Statistical Inference as Severe Testing:

How to Get Beyond the Statistics Wars

Deborah G. Mayo

Abstract for Book

By disinterring the underlying statistical philosophies this book sets the stage for understanding and finally getting beyond today's most pressing controversies revolving around statistical methods and irreproducible findings. *Statistical Inference as Severe Testing* takes the reader on a journey that provides a non-technical "how to" guide for zeroing in on the most influential arguments surrounding commonly used–and abused– statistical methods. The book sets sail with a tool for telling what's true about statistical controversies: If little if anything has been done to rule out flaws in taking data as evidence for a claim, then that claim has not passed a stringent or *severe test*. In the severe testing account, probability arises in inference, not to measure degrees of plausibility or belief in hypotheses, but to assess and control how severely tested claims are. Viewing statistical inference as severe testing supplies novel solutions to problems of induction, falsification and demarcating science from pseudoscience, and serves as the linchpin for understanding and getting beyond the statistics wars. The book links philosophical questions about the roles of probability in inference to the concerns of practitioners in psychology, medicine, biology, economics, physics and across the landscape of the natural and social sciences.

Keywords for book:

Severe testing, Bayesian and frequentist debates, Philosophy of statistics, Significance testing controversy, statistics wars, replication crisis, statistical inference, error statistics, Philosophy and history of Neyman, Pearson and Fisherian statistics, Popperian falsification

Excursion 1: How to Tell What's True about Statistical Inference

Tour I: Beyond Probabilism and Performance

(1.1) If we're to get beyond the statistics wars, we need to understand the arguments behind them. Disagreements about the roles of probability in statistical inference–holdovers from long-standing frequentist-Bayesian battles–still simmer below the surface of current debates on scientific integrity, irreproducibility, and questionable research practices. Striving to restore scientific credibility, researchers, professional societies, and journals are getting serious about methodological reforms. Some–disapproving of cherry picking and advancing preregistration–are welcome. Others might create obstacles to the critical standpoint we seek. Without understanding the assumptions behind proposed reforms, their ramifications for statistical practice remain hidden. (1.2) Rival standards reflect a tension between using probability (i) to constrain a method's ability to avoid erroneously interpreting data (*performance*), and (ii) to assign degrees of support, confirmation, or plausibility to hypotheses (*probabilism*). We set sail with a tool for telling what's true about statistical inference: If little has been done to rule out flaws in taking data as evidence for a claim, then that claim has not passed a *severe test*. From this minimal severe-testing requirement, we develop a statistical philosophy that goes beyond probabilism and performance. (1.3) We survey the current state of play in statistical foundations.

Excursion 1 Tour I: Keywords

Error statistics, severity requirement: weak/strong, probabilism, performance, probativism, statistical inference, argument from coincidence, Life-off (vs drag down), sampling distribution, cherry-picking

Excursion 1 Tour II: Error Probing Tools vs. Logics of Evidence

Core battles revolve around the relevance of a method's error probabilities. What's distinctive about the severe testing account is that it uses error probabilities evidentially: to assess how severely a claim has passed a test. Error control is necessary but not sufficient for severity. Logics of induction focus on the relationships between given data and hypotheses–so outcomes other than the one observed drop out. This is captured in the Likelihood Principle (LP). Tour II takes us to the crux of central wars in relation to the Law of Likelihood (LL) and Bayesian probabilism. (1.4) Hypotheses deliberately designed to accord with the data can result in minimal severity. The likelihoodist tries to oust them via degrees of belief captured in prior probabilities. To the severe tester, such gambits directly alter the evidence by leading to inseverity. (1.5) If a tester tries and tries again until significance is reached–optional stopping–significance will be attained erroneously with high probability. According to the LP, the stopping rule doesn't alter evidence. The irrelevance of optional stopping is an asset for holders of the LP, it's the opposite for a severe tester. The warring sides talk past each other.

