



# On the Statistical Comparison of Inductive Learning Methods

A. Feelders<sup>†</sup> and W. Verkooijen<sup>††</sup>

University of Twente<sup>†</sup>  
Department of Computer Science  
P.O.Box 217, 7500 AE Enschede  
The Netherlands  
feelders@cs.utwente.nl

Tilburg University<sup>††</sup>  
Department of Economics  
P.O. Box 90153, 5000 LE Tilburg  
The Netherlands  
W.J.H.Verkooijen@kub.nl

**ABSTRACT** Experimental comparisons between statistical and machine learning methods appear with increasing frequency in the literature. However, there does not seem to be a consensus on how such a comparison is performed in a methodologically sound way. Especially the effect of testing multiple hypotheses on the probability of producing a "false alarm" is often ignored.

We transfer multiple comparison procedures from the statistical literature to the type of study discussed in this paper. These testing procedures take the number of tests performed into account, thereby controlling the probability of generating "false alarms". The multiple comparison procedures selected are illustrated on well-known regression and classification data sets.

## 26.1 Introduction

Recent interactions between the statistical and artificial intelligence communities (see e.g. [Han93, CO94]), have led to many studies that compare the performance of empirical statistical and machine learning methods on real-life data sets; examples are the StatLog project ([MST94]), the Santa Fe Time Series Competition ([WG94]), and comparisons reported in numerous journal articles (e.g., [KWR93, RABCK93, WHR90, TAF91, TK92, FG93]).

We observe that there is no consensus in the research community on how such a comparative study is performed in a methodologically sound way.

The ranking of  $k$  preselected methods is usually performed by training (estimating in statistical terminology) the methods on a single data set, and estimating their respective mean prediction errors (MPE) from a hold-out sample. The methods are subsequently ranked according to their estimated MPEs. Some studies use, in our view appropriately, statistical significance testing in order to make this ranking. However, the effect that comparing multiple methods has on the probability of generating a "false alarm" (claiming

---

<sup>†</sup>*Learning from Data: AI and Statistics V.* Edited by D. Fisher and H.-J. Lenz. ©1996 Springer-Verlag.

that one method is better than another, when in fact it is not) is to our knowledge ignored in the literature.

The statistical analysis of comparative studies, method-ranking in particular, is addressed in this paper. Specifically, we address *methodological* issues of studies in which the performance of several regression and classification methods are compared on real-life data sets.

## 26.2 Ranking Methods by Significance Testing

The ranking of methods by simply ordering them by the point estimates of their prediction errors should be extended by statistical significance testing. Appropriate tests are those concerned with the difference between means (regression) and proportions (classification). The standard  $t$ -test for testing the difference between two sample means  $\bar{Y}_1$  and  $\bar{Y}_2$  which come from *independent* normal distributed populations, leads to the following confidence interval for the difference

$$\theta_1 - \theta_2 \in [(\bar{Y}_1 - \bar{Y}_2) \pm t_{(\alpha/2, \nu)} \hat{\sigma}_{\text{diff}}] \quad (26.1)$$

where  $\hat{\sigma}_{\text{diff}}$  equals  $\sqrt{\hat{\sigma}_{\bar{Y}_1}^2 + \hat{\sigma}_{\bar{Y}_2}^2}$ . Here,  $\alpha$  denotes the probability of making a Type I error, i.e. of claiming that the population means are different, when in fact they are not. In the standard comparative experiment, however, the MPEs are all estimated from the *same* test sample, which makes them highly correlated. Therefore, a *paired sample t*-test should be used instead. The dependence within the pairs only changes the standard error of the difference  $\hat{\sigma}_{\text{diff}}$ , which now becomes

$$\hat{\sigma}_{\text{diff}} = \sqrt{\hat{\sigma}_{\bar{Y}_1}^2 + \hat{\sigma}_{\bar{Y}_2}^2 - 2 \text{cov}(\bar{Y}_1, \bar{Y}_2)} \quad (26.2)$$

When the variables are positively correlated, the covariance will have a positive value and thus the variance and standard error of a difference between means will be *less* for matched than for unmatched samples. Consequently, the confidence intervals become smaller (given the same  $\alpha$  value), which results in more powerful tests. In conclusion, neglecting the dependence between the samples generally results in tests that are too conservative.

