A inferência estatística múltipla e o problema da inflação do nível de alfa. A ANOVA como exemplo

Teresa Garcia-Marques
Instituto Superior de Psicologia Aplicada

Manuela Azevedo
Instituto Português de Investigação Marítima

Resumo. – A análise de resultados de um estudo que envolva mais de duas condições experimentais sugere um conjunto de comparações entre os diferentes grupos. Estas comparações envolvem um problema de inferência estatística múltipla que se relaciona com a inflação da probabilidade de cometer um erro de Tipo I. Este artigo pretende clarificar tal problema, referindo alguns procedimentos estatísticos que se demonstra serem mais eficazes nas situações de inferência múltipla, mais comuns em Psicologia.

A investigação em psicologia baseia-se cada vez mais em delineamentos experimentais que envolvem o estudo de mais do que dois grupos naturais, mais do que duas condições experimentais ou mais do que duas variáveis de medida, vendo-se o investigador frequentemente envolvido num problema de inferência estatística múltipla. No entanto, este nem sempre se apercebe de todas as implicações. Nos últimos cinquenta anos, não só se têm desenvolvido diferentes procedimentos especificamente orientados para fazer face a este problema, como a literatura de investigação em psicologia tem sido sensibilizada à sua utilização. O mesmo tipo de evolução pode ser verificado com a actualização dos diferentes programas estatísticos para análise e tratamento de dados, que integram nas suas versões mais recentes um maior número de procedimentos para comparações múltiplas.

O presente artigo pretende clarificar o que se entende por um problema de inferência estatística múltipla, reportando-o às situações mais vulgarmente encontradas na análise de resultados decorrentes de investigações em psicologia. Não será, no entanto, exaustivo na referência aos diferentes métodos ou procedimentos especificamente concebidos para lidar com situações de comparações múltiplas. Assim, serão apenas referidos alguns dos procedimentos que se demonstram serem mais eficazes/adequados a algumas das situações de inferência múltipla, mais comuns em psicologia. De facto, é necessário atender a que os procedimentos desenvolvidos se baseiam num conjunto de pressupostos que de alguma forma limitam a sua eficiência, de-

finem contextos de aplicação mais ou menos específicos e impõem uma posição mais ou menos conservadora (no sentido de terem forte probabilidade de não rejeitarem a hipótese nula). Procuraremos não só apresentar e descrever estes procedimentos, mas também fornecer alguma informação que facilite ao investigador a selecção do mais apropriado à sua situação concreta.

1. *O problema de inferência estatística múltipla.*

Na presença de um conjunto de dados, o investigador coloca uma série de questões relativas tanto às suas características específicas como às suas características relacionais. A resposta a estas questões, regra geral, envolve a estimação de um grande número de parâmetros (médias, desvios padrões, correlações, etc.), bem como o teste de várias hipóteses estatísticas (a análise da significância dos efeitos observados). O conjunto dos dados a que o investigador tem acesso não é independente da forma como planeou/ /delineou o seu estudo. Este delineamento ou planeamento experimental permite o estudo de uma rede de relações com um significado teórico ou prático, a hipótese de estudo. O investigador espera que os dados se apresentem com o padrão definido nessa hipótese; por exemplo, o investigador espera que se definam certas diferenças entre as médias de apenas alguns dos diferentes grupos estudados. A presença de um ou outro padrão é que é relevante para as conclusões que o investigador pretende retirar, pelo que as diferentes questões que o investigador coloca aos dados não são verdadeiramente independentes (sê-lo-iam se cada questão estivesse associada a um delineamento experimental diferente). Não sendo independentes, as várias questões são abordadas simultaneamente, sob uma mesma amostra, pelo que devem ser testadas em simultâneo ou, pelo menos, por um método ou procedimento que tenha em conta as suas relações. Vejamos o caso de um delineamento factorial (n x r): o procedimento habitual para testar o modelo ANOVA (Análise de Variância) consiste em testar em simultâneo os efeitos principais e a sua interacção. O procedimento do teste F da ANOVA fornece-nos uma informação muito geral sobre estes efeitos, ou seja, uma informação relativa à probabilidade de o efeito se verificar por acaso, traduzido no grau de confiança atribuído à rejeição da hipótese nula. Este tipo de análise estatística não nos esclarece, contudo, sobre as características específicas deste efeito, pois não permite, *per se*, a realização de com-

parações específicas entre médias. Caracterizar um efeito complexo (ex.: uma diferença que envolva 3 ou mais médias), identificado por uma ANOVA, implica a realização de um conjunto de comparações ou análises parciais com vista a obter informação mais particularizada. Deste modo, deverão realizar-se testes múltiplos, com o objectivo de compreender o verdadeiro padrão emergente dos dados, o que requer uma abordagem na área da inferência estatística múltipla.

Os problemas relacionados com a inferência múltipla são tanto mais frequentes, quanto maior o número de variáveis envolvidas no estudo e quanto menor a precisão com que as hipóteses do investigador são formuladas. A utilização de estatísticas como a média e a variância, para resumir os dados relativos a cada uma das variáveis, ou do coeficiente de correlação para ilustrar a relação entre duas variáveis vai, regra geral, para além do mero papel descritivo, pois frequentemente se utilizam estas estatísticas para fazer inferências sobre os parâmetros da população. De facto, é frequente em investigações em psicologia encontrarem-se as seguintes atitudes: 1) procurar, «como que apalpando no escuro», quais os parâmetros significativos (significativamente diferentes de zero); 2) comparar entre si os desempenhos de diferentes grupos (testar um conjunto de elevado número de efeitos) numa dimensão; ou 3) comparar dois grupos quaisquer, relativamente a um grande conjunto de variáveis com o fim de desvendar diferenças em cada variável e nas relações entre variáveis. Qualquer destas situações levanta problemas relacionados com a inferência múltipla, para os quais nem sempre o investigador está alertado.

Examinemos, em primeiro lugar, o que se entende por problema de inferência estatística no que concerne o facto de este envolver a possibilidade de erro. Pode definir-se um problema de inferência estatística do seguinte modo: *Admitindo a probabilidade de obtermos um determinado resultado por mero acaso, queremos averiguar, testar, se essa probabilidade é suficientemente reduzida, de forma a considerarmos o dado obtido como efeito ou característica consistente.* Assim, estipulamos *a priori* que se podermos afirmar que o efeito observado (padrão relacional dos dados) naquela amostra ocorre por acaso apenas 5 por cento das vezes (o nosso alfa nominal) em que observamos amostras obtidas em condições idênticas, então, o efeito surgiu por uma outra razão que não a do mero acaso. Nesta situação, afirmamos ter 95 por cento de confiança de que se trata de um efeito verdadeiro, consistente, capaz de ser replicado. Claro está que se o resul-

tado obtido for a concretização do acaso (o que será sempre desconhecido para nós), estamos a cometer um erro. Ao escolher um nível de significância (alfa) de 5 por cento, estamos a considerar não ser grave a possibilidade de errarmos 5 vezes em 100. Este tipo de erro, o de rejeitar a nulidade de um efeito (especificado na hipótese nula) quando esse efeito realmente não existe, denomina-se erro de primeira espécie ou, mais vulgarmente, *Erro Tipo I.*

