Estatística: Aplicação ao Sensoriamento Remoto

SER 204 - ANO 2024

Simulação Estocástica

Camilo Daleles Rennó
camilo.renno@inpe.br
http://www.dpi.inpe.br/~camilo/estatistica/

Simulação

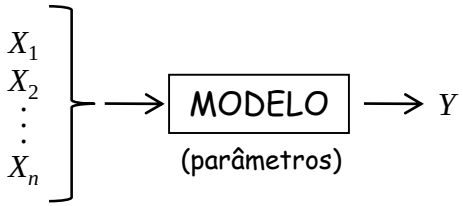
O que é Simulação?

é um experimento realizado a partir de modelos (reais ou virtuais)

Pode ser:

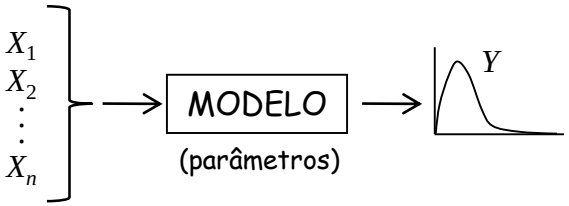
a) determinística: as entradas do modelo são fixas e para uma determinada combinação de valores de entrada o resultado final é sempre o mesmo

b) estocástica (ou probabilística): o modelo e/ou as entradas incorporam variações aleatórias de modo que os resultados são diferentes a cada simulação



entrada fixa + modelo determinístico

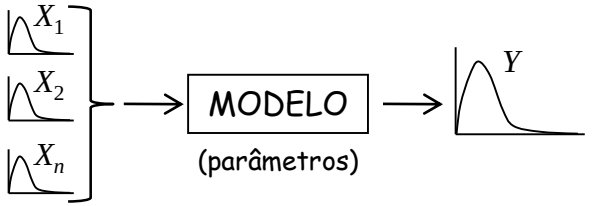
$$Y = 2X_1 + 3X_2 - X_3$$



entrada fixa + modelo estocástico

$$Y = 2X_1 + 3X_2 - X_3 + \varepsilon$$

ε é um v.a.



entrada estocástica + modelo determinístico

$$Y = 2(X_1+\varepsilon_1) + 3(X_2+\varepsilon_2) - (X_3 + \varepsilon_3)$$

ε_i são v.a. independentes

Simulação

Para que fazer Simulação?

a) gerar amostras de uma v.a. cuja distribuição é conhecida
   ex: testar métodos que podem ser influenciados pelo tipo de dado de entrada

b) avaliar propagação de incertezas (quando a solução analítica é inviável)
   ex: avaliar os resultados da combinação não linear de muitas variáveis

c) avaliar cenários futuros (resultados possíveis)
   ex: experimentar "em laboratório" condições inexistentes ou raras

d) testar a sensibilidade de parâmetros de um modelo (ou distribuição)
   ex: identificar quais parâmetros afetam mais os resultados finais

e) estimar pontualmente ou por intervalo um determinado resultado de um modelo
   ex: descrever estatisticamente os possíveis resultados de um experimento

f) testar a significância de um resultado num teste de hipótese
   ex: avaliar se uma hipótese é válida ou não sem utilizar testes estatísticos clássicos

Simulação de Monte Carlo

É um método que utiliza sequencias de números aleatórios para descrever o comportamento de uma ou mais variáveis ou a combinação das mesmas, bastando para isso conhecer a Função de Probabilidade de cada variável (ou a FDP, no caso de uma v.a. contínua).

É particularmente útil quando o modelo é complexo, não-linear, ou quando envolve muitas variáveis de entrada (com diferentes graus de incerteza), o que dificultaria uma solução analítica.

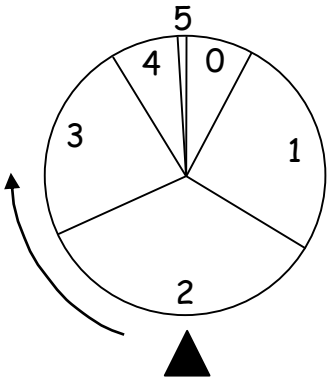
Através de um grande número de repetições (acima de 1000), garante-se que praticamente todas as combinações de entradas sejam avaliadas.

Simulação de Monte Carlo

O termo Monte Carlo foi dado em homenagem a roleta, jogo muito popular de Monte Carlo, Mônaco.

X ~ Binomial ($n = 5$; $p = 0{,}4$)

x	$P(X = x)$
0	7,78%
1	25,92%
2	34,56%
3	23,04%
4	7,68%
5	1,02%



área de cada fatia é proporcional a probabilidade do valor correspondente

roda-se a roleta e, ao parar, anota-se o valor obtido
repete-se o procedimento até que se consiga o número desejado de simulações
quanto maior o número de simulações, mais a proporção relativa de cada resultado possível se aproximará de sua probabilidade

Geração de Números Aleatórios

Originalmente os números aleatórios eram gerados usando dados, roletas, tabelas, etc.

Atualmente os computadores são usados para gerar números chamados pseudo-aleatórios, que constituem uma sequencia de valores que, embora sejam gerados de forma determinística, simulam valores independentes de uma v.a. uniforme contínua [0,1].

Qualquer v.a. pode ser simulada a partir de uma v.a. uniforme contínua [0,1] desde que se conheça sua função de distribuição acumulada $F(x) = P(X \leq x)$.

Geração de Números Aleatórios

Procedimento Geral:

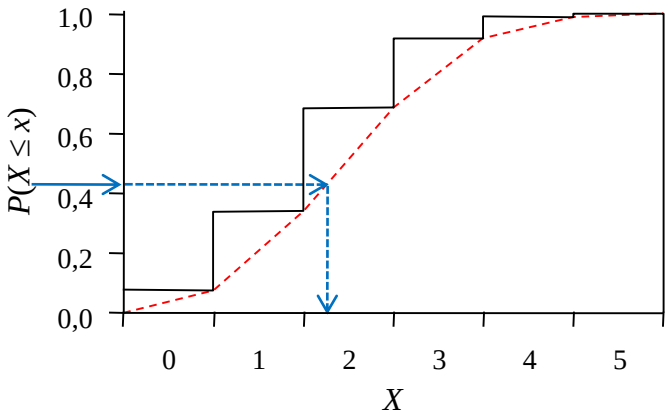
a) gera-se um número aleatório de uma v.a. uniforme contínua [0,1], u

b) determina-se qual é o menor valor da v.a. desejada cuja probabilidade seja maior ou igual a u

c) repete-se os procedimentos a) e b) até que n valores tenham sido obtidos

X ~ Binomial ($n = 5$; $p = 0{,}4$)

x	$P(X = x)$	$P(X \leq x)$
0	7,78%	7,78%
1	25,92%	33,70%
2	34,56%	68,26%
3	23,04%	91,30%
4	7,68%	98,98%
5	1,02%	100,00%



Ex: se $u = 0{,}4367$ (43,67%)

$x = 2$

Geração de Números Aleatórios

Procedimento Geral:

a) gera-se um número aleatório de uma v.a. uniforme contínua [0,1], u

b) determina-se qual é o menor valor da v.a. desejada cuja probabilidade seja maior ou igual a u

