Avaliação de Geradores de Números Pseudoaleatórios Através de Técnicas da Teoria da Informação

Marcelo Q. A. Oliveira[1], **Tamer S. G. Cavalcante**[1],
Heitor S. Ramos[1], **Osvaldo Rosso**[1], **Alejandro C. Frery**[1]

[1]Laboratório de Computação Cienífica e Análise Numérica
Universidade Federal de Alagoas (LaCCAN-UFAL)
Campus A.C. Simões, Av Lourival Melo Mota, S/N - Tabuleiro do Martins
CEP: 57072-900 – Maceió - AL – Brazil

{marceloqao,tamersgc,heitor.ramos,oarosso,acfrery}@gmail.com

Resumo. *Com a evolução e crescimento do número de dispositivos conectados à rede, acompanhamos um aumento vertiginoso na quantidade de dados armazenados. Por outro lado, esta evolução não é acompanhada pela capacidade de análise destes dados através das técnicas convencionais, o que impulsiona novas áreas como Big Data. Junto com estas novas áreas surgem novos desafios como o de executar experimentos e análises estatísticas em bases de dados cada vez maiores, o que traz à tona problemas correlatos como o de gerar números aleatórios e, consequentemente, avaliar tais geradores nesses cenários desafiadores. Dentre uma vasta coleção de testes para geradores de números aleatórios, este trabalho propõe uma nova abordagem de teste não-paramétrico que utiliza técnicas de teoria da informação.*

1. Introdução

Números aleatórios perfazem uma das partes mais importantes em aplicações computacionais nos vários campos do conhecimento, como aborda [Knuth 1998]. Existem dois tipos básicos de geradores: Geradores de Números Aleatórios, do inglês *Random Number Generators* (RNGs) e Geradores de Números Pseudoaleatórios, do inglês *Pseudorandom Number Generators* (PRNGs). Computadores são máquinas determinísticas e, portanto, não é possível gerar sequências aleatórias de forma algorítmica. Para construir RNGs utiliza-se uma fonte não determinística de dados juntamente com algumas funções de processamento para produzir as sequências. A maior parte dos RNGs utiliza fenômenos físicos naturais como decaimento radioativo, ruídos térmicos em semicondutores, amostras de som em um local ruidoso, ruído no espectro eletromagnético, dentre outros; tais fontes de aleatoriedade não fazem parte do hardware usual de computadores. Como consequência, quando há necessidade de se dispor de dados que exibam aleatoriedade, lança-se mão de PRNGs.

A maneira mais conveniente e confiável de se gerar números pseudoaleatórios em máquinas determinísticas é através de algoritmos que produzem sequências com comportamento semelhante às produzidas por RNGs. Tais algoritmos produzem sequências de números não aleatórios, mas que, sob certas condições, comportam-se como aleatórios. Como define [L'Ecuyer 2007], essas sequências podem ser chamadas pseudoaleatórias, e os programas utilizados em sua produção de geradores de números pseudoaleatórios.

Esses geradores são geralmente mais convenientes do que os RNGs pois não necessitam de hardware adicional e possibilitam a reprodutibilidade.

Os PRNGs são a principal fonte de aleatoriede em jogos por computador, e no desenvolvimento de técnicas computacionais intensivas como os métodos Monte Carlo [Cipra 2000] e MCMC – *Monte Carlo Markov Chain* [Diaconis 2009]. Pela sua importância nessas e em outras técnicas, torna-se fundamental avaliar as suas propriedades.

2. Testes de Aleatoriedade

Existem duas abordagens para testar-se a capacidade de um PRNG gerar sequências de qualidade. Segundo [L'Ecuyer 1992], há testes teóricos e empíricos. Os teóricos são específicos para cada tipo de gerador, pois o analisam através das suas propriedades matemáticas. Já os testes empíricos valem-se de técnicas estatísticas para avaliar o quão boas são as sequências produzidas por um determinado gerador.

Uma sequência numérica é dita estatisticamente aleatória quando não possui padrões perceptíveis ou algum tipo de comportamento regular. Testes estatísticos são métodos de avaliar a qualidade da sequência aleatória avaliando se a distribuição do conjunto de dados experimentais adere a uma distribuição uniforme. Neste trabalho proporemos um teste não paramétrico baseado em ferramentas da teoria da informação.

