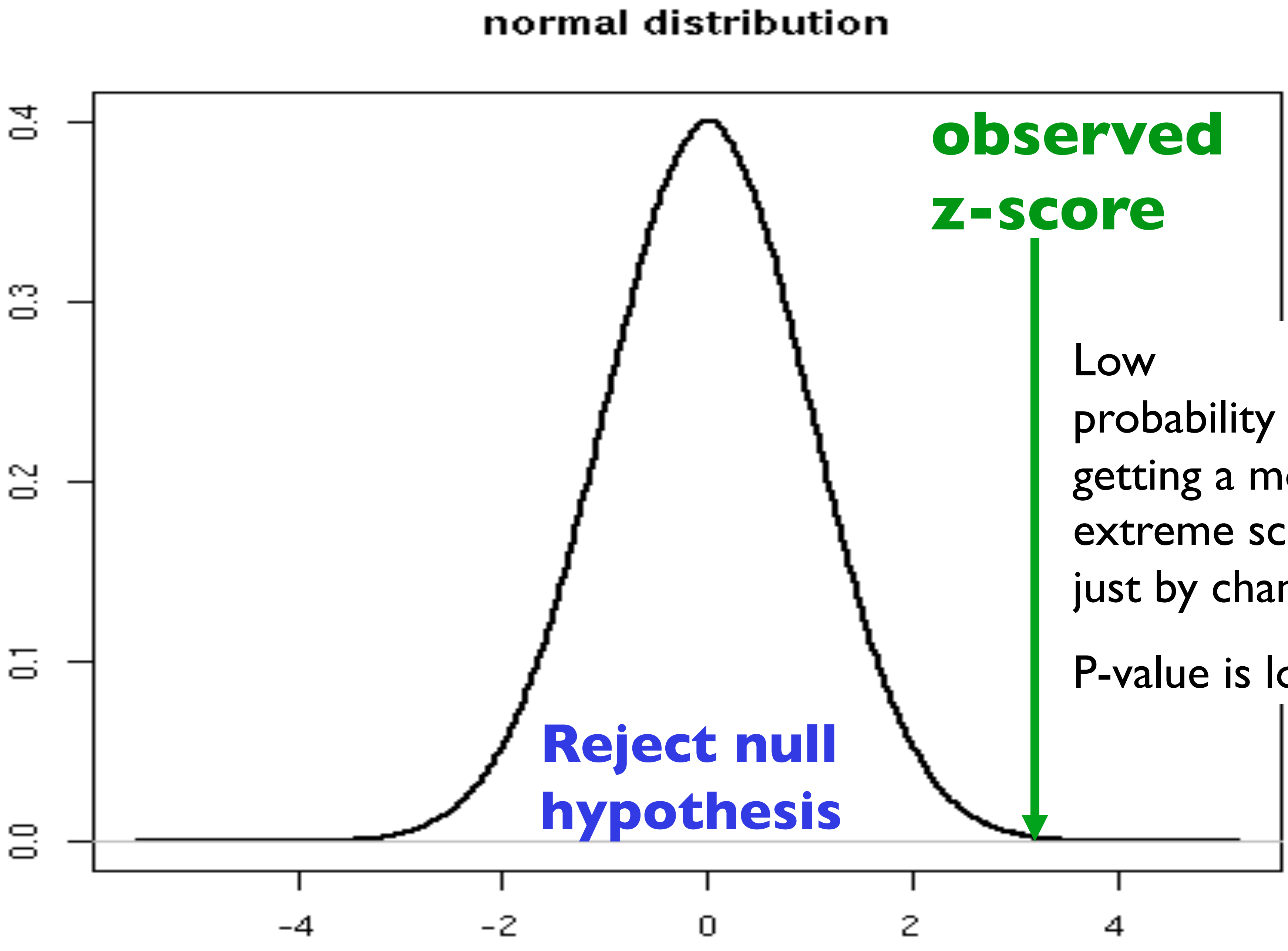
P-value is high!



observed z-score

Also a high probability of getting a more extreme score just by chance

P-value is also high!



