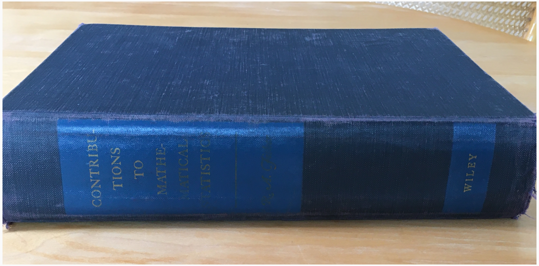
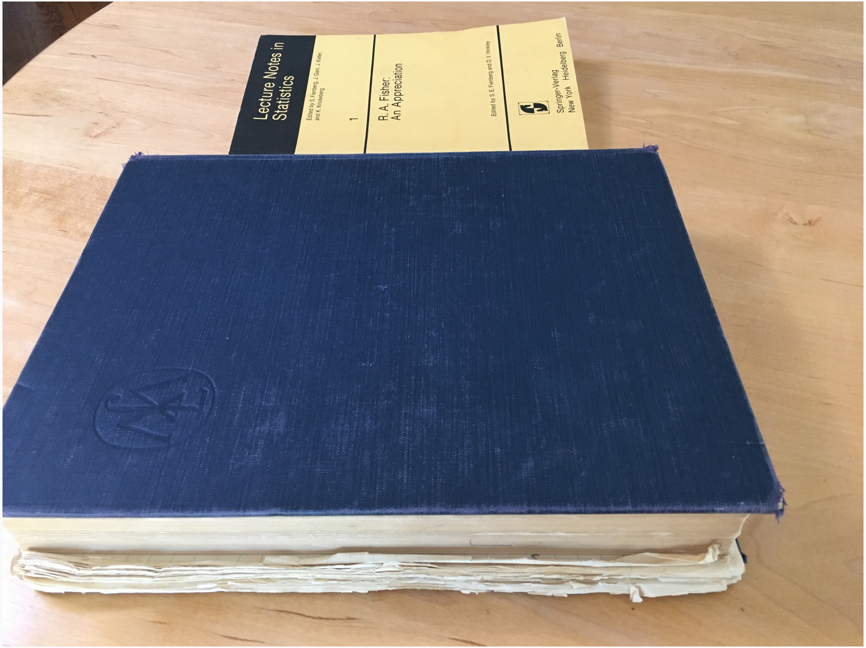
Fisher's contributions to mathematical statistics

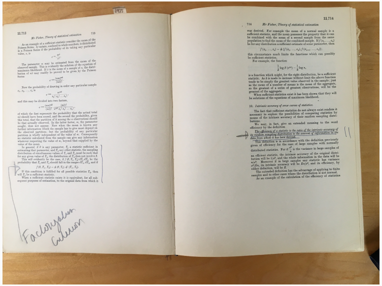
Nancy Reid
University of Toronto

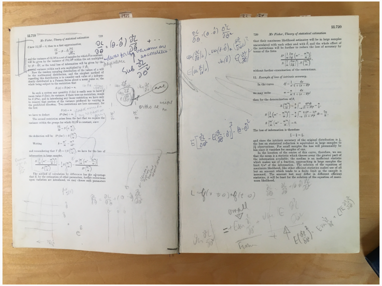
April 21 2022

Gonville & Caius
UNIVERSITY OF CAMBRIDGE




Lecture Notes in Statistics
Edited by S. Fienberg, J. Gani, J. Kiefer, and K. Krickeberg
1
R. A. Fisher: An Appreciation
Edited by S. E. Fienberg and D. V. Hinkley
Springer-Verlag New York Heidelberg Berlin



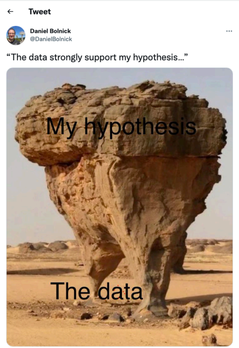


~~Mathematical~~ Theoretical statistics

- Statistics needs a healthy interplay between theory and applications
  - theory meaning foundations, rather than theoretical analysis of specific techniques

- Foundations?
- "A solid base, on which rests a large structure" OED
  - must be continually tested against new applications

  - "the practical application of general theorems is a different art from their establishment by mathematical proof"

