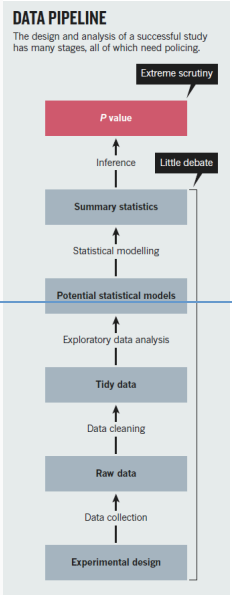
Statistical Theory for Data Science

STA2212H S LEC9101

Week 10

March 17 2026



COMMENT

P values are just the tip of the iceberg

Leek & Peng, Nature, 2015

Ridding science of shoddy statistics will require scrutiny of every step, not merely the last one, say **Jeffrey T. Leek** and **Roger D. Peng**.

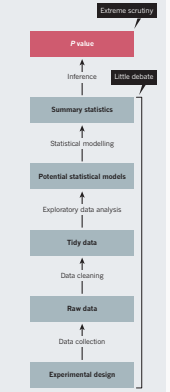
There is no statistic more maligned than the *P* value. Hundreds of papers and blogposts have been written about what some statisticians deride as 'null hypothesis significance testing' (NHST; see, for example, go.nature.com/pfvgqe). NHST deems whether the results of a data analysis are important on the basis of whether a summary statistic (such as a *P* value) has crossed a threshold. Given the discourse, it is no surprise that some hailed as a victory the banning of NHST methods (and all of statistical inference) in the journal *Basic and Applied Social Psychology* in February[1].

Such a ban will in fact have scant effect on the quality of published science. There are many stages to the design and analysis of a successful study (see 'Data pipeline'). The last of these steps is the calculation of an inferential statistic such as a *P* value, and the application of a 'decision rule' to it (for example, $P < 0.05$). In practice, decisions that are made earlier in data analysis have a much greater impact on results — from experimental design to batch effects, lack of adjustment for confounding factors, or simple measurement error. Arbitrary levels of statistical significance can be achieved by changing the ways in which data are cleaned, summarized or modelled[2].

P values are an easy target: being widely used, they are widely abused. But, in practice, deregulating statistical significance opens the door to even more ways to game statistics — intentionally or unintentionally — to get a result. Replacing *P* values with Bayes factors or another statistic is ultimately about choosing a different trade-off of true positives and false positives. Arguing about the *P* value is like focusing on a single misspelling, rather than on the faulty logic of a

DATA PIPELINE

The design and analysis of a successful study has many stages, all of which need policing.



analysis is taught through an apprenticeship model, and different disciplines develop their own analysis subcultures. Decisions are based on cultural conventions in specific communities rather than on empirical evidence. For example, economists call data measured over time 'panel data', to which they frequently apply mixed-effects models. Biomedical scientists refer to the same type of data structure as 'longitudinal data', and often go at it with generalized estimating equations.

Statistical research largely focuses on mathematical statistics, to the exclusion of the behaviour and processes involved in data analysis. To solve this deeper problem, we must study how people perform data analysis in the real world. What sets them up for success, and what for failure? Controlled experiments have been done in visualization[3] and risk interpretation[4] to evaluate how humans perceive and interact with data and statistics. More recently, we and others have been studying the entire analysis pipeline. We found, for example, that recently trained data analysts do not know how to infer *P* values from plots of data[5], but they can learn to do so with practice.

The ultimate goal is evidence-based data analysis[6]. This is analogous to evidence-based medicine, in which physicians are encouraged to use only treatments for which efficacy has been proved in controlled trials. Statisticians and the people they teach and collaborate with need to stop arguing about *P* values, and prevent the rest of the iceberg from sinking science. ■

Jeffrey T. Leek *and* **Roger D. Peng** *are associate professors of biostatistics at the Johns Hopkins Bloomberg School of Public*

This just in

March 17, 2026

SIGMA XI
THE SCIENTIFIC RESEARCH SOCIETY

SmartBrief

Your World of Science News SIGN UP · SHARE

Top Story

CERN finds new particle with two charm quarks

(Ronald Patrick/Getty Images)

Researchers at CERN's Large Hadron Collider have discovered a new particle, Xicc+, containing two charm quarks and a down quark. This heavier, proton-like particle had a predicted lifetime much shorter than similar particles, making it difficult to detect. The discovery, which has a statistical significance of over seven sigma, resolves a 20-year-old mystery regarding the mass of such particles and highlights the capabilities of recent upgrades to the LHC.

Full Story: New Scientist (3/17)

SHARE

This just in

NewScientist

Sign in | Enter search keywords

News Features Newsletters Podcasts Video Comment Culture Games | This week's magazine

Health Space Physics Technology Environment Mind Humans Life Mathematics Chemistry Earth Society

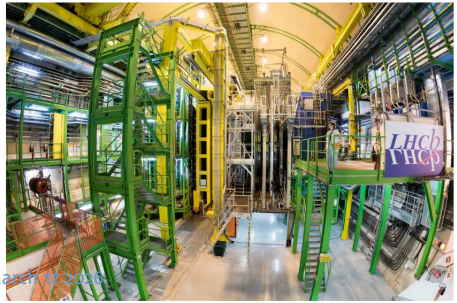
Physics

Particle discovered at CERN solves a 20-year-old mystery

Physicists working on the LHCb experiment have spotted an elusive and fleeting particle, a heavier and more charming cousin to the proton, that has been sought for decades

By Alex Wilkins

17 March 2026



The LHCb experiment cavern at CERN

2007

Today

1. March 24: Guest Lecture, Professor Radu Craiu — models for dependent data
2. Recap: graphical models
3. Nonparametric methods
4. Project papers: introductions

