https://doi.org/10.4322/2357-9730.89242

Estatística descritiva: perguntas que você sempre quis fazer, mas nunca teve coragem

Descriptive statistics: questions you have always wanted to ask, but never had the courage to

Aline Castello Branco Mancuso[1], Stela Maria de Jezus Castro[1,2], Luciano Santos Pinto Guimarães[1], Vanessa Bielefeldt Leotti[1,2], Vânia Naomi Hirakata[1], Suzi Alves Camey[1,2]

Clin Biomed Res. 2018;38(4):414-418

1 Unidade de Bioestatística, Grupo de Pesquisa e Pós-graduação (GPPG), Hospital de Clínicas de Porto Alegre (HCPA). Porto Alegre, RS, Brasil.

2 Departamento de Estatística, Instituto de Matemática, Universidade Federal do Rio Grande do Sul (UFRGS). Porto Alegre, RS, Brasil.

Autor correspondente:
Aline Castello Branco Mancuso
l-bioestatistica@hcpa.edu.br
Unidade de Bioestatística, Grupo de Pesquisa e Pós-graduação (GPPG), Hospital de Clínicas de Porto Alegre (HCPA)
Rua Ramiro Barcelos, 2350.
90035-007, Porto Alegre, RS, Brasil.

RESUMO

A revista do HCPA (*Clinical & Biomedical Research*) está reabrindo a seção de Bioestatística com o intuito de apresentar artigos explicativos, conceituais ou tutoriais, de modo a elucidar os leitores sobre os mais diversos temas estatísticos. Neste contexto, este artigo será o primeiro de uma série que tem como objetivo responder algumas das questões mais levantadas por pesquisadores da área da saúde. Começando pela Estatística Descritiva, alguns conceitos são esclarecidos e diversas referências são indicadas para o estudo do tema e para análises em SPSS ou R-project.

Palavras-chave: *Estatística descritiva; programa estatístico; desvio-padrão; erro-padrão; SPSS; R-project*

ABSTRACT

The HCPA journal (Clinical & Biomedical Research) is reopening its Biostatistics section with the aim of presenting readers with explanatory, conceptual or tutorial articles on a wide range of statistical topics. In this context, this is the first in a series of articles seeking to answer some of the questions raised by health researchers. Starting with descriptive statistics, some concepts are introduced and several references are indicated for those interested in studying the topic and performing analyses in SPSS or R-project.

Keywords: *Descriptive statistics; statistical software; standard deviation; standard error; SPSS; R-project*

Com o crescimento e desenvolvimento tecnológico uma nova Era se desenhou: a Era do *Big Data*. Nesta nova era, não apenas novas metodologias de análise estatística estão surgindo, mas técnicas consagradas também estão sendo repensadas e aprimoradas. Para acompanhar este desenvolvimento é preciso não apenas estar em constante aprendizagem, mas também ter uma base teórica consolidada, de modo a garantir o entendimento e compreensão de conceitos básicos.

Neste contexto, a revista do HCPA (*Clinical & Biomedical Research*) está reabrindo a seção de Bioestatística com o objetivo de apresentar artigos explicativos, conceituais ou tutoriais, de modo a elucidar os leitores sobre os mais diversos temas estatísticos. Com o intuito de publicar artigos adequados à necessidade dos pesquisadores no atual cenário estatístico, uma enquete em formulário eletrônico foi encaminhada aos pesquisadores do HCPA. Foram realizadas apenas duas perguntas: 1) quais assuntos gostariam que fossem abordados na seção de Bioestatística e 2) qual o nível de instrução do respondente.

ISSN 2357-9730

Entre os 140 respondentes, além de sugestões de temas estatísticos, também surgiram muitas perguntas, das quais grande parte versavam sobre assuntos básicos de estatística ou já abordados na literatura. Diante desta realidade, uma série de artigos será apresentada, respondendo e sugerindo referências para o melhor entendimento das principais dúvidas observadas na enquete. No entanto, estes artigos não terão como objetivo ensinar como calcular medidas estatísticas ou fazer testes, apenas auxiliar na interpretação e uso dos mesmos.

Começando pelo que se entende ser a primeira análise estatística de qualquer estudo ou pesquisa desenvolvida, este artigo responde algumas das principais questões levantadas sobre Estatística Descritiva:

ANÁLISE DESCRITIVA: QUAIS OS PRINCIPAIS PONTOS A SEREM AVALIADOS?

