Comparação de Grupos "versus" Comparações Emparelhadas

J. M. POMPEU MEMORIA (*)

Um dos casos mais simples e comuns em experimentação é aquele em que se deseja saber qual o melhor de dois tratamentos. O experimentador dispõe de um conjunto de observações $x_1, x_2, \ldots, x_n$ referentes a um tratamento A e $x_1', x_2', \ldots, x_n'$, referentes a um outro tratamento B. Sob o ponto de vista estatístico trata-se da verificação da significância da diferença de duas médias, $\bar{x}$ referente a A e $\bar{x}'$ referente a B. Nestas situações surge a relevante questão: Quando as duas séries de observações devem ser tratadas como dois grupos independentes e em que casos os dados podem ser emparelhados?

Snedecor (5) considera esta decisão de suprema importância, podendo mesmo determinar o sucesso ou insucesso do experimento e embora teça considerações sôbre o assunto, declara não haver fórmula universal para decidir qual das duas alternativas é a mais conveniente. O autor do presente trabalho tentará provar, por considerações teóricas, quando há vantagem de um método em relação ao outro e estabelecer uma regra geral. Evidentemente isto não eliminará em nenhuma circunstância o julgamento do experimentador fundado na natureza do material experimental e no modo em que o experimento foi conduzido; servirá, entretanto, como um critério auxiliar que pretende pôr a questão em termos quantitativos.

Em primeiro lugar serão considerados os dois processos isoladamente, depois um comparado com o outro e, finalmente, a aplicação a um exemplo prático.

(*) Eng. Agr., M. Sc., Prof. adjunto de Estatística Experimental da E.S.A. da U.R.E.M.G.

Método I — Comparação de Grupos

Neste caso, os dados x_i ($i = 1, 2, \ldots, n$) em relação ao tratamento A e os dados x'_j ($j = 1, 2, \ldots, n'$) em relação ao tratamento B, são considerados como duas amostras independentes tiradas da mesma população. O experimentador estabelece a hipótese de que não há diferença entre as duas médias $\bar{x}$ e $\bar{x}'$ (hipótese de nulidade) e na base dos resultados obtidos aceitará ou rejeitará esta hipótese. Nos resultados experimentais nunca se encontra exatamente $|\bar{x} - \bar{x}'| = 0$, porém valores que flutuarão em torno de zero, devido aos erros de amostragem. Seja a o valor desta diferença, far-se-á então um teste estatístico para se determinar qual é a probabilidade de se obter $|\bar{x} - \bar{x}'| \geq a$, na suposição de que o valor real desta diferença é nulo. Se a probabilidade encontrada for pequena, o experimentador é tentado a rejeitar a hipótese de nulidade e em caso contrário a aceitá-la. O êrro de se rejeitar esta hipótese, se verdadeira, pode ser feito tão pequeno quanto se queira, a critério do experimentador, sendo consagrados pela prática os limites de 5% e 1%. Não é, entretanto, conveniente diminuir demasiadamente êste êrro, pois, aumenta-se o êrro de aceitar uma hipótese falsa, i.e., de que não há diferença quando na verdade ela existe.

A prova estatística apropriada é o teste t, que obedece à distribuição do mesmo nome, descoberta por Student, sendo $t = \dfrac{|\bar{x} - \bar{x}|}{s(\bar{x} - \bar{x}')}$ (1) com $n + n' - 2$ graus de liberdade.

A validez dêste teste é assegurada para o caso dos itens distribuirem-se normalmente, mas é sabido que mesmo se esta distribuição não for normal, a distribuição das médias tende para a normalidade à medida que aumenta o tamanho da amostra. Isto nem sempre acontece em amostras pequenas, entretanto, a experiência tem mostrado que o teste não é muito sensível a moderadas discrepâncias da normalidade. A condição de normalidade é, apesar disso, sempre pressuposta. No caso em aprêço, as duas séries de da-

dos foram consideradas como duas amostras independentes tiradas da mesma população normal.

