Teste t para 2 amostras

Visão geral

Um teste t para 2 amostras pode ser usado para comparar se dois grupos independentes diferem. Esse teste é derivado sob as suposições de que ambas as populações são normalmente distribuídas e têm variâncias iguais. Embora a suposição de normalidade não seja crítica (Pearson, 1931; Barlett, 1935; Geary, 1947), a suposição de variâncias iguais é crítica se os tamanhos amostrais são muito diferentes (Welch, 1937; Horsnell, 1953).

Alguns praticantes primeiro realizam um teste preliminar para avaliar a igualdade de variâncias antes de realizar o procedimento t clássico para 2 amostras. Esta abordagem, contudo, tem graves desvantagens, devido a esses testes de variância estarem sujeitos a suposições e limitações importantes. Por exemplo, vários testes de variâncias iguais, tais como o clássico teste F, são sensíveis a desvios de normalidade. Outros testes, que não contam com a suposição de normalidade, como Levene/Brown-Forsythe, têm baixo poder para detectar uma diferença entre variâncias.

B.L. Welch desenvolveu um método de aproximação para comparar as médias de duas populações normais independentes, quando suas variâncias não são necessariamente iguais (Welch, 1947). Como o teste t modificado de Welch não é derivado sob a suposição de variâncias iguais, ele permite aos usuários comparar as médias de duas populações sem primeiro ter que testar a igualdade das variâncias.

Neste documento, comparamos o método t modificado de Welch com o procedimento t clássico para 2 amostras e determinamos qual procedimento é o mais confiável. Descrevemos também as seguintes verificações nos dados que são automaticamente realizadas e exibidas no Assistente de Cartão de Relatórios e explicamos como elas afetam os resultados da análise:

- Normalidade
- Dados atípicos
- Tamanho amostral

Método de teste t para 2 amostras

Teste t clássico para 2 amostras versus teste t de Welch

Se os dados são provenientes de duas populações normais com as mesmas variâncias, o teste t clássico para 2 amostras é tão ou mais poderoso que o teste t de Welch. A suposição de normalidade não é crítica para o procedimento clássico (Pearson, 1931; Barlett, 1935; Geary, 1947), mas a suposição de igualdade de variâncias é importante para assegurar resultados válidos. Mais especificamente, o procedimento clássico é sensível à suposição de igualdade de variâncias iguais quando os tamanhos amostrais diferem independentemente de quão grandes são as amostras (Welch, 1937; Horsnell, 1953). Na prática, porém, a suposição de igualdade de variâncias raramente é verdadeira, o que pode levar a taxas de erro Tipo I mais elevadas. Portanto, se o teste t clássico para 2 amostras é usado quando duas amostras têm variâncias diferentes, o teste é mais propenso a produzir resultados incorretos.

O teste t de Welch é uma alternativa viável para o teste t clássico porque não assume igualdade de variâncias e, por conseguinte, é insensível a variâncias diferentes para todos os tamanhos amostrais. No entanto, o teste t de Welch é baseado em uma aproximação e o seu desempenho em amostras de pequenas dimensões pode ser questionável. Quisemos determinar se o teste t de Welch, ou o teste t clássico para 2 amostras é o teste mais confiável e prático para utilizar no Assistente.

Objetivo

Quisemos determinar, através de estudos de simulação e derivações teóricas, se o teste t de Welch ou o teste t clássico para 2 amostras é mais confiável. Mais especificamente, desejamos examinar:

- As taxas de erro Tipo I e Tipo II de ambos o teste t clássico para 2 amostras e do teste t de Welch em diversos tamanhos de amostras, quando os dados são normalmente distribuídos e as variâncias são iguais.
- As taxas de erro Tipo I e Tipo II do teste t de Welch para experimentos não balanceados e com variâncias diferentes para os quais o teste t clássico para 2 amostras falha.

Método

Nossas simulações enfocam três áreas:

- Foram comparados os resultados das simulações do teste t clássico para 2 amostras e do teste t de Welch sob diversas suposições do modelo, incluindo a normalidade, não normalidade, variâncias iguais, variâncias não iguais, experimentos balanceados e não balanceados. Para obter mais detalhes, consulte o Apêndice A.

- Derivamos a função poder para o teste t de Welch e a comparamos com a função poder do teste t clássico para 2 amostras. Para obter mais detalhes, consulte o Apêndice B.
- Estudamos os efeitos da não normalidade na função poder teórica do teste t de Welch.

Resultados

Quando as suposições do modelo t clássico para 2 amostras são mantidas, o teste t de Welch funciona tão bem ou quase tão bem quanto o teste t clássico para 2 amostras, exceto para experimentos pequenos não balanceados. No entanto, o teste t clássico para 2 amostras também pode ter um desempenho pobre quando experimentos são pequenos e não balanceados, devido à sua sensibilidade à suposição de igual variância. Além disso, em configurações práticas, é difícil estabelecer que duas populações têm exatamente a mesma variância. Portanto, a superioridade teórica do teste clássico para 2 amostras sobre o teste t de Welch tem pouco ou nenhum valor prático. Por esse motivo, o Assistente usa o teste t de Welch para comparar as médias de duas populações. Para os resultados detalhados da simulação, consulte os Apêndices A, B e C.

Verificações de dados

Normalidade

O teste t de Welch, o método usado no Assistente para comparar as médias de duas populações independentes, é derivado sob a suposição de que as populações são normalmente distribuídas. Felizmente, mesmo quando os dados não são normalmente distribuídos, o teste t de Welch funciona bem se as amostras forem grandes o bastante.

Objetivo

Queríamos determinar o quão próximos os níveis simulados de significância para o método Welch e o teste t clássico para 2 amostras correspondiam ao nível alvo de significância (taxa de erro Tipo I) de 0,05.

Método

Realizamos simulações do teste t de Welch e o teste t clássico para 2 amostras em 10.000 pares de amostras independentes gerados de populações normais, assimétricas e contaminadas (variâncias iguais e diferentes). As amostras foram de diversos tamanhos. A população normal serve como uma população de controle para fins de comparação. Para cada condição, calculamos os níveis de significância simulada e os comparamos com o alvo ou nível de significância nominal de 0,05. Se o teste tiver bom desempenho, os níveis de significância simulados deve estar próximos de 0,05.

Resultados

Para amostras moderadas ou grandes, o teste t de Welch mantém suas taxas de erro Tipo I para dados normais e também não normais. Os níveis de significância simulados estão perto do nível de significância alvo quando ambos os tamanhos de amostras são de 15, no mínimo. Consulte o Apêndice A para obter mais detalhes.

Como o teste apresenta bom desempenho com amostras relativamente pequenas, o Assistente não testa a normalidade dos dados. Em vez disso, ele verifica o tamanho das amostras e exibe os seguintes indicadores de status no Cartão de Relatórios.

Status	Condição
	Ambos os tamanhos amostrais são, no mínimo, de 15; a normalidade não é um problema.
	No mínimo um dos tamanhos amostrais < 15; a normalidade pode ser um problema.

Dados atípicos

Dados atípicos são valores de dados extremamente grandes ou pequenos, também conhecidos como outliers. Dados atípicos podem ter uma forte influência nos resultados da

análise. Quando a amostra é pequena, eles podem afetar as chances de encontrar resultados estatisticamente significativos. Dados atípicos podem indicar problemas com a coleta de dados ou comportamento incomum de um processo. Portanto, muitas vezes, vale a pena investigar esses pontos de dados e eles devem ser corrigidos quando possível.

Objetivo

Queríamos desenvolver um método para verificar valores de dados que são muito grandes ou muito pequenos, em relação à amostra geral, e que podem afetar os resultados da análise

Método

Desenvolvemos um método para verificar os dados atípicos com base no método descrito por Hoaglin, Iglewicz e Tukey (1986) para identificar outliers em boxplots.

Resultados

O Assistente identifica um ponto de dados tão incomum se ele estiver mais de 1,5 vezes a amplitude interquartílica além do quartil inferior ou superior da distribuição. Os quartis inferior e super são os percentis 25º e 75º dos dados. O amplitude interquartílica é a diferença entre os dois quartis. Esse método funciona bem mesmo quando há vários outliers, porque ele possibilita a detecção de cada outlier específico.

Os outliers tendem a ter uma influência na função poder somente quando os tamanhos amostrais são muito pequenos. Em geral, quando outliers estão presentes, os valores de poder observados tendem a ser um pouco mais altos do que os valores de poder teóricos alvo. Esse padrão pode ser visto na Figura 10 no Apêndice C onde as curvas de poder teóricas e simuladas não estão razoavelmente próximas até que o tamanho amostral mínimo alcance 15.

Ao verificar dados atípicos, o Cartão de Relatórios do Assistente do teste t para 2 amostras exibe os seguintes indicadores de status:

Status	**Condição**
	Não há pontos de dados atípicos.
	Pelo menos um ponto de dados é incomum e pode afetar os resultados do teste.

Tamanho amostral

Tipicamente, um teste de hipóteses é realizado para coletar evidências para rejeitar a hipótese nula de “nenhuma diferença”. Se as amostras forem muito pequenas, o poder do teste pode não ser adequando para detectar uma diferença entre as médias, quando esta realmente existir, o que resulta em um erro Tipo II. Portanto, é crucial assegurar que os

tamanhos amostrais sejam suficientemente grandes para detectar diferenças importantes na prática, com alta probabilidade.

Objetivo

Se os dados atuais não fornecem evidência suficiente contra a hipótese nula, queremos determinar se os tamanhos amostrais são grandes o suficiente para o teste detectar diferenças práticas de interesse com alta probabilidade. Apesar do objetivo do planejamento do tamanho amostral ser assegurar que as amostras sejam grandes o suficiente para detectar diferenças importantes com alta probabilidade, as amostras não devem ser tão grandes que diferenças inexpressivas tornem-se estatisticamente significativas com alta probabilidade.

Método

A análise de poder e de tamanho amostral é baseada na função poder teórico do teste específico que é usado para realizar a análise estatística. Para o teste t de Welch, esta função poder depende dos tamanhos amostrais, da diferença entre entre as médias das duas populações e e das variâncias verdadeiras das duas populações. Para obter mais detalhes, consulte o Apêndice B.

Resultados

Quando os dados não fornecem evidência suficiente contra a hipótese nula, o Assistente calcula diferenças práticas que podem ser detectadas com uma probabilidade de 80% e de 90% para os tamanhos amostrais dados. Além disso, se o usuário fornecer uma diferença prática particular de interesse, o Assistente calcula tamanhos amostrais que produzem uma chance de 80% e de 90% de detecção da diferença.

Não há resultado geral para relatar porque os resultados dependem das amostras específicas do usuário. Contudo, você pode consultar os Apêndices B e C para obter mais informações sobre a função do poder do teste de Welch.

Ao verificar o poder e o tamanho amostral, o Cartão de Relatórios do Assistente do teste t para 2 amostras exibe os seguintes indicadores de status:

Status	Condição
	O teste encontra uma diferença entre as médias, portanto, o poder não é um problema. OU O poder é suficiente. O teste não encontrou uma diferença entre as médias, mas a amostra é grande o suficiente para fornecer pelo menos uma chance de 90% de detecção da diferença dada.
	O poder pode ser suficiente. O teste não encontrou uma diferença entre as médias, mas a amostra é grande o suficiente para fornecer uma chance de 80% a 90% de detecção da diferença dada. O tamanho da amostra necessário para alcançar o poder de 90% é indicado.

