IFUSP
Instituto de Física da USP

Estudo da Probabilidade de Jogadas Repetidas

Grupo 18

Heitor Pilotto
8944752

M. Eduarda Carvalho
9849010

Renan Hirayama
8945325

Instituto de Física - USP
Prof. Zwinglio Guimarães

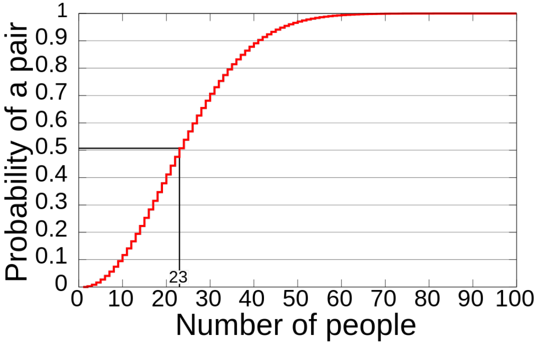
12 de Dezembro de 2017

Motivação

TEFE
2017-2

H. Pilotto
M. Carvalho
R. Hirayama

- Paradoxo do aniversário:



(*Wikipedia*)

H. Pilotto
M. Carvalho
R. Hirayama

Objetivo

- Generalizar o paradoxo do aniversário para um número arbitrário de D "dias"e de R "aniversariantes";

H. Pilotto
M. Carvalho
R. Hirayama

Objetivo

- Generalizar o paradoxo do aniversário para um número arbitrário de D "dias"e de R "aniversariantes";
- Pensando em dados, modelar FDP $\mathrm{P}(n; D, R)$

Teoria

H. Pilotto
M. Carvalho
R. Hirayama

Princípio da casa de pombo

Há um número mínimo n_{min} e um máximo n_{max} de jogadas necessárias para atingir as repetições desejadas.

Teoria

H. Pilotto
M. Carvalho
R. Hirayama

Princípio da casa de pombo

Há um número mínimo n_{min} e um máximo n_{max} de jogadas necessárias para atingir as repetições desejadas.

No nosso caso,

$$n_{min} = R \tag{1}$$

$$n_{max} = D(R-1)+1 \tag{2}$$

H. Pilotto
M. Carvalho
R. Hirayama

Usamos análise combinatória para determinar $\mathrm{P}(n; D, R = 2)$:

$\rightarrow$ D=4: [de acordo com (1), $n_{min} = R = 2$]

H. Pilotto
M. Carvalho
R. Hirayama

Teoria
Análise combinatória

Usamos análise combinatória para determinar $\mathrm{P}(n; D, R = 2)$:

$\rightarrow$ D=4: [de acordo com (1), $n_{min} = R = 2$]

- Jogada 2: $\frac{4}{4}\frac{1}{4} = \frac{4}{4^2}$ Na primeira jogada pode ser qualquer lado, na segunda tem que ser o lado que caiu na primeira;

H. Pilotto
M. Carvalho
R. Hirayama

Usamos análise combinatória para determinar $\mathrm{P}(n; D, R = 2)$:

$\rightarrow$ D=4: [de acordo com (1), $n_{min} = R = 2$]

- Jogada 2: $\frac{4}{4}\frac{1}{4} = \frac{4}{4^2}$ Na primeira jogada pode ser qualquer lado, na segunda tem que ser o lado que caiu na primeira;
- Jogada 3: $\frac{4}{4}\frac{3}{4}\frac{2}{4} = \frac{4 \cdot 3 \cdot 2}{4^3}$ Na primeira jogada pode ser qualquer lado, na segunda não pode ser o lado que caiu na primeira, na terceira tem que ser um dos lados que caíram anteriormente;

H. Pilotto
M. Carvalho
R. Hirayama

Usamos análise combinatória para determinar $\mathrm{P}(n; D, R = 2)$:

$\rightarrow$ D=4: [de acordo com (1), $n_{min} = R = 2$]

- Jogada 2: $\frac{4}{4}\frac{1}{4} = \frac{4}{4^2}$ Na primeira jogada pode ser qualquer lado, na segunda tem que ser o lado que caiu na primeira;
- Jogada 3: $\frac{4}{4}\frac{3}{4}\frac{2}{4} = \frac{4 \cdot 3 \cdot 2}{4^3}$ Na primeira jogada pode ser qualquer lado, na segunda não pode ser o lado que caiu na primeira, na terceira tem que ser um dos lados que caíram anteriormente;

$\vdots$

H. Pilotto
M. Carvalho
R. Hirayama

Usamos análise combinatória para determinar $\mathrm{P}(n; D, R = 2)$:

$\rightarrow$ D=4: [de acordo com (1), $n_{min} = R = 2$]

