SGESTAT - Software para cálculos e gráficos estatísticos: uma ferramenta computacional para o apoio do ensino-aprendizagem de Estatística

Sidney C. Ferrari[1], Rosemeiry C. Prado[2], Fernando A. Paulo[3], Rodrigo M. Alexandre[4]

[1]Departamento de Ciência da Computação e Matemática Computacional – Instituto de Ciência da Computação e Matemática Computacional – ICMC – Universidade de São Paulo (USP) – Campus de São Carlos – SP – 13566-590 – São Carlos – SP - Brazil

[2]Programa de Pós-Graduação em Educação para a Ciência - UNESP – Universidade Estadual Paulista – Júlio de Mesquita Filho – Av. Eng. Luiz Edmundo Carrijo Coube, 14-01 - Vargem Limpa – Bauru – SP – Brazil

[3]Faculdade de Tecnologia do Estado de São Paulo – Fatec – Av. Vitalina Marcusso, 1400 – Campus Universitário – Ourinhos – SP - Brazil

[4]Faculdade de Tecnologia do Estado de São Paulo – Fatec – Av. Vitalina Marcusso, 1400 – Campus Universitário – Ourinhos – SP - Brazil

ferrarisc@gmail.com,rosecprado@zipmail.com.br, {fernando.paulo, rodrigo.alexandre}@fatec.sp.gov.br

Abstract. *The purpose of this study is to present a module for simple linear regression SGESTAT software, with the aim of providing a tool that encompasses much of the curriculum of undergraduate courses in Statistics, since statistical software for data manipulation and analysis can be large contribution to the process of learning content. Initially we tried to do a literature review in order to get a theoretical basis in conjunction with the study of the main tools for software development. The results presented by SGESTAT, as the scatter diagram and the regression model contributed to the analysis and interpretation of data and enabled a comparative study with the software R.*

Resumo. *A proposta deste trabalho é apresentar um módulo de regressão linear simples no software SGESTAT, com o objetivo de disponibilizar uma ferramenta que englobe grande parte do conteúdo programático de Estatística em cursos de graduação, visto que, softwares estatísticos para manipulação e análises de dados podem ser de grande contribuição no processo de aprendizagem de conteúdos. Inicialmente procurou-se fazer uma revisão literária, a fim de obter um embasamento teórico em conjunto com o estudo das principais ferramentas para o desenvolvimento do software. Os resultados apresentados pelo SGESTAT, como o diagrama de dispersão e o modelo de regressão contribuíram para as análises e interpretações dos dados e possibilitaram um estudo comparativo com o software R.*

1. Introdução

Apesar da importância do estudo de Estatística, o aprendizado dessa disciplina ainda passa por diversas dificuldades. Fernandez e Selau (1999); Mantovani e Viana (2004) e Gracio e Oliveira (2005) apud Cazorla (2008) destacam que, o aprendizado de estatística sofre de vários problemas, como a base matemática precária dos alunos, que em muitos casos é motivada pela falta de atitudes positivas em relação à Estatística e Matemática, e também a

ênfase exagerada em cálculos, linguagens e métodos estatísticos. Existem diversos *softwares* estatísticos no mercado, porém, a maioria é destinada ao âmbito comercial, necessitando adquirir licença para sua utilização, como por exemplo, o Minitab, SPSS e SAS. Os que não são comerciais precisam adquirir licença e em muitos casos, são de difícil utilização, como é o caso do *software* R, com uma interface pouco intuitiva ao usuário, ou seja, requer um conhecimento mais apurado do profissional da área.

A proposta deste trabalho é implementar um módulo de Regressão Linear Simples, no *software* SGESTAT. Pires (2001) deu início a primeira etapa do desenvolvimento do *software*, enfatizando a parte de planejamento, priorizando o levantamento e armazenamento de informações; na segunda etapa, houve a necessidade de tratar a estatística descritiva, que é a descrição dos dados, sejam eles de uma amostra ou população, organizando em tabelas e gráficos; na terceira e quarta partes foram implementadas as principais distribuições de probabilidade, discretas e contínuas, como a binomial, Poisson, normal, uniforme, gama, exponencial etc.; na quinta etapa foram implementados os testes de hipóteses para média e proporção em uma amostra. A motivação inicial foi dar prosseguimento aos projetos anteriores, acrescentando novas funcionalidades que possam ser utilizadas pelos alunos da disciplina de Estatística dos cursos de graduação.

