

# The “Miss rate” for the analysis of gene expression data

Jonathan Taylor\*

Robert Tibshirani †

and Bradley Efron‡

## Abstract

Multiple testing issues are important in gene expression studies, where typically thousands of genes are compared over two or more experimental conditions. The false discovery rate has become a popular measure in this setting. Here we discuss a complementary measure, the “miss rate” and show how to estimate it in practice.

---

\*Department of Statistics, Stanford University, Stanford, CA 94305;  
jtaylor@stat.stanford.edu

†Department of Health Research & Policy and Department of Statistics, Stanford University, Stanford, CA 94305; tibs@stanford.edu

‡Departments of Statistics, and Health Research & Policy, Stanford University, Stanford, CA 94305; brad@stat.stanford.edu

# 1 Introduction

We discuss the problem of identifying differentially expressed genes from a set of microarray experiments. This problem has received much attention lately— see Dudoit et al. (2003) for a nice summary. The false discovery rate (FDR) (Benjamini & Hochberg 1985) has become a popular error measure in this setting, see e.g. Tusher et al. (2001), Efron et al. (2001), Storey (2002*a*), Storey & Tibshirani (2003), Genovese & Wasserman (2003). In this short paper, we introduce the “miss rate”, which is the complement of the FDR. It is the proportion of genes that are truly differentially expressed, among those declared non-significant. We show how to estimate the miss rate in practice, and discuss its properties both numerically and from a mathematical point of view.

## 2 T-statistics, thresholding and the False Discovery Rate

Suppose we have  $m$  genes measured on  $n$  arrays, under two different experimental conditions. Let  $\bar{x}_{i1}$  and  $\bar{x}_{i2}$  be the average gene expression for gene  $i$  under conditions 1 and 2, and let  $s_i$  be the pooled standard deviation for

gene  $i$ :

$$s_i = [(1/n_1 + 1/n_2) \cdot \sum_1 (x_{ij} - \bar{x}_{i1})^2 + \sum_2 (x_{ij} - \bar{x}_{i2})^2]^{1/2}$$

Here  $n_k$  is the number of arrays in condition  $k$ , and each summation is taken over its respective group. Then a reasonable test statistic for assessing differential gene expression is the standard (unpaired) t-statistic:

$$T_i = \frac{\bar{x}_{i2} - \bar{x}_{i1}}{s_i \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

For simplicity, our discussion focusses on the two-sample problem and the unpaired T-statistic, but it applies equally well to other settings and test statistics.

Using the statistic  $T_i$ , we can simply compute its value for each gene, choose a threshold  $c$  and then declare significant all genes satisfying  $|T_i| > c$ .

Table 1: *Possible Outcomes from  $m$  Hypothesis Tests*

	Accept	Reject	Total
Null True	$U$	$V$	$m_0$
Alternative True	$Q$	$S$	$m_1$
	$W$	$R$	$m$

Table 1 displays the various outcomes when testing  $m$  genes. The quantity  $V$  is the number of false positives (Type I errors), while  $R$  is the total number

of hypotheses rejected. The false discovery rate (FDR) is the expected value of  $V/R$ .

Consider for example the microarray data taken from Golub et al. (1999). It consists of the expression of 6087 genes in 38 leukemia patient samples: 27 with ALL and 11 with AML. The objective is to find genes whose expression differs across the two types of leukemia.

A histogram of the 6087  $T_i$  values is shown in Figure 1: they range from  $-7.5$  to  $10.1$ . If the  $T_i$  values were normally distributed, we could consider any value  $> 2$  in absolute value to be significantly large. But with more than 6000 genes, we would expect many to have  $|T_i| > 2$  just by chance.

