



Matemática Multimídia

ANÁLISE DE DADOS
E PROBABILIDADE



GUIA DO PROFESSOR



Software

Medidas do corpo – gráficos de dispersão

Objetivos da unidade

1. Analisar representação gráfica de dados estatísticos;
2. Familiarizar o aluno com gráfico de dispersão e análise estatística bivariada;
3. Utilizar o conceito de correlação linear.

REQUISITOS DE SOFTWARE Navegador moderno (Internet Explorer 7.0+ ou Firefox 3.0+), Adobe Flash Player 9.0+ e máquina Java 1.5+.

RESTRIÇÕES DE ACESSIBILIDADE Este software não possui recurso nativo de alto contraste nem possibilita navegação plena por teclado.

LICENÇA Esta obra está licenciada sob uma licença Creative Commons



UNICAMP



FUNDO NACIONAL
DE DESENVOLVIMENTO
DA EDUCAÇÃO

Secretaria de
Educação a Distância

Ministério da
Ciência e Tecnologia

Ministério
da Educação



Medidas do corpo – gráficos de dispersão

GUIA DO PROFESSOR

Sinopse

Neste software, o aluno irá estudar análise exploratória de dados para duas variáveis: número do calçado e altura. A relação entre essas duas variáveis quantitativas será analisada através do chamado gráfico de dispersão e do coeficiente de correlação linear.

Conteúdos

- Estatística, Interpretação de Gráficos e Dados;
- Gráficos bivariados para duas variáveis quantitativas: gráfico de dispersão;
- Coeficiente de correlação linear.

Objetivos

1. Analisar representação gráfica de dados estatísticos;
2. Familiarizar o aluno com gráfico de dispersão e análise estatística bivariada;
3. Utilizar o conceito de correlação linear.

Duração

Uma aula dupla.

Recomendação de uso

Sugerimos que os dados a serem utilizados neste software sejam previamente coletados e registrados em uma tabela pelos alunos.

Material relacionado

- Softwares: Medidas do Corpo – Gráficos Univariados; Medidas do Corpo – BoxPlot;
- Vídeos: Ação, Reação, Correlação; Exames de Gabriela; Expresso Lanches;
- Experimento: Variáveis Antropométricas.



Introdução

A pesquisa científica é um processo de aprendizagem. Neste contexto, os métodos estatísticos entregam ferramentas matemáticas que permitem otimizar esse processo.

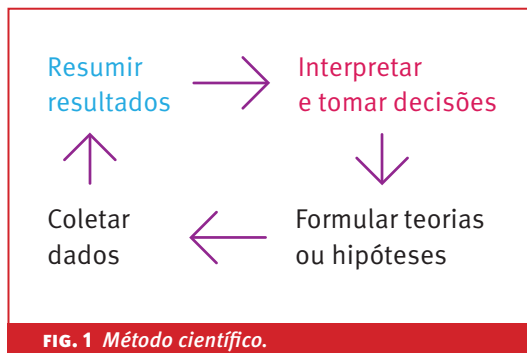
Basicamente, podemos identificar os seguintes estágios de uma pesquisa científica:

1. Formulação de uma hipótese, que terá certas consequências;
2. Amostragem ou coleta de dados;
3. Resumo, representação gráfica e comparação dos dados obtidos com o que seria de se esperar de acordo com a hipótese estabelecida;
4. Aceitação ou rejeição da hipótese. No caso de rejeição, uma nova hipótese é formulada. No caso de aceitação, a hipótese é mantida até que novas amostras determinem sua rejeição.

Essas etapas formam um ciclo iterativo entre o avanço teórico (hipótese) e os procedimentos de obtenção de dados.

A formulação das hipóteses está relacionada ao levantamento de possíveis respostas para um problema específico.

Ao coletar dados, estamos interessados em obter informação que permita manter a validade de uma hipótese ou que entregue evidências suficientes para a sua rejeição e consequente formulação de novas hipóteses, que serão testadas com uma nova coleta de dados e assim por diante.



O software

Estrutura do software

Neste software, pretendemos abordar uma possível representação gráfica para a análise da relação entre duas variáveis quantitativas, como descrito brevemente na Introdução do software e na introdução da ATIVIDADE 1, parte 2.

Na ATIVIDADE 1, o aluno deverá fornecer os dados que serão utilizados ao longo do software. Caso outro software desta mesma sequência já tenha sido utilizado no mesmo computador, os dados serão carregados automaticamente. O experimento Variáveis Antropométricas também utiliza um conjunto de dados semelhante.