This week

- Toronto Data Workshop, Wednesday 18 March, 12.00 (EDT) on Zoom
  Zoro Wang, Carnegie-Mellon University : “AI agents at work – but whose work?”
- Statistics Colloquium Thursday 19 March, 11am Hydro 9014
  Roger Peng, UT Austin: “Building data analysis proofs”

Recap: graphical models

- graphs are useful for visualizing multivariate distributions — mv normal; multinomial
- directed acyclic graphs encode independence among components
  all edges have arrows; no cycles
- building a joint distribution from conditional pieces relies on a Markov property
  $W \perp \widetilde{W}$ | parents, descendants
- DAGs can be linked to causality with special notation for interventions — AoS 17.8
- graphical models are widely used with complex multivariate data
  e.g. neural networks, spatial processes

Graphical models: example

Drton & Maathuis, 2014

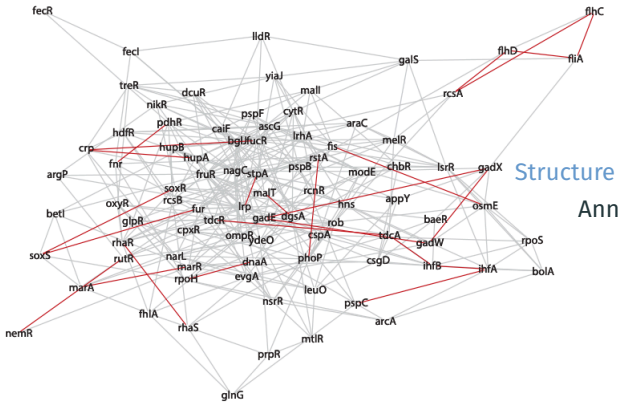


Figure 2

Estimated conditional independence graph in a Gaussian copula model for data on the expression of 87 transcription factors in *Escherichia coli*. The 24 red edges correspond to pairs of transcription factors that are known to interact.

Structure learning in graphical modelling

Ann. Rev. Statist. Applic. 4, 365–393.

Recap: unmeasured confounding

- in observational studies, apparent treatment effects might be “explained away” by adjusting for other properties of the units
- if they are not, then under an assumption of no unmeasured confounding, evidence of a causal treatment effect is strengthened
- but this assumption cannot be tested by the data in the observational study
- Cornfield’s lemma:
  - if a third confounding variable explains an observed association between exposure and outcome variables,
  - the association between the exposure and the confounder, and the confounder and the outcome,
  - must be at least as strong as the association between the exposure and the outcome

as measured by the risk ratio

Cornfield et al. (1959) reprinted in *Int. J. Epidemiology* **38** (2009);
see also VanderWeele & Ding (2017) *Ann. Internal Medicine*

- plug-in estimation: if $\theta = T(F)$, then $\hat{\theta} = T(\widehat{F}_n)$ $\widehat{F}_n(t) = \frac{1}{n}\Sigma_i 1\{X_i \leq t\}$
- parameters of the form $T(F)$ are called statistical functionals
- Example: $\mu = \mathrm{E}(X) = \int x dF(x) = T(F)$

$$\hat{\mu} = \int x d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Example: $\kappa_3 = \mathrm{E}(X-\mu)^3 = \int (x-\mu)^3 dF(x)$

$$\hat{\kappa}_3 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^3$$

- Example: $\theta = T(F) = F^{-1}(p)$ pth quantile

$$\hat{\theta} = T(\widehat{F}_n) = \widehat{F}_n^{-1}(p)$$

A very clever plug-in estimate

- suppose $\theta = T(F)$ and $\hat{\theta} = T(\widehat{F}_n)$, as above, and we want to estimate $\text{var}_F(\hat{\theta})$
- write $\text{var}_F(\hat{\theta}) = V(F)$, then use plug-in again

$$\widehat{var}(\hat{\theta}) = V(\widehat{F}_n)$$

- If $V(F)$ is very complicated, or unknown, simulate many samples from $\widehat{F}_n$, get many estimates of θ, use the simulation variance

e.g. $\theta = \kappa_3; \theta = F^{-1}(p)$

A very clever plug-in estimate

- suppose $\theta = T(F)$ and $\hat{\theta} = T(\widehat{F}_n)$, as above, and we want to estimate $\text{var}_F(\hat{\theta})$
- write $\text{var}_F(\hat{\theta}) = V(F)$, then use plug-in again

$$\widehat{var}(\hat{\theta}) = V(\widehat{F}_n)$$

- If $V(F)$ is very complicated, or unknown, simulate many samples from $\widehat{F}_n$, get many estimates of θ, use the simulation variance

e.g. $\theta = \kappa_3; \theta = F^{-1}(p)$

- each simulated sample $X^*_{1b}, \ldots, X^*_{nb}$ and each estimate $\hat{\theta}^*_b$

$$\widehat{var}(\hat{\theta}) \doteq \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^*_b - \bar{\hat{\theta}}^*_b)^2$$

- each simulated sample is a sample with replacement from the original data $X_1, \ldots, X_n$

§8.2: slightly different notaton

Real world $F \Longrightarrow X_1, \ldots, X_n \Longrightarrow T(F_n) = \hat{\theta}_n$

Bootstrap world $\widehat{F}_n \Longrightarrow X_1^*, \ldots, X_n^* \Longrightarrow T(\widehat{F}_n^*) = \hat{\theta}_n^*$ repeat B times