- Jogada 2: $\frac{4}{4}\frac{1}{4} = \frac{4}{4^2}$ Na primeira jogada pode ser qualquer lado, na segunda tem que ser o lado que caiu na primeira;
- Jogada 3: $\frac{4}{4}\frac{3}{4}\frac{2}{4} = \frac{4 \cdot 3 \cdot 2}{4^3}$ Na primeira jogada pode ser qualquer lado, na segunda não pode ser o lado que caiu na primeira, na terceira tem que ser um dos lados que caíram anteriormente;

  $\vdots$

- Jogada 5: $\frac{4}{4}\frac{3}{4}\frac{2}{4}\frac{1}{4}\frac{4}{4} = \frac{4! \cdot 4}{4^5}$

H. Pilotto
M. Carvalho
R. Hirayama

Usamos análise combinatória para determinar $\mathrm{P}(n; D, R = 2)$:

$\rightarrow$ D=4: [de acordo com (1), $n_{min} = R = 2$]

- Jogada 2: $\frac{4}{4}\frac{1}{4} = \frac{4}{4^2}$ Na primeira jogada pode ser qualquer lado, na segunda tem que ser o lado que caiu na primeira;
- Jogada 3: $\frac{4}{4}\frac{3}{4}\frac{2}{4} = \frac{4 \cdot 3 \cdot 2}{4^3}$ Na primeira jogada pode ser qualquer lado, na segunda não pode ser o lado que caiu na primeira, na terceira tem que ser um dos lados que caíram anteriormente;

$\vdots$

- Jogada 5: $\frac{4}{4}\frac{3}{4}\frac{2}{4}\frac{1}{4}\frac{4}{4} = \frac{4! \cdot 4}{4^5}$

Teoria
Análise combinatória

Usamos análise combinatória para determinar $\mathrm{P}(n; D, R = 2)$:

$\rightarrow$ D=4: [de acordo com (1), $n_{min} = R = 2$]

- Jogada 2: $\frac{4}{4}\frac{1}{4} = \frac{4}{4^2}$ Na primeira jogada pode ser qualquer lado, na segunda tem que ser o lado que caiu na primeira;
- Jogada 3: $\frac{4}{4}\frac{3}{4}\frac{2}{4} = \frac{4 \cdot 3 \cdot 2}{4^3}$ Na primeira jogada pode ser qualquer lado, na segunda não pode ser o lado que caiu na primeira, na terceira tem que ser um dos lados que caíram anteriormente;

  $\vdots$
- Jogada 5: $\frac{4}{4}\frac{3}{4}\frac{2}{4}\frac{1}{4}\frac{4}{4} = \frac{4! \cdot 4}{4^5}$

$$Rightarrow \mathrm{P}(n; 4, 2) = \frac{(n-1) \cdot 4!}{(4+1-n)! \cdot 4^n} \tag{3}$$

H. Pilotto
M. Carvalho
R. Hirayama

Usamos análise combinatória para determinar $\mathrm{P}(n; D, R = 2)$:

$\rightarrow$ D=6:

- Jogada 2: $\frac{6}{6}\frac{1}{6} = \frac{6}{6^2}$
- Jogada 3: $\frac{6}{6}\frac{5}{6}\frac{1}{6} = \frac{6 \cdot 5 \cdot 1}{6^3}$

$\vdots$

- Jogada 7: $\frac{6}{6}\frac{5}{6}\frac{4}{3}\frac{3}{6}\frac{2}{6}\frac{1}{6} = \frac{6! \cdot 6}{6^7}$

H. Pilotto
M. Carvalho
R. Hirayama

Usamos análise combinatória para determinar $\mathrm{P}(n; D, R = 2)$:

$\rightarrow$ D=6:

- Jogada 2: $\frac{6}{6}\frac{1}{6} = \frac{6}{6^2}$
- Jogada 3: $\frac{6}{6}\frac{5}{6}\frac{1}{6} = \frac{6 \cdot 5 \cdot 1}{6^3}$

$$\vdots$$

- Jogada 7: $\frac{6}{6}\frac{5}{6}\frac{4}{3}\frac{3}{6}\frac{2}{6}\frac{1}{6} = \frac{6! \cdot 6}{6^7}$

$$\Rightarrow \mathrm{P}(n; 6, 2) = \frac{(n-1) \cdot 6!}{(6+1-n)! \cdot 6^n} \tag{4}$$

De (3) e (4), é razoável inferir que, para $R = 2$, temos

$$\mathrm{P}(n; D, 2) = \frac{(n-1) \cdot D!}{(1 + D - n)! \cdot D^n} \tag{5}$$

Porém precisamos confirmar esse modelo. Para isso, fazemos primeiramente uma simulação, seguida de um experimento real.

Ilustração

H. Pilotto
M. Carvalho
R. Hirayama

Jogamos o dado de D lados n vezes até repetir algum lado R vezes:

▸ Vídeo no *YouTube*

H. Pilotto
M. Carvalho
R. Hirayama

Nosso programa consiste em criar um vetor com D zeros e sortear um inteiro aleatório entre 0 e D. A posição sorteada é incrementada em 1 e o programa checa se algum elemento do vetor vale R. No caso afirmativo, a simulação acaba e começa outro jogo, guardando n, a soma de todos os elementos do vetor. No caso negativo, ocorre outra iteração.
Chamamos a atenção para o fato de que o programa não necessita que $R = 2$, auxiliando na generalização.

