COMPARAÇÃO DOS PODERES DOS TESTES T DE STUDENT E MAN-WHITNEY WILCOXON PELO MÉTODO DE MONTE CARLO

André Vilela Komatsu[1]

Resumo: *O cálculo do tamanho amostral e a escolha do teste de hipótese são elementos importantes para o delineamento de uma pesquisa comparativa. O presente estudo utilizou o método de Monte Carlo para determinar a acurácia dos testes t de student e Man-Whitney Wilcoxon para populações com distribuição Normal e Qui-quadrado e diferentes desvios-padrão e tamanho do efeito. Os resultados indicam que quanto maior o desvio-padrão ou menor o tamanho do efeito, maior deve ser o tamanho amostral. O desempenho dos testes em indicar corretamente as diferenças foi parecido quando as amostras vieram de populações com distribuição normal, enquanto que para populações com distribuição qui-quadrado o teste de Mann-Whitney precisou de bem menos observações para atingir o mesmo poder que o teste t. Sugere-se que a escolha do teste seja feita com base em características da população, estimadas pela amostra.*

Palavras-chave: Teste de hipóteses, Teste t de student, Teste de Man-Whitney.

Abstract: *Calculation of the sample size and the choice of the hypothesis test are important elements for the comparative research design. The present study perform the Monte Carlo method to determine the accuracy of Student's t test and Man-Whitney Wilcoxon test for populations with Normal and Chi-square distribution, different standard deviations and effect size. The results indicate that the larger the standard deviation or the smaller the effect size, the larger the sample size should be. The performance of the tests in correctly indicating the differences was similar when the samples came from populations with normal distribution, whereas for populations with chi-square distribution the Mann-Whitney test required far fewer observations to reach the same power as the t test. It is suggested that the choice of the test be made based on the real characteristics of the population, estimated by the sample under analysis.*

Keywords: Hypothesis test, T test, Man-Whitney test.

1. Introdução

Um dos grandes desafios em várias áreas de pesquisa é estimar o tamanho amostral necessário e escolher o melhor teste de hipótese para identificar diferenças significativas entre as médias de duas ou mais populações. Em termos gerais, há um conhecimento difundido de que testes paramétricos são sempre mais poderosos do que os não-paramétricos, e de que a partir de 30 observações pode e deve-se optar por técnicas paramétricas. Tais afirmações podem ser facilmente encontradas em diversas apresentações de slides, manuais e livros-texto disponíveis na internet. Contudo, não é tarefa simples demonstrar a validade dessas suposições para problemas reais, e muitos

[1] Universidade de São Paulo e-mail: avk@usp.br

pesquisadores podem estar escolhendo determinados métodos, em detrimento de outros, para testar suas hipóteses com base no senso comum.

A grande aplicabilidade da inferência estatística é obter uma informação acurada de uma característica de uma população sem que se precise coletar informações de todos os seus indivíduos. Assim, utiliza-se um subconjunto de dados observados (amostra) para inferir características do conjunto maior (população). Dado um conjunto de observações independentes (amostra aleatória) de uma população cuja distribuição pertença a alguma família de distribuições, como a Normal ou a de Poisson, a inferência estatística paramétrica pode ser utilizada para testar hipóteses sobre, ou estimar, parâmetros desconhecidos, como a média e o intervalo de confiança. No entanto, conforme apontam Sprent e Smeeton (2007), a inferência paramétrica nem sempre é apropriada ou nem sempre é possível de ser realizada, acrescentando-se ao fato de que para a maior parte dos problemas estatísticos não importa se o método paramétrico ou o não paramétrico é apropriado, pois o que queremos deduzir depende de quais pressupostos podem ser feitos de forma válida.