Example 8.7

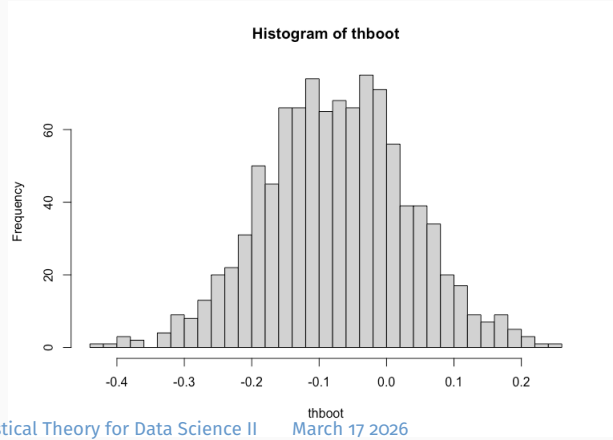
```
> AOSdata
  placebo   old   new     Z     Y
1    9243 17649 16449  8406 -1200
2    9671 12013 14614  2342  2601
3   11792 19979 17274  8187 -2705
4   13357 21816 23798  8459  1982
5    9055 13850 12560  4795 -1290
6    6290  9806 10157  3516   351
7   12412 17208 16570  4796  -638
8   18806 29044 26325 10238 -2719
```

$$Z = O - P, Y = N - O, \quad \frac{N-O}{O-P} = \frac{N-P}{O-P} - 1$$

```
> that <- mean(AOSdata$Y)/mean(AOSdata$Z)
> that
[1] -0.0713061
> thboot <- rep(0,1000)
> for(i in 1:1000){
+ Ystar <- sample(AOSdata$Y, 8, replace=TRUE)
+ Zstar <- sample(AOSdata$Z, 8, replace = TRUE
+ )
+ thboot[i] <- mean(Ystar)/mean(Zstar)}

> sqrt(var(thboot))
[1] 0.1077168
> that + sqrt(var(thboot))*c(-2,2)
[1] -0.2867397  0.1441275
```

```
> hist(thboot, breaks = 30); sort(thboot)[c(25,975)]
[1] -0.2856603  0.1445028
```



Bootstrap confidence intervals

- previous slide used $0.025, 0.975$ percentiles of bootstrap dist'n — percentile method
- studentized confidence intervals use dist'n of $(\hat{\theta}_b^* - \hat{\theta})/\hat{\sigma}^*$ — VanderVaart 23.1
- percentile method is invariant to monotone transformation $\varphi = \varphi(\theta)$

Bootstrap confidence intervals

- previous slide used $0.025, 0.975$ percentiles of bootstrap dist'n — percentile method
- studentized confidence intervals use dist'n of $(\hat{\theta}_b^* - \hat{\theta})/\hat{\sigma}^*$ — VanderVaart 23.1
- percentile method is invariant to monotone transformation $\varphi = \varphi(\theta)$

- studentized confidence intervals have correct coverage if both — VderV 23.2

$$\begin{aligned} \mathrm{pr}\{(\hat{\theta} - \theta)/\hat{\sigma} \leq t\} &\to F(t), \\ \mathrm{pr}\{\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^* \leq t | \hat{F}_n\} &\to F(t) \end{aligned}$$

- equivalent to

$$\sup_x \left| \mathrm{pr}\left(\frac{\hat{\theta} - \theta}{\hat{\sigma}} \leq x \mid F \right) - \mathrm{pr}\left(\frac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}^*} \leq x \mid \widehat{F}_n \right) \right| \xrightarrow{p} 0$$

proof involves triangular arrays

Bootstrap sampling

```
library("tidyverse")
leukemia_big<- read.csv ("http://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv")
leukemia_big[136,] %>% select(starts_with("ALL")) %>% as.numeric() -> all136
leukemia_big[136,] %>% select(starts_with("AML")) %>% as.numeric() -> aml136
t.test(all136,aml136, var.equal = TRUE)

95 percent confidence interval:
 -0.32817995 -0.06680742

> median(all136)-median(aml136)
[1] -0.235093
> mean(all136)-mean(aml136)
[1] -0.1974937
```

Bootstrap sample has 25 draws with replacement from AML; 47 draws with replacement from ALL

i.e. tailored to the two-sample problem

Bootstrap sampling

- sampling from $\widehat{F}_n(\cdot)$ only appropriate for i.i.d. data
- variations have been developed for time series, regression, censored data, etc.

- the parametric bootstrap resamples from $F(\cdot; \hat{\theta})$

Bootstrap sampling

- sampling from $\widehat{F}_n(\cdot)$ only appropriate for i.i.d. data
- variations have been developed for time series, regression, censored data, etc.

- the parametric bootstrap resamples from $F(\cdot; \hat{\theta})$

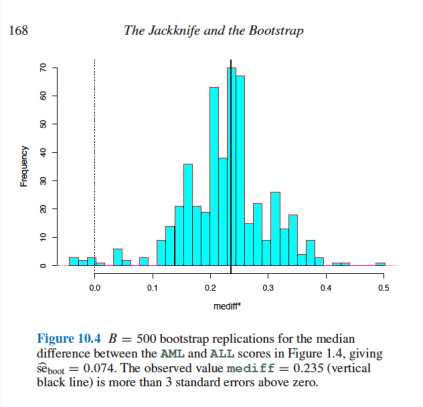


Figure 10.4 $B = 500$ bootstrap replications for the median difference between the `AML` and `ALL` scores in Figure 1.4, giving $\widehat{\text{se}}_{\text{boot}} = 0.074$. The observed value `mediff` $= 0.235$ (vertical black line) is more than 3 standard errors above zero.

