In Praise of Small Data

Statistical Science and Data Science

Nancy Reid
University of Toronto

Department of Statistics, OSU
April 11 2019

Outline

Statistics at a Crossroads

Examples: Statistics in the news

Statistical theory

Statistics and data science

Statistics at a Crossroads

Statistics at a crossroads

NSF award

- NSF workshop and report
- "... we are at a crossroads with an unprecedented opportunity to modernize ... to become the major player in data science, but also with a non-ignorable risk to make ourselves obsolete in the broad community of data science."
- "... critical question, where do we go from here?"

… Challenges and Opportunities

Tukey 1962

The Annals of Mathematical Statistics

Vol. 33, No. 1, Mar., 1962

Published by: Institute of Mathematical Statistics
https://www.jstor.org/stable/i312810

Journal Home Page

"The future of data analysis can … lead to the provision of a great service to all fields of science and technology. Will it? That remains to … our willingness to take up the rocky road of real problems in preferences to the smooth road of unreal assumptions … Who is for the challenge?"

THE FUTURE OF DATA ANALYSIS[1]

By John W. Tukey

Princeton University and Bell Telephone Laboratories

FIELDS
THE FIELDS INSTITUTE
THEMATIC PROGRAM ON STATISTICAL INFERENCE, LEARNING, AND MODELS FOR
BIG DATA
JANUARY - JUNE, 2015
PROGRAM
JANUARY 12 – 23, 2015
Opening Conference and Boot Camp
Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Bin Yu
JANUARY 26 – 30, 2015
Workshop on Big Data and Statistical Machine Learning
Organizing committee: Ruslan Salakhutdinov (Chair), Dale Schuurmans, Yoshua Bengio, Hugh Chipman, Bin Yu
FEBRUARY 9 – 13 , 2015
Workshop on Optimization and Matrix Methods in Big Data
Organizing Committee: Stephen Vavasis (Chair), Anima Anandkumar, Petros Drineas, Michael Friedlander , Nancy Reid, Martin Wainwright
FEBRUARY 23 – 27, 2015
Workshop on Visualization for Big Data: Strategies and Principles
Organizing Committee: Nancy Reid (Chair), Susan Holmes, Snehelata Huzurbazar, Hadley Wickham, Leland Wilkinson
MARCH 23 – 27, 2015
Workshop on Big Data in Health Policy
Organizing Committee: Lisa Lix (Chair), Constantine Gatsonis , Sharon-Lise Normand
APRIL 13 – 17, 2015
Workshop on Big Data for Social Policy
Organizing Committee: Sallie Keller (Chair), Robert Groves, Mary Thompson
JUNE 13 – 14, 2015
Closing Conference
Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Hugh Chipman, Ruslan Salakhutdinov, Yoshua Bengio, Richard Lockhart
to be held at AARMS of Dalhousie University
GRADUATE COURSES
JANUARY TO APRIL 2015
Large Scale Machine Learning
Instructor: Ruslan Salakhutdinov (University of Toronto)
JANUARY TO APRIL 2015
Topics in Inference for Big Data
Instructors: Nancy Reid (University of Toronto), Mu Zhu (University of Waterloo)
This thematic program emphasizes both applied and theoretical aspects of statistical inference, learning and models in big data. The opening conference will serve as an introduction to the program, concentrating on overview lectures and background preparation. Workshops throughout the program will highlight cross-cutting themes, such as learning and visualization, as well as focus themes for applications in the social, physical and life sciences. It is expected that all activities will be webcast using the FieldsLive system to permit wide participation. Allied activities planned include workshops at PIMS in April and May and CRM in May and August.
ORGANIZING COMMITTEE
Yoshua Bengio (Montréal)
Hugh Chipman (Acadia)
Sallie Keller (Virginia Tech)
Lisa Lix (Manitoba)
Richard Lockhart (Simon Fraser)
Nancy Reid (Toronto)
Ruslan Salakhutdinov (Toronto)
INTERNATIONAL ADVISORY COMMITTEE
Constantine Gatsonis (Brown)
Susan Holmes (Stanford)
Snehelata Huzurbazar (Wyoming)
Nicolai Meinshausen (ETH Zurich)
Dale Schuurmans (Alberta)
Robert Tibshirani (Stanford)
Bin Yu (UC Berkeley)
For more information, allied activities off-site, and registration, please visit:
www.fields.utoronto.ca/programs/scientific/14-15/bigdata

expectations
Consumer 3D Printing
Gamification
Wearable User Interfaces
Big Data
Complex-Event Processing
Natural-Language Question Answering
Internet of Things
Content Analytics
Speech-to-Speech Translation
In-Memory Database Manag
Virtual Assistants
Mobile Robots
3D Scanners
Neurobusiness
Biochips
Autonomous Vehicles
Prescriptive Analytics
Affective Computing
Electrovibration
Volumetric and Holographic Displays
Human Augmentation
Brain-Computer Interface
3D Bioprinting
Quantified Self
Quantum Computing
Smart Dust
Bioacoustic Sensing
Augmented Reality
Machine-to-Machine Co
Mobile Health Monitorin
NFC
Mesh Networks: Sensor
Cloud Computing
In-Memo
Virtual Reali
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment

Big Data

2014


THE FIELDS INSTITUTE

FIELDS

THEMATIC PROGRAM ON STATISTICAL INFERENCE, LEARNING, AND MODELS FOR BIG DATA

JANUARY – JUNE, 2015

PROGRAM

JANUARY 12 – 23, 2015
Opening Conference and Boot Camp
Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Bin Yu

JANUARY 26 – 30, 2015
Workshop on Big Data and Statistical Machine Learning
Organizing committee: Ruslan Salakhutdinov (Chair), Dale Schuurmans, Yoshua Bengio, Hugh Chipman, Bin Yu

FEBRUARY 9 – 13, 2015
Workshop on Optimization and Matrix Methods in Big Data
Organizing Committee: Stephen Vavasis (Chair), Anima Anandkumar, Petros Drineas, Michael Friedlander, Nancy Reid, Martin Wainwright

FEBRUARY 23 – 27, 2015
Workshop on Visualization for Big Data: Strategies and Principles
Organizing Committee: Nancy Reid (Chair), Susan Holmes, Snehelata Huzurbazar, Hadley Wickham, Leland Wilkinson

MARCH 23 – 27, 2015
Workshop on Big Data in Health Policy
Organizing Committee: Lisa Lix (Chair), Constantine Gatsonis, Sharon-Lise Normand

APRIL 13 – 17, 2015
Workshop on Big Data for Social Policy
Organizing Committee: Sallie Keller (Chair), Robert Groves, Mary Thompson

JUNE 13 – 14, 2015
Closing Conference
Organizing Committee: Nancy Reid (Chair), Sallie Keller, Lisa Lix, Hugh Chipman, Ruslan Salakhutdinov, Yoshua Bengio, Richard Lockhart
to be held at AARMS of Dalhousie University

GRADUATE COURSES

JANUARY TO APRIL 2015
Large Scale Machine Learning
Instructor: Ruslan Salakhutdinov (University of Toronto)

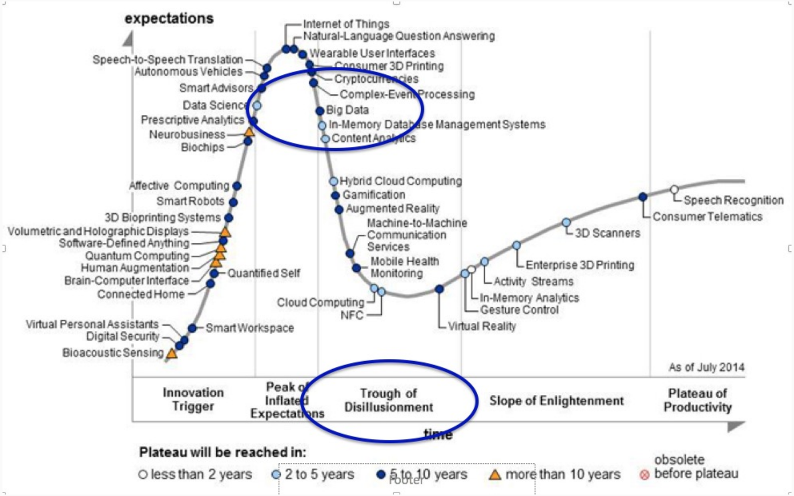
JANUARY TO APRIL 2015
Topics in Inference for Big Data
Instructors: Nancy Reid (University of Toronto), Mu Zhu (University of Waterloo)

This thematic program emphasizes both applied and theoretical aspects of statistical inference, learning and models in big data. The opening conference will serve as an introduction to the program, concentrating on overview lectures and background preparation. Workshops throughout the program will highlight cross-cutting themes, such as learning and visualization, as well as focus themes for applications in the social, physical and life sciences. It is expected that all activities will be webcast using the FieldsLive system to permit wide participation. Allied activities planned include workshops at PIMS in April and May and CRM in May and August.