Simulação
Exemplo

TEFE
2017-2

H. Pilotto
M. Carvalho
R. Hirayama

$$\begin{bmatrix}0\\0\\0\\0\\0\\0\end{bmatrix}$$

$$\longrightarrow \begin{bmatrix}1\\0\\0\\0\\0\\0\end{bmatrix}$$

$$\longrightarrow \begin{bmatrix}1\\0\\1\\0\\0\\0\end{bmatrix}$$

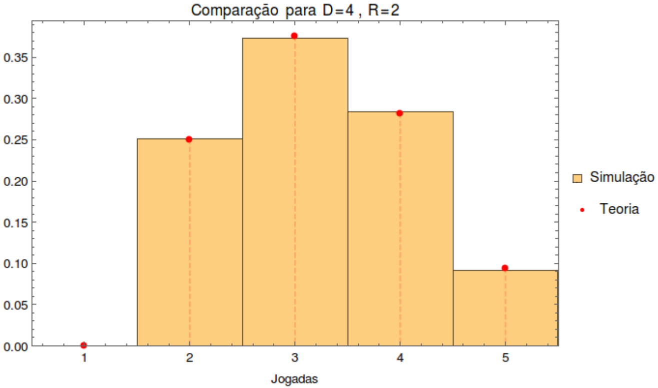
$$\longrightarrow \begin{bmatrix}1\\0\\1\\0\\0\\1\end{bmatrix}$$

$$\longrightarrow \begin{bmatrix}1\\0\\2\\0\\0\\1\end{bmatrix}$$

Nesse exemplo, $D = 6$, $R = 2$ e $n = 4$.

Simulação
Confirmação da análise teórica

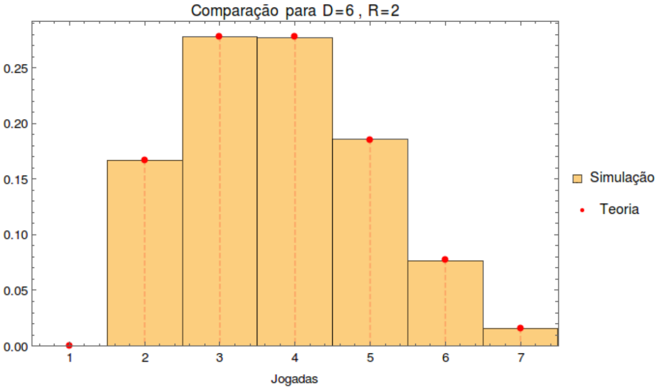
H. Pilotto
M. Carvalho
R. Hirayama



$$|z| = \left| \frac{\mu_T - \mu_S}{\sigma_{m_S}} \right| = 0.81$$

Simulação
Confirmação da análise teórica

H. Pilotto
M. Carvalho
R. Hirayama



$$|z| = 0.03$$

H. Pilotto
M. Carvalho
R. Hirayama

Simulação
Evolução e convergência

Até então, usamos 10^5 simulações de jogos para fazer os histogramas anteriores. Porém isso toma muito tempo...

H. Pilotto
M. Carvalho
R. Hirayama

Até então, usamos 10^5 simulações de jogos para fazer os histogramas anteriores. Porém isso toma muito tempo...

- Quando parar o programa?

H. Pilotto
M. Carvalho
R. Hirayama

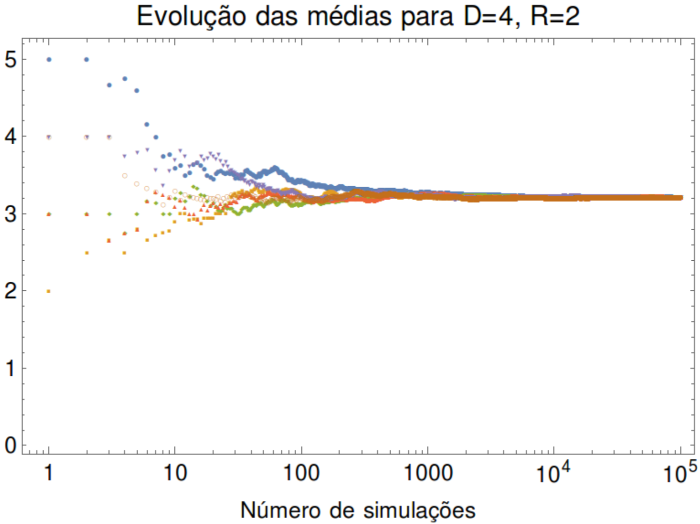
Até então, usamos 10^5 simulações de jogos para fazer os histogramas anteriores. Porém isso toma muito tempo...

- Quando parar o programa?

