BEANPLOT UMA NOVA FERRAMENTA GRÁFICA

BEANPLOT A NEW GRAPHICAL TOOL

Suzi Alves Camey[1-4], Luciana Neves Nunes[1,2], Luciane Nascimento Cruz[3,4]

RESUMO

A estatística descritiva é uma poderosa ferramenta para se analisar conjuntos de dados, entretanto é muito pouco utilizada. Uma análise descritiva bem conduzida pode evitar vários problemas que podem ocorrer em análises mais complexas, além de fornecer um retrato da amostra em estudo. Na estatística descritiva existem os métodos gráficos, que se bem empregados, são bem mais informativos que tabelas. Dentre os tipos de gráficos mais conhecidos existem o *boxplot*, histograma, gráfico de dispersão, ou de barras. O objetivo desse artigo é descrever um novo tipo de gráfico chamado *beanplot* que pode ser feito no aplicativo R. Através de exemplos é mostrado como fazer o *beanplot* no R e como interpretar seus resultados. Nesse gráfico podemos representar várias informações sobre variáveis quantitativas, tais como: média, mediana, distribuição dos dados, etc. Além disso, através desse gráfico podemos comparar distribuições de diversas variáveis ou da mesma variável em diferentes grupos.

Palavras-chave: *Beanplot; métodos gráficos; estatística descritiva; R.*

ABSTRACT

Descriptive analysis is a powerful tool to analyze data sets, but is rarely used. It can avoid many problems that can occur in more complex analyses, providing a picture of the sample under study. Some graphical methods are much more informative than tables. There are several types of graphics which are well known: boxplot, histogram, scatter plot, or bar plot. The aim of this paper is to present a new type of graph called beanplot, describing the steps to build the graphs using the statistical software R. Besides, some examples are presented to discuss how to interpret the results. Through beanplot graphs, it is possible to represent a plenty of information regarding quantitative variables, such as mean, median, distribution of data, etc. Moreover, through this graphic we compare distributions of several variables or the same variable in different groups.

Keywords: *Beanplot; graphical methods; descriptive analysis; R*

Rev HCPA 2010;30(2):185-191

A análise descritiva de um conjunto de dados é uma das ferramentas mais importantes da estatística, apesar de raramente ser utilizada. Quando bem feita, previne vários problemas que podem ocorrer em análises mais complexas, pois permite visualizar muitas características que em outras análises podem passar despercebidas.

Uma importante parte da estatística descritiva inclui os gráficos, que fornecem uma visualização de características dos dados. Gráficos como histogramas, ramos-e-folhas, gráficos de dispersão, de barras, ou *boxplots* são frequentemente usados para análises univariadas e para comparações de distribuições de dados. Entretanto, quando é necessário comparar diferentes grupos, o uso de gráficos como histogramas ou ramos-e-folhas acarreta o problema de ocupação de muito espaço para se atingir tal objetivo. Uma possibilidade para se fazer a comparação de diferentes distribuições é a utilização de *boxplots*, porém certos problemas podem surgir em casos particulares, e geralmente estão associados com interpretações inapropriadas e não com a técnica em si. Muitas vezes esses erros de interpretações vêm de analistas não-estatísticos que tentam ganhar mais informações do que o gráfico contém. Outro problema que pode surgir nos *boxplots* é na detecção de *outliers*, principalmente para distribuições não-normais. Mesmo que a distribuição seja normal a detecção de *outliers* pode se tornar difícil pois, à medida que aumenta o número de observações, pode aumentar o número de *outliers* detectados.

Para prevenir esses problemas foram desenvolvidas variações dos *boxplots* que incluem no gráfico informações adicionais disponíveis (1-3). Essas variações foram criadas para facilitar a interpretação e fornecer informações acerca dos dados que diminuíssem a possibilidade de erros nas interpretações. Porém, mesmo essas variações têm problemas, principalmente quando o objetivo é a comparação de distribuições de diferentes grupos. Em 2008, Kampstra sugeriu um novo gráfico chamado *beanplot*, que é a combinação de um gráfico de dispersão unidimensional com uma curva de densidade estimada. Em tal gráfico, não existe o problema de detecção de *outliers*, pois todas as observações ficam visíveis e a complicação que poderia surgir pelo uso do conceito de quartis também de-

1. Unidade de Bioestatística, Grupo de Pesquisa e Pós-Graduação, Hospital de Clínicas de Porto Alegre.
2. Departamento de Estatística, Instituto de Matemática, Universidade Federal do Rio Grande do Sul (UFRGS).
3. Programa de Pós-Graduação em Epidemiologia, Faculdade de Medicina, UFRGS.
4. Instituto Nacional de Ciência e Tecnologia para Avaliação de Tecnologias em Saúde (IATS)-CNPq/Brasil.
Contato: Suzi Camey. E-mail: camey@mat.ufrgs.br (Porto Alegre, RS, Brasil).

saparece, porque simplesmente a média é usada como medida para resumir os conjuntos de dados (4).