Often the estimated MPEs of more than two, say  $k$ , methods are being compared. The first idea that comes to mind is to test each possible difference by a paired  $t$ -test with probability of Type I error of size  $\alpha$ . The problem with this approach is that the probability of making at least one Type I error over the whole family of  $t$ -tests (one test per pair of methods being compared) exceeds  $\alpha$  by an amount that increases with the number of tests made. For  $J$  *statistically independent* tests the probability of making at least one Type I error, better known as the *familywise error rate* (FWE), is  $1 - (1 - \alpha)^J$ . When  $J$  is large, say 20, this can be a large probability; for  $\alpha = 0.05$  there is a probability of 0.64 for one or more Type I errors. This means that the probability that we incorrectly claim to have found one or more significant differences is 64%. Such an incorrect claim is called a "false alarm". When the tests are statistically dependent of each other, such as pairwise difference tests, the FWE becomes even larger. Thus, when enough pairwise tests are performed one will with high probability find one or more "significant" differences, i.e. false alarms. This

Observations	Functions						Total
	$f_1$	$f_2$	...	$f_i$	...	$f_k$	
1	$Y_{11}$	$Y_{12}$	...	$Y_{1i}$	...	$Y_{1k}$	$Y_{1.}$
2	$Y_{21}$	$Y_{22}$	...	$Y_{2i}$	...	$Y_{2k}$	$Y_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$j$	$Y_{j1}$	$Y_{j2}$	...	$Y_{ji}$	...	$Y_{jk}$	$Y_{j.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$n$	$Y_{n1}$	$Y_{n2}$	...	$Y_{ni}$	...	$Y_{nk}$	$Y_{n.}$
Total	$\overline{Y}_{.1}$	$\overline{Y}_{.2}$	...	$\overline{Y}_{.i}$	...	$\overline{Y}_{.k}$	
Means	$\overline{Y}_{.1}$	$\overline{Y}_{.2}$	...	$\overline{Y}_{.i}$	...	$\overline{Y}_{.k}$	

TABLE 26.1. One-way repeated measures lay-out.

problem is known as the *multiplicity effect* or *selection effect*. Statistical procedures have been designed to take into account and properly control for the multiplicity effect; they are called *multiple comparison procedures*.

A crude approach to deal with the multiplicity effect is the Bonferroni method <sup>2</sup>, which rejects the pairwise null hypothesis  $\theta_i - \theta_{i'} = 0$  when the  $p$ -value is less than  $\alpha/J$ , where  $\alpha$  is the preset FWE level and  $J$  is the number of tests. This method neglects the dependency between the pairwise difference tests, and it further assumes normality of the data.

There are many alternative tests, ranging from slight adjustments of the Bonferroni method to very sophisticated techniques. The existing comparison procedures can roughly be categorised as analytical [HT87] or resampling based [WY93]. The former approach requires certain distributional assumptions of the underlying statistical model and typically uses table lookup to make a probability statement. The latter approach generates empirical distributions of the relevant statistics by resampling from the data set at hand, thereby removing the risk of making faulty statements due to unsatisfied assumptions. Evidently, the resampling approach involves much more computation than the analytical approach. In this paper we restrict ourselves to the analytical approach.

The characteristics of a particular experimental design often prescribe adjustments to general tests for differences or make special purpose tests necessary. The experimental design that captures the subject of this study is the *one-way repeated measures design*, which is displayed in Table 26.1. In such designs, blocks consisting of a random sample of, say,  $n$  experimental units drawn from a large population constitute the random factor. Each unit is measured under  $k$  different conditions. The conditions of measurements are fixed in advance, and constitute the treatment factor. In the terminology of this study, experimental units correspond to the observations from the test set, and the treatment factor corresponds with the regression or classification model type.

---

<sup>2</sup>This method originally due to R.A. Fisher ([Fis35]) is popularly known as the Bonferroni method since it uses the Bonferroni inequality (which says that the probability of a union of events is less than the sum of the individual event probabilities).

