COMPARAÇÃO DOS TESTES DE ADERÊNCIA À NORMALIDADE KOLMOGOROV-SMIRNOV, ANDERSON-DARLING, CRAMER–VON MISES E SHAPIRO-WILK POR SIMULAÇÃO

Vanessa Bielefeldt Leotti, Universidade Federal do Rio Grande do Sul, vleotti@yahoo.com.br
Alan Rodrigues Birck, Universidade Federal do Rio Grande do Sul, alanbirck@hotmail.com
João Riboldi, Universidade Federal do Rio Grande do Sul, riboldi@mat.ufrgs.br

RESUMO: Uma grande quantidade de métodos estatísticos supõe que os dados provenham de uma Distribuição Normal, pois este fato permite que seja realizada a maioria das técnicas de inferência estatística conhecidas. Para avaliar o atendimento à esta suposição, existem testes não-paramétricos como Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling e Shapiro-Wilk, que verificam se a distribuição de um conjunto de dados adere à Distribuição Normal. Neste trabalho, através de simulação Monte Carlo, a eficiência desses quatro testes foi comparada, sob diferentes distribuições e diferentes tamanhos de amostra. As simulações foram feitas no software SAS, com 1000 replicações para cada tamanho de amostra e distribuição especificada. Os resultados mostraram equivalência dos quatro testes para dados normais, com exceção do critério de Kolmogorov-Smirnov, que se mostrou inferior, e para dados não-normais o teste de Shapiro-Wilk mostrou-se sempre superior, concluindo-se então que este é aparentemente o melhor teste de aderência à normalidade.
Palavras chave: Normalidade, testes de aderência, simulação.

1. INTRODUÇÃO

Uma variável aleatória, seja idade de um grupo de pessoas, seja, por exemplo, o tempo de vida útil de determinado equipamento, assume uma distribuição de freqüências específica. As distribuições de freqüências podem apresentar formas variadas (Callegari-Jacques).

Na literatura estatística encontramos muitas distribuições teóricas. Essas são modelos que procuram representar o comportamento de determinado evento em função da freqüência de sua ocorrência. No caso das variáveis contínuas, esse evento será um intervalo de valores. As distribuições de freqüências são, em verdade, distribuições de probabilidade, onde para um evento teremos uma probabilidade de ocorrência associada. Em outras palavras, podemos inferir com que probabilidade determinado evento pode ocorrer novamente.

Os modelos tornam certas inferências exeqüíveis e por vezes mais poderosas. A modelagem é um recurso amiúde utilizado. Não podemos medir o volume de determinado sólido se não considerarmos que *supostamente* seja construído de igual maneira a um modelo pré-concebido. Uma vez atribuída sua forma a um modelo podemos estimar, através de certos parâmetros, seu volume com

razoável grau de aproximação. É praticamente impossível medir o volume de uma nuvem que passeia pelo céu. Entretanto, se concebermos sua forma como um elipsóide, podemos estimar seu volume de maneira algébrica, sabendo apenas sua altura, largura e profundidade.

Não é diferente com as distribuições de probabilidade. Assumir que determinado grupo de dados se distribui conforme um modelo nos permite realizar estimativas sem precisar da totalidade das informações. Invariavelmente, nos surge uma dúvida: Como estimar se a distribuição de um grupo de dados concorda com um particular modelo teórico?

Nosso objeto de estudo é a mais importante distribuição de probabilidade: distribuição normal ou gaussiana. É sabido que a suposição de normalidade na distribuição dos dados é exigida para a realização de muitos métodos estatísticos, assim como a suposição de independência entre as observações. A pergunta permanece: de quais maneiras podemos estimar se distribuição dos dados que estamos estudando se ajusta a uma distribuição normal?

