Brazilian Journal of Videoendoscopic Surgery

Noções Básicas de Bioestatística

Basics of Biostatistics

MARCO AURELIO PINHO DE OLIVEIRA[1]; RAPHAEL CÂMARA MEDEIROS PARENTE[2]

[1.] Doutor em Epidemiologia pelo Instituto de Medicina Social da Universidade do Estado do Rio de Janeiro. Chefe do Setor de Ginecologia da UERJ; [2.] Doutor em Ginecologia pela UNIFESP (Reprodução Humana). Mestre em Epidemiologia pelo Instituto de Medicina Social da Universidade do Estado do Rio de Janeiro.

Bras. J. Video-Sur, 2010, v. 4, n. 1: 005-008

Quando se inicia a análise estatística dos dados a primeira pergunta óbvia é: "o que quer dizer estatística?". Simploriamente, a estatística significa o conjunto de relações calculadas com base nos dados de uma amostra adequada, que deve ser parte representativa de uma população.

Nós podemos dividir a estatística, didaticamente, em dois grupos: 1- Descritiva; 2-Inferencial. Na estatística descritiva, o objetivo é simplesmente descrever a amostra em questão. A descrição normalmente é feita na tentativa de se resumir os dados obtidos, seja através das freqüências em percentual, médias e desvios padrão ou gráficos. Na maioria dos trabalhos científicos o que se vê é apenas esta estatística descritiva. Estes trabalhos na sua maioria se limitam a revisões de prontuários ou fichas apropriadas, e não envolvem hipóteses a serem testadas. O papel da estatística inferencial é transferir, generalizar as conclusões da amostra para a população. Para sermos mais objetivo, o interesse maior no dia-a-dia é de comparar dados entre dois ou mais grupos para saber se houve diferença estatisticamente significativa. Vale a pena comentar um pouco sobre o que é significância. Se alguém disser que a chance de algo acontecer é de 1 em 100 (probabilidade de 0,01 ou $p = 0,01$), isto é pode ser considerado muito ou pouco? Depende. Se esta for a probabilidade de um avião cair, há de se concordar que é alta. Mas, se esta for a chance de falha na melhora da cefaléia após a tomada de uma aspirina, a probabilidade da falha é baixa. Quem estipula o nível de significância é o pesquisador. No meio acadêmico ficou tradicionalmente estipulado que se a chance de algo ocorrer é menor que 5 % ($p < 0,05$) então ela é pouco provável de acontecer. Por exemplo, no estudo de um novo diurético distribuímos aleatoriamente 30 pessoas para o grupo de medicamento ativo e 30 pessoas para o grupo placebo (medicamento inerte). A média do volume urinário em 24 horas foi de 3600 ml no primeiro grupo e de 3400 ml no segundo grupo.

Como existe a diferença de 200 ml, em média, logo podemos afirmar que o medicamento realmente funciona como diurético?. Claro que não! É necessário realizar o teste estatístico apropriado (neste caso poderia ser o t de student) e ver qual é a probabilidade desta distribuição ter ocorrido apenas ao acaso. No momento da composição das amostras, pode ser que por acaso tenhamos escolhido para o grupo medicamento ativo os indivíduos que naturalmente apresentam maior diurese nas 24 horas - ou será que isso não ocorreu e o medicamento foi realmente eficaz? Para ajudar nesta decisão, os testes estatísticos são usados para que possamos saber, num determinado estudo, qual a probabilidade da distribuição ter ocorrido apenas pelo acaso. Após a realização do teste de t de student, verificamos que a probabilidade de encontrarmos uma diferença de 200 ml (1600 ml - 1400 ml) nesta amostra de 60 (30 + 30) pessoas é de 3 % ($p = 0,03$), portanto $p < 0,05$. Como já foi colocado, nós consideramos esta ocorrência pouco provável, ou seja, é pouco provável ($p = 0,03$) que esta distribuição tenha ocorrido pelo acaso, logo, devemos ter outra explicação para a questão e até que se prove o contrário a diferença de 200 ml na média foi por causa do medicamento ativo. E atenção: ainda temos 3 % de chance desta diferença de ter sido pelo acaso e não pelo medicamento ativo - esse é o risco (erro tipo alfa ou tipo I) que se corre nos testes de hipóteses. Porém, se após a realização do teste de t de student nós encontrássemos $p = 0,15$ ($p > 0,05$) ao invés de $p = 0,03$, chegaríamos à conclusão de que a chance da distribuição ter sido ao acaso não é peque-

na ($p > 0,05$), portanto não poderíamos afirmar que o medicamento ativo teve efeito. Neste caso, por conta do resultado ser não-significativo, deve-se observar o poder do teste estatístico, que deve ser calculado *a priori* (antes da realização do estudo). Quanto menor a amostra, menor o poder para se afirmar que o tratamento não funciona, ou seja, o tratamento pode ser de fato eficaz, porém o pequeno número de participantes na amostra não é permite atingir a significância estatística. Se o poder for menor que 80% (existem fórmulas específicas para calculá-lo) podemos estar diante de um $p > 0,05$ falso, ou seja, p poderia ser menor que 0,05, porém a amostra pode ter sido pequena para atingir tal probabilidade - erro tipo II ou beta.

