Planning of Studies

Statistical thinking for Big Data

Outline

- Hoerl et al. (2014) Applying statistical thinking to 'Big Data' problems. Wiley Interdisciplinary Review in Computational Statistics.
- Lazer et al. (2014) The parable of Google flu: traps in big data analysis. Science, 14 March
- Paul, et al. (2014) Twitter improves influenza forecasting. PLOS Current Outbreaks, 28 October
- Planning of studies: surveys, experiments, observational studies. Cox & Donnelly, Principles of Applied Statistics

Advanced Review

Applying statistical thinking to 'Big Data' problems

Roger W. Hoerl,[1]* Ronald D. Snee[2] and Richard D. De Veaux[3]

Much has been written recently about 'Big Data' and the new possibilities that mining this vast amount of data brings. It promises to help us understand or predict everything from the Higgs boson to what a customer might purchase next from Amazon. As with most new phenomena, it is hard to sift through the hype and promotion to understand what is actually true and what is actually useful. One implicit or even explicitly stated assumption in much of the Big Data literature is that statistical thinking fundamentals are no longer relevant in the petabyte age. However, we believe just the opposite. Fundamentals of good modeling and statistical thinking are crucial for the success of Big Data projects. Sound statistical practices, such as ensuring high-quality data, incorporating sound domain (subject matter) knowledge, and developing an overall strategy or plan of attack for large modeling problems, are even more important for Big Data problems than small data problems.

How to cite this article:
WIREs Comput Stat 2014, 6:222–232. doi: 10.1002/wics.1306

Keywords: data mining; statistical engineering; analytics; machine learning

Applying Statistical Thinking to 'Big Data' problems

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

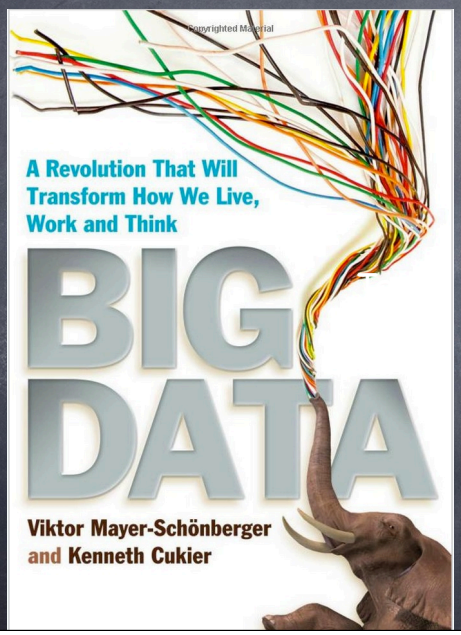
By Chris Anderson 06.23.08

Illustration: Marian Bantjes

THE PETABYTE AGE:
Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the era of

"All models are wrong, but some are useful."

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies



The world is `shifting from causation to correlation'

Some warnings

- Duke Genomics Center: "forensic bio-informatics" (Baggerly & Coombes, AOAS '09)
- Lehman Brothers analytic models to predict defaults "implicit assumption that the future would behave like the past"
- Amazon automatic pricing: The Making of a Fly $1,730,045 -> $23,698,656 -> $106
- Netflix Prize: "the additional accuracy gains … did not seem to justify the engineering effort"
- (Google flu trends -- Lazer et al.)

Statistical thinking

- All work occurs in a system of interconnected processes
- Variation exists in all processes
- Understanding and reducing variation are keys to success
- Building blocks:
  - Problem statement; Process understanding; Analysis strategy; Sources of variation; Quality of data; Domain knowledge; Modeling process; Sequential approach

Data quality

- 80% of data science is 'data wrangling'
- Missing data, outliers, mis-measurements, coding errors, ...
- Keeping track of data manipulations -- 99 for NA, e.g.
- Automated 'cleaning' -- filters should be well understood
- Observational studies vs randomized experiments

Domain knowledge

- Actionable conclusions
- Example in nutrition: a designed experiment that shows eating a particular food consistently leads to elevated blood pressure could be 'actionable'
- An observational study that indicated a correlation would be of interest, but not likely actionable

