Statistical modelling

Christian P. Robert

Université Paris Dauphine, IUF, & University of Warwick
https://sites.google.com/view/statistical-modelling

Licence MI2E, année 2020–2021

https://sites.google.com/view/statistical-modelling

About

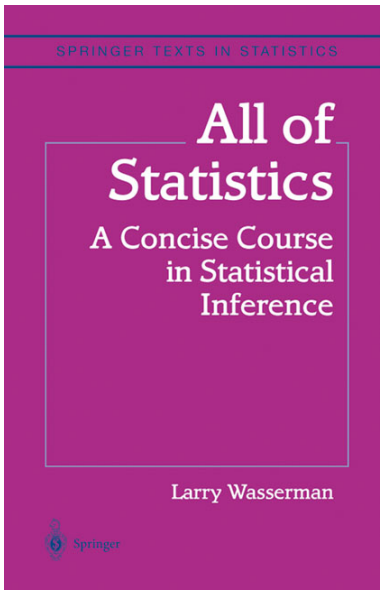
This course is the first part of the L3 statistics course, the second part being devoted to tests and model choice, taught by Marc Hoffmann. It covers the fundamentals of parametric statistics, both from mathematical and methodological points of view. With some forays into computational statistics. The main theme is that modelling is an inherent part of the statistical practice, rather than an

Contents

1. Statistics, the what and why
2. Probabilistic models for statistics
3. Glivenko-Cantelli theorem, Monte Carlo principles, and the bootstrap
4. Likelihood function, statistical information, and likelihood inference

Outline

1. the what and why of statistics
2. statistical models
3. bootstrap estimation
4. Likelihood function and inference
5. Decision theory and Bayesian analysis

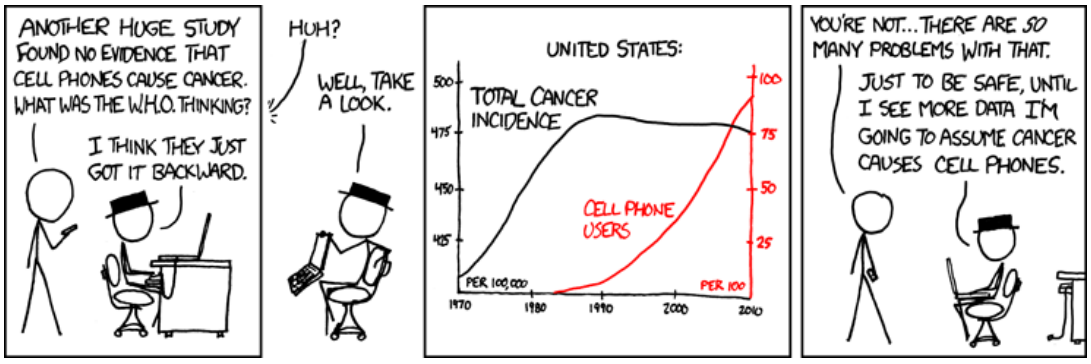


Chapter 0 : the what and why of statistics

What is statistics?

Many notions and usages of statistics, from description to action:

- summarising data
- extracting significant patterns from huge datasets
- exhibiting correlations
- smoothing time series
- predicting random events
- selecting influential variates
- making decisions
- identifying causes
- detecting fraudulent data



[xkcd]

What is statistics?

Many approaches to the field

- algebra
- data mining
- mathematical statistics
- machine learning
- computer science
- econometrics
- psychometrics

[xkcd]

Definition(s)

Given data $x_1, \ldots, x_n$, possibly driven by a probability distribution F, the goal is to infer about the distribution F with theoretical guarantees when n grows to infinity.

- data can be of arbitrary size and format
- driven means that the x_i's are considered as realisations of random variables related to F
- sample size n indicates the number of [not always exchangeable] replications
- distribution F denotes a probability distribution of a known or unknown transform of x_1
- inference may cover the parameters driving F or some functional of F
- guarantees mean getting to the "truth" or as close as possible to the "truth" with infinite data
- "truth" could be the entire F, some functional of F or some decision involving F

Definition(s)

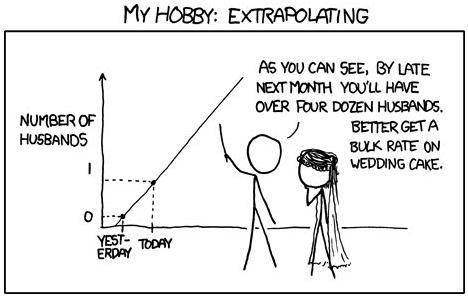
Given data $x_1, \ldots, x_n$, possibly driven by a probability distribution F, the goal is to infer about the distribution F with theoretical guarantees when n grows to infinity.

