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

Statistical Modelling

Dauphine UNIVERSITÉ PARIS | PSL Statistical Modelling

About

Contents

References

Courses Notes

Tutorial

Practical

Archives

Instructors

About

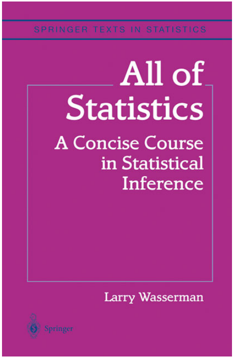
This course is the first part of the L3 statistics course, the second part being devoted to tests and model choice, taught by Marc Hoffmann. It covers the fundamentals of parametric statistics, both from mathematical and methodological points of view. With some forays into computational statistics. The 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



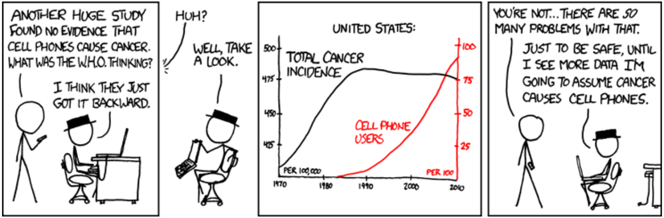
Chapter 0 : the what and why of statistics

1. the what and why of statistics
   - What?
   - Examples
   - Why?

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

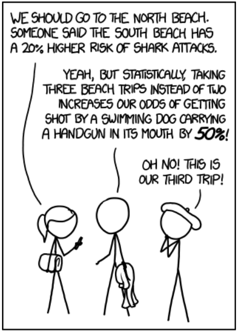


[xkcd]

What is statistics?

Many approaches to the field

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



REMINDER: A 50% INCREASE IN A TINY RISK IS STILL TINY.

[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)

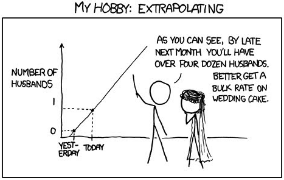
Given data $x_1, \ldots, x_n$, possibly driven by a probability distribution F, the goal is to infer about the distribution F with theoretical guarantees when n grows to infinity.

- 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)

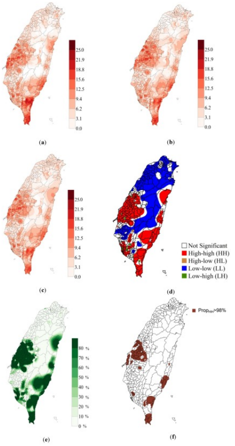
"Model produces data. The data does not produce the model."
—P. Westfall and K. Henning—

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

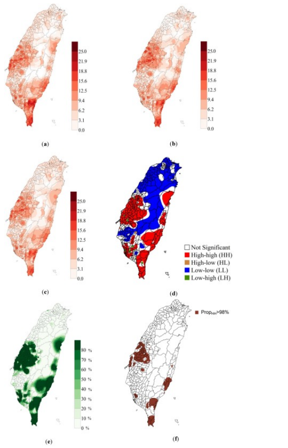
Mortality from oral cancer in Taiwan:

Model chosen to be

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

[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

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

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

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

Corresponding data model is:

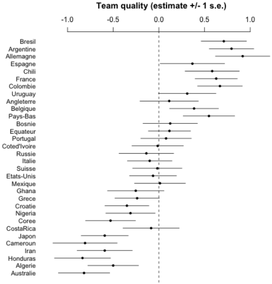
$$y_{ij} \sim \mathcal{N}(a_i - a_j, \sigma_y),$$

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

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

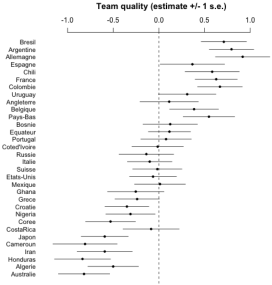
Potential outliers led to fatter tail model:

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

Nate Silver's prior scores

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

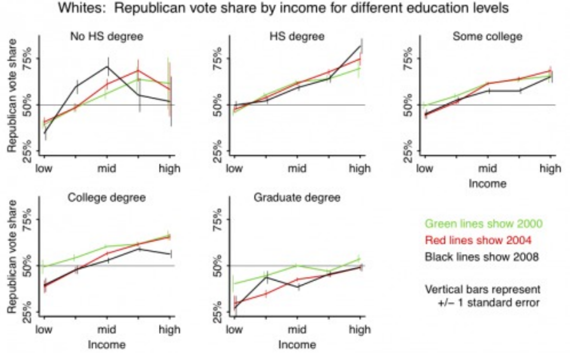
[A. Gelman, blog, 13 July 2014]



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."



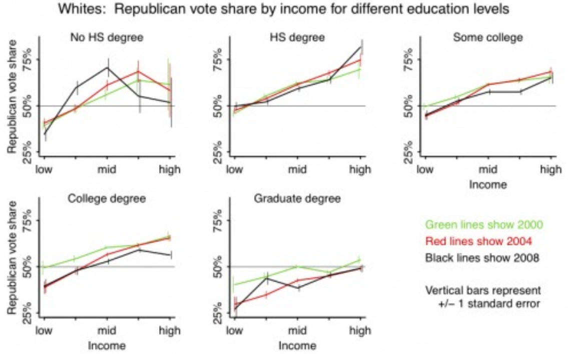
Example 3: American voting patterns

"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."



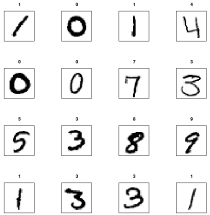
Example 4: Automatic number recognition

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

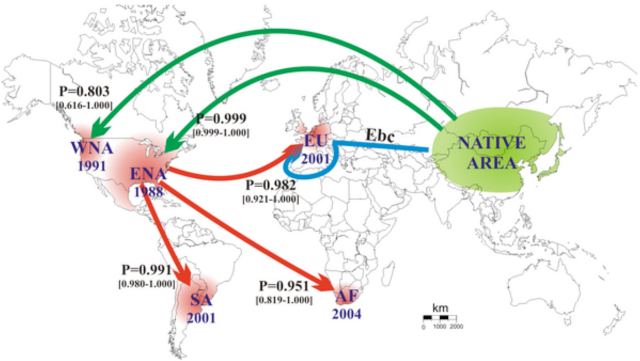
Several studies in recent years have shown the harlequin conquering other ladybirds across Europe. In the UK scientists found that seven of the eight native British species have declined. Similar problems have been encountered in Belgium and Switzerland.

[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.]

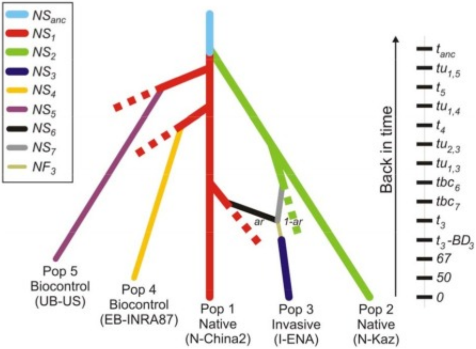
Example 5: Asian beetle invasion



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

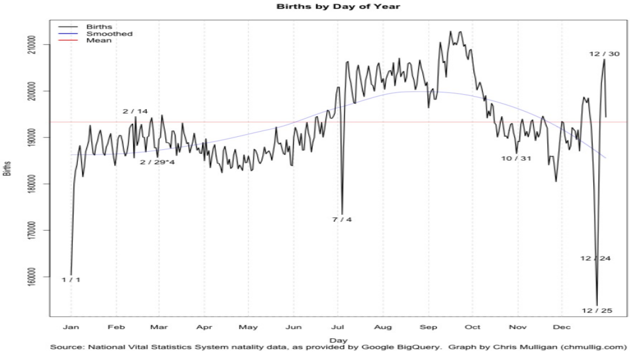


Most likely scenario of evolution, based on data:
samples from five populations (18 to 35 diploid individuals per sample), genotyped at 18 autosomal microsatellite loci, summarised into 130 statistics

[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, ...)

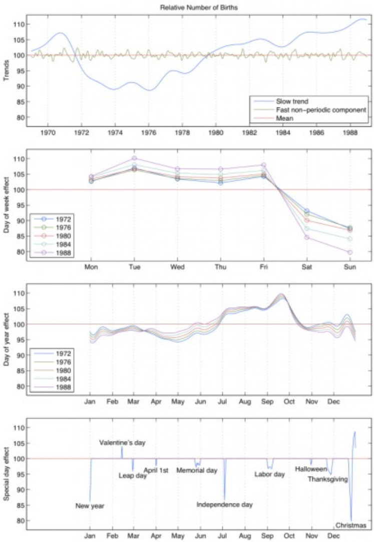
> *The data could be cleaned even further. Here's how I'd start: go back to the data for all the years and fit a regression with day-of-week indicators (Monday, Tuesday, etc), then take the residuals from that regression and pipe them back into [my] program to make a cleaned-up graph. It's well known 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?

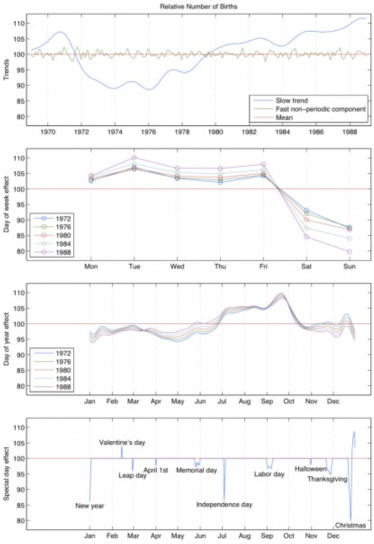
I modeled the data with a Gaussian process with six components:

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?



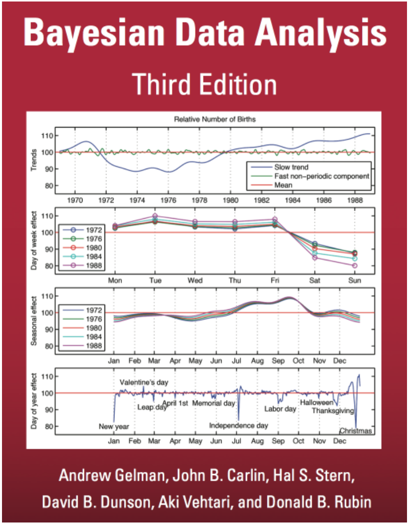
- *Day of the week effect has been increasing in 80's*
- *Day of year effect has changed only a little during years*
- *22nd to 31st December is strange time*

[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

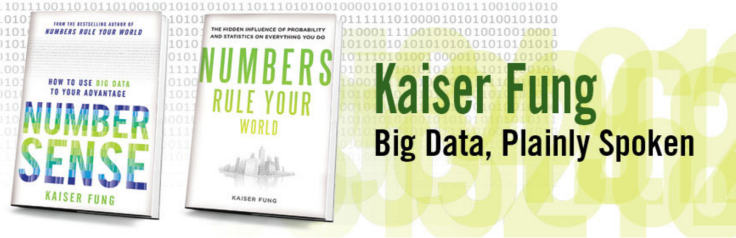
Warning: This does not mean statistical methodology only applies when the model is correct

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

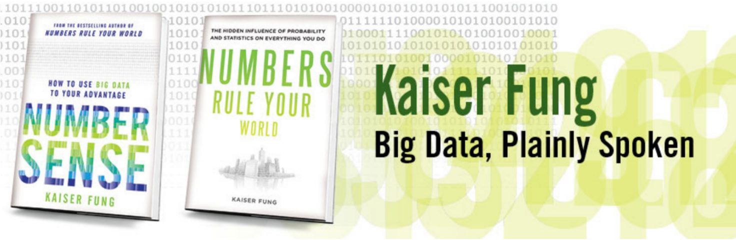
Six quotes from Kaiser Fung



- 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]

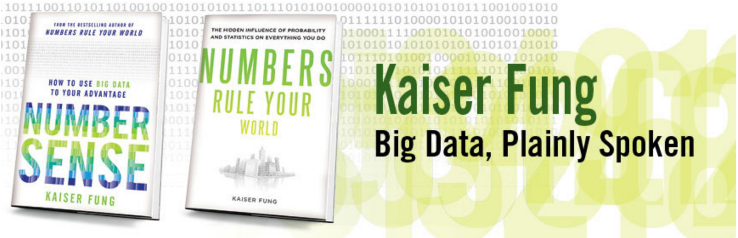
Six quotes from Kaiser Fung



- 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]

Chapter 1 : statistical vs. real models

- Statistical models
- Quantities of interest
- Exponential families

Statistical models

For most of the course, we assume that the data is a random sample $x_1, \ldots, x_n$ and that

$$X_1, \ldots, X_n \sim F(x)$$

as i.i.d. variables or as transforms of i.i.d. variables

[observations versus Random Variables]

Motivation:

Repetition of observations increases information about F, by virtue of probabilistic limit theorems (LLN, CLT)

Statistical models

For most of the course, we assume that the data is a random sample $x_1, \ldots, x_n$ and that

$$X_1, \ldots, X_n \sim F(x)$$

as i.i.d. variables or as transforms of i.i.d. variables

Motivation:

Repetition of observations increases information about F, by virtue of probabilistic limit theorems (LLN, CLT)

Warning 1: Some aspects of F may ultimately remain unavailable

Statistical models

For most of the course, we assume that the data is a random sample $x_1, \ldots, x_n$ and that

$$X_1, \ldots, X_n \sim F(x)$$

as i.i.d. variables or as transforms of i.i.d. variables

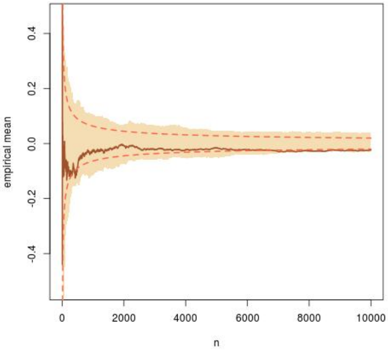
Motivation:

Repetition of observations increases information about F, by virtue of probabilistic limit theorems (LLN, CLT)

Warning 2: The model is always wrong, even though we behave as if...

Limit of averages

Case of an iid sequence $X_1, \ldots, X_n \sim \mathcal{N}(0, 1)$



Evolution of the range of $\bar{X}_n$ across 1000 repetitions, along with one random sequence and the theoretical 95% range

Limit theorems

Law of Large Numbers (LLN)

If $X_1, \ldots, X_n$ are i.i.d. random variables, with a well-defined expectation $\mathbb{E}[X]$

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{\text{prob}} \mathbb{E}[X]$$

[proof: see Terry Tao's "What's new", 18 June 2008]

Limit theorems

Law of Large Numbers (LLN)

If $X_1, \ldots, X_n$ are i.i.d. random variables, with a well-defined expectation $\mathbb{E}[X]$

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{\text{a.s.}} \mathbb{E}[X]$$

[proof: see Terry Tao's "What's new", 18 June 2008]

Limit theorems

Law of Large Numbers (LLN)

If $X_1, \ldots, X_n$ are i.i.d. random variables, with a well-defined expectation $\mathbb{E}[X]$

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{\text{a.s.}} \mathbb{E}[X]$$

Central Limit Theorem (CLT)

If $X_1, \ldots, X_n$ are i.i.d. random variables, with a well-defined expectation $\mathbb{E}[X]$ and a finite variance $\sigma^2 = \mathrm{var}(X)$,

$$\sqrt{n}\left\{\frac{X_1 + \ldots + X_n}{n} - \mathbb{E}[X]\right\} \xrightarrow{\text{dist.}} N(0, \sigma^2)$$

[proof: see Terry Tao's “What's new”, 5 January 2010]

Limit theorems

Central Limit Theorem (CLT)

If $X_1, \ldots, X_n$ are i.i.d. random variables, with a well-defined expectation $\mathbb{E}[X]$ and a finite variance $\sigma^2 = \mathrm{var}(X)$,

$$\sqrt{n}\left\{\frac{X_1 + \ldots + X_n}{n} - \mathbb{E}[X]\right\} \xrightarrow{\text{dist.}} N(0, \sigma^2)$$

[proof: see Terry Tao's "What's new", 5 January 2010]

Continuity Theorem

If

$$X_n \xrightarrow{\text{dist.}} a$$

and g is continuous at a, then

$$g(X_n) \xrightarrow{\text{dist.}} g(a)$$

Limit theorems

Central Limit Theorem (CLT)

If $X_1, \ldots, X_n$ are i.i.d. random variables, with a well-defined expectation $\mathbb{E}[X]$ and a finite variance $\sigma^2 = \mathrm{var}(X)$,

$$\sqrt{n}\left\{\frac{X_1 + \ldots + X_n}{n} - \mathbb{E}[X]\right\} \xrightarrow{\text{dist.}} N(0, \sigma^2)$$

[proof: see Terry Tao's "What's new", 5 January 2010]

Slutsky's Theorem

If X_n, Y_n, Z_n converge in distribution to X, a, and b, respectively, then

$$X_n Y_n + Z_n \xrightarrow{\text{dist.}} aX + b$$

Limit theorems

Central Limit Theorem (CLT)

If $X_1, \ldots, X_n$ are i.i.d. random variables, with a well-defined expectation $\mathbb{E}[X]$ and a finite variance $\sigma^2 = \mathrm{var}(X)$,

$$\sqrt{n}\left\{\frac{X_1 + \ldots + X_n}{n} - \mathbb{E}[X]\right\} \xrightarrow{\text{dist.}} N(0, \sigma^2)$$

[proof: see Terry Tao's "What's new", 5 January 2010]

Delta method's Theorem

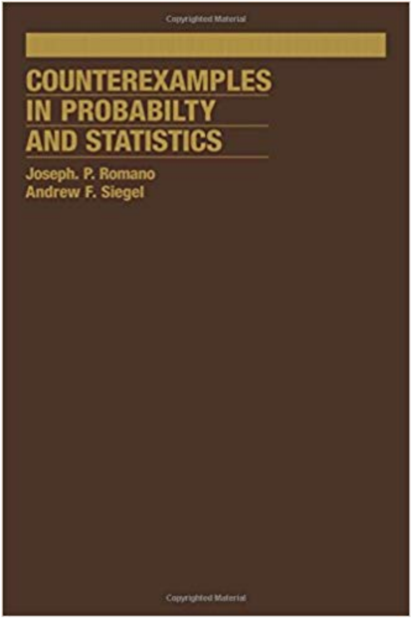
If

$$\sqrt{n}\{X_n - \mu\} \xrightarrow{\text{dist.}} N_p(0, \Omega)$$

and $g : \mathbb{R}^p \to \mathbb{R}^q$ is a continuously differentiable function on a neighbourhood of $\mu \in \mathbb{R}^p$, with a non-zero gradient $\nabla g(\mu)$, then

$$\sqrt{n}\{g(X_n) - g(\mu)\} \xrightarrow{\text{dist.}} N_q(0, \nabla g(\mu)^T \Omega \nabla g(\mu))$$

Entertaining read



Exemple 1: Binomial sample

Case # 1: Observation of i.i.d. Bernoulli variables

$$X_i \sim \mathcal{B}(p)$$

with unknown parameter p (e.g., opinion poll)
Case # 2: Observation of independent Bernoulli variables

$$X_i \sim \mathcal{B}(p_i)$$

with unknown and different parameters p_i (e.g., opinion poll, flu epidemics)
Transform of i.i.d. $U_1, \ldots, U_n$:

$$X_i = \mathbb{I}(U_i \leqslant p_i)$$

Exemple 1: Binomial sample

Case # 1: Observation of i.i.d. Bernoulli variables

$$X_i \sim \mathcal{B}(p)$$

with unknown parameter p (e.g., opinion poll)

Case # 2: Observation of conditionally independent Bernoulli variables

$$X_i | z_i \sim \mathcal{B}(p(z_i))$$

with covariate-driven parameters $p(z_i)$ (e.g., opinion poll, flu epidemics)

Transform of i.i.d. $U_1, \ldots, U_n$:

$$X_i = \mathbb{I}(U_i \leqslant p_i)$$

Parametric versus non-parametric

Two classes of statistical models:

- Parametric when F varies within a family of distributions indexed by a parameter θ that belongs to a finite dimension space Θ:

$$F \in \{F_\theta,\ \theta \in \Theta\}$$

  and to "know" F is to know which θ it corresponds to (identifiability);
- Non-parametric all other cases, i.e. when F is not constrained in a parametric way or when only some aspects of F are of interest for inference

Trivia: Machine-learning does not draw such a strict distinction between classes

Parametric versus non-parametric

Two classes of statistical models:

- Parametric when F varies within a family of distributions indexed by a parameter θ that belongs to a finite dimension space Θ:

$$F \in \{F_\theta,\ \theta \in \Theta\}$$

  and to "know" F is to know which θ it corresponds to (identifiability);
- Non-parametric all other cases, i.e. when F is not constrained in a parametric way or when only some aspects of F are of interest for inference

Trivia: Machine-learning does not draw such a strict distinction between classes

Non-parametric models

In non-parametric models, there may still be constraints on the range of F's as for instance

$$\mathbb{E}_F[Y|X = x] = \Psi(\beta^T x), \ \mathsf{var}_F(Y|X = x) = \sigma^2$$

in which case the statistical inference only deals with estimating or testing the constrained aspects or providing prediction.
Note: Estimating a density or a regression function like $\Psi(\beta^T x)$ is only of interest in a restricted number of cases

Parametric models

When $F = F_\theta$, inference usually covers the whole of the parameter θ and provides

- point estimates of θ, i.e. values substituting for the unknown "true" θ
- confidence intervals (or regions) on θ as regions likely to contain the "true" θ
- testing specific features of θ (true or not?) or of the whole family (goodness-of-fit)
- predicting some other variable whose distribution depends on θ

$$z_1, \ldots, z_m \sim G_\theta(z)$$

Inference: all those procedures depend on the sample $(x_1, \ldots, x_n)$

Parametric models

When $F = F_\theta$, inference usually covers the whole of the parameter θ and provides

- point estimates of θ, i.e. values substituting for the unknown "true" θ
- confidence intervals (or regions) on θ as regions likely to contain the "true" θ
- testing specific features of θ (true or not?) or of the whole family (goodness-of-fit)
- predicting some other variable whose distribution depends on θ

$$z_1, \ldots, z_m \sim G_\theta(z)$$

Inference: all those procedures depend on the sample $(x_1, \ldots, x_n)$

Example 1: Binomial experiment again

Model: Observation of i.i.d. Bernoulli variables

$$X_i \sim \mathcal{B}(p)$$

with unknown parameter p (e.g., opinion poll)

Questions of interest:

1. likely value of p or range thereof
2. whether or not p exceeds a level p_0
3. how many more observations are needed to get an estimation of p precise within two decimals
4. what is the average length of a "lucky streak" (1's in a row)

Exemple 2: Normal sample

Model: Observation of i.i.d. Normal variates

$$X_i \sim N(\mu, \sigma^2)$$

with unknown parameters μ and $\sigma > 0$ (e.g., blood pressure)

Questions of interest:

1. likely value of μ or range thereof
2. whether or not μ is above the mean η of another sample $y_1, \ldots, y_m$
3. percentage of extreme values in the next batch of m x_i's
4. how many more observations to exclude $\mu = 0$ from likely values
5. which of the x_i's are outliers

Quantities of interest

Statistical distributions (incompletely) characterised by (1-D) moments:

- central moments

$$\mu_1 = \mathbb{E}\left[X\right] = \int x dF(x) \quad \mu_k = \mathbb{E}\left[(X - \mu_1)^k\right] \ \ k > 1$$

- non-central moments

$$\xi_k = \mathbb{E}\left[X^k\right] \ \ k \geqslant 1$$

- α quantile

$$\mathbb{P}(X < \zeta_\alpha) = \alpha$$

and (2-D) moments

$$\mathrm{cov}(X^i, X^j) = \int (x^i - \mathbb{E}[X^i])(x^j - \mathbb{E}[X^j]) dF(x^i, x^j)$$

Note: For parametric models, those quantities are transforms of the parameter θ

Example 1: Binomial experiment again

Model: Observation of i.i.d. Bernoulli variables

$$X_i \sim \mathcal{B}(p)$$

Single parameter p with

$$\mathbb{E}[X] = p \;\; \mathrm{var}(X) = p(1-p)$$

[somewhat boring...]

Median and mode

Example 1: Binomial experiment again

Model: Observation of i.i.d. Binomial variables

$$X_i \sim \mathcal{B}(n, p) \ \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Single parameter p with

$$\mathbb{E}[X] = np \ \mathrm{var}(X) = np(1-p)$$

[somewhat less boring!]

Median and mode

Example 2: Normal experiment again

Model: Observation of i.i.d. Normal variates

$$X_i \sim N(\mu, \sigma^2) \quad i = 1, \ldots, n\,,$$

with unknown parameters μ and $\sigma > 0$ (e.g., blood pressure)

$$\mu_1 = \mathbb{E}[X] = \mu \ \mathrm{var}(X) = \sigma^2 \ \mu_3 = 0 \ \mu_4 = 3\sigma^4$$

Median and mode equal to μ

Exponential families

Class of parametric densities with nice analytic properties

Start from the normal density:

$$\begin{aligned}\varphi(x;\theta) &= \frac{1}{\sqrt{2\pi}} \exp\left\{x\theta - {}^{x^2}\!/_2 - {}^{\theta^2}\!/_2\right\} \\ &= \frac{\exp\{-\theta^2/2\}}{\sqrt{2\pi}} \underbrace{\exp\{x\theta\}}_{x \text{ meets } \theta} \exp\left\{{}^{-x^2}\!/_2\right\}\end{aligned}$$

where θ and x only interact through single exponential product

Exponential families

Class of parametric densities with nice analytic properties

Definition

A parametric family of distributions on $\mathcal{X}$ is an exponential family if its density with respect to a measure ν satisfies

$$f(x|\theta) = c(\theta)h(x)\underbrace{\exp\{T(x)^{\mathrm{T}}\tau(\theta)\}}_{\text{scalar product}}, \theta \in \Theta,$$

where $T(\cdot)$ and $\tau(\cdot)$ are k-dimensional functions and $c(\cdot)$ and $h(\cdot)$ are positive unidimensional functions.

