COMPARAÇÃO DE VALORES MÉDIOS POPULACIONAIS: UM PROBLEMA COMUM

*Sara Morgado Nunes**
*João Renato Sebastião**

RESUMO

Uma das questões que surge com mais frequência em Estatística Aplicada é a comparação de valores médios populacionais. Neste trabalho procede-se a uma breve revisão de alguns testes paramétricos e não paramétricos mais utilizados na abordagem desta questão e apresentam-se também os procedimentos bootstrap enquanto metodologia alternativa eficaz no tratamento do problema em questão.

Num estudo de simulação aplica-se o Teste T, o teste de Mann-Whitney-Wilcoxon e o bootstrap a amostras geradas artificialmente, com o objectivo de comparar valores médios populacionais e confrontar os resultados das metodologias aplicadas.

1. INTRODUÇÃO

A igualdade de dois valores médios populacionais desconhecidos constitui com frequência uma hipótese a testar em problemas que emergem em estudos das mais diversas áreas. Imagine-se que se pretende comparar os tempos médios de duração de dois tipos de lâmpadas obtidas através de dois processos de fabrico diferentes *A* e *B*. Mais concretamente, pretende-se saber se o tempo médio de duração das lâmpadas obtidas pelo processo *A* difere significativamente do tempo médio de duração das lâmpadas obtidas pelo processo *B*, ou seja, testar a hipótese "$\mu_A = \mu_B$", onde μ_A e μ_B denotam o tempo médio de duração das lâmpadas obtidas pelos processos *A* e *B*, respectivamente. Uma vez que se desconhecem os valores médios populacionais, tal comparação poderá ser baseada em estimativas dos tempos médios de duração de ambos os processos.

Formalizando o problema, considerem-se duas populações contínuas *X* e *Y* com valores médios desconhecidos μ_X e μ_Y, respectivamente. A partir das populações *X* e *Y* é possível obter amostras aleatórias independentes de dimensão *m* e *n*, respectivamente: $X_1, X_2, \ldots, X_m$ e $Y_1, Y_2, \ldots, Y_n$. Com base nestas duas amostras pretende-se testar a hipótese nula $H_0 : \mu_X = \mu_Y$ *versus* a hipótese alternativa $H_1 : \mu_X \neq \mu_Y$ (teste bilateral), $H_1 : \mu_X < \mu_Y$ (teste unilateral à esquerda) ou $H_1 : \mu_X > \mu_Y$ (teste unilateral à direita).

O objectivo de qualquer teste estatístico é fornecer ferramentas que permitam validar ou rejeitar a hipótese nula, com base na informação obtida a partir das amostras aleatórias disponíveis. Assim, a ideia subjacente a qualquer teste de hipóteses é especificar o conjunto de valores que conduzem à rejeição de H_0, constituindo tal conjunto a região de rejeição do

* Docente da Escola Superior de Gestão do Instituto Politécnico de Castelo Branco.

teste. Porém, qualquer que seja a decisão final, existe sempre o risco de se estar a tomar a decisão errada. Na verdade, podem ser cometidos dois tipos de erros: rejeitar H_0, sendo H_0 verdadeira (Erro Tipo I) ou não rejeitar H_0, sendo H_0 falsa (Erro Tipo II).

Notando por *a* a probabilidade de se cometer um Erro Tipo I e por *b* a probabilidade de se cometer um Erro Tipo II, tem-se

$$\alpha = P(Erro\ Tipo\ I) = P(rejeitar\ H_0 \mid H_0\ verdadeira) \quad \text{e}$$
$$\beta = P(Erro\ Tipo\ I\ ) = P(manter\ H_0 \mid H_0\ falsa) \ .$$

Daqui resulta que

$$1-\alpha = P(manter\ H_0 \mid H_0\ verdadeira)$$
$$\text{e} \quad 1-\beta = P(rejeitar\ H_0 \mid H_0\ falsa) \ .$$

Assim, o ideal seria conseguir que

$$P(rejeitar\ H_0 \mid H_0\ falsa) \to 1$$
$$\text{e} \quad P(rejeitar\ H_0 \mid H_0\ verdadeira) \to 0 \ .$$

Enquanto *a* denota o nível de significância do teste e fixa-se geralmente em *5%*, o valor de *b* é mais difícil de determinar e está relacionado com o poder do teste. Chama-se potência do teste ao valor $1-\beta$, isto é, à probabilidade de se rejeitar H_0 quando esta é falsa. Assim, não existe um único valor para *b* mas infinitos, dando origem à função potência.

Para dar resposta ao problema da comparação de dois valores médios populacionais, é frequente recorrer-se aos testes paramétricos quando é possível assumir que as populações em causa são Normais ou aos testes não paramétricos caso a Normalidade não esteja assegurada. Porém, na prática, estas soluções nem sempre são viáveis por não se verificarem alguns pressupostos necessários à aplicação dessas técnicas, o que ocorre frequentemente em dados do âmbito das Ciências Sociais. Como alternativa aos testes de hipóteses conhecidos, é possível recorrer às técnicas de Monte Carlo que consistem em gerar amostras artificialmente com o objetivo de dados comportamentos da estatística em causa. O recurso a estas técnicas computacionalmente intensivas tem aumentado consideravelmente à medida em que possibilitam a Inferência Estatística sem que para isso seja necessário recorrer a cálculos matemáticos complexos.