Muitos delineamentos de pesquisa em psicologia, educação e ciências sociais implicam comparar dois grupos em relação a alguma característica mensurada. Neste cenário, uma das dúvidas que muitos pesquisadores se deparam no momento de testar suas hipóteses é qual teste escolher, considerando a desejabilidade de se obter maior poder em detectar diferenças e diminuir os riscos de falsos positivos. Para comparar duas amostras independentes, os pesquisadores têm recorrido principalmente a duas ferramentas estatísticas: o teste t de student e o teste de Mann-Whitney (ou Mann-Whitney WIlcoxon). Uma busca no Portal de Periódicos CAPES mostra que há 3.177 artigos mencionando o teste de Mann-Whitney e 5.691 mencionando o teste t de student, sugerindo uma preferência dos pesquisadores pela alternativa paramétrica. A pergunta que este artigo pretende responder é: essa preferência é justificada? Ou seja, o teste t é mais confiável que o teste de Mann-Whitney?

Sendo assim, o objetivo do presente estudo é estimar o poder dos testes t de student e Mann-Whitney Wilcoxon por meio do método de Monte Carlo na comparação de duas amostras vindas de populações com 10 milhões de elementos cada, utilizando três níveis de desvios-padrão (pequeno, médio e grande), quatro níveis de magnitude de diferença entre as populações (pequeno, médio, grande e muito grande) e duas distribuições de probabilidade (normal e qui-quadrado). As magnitudes das diferenças foram estimadas com base no coeficiente d de Cohen (Cohen, 1988).

2. Método

Foram geradas duas populações de 10 milhões de elementos. Em seguida, foram selecionadas aleatoriamente mil amostras de tamanho n, com n variando de 5 a k, sendo k o número em que a porcentagem de acertos do teste superasse 95% em pelo menos dois de três números consecutivos (optou-se por este método para reduzir ainda mais a probabilidade do teste atingir 95% de acertos ao acaso). Quando as porcentagens de acerto dos dois testes atingiram os 95% de acertos, o k em questão foi registrado como o número amostral ótimo. Esse procedimento foi executado para dois tipos de distribuição (normal e qui-quadrado), três desvios-padrão diferentes (pequeno, médio e grande) e quatro tamanhos de diferença (pequeno, médio, grande e muito grande) entre as verdadeiras médias populacionais. A tabela 1 mostra as características das populações em cada uma das 24 (duas distribuições vezes três desvios-padrão vezes quatro tamanhos de efeito) situações comparadas.

O algoritmo utilizado para gerar as populações, estimar a magnitude da diferença e selecionar aleatoriamente as mil amostras e compará-las foi desenvolvido pelo autor do presente estudo na linguagem R e está disponível no endereço do GitHub: https://github.com/andrevk/SBBHQK/.

Tabela 1: *Características das duas populações comparadas*

	Distribuição Normal				
	Pop. 1			**Pop. 2**	
	M	**DP**		**M**	**DP**
Desvio pequeno	0	1	vs.	0,2	1
Desvio Médio	0	2	vs.	0,5	2
Desvio Grande	0	3	vs.	0,8	3
	Distribuição Qui-quadrado (npc = 4)				
	Pop. 1			**Pop. 2**	
	M	**DP**		**M**	**DP**
Desvio pequeno	5	2	vs.	6	2
Desvio Médio	5	3	vs.	7	3
Desvio Grande	5	4	vs.	8	4

Nota: cada linha de comparação foi repetida quatro vezes, uma para cada nível de tamanho do efeito (pequeno = 0,2; médio = 0,5; grande = 0,8; e muito grande = 1,4;), totalizando 24 comparações.

3. Resultados

A figura 1 mostra a porcentagem de acertos dos testes *t de student* (linha contínua) e *Man-Whtiney Wilcoxon* (linha pontilhada) na comparação das médias de duas populações cujas variáveis são normalmente distribuídas. Nos casos em que o desvio-

padrão mantém-se constante (linhas verde e azul), nota-se que, quando o tamanho do efeito entre as médias é grande (0,8), são necessárias 41 observações (verde) para obter um índice de 95% de acertos, mas quando o tamanho do efeito é pequeno (0,2), são necessárias 700 observações (azul) para obter o mesmo índice. Nos casos em que o tamanho do efeito mantém-se constante (linhas azul e roxo), nota-se que, quando o desvio-padrão da população é grande (DP = 3), são preciso 2.000 observações (roxo) para obter 95% de acertos. Observa-se também que as porcentagens de acertos do teste t são sempre ligeiramente maiores que as do teste de *Man-Whitney Wilcoxon*, independentemente do tamanho da amostra.