Function $c(\cdot)$ is redundant, being defined by normalising constraint:

$$c(\theta)^{-1} = \int_{\mathcal{X}} h(x)\exp\{T(x)^{\mathrm{T}}\tau(\theta)\}d\nu(x)$$

Exponential families (examples)

Example 1: Binomial experiment again

Binomial variable

$$X \sim \mathcal{B}(n, p) \quad \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

can be expressed as

$$\mathbb{P}(X = k) = (1-p)^n \binom{n}{k} \exp\{k \log(p/(1-p))\}$$

hence

$$c(p) = (1-p)^n, \; h(x) = \binom{n}{x}, \; T(x) = x, \; \tau(p) = \log(p/(1-p))$$

Exponential families (examples)

Example 1: Binomial experiment again

Binomial variable

$$X \sim \mathcal{B}(n, p) \quad \mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

can be expressed as

$$\mathbb{P}(X = k) = (1 - p)^n \binom{n}{k} \exp\{k \log(p/(1 - p))\}$$

hence

$$c(p) = (1 - p)^n\,,\ h(x) = \binom{n}{x}\,,\ T(x) = x\,,\ \tau(p) = \log(p/(1 - p))$$

Exponential families (examples)

Example 2: Normal experiment again

Normal variate

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

with parameter $\theta = (\mu, \sigma^2)$ and density

$$\begin{aligned}
f(x|\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(x-\mu)^2/2\sigma^2\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-x^2/2\sigma^2 + x\mu/\sigma^2 - \mu^2/2\sigma^2\} \\
&= \frac{\exp\{-\mu^2/2\sigma^2\}}{\sqrt{2\pi\sigma^2}} \exp\{-x^2/2\sigma^2 + x\mu/\sigma^2\}
\end{aligned}$$

hence

$$c(\theta) = \frac{\exp\{-\mu^2/2\sigma^2\}}{\sqrt{2\pi\sigma^2}}, \; T(x) = \begin{pmatrix} x^2 \\ x \end{pmatrix}, \; \tau(\theta) = \begin{pmatrix} -1/2\sigma^2 \\ \mu/\sigma^2 \end{pmatrix}$$

natural exponential families

reparameterisation induced by the shape of the density:

Definition

In an exponential family, the natural parameter is $\tau(\theta)$ and the natural parameter space is

$$\Theta = \left\{ \tau \in \mathbb{R}^k ; \int_{\mathcal{X}} h(x) \exp\{T(x)^\mathrm{T}\tau\} d\nu(x) < \infty \right\}$$

Example For the $\mathcal{B}(m, p)$ distribution, the natural parameter is

$$\theta = \log\{p/(1-p)\}$$

and the natural parameter space is $\mathbb{R}$

natural exponential families

reparameterisation induced by the shape of the density:

Definition

In an exponential family, the natural parameter is $\tau(\theta)$ and the natural parameter space is

$$\Theta = \left\{ \tau \in \mathbb{R}^k ; \int_{\mathcal{X}} h(x) \exp\{T(x)^{\mathrm{T}}\tau\} d\nu(x) < \infty \right\}$$

Example For the $\mathcal{B}(m, p)$ distribution, the natural parameter is

$$\theta = \log\{p/(1-p)\}$$

and the natural parameter space is $\mathbb{R}$

regular and minimal exponential families

Possible to add and (better!) delete useless components of T:

Definition

A regular exponential family corresponds to the case where Θ is an open set.
A minimal exponential family corresponds to the case when the $T_i(X)$'s are linearly independent, i.e.

$$\mathbb{P}_\theta(\alpha^T T(X) = \text{const.}) = 0 \quad \text{for } \alpha \neq 0 \quad \theta \in \Theta$$

Also called non-degenerate exponential family
Usual assumption when working with exponential families

regular and minimal exponential families

Possible to add and (better!) delete useless components of T:

Definition

A regular exponential family corresponds to the case where Θ is an open set.
A minimal exponential family corresponds to the case when the $T_i(X)$'s are linearly independent, i.e.

$$\mathbb{P}_\theta(\alpha^T T(X) = \text{const.}) = 0 \quad \text{for } \alpha \neq 0 \quad \theta \in \Theta$$

Also called non-degenerate exponential family
Usual assumption when working with exponential families

Illustrations

- For a Normal $\mathcal{N}(\mu, \sigma^2)$ distribution,

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}}\,\frac{1}{\sigma}\,\exp\{-\,{}^{x^2}\!/{}_{2\sigma^2} + {}^{\mu}\!/{}_{\sigma^2}\,x - \,{}^{\mu^2}\!/{}_{2\sigma^2}\}$$

  means this is a two-dimensional minimal exponential family

- For a fourth-power distribution

$$f(x|\mu) = C(\theta)\,\exp\{-(x-\theta)^4\}\} \propto e^{-x^4}\,e^{4\theta^3 x - 6\theta^2 x^2 + 4\theta x^3 - \theta^4}$$

  implies this is a three-dimensional minimal exponential family

[Exercise: find C]

convexity properties

Highly regular densities

Theorem

The natural parameter space Θ of an exponential family is convex and the inverse normalising constant $c^{-1}(\theta)$ is a convex function.

Example For $\mathcal{B}(n, p)$, the natural parameter space is $\mathbb{R}$ and the inverse normalising constant $(1 + \exp(\theta))^n$ is convex

convexity properties

Highly regular densities

Theorem

The natural parameter space Θ of an exponential family is convex and the inverse normalising constant $c^{-1}(\theta)$ is a convex function.

Example For $\mathcal{B}(n, p)$, the natural parameter space is $\mathbb{R}$ and the inverse normalising constant $(1 + \exp(\theta))^n$ is convex

analytic properties

Lemma

If the density of X has the minimal representation

$$f(x|\theta) = c(\theta)h(x)\exp\{T(x)^{\mathrm{T}}\theta\}$$

then the natural statistic $Z = T(X)$ is also distributed from an exponential family and there exists a measure ν_T such that the density of Z $[= T(X)]$ against ν_T is

$$f(z;\theta) = c(\theta)\exp\{z^{\mathrm{T}}\theta\}$$

analytic properties

Theorem

If the density of $Z = T(X)$ against ν_T is $c(\theta)\exp\{z^T\theta\}$, if the real value function φ is measurable, with

$$\int |\varphi(z)| \exp\{z^T\theta\} \, d\nu_T(z) < \infty$$

on the interior of Θ, then

$$f: \; \theta \to \int \varphi(z) \exp\{z^T\theta\} \, d\nu_T(z)$$

is an analytic function on the interior of Θ and

$$\nabla f(\theta) = \int z\varphi(z) \exp\{z^T\theta\} \, d\nu_T(z)$$

moments of exponential families

Normalising constant $c(\cdot)$ generating all moments

Proposition

If $T(\cdot) : \mathcal{X} \to \mathbb{R}^d$ and the density of $Z = T(X)$ is $\exp\{z^T\theta - \psi(\theta)\}$, then

$$\mathbb{E}_\theta\left[\exp\{T(x)^T u\}\right] = \exp\{\psi(\theta + u) - \psi(\theta)\}$$

and $\psi(\cdot)$ is the cumulant generating function.

[Laplace transform]

moments of exponential families

Normalising constant $c(\cdot)$ generating all moments

Proposition

If $T(\cdot): \mathcal{X} \to \mathbb{R}^d$ and the density of $Z = T(X)$ is $\exp\{z^T\theta - \psi(\theta)\}$, then

$$\mathbb{E}_\theta[T_i(X)] = \frac{\partial\psi(\theta)}{\partial\theta_i} \quad i = 1, \ldots, d,$$

and

$$\mathbb{E}_\theta\left[T_i(X)\, T_j(X)\right] = \frac{\partial^2\psi(\theta)}{\partial\theta_i\partial\theta_j} \quad i, j = 1, \ldots, d$$

Sort of integration by part in parameter space:

$$\int \left\{T_i(x) + \frac{\partial}{\partial\theta_i}\log c(\theta)\right\} c(\theta)h(x)\exp\{T(x)^T\theta\}d\nu(x) = \frac{\partial}{\partial\theta_i}1 = 0$$

Sample from exponential families

Take an exponential family

$$f(x|\theta) = h(x)\ \exp\left\{\tau(\theta)^{\mathrm{T}} T(x) - \psi(\theta)\right\}$$

and id sample $x_1, \ldots, x_n$ from $f(x|\theta)$.

Then

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} h(x_i)\ \exp\left\{\tau(\theta)^{\mathrm{T}} \sum_{i=1}^{n} T(x_i) - n\psi(\theta)\right\}$$

Remark

For an exponential family with summary statistic $T(\cdot)$, the statistic

$$S(X_1, \ldots, X_n) = \sum_{i=1}^{n} T(X_i)$$

is sufficient for describing the joint density

Sample from exponential families

Take an exponential family

$$f(x|\theta) = h(x)\ \exp\left\{\tau(\theta)^{\mathrm{T}} T(x) - \psi(\theta)\right\}$$

and id sample $x_1, \ldots, x_n$ from $f(x|\theta)$.
Then

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} h(x_i)\ \exp\left\{\tau(\theta)^{\mathrm{T}} \sum_{i=1}^{n} T(x_i) - n\psi(\theta)\right\}$$

Remark

For an exponential family with summary statistic $T(\cdot)$, the statistic

$$S(X_1, \ldots, X_n) = \sum_{i=1}^{n} T(X_i)$$

is sufficient for describing the joint density

connected examples of exponential families

Example

Chi-square χ^2_k distribution corresponding to distribution of $X_1^2 + \ldots + X_k^2$ when $X_i \sim \mathcal{N}(0,1)$, with density

$$f_k(z) = \frac{z^{k/2-1}\exp\{-z/2\}}{2^{k/2}\Gamma(k/2)} \; z \in \mathbb{R}_+$$

connected examples of exponential families

Counter-Example

Non-central chi-square $\chi^2_k(\lambda)$ distribution corresponding to distribution of $X_1^2 + \ldots + X_k^2$ when $X_i \sim \mathcal{N}(\mu, 1)$, with density

$$f_{k,\lambda}(z) = {}^1\!/_2 \left({}^z\!/_\lambda\right)^{k/4 - 1/2} \exp\{-(z+\lambda)/2\} I_{k/2-1}(\sqrt{z\lambda}) \quad z \in \mathbb{R}_+$$

where $\lambda = k\mu^2$ and I_ν Bessel function of second order

connected examples of exponential families

Counter-Example

Fisher $\mathcal{F}_{n,m}$ distribution

corresponding to the ratio

$$Z = \frac{Y_n/n}{Y_m/m} \; Y_n \sim \chi^2_n, \; Y_m \sim \chi^2_m \,,$$

with density

$$f_{m,n}(z) = \frac{\left(n/m\right)^{n/2}}{B\left(n/2, m/2\right)} z^{n/2-1} \left(1 + n/m z\right)^{-n+m/2} \; z \in \mathbb{R}_+$$

Example

Ising $\mathcal{B}e(n/2, m/2)$ distribution corresponding to the distribution of

$$Z = \frac{nY}{nY + m} \text{ when } Y \sim \mathcal{F}_{n,m}$$

has density

$$f_{m,n}(z) = \frac{1}{B(n/2, m/2)} \, z^{n/2-1} \, (1-z)^{m/2-1} \; z \in (0,1)$$

Counter-Example

Laplace double-exponential $\mathcal{L}(\mu, \sigma)$ distribution corresponding to the rescaled difference of two exponential $\mathcal{E}(\sigma^{-1})$ random variables,

$$Z = \mu + X_1 - X_2 \text{ when } X_1, X_2 \overset{\sim}{\text{iid}} \mathcal{E}(\sigma^{-1})$$

has density

$$f(z; \mu, \sigma) = \frac{1}{\sigma} \exp\{-\sigma^{-1}|x - \mu|\}$$

chapter 2 : the bootstrap method

- Introduction
- Glivenko-Cantelli Theorem
- The Monte Carlo method
- Bootstrap
- Parametric Bootstrap

Motivating example

Case of a random event with binary (Bernoulli) outcome $Z \in \{0, 1\}$ such that $\mathbb{P}(Z = 1) = p$

Observations $z_1, \dots, z_n$ (iid) put to use to approximate p by

$$\hat{p} = \hat{p}(z_1, \dots, z_n) = {}^1\!/_n \sum_{i=1}^{n} z_i$$

Illustration of a (moment/unbiased/maximum likelihood) estimator of p

intrinsic statistical randomness

inference based on a random sample implies uncertainty

Since it depends on a **random** sample, an estimator

$$\delta(X_1, \ldots, X_n)$$

also is a **random** variable

Hence “error” in the reply: an estimator produces a different estimation of the same quantity θ each time a new sample is used (data does produce the model)

intrinsic statistical randomness

inference based on a random sample implies uncertainty

Since it depends on a **random** sample, an estimator

$$\delta(X_1, \ldots, X_n)$$

also is a **random** variable

Hence "error" in the reply: an estimator produces a different estimation of the same quantity θ each time a new sample is used (data does produce the model)

intrinsic statistical randomness

inference based on a random sample implies uncertainty

Since it depends on a **random** sample, an estimator

$$\delta(X_1, \ldots, X_n)$$

also is a **random** variable

Hence "error" in the reply: an estimator produces a different estimation of the same quantity θ each time a new sample is used (data does produce the model)

infered variation

inference based on a random sample implies uncertainty

Question 1 :

How much does $\delta(X_1, \ldots, X_n)$ vary when the sample varies?

Question 2 :

What is the variance of $\delta(X_1, \ldots, X_n)$?

Question 3 :

What is the distribution of $\delta(X_1, \ldots, X_n)$?

infered variation

inference based on a random sample implies uncertainty

Question 1 :

How much does $\delta(X_1, \ldots, X_n)$ vary when the sample varies?

Question 2 :

What is the variance of $\delta(X_1, \ldots, X_n)$?

Question 3 :

What is the distribution of $\delta(X_1, \ldots, X_n)$?

infered variation

inference based on a random sample implies uncertainty

Question 1 :

How much does $\delta(X_1, \ldots, X_n)$ vary when the sample varies?

Question 2 :

What is the variance of $\delta(X_1, \ldots, X_n)$?

Question 3 :

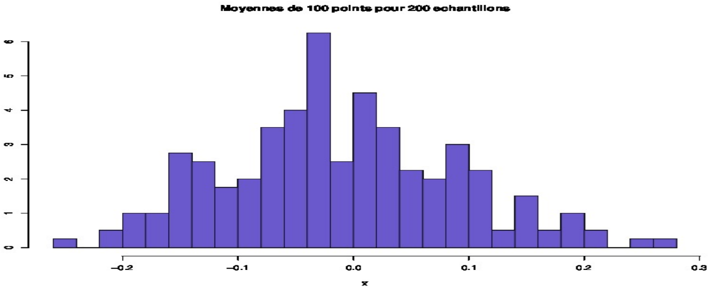
What is the distribution of $\delta(X_1, \ldots, X_n)$?

infered variation

Example (**Normal sample)**

Take $X_1, \dots, X_{100}$ a random sample from $\mathcal{N}(\theta, 1)$. Its mean θ is estimated by

$$\hat{\theta} = \frac{1}{100} \sum_{i=1}^{100} X_i$$



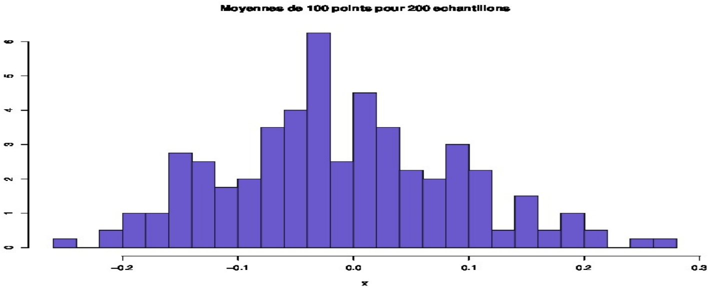
Variation compatible with the (known) theoretical distribution $\hat{\theta} \sim \mathcal{N}(\theta, 1/100)$

infered variation

Example (**Normal sample)**

Take $X_1, \ldots, X_{100}$ a random sample from $\mathcal{N}(\theta, 1)$. Its mean θ is estimated by

$$\hat{\theta} = \frac{1}{100} \sum_{i=1}^{100} X_i$$



Variation compatible with the (known) theoretical distribution $\hat{\theta} \sim \mathcal{N}(\theta, 1/100)$

Associated difficulties (illustrations)

- Observation of a **single** sample $x_1, \ldots, x_n$ in most cases
- The sampling distribution F is often unknown
- The evaluation of the average variation of $\delta(X_1, \ldots, X_n)$ is paramount for the construction of confidence intervals and for testing/answering questions like

$$H_0 : \theta \leqslant 0$$

- In the **normal** case, the **true** θ stands with high probability in the interval

$$[\hat{\theta} - 2\sigma, \hat{\theta} + 2\sigma] \,.$$

$\boxed{\textbf{Quid of } \sigma \textbf{ ?!}}$

Associated difficulties (illustrations)

- Observation of a **single** sample $x_1, \ldots, x_n$ in most cases
- The sampling distribution F is often unknown
- The evaluation of the average variation of $\delta(X_1, \ldots, X_n)$ is paramount for the construction of confidence intervals and for testing/answering questions like

$$H_0 : \theta \leqslant 0$$

- In the **normal** case, the **true** θ stands with high probability in the interval

$$[\hat{\theta} - 2\sigma, \hat{\theta} + 2\sigma] \,.$$

$\boxed{\textbf{Quid of } \sigma \textbf{ ?!}}$

Associated difficulties (illustrations)

- Observation of a **single** sample $x_1, \ldots, x_n$ in most cases
- The sampling distribution F is often unknown
- The evaluation of the average variation of $\delta(X_1, \ldots, X_n)$ is paramount for the construction of confidence intervals and for testing/answering questions like

$$H_0 : \theta \leqslant 0$$

- In the **normal** case, the **true** θ stands with high probability in the interval

$$[\hat{\theta} - 2\sigma, \hat{\theta} + 2\sigma] .$$

$\boxed{\textbf{Quid of } \sigma \textbf{ ?!}}$

Associated difficulties (illustrations)

- Observation of a **single** sample $x_1, \ldots, x_n$ in most cases
- The sampling distribution F is often unknown
- The evaluation of the average variation of $\delta(X_1, \ldots, X_n)$ is paramount for the construction of confidence intervals and for testing/answering questions like

$$H_0 : \theta \leqslant 0$$

- In the **normal** case, the **true** θ stands with high probability in the interval

$$[\hat{\theta} - 2\sigma, \hat{\theta} + 2\sigma] .$$

$\boxed{\textbf{Quid of } \sigma \textbf{ ?!}}$

Associated difficulties (illustrations)

- Observation of a **single** sample $x_1, \ldots, x_n$ in most cases
- The sampling distribution F is often unknown
- The evaluation of the average variation of $\delta(X_1, \ldots, X_n)$ is paramount for the construction of confidence intervals and for testing/answering questions like

$$H_0 : \theta \leqslant 0$$

- In the **normal** case, the **true** θ stands with high probability in the interval

$$[\hat{\theta} - 2\sigma, \hat{\theta} + 2\sigma] \, .$$

$\boxed{\textbf{Quid of } \sigma \textbf{ ?!}}$

Estimation of the repartition function

Extension/application of the LLN to the approximation of the cdf:

For an i.i.d. sample $X_1, \ldots, X_n$, empirical cdf

$$\begin{aligned}\hat{F}_n(x) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{]-\infty,x]}(X_i) \\ &= \frac{\text{card}\,\{X_i;\, X_i \leqslant x\}}{n},\end{aligned}$$

Step function corresponding to the empirical distribution

$$1/n \sum_{i=1}^{n} \delta_{X_i}$$

where δ Dirac mass

Estimation of the repartition function

Extension/application of the LLN to the approximation of the cdf:
For an i.i.d. sample $X_1, \ldots, X_n$, empirical cdf

$$\begin{aligned}\hat{F}_n(x) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{]-\infty,x]}(X_i)\\ &= \frac{\mathrm{card}\,\{X_i;\, X_i \leqslant x\}}{n}\,,\end{aligned}$$

Step function corresponding to the empirical distribution

$$1/n \sum_{i=1}^{n}\delta_{X_i}$$

where δ Dirac mass

convergence of the empirical cdf

Glivenko-Cantelli Theorem

$$\|\hat{F}_n - F\|_\infty = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0$$

[Glivenko, 1933;Cantelli, 1933]

$\hat{F}_n(x)$ **is a convergent estimator of the cdf** $F(x)$

convergence of the empirical cdf

Dvoretzky–Kiefer–Wolfowitz inequality

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x) - F(x)\right| > \varepsilon\right) \leqslant e^{-2n\varepsilon^2}$$

for every $\varepsilon \geqslant \varepsilon_n = \sqrt{1/2n \ln 2}$

[Massart, 1990]

$\hat{F}_n(x)$ **is a convergent estimator of the cdf** $F(x)$

convergence of the empirical cdf

Donsker's Theorem

The sequence

$$\sqrt{n}(\hat{F}_n(x) - F(x))$$

converges in distribution to a Gaussian process G with zero mean and covariance

$$\text{cov}[G(s), G(t)] = \mathbb{E}[G(s)G(t)] = \min\{F(s), F(t)\} - F(s)F(t).$$

[Donsker, 1952]

$\hat{F}_n(x)$ **is a convergent estimator of the cdf** $F(x)$

statistical consequences of Glivenko-Cantelli

Moments

$$\mathbb{E}[\hat{F}_n(x)] = F(x)$$

$$\mathrm{var}[\hat{F}_n(x)] = \frac{F(x)(1 - F(x))}{n}$$

statistical consequences of Glivenko-Cantelli

Confidence band

If

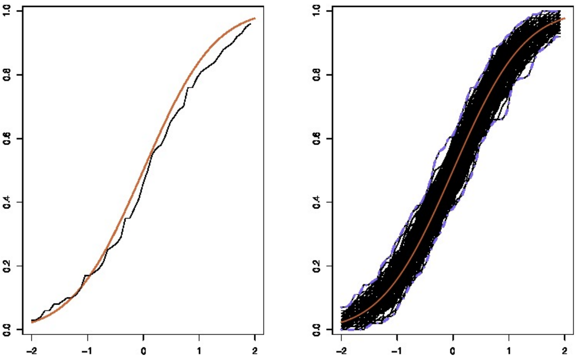
$$L_n(x) = \max\left\{\hat{F}_n(x) - \epsilon_n, 0\right\}, U_n(x) = \min\left\{\hat{F}_n(x) + \epsilon_n, 1\right\},$$

then, for $\varepsilon_n = \sqrt{1/2n \ln 2/\alpha}$,

$$\mathbb{P}\Big(L_n(x) \leqslant F(x) \leqslant U_n(x) \text{ for all } x\Big) \geqslant 1 - \alpha$$

Glivenko-Cantelli in action

Example (**Normal sample**)



Estimation of the cdf $\mathbb{F}$ from a normal sample of 100 points and variation of this estimation over 200 normal samples

Properties

- Estimator of a *non-parametric* nature : it is not necessary to know the distribution or the shape of the distribution of the sample to derive this estimator

  © it is always available

- **Robustess versus efficiency:** If the [parameterised] shape of the distribution is known, there exists a better approximation based on this shape, but if the shape is wrong, the parametric result can be completely off!

Properties

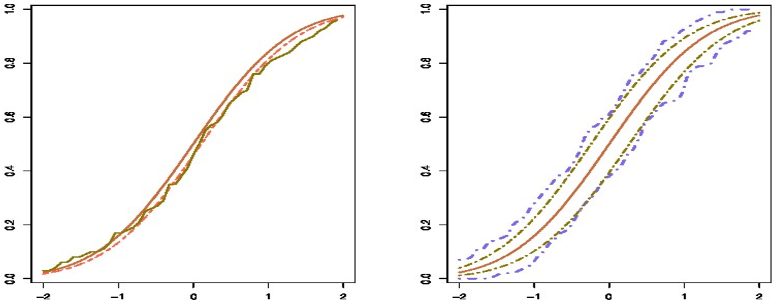
- Estimator of a *non-parametric* nature : it is not necessary to know the distribution or the shape of the distribution of the sample to derive this estimator

  © **it is always available**

- **Robustess versus efficiency:** If the [parameterised] shape of the distribution is known, there exists a better approximation based on this shape, but if the shape is wrong, the parametric result can be completely off!

