ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Statistical Significance

D.R. Cox

Nuffield College, Oxford University, Oxford OX1 1NF, United Kingdom;
email: david.cox@nuffield.ox.ac.uk

Annu. Rev. Stat. Appl. 2020. 7:1–10

First published as a Review in Advance on August 16, 2019

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-031219-041051

Keywords

statistical significance, probability

Abstract

A broad review is given of the role of statistical significance tests in the analysis of empirical data. Four main types of application are outlined. The first, conceptually quite different from the others, concerns decision making in such contexts as medical screening and industrial inspection. The others assess the security of conclusions. The article concludes with an outline discussion of some more specialized points.

1. INTRODUCTION

Formal probabilistic elements are one component involved in deploying statistical concepts in research. Other broadly statistical ideas aid the precise formulation of research questions for study, the planning of investigations, issues of metrology, and then, depending on the nature of the study, various stages of data collection and analysis. This culminates in the formulation of conclusions, often at various levels of security. These may lead to recommendations for action, including themes for further study. Significance tests may be a component of a number of these phases but are most prominent in the intermediate and concluding phases of analysis, and we focus on these phases in the following.

Now, to claim a result to be highly significant, or even just significant, sounds like enthusiastic endorsement, whereas to describe a result as insignificant is surely dismissive. To help avoid such misinterpretations, the qualified terms statistically significant or statistically insignificant should, at the risk of some tedium, always be used.

A crucial point is that while the pressures from mathematical theory are toward unification and apparent generality, there are a number of quite distinct types of application of significance tests, and distinguishing between these is central to fruitful application and understanding.

2. CRITICAL DEFINITIONS

Suppose that we study a system with haphazard variation and are interested in a hypothesis, H, about the system. We find a test quantity, a function $t(y)$ of data y, such that if H holds, $t(y)$ can be regarded as the observed value of a random variable $t(Y)$ having a distribution under H that is known numerically to an adequate approximation, either by mathematical theory or by computer simulation. Often the distribution of $t(Y)$ is known also under plausible alternatives to H, but this is not necessary. It is enough that the larger the value of $t(y)$, the stronger the pointer against H. These are already formidable abstractions. In general we suppose specified an ordering or partial ordering of the possible values of $t(y)$ in terms of decreasing consistency with H. For ease of exposition we largely consider the one-sided case where the larger the value of $t(y)$, the more the data point against H.

The procedure is now as follows. Given the observed y_{obs}, say, we calculate

$$p_{\text{obs}} = \mathrm{P}\{t(Y) \geq t_{\text{obs}}; H\}, \qquad 1.$$

where $t_{\text{obs}} = t(y_{\text{obs}})$, to be called the observed significance level or p-value. An immediate consequence of the definition is that for continuously distributed random variables, p_{obs} is uniformly distributed on $(0, 1)$ when H holds.

Often $t(\cdot)$ will be some simple measure, for example, of the difference between the locations of two sets of data or based on the comparison of observed and fitted frequencies.

As a less direct example, suppose that Y is normally distributed with known standard deviation and unknown mean μ and that the hypothesis H is that μ is a nonnegative integer representing, for example, the number of carbon atoms in a complex organic molecule. Then $t(y)$ is the distance of the data mean from the nearest nonnegative integer. Any specified nonnegative integer may then be tested for consistency with the data. In particular it may happen that no nonnegative integer is consistent with the data at a reasonable level, suggesting potential incompatibility of data and theory or underestimation of the error variability. More broadly the choice of test statistic may be based sometimes on qualitative reasonableness or sometimes on theoretical considerations of optimal sensitivity in more formalized contexts.

