Module 11: Cross-validation and the Control of Error Rates

As is your sort of mind, so is your sort of search; you'll find what you desire.
— Robert Browning (1812–1889)

Abstract: This module emphasizes what might be termed "the practice of safe statistics." The discussion is split into three parts: (1) the importance of cross-validation for any statistical method that relies on an optimization process based on a given data set (or sample); (2) the need to exert control on overall error rates when carrying out multiple testing, even when that testing is done only implicitly; (3) in the context of "big data" and associated methods for "data mining," the necessity of some mechanism for ensuring the replicability of "found results."

Contents

1 Cross-validation

Many texts in statistics that include a discussion of (multiple) regression and related techniques give little weight to the topic of cross-validation, which we believe is crucial to the appropriate (and "safe") use of these methods.[1] Cross-validation might be discussed under the rubric of how does a result found for a particular sample of data "hold up" in a new sample. As a general illustration, consider (multiple) regression where the interest is in predicting a single dependent measure, Y, from a linear combination of K independent variables, $X_1, \ldots, X_K$. As a measure of how well a regression equation does in the sample, we typically use the squared correlation (R^2) between the values on Y and those predicted from the regression equation, say, $\hat{Y}$. This is a measure of how well an equation does on the same data set from which it was derived, typically through an optimization process of least-squares. Our real interest, however, may be in how well or badly the sample equation works generally. The sample equation has been optimized with respect to the particular data at hand, and therefore, it might be expected that the squared correlation is "inflated." In other words, the concern is with sample equation performance in a new group; this is the quintessential question of "cross-validation."

[1] We have one salient example in Module 2 where a lack of cross-validation lead to overly-optimistic estimates of how well actuarial predictions of violence could be made. This was the development of the COVR instrument in the MacArthur Study of Mental Disorder and Violence. In the training sample, 1 out of 3 predictions of "violence" were wrong; in the one small cross-validation study done somewhat later using completely "new" data, 2 out of 3 predictions of "violence" were incorrect. In fact, the COVR even failed to be clinically efficient in the Meehl and Rosen sense – the diagnostic test was outperformed by prediction using simple base rates.

There are several general strategies that can be used to approach the task of cross-validation:

a) Get *new* data and use the sample equation to predict Y and calculate the squared correlation between Y and $\hat{Y}$; denoting this squared correlation by R^2_{new}, the difference $R^2 - R^2_{new}$ is called "shrinkage" and measures the drop in how well one can predict with new data. The chief problem with this first approach is that new data may be "hard to come by" and/or very expensive.

b) We can first split the original sample into two parts; obtain the equation on one part (the "training set") and test how well it does on the second (the "test set"). This is a common method of cross-validation; the only possible down-side is when the original sample is not very big, and the smaller training sample might produce a more unstable equation than desirable.

c) *Sample reuse methods*: here, the original sample is split into K parts, with the equation obtained with $K-1$ of the parts aggregated together and then tested on the one part left out. This process is repeated K times, leaving one of the K parts out each time; it is called K-fold cross-validation. Given the increased computational power that is now readily available, this K-fold cross-validation is close to being a universal default option (and with K usually set at around 10).

At the extreme, if n subjects are in the original sample, n-fold cross-validation would leave one person out at a time. For this person left out, say person i, we obtain $\hat{Y}_i$ and then calculate the squared correlation between the Y_i's and $\hat{Y}_i$'s to see how well we cross-validate

with a "new" sample. Each equation is constructed with $n-1$ subjects so there should be more stability present than in approach (b).

1.1 An Example of a Binary Classifier

The term *discrimination* (in a nonpejorative statistical sense) can refer to the task of separating groups through linear combinations of variables maximizing a criterion, such as an F-ratio. The linear combinations themselves are commonly called Fisher's linear discriminant functions. The related term *classification* refers to the task of allocating observations to existing groups, typically to minimize the cost and/or probability of misclassification. These two topics are intertwined, but here we briefly comment on the topic of classification when there are two groups (or in the current jargon, we will construct a "binary classifier").

