ISSN: 2317-0840

# Pacote em ambiente R para automatizar estatísticas descritivas

*Emmanuel Arnhold*

*Escola de Veterinária e Zootecnia, Campus Samambaia, Rodovia Goiânia-Nova Veneza, Caixa Postal 131, CEP 74001-970, Goiânia/GO. E-mail: earnhold@pq.cnpq.br*

**Resumo:** Objetivou-se criar pacote em ambiente R, denominado `ds` com título "Descriptive Statistics", a fim de disponibilizar funções para automatizar diversas estatísticas descritivas. O pacote dispõe a função `gds()`, que retorna, para uma única variável ou um conjunto de variáveis, diversas medidas de estatística descritiva, como a média, o máximo, o mínimo, a mediana, a média mais 1, 2 e 3 desvios padrões, a média menos 1, 2 e 3 desvios padrões, os quantis 0,14%, 0,28%, 15,87%, 84,14%, 97,73%, 99,87%, o tamanho amostral (n), a amplitude, a variância, o desvio padrão, o erro padrão da média, o coeficiente de variação, a assimetria, a curtose e o valor de probabilidade pelo teste de normalidade de Shapiro-Wilk. O pacote também dispõe a função `dscor()`, que estima a correlação de Pearson ou Spearman para um conjunto de variáveis, e testa as correlações obtidas utilizando o teste t. O pacote ainda disponibiliza a função `dplot()` para observar um conjunto de variáveis quantitativas em um gráfico de dispersão e da função `tables()` que tabula variáveis qualitativas, apresentado as freqüências absolutas e percentuais e realizando testes Qui-quadrado e Exato de Fisher.

**Palavras-chave:** software, estatística, análise de dados

**Abstract:** The objective was to create package in R environment, called `ds` with title "Descriptive Statistics" in order to provide functions to automate various descriptive statistics. The package comprises a `gds()` function, which returns to a single variable or set of variables, several measures of descriptive statistics such as average, maximum, minimum, the median, average over 1, 2 and 3 standard deviations of the mean minus 1, 2 and 3 standard deviations of the quantiles 0.14%, 0.28%, 15.87%, 84.14%, 97.73%, 99.87%, the sample size (n), the amplitude, variance, standard deviation, standard error of the mean, the coefficient of variation in asymmetry, kurtosis and probability value of the Shapiro-Wilk normality test. The package also has the `dscor()` function, which estimates the Pearson or Spearman correlation for a set of variables, and test correlations using the t test. The package also provides the function `dplot()` to observe a set of quantitative variables on a scatter plot and function `tables()` which tabulates qualitative variables, presented as absolute frequencies and percentage and performing Chi-square and Fisher exact tests.

**Index terms:** software, statistics, data analysis

## Introdução

O ambiente R (R CORE TEAM, 2013) foi criado em 1996 por Ross Ihaka e Robert Gentleman na universidade de Auckaland, Nova Zelândia (PETERNELLI; MELLO, 2011). Foi posteriormente desenvolvido por colaboradores de vários locais do mundo. Entre suas vantagens tem-se a possibilidade de ampliação de suas funções, devido à fácil programação e ao sistema de uso de "pacotes", que são complementos contendo funções específicas, ampliando enormemente a capacidade de análises. Um conjunto de pacotes básicos é incluído com a instalação do R. Porém, muitos outros pacotes estão disponíveis. Atualmente existem mais de 7000 pacotes para usos em diversas áreas do conhecimento. Alguns destes pacotes são bastante importantes e de uso mais geral, como o pacote "multcomp" (HOTHORN; BRETZ; WESTFALL, 2008), com funções que realizam testes de contrastes. Outros pacotes são se uso mais específico, como o "pedigreemm" (VASQUEZ et al., 2006), utilizado no melhoramento genético com uso de matriz de parentesco em modelos lineares generalizados mistos.

O R torna-se, portanto, importante ferramenta tecnológica na análise e manipulação de dados, realizando análise de variância, testes estatísticos, operações matemáticas, simulação, modelagem linear e não linear, séries temporais, análise de sobrevivência, análise multivariada, entre outras, além de apresentar facilidade na elaboração de gráficos, no qual o usuário tem pleno controle.

Existem inúmeras funções disponíveis no R base e em pacotes para realizar análises de estatística descritiva. No entanto, as funções disponíveis no R base realizam análises de medidas isoladas como a média, variância e desvio padrão. Para obter a média para um conjunto de variáveis, pode-se utilizar a função `mean()` em conjunto com a `lapply()` ou `sapply()`, por exemplo. Fazer esta programação para várias medidas de estatística descritiva além de demandar tempo adicional, pode ser um tanto complicada para usuários menos experientes. O pacote R denominado *ExpDes* (FERREIRA et al., 2013), também automatiza análises que podem ser feitas com o R base, e tem tido bastante sucesso como ferramenta de análise de dados, sendo utilizado tanto por usuários menos experientes e até por usuários mais experientes buscando praticidade na análise de dados.