Como escolher o teste estatístico apropriado

Como já sabemos para o que serve o p fornecido pelos testes estatísticos, vamos nos preocupar agora em quando utilizar determinado teste. Para isto é fundamental que saibamos qual o nível de mensuração das variáveis envolvidas. Podemos dividir em três grupos: 1- Nominal ; 2 - Ordinal; 3 - Intervalar/Razão. Na variável nominal, o número não vale como número e sim como categoria, por exemplo: 1 = solteiro; 2 = casado; 3 = divorciado 4 = desquitado e 5 = viúvo. Não se pode somar, subtrair ou tirar médias. Esses números representam apenas categorias diferentes. Os testes mais usados nestes casos são o qui-quadrado (X^2) e o teste de Fisher, este usado principalmente para amostras muito pequenas. Na variável ordinal, os números já podem ser ordenados (p.ex. do menor para o maior), porém não se deve tirar média ou desvio padrão, como p.ex.: na classificação da endometriose, a paciente que recebe 40 pontos não tem o dobro de endometriose do que a paciente que recebeu 20 pontos, porém pode-se dizer que a primeira tem mais endometriose que a segunda. Outro exemplo é a pontuação que se dá para dor no pós-operatório (fraca = 1; média =2, etc..). Os testes mais usados são o U de Mann-Whitney (para dois grupos) e o teste de Kruskal - Wallis (três ou mais grupos). Estes testes não se utilizam de parâmetros da população (não requerem, por exemplo, distribuição normal) e são denominados de não-paramétricos. O terceiro grupo inclui variáveis intervalares e de razão (a diferença básica é que na razão o zero é absoluto (p.ex, peso) e na intervalar o zero é relativo (p.ex., temperatura em Celsius) - os testes estatísticos costumam ser os mesmos para esses dois tipo de variáveis. Neste grupo os números são realmente números, podendo-se somar, subtrair, dividir, multiplicar, tirar médias e desvio padrão. Podem ser contínuos (p.ex. peso em Kg) ou descontínuos, p.ex. número de filhos (1, 2, 3, etc..). Nestes casos, 4 Kg é o dobro de 2 Kg, assim como quatro filhos é o dobro de dois. Os testes mais usados são o t de student (para dois grupos) e o teste de análise de variância (três ou mais grupos). Como estes testes utilizam parâmetros da população (notadamente média e desvio padrão, assumindo que a população apresente uma distribuição normal), eles são chamados de testes paramétricos.

Entendendo intervalo de confiança

Outro assunto que merece ser abordado é o intervalo de confiança. Para que possamos entender o intervalo de confiança é necessário o conhecimento prévio do erro padrão da média. Já foi comentado que o pesquisador trabalha com amostras de uma população, e que através dos dados destas amostras deseja conhecer a população (extrapolação dos dados ou generalização). As melhores amostras são aquelas selecionadas aleatoriamente da população em questão. Acontece que estas amostras são diferentes uma das outras. Por exemplo, digamos que um pesquisador A deseja saber qual é o peso médio dos médicos de um determinado hospital. Neste hospital trabalham 100 médicos de cinco especialidades diferentes (a , b , c , d , e), com 20 médicos cada . O pesquisador A resolve selecionar ao acaso, cinco médicos de cada especialidade, totalizando 25 médicos - amostra estratificada por especialidade. A média encontrada foi de 68 Kg. Outro pesquisador, chamado de B, resolve fazer um estudo idêntico ao do A. Ele encontrou uma média de 70 Kg já que obviamente os indivíduos selecionados ao acaso não foram os mesmos. O pesquisador C num estudo idêntico encontrou 72 Kg de média. Existe alguma coisa errada com as médias encontradas? Não, apenas os indivíduos selecionados ao acaso não são os mesmos nas três pesquisas. Portanto, quando um pesquisador seleciona a sua amostra, ele sabe que existem muitas outras amostras e que vão fornecer médias diferentes da que ele vai encontrar. O número de amostras diferentes é praticamente infinito. Se continuássemos a fazer outras pesquisas idênticas, teríamos várias médias (p.ex., 66 Kg, 68 Kg, 70 Kg, 72 Kg e 74 Kg) que no seu conjunto apresentam a propriedade da distribuição normal. Existe uma propriedade estatística que diz que