- Efron and co-authors improved percentile confidence intervals using bias correction and acceleration
- resulting intervals (bounds) are second-order correct

BC_a CIs

$$\mathrm{pr}_F(\theta \leq \hat{\theta}^{\alpha}_{BCa}) = \alpha + \frac{c}{n} + o(\frac{1}{n})$$

- Efron and co-authors improved percentile confidence intervals using bias correction and acceleration — BC_a CIs
- resulting intervals (bounds) are second-order correct

$$\mathrm{pr}_F(\theta \leq \hat{\theta}^{\alpha}_{BCa}) = \alpha + \frac{c}{n} + o(\frac{1}{n})$$

- studentized bootstrap bounds are also second-order correct
- implicitly, bootstrap implement higher-order asymptotics — not clear why
- proof relies on Edgeworth expansion etc.
- VderV p.338 “The proofs of these assertions are somewhat technical”

- Efron and co-authors improved percentile confidence intervals using bias correction and acceleration — BC_a CIs
- resulting intervals (bounds) are second-order correct

$$\mathrm{pr}_F(\theta \leq \hat{\theta}^{\alpha}_{BCa}) = \alpha + \frac{c}{n} + o(\frac{1}{n})$$

- studentized bootstrap bounds are also second-order correct
- implicitly, bootstrap implement higher-order asymptotics — not clear why
- proof relies on Edgeworth expansion etc.
- VderV p.338 “The proofs of these assertions are somewhat technical”

- `library(boot)` is very reliable — Angelo Canty
- see also Davison & Hinkley (1997) *Bootstrap Methods and their Application*.

The Annals of Statistics
1979, Vol. 7, No. 1, 1–26

THE 1977 RIETZ LECTURE

BOOTSTRAP METHODS: ANOTHER LOOK AT THE JACKKNIFE

BY B. EFRON

Stanford University

We discuss the following problem: given a random sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from an unknown probability distribution F, estimate the sampling distribution of some prespecified random variable $R(\mathbf{X}, F)$, on the basis of the observed data $\mathbf{x}$. (Standard jackknife theory gives an approximate mean and variance in the case $R(\mathbf{X}, F) = \theta(\hat{F}) - \theta(F)$, θ some parameter of interest.) A general method, called the "bootstrap," is introduced, and shown to work satisfactorily on a variety of estimation problems. The jackknife is shown to be a linear approximation method for the bootstrap. The exposition proceeds by a series of examples: variance of the sample median, error rates in a linear

Part II Report

Project Guidelines STA 2212S: Statistical Theory for Data Science 2026

Presentation on March 31, 2026.
Report submission due April 14, 2026.

Part 1: Presentation

On the last day of class (March 31), you will present your final project. This includes:

- Uploading your slide deck to Quercus by 4 pm on March 31

Taylor & Fr
Taylor & Francis Gro

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Cross-Validation: What Does It Estimate and How Well Does It Do It?

Stephen Bates, Trevor Hastie & Robert Tibshirani

To cite this article: Stephen Bates, Trevor Hastie & Robert Tibshirani (2024) Cross-Validation: What Does It Estimate and How Well Does It Do It?, Journal of the American Statistical Association, 119:546, 1434-1445, DOI: 10.1080/01621459.2023.2197686

Cross-validation

- training data $(X_1, Y_1), \ldots, (X_n, Y_n)$
- build a regression model e.g. $f(x; \hat{\theta})$
- validation data $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+N}, Y_{n+N})$
- validation error

or some other loss function

$$\frac{1}{N} \sum_{j=1}^{N} \{Y_{n+j} - f(X_{n+j}; \hat{\theta})\}^2$$

- compare various models f by comparing their validation error

- k-fold cross-validation makes validation sets from the original observations

MSE_i validation error on ith holdout set

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$

[1] Introduction to Statistical Learning link

Cross-validation

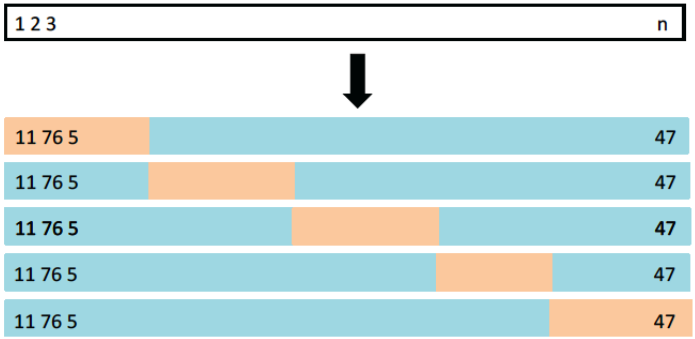


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

Cross-validation as an estimator

Bates et al 2023

- Out-of-sample error

over distribution of (X_{n+1}, Y_{n+1}) | sample

$$ERR_{XY} = \mathrm{E}\left[\left\{f(X_{n+1}; \hat{\theta}) - Y_{n+1}\right\}^2 \mid X, Y\right]$$

- average error

over distribution of $(Y_1, X_1), \ldots, (Y_n, X_n)$

$$ERR = \mathrm{E}(ERR_{XY})$$

be careful with Es

- ERR_{XY}: error of the model on our training data
- ERR: error of the fitting method on same-sized datasets
- Bates et al. argue that CV is a better estimate of ERR than it is of ERR_{XY}
- Suggest concentrating on §3, with 1 example from each of §5, 6.

Biometrika (2024), **111**, 2, *pp.* 677–689 https://doi.org/10.1093/biomet/asad044
Advance Access publication 18 July 2023

An anomaly arising in the analysis of processes with more than one source of variability

By H. S. BATTEY

Department of Mathematics, Imperial College London, South Kensington Campus, London SW7 2AZ, U.K.
h.battey@imperial.ac.uk

and PETER MCCULLAGH

Department of Statistics, University of Chicago, 5747 S Ellis Ave, Chicago, Illinois 60637, U.S.A.