Com base nessas considerações, o problema desse trabalho pode ser resumido na questão: o *software* SGESTAT possui todas as funcionalidades que abrangem grande parte do conteúdo de Estatística no ensino de graduação?

O projeto teve como objetivo implementar novas funcionalidades ao *software*, desenvolvido em projetos anteriores, como a Regressão Linear Simples, apresentar o diagrama de dispersão, calcular os coeficientes de correlação de Pearson (r) e de determinação (R^2) e, apresentar a reta da equação de regressão, a fim de auxiliar os alunos com uma ferramenta que possibilite comparar os resultados de exercícios feitos em aula, e obter suas respectivas análises simplificadas e, também permitir ao aluno formular exercícios similares, apenas substituindo os parâmetros de entrada.

2. Revisão Bibliográfica

Num primeiro momento, procurou-se fazer uma breve revisão da literatura relacionada ao tema em estudo, enfatizando os principais trabalhos da área, a fim de obter um embasamento teórico que contribuiu para o desenvolvimento do módulo de regressão linear simples no *software* SGESTAT. Em seguida, pesquisaram-se trabalhos correlatos, com o objetivo de contribuir, para desenvolver uma ferramenta que fosse útil aos estudantes de graduação que fazem a disciplina de Estatística. Para uma melhor compreensão, dividiu-se esta revisão em dois itens, cada um deles concentrando os seus aspectos fundamentais.

A regressão linear simples

Carvalho e Campos (2008) destacam que, existem cinco modelos de gráficos, ou seja, cinco resultados possíveis de apresentação do diagrama de dispersão. O primeiro diagrama é quando < 1, apresentando uma correlação positiva; o segundo diagrama é quando -1 0, apresentando uma correlação negativa; o terceiro diagrama é quando há uma correlação perfeita positiva, com 1; o quarto diagrama é quando há uma correlação perfeita negativa, com r 1; e, finalmente, a quinta e última forma de apresentar o diagrama de dispersão, quando 0, nesse caso há uma ausência de correlação.

O diagrama de dispersão sugere o tipo de regressão que melhor se ajusta aos dados analisados. Lapponi (2005) destaca que por meio da regressão linear simples, é possível determinar a melhor reta que explica essa relação entre a variável dependente e a variável

independente**.** A equação linear que representa o modelo de regressão linear simples é dada pela equação: Y

Onde:

= valor da variável dependente na i-ésima tentativa ou observação;

= primeiro parâmetro da equação de regressão, o qual indica o valor de Y quando X = 0;

x = segundo parâmetro da equação de regressão, chamado de coeficiente de regressão, que indica a inclinação da linha de regressão;

= erro de amostragem aleatória na i-ésima tentativa.

As fórmulas para calcular as estimativas de β, são representadas por b, sendo:

$\frac{\Sigma(x_i) - (\Sigma x_i).(\Sigma y_i)}{\Sigma x \quad (\Sigma x_i)^2}$ e $\frac{\Sigma y \quad \Sigma x}{}$

Os valores (x_1, y_1), (x_2, y_2), ..., $(x \quad)$ correspondem às amostras efetivamente observadas. Com os parâmetros definidos, pode-se determinar a seguinte equação de regressão: $\hat{y}$

Após a obtenção da equação da regressão linear necessita-se saber o quanto essa regressão se ajusta aos dados analisados. Essa precisão do ajuste é obtida pelo coeficiente de determinação que fornece a confiabilidade que teremos nas possíveis previsões que poderão ser feitas com o modelo.

Moore (2005) explica que a correlação *r* faz a descrição da intensidade de uma relação linear. Em se tratando de regressão linear, essa descrição pode ser feita de forma mais específica, ou seja, eleva-se ao quadrado a correlação *r* (r^2).