Existem disponíveis alguns testes para verificar a suposição de normalidade dos dados, Anderson-Darling, Cramer-von Mises, Kolmogorov-Smirnov e Shapiro-Wilk, bem como recursos gráficos, como histograma e normal plot. Essas estatísticas têm metodologia diferente para realização do teste de hipóteses, assim se faz necessário um estudo mais aprofundado de qual destas é mais adequada.

Os testes de Anderson-Darling, Cramer-von Mises e Kolmogorov-Smirnov são baseados na função de distribuição empírica (FDE) dos dados, e apresentam vantagens sobre o teste de aderência qui-quadrado, incluindo maior poder e invariância em relação aos pontos médios dos intervalos escolhidos. O teste de Kolmogorov-Smirnov pertence à classe suprema de estatísticas baseadas na FDE, pois trabalha com a maior diferença entre a distribuição empírica e a hipotética. Os testes Anderson-Darling e Cramer-von Mises pertencem à classe quadrática de estatísticas baseadas na FDE, pois trabalham com as diferenças quadráticas entre a distribuição empírica e a hipotética. No entanto, o teste de Shapiro-Wilk baseia-se nos valores amostrais ordenados elevados ao quadrado e tem sido o teste de normalidade preferido por mostrar ser mais poderoso que diversos testes alternativos.

2. METODOLOGIA

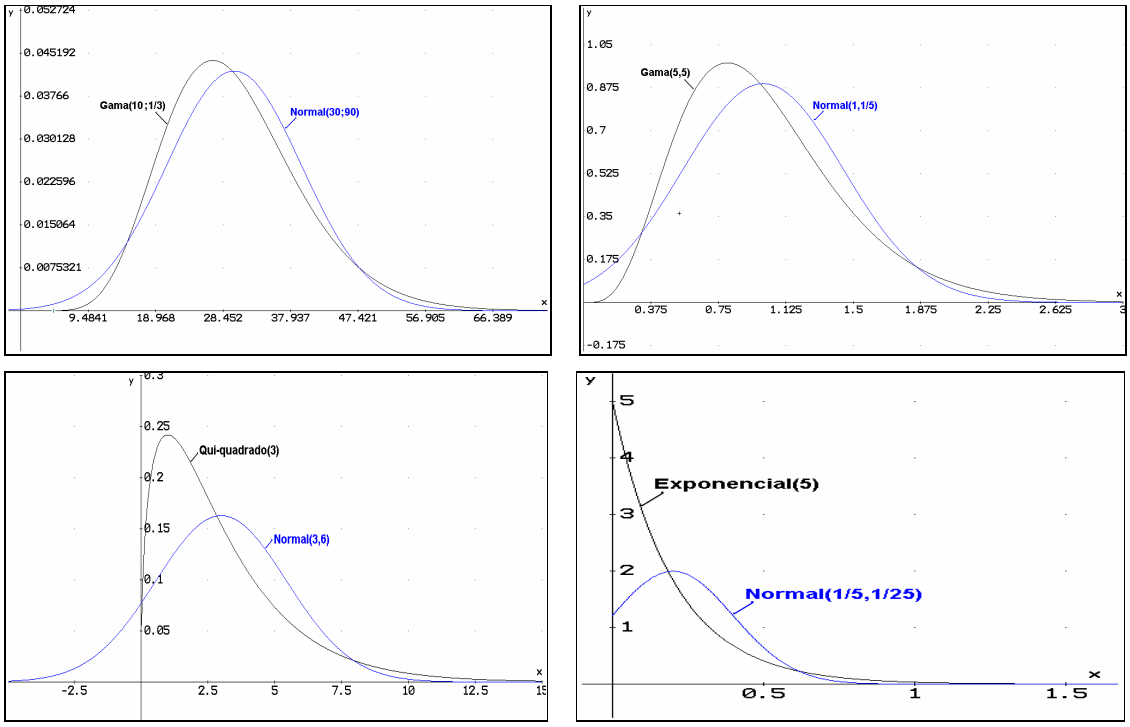
Primeiramente, definiram-se os tamanhos de amostras e distribuições a serem estudadas. Os tamanhos de amostra escolhidos foram: n=15, 30, 50 e 100. As distribuições simuladas foram:

- $Normal(0,1)$: para avaliar a taxa de erro tipo I dos testes (rejeição de H_0 dada H0 verdadeira);
- $Gama\left(10, \frac{1}{3}\right)$: por ser uma distribuição aproximadamente simétrica;
- $Gama(5,5)$: por ser uma distribuição levemente assimétrica à esquerda;

- $Qui-quadrado(3)$: por ser uma distribuição acentuadamente assimétrica à esquerda;
- $Exponencial(5)$: por ser uma distribuição assimétrica e monótona decrescente.

Os gráficos abaixo, gerados no software Derive, mostram as funções densidade de probabilidade das distribuições acima especificadas, bem como as das distribuições normais, com média e variância correspondente.

Comparação das distribuições geradas com a Normal



A escolha destas distribuições levou em conta as possibilidades do software SAS e o fato de que não haveria sentido em simular distribuições como a Binomial (que é uma distribuição de variáveis discretas) ou a Beta (que é uma distribuição de variáveis binárias).

Para cada tamanho de amostra e distribuição escolhida, fez-se 1000 replicações no software. Para cada amostra, aplicavam-se os testes e gravava-se o *p-value* resultante cada um. Após as 1000 replicações terem sido geradas, analisava-se a freqüência com que os *p-values* encontravam-se nas classes:

- $0 \leq p-value \leq 0{,}01$: classe que corresponde à rejeição de H_0: Os dados provém de uma distribuição Normal, com nível de significância $\alpha = 0{,}01$.
- $0{,}01 < p-value \leq 0{,}05$: conjuntamente com a classe anterior, corresponde à rejeição de H_0 com $\alpha = 0{,}05$.

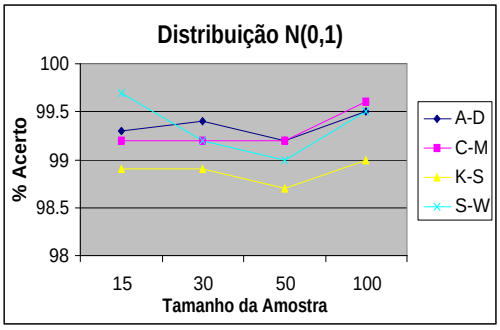
- $0{,}05 < p-value \leq 0{,}10$: conjuntamente com as duas classes anteriores, corresponde à rejeição de H_0 com $\alpha = 0{,}10$.
- $0{,}10 < p-value \leq 1$: corresponde à não rejeição de H_0 com $\alpha = 0{,}10$.

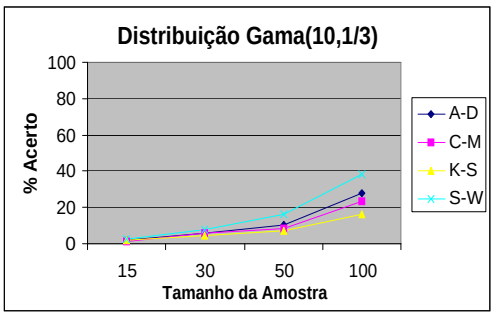
Neste trabalho adotou-se o nível de significância 0,01, pois em testes de suposições de modelos, querer se rejeitar o atendimento à suposição somente em casos extremos.

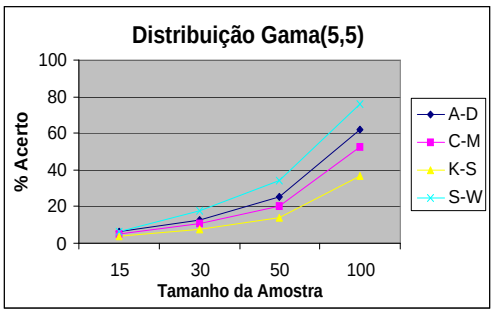
Para avaliar a precisão dos testes em relação ao tamanho da amostra, fizeram-se gráficos no Excel com os tamanhos de amostra no eixo horizontal, e o percentual de acerto do teste no eixo vertical. Quando a distribuição é Normal, este percentual de acerto eqüivale às freqüências percentuais dos *p-values* pertencentes às três últimas classes, ou seja, o percentual de não rejeição de H_0; caso contrário, é o percentual de *p-values* pertencentes à primeira classe, ou seja, o percentual de rejeição de H_0.

