Non-Experimental Data, Hypothesis Testing, and the Likelihood Principle: A Social Science Perspective

Other titles in Foundations and Trends® in Econometrics

A Simple Method for Predicting Covariance Matrices of Financial Returns
Kasper Johansson, Mehmet G. Ogut, Markus Pelger, Thomas Schmelzer and Stephen Boyd
ISBN: 978-1-63828-308-9

Factor Extraction in Dynamic Factor Models: Kalman Filter Versus Principal Components
Esther Ruiz and Pilar Poncela
ISBN: 978-1-63828-096-5

Performance Analysis: Economic Foundations and Trends
Valentin Zelenyuk
ISBN: 978-1-68083-866-4

Experimetrics: A Survey
Peter G. Moffatt
ISBN: 978-1-68083-792-6

Climate Econometrics: An Overview
Jennifer L. Castle and David F. Hendry
ISBN: 978-1-68083-708-7

Non-Experimental Data, Hypothesis Testing, and the Likelihood Principle: A Social Science Perspective

Tom Engsted
Aarhus University
tengsted@econ.au.dk

Jesper W. Schneider
Aarhus University
jws@ps.au.dk

the essence of knowledge
Boston — Delft

Foundations and Trends® in Econometrics

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

T. Engsted and J. W. Schneider. *Non-Experimental Data, Hypothesis Testing, and the Likelihood Principle: A Social Science Perspective*. Foundations and Trends® in Econometrics, vol. 13, no. 1, pp. 1–66, 2024.

ISBN: 978-1-63828-325-6

Foundations and Trends® in Econometrics
Volume 13, Issue 1, 2024

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Econometrics publishes survey and tutorial articles in the following topics:

- Econometric Models
- Simultaneous Equation Models
- Estimation Frameworks
- Biased Estimation
- Computational Problems
- Microeconometrics
- Treatment Modeling
- Discrete Choice Modeling
- Models for Count Data
- Duration Models
- Limited Dependent Variables
- Panel Data
- Time Series Analysis
- Latent Variable Models
- Qualitative Response Models
- Hypothesis Testing
- Econometric Theory
- Financial Econometrics
- Measurement Error in Survey Data
- Productivity Measurement and Analysis
- Semiparametric and Nonparametric Estimation
- Bootstrap Methods
- Nonstationary Time Series
- Robust Estimation

Information for Librarians

Foundations and Trends® in Econometrics, 2024, Volume 13, 4 issues. ISSN paper version 1551-3076. ISSN online version 1551-3084. Also available as a combined paper and online subscription.

Contents

1 Introduction **3**

2 Testing Methodologies **9**
2.1 Frequentist Tests 9
2.2 The Likelihood Principle 16
2.3 The Bayes Factor 18
2.4 Implicit Priors in Frequentist Tests, and False Discoveries . 27

3 Non-Experimental Social Science Data **32**
3.1 Super-Populations 35
3.2 Haavelmo's Framework 38
3.3 Informal Measures of Fit for Misspecified Models 44

4 An Alternative Paradigm for the Social Sciences **46**

References **57**

Non-Experimental Data, Hypothesis Testing, and the Likelihood Principle: A Social Science Perspective

Tom Engsted[1] and Jesper W. Schneider[2]

[1]*Department of Economics and Business Economics, Aarhus University, Denmark; tengsted@econ.au.dk*
[2]*Danish Centre for Studies in Research & Research Policy, Department of Political Science, Aarhus University, Denmark; jws@ps.au.dk*

ABSTRACT

We argue that frequentist hypothesis testing – the dominant statistical evaluation paradigm in empirical research – is fundamentally unsuited for analysis of the non-experimental data prevalent in economics and other social sciences. Frequentist tests comprise incompatible repeated sampling frameworks that do not obey the Likelihood Principle (LP). For probabilistic inference, methods that are guided by the LP, that do not rely on repeated sampling, and that focus on model comparison instead of testing (e.g., subjectivist Bayesian methods) are better suited for passively observed social science data and are better able to accommodate the huge model uncertainty and highly approximative nature of structural models in the social sciences. In addition to

*We thank Esben Høg, Søren Johansen, Jan P. de Ruiter, Dick Startz, Eric-Jan Wagenmakers, Allan Würtz, and an anonymous referee for useful comments on an earlier version of this work.

