Anais do Conic-Semesp. Volume 1, 2013 - Faculdade Anhanguera de Campinas - Unidade 3. ISSN 2357-8904

TÍTULO: TESTES ESTATÍSTICOS PARA DIAGNÓSTICO DA NORMALIDADE DE UMA DISTRIBUIÇÃO DE DADOS

CATEGORIA: CONCLUÍDO

ÁREA: CIÊNCIAS SOCIAIS APLICADAS

SUBÁREA: CIÊNCIAS CONTÁBEIS

INSTITUIÇÃO: FACULDADE ANHANGUERA DE TAUBATÉ

AUTOR(ES): ANA ISA RIBEIRO FERREIRA, AUYRA MICHELE MARIANO DE SOUZA, DANIELA TATIANE DA SILVA

ORIENTADOR(ES): EVERALDO DE BARROS

Realização: SEMESP Apoio: Anhanguera

Testes Estatísticos para Diagnóstico da Normalidade de uma Distribuição de Dados

Resumo

Dados estatísticos representam observações de um dado fenômeno. A partir da coleta de amostras representativas de uma determinada população em estudo, os subconjuntos de dados estatísticos são inicialmente classificados e caracterizados. Posteriormente, por meio da estatística inferencial, decisões são tomadas sobre uma população com base nos subconjuntos de dados extraídos da população. Análises estatísticas da população pesquisada são rotineiramente realizadas partindo do pressuposto que os dados coletados seguem uma distribuição normal de probabilidade. Dentre as distribuições de probabilidade contínuas, a distribuição normal apresenta diversas vantagens, pois, é completamente definida conhecendo-se a média e o desvio-padrão da variável aleatória. Ainda, muitos fenômenos observados na natureza são caracterizados por uma variável aleatória contínua e descrita por uma distribuição normal de probabilidade. Todavia, para uma correta análise estatística, torna-se imperativo que um diagnóstico da variável aleatória seja realizado para avaliar a distribuição dos dados, evitando desta forma uma suposição incorreta de normalidade dos dados coletados. Neste artigo, testes estatísticos para diagnosticar a normalidade em um conjunto de observações, são discutidos. Uma pesquisa com abordagem qualitativa, com objetivo exploratório, descritivo, utilizando o procedimento técnico classificado como estudo de caso, é proposta, por meio da aplicação de testes de normalidade em um conjunto de observações de um fenômeno físico aleatório.

Introdução

A maior parte dos fenômenos observados na natureza pode ser caracterizada por uma variável aleatória contínua e descrita por uma distribuição normal de probabilidade. Dentre as principais características de uma distribuição normal, podemos incluir que a forma da distribuição é em formato de sino determinada pela média e pelo desvio padrão, simétrica em torno da média; o ponto máximo da distribuição é determinado pela média; a média, a mediana e a moda têm o

mesmo valor; a área total sob a curva normal é igual à unidade e 99,7% da área sob a curva encontra-se entre o intervalo de mais ou menos três desvios-padrão.

Devido às vantagens de a distribuição normal ser completamente definida conhecendo-se a média e o desvio-padrão do conjunto de dados observados, esta distribuição é largamente aplicada nas análises estatísticas de variáveis aleatórias contínuas. Muitas análises estatísticas são baseadas no pressuposto de que os dados seguem uma distribuição normal. Todavia, torna-se recomendável que testes de normalidade sejam realizados para garantir que o conjunto de dados da variável aleatória investigada possa ser descrito por uma distribuição normal de probabilidade.

Isto posto, os interesses deste artigo são destinados a realizar um estudo sobre testes para diagnosticar se uma determinada variável aleatória pode ser representada por uma distribuição normal.

Objetivos

Esta pesquisa possui os seguintes objetivos:

Objetivo geral

Desenvolver estudos relacionados à estatística e a distribuição normal de probabilidade.

Objetivos específicos

Aplicar testes de normalidade em um conjunto de observações e avaliar se este conjunto pode ou não ser representado por uma distribuição normal.

Metodologia

A metodologia para o desenvolvimento da pesquisa está relacionada à classificação e à delimitação da pesquisa a ser desenvolvida.

Para os procedimentos metodológicos deste artigo é adotada uma pesquisa básica com o objetivo descritivo, com abordagem qualitativa, utilizando os procedimentos de pesquisa bibliográfica e de estudo de caso, discutidos por Godoy (1995a; 1995b), Gil (2009) e Jung (2010).