Podemos analisar como a média de um histograma evolui para determinar um número suficiente de simulações (cada tipo de marcador representa um histograma independente).

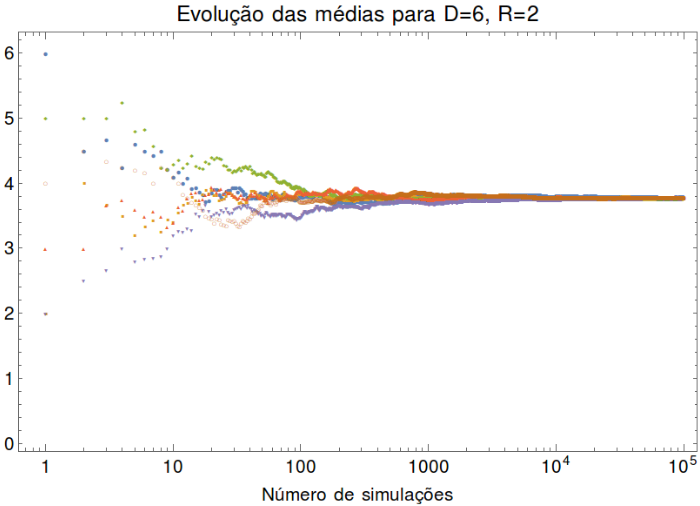
TEFE
2017-2

H. Pilotto
M. Carvalho
R. Hirayama



Simulação
Evolução e convergência

H. Pilotto
M. Carvalho
R. Hirayama



H. Pilotto
M. Carvalho
R. Hirayama

Vemos que para ambos os casos, algo em torno de 1000 simulações seriam suficientes. Por segurança, usamos $N_{sim} = 10^4$ daqui para frente.
Ademais, podemos usar isso para obter a valiosa informação de quando é razoável parar de jogar o dado num experimento real.

H. Pilotto
M. Carvalho
R. Hirayama

É interessante notar que a dispersão dos gráficos de evolução das médias é simplesmente o desvio-padrão da média,

$$\sigma_m \propto \frac{1}{\sqrt{N_{sim}}}$$

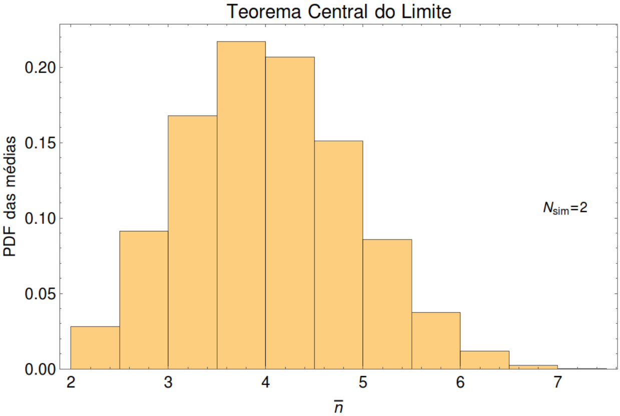
Podemos visualizar esse efeito olhando como o histograma das médias - que tende a uma gaussiana, pelo TCL - evolui com o número de simulações usados no cálculo de cada média.

H. Pilotto
M. Carvalho
R. Hirayama

Simulação

Evolução e convergência

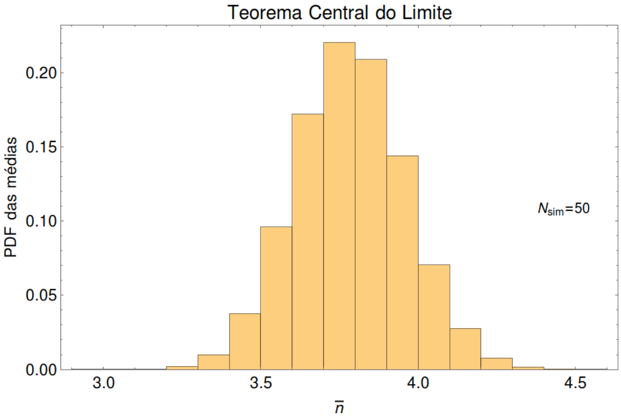
▸ Vídeo da evolução



Simulação
Evolução e convergência

H. Pilotto
M. Carvalho
R. Hirayama

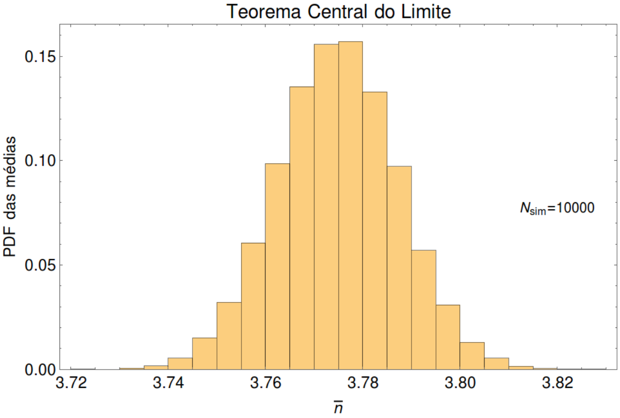
▸ Vídeo da evolução



TEFE
2017-2

H. Pilotto
M. Carvalho
R. Hirayama

▸ Vídeo da evolução



Experimento

H. Pilotto
M. Carvalho
R. Hirayama

Com a análise anterior, podemos considerar que simulação e teoria são compatíveis. Porém queremos verificar que a realidade aceita nosso modelo. Para tal, fizemos o experimento fisicamente para o caso $D = 6$.