A análise descritiva, ou síntese numérica, é a etapa inicial de qualquer estudo. Podendo também ser considerada a mais importante, visto que é a partir dela que definimos como e qual análise será utilizada. Erros, nesta primeira etapa, podem invalidar todas as demais.

A descrição dos dados tem como objetivo básico resumir uma série de valores de mesma natureza através de um conjunto de ferramentas e técnicas: tabelas, gráficos, medidas (estatísticas) de variabilidade e de tendência central que ajudam na produção de uma visão global dos dados[1]. Bussab e Morettin[2], Farber e Larson[3] e Crespo[4] apresentam algumas das principais medidas-resumo.

No entanto, não basta resumir. Uma análise cuidadosa destes resultados é crucial. Um dos principais pontos a serem avaliados é a existência de erros de digitação. Nas variáveis categóricas é fácil detectá-los através da tabela de frequências, já nas variáveis quantitativas é importante verificar se o mínimo e o máximo observado estão conforme o esperado. Cabe salientar que valores codificados como *missings* (dados faltantes) até podem estar inclusos nas tabelas de frequências (devidamente identificados), mas não devem ser incluídos nas estatísticas descritivas das variáveis quantitativas.

Outro importante ponto a ser avaliado é a presença de *outliers* (valores atípicos). Aconselha-se, primeiramente, verificar se tal valor foi corretamente registrado, posteriormente, qual a possível causa desta anomalia e se deve ou não ser excluído. Esta qualificação deve ser cautelosa, pois o mesmo pode viesar os resultados, mas também pode ser exatamente o que se está procurando. O *Boxplot* (diagrama de caixas) é uma das ferramentas mais conhecidas para auxiliar na detecção dos mesmos, mas muitos outros métodos e técnicas podem ser encontradas na literatura. Hawkis[5] é uma referência nesta área e Figueira[6] cita diferentes referências.

COMO APRESENTAR OS RESULTADOS DA ANÁLISE DESCRITIVA?

A primeira tabela do capítulo de resultados de um artigo científico é, geralmente, a tabela que descreve a amostra estudada. Ou seja, a tabela que apresenta os resultados descritivos. No meio acadêmico, ela é conhecida, carinhosamente, como "tabela 1". A Tabela 1 apresenta um exemplo fictício para os resultados descritivo de um ensaio clínico, por exemplo.

A Tabela 1 apresenta a média e o desvio padrão para a IDADE (variável simétrica), a mediana e os intervalos interquartílicos (P25% - P75%) para o PESO (variável assimétrica) e a frequência absoluta e relativa para SEXO e TABAGISMOS (variáveis categóricas). Observa-se, também, que quando existem apenas duas categorias em uma variável é desnecessária a apresentação de ambas, visto que é possível obter o percentual masculino, por exemplo, através dos resultados feminino.

A mesma tabela pode ser adaptada para os demais delineamentos. Quando a amostra é única e não existem grupos a serem comparados, basta excluir as duas colunas centrais e proceder de forma análoga.

Tabela 1: Exemplo de apresentação para análise descritiva.

Características	Grupo A 133 (54,7%)	Grupo B 110 (45,3%)	Total
Idade	42 ± 15,3	50 ± 11,8	47 ± 13,1
Sexo Feminino	72 (54,1%)	58 (52,7%)	130 (53,5%)
Peso	73 (63-83)	80 (75-85)	75 (65-85)
Tabagismo			
Fumante	5 (3,8%)	10 (9,1%)	17 (7,0%)
Ex-Fumante	25 (18,8%)	23 (20,9%)	46 (18,9%)
Nunca Fumou	103 (77,4%)	77 (70,0%)	180 (74,1%)

COMO REALIZAR ANÁLISE DESCRITIVA NO R?

Iniciado em 1993, o software R-project[7] vem ganhando cada vez mais adeptos. Além de ser uma plataforma inteiramente gratuita para análises estatísticas básicas e avançadas, é uma ferramenta extremamente poderosa, pois sua capacidade vem sendo constantemente aumentada através de pacotes (*packages*) criados por qualquer colaborador interessado em compartilhar seus programas.