Para calcular t em (1) precisa se determinar s $(\bar{x} - \bar{x}')$, o desvio padrão da diferença das médias. A variância na diferença das médias $s^2(\bar{x} - \bar{x}')$ é igual à soma das variâncias das médias, porque as duas amostras foram tiradas independentemente uma da outra, isto foi garantido pelo modo como elas foram sorteadas. Tem-se, então :

$$s^2(\bar{x} - \bar{x}') = \frac{s^2}{n} + \frac{s^2}{n'} \quad (2)$$, onde s^2, a variância dos itens, é estimada por

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{j=1}^{n'}(x'_j - \bar{x}')^2}{n + n' - 2} \quad (3)$$

Na fórmula (3) utilizou-se uma estimativa combinada da variância, onde a soma dos quadrados dos desvios foram obtidas separadamente em cada uma das amostras. E' preciso não admitir, como bem salienta Fisher (2) que o método acarreta a suposição de igualdade das duas variâncias nos dois tratamentos. Esta suposição é inerente à hipótese verificada, i.e., de que as duas amostras foram tiradas da mesma população normal. Depois das substituições a fórmula (1) reduz-se a

$$t = \frac{(\bar{x} - \bar{x}')\sqrt{n + n' - 2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{j=1}^{n'}(x'_j - \bar{x}')^2}} \cdot \frac{\sqrt{nn'}}{\sqrt{n + n'}} \quad (4)$$

No caso particular de n = n' as fórmulas (2), (3) e (4) transformam-se, respectivamente, em (5), (6) e (7) :

$$s^2_{(\bar{x} - \bar{x}')} = \frac{2\,s^2}{n} \quad (5), \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 - \sum_{j=1}^{n}(x'_j - \bar{x}')^2}{2\,(n-1)} \quad (6)$$

$$t = \frac{(\bar{x} - \bar{x}')\sqrt{n(n-1)}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{j=1}^{n}(x'_j - \bar{x}')^2}} \quad (7)$$

. Êste método equivale à análise de variância com

um único critério de classificação, onde a variação dentro dos tratamentos constitui o êrro com $2(n-1)$ graus de liberdade.

Método II — Comparações Emparelhadas

Suponha-se que no experimento em consideração, os dados possam ser tomados aos pares: $x_1\ x_1'$, $x_2\ x_2'$ ..., $x_n\ x_n'$. Neste caso $n = n'$ forçosamente, cada valor x_i está associado a um valor x'_j $(i, j = 1, 2, \ldots, n)$. Os pares assim formados devem corresponder a um significado físico, podendo ser canteiros próximos, indivíduos da mesma ninhada, gêmeos idênticos, o próprio indivíduo submetido a dois tratamentos diferentes, etc. Pelo método I, a formação de pares é imaterial.

Aqui, como no caso anterior, a hipótese de nulidade é que $|\bar{x} - \bar{x}'| = 0$. Acha-se para cada par um conjunto de diferenças $x_i - x'_j$ $(i, j = 1, 2, \ldots, n)$ iguais, respectivamente, a $d_1, d_2, \ldots, d_n$ cuja média $\bar{d} = 0$ pela hipótese de nulidade, pois que $\bar{d} = \bar{x} - \bar{x}'$, conforme a prova dada abaixo:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i,j=1}^{n}(x_i - x'_j) = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{j=1}^{n} x'_j =$$

$= \bar{x} - \bar{x}'$. Donde se tem $t = \dfrac{\bar{d}}{s_{\bar{d}}}$ que obedece à distribuição de Student com $n-1$ graus de liberdade, pois só se dispõe de n diferenças. A variância da média das diferenças $s^2_{\bar{d}}$ é igual a $s^2_{\bar{d}} = \dfrac{s^2_d}{n}$ e $s^2_d = \dfrac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}$, donde t pode ser achado pela fórmula (8)

$$t = \frac{\bar{d}\sqrt{n(n-1)}}{\sqrt{\sum_{i=1}^{n}(d_i - \bar{d})^2}} \qquad (8)$$

Êste método equivale à análise de variância com dois critérios de classificação, um devido aos tratamentos e o outro causado pela formação de pares. Pelo método I a varia-

ção entre os pares ficou incluida na variação dentro dos tratamentos. No método II, a variação entre os pares foi eliminada do êrro. À esta fonte de variação corresponde n — 1 graus de liberdade, de modo que o êrro ficou com os n — 1 graus de liberdade restantes.