Status	Condição
	O poder pode não ser suficiente. O teste não encontrou uma diferença entre as médias, e a amostra é grande o suficiente para fornecer uma chance de 60% a 80% de detecção da diferença dada. Os tamanhos amostrais necessários para alcançar o poder de 80% e o poder de 90% estão indicados.
	O poder não é suficiente. O teste não encontrou uma diferença entre as médias, e a amostra não é grande o suficiente para fornecer pelo menos uma chance de 60% de detecção da diferença dada. Os tamanhos amostrais necessários para alcançar o poder de 80% e o poder de 90% estão indicados.
	O teste não encontrou uma diferença entre as médias. Você não especificou uma diferença prática entre as médias para detecção; portanto, o relatório indica as diferenças que você pôde detectar com uma chance de 80% e de 90%, com base no seu alfa, desvios padrão e tamanhos amostrais.

Referências

Arnold, S. F. (1990). *Mathematical Statistics*. Englewood Cliffs, NJ*:* Prentice Hall, Inc.

Aspin, A. A. (1949). Tables for Use in Comparisons whose Accuracy Involves Two Variances, Separately Estimated, *Biometrika, 36*, *290-296*.

Bartlett, M. S. (1935). The effect of non-normality on the t-distribution. *Proceedings of the Cambridge Philosophical Society*, *31*, *223-231*.

Box, G. E. P. (1953). Non-normality and Tests on Variances, *Biometrika, 40*, *318-335*.

Geary, R. C. (1947). Testing for Normality, *Biometrika, 34*, *209-242*.

Hoaglin, D. C., Iglewicz, B. e Tukey, J. W. (1986). Performance of Some Resistant Rules for Outlier Labeling. *Journal of the American Statistical Association*, *81*, *991-999*.

Horsnell, G. (1953). The effect of unequal group variances on the F test for homogeneity of group means. *Biometrika, 40*, *128-136*.

James, G. S. (1951). The comparison of several groups of observations when the ratios of the populations variances are unknown, *Biometrika, 38*, *324-329*.

Kulinskaya, E. Staudte, R. G. e Gao, H. (2003). Power Approximations in Testing for unequal Means in a One-Way Anova Weighted for Unequal Variances, *Communication in Statistics, 32(12), 2353-2371*.

Lehmann, E. L. (1959). *Testing statistical hypotheses*. New York, NY: Wiley.

Neyman, J., Iwaszkiewicz, K. & Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation, *Journal of the Royal Statistical Society, Series B, 2, 107-180*.

Pearson, E. S. (1931). The Analysis of variance in case of non-normal variation, *Biometrika, 23*, *114-133*.

Pearson, E.S. & Hartley, H.O. (Eds.). (1954). *Biometrika Tables for Statisticians*, Vol. I. London: Cambridge University Press.

Srivastava, A. B. L. (1958). Effect of non-normality on the power function of t-test, *Biometrika*, *45, 421-429*.

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika, 38*, *330-336*.

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika, 34*, *28-35*.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal, *Biometrika, 29*, *350-362*.

Wolfram, S. (1999). *The Mathematica Book (*4th ed.). Champaign, IL: Wolfram Media/Cambridge University Press.

Apêndice A: impacto da não normalidade e da heterogeneidade no teste t clássico para 2 amostras e no teste t de Welch

Realizamos vários estudos de simulação planejados para comparar o teste t clássico para 2 amostras e o teste t de Welch sob diferentes suposições do modelo.

Estudo de simulação A

Realizamos o estudo em três partes:

- Na primeira parte do estudo, exploramos a sensibilidade do teste t clássico para 2 amostras e o teste t de Welch para a suposição de variância igual quando a suposição de normalidade é verdadeira. Duas amostras foram geradas a partir de duas populações normais independentes. A primeira amostra, a amostra de base, foi gerada a partir de uma população normal com média 0 e desvio padrão $\sigma_1 = 2$, $N(0, 2)$. A segunda amostra também foi gerada a partir de uma população normal com média 0, mas com o desvio padrão σ_2 escolhido de forma que o $\rho = \sigma_2/\sigma_1$ 0,5, 1,0, 1,5 e 2. Em outras palavras, a segunda amostra foi gerada a partir das populações $N(0, 1)$, $N(0, 2)$, $N(0, 3)$ e $N(0, 4)$ respectivamente. Além disso, o tamanho da amostra de base em cada caso foi fixada em $n_1 = 5, 10, 15, 20$ e para cada n_1 dado, o tamanho da segunda amostra, n_2 foi escolhido de forma que a razão dos tamanhos das amostras, $r = n_2/n_1$ foi aproximadamente igual a 0,5, 1, 1,5 e 2,0.

  Para cada um desses experimentos de 2 amostras, geramos 10.000 pares de amostras independentes a partir das respectivas populações. Em seguida, foi realizado o teste t clássico para 2 amostras e o teste t de Welch em cada um dos 10.000 pares de amostras para testar a hipótese nula de nenhuma diferença entre as médias. Como a diferença verdadeira entre as médias é nula, a fração das 10.000 réplicas para as quais a hipótese nula é rejeitada, representa o nível simulado de significância do teste. Uma vez que o nível de significância alvo de cada um dos testes é $\alpha = 0{,}05$, o erro de simulação associado com cada teste e cada experimento está em torno de 0,2%.

- Na segunda parte, investigamos o impacto da não normalidade, especificamente a assimetria, nos níveis de significância simulados dos dois testes. Esta simulação foi criada da mesma forma que a simulação anterior, exceto que a amostra de base foi gerada a partir da distribuição qui-quadrado com 2 graus de liberdade, Qui (2) e as segundas amostras foram geradas de outras distribuições qui-quadrado de forma que $\rho = \sigma_2/\sigma_1$ assume os valores 0,5, 1,0, 1,5 e 2. A diferença hipotética entre as médias foi definida para ser a verdadeira diferença entre as médias de populações pai.

- Na terceira parte, analisamos o efeito de outliers sobre o desempenho dos dois testes t. Por este motivo, as duas amostras foram geradas a partir de distribuições normais contaminadas. A população normal contaminada $\mathrm{CN}(\mathrm{p},\sigma)$ é uma mistura de duas populações normais: a população $\mathrm{N}(0,1)$ e a população $\mathrm{N}(0,\sigma)$ normal. Definimos a distribuição normal contamidada como:

$$CN(p,\sigma) = pN(0,1) + (1-p)N(0,\sigma)$$

onde p é o parâmetro de mistura e $1-\mathrm{p}$ é a proporção de contaminação ou proporção de outliers. É fácil mostrar que se X é distribuído como $CN(p,\sigma)$ então sua média é $\mu_X = 0$ e seu desvio padrão é $\sigma_X = \sqrt{p+(1-p)\sigma^2}$.

A amostra de base foi gerada a partir de $CN(0{,}8,4)$ e a segunda amostra foi gerada a partir de $CN(0{,}8,\sigma)$ normal contaminada. O parâmetro σ foi escolhido para que a razão dos desvios padrão das duas populações (contaminadas) $\rho = \sigma_2/\sigma_1$ seja igual a 0,5, 1,0, 1,5 e 2, exatamente como nas partes I e II. Devido a $\sigma_1 = \sqrt{0{,}8 + (1-0{,}8)*16} = 2{,}0$, isso resulta na escolha de $\sigma = 1,4,\ 6{,}40, 8{,}72$, respectivamente Em outras palavras, as segundas amostras foram geradas a partir de $CN(0{,}8,1)$, $CN(0{,}8,4)$, $CN(0{,}8,6{,}4)$ e $CN(0{,}8,8{,}72)$. A seguir, realizamos as simulações conforme descrito na Parte I.

Os resultados do estudo estão organizados na Tabela 1, e exibidos nas Figuras 1, 2, e 3.

Resultados e resumo

Em geral, os resultados da simulação apoiam os resultados teóricos em que, sob a suposição de normalidade e variâncias iguais, o teste t clássico para 2 amostras produz níveis de significância que são próximos do nível alvo, mesmo quando os tamanhos amostrais são pequenos. A segunda coluna de gráficos na Figura 1 exibe os níveis de significância simulados nos experimentos nos quais as variâncias das duas populações normais são iguais. As curvas dos níveis de significância simulados com base no teste t clássico para 2 amostras são indistinguíveis das linhas de nível alvo.

As tabelas a seguir mostram os níveis de significância simulados de dois testes bilaterais para ambos o teste t para 2 amostras e o teste t de Welch, cada um com $\alpha = 0{,}05$ baseado em pares de amostras geradas a partir da população normal, populações assimétricas (qui-quadrado) e de populações normais contaminadas. Os pares de amostras são da mesma família de distribuição, mas as variâncias das populações pai respectivas não são necessariamente iguais.

Tabela 1 Níveis de significância simulados de testes bilaterais (teste t clássico para 2 amostras e teste t de Welch, cada um com $\alpha = 0{,}05$) para n = 5.

			Pop. base.: N(0, 2) 2a. pop: N(0, σ_2)				População de base: Qui(2) 2a. pop.: Qui-quadrado				Pop. base: CN(0,8, 4) 2a. pop.: CN(0,8, σ)			
		$\frac{\sigma_2}{\sigma_1}$	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0
n_2	$\frac{n_2}{n_1}$	Mét.	$n_1 = 5$				$n_1 = 5$				$n_1 = 5$			
3	.6	2T	0,035	0,050	0,079	0,105	0,058	0,042	0,078	0,113	0,031	0,036	0,035	0,034
		Welch	0,035	0,039	0,049	0,055	0,048	0,029	0,055	0,063	0,029	0,024	0,021	0,020
5	1,0	2T	0,061	0,052	0,054	0,058	0,086	0,036	0,054	0,064	0,035	0,031	0,025	0,023
		Welch	0,048	0,042	0,044	0,047	0,066	0,021	0,040	0,050	0,027	0,023	0,018	0,016
8	1,6	2T	0,096	0,048	0,033	0,027	0,133	0,041	0,033	0,032	0,059	0,037	0,029	0,024
		Welch	0,050	0,045	0,043	0,042	0,094	0,034	0,032	0,041	0,034	0,029	0,026	0,022
10	2,0	2T	0,118	0,055	0,034	0,025	0,139	0,041	0,028	0,024	0,073	0,041	0,028	0,023
		Welch	0,052	0,051	0,050	0,051	0,097	0,041	0,033	0,042	0,035	0,032	0,028	0,025

Tabela 2 Níveis de significância simulados de teste bilaterais (teste t clássico para 2 amostras e teste t de Welch, cada um com $\alpha = 0{,}05$) para n = 10

			Pop. base.: N(0, 2) 2a. pop: N(0, σ_2)				População de base: Qui(2) 2a. pop.: Qui-quadrado				Pop. base: CN(0,8, 4) 2a. pop.: CN(0,8, σ)			
		$\frac{\sigma_2}{\sigma_1}$	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0
n_2	$\frac{n_2}{n_1}$	Met.	$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
5	0,5	2T	0,020	0,050	0,081	0,112	0,039	0,044	0,091	0,123	0,021	0,035	0,045	0,047
		Welch	0,046	0,048	0,050	0,050	0,043	0,047	0,067	0,063	0,034	0,028	0,022	0,019
10	1,0	2T	0,057	0,051	0,053	0,055	0,068	0,044	0,053	0,054	0,043	0,042	0,037	0,032
		Welch	0,051	0,049	0,049	0,049	0,062	0,037	0,046	0,049	0,039	0,038	0,032	0,027
15	1,5	2T	0,088	0,048	0,034	0,029	0,100	0,043	0,032	0,032	0,064	0,040	0,028	0,021
		Welch	0,050	0,048	0,047	0,048	0,074	0,044	0,041	0,046	0,035	0,037	0,035	0,031
20	2	2T	0,110	0,048	0,026	0,019	0,133	0,042	0,026	0,022	0,093	0,046	0,029	0,019
		Welch	0,048	0,047	0,045	0,046	0,083	0,050	0,044	0,049	0,036	0,039	0,040	0,038