In the simple two-group situation, there are two populations, π_1 and π_2; π_1 is assumed to be characterized by a normal distribution with mean μ_1 and variance σ_X^2 (the density is denoted by $f_1(x)$); π_2 is characterized by a normal distribution with mean μ_2 and (common) variance σ_X^2 (the density is denoted by $f_2(x)$). Given an observation, say x_0, we wish to decide whether it should be assigned to π_1 or to π_2. Assuming that $\mu_1 \leq \mu_2$, a criterion point c is chosen; the rule then becomes: allocate to π_1 if $x_0 \leq c$, and to π_2 if $> c$. The probabilities of misclassification are given in the following chart:

		True State	
		π_1	π_2
Decision	π_1	$1-\alpha$	β
	π_2	α	$1-\beta$

In the terminology of our previous usage of Bayes' rule to obtain the positive predictive value of a test, and assuming that π_1 refers to a person having "it," and π_2 to not having "it," the sensitivity of the test is $1-\alpha$ (true positive); specificity is $1-\beta$, and thus, β refers to a false negative and α to a false positive.

To choose c so that $\alpha+\beta$ is smallest, select the point at which the densities are equal. A more complicated way of stating this decision rule is to allocate to π_1 if $f_1(x_0)/f_2(x_0) \geq 1$; if < 1, then allocate to π_2. Suppose now that the prior probabilities of being drawn from π_1 and π_2 are p_1 and p_2, respectively, where $p_1+p_2=1$. If c is chosen so the Total Probability of Misclassification (TPM) is minimized (that is, $p_1\alpha+p_2\beta$), the rule would be to allocate to π_1 if $f_1(x_0)/f_2(x_0) \geq p_2/p_1$; if $< p_2/p_1$, then allocate to π_2. Finally, to include costs of misclassification, $c(1|2)$ (for assigning to π_1 when actually coming from π_2), and $c(2|1)$ (for assigning to π_2 when actually coming from π_1), choose c to minimize the Expected Cost of Misclassification (ECM), $c(2|1)p_1\alpha + c(1|2)p_1\beta$, by the rule of allocating to π_1 if $f_1(x_0)/f_2(x_0) \geq (c(1|2)/c(2|1))(p_2/p_1)$; if $< (c(1|2)/c(2|1))(p_2/p_1)$, then allocate to π_2.

Using logs, the last rule can be restated:

allocate to π_1 if $\log(f_1(x_0)/f_2(x_0)) \geq \log((c(1|2)/c(2|1))(p_2/p_1))$. The left-hand side is equal to

$(\mu_1 - \mu_2)(\sigma_X^2)^{-1}x_0 - (1/2)(\mu_1 - \mu_2)(\sigma_X^2)^{-1}(\mu_1 + \mu_2)$,
so the rule can be rephrased further:

allocate to π_1 if

$$x_0 \leq \{(1/2)(\mu_1 - \mu_2)(\sigma_X^2)^{-1}(\mu_1 + \mu_2) - \log((c(1|2)/c(2|1))(p_2/p_1))\{\frac{\sigma_X^2}{-(\mu_1 - \mu_2)}\}$$

or

$$x_0 \leq \{(1/2)(\mu_1+\mu_2) - \log((c(1|2)/c(2|1))(p_2/p_1))\}\{\frac{\sigma_X^2}{(\mu_2 - \mu_1)}\} = c\,.$$

If the costs of misclassification are equal (that is, $c(1|2) = c(2|1)$), then the allocation rule is based on classification functions: allocate to π_1 if

$$[\frac{\mu_1}{\sigma_X^2}x_0 - (1/2)\frac{\mu_1^2}{\sigma_X^2} + \log(p_1)] - [\frac{\mu_2}{\sigma_X^2}x_0 - (1/2)\frac{\mu_2^2}{\sigma_X^2} + \log(p_2)] \geq 0\,.$$

The classifier just constructed has been phrased using population parameters, but to obtain a sample-based classifier, estimates are made for the population means and variances. Alternatively, a "dummy" binary dependent variable Y ($= 0$ for an observation in group 1; $= 1$ for an observation in group 2) can be predicted from X; the sample-based classifier is obtained in this way. Also, this process of using a binary Y but with K independent variables, $X_1, \ldots, X_K$, leads to a binary classifier based on more than one independent variable (and to what is called Fisher's linear discriminant function).[2]

[2]In the terminology of signal detection theory and the general problem of yes/no diag-

In moving to the sample where estimated quantities (sample means, variances, and covariances) are used for the population parameters, we can do more than just hope that the (sample) classification rule does well by carrying out a cross-validation. First, a misclassification table can be constructed based on simple resubstitution of the original data into the sample classification rule (where n_1 observations are in group π_1 and n_2 are in group π_2):