We proceed by considering rules of the form  $|T_i| > c$ , for various values of the cutpoint  $c$ , and estimating the FDR of each rule by taking random permutations of the class labels. Here are the details:

1. Create  $K$  permutations of the data, producing T-statistics  $T_i^k$ , for features  $i = 1, 2, \dots, p$  and permutations  $k = 1, 2, \dots, K$ .
2. For a range of values of the cutpoint  $C$ , let  $\hat{R} = \sum_i I(|T_i| > C)$ ,  $\hat{V} = (1/K) \sum_{i,k} I(|T_i^k| > C)$ . Let  $\pi_0 = m_0/m$ , the true proportion of null genes among the  $m$
3. Estimate the FDR by  $\widehat{\text{FDR}} = \pi_0 \hat{V} / \hat{R}$ .

Of course  $\pi_0$  is unknown: we can estimate it in a number of ways. Here

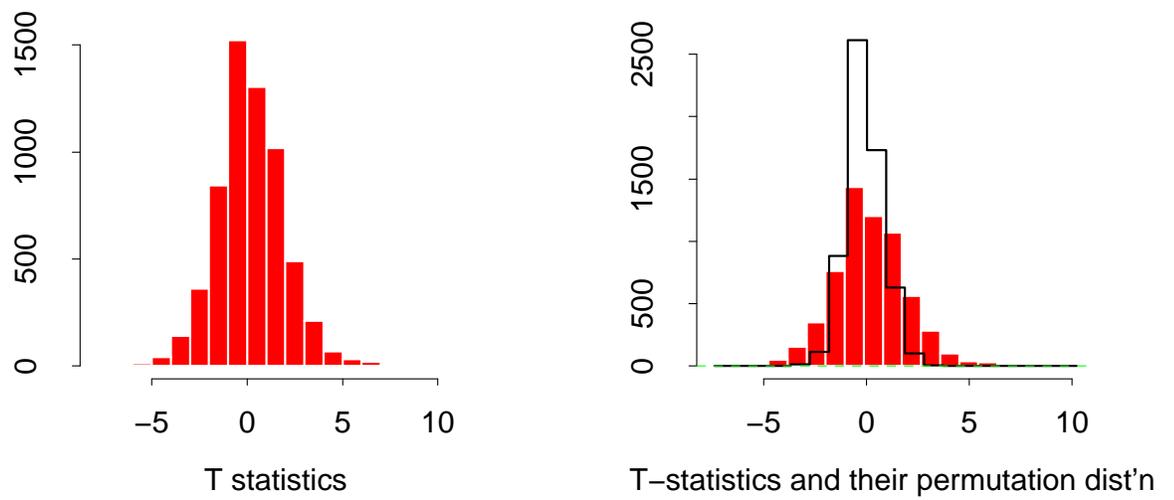


Figure 1: *Microarray example: top left panel shows a histogram of the 6087 T-statistics. In the top right panel we have overlaid the histogram of the T-statistics from the 100 data permutations.*

is one simple approach, from Storey (2002a). Let  $(q_{.25}, q_{.75})$  be the quartiles of the T-statistics from the permuted datasets. Let  $\hat{\pi}_0 = \#\{T_i \in (q_{.25}, q_{.75})\}/(.5m)$ , and set  $\hat{\pi}_0 = \min(\hat{\pi}_0, 1)$ .

In our example, if we take  $c = 2.9$ , we get  $R = 609$ ,  $\hat{V} = 31.9$ ,  $\hat{\pi}_0 = .70$ , giving  $\widehat{\text{FDR}} = .037$

### 3 The Miss Rate

Having derived a list of genes from a using a rule like  $|T_i| > c$ , it is of interest to estimate some sort of false negative rate. Looking at Table 1, the quantity  $E(Q/W)$  is what Genovese & Wasserman (2003) call the false non-discovery rate. This quantity is the proportion of false negatives among all genes with  $|T_i| < c$ . Since the vast majority of these genes have values of the T-statistic near zero, this quantity would not usually be of practical interest. Consider instead some cutpoint  $c_0 < c$ , chosen for example so that say 5% of the values  $|T_i|$  lie in  $(c_0, c)$ . Then we call the *miss rate* the expected proportion of genes in  $(c_0, c)$  that are non-null.