Devido a esta extensa lista de colaboradores, é possível encontrar diversas formas de se realizar uma mesma análise estatística no R. Mais pontualmente, para a análise descritiva, o pacote *DescTools*[8] é bastante prático. Por exemplo, se o arquivo de dados se chama 'banco', ao se executar o comando '*Desc(banco)*' tem-se uma descrição de cada variável através de medidas estatísticas e gráficos, dado que o banco esteja corretamente formatado.

Além dos diversos e mais variados *blogs* já criados pelos "*R-users"*, existem muitas tutoriais para estatística básica disponíveis *online*. Já na literatura, cita-se Field et al.[9], Mello e Peternelli[10] e Jelihovschi[11].

COMO REALIZAR ANÁLISE DESCRITIVA NO SPSS?

Distribuído pela IBM, empresa classificada como visionária ou líder em diversos ramos pelo Quadrante Mágico de Gartner de 2018, o SPSS é um dos softwares mais populares em análise estatística. Além do já citado Field[12], atualmente é possível encontrar diversos tutoriais disponíveis *online*, mas destaca-se aqui o site da Unidade de Bioestatística do GPPG/HCPA[13]. O mesmo também apresenta um tutorial de como importar o banco de dados para SPSS e de diversas outras análises estatísticas.

QUAL A DIFERENÇA ENTRE DESVIO-PADRÃO, VARIÂNCIA E ERRO-PADRÃO?

O contraste entre os termos 'desvio padrão' e 'erro padrão' reflete a distinção entre 'descrição de dados' e 'inferência'. Em outros termos: o desvio padrão e a variância são medidas de variabilidade das observações; já o erro padrão pode ser entendido como uma medida de precisão da média da amostra.

Em uma amostra, o **desvio padrão (dp)** de uma variável é uma estimativa da variabilidade desta variável na população de onde a amostra foi retirada. Independente da distribuição da variável, cerca de 95% dos valores observados estão entre 2 desvios padrões abaixo e acima da média desta variável. Para uma variável com distribuição Gaussiana, também conhecida como distribuição Normal (que é uma distribuição simétrica), os outros 5% estarão igualmente distribuídos abaixo e acima destes limites. Já para uma variável assimétrica, os valores que estão fora destes limites podem estar concentrados em uma das extremidades. Nestas situações, costuma-se utilizar outra estatística como medida resumo da variabilidade: a amplitude interquartílica[14].

A Figura 1 ilustra três possíveis situações. Por exemplo, foi medido o PESO (em kg) em uma amostra de 100 indivíduos, cuja média foi igual a 65 kg e o desvio padrão foi igual a 5 kg. Em outras