Neste trabalho começam por apresentar-se as "ferramentas" estatísticas tradicionais mais utilizadas na abordagem da questão da comparação de dois valores médios populacionais e, de seguida, a metodologia bootstrap como solução alternativa para o problema apresentado. Na última parte apresentam-se os resultados de um estudo de simulação levado a efeito no *software* estatístico *R-Project*, que assentou na comparação de valores médios com base nas metodologias paramétricas, não paramétricas e bootstrap, recorrendo a amostras geradas artificialmente.

2. TESTES PARAMÉTRICOS E NÃO PARAMÉTRICOS

Os Testes de Hipóteses são amplamente utilizados na Inferência Estatística para validar ou rejeitar afirmações relativas a parâmetros populacionais. No vasto grupo dos Testes de Hipóteses há a considerar os testes paramétricos e os não paramétricos. Os testes paramétricos exigem, em geral, a verificação de pressupostos mais sólidos que os não paramétricos e são aplicáveis unicamente a variáveis quantitativas, enquanto os testes não paramétricos podem aplicar-se a qualquer variável de nível ordinal.

Veja-se como a metodologia paramétrica poderia dar resposta ao problema da comparação de dois valores médios populacionais.

Quando se tem $X_1, X_2, \ldots, X_m$ e $Y_1, Y_2, \ldots, Y_n$ provenientes de populações $N(\mu_X, \sigma_Y^2)$ e $N(\mu_X, \sigma_Y^2)$, respectivamente, com variâncias populacionais σ_X^2 e σ_Y^2

conhecidas, padroniza-se a variável aleatória $\bar{X}-\bar{Y}$, onde $\bar{X}$ e $\bar{Y}$ denotam as respectivas médias amostrais, e recorre-se à estatística de teste

$$Z_0 = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n}}} \sim N(0,1) \qquad \textbf{(1)}$$

Fixando *a* como o nível de significância do teste, podem obter-se os valores críticos $-z_{\alpha/2}$ e $z_{\alpha/2}$ que denotam os quantis de probabilidade $\alpha/2$ e $1-\alpha/2$ da distribuição *Normal (0,1)*. Para um teste bilateral rejeita-se H_0 quando $Z_0 < -z_{\alpha/2}$ ou $Z_0 > z_{\alpha/2}$, ficando assim definidas as áreas de rejeição e aceitação da hipótese nula.

Porém, na prática, raramente se conhecem as variâncias populacionais. Assim, quando σ_X^2 e σ_Y^2 são desconhecidas mas se pode considerar que são da mesma ordem de grandeza, estão reunidas as condições para a aplicação do teste *"t" de Student* para a diferença de médias. Neste caso estima-se σ_X^2 e σ_Y^2 a partir dos dados amostrais e obtém-se a estatística de teste

$$t_0 = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{1}{m}+\frac{1}{n}}\sqrt{\frac{(m-1)S_X^2+(n-1)S_Y^2}{m+n-2}}} \sim t_{m+n-2} \qquad \textbf{(2)}$$

onde S_X^2 e S_Y^2 denotam as variâncias amostrais corrigidas das respectivas amostras e t_{m+n-2} a distribuição *t-Student* com *m+n-2* graus de liberdade. Neste caso, para um teste bilateral, rejeita-se H_0 quando $t_0 < -t_{m+n-2;\alpha/2}$ ou $t_0 > t_{m+n-2;\alpha/2}$, sendo $t_{m+n-2;\alpha/2}$ o quantil de probabilidade $1-\alpha/2$ da distribuição t_{m+n-2}.

A situação em que não se pode considerar $\sigma_X^2 = \sigma_Y^2$, constitui o clássico problema de Behrens-Fisher. Best e Rayner (1987) demonstram o valor prático da solução de Welch que

$$\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}}}$$

se baseia no facto da estatística seguir uma distribuição *t-Student*. Porém, vários estudos demonstraram que a solução de Welch nem sempre conduz a resultados válidos uma vez que não satisfaz $P(p=\alpha)=\alpha$ para certas combinações de *m* e *n*.

Quando as amostras disponíveis possuem dimensão superior a 30, o Teorema do Limite Central permite demonstrar que variável aleatória $\bar{X}-\bar{Y}$ segue uma distribuição *Normal*, bastando assim substituir na expressão **(1)** σ_X^2 e σ_Y^2 por S_X^2 e S_Y^2, respectivamente.

O objectivo de qualquer teste paramétrico é inferir sobre o parâmetro populacional a partir da sua estimativa numa amostra. Dado que o processo de estimação envolve a amostra, a distribuição amostral e a população, são necessários alguns pressupostos que permitam assegurar a compatibilidade destas componentes, havendo portanto a necessidade de se assumir uma distribuição subjacente. Assim, a maior limitação dos testes apresentados é a suposição de Normalidade.

