FACULTÉ DES SCIENCES ET TECHNOLOGIES | Université de Lille

Variance Reduction Methods

Challenges in the Context of Rare Events

Reda Mdair

2024

Contents

1 Introduction

Monte Carlo simulations represent an essential approach for estimating complex probabilistic quantities, particularly when analytical expressions are inaccessible. Their conceptual simplicity and guaranteed convergence make them powerful and versatile methods. However, their practical efficiency is often limited by a major drawback: the variance of the resulting estimators can be high, making the results imprecise unless the sample size is significantly increased.

This issue becomes even more critical in the context of rare events, where the probability of interest is extremely low. In such cases, a naive Monte Carlo approach can be highly inefficient, requiring billions of simulations to obtain a usable estimate. This is where variance reduction techniques come into play.

The objective of this project is to explore these techniques and demonstrate how they can drastically enhance estimation accuracy while reducing computational cost. We will specifically examine three major methods:
- **Importance Sampling (IS)**, which modifies the simulation distribution to better capture the critical regions of the rare event space.
- **Control Variates (CV)**, which leverage correlations between random variables to reduce estimator variance.
- **Stratified Sampling (SS)**, which partitions the sampling space into homogeneous strata to obtain more precise estimates by grouping simulations.

Through a rigorous theoretical approach, we will illustrate how these techniques improve the efficiency of Monte Carlo simulations, particularly for rare events. The final application will demonstrate their relevance in a real-world scenario, anchored in the domain of sports, by evaluating the probability of an exceptional score in a football match.

2 Variance Reduction Methods

Monte Carlo methods are widely used approaches for estimating rare event probabilities. The probability associated with such events can be expressed as an integral of the form:

$$I = E_f[h(X)] = \int h(x)f(x)dx \quad (*)$$

where f is the probability density function of the random variable X, and h is a function of X. This integral can be estimated using the classical Monte Carlo method, which follows the principle:

Definition (Classical Monte Carlo)

Let $(X_1, \ldots, X_n)$ be a sample of n i.i.d. random variables drawn from the distribution f. The classical Monte Carlo (MCC) estimator of $E_f[h(X)]$ is given by:

$$\delta_{MCC} = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \qquad (2.0)$$

This method is a fundamental result in probability theory and relies on the strong law of large numbers. Its advantage is that the MCC estimator is unbiased and strongly consistent.

However, a major limitation of Monte Carlo methods is that their precision depends on the variance of the estimator. This variance can be particularly high in the context of rare events.
To address this challenge, various variance reduction techniques have been developed to improve the efficiency of classical Monte Carlo simulations in estimating extreme events.

2.1 Importance Sampling

The principle of Importance Sampling (IS) consists of simulating not according to the target density f of X, but rather according to an instrumental density $\tilde{f}$. This instrumental density must be carefully chosen to improve the performance of Monte Carlo-based estimators.
A concrete application of this method will be presented in Section 5.

Recall that we aim to approximate $E_f(h(X))$, where X is a random variable with density f. For any density $\tilde{f}$, we can rewrite equation (2.0) as follows:

$$I = E_f[h(X)] = \int h(x)f(x)dx = \int \left(\frac{h(x)f(x)}{\tilde{f}(x)}\right)\tilde{f}(x)dx$$

$$= E_{\tilde{f}}\left[\frac{h(Y)f(Y)}{\tilde{f}(Y)}\right]$$

where Y is a random variable following the distribution $\tilde{f}$.

Since the Monte Carlo estimator (MCC) is given by $\delta_{MCC} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$, where $(X_1, \ldots, X_n)$ is a sample of size n drawn from f, the Importance Sampling (IS) estimator of $E_f(h(X))$ is constructed as follows:

$$\boxed{\delta_{IS} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{h(Y_i)f(Y_i)}{\tilde{f}(Y_i)}\right]} \tag{2.1}$$

where $(Y_1, \ldots, Y_n)$ is a sample of size n drawn from $\tilde{f}$.

For this estimator to be well-defined, we must ensure that the support of hf is included in that of $\tilde{f}$ to avoid division by zero. In other words, the necessary condition is: $h(y)f(y) > 0 \Rightarrow \tilde{f}(y) > 0$.

Bias of the estimator: The IS estimator is unbiased, provided that $supp(hf) \subseteq supp(\tilde{f})$.

Convergence of the estimator: By applying the strong law of large numbers to the sequence of i.i.d. random variables $(\frac{h(Y_i)f(Y_i)}{\tilde{f}(Y_i)})_{i\geq 1}$ (with finite expectation under $\tilde{f}$), we obtain:

$$\delta_{IS} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{h(Y_i)f(Y_i)}{\tilde{f}(Y_i)}\right] \xrightarrow[n\to\infty]{a.s.} E_{\tilde{f}}\left[\frac{h(Y)f(Y)}{\tilde{f}(Y)}\right] = I.$$

We now compare the variance of the Importance Sampling estimator to that of the MCC estimator. Let X and Y be random variables with respective densities f and $\tilde{f}$. We obtain the following results:

- $Var(\delta_{MCC}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} h(X_i)\right)$
  $= \frac{1}{n}Var(h(X))$
  $= \frac{1}{n}(E_f[h(X)^2] - I^2)$

- $Var(\delta_{IS}) = Var\left(\frac{1}{n}\sum_{i=1}^{n}\frac{h(Y_i)f(Y_i)}{\tilde{f}(Y_i)}\right)$
  $= \frac{1}{n}Var\left(\frac{h(Y)f(Y)}{\tilde{f}(Y)}\right)$
  $= \frac{1}{n}\left(E_{\tilde{f}}\left[\frac{h(Y)^2 f(Y)^2}{\tilde{f}(Y)^2}\right] - I^2\right)$
  $= \frac{1}{n}\left(E_f\left[\frac{h(X)^2 f(X)}{\tilde{f}(X)}\right] - I^2\right)$
  $= \frac{1}{n}\left(E_f[h(X)^2 w(X)] - I^2\right)$

where we have introduced the importance weight notation:

$$w(X) = \frac{f(X)}{\tilde{f}(X)}.$$

These two variances differ only by the importance weight. Thus, an appropriate choice of instrumental density $\tilde{f}$ should minimize the term $E_f[h(X)^2 w(X)]$.

