False Discovery Rates (FDRs)

Jim McNicol
BioSS
Scottish Crop Research Institute

1. Motivating Example, Email

Developing a Filter to detect SPAM email.

Based on 100,000 spam (Alert) emails and 100,000 genuine (OK) emails grabbed from internet.

Filters not perfect. Best we can do is:
correctly identify 80,000 of Alerts and 95,000 of OKs

For a new email we have two hypotheses,
Ho: email status is OK
Ha: email status is Alert

The Filter is a test.
It examines the text and Decides whether the email status is OK (Do) or Alert (Da).

The Filter has been constructed so that we expect

Pr (Da|Ho) = 5,000/100,000 or 5% (Significance)
Pr (Da|Ha) = 80,000/100,000 or 80% (Power)

These are the two error rates associated with the test.

Applying the Filter

1000 Academic emails:
Filter identified 100 Alerts.
All 1000 emails were later examined individually with the following results:

		Decision	
		Do	Da
Truth	Ho:OK	850	20
	Ha:Alert	50	80

Among the 100 Alerts 20 were False Positives.

Pr (Ho|Da) = 20/100 or 20%.

This is the False Discovery Rate.

10,000 Commercial emails:
Filter identified 1,000 Alerts.
All 10,000 emails were later examined individually with the following results:

		Decision	
		Do	Da
Truth	Ho	8500	100
	Ha	500	900

Among the 1000 Alerts 100 were False (Positives).

Pr (Ho|Da) = 100/1000 or 10%.

This is the False Discovery Rate.

2. Reminder of Statistical Hypothesis Testing

Hypotheses

In simple statistical hypothesis testing there are two hypotheses:

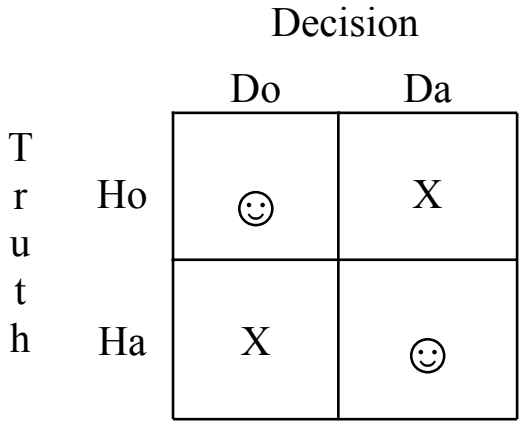
Ho, the null hypothesis
and
Ha, the alternative hypothesis.

We look to the data to indicate which of these hypotheses is more likely to be true.

On the basis of the data values a decision is made to accept *Ho* as true, or *Ha* as true. These decisions we refer to as *Do* and *Da*.

The decision we make can be correct or wrong.

The Figure below summarises the situation.



Error Probabilities

There are two error probabilities associated with this decision making process:

Firstly there is the probability of deciding the alternative hypothesis is true when in fact the null is true.

This is written $\Pr(Da \mid Ho)$. Then there is the probability of deciding the null hypothesis is true when in fact the alternative is true. This is written $\Pr(Do \mid Ha)$.

Simple Statistical Significance Tests

A formal statistical test is carried out, assuming *Ho* to be true. This test produces a significance value, *p.*

On the basis of the significance value, *p*, of the test statistic, a decision is made to accept either *Ho* or *Ha*.

Small values of *p* indicate strong evidence in the data away from *Ho* towards *Ha* and we accept *Ha* when *p* is less than some pre-determined *critical* value of p, p_{crit}.

Link between Error Probabilities and Significance Tests

In fact p_{crit} is $\Pr(Da \mid Ho)$.

Often p_{crit} is chosen so that $\Pr(Da \mid Ho)$ is 0.05.

That is, the probability of rejecting the null hypothesis, when the null is true, is 5%.

Once a value for $\Pr(Da \mid Ho)$ is chosen, then $\Pr(Do \mid Ha)$ can be derived. Even for simple problems this can be technically demanding.

$\Pr(Do \mid Ha)$ is often very much greater than $\Pr(Da \mid Ho)$ and can be as high as 0.50 when $\Pr(Da \mid Ho)$ is set at 0.05.

Generally the lower we set $\Pr(Da \mid Ho)$ the greater $\Pr(Do \mid Ha)$ becomes. In every hypothesis testing situation there is always a trade-off between these two error probabilities.

Traditional statistical tests are constructed to have a relatively low probability of mistakenly rejecting the null hypothesis, and a relatively high probability of mistakenly rejecting the alternative hypothesis.

Mistakenly rejecting *Ho* is usually considered a more serious error than mistakenly rejecting *Ha*.