- data can be of arbitrary size and format
- driven means that the x_i's are considered as realisations of random variables related to F
- sample size n indicates the number of [not always exchangeable] replications
- distribution F denotes a probability distribution of a known or unknown transform of x_1
- inference may cover the parameters driving F or some functional of F
- guarantees mean getting to the "truth" or as close as possible to the "truth" with infinite data
- "truth" could be the entire F, some functional of F or some decision involving F

Warning: models are neither true nor real

Data most usually comes without a model, which is a mathematical construct intended to bring regularity and reproducibility, in order to draw inference

> *"All models are wrong but some are more useful than others"*
> *—George Box—*



Usefulness is to be understood as having explanatory or predictive abilities

Warning (2)

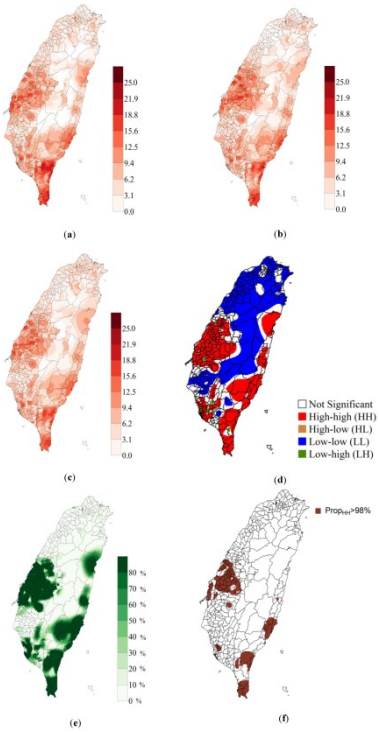
> *"Model produces data. The data does not produce the model."*
> *—P. Westfall and K. Henning—*

Meaning that

- a single model cannot be associated with a given dataset, no matter how precise the data gets
- but models can be checked by opposing artificial data from a model to observed data and spotting potential discrepancies

© Relevance of [computer] simulation tools relying on probabilistic models

Example 1: spatial pattern



(a) and (b) mortality in the 1st and 8th realizations; (c) mean mortality; (d) LISA map; (e) area covered by hot spots; (f) mortality distribution with high reliability

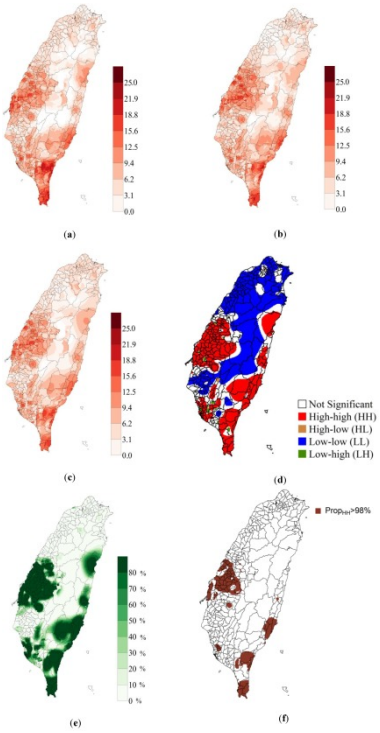
Mortality from oral cancer in Taiwan:

Model chosen to be

$$Y_i \sim \underbrace{\mathcal{P}(m_i)}_{\text{Poisson}} \quad \log m_i = \log E_i + a + \epsilon_i$$

[Lin et al., 2014, Int. J. Envir. Res. Pub. Health]

Example 1: spatial pattern



(a) and (b) mortality in the 1st and 8th realizations; (c) mean mortality; (d) LISA map; (e) area covered by hot spots; (f) mortality distribution with high reliability

Mortality from oral cancer in Taiwan:

Model chosen to be

$$Y_i \sim \mathcal{P}(m_i) \;\; \log m_i = \log E_i + a + \epsilon_i$$

where

- Y_i and E_i are observed and age/sex standardised expected counts in area i
- a is an intercept term representing the baseline (log) relative risk across the study region
- noise ϵ_i spatially structured with zero mean

[Lin et al., 2014, Int. J. Envir. Res. Pub. Health]

Example 2: World cup predictions

If team i and team j are playing and score y_i and y_j goals, resp., then the data point for this game is

$$y_{ij} = \text{sign}(y_i - y_j) \times \sqrt{|y_i - y_j|}$$

Corresponding data model is:

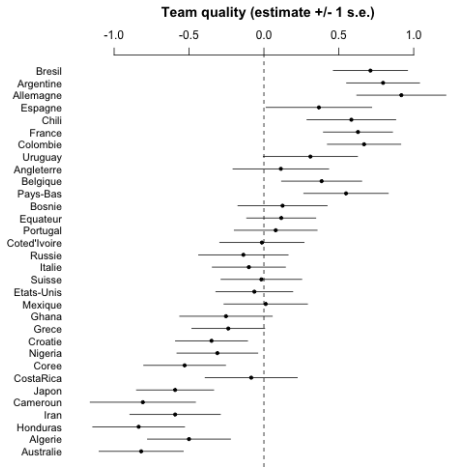
$$y_{ij} \sim \mathcal{N}(a_i - a_j, \sigma_y),$$

where a_i and a_j ability parameters and σ_y scale parameter estimated from the data

