# **Mathematical Statistics II**

## STA2212H S LEC9101



### More on vaccines

### **Boyang Zhao's blog post**



### Vaccine efficacy

The way we measure vaccine efficacy is defined as follows,

vaccine efficacy (VE) = 1 - R

#### Boyang Zhao

O Netherlands

Twitter

LinkedIn

O GitHub

Senior data scientist at ING Bank Computational biology, genomics, systems biology, finance, banking @MIT PhD The R can be ratio of risks (RR, risk ratio); rates (RR, hicklence rate ratio); or hazards (RR, hazard ratio). Because of the ratios, we see that vaccine efficacy is a relative measure - in how much relative reduction in infection or disease in the vaccinated group compared to the unvaccinated group. A VE of 90% means there are 90% fewer cases in the vaccinated group compared to the placebo group. We mill use the subscripts – and <sub>p</sub> to denote vaccinated and placebo groups, respectively; but obviously the discussion is applicable for comparing between any two treatment arms - not necessarily have to be a placebo.

#### With RR

$$\mathrm{VE} = 1 - \mathrm{RR} = 1 - rac{c_v/N_v}{c_p/N_p}$$

where  $N_{v}$  and  $N_{p}$  are the total number of participants in the vaccinated and placebo group, respectively.

With IRR

$$\mathrm{VE} = 1 - \mathrm{IRR} = 1 - rac{c_v/T_v}{c_p/T_p}$$

where  $T_{\boldsymbol{v}}$  and  $T_{\boldsymbol{p}}$  are the time-person years for the vaccinated and placebo group, respectively.

With HR

$$\mathrm{VE} = 1 - \mathrm{HR} = 1 - rac{\lambda_v}{\lambda_p}$$

where  $\lambda_v$  and  $\lambda_p$  are the hazard rates for the vaccinated and placebo group, respectively. This measures the relative reduction in the hazard of infection. The hazard ratio can be

### link



 $\theta = (1 - VE)/(2 - VE)$ 

## **Polling analysis**

POLLING

### US election polls: a quick postmortem

How did the 2020 US presidential election polls really do? **Ole J. Forsberg** gives his assessment

The American Association for Public Opinion Research (AAPOR) is expected to produce a report early this year that explores the strengths and weaknesses of the polis in the zoao US election cycle. The polls were criticised in some quarters immediately after the election, when it became clear that Donald Trump had done better than expected and that Joseph R. Biden Jr's margin of victory in the popular vote was not as large as anticipated.'

In preparation for this report, I wanted to provide some insight into the polls and some suggestions of my own for moving forward. Specifically, I hope to convince polling houses to use some type of model averaging – or even Bayesian methods – to closing weeks of the campaign. The first source of error, faulty weighting, is extremely important for polling houses to take seriously. While the number of US polling houses taking education level into consideration increased in 2020, the education characteristics of the voting population remain uncertain.

"Sty votes" – the second source of error – may be more myth than reality (zgigh,ht/pNIb6R). But whether shy or not, there are some voters who either choose not to respond to polls, or who choose not to answer honesity when surveyed. Pollsters need to address this, either by asking additional questions to model respondent preference for those who choose not to say how they will vote, or by finding new ways to encourage the public to interpretation, not of polling. The mistake happens in how we interpret a poll result such as "A8% Biden, 44% Trump". Do we focus on the two-party vote and claim that Biden is ahead, or do we acknowledge that there is a sizeable portion of voters - 8% – who may only decide how to vote once in the polling booth? Clearly, the latter interpretation is more appropriate, but it makes for a less straightforward story, so these undecided voters tend to be overlooked in media reports.

### Missing data

The majority of polls in the 2020 election cycle contained just three response options for those asked about their intended vote: "Biden", "Trump", and "undecided". The implied fourth mentioned earlier constitute a huge amount of missing data about voting intention. Ignoring these missing data leads to false precision in the polls' assessment of the state of the election.

While some undecided voters ultimately will not vote, many will eventually decide between the two candidates. This increases the uncertainty in polling estimates beyond what is reported in terms of confidence intervals and margins of error. As a result, when those late-deciding voters finally vote, polls may look very wrong.

To illustrate this point, compare the polls in the final two weeks of the 2020 election to the final election result (Table 1). In this sample of 174 polls, the actual Biden vote was within the polls' margins of error 85% of the time while the actual Trump vote was within the polls' margins of error only 43% of the time. For the 57% of confidence intervals that missed Trump's actual vote, they were always too low, never too high meaning that the polls consistently underestimated Trump's final vote. The 15% of confidence intervals

Table 1: Results from comparing candidate support levels in polls from the last two weeks of the US presidential election with the actual outcome of the election (vote share). Polls are a mix of state-level and national polls from a variety of polling houses, using a variety of methods.

