New operational instruments for statistical exploration (=NOISE)

Christian P. Robert

Université Paris Dauphine
http://www.ceremade.dauphine.fr/~xian

Licence MI2E, 2010–2011

Outline

Chapter 1 :
Simulation of random variables

Introduction

Necessity to "reproduce chance" on a computer

- Evaluation of the behaviour of a complex system (network, computer program, queue, particle system, atmosphere, epidemics, economic actions, &tc)
- Determine probabilistic properties of a new statistical procedure or under an unknown distribution [bootstrap]
- Validation of a probabilistic model
- Approximation of an expectation/integral for a non-standard distribution [Law of Large Numbers]
- Maximisation of a weakly regular function/likelihood

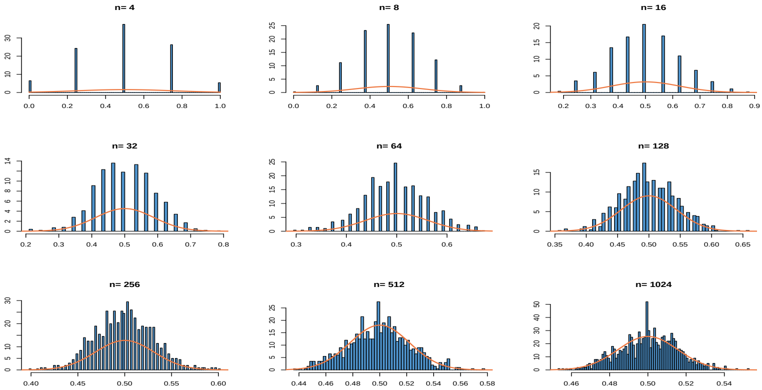
Example (TCL for the binomial distribution)

If

$$X_n \sim \mathcal{B}(n, p)\,,$$

X_n converges in distribution to the normal distribution:

$$\sqrt{n}\,(X_n/n - p) \overset{n\to\infty}{\rightsquigarrow} \mathcal{N}\left(0, \frac{p(1-p)}{n}\right)$$

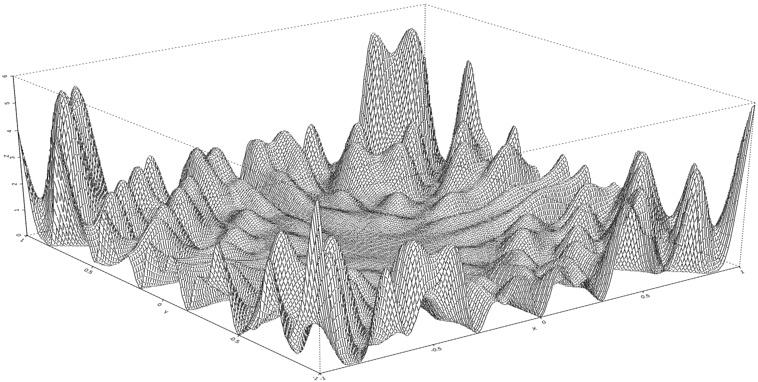

n= 4
n= 8
n= 16
n= 32
n= 64
n= 128
n= 256
n= 512
n= 1024

Example (Stochastic minimisation)

Consider the function

$$\begin{aligned} h(x,y) &= (x\sin(20y)+y\sin(20x))^2\cosh(\sin(10x)x) \\ &+ (x\cos(10y)-y\sin(10x))^2\cosh(\cos(20y)y)\ , \end{aligned}$$

to be minimised. (I know that the global minimum is 0 for $(x,y)=(0,0)$.)


Z
0
1
2
3
4
5
6
Y
1
0.5
0
-0.5
-1
X
-1
-0.5
0
0.5
1

Example (Stochastic minimisation (2))

Instead of solving the first order equations

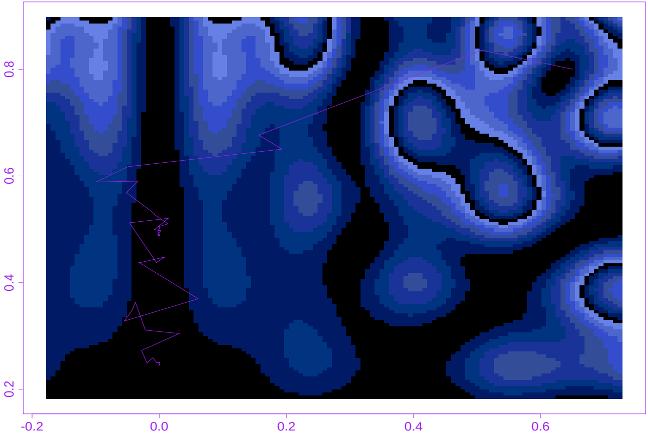
$$\frac{\partial h(x,y)}{\partial x} = 0\,, \quad \frac{\partial h(x,y)}{\partial y} = 0$$

and of checking that the second order conditions are met, we can generate a random sequence in $\mathbb{R}^2$

$$\theta_{j+1} = \theta_j + \frac{\alpha_j}{2\beta_j}\,\Delta h(\theta_j, \beta_j \zeta_j)\,\zeta_j$$

where

- ⋄ the ζ_j's are uniform on the unit circle $x^2 + y^2 = 1$;
- ⋄ $\Delta h(\theta, \zeta) = h(\theta + \zeta) - h(\theta - \zeta)$;
- ⋄ (α_j) and (β_j) converge to 0

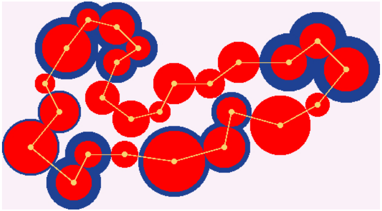


Case when $\alpha_j = 1/10 \log(1+j)$ **et** $\beta_j = 1/j$

The traveling salesman problem

A classical allocation problem:

- Salesman who needs to visit n cities
- Traveling costs between pairs of cities known [and different]
- Search of the optimum circuit



An NP-complete problem

The traveling salesman problem is an example of mathematical problems that require **explosive** resolution times

Number of possible circuits $n!$ and exact solutions available in $\mathrm{O}(2^n)$ time

Numerous practical consequences (networks, integrated circuit design, genomic sequencing, &tc.)

Procter & Gamble competition, 1962

An open problem

Exact solution for $15,112$ German cities found in 2001 in 22.6 CPU years.

Exact solution for the $24,978$ Sweedish cities found in 2004 in 84.8 CPU years.

Resolution via simulation

The **simulated annealing** algorithm:
Repeat

- Random modifications of parts of the original circuit with cost C_0
- Evaluation of the cost C of the new circuit
- Acceptation of the new circuit with probability

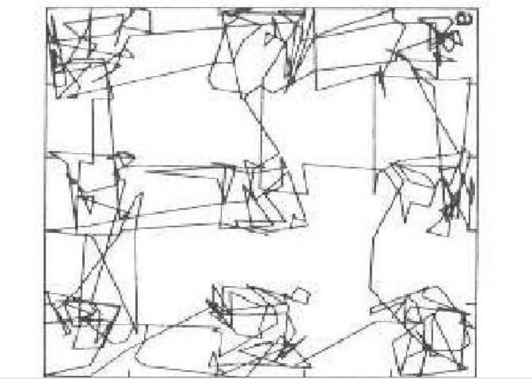
$$\exp\left\{\frac{C_0 - C}{T}\right\} \wedge 1$$

T, **temperature**, is progressively reduced

[Metropolis, 1953]

Illustration

Example (400 cities)



$T = 1.2$

Option pricing

Complicated computation of expectations/average values of options, $\mathbb{E}[C_T]$, necessary to evaluate the entry price $(1+r)^{-T}\,\mathbb{E}[C_T]$

Example (European options)

Case when

$$C_T = (S_T - K)^+$$

with

$$S_T = S_0 \times Y_1 \times \cdots \times Y_T\,, \mathsf{Pr}(Y_i = u) = 1 - \mathsf{Pr}(Y_i = d) = p\,.$$

Resolution via the simulation of the binomial rv's Y_i

Option pricing (cont'd)

Example (Asian options)

Continuous time model where

$$C_T = \left(\frac{1}{T} \int_0^T S(t) \mathrm{d}t - K \right)^+ \approx \left(\frac{1}{T} \sum_{n=1}^{T} S(n) - K \right)^+ ,$$

with

$$S(n+1) = S(n) \times \exp\left\{ \Delta X(n+1) \right\}, \Delta X(n) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) .$$

Resolution via the simulation of the normal rv's ΔX_i

Pseudo-random generator

Pivotal element of simulation techniques: they all require the availability of uniform $\mathscr{U}(0,1)$ random variables via transformations

Definition (**Pseudo-random generator**)

Un *Pseudo-random generator* is a **deterministic** Ψ from $]0,1[$ to $]0,1[$ such that, for any starting value u_0 and any n, the sequence

$$\{u_0, \Psi(u_0), \Psi(\Psi(u_0)), \ldots, \Psi^n(u_0)\}$$

behaves (statistically) like an iid sequence $\mathscr{U}(0,1)$

¡Paradox!

While avoiding randomness, the deterministic sequence $(u_0, u_1 = \Psi(u_0), \ldots, u_n = \Psi(u_{n-1}))$ must resemble a random sequence!

In R, use of the procedure

```
runif( )
```

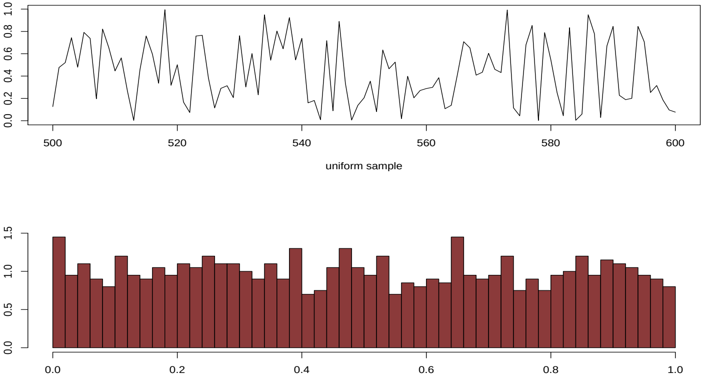
Description:

'runif' generates random deviates.

Example:

u = runif(20)

'Random.seed' is an integer vector, containing the random number generator (RNG) state for random number generation in R. It can be saved and restored, but should not be altered by the user.


0.0
0.2
0.4
0.6
0.8
1.0
500
520
540
560
580
600
uniform sample
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0

In C, use of the procedure

rand() / random()

SYNOPSIS

include <stdlib.h>

long int random(void);

DESCRIPTION

The random() function uses a non-linear additive feedback random number generator employing a default table of size 31 long integers to return successive pseudo-random numbers in the range from 0 to RAND MAX. The period of this random generator is very large, approximately 16*((2**31)-1).

RETURN VALUE

random() returns a value between 0 and RAND MAX.

En Scilab, use of the procedure

rand()

rand() : with no arguments gives a scalar whose value changes each time it is referenced. By default, random numbers are uniformly distributed in the interval (0,1). rand('normal') switches to a normal distribution with mean 0 and variance 1. rand('uniform') switches back to the uniform distribution

EXAMPLE

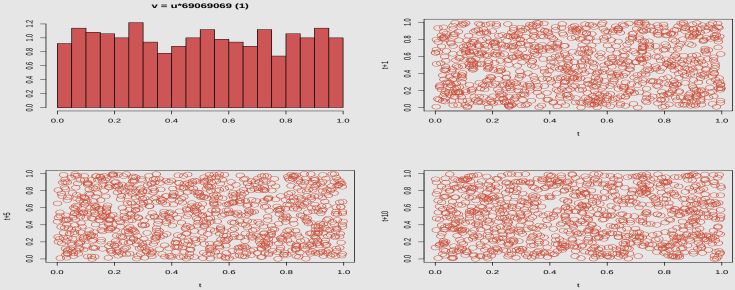
x=rand(10,10,'uniform')

Example (A standard uniform generator)

The congruencial generator

$$D(x) = (ax + b) \bmod (M + 1).$$

has a period of M for proper choices of (a, b) and becomes a generator on $]0, 1[$ when dividing by $M + 2$



Conclusion :

Use the appropriate random generator on the computer or the software at hand instead of constructing a random generator of poor quality

Distributions different from the uniform distribution (1)

A problem formaly solved since

Theorem (Generic inversion)

If U is a uniform random variable on $[0,1)$ and if F_X is the cdf of the random variable X, then $F_X^{-1}(U)$ is distributed like X

Proof. Indeed,

$$P(F_X^{-1}(U) \leq x) = P(U \leq F_X(x)) = F_X(x)$$

Note. When F_X is not strictly increasing, we can take

$$F_X^{-1}(u) = \inf \{x; F_X(x) \geq u\}$$

Applications...