- linear model $y = X\beta + Zb + \epsilon, \quad b \sim N_q(0, \Omega_b), \epsilon \sim N_n(0, \Omega)$ SM notation
- distribution: $y \mid b \sim N_n(X\beta + Zb, \Omega), \quad y \sim N_n(X\beta, Z\Omega_b Z^{\mathrm{T}} + \Omega)$
- when $\Omega = \sigma^2 I$, let $\sigma^2 \Upsilon^{-1} \equiv Z\Omega_b Z^{\mathrm{T}} + \Omega$
- parameters in Υ denoted ψ

$$\ell(\beta, \sigma, \psi) = -\frac{1}{2\sigma^2}(y - X\beta)^T \Upsilon(\psi)(y - X\beta) - \frac{n}{2}\log\sigma^2 + \frac{1}{2}\log|\Upsilon(\psi)|$$

- $\hat{\beta}_\psi = (X^{\mathrm{T}}\Upsilon(\psi)X)^{-1}X^{\mathrm{T}}\Upsilon(\psi)y, \quad \hat{\sigma}^2_\psi = \frac{1}{n}(y - X\hat{\beta}_\psi)^{\mathrm{T}}\Upsilon(\psi)(y - X\hat{\beta}_\psi)$
- $\ell_{\mathrm{p}}(\psi) = \ell(\hat{\beta}_\psi, \hat{\sigma}_\psi, \psi)$
- "Unfortunately life is not so simple."

boundary values, REML

$$\ell_{REML}(\beta, \sigma, \psi) = -\frac{1}{2\sigma^2}(y - X\beta)^T \Upsilon(\psi)(y - X\beta) - \frac{n-p}{2}\log\sigma^2 + \frac{1}{2}\log|\Upsilon(\psi)| - \log|X^{\mathrm{T}}\Upsilon X|$$

R for residual

- Battey/McCullagh use different notation: $Z\Omega_b Z^{\mathrm{T}} + \Omega \longrightarrow \sum_{u=0}^{s} \theta_u V_u$ $V_0 = I$; i.e. $\theta_0 = \sigma^2$
  but see their §5
- use only ℓ_{REML}
- focus on testing values of θ (variance components): $H_0 : \theta_s = 0$
- Wald statistic

$$(\hat{\theta}_s - 0)/\{I^{ss}(\hat{\boldsymbol{\theta}})\}^{1/2}$$

  they use the squared version $\hat{\theta}_s^2 / I^{ss}(\hat{\theta})$

- Likelihood ratio statistic

$$\Lambda = 2\{\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}^{(0)})\}$$

- It would be enough to do §3 or §5, with some numerical work from §6.1 or 6.2

Biometrika (2023), **110**, 1, *pp.* 83–99 https://doi.org/10.1093/biomet/asac021
Advance Access publication 5 April 2022

Testing generalized linear models with high-dimensional nuisance parameters

BY JINSONG CHEN

College of Applied Health Sciences, University of Illinois at Chicago, 1919 W Taylor St, Chicago, Illinois 60612, U.S.A.

jinsongc@uic.edu

QUEFENG LI

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.

- exponential family density:

$$f(y;\eta) = \exp\{y\eta - b(\eta) + c(y)\}, \quad \mu = \mathrm{E}(y) = b'(\eta)$$

- linear model $g(\mu) = Z^{\mathrm{T}}\gamma + W^{\mathrm{T}}\beta$ nuisance, interest
- interested in testing $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ γ unspecified
- dimensions of β, p_β and γ, p_γ allowed to diverge with n
- data $(y_1, z_1, w_1), \ldots, (y_n, z_n, w_n)$
- profile 'style': estimate γ under H_0, giving $\widehat{\gamma}_\phi$ (in their notation) $\widehat{\gamma}^{(0)}$

$$\widehat{\gamma}_\phi = \arg\min_\gamma \left(-\frac{1}{n}\sum_{i=1}^{n}[y_i\eta(z_i^{\mathrm{T}}\gamma) - b(\eta_i(z^{\mathrm{T}}\gamma)\}] + \zeta|\sum_{j=1}^{p_\gamma}|\gamma_j| \right)$$

- enough to give results and discuss assumptions; can omit asymptotic power discussion; can focus on one simulation scenario

The Annals of Applied Statistics
2008, Vol. 2, No. 4, 1360–1383
DOI: 10.1214/08-AOAS191

A WEAKLY INFORMATIVE DEFAULT PRIOR DISTRIBUTION FOR LOGISTIC AND OTHER REGRESSION MODELS

BY ANDREW GELMAN, ALEKS JAKULIN, MARIA GRAZIA PITTAU AND YU-SUNG SU

Columbia University, Columbia University, University of Rome, and City University of New York

We propose a new prior distribution for classical (nonhierarchical) logistic regression models, constructed by first scaling all nonbinary variables to have mean 0 and standard deviation 0.5, and then placing independent

Choosing priors

- logistic regression model

$$y_i \sim Bernoulli(p_i), \quad \log \frac{p_1}{1-p_i} = x_i^T \beta, \quad i = 1, \ldots, n$$

- if some linear combination of x_i's perfectly separates $y_i = 1$ from $y_i = 0$, then model is not identifiable

separable

- in practice, one or more components of $\hat{\beta}$ will be very large, as will $\hat{se}(\hat{\beta})$
- putting an informative prior on β avoids unbounded likelihood
- authors suggest independent Cauchy priors on each regression coefficient
- requires re-scaling to make coefficients comparable
- can ignore "other models" section; choose just one example, don't need detail on computation

Statistics and Computing (2024) 34:57
https://doi.org/10.1007/s11222-023-10366-5

ORIGINAL PAPER

Detecting and diagnosing prior and likelihood sensitivity w power-scaling

Noa Kallioinen[1] · Topi Paananen[1] · Paul-Christian Bürkner[2] · Aki Vehtari[1]

Received: 22 May 2023 / Accepted: 17 November 2023 / Published online: 31 December 2023

Power scaling

- usual posterior

$$\pi(\theta \mid y) \propto L(\theta; y)\pi(\theta)$$

- tempered likelihood

$$\tilde{\pi}(\theta \mid y) \propto L(\theta; y)^{\alpha}\pi(\theta)$$

- typically $\alpha < 1$
- proposed for Bayes inference in models with very many parameters

- this paper also considers

$$\breve{\pi}(\theta \mid y) \propto L(\theta; y)\pi(\theta)^{\alpha}$$

- Suggest emphasis on sensitivity; computational aspects can be omitted; one simulation and one case study, e.g.