Portanto, o coeficiente de determinação é uma medida que serve para descrever a proporção da variável y que pode ser explicada por variações em x. A fórmula é apresentada a razão entre a variação explicada e a variação total:

$\frac{\Sigma(\hat{y} - \bar{y})}{\Sigma(y - \bar{y})}$ _____

Para Carvalho e Campos (2008), a correlação linear é simples por se tratar apenas de duas variáveis x e y, e para determinar a ocorrência de correlação linear simples, isto é, se há relação entre elas é preciso, de maneira genérica, que sejam respondidas duas perguntas: Existe alguma força unindo estas duas variáveis? Caso exista esta força, como se comporta uma variável em relação à outra?

Caso as respostas sejam afirmativas, haverá correlação entre as variáveis.

Carvalho e Campos (2008) destacam que, os valores do coeficiente de correlação estão sempre entre o intervalo de -1 e +1, onde o valor +1 indica que as variáveis X e Y têm uma correlação linear e positiva, isto é, todos os pontos dos dados em uma reta tem uma inclinação positiva. E quando o valor for -1 indica que as variáveis X e Y têm correlação linear e negativa, com todos os pontos em uma reta tendo uma inclinação negativa. E quando os valores do coeficiente de correlação ficarem próximos de zero indicam que as variáveis X e Y não se correlacionam linearmente.

"É importante ressaltar que o conceito de correlação refere-se a uma associação numérica entre duas variáveis, não implicando, relação de causa-e-efeito, ou mesmo uma estrutura com interesses práticos." (BARBETTA; REIS; BORNIA, 2009, p.316).

A utilização de ferramenta computacional no ensino-aprendizagem de Estatística

Analisando a literatura pesquisada foi possível fazer um levantamento das principais contribuições existentes, em relação a soluções, seja de ferramentas, ou métodos e recursos que fujam do ensino tradicional em sala de aula. Ferreira e Cymrot (2012), em seus estudos sobre o uso de *software* estatístico na área de engenharia elétrica, abordam a importância de se utilizar uma ferramenta estatística para manipular dados com mais precisão e rapidez. Um outro aspecto que as autoras destacaram, foi o fato de que é preciso saber escolher uma ferramenta adequada à área estudada, contribuindo para a formação dos futuros engenheiros, pois no ambiente de trabalho serão exigidos tais conhecimentos. Ainda segundo as autoras, a utilização de *softwares* estatísticos também proporciona um importante ganho educacional, por estimular métodos de simulação.

Em relação ao desenvolvimento dessas ferramentas voltadas à Estatística, para utilização de discentes e docentes nos cursos de graduação, são poucos os trabalhos relacionados. Rebelo (2004), propôs o desenvolvimento de um módulo no *software* SEstat.Net, na plataforma W*eb*, percebendo a necessidade de professores e alunos ter em uma alternativa de ensinar e aprender a análise de dados estatísticos, mais especificamente, regressão linear simples.

3. Metodologia

O projeto foi desenvolvido, após a fase de revisão bibliográfica, com o auxílio de várias ferramentas. Num primeiro momento, realizou-se a modelagem UML do sistema.

A UML (*Unified Modeling Language*) é uma linguagem utilizada para modelar sistemas orientados a objetos. Fowler (2005, p.25), "é uma família de notações gráficas, apoiada por um metamodelo único, que ajuda na descrição e no projeto de sistemas de software, particularmente daqueles construídos utilizando estilo orientado a objetos (OO)". Na modelagem do módulo de regressão linear simples foram utilizados os diagramas de classes e os casos de uso, a fim de facilitar a codificação.

Na segunda etapa foi criação do banco de dados, utilizando o SGBD MySQL, que é um Sistema Gerenciador de Banco de Dados Relacional. O MySQL 5.1.13 foi o SGBD escolhido nesse projeto por se tratar de uma ferramenta completa, possui licença de uso gratuita e de fácil utilização.