ORGANIZING COMMITTEE
Yoshua Bengio (Montréal)
Hugh Chipman (Acadia)
Sallie Keller (Virginia Tech)
Lisa Lix (Manitoba)
Richard Lockhart (Simon Fraser)
Nancy Reid (Toronto)
Ruslan Salakhutdinov (Toronto)

INTERNATIONAL ADVISORY COMMITTEE
Constantine Gatsonis (Brown)
Susan Holmes (Stanford)
Snehelata Huzurbazar (Wyoming)
Nicolai Meinshausen (ETH Zurich)
Dale Schuurmans (Alberta)
Robert Tibshirani (Stanford)
Bin Yu (UC Berkeley)

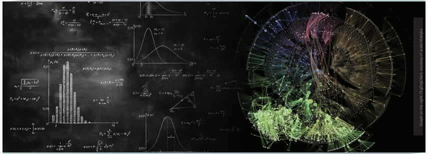
For more information, allied activities off-site, and registration, please visit:
www.fields.utoronto.ca/programs/scientific/14-15/bigdata


expectations
Internet of Things
Natural-Language Question Answering
Wearable User Interfaces
Consumer 3D Printing
Cryptocurrencies
Complex-Event Processing
Big Data
In-Memory Database Management Systems
Content Analytics
Speech-to-Speech Translation
Autonomous Vehicles
Smart Advisors
Data Science
Prescriptive Analytics
Neurobusiness
Biochips
Affective Computing
Smart Robots
3D Bioprinting Systems
Volumetric and Holographic Displays
Software-Defined Anything
Quantum Computing
Human Augmentation
Brain-Computer Interface
Connected Home
Quantified Self
Virtual Personal Assistants
Digital Security
Bioacoustic Sensing
Smart Workspace
Hybrid Cloud Computing
Gamification
Augmented Reality
Machine-to-Machine Communication Services
Mobile Health Monitoring
Cloud Computing
NFC
Virtual Reality
Gesture Control
In-Memory Analytics
Activity Streams
Enterprise 3D Printing
3D Scanners
Consumer Telematics
Speech Recognition
As of July 2014
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau
Footer

Data Science

2018

Statistical Inference, Learning and Models in Data Science



September 24 - 27, 2018 at **THE FIELDS INSTITUTE**
September 28, 2018 at **MARS**

This is a retrospective workshop for the 2015 thematic program *Statistical Models, Learning and Inference* for Big Data. We will reflect on recent progress and the shift in emphasis to data science in the intervening three years.

Invited Speakers

Edoardo Airoldi, *Harvard University*
Jimmy Ba, *University of Toronto*
Jelena Bradic, *University of California*
Fanny Chevalier, *University of Toronto*
Michael Correll, *Tableau*
Debbie Dupuis, *HEC Montreal*
Ruth Etzioni, *Fred Hutchinson Cancer Research Center*
Mark Fox, *University of Toronto*
Marzyeh Ghassemi, *MIT*
Laura Hatfield, *Harvard Medical School*
Heike Hofmann, *Iowa State University*
Eric Kolaczyk, *Boston University*
Todd Kuffner, *Washington University*
Simon Lacoste-Julien, *University of Montreal*
Rahul Mazumder, *MIT Sloan School*
Isabel Meirelles, *OCAD University*
Raymond Ng, *University of British Columbia*
Sofia Olhede, *University College London*
George Paliouras, *IIT Athens*
Greg Ridgeway, *University of Pennsylvania*
Veronika Rockova, *University of Chicago*
Mark Schmidt, *University of British Columbia*
Ravi Shroff, *New York University*
Nathan Srebro, *Toyota Technical Institute*
Yaoliang Yu, *University of Waterloo*
Francis Zwiers, *University of Victoria*

... more speakers on the Industry Day, on Friday September 28!

Organizing Committee

Fanny Chevalier, *University of Toronto*
David Duvenaud, *University of Toronto*
Sallie Keller, *Virginia Tech*
Lisa Lix, *University of Manitoba*
Nancy Reid, *University of Toronto*
Nathan Taback, *University of Toronto*
Stephen Vavasis, *University of Waterloo*

Statistics and Probability Letters 136 (2018) 42–45
Contents lists available at ScienceDirect
Statistics and Probability Letters
journal homepage: www.elsevier.com/locate/stapro
ELSEVIER

The role of Statistics in the era of big data

Edited by Laura Sangalli
Volume 136, Pages 1-170 (May 2018)

Actions for selected articles
Select all / Deselect all
Download PDFs
Export citations
Show all article previews

Full text access
Editorial Board
Page ii
Download PDF

Editorial • Full text access
The role of Statistics in the era of Big Data
Laura M. Sangalli
Pages 1-3
Download PDF

Short communication • Full text access
Statistics in the big data era: Failures of the machine
David B. Dunson
Pages 4-9
Download PDF Article preview

Short communication • Full text access
On the role of statistics in the era of big data: A call for a debate
Piercesare Secchi
Pages 10-14

Examples: Statistics in the news

Example 1: human longevity

NEWS · 28 JUNE 2018

There's no limit to longevity, says study that revives human lifespan debate

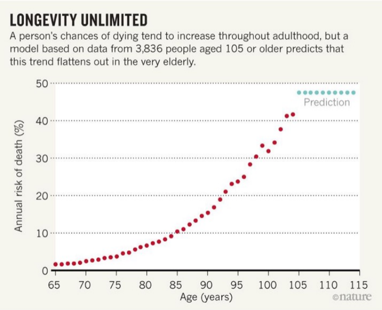
Death rates in later life flatten out and suggest there may be no fixed limit on human longevity, countering some previous work.

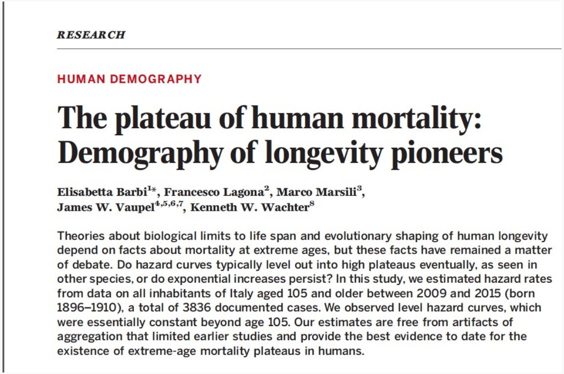
Elie Dolgin

Nature News
June 28 2018

"the study included fewer than 100 people who lived to 110 or beyond"
"even small inaccuracies in the Italian longevity records could lead to a spurious conclusion"

E. Dolgin, Nature





RESEARCH

HUMAN DEMOGRAPHY

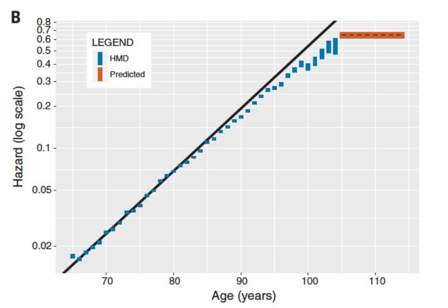
The plateau of human mortality: Demography of longevity pioneers

Elisabetta Barbi[1*], Francesco Lagona[2], Marco Marsili[3], James W. Vaupel[4,5,6,7], Kenneth W. Wachter[8]

Theories about biological limits to life span and evolutionary shaping of human longevity depend on facts about mortality at extreme ages, but these facts have remained a matter of debate. Do hazard curves typically level out into high plateaus eventually, as seen in other species, or do exponential increases persist? In this study, we estimated hazard rates from data on all inhabitants of Italy aged 105 and older between 2009 and 2015 (born 1896–1910), a total of 3836 documented cases. We observed level hazard curves, which were essentially constant beyond age 105. Our estimates are free from artifacts of aggregation that limited earlier studies and provide the best evidence to date for the existence of extreme-age mortality plateaus in humans.