Testing Random Effects for Binomial Data

Lucas Kania[†] Larry Wasserman[†,‡] Sivaraman Balakrishnan[†,‡]

[†]Department of Statistics and Data Science, Carnegie Mellon University
[‡]Machine Learning Department, Carnegie Mellon University

lucaskania@cmu.edu, {larry,siva}@stat.cmu.edu

November 4, 2025

Abstract

In modern scientific research, small-scale studies with limited participants are increasingly common. However, interpreting individual outcomes can be challeng-

Random effects binomial

- model: $X_i \mid p_i \sim Binom(t, p_i),\ p_i \sim \pi,\ i = 1, \ldots, n$ — π: variation over i in prob success
- problem: test $H_0 : \pi = \pi_0$ — goodness-of-fit
- e.g. $\pi_0 = \delta_{p_0}$, point mass at unknown p_0 — homogeneity testing
- emphasize finite-sample (fixed n) properties
- define alternative hypothesis as $H_1 : W_1(\pi, \pi_0) \geq \epsilon$

$$W_1(\pi, \pi_0) = \sup_{f \in class} \mathrm{E}_{p \sim \pi}\{f(p)\} - \mathrm{E}_{p \sim \pi_0}\{f(p)\}$$

- require test to have type I error α
- find test that maximizes type 2 error under the "worst case null, π_0"
- Can concentrate on either goodness-of-fit testing or homeogeneity testing; if g-o-f can emphasize plug-in test; can skip proofs and §s 4,5; one example

Pattern Recognition 107 (2020) 107501

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Handling incomplete heterogeneous data using VAEs

Alfredo Nazábal [a,*], Pablo M. Olmos [b], Zoubin Ghahramani [c,d], Isabel Valera [e,f]

[a] The Alan Turing Institute, London, United Kingdom
[b] University Carlos III, Madrid, Spain
[c] University of Cambridge, Cambridge, United Kingdom
[d] Uber AI Labs, San Francisco, US
[e] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[f] Department of Computer Science, Saarland University, Saarbrücken, Germany

Incomplete heterogeneous data and VAEs

- heterogeneous data: mix of D continuous, discrete, nominal, ordinal observations $\boldsymbol{x}_n = (x_{n1}, \ldots, x_{nd}),\ n = 1, \ldots, N$
- in each observation $\boldsymbol{x}_n$, entries may be MCAR: $\boldsymbol{x}_n = (\boldsymbol{x}_n^o, \boldsymbol{x}_n^m)$
- model for the data (with latent variables $\boldsymbol{z}_n \in \mathbb{R}^K$) decoder/generative model

$$p(\boldsymbol{x}_n, \boldsymbol{z}_n) = p(\boldsymbol{z}_n) \prod_d p(x_{nd} \mid \boldsymbol{z}_n), \quad \boldsymbol{z}_n \sim N_k(0, I)$$

- approximation to the posterior[2] encoder, variational approximation

$$q(\boldsymbol{z}_n, \boldsymbol{x}_n) \approx p(\boldsymbol{x}_n, \boldsymbol{z}_n)$$

- q designed to miminize (?) variational lower bound

$$\mathrm{E}_q\{\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q(\boldsymbol{z} \mid \boldsymbol{x})\}$$

[2] Kingma & Welling 2019 Link

… Incomplete heterogeneous data and VAEs

- authors introduce versions appropriate for MCAR,

$$p(\boldsymbol{x}_n \mid z_n) = \prod_{d \in O_n} p(x_{nd} \mid \boldsymbol{z}_n) \prod_{d \in M_n} p(x_{nd} \mid \boldsymbol{z}_n),$$

$$q(\boldsymbol{z}_n, \boldsymbol{x}_n^m \mid \boldsymbol{x}_n^o) = q(\boldsymbol{z}_n \mid x_n^o) \prod_{d \in M_n} p(x_{nd} \mid \boldsymbol{z}_n)$$

- densities p and q have parameters (here $\boldsymbol{h}$ I think)
- p is the likelihood function or the posterior density; q is an approximation to it

- Enough to focus on creating the model as in §2, could mention improvements (HI-VAE) in §3 at high level, one example from §4 enough

Annual Review of Statistics and Its Application

Quantile Regression for Survival Data

Limin Peng

Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia 30322, USA; email: lpeng@emory.edu

Quantile regression for survival data

- Recall: quantile regression Koenker[3]
  $\widehat{\beta}_\tau = \arg\min_\beta \rho_\tau(y_i - x_i^T\beta), \quad \rho_\tau(u) = u\{\tau - I(u \le 0)\}$
- survival time T, define $Y = \log T$,

$$Q_Y(\tau \mid \boldsymbol{Z}) \equiv \inf\{t : \mathrm{pr}(Y \le t \mid \boldsymbol{Z}) \ge \tau\} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_0(\tau), \quad 0 < \tau_L \le \tau \le \tau_U < 1$$

- data $(\tilde{T}_i, \Delta_i, \boldsymbol{Z}_i), i = 1, \ldots, n; \quad \tilde{T}_i = \min(Y_i, C_i), \quad \Delta_i = 1\{Y_i \le C_i\}, \quad \tilde{Y}_i = \log \tilde{T}_i$
- with no censoring estimating equation is $\sum_{i=1}^n \rho_\tau\{\tilde{Y}_i - Z_i^{\mathrm{T}}\beta(\tau)\} = 0$ $Y_i = \log T_i$, typo
- if censoring times known, this becomes $\sum_{i=1}^n \rho_\tau\{\tilde{Y}_i - \min(Z_i^{\mathrm{T}}\beta(\tau), U_i)\} = 0$ $U_i = \log C_i$
- variations for random censoring
- omit §3,4 can omit martingale and data augmentation in §2, 1 example from §5

[3] Koenker, Roger, and Kevin F. Hallock (2001). Quantile Regression. *J. Econ. Persp.* **15**, 143–56.