Depois de rever cerca de 400 conjuntos de dados recolhidos em estudos do âmbito

das Ciências Sociais, Micceri (1989) concluiu que a maior parte destes dados não seguiam distribuições Normais. Breckler (1990) examinou 72 artigos do domínio da Psicologia e observou que apenas em 19% era possível assumir o pressuposto de Normalidade e que em menos de 10% dos estudos se havia tido o cuidado de observar se este pressuposto estava a ser violado. Keselman et al (1998) referem que muitos investigadores raramente verificam se os requisitos na aplicação dos testes paramétricos são satisfeitos. É um facto que os dados recolhidos no âmbito das Ciências Sociais geralmente violam os pressupostos paramétricos.

Se não for possível assumir que as populações *X* e *Y* seguem distribuições Normais, ou a homogeneidade das variâncias ou ainda se as amostras em causa são de pequena dimensão ou as variáveis são de nível ordinal, não se verificam os pressupostos paramétricos. Nestas circunstâncias, pode optar-se por um teste não paramétrico. Se por um lado este tipo de testes são, em geral, menos potentes que os testes paramétricos, por outro, os testes não paramétricos baseiam-se na ordem das observações, o que parece constituir uma vantagem para dados do âmbito das Ciências Sociais que são muitas vezes expressos em escalas de Likert.

O teste não paramétrico mais adequado para comparar os valores médios de duas populações foi estudado por Mann e Whitney (1947) e descrito por Wilcoxon, ficando assim conhecido por Teste de Mann-Whitney-Wilcoxon. Apesar de se tratar de um teste não paramétrico, à aplicação do Teste de Mann-Whitney-Wilcoxon está subjacente o pressuposto de que as populações em causa possuem a mesma forma, diferindo apenas, eventualmente, nos parâmetros de localização m_X e m_Y.

O Teste de Mann-Whitney-Wilcoxon baseia-se nas ordens atribuídas às observações disponíveis depois de ordenadas. Assim, dadas duas amostras aleatórias $X_1, X_2, \ldots, X_m$ e $Y_1, Y_2, \ldots, Y_n$, começam por se ordenar as $m+n$ observações de forma crescente, atribuindo-se um número de ordem a cada uma. No caso de existirem valores empatados, atribui-se-lhes a média das ordens que lhes corresponderiam caso fossem distintos. A estatística de teste é dada pela soma das ordens relativas a uma das amostras. Para simplificar convenciona-se escolher a amostra de menor dimensão. Seja então $X_1, X_2, \ldots, X_m$ a amostra de menor dimensão, isto é, $m \leq n$, e U a *soma das ordens da amostra X*. Mann e Whitney demonstraram que U tem distribuição aproximadamente Normal para $n > 8$, com valor médio e variância dados por

$$m_U = \frac{m(m+n+1)}{2} \qquad s_U^2 = \frac{mn(m+n+1)}{12}$$

Logo,

$$Z_U = \frac{U - \frac{m(m+n+1)}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \sim N(0,1)$$

Assim, para um teste bilateral, rejeita-se H_0 quando $Z_U < -z_{\alpha/2}$ ou $Z_U^{(3)} > z_{\alpha/2}$.

Os procedimentos não paramétricos são muitas vezes criticados pela perda de informação decorrente da transformação dos dados em ordens. Por outro lado, sabe-se que, em geral, os testes não paramétricos são menos potentes que os testes paramétricos, ou seja, a probabilidade de se rejeitar H_0 quando esta é falsa é superior nos testes paramétricos.

3. A METODOLOGIA BOOTSTRAP

A difusão em larga escala dos meios informáticos veio contribuir para o desenvolvimento e expansão de metodologias computacionalmente intensivas. Actualmente os meios informáticos e *softwares* disponíveis colocam ao alcance de todos instrumentos eficazes na

resolução de muitos problemas estatísticos sem ser necessário o recurso a fórmulas matemáticas complexas. Uma dessas revolucionárias metodologias é o bootstrap que permite também dar resposta ao problema inicialmente apresentado. O bootstrap está ligado às técnicas de Monte Carlo cujo nome resulta da analogia com as casas de jogo na Riviera Francesa. Ao estudarem como tornar máxima a probabilidade de ganhar, alguns jogadores recorriam a simulação para analisar a ocorrência de cada caso. Hoje a simulação Monte Carlo constitui uma técnica muito conhecida no âmbito da Estatística e largamente utilizada em todo o mundo.

O princípio básico da metodologia bootstrap proposta por Efron (1979) é a utilização repetida da amostra original de modo a obter várias estimativas do parâmetro de interesse que serão depois usadas para inferir acerca das suas características. Esta metodologia surgiu inicialmente numa tentativa de estimar o erro padrão de uma determinada estatística mas logo foi aperfeiçoada por Efron e Tibshirani (1986, 1993) que a apresentam como uma técnica de base computacional adequada a vários problemas de estimação. O bootstrap veio, sem dúvida, revolucionar toda a Estatística e actualmente aplica-se com frequência nas mais diversas áreas de investigação.