Science

June 29 2018



"We observed level hazard curves, which were essentially constant beyond age 105"

"... provide the best evidence to date for the existence of extreme-age mortality plateaus"

" 'This study is unlikely to be the last word on the age-limit dispute', says Haim Cohen, a molecular biologist at Bar-Ilan University in Ramut-Gan Israel. 'I'm sure that the debate is going to continue'."

Dolgin, Nature, June 2018

PLOS | BIOLOGY

"The capacity for data entry and age inflation errors provides a sufficient model to explain late-life mortality patterns observed by Barbi and colleagues"

FORMAL COMMENT

Plane inclinations: A critique of hypothesis and model choice in Barbi et al

Saul Justin Newman *

Research School of Biology, The Australian National University, Acton, ACT, Australia

* saul.newman@anu.edu.au

Abstract

This study highlights how the mortality plateau in Barbi and colleagues can be generated by low-frequency, randomly distributed age-misreporting errors. Furthermore, sensitivity of the

"… claims of Barbi and colleagues rest on nearly 4,000 carefully validated cases from an established registration system. A critique like Newman's, … can hardly carry force."

PLOS | BIOLOGY

FORMAL COMMENT

Hypothetical errors and plateaus: A response to Newman

Kenneth W. Wachter *

Department of Demography, University of California, Berkeley, California, United States of America

* wachter@demog.berkeley.edu

Abstract

Newman questions recent claims about a plateau in mortality rates for Italians beyond age 105 on the basis of a hypothetical model. His model implies implausibly high error rates for

... is there a limit to human longevity?

- claims that age-misreporting can generate spurious late life plateaus
- Barbi et al (2018) fit a parametric model and used likelihood ratio test to compare to a constant hazard

  for age > 105
- Newman argued that a modelling choice they made influenced their results
- "of the 861 ... combinations tested, the model selected by Barbi et al generated the single largest late-life mortality plateau"

- statistics: Gompertz model, LRT, power analysis

  $h(x) = ae^{bx}e^{\beta_1 C + \beta_2 M}$
- **data science:** 861 such fits, plus simulated errors

  with probabilities ranging from 10^{-3} to 10^{-6}
- **domain science:** all inhabitants of Italy aged ≥ 105 years 2009–2015 (3836 cases) + Human Mortality Database

… is there a limit to human longevity?

Barbi et al 2018

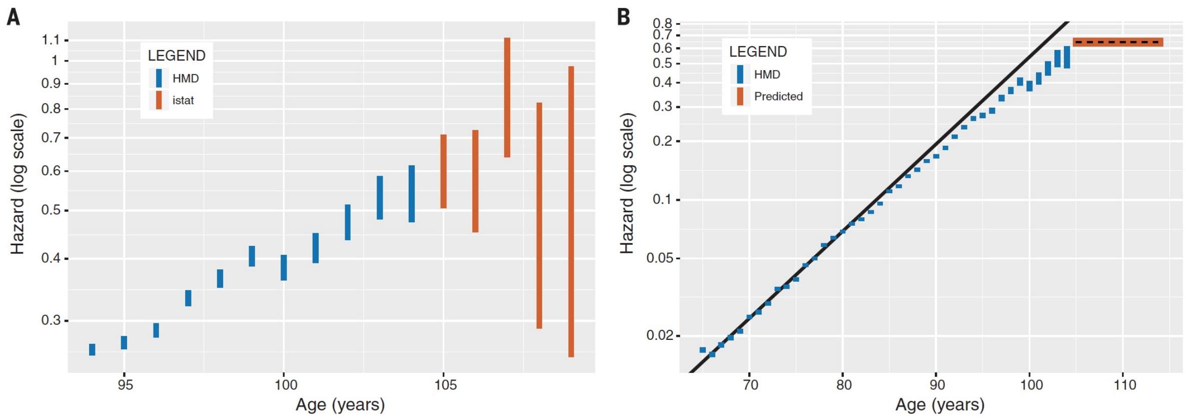


Fig. 1. Yearly hazards on a logarithmic scale for the cohort of Italian women born in 1904. Confidence intervals were derived from Human Mortality Database (HMD) data for ages up to 105 and from ISTAT data beyond age 105. (**A**) Closeup with 95% confidence intervals based solely on single-cohort data. (**B**) Broad view with estimated plateau beyond age 105 (black dashed line) and 95% confidence bands (orange) predicted from the model parameters based on the full ISTAT database, along with a straight-line prediction (black) from fitting a Gompertz model to ages 65 to 80.

Example 2: wildfire

THE GLOBE AND MAIL CANADA SUBSCRIBE REGISTER LOG IN AdChoices

B.C. wildfires stoked by climate change, likely to become worse: study

Jeff Lewis
Jan 8 2019
Globe & Mail

JEFF LEWIS › ENVIRONMENT REPORTER
PUBLISHED JANUARY 8, 2019
UPDATED 18 HOURS AGO

A helicopter flies over a wildfire southwest of the town of Cache Creek, B.C., on July 18, 2017.

BEN NELMS/REUTERS

TRENDING

1. OPINION
   As parents of complex special-needs kids, we know inclusive education doesn't work
   PHIL RICHMOND AND HAYLEY AVRUSKIN
2. New Canadian telescope detecting more brief, powerful radio blasts from far beyond our galaxy
3. Jagmeet Singh gets his chance as Trudeau calls three by-elections, including in Burnaby South
4. BMO slices 1,000 points from its Toronto stock market forecast
5. Toronto's Vena secures $115-million in financing from U.S. private-equity firms

Event attribution

Example 2

AGU100 ADVANCING EARTH AND SPACE SCIENCE

Earth's Future

RESEARCH ARTICLE

10.1029/2018EF001050

Attribution of the Influence of Human-Induced Climate Change on an Extreme Fire Season

M. C. Kirchmeier-Young[1,2], N. P. Gillett[2], F. W. Zwiers[1], A. J. Cannon[3], and F. S. Anslow[1]

[1]Pacific Climate Impacts Consortium, University of Victoria, Victoria, British Columbia, Canada, [2]Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada, Victoria, British Columbia, Canada, [3]Climate Research Division, Environment and Climate Change Canada, Victoria, British Columbia, Canada

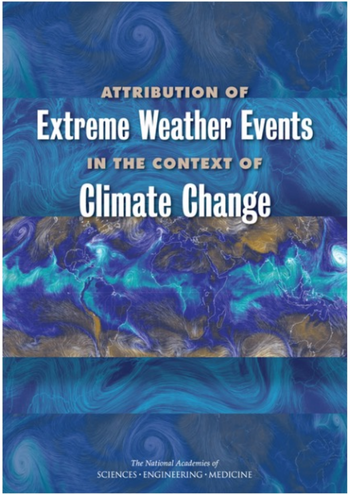
Key Points:

- An event attribution analysis is performed for the record-breaking wildfire season of 2017 in BC
- Anthropogenic climate change greatly increased the likelihood of extreme warm temperatures and high fire risk
- A strong anthropogenic climate change contribution is also found for the large area burned

Supporting Information:

- Supporting Information S1

Abstract A record 1.2 million ha burned in British Columbia, Canada's extreme wildfire season of 2017. Key factors in this unprecedented event were the extreme warm and dry conditions that prevailed at the time, which are also reflected in extreme fire weather and behavior metrics. Using an event attribution method and a large ensemble of regional climate model simulations, we show that the risk factors affecting the event, and the area burned itself, were made substantially greater by anthropogenic climate change. We show over 95% of the probability for the observed maximum temperature anomalies is due to



The relatively young science of extreme event attribution seeks to tease out the influence of human-caused climate change from other factors, such as natural sources of variability like El Niño, as contributors to individual extreme events.