Nate Silver's prior scores

$$a_i \sim \mathcal{N}(b \times \text{prior score}_i, \sigma_a)$$

[A. Gelman, blog, 13 July 2014]



Resulting confidence intervals

Example 2: World cup predictions

If team i and team j are playing and score y_i and y_j goals, resp., then the data point for this game is

$$y_{ij} = \text{sign}(y_i - y_j) \times \sqrt{|y_i - y_j|}$$

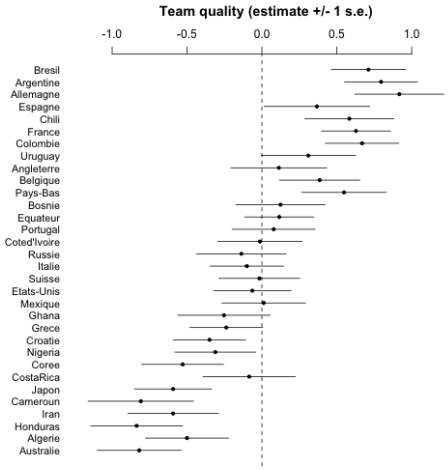
Potential outliers led to fatter tail model:

$$y_{ij} \sim \mathcal{T}_7(a_i - a_j, \sigma_y),$$

Nate Silver's prior scores

$$a_i \sim \mathcal{N}(b \times \text{prior score}_i, \sigma_a)$$

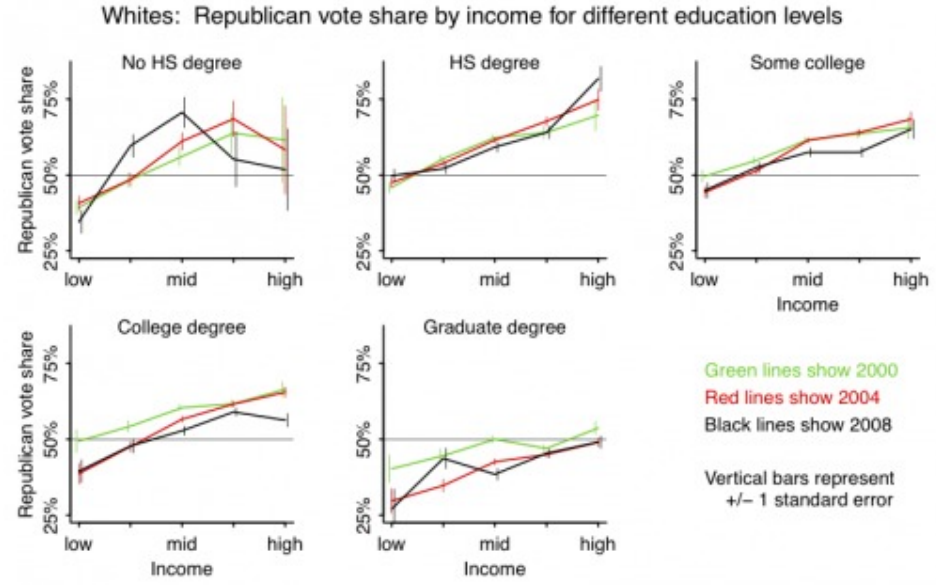
[A. Gelman, blog, 13 July 2014]



Resulting confidence intervals

Example 3: American voting patterns

"Within any education category, richer people vote more Republican. In contrast, the pattern of education and voting is nonlinear."



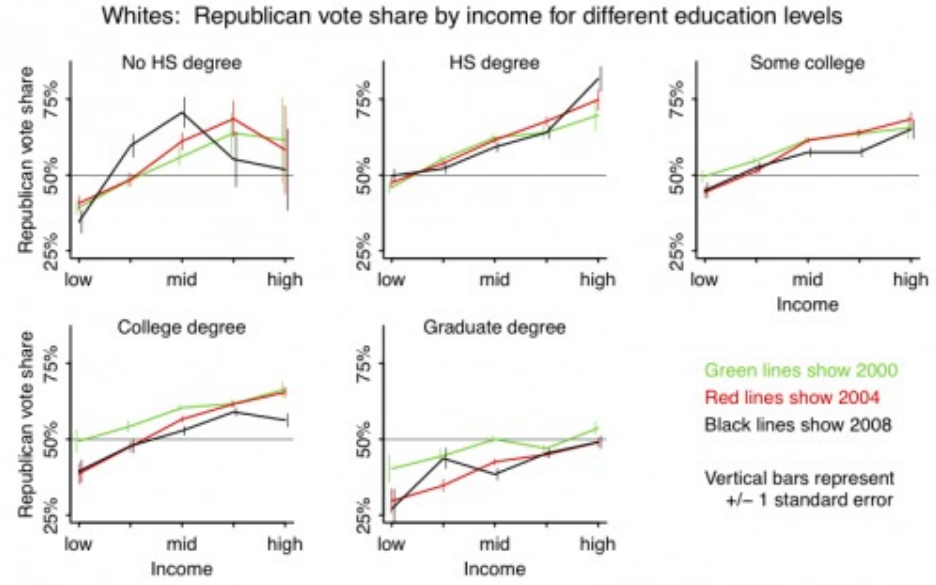
Example 3: American voting patterns

"Within any education category, richer people vote more Republican. In contrast, the pattern of education and voting is nonlinear."
"There is no plausible way based on these data in which elites can be considered a Democratic voting bloc. To create a group of strongly Democratic-leaning elite whites using these graphs, you would need to consider only postgraduates (...), and you have to go down to the below-$75,000 level of family income, which hardly seems like the American elites to me."