Quando os testes apresentados na secção anterior não se mostram adequados, o bootstrap pode revelar-se uma técnica útil no contexto dos testes de hipóteses. Tendo em consideração o problema inicialmente apresentado, Efron e Tibshirani (1993) e Davison e Hinkley (1997), descrevem a aplicação do bootstrap à comparação de valores médios populacionais através do seguinte algoritmo:

1) A partir de cada uma das amostras iniciais $x_1, x_2, \ldots, x_m$ e $y_1, y_2, \ldots, y_n$ obtêm-se, com reposição, B amostras bootstrap independentes de dimensão m e n, respectivamente $x_1^{*b}, x_2^{*b}, \ldots, x_m^{*b}$ e $y_1^{*b}, y_2^{*b}, \ldots, y_n^{*b}$, $b=1,\ldots,B$. Na prática convém ter $B \geq 1000$.

2) Em cada uma das B amostras simuladas calculam-se as médias $\bar{x}^{*b}$ e $\bar{y}^{*b}$, e a respectiva diferença $\bar{x}^{*b} - \bar{y}^{*b}$, $b=1,\ldots,B$.

3) Aproxima-se a distribuição da estatística teste

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}} \qquad \textbf{(4)}$$

pela distribuição empírica T^*, simulada a partir dos valores

$$t^* = \frac{(\bar{x}^{*b} - \bar{y}^{*b}) - (\bar{x} - \bar{y})_{obs}}{\sqrt{\frac{S_X^{2*b}}{m} + \frac{S_Y^{2*b}}{n}}} \qquad \textbf{(5)}$$

onde S_X^{2*b} e S_Y^{2*b} denotam as variâncias amostrais corrigidas nos conjuntos de dados simulados e $(\bar{x} - \bar{y})_{obs}$ denota a diferença das médias amostrais nas amostras originais.

4) Fixando um nível de significância a, estima-se $\hat{t}_a$, o quantil de probabilidade a.

5) Para um teste de hipóteses unilateral à esquerda, rejeita-se H_0 quando $T_{obs} < \hat{t}_a$, sendo T_{obs} o valor da estatística de teste obtido a partir das amostras originais.

Em geral, os *softwares* estatísticos recorrem ao chamado *p-value* e não à estatística

de teste para tomar a decisão estatística. O *p-value* representa o menor valor de *a* para o qual os dados observados evidenciam que a hipótese nula deve ser rejeitada, fornecendo assim uma maior informação na medida em que permite analisar até que ponto os dados observados discordam da hipótese nula.

Para um teste unilateral à esquerda, o *p-value* é dado por

$$p = \frac{1+\#\left\{T^{*} \leq T_{obs}\right\}}{B+1} \qquad (6)$$

onde $\#\left\{T^{*} \leq T_{obs}\right\}$ denota o número de valores T^{*} que são inferiores ou iguais ao valor T_{obs}. Fixado um nível de significância *a*, rejeita-se H_0 quando $a>p$.

4. ESTUDO DE SIMULAÇÃO

Com o objectivo de comparar o desempenho das metodologias apresentadas enquanto meio de resposta ao problema da comparação de dois valores médios populacionais, levou-se a efeito, no *software* estatístico *R-Project*, um Estudo de Simulação que consistiu na exploração de três situações distintas:

- Na situação 1 consideram-se amostras provenientes de populações Normais com variâncias idênticas, sendo portanto um caso em que é válida a aplicação do Teste T;
- Na situação 2 consideram-se amostras provenientes de populações que, não sendo Normais, possuem a mesma forma, ou seja, falham os pressupostos de Normalidade mas verificam-se as premissas necessárias à aplicação do Teste de Mann-Whitney-Wilcoxon;
- Na situação 3 consideram-se amostras para as quais não estão verificados os pressupostos de Normalidade nem a identidade da forma das populações em causa, tratando-se portanto de um caso a que não pode ser aplicado o Teste T ou o Teste de Mann-Whitney-Wilcoxon.

4.1. SITUAÇÃO 1

Suponha-se que duas máquinas de empacotamento A e B de uma fábrica de rações fornecem embalagens cujo peso tem uma distribuição Normal. O director de produção suspeita que o peso médio das embalagens resultantes da máquina A é inferior ao peso médio das embalagens resultantes da máquina B. Para confirmar a sua suspeita recolheu duas amostras aleatórias A e B de embalagens resultantes das duas máquinas e obteve os seguintes pesos (em *kg*):

A	50.068	50.019	49.821	50.027	50.179	49.946	50.164	50.070	50.262
	50.087	49.925	49.820	50.032	50.004	49.960	50.003	50.051	50.229
	50.115	50.030	50.088	49.873	50.020	49.997	49.947		
B	50.671	50.503	50.315	50.586	50.479	50.583	50.480	50.463	50.559
	50.498	50.756	50.517	50.729	50.554	50.448	50.423	50.365	50.310
	50.557	50.619	50.384	50.598	50.496	50.418	50.523		

As hipóteses a testar são H_0: *O peso médio das embalagens produzidas pela máquina*

A não difere significativamente do peso médio das embalagens produzidas pela máquina B, isto é, $\mu_A = \mu_B$ *versus* H_1*: O peso médio das embalagens produzidas pela máquina A é inferior ao peso médio das embalagens produzidas pela máquina B*, isto é, $\mu_A < \mu_B$. Tratando-se de populações Normais, estão verificados os pressupostos para a aplicação do Teste T.

