FDR and ROC: Similarities, Assumptions, and Decisions

1. Why FDR and ROC?

It is a privilege to have been asked to introduce this collection of papers appearing in *Statistica Sinica*. The papers fall into two topical areas, receiver operating characteristic (ROC) analyses (Cai and Dodd; Song and Zhou) and various flavors of false discovery rates (FDRs) in multiple hypothesis testing applications (Chi and Tan; Craiu and Sun; Ge, Sealfon and Speed; Guan, Wu and Zhao; Sarkar, Zhou and Ghosh). The common thread is "decision-making," and I congratulate the editors of *Statistica Sinica* for bringing this topic to the fore.

There are myriads of flavors of multiple hypothesis testing, with questions of which resampling method to use (see Ge et al. in this issue), and even more broadly, which type I error rate definition to use (strong, weak, familywise, false discovery rate or proportion; see Chi and Tan of this issue). These questions are interesting and remain active in current areas of research. However, to make our methods useful, we must move toward answering the simple question that every scientist has: Which method is best? In the case of hypothesis testing for gene expression data, for example, scientists want to know which genes are "interesting," and which are "not interesting," borrowing Efron's (2004) terminology. As in ROC analysis (Cai and Dodd; Song and Zhou), the answer should involve both types of decision errors (Craiu and Sun; Sarkar et al.). The logical next step is to consider losses resulting from both types of errors; Sarkar et al. provide a good step in this direction.

2. Specific Comparisons

In ROC analysis, there is a statistical measure T reflecting a true underlying state δ, with $\delta = 1$ often denoting "diseased" and $\delta = 0$ "not diseased." Goals are to choose a threshold c for making the classification $\delta = 1(0)$ when $T > c$ ($\leq c$), and to characterize the benefits of such a procedure. Hypothesis testing has a similar structure: there is a statistical measure T reflecting a true underlying state, δ, with $\delta = 0$ denoting "H_0 true" and $\delta = 1$ denoting "H_0 false." Goals are similar to those for ROC analysis.

Main differences are: (i) in ROC analysis, T is an observation-specific measure, for example, level of prostate-specific antigen (PSA) in a blood sample from a male patient, whereas in hypothesis testing, T is an aggregate across all observations in a data set, for example, the standardized difference between average PSA for treatment and control groups of male patients; (ii) in ROC analysis, δ is again an observation-specific measure, for example, existence or non-existence of prostate cancer in the male patient whose PSA was measured, whereas in hypothesis testing, δ is a parameter governing the production of the existing data, for example, an indicator of whether the "population average" PSA levels differ; and (iii) in ROC analysis the case $T \leq c$ is classified directly as "non-diseased," whereas in hypothesis testing it is classified indirectly as "fail to reject H_0".

In either case, there is a concern for both types of errors, although typically in hypothesis testing there is a stronger emphasis on type I errors, simply because they can be estimated and controlled tractably, while making fewer assumptions. Concern for type II errors drives research into methods with lower type II error rates while controlling type I errors. In contrast, ROC analysis has as its main measure area under the curve (AUC), which implicitly treats both error types symmetrically.

3. Assumptions

With ROC analysis, data are available where it is known precisely whether $\delta = 0$ or $\delta = 1$. Hypothesis testing often puts $\delta = I(\theta = 0)$ where θ is a difference between means such as $\theta = \mu_1 - \mu_2$, and it is not possible to determine δ since θ is unobservable. The emphasis on Type I errors occurs because $T \left| \left\{ \delta = 0 \right\} \right.$ typically has a known distribution under minimal assumptions, whereas the distribution of $T \left| \left\{ \delta = 1 \right\} \right.$ requires the assumption of a specific value or Bayesian prior for θ. Statisticians are often averse to making extra assumptions, since they limit applicability. However, with limited applicability comes greater potential utility, as methods are often much more useful when the assumptions happen to be true.