Tabela 3 Níveis de significância simulados de teste bilaterais (teste t clássico para 2 amostras e teste t de Welch cada um com $\alpha = 0{,}05$) para n = 15

			Pop. base.: N(0, 2) 2a. pop: N(0, σ_2)				População de base: Qui(2) 2a. pop.: Qui-quadrado				Pop. base: CN(0,8, 4) 2a. pop.: CN(0,8, σ)			
		$\frac{\sigma_2}{\sigma_1}$	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0
n_2	$\frac{n_2}{n_1}$	Met.	$n_1 = 15$				$n_1 = 15$				$n_1 = 15$			
8	0,53	2T	0,021	0,050	0,083	0,110	0,036	0,041	0,089	0,114	0,022	0,044	0,056	0,062
		Welch	0,050	0,051	0,051	0,050	0,047	0,049	0,067	0,062	0,044	0,036	0,027	0,022
15	1,0	2T	0,049	0,047	0,050	0,053	0,064	0,046	0,051	0,061	0,045	0,045	0,041	0,037
		Welch	0,045	0,046	0,049	0,048	0,060	0,042	0,048	0,057	0,042	0,043	0,039	0,033
23	1,53	2T	0,081	0,049	0,033	0,028	0,103	0,042	0,036	0,030	0,075	0,048	0,033	0,024
		Welch	0,048	0,049	0,048	0,050	0,071	0,042	0,048	0,050	0,042	0,045	0,044	0,041
30	2,0	2T	0,111	0,050	0,028	0,018	0,123	0,049	0,027	0,020	0,100	0,046	0,025	0,016
		Welch	0,049	0,051	0,051	0,053	0,074	0,056	0,045	0,047	0,039	0,044	0,042	0,040

Tabela 4 Níveis de significância simulados de testes bilaterais (teste t clássico para 2 amostras teste t de Welch cada um com $\alpha = 0{,}05$) para n = 20

			Pop. base.: N(0, 2) 2a. pop: N(0, σ_2)				População de base: Qui(2) 2a. pop.: Qui-quadrado				Pop. base: CN(0,8, 4) 2a. pop.: CN(0,8, σ)			
		$\frac{\sigma_2}{\sigma_1}$	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0
n_2	$\frac{n_2}{n_1}$	Mét.	$n_1 = 20$				$n_1 = 20$				$n_1 = 20$			
10	0,5	2T	0,019	0,052	0,087	0,115	0,028	0,048	0,087	0,119	0,021	0,048	0,067	0,079
		Welch	0,050	0,054	0,053	0,053	0,044	0,054	0,061	0,061	0,048	0,042	0,035	0,028
20	1,0	2T	0,048	0,049	0,052	0,053	0,057	0,046	0,052	0,056	0,049	0,044	0,042	0,040
		Welch	0,045	0,049	0,051	0,050	0,055	0,044	0,050	0,052	0,047	0,042	0,040	0,037
30	1,5	2T	0,086	0,054	0,039	0,032	0,098	0,047	0,035	0,033	0,075	0,047	0,033	0,022
		Welch	0,054	0,054	0,053	0,052	0,068	0,047	0,051	0,053	0,041	0,043	0,044	0,042
40	2,0	2T	0,107	0,049	0,026	0,016	0,123	0,046	0,027	0,019	0,107	0,047	0,026	0,016
		Welch	0,048	0,049	0,046	0,047	0,070	0,054	0,046	0,045	0,044	0,043	0,043	0,042

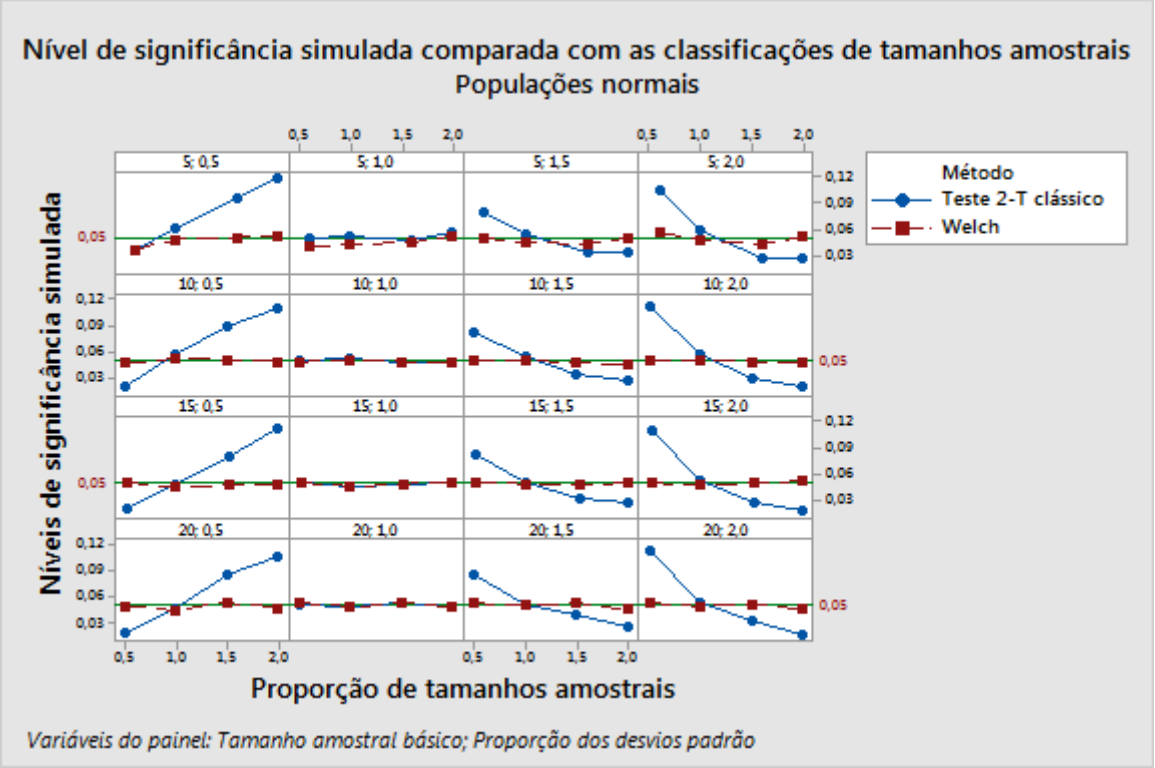

Figura 1 Níveis de significância simulados de testes bilaterais (teste t clássico para 2 amostras e teste t de Welch, cada um com $\alpha = 0{,}05$) baseados em pares de amostras geradas a partir de duas populações normais com variâncias iguais ou diferentes plotadas versus uma razão de tamanhos amostrais.

Os resultados da simulação mostram que para amostras relativamente pequenas o teste t clássico para 2 amostras é robusto contra não normalidade, mas é sensível à suposição de variâncias iguais, exceto se o experimento com duas amostras for quase balanceado. Isso é mostrado graficamente nas Figuras 1, 2 e 3. As curvas do nível de significância simulado, baseado no teste t clássico para 2 amostras cruzam a linha do nível alvo no ponto onde a razão dos tamanhos amostrais é 1,0, mesmo quando as variâncias são muito diferentes. Para todas as três famílias de distribuições (populações normal, qui-quadrado e normal contaminada), se os tamanhos amostrais forem diferentes, os níveis de significância simulados do teste t clássico para 2 amostras estarão próximos do nível alvo apenas quando as variâncias forem iguais. Isso está ilustrado na segunda coluna de gráficos em cada uma das figuras 1, 2 e 3.

O desempenho do teste t clássico é indesejável quando o experimento é não balanceado e as variâncias são diferentes. Mesmo pequenas disparidades entre as variâncias são problemáticas. Para aqueles experimentos não balanceados de variâncias diferentes, a normalidade dos dados não melhora os níveis de significância simulados. Na verdade, os níveis de significância simulados se encontram fora do nível alvo uma vez que os tamanhos amostrais aumentam independentemente da população pai. Quando a amostra maior é gerada a partir da população com uma variância maior, os níveis de significância simulados são menores do que o nível alvo. Quando as amostras maiores são geradas a partir da população com variância menor, os níveis simulados são maiores do que os níveis alvo.

Arnold (1990, página 372) faz um comentário similar ao examinar a distribuição assintótica da estatística do teste t clássico para 2 amostras sob a suposição de variâncias diferentes.

O teste t de Welch para 2 amostras, por outro lado, é insensível a desvios da suposição de igualdade de variâncias, conforme ilustrado nas Figuras 1, 2 e 3. Isso não é surpreendente uma vez que o teste t de Welch não é derivado sob a suposição de igualdade de variâncias. A suposição normal da qual o teste t de Welch é derivado parece ser importante somente quando o menor dos dois tamanhos amostrais é muito pequeno. Para amostras maiores, contudo, o teste torna-se robusto a desvios da suposição de normalidade. Isso está ilustrado nas Figuras 2 e 3, onde os níveis de significância simulados permanece consistentemente próximo do nível alvo, quando o tamanho mínimo das duas amostras é 15. Quando ambas as amostras são geradas a partir da distribuição qui-quadrado com 2 graus de liberdade e o tamanho de ambas as amostras é 15, o nível de significância simulado é 0,042 (consulte a Tabela 3).

Outliers também não parecem afetar o desempenho do teste t de Welch quando o tamanho mínimo das duas amostras é grande o suficiente. A Tabela 3 e a Figura 3 mostram que quando o tamanho mínimo das duas amostras é, no mínimo, 15, os níveis de significância simulados estão próximos do nível alvo (os níveis de significância simulados são 0,045, 0,045, 0,041, 0,037 quando a razão dos desvios padrão é 0,5, 1,0, 1,5 e 2,0 respectivamente).

Esses resultados mostram que para fins mais práticos, o teste t de Welch para 2 amostras tem melhor desempenho do que o teste t clássico para 2 amostras em termos de seus níveis de significância simulados ou taxa de erros do Tipo I.

