What is the "Multiple Testing Problem" and what can we do about it?

Susan E. Hodge[1], Lisa J. Strug[2]

Based on a talk given at SISMEC in Pavia, Italy, September 2009
[1] Division of Statistical Genetics, Department of Biostatistics, Columbia Mailman School of Public Health; Department of Psychiatry, Columbia School of Physicians & Surgeons; and Division of Epidemiology, New York State Psychiatric Institute, New York, NY, USA
[2] Child Health Evaluative Sciences, The Hospital for Sick Children, and The Dalla Lana School of Public Health, University of Toronto

This work supported in part by grant #MH-48858 from the National Institute of Mental Health (SEH) and by grant #HG-004314 from the National Institutes of Health, The Natural Sciences and Engineering Research Council of Canada, and the Early Researcher award program of the Ontario Ministry of Research and Innovation LJS).

Corresponding Author:
Dr. Susan E. Hodge
NYSPI, Unit 24, 1051 Riverside Drive, New York, NY 10032, USA
Phone: (212) 543-5606 - fax (212) 368-3534
e-mail: seh2@columbia.edu

Summary
In this paper, (1) we review what the statistical "multiple testing problem" is, for example, as it applies to genomewide linkage or association analysis of human genetic diseases. (2) Then we describe a different paradigm for statistical inference – the "Evidential" paradigm, as developed and advocated by R. Royall. A feature of this paradigm is that one *decouples* the *measure of evidence* from the *error probabilities*, and we explain what is meant by each of the italicized terms. (3) We show how the core root of the multiple testing problem is precisely the confounding of error probabilities with evidence measures. Thus, the Evidential paradigm, since it separates those two concepts, can help us deal with the multiple testing problem in a more logically consistent way.

KEY WORDS: *multiple testing, Evidentialism, statistical paradigms*

1. Introduction

By way of introduction, suppose two collaborators are interested in the possible genetic contributions to disease X. Investigator 1 has her favourite gene G, which is the only gene she is investigating, whereas investigator 2 has no *a priori* hypotheses about genetic contributors. Also suppose that genome-wide SNP data are available. Investigator 1 examines the association between disease X and gene G and observes positive evidence, with a *p*-value of 0.001, which she interprets as "strong" evidence. Investigator 2 examines association between disease X and each of 550,000 genetic markers; he observes positive evidence of association with that same gene G, and with the same *p*-value of 0.001, but he interprets this as only "weak" evidence.

This simple example illustrates how the same *evidence* can lead to different conclusions, – "strong" vs. "weak" in the above example – and thus provides an intuitively appealing example of the multiple testing problem, one we see often in human genetics. This situation leads us into logical paradoxes and inconsistencies. However, thinking about statistical evidence in a different way – from the perspective of the "evidential" paradigm – can help resolve these paradoxes. In this brief paper we illustrate these principles with genetic examples.

2. Evidential Paradigm

The fundamental feature and appeal of the "evidential" paradigm, in the context of multiple testing, is that it "decouples" "measures of evidence" from "error probabilities." In the next three subsections, we explain what we mean by each of these terms – measures of evidence (Sec. 2.1), error probabilities (Sec. 2.3), and decoupling (Sec. 2.3).

Measures of Evidence

We define the "likelihood ratio" (LR) as our measure of evidence, where the LR is defined as the likelihood ratio between two simple hypotheses, namely,

$$LR \equiv \frac{L(Hypothesis\ 1 \mid data)}{L(Hypothesis\ 2 \mid data)} = \frac{L(H_1)}{L(H_0)}.$$

The likelihood is technically defined as "proportional to probability" up to a multiplicative constant [1]. The point is that the *ratio* of likelihoods between two simple hypotheses indicates which hypothesis is better supported by the observed data. If the data we observe are more rare (i.e., much more "surprising") under one hypothesis than another, then the LR reflects that fact. This is explicitly stated by the well-known *Law of Likelihood* [2-3].

