IA369P – Tópicos em Engenharia de Computação VI

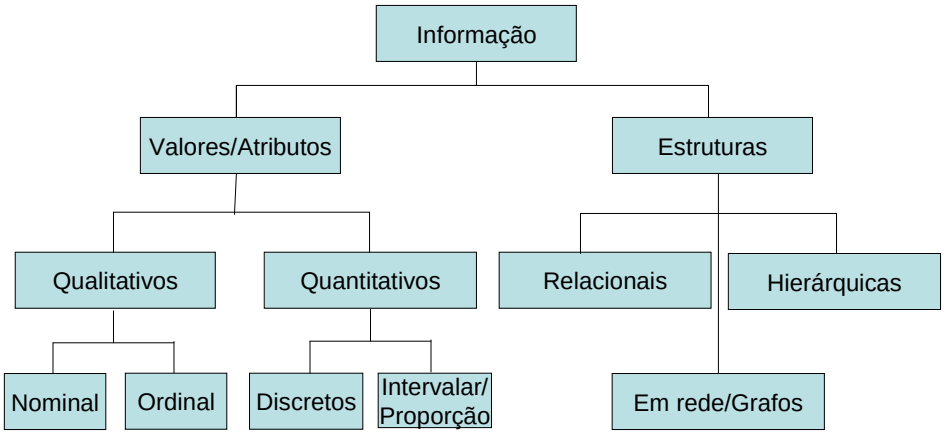
Visualização de Informação: Algoritmos

Análise de Dados

Capítulos 1 e 2 do livro-texto Cleveland

Informação

Uma Classificação



Tipos de Atributos

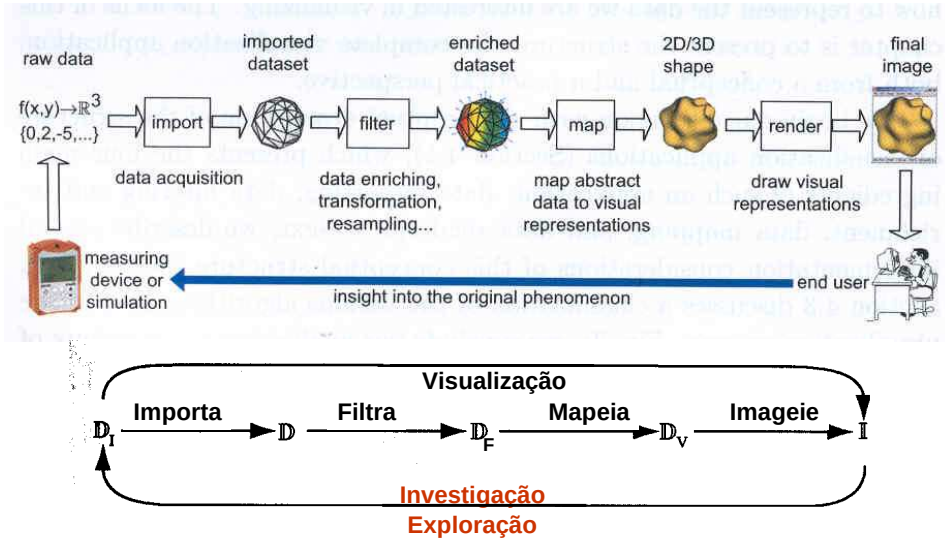
Atributos	Características	Domínio	Operações	Exemplos
Nominal	Dados categóricos que não apresentam uma ordem intrínseca	Conjunto não-ordenado	Comparações (igualdade)	nomes de objetos, números de identificação
Ordinal	Dados categóricos que apresentam uma ordem	Conjunto ordenado	Comparações (igualdade, maior, menor)	classificação de uma avaliação
Discretos	Dados contáveis	Domínio de números interios	Operações sobre inteiros	Linhas de um programa, quantidade de caracteres em um texto
Intervalar + proporção	Dados não contáveis	Domínio de números reais	Operações sobre reais	tempo, altura, distância

Análise de Dados

Coletar, modelar e transformar dados com o objetivo de enfatizar informação relevante, sugerir hipóteses, facilitar a elaboração de conclusões e suportar tomadas de decisão.

Será que a visualização também ajuda entender os dados coletados, fazer inferências e elaborar conjeturas?

Modelo Conceitual



Mapeamento

- Mapeamento de informação visual e verbal
  - 2 atributos → (x,y)
  - 3 atributos → (x,y,z)
  - 4 atributos → (x,y,z,cor)
  - 5 atributos → (x,y,z,matiz,luminância)
  - k atributos → vetores de dimensão k
    - Utilizar tabelas com k colunas
    - utilizar mais atributos gráficos
    - substituir dados relacionais/estruturais por diagramas
    - decompor vetores em k escalares
    - projetar vetores em dimensão 2 ou 3
    - posicionar os dados em dimensão 2 ou 3, preservando distância entre os vetores correspondentes