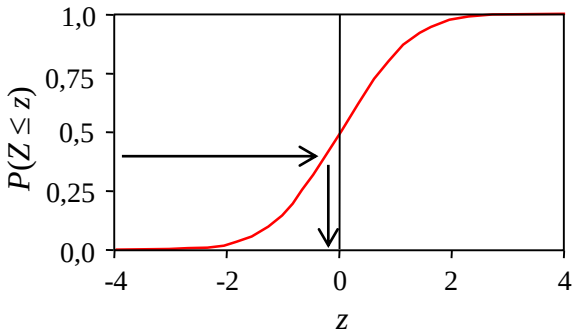
c) repete-se os procedimentos a) e b) até que n valores tenham sido obtidos

Sorteio de 8 valores $X \sim \text{Normal}(\mu = 10;\ \sigma^2 = 4)$

$u(0,1)$	$Z \sim N(0,1)$	$X \sim N(10,4)$
0,4138	-0,2177	9,5645
0,9155	1,3753	12,7505
0,6218	0,3101	10,6202
0,3848	-0,2929	9,4142
0,1058	-1,2492	7,5017
0,2763	-0,5938	8,8124
0,3855	-0,2910	9,4180
0,8036	0,8546	11,7092

$$Z = \frac{X - 10}{2} \sim N(0,1)$$

$$X = 2Z + 10$$



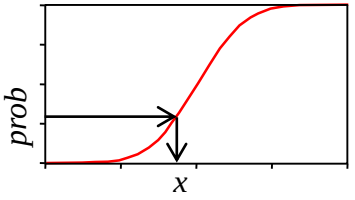
Distribuições e Números Aleatórios no Excel

Distribuição	$f(x)$* cumulativo = 0	$F(x)$ cumulativo = 1	$F^{-1}(prob)$**	Aleatório
Beta	DIST.BETA	DIST.BETA	INV.BETA	
Binomial	DISTR.BINOM	DISTR.BINOM	INV.BINOM	
Binomial Negativa	DIST.BIN.NEG.N	DIST.BIN.NEG.N		
Chi-quadrado	DIST.QUIQUA	DIST.QUIQUA	INV.QUIQUA	
Exponencial	DISTR.EXPON	DISTR.EXPON		
F	DIST.F	DIST.F	INV.F	
Gama	DIST.GAMA	DIST.GAMA	INV.GAMA	
Hipergeométrica	DIST.HIPERGEOM.N	DIST.HIPERGEOM.N		
Log Normal	DIST.LOGNORMAL.N	DIST.LOGNORMAL.N	INV.LOGNORMAL	
Normal	DIST.NORM	DIST.NORM	INV.NORM.N	
Normal Padrão	DIST.NORMP	DIST.NORMP	INV.NORMP.N	
Poisson	DIST.POISSON	DIST.POISSON		
t de Student	DIST.T	DIST.T	INV.T	
Uniforme Contínua [0,1]				ALEATÓRIO
Uniforme Discreta				ALEATÓRIOENTRE
Weibull	DIST.WEIBULL	DIST.WEIBULL		

*para v.a. contínuas, f(x) representa a Função Densidade de Probabilidade, e só é útil no caso de se "desenhar" a distribuição

**Se $prob = F(x)$

então: $x = F^{-1}(prob)$



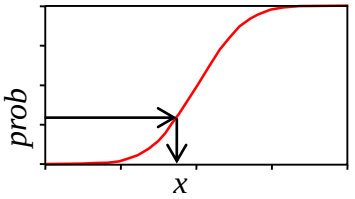
Distribuições e Números Aleatórios no R

Distribuição	$f(x)$*	$F(x)$	$F^{-1}(prob)$**	**Aleatório**
Beta	dbeta	pbeta	qbeta	rbeta
Binomial	dbinom	pbinom	qbinom	rbinom
Binomial Negativa	dnbinom	pnbinom	qnbinom	rnbinom
Chi-quadrado	dchisq	pchisq	qchisq	rchisq
Exponencial	dexp	pexp	qexp	rexp
F	df	pf	qf	rf
Gama	dgamma	pgamma	qgamma	rgamma
Geométrica	dgeom	pgeom	qgeom	rgeom
Hipergeométrica	dhyper	phyper	qhyper	rhyper
Log Normal	dlnorm	plnorm	qlnorm	rlnorm
Logística	dlogis	plogis	qlogis	rlogis
Normal	dnorm	pnorm	qnorm	rnorm
Poisson	dpois	ppois	qpois	rpois
t de Student	dt	pt	qt	rt
Uniforme	dunif	punif	qunif	runif
Weibull	dweibull	pweibull	qweibull	rweibull

*para v.a. contínuas, f(x) representa a Função Densidade de Probabilidade, e só é útil no caso de se "desenhar" a distribuição

**Se $prob = F(x)$

então: $x = F^{-1}(prob)$



Avaliação das Simulações

- Estimação da Função de Probabilidade através das frequências relativas observadas (variáveis discretas) ou da Função de Probabilidade Acumulada (variáveis discretas e contínuas)

  histogramas e gráficos de frequência acumulada

- Métricas de tendência central e de dispersão:

  média, desvio padrão, mediana, quantis, amplitude, mínimo/máximo, etc

- Intervalos de Credibilidade (não confundir com intervalos de confiança!)

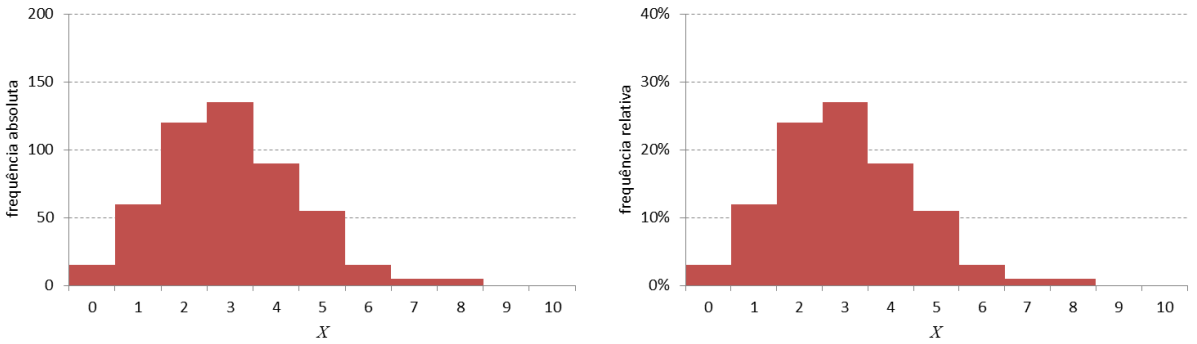
- Box-plot

  mediana, 1º e 3º quartis e valores extremos (*outliers*)

Histograma

Muito utilizado para v.a. discretas e indica a frequência absoluta ou relativa de cada valor observado na amostra. Recomenda-se que não haja espaços entre as colunas para que o histograma não seja confundido com um gráfico de barras.

Exemplo: simulação de 500 valores de uma v.a. discreta X [0, 10].