Tom Engsted and Jesper W. Schneider (2024), "Non-Experimental Data, Hypothesis Testing, and the Likelihood Principle: A Social Science Perspective", Foundations and Trends® in Econometrics: Vol. 13, No. 1, pp 1–66. DOI: 10.1561/0800000048.

formal probabilistic inference, informal model evaluation along relevant substantive and practical dimensions should play a leading role. We sketch the ideas of an alternative paradigm containing these elements.

Keywords: Frequentist versus Bayesian analysis; observational social science data; super-populations; Haavelmo's framework; misspecified models; formal statistical versus informal model evaluation.

JEL Codes: B23, B41, C11, C12, C18, C52.

1

Introduction

The presumed replication crisis in many empirical sciences has revitalized methodology discussions and led to renewed emphasis on model construction and evaluation, the proper conduct of statistical analysis, and the pros and cons of different statistical methodologies (e.g., Ioannidis, 2005; Nuzzo, 2014; Benjamin *et al.*, 2018; Amrhein *et al.*, 2019; Wasserstein *et al.*, 2019; McShane *et al.*, 2019). To highlight the importance of these issues, The American Statistical Association recently took the unprecedented step of issuing an official statement addressing the widespread misunderstandings and misuse of statistical inference in empirical research (Wasserstein and Lazar, 2016). In this monograph we offer new perspectives on these discussions with emphasis on the special problems and challenges facing social scientists.

Empirical analyses in the social sciences are typically based on non-experimental, passively observed data samples that are not directly repeatable in the same way as in randomized controlled experiments – the "gold standard" of traditional statistics. Nonetheless, statistical analysis of observational social science data most often takes place within the classical frequentist statistical paradigm that builds on the "principle of repeated sampling" where "*statistical procedures are to be* *assessed by their behaviour in hypothetical repetitions under the same conditions*" and where "*measures of uncertainty are to be interpreted as hypothetical frequencies in long run repetitions*" (Cox and Hinkley, 1974, p. 45).

In many cases empirical social science researchers apply the classical estimation and test procedures on their observational data uncritically and with no discussion of possible inadequacies. In the few instances where the distinction between the underlying repeated sampling/controlled experiments framework in frequentist theory and the actual observational data is noted and discussed, the analysis is typically justified by reference to "super-populations," or to the work of Haavelmo (1944) who argued that frequentist likelihood and Neyman-Pearson procedures are applicable also for non-experimental social science data.

Another fundamental problem for social scientists is that since social and behavioral relationships are extremely complex and notoriously unstable, models of social and economic behavior must be highly stylized and built on many simplifying and "unrealistic" assumptions. This implies that when we take our models to the data, they do not fit well into the classical statistical paradigm where deviations between model and data reflect pure random error. The methodology of testing (and rejecting) statistical hypotheses can be considered natural within the traditional Popperian falsificasionist paradigm, but the problem with applying this paradigm in the social sciences is that our models by construction are false to an extent that it is relatively easy to reject them. This does not mean, however, that the models may not contain important elements of truth. The statistician George Box's famous bonmot "*All models are wrong, but some are useful*" is particularly relevant for models of human behavior in sociology, psychology, political science, management and economics. Another way to put this is: "*It is not easy to construct an interesting economic theory which cannot be rejected out of hand*" (Keuzenkamp, 2000, p. 9).

In addition, it has become clear that the uncertainty surrounding statements and predictions from our empirical models is much higher than what can be measured from the traditional standard errors and confidence intervals of the estimated parameters in these models. As an example, to our knowledge not a single econometric model – published before 2008 – predicted the financial crisis in 2008–2009 and the subsequent worldwide recession and economic turmoil. The few economists who were able to foresee (parts of) what was coming (e.g., Shiller, 2005, preface and ch. 2), did not use sophisticated statistical or econometric models but simple descriptive analyses and basic common sense.[1] There is a real uncertainty associated with our formal empirical models that is much larger than what we usually acknowledge, and such model uncertainty does not fit easily into the traditional statistical/econometric framework.

