Exploração de dados baseada em Estatística Simples e Regressões

Luiz Paulo Cravo Júnior[1], Ramon Rotatori Viana[2], Gabriel Rodrigues Hickel[3]

[1,2,3]Universidade do Vale do Paraíba/Faculdade de Ciência da Computação, Avenida Shishima Hifumi, 2911, Urbanova, luizpaulo@univap.br, ramon_sjc@yahoo.com.br, hickel@univap.br

Resumo - Este artigo descreve a criação de ferramentas para exploração de dados com métodos que executam cálculos estatísticos de forma a orientar e dar uma melhor visão acerca dos dados contidos na base de dados do Observatório Virtual da Univap. Os cálculos estatísticos vão desde os mais simples até aos mais complexos, chegando aos cálculos realizados que utilizam o princípio da regressão linear, podendo, assim, calcular estatísticas de dados futuros baseadas nas informações já contidas e armazenadas no Banco de Dados atual. A linguagem de programação, PHP (Processador de hipertextos, do inglês, Hypertext Preprocessor), foi utilizada para a implementação dos cálculos abordados neste Projeto.

Palavras-chave: PHP, Estatística, Regressões, Estrelas, Observatório Virtual
Área do Conhecimento: Ciências Exatas e da Terra

Introdução

Do latim *statisticum collegium*, o termo estatística surgiu e teve sua definição estabelecida no século XIX pela Enciclopédia Britânica. A palavra foi proposta pela primeira vez no século XVII, na Universidade de Lena e adotada por acadêmicos nesta época. O termo é utilizado para definir uma análise de dados e aplicado de forma matemática sobre aquilo que queremos observar.

A estatística foi implantada com a finalidade de se obter um consenso sobre o que as observações analisadas nos dizem sobre o mundo que observamos, bem como auxiliar uma tomada de decisões baseada em dados. Também é utilizada sobre as variações de amostras analisadas a fim de estimar valores de probabilidades de variações, erros ou valores previstos.

Os cálculos estatísticos, neste caso, serão utilizados para verificar as variações, diferenças e comparar dados do Observatório Virtual da Univap a fim de fazer comparações entre os dados que serão analisados bem como estimar valores e dados futuros para análises e pesquisas relevantes na área.

Serão aplicados os conhecimentos na linguagem de programação PHP (Processador de Hipertextos, do inglês, Hypertext Preprocessor) a fim de calcular os dados estatísticos do Observatório Virtual da Univap, com o objetivo de informar ao usuário do Sistema, os dados previamente selecionados para análise.

Os cálculos estatísticos, neste caso, serão utilizados para verificar as variações, diferenças e comparar dados do Observatório Virtual da Univap a fim de fazer comparações entre os dados que serão analisados bem como estimar valores e dados futuros para análises e pesquisas relevantes na área.

Inicialmente, os cálculos estatísticos foram utilizados em Universidades com Professores e Mestres a fim de estudar certas amostragens de dados através de um único resultado que pudesse representar o conjunto de dados ou sua variação, entre outros.

Materiais e Métodos

A disciplina Inicialmente, os cálculos estatísticos foram utilizados em Universidades com Professores e Mestres a fim de estudar certas amostragens de dados através de um único resultado que pudesse representar o conjunto de dados ou sua variação, entre outros.

Para se chegar ao objetivo proposto durante o desenvolvimento do Módulo de Exploração, o qual foi dividido em três etapas para sua elaboração e implementação, seguiu-se as seguintes etapas:

- 1ª etapa:- Realizar estudos acerca dos cálculos

  Nesta primeira etapa, foram realizados várias pesquisas procurando conhecer as fórmulas, métodos e limitações dos cálculos.

- 2ª etapa:- Implementação do Sistema
  Nesta etapa, deu-se início a parte de programação, onde os estudos realizados na primeira etapa foram de suma importância e possibilitaram esta etapa, de desenvolvimento.

- 3ª etapa:- Testes
  A terceira etapa foi a etapa em que foram realizados os testes dos módulos programados e onde foram corrigidas eventuais e pequenas falhas que ainda ocorriam no sistema.

PHP – Processador de Hipertextos

O A linguagem de programação utilizada foi o PHP. O PHP surgiu em meados de 1994 e foi inicialmente utilizado como um pacote de ferramentas para uma página web. O PHP é uma linguagem de criação de scripts embutida em HTML no servidor. Podemos pensar no PHP como "uma coleção de supertags de HTML que permitem adicionar funções do servidor às suas páginas da Web".