Consensus Report

National Academy of Sciences Engineering and Medicine

- “anthropogenic climate change increased the area burned by a factor of 7 - 11”
- “We use a large ensemble of CanRCM4 … consisting of 50 realizations on a 50-km grid. Each realization is driven by a member of the CanESM2 … large ensemble … We utilize data from 1961 to 2020.”
- “A data set of gridded maximum (and minimum) temperature and precipitation anomalies was created by interpolating monthly values calculated from surface station observations relative to a 30-year climatology. Observational data was acquired from numerous sources and interpolated using a thin plate spline methodology.”
- “… values for each year and large ensemble realization were pooled together for two time periods: 1961-1970 and 2011-2020, resulting in 500 values for each decade (10 years x 50 realizations). ”

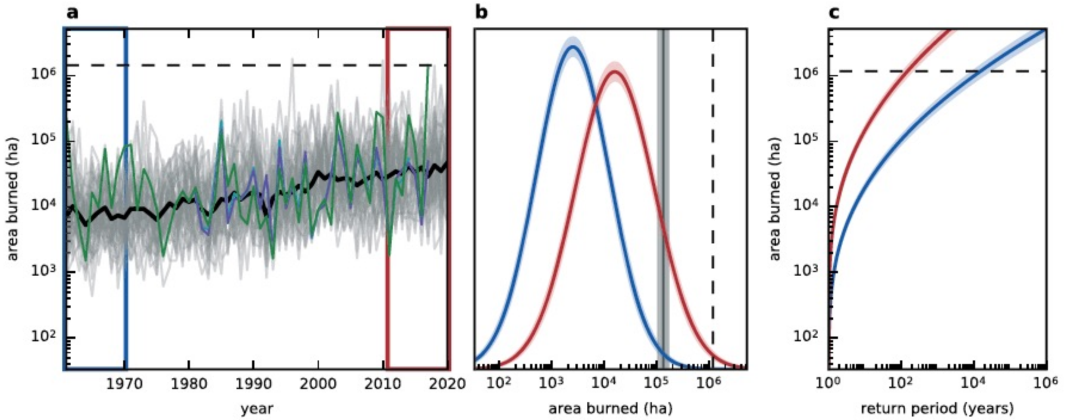
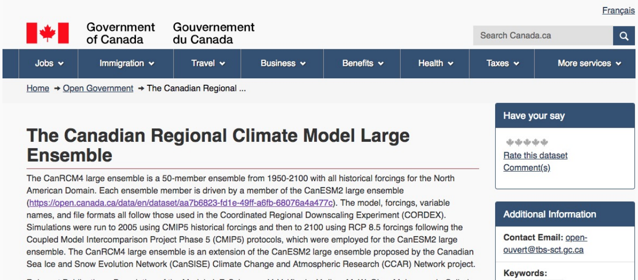


Figure 5. Time series (a, log scale) of regression-predicted annual burned area in the BC Southern Cordillera for bias-corrected CanRCM4 realizations (gray) and ensemble mean (bold), reanalysis (turquoise/purple), and observations (green). The dashed line marks the observed 2017 value. Probability distributions (b) for area burned amounts (log scale) from decades outlined in corresponding colors in (a). The gray bar indicates the area burned amount in the distribution with reduced anthropogenic influence (blue) of a corresponding percentile to the 2017 amount (dashed line) in the

- complex computer simulation of global climate — mathematics, numerical analysis
- creation of regional climate scenarios — mathematics, statistics
- combined with available observational data — statistics, data science
- modelled with regression and kernel density estimation — statistics, mathematics


Français

Government of Canada | Gouvernement du Canada

Search Canada.ca

Jobs | Immigration | Travel | Business | Benefits | Health | Taxes | More services

Home → Open Government → The Canadian Regional …

The Canadian Regional Climate Model Large Ensemble

The CanRCM4 large ensemble is a 50-member ensemble from 1950-2100 with all historical forcings for the North American Domain. Each ensemble member is driven by a member of the CanESM2 large ensemble (https://open.canada.ca/data/en/dataset/aa7b6823-fd1e-49ff-a6fb-68076a4a477c). The model, forcings, variable names, and file formats all follow those used in the Coordinated Regional Downscaling Experiment (CORDEX). Simulations were run to 2005 using CMIP5 historical forcings and then to 2100 using RCP 8.5 forcings following the Coupled Model Intercomparison Project Phase 5 (CMIP5) protocols, which were employed for the CanESM2 large ensemble. The CanRCM4 large ensemble is an extension of the CanESM2 large ensemble proposed by the Canadian Sea Ice and Snow Evolution Network (CanSISE) Climate Change and Atmospheric Research (CCAR) Network project.

Have your say

Rate this dataset

Comment(s)

Additional Information

Contact Email: open-ouvert@tbs-sct.gc.ca

Keywords:

Bad Diets Are Responsible For More Deaths Than Smoking, Global Study Finds

April 3, 2019 · 6:31 PM ET

Heard on All Things Considered

ALLISON AUBREY

Example 3: Diet and mortality

Bad Diets Are Responsible For More Deaths Than Smoking, Global Study Finds

April 3, 2019 · 6:31 PM ET
Heard on All Things Considered

ALLISON AUBREY

… diet and disease

GBD Diet Collaborators 2019

THE LANCET

Log in Register Subscribe Claim

ARTICLES | ONLINE FIRST

PDF [1 MB] Figures Save Share Reprints Request

Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017

GBD 2017 Diet Collaborators

Open Access • Published: April 03, 2019 • DOI: https://doi.org/10.1016/S0140-6736(19)30041-8

Check for updates

PlumX Metrics

Summary
Introduction
Methods
Results
Discussion

Summary

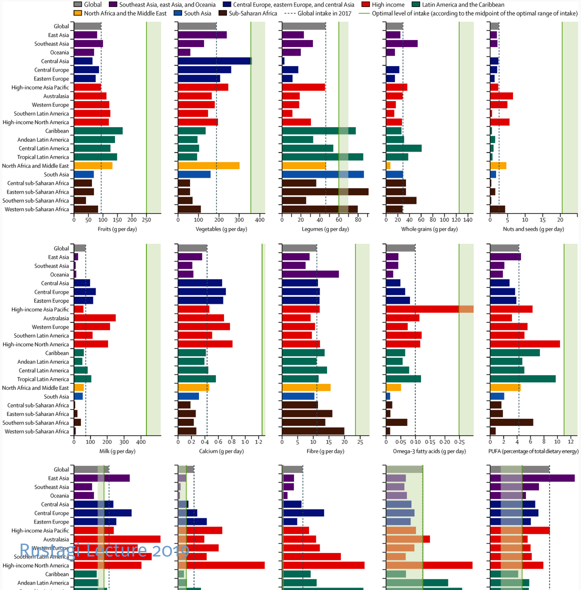
Background

Suboptimal diet is an important preventable risk factor for non-communicable diseases (NCDs); however, its impact on the burden of NCDs has not been systematically evaluated. This study aimed to evaluate the consumption of major foods and nutrients across 195 countries and to quantify the impact of their suboptimal intake on NCD mortality and morbidity.