Excursion 1 Tour II: Keywords

Statistical significance: nominal vs actual, Law of likelihood, Likelihood principle
Inductive inference, Frequentist/Bayesian, confidence concept, Bayes theorem, default/non-subjective Bayesian, stopping rules/optional stopping, argument from intentions

Excursion 2: Taboos of Induction and Falsification

Tour I: Induction and Confirmation

The roots of rival statistical accounts go back to the logical Problem of Induction. (2.1) The logical problem of induction is a matter of finding an argument to justify a type of argument (enumerative induction), so it is important to be clear on arguments, their soundness versus their validity. Given that any attempt to solve the logical problem of induction leads to circularity, philosophers turned instead to building logics that seemed to capture our intuitions about induction, e.g., Carnap's confirmation theory. There's an analogy between contrasting views in philosophy and statistics: Carnapian confirmation is to Bayesian statistics, as Popperian falsification is to frequentist error statistics. Logics of confirmation take the form of probabilisms, either in the form of raising the probability of a hypothesis, or arriving at a posterior probability. (2.2) The contrast between these types of probabilisms, and the problems each is found to have in confirmation theory is directly relevant to the types of probabilisms in statistics. Notably, Harold Jeffreys' non-subjective Bayesianism, and current spin-offs, share features with Carnapian inductive logics. We examine problems of irrelevant conjunctions: if $\boldsymbol{x}$ confirms H, it confirms (H & J) for any J.

Tour I: keywords

asymmetry of induction and falsification, argument, sound and valid, enumerative induction (straight rule), confirmation theory (and formal epistemology), statistical affirming the consequent, guide to life, problem of induction, irrelevant conjunction, likelihood ratio, old evidence problem

Excursion 2 Tour II: Falsification, Pseudoscience, Induction

Tour II visits Popper, falsification, corroboration, Duhem's problem (what to blame in the case of anomalies) and the demarcation of science and pseudoscience (2.3). While Popper comes up short on each, the reader is led to improve on Popper's notions. Central ingredients for our journey are put in place via souvenirs: a framework of models and problems, and a post-Popperian language to speak about inductive inference. Defining a severe test, for Popperians, is linked to when data supply novel evidence for a hypothesis: family feuds about defining novelty are discussed (2.4). We move into Fisherian significance tests and the crucial requirements he set: isolated significant results are poor evidence of a genuine effect, and statistical significance doesn't warrant substantive, e.g., causal inference (2.5). Applying our new demarcation criterion to a plausible effect (males are more likely than females to feel threatened by their partner's success), we argue that a real revolution in psychology will need to be more revolutionary than at present. Whole inquiries might have to be falsified, their measurement schemes questioned (2.6). The Tour's pieces are synthesized in (2.7), where a guest lecturer explains how to solve the problem of induction now, having redefined induction as severe testing.

Excursion 2 Tour II: keywords

Corroboration, Demarcation of science and pseudoscience, Falsification, Duhem's problem Novelty, Biasing selection effects, Simple significance tests, Fallacies of rejection, NHST, Reproducibility and replication

Excursion 3: Statistical Tests and Scientific Inference

Tour I: Ingenious and Severe Tests

We move from Popper to the development of statistical tests (3.2) by way of a gallery on (3.1): Data Analysis in the 1919 Eclipse tests of the General Theory of Relativity (GTR). The tour opens by honing in on where the main members of our statistical cast are in 1919: Fisher, Neyman and Pearson. From the GTR episode, we identify the key elements of a statistical test– the steps we find in E. Pearson's opening description of tests in (3.2). The typical (behavioristic) formulation of N-P tests is as mechanical rules to accept or reject claims with good long run error probabilities. The severe tester breaks out of the behavioristic prison. The classical testing notions–Type I and II errors, power, consistent tests–are shown to grow out of requiring of probative tests. Viewing statistical inference as severe testing, we explore how members of the Fisherian tribe can do all N-P tests do (3.3). We consider the frequentist principle of evidence FEV (Mayo and Cox) and the divergent interpretations that are called for by Cox's taxonomy of null hypotheses. The last member of the taxonomy–substantively based null hypotheses–returns us to the opening episode of GTR.