Estatística

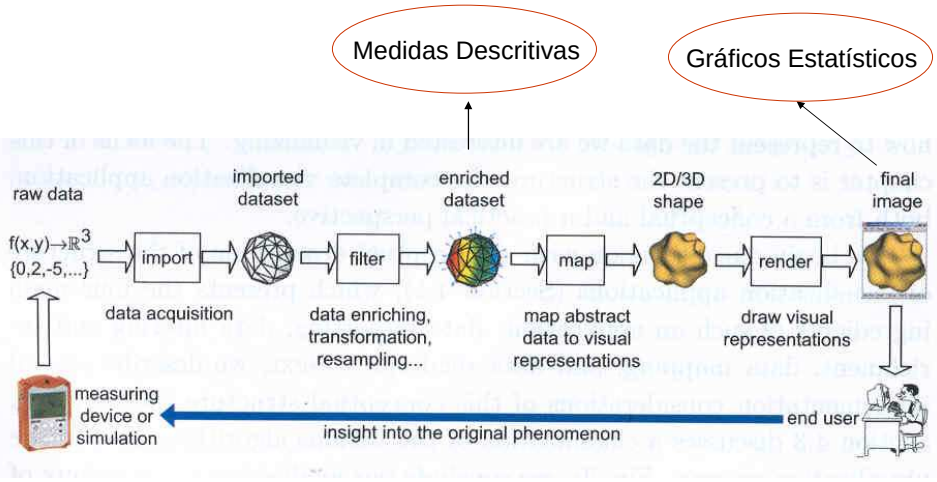
- A estatística utiliza-se das teorias probabilísticas para explicar a frequência da ocorrência de eventos, tanto em estudos observacionais quanto em experimento, modelar a aleatoriedade e a incerteza de forma a estimar ou possibilitar a previsão de fenômenos futuros, conforme o caso.
- A estatística é uma ferramenta matemática que nos informa sobre o quanto de erro nossas observações apresentam sobre a realidade observada.
- O objetivo da estatística é a produção da melhor informação a partir de dados disponíveis.
- Duas áreas:
  - Estatística descritiva: ocupa-se com a descrição dos dados
  - Inferência estatística: com base na Teoria das Probabilidades, fazer afirmações a partir de um conjunto de dados.

http://pt.wikipedia.org/wiki/Estat%C3%ADstica

População e Amostra

Lei de Bernouilli (1654-1705), ou **primeira lei dos grandes números**: "É muito pouco provável que, se efetuarmos um número suficientemente grande de experiências, a freqüência relativa de um acontecimento se afaste muito da sua probabilidade (de sucesso)."
Segunda lei dos grandes números: "À medida que o número de repetições de um experimento aleatório cresce, maior tende a ser o valor absoluto da diferença entre a freqüência absoluta experimental de um sucesso e a freqüência absoluta teórica (esperada)."

Modelo Conceitual



Medidas Descritivas

- Dados Qualitativos
  - Frequência em cada categoria
    - Função de Distribuição Acumulada:descreve completamente a distribuição da probabilidade de uma variável aleatória de valor real *X*
- Dados Quantitativos
  - Medidas-resumo = medidas de posição + medidas de dispersão
    - Média: razão entre a soma dos valores de dados e o número de ocorrências .
    - Mediana: valor do dado central, depois de ordenarmos os dados por ordem crecente ou decrescente
    - Variância: razão entre a soma dos quadrados dos desvios de cada dado em relação à média e o número de ocorrências.
    - Desvio-padrão: raíz quadrada da variância.
    - Erro-padrão: média das amostras de tamanho n em relação em relação à média populacional. Razão entre o desvio-padrão e raíz do tamanho amostral n.
    - Quantis:são pontos de corte que determinam as fronteiras entre subconjuntos consecutivos. o *k*-ésimo *q*-quantil é o valor *x* tal que a probabilidade de um evento da variável aleatória será inferior *x* é no máximo *k*/*q* e a probabilidade da variável aleatória ser superior ou igual a *x* é pelo menos *(q-k)*/*q*.
    - Máximo e Mínimo: máximo e mínimo dos valores.

Exemplo

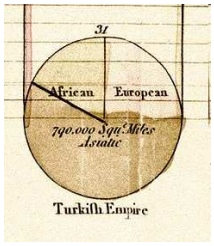
- Uma amostra: {2,3,5,7,9}
  - Média: $\frac{2+3+5+7+9}{5} = 5.2$
  - Mediana: 5
  - Variância: $\frac{3.8^2 + 1.8^2 + (-0.2)^2 + (-2.2)^2 + (-3.2)^2}{5} = 6.56$
  - Desvio-padrão: $\sqrt{6.56} = 2.56$
  - Erro-padrão: $\frac{2.56}{\sqrt{5}} = 1.14$
  - Quantis:
    - Primeiro quartil ($Q_{1/4}$): 2.5
    - Segundo quartil ($Q_{2/4}$): 5
    - Terceiro quartil ($Q_{3/4}$): 8
  - Máximo e Mínimo: 9 e 2

Exercícios

- Determine as medidas descritivas para as seguintes amostras:
  - {6, 47, 49, 15, 42, 41, 7, 39, 43, 40, 36} (Quantil: decis = 10-quantil)
  - {7, 15, 36, 39, 40, 41} (Quantil: quartil = 4-quantil)
  - {57.0, 62.9, 63.5, 64.1, 66.5, 67.1, 73.6., 89.0} (Quantil: mediana=2-quantil)

Gráficos Estatísticos

- Gráficos Estatísticos: representações gráficas para
  - explorar/inspecionar o conteúdo de um volume de dados
  - achar a estrutura intrínseca (relações) de um volume de dados
  - validar as suposições em modelos estatísticos
  - visualizar os resultados de uma análise estatística
- Gráficos estatísticos famosos:
  - http://en.wikipedia.org/wiki/Statistical_graphics



