6. Sums of Random Variables

- mean and variance
- PDF of sums of independent RVs
- laws of large numbers
- central limit theorems

Mean and Variance

let $X_1, X_2, \ldots, X_n$ be a sequence of RVs

regardless of statistical dependence, we have

$$\mathbf{E}[X_1 + X_2 + \cdots + X_n] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \cdots + \mathbf{E}[X_n]$$

the variance of a sum of RVs is, however, NOT equal to the sum of variances

$$\mathbf{var}(X_1 + X_2 + \cdots + X_n) = \sum_{k=1}^{n} \mathbf{var}(X_k) + \sum_{j=1}^{n} \sum_{k=1}^{n} \mathbf{cov}(X_j, X_k)$$

If $X_1, X_2, \ldots, X_n$ are *uncorrelated*, then

$$\mathbf{var}(X_1 + X_2 + \cdots + X_n) = \mathbf{var}(X_1) + \mathbf{var}(X_2) + \cdots + \mathbf{var}(X_n)$$

PDF of sums of independent RVs

consider the sum of n independent RVs

$$S_n = X_1 + X_2 + \cdots + X_n$$

the characteristic function of S_n is

$$\begin{aligned}
\Phi_S(\omega) &= \mathbf{E}[e^{j\omega S_n}] = \mathbf{E}[e^{j\omega(X_1+X_2+\cdots+X_n)}] \\
&= \mathbf{E}[e^{j\omega X_1}]\cdots\mathbf{E}[e^{j\omega X_n}] \\
&= \Phi_{X_1}(\omega)\cdots\Phi_{X_n}(\omega)
\end{aligned}$$

thus the pdf of S_n is found by finding the inverse Fourier of $\Phi_S(\omega)$:

$$f_S(X) = \mathcal{F}^{-1}[\Phi_{X_1}(\omega)\cdots\Phi_{X_n}(\omega)]$$

Example

find the pdf of a sum of n independent exponential RVs

all exponential variables have parameter α

the characteristic function of a single exponential RV is

$$\Phi_X(\omega) = \frac{\alpha}{\alpha - \mathrm{j}\omega}$$

the characteristic function of the sum is

$$\Phi_S(\omega) = \left(\frac{\alpha}{\alpha - \mathrm{j}\omega}\right)^n$$

we see that S_n is an n-Erlang RV

Sample mean

let X be an RV with $\mathbf{E}[X] = \mu$ (unknown)

$X_1, X_2, \ldots, X_n$ denote n independent, repeated measurements of X

X_j's are *independent, identically distributed* (i.i.d.) RVs

the **sample mean** of the sequences is used to estimate $\mathbf{E}[X]$:

$$M_n = \frac{1}{n} \sum_{j=1}^{n} X_j$$

two statistical quantities for characterizing the sample mean's properties:

- $\mathbf{E}[M_n]$: we say M_n is unbiased if $\mathbf{E}[M_n] = \mu$
- $\mathbf{var}(M_n)$: we examine this value when n is large

the sample mean is an **unbiased estimator** for μ:

$$\mathbf{E}[M_n] = \mathbf{E}\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right] = \frac{1}{n}\sum_{j=1}^{n}\mathbf{E}[X_j] = \mu$$

suppose $\mathbf{var}(X) = \sigma^2$ (true variance)

since X_j's are i.i.d, the variance of M_n is

$$\mathbf{var}(M_n) = \frac{1}{n^2}\sum_{j=1}^{n}\mathbf{var}(X_j) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

hence, the variance of the sample mean approaches zero as the number of samples increases

Weak Law of Large Numbers

let $X_1, X_2, \ldots, X_n$ be a sequence of iid RVs with finite mean $\mathbf{E}[X] = \mu$ and variance σ^2

