Tour III Auditing: Biasing Selection Effects and Randomization

> This account of the rationale of induction is distinguished from others in that it has as its consequences two rules of inductive inference which are very frequently violated . . .The first of these is that the sample must be a random one . . .The other rule is that the character [to be studied] must not be determined by the character of the particular sample taken. (Peirce 1.95)

The biggest source of handwringing about statistical inference these days boils down to the fact it has become very easy to infer claims that have been subjected to insevere tests. High-powered computerized searches and data trolling permit sifting through reams of data, often collected by others, where in fact no responsible statistical assessments are warranted. "We're more fooled by noise than ever before, and it's because of a nasty phenomenon called 'big data'. With big data, researchers have brought cherry picking to an industrial level" (Taleb 2013). Selection effects alter a method's error probabilities and yet a fundamental battle in the statistics wars revolves around their relevance (Section 2.4). We begin with selection effects, and our first stop is to listen in on a court case taking place.

Dr. Paul Hack, CEO of Best Drug Co., is accused of issuing a report on the benefits of drug X that exploits a smattering of questionable research practices (QRPs): It ignores multiple testing, uses data-dependent hypotheses, and is oblivious to a variety of selection effects. What happens shines a bright spotlight on a mix of statistical philosophy and evidence. The case is fictional; any resemblance to an actual case is coincidental.

Round 1 The prosecution marshals their case, calling on a leader of an error statistical tribe who is also a scientist at Best: *Confronted with a lack of statistically significant results on any of 10 different prespecified endpoints (in randomized trials on Drug X), Dr. Hack proceeded to engage in a post-data dredging expedition until he unearthed a subgroup wherein a nominally statistically significant benefit [B] was found. That alone was the basis for a report to share-holders and doctors that Drug X shows impressive benefit on factor B.* Colleagues called up by the prosecution revealed further details: *Dr. Hack had ordered his chief data analyst to "shred and dice the data into tiny julienne slices until he got some positive results" sounding like the adman for*

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

Chop-o-Matic. The P-value computed from such a post-hoc search cannot be regarded as an actual P-value, yet Dr. Hack performed no adjustment of P-values, nor did he disclose the searching expedition had been conducted. Moreover, we learn, *the primary endpoint describing the basic hypothesis about the mechanism by which Drug X might offer benefits attained a non-significant P-value of 0.52. Despite knowing the FDA would almost certainly not approve drug X based on post-hoc searching, Dr. Hack optimistically reported on profitability for Best, thanks to the "positive" trials on drug X.*

Anyone who trades Biotech stocks knows that when a company reports: 'We failed to meet primary and perhaps secondary endpoints,' the stock is negatively affected, at times greatly. When one company decides to selectively report or be overly optimistic, it's unfair to patients and stockholders.

Round 2 Next to be heard from are defenders of Dr. Hack: *There's no need to adjust for post-hoc data dredging, the fact that significance tests require such adjustments is actually one of their big problems. What difference does it make if Dr. Hack intended to keep trying and trying again until he found something? Intentions are irrelevant to the import of data.* Others insist that: *the position on cherry picking is open to debate, depending on one's philosophy of evidence. For the courts to take sides would set an ominous precedent.* They cite upstanding statisticians who can attest to the irrelevance of such considerations.

Round 3 A second wave of Hack's defenders (which could be the same as in Round 2) pile on, with a list of reasons for *P*-phobia: *Significance levels exaggerate evidence, force us into dichotomous thinking, are sensitive to sample size, aren't measures of evidence because they aren't comparative reports, and violate the likelihood principle.* Even expert prosecutors, they claim, construe a *P*-value as the probability the results are due to chance, which is to commit the prosecutor's fallacy (misinterpreting *P*-values as posterior probabilities), so they are themselves confused.

Dr. Hack's lawyer jumps at the opening before him: *You see there is disagreement among scientists, at a basic philosophical level. To hold my client accountable would be no different than banning free and open discussion of rival interpretations of data amongst scientists.*

Dr. Hack may not get off the hook in the end – at least in fields where best practice manuals encode taking account of selection effects – but that could change if enough people adopt the stance of friends of Hack. In any event, it is not too much of a caricature of actual debates taking place. You, the citizen scientist, have the tough job of sifting through the cacophony. Severity principle in hand, you can at least decipher where the controversy is coming from. To limber up you might disinter the sources of claims of Round 3.

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

4.6 Error Control is Necessary for Severity Control

> To base the choice of the test of a statistical hypothesis upon an inspection of the observations is a dangerous practice; a study of the configuration of a sample is almost certain to reveal some feature, or features, which are exceptional if the [chance] hypothesis is true. (Pearson and Chandra Sekar 1936, p. 127)

> The likelihood principle implies. . .the irrelevance of predesignation, of whether an hypothesis was thought of beforehand or was introduced to explain the known effects. (Rosenkrantz 1977, p. 122)

Here we encounter the same source of tribal rivalry first spotted in Excursion I with optional stopping (Tour II). Yet we also allowed that data dependencies, double counting and non-novel results are not always problematic. The advantage of the current philosophy of statistics is that it makes it clear that the problem – *when it is a problem* – is that these gambits alter how well or severely probed claims are. We defined problematic cases as those where data or hypotheses are selected or generated, or a test criterion is specified, in such a way that the minimal severity requirement is violated, altered (without the alteration being mentioned), or unable to be assessed (Section 2.4).

Because they alter the severity, they must be taken account of in auditing a result, which includes checking for (i) selection effects, (ii) violations of model assumptions, and (iii) obstacles to any move from statistical to substantive causal or other theoretical claims.

There is no point in raising thresholds for significance if your methodology does not pick up on biasing selection effects. Yet, surprisingly, that is the case for many of the methods advocated by critics of significance tests, and related error statistical methods.

> Two problems that plague frequentist inference: multiple comparisons and multiple looks, or, as they are more commonly called, *data dredging* and peeking at the data. The frequentist solution to both problems involves adjusting the *P* value . . . But adjusting the measure of evidence because of considerations that have nothing to do with the data defies scientific sense, belies the claim of 'objectivity' that is often made for the *P* value. (Goodman 1999, p. 1010)

This is epidemiologist Steven Goodman.[1] To his credit, he recognizes the philosophical origins of his position.

Older arguments from Edwards, Lindman, and Savage (1963) (E, L, & S) live on in contemporary forms. The Bayesian psychologist Eric-Jan Wagenmakers tells us:

[1] Currently a co-director of the Meta-Research Innovation Center at Stanford (METRICS).

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

> [I]f the sampling plan is ignored, the researcher is able to always reject the null hypothesis, even if it is true. This example is sometimes used to argue that any statistical framework should somehow take the sampling plan into account. Some people feel that 'optional stopping' amounts to cheating . . . This feeling is, however, contradicted by a mathematical analysis. (2007, p. 785)

Being contradicted by mathematics is a heavy burden to overcome. Look closely and you'll see we are referred to E, L, & S. But the "proof" assumes the Likelihood Principle (LP) by which error probabilities drop out (Section 1.5). Error probabilities and severity are altered, but if your account has no antennae to pick up on them, then, to you, there's no effect.

Holders of the LP point fingers at error statisticians for worrying about "the sample space" and "intentions." To leaders of movements keen to rein in researcher flexibility, by contrast, a freewheeling attitude toward data-dependent hypotheses and stopping rules is pegged as a major source of spurious significance levels. Simmons, Nelson, and Simonsohn (2011) list as their first requirement for authors: "Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article" (p. 1362). I'd relax it a little, requiring they report how their stopping plan alters the relevant error probability. So let me raise an "either or" question: Either your methodology picks up on influences on error probing capacities of methods or it does not. If it does, then you are in sync with the minimal severity requirement. We may compare our different ways of satisfying it. If it does not, then we've hit a crucial nerve. If you care, but your method fails to reflect that concern, then a supplement is in order. Opposition in methodology of statistics is fighting over trifles if it papers over this crucial point. If there is to be a meaningful "reconciliation," it will have to be here.