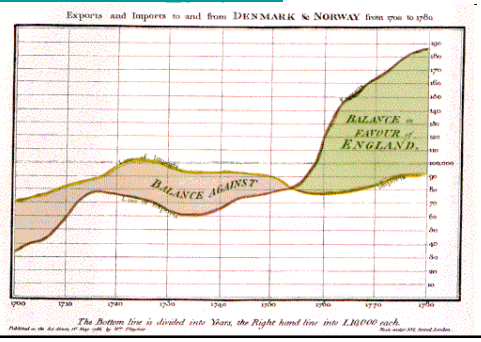
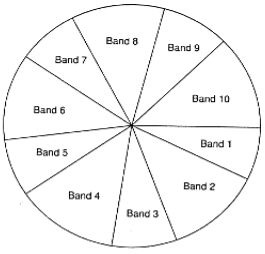
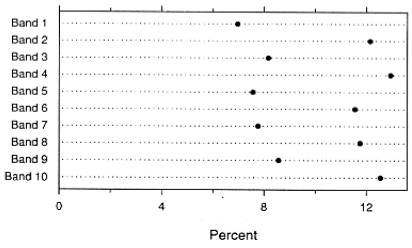


Gráfico de Pontos X Gráfico-Pizza

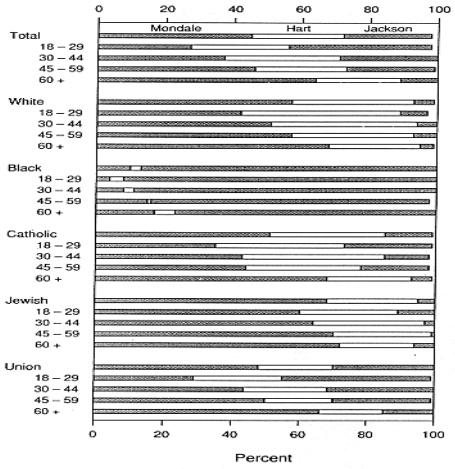


4.19 PIE CHART. The pie chart falls in the category of a pop chart — a graphical method used frequently in the mass media and certain business presentations but far less in science and technology. Both table look-up and pattern perception are less efficient for pie charts than for dot plots.



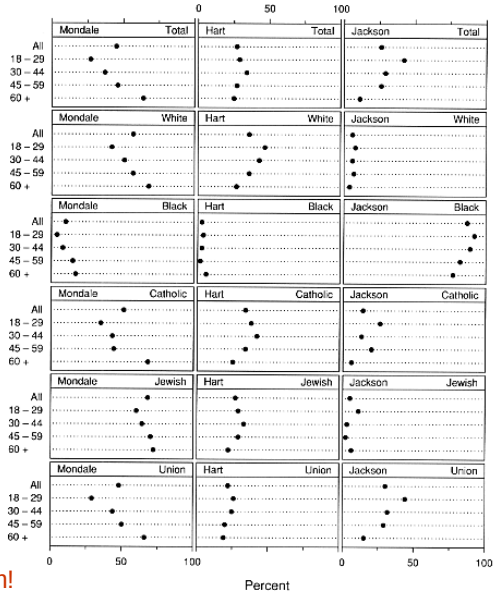
4.20 DOT PLOT. The data from Figure 4.19 are graphed by a dot plot. Patterns emerge that cannot be decoded from Figure 4.19.

Gráfico de Barras Particionadas



4.21 DIVIDED BAR CHART. A divided bar chart is used to show the percentage of the vote for three candidates in the 1984 New York Democratic primary election. The Mondale values are graphed by position along a common scale, but the Hart values and the Jackson values are not and our visual decoding of these latter two sets of values is less accurate than for the Mondale values.

Gráfico de Pontos de Correspondência Múltipla

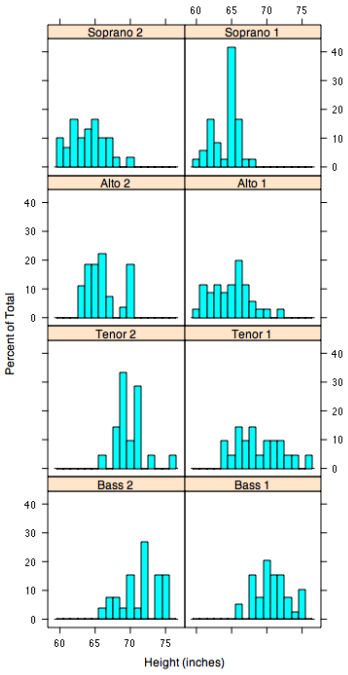


Referencial de comparação comum!

4.22 MULTIWAY DOT PLOT. The data from Figure 4.21 are graphed by a multiway dot plot. Now the Hart values and the Jackson values are encoded by position along a common scale. Now we can perceive a Hart age pattern.

