Análise estatística: Série Entendendo a Pesquisa Clínica 1

Statistical analysis: Understanding Clinical Research Series 1

Marco Aurelio Pinho de Oliveira[1]
Luis Guillermo Coca Velarde[2]
Renato Augusto Moreira de Sá[3]

Palavras-chave
Análise estatística
Interpretação estatística de dados
Estatística como assunto

Keywords
Statistical analysis
Data interpretation, statistical
Statistics as topic

Resumo

Demandas crescentes de tempo dos ginecologistas e obstetras reduzem as suas oportunidades para ficar a par da literatura médica. Em contrapartida, alguns referem que não conseguem fazer a leitura de forma crítica. Acreditamos que, fornecendo informações sobre os métodos de pesquisa habitualmente apresentado para os médicos, possibilitaremos um maior senso crítico e condições para reflexão sobre o estudo publicado.

Abstract

Increasing time demands of gynecologists and obstetricians reduce their opportunities to stay abreast of the medical literature. However many doctors refer that they are not able to perform a critical reading. We believe that by providing information about the research methods, we will make possible a greater critical sense and also conditions for reflection about published studies

[1]Professor Adjunto de Ginecologia da Universidade do Estado do Rio de Janeiro (UERJ) – Rio de Janeiro (RJ), Brasil.
[2]Professor Associado do Departamento de Estatística da Universidade Federal Fluminense (UFF) – Niterói (RJ), Brasil.
[3]Professor Associado de Obstetrícia da UFF – Niterói (RJ), Brasil.
Endereço para correspondência: Renato Augusto Moreira de Sá – Universidade Federal Fluminense – Pós Graduação em Ciências Médicas – Avenida Marques de Paraná, 303, 4º andar do prédio anexo – CEP: 24033-900 – Niterói (RJ), Brasil – E-mail: rsa@cpdt.com.br
Conflito de interesses: não há.

Introdução

Demandas crescentes de tempo dos ginecologistas e obstetras reduzem as suas oportunidades para ficar a par da literatura médica. Em contrapartida, alguns referem que não conseguem fazer a leitura de forma crítica. Acreditamos que, fornecendo informações sobre os métodos de pesquisa habitualmente apresentado para os médicos, possibilitaremos um maior senso crítico e condições para reflexão sobre o estudo publicado. Desta forma, a revista FEMINA publicará uma série de artigos com este objetivo, intitulada "Entendendo a Pesquisa Clínica". O primeiro, que apresentamos agora, diz respeito às análises estatísticas. Ao final de cada um dos artigos da série, serão sugeridas leituras complementares.

Quando se inicia a análise estatística dos dados a primeira pergunta óbvia é: "o que quer dizer estatística?". Simploriamente, a estatística se refere ao conjunto de relações calculadas com base nos dados de uma amostra adequada, que deve ser parte representativa de uma população. Ela é necessária devido a condições de incerteza inerentes a situações de tomada de decisão em que não é possível ter informação de toda a população[1].

Entendendo a análise estatística

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 frequências 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. Adicionalmente, a estatística descritiva permite munir a análise inferencial de evidências de possíveis associações na amostra que precisarão ser testadas para sua correspondente generalização.

O papel da estatística inferencial é transferir, generalizar as conclusões da amostra para a população. Para sermos mais objetivos, o interesse maior no dia a dia é de comparar dados entre dois ou mais grupos para saber se houve diferença significativa[1,2].

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), isto pode ser considerado muito ou pouco? A resposta é "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.

As decisões tomadas por inferência estão associadas com probabilidades que tentam minimizar a possibilidade de errar ao tomar uma decisão incorreta. Uma destas probabilidades, talvez a mais conhecida, é o nível de significância. Quem estipula este nível de significância é o pesquisador. No meio acadêmico ficou tradicionalmente estipulado que se a chance de decidir por uma significativa diferença quando esta não existe 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 3.600 mL no primeiro grupo e de 3.400 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 diferença 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 diferença ter ocorrido apenas pelo acaso. Após a realização do teste *t* de Student, verificamos que a probabilidade de encontrarmos uma diferença de 200 mL (1.600–1.400 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 I) que se corre nos testes de hipóteses. Porém, se após a realização do teste *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 diferença ter sido ao acaso não é pequena ($p>0,05$), 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);
- É comum que os menos afeitos à estatística confundam o valor de alfa com o erro. Alfa (α) é a probabilidade de co-

meter erro tipo I. Na verdade, deveria dizer que o valor de p é 0,03. Definimos o valor de p como a probabilidade de observar aqueles dados caso a hipótese nula seja verdadeira (ou seja, não teria diferenças entre os grupos). Neste caso será a probabilidade de observar uma diferença de 200 mL quando, na verdade, não existe efeito diurético significativamente maior que o placebo.