I'm raising this in a deliberate stark, provocative fashion to call attention to an elephant in the room (or on our ship). The widespread concern of a "crisis of replication" has nearly everyone rooting for predesignation of aspects of the data analysis, but if they're not also rooting for error control, what are they cheering for? Let's allow there are other grounds to champion predesignation, so don't accuse me of being unfair . . . *yet*. The tester sticks her neck out: she requires it to have a direct effect on inferential measures.

Paradox of Replication

> We often hear it's too easy to obtain small p-values, yet replication attempts find it difficult to get small p-values with preregistered results. This shows the problem isn't p-values but failing to adjust them for cherry picking, multiple testing, post-data subgroups and other *biasing selection effects*. (Mayo 2016, p. 1)

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

Criticism assumes a standard. If you criticize hunting and snooping because of the lack of control of false positives (Type I errors), then the assumption is that those matter. Suppose someone claims it's too easy to satisfy standard significance thresholds, while chiming in with those bemoaning the lack of replication of statistically significant results.

Critic of tests: It's too easy to get low significance levels.
You: Why is it so hard to get small P-values in replication research?

Wait for their answer.

Critic: Aside from expected variability in results, there was likely some P-hacking, cherry picking, and other QRPs in the initial studies.
You: So, I take it you want methods that pick up on these biasing effects, and you favor techniques to check or avoid them.
Critic: Actually I think we should move to methods where selection effects make no difference (Bayes factors, Bayesian posteriors).

Now what? One possibility is a belief in magical thinking, that ignoring biasing effects makes them disappear. That doesn't feel very generous. Or maybe the unwarranted effects really do disappear. Not to the tester. Consider the proposals from Tour II: Bayes ratios, with or without their priors. Imagine my low P-value emerged from the kind of data dredging that threatens the P-value's validity. I go all the way to getting a z-value of 2.5, apparently satisfying Johnson's approach. I erect the maximally likely alternative, it's around 20 times more likely than the null, and am entitled, on this system, to infer a posterior of 0.95 on H_{max}. Error statistical testers would complain that the probability of finding a spurious P-value this way is high; if they are right, as I think they are, then the probability of finding a spurious posterior of 0.95 *is just as high*. That is why Bayes factors are at odds with the severity requirement. I'm not saying in principle they couldn't be supplemented in order to control error probabilities – nor that if you tell Bayesians what you want, they can't arrange it (with priors, shrinkage, or what have you). I'm just saying that the data-dredged hypothesis that finds its way into a significance test can also find its way into a Bayes factor. There's one big difference. I have error statistical grounds to criticize the former. If I switch tribes to one where error probabilities are irrelevant, my grounds for criticism disappear.

The valid criticism of our imaginary Dr. Hack, in Round 1, is this: he purports to have found an effect that would be difficult to generate if there were no genuine discrepancy from the null, when in fact it is easy to generate it. It is frequently brought about in a world where the null hypothesis is true.

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

The American Statistical Association's statement on *P*-values (2016, p. 131) correctly warns, "[c]onducting multiple analyses of the data and reporting only those with certain *p*-values" leads to spurious statistical levels. Their validity is lost, and the alarm goes off when we audit. When Hack's defenders maintain that scientists should not be convicted for engaging in the all-important task of exploration, you can readily agree. But you can still insist the results of explorations be independently tested or separately defended. If you're not controlling error probabilities, however, there's no alarm bell. This leads to the next group, Round 2, declaring that it makes no sense to adjust measures of evidence "because of considerations that have nothing to do with the data," thereby denying the initial charge against poor Dr. Hack, or should I say, lucky Dr. Hack, because the whole matter has now come down to something "quasi-philosophical," a murky business at best. Round 3 piles on with '*P*-values are invariably misinterpreted', and no one really likes them much anyway. The *P*-value is the most unpopular girl in the class and I wouldn't even take you to visit *P*-value tribes – or "cults" as some call them[2] – I prefer speaking of observed significance levels, if it weren't that they suddenly have occupied so much importance in the statistics wars.

You might say that even if some people deny that selection effects actually alter the "evidence," the question of whether they can be ignored in interpreting data in legal or policy settings is not open for debate. After all, statutes in law and medicine require taking them into account. For example, the *Reference Manual on Scientific Evidence* for lawyers makes it clear that finding post-data subgroups that show impressive effects – when primary and secondary endpoints are insignificant – calls for adjustments. In a chapter by David Kaye and David Freedman, they emphasize the importance of asking:

> *How many tests have been performed?* Repeated testing complicates the interpretation of significance levels. If enough comparisons are made, random error almost guarantees that some will yield 'significant' findings, even when there is no real effect . . .
>
> If a fair coin is tossed a few thousand times, it is likely that at least one string of ten consecutive heads will appear. . . . Ten heads in the first ten tosses means one thing [evidence the coin is biased]; a run of ten heads somewhere along the way to a few thousand tosses of a coin means quite another. (Kaye and Freedman 2011, pp. 127–8)

Nevertheless, statutes can be changed if their rationale is overturned. The issue is not settled. There are regularly cases where defenders emphasize the lack of consensus and argue precisely as defenders in Round 2.[3]

[2] For example, Ziliak and McCloskey's *The Cult of Significance* (2008a).

[3] A case that went to the Supreme Court of the United States (SCOTUS): "There is no consensus whether, when, or how to adjust p-values or Type I error rates for multiple testing . . . the issue is not settled among scientists." Rothman, Lash, and Schachtman 2013, pp. 21–2.

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

Error Control is Only for Long Runs. But wait a minute. I am overlooking the reasons given for ignoring error control, some even in direct reply to Mayo (2016):

> Bayesian analysis does not base decisions on error control. Indeed, Bayesian analysis does not use sampling distributions … As Bayesian analysis ignores counterfactual error rates, it cannot control them. (Kruschke and Liddell 2017, pp. 13, 15)

This recognition may be a constructive step, admitting lack of error control. But on their view, caring about error control could only matter if you were in the business of performance in the long run. Is this true? By the way, the error control is actual; counterfactuals are used in determining what they are. Kruschke and Liddell continue:

> [I]f the goal is specifically to control the rate of false alarms when the decision rule is applied repeatedly to imaginary data from the null hypothesis, then, by definition, the analyst must compute a *p* value and corresponding CIs. . . . the analyst must take into account the exact stopping and testing intentions. . . . any such procedure does not yield the credibility of parameter values … but instead yields the probability of imaginary data . . .. (ibid., p. 15)

What's this about imaginary data? Suppose a researcher reports data ***x*** claiming to show impressive improvement in an allergic reaction among patients given a drug. They don't report that they dredged different measures of improvement or searched through post-hoc subgroups among patients. The actual data ***x*** only showed positive results on allergic reaction, let's say. Why report imaginary data that could have resulted but didn't? Looking at imaginary data sounds really damning, but when we see what they mean, we're inclined to regard it a scandal to *hide* such information. The problem is not about long runs – lose the alleged time aspect. I can simulate the sampling distribution to see the relative frequencies any day of the week. I can show you if a method had very little capability of reporting bupkis when it should have reported bupkis (nothing). It didn't do its job today. If Kruschke and Liddell's measure of credibility ignores what's regarded as QRPs in common statutes of evidence, you may opt for error control. True, error control is a necessary, not a sufficient, condition for a severity assessment. It follows from the necessity that an account with poor error control can neither falsify nor corroborate with severity.

Direct Protection Requires Skin off Your Nose. Testers are prepared to admit a difference in goals. Rival tribes may say lack of error control is no skin off their noses, since what they want are posterior probabilities and Bayes factors. Severe testers, on the other hand, appeal to statistics for protection

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

against being misled by a cluster of biases and mistakes. A method *directly protects* against a QRP only insofar as its commission *is* skin off its nose. It must show up in the statistical analysis. More than that, there must be a general rationale for the concern. That a Bayesian probabilist avoids selection effects that hurt severity doesn't automatically mean they do so directly because of the severity violation. Moreover, it shouldn't be an option whether to take account of them or not, they must be. The tester holds this even granting there are cases where auditing shows no damage to severity.