O PHP é um módulo oficial do servidor http Apache. Com ele, é possível coletar dados de um formulário, gerar páginas dinamicamente ou enviar e receber cookies, gerando páginas da internet dinâmicas, possibilitando maior interação com o usuário. O código PHP fica "escondido" nas páginas dos usuários, já que os resultados das páginas PHP é o HTML puro, que é exibido para o usuário na tela da página da internet.

MySQL & SQL – Linguagem de Consulta Estruturada

Como foi utilizada uma grande quantidade de informações no banco de dados do Observatório Virtual da Univap, foi preciso um Servidor MySQL, pelo fato de que ele pode ser usado em sistemas de produção com alta carga e missão crítica ou pode ser embutido em programa de uso em massa, como neste caso, onde serão poderão haver múltiplos usuário acessando os dados do sistema para realizar consultas, estimar valores e outros.

O servidor de banco de dados MySQL é extremamente rápido, confiável, e fácil de usar. Por ser um aplicativo de licença *Open Source*, os usuários podem contribuir com falhas de segurança e outras, o que faz com que ele seja um excelente servidor de banco de dados.

Os Cálculos

A média, que pode ser definida como sendo o quociente da soma dos calores de uma característica quantitativa pela quantidade total de valores. A média nada mais é do que a soma dos elementos dividido pela quantidade de elementos. A fórmula que define a média é:

$$\overline{x} = \frac{\sum_{i=1}^{n}(X)}{n}$$

A moda é definida como sendo a característica quantitativa discreta de freqüência mais elevada, ou seja, a moda é o valor que mais se repete dentro de um conjunto de elementos ou de dados analisados. A moda representa o valor mais "bem cotado" dentro de uma amostra. Para calculá-la, não existem fórmulas, mas sim métodos para realizarem as buscar e verificação de quais e quantas vezes determinados elementos se repetem a fim de serem classificados como sendo a moda de uma quantidade amostral de elementos.

A mediana é definida como sendo o número real que divide em duas partes de igual freqüência os elementos de uma distribuição estatística. A mediana nada mais é do que o valor central que divide um conjunto em duas metades iguais, possuindo, assim como a moda, métodos que buscam seu resultado, não determinado por fórmulas pré-estabelecidas.

O Desvio Médio é o desvio absoluto (diferença) dos valores analisados (individualmente) em relação à mediana ou a outro tipo de média. O Desvio Médio é o resultado da soma de todos os elementos subtraídos da média e depois de somados, divididos pelo número de elementos. Defini-se o Desvio Médio como:

$$D.M. = \frac{\sum_{i=1}^{n}|X_i - X|}{n}$$

A variância é definida como a média aritmética dos desvios em relação a média. A variância é calculada como sendo o resultado do quadrado da soma de todos os elementos subtraídos da média e depois de somados, divididos pelo número de elementos. É representado desta forma:

$$Var. = \frac{\sum_{i=1}^{n}(X_i - X)^2}{n}$$

O Desvio Padrão é definido como sendo a raiz quadrada do resultado do quadrado da soma de todos os elementos subtraídos da média e depois de somados, divididos pelo número de elementos, ou seja, a raiz quadrada da variância. É representada como:

$$D.P. = \sqrt{\frac{\sum_{i=1}^{n}(X_i - X)^2}{n}}$$

A Amplitude é definida como sendo o tamanho total da amostragem analisada. É calculada subtraindo-se do maior, o menor dos valores

analisados, resultando no tamanho amostral, ou seja, é o comprimento da amostra analisada. Pode ser representada como:

$$\text{Amp.} = \text{maior(x)} - \text{menor(x)}$$

O Teste-t de Student é utilizado para testar a igualdade entre as médias de dois conjuntos de números, isto é, ele verifica se um conjunto de números é significativamente maior ou menor que o outro. É definido como:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Já a Análise de Variância, divide a variabilidade em variabilidade Entre Grupos e variabilidade Dentro de Grupos, e compara as duas. Quanto maior for a primeira comparada à segunda, maior é a evidência de que existe variabilidade entre os grupos analisados, ou seja, que possuem médias diferentes.