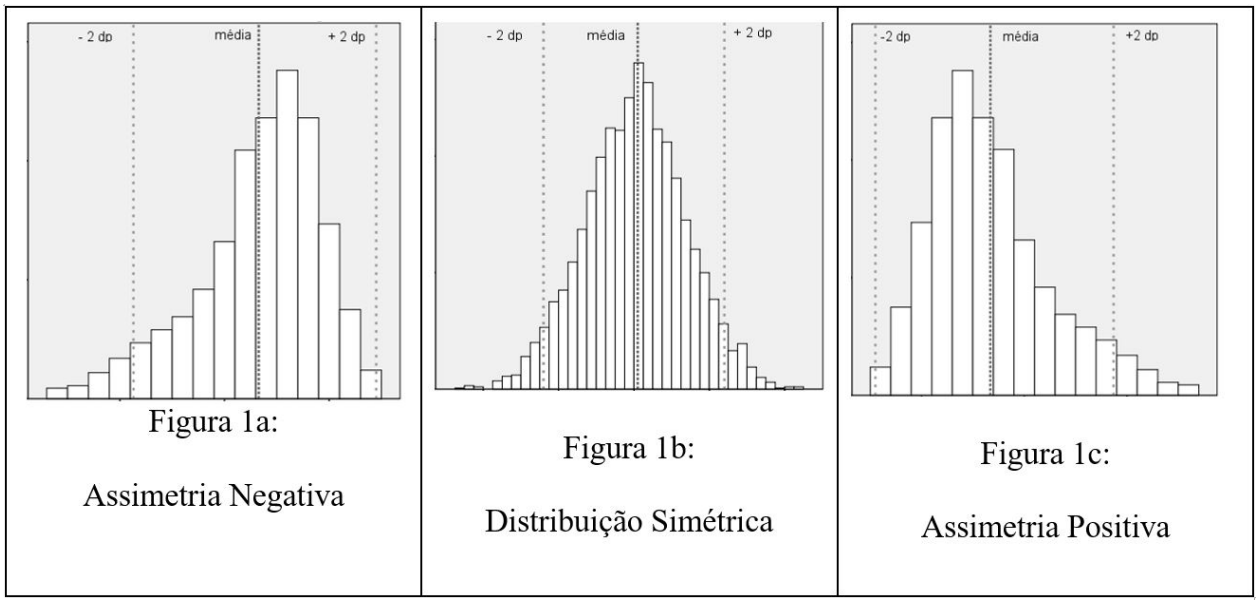

Figura 1: Exemplos de distribuições.

palavras, cerca de 95 indivíduos têm seu peso entre 55 kg e 75 kg (a média mais ou menos dois desvios padrões). Já a **variância** nada mais é do que o quadrado do desvio padrão. No exemplo, a variância do peso será de 25 kg^2.

Contudo, geralmente o interesse não está na média de uma particular amostra, e sim na média da variável para todos os indivíduos da população de onde a amostra foi retirada. No entanto, um censo (coleta de informações de toda a população em estudo) nem sempre é possível. Na prática, os dados são coletados apenas de uma única amostra desta população de interesse. Mas, inúmeras amostras diferentes de mesmo tamanho (n) podem ser coletadas da mesma população e diferentes médias podem ser observadas. A Figura 1 ilustra um processo de selecionar amostras de uma população para o estudo do PESO.

Na Figura 2, percebe-se que diferentes amostras podem apresentar diferentes médias. Este conjunto de diferentes médias é descrito pela chamada "Distribuição Amostral de Médias". Esta distribuição amostral apresenta o comportamento das amostras da população, com a média e o desvio padrão do conjunto de médias, agora também tratado como variável. No exemplo da Figura 1, 5 médias foram observadas, cuja média (das médias) foi 70 kg e o desvio padrão (das médias) foi 11,83 kg. E é este desvio padrão da estimativa da média que chamamos de **erro padrão** (ou erro padrão da média). Como o **erro padrão** é um tipo de desvio padrão, a confusão entre os dois termos é comum e compreensível[12].

No entanto, se não é possível analisarmos inúmeras amostras, como calcular o erro padrão? Para tal, pode-se estimar o erro padrão (EP) através do desvio padrão (DP) e do tamanho (n) de uma amostra, tal que $EP = DP/\sqrt{n}$. Portanto a magnitude do erro padrão decresce à medida que o tamanho da amostra aumenta. Por outro lado, a magnitude do desvio padrão não é influenciada pelo tamanho da amostra.

Voltando ao exemplo anterior de 100 indivíduos com média de 65 ± 5 kg, o erro padrão da média do peso é $EP = 5/\sqrt{100} = 5/10 = 0,5$, logo em cerca de 95% das amostras de tamanho 100 da mesma população, o peso médio estará entre 64 kg e 66 kg, a média (65 kg) mais ou menos 2 erros padrões (1kg).

Uma melhor compreensão sobre a diferença entre os termos também pode ser vista em Altman e Bland[14] e Field[12].

QUAIS AS PRINCIPAIS REFERÊNCIAS PARA ESTATÍSTICA BÁSICA E CONCEITOS INICIAIS: MÉDIA, MEDIANA, DESVIO, ETC?

Atualmente, o conhecimento de estatística básica é fundamental para qualquer profissional, seja ele da saúde ou não, seja do meio acadêmico ou não. Desde notícias esportivas, manchetes em jornais, revistas, bulas, etc à artigos acadêmicos, os conceitos básicos tornam-se essenciais para a correta compreensão dos mesmos. Até mesmo para a tomada de decisões.