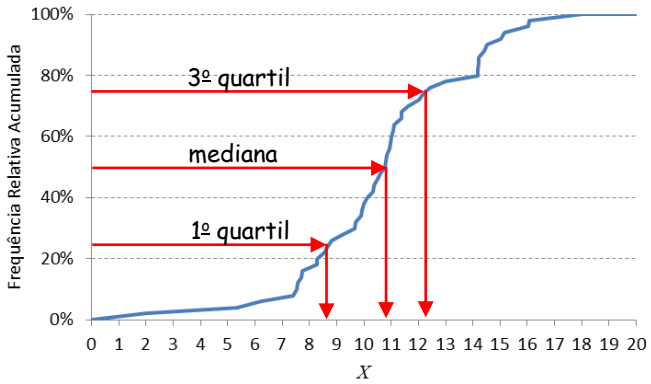


Para v.a. contínuas, é necessário dividir os dados em classes com intervalos regulares ou não. A definição do número e da largura dos intervalos é fundamental para a boa representação da distribuição amostral. No entanto, esta definição é bastante arbitrária, dependendo basicamente do tamanho e da variação encontrada na amostra. Para isso, muitas fórmulas são encontradas na literatura (p.ex. Sturges, Rice, Doane, Scott e Freedman-Diaconi).

Frequência Acumulada

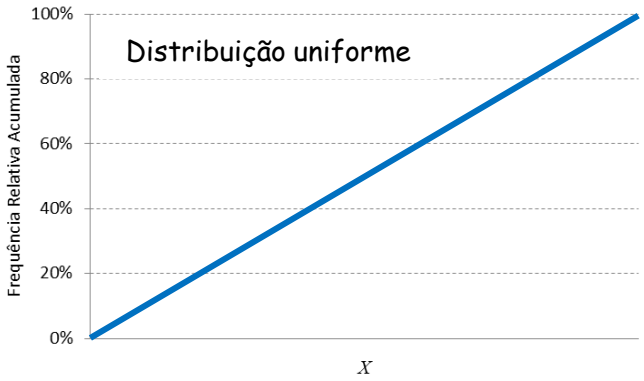
É uma alternativa bastante utilizada para representar a distribuição de valores simulados de v.a. contínuas pois não requer a discretização em classes de faixas de valores como no histograma. Além disso, pode ser usada para identificar rapidamente a mediana e quantis.

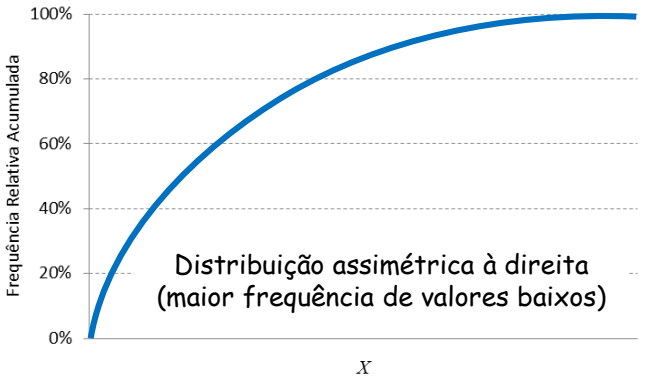
Exemplo: simulação de 50 valores de uma v.a. contínua X.

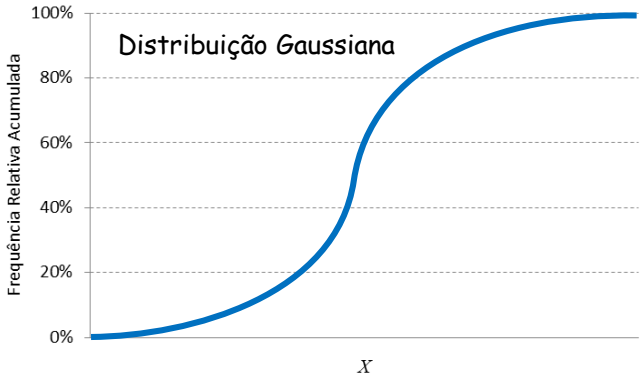


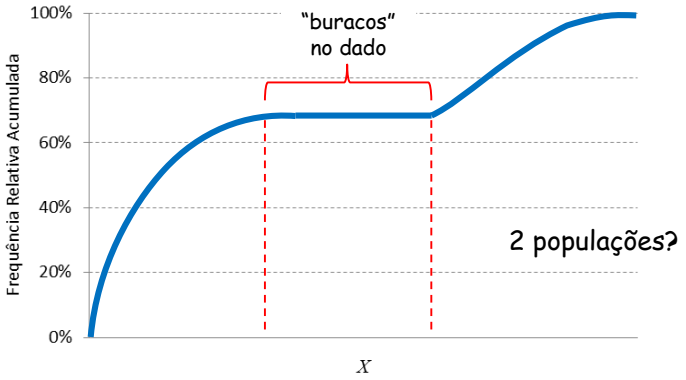
É muito útil na comparação com distribuições teóricas mas requer um certo "treinamento" para enxergar as características peculiares de cada distribuição como assimetrias e valores raros, por exemplo.

Frequência Acumulada







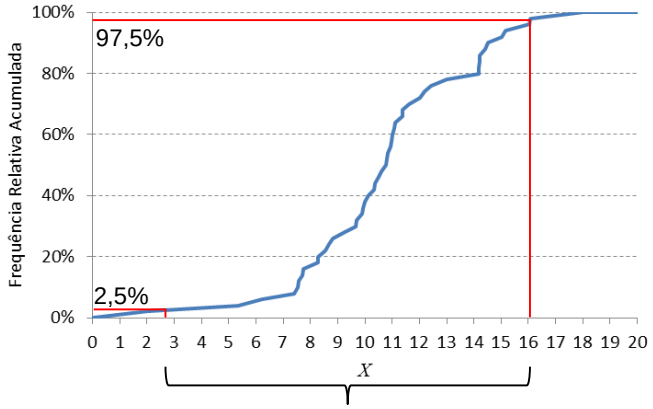


Intervalo de Credibilidade

Corresponde a um intervalo no qual há uma determinada probabilidade de um valor não observado estar contido

Há muitos intervalos de credibilidade que podem ser obtidos a partir de um conjunto de valores, sendo o mais comum, aquele obtido descartando-se uma determinada porcentagem dos valores extremos (menores e maiores)

Por exemplo, para um Intervalo de Credibilidade de 95%, descarta-se 2,5% dos valores extremos. O intervalo é definido pelos valores mínimo e máximo de novo conjunto de valores



$$P(2{,}7 < X < 16{,}1) = 95\%$$

Intervalo de Credibilidade de 95%

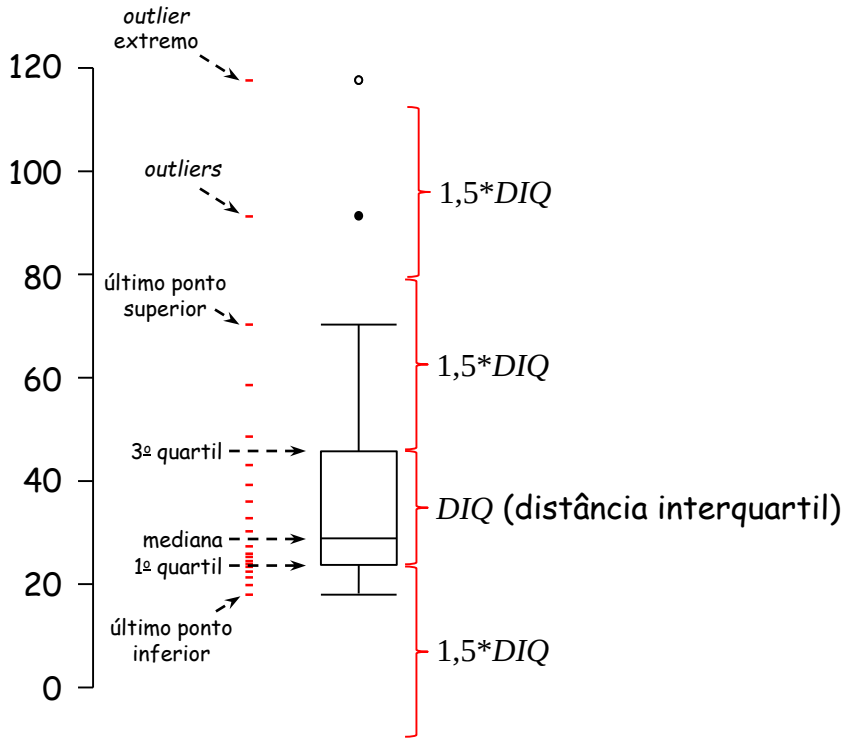
Boxplot

É uma ótima alternativa para mostrar graficamente a dispersão de valores de uma simulação e são muito úteis para comparar conjuntos de dados pois causam grande impacto visual e são fáceis de entender.