### 26.3 Pairwise Comparisons for Regression

The general setting of this section is Table 1 (the one-way repeated measures design with  $k$  different prediction models that predict the observations from the *same* random test set of size  $n$ ). The deviation of the predicted value from the true value is assumed to be measured as squared error, but any other error measure could be inserted equally well (e.g. absolute error). When the observations are not randomly drawn from a population, but result from a (highly) autocorrelated time series, the subsequent approach seems not to be justified. Diebold and Mariano [DM94] discuss the comparison of predictive accuracy of two time series models; they leave the multiple comparison problem for further research.

Let  $\mathbf{Y}_j = (Y_{j1}, Y_{j2}, \dots, Y_{jk})$  denote the vector of prediction errors for the  $j$ th observation ( $1 \leq j \leq n$ ). The following model is assumed:

$$\mathbf{Y}_j = \mathbf{M}_j + \mathbf{E}_j \quad (1 \leq j \leq n) \quad (26.3)$$

where all the  $\mathbf{M}_j = (M_{j1}, M_{j2}, \dots, M_{jk})$  and  $\mathbf{E}_j = (E_{j1}, E_{j2}, \dots, E_{jk})$  are distributed independently of each other as  $k$ -variate normal vectors, the former with mean vector  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$  (the vector of model effects) and variance-covariance matrix  $\boldsymbol{\Sigma}_0$ , and the latter with mean vector  $\mathbf{0}$  and variance-covariance matrix  $\sigma^2 \mathbf{I}$ . Thus the  $\mathbf{Y}_j$ 's are independent and identically distributed (i.i.d.)  $N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$  random vectors where  $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 + \sigma^2 \mathbf{I}$ .

Exact procedures for making pairwise comparisons among the  $\theta_i$ 's can be constructed if we impose special restrictions on the form of  $\boldsymbol{\Sigma}$ . The least restrictive of such models is the *spherical model*, which assumes that all pairwise differences of the sample means of the regression models have the same variance (for more details see [HT87, CH90, WBM91, Hay88]). In practice, however, this assumption will rarely be satisfied [HT87, Hay88].

Therefore Hochberg and Tamhane [HT87, page 215] propose a test, in case one is unsure about the sphericity assumption. They propose the following approximate  $100(1 - \alpha)\%$  simultaneous confidence intervals for the pairwise differences  $\theta_i - \theta_{i'}$ :

$$\theta_i - \theta_{i'} \in \left[ \bar{Y}_{.i} - \bar{Y}_{.i'} \pm |M|_{k^*, n-1}^{(\alpha)} \sqrt{\frac{S_{ii} + S_{i'i'} - 2S_{ii'}}{n}} \right] \quad (1 \leq i < i' \leq k) \quad (26.4)$$

where  $|M|_{k^*, n-1}^{(\alpha)}$  is the upper  $\alpha$  point of the Studentized maximum modulus distribution (see [HT87, Table 6]) with parameter  $k^* = k(k-1)/2$  and degrees of freedom  $n-1$ ; and where  $S_{ii'}$  is the estimated (co)variance between  $Y_i$  and  $Y_{i'}$ :

$$S_{ii'} = \frac{\sum_{j=1}^n (Y_{ji} - \bar{Y}_{.i})(Y_{ji'} - \bar{Y}_{.i'})}{n-1} \quad (1 \leq i, i' \leq k) \quad (26.5)$$

The Studentized maximum modulus distribution is defined as the distribution of

$$\max_{1 \leq i \leq k} |T_i|$$

where  $|T_i|$  is the modulus of the  $k$ -variate  $t$ -distribution with  $\nu$  degrees of freedom and common correlation of zero.

We use an empirical analysis of the *Boston housing data*<sup>3</sup> to illustrate this procedure. We compare the performance of four regression models, designated  $f_1$  through  $f_4$ . The first model  $f_1$  is a linear model trained with OLS; models  $f_2$  through  $f_4$  are feed-forward neural networks with two, four, and six hidden units respectively. The data set is split into two parts: the first part (400) is used to estimate the parameters of the model; the second part (106) is used to measure the model's performance. The observed average squared prediction error  $\bar{Y}_{.i}$  for method  $f_i$  ( $i = 1, 2, 3, 4$ ) are  $\bar{Y}_{.1} = 6.38\text{e-}3$ ;  $\bar{Y}_{.2} = 3.49\text{e-}3$ ;  $\bar{Y}_{.3} = 2.94\text{e-}3$ ; and  $\bar{Y}_{.4} = 3.00\text{e-}3$ .

