VERIFICAÇÃO DA INTEGRIDADE DOS DADOS RESULTANTES DE ANÁLISES DE DADOS OBTIDAS POR MEIO DO MICROSOFT EXCEL

Matheus Percário Bruder, Gustavo Tarrafil, Cícero Roberto Nobre de Oliveira Alcântra, Daniel Mitio Mori, Lucas Henrique Carneiro Leao Couto, Kevin Barrios, Kenzo Ishibashi

RESUMO

A pesquisa foi executada a fim de detectar a existência de divergência entre os resultados obtidos através da análise de dados feita pelo Microsoft Excel, quando comparados aos resultados das mesmas análises estatísticas realizadas pelos *Microcal Origin* 8.0 e R. As covariáveis utilizadas para encontrar a integridade dos dados do software, foram: análise de variância, análise de regressão linear, média, moda, mediana, desvio padrão e coeficiente de variação. Houveram diferenças numéricas entre os resultados, porém não significativas, portanto, o Excel pode ser utilizado para o uso empresarial, doméstico e afins.

INTRODUÇAO

Nos últimos anos, diversos softwares pessoais vêm sendo utilizados e desenvolvidos, para a análise de dados de levantamentos complexos [1]. Novos softwares estão sendo estudados para elaboração de instrumentos, suportes e apoio à alunos e professores do ensino básico e secundário [2].

Antigamente, a análise dos dados era realizada somente por profissionais específicos da área de estatística. No entanto, atualmente, devido a popularidade dos métodos e softwares estatísticos, bem como, o desenvolvimento de interfaces mais “amigáveis”, muitas pessoas estão optando por realizar suas próprias análises [1]. A comunidade acredita que a utilização de soluções não especificas, tais como análises de dados feitas no Microsoft Excel ou planilhas eletrônicas de um modo geral, fazem com que a análise perca parte de sua credibilidade [3]. Uma variedade de softwares tem sido criada para auxiliar na análise de dados qualitativos, tais programas também são alvos de intensos debates [3].

O Excel é reconhecido como ferramenta para análise de dados e como um excelente suporte didático para o ensino. Uma vantagem do Excel é que está instalado na maioria dos computadores pessoais, e possui ferramentas que inclui, entre outros, estatísticas descritivas, ANOVA e regressão linear. Entretanto, não tem muita flexibilidade e a execução dos comandos não é simples, seja usando as ferramentas ou os programas integrados [4].

Pacotes estatísticos não ganharam a mesma popularidade. Um dos principais motivos era a interface pouco “amigável”. Até pouco tempo, todas as análises eram executadas por comandos, o que não contribuía para com aqueles menos familiarizados com programas computacionais [4].

O objetivo foi verificar a integridade dos resultados provenientes de análises de dados realizadas pelo Microsoft Excel, o qual serviu como base para comparação dos resultados com outros dois softwares específicos, o *Microcal Origin* 8.0 e o R [1].

MATERIAL E MÉTODO

Os programas, Microsoft Excel, *Microcal Origin* 8.0 e R, foram utilizados no sistema operacional Microsoft Windows. Excel: é um editor de planilhas (Folhas de Cálculo) produzido pela Microsoft existente a mais de trinta anos no mercado [2][4]. *Origin*: é um programa de computador próprio para gráficos científicos interativos e análise de dados, produzido pela *OriginLab* Corporation [5]. R: é um resultado do trabalho em conjunto de colaboradores de diversas partes do mundo. Foi escrito por Robert Gentleman e Ross Ihaka do Departamento de Estatística da Universidade de *Auckland* [4].

Para realizar a pesquisa, foram utilizados dados fornecidos pelo Laboratório de Biofísica Aplicada da UNESP de Botucatu. Os dados são relativos às medições de densidade de madeira clonal de *Eucalyptus ssp*, a fim de determinar a época certa de colheita para indústrias. Para promover maior variabilidade de resultados foram criados banco de dados de 4000 amostras, retiradas de um banco de dados de 100.000 amostras, em que foram utilizadas para a retirada aleatória de três banco de dados contendo 2000, 500 e 200 amostras. a partir dos três bancos de dados criados foram criadas subdivisões.

Na Figura 1, cada banco de dados foi subdividido em vinte séries, as quais possuem amostras aleatórias do banco de dados padrão de 4000 amostras: Sendo (Banco de dados: 2000 amostras: subdivisões: Série 1, Série 2, ..., Série 20); (Banco de dados: 500 amostras: subdivisões: Série 1, Série 2, ..., Série 20); (Banco de dados: 200 amostras: subdivisões: Série 1, Série 2, ..., Série 20).