H. Pilotto
M. Carvalho
R. Hirayama

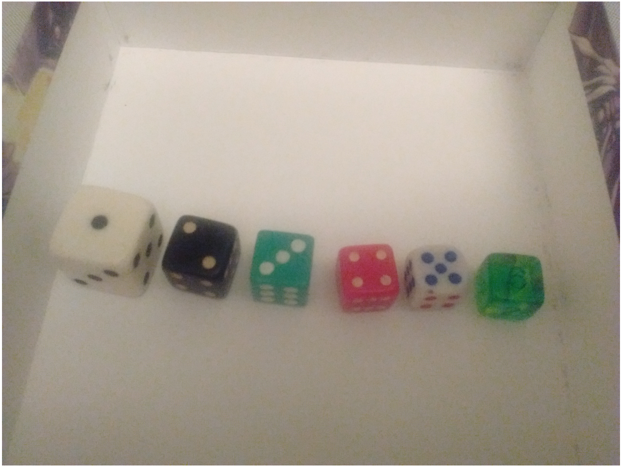
Experimento
Método para D=6, R=2

Ao invés de jogar um dado até n_{max} vezes, podemos ordenar previamente $n_{max} - 1$ dados distinguíveis e jogar todos de uma vez, o que economiza tempo. O "-1"se deve ao fato de que, se não repetiu até o $(n_{max} - 1)$-ésimo dado, o próximo com certeza repetiria, de acordo com (2).
No caso de $D = 6$, $R = 2$, precisamos de 6 dados, que foram ordenados:

Experimento
Método para D=6, R=2

TEFE
2017-2

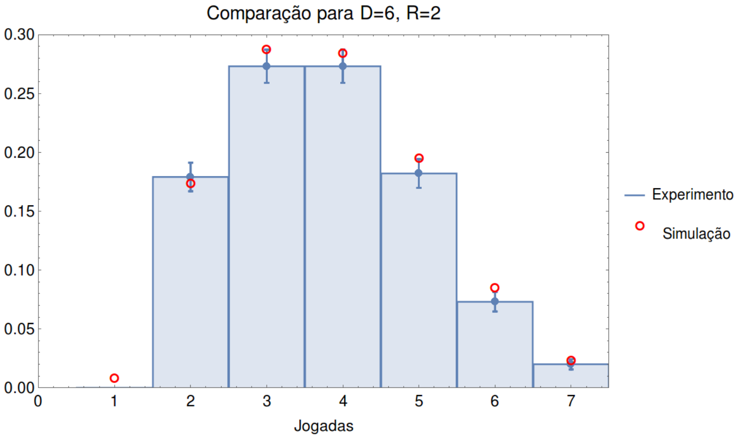
H. Pilotto
M. Carvalho
R. Hirayama



▸ Vídeo do lançamento

H. Pilotto
M. Carvalho
R. Hirayama

Experimento
Resultado



$$|z| = 0.48; \;\; \chi^2 = 4.4$$

H. Pilotto
M. Carvalho
R. Hirayama

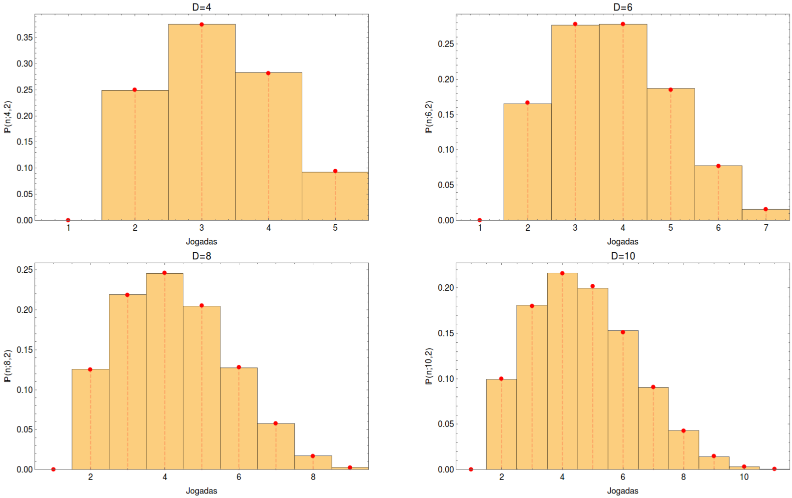
Ambos teste-z e teste de χ^2 (o NGL é simplesmente o número de bins não-nulos, pois não há ajuste) assumem valores dentro do esperado mostrando que simulação e experimento - e por consequência teoria - são razoavelmente compatíveis, ou seja, o modelo (5) é adequado.

$$|z| = 0.48; \;\; \chi^2 = 4.4$$

Deste modo, podemos usar a simulação para prever as distribuições para outros D's.