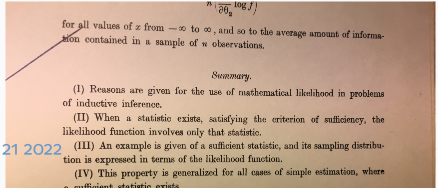


Fisher 1922 1925 1934

- 1922: On the mathematical foundations of theoretical statistics: statistic/parameter, estimation, consistent, sufficient, efficient, likelihood, maximum likelihood estimate, information, intrinsic accuracy

- 1925: Theory of statistical estimation: all of the above, scoring algorithm, loss of information, ancillary

- 1934: Two new properties of mathematical likelihood: conditional inference, location model, ancillary configuration, recovery of information, exponential family, distribution of sufficient estimate, uniformly most powerful tests



for all values of x from $-\infty$ to ∞, and so to the average amount of information contained in a sample of n observations.

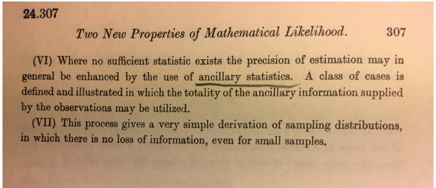
Summary.

(I) Reasons are given for the use of mathematical likelihood in problems of inductive inference.

(II) When a statistic exists, satisfying the criterion of sufficiency, the likelihood function involves only that statistic.

(III) An example is given of a sufficient statistic, and its sampling distribution is expressed in terms of the likelihood function.

(IV) This property is generalized for all cases of simple estimation, where a sufficient statistic exists.



24.307

Two New Properties of Mathematical Likelihood. 307

(VI) Where no sufficient statistic exists the precision of estimation may in general be enhanced by the use of ancillary statistics. A class of cases is defined and illustrated in which the totality of the ancillary information supplied by the observations may be utilized.

(VII) This process gives a very simple derivation of sampling distributions, in which there is no loss of information, even for small samples.

Likelihood Inference

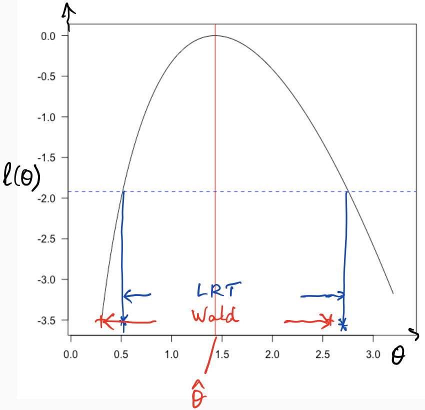
- Model: $Y \sim f(y;\theta), \theta \in \mathbb{R}^p, y \in \mathbb{R}^n$
- Likelihood function: $L(\theta; y) \propto f(y;\theta)$
- Maximum likelihood estimator $\hat\theta = \hat\theta(y) = \arg\sup_\theta L(\theta; y) = \arg\sup_\theta \log\{L(\theta; y)\}$
- Observed and expected Fisher information $\ell(\theta; y) = \log L(\theta; y)$
  $j(\theta) = -\partial^2\ell(\theta; y)/\partial\theta^2, \quad i(\theta) = \mathrm{E}\{-\partial^2\ell(\theta; Y)/\partial\theta^2\}$
- in "large samples" $\hat\theta \mathrel{\dot\sim} N_p\{\theta, j^{-1}(\hat\theta)\}$ equivalently $N_p\{\theta, i^{-1}(\theta)\}$
- in "large samples" $\ell'(\theta) \mathrel{\dot\sim} N_p\{0, i^{-1}(\theta)\}$

...Likelihood inference

- in "large samples" $\hat{\theta} \dot\sim N_p\{\theta, j^{-1}(\hat{\theta})\}$
- in "large samples" $\ell'(\theta) \dot\sim N_p\{0, i^{-1}(\theta)\}$
- in "large samples" $2\{\ell(\hat{\theta}) - \ell(\theta)\} \dot\sim \chi^2_p$

- $y = (y_1, \ldots, y_n); \quad y_i \sim Gamma(\theta, 1)$
- $L(\theta) = \prod_{i=1}^n y_i^{\theta-1} e^{-y_i} / \Gamma(\theta)$