	$f_1$	$f_2$	$f_3$	$f_4$
$f_1$	1.68e-04	3.76e-05	3.46e-05	4.02e-05
$f_2$	–	3.01e-05	1.54e-05	1.27e-05
$f_3$	–	–	2.24e-05	2.67e-05
$f_4$	–	–	–	4.15e-05

TABLE 26.2. The  $S_{ii'}$  matrix for the Boston housing case.

	$\theta_2$	$\theta_3$	$\theta_4$
$\theta_1$	[5.9e-4 , 5.2e-3]	[1.1e-3 , 5.7e-3]	[1.1e-3 , 5.8e-3]
$\theta_2$	–	[-4.3e-4 , 1.5e-3]	[-9.2e-4 , 1.9e-3]
$\theta_3$	–	–	[-7.2e-4 , 6.2e-3]

TABLE 26.3. All pairwise confidence intervals.

Suppose that it is of interest to make all pairwise comparisons among the four methods, using a Type I familywise error rate  $\alpha = 0.10$ . In Table 26.2 the  $S_{ii'}$  values are displayed in the cells, calculated according to (26.5). Obviously, we have  $|M|_{6,105}^{(0.1)}$ , which equals 2.135 ([HT87, Table 6]), and we construct the confidence intervals for the pairwise differences  $\theta_i - \theta_{i'}$  according to (26.4). Table 26.3 shows that the linear model performs significantly worse than the three neural network models: the confidence intervals cover positive values only. Among the neural network models no significant differences can be observed; there is no statistical evidence for preferring neural network models with more than two hidden units.

## 26.4 Pairwise Comparisons for Classification

In this section we discuss significance testing for the comparison of two or more *classification* methods. Again we notice that it is not appropriate to use a standard test based on the assumption of independent samples. Instead, we use, as suggested by Ripley ([Rip93]), McNemar's test ([MM77]), when only two classification methods are compared. This test is normally used to test for differences between proportions in paired sample designs. The comparisons performed in this section should not be considered as serious evaluations of the methods involved; they are purely illustrative.

<sup>3</sup>This data set can be obtained by anonymous ftp from `lib.stat.cmu.edu` with user `statlib`

	$I_{nn}$	$C_{nn}$	Total
$I_{lda}$	61	23	84
$C_{lda}$	32	268	300
Total	93	291	384

TABLE 26.4. Incorrect and Correct classifications of lda and nn

### 26.4.1 Testing a single hypothesis

A hypothesis test involving the application of linear discriminant analysis and a feed-forward neural network to the *diabetes* data set<sup>4</sup> illustrates the use of McNemar's test. The 768 observations in this dataset were divided in a training set and a test set of 384 observations each.

Table 26.4 summarizes the result of using the linear discriminant function (lda) and a neural network (nn) estimated on the training set to classify the observations in the test set. The cells  $(C_{lda}, C_{nn})$  and  $(I_{lda}, I_{nn})$  contain the number of cases classified correctly and incorrectly by both the linear discriminant function and the neural network respectively. Since we want to test

$$H_0 : \text{MPE}_{lda} = \text{MPE}_{nn} \quad \text{against} \quad H_a : \text{MPE}_{lda} \neq \text{MPE}_{nn},$$

only the cells  $(I_{lda}, C_{nn})$  and  $(C_{lda}, I_{nn})$  of this table are of interest. Here MPE is defined as the proportion of misclassifications made on the population of interest. When an observation falling in the  $(I_{lda}, C_{nn})$  cell of this table is defined as a success, then the number of successes is binomially distributed with  $n = (I_{lda}, C_{nn}) + (C_{lda}, I_{nn}) = 55$  and  $\pi = 0.5$ , under the null hypothesis. Observing 23 successes out of 55 trials (with  $\pi = 0.5$ ) has a  $p$ -value 0.28; this  $p$ -value was calculated using an exact binomial test. According to any conventional significance level, we should conclude that  $H_0$  cannot be rejected. Whether or not conventional significance levels are appropriate in this kind of hypothesis test is debatable. It is not clear that a Type I error is the most severe error one can make.