"We assume *Ho* is true unless there is strong evidence in the data to suggest otherwise."

3. Multiple Testing and Error Rates

Consider the situation where we want to apply the same statistical test many times, for example a simple microarray experiment with 'infected tissue' and 'healthy tissue' arrays, where a t-test is used to compare expression levels for each gene.

Suppose *m* such statistical tests have been carried out, resulting in m significance values, p_i, $i=1\ldots m$. A cut-off value for *p* is selected, p_{crit}, and for all spots with *p* less than p_{crit} the decision is made to accept the alternative hypothesis, *Da*. For *p* greater than p_{crit} the null hypothesis is accepted, *Do*.

Let *R* denote the number of alternative hypotheses accepted in this way. Then *m-R* null hypotheses are accepted. Ideally, among the *R* decisions, *Da*, most will be correct. Inevitably some tests will indicate 'differential expression' when none exits. The proportion of the *R* decisions *Da* which are incorrect is the False Discovery Rate (FDR). Similarly some of the *m-R* decisions *Do* will be incorrect and this proportion we call the False Rejection Rate (FRR).

This situation of applying the same statistical test to many sets of data is referred to as **Multiple Testing**, as opposed to Multiple Comparisons, where all pairwise comparisons among levels of a single treatment are tested.

The Figure below summarises the multiple testing situation.

From the Figure the **False Discovery Rate, FDR = b/R.**

Note that in practice only m and R in the Figure are known. The False Discovery Rate, b*/R* and the False Rejection Rate, c*/(m-R)* must therefore be estimated. Further, the value of *R* depends on the value chosen for p_{crit} and so any estimates FDR and FRR also depend on p_{crit}.

