Os Caminhos da Estatística e suas Incursões pela Epidemiologia

The Paths of Statistics and its Incursions through Epidemiology

Celia L. Szwarcwald [1]
Euclides A. de Castilho [1]

SZWARCWALD, C. L. & CASTILHO, E. A. de ***The Paths of Statistics and its Incursions through Epidemiology.*** *Cad. Saúde Públ., Rio de Janeiro, 8 (1): 05-21, jan/mar, 1992.*

In this paper the development of Statistics is contemplated from its probabilistic fundamentals until the current studies of time and space "dependence". Some applications of the quantitative method in the epidemiologic approach are evaluated. An attempt is made to establish some limits to the current statistical techniques through the discussion of theoretical assumptions and their adequacy to analyse empirical data. The development (or generalization) of new procedures that could possibly help to overcome methodological difficulties that are still found in various analysis of causal inference in Epidemiology is emphasized.

Keywords: *Statistics; Applied Statistics; History of Statistics; Biostatistics; Statistics/Epidemiology Relationships*

O DESENVOLVIMENTO DA ESTATÍSTICA

A História da Probabilidade

O homem traz consigo a idéia de "chance" desde os mais remotos tempos. Evidências estão nos jogos de aposta, referenciados em toda a história da humanidade, e nos "instrumentos da sorte", encontrados em sítios arqueológicos de grande antiguidade. Imagina-se que a noção intuitiva de probabilidade estaria presente no curso dos jogos, influenciando o apostador nas suas estratégias e decisões (Davis, 1955). No entanto, até meados do século XVI, a grande maioria dos pensadores negava a existência da "chance" nos fenômenos naturais. Mesmo diante do seu reconhecimento, era excluída como objeto do discurso racional. Aristóteles identificava "chance" como "a classe de tudo que é indefinido, inescrutável ao intelecto humano". Na mesma linha de pensamento, séculos mais tarde, o mistério da "chance" ainda era explicado como uma deficiência de nosso conhecimento, que, limitado pela inteligência, era incapaz de apreender todas as causas de ocorrência dos eventos e suas possíveis interações simultâneas (Neuts, 1973).

Os primeiros problemas de probabilidade aparecem no período da Renascença e relacionam-se apenas aos jogos de azar. As soluções da "geometria do dado" são apresentadas por matemáticos franceses no século XVII, destacando-se particularmente Blaise Pascal e Pierre Fermat (Davis, 1955; Kendall, 1956). Utilizando elementos de análise combinatória no cálculo de probabilidades, Jakob Bernouilli dá continuidade a esses estudos. Entre suas contribuições, sobressaem-se a distribuição que leva seu nome e a "lei fraca dos grandes números", mais conhecida como "tentativas independentes de Bernouilli" (Neuts, 1973).

O desenvolvimento do pensamento probabilístico moderno está, sem dúvida, estreitamente relacionado à ascensão do método empírico nas pesquisas científicas. Revolucionando o pensamento de sua época, Francis Bacon, ao final do século XVII, enfatiza o papel da experiência no processo de geração do conhecimento e propõe a indução como método de investigação (Demo, 1989). A necessidade de expressar o

[1] *Fundação Oswaldo Cruz. Avenida Brasil, 4365, Rio de Janeiro, RJ, 21045, Brasil.*

grau de incerteza na ocorrência dos experimentos e de explicar o fato de duas experiências iguais poderem ter resultados diferentes leva ao reconhecimento da racionalidade probabilística em eventos da natureza. A pesquisa em probabilidade no século XVIII culmina com o notável trabalho de Pierre-Simon de Laplace, "Theorie Analitique de Probabilités". À luz da concepção do cientificismo, rapidamente amplia-se o domínio de abrangência do cálculo probabilístico. Este torna-se indispensável para lidar com dados relativos a temas de interesse social e econômico, como administração das finanças públicas, saúde coletiva, conduta de eleições e seguro de vida. Surgem as primeiras idéias do positivismo e Condorcet propõe uma "ciência natural da sociedade", isto é, uma "matemática social" baseada no cálculo das probabilidades (Lowy, 1991).

De Laplace até o início do século XX, pouco se acrescenta à teoria das probabilidades. Os raros avanços estão principalmente relacionados ao desenvolvimento de técnicas estatísticas e à análise de erros experimentais (Neuts, 1973).

Durante a primeira metade do século XX, a preocupação dominante da pesquisa matemática é com o tratamento abstrato e a axiomatização de vários de seus ramos. Após a descoberta de Komolgorov, em 1903, de que a probabilidade poderia ser considerada uma "medida" (em termos matemáticos), os vagos fundamentos teóricos são reformulados sob um outro referencial, a "teoria das medidas", bem mais poderoso conceitualmente (Ash, 1972).

Destacam-se como contribuições da moderna concepção a "lei forte dos grandes números" e a demonstração do "teorema do limite central", por J. W. Lindeberg, em 1922 (Feller, 1968).

No que diz respeito ao campo aplicativo, pouco a pouco os modelos determinísticos são substituídos pelos probabilísticos e tornam-se habituais no estudo de diferentes fenômenos. Introduzida inicialmente na teoria da dinâmica dos gases, a teoria das probabilidades desempenha, hoje, papel importante na física quântica e invade os domínios da teoria atômica (Neuts, 1973).

Em anos mais recentes, a pesquisa na área de probabilidades tem se concentrado no estudo da "dependência". A generalização dos processos de Poisson e das cadeias de Markov dá origem à teoria dos processos estocásticos, cuja amplitude e variedade de aplicações parecem ser inesgotáveis (Narayan Bhat, 1972).

O Objeto da Estatística Através do Tempo

A palavra "estatística" é derivada de *status*, em latim, e significa, na sua origem, o "estudo do estado". Inicialmente, no século XVI, pensada pelos ingleses como uma ciência política, destinava-se a descrever características de um país, tais como população, área, riquezas e recursos naturais (Laurenti et al., 1985; Yule & Kendall, 1950). Deste papel histórico, origina-se a sua função de caracterização numérica de uma série de informações populacionais. Com esta abordagem, o termo é utilizado no plural, como as "estatísticas de saúde", as "estatísticas de mortalidade", as "estatísticas do registro civil", entre outras (Berquó et al., 1984; Yule & Kendall, 1950).

Os estudos desenvolvidos por Pierre-Simon de Laplace e Carl Friedrich Gauss, no início do século XIX, transformam a concepção da Estatística. Com a visão de uma teoria dos erros, passa a ser amplamente aplicada a dados experimentais (Yule & Kendall, 1950). Sistematiza-se a análise dos desvios em relação à média em medidas repetidas de uma quantidade. São elaborados conceitos da teoria da estimação, como o método de mínimos quadrados por Gauss, e o primeiro intervalo de confiança , em 1812, em um trabalho de Laplace (Lehmann, 1959) [Apesar de sua dedução correta, o autor considerava o parâmetro como uma variável ao atribuir-lhe a probabilidade de recair no intervalo. A interpretação apropriada data de um século mais tarde, devida a E. B. Wilson, em 1927, e H. Hotelling, em 1931 (Lehmann, 1959). Desafortunadamente, até os dias presentes, com muita freqüência, o conceito é erradamente aplicado].

Na segunda metade do século XIX, a teoria estatística passa a ser enunciada a partir de generalizações das propriedades observadas em amostras grandes. São pesquisadas famílias de funções matemáticas que se aproximem das distribuições de freqüências empíricas (Steel & Torrie, 1981). Na Alemanha, prioriza-se o estudo pelo coletivo, originando-se os princípios da Estatística Descritiva, ramo da Estatísti-

ca que tem a função de organizar os dados, resumindo-os numa série de medidas, gráficos e tabelas para enfatizar as características essenciais do conjunto (Rankin, 1966). Nomes de destaque desta época são os de Francis Galton e Karl Pearson. O primeiro, por meio de experimentos em Genética, estuda a distribuição normal bivariada, propõe o coeficiente de correlação como medida de associação e descobre algumas características das distribuições condicionais, como a regressão linear e a homoscedasticidade (Anderson, 1958). Por sua vez, Karl Pearson desenvolve a teoria e a aplicação de diferentes tipos de correlação à pesquisa biológica. Seus estudos concentram-se na procura de distribuições teóricas, publicando, em 1900, a famosa estatística qui-quadrado para o teste de adequação dos dados às distribuições de probabilidades. É fundador da revista *Biometrika* e de uma escola de Estatística, vindo estimular a produção de novos conhecimentos na área (Walker, 1958).

Um aluno de Karl Pearson, de nome William S. Gosset, dedica-se ao estudo de pequenas amostras e das distribuições do desvio-padrão, da razão entre a média e o desvio padrão e do coeficiente de correlação amostral. Seus resultados são divulgados na *Biometrika*, em 1908, sob o pseudônimo de Student, porque, por razões contratuais de trabalho, suas publicações não podiam ser individualizadas (Steel & Torrie, 1981).

Por outro lado, problemas conceituais apresentados pelo matemático alemão Wilhelm Lexis colocam em questionamento, na mesma época, o interesse apenas pelo coletivo. Ao estudar anualmente a razão de sexo no nascimento, através de estatísticas vitais, Lexis mostra, por meio de resultados empíricos, a consistência da suposição de que a determinação do sexo é governada por um simples mecanismo de chance, como o procedimento "cara-coroa". Isto renova o esforço à procura de mecanismos de chance atuando nos indivíduos para produzir as observadas características coletivas (Rankin, 1966). Nos anos 20, George Polyá constrói um sistema de mecanismos de chance que pode gerar quase todas as distribuições propostas por Karl Pearson. O objeto da Estatística move-se do estudo do coletivo à construção dos mecanismos de chance, ou dos modelos estocásticos dos fenômenos. Esta idéia é explicitamente expressa por Émile Borel: "O problema básico da estatística matemática é inventar um sistema de simples mecanismos de chance, tais que as probabilidades determinadas por este sistema concordem com as freqüências relativas observadas dos vários detalhes do fenômeno estudado" (Rankin, 1966). No decorrer do século XX, o campo indicado pela definição de Borel cresce em importância, concomitante à produção de considerável literatura em processos estocásticos, constituindo-se, atualmente, em um dos capítulos da teoria das probabilidades (Feller, 1968).

Inferência Estatística: um Produto do Século XX

Enquanto a concepção estatística dos sistemas de mecanismos de chance caía em processo de desuso, esforço crescente era atribuído aos problemas de estimação e à dedução das distribuições de probabilidades, sobressaindo-se notavelmente a obra de Ronald A. Fisher (Hotelling, 1951). São devidas a ele várias contribuições de uso atual e amplamente divulgadas, entre elas o método da estimação por máximo-verossimilhança e a distribuição da razão entre variâncias, denominada posteriormente por G. W. Snedecor distribuição "F", em sua homenagem (Remington & Schork, 1970). Fundamentando-se no princípio da aleatorização à experimentação agrícola, Fisher desenvolve as bases dos "desenhos de experimentos". Problemas de classificação em Botânica o levam à proposição da função discriminante, em 1936. No livro clássico de C. Radhakrishna Rao, há mais de vinte citações referentes à sua autoria de procedimentos de estimação e análise (Rao, 1973).

Simultaneamente aos progressos na teoria da estimação, o pensamento estatístico da primeira metade do século XI tem seu interesse voltado à solução dos problemas de testes de hipóteses.

Referências vagas à "significância" datam dos séculos XVIII e XIX. Em 1900, Karl Pearson utiliza o conhecido teste qui-quadrado. Porém, somente em 1928 são introduzidos os conceitos de erro de primeira e segunda espécies, por Jerzy Neyman e Egon S. Pearson. Primeiros a

reconhecer que a decisão de um teste deve envolver considerações não só sobre a hipótese, mas também sobre as alternativas, estes dois autores tiveram marcante influência nos rumos da Estatística contemporânea (Lehmann, 1959).

Em meados dos anos 30, não fugindo ao tratamento axiomático da Matemática a todos os seus ramos, é dada à Estatística nova formulação teórica. J. Neyman e E. S. Pearson apresentam a teoria da inferência estatística, em 1936, apta a considerar os testes de hipóteses com a precisão e o rigor impostos pela Matemática moderna (Lehmann, 1959). De alta repercussão acadêmica, a teoria matemática de Neyman-Pearson vem a referendar o campo de pesquisa teórica, a Estatística Matemática, tratada como uma disciplina matemática na qual a probabilidade é a ferramenta básica (Hoel, 1980). Os testes de hipóteses são apreciados, à luz da teoria dos jogos, pioneiramente por Abraham Wald, em 1940. Reconhecendo as vantagens do ponto de vista conceitual, estende a abordagem da teoria dos jogos, originalmente proposta para aplicações em Economia, ao domínio estatístico. Assim generalizada, passa a ser denominada teoria da decisão (Fergunson, 1967). Utilizando a linguagem de jogos, o espaço dos parâmetros populacionais a serem testados é o conjunto dos possíveis resultados de um jogo, enquanto as decisões estatísticas são as alternativas ou estratégias do jogador. Busca-se a "melhor" opção através do conhecimento adquirido com informações pesquisadas por meio da experimentação. A qualificação de "melhor" tem o sentido de minimizar a probabilidade de erro (a perda) conseqüente à decisão tomada (Fergunson, 1967). Outro grande legado de A. Wald é a chamada análise seqüencial, muito utilizada em problemas que envolvem controle de qualidade (Wolfowitz, 1952).

A Importância da Amostragem

A influência da inferência estatística extravasa o plano teórico. A união da velha estatística à nova teoria probabilística amplia sobremaneira a sua aplicação à análise de dados empíricos. Agora é possível responder a questionamentos relativos a parâmetros populacionais através de um pequeno subconjunto, a amostra.

Em procedimento tipicamente indutivo, chegando-se a conclusões sobre uma população a partir do estudo de uma amostra, a técnica de amostragem torna-se essencial. Surge o problema de selecionar uma amostra, o mais representativa da população total, diante das limitações de custos e das possibilidades de perda de precisão na estimativa dos parâmetros.

As técnicas de amostragem estão indispensavelmente vinculadas ao nome de W. G. Cochran, que as sistematizou em 1953 (Cochran, 1953). Embora de freqüente emprego em investigações populacionais, nem sempre o tratamento analítico dos dados é adequado ao tipo de procedimento utilizado para a seleção das unidades experimentais, resultando em sérios vieses de interpretação. Com esta perspectiva, um seguro objeto de estudo da Estatística aplicada nos próximos anos será o desenvolvimento de métodos de estimação e inferência compatíveis com as diferentes técnicas de amostragem. Vale insistir que esta questão não vem recebendo a devida consideração e são inúmeros os exemplos de inferências incorretas, conseqüentes ao corriqueiro tratamento de que sempre está-se diante de amostras aleatórias simples.

A Estatística Recente

A partir dos anos 40, a pesquisa estatística se volta para solucionar problemas envolvendo variados aspectos da inferência, cada um tendo a sua aplicação a situações específicas. Os testes de hipóteses para médias, variâncias e proporções, a teoria dos testes uniformemente mais poderosos, o processo de inclusão (exclusão) de variáveis nos modelos de regressão são algumas das formas de inferência de uso consagrado (Rao, 1973).

Nesta mesma linha, encontram-se os "métodos não paramétricos", mais apropriadamente denominados "livres de distribuição". Constituem-se em testes de hipóteses cuja aplicação independe dos pressupostos teóricos da estatística paramétrica, inclusive no que diz respeito à distribuição da variável aleatória em estudo. Apesar de apresentarem as vantagens de suposições teóricas mais flexíveis, os testes não paraméticos podem, por vezes, ser pouco sensíveis, deixando passar desapercebidas

características quantitativas importantes das informações (Rao, 1973; Remington & Schork, 1970).

Estimulada pelos seus campos de aplicação, ao lado das facilidades de processamento introduzidas pela informática, a Estatística tem enfatizado, ultimamente, o desenvolvimento dos procedimentos multivariados. Classicamente baseados na distribuição multinormal, expandiram-se anos mais tarde também à função multinomial (Anderson, 1958; Bishop, Finberg & Holland, 1975). O conceito matemático de "combinação linear" é introduzido para descrever as relações entre uma variável resposta e um conjunto de variáveis independentes ou explicativas. Entre os modelos mais conhecidos estão os de regressão múltipla, análise de variância e covariância e a função discriminante. No caso de multiplicidade de respostas, as principais técnicas são as de correlação canônica, de discriminação de vários grupos e de análises de variância e covariância multivariadas (Green, 1978; Searl, 1971).

Nos anos 70/80, são propostos os modelos log-lineares para a análise de dados categóricos, onde os logaritmos das probabilidades dos estados multinomiais são expressos como combinação linear de efeitos principais e de interação entre os fatores (Bishop, Finberg & Holland, 1975; Haberman, 1978). Capaz de lidar com os dois tipos de variáveis independentes, contínuas e discretas, a regressão logística representa o logito da probabilidade condicional do sucesso de uma resposta binária como uma função linear (Cox, 1970). Embora de formas diferentes, todos estes modelos enfocam aspectos de explicação para uma variável considerada como dependente de outras. Já os procedimentos multivariados de análise fatorial, componentes principais, análise de correspondências e análise de conglomerados têm abordagem diferente. A ênfase é dada à análise de interdependência no conjunto total de variáveis (Green, 1978). Os três primeiros são denominados redutores do espaço multivariado, pois têm o objetivo de representar as informações originais por meio de um número menor de variáveis que o considerado inicialmente. A análise de conglomerados também é um procedimento simplificador, porém, neste caso, a redução procede-se no número de objetos e não nas dimensões do espaço (Green, 1978).

De maneira bem resumida, o temário da análise multivariada pode ser assim subdividido: de mensuração da dependência entre variáveis; de analogia à inferência univariada; de redução das dimensões do espaço; de classificação e agrupamento das unidades experimentais (Anderson, 1958). Tais métodos se propõem a analisar observações coletadas num corte de tempo. A interpretação corresponde, assim, à imagem das observações num dado momento, sem apreender sua evolução temporal.

Sob a consideração de que a explicação de certos fenômenos envolve o estudo do seu acompanhamento temporal, uma das vertentes da pesquisa estatística atual objetiva a proposição de modelos que incluam a possibilidade de análise da "dependência no tempo". Neste sentido, desenvolvem-se os modelos de séries temporais, com o reconhecimento explícito da importância da seqüência das observações no tempo. No caso de uma estrutura probabilística, isto é, as flutuações irregulares apresentarem propriedades estatísticas de variabilidade, as séries constituem-se em processos estocásticos. As informações sucessivas são dependentes das anteriores, fazendo-se necessária a introdução de novos conceitos, como o de auto-correlação para medir a dependência de observações da mesma variável em tempos diferentes (Anderson, 1971). Embora haja o reconhecimento geral de sua importância, as séries temporais ainda possuem domínio restrito de aplicação. Sua utilização tem sido limitada à interpretação de séries econômicas, com propósitos predominantemente preditivos.

Os estudos da dependência no tempo inspiram os adeptos da Geografia Quantitativa às análises da dependência no espaço. A produção de métodos é acelerada graças à constatação que as técnicas estatísticas convencionais, baseadas na independência das unidades experimentais, mostram-se impróprias ao tratamento dos dados geográficos que exibem tipicamente ordenação sistemática no espaço (Hammond & McCullagh, 1978; Johnston, 1978). Dada a similaridade dos problemas de dependência nos

domínios do tempo e do espaço, muitos dos métodos de inferência temporal têm sido adaptados para análise das distribuições espaciais. Entretanto, enquanto a medida de auto-correlação no tempo é um problema unidimensional, a interdependência entre observações espaciais pode ser multidimensional, resultando em questões bem mais complexas e ainda não de todo resolvidas (Hammond & McCullagh, 1978). Mais recentemente, a articulação do interesse econométrico na dependência temporal e do geográfico na dependência espacial origina a elaboração de séries espaço-temporais que incluem parâmetros que variam em ambos os domínios (Cliff & Hagget, 1979; Raubertas, 1988; Tango, 1984).

O Paradoxo Estatístico

Embora de uso amplamente estabelecido, a teoria preconizada por J. Neyman e E. S. Pearson é até hoje geradora de controvérsias. Muitos estatísticos de renome, desde a elaboração conceitual dos testes de hipóteses, questionam a validade do estabelecimento de um nível de significância como forma de decisão (Rao, 1973). Os debatedores argumentam que a decisão estatística é tomada sem levar em consideração a probabilidade *a priori* da hipótese nula (Fisher, 1956; Jeffreys, 1948; Savage, 1954).

A contradição entre o procedimento de inferência e a existência de uma distribuição *a priori* da hipótese nula fica evidente no trabalho de Lindley, denominado pelo próprio autor como o "paradoxo estatístico" (Lindley, 1957). Por meio do teste habitual para a média de uma distribuição normal, considerando uma amostra aleatória de tamanho "n", Lindley demonstra que um determinado valor de "n" pode ser sempre encontrado tal que:

a) O valor da média é significativamente diferente ao proposto na hipótese nula ao nível de α %;

b) A probabilidade *a posteriori* de que a hipótese nula é verdadeira é (100 - α)%.

Este é o paradoxo. Sendo α pequeno, por exemplo 5%, a interpretação do primeiro resultado é decidir que a média é significativamente diferente do valor especificado na hipótese nula, enquanto pelo segundo existem boas razões de se acreditar na igualdade (Lindley, 1957). Indaga-se, então, o porquê do uso consagrado do nível de significância em papel decisório. A resposta é dada também por Lindley, que demonstra que para a suposição da probabilidade a priori igual a 50%, o paradoxo só vem a ocorrer para amostras relativamente grandes (Lindley, 1957). O problema trazido à compreensão dos usuários da área de saúde é muito bem examinado por Browner e Newman (Browner & Newman, 1987). A analogia é feita a um teste de diagnóstico cujos resultados podem ser positivos ou negativos. A veracidade das hipóteses nula e alternativa correspondem à ausência e à presença da doença,respectivamente. A probabilidade de rejeição da hipótese nula quando ela é verdadeira (o nível de significância) é relacionada à falso-positividade, enquanto o poder do teste, à sensibilidade. Como nos testes de diagnóstico, os autores apontam as vantagens da análise bayesiana na interpretação dos resultados, baseados nos seguintes fatos: os valores do nível de significância descritivo ("p") podem ser maiores do que 5%, mas produzirem valores preditivos sugestivos de que a hipótese nula é falsa; os valores de "p" podem ser menores do que 5%, mas não se mostrarem aptos a estabelecer a veracidade da hipótese alternativa.

Desde a avaliação crítica da teoria de Neyman-Pearson, propostas alternativas têm sido elaboradas para o tratamento dos testes de hipóteses, constituindo-se nas denominadas escolas de inferência estatística (Oakes, 1990). Entre as principais está a fisheriana, cuja argumentação é baseada na probabilidade fiducial e que também tem sido sujeita a diversas objeções (Rao, 1973). O desenvolvimento da escola bayesiana, em época mais recente, expõe novamente ao debate os fundamentos da inferência estatística (Phillips, 1973).

As Ilusões da Estatística

As estatísticas há muito ultrapassaram o domínio da ciência. Utilizadas por toda parte, são muitas vezes enganosas, dependendo do propósito com que estão sendo abordadas. Apresentadas pela mídia na intenção de impres-

sionar o espectador, são calculadas freqüentemente de maneira inadequada. É o caso, por exemplo, da taxa de acidentes de trânsito fatais dada por unidade de tempo e não pelo número de habitantes da população.

Muitas vezes, com propósitos de mascarar certos aspectos das informações, as medidas de tendência central são escolhidas intencionalmente. São os casos clássicos do emprego da mediana, quando não se deseja levar em consideração os valores extremos das observações, e da média geométrica, para produzir um indicador de menor magnitude que o aritmeticamente calculado. Um fato que ficou conhecido no Brasil, no governo Figueiredo, em 1983, foi a decisão de que o índice nacional de preços ao consumidor (INPC) passaria a ser estimado como média geométrica dos seus componentes, produzindo, desta forma, um número (ilusoriamente) mais baixo do que aqueles anteriormente usados.

Artifícios de representação também podem ser realizados através de procedimentos gráficos. Para enfatizar uma tendência crescente em um sistema cartesiano, basta comprimir a escala horizontal e ampliar a vertical que a visão de aclive será muito mais acentuada (Remington & Schork, 1970). A este respeito, Huff apresenta diversas situações que conduzem a enganos de interpretação (Huff, 1954).

Contudo, a estimativa de estatísticas de maneira incorreta nem sempre é intencional, ocorrendo, em algumas ocasiões, por falhas nas informações em que são baseadas. Diante do desconhecimento da existência de subenumeração do número de nascidos vivos nos censos decenais, por exemplo, a taxa de natalidade do Brasil seria subestimada se calculada a partir dos dados censitários publicados pela FIBGE.

Vieses de interpretação na investigação científica são também raramente propositais. Decorrem, geralmente, pelo desenho inapropriado do experimento, inadequação do método de análise ou pela superficialidade na explicação dos resultados. Vários periódicos médicos apresentam artigos de revisão sobre trabalhos publicados que contêm aplicação de técnicas estatísticas a estudos clínicos. Uma ampla pesquisa, por exemplo, foi organizada pelos editores do *New England Journal of Medicine*. O estudo teve o objetivo de determinar os métodos estatísticos utilizados e se estavam sendo apropriada e corretamente aplicados. Em uma análise de mais de mil artigos publicados na revista, mostrou-se o uso insuficiente das técnicas multivariadas e da modelagem estatística; que o poder dos testes de hipóteses foi apresentado em somente 2% dos trabalhos analisados; e a necessidade de maior divulgação das técnicas estatísticas para a seleção mais adequada do método de análise (Bailar & Mosteller, 1986).

No que concerne à utilização da Estatística para demonstração de uma hipótese por meio da experimentação, é preciso ressaltar que a estatística não "prova" nada. Através de seus procedimentos descritivos, estimadores e inferenciais, ela apenas auxilia o pesquisador a tomar uma decisão. Um dos grandes mitos da Estatística é o nível de significância descritivo do teste, o valor de "p". A ele atribui-se tanto o papel de demonstrador matemático-empírico como o de destruidor de teorias, sem que sejam observados o tamanho da amostra, o poder do teste ou a probabilidade *a posteriori* da hipótese nula ser verdadeira (Greenland, 1988). Desde que as estatíticas de decisão são função crescente do número de observações, quanto maior o tamanho da amostra, maior a probabilidade de rejeição da hipótese. Sendo assim, as formulações das hipóteses nula e alternativa é que devem governar o delineamento da investigação, o tamanho da amostra e o procedimento de coleta das informações. Esses, por sua vez, conduzem à escolha do método adequado de análise.

Todavia, ainda que toda a análise quantitativa tenha sido procedida corretamente, os resultados devem ser sujeitos à contemplação cautelosa. Embora significativos estatisticamente, podem não seguir nenhuma lógica de explicação. A Estatística não é a "benção final" das evidências encontradas na pesquisa. Pelo contrário, o maior poder da metodologia estatística reside em tirar dos dados o seu máximo potencial de informação. Acredita-se que os procedimentos descritivos do comportamento de cada variável e a compreensão da estrutura de interdependência, constituindo-se no que se chama "o entrar nos dados", em permanente

referência à natureza do objeto em estudo, são os passos mais importantes na análise interpretativa dos resultados de um experimento.

A ESTATÍSTICA NA EPIDEMIOLOGIA

As Estatísticas Demógrafo-Sanitárias

O sistema atual de registro civil é resultante de um processo evolutivo que se inicia com a transcrição de dados de batizados, enterros e casamentos pelo clero nos registros paroquiais (Laurenti et al., 1985). Em princípios do século XVI, em função da epidemia da peste, os registros de mortes semanais tornam-se obrigatórios em Londres. Aos poucos, óbitos por outras causas também são incluídos e o sistema é estendido a todas as paróquias da Inglaterra (Pollard et al., 1974). Transformados em séries mais regulares no século seguinte, fundamentam os estudos de John Graunt, primeiro a perceber a importância da análise quantitativa dos eventos vitais. Na publicação *Observations upon the bills of mortality*, em 1662, Graunt introduz o princípio da razão de regularidade estatística, observa uma razão de sexo ao nascimento constante, reconhece padrões sazonais e diferenças urbano-rurais no comportamento das taxas brutas de mortalidade e tem o mérito de construir a primeira tábua de vida. William Petty converte seu trabalho nas bases da "aritmética política", que pouco a pouco passa a ser conhecida como Demografia (Laurenti et al., 1985; Pollard et al., 1974).

Somente a partir do século XIX, quando a responsabilidade do registro dos eventos vitais transfere-se da Igreja para o Estado e estabelece-se, de forma legal, a sua obrigatoriedade em vários países, são impulsionados os estudos demográficos. Surgem também as primeiras análises de morbidade na Inglaterra e nos Estados Unidos, introduzindo-se a abordagem de doenças pelo método quantitativo (Barreto, 1990). Em 1839, William Farr, na função de compilador do sistema oficial de registros na Inglaterra, estabelece a coleta sistemática de informações sobre morbidade e mortalidade (Laurenti et al., 1985). Primeiro estatístico médico, Farr faz uso do registro civil para o estudo de doenças e propõe uma forma de classificá-las com uniformidade internacional (OMS, 1978).

Desde Farr até os dias de hoje, vários indicadores e procedimentos de análise foram desenvolvidos com o objetivo de traçar o perfil nosológico de uma população. Atualmente, esta tarefa é de competência da Estatística Demógrafo-Sanitária, mais conhecida como Estatística Vital, embora esta última denominação não esteja de acordo com a definição das Nações Unidas, que lhe atribui somente o tratamento dos eventos vitais (Laurenti et al., 1985). De certa forma, constitui-se na estatística descritiva da saúde, tendo a função de construir medidas numéricas que caracterizem séries de dados vitais (nascimentos, óbitos e perdas fetais) e de informações relativas a doenças e a serviços (Laurenti et al., 1985). A construção dos indicadores de saúde a partir de dados secundários está relacionada à qualidade dos sistemas de informações. Muitas vezes incompletos e descontínuos, não permitem um adequado tratamento estatístico dos dados.

Os vínculos com a Demografia permanecem estreitos. Em primeiro plano, manifestam-se pelo interesse mútuo nos aspectos dinâmicos das sociedades (fecundidade, mortalidade e migração) e naqueles relativos à composição das populações segundo sexo, idade, situação de domicílio, entre outros. Em segundo, pela necessidade de desenvolvimento de técnicas demográficas, quer seja para estimativas de denominadores das taxas de morbi-mortalidade, quer seja para mensuração indireta de indicadores em populações com sistemas de registro incompletos.

No que diz respeito à abordagem conceitual, o interesse atual tem sido na proposição de indicadores mais sensíveis à percepção da saúde de uma população. Partindo do princípio de que a ausência de doença não implica necessariamente na presença de saúde, alguns pesquisadores dedicam-se a tentativas de definições de saúde no sentido positivo (Goldberg, 1990).

No tocante à metodologia de avaliação das estatísticas demógrafo-sanitárias de uma população, a sua evolução num certo período de tempo encontra instrumental nos procedimentos de séries temporais, que permitem a determinação dos componentes de tendência, periodici-

dade e sazonalidade. Já a análise das distribuições espaciais tem tido aproximações recentes com os modelos utilizados pela Geografia Quantitativa e vem demonstrando interessantes resultados (Breslow & Enstrom, 1974; Cook & Pocock, 1983).

A Epidemiologia e o Método Indutivo Estatístico

O termo Bioestatística aparece primeiramente em 1923, em substituição à expressão "estatísticas vitais" (Berquó et al., 1984). Tem hoje significado mais abrangente e é considerada como a disciplina que trata da aplicação dos procedimentos estatísticos, descritivos e inferenciais aos problemas biológicos (Remington & Schork, 1970). Sua aplicação às ciências médicas é particularmente impulsionada por influência da publicação de Bradford Hill, *Principles of Medical Statistics*, em 1937 (Berquó et al., 1984).

No que se refere à análise de dados epidemiológicos, a história da utilização do método indutivo quantitativo é estreitamente relacionada à questão da causalidade e à forma com que esta é tratada ao longo do tempo. Embora seja atualmente uma das grandes fomentadoras da Bioestatística, a Epidemiologia só vem a adotá-la como metodologia analítica em meados do presente século, a partir da consagração da teoria de multicausalidade (Barreto, 1990).

A abordagem de associações entre fatores ambientais e doença aparece desde o século XIX. Vários pesquisadores, naquela época, além da caracterização quantitativa da situação de saúde de populações selecionadas, analisavam comunidades quanto às suas condições de saneamento, moradia, ocupação e nutrição (Susser, 1985).

Mas as investigações em populações tiveram seu desenvolvimento enfraquecido nas primeiras décadas do século XX. A "teoria do germe" que se impôs sobre a "teoria miasmática" adotou o critério laboratorial como o único válido para a verificação das hipóteses de unicausalidade (Barreto, 1990; Susser, 1985). A quantificação adquire novamente papel importante a partir dos progressos obtidos na concepção da multicausalidade para doenças infecciosas. Surgem os modelos matemáticos contemplando o agente causal e os fatores ambientais relacionados à sua transmissão (Barreto, 1990).

Procurando novos caminhos para ampliar sua capacidade explicativa na determinação das enfermidades, a Epidemiologia encontra na inferência estatística o instrumental adequado para o teste de suas hipóteses. A teoria da decisão enquadra-se perfeitamente no espírito positivista do raciocínio epidemiológico da época, apresentando meios de "provar" empiricamente relações causais conjecturadas teoricamente (Almeida Filho, 1989).

Nos anos 60, os avanços na informática permitem o processamento de grandes massas de dados, estimulando a realização de investigações populacionais. Divulga-se o emprego das técnicas multivariadas, que embora tivessem sido deduzidas na década de 30, só agora podem ser usadas na prática. Surgem *softwares* ditos próprios para o tratamento de informações quantitativas das ciências sociais. Intensifica-se a aplicação dos modelos lineares à interpretação das associações epidemiológicas. Fortalecem-se os laços interdisciplinares, ocorre a chamada "matematização da Epidemiologia" (Almeida Filho, 1989).

A incapacidade interpretativa dos modelos determinísticos causais na explicação das doenças crônicas, em predomínio nos países industrializados, conduz os epidemiologistas à elaboração de novas propostas conceituais e metodológicas. À luz do conceito de risco, ao invés do determinismo do efeito, passa a ser avaliada a probabilidade de ocorrência da doença. São formulados desenhos de estudos alternativos que solicitam procedimentos estatísticos específicos (Breslow & Day, 1980; Breslow & Day, 1987). Para cada delineamento experimental, são criadas técnicas de estimação e análise, a regressão linear é trocada pela logit-linear, a produção de programas para microcomputadores é acelerada.

Nos países centrais, proliferam estudos dispendiosos, com amostras enormes para possibilitar o controle de inúmeras variáveis intervenientes. Em ocasiões não raras, entretanto, a estimativa do risco não se diferencia expressivamente da unidade, ao ponto de se

acreditar convictamente na decisão inferencial de rejeição da hipótese nula. Ao não se conseguir realizar a distinção entre os significados estatístico e epidemiológico da associação, a conduta adotada é a de repetição do experimento para, somente à evidência de respostas semelhantes, estabelecê-la como verdadeira (Knekt et al., 1988; *UK National Case-Control Study Group*, 1989). Muito esforço é consumido para a produção relativamente pobre de conhecimentos.

No decorrer das últimas décadas, os paradigmas da pesquisa epidemiológica têm sido expostos a intensos debates. O estabelecimento da causalidade através dos modelos tradicionais vem sendo colocado em questionamento, principalmente no que diz respeito à compreensão dos problemas de saúde cujos determinantes estão no interior das organizações sociais (Sabroza, 1990). Esta situação, amplamente discutida por diversos autores da América Latina (Sergio Arouca, Jaime Breilh e Asa Cristina Laurell, entre outros), enfatiza o inadequado tratamento de atributos coletivos como sendo passíveis de uma expressão individual (Almeida Filho, 1989; Costa, 1990; Nunes, 1985). É curioso que este reducionismo na prática se faz, na verdade, de modo mais acentuado, pois a quase totalidade dos estudos que se dizem capazes de lidar com a causalidade o fazem com base em procedimentos estatísticos que assumem relações lineares (ou logit-lineares) entre as variáveis.

Os Processos Estocásticos

Já em princípios do século XX, a Epidemiologia buscava na Matemática a solução de seus modelos teóricos de multicausalidade de doenças infecciosas. Ignoradas as variações randômicas e baseando-se na consideração que o processo saúde-doença era governado apenas por leis dinâmicas, surgem os modelos matemáticos determinísticos para representação das epidemias (Bailey, 1964).

Anos mais tarde, com a identificação de que os eventos mórbidos são sujeitos à chance, paralelamente ao avanço na teoria das probabilidades, a modelagem é aperfeiçoada e passam a ser utilizados os processos estocásticos. O uso do adjetivo "estocástico", sinônimo de probabilístico, tem o propósito de enfatizar o aspecto aleatório da ocorrência dos fenômenos, em constraste com as antigas formulações determinísticas. Estas, contudo, são legítimas no caso de populações grandes, quando pode-se assumir que as flutuações estatísticas são suficientemente pequenas para serem ignoradas, além de considerar-se útil a sua abordagem, anterior à probabilística, pela sua capacidade explicativa à dinâmica do processo (Bartlett, 1960).

De maneira formal, um modelo estocástico é aquele que especifica a distribuição de probabilidades de uma variável (vetor) aleatória (o) sobre uma classe de situações de interesse em cada ponto do tempo. A sucessão de estados ou de mudanças, concebida como contínua no tempo, constitui-se no processo estocástico (Iosifescu & Tautu, 1973). Dito estacionário quando a sua estrutura probabilística é constante no tempo, o seu estudo teórico constitui-se num dos temas abordados pelos procedimentos de séries temporais, quando estas são geradas por um modelo subdividido em uma tendência determinística e uma parte aleatória com a propriedade de invariância (Anderson, 1971). Em contraposição está o processo evolucionário, cuja primeira formulação matemática foi realizada por Francis Galton, no final do século XIX, interessado particularmente na probabilidade de extinção das famílias de nobre posição na Inglaterra. Em 1924, G. Udny Yule deduz o "modelo puro de nascimentos-mortes" numa população (Iosifescu & Tautu, 1973).

Desde então, os processos estocásticos têm sido utilizados para representar a evolução de vários fenômenos biológicos, como o crescimento de populações, migração, competição entre espécies, flutuações na composição genética de populações (como mutação e seleção), além dos sistemas fisiológicos de múltiplos compartimentos e dos processos epidêmicos (Iosifescu & Tautu, 1973).

Estes últimos têm sido de interesse permanente para a explicação dos mecanismos de transmissão de certas doenças (Bailey, 1964; Bartlett, 1960; Iosifescu & Tautu, 1973). O grau de complexidade dos modelos depende do número de categorias que compõem a população epidêmica, porém pelo menos dois componentes são sempre necessários, os infectados

e os suscetíveis, cujas relações determinam a dinâmica do processo. A intratabilidade matemática dos modelos mais sofisticados vem sendo superada por procedimentos de simulação.

Atenção tem se dirigido recentemente à modelagem de dinâmica de doenças como a AIDS (Castillo-Chavez, 1989) e aos processos que objetivam descrever a propagação espacial das epidemias (Cliff & Hagget, 1979).

As Medidas de Associação Estatística

A Epidemiologia tem na causalidade,como já dito, uma de suas questões fundamentais. O problema que permanentemente se coloca é o da mensuração das relações causais. Afora a questão da possibilidade de se quantificar os determinantes sociais do processo saúde-doença, mesmo no âmbito da chamada epidemiologia clássica, o seu modo de trabalho com as ditas relações causais merece algumas reflexões a partir do corpo teórico da Estatística. Desde o conceito de probabilidade condicional, passando pelo coeficiente de correlação e pelo qui-quadrado de Pearson até a dependência no tempo e no espaço dos dias de hoje, a preocupação com a "dependência" entre dois atributos tem despertado interesse constante.

Em termos teóricos, duas variáveis são independentes se e somente se a distribuição de probabilidades condicional da primeira, dada a segunda, é igual à distribuição marginal da primeira (Hoel et al., 1971). Esta noção de "dependência" pode ser visualizada através da análise de uma tabela de contingência, quando as variáveis são consideradas associadas se as distribuições multinomiais forem significativamente diferentes para dois níveis da resposta; pode ser traduzida pelo risco relativo ou pelo *odds ratio* iguais a 1 na situação de independência; ou, ainda, na construção da teoria de regressão múltipla no caso de multinormalidade, onde a média da distribuição condicional é um modelo linear das variáveis preditoras e a reta é constante quando há independência.

Um conceito mais intuitivo de mensuração de "dependência" é o de covariância. Tem o sentido de examinar o comportamento conjunto em comparação à multiplicação dos isolados. Se há independência, a covariância é nula (Hoel et al., 1971). As primeiras medidas do grau de dependência entre duas variáveis aleatórias foram propostas através do coeficiente de correlação, descrito como a covariância padronizada pelo produto dos desvios-padrão de cada uma. Pela desigualdade de Schwarz, demonstra-se que seu valor absoluto é limitado pela unidade. A magnitude da associação é, então, medida dentro de um intervalo de extremo inferior zero (nenhuma associação) até o ponto máximo de um (Hoel et al., 1971).

Em 1944, H. E. Daniels dá uma interpretação geométrica da independência, representando-a pela ortogonalidade de dois vetores no espaço euclidiano. Neste contexto, a medida de correlação corresponde ao cosseno do ângulo formado pelos vetores aleatórios em consideração. A associação máxima, quando o cosseno é igual a um, é referida à colinearidade, em oposição à perpendicularidade, situação de cosseno zero e ausência de correlação. Daniels demonstra, ainda, que as medidas de associação tradicionais, como os coeficientes de correlação de Pearson, Spearman e de Kendall, além do coeficiente de contingência média, podem ser expressos por meio de cossenos de ângulos entre vetores de coordenadas convenientemente escolhidas (Daniels, 1944).

Leo A. Goodman é outro autor contemporâneo que contribui expressivamente ao problema de medir associações em variáveis categóricas ordinais. Objetivando captar o efeito da ordenação dos níveis de cada um dos fatores, propõe medidas baseadas na "redução proporcional dos erros" na predição da resposta. Os erros são respectivos a duas situações, a de ausência de informações sobre a variável preditora, relativamente a uma segunda, diante do conhecimento prévio do valor da variável independente (Goodman, 1979).

Na procura de critérios de escolha de medidas de associação adequadas às análises quantitativas das pesquisas sociológicas, Herbert L. Costner, em 1965, propõe adotar aquelas que pudessem ser estabelecidas por meio da redução proporcional no erro de predição (Costner, 1965). É possível demonstrar que a definição geométrica de Daniels, atribuída à correlação (como o cosseno do ângulo formado

pelos vetores aleatórios), tem uma interpretação de "redução proporcional no erro".

Assim, as atuais propostas de estatísticas para medir associações entre variáveis têm sido baseadas na definição de Daniels. Sendo o cosseno de um ângulo em um espaço vetorial expresso como razão de um produto interno dos vetores (covariância) pelo produto das normas (desvios-padrão), as formulações generalizadas têm evoluído em duas direções: convenientes escolhas de funções de coordenadas vetoriais no espaço euclidiano e definição de um produto interno adequado em um espaço de Hilbert (Ash, 1972), possibilitando a extensão para espaços infinito-dimensionais. Esta última aproximação foi considerada por T. W. Anderson no estudo de predição de processos estocásticos estacionários no tempo (Anderson, 1971). É fato por demais conhecido que a significância da correlação estatística é insuficiente para indicar dependência no sentido epidemiológico. Vários autores têm se preocupado inclusive em estabelecer critérios, de tal modo que na ocorrência da associação estatística, seja possível determinar se ela é, de fato, causal (Hill, 1965). Entretanto, os epidemiologistas, perante os problemas de causalidade, têm mostrado atitudes díspares. Não só a significância estatística tem sido apresentada freqüentemente como evidência de uma relação causal, como também à inexistência de correlação estatística, a hipótese epidemiológica é descartada de imediato. Em divergência a estas condutas, é preciso ressaltar que para determinadas distribuições de probabilidades, as variáveis aleatórias podem ser não correlacionadas, mas dependentes (Hoel et al., 1971). Salienta-se, ainda, que é usual considerar as variáveis contínuas como normalmente distribuídas, acarretando em mensurar a associação entre elas por meio de modelos lineares. Desta maneira, se a regressão for quadrática, provavelmente será encontrada uma correlação de baixa magnitude.

Na prática, o que vem ocorrendo é o emprego automático dos modelos multivariados lineares (ou logit-lineares), sem análise prévia ou qualquer representação gráfica das relações de dependência no conjunto de informações. Os testes para correlações parciais das variáveis contínuas ou as estastísticas de máximo-verossimilhança correspondentes à inclusão de variáveis nos modelos logísticos são os critérios estabelecidos pelos epidemiologistas para o julgamento de suas hipóteses. Percorrendo todos os significados das medidas de associação estatística ao longo do tempo, sua interpretação como redução proporcional no erro de predição e suas generalizações, indaga-se o porquê desta utilização tão restrita em vista do leque de possibilidades existentes.

Os Modelos de Regressão

O objetivo de uma análise estatística utilizando a técnica de construção de modelos é, em geral, o de encontrar a melhor adequação (no sentido de minimizar o erro de predição) através do menor número possível de variáveis (Draper & Smith, 1966). Este propósito, no entanto, está longe de satisfazer os objetivos da Epidemiologia na procura dos determinantes ou dos fatores de risco de um problema de saúde. Em primeiro lugar, o princípio da parcimônia, se é conveniente ao intuito preditivo na diminuição dos custos e esforços em obter informações, é, pelo contrário, insatisfatório para uma interpretação plausível das relações entre as variáveis. A economia de variáveis consiste, na verdade, em minimizar o caminho explicativo de um evento ao outro (Li, 1975).

Uma segunda colocação que se impõe refere-se ao fato de que, nos procedimentos de regressão, as variáveis explicativas são tratadas com equanimidade, resultando num modelo em que a resposta é determinada pela adição de efeitos, sem a interpretação do fenômeno. As decisões de inclusão (exclusão) de fatores são puramente estatísticas e, como recomendado em procedimentos com comparações múltiplas, baseadas na diminuição do nível de significância. Ao final de todas as etapas, nada se sabe sobre o poder de cada teste de hipótese causal, muito menos pondera-se sobre suas probabilidades *a priori*. Além disso, em diversas ocasiões, um coeficiente de correlação múltipla baixo é considerado como aceitável, ou seja, grande parte da variabilidade da resposta é atribuída ao acaso.

O método conhecido como a "análise de trajetórias" é uma forma de regressão estruturada onde um diagrama especifica a natureza da

estrutura proposta. É de acordo com este diagrama que a análise subseqüente é realizada (Li, 1975). No caso do desconhecimento prévio do delineamento do circuito causal, vários esquemas podem ser propostos, considerando os possíveis papéis das variáveis como "de confundimento", "intermediárias" ou " modificadoras de efeito" (Breslow & Day, 1980; Morgenstern, 1989). Criado por Sewell Wright, em 1921, para análise de diagramas genealógicos, teve seu emprego divulgado por O. D. Duncan nas ciências sociais (Li, 1975). Sob o nome de "teoria dos grafos", tem vasto campo de aplicação na Pesquisa Operacional, com o objetivo de otimização dos fluxos de organização, como as redes de comunicação e transporte (Berge & Ghouila-Houri, 1962). Apesar de se constituir num procedimento bem mais apropriado para a construção de uma estrutura causal compatível com os dados observados, tem pouca repercussão ainda entre os epidemiologistas.

A Interpretação Estatística de Risco

O conceito de risco, fundamental à Epidemiologia moderna, é definido como "a probabilidade de um indivíduo de uma população vir a desenvolver a doença durante um dado período de tempo" (Morgenstern, 1989). A partir desta concepção probabilística, novas medidas de associação são adotadas, como o "risco relativo" e a "razão dos produtos cruzados" (*odds ratio*). O grau de dependência é avaliado pelo afastamento destas medidas da unidade (Fleiss, 1973). A resposta determinística é transformada numa probabilística, o risco (ou uma função do risco) passa a ser utilizado como variável dependente dos modelos de regressão, a causa torna-se o "fator de risco".

Em virtude de sua fácil interpretação, o modelo logístico tem sido um método de análise amplamente difundido na pesquisa epidemilógica. No caso de uma só covariável, o coeficiente angular da reta corresponde à razão dos produtos cruzados. Extensão feita ao caso politômico, os parâmetros da regressão representam os *odds ratio* em relação a uma categoria de referência (Hosmer & Lemeshow, 1989). Estatisticamente, a variável dependente tem distribuição Bernouilli (ausência ou presença da doença) e a sua esperança condicional, igual à probabilidade do sucesso, é descrita como uma função logística das variáveis preditoras. Sob a suposição de independência das unidades experimentais, os erros do modelo seguem uma distribuição binomial (Hosmer & Lemeshow, 1989).

Desta forma, este processo de "modelagem" dos dados é tipicamente um procedimento de análise de mecanismos individuais independentes que, somando-se, produzem o efeito coletivo. Assinala-se, portanto, novamente o despropósito de incluir nos modelos variáveis mensuradas em grupos (onde as observações podem ser dependentes), fugindo ao pressuposto de independência dos erros da regressão. Ressalve-se, também, que a definição de "grupo de risco" ("grupo populacional em que se encontra um risco relativo de uma dada condição maior do que 1,0") (Almeida Filho, 1989) não tem qualquer suporte na teoria dos modelos estatísticos. Probabilisticamente, "grupo de risco" é a união de indivíduos, supostamente independentes, que apresentam um determinado atributo, chamado "fator de risco" pelos epidemiologistas.

Medidas em Grupos de Observações: a Falácia Ecológica e o Problema da Unidade de Análise

Em análise de correlações entre variáveis relativas a grupos de indivíduos, ao invés dos próprios indivíduos, falsos juízos podem ocorrer se as inferências "entre grupos" (ecológicas) são supostamente válidas para "dentro dos grupos" (Piantadosi et al., 1988). O problema de interpretação na análise das associações ecológicas foi apontado pioneiramente por W. S. Robinson, que lhe deu o nome de "falácia ecológica" (Robinson, 1950). Desde então, esta questão tem sido abordada por diversos autores. Alguns apontam para situações onde sérios erros seriam introduzidos em inferências sobre indivíduos por meio de estudos ecológicos (Morgenstern, 1982). Outros delineiam circunstâncias onde tais inferências estariam justificadas (Richardson et al., 1987).

A relação matemática entre as correlações

ecológica e individual, embora proposta também por Robinson, foi demonstrada apenas recentemente (Piantadosi et al., 1988). Consiste em descrever o coeficiente de regressão entre dois fatores como soma ponderada dos coeficientes angulares "dentro" e "entre" grupos. Assim, comprova-se que na ausência de dados individuais não é possível a estimativa da "verdadeira" associação (a "total") e que apenas na igualdade dos parâmetros "dentro" e "entre" a correlação é expressa pela chamada correlação ecológica.

Porém, este não é o único problema de uma análise ecológica. A questão da modificação do agrupamento de observações é outro ponto para reflexão. Foi identificada por G. U. Yule e M. G. Kendall, em 1950, que assinalaram: "nós não podemos perder de vista que nossos resultados dependem da unidade de análise" (Yule & Kendall, 1950). Em teoria, existe uma infinidade de maneiras na qual uma área pode ser dividida, apesar dos dados serem apresentados para um particular conjunto de subdivisões. Estas podem ser recombinadas de tal forma a constituir regiões numa nova escala. Para cada uma das alternativas, os coeficientes de correlação tomam valores diferentes, acarretando em distintas possibilidades de interpretação. Este é o denominado "problema da modificação da unidade de área", abordado recentemente por S. Openshaw e P. J. Taylor em estudos de distribuições espaciais (Openshaw & Taylor, 1979).

Modelos em Perspectiva

Diante dos problemas metodológicos encontrados para testar muitas das hipóteses de multicausalidade de interesse epidemiológico atual, resta recorrer ao desenvolvimento de modelos estatísticos mais apropriados. Apesar das limitações da Estatística como instrumental analítico dos diversos campos de indagação da Epidemiologia, entende-se que o esforço deverá ser dirigido à procura de modelos que permitam avaliar os agravos de saúde na sua maior complexidade, seja nos mecanismos unitários que produzem as características coletivas, seja nos processos coletivos que influenciam o fenômeno que vem a ocorrer no indivíduo.

Desta forma, vislumbram-se algumas perspectivas, como a análise em desenhos hierarquizados, onde possa ser considerado o nível de atuação de cada variável em estudo. O processo amostral, determinado pela hierarquização dos fatores, seria realizado, então, em quantos estágios se fizessem necessários. Em cada etapa, as unidades experimentais seriam supostamente dependentes, expressando-se a matriz de variâncias-covariâncias do vetor de observações como uma matriz não diagonal, cujos elementos que não pertencessem à diagonal principal (as covariâncias) fossem funções da correlação intra-classe. O progresso da resolução estatística estará em formular a partição da correlação total na estrutura especificada.

Já para os estudos ecológicos, onde a intenção da análise resida apenas nas inferências para as unidades amostradas e não para os indivíduos, é freqüente o interesse pelas representações espaciais (mapas) das patologias. O coeficiente de correlação, como utilizado tradicionalmente "ponto a ponto", não capta os efeitos de aglomeração ou de propagação dos fenômenos. Releva-se, deste modo, a generalização dos processos estocásticos no domínio do tempo para o domínio do espaço, elaborando métodos de estimação de medidas de associação entre distribuições espaciais (Clifford et al., 1989).

No mesmo contexto, uma outra possibilidade é a construção de coeficientes de correlação em espaços de Hilbert, conforme já referido, mediante a definição adequada de um produto interno. Neste caso, a extensão da teoria de regressão entre modelos temporais para modelos espaciais seria realizada por meio da escolha de um eixo direcional unidimensional, como, por exemplo, a distância dos pontos do espaço a um determinado ponto considerado como origem.

Diante do propósito contínuo de elaboração de modelos que traduzam o real à linguagem matemática, acredita-se que uma outra possível vertente de pesquisa estatística será a procura de modelos que contemplem a compreensão do processo evolutivo a que estão sujeitas as distribuições dos fenômenos.

Por outro lado, a abrangência do comportamento temporal dos mecanismos explicativos

aliados à chance gera modelos cada vez mais complexos. Entende-se,portanto, que um dos rumos a ser seguido é a procura de instrumental, no interior da própria Matemática, que venha a simplificar a resolução de tais problemas.

AGRADECIMENTOS

A autora CLS agradece à OPAS, especificamente ao Dr. Moises Goldbaum, por ter concedido a oportunidade de sua participação no curso *Advanced Statistical Methods in Cancer Epidemiology-IARC*, 1989, que forneceu subsídios para a elaboração de parte deste trabalho, sobretudo nos itens referentes aos modelos estatísticos utilizados atualmente pela Epidemiologia.

RESUMO

SZWARCWALD, C. L. & CASTILHO, E. A. de **Os Caminhos da Estatística e suas Incursões pela Epidemiologia.** Cad. Saúde Públ., Rio de Janeiro, 8 (1): 05-21, jan/abr, 1992.

Neste trabalho, contempla-se o desenvolvimento da Estatística, desde suas origens probabilísticas até os atuais modelos de "dependência" no tempo e no espaço. Avalia-se a evolução do método quantitativo na abordagem epidemiológica, como também procura-se estabelecer limites das técnicas estatísticas habituais, discutindo-se suas suposições teóricas e sua adequação ao tratamento analítico das informações. Enfatizam-se a importância do desenvolvimento e/ou generalização de procedimentos que possam ajudar a superar as dificuldades metodológicas ainda encontradas em diversos estudos de inferência causal em Epidemiologia.

Palavras-Chave: Estatística; Estatística Aplicada; História da Estatística; Bioestatística; Relações Estatística/Epidemiologia

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA FILHO, N., 1989. *Epidemiologia sem Números (Introdução Crítica à Ciência Epidemiológica)*. Rio de Janeiro: Editora Campus.

ANDERSON, T. W., 1958. *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons.

_____, 1971. *The Statistical Analysis of Time Series*. New York: John Wiley & Sons.

ASH, R. B., 1972. *Real Analysis and Probability*. New york: Academic Press.

BAILAR, J. C. & MOSTELLER, F. (Ed.), 1986. *Medical Uses of Statistics*. Waltham, Massachussets: NEJM Books.

BAILEY, N. T. J., 1964. *The Elements of Stochastic Processes with Applications to the Natural Sciences*. New York: John Wiley & Sons.

BARRETO, M. L., 1990. A Epidemiologia, sua história e crises: notas para pensar o futuro. In: *Epidemiologia Teoria e Objeto* (D. C. Costa, org.), pp. 19-38, São Paulo: Hucitec-Abrasco.

BARTLETT, M. S., 1960. *Stochastic Population Models in Ecology and Epidemiology*. London: Methuen.

BERGE, C. & GHOUILA-HOURI, A., 1962. *Programmes, Jeux et Réseaux de Transport*. Paris: Dunod.

BERQUÓ, E. S.; SOUZA, J. M. P. & GOTLIEB, S. L. D., 1984. *Bioestatística*. São Paulo: E.P.M..

BISHOP, Y.; FINBERG, S. & HOLLAND, P., 1975. *Discrete Multivariate Analysis*. Cambridge: MIT Press.

BRESLOW, N. E. & DAY, N. E., 1980. *Statistical Methods in Cancer Research v.1 - The Analysis of Case-Control Studies*. IARC scientific publication nº 32, Lyon, International Agency for Research on Cancer.

_____, 1987. *Statistical Methods in Cancer Research v.2 - The Design and Analysis of Cohort Studies*. IARC scientific publication nº 82, Lyon, International Agency for Research on Cancer.

BRESLOW, N. E. & ENSTROM, J. E., 1974. Geographic correlations between cancer mor-tality rate and alcohol-tobacco consumption in the United States. *Journal of the National Cancer Institute*, 53: 631-639.

BROWNER, W. S. & NEWMAN, T. B., 1987. Are all significant "p" values created equal? The analogy between diagnostic tests and clinical research. *Journal of the American Medical Association*, 257: 2459-2463.

CASTILLO-CHAVEZ, C. (Ed.), 1989. *Mathematical and Statistical Approaches to AIDS Epidemiology*. Berlin: Springer-Verlag.

CLIFF, A .D. & HAGGET, P., 1979. Geographical aspects of epidemic diffusion in closed communities. In: *Statistical Applications in the Spatial Sciences* (N. Wrigley, ed.), pp. 5-44, London: Pion Limited.

CLIFFORD, P.; RICHARDSON, S. & HEMON, D., 1989. Assessing the significance of the correlation between two spatial processes. *Biometrics*, 45: 123-134.

COCHRAN, W. G., 1953. *Sampling Techniques*. New York: John Wiley & Sons.

COSTA, D. C. (Org.), 1990. *Epidemiologia Teoria e Objeto*. São Paulo: Hucitec/Abrasco.

COOK, D. G. & POCOCK, S. J., 1983. Multiple regression in geographic mortality studies with allowance for spatially correlated errors. *Biometrics*, 39: 361-371.

COSTNER, H. L., 1965. Criteria for measures of association. *American Sociological Review*, 30: 341-353.

COX, D. R., 1970. *Analysis of Binary Data*. London: Methuen.

DANIELS, H. E., 1944. The relation between measures of correlation in the universe of sample permutations. *Biometrika*, 33: 129-135.

DAVIS, F. N., 1955. Dicing and Gaming (a note on the history of probability). *Biometrika*, 42: 1-15.

DEMO, P., 1989. *Metodologia Científica em Ciências Sociais*. São Paulo: Editora Atlas.

DRAPER, N. R. & SMITH, H., 1966. *Applied Regression Analysis*. New York: John Wiley & Sons.

FELLER, W., 1968. *An Introduction to Probability Theory and Its Applications*. 3rd edition, New York: John Wiley & Sons.

FERGUNSON, T. S., 1967. *Mathematical Statistics (a decision theory approach)*. New York: Academic Press.

FISHER, R. A., 1956. *Statistical Method and Scientific Inference*. Edinburgh: Oliver and Boyd.

FLEISS, J. L., 1973. *Statistical Methods for Rates & Proportions*. New York: John Wiley & Sons.

GOLDBERG, M., 1990. Este obscuro objeto da Epidemiologia. In: *Epidemiologia Teoria e Objeto* (D. C. Costa, org.), pp. 87-136, São Paulo: Hucitec
Abrasco

GOODMAN, L. A., 1979. Simple models for the analysis of association in cross-classification having ordered categories. *Journal of the American Statistics Association*, 74: 537-552.

GREEN, P. E., 1978. *Analysing Multivariate Data*. Hinsdale, Illinois: The Dryden Press.

GREENLAND, S., 1988. On sample-size and power calculations for studies using confidence intervals. *American Journal of Epidemiology*, 128: 231-237.

HABERMAN, S. J., 1978. *Analysis of Qualitative Data*. New York: Academic Press.

HAMMOND, R. & MC CULLAGH, P. S., 1978. *Quantitative Techniques in Geography: an Introduction*. Oxford: Clarendon Press.

HILL, A. B., 1965. *Principles of Medical Statistics*. New York: Oxford University Press.

HOEL, P. G.; PORT, S. C. & STONE, C. J., 1971. *Introduction to Probability Theory*. Boston: Houghton Mifflin Company.

HOEL, P. G., 1980. *Estatística Matemática*. Rio de Janeiro: Editora Guanabara Dois.

HOSMER, D. W. & LEMESHOW, S., 1989. *Applied Logistic Regression*. New York: John Wiley & Sons.

HOTELLING, H., 1951. The impact of R. A. Fisher on statistics. *Journal of the American Statistics Association*, 46: 35-46.

HUFF, D., 1954. *How to Lie with Statistics*. New York: W. W. Norton.

IOSIFESCU, M. & TAUTU, P., 1973. *Stochastic Processes and Applications in Biology and Medicine*. New York: Springer-Verlag.

JEFFREYS, H., 1948. *Theory of Probability*. 2nd ed., Oxford: Clarendon Press.

JOHNSTON, R. J., 1978. *Multivariate Statistical Analysis in Geography*. London: Longman.

KENDALL, M. G., 1956. Studies in the history of probability and statistics: II. *Biometrika*, 43: 1-14.

KNEKT, P.; REUNANEN, A.; AROMAA, A.; HELIOVAARA, M. & HAKAMA, M., 1988. Serum cholesterol and risk of cancer in a cohort of 39,000 men and women. *Journal of Clinical Epidemiology*, 41: 519-530.

LAURENTI, R.; JORGE, M. H. P. M.; LEBRÃO, M. L. & GOTLIEB, S. L. D., 1985. *Estatísticas de Saúde*. São Paulo: Editora Pedagógica e Universitária Ltda.

LEHMANN, E. L., 1959. *Testing Statistical Hypotheses*. New York: John Wiley & Sons.

LI, C. C., 1975. *Path Analysis-a Primer*. Pacific Grove, California: The Boxwood Press.

LINDLEY, D. V., 1957. A statistical paradox. *Biometrika*, 44: 187-192.

LOWY, M., 1991. *Ideologias e Ciência Social - Elementos para uma Análise Marxista*. São Paulo: Cortez Editora.

MORGENSTERN, H., 1982. Uses of ecologic analysis in epidemiologic research. *American Journal of Public Health*, 72: 1336-1344.

MORGENSTERN, H., 1989. Epidemiologic Methods, class notes (Mimeo.).

NARAYAN BHAT, U., 1972. *Elements of Applied Stochastic Processes*. New York: John Wiley & Sons.

NEUTS, M. F., 1973. *Probability*. Boston: Allyn and Bacon Inc..

NUNES, E. D. (Org.), 1985. *As Ciências Sociais em Saúde na América Latina: tendências e perspectivas*. Brasília: OPAS.

OAKES, M., 1990. *Statistical Inference*. Chestnut Hill, MA: Epidemiology Resources Inc.

OPENSHAW, S. & TAYLOR, P. J., 1979. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: *Statistical Applications in the Spatial Sciences* (N. Wrigley, ed.), pp. 128-144, London: Pion Limited.

ORGANIZAÇÃO MUNDIAL DA SAÚDE, 1978. *Classificação Internacional de Doenças, Lesões e Causas de Óbitos: 9ª revisão*. Vol. 1. São Paulo, Centro da OMS para classificação de doenças em Português.

PIANTADOSI, S.; BYAR, D. P. & GREEN, S. B., 1988. The ecological fallacy. *American Journal of Epidemiology*, 127: 893-900.

PHILLIPS, L. D., 1973. *Bayesian Statistics for Social Scientists*. London: Nelson.

POLLARD, A. H. ; YUSUF, F. & POLLARD, G. N., 1974. *Demographic Techniques*. Sydney: Pergamon Press.

RANKIN, B., 1966. The history of probability and the changing concept of the individual. *Journal of the History of Ideas*, 27: 483-504.

RAO, C. R., 1973. *Linear Statistical Inference and Its Applications*. New York: John Wiley & Sons.

RAUBERTAS, R. F., 1988. Spatial and temporal analysis of disease occurrence for detection of clustering. *Biometrics*, 44: 1121-1129.

REMINGTON, R. D. & SCHORK, M. A., 1970. *Statistics with Applications to the Biological and Health Sciences*. Englewoods Cliffs, New Jersey: Prentice-Hall.

RICHARDSON, S.; STUCKER, I. & HEMON, D., 1987. Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology*, 16: 111-120.

ROBINSON, W. S., 1950. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15: 351-357.

SABROZA, P. C., 1990. Prefácio. In: *Epidemiologia Teoria e Objeto* (D. C. Costa, org.), pp. 7-10, São Paulo: Hucitec/Abrasco.

SAVAGE, L. J., 1954. *The Foundations of Statistics*. London: Routledge and Kegan Paul.

SEARL, S. R., 1971. *Linear Models*. New York: John Wiley & Sons.

STEEL, R. G. D. & TORRIE, J. H., 1981. *Principles and Procedures of Statistics (a biometrical approach)*. Singapore: Mc Graw-Hill.

SUSSER, M., 1985. Epidemiology in the United States after World War II: the evolution of technique. *Epidemilogic Reviews*, 7: 147-177.

TANGO, T., 1984. The detection of disease clustering in time. *Biometrics*, 40: 15-26.

UK NATIONAL CASE-CONTROL STUDY GROUP, 1989. Oral contraceptive use and breast cancer risk in young women. *The Lancet*, May 6: 973-982.

WALKER, H. M., 1958. The contributions of Karl Pearson. *Journal of the American Statistics Association*, 53: 11-27.

WOLFOWITZ, J., 1952. Abraham Wald, 1902-1950. *Annals of Mathematical Statistics*, 23: 1-13.

YULE, G. U. & KENDALL, M. G., 1950. *An Introduction to the Theory of Statistics*. London: Charles Griffin.