Portanto, desenvolveu-se o pacote `ds`, com título "Descriptive Statistics". O objetivo foi disponibilizar funções em ambiente R a fim de obter, de forma prática, diversas análises de estatística descritiva.

## Pacote `ds`

O pacote "ds" dispõe da função "gds(data)", que possui único argumento denominado "data", que informa o conjunto de dados que deve ser uma tabela em objeto R da classe "data.frame" ou "matrix". Nesta tabela cada coluna refere-se a uma variável quantitativa (numérica).

A função retorna uma tabela (objeto R da classe "data.frame") que poder ser facilmente salva em arquivos txt ou xls. Esta tabela contém, para cada variável, a média, o máximo, o mínimo, a mediana, a média mais 1, 2 e 3 desvios padrões, a média menos 1, 2 e 3 desvios padrões, os quantis 0,14%, 0,28%, 15,87%, 84,14%, 97,73%, 99,87%, o tamanho amostral (n), a amplitude, a variância, o desvio padrão, o erro padrão da média, o coeficiente de variação, a assimetria, a curtose e o valor de probabilidade pelo teste de normalidade de Shapiro-Wilk.

Caso ocorra a falta de algum dado utiliza-se `NA` no lugar do dado faltante e a análise

ocorre já considerando `NA`.

Para exemplificar vamos utilizar dados de Kaps e Lamberson (2009), estimando estatísticas descritivas para o Peso do Coração (PC), em grama, e a Circunferência do Coração (CC), em cm, para 10 vacas Gir.

```
PC<-c(641, 620, 633, 651, 640, 666, 650, 688, 680, 670)
CC<-c(205, 212, 213, 216, 216, 217, 218, 219, 221, 226)
```

Criando uma tabela chamada de “dados”:

```
dados<-data.frame(PC,CC)
```

Realizando a análise com a função “gds” abaixo se obtém o resultado expresso na Figura 1.

```
gds(dados)
```

```
Console ~/Área de Trabalho/
>
> gds(dados)
                                      PC       CC
Mean                            653.9000 216.3000
Maximum                         688.0000 226.0000
Minimum                         620.0000 205.0000
Median                          650.5000 216.5000
Mean + 3 standard deviation     718.8531 233.1553
Mean + 2 standard deviation     697.2020 227.5368
Mean + 1 standard deviation     675.5510 221.9184
Mean - 1 standard deviation     632.2490 210.6816
Mean - 2 standard deviation     610.5980 205.0632
Mean - 3 standard deviation     588.9469 199.4447
Quantile (99.87%)               687.9064 225.9415
Quantile (97.73%)               686.3656 224.9785
Quantile (84.14%)               675.7260 220.1452
Quantile (15.87%)               635.9981 212.4283
Quantile (2.28%)                622.6676 206.4364
Quantile (0.14%)                620.1638 205.0882
n                                10.0000  10.0000
Range                            68.0000  21.0000
Variance                        468.7667  31.5667
Standard deviation               21.6510   5.6184
Standard error of the mean        6.8467   1.7767
Coefficient of variation (%)      3.3111   2.5975
Skewness                          0.1197  -0.3414
Kurtosis                          1.9561   3.2421
P-value (Shapiro-Wilk)            0.9027   0.8684
```

Figura 1. Estatísticas descritivas obtidas para as variáveis Peso do Coração (PC), em grama, e Circunferência do Coração (CC), em cm, através da função `gds` do pacote R `ds`

O pacote `ds` também dispõe da função `dscor(data, method=1, option=1)`, que possui três argumentos. O argumento denominado `data` informa o conjunto de dados que deve ser uma tabela em objeto R da classe "data.frame" ou "matrix". Nesta tabela, cada coluna refere-se a uma variável quantitativa ou numérica, as quais se desejam estimar correlações. O argumento `method` define o método de obtenção da correlação, sendo o método 1 (`method=1`) o padrão, retornando a correlação de Pearson, e o método 2 (`method=2`) a correlação de Spearman. O terceiro argumento define a forma de apresentação das correlações e dos valores de probabilidade pelo teste t. Na opção 1 (`option=1`) a função retorna uma tabela ("data.frame") onde a primeira coluna identifica os pares de variáveis, a segunda coluna a correlação estimada para cada par e a terceira o valor de probabilidade pelo teste t. Na opção 2 (`option=2`) a função retorna uma matriz (objeto R da classe "matrix"), sendo as correlações localizadas acima da diagonal e os valores de probabilidade abaixo da diagonal. Para esta função, caso ocorra a falta de algum dado utiliza-se `NA` no lugar do dado faltante e a análise também ocorre já considerando `NA`.