Na parte 2 da ATIVIDADE 1, o aluno analisará um gráfico de dispersão para as variáveis Índice de Massa Corpórea (IMC) e Índice de Desenvolvimento Humano (IDH). Os dados são referentes a uma amostra de adolescentes, com idades entre 10 e 18 anos, matriculados na rede pública de ensino de 93 municípios do estado do Paraná.

Na ATIVIDADE 2, o aluno deverá construir o gráfico de dispersão para as variáveis “número do calçado” e “altura”, com os dados coletados na classe, analisando graficamente a existência de associação entre as variáveis, de acordo com os valores obtidos na amostra.

A existência de associação pode ser quantificada com a medida-resumo coeficiente de correlação linear, obtido a partir das médias e desvios-padrão de cada uma das variáveis. O aluno deverá ser capaz de obter e interpretar o valor desse coeficiente.



TELA 1 *Mapa do software.*













1 Variáveis quantitativas

ATIVIDADE

Começamos esta atividade registrando os dados obtidos na classe, referentes às variáveis: “gênero”, “número de calçado” e “altura”. O aluno deve preencher a tabela com os dados obtidos.

A variável qualitativa “gênero” tem como possível resposta duas categorias não ordenadas: “F”, feminino, e “M”, masculino. A variável quantitativa “altura” deve ser registrada em cm, e a variável quantitativa “número de calçado” será um número inteiro não negativo.



Remover linha	Gênero	Número do calçado	Altura (cm)	
	F	31	145	✓
	F	36	150	✓
	F	32	143	✓
	F	34	144	✓
	F	33	143	✓
	F	33	142	✓
	M	35	150	✓
	M	36	153	✓
	M	36	153	✓
	M	37	152	✓
	F	34	145	✓
	F	35	151	✓

Linhas válidas: 25

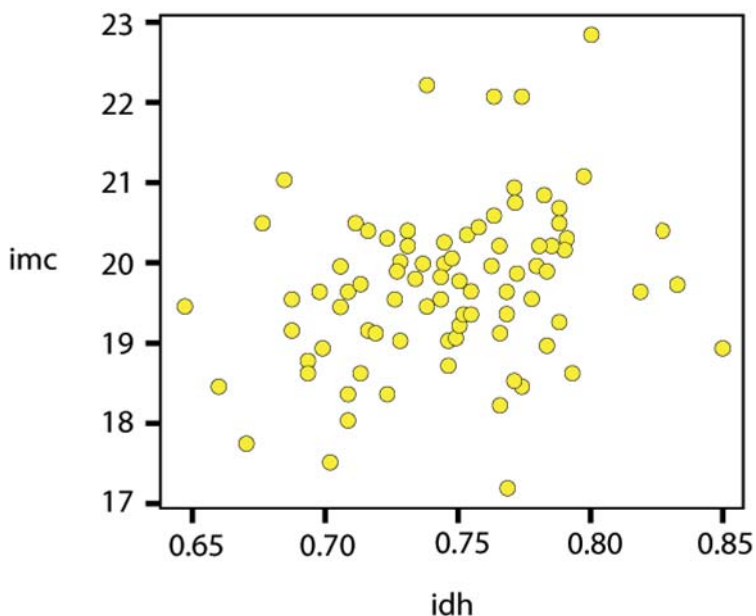
Adicionar 5 linhas

Salvar dados

Limpar dados

TELA 2

Na parte 2 da AIVIDADE 1, apresentamos um exemplo de gráfico de dispersão com dados do IMC e do IDH, referentes a uma amostra de adolescentes, com idades entre 10 e 18 anos, matriculados na rede pública de ensino de 93 municípios do estado do Paraná.



TELA 3

Observemos que no gráfico há uma leve associação positiva entre as variáveis, indicada por uma tênue tendência crescente da nuvem de pontos, ou seja, quanto maior o IDH, maior é em média o IMC.

O ajuste de uma curva que modele adequadamente as relações entre as variáveis é chamado regressão entre as variáveis e tem como um dos objetivos fazer previsões ou diagnósticos sobre uma das variáveis (usualmente chamada variável-resposta ou dependente) a partir do valor da outra variável (chamada covariável ou variável independente).

O modelo de regressão mais simples entre duas variáveis é o modelo linear, isto é, assumimos que uma reta pode modelar bem a relação entre as variáveis.

O gráfico de dispersão entrega um primeiro indício (visual) da qualidade de um ajuste linear. O segundo indício usualmente utilizado para quantificar esse ajuste é o coeficiente de correlação linear, r , que será obtido na ATIVIDADE 2.