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

$$F_X(x) = \sum_{i \leq x} \binom{n}{i} p^i (1-p)^{n-i}$$

and $F_X^{-1}(u)$ can be obtained numerically

- Exponential distribution, $\mathscr{E}xp(\lambda)$,

$$F_X(x) = 1 - \exp(-\lambda x) \qquad \text{et} \qquad F_X^{-1}(u) = -\log(\mathbf{u})/\lambda$$

- Cauchy distribution, $\mathscr{C}(0,1)$,

$$F_X(x) = \frac{1}{\pi}\arctan(x) + \frac{1}{2} \qquad \text{et} \qquad F_X^{-1}(u) = \tan(\pi(u - 1/2))$$

Other transformations...

[Hint]

Find transforms linking the distribution of interest with simpler/know distributions

Example (Box-Müller transform)

For the normal distribution $\mathscr{N}(0,1)$, if $X_1, X_2 \overset{i.i.d.}{\sim} \mathscr{N}(0,1)$,

$$X_1^2 + X_2^2 \sim \chi_2^2, \qquad \arctan(X_1/X_2) \sim \mathscr{U}([0,2\pi])$$

[Jacobian]

Since the χ_2^2 distribution is the same as the $\mathscr{E}xp(1/2)$ distribution, using a cdf inversion produces

$$X_1 = \sqrt{-2\,\log(U_1)}\,\sin(2\pi U_2) \qquad X_2 = \sqrt{-2\,\log(U_1)}\,\cos(2\pi U_2)$$

Example

Student's t and Fisher's F distributions are natural byproducts of the generation of the normal and of the chi-square distributions.

Example

The Cauchy distribution can be derived from the normal distribution as: if $X_1, X_2 \overset{i.i.d.}{\sim} \mathscr{N}(0,1)$, then $X_1/X_2 \sim \mathscr{C}(0,1)$

Example

The Beta distribution $\mathscr{B}(\alpha, \beta)$, with density

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}\,,$$

can be derived from the Gamma distribution by: if $X_1 \sim \mathscr{G}a(\alpha, 1)$, $X_2 \sim \mathscr{G}a(\beta, 1)$, then

$$\frac{X_1}{X_1 + X_2} \sim \mathscr{B}(\alpha, \beta)$$

Multidimensional distributions

Consider the generation of

$$(X_1, \ldots, X_p) \sim f(x_1, \ldots, x_p)$$

in $\mathbb{R}^p$ with components that are not necessarily independent

Cascade rule

$$f(x_1, \ldots, x_p) = f_1(x_1) \times f_{2|1}(x_2|x_1) \ldots \times f_{p|-p}(x_p|x_1, \ldots, x_{p-1})$$

Implementation

Simulate for $t = 1, \ldots, T$

1. $X_1 \sim f_1(x_1)$
2. $X_2 \sim f_{2|1}(x_2|x_1)$

$\vdots$

p. $X_p \sim f_{p|-p}(x_p|x_1, \ldots, x_{p-1})$

Distributions different from the uniform distribution (2)

- F_X^{-1} rarely available
- implemented algorithm in a resident software only for standard distributions
- inversion lemma does not apply in larger dimensions
- new distributions may require fast resolution

The Accept-Reject Algorithm

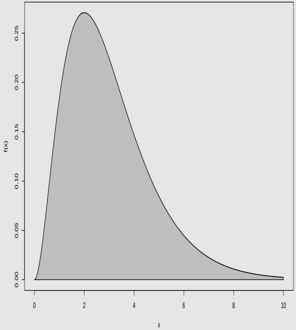
Given a distribution with density f to be simulated

Theorem (Fundamental theorem of simulation)

The uniform distribution on the sub-graph

$$\mathscr{S}_f = \{(x,u); 0 \leq u \leq f(x)\}$$

produces a marginal in x with density f.



Proof :

Marginal density given by

$$\int_0^\infty \mathbb{I}_{0\leq u\leq f(x)} \mathrm{d}u = f(x)$$

and independence from the normalisation constant

Example

For a normal distribution, we just need to simulate (u, x) at random in

$$\{(u, x);\ 0 \leq u \leq \exp(-x^2/2)\}$$

Accept-reject algorithm

1. Find a density g that can be simulated and such that
$$\sup_x \frac{f(x)}{g(x)} = M < \infty$$
2. Generate
$$Y_1, Y_2, \ldots \overset{i.i.d.}{\sim} g\,, \qquad U_1, U_2, \ldots \overset{i.i.d.}{\sim} \mathscr{U}([0,1])$$
3. Take $X = Y_k$ where
$$k = \inf\{n\,; U_n \leq f(Y_n)/Mg(Y_n)\}$$

Theorem (Accept–reject)

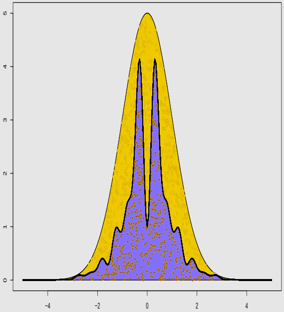
The random variable produced by the above stopping rule is distributed form f_X

Proof (1) : We have

$$
\begin{aligned}
P(X \leq x) &= \sum_{k=1}^{\infty} P(X = Y_k \, , Y_k \leq x) \\
&= \sum_{k=1}^{\infty} \left(\mathbf{1} - \frac{\mathbf{1}}{\mathbf{M}}\right)^{\mathbf{k-1}} P(U_k \leq f(Y_k)/Mg(Y_k) \, , Y_k \leq x) \\
&= \sum_{k=1}^{\infty} \left(1 - \frac{1}{M}\right)^{k-1} \int_{-\infty}^{x} \int_{0}^{f(y)/Mg(y)} du \; g(y) \mathrm{d}y \\
&= \sum_{k=1}^{\infty} \left(1 - \frac{1}{M}\right)^{k-1} \frac{1}{M} \int_{-\infty}^{x} f(y) \mathrm{d}y
\end{aligned}
$$

Proof (2)

If (X, U) is uniform on $A \supset B$, the distribution of (X, U) restricted to B is uniform on B.



Properties

- Does not require a normalisation constant
- Does not require an exact upper bound M
- Allows for the recycling of the Y_k's for another density f (note that rejected Y_k's are no longer distributed from g)
- Requires on average M Y_k's for one simulated X (efficiency measure)

Example

Take $f(x) = \exp(-x^2/2)$ et $g(x) = 1/(1+x^2)$

$$\frac{f(x)}{g(x)} = (1+x^2)\, e^{-x^2/2} \leq 2/\sqrt{e}$$

Probability of acceptance $\sqrt{e/2\pi} = 0.66$

Theorem (Envelope)

If there exists a density g_m, a function g_l and a constant M such that

$$g_l(x) \leq f(x) \leq M g_m(x) ,$$

then

1. Generate $X \sim g_m(x)$, $U \sim \mathcal{U}_{[0,1]}$;
2. Accept X if $U \leq g_l(X)/M g_m(X)$;
3. else, accept X if $U \leq f(X)/M g_m(X)$

produces random variable distributed from f.

Uniform ratio algorithms

Slice sampler

Result :

Uniform simulation on

$$\{(u, v);\ 0 \leq u \leq \sqrt{2f(v/u)}\}$$

produces

$$X = V/U \sim f$$

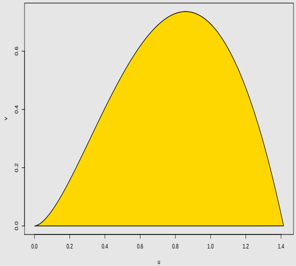
Proof :

Change of variable $(u, v) \rightarrow (x, u)$ with Jacobian u and marginal distribution of x provided by

$$x \sim \int_0^{\sqrt{2f(x)}} u \, \mathrm{d}u = \frac{\sqrt{2f(x)}^2}{2} = f(x)$$

Example

For a normal distribution, simulate (u, v) at random in



$$\{(u,v);\ 0 \leq u \leq \sqrt{2}\, e^{-v^2/4u^2}\} = \{(u,v);\ v^2 \leq -4\, u^2\ \log(u/\sqrt{2})\}$$

Slice sampler

If a uniform simulation on

$$\mathfrak{G} = \{(u, x);\, 0 \leq u \leq f(x)\}$$

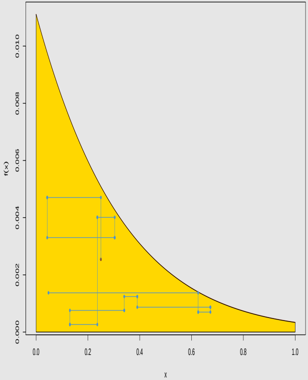
is too complex [because of the inversion of x into $u \leq f(x)$], we can use instead a **random walk** on $\mathfrak{G}$:

Slice sampler

Simulate for $t = 1, \ldots, T$

1. $\omega^{(t+1)} \sim \mathscr{U}_{[0, f(x^{(t)})]}$;
2. $x^{(t+1)} \sim \mathscr{U}_{\mathfrak{S}^{(t+1)}}$, where

$$\mathfrak{S}^{(t+1)} = \{y;\ f(y) \geq \omega^{(t+1)}\}.$$



Justification

The **random walk** is exploring uniformly $\mathfrak{G}$:

If

$$(U^{(t)}, X^{(t)}) \sim \mathscr{U}_{\mathfrak{G}}\,,$$

then

$$(U^{(t+1)}, X^{(t+1)}) \sim \mathscr{U}_{\mathfrak{G}}\,.$$

Proof:

$$
\begin{aligned}
&\mathsf{Pr}((U^{(t+1)}, X^{(t+1)}) \in A \times B) \\
= &\int \int \int_B \int_A \mathbb{I}_{0 \le u \le f(x)} \, \frac{\mathbb{I}_{0 \le u' \le f(x)}}{f(x)} \, \frac{\mathbb{I}_{f(x') \ge u'}(x')}{\int \mathbb{I}_{f(y) \ge u'} \mathrm{d}y} \, d(x, u, x', u') \\
= &\int \int_B \int_A f(x) \, \frac{\mathbb{I}_{0 \le u' \le f(x)}}{f(x)} \, \frac{\mathbb{I}_{f(x') \ge u'}(x')}{\int \mathbb{I}_{f(y) \ge u'} \mathrm{d}y} \, d(x, x', u') \\
= &\int \mathbb{I}_{f(x) \ge u'} \mathrm{d}x \int_B \int_A \frac{\mathbb{I}_{f(x') \ge u'}(x')}{\int \mathbb{I}_{f(y) \ge u'} \mathrm{d}y} \, d(x', u') \\
= &\int_B \int_A \mathbb{I}_{f(x') \ge u' \ge 0} \, d(x', u')
\end{aligned}
$$

Example (**Normal distribution**)

For the standard normal distribution,

$$f(x) \propto \exp(-x^2/2),$$

a slice sampler is

$$\begin{aligned} \omega | x &\sim \mathscr{U}_{[0, \exp(-x^2/2)]} \,, \\ X | \omega &\sim \mathscr{U}_{[-\sqrt{-2\log(\omega)}, \sqrt{-2\log(\omega)}]} \end{aligned}$$