Comparação dos dois métodos

Na base dos resultados obtidos vê-se fàcilmente que as fórmulas (7) e (8) de t, diferem apenas quanto ao desvio padrão, pois que a média das diferenças $\bar{d}$ é igual à diferênça das médias $\bar{x} - \bar{x}'$, conforme já foi provado.

No método I a variância $s^2_{(\bar{x}-\bar{x}')}$ pode ser expressa conforme o desenvolvimento abaixo:

$$s^2_{(\bar{x}-\bar{x}')} = \frac{2\,s^2}{n} = \frac{2}{n} \cdot \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 - \sum_{j=1}^{n}(x'_j-\bar{x}')^2}{2\,(n-1)} =$$

$$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n} x_i^2 + \sum_{j=1}^{n} x'^2_j - \frac{1}{n}\left(\sum_{i=1}^{n} x\right)^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x'^2\right)\right] \quad (9)$$

No método II tem-se $s^2_{\bar{d}} = \frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{n(n-1)} = \frac{\sum_{i=1}^{n} d_i^2 - \frac{\left(\sum_{i=1}^{n} d\right)^2}{n}}{n\,(n-1)} =$

$$(*) = \frac{1}{n\,(n-1)}\left\{\underset{i,j=1}{\overset{n}{S}}(x_i - x'_j)^2 - \frac{1}{n}\left[\underset{j=1}{\overset{n}{S}}(x_i - x'_j)\right]^2\right\} =$$

$$\frac{1}{n\,(n-1)}\left\{\underset{i=1}{\overset{n}{S}} x_i^2 + \underset{j=1}{\overset{n}{S}} x'^2_j - \frac{1}{n}\left(\underset{i=1}{\overset{n}{S}} x\right)^2 - \frac{1}{n}\left(\underset{j=1}{\overset{n}{S}} x'\right)^2 - \right.$$

$$\left. - 2\left[\underset{i,j=1}{\overset{n}{S}} x_i x'_j - \frac{1}{n}\left(\underset{i=1}{\overset{n}{S}} x\right)\left(\underset{j=1}{\overset{n}{S}} x'\right)\right]\right\} = \frac{1}{n\,(n-1)}$$

$$\left[\underset{i=1}{\overset{n}{S}} x_i^2 + \underset{j=1}{\overset{n}{S}} x'^2 - \frac{1}{n}\left(\underset{i=1}{\overset{n}{S}} x\right)^2 - \frac{1}{n}\left(\underset{j=1}{\overset{n}{S}} x'\right)^2\right] -$$

(*) Por falta do sinal Σ foi usado "S" em sua substituição.

$$-\frac{2}{n\,(n-1)}\sum_{i,j=1}^{n}(x_i-\bar{x})\,(x_j'-\bar{x}')\quad(10)\text{, pois }\frac{\sum_{i=1}^{n}x_i}{n}=\bar{x}$$

e $\dfrac{\sum_{j=1}^{n}x'}{n}=\bar{x}'$.

