

ESTIMATION OF DETERMINISTIC PARAMETERS: LARGE NUMBERS LAW AND SAMPLE ESTIMATORS



Fulvio GINI

Dipartimento di Ingegneria dell'Informazione

University of Pisa

Via G. Caruso 16, I-56122, Pisa, Italy

fulvio.gini@unipi.it





The Estimation Problem and Properties of an Estimator

■ **Problem:** Given the observed signal $x(t)$ for $t \in [0, T)$, that is a realization of the random process $X(t)$, we want to determine the value of a parameter θ that is not directly observable but is related in same way to $X(t)$, i.e. all the process realizations are affected by the value of θ , so also the particular realization we observe in $[0, T)$.

Observed signal: $X(t; \theta) \quad t \in [0, T)$

■ **Mathematical model of the data:** the first step is to derive a mathematical model of the data that models the dependence of the observed signal on the unknown parameter to be estimated.

■ **Discrete representation of the observed signal:** to derive the mathematical model we are looking for, first we have to represent the continuous-time observed signal in discrete form (basis expansion).

■ Given the characteristics of the observed signal, we first determine a **basis** Ψ that allows to represent all possible realizations of process $X(t)$, then we derive the corresponding image vector \mathbf{X} .

■ Discrete representation of the observed signal:

Observed signal: $X(t; \theta) \quad t \in [0, T) \xLeftrightarrow{\Psi} \mathbf{X}(\theta) \quad \text{image vector}$

■ **The data model:** being $X(t)$ a random process, the image vector \mathbf{X} is a random vector, so its statistical behavior is completely characterized by its joint probability density function (pdf):

Observed data in discrete form: $\mathbf{X}(\theta) \Rightarrow f_{\mathbf{X}}(\mathbf{x}; \theta)$

■ **The estimation problem:** given the observed data vector \mathbf{x} , determine a measure (as much accurate as possible) of the unknown parameter θ :

$\hat{\theta} = g(\mathbf{x}) \rightarrow$ The estimator is a function of the data vector \mathbf{x}
 $\rightarrow \hat{\theta}$ is a random variable (r.v.)

The Estimation Problem

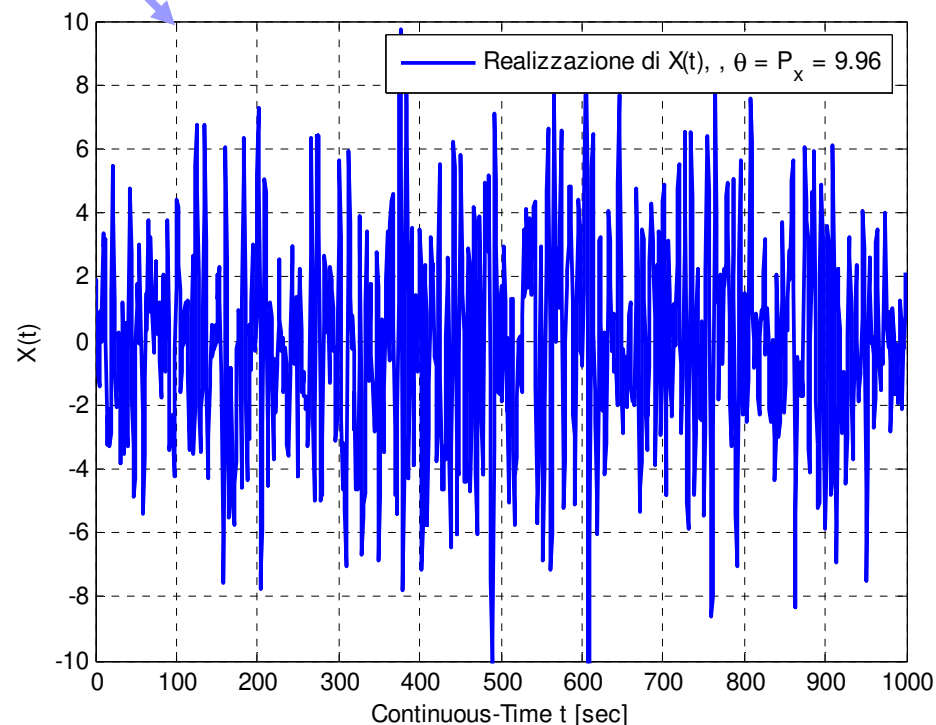
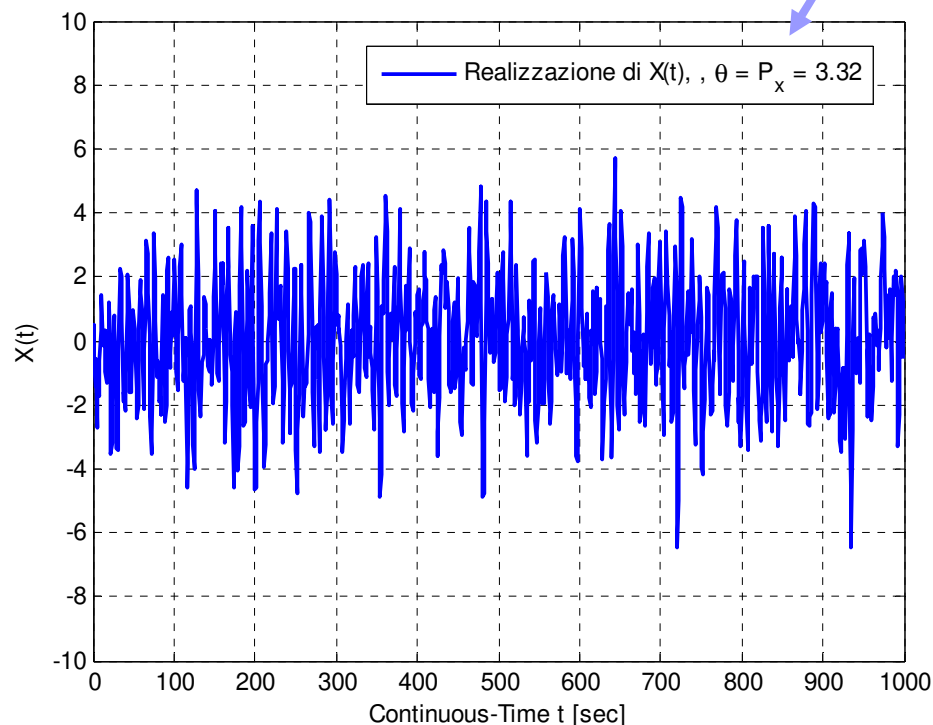
5

■ *Example:* realizations of a Gaussian process $X(t)$ where θ is the power of $X(t)$.

Observed signal: $X(t; \theta) \quad t \in [0, T) \quad \theta = P_x = E\{X^2(t)\}$

$$\theta = P_x = 3.32$$

$$\theta = P_x = 9.96$$



■ Discrete representation of the observed signal:

Observed signal: $X(t; \theta) \quad t \in [0, T) \xleftrightarrow{\Psi} \mathbf{X}(\theta)$ image vector

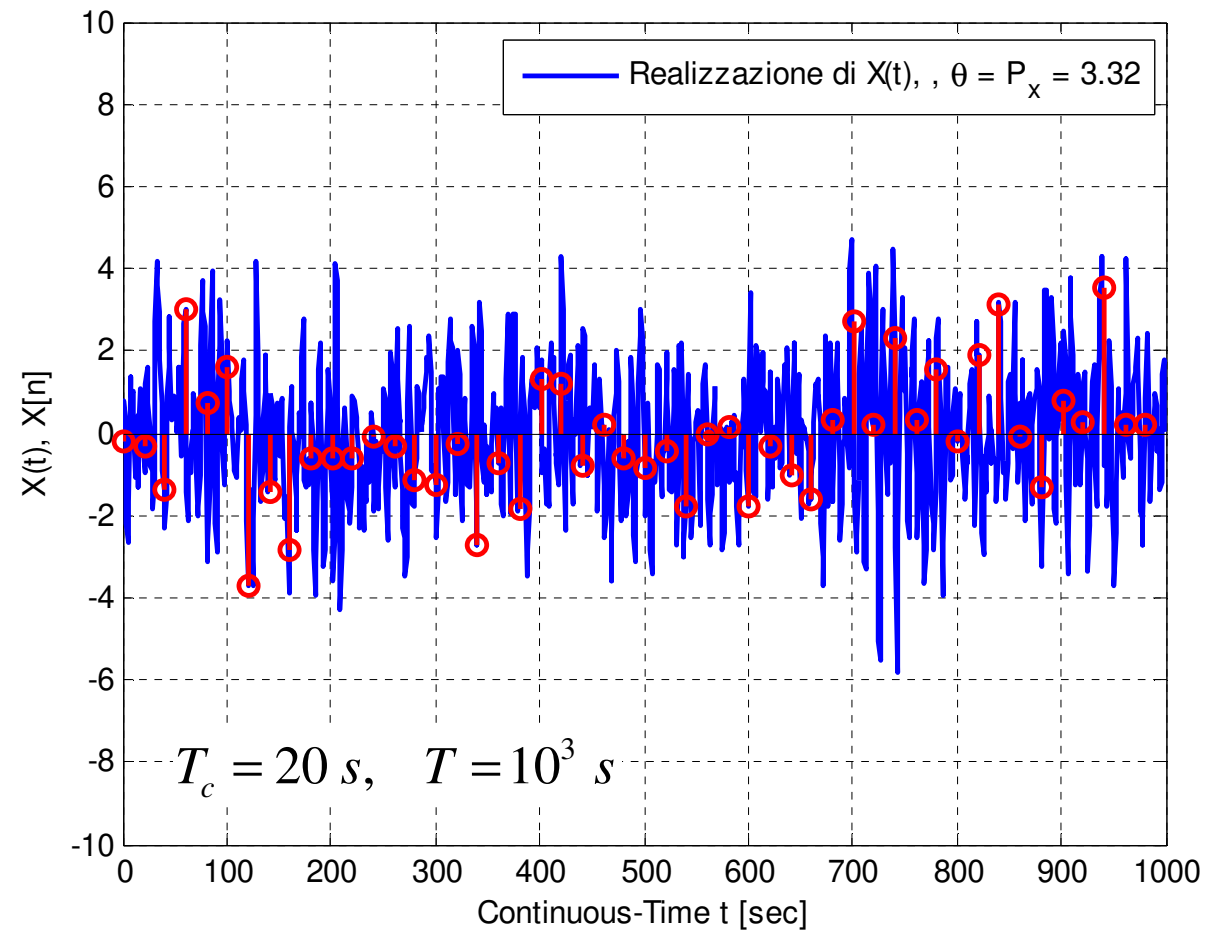
■ If the process $X(t)$ has rigorously finite bandwidth B or it has essential bandwidth B at level α , with $N=2BT \gg 1$ such that $\alpha \ll 1$, we can represent $X(t)$ in discrete form by using the orthonormal basis of dimension $N=2BT$ of *sinc* functions, i.e. the continuous-time observed signal is discretized by collecting $N=2BT$ samples at time intervals multiples of $T_c=1/2B$ within the observation interval $[0, T)$:

$$\begin{aligned} \mathbf{X} &= [X[0] \quad X[1] \quad \cdots \quad X[N-1]]^T \\ &= \frac{1}{\sqrt{2B}} [X(0) \quad X(T_c) \quad \cdots \quad X((N-1)T_c)]^T \end{aligned}$$

$$\text{Dimension of the image vector: } N = \frac{T}{T_c} = 2BT$$

The Estimation Problem

7



$$\mathbf{X} = \frac{1}{\sqrt{2B}} \begin{bmatrix} X(0) & X(T_c) & \cdots & X((N-1)T_c) \end{bmatrix}^T, \quad N = 2BT = 50$$

- **Estimator:** $\hat{\theta} = g(\mathbf{x})$ is a function of the observed data vector \mathbf{X} , hence it is a random variable.
- The full statistical characterization of the estimator is provided by its pdf.
- If we know the pdf of the data vector, we can derive the pdf of the estimator by applying the **fundamental theorem** for the transformation of random variables or the method of the cumulative distribution function:

$$\mathbf{X}(\theta) \Rightarrow f_{\mathbf{X}}(\mathbf{x}; \theta) \Rightarrow f_{\hat{\theta}}(\hat{\theta}; \theta)$$

- The statistical behavior of the estimator depends on the entire pdf. However, it is common practice to use as **quality index** of an estimator a specific moment of the pdf, that is called **mean square error (MSE)** and represents the statistical power of the estimation error.
- Also, the mean value of the estimation error (**bias**) plays an important role.

■ **Estimation error:** $\varepsilon(\mathbf{x}) \triangleq \theta - \hat{\theta}(\mathbf{x}) \rightarrow \varepsilon$ is a random variable

■ **Mean Square Error (MSE), variance, and bias** of the estimate of a parameter θ are mutually related. Assume that the parameter is **deterministic** (we will analyze the case of random parameters separately):

$$MSE\{\hat{\theta}\} \triangleq E\left\{\left|\theta - \hat{\theta}\right|^2\right\}, \quad \text{var}\{\hat{\theta}\} \triangleq E\left\{\left|\hat{\theta} - E\{\hat{\theta}\}\right|^2\right\}, \quad b\{\hat{\theta}\} \triangleq E\{\theta - \hat{\theta}\}$$



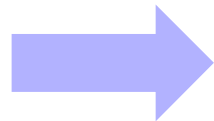
$$MSE\{\hat{\theta}\} = E\left\{\left|\theta - \hat{\theta}\right|^2\right\} = \int_{-\infty}^{+\infty} \left|\theta - \hat{\theta}\right|^2 f_{\hat{\theta}}(\hat{\theta}; \theta) d\hat{\theta} = \iiint \left|\theta - \hat{\theta}(\mathbf{x})\right|^2 f_{\mathbf{x}}(\mathbf{x}; \theta) d\mathbf{x}$$

■ Note that in general the MSE depends on the value of θ .

- **Bias** and **Mean Square Error (MSE)** of an estimator are the first and second order moments of the estimation error, which is a random variable:

Estimation error:

$$\varepsilon(\mathbf{x}) \triangleq \theta - \hat{\theta}(\mathbf{x})$$



$$\left\{ \begin{array}{l} b\{\hat{\theta}\} \triangleq E\{\varepsilon\} = E\{\theta - \hat{\theta}\} = \theta - E\{\hat{\theta}\} \\ MSE\{\hat{\theta}\} \triangleq E\{|\varepsilon|^2\} = E\{|\theta - \hat{\theta}|^2\} \end{array} \right.$$

- **MSE**, **variance**, and **bias** of the estimate of θ are related as follows:

$$\begin{aligned} MSE\{\hat{\theta}\} &= E\{|\theta - \hat{\theta}|^2\} = E\{|\theta - E\{\hat{\theta}\} + E\{\hat{\theta}\} - \hat{\theta}|^2\} \\ &= E\{|\theta - E\{\hat{\theta}\}|^2\} + E\{|E\{\hat{\theta}\} - \hat{\theta}|^2\} + 2E\{(\theta - E\{\hat{\theta}\})(E\{\hat{\theta}\} - \hat{\theta})\} \\ &= |\theta - E\{\hat{\theta}\}|^2 + E\{|E\{\hat{\theta}\} - \hat{\theta}|^2\} = |b\{\hat{\theta}\}|^2 + \text{var}\{\hat{\theta}\} \end{aligned}$$

$$MSE\{\hat{\theta}\} = \left|b\{\hat{\theta}\}\right|^2 + \text{var}\{\hat{\theta}\}$$

- The MSE (statistical power of the estimation error) can be reduced by reducing the bias (mean value of the estimation error) and/or by reducing the error variance (the variance is a measure of dispersion of a r.v. around its mean value).
- An estimator that is unbiased and has the minimum variance is called **Minimum Variance Unbiased (MVU)** estimator.
- An estimator of θ that has MSE lower than another estimator of θ is said to be more **efficient** than the other.
- **Questions:** Does it exist an estimator that is the **most efficient**, i.e. an estimator having the **lowest MSE**? The answer is NO if we are talking about deterministic parameters. Whereas, the answer is YES for random parameters.

- Any estimator of the unknown parameter θ is a function of the observed data vector \mathbf{x} and so also of the data size N . If we want to stress this dependency we can add the subscript N :

$$\hat{\theta}_N \equiv \hat{\theta}_N(\mathbf{x})$$

- An estimator is **unbiased** if the mean value of the estimator is equal to the true value of the parameter, i.e. if the error is zero mean.
- If instead the mean value of the error is different from zero, it is called **bias** and the estimator is **biased**.

$$\text{Bias: } b\{\hat{\theta}_N(\mathbf{x})\} \triangleq E\{\varepsilon_N(\mathbf{x})\} = E\{\theta - \hat{\theta}_N(\mathbf{x})\} = \theta - E\{\hat{\theta}_N(\mathbf{x})\}$$

$$\text{Unbiased estimator: } b\{\hat{\theta}_N(\mathbf{x})\} = 0 \rightarrow E\{\hat{\theta}_N(\mathbf{x})\} = \theta$$

$$\text{Biased estimator: } b\{\hat{\theta}_N(\mathbf{x})\} \neq 0 \rightarrow E\{\hat{\theta}_N(\mathbf{x})\} \neq \theta$$

- An estimator is **asymptotically unbiased** if the bias goes to zero when the data size N grows to infinity:

$$\lim_{N \rightarrow \infty} b\{\hat{\theta}_N(\mathbf{x})\} = 0 \quad \Leftrightarrow \quad \lim_{N \rightarrow \infty} E\{\hat{\theta}_N(\mathbf{x})\} = \theta$$

- An estimator is **consistent** if the estimate converges **in probability** to the true value of the parameter when the data size N grows to infinity:

$$\lim_{N \rightarrow \infty} \Pr\left\{\left|\hat{\theta}_N(\mathbf{x}) - \theta\right| \geq \varepsilon\right\} = 0 \quad \forall \varepsilon > 0$$

- Usually, it is not easy to verify the **convergence in probability** of an estimator. It is easier to verify the **mean square convergence**: this kind of convergence is guaranteed when the estimator is asymptotically unbiased and the estimate variance goes to zero when the data size N grows to infinity.

■ An estimator converges in **mean square sense (m.s.s.)** to the true value of the parameter if:

$$\lim_{N \rightarrow \infty} E \left\{ \left| \hat{\theta}_N(\mathbf{x}) - \theta \right|^2 \right\} = \lim_{N \rightarrow \infty} MSE \left\{ \hat{\theta}_N(\mathbf{x}) \right\} = 0$$

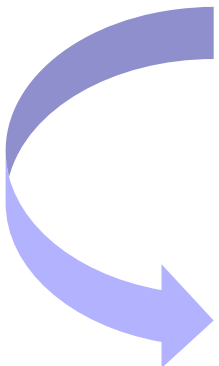
■ Now, since it holds true that:

$$E \left\{ \left| \hat{\theta}_N(\mathbf{x}) - \theta \right|^2 \right\} = MSE \left\{ \hat{\theta}_N(\mathbf{x}) \right\} = \left| b \left\{ \hat{\theta}_N(\mathbf{x}) \right\} \right|^2 + \text{var} \left\{ \hat{\theta}_N(\mathbf{x}) \right\}$$

→ **convergence in mean square sense** is guaranteed when the estimator is asymptotically unbiased and the estimation variance goes to zero when the data size N grows to infinity, as previously stated.

■ If an estimator converges in mean square sense, it also converges in probability. The reverse is not true, i.e. convergence in probability does not imply mean square convergence.

■ *Proof.* If we apply the **Bienaymé-Tchebycheff inequality**:


$$\begin{aligned} X &\equiv \hat{\theta}_N \\ a &\equiv \theta \end{aligned} \quad \Pr\{|X - a| \geq \varepsilon\} \leq \frac{E\{|X - a|^2\}}{\varepsilon^2}$$
$$\lim_{N \rightarrow \infty} \Pr\{|\hat{\theta}_N - \theta| \geq \varepsilon\} \leq \frac{\lim_{N \rightarrow \infty} E\{|\hat{\theta}_N - \theta|^2\}}{\varepsilon^2} = 0$$

■ Since the limit on the left-hand side cannot be negative, it must necessarily be zero, so mean square convergence implies convergence in probability.



Taxonomy of the Estimation Problems

- The first big distinction is between **deterministic parameters** and **random parameters**:

Deterministic parameters:

- probability of an event $p = \Pr\{A\}$
- moments of a distribution (mean value, variance, skewness, kurtosis, *etc.*)
- parameters of a distribution
- parameters of a signal (amplitude, phase, frequency, time of arrival, *etc.*)
- moments of a random vector (mean vector, correlation matrix, *etc.*)
- moments of a random signal (mean value function, ACF, PSD, *etc.*)

ACF=Autocorrelation Function, PSD=Power Spectral Density

Random parameters:

- parameters of a signal of which we known the a-priori pdf
- random signals (values of a realization of a random process)

- According to the class to which the parameter/signal belongs to, different methods have been developed. We will investigate the principal ones:

Deterministic parameters:

- { Sample estimates (application of the Law of large numbers)
- { Method of Moments (MM)
- { Maximum Likelihood (ML) method
- { Least Squares (LS): Linear LS or Non Linear LS (NLLS)

Random parameters:

- { Bayesian estimators (MMSE, LMMSE, MAP)
- { LMMSE (for filtering/smoothing/prediction of random signals)

MMSE = Minimum Mean Square Error

LMMSE = Linear Minimum Mean Square Error

MAP=Maximum A Posteriori



Large Numbers Law and Measure of the Probability of an Event

Deterministic parameters:

- { probability of an event $p = \Pr\{A\}$
- { moments of a distribution (mean value, variance, skewness, kurtosis, *etc.*)
- { parameters of a distribution
- { parameters of a signal (amplitude, phase, frequency, time of arrival, *etc.*)
- { moments of a random vector (mean vector, correlation matrix, *etc.*)
- { moments of a random signal (mean value function, ACF, PSD, *etc.*)

ACF=Autocorrelation Function, PSD=Power Spectral Density

Deterministic parameters:

- { Sample estimators (application of the Law of large numbers)
- { Method of Moments (MM)
- { Maximum Likelihood (ML) method
- { Least Squares (LS): Linear LS or Non Linear LS (NLLS)

- A **random experiment** is a process by which we observe something uncertain. After the experiment, the result of the random experiment is known.
- An **outcome** is a result of a random experiment. The set Ω of all possible outcomes is called the **sample space**. Thus, in the context of a random experiment, the sample space Ω is our universal set.
- Some examples of random experiments and their sample spaces:
 - toss a coin; sample space: $\Omega=\{H,T\}$ (heads, tails)
 - roll a die; sample space: $\Omega=\{1,2,3,4,5,6\}$
 - observe the number of iPhones sold by an Apple store in Pisa in 2018; sample space: $\Omega=\{0,1,2,3, \dots\}$
- When we repeat a random experiment several times, we call each one of them a **trial**. Thus, a trial is a particular performance of a random experiment. Note that the sample space is defined based on how you define your random experiment.

■ It is given a **random experiment** described by the triad $S = (\Omega, F, P)$, where:

Ω = sample space (union of all elementary outcomes)

F = set of all events (all possible outcomes and unions of them)

P = probability map (it associates a probability to any event in F)

■ Assume we are interested on a particular event $A \in F$, which has **probability of occurrence** p , i.e. $\Pr\{A\}=p$.

■ **Problem:** we want to measure, i.e. estimate, the probability of occurrence p of the event A having the possibility to run the experiment N times under the same conditions and in such a way that the outcome of each trial is independent from the outcomes in the other trials.

■ **Observed data:** they are given by the N outputs of the N trials of the random experiment. In particular, at the end of each trial, we observe if the event A occurs or not. Note that in this case the data are already in discrete format.

- Let us now define the following random variables (r.v.'s):

$$X_i \triangleq \begin{cases} 0 & \text{if the event } A \text{ does not occur in the } i\text{-th trial} \\ 1 & \text{if the event } A \text{ occurs in the } i\text{-th trial} \end{cases}$$

$$K \triangleq \sum_{i=1}^N X_i \quad \text{Number of times } A \text{ occurs over the } N \text{ trials}$$

- The $\{X_i\}$ are our observed data. They are binary random variables.
- Due to assumption that the N trials are run under the same conditions and each one independently from the others, we have that the $\{X_i\}$ are *independent and identically distributed* (IID) random variables.
- As a consequence K is a **binomial** r.v. of parameters N and p .

