Journal of the International Neuropsychological Society (1999), **5**, 510–517.

Extending the *p*-plot: Heuristics for multiple testing

IAN ABRAMSON,[1] TANYA WOLFSON,[2] THOMAS D. MARCOTTE,[2]
IGOR GRANT,[2,3] AND THE HNRC GROUP[2]

[1]Department of Mathematics, University of California at San Diego
[2]Department of Psychiatry, University of California at San Diego
[3]Psychiatry Service, VA San Diego Healthcare System

(RECEIVED October 9, 1997; REVISED August 17, 1998; ACCEPTED October 28, 1998)

Abstract

In the problem of large-scale multiple testing the *p*-plot is a graphically based competitor to the notoriously weak Bonferroni method. The *p*-plot is less stringent and more revealing in that it gives a gauge of how many hypotheses are decidedly false. The method is elucidated and extended here: the bootstrap reveals bias and sampling error in the usual point estimates, a bootstrap-based confidence interval for the gauge is given, as well as two acceptably powerful blanket tests of significance. Guidelines for use are given, and interpretational pitfalls pointed out, in the discussion of a case study linking premortem neuropsychological and postmortem neuropathologic data in an HIV cohort study. (*JINS*, 1999, *5*, 510–517)

Keywords: Multiple testing, *P*-plot, Bonferroni, Bootstrap

INTRODUCTION

It is common in both observational and controlled studies to record many variables on a set of participants, usually split into several groups. The questions at hand may be of any form involving correlations, group mean comparisons, or many other sets of statistics, but a common theme is to find the significant results among them with some control on overall error rate.

Common to such studies is the problem inherent in multiple testing. Suppose N simultaneous independent tests of null hypotheses H_i are conducted, each test with the type I error probability at some level α (often .05). Type I error means rejecting the null hypothesis when it is true, in other words getting a spurious significant p value on a test. Multiple testing is frustrated by the fact that the overall probability of type I error for the entire collection of tests is higher than α. How much higher depends on the number of tests run, but it can approach 100% if there are enough of them. Assuming that all of the null hypotheses in our collection are true, the overall type I error probability is the chance of raising any flag when none should be raised. Computing this probability is simple:

$$\begin{aligned}\text{overall probability} &= P\{\text{at least one } H_i \text{ rejected}\}\\ &= 1 - P\{\text{no } H_i \text{ rejected}\}\\ &= 1 - (1-\alpha)^{(\text{number of tests})}\end{aligned}$$

For 100 tests at .05 this probability is almost 1:.994. There are no simple cures for this problem.

Why Not Bonferroni?

The Bonferroni procedure is a fundamental but notoriously weak approach to multiple testing. Simple to state and use, the basic form of the procedure for handling N simultaneous tests of H_i entails conducting each at level α/N, and rejecting (at level α) an implicit "grand null" hypothesis H_0 that all the nulls H_i are true, if and only if at least one of the H_i is rejected (at level α/N). The idea is to limit the overall type I error probability by α. Since there are many chances to raise a red flag mistakenly, the (null) probability of each gets limited to α/N. (The overall level is then demonstrably no higher than α, but an exact determination is affected by dependencies between the tests, seldom known.)

Dr. Erin Bigler served as Action Editor during the course of this review.

Reprint requests to: Tanya Wolfson, HIV Neurobehavioral Research Center, 2760 5th Avenue Ste 200, San Diego CA 92103. E-mail: twolfson@ucsd.edu

The problem is that while there may have been decent power to detect likely departures from many H_i at the α level, the power to detect those departures at the α/N level can be so poor as to render the procedure useless.

Can Bonferroni Be Improved?

Weakness is intrinsic to methods that are based on the Bonferroni point of view. Minor improvements are possible:

- Bonferroni is slightly more conservative even than Bonferroni needs to be if the tests of H_i are only weakly dependent (seldom verifiable). In that case if either α is small (it almost always is), or N is large, each H_i could be tested at the $(\alpha + \alpha^2/2)/N$ level for the same overall level α. This is only slightly less stringent and the improvement is negligible.
- There is a more-eggs-in-some-baskets approach: Instead of impartially splitting the overall α into N equal parts one can assign larger significance levels to the tests thought (*a priori*) to have the best power, smaller ones going to the rest. They must still add up to α and (the rub), it must be done prospectively. Arbitrariness of the division enters here, and with it the appearance of having chosen it expediently.
- Perhaps the best prospect for getting more benefit from Bonferroni is to understand it better, and not (as people often do) have N count every test in sight if the grand null hypothesis so formed makes no sense. There isn't necessarily just one grand null per paper. Depending on one's scientific intent one can tackle several separately interesting clusters or domains of questions, each involving multiple testing to be handled by Bonferroni with a modest N, and not unduly concern oneself with the chance of a false pronouncement somewhere (anywhere) in any one of the clusters all taken together—for which the total N would be much larger. This is especially valid when the work is exploratory and intended to motivate further data gathering or finer analysis, rather than being the last word on an inference problem. Valuable information in large multivariate databases would go to waste if they were to be rigidly analyzed with all multiple comparisons safeguards in place. Data mining and fishing expeditions are dirty words, but tempered with an awareness of the fallacies they can lead to, and supported by honest documentation, it is not a scientific crime to scan a profusion of p values, regarding them roughly as indicators of falsehood of corresponding nulls, descriptive data in their own right, and go from there. The idea of p values as data brings us to the p-plot.

A BRIEF SURVEY

There are many approaches to the multiple comparisons problem, including refinements on the basic Bonferroni technique, and forerunners of p-plots. Hills (1969), suggests a half-normal plot for assessing large correlation matrices in particular. In this graphical approach, ranked Fisher-transformed correlations are plotted against normal Z-quantiles. The underlying assumption, like that for the p-plots, is that transformed correlations not significantly different from zero should lie (suitably plotted) about a straight line through the origin, while those significantly larger should produce visible bending. Simes (1986), modified the basic Bonferroni approach by using ranked p values. The basic rule of rejecting the "grand null" hypothesis H_0 if any one out of p values is less than α/N, was changed to a more lenient rule of rejecting H_0 if the jth ordered p value $p_{(j)}$ is smaller than $j\alpha/N$. The approach, while supported by a rationale, is not always appropriate, as striking counterexamples can be constructed to illustrate. Hommel (1988) further suggested a decision strategy for individual hypotheses in the Simes (1986) procedure. Recent advances are due to Hochberg and Benjamini (1995), who describe an approach in which the false detection rate (which they argue should be per test performed) is controlled. Zhang et al. (1997) offers a very good survey of recent developments in the field.

THE IDEA OF A *P*-PLOT

Bonferroni wastes information. Imagine scanning the p values from a set of N tests. Applying Bonferroni one would find the smallest value, compare it with α/N, and if it is smaller, reject the grand null. All the usable information in the p values is distilled down to the smallest of them. The p-plot and virtually all other multiple comparisons methods put the other p values to use as well; even the large ones can inform in some way, as we will see.

It is a theorem (roughly stated) that a p value is uniformly distributed between zero and 1 if a simple null hypothesis is true. This means that if one could repeat the whole experiment over and over, basing a new p value each time on the replicate datasets, their histogram would approximate a boxcar shape from zero to 1 if the null were true (and the sampled p values would be less than .05, for instance, 5% of the time—the false alarm or type I error rate). If the null is false this so-called sampling distribution for the p values is "stochastically smaller"—more bunched up to the left, and with a narrower right tail. Then the relative area to the left of .05 exceeds 5%. It is in fact the power of the test at the 5% level.

It follows then, brushing a mathematical technicality aside for the moment (see Cautionary Notes below), that if every null from a set of tests is true, the p values should look like a sample from the uniform distribution, whereas if "something is going on," that is, the grand null is false, the p values should look like a mixture of samples, a blend of a uniform component, corresponding to the true nulls, and stochastically smaller components, corresponding to the false nulls (none of which may actually have made the stringent Bonferroni cut). There is a graphical diagnostic called the QQ-plot (Rice, 1988) for detecting whether the so-called empirical distribution of p values is pure uniform or such a

https://doi.org/10.1017/S1355617799566046 Published online by Cambridge University Press

mixture. It is a testing problem in its own right, a kind of goodness of fit test, and the p values themselves become the data for it.

The p-plot (of which Figure 2 will be an enhanced example) is a close adaptation of the standard QQ-plot, and is constructed by plotting the number of p values in the set exceeding the variable p, not against p, but against $1 - p$, which is a bit peculiar, but it is the convention. The curve climbs from zero at the origin to N at $1 - p = 1$, taking jumps of 1 at every point that is ($1 -$ an observed p value). Under the grand null it will track a straight line of slope N through the origin with minor random irregularities. Otherwise the p values will follow a non-uniform distribution and the curve will bend systematically upwards, that is, track a convex curve. When N is large the irregularities will not mask the shape of the curve, and it is easy to distinguish a convex curve from a straight line and hence assess departure from the grand null. But there will always be some randomness in the sample of p values, and if N is small the spacings between the points become large, which, on "joining the dots," could create the illusion of systematic bending even if there isn't any. The graphic may then mislead, or at least not be striking, although the overall quantitative significance test described later would still be valid (if weak). Guidelines for a threshold N are hard to give, but the procedure would not be recommended for N less than about 10 (in which case the Bonferroni method itself might be tolerably powerful).