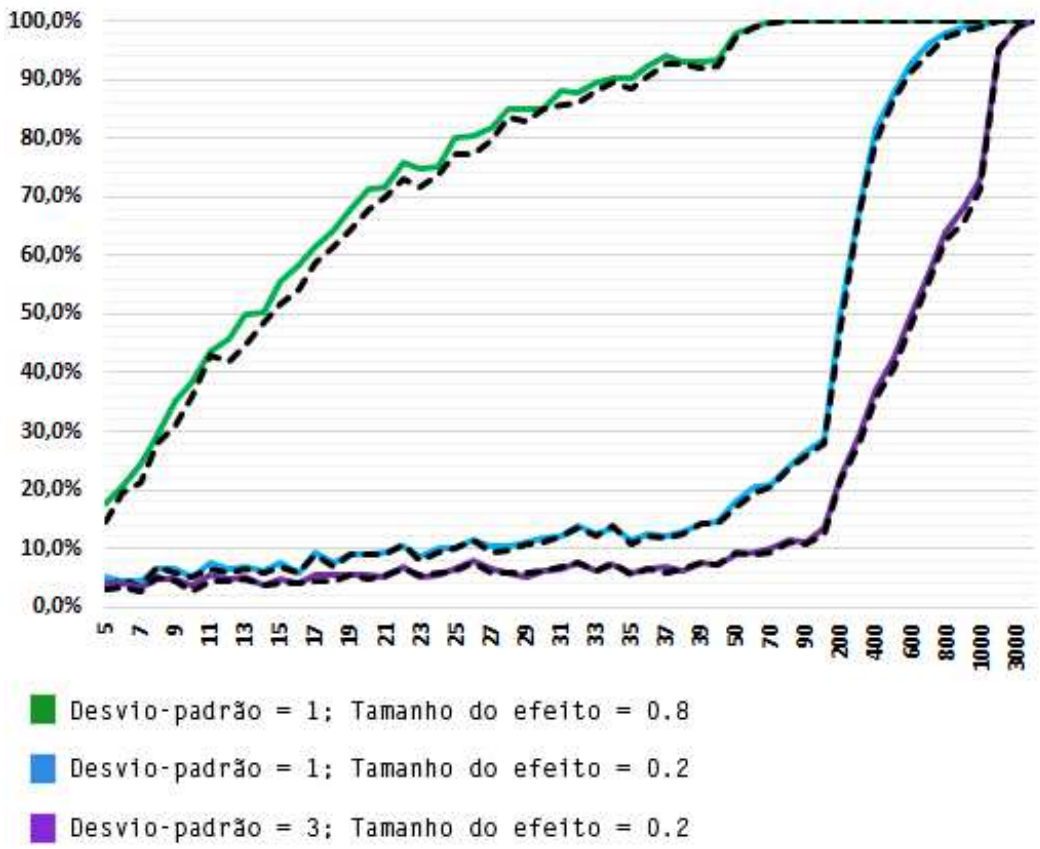

Figura 1: Aumento da amostra e porcentagem de acertos dos testes para diferentes desvios-padrão e tamanho do efeito (distribuição normal).

A figura 2 mostra as porcentagens de acerto dos testes *t de student* (linha contínua) e *Man-Whitney Wilcoxon* (linha pontilhada) na comparação das médias de duas populações cujas variáveis possuem distribuição qui-quadrado, com diferentes graus de liberdade e tamanho do efeito entre as médias. Os resultados seguem o mesmo padrão da figura 1, na qual populações com maior dispersão e menor tamanho do efeito precisam de maior quantidade de observações até atingir a taxa de 95% de acertos. No

entanto, para as distribuições qui-quadrado aqui apresentadas, o teste não paramétrico *Man-Whitney Wilcoxon* apresentou melhor desempenho que o teste *t de student*, independentemente do tamanho da amostra.