		Decision: Do	Decision: Da	
Truth	Ho	a	b	
	Ha	c	d	
		m-R	R=b+d	m

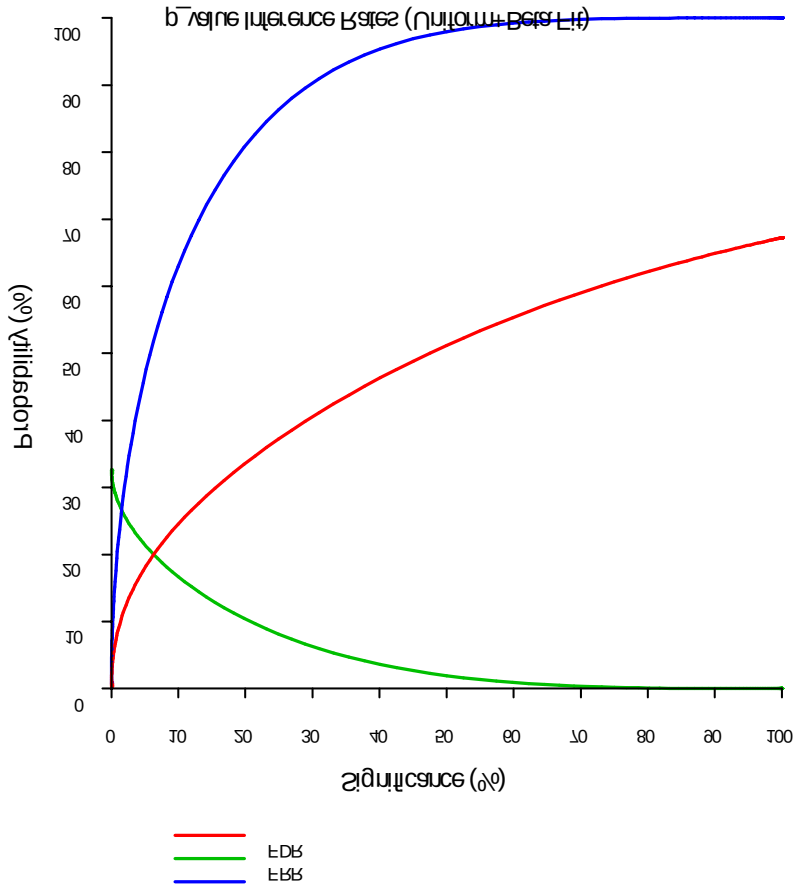
4. Microarray Example

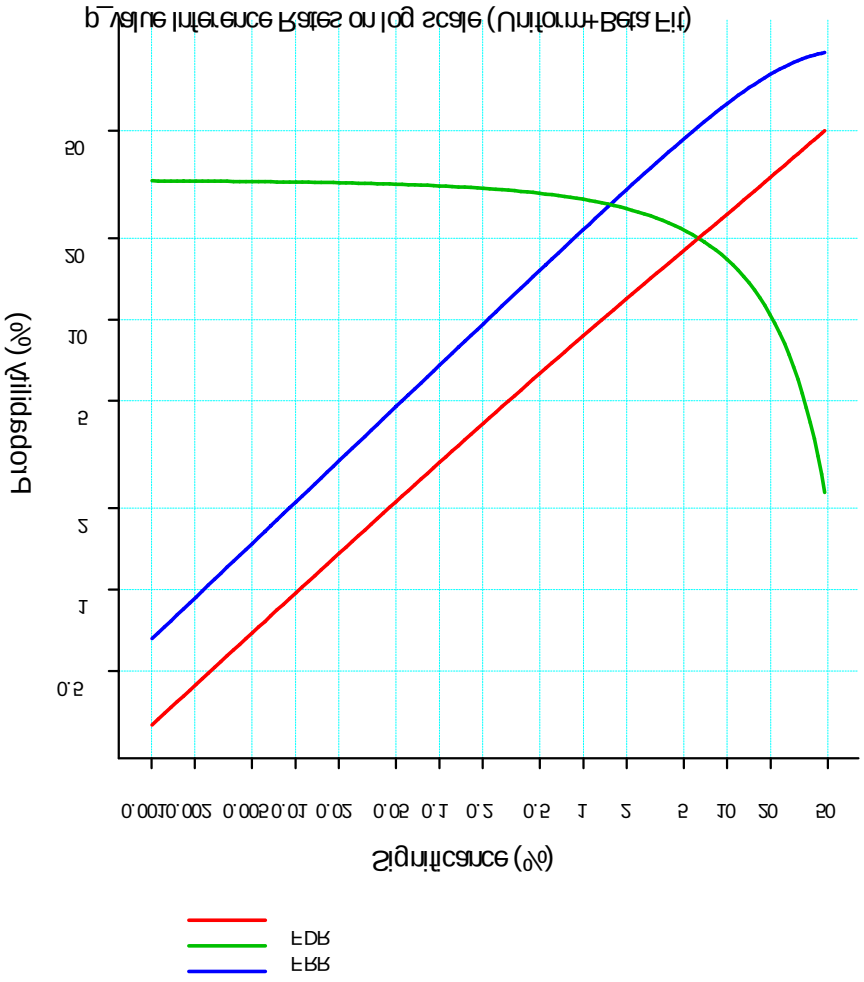
A microarray experiment by Hedenfalk et al. (2000) tried to identify to find genes that were differentially expressed between two mutation-positive tumours, BRCA1 and BRCA2. There were 7 and 8 arrays respectively of the two tissue types. (Storey and Tibshirani (2003) used a Behrens-Fisher test statistic t_i for each of 3170 genes and a significance level, p_i, was derived by permutations of array labels. The data along with the *p* values are available at http://research.nhgri.nih.gov/microarray/NEJM_Supplement/.) hedenfalk.xls shows the results of using a simple t-test to compare the tissue types. Corresponding to each p-value there is a FDR value. This is the estimate of the False Discovery rate if we used the corresponding p-value as p_{crit}

For example, for p=0.050000, (gene 2664) the estimated FDR is 0.177656, or 17.8%. Similarly for gene 2087 p=0.01 and the estimated FDR is 8%.

In practice these figures mean that if we use a significance level of 5% to identify genes which are differentially expressed between the two tissue types, we estimate that 17.7% of the identified genes will be false positives. On the other hand a significance filter of 1% will generate 8% false positives.

The figures below show how the FDR changes with p_{crit}





5. Exercises

The following exercise are based on the results in Tuomainen et al.(2006).

5.1 Thlaspi caerulescens is a good model species for studying metal hyperaccumulation in plants. Proteomic profiling (2D gels) was used to identify differences in protein intensities among 3 *T. caerulescens* accessions subjected to 5 different root exposures involving Zn and Cd. There were 3 reps of each of the 15 treatment combinations, giving 45 gels in total. Proteins were separated using two-dimensional electrophoresis and stained with SYPRO Orange. Intensity values and quality scores were obtained for each spot by using PDQuest software.

Exp.xls contains the (doctored) p-values and estimated FDR values for each of the 291 spots from an ANOVA to detect differences among 5 exposure treatments.

Plot the FDRs against the p-values and confirm that the plot has approximately the same shape as the Hedenfalk graph.

From the spreadsheet what is the FDR corresponding to a significance cut-off (p_{crit}) of 5%? Using a 5% significance cutoff how many spots are significant? How many of these would expect to be false positives?

Repeat with $p_{crit} = 0.01$.