		True State	
		π_1	π_2
Decision	π_1	a	b
	π_2	c	d
sums		n_1	n_2

The apparent error rate (APR) is $(b + c)/n$, which is overly optimistic because it is optimized with respect to this sample. A K-fold cross-validation would give a less optimistic estimate; for example, letting $K = n$ and using the "hold out one-at-a-time" strategy, the following misclassification table might be obtained:

nostic decisions as discussed in Module 4, a plot of sensitivity (true positive probability) on the y-axis against 1− specificity on the x-axis as c varies, is an ROC curve (for Receiver Operating Characteristic). This ROC terminology originated in World War II in detecting enemy planes by radar (group π_1) from the noise generated by random interference (group π_2). The ROC curve is bowed from the origin of (0, 0) at the lower-left corner to (1.0, 1.0) at the upper right; it indicates the trade-off between increasing the probability of true positives and the increase of false positives. Generally, the adequacy of a particular diagnostic decision strategy is measured by the area under the ROC curve, with areas closer to 1.0 being better; that is, steeper bowed curves hugging the left wall and the top border of the square box. For a comprehensive introduction to diagnostic processes, see Swets, Dawes, and Monahan (2000).

		True State	
		π_1	π_2
Decision	π_1	a*	b*
	π_2	c*	d*
sums		n_1	n_2

To estimate the actual error rate (AER), we would use $(b^* + c^*)/n$, and would expect it to be greater than the APR.

2 Problems With Multiple Testing

A difficulty encountered with the use of automated software analyses is that of multiple testing, where the many significance values provided are all given as if each were obtained individually without regard for how many tests were performed. This situation gets exacerbated when the "significant" results are then culled, and only these are used in further analysis. A good case in point is reported in the next section on odd correlations where highly inflated correlations get reported in fMRI studies because an average is taken only over those correlations selected to have reached significance according to a stringent threshold. Such a context is a clear violation of a dictum given in many beginning statistics classes: you cannot legitimately test a hypothesis on the same data that first suggested it.

Exactly the same issue manifests itself, although in a more subtle, implicit form, in the modern procedure known as data mining. Data mining consists of using powerful graphical and algorithmic methods

to view and search through high-dimensional data sets of moderate-to-large size, looking for interesting features. When such a feature is uncovered, it is isolated and saved. Implicit in the search, however, are many comparisons that the viewer makes and decides are not interesting. Because the searching and comparing is done in real time, it is difficult to keep track of how many "insignificant" comparisons were discarded before alighting on a significant one. Without knowing how many, we cannot judge the significance of the interesting features found without an independent confirmatory sample. Such independent confirmation is all too rarely done.

Uncontrolled data mining and multiple testing on some large (longitudinal) data sets can also lead to results that might best be labeled with the phrase "the oat bran syndrome." Here, a promising association is identified; the relevant scientists appear in the media and on various cable news shows; and an entrepreneurial industry is launched to take advantage of the supposed findings. Unfortunately, some time later, contradictory studies appear, possibly indicating a downside of the earlier recommendations, or at least no replicable effects of the type reported previously. The name "the oat bran syndrome" results from the debunked studies from the 1980s that had food manufacturers adding oat bran to absolutely everything, including beer, to sell products to people who wanted to benefit from the fiber that would supposedly prevent cancer.

To be more formal about the problem of multiple testing, suppose there are K hypotheses to test, $H_1, \ldots, H_K$, and for each, we set the criterion for rejection at the fixed Type I error value of α_k, $k = 1, \ldots, K$. If the event A_k is defined as the incorrect rejection of H_k

(that is, rejection when it is true), the Bonferroni inequality gives

$$P(A_1 \text{ or } \cdots \text{ or } A_K) \leq \sum_{k=1}^{K} P(A_k) = \sum_{k=1}^{K} \alpha_k \ .$$

Noting that the event (A_1 or $\cdots$ or A_K) can be verbally restated as one of "rejecting incorrectly *one or more* of the hypotheses," the experiment-wise (or overall) error rate is bounded by the sum of the K α values set for each hypothesis. Typically, we let $\alpha_1 = \cdots = \alpha_K = \alpha$, and the bound is then $K\alpha$. Thus, the usual rule for controlling the overall error rate through the Bonferroni correction sets the individual αs at some small value such as $.05/K$; the overall error rate is then guaranteed to be no larger than .05.