In this monograph we discuss the conceptual and interpretational problems of classical frequentist tests in the context of observational non-experimental data, and the justifications, if any, social scientists typically have advanced for the suitability of frequentist tests on such data (Sections 2 and 3). The basic question is: does it make sense to apply frequentist testing procedures, that fundamentally build on repeated sampling, on social science data that are fundamentally non-repeatable? We compare the frequentist testing framework with the Bayesian framework of testing statistical hypotheses and comparing models based on Bayes factors. In contrast to the frequentist framework, the Bayesian framework does not rely on repeated sampling but instead follows the so-called "Likelihood Principle" (LP). It is well-known among statisticians – but not among social scientists – that frequentist testing procedures conflict with the LP according to which likelihood functions that are proportional to each other should lead to the same statistical inference (Berger and Wolpert, 1988). The LP implies that a model's likelihood function contains all relevant information from a given sample about the model parameters. Frequentist tests do not obey this principle because they involve tail area probabilities of hypothetical data that are not part of the likelihood function (e.g., the classical p-value measures the probability of the observed data *or more extreme*

[1] Another more recent example is the complete surprise to everyone – including econometric inflation forecasters – of the spike in worldwide inflation starting in 2021 and continuing during 2022.

data under the null hypothesis, cf. Section 2.1).[2] Not obeying the LP has profound implications for the proper conduct of frequentist tests, implications that are most often not recognized by empirical social scientists. In theory, frequentist tests require a pre-specified and fixed sampling plan to an extent that is close to meaningless, at least when dealing with observational social science data (Berger and Wolpert, 1988; Wagenmakers, 2007). Methods that obey the LP are more flexible in this respect because they do not rely on the exact sampling plan, but only on the likelihood function.[3]

Others before us have discussed these issues. The problems with frequentist tests are well-described in the statistics literature, and the special challenges facing social scientists working with non-experimental data are also well-known. For example, the fundamental problem of model uncertainty is the underlying motivation for Leamer's (1978, 1983) "extreme bounds analysis." Nonetheless, it seems to us that these problems are either forgotten or neglected in much of todays empirical work. Earlier, both of us have expressed concerns about the dominance of the frequentist testing paradigm, Schneider (2013, 2015, 2016, 2018) in the fields of information science and scientometrics, and Engsted (2002, 2009) in economics and econometrics; concerns not least spawned by long-term experience with applying frequentist tests in our own empirical research. Since the 1980s alternatives to "statistical significance" as the main model evaluation tool have appeared in the economics literature, and in Engsted (2002, 2009) one of us expressed the belief that such alternatives – that focus more on "economic significance" – would gradually replace statistical significance. Unfortunately, this has not happened in general, albeit in a few sub-fields.[4] The classical

[2]The Likelihood Principle is not to be confused with the "principle of likelihood in testing hypotheses" described in Neyman and Pearson (1933, p. 295), which consists in comparing a likelihood ratio to a "critical region," cf. Section 2.1.

[3]We do not discuss the differences between frequentist and Bayesian estimation procedures because these differences are not nearly as profound as the differences between frequentist and Bayesian testing procedures.

[4]Engsted (2009) commented on Ziliak and McCloskey (2008) who criticized the practice of econometrics. In retrospect, the last fifteen years have proven Ziliak and McCloskey mostly right in their scepticism about the future of econometric practice. The "null ritual" described by Gigerenzer (2004) is still widely practiced.

frequentist testing paradigm continues to dominate empirical scientific work, including research in economics and other social sciences, and statistical significance at the 5% level remains a target that researchers – and journal editors and reviewers – strongly emphasize, albeit often implicitly (Harvey, 2017; Andrews and Kasy, 2019), leading to what Gigerenzer and Marewski (2015) call "surrogate science."[5] New and innovative tools also embrace this paradigm. For example, the classical p-value "*has become firmly embedded in the minds and habits of machine learning researchers*" (Berrar, 2022, pp. 1102–1103). Given the massive conceptual and interpretational problems with frequentist tests, this is an unfortunate state of affairs. We believe that there is still a need to address these matters; hence, this monograph.[6]