In detail, consider the definitions in Table 2. The miss rate is defined to be

$$\text{MR}(c, c_0) = E\left(\frac{Q_0}{W_0}\right) \tag{1}$$

For example taking  $c_0 = 2.46$  gives  $W_0 = 305$  genes with values of  $T_i$  in

Table 2:

	$ T_i $ in $(c_0, c)$	$ T_i  > c$ ( Reject)	Total
Null True	$U_0$	$V$	$m_0$
Alternative True	$Q_0$	$S$	$m_1$
	$W_0$	$R$	$m$

$(c_0, c)$ . The estimated miss rate for this interval, calculated in a way described below, is 85.3%. Thus we estimate that  $.853 \cdot 305 = 260.2$  of these 357 genes are non-null, i.e differentially expressed across the two groups.

The miss rate MR is estimated using the same information gathered for the estimation of the FDR. With  $W_0$  equal to the number of  $|T_i|$  in  $(c_0, c)$  and  $\hat{U}_0$  equal to the average number of permutation values  $|T_i^K|$  in  $(c_0, c)$ , we define

$$\widehat{\text{MR}}(c_0, c) = \frac{W_0 - \hat{\pi}_0 U_0}{W_0} \quad (2)$$

The miss rate serves as a useful cautionary statistic. The estimated FDR is low here (3.7%), so we are happy that only few among our list of 609 genes are false positives. However among the next best 305 genes (all declared non-significant), an estimated 85.3% are actually non-null.

Table 3 gives the estimated miss rates for successive intervals below the cutpoint of 2.89, each containing 5% of the genes. We see that the miss rate

Interval	Number of genes	Miss Rate
(2.45, 2.89)	305	.853
(2.15, 2.46)	304	.604
(1.91, 2.15)	304	.274

Table 3: *Estimated miss rates, Leukemia example*

does not become low until we get down to values of the T-statistic around 2.0

When estimating both the FDR and miss rate, it is possible to obtain values either  $< 0$  or  $> 1$ . In each case the corresponding estimate is set to 0 or 1 respectively.

There is a close relationship between the miss rate and the *local false discovery rate* (fdr) defined in Efron et al. (2001) and Efron & Tibshirani (2002). The local fdr is the false discovery rate in an infinitesimal interval  $(c - \epsilon, c)$ . The miss rate is 1 minus the local fdr, taken over a large interval  $(c_0, c)$ . In defining the missrate, we have focus on a larger interval to aid in interpretability.

Another important issue is the choice of the distribution for the test statistics under the null hypothesis. Here we have used a permutation distribution for the null, which is simple an convenient. But as shown in Efron (2004), this can sometimes be under-dispersed, resulting in under-estimation of the

cutpoint	> 1.4	> 1.6	> 2.0	> 2.7	> 4.1
FDR	0.601	0.504	0.348	0.107	.005
$\widehat{\text{FDR}}$	0.600	0.500	0.349	0.107	.006
# genes $R$	250	200	150	100	50
MR	0.008	0.030	0.169	0.791	
$\widehat{\text{MR}}$	0.054	0.084	0.174	0.792	
$\hat{\pi}_0 U_0$	49.21	45.89	41.82	11.55	
# genes $W_0$	50	50	50	50	

Table 4: *FDR and MR results for simulated data, Example 1. The cutpoints represent the 75, 80, 85, 90 and 95 percentiles of  $|T_i|$*

false discovery rate (and that is probably the case in this example). Efron (2004) discusses alternative methods for generating the null distribution of the test statistics.

## 4 A simulation study

We simulated data from  $p = 1000$  genes and  $n = 40$  samples, in two groups of size 20, All values were generated i.i.d.  $N(0,1)$  except for the first 100 genes in samples 21–40, which were  $N(1.25, 1)$ . Table 4 shows the results averaged over 50 simulations.

The simulation standard errors are  $< .01$  for FDR, MR and their estima-

tors. In general both  $\widehat{\text{FDR}}$  and  $\widehat{\text{MR}}$  do a reasonably good job of estimating the false discovery rate and miss rate, respectively. When MR is low, the  $\widehat{\text{MR}}$  over-estimates it on average, due to the truncation of  $\widehat{\text{MR}}$  at zero. We also note the estimate  $\hat{\pi}_0$  averaged .91, close to the actual value  $\pi_0 = .90$ .