Analysis strategies

- Prediction on a test set is often not adequate for understanding
- Test sets are typically subsamples of original data set; can't incorporate changes over time
- Model validation requires ongoing evaluation, including applicati
- on to new data sets collected under different conditions
- What scientific questions can this data set help to answer?
- As opposed to 'what is the best model for this data?'
- "A model is a how, not a what"

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

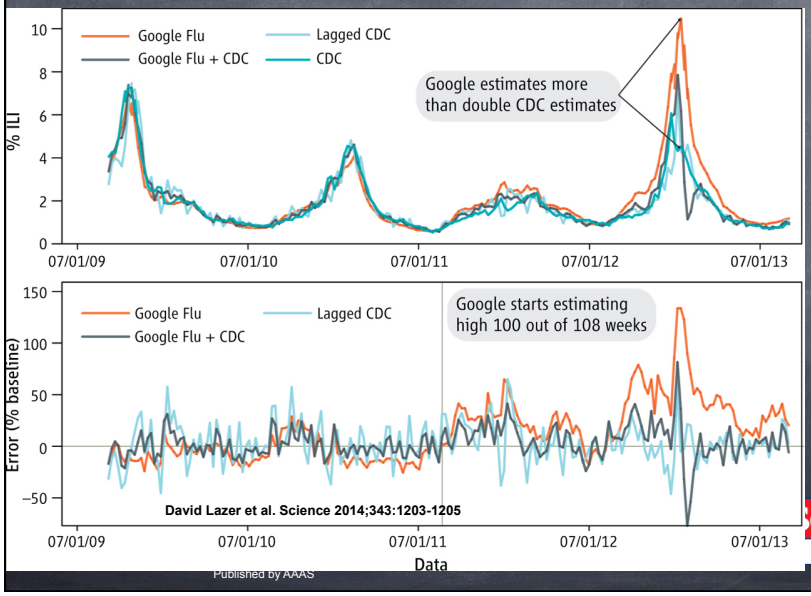
The problems we identify are not limited to GFT. Research on whether search or social media can predict *x* has become commonplace (*5*–*7*) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories (*8*). We explore two ability and dependencies among data (*12*). The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated in 2009, the comparative value of the algorithm as a stand-alone flu monitor is questionable. A study in 2010 demonstrated that GFT accuracy was not much better than a fairly simple projection forward using already available (typically on a 2-week lag)

GFT overestimation.GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%.



David Lazer et al. Science 2014;343:1203-1205

Published by AAAS

Lazer et al

- initial version (2009) found best matches among 50m search terms, to fit 1152 data points
- part flu detector, part winter detector
- updated in 2009; but persistently over-estimated flu prevalence
- errors not randomly distributed: temporal autocorrelation + seasonality
- errors large than those using 2-week lagged data from CDC
- but errors in both can be reduced by combining

Lazer et al

Reasons

- all empirical research stands on a foundation of measurement
- is the instrumentation capturing the theoretical construct of interest?
- is measurement stable and comparable across cases and time?
- are measurement errors systematic?
- in improving its service to customers, Google is also changing the data-generating process

Lazer et al

Lessons

- failure of transparency and replicability
- use Big Data to understand the unknown, (e.g. local level predictions)
- the (search) algorithms are continually changing, both for science and for the business model
- size isn't everything! perhaps it's time we focussed on an "all data revolution"

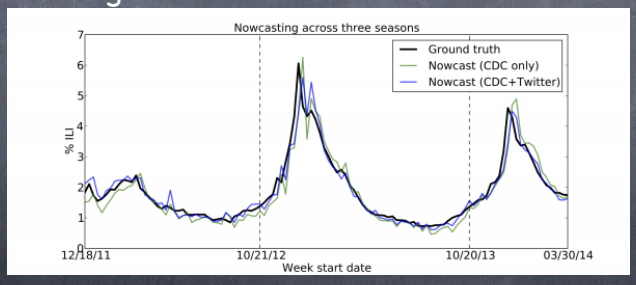
Significance Magazine, from Financial Times,

Harford on GFT

- Google engineers weren't trying to figure out what caused what. They were merely finding statistical patter
- figuring out what is correlated with what is much cheaper and easier
- possible the news had many stories about flu in Dec 2012 that provoked internet searches
- or that Google's own search algorithm moved the goalposts

Updates

- Oct 31, 2014: from Google "Google Flu Trends gets a brand new engine"
- Oct 28, 2014 "Twitter improves influenza forecasting"