Histograma

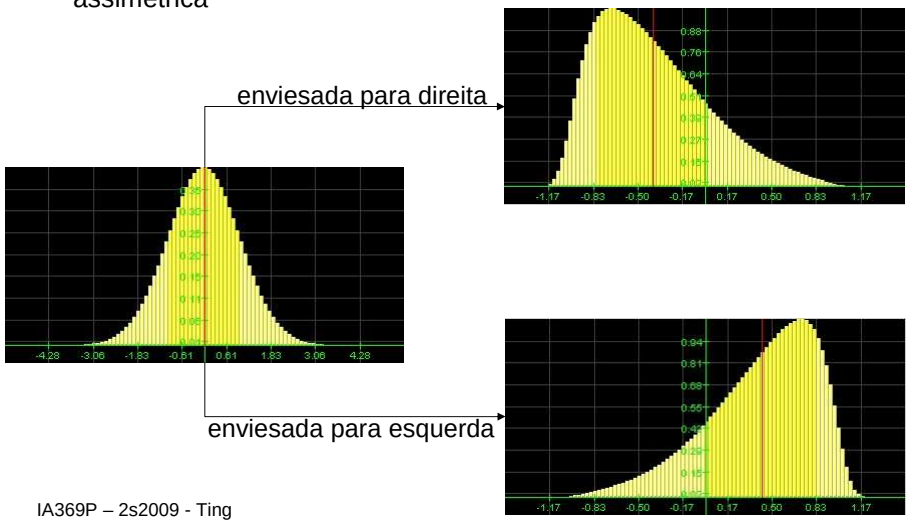
- Gráfico de barras de distribuição de frequência de um volume de dados.



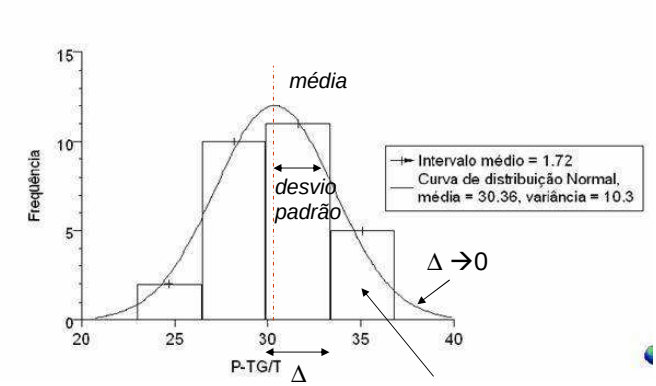
http://osiris.sunderland.ac.uk/~cs0her/Statistics/UsingLatticeGraphicsInR.htm

Histograma

- Distribuição enviesada: distribuição de frequência acentuadamente assimétrica

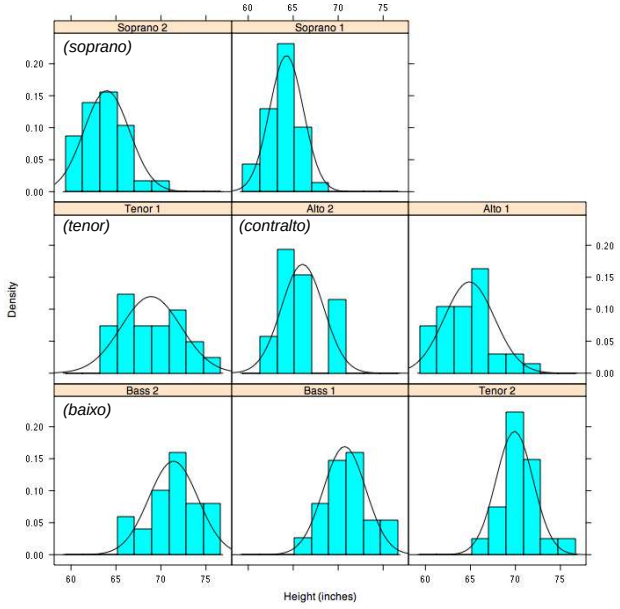


Distribuição de Frequência Contínua



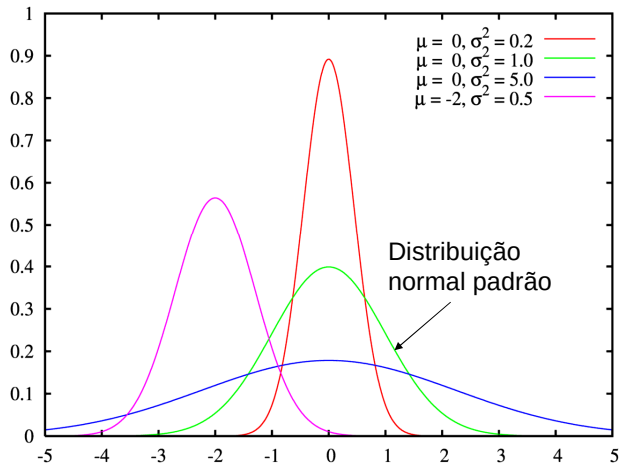
Probabilidade de ocorrência
Área Total = 1 ≡100%

Distribuição de Frequência Contínua



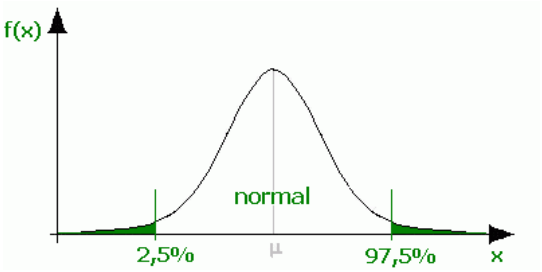
Distribuição Normal

Função de densidade de probabilidade: $f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{(-\frac{(x-\mu)^2}{2\sigma^2})}$



Distribuição Normal

- Um valor tem ocorrência normal se está entre 95% da área sob curva da distribuição normal.

