Statistical Analysis of Simulation Data

Bo Li, Spring 2019

(A) Data Processing
(B) I.I.D. Output
(C) Stationary Output
(D) Asymptotically stationary Output
(E) Empirical CDF
(F) Kernel Density Estimation

(A) Data Processing

Visualization

Histogram: Divide an underlying interval in $\mathbb{R}$ into small intervals. Count how many data points in each small interval. With each small interval, plot a rectangle with height or ~~volume~~ area the number or frequency of the data points in that interval.

Scatter plot: Plot data points in $\mathbb{R}^d$ for $d=1, 2$, or 3.

Empirical CDF: (Cummulative ~~Density~~ Distribution Function). Approximation of a true CDF, represented as graph of a one-variable function.

Density plot: Approximation of a true PDF for a real random variable.

Data Characterizing Numbers

Let $X_1, \cdots, X_N$ be i.i.d. random varibles in $\mathbb{R}^d$.
The sample mean is
$$\frac{1}{N}\sum_{i=1}^{N} X_i .$$
If $d=1$ and $X_{(1)} \le \cdots \le X_{(N)}$, then the sample median is
$$\begin{array}{ll} X_{((N+1)/2)} & \text{if } N \text{ is odd,} \\ [X_{(N/2)} + X_{(N/2+1)}]/2 & \text{if } N \text{ is even.} \end{array}$$
The range is $X_{(N)} - X_{(1)}$.
The sample variance for $X_1, \cdots X_N \in \mathbb{R}^d$ is
$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i^2 - N\bar{X}^2\right)$$
where $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ and $A^2 = \|A\|^2$ $(A \in \mathbb{R}^d)$.

The sample standard deviation is S. (with S^2 the sample variance).

The sample ~~k-th~~ moment is $\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^k$ with $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$.

For $d=1$, The sample γ-quantile or $\gamma \times 100$ percentile of $X_1, \cdots, X_N$ is $X_{(\lceil \gamma N \rceil)}$, where $X_{(1)} \le \cdots \le X_{(N)}$ and $\lceil \alpha \rceil$ is the smallest integer $\ge \alpha$.

Let $(X_1, Y_1), \cdots, (X_N, Y_N) \in \mathbb{R}^m \times \mathbb{R}^m$ be i.i.d. random vectors sampled from a bivariate distribution. The sample covariance is

$$\frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X}) \cdot (Y_i - \bar{Y}),$$

where $\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$ and $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$. The sample correlation ~~coe~~ coefficient is

$$\frac{\sum_{i=1}^{N} (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}}.$$

Let X be a random variable of finite expectation $\mu = \mathbb{E}(X)$. Suppose ~~a~~ an approximate $\hat{\mu} \approx \mu$ and $r > 0$ satisfy

$$\mathbb{P}\{ \mu \in [\hat{\mu} - r, \hat{\mu} + r] \} = 0.95 \text{ (or } \alpha)$$

then we say that the interval $[\hat{\mu} - r, \hat{\mu} + r]$ is of 95% (or α percentage) confidence

(B) I. I. D. Output.

Let $X_1, \cdots, X_N$ be random variables in $\mathbb{R}$, i.i.d. according to a density f, with particularly $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = Var(X_1)$, both finite. Denote $\bar{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i$. Then

$$\mathbb{E}(\bar{X}_N) = \mu \quad \text{and} \quad Var(\bar{X}_N) = \frac{\sigma^2}{N}.$$

The central limit theorem implies that for large N,

$$\frac{\bar{X}_N - \mu}{\sqrt{\sigma^2/N}} \sim N(0,1) \text{ approximately.}$$

One can verify for the sample variance

$$S_N^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X}_N)^2$$

that $\mathbb{E}(S_N^2) = \sigma^2$ (exactly) and $\lim_{N\to\infty} S_N^2 = \sigma^2$ with probability 1. Thus for $N \gg 1$,

$$\frac{\tilde{X}_N - \mu}{\sqrt{S_N^2/N}} \sim N(0,1) \text{ approximately.}$$

Let $\Phi(x)$ be the CDF of the standard normal distribution. Given $\alpha \in (0,1)$, let z_α be the unique real number such that $1 - \Phi(z_\alpha) = \alpha$, i.e.,

$$1 - \alpha = \Phi(z_\alpha) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_\alpha} e^{-\frac{x^2}{2}}\,dx = \int_{-\infty}^{z_\alpha}\phi(x)\,dx$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

Let $r > 0$ and $\gamma \in (0,1)$. We have

$$\mathbb{P}\left(\mu \in \left[\bar{X}_N - r\frac{S_N}{\sqrt{N}},\ \bar{X}_N + r\frac{S_N}{\sqrt{N}}\right]\right) = \gamma$$

$$\iff \mathbb{P}\left(\frac{\bar{X}_N - \mu}{S_N/\sqrt{N}} \in [-r, r]\right) = \gamma$$

$$\iff \int_{-r}^{r} \overset{\phi}{\cancel{\Phi}}(x)\,dx = \gamma$$

$$\iff \frac{\gamma+1}{2} = \int_{-\infty}^{r} \overset{\phi}{\cancel{\Phi}}(x)\,dx = \Phi(r)$$

$$\left[\text{using: } \int_{-\infty}^{\infty} \overset{\phi}{\cancel{\Phi}}(x)\,dx = 1 \text{ and } \underset{\phi}{\cancel{\Phi}}(-x) = \underset{\phi}{\cancel{\Phi}}(x)\right]$$

Therefore $[\bar{X}_N - r S_N/\sqrt{N}, \bar{X}_N + r S_N/\sqrt{N}]$ is approximately (when $N \gg 1$) a γ-percentage confidence interval for μ, where $r > 0$ is determined by $r = \Phi^{-1}\left(\frac{\gamma+1}{2}\right)$.