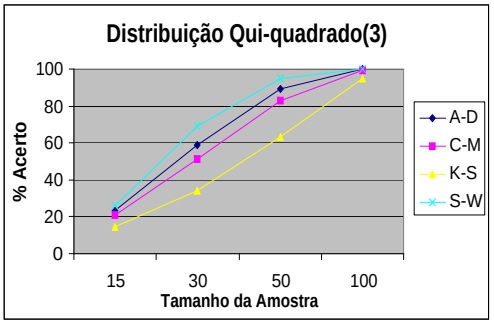
3. RESULTADOS E DISCUSSÃO

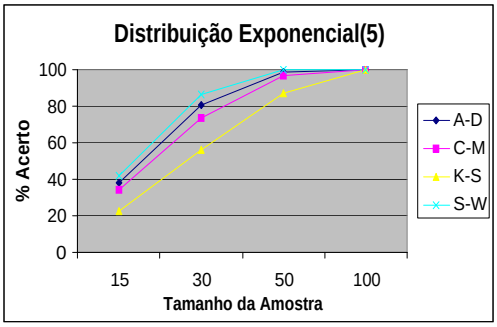
Os resultados foram condensados de forma gráfica, com gráficos de linha relacionando o tamanho de amostra e o percentual de acerto para cada teste. A seguir estes são apresentados:











Através dos gráficos de comparação dos resultados, percebe-se que, quando o conjunto de dados provém de uma distribuição Normal e independente do tamanho da amostra, todos os testes têm percentual de acerto maior que 98,5%. De uma maneira geral, os quatro critérios se equivalem em eficiência, com exceção do teste Kolmogorov-Smirnov que claramente mostrou-se inferior aos demais.

Observa-se que os quatro testes tiveram baixo desempenho quando a distribuição dos dados era aproximadamente simétrica, mas não-normal. Mesmo quando n=100 todos tiveram percentual de acerto abaixo de 40%.

Quanto mais assimétrica a distribuição, melhor era o desempenho dos testes, já que o percentual de acerto era maior mesmo com amostras menores.

Para todas as distribuições não-normais geradas, evidenciou-se uma superioridade do teste de Shapiro-Wilk em relação aos demais, e inferioridade do teste de Kolmogorov-Smirnov.

4. CONCLUSÕES

Como para dados normais os quatro critérios mostraram-se equivalentes, com exceção do critério de Kolmogorov-Smirnov, e para dados não-normais o teste de Shapiro-Wilk mostrou-se sempre superior, conclui-se que este é aparentemente o melhor teste de aderência à normalidade.

Quando a distribuição dos dados for aproximadamente simétrica, sugere-se utilizar um nível de significância para o teste não tão rigoroso como $\alpha = 0{,}01$, pois o desempenho dos testes será melhor.

Mais estudos poderiam ser feitos, considerando-se outras distribuições, outros parâmetros destas distribuições e outros tamanhos de amostra.

REFERÊNCIAS BIBLIOGRÁFICAS

StatSoft, Inc. (2004). *Electronic Statistics Textbook*. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/stathome.html.

SAS Institute Inc., Cary, NC, USA. *Help do software SAS System for Windows V8*.

Kirkwood, Betty R (1989) *Essentials of Medical Statistics*. Blackwell Scientific Publications. Londres.

Callegari-Jacques, Sídia M. (2003) *Bioestatística: princípios e aplicações*. Artmed. Porto Alegre.