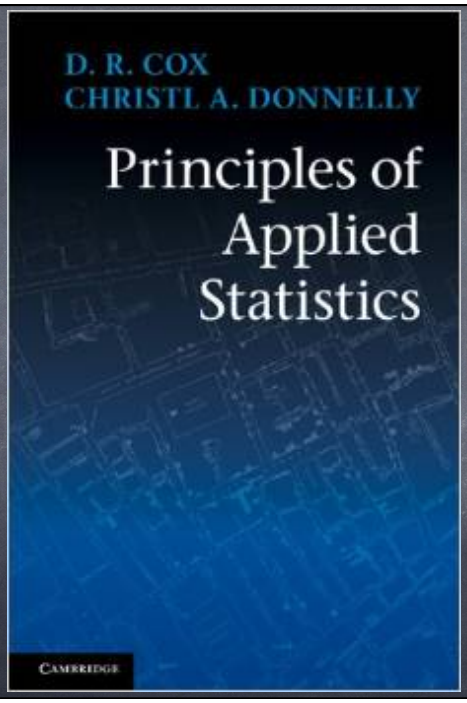


Paul MJ, Dredze M, Broniatowski D. Twitter Improves Influenza Forecasting.
PLOS Currents Outbreaks. 2014 Oct 28. Edition 1.
doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.

Updates

3:30-4:30	**Yulia Gel,** University of Waterloo *The Role of Modern Social Media Data in Surveillance and Prediction of Infectious Diseases: from Time Series to Networks*
4:30	Cash Bar Reception

- initialize our model by health-related tweets, conditionally on the available information about location and socio-demographics of a Twitter user
- simulate possible scenarios of infection spread via the population network
- but, (Harford): "Twitter users are not representative of the population as a whole. (According to the Pew research project, in 2013, US-based Twitter users were disporportionately young, urban or suburban, and black.)"
- and, "Statisticians are scrambling to develop new methods to seize the opportunity of big data. Such new methods are essential but they will work by building on the old statistical lessons, not by ignoring them"



Planning of Studies

Cox & Donnelly 2011

General Concepts Chapter 1

- the need for statistical analysis typically arises from the presence of unexplained and haphazard variation; a combination of natural variability and measurement or other error (e.g. blood pressure)
- the "ideal sequence":
  - formulation of research questions,
  - search for relevant data;
  - design and implementation of investigations to obtain appropriate data;
  - analysis of the data;
  - interpretation of the results
- an extreme departure: a large body of data become available
- various methods may uncover possible relationships of interest
- conclusions likely to be tentative and in need of independent confirmation

General Concepts Chapter 1

- it is essential to be clear at the design stage broadly how the data are to be analysed
- because conclusions are publicly more convincing
- reduces the chance that the data cannot be satisfactorily analysed
- it will often be unrealistic and potentially dangerous to follow an initial plan unswervingly
- in major studies with long time frames it will be wise to list the possible data configurations that might arise

What are the principles of applied statistics?

Chapter 1

- formulation and clarification of focused research questions of subject-matter importance
- design of individual investigations and sequences of investigations that produce secure answers and open new possibilities
- production of effective and reliable measurement procedures
- development of simple (or if-needed not-so-simple) methods of analysis
- effective presentation of conclusions
- structuring analyses to facilitate their interpretation in subject-matter terms

Design of Studies

Chapter 2

Common objectives

- to avoid systematic error, that is distortion in the conclusions arising from sources that do not cancel out in the long run
- to reduce the non-systematic (random) error to a reasonable level by replication and other techniques
- to estimate realistically the likely uncertainty in the final conclusions
- to ensure that the scale of effort is appropriate

Design of Studies — Chapter 2

Types of study

- experiment: investigator has complete control over treatment assignment
- observational study
  - sample surveys
  - retrospective and prospective studies
  - secondary analysis of available data
- census
- meta-analysis: statistical assessment o fa collection of studies on the same topic

Design of Studies — Chapter 2

Avoidance of systematic error

- distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run
- can arise through the measuring process, or from the spatial or temporal arrangement of units
- which can often be avoided by design, or adjustment in the analysis
- can arise by the entry of personal judgment into some aspect of the data collection or analysis
- can be avoided by randomization and blinding