Capitalizing on Chance

Gather round to listen to Hanan Selvin, a sociologist writing in 1957, reprinted in Morrison and Henkel (1970):

> [W]hen the hypotheses are tested on the same data that suggested them and when tests of significance are based on such data, then a spurious impression of validity may result. The computed level of significance may have almost no relation to the true level ... Suppose that twenty sets of differences have been examined, that one difference seems large enough to test and that this difference turns out to be 'significant at the 5 percent level.' Does this mean that differences as large as the one tested would occur by chance only 5 percent of the time when the true difference is zero? The answer is *no*, because the difference tested has been *selected* from the twenty differences that were examined. The actual level of significance is not 5 percent, but 64 percent! (Selvin 1970, p. 104)

Selvin would give Dr. Hack a hard time: to ignore a variety of selection effects results in a fallacious computation of the *actual* significance level associated with a given inference. Each of the 20 hypotheses is on different but closely related effects.

Suppose the single property found to have an impressive departure is hypothesis 13 of the 20. The possible results now are the possible factors that might be found to show a 2 standard deviation departure (for a two-sided 0.05 test) from the null. Thus the Type I error probability is the probability of finding at least one such significant difference out of 20, even though all 20 nulls are true. The probability that this procedure yields erroneous rejections differs from, and will be much greater than, 0.05. There are different and many more ways that one can err in this procedure than in testing a single prespecified null hypothesis, and this influences the actual P-value. The *nominal* (or computed) level for H_{13} would ignore the selection and report the P-value associated with a 2 standard deviation difference, 0.05.

Assuming 20 independent samples are drawn from populations having true discrepancies of zero, the probability of attaining at least one nominally statistically significant outcome (in either direction) with N independent

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

tests at the α significance level is $1 - (1 - \alpha)^N$. Test results are treated like Bernoulli trials with probability of "success" (a 0.05 rejection) equal to 0.05 on each trial. The probability of getting no successes in 20 independent trials is $(0.95)^{20}$.

> Pr(Test rejects at least one H_i at level 0.05; all H_i true) $= 1 - (1 - 0.05)^{20}$ $= 0.64$.

This would give the *actual* significance level, if we could assume independence; in practice this wouldn't hold, so it would be a conservative value. This is the experiment-wide significance level or *family-wise error rate* (FWER). In terms of P-values, you'd need to compute the probability that the smallest of 20 is as small as yours. The *Bonferroni correction* can ensure this is at most α^* by setting the α-level for each test at α^*/N. Alternatively the correction may be attained by multiplying the P-value by N. You could hunt through N hypotheses and be an "honest hunter," and report the adjusted P-value. Some find the Bonferroni adjustment too strict, not to mention its assumption of independence is unlikely to hold. There's a large literature, and myriad ways to adjust, appropriate for different problem situations. Juliet Shaffer has done considerable work on this (Shaffer 1995).

The need for an adjustment recognizes how various tactics lead reported error probabilities to mischaracterize how readily the test would alert us to blatant deceptions. Skeptical of the inferred inference, we ask: how frequently would this method have alerted me to erroneous claims of this form? If it would almost never alert me to such mistakes, I deny it is good evidence for this particular claim. This is to use error statistical probabilities to assess severity. Even where we would not place much stock on the precise corrected P-value, it's important to have an alert that the unaudited P-value is invalid or questionable.

This illustrates how error statistical methods directly protect against such biasing selection effects: revealing how we could be fooled, as well as self-correct. For the severe tester, outputting H_{13}, ignoring the non-significant others, renders H_{13} poorly tested. You might say, but look there's the hypothesis H_{13}, and data $\boldsymbol{x}$ – shouldn't it speak for itself? No. That's just the evidential-relationship or logicist in you coming out. As with all problematic cases, it's not that the method of hypothesis testing has been refuted or even found flawed. It's the opposite. It's doing its work. Because the method's requirements are distorted so that their logic breaks down, it outputs "spurious," just as it should. The types of cases calling for adjustment tend to be cases where there is a related group of hypotheses – such as effects of a drug. The concern is treating the hypotheses disinterred in explorations in just the *same way* as if they were predesignated.

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

Adjustment for Selection Goes against Scientific Norms. Epidemiologist Kenneth Rothman denies an adjustment for selection is appropriate because it depends on the assumption that the variable of interest (e.g., treatment with drug X) is unrelated to all 20 (or however many) of the effect variables searched. He calls it the "universal null," which he views as untenable. Why? "[N]o empiricist could comfortably presume that randomness underlies the variability of all observations. Scientists presume instead that the universe is governed by natural laws"(1990, p. 45).

This is interesting; let's examine it. First, the universal null is an i-assumption (argumentative) only. Second, the universal null does not say that observed outcomes are not governed by laws. Suppose the one nominally significant factor is improved mortality (while 19 others are non-significant). Each death has a reason or cause. It's because death can come about for so many different reasons that there's variability, and we try to root out those due to (or at least systematically associated with) the treatment variable. What's alleged to be due to chance is that the group assigned drug X happens to do nominally better on one factor. Reflect on some of the cases we've visited in our two first excursions: the Texas Sharpshooter, the Pickrite Stock method, Dr. Playfair, Lady Tasting Tea, and the angst of Diederik Stapel's review committee. They're all a bit different but share canonical features.

The Texas Sharpshooter vividly shows how selection effects can change the process responsible for your observations. A silly version described by Goldacre is this: "Imagine I am standing near a large wooden barn with an enormous machine gun. I place a blindfold over my eyes and laughing maniacally I fire off many thousands and thousands of bullets into the side of the barn." Circling a cluster of closely placed bullet holes, he declares it evidence of his marksmanship (2008, p. 258). The skill that he's allegedly testing and making inferences about is his ability to shoot when the target is given and fixed, but that's not the skill actually responsible for *the resulting high score*. That would be so even without the blindfold. It's the high score that's due to chance. Analogously, in searching the data, if you draw a line around those treated who happen to show the beneficial effect, you're influencing what's producing your overall score. Given the variability of the outcomes, it's fairly probable that at least one of 20 factors shows a nominally significant association – say decreased mortality – even if drug X does not systematically improve *any* of the 20 outcomes searched in the populations of interest. You may question the plausibility of this universal null, but that doesn't block its argumentative role. It serves as a canonical representation of deception, or a blatant error scenario for which statistical models are so apt. For a severe tester, the mere fact that your method doesn't distinguish a case where the high score is due to the efficacy of the drug and one where it's due to post hoc hunting, suffices to discredit the resulting inference. You haven't sincerely tried to avoid deception. Other data might later turn up to warrant the claimed benefit; this does not stop us from needing to say something about this one data analysis: It's BENT.

False Discovery Rates. The Bonferroni, or a number of related adjustments, is relevant when the nominally significant difference is the basis for a specific inference. What if you're just trying to get some factors for subsequent severe scrutiny, and the concern is with overlooking genuine effects (Type II errors)? For example, the analysis of microarray data involves analyzing thousands of genes to see which ones are "on" or "expressed" in healthy versus diseased tissue. These are called genome-wide association studies (GWAS). A GWAS might test tens of thousands of null hypotheses that disease status and a given genomic expression are statistically independent. The required P-value using the conservative Bonferroni adjustment would be so tiny that you'd miss out on genes worth following up. Background knowledge might suggest a small proportion of the null hypotheses are false, and a culling procedure is needed. A novel approach by Yosef Benjamini and Yoav Hochberg (1995) uses what's called the *false-discovery rate* (FDR): the expected proportion of the N hypotheses tested that are falsely rejected.[4]

Let R be the number of rejected null hypotheses and V the number of erroneously rejected hypotheses. Then the proportion of falsely rejected null hypotheses is V/R. R is observable, V is not, but we can compute the expected value of V/R:

> FDR: the expected proportion of the hypotheses tested that are falsely rejected, $E(V/R)$.