5.2 The spreadsheet also contains the (fictitious) results of a follow-up confirmatory test. Test=1 indicates that differences in response for that spot were confirmed; test=0 indicates that differences were not found. Using the results of this test as 'truth' what was the true FDR when using $p_{crit} = 0.05$, or when using $p_{crit} = 0.01$?

6. Basis of FDR Estimation

There are many methods used to estimate the FDR directly or derive upper bounds for it. The different methods for direct estimation seem to produce generally similar results.

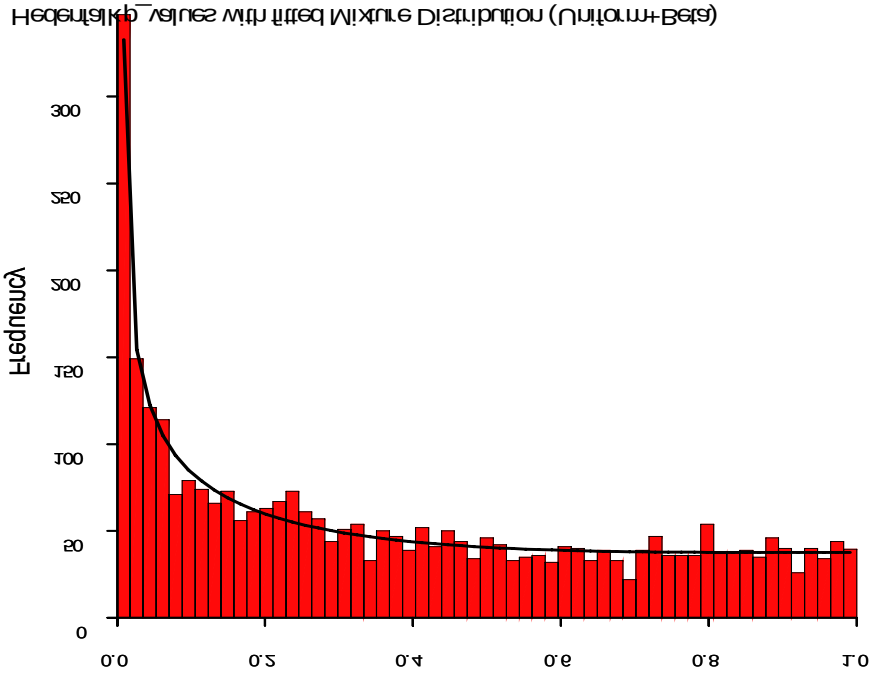
We describe a simple method based on 'mixture distributions'. This is the method used in Genstat.

When the same statistical test is applied to many different data sets, each application generates a p-value. For all the data sets where the null hypothesis, Ho, is true the p-values will be uniformly distributed in the region [0,1]. For the data sets where Ha is true the p-values will most often be small, close to zero, but with an unknown distribution. Combining these two sets of p-values gives a 'mixture distribution'

The parameters of this mixture distribution can be estimated and from these estimates it is straightforward to estimate the False Discovery Rate Pr(Ho|Da). The Ha p-values are modelled by a Beta or a Gamma distribution.

This method treats the p-values as a random sample from the mixture distribution and by implication assumes the p-values are all statistically independent. In practice this will not be the case but this lack of independence does not matter too much for many data sets, particularly when the sample size is large.

The figure below shows the Hedenfalk p-values and their fitted Mixture distribution. The histogram of p-values is, in itself, very useful diagnostic tool. Flat histograms indicate there are very few significant results. The Hedenfalk histogram is a common shape. Histograms with relatively high counts to the right (many values close to 1) suggest something 'wrong' with the data, either that the assumptions underlying the statistical test are invalid or that the p-values are strongly dependent. In either case further investigation is needed.



7. Genstat Procedure for estimating FDR

The Genstat code to derive the FDR from the Hedenfalk p-values is

```
fdrmixture [distribution=beta] probabilities=p_value; FDR=heden_fdr
```

This uses the p-values as input and generates FDRs corresponding to each p-value. For example, the first gene has an associated p-value of 0.012233. If 0.012233 was used as the significance cut-off, the FDR would be 0.089529.

The procedure automatically generates a histogram of the p-values and we strongly recommend using this histogram both as an aid to understanding your data and as a diagnostic tool.