- “this study aimed to evaluate the consumption of major foods and nutrients across 195 countries and to quantify the impact of their suboptimal intake”
- “we estimated the proportion of disease-specific burden attributable to each dietary risk factor”
- “analysis included the intake of each dietary factor, the effect size of the dietary factor on disease endpoint, and the level of intake associated with the lowest risk of mortality”
- “... we calculated the number of deaths and DALYs attributable to diet for each disease outcome”
- “In 2017, 11 million (95% uncertainty interval [UI] 10–12) deaths and 255 million (234–274) DALYs were attributable to dietary risk factors.”
- “High intake of sodium (3 million [1–5] deaths and 70 million [34–118] DALYs), low intake of whole grains (3 million [2–4] deaths and 82 million [59–109] DALYs), and low intake of fruits (2 million [1–4] deaths and 65 million [41–92] DALYs) were the leading dietary risk factors”

… diet

Figure 1



dietary intake (age-standardized)

colours = regions

one panel per food

... diet

Figure 1

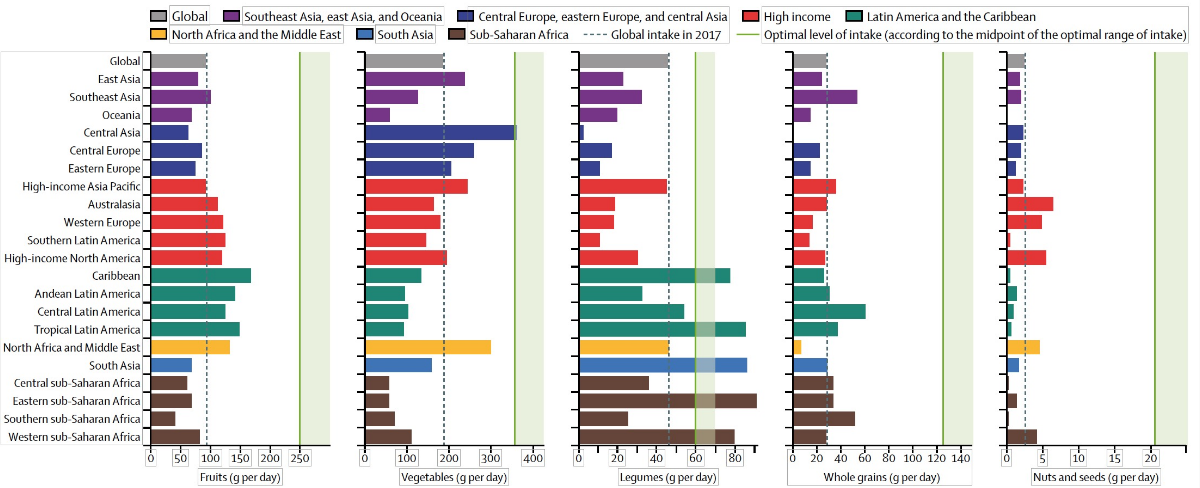


Figure 3

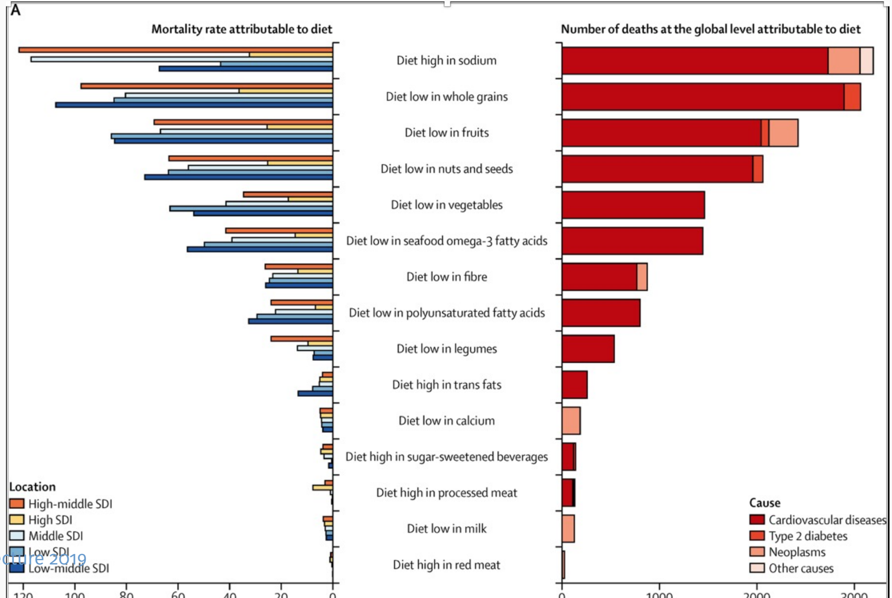
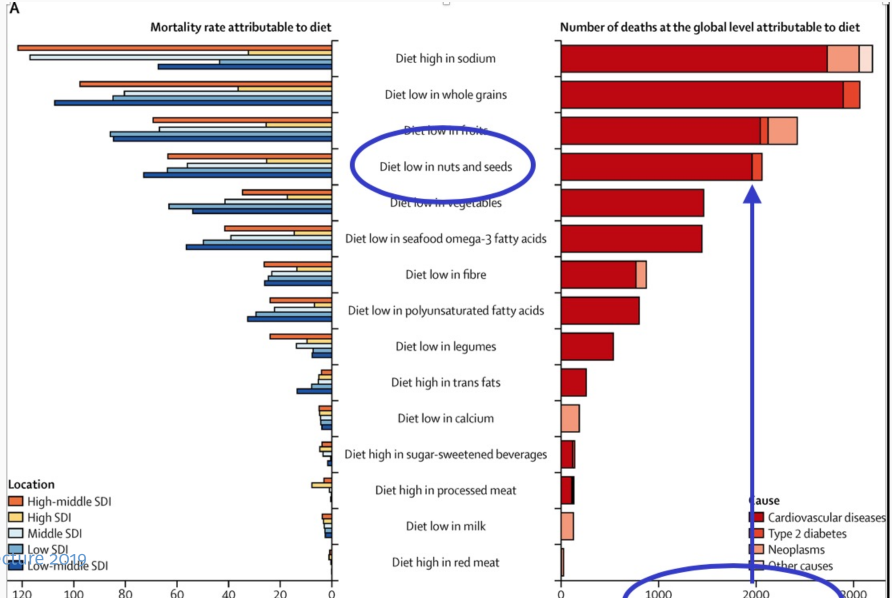

A
Mortality rate attributable to diet
Number of deaths at the global level attributable to diet
Diet high in sodium
Diet low in whole grains
Diet low in fruits
Diet low in nuts and seeds
Diet low in vegetables
Diet low in seafood omega-3 fatty acids
Diet low in fibre
Diet low in polyunsaturated fatty acids
Diet low in legumes
Diet high in trans fats
Diet low in calcium
Diet high in sugar-sweetened beverages
Diet high in processed meat
Diet low in milk
Diet high in red meat
Location
High-middle SDI
High SDI
Middle SDI
Low SDI
Low-middle SDI
Cause
Cardiovascular diseases
Type 2 diabetes
Neoplasms
Other causes
120
100
80
60
40
20
0
0
1000
2000
3000

Figure 3


A
Mortality rate attributable to diet
Number of deaths at the global level attributable to diet
Diet high in sodium
Diet low in whole grains
Diet low in fruits
Diet low in nuts and seeds
Diet low in vegetables
Diet low in seafood omega-3 fatty acids
Diet low in fibre
Diet low in polyunsaturated fatty acids
Diet low in legumes
Diet high in trans fats
Diet low in calcium
Diet high in sugar-sweetened beverages
Diet high in processed meat
Diet low in milk
Diet high in red meat
Location
High-middle SDI
High SDI
Middle SDI
Low SDI
Low-middle SDI
Cause
Cardiovascular diseases
Type 2 diabetes
Neoplasms
Other causes
120
100
80
60
40
20
0
0
1000
2000
3000

Health Effects of Dietary Risks in 195 Countries: Findings from the Global Burden of Diseases Study 2017

Supplementary appendix

Table of contents

$$PAF_{asct} = \frac{\int_l^u RR_{as}(x)P_{asct}(x)dx - RR_{as}(TMREL)}{\int_l^u RR_{as}(x)P_{asct}(x)dx}$$

$$Total\ attributable\ deaths_{asct} = \sum_{o=1}^{w} Death_{oasct}\, PAF_{oasct}$$

Assuming no correlation between dietary factors, we estimated the PAFs and number of deaths for the overall effect of all dietary factors relevant to each outcome based on the following equations:

$$PAF_{oasct} = 1 - \prod_{i}^{n}(1 - PAF_{ioasct})$$

$$Total\ attributable\ deaths_{asct} = \sum_{o=1}^{w} Death_{oasct}\, PAF_{oasct}$$

Supplemental Table 2. Citation of the epidemiological studies used to evaluate the causal relationship between dietary risk-outcome pairs in Supplemental Table 1.