[A. Gelman, blog, 23 March 2012]

Example 3: American voting patterns

"Within any education category, richer people vote more Republican. In contrast, the pattern of education and voting is nonlinear."



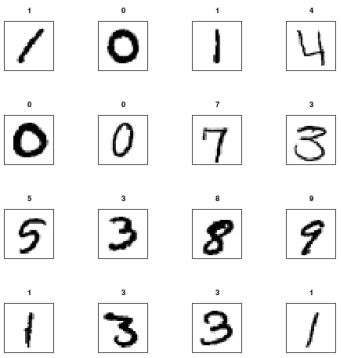
Example 4: Automatic number recognition

Reading postcodes and cheque amounts by analysing images of digits

Classification problem: allocate a new image (1024x1024 binary array) to one of the classes 0,1,...,9

Tools:

- linear discriminant analysis
- kernel discriminant analysis
- random forests
- support vector machine
- deep learning



Example 5: Asian beetle invasion

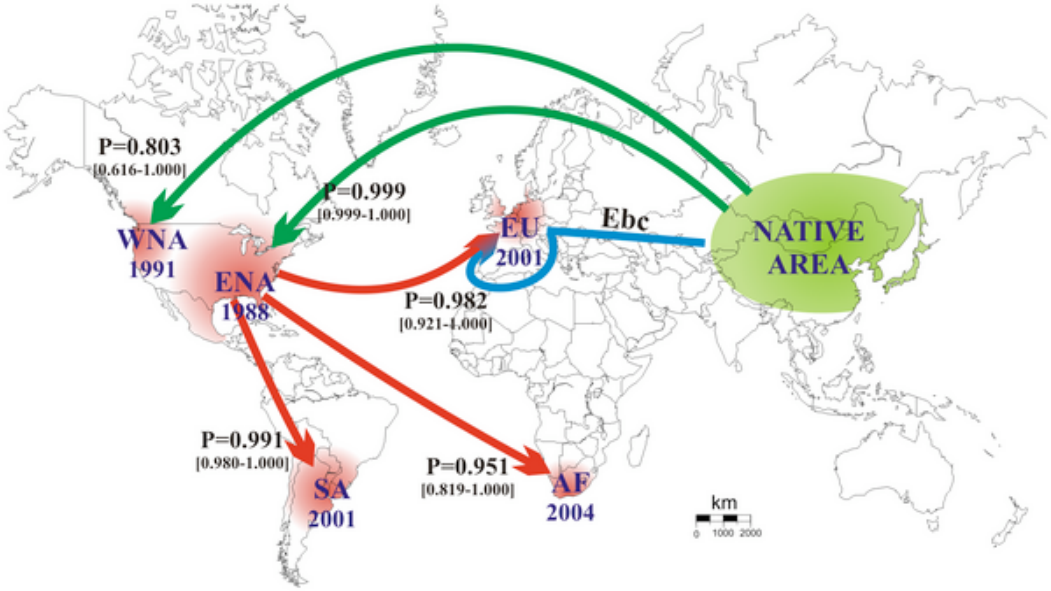
Several studies in recent years have shown the harlequin conquering other ladybirds across Europe. In the UK scientists found that seven of the eight native British species have declined. Similar problems have been encountered in Belgium and Switzerland.

[BBC News, 16 May 2013]

- How did the Asian Ladybird beetle arrive in Europe?
- Why do they swarm right now?
- What are the routes of invasion?
- How to get rid of them (biocontrol)?

[Estoup et al., 2012, Molecular Ecology Res.]