Quanto menor a amostra, menor o poder, isto é, menor a probabilidade de tomar a decisão correta quando o tratamento é realmente eficaz. Ou seja, o tratamento pode ser de fato eficaz; porém, o pequeno número de participantes na amostra pode não permitir 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, que nos levaria a tomar a decisão errada de que o medicamento não é eficaz, ou seja, p poderia ter sido menor que 0,05. Porém, como a probabilidade de erro tipo II é grande, decorrente de um poder baixo, podemos estar afirmando que o medicamento é eficaz quando, na verdade, o é, com probabilidade baixa[1-3].

Como escolher o teste estatístico apropriado

Como já sabemos para o quê serve o p fornecido pelos testes estatísticos, vamos nos preocupar agora com a escolha do teste adequado[4]. Para isto, é fundamental que saibamos qual o nível de mensuração das variáveis envolvidas. Podemos dividir, estas variáveis, em três grupos: 1 – Nominal; 2 – Ordinal; 3 – Numéricos[1,3].

Na variável nominal, observamos características que, às vezes, representamos por números, mas 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 deste tipo de variáveis. Esses números representam apenas categorias diferentes. Os testes mais usados nestes casos em que queremos analisar a associação entre duas variáveis nominais são o qui-quadrado (χ^2) e o teste de Fisher, este usado principalmente para amostras muito pequenas[3].

Na variável ordinal, as categorias apresentam uma relação de hierarquia ou ordenação e, assim, os números já podem ser ordenados (por exemplo,. do menor para o maior). Porém, não trazem informação útil, como na classificação da endometriose, a paciente que recebe 40 pontos não tem o dobro de endometriose do que a paciente que recebeu 20 pontos. Entretanto, 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); em ambos, casos é necessário que os grupos que estão sendo comparados não apresentem medições provenientes dos mesmos indivíduos. Estes testes não se utilizam de distribuições de probabilidades para representar a população (não requerem, por exemplo, distribuição normal) e são denominados de não-paramétricos.

O terceiro grupo está formado pelas variáveis numéricas. Estas podem ser contínuas, que são geralmente provenientes de mensurações, e as discretas, decorrentes de contagens. Para exemplificar as primeiras, podemos citar o peso medido em quilos; e as outras, o número de filhos. 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 pressupõem uma distribuição normal para os dados, eles são chamados de testes paramétricos. Caso as medidas dos diferentes grupos sejam provenientes de um mesmo conjunto de indivíduos, será necessário escolher testes específicos para o caso de dados que em estatística são conhecidos como pareados. Isto acontece quando, no exemplo da avaliação do medicamento diurético, cada indivíduo tiver seu volume urinário comparado antes e depois de utilizado o medicamento que está sendo testado. Por outro lado, a propriedade de normalidade da distribuição dos dados precisa ser testada, o que leva a utilizar testes de aderência como Shapiro-Wilks ou Kolmogorov-Smirnov. Caso estes testes não aceitem a hipótese de normalidade para os dados, se faz necessário a escolha de versões não paramétricas dos testes citados anteriormente, como os já citados Mann-Whitney e Kruskal-Wallis[3].

Entendendo intervalo de confiança

Outro assunto que merece ser abordado é o intervalo de confiança (IC)[4]. 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 as características da 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[3].

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 muito grande. Se continuássemos a fazer outras pesquisas idênticas, teríamos várias médias (por exemplo, 66, 68, 70, 72 e 74 kg) que, no seu conjunto e sob determinadas condições, apresentam a propriedade de terem 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 se distribuir 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 nas médias de amostras de um determinado tamanho, pois, como já discutimos, a média de uma amostra não é necessariamente idêntica à média real da população[1-3]. O intervalo de confiança está definido por um par de números que, com certo grau de confiança, medido pelo chamado "coeficiente de confiança", contém o verdadeiro valor do parâmetro ou característica populacional que no caso é a média. Habitualmente, se utiliza o intervalo de 95% de confiança (IC95%) (α=5%).