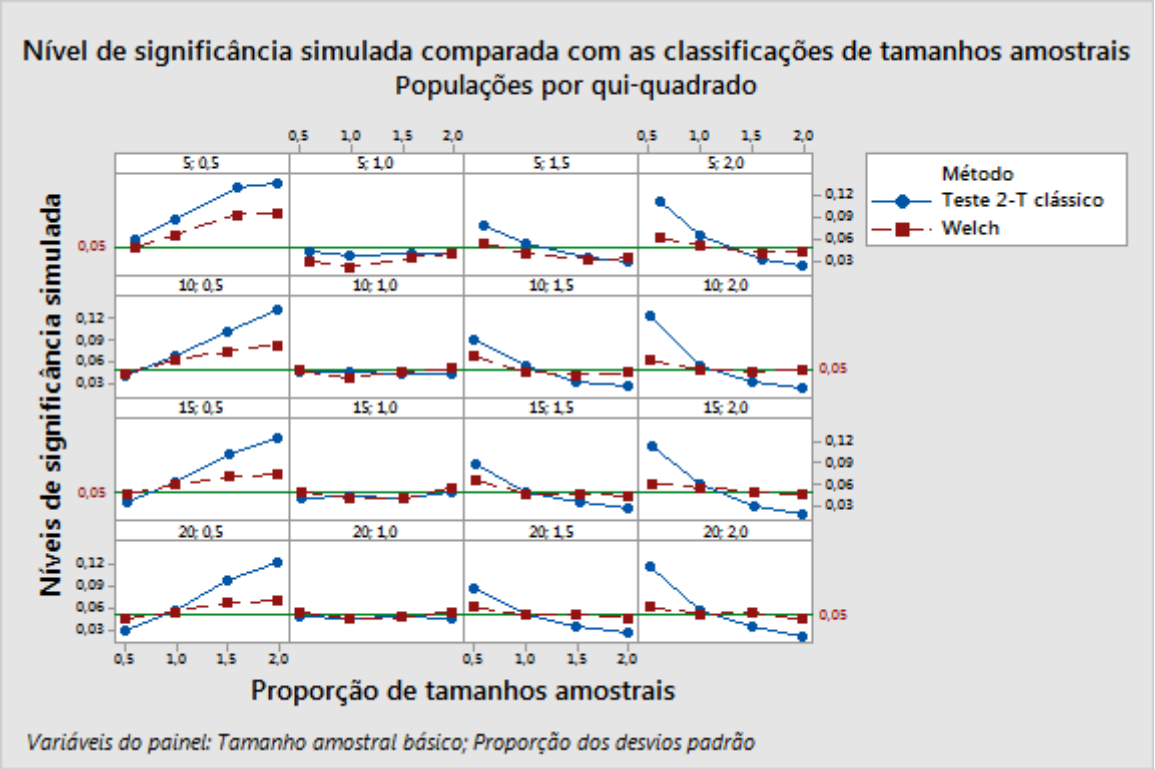

Figura 2 Níveis de significância simulados de testes bilaterais (teste t clássico para 2 amostras e teste t de Welch), baseados em pares de amostras geradas a partir de duas

populações normais com variâncias iguais ou diferentes plotadas versus a razão de tamanhos amostrais.

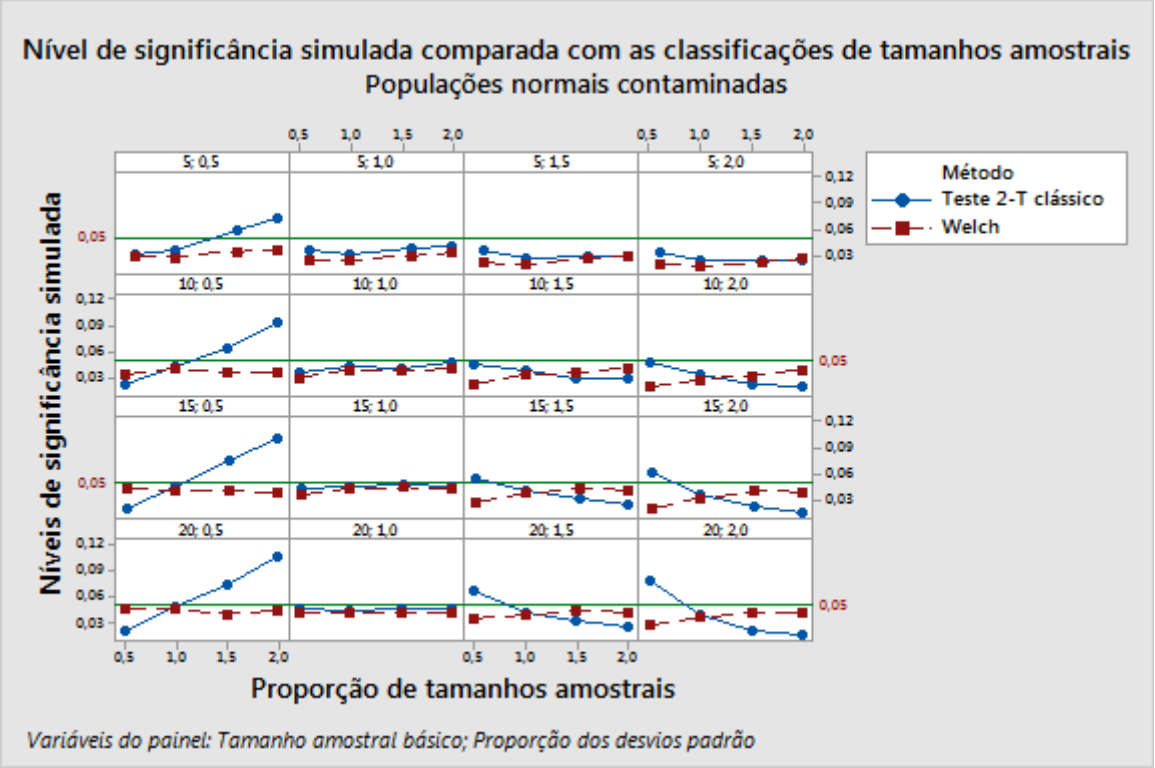

Figura 3 Níveis de significância simulados de testes bilaterais (teste t clássico para 2 amostras e teste t de Welch), baseados em pares de amostras geradas a partir de duas populações normais com variâncias iguais ou diferentes plotadas versus a razão de tamanhos amostrais.

Apêndice B: comparação das funções de poder dos dois testes

Queríamos determinar as condições sob as quais a função poder do teste t de Welch pode ser igual ou aproximadamente igual à função poder do teste t clássico para 2 amostras.

Em geral, as funções de poder dos testes t (1 amostra ou 2 amostras) são bem conhecidas e discutidas em várias publicações (Pearson and Hartley, 1952; Neyman et al., 1935; Srivastava, 1958). O teorema a seguir declara a função poder de cada uma das três diferentes hipóteses alternativas em experimentos de duas amostras.

TEOREMA B1

Sob as suposições de normalidade e igualdade de variâncias, a função poder de um teste t bilateral para duas amostras que tem um tamanho nominal α pode ser expressa como uma função dos tamanhos amostrais e da diferença $\delta = \mu_1 - \mu_2$ como

$$\pi(n_1, n_2, \delta) = 1 - F_{d_C,\lambda}\left(t_{d_C}^{\alpha/2}\right) + F_{d_C,\lambda}\left(-t_{d_C}^{\alpha/2}\right)$$

onde $F_{d_C,\lambda}(.)$ é o F.D.A. da distribuição não central com $d_C = n_1 + n_2 - 2$ graus de liberdade e parâmetro de não centralidade

$$\lambda = \frac{\delta}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

Além disso, a função poder associada à hipótese alternativa $\mu_1 > \mu_2$ é dada como

$$\pi(n_1, n_2, \delta) = 1 - F_{d_C,\lambda}\left(t_{d_C}^{\alpha}\right)$$

Por outro lado, ao testar contra a alternativa $\mu_1 < \mu_2$ o poder é expresso como

$$\pi(n_1, n_2, \delta) = F_{d_C,\lambda}\left(-t_{d_C}^{\alpha}\right)$$

Apesar do resultado no teorema acima ser bem conhecido, a função poder do teste baseado no teste t modificado de Welch não foi especificamente discutida na literatura. Uma aproximação pode ser deduzida da função poder aproximada dada para o modelo da ANOVA com um fator (consulte Kulinskaya et. al, 2003). Infelizmente, esta função poder é aplicável apenas a alternativas bilaterais. Entretanto, o experimento de duas amostras é um caso tão especial que uma abordagem diferente pode ser adotada para obter a função poder (exata) do teste t de Welch para cada uma das três alternativas. Essas funções são dadas no seguinte teorema.

TEOREMA B2

Sob a suposição de que as populações são normalmente distribuídas (mas não necessariamente com a mesma variância), a função de poder de um teste t bilateral de Welch que tem um tamanho nominal α pode ser expressa como uma função dos tamanhos amostrais e e da diferença $\delta = \mu_1 - \mu_2$ como

$$\pi_W(n_1, n_2, \delta) = 1 - G_{d_W,\lambda_W}\left(t_{d_W}^{\alpha/2}\right) + G_{d_W,\lambda_W}\left(-t_{d_W}^{\alpha/2}\right)$$

onde $G_{d,\lambda}(.)$ é o F.D.A. da distribuição t não central com d_W graus de liberdade dada como

$$d_W = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2(n_1-1)} + \frac{\sigma_2^4}{n_2^2(n_2-1)}}$$

e parâmetro de não-centralidade

$$\lambda_W = \frac{\delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Para as alternativas unilaterais, as funções de poder são dadas como

$$\pi_W(n_1, n_2, \delta) = 1 - G_{d_W,\lambda_W}\left(t_{d_W}^{\alpha}\right)$$

e

$$\pi_W(n_1, n_2, \delta) = G_{d_W,\lambda_W}\left(-t_{d_W}^{\alpha}\right)$$

para teste da hipótese nula contra o $\mu_1 > \mu_2$alternativo e para testar a hipótese nula versus a alternativa $\mu_1 < \mu_2$, respectivamente.

A prova do resultado é dada no Apêndice D.

Antes de compararmos essas duas funções de poder, observe que devido ao teste t clássico para 2 amostras ser derivado sob a suposição adicional de que as variâncias das populações são iguais, as funções de poder teóricas dos dois testes devem ser comparadas quando esta segunda suposição se verifica para o teste t de Welch.

Em teoria, sabemos que sob a normalidade e suposições de variâncias iguais,

$\pi(n_1, n_2, \delta) \geq \pi_W(n_1, n_2, \delta)$ para todos os n_1, n_2, δ

O próximo resultado declara condições sob as quais as duas funções são (aproximadamente) iguais.

TEOREMA B3

Sob as suposições de normalidade e igualdade de variâncias temos o seguinte:

1. Se $n_1 \sim n_2$ então $\pi(n_1, n_2, \delta) \sim \pi_W(n_1, n_2, \delta)$ para cada diferença δ. Em particular, se $n_1 = n_2$ então $\pi(n_1, n_2, \delta) = \pi_W(n_1, n_2, \delta)$ para cada diferença δ, para que o teste t de Welch seja tão poderoso quanto o teste t clássico para 2 amostras.
2. Se n_1e $n_2$forem pequenas e $n_1 \neq n_2$então o teste t de Welch tem menos poder que o teste t clássico para 2 amostras. Contudo, se n_1e n_2 forem maiores do que $\pi(n_1, n_2, \delta) \sim \pi_W(n_1, n_2, \delta)$ (independentemente da diferença entre os tamanhos amostrais).

A prova do resultado é fornecida no Apêndice E.

Sob a suposição da igualdade de variâncias, os parâmetros de não centralidade associados com as funções de poder dos dois testes são idênticos. A diferença entre as funções de poder só pode ser atribuída à diferença entre seus respectivos graus de liberdade. Da teoria, sabemos que sob as suposições declaradas o teste t clássico é UMP (uniformemente mais poderoso) e, portanto, tem mais graus de liberdade. O ponto dos resultados acima,

entretanto, é que se o experimento for balanceado ou aproximadamente balanceado, então as funções de poder são idênticas ou aproximadamente idênticas. O único caso onde o teste t clássico é notadamente mais poderoso que o teste t de Welch é quando o experimento é marcadamente não balanceado e as amostras são pequenas. Infelizmente, também acontece de ser o caso onde o teste t clássico para 2 amostras é particularmente sensível à suposição de igualdade de variâncias conforme mostrado no Apêndice A. Como resultado, a função poder do teste t de Welch é a função mais confiável para fins práticos.