Calculando-se a estatística de teste definida em **(2)**, obteve-se $t_0=-4.304$. Para um nível de significância de $a=0.05$, o valor crítico é dado por $-t_{18}(0.05)=-1.677$. Tratando-se de um teste unilateral à esquerda, rejeita-se H_0 dado que $t_0=-4.304 < -t_{18}(0.05)= -1.677$, ou seja, admite-se que o peso médio das embalagens produzidas pela máquina A é inferior ao peso médio das embalagens produzidas pela máquina B.

Outra forma de dar resposta ao problema consistiu em testar as hipóteses formuladas através da metodologia bootstrap. Para tal, começaram por gerar-se 999 amostras simuladas nas quais se calcularam as réplicas bootstrap definidas em **(5)**, obtendo-se a distribuição empírica da estatística de teste definida em **(4)**:

Fig.1 - Histograma bootstrap das 999 réplicas bootstrap t^* para a situação 1.

Considerando um nível de significância de $a=0.05$, tem-se $t^*_{(0.05)} = -1.735$ (assinalado na Fig.1). Dado que o valor da estatística de teste nas amostras observadas é $T_{obs} = -5.216$, então rejeita-se H_0 uma vez que $T_{obs} = -5.216 < t^*_{(0.05)} = -1.735$.

Como atrás se referiu, o cálculo do *p-value* definido em **(6)** constitui

uma forma alternativa para se chegar à decisão estatística. Para esta situação encontrou-se *p-value=0.001*, o qual, sendo inferior ao nível de significância *a=0.05*, indica que deve rejeitar-se H_0.

Como se viu, ambas as metodologias conduziram à rejeição de H_0. Com o objectivo de obter mais informação acerca do desempenho das duas metodologias aplicadas, compararam-se as funções potência resultantes de ambos os testes, para um nível de significância *a=0.05*:

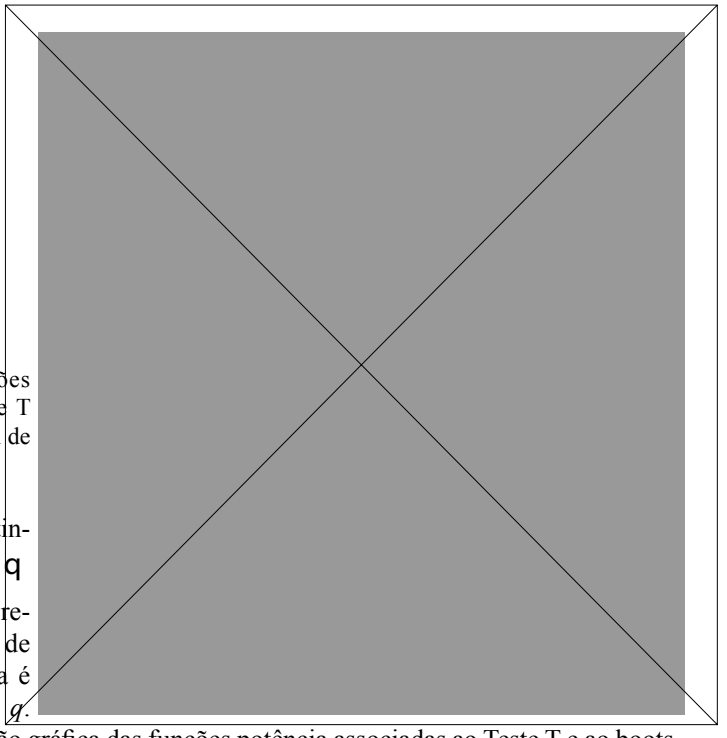

Fig.2 – Gráfico das funções potência associadas ao Teste T e ao bootstrap, para um nível de significância *a=0.05*.

Note-se que, admitindo que se tem $m_B - m_A = q$, a função potência *f(q)*, representa a probabilidade de se rejeitar H_0 quando esta é falsa, para cada valor de *q*. Observando a representação gráfica das funções potência associadas ao Teste T e ao bootstrap, não se encontram diferenças significativas, pelo que se conclui que, para a situação apresentada, as metodologias utilizadas, conduzem a resultados semelhantes.

4.2 SITUAÇÃO

Suponha-se que um grupo de cientistas pretende estudar se sujeitos que residem a altitudes elevadas possuem maior concentração de hemoglobina no sangue que sujeitos que se encontram a baixas altitudes. Para tal, seleccionaram aleatoriamente 25 habitantes de uma vila costeira (X) e 25 habitantes de uma vila situada a 2000 metros de altitude (Y). Após se medir o nível de hemoglobina a cada sujeito, obtiveram-se as seguintes concentrações (em *g/dl*):

X	21.62	9.73	13.33	13.72	8.42	16.31	7.7	4.98	11.63
	4.84	6.45	6.37	14.54	18.02	7.68	11.14	6.52	8.23
	10.7	11.61	18.03	8.81	4.32	6.28	11.71		

Y	19.82	19.58	21.82	9.35	11.49	11.34	9.41	9.53	17.54
	24.2	20.22	19.25	16.02	14.49	11.1	14.66	16.03	19.59
	11.47	16.91	15.66	10.01	18.28	24.31	25.81		

As hipóteses a testar são H_0: *A concentração média de hemoglobina não difere significativamente entre sujeitos que habitam a altitudes elevadas ou baixas*, isto é, $m_A = m_B$ *versus* H_1: *A concentração média de hemoglobina em sujeitos que habitam a baixas altitudes é inferior à dos sujeitos que habitam a altitudes elevadas,* isto é, $m_A < m_B$.