Benjamini and Hochberg cite a study from the literature where a new treatment is compared to an existing one in a randomized trial of patients with heart disease. (I'm omitting details.) Although 15 different tests are run, the study does not take account of multiple testing. To apply the FDR method, you rank the 15 P-values from lowest to highest (p. 295):

> 0.0001, 0.0004, 0.0019, 0.0095, 0.0201, 0.0278, 0.0298, 0.0344, 0.0459, 0.3240, 0.4262, 0.5719, 0.6528, 0.7590, 1.000.

[4] This should not be confused with use of this term as a posterior probability, e.g., David Colquhoun (2014). There's no prior used in Benjamini and Hochberg.

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

Each gets an index number *i* from 1 to 15: p_1= 0.0001, p_2 = 0.0004, . . ., p_{15} = 1.000, etc. The Bonferroni correction is 0.05/15 = 0.0033. So it only allows rejecting the hypotheses corresponding to those with the three lowest *P*-values. These happen to be reduced allergic reaction and two aspects of bleeding. The fourth corresponds to mortality, so one can't infer a reduction in mortality at the Bonferroni adjusted level.

Let's compare the Bonferroni to applying the FDR method. Choose *Q*, the desired FDR – before collecting the data – and compute for each p_i its Benjamini and Hochberg critical value: $Q_i/15$. They illustrate with the FDR value of 0.05. Starting with p_{15}, look for the first P-value to satisfy $p_i < 0.05i/15$. In this case it's p_4:

$$p_4 = 0.0095 \le 4(0.05)/15 = 0.013.$$

Thus, the null hypothesis corresponding to p_4 – no improved mortality – may be rejected using the FDR. Generally, *Q* is chosen to be much higher than 0.05, say 0.25. In that case you'd look for the first *P*-value, starting at p_{15}, to satisfy $p_i < 0.25i/15$. Here it's p_9: $p_9 = 0.0459 \le 9(0.25)/15 = 0.15$. So all null hypotheses corresponding to ranks 1–9 would be rejected using the FDR with $Q = 0.25$.

In screening genes, the immediate task is largely performance: controlling the noise in the network, balancing false positives with low power to detect genes worth following up, rather than inferring, for any specific gene, that there is evidence of its association to a disease.[5] Here the FDR isn't serving as an inferential measure. Another context Benjamini and Hochberg give, where the FDR seems more apt than the FWER, is where an overall decision to recommend a new treatment is correct, so long as it gives improvement on any one of *N* features:

> We wish therefore to make as many discoveries as possible (which will enhance a decision in favour of the new treatment), subject to control of the FDR. . . . a small proportion of errors will not change the overall validity of the conclusion. (ibid., p. 292)

Here the "mistake" would be failing to recommend the new treatment even though it's better on any of these *N* factors (and presumably no worse). Here the FDR gives the relevant assessment for severity, given what counts as a mistaken inference.

In general, it makes most sense to report the nominal *P*-value associated with each hypothesis and then indicate if it's rejectable according to the chosen

[5] The FDR reflects "the concern to assure reproducible results in the face of selection effects", in contrast to making personal decisions as to what leads to follow (Benjamini 2008, p. 23).

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

adjustment method and level. Despite the importance I place on avoiding or correcting for selection effects, I do not think much weight should be attached to any particular selected *P*-value adjustment. It may suffice to register the spuriousness of a nominal *P*-value. Even searching can yield enough significant results so that the error probability is low. Go back to Selvin who is alluding to FWER (under the assumption of all true nulls):

> We have seen that the probability of *at least one* difference 'significant' at the 5 percent level is 0.64. By similar calculation it can be shown that the probability of at least *two* 'significant' differences is 0.26 . . . at least three is 0.07, . . . at least four is 0.01. (Selvin 1970, pp. 104–5)

The set of differences combined could be seen to be significant at the 1 percent level. Whether this translates into the relevant assessment for severity depends on the context. One may be able to argue, in considering a group of related hypotheses, that if enough non-nulls are found nominally significant, the overall error probability can be low. There is still the problem of actually computing the overall *P*-values, taking into account correlated biases and dependencies of the factors hunted.[6]

The current state of play in dealing with multiple testing is fascinating and growing daily, but would take us too far afield to discuss. Most embody the performance oriented goal of error control. Westfall and Young (1993), in one of the earliest treatments on *P*-value adjustments, develop resampling methods to adjust for both multiple testing and multiple modeling (e.g., selecting variables to be included in a regression model). Others devise model-building techniques that control error probabilities as they build (David Hendry 2011). In some cases, it's deemed more appropriate to combine tests through meta-analysis, rather than assess individual inferences, especially where the tests can be considered to be testing the same or appropriately related hypotheses.

There is always room for an appeal procedure. Appealing to everything we know (big picture inference), even the hunter for significance may succeed in arguing that the mistaken interpretation has been avoided *in the case at hand.* Just because the formal statistics has run its course, there are other informal arguments we can turn to. The severe tester builds formal and informal repertoires to classify how the capacities of tests are affected by one or another selection effect. Leading the way is what counts as erroneously solving the

[6] Efron (2013, p. 142) describes a study on 102 men, 52 with prostate cancer and 50 controls, in which over 6000 genes are studied. This is akin to 6000 significance tests. His alternative formulation is empirical Bayesian, where the relative frequency of genes of a certain type is known.

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

problem at hand. We then consider if the capacity of the test to avoid such mistakes is compromised. The fact that our analysis depends on context should not make us feel that we do not know what we are talking about. It's the onus of the researcher to demonstrate the validity of her inference. Inability to compute error probabilities, even approximately, entails low severity in our system.

Exhibit (ix): Infant Training. In the late 1940s, William Sewell sought to investigate a variety of infant training experiences regarding nursing, weaning, and toilet training that, according to widely accepted Freudian psychological theories of the day, were thought advantageous for a child's personality adjustment. Leslie Kish's (1959) discussion is another classic we find in Morrison and Henkel (1970, pp. 127–41).

The researchers in the infant training study conducted 460 statistical significance tests! Out of these 18 were found statistically significant at the 0.05 level or beyond, 11 in the direction expected by the popular psychological account. Sewell denies that we should be just as impressed with the 11 statistically significant results as we would be if they were the only 11 hypotheses to be tested. As Kish points out: "By chance alone one would expect 23 'significant differences' at the 5 percent level. A 'hunter' would report either the 11 or the 18 and not the hundreds of 'misses'" (ibid., p. 138, note 12). A hunter is one who denies the need to take account of the searching.

What's intriguing about this study is that the hunting expedition led to negative results, adjusted to take account of the searching. Here's Sewell (1952, p. 158): "On the basis of the results of this study, the general null hypothesis that the personality adjustments and traits of children who have undergone varying training experiences do not differ significantly cannot be rejected." There were six main hypotheses. For example:

- "*The personality adjustments of the children who were breast fed do not differ* [statistically] *significantly from those of the children who were bottle fed* cannot be rejected." None of the 46 tests were significant (ibid., p. 156).
- *The personality adjustments and traits of the children who were weaned gradually do not differ* [statistically] *significantly from those of the children weaned abruptly* cannot be rejected on the basis of the statistical evidence." Only two of 46 tests were significant (ibid.).

And so it went for children with late versus early induction to bowel training, for those not punished versus those punished for toilet training accidents and others besides (ibid.).

While recognizing the need for a more controlled study in order to falsify the claims, Sewell concludes:

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

> [Psychologists and counselors] have strongly advocated systems of infant care which they believe follow logically from the Freudian position. . . . breast feeding, a prolonged period of nursing, gradual weaning, a self-demand schedule, easy and late bowel and bladder training, . . . freedom from punishment . . . They have assumed that these practices will promote the growth of secure and unneurotic personalities. (ibid., p. 151)

> Certainly, the results of this study cast serious doubts on the validity of the psychoanalytic claims regarding the importance of the infant [training] . . . (ibid., pp. 158–9).

