Aspectos básicos da descrição e sumarização de informações em medicina

Mário B. Wagner
Doutor em Epidemiologia (Universidade de Londres)
Professor Adjunto, Departamento de Medicina Social,
Faculdade de Medicina, Universidade Federal do Rio Grande do Sul

Fonte:
Jornal de Pediatria 1998; **74**:71-76.

Resumo
Objetivo: Introduzir conceitos essenciais sobre descrição e sumarização de dados. Métodos: Revisão de diversos livros estatísticos, selecionando aspectos a considerar na descrição e sumarização de dados. Resultados: As variáveis foram classificadas como dependente (desfecho) e independente (exposição). Os níveis de medida descritos foram nominal, ordinal, intervalar e de razão. Pontos importantes mencionados na elaboração de tabelas incluíram um título claro e informativo, número adequado de intervalos de classe e categorias mutuamente exclusivas. Quanto aos gráficos, as escalas devem ser claras e sempre que possível o valor zero deve estar presente no eixo vertical. Medidas de tendência central, como a média e a mediana, descrevem o "valor típico" dos dados. Medidas de dispersão, como o desvio padrão e a amplitude entre quartis, estimam a variabilidade em torno da tendência central. Conclusões: A descrição e a sumarização de dados formam uma parte essencial da análise de dados, freqüentemente negligenciada pelos pesquisadores. É universalmente aceito que o nível de medida de uma variável e a forma da distribuição dos dados determinam o tipo de análise e ser conduzida. Os gráficos fornecem uma visão geral dos dados e as tabelas concentram-se nos detalhes. Os sumarizadores numéricos (medidas de tendência central e dispersão) são fundamentais para a descrição e a inferência estatística. A média está indicada quando os dados possuem pelo menos um nível de medida intervalar e seguem uma distribuição normal. Neste tipo de dados use os testes paramétricos. Para dados ordinais ou assimétricos considere o uso da mediana e dos testes não-paramétricos.

Abstract
Objective: To introduce essential concepts of data description and summarisation. Methods: Review of a number of medical statistics textbooks, selecting the aspects to be considered when performing data description and summarisation. Results: Variables were classified as dependent (outcome) and independent (exposure). The scales of measurement described were nominal, ordinal, interval and ratio. Important points mentioned when using tables included clear and informative title, adequate number of class intervals, and mutually exclusive categories. Regarding graphs, the scales should be clearly indicated and whenever possible the zero point should be visible on the vertical axis. Measures of central tendency, for instance the mean and the median, describe the "typical value" of a group of data. Measures of dispersion, such as the standard deviation and the interquartile range, estimate the spread around the central tendency. Conclusions: Data description and summarisation are an essential part of data analysis which is frequently overlooked by researchers. It is widely accepted that the scale of measurement of a variable and the underlying shape of data will determine the type of statistical analysis to be performed. Graphs should be used for a general view of data. Frequency tables are meant for details. Numerical summaries (measures of central tendency and dispersion) are fundamental for describing data and statistical inference. Means are always preferred when data are at least on an interval scale and follow a normal distribution. On this type of data use parametric tests. For ordinal or skewed data consider the median and non-parametric tests.

Introdução

Os artigos publicados em revistas biomédicas freqüentemente contém termos específicos do domínio da epidemiologia e da bioestatística. Para entender adequadamente esses artigos o leitor deve estar familiarizado com os princípios fundamentais do método epidemiológico e dos processos utilizados em bioestatística. Isto não significa, no entanto, que deva conhecer todos os detalhes técnicos dos cálculos estatísticos. Em uma época onde encontramos com facilidade poderosos computadores pessoais e sofisticados pacotes estatísticos é muito mais importante conhecer a indicacão e adequação dos testes estatísticos do que saber como executá-los.