Hill (1985) and Poirier (1988) encouraged economists to apply subjectivist statistical methods that obey the LP. We think it is time to repeat this advice. We end the monograph by presenting (in Section 4) some Bayesian inspired ideas of an alternative formal paradigm that is guided by the LP and does not involve model choice based on hypothesis testing in the traditional sense. Instead it focuses on comparing models probabilistically based on combining personal prior views of model uncertainty with the information in the data. We believe that such a paradigm is more transparent, more flexible, and better reflects the way social scientists think and talk about social, behavioral, and economic models. In addition, we believe this paradigm addresses model uncertainty and the highly approximative nature of social science models in a more satisfactory way than the traditional paradigm does. The alternative paradigm does not rely on any repeated sampling aspects (or similar, like the tail area probability of hypothetical data in the p-value)

[5]Some journals have begun to downplay conventional significance levels. For example, the AER Guidelines for Accepted Articles states: "Do not use asterisks to denote significance of estimation results. Report standard errors in parentheses."

[6]The so-called "credibility revolution" in empirical microeconomics, with its strong focus on research design using experimental and quasi-experimental methods, has been seen as a big step forward in securing a more trustworthy empirical practice (Angrist and Pischke, 2010). It is noteworthy, however, that p-hacking and publication bias with strong reliance on conventional significance thresholds continue to dominate also this field (Brodeur *et al.*, 2020).

and thereby fit more naturally with the non-experimental observational data that social scientists typically work with.

We emphasize, however, that when it comes to the *evaluation* of structural models in the social sciences, a formal statistical framework (whether Bayesian, frequentist, or otherwise) should in our view play only a secondary role. Informal measures of fit with focus on substantive and practical significance are to be preferred over measures based on statistical significance. We elaborate these thoughts in Section 4.

References

Amrhein, V., S. Greenland, B. McShane, and + 800 Signatories (2019). "Retire statistical significance". *Nature*. 567: 305–307. DOI: 10.1038/d41586-019-00857-9.

An, S. and F. Schorfheide (2007). "Bayesian analysis of DSGE models". *Econometric Reviews*. 26: 113–172. DOI: 10.1080/07474930701220071.

Andrews, I. and M. Kasy (2019). "Identification of and correction for publication bias". *American Economic Review*. 109: 2766–2794. DOI: 10.1257/aer.20180310.

Angrist, J. D. and J.-S. Pischke (2010). "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics". *Journal of Economic Perspectives*. 24: 3–30. DOI: 10.1257/jep.24.2.3.

Barillas, F. and J. Shanken (forthcoming). "Comparing priors for comparing asset pricing models". *Journal of Finance*.

Barillas, F. and J. Shanken (2018). "Comparing asset pricing models". *Journal of Finance*. 73: 715–754. DOI: 10.1111/jofi.12607.

Barnett, V. (1999). *Comparative Statistical Inference*. 3rd ed. New York: John Wiley & Sons.

Benjamin, D. J. *et al.* (2018). "Redefine statistical significance". *Nature Human Behaviour*. 2: 6–10.

Berger, J. O. (2003). "Could Fisher, Jeffreys and Neyman have agreed on testing? (with discussion)". *Statistical Science*. 18: 1–32.

Berger, J. O. and M. Delampady (1987). "Testing precise hypotheses (with discussion)". *Statistical Science*. 2: 317–352.

Berger, J. O. and T. Sellke (1987). "Testing a point null hypothesis: The irreconcilability of p values and evidence (with discussion)". *Journal of the American Statistical Association*. 82: 112–139.

Berger, J. O. and R. L. Wolpert (1988). *The Likelihood Principle*. Monograph Series, Vol. 6. Hayward, California: Institute of Mathematical Statistics.

Berk, R. A. and D. A. Freedman (2003). "Statistical assumptions as empirical commitments". In: *Law, Punishment, and Social Control: Essaues in Honor of Sheldon Messinger*. Ed. by T. G. Blomberg and S. Cohen. New York: Aldine de Gruyter.

Berk, R. A., B. Western, and R. E. Weiss (1995a). "Statistical inference for apparent populations (with discussion)". *Sociological Methodology*. 25: 421–485.

Berk, R. A., B. Western, and R. E. Weiss (1995b). "Reply to Bollen, Firebaugh, and Rubin". *Sociological Methodology*. 25: 481–485. DOI: 10.2307/271077.

Berrar, D. (2022). "Using p-values for the comparison of classifiers: Pitfalls and alternatives". *Data Mining and Knowledge Discovery*. 36: 1102–1139. DOI: 10.1007/s10618-022-00828-1.