Table 5 shows a second simulation example, as above but with 2000 genes, with the first 300 genes differing in mean by 0.25 units in the second set of 20 samples. Again, both estimates are accurate enough to be informative in practice.

cutpoint	> 1.3	> 1.5	> 1.7	> 2.0	> 2.5
FDR	0.637	0.595	0.541	0.465	.349
$\widehat{\text{FDR}}$	0.697	0.654	0.593	0.515	.401
# genes $R$	250	200	150	100	50
MR	0.194	0.244	0.306	0.419	
$\widehat{\text{MR}}$	0.133	0.162	0.252	0.371	
$\hat{\pi}_0 U_0$	86.99	83.83	74.81	62.93	
# genes $W_0$	100	100	100	100	

Table 5: *FDR and MR results for simulated data, Example 2. The cutpoints represent the 75, 80, 85, 90 and 95 percentiles of  $|T_i|$*

Table 6 shows the results of a third simulation study. The setup is the same as in Example 1, except that the 1000 genes have been divided

cutpoint	> 1.4	> 1.6	> 2.0	> 2.7	> 4.1
FDR	0.603	0.508	0.357	0.121	.012
$\widehat{\text{FDR}}$	0.604	0.503	0.360	0.122	.012
# genes $R$	250	200	150	100	50
MR	0.014	0.039	0.171	0.769	
$\widehat{\text{MR}}$	0.052	0.105	0.179	0.768	
$\hat{\pi}_0 U_0$	50.42	45.645	41.81	11.55	
# genes $W_0$	50	50	50	50	

Table 6: *FDR and MR results for simulated data, Example 3 (correlated data)* The cutpoints represent the 75, 80, 85, 90 and 95 percentiles of  $|T_i|$

into 20 blocks of 50 consecutive genes. Within each block  $k$ , we add  $\theta_k = 5 \cdot (|z_1|, |z_2|, \dots, |z_{40}|)$  to the expression values for each gene, where  $z_j$  is a standard Gaussian variate. This makes the pairwise correlation of genes in a block equal to about 0.35. This correlation has little effect on the results.

## 5 Some Theory

Our estimate of the miss rate is

$$\widehat{\text{MR}}(c_0, c) = 1 - \hat{\pi}_0 \cdot \frac{\widehat{U}_0}{W_0}$$

where  $\widehat{U}_0$  is the average number of permutation values of the  $|T_i^K|$  in the interval  $(c_0, c)$ .

Under some reasonable conditions, our estimate of MR will be consistent and (asymptotically) conservative. The main assumption that is required is that as the number of genes  $m$  grows and the number of permutation samples  $K$  grows, the empirical distribution functions

$$\widehat{F}(t) = \frac{1}{m} \sum_{i=1}^m I(|T_i| \leq t)$$

$$\widehat{F}_0(t) = \frac{1}{K \cdot m} \sum_{i=1}^m \sum_{k=1}^K I(|T_i^k| \leq t)$$

converge uniformly to non-random limits, say  $F(t)$  and  $F_0(t)$  and that the proportion of true nulls  $\pi_{0,m} = m_0/m$  converges to some limiting proportion  $0 < \pi_0 < 1$ .

The simplest example under which these conditions are satisfied are when the genes are independent; the null distribution of each  $T_i$  is the same; and the active genes are drawn i.i.d. from a mixture distribution so that the alternative distribution of the  $T$ 's are also identical (Genovese & Wasserman 2002, Storey & Tibshirani 2001). For a more precise description of the necessary conditions, the interested reader is referred to (Storey et al. 2004, Storey & Tibshirani 2001).

The distribution function  $F_0$  can be thought of as the “null” distribution of a typical inactive gene. For many models, this null distribution is the same

across genes, but, in general it is possible that the null distribution is different across genes, in which case  $F_0$  is the mixture of these null distribution across genes. The distribution function  $F$  can be thought of as a mixture which puts weight  $\pi_0$  on  $F_0$  – the “null” distribution of  $T_i$  and weight  $(1 - \pi_0)$  on the “alternative” component  $F_1$ .