O objetivo desse artigo é descrever o *beanplot*, um novo gráfico que fornece várias informações sobre a distribuição de variáveis quantitativas.

O *BEANPLOT*

O nome *beanplot* vem de "vagem" (*green beans*). A curva de densidade estimada pode ser vista como uma vagem de feijão, enquanto o gráfico de dispersão mostra os grãos dentro da vagem. Os grãos representam as medidas individuais, como em um gráfico de dispersão unidimensional (1-d). O *beanplot,* portanto, combina uma curva de densidade com um gráfico de dispersão 1-d.

A forma da densidade usada é um polígono dado por uma curva de densidade estimada da Normal e sua versão espelhada. Frequentemente, esse polígono tem o formato semelhante ao de um violino, e também é usado no *violin plot*. No pacote *beanplot*, a densidade é estimada através da função *density*. Os detalhes sobre essa função podem ser obtidos com o comando *density*.

Enquanto o *boxplot* e suas variações expõem a mediana no gráfico, o *beanplot* pode exibir a média do conjunto. Essa é uma vantagem do *beanplot,* porque a média é uma medida mais fácil de interpretar. Quando mais de um grupo de sujeitos ou de variáveis são plotados simultaneamente, pode-se exibir também a média geral do conjunto de dados, outra vantagem sobre o *boxplot*.

Se as vagens representam um único grupo de sujeitos elas são simétricas, mas se o conjunto de dados que está sendo analisado tem dois subgrupos, por exemplo, homens e mulheres, é possível representá-los na mesma vagem. Dessa forma cada subgrupo será um dos lados de uma vagem tornando-a assimétrica na maioria dos casos.

Como instalar o R e o beanplot

O R (5) é um programa livre que pode ser obtido em http://www.r-project.org/. Após instalar e executar o R é necessário instalar a biblioteca *beanplot*. Esse processo consiste de **DUAS** etapas:

1. Instalar:

```
install.packages("beanplot", repos = "http://cran-r.c3sl.ufpr.br/")
```

2. Carregar:

```
library (beanplot)
```

Para executar essas duas etapas, basta copiar os comandos acima (em negrito) e colar no console do R. A Figura 1 ilustra essas etapas.