Design of Studies Chapter 2

Control and estimation of random error

- statistical analysis is particularly important in investigations in which haphazard variation plays an important role
- we can lessen the impact of haphazard variation by:
  - use of artificially uniform material
  - arranging that the comparisons of main interest compare like with like
  - inclusion of background variables
  - replication
- these may impact generalizability, so depend on the context

Design of Studies Chapter 2

Scale of effort

- how big should my sample be?
- key observation: $\mathrm{var}(\bar{y}_1 - \bar{y}_2) = 2\sigma^2/m$
- if the standard error of the most important comparison is to be less than c, then need $m \approx 2\sigma^2/c^2$
- c will be to some extent determined by the magnitude of differences of interest
- note much simpler than considering power of a test

Design of Studies — Chapter 2

Factorial Principle

- it is often useful, even beneficial, to investigate several different aspects of a system simultaneously
- the classical experiment of this type is a factorial experiment

Block	b	ac	1	c	abc	a	bc	ab
Block 2	abc	a	ac	ab	1	bc	b	c
Block 3	a	c	ac	bc	1	b	abc	ab

Sample Surveys — Chapter 3

- target population
- goal is to estimate one or more features of the population
- example: election polls
- example: Labour Force Survey
- example: National Household Survey

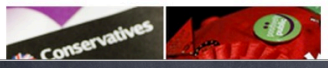
ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

ASA
AMERICAN STATISTICAL ASSOCIATION

significance

The political pollsters' report card: the statisticians' verdict

Written by Deirdre Toher & Robin Evans on 11 March 2015. Posted in Politics



Election season is now well underway in the UK and every scandal, botched interview and policy announcement is being

YouGov has recruited a huge online panel of 360,000 adults to take its surveys, but…

YouGov (pollster for *The Sun* and *The Sunday Times*)

YouGov has recruited a huge online panel of 360,000 adults to take its surveys, but this group is to some extent self-selecting, and not representative of the entire population of voters. When it launches a new poll, YouGov invites a much smaller sub-sample of the panel to take part, and this group is chosen to be representative of people based on age, gender, social class and newspaper readership.

Afterwards the responses are checked against various other surveys and re-weighted if necessary. Respondents are encouraged to take part by a small cash incentive - this is designed to increase the response rates and make sure that not all participants are political animals. A possible disadvantage is that it might increase the response rate differently among different groups. For example, if someone doesn't respond because they're busy it might not make any difference if you offer them money.

Populus

Populus also uses online polling with similar methodology to YouGov and a panel of around 100,000 people, but the outcome appears to be systematically different. During January, Populus was consistently recording around a 2%-3% Labour lead whereas YouGov suggested a dead heat. This is presumably down to a combination of the different make-up of their panels and the weighting they use. This illustrates the limits of all polling - there isn't much point in getting a massive sample if it won't be representative.

Like most pollsters, Populus and YouGov both prompt voters with a list of parties which includes UKIP and the SNP or Plaid Cymru (in Scotland and Wales respectively), but not the Green Party. This may cause a slight reduction in the proportion of people who say they vote Green, but it is unlikely to be dramatic.

Sample Surveys Chapter 3

Key concepts:

- target population
- sampling frame – implicit or explicit list of population members
- randomized selection of samples from the population
  - to avoid systematic error
  - as a basis for inference about the population
- SRS: sample $y_1, \ldots, y_n$ population $y_1, \ldots, y_N$
- Total in population estimated by

$$\frac{N}{n} \sum_{i=1}^{n} y_i = N\bar{y}$$

Sample Surveys Chapter 3

Improving precision

- stratification: population $N_1, \ldots, N_h$
- sample sizes: $n_1, \ldots, n_h$
- SRS within strata
- Total in population estimated by

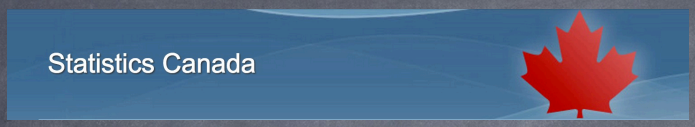
$$\sum_{h,i} \frac{N_h}{n_h} y_{hi}$$

- also possible to adjust sample means using auxiliary variables, either by matching the means, or by regression
- similar to randomized block design

Sample Surveys Chapter 3

More complex strategies

- multi-stage cluster sampling
- panel studies over time – rotating in and out
- sampling without a frame
  - genetics of rare human diseases – probands
  - hard to reach populations – snowball sampling
  - capture-recapture sampling
  - selective data collection



- This is a sample survey with a cross-sectional design.

