Revista Eletrônica Nutritime, v.5, n° 6, p.777-788 Novembro/Dezembro 2008.

Artigo Número 76

TESTES ESTATÍSTICOS PARA COMPARAÇÃO DE MÉDIAS

Andréia Fróes Galuci Oliveira[1]

INTRODUÇÃO

Quando a análise de variância de um experimento mostra que as médias de tratamento não são estatisticamente iguais, é apenas lógico perguntar quais são as médias que diferem entre si. O pesquisador em geral gostaria de aplicar um teste para comparar médias, duas a duas. Considere um experimento para comparar três tratamentos, A, B e C. Se a análise de variância mostrar que as médias desses tratamentos não são estatisticamente iguais, é bastante possível que o pesquisador procure um teste estatístico para comparar as médias de A e B, A e C, B e C. Mas como se faz à comparação de médias, duas a duas?

O pesquisador precisa de um método que forneça a diferença mínima significante entre duas médias. Essa diferença seria o instrumento de medida. Toda vez que o valor absoluto da diferença entre duas médias é igual ou maior do que a diferença mínima significante, as médias são consideradas estatisticamente diferentes, ao nível de significância estabelecida (VIEIRA et al., 1989).

A estratégia da análise estatística se define a partir dos objetivos do pesquisador e das condições experimentais então existentes.

A natureza quantitativa dos tratamentos empregados pode sugerir o estudo de modelos (ou funções) que venham a trazer alguma informação adicional ao pesquisador, mas este procedimento, por si só, não substitui a comparação das médias obtidas no ensaio (SAMPAIO, 2002).

Segundo VIEIRA et al. (1989), a comparação de médias só pode ser feita após a análise de variância. Isto porque todos os procedimentos para obter a d.m.s. exigem o cálculo do quadrado médio do resíduo. Mas a análise de variância também dá o valor de F, que permite decidir se as médias são ou não iguais, a determinado nível de significância.

A Análise de Variância é um método suficientemente poderoso para poder identificar diferenças entre as médias populacionais devidas a várias causas atuando simultaneamente sobre os elementos da população (COSTA NETO, 1977).

A possibilidade de utilizar um dos vários testes estatísticos existentes para a comparação de médias, de alguma maneira, seduz o experimentador que, entre comedido e inovador, deseja optar por um deles.

A eleição de um teste pelo pesquisador deve contemplar aquele cujas conclusões que advieram de seu uso sejam menos sujeitas a erros tidos como indesejáveis (SAMPAIO, 2002).

Segundo FONSECA et al. (1982) o método de análise de variância indica a aceitação ou rejeição da hipótese de igualdade das médias. Se a hipótese de nulidade (Ho) for rejeitada, estaremos admitindo que, pelo menos, uma das médias é diferente das demais. Surge, contudo a questão: Quais médias devem ser consideradas diferentes?

[1] Zootecnista, Doutoranda em Produção Animal, Universidade Estadual de Maringá – froesgaluci@hotmail.com

Existem alguns testes para a solução dessa questão, que serão apresentados a seguir.

CONSIDERAÇÕES GERAIS

Segundo SAMPAIO (2002), as situações que conduzem à comparação de valores médios precisam considerar sempre dois aspectos essenciais para a escolha do teste adequado:

1. O primeiro deles refere-se à caracterização da resposta a ser medida (variável alvo) quanto a sua natureza (qualitativa ou quantitativa), a sua distribuição (normal ou não), a sua continuidade (contínua ou discreta) e a sua instabilidade (muito ou pouco instável).
2. O segundo aspecto a ser considerado é que, após a utilização de um teste de comparação de médias, existem dois tipos de erro reconhecidos pela teoria estatística: o erro tipo I (atribuir uma significância quando ela realmente não existir) e o erro tipo II (atribuir uma equivalência quando realmente houver uma diferença significativa).