■ **Mean value** and **variance** of the IID binary random variables $\{X_i\}$

$$X_i \triangleq \begin{cases} 0 & \text{if the event } A \text{ does not occur in the } i\text{-th run} \\ 1 & \text{if the event } A \text{ occurs in the } i\text{-th run} \end{cases}$$

$$\Pr\{A\} = p \Rightarrow \Pr\{X_i = 1\} = \Pr\{A\} = p, \quad \Pr\{X_i = 0\} = \Pr\{\bar{A}\} = 1 - p$$

Probability mass function (pmf):

$$p_{X_i}(x) = \Pr\{X_i = x\} = (1 - p)\delta[x] + p\delta[x - 1], \quad \forall i$$

Kronecker's delta function

$$\left\{ \begin{array}{l} \eta_X = E\{X_i\} = (1 - p) \cdot 0 + p \cdot 1 = p \\ P_X = E\{X_i^2\} = (1 - p) \cdot 0^2 + p \cdot 1^2 = p \\ \sigma_X^2 = \text{var}\{X_i\} = E\{(X_i - \eta_X)^2\} = P_X - \eta_X^2 = p - p^2 = p(1 - p) \end{array} \right.$$

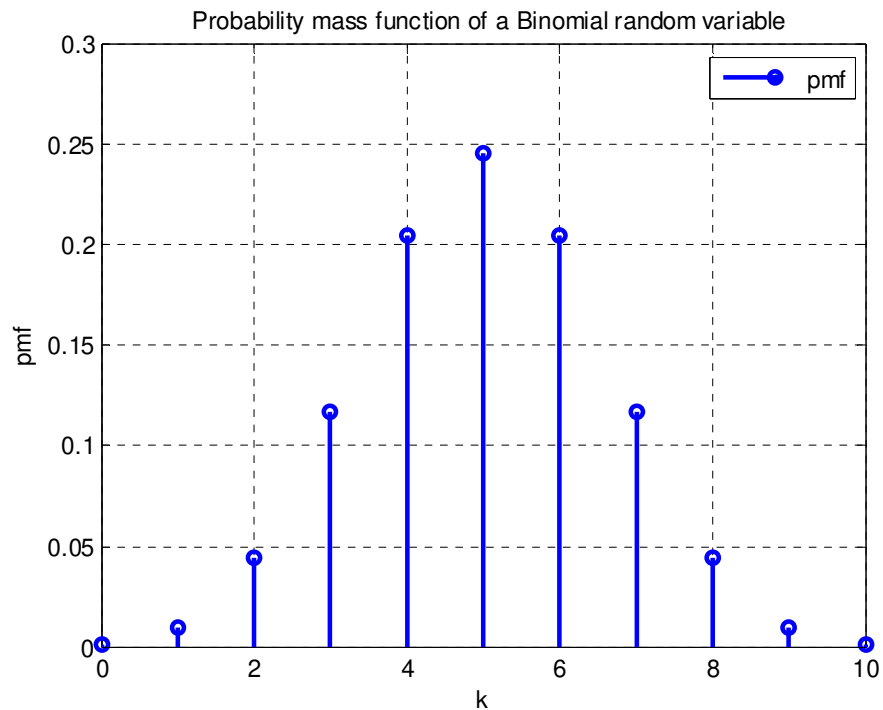
- **Mean value** and **variance** of the binomial random variable K :

$$K \triangleq \sum_{i=1}^N X_i \quad \Rightarrow \quad K \in B(p, N)$$

pmf :

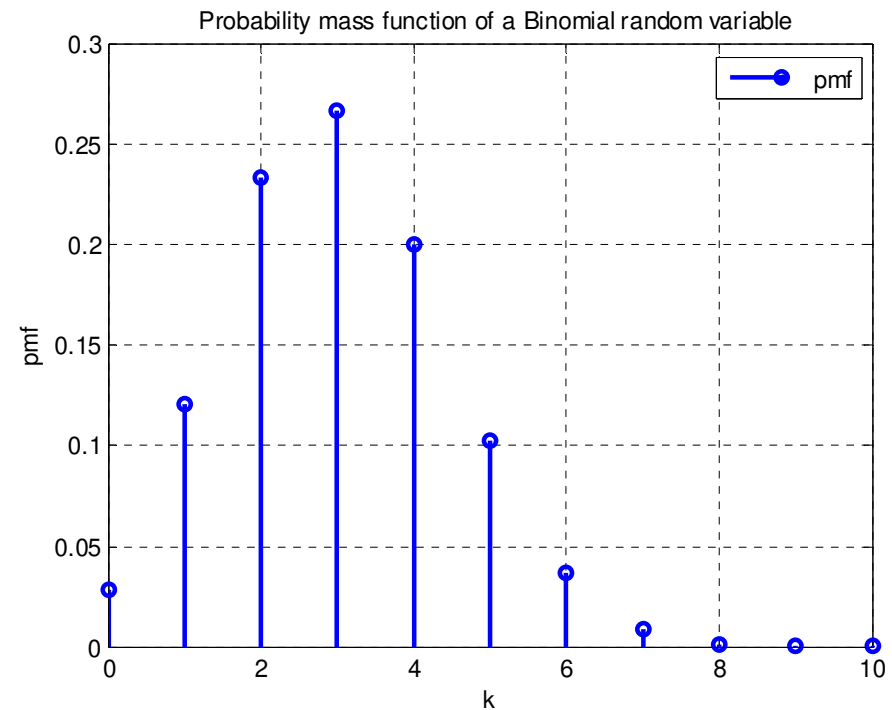
$$p_K(k) = P\{K = k\} = \binom{N}{k} p^k (1-p)^{N-k}, \quad 0 < p < 1, \quad 0 \leq k \leq N$$

$$\left\{ \begin{array}{l} \eta_K = m_K(1) = E\{K\} = \sum_{k=0}^N k p_K(k) = Np \\ \sigma_K^2 = \mu_K(2) = \text{var}\{K\} = E\{(K - \eta_K)^2\} = E\{K^2\} - \eta_K^2 \\ \quad = \sum_{k=0}^N k^2 p_K(k) - \eta_K^2 = Np(1-p) \end{array} \right.$$



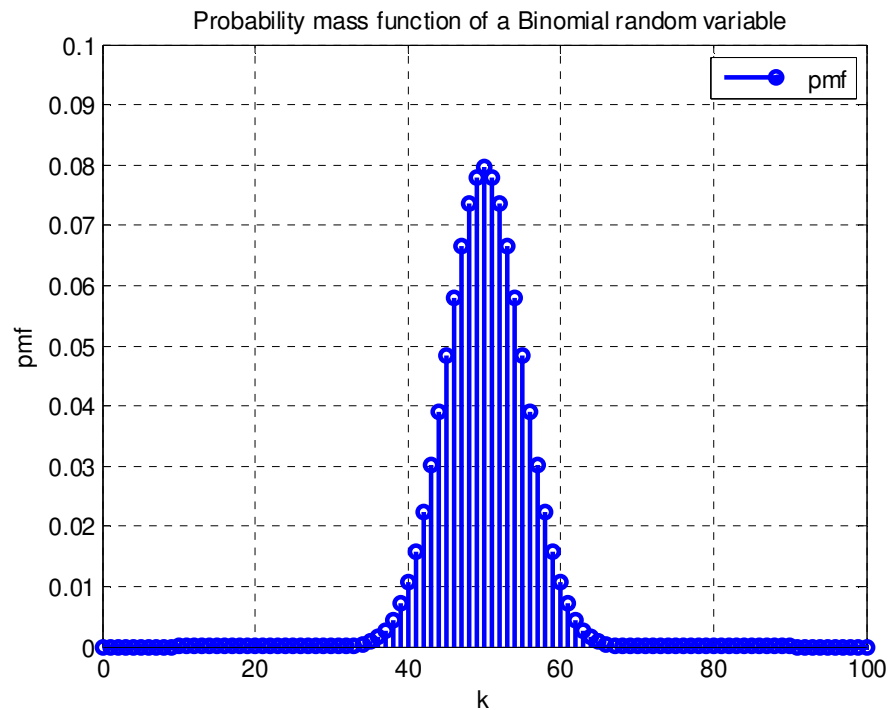
$$K \in B(0.5, 10)$$

$$\eta_K = Np = 5$$



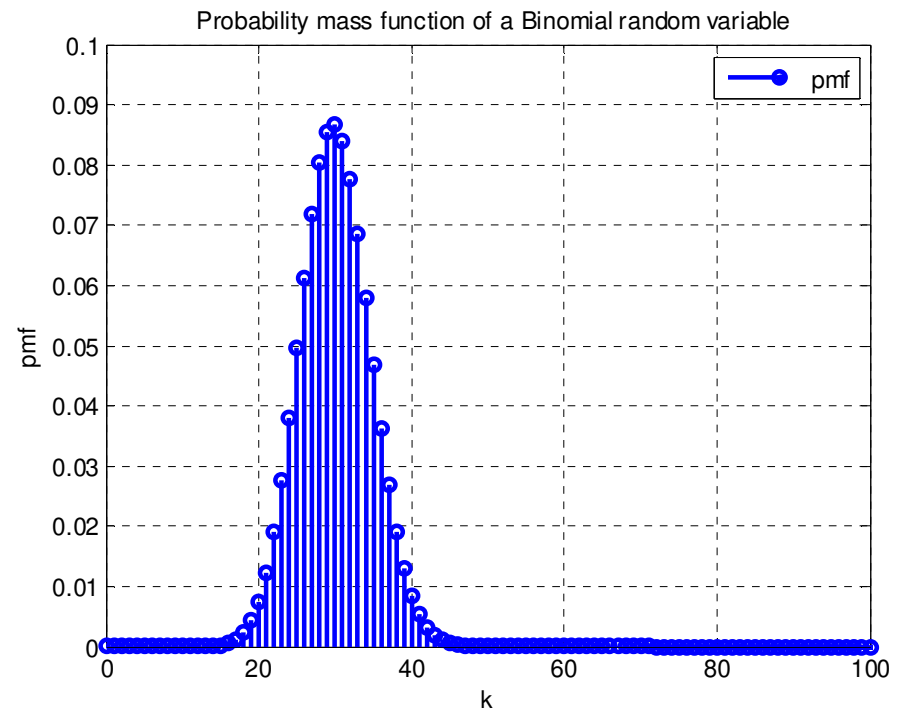
$$K \in B(0.3, 10)$$

$$\eta_K = Np = 3$$



$$K \in B(0.5, 100)$$

$$\eta_K = Np = 50$$



$$K \in B(0.3, 100)$$

$$\eta_K = Np = 30$$

- **Mean value** and **variance** of the binomial r.v. K :
- There are three different methods to calculate mean and variance of a binomial r.v., we see only one of them:

$$K \in B(p, N) \quad \Leftrightarrow \quad K = \sum_{i=1}^N X_i, \quad X_i \in \{0, 1\}, \quad \Pr\{X_i = 1\} = p$$

$$\begin{cases} \eta_X = E\{X_i\} = p \\ \sigma_X^2 = \text{var}\{X_i\} = p(1-p) \end{cases}$$



On the average, the event A occurs Np times over the N trials.

$$\eta_K = E\{K\} = E\left\{\sum_{i=1}^N X_i\right\} = \sum_{i=1}^N E\{X_i\} = \sum_{i=1}^N p = Np$$





$$\begin{aligned}\sigma_K^2 &= \text{var}\{K\} = E\{(K - Np)^2\} = E\left\{\left(\sum_{i=1}^N X_i - Np\right)^2\right\} \\&= E\left\{\left(\sum_{i=1}^N (X_i - p)\right)^2\right\} = E\left\{\sum_{i=1}^N (X_i - p) \sum_{k=1}^N (X_k - p)\right\} \\&= E\left\{\sum_{i=1}^N \sum_{k=1}^N (X_k - p)(X_i - p)\right\} = \sum_{i=1}^N \sum_{k=1}^N E\{(X_k - p)(X_i - p)\} \\&= \sum_{i=1}^N E\{(X_i - p)^2\} + \sum_{i=1}^N \sum_{\substack{k=1 \\ k \neq i}}^N E\{(X_k - p)(X_i - p)\} \\&= \sum_{i=1}^N E\{(X_i - p)^2\} = \sum_{i=1}^N \text{var}\{X_i\} = Np(1 - p)\end{aligned}$$

■ Note that:
$$\sum_{i=1}^N \sum_{\substack{k=1 \\ k \neq i}}^N E\{(X_k - p)(X_i - p)\} = 0$$

because X_k and X_i are independent, and as a consequence, they are also uncorrelated:

$$E\{(X_k - p)(X_i - p)\} = E\{(X_k - p)\}E\{(X_i - p)\} = 0 \quad \text{for } i \neq k$$

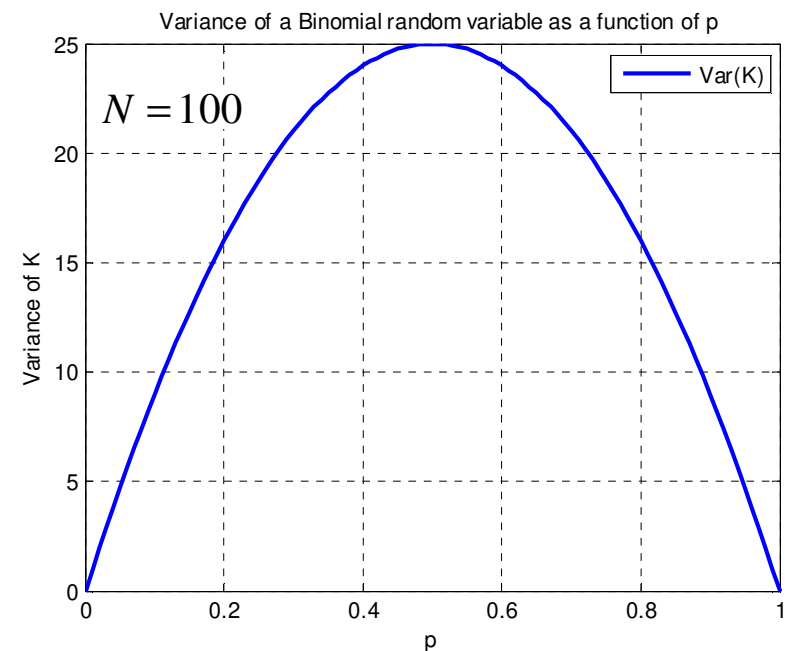
■ In summary, if the N r.v.'s are uncorrelated \rightarrow the variance of the sum is equal to the sum of the variances:

$$\sigma_K^2 = \text{var}\{K\} = \text{var}\left\{\sum_{i=1}^N X_i\right\} = \sum_{i=1}^N \text{var}\{X_i\} = Np(1-p)$$

■ Note that the maximum of the variance of K (maximum spread of values around the mean) is $N/4$ and it is obtained when $p=0.5$, i.e. when the event A and its complementary event have the same probability of occurrence (maximum uncertainty about the outcome of each trial).

■ The variance of K is 0 when $p=0$ or $p=1$ (no spread of values around the mean, no uncertainty about the outcome of the trial). In this case, K assumes with probability 1 its mean value.

$$K \in B(p, N) \rightarrow \sigma_K^2 = Np(1-p)$$



- There is an important theorem about binomial random variables that we will use in the following: the **De Moivre-Laplace Theorem**.
- It is a special case of the **Central-Limit Theorem (CLT)**: it states that the normal distribution may be used as an approximation to the binomial distribution under certain conditions.
- In particular, the theorem shows that the **probability mass function (pmf)** of the random number K of "successes" observed in a series of N independent Bernoulli trials, each having probability p of success, converges to the **probability density function (pdf)** of the normal distribution with mean Np and variance $Np(1-p)$:

$$K \xrightarrow{N \rightarrow \infty} X, \quad \text{where} \quad X \in \mathcal{N}(Np, Np(1-p))$$

convergence in distribution

Convergence in distribution:

$$F_K(k) \triangleq \Pr\{K \leq k\} = \sum_{i=0}^k \Pr\{K = i\} = \sum_{i=0}^k \binom{N}{i} p^i (1-p)^{N-i} \xrightarrow{N \rightarrow \infty} F_X(k)$$

$$\text{where } F_X(k) \triangleq \Pr\{X \leq k\} = 1 - Q\left(\frac{k - Np}{\sqrt{Np(1-p)}}\right)$$

- where Q is the Q error function.
- In other words, if N is large enough, with good accuracy we can approximate the discrete r.v. K as a continuous Gaussian r.v. having the same mean and variance as K :

$$K \in B(p, N), \quad \text{if } Np(1-p) \gg 1: \quad K \in \mathcal{N}(Np, Np(1-p))$$

- The approximation requires $Np(1-p) \gg 1$, but it is already good for $Np(1-p) \geq 10$.

■ Standard Gaussian (or normal) r.v.:

$$X \in \mathcal{N}(0,1) \rightarrow f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

$$\Phi(x) \triangleq F_X(x) = \Pr\{X \leq x\} = \int_{-\infty}^x f_X(z) dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{z^2}{2}\right) dz$$

■ $\Phi(x)$ is not available in closed-form, it must be calculated numerically.
In MATLAB we use the script **normcdf.m**.

■ Frequently, we use the **Q function** defined as:

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} \exp\left(-\frac{z^2}{2}\right) dz = 1 - \Phi(x)$$

in MATLAB: **qfunc.m**

- For a **Gaussian (or normal) r.v.** with given mean η and variance σ^2 :

$$X \in \mathcal{N}(\eta, \sigma^2)$$

$$F_X(x) = \Pr\{X \leq x\} = \int_{-\infty}^x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(z-\eta)^2}{2\sigma^2}\right] dz$$

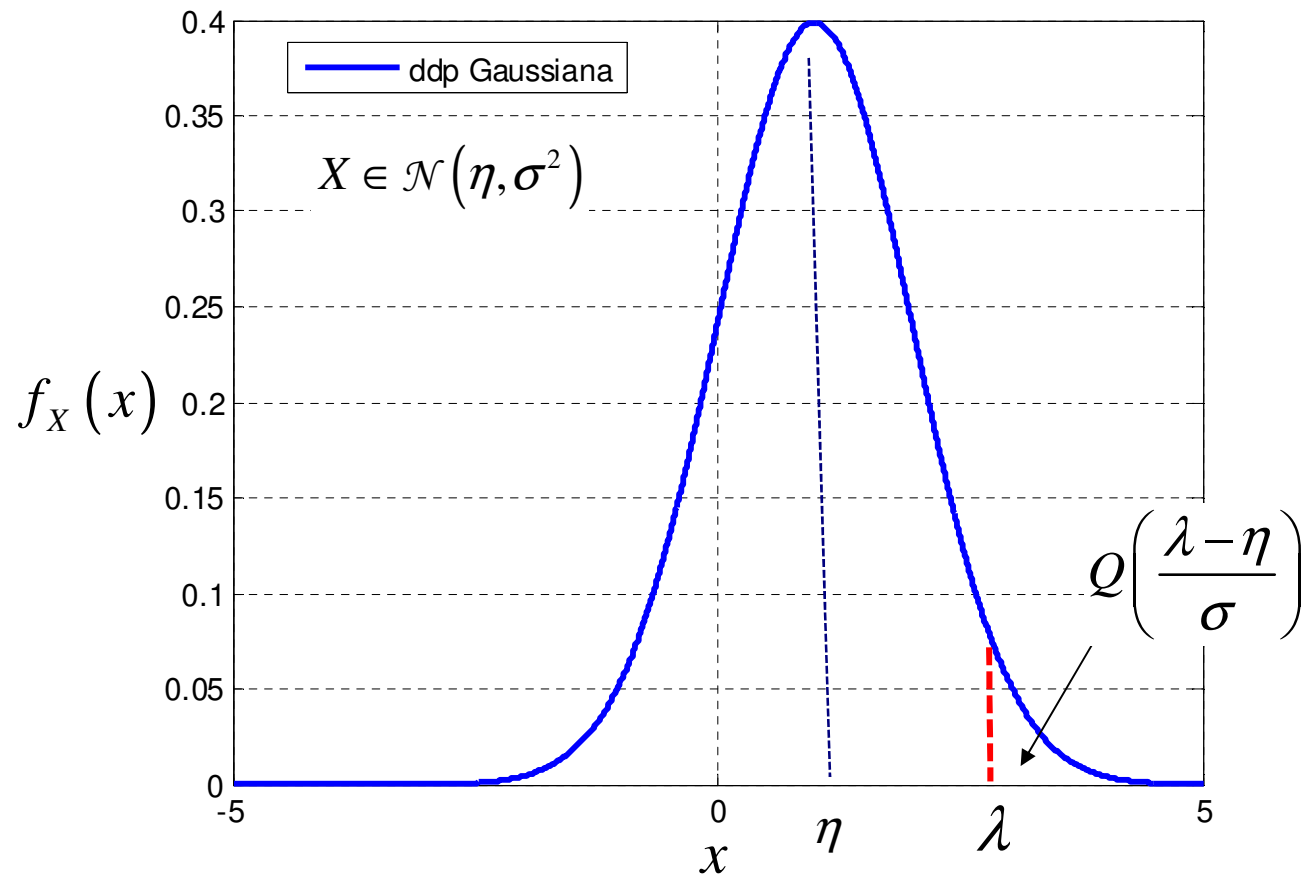
$$\left[\text{by setting: } y \triangleq \frac{z-\eta}{\sigma} \Rightarrow dy = \frac{dz}{\sigma} \right]$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\eta)/\sigma} \exp\left(-\frac{y^2}{2}\right) dy = \Phi\left(\frac{x-\eta}{\sigma}\right) = 1 - Q\left(\frac{x-\eta}{\sigma}\right)$$

$$\Rightarrow \Pr\{X > \lambda\} = 1 - F_X(\lambda) = Q\left(\frac{\lambda - \eta}{\sigma}\right)$$

- For a **Gaussian (or normal) r.v.** having mean η and variance σ^2 :

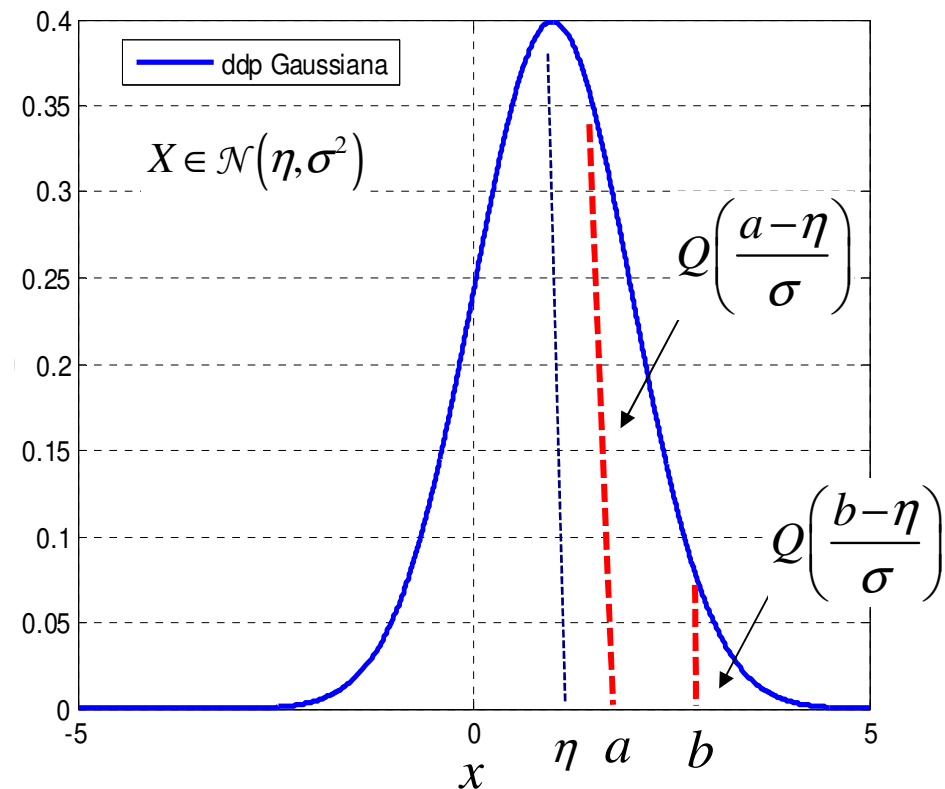
$$\Pr\{X > \lambda\} = 1 - F_X(\lambda) = Q\left(\frac{\lambda - \eta}{\sigma}\right)$$



- In general, for a **Gaussian (or normal) r.v.** with mean η and variance σ^2 , the probability of any event we are interested can be easily calculated by using the Q function:

$$\begin{aligned}\Pr\{a < X \leq b\} \\&= \int_a^b f_X(\alpha) d\alpha = F_X(b) - F_X(a) \\&= 1 - Q\left(\frac{b-\eta}{\sigma}\right) - \left(1 - Q\left(\frac{a-\eta}{\sigma}\right)\right) \\&= Q\left(\frac{a-\eta}{\sigma}\right) - Q\left(\frac{b-\eta}{\sigma}\right)\end{aligned}$$