Note

The technique also operates when f is replaced with

$$\varphi(x) \propto f(x)$$

It can be generalised to the case when f is decomposed in

$$f(x) = \prod_{i=1}^{p} f_i(x)$$

Example (Truncated normal distribution)

If we consider instead the truncated $\mathcal{N}(-3,1)$ distribution restricted to $[0,1]$, with density

$$f(x) = \frac{\exp(-(x+3)^2/2)}{\sqrt{2\pi}[\Phi(4)-\Phi(3)]} \propto \exp(-(x+3)^2/2) = \varphi(x)\,,$$

a slice sampler is

$$\begin{array}{rcl} \omega|x & \sim & \mathscr{U}_{[0,\exp(-(x+3)^2/2)]}\,, \\ X|\omega & \sim & \mathscr{U}_{[0,1\wedge\{-3+\sqrt{-2\log(\omega)}\}]} \end{array}$$

The Metropolis–Hastings algorithm

Generalisation of the slice sampler to situations when the slice sampler cannot be easily implemented

Idea

Create a sequence $(X_n)_n$ such that, for n 'large enough', the density of X_n is close to f

The Metropolis–Hastings algorithm (2)

If f is the density of interest, we pick a proposal conditional density

$$q(y|x)$$

such that

- it is easy to simulate
- it is positive everywhere f is positive

Metropolis–Hastings

For a current value $X^{(t)} = x^{(t)}$,

1. Generate $Y_t \sim q(y|x^{(t)})$.
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with proba. } \rho(x^{(t)}, Y_t), \\ x^{(t)} & \text{with proba. } 1 - \rho(x^{(t)}, Y_t), \end{cases}$$

where

$$\rho(x, y) = \min \left\{ \frac{f(y)}{f(x)} \, \frac{q(x|y)}{q(y|x)} \, , 1 \right\} .$$

Properties

- Always accept moves to y_t's such that

$$\frac{f(y_t)}{q(y_t|x_t)} \geq \frac{f(x_t)}{q(x_t|y_t)}$$

- Does not depend on normalising constants for both f and $q(\cdot|x)$ (if the later is independent from x)
- Never accept values of y_t such that $f(y_t) = 0$
- The sequence $(x^{(t)})_t$ can take repeatedly the same value
- The $X^{(t)}$'s are dependent (Marakovian) random variables

links

Justification

Joint distribution of $(X^{(t)}, X^{(t+1)})$

If $X^{(t)} \sim f(x^{(t)})$,

$$\begin{aligned}(X^{(t)}, X^{(t+1)}) \quad \sim \quad & f(x^{(t)}) \Big\{ \rho(x^{(t)}, x^{(t+1)}) \times q(x^{(t+1)}|x^{(t)}) \\ & \qquad\qquad\qquad [Y_t \text{ accepted}] \\ & + \int \Big[1 - \rho(x^{(t)}, y)\Big] \, q(y|x^{(t)}) \, \mathrm{d}y \, \mathbb{I}_{x^{(t)}}(x^{(t+1)}) \Big\} \\ & \qquad\qquad\qquad [Y_t \text{ rejected}]\end{aligned}$$

Balance condition

$$\begin{aligned} f(x) \times \rho(x,y) \times q(y|x) &= f(x) \min\left\{ \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}, 1 \right\} q(y|x) \\ &= \min\left\{ f(y)q(x|y), f(x)q(y|x) \right\} \\ &= f(y) \times \rho(y,x) \times q(x|y) \end{aligned}$$

Thus **the distribution of $(X^{(t)}, X^{(t+1)})$ as the distribution of $(X^{(t+1)}, X^{(t)})$:** if $X^{(t)}$ has the density f, then so does $X^{(t+1)}$

Link with slice sampling

The slice sampler is a very special case of Metropolis-Hastings algorithm where the acceptance probability is always 1

1. for the generation of U,

$$\frac{\mathbb{I}_{0\leq u'\leq f(x)}}{\mathbb{I}_{0\leq u\leq f(x)}} \times \frac{f(x)^{-1}\,\mathbb{I}_{0\leq u'\leq f(x)}}{f(x)^{-1}\,\mathbb{I}_{0\leq u\leq f(x)}} = 1$$

[joint density] [conditional density]

2. pour la génération de X,

$$\frac{\mathbb{I}_{0\leq u\leq f(y)}}{\mathbb{I}_{0\leq u\leq f(x)}} \times \frac{\mathbb{I}_{\{z;u\leq f(z)\}}(x)}{\mathbb{I}_{\{z;u\leq f(z)\}}(y)}\,\frac{\int_{\{z;u\leq f(z)\}} f(z)\,\mathrm{d}z}{\int_{\{z;u\leq f(z)\}} f(z)\,\mathrm{d}z} = 1$$

[joint density] [conditional density]

Independent proposals

Proposal q independent from $X^{(t)}$, denoted g as in Accept-Reject algorithms.

Independent Metropolis-Hastings

For the current value $X^{(t)} = x^{(t)}$,

1. Generate $Y_t \sim g(y)$
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with proba. } \min\left\{ \dfrac{f(Y_t)\, g(x^{(t)})}{f(x^{(t)})\, g(Y_t)}\,, 1 \right\}, \\ x^{(t)} & \text{otherwise.} \end{cases}$$

Properties

- Alternative to Accept-Reject
- Avoids the computation of $\max f(x)/g(x)$
- Accepts more often than Accept-Reject
- If x_t achieves $\max f(x)/g(x)$, this is almost identical to Accept-Reject
- Except that the sequence (x_t) is not independent

Example (Gamma distribution)

Generate a distribution $\mathcal{G}a(\alpha, \beta)$ from a proposal $\mathcal{G}a(\lfloor\alpha\rfloor, b = \lfloor\alpha\rfloor/\alpha)$, where $\lfloor\alpha\rfloor$ is the integer part of α (this is a sum of exponentials)

1. Generate $Y_t \sim \mathcal{G}a(\lfloor\alpha\rfloor, \lfloor\alpha\rfloor/\alpha)$
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with prob. } \left(\dfrac{Y_t}{x^{(t)}} \exp\left\{\dfrac{x^{(t)} - Y_t}{\alpha}\right\}\right)^{\alpha-\lfloor\alpha\rfloor} \\ x^{(t)} & \text{else.} \end{cases}$$

Random walk Metropolis–Hastings

Proposal

$$Y_t = X^{(t)} + \varepsilon_t,$$

where $\varepsilon_t \sim g$, independent from $X^{(t)}$, and g *symmetrical*
Instrumental distribution with density

$$g(y - x)$$

Motivation

local perturbation of $X^{(t)}$ / exploration of its neighbourhood

Random walk Metropolis–Hastings

Starting from $X^{(t)} = x^{(t)}$

1. Generate $Y_t \sim g(y - x^{(t)})$
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with proba. } \min\left\{1, \dfrac{f(Y_t)}{f(x^{(t)})}\right\}, \\ & \quad \text{[symmetry of } g\text{]} \\ x^{(t)} & \text{otherwise} \end{cases}$$

Properties

- Always accepts higher point and sometimes lower points (see gradient algorithm)
- Depends on the dispersion de g
- Average robability of acceptance

$$\varrho = \int\int \min\{f(x), f(y)\} g(y-x)\, \mathrm{d}x \mathrm{d}y$$

- close to 1 if g has a small variance **[Danger!]**
- far from 1 if g has a large variance **[Re-Danger!]**

Example (**Normal distribution**)

Generate $\mathcal{N}(0,1)$ based on a uniform perturbation on $[-\delta,\delta]$

$$Y_t = X^{(t)} + \delta\omega_t$$

Acceptance probability

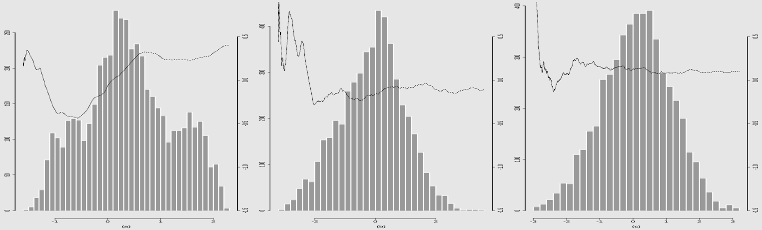
$$\rho(x^{(t)}, y_t) = \exp\{(x^{(t)^2} - y_t^2)/2\} \wedge 1.$$

Example (Normal distribution (2))

Statistics based on 15000 simulations

δ	0.1	0.5	1.0
mean	0.399	-0.111	0.10
variance	0.698	1.11	1.06

When $\delta \uparrow$, faster exploration of the support of f.



3 samples with $\delta = 0.1, 0.5$ and 1.0, with convergence of empirical averages (over 15000 simulations).

Missing variable models

Special case when the density to simulate can be written as

$$f(x) = \int_{\mathcal{Z}} \tilde{f}(x, z) \mathsf{d}z$$

The random variable Z is then called **missing data**

Completion principe

Idea

Simulate $\tilde{f}$ produces simulations from f

If

$$(X, Z) \sim \tilde{f}(x, z)\,,$$

marginaly

$$X \sim f(x)$$

Data Augmentation

Starting from $x^{(t)}$,

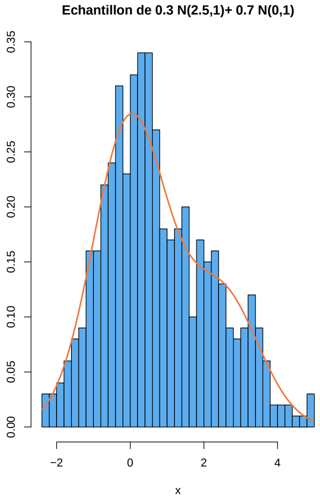
1. Simulate $Z^{(t+1)} \sim \tilde{f}_{Z|X}(z|x^{(t)})$;
2. Simuleater $X^{(t+1)} \sim \tilde{f}_{X|Z}(x|z^{(t+1)})$.

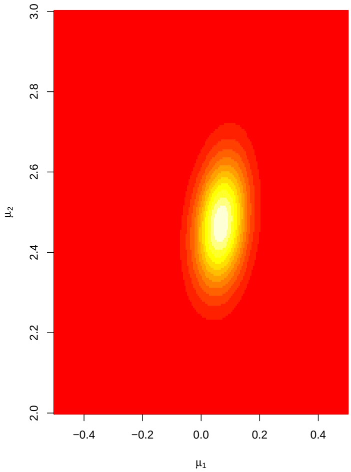
Example (**Mixture of distributions**)

Consider the simulation (in $\mathbb{R}^2$) of the density $f(\mu_1, \mu_2)$ proportional to

$$e^{-\mu_1^2-\mu_2^2} \times \prod_{i=1}^{100} \left\{0.3\, e^{-(x_i-\mu_1)^2/2} + 0.7\, e^{-(x_i-\mu_2)^2/2}\right\}$$

when the x_i's are given/observed.