- $\psi(\hat{\theta}) = \frac{1}{n} \sum \log(y_i)$ $\qquad \psi = \log \Gamma'$
- $j(\hat{\theta}) = n\psi'(\hat{\theta})$



Small samples

- can we find the exact distribution of the maximum likelihood estimator?
- special case 1: location model $Y_i \sim f(y_i - \theta), i = 1, \ldots, n; \theta \in \mathbb{R}$ Fisher 1934
- ancillary statistic $a = (y_1 - \hat{\theta}, \ldots, y_n - \hat{\theta})$ $\sum(\partial/\partial\theta)\log\{f(y_i;\hat{\theta})\} = 0$
- special case 2: exponential family model $Y_i \sim \exp\left\{s(y)^T\theta - nc(\theta)\right\}h(y)$ Fisher 1925
- sufficient statistic $s = s(y)$ is ‘matched’ to θ same dimension
- maximum likelihood estimate is sufficient likelihood map is sufficient

- can we find the exact distribution of the maximum likelihood estimator?

- special case 1: $y \to (\hat{\theta}, a)$ Fisher 1934

$$f(\hat{\theta} \mid a; \theta) = \frac{L(\theta; \hat{\theta}, a)}{\int L(\theta; \hat{\theta}, a) d\theta} = \frac{\exp\{\ell(\theta; \hat{\theta}, a)\}}{\int \exp\{\ell(\theta; \hat{\theta}, a)\} d\theta}$$

- special case 2: $y \to s$

$$f(s; \theta) = \exp\{s^T\theta - nc(\theta)\}\tilde{h}(s) \qquad s = nc'(\hat{\theta}), \quad \tilde{h}(s) = \int ...dy$$

- general case $1 + O(n^{-3/2})$

$$f(\hat{\theta}; \theta \mid a) \doteq c|j(\hat{\theta})|^{1/2} \exp\{\ell(\theta; \hat{\theta}, a) - \ell(\hat{\theta}; \hat{\theta}, a)\}$$

… Small samples

- we know the distribution of the maximum likelihood estimator

  to a high order of approximation

- how do we use this for inference?

- find values of θ that are consistent with our observed data — confidence intervals

- find the probability for a given θ_0 of observing a result
  "as or more extreme than our observed data" (F 1925) — p-values

- computationally feasible? — $f(\hat\theta;\theta \mid a) \doteq c|j(\hat\theta)|^{1/2}\exp\{\ell(\theta;\hat\theta,a)-\ell(\hat\theta;\hat\theta,a)\}$

- models with many parameters: $\theta=(\psi,\lambda)$, $\quad \ell_{\mathrm{p}}(\psi)=\ell(\psi,\hat\lambda_\psi)$

  profile log-likelihood

Models with many parameters

- statistical models with many parameters $\theta = (\psi, \lambda)$

  parameter of interest
  nuisance parameter(s)

- profile or concentrated log-likelihood function $\ell_{\text{prof}}(\psi) \equiv \ell(\psi, \hat{\lambda}_\psi; y)$

  $\hat{\lambda}_\psi = \arg\sup_\lambda L(\psi, \lambda)$

- now use "large samples" theory on $\ell_{prof}(\psi)$

  approximation can be very poor

- can sometimes isolate parameters in a marginal or conditional distribution

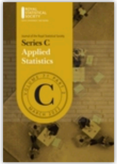
  e.g. $$f(y; \psi, \lambda) \propto f_c(s \mid t; \psi) f(t; \psi, \lambda)$$

  Fisher's exact test

- can approximate this conditional likelihood with relatively simple adjustments

  B-N 1983; Cox R 1987

$$\ell_{\text{mod}}(\psi) = \underbrace{\ell_{\text{prof}}(\psi)}_{O(n)} - \underbrace{\frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|}_{O(1)} + \underbrace{M(\psi)}_{O(1/\sqrt{n})}$$



Volume 71, Issue 2

Pages: **269-492**
March 2022

‹ Previous Issue

Articles

Most Recent | Most Cited

Open Access
Assessing seismic origin of geological features by fitting equidistant parallel lines
P.E. Jupp, I.B.J. Goudie, R.A. Batchelor, R.J.B. Goudie
First Published: 17 April 2022
Abstract | Full text | PDF | References | Request permissions