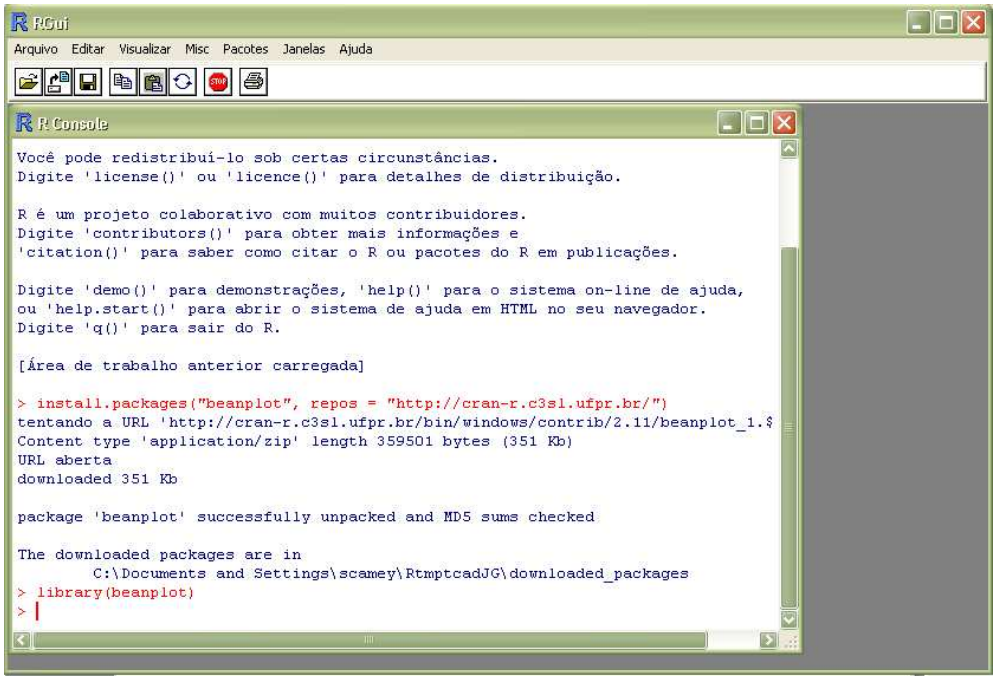

Figura 1 - Console do R após iniciá-lo com os comandos para instalar e carregar o `beanplot`.

Os comandos necessários para gerar cada gráfico serão descritos na próxima seção previamente a cada exemplo. Devemos lembrar que o R faz distinção entre letras maiúsculas de minúsculas, muitas vezes o comando não funciona porque foi digitado com letra maiúscula quando deveria ter sido digitado com letra minúscula e vice-versa. Os gráficos gerados na análise podem ser copiados e colados em documentos do Word ou PowerPoint.

EXEMPLOS DE CONSTRUÇÃO E INTERPRETAÇÃO DOS GRÁFICOS

Os dados utilizados nesse artigo são uma subamostra de 279 sujeitos oriundos da tese de doutorado "Medidas de Qualidade de Vida e Utilidade em uma Amostra da População Brasileira". Nesta tese o WHOQOL-Breve e o SF-36 foram aplicados a uma amostra aleatória da população geral de Porto Alegre. Os participantes eram pessoas alfabetizadas com idade entre 20 a 64 anos de idade (6).

Para ajudar a entender como o *beanplot* é construído, vamos começar com o gráfico da densidade normal estimada por kernel do Domínio 4 (Meio Ambiente) do WHOQL-Breve. A Figura 2 ilustra a densidade estimada a partir dos valores observados deste domínio na amostra de 279 sujeitos.

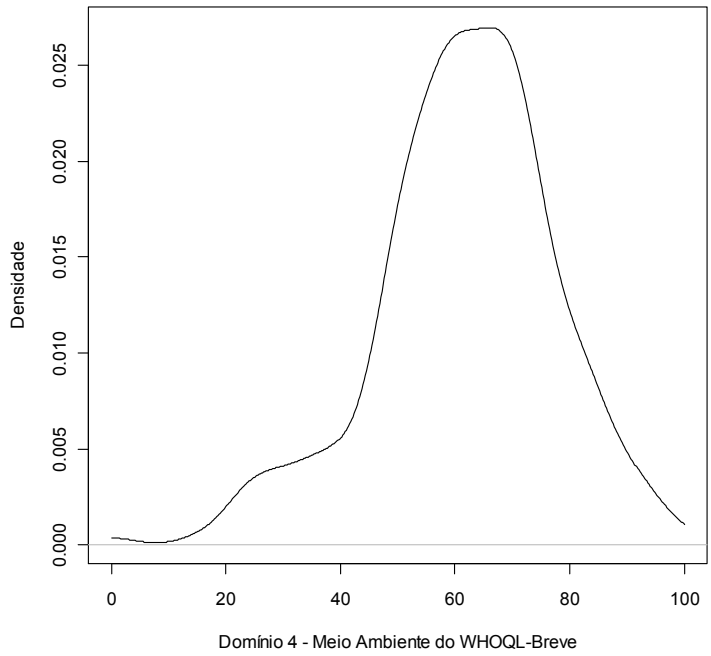

Figura 2 - Gráfico da distribuição estimada por kernel do Domínio 4 - Meio Ambiente do WHOQL-Breve.

O comando para gerar a Figura 2 é:

```
plot(density(ENVIR, from=0, to=100),main=" ", ylab="Densidade",
xlab="Domínio 4 - Meio Ambiente do WHOQL-Breve")
```

Detalhando o comando acima:

- função **`plot`** – desenha o gráfico;
- função **`density`** - estima a densidade;
- **`ENVIR`** - nome da variável Domínio 4 - Meio Ambiente no banco de dados;
- **`from=0, to=100`** – determina o valor mínimo e o valor máximo do escore;
- **`main`**, **`ylab`** e **`xlab`** - especificam os títulos: principal do gráfico, do eixo y e do eixo x, respectivamente.