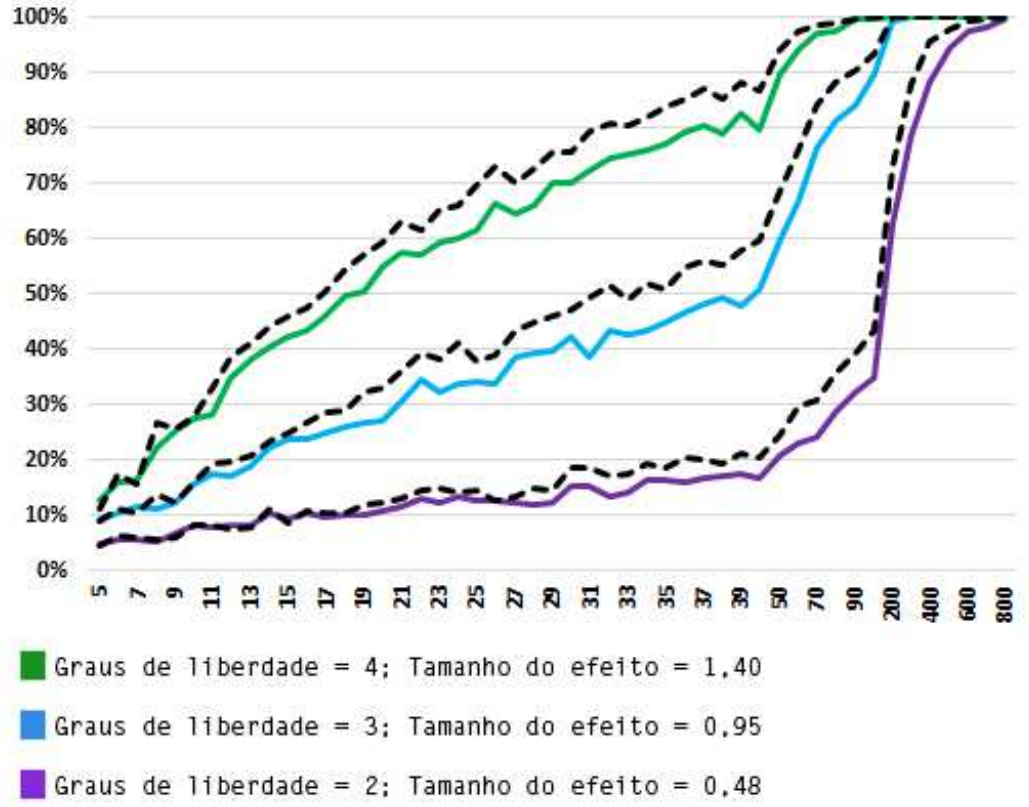

Figura 2. Aumento da amostra e porcentagem de acertos dos testes para diferentes desvios-padrão e tamanho do efeito (distribuições qui-quadrado).

A tabela 1 mostra o tamanho da amostra necessária para que os testes paramétrico e não paramétrico atinjam 95% de acertos, variando a distribuição, o desvio-padrão e o tamanho do efeito da diferença entre as médias populacionais. Nas comparações entre as populações cujas características estão normalmente distribuídas, o *teste t* necessita de ligeiramente menos observações do que o teste de *Man-Whitney*, independentemente do desvio-padrão ou da magnitude da diferença. Inversamente, para a distribuição qui-quadrado, o teste de *Man-Whitney* precisou de menos observações que o *teste t*, sendo que em alguns casos o teste paramétrico precisou de quase o dobro de observações, diferenças maiores do que as observados na distribuição normal.