		Confidence interval hits		Average miss (standard error)	
Source	n	Biden	Trump	Biden	Trump
All polls	174	85% (79% to 90%)	43% (35% to 50%)	-0.09	+2.41
Online only	23	96% (78% to 99%)	30% (13% to 53%)	-0.79	+2.21
Online + telephone	26	92% (75% to 99%)	54% (33% to 73%)	-0.78	+2.24
Telephone only	125	82% (74% to 88%)	42% (34% to 52%)	-0.18	+2.48
University	60	92% (82% to 97%)	27% (16% to 40%)	-0.10	+2.99
Non-university	114	82% (73% to 88%)	51% (41% to 60%)	-0.09	+2.10
Partisan	52	79% (65% to 89%)	75% (61% to 86%)	+0.62	+1.33
Non-partisan	122	88% (81% to 93%)	29% (21% to 38%)	-0.40	+2.87

"Personally, I favour the Bayesian solution because it provides a solid statistical structure for estimation and communication of results."

# Latest issue of Applied Statistics (JRSS C)





Journal of the Royal Statistical Society: Series C (Applied Statistics) Volume 70, Issue 2

Pages: 249-506

March 2021

#### ISSUE INFORMATION

### Den Access

Pres Access issue Information Pages: 249-250   First Published: 08 March 2021	Future proofing a building design using history matching inspired level-set techniques Evan Baker, Peter Challenor, Matt Earnes Pages: 335-350 I First Published: 19 December 2020
DRIGINAL ARTICLES	Recurrent events modelling of haemophilia bleeding events Andrew C. Titman, Martin J. Wolfsegger, Thomas F. Jaki Pages: 351-371 I First Published: 07 January 2021
In Zhang, Inyoung Kim Pages: 251-269 I First Published: 01 December 2020	Multiscale null hypothesis testing for network-valued data: Analysis of brain networks of patients with autism
Quantile-frequency analysis and spectral measures for dia with nonlinear dynamics	lienia Lovato, Alessia Pini, Aymeric Stamm, Maxime Taquet, Simone Vantini Pages: 372-397 I First Published: 22 January 2021
Ta-Hsin Li Pages: 270-290 I First Published: 22 November 2020	ⓓ open Access Bayesian semi-parametric G-computation for causal inference in a cohort study with MAR dropout and death

Maria Josefsson Michael J Daniels

# Annals of Applied Statistics 2019 STATISTICS

AN OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS

Article Modelling multilevel spatial behaviour in binary-mark muscle fibre ION CORNWALL AND PHILIP W. SHEARD 1329 Identifying and estimating principal causal effects in a multi-site trial of Early College Imputation and post-selection inference in models with missing data: An application to colorectal cancer surveillance guidelines LIN LIU, YUOI OIU, LOKI NATARAJAN AND KAREN MESSER 1370 M A hidden Markov model approach to characterizing the photo-switching behavior of fluorophores ..... LEKHA PATEL, NILS GUSTAFSSON, YU LIN, RAIMUND OBER. RICARDO HENRIQUES AND EDWARD COHEN 1397 Identifying multiple changes for a functional data sequence with application to freeway. traffic segmentation ..... JENG-MIN CHIOU, YU-TING CHEN AND TAILEN HSING 1430 44 The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies .... JOHANN GAGNON-BARTSCH AND YOTAM SHEM-TOY. 1464 Spatio-temporal short-term wind forecast: A calibrated regime-switching method AHMED AZIZ EZZAT, MIKYOUNG JUN AND YU DING 1484 Network modelling of topological domains using Hi-C data ...... Y. X. RACHEL WANG, PURNAMRITA SARKAR, OANA URSU, ANSHUL KUNDAJE AND PETER J. BICKEL 1511 Fast dynamic nonparametric distribution tracking in electron microscopic data Kalman Rifer YANJUN OJAN, JJANHUA Z. HUANG, CHIWOO PARK AND YU DING 1537 Distributional regression forests for probabilistic precipitation forecasting in complex AND ACHIM ZEILEIS 1564 A Modeling seasonality and serial dependence of electricity price curves with warping 3 RCPnorm: An integrated system of random-coefficient hierarchical regression models for OIWEI LI, WEI LU, XIMING TANG, IGNACIO WISTUBA AND YANG XIE 1617 Network classification with applications to brain connectomics JESÚS D. ARROYO RELIÓN, DANIEL KESSLER, ELIZAVETA LEVINA AND STEPHAN F. TAYLOR 1648 Sequential decision model for inference and prediction on nonuniform hypergraphs with application to knot matching from computational forestry ...... SEONG-HWAN JUN 3 SAMUEL W. K. WONG, JAMES V. ZIDEK AND ALEXANDRE BOUCHARD-CÔTE 1678 A Bayesian mark interaction model for analysis of tumor pathology images cont. priots L . sims OIWELLL XINLEI WANG, FAMING LIANG AND GUANGHUA XIAO, 1708 A hierarchical Bayesian model for single-cell clustering using RNA-sequencing data VIVILIU JOSHUAL, WARREN AND HONGYU ZHAO 1733 Vol. 13 No. 3-September 2019