OS TESTES ESTATÍSTICOS

Todos os testes atualmente em voga fornecem o mesmo resultado quando existem apenas dois tratamentos e, portanto um só contraste a ser avaliado (1 gl ou 1 grau de liberdade).

Os testes de comparação múltipla procuram atender aos interesses da experimentação moderna que utiliza um número maior de grupos investigados devidos às limitações amostral e de infra-estrutura, este número varia entre 3 e 6 na grande maioria dos ensaios, exceto nas competições de variedades em agricultura e em geral em esquemas fatoriais.

Os esquemas fatoriais são casos especiais onde cada fator estudado se apresenta com 2 a 4 níveis, aqui também limitados para atender restrições orçamentárias. O estudo fundamental das interações entre os fatores exigirá a comparação dos níveis de um fator dentro de níveis fixos dos demais. As comparações racionais envolverão, portanto, muito menos médias (de 2 a 4) do que aquelas subitamente sugeridas pelo fatorial completo.

A teoria estatística sempre deixou claro que frente a vários tratamentos (t) apenas os (t-1) graus de liberdade poderiam ser decompostos e testados enquanto na realidade existiam t(t-1)/2 comparações possíveis. Os pesquisadores sempre se interessaram mais por estas t(t-1)/2 comparações entre médias, já que os t-1 contrastes permitidos nem sempre retratavam a comparação de dois tratamentos e sim de grupos deles.

Na realidade percebeu-se que com o aumento do número de tratamento (e, portanto da amplitude da menor para a maior média) aumentava a probabilidade de se detectarem diferenças significativas entre t(t-1)/2 contrastes possíveis. Neste caso, o erro I estaria sendo beneficiado à medida que a distância entre duas médias fosse aumentando.

Todos os estudos desenvolvidos a partir deste alerta preocuparam-se em controlar aquele tipo de erro em particular, criando testes mais rigorosos no controle do erro I e conseqüentemente mais permissivos no do erro II.

Com base no exposto em considerações gerais, passaremos a tecer comentários sobre os testes mais freqüentemente elegidos na experimentação (SAMPAIO, 2002).

O Teste F

Em 1924, Fisher apresentou a fundamentação básica que permitiu a posterior formulação do teste F por Snedecor. Estruturado para avaliar a variação média de uma determinada fonte em relação à variação individual, ou seja:

F = Variância da fonte testada/ Variância do resíduo

O valor assim calculado de F deve ser comparado àquele tabelado segundo os graus de liberdade da fonte testada e do resíduo, respectivamente nas colunas e linhas daquela tabela.

Quando a fonte sendo testada não for uma interação e se redigir a apenas um grau de liberdade, o teste F é adequadamente aplicado ao teste t de Student ($t^2 = F$).

Segundo SAMPAIO (2002), uma fonte de variação com apenas 1 gl pode advir das seguintes situações:

1. Existem apenas dois tratamentos, e, portanto um único contraste a ser feito, equivalente a 1 gl. A significância ou não do teste F executado na análise de variância traduzirá ou não a diferença significativa entre os dois tratamentos.
2. Existem mais de dois tratamentos (t tratamentos, por exemplo), e, portanto existirão t-1 contrastes ortogonais, cada um correspondente a 1 gl. Alguns desses contrastes poderão definir a comparação direta entre dois tratamentos, outros envolverão mais de um deles. Para aqueles contrastes, prevalece o comentário feito no item "1" anterior. Para a os últimos, independentemente do resultado do teste F, as conclusões deverão ser feitas discretamente, obedecendo à natureza do contraste.

Segundo VIEIRA et al., (1989). Para obter a diferença mínima significante (d.m.s.) estabelecida pelo teste t, basta calcular:

$$d.m.s. = t \sqrt{\frac{2 \cdot QMR}{r}}$$

onde t é um valor dado em tabela, QMR é o quadrado médio do resíduo da análise de variância e r é o número de repetições de cada tratamento. Toda vez que o valor absoluto da diferença entre duas médias é igual ou maior do que o valor da d.m.s., as médias são estatisticamente diferentes.