A Figura 3 representa um *stripchart* do mesmo conjunto de dados. Esse gráfico nada mais é do que um gráfico de dispersão unidimensional, onde cada ponto representa a posição de um valor observado, sem considerar a frequência de cada valor observado. Ele é obtido através do seguinte comando:

```
stripchart(ENVIR,xlab="Domínio 4 - Meio Ambiente do WHOQL-Breve")
```

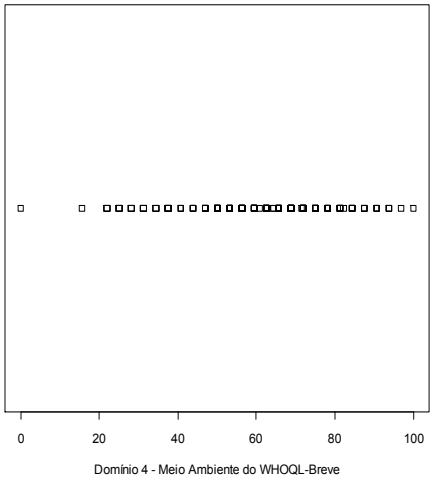

Figura 3 - *Stripchart* do Domínio 4 - Meio Ambiente do WHOQL-Breve

O *beanplot* utiliza esse gráfico de dispersão unidimensional com pequenas linhas, no lugar dos quadrados. Essas pequenas linhas são feitas de uma cor diferente da área da densidade, para garantir a visibilidade da forma da densidade.

Agregando as Figuras 2 e 3 temos então o *beanplot*, como vemos na Figura 4.

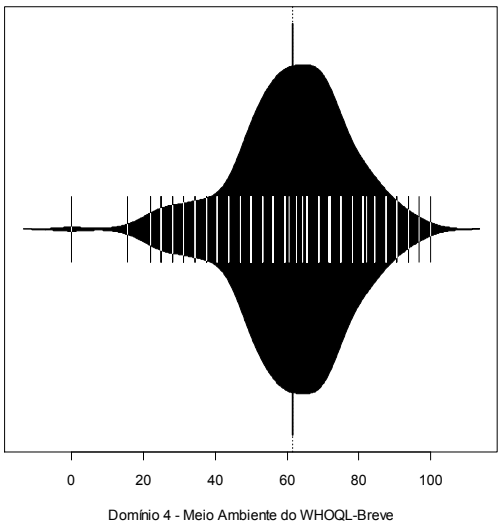

Figura 4 - *Beanplot* do Domínio 4 - Meio Ambiente do WHOQOL-Breve.

O gráfico é gerado a partir do comando:

```
beanplot(ENVIR, horizontal=T, xlab="Domínio 4 - Meio Ambiente do WHOQL-Breve",
method="overplot")
```

onde a opção **`horizontal=T`** determina que o gráfico seja feito na posição horizontal, essa opção não é a padrão, pois geralmente construímos o *beanplot* na posição vertical para facilitar a comparação entre variáveis os subgrupos de sujeitos. Já a opção **`method="overplot"`** mantém fixo o comprimento das linhas verticais independente do número de observações com o mesmo valor, ela também não é padrão na função ***`beanplot`***.

Na Figura 4 percebe-se que a densidade estimada vai além dos valores mínimos e máximos do Domínio 4. Isso acontece porque ela representa uma suavização do polígono gerado pela distribuição empírica dos dados. As linhas verticais mais finas representam as posições de cada uma das observações. A linha vertical mais grossa, nesse caso, é a média das observações que pode ser substituída pela mediana ou quintis.

A Figura 5 ilustra um *beanplot* para diversas variáveis simultaneamente.

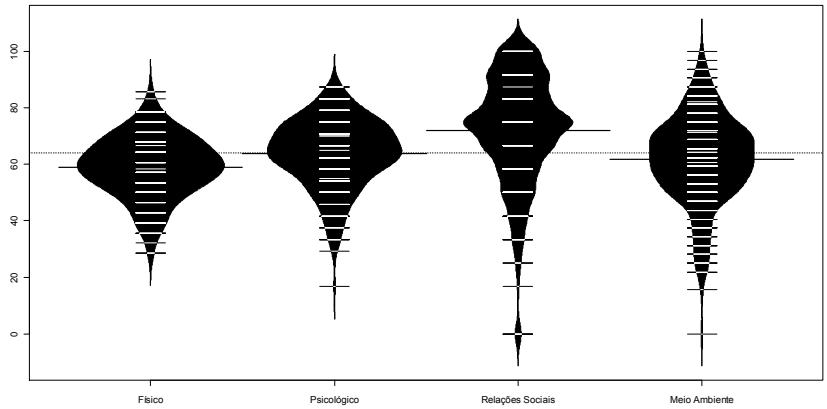

Figura 5 - Beanplot dos quatro domínios do WHOQL-Breve.