Tour I: keywords

eclipse test, statistical test ingredients, Type I & II errors, power, P-value, uniformly most powerful (UMP); severity interpretation of tests, severity function, frequentist principle of evidence FEV; Cox's taxonomy of nulls

Excursion 3 Tour II: It's The Methods, Stupid

Tour II disentangles a jungle of conceptual issues at the heart of today's statistical wars. (3.4) unearths the basis for counterintuitive inferences thought to be licensed by Fisherian or N-P tests. These howlers and chestnuts show: the need for an adequate test statistic, the difference between implicationary and actual assumptions, and the fact that tail areas serve to raise, and not lower, the bar for rejecting a null hypothesis. Stop (3.5) pulls back the curtain on an equivocal use of "error probability". When critics allege that Fisherian P-values are not error probabilities, they mean Fisher wanted an evidential not a performance interpretation–this is a philosophical not a mathematical claim. In fact, N-P and Fisher used P-values in both ways. Critics argue that P-values are for evidence, unlike error probabilities, but in the next breath they aver P-values aren't good measures of evidence either, since they disagree with probabilist measures: likelihood ratios, Bayes Factors or posteriors (3.6). But the probabilist measures are inconsistent with the error probability ones. By claiming the latter are what's wanted, the probabilist begs key questions, and misinterpretations are entrenched.

Excursion 3 Tour II keywords

howlers and chestnuts of statistical tests, Jeffreys tail area criticism, two machines with different positions, weak conditionality principle, likelihood principle, long run performance vs probabilism, Neyman vs Fisher, hypothetical long-runs, error probability$_1$ and error probability $_2$, incompatibilism (Fisher & Neyman-Pearson must be separated)

Excursion 3 Tour III: Capability and Severity: Deeper Concepts

A long-standing family feud among frequentists is between hypotheses tests and confidence intervals (CIs). In fact there's a clear duality between the two: the parameter values within the (1 – α) CI are those that are not rejectable by the corresponding test at level α. (3.7) illuminates both CIs and severity by means of this duality. A key idea is arguing from the capabilities of methods to what may be inferred. In (3.8) we reopen a highly controversial matter of interpretation in relation to statistics and the 2012 discovery of the Higgs particle based on a "5 sigma observed effect". Because the 5-sigma standard refers to frequentist significance testing, the discovery was immediately imbued with controversies that, at bottom, concern statistical philosophy. Some Bayesians even hinted it was "bad science". One of the knottiest criticisms concerns the very meaning of the phrase: "the probability our data are merely a statistical fluctuation". Failing to clarify it may impinge on the nature of future big science inquiry. The problem is a bit delicate, and my solution is likely to be provocative. Even rejecting my construal will allow readers to see what it's like to switch from wearing probabilist, to severe testing, glasses.

Excursion 3 Tour III: keywords

confidence intervals, duality of confidence intervals and tests
rubbing off interpretation, confidence level, Higg's particle, look elsewhere effect, random fluctuations, capability curves, 5 sigma, beyond standard model physics (BSM)

Excursion 4: Objectivity and Auditing

Tour I: The Myth of "The Myth of Objectivity"

Blanket slogans such as "all methods are equally objective and subjective" trivialize into oblivion the problem of objectivity. Such cavalier attitudes are at odds with the moves to take back science. The goal of this tour is to identify what there is in objectivity that we won't give up, and shouldn't. While knowledge gaps leave room for biases and wishful thinking, we regularly come up against data that thwart our expectations and disagree with predictions we try to foist upon the world. This pushback supplies objective constraints on which our critical capacity is built. Supposing an objective method is to supply formal, mechanical, rules to process data is a holdover of a discredited logical positivist philosophy.
Discretion in data generation and modeling does not warrant concluding: statistical inference is a matter of subjective belief. It is one thing to talk of our models as objects of belief and quite another to maintain that our task is to model beliefs. For a severe tester, a statistical method's objectivity requires the ability to audit an inference: check assumptions, pinpoint blame for anomalies, falsify, and directly register how biasing selection effects–hunting, multiple testing and cherry-picking–alter its error probing capacities.