When the discreteness of the distribution of $t(Y)$ is important, there are possible modifications of the definition of p-value just given. One would be to replace the greater than or equals sign in the definition by strictly greater than. Some rewording of the explanation of hypothetical meaning would be required. Another possibility is to define p^* by

$$p^*_{\text{obs}} = \text{P}\{t(Y) > t_{\text{obs}}; H\} + \frac{1}{2}\text{P}\{t(Y) = t_{\text{obs}}; H\}.$$

This has some formal advantage for problems with discrete distributions in making the distribution of p^* under H closer to the continuous uniform distribution on $(0, 1)$ holding for continuously distributed data. For most direct interpretations the definition given in Equation 1 remains preferable. If intrinsic nonlinearities are present, care is needed to ensure the appropriateness of the listing of the potential sample points in order of decreasing consistency with H.

In its primary role, the p-value is an indicator of consistency with H in the respect tested, not a measure of how large or important any departure may be, and obviously not a probability that H is correct. As with other analytical devices, we ask: What is the interpretation of the particular value we obtain? The operational meaning of a particular value—say, for illustration, $p = 0.07$—is as follows. If we were to take this value to be just decisive evidence against H, then in about 7% of a hypothetical long run of such situations in which H is true, it would be falsely rejected. At this stage, at least, the accept-reject decision is purely notional to illustrate meaning.

For example, consider the measurement of height. If pressed as to what a particular numerical measurement means, the answer would involve comparison with an international standard. Unlikely to be realized physically though that comparison may be, it serves to establish the broad comparability of measurements made in different times and places. There is a similar role for p-values. Given, of course, certain specific assumptions, the p-value is an assessment of uncertainty, calibrated by a process that, so far in the discussion, is to be taken as defining meaning, not as a specification for immediate use.

To continue the discussion demands a distinction between a number of quite different situations.

3. DISTINCTIVE TYPES OF APPLICATION

3.1. Preliminary Remarks

We now distinguish four broad situations in which the calculation of p-values is potentially useful. The first corresponds closely to the considerations underlying the Neyman–Pearson theory of testing hypotheses (Neyman & Pearson 1928, 1967), and the next two may often have interval estimation considerations attached to them, whereas the fourth, in spirit, belongs more to exploratory analysis. It may be argued that much recent controversy over significance tests stems from a failure to make such distinctions.

A rather separate distinction centers on the nature of probability in these considerations, and we defer discussion of this to Section 4.

3.2. Two-Decision Situation

The following specification is essentially that familiar from the Neyman–Pearson theory of testing hypotheses. We have a hypothesis H about the distribution of observed random variables and one or more alternative hypotheses. On the basis of data y, we have to decide either to accept or to reject H. In the formulation of Section 2, the test statistic, $t(y)$, and a critical value for it are chosen,

the latter in light of the probability of falsely rejecting H when true, the size. The test statistic itself is chosen to maximize power, the probability of detecting a departure of specified magnitude when present. The endpoint of use is not the assessment of uncertainty in a conclusion but a decision to accept or reject. For formal discussion of this, the distribution of the data has to be specified both for H and for the alternatives.

The following examples are close to this description. A simplified model of some forms of health screening is that individuals are tested annually, the test outcome is scored, and on the basis of that score, each individual is classified as either satisfactory (return next year!) or unsatisfactory (return tomorrow for further examination!). This is, in the first place, a two-decision context. Broadly similar situations arise in other areas, for example, in the auditing of accounts by monetary unit sampling and in industrial quality control through acceptance procedures.

In these examples, definition of the procedure is directly by performance in repeated application, but the probabilities involved are intended to be relevant to each specific application.

We discuss separately in Section 10 the use of power in the choice of study size.

3.3. A Subject-Matter Hypothesis

In the previous subsection a routine repetitive decision between acceptance or rejection was considered. Now we suppose that the hypothesis H is of specific subject-matter interest, and clear evidence of a departure from it in the particular study under analysis is of intrinsic concern. In some contexts H represents conventional theory or expectation, and departure from it is an alarm signal. In others departure represents a novel discovery of intrinsic interest. For an appropriate test criterion $t(y)$, we use p_{obs} as an indicator for evidence of departure. In this context, the p-value is an objective measure of uncertainty, clearly calibrated against performance under hypothetical repetition. At this point the choice of a critical value for decision making, for example, to decide on submission for publication (yes or no) is not the concern. Depending somewhat on the particular field of study, there are broadly agreed-upon conventional guidelines, such as that $p_{\text{obs}} > 0.1$ represents reasonable consistency with H. But that is a separate issue from the use of p_{obs} as a device for communicating to the reader a level of security.