	Banco de Dados	**Séries aleatórias de dados a serem análisadas**										
Série	**Densidade de Madeira**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**...S20**
1	**647**	571	709	395	670	556	523	572	458	485	512	
2	**621**	447	583	595	426	500	541	607	630	512	526	
3	**677**	539	617	474	426	520	630	643	450	633	649	
4	**586**	483	370	426	478	521	472	486	454	479	453	
5	**694**	427	485	588	443	467	585	631	509	546	613	
6	**541**	631	521	661	626	583	610	589	512	697	479	
7	**513**	661	593	512	644	570	519	455	633	639	557	
8	**623**	577	495	482	574	516	538	449	602	407	624	
9	**485**	571	472	506	470	476	628	592	444	492	376	
10	**495**	537	756	545	504	435	503	540	603	659	519	
...4000												

Figura 1 – Banco de dados com 4000 amostras de densidade de madeira clonal de *Eucalyptus ssp*. Dado: banco de dados utilizado para obter bancos de dados de 2000, 500 e 200 amostras de forma aleatória.

Para os três softwares de análise, – Microsoft Excel, *Microcal Origin* 8.0 e programa R – foram utilizados para analisar os mesmos cálculos, das covariáveis: média, moda, mediana, desvio padrão, coeficiente de variação, regressão linear e análise de variância. Porém, com uso de séries distintas, porém os cálculos de regressão linear e análise de variância necessitam de mais de um parâmetro de comparação.

Os cálculos estatísticos mais simples, como, a média, moda, mediana, desvio padrão e o coeficiente de variação, foram executados para todas as vinte séries dos três bancos de dados com, respectivamente, 2000, 500 e 200 amostras.

De maneira análoga foi realizado o cálculo de regressão linear, o qual necessita de dois parâmetros para comparação. Dessa forma, foram elaboradas dez regressões lineares, entre a série 1 e a série 2, entre a série 3 e a série 4 e assim por diante, até a série 20. Sendo: regressões lineares (série X = SX: como, *S1, S2; S3, S4; ...; S19, S20).* após efetuar a regressão linear para obter um padrão no processo de comparação, os valores coletados foram R^2, R^2 ajustado e valor F.

Os cálculos de análise de variância seguiram o mesmo procedimento das regressões lineares. A única diferença foi que para a análise de variância, utilizou-se três parâmetros diferentes de formas aleatórias. Sendo: análises de variância (série X = SX: como, *S7, S16, S2*; *S15, S20, S3*; *S5, S4, S16).* Visando uma padronização dos resultados, para as dez análises de variância os resultados comparados, foram: Valor de P e F.

A fim de conseguir uma maior exatidão, os resultados obtidos pelos três softwares foram padronizados com cinco casas decimais.

RESULTADOS

As análises realizadas pelos três softwares, Microsoft Excel, *Microcal Origin. 8.0* e programa R, apresentaram diferenças numéricas, porém, não significativas.

A fim de proporcionar maior exatidão, os resultados foram padronizados com cinco casas decimais, pois, outros trabalhos foram realizados com softwares distintos com três casas decimais [1].

Para cada banco de dados com, respectivamente, 2000, 500 e 200 amostras aleatórias de densidades de madeira, as covariáveis analisadas, apresentaram resultados idênticos pelos softwares. O único parâmetro não estimado, foi a moda no software "R", devido a indisponibilidade da função pelo programa.

Resultado do banco de dados com 2000 amostras

EXCEL E ORIGIN 8.0			**R**		
	Média	**Desvio Padrão**		**Média**	**Desvio Padrão**
Média	542,54405	1,51487	**Média**	542,54400	1,51480
Mediana	534,46906	1,96444	**Mediana**	534,46900	1,96440
Desvio Padrão	83,43070	1,20596	**Desvio Padrão**	83,43070	1,20590
Moda	508,10988	66,31591	**Moda***		
Coeficiente de Variação	15,37750	0,20425	**Coeficiente de Variação**	15,37750	0,20420

Resultado do banco de dados com 500 amostras

EXCEL E ORIGIN 8.0			**R**		
Média	541,24076	3,45928	**Média**	541,24070	3,45920
Mediana	531,98210	4,34279	**Mediana**	531,98200	4,34270
Desvio Padrão	84,23118	2,54865	**Desvio Padrão**	84,23110	2,54860
Moda	528,21852	93,04817	**Moda***		
Coeficiente de Variação	15,56274	0,46635	**Coeficiente de Variação**	15,56270	0,46630