Tour I: keywords

objective vs. subjective, objectivity requirements, auditing, dirty hands argument
logical positivism; default Bayesians, equipoise assignments, (Bayesian) wash-out theorems, degenerating program, epistemology: internal/external distinction

Excursion 4 Tour II: Rejection Fallacies: Whose Exaggerating What?

We begin with the *Mountains out of Molehills Fallacy* (large *n* problem): The fallacy of taking a (P-level) rejection of H_0 with larger sample size as indicating greater discrepancy from H_0 than with a smaller sample size. (4.3). The Jeffreys-Lindley paradox shows with large enough *n*, a .05 significant result can correspond to assigning H_0 a high probability .95. There are family feuds as to whether this is a problem for Bayesians or frequentists! The severe tester takes account of sample size in interpreting the discrepancy indicated. A modification of confidence intervals (CIs) is required.
It is commonly charged that significance levels overstate the evidence against the null hypothesis (4.4, 4.5). What's meant? One answer considered here, is that the P-value can be smaller than a posterior probability to the null hypothesis, based on a lump prior (often .5) to a point null hypothesis. There are battles between and within tribes of Bayesians and frequentists. Some argue for lowering the P-value to bring it into line with a particular posterior. Others argue the supposed exaggeration results from an unwarranted lump prior to a wrongly formulated null.We consider how to evaluate reforms based on bayes factor standards (4.5). Rather than dismiss criticisms of error statistical methods that assume a standard from a rival account, we give them a generous reading. Only once the minimal principle for severity is violated do we reject them. Souvenir R summarizes the severe tester's interpretation of a rejection in a statistical significance test. At least 2 benchmarks are needed: reports of discrepancies (from a test hypothesis) that are, and those that are not, well indicated by the observed difference.

Keywords:

significance test controversy, mountains out of molehills fallacy, large n problem, confidence intervals, P-values exaggerate evidence, Jeffreys-Lindley paradox, Bayes/Fisher disagreement, uninformative (diffuse) priors, Bayes factors, spiked priors, spike and slab, equivocating terms, severity interpretation of rejection (SIR)

Excursion 4 Tour III: Auditing: Biasing Selection Effects & Randomization

Tour III takes up Peirce's "two rules of inductive inference": predesignation (4.6) and randomization (4.7). The Tour opens on a court case transpiring: the CEO of a drug company is being charged with giving shareholders an overly rosy report based on post-data dredging for nominally significant benefits. Auditing a result includes checking for (i) selection effects, (ii) violations of model assumptions, and (iii) obstacles to moving from statistical to substantive claims. We hear it's too easy to obtain small *P*-values, yet replication attempts find it difficult to get small *P*-values with preregistered results. I call this the *paradox of replication*. The problem isn't *P*-values but failing to adjust them for cherry picking and other *biasing selection effects*. Adjustments by Bonferroni and false discovery rates are considered. There is a tension between popular calls for preregistering data analysis, and accounts that downplay error probabilities. Worse, in the interest of promoting a methodology that rejects error probabilities, researchers who most deserve lambasting are thrown a handy line of defense. However, data dependent searching need not be pejorative. In some cases, it can improve severity. (4.6)

Big Data cannot ignore experimental design principles. Unless we take account of the sampling distribution, it becomes difficult to justify resampling and randomization. We consider RCTs in development economics (RCT4D) and genomics. Failing to randomize microarrays is thought to have resulted in a decade lost in genomics. Granted the rejection of error probabilities is often tied to presupposing their relevance is limited to long-run behavioristic goals, which we reject. They are essential for an epistemic goal: controlling and assessing how well or poorly tested claims are. (4.7)

Keywords

error probabilities and severity, predesignation, biasing selection effects, paradox of replication, capitalizing on chance, bayes factors, batch effects, preregistration, randomization: Bayes-frequentist rationale, bonferroni adjustment, false discovery rates, RCT4D, genome-wide association studies (GWAS)