Apesar da maioria dos médicos não considerar a pesquisa como uma de suas atividades principais, muitos estão indiretamente em contato com o processo de produção científica através da leitura de artigos em revistas biomédicas. Para uma boa leitura necessitam conhecimentos básicos de epidemiologia e bioestatística, de modo que possam diferenciar bons e maus artigos, determinar a validade de suas conclusões e entender as limitações dos estudos. Além disso, o conhecimento do método epidemiológico e dos procedimentos estatísticos é fundamental para um maior envolvimento com pesquisa, auxiliando no delineamento e execução de estudos, bem como na escolha das técnicas mais apropriadas para a análise de dados.

Variáveis e Dados

As pesquisas biomédicas, quando reduzidas aos seus elementos essenciais, podem ser consideradas como estudos de relações entre variáveis (1). Usualmente são estudadas as diferenças ou associações entre as variáveis observadas nos pacientes. Assim, podemos definir *variável* simplesmente como sendo toda característica ou condição que pode ser observada ou avaliada. Toda variável é passível de modificar-se, ou seja, apresentar variação em seus valores de paciente a paciente ou no mesmo paciente de um momento a outro (1,2,3). Chama-se *dado* o valor resultante da mensuração ou avaliação de uma variável em um paciente.

Tipos de variáveis e seus níveis de medida

As variáveis podem ser classificadas de diversas formas. Nos estudos biomédicos é comum diferenciarmos uma variável dependente e outra independente (2). De modo simplificado podemos considerar a *variável dependente* como sendo o desfecho de interesse e a *variável independente* como sendo o fator em estudo que pode estar associado com o desfecho. Por exemplo, se desejássemos avaliar a associação entre a poluição atmosférica e o baixo peso ao nascer, o grau de exposição das mães aos poluentes atmosféricos seria a variável independente (fator em estudo) e a ocorrência de recém-nascidos (RNs) de baixo peso seria a variável dependente (desfecho).

Tanto a variável dependente como a independente podem assumir uma série de valores específicos de acordo com as características próprias dos dados gerados. Assim, uma abordagem freqüentemente utilizada em estatística envolve classificar as variáveis de acordo com seus níveis de medida (1,2). Abaixo apresentamos os quatro tipos básicos de escalas ou níveis de medida listados em ordem crescente de refinamento no processo de mensuração.

> *Variável nominal*: Uma variável com nível de medida nominal compõe-se de duas ou mais categorias que não possuem nenhuma relação hierárquica entre si. A variável quando possui apenas duas categorias ou estados é também chamada de dicotômica, por exemplo, masculino ou feminino, curado ou não curado, grávida ou não grávida, exposto ou não exposto, vivo ou morto. Entretanto, as variáveis nominais comumente apresentam três ou mais categorias sendo então chamadas de politômicas ou polinomiais. Como exemplo podem ser citados o sistema de grupamento sangüíneo ABO em quatro categorias (A, B, AB e O), a cor dos olhos em cinco categorias (preto, marrom, azul, verde e mista) e o diagnóstico principal pela Classificação Internacional das Doenças (CID-10) que pode assumir qualquer uma de algumas centenas de categorias diagnósticas possíveis.

Variável ordinal: Variável que apresenta categorias que podem ser ordenadas de acordo com algum sistema de graduação, mas as diferenças entre as categorias não podem ser consideradas iguais. Um exemplo de variável ordinal é a avaliação do estado geral de um paciente em mau (MEG), regular (REG) ou bom estado geral (BEG). Neste caso não se pode afirmar que a diferença entre um paciente REG e um MEG é a mesma do que entre um BEG e um REG.

Variável intervalar: É uma variável em que se estabelece para cada mensuração um valor de uma escala supostamente ilimitada de valores que estão igualmente espaçados entre si. Neste tipo de escala, apesar dos intervalos serem regulares o valor zero é uma convenção, ou seja, o zero não representa a ausência total da variável que está sendo medida. Por exemplo, 0 ^{o}C não significa ausência de temperatura, mas simplesmente que a temperatura equivalente ao ponto de congelamento da água foi atingida.