Risk	Outcome	Citation/Note
Diet high in processed meat	Colon and rectum cancer	World Cancer Research Fund, American Institute for Cancer Research, Imperial College London. WCRF/AICR Systematic Literature Review Continuous Update Project Report: The Associations between Food, Nutrition and Physical Activity and the Risk of Colorectal Cancer. Oct 2010.
Diet high in processed meat	Ischaemic heart disease	Micha R, Wallace SK, Mozaffarian D. Red and processed meat consumption and risk of incident coronary heart disease, stroke, and diabetes mellitus: a systematic review and meta-analysis. Circulation 2010; 121: 2271–83.
Diet high in processed meat	Diabetes mellitus	Pan A, Sun Q, Bernstein AM, et al. Red meat consumption and risk of type 2 diabetes: 3 cohorts of US adults and an updated meta-analysis. Am J Clin Nutr 2011; 94: 1088–96.
Diet high in sugar-sweetened beverages	Diabetes mellitus	Imamura F, O'Connor L, Ye Z, *et al.* Consumption of sugar sweetened beverages, artificially sweetened beverages, and fruit juice and incidence of type 2 diabetes: systematic review, meta-analysis, and estimation of population attributable fraction. *BMJ* 2015; 351: h3576.
Diet high in sugar-sweetened beverages	Ischaemic heart disease	Xi B, Huang Y, Reilly KH, et al. Sugar-sweetened beverages and risk of hypertension and CVD: a dose-response meta-analysis. Br J Nutr 2015; 113: 709–17.
Diet low fibre	Colon and rectum cancer	World Cancer Research Fund, American Institute for Cancer Research, Imperial College London. WCRF/AICR Systematic Literature Review Continuous Update Project Report: The Associations between Food, Nutrition and Physical Activity and the Risk of Colorectal Cancer. Oct 2010.
Diet low fibre	Ischaemic heart disease	Threapleton DE, Greenwood DC, Evans CE, et al. Dietary fibre intake and risk of cardiovascular disease: systematic review and meta-analysis. BMJ (Clinical research ed) 2013; 347: f6879.
Diet low in calcium	Colon and rectum cancer	World Cancer Research Fund, American Institute for Cancer Research, Imperial College London. WCRF/AICR Systematic Literature Review Continuous Update Project Report: The Associations between Food, Nutrition and Physical Activity and the Risk of Colorectal Cancer. Oct 2010.
Diet low in seafood omega-3 fats	Ischaemic heart disease	Chowdhury R, Stevens S, Gorman D, et al. Association between fish consumption, long chain omega 3 fatty acids, and risk of cerebrovascular disease: systematic review and meta-analysis. BMJ (Clinical research ed) 2012; 345: e6698.
Diet low in polyunsaturated fats	Ischaemic heart disease	Farvid MS, Ding M, Pan A, et al. Dietary linoleic acid and risk of coronary heart disease: a systematic review and meta-analysis of prospective cohort studies. Circulation 2014; 130: 1568–78.
Diet low in polyunsaturated fats	Ischaemic heart disease	Mozaffarian D, Micha R, Wallace S. Effects on coronary heart disease of increasing polyunsaturated fat in place of saturated fat: a systematic review and meta-analysis of randomized controlled trials. PLoS Med 2010; 7: e1000252.
Diet high in trans fats	Ischaemic heart disease	Mozaffarian D, Clarke R. Quantitative effects on cardiovascular risk factors and coronary heart disease risk of replacing partially hydrogenated vegetable oils with other fats and oils. Eur J Clin Nutr. 2009; 63(Suppl 2): S22-33.
Diet high in trans fats	Ischaemic heart disease	http://www.bmj.com/content/bmj/suppl/2015/08/11/bmj.h3978.DC1/sour025275.ww2_default.pdf; pg. 44

iet high in trans fats	Ischaemic heart disease	Mozaffarian D, Clarke R. Quantitative effects on cardiovascular risk factors and coronary heart disease risk of replacing partially hydrogenated vegetable oils with other fats and oils. Eur J Clin Nutr. 2009; 63(Suppl 2): S22-:
iet high in trans fats	Ischaemic heart disease	http://www.bmj.com/content/bmj/suppl/2015/08/11/bmj.h3978.DC1/sour025275.ww2_default.pdf; pg. 44
iet high in sodium and high ystolic blood pressure	n/a	Aburto NJ, Ziolkovska A, Hooper L, Elliott P, Cappuccio FP, Meerpohl JJ. Effect of lower sodium intake on hea systematic review and meta-analyses. BMJ 2013; 346: f1326.
iet high in sodium	Stomach cancer	World Cancer Research Fund, American Institute for Cancer Research. Food, Nutrition, Physical Activity, and th Prevention of Cancer: a Global Perspective. Washington DC: AICR, 2007.
iet high in sodium	Stomach cancer	D'Elia, Lanfranco, Giovanni Rossi, Renato Ippolito, Francesco P. Cappuccio, and Pasquale Strazzullo. 2012. "Habitual Salt Intake and Risk of Gastric Cancer: A Meta-Analysis of Prospective Studies." Clinical Nutrition31 489–98. doi:10.1016/j.clnu.2012.01.003.
iet low in nuts and seeds	Ischaemic heart disease and diabetes mellitus	Experimental evidence on the relationship of nuts with ischaemic heart disease and diabetes mellitus come from t PREDIMED trial; a randomized trial consisting of three arms: a Mediterranean diet with extra-virgin olive oil, a Mediterranean diet with nuts, and a control diet. Given that the intake of dietary factors other than nuts changed i intervention arms of this trial, the observed effect might be fully attributable to nuts.
iet high in sodium	Cardiovascular diseases	Evidence on the direct effect of sodium on cardiovascular disease mainly comes from prospective cohort studies. Considering that, in GBD, we have only evaluated the effect of sodium mediated through systolic blood pressure did not present epidemiologic evidence on the direct effect of sodium on cardiovascular disease in this table. Evic on the effect of sodium on systolic blood pressure mostly comes from randomized controlled trials. While some c studies evaluated the relationship between sodium and systolic blood pressure, we did not identify a systematic evaluation of these studies.

The PREDIMED trial

PREDIMED trial

predimed trial - Google Search

predimed trial **retraction**

predimed trial **nejm**

predimed trial **mediterranean diet**

predimed plus trial

… PREDIMED trial

Estruch et al. NEJM 2013, 2018

The NEW ENGLAND JOURNAL of MEDICINE

CORRESPONDENCE

Retraction and Republication: Primary Prevention of Cardiovascular Disease with a Mediterranean Diet. N Engl J Med 2013;368:1279-90.

... PREDIMED trial

Agarwal & Ionnides BMJ 2019

thebmj

BMJ 2019;364:l341 doi: 10.1136/bmj.l341 (Published 7 February 2019)

ANALYSIS

Check for updates

PREDIMED trial of Mediterranean diet: retracted, republished, still trusted?