Example (**Normal sample)**

cdf of $\mathcal{N}(\theta, 1)$, $\Phi(x - \theta)$



Estimation of $\Phi(\cdot - \theta)$ by $\hat{F}_n$ and by $\Phi(\cdot - \hat{\theta})$ based on 100 points and maximal variation of those estimations over 200 replications

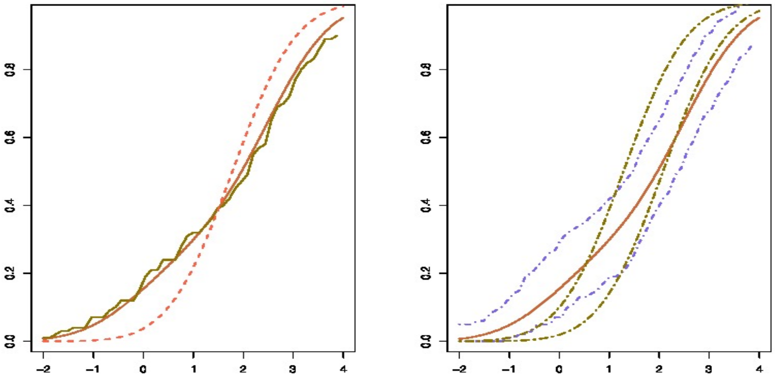
Example (Non-normal sample)

Sample issued from

$$0.3\mathcal{N}(0,1) + 0.7\mathcal{N}(2.5,1)$$

wrongly allocated to a normal distribution $\Phi(\cdot - \theta)$

parametric versus non-parametric inference



Estimation of F by $\hat{F}_n$ and by $\Phi(\cdot - \hat{\theta})$ based on 100 points and maximal variation of those estimations over 200 replications

Extension to functionals of F

For any quantity $\theta(F)$ depending on F, for instance,

$$\theta(F) = \int h(x)\, dF(x)\,,$$

[Functional of the cdf]

use of the plug-in approximation $\theta(\hat{F}_n)$, for instance,

$$\begin{aligned} \widehat{\theta(F)} &= \int h(x)\, d\hat{F}_n(x) \\ &= {}^1\!/\!_n \sum_{i=1}^{n} h(X_i) \end{aligned}$$

[Moment estimator]

Extension to functionals of F

For any quantity $\theta(F)$ depending on F, for instance,

$$\theta(F) = \int h(x)\,\mathrm{d}F(x)\,,$$

[Functional of the cdf]

use of the plug-in approximation $\theta(\hat{F}_n)$, for instance,

$$\begin{aligned}\widehat{\theta(F)} &= \int h(x)\,\mathrm{d}\hat{F}_n(x) \\ &= {}^1\!/_n \sum_{i=1}^{n} h(X_i)\end{aligned}$$

[Moment estimator]

examples

variance estimator

If

$$\theta(F) = \mathrm{var}(X) = \int \big(x - \mathbb{E}_F[X]\big)^2 dF(x)$$

then

$$\begin{aligned}\theta(\widehat{F}_n) &= \int \big(x - \mathbb{E}_{\widehat{F}_n}[X]\big)^2 d\widehat{F}_n(x)\\ &= {}^1\!/_n \sum_{i=1}^{n} \big(X_i - \mathbb{E}_{\widehat{F}_n}[X]\big)^2 = {}^1\!/_n \sum_{i=1}^{n} \big(X_i - \bar{X}_n\big)^2\end{aligned}$$

which differs from the (unbiased) sample variance

$${}^1\!/_{n-1} \sum_{i=1}^{n} \big(X_i - \bar{X}_n\big)^2$$

examples

median estimator

If $\theta(F)$ is the median of F, it is defined by

$$\mathbb{P}_F(X \leqslant \theta(F)) = 0.5$$

$\theta(\widehat{F}_n)$ is thus defined by

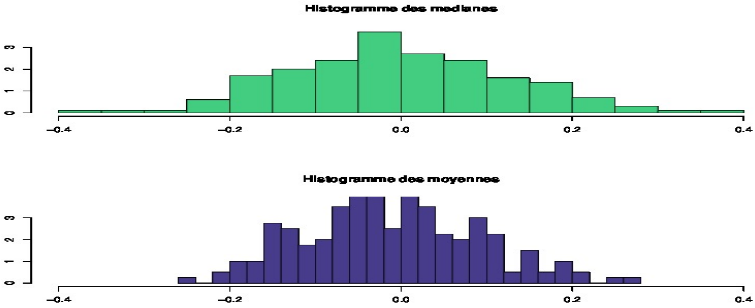
$$\mathbb{P}_{\widehat{F}_n}(X \leqslant \theta(\widehat{F}_n)) = {}^1\!/_n \sum_{i=1}^{n} \mathbb{I}(X_i \leqslant \theta(\widehat{F}_n)) = 0.5$$

which implies that $\theta(\widehat{F}_n)$ is the median of $X_1, \ldots, X_n$, namely $X_{(n/2)}$

median estimator

Example (**Normal sample)**

θ also is the median of $\mathcal{N}(\theta, 1)$, hence another estimator of θ is the median of $\hat{F}_n$, i.e. the median of $X_1, \ldots, X_n$, namely $X_{(n/2)}$



Comparison of the variations of sample means and sample medians over 200 normal samples

q-q plots

Graphical test of adequation for dataset $x_1, \ldots, x_n$ and targeted dsitribution F:

Plot sorted $x_1, \ldots, x_n$ against $F^{-1}(1/n+1), \ldots, F^{-1}(n/n+1)$

Example

Normal $\mathcal{N}(0, 1)$ sample against

- $\mathcal{N}(0, 1)$
- $\mathcal{N}(0, 2)$
- $\mathcal{E}(3)$

theoretical distributions

q-q plots

Graphical test of adequation for dataset $x_1, \ldots, x_n$ and targeted dsitribution F:

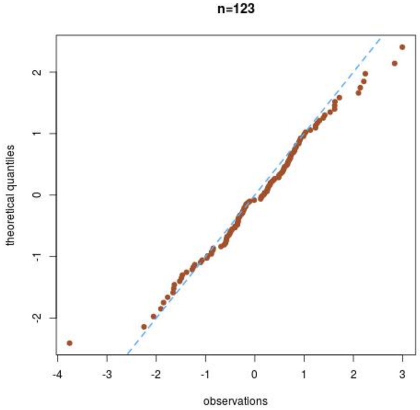
Plot sorted $x_1, \ldots, x_n$ against $F^{-1}(1/n+1), \ldots, F^{-1}(n/n+1)$

Example

Normal $\mathcal{N}(0,1)$ sample against

- $\mathcal{N}(0,1)$
- $\mathcal{N}(0,2)$
- $\mathcal{E}(3)$

theoretical distributions



q-q plots

Graphical test of adequation for dataset $x_1, \ldots, x_n$ and targeted dsitribution F:

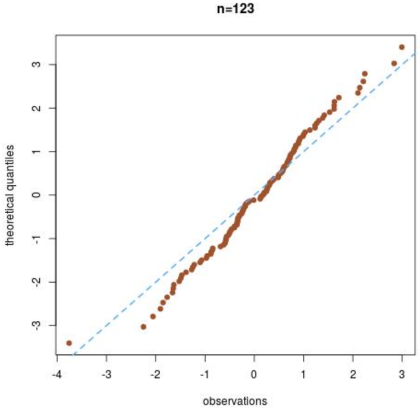
Plot sorted $x_1, \ldots, x_n$ against $F^{-1}(1/n+1), \ldots, F^{-1}(n/n+1)$

Example

Normal $\mathcal{N}(0,1)$ sample against

- $\mathcal{N}(0,1)$
- $\mathcal{N}(0,2)$
- $\mathcal{E}(3)$

theoretical distributions



q-q plots

Graphical test of adequation for dataset $x_1, \dots, x_n$ and targeted dsitribution F:

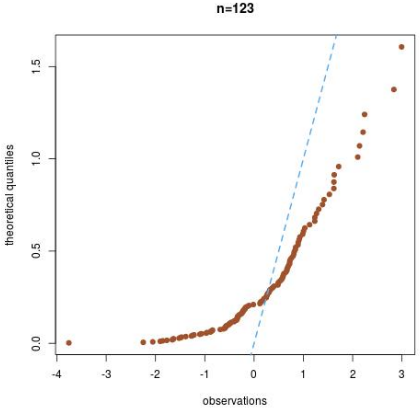
Plot sorted $x_1, \dots, x_n$ against $F^{-1}(1/n+1), \dots, F^{-1}(n/n+1)$

Example

Normal $\mathcal{N}(0, 1)$ sample against

- $\mathcal{N}(0, 1)$
- $\mathcal{N}(0, 2)$
- $\mathcal{E}(3)$

theoretical distributions



basis of Monte Carlo simulation

Recall the

Law of large numbers

If $X_1, \ldots, X_n$ simulated from f,

$$\widehat{\mathbb{E}[h(X)]}_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{\text{a.s.}} \mathbb{E}[h(X)]$$

Result fundamental for the use of computer-based simulation

basis of Monte Carlo simulation

Recall the

Law of large numbers

If $X_1, \ldots, X_n$ simulated from f,

$$\widehat{\mathbb{E}[h(X)]}_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i) \xrightarrow{\text{a.s.}} \mathbb{E}[h(X)]$$

Result fundamental for the use of computer-based simulation

computer simulation

Principle

- produce by a computer program an arbitrary long sequence

$$x_1, x_2, \ldots \overset{\text{iid}}{\sim} F$$

- exploit the sequence *as if* it were a truly iid sample

© Mix of algorithmic, statistics, and probability theory

computer simulation

Principle

- produce by a computer program an arbitrary long sequence

$$x_1, x_2, \ldots \overset{\text{iid}}{\sim} F$$

- exploit the sequence as if it were a truly iid sample

© **Mix of algorithmic, statistics, and probability theory**

Monte Carlo simulation in practice

- For a given distribution $\mathbb{F}$, call the corresponding pseudo-random generator in an arbitrary computer language

```
> x=rnorm(10)
> x
 [1] -0.02157345 -1.13473554  1.35981245 -0.88757941  0
 [7] -0.74941846  0.50629858  0.83579100  0.47214477
```

- use the sample as a statistician would do

```
> mean(x)
[1] 0.004892123
> var(x)
[1] 0.8034657
```

to approximate quantities related with $\mathbb{F}$

Monte Carlo integration

Approximation of integrals related with F:

Law of large numbers

If $X_1, \dots, X_n$ simulated from f,

$$\hat{\mathfrak{I}}_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i) \xrightarrow{\text{a.s.}} \mathfrak{I} = \int h(x)\, dF(x)$$

Convergence a.s. as $n \to \infty$

Monte Carlo principle

1. Call a computer pseudo-random generator of F to produce $x_1, \dots, x_n$
2. Approximate $\mathfrak{I}$ with $\hat{\mathfrak{I}}_n$
3. Check the precision of $\hat{\mathfrak{I}}_n$ and if needed increase n

Monte Carlo integration

Approximation of integrals related with F:

Law of large numbers

If $X_1, \dots, X_n$ simulated from f,

$$\hat{\mathfrak{I}}_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i) \xrightarrow{\text{a.s.}} \mathfrak{I} = \int h(x)\, dF(x)$$

Convergence a.s. as $n \to \infty$

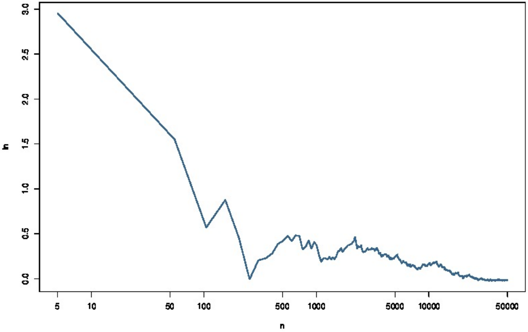
Monte Carlo principle

1. Call a computer pseudo-random generator of F to produce $x_1, \dots, x_n$
2. Approximate $\mathfrak{I}$ with $\hat{\mathfrak{I}}_n$
3. Check the precision of $\hat{\mathfrak{I}}_n$ and if needed increase n

example: normal moment

For a Gaussian distribution, $\mathbb{E}[X^4] = 3$. Via Monte Carlo integration,

n	5	50	500	5000	50,000	500,000
$\widehat{\mathfrak{I}}_n$	1.65	5.69	3.24	3.13	3.038	3.029



How can one approximate the distribution of $\theta(\hat{F}_n)$?

Given an estimate $\theta(\hat{F}_n)$ of $\theta(F)$, its variability is required to evaluate precision

bootstrap principle

Since

$$\theta(\hat{F}_n) = \theta(X_1, \ldots, X_n) \quad \textbf{with} \quad X_1, \ldots, X_n \overset{\textbf{iid}}{\sim} F$$

replace F **with** $\hat{F}_n$ **:**

$$\theta(\hat{F}_n) \approx \theta(X_1^*, \ldots, X_n^*) \quad \textbf{with} \quad X_1^*, \ldots, X_n^* \overset{\textbf{iid}}{\sim} \hat{F}_n$$

How can one approximate the distribution of $\theta(\hat{F}_n)$?

Given an estimate $\theta(\hat{F}_n)$ of $\theta(F)$, its variability is required to evaluate precision

bootstrap principle

Since

$$\theta(\hat{F}_n) = \theta(X_1, \ldots, X_n) \quad \textbf{with} \quad X_1, \ldots, X_n \overset{\textbf{iid}}{\sim} F$$

replace F **with** $\hat{F}_n$ **:**

$$\theta(\hat{F}_n) \approx \theta(X_1^*, \ldots, X_n^*) \quad \textbf{with} \quad X_1^*, \ldots, X_n^* \overset{\textbf{iid}}{\sim} \hat{F}_n$$

How can one approximate the distribution of $\theta(\widehat{F}_n)$?

Given an estimate $\theta(\widehat{F}_n)$ of $\theta(F)$, its variability is required to evaluate precision

bootstrap principle

Since

$$\theta(\widehat{F}_n) = \theta(X_1, \ldots, X_n) \quad \textbf{with} \quad X_1, \ldots, X_n \overset{\textbf{iid}}{\sim} F$$

replace F **with** $\widehat{F}_n$ **:**

$$\theta(\widehat{F}_n) \approx \theta(X_1^*, \ldots, X_n^*) \quad \textbf{with} \quad X_1^*, \ldots, X_n^* \overset{\textbf{iid}}{\sim} \widehat{F}_n$$

bootstrap principle

Since

$$\theta(\hat{F}_n) = \theta(X_1, \ldots, X_n) \quad \textbf{with} \quad X_1, \ldots, X_n \overset{\textbf{iid}}{\sim} F$$

replace F **with** $\hat{F}_n$ **:**

$$\theta(\hat{F}_n) \approx \theta(X_1^*, \ldots, X_n^*) \quad \textbf{with} \quad X_1^*, \ldots, X_n^* \overset{\textbf{iid}}{\sim} \hat{F}_n$$

illustration: bootstrap variance

For a given estimator $\theta(\hat{F}_n)$, a random variable, its (true) variance is defined as

$$\sigma^2 = \mathbb{E}_F\left[(\theta(\hat{F}_n) - \mathbb{E}_F[\theta(\hat{F}_n)])^2\right]$$

bootstrap approximation

$$\mathbb{E}_{\widehat{F}_n}\left[(\theta(\widehat{\hat{F}_n}) - \mathbb{E}_{\widehat{F}_n}[\theta(\hat{F}_n)])^2\right] = \mathbb{E}_{\widehat{F}_n}\left[\theta(\widehat{\hat{F}_n})^2\right] - \theta(\hat{F}_n)^2$$

meaning that the random variable $\theta(\widehat{\hat{F}_n})$ in the first expectation is now a transform of

$$X_1^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \widehat{F}_n$$

while the second $\theta(\hat{F}_n)$ is the original estimate

screen snapshot

bootstrap

/ˈbuːtstrap/

noun

noun: **bootstrap**; plural noun: **bootstraps**

1. a loop at the back of a boot, used to pull it on.
2. COMPUTING

   a technique of loading a program into a computer by means of a few initial instructions which enable the introduction of the rest of the program from an input device.
3. the technique of starting with existing resources to create something more complex and effective.

   "we see the creative act as a bootstrap process"

verb

verb: **bootstrap**; 3rd person present: **bootstraps**; gerund or present participle: **bootstrapping**; past tense: **bootstrapped**; past participle: **bootstrapped**

1. COMPUTING

   fuller form of boot[1] (sense 2 of the verb).
2. start up (an Internet-based business or other enterprise) with minimal financial resources.
   - get (oneself or something) into or out of a situation using existing resources.

     "the company is bootstrapping itself out of a marred financial past"

Remarks

- `bootstrap` because the sample itself is used to build an evaluation of its own distribution
- a bootstrap sample is obtained by n samplings with replacement in $(X_1, \ldots, X_n)$
- that is, X_1^* sampled from $(X_1, \ldots, X_n)$, then X_2^* independently sampled from $(X_1, \ldots, X_n)$, ...
- a bootstrap sample can thus take n^n values (or $\binom{2n-1}{n}$ values if the order does not matter)
- combinatorial complexity prevents analytic derivations

Remarks

- `bootstrap` because the sample itself is used to build an evaluation of its own distribution
- a bootstrap sample is obtained by n samplings with replacement in $(X_1, \ldots, X_n)$
- that is, X_1^* sampled from $(X_1, \ldots, X_n)$, then X_2^* independently sampled from $(X_1, \ldots, X_n)$, ...
- a bootstrap sample can thus take n^n values (or $\binom{2n-1}{n}$ values if the order does not matter)
- combinatorial complexity prevents analytic derivations

bootstrap by simulation

Implementation

Since $\hat{F}_n$ is known, it is possible to **simulate** from $\hat{F}_n$, therefore one can approximate the distribution of $\theta(X_1^*, \ldots, X_n^*)$ [instead of $\theta(X_1, \ldots, X_n)$]

The distribution corresponding to

$$\hat{F}_n(x) = \text{card}\{X_i; X_i \leqslant x\}/n$$

allocates a probability of $1/n$ to each point in $\{x_1, \ldots, x_n\}$:

$$\Pr^{\hat{F}_n}(X^* = x_i) = 1/n$$

Simulating from $\hat{F}_n$ is equivalent to sampling **with replacement** in $(X_1, \ldots, X_n)$

[in R, sample(x,n,replace=TRUE)]

bootstrap algorithm

Monte Carlo implementation

1. For $b = 1, \ldots, B$,
   1. generate a sample $X_1^b, \ldots, X_n^b$ from $\hat{F}_n$
   2. construct the corresponding value

$$\hat{\theta}^b = \theta(X_1^b, \ldots, X_n^b)$$

2. Use the sample

$$\hat{\theta}^1, \ldots, \hat{\theta}^B$$

   to approximate the distribution of

$$\theta(X_1, \ldots, X_n)$$

bootstrap algorithm

Monte Carlo implementation

1. For $b = 1, \ldots, B$,
   1. generate a sample $X_1^b, \ldots, X_n^b$ from $\hat{F}_n$
   2. construct the corresponding value

$$\hat{\theta}^b = \theta(X_1^b, \ldots, X_n^b)$$

2. Use the sample

$$\hat{\theta}^1, \ldots, \hat{\theta}^B$$

to approximate the distribution of

$$\theta(X_1, \ldots, X_n)$$

bootstrap algorithm

Monte Carlo implementation

1. For $b = 1, \dots, B$,
   1. generate a sample $X_1^b, \dots, X_n^b$ from $\hat{F}_n$
   2. construct the corresponding value

$$\hat{\theta}^b = \theta(X_1^b, \dots, X_n^b)$$

2. Use the sample

$$\hat{\theta}^1, \dots, \hat{\theta}^B$$

   to approximate the distribution of

$$\theta(X_1, \dots, X_n)$$

mixture illustration

- Observation of a sample [here simulated from $0.3\mathcal{N}(0,1) + 0.7\mathcal{N}(2.5,1)$ as illustration]

```
> x=rnorm(250)+(runif(250)<.7)*2.5 #n=250
```

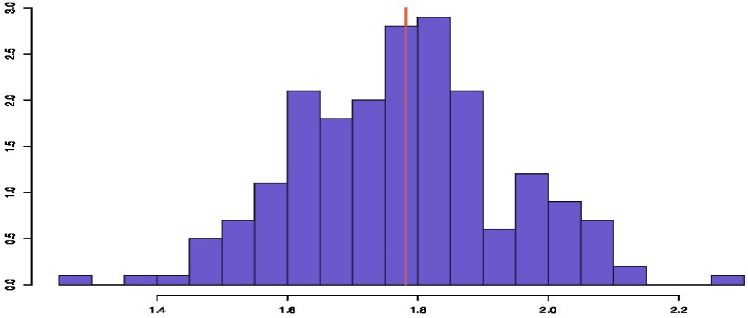
- Interest in the distribution of $\bar{X} = {}^1\!/_n \sum_i X_i$

```
> xbar=mean(x)
[1] 1.73696
```

- Bootstrap sample of $\bar{X}^*$

```
> bobar=rep(0,1000) #B=1000
> for (t in 1:1000)
+ bobar[t]=mean(sample(x,250,rep=TRUE))
> hist(bobar)
```

Example (**Sample** $0.3\mathcal{N}(0,1) + 0.7\mathcal{N}(2.5,1)$)



Variation of the empirical means over 200 bootstrap samples versus observed average

mixture illustration

Example (Derivation of the average variation)

For an estimator $\theta(X_1, \ldots, X_n)$, the standard deviation is given by

$$\eta(F) = \sqrt{E^F\left[\{\theta(X_1, \ldots, X_n) - E^F[\theta(X_1, \ldots, X_n)]\}^2\right]}$$

and its bootstrap approximation is

$$\eta(\hat{F}_n) = \sqrt{E^{\hat{F}_n}\left[\{\theta(X_1, \ldots, X_n) - E^{\hat{F}_n}[\theta(X_1, \ldots, X_n)]\}^2\right]}$$

mixture illustration

Example (Derivation of the average variation)

For an estimator $\theta(X_1, \ldots, X_n)$, the standard deviation is given by

$$\eta(F) = \sqrt{E^F \left[\{\theta(X_1, \ldots, X_n) - E^F[\theta(X_1, \ldots, X_n)]\}^2\right]}$$

and its bootstrap approximation is

$$\eta(\hat{F}_n) = \sqrt{E^{\hat{F}_n} \left[\{\theta(X_1, \ldots, X_n) - E^{\hat{F}_n}[\theta(X_1, \ldots, X_n)]\}^2\right]}$$

mixture illustration

Example (Derivation of the average variation)

Approximation itself approximated by Monte-Carlo:

$$\hat{\eta}(\hat{F}_n) = \left({}^1\!/_B \sum_{b=1}^{B} (\theta(X_1^b, \ldots, X_n^b) - \bar{\theta})^2 \right)^{1/2}$$

where

$$\bar{\theta} = {}^1\!/_B \sum_{b=1}^{B} \theta(X_1^b, \ldots, X_n^b)$$

bootstrap confidence intervals

Several ways to implement the bootstrap principle to get confidence intervals, that is intervals $C(X_1, \dots, X_n)$ on $\theta(F)$ such that

$$\mathbb{P}\big(C(X_1, \dots, X_n) \ni \theta(F)\big) = 1 - \alpha$$

[$1 - \alpha$-level confidence intervals]

1. rely on the normal approximation

$$\theta(\widehat{F}_n) \approx N(\theta(F), \eta(F)^2)$$

   and use the interval

$$\big[\theta(\widehat{F}_n) + z_{\alpha/2}\eta(\widehat{F}_n), \theta(\widehat{F}_n) - z_{\alpha/2}\eta(\widehat{F}_n)\big]$$

bootstrap confidence intervals

Several ways to implement the bootstrap principle to get confidence intervals, that is intervals $C(X_1, \ldots, X_n)$ on $\theta(F)$ such that

$$\mathbb{P}\big(C(X_1, \ldots, X_n) \ni \theta(F)\big) = 1 - \alpha$$

[$1 - \alpha$-level confidence intervals]

2 generate a bootstrap approximation to the cdf of $\theta(\widehat{F}_n)$

$$\widehat{H}(r) = {}^1\!/_B \sum_{b=1}^{B} \mathbb{I}(\theta(X_1^b, \ldots, X_n^b) \leqslant r)$$

and use the interval

$$\left[\widehat{H}^{-1}(\alpha/2), \widehat{H}^{-1}(1 - \alpha/2)\right]$$

which is also

$$\left[\theta^*_{(n\{\alpha/2\})}, \theta^*_{(n\{1-\alpha/2\})}\right]$$

bootstrap confidence intervals

Several ways to implement the bootstrap principle to get confidence intervals, that is intervals $C(X_1, \ldots, X_n)$ on $\theta(F)$ such that

$$\mathbb{P}\big(C(X_1, \ldots, X_n) \ni \theta(F)\big) = 1 - \alpha$$

[$1 - \alpha$-level confidence intervals]

3 generate a bootstrap approximation to the cdf of $\theta(\widehat{F}_n) - \theta(F)$,

$$\widehat{H}(r) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}((\theta(X_1^b, \ldots, X_n^b) - \theta(\widehat{F}_n) \leqslant r)$$

and use the interval

$$\big[\theta(\widehat{F}_n) - \widehat{H}^{-1}(1 - \alpha/2), \theta(\widehat{F}_n) - \widehat{H}^{-1}(\alpha/2)\big]$$

which is also

$$\big[2\theta(\widehat{F}_n) - \theta^*_{(n\{1-\alpha/2\})}, 2\theta(\widehat{F}_n) - \theta^*_{(n\{\alpha/2\})}\big]$$

exemple: median confidence intervals

Take $X_1, \ldots, X_n$ an iid random sample and $\theta(F)$ as the median of F, then

$$\theta(F_n) = X_{(n/2)}$$

```
> x=rnorm(123)
> median(x)
[1] 0.03542237
> T=10^3
> bootmed=rep(0,T)
> for (t in 1:T) bootmed[t]=median(sample(x,123,rep=TRUE))
> sd(bootmed)
[1] 0.1222386
> median(x)-2*sd(bootmed)
[1] -0.2090547
> median(x)+2*sd(bootmed)
[1] 0.2798995
```

exemple: median confidence intervals

Take $X_1, \ldots, X_n$ an iid random sample and $\theta(F)$ as the median of F, then

$$\theta(F_n) = X_{(n/2)}$$

```
> x=rnorm(123)
> median(x)
[1] 0.03542237
> T=10^3
> bootmed=rep(0,T)
> for (t in 1:T) bootmed[t]=median(sample(x,123,rep=TRUE))
> quantile(bootmed,prob=c(.025,.975))
      2.5%      97.5%
-0.2430018  0.2375104
```

exemple: median confidence intervals

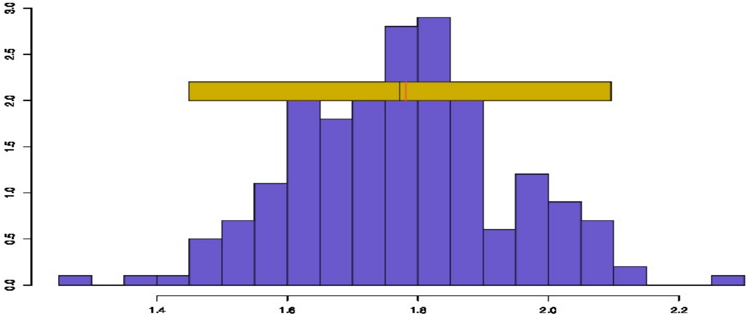
Take $X_1, \ldots, X_n$ an iid random sample and $\theta(F)$ as the median of F, then

$$\theta(F_n) = X_{(n/2)}$$

```
> x=rnorm(123)
> median(x)
[1] 0.03542237
> T=10^3
> bootmed=rep(0,T)
> for (t in 1:T) bootmed[t]=median(sample(x,123,rep=TRUE))
> 2*median(x)-quantile(bootmed,prob=c(.975,.025))
     97.5%       2.5%
-0.1666657  0.3138465
```

example: mean bootstrap variation

Example (**Sample** $0.3\mathcal{N}(0,1) + 0.7\mathcal{N}(2.5,1)$)



