Statistical modelling

Christian P. Robert

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

Licence MI2E, année 2020–2021
Getting Started

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 end in itself.

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

[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 $F$
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 \textit{without} a model, which is a mathematical construct intended to bring regularity and reproducibility, in order to draw inference

\begin{quote}
\textit{``All models are wrong but some are more useful than others''}
—George Box—
\end{quote}

Usefulness is to be understood as having explanatory or predictive abilities
“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

Mortality from oral cancer in Taiwan:
Model chosen to be

\[ Y_i \sim \mathcal{P}(m_i) \quad \log m_i = \log E_i + \alpha + \epsilon_i \]


(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
Example 1: spatial pattern

Mortality from oral cancer in Taiwan:
Model chosen to be

\[ Y_i \sim \mathcal{P}(m_i) \quad \log m_i = \log E_i + \alpha + \epsilon_i \]

where

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

Example 2: World cup predictions

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

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

Corresponding data model is:

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

where \( a_i \) and \( a_j \) ability parameters and \( \sigma_y \) scale parameter estimated from the data

Nate Silver’s prior scores

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

[\text{A. Gelman, blog, 13 July 2014}]
Example 2: World cup predictions

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

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

Potential outliers led to fatter tail model:

$$y_{ij} \sim T_7(a_i - a_j, \sigma_y),$$

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.”
“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.

[BBN 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.

[Last name & 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

![Births by Day of Year](image)

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)
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
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
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]
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]
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
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)
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
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 $\mathbb{E}[X]$

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

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 $\sigma^2 = \text{var}(X)$,

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

[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 $E[X]$ and a finite variance $\sigma^2 = \text{var}(X)$,

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

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

Continuity Theorem
If

$$
X_n \overset{\text{dist.}}{\to} a
$$

and $g$ is continuous at $a$, then

$$
g(X_n) \overset{\text{dist.}}{\to} g(a)
$$
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 $\sigma^2 = \text{var}(X)$,

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

[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 $X, a, \text{ and } b$, respectively, then

$$X_nY_n + Z_n \xrightarrow{\text{dist.}} aX + b$$
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 $\sigma^2 = \text{var}(X)$,

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

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

Delta method’s Theorem

If $\sqrt{n}(X_n - \mu) \overset{\text{dist.}}{\rightarrow} N_p(0, \Omega)$ and $g : \mathbb{R}^p \rightarrow \mathbb{R}^q$ is a continuously differentiable function on a neighbourhood of $\mu \in \mathbb{R}^p$, with a non-zero gradient $\nabla g(\mu)$, then

$$\sqrt{n} \{g(X_n) - g(\mu)\} \overset{\text{dist.}}{\rightarrow} N_q(0, \nabla g(\mu)^T \Omega \nabla g(\mu))$$
Entertaining read
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. \( U_1, \ldots, U_n \): \n
\[ X_i = \mathbb{I}(U_i \leq p_i) \]
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. \( U_1, \ldots, U_n \):

\[ X_i = \mathbb{I}(U_i \leq p_i) \]
Two classes of statistical models:

- **Parametric** when $F$ varies within a family of distributions indexed by a parameter $\theta$ that belongs to a finite dimension space $\Theta$:

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

and to “know” $F$ is to know which $\theta$ 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 $\theta$ that belongs to a finite dimension space $\Theta$:

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

  and to “know” $F$ is to know which $\theta$ 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

$$E_F[Y|X = x] = \Psi(\beta^T x), \quad \text{var}_F(Y|X = x) = \sigma^2$$

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.
When \( F = F_\theta \), inference usually covers the whole of the parameter \( \theta \) and provides

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

\[
z_1, \ldots, z_m \sim G_\theta(z)
\]

Inference: all those procedures depend on the sample \((x_1, \ldots, x_n)\)
When $F = F_\theta$, inference usually covers the whole of the parameter $\theta$ and provides

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

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

**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

\[ X_i \sim B(p) \]

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 \( \mu \) and \( \sigma > 0 \) (e.g., blood pressure)

Questions of interest:

1. likely value of \( \mu \) or range thereof
2. whether or not \( \mu \) is above the mean \( \eta \) 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**
  \[ \mu_1 = \mathbb{E} [X] = \int x dF(x) \]
  \[ \mu_k = \mathbb{E} [(X - \mu_1)^k] \quad k > 1 \]

- **non-central moments**
  \[ \xi_k = \mathbb{E} [X^k] \quad k \geq 1 \]

- **\( \alpha \) quantile**
  \[ \mathbb{P}(X < \xi_\alpha) = \alpha \]

and (2-D) moments

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

**Note:** For parametric models, those quantities are transforms of the parameter \( \theta \)
Example 1: Binomial experiment again

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

\[ X_i \sim \mathcal{B}(p) \]

Single parameter \( p \) with

\[ \mathbb{E}[X] = p \quad \text{var}(X) = p(1 - p) \]

[somewhat boring...]

Median and mode
Example 1: Binomial experiment again

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

\[ X_i \sim B(n, p) \quad P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \]

Single parameter \( p \) with

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

[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 \( \mu \) and \( \sigma > 0 \) (e.g., blood pressure)

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

Median and mode equal to \( \mu \)
Exponential families

Class of parametric densities with nice analytic properties

Start from the normal density:

\[
\phi(x; \theta) = \frac{1}{\sqrt{2\pi}} \exp \left\{ x\theta - x^2/2 - \theta^2/2 \right\}
\]

\[
= \frac{\exp\{ -\theta^2/2\}}{\sqrt{2\pi}} \underbrace{\exp\{x\theta\} \exp\{-x^2/2\}}_{x \text{ meets } \theta}
\]

where \( \theta \) 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 $X$ is an exponential family if its density with respect to a measure $\nu$ satisfies

$$f(x|\theta) = c(\theta)h(x)\exp\{T(x)^T\tau(\theta)\}, \theta \in \Theta,$$

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:

$$c(\theta)^{-1} = \int_X h(x)\exp\{T(x)^T\tau(\theta)\}d\nu(x)$$
Exponential families (examples)

Example 1: Binomial experiment again

Binomial variable

\[ X \sim \mathcal{B}(n, p) \quad \Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \]

can be expressed as

\[ \Pr(X = k) = (1 - p)^n \binom{n}{k} \exp\{k \log(p/(1 - p))\} \]

hence

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

\[ c(p) = (1 - p)^n, \quad h(x) = \binom{n}{x}, \quad T(x) = x, \quad \tau(p) = \log(p/(1 - p)) \]
Example 2: Normal experiment again

Normal variate

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

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

\[
\begin{align*}
f(x|\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\} \\
&= \frac{\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2}\right\}
\end{align*}
\]

hence

\[
\begin{align*}
c(\theta) &= \frac{\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}}{\sqrt{2\pi\sigma^2}} \\
T(x) &= \begin{pmatrix} x^2 \\ x \end{pmatrix} \\
\tau(\theta) &= \begin{pmatrix} -\frac{1}{2\sigma^2} \\ \frac{\mu}{\sigma^2} \end{pmatrix}
\end{align*}
\]
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

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

**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

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

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

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

and the natural parameter space is $\mathbb{R}$
Possible to add and (better!) delete useless components of $T$:

**Definition**

A **regular exponential family** corresponds to the case where $\Theta$ 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
Possible to add and (better!) delete useless components of $T$:

**Definition**

A **regular exponential family** corresponds to the case where $\Theta$ 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,

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

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 $C$]
Highly regular densities

**Theorem**

The natural parameter space $\Theta$ 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.
Highly regular densities

**Theorem**

The natural parameter space $\Theta$ 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.
Lemma

If the density of $X$ has the minimal representation

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

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

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

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

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

on the interior of $\Theta$, then

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

is an analytic function on the interior of $\Theta$ and

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

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

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

[Laplace transform]
Normalising constant $c(\cdot)$ generating all moments

**Proposition**

If $T(\cdot) : \mathcal{X} \rightarrow \mathbb{R}^d$ 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

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

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

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

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

Then

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

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

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

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

Then

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

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
Example

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

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

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

where $\lambda = k\mu^2$ and $I_\nu$ Bessel function of second order
connected examples of exponential families

Counter-Example

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

corresponding to the ratio

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

with density

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

Ising $\text{Be}(n/2, m/2)$ distribution corresponding to the distribution of

$$Z = \frac{nY}{nY + m} \quad \text{when} \quad Y \sim F_{n,m}$$

has density

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

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

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

has density

$$f(z; \mu, \sigma) = \frac{1}{\sigma} \exp\{-\sigma^{-1}|x - \mu|\}$$
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 $z_1, \ldots, z_n$ (iid) put to use to approximate $p$ by

$$\hat{p} = \hat{p}(z_1, \ldots, z_n) = \frac{1}{n} \sum_{i=1}^{n} z_i$$

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 $\theta$ 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 \( \theta \) each time a new sample is used (data does produce the model)
intrinsic statistical randomness

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 $\theta$ each time a new sample is used (data does produce the model)
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 \( \delta(X_1, \ldots, X_n) \)?
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 $\delta(X_1, \ldots, X_n)$?
Example (**Normal sample**)

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

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

Variation compatible with the (known) theoretical distribution

$\hat{\theta} \sim \mathcal{N}(\theta, 1/100)$
Example (**Normal sample**)

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

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

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

$$H_0 : \theta \leq 0$$

- In the **normal** case, the **true** $\theta$ stands with high probability in the interval

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

**Quid of $\sigma$ ?!**
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

$$H_0 : \theta \leq 0$$

In the **normal** case, the **true** $\theta$ stands with high probability in the interval

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

**Quid of $\sigma$?!**
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

$$H_0 : \theta \leq 0$$

- In the **normal** case, the **true** $\theta$ stands with high probability in the interval

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

**Quid of $\sigma$?!**
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

\[ H_0 : \theta \leq 0 \]

- In the **normal** case, the **true** \(\theta\) stands with high probability in the interval

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

**Quid of \(\sigma\) ?!**
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

\[
H_0 : \theta \leq 0
\]

In the **normal** case, the **true** \( \theta \) stands with high probability in the interval

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

**Quid of \( \sigma \) ?!**
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

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{]-\infty, x]}(X_i)$$

$$= \frac{\text{card}\{X_i; X_i \leq x\}}{n},$$

Step function corresponding to the empirical distribution

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

where $\delta$ 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

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{-\infty, x]}(X_i)$$

$$= \frac{\text{card}\{X_i; X_i \leq x\}}{n},$$

Step function corresponding to the empirical distribution

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

where $\delta$ Dirac mass
convergence of the empirical cdf

Glivenko-Cantelli Theorem

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

[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

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

for every \( \varepsilon \geq \varepsilon_n = \sqrt{\frac{1}{2n} \ln 2} \)