Este problema generaliza-se e agrava-se quando se procede, não a uma inferência, mas a um conjunto de inferências relativas a um mesmo padrão de resultados. A questão em causa pode colocar-se nos seguintes termos: se a probabilidade de cometer um erro Tipo I num teste é de 5 por cento, qual será a probabilidade de cometer, pelo menos, um destes tipos de erro em seis testes? Tomemos o exemplo corrente da comparação do desempenho de seis grupos experimentais com um grupo controlo. Tratando-se da comparação de pares de médias em grupos independentes recorremos (ingenuamente) a um teste-*t* para amostras não emparelhadas. Considere-se que, a um nível de significância de 5 por cento, rejeitamos a hipótese nula (*Ho*) nas seis comparações. Qual será, no entanto, a probabilidade de não estarmos a cometer um erro Tipo I ao afirmar serem todos os grupos experimentais diferentes do grupo controlo, i.e., serem as diferenças todas significativas? O cálculo é simples: trata-se da probabilidade de não cometermos um erro Tipo I no primeiro teste, de não cometermos um erro Tipo I no segundo teste, de não cometermos um erro Tipo I no terceiro teste, ..., de não cometermos um erro Tipo I no sexto teste. Note-se que, deste modo, estamos a realizar cada teste como se os outros não existissem, pois os testes estão a ser tratados como «testes independentes». A probabilidade da conjunção de acontecimentos independentes facilmente se demonstra ser idêntica ao produto das probabilidades de cada um desses acontecimentos (ver, por exemplo, Murteira, 1990) pelo que a probabilidade de não cometermos, nesta análise, nenhum erro Tipo I é dada por:

$$(1-0.05)\ (1-0.05)\ (1-0.05)\ (1-0.05)\ (1-0.05)\ (1-0.05) = (1-0.05)^6 = 0.735$$

donde a probabilidade de cometer pelo menos um erro Tipo I, será

$$1 - 0.735 = 0.265.$$

O valor de «alfa» que pretendíamos fixado em 5 por cento (alfa nominal) foi inflacionado para 26,5 por cento (alfa real). Note-se que este caso se refere a seis testes estatísticos realizados sobre um único conjunto de dados e que muitas das investi-

gações recolhem informação sobre um conjunto imenso de variáveis nas quais, por exemplo, se pretendem comparar dois ou mais grupos. Procuram-se as diferenças e para tal realiza-se um grande número de testes. No entanto, habitualmente os resultados apresentados só se reportam aos testes significativos mascarando a probabilidade de erro Tipo I que, na realidade, é bastante mais elevada do que transparece. De uma forma geral, a proporção de erro Tipo I que se comete ao realizar um número *c* de comparações (testes) encontra-se no intervalo[1] (definido pela desigualdade de Bonferroni) compreendido por

$[\alpha; 1-(1-\alpha)^c]$.

A posição da proporção de erro Tipo I neste intervalo é determinada pelo grau de dependência dos efeitos testados (regra geral, uma incógnita). Se os efeitos testados independentemente nas *c* comparações estiverem perfeitamente correlacionados, não se verificará qualquer inflação do erro Tipo I, que será idêntico ao alfa nominal. O extremo superior do intervalo é ilustrado pelo exemplo dado anteriormente (independência total dos efeitos), envolvendo a realização de comparações ortogonais. Qualquer conjunto de comparações que não seja totalmente ortogonal envolve uma inflação do alfa, situada num ponto desconhecido deste intervalo. Em estudos que envolvem múltiplas variáveis dependentes ou múltiplas variáveis de medida[2] a probabilidade de erro Tipo I é também uma incógnita: tendo todas as variáveis sido obtidas dos mesmos sujeitos, estas encontram-se correlacionadas de uma forma desconhecida, pelo que os testes separados não são realmente independentes. Quando as variáveis estão perfeitamente correlacionadas, as múltiplas análises univariadas correspondem a repetições do mesmo teste, pelo que se um erro foi cometido na primeira análise, este vai verificar-se repetidamente. Assim, em caso de dependência total, o erro será sempre alfa (pelo que o alfa real corresponde ao nominal) aproximando-se do outro extremo do intervalo, ou seja, do valor $1-(1-\alpha)^c$ na medida em que se verifica uma maior independência das variáveis em estudo.

1.1. Considerações prévias.

Antes de nos referirmos aos principais procedimentos desenvolvidos com vista a fazer face ao problema de inflação do alfa, é necessário estabelecer alguns conceitos básicos e nomenclatura.

1.2. Diferentes categorias de erro Tipo I.

Podemos considerar a existência de diferentes categorias de proporção de erro Tipo I, em situações que envolvem comparações múltiplas. Consideram-se essecialmente três tipos de proporção de erro (por exemplo Ryan, 1959; Klockars e Sax, 1986; Zwick, 1993), que a seguir se indicam.

Proporção de erro por comparação (αc). Trata-se da probabilidade de rejeitar erradamente *Ho* numa comparação particular entre dois parâmetros e corresponde à probabilidade de cometer um erro Tipo I em cada teste estatístico. Numa perspectiva frequencista, trata-se da percentagem de comparações, em tudo idênticas a esta (replicações experimentais), que se espera serem significativas por mero acaso.

Proporção de erro experimental (experimentwise) (αEW) *e erro familiar* (familywise) (αFW). Trata-se da probabilidade de se associarem conclusões erradas aos resultados de uma dada experiência, ou seja, de se cometer pelo menos um erro Tipo I ao concluir que os resultados do estudo definem um padrão específico (onde se destacam certas diferenças numa ou outra direcção) após a realização de comparações múltiplas. Visto a leitura destes dados experimentais, como um todo, envolver um grande número de comparações, temos para o estudo global, ou seja, para as suas conclusões, uma probabilidade de erro experimental (*experimentwise*) regra geral superior ao erro estipulado por comparação. E, como vimos anteriormente, quanto maior o número de comparações maior a possibilidade de rejeitarmos erradamente *Ho*. Se as comparações forem realizadas para determinados subconjuntos de observações, nomeadamente aqueles que definem um efeito particular, então a proporção de erro associada é designada por *familywise error rate*. Esta terminologia, com origem em Tukey (1953), tem em conta o facto de as comparações poderem pertencer a famílias distintas daquelas que ocorrem quando os estudos são multidimensionais, isto é, quando envolvem mais do que uma «variável independente» e, portanto, mais que uma dimensão[3]. No caso unidimensional (observações para um único efeito) a *familywise error rate* corresponde ao que se designa por *experimentwise error rate*.

Proporção de erro por experiência (αPe). A proporção de erro por experiência, assim designado na literatura inglesa, representa o valor esperado do número de erros que se encon-

tram associados a uma dada experiência. Dependendo do número de comparações que a análise dos resultados de um estudo envolve temos, em média, x comparações onde se rejeitou erradamente *Ho*; esse número x designa a proporção de erro por experiência.

Querer controlar este tipo de erro traduz uma preocupação diferente daquela de controlar o erro experimental. Neste último, a ideia é a de que qualquer conclusão errada põe em causa toda a experiência (Ryan, 1959), como é naturalmente o caso em que se testa a validade de um modelo teórico ao qual se associa um dado padrão de resultados. O controlo da proporção de erro por experiência tem subjacente a ideia de que há que reduzir ao máximo o número de afirmações erradas associadas a um estudo. Nesta perspectiva, não seria problemático cometer, por exemplo, um erro Tipo I. Note-se, no entanto, que não se pode identificar qual das conclusões do estudo está errada. A decisão sobre qual a proporção de erro a controlar fica facilitada pelo facto de se usar o mesmo procedimento para o controlo de ambos os erros (Ryan, 1959).

No texto que se segue consideraremos a seguinte nomenclatura relativamente às diferentes categorias de erro:

α proporção de erro Tipo I que o investigador estabelece como nível de significância pretendido na análise (5 ou 1 por cento, tradicionalmente)

αEW proporção de erro experimental

α_c proporção de erro por comparação ou teste

c número de comparações ou testes a realizar

αPe proporção de erro por experiência.

1.3. Potência de teste.

A preocupação deste artigo é fazer face ao chamado erro Tipo I, que ocorrerá quando rejeitamos *Ho*, sendo *Ho* verdadeira. No entanto, um teste estatístico pode levar-nos a não rejeitar *Ho* quando na realidade *Ho* é falsa. Neste caso estaremos a cometer outro tipo de erro, o erro de segunda espécie ou erro Tipo II (designado por β).