The problem of multiple testing and the failure to practice "safe statistics" appears in both blatant and more subtle forms. For example, companies may suppress unfavorable studies until those to their liking occur. A possibly apocryphal story exists about toothpaste companies promoting fluoride in their products in the 1950s and who repeated studies until large effects could be reported for their "look Ma, no cavities" television campaigns. This may be somewhat innocent advertising hype for toothpaste, but when drug or tobacco companies engage in the practice, it is not so innocent and can have a serious impact on our collective health. It is important to know how many things were tested to assess the importance of those reported. For example, when given only those items from some inventory or survey that produced significant differences between groups, be very wary!

People sometimes engage in a number of odd behaviors when doing

multiple testing. We list a few of these below in summary form:

(a) It is not legitimate to do a Bonferroni correction post hoc; that is, find a set of tests that lead to significance, and then evaluate just this subset with the correction;

(b) Scheffé's method (and relatives) are the only true post-hoc procedures to control the overall error rate. An unlimited number of comparisons can be made (no matter whether identified from the given data or not), and the overall error rate remains constant;

(c) You cannot look at your data and then decide which planned comparisons to do;

(d) Tukey's method is not post hoc because you actually plan to do all possible pairwise comparisons;

(e) Even though the comparisons you might wish to test are independent (such as those defined by orthogonal comparisons), the problem of inflating the overall error rate remains; similarly, in performing a multifactor analysis of variance (ANOVA) or testing multiple regression coefficients, all of the tests carried out should have some type of control imposed on the overall error rate;

(f) It makes little sense to perform a multivariate analysis of variance before you go on to evaluate each of the component variables. Typically, a multivariate analysis of variance (MANOVA) is completely noninformative as to what is really occurring, but people proceed in any case to evaluate the individual univariate ANOVAs irrespective of what occurs at the MANOVA level; we may accept the null hypothesis at the overall MANOVA level but then illogically

ask where the differences are at the level of the individual variables. Plan to do the individual comparisons beforehand, and avoid the uninterpretable overall MANOVA test completely.

We cannot leave the important topic of multiple comparisons without at least a mention of what is now considered one of the more powerful methods currently available: the False Discovery Rate (Benjamini & Hochberg, 1995). But even this method is not up to the most vexing of problems of multiplicity. We have already mentioned data mining as one of these; a second problem arises in the search for genetic markers. A typical paradigm in this crucial area is to isolate a homogeneous group of individuals, some of whom have a genetic disorder and others do not, and then to see if one can determine which genes are likely to be responsible. One such study is currently being carried out with a group of 200 Mennonites in Pennsylvania. Macular degeneration is common among the Mennonites, and this sample was chosen so that 100 of them had macular degeneration and a matched sample of 100 did not. The genetic structure of the two groups was very similar, and so the search was on to see which genes were found much more often in the group that had macular degeneration than in the control group. This could be determined with a t-test. Unfortunately, the usefulness of the t-test was diminished considerably when it had to be repeated for more than 100,000 separate genes. The Bonferroni inequality was no help, and the False Discovery Rate, while better, was still not up to the task. The search still goes on to find a better solution to the vexing problem of multiplicity.[3]

[3]The probability issues involved with searching through the whole genome are discussed in: "Nabbing Suspicious SNPS: Scientists Search the Whole Genome for Clues to Common

3 Odd Correlations

A recent article (Vul et al. 2009) in *Perspectives on Psychological Science*, has the intriguing title, "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition" (renamed from the earlier and more controversial "Voodoo Correlations in Social Neuroscience"; note that the acronym fMRI stands for functional Magnetic Resonance Imaging, and is always written with a lower-case letter "f"). These authors comment on the extremely high (for example, greater than .80) correlations reported in the literature between brain activation and personality measures, and point out the fallaciousness of how they were obtained. Typically, huge numbers of separate correlations were calculated, and only the mean of those correlations exceeding some threshold (based on a very small significance level) are reported. It is tautological that these correlations selected for size must then be large in their average value. With no cross-validation attempted to see the shrinkage expected in these measures on new samples, we have sophistry at best. Any of the usual understanding of yardsticks provided by the correlation or its square, the proportion of shared variance, are inappropriate. In fact, as noted by Vul et al. (2009), these inflated mean correlations typically exceed the upper bounds provided by the correction for attenuation based on what the reliabilities should be for the measures being correlated.