Histogram of the x_i's and level set of $f(\mu_1, \mu_2)$

Completion (1)

Replace every sum in the density with an integral:

$$0.3\,e^{-(x_i-\mu_1)^2/2}+0.7\,e^{-(x_i-\mu_2)^2/2}=\int\Big(\mathbb{I}_{[0,0.3\,e^{-(x_i-\mu_1)^2/2}]}(u_i)$$

$$+\mathbb{I}_{[0.3\,e^{-(x_i-\mu_1)^2/2},0.3\,e^{-(x_i-\mu_1)^2/2}+0.7\,e^{-(x_i-\mu_2)^2/2}]}(u_i)\Big)\,\mathrm{d}u_i$$

and simulate $((\mu_1,\mu_2),(U_1,\ldots,U_n))=(X,Z)$ via Data Augmentation

Completion (2)

Replace the U_i's by the ξ_i's, where

$$\xi_i = \begin{cases} 1 & \text{si } U_i \leq 0.3\, e^{-(x_i-\mu_1)^2/2}, \\ 2 & \text{sinon} \end{cases}$$

Then

$$\begin{aligned} \Pr(\xi_i = 1|\mu_1, \mu_2) &= \frac{0.3\, e^{-(x_i-\mu_1)^2/2}}{0.3\, e^{-(x_i-\mu_1)^2/2} + 0.7\, e^{-(x_i-\mu_2)^2/2}} \\ &= 1 - \Pr(\xi_i = 2|\mu_1, \mu_2) \end{aligned}$$

Conditioning (1)

The conditional distribution of $Z = (\xi_1, \ldots, \xi_n)$ given $X = (\mu_1, \mu_2)$ is given by

$$\begin{aligned}\Pr(\xi_i = 1|\mu_1, \mu_2) &= \frac{0.3\, e^{-(x_i-\mu_1)^2/2}}{0.3\, e^{-(x_i-\mu_1)^2/2} + 0.7\, e^{-(x_i-\mu_2)^2/2}} \\ &= 1 - \Pr(\xi_i = 2|\mu_1, \mu_2)\end{aligned}$$

Conditioning (2)

The conditional distribution of $X = (\mu_1, \mu_2)$ given $Z = (\xi_1, \ldots, \xi_n)$ is given by

$$
\begin{aligned}
(\mu_1, \mu_2)|Z \quad &\sim \quad e^{-\mu_1^2 - \mu_2^2} \times \prod_{\{i;\xi_i=1\}} e^{-(x_i-\mu_1)^2/2} \times \prod_{\{i;\xi_i=2\}} e^{-(x_i-\mu_2)^2/2} \\
&\propto \quad \exp\left\{-(n_1+2)\left(\mu_1 - \frac{n_1\hat{\mu}_1}{n_1+2}\right)^2 /2\right\} \\
&\qquad \times \exp\left\{-(n_2+2)\left(\mu_2 - \frac{n_2\hat{\mu}_2}{n_2+2}\right)^2 /2\right\}
\end{aligned}
$$

where n_j is the number of ξ_i's equal to j and $n_j\hat{\mu}_j$ is the sum of the x_i's associated with those ξ_i equal to j

[Easy!]

Chapter 2 :
Monte Carlo Methods & EM algorithm

- Introduction
- Integration by Monte Carlo method
- Importance functions
- Acceleration methods

Uses of simulation

1. integration

$$\mathfrak{I} = \mathbb{E}_f[h(X)] = \int h(x) f(x) dx$$

2. limiting behaviour/stationarity of complex systems
3. optimisation

$$\arg\min_x h(x) = \arg\max_x \exp\{-\beta h(x)\} \qquad \beta > 0$$

Example (Propagation of an epidemic)

On a grid representing a region, a point is given by its coordinates x, y

The probability to catch a disease is

$$P_{x,y} = \frac{\exp(\alpha + \beta \cdot n_{x,y})}{1 + \exp(\alpha + \beta \cdot n_{x,y})} \, \mathbb{I}_{n_{x,y}>0}$$

if $n_{x,y}$ denotes the number of neighbours of (x, y) who alread have this disease

The probability to get healed is

$$Q_{x,y} = \frac{\exp(\delta + \gamma \cdot n_{x,y})}{1 + \exp(\delta + \gamma \cdot n_{x,y})}$$

Example (Propagation of an epidemic (2))

Question

Given $(\alpha, \beta, \gamma, \delta)$, what is the speed of propagation of this epidemic? the average duration? the number of infected persons?

Monte Carlo integration

Law of large numbers

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

$$\hat{\mathfrak{I}}_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i) \longrightarrow \mathfrak{I}$$

Central Limit Theorem

Evaluation of the error b

$$\hat{\sigma}_n^2 = \frac{1}{n^2} \sum_{i=1}^{n} (h(X_i) - \hat{\mathfrak{I}})^2$$

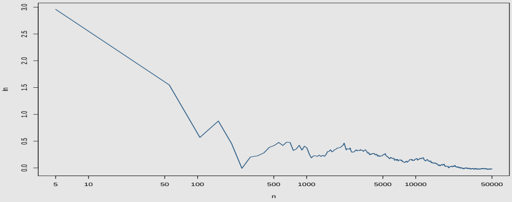
and

$$\hat{\mathfrak{I}}_n \approx \mathscr{N}(\mathfrak{I}, \hat{\sigma}_n^2)$$

Example (Normal)

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

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



Example (Cauchy / Normal)

Consider the joint model

$$X|\theta \sim \mathscr{N}(\theta, 1), \quad \theta \sim \mathcal{C}(0, 1)$$

Once X is observed, θ is estimated by

$$\delta^\pi(x) = \frac{\displaystyle\int_{-\infty}^{\infty} \frac{\theta}{1+\theta^2} e^{-(x-\theta)^2/2} d\theta}{\displaystyle\int_{-\infty}^{\infty} \frac{1}{1+\theta^2} e^{-(x-\theta)^2/2} d\theta}$$

Example (Cauchy / Normal (2))

This representation of δ^π suggests using iid variables

$$\theta_1, \cdots, \theta_m \sim \mathscr{N}(x, 1)$$

and to compute

$$\hat{\delta}^\pi_m(x) = \frac{\sum_{i=1}^m \dfrac{\theta_i}{1+\theta_i^2}}{\sum_{i=1}^m \dfrac{1}{1+\theta_i^2}} \,.$$

By vurtue of the **Law of Large Numbers**,

$$\hat{\delta}^\pi_m(x) \longrightarrow \delta^\pi(x) \quad \text{quand } m \longrightarrow \infty.$$

Example (Normal cdf)

Approximation of the normal cdf

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

by

$$\hat{\Phi}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i \le t},$$

based on a sample of size n $(X_1, \ldots, X_n)$, generated by the algorithm of Box-Muller.

Example (Normal cdf(2))

- Variance

$$\Phi(t)(1-\Phi(t))/n,$$

  since the variables $\mathbb{I}_{X_i \leq t}$ are iid Bernoulli$(\Phi(t))$.
- For t close to $t=0$ thea variance is about $1/4n$: a precision of four decimals requires on average

$$\sqrt{n} = \sqrt{2}\; 10^4$$

  simulations, thus, 200 **millions of iterations.**
- Larger [absolute] precision in the tails

Example (Normal cdf(3))

n	0.0	0.67	0.84	1.28	1.65	2.32	2.58	3.09	3.72
10^2	0.485	0.74	0.77	0.9	0.945	0.985	0.995	1	1
10^3	0.4925	0.7455	0.801	0.902	0.9425	0.9885	0.9955	0.9985	1
10^4	0.4962	0.7425	0.7941	0.9	0.9498	0.9896	0.995	0.999	0.9999
10^5	0.4995	0.7489	0.7993	0.9003	0.9498	0.9898	0.995	0.9989	0.9999
10^6	0.5001	0.7497	0.8	0.9002	0.9502	0.99	0.995	0.999	0.9999
10^7	0.5002	0.7499	0.8	0.9001	0.9501	0.99	0.995	0.999	0.9999
10^8	0.5	0.75	0.8	0.9	0.95	0.99	0.995	0.999	0.9999

Evaluation of normal quantiles by Monte Carlo based on n normal generations

Importance functions

Alternative representation :

$$\mathfrak{I} = \int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx$$

Thus, if $Y_1, \ldots, Y_n$ simuated from g,

$$\tilde{\mathfrak{I}}_n = \frac{1}{n}\sum_{i=1}^{n} h(Y_i)\frac{f(Y_i)}{g(Y_i)} \longrightarrow \mathfrak{I}$$

Appeal

- Works for all g's such that

$$\text{supp}(g) \supset \text{supp}(f)$$

- Possible improvement of the variance
- Recycling of simulations $Y_i \sim g$ for other densities f
- Usage of simple distributions g

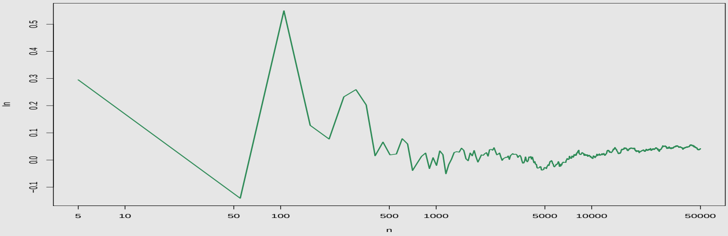
Example (Normal)

For the normal distribution and the approximation of $\mathbb{E}[X^4]$,

$$\int_{-\infty}^{\infty} x^4 e^{-x^2/2} dx \overset{[y=x^2]}{=} 2\int_0^{\infty} y^{3/2} \frac{1}{2} e^{-y/2} dy$$

suggests using $g(y) = \exp(-y/2)/2$

n	5	50	500	5000	50000
$\tilde{\mathfrak{I}}_n$	3.29	2.89	3.032	2.97	3.041



Choice of the importance function

The "best" g function depends on the density f *and* on the h function

Theorem (Optimal importance)

The choice of g that minimises the variance of $\tilde{\mathfrak{I}}_n$ is

$$g^\star(x) = \frac{|h(x)|f(x)}{\mathfrak{I}}$$

Remarks

- Finite variance only if

$$\mathbb{E}_f\left[h^2(X)\frac{f(X)}{g(X)}\right] = \int_{\mathcal{X}} h^2(x)\ \frac{f(X)}{g(X)}\, dx < \infty\ .$$

- **Null** variance for $g^\star$ if h s positive (!!)
- $g^\star$ depends on the very $\mathfrak{I}$ we are trying to estimate (??)
- Replacement of $\tilde{\mathfrak{I}}_n$ by the harmonic mean

$$\check{\mathfrak{I}}_n = \frac{\sum_{i=1}^n h(y_i)/|h(y_i)|}{\sum_{i=1}^n 1/|h(y_i)|}$$

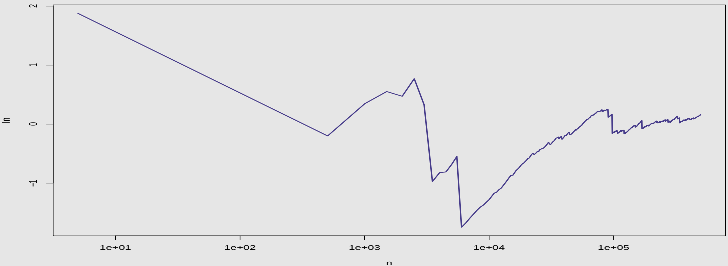
(numerator and denominator are convergent)
often **poor** (infinite variance)

Example (Normal)

For the normal distribution and the approximation of $\mathbb{E}[X^4]$, $g^\star(x) \propto x^4 \exp(-x^2/2)$, distribution of the squared root of a $\mathscr{G}a(5/2, 1/2)$ rv

[Exercise]

n	5	50	500	5,000	50,000	500,000
$\check{\mathfrak{I}}_n$	4.877	2.566	2.776	2.317	2.897	3.160



Example (Student's t)

$X \sim \mathcal{T}(\nu, \theta, \sigma^2)$, with density

$$f(x) = \frac{\Gamma((\nu+1)/2)}{\sigma\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{(x-\theta)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2} .$$

Take $\theta = 0$, $\sigma = 1$ and

$$\mathfrak{I} = \int_{2.1}^{\infty} x^5 f(x) dx$$

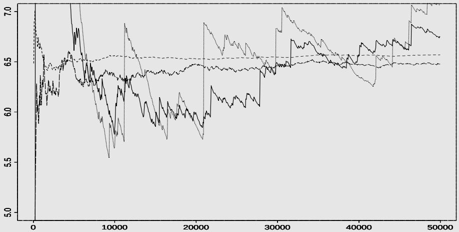
is the integral of interest

Example (Student's t (2))

- Choice of importance functions
  - f, since $f = \frac{\mathscr{N}(0,1)}{\sqrt{\chi^2_\nu/\nu}}$
  - Cauchy $\mathcal{C}(0,1)$
  - Normal $\mathcal{N}(0,1)$
  - $\mathscr{U}([0, 1/2.1])$

Results:

- Uniform optimal
- Cauchy OK
- f and Normal poor



Correlated simulations

Negative correlation...