Tabela 2: *Tamanho amostral necessário para os testes t e Man-Whitney acertarem 95% das comparações entre duas populações com 10.000.000 elementos*

Tamanho do efeito:	**Pequeno (d = 0.2)**		**Médio (d = 0.5)**		**Grande (d = 0.8)**		**Muito Grande (d = 1.4)**	
	t	**w**	**t**	**w**	**t**	**w**	**t**	**w**
Distribuição Normal								
DP = 1,0	660	666	106	108	41	43	15	16
DP = 2,0	1295	1314	209	219	82	82	28	28
DP = 3,0	2044	2044	485	486	120	129	41	42
Distribuição Qui-quadrado								
2.43 > DP > 1	>930*	523	256	145	71	45	27	19
3.5 > DP > 2	>930*	803	275	151	111	67	40	27
4 > DP > 3	>2010*	1198	350	212	135	97	50	39

Nota: t = teste *t de student*; w = teste *Man-Whitney Wilcoxon*
* Para diferenças de magnitude pequena, o teste t mostrou-se irregular para atingir 95% de acertos na distribuição qui-quadrado. A partir dos números assinalados, o teste t possui índice de acerto próximo aos 95%.

6. Considerações finais

O presente estudo demonstrou, pelo método de Monte Carlo, que quanto maior o desvio-padrão da população ou menor a magnitude da diferença entre as médias de duas populações, maior deverá ser o tamanho da amostra para obter índices aceitáveis de confiabilidade. Isso significa que não há um tamanho amostral fixo (como a regra de ouro de 30 observações) para garantir que o resultado do teste seja confiável. É necessário, portanto, estimar a variância da população e o tamanho do efeito esperado para que o cálculo amostral seja feito com precisão, como sugere a fórmula discutida por Charan e Biswas (2013). Na prática, uma solução razoável para essa estimativa é considerar as variâncias e o tamanho do efeito das próprias amostras colhidas.

Também pode-se observar que a diferença entre a eficácia dos testes *t de student* e *Man-Whitney Wilcoxon* independe do tamanho da amostra, mas, sim, da verdadeira distribuição da população original. Sendo assim, o teste não paramétrico deve ser preferido quando a distribuição é cortada em um dos lados, como por exemplo não se pode obter valores menores que zero e não há limite (ou o limite é grande) para valores positivos, ou quando, por quaisquer motivos, a mediana é a medida que melhor representa a característica da população.

Com as configurações – tamanho do feito, tamanho do desvio-padrão e distribuição – utilizadas no presente estudo, constata-se que o fator mais importante

para identificar diferenças significativas em níveis confiáveis é a real diferença entre as médias da população (tamanho do efeito). Na distribuição normal, um tamanho de efeito pequeno necessita-se de pelo menos 660 observações, enquanto que para um efeito muito grande bastam 16. Essa grande diferença de observações necessárias também se observa na distribuição qui-quadrado. Sabe-se que, em muitas áreas de pesquisa, coletar centenas de observações pode ser bastante custoso ou demorado. Sendo assim, os pesquisadores devem estar cientes de que, para detectar uma diferença que na realidade é pequena, é necessário obter uma amostra numerosa, e que, nos casos em se constatam diferenças significativas utilizando amostras pequenas, a chance de falso positivo pode ser grande se a real diferença for de fato pequena.

Por fim, indica-se que a escolha do teste de hipótese a ser utilizado deve ser feita com base nas verdadeiras características da população, que podem ser estimadas pela própria amostra em análise. E um recurso que pode ser empregado para projetar o tamanho amostral ou definir o melhor teste para a comparação em questão é o algoritmo disponibilizado por este estudo, que pode ser usado para simular outras combinações de desvios-padrão, tamanho de efeito e família da distribuição além das apresentadas aqui.

Referências

[1] SPRENT, P.; SMEETON, N. **Applied nonparametric statistical methods**. Boca Raton (Fla.): Chapman & Hall/CRC. 2007.

[2] COHEN, J. **Statistical power analysis for the behavioral sciences**. New York: Psychology Press. 1998.

[3] R Core Team. R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2017

[4] CHARAN, J.; BISWAS, T. (2013). **How to calculate sample size for different study designs in medical research?** Indian Journal of Psychological Medicine, 35(2), 121. doi:10.4103/0253-7176.116232.