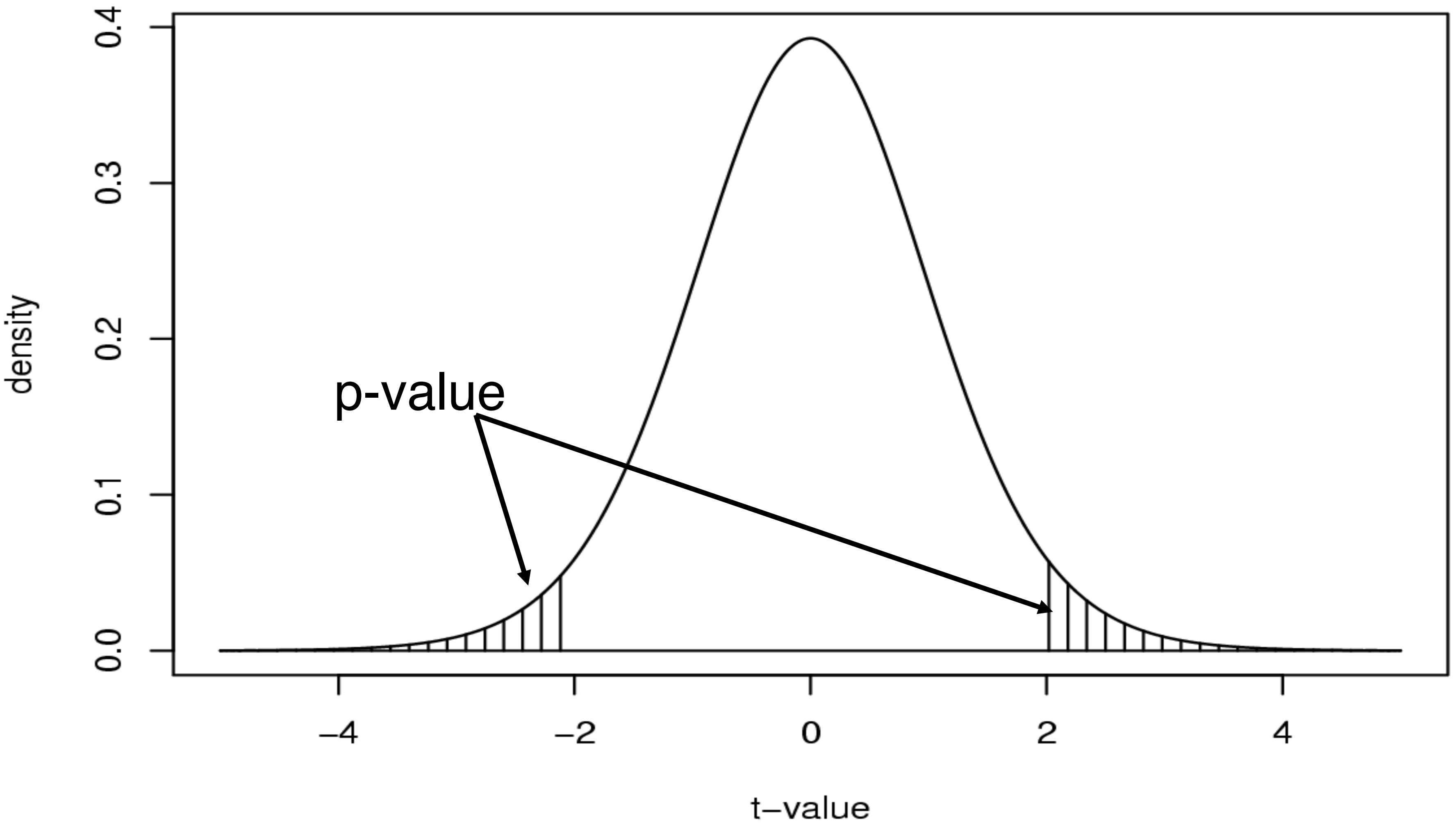
observed z-score

Low probability of getting a more extreme score just by chance

P-value is low

Reject null hypothesis

The p-values for two-sided tests



The **z-score** is based on the assumption that you are sampling from a normal distribution.

The normality assumption doesn't matter so much when you sample large numbers of observations.

The law of large numbers takes care of things for us. In the long run, the mean of a large number of observations tends to be distributed normally.

The z-score assumes that you **know the variance** of the null distribution - in practise this is rarely the case!!

To account for the increased uncertainty in our sample mean and variance estimates we use a **t-distribution** (not a normal distribution), thus we perform what is called a **t-test**

The distribution of the t-test is similar to the normal distribution, except **the tails of the bell cure are wider**.

