

## TÉCNICAS DE ANÁLISES EXPLORATÓRIAS DE DADOS

### AUTORES

ALFREDO RIBEIRO DE FREITAS<sup>1</sup>

<sup>1</sup> Pesquisador da Embrapa Pecuária Sudeste, C.P. 339, 13560-970. São Carlos, SP. ribeiro@cppse.embrapa.br

### RESUMO

O objetivo foi utilizar em dados de desenvolvimento ponderal de bovinos Nelore, oriundos da Associação Brasileira de Criadores de Zebu - ABCZ, as técnicas de análises exploratórias: a) Momentos e Quantis: média aritmética, desvio padrão, erro-padrão da média, assimetria, curtose, etc; b) Testes de Normalidade: Kolmogorov-Smirnov, Cramer-von Mises e Anderson-Darling; c) Testes de Localização: Teste t de Student, Teste do Sinal e Teste das Ordens Assinaladas; d) Medidas Robustas de Escalas: Amplitude Interquartilica, Diferença Média de Gini, Desvio Absoluto da Mediana, Sn e Qn e e) Médias Winsorizadas e Trimmed. Todas as estatísticas obtidas, associadas aos dados de pesagens, indicam que estes não se ajustam à distribuição normal.

### PALAVRAS-CHAVE

bovinos de corte, desenvolvimento ponderal, Medidas Robustas de Escalas, Médias Winsorizadas e Trimmed, Momentos e Quantis, Testes de Normalidade

### TITLE

EXPLORATORY DATA ANALYSIS TECHNIQUES

### ABSTRACT

The objective of this study was to analyze Nelore performance data, obtained at the National Archive of Brazilian Zebu Breeders Association (ABCZ), by using the following exploratory data analysis techniques: a) Moments and Quantiles: Mean, standard deviation, skewness, kurtosis, standard error of the mean, etc; b), Tests for Normality: Kolmogorov-Smirnov, Cramer-von Mises statistic and Anderson-Darling; c) Tests for Location: Student's t statistic, Sign statistic and Signed Rank test; d) Robust Measures of Scale: Interquartile range, Gini's mean difference, Median Absolute Deviation, Sn and Qn and e) Trimmed and Winsorized means. All the statistics associated to the weight data, showed that these do not fit to a normal distribution.

### KEYWORDS

beef cattle, Location Tests, Moments and Quantiles, performance data, Robust Measures of Scale, Tests for Normality

### INTRODUÇÃO

Com o crescimento dos recursos computacionais em todas as áreas (PEARSON, 2001), o volume de dados tem crescido enormemente. A quantidade de informação no mundo duplica a cada 20 meses, e o tamanho e a quantidade dos bancos de dados crescem com velocidade ainda maior. Entretanto, este crescimento tem mantido relação inversa com a qualidade. Mesmo as instituições de pesquisas, que tradicionalmente tem maior rigor na coleta e análise de dados, vem enfrentando dificuldades crescentes, pois com as facilidades computacionais, grande parte dos dados armazenados e analisados, não são coletados de experimentos planejados, o que dificulta testar hipóteses de interesse. A maioria destes dados, principalmente devido à problemas de coletas, apresenta outliers, correlações absurdas, afastamento da distribuição normal, heterogeneidade de variâncias, grau

acentuado de assimetria e curtose, entre outras. Estas anomalias podem facilmente omitir resultados importantes quando se utilizam procedimentos tradicionais de análises e/ou quando realizadas por profissionais inexperientes. Como exemplos de dados que apresentam tais anomalias, podem citar as pesagens de bovinos zebuínos (FREITAS et al., 2000). Entretanto, como a qualidade dos dados colhidos no campo e a sua posterior análise são fundamentais para o sucesso de um programa de melhoramento genético animal, para se proceder a um refinamento metodológico e obter estimativas de parâmetros genéticos confiáveis, é necessário utilizar adequadamente as técnicas de análises exploratórias de dados. O objetivo deste trabalho é utilizar técnicas de análises exploratórias para avaliar as distribuições associadas à dados de nove pesagens de bovinos zebuínos Nelore, realizadas em intervalos trimestrais, até os dois anos de idade