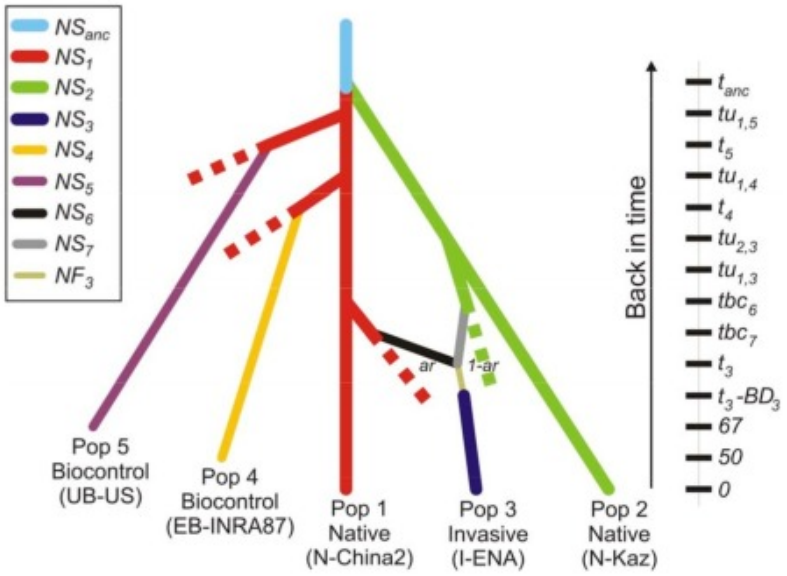
Example 5: Asian beetle invasion



For each outbreak, the arrow indicates the most likely invasion pathway and the associated posterior probability, with 95% credible intervals in brackets

[Lombaert & al., 2010, PLoS ONE]

Example 5: Asian beetle invasion

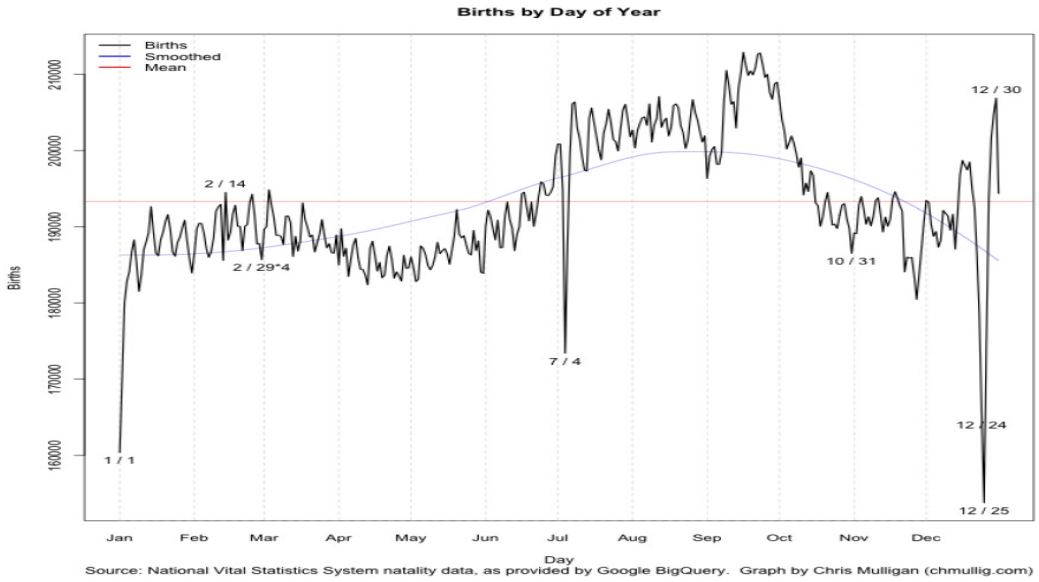


Most likely scenario of evolution, based on data:
samples from five populations (18 to 35 diploid individuals per sample), genotyped at 18 autosomal microsatellite loci, summarised into 130 statistics

[Lombaert & al., 2010, PLoS ONE]

Example 6: Are more babies born on Valentine's day than on Halloween?

Uneven pattern of birth rate across the calendar year



with large variations on heavily significant dates (Halloween, Valentine's day, April fool's day, Christmas, ...)

Example 6: Are more babies born on Valentine's day than on Halloween?

Uneven pattern of birth rate across the calendar year with large variations on heavily significant dates (Halloween, Valentine's day, April fool's day, Christmas, ...)

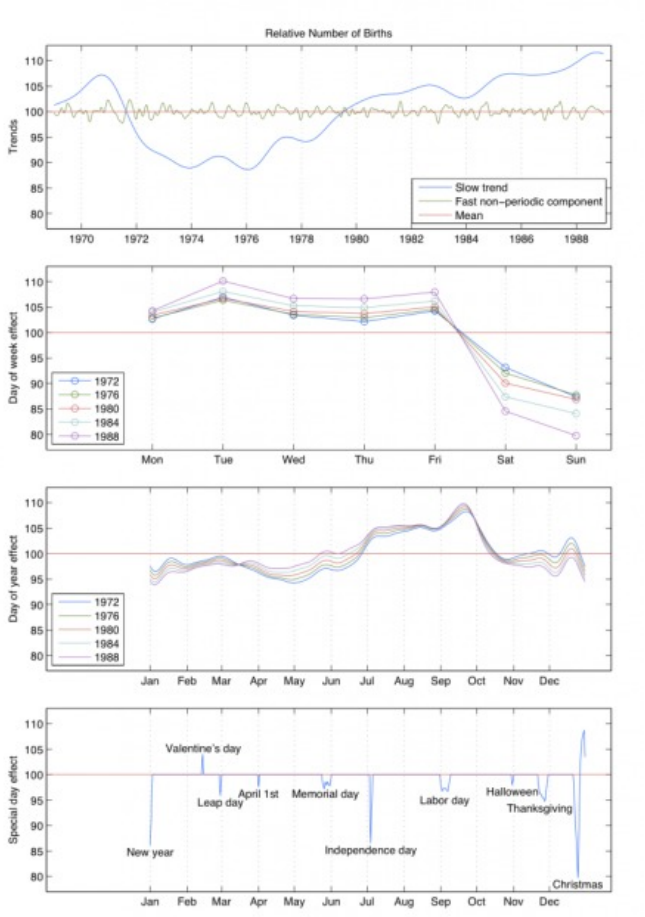
> *The data could be cleaned even further. Here's how I'd start: go back to the data for all the years and fit a regression with day-of-week indicators (Monday, Tuesday, etc), then take the residuals from that regression and pipe them back into [my] program to make a cleaned-up graph. It's well known that births are less frequent on the weekends, and unless your data happen to be an exact 28-year period, you'll get imbalance, which I'm guessing is driving a lot of the zigzagging in the graph above.*