The density that minimizes the variance has an explicit solution, given by the following lemma:

Lemma (Optimal Instrumental Density)
Let Y be a random variable with density $\tilde{f}$ satisfying $E_{\tilde{f}}[h(Y)^2 w(Y)^2] < \infty$. Then, the density that minimizes the variance is given by:

$$\tilde{f}(y) = \frac{|h(y)|f(y)}{\int |h(y)|f(y)dy}.$$

Proof:

$$\begin{aligned} Var\left(h(Y)\frac{f(Y)}{\tilde{f}(Y)}\right) &= E_{\tilde{f}}\left[h(Y)^2\frac{f(Y)^2}{\tilde{f}(Y)^2}\right] - E_{\tilde{f}}\left[h(Y)\frac{f(Y)}{\tilde{f}(Y)}\right]^2 \\ &= E_{\tilde{f}}\left[h(Y)^2 w(Y)^2\right] - \left(\int h(y)f(y)dy\right)^2. \end{aligned}$$

Since the second term does not depend on $\tilde{f}$, minimizing the variance of the IS estimator reduces to minimizing the quantity $E[w(Y)^2 h(Y)^2]$. By Jensen's inequality, we obtain:

$$E[w(Y)^2 h(Y)^2] = E[(w(Y)|h(Y)|)^2] \geq E[w(Y)|h(Y)|]^2 = E[|h(X)|]^2.$$

Equality holds if and only if $w(Y)|h(Y)|$ is a constant almost surely. Since Y follows the law $\tilde{f}$, this implies the existence of a constant c such that, for all y where $\tilde{f}(y) > 0$,

$$w(y)|h(y)| = c \Leftrightarrow \tilde{f}(y) = \frac{|h(y)|f(y)}{c} \Leftrightarrow \tilde{f}(y) = \frac{|h(y)|f(y)}{\int |h(y)|f(y)dy}.$$

where we used the fact that $\tilde{f}$ is a density to determine the constant c. □

However, this optimal density is impractical as it depends on the very quantity I we aim to estimate. Nonetheless, this result provides valuable insight: a significant variance reduction can be achieved by selecting $\tilde{f}$ such that the ratio $\frac{|h|f}{\tilde{f}}$ remains nearly constant with finite variance.

Quality Assessment of the Generated Samples:

Beyond comparing the IS estimator to the MCC estimator in terms of variance, it is also useful to evaluate the adequacy of the chosen instrumental density $\tilde{f}$. A common criterion for this assessment is the *effective sample size* (ESS), defined as:

$$ESS = \frac{(\sum_{i=1}^n w_i)^2}{\sum_{i=1}^n w_i^2}.$$

This quantity equals n if all importance weights are equal to 1, which corresponds to the case where $\tilde{f} = f$ (i.e., the MCC criterion). The ESS thus represents the number of observations that effectively contribute to the estimation. A value close to n indicates low variance in the importance weights, leading to a more reliable

estimation.

Limitations of Importance Sampling:

The choice of the instrumental distribution is subject to a fundamental constraint. Specifically, the variance of the estimator δ_{IS} may not be finite. If $Var_f(h(X)) < \infty$, it suffices for $\tilde{f}$ to have heavier tails than f (or equivalently, for the ratio $\frac{f}{\tilde{f}}$ to be bounded) to avoid this issue. If the ratio is unbounded, the importance weights will exhibit high variability, leading to erratic fluctuations across iterations.

2.2 Control Variates

The control variate (CV) method is a widely used variance reduction technique. Its key feature is leveraging the correlation between random variables. In particular, we will see that, in a specific case, it necessarily reduces the variance of the classical estimator.

In this section, we define $\theta = E[X]$ as the quantity of interest to estimate. Let Y be a random variable with a known mean $y = E[Y]$. The control variate method consists of generating $(X_1, Y_1), ..., (X_n, Y_n)$, which are i.i.d. pairs following the same distribution as (X, Y), and combining the empirical means $\bar{X_n}$ and $\bar{Y_n}$ to improve the classical estimator: $\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

We define the control variate estimator δ_{CV} as:

$$\boxed{\delta_{CV} = \frac{1}{n}\sum_{i=1}^{n}(X_i - c(Y_i - y)) = \bar{X_n} - c(\bar{Y_n} - y)} \tag{2.2}$$

where the constant c is freely chosen. By default, a naive choice is $c = 1$.

Bias of the estimator:

$$\begin{aligned} E[\delta_{CV}] &= E[\bar{X_n}] - cE[\bar{Y_n}] + cE[y] \\ &= E[X] \end{aligned}$$

Thus, the CV estimator is unbiased.

Convergence of the estimator:

Applying the strong law of large numbers to the i.i.d. sequences $(X_i)_{1\leq i\leq n}$ and $(Y_i)_{1\leq i\leq n}$, we obtain:

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{a.s.} E[X] \\ \frac{1}{n}\sum_{i=1}^{n} Y_i - y \xrightarrow[n\to\infty]{a.s.} 0 \end{cases} \quad \Rightarrow \delta_{CV} \xrightarrow[n\to\infty]{a.s.} E[X].$$

Variance of the estimator:

$$\begin{aligned} Var(\delta_{CV}) &= Var(\bar{X_n} - c(\bar{Y_n} - E[Y])) \\ &= Var(\bar{X_n}) + c^2 Var(\bar{Y_n}) - 2cCov(\bar{X_n}, \bar{Y_n}) \\ &= Var(\bar{X_n}) + \frac{1}{n}\left(c^2 Var(Y) - 2cCov(X,Y)\right) \end{aligned}$$

This expression admits an optimal choice of c that minimizes the variance, given by the following proposition:

Proposition (Optimal constant c)
The control variate estimator δ_{CV} has a variance lower than that of the classical estimator $\bar{X_n}$ if and only if:

$$c^2 Var(Y) - 2cCov(X,Y) < 0.$$

The minimum variance estimator δ^*_{CV} is obtained for:

$$c^* = \underset{c\in\mathbb{R}}{\operatorname{argmin}}(Var(\delta_{CV}(c))) = \frac{Cov(X,Y)}{Var(Y)}.$$

Proof: Let $V = Var(\delta_{CV}(c))$. This function is polynomial and satisfies:

$$\frac{\partial V}{\partial c} = \frac{1}{n}(2cVar(Y) - 2Cov(X,Y)) \quad \text{where } \frac{\partial V}{\partial c} = 0 \Leftrightarrow c = \frac{Cov(X,Y)}{Var(Y)}.$$

$$\frac{\partial^2 V}{\partial c^2} = \frac{2}{n}Var(Y) > 0.$$

Since V is strictly convex, c^* is indeed the variance-minimizing value. □

The optimal choice of c^* leads to the variance:

$$Var(\delta^*_{CV}) = Var(\bar{X_n})(1 - \rho(X,Y)^2),$$

where $\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ denotes the correlation coefficient between X and Y. This result highlights that:

$$Var(\delta^*_{CV}) \leq Var(\bar{X_n}).$$

Thus, the minimum variance estimator δ^*_{CV} always has a lower variance than $\bar{X_n}$ whenever X and Y are correlated. This variance reduction is more pronounced as the correlation increases.

How to determine $\mathbf{c^*}$ without knowing $\mathbf{Cov(X,Y)}$ and $\mathbf{Var(Y)}$?
In most cases, these quantities are unknown. Instead, we use the sample-based estimator:

$$\hat{c}^*_n = \frac{\sum_{i=1}^n (Y_i - y)(X_i - \bar{X_n})}{\sum_{i=1}^n (Y_i - y)^2}.$$

Connection with Linear Regression:

There is also an interesting relationship between control variates and regression analysis. In fact, computing the optimal estimator δ^*_{CV} is equivalent to performing a regression of X on the control variable Y.

Consider the following model:

$$X = aY + b + \epsilon,$$

where a is the regression coefficient, b is the intercept, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ represents the random error.
The ordinary least squares (OLS) estimators of a and b are given by:

$$\hat{a} = \frac{\widehat{Cov(X,Y)}}{\widehat{Var(Y)}} = \frac{\sum_{i=1}^n (Y_i - y)(X_i - \bar{X_n})}{\sum_{i=1}^n (Y_i - y)^2}, \qquad \hat{b} = \hat{X_n} - \hat{a}\hat{Y_n}.$$

The link with the control variate method lies in the relation: $c^* = \hat{a}$. By computing $\hat{b} + \hat{a}y$, we obtain:

$$\begin{aligned} \hat{b} + \hat{a}y &= \hat{X_n} - \hat{a}\hat{Y_n} + \hat{a}y \\ &= \hat{X_n} - \hat{a}(\hat{Y_n} - y) \\ &= \hat{X_n} - c^*(\hat{Y_n} - y) \\ &= \delta^*_{CV}. \end{aligned}$$

Thus, the linear regression model evaluated at $y = E[Y]$ provides the optimal control variate estimator.

2.2.1 Numerical Example (Estimation Using Control Variates)

Consider a random variable $U \sim \mathcal{U}[0,1]$, and let us estimate the quantity:

$$I = E\left[\sin\left(\frac{\pi}{2}U\right)\right].$$

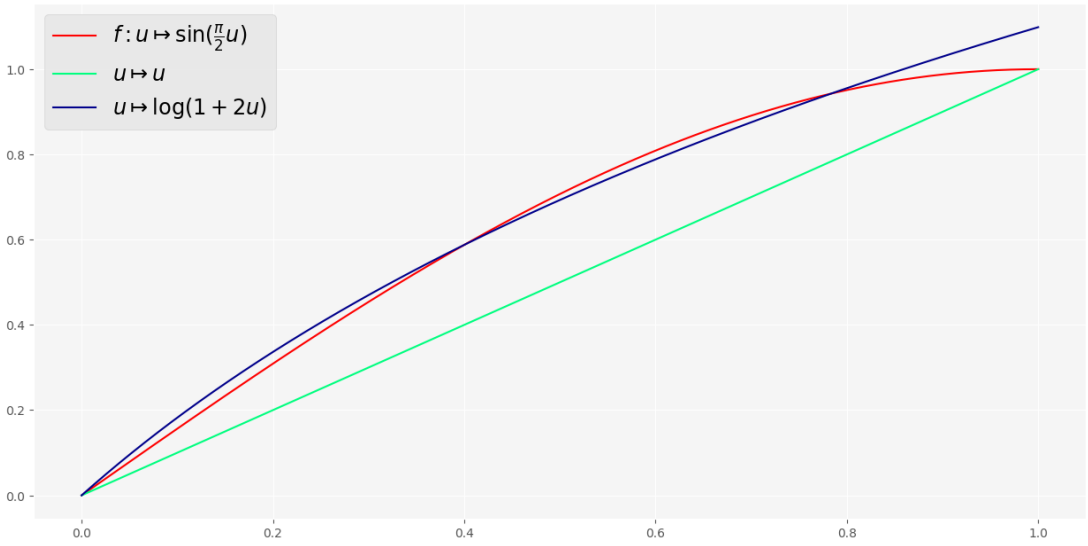
The exact value is 0.6366198, obtained using the *integrate* function in R.
For $u \in [0,1]$, let us define $f : u \mapsto \sin(\frac{\pi}{2}u)$ as the function associated with the quantity of interest. We propose a linear approximation $u \mapsto u$ and a second, less intuitive approximation $u \mapsto \ln(1+2u)$. All three functions are well-defined on $[0,1]$ and take the same values at the interval boundaries.