Na terceira etapa do desenvolvimento, foi a escolha da linguagem de programação. Neste caso, definiu-se o Java SE (*Desktop*), versão 7, como sendo a linguagem de programação para o desenvolvimento do módulo de regressão simples. Java é uma linguagem robusta, e independente de plataforma, ou seja, o mesmo código roda em sistemas operacionais diferentes. Alguns dos pontos fortes da linguagem são a manutenibilidade e reusabilidade de códigos, pelo fato de ser uma linguagem 100 % orientada a objetos.

A IDE para implementação dos códigos Java foi o NetBeans 8.0, por ser uma ferramenta gratuita e que atende bem as funcionalidades de escrever, compilar e debugar. É um ambiente de desenvolvimento, *open-source* escrito em Java.

Na quarta e última etapa realizaram-se os testes no sistema para validar os requisitos funcionais, para saber se o *software* é executado conforme foi especificado. Também foram realizados testes para o tratamento de erro do sistema, ou seja, como o *software* interage com o usuário no momento que surge o erro.

Vale ressaltar que a maior dificuldade encontrada no desenvolvimento foi a utilização da biblioteca JFreeChart do Java, que é responsável pela geração de gráficos.

4. Resultados e Discussão

A proposta do artigo foi de desenvolver uma ferramenta para que qualquer aluno do curso de graduação, independentemente do curso que estiver fazendo, e que esteja estudando a disciplina de Estatística, ou ainda, já estudou e precisa realizar análises estatísticas em trabalhos acadêmicos, como artigos e iniciação científica, por exemplo, possa utilizá-la sem maiores dificuldades.

A validação do SGESTAT foi por meio de um estudo comparativo com o R, que é um *software* livre (*open source*) onde há uma comunidade que faz as atualizações. O R é uma linguagem orientada a objetos criada em 1996 por Ross Ihaka e Robert Gentleman. Este *software* aliado a um ambiente integrado permite a manipulação de dados, realização de cálculos e geração de gráficos. Em relação à liberdade na utilização, Guesser (2009) destaca que o *software* livre, oferece uma possibilidade de ação muito mais abrangente aos seus usuários. Na comparação, observou-se o envolvimento da interação entre o usuário e o *software*, por meio de uma interface. Quanto mais intuitiva e simples for a interface, maior é a produtividade dos trabalhos realizados, melhor é a análise e interpretação dos resultados.

Um exemplo utilizando o SGESTAT

Para ilustrar a geração do diagrama de dispersão e sua reta de regressão, utilizaram-se alguns dados estatísticos de regressão linear simples, extraídos da literatura, com o objetivo de validar o *software* e mostrar a sua contribuição no aprendizado de Estatística, mais especificamente, em regressão linear simples, uma vez que a visualização da parte gráfica pode facilitar o entendimento dos resultados.

Exemplo 1 – Para examinar a relação de uma loja, medido em unidades de pés quadrados, e suas vendas anuais, foi selecionada uma amostra de 14 lojas. A Tabela 1 sintetiza os dados para essas 14 lojas.

Tabela 1 – Amostra de 14 lojas

Loja	Área em Pés Quadrados (Milhares)	Vendas Anuais (em milhões de dólares)	Loja	Área em Pés Quadrados (Milhares)	Vendas Anuais (em milhões de dólares)
1	1,7	3,7	**8**	1,1	2,7
2	1,6	3,9	**9**	3,2	5,5
3	2,8	6,7	**10**	1,5	2,9
4	5,6	9,5	**11**	5,2	10,7
5	1,3	3,4	**12**	4,6	7,6
6	2,2	5,6	**13**	5,8	11,8
7	1,3	3,7	**14**	3,0	4,1

Validando o SGESTAT na prática, primeiramente o aluno deve digitar os valores das variáveis X (área em pés quadrados – em milhares) e Y (vendas anuais – em milhões de dólares), que são os pares ordenados apresentados na tabela. O *software* exibe uma mensagem de alerta ao usuário, caso a quantidade de elementos digitados, tanto de X quanto de Y sejam diferentes, por ser tratar de valores que representam pares ordenados e retornando o foco do "cursor" na variável que estiver com a quantidade de elementos maior. Portanto, o total de elementos de cada variável deve ser igual. A funcionalidade do botão "adicionar" é dinâmica, ou seja, caso o aluno tenha digitado algum valor errado, seja em X ou em Y, basta apenas clicar no valor digitado erroneamente, feito isso o botão "adicionar" muda para o nome "alterar", deixando a aparência da tela mais "limpa".