a média de todas estas médias é igual à média da população, ou seja, a média verdadeira, caso fossem pesados todos os 100 médicos. Digamos que um outro pesquisador D com mais tempo resolveu medir o peso de todos os médicos e encontrou 70 Kg de média. As várias médias encontradas nas amostras pelos outros pesquisadores vão ter distribuição normal em torno da média real da população. Nós sabemos que é 70 Kg graças ao pesquisador D.

O desvio padrão das possíveis médias é chamado de erro padrão da média (EPM) ou standard error of the mean (SEM). Este erro expressa a variabilidade que pode ser encontrada na média de uma amostra de um determinado tamanho, pois, como já discutimos, a média de uma amostra normalmente não é idêntica à média real da população. O intervalo de confiança nada mais é que o grau de confiança que o pesquisador tem que a média da população (média verdadeira) está contida naquele intervalo. Habitualmente se utiliza o intervalo de confiança de 95% (a=5%). O pesquisador A, que encontrou uma média de 68Kg na sua amostra, diria que a média da população (100 médicos) deve estar entre 68 Kg e mais ou menos algum erro. Este erro pode ser calculado usando-se o valor correto da distribuição t para um intervalo de confiança de 95%, ou um a = 5 %. Para uma amostra de 25 indivíduos o valor fornecido pela tabela da distribuição t é igual a 2,06. Este valor deve ser multiplicado pelo erro padrão da média (EPM), que pode ser calculada dividindo-se o desvio padrão da amostra pela raiz quadrada do número de indivíduos na amostra. Se o EPM fosse igual a 1, o erro seria igual a 2,06. Portanto teríamos 95% de certeza que a média da população estaria entre 68 + 2,06kg, ou seja, aproximadamente entre 66 e 70kg (neste caso o intervalo de 95% incluiu o valor verdadeiro – 70kg).

Não devemos confundir o EPM com o desvio padrão (DP) ou standard deviation (SD). O primeiro, como já foi explicado, expressa a variabilidade, a incerteza, da média obtida através de uma amostra. O DP, expressa a variabilidade dos indivíduos (e não das médias) selecionados em torno da média da amostra. No caso do pesquisador A, o DP é calculado da seguinte forma: pegar o peso de cada um dos 25 médicos escolhidos, subtrair da média encontrada (68 Kg), e elevar ao quadrado esta diferença. Se um indivíduo pesa 98kg, você deve subtrair 98 Kg - 68 Kg e elevar este resultado ao quadrado, ou seja, 30^2. Em seguida deve ser feita a soma de todas essas diferenças e dividir pelo número de indivíduos menos um (nesse caso seria 25-1 = 24). Este valor é chamado de variância. Depois disso basta encontrar a raiz quadrada da variância. Este número é o desvio padrão da amostra. Como foi colocado anteriormente, para obter o EPM basta dividir o DP pela raiz quadrada de N (neste caso seria a raiz quadrada de 25).

Quanto menor a amostra maior será o intervalo de confiança, com conseqüente menor credibilidade do valor encontrado. Por exemplo, digamos que o pesquisador A encontrou 68 Kg de média e um intervalo de confiança de 95% de $\pm$ 2 kg . Portanto, ele pode ter uma confiança de 95% que a média da população se encontra entre 66 Kg e 70 Kg. Neste exemplo, a média verdadeira (70 Kg) realmente se encontra neste intervalo. Se ao invés de 5 médicos, ele selecionasse apenas 1 médico de cada especialidade (total de 5 médicos) e por acaso encontrasse a mesma média de 68 kg, o intervalo de confiança de 95% poderia subir, por exemplo, de $\pm$ 2 kg para $\pm$ 8 kg e o pesquisador teria que publicar seu resultado como 68 $\pm$ 8 Kg (IC 95%), que inclui também a média verdadeira. O problema é que na maioria das vezes nós não sabemos qual é a média verdadeira e, quanto menos incerteza, refletida pelo menor o intervalo de confiança, melhor.