Segundo Gil (2009), o delineamento refere-se ao planejamento da pesquisa em sua dimensão mais ampla, que envolve tanto a diagramação quanto a previsão de análise e interpretação de coleta de dados. Nesta pesquisa, são atribuídas as seguintes delimitações:

- Aplicar um instrumento proposto na literatura para diagnosticar a normalidade em um conjunto de observações.
- Avaliar a normalidade dos dados por meio da aplicação do teste estatístico de Anderson-Darling (BPI CONSULTING, 2009; 2011) e obter a função de distribuição de probabilidade para os conjuntos de dados que seguem uma distribuição normal.

Desenvolvimento

Nesta seção são apresentadas a fundamentação teórica da pesquisa, a definição das V. A. contínuas e a formulação do teste de Anderson-Darling para normalidade.

Fundamentação Teórica

Estatística é definida por Larson e Farber (2007) como a ciência que se ocupa de coletar, organizar, analisar e interpretar dados a fim de tomar decisões. O conjunto de todos os resultados, respostas, medidas ou contagens que são de interesse é denominado população, e o subconjunto de uma população é definido como amostra. Um conjunto de dados que representa um fenômeno físico pode ser classificado como determinístico ou aleatório. No primeiro caso, os dados podem ser representados por uma formulação matemática, enquanto que no segundo caso os dados são descritos por suas propriedades estatísticas, pois, não apresentam previsibilidade.

Em uma revisão dos princípios da teoria da probabilidade (LARSON; FARBER, 2007; TAVARES, 2007; BENDAT; PIERSOL, 1986), temos que um experimento probabilístico é definido como uma ação por meio da qual as contagens, medidas ou respostas são obtidas, e o conjunto de todos os resultados possíveis é definido como espaço amostral. Um evento consiste em um subconjunto do espaço amostral. Temos ainda que uma variável aleatória (V. A.) representa um valor numérico associado a cada um dos resultados de um experimento probabilístico, sendo determinado por uma possibilidade. Uma V. A. é classificada como discreta, para um número finito de resultados, e, contínua, quando representada por um intervalo sobre o eixo real. Denotando $x(k)$ uma V. A. de interesse, para cada valor fixo de x, o evento aleatório $x(k) \leq x$ é definido como o conjunto de possíveis resultados de k. Nestes termos, a função de

distribuição de probabilidade P(x) é definida com a probabilidade a qual é atribuída um conjunto de pontos k que satisfaz a desigualdade x(k) ≤ x.

Uma V. A. contínua possui uma distribuição de probabilidade contínua. Dentre as várias distribuições de probabilidade contínuas, a distribuição normal, também chamada distribuição de Gauss ou Gaussiana, é a mais importante distribuição em estatística para modelar conjuntos de medidas observados na natureza. Portanto, a condução de testes para avaliação da normalidade de dados é vital para o sucesso de uma análise estatística.

Definição das V. A. Contínuas da Pesquisa

A definição da V. A. contínua ocorreu mediante a consulta de uma base de dados reportada na literatura (JOHNSON,1996). No conjunto de dados disponibilizado pelo autor para estimação da quantidade de gordura corporal, foram observadas as seguintes variáveis em uma amostra de 252 pessoas do sexo masculino: idade, peso, altura, circunferência do pescoço, circunferência do tórax, circunferência do abdome, circunferência dos quadris, circunferência da coxa, circunferência do joelho, circunferência do tornozelo, circunferência do bíceps, circunferência do antebraço e circunferência do punho. A partir desta base de dados, as amostras de cada V. A. contínua foram avaliadas para diagnóstico da normalidade dos dados.

Teste de Anderson-Darling para Normalidade

O teste de Anderson-Darling é um teste estatístico utilizado para verificar se um conjunto de dados é proveniente de uma determinada distribuição de probabilidade. Para Moraes, Ferreira e Balestrassi (2005) no teste de Anderson-Darling “considera-se normal a distribuição que apresentar *p-value* maior que 0,05, o que significaria uma probabilidade maior que 5% em cometer erro, ao rejeitar a hipótese de normalidade da distribuição em análise.”

As duas hipóteses para o teste de Anderson-Darling para uma distribuição normal são expressas por (SILVA; ARAÚJO; COSTA FILHO, 2010):

H_0: os dados seguem uma distribuição de probabilidade normal.

H_1: os dados não seguem uma distribuição de probabilidade normal.

A hipótese nula é a que os dados são normalmente distribuídos e hipótese alternativa é a que os dados não seguem uma distribuição normal. Se o nível de

significância (*p-value*) for pequeno (p ≤ 0,05), a hipótese nula é rejeitada e conclui-se que os dados analisados não seguem uma distribuição normal. A estatística teste de Anderson-Darling é expressa por (BPI CONSULTING, 2011):

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln F(X_i) + \ln(1 - F(X_{n-i+1}))\right]$$

onde *n* é tamanho da amostra, *F* é a função de distribuição acumulada para a distribuição específica e *i* é a *i*-ésima amostra quando os dados estão ordenados em ordem crescente.