An amusing critique of fMRI studies that fail to correct for multiple comparisons and control false positives involves the scan of a dead salmon's brain and its response to human emotions ("Trawling

Diseases" (Regina Nuzzo, *ScienceNews*, June 21, 2008).

the Brain," Laura Sanders, December 19, 2009, *ScienceNews*). The original article was published in the *Journal of Serendipitous and Unexpected Results* (Craig Bennett, et al., 2010, *1*, 1–6), with the long title "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction." This tongue-in-cheek piece provides a cautionary lesson for anyone involved with the interpretation of fMRI research. A dead salmon's brain can display much of the same beautiful red-hot areas of activity in response to emotional scenes flashed to the (dead) salmon that would be expected for (alive) human subjects. We give the abstract below.

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of at least one false positive is almost certain. Proper correction for multiple comparisons should be completed during the analysis of these datasets, but is often ignored by investigators. To highlight the danger of this practice we completed an fMRI scanning session with a post-mortem Atlantic Salmon as the subject. The salmon was shown the same social perspective taking task that was later administered to a group of human subjects. Statistics that were uncorrected for multiple comparisons showed active voxel clusters in the salmon's brain cavity and spinal column. Statistics controlling for the familywise error rate (FWER) and false discovery rate (FDR) both indicated that no active voxels were present, even at relaxed statistical thresholds. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum cluster sizes ($k > 8$) is an ineffective control for multiple comparisons. We further argue that the vast majority of fMRI studies should be utilizing proper multiple comparisons correction as standard practice when thresholding their data.

For conducting the "dead-salmon" study, the main authors, Craig Bennett and Michael Miller, received a 2012 Ig Nobel prize. They

were interviewed shortly thereafter by Scott Simon for NPR's Weekend Edition. The transcript of this interview follows:

Host Scott Simon speaks with Craig Bennett and Michael Miller about being awarded a 2012 Ig Nobel prize for their paper on the brain waves of dead Atlantic Salmon, published in the *Journal of Serendipitous and Unexpected Results*.

SCOTT SIMON, HOST:

In a couple weeks, the prestigious Nobel Prizes will be announced. But this week, the Ig Nobels honored the silliest discoveries of 2012. A study on the physics of the ponytail; a paper on why coffee spills when you walk; and a prize for a group of psychologists who scanned the brain of an unpromising patient: a deceased Atlantic salmon. Even more unlikely were their findings: the dead fish had thoughts. Who knows – maybe dreams. Craig Bennett did the experiment and accepted the award with good humor, and a couple of fish jokes.

CRAIG BENNETT: Some have called functional neuroimaging, which is an important method for studying the human brain, a fishing expedition. Some have even called the results a red herring. But ...

SIMON: Craig Bennett and his colleague, Dr. Michael Miller, joins us now from studios at Harvard University. Gentlemen, thanks for being with us.

MICHAEL MILLER: Thank you, Scott.

: Yeah, it's good to be here.

SIMON: Is there any defensible reason to study the brain of a dead fish?

MILLER: Well, not for genuine, functional brain activities there's not.

: We wanted to illustrate kind of the absurdity of improper statistical approaches, that you can find false positives, or what is essentially garbage results. And using the incorrect statistical approach you can actually see that there are voxels of activity in the dead, frozen salmon's brain.

MILLER: You know, while the salmon was in the scanner, we were doing the testing exactly like a human would have been in there.

SIMON: I'm sorry, did you say to the postmortem salmon, just press this button in case you get antsy?

: We actually did, because we were also training our research assistants on the proper methods on how to interact with humans. And so not only did we give the experimental instructions to the salmon but we also were on the intercom asking if the salmon was OK throughout the experiment.

SIMON: Did you just go into Legal Seafood and say give me a mackerel - forgive me, an Atlantic salmon?

MILLER: It was a Saturday morning and we were conducting the testing very early so that we didn't interrupt the running of humans later in the day. So, I walked into the local supermarket at 6:30 in the morning, and I said, excuse me, gentlemen, I need a full-length Atlantic salmon. And I'm not a morning person, I just kind of added - for science. And they kind of looked at me funny, but then they were like, you know, we'll be happy to oblige. That'll be $27.50, and before I knew it, I had a full-length Atlantic salmon that was ready to scan.

SIMON: Gentlemen, I'm sorry if this question sounds indelicate, but when your experimentation was done, grilled or poached?

: Baked. That was dinner that night.

(LAUGHTER)

SIMON: Well, science was served, I expect, right?