Occasionally the estimation procedure 'fails to converge'. The three common solutions are

- Allow more iterations in the numerical process
- Switch to a Gamma distribution for modeling the Ha p-values
- Specify initial values for the three parameters, φ, A and B. φ estimates the proportion of Ho values. When this proportion is not close to the default of 0.9 there will be convergence problems. A and B are the parameters of the Beta or Gamma distribution.

```
fdrmixture [distribution=gamma; maxcycle=100; initial=!(0.8,0.2,2)]\
probabilities=p_value; FDR=heden_fdr
```

This Genstat procedure also estimates the False Rejection Rate, FRR, ie the probability of rejecting a gene that is a true positive, Pr(Ha|Do).

```
fdrmixture [distribution=beta] \
```

```
probabilities=p_value; FDR=heden_fdr; FRR=heden_frr.
```

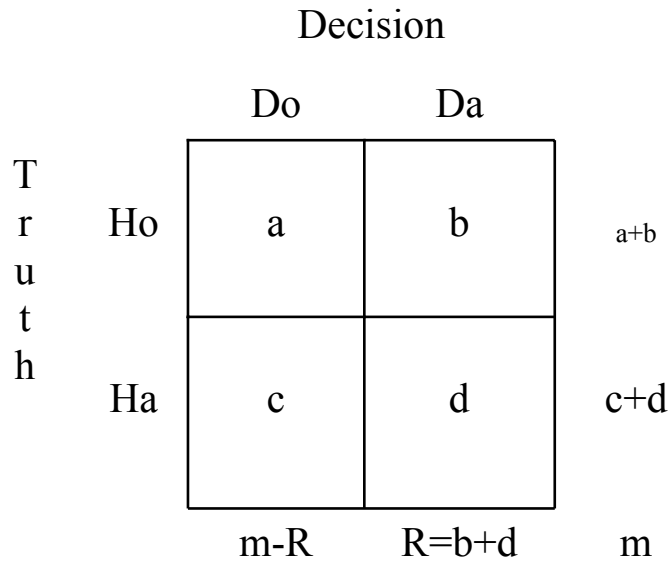
8. Exercises using the Genstat Procedure

8.1 Estimate the Hedenfalk FDRs using a Gamma rather than Beta mixture and confirm that this makes little difference to the FDR estimates. Note that this is not true for all data sets.

8.2 An experiment was conducted to assess potato metabolic profile response to two factors A and B. 283 metabolic compounds were measured and the significance values from an ANOVA to detect differences in response to the levels of factor A, factor B and the A.B interaction are given in the file 'twofactors,xls'.
By comparing the histograms of p-values for A, B and A.B what can you deduce about the FDRs for A, B and A.B?
Use the Genstat procedure to derive FDRs for A, B and A.B
What are the FDRs for A and B when you select metabolites on a basis of 5% significance?

9. Related Jargon

In the context of deciding which of two hypotheses is correct there is a range of confusing jargon. Here is a summary.



$$\Pr(Da|Ho) = b/(a+b) = \text{Type I Error} = \alpha = \text{False Positive Rate}$$

$$\Pr(Do|Ha) = c/(c+d) = \text{Type II Error} = \beta = \text{False Negative Rate}$$

$$\Pr(Do|Ho) = a/(a+b) = \text{Specificity} = 1-\alpha = \text{True Negative Rate}$$

$$\Pr(Da|Ha) = d/(c+d) = \text{Senstivity} = 1-\beta = \text{True Positive Rate}$$

All of these terms describe theoretical properties of the statistical testing procedure.

A Receiver Operator Characteristic Curve (ROC Curve) is a plot showing how $1-\beta$ (sensitivity) changes as a function of α (1-specificity).

$$\Pr(Ho|Da) = b/R = \text{False Discovery Rate (FDR)}$$

$$\Pr(Ha|Do) = c/(m-R) = \text{False Rejection Rate.}$$

These probabilities depend on both the properties of the statistical test and the data to which they are applied. They can be estimated only when estimates of Pr (Ho) and Pr (Ha), (a+b)/m and (c+d)/m respectively, are available for that data set.

10. References

Allison,D.,B., Gadbury,G.,L., Heo,M., Fernández,J.,R., Lee,C., Prolla,T.,A., Weindruch,R. (2002) A mixture model approach to the analysis of microarray gene expression data. *Computatonal Statistics and Data Analysis*, **39**, 1-20.

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner,M., Simon,R., Meltzer,P., Guterson,B., Esteller,M., Kallioniemi,O.P. *et al*. (2001) Gene-expression profiles in hereditary breast cancer. *N Engl J Med,* **22**, 539-548.

Storey, J.D., and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci*, **100**, 9940-9445.

Tuomainen M.H., Nunan N, Lehesranta S.J., Tervahauta A.I., Hassinen V.H., Schat H., Koistinen K.M., Auriola S., McNicol J., Kärenlampi S.O. (2006) Multivariate analysis of protein profiles of metal hyperaccumulator Thlaspi caerulescens accessions. Proteomics, **6**, 3696-3706