The original source (Schweder & Spjotvoll, 1982) contains full details on the basic construction of the p-plot, which we will not repeat, except to point out that a key step in the quantitative interpretation of the plot is to fit a straight line through the origin, approximating the left part of the curve. Precise data-driven guidelines on where the "left part" ends are hard to give, but loosely speaking it should not extend beyond where the bending becomes obvious. Instead of choosing an abrupt right cutoff for the points on the p-plot to which the regression is fitted (which inevitably involves some arbitrariness and visual judgement), one could fit a weighted least squares regression with weights decaying according to a sigmoidal pattern across the plot.

The slope of the line approximating the left part of the curve is an estimate of the number of true or approximately true null hypotheses. To put this on a slightly firmer mathematical footing than it appears at first, if indeed M of the N hypotheses is true, and is "large", and all the false hypotheses are decidedly false, so that they contribute a very thin-tailed component to the sampling distribution mixture of the p values, then the slope of the tangent of the p-plot through the origin is indeed a reasonable estimate of M.

AN EXAMPLE

A look at a real problem serves at this point to motivate our extensions of the p-plot. The data that are used as examples here are derived from a correlational study that has recently been reported (Masliah et al., 1997), although the applicability of the method is much wider (and includes, for instance, the common problem of multiple testing of location shifts).

Investigators from the Neurobehavioral and Neuropathology cores at the HIV Neurobehavioral Research Center (HNRC) each gathered a sample of multivariate data vectors corresponding to (and linked by) the participants in the study. The Neurobehavioral Core of the HNRC studies the processes underlying HIV-related neuropsychological impairment. The objective of the core is to better characterize the mechanisms and neuromedical outcomes of neuropsychological impairment frequently seen in patients infected with HIV-1. The Neuropathology Core of the HNRC was established in an effort to link premortem and postmortem information. The core performs postmortem studies and coordinates the diagnosis of brain tissue in persons who have died from AIDS. Its data provide key covariates in exploring correlations of post- with premortem findings from the various cores in the HNRC.

The combined data vectors from the investigators of both cores are used to answer the question of what pathological changes in the brain are associated with cognitive deficits in persons with AIDS. Previous studies examining postmortem brain tissue have been inconclusive on relationships between neuropsychological impairment and the amount of virus in the brain, neuronal loss and other pathological markers. For the present study the investigators sought to determine whether the amount of neuronal branching was significantly related to premortem neurobehavioral functioning. Dendritic injury was assessed by immunohistochemical staining of brain tissue using monoclonal antibodies to microtubule associated protein 2 (MAP2) which is considered a specific dendritic marker. The investigators determined the percent area occupied by MAP2-immunolabeled dendrites and compared it to the density of a presynaptic (axonal) marker, synaptophysin (SYN). Twenty participants with neuropsychological evaluations within 18 months of death were involved in this study. Presynaptic and dendritic staining was performed on three brain regions believed to be affected by HIV: midfrontal region, putamen and globus pallidus (MF, PUT, and GP). The subset of neuropsychological tests selected by the Neurobehavioral Core is commonly used to develop clinical ratings for eight functional neuropsychological domains (Verbal, Abstraction, Perceptual–Motor, Attention, Learning, Memory, Motor and Sensory) as well as a rating of overall neuropsychological functioning (Global). In this example T scores from tests rather than the neuropsychological domains are used, since the investigators are often interested in the relationship between damage present in neuroanatomical regions and cognitive functioning as assessed by specific neuropsychological tests.

The analysis could be presented as a large correlation matrix, or as a large matrix of corresponding p values (shown in Table 1, and as a histogram in Figure 1), based on the standard approximation to Fisher's z-transform.