Note: $Q(z) = 1 - Q(-z)$

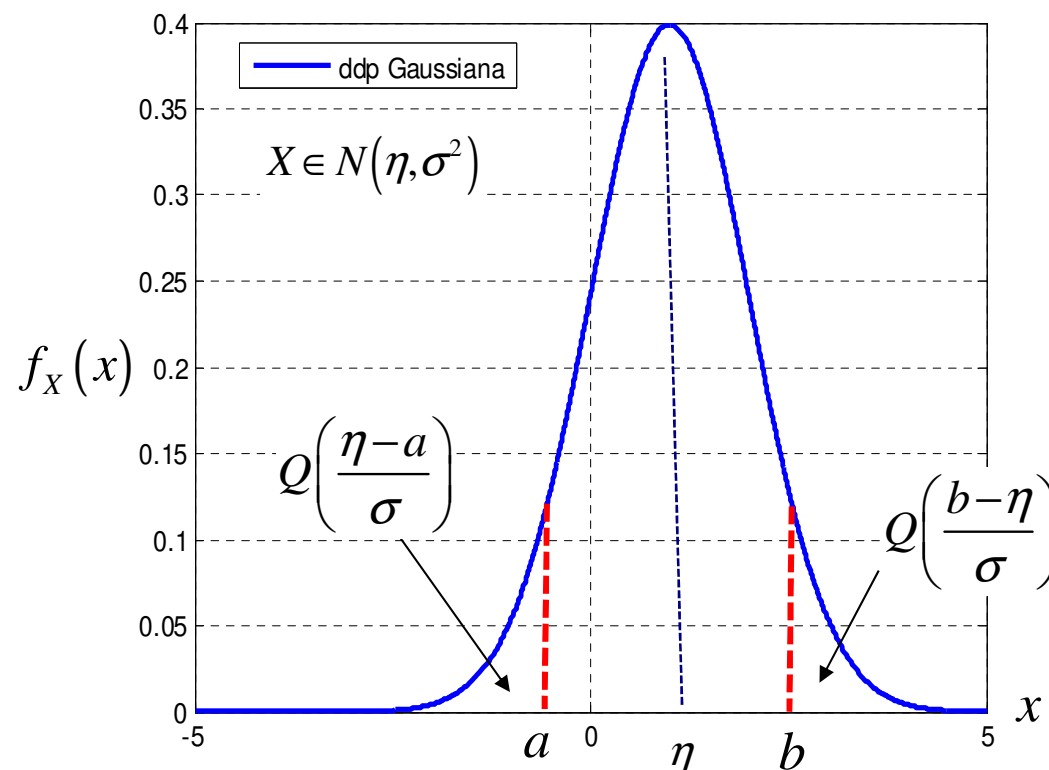


Measure of the Probability of an Event

38

■ **Gaussian (or normal) r.v.** with mean η and variance σ^2 : the probability that the r.v. assumes a value that deviates from the mean value less than k times the standard deviation σ is given by:

$$\Pr\{|X - \eta| \leq k\sigma\} = \Pr\{\eta - k\sigma \leq X \leq \eta + k\sigma\} = \int_{\eta - k\sigma}^{\eta + k\sigma} f_X(z) dz$$



$$a = \eta - k\sigma$$
$$b = \eta + k\sigma$$

38

$$\begin{aligned}\Pr\{|X - \eta| \leq k\sigma\} &= \int_{\eta - k\sigma}^{\eta + k\sigma} f_X(z) dz \\ &= 1 - Q\left(\frac{(\eta + k\sigma) - \eta}{\sigma}\right) - Q\left(\frac{\eta - (\eta - k\sigma)}{\sigma}\right) \\ &= 1 - 2Q(k), \quad k = 1, 2, 3, \dots\end{aligned}$$

$$\Pr\{|X - \eta| \leq \sigma\} = 1 - 2Q(1) \cong 0.683$$

$$\Pr\{|X - \eta| \leq 2\sigma\} = 1 - 2Q(2) \cong 0.956$$

$$\Pr\{|X - \eta| \leq 3\sigma\} = 1 - 2Q(3) \cong 0.997$$

Measure of the Probability of an Event

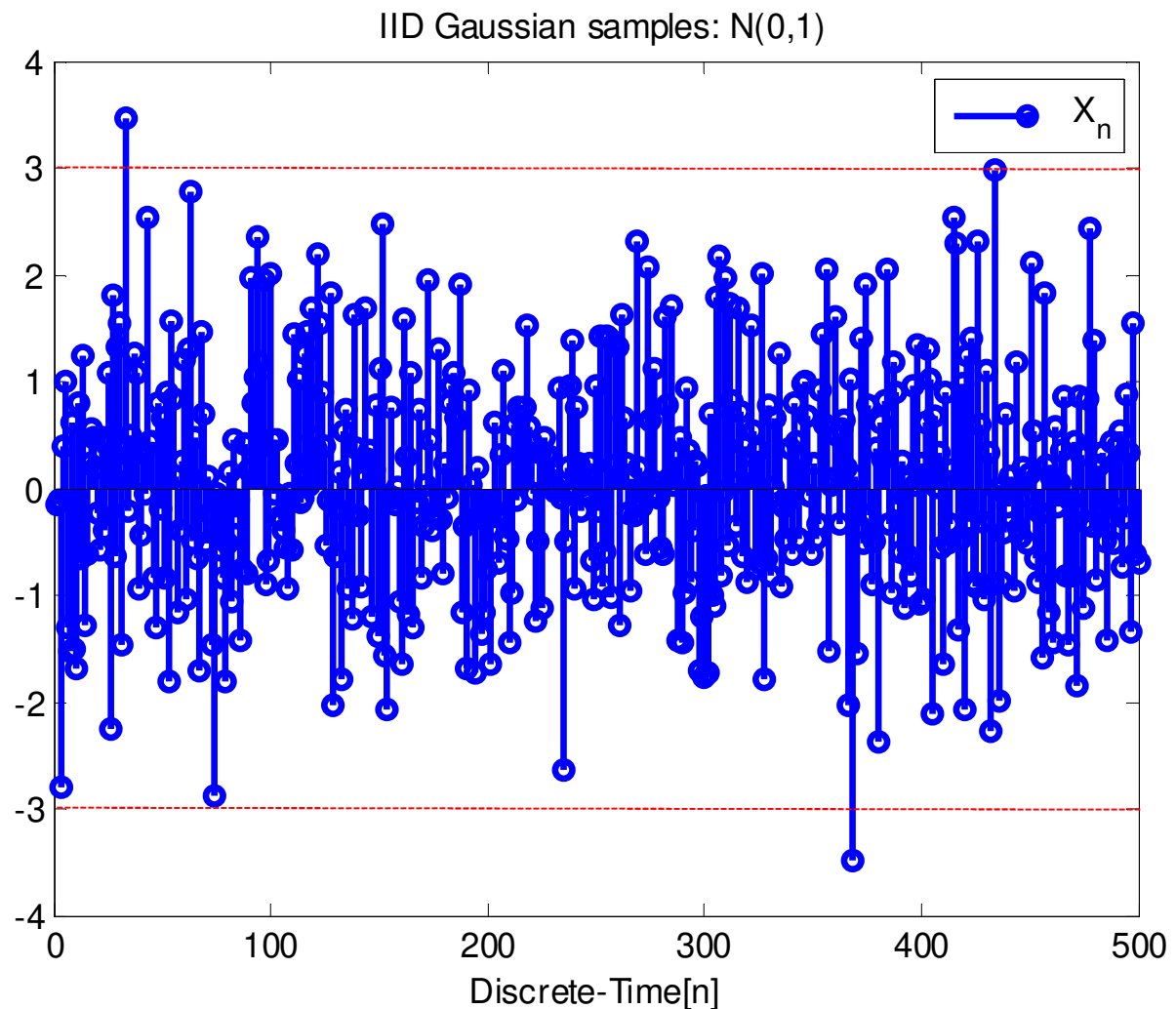
40

500 independent realizations of a r.v. $X \in \mathcal{N}(0,1) \rightarrow \eta = 0, \sigma^2 = 1$

$$\Pr\{|X| \leq 1\} \cong 0.683$$

$$\Pr\{|X| \leq 2\} \cong 0.956 \quad \times^c$$

$$\Pr\{|X| \leq 3\} \cong 0.997$$



■ Frequency of occurrence F of an event A :

$$F \triangleq \frac{K}{N} = \frac{1}{N} \sum_{i=1}^N X_i \quad \left[\text{Aritmetic mean of the observed data } \{X_i\} \right]$$

■ Mean value and variance of F :

$$\eta_F = E\{F\} = \frac{E\{K\}}{N} = p$$


$$\begin{aligned} \sigma_F^2 &= E\{(F - \eta_F)^2\} = E\left\{\left(\frac{K}{N} - p\right)^2\right\} = E\left\{\left(\frac{K - Np}{N}\right)^2\right\} \\ &= \frac{E\{(K - Np)^2\}}{N^2} = \frac{\text{var}\{K\}}{N^2} = \frac{p(1-p)}{N} \end{aligned}$$

- F is a discrete r.v. that can assume values $f_k = k/N$, for $k=0,1,2, \dots, N$.

Since $p_F(f_k) \triangleq \Pr\left\{F = f_k = \frac{k}{N}\right\} = \Pr\{K = k\} \Rightarrow p_F(f_k) = p_K(k)|_{k=Nf_k}$

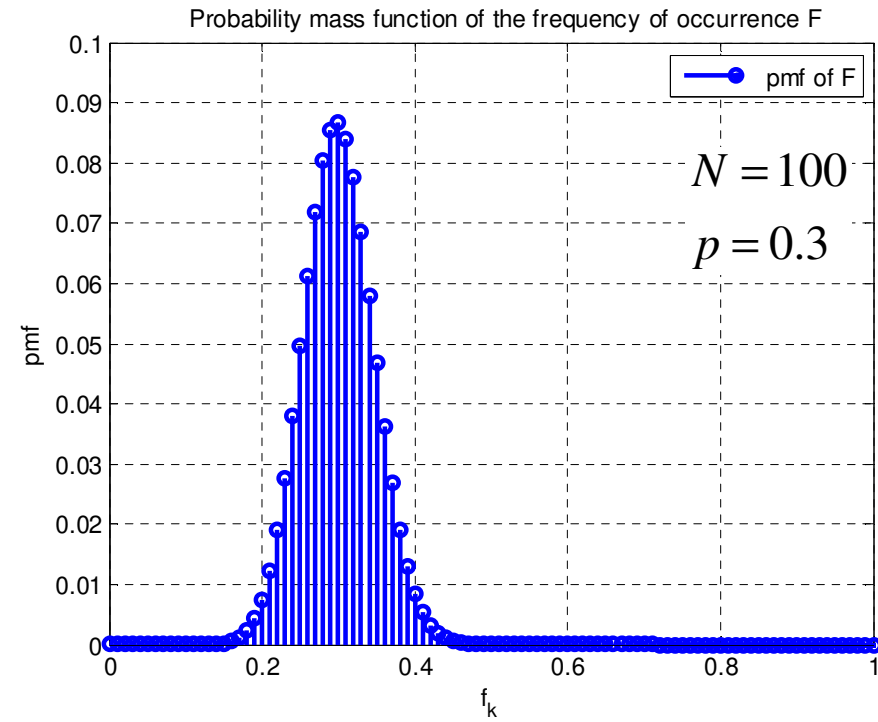
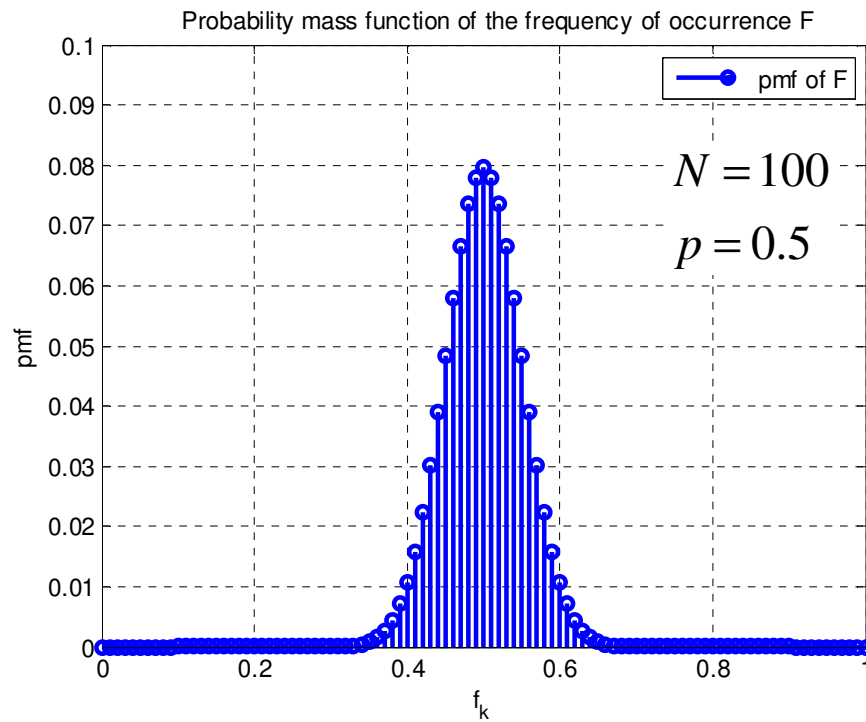
- Hence, the **probability mass function (pmf)** of F is given by:

$$p_K(k) = \Pr\{K = k\} = \binom{N}{k} p^k (1-p)^{N-k}, \quad 0 \leq k \leq N$$


$$p_F(f_k) = \Pr\{F = f_k\} = \binom{N}{Nf_k} p^{Nf_k} (1-p)^{N(1-f_k)}, \quad f_k = \frac{k}{N}, 0 \leq k \leq N$$

Measure of the Probability of an Event

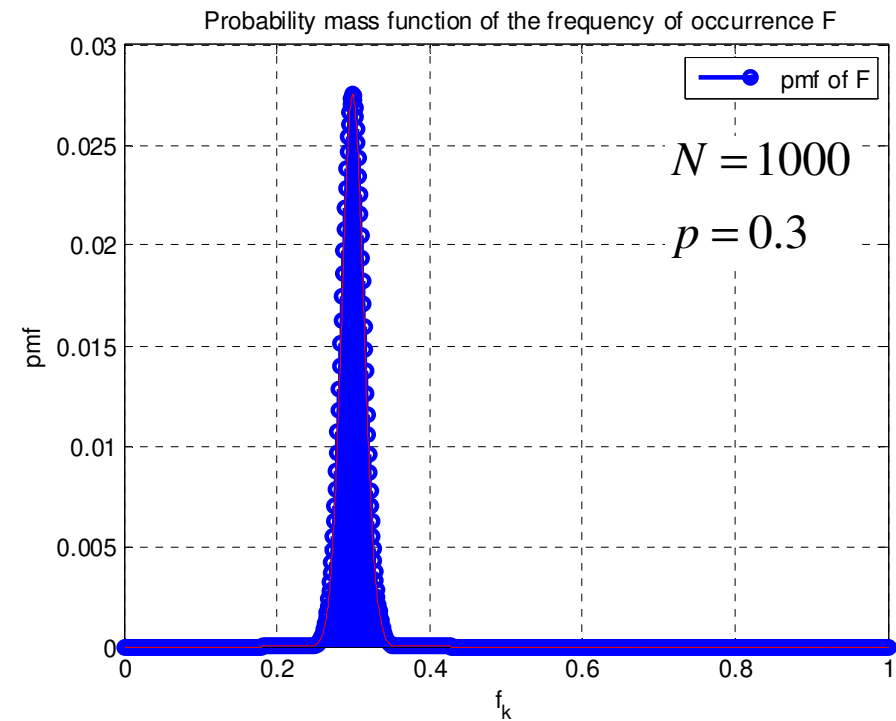
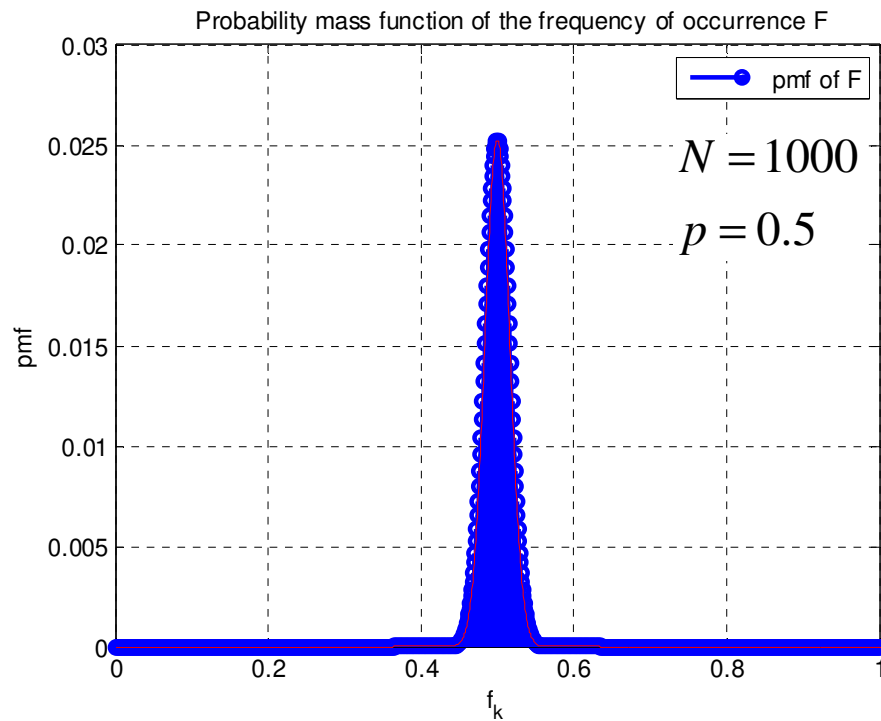
43



$$\eta_F = E\{F\} = p, \quad \sigma_F^2 = E\{(F - \eta_F)^2\} = \frac{p(1-p)}{N}$$

Measure of the Probability of an Event

44



$$\eta_F = E\{F\} = p, \quad \sigma_F^2 = E\{(F - \eta_F)^2\} = \frac{p(1-p)}{N}$$

$$\eta_F = p, \quad \lim_{N \rightarrow \infty} \sigma_F^2 = \lim_{N \rightarrow \infty} \frac{p(1-p)}{N} = 0$$

■ Hence: $F \xrightarrow{N \rightarrow \infty} p$

$$F = \langle X_i \rangle = \frac{1}{N} \sum_{i=1}^N X_i \xrightarrow{N \rightarrow \infty} E\{X_i\} = p$$

which means that the frequency of occurrence F of an event A (arithmetic mean of the N observed data X_i) converges to the probability p of the event (statistical mean of the X_i) when the number N of observed data grows to infinity.

■ The *convergence* is both in mean square sense and in probability.



Convergence in mean square sense

$$\lim_{N \rightarrow \infty} E\{|F - \eta_F|^2\} = \lim_{N \rightarrow \infty} \sigma_F^2 = \lim_{N \rightarrow \infty} \frac{p(1-p)}{N} = 0$$

Convergence in probability

$$\lim_{N \rightarrow \infty} \Pr\{|F - \eta_F| > \varepsilon\} = 0 \quad \forall \varepsilon > 0$$

■ We have already seen that convergence in **mean square sense** implies convergence in **probability** thanks to the **Bienaymé-Tchebycheff inequality**:

■ BERNOULLI THEOREM (OR LARGE NUMBERS LAW)

■ The frequency of occurrence F of an event A , whose probability of occurrence is p , converges in probability and in mean square sense to p when the number N of independent trials goes to infinity:

$$\lim_{N \rightarrow \infty} F = p$$

■ Hence, this suggests that the probability p can be estimated by calculating the frequency of occurrence F of the event A over a number N of independent trials of the random experiment.

■ This estimator is **linear**, **unbiased** and **consistent**.

- This estimator is **linear**, i.e. a linear function of the N data $\{X_i\}$, **unbiased** and **consistent**:

$$\hat{p} = F = \frac{1}{N} \sum_{i=1}^N X_i$$

$$E\{\hat{p}\} = E\{F\} = \eta_F = p \rightarrow \text{unbiased}$$

$$MSE\{\hat{p}\} = E\{(\hat{p} - p)^2\} = E\{(F - \eta_F)^2\} = \sigma_F^2 = \frac{p(1-p)}{N}$$

$$\lim_{N \rightarrow \infty} MSE\{\hat{p}\} = \lim_{N \rightarrow \infty} \sigma_F^2 = \lim_{N \rightarrow \infty} \frac{p(1-p)}{N} = 0 \rightarrow \text{consistent}$$

- De Moivre-Laplace theorem \rightarrow the estimator is also **asymptotically Gaussian**:

$$Np(1-p) \gg 1 \rightarrow \hat{p} \overset{a}{\in} \mathcal{N}\left(p, \frac{p(1-p)}{N}\right)$$

Measure of the Probability of an Event

49

■ *Example:* we have a sequence of binary symbols $\{A_k\}$ generated by a **discrete source** and we want to estimate the probability p of symbol "1".

$$A_k \in \{0,1\} \quad IID$$

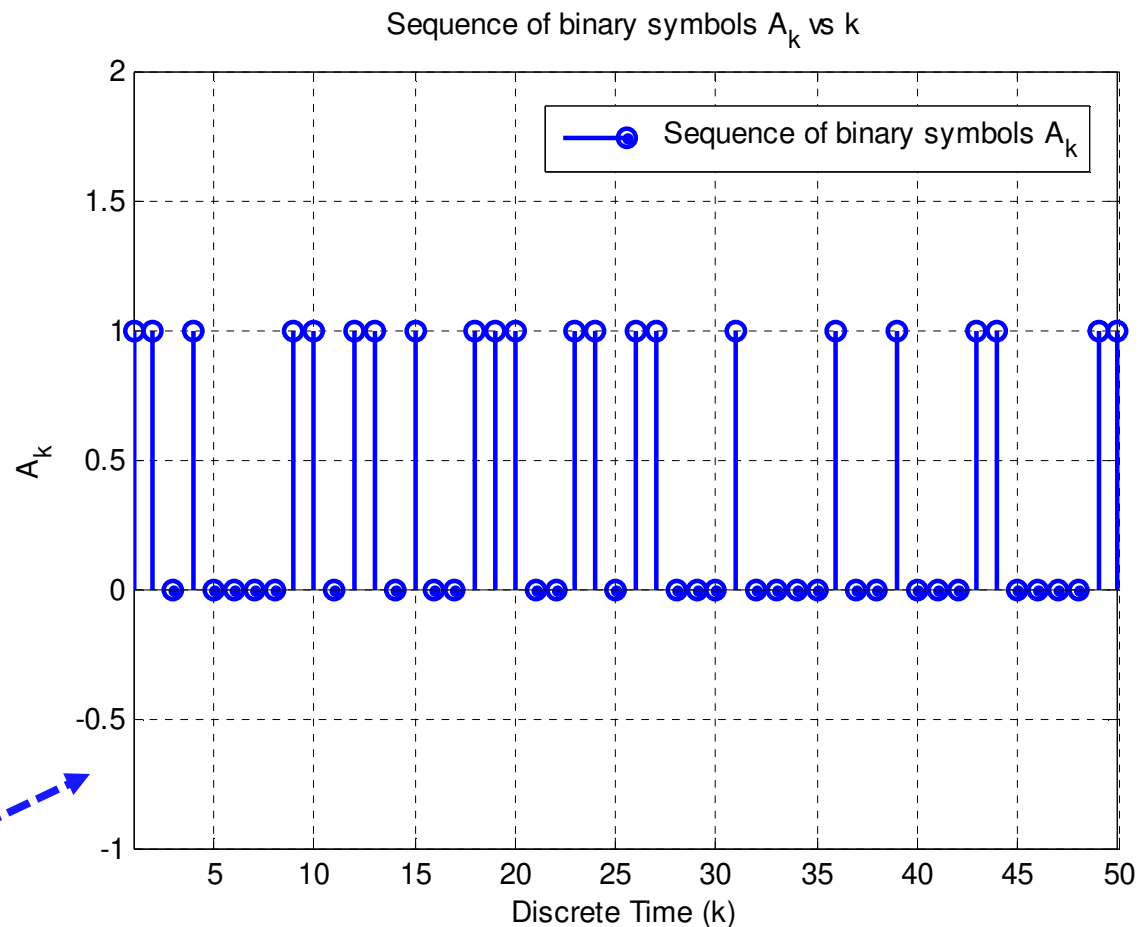
$$\Pr\{A_k = 1\} = p$$

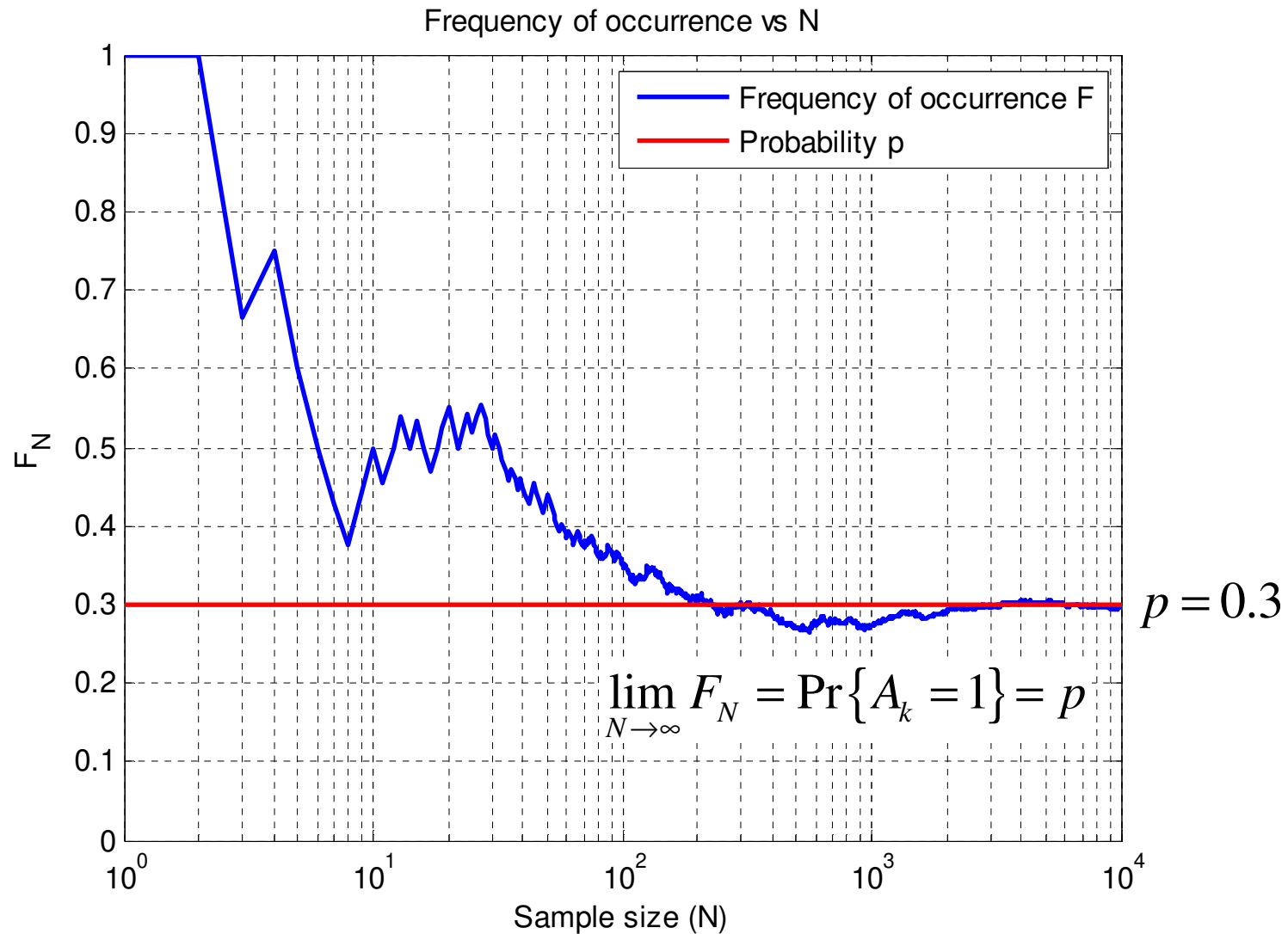
$$\Pr\{A_k = 0\} = 1 - p$$

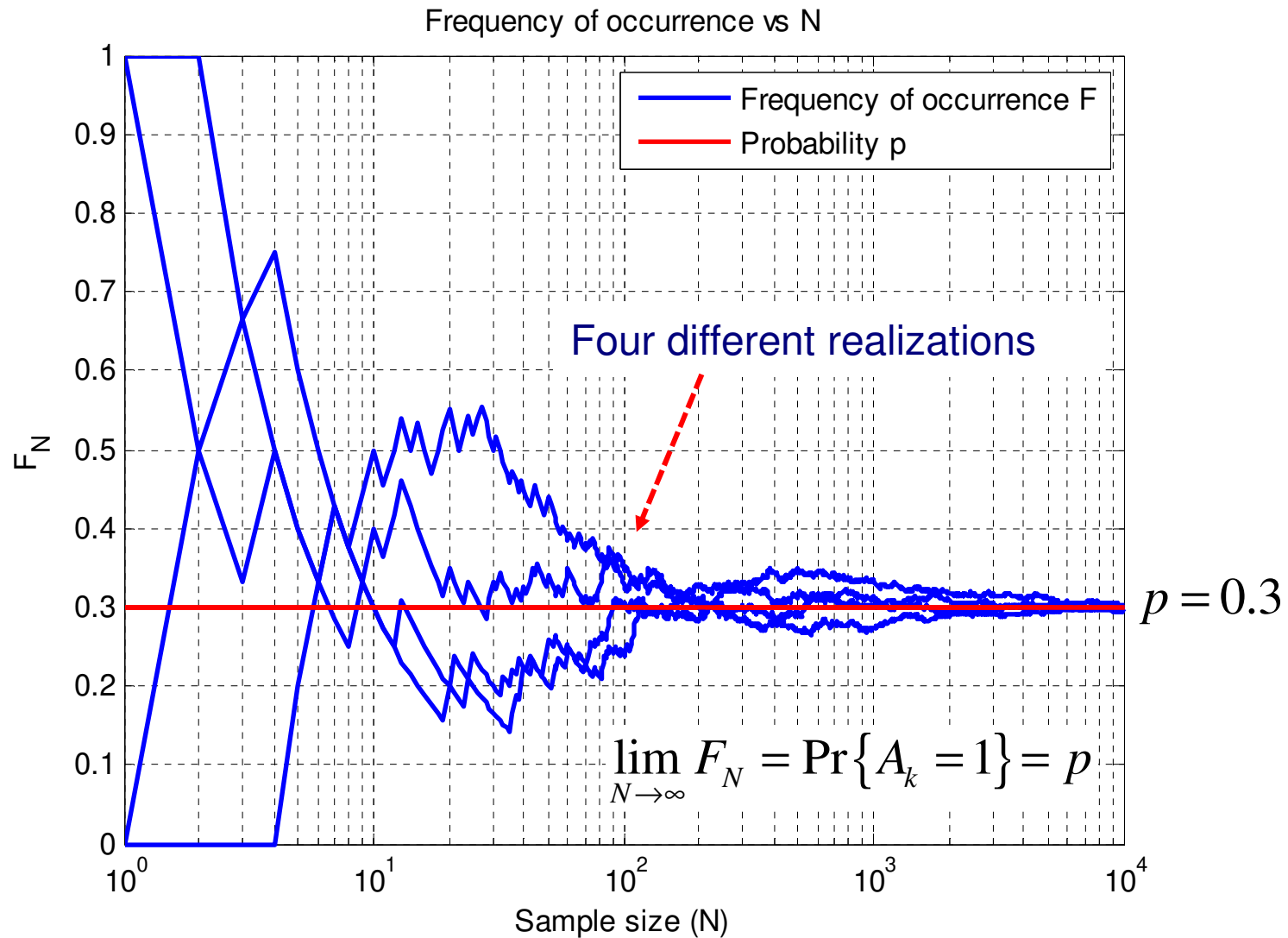
$$N = 10^4$$

$$p = 0.3$$

■ First 50 symbols







- *Problem:* How to choose the number of observations N ?
- A possible criterion is to look for the minimum number of observed data which guarantees that the standard deviation of the error is a given fraction α of the true value p , e.g. $\alpha=0.1$, i.e. 10% of p . This is equivalent to say that the MSE should be lower than $\alpha^2 p^2$:

$$\sigma_{\varepsilon} = \sqrt{E\{(F - p)^2\}} \leq \alpha p \rightarrow E\{(F - p)^2\} = \text{MSE}\{\hat{p}\} \leq \alpha^2 p^2$$
$$\rightarrow \frac{E\{(F - p)^2\}}{p^2} = E\left\{\left(\frac{F - p}{p}\right)^2\right\} = E\{\varepsilon_r^2\} \leq \alpha^2$$

$$\text{Relative error: } \varepsilon_r \triangleq \frac{F - p}{p} \Rightarrow E\{\varepsilon_r^2\} = \frac{E\{(F - p)^2\}}{p^2} = \frac{p(1 - p)}{Np^2} = \frac{1 - p}{Np} \leq \alpha^2$$

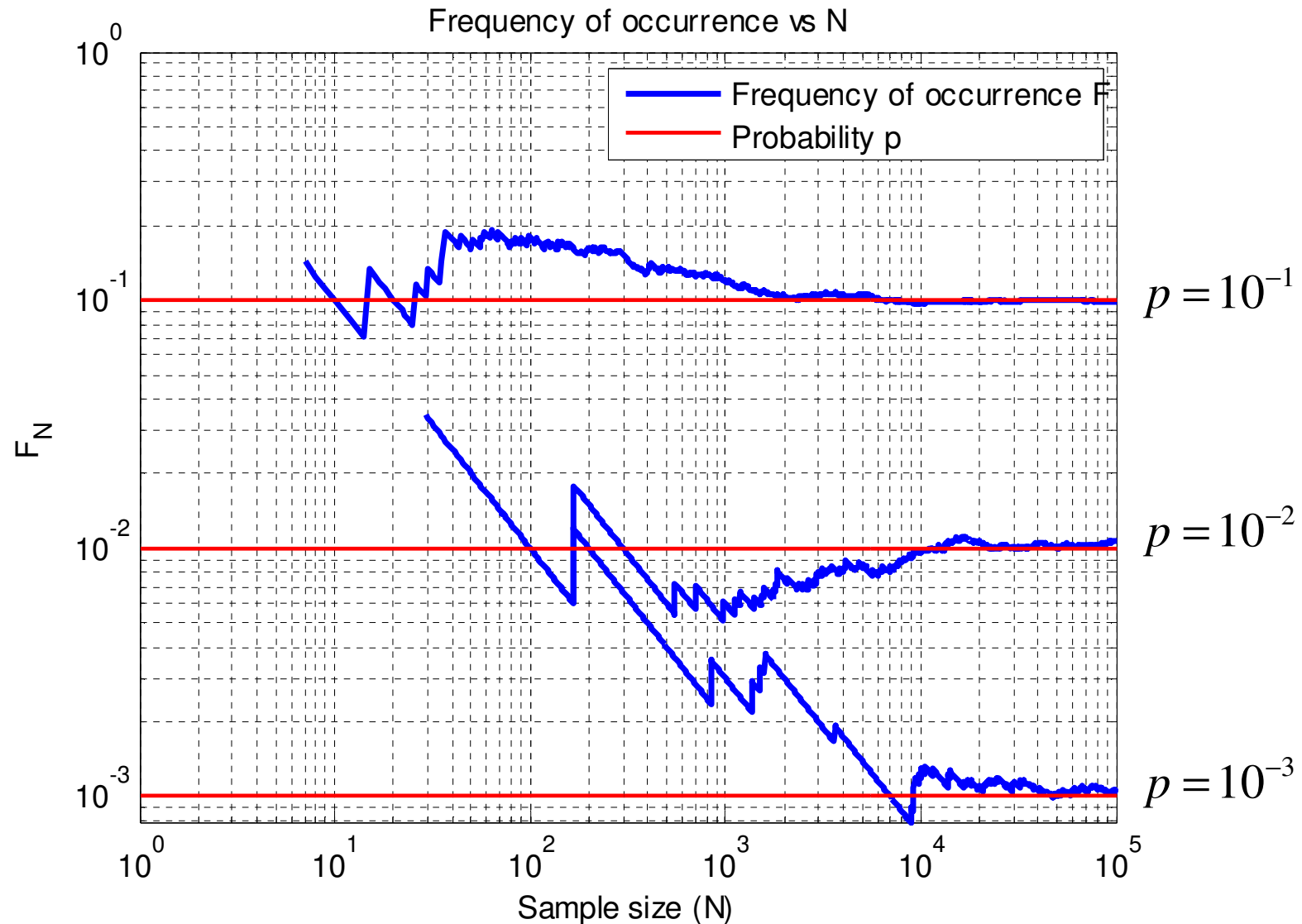
■ *Example:* $\alpha=0.1$, i.e. the standard deviation of the error should be lower than 10% of p .

$$\alpha = 0.1 \rightarrow E\{\varepsilon_r^2\} = \frac{1-p}{Np} \leq 0.01 \Rightarrow N \geq \frac{1-p}{0.01p} = \frac{100}{p}(1-p)$$

$$\forall p \in (0,1): N \geq \frac{100}{p} > \frac{100}{p}(1-p), \quad \text{if } p \ll 1: N \geq \frac{100}{p}$$

■ **Rule of thumb:** N should be about two orders of magnitude greater than $1/p$.

■ As a consequence, the lower is the probability p we want to estimate and the more data we need, if we want to achieve the required accuracy (10% in this case).





Measure of the Moments of a Random Variable

- We proved that if $\{X_i\}$'s are N independent and identically distributed (IID) binary random variables, when N goes to infinity the arithmetic mean converges to the statistical mean (ensemble mean):

$$F = \langle X_i \rangle = \frac{1}{N} \sum_{i=1}^N X_i \xrightarrow{N \rightarrow \infty} E\{X_i\} = p$$

- The convergence is in mean square sense and in probability.
- **Question:** Is this true only for IID binary random variables?
- As we will see in the following, the answer is NO. The only necessary condition is that the N r.v.'s are **identically distributed**.
- We can use this result to **estimate the mean value** of a general random variable X (not necessarily a binary r.v.).

- Let us define Z the arithmetic mean, also termed **Sample Mean**, of the $\{X_j\}$'s, that are N realizations of the r.v. X of which we want to estimate the mean value η_x .
- X can have **any probability density function** (pdf), it should not necessarily be a binary r.v.; it can be either a discrete or a continuous r.v.
- We want to investigate if the **Sample Mean** of the N observed data can be used as an **estimator** of the **mean value** of X .

$$\eta_x \triangleq E\{X\}, \quad \hat{\eta}_x = \langle X_i \rangle = \frac{1}{N} \sum_{i=1}^N X_i$$


- Is the **Sample Mean** a good estimator of η_x ?
- To answer this question we have to calculate its **bias** and **MSE**.
- We assume that the N r.v. are identically distributed, but not necessarily independent.

$$E\{\hat{\eta}_X\} = E\left\{\frac{1}{N} \sum_{i=1}^N X_i\right\} = \frac{1}{N} \sum_{i=1}^N E\{X_i\} = \frac{1}{N} \sum_{i=1}^N \eta_X = \eta_X$$

identically distributed \rightarrow same mean

■ Hence, the mean value of the **Sample Mean** is equal to the statistical mean of X , as a consequence, the Sample Mean is an **unbiased** estimator of the mean of the r.v. X .


■ In other words, the mean value of the estimation error is zero:


$$\begin{aligned}\varepsilon &\triangleq \eta_X - \hat{\eta}_X \\ E\{\varepsilon\} &= E\{\eta_X - \hat{\eta}_X\} = \eta_X - E\{\hat{\eta}_X\} = \eta_X - \eta_X = 0 \\ MSE\{\hat{\eta}_X\} &\triangleq E\{\varepsilon^2\} = E\{(\hat{\eta}_X - \eta_X)^2\} = \text{var}\{\hat{\eta}_X\}\end{aligned}$$

$$\begin{aligned}
 MSE\{\hat{\eta}_X\} &= \text{var}\{\hat{\eta}_X\} = E\{(\hat{\eta}_X - \eta_X)^2\} = E\left\{\left(\frac{1}{N} \sum_{i=1}^N X_i - \eta_X\right)^2\right\} \\
 &= E\left\{\left(\frac{1}{N} \sum_{i=1}^N (X_i - \eta_X)\right)^2\right\} = \frac{1}{N^2} E\left\{\sum_{i=1}^N (X_i - \eta_X) \sum_{k=1}^N (X_k - \eta_X)\right\} \\
 &= \frac{1}{N^2} \sum_{i=1}^N \sum_{k=1}^N E\{(X_k - \eta_X)(X_i - \eta_X)\} = \frac{1}{N^2} \sum_{i=1}^N \sum_{k=1}^N c_{X_k X_i} \\
 &= \frac{1}{N^2} \sum_{i=1}^N \sigma_X^2 + \frac{1}{N^2} \sum_{i=1}^N \sum_{\substack{k=1 \\ k \neq i}}^N c_{X_k X_i} = \frac{\sigma_X^2}{N} + \frac{1}{N^2} \sum_{i=1}^N \sum_{\substack{k=1 \\ k \neq i}}^N c_{X_k X_i}
 \end{aligned}$$

where $c_{X_k X_i} \triangleq \text{cov}\{X_k, X_i\} = E\{(X_k - \eta_X)(X_i - \eta_X)\}$

■ If the $\{X_i\}$'s are independent, they are also uncorrelated, so we have:


$$c_{X_k X_i} = \text{cov}\{X_k, X_i\} = E\{(X_k - \eta_X)(X_i - \eta_X)\} = \begin{cases} \sigma_X^2 & i = k \\ 0 & i \neq k \end{cases}$$

$$MSE\{\hat{\eta}_X\} = \frac{\sigma_X^2}{N}$$

■ If the variance of X is finite, the MSE of the Sample Mean tends to zero when N goes to infinity!

$$E\{\hat{\eta}_X\} = \eta_X \rightarrow \text{unbiased}$$

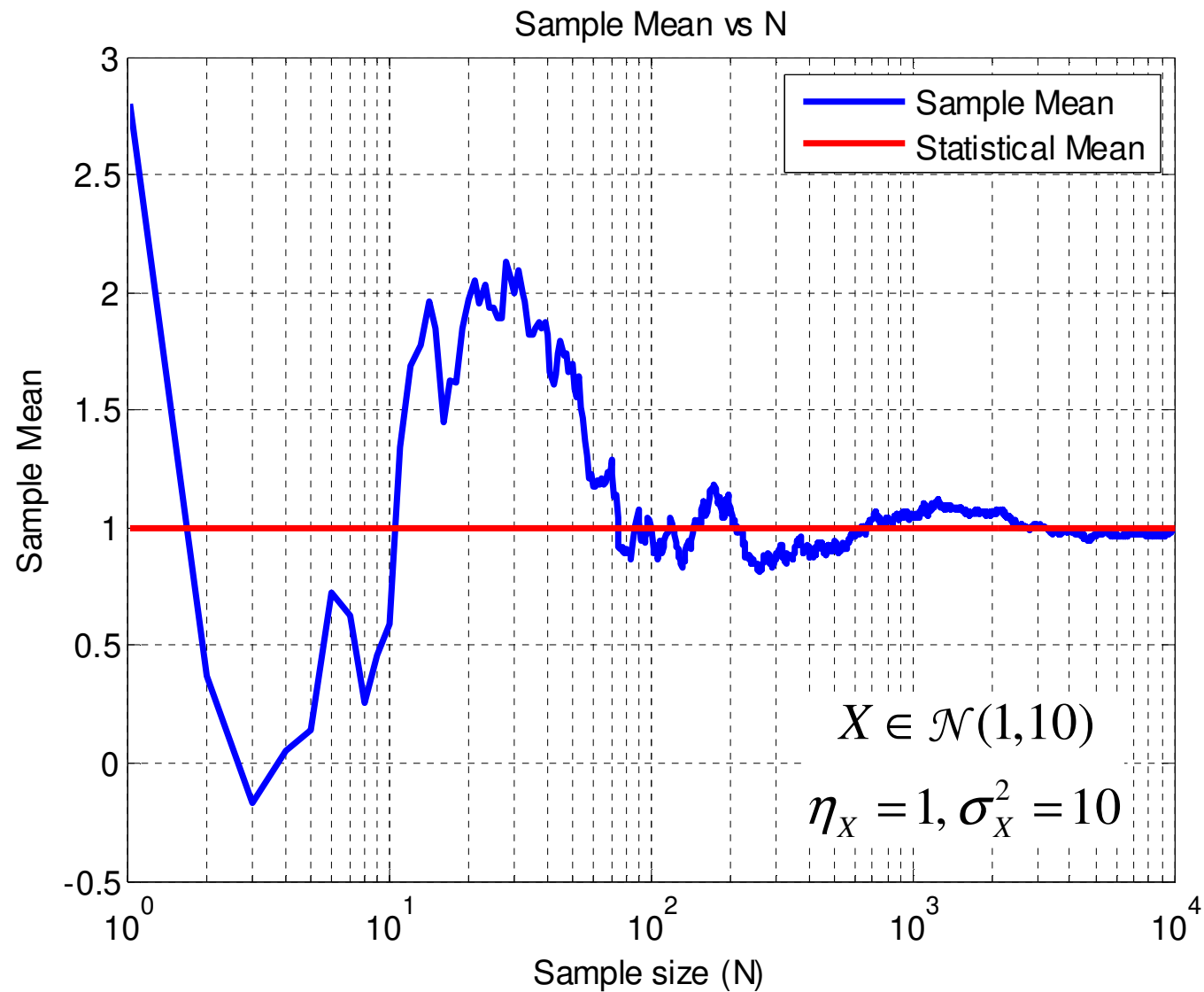
$$MSE\{\hat{\eta}_X\} = E\{(\hat{\eta}_X - \eta_X)^2\} = \frac{\sigma_X^2}{N}$$

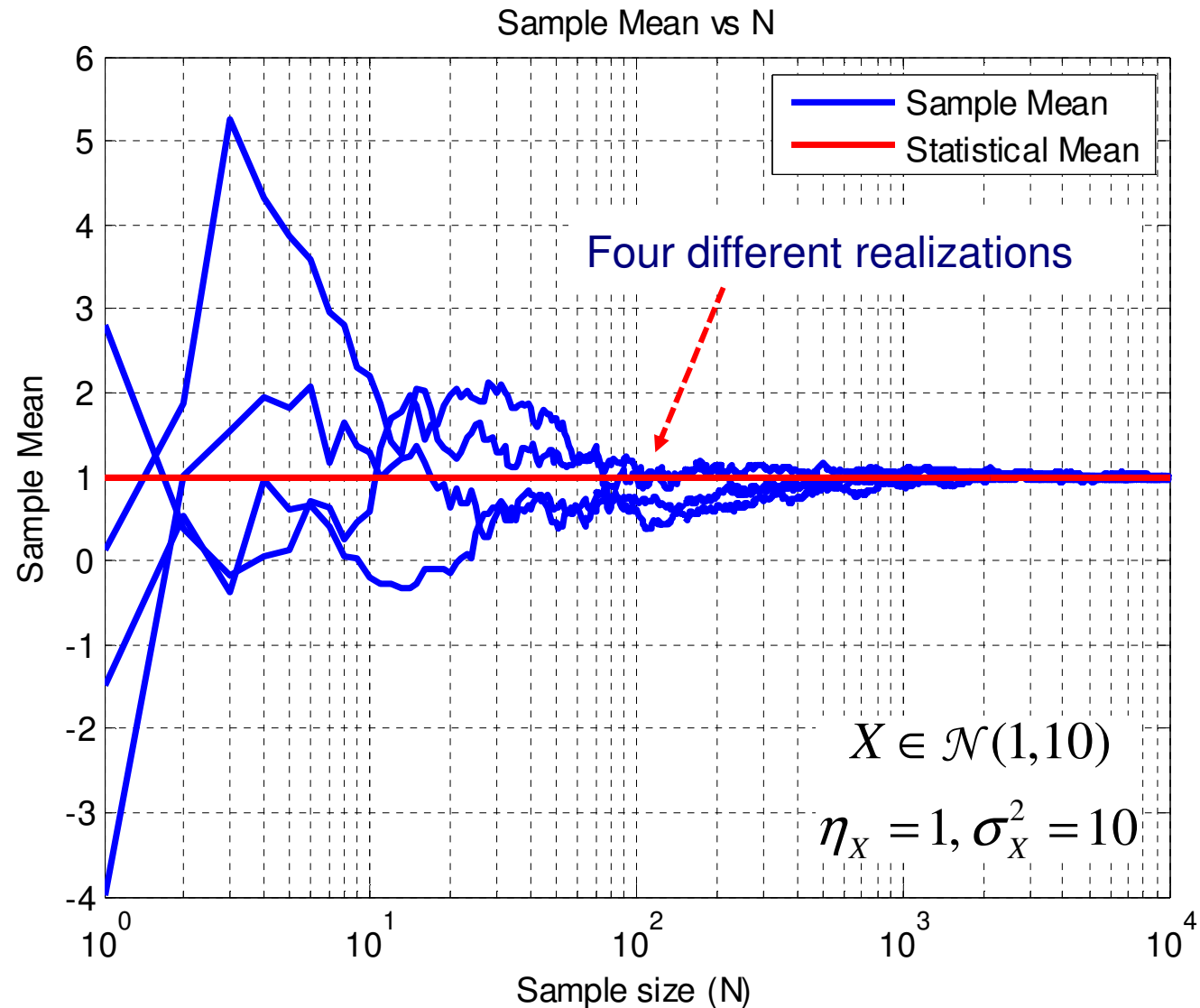
$$\lim_{N \rightarrow \infty} MSE\{\hat{\eta}_X\} = \lim_{N \rightarrow \infty} \frac{\sigma_X^2}{N} = 0 \rightarrow \text{consistent}$$

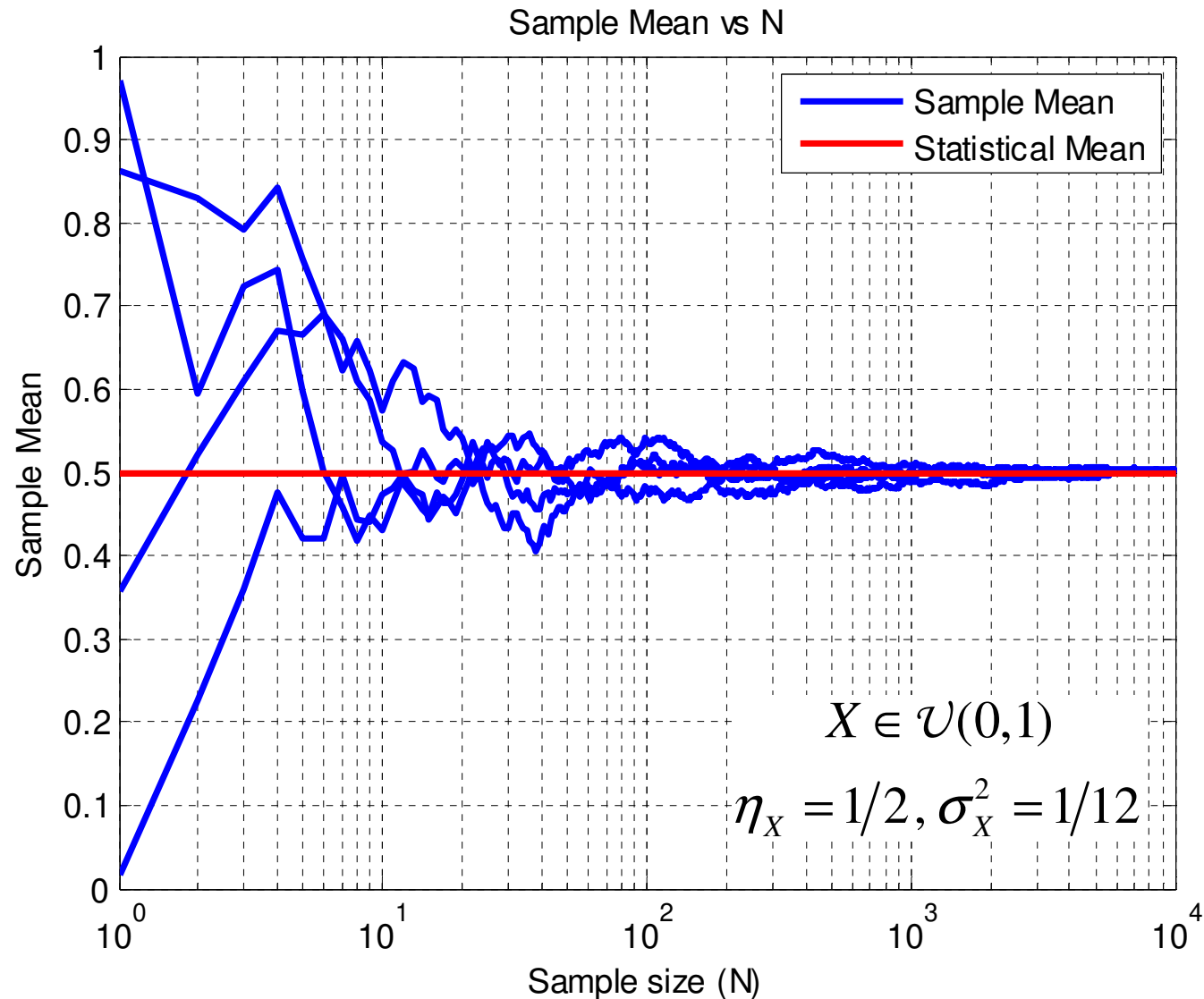
- In summary, we proved that if the variance of X is finite and the observed data are IID, the **Sample Mean** converges in mean square sense, and so also in probability, to the statistical mean of X .
- We will analyze separately the case of dependent data.
- The **mean square value** of the relative error is:

$$\left\{ \begin{array}{l} \varepsilon_r \triangleq \frac{\varepsilon}{\eta_X} = \frac{\eta_X - \hat{\eta}_X}{\eta_X} \\ \sigma_{\varepsilon_r}^2 = E \left\{ \left(\frac{\eta_X - \hat{\eta}_X}{\eta_X} \right)^2 \right\} = \frac{E \{ (\eta_X - \hat{\eta}_X)^2 \}}{\eta_X^2} = \frac{MSE \{ \hat{\eta}_X \}}{\eta_X^2} = \frac{\sigma_X^2}{\eta_X^2 N} = \frac{1}{\gamma N} \end{array} \right.$$

where $\gamma \triangleq \frac{\eta_X^2}{\sigma_X^2}$ is a sort of Signal to Noise power Ratio (SNR)