In addition to their astuteness in taking account of searching, notice they're not reticent in reporting negative results. Since few if any were statistically significant, Fisher's requirement for demonstrating an effect fails. This casts serious doubt on what was a widely accepted basis for recommending Freudian-type infant training.[7] The presumption of a genuine effect is statistically falsified, or very close to it.

When Searching Doesn't Damage Severity: Explaining a Known Effect

What is sometimes overlooked is that the problem is not that a method is guaranteed to output some claim or other that fit the data, the problem is doing so unreliably. Some criticisms of adjusting for selection are based on examples where severity is *improved* by searching. We should not be tempted to run together examples of pejorative data-dependent hunting with cases that are superficially similar, but unproblematic. For example, searching for a DNA match with a criminal's DNA is somewhat akin to finding a statistically significant departure from a null hypothesis: "one searches through data and concentrates on the one case where a 'match' with the criminal's DNA is found, ignoring the non-matches." (Mayo and Cox 2006; p. 94) Isn't an error statistician forced to adjust for hunting here as well? No.

In illicit hunting and cherry picking, the concern is that of inferring a genuine effect, when none exists; whereas "here there is a known effect or specific event, the criminal's DNA, and reliable procedures are used to track down the specific cause or source" – or so we assume with background knowledge of a low "erroneous match" rate. "The probability is high that we would not obtain a match with person *i*, if *i* were not the criminal"; so, by the severity criterion (or FEV), finding the match is good evidence that *i* is the

[7] Some recommend that researchers simply state "with maximum clarity and precision, which hypotheses were developed entirely independently of the data and those which were not, so readers will know how to interpret the results" (Bailar 1991). I concur with Westfall and Young (1993, p. 20) that it is doubtful that a reader will know how to interpret a report: the 11 favorable results have been selected from the 200 tests. Isn't that the purpose of a statistical analysis?

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

criminal. "Moreover, each non-match found, by the stipulations of the example, virtually excludes that person." Thus, the more such negative results, the stronger is the evidence against *i* when a match is finally found. Negative results fortify the inferred match. Since "at most one null hypothesis of innocence is false, evidence of innocence on one individual increases, even if only slightly, the chance of guilt of another" (ibid).

Philip Dawid (2000, p. 325) invites his readers to assess whether they are "intuitive Bayesians or intuitive frequentists" by "the extreme case that the data base contains records on everyone in the population." One could imagine this put in terms of 'did you hear about the frequentist who thought finding a non-match with everyone but Sam is poor evidence that Sam is guilty?' The criticism is a consequence of blurring pejorative and non-pejorative cases. It would be absurd to consider the stringency of the probe as diminishing as more non-matches are found. Thus we can remain intuitive frequentist testers. A cartoon shows a man finding his key after searching with the caption "always the last place you look." Searching for your lost key is like the DNA search for a criminal. (Note the echoes with the philosophical dispute about the relevance of novel predictions; Section 2.4.)

If an effect is known to be genuine, then a sought-for and found explanation needn't receive a low severity. Don't misunderstand: not just any explanation of a known effect passes severely, but one mistake – spurious effect – is already taken care of. Once the deflection effect was found to be genuine, it had to be a constraint on theorizing about its cause. In other cases, the trick is to hunt for a way to make the effect manifest in an experiment. When teratogenicity was found in babies whose mothers had been given thalidomide, it took them quite some time to find an animal in which the effect showed up: finally it was replicated in New Zealand rabbits! It is one thing if you are really going where the data *take you*, as opposed to subliminally taking the data where you want them to go. Severity makes the needed distinctions.

Renouncing Error Probabilities Leaves Trenchant Criticisms on the Table

Some of the harshest criticisms of frequentist error-statistical methods these days rest on principles that the critics themselves purport to reject. An example is for a Bayesian to criticize a reported *P*-value on the grounds that it failed to adjust for searching, while denying searching matters to evidence. If it is what I call a "for thee and not for me" argument, the critic is not being inconsistent. She accepts the "I don't care, but you do" horn. When a Bayesian, like I. J. Good, says a Fisherian but not a Bayesian can cheat with optional stopping,

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

he means that error probabilities aren't the Bayesian's concern (Section 1.5). (Error probabilities without a subscript always refer to error probability$_1$ (Section 3.6) from frequentist methods.) It's perfectly fair to take the "we don't care" horn of my dilemma, and that's a great help in getting beyond the statistics wars. What's unfair is dismissing those who care as fetishizing imaginary data. Critics of error statistics should admit to consequences sufficiently concerning to be ensconced in statutes of best practices. At least if the statistics wars are to become less shrill and more honest.

What should we say about a distinct standpoint you will come across? First a critic berates a researcher for reporting an unadjusted *P*-value despite hunting and multiple testing. Next, because the critic's own methodology eschews the error statistical rationale on which those concerns rest, she is forced to switch to different grounds for complaining – generally by reporting disbelief in the effect that was hunted. Recall Royall blocking "the deck is made up of aces of diamonds," despite its being most likely (given the one ace of diamonds drawn), by switching to the belief category (Section 1.4). That might work in such trivial cases. In others, it weakens the intended criticism to the point of having it obliterated by those who deserve to be charged with severe testing crimes.

Exhibit (x): Relinquishing Their Strongest Criticism: Bem. There was an ESP study that got attention a few years back (Bem 2011). Anyone choosing to take up an effect that has been demonstrated only with questionable research practices or has been falsified must show they have avoided the well-known tricks. But Bem openly admits he went on a fishing expedition to find results that appear to show an impressive non-chance effect, which he credits to ESP (subjects did better than chance at predicting which erotic picture they'll be shown in the future). The great irony is that Wagenmakers et al. (2011), keen as they are to show "Psychologists Must Change the Way They Analyze Their Data" and trade significance tests for Bayes factors, relinquish their strongest grounds for criticism. While they mention Bem's *P*-hacking (fishing for a type of picture subjects get right most often), this isn't their basis for discrediting Bem's results. After all, Wagenmakers looks askance at adjusting for selection effects:

> *P* values can only be computed once the sampling plan is fully known and specified in advance. In scientific practice, few people are keenly aware of their intentions, particularly with respect to what to do when the data turn out not to be significant after the first inspection. Still fewer people would adjust their *p* values on the basis of their intended sampling plan. (Wagenmakers 2007, p. 784)

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

Rather than insist they ought to adjust, Wagenmakers dismisses a concern with "hypothetical actions for imaginary data" (ibid.). To criticize Bem, Wagenmakers et al. (2011) resort to a default Bayesian prior that makes the null hypothesis comparatively more probable than a chosen alternative (along the lines of Excursion 4, Tour II). Not only does this forfeit their strongest criticism, they give Bem et al. (2011) a cudgel to thwack back at them:

> Whenever the null hypothesis is sharply defined but the prior distribution on the alternative hypothesis is diffused over a wide range of values, as it is in … Wagenmakers et al. (2011), it boosts the probability that *any* observed data will be higher under the null hypothesis than under the alternative. This is known as the Lindley-Jeffreys paradox: A frequentist analysis that yields strong evidence in support of the experimental hypothesis can be contradicted by a misguided Bayesian analysis that concludes that the same data are more likely under the null. (p. 717)

Instead of getting flogged, Bem is positioned to point to the flexibility of getting a Bayes factor in favor of the null hypothesis. Rather than showing psychologists should switch, the exchange is a strong argument for why they should stick to error statistical requirements.

P-Values Can't Be Trusted Except When Used to Argue That *P*-values Can't Be Trusted!

There is more than a whiff of inconsistency in proclaiming *P*-values cannot be trusted while in the same breath extolling the uses of statistical significance tests and *P*-values in mounting criticisms of significance tests and P-values. Isn't a critic who denies the entire error statistical methodology, significance test, N-P tests, and confidence intervals, also required to forfeit the results those methods give when they just happen to criticize a given of the tests? How much more so when those criticisms are the basis for charging someone with fraud. Yet that is not what we see.