[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 \( F(x) \)
Moments

\[ \mathbb{E}[\hat{F}_n(x)] = F(x) \]
\[ \text{var}[\hat{F}_n(x)] = \frac{F(x)(1 - F(x))}{n} \]
Confidence band

If

\[ L_n(x) = \max\{\hat{F}_n(x) - \epsilon_n, 0\}, \quad U_n(x) = \min\{\hat{F}_n(x) + \epsilon_n, 1\}, \]

then, for \( \epsilon_n = \sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}} \),

\[ \mathbb{P}\left( L_n(x) \leq F(x) \leq U_n(x) \text{ for all } x \right) \geq 1 - \alpha \]
Glivenko-Cantelli in action

Example (Normal sample)

Estimation of the cdf $F$ 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*

- **Robustness 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 \textit{non-parametric} nature: it is not necessary to know the distribution or the shape of the distribution of the sample to derive this estimator. \textcopyright it is always available

- \textbf{Robustness 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!
parametric versus non-parametric inference

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
parametric versus non-parametric inference

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,

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

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

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

$$= \frac{1}{n} \sum_{i=1}^{n} h(X_i)$$

[Functional of the cdf]
[Moment estimator]
Extension to functionals of $F$

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

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

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

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

$$= \frac{1}{n} \sum_{i=1}^{n} h(X_i)$$

[Functional of the cdf]

[Moment estimator]
If
\[ \theta(F) = \text{var}(X) = \int (x - \mathbb{E}_F[X])^2 \, dF(x) \]
then
\[ \theta(\hat{F}_n) = \int (x - \mathbb{E}_{\hat{F}_n}[X])^2 \, d\hat{F}_n(x) \]
\[ = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mathbb{E}_{\hat{F}_n}[X])^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \]
which differs from the (unbiased) sample variance
\[ \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \]
If $\theta(F)$ is the median of $F$, it is defined by

$$P_F(X \leq \theta(F)) = 0.5$$

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

$$P_{\hat{F}_n}(X \leq \theta(\hat{F}_n)) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq \theta(\hat{F}_n)) = 0.5$$

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

\(\theta\) also is the median of \(\mathcal{N}(\theta, 1)\), hence another estimator of \(\theta\) 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 distribution $F$:
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 distribution $F$:
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 distribution $F$:
Plot sorted $x_1, \ldots, x_n$ against $F^{-1}(1/(n+1)), \ldots, F^{-1}(n/(n+1))$

Example

Normal $N(0, 1)$ sample against
- $N(0, 1)$
- $N(0, 2)$
- $E(3)$
theoretical distributions
q-q plots

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

$$\hat{E}\left[h(X)\right]_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i) \overset{\text{a.s.}}{\rightarrow} E[h(X)]$$

Result fundamental for the use of computer-based 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
Principle

- produce by a computer program an arbitrary long sequence

\[ x_1, x_2, \ldots \sim \text{iid} \]

- 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 $F$, call the corresponding pseudo-random generator in an arbitrary computer language

```r
> x=rnorm(10)
> x
[1] -0.02157345 -1.13473554 1.35981245 -0.88757941 0.70356394 -1.03538265
[7] -0.74941846 0.50629858 0.83579100 0.47214477
```

use the sample as a statistician would do

```r
> mean(x)
[1] 0.004892123
> var(x)
[1] 0.8034657
```
to approximate quantities related with $F$
Monte Carlo integration

Approximation of integrals related with $F$:

**Law of large numbers**

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

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

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

**Monte Carlo principle**

1. Call a computer pseudo-random generator of $F$ to produce $X_1, \ldots, X_n$
2. Approximate $\mathcal{I}$ with $\hat{I}_n$
3. Check the precision of $\hat{I}_n$ and if needed increase $n$
Approximation of integrals related with $F$:

**Law of large numbers**

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

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

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

**Monte Carlo principle**

1. Call a computer pseudo-random generator of $F$ to produce $X_1, \ldots, X_n$
2. Approximate $J$ with $\hat{J}_n$
3. Check the precision of $\hat{J}_n$ and if needed increase $n$
example: normal moment

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

<table>
<thead>
<tr>
<th>( n )</th>
<th>5</th>
<th>50</th>
<th>500</th>
<th>5000</th>
<th>50,000</th>
<th>500,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{I}_n )</td>
<td>1.65</td>
<td>5.69</td>
<td>3.24</td>
<td>3.13</td>
<td>3.038</td>
<td>3.029</td>
</tr>
</tbody>
</table>

![Graph showing the convergence of \( \hat{I}_n \) to \( \mathbb{E}[X^4] \) as \( n \) increases.]
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

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

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

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

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

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

$$\theta(\hat{F}_n) \approx \theta(X^*_1, \ldots, X^*_n) \quad \text{with} \quad X^*_1, \ldots, X^*_n \sim \hat{F}_n$$
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

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

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

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

\[ \theta(\hat{F}_n) = \theta(X_1, \ldots, X_n) \quad \text{with} \quad X_1, \ldots, X_n \overset{\text{iid}}{\sim} F \]

replace \( F \) with \( \hat{F}_n \):

\[ \theta(\hat{F}_n) \approx \theta(X_1^*, \ldots, X_n^*) \quad \text{with} \quad X_1^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \hat{F}_n \]
For a given estimator $\theta(\hat{F}_n)$, a random variable, its (true) variance is defined as

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

bootstrap approximation

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

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

$$X_1^*, \ldots, X_n^* \sim \hat{F}_n$$

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

/ˈbuːtstrɑːp/ ①

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”
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 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
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

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

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)]
Monte Carlo implementation

1. For \( b = 1, \ldots, B \),
   - generate a sample \( X_1^b, \ldots, X_n^b \) from \( \hat{F}_n \)
   - 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
     \[
     \theta(X_1, \ldots, X_n)
     \]
Monte Carlo implementation

1. For $b = 1, \ldots, B$,
   - generate a sample $X_1^b, \ldots, X_n^b$ from $\hat{f}_n$
   - 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

   $\theta(X_1, \ldots, X_n)$
Monte Carlo implementation

1. For $b = 1, \ldots, B$,
   
   - generate a sample $X_1^b, \ldots, X_n^b$ from $\hat{F}_n$
   - 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

   $$\theta(X_1, \ldots, X_n)$$
Observation of a sample [here simulated from $0.3N(0,1) + 0.7N(2.5,1)$ as illustration]

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

Interest in the distribution of $\bar{X} = \frac{1}{n} \sum X_i$

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

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

```r
> bobar=rep(0,1000)  #B=1000
> for (t in 1:1000)
+ bobar[t]=mean(sample(x,250,rep=TRUE))
> hist(bobar)
```
Example (Sample $0.3N(0, 1) + 0.7N(2.5, 1)$)

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

\[
\eta(\hat{F}_n) = \sqrt{E^{\hat{F}_n} \left[ \{ \theta(X_1, \ldots, X_n) - E^{\hat{F}_n}[\theta(X_1, \ldots, X_n)] \}^2 \right]}
\]
Example (Derivation of the average variation)

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

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

and its bootstrap approximation is

$$
\eta(\hat{F}_n) = \sqrt{\mathbb{E}^{\hat{F}_n} \left[ \{\theta(X_1, \ldots, X_n) - \mathbb{E}^{\hat{F}_n}[\theta(X_1, \ldots, X_n)]\}^2 \right]}
$$
Example (Derivation of the average variation)

Approximation itself approximated by Monte-Carlo:

\[ \hat{\eta}(\hat{F}_n) = \left( \frac{1}{B} \sum_{b=1}^{B} (\theta(X^b_1, \ldots, X^b_n) - \bar{\theta})^2 \right)^{1/2} \]

where

\[ \bar{\theta} = \frac{1}{B} \sum_{b=1}^{B} \theta(X^b_1, \ldots, X^b_n) \]
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

\[
P(C(X_1, \ldots, X_n) \ni \theta(F)) = 1 - \alpha
\]

[1 − \( \alpha \)-level confidence intervals]

1 rely on the normal approximation

\[
\theta(\hat{F}_n) \approx N(\theta(F), \eta(F)^2)
\]

and use the interval

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

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

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

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

$$\hat{H}(r) = 1/B \sum_{b=1}^{B} \mathbb{I}(\theta(X^b_1, \ldots, X^b_n) \leq r)$$

and use the interval

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

which is also

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

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

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

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

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

and use the interval

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

which is also

$$[2\theta(\hat{F}_n) - \theta_{(n{1-\alpha/2})}, 2\theta(\hat{F}_n) - \theta_{(n{\alpha/2})}]$$
Take $X_1, \ldots, X_n$ an iid random sample and $\theta(F)$ as the median of $F$, then

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

```r
> 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

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