Two samples $(X_1, \ldots, X_m)$ and $(Y_1, \ldots, Y_m)$ distributed from f in order to estimate

$$\mathfrak{I} = \int_{\mathbb{R}} h(x) f(x) dx \,.$$

Both

$$\hat{\mathfrak{I}}_1 = \frac{1}{m} \sum_{i=1}^{m} h(X_i) \quad \text{et} \quad \hat{\mathfrak{I}}_2 = \frac{1}{m} \sum_{i=1}^{m} h(Y_i)$$

have mean $\mathfrak{I}$ and variance σ^2

Correlated simulations (2)

...reduices the variance

The variance of the average is

$$\mathrm{var}\left(\frac{\hat{\mathfrak{I}}_1+\hat{\mathfrak{I}}_2}{2}\right)=\frac{\sigma^2}{2}+\frac{1}{2}\mathrm{cov}(\hat{\mathfrak{I}}_1,\hat{\mathfrak{I}}_2).$$

Therefore, if both samples are **negatively correlated**,

$$\mathrm{cov}(\hat{\mathfrak{I}}_1,\hat{\mathfrak{I}}_2)\leq 0\,,$$

they do better than two independent samples with the same size

Antithetic variables

Construction of negatively correlated variables

1. If f symmetric about μ, take $Y_i = 2\mu - X_i$
2. If $X_i = F^{-1}(U_i)$, take $Y_i = F^{-1}(1 - U_i)$
3. If $(A_i)_i$ is a partition of $\mathcal{X}$, partitionned sampling takes X_j's in each A_i (requires the knowledge of $\Pr(A_i)$)

Control variates

Take

$$\mathfrak{I} = \int h(x) f(x) dx$$

to be computer and

$$\mathfrak{I}_0 = \int h_0(x) f(x) dx$$

already known

We nonetheless estimate $\mathfrak{I}_0$ by $\hat{\mathfrak{I}}_0$ (and $\mathfrak{I}$ by $\hat{\mathfrak{I}}$)

Control variates (2)

Combined estimator

$$\hat{\mathfrak{I}}^* = \hat{\mathfrak{I}} + \beta(\hat{\mathfrak{I}}_0 - I_0)$$

$\hat{\mathfrak{I}}^*$ is unbiased for $\mathfrak{I}$ et

$$\text{var}(\hat{\mathfrak{I}}^*) = \text{var}(\hat{\mathfrak{I}}) + \beta^2 \text{var}(\hat{\mathfrak{I}}) + 2\beta \mathrm{cov}(\hat{\mathfrak{I}}, \hat{\mathfrak{I}}_0)$$

Control variates (3)

Optimal choice of β

$$\beta^\star = -\frac{\mathrm{cov}(\hat{\mathfrak{I}}, \hat{\mathfrak{I}}_0)}{\mathsf{var}(\hat{\mathfrak{I}}_0)} \; ,$$

with

$$\mathsf{var}(\hat{\mathfrak{I}}^\star) = (1 - \rho^2)\, \mathsf{var}(\hat{\mathfrak{I}}) \; ,$$

where ρ correlation between $\hat{\mathfrak{I}}$ and $\hat{\mathfrak{I}}_0$

Example (Approximation of quantiles)

Consider the evaluation of

$$\varrho = \mathsf{Pr}(X > a) = \int_a^\infty f(x)dx$$

by

$$\hat{\varrho} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i > a), \qquad X_i \overset{\text{iid}}{\sim} f$$

with $\mathsf{Pr}(X > \mu) = \frac{1}{2}$

Example (Approximation of quantiles (2))

The control variate

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i > a) + \beta\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i > \mu) - \mathsf{Pr}(X > \mu)\right)$$

improves upon $\hat{\varrho}$ if

$$\beta < 0 \quad \text{et} \quad |\beta| < 2\frac{\mathrm{cov}(\delta_1,\delta_3)}{\mathsf{var}(\delta_3)} = 2\frac{\mathsf{Pr}(X > a)}{\mathsf{Pr}(X > \mu)}.$$

Integration by conditioning

Take advantage of the inequality

$$\mathbf{var}(\mathbb{E}[\delta(\mathbf{X})|\mathbf{Y}]) \leq \mathbf{var}(\delta(\mathbf{X}))$$

also called **Rao-Blackwell Theorem**

Consequence :

If $\hat{\mathfrak{I}}$ is an unbiased estimator of $\mathfrak{I} = \mathbb{E}_f[h(X)]$, with X simulated from the joint density $\tilde{f}(x, y)$, where

$$\int \tilde{f}(x, y) dy = f(x),$$

the estimator

$$\hat{\mathfrak{I}}^* = \mathbb{E}_{\tilde{f}}[\hat{\mathfrak{I}}|Y_1, \ldots, Y_n]$$

dominates $\hat{\mathfrak{I}}(X_1, \ldots, X_n)$ in terms of variance (and is also unbiased)

Example (Mean of a Student's t)

Consider

$$\mathbb{E}[h(x)] = \mathbb{E}[\exp(-x^2)] \quad \text{avec} \quad X \sim \mathscr{T}(\nu, 0, \sigma^2)$$

Student's t distribution can be simulated as

$$X|y \sim \mathcal{N}(\mu, \sigma^2 y) \qquad \text{and} \qquad Y^{-1} \sim \chi^2_\nu.$$

Example (Mean of a Student's t (2))

The empirical average

$$\frac{1}{m}\sum_{j=1}^{m} \exp(-X_j^2)\ ,$$

can be improved based on the joint sample

$$((X_1, Y_1), \ldots, (X_m, Y_m))$$

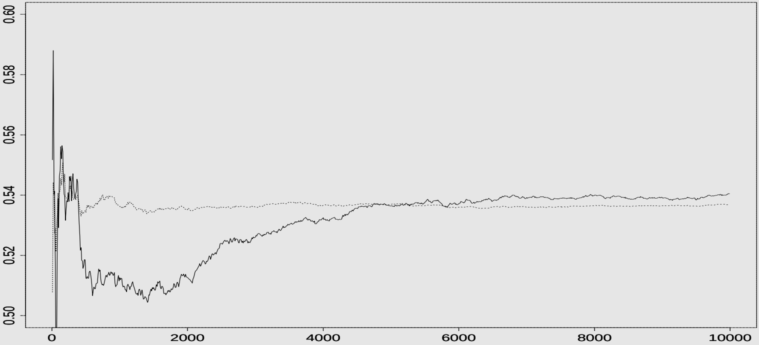
since

$$\frac{1}{m}\ \sum_{j=1}^{m}\ \mathbb{E}[\exp(-X^2)|Y_j] = \frac{1}{m}\ \sum_{j=1}^{m}\ \frac{1}{\sqrt{2\sigma^2 Y_j + 1}}$$

is the conditional expectation

Example (Mean of a Student's t (3))

In this special case, the precision is **ten times** higher



Estimators of $\mathbb{E}[\exp(-X^2)]$: empirical average (full lines) versus conditional expectation (dotted line) for $(\nu, \mu, \sigma) = (4.6, 0, 1)$.

Chapter 3 : The Bootstrap Method

- Introduction
- Glivenko-Cantelli's Theorem
- Bootstrap
- Parametric Bootstrap

Intrinsic randomness

Estimation from a random sample means uncertainty

Since based on a **random** sample, an estimator

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

also is a **random** variable

Average variation

Question 1 :

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

Question 2 :

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

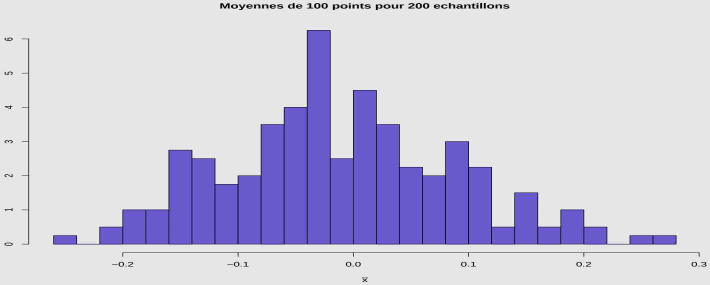
Question 3 :

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 θ is estimated by

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



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

Associated problems

- Observation of a **single** sample in most cases
- The sampling distribution 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

$$\mathrm{H}_0 \;:\; \theta \leq 0$$

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

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

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

Estimation of the repartition function

Extension/application of the LLN to the approximation of the cdf:
For a sample $X_1, \ldots, X_n$, if

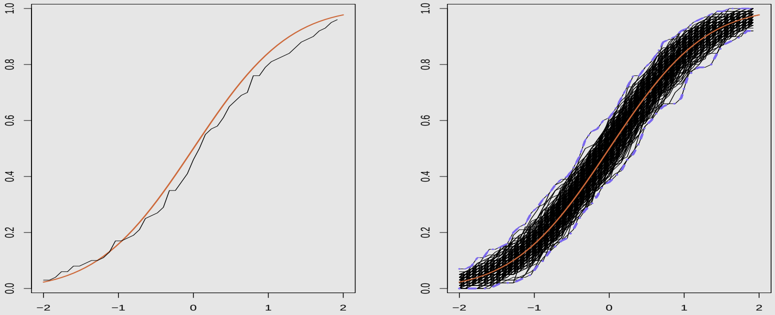
$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{]-\infty, X_i]}(x) = \frac{\text{card}\,\{X_i;\; X_i \leq x\}}{n},$$

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

[Glivenko–Cantelli]

$$\boxed{\hat{F}_n(x) \longrightarrow \Pr(X \leq x)}$$

Example (**Normal sample)**



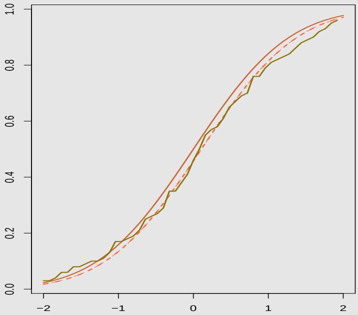
Estimation of the cdf F from a normal sample of 100 points and variation of this estimation over 200 normal samples

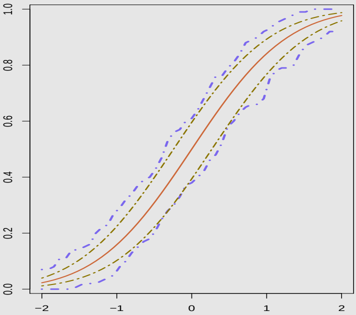
Properties

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

Example (**Normal sample)**

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





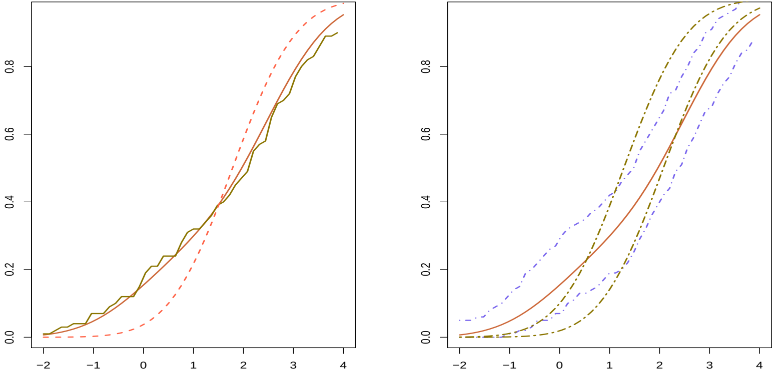
Estimation of $\Phi(\cdot-\theta)$ by $\hat{F}_n$ and by $\Phi(\cdot-\hat{\theta})$ based on 100 points and maximal variation of thoses estimations over 200 replications