This makes sense because we are only estimating the variance, from a small number of observations, so we'll get more extreme scores by chance.

As n gets really large, our sample variance estimate gets better and the t-distribution gets closer to the normal (0,1).

But what do we do, if we don't think we are sampling from a normal distribution, or if we have outliers.

We cannot assume a distribution for the mean - we assume no distribution - we use **non-parametric statistics**.

non-parametric statistics in brief

Non-parametric tests usually rely on statistics based on the rank of observations.

<u>Wilcoxon signed rank test : 1 sample test</u>

<u>Wilcoxon rank sum test : 2 sample test</u>
(tests if two samples are equivalent by checking that the ranks of all observations are equally distributed between the two samples)

<u>Kruskall Wallis test : n samples test</u>
(tests if many samples are equal)

All of these tests require a reasonable sample size (not usual with microarray data!)

Issues with classical statistics for microarray data

Differential Expression

- Compare gene expression under different conditions (treated vs untreated, wild type vs knockout, normal vs diseased tissue)

- Differential expression (DE) as list-making exercise: rank genes according to likelihood of (evidence for) DE

- Trade off: list length vs false positive (type 1 error) and false negative (type 2 error).

- What determines fold-change threshold?

- Some p-value for assessing significance would be nice . . .

Differential Expression

Null hypothesis (H_0): gene not DE (M=0).

- First issue: relative gene expression measurements (observed M-values) are not emitted from a normal distribution; in fact M-values have a very heavy tailed distribution (i.e. we get a lot more observations in the tails of the bell curve).

- If we only had one gene, a t-test would be fine, but because we have thousands of t-tests (and each gene comes from a different distribution - different variance) then…

Multiple comparison problem

- $p < 0.05$ means < 1/20 chance of rejecting H_0 when it should be accepted. So what happens if we do 1000+ tests?

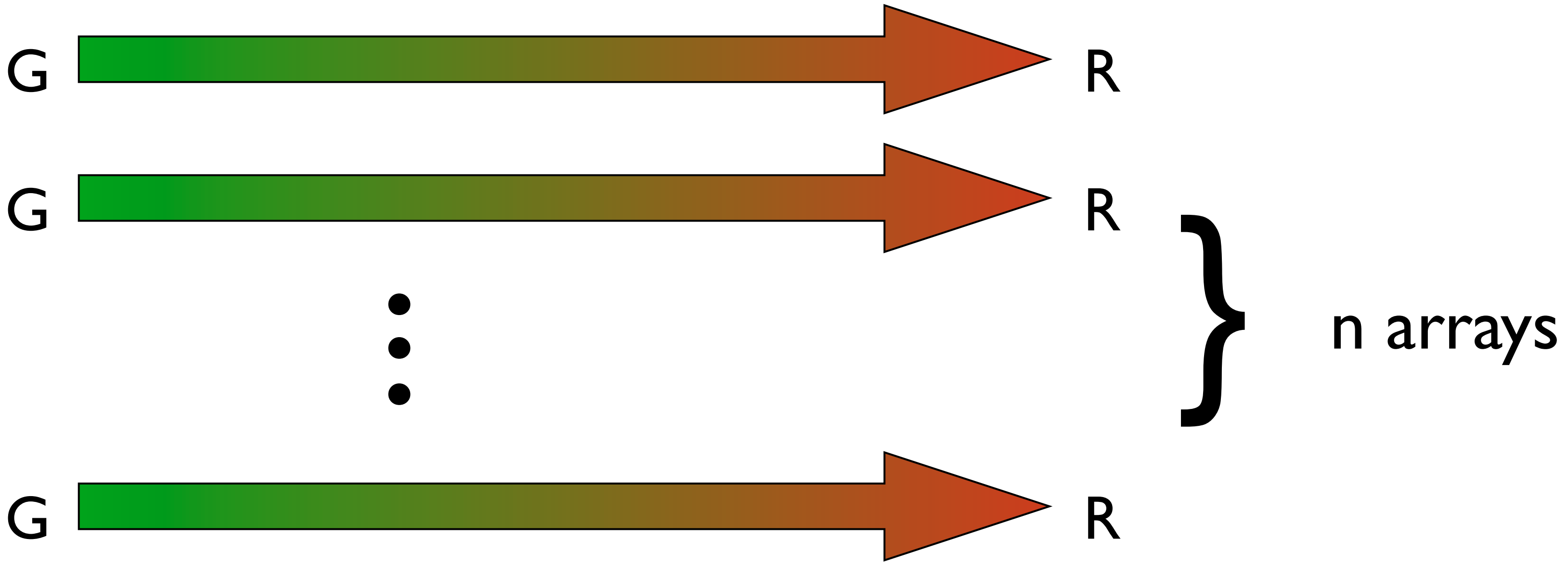
- Smaller p ⇒ harder to reject H_0 (easier to get false negative), but more believable.