Uri Simonsohn became a renowned fraud-buster by inventing statistical tests to rigorously make out his suspicions of the work of social psychologists Dirk Smeesters, Lawrence Sanna, and others – "based on statistics alone"– as one of his titles reads. He shows the researcher couldn't have gotten so little variability, or the results are too good to be true – along with a fastidious analysis of numerous papers to rule out, statistically, any benign explanations (Simonsohn 2013). Statistician Richard Gill, often asked for advice on such cases, notes: "The methodology here is not new. It goes back to Fisher (founder of modern statistics) in the 30's. . . The tests of goodness of fit were, again and again, too good" (2014). I expected that tribes who deny the evidential weight

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

of significance tests would come to the defense of the accused, but (to my knowledge) none did.

Note, too, that the argument of the fraud-busters underscores the severity rationale for the case at hand. Critics called in to adjudicate high-profile cases of suspected fraud are not merely trying to ensure they will rarely erroneously pinpoint frauds in the long run. They are making proclamations on the specific case at hand – and in some cases, a person's job depends on it. They will use a cluster of examples to mount a strong argument from coincidence that the data in front of us could not have occurred without finagling. Other tools are used to survey a group of significance tests in a whole field, or by a given researcher. For instance, statistical properties of *P*-values are employed to ascertain if too many *P*-values at a given level are attained. These are called *P*-curves (Simonsohn et al. 2014). Such fraud detection machines at most give an indication about a field or group of studies. Of course, once known, they might themselves be gamed. But it's an intriguing new research field; and it is an interesting fact that when scientists need to warrant serious accusations of bad statistics, if not fraud, they turn to the error statistical reasoning and to statistical tests. If you got rid of them, they'd only have to be reinvented by those who insist on holding others accountable for their statistical inferences.

Exhibit (xi): How Data-dependent Selections Invalidate Error Probability Guarantees. It can be shown that a statistical method directly protects against data-dependent selections by demonstrating how they can cause a breakdown in methods. Philosopher of science Ronald Giere considers Neyman–Pearson interval estimation for a Binomial proportion. If assumptions are met, the sample mean will differ by 2 standard deviations from the true value of θ less than 5 percent of the time, approximately. Giere shows how to make the probability of successful estimates not 0.95 but 0! "This will be sufficient to prove the point [the inadmissibility of this method] because Neyman's theory asserts that the average ratio of success is independent of the constitutions of the populations examined" (Giere 1969, p. 375). Take a population of *A*'s and to each set of *n* members assign a shared property. The full population has *U* members where $U > 2n$. Then arbitrarily assign this same property to $U/2 - n$ additional members.

> Given a sufficient store of logically independent properties, this can be done for all possible combinations of *n* *A*'s. The result is a population so constructed that while every possible *n*-membered sample contains at least one apparent regularity [a property

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

> shared by all members of the sample] every independent property has an actual ratio of exactly one-half in the total population. (ibid., p. 376)[8]

The bottom line is, showing how you can distort error probabilities through the efforts of finagling shows the *value* of these methods. It's hard to see how accounts that claim error probabilities are irrelevant can supply such direct protection, although they may *indirectly* block the same fallacies. This remains to be shown.

Souvenir S: Preregistration and Error Probabilities

"One of the best-publicized approaches to boosting reproducibility is preregistration ... to prevent cherry picking statistically significant results" (Baker 2016, p. 454). It shouldn't be described as too onerous to carry out. Selection effects alter the outcomes in the sample space, showing up in altered error probabilities. If the sample space (and so error probabilities) is deemed irrelevant post-data, the direct rationale for preregistration goes missing. Worse, in the interest of promoting a methodology that downplays error probabilities, researchers who most deserve lambasting are thrown a handy line of defense. Granted it is often presupposed that error probabilities are relevant only for long-run performance goals. I've been disabusing you of that notion. Perhaps some of the "never error probabilities" tribe will shift their stance now: 'But Mayo, using error probabilities for severity, differs from the official line, which is all about performance.' One didn't feel too guilty denying a concern with error probabilities before. If viewing statistical inference as severe tests yields such a concession, I will consider my project a success. Actually, my immediate goal is less ambitious: to show that looking through the severity tunnel lets you unearth the crux of major statistical battles. In the meantime, no fair critic of error statistics should proclaim error control is all about hidden intentions that a researcher can't be held responsible for. They should be.

4.7 Randomization

> The purpose of randomisation ... is to guarantee the validity of the test of significance, this test being based on an estimate of error made possible by replication. (Fisher [1935b]1951, p. 26)

> The problem of analysing the idea of randomization is more acute, and at present more baffling, for subjectivists than for objectivists, more baffling because an ideal subjectivist would not need randomization at all. He would

[8] For a miniature example, if $U = 6$ (there are 6 A's in the population) and $n = 2$, there are 15 possible pairs. Each pair is given a property and so is one additional member.

> simply choose the specific layout that promised to tell him the most. (Savage 1962, p. 34)

> Randomization is a puzzle for Bayesians. The intuitive need for randomization is clear, but there is a standard result that Bayesians need not randomize. (Berry and Kadane 1997, p. 813)

> Many Bayesians (though there are some very prominent exceptions) regard it as irrelevant and most frequentists (again there are some exceptions) consider it important. (Senn 2007, p. 34)

There's a nagging voice rarely heard from in today's statistical debates: if an account has no niche for error statistical reasoning, what happens to design principles whose primary purpose is to afford it? Randomization clearly exploits counterfactual considerations of outcomes that could have occurred, so dear to the hearts of error statisticians.

> Some of the greatest contributions of statistics to science involve adding additional randomness and leveraging that randomness. Examples are randomized experiments, permutation tests, cross-validation and data-splitting. These are unabashedly frequentist ideas and, while one can strain to fit them into a Bayesian framework, they don't really have a place in Bayesian inference. (Wasserman 2008, p. 465)

One answer is to recognize that, apart from underwriting significance tests and the estimation of the standard error, randomization also has a role in preventing types of biased selections, especially where the context requires convincing others. Although these and other justifications are interesting and important, their defenders tend to regard them as subsidiary and largely optional uses for randomization.

Randomization country is huge; one scarcely does it justice in a single afternoon's tour. Moreover, a deeper look would require I call in a more expert field guide. A glimpse will shed light on core differences that interest us. Let's focus on the random allocation of a treatment or intervention, in particular in comparative treatment-control studies.

The problem with attributing Mary's lack of dementia (by age 80) to her having been taking HRT since menopause is that we don't know what her condition would have been like if she had not been so treated. Moreover, she's just one case and we're interested in treatment effects that are statistical. A factor is sometimes said to statistically contribute to some response where the response on average in the experimental population of treateds would be higher (or lower) than it would have been had they not been treated – in effect comparing two counterfactual populations. Randomized control experiments let us peer into these counterfactual populations by finding out about the difference between the average response in the treated group μ_T and the

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

average response among a control group μ_C. With randomized control trials (RCTs), there is a deliberate introduction of a probabilistic assignment of a treatment of interest, using known chance mechanisms, such as a random number generator. Letting $\Delta = \mu_T - \mu_C$, one may consider what Cox calls a strong ("no effect") null: that the average response is no different or no greater among the treated than among the control group H_0: $\Delta = 0$ (vs. H_1: $\Delta \neq 0$), or a one-sided null: H_0: $\Delta \leq 0$ vs. H_1: $\Delta > 0$. We observe $\overline{x}_T - \overline{x}_C$, where $\overline{x}_T$ and $\overline{x}_C$ are the observed sample means in the treated and control groups, forming the standard test statistic $d^* = (d - \Delta)/\text{SE}$, where SE is the standard error ($d = \overline{x}_T - \overline{x}_C$). Thanks to randomized assignment, we can estimate the standard error and the sampling distribution of d^*.