For a simple example, imagine that across the hall from your office is a lab that is shared by two colleagues, Amy and Robert. When either one is working there alone, that person listens to music on the radio. If Amy is there alone, she listens about half the time (50%) to classical music, and 50% to jazz, whereas Robert greatly prefers jazz, which he chooses 98% of the time, listening to classical music only 2% of the time. You come in to work one morning and hear classical music coming out of the lab. Which colleague is more likely to be in the lab? The answer of course is Amy, because the observation of classical music is much less rare or surprising if Amy is there than if Robert is there. A formal analysis could proceed as follows:

H_1: Amy is in lab, vs. H_0: Robert is in lab.

$$LR = \frac{L(Amy\ hypothesis \mid data)}{L(Robert\ hypothesis \mid data)} = \frac{.50}{.02} = 25.$$

This value of 25 for the LR can be interpreted as, "The Amy hypothesis is 25 times 'more likely' than the Robert hypothesis," or, equivalently, "The observation of classical music is 25 times 'less surprising' if the Amy hypothesis is true than if the Robert hypothesis is true."

These concepts form, indirectly, the basis of every statistical test or procedure and are explicitly integral to genetic linkage studies. Here is an example from linkage analysis. The goal is to determine whether the disease locus of interest is "linked" to a known marker locus. By "linked" we mean that the disease and marker loci are located on the same chromosome and are reasonably close together. We measure the distance between any two loci by what is called the recombination fraction (usually denoted θ). From biology we know that a value of $\theta = 0.5$ corresponds to independent assortment (Mendel's Second Law) and thus represents *lack* of linkage, whereas values of $\theta < 0.5$ represent linkage. In current genetic research, we usually work with smaller values of θ, such as $\theta < 0.1$ and even $\theta < 0.01$, since we currently have so many markers available in the human genome. Thus, we might set up these two hypotheses:

H_0: $\theta = 0.5$ (i.e., no linkage) vs. H_1: $\theta = 0.05$.

The corresponding likelihood ratio is:

$$LR = \frac{L(\theta = 0.05)}{L(\theta = 0.5)}$$

As a simple example, say we had observed a family with 10 children, *none* of them representing recombination from the parents. If H_1 is true, the likelihood (probability) of 10 nonrecombinant children is given by the probability of a single nonrecombinant ($1-\theta$, i.e., 0.95), raised to the 10th power, i.e., $(.95)^{10}$. In contrast, under the null hypothesis of no linkage, that probability is simply one-half raised to the 10th power, or $(.5)^{10}$. Thus the ratio of these two probabilities, the LR is

$$LR = \frac{(.95)^{10}}{(.5)^{10}} = 613.1$$

This says that the hypothesis of linkage with θ=.05 is 613 times more likely than hypothesis of no-linkage. Another way to express this is that it would be 613 times "more surprising" or "more rare" to observe this family if there is no linkage than if there is linkage with $\theta = 0.05$.

For convenience, geneticists work with the logs of the LRs, rather than the LRs themselves. The lod score is defined as $\text{Lod}(\theta) \equiv \log_{10}(\text{LR for that value of } \theta)$. In the above example, $\text{Lod}(\theta{=}.05) = \log_{10}(613.1) = 2.79$.

To use the LR as a measure of evidence, we need to set some criterion or cutoff value that will represent convincing evidence favouring one hypothesis over another. To that end, we choose a value of "k" (where $k > 1$) and agree that a LR $\geq k$ represents "strong" evi-

dence in favor of H_1, and that LR ≤ 1/*k* represents "strong" evidence in favor of H_0. If the LR falls between those two values, that represents "weak" evidence. Possible values of *k* could be 8, 32, 100, 1,000, and each of these has some historical or scientific application. Traditionally in linkage analysis, a Lod score of 3.0 (i.e., LR of 1,000) has been to taken to constitute proof of linkage.