Simulação
Para outros D's

H. Pilotto
M. Carvalho
R. Hirayama

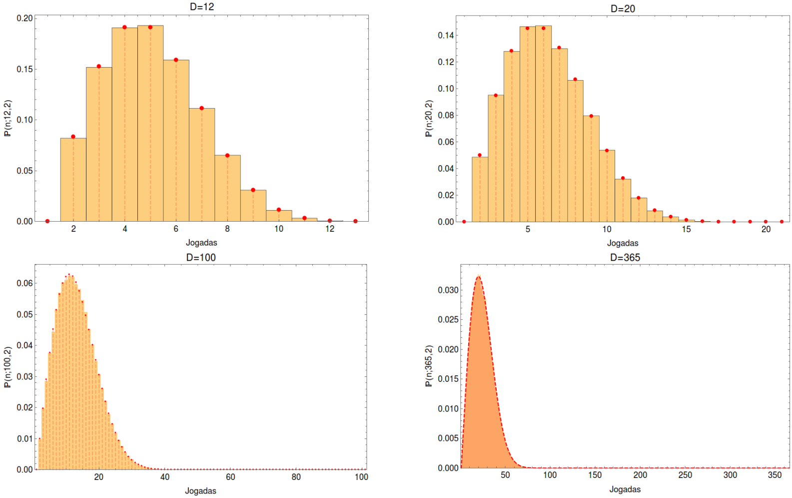


Simulação

Para outros D's

TEFE
2017-2

H. Pilotto
M. Carvalho
R. Hirayama



Simulação
Para outros D's

H. Pilotto
M. Carvalho
R. Hirayama

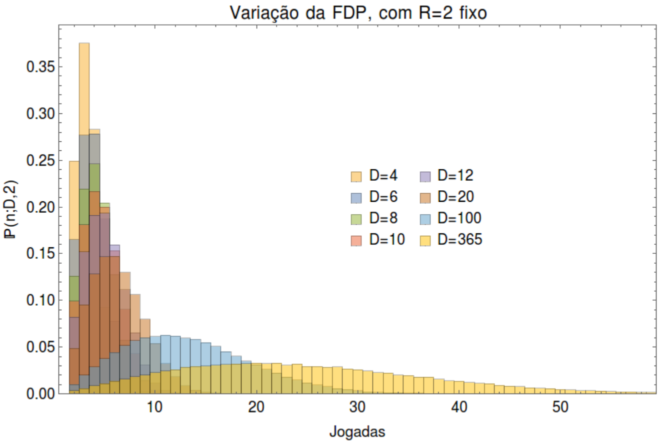


Tabela: Valores de z para R=2

D	4	6	8	10	12	20	100	365
z	0.81	0.03	1.6	0.16	1.3	0.7	1.5	0.7

Simulação

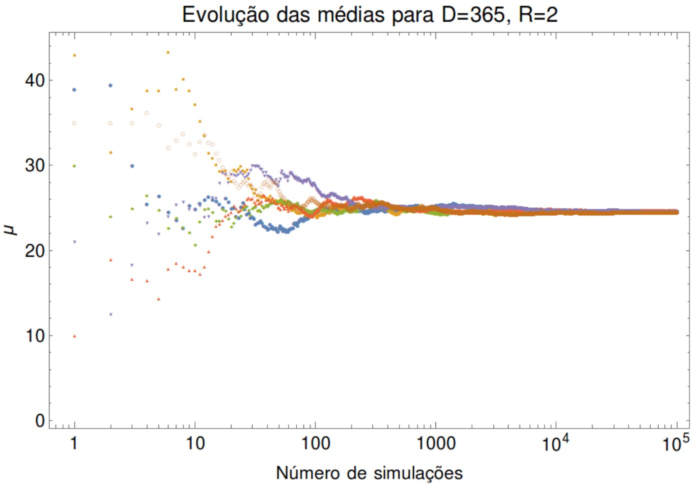
Evolução e convergência

Podemos confiar que todos os histogramas do slide anterior são representativos, pois olhando para a evolução das médias com $D = 365$, vemos que em $N_{sim} = 10^4$ o programa já convergiu. Intuitivamente, para D's menores, seriam necessárias ainda menos simulações.