Com o objectivo de se obter alguma informação acerca da forma das distribuições em causa, procedeu-se à observação dos histogramas das amostras X e Y, que apresentam a seguinte forma:

Fig.3 - Histogramas dos dados observados nas amostras **X** e **Y**.

Os histogramas da Fig.3 indiciam que não se verifica o pressuposto de Normalidade e portanto a aplicação de um teste paramétrico não se adequa a esta situação. Assim, optou-se por recorrer ao Teste de Mann-Whitney-Wilcoxon.

Da aplicação do Teste de Mann-Whitney-Wilcoxon resultou a estatística de teste definida em **(3)**, Z_U=*-3.368* . Para um nível de significância de *a=0.05*, o valor crítico é dado por *-z(0.05)=-1.645*. Visto tratar-se de um teste unilateral à esquerda, o facto de se ter Z_U=*-3.368* < *-z(0.05)=-1.645* conduz à rejeição de H_0, ou seja, admite-se que a concentração média de hemoglobina em indivíduos que habitam a baixas altitudes é de facto inferior à de indivíduos que habitam a altitudes elevadas, com um nível de confiança de 95%.

De forma análoga ao que se fez na Situação 1, também aqui se abordou o problema apresentado através da metodologia bootstrap. Calculando a estatística de teste nas amostras observadas, obteve-se $T_{obs} = -4.283$ a qual, sendo inferior ao valor do quantil de probabi-

lidade 0.95, $\hat{t}_{(6)} = 1.760$, conduz igualmente à rejeição de H_0. Se se optasse pelo cálculo do *p-value*, obter-se-ia *p-value=0.001* o qual, sendo inferior ao nível de significância *a=0.05*, levaria à mesma decisão estatística.

Como também aqui ambas as metodologias conduziram à rejeição de H_0, optou-se pela comparação das funções potência resultantes de ambos os testes, para um nível de significância *a=0.05*:

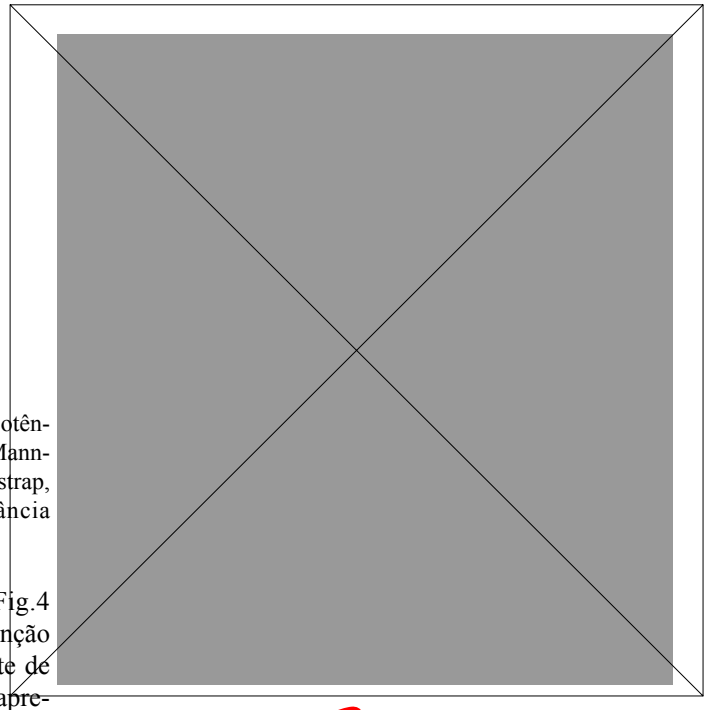

Fig.4 – Gráfico das funções potência associadas ao Teste de Mann-Whitney-Wilcoxon e ao bootstrap, para um nível de significância *a=0.05*.

A observação da Fig.4 permite constatar que a função potência associada ao Teste de Mann-Whitney-Wilcoxon apresenta valores significativamente inferiores à função potência associada ao bootstrap. Este facto permite concluir que a probabilidade de se rejeitar H_0 quando esta é falsa é, em geral, mais elevada para o bootstrap que para o Teste de Mann-Whitney-Wilcoxon, ou seja, o teste bootstrap revela-se mais potente que o Teste de Mann-Whitney-Wilcoxon.