Local asymptotics of selection models with applications in Bayesian selective inference

Daniel García Rasines[1*] G. Alastair Young[2†]

April 15, 2025

Selective inference

- we fit several models and then select one, for inference about its parameters
- 'double-dipping': same data is used for model selection and for inference
- how to correct the inference?
  - make your method valid for any possible selection
  - make inference conditional on the selection — conditional
  - split the sample; select on one split, fit on the other — info splitting
- model supposes a r.v. U_n determining the selection, data $\boldsymbol{y}^n = (y_1, \ldots, y_n)$ *i.i.d.*

  e.g. $U_n = 1\{\bar{Y}_n > 0\}$
- likelihood function

$$f(\boldsymbol{y}^n; \theta) = \frac{f(u_n \mid \boldsymbol{y}^n)}{f(u_n; \theta)} \prod_{i=1}^{n} f(y_i; \theta) = \frac{p_n(\boldsymbol{y}^n)}{\varphi_n(\theta)} \prod_{i=1}^{n} f(y_i; \theta)$$

- Gaussian selection model if $f(y_i; \theta)$ is normal
- asymptotic expansion, §3.1, perhaps mention Bayesian version

Nonparametric inference under shape constraints: past, present and future

Richard J. Samworth*

Abstract. We survey the field of nonparametric inference under shape constraints, providing a historical overview and a perspective on its current state. An outlook and some open problems offer thoughts on future directions.

1 Introduction. Traditionally, we think of statistical methods as being divided into parametric approaches, which can be restrictive, but where estimation is typically straightforward (e.g. using maximum likelihood), and nonparametric methods, which are more flexible but often require careful choices of tuning parameters. Nonparametric inference under shape constraints sits somewhere in the middle, seeking in some ways the best of both worlds. The origins of the field are often traced to Grenander (1956), who proved that there exists a

Nonparametric inference under shape constraints

- Suppose $X_1, \ldots, X_n \in \mathbb{R}$ are i.i.d. with density $f(\cdot)$ — with respect to some measure
- Recall kernel density estimator of $f(\cdot)$

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

- requires specification of tuning parameter h and kernel function $K(\cdot)$
- if we restrict f to have 'nice' properties, can we find better estimators
- e.g. f is monotonically decreasing on $[0, \infty)$ — Grenander class
- e.g. $\log f$ is concave — log-concave class
- example result: for a measure Q, suppose $L(g, Q) = \int_0^\infty \log g dQ$ exists; for what Q does $\sup_g L(g, Q)$ exist; and what does the best g look like?
- one of Grenander/log-concave; main results; no proofs

Journal of Data Science 24 (1), 53–85
January 2026

DOI: 10.6339/25-JDS1211
Data Science Reviews

Causal Inference: A Tale of Three Frameworks

Linbo Wang[1,*], Thomas S. Richardson[2], and James M. Robins[3]

[1]*Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada*
[2]*Department of Statistics, University of Washington, Seattle, WA, U.S.A.*
[3]*Harvard T.H. Chan School of Public Health, Boston, MA, U.S.A.*

Abstract

Causal inference is a central goal across many scientific disciplines. Over the past several decades, three major frameworks have emerged to formalize causal questions and guide their analysis: the potential outcomes framework, structural equation models, and directed acyclic graphs. Although these frameworks differ in language, assumptions, and philosophical orientation, they often lead to compatible or complementary insights. This paper provides a comparative introduc-

Causal inference: Three frameworks

1. potential outcomes framework $A = 0, 1$ potential outcomes $Y(0), Y(1)$, covariates L
   $ACE \equiv \mathrm{E}\{Y(1)\} - \mathrm{E}\{Y(0)\}$, $CATE \equiv \mathrm{E}\{Y(1) \mid L\} - \mathrm{E}\{Y(0) \mid L\}$ March 3 and 10

2. structural equation models (nonparametric) ϵ are noise terms

   $$L \sim f_L(\epsilon_L); \quad A \sim f_A(L, \epsilon_A); \quad Y \sim f_Y(L, A, \epsilon_Y)$$

   linked to "interventional world" using "do operator": AoS §17.8

   $$L \sim f_L(\epsilon_L); \quad A = a; \quad Y \sim f_Y(L, a, \epsilon_Y)$$

   $ACE = \mathrm{E}\{f_Y(L, 1, \epsilon_Y)\} - \mathrm{E}\{f_Y(L, 0, \epsilon_Y)\}$

3. directed acyclic graphs

enough to link 1. to 2. or to 3.; **or** to present an overview of DAGs

Methods for correcting inference based on outcomes predicted by machine learning

Siruo Wang[a], Tyler H. McCormick[b,c], and Jeffrey T. Leek[a,1]

[a]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; [b]Department of Statistics, University of Washington, Seattle, WA 98195; and [c]Department of Sociology, University of Washington, Seattle, WA 98195

Edited by Robert Tibshirani, Stanford University, Stanford, CA, and approved October 6, 2020 (received for review January 24, 2020)

Many modern problems in medicine and public health leverage machine-learning methods to predict outcomes based on observable covariates. In a wide array of settings, predicted outcomes are used in subsequent statistical analysis, often without accounting for the distinction between observed and predicted outcomes. We call inference with predicted outcomes postprediction inference. In this paper, we develop methods for correcting statistical inference using outcomes predicted with arbitrarily complicated machine-learning models including random forests and deep neural nets. Rather than trying to derive the correction from first principles for each machine-learning algorithm, we observe that there is typically a low-dimensional and easily modeled represen-

known inheritance patterns for the disease. The predicted outcome can be used in place of the observed Alzheimer's status when performing a genome-wide association study (15).