Example (**Non-normal sample**)

Sample issued from

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

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



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

Extension to functionals of F

For any quantity of the form

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

[Functional of the cdf]

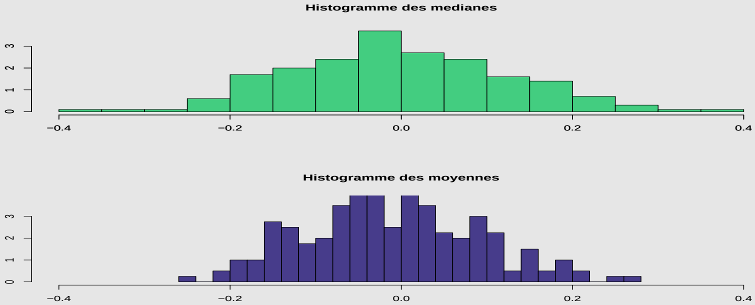
use of the approximation

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

[Moment estimator]

Example (**Normal sample)**

Since θ also is the median of $\mathcal{N}(\theta, 1)$, $\hat{\theta}$ can be chosen as the median of $\hat{F}_n$, **equal to** 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

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

Principle

Since

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

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

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

Implementation

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

The distribution corresponding to

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

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

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

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

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

Monte Carlo implementation

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

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

2. Use the sample

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

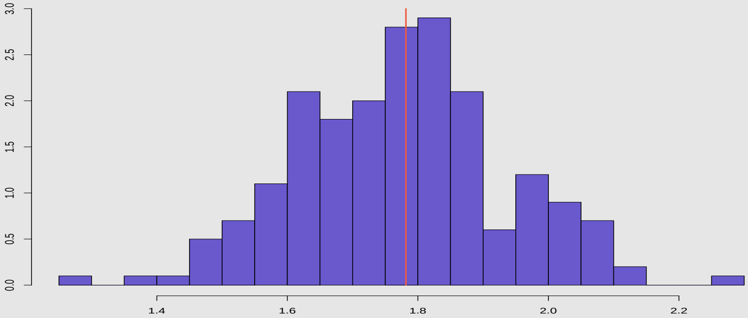
   to approximate the distribution of

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

Notes

- `bootstrap`
  the sample itself is used to build an evaluation of its distribution
  [Adventures of the Munchausen Baron]
- a bootstrap sample is obtained via n samplins with replacement in $(X_1, \ldots, X_n)$
- this sample can then take n^n values (or $\binom{2n-1}{n}$ values if the order does not matter)

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



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

Example (**Derivation of the average variation)**

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

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

and its bootstrap approximation is

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

Example (**Derivation of the average variation (2)**)

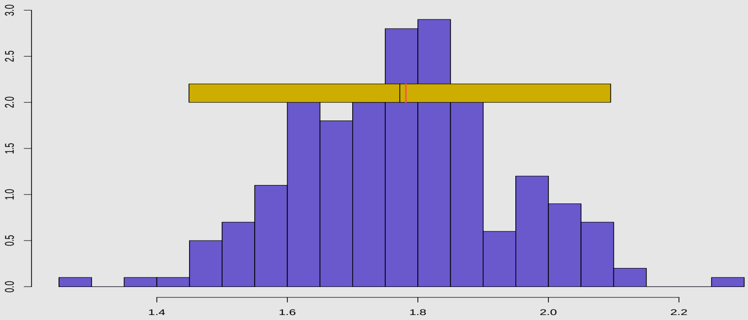
Approximation itself approximated by

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

where

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

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



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

Example (**Normal sample)**

Sample

$$(X_1, \ldots, X_{100}) \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$$

Comparison of the confidence intervals

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

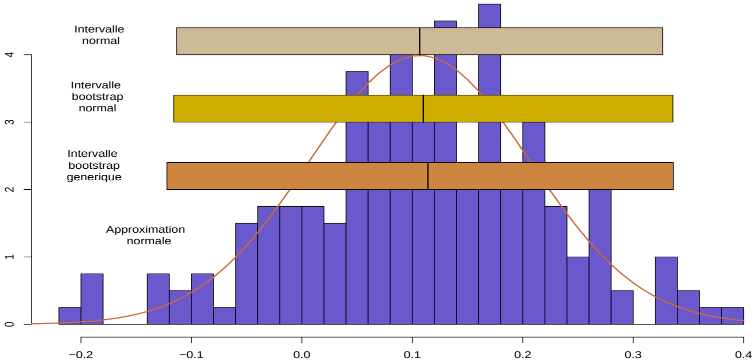
[normal approximation]

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

[normal bootstrap approximation]

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

[generic bootstrap approximation]



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

Parametric Bootstrap

If the parametric shape of F is known,

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

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

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

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

[Cf Example 46]

Parametric Bootstrap

Approximation of the distribution of

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

by the distribution of

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

May avoid simulation approximations in some cases

Example (**Exponential Sample**)

Take

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{E}\mathsf{xp}(\lambda)$$

and $\lambda = 1/\mathsf{E}_\lambda[X]$ 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

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

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

Questions :

- What is the bias

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

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

Bootstrap evaluation of the bias

Example (Exponential Sample (3))

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

[parametric version]

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

[non-parametric version]

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

In the first (parametric) version,

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

and

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

therefore the bias is **analytically** evaluated as

$$-\lambda/n - 1$$

and estimated by

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

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

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

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

which achieves the **"wrong"** sign

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

Construction of a confidence interval on λ

By parametric bootstrap,

$$\mathsf{Pr}_\lambda\left(\hat\lambda_1 \le \lambda \le \hat\lambda_2\right) = \mathsf{Pr}\left(\omega_1 \le \lambda/\hat\lambda \le \omega_2\right) = 0.95$$

can be deduced from

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

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

$$[\hat\lambda_1, \hat\lambda_2] = [0.452, 0.580]$$

Example (**Exponential Sample (7)**)

In nonarametric bootstrap, one replaces

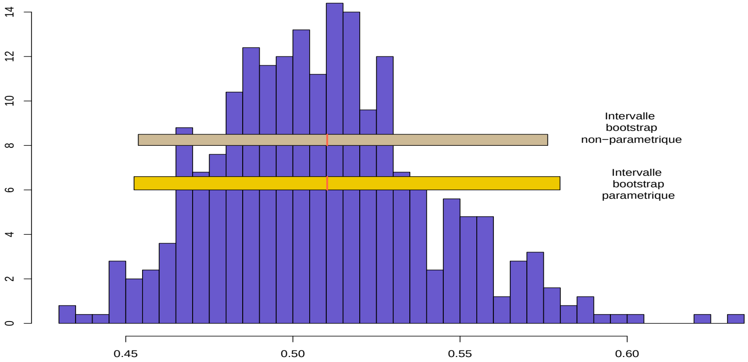
$$\mathsf{Pr}_F\left(q(.025) \le \lambda(F) \le q(.975)\right) = 0.95$$

with

$$\mathsf{Pr}_{\hat{F}_n}\left(q^*(.025) \le \lambda(\hat{F}_n) \le q^*(.975)\right) = 0.95$$

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

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


14
12
10
8
6
4
2
0
0.45
0.50
0.55
0.60
Intervalle bootstrap non−parametrique
Intervalle bootstrap parametrique

Example (**Student Sample)**

Take

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathfrak{T}(5, \mu, \tau^2) \overset{\text{def}}{=} \mu + \tau \, \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_5/5}}$$

μ and τ could be estimated by

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

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

Problem

$\hat{\mu}_n$ is not distributed from a Student $\mathfrak{T}(5, \mu, \tau^2/n)$ distribution
The distribution of $\hat{\mu}_n$ ccan be reproduced by bootstrap sampling

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

Comparison of confidence intervals

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

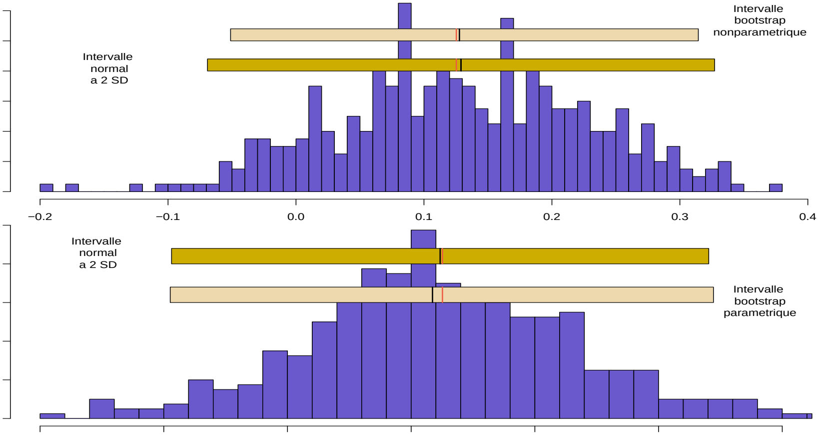
[normal approximation]

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

[parametric boostrap approximation]

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

[non parametric boostrap approximation]



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

Chapter 4 : Rudiments of Nonparametric Statistics

- Introduction
- Density Estimation
- Nonparametric tests

Probleme :

How could one conduct a statistical inference when the distribution of the data $X_1, \ldots, X_n$ is unknown?

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$$

with F **unknown**

Nonparametric setting in opposition to the **parametric** case when $F(\cdot) = G_\theta(\cdot)$ with only θ unknown

Nonparametric Statistical Inference

- Estimation of a quantity that depends on F

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

- Decision on an hypothesis about F

$$F \in \mathcal{F}_0\,? \qquad F == F_0\,? \qquad \theta(F) \in \Theta_0\,?$$

- Estimation of functionals of F

$$F \qquad f(x) = \frac{dF}{dx}(x) \qquad \mathsf{E}_F[h(X_1)|X_2 = x]$$

Density Estimation

To estimate

$$f(x) = \frac{dF}{dx}(x)$$

[density of X]

a natural solution is

$$\hat{f}_n(x) = \frac{d\hat{F}_n}{dx}(x)$$

but

$\hat{F}_n$ **cannot be differentiated!**

Histogram Estimation

A first solution is to reproduce the stepwise constant structure of $\hat{F}_n$ pour f

$$\hat{f}_n(x) = \sum_{i=1}^{k} \omega_i \mathbb{I}_{[a_i, a_{i+1}[}(x) \qquad a_1 < \ldots < a_{k+1}$$

by picking the ω_i's such that

$$\sum_{i=1}^{k} \omega_i (a_{i+1} - a_i) = 1 \quad \text{et} \quad \omega_i (a_{i+1} - a_i) = \widehat{P_F}(X \in [a_i, a_{i+1}[)$$

Histogram Estimation (cont'd)

For instance,

$$\begin{aligned}\omega_i(a_{i+1} - a_i) &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{[a_i, a_{i+1}[}(X_i) \\ &= \hat{F}_n(a_{i+1}) - \hat{F}_n(a_i)\end{aligned}$$

[bootstrap]

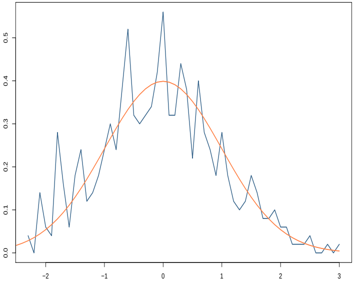
is a converging estimator of $P_F(X \in [a_i, a_{i+1}[)$

[Warning: side effects!]

hist(x)$density

With **R**, hist(x)$density provides the values of ω_i and hist(x)$breaks the values of the a_i's

It is better to use the values produced by hist(x)$density to build up a stepwise linear function by plot(hist(x)$density) rather than to use a stepwise constant function.



