MPEI

Soma e Combinação Linear de Variáveis Aleatórias

Funções de Variáveis Aleaórias

Motivação

- Se somarmos duas variáveis aleatórias X_1 e X_2 quais as características da variável aleatória $S = X_1 + X_2$?

  – Em termos de momentos ?
    - Em especial média e variância

  – Em termos de função de distribuição ?

- E no caso geral $S_n = X_1 + X_2 + \cdots + X_n$?

Média da soma de n variáveis

- Sejam $X_1, X_2, \ldots, X_n$, n variáveis aleatórias e $S_n = X_1 + X_2 + \cdots + X_n$ a sua soma
- Teorema: **A média da soma de *n* variáveis é igual à soma das médias**
- Demonstração

$$
\begin{aligned}
E\left[S_n\right] &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (\sum_{j=1}^{n} x_j) f_{X_1 \ldots X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_n = \\
&= \sum_{j=1}^{n} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} x_j f_{X_1 \ldots X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_n = \\
&= \sum_{j=1}^{n} \int_{-\infty}^{+\infty} x_j f_{X_j}(x_j) dx_j = \sum_{j=1}^{n} E\left[X_j\right]
\end{aligned}
$$

Variância da soma de n variáveis

- Considerando da mesma forma $S_n = X_1 + X_2 + \cdots + X_n$:
- Teorema: **A variância da soma de *n* variáveis aleatórias é dada pela soma de todas as variâncias e covariâncias**

$$Var(S_n) = \sum_{i=1}^{n} Var(X_i) + \sum_{\substack{j=1 \\ j\neq k}}^{n} \sum_{k=1}^{n} Cov(X_j, X_k)$$

- Demonstração:

$$\begin{aligned} Var\left(S_n\right) &= E\left[\sum_{j=1}^{n}\left(X_j - E\left[X_j\right]\right)\sum_{k=1}^{n}\left(X_k - E\left[X_k\right]\right)\right] = \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} E\left[\left(X_j - E\left[X_j\right]\right)\right] E\left[\left(X_k - E\left[X_k\right]\right)\right] \end{aligned}$$

Variância da soma de *n* variáveis

- Se as variáveis **são independentes,** $Cov(X_j, X_k) = 0$, para todo o $j \neq k$, pelo que:

- $\boldsymbol{Var(S_n) = \sum_{i=1}^{n} Var(X_i)}$
  - **Variância da soma igual a soma das variâncias**

- Se para além de independentes forem **identicamente distribuídas (IID)**
  e tivermos $E[X_i] = \mu$ e $Var(X_i) = \sigma^2, i = 1,2, \ldots, n$
  a média e variância da soma são dadas por:

- $\mathrm{E}[S_n] = n\,\mu$ e $Var(S_n) = n\,\sigma^2$

Função de distribuição da soma de 2 variáveis aleatórias independentes

- Caso discreto (2 v.a. Discretas X e Y)
- Fazendo $\mathrm{Z} = X + Y$
- $p_Z(z) = P(X + Y = z)$

$=\sum_{\{(x,y)|x+y=z\}} P(X = x, Y = y)$

$$= \sum_x P(X = x, Y = z - x)$$

$= \sum_x p_X(x) \quad p_Y(z - x)$; devido à indep.