Para podermos afirmar validamente que uma hipótese é verdadeira, devemos assegurar-nos de que tanto o risco de primeira espécie, α, como o de segunda espécie, β, são suficientemente pequenos. Enquanto $1-\alpha$ se refere ao coeficiente de confiança depositado na decisão de rejeição de *Ho* (probabilidade de não estar a cometer, por acaso, um erro Tipo I) $1-\beta$ refere-se à potência de teste

ou capacidade do teste detectar o efeito estudado (probabilidade de não estarmos a cometer, por acaso, um erro Tipo II, quando o resultado do nosso teste sugere a não rejeição da *Ho*) (Cohen, 1988).

Em investigação, regra geral, assumimos uma *atitude conservadora*, na medida em que só aceitamos um efeito como tal, caso a probabilidade de o observar, por mero acaso, seja realmente muito pequena, i.e, impondo um reduzido risco de erro Tipo I (α=5%, 1%). Não devemos esquecer, no entanto, que uma atitude conservadora reduz a potência das comparações que levamos a cabo, pelo que a não rejeição de *Ho* não nos permitirá concluir sobre a validade da conclusão de *ausência de efeito*. A relevância deste ponto, nesta nossa análise, reside no facto de as diferentes técnicas ou abordagens ao problema de inflação do alfa diferirem quanto ao seu papel mais ou menos conservador. Um procedimento muito conservador que nos sugerisse a rejeição de *Ho*, não acarretaria nenhum problema de interpretação. Mas se a sugestão fosse de não rejeição de *Ho*, ficar-nos-ia a dúvida de estarmos ou não a utilizar um instrumento suficientemente potente para detectar a presença de um efeito, caso ele exista na realidade.

Uma sugestão oferecida por Ramsey (1981) e também referida por Zwick (1993) é a de considerar que, tal como existem diversos erros Tipo I, existem igualmente diferentes definições de potência: a potência associada à detecção de uma ou mais diferenças entre médias (potência por comparação) e a potência associada à detecção de todas as diferenças que possam ser discriminadas na experiência (a potência experimental).

1.4. Comparações planeadas *vs* comparações «Post hoc».

Uma distinção que, de certo modo, influencia a selecção de uma abordagem anti-inflaccionária dos erros Tipo I é a distinção entre comparações planeadas ou *a priori* e comparações não-planeadas (exploratórias), *post hoc* ou *a posteriori*.

As comparações múltiplas surgem na análise dos resultados de uma experiência como análises parciais dos dados: testes de significância dos parâmetros, contrastes específicos (ortogonais ou não) que envolvam quer pares de estatísticas quer conjuntos mais complexos (ex.: anovas dentro de manovas).

Uma comparação k (k=1, ..., c número de comparações) com um grau de liberdade, é entendida como uma combinação linear de médias ou contraste, $L_k = a_i X_i + ... + + a_m X_m$, sendo a_i o coeficiente associado ao grupo i e X_i a média do grupo i (a

ausência de um grupo refere-se à associação de um coeficiente zero) e em que $\Sigma\ (n_i\ a_i) = 0$. Um tipo de contraste especial é o que resulta da atribuição de pesos 1 e -1 a apenas duas médias (*pairwise comparisons*). Os outros tipos de contraste são frequentemente designados de contrastes ou comparações complexas. No caso de homogeneidade de variâncias temos que o erro padrão da comparação é dado por

$$se_k = \sqrt{MSe\,\Sigma\frac{a_i^2}{n_i}}$$

em que *MSe* representa o quadrado médio do erro residual da ANOVA. No caso de uma simples comparação de médias em grupos de igual dimensão o contraste virá: $L=X_1-X_2$ e o erro padrão da contraste dado por

$$se_k = \sqrt{\frac{2MSe}{n}}\ .$$

A estatística de teste associada à hipótese nula, *Ho*: $L_k=0$, é definida por

$$t_k = \frac{L_k}{se_k}\ .$$

Na realização de uma investigação há que distinguir, na prática, duas situações: 1) o plano da investigação associa-se claramente a um conjunto de hipóteses empíricas directamente traduzidas em hipóteses estatísticas, as quais definem comparações ou testes *a priori* e, portanto, comparações planeadas (saliente-se que estas comparações são, regra geral, de número reduzido); e 2) o investigador tem pela frente um conjunto de dados, aguça a sua curiosidade e procura (explora) padrões consistentes, não previamente definidos na sua hipótese de estudo; a procura destes padrões resulta na realização de testes *post hoc* (testes planeados apenas após a observação dos dados e como tal concebidos *a posteriori*). Os estudos exploratórios ou estudos que formulam hipóteses sobre relações muito vagas são típicos desta situação, originando um grande número de comparações *post hoc*. Como exemplo de hipótese vaga considere-se a hipótese de não existência de homogeneidade de *x* grupos relativamente a um conjunto de *y* medidas. Caso se pretenda entender a origem da heterogeneidade, ter-se-á de levar a cabo um conjunto de análises parciais dos dados, realizando comparações múltiplas sob as quais a hipótese de rejeição ou não de *Ho* não havia sido formulada. O que distingue, verdadeiramente, a proporção de erro associado a comparações *a priori* com a proporção de erro associado a comparações *a posteriori* é o número de comparações que uma e outra análise suscitam (Ryan, 1959). Análises *post hoc* envolvem a realização de *todas* as comparações possíveis entre as estatísticas (por exemplo: médias) que sumarizam a informação contida nos dados. Na realidade, os testes planeados só estão em vantagem relativamente aos não-planeados quando o seu número é bastante inferior ao número total de compa-

rações que podem ser levadas a cabo sob um determinado delineamento experimental.

Análises *post hoc* são conduzidas, regra geral, após a realização do *omnibus* F teste, ou seja, o teste F global, que envolve todas as estatísticas que posteriormente serão sujeitas a comparações. Caso o teste F não seja significativo, qualquer comparação múltipla «significativa» traduz um erro Tipo I e só ocorrerá se não houve um controlo adequado da inflação do alfa. Assim, sem recurso a procedimentos específicos para testes múltiplos, o risco de cometer erros Tipo I será inferior se as análises *post hoc* estiverem associadas a comparações precedidas de um teste F significativo (por exemplo, Wilcox, 1987). A realização do teste global no caso de comparações planeadas, não faz sentido, a não ser que a hipótese do experimentador seja de tal modo vaga que apenas aponte para a existência de qualquer diferença, sem especificar qual.

1.5. Dependência *vs* independência das comparações.

Uma das preocupações que acompanha qualquer realização de um conjunto de comparações é o grau de relação que estas estabelecem entre si: comparações ortogonais ou independentes são diferenciadas daquelas que fornecem informação redundante ou correlacionada.

Em r condições experimentais (níveis do factor) definem-se r-1 *contrastes ortogonais* (que corresponde ao número de graus de liberdade associados a esse factor). Garante-se a ortogonalidade dos contrastes realizados ao verificar-se que a soma dos produtos dos pesos atribuídos aos i níveis do factor (a_i) for zero. Assim, quaisquer j contrastes (j=1, .., r-1) realizados para X_i definidos pelos pesos a_{ij} são ortogonais se

$$\sum_{i=1}^{r}\prod_{j=1}^{r-1} a_{ij} = 0$$

independentemente do peso específico associado a cada factor que se exemplifica com os seguintes dois conjuntos de contrastes ortogonais realizados para três grupos:

i)	+1	-1/2	-1/2	*ii*)	+1	0	-1
	0	+1	-1		-1/2	+1	-1/2
Πa_{ij} =	0	-1/2	+1/2	Πa_{ij} =	-1/2	0	+1/2
	$\Sigma(\Pi a_{ij}) = 0$				$\Sigma(\Pi a_{ij}) = 0.$		

Cada conjunto de contrastes ortogonais refere questões independentes, não correlacionadas. Trata-se de uma repartição específica do efeito (principal ou interacção) na medida em que a soma dos quadrados associada ao efeito com r-1 graus de liberdade é repartida em r-1 comparações com 1 grau de liberdade, esgotando-se toda a informação independente contida nos dados, ou seja,

$$SS_A = \sum_{j=1}^{r-1} SS_{Lj}$$

em que $$SS_{Lj} = \left(\frac{Lj}{\sqrt{\frac{a_i}{n_i}}}\right)^2 \Leftrightarrow \left(t_{(1)}\sqrt{MSe}\right)^2.$$