```r
> 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

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

```r
> 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 $0.3N(0, 1) + 0.7N(2.5, 1)$)

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

Sample

\[(X_1, \ldots, X_{100}) \sim \mathcal{N}(\theta, 1)\]

Comparison of the confidence intervals

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

[normal approximation]

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

[normal bootstrap approximation]

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

[generic bootstrap approximation]
Example (Normal sample)

Sample

\((X_1, \ldots, X_{100}) \sim \mathcal{N}(\theta, 1)\)

Comparison of the confidence intervals

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

[normal approximation]

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

[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 \( X_1, \ldots, X_n \sim U(0, \theta) \) then

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

Using bootstrap, distribution of \( \frac{n-1}{n} \theta(\hat{F}_n) \) far from truth

\[
f_{\text{max}}(x) = nx^{n-1}/\theta^n \mathbb{1}_{(0,\theta)}(x)
\]
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 $n^{-1/n} \theta(\hat{F}_n)$ far from truth

$$
f_{\text{max}}(x) = nx^{n-1}/\theta^n \mathbb{I}_{(0,\theta)}(x)
$$
If the parametric shape of $F$ is known,

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

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 $\lambda$

[Cf Example 3]
Parametric Bootstrap

If the parametric shape of F is known,

\[ F(\cdot) = \Phi_\lambda(\cdot) \quad \lambda \in \Lambda, \]

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 \( \lambda \)

[Cf Example 3]
Approximation of the distribution of
\[ \theta(X_1, \ldots, X_n) \]
by the distribution of
\[ \theta(X_1^*, \ldots, X_n^*) \]
\[ X_1^*, \ldots, X_n^* \overset{iid}{\sim} \Phi_{\hat{\lambda}_n} \]
May avoid Monte Carlo simulation approximations in some cases
Approximation of the distribution of 

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

by the distribution of 

$$\theta(X_1^*, \ldots, X_n^*) \quad X_1^*, \ldots, X_n^* \overset{iid}{\sim} \Phi_{\lambda_n}$$

May avoid Monte Carlo simulation approximations in some cases
Example (Exponential Sample)

Take

\[ X_1, \ldots, X_n \sim \text{iid} \exp(\lambda) \]

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 )
Take
\[ X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda) \]
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

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

- 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))

\[ \hat{\lambda}(x_1, \ldots, x_n) - E_{\hat{\lambda}(x_1, \ldots, x_n)}[\hat{\lambda}(X_1, \ldots, X_n)] \]

[parametric version]

\[ \hat{\lambda}(x_1, \ldots, x_n) - E_{f_n}[\hat{\lambda}(X_1, \ldots, X_n)] \]

[non-parametric version]
Example (Exponential Sample (4))

In the first (parametric) version,

\[ \frac{1}{\hat{\lambda}(X_1, \ldots, X_n)} \sim \mathcal{G}(n, n\lambda) \]

and

\[ E_\lambda[\hat{\lambda}(X_1, \ldots, X_n)] = \frac{n}{n-1} \lambda \]

due to the bias is **analytically** evaluated as

\[ -\frac{\lambda}{n - 1} \]

and estimated by

\[ -\frac{\hat{\lambda}(X_1, \ldots, X_n)}{n - 1} = -0.00787 \]
Example (Exponential Sample (4))

In the first (parametric) version,

\[ \frac{1}{\hat{\lambda}(X_1, \ldots, X_n)} \sim \text{Ga}(n, n\lambda) \]

and

\[ E_\lambda[\hat{\lambda}(X_1, \ldots, X_n)] = \frac{n}{n - 1}\lambda \]

due to the exponential distribution properties.

Therefore the bias is **analytically** evaluated as

\[ -\frac{\lambda}{n - 1} \]

and estimated by

\[ -\frac{\hat{\lambda}(X_1, \ldots, X_n)}{n - 1} = -0.00787 \]
Example **(Exponential Sample (5))**  
In the second (nonparametric) version, evaluation by Monte Carlo,  

\[ \hat{\lambda}(x_1, \ldots, x_n) - E_{\hat{f}_n}[\hat{\lambda}(X_1, \ldots, X_n)] = 0.00142 \]

which achieves the \textbf{“wrong”} sign
Example (**Exponential Sample (6)**)

**Construction of a confidence interval on \( \lambda \)**

By parametric bootstrap,

\[
\Pr_{\lambda} (\hat{\lambda}_1 \leq \lambda \leq \hat{\lambda}_2) = \Pr (\omega_1 \leq \lambda/\hat{\lambda} \leq \omega_2) = 0.95
\]

can be deduced from

\[
\lambda/\hat{\lambda} \sim \text{Ga}(n, n)
\]

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

\[
[\hat{\lambda}_1, \hat{\lambda}_2] = [0.452, 0.580]
\]
Example (**Exponential Sample (7)**)

In nonparametric bootstrap, one replaces

\[
\Pr_F (q(.025) \leq \lambda(F) \leq q(.975)) = 0.95
\]

with

\[
\Pr_{\hat{F}_n} (q^*(.025) \leq \lambda(\hat{F}_n) \leq q^*(.975)) = 0.95
\]

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 (Exponential Sample (7))

In nonparametric bootstrap, one replaces

\[
\Pr_F (q(.025) \leq \lambda(F) \leq q(.975)) = 0.95
\]

with

\[
\Pr_{\hat{F}_n} (q^*(.025) \leq \lambda(\hat{F}_n) \leq q^*(.975)) = 0.95
\]

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 (Student Sample)

Take

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

\( \mu \) and \( \tau \) could be estimated by

\[
\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \hat{\tau}_n = \sqrt{\frac{5 - 2}{5}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2} = \sqrt{\frac{5 - 2}{5}} \hat{\sigma}_n
\]
Example (Student Sample (2))

**Problem** \( \hat{\mu}_n \) is not distributed from a Student \( \mathcal{S}(5, \mu, \tau^2/n) \) distribution

The distribution of \( \hat{\mu}_n \) can be reproduced by bootstrap sampling
Example (Student Sample (3))
Comparison of confidence intervals

\[ [\hat{\mu}_n - 2 \times \hat{\sigma}_n/10, \hat{\mu}_n + 2 \times \hat{\sigma}_n/10] = [-0.068, 0.319] \]  
[normal approximation]

\[ [q^*(0.05), q^*(0.95)] = [-0.056, 0.305] \]  
[parametric bootstrap approximation]

\[ [q^*(0.05), q^*(0.95)] = [-0.094, 0.344] \]  
[non parametric bootstrap approximation]
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 \( f_\theta \) [wrt a fixed measure \( \nu \)], 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 \( \nu \) 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 \( f_\theta \) [wrt a fixed measure \( \nu \)], 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 \( \nu \) 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 $x_1, \ldots, x_n$ is the function

$$L : \Theta \rightarrow \mathbb{R}_+$$

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

same formula as density but different space of variation
The likelihood

Definition (likelihood function)

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

$$L : \Theta \rightarrow \mathbb{R}_+$$

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

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 \( \theta \)

\[ \theta = 3 \]
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 \( \theta \)

\[ x = 3 \]
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 $\theta$

$\theta = 2$
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 $\theta$

$x = 2$
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 \( \theta \)

\[\theta = 1/2\]
Example: density function versus likelihood function

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

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

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

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

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

$x = 1/2$
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, $p_A$:

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

- 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]
Special case when a random variable $X$ may take specific values $a_1, \ldots, a_k$ and a continuum of values $\mathcal{A}$

**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 continuum of values $\mathcal{A}$

**Example:** Tobit model where $y \sim \mathcal{N}(X^T \beta, \sigma^2)$ but $y^* = y \times \mathbb{I}\{y \geq 0\}$ observed
Special case when a random variable $X$ may take specific values $a_1, \ldots, a_k$ and a continuum of values $\mathcal{A}$.

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

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

Results in likelihood

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

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 $\theta$
- 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 $x_1, \ldots, x_n \sim P(3)$ 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
Likelihood functions for $\chi_1, \ldots, \chi_n \sim \mathcal{N}(0, 1)$ as sample varies
why concentration takes place

Consider

\[ x_1, \ldots, x_n \overset{iid}{\sim} F \]

Then

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

and by \( \text{LLN} \)

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

Lemma

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

\[
\int_{\mathcal{X}} \log \frac{f(x)}{f(x|\theta)} \, dF(x)
\]

© Member of the family closest to true distribution
why concentration takes place

by LLN

\[ \frac{1}{n} \sum_{i=1}^{n} \log f(x_i|\theta) \xrightarrow{LLN} \int_X \log f(x|\theta) \, dF(x) \]

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

\[ \int_X \log \frac{f(x)}{f(x|\theta)} \, dF(x) \]

© Member of the family closest to true distribution
Score function

Score function defined by

\[ \nabla \log L(\theta|x) = \left( \frac{\partial}{\partial \theta_1} L(\theta|x), \ldots, \frac{\partial}{\partial \theta_p} L(\theta|x) \right) / L(\theta|x) \]

Gradient (slope) of likelihood function at point \( \theta \)

**Lemma**

When \( X \sim F_\theta \),

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

Score function defined by

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

Gradient (slope) of likelihood function at point $\theta$

**Lemma**

When $X \sim F_\theta$,

$$E_\theta[\nabla \log L(\theta|X)] = 0$$
Score function

Score function defined by