For example, $\gamma = 95\% = 0.95$. $\frac{\gamma+1}{2} = 0.975$.
$r \approx 1.96$.

Generalization to vector-valued random variables.

Let $\vec{X}_1, \cdots, \vec{X}_N, \cdots$ be random variables in $\mathbb{R}^d$ i.i.d, with the distribution f. Suppose the common expectation and variance are

$$\vec{\mu} = \mathbb{E}(\vec{X}_1) \quad \text{and} \quad \sigma^2 = Var(\vec{X}_1)$$

both finite. Then, an approximate $\gamma \in (0,1)$ confidence region for $\vec{\mu}$ is

$$\left\{ \vec{x} \in \mathbb{R}^d : (\bar{\vec{X}}_N - \vec{x})^T (\hat{\Sigma})^{-1} (\bar{\vec{X}}_N - \vec{x}) \le \frac{\chi^2_{d,\gamma}}{N} \right\},$$

where $$\hat{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N} (\vec{X}_i - \bar{\vec{X}}_N) \cdot (\vec{X}_i - \bar{\vec{X}}_N)^T$$

is the sample covariance matrix, and $\chi^2_{d,\gamma}$ is the γ-quantile of χ^2_d distribution.

The Delta Method

Suppose $\vec{X}_1, \ldots, \vec{X}_N, \ldots \overset{\text{i.i.d}}{\sim} f : \mathbb{R}^d \to [0,\infty)$. Then

$$\sqrt{N}\left(\overline{\vec{X}_N} - \vec{\mu}\right) \xrightarrow{\text{dist.}} \vec{K} \sim \mathcal{N}(\vec{0}, \Sigma),$$

where $\overline{\vec{X}_N} = \frac{1}{N}\sum_{i=1}^{N} \vec{X}_i$. Then, for any C^1-function $\vec{g}$, we have,

$$\sqrt{N}\left(\vec{g}(\overline{\vec{X}_N}) - \vec{g}(\vec{\mu})\right) \xrightarrow{\text{dist.}} \vec{R} \sim \mathcal{N}(\vec{0}, J\Sigma J^T),$$

where

$$J = J_{\vec{g}}(\vec{\mu}) = \left(\frac{\partial g_i(\vec{\mu})}{\partial x_j}\right)$$

is the Jacobi matrix of $\vec{g}$ at $\vec{\mu}$.

Simple proof.

$$\vec{g}(\overline{\vec{X}_N}) = \vec{g}(\vec{\mu}) + J_{\vec{g}}(\vec{\mu})(\overline{\vec{X}_N} - \vec{\mu}) + O(\|\overline{\vec{X}_N} - \vec{\mu}\|^2).$$

As $N \to \infty$,

$$\sqrt{N}\left(\vec{g}(\overline{\vec{X}_N}) - \vec{g}(\vec{\mu})\right) \approx \sqrt{N}\, J_{\vec{g}}(\vec{\mu})\,(\overline{\vec{X}_N} - \vec{\mu}) \longrightarrow J_{\vec{g}}(\vec{\mu})\vec{K} = \vec{R},$$

where $\vec{K} \sim \mathcal{N}(\vec{0}, \Sigma)$. Thus $\vec{R} = J_{\vec{g}}(\vec{\mu})\vec{K} \sim \mathcal{N}(\vec{0}, J\Sigma J^T)$, $J = J_{\vec{g}}(\vec{\mu})$.

(C) Stationary Output

We now consider a stationary stochastic process $X_1, X_2, \cdots$, where each $X_j \in \mathbb{R}$ is a random variable. We recall that the sationarity means that, for any integers $n \geq 1$ and $k \geq 1$, and any $x_1, \cdots, x_n \in \mathbb{R}$, the joint distributions

$$\mathbb{P}(X_1 < x_1, \cdots, X_n < x_n) = \mathbb{P}(X_{k+1} < x_1, \cdots, X_{k+n} < x_n).$$

In particular,

$$\mathbb{P}(X_k < x) = \mathbb{P}(X_1 < x) \quad \forall k \geq 1, \forall x \in \mathbb{R},$$

i.e., all X_k $(k \geq 1)$ have the same distribution. We set $\mu = \mathbb{E}(X_k)$, $\sigma^2 = Var(X_k)$ $(k \geq 1)$, both assumed to be finite. Moreover, for any $i \in \mathbb{N}$ and any $k \in \mathbb{N}$,

$$Cov(X_i, X_{i+k})$$

is independent of i.

By the Law of Large Numbers and the Central Limit Theorem, we have with $\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ that

$$\lim_{N \to \infty} \overline{X}_N = \mu \quad \text{with probability } 1,$$

$$\frac{\overline{X}_N - \mu}{\sqrt{Var(\overline{X}_N)}} \sim N(0,1) \quad \text{approximately for } N \gg 1.$$

If $\hat{V}_N$ is a good estimator of $Var(\overline{X}_N)$, then, for $\gamma \in (0,1)$, we can determine $r > 0$, with $r = \Phi^{-1}\left(\frac{\gamma+1}{2}\right)$, $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$,

So that
$$\left[\bar{X}_N - r_\gamma\sqrt{\hat{V}_N},\ \bar{X}_N + r_\gamma\sqrt{\hat{V}_N}\right]$$
is a γ-confindence interval for μ.