The LFS uses a probability sample that is based on a stratified multi-stage design. Each province is divided into large geographic stratum. The first stage of sampling consists of selecting smaller geographic areas, called clusters, from within each stratum. The second stage of sampling consists of selecting dwellings from within each selected cluster.

The LFS uses a rotating panel sample design so that selected dwellings remain in the LFS sample for six consecutive months. Each month about 1/6th of the LFS sampled dwellings are in their first month of the survey, 1/6th are in their second month of the survey, and so on. One feature of the LFS sample design is that each of the six rotation groups can be used as a representative sample by itself.

Within selected dwellings, basic demographic information is collected for all household members. Labour force information is collected for all civilian household members who are aged 15 and over.

Recently, the monthly LFS sample size has been approximately 56,000 households, resulting in the collection of labour market information for approximately 100,000 individuals. It should be noted that the LFS sample size is subject to change from time to time in order to meet data quality or budget requirements.

Sample Surveys

Chapter 3

EDA of a Fire Weather Risk Index

Exploring an FWI Time Series

Data:

- A set of daily FWI series
- 1963 – 2004 fire seasons
- Sample of six weather stations

wxstn_id	wx_date	fwi	year	month	day	julian	doy
10100	1963-05-05	5	1963	5	5	1220	124
10100	1963-05-06	5	1963	5	6	1221	125
10100	1963-05-07	12	1963	5	7	1222	126
...							
10100	0011-11-20	0.90	2004	11	11	16386	315
10100	0011-12-20	0.39	2004	11	12	16387	316

Thursday January 15: Environmental Science

9:30-10:30	**Charmaine Dean**, Western University *Wildfire and Forest Disease Prediction to Inform Forest Management: Statistical Science Challenges*
10:30-11:00	Coffee break
11:00-12:00	**Doug Woolford**, Wilfrid Laurier University *Exploratory data analysis, visualization and modelling methods for large data in forest fire science*

New: Gdn & Economist

New study estimating number of dead in Iraq hotly contested

BY ESTANISLAO OZIEWICZ

More than 650,000 Iraqis have died as a result of the 2003 invasion and ensuing violence, according to a new study directed by a U.S. public-health expert.

But the results of the study, published in the on-line edition of the leading British medical journal The Lancet, are as controversial as an earlier survey by researchers from the Johns Hopkins Bloomberg School of Public Health that attributed nearly 100,000 deaths to the first 18 months of the war.

"We estimate that almost 655,000 people — 2.5 per cent of the population — have died in Iraq," the authors of the study conclude. "Although such death rates might be common in times of war, the combination of long duration and tens of millions of people affected has made this the deadliest international conflict of the 21st century."

U.S. President George W. Bush dismissed the report. "The methodology is pretty well discredited," he said yesterday. Similarly, Iraqi government spokesman Al-al-Dabbagh told Reuters. "These numbers are exaggerated and not precise."

Mr. Bush has previously put the number of Iraqi deaths at 30,000. He reaffirmed that number yesterday.

"I stand by the figure," he said. "Six hundred thousand or whatever they guessed at . . . it's not credible."

Even some less self-interested and partisan bodies are skeptical of the numbers, partly because they are many times higher than other apparently independent estimates. For example, Iraq Body Count, an Internet-based organization, put the death count at 48,693 yesterday.

'If there is surprise about the size of the figure, it has more to do with our existing death tolls. The conventional wisdom is based on shoddy information.'

Sarah Leah Whitson, New York-based Human Rights Watch

But Human Rights Watch defended the study whose lead author is Gilbert Burnham, co-director of the Centre for Refugee and Disaster Response at Johns Hopkins.

"If there is surprise about the size of the figure, it has more to do with our existing death tolls," said Sarah Leah Whitson, of New York-based Human Rights Watch. "The conventional wisdom is based on shoddy information."

The study was funded by the Massachusetts Institute of Technology.

The number of Iraqis who have died above the normal rates since March of 2003 includes deaths from all causes, including those due to a rise in certain diseases and illnesses, the study said.