$$F = \frac{\overline{SS_b}}{\overline{SS_w}}$$

O cálculo realizado com regressões é muito utilizado para estimar dados futuros e pesquisar dados não calculados, obtendo-se um valor estimado e bem próximo do valor real, através de análises realizadas com valores da amostra em questão. Para realizar tais cálculos, são realizados cálculos desta forma:

$$Estimação = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} * (índice\ relacionado) + \bar{y} - slope \cdot \bar{x}$$

Resultados

O Sistema implementado, chamado Exploração de Dados baseada em Estatística simples e regressões, para ser utilizado dentro do Observatório Virtual da Univap, tem como principal objetivo informar ao usuário, dados das tabelas do Banco de Dados do Observatório Virtual da Univap de acordo com as opções que ele escolher. O Observatório conta com vários outros módulos que realizam desde cálculos até consultas a banco de dados externos.

Para utilizar o módulo da exploração de dados baseada em estatística, o usuário tem, obrigatoriamente, de seguir o fluxograma:

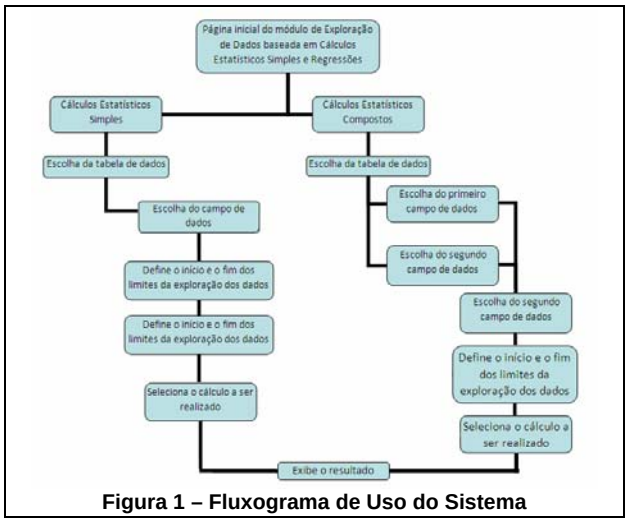

Figura 1 – Fluxograma de Uso do Sistema

A tela principal deste módulo desenvolvido pode ser observada na figura 1 abaixo.

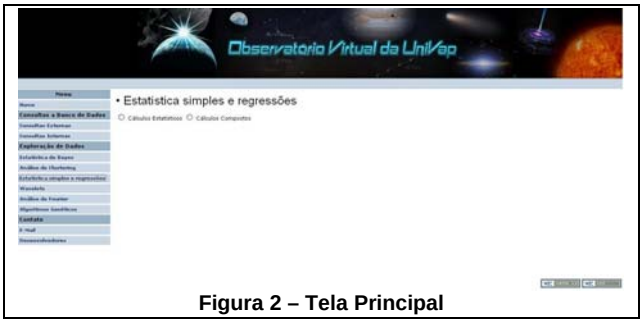

Figura 2 – Tela Principal

Após selecionado o tipo de cálculos que se deseja realizar (Cálculos Estatísticos Simples ou Compostos), o usuário verá a tela 2.a caso tenha selecionado a primeira opção e a tela 2.b caso tenha selecionado a segunda opção.

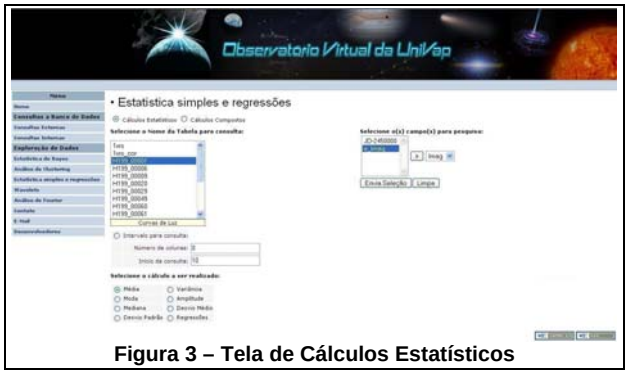

Figura 3 – Tela de Cálculos Estatísticos

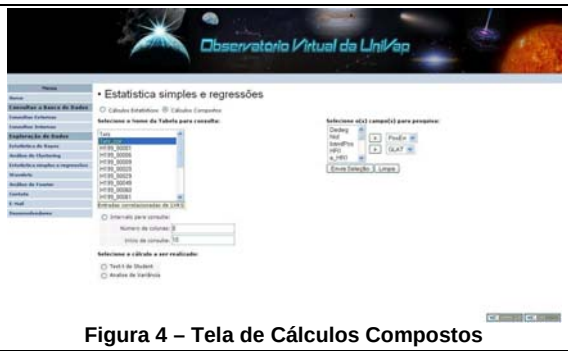

Figura 4 – Tela de Cálculos Compostos

Discussão

O software atende a todas as especificações do usuário e seguiu às três etapas do projeto definidas pela equipe de desenvolvimento no início da disciplina.