Example 6: Are more babies born on Valentine's day than on Halloween?

I modeled the data with a Gaussian process with six components:

1. *slowly changing trend*
2. *7 day periodical component capturing day of week effect*
3. *365.25 day periodical component capturing day of year effect*
4. *component to take into account the special days and interaction with weekends*
5. *small time scale correlating noise*
6. *independent Gaussian noise*

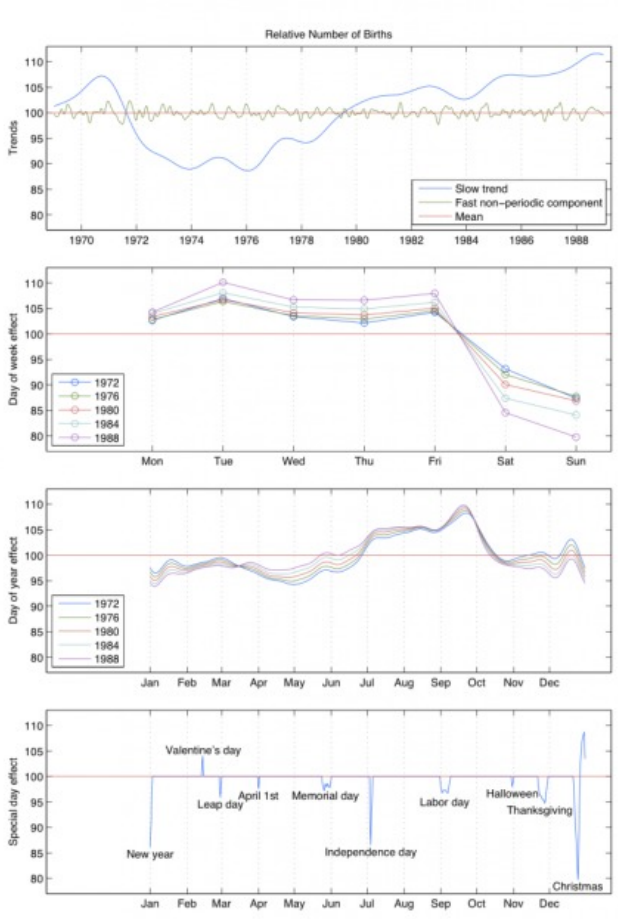
[A. Gelman, blog, 12 June 2012]



Example 6: Are more babies born on Valentine's day than on Halloween?

- *Day of the week effect has been increasing in 80's*
- *Day of year effect has changed only a little during years*
- *22nd to 31st December is strange time*

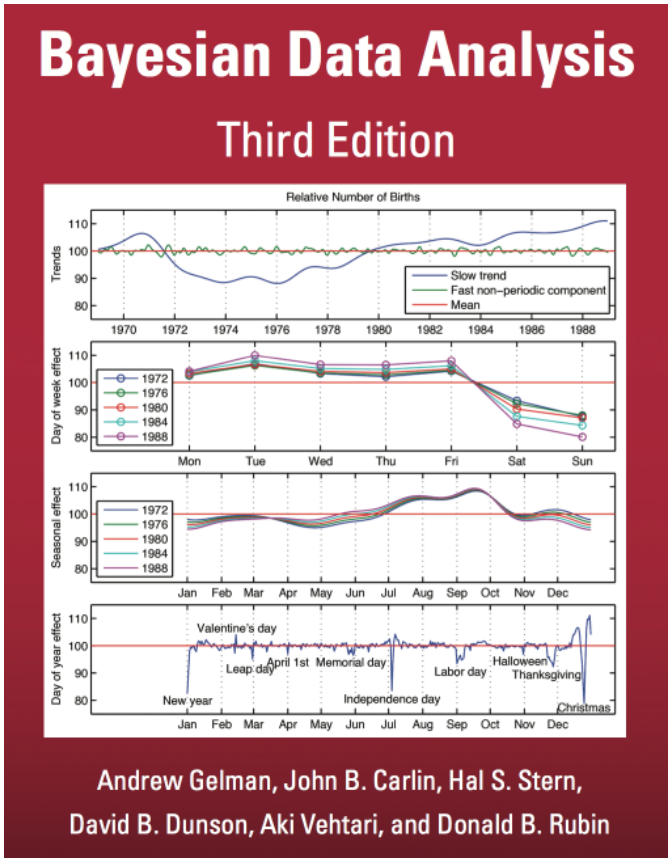
[A. Gelman, blog, 12 June 2012]



Example 6: Are more babies born on Valentine's day than on Halloween?