Histogram estimator for $k = 45$ and 450 normal observations

Probabilist Interpretation

Starting with stepwise constant functions, the resulting approximation of the distribution is a weighted sum of uniforms

$$\sum_{i=1}^{k} \pi_i \mathcal{U}([a_i, a_{i+1}])$$

Equivalent to a stepwise linear approximation of the cdf

$$\tilde{F}_n(x) = \sum_{i=1}^{n} \pi_i \frac{x - a_i}{a_{i+1} - a_i} \mathbb{I}_{[a_i, a_{i+1}[}(x)$$

Drawbacks

- Depends on the choice of the partition $(a_i)_i$, often based on the data itself (see **R**)
- Problem of the endpoints a_1 and a_{k+1} : while not infinite (why?), they still must approximate the support of f
- k and $(a_i)_i$ must depend on n to allow for the convergence of $\hat{f}_n$ toward f
- **but...** $a_{i+1} - a_i$ must not decrease too fast to 0 to allow for the convergence of π_i: there must be enough observations in each interval $[a_i, a_{i+1}]$

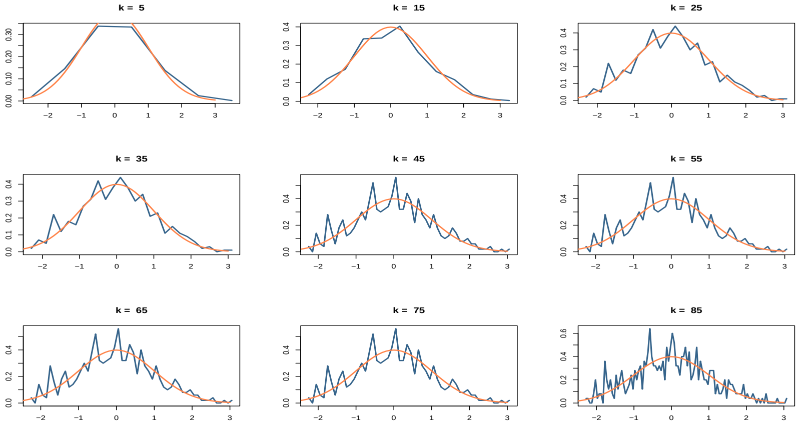
Scott bandwidth

"Optimal" selection of the width of the classes :

$$h_n = 3.5\,\hat{\sigma}\,n^{-1/3} \quad \text{et} \quad h_n = 2.15\,\hat{\sigma}\,n^{-1/5}$$

provide the right width $a_{i+1} - a_i$ (nclass = range(x) / h) for a stepwise constant $\hat{f}_n$ and a stepwise linear f_n, respectively. (In the sense that they ensure the convergence of $\hat{f}_n$ toward f when n goes to ∞.)

[nclass=9 and nclass=12 in the next example]



Variation of the histogram estimators as a function of k for a normal sample with 450 observations

Kernel Estimator

Starting with the definition

$$f(x) = \frac{dF}{dx}(x)\,,$$

we can also use the approximation

$$\begin{aligned}\hat{f}(x) &= \frac{\hat{F}_n(x+\delta) - \hat{F}_n(x-\delta)}{2\delta} \\ &= \frac{1}{2\delta n}\sum_{i=1}^{n}\{\mathbb{I}_{X_i<x+\delta} - \mathbb{I}_{X_i<x-\delta}\} \\ &= \frac{1}{2\delta n}\sum_{i=1}^{n}\mathbb{I}_{[-\delta,\delta]}(x-X_i)\end{aligned}$$

when δ is small enough.

[Positive point : $\hat{f}$ is a density]

Analytical and probabilistic interpretation

With this approximation

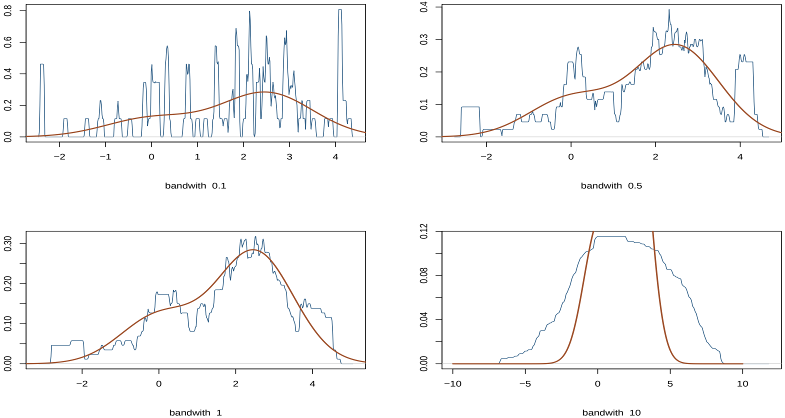
$$\hat{f}_n(x) = \frac{\#\,\text{observations close to } x}{2\delta n}$$

Particular case of an histogram estimator where the a_i's are like $X_j \pm \delta$

Representation of $\hat{f}_n$ as a weighted sum of uniforms

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{U}([X_i - \delta, X_i + \delta])$$

[Note connection with bootstrap]



Variation of uniform kernel estimators as a function of δ for a non-normal sample of 200 observations

Extension

Instead of a uniform approximation around each X_i, we can use a smoother distribution:

$$\hat{f}(x) = \frac{1}{\delta n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{\delta}\right)$$

where K is a probability density (**kernel**) and δ a scale factor that is small enough.

With **R**, density(x)

Kernel selection

All densities are a priori acceptable. In practice (and with **R**, usage of

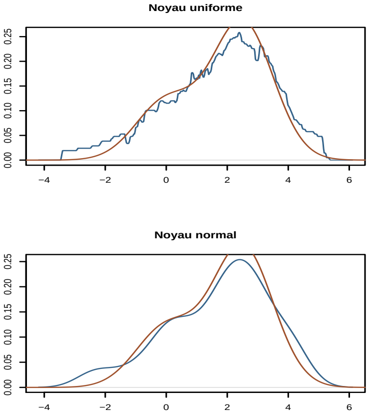
- the normal/Gaussian kernel [kernel="gaussian" or "g"]
- the Epanechnikov's kernel [kernel="epanechnikov" or "e"]

$$K(y) = C\,\{1 - y^2\}^2\,\mathbb{I}_{[-1,1]}(y)$$

- the triangular kernel [kernel="triangular" or "t"]

$$K(y) = (1 + y)\mathbb{I}_{[-1,0]}(y) + (1 - y)\mathbb{I}_{[0,1]}(y)$$

Conclusion : Very little influence on the estimation of f (except for the uniform kernel [kernel="rectangular" or "r"]).

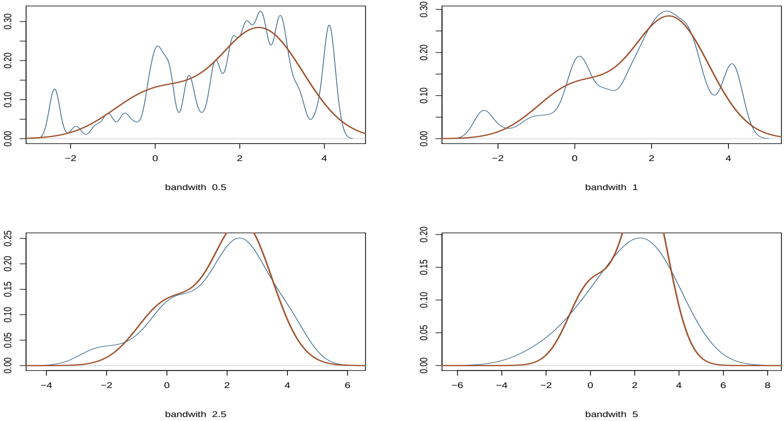


Variation of the kernel estimates with the kernel for a non-normal sample of 200 observations

Convergence to f

The choice of the **bandwidth** δ is crucial!

- If δ large, many X_i contribute to the estimation of $f(x)$
  [Over-smoothing]
- If δ small, few X_i contribuent to the estimation of $f(x)$
  [Under-smoothing]



Variation of $\hat{f}_n$ as a function of δ for a non-normal sample of 200 observations

Optimal bandwidth

When considering the averaged integrated error

$$d(f, \hat{f}_n) = \mathbb{E}\left[\int \{f(x) - \hat{f}_n(x)\}^2 \, \mathrm{d}x\right] ,$$

there exists an optimal choice for the bandwidth δ, denoted h_n to indicate its dependance on n.

Optimal bandwidth (cont'd)

Using the decomposition

$$\int \left\{ f(x) - \mathbb{E}\left[\hat{f}(x)\right] \right\}^2 \mathrm{d}x + \int \mathrm{var}\{\hat{f}(x)\} \,\mathrm{d}x \,,$$

[Bias^2+variance]

and the approximations

$$f(x) - \mathbb{E}\left[\tilde{f}(x)\right] \simeq \frac{f''(x)}{2}\, h_n^2$$

$$\mathbb{E}\left[\frac{\exp\{-(X_i - x)^2/2h_n^2\}}{\sqrt{2\pi} h_n} \right] \simeq f(x) \,,$$

[Exercise]

Optimal bandwidth (cont'd)

we deduce that the bias is of order

$$\int \left\{ \frac{f''(x)}{2} \right\}^2 \mathrm{d}x \, h_n^4$$

and that the variance is approximately

$$\frac{1}{nh_n\sqrt{2\pi}} \int f(x) \, \mathrm{d}x = \frac{1}{nh_n\sqrt{2\pi}}$$

[Exercise]

Optimal bandwidth (end'd)

Therefore, the error goes to 0 when n goes to ∞ if

1. h_n goes to 0 *and*
2. nh_n goes to infinity.

The optimal bandwidth is given by

$$\hat{h}_n^\star = \left(\sqrt{2\pi} \int \left\{f''(x)\right\}^2 \, \mathrm{d}x \, n\right)^{-1/5}$$

Empirical bandwidth

Since the optimal bandwidth depends on f, unknown, we can use an approximation like

$$\hat{h}_n = \frac{0.9 \min(\hat{\sigma}, \hat{q}_{75} - \hat{q}_{25})}{(1.34n)^{1/5}},$$

where $\hat{\sigma}$ is the empirical standard deviation and $\hat{q}_{25}$ and $\hat{q}_{75}$ are the estimated 25% and 75% quantiles of X.

Note : The values 0.9 and 1.34 are chose for the normal case.

Warning! This is not the defect bandwidth in **R**

The perspective of statistical tests

Given a question about F, such as

Is F equal to F_0, a known distribution ?

the statistical answer is based on the data

$$X_1, \ldots, X_n \sim F$$

to decide whether **yes or no** the question **[the hypothesis]** is compatible with this data

The perspective of statistical tests (cont'd)

A **test procedure** (or statistical test) $\varphi(x_1, \dots, x_n)$ is taking values in $\{0, 1\}$ (for yes/no)

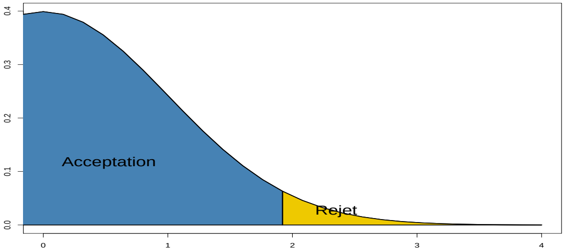
When deciding about the question on F, there are two types of errors:

1. refuse the hypothesis erroneously (Type I)
2. accept the hypothesis erroneously (Type II)

Both types of errors must then be balanced

The perspective of statistical tests (cont'd)

In pratice, a choice is made to concentrate upon type I errors and to reject the hypothesis only when the data is **significantly** incompatibles with this hypothesis.



To accept an hypothesis after a test only means that the data has not rejected this hypothesis !!!

Comparison of distributions

Example (Two equal distributions?)