Ilustramos os resultados do Teorema B3 através do seguinte exemplo, onde as duas populações normais têm o mesmo desvio padrão de 3. Os valores de poder baseados nas funções de poder (bilaterais) do Teorema B1 e do Teorema B2 são calculados sob os seguintes quatro cenários:

1. Ambas as amostras são pequenas, mas têm o mesmo tamanho ($n_1 = n_2 = 10$).
2. Ambas as amostras são pequenas, mas uma amostra é duas vezes maior do que a a outra amostra
   ($n_1 = 10, n_2 = 20$).
3. Uma amostra é pequena e a outra amostra é moderada em tamanho, mas a amostra moderada é quatro vezes maior do que a amostra menor ($n_1 = 10, n_2 = 40$).
4. Uma amostra é moderada em tamanho e a outra é grande, mas a amostra maior é quatro vezes maior do que a amostra moderada ($n_1 = 50, n_2 = 200$).

Supondo-se que $\alpha = 0{,}05$ para ambos os testes, as funções de poder são avaliadas em cada cenário na diferença $\delta = 0{,}0, 0{,}5, 1{,}0, 1{,}5, 2{,}0, \ldots 5{,}0$. Os resultados são exibidos na Tabela 5 e as funções são plotadas na Figura 4.

Tabela 5 Comparação das funções de poder teórico de testes t clássicos bilaterais para 2 amostras e testes t bilaterais de Welch, $\alpha = 0{,}05$. Os tamanhos amostrais, n_1 e n_2, são fixos e as funções de poder são avaliados nas diferenças δ que vão de 0,0 a 5,0.

$\boldsymbol{\delta}$	0,0	0,5	1,0	1,5	2,0	2,5	3	3,5	4	4,5	5,0
$\boldsymbol{n_1 = n_2 = 10}$											
$\boldsymbol{\pi(n_1, n_2, \delta)}$	0,05	0,064	0,109	0,185	0,292	0,422	0,562	0,694	0,805	0,887	0,941
$\boldsymbol{\pi_W(n_1, n_2, \delta)}$	0,05	0,064	0,109	0,185	0,292	0,422	0,562	0,694	0,805	0,887	0,941
$\boldsymbol{n_1 = 10, n_2 = 20}$											
$\boldsymbol{\pi(n_1, n_2, \delta)}$	0,05	0,070	0,132	0,239	0,383	0,547	0,703	0,828	0,913	0,962	0,986
$\boldsymbol{\pi_W(n_1, n_2, \delta)}$	0,05	0,070	0,129	0,231	0,371	0,531	0,686	0,813	0,902	0,955	0,982
$\boldsymbol{n_1 = 10, n_2 = 40}$											
$\boldsymbol{\pi(n_1, n_2, \delta)}$	0,05	0,075	0,152	0,283	0,455	0,637	0,791	0,899	0,959	0,986	0,996
$\boldsymbol{\pi_W(n_1, n_2, \delta)}$	0,05	0,072	0,142	0,261	0,419	0,592	0,748	0,865	0,938	0,976	0,992

δ	0,0	0,5	1,0	1,5	2,0	2,5	3	3,5	4	4,5	5,0
$n_1 = 50, n_2 = 200$											
$\pi(n_1, n_2, \delta)$	0,05	0,182	0,556	0,883	0,987	0,999	1.	1.	1.	1.	1.
$\pi_W(n_1, n_2, \delta)$	0,05	0,180	0,548	0,877	0,986	0,999	1.	1.	1.	1.	1.

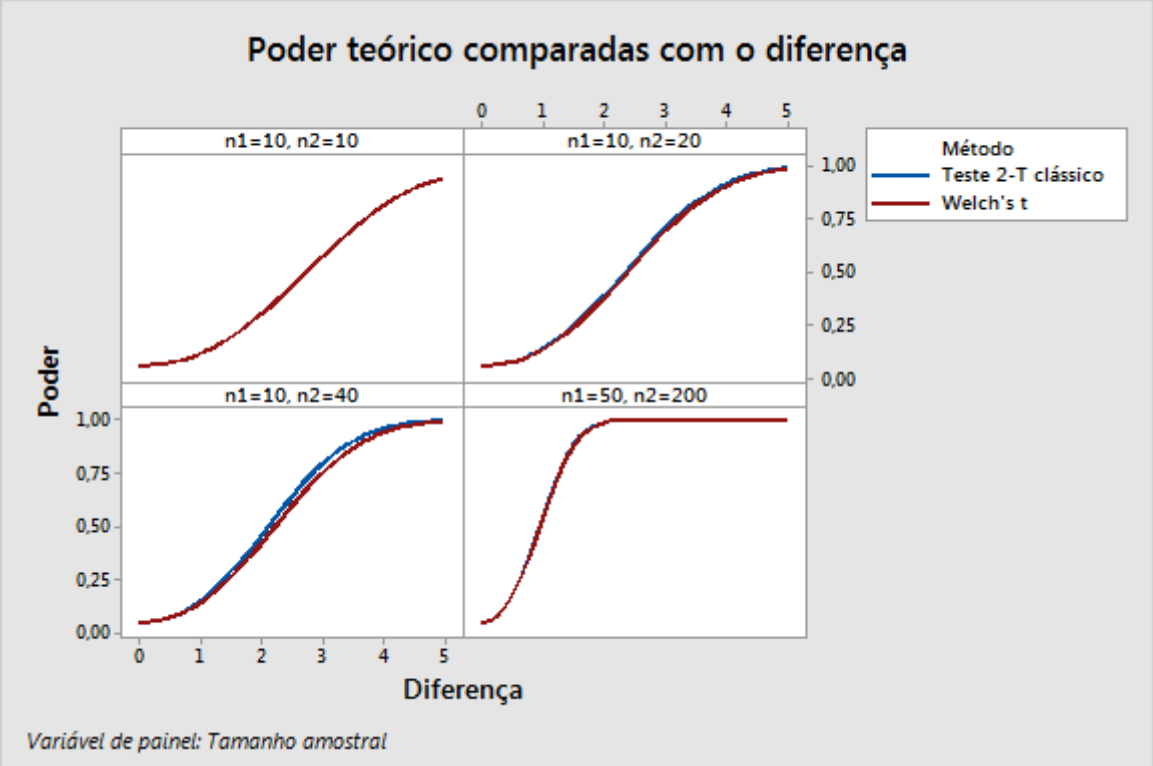

Figura 4 Gráficos de funções de poder teórico de testes t clássicos bilaterais para 2 amostras e testes t bilaterais de Welch versus δ diferença a ser detectada entre as médias. Ambos os testes usam $\alpha = 0{,}05$. As populações supostas são normais com o mesmo desvio padrão de 3.

Estudo de simulação A

O propósito deste estudo de simulação é comparar os níveis de poder associados com o teste t clássico para
2 amostras com níveis de poder associados ao teste t de Welch para 2 amostras em experimentos balanceados onde supõe-se que as variâncias sejam diferentes. Os experimentos nestes estudos são similares aos discutidos no Apêndice A.

No primeiro grupo de experimentos, foram gerados pares de amostras de tamanhos iguais a partir das populações normais com variâncias diferentes. A população de base foi fixada para ser $N(0, 2)$ e as segundas populações normais foram escolhidas da tal modo que a razão dos desvios padrão $\rho = \sigma_2/\sigma_1$ fossem iguais a 0,5, 1,5 e 2. Similarmente, em um segundo grupo, as duas amostras foram geradas a partir de distribuições qui-quadrado com variâncias diferentes (a população de base é Chi(2)). No último conjunto de experimentos os pares

amostrais foram gerados a partir da distribuição normal contaminada (a população de base CN(.8,4)) como definido anteriormente no Apêndice A.

Para cada conjunto de experimentos, calculamos os níveis de poder simulado (para uma dada diferença detectável δ) associada com cada teste para os tamanhos amostrais $n = n_1 = n_2 = 5, 10, 15, 20, 25, 30$. Em cada experimento, o nível de poder simulado foi calculado como a proporção de instâncias em que a hipótese nula foi rejeitada quando ela era falsa. Para todos os experimentos, a diferença entre as médias foi especificada em uma unidade do padrão na população de base (a primeira das duas amostras). Mais especificamente, fixamos $\delta = 1{,}0 \times \sigma_1 = 2{,}0$ porque ele é relativamente pequeno para todas as três famílias de distribuições neste estudo. Os resultados das simulações são relatados na Tabela 2.2 e exibidos na Figura 2.2a, 2.2b Figura e Figura 2.2c.

Resultados e resumo

Os resultados apresentados na Tabela 6 e na Figura 4 mostram que, sob a suposição de igualdade de variâncias, as funções teóricas de poder são idênticas em experimentos balanceados, como indicado no Teorema 2.3. Além disso, quando os tamanhos amostrais são relativamente pequenos, mas quase do mesmo tamanho, as duas funções produzem valores de poder que são aproximadamente iguais. É somente quando as amostras são relativamente pequenas e uma amostra é aproximadamente quatro vezes maior que a outra amostra que algumas diferenças perceptíveis entre as funções de poder começam a surgir (por exemplo, quando $n_1 = 10$, $n_2 = 40$). Mesmo neste caso, os valores de poder teóricos baseados no teste t clássico para 2 amostras são apenas ligeiramente mais altos do que os valores de poder com base no teste t de Welch. Por fim, quando os experimentos são notoriamente não balanceados, mas as amostras são (relativamente) grandes, as duas funções de poder são essencialmente idênticas, como declarado no Teorema B3.

Além disso, em experimentos balanceados com variâncias desiguais, os dois testes produzem valores de poder que são praticamente idênticos. Em amostras muito pequenas ($n < 10$), contudo, o teste t clássico para 2 amostras apresenta desempenho ligeiramente melhor.

Tabela 6 Comparação dos níveis de poder simulados do teste t clássico para 2 amostras e o teste de Welch em experimentos balanceados de variâncias diferentes

		População de base: N(0, 2)			População de base: Qui(2)			População de base: CN(0,8, 4)		
n	$\frac{\sigma_2}{\sigma_1}$	0,5	1,5	2,0	0,5	1,5	2,0	0,5	1,5	2,0
5	2T	0,431	0,196	0,152	0,555	0,281	0,215	0,579	0,373	0,335
	Welch	0,366	0,166	0,119	0,424	0,25	0,184	0,521	0,32	0,283
10	2T	0,77	0,385	0,27	0,846	0,438	0,324	0,79	0,51	0,435
	Welch	0,747	0,372	0,253	0,832	0,427	0,308	0,776	0,493	0,417
15	2T	0,916	0,539	0,387	0,948	0,565	0,424	0,898	0,615	0,508

		População de base: N(0, 2)			População de base: Qui(2)			População de base: CN(0,8, 4)		
	Welch	0,908	0,532	0,375	0,945	0,557	0,413	0,891	0,605	0,497
20	2T	0,971	0,682	0,497	0,982	0,68	0,521	0,952	0,702	0,573
	Welch	0,969	0,677	0,487	0,981	0,676	0,511	0,947	0,697	0,563
25	2T	0,99	0,779	0,591	0,994	0,765	0,605	0,98	0,783	0,641
	Welch	0,99	0,777	0,582	0,994	0,762	0,597	0,979	0,778	0,636
30	2T	0,998	0,851	0,675	0,998	0,826	0,676	0,994	0,839	0,699
	Welch	0,998	0,849	0,67	0,998	0,824	0,668	0,994	0,836	0,694