Excursion 4 Tour IV: More Auditing: Objectivity and Model Checking

While all models are false, it's also the case that no useful models are true. Were a model so complex as to represent data realistically, it wouldn't be useful for finding things out. A statistical model is useful by being *adequate for a problem, meaning* it enables controlling and assessing if purported solutions are well or poorly probed and to what degree. We give a way to define severity in terms of solving a problem.(4.8) When it comes to testing model assumptions, many Bayesians agree with George Box (1983) that "it requires frequentist theory of significance tests" (p. 57). Tests of model assumptions, also called misspecification (M-S) tests, are thus a promising area for Bayes-frequentist collaboration. (4.9) When the model is in doubt, the likelihood principle is inapplicable or violated. We illustrate a non-parametric bootstrap resampling. It works without relying on a theoretical probability distribution, but it still has

assumptions. (4.10). We turn to the M-S testing approach of econometrician Aris Spanos.(4.11) I present the high points for unearthing spurious correlations, and assumptions of linear regression, employing 7 figures. M-S tests differ importantly from model selection–the latter uses a criterion for choosing among models, but does not test their statistical assumptions. They test fit rather than whether a model has captured the systematic information in the data.

Keywords

adequacy for a problem, severity (in terms of problem solving), model testing/misspecification (M-S) tests, likelihood principle conflicts, bootstrap, resampling, Bayesian p-value, central limit theorem, nonsense regression, significance tests in model checking, probabilistic reduction, respecification

Excursion 5: Power and Severity

Tour I: Power: Pre-data and Post-data

The power of a test to detect a discrepancy from a null hypothesis H_0 is its probability of leading to a significant result if that discrepancy exists. Critics of significance tests often compare H_0 and a point alternative H_1 against which the test has high power. But these don't exhaust the space. Blurring the power against H_1 with a Bayesian posterior in H_1 results in exaggerating the evidence. (5.1) A drill is given for practice (5.2). As we learn from Neyman and Popper: if data failed to reject a hypothesis H, it does not corroborate H unless the test probably would have rejected it if false. A classic fallacy is to construe no evidence against H_0 as evidence of the correctness of H_0. It was in the list of slogans opening Excursion 1. H is corroborated severely only if, and only to the extent that, it passes a test it probably would have failed, if false. By reflecting this reasoning, power analysis avoids such fallacies, but it's too coarse. Severity analysis follows the pattern but is sensitive to the actual outcome (it uses what I call *attained power*). (5.3) Using severity curves we read off assessments for interpreting non-significant results in a standard test. (5.4)

Tour I: keywords

power of a test, attained power (and severity), fallacies of non-rejection, severity curves, severity interpretation of negative results (SIN), power analysis, Cohen and Neyman on power analysis, retrospective power

Excursion 5 Tour II: How not to Corrupt Power

We begin with objections to power analysis, and scrutinize accounts that appear to be at odds with power and severity analysis.(5.5) Understanding power analysis also promotes an improved construal of CIs: instead of a fixed confidence level, several levels are needed, as with confidence distributions. Severity offers an evidential assessment rather than mere coverage probability. We examine an influential new front in the statistics wars based on what I call the diagnostic model of tests. (5.6) The model is a cross between a Bayesian and frequentist analysis. To get the priors, the hypothesis you're about to test is viewed as a random sample from an urn of null hypotheses, a high proportion of which are true. The analysis purports to explain the replication crisis because the proportion of true nulls amongst hypotheses rejected may be higher than the probability of rejecting a null hypothesis given it's true. We question the assumptions and the altered meaning of error probability (error probability$_2$ in 3.6). The Tour links several arguments that use probabilist measures to critique error statistics.