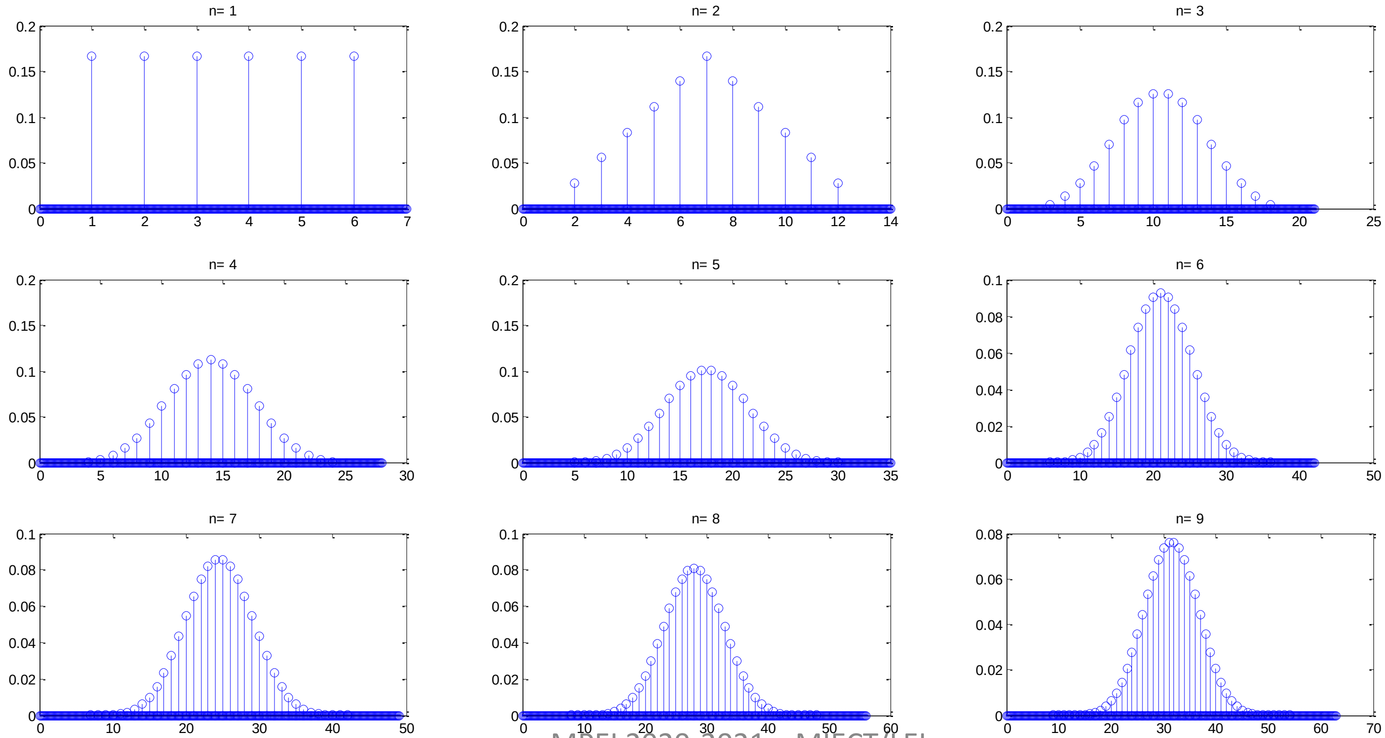
$= p_X\ (x) \ * \ p_Y(z)$

demoConvolucao.m

- Que é a **convolução** discreta de p_X e p_Y

Exemplo (em Matlab)