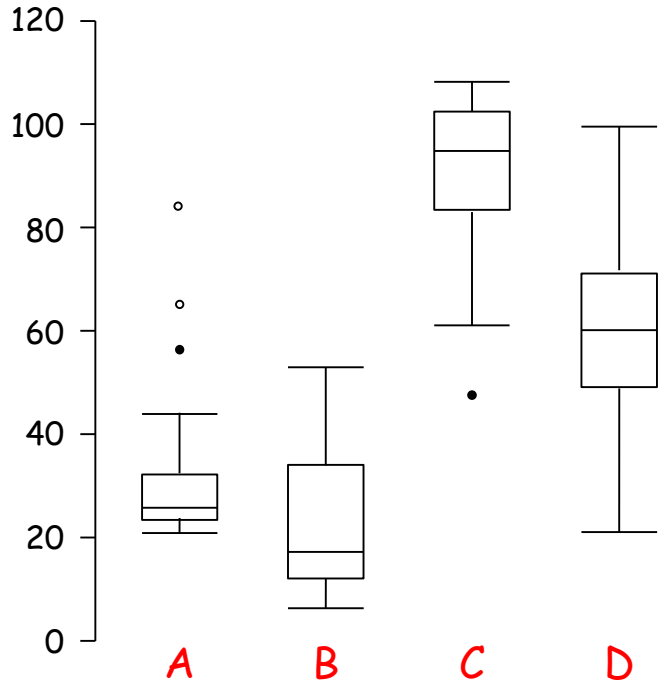
Há muitas variações de boxplot, mas em geral representam:

a) mediana
b) 1º e 3º quartis
c) mínimos e máximos
d) valores raros ("outliers")

Ex: simulação com 20 valores



Boxplot



a) qual é a distribuição mais simétrica?

D

b) qual é a distribuição mais assimétrica?

A

c) quais as 2 distribuições cujos valores mais se confundem entre si?

A e B

d) quais as 2 distribuições cujos valores mais se distinguem entre si?

B e C

Exemplo de Aplicação 1

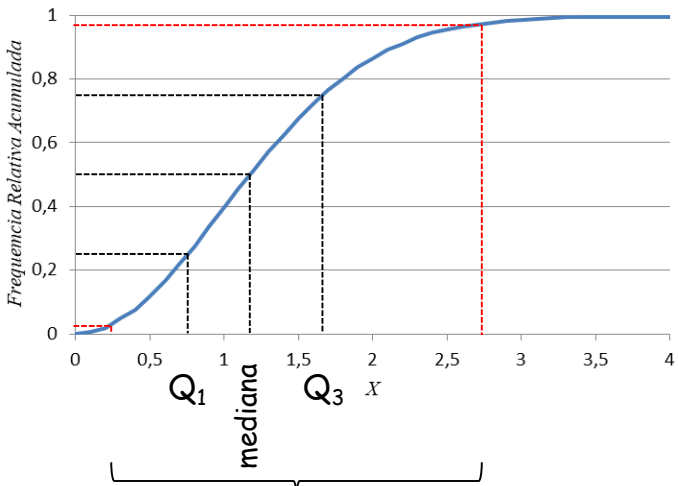
Estimar a função de probabilidade acumulada de uma v.a. resultante da raiz quadrada da soma quadrática de duas v.a. independentes normalmente distribuídas, ambas com $\mu = 0$ e $\sigma = 1$. Construa o intervalo de credibilidade de 95%.

$$X = \sqrt{Z_1^2 + Z_2^2}$$

U_1	Z_1	U_2	Z_2	X
0,3264	-0,4498	0,5952	0,2410	0,5103
0,0350	-1,8119	0,8684	1,1187	2,1295
0,7561	0,6938	0,6633	0,4215	0,8118
0,5314	0,0789	0,4063	-0,2371	0,2499
0,5864	0,2184	0,2498	-0,6751	0,7095
…	…	…	…	…
0,1881	-0,8850	0,9408	1,5614	1,7947

→ repetido 10000 vezes

X	FRAcum
0	0,0000
0,1	0,0046
0,2	0,0172
0,3	0,0407
0,4	0,0688
…	…
4	0,9959



$P(0{,}229 < X < 2{,}721) = 95\%$

Intervalo de Credibilidade de 95%

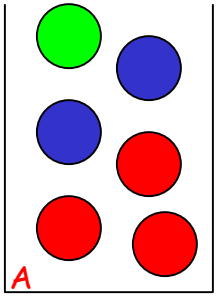
No Excel:
U1 = ALEATÓRIO()
Z1 = INV.NORMP.N(U1)

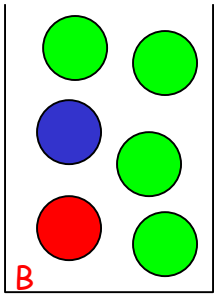
(ver Simulacao_exemplo1.xlsx)

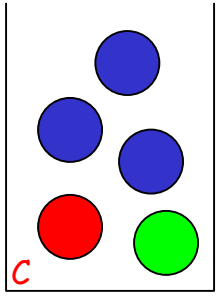
No R:
U1 <- runif(n=10000)
Z1 <- qnorm(U1)

ou Z1 <- rnorm(n=10000)

Exemplo de Aplicação 2



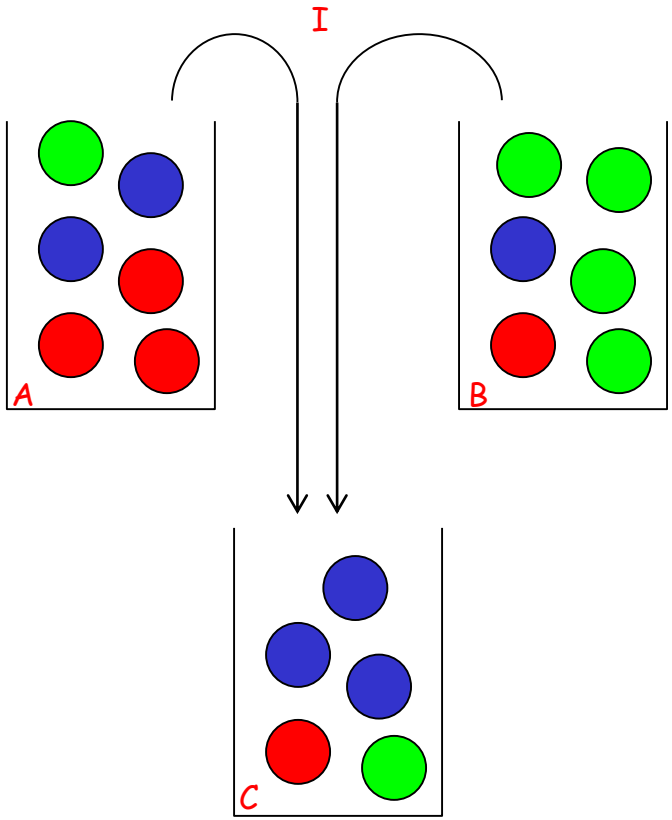




Estimar função de probabilidade de um experimento complexo (urnas)