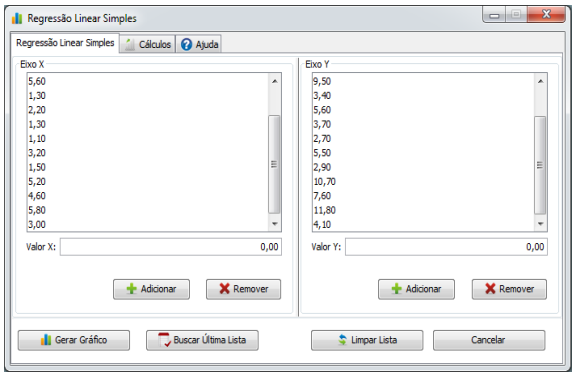

Figura 1 – Tela principal com os parâmetros de entrada.

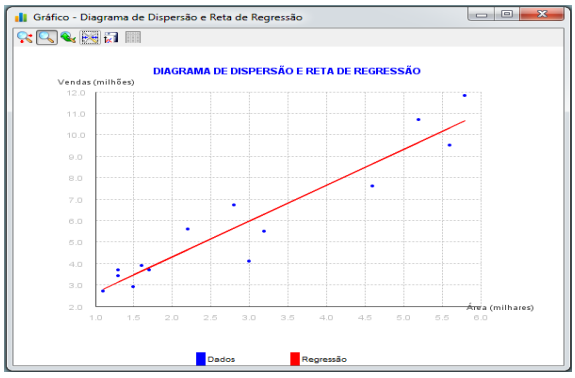

Figura 2 – Diagrama de dispersão e reta de regressão.

Com os dados digitados corretamente nas variáveis X e Y, o aluno pode dar seguimento a geração do diagrama de dispersão e a reta da equação. As figuras 1 e 2 apresentam, respectivamente, a tela principal do SGESTAT com os valores digitados e o gráfico gerado após o usuário ter clicado no botão "Gerar Gráfico".

Um exemplo utilizando o R

A título de ilustração foi utilizado o mesmo Exemplo 1 no *software* R. Com o R iniciado, encontra-se o símbolo " > " em vermelho, que é o *prompt* do R (também conhecido como R console), indicando que o R está pronto para receber seus comandos. Como mencionado anteriormente, o R é orientado a objetos, portanto, para utilizar as variáveis X e Y, que são os parâmetros de entrada, inicialmente, devem-se criar dois objetos, um que conterá o valor de X (Pés Quadrados - em milhares) e outro para Y (Vendas Anuais - em milhões de dólares). As figuras 3 e 4 demostram a representação dos objetos criados no *prompt* do R e o diagrama de dispersão com a sua reta de regressão:

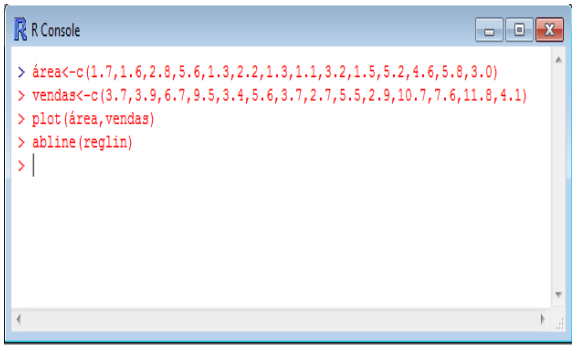

Figura 3 – Comando para gerar gráfico.