Diversas são as suítes de testes na literatura: [Kendall and Babington-Smith 1938, Knuth 1998, Marsaglia 1995, NIST 1999] e [L'Ecuyer and Simard 2007]. Essas suítes são baterias de testes para cada sequência de entrada. Essa abordagem pode ter o inconveniente de requisitar muito tempo para realização de todos os testes. Adicionalmente, alguns testes são difíceis de interpretar e podem apresentar resultados conflitantes (uma mesma sequência pode passar em um teste e ser reprovada em outro).

As suítes mencionadas formulam testes de hipóteses para verificar se a distribuição da sequência aleatória de entrada é aderente a alguma distribuição conhecida. Por exemplo, a *Overlapping permutations* [Marsaglia 1995] analiza sequências de cinco números aleatórios consecutivos para verificar se as 120 permutações possíveis são igualmente frequentes na sequência de entrada. Dessa maneira, os dados são reorganizados e testados contra uma distribuição conhecida, neste caso, a distribuição uniforme.

3. Fundamentos de Teoria da Informação

A proposta deste trabalho é baseada no trabalho de [Larrondo et al. 2013], que mostra uma maneira simples e compacta de descrever o comportamento de sequências aleatórias através de métricas da teoria da informação, segundo apresentamos a seguir.

A Entropia de Shannon [Shannon 1948] é uma medida de desordem. Dada uma função de probabilidade $P = \{p_i : i = 1, \ldots, M\}$ sobre M valores, a medida de informação logarítmica de Shannon é $S[P] = -\sum_{i=1}^{M} p_i \log p_i$. Essa medida é relacionada com a informação associada ao processo físico descrito por P. Se $S[P] = 0$, então o conhecimento sobre o processo descrito pela distribuição de probabilidade é máximo e os possíveis resultados podem ser previstos com absoluta certeza. Por outro lado, o conhecimento é mínimo para a distribuição uniforme, i.e. para $p_i = M^{-1}$ para todo i, uma vez que cada resultado apresenta a mesma probabilidade de ocorrência dos outros.

López et. al. [López-Ruiz et al. 1995] introduziram a Complexidade Estatística, modificada depois por [Lamberti et al. 2004]: $\mathcal{C}_{JS}[P] = \mathcal{Q}_J[P, P_e]\mathcal{H}_S[P]$, em que $\mathcal{H}_S[P] = S[P]/S_{\max}$ é a Entropia de Shannon Normalizada ($\mathcal{H}_S \in [0, 1]$) com $S_{\max} = S[P_e] = \log M$, P_e a distribuição uniforme, e o desequilíbrio $\mathcal{Q}_J$ é definido em termos da divergência de Jensen-Shannon. Ou seja, $\mathcal{Q}_J[P, P_e] = \mathcal{Q}_0\mathcal{J}[P, P_e]$, com $\mathcal{J}[P, P_e] = S\left[(P + P_e)/2\right] - (S[P] - S[P_e])/2$ e $\mathcal{Q}_0$ é uma constante de normalização igual ao inverso do valor máximo possível de $\mathcal{J}[P, P_e]$, de modo que $\mathcal{Q}_J \in [0, 1]$.

O valor da Complexidade de um sistema é nulo nas duas situações opostas extremas: ou no conhecimento perfeito, ou na aleatoriedade completa. Qualquer outro tipo de sistema se situará entre essas configurações extremas.

A avaliação da Entropia e da Complexidade Estatística exige a definição preliminar de uma distribuição de probabilidade P. Bandt e Pompe introduziram um método simples para definir essa distribuição de probabilidade a partir de séries temporais, levando em consideração a causalidade temporal da dinâmica dos processos [Bandt and Pompe 2002]. Neste trabalho, utilizamos o plano entropia-complexidade (HC) para avaliar a qualidade dos geradores pseudoaleatórios: o espaço de representação da Complexidade Estatística $\mathcal{C}_{JS}$ em função da Entropia $\mathcal{H}_S$ do sistema, utilizando a distribuição de probabilidade de [Bandt and Pompe 2002] para estimar os quantificadores.