Interval of bootstrap variation at $\pm 2\hat{\eta}(\hat{F}_n)$ **and average of the observed sample**

example: mean bootstrap variation

Example (Normal sample)

Sample

$$(X_1, \ldots, X_{100}) \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$$

Comparison of the confidence intervals

$$[\bar{x} - 2 * \hat{\sigma}_x/10, \bar{x} + 2 * \hat{\sigma}_x/10] = [-0.113, 0.327]$$

[normal approximation]

$$[\bar{x}^* - 2 * \hat{\sigma}^*, \bar{x}^* + 2 * \hat{\sigma}^*] = [-0.116, 0.336]$$

[normal bootstrap approximation]

$$[q^*(0.025), q^*(0.975)] = [-0.112, 0.336]$$

[generic bootstrap approximation]

example: mean bootstrap variation

Example (Normal sample)

Sample

$$(X_1, \ldots, X_{100}) \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$$

Comparison of the confidence intervals

$$[\bar{x} - 2 * \hat{\sigma}_x/10, \bar{x} + 2 * \hat{\sigma}_x/10] = [-0.113, 0.327]$$

[normal approximation]

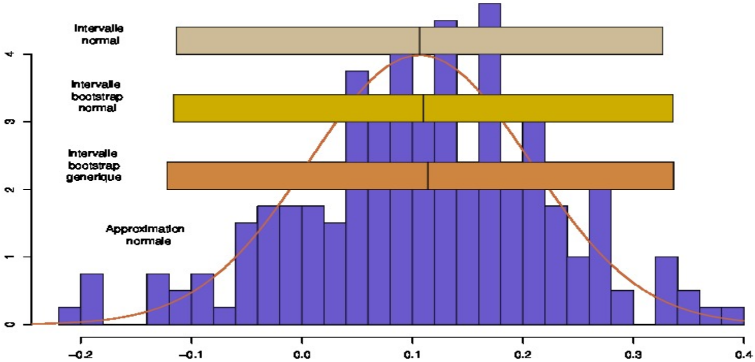
$$[\bar{x}^* - 2 * \hat{\sigma}^*, \bar{x}^* + 2 * \hat{\sigma}^*] = [-0.116, 0.336]$$

[normal bootstrap approximation]

$$[q^*(0.025), q^*(0.975)] = [-0.112, 0.336]$$

[generic bootstrap approximation]

example: mean bootstrap variation



Variation ranges at 95% for a sample of 100 points and 200 bootstrap replications

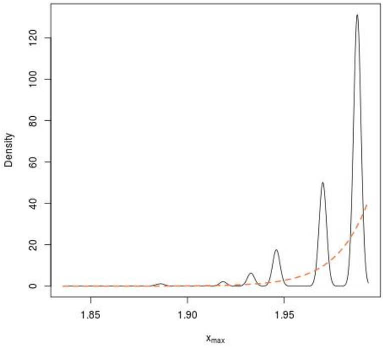
a counter-example

Consider $X_1, \ldots, X_n \sim \mathcal{U}(0, \theta)$ then

$$\theta = \theta(F) = \mathbb{E}_\theta \left[\frac{n}{n-1} X_{(n)} \right]$$

Using bootstrap, distribution of ${}^{n-1}/_{n}\,\theta(\widehat{F}_n)$ far from truth

$$f_{\max}(x) = {}^{n\,x^{n-1}}/_{\theta^n}\, \mathbb{I}_{(0,\theta)}(x)$$



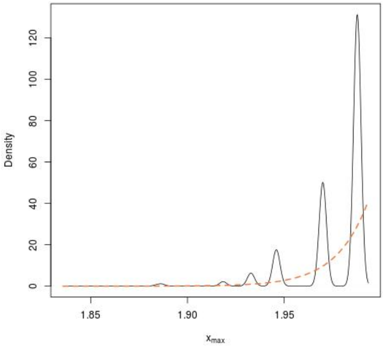
a counter-example

Consider $X_1, \ldots, X_n \sim \mathcal{U}(0, \theta)$ then

$$\theta = \theta(F) = \mathbb{E}_\theta \left[\frac{n}{n-1} X_{(n)} \right]$$

Using bootstrap, distribution of ${}^{n-1}/_{n}\theta(\widehat{F}_n)$ far from truth

$$f_{\max}(x) = {}^{n\,x^{n-1}}/_{\theta^n}\, \mathbb{I}_{(0,\theta)}(x)$$



Parametric Bootstrap

If the parametric shape of F is known,

$$F(\cdot) = \Phi_\lambda(\cdot) \qquad \lambda \in \Lambda\,,$$

an evaluation of F more efficient than $\hat{F}_n$ is provided by

$$\Phi_{\hat{\lambda}_n}$$

where $\hat{\lambda}_n$ is a convergent estimator of λ

[Cf Example 3]

Parametric Bootstrap

If the parametric shape of F is known,

$$F(\cdot) = \Phi_\lambda(\cdot) \qquad \lambda \in \Lambda\,,$$

an evaluation of F more efficient than $\hat{F}_n$ is provided by

$$\Phi_{\hat{\lambda}_n}$$

where $\hat{\lambda}_n$ is a convergent estimator of λ

[Cf Example 3]

Parametric Bootstrap

Approximation of the distribution of

$$\theta(X_1, \ldots, X_n)$$

by the distribution of

$$\theta(X_1^*, \ldots, X_n^*) \qquad X_1^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \Phi_{\hat{\lambda}_n}$$

May avoid Monte Carlo simulation approximations in some cases

Parametric Bootstrap

Approximation of the distribution of

$$\theta(X_1, \ldots, X_n)$$

by the distribution of

$$\theta(X_1^*, \ldots, X_n^*) \qquad X_1^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \Phi_{\hat{\lambda}_n}$$

May avoid Monte Carlo simulation approximations in some cases

example of parametric Bootstrap

Example (Exponential Sample)

Take

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{E}\text{xp}(\lambda)$$

and $\lambda = 1/\mathbb{E}_\lambda[X]$ to be estimated

A possible estimator is

$$\hat{\lambda}(x_1, \ldots, x_n) = \frac{n}{\sum_{i=1}^n x_i}$$

but this estimator is biased

$$\mathbb{E}_\lambda[\hat{\lambda}(X_1, \ldots, X_n)] \neq \lambda$$

example of parametric Bootstrap

Example (Exponential Sample)

Take

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{E}\text{xp}(\lambda)$$

and $\lambda = 1/E_\lambda[X]$ to be estimated
A possible estimator is

$$\hat{\lambda}(x_1, \ldots, x_n) = \frac{n}{\sum_{i=1}^n x_i}$$

but this estimator is biased

$$E_\lambda[\hat{\lambda}(X_1, \ldots, X_n)] \neq \lambda$$

Example (**Exponential Sample (2)**)

Questions :

- What is the bias

$$\lambda - E_\lambda[\hat{\lambda}(X_1, \ldots, X_n)]$$

  of this estimator ?
- What is the distribution of this estimator ?

Example (**Exponential Sample (2)**)

Questions :

- What is the bias

$$\lambda - E_\lambda[\hat{\lambda}(X_1, \ldots, X_n)]$$

  of this estimator ?
- What is the distribution of this estimator ?

Bootstrap evaluation of the bias

Example (**Exponential Sample (3)**)

$$\hat{\lambda}(x_1, \ldots, x_n) - \mathrm{E}_{\hat{\lambda}(x_1,\ldots,x_n)}[\hat{\lambda}(X_1, \ldots, X_n)]$$

[parametric version]

$$\hat{\lambda}(x_1, \ldots, x_n) - \mathrm{E}_{\hat{F}_n}[\hat{\lambda}(X_1, \ldots, X_n)]$$

[non-parametric version]

example: bootstrap bias evaluation

Example (**Exponential Sample (4)**)

In the first (parametric) version,

$$1/\hat{\lambda}(X_1, \ldots, X_n) \sim \mathcal{G}a(n, n\lambda)$$

and

$$\mathbb{E}_\lambda[\hat{\lambda}(X_1, \ldots, X_n)] = \frac{n}{n-1}\lambda$$

therefore the bias is **analytically** evaluated as

$$-\lambda/n-1$$

and estimated by

$$-\frac{\hat{\lambda}(X_1, \ldots, X_n)}{n-1} = -0.00787$$

example: bootstrap bias evaluation

Example (**Exponential Sample (4)**)

In the first (parametric) version,

$$1/\hat{\lambda}(X_1, \ldots, X_n) \sim \mathcal{G}a(n, n\lambda)$$

and

$$\mathbb{E}_\lambda[\hat{\lambda}(X_1, \ldots, X_n)] = \frac{n}{n-1}\lambda$$

therefore the bias is **analytically** evaluated as

$$-\lambda/n - 1$$

and estimated by

$$-\frac{\hat{\lambda}(X_1, \ldots, X_n)}{n-1} = -0.00787$$

example: bootstrap bias evaluation

Example (**Exponential Sample (5)**)

In the second (nonparametric) version, evaluation by Monte Carlo,

$$\hat{\lambda}(x_1, \ldots, x_n) - \mathbb{E}_{\hat{F}_n}[\hat{\lambda}(X_1, \ldots, X_n)] = 0.00142$$

which achieves the **"wrong"** sign

Example (**Exponential Sample (6)**)

Construction of a confidence interval on λ

By parametric bootstrap,

$$\Pr_\lambda\left(\hat{\lambda}_1 \leqslant \lambda \leqslant \hat{\lambda}_2\right) = \Pr\left(\omega_1 \leqslant \lambda/\hat{\lambda} \leqslant \omega_2\right) = 0.95$$

can be deduced from

$$\lambda/\hat{\lambda} \sim \mathcal{G}a(n, n)$$

[In R, qgamma(0.975,n,1/n)]

$$[\hat{\lambda}_1, \hat{\lambda}_2] = [0.452, 0.580]$$

example: bootstrap bias evaluation

Example (Exponential Sample (7))

In nonarametric bootstrap, one replaces

$$\Pr_F\left(q(.025) \leqslant \lambda(F) \leqslant q(.975)\right) = 0.95$$

with

$$\Pr_{\hat{F}_n}\left(q^*(.025) \leqslant \lambda(\hat{F}_n) \leqslant q^*(.975)\right) = 0.95$$

Approximation of quantiles $q^*(.025)$ and $q^*(.975)$ of $\lambda(\hat{F}_n)$ by bootstrap (Monte Carlo) sampling

$$[q^*(.025), q^*(.975)] = [0.454, 0.576]$$

example: bootstrap bias evaluation

Example (Exponential Sample (7))

In nonarametric bootstrap, one replaces

$$\Pr_F\left(q(.025) \leqslant \lambda(F) \leqslant q(.975)\right) = 0.95$$

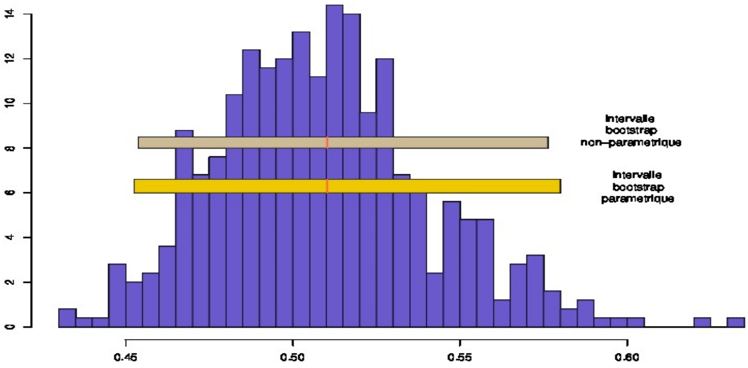
with

$$\Pr_{\hat{F}_n}\left(q^*(.025) \leqslant \lambda(\hat{F}_n) \leqslant q^*(.975)\right) = 0.95$$

Approximation of quantiles $q^*(.025)$ and $q^*(.975)$ of $\lambda(\hat{F}_n)$ by bootstrap (Monte Carlo) sampling

$$[q^*(.025), q^*(.975)] = [0.454, 0.576]$$

example: bootstrap bias evaluation



example: bootstrap distribution evaluation

Example (Student Sample)

Take

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathfrak{T}(5, \mu, \tau^2) \overset{\text{def}}{=} \mu + \tau \frac{\mathcal{N}(0,1)}{\sqrt{\chi_5^2/5}}$$

μ and τ could be estimated by

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \begin{aligned} \hat{\tau}_n &= \sqrt{\frac{5-2}{5}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2} \\ &= \sqrt{\frac{5-2}{5}}\,\hat{\sigma}_n \end{aligned}$$

example: bootstrap distribution evaluation

Example (**Student Sample (2)**)

Problem $\hat{\mu}_n$ is not distributed from a Student $\mathfrak{T}(5, \mu, \tau^2/n)$ distribution

The distribution of $\hat{\mu}_n$ can be reproduced by bootstrap sampling

example: bootstrap distribution evaluation

Example (**Student Sample (3)**)

Comparison of confidence intervals

$$[\hat{\mu}_n - 2 * \hat{\sigma}_n/10, \hat{\mu}_n + 2 * \hat{\sigma}_n/10] = [-0.068, 0.319]$$

[normal approximation]

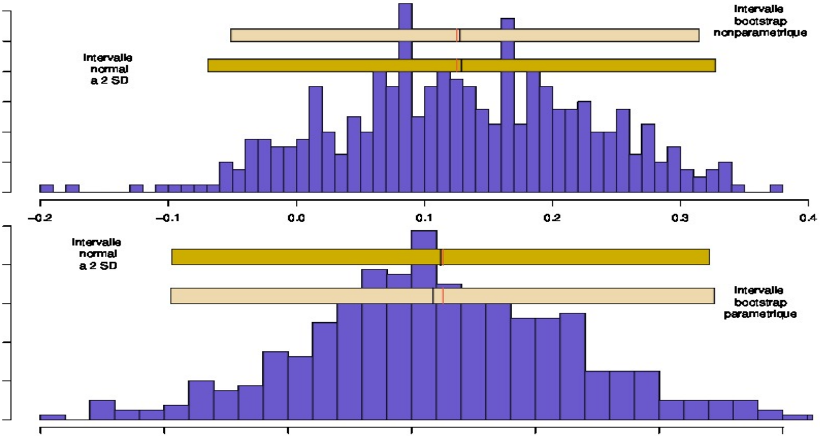
$$[q^*(0.05), q^*(0.95)] = [-0.056, 0.305]$$

[parametric boostrap approximation]

$$[q^*(0.05), q^*(0.95)] = [-0.094, 0.344]$$

[non parametric boostrap approximation]

example: bootstrap distribution evaluation



95% **variation interval for a** 150 **points sample with** 400 **bootstrap replicas** *(top)* **nonparametric and** *(bottom)* **parametric**

Chapter 3 :
Likelihood function and inference

4. Likelihood function and inference
    - The likelihood
    - Information and curvature
    - Sufficiency and ancilarity
    - Maximum likelihood estimation
    - Non-regular models
    - EM algorithm

The likelihood

Given an usually parametric family of distributions

$$F \in \{F_\theta,\ \theta \in \Theta\}$$

with densities f_θ [wrt a fixed measure ν], the density of the iid sample $x_1, \ldots, x_n$ is

$$\prod_{i=1}^{n} f_\theta(x_i)$$

Note In the special case ν is a counting measure,

$$\prod_{i=1}^{n} f_\theta(x_i)$$

is the probability of observing the sample $x_1, \ldots, x_n$ among all possible realisations of $X_1, \ldots, X_n$

The likelihood

Given an usually parametric family of distributions

$$F \in \{F_\theta,\ \theta \in \Theta\}$$

with densities f_θ [wrt a fixed measure ν], the density of the iid sample $x_1, \ldots, x_n$ is

$$\prod_{i=1}^{n} f_\theta(x_i)$$

Note In the special case ν is a counting measure,

$$\prod_{i=1}^{n} f_\theta(x_i)$$

is the probability of observing the sample $x_1, \ldots, x_n$ among all possible realisations of $X_1, \ldots, X_n$

The likelihood

Definition (likelihood function)

The likelihood function associated with a sample $x_1, \dots, x_n$ is the function

$$\begin{aligned} L : \Theta &\longrightarrow \mathbb{R}_+ \\ \theta &\longrightarrow \prod_{i=1}^{n} f_\theta(x_i) \end{aligned}$$

same formula as density but different space of variation

The likelihood

Definition (likelihood function)

The likelihood function associated with a sample $x_1, \dots, x_n$ is the function

$$\begin{aligned} L : \Theta &\longrightarrow \mathbb{R}_+ \\ \theta &\longrightarrow \prod_{i=1}^{n} f_\theta(x_i) \end{aligned}$$

same formula as density but different space of variation

Example: density function versus likelihood function

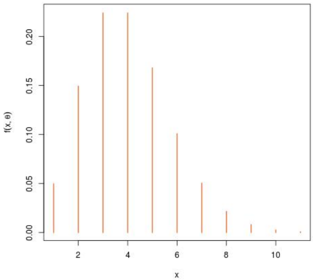
Take the case of a Poisson density
[against the counting measure]

$$f(x;\theta) = \frac{\theta^x}{x!}\, e^{-\theta}\, \mathbb{I}_{\mathbb{N}}(x)$$

which varies in $\mathbb{N}$ as a function of x
versus

$$L(\theta;x) = \frac{\theta^x}{x!}\, e^{-\theta}$$

which varies in $\mathbb{R}_+$ as a function of θ



$\theta = 3$

Example: density function versus likelihood function

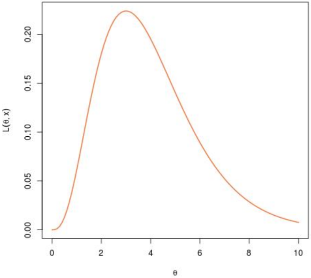
Take the case of a Poisson density
[against the counting measure]

$$f(x;\theta) = \frac{\theta^x}{x!}\, e^{-\theta}\, \mathbb{I}_{\mathbb{N}}(x)$$

which varies in $\mathbb{N}$ as a function of x
versus

$$L(\theta;x) = \frac{\theta^x}{x!}\, e^{-\theta}$$

which varies in $\mathbb{R}_+$ as a function of θ



$x = 3$

Example: density function versus likelihood function

Take the case of a Normal $\mathcal{N}(0, \theta)$ density [against the Lebesgue measure]

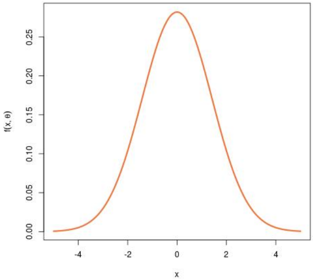
$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-x^2/2\theta}\, \mathbb{I}_{\mathbb{R}}(x)$$

which varies in $\mathbb{R}$ as a function of x

versus

$$L(\theta;x) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-x^2/2\theta}$$

which varies in $\mathbb{R}_+$ as a function of θ



$\theta = 2$

Example: density function versus likelihood function

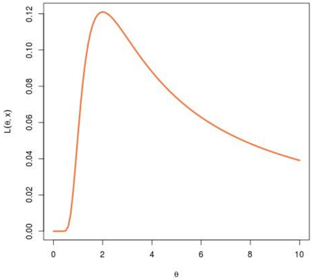
Take the case of a Normal $\mathcal{N}(0, \theta)$ density [against the Lebesgue measure]

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-x^2/2\theta}\, \mathbb{I}_{\mathbb{R}}(x)$$

which varies in $\mathbb{R}$ as a function of x
versus

$$L(\theta;x) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-x^2/2\theta}$$

which varies in $\mathbb{R}_+$ as a function of θ



$x = 2$

Example: density function versus likelihood function

Take the case of a Normal $\mathcal{N}(0, 1/\theta)$ density [against the Lebesgue measure]

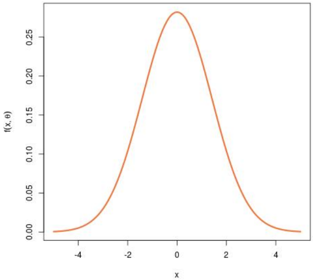
$$f(x;\theta) = \frac{\sqrt{\theta}}{\sqrt{2\pi}}\, e^{-x^2\theta/2}\, \mathbb{I}_{\mathbb{R}}(x)$$

which varies in $\mathbb{R}$ as a function of x

versus

$$L(\theta;x) = \frac{\sqrt{\theta}}{\sqrt{2\pi}}\, e^{-x^2\theta/2}\, \mathbb{I}_{\mathbb{R}}(x)$$

which varies in $\mathbb{R}_+$ as a function of θ



$\theta = 1/2$

Example: density function versus likelihood function

Take the case of a Normal $\mathcal{N}(0, 1/\theta)$ density [against the Lebesgue measure]

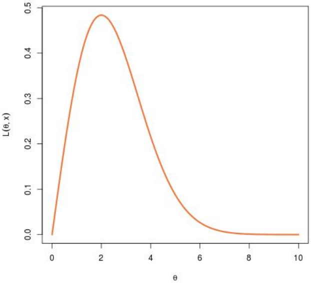
$$f(x;\theta) = \frac{\sqrt{\theta}}{\sqrt{2\pi}}\, e^{-x^2\theta/2}\, \mathbb{I}_{\mathbb{R}}(x)$$

which varies in $\mathbb{R}$ as a function of x

versus

$$L(\theta;x) = \frac{\sqrt{\theta}}{\sqrt{2\pi}}\, e^{-x^2\theta/2}\, \mathbb{I}_{\mathbb{R}}(x)$$

which varies in $\mathbb{R}_+$ as a function of θ



$x = 1/2$

Example: Hardy-Weinberg equilibrium

Population genetics:

- Genotypes of biallelic genes AA, Aa, and aa
- sample frequencies n_{AA}, n_{Aa} and n_{aa}
- multinomial model $\mathcal{M}(n; p_{AA}, p_{Aa}, p_{aa})$
- related to population proportion of A alleles, p_A:

$$p_{AA} = p_A^2\,,\ p_{Aa} = 2p_A(1-p_A)\,,\ p_{aa} = (1-p_A)^2$$

- likelihood

$$L(p_A|n_{AA}, n_{Aa}, n_{aa}) \propto p_A^{2n_{AA}}[2p_A(1-p_A)]^{n_{Aa}}(1-p_A)^{2n_{aa}}$$

[Boos & Stefanski, 2013]

mixed distributions and their likelihood

Special case when a random variable X may take specific values $a_1, \dots, a_k$ and a continum of values $\mathfrak{A}$

Example: Rainfall at a given spot on a given day may be zero with positive probability p_0 [it did not rain!] or an arbitrary number between 0 and 100 [capacity of measurement container] or 100 with positive probability p_{100} [container full]

mixed distributions and their likelihood

Special case when a random variable X may take specific values $a_1, \ldots, a_k$ and a continum of values $\mathfrak{A}$

Example: Tobit model where $y \sim \mathcal{N}(X^T\beta, \sigma^2)$ but $y^* = y \times \mathbb{I}\{y \geqslant 0\}$ observed

mixed distributions and their likelihood

Special case when a random variable X may take specific values $a_1, \ldots, a_k$ and a continum of values $\mathfrak{A}$

Density of X against composition of two measures, counting and Lebesgue:

$$f_X(a) = \begin{cases} \mathbb{P}_\theta(X = a) & \text{if } a \in \{a_1, \ldots, a_k\} \\ f(a|\theta) & \text{otherwise} \end{cases}$$

Results in likelihood

$$L(\theta|x_1, \ldots, x_n) = \prod_{j=1}^{k} \mathbb{P}_\theta(X = a_i)^{n_j} \times \prod_{x_i \notin \{a_1, \ldots, a_k\}} f(x_i|\theta)$$

where n_j # observations equal to a_j

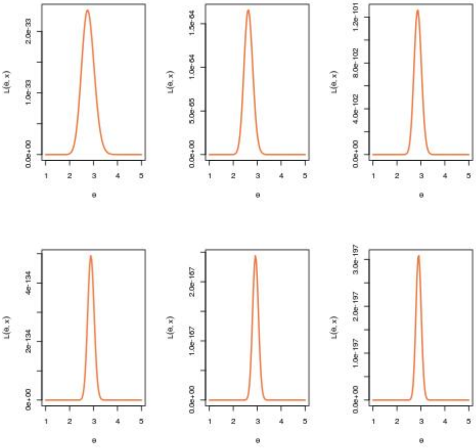
Enters Fisher, Ronald Fisher!