Birnbaum, A. (1962). "On the foundations of statistical inference". *Journal of the American Statistical Association*. 57: 269–306. DOI: 10.1080/01621459.1962.10480660.

Bollen, K. A. (1995). "Apparent and nonapparent significance tests". *Sociological Methodology*. 25: 459–468. DOI: 10.2307/271074.

Brodeur, A., N. Cook, and A. Heyes (2020). "Methods matter: p-hacking and publication bias in causal analysis in economics". *American Economic Review*. 110: 3634–3660. DOI: 10.1257/aer.20190687.

Campbell, J. Y. and R. J. Shiller (1987). "Cointegration and test of present value models". *Journal of Political Economy*. 95: 1062–1088. DOI: 10.1086/261502.

Chib, S., X. Zeng, and L. Zhao (2020). "On comparing asset pricing models". *Journal of Finance*. 75: 551–577. DOI: 10.1111/jofi.12854.

Christ, C. F. (1994). "The Cowles Commission's contributions to econometrics at Chicago, 1935–1955". *Journal of Economic Literature*. 32: 30–59.

Christensen, R. (2005). "Testing Fisher, Neyman, Pearson, and Bayes". *The American Statistician*. 59: 121–126. DOI: 10.1198/000313005X20871.

Cox, D. R. (1958). "Some problems connected with statistical inference". *Annals of Mathematical Statistics*. 29: 357–372. DOI: 10.1214/aoms/1177706618.

Cox, D. R. and D. R. Hinkley (1974). *Theoretical Statistics*. London: Chapmann and Hall.

Cremers, K. J. M. (2002). "Stock return predictability: A Bayesian model selection perspective". *Review of Financial Studies*. 15: 1223–1249. DOI: 10.1093/rfs/15.4.1223.

Del Negro, M., F. Schorfheide, F. Smets, and R. Wouters (2007). "On the fit of new Keynesian models (with discussion)". *Journal of Business & Economic Statistics*. 25: 123–162. DOI: 10.1198/073500107000000016.

Denton, F. T. (1988). "The significance of significance: Rhetorical aspects of statistical hypothesis testing in economics". In: *The Consequences of Economic Rhetoric*. Ed. by A. Klamer, D. N. McCloskey, and R. M. Solow. Cambridge University Press.

Edwards, W., H. Lindman, and L. J. Savage (1963). "Bayesian statistical inference for psychological research". *Psychological Review*. 70: 193–242. DOI: 10.1037/h0044139.

Eichenbaum, M. (1995). "Some comments on the role of econometrics in economic theory". *Economic Journal*. 105: 1609–1621. DOI: 10.2307/2235122.

Engsted, T. (2002). "Measures of fit for rational expectations models". *Journal of Economic Surveys*. 16: 301–355. DOI: 10.1111/1467-6419.00171.

Engsted, T. (2009). "Statistical vs. economic significance in economics and econometrics: Further comments on McCloskey and Ziliak". *Journal of Economic Methodology*. 16: 393–408. DOI: 10.1080/13501780903337339.

Faust, J. and C. H. Whiteman (1997). "General-to-specific procedures fitting a data-admissible, theory-inspired, congruent, parsimonious, encompassing, weakly-exogenous, identified, structural model to the DGP: A translation and critique". *Carnegie-Rochester Conference Series on Public Policy*. 47: 121–161.

Fernandez-Villaverde, J. and J. F. Rubio-Ramirez (2004). "Comparing dynamic equilibrium models to data: A Bayesian approach". *Journal of Econometrics*. 123: 153–187. DOI: 10.1016/j.jeconom.2003.10.031.

Fisher, R. A. (1922). "On the mathematical foundations of theoretical statistics". *Philosophical Transactions of the Royal Society of London A*. 222: 309–368. DOI: 10.1098/rsta.1922.0009.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Fisher, R. A. (1955). "Statistical methods and scientific induction". *Journal of the Royal Statistical Society, Series B*. 17: 69–78.

Fisher, R. A. (1959). *Statistical Methods and Scientific Inference*. 2nd ed. Edinburgh: Oliver & Boyd.

Gelman, A. and C. R. Shalizi (2013). "Philosophy and the practice of Bayesian statistics (with discussion)". *British Journal of Mathematical and Statistical Psychology*. 66: 8–80. DOI: 10.1111/j.2044-8317.2011.02037.x.