Arnav Agarwal and **John P A Ioannidis** consider what we can learn from the retraction and republication of an influential trial of Mediterranean diet

Arnav Agarwal *research associate* [1], John P A Ioannidis *professor* [2]

… diet and mortality

- 15 dietary risk factors
- consumption: nutrition surveys supplemented with sales data
- measurement of consumption: 24h dietary recall if available
- "spatiotemporal Gaussian process regression to estimate mean intake by age, sex, country and year"
- "for each diet-disease pair … published meta-analyses of prospective observational studies .. to estimate the relative risk of mortality …"
- relative risk converted to population attributable fraction
- then converted to mortality estimates

re nuts and seeds: PREDIMED – risk reduction of approximately 3 major cardiovascular events per 1000 person-years among adults 55 – 80 years old at high cv risk, in Spain

The NEW ENGLAND JOURNAL of MEDICINE ⟶ observational studies ⟶ THE LANCET ⟶ npr

All coverage

Radio Canada International - ENGLISH

Poor diet responsible for 1 in 5 deaths: study

5 days ago

Ahram Online

One in five deaths worldwide linked to unhealthy diet - Health - Life & Style

5 days ago

EurekAlert

The Lancet: Globally, 1 in 5 deaths are associated with poor diet

6 days ago

Deutsche Welle

Poor eating habits killing millions globally, study says

5 days ago · International

The Times of Israel

Israel has lowest rate of diet-related deaths in the world, major study finds

5 days ago · International

Mirror Online

Bangers and burgers 'cause 4,000 DEATHS a year' study claims

6 days ago

American Council on Science and Health

Can We Eat Our Way to Health (Or at Least Avoid Dying)?

5 days ago

Daijiworld.com

Poor diet causes hundreds of deaths in India

5 days ago

Yahoo News

The diet mistakes that could be killing us

5 days ago

The National

Death by diet more likely than by smoking, drugs or blood pressure, finds global study

5 days ago · International

The Week UK

The countries with the healthiest and unhealthiest diets revealed

5 days ago

Daily Mail

Poor diet kills more than smoking and hypertension, Bill Gates study reveals

6 days ago

Bloomberg

It Isn't Just Meat That's Killing You

5 days ago

National Post

Poor diets heavy in salt and sugar kill one in five people, global study finds

5 days ago

Lab Manager Magazine

One in five Deaths Associated with Poor Diet Globally

5 days ago

The Straits Times

1 in 5 deaths globally due to poor diet; study warns eating too much sugar, salt and meat killing millions

DQkFTRWdvSUwyMHZNRGx1YlY4U0FtVnVHZ0pEUVNnQVAB?hl=en-CA&gl=CA&ceid=CA%3Aen

Global NEWS | Toronto Change Location

TIME

UK DAILY

CNN

What we aren't eating is killing us, global study finds

By Sandee LaMotte, CNN

Updated 7:40 PM ET, Wed April 3, 2019

One in five people are eating themselves to an earl death: Global study

Ashley May, USA TODAY Published 9:04 p.m. ET April 3, 2019 | Updated 9:57 p.m. ET April 3, 2019

THE TIMES OF ISRAEL

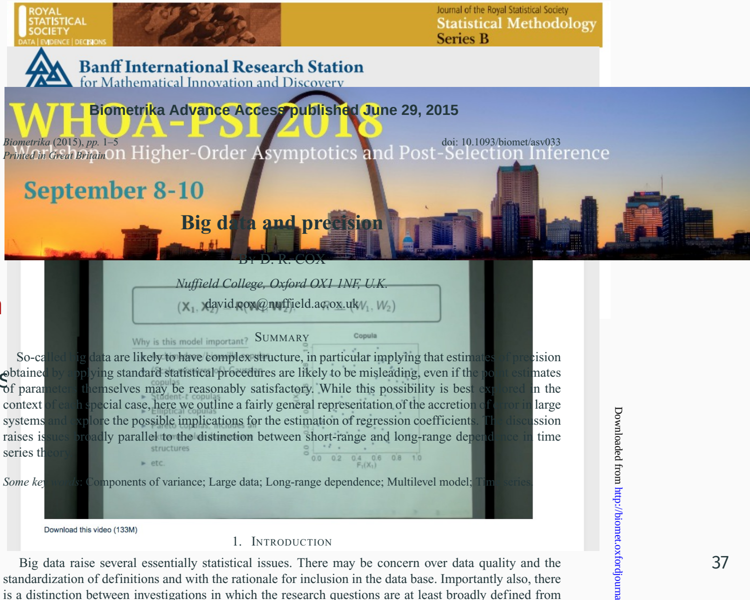
Israel has lowest rate of diet-related deaths in the world, major study finds

Analysis published in The Lancet finds fruit, vegetables, nuts and seeds are instrumental in avoiding deadly diseases

Statistical theory

Statistical theory

- causality
- data on networks
- multivariate extremes
- quantile regression
- high-dimensional inference
- model selection
- sparsity
- inference after model selection
- multivariate responses
- nonparametric, robust methods
- foundations
- ...



Statistical theory

- causality
- data on networks
- multivariate extremes
- quantile regression
- **high-dimensional inference**
- model selection
- sparsity
- inference after model selection
- multivariate responses
- nonparametric, robust methods
- **foundations**
- ...

Foundations

- how to get from data to conclusions
- with generalizable strategies
- what principles do we use to develop these strategies
- how are these strategies to be evaluated — efficiency, precision
- probability to describe physical haphazard variability — frequentist

  subject to empirical validation
- probability to describe the uncertainty of knowledge — Bayesian

  degree of belief

Bayesian, Fiducial, and Frequentist (BFF) Conferences

APPLY

***** Deadline for applications for this workshop is March 20, 2019 *****

Applications received after March 20th are subject to availability.

Location

This workshop will be held at Penn Pavilion on the campus of Duke University.

Role of Probability

R & Cox 2013

- probability to describe physical haphazard variability — frequentist
    - probabilities represent features of the "real" world in somewhat idealized form
    - subject to empirical test and improvement
- probability to describe the uncertainty of knowledge — Bayesian
    - measures rational or "impersonal" degree of belief,

      or — Jeffreys, 1939,1961
    - measures a particular person's degree of belief

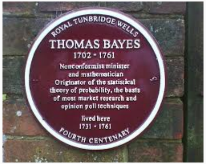
      F.P. Ramsey, 1926
    - linked to personal decision making

- confidence intervals or *p*-values refer to empirical probabilities

"[7 – 11]"

- inference is assessed by behaviour of the procedure under hypothetical repetition

- the Bayesian approach to inference describes uncertainty of knowledge
- this can be interpreted empirically by appeal to a notion of calibration



Statistical theory

- causality
- data on networks
- multivariate extremes
- quantile regression
- **high-dimensional inference**
- model selection
- sparsity
- inference after model selection
- multivariate responses
- nonparametric, robust methods
- **foundations**
- ...

High-dimensional asymptotic theory

- $f(y;\theta)$, $y \in \mathbb{R}^n$, $\theta \in \mathbb{R}^p$ — $y_1, \ldots, y_n$ independent
- classical: p fixed, $n \to \infty$ — $\sqrt{n}(\hat{\theta} - \theta)V^{-1/2} \xrightarrow{d} N_p(0, I)$
- semi-classical: $p_n/n \to 0$, or $p_n^{3/2}/n \to 0$ — Huber 70, Portnoy 80s; Sartori 90s, Lunardon '18, ...
- moderate dimension $p_n/n \to \kappa \in (0,1)$ — Candes '17, Lei/Bickel/El Karaoui '18, Coolen et al. '19
- high dimension $p_n \sim n^\alpha$ — $p > n$
- ultra-high dimension $p_n \sim e^n$

Moderate dimension

Lei, El-Karoui, Bickel 18

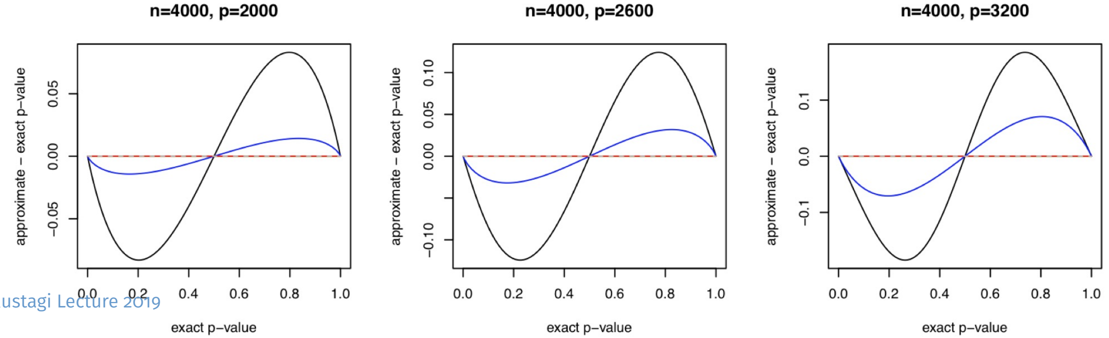
- $\hat{\beta} = \arg\min \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - x_i^{\mathrm{T}} \beta)$ *M*-estimator
- coordinate-wise asymptotic normality

$$\max_j d_{TV} \left\{ \mathcal{L} \left(\frac{\hat{\beta}_j - \mathrm{E}(\hat{\beta}_j)}{\sqrt{\mathrm{var}(\hat{\beta}_j)}} \right), N(0,1) \right\} = o(1)$$

- "For instance for least-squares, standard degrees of freedom adjustments effectively take care of many dimensionality-related problems"
- in least squares, 'standard degrees of freedom adjustments' can be derived using higher order asymptotics for p fixed
- e.g. $n = 50, p = 30$ or $n = 500, p = 300$ moderate or classical?