(ver Simulacao_exemplo2.xlsx)

Exemplo de Aplicação 2

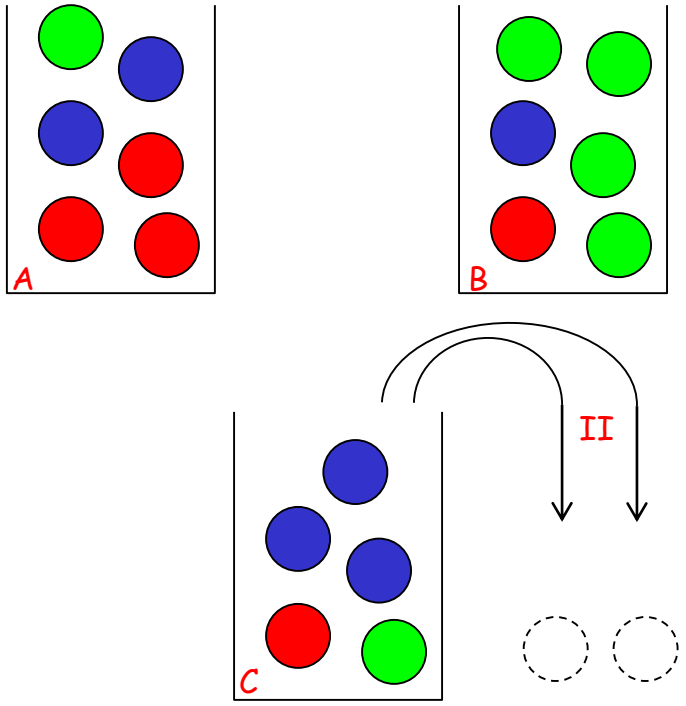


Etapas:

I) Das urnas A e B, sorteia-se uma bola de cada. As duas bolas são colocadas na urna C

(ver Simulacao_exemplo2.xlsx)

Exemplo de Aplicação 2



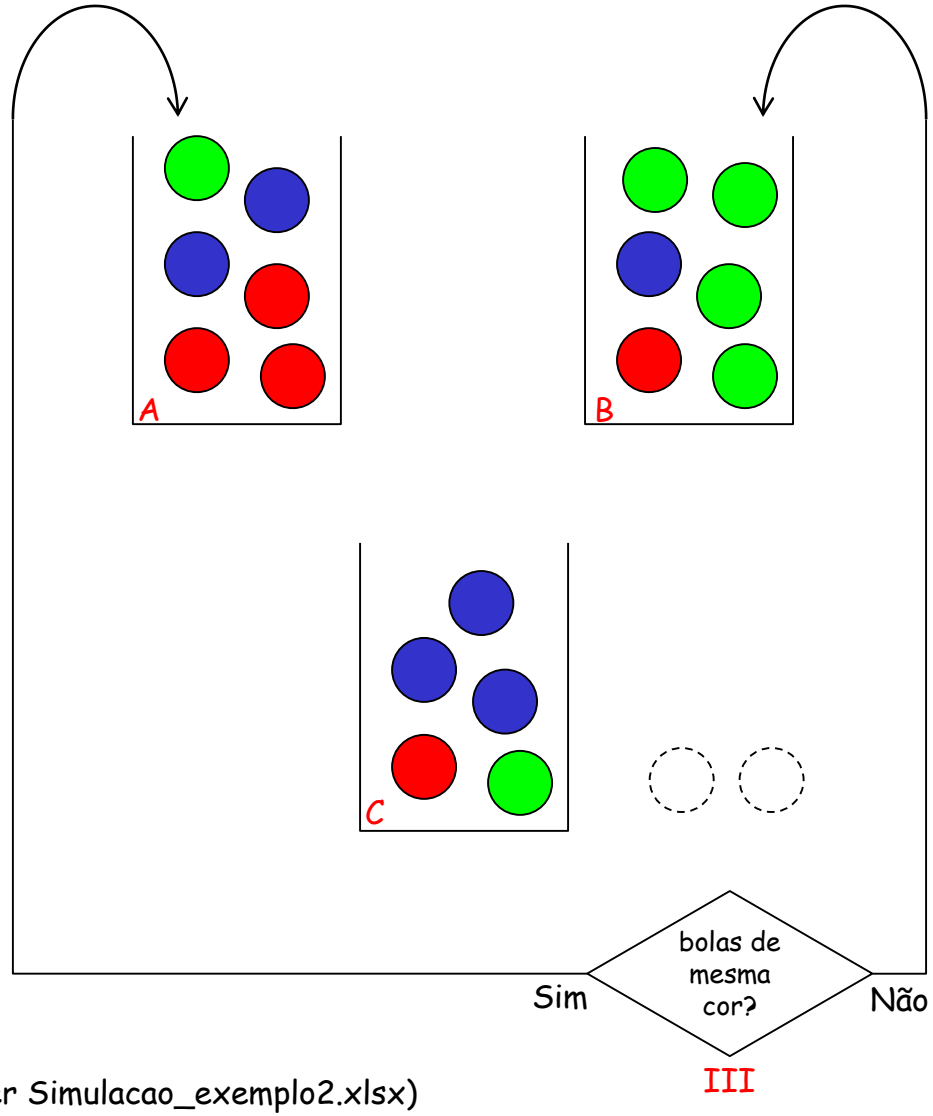
Etapas:

I) Das urnas A e B, sorteia-se uma bola de cada. As duas bolas são colocadas na urna C

II) Da urna C, sorteiam-se duas bolas (sem reposição)

(ver Simulacao_exemplo2.xlsx)

Exemplo de Aplicação 2



Etapas:

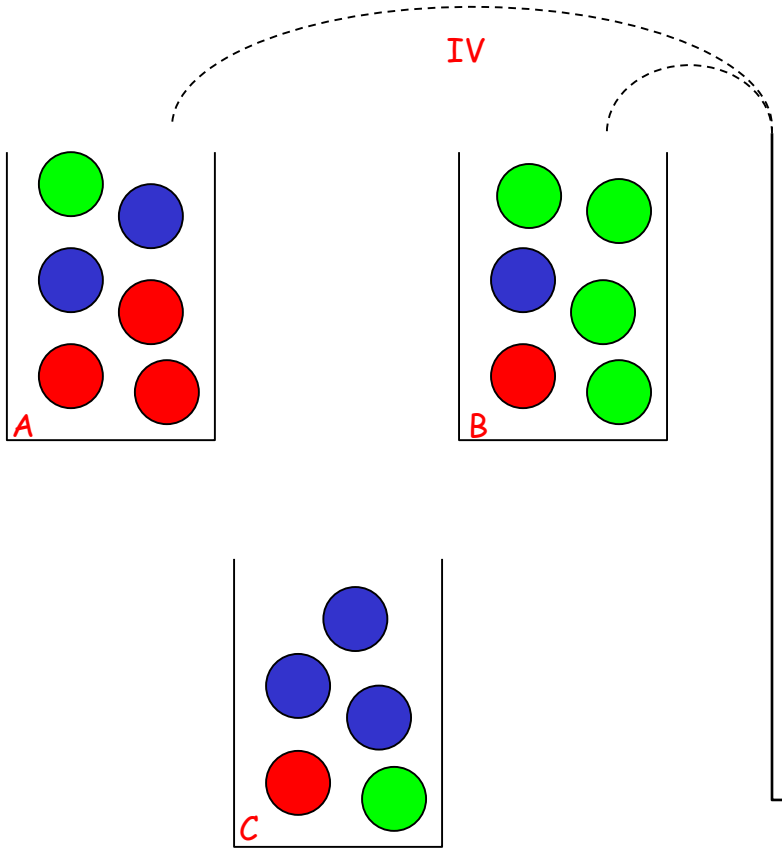
I) Das urnas A e B, sorteia-se uma bola de cada. As duas bolas são colocadas na urna C

II) Da urna C, sorteiam-se duas bolas (sem reposição)

III) Se as bolas forem da mesma cor, ambas são colocadas na urna A. Caso contrário, ambas são colocadas na urna B

(ver Simulacao_exemplo2.xlsx)

Exemplo de Aplicação 2



Etapas:

I) Das urnas A e B, sorteia-se uma bola de cada. As duas bolas são colocadas na urna C

II) Da urna C, sorteiam-se duas bolas (sem reposição)

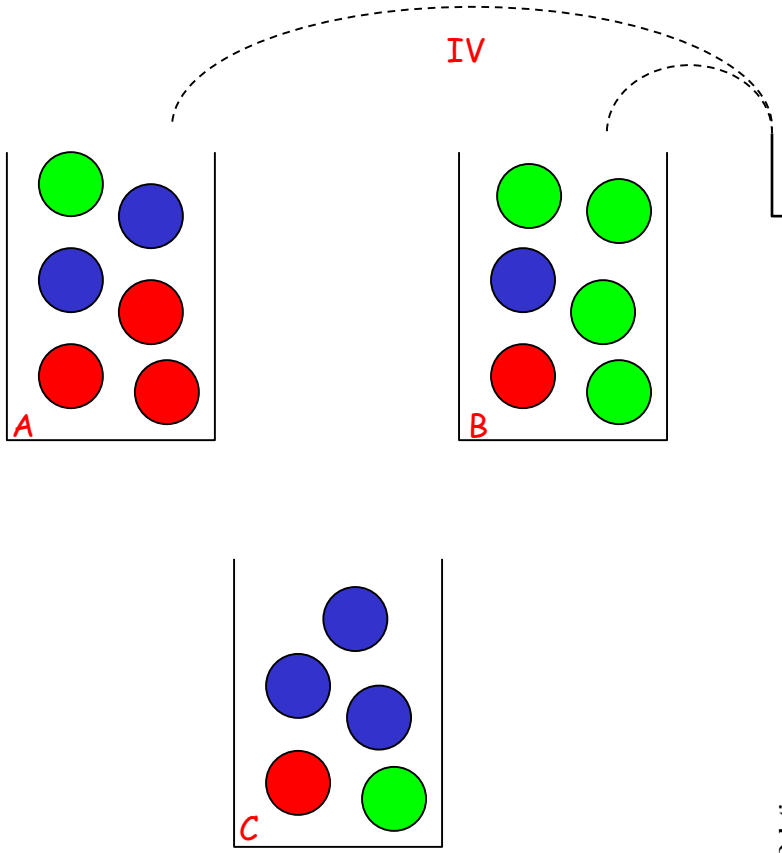
III) Se as bolas forem da mesma cor, ambas são colocadas na urna A. Caso contrário, ambas são colocadas na urna B

IV) Escolhe-se aleatoriamente a urna A ou B e dela retiram-se 5 bolas (sem reposição)

Definindo-se X como o número de bolas azuis nas 5 observações, qual a distribuição dos valores de X?

(ver Simulacao_exemplo2.xlsx)

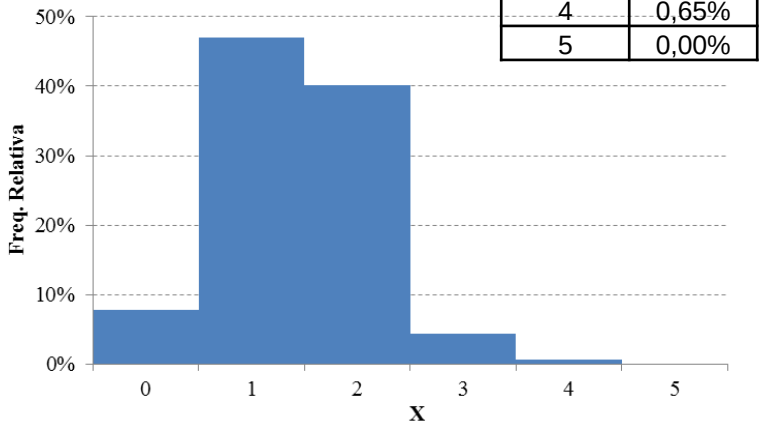
Exemplo de Aplicação 2



Definindo-se X como o número de bolas azuis nas 5 observações, qual a distribuição dos valores de X?

Após 10.000 simulações, obteve-se como resultado:

X	Freq.Rel.
0	7,85%
1	47,00%
2	40,19%
3	4,31%
4	0,65%
5	0,00%



(ver Simulacao_exemplo2.xlsx)

Exemplo de Aplicação 2 no R

```
>A<-c("R","R","R","G","B","B")
>B<-c("R","G","G","G","G","B")
>C<-c("R","G","B","B","B")

>n<-10000
>p<-rep(0,6)
>for (i in 1:n) {
>Af<-A
>Bf<-B
>sorteio1<-c(sample(A,size=1),sample(B,size=1))
>Cf<-c(C,sorteio1)

>sorteio2<-sample(Cf,size=2)
>if (sorteio2[1] == sorteio2[2]) Af<-c(A,sorteio2) else Bf<-c(B,sorteio2)
>if (runif(1,0,1) < 0.5) sorteio3<-sample(Af,5) else sorteio3<-sample(Bf,5)
>nB<-length(which(sorteio3 == "B"))
>p[nB+1]<-p[nB+1]+1
>}
>p<-p/n
>p

[1] 0.0817 0.4702 0.3939 0.0474 0.0068 0.0000
```

← Valores podem mudar a cada simulação mas tendem a se estabilizar se n for muito grande