\[ \nabla \log L(\theta|x) = \left( \frac{\partial}{\partial \theta_1} L(\theta|x), \ldots, \frac{\partial}{\partial \theta_p} L(\theta|x) \right) / L(\theta|x) \]

Gradient (slope) of likelihood function at point \( \theta \)

Lemma

When \( X \sim F_\theta \),

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

Reason:

\[ \int_X \nabla \log L(\theta|x) \, dF_\theta(x) = \int_X \nabla L(\theta|x) \, dx = \nabla \int_X \, dF_\theta(x) \]
Score function

Score function defined by

\[ \nabla \log L(\theta|x) = \left( \frac{\partial}{\partial \theta_1} L(\theta|x), \ldots, \frac{\partial}{\partial \theta_p} L(\theta|x) \right) / L(\theta|x) \]

Gradient (slope) of likelihood function at point \( \theta \)

**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

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

Gradient (slope) of likelihood function at point $\theta$

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 $\theta$
Score function

Score function defined by

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

Gradient (slope) of likelihood function at point $\theta$

Lemma

When $\mathbf{X} \sim F_\theta$,

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

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

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

Often called Fisher information

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

Sometimes denoted \( \mathcal{I}_X \) to stress dependence on distribution of \( X \)
Another notion attributed to Fisher [more likely due to Edgeworth]

Information: covariance matrix of the score vector

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

Often called Fisher information

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

Sometimes denoted \( \mathcal{I}_X \) to stress dependence on distribution of \( X \)
Fisher’s information matrix

Second derivative of the log-likelihood as well

Lemma

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

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

Hence

\[
I_{ij}(\theta) = -\mathbb{E}_{\theta} \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(X|\theta) \right]
\]
Binomial $\mathcal{B}(n, p)$ distribution

$$f(x|p) = \binom{n}{x} p^x (1 - p)^{n-x}$$

$$\frac{\partial}{\partial p} \log f(x|p) = \frac{x}{p} - \frac{n-x}{1-p}$$

$$\frac{\partial^2}{\partial p^2} \log f(x|p) = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}$$

Hence

$$I(p) = \frac{np}{p^2} + \frac{n-np}{(1-p)^2}$$

$$= \frac{n}{p(1-p)}$$
Multinomial $\mathcal{M}(n; p_1, \ldots, p_k)$ distribution

$$f(x|\mathbf{p}) = \binom{n}{x_1 \cdots x_k} p_1^{x_1} \cdots p_k^{x_k}$$

$$\frac{\partial}{\partial p_i} \log f(x|\mathbf{p}) = \frac{x_i}{p_i} - \frac{x_k}{p_k}$$

$$\frac{\partial^2}{\partial p_i \partial p_j} \log f(x|\mathbf{p}) = -\frac{x_k}{p_k^2}$$

$$\frac{\partial^2}{\partial p_i^2} \log f(x|\mathbf{p}) = -\frac{x_i}{p_i^2} - \frac{x_k}{p_k^2}$$

Hence

$$\mathcal{I}(\mathbf{p}) = n \begin{pmatrix} 1/p_1 + 1/p_k & \cdots & 1/p_k \\ 1/p_k & \cdots & 1/p_k \\ \vdots \\ 1/p_k & \cdots & 1/p_{k-1} + 1/p_k \end{pmatrix}$$
Multinomial $\mathcal{M}(n; p_1, \ldots, p_k)$ distribution

$$f(x|p) = \binom{n}{x_1 \ldots x_k} p_1^{x_1} \cdots p_k^{x_k}$$

$$\frac{\partial}{\partial p_i} \log f(x|p) = \frac{x_i}{p_i} - \frac{x_k}{p_k}$$

$$\frac{\partial^2}{\partial p_i \partial p_j} \log f(x|p) = -\frac{x_k}{p_k^2}$$

$$\frac{\partial^2}{\partial p_i^2} \log f(x|p) = -\frac{x_i}{p_i^2} - \frac{x_k}{p_k^2}$$

and

$$\mathcal{J}(p)^{-1} = \frac{1}{n} \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{k-1} \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_{k-1} \\ \cdots & \cdots & \cdots & \cdots \\ -p_1p_{k-1} & -p_2p_{k-1} & \cdots & p_{k-1}(1-p_{k-1}) \end{pmatrix}$$
Illustrations

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

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

$$\frac{\partial}{\partial \mu} \log f(x|\theta) = \frac{x-\mu}{\sigma^2}$$

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

$$\frac{\partial^2}{\partial \mu^2} \log f(x|\theta) = -\frac{1}{\sigma^2}$$

$$\frac{\partial^2}{\partial \mu \partial \sigma} \log f(x|\theta) = -2 \frac{x-\mu}{\sigma^3}$$

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

Hence

$$\mathcal{I}(\theta) = \frac{1}{\sigma^2} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$
Additive features translating as accumulation of information:

- If $X$ and $Y$ are independent, $I_X(\theta) + I_Y(\theta) = I_{(X,Y)}(\theta)$
- $I_{X_1,\ldots,X_n}(\theta) = nI_{X_1}(\theta)$
- If $X = T(Y)$ and $Y = S(X)$, $I_X(\theta) = I_Y(\theta)$
- If $X = T(Y)$, $I_X(\theta) \leq I_Y(\theta)$

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

$$I(\theta) = \begin{bmatrix} \frac{\partial \eta}{\partial \theta} \end{bmatrix}^T I(\eta) \begin{bmatrix} \frac{\partial \eta}{\partial \theta} \end{bmatrix}$$

"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, $\mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta) = \mathcal{I}_{(X,Y)}(\theta)$
- $\mathcal{I}_{X_1,\ldots,X_n}(\theta) = n\mathcal{I}_{X_1}(\theta)$
- if $X = T(Y)$ and $Y = S(X)$, $\mathcal{I}_X(\theta) = \mathcal{I}_Y(\theta)$
- if $X = T(Y)$, $\mathcal{I}_X(\theta) \leq \mathcal{I}_Y(\theta)$

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

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

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

$$\mathcal{J}(\theta) = \left\{ \frac{\partial \eta}{\partial \theta} \right\}^T \mathcal{J}(\eta) \left\{ \frac{\partial \eta}{\partial \theta} \right\}$$

"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]
Back to the Kullback–Leibler divergence

\[ D(\theta', \theta) = \int_X f(x|\theta') \log \frac{f(x|\theta')}{f(x|\theta)} \, dx \]

Using a second degree Taylor expansion

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

approximation of divergence:

\[ D(\theta', \theta) \approx \frac{1}{2} (\theta - \theta')^T \mathcal{I}(\theta')(\theta - \theta') \]

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

Back to the Kullback–Leibler divergence

\[ \mathcal{D}(\theta', \theta) = \int_X f(x|\theta') \log \frac{f(x|\theta')}{f(x|\theta)} \, dx \]

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:

\[ \mathcal{D}(\theta', \theta) \approx \frac{1}{2} (\theta - \theta')^T \mathcal{I}(\theta')(\theta - \theta') \]

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

Back to the Kullback–Leibler divergence

\[ \mathcal{D}(\theta', \theta) = \int_X f(x|\theta') \log \frac{f(x|\theta')}{f(x|\theta)} \, dx \]

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:

\[ \mathcal{D}(\theta', \theta) \approx \frac{1}{2} (\theta - \theta')^T \mathcal{I}(\theta')(\theta - \theta') \]

[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. $f(x|\theta)$,

$$\frac{1}{\sqrt{n}} \nabla \log L(\theta|X_1, \ldots, X_n) \approx \mathcal{N}(0, I_{X_1}(\theta))$$

[at the “true” $\theta$]

Notation $I_1(\theta)$ stands for $I_{X_1}(\theta)$ and indicates information associated with a single observation
Central limit law of the score vector
Given $X_1, \ldots, X_n$ i.i.d. $f(x|\theta)$,

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

[at the “true” $\theta$]

Notation $\mathcal{I}_1(\theta)$ stands for $\mathcal{I}_{X_1}(\theta)$ and indicates information associated with a single observation
What if a transform of the sample $S(X_1, \ldots, X_n)$ contains all the information, i.e.

$$I(X_1, \ldots, X_n)(\theta) = I_S(X_1, \ldots, X_n)(\theta)$$

uniformly in $\theta$?

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

What if a transform of the sample

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

contains all the information, i.e.

$$\mathcal{I}(X_1, \ldots, X_n)(\theta) = \mathcal{I}_S(X_1, \ldots, X_n)(\theta)$$

uniformly in $\theta$?

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 $X_1, \ldots, X_n$]
What if a transform of the sample

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

contains all the information, i.e.

$$\mathcal{I}(X_1, \ldots, X_n)(\theta) = \mathcal{I}_S(X_1, \ldots, X_n)(\theta)$$

uniformly in $\theta$?

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

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

Hence

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

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

\[
L(p|x_1, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i} (1 - p)^{n-x_i} = \frac{p}{1-p} \sum_{i} x_i (1 - p)^n
\]

Hence

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

is sufficient
Illustrations

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

$$L(\mu, \sigma|x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\{-\frac{(x_i - \mu)^2}{2\sigma^2}\}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x}_n + \bar{x}_n - \mu)^2\right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\bar{x}_n - \mu)^2\right\}$$

Hence

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

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:

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

Lemma

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

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

is sufficient
Sufficiency and exponential families

Both previous examples belong to exponential families

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

Generic property of exponential families:

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

**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 $\mathcal{U}(0, \theta)$
If \( X_1, \ldots, X_n \) are iid random variables from a density \( f(\cdot|\theta) \) whose support does not depend on \( \theta \) and verifying the property that there exists an integer \( n_0 \) such that, for \( n \geq n_0 \), 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 $\theta$ and verifying the property that there exists an integer $n_0$ such that, for $n \geq n_0$, 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 \textit{minimal sufficient} if it is a function of any other sufficient statistic