Gigerenzer, G. (2004). "Mindless statistics". *Journal of Socio-Economics*. 33: 587–606. DOI: 10.1016/j.socec.2004.09.033.

Gigerenzer, G. and J. N. Marewski (2015). "Surrogate science: The idol of a universal method for scientific inference". *Journal of Management*. 41: 421–440. DOI: 10.1177/0149206314547522.

Gorroochurn, P. (2016). *Classic Topics on the History of Modern Mathematical Statistics: From Laplace to More Recent Times*. John Wiley & Sons.

Greenland, S. (2005). "Multiple-bias modelling for analysis of observational data (with discussion)". *Journal of the Royal Statistical Society, Series A*. 168: 267–306.

Greenland, S. (2023). "Divergence versus decision P-values: A distinction worth making in theory and keeping in practice: Or, how divergence P-values measure evidence even when decision P-values do not". *Scandinavian Journal of Statistics*. 50: 54–88. DOI: 10.1111/sjos.12625.

Greenland, S. and C. Poole (2013). "Living with statistics in observational research". *Epidemiology*. 24: 73–78. DOI: 10.1097/EDE.0b013e3182785a49.

Haavelmo, T. (1944). "The probability approach in econometrics". *Econometrica (Supplement)*. 12: 1–118.

Hansen, L. P. and R. Jagannathan (1991). "Implications of security market data for models of dynamic economies". *Journal of Political Economy*. 99: 225–262.

Hansen, L. P. and R. Jagannathan (1997). "Assessing specification errors in stochastic discount factor models". *Journal of Finance*. 52: 557–590.

Hansen, L. P. and T. J. Sargent (1980). "Formulating and estimating dynamic linear rational expectations models". *Journal of Economic Dynamics and Control*. 2: 7–46.

Harvey, C. R. (2017). "Presidential address: The scientific outlook in financial economics". *Journal of Finance*. 72: 1399–1440.

Harvey, C. R., Y. Liu, and A. Saretto (2020). "An evaluation of alternative multiple testing methods for finance applications". *Review of Asset Pricing Studies*. 10: 199–248. DOI: 10.1093/rapstu/raaa003.

Heckmam, J. J. (1992). "Haavelmo and the birth of modern econometrics: A review of 'The History of Econometric Ideas' by Mary Morgan". *Journal of Economic Literature*. 30: 876–886.

Heckman, J. J. (2000). "Causal parameters and policy analysis in economics: A twentieth century retrospective". *Quarterly Journal of Economics*. 115: 45–97. DOI: 10.1162/003355300554674.

Hendry, D. F. (1980). "Econometrics—Alchemy or science?" *Economica*. 47: 387–406. DOI: 10.2307/2553385.

Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford University Press.

Hendry, D. F., E. E. Leamer, and D. J. Poirier (1990). "The ET dialogue: A conversation on econometric methodology". *Econometric Theory*. 6: 171–261. DOI: 10.1017/S0266466600005119.

Hill, B. M. (1985). "Some subjective Bayesian considerations in the selection of models (with discussion)". *Econometric Reviews*. 4: 191–246. DOI: 10.1080/07474938608800083.

Hirschauer, N., S. Grüner, and O. Mußhoff (2022). *Fundamentals of Statistical Inference: What is the Meaning of Random Error?* Springer.

Hoover, K. D. (2014). "On the reception of Haavelmo's econometric thought". *Journal of the History of Economic Thought*. 36: 45–65. DOI: 10.1017/S1053837214000029.

Hoover, K., S. Johansen, and K. Juselius (2008). "Allowing the data to speak freely: The macroeconometrics of the cointegrated vector autoregression". *American Economic Review, Papers & Proceedings*. 98: 251–255. DOI: 10.1257/aer.98.2.251.

Hoover, K. and K. Juselius (2015). "Trygve Haavelmo's experimental methodology and scenario analysis in a cointegrated vector autoregression". *Econometric Theory*. 31: 249–274. DOI: 10.1017/S0266466614000292.

Hubbard, R. and M. J. Bayarri (2003). "Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing (with discussion)". *The American Statistician*. 57: 171–182. DOI: 10.1198/0003130031856.