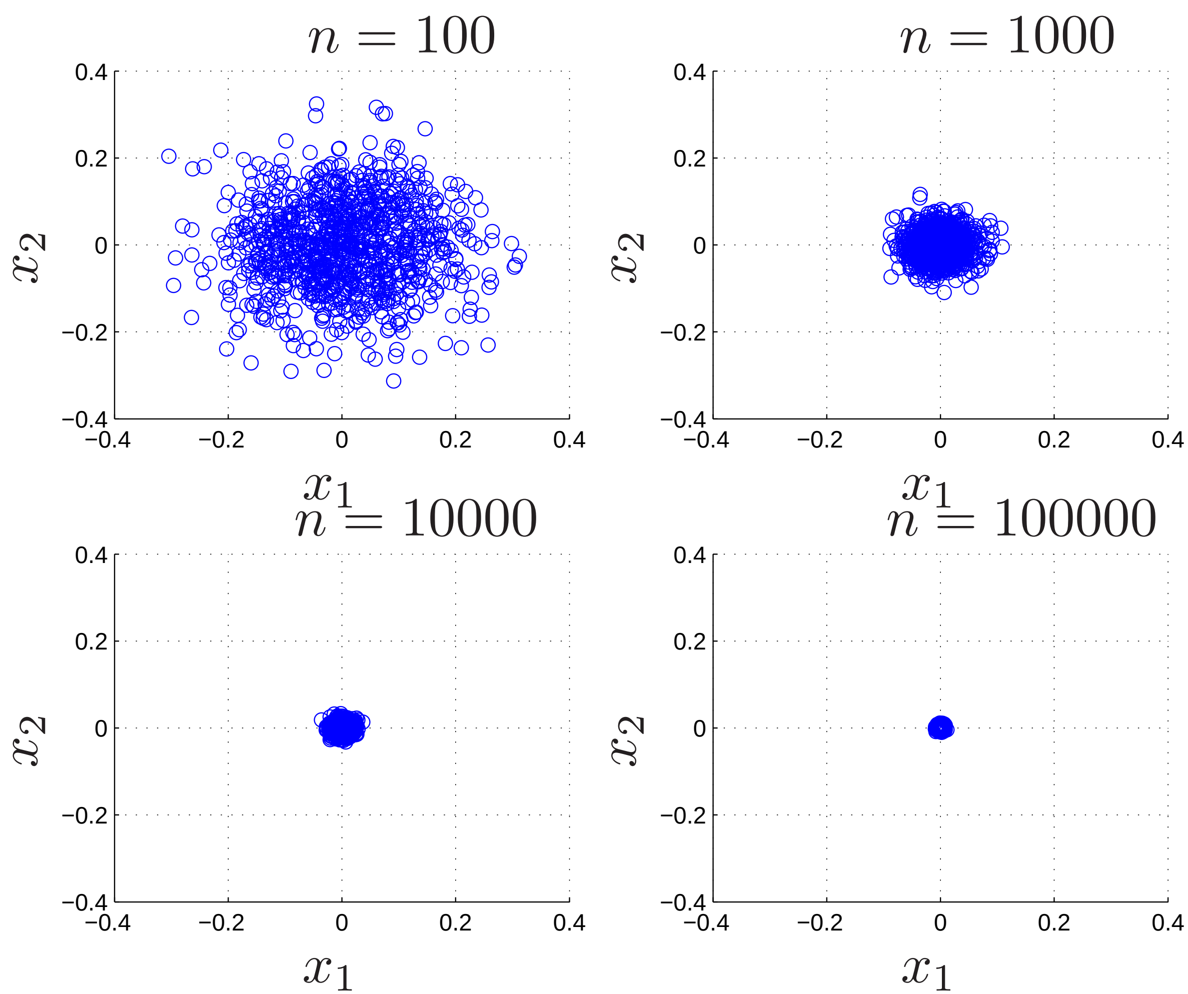
for any $\epsilon > 0$,

$$\lim_{n\to\infty} P[\ |M_n - \mu| < \epsilon\ \ ] = 1$$

- for large enough n, the sample mean will be close to the true mean with high probability
- *Proof.* apply Chebyshev inequality:

$$P[|M_n - \mu| \geq \epsilon] \leq \frac{\sigma^2}{n\epsilon^2} \quad \Longrightarrow \quad P[|M_n - \mu| < \epsilon] \geq 1 - \frac{\sigma^2}{n\epsilon^2}$$

scattergram of 1000 realizations of the sample mean



- M_n's are computed from 2-dimensional Gaussian with zero mean
- as n increases, the probability of M_n's are concentrated at zero is high

Strong Law of Large Numbers

let $X_1, X_2, \ldots, X_n$ be a sequence of iid RVs with finite mean $\mathbf{E}[X] = \mu$ and finite variance, then

$$P[\lim_{n\to\infty} M_n = \mu] = 1$$

- M_k is the sequence of sample mean computed using X_1 through X_k
- with probability 1, every sequence of sample mean calculations will eventually approach and stay close to $\mathbf{E}[X] = \mu$
- the strong law implies the weak law

Central Limit Theorem

let $X_1, X_2, \ldots, X_n$ be a sequence of iid RVs with

finite mean $\mathbf{E}[X] = \mu$ and finite variance σ^2

let S_n be the sum of the first n RVs in the sequences:

$$S_n = X_1 + X_2 + \cdots + X_n$$

and define

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

then

$$\lim_{n\to\infty} P(Z_n \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$$

as n becomes large, the CDF of normalized S_n approaches Gaussian distribution

Proof of Central Limit Theorem

first note that

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} (X_k - \mu)$$

the characteristic function of Z_n is given by

$$\begin{aligned}
\Phi_{Z_n}(\omega) &= \mathbf{E}[e^{\mathrm{j}\omega Z_n}] = \mathbf{E}\left[\exp \frac{\mathrm{j}\omega}{\sigma\sqrt{n}} \sum_{k=1}^{n} (X_k - \mu)\right] \\
&= \mathbf{E}\left[\prod_{k=1}^{n} e^{\mathrm{j}\omega (X_k - \mu)/\sigma\sqrt{n}}\right] \\
&= \left(\mathbf{E}[e^{\mathrm{j}\omega (X - \mu)/\sigma\sqrt{n}}]\right)^n
\end{aligned}$$

(using the fact that X_k's are iid)