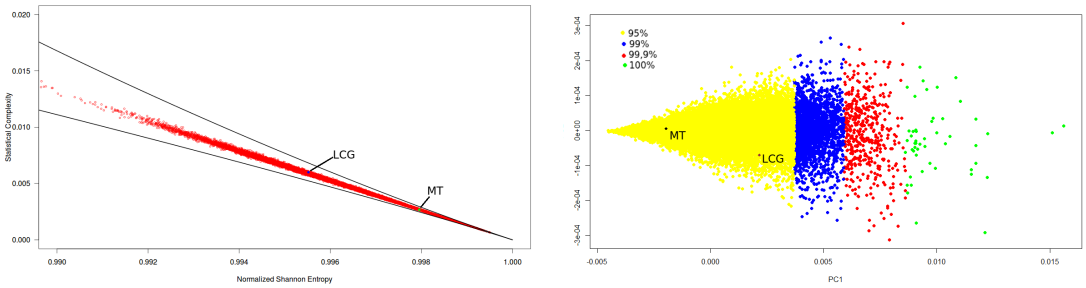
A Complexidade está limitada por valores mínimo $C_{\min}$ e máximo $C_{\max}$ para cada valor de $\mathcal{H}_S$. Esses valores podem ser calculados por meio uma análise geométrica do espaço de probabilidade [Martín et al. 2006], e só dependem de como a Entropia e o Desequilíbrio são calculados.

O Plano Complexidade-Entropia tem sido utilizado para caracterizar diferentes tipos de processos. [Rosso et al. 2007] o utilizaram para distinguir sistemas caóticos de processos estocásticos. [Larrondo et al. 2013] descrevem a primeira abordagem para avaliação de sequências pseudoaleatórias através do plano HC, entretanto, não foi criado um teste estatístico. Sabemos que a sequência ideal seria mapeada em $(1, 0)$ no plano HC, mas ao gerarmos sequências finitas, não há como obter esse valor, por melhor que seja o gerador. Dessa forma, dada uma sequência pseudoaleatória, queremos realizar um teste de hipóteses confrontando-a com uma sequência verdadeiramente aleatória obtida a partir de um RNG que é usado como referência.

4. Teste Não-Paramétrico Baseado em Ferramentas de Teoria da Informação

Com o objetivo de ter uma referência foram utilizados dados oriundos de um gerador real. Os dados foram fornecidos pelo grupo de Processamento de Informação Quântica do Instituto de Tecnologia Max Plank, num arquivo binário de aproximadamente $200\,\mathrm{Mibit}$ obtido segundo o processo descrito em [Gabriel et al. 2010]. Tais dados foram mapeados como uma sequência de 10^8 números aleatórios no intervalo $(0, 1)$, e então particionados em 10^5 sequências de 10^3 elementos cada uma. Posteriormente, foram calculados os valores da entropia e da complexidade estatística para cada uma das subsequências, resultando em 10^5 pares de pontos no plano (H, C).

Como apontado por [Larrondo et al. 2013] uma sequência aleatória ideal produziria o valor $(1, 0)$ no plano HC. Dessa maneira, a fim de avaliar a qualidade de um



(a) Plano HC (b) PCA

Figura 1. Teste não paramétrico para uma sequência de 1000 pontos dos geradores Mersenne Twister e Congruencial Linear

PRNG qualquer, elaboramos um teste de hipóteses não-paramétrico para medir a qualidade da sequência gerada pelo PRNG através da posição do ponto observado no plano (H, C). A medida é feita comparando o ponto com aqueles obtidos das sequências de mesmo tamanho produzidas pelo RNG de [Gabriel et al. 2010]. Neste trabalho testamos duas sequências de tamamnho 10^3 produzidas pelos geradores Mersenne Twister (MT) e Congruencial Linear (LCG), ambos implementados no `R` [R Core Team 2015].