Contrastes Ortogonais

Se existirem cinco tratamentos de natureza qualitativa, apenas 4 gl poderão ser decompostos ortogonalmente (ou seja, sem confundimentos) de modo que as somas de quadrados individuais somadas correspondam a SQTratamento com 4 gl.

Portanto, se existirem mais de dois tratamentos, o teste F só fornecerá uma informação efetiva quando o contraste utilizado envolver apenas dois grupos experimentais.

Uma fonte de variação com mais de 1 grau de liberdade pode ser testada pela razão de variâncias F, entretanto o que estará sendo julgado é a variação média desta fonte. O resultado do teste F avaliará, portanto a magnitude dessa variação média, mas não deixará entrever se existe algum contraste que poderia ser significativo.

Contudo, percebemos que quando utilizamos o teste F para mais de um grau de liberdade, algumas comparações significativas poderão passar despercebidas.

QUADRO 1. Médias segundo o tratamento e comparações.

Tratamento	Médias	Comparações
B	41	A
A	38	Ab
E	33	Abc
C	25	Bc
D	24	C

Fonte: SAMPAIO, 2002.

Nos modelos matemáticos lineares com mais de uma variável independente, cada um desses elementos no modelo (Quadro 1) assume apenas 1 gl, razão pela qual caberia a utilização do teste F. Cada variável independente no modelo seria um contraste, com 1 gl. Esses contrastes poderiam ser de origem qualitativa, mais também podem ser de origem quantitativa. Isto ocorre quando os tratamentos estudados são níveis eqüidistantes e variáveis de um mesmo fator. Neste caso, embora fosse possível sugerir qualquer grupo de contrastes ortogonais, o mais adequado seria investigar o tipo de função observada através dos coeficientes dos polinômios ortogonais.

O teste de F, portanto, para ser devidamente aplicado, exige um profundo conhecimento por parte do pesquisador das condições envolvidas que caracterizam o contraste (ou fonte) por ele testado (SAMPAIO, 2002).

O Teste t

Segundo PIMENTEL GOMES (2000), outro teste clássico é o teste t, que pode ser usado para comparar médias. Como requisitos para sua aplicação conscienciosa temos porém, os seguintes:

1. As comparações feitas pelo teste t devem ser escolhidas antes de serem examinados os dados;
2. Pode-se fazer no máximo tantas comparações quantos são os graus de liberdade para tratamentos, e os contrastes devem ser ortogonais.

Segundo VIEIRA et al. (1989), para obter a diferença mínima significante (d.m.s.) estabelecida pelo teste t, basta calcular:

$$d.m.s. = t \sqrt{2 \cdot QMR / r}$$

onde t é um valor dado em tabela, QMR é o quadrado médio do resíduo da análise de variância e r é o número de repetições de cada tratamento. Toda vez que o valor absoluto da diferença entre duas médias é igual ou maior que o valor da d.m.s., as médias são estatisticamente diferentes.

O Teste t de Student

Embora os estudos iniciais da distribuição de t fossem encetados por Gosset em 1908, citado por SAMPAIO (2002), o domínio, aplicação e divulgação deste teste deveram-se novamente a Fisher em 1926, citado por SAMPAIO (2002).

Duas médias A e B obtidas de grupos experimentais com r_A e r_B observações respectivamente podem ser comparadas pela relação.

$$t = \frac{A - B}{(s_e^2 / r_A) + (s_e^2 / r_B)}$$

onde s_e^2 é a variância do resíduo estimada pela análise de variância.

A existência de vários tratamentos conduzirá à mesma estimativa de s_e^2 , mas à medida que as médias se distanciarem uma das outras, haverá maior chance de ocorrência do erro tipo I. Beneficiar a ocorrência deste erro significa utilizar um teste mais sensível para detectar diferenças significativas.