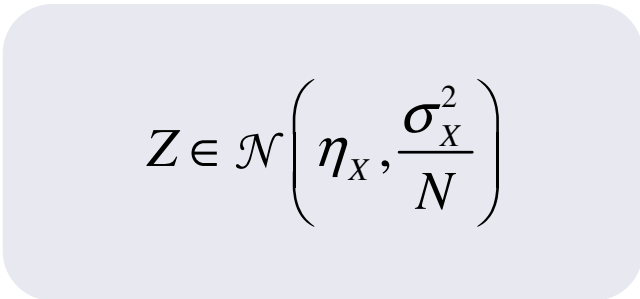








- If X is a Gaussian r.v., since the **Sample Mean** Z is a linear combination of Gaussian r.v.'s, it is also Gaussian distributed:

$$X \in \mathcal{N}(\eta_X, \sigma_X^2) \Rightarrow Z = \frac{1}{N} \sum_{i=1}^N X_i \text{ is a Gaussian r.v.}$$

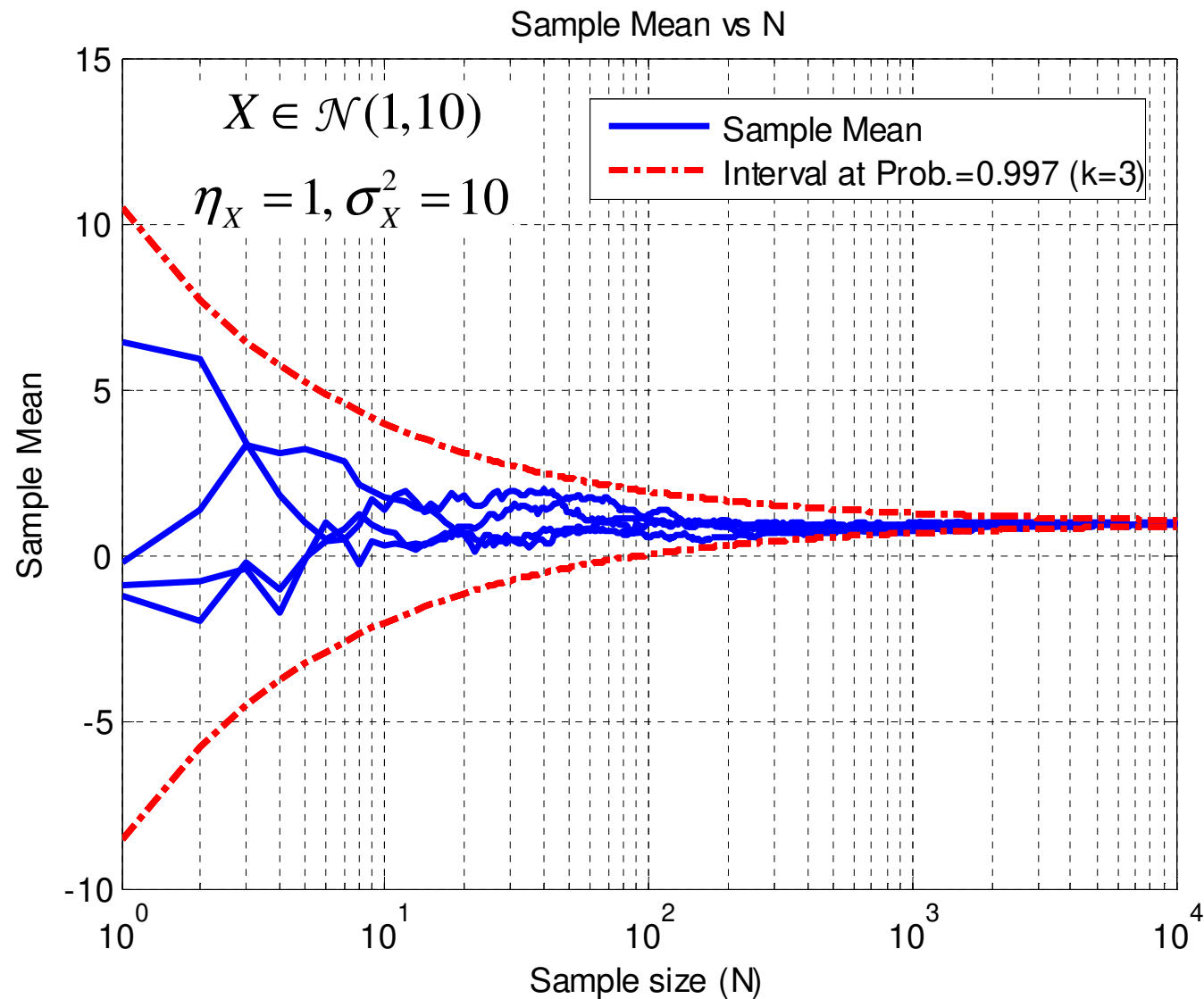

$$Z \in \mathcal{N}\left(\eta_X, \frac{\sigma_X^2}{N}\right)$$

- Even if X is not Gaussian distributed, thanks to the **Central-Limit Theorem**, for large N we can approximate Z as a Gaussian random variable.

- Even if X is not Gaussian distributed, thanks to the **Central-Limit Theorem**, Z is asymptotically Gaussian distributed, i.e. for $N \rightarrow \infty$.
- Hence, if N is large, even if finite, the approximation is accurate and we can rely on the Gaussian assumption to calculate the probabilities of events of interest:


$$\hat{\eta}_X = Z = \frac{1}{N} \sum_{i=1}^N X_i \stackrel{a}{\in} \mathcal{N}\left(\eta_X, \frac{\sigma_X^2}{N}\right)$$
$$\Pr\left\{\left|\hat{\eta}_X - \eta_X\right| \leq k \frac{\sigma_X}{\sqrt{N}}\right\} = 1 - 2Q(k) \cong \begin{cases} 0.683 & k=1 \\ 0.956 & k=2 \\ 0.997 & k=3 \end{cases}$$


- The length of this interval goes to 0 when N goes to infinity.



- How to choose N ? One possible criterion is to choose the minimum value which guarantees that the standard deviation of the relative error is lower than a certain value, e.g. 10%:


$$\sigma_{\varepsilon_r} = \sqrt{E \left\{ \left(\frac{\eta_X - \hat{\eta}_X}{\eta_X} \right)^2 \right\}} = \sqrt{\frac{1}{N\gamma}} \leq 0.1 \Rightarrow N \geq \frac{100}{\gamma}$$

- The value of N is inversely proportional to the ratio γ , which is a sort of Signal to Noise power Ratio (SNR):

$$\gamma \triangleq \frac{\eta_X^2}{\sigma_X^2}$$

- If we know the pdf of Z , we can calculate the minimum value of N such that the absolute value of the relative error is lower of a given value δ , e.g. $\delta=0.1$ i.e. 10%, with a desired probability α .

$$\hat{\eta}_X \overset{a}{\in} \mathcal{N}\left(\eta_X, \frac{\sigma_X^2}{N}\right)$$


$$\Pr\left\{\left|\frac{\hat{\eta}_X - \eta_X}{\eta_X}\right| \leq \delta\right\} = \Pr\{|\hat{\eta}_X - \eta_X| \leq \delta\eta_X\} = \alpha$$

$$\Pr\{|\hat{\eta}_X - \eta_X| \leq \delta\eta_X\} = 1 - 2Q\left(\frac{\delta\eta_X}{\sigma_X/\sqrt{N}}\right) = \alpha$$

$$\Rightarrow \frac{\delta\eta_X\sqrt{N}}{\sigma_X} = Q^{-1}\left(\frac{1-\alpha}{2}\right) \Rightarrow N = \left[\frac{\sigma_X}{\delta\eta_X} Q^{-1}\left(\frac{1-\alpha}{2}\right)\right]^2$$

$$N = \left[\frac{\sigma_X}{\delta \eta_X} Q^{-1} \left(\frac{1-\alpha}{2} \right) \right]^2$$

- Set e.g. $\delta=0.1$ and $\alpha=0.95$:

$$N = \left[\frac{\sigma_X}{0.1 \eta_X} Q^{-1} \left(\frac{1-0.95}{2} \right) \right]^2 = \frac{10^2 \sigma_X^2}{\eta_X^2} [1.96]^2 = 384 \cdot \frac{\sigma_X^2}{\eta_X^2} = \frac{384}{\gamma}$$

- The value of N is again inversely proportional to the ratio γ .
- Clearly, if this ratio does not return an integer number, we choose the smallest integer greater than this ratio.

- Let us now investigate the behavior of the **Sample Mean** in the case of dependent data, i.e. the N observed samples are identically distributed but not independent:

$$\hat{\eta}_X = \frac{1}{N} \sum_{i=1}^N X_i$$

$$E\{\hat{\eta}_X\} = E\left\{\frac{1}{N} \sum_{i=1}^N X_i\right\} = \frac{1}{N} \sum_{i=1}^N E\{X_i\} = \eta_X$$

- Hence, independently from the fact that $\{X_i\}$ are independent or not, the **Sample Mean** is an **unbiased estimator** of the statistical mean of X .
- Let us now calculate the MSE: since the estimator is unbiased, the MSE coincides with the variance of the estimator.

- We already derived the following result for the MSE (p. 59):

$$MSE\{\hat{\eta}_X\} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N c_{X_i X_j} = \frac{\sigma_X^2}{N} + \frac{1}{N^2} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N c_{X_j X_i}$$

$$c_{X_i X_j} \triangleq E\{(X_i - \eta_X)(X_j - \eta_X)\}, \quad i, j = 1, 2, \dots, N$$

- The **Sample Mean** is a **consistent** estimator if and only if (*iff*):


$$\lim_{N \rightarrow \infty} MSE\{\hat{\eta}_X\} = \lim_{N \rightarrow \infty} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N c_{X_i X_j} = \lim_{N \rightarrow \infty} \left(\frac{\sigma_X^2}{N} + \frac{1}{N^2} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N c_{X_j X_i} \right) = 0$$

- When does this happen? **The independence assumption is not necessary.**

- We proved that if the $\{X_i\}$ are IID we have that:

$$\langle X_i \rangle = \frac{1}{N} \sum_{i=1}^N X_i \xrightarrow{N \rightarrow \infty} E\{X_i\}, \quad \text{if } \text{var}\{X_i\} < +\infty$$

- How can we extend this result if we want to estimate other **statistical indexes**, e.g. the *variance*, the *skewness* and the *kurtosis*?
- If we define a new r.v. $S=g(X)$ and we apply the previous results to S we get:


$$\langle S_i \rangle = \frac{1}{N} \sum_{i=1}^N S_i \xrightarrow{N \rightarrow \infty} E\{S_i\}, \quad \text{if } \text{var}\{S_i\} < +\infty$$
$$\langle g(X_i) \rangle = \frac{1}{N} \sum_{i=1}^N g(X_i) \xrightarrow{N \rightarrow \infty} E\{g(X_i)\}, \quad \text{if } \text{var}\{g(X_i)\} < +\infty$$

- In the case we want to estimate the **variance**:

$$\sigma_X^2 = E\{(X - \eta_X)^2\} \Rightarrow S \triangleq g(X) = (X - \eta_X)^2$$

- The previous result suggests to estimate the variance as follows:

$$\hat{\sigma}_X^2 = \langle (X_i - \eta_X)^2 \rangle = \frac{1}{N} \sum_{i=1}^N (X_i - \eta_X)^2$$

- The problem is that also the mean is usually unknown, so we should replace the mean with its estimate, i.e. the Sample Mean:

$$\hat{\sigma}_X^2 = \langle (X_i - \hat{\eta}_X)^2 \rangle = \frac{1}{N} \sum_{i=1}^N (X_i - \hat{\eta}_X)^2$$

$$\text{where } \hat{\eta}_X = \frac{1}{N} \sum_{k=1}^N X_k$$

- The estimator of the variance we derived is called **Sample Variance**.
- To derive the bias, it is useful to express it in a different way:

$$\begin{aligned}\hat{\sigma}_X^2 &= \frac{1}{N} \sum_{i=1}^N (X_i - \hat{\eta}_X)^2 \\ \text{Sample Variance} &= \frac{1}{N} \sum_{i=1}^N X_i^2 - 2 \cdot \frac{1}{N} \sum_{i=1}^N X_i \hat{\eta}_X + \frac{1}{N} \sum_{i=1}^N \hat{\eta}_X^2 \\ &= \frac{1}{N} \sum_{i=1}^N X_i^2 - 2\hat{\eta}_X \cdot \frac{1}{N} \sum_{i=1}^N X_i + \hat{\eta}_X^2 \\ &= \frac{1}{N} \sum_{i=1}^N X_i^2 - \hat{\eta}_X^2\end{aligned}$$

Sample Power Sample Mean squared

■ We derive now the mean value and the variance of the **Sample Variance**, and then its bias and MSE.

■ **Assumption:** the observed data $\{X_j\}$ are IID.

■ Mean value and variance of the Sample Variance:

$$\hat{\sigma}_X^2 = \frac{1}{N} \sum_{i=1}^N X_i^2 - \hat{\eta}_X^2$$

$$E\{\hat{\sigma}_X^2\} = E\left\{\frac{1}{N} \sum_{i=1}^N X_i^2 - \hat{\eta}_X^2\right\} = E\left\{\frac{1}{N} \sum_{i=1}^N X_i^2\right\} - E\{\hat{\eta}_X^2\}$$

$$= \frac{1}{N} \sum_{i=1}^N E\{X_i^2\} - \left(\text{var}\{\hat{\eta}_X\} + \left(E\{\hat{\eta}_X\}\right)^2\right)$$

$$= \frac{1}{N} \sum_{i=1}^N (\sigma_X^2 + \eta_X^2) - \left(\frac{\sigma_X^2}{N} + \eta_X^2\right) = \sigma_X^2 - \frac{\sigma_X^2}{N} = \left(1 - \frac{1}{N}\right) \sigma_X^2$$

- The **Sample Variance** is a **biased** estimator of the variance:

$$E\{\hat{\sigma}_X^2\} = \sigma_X^2 \left(1 - \frac{1}{N}\right) \neq \sigma_X^2$$

$$b\{\hat{\sigma}_X^2\} \triangleq E\{\varepsilon\} = E\{\sigma_X^2 - \hat{\sigma}_X^2\} = \sigma_X^2 - \sigma_X^2 \left(1 - \frac{1}{N}\right) = \frac{\sigma_X^2}{N} \neq 0$$

- The **Sample Variance** is **biased** but **asymptotically unbiased**:

$$\lim_{N \rightarrow \infty} b\{\hat{\sigma}_X^2\} = \lim_{N \rightarrow \infty} \frac{\sigma_X^2}{N} = 0$$

■ The derivation of the **variance** of the **Sample Variance** is quite tedious, but not conceptually difficult. We skip here the derivation and we provide only the final result:

$$\hat{\sigma}_X^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \hat{\eta}_X)^2$$

$$E\{\hat{\sigma}_X^2\} = \sigma_X^2 \left(1 - \frac{1}{N}\right)$$

$$\begin{aligned} \text{var}\{\hat{\sigma}_X^2\} &= E\left\{\left(\hat{\sigma}_X^2 - E\{\hat{\sigma}_X^2\}\right)^2\right\} = E\left\{\left(\frac{1}{N} \sum_{i=1}^N (X_i - \hat{\eta}_X)^2 - \sigma_X^2 \left(1 - \frac{1}{N}\right)\right)^2\right\} \\ &= \frac{\mu_X(4) - \sigma_X^4}{N} - \frac{2(\mu_X(4) - 2\sigma_X^4)}{N^2} + \frac{\mu_X(4) - 3\sigma_X^4}{N^3} \end{aligned}$$

$\mu_X(4) \triangleq E\{(X - \eta_X)^4\}$ is the 4th order central moment of X

- If the 4th order central moment is finite, i.e. $\mu_X(4) < +\infty$

$$\lim_{N \rightarrow \infty} \text{var}\{\hat{\sigma}_X^2\} = \lim_{N \rightarrow \infty} \left[\frac{\mu_X(4) - \sigma_X^4}{N} - \frac{2(\mu_X(4) - 2\sigma_X^4)}{N^2} + \frac{\mu_X(4) - 3\sigma_X^4}{N^3} \right] = 0$$

$$b\{\hat{\sigma}_X^2\} = \frac{\sigma_X^2}{N}$$

$$MSE\{\hat{\sigma}_X^2\} = \text{var}\{\hat{\sigma}_X^2\} + \left(b\{\hat{\sigma}_X^2\}\right)^2 \xrightarrow{N \rightarrow \infty} 0$$


- In summary, under the IID assumption, the **Sample Variance** is a **consistent** estimator of the variance.
- It is **biased** but **asymptotically unbiased**.

- If X is a **Gaussian** random variable:

$$X \in \mathcal{N}(\eta_X, \sigma_X^2) \rightarrow \mu_X(4) = 3\sigma_X^4$$

$$\text{var}\{\hat{\sigma}_X^2\} = \frac{2\sigma_X^4}{N} - \frac{2\sigma_X^4}{N^2} = \frac{2\sigma_X^4}{N} \left(1 - \frac{1}{N}\right) \stackrel{N \gg 1}{\cong} \frac{2\sigma_X^4}{N}$$

$$b\{\hat{\sigma}_X^2\} = \frac{\sigma_X^2}{N}$$


$$\begin{aligned} \text{MSE}\{\hat{\sigma}_X^2\} &= \text{var}\{\hat{\sigma}_X^2\} + \left(b\{\hat{\sigma}_X^2\}\right)^2 = \frac{2\sigma_X^4}{N} \left(1 - \frac{1}{N}\right) + \frac{\sigma_X^4}{N^2} \\ &= \frac{2\sigma_X^4}{N} \left(1 - \frac{1}{2N}\right) \stackrel{N \gg 1}{\cong} \frac{2\sigma_X^4}{N} \end{aligned}$$

- How can we get an **unbiased** estimator of the variance?

$$\tilde{\sigma}_X^2 = \alpha \hat{\sigma}_X^2 = \frac{\alpha}{N} \sum_{i=1}^N (X_i - \hat{\eta}_X)^2, \quad \text{where } \alpha \text{ is such that } E\{\tilde{\sigma}_X^2\} = \sigma_X^2$$

$$E\{\tilde{\sigma}_X^2\} = E\{\alpha \hat{\sigma}_X^2\} = \alpha \sigma_X^2 \left(1 - \frac{1}{N}\right) = \sigma_X^2 \quad \Rightarrow \quad \alpha = \left(1 - \frac{1}{N}\right)^{-1} = \frac{N}{N-1}$$

$$\tilde{\sigma}_X^2 = \frac{N}{N-1} \hat{\sigma}_X^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \hat{\eta}_X)^2 \quad [\text{unbiased estimator of } \sigma_X^2]$$

- We observe that if $N \gg 1$ there is no significant difference between the two estimators (biased and unbiased). The difference may be significant for small N .

$$\tilde{\sigma}_X^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \hat{\eta}_X)^2 = \frac{N}{N-1} \hat{\sigma}_X^2$$

$$\begin{aligned} MSE\{\tilde{\sigma}_X^2\} &= \text{var}\{\tilde{\sigma}_X^2\} + \left(b\{\tilde{\sigma}_X^2\}\right)^2 \\ &= \text{var}\left\{\frac{N}{N-1} \hat{\sigma}_X^2\right\} = \left(\frac{N}{N-1}\right)^2 \text{var}\{\hat{\sigma}_X^2\} \\ &= \left(\frac{N}{N-1}\right)^2 \frac{2\sigma_X^4}{N} \left(1 - \frac{1}{N}\right) = \frac{2\sigma_X^4}{N-1} \end{aligned}$$

$$MSE\{\tilde{\sigma}_X^2\} = \frac{2\sigma_X^4}{N-1} = \frac{2\sigma_X^4}{N} \left(1 + \frac{1}{N-1}\right) > \frac{2\sigma_X^4}{N} \left(1 - \frac{1}{2N}\right) = MSE\{\hat{\sigma}_X^2\}$$

■ The unbiased estimator has higher MSE than the biased one, hence it is less efficient.

- In summary, we proved that:

$$\langle g(X_i) \rangle = \frac{1}{N} \sum_{i=1}^N g(X_i) \xrightarrow{N \rightarrow \infty} E\{g(X_i)\} \quad \text{if } \text{var}\{g(X_i)\} < +\infty$$

- This suggests that if we want to estimate a parameter $\delta = E\{g(X)\}$, and we can observe N independent realizations of the r.v. X , a consistent estimator of δ is given by the **Sample Estimator**:

$$\hat{\delta} = \langle g(X_i) \rangle = \frac{1}{N} \sum_{i=1}^N g(X_i) \xrightarrow{N \rightarrow \infty} \delta = E\{g(X_i)\} \quad \text{if } \text{var}\{g(X_i)\} < +\infty$$

- The assumption that the N data $\{X_i\}$ are independent is a sufficient condition for consistency, but not necessary.
- We can use Sample Estimators to estimate any ordinary/central moment of X .

$$\delta = E\{g(X_i)\}, \quad \text{var}\{g(X_i)\} < +\infty, \quad \hat{\delta} = \frac{1}{N} \sum_{i=1}^N g(X_i)$$

$$E\{\hat{\delta}\} = E\left\{\frac{1}{N} \sum_{i=1}^N g(X_i)\right\} = \frac{1}{N} \sum_{i=1}^N E\{g(X_i)\} = \frac{1}{N} \sum_{i=1}^N \delta = \delta$$

$$\text{var}\{\hat{\delta}\} = \text{var}\left\{\frac{1}{N} \sum_{i=1}^N g(X_i)\right\} = \frac{1}{N^2} \sum_{i=1}^N \text{var}\{g(X_i)\} = \frac{\text{var}\{g(X_i)\}}{N}$$

$$\lim_{N \rightarrow \infty} \text{MSE}\{\hat{\delta}\} = \lim_{N \rightarrow \infty} \text{var}\{\hat{\delta}\} = \lim_{N \rightarrow \infty} \frac{\text{var}\{g(X_i)\}}{N} = 0$$

■ where we used the fact that the N observed data $\{X_i\}$ are Independent and Identically Distributed (IID).

■ **Ordinary moments:** $m_X(n) \triangleq E\{X^n\} \Rightarrow \hat{m}_X(n) = \frac{1}{N} \sum_{i=1}^N X_i^n$

■ **Central moments:** $\mu_X(n) \triangleq E\{(X - \eta_X)^n\} \Rightarrow \hat{\mu}_X(n) = \frac{1}{N} \sum_{i=1}^N (X_i - \hat{\eta}_X)^n$


■ This introduces a bias!


■ It is important to note that in the Gaussian case (but it is true in general):

$$\frac{2\sigma_X^4}{N} \left(1 - \frac{1}{2N}\right) > \frac{\sigma_X^2}{N} \rightarrow \text{MSE}(\hat{\sigma}_X^2) > \text{MSE}(\hat{\eta}_X)$$

■ Given the data size N , the MSE is larger for higher-order moments. In other words, the higher is the order n of the moment we want to estimate and the more data we need to achieve a desired estimation accuracy.

Definition of *Skewness* and *Kurtosis*:



$$\gamma_{3X} \triangleq \frac{\mu_X(3)}{(\mu_X(2))^{3/2}}$$



$$\gamma_{4X} \triangleq \frac{\mu_X(4)}{(\mu_X(2))^2} - 3$$

where $\mu_X(2) = \sigma_X^2$ is the variance of X

- The *skewness* is the normalized 3rd order central moment of a r.v., it is a measure of *asymmetry* of a distribution.
- All symmetric distributions have zero *skewness*, so the *skewness* of a Gaussian r.v. is zero.
- If the *skewness* is different from zero (neglecting the measurement errors), the observed data cannot be Gaussian.

Definition of *Skewness* and *Kurtosis*:



$$\gamma_{3X} \triangleq \frac{\mu_X(3)}{(\mu_X(2))^{3/2}}$$



$$\gamma_{4X} \triangleq \frac{\mu_X(4)}{(\mu_X(2))^2} - 3$$

$$X \in \mathcal{N}(\eta_X, \sigma_X^2) \rightarrow \mu_X(4) = 3\sigma_X^4 \rightarrow \frac{\mu_X(4)}{(\mu_X(2))^2} = 3 \rightarrow \gamma_{4X} = 0$$

- The *kurtosis* is the normalized 4th order central moment of a r.v. scaled in such a way that it is zero for a Gaussian r.v.
- It is a measure of *flatness* of the distribution w.r.t. the Gaussian one.
- If the *kurtosis* is positive the pdf has longer tails than the Gaussian one, so the probability to observe large values is higher than in the Gaussian case. These kind of pdf's are called **heavy-tailed distributions** (e.g. the Laplace, the Weibull, the log-normal, the K, the *t*-distribution are heavy-tailed).

Definition of *Skewness* and *Kurtosis*:


$$\gamma_{3X} \triangleq \frac{\mu_X(3)}{(\mu_X(2))^{3/2}}$$


$$\gamma_{4X} \triangleq \frac{\mu_X(4)}{(\mu_X(2))^2} - 3$$

- The *skewness* and *kurtosis* are indicators of non Gaussianity of the data: if they are different from zero the data cannot be Gaussian distributed.
- We can build statistical tests based on the sample estimates of the *skewness* and *kurtosis* to decide whether the data are Gaussian distributed or not.

- To better understand the meaning of the kurtosis, let us address the following problem: we want to select one of the following three symmetric distributions to model our data:

$$\text{Gaussian: } f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\eta_X)^2}{2\sigma^2}}$$

$$\text{Laplace: } f_Y(y) = \frac{\lambda}{2} e^{-\lambda|y-\eta_Y|}$$

$$\text{Uniform: } f_Z(z) = \frac{1}{\alpha} \text{rect}\left(\frac{z-\eta_Z}{\alpha}\right)$$

- We estimate the mean from the data, e.g. by using the Sample Mean estimator, and we find that the mean is 0 (neglecting the estimation error).
- All the three distributions can model zero mean data by setting:

$$\eta_X = \eta_Y = \eta_Z = 0$$

- Hence, the problem becomes to select one of the following three zero mean distributions:

$$\text{Gaussian: } f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

$$\text{Laplace: } f_Y(y) = \frac{\lambda}{2} e^{-\lambda|y|}$$

$$\text{Uniform: } f_Z(z) = \frac{1}{\alpha} \text{rect}\left(\frac{z}{\alpha}\right)$$

- The next step is to estimate the variance of the data, e.g. by using the Sample Variance estimator.