However, unlike the i.i.d. case, the sample variance
$$S_N^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X}_N)^2$$
is no longer a good estimator of $Var(\bar{X}_N)$ in general. It is still true that
$$\lim_{N\to\infty} S_N^2 = Var(X_1) = \sigma^2 \text{ with prob. 1.}$$
But in general:
$$Var(\bar{X}_N) \neq Var(X_1)/N.$$
Since X_i's are not necessary independent.

Observe that
$$Var(\bar{X}_N) = \frac{1}{N^2} Var(X_1 + \cdots + X_N)$$
$$= \frac{1}{N^2}\sum_{i,j=1}^{N} Cov(X_i, X_j)$$

Define
$$c(k) = Cov(X_i, X_{i+k}) \quad \forall i \geq 1, \forall k \geq 0.$$

- $c(k)$ is independent of $i \geq 1$ by the stationarity.
- $c(0) = Cov(X_i, X_i) = Var(X_i) = \sigma^2 \quad (\forall i \geq 1).$
- If $k \in \mathbb{Z}$ and $k < 0$, we define
$$c(k) = c(-k)$$

This definition is consistent with the case that $k>0$. If $k \in \mathbb{Z}$, $k<0$, we can choose $i \in \mathbb{Z}$ s.t. $i+k>0$. Hence,

$$\begin{aligned} c(-k) &= Cov(X_i, X_{i-k}) \\ &= Cov(X_{i+k}, X_i) \\ &= Cov(X_i, X_{i+k}) \\ &= c(k). \end{aligned}$$

The function $c(k)$ $(k \in \mathbb{Z})$ is <u>the (auto) covariance function</u> of the process X_n $(n=1,2,\cdots)$.

Now, we have

$$\begin{aligned} Var(\bar{X}_N) &= \frac{1}{N^2} \sum_{i,j=1}^{N} c(j-i) \\ &= \frac{1}{N^2}\left[N c(0) + (N-1)c(1) + (N-1)c(-1) + \cdots\right] \\ &= \frac{1}{N^2} \sum_{k=-N}^{N} (N-|k|)\, c(k) \\ &= \frac{1}{N} \sum_{k=-N}^{N} \left(1 - \frac{|k|}{N}\right) c(k) \end{aligned}$$

<u>Proposition</u> Assume

$$V := \sum_{k=-\infty}^{\infty} c(k) = \sum_{k=-\infty}^{\infty} Cov(X_1, X_{k+1})$$

converges absolutely. Then

$$\lim_{N\to\infty} N\, Var(\bar{X}_N) = V.$$

<u>Proof</u>. By the above calculations, we have

$$\begin{aligned} N\, Var(\bar{X}_N) &= \sum_{k=-N}^{N} \left(1 - \frac{|k|}{N}\right) c(k) \\ &= \sum_{k=-N}^{N} c(k) - \frac{1}{N} \sum_{k=-N}^{N} |k|\, c(k). \end{aligned}$$

Since $\sum_{k=-\infty}^{\infty} c(k) = V$ with absolute convergence it suffices to show that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=-N}^{N} |k|\,|c(k)| = 0. \qquad (*)$$

$\forall \varepsilon > 0$. Since $\sum_{-\infty}^{\infty} |c(k)| < \infty$, there exists $N_0 \in \mathbb{N}$

Such that

$$\sum_{|k|>N_0} |c(k)| < \varepsilon.$$

Thus, for $N > N_0$,

$$\sum_{k=-N}^{N} \frac{|k|}{N} |c(k)| = \frac{1}{N} \sum_{|k|\le N_0} |k|\, c(k) + \sum_{|k|>N_0} \frac{|k|}{N} |c(k)|$$

$$\le \frac{1}{N} \sum_{|k|\le N_0} |k|\, c(k) + \sum_{|k|>N_0} |c(k)|$$

$$\le \frac{1}{N} \sum_{|k|\le N_0} |k|\, c(k) + \varepsilon.$$

Thus $\limsup_{N\to\infty} \sum_{k=-N}^{N} \frac{|k|}{N} |c(k)| \le \varepsilon.$

Hence, since $\varepsilon > 0$ is arbitrary, $(*)$ is true. $\square$

<u>The Covariance Method</u> We continue our discussions and now study how to estimate $V = c(0) + 2\sum_{k=1}^{\infty} c(k)$, where $c(k) = \mathrm{Cov}(X_i, X_{i+k})$.

Let $\hat{c}_N(k) = \frac{1}{N-k} \sum_{j=1}^{N-k} (X_j - \bar{X}_N)(X_{j+k} - \bar{X}_N).$

(This is a random variable.) Then for each k,

$$\lim_{N\to\infty} \hat{c}_N(k) = c(k) \text{ with probability } 1.$$

A natural estimator of V is then

$$\hat{V}_N^* = \hat{c}_N(0) + 2\sum_{k=1}^{N-1} \hat{c}_N(k).$$

But, it turns out this is a bad estimator.

Here is an example. $N = 1{,}000$, $c(k) \approx e^{-|k|/10}$. Then $\hat{c}_N(k) \approx 0$ for $k \geq 100$. Hence, $\sum_{i=101}^{1{,}000} \hat{c}_N(k)$ is mostly "noise". Heuristically, $Var(\hat{c}_N(k)) \approx O(1/N)$. So, $Var(\hat{V}_N^*) \approx N \cdot O(N^{-1}) = O(1)$. So, the variance of $\hat{V}_N^*$ does not converge to 0.