Lemma

For a minimal exponential family representation

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

\( S(X_1) + \ldots + S(X_n) \) 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

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

\( S(X_1) + \ldots + S(X_n) \) is minimal sufficient
Opposite of sufficiency:

Aancillarity

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 $\theta$.

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 \( \theta \).

*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
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 $\theta$.

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 $X_1, \ldots, X_n \overset{iid}{\sim} U(0, \theta)$, $A(X_1, \ldots, X_n) = (X_1, \ldots, X_n)/X_{(n)}$ is ancillary.

2. If $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$,

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

is ancillary.

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

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

[see, e.g., rank tests]
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 $\Psi$ such that $E_\theta[\Psi(A(X_1, \ldots, X_n))] = 0$ for all $\theta$'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 $\theta$ 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 $\Psi$ such that $E_{\theta}[\Psi(A(X_1, \ldots, X_n))] = 0$ for all $\theta$’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 $\theta$ 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 $\sigma$ is known, $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ 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.
Example 1

If $X = (X_1, \ldots, X_n)$ is a random sample from the Normal distribution $\mathcal{N}(\mu, \sigma^2)$ when $\sigma$ is known, $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ 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 \( U(\theta, \theta + 1) \) distribution, \( X \) and \( \lfloor X \rfloor \) are independent, but while \( X \) is complete and sufficient, \( \lfloor X \rfloor \) 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.
Let $X$ be distributed as

<table>
<thead>
<tr>
<th>$x$</th>
<th>$p_x$</th>
<th>$-5$</th>
<th>$-4$</th>
<th>$-3$</th>
<th>$-2$</th>
<th>$-1$</th>
<th>$1$</th>
<th>$2$</th>
<th>$3$</th>
<th>$4$</th>
<th>$5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha' p^2 q$</td>
<td>$\alpha' p q^2$</td>
<td>$p^3/2$</td>
<td>$q^3/2$</td>
<td>$\gamma' p q$</td>
<td>$\gamma' p q$</td>
<td>$q^3/2$</td>
<td>$p^3/2$</td>
<td>$\alpha p q^2$</td>
<td>$\alpha p^2 q$</td>
<td></td>
</tr>
</tbody>
</table>

with

$$\alpha + \alpha' = \gamma + \gamma' = \frac{2}{3}$$

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

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

1. an estimator of \( \theta \) is a statistic \( T(X_1, \ldots, X_N) \) or \( \hat{\theta}_n \) providing a [reasonable] substitute for the unknown value \( \theta \).
2. an estimate of \( \theta \) 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 \( \mu \) and \( \hat{\mu}_N = 2.014 \) is an estimate
When given a parametric family $f(\cdot | \theta)$ and a sample supposedly drawn from this family

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

1. **an estimator** of $\theta$ is a statistic $T(X_1, \ldots, X_N)$ or $\hat{\theta}_n$ providing a [reasonable] substitute for the unknown value $\theta$.

2. **an estimate** of $\theta$ 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 $\mu$ and $\hat{\mu}_N = 2.014$ is an estimate
Rao–Blackwell Theorem

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

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

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

\[
\text{var}_\theta(\delta_1(X)) \leq \text{var}_\theta(\delta(X))
\]

[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 $\theta$ and $T = T(X)$ is a sufficient statistic, then

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

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

$$\text{var}_\theta(\delta_1(X)) \leq \text{var}_\theta(\delta(X))$$

[Rao, 1945; Blackwell, 1947]

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

Estimator $\delta_0$

- 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 $\delta_0$
- 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 $\Psi(\theta)$,

$$\delta_0(X) = \mathbb{E}_\theta[\delta(X)|S(X)]$$
If $\hat{\theta}$ is an estimator of $\theta \in \mathbb{R}$ with bias

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

then

$$\text{var}_\theta(\hat{\theta}) \geq \frac{[1 + b'(\theta)]^2}{J(\theta)}$$

[Fréchet, 1943; Darmois, 1945; Rao, 1945; Cramér, 1946] variance of any unbiased estimator at least as high as inverse Fisher information.
If \( \hat{\theta} \) is an estimator of \( \theta \in \mathbb{R} \) with bias

\[
b(\theta) = E_\theta [\hat{\theta}] - \theta
\]

then

\[
\text{var}_\theta (\hat{\theta}) \geq \frac{[1 + b'(\theta)]^2}{I(\theta)}
\]

[Fréchet, 1943; Darmois, 1945; Rao, 1945; Cramér, 1946] variance of any unbiased estimator at least as high as inverse Fisher information.
If $\delta = \delta(X)$ unbiased estimator of $\Psi(\theta)$, then

$$\text{var}_\theta(\delta) \geq \frac{[\Psi'(\theta)]^2}{\mathcal{I}(\theta)}$$

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

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

And Cauchy-Schwarz implies

$$\text{cov}_\theta(Z, \delta)^2 \leq \text{var}_\theta(\delta)\text{var}_\theta(Z) = \text{var}_\theta(\delta)\mathcal{I}(\theta)$$
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

\[
L(\hat{\theta}_N|X_1, \ldots, X_N) \geq L(\theta_N|X_1, \ldots, X_N) \quad \text{for all } \theta \in \Theta
\]

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 \( \theta \) but makes observation \((x_1, \ldots, x_N)\) most likely...
Given the concentration property of the likelihood function, reasonable choice of estimator as mode:

**MLE**

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

\[
L(\hat{\theta}_N|X_1, \ldots, X_N) \geq L(\theta_N|X_1, \ldots, X_N) \quad \text{for all } \theta \in \Theta
\]

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 \( \theta \) 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

\[
L(\hat{\theta}_N|X_1, \ldots, X_N) \geq L(\theta_N|X_1, \ldots, X_N) \quad \text{for all } \theta \in \Theta
\]

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 \( \theta \) 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 $\theta$, then

$$\hat{\xi}_N^{MLE} = h(\hat{\theta}_N^{MLE})$$

[estimator of transform = transform of estimator]

By extension, if $\xi = h(\theta)$ is any transform of $\theta$, then

$$\hat{\xi}_N^{MLE} = h(\hat{\theta}_n^{MLE})$$

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 \( \theta \), then

\[
\hat{\xi}^\text{MLE}_N = h(\hat{\theta}^\text{MLE}_N)
\]

[estimator of transform = transform of estimator]

By extension, if \( \xi = h(\theta) \) is any transform of \( \theta \), then

\[
\hat{\xi}^\text{MLE}_N = h(\hat{\theta}^\text{MLE}_n)
\]

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 \( \hat{\theta}_n^{\text{MLE}} \)

Case of \( x_1, \ldots, x_n \sim \mathcal{N}(\mu, 1) \)

3. [with \( \tau = +\infty \)]
Unicity of maximum likelihood estimate

Depending on regularity of $L(\cdot|x_1, \ldots, x_N)$, there may be

1. several or an infinity of MLE's [or of solutions to likelihood equations]

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. no MLE at all

2. 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 $\ell(\theta) = \log L(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_n)$:

**Lemma**

If $\Theta$ is connected and open, and if $\ell(\cdot)$ is twice-differentiable with

$$\lim_{\theta \to \partial \Theta} \ell(\theta) < +\infty$$

and if $\mathbf{H}(\theta) = \nabla \nabla^T \ell(\theta)$ 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 \( \Theta \) 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
Lemma

If $f(\cdot|\theta)$ is a minimal exponential family

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

with $T(\cdot)$ one-to-one and twice differentiable over $\Theta$, if $\Theta$ 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 $S(x) = \mathbb{E}_\theta[S(X)]$
lemma

If $\Theta$ 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.
Uniform $\mathcal{U}(0, \theta)$ likelihood

$$L(\theta|x_1, \ldots, x_n) = \theta^{-n} \mathbb{1}_{\theta > \max_i x_i}$$

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

$$\hat{p}_{n}^{\text{MLE}} = \bar{X}_n$$
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

$$(\hat{\mu}_n^{MLE}, \hat{\sigma}_n^{MLE}) = \left( \bar{X}_n, \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right)$$
The fundamental theorem of Statistics

fundamental theorem

Under appropriate conditions, if \((X_1, \ldots, X_n) \overset{iid}{\sim} f(x|\theta)\), if \(\hat{\theta}_n\) is solution of \(\nabla \log f(X_1, \ldots, X_n|\theta) = 0\), then

\[
\sqrt{n}\{\hat{\theta}_n - \theta\} \overset{d}{\longrightarrow} N_p(0, \mathcal{I}(\theta)^{-1})
\]

Equivalent of CLT for estimation purposes

- \(\mathcal{I}(\theta)\) can be replaced with \(\mathcal{I}(\hat{\theta}_n)\)
- or even \(\hat{\mathcal{I}}(\hat{\theta}_n) = -\frac{1}{n} \sum_i \nabla \nabla^T \log f(x_i|\hat{\theta}_n)\)
The fundamental theorem of Statistics

fundamental theorem

Under appropriate conditions, if \((X_1, \ldots, X_n) \overset{iid}{\sim} f(x|\theta)\), if \(\hat{\theta}_n\) is solution of \(\nabla \log f(X_1, \ldots, X_n|\theta) = 0\), then

\[
\sqrt{n}\{\hat{\theta}_n - \theta}\overset{\mathcal{L}}{\rightarrow} \mathcal{N}_p(0, \mathcal{I}(\theta)^{-1})
\]

Equivalent of CLT for estimation purposes

- \(\mathcal{I}(\theta)\) can be replaced with \(\mathcal{I}(\hat{\theta}_n)\)
- or even \(\hat{\mathcal{I}}(\hat{\theta}_n) = -\frac{1}{n} \sum_i \nabla \nabla^T \log f(x_i|\hat{\theta}_n)\)