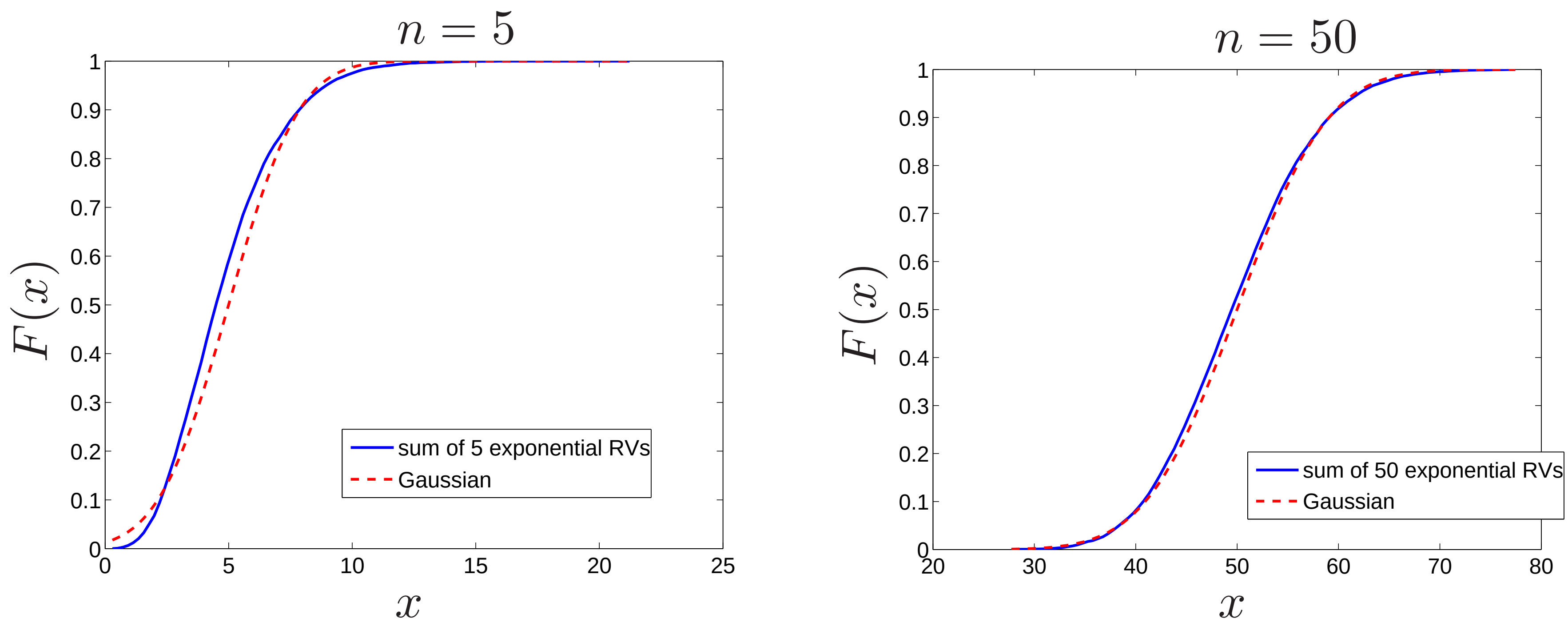
expanding the exponential expression gives

$$\begin{aligned}\mathbf{E}[e^{\mathrm{j}\omega(X-\mu)/\sigma\sqrt{n}}] &= \mathbf{E}\left[1 + \frac{\mathrm{j}\omega}{\sigma\sqrt{n}}(X-\mu) + \frac{(\mathrm{j}\omega)^2}{2!n\sigma^2}(X-\mu)^2 + \dots\right] \\ &\approx 1 - \frac{\omega^2}{2n}\end{aligned}$$

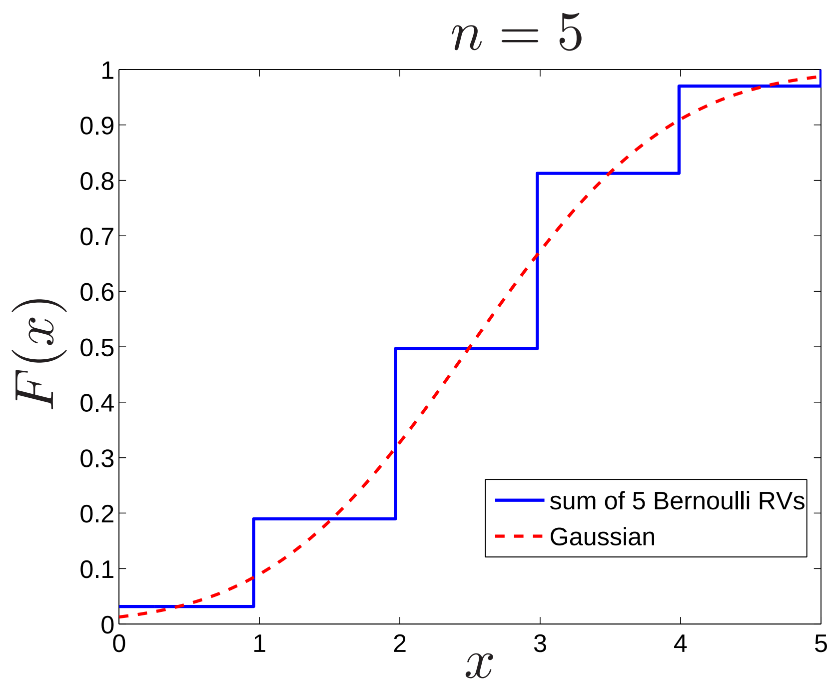
(the higher order term can be neglected as n becomes large)