This display is strikingly uninformative. The Bonferroni overall p value for the grand null hypothesis is the smallest

https://doi.org/10.1017/S1355617799566046 Published online by Cambridge University Press

Table 1. *P* values for 150 correlations

Test	MAP2 midfront	MAP2 putamen	MAP2 glob pall	SYN midfront	SYN putamen	SYN glob pall
WAIS–R Vocab	.405	.803	.949	.478	.024	.121
WAIS–R Blocks	.007	.454	.049	.029	.340	.585
WAIS–R Digit Sym	.007	.927	.701	.296	.947	.900
WAIS–R Digit Span	.151	.841	.645	.587	.678	.807
Boston Naming	.288	.194	.184	.237	.550	.943
Fluency–Category	.017	.181	.315	.415	.639	.707
Thurstone Word	.001	.778	.445	.041	.195	.567
FAS	.00013	.846	.986	.167	.879	.922
Category Test	.010	.262	.491	.183	.077	.221
Trails A	.262	.892	.674	.498	.882	.802
Trails B	.032	.356	.322	.050	.231	.732
Story Lrn Memory	.057	.196	.164	.609	.882	.827
Story %Loss	.607	.105	.333	.690	.255	.763
Figure Lrn Memory	.005	.222	.953	.083	.983	.794
Figure %Loss	.361	.366	.427	.062	.156	.092
Sens Percep Exam	.803	.609	.675	.540	.851	.855
Grip Dom	.048	.418	.469	.578	.394	.384
Grip Non–Dom	.083	.316	.318	.570	.522	.482
Pegs Dom	.032	.057	.172	.176	.188	.239
Pegs Non–Dom	.027	.083	.262	.139	.141	.184
Tapping Dom	.026	.162	.074	.031	.085	.051
Tapping Non–Dom	.126	.350	.159	.314	.394	.226
Digit Vigil (Err)	.305	.289	.959	.822	.971	.311
Digit Vigil (Time)	.049	.262	.430	.118	.229	.316
PASAT	.084	.754	.463	.243	.556	.808

Note. WAIS–R Vocab = Wechsler Adult Intelligence Scale–Revised Vocabulary; WAIS–R Digit Span = WAIS–R Digit Span; WAIS–R Blocks = WAIS–R Block Design; WAIS–R Dig Sym = WAIS–R Digit Symbol; Boston Naming = Boston Naming Test; Fluency Category = Category Fluency Test; Thurstone Word = Thurstone Word Fluency; FAS = Controlled Oral Word Association Test; Category Test = Category Test (Errors); Trails A = Trail Making Test A (Time); Trails B = Trail Making Test B (Time); Story Lrn Memory = Story Memory Learning; Story %Loss = Story Memory Retention (Loss); Figure Lrn Memory = Figure Memory Learning; Figure %Loss = Figure Memory Retention (Loss); Sens Percep Exam = Sensory-Perceptual Examination; Grip Dom = Hand Dynamometer (Dominant Hand); Grip Non-Dom = Hand Dynamometer (Non-Dominant Hand); Pegs Dom = Grooved Pegboard Test (Dominant Hand; Time); Pegs Non-Dom = Grooved Pegboard Test (Non-Dominant Hand; Time); Tapping Dom = Finger Tapping Test (Dominant Hand); Tapping Non-Dom = Finger Tapping Test (Non-Dominant Hand); Digit Vigil (Err) = Digit Vigilance Test (Errors); Digit Vigil (Time) = Digit Vigilance Test (Time); PASAT = Paced Auditory Serial Addition Test.

p value observed multiplied by *N*, in our case it is .00013 × 150, which gives an overall *p* value of .02. We can in fact reject the grand null, but at a *p* value that is a poor reflection of the obviously overwhelming evidence against it. Bonferroni also gives us a *p*-value cutoff for significance, α/N. In our case the *p*-value cutoff would be .0003. Only one out of our 150 *p* values survives Bonferroni. The *p*-plot however (Figure 2) shows the unmistakable convex bend. A line fitted through the origin to the leftmost 80% of the points is superimposed. Its slope is 114, and the rough conclusion is that 114 of the null hypotheses are close to true. About 36 must then be false—but see Cautionary Notes, below.

Gauging Bias and Uncertainty From the *P*-Plot: The Bootstrap Extension

The point estimate 114 is of course uncertain. How can we gauge its sampling variation and fit a confidence interval? On the face of it, this is a theoretical nightmare. The sampled correlations are based on variable sample sizes according to the missing value patterns across the data, there are unknown mutual dependencies between the tests, and the true correlation patterns are unknown. A problem made in heaven, in other words, for the bootstrap. The rough idea of the bootstrap is to mimic the sampling variation and bias in an estimation procedure by repeated sampling from the original sample of observations (multivariate in this case). Each sample will contain some of the original points repeated, and others not at all. There is a theoretical basis for the intuitive idea, which lies at the heart of the bootstrap, that the original dataset is to the population (of which it is a crude reflection), as the bootstrap samples are to the original sample. From this it is possible to estimate the sampling properties of the procedure (not only dispersion, but correctable bias as well), and fit confidence intervals with very accurate coverage probabilities, even under nonstandard distri-

https://doi.org/10.1017/S1355617799566046 Published online by Cambridge University Press