As a (frequentist) example, consider the multiple hypothesis testing application with two-sample, m-dimensional multivariate data, common in gene expression studies, $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{n_x} \overset{iid}{\sim} F_{\boldsymbol{X}}$; $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_{n_y} \overset{iid}{\sim} F_{\boldsymbol{Y}}$, with the $\boldsymbol{X}$'s independent of the $\boldsymbol{Y}$'s. Let $E(\boldsymbol{X}_1) = \left[\mu_1^{(\boldsymbol{X})} \ldots \mu_m^{(\boldsymbol{X})} \right]'$, $E(\mathbf{Y}_1) = \left[\mu_1^{(\boldsymbol{Y})} \ldots \mu_m^{(\boldsymbol{Y})} \right]'$ and $\theta_j = \mu_j^{(\boldsymbol{X})} - \mu_j^{(\boldsymbol{Y})}$, $j = 1,\ldots, m$.

Hypotheses are $H_{0j}:\theta_j = 0$; all of the papers in this issue devoted to FDR can be applied to this problem, either directly or indirectly.

Consider the related problem of testing $H_0 := \bigcap_{j=1}^{m} H_{0j}$. Typically $m>>n$ and standard multivariate methods are not appropriate. However, one can test H_0 by resampling

$$T = \max_{j=1,\ldots,m} T_j\,,$$

where T_j is the separate-variance t statistic

$$T_j = \frac{\left|\bar{X}_j - \bar{Y}_j\right|}{\left(s_{Xj}^2/n_x + s_{Yj}^2/n_y\right)^{1/2}}\,.$$

Troendle, Korn and McShane (2004) compared two types of bootstraps, a “within-sample” bootstrap, and a “pooled-sample” bootstrap, as well as permutation resampling, to estimate the null distribution of T. The “within-sample” method makes no assumptions on the distributions F_X and F_Y other than finite variances, the “pooled sample” assumes additionally that F_X and F_Y differ only in location, and the permutation method assumes that H_0 implies that the distribution of $\left\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{n_x},\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_{n_y}\right\}$ is exchangeable. The latter two models imply that the “subset pivotality” assumption holds (see Ge et al. in this issue and Westfall and Troendle (2008), for details on the subset pivotality assumption). Troendle et al.’s finding was that the type I error rate for the “within-sample” method was far from nominal levels for typical genomic applications with small n, large m, in the 0.001–0.006 range for a nominal α =0.05 test. Type I error rate was somewhat large for the pooled sample method, 0.077–0.101, and nearly exact for the permutation method. Because the within sample method had such low true Type I error rates, it suffered great power loss, 0.26 –0.56 compared to 0.83–0.86 for the pooled bootstrap and 0.79 for the permutation method.

The conclusion is well known ─ assumptions can help greatly. On the other hand, it is true that when the assumptions are badly violated, the pooled method can behave poorly (Westfall, (2003)).

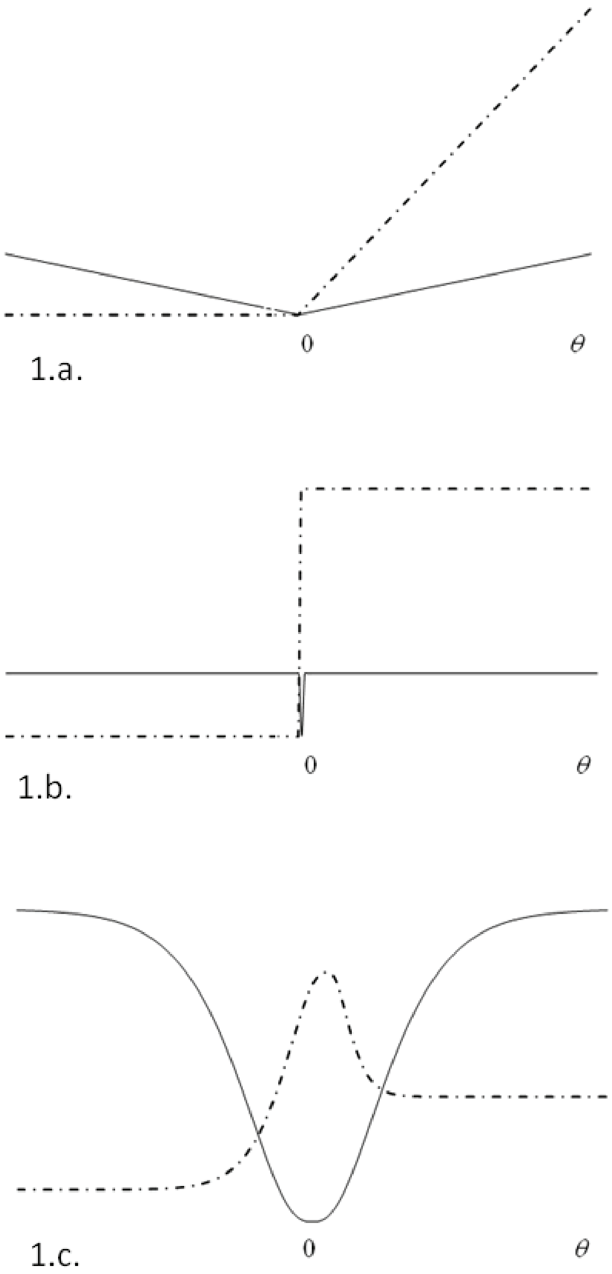

Figure 1: Three loss functions as indicated in 1.a., Waller-Duncan, 1.b., 0-1 Loss, and 1.c., specialized. Loss for claiming a negative θ is a dotted line, loss for claiming θ not scientifically different from 0 is solid.

4. Loss Functions and Decisions

Assumptions are unavoidable, and are in fact desirable for decision-making. For ROC analysis, Vickers (2008) discusses accuracy metrics such as error rates and AUC, and concludes that we need to go one step farther and assume specific losses (or equivalently, benefits, in Vickers' analysis), so that the resulting decision can be called "best." For hypothesis testing, Figure 1 shows examples of loss functions applicable to the problem of determining sign of θ, whether positive, negative, or not scientifically different from 0. Historically, these loss functions have been chosen for analytic tractability; examples include the Waller-Duncan loss functions shown in Figure 1.a (Waller and Duncan, (1969)), and the 0-1 loss functions shown in Figure 1.b. (Lewis and Thayer, (2004)). Tractability is not an issue with modern Bayesian methods, as one can simulate from the posterior distribution of θ, plug into the loss functions, and select the action with the minimum average loss; an example is given in Westfall, Tobias, Rom, Wolfinger, and Hochberg (1999, pp. 278-281). In the case of gene expression data, "loss to science" in terms of speed to discovery might take a more exotic form, such as in Figure 1.c. Identifying reasonable loss functions will require much discussion and consensus among scientists and statisticians.

When decision rules are framed in terms of losses, insights are gained. For example, familywise error rate (FWER) controlling methods that are generally considered inappropriate for large-scale multiple testing are optimal for certain kinds of loss functions (Lu and Westfall (2008)), and the usual multiple testing methods are inadmissible relative to a recent proposal of Efron (2004) as $n \to \infty$. But ultimately, the goal will be to advance science, again, by answering the question "Which method is best?" as directly as possible, using strong but defensible assumptions concerning losses and priors.

5. Conclusion

It is my pleasure to comment on this special issue with the unifying theme "decision-making." Hopefully it will spur continued research in how to provide scientists with the best tools for making decisions using their data.

References

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99**, 96-104.

Lewis, C. and Thayer, D. T. (2004). A loss function related to the FDR for random effects multiple comparisons. *J. Stat. Plan. Infer.* **125**, 49-58.

Lu, Y. and Westfall, P. H. (2008). Is Bonferroni admissible for large *m*? *Amer. J. Math. Management Sci.* (To appear).

Troendle, J. F., Korn, E. L. and McShane L. M. (2004). An example of slow convergence of the bootstrap in high dimensions. *Amer. Statist.* **58**, 25-29.

Vickers, A. J. (2008). Decision analysis for the evaluation of diagnostic tests, prediction models and molecular markers. *Amer. Statist.* **62** (4) (To appear).

Waller, R. A. and Duncan, D. B. (1969). A Bayes rule for the symmetric multiple comparisons problem. *J. Am. Statist. Assoc.* **64**, 1484-1503.

Westfall, P. H., Tobias, R., Rom, D., Wolfinger, R. and Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests using the SAS® System, SAS®*Institute Inc. Books by Users.

Westfall, P. H. (2003). Comment on “Resampling-based multiple testing for microarray data analysis,” by Y. Ge, S. Dudoit and T. P. Speed. *Test* **12**, 60-65.

Westfall, P. H., and Troendle, J. F. (2008). Multiple testing with minimal assumptions. *Biom. J.* (To appear).

— Peter Westfall