## MATERIAL E MÉTODOS

Foram analisados dados de nove pesagens: PN e P1 a P8, de bovinos da raça Nelore Nelore, oriundos da Associação Brasileira de Criadores de Zebu - ABCZ; as pesagens foram avaliadas em intervalos trimestrais, do nascimento aos dois anos de idade. Serão utilizadas cinco técnicas de análises exploratórias, cujas teorias são descritas no módulo INSIGHT do SAS (SAS, 2000) a) "Momentos e Quantis": os momentos incluem a média aritmética, variância, desvio padrão, erro padrão da média, coeficiente de variação e medida de assimetria e de curtose; os quantis incluem o valor mínimo e máximo da amostra; a amplitude, a mediana, a moda, os quartis, decis e percentis; b) "Testes de normalidade": Kolmogorov-Smirnov (K-S), Cramer-von Mises e Anderson-Darling; usam estatísticas baseadas em uma função de distribuição empírica. O teste de K-S avalia a discrepância entre uma distribuição empírica e a distribuição Normal considerada como referência. Em todos os testes, a hipótese nula a ser testada é que os dados em estudo corresponde a uma amostra aleatória proveniente de uma distribuição normal; c) "Testes de Locação": Testa a hipótese de que a média/mediana de um conjunto de dados difere ou não de um parâmetro teórico (média/mediana). Se os dados são provenientes de uma população aproximadamente normal usa-se o teste t de Student; se os dados não tem distribuição definida, utilizam-se os testes não-paramétricos: Teste do Sinal e das Ordens Assinaladas. A hipótese nula a ser testada é que a média/mediana obtida da amostra não difere do mesmo parâmetro obtido de uma distribuição teórica (geralmente a normal), usada como referência; d) Medidas Robustas de Escalas: O desvio-padrão amostral (DP) é comumente usado para obter inferências de uma população. Entretanto, como ele é sensível à "outliers", é importante o uso de estimadores robustos para esta estatística, principalmente para grandes arquivos de dados coletados em condições de campo, como no presente estudo. São utilizadas cinco estatísticas para se obter um estimador robusto do DP: Intervalo Interquartilico; Diferença Média de Gini (G); Desvio Absoluto da Mediana-DAM;  $S_n$  e  $Q_n$ ; e) "Médias Winsorizadas e Trimmed": Quando "outliers" ou valores extremos estão presentes nos dados de uma amostra, as médias obtidas com a eliminação de uma percentagem de valores extremos, denominadas de médias "trimmed" e "Winsorized", são estimativas robustas da média da população e são insensíveis à "outliers".

## RESULTADOS E DISCUSSÃO

Tabela 1 resume as técnicas de análises exploratórias para cada pesagem. Os coeficientes de assimetria foram positivos e crescentes, indicando que a cauda da curva de frequência é viesada à direita; esperar-se-ia a relação Média > Mediana > Moda. Como estes dados possuem vícios de pesagens (FREITAS et al., 2000), com distribuição plurimodal, este fato pode ser a causa dessa relação ter ocorrido apenas para P3. Os coeficientes de curtose foram positivos de P1 a P8, mostrando que o pico de frequência da distribuição dos pesos é mais pontiaguda que a da normal. A hipótese de nulidade ( $H_0$ ) foi rejeitada em todos os testes de normalidade, indicando que os dados de pesagens, não são uma amostra aleatória proveniente de uma distribuição normal. A assimetria, a curtose e a não-normalidade dos dados, influenciam as inferências obtidas, a estimação dos efeitos fixos e a heterogeneidade de variância do erro (COCHRAN e COX, 1978; BROWNIE et al., 1990, AZZALINI e DALLA VALLE, 1996). Quanto aos testes de locação, todos rejeitaram  $H_0$  ao nível de