As an example, in the spectacular studies at the Large Hadron Collider at CERN, particles were accelerated around a 25-km underground system in Switzerland and France, and after collision, the spectrum of the energies of the resulting particles was recorded. When energy was grouped into fairly narrow bands, Poisson-distributed counts were obtained showing a smooth trend of frequency with energy. One objective was to find, if it existed, evidence for the Higgs boson. The presence of this particle would generate an excess frequency in one prespecified energy band. Here, there was a very clearly specified point null hypothesis, that the Higgs does not exist, the observations then forming pure noise of a particular structure. There is no specific relevant quantitative alternative.

Other examples in the natural sciences concern testing consistency with well-established relations, such as that of the dependence on temperature of the rate of a chemical reaction. A rather different type of example concerns the comparison of adverse reaction reports on two medical treatments in a situation where any differences between the treatments are initially thought minor and unlikely to influence the occurrence of adverse events.

3.4. Dividing Hypotheses

A quite different situation arises when the notional hypothesis H divides the possible parameter values into two sets, for example, A giving larger mean response than B versus B giving larger mean

response than A, with no particular direct prior expectation in the hypothesis that they are the same. Have the data clearly established the direction of an effect? A two-sided test of significance may be used. In effect, the outcome of the test specifies the level at which a confidence interval for the difference contains only points of the same sign.

3.5. Tests of Model Adequacy

The final group of applications is the most informal. Particularly in fitting relatively complicated models, and perhaps especially when there is little experience of the field on which to draw, there may be many aspects of the assumed model that need scrutiny. Examples are checks of assumptions of linearity; of the absence of interaction effects; of distributional form, for example normality; and of stability of variance, as well as absence of seriously anomalous observations. Significance tests play an important, if largely informal, role in this aspect of statistical analysis, signaling which apparently anomalous features of data are beyond those expected under the inevitably oversimplified initial model. Judgement is likely to be required in deciding which anomalies are likely, in the specific application, to be important enough to warrant a modification, possibly major, of the original analysis.

4. ANALYSES BASED ON STUDY RANDOMIZATION

In the discussion so far, probability enters to represent the haphazard character of some parts of the variation present. In some contexts of carefully planned experimental or observational investigation based on appropriate sampling, analysis may be based directly on the randomization used in design. Thus, in a randomized experiment to compare, say, a treatment with a control, a null hypothesis might be taken to be that each study individual has a response totally unaffected by the condition allocated to it and other individuals. A simple extension deals with confidence intervals for a hypothesized additive treatment effect. The configuration of observations recorded can, under such an assumption, then be compared with that which would have been observed under a different outcome of the allocation process. Somewhat similar arguments apply in a sampling context when individuals are chosen at random from a precisely specified target population. With current computational resources the enumerations required are feasible, whereas earlier, the argument was largely restricted to small problems or required mathematical approximations.

Except when the study individuals are chosen by well-defined selection probabilities from a specific target population, randomization inferences do not formally allow generalization. They focus on what really happened when the current set of data were generated.

5. ONE- AND TWO-SIDED TESTS

An elementary but sometimes treacherous point concerns the distinction between one- and two-sided tests. In some situations the only departures of interest from H are in one direction, as assessed by the test statistic $t(y)$. The illustration concerning the Higgs boson is an example. In other cases departures from H may reasonably be treated symmetrically, and the test statistic t may be taken as the square of a suitably defined one-sided test statistic.