A better estimator of V is

$$\hat{V}_{N,L} = \hat{c}_N(0) + 2 \sum_{k=1}^{L} \hat{c}_N(k),$$

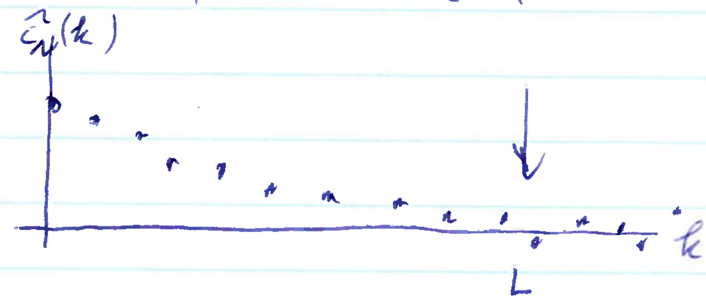
where $L \gg 1$ is fixed! In this case, since $N\,Var(\bar{X}_N) \to V$ as $N \to \infty$, $\hat{V}_{N,L}/N$ is a good estimator of $Var(\bar{X}_N)$. Hence, since

$$\frac{\bar{X}_N - \mu}{\sqrt{Var(\bar{X}_N)}} \sim N(0,1) \quad \text{approximately for } N \gg 1,$$

The interval $\left[\bar{X}_N - r\sqrt{\frac{\hat{V}_{N,L}}{N}},\ \bar{X}_N + r\sqrt{\frac{\hat{V}_{N,L}}{N}}\right]$, where $r = \Phi^{-1}\left(\frac{\gamma+1}{2}\right)$ with $\gamma \in (0,1)$, is a γ-confidence interval for μ.

How to choose L?

Method 1. Choose L such that $\hat{c}_N(k)$ is indistinguishable from noise for all $k > L$.



Method 2: Self-consistent windowing.

Define $\tau = V / Var(X_1) = V/\sigma^2$, where $V = \sum_{k=-\infty}^{\infty} C(k) = \sum_{k=-\infty}^{\infty} Cov(X_0, X_k)$. Then

$$\tau = \sum_{k=-\infty}^{\infty} \frac{Cov(X_0, X_k)}{\sigma^2} = \sum_{k=-\infty}^{\infty} \underset{\uparrow}{Corr}(X_0, X_k)$$

Corr = Correlation coeff.

$$= 1 + 2 \sum_{k=1}^{\infty} Corr(X_0, X_k).$$

If $\{X_i\}$ were i.i.d, then $\tau = 1$. In general,

$$Var(\bar{X}_{\tau N}) \approx \frac{V}{\tau N} \approx \frac{\sigma^2}{N}.$$

Thus τN observations from $\{X_i\}$ gives approximately the same variance (hence the same size of confidence interval) that we would get if we had been able to generate N i.i.d. samples.

Now, choose some L_1 for L as our first guess. (e.g., using Method (1)). Set $\hat{\tau}_{N,L_1} = \frac{\hat{V}_{N,L_1}}{\hat{C}(0)}$. Then, choose $L \geq 5\hat{\tau}_{N,L_1}$ and $L \geq L_1$.

The Method of Batch Means for Determining Confidence Intervals

Again, $X_1, X_2, \cdots$ is a stationary process. Divide $X_1, \cdots, X_N$ ($N >> 1$) into non overlapping subsections of equal length. Each of such subsections is called a batch. Let b be the common batch length. So, each batch has N/b consecutive X_i's. (Assume N is an integer multiple of b.) Let $L = N/b$. So, the kth batch is the subsequence/subsection

$$X_{(k-1)L+1}, \cdots, X_{kL}. \quad (k = 1, \cdots, b).$$

Define for each $k \in \{1, \cdots, b\}$,

$$Y_k = \frac{1}{L} \sum_{i=(k-1)L+1}^{kL} X_i.$$

If L is sufficiently large, then:

(1) $Y_1, \cdots, Y_b \sim N(\mu, V/L)$ approximately;

(2) $Y_1, \cdots, Y_b$ are approximately independent.

Now, using the classical statistics applied to $Y_1, \cdots, Y_b$, we obtain a $\gamma \in (0,1)$ confidence interval for μ:

$$\left[\bar{Y}_b - r\sqrt{S_b^2/b}, \ \bar{Y} + r\sqrt{S_b^2/b} \right],$$

where $\bar{Y}_b = \frac{1}{b} \sum_{i=1}^{b} Y_i$, $r = \Phi^{-1}\left(\frac{\gamma+1}{2} \right)$, $\Phi(x)$ is the CDF of $N(0,1)$, i.e., $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$, and

$$S_b^2 = \frac{1}{b-1} \sum_{i=1}^{b} (Y_i - \bar{Y}_b)^2.$$

<u>The Regenerative Method</u> Let $X_k = h(W_k)$ $(k = 1, 2, \cdots)$ be a stationary process, where $h : \mathbb{R} \to \mathbb{R}$ is a continuous function, and W_k $(k = 1, 2, \cdots)$ is a Markov chain on some state space S with stationary/invariant distribution π.