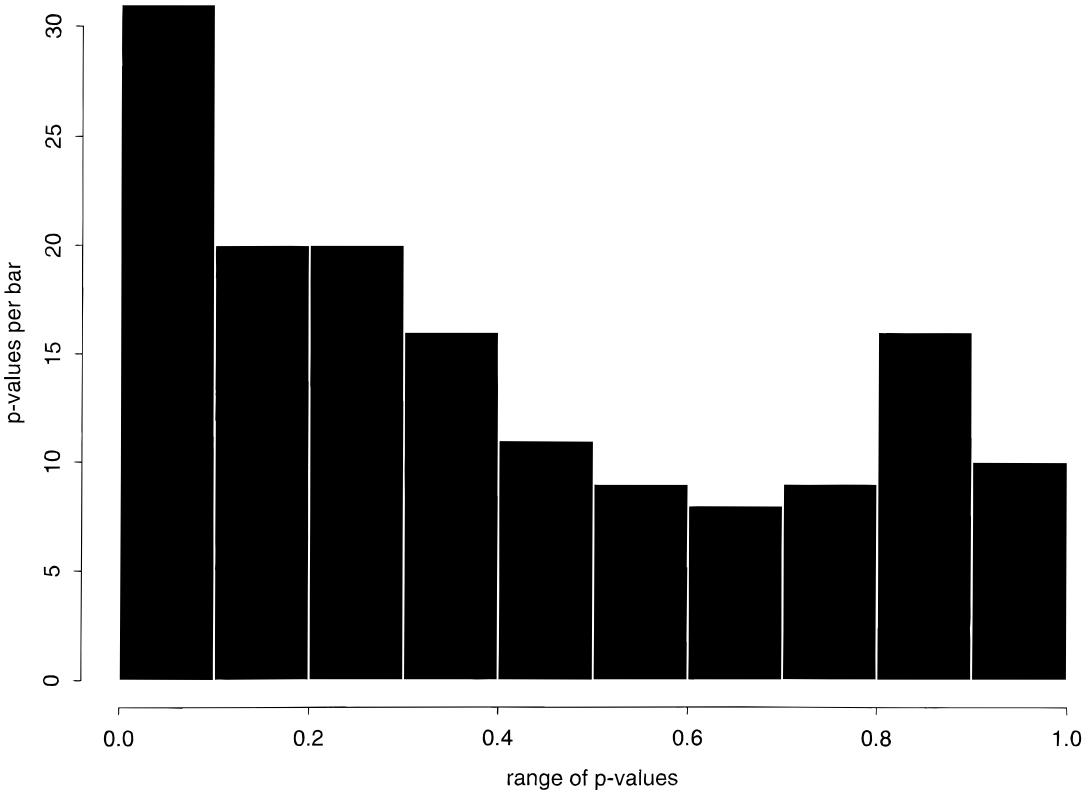

Fig. 1. Histogram of *p* values from 150 correlations.

butional assumptions. There are several such refined interval estimation techniques in use now. The one used here is called BCA (bias correction and acceleration), and is described, along with a full exposition of the bootstrap, in Efron and Tibshirani (1993). In this case a histogram of the "bootstrap distribution" that emerges for the number of true nulls is shown in Figure 3. It is based on 1500 bootstrap replicates. There is some bias in the estimation technique, evident from the off-central placing of the original point estimate 114. Since the bootstrap samples are to the original sample as the original sample to the population, the bias is stronger in the bootstrap. Thus the confidence interval corrected for bias is even further away from the center of the bootstrap samples than the original estimate. Our BCA 95% confidence interval turns out to be (100, 143). It adjusts for the bias as well as for nonnormality of the sampling distribution. This interval is in itself strong evidence against the "grand null" hypothesis that all 150 nulls are true. A more quantitative approach to this question is given in the next section.