Como foi referido, se todas as comparações estiverem perfeitamente correlacionadas, então, ou são significativas ou são não-significativas, pelo que $\alpha EW=\alpha_c$. Quando se compara uma média com muitas outras (médias), esta média surge num número elevado de testes, donde esses testes não são independentes. O número de comparações ortogonais passíveis de serem realizadas num dado delineamento experimental é sempre menor do que o número de comparações possíveis. No entanto, é ele e só ele que determina a máxima inflação possível do erro experimental; todas as outras análises lhe são redundantes. Mas, se o número de comparações ortogonais afecta directamente o erro experimental, não devemos esquecer que a sua dependência afecta o valor esperado do erro por experiência. O número de conclusões erradas, retiradas da análise dos dados, será superior. «Num conjunto de comparações ortogonais é possível ter mais que um erro Tipo I, mas admitamos que isso seria algo raro. Com comparações não-ortogonais, um erro Tipo I ocorrido numa comparação poderá ser replicado a outras comparações do mesmo conjunto. Apesar de este facto conduzir a um maior número de erros por experiência, quando um erro realmente ocorre, acontece que será menor o número de experiências com qualquer erro que seja» (Klockars e Sax, 1986, p. 36).

3. *Procedimentos para a realização de comparações múltiplas.*

Apresentamos de seguida um fluxograma decisional relativo a diferentes técnicas ou procedimentos de comparações múltiplas (tendo por base o fluxograma oferecido in Hopkins e Chadbourn, 1967 e apresentado in Keppel, 1973), que se centra unicamente nas abordagens paramétricas, não multivariadas, que visam a comparação de médias em condições de normalidade e homocedastidade. Assim, não se referem as técnicas desenvolvidas especificamente para a condição de heteroscedastidade, para análises não-paramétricas e para alguns testes específicos (como por exemplo, o teste simultâneo de significância para uma matriz de correlações, apresentado in Steiger, 1980, e também referido in Cohen e Cohen, 1983).

Os testes de comparações de médias baseam-se essencialmente em três tipos de estatísticas: *t* ou *student's t statistic*, *q* ou *studentized range statistic*, e *F* ou *variance ratio statistic*, que se encontram relacionadas. Por agora lembramos apenas que a estatística *t* nos fornece uma medida padronizada da distância entre médias,

diferente da medida q, visto a primeira utilizar como unidade o erro padrão da diferença entre as médias (*i*) e a segunda o erro padrão da média global (*ii*):

$$i)\ t = \frac{\overline{X}_1 - \overline{X}_2}{s_{(\overline{x}_1 - \overline{x}_2)}} \qquad ii)\ q = \frac{\overline{X}_1 - \overline{X}_2}{s(\overline{X})}.$$

A diferença nos denominadores é ditada por uma diferença na natureza da questão a que ambas dizem respeito (Klockars e Sax, 1986): 1) qual a magnitude da diferença entre médias relativamente ao padrão da variação da diferença entre médias; 2) qual a magnitude da diferença entre médias relativamente ao padrão de variação de médias. Se considerarmos que $s_{(\overline{x}_1 - \overline{x}_2)} = \sqrt{s^2(\overline{x}_1) + s^2(\overline{x}_2)}$, então, para o caso de amostras de igual dimensão e variância temos $s_{(\overline{x}_1 - \overline{x}_2)} = \sqrt{2\, s^2(\overline{x})}$. Desta forma, $q = \sqrt{2}\, t$, tendo ainda presente que $F_{[1,\upsilon]} = t^2_{[\upsilon]}$, então também $q_{[\upsilon]} = \sqrt{2F_{[1,\upsilon]}}$ e $F = \frac{q^2}{2}$.

Sempre que o consideremos necessário, faremos referência ao valor crítico do teste subjacente ao procedimento de controlo de inflação do erro Tipo I, que irá ser apresentado, lembrando que ele se associa a uma destas três estatísticas e, logo, a uma das distribuições de probabilidade t, q ou F.

Apresentaremos cada método ou procedimento, de forma a que o Leitor possa fazer os cálculos que cada um envolve, realizando breves ajustamentos do *output* do seu programa de estatística. A maioria dos programas estatísticos permite a realização de diferentes tipos de comparações de estatísticas, incluindo alguns dos procedimentos de controlo de erro Tipo I, como a comparação de todos os pares de médias (*pairwise methods*) em caso de homocesdasticidade: Tukey Honestly Significant Difference Test (HSD), Tukey Wholly Significant Difference Test (WSD), Least Significant Diference Test (LSD)[4], Scheffé, Ryan Q, Duncan Test, Newman-Keuls Test[5], entre outros. Os programas estatísticos podem ser mais ou menos limitados na flexibilidade com que permitem a utilização destes métodos (dificuldades em estender a comparações mais complexas e a comparações previamente delimitadas).

Os procedimentos serão referenciados relativamente a uma família de efeitos, pelo que o controlo do erro *experimentwise* coincidirá com o controlo do erro *familywise*. Uma ilustração numérica destes procedimentos (apresentada no final do artigo) será realizada no contexto de um delineamento experimental factorial, para que a distinção entre estes dois tipos de erro seja melhor entendida.

Vejamos, então, o sistema decisional aqui representado. Em primeiro lu-

gar, há que estabelecer a proporção de erro que se deseja ver associada às conclusões da experiência, ou seja, estabelecer o nível de alfa. Como regra de algibeira guardemos a ideia, explicitada mais adiante, de que é preferível estabelecer um alfa mais elevado do que o habitual 1 ou 5 por cento.

O número de comparações ou contrastes a realizar vai igualmente determinar a inflação de alfa associada a uma ou outra técnica. A caracterização da análise comparativa, como tendo sido estabelecida *a priori* ou *a posteriori*, é importante neste sistema decisional, não só porque vai determinar o número de comparações a realizar, mas também porque vai definir os tipos específicos de contrastes a levar a cabo.

Centremo-nos, em primeiro lugar, nos procedimentos directamente baseados na desigualdade de Bonferroni. Estes procedimentos, mais adequados a situações de comparações planeadas, têm particular importância quando o número de comparações a realizar é reduzido. Uma das suas grandes vantagens é a flexibilidade de aplicação a qualquer comparação: as que envolvem dois grupos (*pairwise comparisons*) as que envolvem mais que dois grupos (comparações complexas) ou ainda no caso destas comparações serem multivariadas.

Método de Dunn-Bonferroni. Dunn (1961) propôs o ajustamento do valor crítico associado a um contraste linear, recorrendo à desigualdade de Bonferroni (Wilcox, 1987). Assim, este procedimento é designado quer pelo nome da autora (Dunn), quer pela referência à desigualdade probabilística que lhe serve de base (Bonferroni), quer ainda por uma combinação das duas designações (Dunn-Bonferroni).

Consiste em testar cada comparação ao nível α/c, tratando-se de uma repartição do nível de significância pelas diferentes comparações que se pretendem realizar. Pela desigualdade de Bonferroni fica garantido que αEW não excederá o valor estipulado de alfa. Não é, contudo, necessário que esta partição seja equitativa, cabendo ao investigador decidir como quer distribuir o nível de, tendo em conta as suas hipóteses e a preocupação com os erros Tipo I e Tipo II.

Visto o procedimento Dunn-Bonferroni se tornar cada vez mais conservador à medida que o número de comparações aumenta, é particularmente aconselhável quando esse número é reduzido, sendo assim mais adequado a comparações planeadas. No caso de três comparações, por exemplo, (c=3) testamos três hipóteses nulas, Ho, $H'o$ e $H''o$ às quais associamos os níveis de significância p, p' e p'', respectivamente. O nível

α pode ser repartido equatitativamente (*i*) ou não (*ii*), da seguinte forma:

i) α=0.05: α_c =0.05/3 $\Rightarrow p > 0.05/3$ ou $p < 0.05/3$

ii) α=0.05: α_1 =0.02 α_2 =0.01 α_3 =0.02.