Based on these approximations, we propose two control variates, whose performances will be compared:

$$Y_1 = U \qquad Y_2 = \ln(1+2U).$$

Their expectations are well known and given by:

$$E[Y_1] = \frac{1}{2} = y_1 \qquad E[Y_2] = 0.6479184 = y_2.$$



Thus, we obtain two equivalent representations of I:

$$I = E\left[\sin\left(\frac{\pi}{2}U\right) - (U - y_1)\right] = E\left[\sin\left(\frac{\pi}{2}U\right) - (\ln(1+2U) - y_2)\right].$$

We define two control variate estimators as follows:

$$\delta^1_{CV} = \frac{1}{n}\sum_{i=1}^{n}\left[\sin\left(\frac{\pi}{2}U_i\right) - \left(U_i - \frac{1}{2}\right)\right] \quad \delta^2_{CV} = \frac{1}{n}\sum_{i=1}^{n}\left[\sin\left(\frac{\pi}{2}U_i\right) - (\ln(1+2U_i) - 0.65)\right],$$

where $(U_1, ..., U_n)$ are i.i.d. random variables following the same distribution as U.

Remark: We have chosen $c = 1$ as the default constant.

Next, we compare the variance of the estimators δ^1_{CV} and δ^2_{CV} with that of the classical Monte Carlo estimator:

$$\delta_{MCC} = \frac{1}{n}\sum_{i=1}^{n}\sin\left(\frac{\pi}{2}U_i\right).$$

For $n = 10,000$, the results obtained using R are presented in the following table:

	δ_{MCC}	δ^1_{CV}	δ^2_{CV}
Mean	0.6356	0.6354	0.6361
Variance	9.57×10^{-2}	4.14×10^{-3}	6.88×10^{-4}

The estimator δ^2_{CV} provides the best approximation of the exact value of I, exhibiting the lowest variance. This result is expected when analyzing the functions associated with the control variates. Indeed, the function $u \mapsto \ln(1+2u)$ closely approximates the target function $f : u \mapsto \sin(\frac{\pi}{2}u)$.
This leads to a strong correlation between the random variables $\sin(\frac{\pi}{2}U)$ and

$\ln(1+2U)$, which in turn results in a significantly reduced variance for the estimator δ^2_{CV}.
As for the estimator δ^1_{CV}, it remains more accurate than the MCC estimator, with a significantly lower variance, despite the fact that the linear function $u \mapsto u$ only moderately approximates the function f.

2.3 Stratification

Consider $\mathcal{X}$ as the set of values taken by a random variable X, and let $(\mathcal{X}_1, ..., \mathcal{X}_N)$ be a partition into N groups (called **strata**) of $\mathcal{X}$:

$$\mathbb{P}[X \in \bigcup_{i=1}^{N} \mathcal{X}_i] = 1 \qquad \text{with} \quad \mathcal{X}_i \bigcap \mathcal{X}_j = \emptyset, \;\; i \neq j.$$

The stratification method consists of forming homogeneous groups and then combining the information contained in each stratum by performing independent sampling within them.

By defining $p_i = \mathbb{P}(X \in \mathcal{X}_i)$, our quantity of interest can be rewritten as:

$$I = E[h(X)] = \sum_{i=1}^{N} E[h(X)|X \in \mathcal{X}_i]\mathbb{P}(X \in \mathcal{X}_i) = \sum_{i=1}^{N} p_i E[h(X)|X \in \mathcal{X}_i].$$

This formulation requires the ability to sample from conditional distributions. If this is feasible, we can approximate each conditional mean and obtain the stratified estimator:

$$\boxed{\delta_{ST} = \sum_{i=1}^{N} p_i \left[\frac{1}{n_i} \sum_{j=1}^{n_i} h\left(X_j^{(i)}\right) \right]} \tag{2.3}$$

where, for each $i \in \{1, \ldots, N\}$, $(X_1^{(i)}, X_2^{(i)}, ..., X_{n_i}^{(i)})$ are i.i.d. samples from the conditional distribution $\mathcal{L}(X|X \in \mathcal{X}_i)$, and $(n_i)_{1 \leq i \leq N}$ (referred to as **allocations**) represent the number of samples drawn in each stratum, subject to the constraint $n_1 + ... + n_N = n$ (total number of simulations).

Bias of the estimator: For any $i \in \{1, \ldots, N\}$, we have:

$$\begin{aligned} E\left[\frac{1}{n_i} \sum_{j=1}^{n_i} h(X_j^{(i)}) \right] &= \frac{1}{n_i} n_i E[h(X)|X \in \mathcal{X}_i] \mathbf{1}_{\{n_i > 0\}} \\ &= E[h(X)|X \in \mathcal{X}_i] \mathbf{1}_{\{n_i > 0\}}. \end{aligned}$$

Thus, we deduce:

$$\begin{aligned} E[\delta_{ST}] &= \sum_{i=1}^{N} p_i E[h(X)|X \in \mathcal{X}_i] \mathbf{1}_{\{n_i > 0\}} \\ &= E[h(X)] \mathbf{1}_{\{n_i > 0\}}. \end{aligned}$$

The stratified estimator is unbiased, provided that $n_i > 0$ for all $i \in \{1, \dots, N\}$, meaning that at least one sample is drawn in each stratum.

Convergence of the estimator: For $i \in \{1, \dots, N\}$ such that $n_i > 0$, the strong law of large numbers applied to the i.i.d. sequence $(X_j^{(i)})_{1 \leq j \leq n_i}$ gives:

$$\frac{1}{n_i} \sum_{j=1}^{n_i} h\left(X_j^{(i)}\right) \xrightarrow[n_i \to \infty]{a.s.} E[h(X)|X \in \mathcal{X}_i].$$

$$\Rightarrow \delta_{ST} \xrightarrow[n \to \infty]{a.s.} \sum_{i=1}^{N} p_i E[h(X)|X \in \mathcal{X}_i] = I.$$

Thus, the stratified estimator is consistent under the same assumption.