: And science was tasty.

SIMON: Craig Bennett and Michael Miller, University of California Santa Barbara, won the Ig Nobel Prize this week. They joined us from Harvard University. Gentlemen, thanks for being with us.

MILLER: Thank you, Scott.

: Thanks.

SIMON: You can hear more highlights from the Ig Nobel Awards later this fall on a special Thanksgiving edition of NPR's SCIENCE FRIDAY. This is NPR News.

There are several ways to do corrections for multiple comparisons in fMRI. One is through the false discovery method already mentioned (e.g., Benjamini and Hochberg, 1995); another is the class of methods that control the familywise error rate which includes the

Bonferroni correction strategy, random field theory, and a general method based on permutation procedures. This later approach is discussed in detail in "Nonparametric Permutation Tests For Functional Neuroimaging: A Primer with Examples" (Thomas E. Nichols and Andrew P. Holmes; *Human Brain Mapping*, *15*, 2001, 1–25); the abstract for this paper follows:

Requiring only minimal assumptions for validity, nonparametric permutation testing provides a flexible and intuitive methodology for the statistical analysis of data from functional neuroimaging experiments, at some computational expense. ... [T]he permutation approach readily accounts for the multiple comparisons problem implicit in the standard voxel-by-voxel hypothesis testing framework. When the appropriate assumptions hold, the nonparametric permutation approach gives results similar to those obtained from a comparable Statistical Parametric Mapping approach using a general linear model with multiple comparisons corrections derived from random field theory. For analyses with low degrees of freedom, such as single subject PET/SPECT experiments or multi-subject PET/SPECT or fMRI designs assessed for population effects, the nonparametric approach employing a locally pooled (smoothed) variance estimate can outperform the comparable Statistical Parametric Mapping approach. Thus, these nonparametric techniques can be used to verify the validity of less computationally expensive parametric approaches. Although the theory and relative advantages of permutation approaches have been discussed by various authors, there has been no accessible explication of the method, and no freely distributed software implementing it. Consequently, there have been few practical applications of the technique. This article, and the accompanying MATLAB software, attempts to address these issues. The standard nonparametric randomization and permutation testing ideas are developed at an accessible level, using practical examples from functional neuroimaging, and the extensions for multiple comparisons described. Three worked examples from PET and fMRI are presented, with discussion, and comparisons with standard parametric approaches made where appropriate. Practical considerations are given throughout, and rele-

vant statistical concepts are expounded in appendices.

4 Cautionary Summary Comments

As a reminder of the ubiquitous effects of searching/selecting/optimization, and the identification of "false positives," we have mentioned some blatant examples here and in earlier modules—the weird neuroscience correlations; the small probabilities (mis)reported in various legal cases (such as the Dreyfus small probability for the forgery coincidences, or that for the de Berk hospital fatalities pattern); repeated clinical experimentation until positive results are reached in a drug trial—but there are many more situations that would fail to replicate. We need to be ever-vigilant of results obtained by "culling" and then presented as evidence.

A general version of the difficulties encountered when results are culled is labeled the *file-drawer problem*. This refers to the practice of researchers putting away studies with negative outcomes (that is, studies not reaching reasonable statistical significance or when something is found contrary to what the researchers want or expect, or those rejected by journals that will consider publishing only articles demonstrating significant positive effects). The file-drawer problem can seriously bias the results of a meta-analysis, particularly if only published sources are used (and not, for example, unpublished dissertations or all the rejected manuscripts lying on a pile in someone's office). We quote from the abstract of a fairly recent review, "The Scientific Status of Projective Techniques" (Lilienfeld, Wood, & Garb, 2000):

Although some projective instruments were better than chance at detecting child sexual abuse, there were virtually no replicated findings across independent investigative teams. This meta-analysis also provides the first clear evidence of substantial file-drawer effects in the projectives literature, as the effect sizes from published studies markedly exceeded those from unpublished studies. (p. 27)

The general failure to replicate is being continually (re)documented both in the scientific literature and in more public venues. In medicine, there is the work of John Ioannidis:

"Contradicted and Initially Stronger Effects in Highly Cited Clinical Research" (*Journal of the American Medical Association*, 2005, *294*, 218–228);
"Why Most Published Research Findings Are False" (*PLoS Medicine*, 2005, *2*, 696–701).
"Why Most Discovered True Associations Are Inflated" (*Epidemiology*, 2008, *19*, 640–648).[4]