Suppose $\exists$ a state $I \in S$ with $\pi(I)$ not small, i.e., the chain $\{W_k\}_{k=1}^{\infty}$ visits I often. Suppose also that we know exactly when

$W_k = I$, e.g., this holds if $\exists$ no other state J s.t. $h(I) = h(J)$.
The idea of regeneration is to break $\{X_k\}$ into segments beginning and ending with consecutive visits to I. The Markov property implies that these segments are mutually independent, since the process starts fresh ("regenerates") with each visit to I.

Formally, assume $W_1 = I$. (Otherwise, run the chain until the first visit to I, and discard everything before this time.) Let $\sigma_0 = 0$. For each $k \geq 1$, let

$$\sigma_k = \min\{t > \sigma_{k-1} : W_t = I\}$$
$$= \text{the time of the } k\text{th visit to } I.$$

The "kth segment" is the part of process from $\sigma_{k-1}+1$ to σ_k (inclusive).

We want to estimate $\mathbb{E}(X_k) = \mathbb{E}_\pi(h(W_k))$.
Set

$$D_k = \sigma_k - \sigma_{k-1} = \text{duration of } k\text{th segment}.$$

The "regenerative" property implies that $D_1, D_2, \ldots$ are i.i.d. Next, set

$$H_k = \sum_{i=\sigma_{k-1}+1}^{\sigma_k} h(W_i).$$

Then $H_1, H_2, \ldots$ are also i.i.d. For simplicity, assume $N = \sigma_m$ for some m. Then

$$N = D_1 + \cdots + D_m, \quad \sum_{i=1}^{N} X_i = H_1 + \cdots + H_m$$

By the Strong Law of Large Numbers, we obtain

$$\mathbb{E}(X_1) = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$= \lim_{m\to\infty} \frac{\frac{1}{m}(H_1 + \cdots + H_m)}{\frac{1}{m}(D_1 + \cdots + D_m)}$$

$$= \frac{\mathbb{E}(H_1)}{\mathbb{E}(D_1)} \quad \text{with probability 1.}$$

Therefore, we can use the estimator

$$\hat{\mu}_m = \frac{\overline{H}_m}{\overline{D}_m} = \frac{\frac{1}{m}\sum_{i=1}^{m} H_i}{\frac{1}{m}\sum_{i=1}^{m} D_i}$$

for $\mu = \mathbb{E}(X_k)$ $(\forall k \geq 0)$.

Note that this is in general biased as

$$\mathbb{E}(\hat{\mu}_m) \neq \mu$$

Error analysis. By the Central Limit Theorem,

$$\overline{H}_m := \frac{1}{m}\sum_{i=1}^{m} H_i = \mathbb{E}(H_1) + \varepsilon_m,$$

$$\overline{D}_m := \frac{1}{m}\sum_{i=1}^{m} D_i = \mathbb{E}(D_1) + \delta_m,$$

where $(\varepsilon_m, \delta_m)$ is approximately jointly normally distributed. Taylor's expansion leads to

$$\hat{\mu}_m = \frac{\overline{H}_m}{\overline{D}_m} \approx \mathbb{E}(X_1) + \frac{\varepsilon_m}{\mathbb{E}(D_1)} - \delta_m \frac{\mathbb{E}(H_1)}{[\mathbb{E}(D_1)]^2},$$

leading to

$$\mathrm{Var}(\hat{\mu}_m) \approx \frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{H_i}{\overline{D}_m} - D_i \frac{\overline{H}_m}{(\overline{D}_m)^2}\right)^2.$$

Moreover, $\hat{\mu}_m - \mu = O(\frac{1}{m})$.

(1) Asymptotically Stationary Output

Let $\{X_k\}_{k=0}^{\infty}$ be a Markov chain with invariant distribution π. We wish to estimate

$$\mu = \lim_{k\to\infty} \mathbb{E}(X_k)$$

from the simulated samples $X_0, X_1, \cdots$.

The idea ~~t~~ is to choose T large enough so that X_T is close to equilibrium, and discard $X_0, \cdots, X_{T-1}$, and use the previous method for $X_T, X_{T+1}, \cdots$. The period from $k=0$ to $k=T$ is a period of equilibration, or the "burn-in" period. A short such period may result an "initialization bias".

There are no general rules for selecting T. Also, methods of selecting T can be case dependent.

A simple and quick method is to divide the output into b batches (e.g., $20 \le b \le 50$), and plot the batch means. If the mean of the first batch is significantly larger or smaller than all others, then discard the first batch. Otherwise stop. Repeat the next batch.

(E) Empirical CDF

Suppose Monte Carlo simulations produce random variables $X_1, X_2, \cdots$ i.i.d. with the (exact) CDF $F(x)$. (Here, $X_j \in \mathbb{R}$ $\forall j \geq 0$). $F(x)$ is not known. But we wish to use $\{X_k\}_{k=1}^{\infty}$ to find approximations of $F(x)$.

Define
$$\hat{F}_N(x) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}\{X_k \leq x\}$$
$$= \frac{1}{N} \left| \{k : X_k \leq x\} \right| \quad \forall x \in \mathbb{R}.$$

Call $\hat{F}_N(x)$ a <u>random empirical CDF</u>. Let $U_k = F(X_k)$ $(k = 1, 2, \cdots)$. Then $U_k \sim U[0,1]$ and $U_1, \cdots U_n, \cdots$ are i.i.d. Here, we assume F is continuous and strictly increasing. If we denote $u = F(x)$ and $x = F^{-1}(u)$, then
$$\hat{F}_N(x) - F(x) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}\{X_i \leq x\} - F(x)$$
$$= \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}\{U_k \leq u\} - u$$
$$= \hat{G}_N(u) - u,$$

where
$$\hat{G}_N(u) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{1}\{U_k \leq u\},$$
is ~~the~~ called the <u>reduced empirical CDF</u>.