Comparando-se as fórmulas (9) e (10) vê-se que elas diferem apenas pela expressão $\dfrac{2}{n\,(n-1)}\sum_{i,j=1}^{n}(x_i-\bar{x}')\,(x_j-\bar{x}')$. Como o coeficiente de correlação r entre x e x' é dado pela expressão

$$r=\frac{\sum_{i,j=1}^{n}(x_i-\bar{x})\,(x_j'-\bar{x}')}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{j=1}^{n}(x_j'-\bar{x}')^2}}\quad(11)$$

A fórmula (10) pode ser escrita : $s^2_d=s^2_{(\bar{x}-\bar{x}')}-\dfrac{2}{n}$

$$\frac{\sum_{i,j=1}^{n}(x_i-\bar{x})\,(x_j'-\bar{x}')}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{j=1}^{n}(x_j'-\bar{x}')^2}}\quad\frac{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{j=1}^{n}(x_j'-\bar{x}')^2}}{\sqrt{(n-1)\,(n-1)}}\text{ ou}$$

ainda $s^2_d=s^2_{(\bar{x}-\bar{x}')}-\dfrac{2}{n}\,r\,s_x\,s_{x'}\quad(12)$

Os desvios padrões s_x e $s_{x'}$ são iguais pela hipótese de nulidade, pois que foi suposto serem as amostras tiradas da mesma população, faz-se $s_x=s_{x'}=s$ e a fórmula (12) reduz-se a: $s^2_d=s^2_{(\bar{x}-\bar{x}')}-\dfrac{2}{n}\,r\,s^2=\dfrac{2\,s^2}{n}\,(1-r)\quad(13)$

Pela fórmula (13) pode-se bem apreciar a diferença entre as variâncias $s^2_{(\bar{x}-\bar{x}')}$ calculada pelo método I e s^2_d pelo método II.

A mesma fórmula (13) pode ser obtida se as amostras não fossem emparelhadas, porém, consideradas como não independentes. Neste caso a variância da diferença de suas médias $s^2_{(\bar{x} - \bar{x}')}$ seria dado por:

$s^2_{(\bar{x} - \bar{x}')} = s^2_{\bar{x}} + s^2_{\bar{x}'} - 2\, r\, s_{\bar{x}}\, s_{\bar{x}'}$, (caso de amostras correlacionadas) donde, $s^2_{(\bar{x} - \bar{x})} = \frac{s^2}{n} + \frac{s^2}{n} - 2\, r\, \frac{s}{\sqrt{n}}\, \frac{s}{\sqrt{n}}$, na suposição de $s_x = s_{x'}$, e obtem-se então a expressão $s^2_{(\bar{x} - \bar{x}')} = \frac{2\, s^2}{n} (1 - r)$ idêntica à fórmula (13).

Conforme ficou verificado, no emparelhamento foi aplicado implicitamente o princípio de que as duas amostras apresentavam certa correlação.

Agora é simples discutir-se quando há vantagem em se emparelhar ou não os dados. A fórmula (13) mostra claramente que se r for negativo a variância determinada pelo método II é maior, sendo preferível utilizar-se então o método I. Entretanto, se r é positivo, o método do emparelhamento é melhor, pois, sendo menor a variância e portanto o desvio padrão, há mais facilidade em se revelar uma diferença significante. Se $r = 0$, os resultados são equivalentes nos dois métodos.

Na verdade, a situação é mais complicada, pois, no raciocínio acima não foi levado em conta o número de graus de liberdade. Quando êste número cresce o valor de t que determina a significância é menor e além disso é preferível uma estimativa baseada num maior número de graus de liberdade. Como no método dos grupos dispõe-se de um maior número de graus de liberdade é óbvio que as situações de r negativo e de r nulo favorecem êste método. Quando se considera o número de graus de liberdade perdido no emparelhamento, tem-se que mesmo para $r > 0$, o emparelhamento pode ainda não ser compensador para pequenos valores de r.