Para exemplificar vamos utilizar o mesmo exemplo em que foi aplicado na função `gds`, o qual foi denominado `dados`. Realizando a análise com a função `dscor`, com o método de Pearson (`method=1`) e de Spearman (`method=2`), com a saída do resultado em forma de tabela (`option=1`), e alterando para um resultado em forma de matriz (`option=2`), obtendo resultado expresso na Figura 2.

```
dscor(dados, method=1, option=1 )
dscor(dados, method=1, option=2 )
dscor(dados, method=2, option=1 )
dscor(dados, method=2, option=2 )
```

```
Console ~/Área de Trabalho/
> dados
    PC  CC
1  641 205
2  620 212
3  633 213
4  651 216
5  640 216
6  666 217
7  650 218
8  688 219
9  680 221
10 670 226
>
> dscor(dados, method=1, option=1 )
      pairs correlation p-value
1 PC and CC      0.6744  0.0325
> dscor(dados, method=1, option=2 )
       PC     CC
PC 1.0000 0.6744
CC 0.0325 1.0000
> dscor(dados, method=2, option=1 )
      pairs correlation p-value
1 PC and CC      0.8268   <0.01
> dscor(dados, method=2, option=2 )
       PC     CC
PC 1.0000 0.8268
CC 0.0032 1.0000
```

Figura 2. Correlações de Pearson e Spearman estimadas para as variáveis Peso do Coração (PC), e Circunferência do Coração (CC), através da função `dscor` do pacote R `ds`

Para completar, o pacote também disponibiliza a função

```
dplot(data, xlab="Variable x", ylab="Variable y", position=1, colors = TRUE, type="o", mean=TRUE),
```

que possui os argumentos `xlab` e `ylab` para definir os nomes dos eixos x e y, respectivamente. Também possui o argumento "position" que define a posição da legenda e o argumento `colors` que define se o gráfico será colorido (TRUE) ou não (FALSE). O argumento `type` usa as mesmas definições da função `plot()` do R `base`. O argumento `mean` define se o gráfico será gerado com médias (opção `TRUE`) ou com todos os dados (opção `FALSE`). Para exemplificar vamos utilizar dados de Sampaio (2010) (Tabela 1).

Tabela 1. Qualidade média dos ovos (Unidade Haugh) de uma linhagem de aves, estocados a temperatura de 28 ºC, segundo a embalagem e o tempo de estocagem*.

| Embalagem | Tempo (dias) | | | |
|---|---|---|---|---|
| | 7 | 14 | 21 | 28 |
| Sem embalagem | 89 | 85 | 80 | 74 |
| Saco plástico | 92 | 89 | 85 | 81 |
| Película de cera | 96 | 93 | 92 | 90 |

*Sampaio (2010)

Os dados da Tabela 1 podem entrar em formato de tabela (`data.frame`) conforme a Figura 3, gerando o gráfico da Figura 4.

```
Console ~/Área de Trabalho/
>
> dados
  Tempo Sem embalagem Saco plástico Película de cera
1     7            89            92               96
2    14            85            89               93
3    21            80            85               92
4    28            74            81               90
> dplot(dados, xlab="Tempo de estocagem (dias)", ylab="Unidades
Haugh", position=3)
```

Figura 3. Tabela contendo os dados e programação na função `dplot` para gerar gráfico de dispersão dos dados

E ainda, para análise de variáveis qualitativas, o pacote `ds` disponibiliza a função `tables(data)`, com argumento único `data` que deve ser uma tabela (`data.frame` ou `matrix`) contendo as variáveis. A função gera tabelas com a frequência absoluta, freqüência relativa e teste Qui-quadrado para aderência de cada variável (coluna). E também, gera tabelas de contingência associando a primeira variável (primeira coluna) com as demais, apresentando as freqüências absolutas e percentuais, e os testes Qui-quadrado e Exato de Fisher, para testar a associação entre variáveis.