Nearly 60 per cent of the dead were boys and men aged between 15 and 64. "Over the 40-month period of the study, approximately 31 per cent of households attributed the death of their household member to coalition forces," Dr. Burnham told reporters yesterday.

"Of the deaths, we found an increasing proportion were due to car bombs, but the majority were due to gunfire," he said.

Most violent deaths — 56 per cent — were due to gunshots. Air strikes, car bombs and other explosions accounted for 13 per cent to 14 per cent of violent deaths.

The researchers did not ask interviewees whether those they reported dead were civilians or combatants. Dr. Burnham said his study was unable to conclude which deaths were due to sectarian or criminal violence.

The figures are based on a May-June survey conducted by researchers from Johns Hopkins and Al-Mustansiriya University in Baghdad, of 1,849 households including 12,801 household members in 47 randomly selected sites across Iraq. They questioned inhabitants about births, deaths, and migrations. The death rate in Iraq rose to 13.5 per 1000 people per year from 5.5 per year before the invasion, they said.

Economist: The human cost of the war in Iraq -- "A statistical study claims that many more Iraqis have died than was thought"

The Lancet 2006 "Mortality after the 2003 invasion of Iraq: a cross-sectional study"

Science 2006 "Iraqi death estimates called too high: methods faulted"

Johnson et al 2008 J Peace Research
``Bias in epidemiological studies of conflict mortality''

Iraq body count: 48, 693

Burnham et al. (2006): 601,027 (427,000 -- 739,700)

NEJM (2008): 151,000 (104,000 -- 223,000)

- multi-stage cluster sampling
- choose a random cross street to the main street
- select a random household on the cross street to start the process
- interview that house and proceed to adjacent house until 40 houses have been surveyed

Experimental Design Chapter 3

- simplest case: randomized assignment of experimental units to treatments
- experimental unit: smallest subdivision of experimental material such that two different units may experience different treatments
- patient, school, village, patient/period, cage, animal, ...
- randomized assignment to avoid systematic bias and for concealment
- in principle, any observed association between treatment and outcome may be ascribed to the treatment
- since all else is 'equal'
- issue in human trials: noncompliance with treatment - a form of missing data

Experimental Design — Chapter 3

- randomized block designs
- experimental units grouped into homogeneous subsets
  - days, observers, subjects, litters, cages, ...
- treatments are randomized to units within subsets
- subsets of size two are the simplest case – analysis based on differences in responses eliminates block (pair) effects, since all else is 'equal'
- issue in human trials: noncompliance with treatment – a form of missing data
- a notional partition of effects

$$y_{tb} = \bar{y}_{..} + (\bar{y}_{t.} - \bar{y}_{..}) - (\bar{y}_{.b} - \bar{y}_{..}) + (y_{tb} - \bar{y}_{t.} - \bar{y}_{.b} + \bar{y}_{..})$$

Experimental Design — Chapter 3

- leading to a traditional analysis of variance table

Source	DF	SS
Mean	1	$\Sigma\bar{y}_{..}^2$
Blocks	$B-1$	$\Sigma(\bar{y}_{.b} - \bar{y}_{..})^2$
Treatments	$T-1$	$\Sigma(\bar{y}_{t.} - \bar{y}_{..})^2$
Residuals	$(B-1)(T-1)$	$\Sigma(y_{tb} - \bar{y}_{.b} - \bar{y}_{t.} + \bar{y}_{..})^2$

this does not depend on the particular nature of the response, nor on the balance in the experiment, nor on the probability structure

Observational studies Chapter 3

- cross-sectional study – each study individual observed at one time point
- information about the past may be collected at that time (recall bias)
- very difficult to make any causal interpretation, even if the data are collected over a long time period
- example: Canadian Survey on Disability
- sometimes independent cross-sectional samples are taken at several points in time, with the goal of identifying changes over time
- example: "Current and predicted prevalence of obesity in Canada" (CMAJ Open, March 3 2014)

Observational studies Chapter 3

- example: Canadian Survey on Disability

Sampling design and stratification

The sample design can be viewed as a three-phase design where the first two phases were for the selection of the NHS sample itself and the third phase was for the selection of the CSD sample. In the first phase, sample selection of the NHS itself corresponds to a systematic sample of approximately one dwelling out of three across Canada in non-remote areas (N1 regions). The second phase corresponds to subsampling of NHS non respondents in N1 regions, a new procedure that was put in place to mitigate the potential effect of non-response bias due to the higher non-response that resulted from the NHS. In this phase, a subsample of non-respondent dwellings was selected for non-response follow-up (NRFU).