A Blanket *P*-Value

The Bonferroni method furnishes an overall *p* value for the grand null hypothesis—the smallest *p* value observed, multiplied by *N*. The *p*-plot suggests strong competitors based on the geometry of the plot. One possibility is to fit a restricted-range linear regression to the right side of the plot, and to do a one-sided significance test for its having a larger slope than the line through the origin, on which the estimate of the number of true null is based. (Under the grand null these lines should be approximately coincident.) Any dependencies induced by overlapping regression ranges (or by dependencies among the *p* values themselves) should serve to make this test err on the side of conservatism. Another possibility, which does not require the arbitrary choice of a fitting range, and which recognizes that *p*-plots will track a smooth curve, is to fit a quadratic equation through the origin (i.e., without an intercept term), and perform a one-sided significance test on the quadratic coefficient, which is zero under the grand null, and positive if the curve is convex. Both these approaches are illustrated on the accompanying *p*-plot, together with overall *p* values that each gives.

CAUTIONARY NOTES

Having estimated the slope (M) at the left of the *p*-plot, and deduced that approximately $N - M = K$ nulls among the original N are false, it is tempting to scan the original list of *p* values and single out the K hypotheses corresponding to the K smallest, declaring them to be the false ones. This is misleading, and is deprecated. The claim that there are approximately K false nulls among N may be well supported, but the extraction of a particular subset of size K is not. Inevitably many tests will have *p* values falling on the wrong side of the threshold. Our view is that the number K (or the

https://doi.org/10.1017/S1355617799566046 Published online by Cambridge University Press

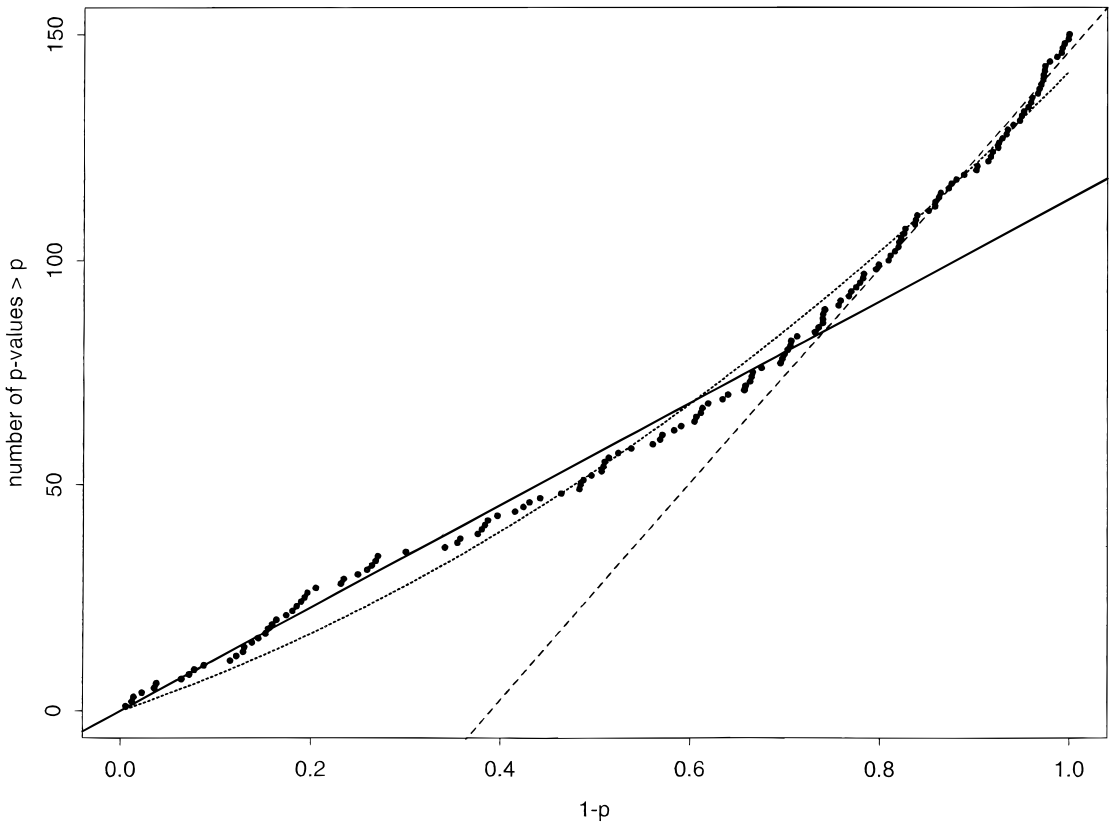

Fig. 2. Plot of *p* values. The solid line on the plot is the least squares line, fit through the leftmost 80% of the points. The dashed line is the least squares fit through the remaining rightmost 20%. Comparison of the slopes of these two lines yields a *p* value $< 10^{-6}$. The dotted curve is the quadratic regression, its quadratic coefficient *p* value $< 10^{-6}$ also. The estimated number of true null hypotheses is equal to the slope of the left least squares line, approximately 114.