Variável de razão: É uma variável que possui as mesmas características da variável com nível de medida intervalar, mas o valor zero não é arbitrário e representa um ponto verdadeiro. Exemplos de variáveis medidas em escalas de razão são altura, peso, tempo, volume, concentrações das mais variadas substâncias no organismo e temperatura em graus Kelvin. Apesar da diferenciação entre variáveis medidas em escalas intervalares e de razão, do ponto de vista estatístisco, elas são geralmente tratadas e analizadas do mesmo modo. O mesmo já não acontece com as variáveis com níveis de medida nominal e ordinal.

A princípio a diferenciação entre os níveis de medida (nominal, ordinal, intervalar e de razão) parece uma tarefa muito simples e de pouca relevância para ser discutida mais a fundo. Em algumas situações, no entanto, a realidade é mais complexa do que se apresenta à primeira vista. Tomemos como ilustração o exemplo da clássica escala de Apgar para avaliação de um RN. A condição fisiológica do RN é avaliada em unidades de Apgar em uma escala que tem como extremos 0 e 10. Estritamente falando não há nenhuma garantia de que a diferença entre um Apgar de 5 e outro de 8 seja igual a diferença entre um Apgar de 7 e outro de 10. Assim, a escala de Apgar é, por definição, uma variável ordinal.

Apesar disso, é comum se observar pesquisadores tratando o Apgar e outras variáveis ordinais semelhantes, como por exemplo o quociente de inteligência (QI) e muitos escores de avaliação psiquiátrica, como se fossem variáveis intervalares. Entendemos que a ocorrência deste fato em variáveis ordinais de escala relativamente ampla não possua uma repercussão maior no tratamento da informação (2). No entanto, deve-se sempre que possível exercitar e respeitar a diferenciação entre os níveis de medida das variáveis, uma vez que até certo ponto eles determinam os tipos de procedimentos estatísticos adequados (3,4).

Descrição de dados

Os dados clínicos oriundos de pacientes ou de investigações médicas podem ser descritos através de tabelas de freqüências, gráficos e sumarizadores numéricos.

Uma *tabela de freqüências* é composta de classes ou intervalos que dividem uma série de dados e registra-se a freqüência dos valores ocorridos dentro de cada classe (3,5). Em outras palavras, a tabela de freqüências descreve a *distribuição de freqüências* dos dados e representa como os valores da variável estudada estão distribuídos em suas classes específicas. Abaixo vemos uma tabela de freqüências por intervalos de classe para os níveis séricos do ácido úrico em 267 indivíduos normais (tabela 1). São apresentadas a freqüência absoluta (n por classe), a freqüência relativa (aqui expressa em percentuais) e o percentual acumulado.

Tabela 1: Níveis séricos de ácido úrico em 267 indivíduos normais

Ácido úrico sérico (mg/100 ml)	Freqüência (n)	%	% acumulado
3,0 - 3,4	2	0,75	0,75
3,5 - 3,9	15	5,62	6,37
4,0 - 4,4	33	12,36	18,73
4,5 - 4,9	40	14,98	33,71
5,0 - 5,4	54	20,23	53,94
5,5 - 5,9	47	17,61	71,55
6,0 - 6,4	38	14,23	85,78
6,5 - 6,9	16	5,99	91,77
7,0 - 7,4	15	5,62	97,39
7,5 - 7,9	3	1,12	98,51
8,0 - 8,4	1	0,37	98,88
8,5 - 8,9	3	1,12	100,00
Total	267	100,00	-