Under the (strong) null hypothesis, the two groups, treated and control, may be regarded as coming from the same population with respect to mean effect, such as age-related dementia. Think about the RCT reasoning this way: if the HRT treatment makes no difference, people in the treated group would have had (or not had) dementia even if they'd been assigned to the control group. Some will get it, others won't (of course we can also consider degrees). Under the null hypothesis, any observed difference would be due to the accidental assignment to group T or C. So, if $\overline{x}_T - \overline{x}_C$ exceeds 0, it's just because more of the people who would have gotten dementia anyway happen to end up being assigned to the treated rather than the control group. Thanks to the random assignment, we can determine the probability of this occurring (under H_0). This is a particularly vivid illustration of a difference "due to chance" – where the chance is the way subjects were assigned to treatment. A statistical connection between the theoretical parameter Δ is created by dint of the design and execution of the experiment.[9]

Bayesians may find a home for randomized assignment (and possibly, double blindness) in the course of demonstrating "utmost good faith" (Senn 2007, p. 35). In the face of suspicious second parties: "Simple randomization is a method which by virtue of its very unpredictability affords the greatest degree of blinding. Some form of randomization is indispensable for any trial in which the issue of blinding is taken seriously..." (ibid., p. 70). Nevertheless, subjective Bayesians have generally concurred with Lindley and Novick that

[9] Stephen Stigler, in his clever *The Seven Pillars of Statistical Wisdom*, discusses some of the experiments performed by Peirce, who first defined randomization. In one, the goal was to test whether there's a threshold below which you can't discern the difference in weights between two objects. Psychologists had hypothesized that there was a minimal threshold "such that if the difference was below the threshold, termed the *just noticeable difference* (jnd), the two stimuli were indistinguishable . . . [Peirce and Jastrow] showed this speculation was false" (Stigler 2016, p. 160). No matter how close in weight the objects were, the probability of a correct discernment of difference differed from ½. Another example of evidence for a "no-effect" null.

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

"[O]ne can do no better than . . . use an allocation which You think is unlikely to have important confounding effects" (Lindley and Novick 1981, p. 52).[10] Still, Berry and Kadane (1997) maintain that despite the "standard result that Bayesians need not randomize" (p. 813) there are scenarios where, because different actors have different subjective goals and beliefs, randomization is the optimal allocation. Say there's two treatments, 1 and 2, where each either shows the response of interest or does not. Dan, who will decide on whether the allocation should be randomized or not, is keen for the result to give a good estimate of the response rates over the whole population, whereas Phyllis, a doctor, believes one type of patient, say healthy ones, does better with treatment 1 than 2, and her goal is giving patients the best treatment (ibid., p. 818). "[I]f Dan has a positive probability that Phyllis, or whoever is allocating, is placing the patients on the two treatments unequally, then randomization is the preferred allocation scheme (optimal)" (ibid.). Presumably, the agent doesn't worry that he unconsciously biases his own study aimed at learning the success rate in the population. For non-subjective Bayesians, there may be an appeal to the fact that "with randomization, the posterior is much less sensitive to the prior. And I think most practical Bayesians would consider it valuable to increase the robustness of the posterior" (Wasserman 2013). An extensive discussion may be found in Gelman et al. (2004). Still, as I understand it, it's not the *direct* protection of error probabilities that drives the concern.

Randomization and the Philosophers

> It seems surprising that the value of randomisation should still be disputed at this stage, and of course it is not disputed by anybody in the business. There is, though, a body of philosophers who do dispute it. (Colquhoun 2011, p. 333)

It's a bit mortifying to hear Colquhoun allude to "subversion by philosophers of science" (p. 321). Philosophical arguments against randomization stem largely from influential Bayesian texts (e.g., Howson and Urbach 1993), "On the Bayesian view, randomization is optional, and the essential condition is for the comparison groups in a clinical trial to be adequately matched on factors believed to have prognostic significance" (Howson and Urbach 1993, p. 378). A criticism philosophers often raise is due to the possibility of unknown confounding factors that differentiate the treated from the control

[10] Lindley (1982, p. 439) argues that if practitioners would dismiss an allocation that appeared unsatisfactory "one might ask why randomize in the first place?" Just because we can fail to satisfy a statistical assumption, does not imply we shouldn't try to succeed, test if we've failed, and fix problems found.

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

groups (e.g., Worrall 2002, p. S324). As Stephen Senn explains (2013b, p. 1447), such "imbalance," which is expected, will not impugn the statistical significance computations under randomization. The analysis is of a ratio of the between group variability and the within group variability. The said imbalance between groups (the numerator) would also impinge on the within group variability (the denominator). Larger variability will at worst result in larger standard errors, wider confidence intervals, resulting in a non-statistically significant result. The relevance of the unknown factors "is bounded by outcome and if we have randomised, the variation within groups is related to the variation between in a way that can be described probabilistically by the Fisherian machinery" (ibid.). There's an observed difference between groups, and our question is how readily could the observed difference in outcome be generated by chance alone? Senn goes further.

> It is not necessary for the groups to be balanced. In fact, the probability calculation applied to a clinical trial automatically *makes an allowance for the fact that groups will almost certainly be unbalanced*, and if one knew that they were balanced, then the calculation that is usually performed would not be correct. Every statistician knows that you should not analyse a matched pair's design as if it were a completely randomised design. (ibid., p. 1442)

The former randomly assigns a treatment to a pair that have been deliberately matched.

In clinical trials where subjects are assigned as they present themselves, you can't look over the groups for possibly relevant factors, but you can include them in your model. Suppose the sex of the patient is deemed relevant. According to Senn:

(1) If you have sex in the model the treatment estimate is corrected for sex whether or not the design is balanced; balancing makes it more efficient.
(2) Balancing for sex but not having it in the model does not give a valid inference.[11]

RCT4D

A different type of debate is cropping up in fields that are increasingly dabbling with using randomized controlled trials (RCTs) rather than a typical reliance on observational data and statistical modeling. The Poverty Action Lab at MIT led by Abhijit Banerjee and Esther Duflo (2011) is spearheading a major movement in development economics to employ RCTs to test the benefits of

[11] "If you refuse to choose at random between all possible designs in which sex is balanced I will cry fraud." (Senn, private communication; see also Senn 1994, p. 1721)

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

various aid programs for spurring economic growth and decreasing poverty, from bed nets and school uniforms, to micro-loans. For those advocating RCTs in development economics (abbreviated RCT4D), if you want to discover if school uniforms decrease teen pregnancy in Mumbai, you take *k* comparable schools and randomly assign uniforms to some and not to others; at the end of the study, differences in average results are observed. It is hoped thereby to repel criticisms from those who question if there are scientific foundations guiding aid-driven development.

Philosopher of science Nancy Cartwright allows that RCTs, if done very well, can show a program worked in a studied situation, but that's "a long way from establishing that it *will work* in a particular target" (2012, p. 299). A major concern is that what works in Kenya needn't work in Mumbai. In general, the results of RCTs apply to the experimental population and, unless that's a random sample of a given target population, the issue of extrapolating (or external validity) arises. That is true. Merely volunteering to be part of a trial may well be what distinguishes subjects from others.

The conflicting sides here are largely between those who advocate experimental testing of policies and those who think we can do better with econometric modeling coupled with theory. Opponents, such as Angus Deaton, think the attention should be "refocused toward the investigation of potentially generalizable mechanisms that explain why and in what contexts projects can be expected to work" (Deaton 2010, p. 426) via modeling and trial and error. But why not all of the above? Clearly RCTs limit what can be studied, so they can't be the only method. The "hierarchies of evidence" we agree should include explicit recognition of flaws and fallacies of extrapolation.

Giving the mother nutritional information improves child nourishment in city X, but not in city Y where the father does the shopping and the mother-in law decides who eats what. Small classrooms improve learning in one case, but not if applying it means shutting down spaces for libraries and study facilities. I don't see how either the modelers or the randomistas (as they are sometimes called) can avoid needing to be attuned to such foibles. Shouldn't the kind of field trials described in Banerjee and Duflo (2011) reveal clues as to *why* what works in X won't work in Y? Perhaps one of the most valuable pieces of information emerges from talking with and interacting amongst the people involved.