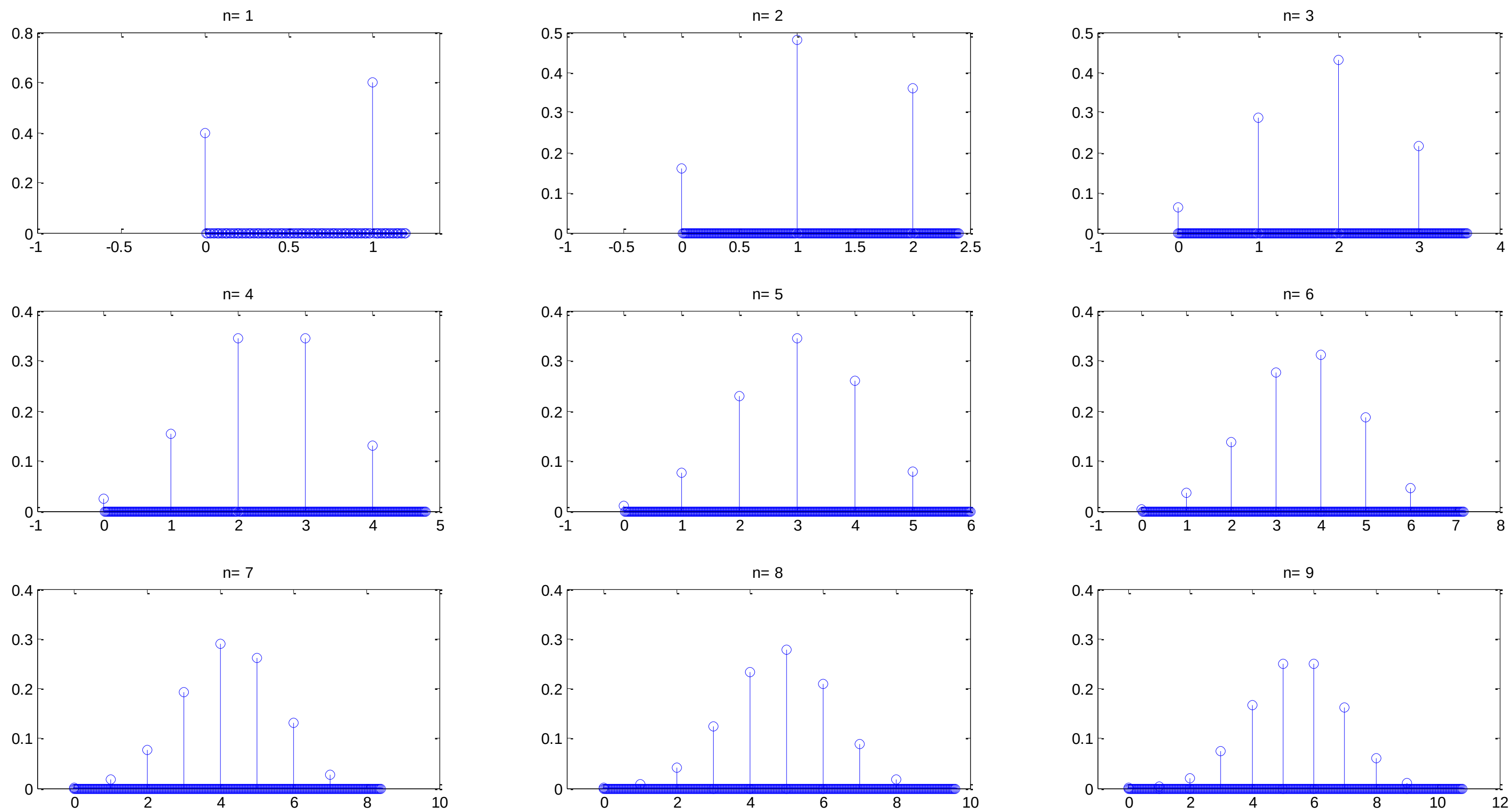
- Usando conv() e a pmf relativa à variável X correspondente ao lançamento de um dado honesto (n=1, 2, ..., 9)



Outro exemplo

demoConvolucaoMoeda.m

- Sendo X relativa ao número de caras num lançamento de moeda não honesta
  - com probabilidade de cara = 0,6