Error probabilities

But if we are to accept using the LR as an evidence measure, we need justification. How does this measure of evidence *behave*? A reliable measure of evidence is one that leads users to draw incorrect conclusions with low probability (i.e., one that has good operating characteristics). Royall [4] has shown that the LR has this property, even if we choose relatively small values of *k* (e.g., *k* = 32).

To show this, define error probabilities, as follows: First, say H_0 is true (no linkage). There are three possible outcomes for our data:

- If LR ≥ *k*, that is a "misleading" outcome, so the probability of this happening, $P[LR \geq k|H_0]$, is an error probability, which we want to minimize.
- If LR ≤ 1/*k*, that is an outcome that leads to the correct conclusion. We call this a "correct" or "strong" outcome, and we want to maximize the probability, $P[LR \leq 1/k|H_0]$, that this will happen.
- What if the LR fall between those two values, i.e., $1/k < LR < k$? This represents an inconclusive outcome; i.e., the data do not give us a clear answer in either direction. We refer to this as a "weak" outcome. It is not exactly an "error" situation, since it does not lead us to an incorrect conclusion, but we still want to minimize its probability, i.e., $P[1/k < LR < k|H_0]$.

Similarly, if H_1 is true (i.e., there is linkage or association), we define the same three outcomes, but now the interpretations are reversed:

- $P[LR \leq 1/k |H_1]$ is an error probability, which we want to minimize.
- $P[LR \geq k|H_1]$ is the probability that the data lead us to the correct conclusion, so we want to maximize it.
- $P[1/k < LR < k|H_1]$ is still the probability of weak evidence and should be minimized.

These probabilities are functions of three quantities: sample size; the choice of alternative hypothesis; and the criterion, *k*. The error probabilities can be controlled by adjusting these values.

These error probabilities are analogous in principle to the more standard frequentist error rates [5]. For example, $P[LR \geq k|H_0]$ mentioned above is analogous to type I error in the classical paradigm, and $P[LR \geq k|H_1]$ is analogous to power. However, we specify and use them differently. Briefly, in the classical paradigm, one collects data, and then one interprets the *p*-value in a dual role as being both an error probability *and* a measure of evidence. In contrast, in the evidential paradigm, one specifies a desirable level of strength of evidence (*k*), then designs one's experiment to maintain acceptable error probabilities. However, after collecting the data, one does not look at the error probabilities any more, but only at the evidence. The error probabilites are relevant only for planning.

To give a homely example: If the weather report in the morning says there is a 75% chance of rain, you decide whether or not to take your umbrella to work, based on that prediction. However, once you've left the house and the day progresses, all that matters is whether it is or is not raining. The 75% prediction is no longer relevant.

Work by Royall [4] analyzes magnitudes of error probabilities when we set *k* at different values, when we vary the alternative hypothesis, and when we control sample size, which can be used to control error. To begin, for a given value of *k*, the error probability cannot exceed 1/*k*, for both $P[LR \geq k|H_0]$ and $P[LR \leq 1/k|H_1]$. This means that if one uses a value of *k* = 32, those errors have an absolute upper bound of 1/32, or ≈0.03. But further, in many cases, with reasonable sample sizes, the maximum value for these probabilities approaches a value much lower than that, namely $\Phi(-\sqrt{2\ln} k)$ for *k* = 32, this value is ≈0.0043. In other words, even *k* as low as 32 can result in very low error probabilities for reasonable sample sizes. (These results apply for simple situations such as outlined here and need to be modified for more complex situations; see also [6-7].)

Since it is relatively easy to keep the probability of misleading evidence low, users of the Evidential Paradigm should focus on lowering the probability of weak evidence, which they can do by a reasonable combination of choosing the alternative hypothesis (H_1) carefully, not setting *k* too high, and increasing sample size.