Para o nível de medida nominal é recomendável, sempre que possível, organizar as categorias por ordem de freqüências decrescentes. Desta forma, as categorias de maior freqüência, e provavelmente maior relevância, serão apresentadas em primeiro lugar. Já para os níveis ordinal, intervalar e de razão deve-se respeitar a ordem entre as categorias (2). A seguir são apresentadas algumas recomendações para a elaboração de tabelas de freqüências. Estas não são regras rígidas e absolutas a serem seguidas incondicionalmente, mas sugestões baseadas na literatura (1,2,3,4) e no bom senso: (a) o título deve ser o mais explicativo possível para o leitor; (b) evitar abreviações, caso isto não seja possível deve-se apresentar uma legenda ou nota de rodapé; (c) o número de classes de uma tabela de freqüências deve oscilar entre 2 e 20, conforme a situação; (d) os limites das classes devem estar de acordo com a precisão de mensuração da variável; (e) os limites inferior e superior das classes não devem induzir ambiguidade no leitor, p.e. "5 a 10 anos" e "10 a 15 anos". Onde colocamos o paciente com 10 anos? (f) as classes preferencialmente, mas não obrigatoriamente, devem ser de mesmo tamanho; (g) ao referir-se a um valor P, é de boa prática mencionar qual o teste estatístico utilizado.

A apresentação gráfica geralmente tem um bom impacto no leitor e facilita o entendimento. No entanto, deve-se evitar o uso exagerado do recurso gráfico, seja na forma de gráficos muito complexos ou através de gráficos desnecessários. Um bom exemplo de uso desnecessário é usar um gráfico de barras ou tipo pizza para representar a distribuição do sexo em uma amostra. Neste caso um simples percentual fornece a mesma informação, sendo mais prático, simples e econômico.

Usualmente, os gráficos apresentam o padrão geral dos dados, deixando os detalhes para as tabelas de freqüências. Para variáveis de nível nominal e ordinal indica-se o uso de gráficos de barra (figura 1a). Em variáveis de nível de medida intervalar e de razão devemos usar o histograma (figura 1b). No histograma a escolha do número de intervalos de classe é fundamental: um número muito pequeno de intervalos de classe e podemos perder informação importante; intervalos de classe em excesso e o padrão básico dos dados ficará obscurecido em uma massa enorme de detalhes. Portanto, na elaboração de um histograma sugere-se a adoção de algo entre 10 a 20 intervalos de classse, mas a escolha final irá depender de uma impressão subjetiva do histograma obtido (2,3).

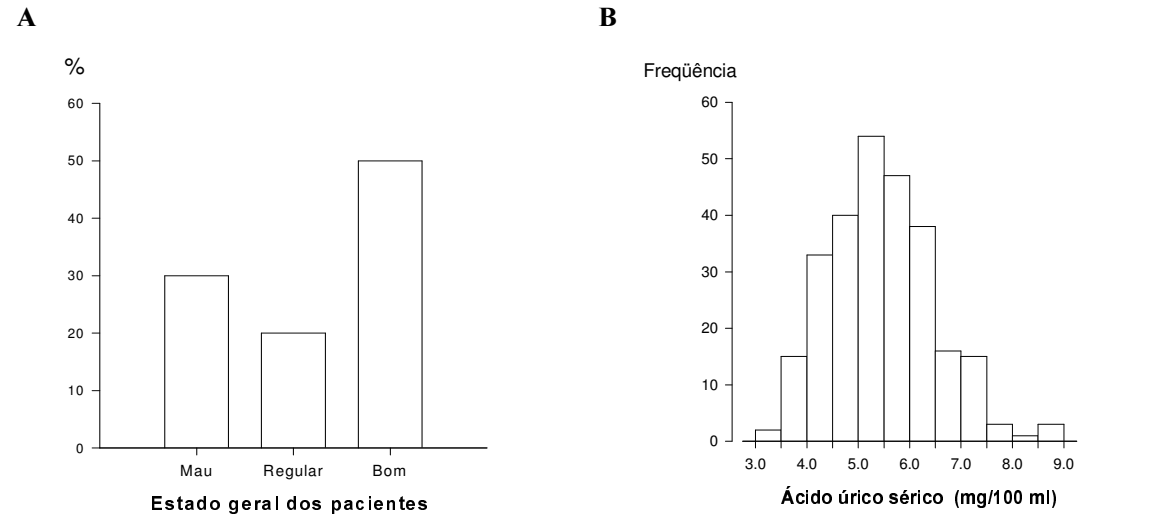

Figura 1: A) Gráfico de barras sobre o estado geral dos pacientes admitidos na emergência do Hospital X; B) Histograma do acido úrico sérico (mg/100 ml) em 267 indivíduos normais.

Sumarizadores numéricos

As medidas de tendência central e as medidas de dispersão podem ser consideradas sumarizadores numéricos. São importantes pelo resumo que fornecem e por possibilitar a utilização de processos e técnicas estatísticas (2,3,6).

Medidas de tendência central

As medidas de tendência central tem como objetivo principal apresentar o "valor típico" dos dados. Abaixo listamos três medidas de tendência central freqüentemente utilizadas.

Média: A média aritmética ou apenas média é simplesmente a soma de todos os dados dividida pelo número de observações. É definida pela fórmula abaixo e lê-se "x barra".

$$\overline{x} = \frac{\sum x}{n}$$

A principal característica da média é que leva em consideração todos os elementos da série e isto a torna particularmente adequada ao processamento estatístico. No entanto, nesta característica reside sua principal desvantagem: a média é influenciada por valores discrepantes. A ocorrência de um único valor discrepante pode alterar substancialmente a média afastando-a de sua função de "valor típico" da série que descreve.

Mediana: A mediana representa aquele valor que divide a série ordenada em duas partes iguais, ou seja, metade dos valores da série estão abaixo da mediana e a outra metade acima. A mediana apresenta a vantagem de não ser efetada por valores discrepantes. No entanto, é considerada menos eficiente no processamento estatístico quando comparada com a média.

Moda: A moda representa simplesmente o valor mais comum. Em tabelas de freqüências e histogramas, onde os dados geralmente estão grupados, a moda corresponde ao valor que tiver a maior freqüência. Distribuições bimodais, ou seja, com dois picos podem assinalar a presença de dois extratos distintos em uma mesma população.

Medidas de dispersão

As medidas de dispersão se referem a quanto os dados de uma série estão agrupados em torno da tendência central

Amplitude: Tecnicamente falando é a diferença entre o valor máximo e o mínimo. Por exemplo, na série 22, 29, 30 e 37 a amplitude é 37 - 22 = 15. Assim, a amplitude é sempre um único valor. Na prática, no entanto, os valores mínimo e máximo são freqüentemente apresentados em seu lugar. Como a amplitude é baseada em somente dois valores (extremos) seu comportamento é muito instável e sujeito a grandes oscilações.

Amplitude entre quartis (AEQ): A amplitude entre quartis é derivada da expressão inglesa *interquartile range* (3). Deve-se relembrar que a mediana divide a série em duas partes iguais, a metade inferior e a metade superior. O quartil inferior, também conhecido como percentil 25 (P25), é a mediana da metade inferior e o quartil superior, ou percentil 75 (P75), é a mediana da metade superior. Assim, a AEQ é a diferença entre P75 e P25 e envolve os dados compreendidos nos 50% centrais da série. Como no caso da amplitude os valores P25 e P75 podem ser apresentados independentemente da própria AEQ. Como lida diretamente com os quartis a AEQ tem menor possibilidade de ser efetada por valores discrepantes do que a amplitude. Por isso, a AEQ é bem mais estável do que a amplitude.

Variância e desvio padrão: São consideradas medidas mais representativas da dispersão ou variabilidade dos dados, uma vez que levam em consideração todos os elementos da série. Para seu cálculo, inicialmente de cada valor x é subtraída a média e a diferença é elevada ao quadrado. A soma é, então, dividida por (n-1), uma quantidade também conhecida como graus de liberdade. Este processo gera um número chamado variância (s^2).

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Infelizmente, ao elevarmos as diferenças ao quadrado (para evitar sinais negativos) a variância final da série fica com sua unidade ao quadrado. Por exemplo, nos dados da tabela 1, s^2 = 1,11 mg/100 ml^2 o que torna difícil compreender-se o que seriam ml^2. Para retornar a unidade original extrai-se a raiz quadrada da variância e obtem-se o desvio padrão (*s* ou *DP)*.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Quando todos os elementos da série são iguais entre si, eles serão iguais à média e o DP = 0. Por outro lado, se os dados estiverem muito dispersos em torno da média o DP será grande. É desta forma que o DP reflete a variabilidade dos dados. Por exemplo, nos dados de ácido úrico o DP = 1,05 mg/100 ml. Em muitas séries de dados acontece de 95% das observações estarem situadas entre a média e dois DP para a esquerda $(\bar{x} - 2DP)$ e a média e dois DP para a direita $(\bar{x} + 2DP)$. Este intervalo é usualmente escrito da seguinte forma: $\bar{x} \pm 2DP$. Esta característica do DP o torna muito útil nas investigações feitas na área biomédica. Isto implicaria, como pode se observar na tabela 1, que a grande maioria dos dados de ácido úrico estaria situada entre 5,42 ± 2 x 1,05, ou seja, entre 3,32 e 7,52 mg/100 ml.

Gráficos de barra de erro e *box plots*

Um tipo de gráfico utilizado em muitas publicações na área biomédica para descrever os dados é o gráfico de barras de erro (figura 2a). Este tipo de gráfico combina uma medida de tendência central - a média - e outra de dispersão, frequentemente o DP (2,3). Em algumas situações podem ser utilizados em conjunto com a média e o erro padrão e o intervalo de confiança, que serão abordadas em outros artigos desta série.

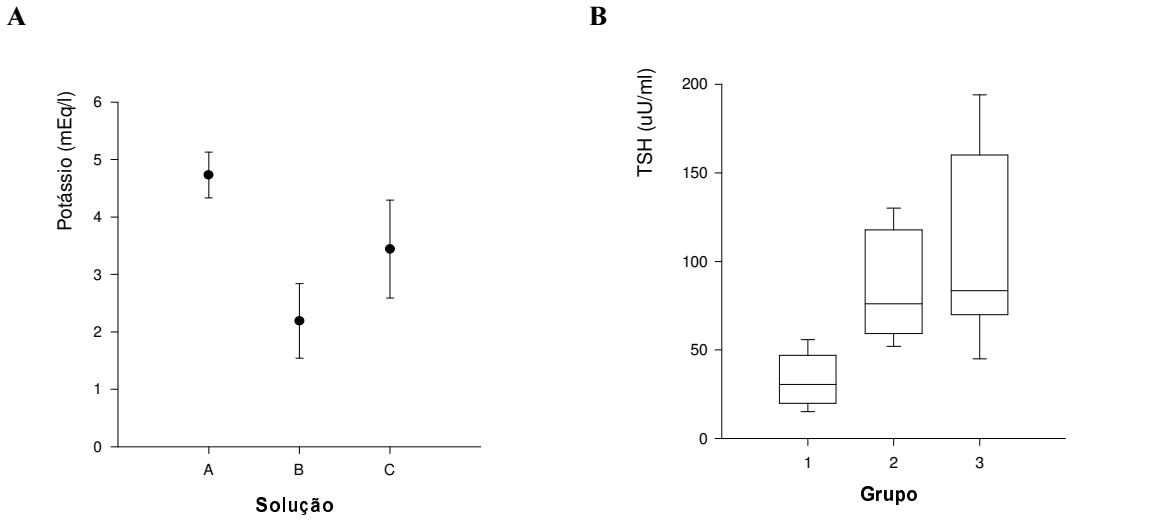

Figura 2: A) Gráfico de barras de erro para a concentração de potássio (mEq/l) em três soluções de nutrição parenteral; B) Box plot de concentrações séricas de TSH (µU/ml) em três grupos de pacientes.

O *box plot* foi descrito pela primeira vez por John Tukey em 1977 (6). É um tipo de gráfico que objetiva apresentar diversas informações sobre o comportamento dos dados e ainda manter uma forma compacta (figura 2b). A mediana da série é representada pela linha horizontal central da caixa (*box*) e os quartis inferior (P25) e superior (P75) pelas linhas inferior e superior que delimitam a caixa.