Sometimes, however, the magnitudes of departures in different directions are not easily comparable. Then it would be more appropriate to define, for each y, two different p-values, one for each type of alternative, and to define the two-sided p-value as twice the smaller of the one-sided values. Reporting only a one-sided value when an effect in either direction should have been considered invites misinterpretation.

6. SUMMARY OF ROLES

In outline, then, we distinguish between the roles of tests in routine repetitive decision making, in communicating aspects of the uncertainty involved in drawing specific conclusions about the primary questions under study, in drawing conclusions about the direction of dependences, and finally in guiding, to some extent informally, the formulation of the questions under investigation.

It might possibly be thought that the general discussion of significance tests should have been preceded by an account of the general principles of statistical inference, including the nature and role of probability in these issues. Indeed, underlying the discussion is a tension between the desire for general principles from which special cases emerge, as contrasted with a need to check that general theory is indeed relevant for each application. As is more broadly the case for theory in science, there is a constant interplay between theory and application, not one-way traffic. An extreme example is quantum mechanics, which is spectacularly successful at prediction of often exotic phenomena—yet its foundations are, at a deep level, controversial.

7. THE ROLE OF PROBABILITY

The discussion above has focused on the distinct roles of significance tests in direct applications. Underpinning this account, however, is the use of probability. While not the first to develop a formal basis, Kolmogorov's (1933) axiomization of probability liberated that topic from discussion of what probability "means" and freed probability theory to develop into a vibrant part of modern pure mathematics. Yet Kolmogorov himself retained an interest in various interpretations of probability, and any discussion of statistical analysis that hinges on probabilistic representations cannot evade discussions of meaning.

The pioneers R.A. Fisher and Harold Jeffreys both used two versions of probability. Jeffreys used the word "chance" to describe the empirical behavior of haphazard sequences and "probability" to assess rational degree of belief in an uncertain proposition given specified evidence. Special forms, possibly improper, of distribution were used to represent initial ignorance. For example, in estimation of the unknown mean μ of a normal distribution, Jeffreys, in essence following Laplace, took the prior distribution of μ to be uniform over a very long interval, in the limit represented by a uniform density over all real numbers. For estimation, but not for hypothesis testing, this approach gives answers appealing from several perspectives, so long as high-dimensional parameter spaces are not involved.

Fisher used the first form of probability and, especially for interpretive purposes, the notion that the data are a random sample from a hypothetical infinite population. Another focus, stemming from aspects of Fisher's work, emphasized likelihood itself as a primary concept (Edwards 1992, Royall 1997). Neyman, certainly in his earlier work, took a monolithic view of probability as referring to frequencies in repetitive contexts; in principle, at least, nothing could be said referring solely to an individual analysis, and he appeared not to cover relatively flexible use of observed p-values. In his later, more applied work, Neyman departed somewhat from that and used p-values flexibly, whereas Fisher paradoxically, in some at least of his work, used 5% significance rather rigidly, although he recognized the arbitrariness of that specific choice. L.J. Savage (1954) advocated a single view, based on probability as concerned with the approach to uncertainty of an individual, You, constrained to self-consistency, so-called coherency. In the present context, he emphasized the modification to the probability of H in passing from prior to posterior distributions.

As noted above, a formally quite distinct route to the incorporation of probability is via the random allocation used in experimental design or in sample selection from a specific target population, the former stemming from Fisher's (1935) discussion of the lady tasting tea.

8. RELATION WITH INTERVAL ESTIMATION

The relation between significance tests and estimation is often quite direct. If interest is focused on one component, θ, say, the null hypothesis specifies that $\theta = \theta_0$, a known value. In strongly specified contexts, formal arguments are available for indicating the most appropriate test statistic, $t(y)$, and this is often directly connected with estimating the parameter θ, in particular with whether a confidence interval for θ at an appropriate level does or does not contain θ_0. Because estimation of such intervals may seem often a richer and more constructive theme than testing a hypothesis, it may seem tempting wherever possible to regard the theory of testing of hypotheses as subsidiary to that of interval estimation.