Define

$$D_N = \sup_{x \in \mathbb{R}} \left| \hat{F}_N(x) - F(x) \right|$$
$$= \sup_{0 \le u \le 1} \left| \hat{G}_N(u) - u \right|.$$

This is called a <u>Kolmogorove statistic</u> of the data. Note that the distribution of D_N does not depend on F (by the second equality).

Here are some properties:

- If $X_{(1)} < \cdots < X_{(N)}$ then
$$\hat{F}_N(X_{(k)}) = \frac{k}{N}, \quad k = 1, \cdots, N.$$
- Binomial distribution:
$$N\hat{F}_N(x) \sim \text{Binomial}(N, F(x)),$$
$$N\hat{G}_N(u) \sim \text{Binomial}(N, u).$$
- Glivenko-Cantelli:
$D_N \xrightarrow{a.s.} 0.$ and hence $\hat{F}_N(x) \xrightarrow{a.e} F(x)$ uniformly in x.
- Central Limit Theorem:
$$\sqrt{N}\left[\hat{F}_N(x) - F(x)\right] \xrightarrow{dist.} Z \sim \mathcal{N}(0, F(x)(1-F(x))).$$
- Conditional ~~Process~~ Poisson: The probability distribution of the reduced empirical cDF ~~$\{\hat{G}_N(u): 0 \le u \le 1\}$, viewed as~~ $\{\hat{G}_N(u): 0 \le u \le 1\}$, viewed as a stochastic process on $[0,1]$, is the same as the conditional distribution of a Poisson process $\{M_u : 0 \le u \le 1\}$ with rate $1/N$ given that $M_1 = N$.

- Brownian bridge: The stochastic process $\{\sqrt{N}[\hat{G}_N(u) - u],\ 0 \le u \le 1\}$ converges in distribution to a Brownian bridge process on $[0,1]$.
- The Kolmogorov distribution:

$$\lim_{N \to \infty} \mathbb{P}(\sqrt{N} D_N \le x) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2(kx)^2}, \quad x > 0.$$

- The confidence interval: An approximate $1-\alpha$ confidence interval for $F(x)$ is

$$\left(\hat{F}_N(x) - z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{N}} \sqrt{\hat{F}_N(x)(1-\hat{F}_N(x))},\ \hat{F}_N(x) - z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{N}} \sqrt{\hat{F}_N(x)(1-\hat{F}_N(x))} \right),$$

where z_γ is such that $1 - \Phi(z_\gamma) = \gamma$ with $\Phi(x)$ is CDF of a random variable $\sim N(0,1)$.

Equivalently, an approximate $1-\alpha$ confidence interval for $F(X_{(k)})$ is

$$\left(\frac{k}{N} - z_{1-\frac{\alpha}{2}} \frac{1}{N} \sqrt{k(1-k/N)},\ \frac{k}{N} - z_{1-\frac{\alpha}{2}} \frac{1}{N} \sqrt{k(1-k/N)} \right).$$

(F) Kernel Density Estimation

This method is for estimating a probability density from simulated data.

Let $X_1, \cdots, X_N$ be independent realizations from an unknown continuous PDF f on some $S \subseteq \mathbb{R}$. Let $K = K(x)$ be a PDF of some random variable on $\mathbb{R}$ and assume it is symmetric $K(-x) = K(x)$ $\forall x \in \mathbb{R}$ (i.e., $K = K(x)$ is an even function. Let $h > 0$. We define

$$\hat{f}_{N,h}(x) = \frac{1}{Nh} \sum_{k=1}^{N} K\left(\frac{x - X_k}{h}\right) \quad \forall x \in R$$

We call $\hat{f}_{N,h}$ a <u>kernel density estimator</u> of f. with the <u>bandwidth</u> $h > 0$. Here, $K(x)$ is called a <u>kernel function</u>

<u>Example</u> The Gaussian kernel density estimator.

$$\hat{f}_{N,h}(x) = \frac{1}{Nh} \sum_{i=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-(x - X_i)^2/(2h^2)} \quad (\forall x \in \mathbb{R}).$$

The function K here is given by

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \forall x \in \mathbb{R}.$$

This is the PDF of a standard normally distributed random variable.

We define the <u>mean integrated squared error</u> (MISE) of a kernel density estimator $\hat{f}_{N,h}$ by

$$\mathrm{MISE}_N(h) = \mathbb{E}_f \int_{-\infty}^{\infty} |\hat{f}_{N,h}(x) - f(x)|^2 dx.$$

We have

$$\mathrm{MISE}_N(h) = \underbrace{\int_{-\infty}^{\infty} \left|\mathbb{E}_f[\hat{f}_{N,h}(x)] - f(x)\right|^2 dx}_{\text{pointwise bias of } \hat{f}_{N,h}} + \underbrace{\int_{-\infty}^{\infty} \mathrm{Var}_f(\hat{f}_{N,h}(x)) dx}_{\text{pointwise variance of } \hat{f}}.$$

We can also use an alternative error criterion in the <u>expected L^1 error</u>:

$$\mathbb{E}_f \int_{-\infty}^{\infty} |\hat{f}_{N,h}(x) - f(x)| dx$$

How to choose a good bandwidth h?