Quando colocadas em ordem decrescente, apenas as médias adjacentes podem ser comparadas sem aquele risco.

Em conseqüência, este teste não deveria ser aplicado quando existisse um grande número de tratamentos. Na prática, a limitação orçamentária de utilização de poucos tratamentos, abona a utilização confortável deste teste desde que o número de tratamentos não ultrapasse quatro.

Por outro lado, o estudo de respostas muito instáveis (CV>30%), onde o erro tipo II é mais freqüente, poderia se valer deste teste para então contrabalançar tal probabilidade.

As médias comparadas por este teste serão diferentes estatisticamente se o valor calculado de t for maior que aquele tabelado segundo os graus de liberdade do erro (SAMPAIO, 2002).

O valor da diferença mínima significativa seria, portanto:

$$d.m.s._{(Student)} = t_{gl\ erro} \sqrt{\frac{s_e^2}{r_A} + \frac{s_e^2}{r_B}}$$

Se a distribuição de uma população é essencialmente normal (com a forma aproximadamente de um sino), então a distribuição de

$$t = \frac{\bar{x} - M}{\frac{s}{\sqrt{n}}}$$

é essencialmente uma distribuição t de Student para todas as amostras de tamanho n. A distribuição t de Student, geralmente conhecida como distribuição t, é utilizada na determinação da valores críticos denotados por $t_{\alpha 2}$ (TRIOLA 1999).

Segundo BONINI et al. (1972) a distribuição t de Student é utilizada para amostras com um número de elementos inferior a 30 (n < 30).

Segundo TRIOLA (1999), a distribuição t de Student é utilizada também quando o σ é desconhecido e quando a população original tem distribuição essencialmente normal.

O Teste de Student – Newman – Keuls (SNK)

Newman (1939) citado por SAMPAIO (2002), apresentou um teste que contornava os inconvenientes do teste t para ensaios com mais de dois tratamentos. Ajustava o valor de t dependendo da distância entre as médias então ordenadas.

Em uma relação decrescente de t médias, duas delas (A e F) apresentarão diferença significativa se:

$$\frac{|A - F|}{\sqrt{s_e^2 / r}} \geq q_i$$

nde s_e^2 é o valor estimado do quadrado médio do resíduo, pela análise de variância, r é o número de repetições, comum a todos os tratamentos e q_i é o valor tabelado proposto (t ajustado) obtido em função da distância entre as médias (i = p + 2, p sendo o número de médias existentes entre as duas médias comparadas, na relação decrescente) e dos graus de liberdade do resíduo, s_e^2 .

Os valores de q_i diminuem com o aumento dos gl, mas aumentam com a distância entre as médias, corrigindo os excessos de erro tipo I.

A diferença mínima significativa entre duas médias com distância i entre elas e com igual número de repetições é:

$$d.m.s._{(SNK)} = q_i \sqrt{s_e^2 / r}$$

Segundo SAMPAIO (2002), quando as médias comparadas A e B apresentarem diferente número de repetições, a diferença mínima significativa seria:

$$d.m.s._{(SNK)} = q_i \sqrt{\frac{s_e^2}{2}\left[\frac{1}{r_A} + \frac{1}{r_B}\right]}$$

O Teste de Tukey

O teste de Tukey, baseado na amplitude total estudentizada (“studentized range”, em inglês) pode ser utilizado para comparar todo e qualquer contraste entre duas médias de tratamentos. O teste é exato e de uso muito simples quando o número de repetições é o mesmo para todos os tratamentos.

No caso de serem diferentes os números de repetições o teste de Tukey pode ainda ser usado, mas então é apenas aproximado (PIMENTEL GOMES, 2000).

No caso de comparações múltiplas entre amostras de tamanhos iguais, o procedimento mais eficiente parece ser o proposto por Tukey, que utiliza valores críticos da *amplitude studentizada,* que denotamos por q (COSTA NETO, 1977).