Excursion 5 Tour II: keywords

confidence distributions, coverage probability, criticisms of power, diagnostic model of tests, shpower vs power, fallacy of probabilistic instantiation, crud factors

Excursion 5 Tour III: Deconstructing the N-P vs. Fisher Debates

We begin with a famous passage from Neyman and Pearson (1933), taken to show N-P philosophy is limited to long-run performance. The play, "Les Miserables Citations" leads to a deconstruction that illuminates the evidential over the performance construal. *(5.7)* To cope with the fact that any sample is improbable in some respect, statistical methods either: appeal to prior probabilities of hypotheses or to error probabilities of a method. Pursuing the latter N-P are led to (i) a prespecified test criterion and (ii) consider alternative hypotheses and power. Fisher at first endorsed their idea of a most powerful test. Fisher hoped fiducial probability would both control error rates of a method – performance – as well as supply an evidential assessment. When confronted with the fact that fiducial solutions disagreed with performance goals he himself had held, Fisher abandoned them. (5.8) He railed against Neyman who was led to a performance construal largely to avoid inconsistencies in Fisher's fiducial probability. The problem we face today is precisely to find a measure that controls error while capturing evidence. This is what severity purports to supply. We end with a connection with recent work on Confidence Distributions.

Excursion 5 Tour III: keywords

Bertrand and Borel debate, Neyman-Pearson test development, behavioristic (performance model) of tests, deconstructing N-P (1933), Fisher's fiducial probabilities, Neyman/Fisher feuds, Neyman and Fisher dovetail, confidence distributions

Excursion 6: (Probabilist) Foundations Lost, (Probative) Foundations Found

Excursion 6 Tour I: What Ever Happened to Bayesian Foundations

Statistical battles often grow out of assuming the goal is a posterior probabilism of some sort. Yet when we examine each of the ways this could be attained, the desirability for science evanesces. We survey classical subjective Bayes via an interactive museum display on Lindley and commentators. (6.1) We durvey a plethora of meanings given to Bayesian priors (6.2) and current family feuds between subjective and non-subjective Bayesians. (6.3) The most prevalent Bayesian accounts are default/non-subjective, but there is no agreement on suitable priors. Sophisticated methods give as many priors as there are parameters and different orderings. They are deemed mere formal devices for obtaining a posterior. How then should we interpret the posterior as an adequate summary of information? While touted as the best way to bring in background, they are simultaneously supposed to minimize the influence of background. The main assets of the Bayesian picture–a coherent way to represent and update beliefs–go by the board.(6.4) The very idea of conveying "the" information in the data is unsatisfactory. It turns on what one wants to know. An answer to: how much a prior would be updated, differs from how well and poorly tested claims are. The latter question, of interest to a severe tester, is not answered by accounts that require assigning probabilities to a catchall factor: science must be open ended.

Tour I: keywords

Classic subjective Bayes, subjective vs default Bayesians, Bayes conditioning, default priors (and their multiple meanings), default Bayesian and the Likelihood Principle, catchall factor

Excursion 6 Tour II: Pragmatic and Error Statistical Bayesians

Tour II asks: Is there an overarching philosophy that "matches contemporary attitudes"? Kass's pragmatic Bayesianism seeks unification by a restriction to cases where the default posteriors match frequentist error probabilities.(6.5) Even with this severe limit, the necessity for a split personality remains: probability is to capture variability as well as degrees of belief. We next consider the falsificationist Bayesianism of Andrew Gelman, and his work with others.(6.6) This purports to be an error statistical view, and we consider how its foundations might be developed. The question of where it differs from our misspecification testing is technical and is left open. Even more important than shared contemporary attitudes is changing them: not to encourage a switch of tribes, but to understand and get beyond the tribal warfare. If your goal is really and truly probabilism, you are better off recognizing the differences than trying to unify or reconcile. Snapshots from the error statistical lens lets you see how frequentist methods supply tools for controlling and assessing how well or poorly warranted claims are. If you've come that far in making the gestalt switch to the error statistical paradigm, a new candidate for an overarching philosophy is at hand. Our Fairwell Keepsake delineates the requirements for a normative epistemology and surveys nine key statistics wars and a cluster of familiar criticisms of error statistical methods. They can no longer be blithely put forward as having weight without wrestling with the underlying presuppositions and challenges collected on our journey. This provides the starting point for any future attempts to refight these battles. The reader will then be beyond the statistics wars. (6.7)

Excursion 6 Tour II: keywords

pragmatic Bayesians, falsificationist Bayesian, confidence distributions, epistemic meaning for coverage probability, optional stopping and Bayesian intervals, error statistical foundations