Para pequenas amostras (n ≤ 200), o valor de AD deve ser ajustado para:

$$AD^* = AD\left(1 + \frac{0,75}{n} + \frac{2,25}{n^2}\right)$$

O nível de significância *p* é estimado a partir do valor de AD^*:

se $AD^* \geq 0,6$; $p = \exp(1,2937 - 5,709(AD^*) + 0,0186(AD^*)^2)$

se $0,34 < AD^* < 0,6$; $p = \exp(0,9177 - 4,279(AD^*) - 1,38(AD^*)^2)$

se $0,2 < AD^* < 0,34$; $p = 1 - \exp(-8,318 + 42,796(AD^*) - 59,938(AD^*)^2)$

se $AD^* \leq 0,2$; $p = 1 - \exp(-13,436 + 101,14(AD^*) - 223,73(AD^*)^2)$

Resultados

Na Tabela 1 são apresentados a média, o desvio-padrão, o valor da estatística teste de Anderson-Darling AD^* e o nível de significância *p*, obtidos para cada V. A. contínua da base de dados investigada. Analisando os resultados obtidos, observa-se que apenas as V. A. joelho, bíceps, antebraço e punho apresentaram um nível de significância *p-value* > 0,05, indicando, pelo critério de Anderson-Darling, que estes dados seguem uma distribuição normal. Uma vez que as demais V. As. apresentaram um baixo nível de significância (p ≤ 0,05), a hipótese de normalidade é rejeitada para estes dados. Após esta análise de normalidade dos dados, confirmou-se que a distribuição normal é apropriada para modelar as V. As. joelho, bíceps, antebraço e punho.

Tabela 1 – Resultados do Teste de Normalidade de Anderson-Darling

V. A.	Valor médio	Desvio-padrão (σ)	AD*	Nível de significância *p*	Distribuição Normal (Sim/Não)
Idade (anos)	41,89	11,67	1,362	0,002	N
Peso (kg)	81,05	13,47	1,455	0,001	N
Altura (m)	1,79	0,10	6,187	0,000	N
Pescoço (cm)	37,89	2,46	0,769	0,046	N
Tórax (cm)	100,13	8,12	0,864	0,027	N
Abdome (cm)	91,79	10,59	0,754	0,050	N
Quadris (cm)	99,95	7,30	2,380	0,000	N
Coxa (cm)	59,67	5,45	1,078	0,008	N
Joelho (cm)	38,58	2,35	0,634	0,098	S
Tornozelo (cm)	23,17	1,77	4,406	0,000	N
Bíceps (cm)	32,31	3,01	0,557	0,150	S
Antebraço (cm)	28,71	1,98	0,353	0,466	S
Punho (cm)	18,18	0,93	0,556	0,151	S

A partir desta afirmação, a função de densidade de probabilidade dessas V. As. pode ser modelada por (BENDAT, 1986):

$$p(x) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{\frac{-(x-\mu_x)^2}{2\sigma_x^2}}$$

onde μ_x é o valor médio e σ_x é o desvio-padrão da V. A. No caso contínuo, temos ainda que:

$$p(x) \geq 0$$

e,

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

A Figura 1 ilustra as funções de densidade de probabilidades obtidas para estas V. As., estimadas a partir do modelo da distribuição normal.