- We cannot discriminate among the three pdf's based on an estimate of the variance, since the three pdf's can have any variance depending on the value of their parameters.
- For example, if the sample estimate of the variance is equal to 1, the parameters of the three distributions are as follows:

$$\sigma_X^2 \triangleq \mu_X(2) = \sigma^2 = 1 \rightarrow \sigma = 1$$

$$\sigma_Y^2 \triangleq \mu_Y(2) = \frac{2}{\lambda^2} = 1 \rightarrow \lambda = \sqrt{2}$$

$$\sigma_Z^2 \triangleq \mu_Z(2) = \frac{\alpha^2}{12} = 1 \rightarrow \alpha = \sqrt{12}$$

- We need to estimate the (central) moments of order higher than 2.
- The central moments of the three distributions are given by:

$$\mu_X(n) = \begin{cases} 0 & n \text{ odd} \\ \sigma^n (n-1)!! & n \text{ even} \end{cases}$$

$$\mu_Y(n) = \begin{cases} 0 & n \text{ odd} \\ \frac{n!}{\lambda^n} & n \text{ even} \end{cases}$$

$$\mu_Z(n) = \begin{cases} 0 & n \text{ odd} \\ \frac{1}{(n+1)} \left(\frac{\alpha}{2} \right)^n & n \text{ even} \end{cases}$$

■ Since the pdf's are symmetric, all the central moments of odd order are 0, so also the *skewness* is 0.

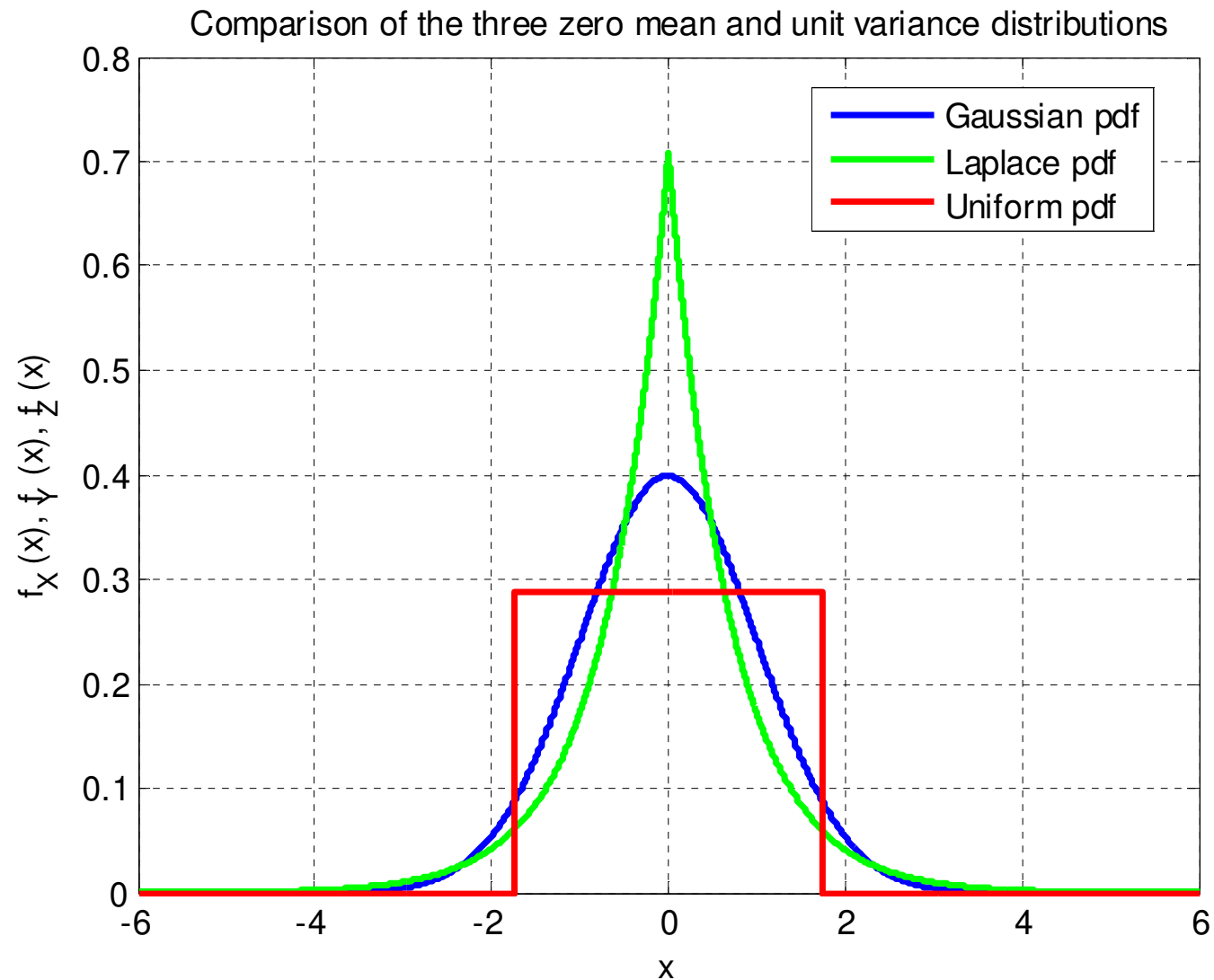
■ As a consequence, we expect to obtain an estimate of the skewness close to 0, whatever of the three is the right pdf model.

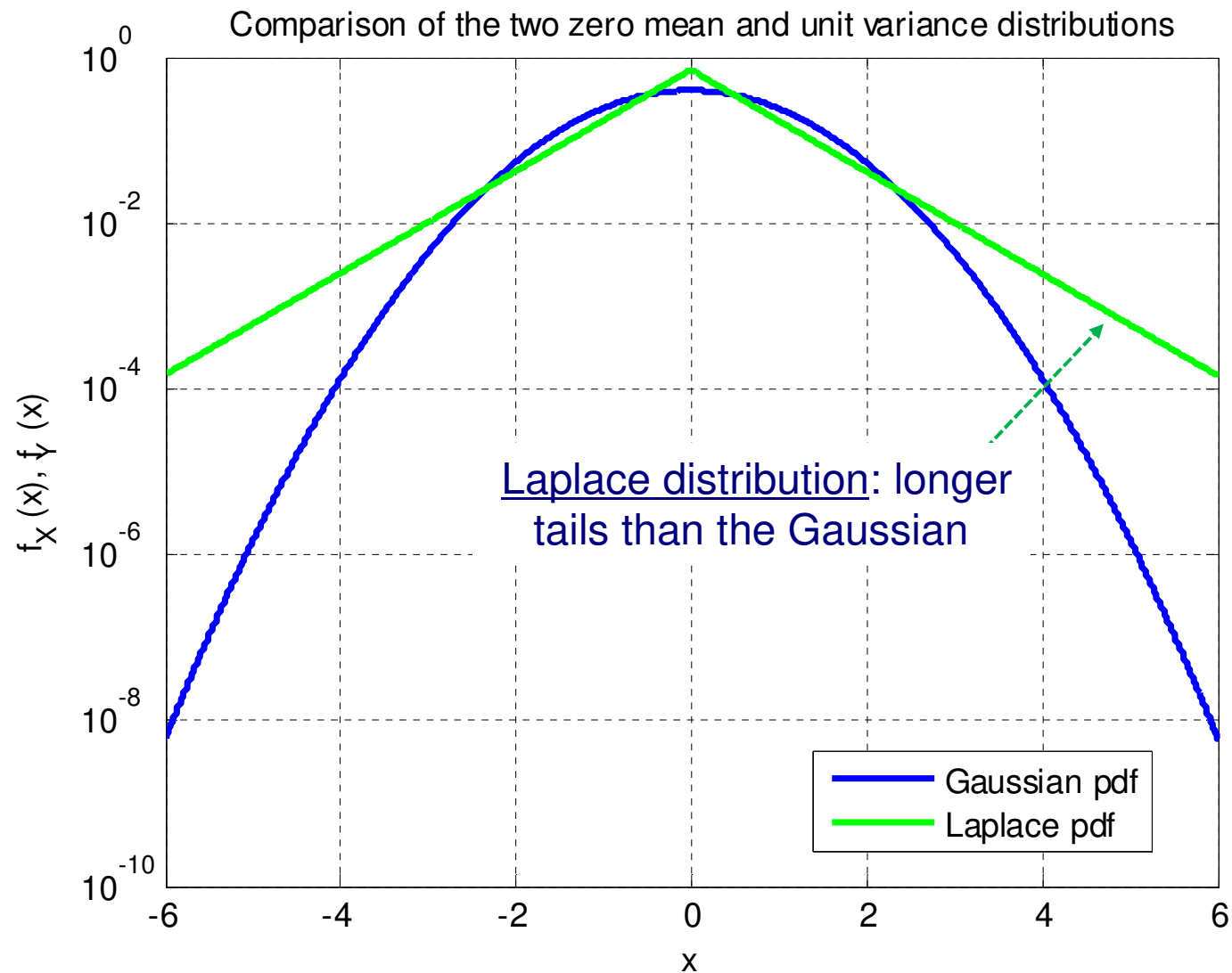
■ Hence, we cannot discriminate among the three pdf's based on an estimate of the *skewness*.

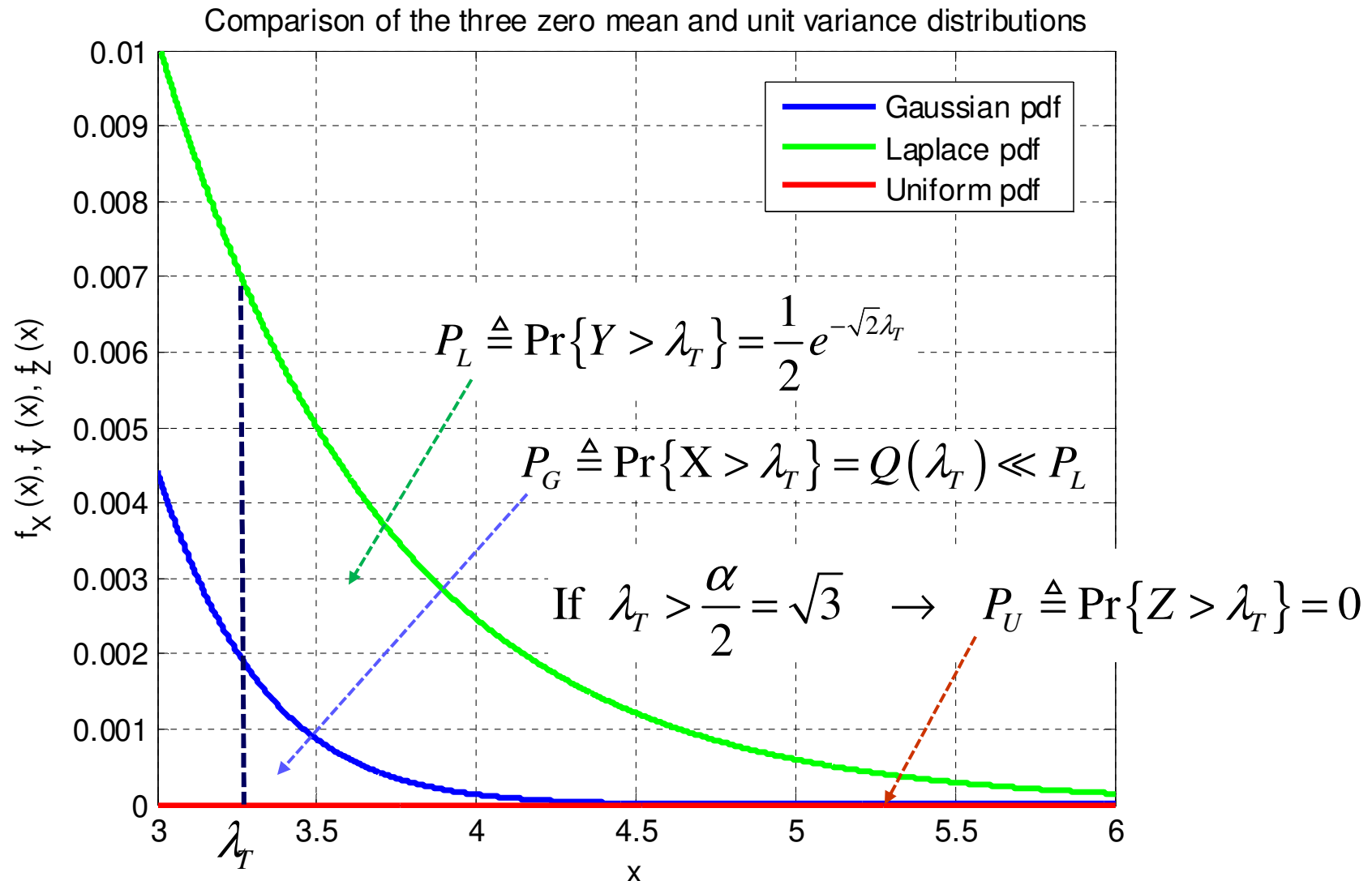
- We should consider the **kurtosis**, which is given by:

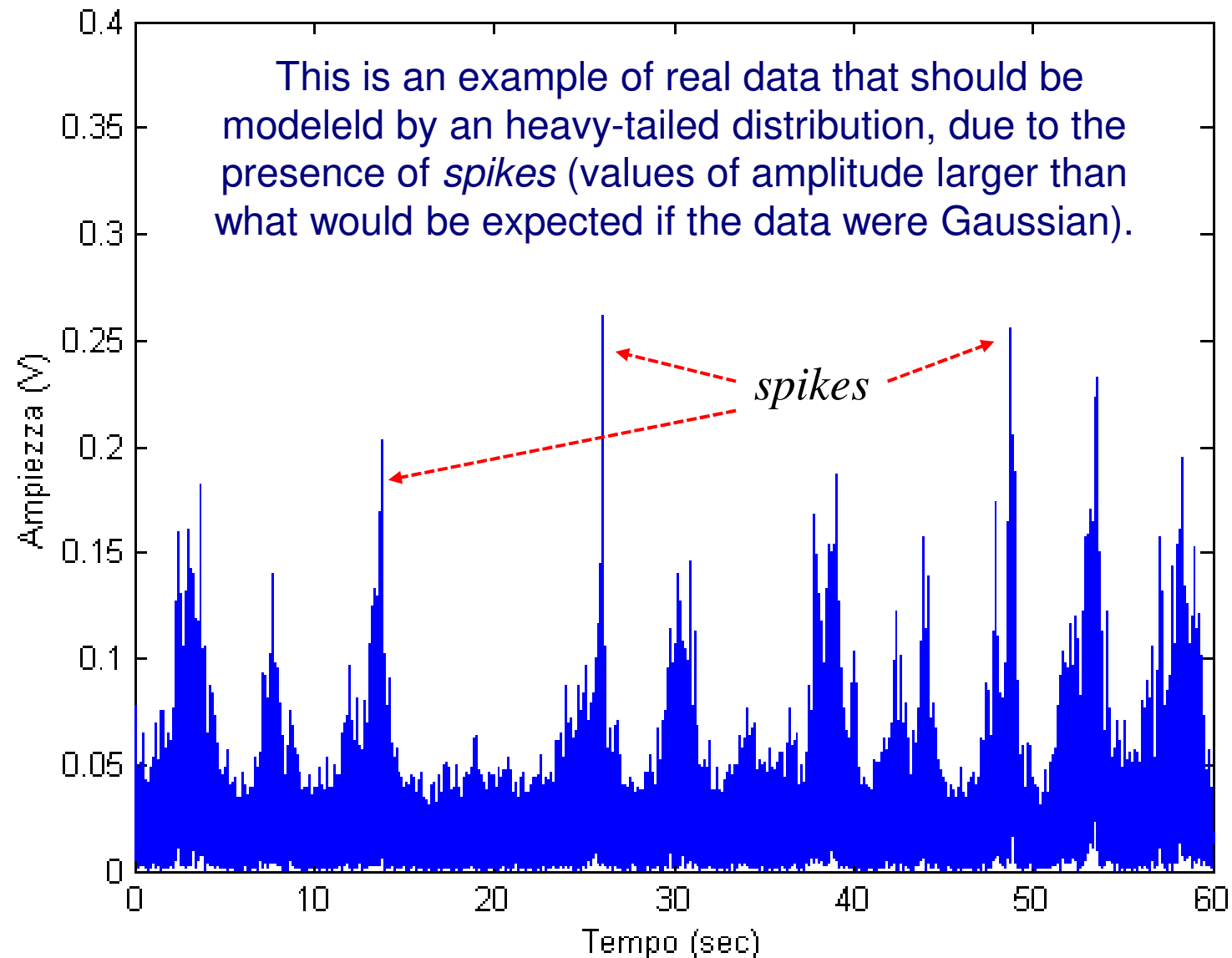
$$\left. \begin{aligned} \mu_X(4) &= 3\sigma^4 \\ \mu_Y(4) &= \frac{24}{\lambda^4} \\ \mu_Z(4) &= \frac{\alpha^4}{80} \end{aligned} \right\} \Rightarrow \left\{ \begin{aligned} \gamma_{4X} &= \frac{3\sigma^4}{(\sigma^2)^2} - 3 = 0 \\ \gamma_{4Y} &= \frac{24/\lambda^4}{(2/\lambda^2)^2} - 3 = 3 \\ \gamma_{4Z} &= \frac{\alpha^4/80}{(\alpha^2/12)^2} - 3 = -1.2 \end{aligned} \right.$$

- Note that the **kurtosis** is an adimensional number, which does not depend on the variance of the pdf (i.e. on the specific values of the parameters of the pdf's).
- The Laplace distribution has **kurtosis**>0. Hence, it is flatter than the Gaussian pdf, i.e. it has longer tails. The reverse is true for the uniform pdf.
- The **kurtosis** can be used to discriminate amongst the three distributions!









■ In summary, we can decide which pdf model to adopt for our observed data, based on the estimate of the first few moments, e.g. the first four, as follows:

■ **Mean value:** $\eta_X \triangleq E\{X\} \Rightarrow \hat{\eta}_X = \frac{1}{N} \sum_{i=1}^N X_i$

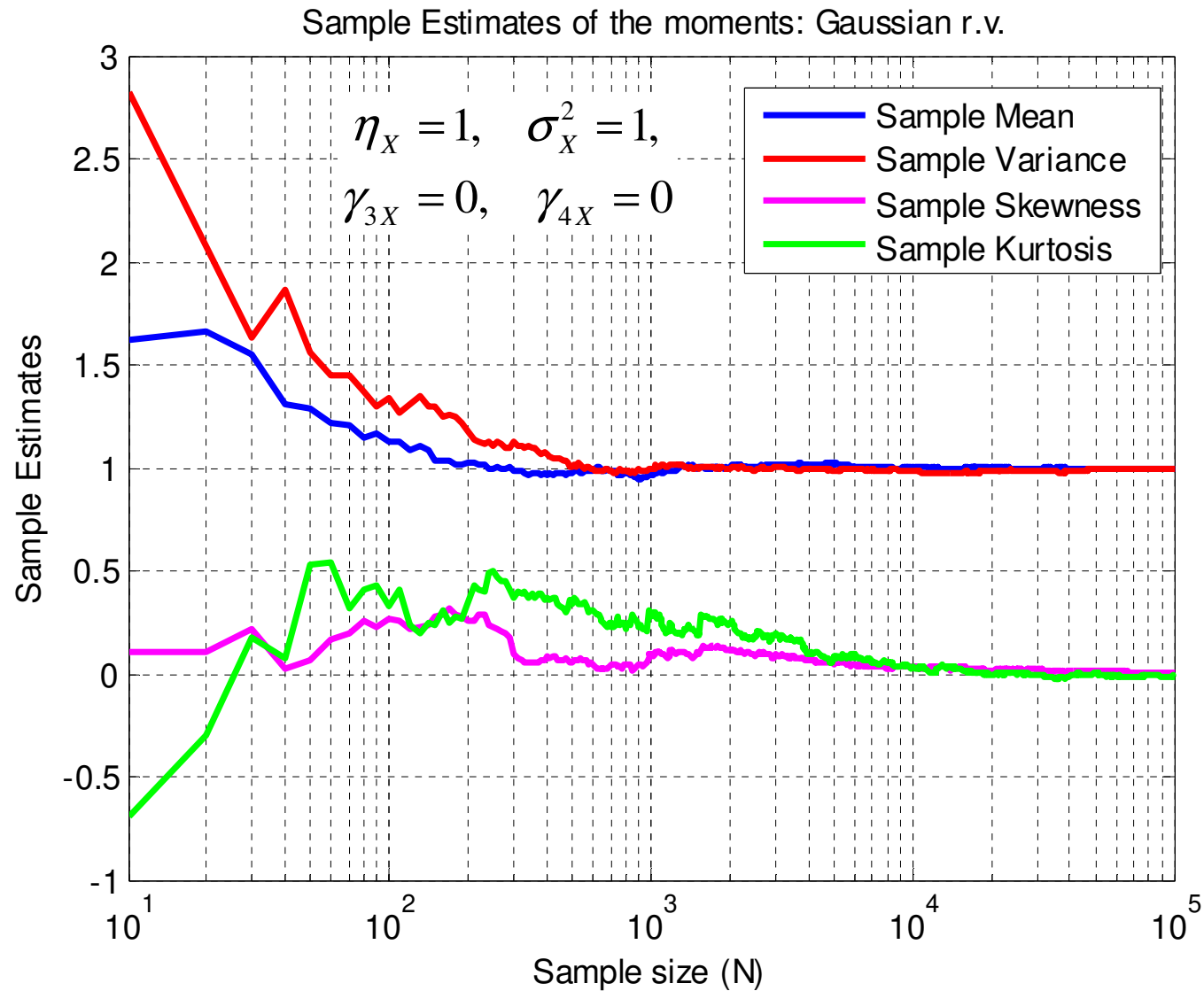
■ **Variance:** $\sigma_X^2 \triangleq \mu_X(2) = E\{(X - \eta_X)^2\} \Rightarrow \hat{\sigma}_X^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \hat{\eta}_X)^2$

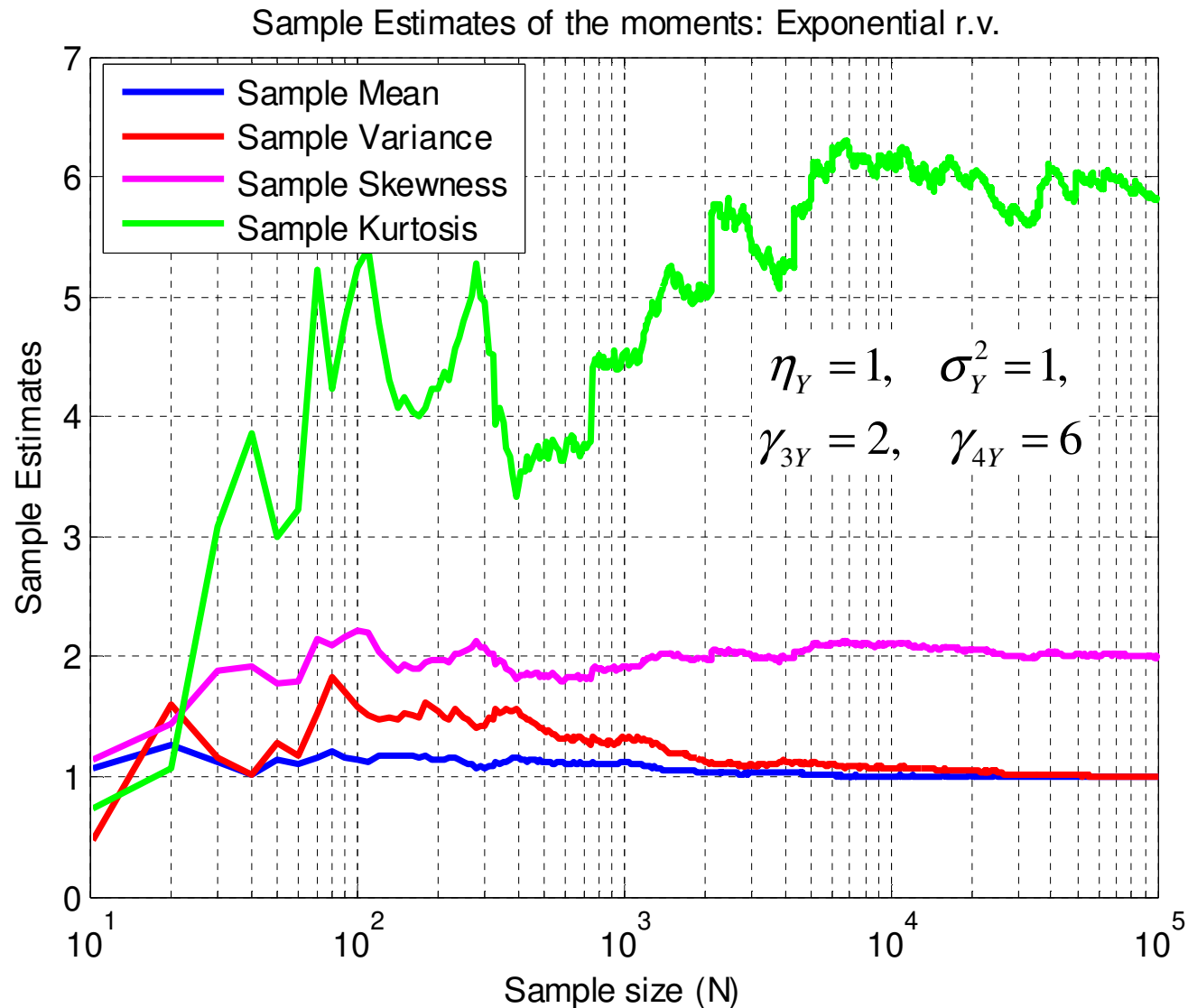
■ **Skewness:**

$$\gamma_{3X} \triangleq \frac{\mu_X(3)}{(\mu_X(2))^{3/2}} = \frac{E\{(X - \eta_X)^3\}}{(\mu_X(2))^{3/2}} \Rightarrow \hat{\gamma}_{3X} = \frac{\frac{1}{N} \sum_{i=1}^N (X_i - \hat{\eta}_X)^3}{(\hat{\sigma}_X^2)^{3/2}}$$