Tukey (1953) citado por SAMPAIO, (2002) considerou muito trabalhoso o procedimento proposto por Newman (1939) citado por SAMPAIO (2002), mas compactuava com a preocupação no controle do erro tipo I. A opção proposta de apenas um valor de diferença mínima significativa, a despeito da existência de várias médias, caracterizou o teste como extremamente rigoroso, que embora controlasse muito bem o erro tipo I, permitia o aparecimento do erro tipo II. O valor único proposto por Tukey coincide com o valor máximo do SNK, 0u seja, equivalente à comparação entre a maior e a menor média.

$$d.m.s._{(Tukey)} = q\sqrt{s_e^2 / r}$$

onde q é o valor tabelado por Tukey em função do número de tratamento e dos graus de liberdade do resíduo.

Para obter o valor da diferença mínima significante (d.m.s.) pelo teste de Tukey basta calcular:

$$d.m.s. = q\sqrt{\frac{QMR}{r}}$$

onde q é o valor dado na tabela ao nível de significância estabelecido, QMR é o quadrado médio do resíduo da análise de variância e r é o número de repetições de cada um dos tratamentos. De acordo com o teste, duas médias são estatisticamente diferentes toda vez que o valor absoluto da diferença entre elas for igual ou maior que a d.m.s. (VIEIRA et al., 1989).

Um teste assim rigoroso aplicado em um ensaio com muitos tratamentos (q aumenta com o número de tratamentos) e envolvendo uma variável muito instável (CV>25%) favorecerá sobremaneira o aparecimento do erro tipo II (SAMPAIO, 2002).

Se as médias comparadas A e B tiverem diferente número de repetições, a d.m.s. seria:

$$d.m.s._{(Tukey)} = q\sqrt{\frac{s_e^2}{2}\left(\frac{1}{r_A} + \frac{1}{r_B}\right)}$$

Segundo COSTA NETO (1939) o método de Tukey é aproximadamente 37% mais eficiente que o de Scheffé, para efeito de comparação das médias duas a duas.

No entanto, muito raramente, pode acontecer que, embora o teste F não tenha sido significativo na análise da variância, obtenha-se um ou mais contrastes significativos pelo teste de Tukey (PIMENTEL GOMES, 2000).

Teste de Scheffé

Segundo PIMENTEL GOMES (2000), o teste de Scheffé só deve ser aplicado quando o teste F tiver dado resultado significativo. Se o valor de F obtido não for significativo, nenhum contraste poderá ser significativo, e, pois, a aplicação do teste de Scheffé não se justifica. Quando, porém, o valor de F obtido é significativo, pelo menos um dos contrastes entre tratamentos será significativo. Mas o contraste em questão pode ser muito complicado ou sem interesse prático. E pode ainda acontecer que nenhum dos contrastes entre duas médias apenas seja significativo.

A flexibilidade proposta por Scheffé (1953) citado por SAMPAIO (2002), para comparar qualquer contraste entre médias e permitindo diferentes números de observações por tratamento definiu um teste um pouco mais rigoroso que aquele que Tukey, merecendo, portanto os mesmos comentários com relação ao perigo aumento do erro tipo II.

A diferença mínima significativa para qualquer contraste (que pode ser a comparação entre apenas duas médias) é

$$d.m.s._{(Scheffé)} = \sqrt{(t - 1)\, F.\, Var(contraste)}$$

onde t é o número de tratamentos, F é o valor tabelado de F com (t – 1) e x graus de liberdade, sendo x os graus de liberdade do resíduo (relativos a s_e^2).

Como a variância de um contraste é definida por:

$$Var_{(contraste)} = s_e^2 \; \Sigma \, C^2_i / r_i$$

onde C_i é o coeficiente do tratamento i com r_i repetições. Para a comparação de duas médias oriundas de um igual número de repetições, tem-se:

$$d.m.s._{(scheffé)} = \sqrt{(t - 1)\, F.\, 2\, s_e^2 / r}$$

Este teste tem a vantagem de utilizar os próprios valores do quadro da Análise de Variância, além de poder ser usado no caso de amostras de tamanhos diferentes (COSTA NETO, 1977).