[4]This particular Ioannidis article covers much more than just the field of medicine; its message is relevant to the practice of probabilistic reasoning in science more generally. The abstract follows:

Newly discovered true (non-null) associations often have inflated effects compared with the true effect sizes. I discuss here the main reasons for this inflation. First, theoretical considerations prove that when true discovery is claimed based on crossing a threshold of statistical significance and the discovery study is underpowered, the observed effects are expected to be inflated. This has been demonstrated in various fields ranging from early stopped clinical trials to genome-wide associations. Second, flexible analyses coupled with selective reporting may inflate the published discovered effects. The vibration ratio (the ratio of the largest vs. smallest effect on the same association approached with different analytic choices) can be very large. Third, effects may be inflated at the stage of interpretation due to diverse conflicts of interest. Discovered effects are not always inflated, and under some circumstances may be deflated – for example, in the setting of late discovery of associations in sequentially accumulated overpowered evidence, in some types of misclassification from measurement error, and in conflicts causing reverse biases. Finally, I discuss potential approaches to this problem. These include being cautious about newly discovered effect

In the popular media, we have the discussion of the "decline effect" by Jonah Lehrer in the *New Yorker* (December 13, 2010), "The Truth Wears Off (Is These Something Wrong With the Scientific Method?)"; or from one of the nation's national newspapers, "Low-Salt Diet Ineffective, Study Finds. Disagreement Abounds" (*New York Times*, Gina Kolata, May 3, 2011). We give part of the first sentence of Kolata's article: "A new study found that low-salt diets increase the risk of death from heart attacks and strokes and do not prevent high blood pressure."

The subtle effects of culling with subsequent failures to replicate can have serious consequences for the advancement of our understanding of human behavior. A recent important case in point involves a gene–environment interaction studied by a team led by Avshalom Caspi (Caspi et al., 2003). A polymorphism related to the neurotransmitter serotonin was identified that apparently could be triggered to confer susceptibility to life stresses and resulting depression. Needless to say, this behavioral genetic link caused quite a stir in the community devoted to mental health research. Unfortunately, the result could not be replicated in a subsequent meta-analysis (could this possibly be due to the implicit culling over the numerous genes affecting the amount of serotonin in the brain?). Because of the importance of this cautionary tale for behavioral genetics research generally, we reproduce below a *News of the Week*

sizes, considering some rational down-adjustment, using analytical methods that correct for the anticipated inflation, ignoring the magnitude of the effect (if not necessary), conducting large studies in the discovery phase, using strict protocols for analyses, pursuing complete and transparent reporting of all results, placing emphasis on replication, and being fair with interpretation of results.

item from *Science*, written by Constance Holden (2009), "Back to the Drawing Board for Psychiatric Genetics":[5]

Geneticists have long been immersed in an arduous and largely fruitless search to identify genes involved in psychiatric disorders. In 2003, a team led by Avshalom Caspi, now at Duke University in Durham, North Carolina, finally landed a huge catch: a gene variant that seemed to play a major role in whether people get depressed in response to life's stresses or sail through. The find, a polymorphism related to the neurotransmitter serotonin, was heralded as a prime example of "gene-environment interaction": whereby an environmental trigger influences the activity of a gene in a way that confers susceptibility. "Everybody was excited about this," recalls Kathleen Merikangas, a genetic epidemiologist at the National Institute of Mental Health (NIMH) in Bethesda, Maryland. "It was very widely embraced." Because of the well-established link between serotonin and depression, the study offered a plausible biological explanation for why some people are so much more resilient than others in response to life stresses.

But an exhaustive new analysis published last week in *The Journal of the American Medical Association* suggests that the big fish may be a minnow at best.

In a meta-analysis, a multidisciplinary team headed by Merikangas and ge-

[5]The general problem of exaggerated initially-found effects for a marker-allele association is discussed by Peter Kraft in his article "Curses – Winner's and Otherwise – in Genetic Epidemiology" (*Epidemiology*, 2008, *19*, 649–651). The abstract follows:

The estimated effect of a marker allele from the initial study reporting the marker-allele association is often exaggerated relative to the estimated effect in follow-up studies (the "winner's curse" phenomenon). This is a particular concern for genome-wide association studies, where markers typically must pass very stringent significance thresholds to be selected for replication. A related problem is the overestimation of the predictive accuracy that occurs when the same data set is used to select a multilocus risk model from a wide range of possible models and then estimate the accuracy of the final model ("over-fitting"). Even in the absence of these quantitative biases, researchers can over-state the qualitative importance of their findings – for example, by focusing on relative risks in a context where sensitivity and specificity may be more appropriate measures. Epidemiologists need to be aware of these potential problems: as authors, to avoid or minimize them, and as readers, to detect them.