ratio K/N) be regarded as an index of intrinsic interest, a gauge, loosely speaking of "how much is going on." There is another more insidious problem with the *p*-plot that precludes a strict theoretical foundation for the method, particularly for the associated significance testing: There are dependencies between the tests, and so between the corresponding *p* values. While it is true that under the grand null the *marginal* distribution of the *p* values is uniform, it does not follow that the empirical distribution of all of them together must necessarily look like a uniform random sample, even under the grand null. There may be "clumping" phenomena, easily imagined in an extreme case where all the tests involved are strongly dependent, with all answering essentially the same question on essentially the same data. (Dependencies between the tests is not the same as dependencies between the variables, which do not disturb the validity of the method, and indeed are sometimes the very target of the investigation.) Truly rigorous use of the *p*-plot depends on there being enough independence among the tests to span the range of *p* values without too much clumping, a condition hard to formulate precisely or verify, but usually in force unless linear redundancy in the data is gross.

WHY *P*-PLOTS?

At this point one is tempted to ask: What good is determining from the *p*-plot that approximately *K* nulls are significant if we cannot reliably determine which subset of *K* it is? Bonferroni, after all, points rigorously to a subset (often empty!) of significant nulls. Unfortunately, in dealing with the type I error control in multiple testing there is a tradeoff between stringency and simplicity. Bonferroni is a stringent method, and needlessly so for an experiment with a large number of simultaneous hypotheses. So the subset it yields will almost always be a small one. Both Bonferroni and the *p*-plots furnish a blanket *p* value for the grand null hypothesis. But where one typically dismisses all results as insignificant after Bonferroni, the overall *p* value obtained from *p*-plots discovers overall significance by more lenient rules. The question of overall significance is of course a straw man. In a first pass at the data, once we determine that we have overall significance, it is of limited value *per se*. The compelling question is, which of our hypotheses are false. Now that we know there are *K*, how do we isolate them? The answer is—only by further data-gathering. The *p*-plot by itself is chiefly an exploratory device. Validating the re-

https://doi.org/10.1017/S1355617799566046 Published online by Cambridge University Press