$P < 0,0001$ , para todas as pesagens, sugerindo que os dados diferem significativamente de uma distribuição normal. A rejeição de  $H_0$  por estes três testes, comprova também que os dados são assimétricos, conforme já discutido. A amplitude dos resultados das medidas robustas de escalas, está apresentada na Tabela 1; todos os estimadores são inferiores ao desvio padrão amostral (DP). O maior estimador robusto, para todas as pesagens, foi o obtido pela Diferença Média de Gini. Estes resultados mostram que o uso tradicional do DP para eliminar dados extremos de uma amostra, ou seja, média  $\pm$  3DP, deva ser substituído pelo seu estimador robusto; raciocínio análogo vale para o uso do DP em rotinas de simulação de dados com distribuição normal. SINGHA e NOCERINOB (2002), trabalhando com dados de contaminantes ambientais censurados, obtiveram estimativas confiáveis de parâmetros populacionais de média e desvio-padrão. Segundos eles, após a eliminação de outliers, é possível obter estimativas robustas em concordância com as correspondentes estimativas clássicas. Em todas as pesagens, as médias aritmética estimadas de Médias Winsorizadas e de Médias Trimmed, somente são boas estimativas da média amostral e da mediana, quando são eliminados 10% dos valores amostrais (5% de cada extremo). Quando suposições de normalidade e de homogeneidade de variância são violadas, procedimentos de comparação múltipla baseados em médias e variâncias obtidas por estes dois estimadores, controlam o erro do tipo I (KESELMAN et al., 1998). Os resultados deste trabalho mostram, por meio dos estimadores obtidos de técnicas de análises exploratórias, que a qualidade dos dados de pesagens de bovinos Nelore, considerando-se as propriedades da distribuição normal, decresce com a idade do animal. Isto deve servir de alerta aos pesquisadores da área de melhoramento animal, uma vez que a tendência atual é estimar parâmetros genéticos por regressão aleatória, utilizando-se todos os dados peso-idade do animal.

## CONCLUSÕES

Dados de pesagens de bovinos da raça Nelore Nelore, avaliados em intervalos trimestrais, do nascimento aos dois anos de idade, não se ajustam à distribuição normal.

A precisão dos estimadores obtidos de técnicas de análises exploratórias e que avaliam as propriedades da distribuição normal, decresce com a idade do animal.

## REFERÊNCIAS BIBLIOGRÁFICAS

1. AZZALINI, A., DALLA VALLE, A. D. 1996. The multivariate skew-normal distribution. "Biometrika", London, v.83, n.4, p. 715-726.
2. BROWNIE, C., BOOS, D.D., OLIVER, J.H. 1990. Modifying the t and ANOVA F tests when treatment is expected to increase variability relative to controls. "Biometrics", v.46, n.1, p.259-266.
3. COCHRAN, W.G., COX, D.F. 1978. "Desenho experimentales". Mexico: Trillas, 661p.
4. FLYNN, A A; GREEN, A J; BOXER, G M; et al. A novel technique, using radioluminography, for the measurement of uniformity of radiolabelled antibody distribution in a colorectal cancer xenograft model. International Journal of Radiation Oncology, Biology, Physics, v.43, n.1. January 1, p.83-189, 1999.
5. FREITAS, A, R. de., SILVA, L.O.C., MACHADO, et al. A qualidade da pesagem de bovinos das raças zebuínas: IV CONGRESSO BRASILEIRO DAS RAÇAS ZEBUÍNAS, Uberaba, MG, "Anais..." Uberaba, 2000. p. 322.
6. HARTWIG, F.; DEARING, B.E. Exploratory data analysis. Series: Quantitative Applications in the Social Science 16 1979. editor: John L. Sullivan. Sage University Paper. 83p.
7. HURLEY, W. J.; LIOR, D. U. Combining expert judgment: On the performance of trimmed mean vote aggregation procedures in the presence of strategic voting. "European Journal of Operational Research", v.140, n.1, 1 July 2002, p.142-147.