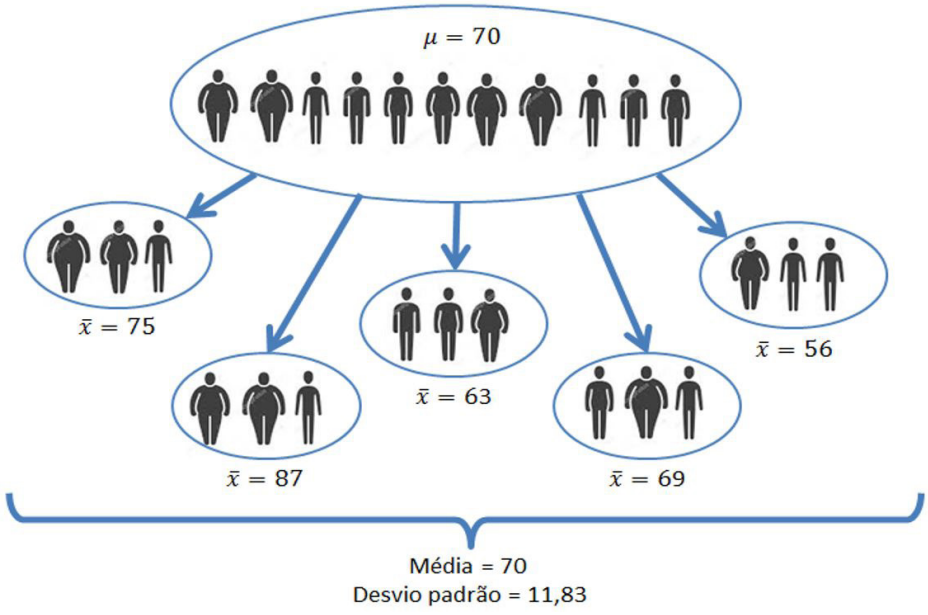

Figura 2: Processo de selecionar amostras de uma população.

Para tal, existem diversas referências: uma bibliografia em inglês muito citada é Zar[15], com mais de 85 mil citações no Google Acadêmico. Em Português, sugere-se os livros de Soares e Siqueira[1], Callegari-Jacques[16] e Vieira[17]. E também é possível encontrar muitos *blogs* sobre o assunto, além de diversos canais com videoaulas no Youtube.

Conflitos de Interesse

Os autores declaram não ter conflitos de interesse

REFERÊNCIAS

1. Soares JF, Siqueira AL. *Introdução à estatística médica*. Belo Horizonte: Departamento de Estatística da UFMG; 2002.
2. Bussab WO, Morettin PA. *Estatística básica*. São Paulo: Saraiva; 2010.
3. Farber L, Larson R. *Estatística aplicada*. 4. ed. São Paulo: Pearson; 2010.
4. Crespo AA. *Estatística fácil*. São Paulo: Saraiva; 2017.
5. Hawkins DM. *Identification of outliers*. Vol. 11. London: Chapman and Hall; 1980. http://dx.doi.org/10.1007/978-94-015-3994-4.
6. Figueira MMC. Identificação de outliers. *Revista Millenium Online, Viseu*, 12, 1998.
7. R Foundation. The R Project for Statistical Computing. Vienna: R Foundation [citado 2019 Jan 15]. Disponível em: www.r-project.org
8. Signorell A. *DescTools: tools for descriptive statistics*. R package version 0.99. Vol. 18. Vienna: R Foundation; 2016.
9. Field A, Miles J, Field Z. *Discovering Statistics Using R*. London: SAGE; 2012.
10. Mello MP, Peternelli LA. *Conhecendo o R: uma visão mais que estatística* [dissertação]. Viçosa: Universidade Federal de Viçosa; 2013.
11. Jelihovschi E. *Análise exploratória de dados usando o R*. Ilhéus: Editus; 2014.
12. Field A. *Descobrindo a estatística usando o SPSS*. 2. ed. São Paulo: Bookman; 2009.
13. Hospital de Clínicas de Porto Alegre (HCPA). Cursos de SPSS. Assessorias Estatísticas [citado 2019 Jan 15]. Disponível em: sites.google.com/a/hcpa.edu.br/bioestatística
14. Altman DG, Bland JM. Standard deviations and standard errors. *BMJ*. 2005;331(7521):903. http://dx.doi.org/10.1136/bmj.331.7521.903. PMid:16223828.
15. Zar JH. *Biostatistical analysis*. Upper Saddle River: Prentice-Hall; 2009.
16. Callegari-Jacques SM. *Bioestatística: princípios e aplicações*. Porto Alegre: Artmed; 2009.
17. Vieira S. *Introdução à bioestatística*. São Paulo: Elsevier; 2015.

Recebido: 27 dez, 2018
Aceito: 15 jan, 2019