# of APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS

#### Articles-Continued from front cover

Incorporating conditional dependence in latent class models for probabilistic record SILL HULAND SHAUN GRANNIS 1753 Bayesian modeling of the structural connectome for studying Alzheimer's Disease ...... ARKAPRAVA ROY, SUBHASHIS GHOSAL, JEFFREY PRESCOTT AND KINGSHUK ROY CHOUDHURY 1791 Wavelet spectral testing: Application to nonstationary circadian rhythms JESSICA & HARGREAVES MARINA I KNIGHT JON W PITCHEORD RACHAEL J. OAKENFULL, SANGEETA CHAWLA, JACK MUNNS AND SETH J. DAVIS 1817 DUSTIN M. LONG. MARIO SIMS, JEFE M. SZYCHOWSKI, YUAN-I MIN-LESLIE A. MCCLURE, GEORGE HOWARD AND NOAH SIMON 184" Approximate inference for constructing astronomical catalogs from images JEFFREY REGIER, ANDREW C. MILLER, DAVID SCHLEGEL, RYAN P. ADAMS, 141 ION D. MCAULIFFE AND PRABHAT 1884 ( Bayesian methods for multiple mediators: Relating principal stratification and causal Ilysis of power plant emission controls D.P. powers; Sims; no hard mediation in the analysis of power plant emission controls CHRISTINE CHOIR AT AND CORWIN M. ZIGLER 1927 Radio-iBAG: Radiomics-based integrative Bayesian analysis of multiplatform genomic sport iby - inducing milli ARVIND U. K. RAO AND VEERABHADRAN BALADANDAYUTHAPANI 1957. Leng, Finty to A semiparametric modeling approach using Bayesian Additive Regression Trees with an hyper- pas application to evaluate heterogeneous treatment effects .... BRET ZELDOW, VINCENT LO RE III AND JASON ROY 1989 BART pior (Pice 6) & N pers; sins; weddinly interest child



- approximate posterior normality
- choosing a prior: subjective, conjugate, flat, convenience
- matching priors; Jeffreys' prior
- multiple parameters, marginal posterior
- Bayesian and frequentist philosophy

AoS §11.1

• empirical and epistemic probability

### RE: Meeting suggestions and new committee members

Today

- 1. Friday: Jeffreys-Lindley paradox (HW 6 (c)); DF re  $\chi^2$ ;
- 2. Bayesian inference overview
- 3. Two One weird examples
- 4. empirical Bayes
- 5. hierarchical Bayes
- Mar 15 5.15 6.15 pm EDT Data Science Speaker Series Jesse Cresswell Machine Learning Scientist, Layer 6 AI at TD "Evaluating Model Performance on Highly Imbalanced Datasets"

AoS 11.9, <del>11.10</del>



## Overview

- all information about  $\theta$  contained in posterior density  $\pi(\theta \mid x^n) = f(x^n \mid \theta)\pi(\theta)/f_{x^n}(x^n)$
- inference about  $\psi(\theta)$  based on marginal posterior
- for comparing two (or more) points  $\theta$  in  $\pi(\theta \mid x^n)$ , don't need marginal distribution of  $X^n$
- for choosing between models, do need marginal distribution of  $X^n$ , as in HW 6 (c)
- Bayesian predictions of future values:

posterior predictive

$$\pi(\mathbf{x}_{new} \mid \mathbf{x}^n) = \int f(\mathbf{x}_{new} \mid \mathbf{x}^n, \theta) \pi(\theta \mid \mathbf{x}^n) d\theta,$$

- probability statements refer to uncertainty of knowledge
- choosing priors can be difficult, and can have large impact in high-dimensional settings
- most applications of Bayesian inference involve sampling from the posterior density
- or approximating the posterior density

normal, Laplace

## **Bayesian computation**

- excellent overview of Bayesian computational methods in SM 11.3
- Laplace approximation of integrals
- importance sampling
- Markov chain Monte Carlo sampling
  - Gibbs sampling
  - Metropolis-Hastings algorithm



Example 11.9 AoS

 $(X_1, R_1, Y_1), \dots, (X_n, R_n, Y_n)$  i.i.d.; parameter  $\theta = (\theta_1, \dots, \theta_B)$ , B very large

Mathematical Statistics II March 10 2021



Mathematical Statistics II March 10 2021