H. Pilotto
M. Carvalho
R. Hirayama

Simulação

Evolução e convergência

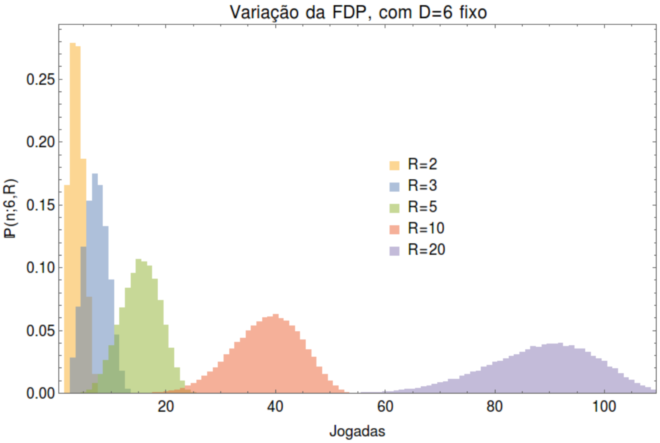


H. Pilotto
M. Carvalho
R. Hirayama

Em seguida, desejamos generalizar (5) para um número maior de repetições. Como a análise combinatória necessária é muito trabalhosa e complicada, podemos extrapolar o programa, aproveitando da boa concordância encontrada anteriormente. Fixando $D = 6$ por simplicidade:

H. Pilotto
M. Carvalho
R. Hirayama

Simulação
Para outros *R*'s

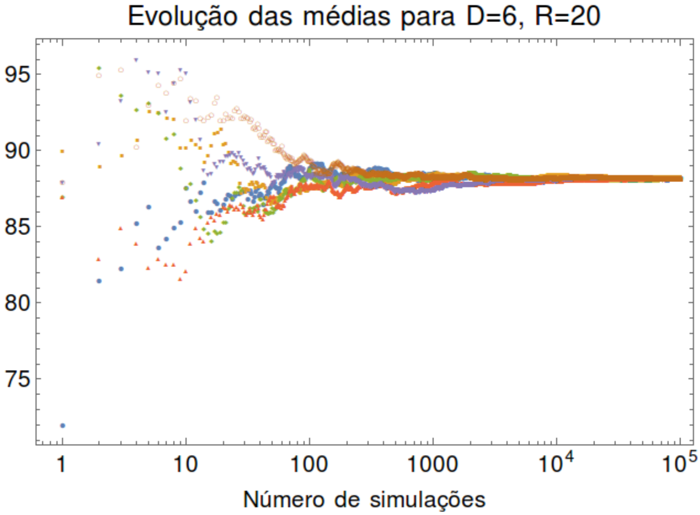


H. Pilotto
M. Carvalho
R. Hirayama

Novamente, os histogramas mostrados são confiáveis em $N_{sim} = 10^4$:

H. Pilotto
M. Carvalho
R. Hirayama

Simulação
Evolução e convergência



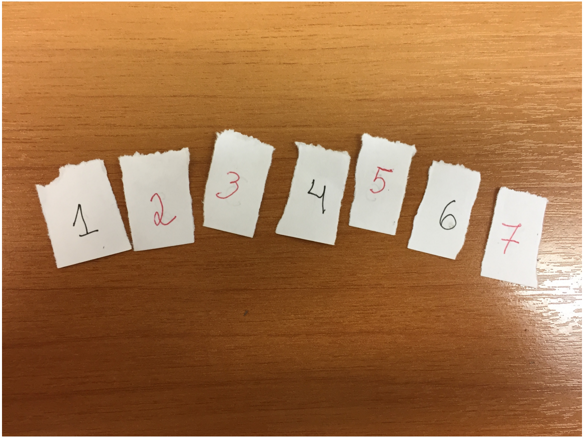
Experimento

Método para D=2, R=3 e 4

Para confirmar a validade do programa para R's diferentes, testamos o caso de 3 e 4 repetições com um "dado"de 2 lados. No caso, usamos pedaços de papel enumerados, em preto de um lado e vermelho do outro, para distinção, semelhante ao experimento anterior com os dados, fazendo isso 1000 vezes. Como as jogadas são independentes, num mesmo lançamento anotamos o resultado de $R = 3$ e $R = 4$.

Experimento

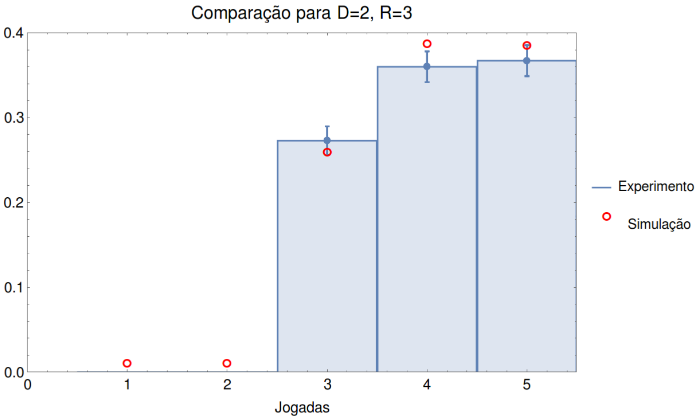
Método para D=2, R=3 e 4



Neste caso, para $R = 3 \rightarrow n = 5$ e para $R = 4 \rightarrow n = 7$.