In the third phase, the CSD sample was selected from the group of individuals who responded to the NHS (including the NRFU subsample) and reported activity limitations within the NHS. The sample was selected so that there would be a sufficiently large sample in each estimation domain.

CMAJ OPEN

Search for Keyword: Advanced Search

Home | Issues | Collections | Information for... | Alerts | About CMAJ Open | Help

Current and predicted prevalence of obesity in Canada: a trend analysis

Laurie K. Twells, PhD[1,2], Deborah M. Gregory, PhD[2,3], Jacinta Reddigan, MSc[2], William K. Midodzi, PhD[2]

Author Affiliations

Correspondence to:
Laurie K. Twells, ltwells@mun.ca

Abstract

Background The prevalence of obesity has increased over the past 3 decades,

« Previous | Next Article »
Table of Contents

This Article

doi: 10.9778/cmajo.20130016
cmajo March 3, 2014 vol. 2 no. 1 E18-E26

Abstract *Free*
Figures Only
» Full Text
Full Text (PDF)

Correction (vol. 2, p. E35)

Online Appendices

Classifications

Research

dr careers
CMA's career centre for physicians

Psychiatrist (Adult General) Locum
Radiologist Locums
Associate / Full Professor - Vice Chair, Education, Department of Medicine
Hospitalist
Chief/Medical Director of Emergency Medicine
Transfusion Medicine Specialist
Hospitalist/Unassigned

We calculated BMIs for adults aged 18 years and older who were not in long-term care using data from Canadian health surveys administered between 1985 and 2011. Calculation of the BMIs was based on self-reported heights and weights.

Observational studies Chapter 3

- Prospective observational study
- Study subjects identified, followed forward in time
- data often has longitudinal structure
- sometimes possible to mirror closely a randomized design
- example: Women's Health Initiative
- "a major 15-year research program to address the most common causes of death, disability and poor quality of life in postmenopausal women -- cardiovascular disease, cancer, and osteoporosis"

Observational studies — Chapter 3

- Retrospective observational study
- useful for rare events – prospective study may require waiting a long time, and accruing an unnecessarily large amount of information on non-cases
- start with cases, and for each case choose one or more controls, either at random, or matched on key features
- recall bias in determining explanatory variables
- called 'choice-based' sampling in econometrics

Observational studies — Chapter 3

Population

	$y=0$	$y=1$
$x=0$	π_{00}	π_{01}
$x=1$	π_{10}	π_{11}

Prospective study

	$y=0$	$y=1$
$x=0$	$\pi_{00}/(\pi_{00}+\pi_{01})$	$\pi_{01}/(\pi_{00}+\pi_{01})$
$x=1$	$\pi_{10}/(\pi_{10}+\pi_{11})$	$\pi_{11}/(\pi_{10}+\pi_{11})$

Retrospective study

	$y=0$	$y=1$
$x=0$	$\pi_{00}/(\pi_{00}+\pi_{10})$	$\pi_{01}/(\pi_{01}+\pi_{11})$
$x=1$	$\pi_{10}/(\pi_{00}+\pi_{10})$	$\pi_{11}/(\pi_{01}+\pi_{11})$

odds ratio in (b) and (c) the same

Bias and confounding

- recall bias
- publication bias
- surveillance bias – high risk groups may be studied more intensively
- ascertainment bias – screening appears to increase the risk in the population
- lead time bias – patients screened for a disease appear to have longer survival times (also called length time bias)
- survivor treatment selection bias – patients who will live longer have more opportunity for treatment
  - example: Redelmeier & Singh "Survival in academy-award winning actors and actresses" *Annals Internal Medicine* 2001
  - see Sylvestre & Hanley "Do Oscar-winners really live longer than successful peers?" *Annals Internal Medicine* 2006

Interpretation Chapter 9

- to what extent can we understand why the data are as they are rather than just describe patterns of variability?
- how generally applicable are the conclusions from a study?
- to what extent are the conclusions applicable in specific instances?
- what is meant by causal statements, and are they justified?
- how can the study conclusions best be integrated with the knowledge base of the field?