- Multiple tests ⇒ need to correct p value.

Moreover, do we have, for each gene, enough replicates for the p-values to be meaningful?

Sample size from replicate arrays

The sample size for each gene is dependent on the amount of replication for each type of array. We must have replication to do statistical analysis of differential expression. We will consider first, the basic problem of all arrays comparing the same two samples (RNA sources).



Gene-wise summaries

- Each gene give a series of log-ratios
- Summarize log-ratios by the average and standard deviation for each gene

$$\begin{tikzcd}
 & M_1, \ldots\ldots, M_n \arrow[dl] \arrow[dr] & \\
M = \text{aveM} & & s = \text{st.dev } M
\end{tikzcd}$$

Summarising replicates to determine differential expression

Obvious thing : average M's

avM

But averages can be driven by outliers

Better than that : account for variability

t = avM / SE

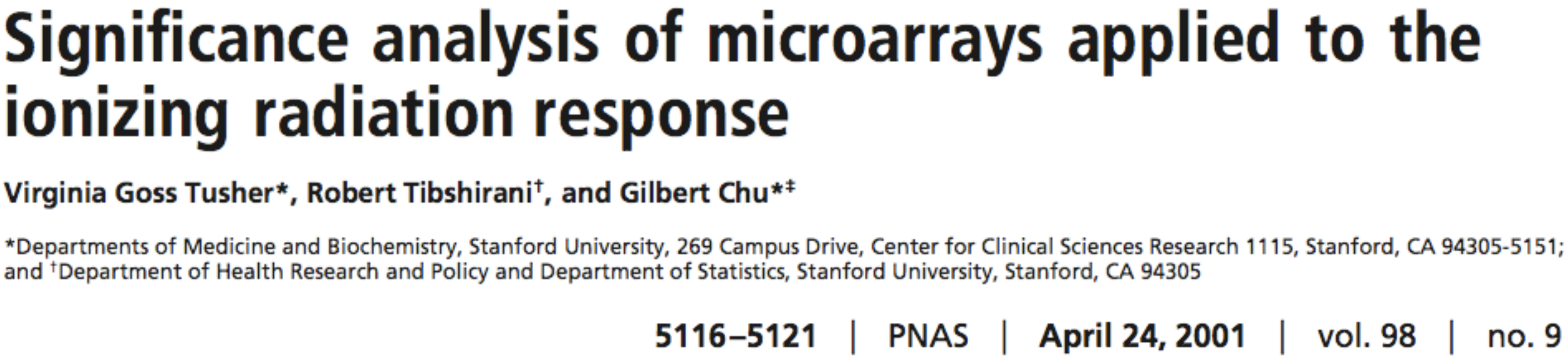
But with 10,000 or so genes, some will have very small SE

Better still : use smoothed SE's

t* = avM / SE*

This is a modified t-statistic (also referred to as a moderated t).