Foram desenvolvidos outros procedimentos com vista a atingir estatutos menos conservadores (ver Klockars e Hancock, 1992). Estes procedimentos envolvem um ajustamento dos níveis de significância dos diferentes testes realizados por uma sequência específica e, como tal, designam-se por procedimentos sequenciais. Foram desenvolvidos por Holm (1979), Shaffer (1986), Hochberg (1988) e Hommel (1988). Como são procedimentos muito semelhantes, faremos apenas referência àqueles propostos por Holm (1979) e Hochberg (1988).

Método de Holm (Sequentially Rejective Test). Trata-se de um procedimento *step-down*, onde se procura rejeitar sucessivamente as hipóteses organizadas sequencialmente pelo valor de *p* (nível de significância do teste) associado a cada estatística. Após a realização das comparações/contrastes/testes específicos, estes são organizados tendo em conta que p_1 designa o valor menor e p_c o valor mais elevado, isto é, de forma a que $p_1 < p_2 < p_3 < \ldots < p_c$. A estes valores associam-se as hipóteses H_1, H_2, H_3, ..., H_c, a testar ao nível α/c-(k-1), onde k=1, ..., c. Se não se rejeita uma determinada hipótese, então não se rejeitam igualmente as de nível superior.

Método de Hochberg. Trata-se de um procedimento *step-up*, pois procura-se reter as hipóteses H_k, às quais se associam níveis de p mais elevados.

Em 1992 Klockars e Hancock apresentaram no *Psychological Bulletin* os resultados de um estudo de simulação, concluindo que qualquer dos métodos baseados na desigualdade de Bonferroni permitem um controlo aceitável do erro Tipo I, independentemente da magnitude dos efeitos estudados. No entanto, tal como o método Dunn-Bonferroni, à medida que o número de comparações aumenta, estes testes tornam-se cada vez mais conservadores, havendo disparidade na sua potência consoante a magnitude de efeito considerada.

Teste de Dunnett. O teste de Dunnett (1955) destina-se unicamente à comparação de um grupo (regra geral, o grupo controlo) com os restantes grupos envolvidos na análise, o que restringe grandemente o número de comparações a realizar. Trata-se de um teste potente, que exerce um razoável controlo sobre possíveis inflações do alfa, não sen-

do, porém, tão conservador como os testes de Tukey e de Scheffé (ver adiante) quando aplicados à mesma situação (Keppel, 1973). A sua eficácia não parece sobrepor-se à dos métodos directamente baseados na desigualdade de Bonferroni, quando o número de grupos experimentais é reduzido[6].

Relativamente aos testes *post hoc* apenas apresentamos três tipos de procedimentos: 1) Tukey, 2) Scheffé, que são testes simultâneos com flexibilidade de aplicação a diferentes tipos de contrastes, e 3) o teste sequencial Q de Ryan, com aplicação restrita ao caso de comparação de duas médias (*pairwise*). As técnicas sequenciais (*stepwise*) são frequentes neste tipo de procedimentos, visto, regra geral, serem mais sensíveis à detecção de efeitos (logo mais potentes) do que os testes simultâneos. No entanto, apesar do seu conservadorismo, os testes simultâneos garantem-nos um controlo mais eficiente sobre alfa (Klockars e Hancock, 1992).

Note-se que a não recomendação do uso destes procedimentos para o caso de comparações planeadas reside no facto de, nestas condições, ser necessário assumir uma atitude conservadora. Esta atitude é inapropriada quando se prevêm efeitos específicos, e apropriada quando o estudo é exploratório ou as suas hipóteses orientadoras são vagas.

Método de Tukey (HSD, Honestly Significant Diference). Num manuscrito inédito (1953), Tukey propõe (Keppel, 1973) uma forma de lidar com o problema de inflação do alfa aplicado à realização de todas as comparações entre pares de médias, através de um procedimento que ficou conhecido pela designação de *Honestly Significant Difference Test* (HSD). Este procedimento, para além de assumir homocedasticidade, requer ainda que o mesmo número de observações seja associado a cada grupo. Em 1956 Kramer (ver Wilcox, 1987, e Zwick, 1993) propõe ligeiras modificações no procedimento de Tukey, de forma a lidar com um número diferente de observações por célula (o conhecido procedimento Tukey-Kramer). Referir-nos-emos aos dois procedimentos em conjunto, considerando o primeiro como um caso especial do segundo.

Atendamos primeiro ao procedimento de Tukey enquanto procedimento *pairwise*. Nesta qualidade tem a característica particular de garantir que o erro experimental coincide exactamente com o alfa nominal (Wilcox, 1987). Por esta razão, é referido como o teste que maior protecção fornece relativamente à inflação do alfa em comparações *pairwise* (Miller,1981; Wilcox, 1987), mantendo-se ao mesmo tempo sensível à detecção de efeitos. Na reali-

zação de *todas* as comparações *pairwise* de um estudo, o procedimento HSD encontra-se em vantagem quer relativamente aos procedimentos baseados na desigualdade de Bonferroni (devido ao seu elevado número), quer mesmo ao procedimento de Scheffé (descrito adiante), na medida em que HSD é mais sensível a detectar diferenças existentes (mais potente) entre pares de médias. Assim, apesar de lhe ser frequentemente associado um estatuto conservador[7] (Kolckars e Sax, 1986), é preferencial relativamente ao método Scheffé, quando se pretende comparar todos os pares de médias (Day e Quinn, 1989; Wilcox, 1987; Winner, 1971) e relativamente aos procedimentos que se baseiam na desigualdade de Bonferroni, quando o número dessas comparações é elevado (quanto maior o número dessas comparações, maior a sua vantagem).

Tukey concebe intervalos de confiança para a magnitude da diferença entre quaisquer duas médias da mesma família, reportando-se a uma distribuição q (*Stundentized Range*). Assim, deve começar-se por comparar a maior diferença de pares das médias com o valor crítico CRT, o que nos reporta à diferença mínima entre aqueles pares de médias que deve ser excedida para que qualquer diferença seja considerada significativa ao nível de alfa nominal.

O valor crítico CRT vem dado por $CRT = q_{r(gl_e);\alpha}\sqrt{\frac{MSe}{n}}$, em que q representa a estatística tabelada para o nível de significância α, com *gle* graus de liberdade (*gle*, graus de liberdade associados a *MSe*) para r número de grupos em estudo; *MSe* (quadrado médio residual) representa o termo erro do modelo ANOVA associado às observações e n o número de observações por grupo.

Este procedimento está inserido na maioria dos programas estatísticos que indicam o valor crítico CRT ou indicam os valores de p associados à estatística de teste q definida como $q = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{MSe}{n}}}$.

Caso o cálculo seja manual, deve comparar-se esta estatística com o valor tabulado de $q_{r(gle);\alpha}$ que segue uma distribuição *Stundentize Range* para o nível de significância α que designarmos. Caso a máxima amplitude da diferença entre as duas médias se apresente estatisticamente significativa, então tem sentido proceder a outras comparações.

A modificação sugerida por Kramer reside na padronização da diferença entre as médias que origina a estatística de teste

$$q = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{MSe}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Como procedimento generalizado a combinações mais complexas de médias da mesma família, o problema associado ao seu estatuto de teste conservador é agravado com o aumento do número de médias que a análise envolve. Tal facto leva a que se deva recorrer preferencialmente ao procedimento de Scheffé (descrito adiante) em abordagens *post hoc* que não sejam *pairwise* (Day e Quinn, 1989; Wilcox, 1987; Winner, 1971).

Método Q de Ryan. Este teste é uma adaptação, realizada por Einot e Gabriel (1975), do teste proposto por Ryan em 1960, para o caso de múltiplas amostras. Estudos de simulação (ver Day e Quinn, 1989) sugerem que este procedimento é o mais potente dos testes de comparações de médias (*pairwise*), exercendo um controlo sobre a inflação do erro Tipo I, semelhante à abordagem de Tukey. Os autores da simulação, Day e Quinn (1989), recomendam o seu uso quando se pretende realizar um grande número de comparações entre pares de médias.