Our own work [6-7] applies these findings to the specific context of linkage analysis. For example, testing θ = 0.05 vs.θ = 0.50 in a sample of 20 sib pairs, and using *k* = 32 as the criterion, the error probabilities are

only 0.0013 when H_0 is true, 0.0018 when H_1 is true. In other work we have also applied these principles to association analysis [8, 9]. The reader is referred to those papers for more details.

Decoupling

"Decoupling" means that conceptually we *separate* the "error probabilities" from the "measure of evidence." Concretely, this means that we first design our experiment to have acceptable error probabilities. But then, after collecting data, we look only at the measure of evidence, and we do not try to make one quantity serve both functions.

Contrast this approach with that of classical hypothesis testing, where the *p*-value is used as both an error probability and a measure of evidence. The *p*-value is technically defined as the "probability of observing results this deviant or more from H_0, if H_0 is true." However, then this same quantity, the *p*-value, is also used as measure of evidence: Investigators commonly say such things as, "a *p*-value of .04 is acceptable but not very strong evidence, whereas a *p*-value of .0001 is strong evidence," and so on.

There are many reasons why the *p*-value is not in fact a good measure of evidence. Here we give just one simple example; for more detailed discussions, see [10, 11].

Example. Say we are testing whether a coin is fair vs. it has "heads" on both sides. Let q represent the probability of a head on a single toss. Then we can formulate H_0: $q = ½$, vs. H_1: $q = 1$. Now say we toss the coin 10 times, and observe "heads" 9 times, "tails" once. Under H_0: this is a rare event, with corresponding *p*-value = 0.011. [The *p*-value in this case equals the probability, under the null hypothesis, of nine heads, plus the probability of all 10 heads, i.e., $10(.5)^{10} + (.5)^{10} = 0.011$]. So if we went by the *p*-value alone, we would reject the null hypothesis, since $p < 0.05$. On the other hand, if H_1 is true, this outcome is not simply rare, it is *impossible*. Thus, given the choice between the two hypotheses, we must choose H_0, for, as Sherlock Holmes says, "How often have I said to you that when you have eliminated the impossible, whatever remains, *however improbable*, must be the truth?" (quoted in [1]).

To summarize, the Evidential Paradigm requires that one uses the LR as a measure of evidence, knowing that error probabilities can be kept small, and that one separates ("decouples") the measure of evidence from the error probabilities.

3. Back to the Multiple Testing Problem

We maintain that the root of the problem is precisely what was discussed above, namely, confounding the error probability with the measure of evidence. We argue that the advantage of "decoupling" is that it lets us consider the multiple testing problem more logically and consistently.

In the classical paradigm, one runs into the kinds of paradoxes alluded to in the Introduction, namely, that it is reasonable to say this: "In a linkage analysis, a likelihood ratio of, say, 1000:1 at locus *X* is taken to represent different evidence, depending on whether we analyzed only locus *X*; or we analyzed ten candidate loci, including *X*; or we analyzed *X* as part of a genome scan of thousands of loci". However this violates common sense. Surely the *evidence* for linkage remains 1000 to 1 in all three situations. What we should be concerned about is that the error probability may differ in these different situations, not that the evidence may differ.

Expressing the concern in statistical terms: Even though the error in any one test is small, when one performs multiple tests, the probability that at least one of these tests has made an error may be large. To express this more precisely, statisticians define the Family-Wise Error Rate:

$$\text{FWER} = P[\text{at least one test yields LR} \geq k \mid H_0]$$

Then, rather than controlling only the error probability of a single test, they try to control this FWER. However, since they use the *p*-value in the dual role of error probability and measure of evidence, when they control the error probability they are perforce "controlling" the evidence as well.

In contrast, with the Evidential paradigm, one decouples the error probability from the evidence measure. If one has to adjust the error probability to ensure that it is small across multiple tests, one does so; but the evidence is what it is and should not be "modified" or "corrected."