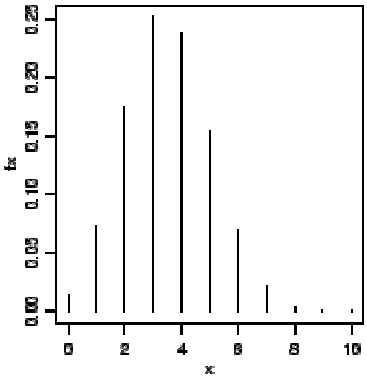


μ ± σ → 68.26%
μ ± 2σ → 95.44%
μ ± 3σ → 99.74%

http://www.ufpa.br/dicas/biome/bionor.htm

Distribuição de Frequência Acumulada

- Frequência Acumulada: soma das freqüências absolutas anteriores de um determinado valor.



Histograma

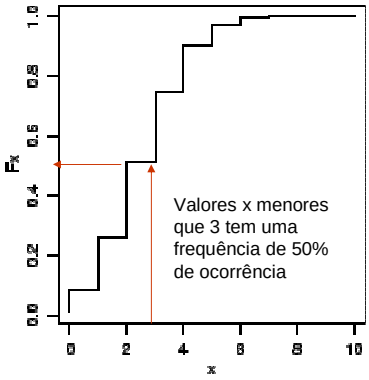
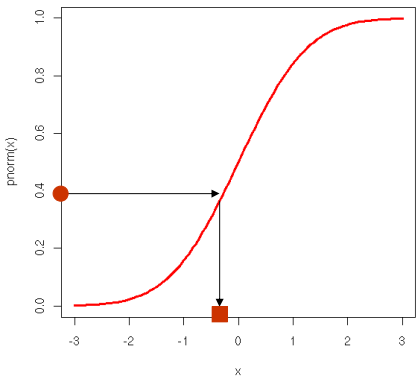


Gráfico de distribuição de frequência acumulada

Função de Quantil

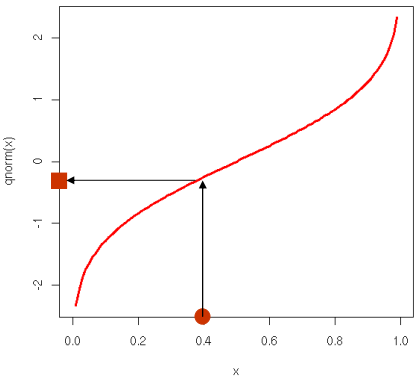
- Função inversa de distribuição de frequência acumulada.

Função de distribuição normal acumulada



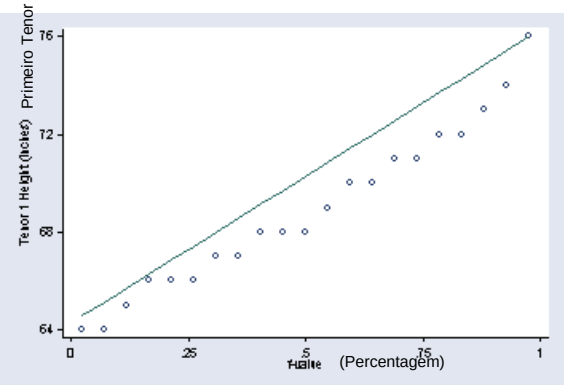
$$pnorm(x) = \int_{X=x_{min}}^{x} f(X,\mu,\sigma) = p$$

Função de quantil da distribuição normal



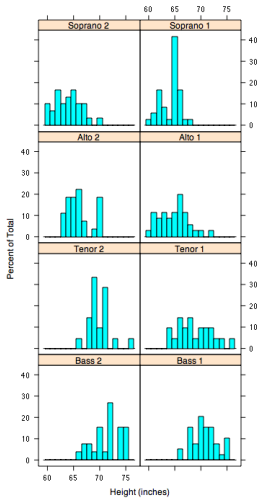
$$qnorm(x) = q_{\mu,\sigma}(f) = pnorm^{-1}(x)$$

Gráfico de Quantil



http://www.ats.ucla.edu/stat/Stata/examples/vizdata/vizdatach2.htm

Gráfico de Quantil



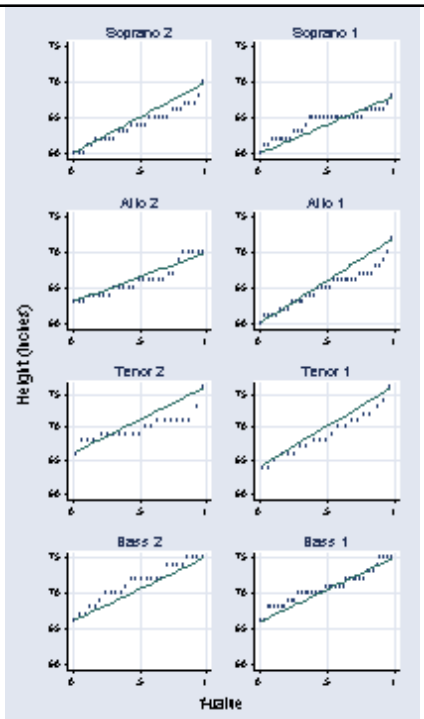


Gráfico QQ

- Em qual dos gráficos a precipitação prevista é maior do que a precipitação observada?

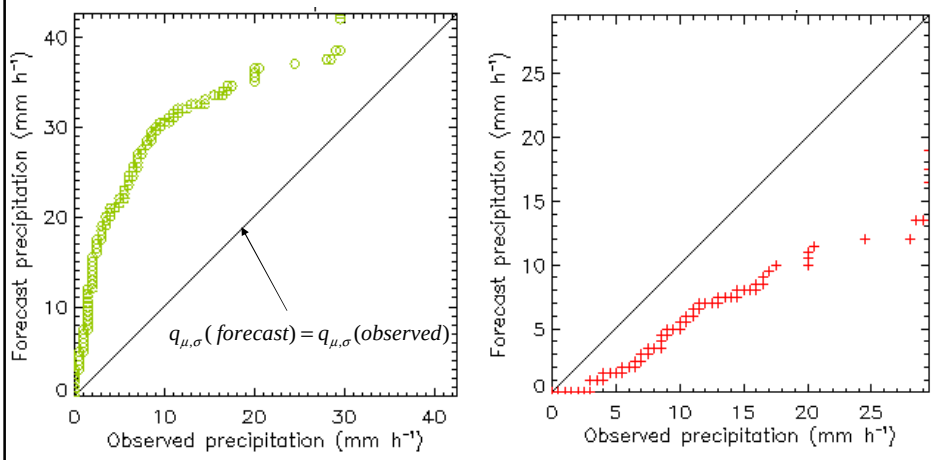
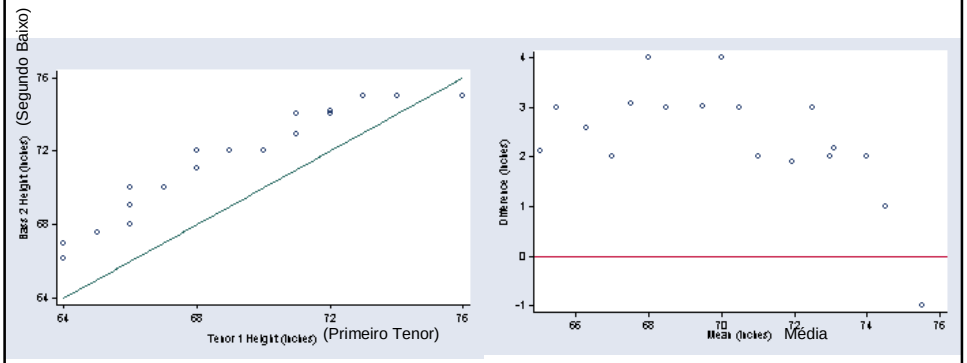


Gráfico de Diferença Média *Tukey*

- Representa visualmente a diferença entre duas distribuições de ocorrências.



http://www.ats.ucla.edu/stat/Stata/examples/vizdata/vizdatach2.htm

Diagrama de Caixas

- *Boxplots*: Representa as medidas-resumo dos dados: os valores centrais e alguma informação a respeito da amplitude deles (5 valores).

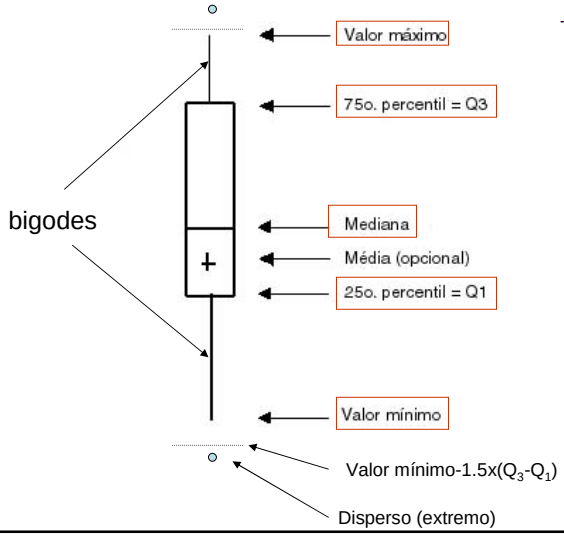
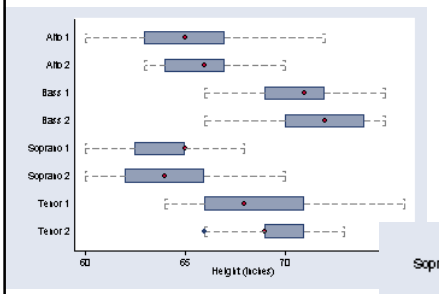


Diagrama de Caixas



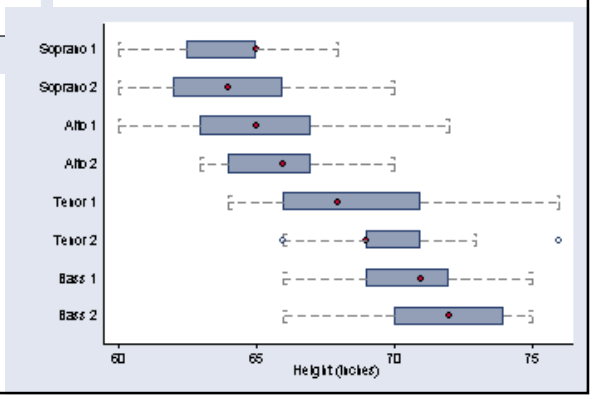


Diagrama de Caixas

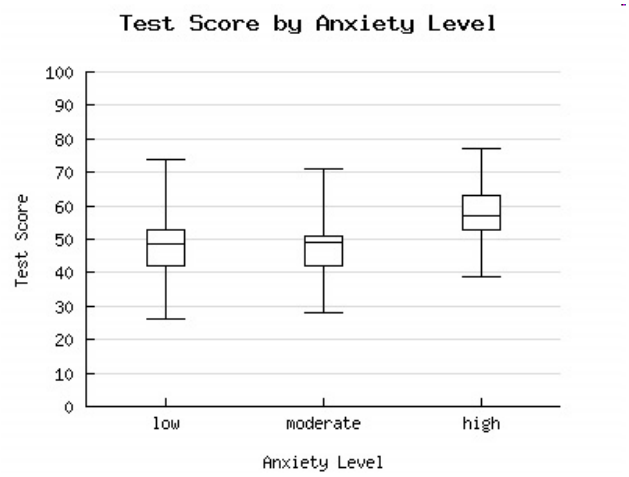
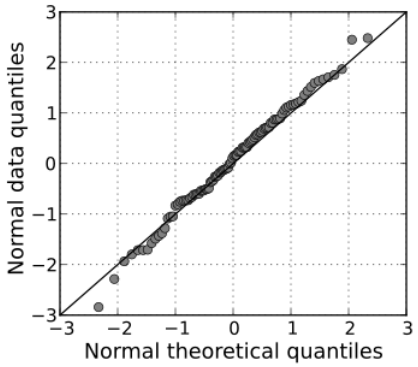


Gráfico QQ Normal

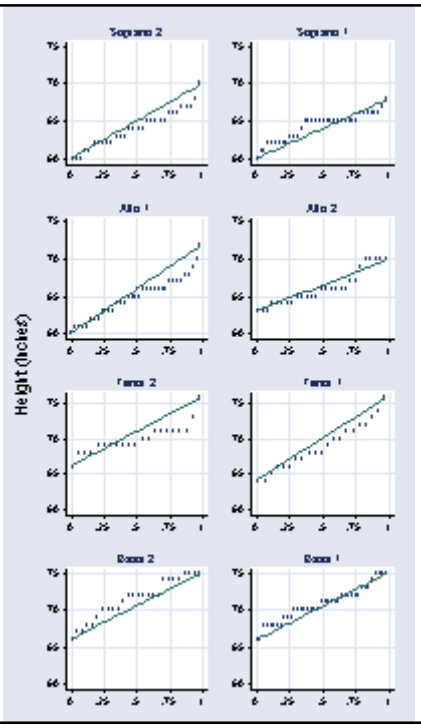
- Relaciona os quantis de um conjunto de dados com os quantis de uma distribuição normal.

$$q_{\mu,\sigma}(f) = \mu + \sigma q_{0,1}(f)$$



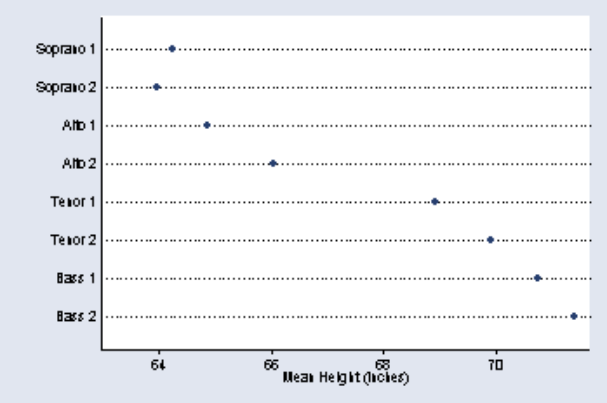
Se o gráfico QQ normal se aproxima a uma reta → a estrutura dos dados avaliados tem uma distribuição normal → Graficamente, pode-se comparar os desvios-padrão das distribuições pela inclinação da reta.

Gráfico QQ Normal



Ajustes e Resíduos

- Ajuste: determinar a função matemática que melhor se aproxima dos dados (média ou mediana).
- Resíduo: diferença entre o valor real e o ajuste