O Teste de Duncan

Duncan (1955) citado por PIMENTEL GOMES, (2000) introduziu um novo teste ou prova para comparação de médias, ao qual chegou depois de uma tentativa anterior. Sua aplicação é bem mais trabalhosa do que a do teste de Tukey, mas se chega a resultados mais detalhados e se discrimina com mais facilidade entre os tratamentos, isto é, o teste de Duncan indica resultados significativos em casos em que o teste de Tukey não permite obter significação estatística. Tal como o teste de Tukey, o de Duncan exige, para ser exato, que todos os tratamentos tenham o mesmo número de repetições.

O teste de Duncan é mais trabalhoso do que o teste t e o de Tukey porque exige o cálculo de diversas diferenças mínimas significantes. Para aplicar o teste de Duncan é preciso primeiro ordenar as médias. Calcula-se então a diferença mínima significante (d.m.s.) para comparar a maior média com a menor. No conjunto ordenado das médias, a comparação entre a maior e a menor média corresponde a um intervalo que abrange

todas as k médias. Se a diferença entre a maior e a menor média é significante, calcula-se outra d.m.s., agora para comparar médias em um intervalo abrangendo k – 1 médias, e assim pó diante.

Sempre que duas médias não são estatisticamente diferentes, não se podem testar as diferenças entre médias que estão no intervalo delimitado por aquelas duas médias. Para fazer o teste é usual escrever as médias em linha e em ordem crescente (ou decrescente). Toda vez que a diferença entre duas médias não é significante sublinha-se o intervalo delimitado por essas duas médias; cada comparação não-significante deve ser indicada por uma linha, distinta das demais; médias sublinhadas por uma mesma linha não são estatisticamente diferentes e as diferenças entre elas não podem ser testadas (VIEIRA et al., 1989).

Duncan (1955) citado por SAMPAIO (2002) sugeriu a utilização de um teste baseando-se na mesma argumentação do teste de Student – Newman – Keuuls. A percepção e o controle do erro tipo I era sem dúvida flagrante no SNK, mas a comparação de médias mais afastadas criava uma oportunidade para o aparecimento do erro tipo II (já que o maior valor da diferença mínima significativa se equipara com aquela do teste Tukey).

Segundo VIEIRA et al. (1989), para obter a d.m.s. aplica-se a fórmula:

$$d.m.s. = z\sqrt{\frac{QMR}{r}}$$

onde z é um valor dado em tabela ao nível de significância estabelecido e para o número de médias abrangidas pelo intervalo delimitado pelas médias em comparação, QMR é o quadrado médio do resíduo da análise de variância e r é o número de repetições. Toda vez que o valor absoluto da diferença entre as médias em comparação é igual ou maior que a d.m.s., conclui-se que as médias são estatisticamente diferentes, ao nível de significância estabelecida.

Segundo Duncan (1955) citado por SAMPAIO (2002), Duncan tentou reduzir as d.m.s. impostas pelas comparações mais dramáticas:

$$d.m.s._{(Duncan)} = q_i\sqrt{s_e^2 / r}$$

onde q_i é o valor tabelado por Duncan (ou seja, um t ajustado) obtido em função da distância entre as duas médias a serem comparadas (i = p + 2, p = número de médias localizadas entre as duas que estão sendo comparadas, quando relacionadas em ordem crescente) e r é o número comum de observações por tratamento. Os valores de q_i não sobem tão rapidamente quanto aqueles do teste Student – Newman – Keuls, controlando assim o aparecimento do erro tipo II na comparação de médias mais afastadas.