Fisher's intuition in the 20's:

- the likelihood function contains the relevant information about the parameter θ
- the higher the likelihood the more likely the parameter
- the curvature of the likelihood determines the precision of the estimation

Concentration of likelihood mode around "true" parameter

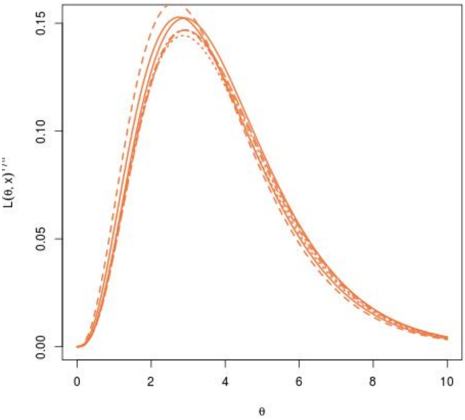
Likelihood functions for $x_1, \ldots, x_n \sim \mathcal{P}(3)$ as n increases



$$n = 40, \ldots, 240$$

Concentration of likelihood mode around "true" parameter

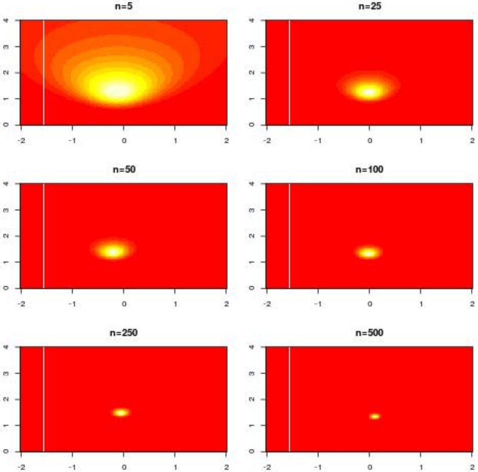
Likelihood functions for $x_1, \ldots, x_n \sim \mathcal{P}(3)$ as n increases



$$n = 38, \ldots, 240$$

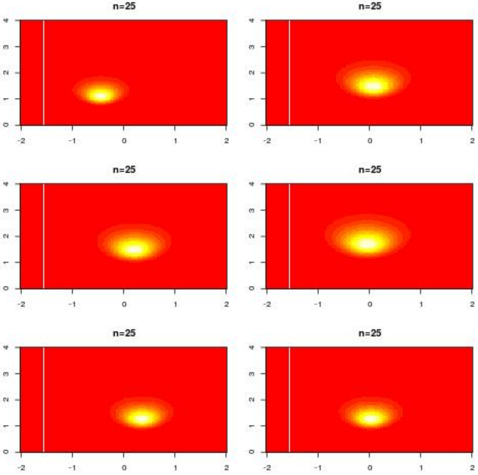
Concentration of likelihood mode around “true” parameter

Likelihood functions for $x_1, \ldots, x_n \sim \mathcal{N}(0, 1)$ as n increases



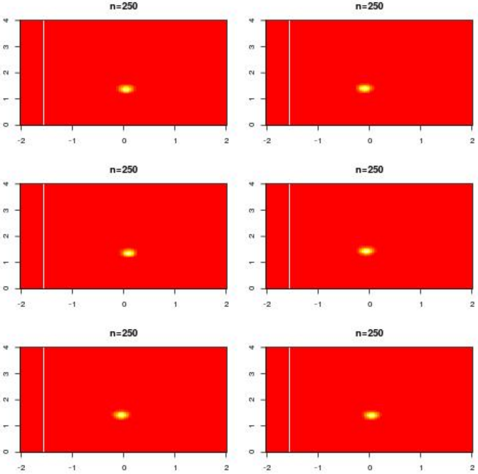
Concentration of likelihood mode around “true” parameter

Likelihood functions for $x_1, \ldots, x_n \sim \mathcal{N}(0, 1)$ as sample varies



Concentration of likelihood mode around "true" parameter

Likelihood functions for $x_1, \ldots, x_n \sim \mathcal{N}(0, 1)$ as sample varies



why concentration takes place

Consider

$$x_1, \ldots, x_n \overset{\text{iid}}{\sim} F$$

Then

$$\log \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

and by LLN

$$1/n \sum_{i=1}^{n} \log f(x_i|\theta) \xrightarrow{\mathcal{L}} \int_{\mathcal{X}} \log f(x|\theta)\, dF(x)$$

Lemma

Maximising the likelihood is asymptotically equivalent to minimising the Kullback-Leibler divergence

$$\int_{\mathcal{X}} \log f(x)/f(x|\theta)\, dF(x)$$

© Member of the family closest to true distribution

why concentration takes place

by LLN

$$^1/_n \sum_{i=1}^{n} \log f(x_i|\theta) \xrightarrow{\mathcal{L}} \int_{\mathcal{X}} \log f(x|\theta)\, dF(x)$$

Lemma

Maximising the likelihood is asymptotically equivalent to minimising the Kullback-Leibler divergence

$$\int_{\mathcal{X}} \log {}^{f(x)}/_{f(x|\theta)}\, dF(x)$$

© Member of the family closest to true distribution

Score function

Score function defined by

$$\nabla \log L(\theta|x) = \left(\partial/\partial\theta_1 L(\theta|x), \ldots, \partial/\partial\theta_p L(\theta|x)\right) \big/ L(\theta|x)$$

Gradient (slope) of likelihood function at point θ

lemma

When $X \sim F_\theta$,

$$\mathbb{E}_\theta[\nabla \log L(\theta|X)] = 0$$

Score function

Score function defined by

$$\nabla \log L(\theta|x) = \left({}^{\partial}\!/\!_{\partial\theta_1} L(\theta|x), \ldots, {}^{\partial}\!/\!_{\partial\theta_p} L(\theta|x)\right) \Big/ L(\theta|x)$$

Gradient (slope) of likelihood function at point θ

lemma

When $X \sim F_\theta$,

$$\mathbb{E}_\theta[\nabla \log L(\theta|X)] = 0$$

Score function

Score function defined by

$$\nabla \log L(\theta|x) = \left(\partial/\partial\theta_1 L(\theta|x), \ldots, \partial/\partial\theta_p L(\theta|x)\right) / L(\theta|x)$$

Gradient (slope) of likelihood function at point θ

lemma

When $X \sim F_\theta$,

$$\mathbb{E}_\theta[\nabla \log L(\theta|X)] = 0$$

Reason:

$$\int_{\mathcal{X}} \nabla \log L(\theta|x)\, dF_\theta(x) = \int_{\mathcal{X}} \nabla L(\theta|x)\, dx = \nabla \int_{\mathcal{X}} dF_\theta(x)$$

Score function

Score function defined by

$$\nabla \log L(\theta|x) = \left(\partial/\partial\theta_1 L(\theta|x), \ldots, \partial/\partial\theta_p L(\theta|x)\right) / L(\theta|x)$$

Gradient (slope) of likelihood function at point θ

lemma

When $X \sim F_\theta$,

$$\mathbb{E}_\theta[\nabla \log L(\theta|X)] = 0$$

Connected with concentration theorem: gradient null on average for true value of parameter

Score function

Score function defined by

$$\nabla \log L(\theta|x) = \left({}^{\partial}\!/\!_{\partial\theta_1} L(\theta|x), \ldots, {}^{\partial}\!/\!_{\partial\theta_p} L(\theta|x) \right) \Big/ L(\theta|x)$$

Gradient (slope) of likelihood function at point θ

lemma

When $X \sim F_\theta$,

$$\mathbb{E}_\theta[\nabla \log L(\theta|X)] = 0$$

Warning: Not defined for non-differentiable likelihoods, e.g. when support depends on θ

Score function

Score function defined by

$$\nabla \log L(\theta|x) = \left({}^{\partial}\!/_{\partial\theta_1} L(\theta|x), \ldots, {}^{\partial}\!/_{\partial\theta_p} L(\theta|x)\right) \big/ L(\theta|x)$$

Gradient (slope) of likelihood function at point θ

lemma

When $X \sim F_\theta$,

$$\mathbb{E}_\theta[\nabla \log L(\theta|X)] = 0$$

Warning (2): Does not imply maximum likelihood estimator is unbiased

Fisher's information matrix

Another notion attributed to Fisher [more likely due to Edgeworth]

Information: covariance matrix of the score vector

$$\mathfrak{I}(\theta) = \mathbb{E}_\theta \left[\nabla \log f(X|\theta) \, \{\nabla \log f(X|\theta)\}^{\mathrm{T}} \right]$$

Often called Fisher information

Measures curvature of the likelihood surface, which translates as information brought by the data

Sometimes denoted $\mathfrak{I}_X$ to stress dependence on distribution of X

Fisher's information matrix

Another notion attributed to Fisher [more likely due to Edgeworth]

Information: covariance matrix of the score vector

$$\mathfrak{I}(\theta) = \mathbb{E}_\theta \left[\nabla \log f(X|\theta) \left\{ \nabla \log f(X|\theta) \right\}^{\mathrm{T}} \right]$$

Often called Fisher information

Measures curvature of the likelihood surface, which translates as information brought by the data

Sometimes denoted $\mathfrak{I}_X$ to stress dependence on distribution of X

Fisher's information matrix

Second derivative of the log-likelihood as well

lemma

If $\mathrm{L}(\theta|x)$ is twice differentiable [as a function of θ]

$$\mathfrak{I}(\theta) = -\mathbb{E}_\theta\left[\nabla^{\mathrm{T}}\nabla \log f(X|\theta)\right]$$

Hence

$$\mathfrak{I}_{ij}(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X|\theta)\right]$$

Illustrations

Binomial $\mathcal{B}(n, p)$ distribution

$$
\begin{aligned}
f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} \\
\partial/\partial p \log f(x|p) &= x/p - {n-x}/{1-p} \\
\partial^2/\partial p^2 \log f(x|p) &= -x/p^2 - {n-x}/{(1-p)^2}
\end{aligned}
$$

Hence

$$
\begin{aligned}
\mathfrak{I}(p) &= {np}/{p^2} + {n-np}/{(1-p)^2} \\
&= n/{p(1-p)}
\end{aligned}
$$

Illustrations

Multinomial $\mathcal{M}(n; p_1, \ldots, p_k)$ distribution

$$
\begin{aligned}
f(x|p) &= \binom{n}{x_1 \cdots x_k} p_1^{x_1} \cdots p_k^{x_k} \\
{}^{\partial}/_{\partial p_i} \log f(x|p) &= {}^{x_i}/_{p_i} - {}^{x_k}/_{p_k} \\
{}^{\partial^2}/_{\partial p_i \partial p_j} \log f(x|p) &= -\,{}^{x_k}/_{p_k^2} \\
{}^{\partial^2}/_{\partial p_i^2} \log f(x|p) &= -\,{}^{x_i}/_{p_i^2} - {}^{x_k}/_{p_k^2}
\end{aligned}
$$

Hence

$$
\mathfrak{I}(p) = n \begin{pmatrix} 1/p_1 + 1/p_k & \cdots & 1/p_k \\ 1/p_k & \cdots & 1/p_k \\ & \ddots & \\ 1/p_k & \cdots & 1/p_{k-1} + 1/p_k \end{pmatrix}
$$

Illustrations

Multinomial $\mathcal{M}(n; p_1, \ldots, p_k)$ distribution

$$
\begin{aligned}
f(x|p) &= \binom{n}{x_1 \cdots x_k} p_1^{x_1} \cdots p_k^{x_k} \\
\partial/\partial p_i \log f(x|p) &= x_i/p_i - x_k/p_k \\
\partial^2/\partial p_i \partial p_j \log f(x|p) &= -x_k/p_k^2 \\
\partial^2/\partial p_i^2 \log f(x|p) &= -x_i/p_i^2 - x_k/p_k^2
\end{aligned}
$$

and

$$
\mathfrak{I}(p)^{-1} = 1/n \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{k-1} \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_{k-1} \\ & \ddots & \ddots & \\ -p_1p_{k-1} & -p_2p_{k-1} & \cdots & p_{k-1}(1-p_{k-1}) \end{pmatrix}
$$

Illustrations

Normal $\mathcal{N}(\mu, \sigma^2)$ distribution

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left\{-{}^{(x-\mu)^2}/{}_{2\sigma^2}\right\} \quad {}^{\partial}/{}_{\partial\mu} \log f(x|\theta) = {}^{x-\mu}/{}_{\sigma^2}$$

$$ {}^{\partial}/{}_{\partial\sigma} \log f(x|\theta) = -{}^{1}/{}_{\sigma} + {}^{(x-\mu)^2}/{}_{\sigma^3} \quad {}^{\partial^2}/{}_{\partial\mu^2} \log f(x|\theta) = -{}^{1}/{}_{\sigma^2}$$

$${}^{\partial^2}/{}_{\partial\mu\partial\sigma} \log f(x|\theta) = -2\,{}^{x-\mu}/{}_{\sigma^3} \quad {}^{\partial^2}/{}_{\partial\sigma^2} \log f(x|\theta) = {}^{1}/{}_{\sigma^2} - 3\,{}^{(x-\mu)^2}/{}_{\sigma^4}$$

Hence

$$\mathfrak{I}(\theta) = {}^{1}/{}_{\sigma^2} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

Properties

Additive features translating as accumulation of information:

- if X and Y are independent, $\mathfrak{I}_X(\theta) + \mathfrak{I}_Y(\theta) = \mathfrak{I}_{(X,Y)}(\theta)$
- $\mathfrak{I}_{X_1,\ldots,X_n}(\theta) = n\mathfrak{I}_{X_1}(\theta)$
- if $X = T(Y)$ and $Y = S(X)$, $\mathfrak{I}_X(\theta) = \mathfrak{I}_Y(\theta)$
- if $X = T(Y)$, $\mathfrak{I}_X(\theta) \leqslant \mathfrak{I}_Y(\theta)$

If $\eta = \Psi(\theta)$ is a bijective transform, change of parameterisation:

$$\mathfrak{I}(\theta) = \left\{\frac{\partial \eta}{\partial \theta}\right\}^{\mathrm{T}} \mathfrak{I}(\eta) \left\{\frac{\partial \eta}{\partial \theta}\right\}$$

"In information geometry, this is seen as a change of coordinates on a Riemannian manifold, and the intrinsic properties of curvature are unchanged under different parametrizations. In general, the Fisher information matrix provides a Riemannian metric (more precisely, the Fisher-Rao metric)." [Wikipedia]

Properties

Additive features translating as accumulation of information:

- if X and Y are independent, $\mathfrak{I}_X(\theta) + \mathfrak{I}_Y(\theta) = \mathfrak{I}_{(X,Y)}(\theta)$
- $\mathfrak{I}_{X_1,\ldots,X_n}(\theta) = n\mathfrak{I}_{X_1}(\theta)$
- if $X = T(Y)$ and $Y = S(X)$, $\mathfrak{I}_X(\theta) = \mathfrak{I}_Y(\theta)$
- if $X = T(Y)$, $\mathfrak{I}_X(\theta) \leqslant \mathfrak{I}_Y(\theta)$

If $\eta = \Psi(\theta)$ is a bijective transform, change of parameterisation:

$$\mathfrak{I}(\theta) = \left\{\frac{\partial\eta}{\partial\theta}\right\}^{\mathrm{T}} \mathfrak{I}(\eta) \left\{\frac{\partial\eta}{\partial\theta}\right\}$$

"In information geometry, this is seen as a change of coordinates on a Riemannian manifold, and the intrinsic properties of curvature are unchanged under different parametrizations. In general, the Fisher information matrix provides a Riemannian metric (more precisely, the Fisher-Rao metric)." [Wikipedia]

Properties

If $\eta = \Psi(\theta)$ is a bijective transform, change of parameterisation:

$$\mathfrak{I}(\theta) = \left\{ \frac{\partial \eta}{\partial \theta} \right\}^{\mathrm{T}} \mathfrak{I}(\eta) \left\{ \frac{\partial \eta}{\partial \theta} \right\}$$

"In information geometry, this is seen as a change of coordinates on a Riemannian manifold, and the intrinsic properties of curvature are unchanged under different parametrizations. In general, the Fisher information matrix provides a Riemannian metric (more precisely, the Fisher-Rao metric)." [Wikipedia]

Approximations

Back to the Kullback–Leibler divergence

$$\mathfrak{D}(\theta', \theta) = \int_{\mathcal{X}} f(x|\theta') \log {}^{f(x|\theta')}/_{f(x|\theta)} \, dx$$

Using a second degree Taylor expansion

$$\log f(x|\theta) = \log f(x|\theta') + (\theta - \theta')^{T} \nabla \log f(x|\theta') \\ + \frac{1}{2}(\theta - \theta')^{T} \nabla\nabla^{T} \log f(x|\theta')(\theta - \theta') + o(||\theta - \theta'||^{2})$$

approximation of divergence:

$$\mathfrak{D}(\theta', \theta) \approx \frac{1}{2}(\theta - \theta')^{T} \mathfrak{I}(\theta')(\theta - \theta')$$

[Exercise: show this is exact in the normal case]

Approximations

Back to the Kullback–Leibler divergence

$$\mathfrak{D}(\theta', \theta) = \int_{\mathcal{X}} f(x|\theta') \log {}^{f(x|\theta')}/_{f(x|\theta)} \, dx$$

Using a second degree Taylor expansion

$$\begin{aligned}\log f(x|\theta) = \log f(x|\theta') + (\theta - \theta')^{\mathrm{T}} \nabla \log f(x|\theta') \\ + \frac{1}{2}(\theta - \theta')^{\mathrm{T}} \nabla\nabla^{\mathrm{T}} \log f(x|\theta')(\theta - \theta') + o(\|\theta - \theta'\|^2)\end{aligned}$$

approximation of divergence:

$$\mathfrak{D}(\theta', \theta) \approx \frac{1}{2}(\theta - \theta')^{\mathrm{T}} \mathfrak{I}(\theta')(\theta - \theta')$$

[Exercise: show this is exact in the normal case]

Approximations

Back to the Kullback–Leibler divergence

$$\mathfrak{D}(\theta', \theta) = \int_{\mathcal{X}} f(x|\theta') \log {}^{f(x|\theta')}/_{f(x|\theta)} \, \mathrm{d}x$$

Using a second degree Taylor expansion

$$\begin{aligned}\log f(x|\theta) = \log f(x|\theta') &+ (\theta - \theta')^{\mathrm{T}} \nabla \log f(x|\theta') \\ &+ \frac{1}{2}(\theta - \theta')^{\mathrm{T}} \nabla\nabla^{\mathrm{T}} \log f(x|\theta')(\theta - \theta') + \mathrm{o}(\|\theta - \theta'\|^2)\end{aligned}$$

approximation of divergence:

$$\mathfrak{D}(\theta', \theta) \approx \frac{1}{2}(\theta - \theta')^{\mathrm{T}} \mathfrak{I}(\theta')(\theta - \theta')$$

[Exercise: show this is exact in the normal case]

First CLT

Central limit law of the score vector
Given $X_1, \ldots, X_n$ i.i.d. $f(x|\theta)$,

$$^1/\sqrt{n}\nabla \log L(\theta|X_1, \ldots, X_n) \approx \mathcal{N}\left(0, \mathfrak{I}_{X_1}(\theta)\right)$$

[at the "true" θ]

Notation $\mathfrak{I}_1(\theta)$ stands for $\mathfrak{I}_{X_1}(\theta)$ and indicates information associated with a single observation

First CLT

Central limit law of the score vector
Given $X_1, \ldots, X_n$ i.i.d. $f(x|\theta)$,

$$^1/\sqrt{n}\nabla \log L(\theta|X_1, \ldots, X_n) \approx \mathcal{N}\left(0, \mathfrak{I}_{X_1}(\theta)\right)$$

[at the "true" θ]

Notation $\mathfrak{I}_1(\theta)$ stands for $\mathfrak{I}_{X_1}(\theta)$ and indicates information associated with a single observation

Sufficiency

What if a transform of the sample

$$S(X_1, \dots, X_n)$$

contains all the information, i.e.

$$\mathfrak{I}_{(X_1,\dots,X_n)}(\theta) = \mathfrak{I}_{S(X_1,\dots,X_n)}(\theta)$$

uniformly in θ?

In this case $S(\cdot)$ is called a sufficient statistic [because it is sufficient to know the value of $S(x_1, \dots, x_n)$ to get complete information]

[A statistic is an arbitrary transform of the data $X_1, \dots, X_n$]

Sufficiency

What if a transform of the sample

$$S(X_1, \dots, X_n)$$

contains all the information, i.e.

$$\mathfrak{I}_{(X_1,\dots,X_n)}(\theta) = \mathfrak{I}_{S(X_1,\dots,X_n)}(\theta)$$

uniformly in θ?

In this case $S(\cdot)$ is called a sufficient statistic [because it is sufficient to know the value of $S(x_1, \dots, x_n)$ to get complete information]

[A statistic is an arbitrary transform of the data $X_1, \dots, X_n$]

Sufficiency

What if a transform of the sample

$$S(X_1, \ldots, X_n)$$

contains all the information, i.e.

$$\mathfrak{I}_{(X_1,\ldots,X_n)}(\theta) = \mathfrak{I}_{S(X_1,\ldots,X_n)}(\theta)$$

uniformly in θ?

In this case $S(\cdot)$ is called a sufficient statistic [because it is sufficient to know the value of $S(x_1, \ldots, x_n)$ to get complete information]

[A statistic is an arbitrary transform of the data $X_1, \ldots, X_n$]

Sufficiency (bis)

Alternative definition:

If $(X_1, \ldots, X_n) \sim f(x_1, \ldots, x_n|\theta)$ and if $T = S(X_1, \ldots, X_n)$ is such that the distribution of $(X_1, \ldots, X_n)$ conditional on T does not depend on θ, then $S(\cdot)$ is a sufficient statistic

Factorisation theorem

$S(\cdot)$ is a sufficient statistic if and only if

$$f(x_1, \ldots, x_n|\theta) = g(S(x_1, \ldots, x_n)|\theta) \times h(x_1, \ldots, x_n)$$

another notion due to Fisher

Sufficiency (bis)

Alternative definition:

If $(X_1, \ldots, X_n) \sim f(x_1, \ldots, x_n|\theta)$ and if $T = S(X_1, \ldots, X_n)$ is such that the distribution of $(X_1, \ldots, X_n)$ conditional on T does not depend on θ, then $S(\cdot)$ is a sufficient statistic

Factorisation theorem

$S(\cdot)$ is a sufficient statistic if and only if

$$f(x_1, \ldots, x_n|\theta) = g(S(x_1, \ldots, x_n)|\theta) \times h(x_1, \ldots, x_n)$$

another notion due to Fisher

Sufficiency (bis)

Alternative definition:

If $(X_1, \ldots, X_n) \sim f(x_1, \ldots, x_n|\theta)$ and if $T = S(X_1, \ldots, X_n)$ is such that the distribution of $(X_1, \ldots, X_n)$ conditional on T does not depend on θ, then $S(\cdot)$ is a sufficient statistic

Factorisation theorem

$S(\cdot)$ is a sufficient statistic if and only if

$$f(x_1, \ldots, x_n|\theta) = g(S(x_1, \ldots, x_n)|\theta) \times h(x_1, \ldots, x_n)$$

another notion due to Fisher

Illustrations

Uniform $\mathcal{U}(0,\theta)$ distribution

$$L(\theta|x_1,\ldots,x_n) = \theta^{-n}\prod_{i=1}^{n}\mathbb{I}_{(0,\theta)}(x_i) = \theta^{-n}\mathbb{I}_{\theta>\max\limits_i x_i}$$

Hence

$$S(X_1,\ldots,X_n) = \max_i X_i = X_{(n)}$$

is sufficient

Illustrations

Bernoulli $\mathcal{B}(p)$ distribution

$$L(p|x_1, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{n-x_i} = \{p/1-p\}^{\sum_i x_i} (1-p)^n$$

Hence

$$S(X_1, \ldots, X_n) = \overline{X}_n$$

is sufficient

Illustrations

Normal $\mathcal{N}(\mu, \sigma^2)$ distribution

$$
\begin{aligned}
L(\mu, \sigma | x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-{}^{(x_i-\mu)^2}/{}_{2\sigma^2}\} \\
&= \frac{1}{\{2\pi\sigma^2\}^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x}_n + \bar{x}_n - \mu)^2\right\} \\
&= \frac{1}{\{2\pi\sigma^2\}^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\bar{x}_n - \mu)^2\right\}
\end{aligned}
$$

Hence

$$
S(X_1, \ldots, X_n) = \left(\overline{X}_n, \sum_{i=1}^{n} (X_i - \overline{X}_n)^2\right)
$$

is sufficient

Sufficiency and exponential families

Both previous examples belong to exponential families

$$f(x|\theta) = h(x) \exp\left\{T(\theta)^T S(x) - \tau(\theta)\right\}$$

Generic property of exponential families:

$$f(x_1, \dots, x_n|\theta) = \prod_{i=1}^n h(x_i) \exp\left\{T(\theta)^T \sum_{i=1}^n S(x_i) - n\tau(\theta)\right\}$$

lemma

For an exponential family with summary statistic $S(\cdot)$, the statistic

$$S(X_1, \dots, X_n) = \sum_{i=1}^n S(X_i)$$

is sufficient

Sufficiency and exponential families

Both previous examples belong to exponential families

$$f(x|\theta) = h(x) \exp\left\{T(\theta)^{\mathrm{T}} S(x) - \tau(\theta)\right\}$$

Generic property of exponential families:

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} h(x_i) \exp\left\{T(\theta)^{\mathrm{T}} \sum_{i=1}^{n} S(x_i) - n\tau(\theta)\right\}$$

lemma

For an exponential family with summary statistic $S(\cdot)$, the statistic

$$S(X_1, \ldots, X_n) = \sum_{i=1}^{n} S(X_i)$$

is sufficient

Sufficiency as a rare feature

Nice property reducing the data to a low dimension transform but...

How frequent is it within the collection of probability distributions?

Very rare as essentially restricted to exponential families

[Pitman-Koopman-Darmois theorem]

with the exception of parameter-dependent families like $\mathcal{U}(0, \theta)$

Sufficiency as a rare feature

Nice property reducing the data to a low dimension transform but...

How frequent is it within the collection of probability distributions?

Very rare as essentially restricted to exponential families

[Pitman-Koopman-Darmois theorem]

with the exception of parameter-dependent families like $\mathcal{U}(0, \theta)$

Sufficiency as a rare feature

Nice property reducing the data to a low dimension transform but...

How frequent is it within the collection of probability distributions?

