Estatística Aplicada à Medicina I – turma A
Análise Exploratória de Dados – Leitura Complementar para o Trabalho e para a Avaliação Escrita (1° TVC)

**"O raciocínio estatístico será, um dia,
tão importante para o exercício eficiente
da cidadania quanto saber ler e escrever"**

H. G. Wells

Professor Ronaldo Rocha Bastos, Ph.D
Departamento de Estatística – ICE/UFJF

Setembro de 2010

Estatística Descritiva / Análise Exploratória de Dados

Gráficos, tabelas e medidas são maneiras de se **descrever** um conjunto de dados. Fazendo uma analogia, poderíamos dizer que o gráfico de ramo-e-folhas (*stem-and-leaf*), por exemplo, é como a **fotografia** de uma pessoa, o de Tukey (*box-plot*) como o **desenho** desta pessoa, retendo apenas os traços essenciais, mais marcantes, e as medidas (*statistics*), uma **descrição escrita** destes traços. Todas estas técnicas têm seu valor, pois cada uma delas realça um aspecto particular, e juntas elas se complementam.

Para poder realizar uma análise exploratória de dados, anote primeiro todos os detalhes que chamarem sua atenção e todas as dúvidas que ocorrerem. Observe, por exemplo:

Quantas observações foram feitas? Quais são os pontos mínimos e máximos? Estes pontos são discrepantes ("outliers")? Você consegue explicá-los? Qual o formato geral do gráfico traçado? É simétrico ou assimétrico? Há espaços sem observações (falhas, buracos) na distribuição dos dados? Em que faixas de valores se encontra a maioria dos dados? Os dados estão dispersos ou aglomerados? Há aglomerados localizados de dados ("clusters")? Onde estão estes "clusters"? Você consegue explicá-los?etc.

Faça também perguntas sobre os dados que recebeu, por exemplo:

Quem forneceu os dados? Como será que eles foram obtidos? Será que eles são realmente válidos? As unidades de medida utilizadas são adequadas? Falta alguma coisa importante nos dados?

Em função dos dados obtidos (se qualitativos: nominais ou ordinais; se quantitativos: discretos ou contínuos) decida-se pelo(s) gráfico(s) e tabela(s) que sejam mais adequados a estes dados e pelas medidas que sejam possíveis de serem calculadas e interpretadas. No caso de medidas, perguntar:

Quais as vantagens desta medida? Existem situações que desaconselham a utilização desta medida? A interpretação da medida é plausível? Outras pessoas saberão interpretar as medidas escolhidas?

Você provavelmente não conseguirá responder a algumas das perguntas acima, mas fazer perguntas deste tipo é essencial para desenvolver sua capacidade de analisar e criticar dados estatísticos. O melhor para organizar todas as idéias é escrever um resumo, ou relatório sucinto, da forma mais clara possível. Organizar os achados da análise, resumir e comunicar as informações numéricas obtidas constituem uma habilidade importante para a carreira de um profissional.

Lembre-se de que a forma de análise de um conjunto de dados depende muito da natureza dos mesmos. Dados **qualitativos**, sejam nominais ou ordinais, permitem certos tipos de análise. Já os dados **quantitativos**, sejam discretos ou contínuos, permitem análises que não são possíveis de se realizar com dados qualitativos.

Uma forma sistemática de se iniciar uma análise exploratória de dados é procurar pelas *seis características de uma base de dados*, a saber: **formato, localização, dispersão, pontos discrepantes, aglomerados e granularidade.**

O **formato** de uma base de dados é, sem dúvida, o fator mais importante para se decidir quais as medidas e os gráficos que melhor descrevem o conjunto de dados. Aqui estamos falando da análise de simetria e caracterização como unimodal, bimodal ou multimodal. Por exemplo, uma distribuição bimodal não fica bem caracterizada ao ser representada por um diagrama de Tukey. (*box plot*).

A **localização** aproximada de uma distribuição é inicialmente estimada visualmente a partir dos gráficos. Posteriormente, após a análise do formato da distribuição, chegamos à escolha da(s) medida(s) de centro mais adequada(s): média, mediana, moda, média truncada, etc.

A **dispersão** mede a quantidade de variação apresentada pelos dados. Novamente devemos partir de uma avaliação meramente visual e, posteriormente, escolher a medida(s) de dispersão mais adequada(s) em função do formato e do propósito para o qual iremos calcular tal medida. Aqui estamos falando da variância, do desvio-padrão, do intervalo interquatílico e da amplitude, embora esta última medida seja pouco robusta.

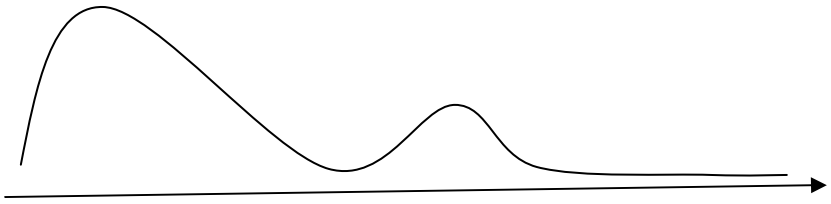
Pontos discrepantes são aqueles valores que se encontram afastados do aglomerado geral formado pelos outros valores do conjunto de dados. Cada ponto discrepante deve ser cuidadosamente analisado, com o objetivo de se verificar a sua representatividade diante da população em estudo (neste caso deve ser mantido) ou sua pouca representatividade ou erro (quando pode ser eliminado). Notar que em alguns casos o ponto discrepante pode ser o valor mais importante da base de dados. Existe um caso verdadeiro de uma análise computadorizada automática ter excluído um ponto discrepante importante: o buraco na camada de ozônio acima do Polo Sul foi detectado por um satélite muito antes do mesmo ser detectado de bases de observação no solo; os valores medidos forma excluídos pelo programa de computador por serem muito menores que os valores que se imaginava possíveis! O diagrama de Tukey é o melhor para a procura e análise de pontos discrepantes. Outros gráficos também podem ser utilizados, lembrando que a grande contribuição do computador é justamente possibilitar que se analise os dados de diferentes formas de maneira rápida e eficiente.

Os **aglomerados** indicam que os dados tendem a se concentrar ao redor de certos valores, formando os agrupamentos que são chamados de aglomerados pelos estatísticos. O gráfico que permite a melhor visualização de aglomerados é justamente o mais simples: o gráfico de pontos.

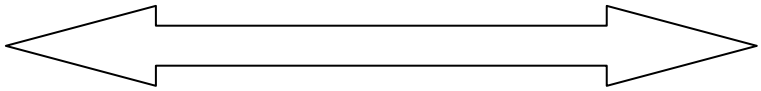
A granularidade indica que apenas valores discretos são permitidos para representar as observações. Isto implica que os dados são realmente discretos ou então que os dados eram contínuos e foram arredondados ou truncados para simplificar a análise. O gráfico de pontos indica bem este fenômeno, com pontos sobrepostos separados por espaços. A observação da granularidade nos permite inferir ou confirmar a forma de coleta dos dados.

As seis características a serem observadas em uma base de dados, de forma esquemática:

- formato



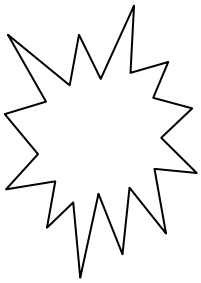
- localização



- dispersão

- pontos discrepantes

- aglomerados



- granularidade