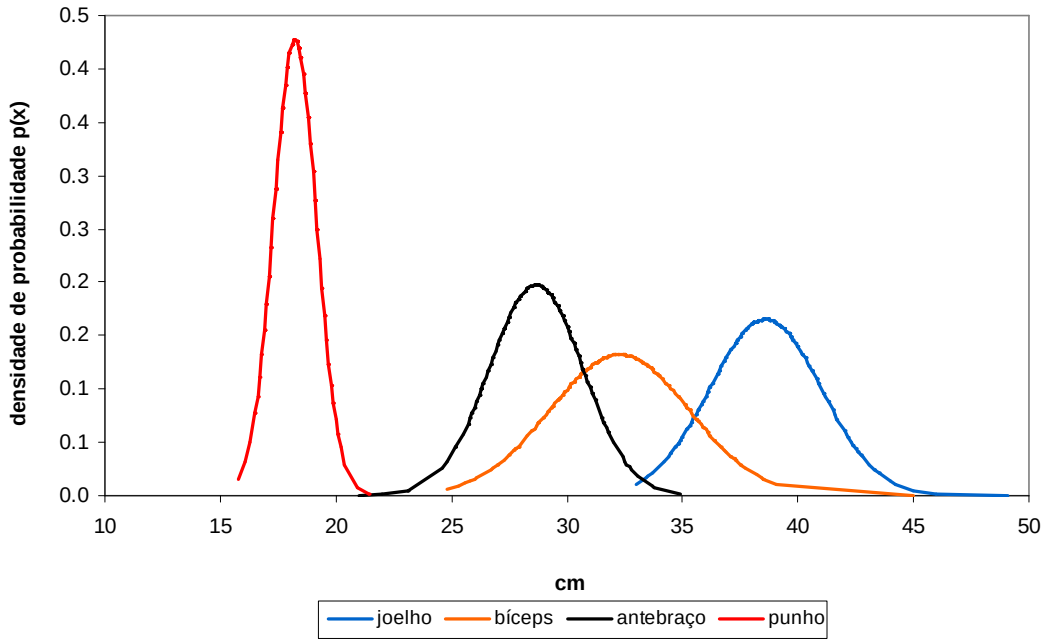


Figura 1 – Função de densidade de probabilidade das V. As. joelho, bíceps, antebraço punho.

A área parcial sob *p(x)*, de $-\infty$ a $+\infty$, para um dado valor *x*, representa a função de distribuição de probabilidade, denotada por *P(x)*, e definida como o conjunto de pontos *k* que satisfaz a desigualdade:

$$x(k) \leq k$$

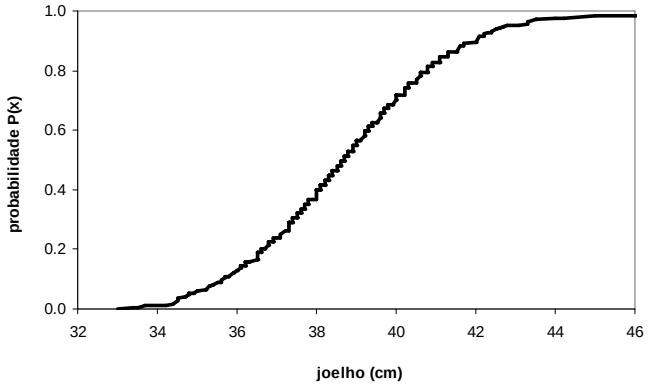
Desta forma, temos que:

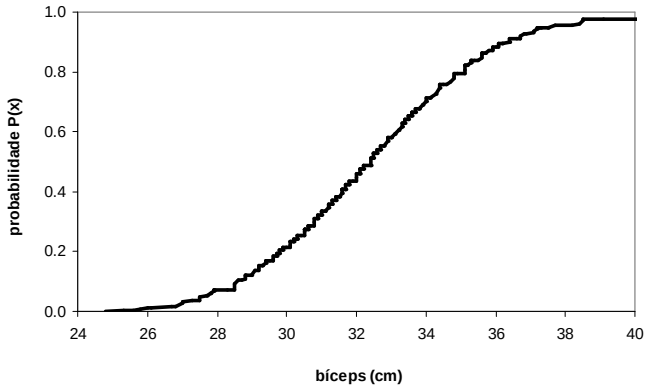
$$P(x) = prob[x(k) \leq x]$$

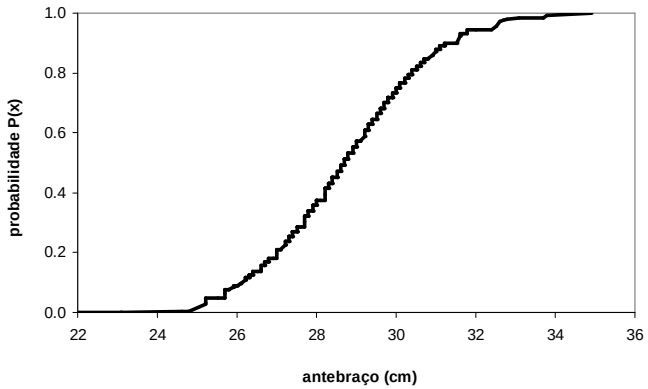
e,

$$\frac{dP(x)}{dx} = p(x)$$

A Figura 2 ilustra as funções de distribuição de probabilidade *P(x)* calculadas para as V. As. joelho, bíceps, antebraço e punho, estimadas a partir da integração da função de densidade de probabilidade *p(x)* de cada V. A. A partir das funções de distribuição de probabilidade *P(x)*, a probabilidade de ocorrência de uma determinada medida para as V. As. pode ser completamente definida. Por exemplo, é observado na Figura 2 que a probabilidade da circunferência de um joelho ser menor que 40 cm é de 70% ($prob[x(k) \leq 40] = 0{,}70$).