Very rare as essentially restricted to exponential families

[Pitman-Koopman-Darmois theorem]

with the exception of parameter-dependent families like $\mathcal{U}(0, \theta)$

Pitman-Koopman-Darmois characterisation

If $X_1, \ldots, X_n$ *are iid random variables from a density* $f(\cdot|\theta)$ *whose support does not depend on* θ *and verifying the property that there exists an integer* n_0 *such that, for* $n \geqslant n_0$*, there is a sufficient statistic* $S(X_1, \ldots, X_n)$ *with fixed [in* n*] dimension, then* $f(\cdot|\theta)$ *belongs to an exponential family*

[Factorisation theorem]

Note: Darmois published this result in 1935 [in French] and Koopman and Pitman in 1936 [in English] but Darmois is generally omitted from the theorem... Fisher proved it for one-D sufficient statistics in 1934

Pitman-Koopman-Darmois characterisation

If $X_1, \ldots, X_n$ *are iid random variables from a density* $f(\cdot|\theta)$ *whose support does not depend on* θ *and verifying the property that there exists an integer* n_0 *such that, for* $n \geqslant n_0$*, there is a sufficient statistic* $S(X_1, \ldots, X_n)$ *with fixed [in* n*] dimension, then* $f(\cdot|\theta)$ *belongs to an exponential family*

[Factorisation theorem]

Note: Darmois published this result in 1935 [in French] and Koopman and Pitman in 1936 [in English] but Darmois is generally omitted from the theorem... Fisher proved it for one-D sufficient statistics in 1934

Minimal sufficiency

Multiplicity of sufficient statistics, e.g., $S'(x) = (S(x), U(x))$ remains sufficient when $S(\cdot)$ is sufficient

Search of a most concentrated summary:

Minimal sufficiency

A sufficient statistic $S(\cdot)$ is minimal sufficient if it is a function of any other sufficient statistic

Lemma

For a minimal exponential family representation

$$f(x|\theta) = h(x)\,\exp\left\{T(\theta)^{\mathrm{T}}S(x) - \tau(\theta)\right\}$$

$S(X_1) + \ldots + S(X_n)$ is minimal sufficient

Minimal sufficiency

Multiplicity of sufficient statistics, e.g., $S'(x) = (S(x), U(x))$ remains sufficient when $S(\cdot)$ is sufficient

Search of a most concentrated summary:

Minimal sufficiency

A sufficient statistic $S(\cdot)$ is minimal sufficient if it is a function of any other sufficient statistic

Lemma

For a minimal exponential family representation

$$f(x|\theta) = h(x) \exp\left\{T(\theta)^{\mathrm{T}} S(x) - \tau(\theta)\right\}$$

$S(X_1) + \ldots + S(X_n)$ is minimal sufficient

Ancillarity

Opposite of sufficiency:

Ancillarity

When $X_1, \ldots, X_n$ are iid random variables from a density $f(\cdot|\theta)$, a statistic $A(\cdot)$ is ancillary if $A(X_1, \ldots, X_n)$ has a distribution that does not depend on θ

Useless?! Not necessarily, as conditioning upon $A(X_1, \ldots, X_n)$ leads to more precision and efficiency:

Use of $F_\theta(x_1, \ldots, x_n | A(x_1, \ldots, x_n))$ instead of $F_\theta(x_1, \ldots, x_n)$

Notion of maximal ancillary statistic

Ancillarity

Opposite of sufficiency:

Ancillarity

When $X_1, \ldots, X_n$ are iid random variables from a density $f(\cdot|\theta)$, a statistic $A(\cdot)$ is ancillary if $A(X_1, \ldots, X_n)$ has a distribution that does not depend on θ

Useless?! Not necessarily, as conditioning upon $A(X_1, \ldots, X_n)$ leads to more precision and efficiency:

Use of $F_\theta(x_1, \ldots, x_n | A(x_1, \ldots, x_n))$ instead of $F_\theta(x_1, \ldots, x_n)$

Notion of maximal ancillary statistic

Ancillarity

Opposite of sufficiency:

Ancillarity

When $X_1, \ldots, X_n$ are iid random variables from a density $f(\cdot|\theta)$, a statistic $A(\cdot)$ is ancillary if $A(X_1, \ldots, X_n)$ has a distribution that does not depend on θ

Useless?! Not necessarily, as conditioning upon $A(X_1, \ldots, X_n)$ leads to more precision and efficiency:

Use of $F_\theta(x_1, \ldots, x_n | A(x_1, \ldots, x_n))$ instead of $F_\theta(x_1, \ldots, x_n)$

Notion of maximal ancillary statistic

Illustrations

1. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{U}(0, \theta)$, $A(X_1, \ldots, X_n) = (X_1, \ldots, X_n)/X_{(n)}$ is ancillary
2. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$,

$$A(X_1, \ldots, X_n) = \frac{(X_1 - \overline{X}_n, \ldots, X_n - \overline{X}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2)}$$

   is ancillary
3. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x|\theta)$, $\text{rank}(X_1, \ldots, X_n)$ is ancillary

```
> x=rnorm(10)
> rank(x)
 [1]  7  4  1  5  2  6  8  9 10  3
```

[see, e.g., rank tests]

Basu's theorem

Completeness

When $X_1, \ldots, X_n$ are iid random variables from a density $f(\cdot|\theta)$, a statistic $A(\cdot)$ is complete if the only function Ψ such that $\mathbb{E}_\theta[\Psi(A(X_1, \ldots, X_n))] = 0$ for all θ's is the null function

Let $X = (X_1, \ldots, X_n)$ be a random sample from $f(\cdot|\theta)$ where $\theta \in \Theta$. If V is an ancillary statistic, and T is complete and sufficient for θ then T and V are independent with respect to $f(\cdot|\theta)$ for all $\theta \in \Theta$.

[Basu, 1955]

Basu's theorem

Completeness

When $X_1, \ldots, X_n$ are iid random variables from a density $f(\cdot|\theta)$, a statistic $A(\cdot)$ is complete if the only function Ψ such that $\mathbb{E}_\theta[\Psi(A(X_1, \ldots, X_n))] = 0$ for all θ's is the null function

Let $X = (X_1, \ldots, X_n)$ be a random sample from $f(\cdot|\theta)$ where $\theta \in \Theta$. If V is an ancillary statistic, and T is complete and sufficient for θ then T and V are independent with respect to $f(\cdot|\theta)$ for all $\theta \in \Theta$.

[Basu, 1955]

some examples

Example 1

If $X = (X_1, \ldots, X_n)$ is a random sample from the Normal distribution $\mathcal{N}(\mu, \sigma^2)$ when σ is known, $\bar{X}_n = {}^1\!/\!{}_n \sum_{i=1}^n X_i$ is sufficient and complete, while $(X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n)$ is ancillary, hence independent from $\bar{X}_n$.

counter-Example 2

Let N be an integer-valued random variable with known pdf $(\pi_1, \pi_2, \ldots)$. And let $S|N = n \sim \mathcal{B}(n, p)$ with unknown p. Then (N, S) is minimal sufficient and N is ancillary.

some examples

Example 1

If $X = (X_1, \ldots, X_n)$ is a random sample from the Normal distribution $\mathcal{N}(\mu, \sigma^2)$ when σ is known, $\bar{X}_n = {}^1\!/\!_n \sum_{i=1}^n X_i$ is sufficient and complete, while $(X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n)$ is ancillary, hence independent from $\bar{X}_n$.

counter-Example 2

Let N be an integer-valued random variable with known pdf $(\pi_1, \pi_2, \ldots)$. And let $S|N = n \sim \mathcal{B}(n, p)$ with unknown p. Then (N, S) is minimal sufficient and N is ancillary.

more counterexamples

counter-Example 3

If $X = (X_1, \ldots, X_n)$ is a random sample from the double exponential distribution $f(x|\theta) = 2\exp\{-|x-\theta|\}$, $(X_{(1)}, \ldots, X_{(n)})$ is minimal sufficient but not complete since $X_{(n)} - X_{(1)}$ is ancillary and with fixed expectation.

counter-Example 4

If X is a random variable from the Uniform $\mathcal{U}(\theta, \theta+1)$ distribution, X and $[X]$ are independent, but while X is complete and sufficient, $[X]$ is not ancillary.

more counterexamples

counter-Example 3

If $X = (X_1, \ldots, X_n)$ is a random sample from the double exponential distribution $f(x|\theta) = 2\exp\{-|x-\theta|\}$, $(X_{(1)}, \ldots, X_{(n)})$ is minimal sufficient but not complete since $X_{(n)} - X_{(1)}$ is ancillary and with fixed expectation.

counter-Example 4

If X is a random variable from the Uniform $\mathcal{U}(\theta, \theta+1)$ distribution, X and $[X]$ are independent, but while X is complete and sufficient, $[X]$ is not ancillary.

last counterexample

Let X be distributed as

x	-5	-4	-3	-2	-1	1	2	3	4	5
p_x	$\alpha' p^2 q$	$\alpha' p q^2$	$p^3/2$	$q^3/2$	$\gamma' p q$	$\gamma' p q$	$q^3/2$	$p^3/2$	$\alpha p q^2$	$\alpha p^2 q$

with

$$\alpha + \alpha' = \gamma + \gamma' = 2/3$$

known and $q = 1 - p$. Then

- $T = |X|$ is minimal sufficient
- $V = \mathbb{I}(X > 0)$ is ancillary
- if $\alpha' \neq \alpha$ T and V are not independent
- T is complete for two-valued functions

[Lehmann, 1981]

Point estimation, estimators and estimates

When given a parametric family $f(\cdot|\theta)$ and a sample supposedly drawn from this family

$$(X_1, \ldots, X_N) \overset{\text{iid}}{\sim} f(x|\theta)$$

1. an estimator of θ is a statistic $T(X_1, \ldots, X_N)$ or $\hat{\theta}_n$ providing a [reasonable] substitute for the unknown value θ.
2. an estimate of θ is the value of the estimator for a given [realised] sample, $T(x_1, \ldots, x_n)$

Example: For a Normal $\mathcal{N}(\mu, \sigma^2)$ sample $X_1, \ldots, X_N$,

$$T(X_1, \ldots, X_N) = \hat{\mu}_n = \overline{X}_N$$

is an estimator of μ and $\hat{\mu}_N = 2.014$ is an estimate

Point estimation, estimators and estimates

When given a parametric family $f(\cdot|\theta)$ and a sample supposedly drawn from this family

$$(X_1, \ldots, X_N) \overset{\text{iid}}{\sim} f(x|\theta)$$

1. an estimator of θ is a statistic $T(X_1, \ldots, X_N)$ or $\hat{\theta}_n$ providing a [reasonable] substitute for the unknown value θ.
2. an estimate of θ is the value of the estimator for a given [realised] sample, $T(x_1, \ldots, x_n)$

Example: For a Normal $\mathcal{N}(\mu, \sigma^2)$ sample $X_1, \ldots, X_N$,

$$T(X_1, \ldots, X_N) = \hat{\mu}_n = \overline{X}_N$$

is an estimator of μ and $\hat{\mu}_N = 2.014$ is an estimate

Rao–Blackwell Theorem

If $\delta(\cdot)$ is an estimator of θ and $T = T(X)$ is a sufficient statistic, then

$$\delta_1(X) = \mathbb{E}_\theta[\delta(X)|T]$$

has a smaller variance than $\delta(\cdot)$

$$\mathrm{var}_\theta(\delta_1(X)) \leqslant \mathrm{var}_\theta(\delta(X))$$

[Rao, 1945; Blackwell, 1947]

mean squared error of Rao–Blackwell estimator does not exceed that of original estimator

Rao–Blackwell Theorem

If $\delta(\cdot)$ is an estimator of θ and $T = T(X)$ is a sufficient statistic, then

$$\delta_1(X) = \mathbb{E}_\theta[\delta(X)|T]$$

has a smaller variance than $\delta(\cdot)$

$$\mathrm{var}_\theta(\delta_1(X)) \leqslant \mathrm{var}_\theta(\delta(X))$$

[Rao, 1945; Blackwell, 1947]

mean squared error of Rao–Blackwell estimator does not exceed that of original estimator

Lehmann–Scheffé Theorem

Estimator δ_0

- unbiased for $\mathbb{E}_\theta[\delta X] = \Psi(\theta)$
- depends on data only through complete, sufficient statistic $S(X)$

is the unique best unbiased estimator of $\Psi(\theta)$

[Lehmann & Scheffé, 1955]

For any unbiased estimator $\delta(\cdot)$ of $\Psi(\theta)$,

$$\delta_0(X) = \mathbb{E}_\theta[\delta(X)|S(X)]$$

Lehmann–Scheffé Theorem

Estimator δ_0

- unbiased for $\mathbb{E}_\theta[\delta X] = \Psi(\theta)$
- depends on data only through complete, sufficient statistic $S(X)$

is the unique best unbiased estimator of $\Psi(\theta)$

[Lehmann & Scheffé, 1955]

For any unbiased estimator $\delta(\cdot)$ of $\Psi(\theta)$,

$$\delta_0(X) = \mathbb{E}_\theta[\delta(X)|S(X)]$$

[Fréchet–Darmois–]Cramér–Rao bound

If $\hat{\theta}$ is an estimator of $\theta \in \mathbb{R}$ with bias

$$b(\theta) = \mathbb{E}_\theta[\hat{\theta}] - \theta$$

then

$$\mathrm{var}_\theta(\hat{\theta}) \geqslant \frac{[1 + b'(\theta)]^2}{\mathfrak{I}(\theta)}$$

[Fréchet, 1943; Darmois, 1945; Rao, 1945; Cramér, 1946]

variance of any unbiased estimator at least as high as inverse Fisher information

[Fréchet–Darmois–]Cramér–Rao bound

If $\widehat{\theta}$ is an estimator of $\theta \in \mathbb{R}$ with bias

$$b(\theta) = \mathbb{E}_\theta[\widehat{\theta}] - \theta$$

then

$$\mathrm{var}_\theta(\widehat{\theta}) \geqslant \frac{[1 + b'(\theta)]^2}{\mathfrak{I}(\theta)}$$

[Fréchet, 1943; Darmois, 1945; Rao, 1945; Cramér, 1946]

variance of any unbiased estimator at least as high as inverse Fisher information

Single parameter proof

If $\delta = \delta(X)$ unbiased estimator of $\Psi(\theta)$, then

$$\mathsf{var}_\theta(\delta) \geqslant \frac{[\Psi'(\theta)]^2}{\mathfrak{I}(\theta)}$$

Take score $Z = \frac{\partial}{\partial\theta} \log f(X|\theta)$. Then

$$\mathsf{cov}_\theta(Z, \delta) = \mathbb{E}_\theta[\delta(X)Z] = \Psi'(\theta)$$

And Cauchy-Schwarz implies

$$\mathsf{cov}_\theta(Z, \delta)^2 \leqslant \mathsf{var}_\theta(\delta)\mathsf{var}_\theta(Z) = \mathsf{var}_\theta(\delta)\mathfrak{I}(\theta)$$

Warning: unbiasedness may be harmful

Unbiasedness is not an ultimate property!

- most transforms $h(\theta)$ do not allow for unbiased estimators
- no bias may imply large variance
- efficient estimators may be biased (MLE)
- existence of UNMVUE restricted to exponential families
- Cramér–Rao bound inaccessible outside exponential families

Maximum likelihood principle

Given the concentration property of the likelihood function, reasonable choice of estimator as mode:

MLE

A maximum likelihood estimator (MLE) $\hat{\theta}_N$ satisfies

$$L(\hat{\theta}_N|X_1,\dots,X_N) \geqslant L(\theta_N|X_1,\dots,X_N) \qquad \text{for all } \theta \in \Theta$$

Under regularity of $L(\cdot|X_1,\dots,X_N)$, MLE also solution of the likelihood equations

$$\nabla \log L(\hat{\theta}_N|X_1,\dots,X_N) = 0$$

Warning: $\hat{\theta}_N$ is not most likely value of θ but makes observation $(x_1,\dots,x_N)$ most likely...

Maximum likelihood principle

Given the concentration property of the likelihood function, reasonable choice of estimator as mode:

MLE

A maximum likelihood estimator (MLE) $\hat{\theta}_N$ satisfies

$$L(\hat{\theta}_N|X_1,\ldots,X_N) \geqslant L(\theta_N|X_1,\ldots,X_N) \qquad \text{for all } \theta \in \Theta$$

Under regularity of $L(\cdot|X_1,\ldots,X_N)$, MLE also solution of the likelihood equations

$$\nabla \log L(\hat{\theta}_N|X_1,\ldots,X_N) = 0$$

Warning: $\hat{\theta}_N$ is not most likely value of θ but makes observation $(x_1,\ldots,x_N)$ most likely...

Maximum likelihood principle

Given the concentration property of the likelihood function, reasonable choice of estimator as mode:

MLE

A maximum likelihood estimator (MLE) $\hat{\theta}_N$ satisfies

$$L(\hat{\theta}_N|X_1,\ldots,X_N) \geqslant L(\theta_N|X_1,\ldots,X_N) \qquad \text{for all } \theta \in \Theta$$

Under regularity of $L(\cdot|X_1,\ldots,X_N)$, MLE also solution of the likelihood equations

$$\nabla \log L(\hat{\theta}_N|X_1,\ldots,X_N) = 0$$

Warning: $\hat{\theta}_N$ is not most likely value of θ but makes observation $(x_1,\ldots,x_N)$ most likely...

Maximum likelihood invariance

Principle independent of parameterisation:

If $\xi = h(\theta)$ is a one-to-one transform of θ, then

$$\hat{\xi}_N^{\mathsf{MLE}} = h(\hat{\theta}_N^{\mathsf{MLE}})$$

[estimator of transform = transform of estimator]

By extension, if $\xi = h(\theta)$ is any transform of θ, then

$$\hat{\xi}_N^{\mathsf{MLE}} = h(\hat{\theta}_n^{\mathsf{MLE}})$$

Alternative of *profile likelihoods* distinguishing between parameters of interest and nuisance parameters

Maximum likelihood invariance

Principle independent of parameterisation:

If $\xi = h(\theta)$ is a one-to-one transform of θ, then

$$\hat{\xi}_N^{\text{MLE}} = h(\hat{\theta}_N^{\text{MLE}})$$

[estimator of transform = transform of estimator]

By extension, if $\xi = h(\theta)$ is any transform of θ, then

$$\hat{\xi}_N^{\text{MLE}} = h(\hat{\theta}_n^{\text{MLE}})$$

Alternative of *profile likelihoods* distinguishing between parameters of interest and nuisance parameters

Unicity of maximum likelihood estimate

Depending on regularity of $L(\cdot|x_1, \ldots, x_N)$, there may be

1. an a.s. unique MLE $\hat{\theta}_n^{\text{MLE}}$
2.
3.

1. Case of $x_1, \ldots, x_n \sim \mathcal{N}(\mu, 1)$
2.
3. [with $\tau = +\infty$]

Unicity of maximum likelihood estimate

Depending on regularity of $L(\cdot|x_1, \ldots, x_N)$, there may be

1.
2. several or an infinity of MLE's [or of solutions to likelihood equations]
3.

1.
2. Case of $x_1, \ldots, x_n \sim \mathcal{N}(\mu_1 + \mu_2, 1)$ [and mixtures of normal]
3. [with $\tau = +\infty$]

Unicity of maximum likelihood estimate

Depending on regularity of $L(\cdot|x_1, \ldots, x_N)$, there may be

1.
2.
3. no MLE at all

1.
2.
3. Case of $x_1, \ldots, x_n \sim \mathcal{N}(\mu_i, \tau^{-2})$ [with $\tau = +\infty$]

Unicity of maximum likelihood estimate

Consequence of standard differential calculus results on $\ell(\theta) = \log L(\theta|x_1, \ldots, x_n)$:

lemma

If Θ is connected and open, and if $\ell(\cdot)$ is twice-differentiable with

$$\lim_{\theta \to \partial\Theta} \ell(\theta) < +\infty$$

and if $H(\theta) = \nabla\nabla^T \ell(\theta)$ is positive definite at all solutions of the likelihood equations, then $\ell(\cdot)$ has a unique global maximum

Limited appeal because excluding local maxima

Unicity of maximum likelihood estimate

Consequence of standard differential calculus results on $\ell(\theta) = \log L(\theta|x_1, \ldots, x_n)$:

lemma

If Θ is connected and open, and if $\ell(\cdot)$ is twice-differentiable with

$$\lim_{\theta \to \partial\Theta} \ell(\theta) < +\infty$$

and if $H(\theta) = \nabla\nabla^T \ell(\theta)$ is positive definite at all solutions of the likelihood equations, then $\ell(\cdot)$ has a unique global maximum

Limited appeal because excluding local maxima

Unicity of MLE for exponential families

lemma

If $f(\cdot|\theta)$ is a minimal exponential family

$$f(x|\theta) = h(x) \exp\left\{T(\theta)^{\mathrm{T}} S(x) - \tau(\theta)\right\}$$

with $T(\cdot)$ one-to-one and twice differentiable over Θ, if Θ is open, and if there is at least one solution to the likelihood equations, then it is the unique MLE

Likelihood equation is equivalent to $S(x) = \mathbb{E}_\theta[S(X)]$

Unicity of MLE for exponential families

lemma

If Θ is connected and open, and if $\ell(\cdot)$ is twice-differentiable with

$$\lim_{\theta \to \partial\Theta} \ell(\theta) < +\infty$$

and if $\mathrm{H}(\theta) = \nabla\nabla^{\mathrm{T}}\ell(\theta)$ is positive definite at all solutions of the likelihood equations, then $\ell(\cdot)$ has a unique global maximum

Illustrations

Uniform $\mathcal{U}(0, \theta)$ likelihood

$$L(\theta|x_1, \ldots, x_n) = \theta^{-n} \mathbb{I}_{\theta > \max_i x_i}$$

not differentiable at $X_{(n)}$ but

$$\hat{\theta}_n^{\text{MLE}} = X_{(n)}$$

[Super-efficient estimator]

Illustrations

Bernoulli $\mathcal{B}(p)$ likelihood

$$L(p|x_1,\ldots,x_n) = \{p/1-p\}^{\sum_i x_i}(1-p)^n$$

differentiable over $(0,1)$ and

$$\hat{p}_n^{\mathsf{MLE}} = \overline{X}_n$$

Illustrations

Normal $\mathcal{N}(\mu, \sigma^2)$ likelihood

$$L(\mu, \sigma | x_1, \ldots, x_n) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\bar{x}_n - \mu)^2\right\}$$

differentiable with

$$(\hat{\mu}_n^{\mathsf{MLE}}, \widehat{\sigma^2}_n^{\mathsf{MLE}}) = \left(\overline{X}_n, \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2\right)$$

The fundamental theorem of Statistics

fundamental theorem

Under appropriate conditions, if $(X_1, \ldots, X_n) \overset{\text{iid}}{\sim} f(x|\theta)$, if $\hat{\theta}_n$ is solution of $\nabla \log f(X_1, \ldots, X_n|\theta) = 0$, then

$$\sqrt{n}\{\hat{\theta}_n - \theta\} \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \mathfrak{I}(\theta)^{-1})$$

Equivalent of CLT for estimation purposes

- $\mathfrak{I}(\theta)$ can be replaced with $\mathfrak{I}(\hat{\theta}_n)$
- or even $\hat{\mathfrak{I}}(\hat{\theta}_n) = -1/n \sum_i \nabla\nabla^T \log f(x_i|\hat{\theta}_n)$

The fundamental theorem of Statistics

fundamental theorem

Under appropriate conditions, if $(X_1, \ldots, X_n) \overset{\text{iid}}{\sim} f(x|\theta)$, if $\hat{\theta}_n$ is solution of $\nabla \log f(X_1, \ldots, X_n|\theta) = 0$, then

$$\sqrt{n}\{\hat{\theta}_n - \theta\} \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \mathfrak{I}(\theta)^{-1})$$

Equivalent of CLT for estimation purposes

- $\mathfrak{I}(\theta)$ can be replaced with $\mathfrak{I}(\hat{\theta}_n)$
- or even $\hat{\mathfrak{I}}(\hat{\theta}_n) = -1/n \sum_i \nabla\nabla^T \log f(x_i|\hat{\theta}_n)$

Assumptions

- θ identifiable
- support of $f(\cdot|\theta)$ constant in θ
- $\ell(\theta)$ thrice differentiable
- [the killer] there exists $g(x)$ integrable against $f(\cdot|\theta)$ in a neighbourhood of the true parameter such that

$$\left|\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k} f(\cdot|\theta)\right| \leqslant g(x)$$

- the following identity stands [mostly superfluous]

$$\mathfrak{I}(\theta) = \mathbb{E}_\theta\left[\nabla \log f(X|\theta)\left\{\nabla \log f(X|\theta)\right\}^{\mathrm{T}}\right] = -\mathbb{E}_\theta\left[\nabla^{\mathrm{T}}\nabla \log f(X|\theta)\right]$$

- $\hat{\theta}_n$ converges in probability to θ [similarly superfluous]

[Boos & Stefanski, 2014, p.286; Lehmann & Casella, 1998]

Inefficient MLEs

Example of MLE of $\eta = ||\theta||^2$ when $x \sim \mathcal{N}_p(\theta, I_p)$:

$$\hat{\eta}^{\text{MLE}} = ||x||^2$$

Then $\mathbb{E}_\eta[||x||^2] = \eta + p$ diverges away from η with p

Note: Consistent and efficient behaviour when considering the MLE of η based on

$$Z = ||X||^2 \sim \chi^2_p(\eta)$$

[Robert, 2001]

Inefficient MLEs

Example of MLE of $\eta = ||\theta||^2$ when $x \sim \mathcal{N}_p(\theta, I_p)$:

$$\hat{\eta}^{\text{MLE}} = ||x||^2$$

Then $\mathbb{E}_\eta[||x||^2] = \eta + p$ diverges away from η with p

Note: Consistent and efficient behaviour when considering the MLE of η based on

$$Z = ||X||^2 \sim \chi^2_p(\eta)$$

[Robert, 2001]

Inefficient MLEs

Example of MLE of $\eta = ||\theta||^2$ when $x \sim \mathcal{N}_p(\theta, I_p)$:

$$\hat{\eta}^{\text{MLE}} = ||x||^2$$

Then $\mathbb{E}_\eta[||x||^2] = \eta + p$ diverges away from η with p

Note: Consistent and efficient behaviour when considering the MLE of η based on

$$Z = ||X||^2 \sim \chi^2_p(\eta)$$

[Robert, 2001]