4.3. SITUAÇÃO 3

Suponhamos que a Direcção Regional de uma dada instituição de crédito pretende averiguar se o valor médio de empréstimos contraídos para habitação pelos clientes da cidade de Castelo Branco é menor que na cidade da Covilhã. Para tal foram seleccionados aleatoriamente 20 clientes da dependência bancária de Castelo Branco e 20 clientes da dependência bancária da Covilhã, registando-se o respectivo montante dos empréstimos contraídos para habitação (em milhares de €):

CB	104.71	99.96	100.79	99.94	100.48	95.68	102.09	98.24	103.16
	102.01	96.73	97.04	96.27	103.19	101.69	100.46	100.44	101.67
	99.78	92.16							

C	101.98	99.11	105.1	101.8	100.03	100.57	99.68	103.49	103.26
	101.97	100.92	100.75	100.88	102.19	100.82	103.56	100.12	99.35
	105.72	100.41							

As hipóteses a testar são H_0: *O volume médio de empréstimos contraídos para habitação pelos clientes de Castelo Branco e da Covilhã não difere significativamente*, isto é, $m_B = m_C$ *versus* H_1: *O volume médio de empréstimos contraídos para habitação pelos clientes de Castelo Branco é inferior ao volume médio contraído na Covilhã*, isto é, $m_B < m_C$.

Numa primeira fase observaram-se os histogramas relativos às amostras CB e C:

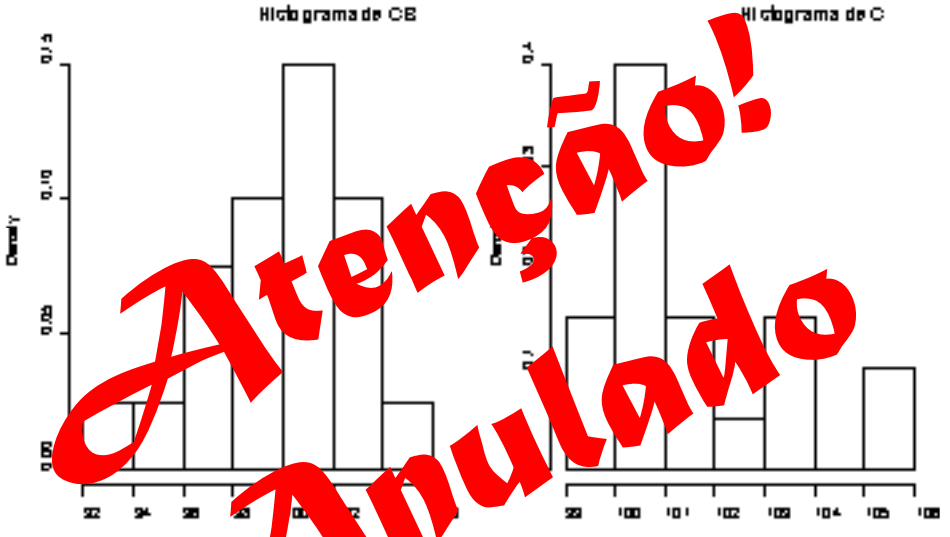

Fig.5 - Histograma dos dados observados nas amostras **CB** e **C**.

Dado não se verificarem os pressupostos necessários quer à aplicação de um Teste Paramétrico, quer à aplicação do Teste de Mann-Whitney-Wilcoxon, o problema em questão foi abordado com base na metodologia bootstrap.

A partir das 999 réplicas bootstrap aproximou-se a distribuição empírica da estatística de teste definida em **(4)**:

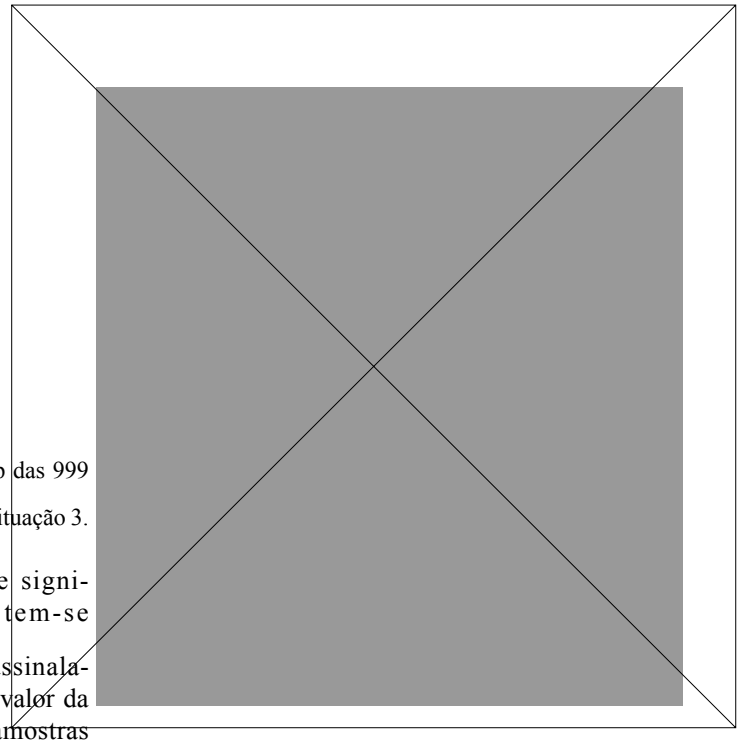

Fig.6 - Histograma bootstrap das 999 réplicas bootstrap t^* para a situação 3.