- *Day of the week effect has been increasing in 80's*
- *Day of year effect has changed only a little during years*
- *22nd to 31st December is strange time*

[A. Gelman, blog, 12 June 2012]



Example 7: Were the 2009 Iranian elections rigged?

Presidential elections of 2009 in Iran saw Mahmoud Ahmadinejad re-elected, amidst considerable protests against rigging.

> *...We'll concentrate on vote counts–the number of votes received by different candidates in different provinces–and in particular the last and second-to-last digits of these numbers. For example, if a candidate received 14,579 votes in a province (...), we'll focus on digits 7 and 9.*
> [B. Beber & A. Scacco, The Washington Post, June 20, 2009]

Similar analyses in other countries like Russia (2018)

Example 7: Were the 2009 Iranian elections rigged?

Presidential elections of 2009 in Iran saw Mahmoud Ahmadinejad re-elected, amidst considerable protests against rigging.

> *The ministry provided data for 29 provinces, and we examined the number of votes each of the four main candidates–Ahmadinejad, Mousavi, Karroubi and Mohsen Rezai–is reported to have received in each of the provinces–a total of 116 numbers.*
> [B. Beber & A. Scacco, The Washington Post, June 20, 2009]

Similar analyses in other countries like Russia (2018)

Example 7: Were the 2009 Iranian elections rigged?

Presidential elections of 2009 in Iran saw Mahmoud Ahmadinejad re-elected, amidst considerable protests against rigging.

> *The numbers look suspicious. We find too many 7s and not enough 5s in the last digit. We expect each digit (0, 1, 2, and so on) to appear at the end of 10 percent of the vote counts. But in Iran's provincial results, the digit 7 appears 17 percent of the time, and only 4 percent of the results end in the number 5. Two such departures from the average–a spike of 17 percent or more in one digit and a drop to 4 percent or less in another–are extremely unlikely. Fewer than four in a hundred non-fraudulent elections would produce such numbers.*
> [B. Beber & A. Scacco, The Washington Post, June 20, 2009]

Similar analyses in other countries like Russia (2018)

Why modelling?

Transforming (potentially deterministic) observations of a phenomenon "into" a model allows for

- detection of recurrent or rare patterns (outliers)
- identification of homogeneous groups (classification) and of changes
- selection of the most adequate scientific model or theory
- assessment of the significance of an effect (statistical test)
- comparison of treatments, populations, regimes, trainings, ...
- estimation of non-linear regression functions
- construction of dependence graphs and evaluation of conditional independence

Assumptions

Statistical analysis is always conditional to some mathematical assumptions on the underlying data like, e.g.,

- random sampling
- independent and identically distributed (i.i.d.) observations
- exchangeability
- stationary
- weakly stationary
- homocedasticity
- data missing at random

When those assumptions fail to hold, statistical procedures may prove unreliable

Warning: This does not mean statistical methodology only applies when the model is correct

Role of mathematics wrt statistics

Warning: This does not mean statistical methodology only applies when the model is correct

Statistics is not [solely] a branch of mathematics, but relies on mathematics to

- build probabilistic models
- construct procedures as optimising criteria
- validate procedures as asymptotically correct
- provide a measure of confidence in the reported results

© This is a mathematical statistics course

Role of mathematics wrt statistics

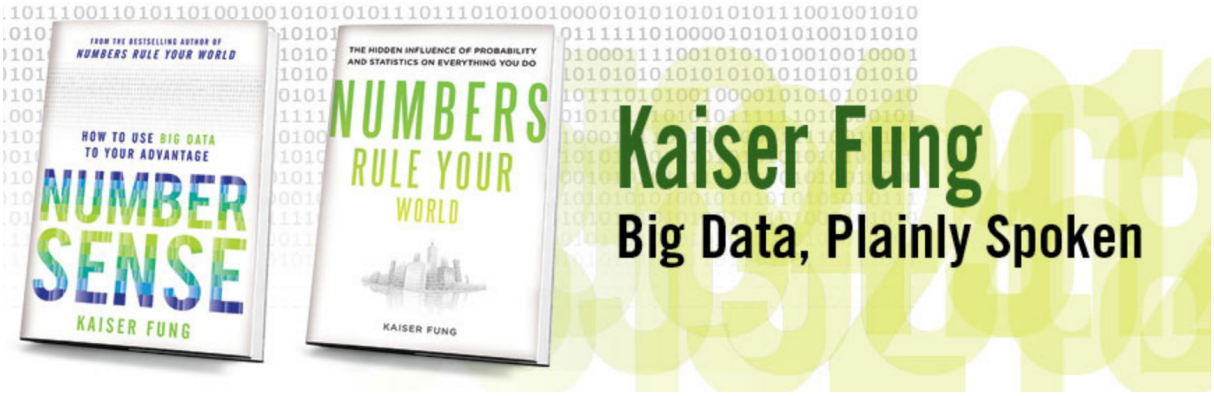
Warning: This does not mean statistical methodology only applies when the model is correct