Inconsistent MLEs

Take $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$ with

$$f_\theta(x) = (1-\theta)\frac{1}{\delta(\theta)}\, f_0\left(\tfrac{x-\theta}{\delta(\theta)}\right) + \theta f_1(x)$$

for $\theta \in [0,1]$,

$$f_1(x) = \mathbb{I}_{[-1,1]}(x) \quad f_0(x) = (1-|x|)\mathbb{I}_{[-1,1]}(x)$$

and

$$\delta(\theta) = (1-\theta)\, \exp\{-(1-\theta)^{-4} + 1\}$$

Then for any θ

$$\hat{\theta}_n^{\text{MLE}} \xrightarrow{\text{a.s.}} 1$$

[Ferguson, 1982; John Wellner's slides, ca. 2005]

Inconsistent MLEs

Consider X_{ij} $i = 1, \ldots, n$, $j = 1, 2$ with $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$. Then

$$\hat{\mu}_i^{\mathsf{MLE}} = {}^{X_{i1}+X_{i2}}/_{2} \quad \widehat{\sigma^2}^{\mathsf{MLE}} = \frac{1}{4n} \sum_{i=1}^{n} (X_{i1} - X_{i2})^2$$

Therefore

$$\widehat{\sigma^2}^{\mathsf{MLE}} \xrightarrow{\text{a.s.}} {}^{\sigma^2}/_{2}$$

[Neyman & Scott, 1948]

Inconsistent MLEs

Consider X_{ij} $i = 1, \ldots, n$, $j = 1, 2$ with $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$. Then

$$\hat{\mu}_i^{\mathsf{MLE}} = {}^{X_{i1}+X_{i2}}/_{2} \qquad \hat{\sigma^2}^{\mathsf{MLE}} = \frac{1}{4n} \sum_{i=1}^{n} (X_{i1} - X_{i2})^2$$

Therefore

$$\hat{\sigma^2}^{\mathsf{MLE}} \xrightarrow{\text{a.s.}} {}^{\sigma^2}/_{2}$$

[Neyman & Scott, 1948]

Note: Working solely with $X_{i1} - X_{i2} \sim \mathcal{N}(0, 2\sigma^2)$ produces a consistent MLE

Likelihood optimisation

Practical optimisation of the likelihood function

$$\theta^\star = \arg\max_\theta L(\theta|\mathbf{x}) = \prod_{i=1}^n g(X_i|\theta).$$

assuming $\mathbf{X} = (X_1, \ldots, X_n) \overset{\text{iid}}{\sim} g(x|\theta)$

- analytical resolution feasible for exponential families

$$\nabla T(\theta) \sum_{i=1}^n S(x_i) = n\nabla\tau(\theta)$$

- use of standard numerical techniques like Newton-Raphson

$$\theta^{(t+1)} = \theta^{(t)} + \mathrm{I}^{\text{obs}}(\mathbf{X}, \theta^{(t)})^{-1}\nabla\ell(\theta^{(t)})$$

with $\ell(.)$ log-likelihood and I^{obs} observed information matrix

Likelihood optimisation

Practical optimisation of the likelihood function

$$\theta^\star = \arg\max_\theta L(\theta|\mathbf{x}) = \prod_{i=1}^n g(X_i|\theta).$$

assuming $\mathbf{X} = (X_1, \ldots, X_n) \overset{\text{iid}}{\sim} g(x|\theta)$

- analytical resolution feasible for exponential families

$$\nabla T(\theta) \sum_{i=1}^n S(x_i) = n\nabla\tau(\theta)$$

- use of standard numerical techniques like Newton-Raphson

$$\theta^{(t+1)} = \theta^{(t)} + \mathrm{I}^{\text{obs}}(\mathbf{X}, \theta^{(t)})^{-1}\nabla\ell(\theta^{(t)})$$

with $\ell(.)$ log-likelihood and I^{obs} observed information matrix

Likelihood optimisation

Practical optimisation of the likelihood function

$$\theta^\star = \arg\max_\theta L(\theta|\mathbf{x}) = \prod_{i=1}^{n} g(X_i|\theta).$$

assuming $\mathbf{X} = (X_1, \ldots, X_n) \overset{\text{iid}}{\sim} g(x|\theta)$

- analytical resolution feasible for exponential families

$$\nabla T(\theta) \sum_{i=1}^{n} S(x_i) = n \nabla \tau(\theta)$$

- use of standard numerical techniques like Newton-Raphson

$$\theta^{(t+1)} = \theta^{(t)} + I^{\text{obs}}(\mathbf{X}, \theta^{(t)})^{-1} \nabla \ell(\theta^{(t)})$$

with $\ell(.)$ log-likelihood and I^{obs} observed information matrix

EM algorithm

Cases where g is too complex for the above to work

Special case when g is a marginal

$$g(x|\theta) = \int_{\mathcal{Z}} f(x, z|\theta)\, \mathrm{d}z$$

Z called latent or missing variable

Illustrations

- censored data

$$X = \min(X^*, a) \qquad X^* \sim \mathcal{N}(\theta, 1)$$

- mixture model

$$X \sim .3\,\mathcal{N}_1(\mu_0, 1) + .7\,\mathcal{N}_1(\mu_1, 1),$$

- desequilibrium model

$$X = \min(X^*, Y^*) \qquad X^* \sim f_1(x|\theta) \quad Y^* \sim f_2(x|\theta)$$

Completion

EM algorithm based on completing data x with z, such as

$$(X, Z) \sim f(x, z|\theta)$$

Z missing data vector and pair (X, Z) complete data vector

Conditional density of Z given x:

$$k(z|\theta, x) = \frac{f(x, z|\theta)}{g(x|\theta)}$$

Completion

EM algorithm based on completing data x with z, such as

$$(X, Z) \sim f(x, z|\theta)$$

Z missing data vector and pair (X, Z) complete data vector

Conditional density of Z given x:

$$k(z|\theta, x) = \frac{f(x, z|\theta)}{g(x|\theta)}$$

Likelihood decomposition

Likelihood associated with complete data (x, z)

$$L^c(\theta|x, z) = f(x, z|\theta)$$

and likelihood for observed data

$$L(\theta|x)$$

such that

$$\log L(\theta|x) = \mathbb{E}[\log L^c(\theta|x, Z)|\theta_0, x] - \mathbb{E}[\log k(Z|\theta, x)|\theta_0, x] \quad (1)$$

for any θ_0, with integration operated against conditionnal distribution of Z given observables (and parameters), $k(z|\theta_0, x)$

There are "two θ's" ! : in (1), θ_0 is a fixed (and arbitrary) value driving integration, while θ both free (and variable)

Maximising observed likelihood

$$L(\theta|\mathbf{x})$$

equivalent to maximise r.h.s. term in (1)

$$\mathbb{E}[\log L^c(\theta|\mathbf{x}, \mathbf{Z})|\theta_0, \mathbf{x}] - \mathbb{E}[\log k(\mathbf{Z}|\theta, \mathbf{x})|\theta_0, \mathbf{x}]$$

There are "two θ's" ! : in (1), θ_0 is a fixed (and arbitrary) value driving integration, while θ both free (and variable)

Maximising observed likelihood

$$L(\theta|\boldsymbol{x})$$

equivalent to maximise r.h.s. term in (1)

$$\mathbb{E}[\log L^c(\theta|\boldsymbol{x}, \boldsymbol{Z})|\theta_0, \boldsymbol{x}] - \mathbb{E}[\log k(\boldsymbol{Z}|\theta, \boldsymbol{x})|\theta_0, \boldsymbol{x}]$$

Intuition for EM

Instead of maximising wrt θ r.h.s. term in (1), maximise only

$$\mathbb{E}[\log \mathrm{L}^c(\theta|\boldsymbol{x}, \boldsymbol{Z})|\theta_0, \boldsymbol{x}]$$

Maximisation of complete log-likelihood impossible since z unknown, hence substitute by maximisation of expected complete log-likelihood, with expectation depending on term θ_0

Intuition for EM

Instead of maximising wrt θ r.h.s. term in (1), maximise only

$$\mathbb{E}[\log \mathrm{L}^c(\theta|\boldsymbol{x}, \boldsymbol{Z})|\theta_0, \boldsymbol{x}]$$

Maximisation of complete log-likelihood impossible since $\boldsymbol{z}$ unknown, hence substitute by maximisation of expected complete log-likelihood, with expectation depending on term θ_0

Expectation–Maximisation

Expectation of complete log-likelihood denoted

$$Q(\theta|\theta_0, x) = \mathbb{E}[\log L^c(\theta|x, \mathbf{Z})|\theta_0, x]$$

to stress dependence on θ_0 and sample x

Principle

EM derives sequence of estimators $\hat{\theta}_{(j)}$, $j = 1, 2, \ldots$, through iteration of **E**xpectation and **M**aximisation steps:

$$Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j-1)}, x) = \max_{\theta} Q(\theta|\hat{\theta}_{(j-1)}, x).$$

Expectation–Maximisation

Expectation of complete log-likelihood denoted

$$Q(\theta|\theta_0, x) = \mathbb{E}[\log L^c(\theta|x, \mathbf{Z})|\theta_0, x]$$

to stress dependence on θ_0 and sample x

Principle

EM derives sequence of estimators $\hat{\theta}_{(j)}$, $j = 1, 2, \ldots$, through iteration of **E**xpectation and **M**aximisation steps:

$$Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j-1)}, x) = \max_{\theta} Q(\theta|\hat{\theta}_{(j-1)}, x).$$

EM Algorithm

Iterate (in m)

1. (*step E*) Compute

$$Q(\theta|\hat{\theta}_{(m)}, x) = \mathbb{E}[\log L^c(\theta|x, Z)|\hat{\theta}_{(m)}, x] \,,$$

2. (*step M*) Maximise $Q(\theta|\hat{\theta}_{(m)}, x)$ in θ and set

$$\hat{\theta}_{(m+1)} = \arg\max_{\theta} \; Q(\theta|\hat{\theta}_{(m)}, x).$$

until a fixed point [of Q] is found

[Dempster, Laird, & Rubin, 1978]

Justification

Observed likelihood

$$L(\theta|\boldsymbol{x})$$

increases at every EM step

$$L(\hat{\theta}_{(m+1)}|\boldsymbol{x}) \geqslant L(\hat{\theta}_{(m)}|\boldsymbol{x})$$

[Exercice: use Jensen and (1)]

Censored data

Normal $\mathcal{N}(\theta, 1)$ sample right-censored

$$L(\theta|x) = \frac{1}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta)^2\right\}[1-\Phi(a-\theta)]^{n-m}$$

Associated complete log-likelihood:

$$\log L^c(\theta|x, z) \propto -\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta)^2 - \frac{1}{2}\sum_{i=m+1}^{n}(z_i-\theta)^2 ,$$

where z_i's are censored observations, with density

$$k(z|\theta, x) = \frac{\exp\{-\frac{1}{2}(z-\theta)^2\}}{\sqrt{2\pi}[1-\Phi(a-\theta)]} = \frac{\varphi(z-\theta)}{1-\Phi(a-\theta)}, \qquad a < z.$$

Censored data

Normal $\mathcal{N}(\theta, 1)$ sample right-censored

$$L(\theta|\mathbf{x}) = \frac{1}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}(x_i - \theta)^2\right\}[1 - \Phi(a - \theta)]^{n-m}$$

Associated complete log-likelihood:

$$\log L^c(\theta|\mathbf{x}, \mathbf{z}) \propto -\frac{1}{2}\sum_{i=1}^{m}(x_i - \theta)^2 - \frac{1}{2}\sum_{i=m+1}^{n}(z_i - \theta)^2 \ ,$$

where z_i's are censored observations, with density

$$k(z|\theta, \mathbf{x}) = \frac{\exp\{-\frac{1}{2}(z - \theta)^2\}}{\sqrt{2\pi}[1 - \Phi(a - \theta)]} = \frac{\varphi(z - \theta)}{1 - \Phi(a - \theta)}, \qquad a < z.$$

Censored data (2)

At j-th EM iteration

$$
\begin{aligned}
Q(\theta|\widehat{\theta}_{(j)},\boldsymbol{x}) &\propto -\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta)^2 - \frac{1}{2}\mathbb{E}\left[\sum_{i=m+1}^{n}(Z_i-\theta)^2 \,\middle|\, \widehat{\theta}_{(j)},\boldsymbol{x}\right] \\
&\propto -\frac{1}{2}\sum_{i=1}^{m}(x_i-\theta)^2 \\
&\quad -\frac{1}{2}\sum_{i=m+1}^{n}\int_a^\infty (z_i-\theta)^2 k(z|\widehat{\theta}_{(j)},\boldsymbol{x})\,dz_i
\end{aligned}
$$

Censored data (3)

Differenciating in θ,

$$n\,\hat{\theta}_{(j+1)} = m\bar{x} + (n-m)\mathbb{E}[Z|\hat{\theta}_{(j)}]\ ,$$

with

$$\mathbb{E}[Z|\hat{\theta}_{(j)}] = \int_a^\infty z k(z|\hat{\theta}_{(j)}, x)\,\mathrm{d}z = \hat{\theta}_{(j)} + \frac{\varphi(a-\hat{\theta}_{(j)})}{1-\Phi(a-\hat{\theta}_{(j)})}.$$

Hence, EM sequence provided by

$$\hat{\theta}_{(j+1)} = \frac{m}{n}\bar{x} + \frac{n-m}{n}\left[\hat{\theta}_{(j)} + \frac{\varphi(a-\hat{\theta}_{(j)})}{1-\Phi(a-\hat{\theta}_{(j)})}\right],$$

which converges to likelihood maximum $\hat{\theta}$

Censored data (3)

Differenciating in θ,

$$n\,\hat{\theta}_{(j+1)} = m\bar{x} + (n-m)\mathbb{E}[Z|\hat{\theta}_{(j)}]\ ,$$

with

$$\mathbb{E}[Z|\hat{\theta}_{(j)}] = \int_a^\infty z k(z|\hat{\theta}_{(j)}, \mathbf{x})\,\mathrm{d}z = \hat{\theta}_{(j)} + \frac{\varphi(a-\hat{\theta}_{(j)})}{1-\Phi(a-\hat{\theta}_{(j)})}.$$

Hence, EM sequence provided by

$$\hat{\theta}_{(j+1)} = \frac{m}{n}\bar{x} + \frac{n-m}{n}\left[\hat{\theta}_{(j)} + \frac{\varphi(a-\hat{\theta}_{(j)})}{1-\Phi(a-\hat{\theta}_{(j)})}\right],$$

which converges to likelihood maximum $\hat{\theta}$

Mixtures

Mixture of two normal distributions with unknown means

$$.3\,\mathcal{N}_1(\mu_0, 1) + .7\,\mathcal{N}_1(\mu_1, 1),$$

sample $X_1, \ldots, X_n$ and parameter $\theta = (\mu_0, \mu_1)$
Missing data: $Z_i \in \{0, 1\}$, indicator of component associated with X_i ,

$$X_i | z_i \sim \mathcal{N}(\mu_{z_i}, 1) \qquad Z_i \sim \mathcal{B}(.7)$$

Complete likelihood

$$\begin{aligned} \log L^c(\theta|x, z) &\propto -\frac{1}{2}\sum_{i=1}^n z_i(x_i - \mu_1)^2 - \frac{1}{2}\sum_{i=1}^n (1 - z_i)(x_i - \mu_0)^2 \\ &= -\frac{1}{2} n_1(\hat{\mu}_1 - \mu_1)^2 - \frac{1}{2}(n - n_1)(\hat{\mu}_0 - \mu_0)^2 \end{aligned}$$

with

$$n_1 = \sum_{i=1}^n z_i\,, \quad n_1\hat{\mu}_1 = \sum_{i=1}^n z_i x_i\,, \quad (n - n_1)\hat{\mu}_0 = \sum_{i=1}^n (1 - z_i) x_i$$

Mixtures

Mixture of two normal distributions with unknown means

$$.3\,\mathcal{N}_1(\mu_0, 1) + .7\,\mathcal{N}_1(\mu_1, 1),$$

sample $X_1, \ldots, X_n$ and parameter $\theta = (\mu_0, \mu_1)$
Missing data: $Z_i \in \{0, 1\}$, indicator of component associated with X_i ,

$$X_i | z_i \sim \mathcal{N}(\mu_{z_i}, 1) \qquad Z_i \sim \mathcal{B}(.7)$$

Complete likelihood

$$\begin{aligned}
\log L^c(\theta|\mathbf{x}, \mathbf{z}) &\propto -\frac{1}{2}\sum_{i=1}^n z_i (x_i - \mu_1)^2 - \frac{1}{2}\sum_{i=1}^n (1 - z_i)(x_i - \mu_0)^2 \\
&= -\frac{1}{2} n_1 (\hat{\mu}_1 - \mu_1)^2 - \frac{1}{2}(n - n_1)(\hat{\mu}_0 - \mu_0)^2
\end{aligned}$$

with

$$n_1 = \sum_{i=1}^n z_i\,, \quad n_1 \hat{\mu}_1 = \sum_{i=1}^n z_i x_i\,, \quad (n - n_1)\hat{\mu}_0 = \sum_{i=1}^n (1 - z_i) x_i$$

Mixtures (2)

At j-th EM iteration

$$Q(\theta|\hat{\theta}_{(j)}, x) = \frac{1}{2}\mathbb{E}\left[n_1(\hat{\mu}_1 - \mu_1)^2 + (n - n_1)(\hat{\mu}_0 - \mu_0)^2|\hat{\theta}_{(j)}, x\right]$$

Differenciating in θ

$$\hat{\theta}_{(j+1)} = \begin{pmatrix} \mathbb{E}\left[n_1\hat{\mu}_1 \left|\hat{\theta}_{(j)}, x\right.\right] \Big/ \mathbb{E}\left[n_1|\hat{\theta}_{(j)}, x\right] \\ \\ \mathbb{E}\left[(n - n_1)\hat{\mu}_0 \left|\hat{\theta}_{(j)}, x\right.\right] \Big/ \mathbb{E}\left[(n - n_1)|\hat{\theta}_{(j)}, x\right] \end{pmatrix}$$

Mixtures (3)

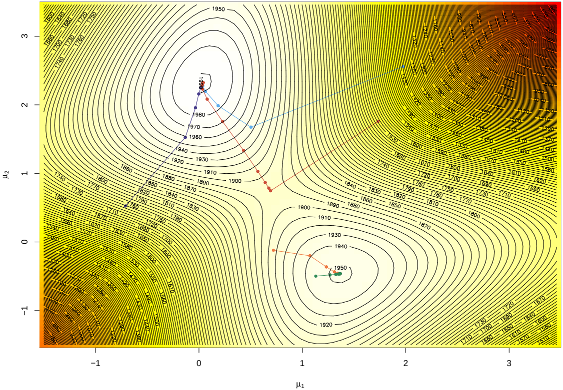
Hence $\widehat{\theta}_{(j+1)}$ given by

$$\begin{pmatrix} \sum_{i=1}^{n} \mathbb{E}\left[Z_i \left|\widehat{\theta}_{(j)}, x_i\right.\right] x_i \Big/ \sum_{i=1}^{n} \mathbb{E}\left[Z_i|\widehat{\theta}_{(j)}, x_i\right] \\ \\ \sum_{i=1}^{n} \mathbb{E}\left[(1 - Z_i) \left|\widehat{\theta}_{(j)}, x_i\right.\right] x_i \Big/ \sum_{i=1}^{n} \mathbb{E}\left[(1 - Z_i)|\widehat{\theta}_{(j)}, x_i\right] \end{pmatrix}$$

Conclusion

Step *(E)* in EM replaces missing data Z_i with their conditional expectation, given x (expectation that depend on $\widehat{\theta}_{(m)}$).

Mixtures (3)



EM iterations for several starting values

Properties

EM algorithm such that

- it converges to local maximum or saddle-point
- it depends on the initial condition $\theta_{(0)}$
- it requires several initial values when likelihood multimodal

Chapter 4 :
Decision theory and Bayesian analysis

A pedestrian example

paired and orphan socks

A drawer contains an unknown number of socks, some of which can be paired and some of which are orphans (single). One takes at random 11 socks without replacement from this drawer: no pair can be found among those. What can we infer about the total number of socks in the drawer?

- sounds like an impossible task
- one observation $x = 11$ and two unknowns, n_{socks} and n_{pairs}
- writing the likelihood is a challenge [exercise]

A pedestrian example

paired and orphan socks

A drawer contains an unknown number of socks, some of which can be paired and some of which are orphans (single). One takes at random 11 socks without replacement from this drawer: no pair can be found among those. What can we infer about the total number of socks in the drawer?

- sounds like an impossible task
- one observation $x = 11$ and two unknowns, n_{socks} and n_{pairs}
- writing the likelihood is a challenge [exercise]

A prioris on socks

Given parameters n_{socks} and n_{pairs}, set of socks

$$\mathcal{S} = \{s_1, s_1, \ldots, s_{n_{\text{pairs}}}, s_{n_{\text{pairs}}}, s_{n_{\text{pairs}}+1}, \ldots, s_{n_{\text{socks}}}\}$$

and 11 socks picked at random from $\mathcal{S}$ give X unique socks.

Rassmus' reasoning

If you are a family of 3-4 persons then a guesstimate would be that you have something like 15 pairs of socks in store. It is also possible that you have much more than 30 socks. So as *a prior* for n_{socks} I'm going to use a negative binomial with mean 30 and standard deviation 15.

On $n_{\text{pairs}}/2n_{\text{socks}}$ I'm going to put a Beta *prior* distribution that puts most of the probability over the range 0.75 to 1.0,

[Rassmus Bååth's Research Blog, Oct 20th, 2014]

A prioris on socks

Given parameters n_{socks} and n_{pairs}, set of socks

$$\mathcal{S} = \{s_1, s_1, \ldots, s_{n_{\text{pairs}}}, s_{n_{\text{pairs}}}, s_{n_{\text{pairs}}+1}, \ldots, s_{n_{\text{socks}}}\}$$

and 11 socks picked at random from $\mathcal{S}$ give X unique socks.

Rassmus' reasoning

If you are a family of 3-4 persons then a guesstimate would be that you have something like 15 pairs of socks in store. It is also possible that you have much more than 30 socks. So as *a prior* for n_{socks} I'm going to use a negative binomial with mean 30 and standard deviation 15.
On $n_{\text{pairs}}/2n_{\text{socks}}$ I'm going to put a Beta *prior* distribution that puts most of the probability over the range 0.75 to 1.0,

[Rassmus Bååth's Research Blog, Oct 20th, 2014]

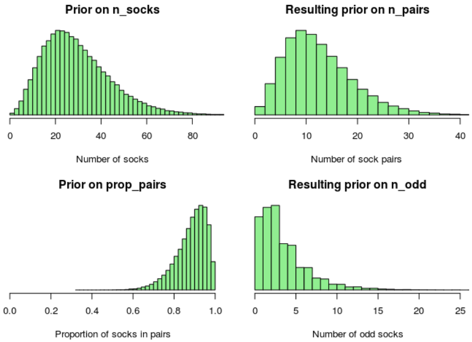
Simulating the experiment

Given a *prior* distribution on n_{socks} and n_{pairs},

$$n_{\text{socks}} \sim \mathcal{N}eg(30, 15) \quad n_{\text{pairs}} | n_{\text{socks}} \sim {}^{n_{\text{socks}}}/_{2}\mathcal{B}e(15, 2)$$

possible to

1. generate new values of n_{socks} and n_{pairs},
2. generate a new observation of X, number of unique socks out of 11.
3. accept the pair $(n_{\text{socks}}, n_{\text{pairs}})$ if the realisation of X is equal to 11



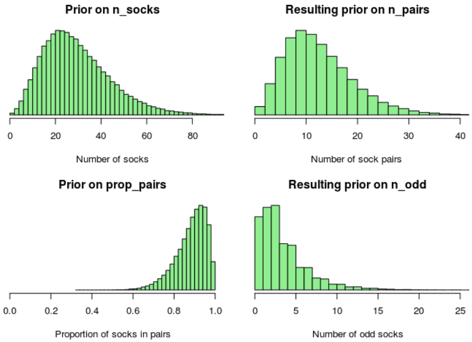
Simulating the experiment

Given a *prior* distribution on n_{socks} and n_{pairs},

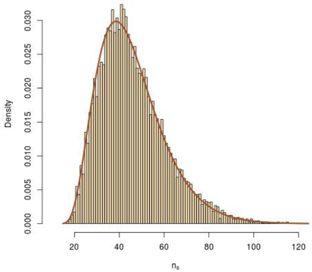
$$n_{\text{socks}} \sim \mathcal{N}eg(30, 15) \quad n_{\text{pairs}}|n_{\text{socks}} \sim {}^{n_{\text{socks}}}/{}_{2}\mathcal{B}e(15, 2)$$

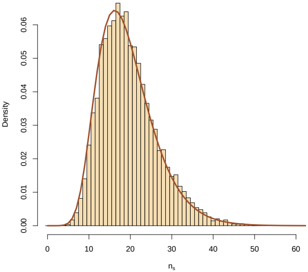
possible to

1. generate new values of n_{socks} and n_{pairs},
2. generate a new observation of X, number of unique socks out of 11.
3. accept the pair $(n_{\text{socks}}, n_{\text{pairs}})$ if the realisation of X is equal to 11



Meaning



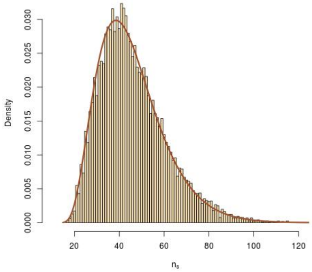


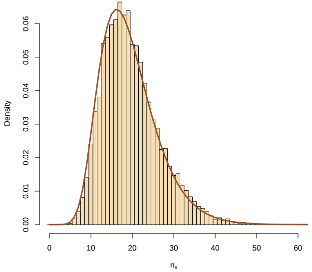
The outcome of this simulation method returns a distribution on the pair $(n_{\text{socks}}, n_{\text{pairs}})$ that is the conditional distribution of the pair given the observation $X = 11$

Proof: Generations from $\pi(n_{\text{socks}}, n_{\text{pairs}})$ are accepted with probability

$$\mathbb{P}\{X = 11 | (n_{\text{socks}}, n_{\text{pairs}})\}$$

Meaning





The outcome of this simulation method returns a distribution on the pair $(n_{\text{socks}}, n_{\text{pairs}})$ that is the conditional distribution of the pair given the observation $X = 11$

Proof: Hence accepted values distributed from

$$\pi(n_{\text{socks}}, n_{\text{pairs}}) \times \mathbb{P}\{X = 11 | (n_{\text{socks}}, n_{\text{pairs}})\} = \pi(n_{\text{socks}}, n_{\text{pairs}} | X = 11)$$

General principle

Bayesian principle Given a probability distribution on the parameter θ called prior

$$\pi(\theta)$$

and an observation x of $X \sim f(x|\theta)$, Bayesian inference relies on the conditional distribution of θ given $X = x$

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int \pi(\theta)f(x|\theta)d\theta}$$

called posterior distribution

[Bayes' theorem]

Thomas Bayes
(FRS, 1701?-1761)

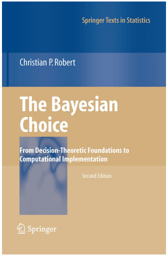
Bayesian inference

Posterior distribution

$$\pi(\theta|x)$$

as distribution on θ the parameter conditional on x the observation used for all aspects of inference

- point estimation, e.g., $\mathbb{E}[h(\theta)|x]$;
- confidence intervals, e.g., $\{\theta;\ \pi(\theta|x) \geqslant \kappa\}$;
- tests of hypotheses, e.g., $\pi(\theta = 0|x)$; and
- prediction of future observations



Central tool... central to Bayesian inference

Posterior defined up to a constant as

$$\pi(\theta|x) \propto f(x|\theta)\,\pi(\theta)$$

- Operates conditional upon the observation(s) $X = x$
- Integrate simultaneously prior information and information brought by x
- Avoids averaging over the unobserved values of X
- Coherent updating of the information available on θ, independent of the order in which i.i.d. observations are collected [domino effect]
- Provides a complete inferential scope and a unique motor of inference

The thorny issue of the prior distribution

Compared with likelihood inference, based solely on

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

Bayesian inference introduces an extra measure $\pi(\theta)$ that is chosen *a priori*, hence subjectively by the statistician based on

- hypothetical range of θ
- guesstimates of θ with an associated (lack of) precision
- type of sampling distribution

Note There also exist reference solutions (see below)

The thorny issue of the prior distribution

Compared with likelihood inference, based solely on

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

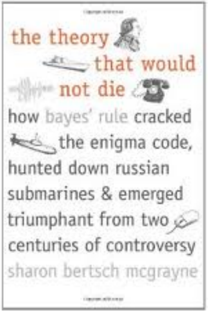
Bayesian inference introduces an extra measure $\pi(\theta)$ that is chosen *a priori*, hence subjectively by the statistician based on

- hypothetical range of θ
- guesstimates of θ with an associated (lack of) precision
- type of sampling distribution

Note There also exist reference solutions (see below)

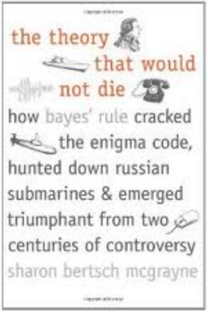
Bayes' example

Billiard ball W rolled on a line of length one, with a uniform probability of stopping anywhere: W stops at p. Second ball O then rolled n times under the same assumptions. X denotes the number of times the ball O stopped on the left of W.