SAM: a modified t-statistic

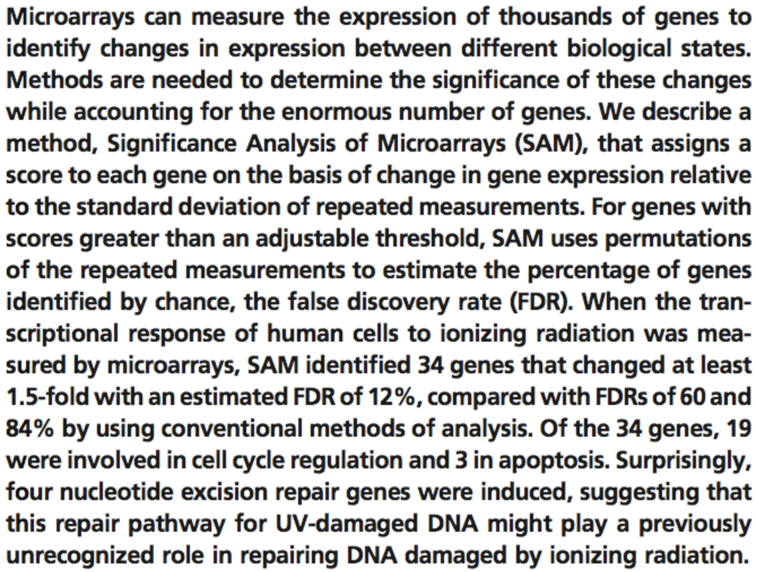


Significance analysis of microarrays applied to the ionizing radiation response

Virginia Goss Tusher[*], Robert Tibshirani[†], and Gilbert Chu[*‡]

[*]Departments of Medicine and Biochemistry, Stanford University, 269 Campus Drive, Center for Clinical Sciences Research 1115, Stanford, CA 94305-5151; and [†]Department of Health Research and Policy and Department of Statistics, Stanford University, Stanford, CA 94305

5116–5121 | PNAS | April 24, 2001 | vol. 98 | no. 9



Microarrays can measure the expression of thousands of genes to identify changes in expression between different biological states. Methods are needed to determine the significance of these changes while accounting for the enormous number of genes. We describe a method, Significance Analysis of Microarrays (SAM), that assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements. For genes with scores greater than an adjustable threshold, SAM uses permutations of the repeated measurements to estimate the percentage of genes identified by chance, the false discovery rate (FDR). When the transcriptional response of human cells to ionizing radiation was measured by microarrays, SAM identified 34 genes that changed at least 1.5-fold with an estimated FDR of 12%, compared with FDRs of 60 and 84% by using conventional methods of analysis. Of the 34 genes, 19 were involved in cell cycle regulation and 3 in apoptosis. Surprisingly, four nucleotide excision repair genes were induced, suggesting that this repair pathway for UV-damaged DNA might play a previously unrecognized role in repairing DNA damaged by ionizing radiation.

SAM: a modified t-statistic

Significance **A**nalysis of **M**icroarrays: very popular method for DE

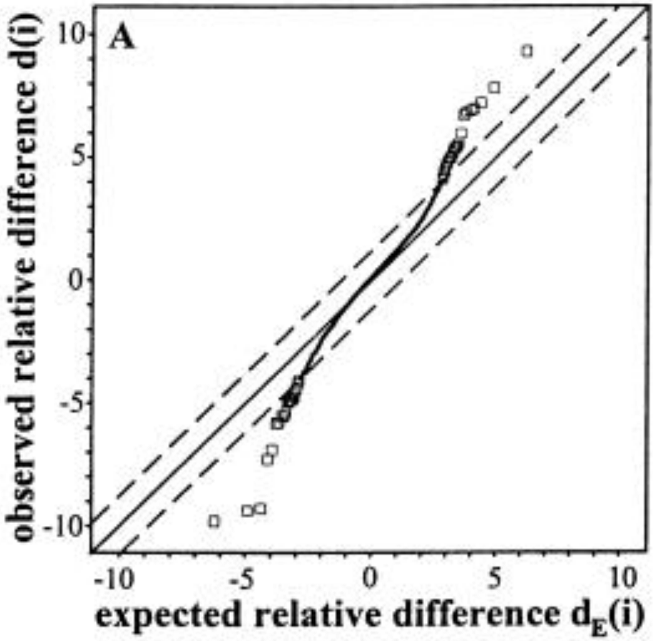
For each gene *i* , calculate:

$$d(i) = \frac{\overline{M}_i}{s_i + s_0}$$

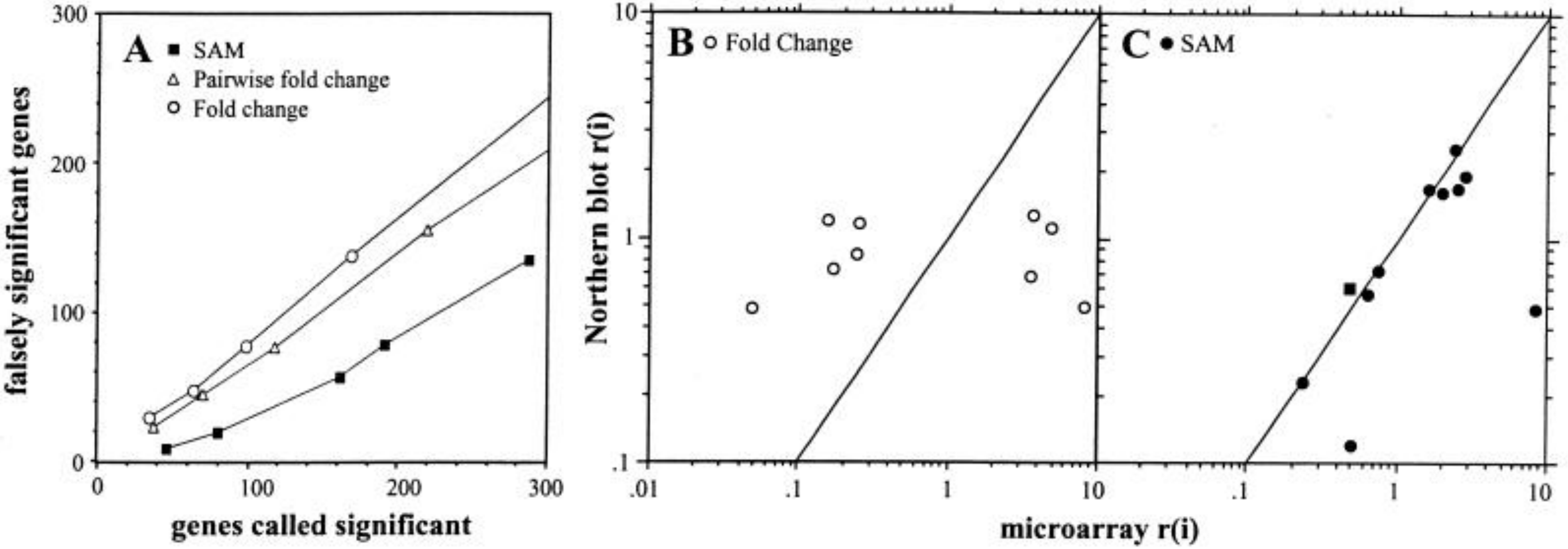
where s_i is the standard deviation of M and s_0 is constant (typically 0-5 %ile of s_i) to minimise variation in *d(i)* with levels of gene expression.

Comparing real *d(i)* with bootstrap samples:

- Rank genes by *d(i)*,so d(1) is largest relative difference.
- For each bootstrap sample (shuffling expression values, *B* bootstraps), calculate $d_p(i)$ and again rank by value, largest first.
- $$d_E(i) = \frac{\sum_p d_p(i)}{B}$$
- Expect $d_E(i) = d(i)$ for most genes.
- $| d(i) - d_E(i) | > \Delta \Rightarrow$ DE candidate.



SAM validation



Tusher VG, Tibshirani R, Chu G. "Significance analysis of microarrays applied to the ionizing radiation response". *Proc Natl Acad Sci U S A*. 2001 Apr 24;98(9):5116-21.