If nothing else, considering the multiple testing paradox from the perspective of the evidential paradigm, as we have outlined here, improves the clarity and precision of thought and of language on this difficult subject. However, one can go further: One can break down the multiple testing problem into two separate situations and say something quantitative about both. Full details are in [7]; here we give a broad outline of these two situations (also see [4] and [12]).

We call the first situation "multiple tests of a single hy-

pothesis". This is the situation in which investigators examine data as they go along, then potentially collect more data, depending on what has been observed so far. It is a sequential approach, and it is in fact what scientists *do*, both in linkage analysis and in most fields of biomedical research. For this situation one can show rigorously that it is legitimate to keep collecting data until the LR gives a clear result, one way or the other. The resultant error probabilities are somewhat higher than they would be for a single test, but they still remain below reasonable upper bounds.

The second situation involves a "single test of of multiple hypotheses". This refers to the familiar situation of, for example, genome scans, in which one analysis (linkage or association) is performed, but that one analysis represents a large number of hypotheses, one for each genetic locus being tested. In this situation there is no absolute upper bound on error probabilities, so more care must be taken. In [7] we explore ways to use sample size to control error probabilities, and we work out efficient ways to use replication as a way to control error probabilities.

As an aside, we note that replication is nothing new; many investigators have argued: "Don't be overly concerned about the *p*-values; focus more on replication." Our work puts that intuitive reaction on a sound logical footing, using the Evidential paradigm.

In summary, we advocate handling the "multiple testing problem" by using the likelihood ratio as a measure of evidence and separating (decoupling) the measure of evidence from the error probabilities. This approach enables investigators to deal with the multiple testing problem more logically and consistently, by separating what they do with the evidence from what they do with the error probabilities.

References

1. Edwards AWF (1992) Likelihood, Expanded Edition. Johns Hopkins, Baltimore.
2. Hogg R, Craig AT (1995) Introduction to Mathematical Statistics. Upper Sattle River: Prentice and Hall.
3. Birnbaum A (1962) On the foundation of statistical inference (with discussion). J Am Stat Assoc 53: 259-326.
4. Royall RM (2000) On the probability of observing misleading statistical evidence (with discussion). J Am Statist Assoc 95: 760–780.
5. Strug LJ, Rohde CA, Corey PN (2007) An introduction to evidential sample size calculations. Am Stat, 61:207-212.
6. Strug LJ, Hodge SE (2006a) An alternative foundation for the planning and evaluation of linkage analysis. I. Decoupling "error probabilities" from "measures of evidence." Hum Hered 61: 166–188.
7. Strug LJ, Hodge SE (2006b) An alternative foundation for the planning and evaluation of linkage analysis. II. Implications for multiple test adjustments. Hum Hered 61: 200–209.
8. Strug LJ, Clarke T, Chiang T, Chien M, Baskurt Z, Li W, Dorfman R, Bali B, Wirrell E, Kugler SL, Mandelbaum DE, Wolf SM, McGoldrick P, Hardison H, Novotny EJ, Ju J, Greenberg DA, Russo JJ, Pal DK (2009) Centrotemporal sharp wave EEG trait in rolandic epilepsy maps to Elongator Protein Complex 4 (ELP4). Eur J Hum Genet 17: 1171-81.
9. Strug LJ, Hodge SE, Chiang T, Pal DK, Corey PN, Rohde C (unpublished observations) A pure likelihood approach to the analysis of genetic association data: An alternative to Bayesian and Frequentist analysis. Eur J Hum Genet, in press
10. Royall R (1994) Statistical Evidence: A Likelihood Paradigm. Chapman & Hall, London.
11. Vieland VJ, Hodge SE (1998) Review of Statistical Evidence: A Likelihood Paradigm, by R. Royall. Am J Hum Genet 63: 283–289.
12. Blume JD (2002) Tutorial in Biostatistics: Likelihood methods for measuring statistical evidence. Stat in Med 21: 2563-2599