RIEE
ANIMAIS E HUMANOS

ARTIGO ORIGINAL

QUESTÕES EM BIOESTATÍSTICA: ALGUMAS REFLEXÕES SOBRE A COMPARAÇÃO DE MÉDIAS

BIOSTATISTICAL QUESTIONS: SOME REFLECTIONS ON THE COMPARISON OF MEANS

*Carlos Alberto Mourão Júnior**

Resumo

O presente artigo analisa criticamente os usos e abusos dos testes estatísticos de comparação entre médias, bem como sugere novos enfoques metodológicos para a coleta e análise de dados objetivando alcançar resultados mais fidedignos.

Palavras-chave

Bioestatística. Pesquisa. Metodologia. Análise Quantitativa. Biometria.

Abstract

This paper critically examines the uses and abuses of statistical tests to compare means, and suggests new methodological approaches to collect and analyze data in order to achieve more reliable results.

Key-words

Biostatistics. Research. Methodology. Quantitative Analysis. Biometry

A comparação de médias entre dois ou mais grupos é uma das modalidades de análise inferencial mais utilizadas na literatura biomédica. Entretanto, apesar da comparação de médias ser um procedimento considerado trivial, corriqueiro e disponível em praticamente todos os software de estatística disponíveis, não é incomum encontrarmos resultados discrepantes mesmo em estudos com desenhos semelhantes, publicados em periódicos conceituados. O motivo das discrepâncias encontradas, as quais motivam o grande número de meta-análises presentes na literatura atual, pode estar no processo de amostragem, no próprio desenho metodológico do estudo ou na escolha equivocada de qual teste estatístico utilizar em cada circunstância. Neste artigo discutiremos alguns usos e abusos dos testes de comparação entre médias.

Para iniciar, vamos imaginar o seguinte estudo hipotético: um pesquisador ministrou aloxano (uma droga que destrói as ilhotas pancreáticas levando ao diabetes melito) a 20 ratos. No dia seguinte dividiu a amostra em dois grupos (A e B), ministrando a cada um deles uma droga com suposto potencial de reverter o diabetes (drogas A e B) e no final de alguns dias dosou a glicemia capilar de jejum dos animais a fim de verificar qual das duas drogas foi mais eficiente em reduzir os níveis de glicose.

Vamos supor que, por motivos operacionais, as glicemias não tenham sido medidas no mesmo dia, e sim ao longo de um período de tempo que variou alguns dias entre a administração do aloxano e a dosagem da glicemia de jejum. De acordo com a pergunta do estudo, a variável resposta é o valor da glicemia de jejum e a variável preditora é a droga utilizada (A ou B). Os dados brutos podem ser vistos no Quadro 1.

Droga utilizada	Glicemia de jejum (md/dl)	Tempo (dias)
A	160	18
A	154	19
A	151	19
A	163	21
A	150	19
A	164	26
A	139	18
A	121	17
A	137	19
A	132	15
B	121	20
B	124	18
B	123	14
B	142	17
B	123	16
B	121	15
B	104	13
B	105	12
B	109	11
B	107	11

Correspondence author: Carlos Alberto Mourão Júnior. Departamento de Fisiologia – ICB. Universidade Federal de Juiz de Fora. 36036-900, Juiz de Fora – MG. camouraojr@gmail.com. (32) 2102-321. (32) 8871-9031.

* Ph.D. Departamento de Fisiologia – Universidade Federal de Juiz de Fora

Received: 10/2010
Accepted:11/2010

Em uma situação como esta a tendência natural seria realizar o teste t para amostras independentes. O teste t forneceu o seguinte resultado: t = 4,98 para 17 graus de liberdade; p < 0,001 e o intervalo de confiança a 95% da diferença entre as médias foi de 16,8 a 41,9 mg/dl. Diante desses resultados a conclusão é de que a probabilidade da hipótese nula ser verdadeira (ou seja, das médias serem iguais na população) é mínima, e portanto pode-se concluir que a droga A é superior à droga B.