Se as médias comparadas A e B contiverem diferente número de repetições, a d.m.s. seria:

$$d.m.s._{(Duncan)} = q_i\sqrt{\frac{s_e^2}{2}\left(\frac{1}{r_A} + \frac{1}{r_B}\right)}$$

Segundo PIMENTEL GOMES (2000), quando o número de médias é avultado (superior a 10, por exemplo) a aplicação do teste de Duncan se torna muito trabalhosa.

O Teste de Dunnett

Segundo VIEIRA et al. (1989), o teste de Dunnett deve ser aplicado toda vez que se pretendem comparar as médias dos tratamentos apenas com a média do controle. Para obter a d.m.s. aplica-se a fórmula:

$$d.m.s. = d\sqrt{\frac{2 \cdot QMR}{r}}$$

onde d é um valor dado em tabela, ao nível de significância estabelecido, QMR é o quadrado médio do resíduo da análise de variância e r é o número de repetições.

Para as comparações múltiplas onde um tratamento serve de referência para os demais, ou seja, deseja-se comparar todos com apenas um. Dunnett (1955) citado por SAMPAIO (2002) sugeriu a seguinte diferença mínima significativa:

$$d.m.s._{(Dunnett)} = D\sqrt{s_e^2 \ \Sigma \ C_i^2 / r_i}$$

onde D é o valor encontrado na tabela de Dunnett proposta em função dos (t – 1) graus de liberdade (t = número de tratamentos) e os graus de liberdade do resíduo (relativos a s_e^2), s_e^2 é a estimativa do quadrado médio do resíduo obtida da análise de variância e C_i é o coeficiente utilizado no contraste i com r_i repetições.

Quando todos os tratamentos têm igual número de observações:

$$d.m.s._{(Dunnett)} = D\sqrt{2 \ s_e^2 / r}$$

que é semelhante ao teste t de Student, exceto pelo valor de D, aqui ajustado para um maior número de tratamentos.

A utilização deste teste, que impede outras comparações não contidas em suas condições iniciais é pouco freqüente dado a esta limitação (SAMPAIO, 2002).

O Teste de Bonferroni

Segundo PIMENTEL GOMES (2000), o teste de Bonferroni é um aperfeiçoamento do teste t. Com efeito, recomendamos que só se aplique o teste t a contrastes escolhidos previamente, antes de serem examinados os dados, e que tais contrastes, em número no máximo igual ao de graus de liberdade para tratamentos, devem ser ortogonais.

ALGUMAS CONSIDERAÇÕES (VANTAGENS E DESVANTAGENS DOS TESTES)

A escolha do método adequado para comparar médias exige, que se leve em consideração tanto o nível de significância como o poder do teste. O nível de significância de um teste é a probabilidade de rejeitar a hipótese de que as médias são iguais, quando esta hipótese é, na realidade, verdadeira. Já o poder do teste é a probabilidade de

rejeitar a hipótese de que as médias são iguais quando esta hipótese é, na realidade, falsa.

Mas a escolha do procedimento adequado para comparar médias exige distinguir ainda nível de significância para comparações de médias e nível de significância para experimentos.

Se for escolhido, para a comparação de médias, ou o teste de Tukey ou o teste de Dunnett, o nível de significância para experimentos será de 5%, mas o nível de significância para comparações de médias será menor que 5%. Por outro lado, se for escolhido, para comparação de médias, ou o teste t ou o teste de Duncan, o nível de significância para comparações de médias será de aproximadamente 5%, mas o nível de significância para experimentos será maior do que 5%. Em compensação, o poder do teste também será maior (VIEIRA et al., 1989).

Segundo PIMENTEL GOMES (2000), os testes de Tukey e de Duncan têm fundamentos muitos semelhantes. O teste de Duncan é, porém, menos conservador, isto é, dá diferenças significativas com mais facilidade. Já o teste de Tukey, mais exigente, temos sempre uma probabilidade de 95% de não apontar como significativa uma diferença realmente nula entre todas as médias de tratamentos. Não é de admirar, pois, que, com n > 2 o teste de Duncan dê resultados significativos em casos em que isto não acontece com o teste de Tukey: é que então o teste de Duncan nos leva a afirmativas erradas com maior freqüência. O mesmo acontece, com perigo muito maior, com o teste t aplicado indiscriminadamente, o que hoje não se aceita mais. O teste de Duncan estabelece, pois, um meio termo entre o rigor um tanto excessivo do teste de Tukey e a falta de rigor exagerado do teste t usado sem as devidas cautelas.