This is just one example of the phenomenon of postprediction inference (postpi). Although common, this approach poses multiple statistical challenges. The predicted outcomes may be biased, or the predicted outcomes may have less variability than the actual outcomes. Standard practice in many applications is to treat predicted outcomes as if they were observed outcomes in subsequent regression models (6, 14–18). As we will show, uncorrected postprediction inference will frequently have deflated standard errors, bias, and inflated false positive rates.

Predicted outcomes from machine learning

- 1. labelled data for training $(y, \boldsymbol{x}) \longrightarrow \hat{f}(\boldsymbol{x})$ — estimates $g\{\mathrm{E}(Y \mid \boldsymbol{x})\}$
- 2. labelled data for testing $(\hat{y}_p, y, \boldsymbol{x})$ — compare prediction to 'truth'
- 3. validation data $(\hat{y}_p, \boldsymbol{x})$
  use validation data in a usual generalized linear model

- doesn't reflect uncertainty in $\hat{y}_p$
- use $r(\hat{y}_p, y)$ from Step 2 to correct regression at Step 3.

- they want a method that is not tuned to the particular m/l algorithm
- suggest a relatively simple model for Step 2. — Linear or logistic regression (I think)
- bootstrap approach recommended to correct bias, variance
- can concentrate on bootstrap correction **or** analytic correction, and on continuous **or** discrete data; may need to refer to supplement

Universal inference

Larry Wasserman[a,b,1,2], Aaditya Ramdas[a,1], and Sivaraman Balakrishnan[a,1]

[a]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213; and [b]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213

Contributed by Larry Wasserman, May 26, 2020 (sent for review December 26, 2019; reviewed by Peter Bühlmann and Robert Tibshirani)

We propose a general method for constructing confidence sets and hypothesis tests that have finite-sample guarantees without regularity conditions. We refer to such procedures as "universal." The method is very simple and is based on a modified version of the usual likelihood-ratio statistic that we call "the split likelihood-ratio test" (split LRT) statistic. The (limiting) null distribution of the classical likelihood-ratio statistic is often intractable when used to test composite null hypotheses in irregular statistical models. Our method is especially appealing for statistical inference in these complex setups. The method we suggest works for any parametric model and also for some nonparametric models, as long as computing a maximum-likelihood estimator (MLE) is feasible under the null. Canonical examples arise in mixture modeling and shape-constrained inference, for which construct-

as $n \to \infty$, where P_{θ^*} denotes the unknown true data-generating distribution.

Constructing tests and confidence intervals for irregular models—where the regularity conditions do not hold—is very difficult (1). An example is mixture models. In this case we observe $Y_1, \ldots, Y_n \sim P$ and we want to test

$$H_0 : \ P \in \mathcal{M}_{k_0} \ \text{versus} \ H_1 : \ P \in \mathcal{M}_{k_1}, \qquad [2]$$

where $\mathcal{M}_k$ denotes the set of mixtures of k Gaussians, with an appropriately restricted parameter space Θ (see for instance ref. 2) and with $k_0 < k_1$. Finding a test that provably controls the type I error at a given level has been elusive. A natural candidate is to base the test on the likelihood-ratio statistic but this turns out to have an intractable limiting distribution (3). As we

Universal Inference

- $Y_1, \ldots, Y_n \sim f(\cdot; \theta) \longrightarrow$ split data into two sets each of size n: (y_1, y_2) they use $p_\theta(y)$
- estimate θ using set 2 $\longrightarrow \hat{\theta}_2$
- construct confidence set for θ using

$$T_n(\theta) = \frac{L(\hat{\theta}_2; y_1)}{L(\theta; y_1)}$$

- confidence set

$$C_n = \{\theta : T_n(\theta) \leq \frac{1}{\alpha}\}$$

- Theorem 1

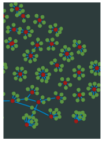
$$P_{\theta_{true}}(\theta_{true} \in C_n) \geq 1 - \alpha$$

… Universal Inference

- coverage guaranteed by Markov inequality
- method for estimating θ can be quite complex — even nonparametric
- there is a hypothesis testing version linked to the confidence set
- question arises about the power of the test
- for a simple case, Davison/Tse shows that the power is quite low
- there is active follow-up work on this approach, e.g. Strieder & Drton (2022) "On the choice of the splitting ratio…" *Electronic J Statist* **16** 6631–6650.

- can ignore hypothesis test part and possibly Chernoff bound; can omit nonparametric. Recommend Davison/Tse power analysis. Can omit §5-7. Enough to do one example from §3

Annual Review of Statistics and Its Application



ISSN: 2326-8298
eISSN: 2326-831X

AIMS AND SCOPE OF JOURNAL: The *Annual Review of Statistics and Its Application* informs statisticians, and users of statistics about major methodological advances and the computational tools that allow for their implementation.

Journal Status	Active
Impact Factor	8.7

Current Volume is OA

Latest Articles

Statistical Methods in Aging Research: Improving Current Practices and Embracing Emerging Approaches

Deependra K. Thapa, Erik S. Parker, Mounika Kandukuri, Xi (Rita) Wang, Thirupathi R. Mokalla, Olivia C. Robertson, Wasiuddin Najam, Andrew E. Teschendorff, Andrew W. Brown, John R. Speakman, Yisheng Peng, Bernard S. Gorman, Heping Zhang, Luis-Enrique Becerra-Garcia, Colby J. Vorland and David B. Allison
Vol. 13 (2026), pp. 493–525

The Enemies of Reliable and Useful Clinical Prediction Models: A Review of Statistical and Scientific Challenges

Ben Van Calster, Maarten van Smeden, Wouter van Amsterdam, Maarten Coemans, Laure Wynants and Ewout W. Steyerberg
Vol. 13 (2026), pp. 465–492

The Natural Value of Treatment and Its Importance for Causal Inference

Aaron L. Sarvet and Mats J. Stensrud