Open Access
Modelling the extremes of seasonal viruses and hospital congestion: The example of flu in a Swiss hospital
Setareh Ranjbar, Eva Cantoni, Valérie Chavez-Demoulin, Giampiero Marra, Rosalba Radice, Katia Jaton
First Published: 13 April 2022
Abstract | Full text | PDF | References | Request permissions

Open Access
The modelling of movement of multiple animals that share behavioural features
Gianluca Mastrantonio

"you are not thinking in the right way"

… the 21st century

- "investigate the use of the marginal likelihood function for model specification"
  Fong & Holmes 2020
- "maximum likelihood estimates are obtained from a multivariate Poisson regression model" Muñoz-Pichardo et al. 2021
- "a penalized likelihood approach to integrate high-dimension subject-level information along low-dimensional aggregate information" Sheng, Huang & Kim 2021

- "explores retrospective and prospective likelihood in terms of power of the score tests" Liu et al. 2020
- "likelihood ratio test for sequential change-point detection" Dette & Gössman 2020
- "proposed a modified profile likelihood method" for genetic association studies
  Zhang et al. 2020
- "a variant of the maximum likelihood estimator using a subset of the data …resulting estimator is still consistent" Ekvall & Jones 2021
- "small-sample bias correction for [the variance of] the maximum likelihood estimator"
  Ozenne et al. 2020
- …

Data is complex

- spatial dependence, nested sampling designs, high-dimensional parametrization
- pseudo-likelihood builds distribution from local characteristics

$$L_{\text{pseudo}}(\theta) = \prod_{j \in \mathcal{N}(y_i)} f(y_i \mid y_j; \theta)$$

- example spatial modelling

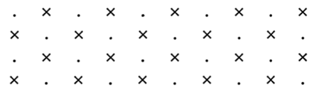
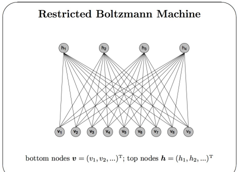


Fig. 1. Coding pattern for a first-order scheme.

Besag 74

- example Boltzmann machine



Zhu

... Data is complex

- spatial dependence, nested sampling designs, high-dimensional parametrization

- composite likelihood combines lower-dimensional marginal densities
-
$$L_{\text{composite}}(\theta) = \prod_{i=1}^{n} \prod_{j<k} f_2(y_{ij}, y_{ik}; \theta)$$

- example longitudinal data – each subject measured at several time points

Renard 04; Kuk& Nott 03

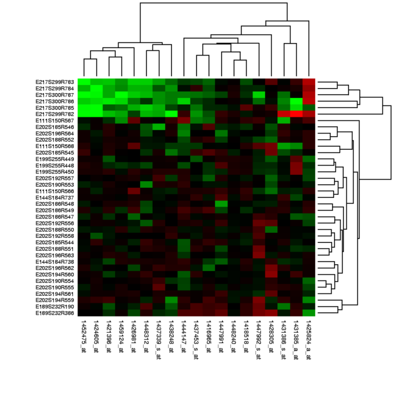
  - latent variable: $w_{ir} = x_{ir}'\beta + z_{ir}'b_i + \epsilon_{ir}, \quad \epsilon_{ir} \sim N(0,1)$
  - binary observations: $y_{ir} = \mathbb{1}(w_{ir} > 0); \quad r = 1, \ldots m; i = 1, \ldots n$ — dependent binary vector y_i
  - probit model: $Pr(y_{ir} = 1 \mid b_i) = \Phi(x_{ir}'\beta + b_i); \quad b_i \sim N_q(0, \Sigma_b)$ — regression

$$L_{composite}(\beta, \Sigma_b; y) = \prod_{i=1}^{n} \prod_{j<k} P_{11,i}^{y_{ir}y_{is}} P_{10,i}^{y_{ir}(1-y_{is})} P_{01,i}^{(1-y_{ir})y_{is}} P_{00,i}^{(1-y_{ir})(1-y_{is})}$$