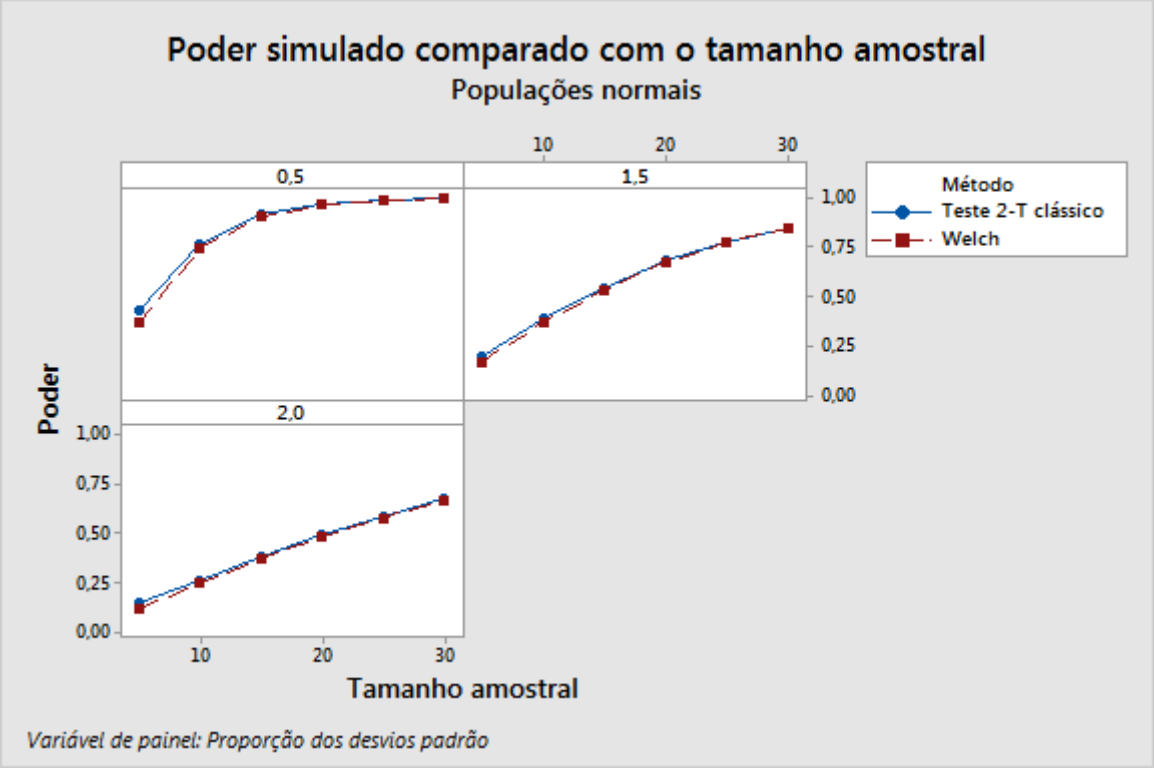

Figura 5 Comparação de níveis de poder simulados do teste t clássico para 2 amostras e do teste t de Welch para 2 amostras em experimentos balanceados de variâncias diferentes. Amostras foram geradas de populações normais com variâncias diferentes de forma que a razão dos desvios padrão é 0,5, 1,5 e 2,0.

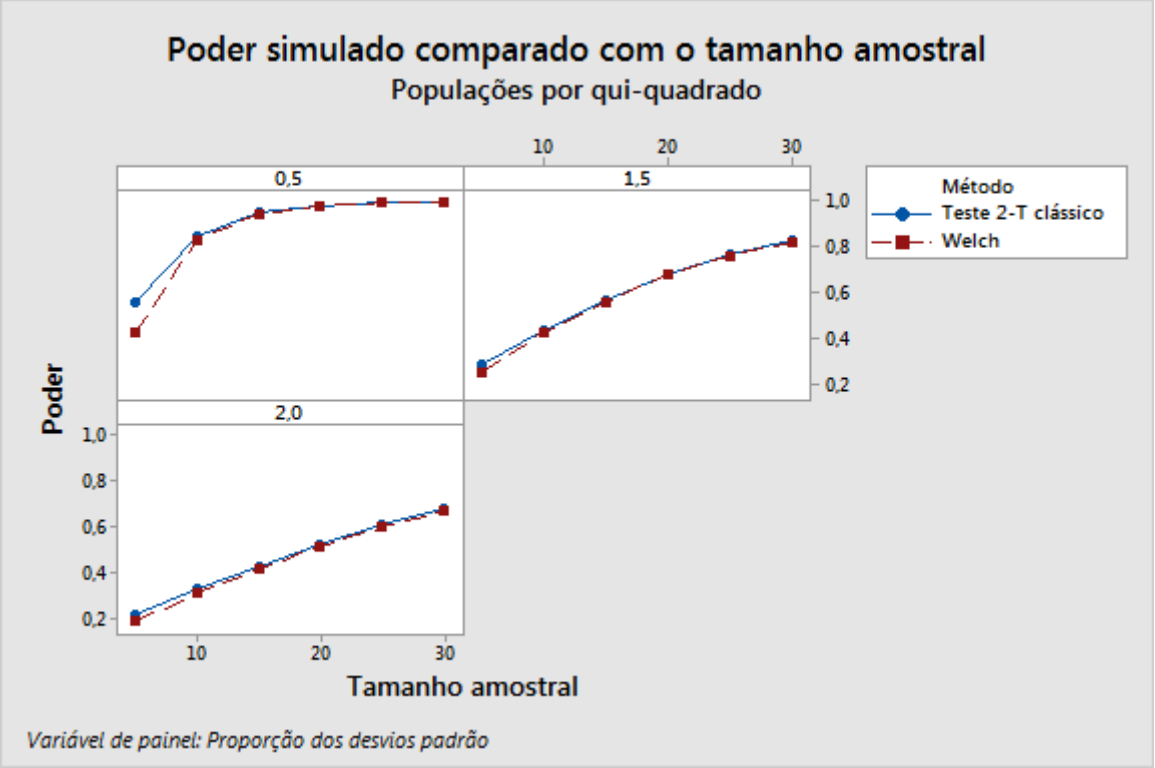

Figura 6 Comparação de níveis de poder simulados do teste t clássico para 2 amostras e do teste t de Welch para 2 amostras em experimentos balanceados de variâncias diferentes. Amostras foram geradas de populações qui-quadrado com variâncias diferentes de forma que a razão dos desvios padrão é 0,5, 1,5 e 2,0.

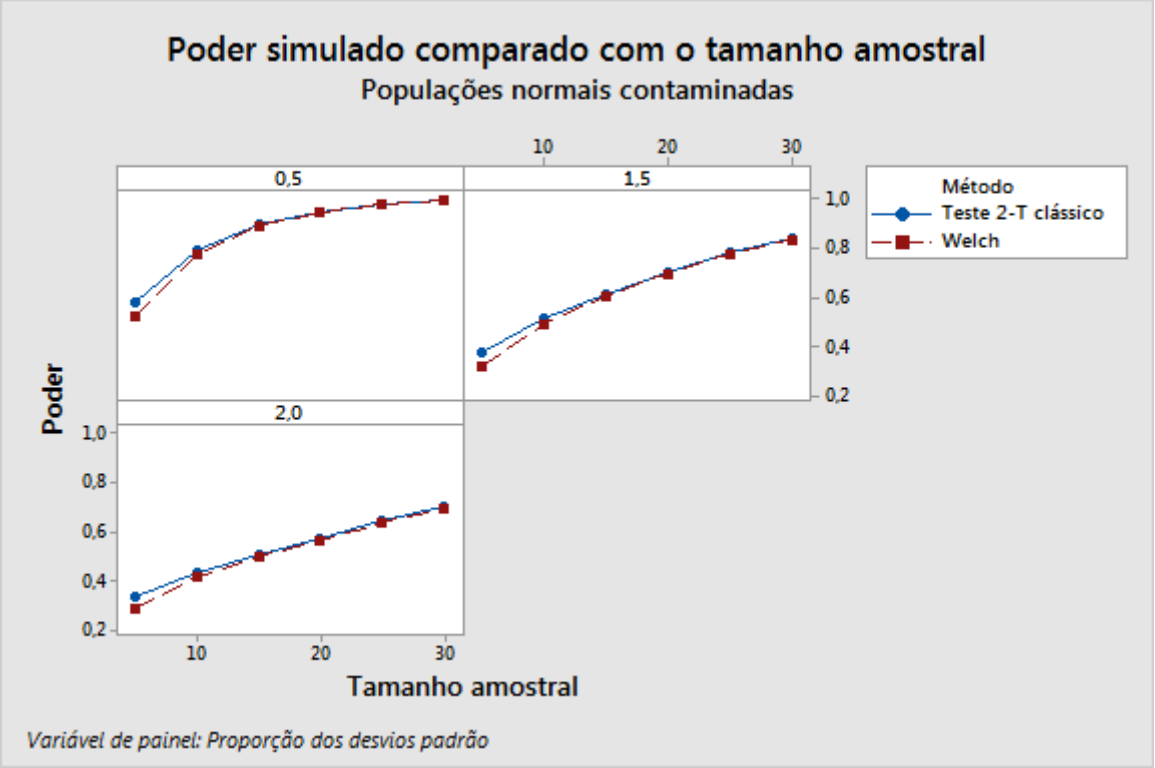

Figura 7 Comparação de níveis de poder simulados do teste t clássico para 2 amostras e do teste t de Welch para 2 amostras em experimentos balanceados de variâncias diferentes. Amostras foram geradas de populações normais contaminadas de variâncias diferentes de forma que a razão dos desvios padrão é 0,5, 1,5 e 2,0.

Apêndice C: Poder e tamanho amostral e sensibilidade à normalidade

No Assistente, o poder da análise para comparação das médias de duas populações está baseado na função poder do teste t de Welch. Caso esta função seja sensível à suposição normal sob a qual ela é derivada, a análise de poder pode produzir conclusões errôneas. Por esse motivo, conduzimos um estudo de simulação para examinar a sensibilidade desta função à suposição normal. A sensibilidade é avaliada como a consistência entre os níveis de poder simulados e os níveis de poder calculados a partir da função poder teórica quando amostras são geradas a partir de distribuições não normais. A distribuição normal serve como o controle da população porque, de acordo com o Teorema B2, os níveis de poder simulados e os níveis de poder teóricos estão mais próximos quando as amostras são geradas de populações normais.

Estudo de simulação C

O estudo é conduzido em três partes usando três distribuições: normal, qui-quadrado e a distribuição normal contaminada. Consulte o Apêndice A para obter mais detalhes. Para cada parte do estudo, o poder simulado é calculado (para os tamanhos de amostra dados n_1 e o n_2 em uma diferença detectável dada δ) conforme a proporção de instâncias quando a hipótese nula foi rejeitada quando era falsa. Em todos os casos a diferença a ser detectada é especificada em uma unidade do padrão na população base. Isto é $\delta = 1{,}0 \times \sigma_1 = 2{,}0$ para todas as três famílias de distribuições neste estudo. Os valores de poder teóricos baseados no teste t de Welch são calculados também para comparação.

Resultados da simulação e resumo

Os resultados mostram que para tamanhos amostrais relativamente pequenos a função poder o teste t de Welch é robusta à suposição de normalidade. Em geral quando o tamanho mínimo das duas amostras é menor do que 15, os valores de poder simulados estão próximos de seus níveis de poder teóricos alvo (consulte as Tabelas de 7 a 10 e as Figuras de 8 a 10).