Variance of the estimator: Under the assumptions of independence and identical distributions, we obtain:

$$\begin{aligned} Var(\delta_{ST}) &= \sum_{i=1}^{N} \frac{{p_i}^2}{{n_i}^2} Var\left(\sum_{j=1}^{n_i} h\left(X_j^{(i)}\right)\right) \\ &= \sum_{i=1}^{N} \frac{{p_i}^2}{{n_i}^2} n_i Var(h(X)|X \in \mathcal{X}_i) \\ &= \sum_{i=1}^{N} \frac{{p_i}^2}{n_i} \sigma_i^2. \end{aligned}$$

where $\sigma_i^2 = Var(h(X)|X \in \mathcal{X}_i) = E[h^2(X)|X \in \mathcal{X}_i] - E[h(X)|X \in \mathcal{X}_i]^2$.
This quantity can be estimated using the empirical variance:

$$\tilde{\sigma_i}^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} h^2\left(X_j^{(i)}\right) - \left(\frac{1}{n_i} \sum_{j=1}^{n_i} h\left(X_j^{(i)}\right)\right)^2.$$

This provides an approximation for the variance:

$$Var(\delta_{ST}) = \sum_{i=1}^{N} \frac{{p_i}^2}{n_i} \tilde{\sigma_i}^2.$$

The variance depends primarily on two factors: the choice of strata $\mathcal{X}_i$ and the allocation of samples $(n_1, \dots, n_N)$.

To minimize variance in stratified sampling, an optimal choice of allocations must be made. A first result ensures variance reduction compared to the MCC estimator: proportional allocation.

Proposition (Proportional Allocation $(n_i)_{1\leq i\leq N}$)

For each stratum $\mathcal{X}_i, i = 1, \ldots, N$, consider an allocation proportional to the stratum probability: $n_i = np_i$.
In this case, the stratified estimator δ^p_{ST} has lower variance than the classical estimator:

$$Var(\delta^p_{ST}) = \frac{1}{n}\sum_{i=1}^{N} p_i\sigma_i^2 \leq Var(\delta_{MCC}).$$

A second result provides the minimum variance by optimizing the choice of allocations, as stated in the following proposition:

Proposition (Optimal Allocation $(n_i)_{1\leq i\leq N}$)

The minimum variance estimator δ^*_{ST} is obtained with the optimal allocation:

$$(n_1^*, \ldots, n_N^*) = \left(n\frac{p_1\sigma_1}{\sum_{i=1}^N p_i\sigma_i}, \ldots, n\frac{p_N\sigma_N}{\sum_{i=1}^N p_i\sigma_i}\right),$$

which results in the variance:

$$Var(\delta^*_{ST}) = \frac{1}{n}\left(\sum_{i=1}^{N} p_i\sigma_i\right)^2 \leq Var(\delta^p_{ST}).$$

This approach depends on the terms σ_i, which are unknown but can be approximated using the previously introduced $\tilde{\sigma}_i$. To leverage this result, we can first conduct an estimation simulation and then use the following quasi-optimal allocation:

$$(n_1^*, \ldots, n_N^*) = \left(n\frac{p_1\tilde{\sigma}_1}{\sum_{i=1}^N p_i\tilde{\sigma}_i}, \ldots, n\frac{p_N\tilde{\sigma_N}}{\sum_{i=1}^N p_i\tilde{\sigma}_i}\right).$$

2.3.1 Numerical Example (Estimation Using Stratification)

Consider a random variable $U \sim \mathcal{U}[0, 1]$ and let us estimate the quantity:

$$I = E[U] = \int_0^1 u\, du.$$

We define five strata, corresponding to equal-length intervals:

$$\mathcal{X}_1 = [0, 0.2];\ \ \mathcal{X}_2 = [0.2, 0.4];\ \ \mathcal{X}_3 = [0.4, 0.6];\ \ \mathcal{X}_4 = [0.6, 0.8];\ \ \mathcal{X}_5 = [0.8, 1].$$

The associated probabilities p_i are all equal:

$$p_i = \mathbb{P}(U \in \mathcal{X}_i) = \frac{1}{5}, \qquad i = 1, \ldots, 5.$$

We perform $n = 10,000$ simulations with proportional allocation:

$$n_i = np_i = \frac{10000}{5} = 2000, \qquad i = 1, \ldots, 5.$$

The estimators obtained using MCC and stratification are:

$$\delta_{MCC} = \frac{1}{n}\left(\sum_{i=1}^{5}\sum_{j=1}^{2000} U_j^{(i)}\right),$$

$$\delta_{ST} = \sum_{i=1}^{5} p_i \left(\frac{1}{n_i}\sum_{j=1}^{n_i} U_j^{(i)}\right) = \sum_{i=1}^{5}\frac{1}{5}\left(\frac{1}{2000}\sum_{j=1}^{2000} U_j^{(i)}\right).$$

The variances of these two estimators are given by:

$$Var(\delta_{MCC}) = \frac{1}{n}\left(E[U^2] - I^2\right) = \frac{1}{n}\left(1 - (1/2)^2\right),$$

$$Var(\delta_{ST}) = \frac{1}{n}\sum_{i=1}^{5} p_i \tilde{\sigma_i}^2 = \frac{1}{n}\sum_{i=1}^{5}\frac{1}{5}\left(\frac{1}{2000}\sum_{j=1}^{2000}\left(U_j^{(i)}\right)^2 - \left(\frac{1}{2000}\sum_{j=1}^{2000} U_j^{(i)}\right)^2\right).$$

The numerical application of these formulas using R yields:

Case $n = 10,000$

	δ_{MCC}	δ_{ST}^p
Mean	0.500168	0.5003644
Variance	8.333333-06	3.300696e-07

Case $n = 100,000$

	δ_{MCC}	δ_{ST}^p
Mean	0.4996247	0.4999157
Variance	8.333333e-07	3.339183e-08

The stratified estimator (ST) provides a better approximation than the Monte Carlo estimator (MCC) as the number of simulations increases. In terms of variance, the advantage is clearly in favor of stratification.

Now, we use the optimal allocation for choosing n_i. The variance of δ_{ST}^* is given by:

$$Var(\delta_{ST}^*) = \frac{1}{n}\left(\sum_{i=1}^{5}\frac{1}{5}\sqrt{\frac{1}{2000}\sum_{j=1}^{2000}\left(U_j^{(i)}\right)^2 - \left(\frac{1}{2000}\sum_{j=1}^{2000} U_j^{(i)}\right)^2}\right)^2.$$

The numerical application yields:

Case $n = 10,000$

	δ_{ST}^*
Variance	3.300587×10^{-7}

Case $n = 100,000$

	δ_{ST}^*
Variance	3.339155×10^{-8}

The optimal allocation logically yields an even lower variance than proportional allocation, although the difference is very small, if not negligible.