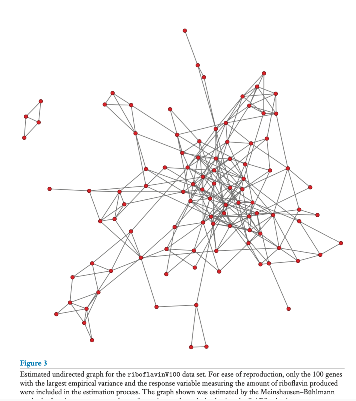
$P_{11,i}$, $P_{10,i}$, etc. evaluated using $\Phi_2(\cdot, \cdot; \rho_{irs})$

… Data is complex

- spatial dependence, nested sampling designs, high-dimensional parametrization



Andrade, Wikipedia



Buhlmann et al 14

High-dimensional parameters

- new limit results, e.g. $\hat{\theta} \xrightarrow{d} N(\theta + \text{bias}, \sigma^2 \times \text{adjustment})$ Sur & Candès, Fan et al.
- higher order approximations allows $p = O(n^\alpha), \alpha < 1/2$ Tang 22
- sparsity – $\mathcal{S} \equiv \{j; \theta_j \neq 0\}; |\mathcal{S}| = s < n$
  - enforce sparsity, e.g. Lasso
  - discover sparsity, e.g. Battey 22
  - isolate parameter(s) of interest Battey & R 22; McCormack et al 19

Various types of ‘likelihood’

- likelihood, marginal likelihood, conditional likelihood, profile likelihood
  adjusted profile likelihood
- pseudo-likelihood, composite likelihood
- semi-parametric likelihood, partial likelihood
- empirical likelihood, penalized likelihood
- bootstrap likelihood, *h*-likelihood, weighted likelihood, quasi-likelihood, local likelihood, sieve likelihood, simulated likelihood

Those pesky *p*-values

- ANDROMEDA Trial: randomized clinical trial to compare two treatments for septic shock

  Hernandez et al 2019

- estimated hazard ratio 0.75 [0.55, 1.02] — after adjusting for confounders
- 2-sided p-value 0.06 — 34.9% vs 43.4% unadjusted

- Discussion: " a peripheral perfusion-targeted resuscitation strategy did not result in a significantly lower 28-day mortality when compared with a lactate level-targeted strategy"

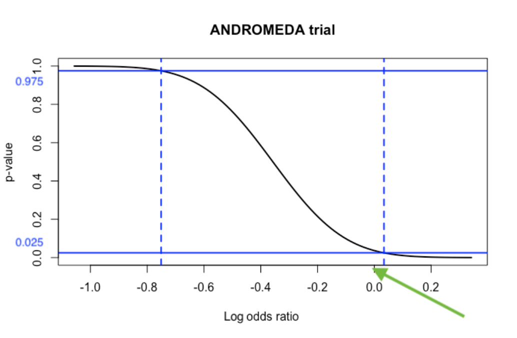
- Abstract: "Among patients with septic shock, a resuscitation strategy targeting normalization of capillary refill time, compared with a strategy targeting serum lactate levels, did not reduce all-cause 28-day mortality."

ANDROMEDA trial

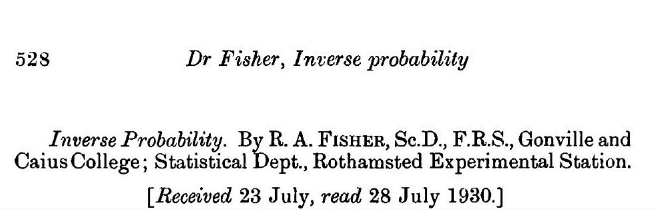
	Died	Lived	
New	74	138	212
Old	92	120	212
Total	166	258	424

2-sided p-value = 0.07

likelihood ratio test
no adjustment for covariates



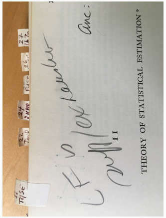
90% confidence interval: $[-0.688, -0.030]$
95% confidence interval: $[-0.751, 0.034]$
99% confidence interval: $[-0.825, 0.107]$



528 *Dr Fisher, Inverse probability*

Inverse Probability. By R. A. Fisher, Sc.D., F.R.S., Gonville and Caius College; Statistical Dept., Rothamsted Experimental Station.