Experimento
Resultados

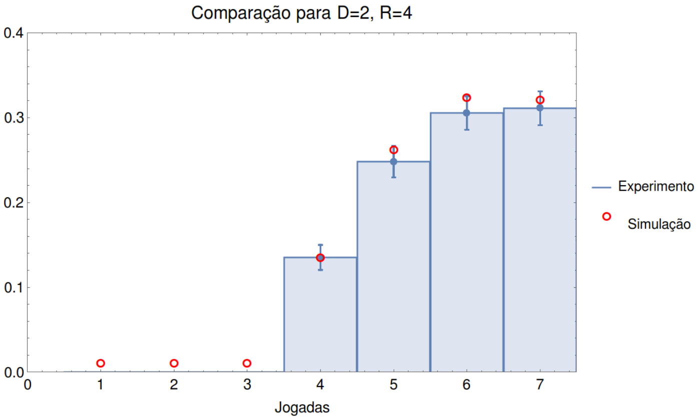
H. Pilotto
M. Carvalho
R. Hirayama



$$|z| = 1.1; \;\; \chi^2 = 2.9$$

H. Pilotto
M. Carvalho
R. Hirayama

Experimento
Resultados



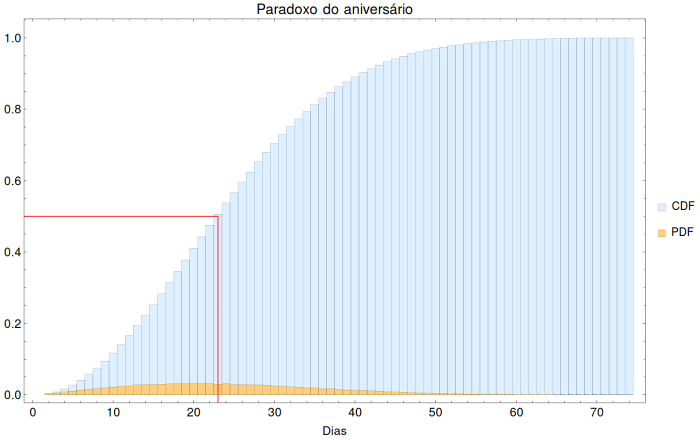
$$|z| = 0.4; \;\; \chi^2 = 1.0$$

H. Pilotto
M. Carvalho
R. Hirayama

Os histogramas indicam uma boa adequação do programa, portanto adquirimos confiança para usá-lo para prever qualquer situação que possa ser formulada deste modo. Por exemplo, podemos resgatar o gráfico do paradoxo do aniversário, encontrado na Wikipedia.

H. Pilotto
M. Carvalho
R. Hirayama

Um ano $\Longrightarrow D = 365$



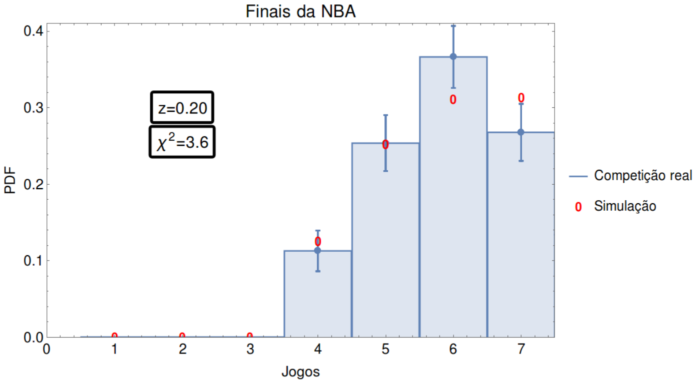
H. Pilotto
M. Carvalho
R. Hirayama

Outra aplicação interessante são as finais da NBA, em que dois times fazem um "melhor de 7", ou seja, quem vencer mais jogos dentre os 7 vence. Isso é análogo ao caso $D = 2$, $R = 4$ do nosso modelo:

Aplicações
Finais da NBA

TEFE
2017-2

H. Pilotto
M. Carvalho
R. Hirayama



(Dados do *Ticketcity*)

Uma possível extensão

H. Pilotto
M. Carvalho
R. Hirayama

- Existe alguma distribuição analítica que descreve bem a pdf genérica $\mathrm{P}(n; D, R)$?

Algo sugerido para essa extensão do trabalho é testar algumas distribuições que se comportam de maneira semelhante, variando seus parâmetros, e comparando-os com os valores respectivos de D e R, o que é possível fazer usando o programa. Isso pode ser usado para obter uma previsibilidade mais rápida dos resultados, já que a computação para valores altos de D e R toma um tempo consideravelmente grande.

Obrigado!

H. Pilotto
M. Carvalho
R. Hirayama

Extra

- Visualização do TCL em ação: Bem legal
- Programa em C++ usado no trabalho (sinta-se livre para modificar): Código