Se o pesquisador quer ter alta chance de rejeitar a hipótese de que as médias são iguais, pode optar pelo teste t ou pelo teste de Duncan. Estes dois testes têm características similares, mas o teste t é mais antigo e, talvez por isso, mais conhecido. Também é de aplicação mais fácil. Entretanto, o pesquisador também pode optar por aplicar o teste de Tukey ou de Dunnett, com nível de significância mais elevado. Estes testes teriam, então, maior poder. Por exemplo, o teste de Tukey a 10% tem maior poder que o teste de Tukey a 5%.

Se o pesquisador só pretende rejeitar a hipótese de que as médias são iguais com muita confiança, deve optar pelo teste de Tukey ou de Dunnett, com baixo nível de significância.

O teste t, apresentado aqui, é uma extensão do teste t de Student, feita por Fisher. Muitos autores consideram errado aplicar o teste t para comparação de médias, duas a duas. Na verdade, não existe propriamente erro: apenas o nível de significância para experimentos se torna, nesse caso, muito elevado (VIEIRA et al., 1989).

Já o teste de Scheffé é ainda mais perigoso desaconselhável para a comparação de duas médias, mas presta bons serviços para provar contrastes mais complicados, e para isto é de uso indicado.

O teste de Bonferroni é muito bom para pequeno número de contrastes, mas excessivamente conservador quando esse número cresce. Por exemplo, com 6 tratamentos se podem obter 15 contrastes diferentes entre duas médias. O teste de Bonferroni aplicado a 10 ou mais desses contrastes é mais conservador do que o de Tukey. Mas não se esqueça que o teste de Bonferroni se pode aplicar também a contrastes mais complexos (PIMENTEL GOMES, 2000).

De qualquer forma, fica aqui um alerta: todos os procedimentos para a comparação de médias têm vantagens e desvantagens. Ainda não existe um teste definitivamente "melhor" que todos os outros. Convém até lembrar que um estatístico

conhecido recomendou não usar testes de hipóteses que se propõem a comparar médias, duas a duas. Já outro estatístico de renome considerou que os procedimentos para comparar médias devem ser vistos mais como indicadores do que como soluções exatas. No entanto, é preciso adotar um procedimento formal para comparar médias. Isto evita que as conclusões fiquem totalmente dependentes da opinião do pesquisador. Mesmo assim, existe uma grande margem de opção tanto na escolha do teste, como no estabelecimento do nível de significância (VIEIRA et al., 1989).

REFERÊNCIAS CONSULTADAS

BONINI, E. E.; BONINI, S. E. **Estatística**. Edições Loyola. São Paulo, 1972. 439p.

COSTA NETO, P. L. O. **Estatística**. São Paulo: Edgard Blücher, 1977. 264p.

FONSECA, J. S.; MARTINS, G. A. M. **Curso de estatística**. 3. ed. São Paulo: Atlas, 1982. 286p.

PIMENTEL GOMES, F. **Curso de estatística experimental.** 14ª ed. Piracicaba – SP: Editora da Universidade de São Paulo, 2000. 477p.

SAMPAIO, I. B. M. **Estatística aplicada à experimentação animal**. 2ª.ed. Belo Horizonte: Fundação de Estudo e Pesquisa em Medicina Veterinária e Zootecnia, 2002. 265p.

TRIOLA, M. F. **Introdução à estatística**. Rio de Janeiro: JC, 1999. 410p.

VIEIRA, S.; HOFFMANN, R. **Estatística experimental.** São Paulo: Atlas, 1989. 175p.