Bayes' example

Billiard ball W rolled on a line of length one, with a uniform probability of stopping anywhere: W stops at p. Second ball O then rolled n times under the same assumptions. X denotes the number of times the ball O stopped on the left of W.

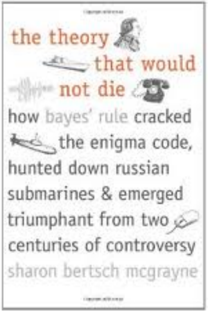


Thomas Bayes' question

Given X, what inference can we make on p?

Bayes' example

Billiard ball W rolled on a line of length one, with a uniform probability of stopping anywhere: W stops at p. Second ball O then rolled n times under the same assumptions. X denotes the number of times the ball O stopped on the left of W.



Modern translation:

Derive the posterior distribution of p given X, when

$$p \sim \mathcal{U}([0,1]) \text{ and } X \sim \mathcal{B}(n,p)$$

Resolution

Since

$$
\begin{aligned}
P(X = x|p) &= \binom{n}{x} p^x (1-p)^{n-x}, \\
P(a < p < b \text{ and } X = x) &= \int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp
\end{aligned}
$$

and

$$
P(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \, dp,
$$

Resolution (2)

then

$$P(a < p < b | X = x) = \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x}\, dp}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x}\, dp} = \frac{\int_a^b p^x (1-p)^{n-x}\, dp}{B(x+1, n-x+1)},$$

i.e.

$$p|x \sim \mathcal{B}e(x+1, n-x+1)$$

[Beta distribution]

Resolution (2)

then

$$\begin{aligned} P(a < p < b|X = x) &= \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x}\, dp}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x}\, dp} \\ &= \frac{\int_a^b p^x (1-p)^{n-x}\, dp}{B(x+1, n-x+1)}\,, \end{aligned}$$

i.e.

$$p|x \sim \mathcal{B}e(x+1, n-x+1)$$

[Beta distribution]

Conjugate priors

Easiest case is when prior distribution is within parametric family

Conjugacy

In this case, posterior inference is tractable and reduces to updating the hyperparameters* of the prior

Example In Thomas Bayes' example, the $\mathcal{B}e(a, b)$ prior is conjugate

*The hyperparameters are parameters of the priors; they are most often not treated as random variables

Conjugate priors

Easiest case is when prior distribution is within parametric family

Conjugacy

Given a likelihood function $L(y|\theta)$, the family Π of priors π_0 on Θ is said to be conjugate if the posterior $\pi(\cdot|y)$ also belong to Π

In this case, posterior inference is tractable and reduces to updating the hyperparameters* of the prior

Example In Thomas Bayes' example, the $\mathcal{B}e(a, b)$ prior is conjugate

---

*The hyperparameters are parameters of the priors; they are most often not treated as random variables

Conjugate priors

Easiest case is when prior distribution is within parametric family

Conjugacy

A family $\mathcal{F}$ of probability distributions on Θ is *conjugate* for a likelihood function $f(x|\theta)$ if, for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$.

In this case, posterior inference is tractable and reduces to updating the hyperparameters* of the prior

Example In Thomas Bayes' example, the $\mathcal{B}e(a, b)$ prior is conjugate

*The hyperparameters are parameters of the priors; they are most often not treated as random variables

Conjugate priors

Easiest case is when prior distribution is within parametric family

Conjugacy

A family $\mathcal{F}$ of probability distributions on Θ is *conjugate* for a likelihood function $f(x|\theta)$ if, for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$.

In this case, posterior inference is tractable and reduces to updating the hyperparameters* of the prior

Example In Thomas Bayes' example, the $\mathcal{B}e(a, b)$ prior is conjugate

*The hyperparameters are parameters of the priors; they are most often not treated as random variables

Exponential families and conjugacy

The family of exponential distributions

$$
\begin{aligned}
f(x|\theta) &= C(\theta)h(x)\exp\{R(\theta)\cdot T(x)\} \\
&= h(x)\exp\{R(\theta)\cdot T(x) - \tau(\theta)\}
\end{aligned}
$$

allows for conjugate priors

$$\pi(\theta|\mu,\lambda) = K(\mu,\lambda)\, e^{\theta.\mu-\lambda\psi(\theta)}$$

Following Pitman-Koopman-Darmois' Lemma, only case [besides uniform distributions]

Exponential families and conjugacy

The family of exponential distributions

$$\begin{aligned} f(x|\theta) &= C(\theta)h(x)\exp\{R(\theta)\cdot T(x)\} \\ &= h(x)\exp\{R(\theta)\cdot T(x) - \tau(\theta)\} \end{aligned}$$

allows for conjugate priors

$$\pi(\theta|\mu,\lambda) = K(\mu,\lambda)\, e^{\theta.\mu-\lambda\psi(\theta)}$$

Following Pitman-Koopman-Darmois' Lemma, only case [besides uniform distributions]

Illustration

Discrete/Multinomial & Dirichlet

If observations consist of positive counts $Y_1, \dots, Y_d$ modelled by a Multinomial $\mathcal{M}(\theta_1, \dots, \theta_p)$ distribution

$$L(y|\theta, n) = \frac{n!}{\prod_{i=1}^d y_i!} \prod_{i=1}^d \theta_i^{y_i}$$

conjugate family is the Dirichlet $\mathcal{D}(\alpha_1, \dots, \alpha_d)$ distribution

$$\pi(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^d \alpha_i)}{\prod_{i=1}^d \Gamma(\alpha_i)} \prod_i^d \theta_i^{\alpha_i - 1}$$

defined on the probability simplex ($\theta_i \geqslant 0, \sum_{i=1}^d \theta_i = 1$), where Γ is the gamma function $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$

Standard exponential families

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	$\mathcal{B}e(\alpha + x, \beta + n - x)$

Standard exponential families [2]

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Negative Binomial $\mathcal{N}eg(m, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	$\mathcal{B}e(\alpha + m, \beta + x)$
Multinomial $\mathcal{M}_k(\theta_1, \ldots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$	$\mathcal{D}(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$
Normal $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}a(\alpha, \beta)$	$\mathcal{G}(\alpha + 0.5, \beta + (\mu - x)^2/2)$

Linearity of the posterior mean

Lemma If

$$\theta \sim \pi_{\lambda,x_0}(\theta) \propto e^{\theta \cdot x_0 - \lambda\psi(\theta)}$$

with $x_0 \in \mathcal{X}$, then

$$\mathbb{E}^{\pi}[\nabla\psi(\theta)] = \frac{x_0}{\lambda}.$$

Therefore, if $x_1, \ldots, x_n$ are i.i.d. $f(x|\theta)$,

$$\mathbb{E}^{\pi}[\nabla\psi(\theta)|x_1, \ldots, x_n] = \frac{x_0 + n\bar{x}}{\lambda + n}$$

Improper distributions

Necessary extension from a prior probability distribution to a prior σ-finite positive measure π such that

$$\int_\Theta \pi(\theta)\,\mathrm{d}\theta = +\infty$$

Improper prior distribution

Note A σ-finite density with

$$\int_\Theta \pi(\theta)\,\mathrm{d}\theta < +\infty$$

can be renormalised into a probability density

Improper distributions

Necessary extension from a prior probability distribution to a prior σ-finite positive measure π such that

$$\int_\Theta \pi(\theta)\, d\theta = +\infty$$

Improper prior distribution

Note A σ-finite density with

$$\int_\Theta \pi(\theta)\, d\theta < +\infty$$

can be renormalised into a probability density

Improper distributions

Necessary extension from a prior probability distribution to a prior σ-finite positive measure π such that

$$\int_\Theta \pi(\theta)\,\mathrm{d}\theta = +\infty$$

Improper prior distribution

Note A σ-finite density with

$$\int_\Theta \pi(\theta)\,\mathrm{d}\theta < +\infty$$

can be renormalised into a probability density

Justifications

Often automatic prior determination leads to improper prior distributions

1. Only way to derive a prior in noninformative settings
2. Performances of estimators derived from these generalized distributions usually good
3. Improper priors often occur as limits of proper distributions
4. More *robust* answer against possible *misspecifications* of the prior
5. Penalization factor

Justifications

Often automatic prior determination leads to improper prior distributions

1. Only way to derive a prior in noninformative settings
2. Performances of estimators derived from these generalized distributions usually good
3. Improper priors often occur as limits of proper distributions
4. More *robust* answer against possible *misspecifications* of the prior
5. Penalization factor

Justifications

Often automatic prior determination leads to improper prior distributions

1. Only way to derive a prior in noninformative settings
2. Performances of estimators derived from these generalized distributions usually good
3. Improper priors often occur as limits of proper distributions
4. More *robust* answer against possible *misspecifications* of the prior
5. Penalization factor

Justifications

Often automatic prior determination leads to improper prior distributions

1. Only way to derive a prior in noninformative settings
2. Performances of estimators derived from these generalized distributions usually good
3. Improper priors often occur as limits of proper distributions
4. More *robust* answer against possible *misspecifications* of the prior
5. Penalization factor

Justifications

Often automatic prior determination leads to improper prior distributions

1. Only way to derive a prior in noninformative settings
2. Performances of estimators derived from these generalized distributions usually good
3. Improper priors often occur as limits of proper distributions
4. More *robust* answer against possible *misspecifications* of the prior
5. Penalization factor

Validation

Extension of the posterior distribution $\pi(\theta|x)$ associated with an improper prior π as given by Bayes's formula

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta},$$

when

$$\int_\Theta f(x|\theta)\pi(\theta)\,d\theta < \infty$$

Validation

Extension of the posterior distribution $\pi(\theta|x)$ associated with an improper prior π as given by Bayes's formula

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta},$$

when

$$\int_\Theta f(x|\theta)\pi(\theta)\,d\theta < \infty$$

Normal illustration

If $x \sim \mathcal{N}(\theta, 1)$ and $\pi(\theta) = \varpi$, constant, the pseudo marginal distribution is

$$m(x) = \varpi \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-(x-\theta)^2/2\right\} d\theta = \varpi$$

and the posterior distribution of θ is

$$\pi(\theta \mid x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-{}^{(x-\theta)^2}/{}_{2}\right\},$$

i.e., corresponds to a $\mathcal{N}(x, 1)$ distribution.

[independent of ω]

Normal illustration

If $x \sim \mathcal{N}(\theta, 1)$ and $\pi(\theta) = \varpi$, constant, the pseudo marginal distribution is

$$m(x) = \varpi \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-(x-\theta)^2/2\right\} d\theta = \varpi$$

and the posterior distribution of θ is

$$\pi(\theta \mid x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-{}^{(x-\theta)^2}/_2\right\},$$

i.e., corresponds to a $\mathcal{N}(x, 1)$ distribution.

[independent of ω]

Normal illustration

If $x \sim \mathcal{N}(\theta, 1)$ and $\pi(\theta) = \varpi$, constant, the pseudo marginal distribution is

$$m(x) = \varpi \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-(x-\theta)^2/2\right\} d\theta = \varpi$$

and the posterior distribution of θ is

$$\pi(\theta \mid x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-{}^{(x-\theta)^2}/{}_2\right\},$$

i.e., corresponds to a $\mathcal{N}(x, 1)$ distribution.

[independent of ω]

Warning

The mistake is to think of them [non-informative priors] as representing ignorance

[Lindley, 1990]

Normal illustration:
Consider a $\theta \sim \mathcal{N}(0, \tau^2)$ prior. Then

$$\lim_{\tau \to \infty} P^\pi(\theta \in [a, b]) = 0$$

for any (a, b)

Warning

> *Noninformative priors cannot be expected to represent exactly total ignorance about the problem at hand, but should rather be taken as reference or default priors, upon which everyone could fall back when the prior information is missing.*
>
> [Kass and Wasserman, 1996]

Normal illustration:

Consider a $\theta \sim \mathcal{N}(0, \tau^2)$ prior. Then

$$\lim_{\tau \to \infty} P^\pi(\theta \in [a, b]) = 0$$

for any (a, b)

Haldane prior

Consider a binomial observation, $x \sim \mathcal{B}(n, p)$, and

$$\pi^*(p) \propto [p(1-p)]^{-1}$$

[Haldane, 1931]

The marginal distribution,

$$\begin{aligned} m(x) &= \int_0^1 [p(1-p)]^{-1} \binom{n}{x} p^x (1-p)^{n-x} dp \\ &= B(x, n-x), \end{aligned}$$

is only defined for $x \neq 0, n$.

[Not recommended!]

Haldane prior

Consider a binomial observation, $x \sim \mathcal{B}(n, p)$, and

$$\pi^*(p) \propto [p(1-p)]^{-1}$$

[Haldane, 1931]

The marginal distribution,

$$\begin{aligned} m(x) &= \int_0^1 [p(1-p)]^{-1} \binom{n}{x} p^x (1-p)^{n-x} dp \\ &= B(x, n-x), \end{aligned}$$

is only defined for $x \neq 0, n$.

[Not recommended!]

The Jeffreys prior

Based on Fisher information

$$\mathfrak{I}(\theta) = \mathbb{E}_\theta \left[\frac{\partial \ell}{\partial \theta^t} \, \frac{\partial \ell}{\partial \theta} \right]$$

Jeffreys prior density is

$$\pi^*(\theta) \propto |\mathfrak{I}(\theta)|^{1/2}$$

Pros & Cons

- relates to information theory
- agrees with most invariant priors
- parameterisation invariant

The Jeffreys prior

Based on Fisher information

$$\mathfrak{I}(\theta) = \mathbb{E}_\theta \left[\frac{\partial \ell}{\partial \theta^t} \, \frac{\partial \ell}{\partial \theta} \right]$$

Jeffreys prior density is

$$\pi^*(\theta) \propto |\mathfrak{I}(\theta)|^{1/2}$$

Pros & Cons

- relates to information theory
- agrees with most invariant priors
- parameterisation invariant

Example

If $x \sim \mathcal{N}_p(\theta, I_p)$, Jeffreys' prior is

$$\pi(\theta) \propto 1$$

and if $\eta = \|\theta\|^2$,

$$\pi(\eta) = \eta^{p/2-1}$$

and

$$\mathbb{E}^\pi[\eta|x] = \|x\|^2 + p$$

with bias $2p$

[Not recommended!]

Example

If $x \sim \mathcal{B}(n, \theta)$, Jeffreys' prior is

$$\mathcal{B}e(1/2, 1/2)$$

and, if $n \sim \mathcal{N}eg(x, \theta)$, Jeffreys' prior is

$$\begin{aligned}
\pi_2(\theta) &= -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right] \\
&= \mathbb{E}_\theta\left[\frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2}\right] = \frac{x}{\theta^2(1-\theta)}, \\
&\propto \theta^{-1}(1-\theta)^{-1/2}
\end{aligned}$$

MAP estimator

When considering estimates of the parameter θ, one default solution is the maximum a posteriori (MAP) estimator

$$\arg\max_{\theta} \ell(\theta|x)\pi(\theta)$$

Motivations

- Most likely value of θ
- Penalized likelihood estimator
- Further appeal in restricted parameter spaces

MAP estimator

When considering estimates of the parameter θ, one default solution is the maximum a posteriori (MAP) estimator

$$\arg\max_{\theta} \ell(\theta|x)\pi(\theta)$$

Motivations

- Most likely value of θ
- Penalized likelihood estimator
- Further appeal in restricted parameter spaces

Illustration

Consider $x \sim \mathcal{B}(n, p)$. Possible priors:

$$\pi^*(p) = \frac{1}{B(1/2, 1/2)} p^{-1/2}(1-p)^{-1/2},$$

$$\pi_1(p) = 1 \quad \text{and} \quad \pi_2(p) = p^{-1}(1-p)^{-1}.$$

Corresponding MAP estimators:

$$\begin{aligned}
\delta^*(x) &= \max\left(\frac{x-1/2}{n-1}, 0\right), \\
\delta_1(x) &= \frac{x}{n}, \\
\delta_2(x) &= \max\left(\frac{x-1}{n-2}, 0\right).
\end{aligned}$$

Illustration [opposite]

MAP not always appropriate:

When

$$f(x|\theta) = \frac{1}{\pi}\left[1 + (x-\theta)^2\right]^{-1},$$

and

$$\pi(\theta) = \frac{1}{2} e^{-|\theta|}$$

then MAP estimator of θ is always

$$\delta^*(x) = 0$$

Prediction

Inference on new observations depending on the same parameter, conditional on the current data

If $x \sim f(x|\theta)$ [observed], $\theta \sim \pi(\theta)$, and $z \sim g(z|x, \theta)$ [unobserved], *predictive* of z is marginal conditional

$$g^{\pi}(z|x) = \int_{\Theta} g(z|x, \theta)\pi(\theta|x)\, d\theta.$$

time series illustration

Consider the AR(1) model

$$x_t = \rho x_{t-1} + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

predictive of x_T is then

$$x_T | x_{1:(T-1)} \sim \int \frac{\sigma^{-1}}{\sqrt{2\pi}} \exp\{-(x_T - \rho x_{T-1})^2 / 2\sigma^2\} \pi(\rho, \sigma | x_{1:(T-1)}) \mathrm{d}\rho \mathrm{d}\sigma \,,$$

and $\pi(\rho, \sigma | x_{1:(T-1)})$ can be expressed in closed form

Posterior mean

Theorem The solution to

$$\arg\min_{\delta} \mathbb{E}^{\pi}\left[\left.||\theta - \delta||^2\right| x\right]$$

is given by

$$\delta^{\pi}(x) = \mathbb{E}^{\pi}\left[\theta|x\right]$$

[Posterior mean = Bayes estimator under quadratic loss]

Posterior median

Theorem When $\theta \in \mathbb{R}$, the solution to

$$\arg\min_{\delta} \mathbb{E}^{\pi}\left[|\theta - \delta| \,|\, x\right]$$

is given by

$$\delta^{\pi}(x) = \text{median}^{\pi}\left(\theta|x\right)$$

[Posterior mean = Bayes estimator under absolute loss]

Obvious extension to

$$\arg\min_{\delta} \mathbb{E}^{\pi}\left[\sum_{i=1}^{p} |\theta_i - \delta| \,\middle|\, x\right]$$

Posterior median

Theorem When $\theta \in \mathbb{R}$, the solution to

$$\arg\min_{\delta} \mathbb{E}^{\pi}\left[|\theta - \delta| \,|\, x\right]$$

is given by

$$\delta^{\pi}(x) = \mathsf{median}^{\pi}\left(\theta | x\right)$$

[Posterior mean = Bayes estimator under absolute loss]

Obvious extension to

$$\arg\min_{\delta} \mathbb{E}^{\pi}\left[\sum_{i=1}^{p} |\theta_i - \delta| \,\middle|\, x\right]$$

Inference with conjugate priors

For conjugate distributions, posterior expectations of the natural parameters may be expressed analytically, for one or several observations.

Distribution	Conjugate prior	Posterior mean
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	$\dfrac{\mu\sigma^2 + \tau^2 x}{\sigma^2 + \tau^2}$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\dfrac{\alpha + x}{\beta + 1}$

Inference with conjugate priors

For conjugate distributions, posterior expectations of the natural parameters may be expressed analytically, for one or several observations.

Distribution	Conjugate prior	Posterior mean
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\dfrac{\alpha + \nu}{\beta + x}$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	$\dfrac{\alpha + x}{\alpha + \beta + n}$
Negative binomial $\mathcal{N}eg(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	$\dfrac{\alpha + n}{\alpha + \beta + x + n}$
Multinomial $\mathcal{M}_k(n; \theta_1, \ldots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$	$\dfrac{\alpha_i + x_i}{\left(\sum_j \alpha_j\right) + n}$
Normal $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}(\alpha/2, \beta/2)$	$\dfrac{\alpha + 1}{\beta + (\mu - x)^2}$

Illustration

Consider

$$x_1,\ \ldots,\ x_n \sim \mathcal{U}([0,\theta])$$

and $\theta \sim \mathcal{P}a(\theta_0, \alpha)$. Then

$$\theta | x_1, \ldots, x_n \sim \mathcal{P}a(\max(\theta_0, x_1, \ldots, x_n), \alpha + n)$$

and

$$\delta^\pi(x_1, \ldots, x_n) = \frac{\alpha + n}{\alpha + n - 1} \max(\theta_0, x_1, \ldots, x_n).$$

HPD region

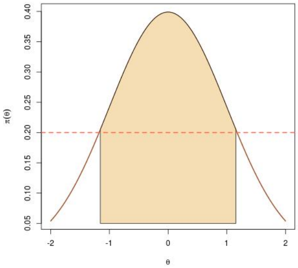
Natural confidence region based on $\pi(\cdot|x)$ is

$$\mathfrak{C}^{\pi}(x) = \{\theta; \pi(\theta|x) > k\}$$

with

$$\mathbb{P}^{\pi}(\theta \in \mathfrak{C}^{\pi}|x) = 1 - \alpha$$

Highest posterior density (HPD) region



HPD region

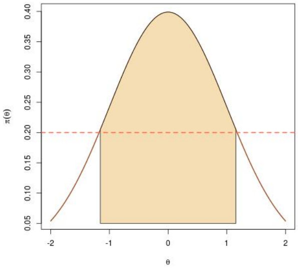
Natural confidence region based on $\pi(\cdot|x)$ is

$$\mathfrak{C}^\pi(x) = \{\theta; \pi(\theta|x) > k\}$$

with

$$\mathbb{P}^\pi(\theta \in \mathfrak{C}^\pi|x) = 1 - \alpha$$

Highest posterior density (HPD) region



Example case $x \sim \mathcal{N}(\theta, 1)$ and $\theta \sim \mathcal{N}(0, 10)$. Then

$$\theta|x \sim \mathcal{N}\left({}^{10}/_{11}x, {}^{10}/_{11}\right)$$

and

$$\begin{aligned}\mathfrak{C}^\pi(x) &= \left\{\theta; |\theta - {}^{10}/_{11}x| > k'\right\} \\ &= \left({}^{10}/_{11}x - k', {}^{10}/_{11}x + k'\right)\end{aligned}$$

HPD region

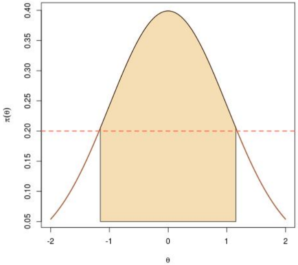
Natural confidence region based on $\pi(\cdot|x)$ is

$$\mathfrak{C}^\pi(x) = \{\theta; \pi(\theta|x) > k\}$$

with

$$\mathbb{P}^\pi(\theta \in \mathfrak{C}^\pi|x) = 1 - \alpha$$

Highest posterior density (HPD) region



Warning Frequentist coverage is not $1 - \alpha$, hence name of credible rather than confidence region

Further validation of HPD regions as smallest-volume $1 - \alpha$-coverage regions