3 Application to a Real-World Context

In this example, our goal is to apply importance sampling to estimate an extreme event in the context of sports. The associated code is provided in Appendix, Section 4. The dataset is available in the same section and can be downloaded via the following link: **Download the CSV Data**.

The focus of our study is LOSC (Lille's football team), and we aim to answer the question: **What is the probability that LOSC wins a match by at least 8 goals in a league competition?**
Notably, this feat has not been achieved by the club since 2005. To address this problem, we collect a sample $(X_1, \ldots, X_n)$ of $n = 746$ observations based on LOSC's last 20 league seasons. The dataset, constructed from publicly available statistics on the official league website `https://ligue1.fr`, is presented in Appendix, Section 4.

The observations X_i in the sample are computed as the difference between the number of goals scored and the number of goals conceded per match:

$$X_i = \begin{cases} > 0 & \text{in case of a win,} \\ = 0 & \text{in case of a draw,} \\ < 0 & \text{in case of a loss.} \end{cases}$$

The main statistical characteristics of the data are as follows:

$$\mu = 0.42, \qquad \sigma = 1.63, \qquad \min(X_i) = -6, \qquad \max(X_i) = 8.$$

The event of interest can be formulated as:

$$I = \mathbb{P}(X \geq 8) = E_f[\mathbf{1}_{X \geq 8}],$$

where X follows an unknown density f, inferred from the observed data.

By constructing the histogram of the observations, we notice a distribution slightly shifted from the standard normal distribution $\mathcal{N}(0, 1)$. The mean and standard deviation of our sample are given by $\mu = 0.42$ and $\sigma = 1.63$.
Thus, it is beneficial to normalize the data to obtain a known distribution:

$$N = \frac{X - \mu}{\sigma} = \frac{X - 0.42}{1.63}.$$

The raw data and the normalized values $(N_i)_{1 \leq i \leq n}$ are then represented as follows:

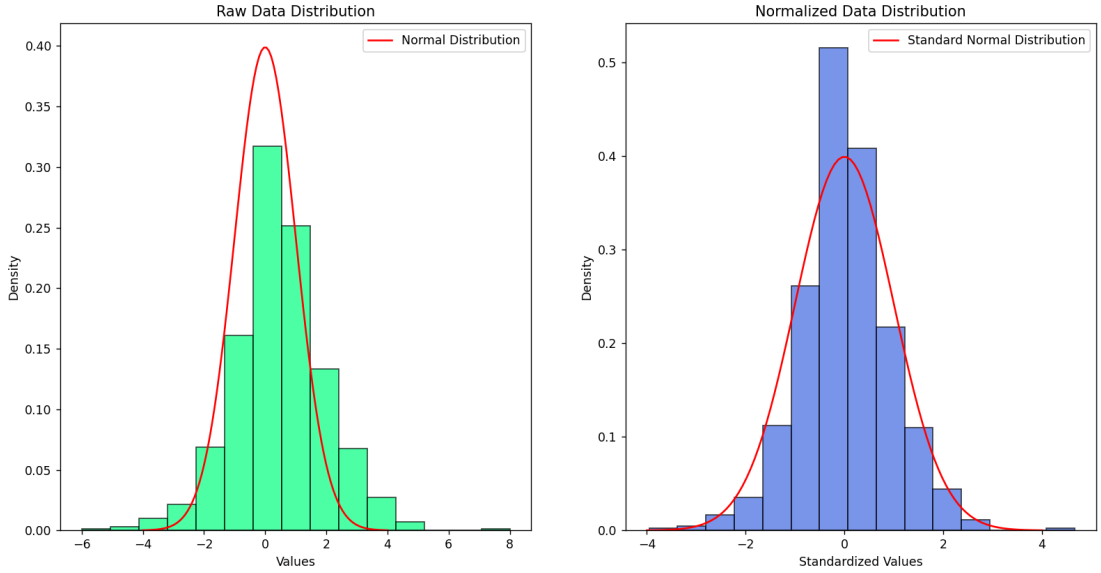


Figure 1: Histograms of Raw and Normalized Data

The second distribution closely resembles a standard normal distribution $\mathcal{N}(0,1)$. We can thus approximate the law of N using this reference. For our estimation, we express N in terms of a known distribution:

$$I = \mathbb{P}(X \geq 8) = \mathbb{P}\left(N \geq \frac{8-\mu}{\sigma}\right) = E_g[\mathbf{1}_{N \geq 4.66}],$$

where g is the density function of N, which serves as our target distribution (in this case, the standard normal distribution).

For the instrumental distribution, we propose a Student's t-distribution with 2 degrees of freedom, given by the density:

$$\tilde{g}(x) = \frac{1}{\sqrt{2\pi}}\Gamma\left(\frac{3}{2}\right)\left(1+\frac{x^2}{2}\right)^{-\frac{3}{2}}.$$

This density function has a support that includes that of g, approximates g well, and assigns more weight to the distribution tails (in this case, the right tail, which is relevant for our study). Thus, it serves as a suitable candidate for our method.

For a sequence $(Y_i)_{1 \leq i \leq n}$ of i.i.d. random variables following the law $\tilde{g}$, we obtain the importance sampling estimator:

$$\delta_{IS} = \frac{1}{n}\sum_{i=1}^{n}\frac{g(Y_i)}{\tilde{g}(Y_i)}h(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{e^{-\frac{1}{2}Y_i^2}}{\Gamma(\frac{3}{2})}\left(1+\frac{Y_i^2}{2}\right)^{\frac{3}{2}}\mathbf{1}_{\{Y_i \geq 4.66\}} = 1.655109 \times 10^{-6}.$$

In contrast, the classical Monte Carlo estimator yields: $\delta_{MC} = 1.340483 \times 10^{-3}$ (only one value in the sample is greater than or equal to 8).