8. KESELMAN, H. J.; LIX, L. M.; KOWALCHUK, R.K. Multiple Comparison Procedures for Trimmed Means. "Psychological Methods". v.3,n.1, March 1998, p.123-141.
9. KRAMER, W B; SAADE, G R; BELFORT, M; et al. A randomized double-blind study comparing the fetal effects of sulindac to terbutaline during the management of preterm labor. "American Journal of Obstetrics and Gynecology". v.180, n.2, Part 1 February 1999,p.396-401. ISSN: 0002-9378.
10. PEARSON, R.K. Exploring process data. Journal of Process Control. v.11, n.2, April 2001, p.179-194.
11. ROGGI, C; SABBIONI, E; MINOIA, C; et al. 1995. Trace element reference values in tissues from inhabitants of the European Union. IX. Harmonization of statistical treatment: blood cadmium in Italian subjects. The Science of the Total Environment ISSN: 0048-9697 . v.166, April 21, p.235-243.
12. SAS Institute 2000. "SAS/INSIGHT" User's Guide. versão 8.2, versão para Windows Cary, NC, USA.
13. SINGHA,A.; NOCERINO, J. Robust estimation of mean and variance using environmental data sets with below detection limit observations. Chemometrics and Intelligent Laboratory Systems.v.60, n.1-2, 28 January 2002, p.69-86

Tabela 1 – Técnicas de análises exploratórias aplicadas à nove pesagens dados de bovinos Nelore: PN (nascimento) e P<sub>1</sub> a P<sub>8</sub>.

Estatística	PN	P <sub>1</sub>	P <sub>2</sub>	P <sub>3</sub>	P <sub>4</sub>	P <sub>5</sub>	P <sub>6</sub>	P <sub>7</sub>	P <sub>8</sub>
	Momentos e Quantis								
ASSIMETRIA	0,99	0,69	0,46	0,45	0,82	0,90	0,92	1,00	1,07
Curtose	4,04	0,70	0,29	0,76	1,40	1,32	1,38	1,63	1,92
locação <sup>1</sup>	1,2=3	2,3,1	2,3,1	1,2,3	2,3,1	2,3,1	3,1,2	3,1,2	2,3,1
	Testes de Normalidade <sup>2</sup>								
K-S	< 0,01	< 0,01	< 0,01	< 0,01	< 0,01	< 0,01	< 0,01	< 0,01	< 0,01
C-V-M	< 0,001	< 0,001	< 0,001	< 0,001	< 0,001	< 0,001	< 0,001	< 0,001	< 0,001
A-D	< 0,001	< 0,001	< 0,001	< 0,001	< 0,001	< 0,001	< 0,001	< 0,001	< 0,001
	Testes de Locação <sup>3</sup>								
Teste t									
SINAL									
Ordens									
	Medidas Robustas de Escalas <sup>4</sup>								
D.P	3,05	26,70	33,08	38,95	49,53	61,96	71,94	79,83	89,59
Mínimo	1,48	26,23	32,20	35,58	42,93	54,85	62,26	66,71	75,13
Máximo	2,75	26,68	33,36	38,37	48,00	59,92	69,43	76,52	85,28
	Médias Trimmed <sup>5</sup>								
(1/2)5%									
	Médias Winsorizadas <sup>5</sup>								
(1/2)5%									

<sup>1</sup> Locação = indica a ordem de ocorrência (média = 1; moda = 2; mediana = 3)<sup>2</sup> K-S = Kolmogorov-Smirnov; C-V-M = Cramer-von Mises; A-D = Anderson-Darling<sup>3</sup> Testes de Locação: Os teste t de Student, Sinal e Ordens Assinaladas rejeitaram a hipótese H<sub>0</sub> ao nível de < 0,0001, para todas as pesagens<sup>4</sup> DP = Desvio padrão amostral; mínimo e máximo, indicam a amplitude dos valores de desvio padrão robusto obtidos por: Amplitude Interquartílica, Diferença Média de Gini(Gini), Desvio Absoluto da Mediana, S<sub>n</sub> e Q<sub>n</sub>; o maior valor foi obtido de Gini.<sup>5</sup> (1/2)5% foram eliminados 5% dos dados de cada extremos da distribuição de freqüência.