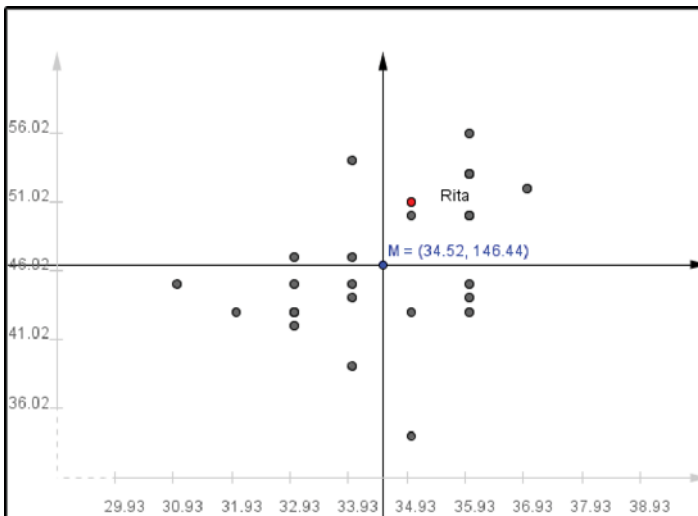


2 Gráfico de dispersão para “números do calçado” e “altura”

Na ATIVIDADE 2, o aluno deverá construir e analisar um gráfico de dispersão relacionando as variáveis “número de calçado” e “altura” de uma pessoa. Este é um gráfico bidimensional em que cada ponto (x, y) representa as mensurações das variáveis para os indivíduos da amostra.

A fim de obter o coeficiente de correlação linear, r , definiremos um novo eixo de coordenadas para os dados obtidos.

A parte 2 da ATIVIDADE 2 consiste em transladar a origem do eixo original para o ponto de médias da amostra, definido como o ponto cujas coordenadas são as médias de cada uma das variáveis. Assim, o aluno deve localizar o ponto móvel disponível na ferramenta no ponto com primeira coordenada igual ao número de calçado médio da amostra, e com segunda coordenada igual à altura média da amostra.

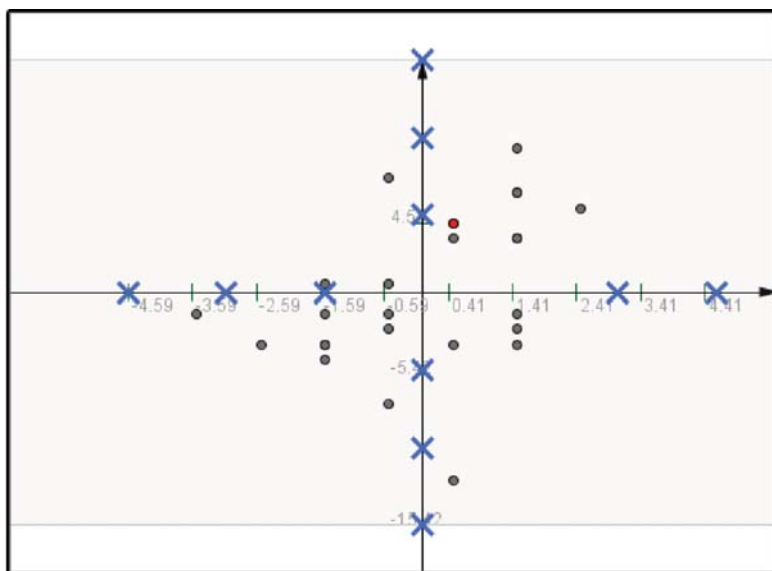


TELA 4

Na parte 3 da ATIVIDADE 2, o aluno reescalará os eixos de modo que a unidade de cada eixo represente um desvio-padrão.

Assim, na nova escala, por exemplo, o ponto $(1,1)$ representará um indivíduo (da amostra ou não) cujo número de calçado está 1 desvio-padrão acima da média da amostra, e cuja altura está 1 desvio-padrão acima da altura média da amostra.

Denotamos por (z_X, z_Y) as coordenadas do ponto (x, y) nesse novo sistema de coordenadas. A variável Z_X é chamada escala-padrão de X .



TELA 5



Definição

O valor médio dos produtos das coordenadas em escala-padrão é o chamado coeficiente de correlação linear, r :

$$r = \frac{1}{n} \sum_{i=1}^n z_{Xi} \cdot z_{Yi}$$

Observemos que, quando a maioria dos pontos está no 1º ou no 3º quadrante, r deve ser positivo, revelando uma associação positiva entre as variáveis, no sentido de que, à medida que uma assume valores maiores, a outra deve aumentar também.