Os resultados estão apresentados na Fig. 1. Na Fig. 1(a) observamos que o ponto produzido pelo gerador MT está mais próximo do gerador ideal $(1, 0)$ que o da sequência gerada por LCG. A Fig. 1(b) mostra os mesmos pontos após a aplicação da transformação de componentes principais, para fins de visualização. Nesta figura, as regiões de confiança para o RNG aos níveis $95\,\%$, $99\,\%$ e $99.9\,\%$ estão identificadas pelas cores amarela, azul e vermelha, respectivamente. Assim, ambos os geradores estão na região de confiança ao $95\,\%$ do RNG, portanto não há evidência suficiente para rejeitar a hipótese de que ambas as sequências de tamanho 10^3 produzidas pelo MT e LCG apresentam propriedades semelhantes à do RNG apresentado. Os p-valores para os testes acima são 0.1349 e 0.850 para os geradores LCG e MT, respectivamente.

5. Conclusão

Neste trabalho estudamos o comportamento dos PRNGs Mersenne Twister (MT) e Congruencial Linear (LCG) com um teste baseado na comparação com os pontos gerados por um RNG, o descrito em [Gabriel et al. 2010], no plano HC. Concluímos que para sequências consideradas pequenas como as de tamanho 10^3 utilizadas neste trabalho, não há evidência estatística para rejeitar a hipótese de que ambos são puramente aleatórios. Em trabalhos futuros, aumentaremos o tamanho das sequências, bem como a quantidade de PRNGs testados para verificar seu comportamento e ainda, disponibilizaremos o teste como um pacote R [R Core Team 2015].

Referências

Bandt, C. and Pompe, B. (2002). Permutation Entropy: A Natural Complexity Measure for Time Series. *Physical Review Letters*, 88(17):174102–174106.

Cipra, B. A. (2000). The best of the 20th century: Editors name top 10 algorithms. *SIAM News*, 33(4):1–2.

Diaconis, P. (2009). The Markov Chain Monte Carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):180–205.

Gabriel, C., Wittmann, C., Sych, D., Dong, R., Mauerer, W., Andersen, U. L., Marquardt, C., and Leuchs, G. (2010). A generator for unique quantum random numbers based on vacuum states. *Nature Photonics*, 4(10):711–715.

Kendall, M. G. and Babington-Smith, B. (1938). Randomness and other random sampling numbers. *Journal of the Royal Statistical Society*, 101:147–166.

Knuth, D. E. (1998). *The Art of Computer Programming, Volume 2: (2Nd Ed.) Seminumerical Algorithms*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA.

Lamberti, P., Martin, M., Plastino, A., and Rosso, O. (2004). Intensive entropic nontriviality measure. *Physica A: Statistical Mechanics and its Applications*, 334(1):119–131.

Larrondo, H. A., De Micco, L., Gonzalez, C. M., Plastino, A., and Rosso, O. A. (2013). Statistical Complexity of Chaotic Pseudorandom Number Generators — BenthamScience. *Concepts and Recent Advances in Generalized Information Measures and Statistics*, pages 283–308.

L'Ecuyer, P. (1992). Testing Random Number Generators. In *Proceedings of the 1992 Winter Simulation Conference*, pages 305–313. IEEE Press.

L'Ecuyer, P. (2007). *Random Number Generation*, pages 93–137. John Wiley & Sons, Inc.

L'Ecuyer, P. and Simard, R. (2007). TestU01: A C Library for Empirical Testing of Random Number Generators. *ACM Transactions on Mathematical Software*, 33(4):Article 22.

López-Ruiz, R., Mancini, H. L., and Calbet, X. (1995). A statistical measure of complexity. *Physics Letters A*, 209(5-6):321–326.

Marsaglia, G. (1995). Diehard. `http://stat.fsu.edu/pub/diehard/`. Acessado em 11/2014.

Martín, M., Plastino, A., and Rosso, O. (2006). Generalized statistical complexity measures: Geometrical and analytical properties. *Physica A: Statistical Mechanics and its Applications*, 369(2):439 – 462.

NIST (1999). Nist statistical test suite. `http://www.itl.nist.gov/div893/staff/soto/jshome.html`. Acessado em 11/2014.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosso, O. A., Larrondo, H. A., Martin, M. T., Plastino, A., and Fuentes, M. A. (2007). Distinguishing noise from chaos. *Physical Review Letters*, 99(15).

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell system technical journal*, 27.