O problema associado a este procedimento relaciona-se, no entanto, com uma vertente prática. Não só não se encontra inserida nos habituais programas estatísticos, como envolve a recorrência a complexas interpolações das tabelas Q. Para o leitor interessado recomendamos o artigo de Day e Quinn (1989), bem como o de Einot e Gabriel (1975).

Método Scheffé. Desenvolvido por Scheffé em 1953, representa o método generalizado a todos os contrastes, mais conhecido e utilizado. Trata-se de um procedimento muito conservador, visto estar concebido para oferecer protecção a um grande número de inferências estatísticas. Por esta razão é recomendado quando o investigador pretende realizar um grande número de comparações sobre o mesmo conjunto de dados.

A estatística que serve de base a este procedimento é uma estatística *F*, pelo que mais próxima da estatística associada ao *omnibus*, teste que se recomenda ser realizado previamente[8]. Caso alguma das relações postulada pelo modelo ANOVA se apresente significativa, então a análise deve prosseguir, com vista a «explicar» o efeito encontrado. Regra geral, explora-se o conjunto de contrastes ortogonais, sendo também frequente a análise de apenas pares de médias[9]. Esta atitude exploratória, sendo a principal fonte de perigo para inflações do alfa, é totalmente segura quando protegida pelo método Scheffé (Miller, 1981).

O procedimento de Scheffé (1953), tal como o de Tukey, assenta na concepção de um novo valor crítico para determinar a amplitude crítica

do intervalo de confiança do efeito observado, relativamente a uma família de efeitos (*Crs*),

$CRs = \sqrt{(r-1)F_{(glA,gle)};\ \alpha_{FW}}\ \sqrt{MSe\sum_i \frac{a_i^2}{n_i}}$, em que r representa o número de grupos definidos pela variável em estudo; F é o valor obtido da tabela F-Snedecor ao nível αFW com glA graus de liberdade do numerador e gle graus de liberdade do denominador (graus de liberdade associados ao efeito residual ou erro); a_i os pesos atribuídos ao grupo i no contraste; MSe o quadrado médio residual do modelo; e n_i o número de observações no grupo i.

Em termos de teste (embora o intervalo de confiança seja muitas vezes necessário, visto que nos pode interessar a magnitude do efeito) temos então que um contraste definido numa família de efeitos, quando elevado ao quadrado, em vez de ser comparado com o valor crítico F, da tabela F-Snedecor, deve ser comparado com o valor que designaremos de Fs, dado por $Fs=(r-1)\ F_{(glA,gle);\ \alpha FW}$.

Note-se que a probabilidade de se verificar um erro Tipo I fica controlada a um nível *familywise* definido por αFW e, só no caso de o modelo ANOVA, a um factor (*one-way*) é que este nível coincidirá com o nível experimental.

4. *Exemplificação numérica.*

Como exemplo, considerem-se os dados relativos ao grau de adaptação ao curso em que se licenciam 120 sujeitos, distribuídos equatitativamente por quatro cursos com frequência de anos diferentes (1º ano, 3º ano e finalistas).

4.1. Análises «a priori».

Imaginemos que este estudo fora guiado por uma teoria que define claramente duas hipóteses específicas: 1) o grau de adaptação ao curso aumenta com os anos já realizados; e 2) os sujeitos encontram-se melhor adaptados nos cursos de ciências humanas do que nos cursos de ciências físicas. Cada uma destas hipóteses é traduzida estatisticamente por um contraste específico, dentro de uma família de efeitos diferente. O número de comparações é reduzido (c=2), como é natural no caso de comparações planeadas (*a priori*), pelo que o procedimento de Dunn-Bonferroni se apresenta como o mais adequado. Lembramos que pressupomos a independência das observações e a normalidade e homocedasticidade das populações de onde foram retiradas as respectivas amostras. Note-se que neste caso existe igual número de observações por célula (5), embora tal não seja

Tabela 1.
Médias observadas.

Tipo de curso	*Finalistas*	*3º ano*	*1º ano*	*Total*
Psicologia	36.0	17.6	11.6	21.7
Sociologia	24.0	37.8	8.6	23.5
Matemática	11.8	4.4	17.6	11.3
Física	9.6	15.8	0.6	8.7
Média	20.4	18.9	9.6	

pressuposto dos procedimentos apresentados.
Pretendendo manter $\alpha EW = 0.05$, utilizaremos $\alpha_c = 0.05/2 = 0.025$, que corresponde ao valor crítico unilateral: $t_{(48);0.025} = 2.01$.

Hipótese 1 *Ho*: L=0 *vs H1*: L > 0

Contraste[10]

$L = (-1)9.6 + (0)18.9 + (1)20.35 =$
$= 20.35 - 9.6 = 10.75$

Erro Padrão

$se_L = \{[(-1)^2+(0)^2+(1)^2]MSe/20\}^{1/2} =$
$= [2(69.52/20)]^{1/2} = (6.95)^{1/2} = 2.64$

Estatística de teste

$t = 10.75/2.64 = 4.08$, pelo que se rejeita *Ho*.

Hipótese 2 *Ho*: L=0 *vs H1*: L > 0

Contraste

$L = (1/2)\ 21.74 + (1/2)\ 23.47 + (-1/2)$
$11.27 + (-1/2)\ 8.67 = 22.61 - 9.97 =$
$= 12.64$

Erro Padrão

$se_L = \{[(0.5)^2+(0.5)^2+(-0.5)^2+(-0.5)^2]$
$MSe/15\}^{1/2} = [1\ (69.517/15)]^{1/2} =$
$= (4.63)^{1/2} = 2.15$

Estatística de teste

$t = 12.64/2.15 = 5.87$, pelo que se rejeita *Ho*.

As hipóteses nulas foram rejeitadas ao nível de significância individual de 2.5 por cento. Assim, temos 95 por cento de confiança ao concluir pela presença dos *dois* efeitos postulados pelas hipóteses em estudo, na medida em que o nosso alfa por experiência é igual ou inferior ao alfa nominal estipulado de início em 5 por cento.
Caso diferente é aquele em que, olhando para o conjunto das médias, queremos estudar os efeitos significativos e suas características específicas. A análise global (*omnibus F test*; ver tabela 2, p. 214) é seguida de análises *post hoc* das quais ilustraremos algumas. Lembramos, porém, o Leitor de que se após a realização da ANOVA pretender realizar um número reduzido de contrastes, poderá ser preferível recorrer a um dos métodos baseados na desigualdade de Bonferroni.

O modelo ANOVA deste plano ou delineamento experimental envolve o estudo de três efeitos, permitindo que sejam realizados em simultâneo três testes independentes sobre os mesmos dados. Se, por alguma razão, queremos para a análise global dos três efeitos manter α=0.05, isto é, se quisermos controlar αEW= 0.05 e não meramente um alfa *familywise* controlado, é razoável recorrer ao procedimento Bonferroni e utilizar αc= 0.05/3 = 0.0167. Ao padrão global dos dados associamos uma probabilidade de erro de 5 por cento (αEW= 0.05) e a cada família de efeitos postulados pelo modelo ANOVA associamos uma probabilidade de erro Tipo I de 1,67 por cento que arredondaremos para 1 por cento (αFW=*0.01*) numa atitude ainda mais conservadora. Da tabela 2 inferimos a presença de três efeitos. Análises *post hoc* ajudam-nos a ilucidar a natureza desses efeitos. Assim, poderemos proceder a comparações *pairwise* no seio de cada família, recorrendo ao método de Tukey (o mais aconselhável para estas análises). Se não nos limitarmos a comparações de pares de médias, estendendo a análise a combinações mais complexas (por exemplo, estudar que efeito específico de interacção é significativo), então, para cada família de efeitos, calcularemos o valor crítico pelo procedimento de Scheffé.

4.2. Análises «post hoc».

Família de efeitos associados à variável curso. Tukey (HSD): todas as comparações *pairwise*. – Comparação de r = 4 médias, αFW = 0.01, $q_{4(48);0.01}$ = = 4.55 (por interpolação) donde $CRT = q_{4(48);0.01}\ [(69.52/15)]^{1/2} = 4.55^* 2.15 = 9.78$.

Concluindo-se, desta forma, que os cursos de Ciências Sociais não diferem entre si (diferença das médias: d = 1.8), bem como os de ciências físicas (d = 2.6), embora entre estes dois tipos de cursos todas as diferenças excedam o valor 9.8.

Scheffé: todas as comparações de médias. – Qualquer contraste deve

Tabela 2.
ANOVA: resultados.

Fonte de variação	*SS*	*gl*	*MS*	*F*	*p*
Curso	2467.250	3	822.417	11.830	0.000
Ano	1361.033	2	680.516	9.789	0.000
Curso-Ano	3411.160	6	568.517	8.178	0.000
Residual	3336.800	48	69.517		

Tabela 3.
Variável curso: valores de p para as comparações *pairwise* obtidas com o método de Tukey.

	Psicologia	*Sociologia*	*Matemática*	*Física*
Psicologia	1.000			
Sociologia	0.941	1.000		
Matemática	0.007	0.001	1.000	
Física	0.001	0.000	0.828	1.000

Tabela 4.
Variável curso: valores de p para as comparações *pairwise* obtidas com o método de Scheffé.

	Psicologia	*Sociologia*	*Matemática*	*Física*
Psicologia	1.000			
Sociologia	0.955	1.000		
Matemática	0.014	0.003	1.000	
Física	0.001	0.000	0.866	1.000

ser comparado com a estatística $Fs = [(4\text{-}1)\, F_{(3,48);0.01}]^{1/2} = 3.56$.
Para as comparações *pairwise* seria equivalente a definir-se a diferença máxima não significativa, como sendo $CRs = Fs\ [2(69.52)/15]^{1/2} = 10.83$, o que facilmente ilustra o maior conservadorismo do teste Scheffé quando aplicado à comparação de pares de médias. Com este teste não se poderia afirmar a existência de diferentes graus de adaptação nos cursos de psicologia e de matemática (onde $d = 10.5$).
Apresenta-se nas tabelas 3 e 4 o *output* do resultado destes contrastes, realizados numa programa estatístico para os procedimentos de Tukey e de Scheffé respectivamente, e que indica os níveis de p associados a cada contraste, como havíamos referido. Estes níveis de p devem ser comparados com o nosso $\alpha FW = 0.01$.

Família de efeitos associados à variável ano. Seriam realizados calculos idênticos aos efectuados para os efeitos da variável curso. Lembramos apenas a diferença existente nos graus de liberdade associados à variável ano (visto ter apenas três grupos) e à dimensão de cada amostra (que, em vez de 15, se refere a 20 sujeitos). O nível crítico de Tukey (tendo $q_{3(48);0.01} = 4.33$) seria de *8.06* e o nível crítico de comparação de duas médias para Scheffé seria de *8.42*. Qualquer comparação complexa seria comparada com a estatística $Fs = 3.19$.

*Família de efeitos associados ao efeito ano*curso.* Explorar o padrão específico do(s) efeito(s) de interacção que é significativo não só envolve um grande número de comparações como, neste caso, envolve comparações de combinações complexas de médias (por exemplo, e considerando as letras como representando as células resultantes de se cruzarem os dois factores: L=1/6(a+b+d+e+i+m)- - 1/6(c+f+g+h+j+l)), pelo que o procedimento mais adequado é, sem dúvida, o de Scheffé. O número de graus de liberdade associados à interacção é dado por (4-1)(3-1) = = 6, pelo que para αFW=1%, Fs = [6 $F_{(6,48);0.01}$] = $(6*3.2)^{1/2}$ = 4.4. No caso de se pretender realizar contrastes específicos entre pares de células, o método Scheffé considerará significativa qualquer diferença que exceda o valor crítico, CRs=4.4 [2(69.52)/ /5]$^{1/2}$ = 4.4*5.27 = *23.2*.

Note-se que, na comparação entre células, os valores usados são médias parcelares e não as médias dos diferentes níveis de cada efeito. Como tal, a utilização do método HSD é polémica. Na literatura (ver Winner, 1971, e Wilcox, 1987) podemos encontrar duas sugestões para o recurso a qualquer comparação de células via HSD: *a*) considerar o número total de células como r = número de médias a comparar, pelo que $CRT = q_{12(48);\alpha FW}$ $(69.517/5)^{1/2}$ = = 20.5. Qualquer comparação entre duas médias (no total 66 comparações possíveis) será permitida sob esta abordagem (Winner,1971); *b*) ter em conta a natureza bivariada da distribuição, recorrendo à tabela de Multivariate Range Distribution (Wilcox, 1987), e calcular dois limites para diferenças não significativas,

CRT = 4.39 (69.52/15)1/2 = *16.36* para as 18 comparações das 4 linhas, ou

CRT = 4.55 (69.52/15)1/2 = *15.59* para as 12 comparações das 3 colunas.

Neste caso apenas se contemplam algumas das comparações possíveis entre todas as células. Qualquer das duas abordagens, sendo *pairwise*, é, como seria de esperar, menos conservadora do que a abordagem de Scheffé.

5. *Conclusão.*

A validade das interpretações dos resultados de investigações é frequentemente posta em causa devido ao número de análises comparativas que a elas estão associadas. Diversos procedimentos garantem a inflação da probabilidade de rejeitarmos a hipótese nula quando esta é falsa. Cabe ao investigador definir o procedimento mais adequado ao seu caso específico, tendo em conta a

natureza e o número das suas hipóteses.

O controlo do erro Tipo I promove, por sua vez, inflações do erro Tipo II, pelo que cabe igualmente ao investigador pensar os seus resultados nesta dupla perspectiva.

Qualquer dos procedimentos aqui apresentados são vantajosos relativamente a uma situação específica. Lembramos que testes como o de Newman-Keuls, o LSD e o Ducan não foram referidos, por serem para a grande maioria dos casos ineficientes e em nenhum caso superiores a pelo menos um dos procedimentos referidos. Os diferentes procedimentos descritos requerem a verificação dos pressupostos de independência das observações, normalidade e homocedasticidade[11] das populações. Se estes pressupostos são violados, tanto as conclusões dos diferentes procedimentos como as do modelo ANOVA podem ser invalidadas. Estudos realizados sobre as consequências da violação destes pressupostos demonstram, no entanto, que, tal como para o teste F-global (teste do modelo ANOVA), os desvios de normalidade têm tanto maior importância quanto maior for a dimensão da amostra, e que os problemas de heterocedasticidade (super-agravados no caso de não-normalidade) são tanto mais graves quanto maior a disparidade do número de observações em cada célula.

[1] Os extremos deste intervalo definem-se pela conhecida desigualdade de Boole (Holm, 1978), cuja advertência da sua pertinência para este campo se deve a Bonferroni. Deste modo, podemos encontrar na literatura a desigualdade referida a qualquer dos dois autores.

[2] Utiliza-se frequentemente a designação *variável dependente* para nos referirmos a uma variável quando esta é observada nas livres concretizações. No entanto, lembramos o Leitor de que, quando procedemos à simples observação de uma variável, esta é uma *variável de medida* e este tipo de variável é apenas designado de dependente se no estudo se manipularem uma ou mais variáveis (ditas independentes). Visto a variação da variável de medida poder estar dependente desta manipulação, esta designa-se, neste e só neste caso, por *variável dependente*.

[3] A preocupação do investigador deve centrar-se no erro experimental se este estiver mais interessado no padrão global dos seus resultados do que em subpadrões específicos. Se, pelo contrário, as hipóteses relativas a cada família de efeitos, passíveis de serem identificadas no plano de estudo, forem vistas pelo investigador como independentes, então este deve centrar a sua preocupação apenas sobre o erro familiar. Salientamos que alguns autores referem-se apenas à necessidade de controlo do *familywise error rate*, sendo este identificado com o erro experimental (*experimentwise*). Esta identificação só tem razão de ser em estudos não factoriais (visto que dois factores definem três famílias diferentes de efeitos) e a decisão de controlo de um tipo de erro ou de outro,

a nosso ver, deve estar dependente dos objectivos do investigador e de cada caso particular

[4] O teste LSD, também designado de Teste Protegido de Fischer, é um teste sequencial que tende a ser apontado como dos mais potentes (Cohen e Cohen, 1975). No entanto, como Zwick (1993) refere, a sua potência aparente resulta de um pobre controlo sobre o erro Tipo I. Quando o número de grupos a submeter a comparação é igual ou superior a 3, o erro experimental não é inflacionado apenas no caso de as médias serem realmente idênticas (isto é, se a hipótese nula for verdadeira). Se a hipótese nula for apenas parcialmente verdadeira (algumas médias diferirem) o procedimento LSD não mantém fixo o valor do erro experimental.

[5] Simulações (Day e Quinn, 1989; ver também Wilcox, 1987, e Zwick, 1993) realizadas com estes testes sugerem a ineficácia do método Newman-Keuls e Duncan (testes sequenciais) na obtenção de um verdadeiro controlo para inflações do erro Tipo I, pelo que o seu uso é consensualmente desaconselhado. Zwick (1993) chama a atenção de que muitos estudos de simulação oferecem resultados deturpados pelo facto de, por exemplo, compararem procedimentos que mantêm o erro por comparação em 5 por cento com aqueles que mantém em 5 por cento o erro experimental (ver também referência de Einot e Gabriel, 1975, e Zwick e Maracuilo, 1984). Desta forma alertamos o Leitor a estar atento, não apenas às conclusões dos estudos de simulação, mas igualmente ao processo pelo qual este foi realizado.

[6] Este procedimento é integrado na maioria dos programas estatísticos, pelo que o seu uso não levanta grandes dificuldades. Caso contrário recomendamos o uso do procedimento Dunn-Bonferroni.

[7] O seu estatuto é conservador relativamente a outras abordagens que se baseiam igualmente na estatística q, mas em termos sequenciais (*stepwise*), como é o caso dos procedimentos de Newman-Keuls e de Duncan (que, como já referimos, são ineficientes em termos de controlo de inflação do alfa). Isto porque o procedimento HSD de Tukey, estipulando como valor crítico para todas as comparações aquele que se baseia no máximo *range*, não atende ao número de «passos» em que as médias se encontram afastadas.

[8] A recomendação advém do facto de o teste global testar em simultâneo todos os contrastes possíveis. Assim, caso este não seja significativo, torna-se inútil qualquer avanço em termos de análises parciais. Como já referimos em nota anterior este procedimento não é afectado em termos de inflação do erro Tipo I, por não se realizar o *omnibus* teste, a sua recomendação advém de razões meramente práticas.

[9] Note-se que o teste F se associa à existência de pelo menos um contraste significativo, embora este não seja obrigatoriamente o que se envolve com comparações de pares de médias.

[10] Este tipo de contraste, onde os diferentes níveis do factor têm subjacente uma variável contínua (exemplo: tempo, idade, ...) e em que os pesos que definem os contrastes se distanciam igualmente, define uma *análise de tendências* (*trend analysis*). Neste caso específico, o contraste define uma tendência linear, cujo residual (tendência

quadrática) estimado é $t = 1.72$ ($SS = 205.41$), que não é significativo.

[11] Veja-se, por exemplo, Zwick (1993) como referência a procedimentos de comparações múltiplas em casos de não-normalidade das populações e/ou desigualdade de variâncias. A mesma fonte apresenta igualmente procedimentos adequados a comparações múltiplas para o caso de não independência das observações, isto é, para o caso de medidas repetidas ou amostras emparelhadas.

Referências

Cohen, J., e Cohen, P. (1983), *Applied multiple regression/correlation analysis for the behavioral sciences* (2ª ed.), Londres, Lawrence Erlbaum.

Cohen, J. (1988), *Statistical power analysis for the behavioral sciences* (2ª ed.), N. J.: Lawrence Erlbaum Associates.

Cox, D. R. (1965), «A remark on multiple comparasion methods», in *Technometrics*, 7 (2), pp. 223-24.

Day, R. W., e Quinn, G. P. (1989), «Comparisons of treatments after an analysis of variance in ecology», in *Ecological Monographs*, 59 (4), pp. 433-63.

Dunn, O. J. (1961), «Multiple comparasions among means», *Journal of the American Statistical Association*, 56, pp. 52-64.

Dunnett, C. W. (1955), «A multiple comparison procedure for comparing several treatments with a control», in *Journal of the American Statistical Association*, 50, pp. 1096-121.

Einot, I., e Gabriel, K. R. (1975), «A study of the powers of several methods of multiple comparisons», in *Journal of the American Statistical Association*, 70, pp. 574-83.

Hochberg, Y. (1988), «A sharper procedure for multiple tests of significance», in *Biometrika*, 75, pp. 800-2.

Hochberg, Y., e Tamhane, A. C. (1987), *Multiple comparison procedures*, Nova Iorque, Wiley.

Holm, S. (1979), «A simple sequential rejective multiple test procedure», in *Scandinavian Journal of Statistics*, 6, pp. 65-70.

Hommel, G. (1988), «A stagewise rejective multiple test procedure based on a modified Bonferroni test», in *Biometrika*, 75, pp. 383-86.

Keppel, G. (1973), *Design and analysis. A research handbook*, N.J., Prentice-Hall.

Klockars, A. J., e Hancock, G. R. (1992), «Power of recent multiple comparison procedures as apllied to a complet set of a planned orthogonal contrasts», in *Psychological Bulletin*, 3, pp. 505-10.

Klockars, A. J., e Sax, G. (1986), *Multiple comparisons*, Series Quantitative Applications in the Social Sciences, Sage University Paper.

Lawrence, B. M. (1990), *Understanding significance testing*, SAGE Publications Inc.

Miller, R. G. Jr (1981), *Simultaneous statistical inference* (2ª ed.), Springer Series in Statistics, Nova Iorque, Springer Verlag.

Murteira, B. J. F. (1990), *Probabilidades e estatística*, vol. I (2ª ed.). McGaw-Hill, Lisboa.

Ramsey, P. H. (1981), «Power of univariate pairwise multiple comparison procedures», in *Psychological Bulletin*, 90, pp. 352-66.

Ryan, T. A. (1959), «Multiple comparisons in psychological research», in *Psychological Bulletin*, 56, pp. 26-47.

Scheffé, H. (1953), «A method for judging all contrasts in the analysis of variance», in *Biometrika*, 40, pp. 87--104.

— (1959), *Analysis of Variance*, Nova Iorque, Wiley.

Shaffer, J. P. (1986), «Modified sequentially rejective multiple test procedures», in *Journal of the American Statistical Association*, 81, pp. 826-31.

Steiger, J. H. (1980), «Tests for comparing elements of a correlation matrix», in *Psychological Bulletin*, 87 (2), pp. 245-51.

Tukey, J. W. (1953), «The problem of multiple comparisons. Unpublished paper», in *Princeton University*, Princeton, N.J.

Wilcox, R. R. (1987), «New design in analysis of variance», in *Annual Review of Psychology*, 38, pp. 29-60.

Winner, B. J. (1971), *Statistical principles in experimental design* (2ª ed.), Tóquio, McGraw-Hill Inc.

Zwick, R. (1993), «Pairwise comparison procedures for one-way analysis of variance», in G. Keren e G. Lewis (ed.), *A Bandbook for data analysis in the Behavioral Sciences. Statistical Issues*, Londres, Lawrence Erlbaum.

Abstract. – Any experiment including more than two conditions raises the problem of statistically comparing conditions one with another. This problem of multiple comparisons is related with the inflation of Type I error probability. This paper delineates this problem and presents statistical procedures specifically designed to deal with it in different situations.