Ajustes

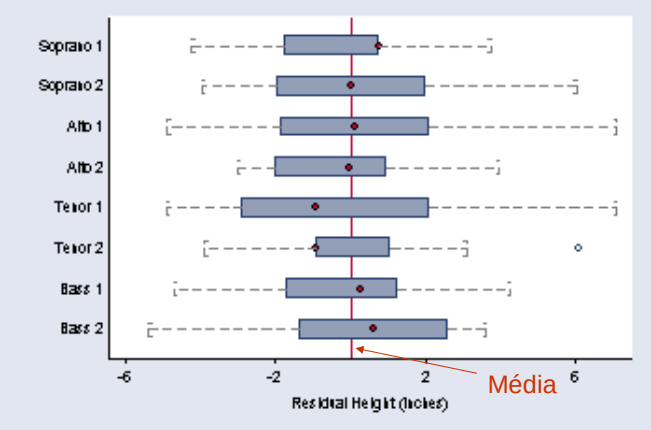
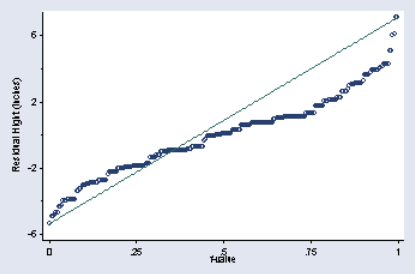
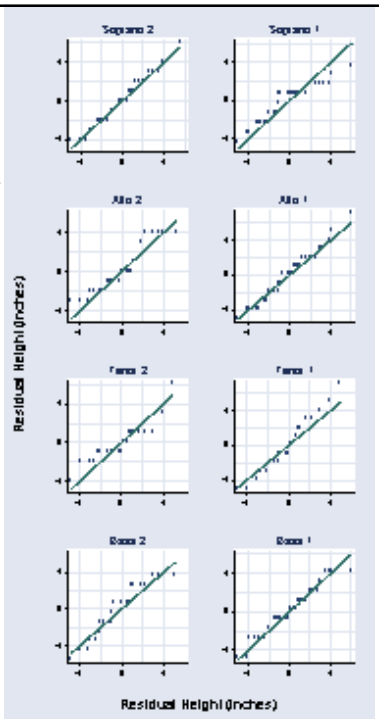


Gráfico QQ

- Distribuição de resíduos dos dados cada grupo de cantores em relação à distribuição dos resíduos de todos os dados





Inferência Estatística

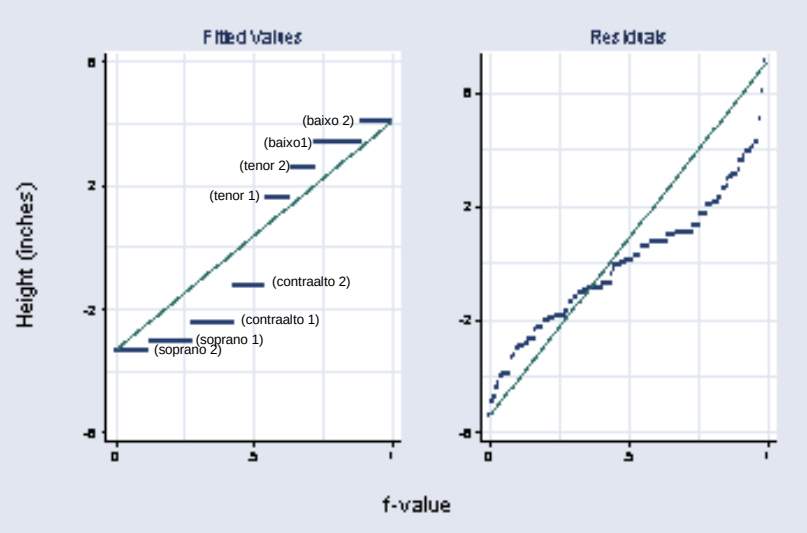
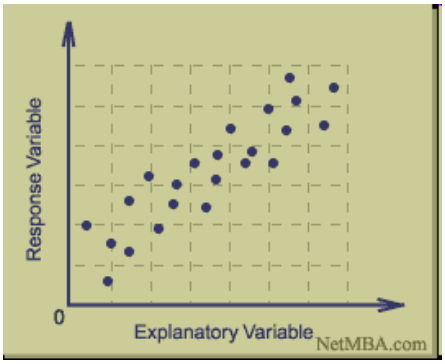
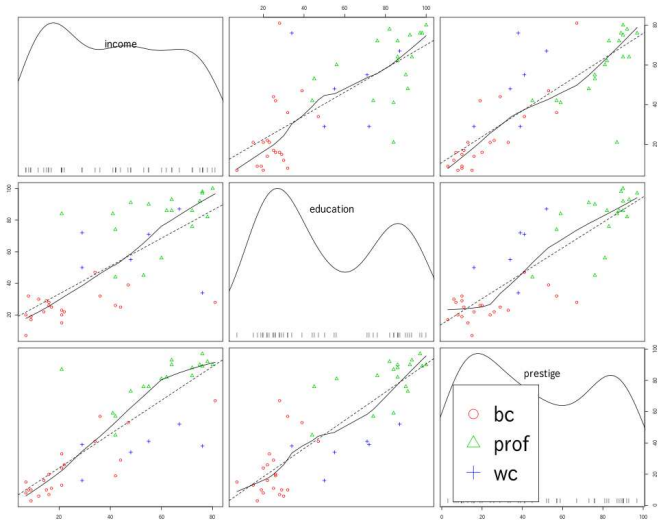


Gráfico de Dispersão

- Permite visualizar a relação entre duas variáveis.



Matriz de Gráficos de Dispersão



Exercícios

1. Qual é a relação entre um histograma e um gráfico de distribuição normal?
2. O que você entende por uma distribuição enviesada?
3. Por que visualmente é mais fácil comparar duas distribuições pelo gráfico QQ do que pelos histogramas?
4. Como se pode descobrir visualmente se os dados em análise tem uma distribuição normal?
5. Como se pode visualmente sintetizar a informação contida em uma matriz de gráficos QQ?
6. O que um gráfico de dispersão pode revelar?