<u>Method 1</u> The Gaussian rule of thumb.

A first-order asymptotic approximation of the MISE of the Gaussian kernel density estimator is

$$\frac{1}{4} h^4 \|f''\|^2_{L^2(\mathbb{R})} + \frac{1}{2Nh\sqrt{\pi}}, \quad N \gg 1.$$

(Here, we assume $f'' \in L^2(\mathbb{R})$.) An asymptotic optimal value of h is then

$$h^* = \left(2N\sqrt{\pi}\|f''\|^2_{L^2(\mathbb{R})}\right)^{-\frac{1}{5}}, \quad N \gg 1.$$

The optimal asymptotic rate of decay of the MISE is

$$MISE_N(h^*) = \frac{5\, \|f''\|_{L^2(\mathbb{R})}^{4/5}}{4^{7/5}\pi^{2/5}} N^{-4/5} + o(N^{-4/5})$$
as $N \to \infty$.

To compute h^*, one needs to estimate $\|f''\|^2_{L^2(\mathbb{R})}$. The Gaussian rule of thumb is to assume that f is the density of the $N(\hat{\mu}, \hat{\sigma}^2)$ distribution, where $\hat{\mu}$ and $\hat{\sigma}^2$ are the sample mean and the sample variance of the data, respectively. In this case,

$$\|f''\|^2_{L^2(\mathbb{R})} = \frac{3}{8\sqrt{\pi}\,\hat{\sigma}^5},$$

and the asymptotic optimal value is

$$h^* := h_{rot} = \left(\frac{4\hat{\sigma}^5}{3N}\right)^{1/5} \approx \sqrt{1.12}\,\hat{\sigma} N^{-1/5},$$
for $N \gg 1$.

Method 2 The Least-Squares Cross Validation

We again work on the selection of an optimal bandwidth for the Gaussian kernel density estimator

$$\hat{f}_{N,h}(x) = \frac{1}{Nh}\sum_{i=1}^{N} \kappa\left(\frac{x - X_i}{h}\right), \quad \forall x \in \mathbb{R},$$

where $\kappa(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

The Least-Squares Cross Validation (LSCV) method defines the optimal bandwidth as an unbiased estimator of the global minimizer of the integrated squared error (ISE), given by

$$ISE_N(h) = \int_{-\infty}^{\infty} |\hat{f}_{N,h}(x) - f(x)|^2 dx.$$

This is a random variable, depending on the particular data. Minimizing $ISE_N(h)$ is equivalent to minimizing

$$\int_{-\infty}^{\infty} |\hat{f}_{N,h}(x)|^2 dx - 2\,\mathbb{E}_f\left(\hat{f}_{N,h}(X)\right).$$

The term $\mathbb{E}_f(\hat{f}_{N,h}(X))$ can be estimated without bias via <u>the cross-validation estimator</u>:

$$\frac{1}{N}\sum_{i=1}^{N} \hat{f}_{N,h}^{(-i)}(X_i),$$

where $\hat{f}_{N,h}^{(-i)}(x)$ is the Gaussian kernel density estimator based on the data points $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_N$. We have

$$\frac{1}{N}\sum_{i=1}^{N} \hat{f}_{N,h}^{(-i)}(X_i) = \frac{1}{hN(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1\\ j\neq i}}^{N} K\left(\frac{X_i - X_j}{h}\right).$$

Let us introduce

$$K_h(x) = \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{x^2}{2h^2}} = \frac{1}{h} K\left(\frac{x}{h}\right), \quad \forall x \in \mathbb{R}.$$

K_h is the PDF for a random variable $Z \sim \mathcal{N}(0, h^2)$.

The cross-validation estimator of $\mathbb{E}_f(\hat{f}_{N,h}(X))$ can be written now as

$$\frac{1}{N}\sum_{i=1}^{N} \hat{f}_{N,h}^{(-i)}(X_i) = \frac{1}{N(N-1)} \cancel{\sum_{i=1}^{N}\sum_{j}} \sum_{\substack{i,j=1\\ i\neq j}}^{N} K_h(X_i - X_j).$$

Note that each X_i is a random variable. Thus,

$$\mathbb{E}_f\left(\frac{1}{N}\sum_{i=1}^{N}\hat{f}_{N,h}^{(-i)}(X_i)\right)$$
$$= \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty} f(x)\,\hat{f}_{N,h}^{(-i)}(x)\,dx$$
$$= \frac{1}{N(N-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{N}\int_{-\infty}^{\infty} f(x)\,K_h(x-X_j)\,dx$$
$$= \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty} f(x)\,K_h(x-X_j)\,dx.$$

But $\mathbb{E}_f\left(\hat{f}_{N,h}(X)\right) = \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty} f(x)\,K_h(x-X_j)\,dx$

Thus, the cross-validation estimator is unbiased.