Caso contínuo

- Sendo X e Y independentes e contínuas
- Fazendo novamente $\mathrm{Z} = X + Y$
- Para obter a função densidade prob. de Z , primeiro obtém-se a f. densidade conjunta de X e Z e depois integra-se

$$
\begin{aligned}
F_{Z|X}(z \mid x) &= \mathbf{P}(Z \le z \mid X = x) = \mathbf{P}(X + Y \le z \mid X = x) = \mathbf{P}(x + Y \le z \mid X = x) = \mathbf{P}(Y \le z - x \mid X = x) \\
&= \mathbf{P}(Y \le z - x) \qquad \text{using the independence of } X \text{ and } Y \\
&= F_Y(z - x)
\end{aligned}
$$

$$
f_{Z|X}(z \mid x) = \frac{d}{dz} F_{Z|X}(z \mid x) = \frac{d}{dz} F_Y(z - x) = f_Y(z - x)
$$

$$
\begin{aligned}
f_Z(z) &= \int_{-\infty}^{\infty} f_{X,Z}(x,z)\,dx = \int_{-\infty}^{\infty} f_X(x) f_{Z|X}(z \mid x)\,dx \\
&= \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\,dx \equiv f_X * f_Y(z)
\end{aligned}
$$

Obtém-se através da convolução, agora contínua

- Para mais informação, ver, por exemplo:
- https://engineering.purdue.edu/~ipollak/ece302/SPRING12/notes/23_GeneralRVs-8_Sums.pdf

Combinações lineares de variáveis aleatórias

- Os **resultados anteriores generalizam-se** facilmente para o caso de termos uma soma pesada (combinação linear) $Y_n = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$
  - Em que $c_1, c_2, \ldots, c_n$ são constantes

- $E[Y_n] = c_1 E[X_1] + c_2 E[X_2] + \cdots + c_n E[X_n]$
- $Var(Y_n) = \sum_{i=1}^{n} {c_i}^2 \, Var(X_i) + \sum_i \sum_{j\,(\neq i)} c_i c_j Cov(X_i, X_j)$

- Se independentes $Var(Y_n) = \sum_{i=1}^{n} {c_i}^2 \, Var(X_i)$

Exemplo de aplicação

- Um semicondutor tem 3 camadas.
- Se a espessura das várias camadas tiver uma variância de 25, 40 e 30 nanómetros quadrados, qual a variância da espessura das 3 camadas ?

- Considerando X_1, X_2, X_3 como as espessuras das camadas e independência
- $Var(X_1 + X_2 + X_3) = Var(X_1) + var(X_2) + var(X_3) = 95$
- Mostrando a propagação das variâncias de todas as camadas para o resultado final
  - Os problemas somam-se ☹

Funções de variáveis aleatórias

A soma (simples ou pesada) de variáveis aleatórias é um caso particular

Funções de v. a. múltiplas

- Muitas vezes encontramos **problemas em que temos uma transformação das v. a.** $X_1, X_2, \dots, X_n$ que produz variáveis aleatórias $Y_1, Y_2, \dots, Y_m$
- O caso mais simples é termos uma função escalar de várias variáveis aleatórias

$$Z = g(X_1, X_2, \dots, X_n)$$

- A função de distribuição acumulada de Z é determinada como sabemos calculando a probabilidade do conjunto $\{Z \leq z\}$

  i.e. A região R_z do espaço n-dimensional tal que

  $R_z = \{\mathbf{x}: g(\mathbf{x}) \leq \mathbf{z}\}$ $\mathbf{x} = (\boldsymbol{x_1}\ \boldsymbol{x_2}\ \dots\ \boldsymbol{x_n})$
- Logo
  - $F_Z(z) = \int_{\mathrm{x} \in R_Z} \dots \int f_X(\mathrm{x})\, d\mathrm{x}$

Expectância de funções de v. aleatórias

- Expectância de uma função de 2 var. aleatórias

  $Z = g(X,Y)$

  $$E[Z] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\, dx\, dy$$

- Para o caso de variáveis discretas :
  - $E[Z] = \sum_i \sum_n g(x_i, y_n)\, p_{X,Y}(x_i, y_n)$

- Resultado generalizável para uma função com um número arbitrário de variáveis aleatórias