Esse gráfico foi gerado pelo seguinte comando:

```
beanplot(PHYS,PSYCH,SOCIAL,ENVIR,names=c("Físico","Psicológico","Relações So-
ciais","Meio Ambiente"),method="overplot")
```

Na Figura 5 as vagens estão na posição vertical e uma linha pontilhada que representa a média dos quatro domínios. A partir do gráfico podem-se fazer várias observações. Uma delas é que o Domínio 3 (Relações Sociais) tem um número menor de valores observados diferentes, pois há bem menos linhas no gráfico de dispersão 1-d e este é o domínio com maior média e dispersão. O Domínio 1 (Físico), por outro lado, é o que apresenta a menor média, uma das menores dispersões e a distribuição mais simétrica.

Como foi visto antes, pode-se utilizar o *beanplot* para comparar subgrupos como no exemplo da Figura 6.

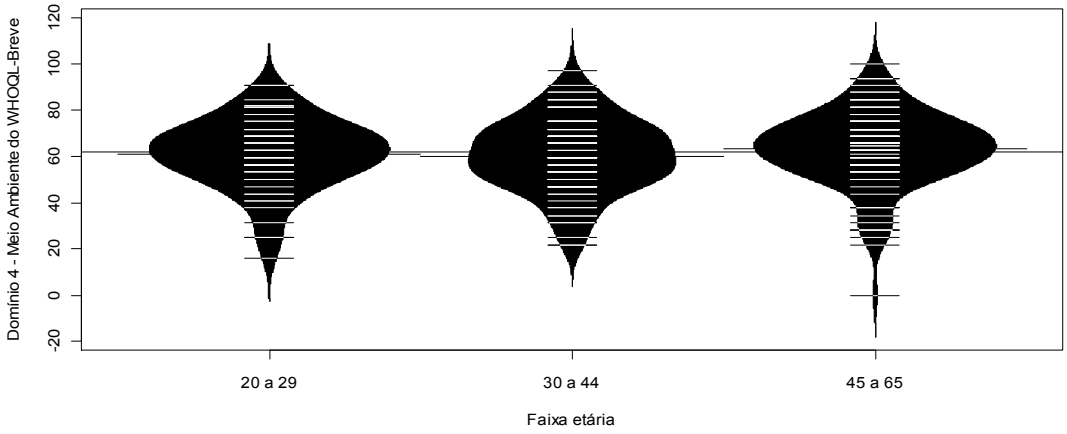

Figura 6 - Beanplot do Domínio 4 - Meio Ambiente do WHOQL-Breve por faixa etária.

Na figura 6 observa-se que quase não há diferença entre as faixas etárias. As distribuições são muito semelhantes nos três grupos. Para obter esse gráfico utiliza-se esse comando:

```
beanplot(ENVIR~idade_rec, xlab="Faixa etária",ylab="Domínio 4 - Meio Ambiente
do WHOQL-Breve",method="overplot")
```

A novidade nesse comando é **~idade_rec**, essa opção informa que a distribuição do Domínio 4 deve ser exibida por categoria de **idade_rec**, nome no banco de dados para a variável faixa etária.

Quando queremos comparar apenas dois subgrupos, uma opção é o *beanplot* assimétrico. Neste tipo de *beanplot*, cada lado da vagem representa um dos subgrupos. A Figura 7 mostra um exemplo onde queremos comparar homens e mulheres ao longo dos domínios do SF-36. Este gráfico é útil para mostrar que os homens apresentam médias maiores em todos os domínios do SF-36. No que se refere à dispersão dos dados, parece não haver diferença ente os grupos. Nele também se observa que, apesar de todos os domínios estarem em uma escala de varia entre 0 e 100, alguns deles, por exemplo, o Domínio Dor, apenas cinco valores são observados: 0, 25, 50, 75 e 100.