There are, however, powerful reasons for not taking sets of confidence intervals as the primary concept, essentially because of the possibility that nested sets of confidence or posterior intervals may not be the appropriate summary of information. An unusual, but not artificial, example concerns the ratio of the means of two normal distributions, where the appropriate summary may be the outside of an interval or even the whole real line. In the latter case the confidence set is not a vacuous statement that the parameter is a real number, but rather a stark warning of the fragility of the data.

9. VERY SMALL *p*-VALUES

It is common, especially in some genomic applications that lead to a very large number of notional p-values, to focus interpretation on very small values of p, usually conveniently expressed as values of $-\log p$ of, say, 10 or more. In a direct sense it is implausible that such very small values of p have a secure face-value interpretation of the kind emphasized in the present article. A possible reconciliation, essentially outside the scope of the current discussion, is to regard the analysis as one of estimation rather than significance testing. For this one might treat each p-value as, in effect, derived from estimating a normally distributed effect of unknown size θ on a scale corresponding to a single observation. Then, approximately, $p = \Phi(-\theta/\sqrt{n})$. That is, the value $\sqrt{n}\Phi^{-1}(1-p)$ can be regarded as estimating, for each of the more extreme studies, a notional standardized magnitude of effect. This may be the basis of tentative comparisons, in particular of conclusions from differing amounts, n, of data.

10. IMPLICATIONS FOR STUDY DESIGN

One common implicit use of significance tests is in the design phase of an investigation, to determine an appropriate scale of investigation, often a sample size. This involves specifying first a hypothesis, often the absence of a difference between two or more treatments or conditions, and then an alternative, a difference important enough to require detectability if present. These specify two probabilities; in the literature, they are called size and power. These are the probabilities of rejecting H when true and when the alternative holds. Given this specification, what sample size is needed? Alternatively, if the maximum feasible sample size is known, is the sensitivity that would be achievable sufficient to make the study viable? In these contexts the use of power is well established by convention, even though, in a sense, the approach is clumsy; many combinations of the defining features lead to the same sample size. Specification of a target standard error for a key contrast is more economical.

A more important issue, however, is that power is virtually irrelevant after the data become available. The study may have high power, and the data may fall in the middle of the indecisive region. Or the study may have low power, and yet the data may be sufficiently extreme to point to

a reasonably clear conclusion. By contrast, the standard error estimated from the data can be compared with that projected in design, and the reason for any serious discrepancy may be explored. That is, the standard error considered at the design phase may still be informative after obtaining the data.

11. SIGNIFICANCE TESTS AND BIG DATA

Particular caution is needed in applying standard types of significance test and related methods to so-called big data. First there is a broad issue of data quality. Some large studies, such as that at CERN mentioned above, provide data undoubtedly of very high quality, as do some large scale epidemiological investigations (Cox et al. 2018). In large studies the standard errors of derived statistics calculated in a standard way will often decrease proportionally to the reciprocal of the square root of sample size and become extremely small. While this may sometimes be reasonable, it seems likely that in many contexts, superficially minor sources of variability and internal correlations among the components of variation present may seriously inflate the errors as assessed by standard statistical methods (Cox 2015). Powerful simulation methods such as the bootstrap provide strong protection against unwarranted assumptions of distributional form but, at least as usually presented, have strong implicit assumptions of mutual independence. In some cases so-called self-similar patterns of variation may be present, for which the standard errors decrease at a rate depending on a power of sample size less than one half.

An important and largely open issue concerns the realistic assessment of precision from such data.