Huntington-Klein, N. *et al.* (2021). "The influence of hidden researcher decisions in applied microeconomics". *Economic Inquiry*. 59: 944–960. DOI: 10.1111/ecin.12992.

Ioannidis, J. P. A. (2005). "Why most published research findings are false". *PLoS Medicine*. 2: 696–701.

Ioannidis, J. P. A., T. D. Stanley, and H. Doucouliagos (2017). "The power of bias in economics research". *Economic Journal*. 127: F236–F265. DOI: 10.1111/ecoj.12461.

Jeffreys, H. (1961). *Theory of Probability*. 3rd ed. London: Oxford University Press.

Kahneman, D. and A. Tversky (1972). "Subjective probability: A judgment of representativeness". *Cognitive Psychology*. 3: 430–454. DOI: 10.1016/0010-0285(72)90016-3.

Kass, R. E. and A. E. Raftery (1995). "Bayes factors". *Journal of the American Statistical Association*. 90: 773–795. DOI: 10.1080/01621459.1995.10476572.

Kass, R. E. and L. Wasserman (1995). "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion". *Journal of the American Statistical Association*. 90: 928–934.

Keuzenkamp, H. A. (1995). "The econometrics of the Holy Grail—A review of econometrics: Alchemy or science? Essays in Econometric Methodology". *Journal of Economic Surveys*. 9: 233–248.

Keuzenkamp, H. A. (2000). *Probability, Econometrics and Truth: The Methodology of Econometrics*. Cambridge University Press.

Keuzenkamp, H. A. and J. R. Magnus (1995). "On tests and significance in econometrics". *Journal of Econometrics*. 67: 5–24.

Kim, J. H. (2019). "Tackling false positives in business research: A statistical toolbox with applications". *Journal of Economic Surveys*. 33: 862–895.

Koopmans, T. C. (1950). *Statistical Inference in Dynamic Economic Models (Cowles Commission Monograph)*. New York: John Wiley & Sons.

Kruschke, J. K. and T. M. Liddell (2018). "The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective". *Psychonomic Bulletin & Review*. 25: 178–206.

Kydland, F. E. and E. C. Prescott (1982). "Time to build and aggregate fluctuations". *Econometrica*. 50: 1345–1370. DOI: 10.2307/1913386.

Larsen, R. J. and M. L. Marx (2012). *An Introduction to Mathematical Statistics and Its Applications*. 5th ed. Pearson.

Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Non Experimental Data*. John Wiley & Sons.

Leamer, E. E. (1983). "Let's take the con out of econometrics". *American Economic Review*. 73: 31–43.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. New York: John Wiley & Sons.

Lehmann, E. L. (1993). "The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two?" *Journal of the American Statistical Association*. 88: 1242–1249. DOI: 10.1080/01621459.1993.10476404.

Lenhard, J. (2006). "Models and statistical inference: The controversy between Fisher and Neyman-Pearson". *The British Journal for the Philosophy of Science*. 57: 69–91. DOI: 10.1093/bjps/axi152.

Lindley, D. V. (1957). "A statistical paradox". *Biometrica*. 44: 187–192. DOI: 10.1093/biomet/44.1-2.187.

Mayo, D. G. (2014). "On the Birnbaum argument for the strong likelihood principle (with discussion)". *Statistical Science*. 29: 227–266.

McCloskey, D. N. (1983). "The rhetoric of economics". *Journal of Economic Literature*. 21: 481–517.

McShane, B. B., D. Gal, A. Gelman, C. Robert, and J. L. Tackett (2019). "Abandon statistical significance". *The American Statistician*. 73: 235–245. DOI: 10.1080/00031305.2018.1527253.

Neyman, J. and E. S. Pearson (1933). "On the problem of the most efficient tests of statistical hypotheses". *Philosophical Transactions of the Royal Society of London, Series A*. 231: 289–337. DOI: 10.1098/rsta.1933.0009.

Nuzzo, R. (2014). "Statistical errors". *Nature*. 506: 150–152. DOI: 10.1038/506150a.

Pena, V. and J. O. Berger (2017). "A note on recent criticism to Birnbaum's theorem". *Working Paper*. URL: https://arxiv.org/abs/1711.08093.