Assumptions

- $\theta$ identifiable
- support of $f(\cdot|\theta)$ constant in $\theta$
- $\ell(\theta)$ thrice differentiable
- [the killer] there exists $g(x)$ integrable against $f(\cdot|\theta)$ in a neighbourhood of the true parameter such that
  \[
  \left| \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} f(\cdot|\theta) \right| \leq g(x)
  \]
- the following identity stands [mostly superfluous]
  \[
  \mathcal{I}(\theta) = \mathbb{E}_\theta \left[ \nabla \log f(X|\theta) \{ \nabla \log f(X|\theta) \}^T \right] = -\mathbb{E}_\theta \left[ \nabla^T \nabla \log f(X|\theta) \right]
  \]
- $\hat{\theta}_n$ converges in probability to $\theta$ [similarly superfluous]

Example of MLE of $\eta = ||\theta||^2$ when $x \sim \mathcal{N}_p(\theta, I_p)$:

$$\hat{\eta}^{\text{MLE}} = ||x||^2$$

Then $E_\eta[||x||^2] = \eta + p$ diverges away from $\eta$ with $p$

Note: Consistent and efficient behaviour when considering the MLE of $\eta$ based on

$$Z = ||X||^2 \sim \chi_p^2(\eta)$$

[Robert, 2001]
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 $\eta$ with $p$

Note: Consistent and efficient behaviour when considering the MLE of $\eta$ based on

$$Z = ||X||^2 \sim \chi_p^2(\eta)$$

[Robert, 2001]
Example of MLE of $\eta = ||\theta||^2$ when $x \sim \mathcal{N}_p(\theta, I_p)$:

$$\hat{\eta}^{MLE} = ||x||^2$$

Then $\mathbb{E}_\eta[||x||^2] = \eta + p$ diverges away from $\eta$ with $p$

**Note:** Consistent and efficient behaviour when considering the MLE of $\eta$ based on

$$Z = ||X||^2 \sim \chi_p^2(\eta)$$

[Robert, 2001]
Inconsistent MLEs

Take $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta(x)$ with

$$f_\theta(x) = (1 - \theta) \frac{1}{\delta(\theta)} f_0(x - \theta/\delta(\theta)) + \theta f_1(x)$$

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 $\theta$

$$\hat{\theta}_n \overset{\text{a.s.}}{\longrightarrow} 1$$

[Ferguson, 1982; John Wellner’s slides, ca. 2005]
Consider $X_{ij}$ $i = 1, \ldots, n$, $j = 1, 2$ with $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$. Then

\[ \hat{\mu}_{i}^{\text{MLE}} = \frac{X_{i1} + X_{i2}}{2} \quad \hat{\sigma}^{2\text{MLE}} = \frac{1}{4n} \sum_{i=1}^{n} (X_{i1} - X_{i2})^2 \]

Therefore

\[ \hat{\sigma}^{2\text{MLE}} \xrightarrow{\text{a.s.}} \frac{\sigma^2}{2} \]

[Neuymon & Scott, 1948]
Inconsistent MLEs

Consider $X_{ij}$ $i = 1, \ldots, n$, $j = 1, 2$ with $X_{ij} \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$\hat{\mu}_{i}^{\text{MLE}} = \frac{X_{i1} + X_{i2}}{2} \quad \hat{\sigma}^2_{i}^{\text{MLE}} = \frac{1}{4n} \sum_{i=1}^{n} (X_{i1} - X_{i2})^2$$

Therefore

$$\hat{\sigma}^2_{\text{MLE}} \xrightarrow{\text{a.s.}} \sigma^2 / 2$$

[Neuman & 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

\[ \theta^* = \arg \max_{\theta} L(\theta|x) = \prod_{i=1}^{n} g(X_i|\theta). \]

assuming \( X = (X_1, \ldots, X_n) \sim \text{iid } g(x|\theta) \)

- 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

\[ \theta^{(t+1)} = \theta^{(t)} + I^{\text{obs}}(X, \theta^{(t)})^{-1} \nabla \ell(\theta^{(t)}) \]

with \( \ell(.) \) log-likelihood and \( I^{\text{obs}} \) observed information matrix
Likelihood optimisation

Practical optimisation of the likelihood function

$$\theta^* = \arg \max_{\theta} L(\theta|x) = \prod_{i=1}^{n} g(X_i|\theta).$$

assuming $X = (X_1, \ldots, X_n) \sim g(x|\theta)$

- 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

$$\theta^{(t+1)} = \theta^{(t)} + I^{\text{obs}}(X, \theta^{(t)})^{-1} \nabla \ell(\theta^{(t)})$$

with $\ell(.)$ log-likelihood and $I^{\text{obs}}$ observed information matrix
Likelihood optimisation

Practical optimisation of the likelihood function

\[ \theta^* = \arg \max_{\theta} L(\theta|x) = \prod_{i=1}^{n} g(X_i|\theta). \]

assuming \( X = (X_1, \ldots, X_n) \overset{iid}{\sim} g(x|\theta) \)

- 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

\[ \theta^{(t+1)} = \theta^{(t)} + I^{\text{obs}}(X, \theta^{(t)})^{-1} \nabla \ell(\theta^{(t)}) \]

with \( \ell(.) \) log-likelihood and \( I^{\text{obs}} \) observed information matrix
EM algorithm

Cases where $g$ is too complex for the above to work

Special case when $g$ is a marginal

$$g(x|\theta) = \int_Z f(x, z|\theta) \, dz$$

$Z$ called latent or missing variable
Illustrations

- Censored data
  \[ X = \min(X^*, a) \quad X^* \sim \mathcal{N}(\theta, 1) \]

- Mixture model
  \[ X \sim .3 \mathcal{N}_1(\mu_0, 1) + .7 \mathcal{N}_1(\mu_1, 1), \]

- Desequilibrium model
  \[ X = \min(X^*, Y^*) \quad X^* \sim f_1(x|\theta) \quad Y^* \sim f_2(x|\theta) \]
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 $\mathbf{x}$ with $\mathbf{z}$, such as

$$(\mathbf{X}, \mathbf{Z}) \sim f(\mathbf{x}, \mathbf{z}|\theta)$$

$\mathbf{Z}$ missing data vector and pair $(\mathbf{X}, \mathbf{Z})$ complete data vector

Conditional density of $\mathbf{Z}$ given $\mathbf{x}$:

$$k(\mathbf{z}|\theta, \mathbf{x}) = \frac{f(\mathbf{x}, \mathbf{z}|\theta)}{g(\mathbf{x}|\theta)}$$
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

\[
\log L(\theta|x) = \mathbb{E}[\log L^c(\theta|x, Z)|\theta_0, x] - \mathbb{E}[\log k(Z|\theta, x)|\theta_0, x] \tag{1}
\]

for any \(\theta_0\), with integration operated against conditionnal distribution of \(Z\) given observables (and parameters), \(k(z|\theta_0, x)\)
There are “two $\theta$’s”! : in (1), $\theta_0$ is a fixed (and arbitrary) value driving integration, while $\theta$ both free (and variable)

Maximising observed likelihood

$$L(\theta|x)$$

equivalent to maximise r.h.s. term in (1)

$$\mathbb{E}[\log L^c(\theta|x, Z)|\theta_0, x] - \mathbb{E}[\log k(Z|\theta, x)|\theta_0, x]$$
There are “two $\theta$’s”! in (1), $\theta_0$ is a fixed (and arbitrary) value driving integration, while $\theta$ both free (and variable)

Maximising observed likelihood

$$L(\theta|x)$$

equivalent to maximise r.h.s. term in (1)

$$\mathbb{E}[\log L^c(\theta|x, Z)|\theta_0, x] - \mathbb{E}[\log k(Z|\theta, x)|\theta_0, x]$$
Instead of maximising wrt $\theta$ r.h.s. term in (1), maximise only

$$E[\log L^c(\theta|x, Z)|\theta_0, x]$$

Maximisation of complete log-likelihood impossible since $z$ unknown, hence substitute by maximisation of expected complete log-likelihood, with expectation depending on term $\theta_0$.
Instead of maximising wrt $\theta$ r.h.s. term in (1), maximise only

$$E[\log L^c(\theta|\mathbf{x}, \mathbf{Z})|\theta_0, \mathbf{x}]$$

Maximisation of complete log-likelihood impossible since $z$ unknown, hence substitute by maximisation of expected complete log-likelihood, with expectation depending on term $\theta_0$
Expectation–Maximisation

Expectation of complete log-likelihood denoted

\[ Q(\theta|\theta_0, x) = \mathbb{E}[\log L_c(\theta|x, Z)|\theta_0, x] \]

to stress dependence on \( \theta_0 \) and sample \( x \)

Principle

EM derives sequence of estimators \( \hat{\theta}_{(j)} \), \( j = 1, 2, \ldots \), through iteration of Expectation and Maximisation 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**

\[ Q(\theta|\theta_0, x) = \mathbb{E}[\log L^c(\theta|x, Z)|\theta_0, x] \]

to stress dependence on \( \theta_0 \) and sample \( x \)

**Principle**

**EM** derives sequence of estimators \( \hat{\theta}(j), j = 1, 2, \ldots, \) through iteration of **Expectation** and **Maximisation** 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 \( \theta \) 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]
Observed likelihood

\[ L(\theta|x) \]

increases at every EM step

\[ L(\hat{\theta}_{(m+1)}|x) \geq L(\hat{\theta}_{(m)}|x) \]

[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

$$k(z|\theta, x) = \frac{\exp\{-\frac{1}{2}(z - \theta)^2\}}{\sqrt{2\pi[1 - \Phi(a - \theta)]}} = \frac{\varphi(z - \theta)}{1 - \Phi(a - \theta)}, \quad a < z.$$
Normal $\mathcal{N}(\theta, 1)$ sample right-censored

$$L(\theta | x) = \frac{1}{(2\pi)^{m/2}} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{m} (x_i - \theta)^2 \right\} \left[ 1 - \Phi(a - \theta) \right]^{n-m}$$

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

$$k(z|\theta, x) = \frac{\exp\{-\frac{1}{2}(z - \theta)^2\}}{\sqrt{2\pi[1 - \Phi(a - \theta)]}} = \frac{\varphi(z - \theta)}{1 - \Phi(a - \theta)}, \quad a < z.$$
Censored data (2)

At j-th EM iteration

\[ Q(\theta|\hat{\theta}_{(j)}, x) \propto -\frac{1}{2} \sum_{i=1}^{m} (x_i - \theta)^2 - \frac{1}{2} \mathbb{E} \left[ \sum_{i=m+1}^{n} (Z_i - \theta)^2 \ \bigg| \ \hat{\theta}_{(j)}, x \right] \]

\[ \propto -\frac{1}{2} \sum_{i=1}^{m} (x_i - \theta)^2 \]

\[ -\frac{1}{2} \sum_{i=m+1}^{n} \int_{a}^{\infty} (z_i - \theta)^2 k(z|\hat{\theta}_{(j)}, x) \, dz_i \]
Differenciating in $\theta$,

$$n \hat{\theta}_{(j+1)} = m \bar{x} + (n - m) \mathbb{E}[Z|\hat{\theta}_{(j)}],$$

with

$$\mathbb{E}[Z|\hat{\theta}_{(j)}] = \int_{\alpha}^{\infty} zk(z|\hat{\theta}_{(j)}, x) \, dz = \hat{\theta}_{(j)} + \frac{\varphi(a - \hat{\theta}_{(j)})}{1 - \Phi(a - \hat{\theta}_{(j)})}.$$ 

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 $\hat{\theta}$.
Differenciating in $\theta$,

$$n \hat{\theta}_{(j+1)} = m \bar{x} + (n - m) \mathbb{E}[Z|\hat{\theta}_{(j)}],$$

with

$$\mathbb{E}[Z|\hat{\theta}_{(j)}] = \int_{a}^{\infty} zk(z|\hat{\theta}_{(j)}, x) \, dz = \hat{\theta}_{(j)} + \frac{\phi(a - \hat{\theta}_{(j)})}{1 - \Phi(a - \hat{\theta}_{(j)})}.$$ 

Hence, EM sequence provided by

$$\hat{\theta}_{(j+1)} = \frac{m}{n} \bar{x} + \frac{n - m}{n} \left[ \hat{\theta}_{(j)} + \frac{\phi(a - \hat{\theta}_{(j)})}{1 - \Phi(a - \hat{\theta}_{(j)})} \right],$$

which converges to likelihood maximum $\hat{\theta}$
Mixtures

Mixture of two normal distributions with unknown means

\[ 0.3 \mathcal{N}_1(\mu_0, 1) + 0.7 \mathcal{N}_1(\mu_1, 1), \]

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 \( X_i \),

\[ X_i | z_i \sim \mathcal{N}(\mu_{z_i}, 1) \quad Z_i \sim \mathcal{B}(0.7) \]

Complete likelihood

\[