Resultado do banco de dados com 200 amostras

EXCEL E ORIGIN 8.0			**R**		
Média	540,57858	6,19726	**Média**	540,57850	6,19720
Mediana	534,72690	6,85665	**Mediana**	534,72680	6,85660
Desvio Padrão	82,97385	4,21631	**Desvio Padrão**	82,97380	4,21630
Moda	549,77673	78,13726	**Moda***		
Coeficiente de Variação	15,34764	0,72825	**Coeficiente de Variação**	15,34760	0,72820

Figura 2 – Resultados da media, moda, mediana, desvio padrão e coeficiente de variação das amostras referentes aos bancos de dados com, respectivamente, 2000, 500 e 200 amostras aleatórias de densidade de madeira clonal de *Eucalyptus ssp*. Dados: O software Excel e Origin estão representados na mesma coluna devido a igualdade de seus resultados.

A regressão linear não apresentou diferença significativa. Os valores de F e R^2 obtidos tanto pelo Microsoft Excel, quanto pelo *Microcal Origin* 8.0 foram idênticos. No entanto, o valor numérico de R^2 ajustado se diferenciou entre os resultados. Tal diferença ocorreu após a segunda casa decimal nos três bancos de dados, conforme Figura 3. Devido a função indisponível que ajusta a regressão linear passando pelo ponto de origem não foi possível calcular a regressão linear pelo programa R.

A análise de variância também não apresentou diferença significativa. Os valores de P e F foram idênticos para os softwares Excel e Origin. Contudo, os mesmos valores de P e F foram diferentes quando calculados por meio do software R, conforme Figura 4.

Resultado do banco de dados com 2000 amostras

EXCEL			ORIGIN 8.0			R		
Regressão Linear	**Média**	**Desvio Padrão**	**Regressão Linear**	**Média**	**Desvio Padrão**	**Regressão**	**Média**	**Desvio Padrão**
R2	0,95389	0,00137	**R2**	0,95389	0,00137	**R2**		
R2 Ajustado*	0,95339	0,00137	**R2 Ajustado***	0,95387	0,00137	**R2 Ajustado**		
Valor F	41389,97	1276,442	**Valor F**	41389,97	1276,442	**Valor F**		

Resultado do banco de dados com 500 amostras

EXCEL			ORIGIN 8.0			R
R2	0,95362	0,00176	**R2**	0,95362	0,00176	**R2**
R2 Ajustado*	0,95162	0,00176	**R2 Ajustado***	0,95353	0,00176	**R2 Ajustado**
Valor F	10275,33	426,1965	**Valor F**	10275,33	426,1965	**Valor F**

Resultado do banco de dados com 200 amostras

EXCEL			ORIGIN 8.0			R
R2	0,95492	0,00309	**R2**	0,95492	0,00309	**R2**
R2 Ajustado*	0,94989	0,00309	**R2 Ajustado***	0,95469	0,00309	**R2 Ajustado**
Valor F	4233,783	305,6460	**Valor F**	4233,78	305,6460	**Valor F**

Figura 3 – Resultados da regressão linear das amostras referentes aos bancos de dados com, respectivamente, 2000, 500 e 200 amostras aleatórias de densidade de madeira clonal de *Eucalyptus ssp*.

Resultado do banco de dados com 2000 amostras

EXCEL E ORIGIN 8.0			R		
Análise de Variancia	**Média**	**Desvio Padrão**	**Análise de Variancia**	**Média**	**Desvio Padrão**
Valor P	0,54776	0,19845	**Valor P**	0,40350	0,17872
Valor F	0,66392	0,38071	**Valor F**	0,88956	0,71404

Resultado do banco de dados com 500 amostras

EXCEL E ORIGIN 8.0			R		
Valor P	0,49613	0,19261	**Valor P**	0,50156	0,35607
Valor F	0,78409	0,45565	**Valor F**	1,08560	1,26578

Resultado do banco de dados com 200 amostras

EXCEL E ORIGIN 8.0			R		
Valor P	0,47971	0,29272	**Valor P**	0,56259	0,28667
Valor F	0,97011	0,77923	**Valor F**	0,68640	0,96965

Figura 4 – Resultados da análise de variância das amostras referentes aos bancos de dados com, respectivamente, 2000, 500 e 200 amostras aleatórias de densidade de madeira clonal de *Eucalyptus ssp*.