Porém, como o tempo entre a administração de aloxano e a dosagem da glicemia de jejum foi diferente entre os animais que compõem a amostra, é importante testar a possibilidade desse tempo se tratar de uma variável de confusão, que poderia ter interferido na variável resposta (glicemia de jejum). De fato, fazendo uma correlação de Pearson entre o tempo e a glicemia de jejum encontramos o seguinte: r = 0,82 e p < 0,001, mostrando que existe uma forte correlação entre as variáveis, ou seja, tudo indica que a variação do tempo pode produzir uma significativa variação na glicemia de jejum, sem levar em consideração a droga utilizada. Podemos concluir então que não é somente a variável droga que pode afetar os valores da glicemia; o tempo também pode afetá-la. Assim sendo, neste caso o tempo fica caracterizado como sendo uma variável de confusão (interveniente). O gráfico de dispersão apresentado no Gráfico 1 ilustra a correlação entre glicemia de jejum e tempo transcorrido entre a administração da droga e a dosagem da glicemia.

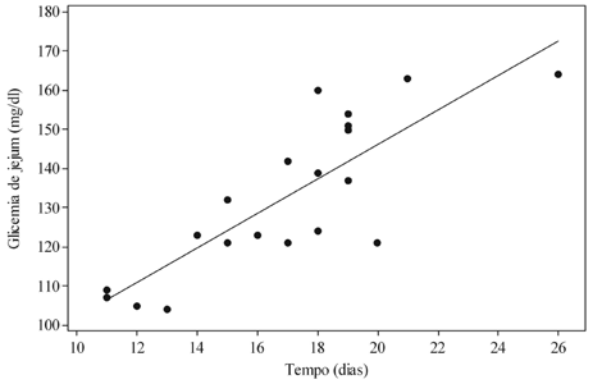

Gráfico 1: correlação entre glicemia de jejum e tempo.

Quando se pretende analisar se uma determinada variável numérica é afetada por um dentre dois fatores utiliza-se a ANOVA com dois fatores para se verificar se existe interação entre os mesmos e se, isoladamente, cada um interfere na variável estudada. Entretanto para que se possa lançar mão da ANOVA com dois fatores (two-way) é necessário que ambos sejam variáveis categóricas, tais como grupo, sexo, tratamento, droga, etc.

Porém, em uma situação como a que estamos analisando, a variável tempo é numérica, e portanto não pode ser tratada como um fator. Por conseguinte, não se aplica nesta caso a two-way ANOVA (tão popular nos experimentos biológicos).

Diante de uma situação como a variação da glicemia em função da droga utilizada, na qual existe uma variável interveniente numérica (no caso, o tempo) torna-se necessária a utilização de um método estatístico pouco visto nas publicações científicas, que é a análise de covariância (ANCOVA), que nada mais é do que a aplicação de uma modelagem linear geral, mesclando ANOVA (análise de variância) com análise de regressão.

Para entender de uma maneira simplificada como funciona a ANCOVA daremos uma explicação geométrica que facilita a compreensão intuitiva desta ferramenta tão importante e tão pouco utilizada na literatura biomédica. Como percebemos que a variável tempo interfere na glicemia de jejum, vamos observar o que acontece quando plotamos a glicemia de jejum em função do tempo, apresentando as respectivas retas de regressão. O resultado pode ser graficamente visualizado no Gráfico 2.

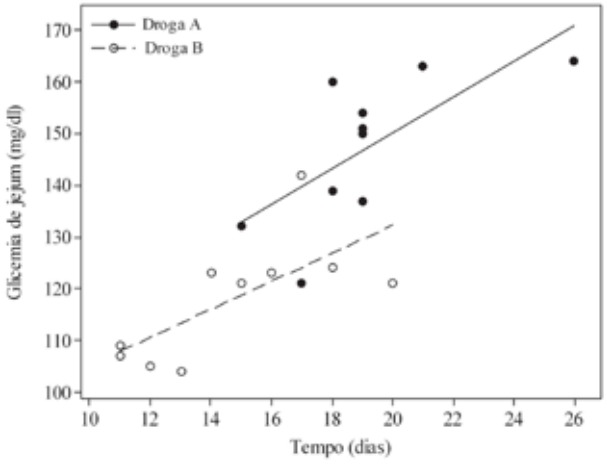

Gráfico 2: Retas de regressão para as drogas A (linha cheia) e B (linha pontilhada).

Observando a figura podemos ver que as retas não são paralelas, ou seja, a diferença entre os níveis de glicemia entre as drogas A e B não é constante ao longo do tempo. A ANCOVA utiliza a mesma modelagem da ANOVA, porém aplica estimadores da regressão linear entre a variável resposta (no caso, a glicemia) e a covariável (no caso, o tempo) a fim de tornar as retas paralelas, tornando assim possível calcular a distância entre as retas. Naturalmente a distância encontrada corresponderá naturalmente à diferença entre as médias, já ajustadas em função da covariável. Em seguida é realizada uma ANOVA convencional com a diferença entre as médias já corrigidas. Fica claro, portanto, que a ANCOVA é um método híbrido, pois se trata de uma ANOVA com dados ajustados pela estimativa da regressão linear. O resultado da ANCOVA no nosso exemplo mostrou um valor de p = 0,011.