The exact probability of the event, obtained using the *"pnorm"* function in R, is approximately 1.55528×10^{-6}. As expected, the importance sampling estimator is significantly closer to the true value than the naive Monte Carlo approach.

Now, we compare the variances of both estimators:

- $Var(\delta_{MC}) = \frac{1}{n} Var\left(\mathbf{1}_{N \geq 4.66}\right) = 1.338686 \times 10^{-3}$
- $Var(\delta_{IS}) = \frac{1}{n} Var\left(\mathbf{1}_{Y \geq 4.66} \frac{g(Y)}{\tilde{g}(Y)}\right) = 9.695668 \times 10^{-10}$

The importance sampling estimator reduces variance by a factor of approximately 10^6 compared to the classical Monte Carlo method, demonstrating the effectiveness of this approach for rare event estimation.

It is also useful to compute the *effective sample size* (ESS) to assess the quality of our choice for $\tilde{g}$. Using its formula, we obtain:

$$ESS = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2} = \frac{\left(\sum_{i=1}^{n} \frac{e^{-\frac{1}{2}Y_i^2}}{\Gamma\left(\frac{3}{2}\right)} \left(1 + \frac{Y_i^2}{2}\right)^{\frac{3}{2}}\right)^2}{\sum_{i=1}^{n} \frac{e^{-Y_i^2}}{\Gamma(\frac{3}{2})^2} \left(1 + \frac{Y_i^2}{2}\right)^3} = 661.3292.$$

This value is close to our sample size $n = 746$, confirming that the choice of the Student's $t(2)$ distribution was appropriate. Moreover, the result is consistent with the low variance observed for the importance sampling estimator.

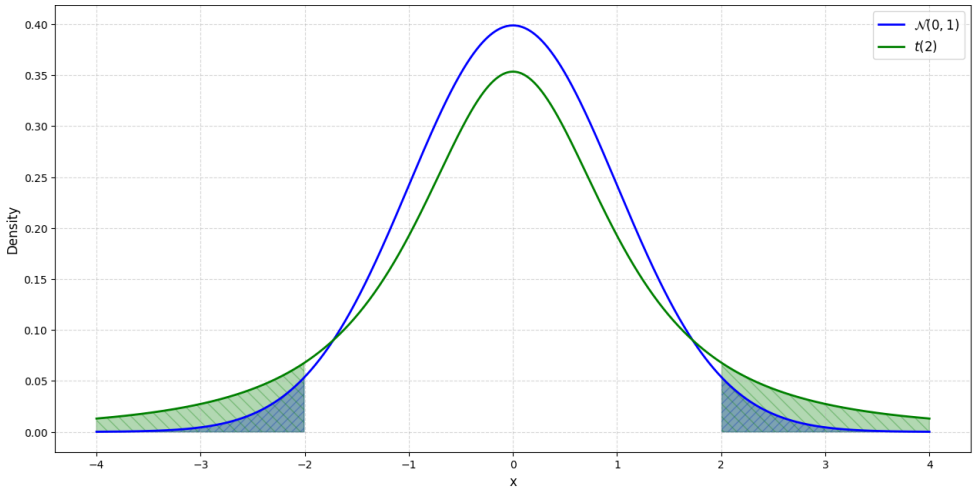


Figure 2: Comparison of the distribution tails between the target and instrumental distributions

4 Appendices

R Code (Numerical Example: Estimation using Control Variates)

```
set.seed(1)

f <- function(u) {
  return(sin(0.5*pi*u))
}

monte_carlo <- function(sample, n) {
  mean_value <- mean(sample)
  variance_value <- ((n-1)/n) * var(sample)
  return(c(mean = mean_value, variance = variance_value))
}

n <- 10^4

sample_U <- runif(n)

sample_MC <- f(sample_U)
sample_CV1 <- f(sample_U) - (sample_U - 1/2)
sample_CV2 <- f(sample_U) - (log(1 + 2*sample_U) - 0.6479184)

result_table <- data.frame(
  MC = monte_carlo(sample_MC, n),
  CV1 = monte_carlo(sample_CV1, n),
  CV2 = monte_carlo(sample_CV2, n)
)

rownames(result_table) <- c("Mean", "Variance")
result_table
```

R Code (Numerical Example: Estimation using Stratification)

```
set.seed(1)

n <- 10^4
num_strata <- 5
strata_values <- seq(0, 1, length.out = num_strata + 1)

sample_unif <- runif(n)

compute_mean <- function(i) {
  return(mean(sample_unif[sample_unif >= strata_values[i]
           & sample_unif < strata_values[i + 1]]))
}
compute_mean_sq <- function(i) {
  return(mean(sample_unif[sample_unif >= strata_values[i]
         & sample_unif < strata_values[i + 1]]^2))
}

mean_stratum <- sapply(1:num_strata, function(i) compute_mean(i))
mean_sq_stratum <- sapply(1:num_strata, function(i) compute_mean_sq
    (i))

estimator_MC <- sum(sample_unif) / n
estimator_ST <- sum((1/5) * mean_stratum)

variance_MC <- (integrate(function(u) u^2 *
        dunif(u, 0, 1), -Inf, Inf)$value - (1/2)^2) / n
variance_ST <- sum((1/5) * (mean_sq_stratum - mean_stratum^2)) / n
variance_ST_min <- sum((1/5) * sqrt(mean_sq_stratum - mean_stratum^
    2))^2 / n

cat("Monte Carlo Estimator:", estimator_MC)
cat("Stratified Sampling Estimator:", estimator_ST)

cat("Monte Carlo Variance:", variance_MC)
cat("Stratified Sampling Variance:", variance_ST)
cat("Minimum Stratified Sampling Variance:", variance_ST_min)
```