O pesquisador A, que encontrou uma média de 68 kg na sua amostra, diria que a média da população (100 médicos) deve estar ao redor de 68 kg e mais ou menos alguma margem de erro. Esta margem de erro pode ser calculada usando-se um valor da distribuição *t* de Student associado ao valor α=5%. Para uma amostra de 25 indivíduos, o que implica usar 24 graus de liberdade, o valor fornecido pela tabela da distribuição t é igual a 2,064. Este valor deve ser multiplicado pelo erro padrão da média (EPM), que pode ser calculado dividindo-se o desvio padrão da amostra pela raiz quadrada do número de indivíduos na amostra. Se o EPM fosse igual a 1, a margem de erro seria igual a 2,064. Portanto, teríamos 95% de certeza que a média da população estaria entre 68±2,064 kg, ou seja, aproximadamente entre 66 e 70 kg (neste caso o intervalo de 95% incluiu o valor verdadeiro – 70 kg).

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[3]. O DP expressa a variabilidade das observações 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 98 kg, você deve subtrair 98-68 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 a amplitude do intervalo de confiança, com consequente menor credibilidade do valor encontrado. Por exemplo, digamos que o pesquisador A encontrou 68 kg de média e uma margem de erro de ±2 kg. Portanto, ele pode ter uma confiança de 95% que a média da população se encontra entre 66 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 ter uma margem de erro maior (por exemplo, de ±2 para ±8 kg) e o pesquisador teria que publicar seu resultado como 68±8 kg (IC95%), 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 pela menor amplitude do intervalo de confiança, melhor.

Problemas comuns com os testes estatísticos

Vamos comentar agora alguns problemas comuns na aplicação dos testes estatísticos[4]. Um dos testes mais usados é o *t* de Student. Este teste é utilizado para comparar médias de 2 grupos quando a variável é numérica e tem uma distribuição normal. Não é adequado usar este teste para variáveis com

mensuração em nível ordinal (por exemplo, pontuar dor no pós-operatório) ou quando os dados da amostra não tenham uma distribuição normal. No caso das variáveis ordinais, cujas categorias são representadas por números, devemos utilizar um teste não-paramétrico similar ao *t* de Student (por exemplo, o teste de Mann-Whitney) ou transformar a variável (log, raiz quadrada, entre outras transformações) para que ela assuma uma distribuição normal.

Outro erro comum no teste *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 habitualmente um nível de significância de 5% para cada comparação, mas o nível de significância geral é obtido de uma conta difícil de realizar, mas que, com certeza, não é 5%. 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, caso esta diferença for significativa, 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 ou os testes de Tukey ou 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[1,5].

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 χ^2, 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 técnicas que consideram diversas variáveis simultaneamente, 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 da influência de várias variáveis ao mesmo tempo sobre uma que é chamada de "resposta" (cada variável influenciadora controla o efeito da outra). Mesmo que o teste *t* de 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 (por exemplo, o paciente faz dieta mais rigorosa) anulou o efeito do medicamento. Isto ocorre porque quase todo efeito do emagrecimento poderia 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. As técnicas estatísticas multivariadas são mais complexas e trabalhosas, necessitando bom conhecimento de estatística para sua aplicação e interpretação. Mal aplicadas e interpretadas, podem confundir mais que ajudar. Porém, sem dúvida, são valiosos recursos na obtenção da verdade científica[2,5].

Leituras suplementares

1. Glantz SA. Primer of Biostatistics. New York: McGraw-Hill; 1997.
2. Greenhalgh T. How to read a paper. London: BMJ Publishing Group; 1997.
3. Munro BH. Statistical Methods for Health Care Research. Philadelphia: Lippincott; 1997.
4. Oliveira MAP, Camara RCM. Noções Básicas de Bioestatística. Brazilian Journal of Videoendoscopic Surgery. 2010;4(1):5-8.
5. Glantz SA, Slinker BK. Primer of Applied Regression and Analyses of Variance. New York: McGraw-Hill: 1990.