Inicialmente pode parecer que, pelo fato da ANCOVA ter também encontrado um resultado estatisticamente significante, ela fez pouca diferença na situação apresentada. Entretanto, convém lembrar que o valor de p precisa ser interpretado dentro do contexto clínico. Em alguns casos uma probabilidade de 1,1% da hipótese nula ser verdadeira pode ter um grande impacto clínico como ocorre, por exemplo, em ensaios com drogas teratogênicas, onde a probabilidade de 1% de uma criança nascer com malformações congênitas é inaceitável.

Porém não nos deixemos iludir pensando que a ANCOVA é a solução mágica para lidar com qualquer variável interveniente numérica. Infelizmente em muitos casos o modelo da ANCOVA não fornece resultados consistentes (tal como ocorre com a ANOVA com dois fatores se a interação for muito expressiva). Por exemplo, se as retas da Figura 2 apresentassem declividades muito diferentes o modelo poderia não ser capaz de realizar o ajuste para torná-las paralelas. Além disso, é fácil perceber que há necessidade da relação das variáveis ser linear (o que nem sempre ocorre com variáveis biológicas). Como fazer então na presença de variáveis intervenientes quando a ANCOVA não se aplica? Tentaremos propor algumas opções para se lidar com essa situação, a qual, lamentavelmente, está muito longe de ser incomum.

Aqui também vale a máxima de que a melhor forma de contornar um problema é evitá-lo, ou seja, as possíveis variáveis intervenientes (de confusão) precisam, ainda na fase de planejamento do estudo, de ser identificadas e, uma vez identificadas, elas deverão ser excluídas (quando possível) ou controladas (balanceadas entre os grupos). No caso do estudo que utilizamos como exemplo, o mais sensato seria utilizar somente animais com o mesmo tempo de exposição ao aloxano, ainda que para isso fosse necessário ampliar a amostra. Em experimentos com animais as variáveis intervenientes podem ser mais facilmente controladas, escolhendo-se, por exemplo, animais irmãos de uma mesma ninhada e do mesmo sexo. Mas e nas pesquisas em humanos, em que podem ocorrer grandes variações genéticas, uma vez que não é possível conseguir amostras compostas somente de irmãos gêmeos?

Neste caso, a melhor estratégia seria trabalhar com amostras pareadas. Infelizmente, talvez por tradição, a maior parte dos estudos publicados utilizam amostras independentes supondo que uma boa amostragem poderia sanar o problema. Entretanto, fazendo uma analogia simplista, trabalhar com amostras independentes pode ser a mesma coisa que comparar maçãs com uvas, uma vez que não há como excluir variabilidades individuais. Talvez isso explique grandes discrepâncias encontradas em ensaios clínicos. Se fossem utilizadas amostras pareadas, isto é, analisando o mesmo indivíduo antes e após uma intervenção, os resultados talvez fossem mais consistentes.

Outro problema que ocorre em muitos estudos, principalmente os multicêntricos, é que tais estudos analisam amostras muito grandes. Existe uma falsa impressão sobre o tamanho da amostra de que "quanto maior melhor." Vejamos como tal afirmação é perigosa. Para tanto, faremos uma simulação para afastar de vez a lamentável ideia de que grandes amostras são a solução para todos os problemas. Como é sabido, basta conhecermos a média, desvio padrão e o número de sujeitos de cada grupo para calcularmos o valor de t.

Vamos analisar uma situação hipotética utilizando o teste t para comparar a variável peso corporal em dois grupos, o primeiro com média de 78 kg e desvio padrão de 12 kg e o segundo com média de 75 kg e desvio padrão de 10 kg. Se cada grupo tiver 30 pacientes o valor de p será 0,297 (não significante), entretanto se cada grupo tiver 200 pacientes o valor de p será de 0,006, considerado extremamente significante, porém refletindo uma diferença clínica de apenas 3 kg. Se o referido estudo estivesse testando drogas antiobesidade será que essa diferença "estatisticamente significante" de apenas 3 kg justificaria afirmar que uma droga é de fato superior à outra? Como podemos ver, o valor de p tem sido hiperestimado, em detrimento do correto julgamento clínico. Nem mesmo o cálculo prévio do tamanho "ideal" da amostra pode substituir o bom senso na interpretação dos resultados, até porque tal cálculo depende de parâmetros da população que normalmente não são conhecidos.