Poirier, D. J. (1988). "Frequentist and subjectivist perspectives on the problems of model building in economics". *Journal of Economic Perspectives*. 2: 121–144. DOI: 10.1257/jep.2.1.121.

Raftery, A. E. (1995). "Bayesian model selection in social research (with discussion)". *Sociological Methodology*. 25: 111–163. DOI: 10.2307/271063.

Raftery, A. E. (1999). "Bayes factors and BIC: Comment on 'A critique of the Bayesian information criterion for model selection". *Sociological Methods & Research*. 27: 411–427. DOI: 10.1177/0049124199027003005.

Schneider, J. W. (2013). "Caveats for using statistical significance tests in research assessments". *Journal of Informetrics*. 7: 50–62. DOI: 10.1016/j.joi.2012.08.005.

Schneider, J. W. (2015). "Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations". *Scientometrics*. 102: 411–432. DOI: 10.1007/s11192-014-1251-5.

Schneider, J. W. (2016). "The imaginarium of statistical inference when data are the population: Comments to Williams and Bornmann". *Journal of Informetrics*. 10: 1243–1248. DOI: 10.1016/j.joi.2016.09.011.

Schneider, J. W. (2018). "NHST is still logically flawed". *Scientometrics*. 115: 627–635. DOI: 10.1007/s11192-018-2655-4.

Schorfheide, F. (2000). "Loss function based evaluation of DSGE models". *Journal of Applied Econometrics*. 15: 645–670. DOI: 10.1002/jae.582.

Schwarz, G. E. (1978). "Estimating the dimension of a model". *Annals of Statistics*. 6: 461–464. DOI: 10.1214/aos/1176344136.

Sellke, T., M. J. Bayarri, and J. O. Berger (2001). "Calibration of p values for testing precise null hypotheses". *The American Statistician*. 55: 62–71. DOI: 10.1198/000313001300339950.

Shiller, R. J. (2005). *Irrational Exuberance*. 2th ed. Princeton University Press.

Smets, F. and R. Wouters (2003). "An estimated stochastic general equilibrium model of the euro area". *Journal of the European Economic Association*. 1: 1123–1175. DOI: 10.1162/154247603770383415.

Smets, F. and R. Wouters (2007). "Shocks and frictions in US business cycles: A Bayesian DSGE approach". *American Economic Review*. 97: 586–606. DOI: 10.1257/aer.97.3.586.

Startz, R. (2014). "Choosing the more likely hypothesis". *Foundations and Trends in Econometrics*. 7: 119–189. DOI: 10.1561/0800000028.

Steel, M. F. J. (2020). "Model averaging and its use in economics". *Journal of Economic Literature*. 58: 644–719. DOI: 10.1257/jel.20191385.

Storey, J. D. (2003). "The positive false discovery rate: A Bayesian interpretation and the q-value". *The Annals of Statistics*. 31: 2013–2035. DOI: 10.1214/aos/1074290335.

Verbeek, M. (2017). *A Guide to Modern Econometrics*. 5th ed. Wiley.

Wagenmakers, E.-J. (2007). "A practical solution to the pervasive problems of p values". *Psychonomic Bulletin & Review*. 14: 779–804.

Wagenmakers, E.-J. (2022). "Approximate objective Bayes factors from p-values and sample size: The $3p\sqrt{n}$ rule". *Working Paper*. University of Amsterdam.

Wagenmakers, E.-J. and A. Ly (2023). "History and nature of the Jeffreys-Lindley paradox". *Archive for History of Exact Sciences*. 77: 25–72.

Wasserstein, R. L. and N. A. Lazar (2016). "The ASA's statement on p-values: Context, process, and purpose". *The American Statistician*. 70: 129–133.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar (2019). "Editorial: Moving to a world beyond '$p < 0.05$". *The American Statistician*. 73: 1–19.

Watson, M. (1993). "Measures of fit for calibrated models". *Journal of Political Economy*. 101: 1011–1044.

Weakliem, D. L. (1999). "A critique of the Bayesian information criterion for model selection". *Sociological Methods & Research*. 27: 359–397.

Western, B. and S. Jackman (1994). "Bayesian inference for comparative research". *American Political Science Review*. 88: 412–423.

Ziliak, S. T. and D. N. McCloskey (2008). *The Cult of Statistical Significance*. The University of Michigan Press.