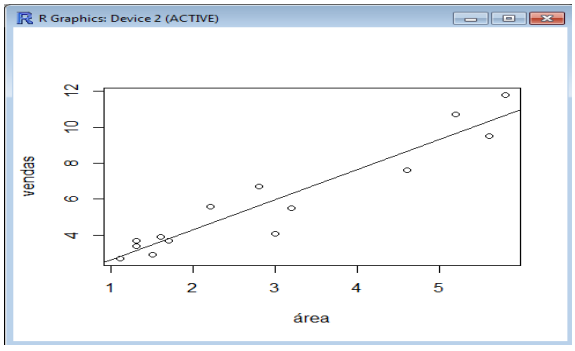

Figura 4 – Diagrama e sua reta de no R.

O comando “plot(área, vendas)” gera o diagrama de dispersão, e a reta de regressão é gerado pelo comando “abline(reglin)”.

Resultados do estudo comparativo entre o SGESTAT e o R

Uma das principais diferenças entre os dois *softwares*, é a forma como é apresentada o resultado final do modelo de regressão, isto é, a reta de regressão. No SGESTAT a equação é vem pronta para o aluno, $y = 0{,}9645 + 1{,}6699x$.

Por outro lado, o R não gera a equação, fornecendo apenas os valores dos coeficientes “a” e “b”, nesse caso o aluno deverá escrever a equação com esses coeficientes. A forma como é apresenta o modelo de regressão nos dois *softwares* é demonstrada nas figuras 5 e 6:

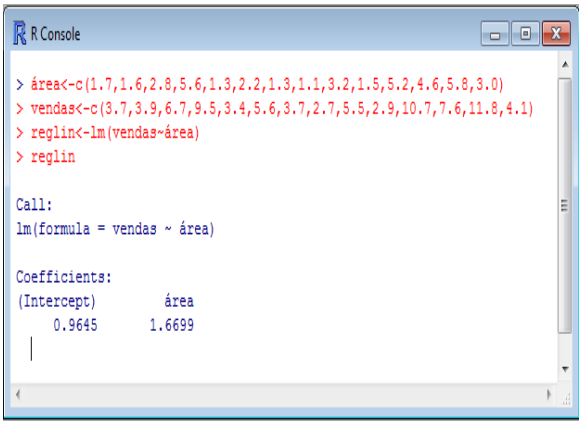

Figura 5 – Modelo de regressão do R.

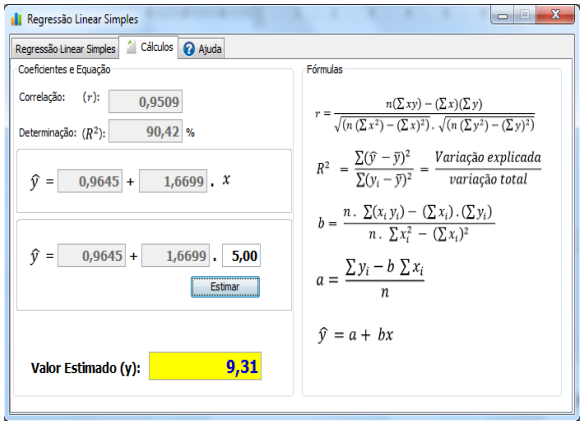

Figura 6 – Modelo de regressão do SGESTAT.

Sob a ótica da utilização do sistema, o SGESTAT cumpriu com seu propósito de ser uma ferramenta para auxílio no ensino-aprendizado da disciplina de Estatística, e que tivesse uma interface gráfica intuitiva para facilitar o uso por parte dos alunos. No caso do R, apesar de ser um *software* com uma grande variedade de recursos e bibliotecas, não possui uma interface gráfica amigável, pois a interação do aluno com a ferramenta ocorre por meio de comandos digitados no *prompt.* Um ponto que é importante ressaltar, apesar do R não exigir que o aluno saiba programar, ele requer um mínimo de conhecimento de orientação a objetos.