As Tabelas 7 a 10 mostram os níveis de poder simulados de um teste t bilateral de Welch com $\alpha = 0{,}05$ baseado em pares de amostras geradas de uma população normal, populações assimétricas (qui quadrado) e populações normais contaminadas. Os pares de amostras são da mesma família de distribuição, mas as variâncias das populações pai não são necessariamente iguais. Os valores de poder teóricos são calculados para comparação.

Tabela 7 Os níveis de poder simulados de um teste t bilateral de Welch com $\alpha = 0{,}05$ para n=5

			População de base: N(0,2)				População de base: Qui(2)				População de base: CN(0,8, 4)			
		$\frac{\sigma_2}{\sigma_1}$	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0
n_2	$\frac{n_2}{n_1}$		$n_1 = 5$				$n_1 = 5$				$n_1 = 5$			
3	0,6	Obs.	0,288	0,158	0,113	0,091	0,432	0,305	0,211	0,149	0,361	0,257	0,234	0,220
		Valor alvo	0,353	0,192	0,116	0,092	0,353	0,192	0,116	0,092	0,353	0,192	0,116	0,092
5	1,0	Obs.	0,370	0,252	0,169	0,121	0,427	0,334	0,248	0,189	0,522	0,380	0,319	0,284
		Valor alvo	0,389	0,286	0,190	0,137	0,389	0,286	0,190	0,137	0,389	0,286	0,190	0,137
8	1,6	Obs.	0,387	0,326	0,242	0,179	0,427	0,364	0,286	0,225	0,573	0,453	0,374	0,319
		Valor alvo	0,400	0,345	0,260	0,193	0,400	0,345	0,260	0,193	0,400	0,345	0,260	0,193
10	2,0	Obs.	0,390	0,351	0,272	0,208	0,421	0,373	0,296	0,235	0,590	0,483	0,394	0,336
		Valor alvo	0,402	0,364	0,291	0,223	0,402	0,364	0,291	0,223	0,402	0,364	0,291	0,223

Tabela 8 Níveis de poder simulados de um teste t bilateral de Welch com $\alpha = 0{,}05$ para n=10

			População de base: N(0, 2)				População de base: Qui(2)				População de base: CN(0,8, 4)			
		$\frac{\sigma_2}{\sigma_1}$	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0
n_2	$\frac{n_2}{n_1}$		$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
5	0,5	Obs.	0,651	0,346	0,197	0,131	0,768	0,493	0,320	0,221	0,689	0,484	0,404	0,358
		Valor alvo	0,666	0,364	0,206	0,139	0,666	0,364	0,206	0,139	0,666	0,364	0,206	0,139
10	1,0	Obs.	0,742	0,556	0,369	0,254	0,831	0,612	0,430	0,308	0,776	0,619	0,496	0,419
		Valor alvo	0,745	0,562	0,337	0,259	0,745	0,562	0,337	0,259	0,745	0,562	0,337	0,259
15	1,5	Obs.	0,765	0,641	0,483	0,358	0,865	0,679	0,511	0,377	0,792	0,679	0,547	0,456

			População de base: N(0, 2)				População de base: Qui(2)				População de base: CN(0,8, 4)			
		$\frac{\sigma_2}{\sigma_1}$	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0
n_2	$\frac{n_2}{n_1}$		$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
		Valor alvo	0,767	0,643	0,483	0,352	0,767	0,643	0,483	0,352	0,767	0,643	0,483	0,352
20	2	Obs.	0,774	0,683	0,549	0,417	0,898	0,737	0,565	0,448	0,797	0,716	0,596	0,490
		Valor alvo	0,777	0,686	0,551	0,422	0,777	0,686	0,551	0,422	0,777	0,686	0,551	0,422

Tabela 9 Níveis de poder simulados de um teste t bilateral de Welch com $\alpha = 0{,}05$ para n=15

			População de base: N(0, 2)				População de base: Qui(2)				População de base: CN(0,8, 4)			
		$\frac{\sigma_2}{\sigma_1}$	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0
n_2	$\frac{n_2}{n_1}$		$n_1 = 15$				$n_1 = 15$				$n_1 = 15$			
8	.53	Obs.	0,857	0,569	0,342	0,229	0,871	0,651	0,421	0,293	0,853	0,632	0,505	0,428
		Valor alvo	0,861	0,568	0,338	0,221	0,861	0,568	0,338	0,221	0,861	0,568	0,338	0,221
15	1,0	Obs.	0,906	0,745	0,535	0,368	0,942	0,763	0,563	0,415	0,891	0,760	0,611	0,500
		Valor alvo	0,910	0,753	0,541	0,379	0,910	0,753	0,541	0,379	0,910	0,753	0,541	0,379
23	1,53	Obs.	0,928	0,831	0,667	0,502	0,975	0,858	0,676	0,517	0,898	0,825	0,698	0,572
		Valor alvo	0,925	0,830	0,670	0,509	0,925	0,830	0,670	0,509	0,925	0,830	0,670	0,509
30	2,0	Obs.	0,933	0,861	0,737	0,589	0,984	0,903	0,750	0,598	0,902	0,847	0,742	0,619
		Valor alvo	0,931	0,863	0,736	0,589	0,931	0,863	0,736	0,589	0,931	0,863	0,736	0,589

Tabela 10 Níveis de poder simulados de um teste t bilateral de Welch com $\alpha = 0{,}05$ para n=20

			População de base: N(0, 2)				População de base: Qui(2)				População de base: CN(0,8, 4)			
		$\frac{\sigma_2}{\sigma_1}$	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0	0,5	1,0	1,5	2,0
n_2	$\frac{n_2}{n_1}$		$n_1 = 20$				$n_1 = 20$				$n_1 = 20$			
10	0,5	Obs.	0,938	0,687	0,426	0,275	0,920	0,698	0,486	0,333	0,923	0,716	0,568	0,476
		Valor alvo	0,941	0,686	0,424	0,277	0,941	0,686	0,424	0,277	0,941	0,686	0,424	0,277
20	1,0	Obs.	0,971	0,866	0,672	0,485	0,981	0,858	0,670	0,506	0,952	0,856	0,696	0,567
		Valor alvo	0,971	0,869	0,673	0,489	0,971	0,869	0,673	0,489	0,971	0,869	0,673	0,489
30	1,5	Obs.	0,977	0,923	0,791	0,629	0,995	0,932	0,785	0,631	0,960	0,908	0,798	0,662
		Valor alvo	0,978	0,922	0,791	0,628	0,978	0,922	0,791	0,628	0,978	0,922	0,791	0,628
40	2,0	Obs.	0,983	0,950	0,858	0,724	0,998	0,966	0,864	0,726	0,958	0,929	0,845	0,725
		Valor alvo	0,981	0,945	0,854	0,719	0,981	0,945	0,854	0,719	0,981	0,945	0,854	0,719

Quando as duas amostras são geradas de populações normais, os valores de poder simulados são consistentes com os valores de poder teóricos, mesmo para amostras muito pequenas. Conforme ilustrado na Figura 7, as curvas de poder teóricas e simuladas são praticamente indistinguíveis. Os resultados são consistentes com o Teorema B2.

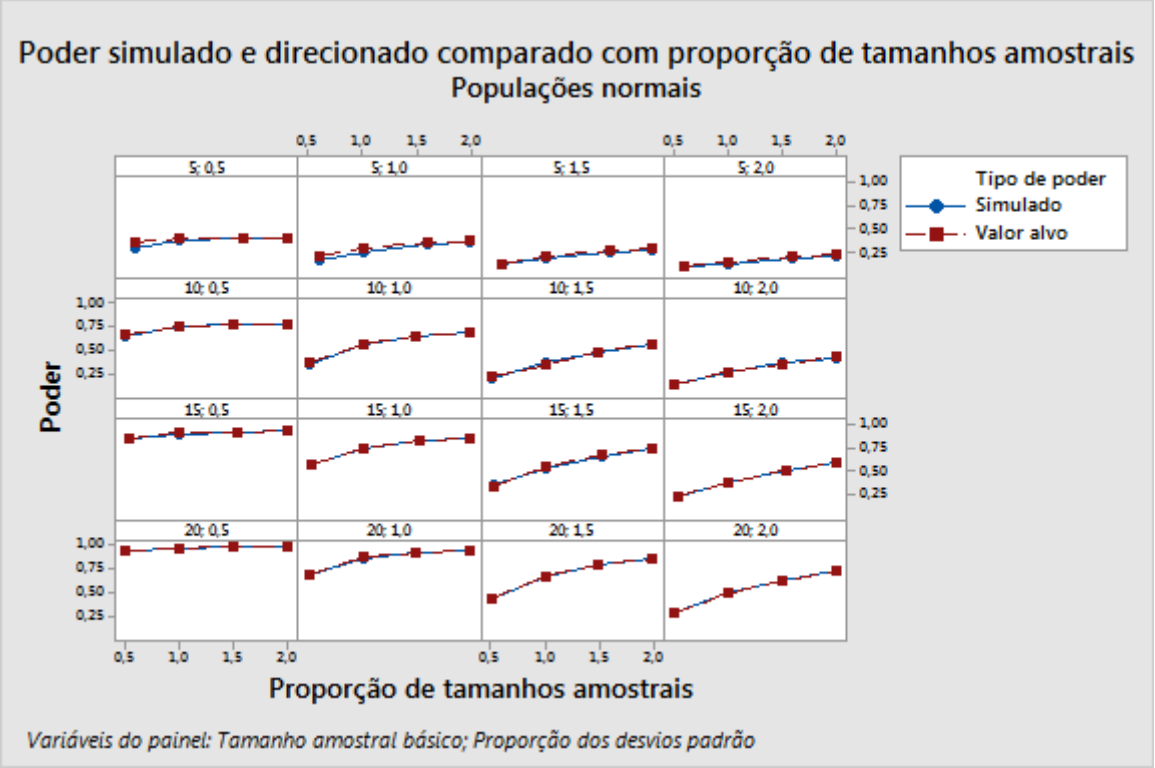

Figura 8 Níveis de poder teóricos alvo e simulados de um teste t bilateral de Welch com $\alpha = 0{,}05$ baseado em pares de amostras gerados de duas populações normais com variâncias iguais ou diferentes plotadas em relação à razão de tamanhos amostrais.

Quando as amostras são geradas de distribuições qui-quadrado assimétricas, os valores de poder simulados são maiores do que os valores de poder teóricos para amostras muito pequenas; contudo, os valores de poder tornam-se mais próximos conforme os tamanhos amostrais aumentam. A Figura 9 mostra que as curvas de poder teóricas e simuladas alvo estão consistentemente próximas quando o tamanho mínimo das duas amostras é, pelo menos, 10. Isso ilustra que dados assimétricos não têm efeito notável na função poder do teste t de Welch, mesmo quando as amostras são relativamente pequenas.