If  $F$  and  $F_0$  are continuous, so that the quantiles of the  $|T_i^k|$  also converge, then the estimate  $\hat{\pi}_{0,m}$  also converges to  $\hat{\pi}_{0,\infty}$ , say, (Storey 2002b, Storey et al. 2004) and

$$\hat{\pi}_{0,\infty} \geq \pi_0.$$

Therefore, asymptotically

$$\widehat{MR}(c_0, c) \leq MR(c_0, c)$$

both in probability and expectation (Efron & Tibshirani 2002), (Storey et al. 2004). Thus, our estimate of  $MR(c_0, c)$  is asymptotically conservative and the true miss rate is actually higher than our estimate on average. In our simulation experiments, the bias in  $\hat{\pi}_0$  was very small, and  $\widehat{MR}$  was usually close to MR on average.

## 6 Discussion

The miss rate (MR) represents a useful cautionary statistic, when interpreting the results of a comparative gene expression study. In situation where

the FDR of a list of significant genes is low, the miss rate of genes that were not quite called significant, can be quite high. The same information used to estimate the FDR can be used to estimate the miss rate. We suggest that MR be routinely reported along with the FDR and the local false discovery rate, in gene expression studies.

**Acknowledgments:** Tibshirani was partially supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183.

## References

- Benjamini, Y. & Hochberg, Y. (1985), ‘Controlling the false discovery rate: a practical and powerful approach to multiple testing’, *J. Royal. Stat. Soc. B.* **85**, 289–300.
- Dudoit, S., Shaffer, J. P. & Boldrick, J. C. (2003), ‘Multiple hypothesis testing in microarray experiments’, *Statistical Science* **18**, 71–103.
- Efron, B. (2004), ‘Large-scale simultaneous hypothesis testing: The choice of a null hypothesis’, *J. Amer. Statist. Assoc* **99**, 96–104.
- Efron, B. & Tibshirani, R. (2002), ‘Microarrays, empirical bayes methods, and false discovery rates’, *Gen. Epi.* .

- Efron, B., Tibshirani, R., Storey, J. & Tusher, V. (2001), ‘Empirical bayes analysis of a microarray experiment’, *J. Amer. Statist. Assoc.* pp. 1151–1160.
- Genovese, C. & Wasserman, L. (2002), ‘Operating characteristics and extensions of the FDR procedure’, *J. Roy. Stat. Soc., Ser. B* **64**, 499–517.
- Genovese, C. & Wasserman, L. (2003), A stochastic process approach to false discovery rates, Technical report, Carnegie Mellon University.
- Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. & Lander, E. (1999), ‘Molecular classification of cancer: class discovery and class prediction by gene expression monitoring’, *Science* **286**, 531–536.
- Storey, J. (2002a), ‘A direct approach to false discovery rates’, *Journal of the Royal Statistical Society B*. pp. 479–498.
- Storey, J. D. (2002b), ‘A direct approach to false discovery rates’, *J. Roy. Stat. Soc., Ser. B* **64**, 479–498.
- Storey, J. D., Taylor, J. E. & Siegmund, D. O. (2004), ‘Strong control, conservative point estimation, and simultaneous conservative consistency

of false discovery rates: A unified approach.', *J. Roy. Stat. Soc., Ser. B* **66**, 187–205.

Storey, J. D. & Tibshirani, R. (2001), Estimating false discovery rates under dependence, with applications to DNA microarrays, Technical Report 2001-28, Stanford University.

Storey, J. & Tibshirani, R. (2003), 'Statistical significance for genomewide studies', *Proc. Natl. Acad. Sci.* **100**, 9440–5.

Tusher, V., Tibshirani, R. & Chu, G. (2001), 'Significance analysis of microarrays applied to transcriptional responses to ionizing radiation', *Proc. Natl. Acad. Sci. USA.* **98**, 5116–5121.