Problemas comuns com os testes estatísticos

Vamos comentar agora alguns problemas comuns na aplicação dos testes estatísticos. Um dos testes mais usados é o t de student. Este teste é utilizado para comparar médias de 2 grupos quando a variável é medida em nível intervalar ou de razão e a amostra tem uma distribuição normal. Não é adequado usar este teste para variáveis com mensuração em nível ordinal (p.ex. pontuar dor no pós-operatório) ou que os dados da amostra não tenham uma distribuição normal. No caso das variáveis ordinais devemos utilizar um teste não-paramétrico similar ao t de student (por exemplo o teste de Mann-Whitney) e no segundo caso podemos usar o Mann-Whitney ou transformar a variável (log , raiz quadrada, entre outras..) para que ela assuma uma distribuição normal. Outro erro comum no teste de t de student é a comparação dois a dois quando se tem três ou mais grupos. Por exemplo, ao se comparar a média de peso de três grupos diferentes (A, B, C) os pesquisadores usaram o t de student para comparar a média do grupo A com a do grupo B, depois B com C e posteriormente A com C. O pesquisador assume habitualmen-

te um erro de 5% para cada comparação, tendo um erro global de 15%, o que é inaceitável. O correto seria usar a análise de variância (ANOVA) para comparar a média dos três grupos e constatar se há diferenças. Com o uso da ANOVA nós podemos detectar que existe uma diferença global, mas não sabemos qual grupo difere de qual. Para saber qual grupo difere dos outros poderíamos usar o teste t de student comparando cada dois grupos, tendo o cuidado de não incorrer no erro de múltiplas comparações. Para isso pode-se usar vários artifícios estatísticos, como a correção de Bonferroni, Tukey e Student-Newman-Keuls, entre outros. Outro erro na escolha dos testes estatísticos é não levar em consideração se os grupos são dependentes (pareados) ou independentes. Existe um teste t de student diferente para cada uma dessas situações. O emprego errôneo pode levar a um falseamento dos resultados e consequentemente das conclusões. Os grupos pareados normalmente se formam pela comparação de um grupo pré-tratamento com o mesmo grupo pós-tratamento.

Para finalizar é importante citar algumas vantagens da análises multivariadas sobre as análises univariadas. Por enquanto comentamos somente sobre testes estatísticos univariados. A desvantagem básica destes testes como o qui-quadrado, Fisher e t de student, é que eles não fazem uma abordagem global do problema. A maioria dos experimentos biológicos são complexos e muitas vezes existem interações entre os fatores causais. Por exemplo, numa pesquisa para determinar se um medicamento é eficaz para perder peso, selecionam-se obesos para o grupo tratamento e grupo controle. Após análise estatística com o teste t de student em relação à diminuição do peso nos dois grupos, verifica-se que o grupo tratamento é superior. Porém, quando se analisa com testes multivariados observa-se que o medicamento em questão não influencia a perda de peso quando se controla (ou se ajusta) o experimento pelo grau de vontade de emagrecer, que foi medido no questionário. Esse controle estatístico é possível com uso de técnicas como a regressão múltipla. Nesta técnica é possível a avaliação de várias variáveis ao mesmo tempo – uma controla o efeito da outra. Mesmo que o teste t student tenha sido aplicado corretamente, a conclusão do teste foi equivocada porque não se levou em consideração outras variáveis que também influenciam na perda de peso. Pela análise univariada a vontade de emagrecer também foi estatisticamente significativa e, por isso, o pesquisador publica que tanto a vontade de emagrecer quanto o medicamento são eficazes. Porém, como foi verificado na análise multivariada, o efeito da vontade de emagrecer (p.ex., o paciente faz dieta mais rigorosa) anulou o efeito do medicamento. Isto ocorre porque quase todo efeito do emagrecimento pôde ser explicado pela vontade de emagrecer e o efeito aditivo do medicamento não foi suficiente para ser significativo. Este cenário só pode ser captado pela técnica multivariada. Os testes estatísticos multivariados são mais complexos e trabalhosos, necessitando bom conhecimento de estatística para sua aplicação e interpretação. Mal aplicados e interpretados podem confundir mais que ajudar. Porém, sem dúvida, são valiosos recursos na obtenção da verdade científica.

LEITURAS SULPEMENTARES

1. Glantz SA – Primer of Biostatistics 4th Edition. McGraw-Hill., New York, 1997.
2. Glantz SA, Slinker BK – Primer of Applied Regression and Analyses of Variance. McGraw-Hill., New York, 1990.
3. Greenhalgh T – How to read a paper. BMJ Publishing Group, London, 1997.
4. Munro BH – Statistical Methods for Health Care Research 3rd Edition. Lippincott, Philadelphia, 1997.

Endereço para Correspondência:
MARCO AURELIO PINHO DE OLIVEIRA
Rua Coelho Neto, 55 / 201
Tel.: (21) 9987-5843
E-mail: maurelio@infolink.com.br