\log L^c(\theta | x, z) \propto -\frac{1}{2} \sum_{i=1}^{n} z_i(x_i - \mu_1)^2 - \frac{1}{2} \sum_{i=1}^{n} (1 - z_i)(x_i - \mu_0)^2
\]

\[ = -\frac{1}{2} n_1 (\hat{\mu}_1 - \mu_1)^2 - \frac{1}{2} (n - n_1)(\hat{\mu}_0 - \mu_0)^2 \]

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 \( X_i \),

\[ X_i|z_i \sim \mathcal{N}(\mu_{z_i}, 1) \quad Z_i \sim \mathcal{B}(.7) \]

Complete likelihood

\[
\begin{align*}
\log L^c(\theta|x, z) & \propto -\frac{1}{2} \sum_{i=1}^{n} z_i(x_i - \mu_1)^2 - \frac{1}{2} \sum_{i=1}^{n} (1 - z_i)(x_i - \mu_0)^2 \\
& = -\frac{1}{2} n_1 (\hat{\mu}_1 - \mu_1)^2 - \frac{1}{2} (n - n_1)(\hat{\mu}_0 - \mu_0)^2
\end{align*}
\]

with

\[
\begin{align*}
n_1 & = \sum_{i=1}^{n} z_i, \quad n_1 \hat{\mu}_1 = \sum_{i=1}^{n} z_i x_i, \quad (n - n_1) \hat{\mu}_0 = \sum_{i=1}^{n} (1 - z_i) x_i
\end{align*}
\]
At $j$-th EM iteration

$$Q(\theta|\hat{\theta}(j), x) = \frac{1}{2} \mathbb{E} [n_1 (\hat{\mu}_1 - \mu_1)^2 + (n - n_1)(\hat{\mu}_0 - \mu_0)^2 | \hat{\theta}(j), x]$$

Differentiating in $\theta$

$$\hat{\theta}_{(j+1)} = \left( \frac{\mathbb{E} [n_1 \hat{\mu}_1 | \hat{\theta}(j), x]}{\mathbb{E} [n_1 | \hat{\theta}(j), x]} \right. \left. \frac{\mathbb{E} [(n - n_1) \hat{\mu}_0 | \hat{\theta}(j), x]}{\mathbb{E} [(n - n_1) | \hat{\theta}(j), x]} \right)$$
Mixtures (3)

Hence $\hat{\theta}_{(j+1)}$ given by

$$
\left( \begin{array}{c}
\sum_{i=1}^{n} \mathbb{E} \left[ Z_i \mid \hat{\theta}_{(j)}, x_i \right] x_i / \sum_{i=1}^{n} \mathbb{E} \left[ Z_i \mid \hat{\theta}_{(j)}, x_i \right] \\
\sum_{i=1}^{n} \mathbb{E} \left[ (1 - Z_i) \mid \hat{\theta}_{(j)}, x_i \right] x_i / \sum_{i=1}^{n} \mathbb{E} \left[ (1 - Z_i) \mid \hat{\theta}_{(j)}, x_i \right]
\end{array} \right)
$$

Conclusion

Step (E) in EM replaces missing data $Z_i$ with their conditional expectation, given $x$ (expectation that depend on $\hat{\theta}_{(m)}$).
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

5 Decision theory and Bayesian analysis
  - Bayesian modelling
  - Conjugate priors
  - Improper prior distributions
  - Bayesian inference
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, $n_{\text{socks}}$ and $n_{\text{pairs}}$
- writing the likelihood is a challenge [exercise]
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, $n_{\text{socks}}$ and $n_{\text{pairs}}$
- writing the likelihood is a challenge [exercise]
A prioris on socks

Given parameters $n_{\text{socks}}$ and $n_{\text{pairs}}$, set of socks

$$S = \{s_1, s_1, \ldots, s_{n_{\text{pairs}}}, s_{n_{\text{pairs}}}, s_{n_{\text{pairs}}+1}, \ldots, s_{n_{\text{socks}}}\}$$

and 11 socks picked at random from $S$ 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 $n_{\text{socks}}$ 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 prioriis on socks

Given parameters $n_{\text{socks}}$ and $n_{\text{pairs}}$, 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 $n_{\text{socks}}$ 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 $n_{\text{socks}}$ and $n_{\text{pairs}}$,

$$n_{\text{socks}} \sim \text{Neg}(30, 15) \quad n_{\text{pairs}} \mid n_{\text{socks}} \sim \frac{n_{\text{socks}}}{2} \text{Be}(15, 2)$$

possible to

1. generate new values of $n_{\text{socks}}$ and $n_{\text{pairs}}$,
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 \( n_{\text{socks}} \) and \( n_{\text{pairs}} \),

\[
\begin{align*}
    n_{\text{socks}} &\sim \text{Neg}(30, 15) \\
    n_{\text{pairs}} | n_{\text{socks}} &\sim n_{\text{socks}} / 2 \text{Be}(15, 2)
\end{align*}
\]

Possible to

1. generate new values of \( n_{\text{socks}} \) and \( n_{\text{pairs}} \),

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
The outcome of this simulation method returns a distribution on the pair \((n_{socks}, n_{pairs})\) that is the conditional distribution of the pair given the observation \(X = 11\).

Proof: Generations from \(\pi(n_{socks}, n_{pairs})\) are accepted with probability

\[
P\{X = 11 | (n_{socks}, n_{pairs})\}
\]
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

\[
\pi(n_{\text{socks}}, n_{\text{pairs}}) \times \mathbb{P}\{X = 11|(n_{\text{socks}}, n_{\text{pairs}})\} = \pi(n_{\text{socks}}, n_{\text{pairs}}|X = 11)
\]
Bayesian principle Given a probability distribution on the parameter $\theta$ called prior

$$\pi(\theta)$$

and an observation $x$ of $X \sim f(x|\theta)$, Bayesian inference relies on the conditional distribution of $\theta$ 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 \( \theta \) 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., \( \{\theta; \pi(\theta|x) \geq \kappa\} \);
- tests of hypotheses, e.g., \( \pi(\theta = 0|x) \); and
- prediction of future observations
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 \( \theta \), independent of the order in which i.i.d. observations are collected \([\text{domino effect}]\)
- 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

\[
L(\theta|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} f(x_i|\theta)
\]

Bayesian inference introduces an extra measure \(\pi(\theta)\) that is chosen \textit{a priori}, hence subjectively by the statistician based on

- hypothetical range of \(\theta\)
- guesstimates of \(\theta\) with an associated (lack of) precision
- type of sampling distribution

\textbf{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 \textit{a priori}, hence subjectively by the statistician based on

- hypothetical range of \( \theta \)
- guesstimates of \( \theta \) with an associated (lack of) precision
- type of sampling distribution

\textbf{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 $W$. 
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

\[
p \sim U([0, 1]) \text{ and } X \sim B(n, p)
\]
Resolution

Since

\[ P(X = x \mid p) = \binom{n}{x} p^x (1 - p)^{n-x} , \]

\[ P(a < p < b \text{ and } X = x) = \int_a^b \binom{n}{x} p^x (1 - p)^{n-x} \, dp \]

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.

\[ p \mid x \sim \text{Be}(x + 1, n - x + 1) \]

[Beta distribution]
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.

$$p | x \sim \text{Be}(x + 1, n - x + 1)$$

[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 $\text{Be}(a, b)$ 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 $\Pi$ of priors $\pi_0$ on $\Theta$ is said to be conjugate if the posterior $\pi(\cdot|y)$ also belong to $\Pi$

In this case, posterior inference is tractable and reduces to updating the hyperparameters* of the prior

Example In Thomas Bayes' example, the $\text{Be}(a, b)$ 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 $\Theta$ 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 $\text{Be}(a,b)$ 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 $F$ of probability distributions on $\Theta$ is *conjugate* for a likelihood function $f(x|\theta)$ if, for every $\pi \in F$, the posterior distribution $\pi(\theta|x)$ also belongs to $F$.

In this case, posterior inference is tractable and reduces to updating the hyperparameters* of the prior

Example In Thomas Bayes’ example, the $\text{Be}(a, b)$ 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

\[
f(x|\theta) = C(\theta)h(x) \exp\{R(\theta) \cdot T(x)\}
= h(x) \exp\{R(\theta) \cdot T(x) - \tau(\theta)\}
\]

allows for conjugate priors

\[
\pi(\theta|\mu, \lambda) = K(\mu, \lambda) e^{\theta \cdot \mu - \lambda \psi(\theta)}
\]

Following Pitman-Koopman-Darmois’ Lemma, only case [besides uniform distributions]
The family of exponential distributions

\[ f(x|\theta) = C(\theta)h(x) \exp\{R(\theta) \cdot T(x)\} \]
\[ = h(x) \exp\{R(\theta) \cdot T(x) - \tau(\theta)\} \]

allows for conjugate priors

\[ \pi(\theta|\mu, \lambda) = K(\mu, \lambda) e^{\theta \cdot \mu - \lambda \psi(\theta)} \]

Following Pitman-Koopman-Darmois’ Lemma, only case [besides uniform distributions]
Illustration

Discrete/Multinomial & Dirichlet

If observations consist of positive counts \( Y_1, \ldots, Y_d \) modelled by a Multinomial \( M(\theta_1, \ldots, \theta_p) \) 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 \( D(\alpha_1, \ldots, \alpha_d) \) distribution

\[
\pi(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{d} \alpha_i)}{\prod_{i=1}^{d} \Gamma(\alpha_i)} \prod_{i=1}^{d} \theta_i^{\alpha_i-1}
\]

defined on the probability simplex \((\theta_i \geq 0, \sum_{i=1}^{d} \theta_i = 1)\), where \( \Gamma \) 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 $N(\theta, \sigma^2)$ | Normal $N(\mu, \tau^2)$ | $N(\rho(\sigma^2\mu + \tau^2x), \rho\sigma^2\tau^2)$ with $\rho = \frac{1}{\sigma^2 + \tau^2}$ |
| Poisson $P(\theta)$ | 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 $\mathcal{B}(\alpha, \beta)$ | $\mathcal{B}(\alpha + x, \beta + n - x)$ |
### Standard exponential families [2]

| $f(x|\theta)$ | $\pi(\theta)$ | $\pi(\theta|x)$ |
|----------------|---------------|-----------------|
| Negative Binomial $\text{Neg}(m, \theta)$ | Beta $\text{Be}(\alpha, \beta)$ | Beta $\text{Be}(\alpha + m, \beta + x)$ |
| Multinomial $\mathcal{M}_k(\theta_1, \ldots, \theta_k)$ | Dirichlet $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$ | Dirichlet $\mathcal{D}(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$ |
| Normal $\mathcal{N}(\mu, 1/\theta)$ | Gamma $\mathcal{G}(\alpha, \beta)$ | $\mathcal{G}(\alpha + 0.5, \beta + (\mu - x)^2/2)$ |
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 \( \sigma \)-finite positive measure \( \pi \) such that

\[
\int_{\Theta} \pi(\theta) \, d\theta = +\infty
\]

Note A \( \sigma \)-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 \(\sigma\)-finite positive measure \(\pi\) such that

\[
\int_{\Theta} \pi(\theta) \, d\theta = +\infty
\]

Improper prior distribution

Note A \(\sigma\)-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 \( \sigma \)-finite positive measure \( \pi \) such that

\[
\int_{\Theta} \pi(\theta) \, d\theta = +\infty
\]

Improper prior distribution

Note A \( \sigma \)-finite density with

\[
\int_{\Theta} \pi(\theta) \, d\theta < +\infty