  $Z = g(\mathbf{X})$ em que $\mathbf{X}$ (bold) é um vector

  $$E[Z] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

Exemplo

- $Z = g(X, Y) = X + Y$ *i.e.* Soma de 2 v.a.
- $E[Z] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f_{X,Y}(x, y)\, dx\, dy$
- $= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, f_{X,Y}(x, y)\, dx\, dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x, y)\, dx\, dy$
- $= \int_{-\infty}^{\infty} x\, f_X(x)\, dx + \int_{-\infty}^{\infty} y f_Y(y)\, dy$
- $= \boldsymbol{E(X) + E(Y)}$
- Resultado válido quer as variáveis **sejam independentes ou não**
- Mostra (conjuntamente com a propriedade de escala, $E(aX) = aE(X)$) que a esperança é um **operador linear**

Momentos de funções de variáveis aleatórias

- Momento de ordem n de uma função escalar de um vector aleatório:

  $Z = g(\mathbf{X})$ em que $\mathbf{X}$ (bold) é um vector

  $$E[\boldsymbol{Z}^{\boldsymbol{n}}] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \boldsymbol{g}^{\boldsymbol{n}}(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$$

- Aplicando para obter a variância temos:
- $\mathrm{Var}[Z] = E[Z^2] - E^2[Z]$
- $= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \boldsymbol{g}^{\mathbf{2}}(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x} - \left(\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \boldsymbol{g}(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}\right)^{\mathbf{2}}$

Média de variáveis aleatórias

- E se a função aplicada a um vetor de n variáveis aleatórias for a média dessas variáveis?

- Interessa-nos em especial o caso em que essas *n* variáveis são independentes e identicamente distribuídas (IID)

Valor esperado da Média

- Se criarmos a variável aleatória relativa à média de *n* variáveis IID X_i,

$$\boldsymbol{M_n = \frac{S_n}{n}},$$

- e assumindo $E[X_i] = \mu$ e $Var(X_i) = \sigma^2$
- teremos :
- $\boldsymbol{E[M_n]} = E\left[\frac{S_n}{n}\right]$
- $= \frac{\sum_i E[X_i]}{n}$
- $= E[X_i] = \boldsymbol{\mu}$

Variância da Média

- **Var($\boldsymbol{M_n}$)= ?**
- **Var[$\boldsymbol{M_n}$]** $= Var\left(\frac{S_n}{n}\right)$
- $= \frac{1}{n^2}\frac{\sum_i Var[X_i]}{1}$
- $= \frac{Var(X_i)}{n} = \boldsymbol{\frac{\sigma^2}{n}}$

- À medida que se aumenta o número de experiências vai diminuindo a variância da estimativa da média

Aplicação interessante de soma de variáveis aleatórias

Contadores estocásticos

Motivação

- Evitar contadores grandes quando o volume de dados é grande
  - Por exemplo: na contagem de células

- Como um contador de n bits contará no máximo até 2^n eventos, será este o limite a ultrapassar

Primeira solução

- Para duplicar o número de eventos que se podem contar, incrementa-se o contador com probabilidade 1/2 cada vez que ocorre um evento

- A ideia *é incrementar o contador metade das vezes (em média)*

…

- Com base na função rand() podemos agora tomar decisões aleatórias com probabilidade 1/2 e portanto construir uma função para incrementar (ou não) o contador:

```
if (rand() < 0.5) then
    incrementar_contador
endif
```

Em Matlab

- Podemos facilmente simular o resultado após 100 eventos:

```
% gera 100 var aleatórias indep em [0,1]
x = rand(1, 100);

% calcular quantas são < 0.5
n = sum(x < 0.5);
```

- n representará o valor do contador após os 100 eventos

Qual é o valor médio do contador após k eventos?