[*Received* 23 July, *read* 28 July 1930.]



$$df = -\frac{\partial}{\partial\theta}F(Y, \theta)d\theta$$

fiducial probability density for θ, given statistic Y

"It is not to be lightly supposed that men of the mental calibre of Laplace and Gauss ... could fall into error on a question of prime theoretical importance, without an uncommonly good reason"

The First Workshop on BFF Inference and Statistical Foundations (BFF 2014)

November 10 – November 14, 2014

BFF 1 – 6

Distributions for parameters

- posterior distribution Bayes 1763
- fiducial probability Fisher 1930
- confidence distribution Cox 1958
- structural probability Fraser 1964
- belief functions Dempster 1967

- objective Bayes
- generalized fiducial inference, fiducial prediction, functional models, “slice-and-dice”, ...
- confidence distributions / curves
- approximate significance functions
- inferential models

Fiducial inference

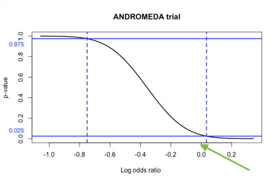
- study the structure of models which give 'valid' fiducial inference — Dawid 22
- change the modelling framework so fiducial arguments can be developed more cohesively — Lang 22
- generalized fiducial density — Hannig 09 ff

$$r(\theta; y^{o}) \propto \underbrace{f(y^{o}; \theta)}_{likelihood} \; \underbrace{J(y^{o}, \theta)}_{\text{“prior”}}$$

- these methods all founder in models with many parameters
- unless each parameter of interest can be "measured separately" — Fisher

Why do we want distributions on parameters?

- inference is intuitive
- combines easily with decision theory
- de-emphasizes point estimation and arbitrary cut-offs



- "it's tempting to conclude that μ is more likely to be near the middle of this interval, and if outside, not very far outside" Cox 2006
- "assigns probability 0.05 to θ lying between the upper endpoints of the 0.90 and 0.95 confidence intervals, etc." Efron 1993
- all inference statements become probability statements about unknowns hmm...

Nature of Probability

- probability to describe physical haphazard variability — aleatory/empirical
  - probabilities represent features of the “real” world in somewhat idealized form
  - subject to empirical test and improvement
  - conclusions of statistical analysis expressed in terms of interpretable parameters
  - enhanced understanding of the data generating process

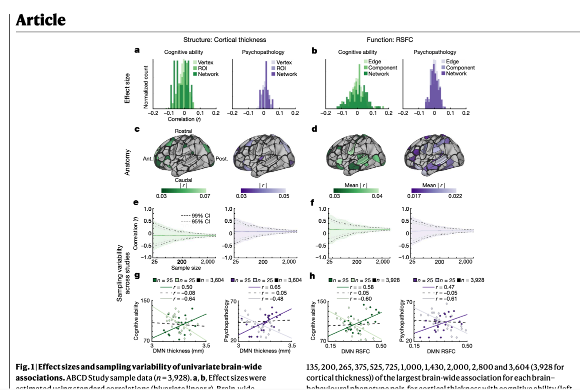
- probability to describe the uncertainty of knowledge — epistemic
  - measures rational, supposedly impersonal, degree of belief, given relevant information — Jeffreys
  - measures a particular person’s degree of belief, subject typically to some constraints of self-consistency — Ramsey, de Finetti, Savage