DISCUSSÃO E CONCLUSÃO

Essa pesquisa foi fundamentada na verificação da integridade dos dados resultantes das análises estatísticas obtidas por meio do Microsoft Excel. Contudo, houveram algumas limitações durante o processamento. A primeira limitação foi verificada simultaneamente nos softwares Origin 8.0 e R, em que ambos “arredondaram” os dados para apenas cinco casas decimais quando foram manipulados para uma planilha eletrônica. A segunda limitação do estudo ocorreu durante a análise de dados do R, pois não possui a função necessária para calcular a moda, tal como, impossibilidade de efetuar a regressão linear.

Trabalhos preocupados com integridade dos softwares [1], apresentaram pesquisas semelhantes, pois a padronização dos resultados é um fator muito importante. A maioria dos resultados foram padronizados com três casas decimais [1]. A análise dos resultados com cinco casas decimais proporcionou a mesma exatidão.

De acordo com a Figura 2, os cálculos referentes a média, moda, mediana, desvio padrão e o coeficiente de variação apresentaram resultados idênticos em todas as análises. Lembrando da limitação do R, o qual forneceu apenas quatro casas decimais. Além disso, a maior e menor média encontradas, foram no banco de dados de 200 amostras, de respectivamente, 551,83596 e 533,23109, que refletiu no maior desvio padrão, os resultados do desvio padrão foram inversamente proporcionais ao tamanho dos bancos de dados.

Os resultados obtidos pela regressão linear podem ser utilizados de forma geral, no uso doméstico, pessoal, empresarial e etc. [1]. Pois apesar da diferença numérica entre os valores de R^2 ajustado, devido aos diferentes cálculos internos de cada software, não existe diferença estatística. Portanto, todos os softwares são recomendados para realizar tais cálculos, desde que, o usuário saiba previamente quais cálculos serão efetuados e, também, como utilizar a interface gráfica do programa de análise de dados escolhido [4]. Com exceção do software R, o qual foi uma limitante e que impossibilitou o cálculo de regressão linear, devido a função indisponível que ajusta a regressão linear passando pelo ponto de origem.

A análise de variância de forma análoga a regressão linear, também apresentou diferenças numéricas nos resultados dos softwares Excel e Origin, quando comparados aos resultados do software R. Isso ocorreu devido aos diferentes cálculos internos realizados por cada um dos programas estatísticos.

Embora o software mais interativo com o usuário tenha sido o *Microcal Origin* 8.0, devido sua interface "amigável", o Excel é o mais utilizado por empresas, para uso doméstico e por pesquisadores, pois trata-se de um software gratuito e de fácil acesso [4][6].

Dessa maneira, conclui-se que houve diferenças numéricas, porém não significativas entre as análises e, principalmente, o Microsoft Excel que pode ser utilizado para realizar análises de dados sem o comprometimento da integridade dos resultados. Portanto, cabe ao usuário escolher o software que lhe apresentar maior familiaridade ou facilidade para aprender [4].

REFERÊNCIAS

[1] SOUSA, M. H; SILVA, N. N. Comparação de softwares para análise de dados de levantamentos complexos. Rev. De Saúde Pública., v.34, n.6, p.646-53, dezembro 2000.

[2] ALVES, H.; CUNHA, L. M. XII – Software Estatístico. Disponível em: <http://homepage.ufp.pt/cmanso/ALEA/Dossier12.pdf>. Acesso em: 11 mai. 2018

[3] LAGE, M. C.; GODOY, A. S. O uso do computador na análise de dados qualitativos: questões emergentes. Rev. Adm. Mackenzie, v.09, n.4, junho 2008.

[4] PAES, A. T. et al. Que programa estatístico utilizar? Rev. Einsten, p.125-7, setembro 2011.

[5] WASS, J. A. OriginPro 8 — Not Just for Graphics Anymore. 2008. Disponível em: <https://www.scientificcomputing.com/article/2008/02/originpro-8-%E2%80%94-not-just-graphics-anymore>. Acesso em: 27 mai. 2018.

[6] NETO, A. A. H.; STEIN, C. E. (2003). Uma abordagem dos testes não-paramétricos com utilização do Excel. Disponível em: <http://home.furb.br/efrain/matematica/minicurso/artigo_11_09_2003.doc>. Acesso em: 15 mai. 2018

AGRADECIMENTOS E OUTROS

Ao Prof. Dr. André F. de Angelis responsável pela disciplina Metodologia do Trabalho Científico pelas aulas expositivas, bem como, pelas orientações ao longo do semestre. Como também, ao Laboratório de Biofísica Aplicada da UNESP de Botucatu e a Edson Marcelo Bruder, por fornecerem o banco de dados relativos às medições de densidade de madeira clonal de *Eucalyptus ssp.*