Statistics is not [solely] a branch of mathematics, but relies on mathematics to

- build probabilistic models
- construct procedures as optimising criteria
- validate procedures as asymptotically correct
- provide a measure of confidence in the reported results

© This is a mathematical statistics course

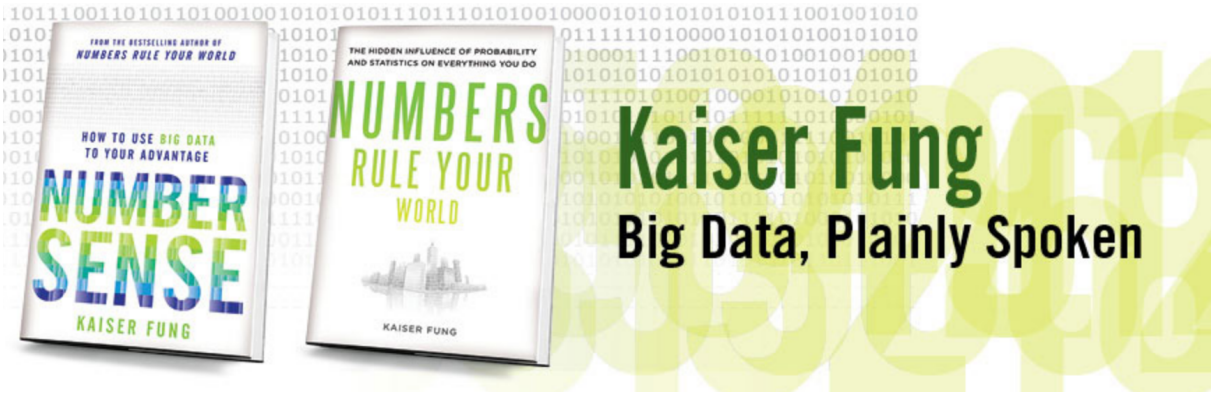
Six quotes from Kaiser Fung



- You may think you have all of the data. You don't.
- One of the biggest myths of Big Data is that data alone produce complete answers.
- Their "data" have done no arguing; it is the humans who are making this claim.

[Kaiser Fung, Big Data, Plainly Spoken blog]

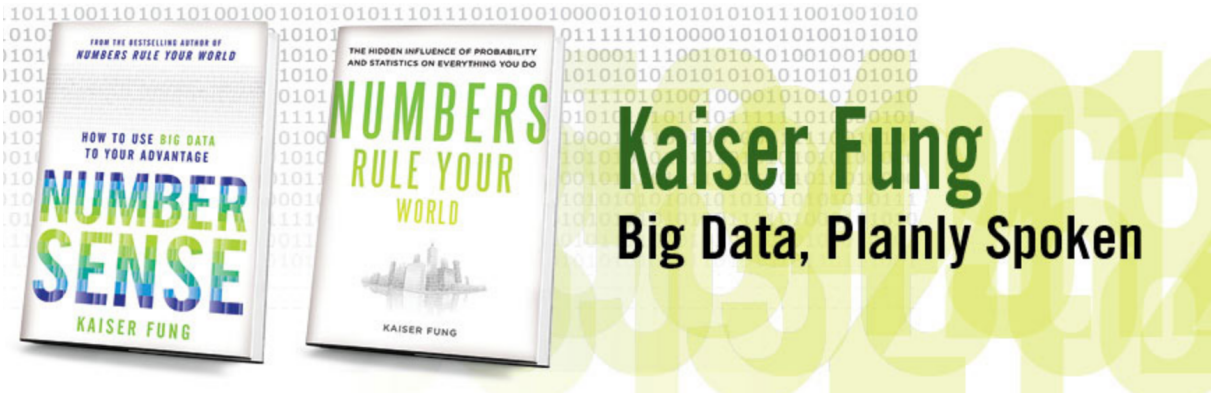
Six quotes from Kaiser Fung



- Before getting into the methodological issues, one needs to ask the most basic question. Did the researchers check the quality of the data or just take the data as is?
- We are not saying that statisticians should not tell stories. Story-telling is one of our responsibilities. What we want to see is a clear delineation of what is data-driven and what is theory (i.e., assumptions).

[Kaiser Fung, Big Data, Plainly Spoken blog]

Six quotes from Kaiser Fung



- The standard claim is that the observed effect is so large as to obviate the need for having a representative sample. Sorry — the bad news is that a huge effect for a tiny non-random segment of a large population can coexist with no effect for the entire population.

[Kaiser Fung, Big Data, Plainly Spoken blog]