Com este módulo do Sistema, o usuário pode fazer consultas no banco de dados, realizar cálculos com base nos dados do banco, comparar dados de tabelas diferentes, e estimar probabilidades.

A linguagem utilizada para o desenvolvimento, PHP, não deixa nada a desejar. Através dele, o sistema é capaz de realizar uma interação com o usuário e dar resultados às consultas realizadas em pouquíssimo tempo.

Conclusão

O trabalho final, resultado de um trabalho de toda a sala de aula teve um bom resultado tendo em vista a complexidade dos "mini-sistemas" que foram desenvolvidos com o objetivo de criar um só sistema com dados, informações e análises em um único local.

A implementação do sistema em PHP foi de grande importância para frisar a importância de disciplinas já cursadas na Faculdade e também para que se visse a importância deste veículo de comunicação tão explorado, que é a internet.

Os conhecimentos a cerca das metodologias utilizadas no projeto puderam ser otimizados. Algumas técnicas de lógica de programação que tiveram de ser utilizadas no projetos e outras que tiveram de ser pesquisadas, serão de vital importância na elaboração de futuros projetos.

Para uma próxima versão, este projeto poderia ter algumas opções a mais quanto às consultas dos dados, a origem dos dados para a realização dos cálculos passando a ter, por exemplo, interação de todas as tabelas em um único cálculo ou de todos os dados. Os métodos para realização dos cálculos também podem mais aprimorados para resultados mais precisos e rápidos.

Referências

-LAROUSSE Cultural, Grande Enciclopédia. Círculo do Livro. Editora Nova Cultural. 1987, Edição Integral, volumes 2, 20, 25, 29 e 30.

-SPIEGEL, Murray R. Estatística. Coleção Shaum, Editora McGraw-Hill do Brasil 1977, 10ª reimpressão.

-MÉDICA, INFORMÁTICA. Departamento de. Faculdade de Medicina da Universidade de São Paulo. Apostila Análise de Variância, MPT 164, edição 2003. Disponível em http://www.dim.fm.usp.br/variancia/index.php. Acesso em 25 abr. 2007.

-PHP, Documentation Group. Free Software Foundation. Licença Pública Geral (GNU). Copyright 1997 - 2003 para o PHP Documentation Group. Disponível em http://www.opencontent.org/openpub/. Acesso em 03 mar. 2007.

-Biometria – Bioestatística - Teste t para dados emparelhados. FDL (Free Documentation License). Leitura Complementar ao Capítulo 5. Disponível em http://www.cultura.ufpa.br/dicas/biome/biomed.htm . Acesso em 28 mar. 2007.

-MÉDICA, INFORMÁTICA. Departamento de. Faculdade de Medicina da Universidade de São Paulo. Apostila Análise de Variância, MPT 164, edição 2003. Disponível em http://www.dim.fm.usp.br/testet/index.php. Acesso em 28 mar. 2007.

-R, Introdução ao. Tutorial do R (Programa para Cálculos Estatísticos). Universidade Federal do Paraná. Disponível em http://leg.ufpr.br/Rtutorial/stats1.html. Acesso em 25 abr. 2007.

-MYSQL, Manual de Referência do. Tradução do "MySQL Reference Manual". Revisão 79, de 03 de set. 2006. Disponível (original) em http://dev.mysql.com/doc/mysql/en. Acesso em 28 fev. 2007.

-SIQUEIRA, Bruno Rodrigues. Curso de PHP. Escola iMasters. Criado em 14 de jul. 2002. Disponível em HTTP://www.imasters.com.br. Acesso em 20 nov. 2005.

-Teknomo, Kardi. Spreadsheet Example of Linear Regression. Microsoft Excel Tutorials. Disponível em http://people.revoledu.com/kardi/tutorial/Regression/RegresionExample.html. Acesso em 21 mai. 2007.