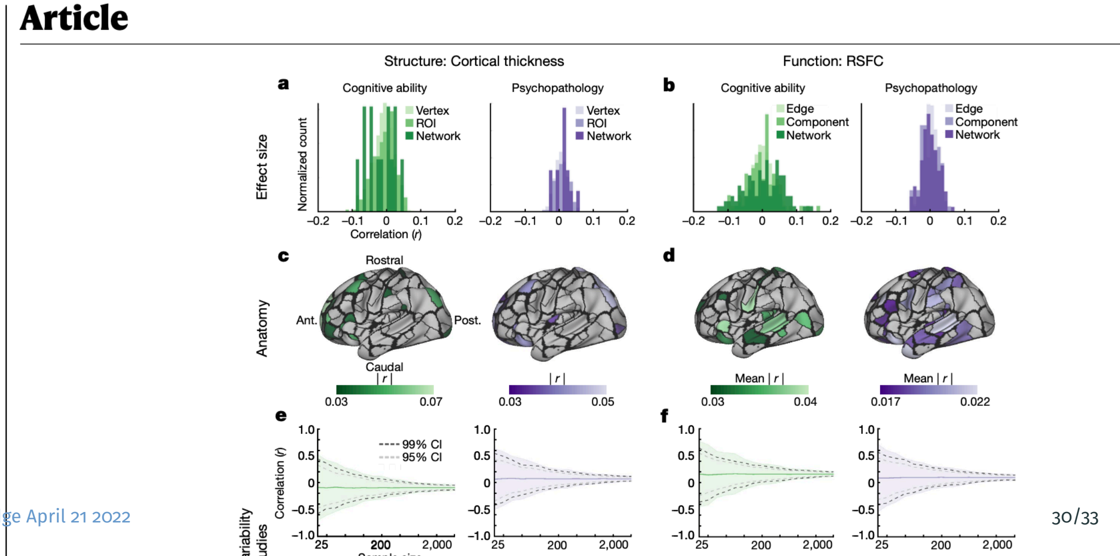
Role of Foundations

- avoid apparent discoveries based on spurious patterns
- shed light on the structure of the problem
- obtain calibrated inferences about interpretable parameters
- provide a realistic assessment of precision
- understand when/why methods work/fail

Needs in applications

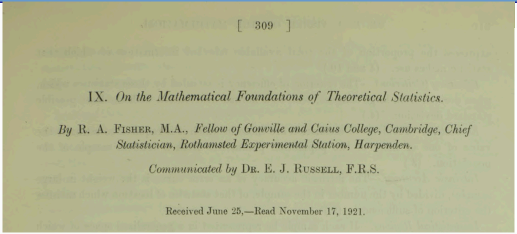
- something that works
- gives ‘sensible’ answers
- not too sensitive to model assumptions
- computable in reasonable time
- provides interpretable parameters




Article
Structure: Cortical thickness
Function: RSFC
a
b
Cognitive ability
Psychopathology
Effect size
Normalized count
Vertex
ROI
Network
Edge
Component
Correlation (r)
−0.2
−0.1
0
0.1
0.2
c
d
Anatomy
Rostral
Caudal
Ant.
Post.
|r|
Mean |r|
0.03
0.07
0.05
0.04
0.017
0.022
e
f
99% CI
95% CI
Correlation (r)
1.0
0.5
−0.5
−1.0
25
200
2,000

Reproducible brain-wide association studies require thousands of individuals

- "relate population variability in brain features" eg functional connectivity
- "and behavioural phenotypes" eg cognitive ability
- "across all univariate brain-wide associations, the largest correlation that replicated out-of-sample was $|r| = 0.16$"
- "at $n = 25$, the 99% confidence interval for univariate associations was $r \pm 0.52$."
- "Bias in favour of significant, larger BWAS effects has limited the publication of null results, perpetuating inflated effect sizes … "



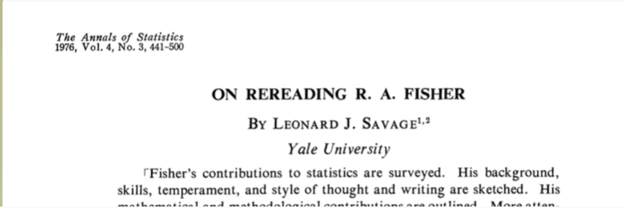
[309]

IX. *On the Mathematical Foundations of Theoretical Statistics.*

By R. A. Fisher, M.A., *Fellow of Gonville and Caius College, Cambridge, Chief Statistician, Rothamsted Experimental Station, Harpenden.*

Communicated by Dr. E. J. Russell, F.R.S.

Received June 25,—Read November 17, 1921.



The Annals of Statistics
1976, Vol. 4, No. 3, 441–500

ON REREADING R. A. FISHER

By Leonard J. Savage[1,2]

Yale University

Fisher's contributions to statistics are surveyed. His background, skills, temperament, and style of thought and writing are sketched. His

Savage: "there is a world of R.A. Fisher at once very near to and very far from the world of modern statisticians … research for the fun of it is abundant and beautiful in Fisher's writings"

Fraser: "One important characteristic of Fisher was his ability to move into new areas of statistics, suggesting concepts and methods … left the theory open to modification and development"

Efron: "This paper makes me happy to be a statistician"

THANK YOU

Gonville & Caius
UNIVERSITY OF CAMBRIDGE