Para um nível de significância de $a=0.05$, tem-se $\hat{t}_{0.05} = -1.579$ (assinalado na Fig.6), enquanto o valor da estatística de teste nas amostras observadas é de $T_{obs} = -2.225$. Dado que se tem $T_{obs} = -2.225 < \hat{t}_{0.05} = -1.579$, rejeita-se H_0.

Em termos de *p-value*, tem-se *p-value*=0.027, valor sendo inferior ao nível de significância $a=0.05$, conduz à rejeição de H_0 e, com uma margem de erro de *5%*, valida-se a hipótese de que o volume médio de empréstimos contraídos para habitação pelos clientes de Castelo Branco é inferior ao volume médio contraído na Covilhã.

A função potência resultante do teste bootstrap, para um nível de significância $a=0.05$ é a seguinte:

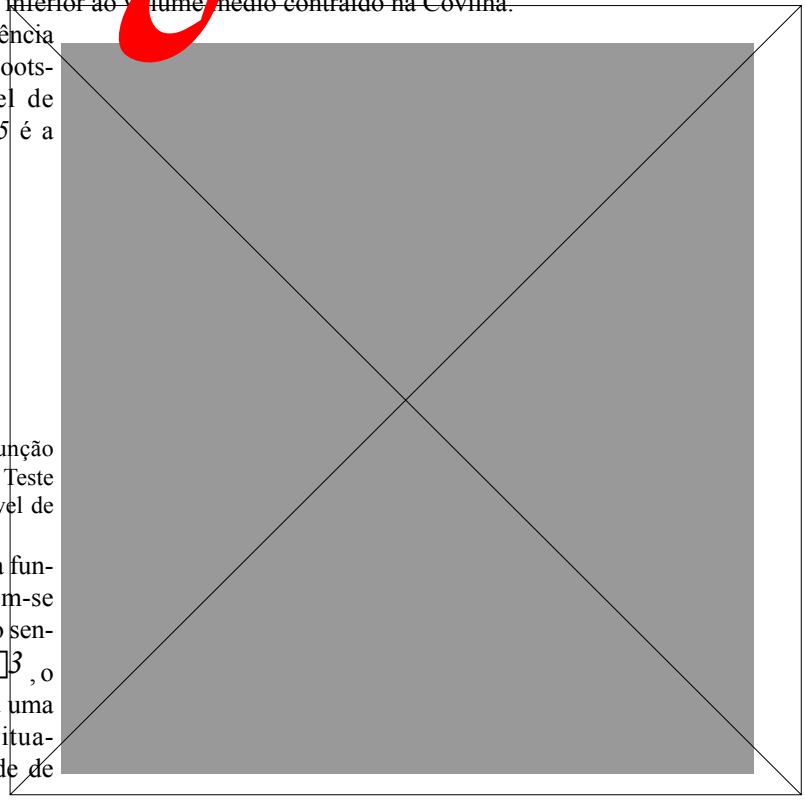

Fig.7 – Gráfico da função potência associada ao Teste bootstrap, para um nível de significância $a=0.05$.

Note-se que a função potência mantém-se no seu valor máximo sensivelmente até $è = -3$, o que indica que, para uma grande parte das situações, a probabilidade de

se rejeitar H_0 quando esta é falsa, é máxima, indiciando que, alterações às diferenças dos valores médios são rapidamente detectadas.

5. CONSIDERAÇÕES FINAIS

Neste trabalho procedeu-se a uma revisão dos testes paramétricos, não paramétricos e bootstrap enquanto metodologias adequadas à abordagem da questão da comparação de dois valores médios populacionais. Apesar de serem, em geral mais potentes que os testes não paramétricos, na prática, os testes paramétricos nem sempre são aplicáveis por não se verificarem os pressupostos necessários. Por sua vez, o bootstrap surge como uma metodologia alternativa aos testes conhecidos cuja aplicação não exige a verificação de qualquer tipo de pressuposto.

Num estudo de simulação comparam-se as funções potência resultantes da aplicação do Teste T, do Teste de Mann-Whitney-Wilcoxon e do bootstrap a amostras geradas artificialmente com o objectivo de comparar os respectivos valores médios populacionais. Conclui-se que, além de ser uma técnica de aplicação muito geral, o bootstrap revela-se, em muitos casos, mais potente que o Teste de Mann-Whitney-Wilcoxon e uma alternativa válida ao Teste T quando os pressupostos de Normalidade não estão assegurados.

REFERÊNCIAS

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, **2**, 291-309.

Best, D. J. e Rayner, J. C. W. (1987). Welch's approximate solution for the Behrens-Fisher Problem. *Technometrics*, **29**, 205-210.

Davison, A. C. e Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, **7**, 1-26.

Efron, B. e Tibshirani, R. J. (1986). Bootstrap methods for standart errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, **1**, 54-77.

Efron, B. e Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, Nova Iorque.

Keselman, H. J., Huberty, C., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D. e Keselman, J. C. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analysis. *Review of Educational Research*, **68**, 350-386.

Micceri, T. (1989). The unicorn, the normal curve and other improbable creatures. *Psychological Bulletin*, **105**, 156-166.

Contactos:
Escola Superior de Gestão de Idanha-a-Nova
Largo do Município
6060 Idanha-a-Nova