■ **Kurtosis:**

$$\gamma_{4X} \triangleq \frac{\mu_X(4)}{(\mu_X(2))^2} - 3 = \frac{E\{(X - \eta_X)^4\}}{(\mu_X(2))^2} - 3 \Rightarrow \hat{\gamma}_{4X} = \frac{\frac{1}{N} \sum_{i=1}^N (X_i - \hat{\eta}_X)^4}{(\hat{\sigma}_X^2)^2} - 3$$



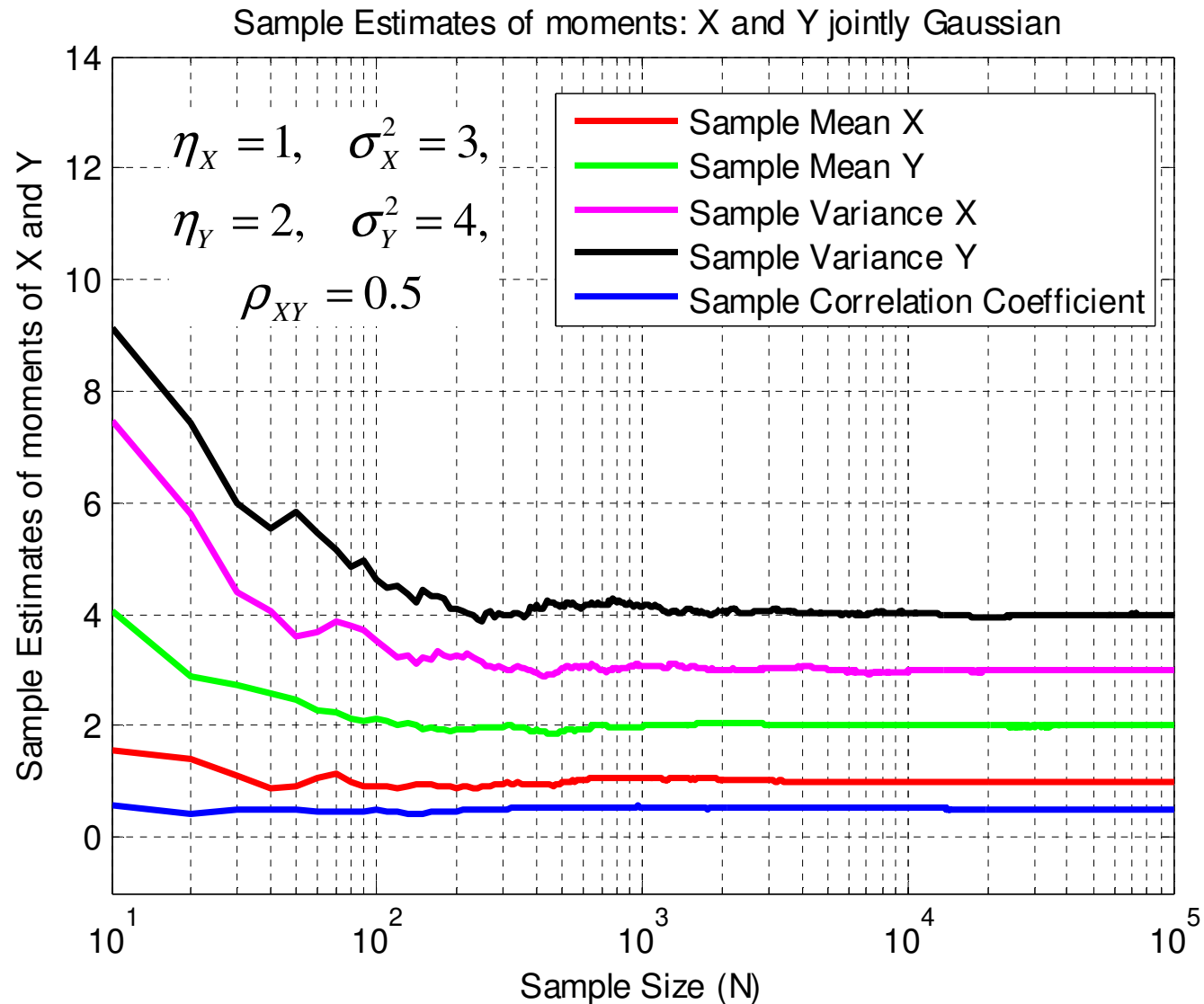


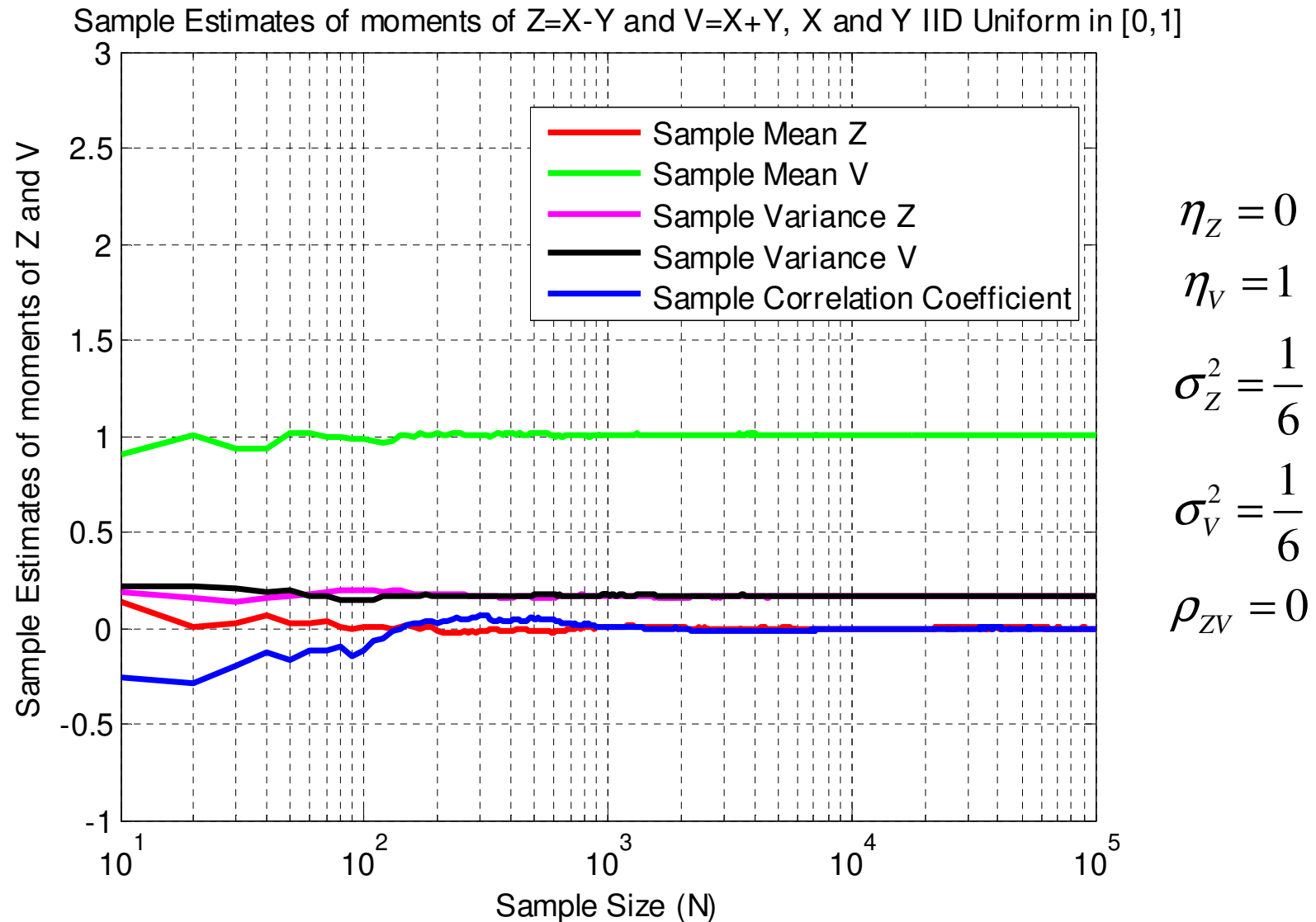
- The same approach can be applied to estimate the **correlation coefficient** between two random variables X and Y , if we can observe N independent realizations of X and Y , i.e. we can collect N couples $\{(X_i, Y_i)\}$:

$$\rho_{XY} = \frac{c_{XY}}{\sigma_X \sigma_Y} = \frac{E\{(X - \eta_X)(Y - \eta_Y)\}}{\sqrt{\sigma_X^2 \sigma_Y^2}} \Rightarrow \hat{\rho}_{XY} = \frac{\frac{1}{N} \sum_{i=1}^N (X_i - \hat{\eta}_X)(Y_i - \hat{\eta}_Y)}{\sqrt{\hat{\sigma}_X^2 \hat{\sigma}_Y^2}}$$

$$\hat{\eta}_X = \frac{1}{N} \sum_{i=1}^N X_i, \quad \hat{\sigma}_X^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \hat{\eta}_X)^2$$

$$\hat{\eta}_Y = \frac{1}{N} \sum_{i=1}^N Y_i, \quad \hat{\sigma}_Y^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{\eta}_Y)^2$$





- The goodness of an estimator is usually measured by its **bias** and **mean square error (MSE)**.
- An alternative way to measure the goodness of an estimator is to calculate its **Confidence Interval (CI)**.
- Confidence intervals were introduced to statistics by Jerzy Neyman in a paper published in 1937.
- **Problem:** Once we have an estimate of our parameter, that is always affected by an error, how far from the estimate can be the true value?
- In other words: In which interval centered around the estimate the true value is confined with a given desired probability α ?
- Note that since the estimate is a realization of a random variable, the event *{the true value is included in an interval $2\Delta_\alpha$ centered around the estimate}* can be verified only with a given probability α .

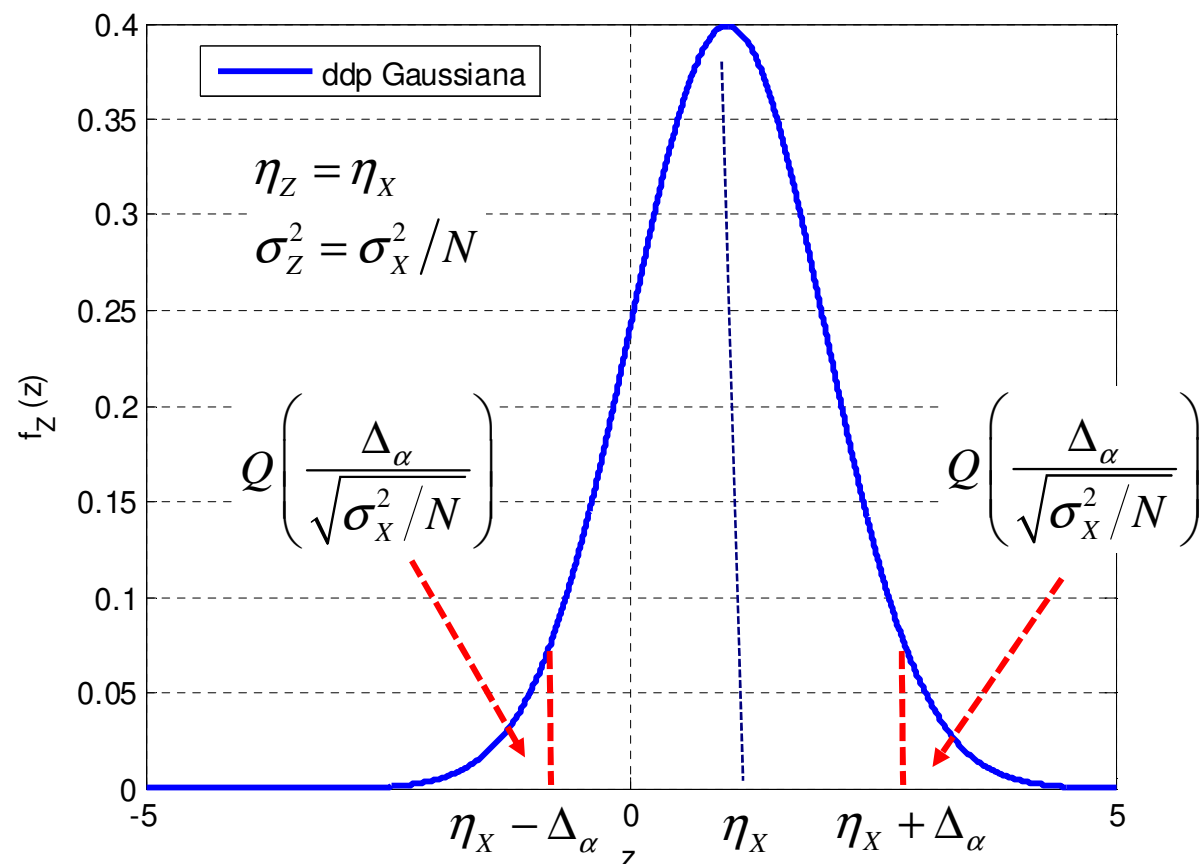
- Note that it is not a point estimate problem but an interval estimate problem.
- The concept of **Confidence Interval** can be applied to the measure of any deterministic parameter (probability, mean value, variance, *skewness*, *kurtosis*, etc.)
- Let us consider the case when the parameter to be estimated is the mean value of the random variable X .
- The confidence interval Δ_α for the estimate of the mean value of X is defined by the following relationship:

$$\Pr\left\{\left|\hat{\eta}_X - \eta_X\right| \leq \Delta_\alpha\right\} = \Pr\left\{\eta_X - \Delta_\alpha \leq \hat{\eta}_X \leq \eta_X + \Delta_\alpha\right\} = \alpha$$

- Assuming that the estimate Z (i.e. the sample mean) is Gaussian distributed, thanks to the Central-Limit Theorem (CLT), we obtain:

$$\begin{aligned}\Pr\left\{\left|\hat{\eta}_X - \eta_X\right| \leq \Delta_\alpha\right\} &= \Pr\left\{\left|Z - \eta_X\right| \leq \Delta_\alpha\right\} \\&= \Pr\left\{\eta_X - \Delta_\alpha \leq Z \leq \eta_X + \Delta_\alpha\right\} \\&= 1 - \Pr\left\{Z < \eta_X - \Delta_\alpha\right\} - \Pr\left\{Z > \eta_X + \Delta_\alpha\right\} \\&= 1 - Q\left(\frac{\Delta_\alpha}{\sqrt{\sigma_X^2/N}}\right) - Q\left(\frac{\Delta_\alpha}{\sqrt{\sigma_X^2/N}}\right) \\&= 1 - 2Q\left(\frac{\Delta_\alpha}{\sqrt{\sigma_X^2/N}}\right) \\&= \alpha\end{aligned}$$

$$\Pr\{Z < \eta_X - \Delta_\alpha\} = Q\left(\frac{\Delta_\alpha}{\sqrt{\sigma_X^2/N}}\right), \quad \Pr\{Z > \eta_X + \Delta_\alpha\} = Q\left(\frac{\Delta_\alpha}{\sqrt{\sigma_X^2/N}}\right)$$



- Solving the equation with respect to Δ_α :

$$1 - 2Q\left(\frac{\Delta_\alpha}{\sqrt{\sigma_X^2/N}}\right) = \alpha \quad \Rightarrow \quad Q\left(\frac{\Delta_\alpha}{\sqrt{\sigma_X^2/N}}\right) = \frac{1-\alpha}{2}$$

$$\frac{\Delta_\alpha}{\sqrt{\sigma_X^2/N}} = Q^{-1}\left(\frac{1-\alpha}{2}\right) \quad \Rightarrow \quad \Delta_\alpha = \frac{\sigma_X}{\sqrt{N}} Q^{-1}\left(\frac{1-\alpha}{2}\right)$$

- For example, if we want to determine the **confidence interval at 95%**, i.e. the interval centered around the estimate within which the true value will be included with probability $\alpha=0.95$:

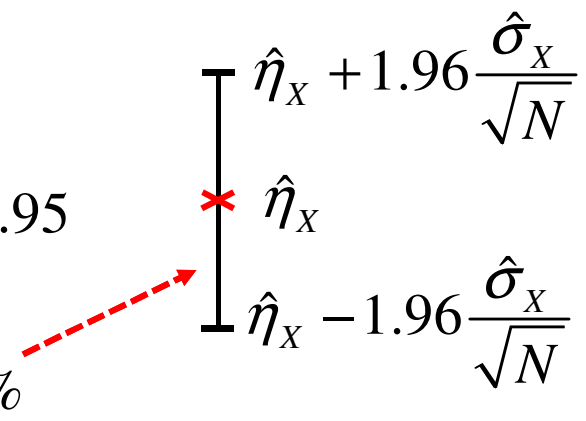
$$\alpha = 0.95 \quad \Rightarrow \quad \Delta_\alpha = \frac{\sigma_X}{\sqrt{N}} Q^{-1}\left(\frac{1-0.95}{2}\right) = 1.96 \frac{\sigma_X}{\sqrt{N}}$$

- The **Confidence Interval at 95%** of the **Sample Mean** estimator of the mean value of X can be expressed by the following equation:

$$\Pr \left\{ \left| \hat{\eta}_X - \eta_X \right| \leq 1.96 \frac{\sigma_X}{\sqrt{N}} \right\}$$

$$= \Pr \left\{ \hat{\eta}_X - 1.96 \frac{\sigma_X}{\sqrt{N}} \leq \eta_X \leq \hat{\eta}_X + 1.96 \frac{\sigma_X}{\sqrt{N}} \right\} = 0.95$$

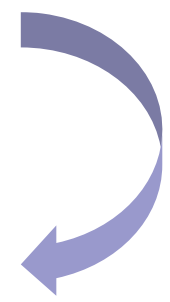
- In practice, also the variance of X is unknown, hence we will use the above equation replacing the unknown variance by the **Sample Variance estimate**:

$$\Pr \left\{ \hat{\eta}_X - 1.96 \frac{\hat{\sigma}_X}{\sqrt{N}} \leq \eta_X \leq \hat{\eta}_X + 1.96 \frac{\hat{\sigma}_X}{\sqrt{N}} \right\} \cong 0.95$$


η_X is somewhere here with prob. of 95%

- The **Confidence Interval at 95%** of the estimator "*Frequency of Occurrence*" of the probability of an event p is:

$$\hat{p} = F = \frac{1}{N} \sum_{i=1}^N X_i, \quad \eta_X = E\{X\} = p, \quad \sigma_X^2 = \text{var}\{X\} = p(1-p)$$


$$\Pr \left\{ \hat{p} - 1.96 \sqrt{\frac{p(1-p)}{N}} \leq p \leq \hat{p} + 1.96 \sqrt{\frac{p(1-p)}{N}} \right\} = 0.95$$


- Since p is unknown, we can proceed in two ways:
 - (i) replace p by its estimate;
 - (ii) consider the **worst case scenario**, that is when the variance is maximum, i.e. $p=1/2$.

- (i) Replace p by its estimate:

$$\Pr \left\{ \hat{p} - 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \leq p \leq \hat{p} + 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \right\} \cong 0.95$$

- (ii) Consider the worst case scenario, i.e. $p=1/2$:

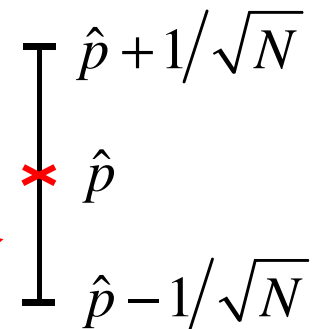
$$\Pr \left\{ \hat{p} - 1.96 \sqrt{\frac{1}{4N}} \leq p \leq \hat{p} + 1.96 \sqrt{\frac{1}{4N}} \right\} \cong 0.95$$

$$\Pr \left\{ \hat{p} - \frac{1}{\sqrt{N}} \leq p \leq \hat{p} + \frac{1}{\sqrt{N}} \right\} \cong 0.95$$
$$\frac{1.96}{\sqrt{4}} = 0.980 \cong 1$$

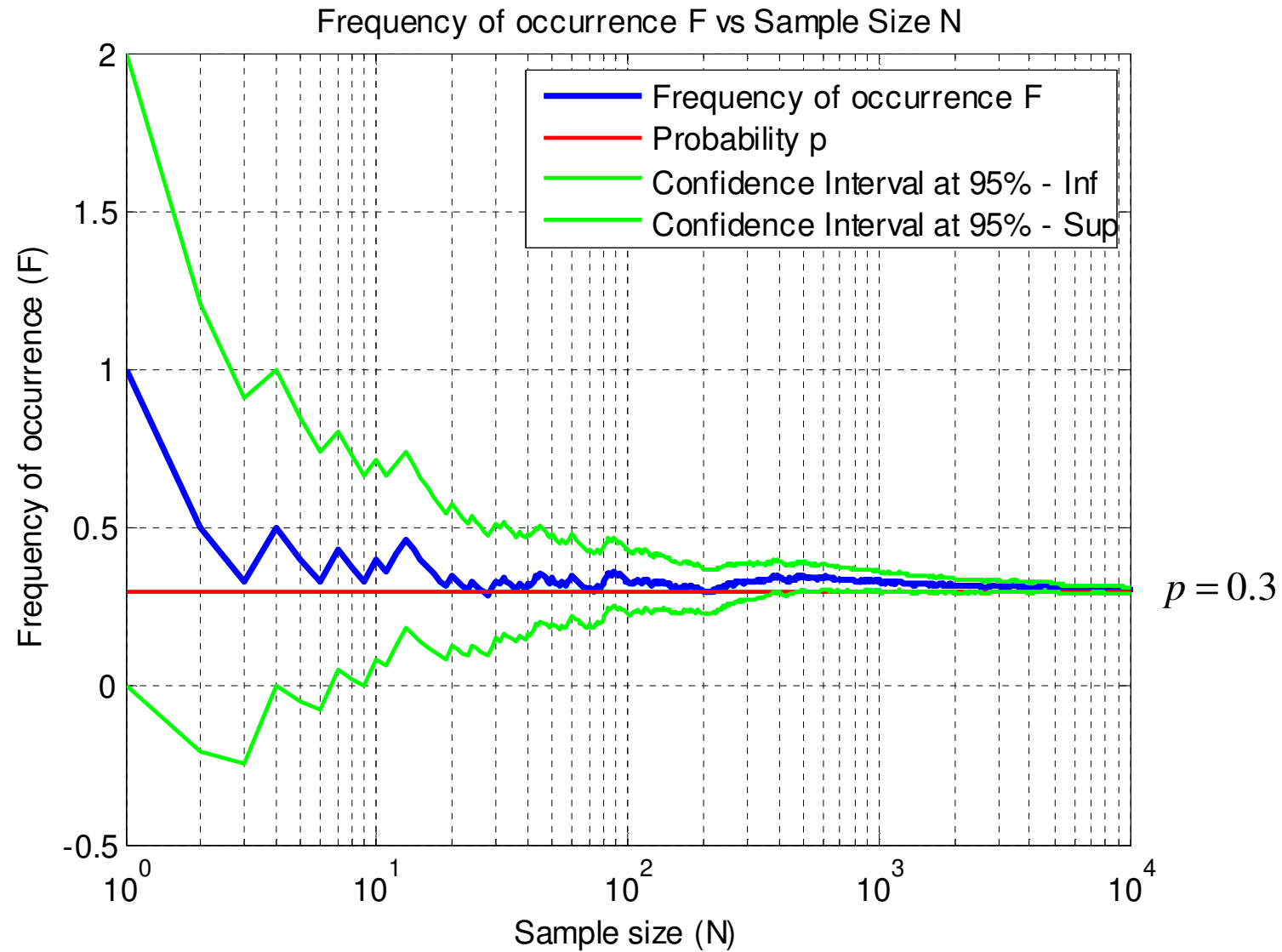
■ In summary: once we get an estimate of p , we can state that the true value of p (which is unknown!) is confined within an interval of amplitude $2\Delta_\alpha$ centered around the estimate with probability α .

■ In the case $\alpha=0.95$, the **Confidence Interval** is $\Delta_{0.95} = \frac{1}{\sqrt{N}}$, so its size tends to 0 when N goes to infinity.

$$\Pr\left\{\hat{p} - \frac{1}{\sqrt{N}} \leq p \leq \hat{p} + \frac{1}{\sqrt{N}}\right\} \cong 0.95$$

p is somewhere here
with prob. of 95%







Measure of the Mean Value and of the Autocorrelation Function of a w.s.s. Random Process

Measure of the Mean Value of a w.s.s. Random Process¹¹⁵

■ Assume $X[n]$ is wide sense stationary (w.s.s.) process, how can we estimate the mean value and the autocorrelation function given N samples of the process: $\{x[0], x[1], \dots, x[N-1]\}$?

■ Let's start from the mean value. We can apply the **Sample Mean estimator**:

$$\eta_x \triangleq E\{X[n]\} \Rightarrow \hat{\eta}_x = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$$

■ The estimator is unbiased, in fact:

$$b\{\hat{\eta}_x\} = E\{\eta_x - \hat{\eta}_x\} = \eta_x - \frac{1}{N} \sum_{n=0}^{N-1} E\{X[n]\} = \eta_x - \frac{1}{N} \sum_{n=0}^{N-1} \eta_x = 0$$

■ then the MSE and the variance coincide.