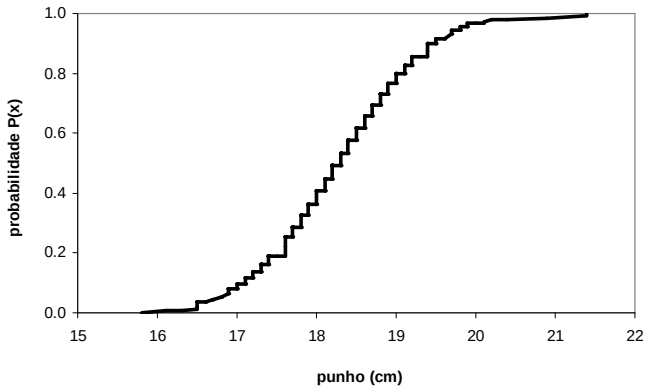


Figura 2 – Função de distribuição de probabilidade das V. As. joelho, bíceps, antebraço punho.

Considerações Finais

A pesquisa desenvolvida permitiu uma avaliação da hipótese de normalidade de dados observados na literatura, por meio do teste estatístico de Anderson-Darling. Das treze V. As. obtidas a partir de uma amostra de 252 indivíduos, o teste de Anderson-Darling confirmou que somente quatro das V. As. seguem uma distribuição normal.

A importância de se aplicar a distribuição normal na análise estatística de um dado conjunto de observações reside no fato de que a função de densidade de probabilidade e a função de distribuição de probabilidade podem ser definidas a partir da estimação da média e do desvio-padrão do conjunto de dados observados. Desta forma, em uma análise estatística, torna-se recomendável que testes de normalidade sejam aplicados para avaliar a hipótese de normalidade dos dados observados.

Fontes Consultadas

BENDAT, J. S.; PIERSOL, A. G. **Random data:** analysis and measurement procedures. Los Angeles: John Wiley & Sons, 1986.

BPI CONSULTING. **Anderson-Darling test for normality**. 2011. Disponível em:<http://www.spcforexcel.com/anderson-darling-test-for-normality>. Acesso em: 12 out. 2010.

BPI CONSULTING. **Normal probability plots**. 2009. Disponível em:

<http://www.spcforexcel.com/normal-probability-plots>. Acesso em: 12 out. 2010.

GIL, A C. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2009.

GODOY, Arild Schmidt. Introdução à pesquisa qualitativa e suas possibilidades. **Revista de Administração de Empresas**. São Paulo, v. 35, n. 2, p. 57-83, mar./abr. 1995a.

GODOY, Arild Schmidt. Introdução à pesquisa qualitativa e suas possibilidades. **Revista de Administração de Empresas**. São Paulo, v. 35, n. 3, p. 20-29, mai./jun. 1995b.

JOHNSON, R. W. Fitting Percentage of Body Fat to Simple Body Measurements. **Journal of Statistics Education**. 1996, v. 4, n. 1. Disponível em: <http://www.amstat.org/publications/jse/>. Acesso em: 17 mai. 2011

JUNG, Carlos F. **Elaboração de projetos de pesquisa aplicados a engenharia de produção**. Taquara: FACCAT, 2010. Disponível em: <http://www.metodologia.net.br>. Acesso em: 3 Abr 2012.

LARSON, R.; FARBER, B. **Estatística aplicada**. São Paulo: Prentice Hall, 2007.

MORAES, Celso Francisco de; FERREIRA, João Roberto; BALESTRASSI, Pedro Paulo. Análise crítica da aplicação de métodos estatísticos em processos definidos por dados que não apresentam distribuição normal. In: SIMPÓSIO DE ENGENHARIA DE PRODUÇÃO, 12., 2005, **Anais**... Bauru: UNESP, 2005.

SILVA, Gilzibene Marques da; ARAÚJO, Adrilayne dos Reis; COSTA FILHO, Galafre Guttemberg da. Análise de Séries Temporais de Pacientes com HIV/AIDS Internados no Hospital Universitário João de Barros Barreto (HUJBB), da Região Metropolitana de Belém, Estado do Pará. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 19., 2010, **Anais**... São Pedro: UNICAMP, 2010.

TAVARES, M. **Estatística aplicada à administração**. Brasília: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES. Universidade Aberta do Brasil – UAB, 2007.