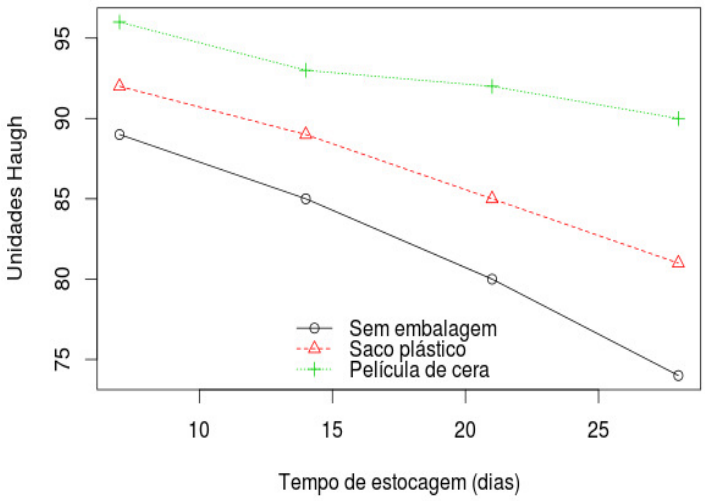


Figura 4. Gráfico gerado pela função `dplot` do pacote `ds`

Para exemplificar vamos utilizar dados de Ferro (2013), com a programação abaixo.

```
sexo<-rep(c("fêmea","macho"), c(69,15))
chifre<-c(rep(c("extremidade grossa", "extremidade afinada"), c(19,
50)),rep(c("extremidade grossa", "extremidade afinada"), c(10, 5)))
mucosa<-c(rep(c("escura", "rosada"), c(51, 18)),rep(c("escura",
"rosada"), c(12, 3)))
dados<-data.frame(sexo,chifre,mucosa)
```

Os dados são relativos a características de bovinos da raça Pantaneira. Utilizando a função `tables` a seguir, tem-se resultado resumido nas Tabelas 2 e 3.

```
tables(dados)
```

Tabela 2. Frequências absolutas e percentuais e teste Qui-quadrado para aderência considerando freqüências esperadas iguais, para o variável sexo, chifre e mucosa de bovinos da raça Pantaneira (dados extraídos de Ferro, 2013)

| Variável | Categorias | Frequências | | P* |
|---|---|---|---|---|
| | | Absoluta | Percentual | |
| Sexo | Macho | 69 | 82,14 | <0,0001 |
| | Fêmea | 15 | 17,86 | |
| Chifre | Extremidade Grossa | 29 | 34,52 | 0,0046 |
| | Extremidade Afinada | 55 | 65,48 | |
| Mucosa | Escura | 63 | 75,00 | <0,0001 |
| | Rosada | 21 | 25,00 | |

* Qui-quadrado para aderência considerando freqüências esperadas iguais

Tabela 3. Tabela de contingência associando a variável sexo as variáveis chifre e mucosa de bovinos Pantaneiros (dados extraídos de Ferro, 2013), com frequências absolutas, freqüências percentuais e valores de probabilidade pelos testes Qui-quadrado e Exato de Fisher

| Variável | Categorias | Sexo | | P* | P** |
|---|---|---|---|---|---|
| | | Fêmea | Macho | | |
| Chifre | Extremidade Grossa | 19 (27,54%) | 10 (66,67%) | 0,0096 | 0,0064 |
| | Extremidade Afinada | 50 (72,46%) | 5 (33,33%) | | |
| Mucosa | Escura | 51 (73,91%) | 12 (80,00%) | 0,8694 | 0,7512 |
| | Rosada | 18 (26,09%) | 3 (20,00%) | | |

* Qui-quadrado para independência. ** Teste exato de Fisher

Assim, pode-se concluir que o pacote `ds` dispõe de funções que podem ser utilizadas com bastante propriedade para o analista de dados que busca praticidade em análises descritivas. Devido, em geral, a maior praticidade e facilidade de uso das funções aqui apresentadas, pode-se considerar como potenciais usuários, principalmente estudantes de graduação e pós-graduação de cursos com pouca afinidade em estatística, matemática e iniciantes no ambiente R.

**Referências**

FERRO, D.A.C., Proposta de grupos de acasalamento para bovinos pantaneiro por meio da caracterização morfológica. **Dissertação de Mestrado**. Universidade Federal de Goiás. 2013. 49p.

FERREIRA, E. B., CAVALCANTI, P. P., NOGUEIRA, D. A. (2013). ExpDes: Experimental Designs pacakge. R package version 1.1.2. http://CRAN.R-project.org/package=ExpDes

HOTHORN, T.; BRETZ, F.; WESTFALL, P. Simultaneous Inference in General Parametric Models. **Biometrical Journal,** v.50, p.346-363, 2008.

KAPS, M.; LAMBERSON, W.R. **Biostatistics for Animal Science: an introductory text.** CABI Publishing, Wallingford, Oxfordshire, UK, 2009. 504p.

PETERNELLI, L.A.; MELLO, M.P. **Conhecendo o R: Uma Visão Estatística**. 1ª ed. Viçosa: Editora UFV, 2011. 185p.

SAMPAIO, I. B. M. **Estatística aplicada a experimentação animal.** 3ª Edição. Belo Horizonte: Editora FEPMVZ, Fundação de Ensino e Pesquisa em Medicina Veterinária e Zootecnia, 2010. 264p.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. http://www.R-project

VAZQUEZ, A.I.; BATES, D.M.; ROSA, G.J.M.; GIANOLA, D; WEIGEL, K.A. Technical note: an R package for fitting generalized linear mixed models in animal breeding. **Journal of Animal Science,** v.88, p.497-504, 2010.