### 26.4.2 Testing Multiple Hypotheses

We will now consider the case where  $k > 2$  classification functions are compared. In this comparison we use the same training and test set as in the above example. We performed a comparative study, including linear discriminant analysis ( $f_1$ ), quadratic discriminant analysis ( $f_2$ ), a classification tree ( $f_3$ ), and two feed-forward neural networks ( $f_4$  and  $f_5$  respectively) which differ only in the value of the weight decay parameter used. All pairwise comparisons are performed, which amounts to a total of  $k^* = 10$  pairwise comparisons. Table 26.1 presents the general lay-out of a study that compares  $k$  classification functions. In this matrix  $Y_{ji}$  has the value zero if  $f_i$  classifies observation  $j$  correctly, and one otherwise.

For our comparative study we have:  $n = 384$ ,  $k = 5$ ,  $\bar{Y}_{.1} = 0.219$ ,  $\bar{Y}_{.2} = 0.232$ ,  $\bar{Y}_{.3} = 0.31$ ,  $\bar{Y}_{.4} = 0.211$ ,  $\bar{Y}_{.5} = 0.229$ .

---

<sup>4</sup>This data set can be obtained by anonymous ftp from [ics.uci.edu](http://ics.uci.edu) in the directory `pub/machine-learning-databases`

	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$
$\theta_1$	[-0.071,0.045]	[-0.149,-0.033]	[-0.05,0.066]	[-0.068,0.048]
$\theta_2$	—	[-0.136,-0.02]	[-0.037,0.079]	[-0.055,0.061]
$\theta_3$	—	—	[0.041,0.157]	[0.023,0.139]
$\theta_4$	—	—	—	[-0.076,0.04]

TABLE 26.5. 95% confidence intervals for all pairwise differences.

The pooled variance of any pairwise difference  $\bar{Y}_{.i} - \bar{Y}_{.i'}$  for this design can be written as ([MM77], p. 180)

$$\hat{\sigma}_{\text{diff}}^2 = \frac{2(k \sum_{j=1}^n Y_{j.} - \sum_{j=1}^n Y_{j.}^2)}{n^2 k(k-1)}$$

We can now construct  $100(1 - \alpha)\%$  simultaneous confidence intervals for all pairwise differences  $\theta_i - \theta_{i'}$  as follows

$$\theta_i - \theta_{i'} \in \left[ \bar{Y}_{.i} - \bar{Y}_{.i'} \pm Z_{k^*, 1-\alpha/2}^\nu \hat{\sigma}_{\text{diff}} \right] \quad (1 \leq i < i' \leq k)$$

where  $\theta_i$  denotes the population proportion of incorrect classifications of  $f_i$ , and  $\nu = n - 1$  denotes degrees of freedom. The distribution of  $Z$  is based on the Student  $t$  distribution, adjusted for the number of comparisons  $k^*$  involved ([Dun61]). Tables for this statistic can be found in ([MM77],[Dun61]). As one would expect, the value of  $Z_{k^*, 1-\alpha/2}^\nu$  increases with the number of comparisons  $k^*$ , leading to wider confidence intervals.

Table 26.5 provides 95% confidence intervals for all pairwise differences. We used  $Z_{10:0.975}^\infty = 2.81$  and  $\hat{\sigma}_{\text{diff}} = 0.0205$ , leading to confidence intervals that are  $2 \times (2.81 \times 0.0205) = 0.116$  wide. If the interval of  $\theta_i - \theta_{i'}$  contains zero, then there is no significant evidence that classification functions  $f_i$  and  $f_{i'}$  differ in their true prediction error. From Table 26.5 we conclude that  $f_3$  (the classification tree) performs significantly worse than *all* other functions; among these other functions no significant difference has been found.

## 26.5 Conclusion and future research

In this paper we proposed statistical procedures to perform studies which compare the predictive accuracy of several regression or classification functions. These procedures explicitly take into account the fact that multiple comparisons are made, and control for the probability of giving a "false alarm". We have attempted to show the relevance of multiple comparison procedures to a type of study that we encountered frequently in the recent Artificial Intelligence (AI) and Machine Learning literature.