- normal theory linear regression $y = X\beta + \sigma\epsilon,\ \epsilon \sim N(0, 1)$
- exact test available based on t-statistic $t = (\hat{\beta}_j - \beta_j)/v_{jj}$ least squares = mle
- all likelihood quantities are functions of t, n and p
- modified log-likelihood root, derived from higher order asymptotics depends only on $t, n - p$ r^* depends on $t, n, n - p$

Plot of differences between approximate and exact p-values for one-sided alternative against the true p-value:



logistic regression

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

- if the MLE exists, then

$p/n \to \kappa \in (0, 1)$

$$\frac{1}{p}\sum_{j=1}^{p}(\hat\beta_j - a_*\beta_j) \longrightarrow 0; \qquad \frac{1}{p}\sum_{j=1}^{p}(\hat\beta_j - a_*\beta_j)^2 \longrightarrow \sigma_*^2$$

- Likelihood Ratio Test for $H : \beta_j = 0$ has scaled χ^2

$$w(\beta_j) = 2\{\ell(\hat\beta) - \ell(\tilde\beta_{(j)})\} \xrightarrow{d} \frac{\kappa\sigma_*^2}{\lambda_*}\chi_1^2$$

- $(a_*, \sigma_*, \lambda_*)$ characterized as the solution of three equations
- e.g. $n = 50, p = 30$ or $n = 500, p = 300$

moderate or classical?

logistic regression

$$\log \frac{p_i}{1-p_i} = x_i^{\mathrm{T}}\beta, \quad y_i \sim \text{Bernoulli}(p_i)$$

- if the MLE exists, then $p/n \to \kappa \in (0,1)$

$$\frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - a_*\beta) \longrightarrow 0; \qquad \frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - a_*\beta)^2 \longrightarrow \sigma_*^2$$

- in logistic regression, change the score equation a little maximum likelihood estimate always exists Firth 93; Kosmidis/Firth 09
- usual limit theory seems to be fine with large p Sartori; Lunardon 18

Statistics and data science

Annual Review of Statistics and Its Application

Volume 1, 2014

Select All

For selected items: Please select

What Is Statistics?

Stephen E. Fienberg

Full Text HTML | Download PDF

About This 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.

What is Statistics?

Fienberg, 2014

- an art and a science
- a set of tools and methods, ... [and] computation
- a way of thinking about data
- what statisticians do and how they think about what they do
- interacts vigorously with astronomy, biology, engineering, geology, medicine and public health, and many social sciences, including political science, law, sociology, psychology, anthropology, archeology, history, ... ”
- ‘A distinguishing feature of the statistics profession, and the methodology it develops, is the focus on a set of cautious principles for drawing scientific conclusions from data.’ Lindsay 2004

Statistical science

Cox & Donnelly 2011

- start with a scientific question
- assess how data could shed light on this
- plan data collection
- consider of sources of variation and how careful planning can minimize their impact

- develop strategies for data analysis: modelling, computation, methods of analysis
- assess the properties of the methods and their impact on the question at hand

- communicate the results: accurately — but not pessimistically
- visualization strategies, conveyance of uncertainties

What is Data Science?

- a course?

- a set of courses?

- a job?

- **a new field of research?**
- **a collaboration?**

… what is data science?

- short-hand for “lots of data”, “complicated data”, “data of uncertain provenance”

- an undergraduate or post-graduate program of training
- a job description

The Alan Turing Institute

- a new multi-disciplinary field of study ??
- combining mathematics, statistics, computer science, domain science

- increased emphasis on privacy, fairness, communication, visualization, impact on policy, workflow and reproducible research

Blake & Olhede, 2016

Data science workflow

data acquisition

data preservation

Making data trustable and usable
Management of data
Modelling and Analysis
Reproducibility
Dissemination and Visualization

Security and privacy

Ethics, policy and social impact

… data science workflow

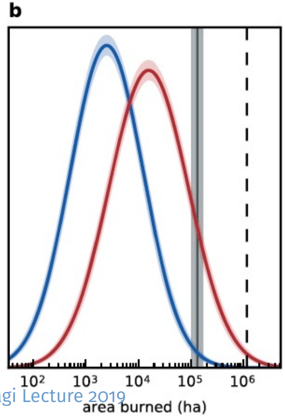
Making data trustable and usable Management of data	provenance, sampling, cleaning, digitizing size, speed, accessibility
Modelling and Analysis Reproducibility Dissemination and Visualization	interpretable vs predictive methods accessibility and impact data, code, output

mathematics statistics computer science domain expertise

Security and privacy	disclosure limitation, anonymization, encryption
Ethics, policy and social impact	fairness and transparency

Statistical Inference

- data $\rightarrow$ conclusions
- data $\rightarrow$ uncertainty about conclusions



“the annual burned areas of this same percentile are smaller by a factor of 7 - 11”

“the best-fit regression model was used, with shading indicating 90% confidence intervals”

Theory of Statistical Inference

- how to get from data to conclusions
- with generalizable strategies
- what principles do we use to develop these strategies
- how are these strategies to be evaluated

  efficiency, precision
- a long history of the subject; using probability to both develop statistical methods and to evaluate their performance

  Bayes, Laplace, Gauss; Student, Fisher, Neyman, Pearson, Jeffreys, ...
- leading to confidence intervals, *p*-values, estimates and standard errors, etc.

In praise of small data

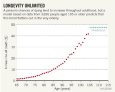
"you don't have as much data as you thought"

- correlation/dependence/heterogeneity/multiple scales

wildfires

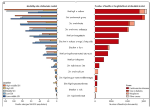
- rare events

extremes



- subgroup analyses/'data slices'

diet and mortality



- complex models/many parameters/high-dimensional inference

sparsity, new asymptotics

- ...

Thank you!

DEPARTMENT OF
STATISTICS

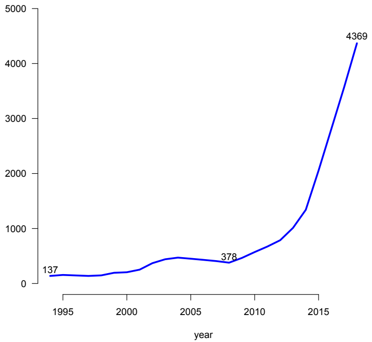
THE OHIO STATE UNIVERSITY

Collaborations

CANSSI INCASS

- Canadian Statistical Sciences Institute
- launched in 2012
- funded 2014–2021 by Natural Sciences and Engineering Research Council
- national scope, virtual institute
- **Collaborative Research Teams**
  multidisciplinary, multi-institution, statistical leadership, scientific engagement

… collaborations



Statistical Sciences

UNIVERSITY OF TORONTO

- statistical genetics
- spatial modelling
- machine learning (with CS)
- visualization (with CS)
- demography (with Sociology)
- astrostatistics (with A and A)
- cognitive neuroscience (with Pyschology)
- data science (with iSchool)
- financial insurance
- actuarial science
- teaching stream

In praise of small data