Causality Chapter 9

- an explanatory variable *C* has a causal impact on outcome on response *Y* if
  - conceptually, *C* could have taken any of its allowable values
  - in an aggregate sense there is evidence that *Y* values at level 1, say, of *C* are systematically different from those that would have been obtained on the same individuals at level 0 of *C*
- counterfactual
- an intrinsic variable cannot be causal in this sense

Causality Chapter 9

- the value of randomization in pointing to causality is in breaking the link between potential confounders and response

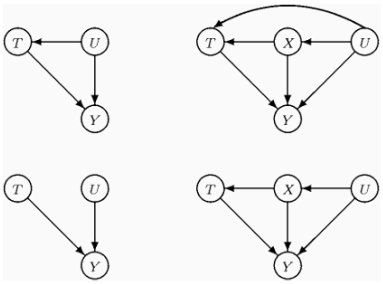


Figure 9.1 Directed acyclic graphs showing consequences of randomization. An arrow from T to Y indicates dependence of Y on T, and so forth. In general both response Y and treatment T may depend on properties U of units (upper left). Randomization (lower left) makes treatments and units independent, so any observed dependence of Y on T cannot be ascribed to joint dependence on U. The upper right graph shows the general dependence of Y, T, and covariates X on U. Randomization makes T and U independent, conditional on X (lower

the control group. The response is to be the blood pressure of an individual measured a fixed time after the drug has first been administered. We calculate the average changes

Causality Chapter 9

- in an observational study, to investigate C as a possible cause for Y we examine the relationship between Y and C conditionally on all variables explanatory to C
- the possibility that there is an unobserved confounder U is considered qualitatively – may be unmeasured and known, or unknown
- two sources of uncertainty: ordering in time, and unobserved confounding

Bradford-Hill Criteria Chapter 9

- strength of the association
- consistency of the association
- specificity of the association
- temporality – causes must precede effects
- biological gradient – dose-response relationship
- (biological) plausibility – consistent with known theory
- coherence
- experiment – based on a suitable (natural) experiment
- analogy

Austin Bradford Hill, "The Environment and Disease: Association or Causation?," *Proceedings of the Royal Society of Medicine*, 58 (1965), 295-300

Bradford-Hill Criteria Chapter 9

- strength of the association
  - larger effects less likely to be due to unobserved confounders
  - Cornfield's Lemma:

$$\frac{f_{AB}}{f_B} > R_A$$

$R_A : risk\ of\ A(>1)$

$f_{AB} : prevalence\ of\ B\ in\ group\ exposed\ to\ A$

$f_B : prevalence\ of\ B\ in\ unexposed$

"Cigarette smokers have a ninefold greater risk of developing lung cancer than nonsmokers. ... Any characteristic proposed as a measure of the postulated cause common to both smoking status and lung-cancer risk must therefore be at least ninefold more prevalent among cigarette smokers than among nonsmokers." (Cornfield et al 1959)

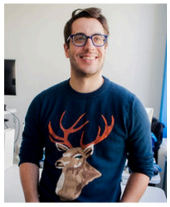
Interpretation Chapter 9

- to what extent are the conclusions applicable in specific instances?

TECHNOLOGY

On the Case at Mount Sinai, It's Dr. Data

By STEVE LOHR MARCH 7, 2015



Jeffrey Hammerbacher uses his finance and tech experience to understand diseases.
Sam Hodgson for The New York Times

Jeffrey Hammerbacher is a number cruncher — a Harvard math major who went from a job as a Wall Street quant to a key role at Facebook to a founder of a successful data start-up.

But five years ago, he was given a diagnosis of bipolar disorder, a crisis that fueled in him a fierce curiosity in medicine — about how the body and brain work and why they sometimes fail. The more he read and talked to experts, the more he became convinced that medicine needed people like him: skilled practitioners of data science who could guide scientific discovery and decision-making.

Now Mr. Hammerbacher, 32, is on the faculty of the Icahn School of Medicine at Mount Sinai, despite the fact that he has no academic training in medicine or biology. He is there because the school has begun an ambitious, well-funded initiative to apply data science to medicine.