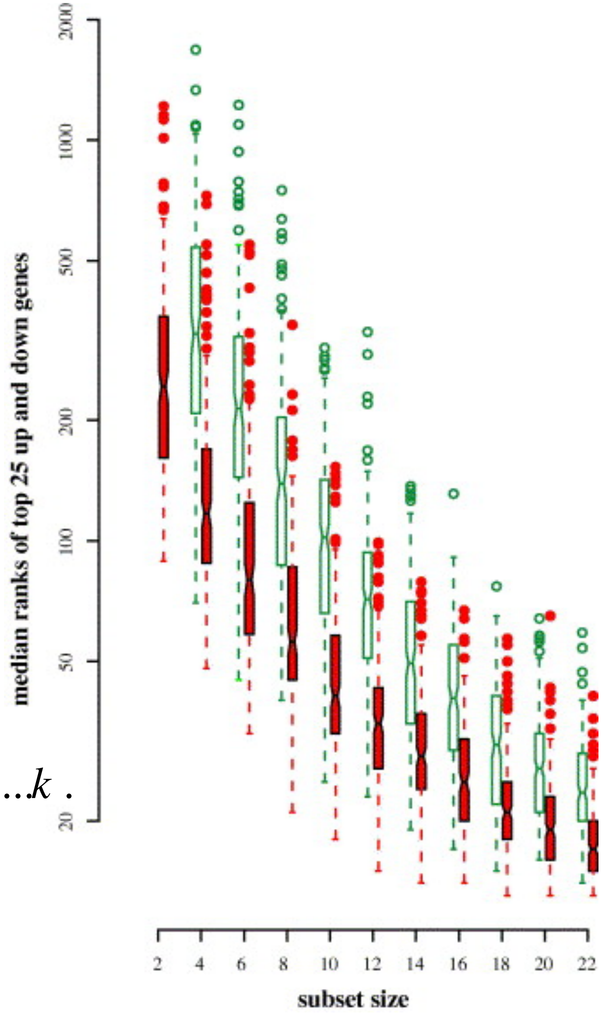
Better than SAM: Rank Products

- Assume 2 colour cDNA chips (single-channels can be accounted for) and compute M of each gene on each chip.
- If gene not DE, very unlikely for gene to be consistently ranked at top of lists sorted by M across replicate chips.
- Looking for up-regulated genes; down-regulated genes handled similarly:

$$RP_g^{up} = \left(\prod_{i=1}^{k} r_{i,g}^{up} \right)^{\frac{1}{k}} \qquad \text{geometric mean}$$

where $r_{i,g}^{up}$ is rank of gene g $(1 = \text{highest};\ n_i = \text{lowest M})$ in chip $i = 1 \ldots k$.

- Significance of rankings assessed using the bootstrap: $q_g = E(RP_g)/rank(g) < FDR$.



Breitling R, Armengaud P, Amtmann A, Herzyk P. "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments". *FEBS Lett*, 2004 Aug 27;573(1-3):83-92.

Even better: the B-statistic (borrowing information from genes)

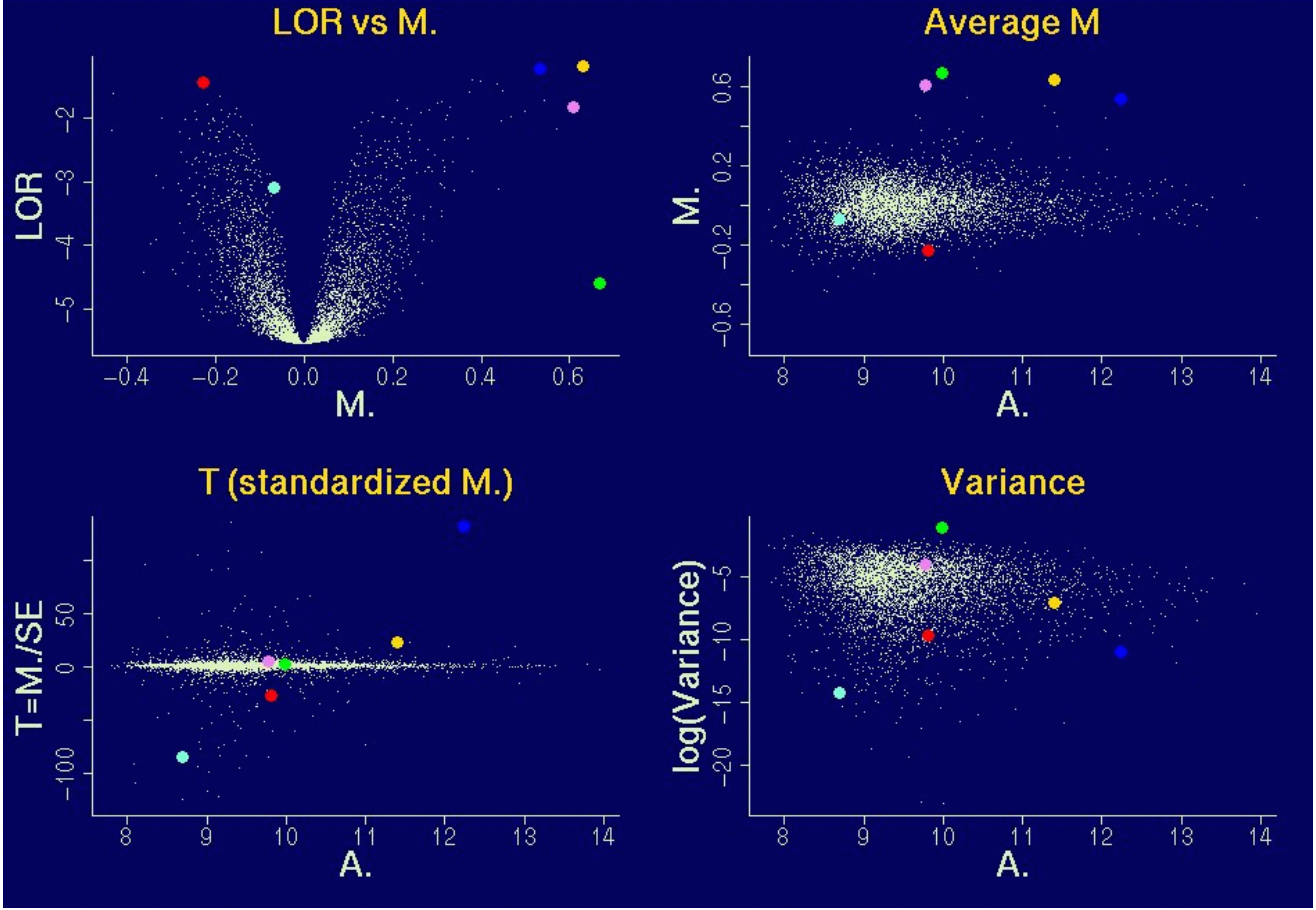
= Similar to a modified t-statistic (smoothes standard errors) =