Although the general difficulties induced by the multiplicity effect and by the dependency among observations are easy to grasp, finding "the right" testing procedure is much more difficult. The literature on the subject is somewhat ambiguous, and requires a high level of statistical knowledge, which AI-researchers do not always possess.

We do not claim that the methods presented here are the best for the given purpose: they are *examples* of tests that can be used for the comparison of the prediction accuracy of different functions. We hope that in future research more attention will be given to this subject, and perhaps more appropriate methods will be found.

One interesting approach may be to expand the empirical analyses by incorporating resampling based multiple comparison techniques. This will enhance the practical relevance for the machine learning engineer, since the assumptions made by analytical approaches may not be satisfied in practical situations. With the computer power currently available, the computational effort required by resampling techniques does not seem to be an obstacle anymore.

## 26.6 REFERENCES

- [CH90] M.J. Crowder and D.J. Hand. *Analysis of repeated measures*. Chapman & Hall, London, 1990.
- [CO94] P. Cheeseman and R.W. Oldford, editors. *Selecting models from data: AI and statistics IV*. Lecture notes in statistics nr. 89. Springer-Verlag, New York, 1994.
- [DM94] F. Diebold and R. Mariano. Comparing predictive accuracy. Technical report, University of Pennsylvania, 1994.
- [Dun61] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56:52–64, 1961.
- [Fis35] R. A. Fisher. *The design of experiments*. Olivier & Boyd, Edinburgh, 1935.
- [FG93] D. Fletcher and E. Goss. Forecasting with neural networks: an application using bankruptcy data. *Information & Management*, 24:159–167, 1993.
- [Han93] D. J. Hand, editor. *Artificial intelligence frontiers in statistics: AI and statistics III*. Chapman & Hall, London, 1993.
- [Hay88] W. Hays. *Statistics*. Holt, Rinehart and Winston, Inc, Fort Worth, 1988.
- [HT87] Y. Hochberg and A. Tamhane. *Multiple comparison procedures*. Wiley & Sons, New York, 1987.
- [KWR93] J. Kim, H. Weistroffer, and R. Redmond. Expert systems for bond rating: a comparative analysis of statistical, rule-based and neural network systems. *Expert Systems*, 10:167–171, 1993.
- [MM77] L. Marascuilo and M. McSweeney. *Nonparametric and distribution-free methods for the social sciences*. Brooks/Cole Publishing Company, Monterey, CA, 1977.
- [MST94] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, editors. *Machine learning, neural and statistical classification*. Ellis Horwood, New York, 1994.
- [Pre95] Lutz Prechelt. A quantitative study of neural network learning algorithm evaluation practices. In *Proc. 4th Intl. Conf. on Artificial Neural Networks*, Cambridge, UK, June 26-28, 1995.

- [RABCK93] A. Refenes, M. Azema-Barac, L. Chen, and S. Karoussos. Currency exchange rate prediction and neural network design strategies. *Neural Computing & Applications*, 1:46–58, 1993.
- [Rip93] B.D. Ripley. Flexible non-linear approaches to classification. In V. Cherkassky, J.H. Friedman, and H. Wechsler, editors, *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*. Springer-Verlag, 1993.
- [TAF91] Z. Tang, C. de Almeida, and P. Fishwick. Time series forecasting using neural networks vs. box-jenkins methodology. *Simulation*, 57:303–310, 1991.
- [TK92] Kar Yan Tam and Melody Y. Kiang. Managerial applications of neural networks: the case of bank failure predictions. *Management science*, 38(7):926–947, 1992.
- [WBM91] B. Winer, D. Brown, and K. Michels. *Statistical principles in experimental design*. McGraw-Hill, New York, 1991.
- [WG94] A. Weigend and N. Gershenfield. *Time series prediction: forecasting the future and understanding the past*. Addison-Wesley, Reading, 1994.
- [WHR90] A. Weigend, A. Huberman, and D. Rumelhart. Predicting the future: a connectionist approach. *International Journal of Neural Systems*, 1(3):193–209, 1990.
- [WY93] P.H. Westfall and S.S. Young. *Resampling-Based Multiple Testing*. John Wiley & Sons, New York, 1993.