Now, we calculate the integral of $|\hat{f}_{N,h}(x)|^2$.

$$\int_{-\infty}^{\infty} |\hat{f}_{N,h}(x)|^2\,dx = \frac{1}{N^2\,2\pi h^2}\sum_{i,j=1}^{N}\int_{-\infty}^{\infty} e^{-\frac{(x-X_i)^2+(x-X_j)^2}{2h^2}}\,dx$$

For fixed X_i, X_j:

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2h^2}\left[(x-X_i)^2+(x-X_j)^2\right]}\,dx$$
$$= \int_{-\infty}^{\infty} e^{-\frac{2}{2h^2}\left(x-\frac{X_i+X_j}{2}\right)^2} e^{-\frac{1}{4h^2}(X_i-X_j)^2}\,dx$$
$$= e^{-\frac{1}{4h^2}(X_i-X_j)^2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\,\frac{h}{\sqrt{2}}\,dy$$
$$= \frac{\sqrt{2\pi}\,h}{\sqrt{2}}\,e^{-\frac{1}{4h^2}(X_i-X_j)^2}.$$

Hence, $\displaystyle\int_{-\infty}^{\infty} |\hat{f}_{N,h}(x)|^2\,dx = \frac{\sqrt{2}}{N^2}\sum_{i,j=1}^{N} K_{2h}(X_i-X_j).$

Define now the LSCV bandwidth h_{LS} to be

$$h_{LS} = \underset{h>0}{\arg\min}\; g(h),$$

where

$$g(h) = \frac{\sqrt{2}}{N^2}\sum_{i,j=1}^{N} K_{2h}(X_i - X_j) - \frac{2}{N(N-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{N} K_h(X_i - X_j).$$

Note that the double-sum can be inefficient in practical implementation.

Back to

$$\hat{f}_{N,h}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right) = \frac{1}{N}\sum_{i=1}^{N} K_h(x - X_i)$$

with $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, and $K_h(x) = \frac{1}{\sqrt{2\pi h^2}} e^{-x^2/(2h^2)}$.

Let

$$K(x,t) = \frac{1}{\sqrt{4\pi t}} e^{-\frac{x^2}{4t}} \qquad (t>0,\ x\in\mathbb{R}).$$

This is the one-dimensional heat kernel. It satisfies

$$\partial_t K = \partial_{xx} K \qquad (t>0,\ x\in\mathbb{R})$$

$\forall \phi \in C_c(\mathbb{R})$: $$\lim_{t\to 0^+}\int_{-\infty}^{\infty} K(x-y,t)\,\phi(y)\,dy = \phi(x) \qquad \forall x\in\mathbb{R}.$$

Let $t = h^2$. Then,

$$\hat{f}_{N,h}(x) = f_N(x,t) = \frac{\sqrt{2}}{N}\sum_{i=1}^{N} K(x - X_i, t)$$

Thus,

$$\begin{cases} \dfrac{\partial f_N}{\partial t} = \dfrac{\partial^2 f_N}{\partial x^2} & (t>0,\ x\in\mathbb{R}). \\ f_N(x,0) = \dfrac{\sqrt{2}}{N}\displaystyle\sum_{i=1}^{N}\delta_{X_i}(x) & \end{cases}$$

where $\delta_{X_i}(x)$ is the Dirac measure at X_i.

This provides a method of calculating $\hat{f}_{N,h}$.

Method 3 Plug-in Bandwidth Selection

This method provides an estimator for $\|f''\|^2 = \|f''\|^2_{L^2(\mathbb{R})}$ in determining the asymptotically optimal bandwidth h^* using the Gaussian kernel density estimation:

$$h^* = \left(2\sqrt{\pi}\, N\, \|f''\|^2\right)^{-1/5}.$$

The estimator of $\|f''\|^2$ is $\|\hat{f}''_{N,h_2}\|^2$ for some $h_2 > 0$, where

$$\hat{f}_{N,h}(x) = \frac{1}{N}\sum_{k=1}^{N} K_h(x - X_k),$$

is the Gaussian kernel density estimator.

In general, we have for $h_k > 0$

$$\|\hat{f}^{(k)}_{N,h_k}\|^2 = \int_{-\infty}^{\infty}\left|\frac{1}{N}\sum_{i=1}^{N} K^{(k)}_{h_k}(x - X_i)\right|^2 dx$$

$$= \frac{1}{N^2}\sum_{i,j=1}^{N}\int_{-\infty}^{\infty} K^{(k)}_{h_k}(x - X_i)\, K^{(k)}_{h_k}(x - X_j)\, dx$$

$$= \frac{(-1)^k\sqrt{2}}{N^2}\sum_{i,j=1}^{N} K^{(2k)}_{2h_k}(X_i - X_j),$$

where $g^{(k)}$ is the kth order derivative of g.

To use $\|\hat{f}^{(2)}_{N,h_k}\|^2$ as an estimator for $\|f''\|^2$, we need to choose h_2. A good choice ~~is~~ for h_k is

$$h_k = \left(\frac{1 + 2^{-j-\frac{1}{2}}}{3}\cdot\frac{(2k-1)!}{N\sqrt{\pi/2}\,\|\hat{f}^{(k+1)}_{N,h_{k+1}}\|^2}\right)^{2/(3+2k)}.$$

To compute $\|\hat{f}^{(k+1)}_{N,h_{k+1}}\|^2$, we need to know h_{k+1}, which in turn requires the estimate h_{k+2}, and so on.

The idea is now to fix ℓ, say $\ell \geq 3$. First compute $\|\hat{f}^{(\ell+2)}_{N,h_{\ell+2}}\|^2$ by assuming f is the normal PDF with the mean and variance estimated from the i.i.d. samples $X_1, \cdots, X_N$. Then, compute $h_{\ell+1}$ using the above formula, which uses $\|\hat{f}^{(\ell+2)}_{N,h_{\ell+2}}\|^2$. Then, compute $\|\hat{f}^{(\ell+1)}_{N,h_{\ell+1}}\|^2$, and then h_ℓ, etc.