12. LARGE NUMBERS OF SIGNIFICANCE TESTS

In some contexts the outcomes $p_1, \ldots, p_m$ of quite large numbers of essentially independent tests need collective assessment. Concentration on $\min(p_1, \ldots, p_m)$ may be an oversimplification. Schweder & Spjøtvoll (1982) suggested study of the frequency distribution of the $\{p_j\}$ near the origin. Visual interpretation is aided by transformation to place more emphasis on the small values. One commonly used possibility is to examine the ordered values of the Gaussian tail areas, $\Phi^{-1}(p_j)$, compared with the expected values of normal order statistics. Benjamini & Hochberg (1995), in an influential contribution, concentrated on a single measure referring to the simultaneous correctness of all statements made.

An alternative is to consider the Rényi decomposition (Rényi 1953) of the transformed values, $z_j = -\log p_j$.

If the random variable P_j is uniformly distributed on $(0, 1)$, then Z_j has a unit exponential distribution. The ordered values

$$Z_{(1)} < Z_{(2)} < \cdots < Z_{(n)}$$

have the representation

$$Z_{(1)} = V_1/n, Z_{(2)} = V_1/n + V_2/(n-1), \ldots, Z_{(n)} = V_1/n + \cdots + V_{n-1}/2 + V_n,$$

where the V_j are independently exponentially distributed with unit mean. This and modifications of it can be the basis for formal and informal analysis of sets of p-values, having the advantage that the small p-values are emphasized. In particular, a modification based on the notion that

the p-values have a null density proportional to $p^{-\gamma}$ for $\gamma < 1$ allows for smooth failure of the assumptions underlying the original calculation of p.

13. A BRIEF HISTORY

It was a temptation easily resisted to start this article with a historical review. Instead the article concludes with some fragmentary remarks about the very extensive literature and background. Hald (2007) has reviewed developments up to about 1935 with final emphasis on the contributions of R.A. Fisher. Stigler (2016) has written with penetrating insight about more recent developments.

Fisher's pathbreaking general contributions, first set out in particular in Fisher (1922, 1926), may be approached from his last book (Fisher 1956), but individual papers give more focus. For accounts putting primary emphasis on likelihood itself, readers are directed to Edwards (1992) and Royall (1997). Neyman & Pearson's (1928, 2967) work on testing hypotheses, which began with the intention of clarifying Fisher's work, is best approached by the fine book of Lehmann (1959); in a very interesting reflection published posthumously, Lehmann (2011) regretted his isolation from the contributions of Fisher.

A definitive approach to statistics from an objective Bayesian viewpoint is given by Jeffreys (1939 and subsequent editions). Some difficulties are implicit, not so much with Jeffreys's use of flat improper priors in a low-dimensional estimation context but in their use in integrating over potential alternative hypotheses when testing a point null hypothesis.

Mayo (2018) has discussed these issues from the perspective of a philosopher of science.

Savage (1954) emphasized a personalistic approach in which probability represents the strength of belief of an individual, You, in an uncertain proposition, required to be coherent, that is, self-consistent. Coherency appeared to take precedence over accord with the real situation. This approach gives a different perspective on the role of hypotheses and their testing. Despite the interest of this approach as addressing the attitude of individuals, it seems largely inappropriate for public representation of uncertainty, the ultimate focus of much statistical work.

Emanating largely from a statement from the American Statistical Association (Wasserstein & Lazar 2016), there has been recent discussion of the supposed role of testing hypotheses, especially, but not only, in the reporting of conclusions from biomedical studies. This emphasizes, differently from the perspective of the present article, the achievement of a preassigned and rather extreme level of significance as a proposed requirement for publication of the results of a study claiming a beneficial effect of, for example, a new medical procedure.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I thank Heather Battey, Anthony Davison, and Nancy Reid for very helpful comments and advice.