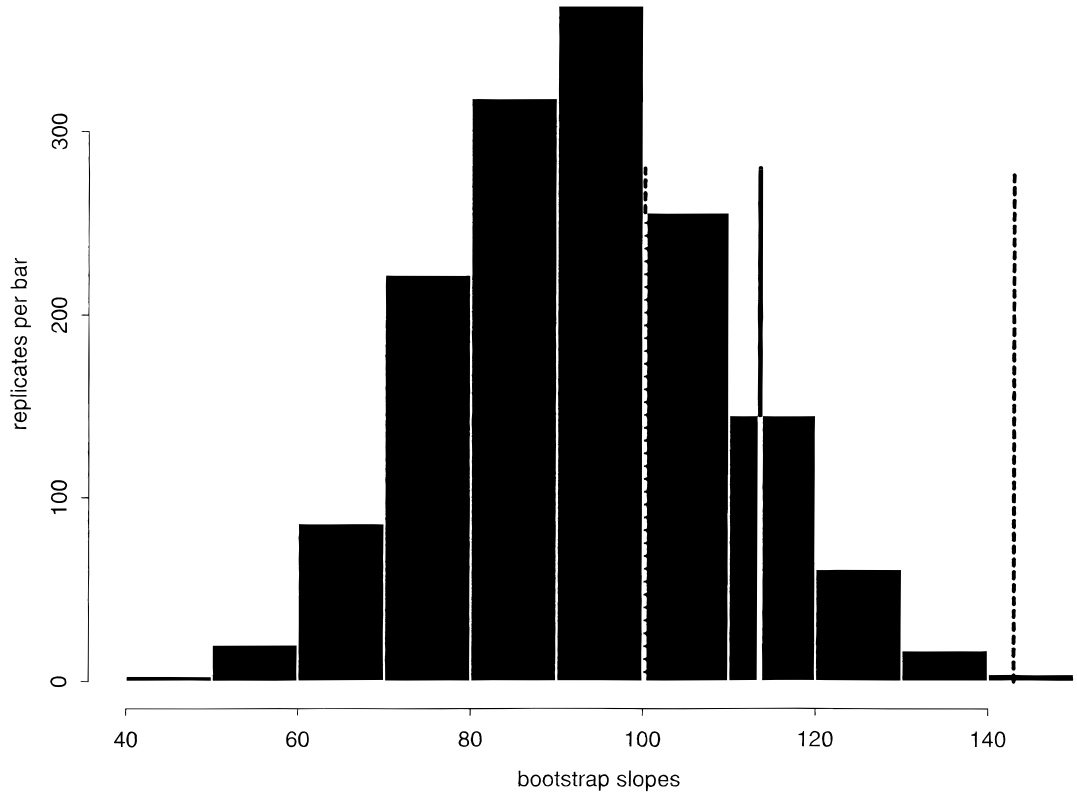

Fig. 3. Distribution of slopes from 1500 bootstrap replicates. The solid line marks the original slope of 114. Surrounding dashed lines mark the 95% bootstrap confidence interval at (100, 143). Note displacement of point estimate 114 from center, revealing correctable bias.

sults generally requires more data. An instructive technique, if there is enough data, is to split it randomly into two halves. Run the algorithm on the first half, and test the resultant subset of flagged hypotheses on the second. The selection of that subset will be based on the estimated K and the K smallest observed p-values. What typically happens using such a "cross-validation" technique (and there are many variations) is that there is a reasonable overlap between the subsets of hypotheses flagged on each of the half—replicate data subsets, but nothing like perfect agreement—which is sobering. No partitioning technique on the hypotheses can reliably separate true from false, which is unachievable, even in the one-sample, one-test situation. Even with this limitation the p-plot is a valuable adjunct to the hunch method in guiding the future of a study. It suggests hypotheses to append to those of key interest (which may not have shown early significance). Studies evolve, and the future focus and choice of data to gather and ways to model them need not be entirely rigid.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Robert K. Heaton, P.I. of the Neurobehavioral Component and Dr. Eliezer Masliah, P.I. of the Neuropathology Component for the use of their data. The HIV Neurobehavioral Research Center (HNRC) is supported by Center award P50-MH 45294 from NIMH. The San Diego HIV Neurobehavioral Research Center [HNRC] group is affiliated with the University of California, San Diego, the Naval Hospital, San Diego, and the San Diego VA Medical Center, and includes Igor Grant, M.D., Director; J. Hampton Atkinson, M.D., Co-Director; Thomas D. Marcotte, Ph.D., Center Manager; James L. Chandler, M.D. and Mark R. Wallace, M.D., Co-Investigators Naval Hospital San Diego; J. Allen McCutchan, M.D., P.I. Neuromedical Component; Stephen A. Spector, M.D., P.I. Virology Component; Robert K. Heaton, Ph.D., P.I. Neurobehavioral Component; Terry Jernigan, Ph.D. and John Hesselink, M.D. ., Co-P.I.s Imaging Component; Eliezer Masliah, M.D., P.I. Neuropathology Component; J. Allen McCutchan, M.D., J. Hampton Atkinson, M.D., and Ronald J. Ellis, M.D., Ph.D., Clinical Trials Component; Daniel R. Masys, M.D., P.I. Data Management Component; Ian Abramson, Ph.D., P.I. Statistics Unit; Julie Nelson, B.A., Data Manager. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the United States Government.

REFERENCES

Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Hills, M. (1969). On looking at large correlations matrices. *Biometrika, 56*, 249–253.

Hochberg, Y. & Benjamini, Y. (1995). Controlling the false dis-

https://doi.org/10.1017/S1355617799566046 Published online by Cambridge University Press

covery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B, 57*, 289–300.
Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75*, 383–386.
Masliah, E., Heaton, R.K., Marcotte, T.D., Ellis, R.J., Wiley, C.A., Mallory, M., Achim, C.L, McCutchan, J.A., Atkinson, J.H., Grant, I., & the HNRC Group (1997). Dendritic injury is a pathological substrate for Human Immunodeficiency Virus-related cognitive disorders. *Annals of Neurology, 42*, 963–972.
Rice, J. (1988). *Mathematical statistics and data analysis.* Monterey, CA: Brooks/Cole.
Schweder, T. & Spjotvoll, E. (1982). Plots of *P*-values to evaluate many tests simultaneously. *Biometrika, 69*, 493–502.
Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika, 73*, 751–754.
Zhang, J., Quan, H., Ng, J., & Stepanavage, M. (1997). Some statistical methods for multiple endpoints in clinical trials. *Controlled Clinical Trials, 18*, 204–221.

https://doi.org/10.1017/S1355617799566046 Published online by Cambridge University Press