Solução Analítica Exemplo 2

$$P(R_{A_I}) = \frac{1}{2} \quad P(G_{A_I}) = \frac{1}{6} \quad P(B_{A_I}) = \frac{1}{3} \quad P(R_{B_I}) = \frac{1}{6} \quad P(G_{B_I}) = \frac{2}{3} \quad P(B_{B_I}) = \frac{1}{6}$$

$$P(CRR) = \frac{1}{2}\frac{1}{6} = \frac{1}{12} \quad P(CRG) = \frac{1}{2}\frac{2}{3} + \frac{1}{6}\frac{1}{6} = \frac{13}{36} \quad P(CRB) = \frac{1}{2}\frac{1}{6} + \frac{1}{3}\frac{1}{6} = \frac{5}{36}$$

$$P(CGG) = \frac{1}{6}\frac{2}{3} = \frac{1}{9} \quad P(CGB) = \frac{1}{6}\frac{1}{6} + \frac{1}{3}\frac{2}{3} = \frac{1}{4} \quad P(CBB) = \frac{1}{3}\frac{1}{6} = \frac{1}{18}$$

$$P(RR_{II}) = \frac{3}{7}\frac{2}{6}\frac{1}{12} + \frac{2}{7}\frac{1}{6}\frac{13}{36} + \frac{2}{7}\frac{1}{6}\frac{5}{36} + \frac{1}{7}\frac{0}{6}\frac{1}{9} + \frac{1}{7}\frac{0}{6}\frac{1}{4} + \frac{1}{7}\frac{0}{6}\frac{1}{18} = \frac{1}{28}$$

$$P(RG_{II}) = 2\left(\frac{3}{7}\frac{1}{6}\frac{1}{12} + \frac{2}{7}\frac{2}{6}\frac{13}{36} + \frac{2}{7}\frac{1}{6}\frac{5}{36} + \frac{1}{7}\frac{3}{6}\frac{1}{9} + \frac{1}{7}\frac{2}{6}\frac{1}{4} + \frac{1}{7}\frac{1}{6}\frac{1}{18}\right) = \frac{103}{756}$$

$$P(RB_{II}) = 2\left(\frac{3}{7}\frac{3}{6}\frac{1}{12} + \frac{2}{7}\frac{3}{6}\frac{13}{36} + \frac{2}{7}\frac{4}{6}\frac{5}{36} + \frac{1}{7}\frac{3}{6}\frac{1}{9} + \frac{1}{7}\frac{4}{6}\frac{1}{4} + \frac{1}{7}\frac{5}{6}\frac{1}{18}\right) = \frac{29}{108}$$

$$P(GG_{II}) = \frac{1}{7}\frac{0}{6}\frac{1}{12} + \frac{2}{7}\frac{1}{6}\frac{13}{36} + \frac{1}{7}\frac{0}{6}\frac{5}{36} + \frac{3}{7}\frac{2}{6}\frac{1}{9} + \frac{2}{7}\frac{1}{6}\frac{1}{4} + \frac{1}{7}\frac{0}{6}\frac{1}{18} = \frac{17}{378}$$

$$P(GB_{II}) = 2\left(\frac{1}{7}\frac{3}{6}\frac{1}{12} + \frac{2}{7}\frac{3}{6}\frac{13}{36} + \frac{1}{7}\frac{4}{6}\frac{5}{36} + \frac{3}{7}\frac{3}{6}\frac{1}{9} + \frac{2}{7}\frac{4}{6}\frac{1}{4} + \frac{1}{7}\frac{5}{6}\frac{1}{18}\right) = \frac{25}{84}$$

$$P(BB_{II}) = \frac{3}{7}\frac{2}{6}\frac{1}{12} + \frac{3}{7}\frac{2}{6}\frac{13}{36} + \frac{4}{7}\frac{3}{6}\frac{5}{36} + \frac{3}{7}\frac{2}{6}\frac{1}{9} + \frac{4}{7}\frac{3}{6}\frac{1}{4} + \frac{5}{7}\frac{4}{6}\frac{1}{18} = \frac{41}{189}$$

Solução Analítica Exemplo 2

$$P(x/ARR \cup B) = \frac{1}{2}\frac{1}{28}\left(\frac{\binom{2}{x}\binom{6}{5-x}}{\binom{8}{5}} + \frac{\binom{1}{x}\binom{5}{5-x}}{\binom{6}{5}}\right) \qquad P(x/AGG \cup B) = \frac{1}{2}\frac{17}{378}\left(\frac{\binom{2}{x}\binom{6}{5-x}}{\binom{8}{5}} + \frac{\binom{1}{x}\binom{5}{5-x}}{\binom{6}{5}}\right)$$

$$P(x/ABB \cup B) = \frac{1}{2}\frac{41}{189}\left(\frac{\binom{4}{x}\binom{4}{5-x}}{\binom{8}{5}} + \frac{\binom{1}{x}\binom{5}{5-x}}{\binom{6}{5}}\right) \qquad P(x/A \cup BRG) = \frac{1}{2}\frac{103}{756}\left(\frac{\binom{2}{x}\binom{4}{5-x}}{\binom{6}{5}} + \frac{\binom{1}{x}\binom{7}{5-x}}{\binom{8}{5}}\right)$$

$$P(x/A \cup BRB) = \frac{1}{2}\frac{29}{108}\left(\frac{\binom{2}{x}\binom{4}{5-x}}{\binom{6}{5}} + \frac{\binom{2}{x}\binom{6}{5-x}}{\binom{8}{5}}\right) \qquad P(x/A \cup BGB) = \frac{1}{2}\frac{25}{84}\left(\frac{\binom{2}{x}\binom{4}{5-x}}{\binom{6}{5}} + \frac{\binom{2}{x}\binom{6}{5-x}}{\binom{8}{5}}\right)$$

$$P(x=0) = \frac{7197}{84672} \qquad P(x=1) = \frac{39343}{84672} \qquad P(x=2) = \frac{33540}{84672}$$

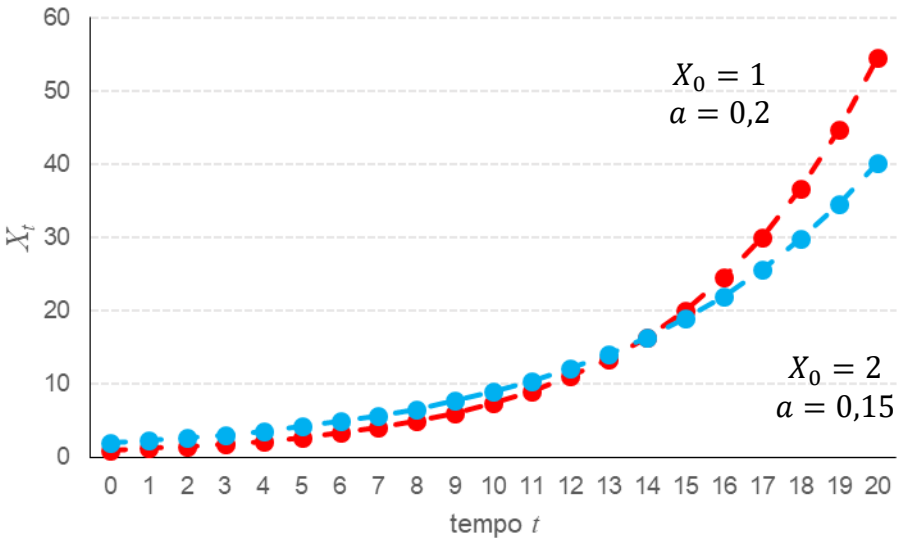
$$P(x=3) = \frac{3936}{84672} \qquad P(x=4) = \frac{656}{84672} \qquad P(x=5) = 0$$

X	FR	Prob
0	7,85%	8,50%
1	47,00%	46,47%
2	40,19%	39,61%
3	4,31%	4,65%
4	0,65%	0,77%
5	0,00%	0,00%

Processos Estocásticos

De modo bastante simplista, pode-se definir um processo estocástico como um conjunto de variáveis aleatórias que descrevem a evolução de um sistema de valores no tempo.