Fisher (3) calculou o que se chama de quantidade de informação e achou ser esta igual a $\frac{\nu + 1}{(\nu + 3)\ s^2}$ quando σ é estimado e igual a $\frac{1}{\sigma^2}$ quando se conhece exatamente σ. Dêste modo a eficiência E do emparelhamento em relação aos grupos pode ser expressa pela fórmula (14), segundo Cochran e Cox (1):

$$E = \frac{(\nu_2 + 1)\ (\nu_1 + 3)\ s^2_{(\bar{x} - \bar{x}')}}{(\nu_1 + 1)\ (\nu_2 + 3)\ s^2_{\bar{d}}} \quad (14)$$

, onde ν significa tanto nesta fórmula como na anterior o número de graus de liberdade, sendo ν_1 referente ao método I e ν_2 ao método II. O ajustamento ao número de graus de liberdade é importante se ν_1 e ν_2 são pequenos, caso contrário utiliza-se

$$E = \frac{s^2_{(\bar{x} - \bar{x}')}}{s^2_{\bar{d}}}$$

Como no emparelhamento há sacrifício da metade do número de graus de liberdade, há apenas n — 1, enquanto que no caso dos grupos dispõe-se de 2 (n — 1), é conveniente considerar a discussão do número de graus de liberdade e estudar a sua influência sôbre a eficiência E, principalmente nos casos de amostras pequenas. Tem-se então que:

$\nu_1 = 2n - 2$, $\nu_1 + 1 = 2n - 1$, $\nu_1 + 3 = 2n + 1$, $\nu_2 = n - 1$, $\nu_2 + 1 = n$, $\nu_2 + 3 = n + 2$

Substituindo-se êstes valores na fórmula (14), assim como o valor de $s^2_{\bar{d}}$ dado pela fórmula (13) e o de $s^2_{(\bar{x} - \bar{x}')}$ dado por $s^2_{(\bar{x} - \bar{x}')} = \frac{2s^2}{n}$, acha-se para E a expressão seguinte $E = \frac{n\ (2n + 1)}{(2n - 1)(n + 2)} \cdot \frac{1}{(1 - r)}$ (15), onde n é o número de pares ou de indivíduos em cada grupo e r é o coeficiente de correlação entre x e x'. Uma aproximação de

(15) pode ser obtida efetuando-se a divisão e desprezando-se os termos em $\frac{k}{n^m}$ para $m \geq 2$, $(k > 1)$. Embora fraca no caso geral, a aproximação é regular se n não for muito pequeno. Obtem-se dêste modo $E = \frac{n - 1}{n(1 - r)}$ (16). Uma simples inspecção desta fórmula mostra que para $r > \frac{1}{n}$ há vantagem em emparelhar os dados e isto corresponderá a pequenos valores positivos de r se n não for muito pequeno, aliás a situação em que a fórmula (16) é razoavelmente aconselhada.

Supondo-se $r = 0,5$ e $n = 10$, a aplicação da fórmula (16) daria $E = 1,80$, o que significa uma eficiência de 180% do emparelhamento em relação à comparação de grupos, ou seja que o emparelhamento melhorou a eficiência de 80%. Isto significa que o método I, i.e. , a comparação de grupos, exigiria 18 observações enquanto o método II necessitaria apenas 10 para fornecer a mesma informação. Houve no primeiro método um desperdício de 8 observações, o que significa maior despeza no experimento. A aplicação da fórmula (15) dá para E o valor 1,84, uma diferença prática pequena em relação a 1,80.

Na prática, como é sabido, os pares são escolhidos de tal modo que há sempre alguma correlação positiva entre x e x', de maneira que o emparelhamento redunda num método mais refinado que o da comparação de grupos. Em caso de dúvida de haver correlação positiva entre x e x', seria aconselhavel o experimentador traçar o diagrama de dispersão (scatter diagram) e por mera inspecção visual verificar se deve supor ou não esta correlação. Seria perda de tempo determinar o valor de r utilizando-se da fórmula (11) e ainda de menor importância, fazer o teste de significância de r, dado a sua precaridade nas pequenas amostras.

Exemplo prático

A aplicação numérica será feita num exemplo citado por Love (4). Trata-se de uma competição de duas variedades de trigo A e B, plantadas em canteiros adjacentes, com 10 repetições.