LITERATURE CITED

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57:289–300

Cox DR. 2015. Big data and precision. *Biometrika* 102:712–16
Cox DR, Kartsonaki C, Keogh R. 2018. Big data: some statistical issues. *Stat. Probab. Lett.* 136:111–15
Edwards AWF. 1992. *Likelihood*. Cambridge, UK: Cambridge Univ. Press. 2nd ed.
Fisher RA. 1922. On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. A* 222:309–68
Fisher RA. 1926. The arrangement of field trials. *J. Minist. Agric.* 33:503–13
Fisher RA. 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd
Fisher RA. 1956. *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd
Hald A. 2007. *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713 to 1935*. New York: Springer
Jeffreys H. 1939. *Theory of Probability*. Oxford, UK: Oxford Univ. Press
Kolmogorov AN. 1933. *Grundbegriffe der Wahrsceinlichkeitsrechnung*. Berlin: Springer
Lehmann EL. 1959. *Testing Statistical Hypotheses*. New York: Wiley. 1st ed.
Lehmann EL. 2011. *Fisher, Neyman, and the Creation of Classical Statistics*. New York: Springer
Mayo D. 2018. *Statistical Inference as Severe Testing*. Cambridge, UK: Cambridge Univ. Press
Neyman J, Pearson ES. 1928. On the use and interpretation of certain test statistics for purposes of statistical inference: part I. *Biometrika* 20A:175–240
Neyman J, Pearson ES. 1967. *Joint Statistical Papers of J. Neyman and E.S. Pearson*. Cambridge, UK: Cambridge Univ. Press
Rényi A. 1953. On the theory of order statistics. *Acta Math. Acad. Sci. Hung.* 4:191–231
Royall R. 1997. *Statistical Evidence: A Likelihood Paradigm*. Boca Raton, FL: Chapman and Hall
Savage LJ. 1954. *Foundations of Statistics*. New York: Wiley
Schweder T, Spjøtvoll E. 1982. Plots of *p*-values to evaluate many tests simultaneously. *Biometrika* 69:493–502
Stigler S. 2016. *The Seven Pillars of Statistical Wisdom*. Cambridge, MA: Harvard Univ. Press
Wasserstein R, Lazar N. 2016. The ASA's statement on *p*-values: context, process and purpose. *Am. Stat.* 70:129–33

Annual Review of Statistics and Its Application

Volume 7, 2020

Contents

Statistical Significance
D.R. Cox 1

Calibrating the Scientific Ecosystem Through Meta-Research
Tom E. Hardwicke, Stylianos Serghiou, Perrine Janiaud, Valentin Danchev, Sophia Crüwell, Steven N. Goodman, and John P.A. Ioannidis 11

The Role of Statistical Evidence in Civil Cases
Joseph L. Gastwirth 39

Testing Statistical Charts: What Makes a Good Graph?
Susan Vanderplas, Dianne Cook, and Heike Hofmann 61

Statistical Methods for Extreme Event Attribution in Climate Science
Philippe Naveau, Alexis Hannart, and Aurélien Ribes 89

DNA Mixtures in Forensic Investigations: The Statistical State of the Art
Julia Mortera 111

Modern Algorithms for Matching in Observational Studies
Paul R. Rosenbaum 143

Randomized Experiments in Education, with Implications for Multilevel Causal Inference
Stephen W. Raudenbush and Daniel Schwartz 177

A Survey of Tuning Parameter Selection for High-Dimensional Regression
Yunan Wu and Lan Wang 209

Algebraic Statistics in Practice: Applications to Networks
Marta Casanellas, Sonja Petrović, and Caroline Uhler 227

Bayesian Additive Regression Trees: A Review and Look Forward
Jennifer Hill, Antonio Linero, and Jared Murray 251

Q-Learning: Theory and Applications
Jesse Clifton and Eric Laber 279

Representation Learning: A Statistical Perspective
Jianwen Xie, Ruiqi Gao, Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu 303

Robust Small Area Estimation: An Overview
Jiming Jiang and J. Sunil Rao 337

Nonparametric Spectral Analysis of Multivariate Time Series
Rainer von Sachs 361

Convergence Diagnostics for Markov Chain Monte Carlo
Vivekananda Roy 387

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics