CAPES, em particular, reconhece a escassez de pessoal pós-graduado em nossa área e parece disposta a colaborar na reversão desse quadro. Uma proposta já disponibilizada é a abertura de cursos de especialização, que fariam um papel de preparação de alunos para o ingresso adequado a cursos de pós-graduação. Outras propostas podem ser postas em discussão e devemos participar ativamente desse processo.

ESTATÍSTICA: A TECNOLOGIA DA CIÊNCIA

Basílio de Bragança Pereira

A pesquisa científica é um processo iterativo de acumulação de conhecimento e envolve a formulação de hipóteses, modelos e teorias, a observação de fenômenos e a verificação e rejeição de hipóteses sobre os mesmos.

A estatística procura tornar esse processo o mais eficiente possível, através de suas técnicas de coleta de dados (amostragem e planejamento de experimentos); apresentação de dados (análise exploratória e descrição: gráficos e tabelas); modelagem (probabilidade e processos estocásticos); análise indutiva (inferência: testes e estimação) e verificação (ajustamento, previsão e controle).

Neste artigo apresenta-se as várias etapas da pesquisa científica e a atuação da estatística nesse processo, concluindo-se que a estatística é a tecnologia da ciência.

1 Introdução: ciência e estatística

Quando se fala em estatística, o público em geral associa a idéia de um trabalho de coletar e armazenar números e dados ou, quando muito, ao cálculo de percentagens e índices a partir desses dados. Entretanto a estatística ou os métodos estatísticos têm um papel muito mais importante na ciência e tecnologia.

A concepção atual de ciência é de aprendizado através de experimentação e dos dados observados segundo o qual a procura das causas, das leis, traduz-se em um processo iterativo de observação do real, de repetição de experimentos, da avaliação quantitativa dos fenômenos em estudo. A pesquisa científica é um processo de aprendizado, obtido pela iteração ilustrada na Fig. 1. Uma hipótese inicial leva por um processo de dedução a certas consequências que são comparadas com os dados.

Quando as conseqüências e os dados falham em concordar, as discrepâncias podem levar, por um processo chamado indução, a modificar a hipótese. Um segundo ciclo na iteração é iniciado. As consequências das novas hipóteses são analisadas e novamente comparadas com dados (velhos ou recentemente obtidos) que novamente leva a mais modificações e ganho de conhecimento. Este conhecimento assim obtido, tem interesse intrínseco para satisfazer nossa curiosidade ou/e para objetivos de decisão. Uma parte de nosso conhecimento é meramente uma descrição do que observamos, a parte mais importante é a generalização ou indução que consiste em fazer inferências de experiências passadas para predizer as futuras.

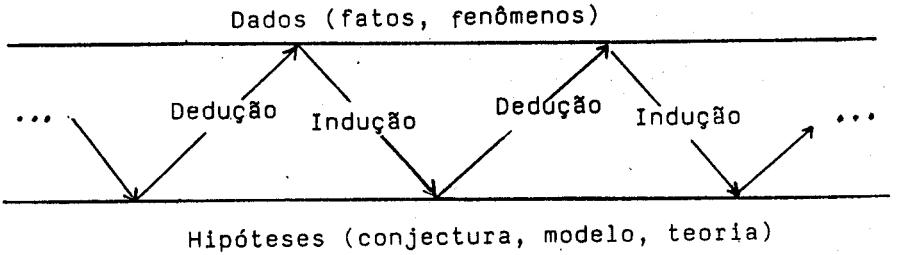

Fig. 1: Processo iterativo de aprendizado

Com relação à ciência e aos métodos estatísticos devemos enfatizar o princípio de unidade da ciência que Karl Pearson explicava da seguinte forma:

"A unidade da ciência é a unidade dos métodos empregados em analizar e aprender através da experiência e dos dados."

e que Harold Jeffreys expressa:

"Deve haver um padrão uniforme de validação para todas as hipóteses, independente de área de conhecimento. Diferentes leis podem ser válidas em diferentes áreas porém elas devem ser testadas pelos mesmos critérios; caso contrário não teremos garantia de que nossas decisões serão aquelas garantidas pelos dados e não meramente o resultado de análise inadequada ou de acreditar no que queiramos acreditar."

Portanto, o princípio da unidade da ciência impõe a utilização dos mesmos padrões de trabalho e de um conjunto unificado de métodos para uso. Este conjunto unificado de métodos constitue o que se entende por Estatística, e neste sentido podemos, sem dúvida, defini-la como a Tecnologia da Ciência.

2 Estrutura da ciência

O que é ciência?

Ciência significa conhecimento. Ciência não se ocupa apenas em coletar fatos sobre fenômenos; além disso ela se impõe a tarefa de explicar o que observamos e predizer o que será observado em circunstâncias que ainda não ocorreram.

Conhecimento científico é obtido observando regularidades em eventos. Se tudo fossem ocorrências aleatórias, então todo evento seria uma nova experiência. Porém, a natureza não se comporta assim, ela contém regularidade e o tema central da atividade científica é explicar tais regularidades. Além disso a identificação de regularidades incomuns e surpreendentes é o ponto principal na produção das grandes descobertas. O cientista se preocupa com as regularidades naturais em fenômenos porque estas sugerem leis e é a possibilidade de descobrir leis científicas qua atrai curiosidade científica.

Leis científicas

Uma lei científica expressa relações entre coisas e pode ser expressa como afirmações ou em termos de formulação matemática.

Hipóteses

Em geral, uma hipótese é considerada como uma proposição de que alguma coisa é verdadeira. Portanto uma hipótese não trata de algo que foi observado mas sobre algo que não foi observado, porém que, em princípio, é observável.

Teorias

Relacionada com leis científicas e hipóteses estão as teorias. Afirmações na forma de leis se referem a coisas ou propriedades de coisas que podem ser observadas. Uma teoria por outro lado, envolve conceitos que não observáveis e usualmente tem significado somente dentro das definições usadas na teoria. Em outras palavras, teorias evolvem conceitos teóricos, por exemplo, a economia usa conceitos como "utilidade", "competição perfeita", "racionalidade" etc. A combinação do conjunto de relações dentro da qual tais conceitos teóricos estão incluídos é chamado de teoria.

Teorias não se constituem apenas de termos teóricos, elas também incorporam leis científicas. Estas leis têm certa generalidade, porém as teorias que incorporam tais leis têm ainda maior generalidade. Em certo sentido as teorias chegam a "realidade" através de leis.

Uma teoria coloca as leis científicas e hipóteses dentro de um contexto que contém conceitos teóricos expressando a visão do cientista sobre como ele concebe a natureza das coisas, além de que se observa em experiementos. Deve-se observar que leis científicas individuais e frequentemente um número de hipóteses dentro de uma teoria, permanecem válidas mesmo a teoria é descartada.

Modelos

Modelos são empregados para facilitar nosso entendimento da realidade. Se lembrarmos como, quando criança, aprendemos a entender o mundo, logo verificamos o papel dos modelos. Modelos de trens, de carros, bonecas, foram fundamentais ao nosso aprendizado. Modelos de situações (simulações) apareciam como brincadeira de professor-aluno, bandido-mocinho etc. Não podíamos tratar com as coisas reais e um modelo tinha que ser suficiente. A coisa real era muito grande, muito complicada, muito cara, muito perigosa. Como adultos pelas mesmas razões construímos modelos de rios para estudar o efeito de uma barragem, de praias para estudar o efeito de um aterro, de aviões para testar em túneis de vento e de forma similar usamos jogos de empresas, jogos de guerra etc.

Muitos dos modelos mencionados são físicos. Entretanto, o estímulo maior para a ciência vem de modelos conceituais ou teóricos. Modelos de economias nacionais foram construídos usando água para representar fluxos ou movimentos de riquezas. É mais fácil e flexível representar o grande número de componentes de uma economia e suas interrelações através de modelos conceituais.

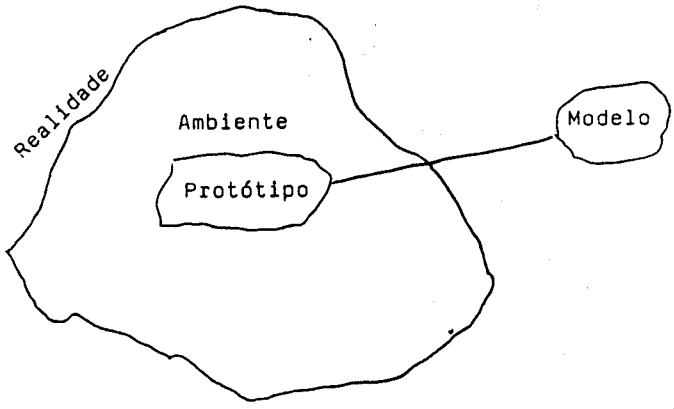

Fig.2: O modelo e a realidade

Não podemos esperar que os modelos sejam a verdadeira descrição da realidade; eles serão apenas aproximações para certos aspectos da realidade. Para ser possível considerar um modelo como uma representação do mundo real, devemos dar atenção a aspectos específicos da realidade. Este subconjunto da realidade consiste de particular objeto, evento, sistema etc. O termo Protótipo pode ser usado. Ao definir um problema separamos a "realidade" em protótipo e o resto que pode ser chamado de Ambiente. A situação é representada na Fig. 2.

A maior parte do ambiente é totalmente irrelevante para o problema em estudo. (A distância de Júpiter não é relevante para uma análise econômica da indústria brasileira). Ao definir o protótipo, muitos fatores relevantes têm que ser desconsiderados para manter o problema em um tamanho razoável. Isto pode ser considerado como o Ambiente Vizinho. Frequentemente verificamos que alguns fatores não considerados, são importantes e devem ser trazidos para o protótipo e alternativamente que fatores no protótipo não são usados, isto é, temos o protótipo errado para a aplicação corrente. Também, frequentemente estaremos examinando como o modelo compara com o protótipo, algumas vezes teremos o modelo errado para o protótipo correto.

Indicamos anteriormente que teorias envolvem conceitos que não são diretamente observáveis. Tem-se então o problema de como a teoria pode ser usada para explicar coisas que são observáveis e além disso predizer eventos que ainda não ocorreram. O problema então é de como relacionar a teoria com coisas que podem ser observadas. Conceitos, hipóteses e leis representam a maneira como a teoria é relacionada a um particular fenômeno. A teoria não pode ser testada diretamente pois a mesma envolve conceitos teóricos que não são observáveis. Entretanto, é possível testar a teoria indiretamente através de um modelo ou modelos. As predições do modelo fornecem um teste indireto da teoria.

Modelos matemáticos e probabilísticos

Em virtude de modelos serem o meio pelo qual analisamos as implicações de

nossas teorias então, em princípio, não há razões para que um modelo não possa ser apresentado verbalmente. A desvantagem de tal procedimento é que seria difícil obter desta forma a precisão que é necessária ao raciocínio dedutivo quando um grande número de variáveis é considerado. Modelos teóricos ou conceituais em ciência geralmente tomam uma forma matemática, porque a matemática é a linguagem da ciência. A matemática permite ao cientista expressar as interrelações de seu modelo precisamente e ao mesmo tempo com extrema generalidade. A evidência que se tem na ciência é que de acordo com o desenvolvimento do assunto o mesmo se torna mais matemático.

Em ciência nem tudo é conhecido com certeza absoluta e nem sempre podemos fazer afirmações com certeza. A razão é que, para muitas coisas da realidade, estamos ainda ignorantes. Isto não significa que não temos informação, porém nos leva a aceitação do acaso ou aleatoriedade e a utilização de modelos mais gerais: modelos probabilísticos. Tais modelos além de representar as regularidades da natureza através de uma componente sistemática ou determinística representam a componente de incerteza através de uma componente probabilística, que mede essa incerteza.

3 A ferramenta da pesquisa científica: os métodos estatísticos

O processo de aprendizado da pesquisa científica na Fig. 1 pode ser visto como um sistema fechado na Fig. 3, na qual a discrepância entre os dados e as conseqüências da hipótese H_1 levam à hipótese H_2; H_2 leva à H_3 e assim por diante.

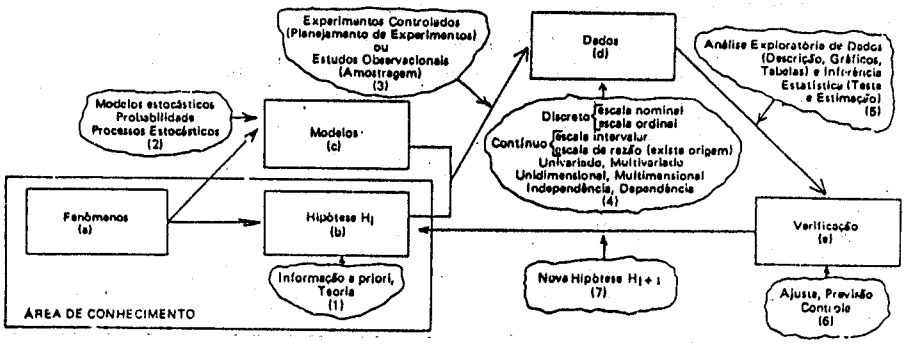

Fig. 3: Processo científico

O trabalho científico é iniciado quando o cientista, ao observar um fenômeno (a), formula hipóteses (b) para explicar o mesmo dentro de uma teoria (1) objetivando descobrir alguma lei científica. Com as hipóteses e escolhendo um modelo (c) para o seu protótipo por um processo dedutivo, ele obtém as conseqüências das hipóteses. Essas conseqüências e os dados reais (d) são comparados verificando-se

se eles estão em concordância (e). Nesta etapa verifica-se não só se o protótipo é correto (5) mas também se o modelo é correto (6). Após essa verificação por um processo de indução, as hipóteses iniciais são modificadas ou novas hipóteses são formuladas, e o processo é então re-iniciado, e assim sucessivamente. Em todas essas etapas do processo iterativo da pesquisa, a Estatística tem um papel fundamental, como se nota na Fig. 3 (2, 3, 4, 5 e 6) e que passamos a descrever.

Os modelos à disposição do cientista são os modelos estocásticos ou probabilísticos (2). Esta é a área de modelos estatísticos tratada na Teoria das Probabilidades para fenômenos estáticos e na Teoria dos Processos Estocásticos para fenômenos dinâmicos. Tendo em vista seu grande desenvolvimento teórico e sua riqueza matemática esta parte atualmente pode ser considerada uma nova disciplina com pesquisadores dedicados exclusivamente à mesma.

Ao coletar as observações (3) o cientista necessariamente procurará fazê-lo de forma que os dados obtidos sejam representativos para seu estudo e de forma mais eficiente possível. Se o cientista tem controle sobre o fenômeno em estudo ele se utiliza das técnicas de Planejamento de Experimentos. Este seria o caso por exemplo de comparar a viscosidade de um produto químico, controlando a pressão e a temperatura do forno. Caso o fenômeno em estudo não permita um controle ele se utiliza das técnicas de Amostragem. Um caso típico seria uma pesquisa eleitoral onde se deseja determinar a intenção de votos dos eleitores de um país.

Dependendo do problema, as observações são feitas em diferentes escalas, número de medições, dimensão e estrutura de dependência (4). Relacionado à escala tem-se: escala nominal (ou classificatória) que é a forma mais fraca de mensuração. Números, símbolos ou nomes são usados para classificar os objetos em celas, contando-se quantos são em cada cela. Por exemplo, descrevendo-se a cor de grãos de areia.

branco	amarelo	cinza	pode ser codificado
1	2	3	ou
45	3	20	ou
-	·	+	etc.

Escala ordinal (ou rank) é usada quando os objetos são identificados como diferentes e também colocados em uma seqüência em relação aos outros. Um exemplo seria a resposta ao efeito de anestésico onde existe uma certa ordenação.

nenhuma dor	pouca dor	muita dor	pode ser codificado
0	1	2	ou
1	50	60	ou
4	3	1	etc.

Escala intervalar é utilizada quando há uma igualdade no comprimento dos intervalos entre as classes. Isto faz com que a razão entre dois intervalos seja

independente da unidade de medida utilizada e da origem, pois ambos são arbitrários. Como exemplo considera-se as medidas de temperatura em centígrados e Fahrenheit que tem diferentes origens (arbitrárias), a razão entre as diferenças de duas temperaturas (intervalos) é a mesma em ambas as escalas:

$°C$	0	20	50	100
$°F$	32	68	122	212

em $°C \frac{100-50}{20-0} = 2,5$
em $°F \frac{212-122}{68-32} = 2,5$

Escala de razão onde há uma origem não arbitrária além das propriedades da escala intervalar, neste caso a razão entre duas medidas é independente da unidade de medida. Massa, comprimento, velocidade, profundidade estão sempre nesta escala: a razão entre medidas de comprimento em centímetros e em polegadas é sempre constante e igual a 2,54. A escala de razão é a mais versátil e poderosa escala pois contém o máximo de informação.

Dois tipos diferentes de medição, atributos e variáveis foram considerados. Atributos são obtidos em escala nominal ou ordinal e tem valores discretos tais como: presença ou ausência ou cinco. Variáveis têm escala intervalar ou de razão e tem uma escala contínua de valores.

Relacionado com o número de medidas do objeto em estudo podemos ter uma observação univariada, quando tomamos uma só medida, por exemplo o peso do objeto. Alternativamente temos uma observação multivariada quando tomamos várias medidas, por exemplo: peso, altura e volume do objeto.

A dimensão da observação é importante quando tratamos de sistemas dinâmicos que evoluem em relação a outras variáveis (tempo, espaço, profundidade etc.). Teremos uma observação unidimensional quando se tem o sistema variando em relação a uma só variável, por exemplo a evolução do Produto Interno Bruto com o tempo. A observação é multidimensional quando a variação se faz em relação a mais de uma variável, por exemplo a evolução do número de casos de sarampo por semana e por municípios de um estado ou a porosidade do solo por profundidade, latitude ou longitude numa bacia petrolífera.

A estrutura de dependência se refere a dependência entre os objetos em estudo. As observações serão independentes se a escolha de um objeto observado não influencia a escolha dos outros objetos observados. Como exemplo temos o caso de pesquisa eleitoral quando a seleção de um eleitor para opinar não é influencida ou influencia a escolha dos outros eleitores a serem ouvidos. As observações serão dependentes quando a ordem ou posição em que foi feita a observação é também importante. Por exemplo no estudo da evolução do Produto Interno Bruto é importante não só ter o valor do mesmo nos diversos anos mas também poder identificar os instantes das observações, no estudo da evolução dos casos de sarampo é importante identificar não só o número de ocorrências como as semanas e os municípios correspondentes.

Após obter as informações, o cientista tem a tarefa de armazená-las, simplificar sua descrição através de listagens, tabelas e gráficos mais convenientes, esta etapa constitui-se no que se chama Análise Exploratória de Dados ou Estatística Descritiva (5). Em seguida o cientista passa à etapa de comparação entre as conseqüências de suas hipóteses e a evidência apresentada nos dados e se utiliza da Inferência Estatística (5). Basicamente, nesta etapa supondo o modelo correto ele procura verificar se suas hipóteses são confirmadas ou não pelos dados e se seu protótipo é correto. Para isso, de acordo com o problema e portanto o modelo utilizado e os dados coletados, ele tem uma quantidade de técnicas de inferência em Séries Temporais, Estatística Multivariada, Estatística Não Paramétrica, Análise de Decisões.

Após esta fase, supondo o protótipo correto é importante se assegurar de que o modelo utilizado na etapa anterior era realmente correto, o cientista passa à fase de Verificação (6). Essa verificação é feita analisando-se as diferenças entre os dados e as respostas do modelo, fazendo previsões e controles com o modelo e verificando se as previsões assim obtidas estão próximas de novas observações (Ajuste e Análise dos Resíduos).

Com o ganho de conhecimento obtido o cientista prossegue formulando novas hipóteses, alterando seus modelos e protótipos e reiniciando outra vez o ciclo.

Deve-se mencionar que dentro das diversas etapas do trabalho estatístico visto acima, embora com os mesmos objetivos, diversos enfoques são disponíveis para a Inferência Estatística dependendo da escola filosófica da Estatística que seja adotada. Entre elas tem-se a Escola Freqüentista, (ou Clássica), a Bayesiana, a de Teoria da Decisão, a da Verossimilhança, a Estrutural, a de Plausibilidade etc.

Finalmente, como foi visto, o cientista deve se preocupar com estatística desde o início de sua pesquisa para que os experimentos que ele realizar ao final sejam úteis na verificação de suas hipóteses. Este aspecto deve ser ressaltado, já que na maioria das vezes o cientista procura um estatístico apenas na fase de inferência estatística, já com os dados observados e muitas vezes a Estatística nada pode oferecer tendo em vista que o experimento não foi realizado adequadamente. Portanto, a preocupação com a Estatística deve estar presente desde o início da pesquisa.

Bibliografia

1. Barnett, V. (1982). *Comparative statistical inference.* 2.ed. Wiley.
2. Box, G.E.P. (1976). Science and statistics. *J. Am. Stat. Ass.*, **71**, 791–799.
3. Box, G.E.P.; Hunter, W.G.; Hunter, J.S. (1978). *Statistics for experimenters.* Wiley.
4. Gilchrist, W. (1984). *Statistical modelling.* Wiley.
5. Healy, M.J.R. (1978). Is statistics a science? *J. R. Stat. Soc. A*, **141**, 385–393.
6. Neal, F.; Shone, R. (1976). *Economic model building.* McMillan.
7. Till, R. (1974). *Statistical methods for the earth scientist. An introduction.* Wiley.

8. Zellner, A. (1984). *Basic issues in econometrics.* Chicago University Press.

9. Pereira, B. de B. (1983). *Correspondência ao Presidente da APEB (Associação dos Profissionais de Estatística do Brasil).* Sugestão sobre Curriculum Mínimo.

10. Pereira, B. de B. (1985). Correspondência: a questão do posicionamento profissional do estatístico. *Boletins do CONRE (Conselho Regional de Estatística, 2ª Região)*, **II**, n. 7 e 8.