R Code (LOSC Application)

```
set.seed(4)

data <- read.csv2("directory_name/LOSC.csv", header=F,
       stringsAsFactors = FALSE, na.strings = c(""))
X <- as.vector(as.matrix(data))
X <- X[!is.na(X)]

n <- length(X)
mu <- mean(X)
sigma <- sd(X)
N <- (X - mu) / sigma

exact_value <- 1 - (pnorm(4.663381, 0, 1))

MC_estimator <- mean(ifelse(X >= 8, 1, 0))

Y <- rt(n, 2)
w <- dnorm(Y) / dt(Y, 2)
IS_estimator <- mean(w * ifelse(Y >= 4.663381, 1, 0))

ESS <- function(n) {
  sum_1 <- 0
  sum_2 <- 0
  for (i in 1:n) {
    sum_1 <- sum_1 + w[i]
    sum_2 <- sum_2 + w[i]^2
  }
  return(sum_1^2 / sum_2)
}

cat("Monte Carlo Estimation:", MC_estimator)
cat("Importance Sampling Estimation:", IS_estimator)
cat("Effective Sample Size (ESS):", ESS(n))

var_MC <- ((n-1)/n) * var(ifelse(N >= 4.663381, 1, 0))
var_IS <- ((n-1)/n) * var(ifelse(Y >= 4.663381, 1, 0) * w)
cat("Monte Carlo Variance:", var_MC)
cat("Importance Sampling Variance:", var_IS)
```

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
J1	2	1	1	0	0	-1	0	0	1	1	0	-1	-1	3	2	1	0	0	3	0
J2	-3	0	4	1	-2	-1	0	-1	0	-1	1	0	1	-3	0	-1	1	-4	0	2
J3	0	-1	-1	0	-1	0	0	1	0	1	2	0	0	-2	3	3	1	0	-6	-3
J4	-1	0	3	0	1	-3	0	1	-1	0	0	1	-3	0	-1	-2	0	1	2	1
J5	1	0	0	0	0	1	3	2	0	-2	2	0	-1	0	1	1	2	-1	-1	0
J6	2	2	-2	-1	1	0	1	0	0	2	0	0	-1	-1	1	0	3	-1	2	-1
J7	2	4	-3	0	2	0	0	0	-2	3	-1	-1	-2	-4	-1	2	4	1	-1	2
J8	4	-1	1	0	1	1	2	0	2	0	1	-1	1	-3	3	0	0	1	1	0
J9	1	2	0	0	0	0	-2	2	0	3	-3	2	-1	0	2	0	0	2	-1	1
J10	1	-1	1	3	0	-1	-2	2	1	1	-1	0	1	-1	1	-1	-1	-1	1	2
J11	-1	0	0	-1	0	4	0	2	2	1	-2	-1	-1	-1	1	3	4	0	3	0
J12	1	2	3	-3	2	2	2	0	1	2	0	0	-1	3	-1	-1	0	-1	1	0
J13	2	0	1	1	-1	2	3	0	-2	0	-2	0	-1	2	0	0	1	0	-1	2
J14	0	2	0	0	3	-2	1	0	-1	1	1	0	0	-3	-2	-2	1	0	0	2
J15	0	-1	0	-2	0	4	0	2	0	1	-1	-2	2	1	0	1	2	0	1	0
J16	0	1	2	0	0	1	3	1	0	1	0	1	1	1	1	1	0	1	2	0
J17	0	-1	0	-1	0	4	1	2	2	-1	0	1	1	-2	0	1	1	1	0	-1
J18	1	0	-2	3	1	3	3	0	0	1	3	3	-2	-3	1	1	-1	0	0	3
J19	0	2	0	1	2	4	0	0	3	0	-1	0	0	0	-1	-4	1	1	4	0
J20	-1	2	0	0	0	2	2	-2	0	-1	1	0	0	1	2	-1	1	2	-1	4
J21	2	-1	1	0	1	-1	1	3	-2	-2	-1	-1	0	-1	1	-2	1	0	0	-2
J22	1	3	1	2	0	1	0	1	-1	0	-1	-2	1	-1	1	1	1	-2	2	3
J23	0	0	0	0	1	0	2	-1	0	-1	0	0	-1	1	4	1	3	-4	2	-2
J24	0	2	-1	0	-2	2	-1	1	2	2	1	1	-1	-3	2	2	2	1	-1	1
J25	0	-2	-1	2	1	1	0	0	2	0	0	0	-1	0	0	-1	0	0	1	0
J26	0	0	-1	0	-2	-1	1	0	2	0	-1	0	1	0	0	3	3	1	0	0
J27	0	2	4	-1	2	0	1	-1	1	1	1	1	-1	-1	1	1	0	4	0	1
J28	-2	1	0	0	1	1	1	4	2	2	1	-3	0	-1	1	1	2	0	2	2
J29	0	4	-2	5	1	-2	2	3	-1	0	3	2	1	0	-1		0	1	2	-1
J30	0	0	-2	1	2	3	-1	1	1	0	-2	1	0	-1	1		-1	0	-1	1
J31	8	0	-3	-2	0	-1	0	-2	5	1	2	1	1	-1	0		1	0	1	1
J32	-2	0	2	0	-1	0	0	3	0	1	1	3	-1	-1	4		2	-1	0	-1
J33	1	0	-2	1	-1	4	5	2	1	1	2	3	-2	0	0		0	-1	3	1
J34	1	0	-1	2	-2	1	1	1	0	0	-5	2	3	-4	5		1	1	-1	0
J35	1	1	-1	1	0	2	1	1	-2	1	2	0	3	2	0		2	-3	0	
J36	2	-2	1	0	0	2	1	3	3	0	-1	1	-2	1	1		3	-1	1	
J37	-2	4	-1	1	1	1	0	-1	0	-2	-4	0	-4	1	5		0	2	1	
J38	2	0	0	0	1	-1	1	3	0	3	3	1	3	-5	-2		1	0	0	

Table 1: Sample Data from Observations (LOSC Application)**Download Link**

5 Bibliography

S. Rasmussen and P. W. Glynn, *Stochastic Simulation: Algorithms and Analysis*, Springer, 2009.
- Chapter 5: "Variance-Reduction Methods"
- Chapter 6: "Rare-Event Simulation"

C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 2004.

R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method*, Wiley, 2016.

Pour la Science, no. 385, French edition of *Scientific American*, November 2009.
- Chapter: "Hasard et incertitude, les défis qu'ils posent".