Given two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, with respective distributions F and G, both unknown
What is the answer to the question

$$F == G \ ?$$

Comparison of distributions (contd)

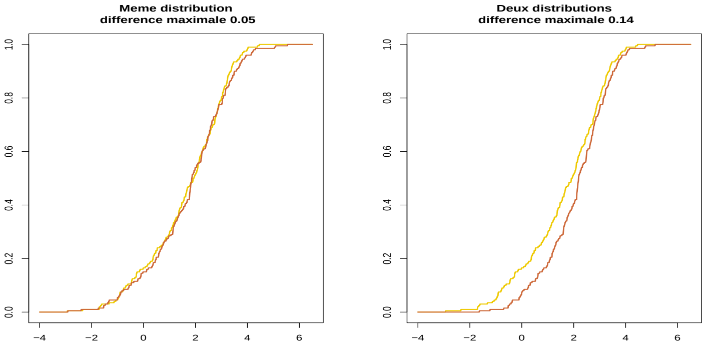
Example (Two equal distributions?)

Idea :

Compare the estimates of F and of G,

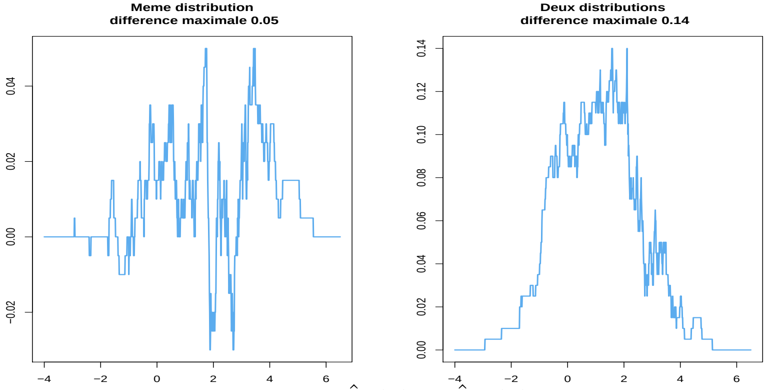
$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i \le x} \qquad \text{et} \qquad \hat{G}_m(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}_{Y_i \le x}$$

The Kolmogorov–Smirnov Statistics



Evaluation via the difference

$$K(m,n) = \max_x \left| \hat{F}_n(x) - \hat{G}_m(x) \right| = \max_{X_i, Y_j} \left| \hat{F}_n(x) - \hat{G}_m(x) \right|$$



Evolution of the difference $\hat{F}_n(x) - \hat{G}_m(x)$ in two cases

The Kolmogorov–Smirnov Statistics (2)

Usage :

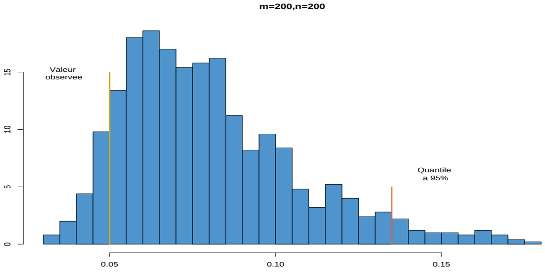
If $K(m,n)$ "large", the distributions F and G are significatively different.
If $K(m,n)$ "small", they cannot be distinguished on the data $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_m$, therefore $F=G$ is acceptable

[Kolmogorov–Smirnov test]

With **R**, ks.test(x,y)

Calibration of the test

For m and n fixed, **if** $F = G$, $K(m,n)$ has a fixed distribution for all F's.
It is thus always possible to reduce the problem to the comparison of two uniform samples and to use simulation to approximate the distribution of $K(m,n)$ and of its quantiles.



Calibration of the test (cont'd)

If the observed $K(m,n)$ is above the 90 or 95 % quantile of $K(m,n)$ under H_0 the value is very unlikely

if $F = G$

and the hypothesis of equality of both distributions is rejected.

Calibration of the test (cont'd)

Example of R output:

Two-sample Kolmogorov-Smirnov test
data: z[, 1] and z[, 2]
D = 0.05, p-value = 0.964
alternative hypothesis: two.sided

p-value = 0.964 means that the probability that $K(m,n)$ is larger than the observed value D = 0.05 is 0.964, thus that the observed value is small under the distribution of $K(m,n)$: **we thus accept the equality hypothesis.**

Test of independance

Example (Independence)

Testing for independance between two rs's X and Y based on the observation of the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$
Question

$$X \perp Y \; ?$$

Rank test

Idea:

If the X_i's are ordered

$$X_{(1)} \leq \dots X_{(n)}$$

the ranks R_i (orders after the ranking of the X_i's) of the corresponding Y_i's

$$Y_{[1]}, \dots, Y_{[n]},$$

must be completely random.

In **R**, rank(y[order(x)])

Rank test (cont'd)

Rank: The vector

$$\mathfrak{R} = (R_1, \ldots, R_n)$$

is called **the rank statistic** of the sample $(Y_{[1]}, \ldots Y_{[n]})$
Spearman's statistic is

$$S_n = \sum_{i=1}^{n} i\, R_i$$

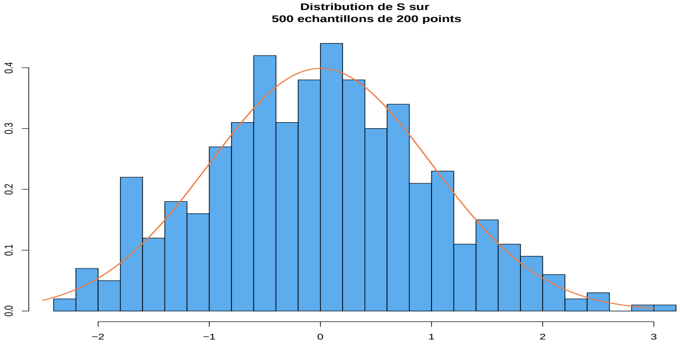
[Correlation between i and R_i]

It is possible to prove that, **if** $X \perp Y$,

$$\mathsf{E}[S_n] = \frac{n(n+1)^2}{4} \qquad \mathrm{var}(S_n) = \frac{n^2(n+1)^2(n-1)}{144}$$

Spearman's statistic

The distribution of S_n is available via [uniform] simulation or via normal approximation



Recentred version of Spearman's statistics and normal approximation

Spearman's statistic (cont'd)

It is therefore possible to find the 5% and 95% quantiles of S_n through simulation and to decide if the observed value of S_n is in-between those quantiles ($=$ Accept independance) or outside ($=$ Reject independance)

Multinomial tests

Example (Chi-square test)

An histogram representation brings a robustified answer to testing problems, like

Is the sample $X_1, \ldots, X_n$ normal $\mathcal{N}(0,1)$?

Idea:

Replace the original problem by its discretised version on intervals $[a_i, a_{i+1}]$

Is it true that

$$P(X_i \in [a_i, a_{i+1}]) = \int_{a_i}^{a_{i+1}} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx \stackrel{\text{def}}{=} p_i \ ?$$

Principle

Multinomial modelling

The problem is always expressed through a multinomial distribution

$$\mathcal{M}_k\left(p_1^0, \ldots, p_k^0\right)$$

or a family of multinomial distributions

$$\mathcal{M}_k\left(p_1(\theta), \ldots, p_k(\theta)\right) \qquad \theta \in \Theta$$

Examples

- For testing the adequation to a standard normal distribution, $\mathcal{N}(0,1)$, k is determined by the number of intervals $[a_i, a_{i+1}]$ and the p_i^0's by

$$p_i^0 = \int_{a_i}^{a_{i+1}} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx$$

- For testing the adequation to a normal distribution, $\mathcal{N}(\theta,1)$, the $p_i(\theta)$ are given by

$$p_i(\theta) = \int_{a_i}^{a_{i+1}} \frac{\exp(-(x-\theta)^2/2)}{\sqrt{2\pi}} dx$$

Examples (cont'd)

- For testing the independance between two random variables, X et Y,

$$X \perp Y \ ?$$

k is the number of cubes $[a_i, a_{i+1}] \times [b_i, b_{i+1}]$, θ is defined by

$$\theta_{1i} = P(X \in [a_i, a_{i+1}]) \quad \theta_{2i} = P(Y \in [b_i, b_{i+1}])$$

and

$$\begin{aligned} p_{i,j}(\theta) &\overset{\text{def}}{=} P(X \in [a_i, a_{i+1}], Y \in [b_i, b_{i+1}]) \\ &= \theta_{1i} \times \theta_{2j} \end{aligned}$$

Chi-square test

A natural estimator for the p_i's is

$$\hat{p}_i = \hat{P}(X \in [a_i, a_{i+1}]) = \hat{F}_n(a_{i+1}) - \hat{F}_n(a_i)$$

[See bootstrap]

The **chi-square statistic** is

$$\begin{aligned} S_n &= n \sum_{i=1}^{k} \frac{(\hat{p}_i - p_i^0)^2}{p_i^0} \\ &= \sum_{i=1}^{k} \frac{(\hat{n}_i - np_i^0)^2}{np_i^0} \end{aligned}$$

when testing the adequation to a multinomial distribution

$$\mathcal{M}_k\left(p_1^0, \ldots, p_k^0\right)$$

Chi-square test (cont'd)

and

$$\begin{aligned} S_n &= n \sum_{i=1}^{k} \frac{(\hat{p}_i - p_i(\hat{\theta}))^2}{p_i(\hat{\theta})} \\ &= \sum_{i=1}^{k} \frac{(\hat{n}_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} \end{aligned}$$

when testing the adequation to a family of multinomial distributions

$$\mathcal{M}_k\left(p_1(\theta), \ldots, p_k(\theta)\right) \qquad \theta \in \Theta$$

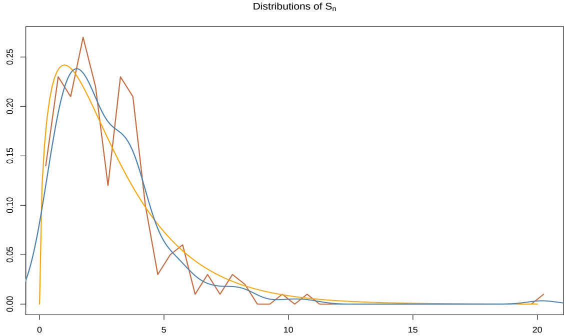
Approximated distribution

For the adequation to a multinomial distribution, the distribution of S_n is approximately (for large n's)

$$S_n \sim \chi^2_{k-1}$$

and for the adequation to a family of multinomial distributions, with $\dim(\theta) = p$,

$$S_n \sim \chi^2_{k-p-1}$$



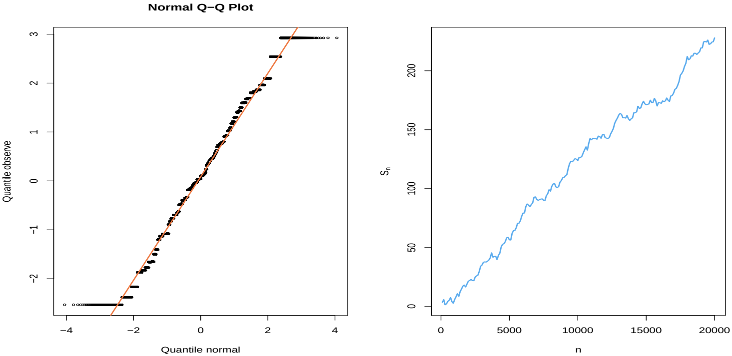
Distribution of S_n for 200 normal samples of 100 points and a test of adequation to $\mathcal{N}(0,1)$ with $k = 4$

Use and limits

The hypothesis under scrutiny is rejected if S_n is too large for a χ^2_{k-1} or χ^2_{k-p-1} distribution

[In **R**, pchisq(S)]

Convergence (in n) to a χ^2_{k-1} (or χ^2_{k-p-1}) distribution is only established for fixed k and (a_i). In pratice, k and (a_i) are determined by the observations, which reduces the validity of the approximation.



QQ-plot of a non-normal sample and evolution of S_n as a function of n for this sample