= It is the log odds ratio of differential expression (LODS, LOR) =

- When there are thousands of genes we can get a better idea of the variability than from just the individual gene variance estimates
- We can't borrow information when there are only a few genes, but when there are tens of thousands of genes we can.
- We want a compromise between individual gene variance estimates and a single variance estimate for all genes.
- The compromise is achieved by empirical Bayes methods which give a weighted combination…

Smyth GK. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments". *Stat Appl Genet Mol Biol*, 2004;3:Article3

B=LOR compared with t and M.



Summary

- Microarray experiments typically have thousands of genes, but only few (1-10) replicates for each gene.
- Averages can be driven by outliers.
- *t*-statistics can be driven by tiny variances.
- B (or moderated t-statistic)
  - use information from all the genes
  - combine the best of *M*. and *t*
  - avoid the problems of *M*. and *t*

Ranking on B could be helpful.

What we want to do is…

- Analyse data all at once
- Use standard deviances not just fold changes
- Use ensemble information to shrink variances
- Assess differential expression for all comparisons together (because microarray experiments will rarely be just a simple comparison between two samples)

Ranking is easier

- How many genes are differentially expressed?
- If there was only one gene, a t-test would give a reliable P-value for judging whether the true log-ratio was zero
- With so many genes, computing absolute P-values on the basis of probability models is problematic
- Much easier to simply rank the genes in order of evidence for differential expression

Why judging significance is hard

- Log-ratios aren't normally distributed, hard to check log-ratios for different genes are correlated in unknown way
- High level of multiple testing means that very small p-values are required – distributional assumptions must hold in extreme tail

Choosing a cut-off

- Could choose a threshold for differential expression if there were known DE and non-DE genes
- Print artificial genes on microarray, then spike corresponding RNA into target RNA before labelling and hybridization
- Choose a cut-off that seem sensible!! Careful and thorough graphical exploration and the choice of ranking statistic are probably the most important aspect to choosing DE. Follow up experimentation that the biologist intends to perform will also play an important role.

Multiple samples

Finding differentially expressed genes when there are more than one kind of array comparison

In classical statistics,

t-test for 1 sample

t-test for 2 samples

regression

anova and

linear models…

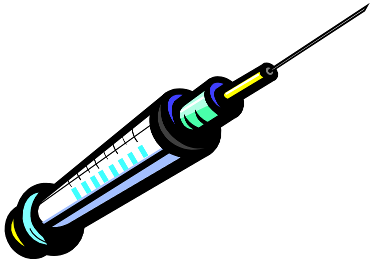
Extensions include dealing with

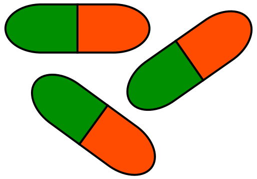
- Replicates within and between slides
- Several effects: use a linear model
- ANOVA: are the effects equal?
- Time series: selecting genes for trends

Is differential expression different between groups
F-statistic

Slides provided by Ingrid Lonnstedt

One-way ANOVA





- Are all the treatments equal?

$$F = \frac{\text{MST}}{\text{MSE}}$$

F-statistic

Note: F statistic in limma has "smoothed" estimates for the denominator

From Ingrid Lonnstedt