Cartwright and Hardie worry that RCTs, being rule oriented, reduce or eliminate the use of necessary discretion and judgment:

> If a rule such as 'follow the RCTs, and do so faithfully' were a good way of deciding about effectiveness, then certainly deliberation is second best (or worse) . . . the

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

> orthodoxy, which is a rules system, discourages decision makers from thinking about their problems, because the aim of rules is to reduce or eliminate the use of discretion and judgment, . . . The aim of reducing discretion comes from a lack of trust in the ability of operatives to exercise discretion well . . . Thus, the orthodoxy not only discourage deliberation, as unnecessary since the rules are superior, but selects in favor of operatives who cannot deliberate. (Cartwright and Hardie 2012, pp. 158–9)

Do rules to prevent QRPs, conscious and unconscious, reflect a lack of trust? Absolutely, even those trying hard to get it right aren't immune to the tunnel vision of their tribe.

The truth is, performing and interpreting RCTs involve enormous amounts of discretion. One of the big problems with experiments in many fields is the way statistical-scientific gaps are filled in interpreting results. In development economics, negative results are hard to hide, but there's still plenty of latitude for post-data explanations. RCTs don't protect you from post-data hunting and snooping. One RCT gave evidence that free school uniforms decreased teenage pregnancy, by encouraging students to remain in school. Here, teen pregnancies serve as a proxy for contracting HIV/AIDS. However, combining uniforms with a curriculum on HIV/AIDS gave negative results.

> In schools that had both the HIV/AIDS and the uniforms programs, girls were no less likely to become pregnant than those in the schools that had nothing. The HIV/AIDS education curriculum, instead of reducing sexual activity . . ., actually *undid* the positive effect of the [free uniforms]. (Banerjee and Duflo 2011, p. 115)

Several different theories are offered. Perhaps the AIDS curriculum, which stresses abstinence before marriage, encouraged marriages and thus pregnancies. Post-data explanations for insignificant results are as seductive here as elsewhere and aren't protected by an RCT without prespecified outcomes. If we are to evaluate aid programs, an accumulation of data on bungled implementation might be the best way to assess what works and why. Rather than scrap the trials, a call for explicit attention to how a program could fail in the new environment is needed. Researchers should also be open to finding that none of the existing models captures people's motivations. Sometimes those receiving aid might just resist being "helped," or being nudged to do what's "rational".

Randomization is no longer at the top of the evidence hierarchy. It' s been supplanted by systematic reviews or meta-analysis of all relevant RCTs. Here too, however, there are problems of selecting which studies to include and their differing quality. Still the need for meta-analysis has promoted "all trials," rather than hiding any in file-drawers, and with the emphasis by the Cochrane collaboration, are clearly not going away. Meta-analytic reviews have received

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

some black eyes, but it's one of the central ways of combining results in frequentist statistics.[12] Our itinerary got too full to visit this important topic; you may catch it on a return tour.

Batch-Effect Driven Spurious Associations

> There is a relatively unknown problem with microarray experiments, in addition to the multiple testing problems [microarray] samples should be randomized over important sources of variation; otherwise p-values may be flawed. Until relatively recently, the microarray samples were not sent through assay equipment in random order. . . . Essentially all the microarray data pre-2010 is unreliable. (Young 2013)

The latest Big Data technologies are not immune from basic experimental design principles. We hear that a decade or more has been lost by failing to randomize microarrays. "Stop Ignoring Experimental Design (or my head will explode)" declares genomics researcher Christophe Lambert (2010). The result is spurious associations due to confounding "to the point, in fact, where real associations cannot be distinguished from experimental artifacts" (ibid., p. 1). Microarray analysis involves a great many steps, plating and processing, and washing and more; minute differences in entirely non-biological variables can easily swamp the difference of interest. Suppose a microarray, looking for genes differentially expressed between diseased and healthy tissue (cases and controls), processes the cases in one batch, say at a given lab on Monday, and the controls on Tuesday. The reported statistically significant differences may be swamped by artifacts – the tiny differences due to different technicians, different reagents, even ozone levels. A "batch" would be a set of microarrays processed in a relatively homogeneous way, say at a single lab on Monday. Batch effects are defined to be systematic non-biological variations between groups of samples (or batches) due to such experimental artifacts. A paper on genetic associations and longevity (Sebastiani et al. 2010) didn't live very long because it turned out the samples from centenarians were differently collected and processed than the control group of average people. The statistically significant difference disappeared when the samples were run on the same batch, showing the observed association to be an experimental artifact.

By randomly assigning the order of cases and controls, the spurious associations vanish! Ideally they also randomize over data collection techniques and balance experimental units to different batches, but "the case/control status is the most important variable to randomize" (Lambert 2010, p. 4). Then,

[12] See for example Ioannidis (2016). For applications, see Cumming (2012) and Senn (2007).

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

corrections due to site and data collection can be made later. But the reverse isn't true. As Fisher said, "To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination:. . . to say what the experiment died of" (Fisher 1938, p. 17). Nevertheless, in an attempt to fix the batch effect driven spurious associations,

> [A] whole class of post-experiment statistical methods has emerged . . . These methods . . . represent a palliative, not a cure. . . . the GWAS research community has too often accommodated bad experimental design with automated post-experiment cleanup. . . . experimental designs for large-scale hypothesis testing have produced so many outliers that the field has made it standard practice to automate discarding outlying data. (Lambert and Black 2012, pp. 196–7)

By contrast, with proper design they find that post-experiment automatic filters are unneeded. In other words, the introduction of randomized design *frees them to deliberate* over the handful of extreme values to see if they are real or artifacts. (This contrasts with the worry raised by Cartwright and Hardie (2012).)

Souvenir T: Even Big Data Calls for Theory and Falsification

> Historically, epidemiology has focused on minimizing Type II error (missing a relationship in the data), often ignoring multiple testing considerations, while traditional statistical study has focused on minimizing Type I error (incorrectly attributing a relationship in data better explained by random chance). When traditional epidemiology met the field of GWAS, a flurry of papers reported findings which eventually became viewed as nonreplicable. (Lambert and Black 2012, p. 199)

This is from Christophe Lambert and Laura Black's important paper "Learning from our GWAS Mistakes: From Experimental Design to Scientific Method"; it directly connects genome-wide association studies (GWAS) to philosophical themes from Meehl, Popper and falsification. In an attempt to staunch the non-replication, they explain, adjusted genome-wide thresholds of significance were required as well as replication in an independent sample (Section 4.6).

However, the intended goal is often thwarted by how this is carried out. "[R]esearchers commonly take forward, say, 20–40 nominally significant signals" that did not meet the stricter significance levels, "then run association tests for those signals in a second study, concluding that all the signals with a p-value ≤.05 have replicated (no Bonferroni adjustment). Frequently 1 or 2 associations replicate – which is also the number expected by random chance" (ibid.). Next these "replicated" cases are combined with the original data "to compute p-values considered genome-wide significant. This method has been

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press

propagated in publications, leading us to wonder if standard practice could become to publish random signals and tell a plausible biological story about the findings" (ibid.).

Instead of being satisfied with a post-data biological story to explain correlations, "[i]f journals were to insist that association studies also suggest possible experiments that could falsify a putative theory of causation based on association, the quality and durability of association studies could increase" (ibid., p. 201). At the very least, the severe tester argues, we should strive to falsify methods of inquiry and analysis. This might at least scotch the tendency Lambert and Black observe, for others to propagate a flawed methodology once seen in respected journals: "[W]ithout a clear falsifiable stance – one that has implications for the theory – associations do not necessarily contribute deeply to science" (ibid., p. 199).

https://doi.org/10.1017/9781107286184.011 Published online by Cambridge University Press