■ The variance can be derived as follows:

$$\begin{aligned}\sigma_{\varepsilon}^2 &= \text{var}\{\hat{\eta}_X\} = \text{MSE}\{\hat{\eta}_X\} = E\{(\eta_X - \hat{\eta}_X)^2\} \\&= E\left\{\left(\eta_X - \frac{1}{N} \sum_{n=0}^{N-1} X[n]\right)^2\right\} = E\left\{\left(\frac{1}{N} \sum_{n=0}^{N-1} \eta_X - \frac{1}{N} \sum_{n=0}^{N-1} X[n]\right)^2\right\} \\&= \frac{1}{N^2} E\left\{\left(\sum_{n=0}^{N-1} (\eta_X - X[n])\right)^2\right\} \\&= \frac{1}{N^2} E\left\{\sum_{n=0}^{N-1} (\eta_X - X[n]) \sum_{k=0}^{N-1} (\eta_X - X[k])\right\} \\&= \frac{1}{N^2} E\left\{\sum_{n=0}^{N-1} \sum_{k=0}^{N-1} (\eta_X - X[k])(\eta_X - X[n])\right\}\end{aligned}$$

Measure of the Mean Value of a w.s.s. Random Process¹¹⁷

$$\begin{aligned}\sigma_{\varepsilon}^2 &= \frac{1}{N^2} E \left\{ \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} (\eta_X - X[k])(\eta_X - X[n]) \right\} \\ &= \frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} E \{ (\eta_X - X[k])(\eta_X - X[n]) \} \\ &= \frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} C_X[n-k] = \frac{1}{N^2} \sum_{m=-(N-1)}^{N-1} (N-|m|) C_X[m] \\ &= \frac{1}{N} \sum_{m=-(N-1)}^{N-1} \left(1 - \frac{|m|}{N} \right) C_X[m]\end{aligned}$$

Triangular weighting window

**Auto-Covariance function
of the process**

■ where we used the “double sum” formula:

$$\sum_{n=0}^{N-1} \sum_{k=0}^{N-1} f[n-k] = \sum_{m=-(N-1)}^{N-1} (N-|m|) f[m]$$

- The **necessary and sufficient condition** in order the **sample mean** to be a **consistent** estimator of the statistical mean of the process is:

$$\lim_{N \rightarrow \infty} \sigma_{\varepsilon}^2 = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{m=-(N-1)}^{N-1} \left(1 - \frac{|m|}{N} \right) C_X[m] = 0$$

- If the process has finite memory, that is $C_X[m]$ has a strictly limited length (e.g. the MA processes), or goes to zero “relatively” fast (e.g. for the stable AR or ARMA processes) then the estimator is consistent, i.e. its MSE goes to zero when the number of data N goes to infinity.

- When does this happen? **The independence assumption is not necessary.**
- **Example 1:** $X[n]$ is an autoregressive process of order 1, AR(1)

$$C_X[m] = R_X[m] = E\{X[n]X[n+m]\} = \sigma_X^2 \rho^{|m|}, \quad \text{where } |\rho| < 1$$

$$MSE\{\hat{\eta}_X\} = \frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} C_X[n-k] = \frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} \sigma_X^2 \rho^{|n-k|}$$

$$= \frac{\sigma_X^2}{N^2} \sum_{m=-(N-1)}^{N-1} (N - |m|) \rho^{|m|} \leq \frac{\sigma_X^2}{N^2} \sum_{m=-(N-1)}^{N-1} N |\rho|^{|m|}$$

$$= \frac{\sigma_X^2}{N} \sum_{m=-(N-1)}^{N-1} |\rho|^{|m|} = \frac{\sigma_X^2}{N} \left(2 \sum_{m=0}^{N-1} |\rho|^m - 1 \right) = \frac{\sigma_X^2}{N} \left(2 \cdot \frac{1 - |\rho|^N}{1 - |\rho|} - 1 \right)$$

- In this case, the **Sample Mean** is a consistent estimator:

$$\lim_{N \rightarrow \infty} MSE\{\hat{\eta}_X\} = \lim_{N \rightarrow \infty} \left[\frac{\sigma_X^2}{N} \left(\frac{1 + |\rho| - 2|\rho|^N}{1 - |\rho|} \right) \right] = 0, \quad \text{if } |\rho| \neq 1$$

- When $\rho=0$, i.e. we go back to the case of independent data:

$$MSE\{\hat{\eta}_X\} \Big|_{\rho=0} = \frac{\sigma_X^2}{N}$$

- Hence, the assumption of data independence is a sufficient condition for the Sample Mean to be consistent, but it is not a necessary condition [*Remember that instead the assumption that the data are identically distributed is necessary*].

- **Example 2:** observed data = constant signal + independent noise

$$X[n] = S + W[n]$$

$$S \in \mathcal{N}(0, \sigma_s^2), \quad W[n] \in \mathcal{N}(0, \sigma_w^2), \quad \{W[n]\}_{n=0}^{N-1} \text{ IID},$$

$W[n]$ white noise, S and $W[n]$ independent

$$\eta_x[n] = E\{X[n]\} = E\{S + W[n]\} = 0,$$

$$\gamma \triangleq \frac{E\{S^2\}}{E\{W^2[n]\}} = \frac{\sigma_s^2}{\sigma_w^2} \quad [\text{SNR}]$$

■ Autocovariance function:

$$\begin{aligned}C_X[m] &= R_X[m] = E\{X[n]X[n+m]\} = E\{(S + W[n])(S + W[n+m])\} \\&= E\{S^2 + SW[n] + SW[n+m] + W[n]W[n+m]\} \\&= E\{S^2\} + E\{W[n]W[n+m]\} \\&= \sigma_S^2 + \sigma_W^2 \delta[m]\end{aligned}$$

■ Variance:

$$\sigma_X^2 = C_X[0] = \sigma_S^2 + \sigma_W^2$$

■ Normalized autocovariance function:

$$\rho_X[m] \triangleq \frac{\text{cov}\{X[n], X[n+m]\}}{\sqrt{\text{var}\{X[n]\} \text{var}\{X[n+m]\}}} = \frac{C_X[m]}{C_X[0]} = \frac{\sigma_S^2 + \sigma_W^2 \delta[m]}{\sigma_X^2}$$

$$= \frac{\sigma_S^2}{\sigma_S^2 + \sigma_W^2} + \frac{\sigma_W^2}{\sigma_S^2 + \sigma_W^2} \delta[m]$$

$$\forall m \neq 0 \Rightarrow \rho \triangleq \rho_X[m] = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_W^2} = \frac{\gamma}{\gamma + 1}$$

ρ is the correlation coefficient between $X[n]$ and $X[n+m]$

$$\forall m \neq 0 \Rightarrow C_X[m] = \rho(\sigma_S^2 + \sigma_W^2) = \rho\sigma_X^2$$

■ Hence, in this case we get:

$$\begin{aligned}
 MSE\{\hat{\eta}_X\} &= \frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} C_X[n-k] = \frac{C_X[0]}{N} + \frac{1}{N^2} \sum_{n=1}^{N-1} \sum_{\substack{k=1 \\ n \neq k}}^{N-1} C_X[n-k] \\
 &= \frac{\sigma_X^2}{N} + \frac{1}{N^2} \sum_{n=1}^{N-1} \sum_{\substack{k=1 \\ n \neq k}}^{N-1} \rho \sigma_X^2 = \frac{\sigma_X^2}{N} + \frac{N(N-1)}{N^2} \rho \sigma_X^2 \\
 &= \frac{\sigma_X^2}{N} + \rho \sigma_X^2 \left(1 - \frac{1}{N}\right)
 \end{aligned}$$

$$\lim_{N \rightarrow \infty} MSE\{\hat{\eta}_X\} = \lim_{N \rightarrow \infty} \left[\frac{\sigma_X^2}{N} + \rho \sigma_X^2 \left(1 - \frac{1}{N}\right) \right] = \rho \sigma_X^2 = \sigma_s^2 \neq 0 \quad \text{if } \rho \neq 0$$

■ In this case, the **Sample Mean** is not a consistent estimator of the mean η_X .

- If the samples $X[n]$ are uncorrelated, then:

$$C_X[m] = \sigma_X^2 \delta[m] \Rightarrow \sigma_\varepsilon^2 = \frac{1}{N} \sum_{m=-(N-1)}^{N-1} \left(1 - \frac{|m|}{N}\right) \sigma_X^2 \delta[m] = \frac{\sigma_X^2}{N}$$

so the MSE decreases as $1/N$.

- Note that if the data are (positively) correlated, then the MSE is larger.

■ **Estimate of the Autocorrelation function (ACF)**, given N samples of the process: $\{x[0], x[1], \dots, x[N-1]\}$.

■ Again, we can apply the **Sample Estimators**:

$$R_X[m] \triangleq E\{X[n]X^*[n+m]\} \Rightarrow \hat{R}_X[m] = \frac{1}{N} \sum_{n=0}^{N-1} x[n]x^*[n+m]$$

■ Note that, since we collect data only from discrete-time 0 to $N-1$, the maximum value for $n+m$ is $N-1$, so the estimator should be expressed as:

$$\tilde{R}_X[m] = \frac{1}{N-m} \sum_{n=0}^{N-m-1} x[n]x^*[n+m], \quad m = 0, 1, 2, \dots, N-1$$

■ where $N-m$ is the number of products that we average. This is called the (unbiased) **Sample Autocorrelation Function**.

- The estimate for negative values of m are obtained by resorting to the Hermitian symmetry property of the ACF:

$$\tilde{R}_X[m] = \tilde{R}_X^*[-m], \quad m = -1, -2, \dots, -(N-1)$$

- This estimator is called the **unbiased Sample ACF**, since it is unbiased for lags $m=0, 1, 2, \dots, N-1$:

$$\begin{aligned} E\{\tilde{R}_X[m]\} &= \frac{1}{N-m} \sum_{n=0}^{N-m-1} E\{X[n]X^*[n+m]\} \\ &= \frac{1}{N-m} \sum_{n=0}^{N-m-1} R_X[m] = R_X[m], \quad \text{for } 0 \leq m \leq N-1 \end{aligned}$$

$$\tilde{R}_x[m] = \frac{1}{N-m} \sum_{n=0}^{N-m-1} x[n] x^*[n+m], \quad m = 0, 1, 2, \dots, N-1$$

- Hence, for $|m| < N$, the error mean value is zero; this justifies the epithet unbiased estimate, somehow.
- The main drawback of this estimator is the fact that, as m increases, the number of products $X[n]X^*[n+m]$ to be averaged for estimating the ACF at lag m (that is the mean value of $X[n]X^*[n+m]$) decreases, since it is equal to $N-m$. Consequently, the error variance increases with m .
- The worst case is when $m=N-1$: in this case $N-m=1$, so only a single product can be evaluated \rightarrow there is no averaging at all. The product $X[N-1]X^*[0]$ can assume very large values, depending on the power of the process, even if $R_x[N-1]$ is very small or equal to zero, then the estimation error can be very large.

- Typically, for processes that do not contain periodic components, the ACF tends to 0 when m increases (with a convergence speed that depends on the bandwidth of the process). To take this into account, the sample estimator can be modified as follows:

$$\hat{R}_x[m] = \frac{1}{N} \sum_{n=0}^{N-m-1} x[n]x^*[n+m], \quad 0 \leq m \leq N-1$$

- This is called the **biased Sample ACF** and it has generally a smaller error than the unbiased one as m tends to $N-1$, since the unbiased estimate can assume large values for m close to $N-1$.
- For small values of m the biased and the unbiased estimators have similar performance, since $N-m \cong N$.

- The bias of the **biased Sample ACF** can be easily derived:

$$\hat{R}_x[m] = \left(\frac{N-m}{N} \right) \tilde{R}_x[m] = \left(1 - \frac{m}{N} \right) \tilde{R}_x[m], \quad \text{for } 0 \leq m \leq N-1$$

$$E\{\hat{R}_x[m]\} = \left(1 - \frac{m}{N} \right) E\{\tilde{R}_x[m]\} = \left(1 - \frac{m}{N} \right) R_x[m]$$

$$b\{\hat{R}_x[m]\} = \frac{m}{N} R_x[m], \quad \text{for } 0 \leq m \leq N-1$$

- The bias tends to increase linearly with m , but if the ACF does not contain periodic components, it typically tends to 0 in such a way that the bias tends to 0.

■ MSE of the biased Sample ACF vs the MSE of the unbiased Sample ACF:

$$\hat{R}_x[m] = \left(1 - \frac{m}{N}\right) \tilde{R}_x[m], \quad \text{for } 0 \leq m \leq N-1$$

$$\begin{aligned} \text{MSE}\{\hat{R}_x[m]\} &= \left(b\{\hat{R}_x[m]\}\right)^2 + \text{var}\{\hat{R}_x[m]\} \\ &= \left(\frac{m}{N} R_x[m]\right)^2 + \left(1 - \frac{m}{N}\right)^2 \text{var}\{\tilde{R}_x[m]\} \\ &= \left(\frac{m}{N} R_x[m]\right)^2 + \left(1 - \frac{m}{N}\right)^2 \text{MSE}\{\tilde{R}_x[m]\}, \quad \text{for } 0 \leq m \leq N-1 \end{aligned}$$

■ The presence of a bias makes the MSE of the biased Sample ACF estimator to increase, but the reduction in variance is such that overall its MSE generally decreases (again, if the ACF does not contain periodic components).

- The **biased Sample ACF** always provides a positive definite estimate, i.e. its Fourier transform is always positive. Instead, the FT of the **unbiased Sample ACF** could be negative for some values of the frequency. This can be a problem since the Power Spectral Density (PSD), that is the FT of the ACF, should always be positive.

- Let us see some numerical results for a **Markov process**, i.e. an AR(1) process:

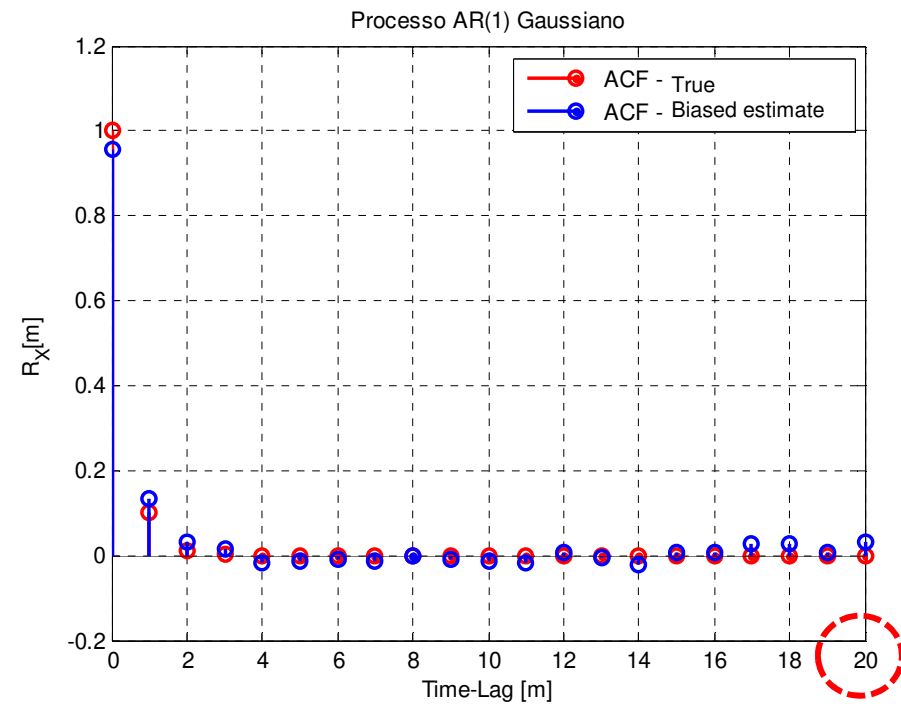
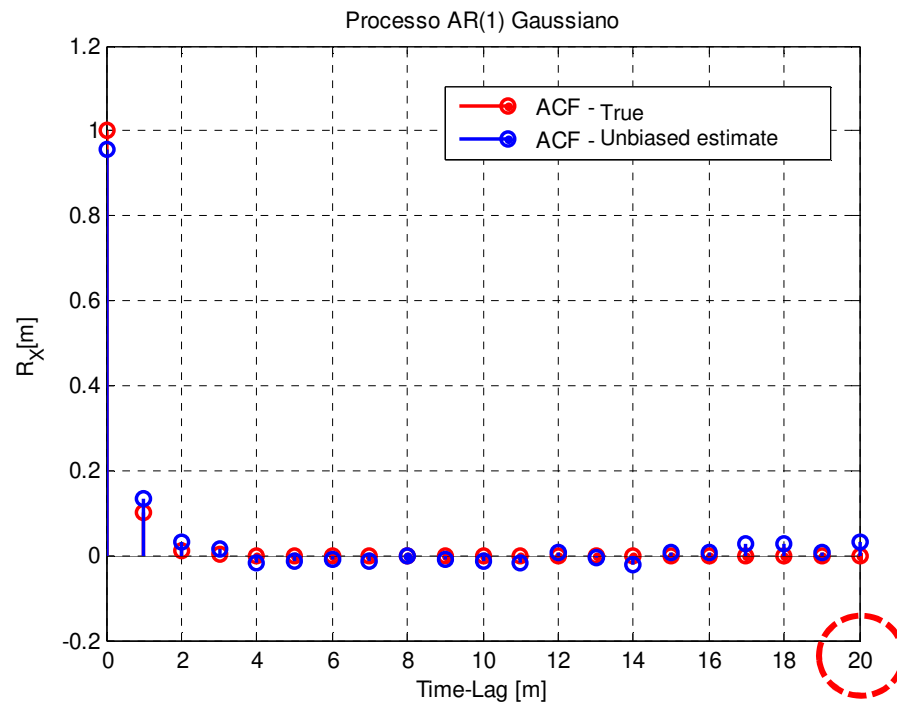
$$R_X[m] \triangleq E\{X[n]X[n+m]\} = \sigma_X^2 \rho^{|m|}, \quad |\rho| < 1$$

- Assume: $\sigma_X^2 = 1$

Estimate of the ACF of a Markov process, AR(1)

$$\rho = 0.1$$

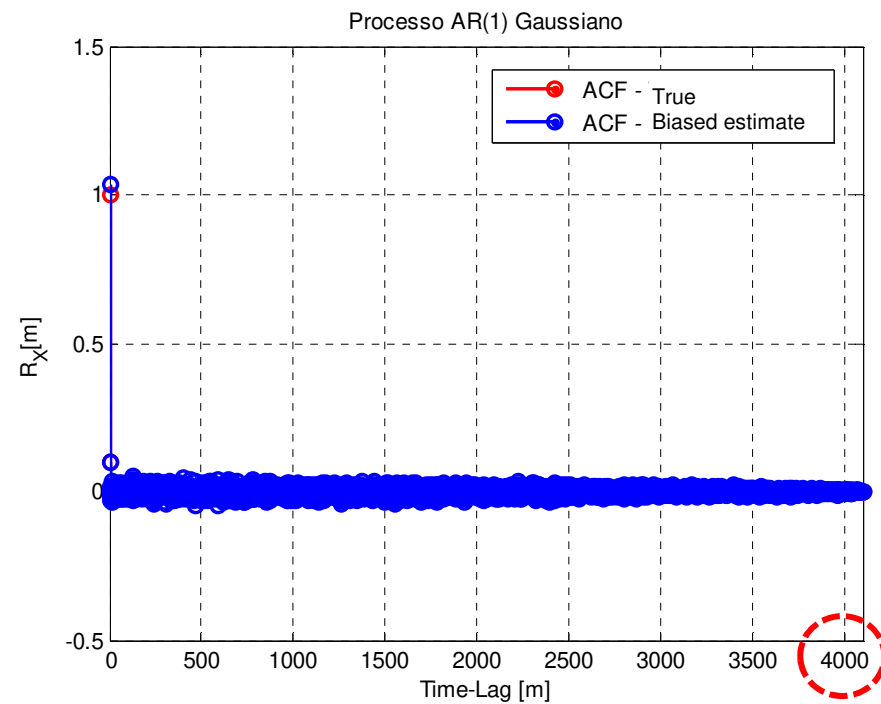
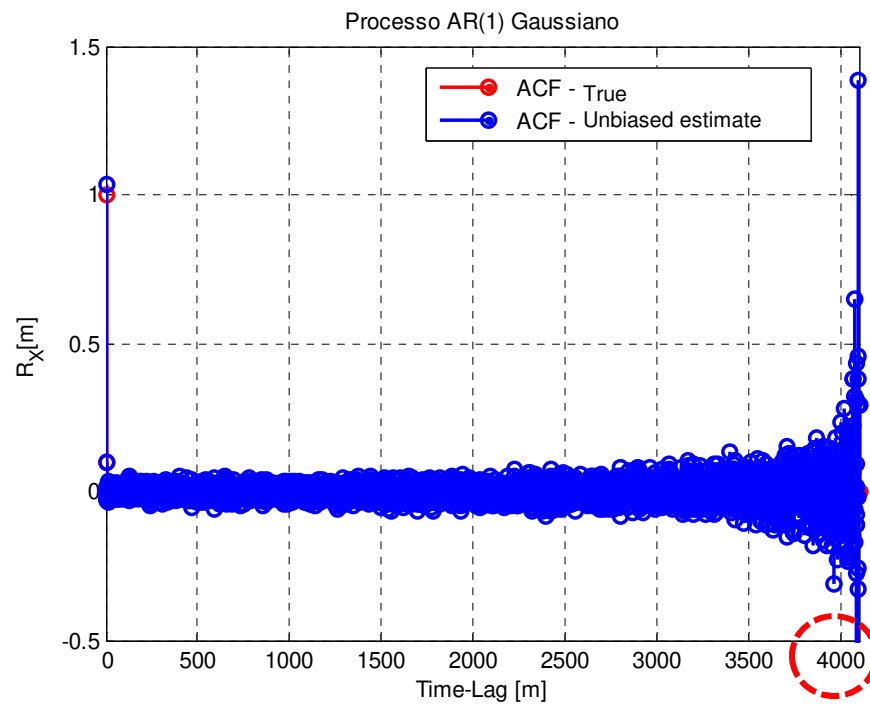
$$N = 4096$$



Estimate of the ACF of a Markov process, AR(1)

$$\rho = 0.1$$

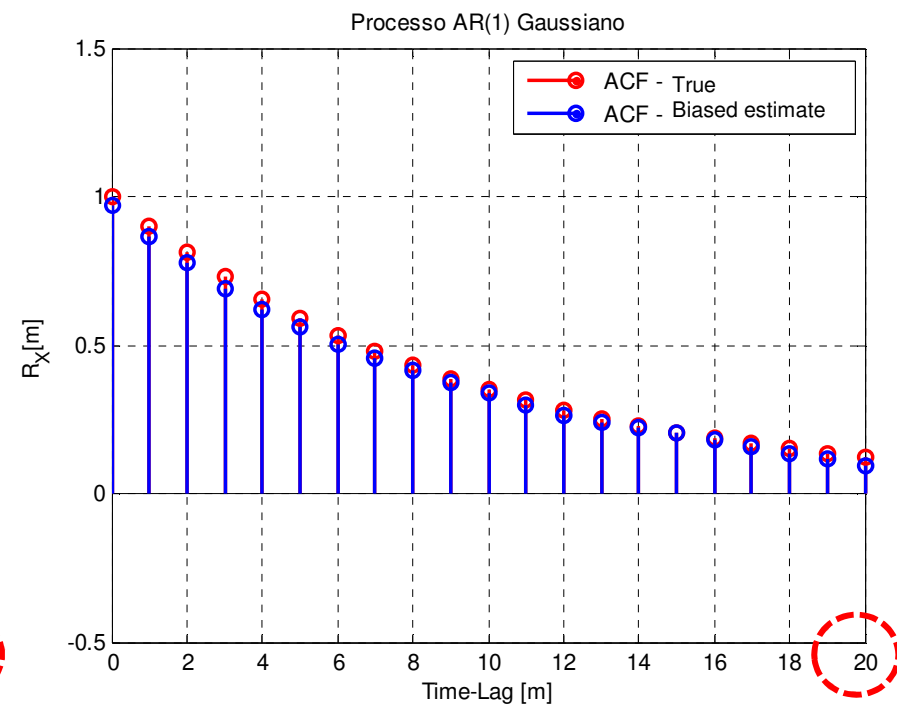
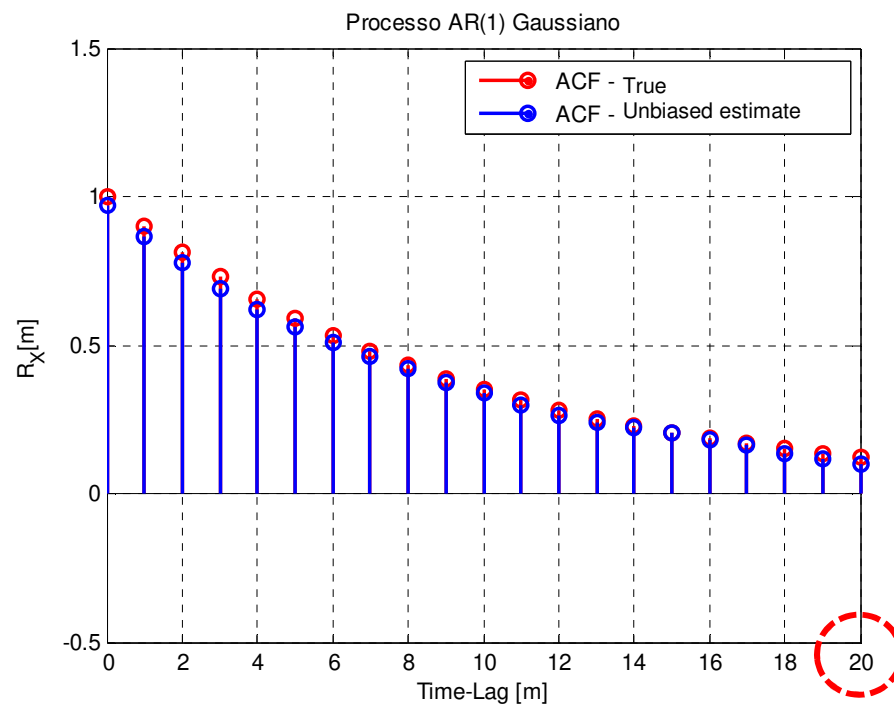
$$N = 4096$$



Estimate of the ACF of a Markov process, AR(1)

$$\rho = 0.9$$

$$N = 4096$$



Estimate of the ACF of a Markov process, AR(1)

$$\rho = 0.9$$

$$N = 4096$$