Se a maioria dos pontos estiver no 2º ou no 4º quadrante, r deve ser negativo, indicando uma associação negativa entre as variáveis: à medida que uma assume valores maiores, a outra deve diminuir.

Se não houver nenhuma tendência crescente nem decrescente, r deve ser próximo de zero, indicando que não há associação entre as variáveis.

Se os pontos observados na amostra do par (X, Y) estiverem todos alinhados, então o coeficiente de correlação linear deve ser igual a 1 ou -1. De fato, neste caso, podemos escrever Y como $aX + b$. Dessa maneira, o valor médio de Y , \bar{Y} , é igual a $a\bar{X} + b$, e o desvio-padrão de Y , dp_y , é igual a $|a| \times dp_x$. Portanto,

$$\begin{aligned} r &= \frac{1}{n} \sum_{i=1}^n Z_{Xi} Z_{Yi} = \frac{1}{n dp_x dp_y} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \\ &= \frac{1}{n dp_x a dp_x} \sum_{i=1}^n |a|(x_i - \bar{x})^2 = \frac{\bar{a}}{a} \end{aligned}$$

No entanto, é importante ressaltar que a análise gráfica deve sempre ser a primeira a ser realizada, porque é ela que entrega a informação mais completa. As medidas-resumo são apenas características numéricas que podem eventualmente não entregar informação tão completa sobre a amostra quanto um bom gráfico.

Fechamento

O foco deste software é o uso de gráficos para tratar as informações coletadas em uma amostra, em particular o uso de gráficos para observar a relação entre duas variáveis quantitativas.

Esta análise pode ser estendida para duas categorias. Por exemplo, poderíamos analisar o gráfico de dispersão entre as variáveis separado a amostra pela variável “gênero” e por turmas diferentes (A, B, C, D) ou por períodos diferentes de uma mesma série (matutino, vespertino e noturno).

Dessa forma, surgem as questões: Existe algum tipo de relação entre as duas variáveis em cada categoria? Este tipo de relação é a mesma nas diferentes categorias? Quanto vale o coeficiente de correlação linear para cada categoria?

Discuta com os alunos a utilidade dessas ferramentas, abordando as soluções obtidas por cada grupo.

Os alunos podem trazer para a aula seguinte um conjunto de dados com informação sobre outras variáveis quantitativas.

Podemos estender também a análise da construção de modelos estatísticos relacionando mais de duas variáveis. Mas isto foge ao escopo desta atividade.

Em linhas gerais, o objetivo da representação gráfica é entregar informação visual clara a respeito das variáveis de interesse. Quanto mais clara for essa informação, melhor é o gráfico.

O software “Gráficos de Barras e de Setores” pode ser útil na realização de uma atividade que explore essas ideias. Ele permite ao usuário plotar diversos tipos de gráfico a partir de um conjunto de dados qualquer.

O software “Histogramas e Quantis” pode ser útil na realização de uma atividade que explore variáveis quantitativas.

Existem outros dois softwares que compartilham com este o mesmo conjunto de dados: Medidas do Corpo – Gráficos Univariados e Medidas do Corpo – Box Plot. Ambos são boas alternativas para complementar as atividades desenvolvidas a partir deste.



Bibliografia

COSTA NETO, Pedro Luiz de Oliveira. **Estatística**. Editora Edgard Blücher, 2002.

MEYER, Paul. Probabilidade: **Aplicações à Estatística**. Livros Técnicos e Científicos Editora, 2003.

Ficha técnica

AUTOR

Laura Leticia Ramos Rifo

PROJETO GRÁFICO

Preface Design



UNIVERSIDADE ESTADUAL DE CAMPINAS

Reitor

Fernando Ferreira Costa

Vice-Reitor

Edgar Salvadori de Decca

Pró-Reitor de Pós-Graduação

Euclides de Mesquita Neto

MATEMÁTICA MULTIMÍDIA

Coordenador Geral

Samuel Rocha de Oliveira

Coordenador de Software

Leonardo Barichello

Coordenador de Implementação

Matias Costa

INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA (IMECC – UNICAMP)

Diretor

Jayme Vaz Jr.

Vice-Diretor

Edmundo Capelas de Oliveira

LICENÇA Esta obra está licenciada sob uma licença Creative Commons 