- O contador é uma variável aleatória, determinada por uma sucessão de experiências aleatórias

- Associando uma variável aleatória a cada evento, de forma a representá-lo probabilisticamente
- Seja X_i a variável aleatória que representa o incremento i, com valor 1 se o contador foi incrementado, e valor zero caso contrário.
- Como $P(X_i = 0)$ e $P(X_i = 1)$ são iguais a 1/2, tem-se

- $E[X_i] = 0 \times P(X_i = 0) + 1 \times P(X_i = 1) = \frac{1}{2}$

Valor médio

- O valor do contador após k eventos é a soma dos k incrementos, $S = X_1 + X_2 + \cdots + X_k$

- E o valor médio:
- $E[S] = E[X_1 + X_2 + \cdots + X_k]$
- $= E[X_1] + E[X_2] + \cdots + E[X_k]$
- $= \frac{1}{2} + \cdots + \frac{1}{2} = \frac{k}{2}$

- Como o valor médio do contador após k eventos é $k/2$, o número de eventos pode ser estimado através do dobro do número registado pelo contador

Variância

- A variância de um qualquer dos X_i é
- $Var(X_i) = E\left[{X_i}^2\right] - \left(E[X_i]\right)^2$

- $E\left[{X_i}^2\right] = 0^2 \times P(X_i = 0) + 1^2 \times P(X_i = 1)$
- $= \frac{1}{2}$

- $Var(X_i) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$

Variância (continuação)

- Como as variáveis X_i são independentes, a variância de S é
- $Var(S) = Var(X_1 + X_2 + \cdots + X_k)$
- $= Var(X_1) + Var(X_2) + \cdots + Var(X_k)$
- $= \frac{1}{4} + \cdots + \frac{1}{4} = \frac{k}{4}$
- O que implica $\sigma = \frac{\sqrt{k}}{2}$
- Para n=10000 teremos:
  - média 5000
  - desvio padrão 50

Comparar com simulação (contadores1.m)

Distribuição de probabilidade

- Pode calcular-se a probabilidade de, após k eventos, o valor do contador ser n.

- Fixemos $k = 4$:
- Teremos $X_1, X_2, X_3 \ e \ X_4$
  – Variáveis binárias que descrevem se o contador é incrementado ou não após o evento 1,2,3 e 4
- O que nos dá 16 possibilidades (2^4)

X_1	X_2	X_3	X_4	valor do contador
0	0	0	0	0
0	0	0	1	1
0	0	1	0	1
0	0	1	1	2
0	1	0	0	1
0	1	0	1	2
0	1	1	0	2
0	1	1	1	3
1	0	0	0	1
1	0	0	1	2
1	0	1	0	2
1	0	1	1	3
1	1	0	0	2
1	1	0	1	3
1	1	1	0	3
1	1	1	1	4

- É agora fácil determinar as probabilidades, por contagem:
- $p(0) = \frac{1}{16}$
- $p(1) = \frac{4}{16}$
- $p(2) = \frac{6}{16}$
- $p(3) = \frac{4}{16}$
- $p(4) = \frac{1}{16}$

Generalizando

- Sendo p a probabilidade de incrementar e $1 - p$ a probabilidade de não incrementar …
- A probabilidade de observar uma soma igual a n após k experiências é:

$$p(n) = \binom{k}{n} p^n \ (1-p)^{k-n}$$

Variante 1

- Como proceder para alargar mais a gama do contador?

- Imaginemos, por exemplo, que se quer multiplicar por 64 essa gama. A solução natural é incrementar com probabilidade $1/64$ em vez de ½

- O valor médio de X_i será agora $\frac{1}{64}$
- E $E[S] = \quad \ldots . = \frac{k}{64}$
- Neste caso, o número de eventos pode ser estimado por $64\, n$, sendo n o valor do contador

Segunda solução

- Neste caso o contador é incrementado com probabilidade cada vez menor à medida que o seu valor aumenta:
- quando o contador contém *n*, a probabilidade de um incremento é 2^{-n}

Eventos	*Valor do contador*	*Número de eventos*
x	1	1
x		
x	2	3
x		
x		
x		
x	3	7
x		
x		
x		
x		
x		
x		
x		
x	4	15