TABELA I

Produções em bushels por acre

	A (x)	B (x')	A — B (x — x' = d)
	41	32	9
	35	31	4
	37	31	6
	33	27	6
	31	29	2
	32	28	4
	28	28	0
	24	24	0
	39	39	0
	32	35	3
Totais	332	304	28

Pelo método I, não se leva em conta o fato de cada par de dados referir-se às produções de canteiros próximos.

Utilizando-se as fórmulas (5), (6) e (7), acha-se para a variância da diferença das médias o valor 4,42, isto é $s^2_{(\bar{x}-\bar{x}')} = 4{,}42$ e portanto $t = \frac{33{,}2 - 30{,}4}{\sqrt{4{,}42}} = 1{,}333$. Êste valor de t com 18 graus de liberdade não é significante no ponto 5% e portanto, o pesquisador seguindo a prática usual aceita a hipótese de que as duas variedades não diferem em produção.

Se ao invés do método I, da comparação de grupos, for usado o método II, das comparações emparelhadas, os dados a serem utilizados serão as diferenças $x - x'$. Empregando-se a fórmula (8) acha-se $t = 2{,}435$, valor significante no ponto 5%, para 9 graus de liberdade. É facil verificar que $\bar{d} = 2{,}8 = \bar{x} - \bar{x}'$ e $s^2_{\bar{d}} = 1{,}33$, portanto $t = \frac{2{,}8}{\sqrt{1{,}33}} = 2{,}435$. Agora a hipótese de nulidade seria rejeitada, isto é aceita-se que a variedade A é superior à B em produção.

Como se vê, êste é um exemplo típico em que o método I conduz a uma conclusão errônea. Qual é a razão? Como as variedades foram plantadas em canteiros próximos, é natural supor que êstes apresentam homogeneidade de fertilidade, de modo que ao se formarem os pares, foi eliminada da variação a parte referente à variação entre os blocos formados de dois canteiros. O que aconteceu aqui é que houve uma variação bem apreciável entre os blocos, o que pode ser determinado fàcilmente pela análise de variância. Por outro lado, os canteiros do mesmo bloco apresentaram uma correlação apreciável. Aplicando-se a fórmula (13), acha-se para r o valor 0,7 dado por $1,33 = 4,42\ (1 - r)$, aproximadamente o mesmo que se usando a fôrmula (11). A diferença é devida ùnicamente às aproximações.

Pode-se medir neste experimento a eficiência do emparelhamento sôbre os grupos. Utilizando-se a fórmula (15) com o valor de $r = 0,7$ acha-se $E = 3,07$, e $E = 3,06$ quando se usa $s\frac{2}{d}$ e $s_{(\bar{x} - \bar{x})}$, segundo a fórmula (14). Com a fórmula aproximada (16), $E = 3,00$. Pràticamente, o emparelhamento trouxe uma melhoria de 200%. No caso em apreço, chegou mesmo a decidir o sucesso do experimento.

Summary

In this paper the author considered group and paired comparisons and has shown by theoretical considerations when one method is preferable to the other. The subject was put in quantitative reasoning and it was proved that when the data of the two groups are positively correlated, pairing is better. Also it was considered the number of degrees of freedom lost in pairing. Finally, with a practical example, it was shown the advantage of pairing in relation to group comparison. This last method was completely fallacious when applied to the given example.

BIBLIOGRAFIA CITADA

(1) Cochran, W. G. and Cox, G. M.
(1950) Experimental Designs, p. 29
John Wiley & Sons, Inc., New York

(2) Fisher, R. A.
(1946) Statistical Methods for Research Workers, pp. 124-125, 10th. edition
Oliver and Boyd. London

(3) Fisher, R. A.
(1947) The Design of Experiments, pp. 234-236, 4th. edition
Hafner-Publishing Co., Inc., New York

(4) Love, H. H.
(1943) Experimental Methods in Agricultural Research, pp. 13-16
University of Puerto Rico, Rio Piedras

(5) Snedecor, G. W.
(1946) Statistical Methods, p. 84, 4th. edition
Iowa State College Press, Ames