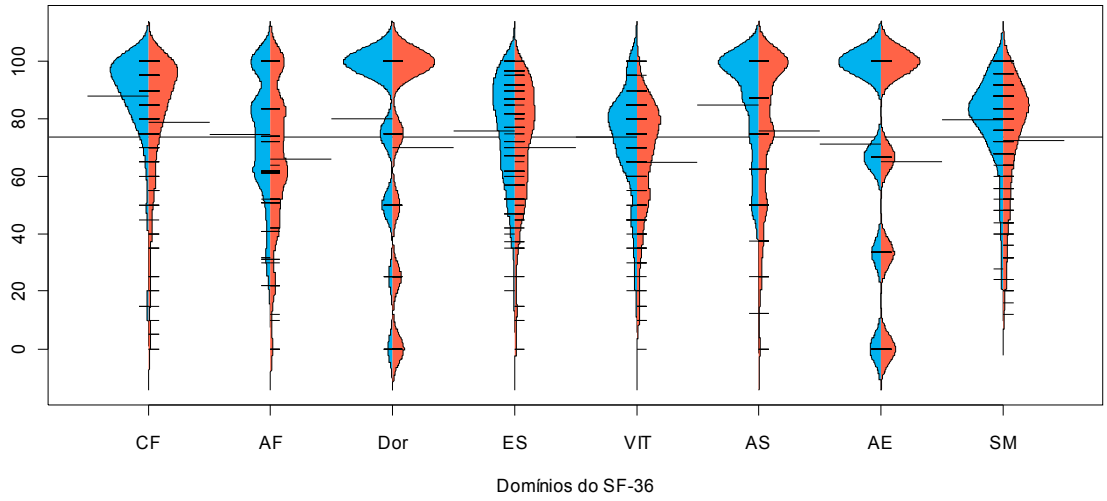

Figura 7 - Beanplot dos domínios do SF-36 por gênero. CF - Capacidade Funcional; AF - Aspectos físicos; ES - Estado geral de Saúde; VIT – Vitalidade; AS - Aspectos sociais; AE - Aspectos emocionais; SM - Saúde mental.

O comando para gerar a Figura 7 é:

```
beanplot(Score_Cap_func~sexo,Score_Dor~sexo,Score_Lim_Asp_Fis~sexo,
       Score_Est_saude~sexo, Score_vitalidade~sexo, Score_social~sexo,
       Score_emocional~sexo, Score_saude_mental~sexo,
        names=c("CF","AF","Dor","ES","VIT","AS","AE","SM"),
        method="overplot",side = "b",col = list("tomato", "deepskyblue2"),
        xlab="Domínios do SF-36")
```

Neste comando aparecem mais duas novidades:

- **side = "b"** – opção necessária para fazer os dois lados da vagem;
- **col = list("tomato", "deepskyblue2") –** opção para selecionar as cores de cada lado da vagem. O R tem inúmeras opções de cores, que podem ser escolhidas pelos nomes, como neste exemplo, ou simplesmente por números.

CONCLUSÃO

Este artigo descreve como fazer o gráfico *beanplot* no R. Este gráfico tem vantagens sobre outros já existentes na literatura, como, por exemplo, o *boxplot.* O *beanplot* é de fácil construção, fácil interpretação e não tem problemas na detecção de *outliers*, pois sua figura mostra todas as observações de um conjunto de dados.

Neste artigo foram mostradas algumas opções básicas do *beanplot*. O comando ? *beanplot* fornece todos detalhes sobre as alternativas do gráfico, como por exemplo, utilizar a mediana, omitir linha da média geral, mudar cores, etc.

REFERÊNCIAS

1. Hintze J, Nelson R. Violin Plots: a box plot-density trace synergism. Am Statistician. 1998;52(2):181-4.
2. McGill R, Tukey J, Larsen W. Variations of box plots. Am Statistician. 1978;32(1):12-6.
3. Box G. Statistics for experimenters: an introduction to design, data analysis, and model building. New York: Wiley; 1978.
4. Kampstra P. Beanplot: a boxplot alternative for visual comparison of distributions. J Statistical Software. 2008;28.
5. R Development Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; Available from: http://www.R-project.org/
6. Cruz L. Medidas de qualidade de vida e utilidade em uma amostra da população brasileira. 2010; Tese de Doutorado, UFRGS, Porto Alegre.

Recebido: 10/05/2010

Aceito: 04/07/2010