neticist Neil Risch of the University of California, San Francisco, reanalyzed data from 14 studies, including Caspi's original, and found that the cumulative data fail to support a connection between the gene, life stress, and depression. It's "disappointing—of all the [candidates for behavior genes] this seemed the most promising," says behavioral geneticist Matthew McGue of the University of Minnesota, Twin Cities.

The Caspi paper concluded from a longitudinal study of 847 New Zealanders that people who have a particular variant of the serotonin transporter gene are more likely to be depressed by stresses, such as divorce and job loss (*Science*, 18 July 2003, pp. 291–293; 386–389). The gene differences had no effect on depression in the absence of adversity. But those with a "short" version of the gene—specifically, an allele of the promoter region of the gene—were more likely to be laid low by unhappy experiences than were those with two copies of the "long" version, presumably because they were getting less serotonin in their brain cells.

Subsequent research on the gene has produced mixed results. To try to settle the issue, Merikangas says, "we really went through the wringer on this paper." The group started with 26 studies but eliminated 12 for various reasons, such as the use of noncomparable methods for measuring depression. In the end, they reanalyzed and combined data from 14 studies, including unpublished data on individual subjects for 10 of them.

Of the 14 studies covering some 12,500 individuals, only three of the smaller ones replicated the Caspi findings. A clear relationship emerged between stressful happenings and depression in all the studies. But no matter which way they sliced the accumulated data, the Risch team found no evidence that the people who got depressed from adverse events were more likely to have the suspect allele than were those who didn't.

Caspi and co-author Terrie Moffitt, also now at Duke, defend their work, saying that the new study "ignores the complete body of scientific evidence." For example, they say the meta-analysis omitted laboratory studies showing that humans with the short allele have exaggerated biological stress responses and are more vulnerable to depression-related disorders such as anxiety and posttraumatic stress disorder. Risch concedes that his team had to omit several supportive studies. That's because, he says, they wanted to focus as

much as possible on attempts to replicate the original research, with comparable measures of depression and stress.

Many researchers find the meta-analysis persuasive. "I am not surprised by their conclusions," says psychiatric geneticist Kenneth Kendler of Virginia Commonwealth University in Richmond, an author of one of the supportive studies that was excluded. "Gene discovery in psychiatric illness has been very hard, the hardest kind of science," he says, because scientists are looking for multiple genes with very small effects.

Dorrett Boomsma, a behavior geneticist at Amsterdam's Free University, points out that many people have questioned the Caspi finding. Although the gene was reported to have an effect on depression only in the presence of life stress, she thinks it is "extremely unlikely that it would not have an independent effect" as well. Yet recent whole-genome association studies for depression, for which scientists scan the genomes of thousands of subjects for tens of thousands of markers, she adds, "do not say anything about [the gene]."

Some researchers nonetheless believe it's too soon to close the book on the serotonin transporter. ... geneticist Joel Gelernter of Yale University agrees with Caspi that the rigorous demands of a meta-analysis may have forced the Risch team to carve away too much relevant material. And NIMH psychiatrist Daniel Weinberger says he's not ready to discount brain-imaging studies showing that the variant in question affects emotion-related brain activity.

Merikangas believes the meta-analysis reveals the weakness of the "candidate gene" approach: genotyping a group of subjects for a particular gene variant and calculating the effect of the variant on a particular condition, as was done in the Caspi study. "There are probably 30 to 40 genes that have to do with the amount of serotonin in the brain," she says. So "if we just pull out genes of interest, ... we're prone to false positives." Instead, she says, most geneticists recognize that whole-genome scans are the way to go. McGue agrees that behavioral gene hunters have had to rethink their strategies. Just in the past couple of years, he says, it's become clear that the individual genes affecting behavior are likely to have "much, much smaller effects" than had been thought.

References

[1] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289–300.

[2] Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington H. L., ... Poulton, R. (2003). Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science*, *301*, 386–399.

[3] Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest*, *1*, 27–66.

[4] Swets, J. A, Dawes, R. M., & Monahan, J. (2000b). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26.

[5] Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.