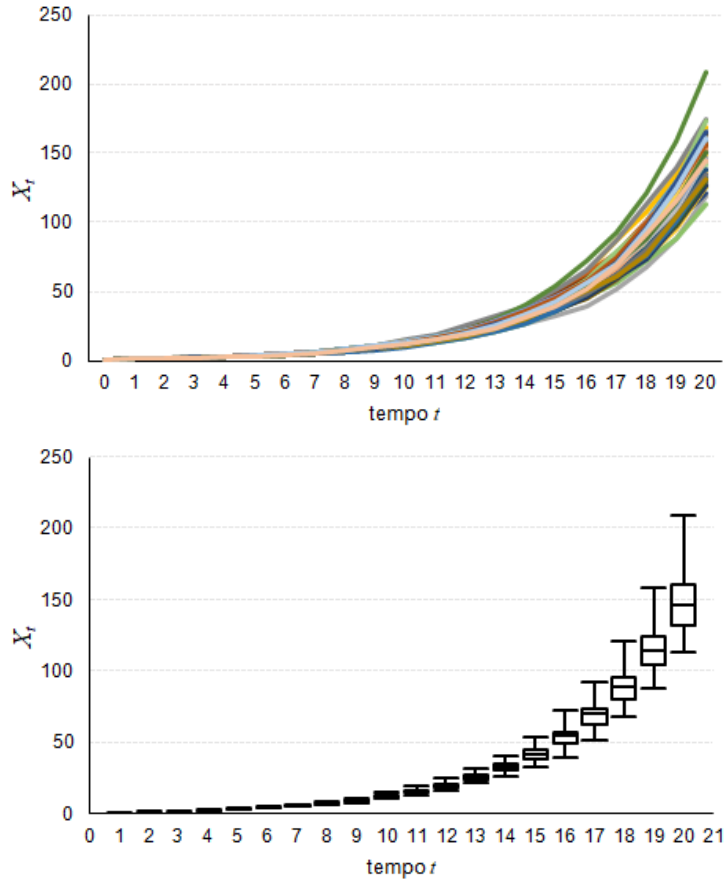
Num processo determinístico, conhecendo-se as condições iniciais, consegue-se com precisão prever como o processo irá evoluir.



$$X_t = X_{t-1} \exp(a) \qquad t > 0$$

Processos Estocásticos

Num processo estocástico, mesmo se conhecendo a condição inicial, existem muitas (as vezes infinitas) direções nas quais o processo pode evoluir.



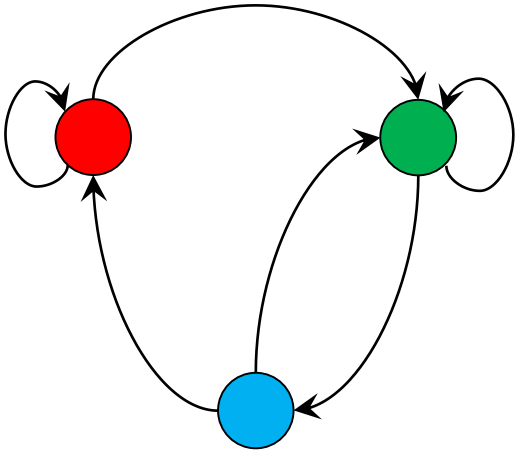
$$X_t = X_{t-1} \exp(a + \varepsilon_t) \qquad t > 0$$

componente aleatório

Cadeia de Markov

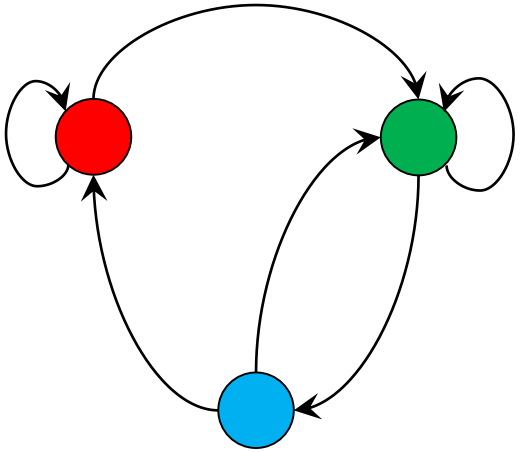
Considere que uma variável aleatória possa assumir somente valores (ou estados) discretos e que seu valor atual dependa exclusivamente de seu valor no tempo anterior (tempo discreto). Além disso, a probabilidade de que seu valor mude de *i* para *j* (transição de estados) é constante no tempo. Este processo é denominado Cadeia de Markov homogênea em tempo discreto. Há muitas variações possíveis, mas esta é a mais comum.

O processo é caracterizado por um espaço de estado, uma matriz de transição descrevendo as probabilidades de transições, e um estado inicial (ou uma distribuição de estados).



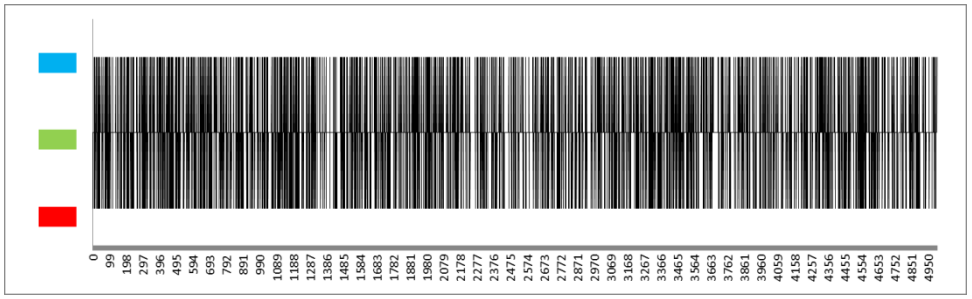
		tempo *t*		
tempo *t*-1		0,25	0,75	0
		0	0,8	0,2
		0,9	0,1	0

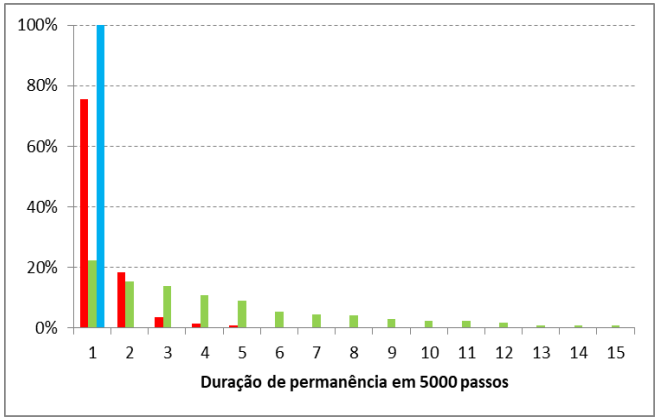
Exemplo de Aplicação 3



tempo t-1 \ tempo t			
	0,25	0,75	0
	0	0,8	0,2
	0,9	0,1	0

Por quanto tempo (passos) o estado permanece o mesmo? Qual a duração de permanência no mesmo estado mais frequente para cada estado considerando que o estado inicial é vermelho e a duração total analisada é 5000 passos? É possível o mesmo estado permanecer o mesmo por mais do que 10 passos?





(ver Simulacao_exemplo3.xlsx)