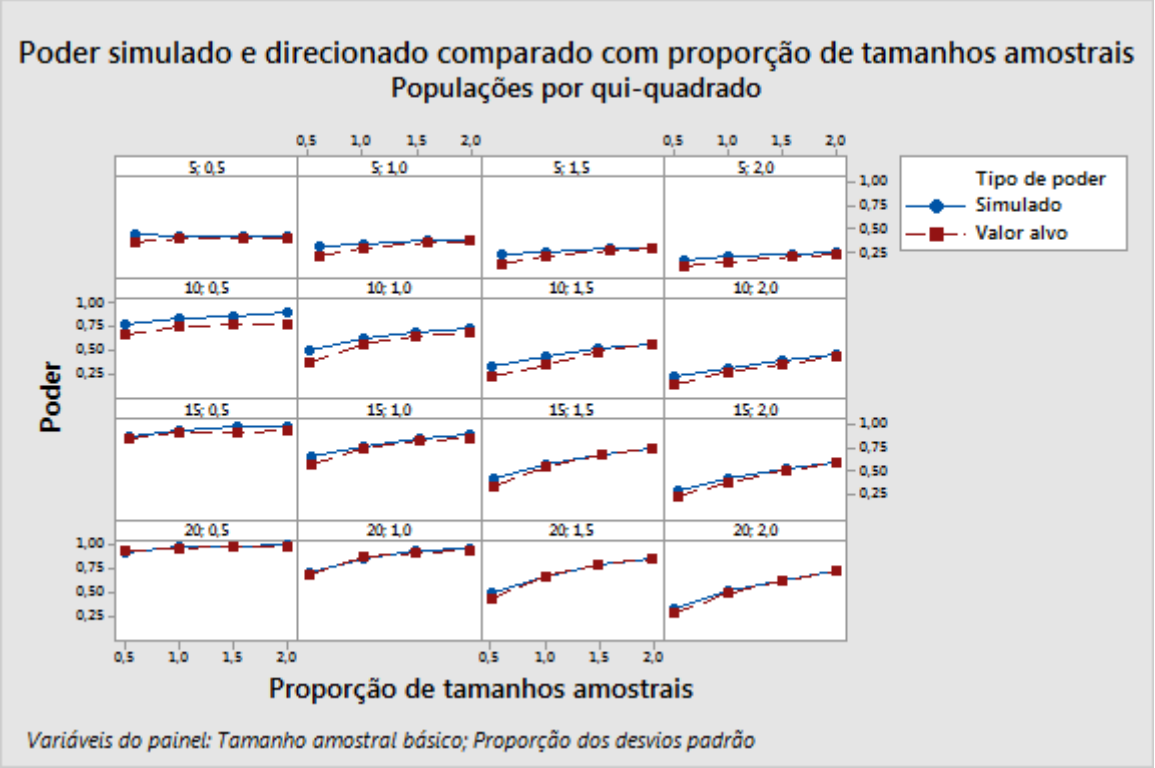

Figura 9 Níveis de poder teóricos alvo e simulados de um teste t bilateral de Welch com $\alpha = 0{,}05$ baseado em pares de amostras geradas a partir de duas populações normais com variâncias iguais ou diferentes plotadas em relação à razão de tamanhos amostrais.

Além disso, outliers tendem a ter uma influência na função poder somente quando os tamanhos amostrais são muito pequenos. Em geral, quando outliers estão presentes os valores de poder simulados tendem a ser um pouco mais altos do que os valores de poder teóricos alvo. Isso está descrito na Figura 10 onde as curvas de poder teóricas e simuladas não são razoavelmente próximas até que o tamanho amostral mínimo alcance 15.

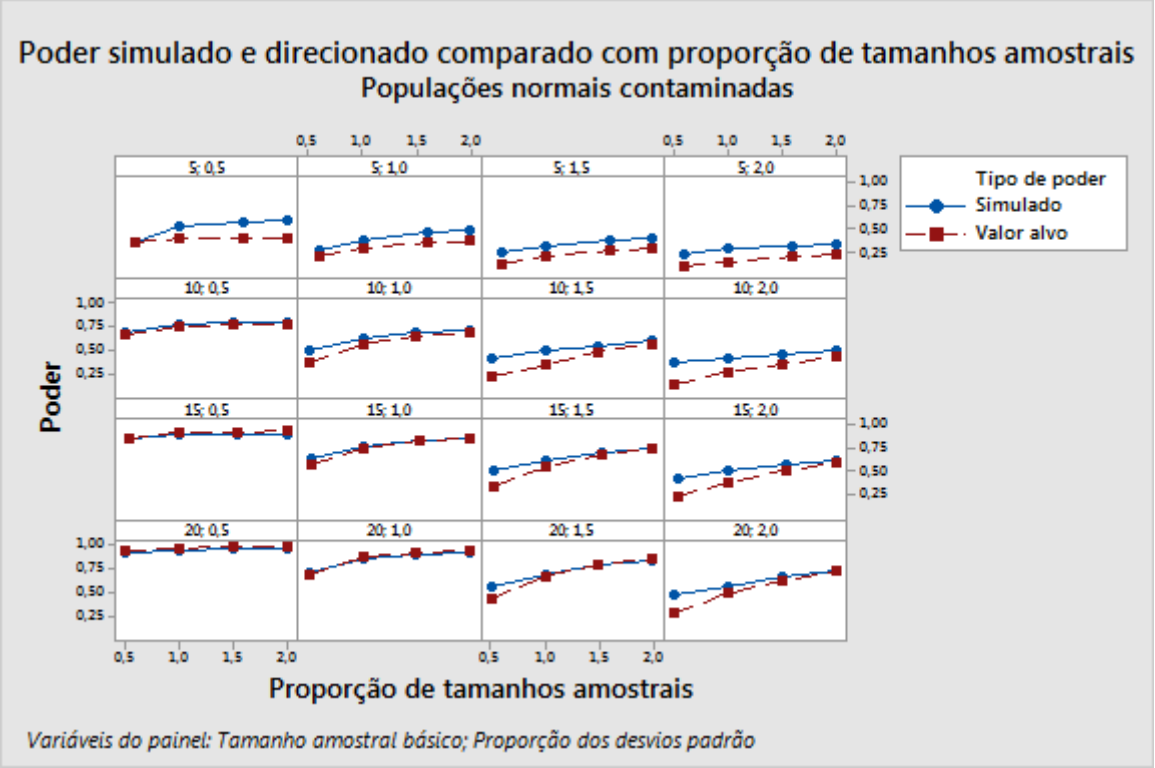

Figura 10 Os níveis de poder teóricos alvo e simulado de um teste t bilateral de Welch com $\alpha = 0{,}05$ baseado em pares de amostras geradas a partir de duas populações normais com variâncias iguais ou diferentes plotadas em relação à razão dos tamanhos amostrais.

Apêndice D: prova do teorema B2

Para o modelo de duas amostras, a abordagem de Welch para derivação da distribuição da estatística do teste

$$t_w(x,y) = \frac{\bar{x} - \bar{y} - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

sob a hipótese nula é baseada em uma aproximação da distribuição de

$$V = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

tão proporcional como uma distribuição qui-quadrado. Mais especificamente,

$$\frac{d_W V}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

é aproximadamente distribuída como uma distribuição qui-quadrado com d_W graus de liberdade onde

$$d_W = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2(n_1-1)} + \frac{\sigma_2^4}{n_2^2(n_2-1)}}$$

(Observe que em uma configuração de uma amostra isso se reduz ao resultado clássico bem conhecido que $(n-1)s^2/\sigma^2 \sim \chi_{n-1}^2$)

Considere o teste da hipótese nula H_o: $\mu_1 = \mu_2$ (ou equivalentemente $\delta = 0$) em comparação à alternativa H_A: $\mu_1 \neq \mu_2$ (ou equivalentemente $\delta \neq 0$)

Sob a hipótese nula, a função poder

$$\pi(n_1, n_2, \delta) = \pi(n_1, n_2, 0) = 1 - \Pr\left(-t_{d_W}^{\alpha/2} \leq \frac{\bar{x} - \bar{y}}{\sqrt{V}} \leq t_{d_W}^{\alpha/2}\right) \approx \alpha$$

onde t_d^α denota o ponto percentual 100 α superior da distribuição t com d graus de liberdade.

Sob a hipótese alternativa,

$$\frac{\bar{x} - \bar{y}}{\sqrt{V}} = \frac{\frac{\bar{x} - \bar{y} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} + \frac{\delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}{\sqrt{\frac{d_W V}{d_W\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}}$$

tem a distribuição t não central aproximada com d_W graus de liberdade com parâmetro de não centralidade

$$\lambda_W = \frac{\delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

devido ao que foi declarado anteriormente

$$\frac{d_W V}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

é aproximadamente distribuída como uma distribuição qui-quadrado com d_W graus de liberdade, e

$$\frac{\bar{x} - \bar{y} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

segue uma distribuição normal padrão.

Segue-se que sob a alternativa,

$$\pi(n_1, n_2, \delta) = 1 - \Pr\left(-t_{d_W}^{\alpha/2} \leq \frac{\bar{x} - \bar{y}}{\sqrt{V}} \leq t_{d_W}^{\alpha/2}\right) \approx 1 - G_{d_W,\lambda_W}\left(t_{d_W}^{\alpha/2}\right) + G_{d_W,\lambda_W}\left(-t_{d_W}^{\alpha/2}\right)$$

onde $G_{d_W,\lambda}(.)$ é o F.D.A. da distribuição t não central com d_W graus de liberdade e parâmetro de não centralidade λ como dado acima.

Apêndice E: prova do teorema B3

Primeiro, observe que d_W pode ser reescrito como

$$d_W = \frac{\left(\frac{1}{n_1} + \frac{\rho^2}{n_2}\right)^2}{\frac{1}{n_1^2(n_1-1)} + \frac{\rho^4}{n_2^2(n_2-1)}}$$

onde $\rho = \sigma_1/\sigma_2$.

Similarmente, o parâmetro de não centralidade associado com a função poder do teste de Welch também pode ser escrito como

$$\lambda_W = \frac{\delta/\sigma_1}{\sqrt{1/n_1 + \rho^2/n_2}}$$

Sob a suposição de igualdade de variâncias, os parâmetros de não centralidade associados com as funções de poder do teste t clássico para 2 amostras e do teste de Welch coincidem. Isto é

$$\lambda = \lambda_W = \frac{\delta}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

onde σ é a variância comum das duas populações. Desta forma, a única diferença nas funções de poder dos dois testes reside na diferença entre os seus respectivos graus de liberdade. Contudo, sob a suposição de igualdade de variâncias, os graus de liberdade associados com a
função poder do teste t de Welch tornam-se

$$d_W = \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^2}{\frac{1}{n_1^2(n_1-1)} + \frac{1}{n_2^2(n_2-1)}} = \frac{(n_1+n_2)^2(n_1-1)(n_2-1)}{n_1^2(n_1-1) + n_2^2(n_2-1)}$$

Pelo Teorema 1, os graus de liberdade relacionados à função poder do teste t clássico para 2 amostras são $d_C = n_1 + n_2 - 2$. Após algumas manipulações algébricas, temos

$$d_C - d_W = \frac{(n_1-n_2)^2(n_1+n_2-1)^2}{n_1^2(n_1-1) + n_2^2(n_2-1)} \geq 0$$

O fato de que $d - d_W \geq 0$ não é surpreendente, porque sabemos que sob a suposição de igualdade das variâncias, o teste t clássico para 2 amostras é UMP (uniformemente mais poderoso), como resultado, devemos esperar que os graus de liberdade associados com esta função poder sejam maiores.

Agora, se $n_1 \sim n_2$ então $d \sim d_W$ e como resultado, as funções de poder têm a mesma ordem de magnitude. Em particular, as funções de poder dos dois testes são idênticas se $n_1 = n_2$. Isso prova a primeira parte do teorema 2.3.

Se $n_1 \neq n_2$, então $d_C - d_W > 0$, de forma que o teste t de Welch tem menos poder do que o teste t clássico para 2 amostras.

Além disso, se as amostras forem maiores, isto é, se $n_1 \to \infty$ e $n_2 \to \infty$ então $d_C \to \infty$ e $d_W \to \infty$ de forma que a distribuição assintótica das estatísticas de teste associadas com ambos os testes é a distribuição normal padrão. Desta forma, os testes são assintoticamente equivalentes e produzem a mesma função poder assintótica.