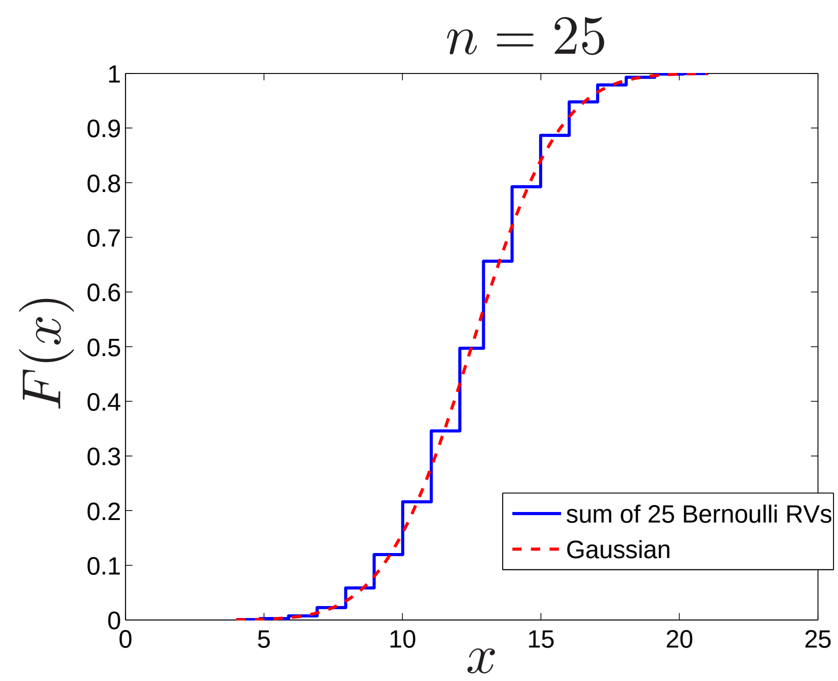
then we obtain

$$\begin{aligned}\Phi_{Z_n}(\omega) &\rightarrow \left(1 - \frac{\omega^2}{2n}\right)^n \\ &\rightarrow e^{-\omega^2/2}, \quad \text{as } n \rightarrow \infty\end{aligned}$$



- blue lines are the CDF of the sum of n exponential RVs with mean $\lambda = 1$ where $n = 5$ (left) and $n = 50$ (right)
- red dashed line is the CDF of a Gaussian RV with the same mean $(n\lambda)$ and variance n/λ^2
- as n increases, the CDF approaches that of Gaussian distribution





- blue lines are the CDFs of the sum of n Bernoulli RVs with $p = 1/2$ where $n = 5$ (left) and $n = 25$ (right)
- red dashed line is the CDF of a Gaussian with mean np and variance $np(1-p)$

Example

the time between events is iid exponential RVs with mean m sec

find the probability that the 1000th even occurs in time interval $(1000 \pm 50)m$

- X_j is the time between events
- S_n is the time of the nth event (then $S_n = X_1 + X_2 + \cdots + X_n$)
- $\mathbf{E}[S_n] = nm$ and $\mathbf{var}(S_n) = nm^2$

the CLT gives

$$\begin{aligned} P(950m &\leq S_{1000} \leq 1050m) \\ &= P\left(\frac{950m - 1000m}{m\sqrt{1000}} \leq Z_n \leq \frac{1050m - 1000m}{m\sqrt{1000}}\right) \\ &\approx \Phi(1.58) - \Phi(-1.58) \end{aligned}$$

References

Chapter 7 in
A. Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, 3rd edition, Pearson Prentice Hall, 2009