\]

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
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
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

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
Extension of the posterior distribution $\pi(\theta|x)$ associated with an improper prior $\pi$ 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

$$\int_\Theta f(x|\theta)\pi(\theta) \, d\theta < \infty$$
Extension of the posterior distribution \( \pi(\theta|x) \) associated with an improper prior \( \pi \) 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

\[
\int_{\Theta} f(x|\theta)\pi(\theta) \, d\theta < \infty
\]
If \( x \sim \mathcal{N}(\theta, 1) \) and \( \pi(\theta) = \varpi \), constant, the pseudo marginal distribution is

\[
m(x) = \varpi \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x - \theta)^2}{2} \right\} \, d\theta = \varpi
\]

and the posterior distribution of \( \theta \) is

\[
\pi(\theta \mid x) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x - \theta)^2}{2} \right\},
\]

i.e., corresponds to a \( \mathcal{N}(x, 1) \) distribution. [independent of \( \varpi \)]
Normal illustration

If $x \sim \mathcal{N}(\theta, 1)$ and $\pi(\theta) = \varpi$, constant, the pseudo marginal distribution is

$$m(x) = \varpi \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x - \theta)^2}{2} \right\} \, d\theta = \varpi$$

and the posterior distribution of $\theta$ is

$$\pi(\theta \mid x) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x-\theta)^2}{2} \right\} ,$$

i.e., corresponds to a $\mathcal{N}(x, 1)$ distribution.

[independent of $\omega$]
If $x \sim \mathcal{N}(\theta, 1)$ and $\pi(\theta) = \varpi$, constant, the pseudo marginal distribution is

$$m(x) = \varpi \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x - \theta)^2}{2} \right\} \, d\theta = \varpi$$

and the posterior distribution of $\theta$ is

$$\pi(\theta \mid x) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(x - \theta)^2}{2} \right\} ,$$

i.e., corresponds to a $\mathcal{N}(x, 1)$ distribution. [independent of $\omega$]
The mistake is to think of them [non-informative priors] as representing ignorance

[ Lindley, 1990 ]

Normal illustration:
Consider a $\theta \sim N(0, \tau^2)$ prior. Then

$$\lim_{\tau \to \infty} \mathbb{P}_{\pi}(\theta \in [a, b]) = 0$$

for any $(a, b)$
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 $(a, b)$
Consider a binomial observation, \( x \sim \mathcal{B}(n, p) \), and

\[
\pi^*(p) \propto [p(1 - p)]^{-1}
\]

[Haldane, 1931]

The marginal distribution,

\[
m(x) = \int_0^1 [p(1 - p)]^{-1} \binom{n}{x} p^x (1 - p)^{n-x} \, dp
\]

\[
= B(x, n - x),
\]

is only defined for \( x \neq 0, n \).
Consider a binomial observation, $x \sim B(n, p)$, and

$$
\pi^*(p) \propto [p(1 - p)]^{-1}
$$

The marginal distribution,

$$
m(x) = \int_0^1 [p(1 - p)]^{-1} \binom{n}{x} p^x (1 - p)^{n-x} \, dp
$$

$$
= B(x, n - x),
$$

is only defined for $x \neq 0, n$.

[Not recommended!]

[Haldane, 1931]
The Jeffreys prior

Based on Fisher information

\[ \mathcal{I}(\theta) = E_\theta \left[ \frac{\partial \ell}{\partial \theta^t} \frac{\partial \ell}{\partial \theta} \right] \]

Jeffreys prior density is

\[ \pi^*(\theta) \propto |\mathcal{I}(\theta)|^{1/2} \]

Pros & Cons

- relates to information theory
- agrees with most invariant priors
- parameterisation invariant
The Jeffreys prior

Based on Fisher information

$$\mathcal{I}(\theta) = \mathbb{E}_\theta \left[ \frac{\partial \ell}{\partial \theta^t} \frac{\partial \ell}{\partial \theta} \right]$$

Jeffreys prior density is

$$\pi^*(\theta) \propto |\mathcal{I}(\theta)|^{1/2}$$

Pros & Cons

- relates to information theory
- agrees with most invariant priors
- parameterisation invariant
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!]
If $x \sim \mathcal{B}(n, \theta)$, Jeffreys’ prior is

\[ \mathcal{B}(1/2, 1/2) \]

and, if $n \sim \mathcal{N}(x, \theta)$, Jeffreys’ prior is

\[
\pi_2(\theta) = -\mathbb{E}_\theta \left[ \frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right] \\
= \mathbb{E}_\theta \left[ \frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2} \right] = \frac{x}{\theta^2(1-\theta)}, \\
\propto \theta^{-1}(1-\theta)^{-1/2}
\]
When considering estimates of the parameter $\theta$, one default solution is the maximum a posteriori (MAP) estimator

$$\arg \max_\theta \ell(\theta|x)\pi(\theta)$$

**Motivations**

- Most likely value of $\theta$
- Penalized likelihood estimator
- Further appeal in restricted parameter spaces
MAP estimator

When considering estimates of the parameter $\theta$, one default solution is the maximum a posteriori (MAP) estimator

$$\arg \max_{\theta} \ell(\theta|x)\pi(\theta)$$

Motivations

- Most likely value of $\theta$
- Penalized likelihood estimator
- Further appeal in restricted parameter spaces
Consider \( x \sim \mathcal{B}(n, p) \). Possible priors:

\[
\pi^*(p) = \frac{1}{\text{B}(1/2, 1/2)} p^{-1/2}(1 - p)^{-1/2},
\]

\[
\pi_1(p) = 1 \quad \text{and} \quad \pi_2(p) = p^{-1}(1 - p)^{-1}.
\]

Corresponding MAP estimators:

\[
\delta^*(x) = \max \left( \frac{x - 1/2}{n - 1}, 0 \right),
\]

\[
\delta_1(x) = \frac{x}{n},
\]

\[
\delta_2(x) = \max \left( \frac{x - 1}{n - 2}, 0 \right).
\]
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 \( \theta \) is always
\[ \delta^*(x) = 0 \]
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

\[
g^{\pi}(z|x) = \int_{\Theta} g(z|x, \theta) \pi(\theta|x) \, d\theta.
\]
Consider the AR(1) model

\[ x_t = \rho x_{t-1} + \epsilon_t \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \]

predictive of \( x_T \) is then

\[ x_T | x_{1:(T-1)} \sim \int \frac{\sigma^{-1}}{\sqrt{2\pi}} \exp\left\{ -\frac{(x_T - \rho x_{T-1})^2}{2\sigma^2} \right\} \pi(\rho, \sigma | x_{1:(T-1)}) \, d\rho \, d\sigma, \]

and \( \pi(\rho, \sigma | x_{1:(T-1)}) \) can be expressed in closed form
**Theorem** The solution to

\[
\arg \min_\delta \mathbb{E}^\pi \left[ \| \theta - \delta \|^2 \mid x \right]
\]

is given by

\[
\delta^\pi(x) = \mathbb{E}^\pi [\theta \mid x]
\]

[Posterior mean = Bayes estimator under quadratic loss]
**Theorem** When $\theta \in \mathbb{R}$, the solution to

$$\arg \min_{\delta} \mathbb{E}^{\pi} [ |\theta - \delta| | x ]$$

is given by

$$\delta^{\pi}(x) = \text{median}^{\pi} (\theta | x)$$

[Posterior mean = Bayes estimator under absolute loss]

Obvious extension to

$$\arg \min_{\delta} \mathbb{E}^{\pi} \left[ \sum_{i=1}^{p} |\theta_i - \delta| | x \right]$$
Theorem When $\theta \in \mathbb{R}$, the solution to

$$\arg \min_{\delta} \mathbb{E}^{\pi} \left[ |\theta - \delta| \mid x \right]$$

is given by

$$\delta^{\pi}(x) = \text{median}^{\pi}(\theta \mid x)$$

[Posterior mean = Bayes estimator under absolute loss]

Obvious extension to

$$\arg \min_{\delta} \mathbb{E}^{\pi} \left[ \sum_{i=1}^{p} |\theta_i - \delta| \mid x \right]$$
For conjugate distributions, posterior expectations of the natural parameters may be expressed analytically, for one or several observations.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Conjugate prior</th>
<th>Posterior mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal $\mathcal{N}(\theta, \sigma^2)$</td>
<td>$\mathcal{N}(\mu, \tau^2)$</td>
<td>$\frac{\mu \sigma^2 + \tau^2 x}{\sigma^2 + \tau^2}$</td>
</tr>
<tr>
<td>Poisson $\mathcal{P}(\theta)$</td>
<td>Gamma $\mathcal{G}(\alpha, \beta)$</td>
<td>$\frac{\alpha + x}{\beta + 1}$</td>
</tr>
</tbody>
</table>
Inference with conjugate priors

For conjugate distributions, posterior expectations of the natural parameters may be expressed analytically, for one or several observations.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Conjugate prior</th>
<th>Posterior mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma ( \mathcal{G}(\nu, \theta) )</td>
<td>Gamma ( \mathcal{G}(\alpha, \beta) )</td>
<td>( \frac{\alpha + \nu}{\beta + \chi} )</td>
</tr>
<tr>
<td>Binomial ( \mathcal{B}(n, \theta) )</td>
<td>Beta ( \mathcal{B}(\alpha, \beta) )</td>
<td>( \frac{\alpha + \chi}{\alpha + \beta + n} )</td>
</tr>
<tr>
<td>Negative binomial ( \mathcal{Neg}(n, \theta) )</td>
<td>Beta ( \mathcal{B}(\alpha, \beta) )</td>
<td>( \frac{\alpha + n}{\alpha + \beta + x + n} )</td>
</tr>
<tr>
<td>Multinomial ( \mathcal{M}_k(n; \theta_1, \ldots, \theta_k) )</td>
<td>Dirichlet ( \mathcal{D}(\alpha_1, \ldots, \alpha_k) )</td>
<td>( \frac{\alpha_i + \chi_i}{\left(\sum_j \alpha_j\right) + n} )</td>
</tr>
<tr>
<td>Normal ( \mathcal{N}(\mu, 1/\theta) )</td>
<td>Gamma ( \mathcal{G}(\alpha/2, \beta/2) )</td>
<td>( \frac{\alpha + 1}{\beta + (\mu - \chi)^2} )</td>
</tr>
</tbody>
</table>
Consider
\[ x_1, ..., x_n \sim U([0, \theta]) \]
and \( \theta \sim Pa(\theta_0, \alpha) \). Then
\[ \theta|x_1, ..., x_n \sim Pa(\max(\theta_0, x_1, ..., x_n), \alpha + n) \]
and
\[ \delta^\pi(x_1, ..., x_n) = \frac{\alpha + n}{\alpha + n - 1} \max(\theta_0, x_1, ..., x_n). \]
Natural confidence region based on \( \pi(\cdot | \chi) \) is

\[
\mathcal{C}^{\pi}(\chi) = \{ \theta; \pi(\theta | \chi) > k \}
\]

with

\[
\mathbb{P}^{\pi}(\theta \in \mathcal{C}^{\pi} | \chi) = 1 - \alpha
\]

Highest posterior density (HPD) region
HPD region

Natural confidence region based on $\pi(\cdot|\chi)$ is

$$\mathcal{C}^\pi(\chi) = \{\theta; \pi(\theta|\chi) > k\}$$

with

$$\mathbb{P}^\pi(\theta \in \mathcal{C}^\pi|\chi) = 1 - \alpha$$

Highest posterior density (HPD) region

Example case $\chi \sim \mathcal{N}(\theta, 1)$ and $\theta \sim \mathcal{N}(0, 10)$. Then

$$\theta|\chi \sim \mathcal{N}(10/11\chi, 10/11)$$

and

$$\mathcal{C}^\pi(\chi) = \{\theta; |\theta - 10/11\chi| > k'\}$$

$$= (10/11\chi - k', 10/11\chi + k')$$
Natural confidence region based on \( \pi(\cdot|x) \) is

\[ \mathcal{C}^{\pi}(x) = \{ \theta; \pi(\theta|x) > k \} \]

with

\[ \mathbb{P}^{\pi}(\theta \in \mathcal{C}^{\pi}|x) = 1 - \alpha \]

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