Finalmente, um procedimento estatístico que deveria ser bem mais utilizado na literatura é a categorização de variáveis numéricas quando existirem pontos de corte já conhecidos e validados. Vejamos um exemplo hipotético: desejamos comparar dois grupos de 20 pacientes, cada um com relação aos níveis plasmáticos de glicose de jejum. No primeiro grupo os valores da glicemia de jejum em mg/dl foram: 210, 119, 79, 181, 181, 128, 203, 145, 108, 187, 119, 172, 173, 112, 168, 169, 204, 127, 197 e 125. No segundo grupo os valores foram: 71, 104, 81, 105, 97, 81, 82, 194, 190, 130, 136, 165, 108, 69, 189, 200, 201, 193, 136 e 158 mg/dl. Como as amostras são homocedásticas (apresentam variâncias relativamente homogêneas) e sua distribuição se aproxima da distribuição normal, foi aplicado o teste t não pareado para comparação de médias, o qual produziu um valor de $p = 0,138$, sugerindo não haver diferença entre os grupos.

Entretanto é fato bem estabelecido que os níveis normais da glicemia de jejum vão até 110 mg/dl. Desta forma, o que de fato nos interessa saber é se a proporção de pacientes com glicemia alterada (diabéticos) é diferente nos dois grupos estudados. No primeiro grupo a proporção de diabéticos era de 18/20 enquanto no segundo grupo era de 11/20. Foi então realizado o teste exato de Fisher, que resultou num valor de $p = 0,030$, mostrando que havia diferença entre os grupos quando se levou em conta o ponto de corte com real significado clínico. Atualmente, a maioria das variáveis biológicas já apresenta limites bem definidos para separar grupos normais de grupos patológicos, podendo inclusive avaliar os impactos de várias estratégias terapêuticas, mas infelizmente a simples comparação de médias continua sendo o procedimento mais utilizado, apesar do risco de se ocultarem resultados importantes como acabamos de ver. Além disso, a categorização em dois grupos permite

análises estatísticas mais sofisticadas como a regressão logística, que é capaz de medir a influência de diversas variáveis, fornecendo inclusive medidas como o odds-ratio no caso de variáveis preditoras categóricas.

Concluindo, a fim de se evitarem interpretações erradas dos testes de diferenças de médias, deixamos as seguintes recomendações: tentar controlar as variáveis de confusão, procurar utilizar amostras pareadas, ter cautela ao interpretar significância estatística em grandes amostras e utilizar testes exatos de proporção quando as variáveis contínuas puderem ser categorizadas através de pontos de corte já validados. Lembre-se: por mais sofisticada que seja a análise estatística inferencial dos dados ela jamais substituirá um planejamento adequado do estudo nem tampouco o bom senso na interpretação dos resultados.

Para um maior aprofundamento sobre o conteúdo apresentado sugerimos a consulta às obras citadas nas referências bibliográficas. Todas as análises estatísticas e simulações apresentadas neste artigo foram realizadas através do pacote estatístico Minitab versão 15.

REFERÊNCIAS

GLANTZ, S.A.; SLINKER, B.K. **Primer of applied regression & analysis of variance**. 2. ed. New York: McGraw-Hill, 2001.

HOSMER, D.W.; LEMESHOW, S. **Applied logistic regression**. New York: Wiley, 1989.

KOERTS, J. **On the theory and application of the general linear model.** Rotterdam: Rotterdam University Press, 1969.

MEYER, R.; KRUEGER, D. **A Minitab guide to statistics**. 2. ed. New Jersey: Prentice Hall, 2001.

MOURAO-JUNIOR, C.A. Bioestatística: armadilhas e como evitá-las. **Boletim do Centro de Biologia da Reprodução**, Juiz de Fora, v. 26, n. 1/2, p. 73-76, 2007.

_____. Questões em bioestatística: o tamanho da amostra. **Revista Interdisciplinar de Estudos Experimentais**, Juiz de Fora, v. 1, n. 1, p. 26-28, 2009.

NETER, J. **Applied linear statistical models; regression, analysis of variance, and experimental designs.** Homewood: R. D. Irwin, 1974.

WILDT, A.R. **Analysis of covariance**. Beverly Hills: Sage Publications, 1978.