5. Considerações finais

Este artigo considera a implementação do módulo de regressão linear simples, com objetivo de contribuir para que o *software* SGESTAT, desenvolvido em projetos anteriores, englobe grande parte do conteúdo programático de Estatística, que é ensinado na maioria dos cursos de graduação. Para demonstrar a utilização do *software*, realizou-se um estudo comparativo com o *software* estatístico R, amplamente utilizado no meio acadêmico. O principal resultado obtido foi a forma simples como é realizada a entrada dos parâmetros obrigatórios, no que tange a regressão linear simples, que são os valores da variável independente X e a variável dependente Y e como resultado do modelo de equação de regressão, utilizando o módulo implementado no *software* SGESTAT.

Logo, o desenvolvimento do módulo de regressão linear simples, no *software* SGESTAT, possibilitou aos alunos usufruírem de uma ferramenta que contempla os recursos necessários para manipular e analisar dados estatísticos sem precisar adquirir licenças de uso, versões limitadas para testes, ou ainda, *softwares* livres com pouca interação, por não possuir uma interface gráfica amigável aos usuários.

Como propostas de trabalhos futuros, essa pesquisa pode ser ampliada, com o desenvolvimento de outros módulos de regressão, uma vez que este trabalho restringe-se à regressão linear simples. Sugere-se criar o módulo que possibilite trabalhar com mais de uma variável independente (X), ou seja, regressão linear múltipla. Outras regressões como polinomiais e quadráticas podem ser incrementadas ao *software* SGESTAT, tornando-o uma ferramenta ainda mais completa, em relação à contribuição no ensino-aprendizagem de Estatística nos cursos de graduação.

Referências

BARBETTA, P. A; Reis, M. M; Bornia, A. C. Estatística: para curso de engenharia e informática. 2 ed São Paulo: Atlas, 2009.

CARVALHO, S; CAMPOS, W. Estatística básica e simplificada. Rio de Janeiro: Elsevier, 2008.

CAZORLA, I. M. O ensino de estatística no Brasil. Sociedade brasileira de educação matemática. Disponível em:<http://www.sbem.com.br/gt_12/arquivos/cazorla.htm >. Acesso em: 06 nov. 2013.

FERNANDEZ, D. e SELAU, L.P.R. Cadeias de Markov aplicado ao curso de Estatística. Anais da Conferência Internacional: Experiências e Perspectivas do Ensino de Estatística – Desafios para o século XXI. Florianópolis, 20, 21 e 22 de setembro de 1999.

FERREIRA, D. S; CYMROT. R. Uso do software R no tratamento estatístico de dados de Engenharia. In: CONGRESSO DE INICIAÇÃO CIENTÍFICA DO INATEL, 22, 2012, Santa Rita do Sapucaí. Santa Rita do Sapucaí: INCITEL, 2012. 07p. 251-257p.

FOWLER, M. UML essencial: um breve guia para a linguagem-padrão de modelagem de objetos. 3 ed. Porto Alegre: Bookman, 2005. p. 25-52.

GRACIO, M. C. C. e OLIVEIRA, E. F. T. O ensino de Estatística na UNESP/Campus de Marília. Educação Matemática em Revista, Ano 11, v. 17, 9-15.

GUESSER, A. H. Software livre & controvérsias tecnocientíficas. Curitiba: Juruá, 2009. p. 40-58.

LAPPONI, C. J. Estatística usando Excel. Rio de Janeiro: Elsevier, 2005. p. 393-418.

MONTOVANI, D. M. N. e VIANA, A. B. N. Ensino de Estatística para cursos de graduação em Administração de Empresas – Novas Perspectivas. VII Seminário de Administração FEA-USP. 2004.

MOORE, D. S. A estatística básica e sua prática. Rio de Janeiro: LTC, 2005.

PIRES, P. C. Software para cálculos e gráficos estatísticos. 2001. ??f. Graduação (Tecnólogo em Processamento de Dados) – Faculdade de Tecnologia do Estado de São Paulo, Ourinhos, 2001.

REBELO, R. A. Planejamento de uma ferramenta computacional de ensino-aprendizagem na análise de regressão, 2004. 130f. Dissertação (Mestrado em Ciências da Computação) - Universidade Federal de Santa Catarina, Florianópolis, 2004.