Esta parte central do gráfico fornece várias informações sobre os dados. A mediana apresenta um estimativa de tendência central; a altura da caixa (AEQ = P75 - P25) é uma estimativa da variabilidade; e a posição da mediana (central ou mais próxima a um dos quartis) indica a presença ou não de assimetria nos dados. As linhas verticais que saem da caixa foram chamadas por Tukey de *whiskers* (significando bigodes de gato em inglês). Em muitos casos, principalmente em séries simétricas e que seguem uma distribuição normal (distribuição teórica freqüentemente encontrada em variáveis intervalares/razão), os *whiskers* representam os valores mínimo e máximo. Nas séries assimétricas os *whiskers* são determinados através de frações da AEQ. Valores discrepantes a tal ponto que ultrapassem os limites estabelecidos por P25 - 1,5AEQ ou P75 + 1,5AEQ são assinalados separadamente no gráfico (2, 6).

Outras fontes de variabilidade

A variabilidade expressa pelo desvio padrão é usualmente denominada variabilidade entre indivíduos. Entretanto, se realizarmos repetidas mensurações em um mesmo indivíduo temos o que se chama de variabilidade entre observações. Este fenômeno geralmente acontece quando realizamos medidas seriadas em um grupo de pacientes.

Adicionalmente podemos ter a variabilidade induzida pelo processo de mensuração, entre observadores, por erro laboratorial ou até de transcrição da informação. Por isso, na investigação da

variabilidade total de um estudo é importante distinguir-se e localizar as diferentes fontes de variabilidade. O pesquisador deve estar consciente dessas fontes e delinear seu estudo e executar as análise de forma adequada.

Considerações finais

Não existe uma técnica estatística que seja aplicável a todas as situações. No que se refere às medidas de tendência central e dispersão devemos ressaltar alguns pontos importantes. Quando os dados forem simétricos e com distribuição próxima à Normal (por exemplo, de forma semelhante ao histograma da figura 1b) a média é na maioria das vezes a melhor medida de tendência central acompanhada do desvio padrão como medida de dispersão. Nesta situação os dados são freqüentemente apresentados como $\bar{x} \pm DP$. Quando outra medida de dispersão for utilizada no lugar do desvio padrão isto deve ser salientado no texto. No caso de dados assimétricos utiliza-se a mediana uma vez que ela não sofre a influência de dados discrepantes. Como medida de dispersão nestes casos são freqüentemente utilizados os valores mínimo e máximo e, mais recentemente, por influência das técnicas propostas por Tukey estão sendo adotados a AEQ ou P25 e P75.

No que se refere aos gráficos devemos atentar fundamentalmente para que as escalas estejam claramente definidas e as unidades especificadas; o eixo dos *y* sempre que possível deve mostrar o valor zero, pois uma falsa impressão de crescimento ou decréscimo pode ocorer quando o zero é omitido do eixo dos *y*; evitar gráficos muito densos com informação demasiada e sempre que possível fornecer medidas apropriadas de variabilidade dos dados.

Enfim, pode-se dizer que os chamados procedimentos e testes paramétricos baseados na média e no desvio padrão devem ser utilizados em variáveis com níveis de medida intervalar e de razão. Por outro lado, quando temos variáveis nominais e ordinais devemos, estritamente falando, utilizar procedimentos e testes não-paramétricos (2 ,3).

Referências

1. Knapp RG & Miller III MC. Clinical Epidemiology and Biostatistics. Malver: Harwarl Publishing Company; 1992.
2. Norman GR & Streiner DL. Biostatistics - The Bare Essentials. London: Mosby; 1994.
3. Campbell, MJ & Machin, D. Medical Statistics - A Commonsense Approach. Chichester: John Wiley & Sons; 1993
4. Kirkwood, BR. Essentials of Medical Statistics. Oxford: Blackwell Scientific Publications; 1988.
5. Zar, JH. Biostatistical Analysis. London: Prentice-Hall International; 1984.
6. Tukey JW. Exploratory data analysis. Reading: Addison-Wesley; 1977.