Test t à 2 échantillons

Généralités

Un test t à 2 échantillons permet de comparer deux groupes indépendants pour déterminer s'ils sont différents. Ce test suppose que les deux populations possèdent des variances égales et qu'elles sont distribuées normalement. Bien que l'hypothèse de normalité ne soit pas critique (Pearson, 1931 ; Barlett, 1935 ; Geary, 1947), l'hypothèse d'égalité des variances l'est lorsque les effectifs d'échantillons sont sensiblement différents (Welch, 1937 ; Horsnell, 1953).

Certains spécialistes effectuent d'abord un test préliminaire pour évaluer l'égalité des variances, puis appliquent la procédure standard du test t à 2 échantillons. Toutefois, cette approche présente de sérieux inconvénients, car ces tests de variance sont soumis à des hypothèses et à des limites importantes. Par exemple, de nombreux tests d'égalité des variances, comme le test F classique, sont sensibles à des écarts par rapport à la normalité. D'autres tests ne s'appuyant pas sur l'hypothèse de normalité, comme le test de Levene/Brown-Forsythe, ne sont pas suffisamment puissants pour détecter une différence entre des variances.

B.L. Welch a développé une méthode d'approximation pour comparer les moyennes de deux populations normales indépendantes lorsque leurs variances ne sont pas nécessairement égales (Welch, 1947). Le test t modifié de Welch ne supposant pas l'égalité des variances, il permet aux utilisateurs de comparer les moyennes de deux populations sans avoir à tester l'égalité des variances au préalable.

Dans cette étude, nous comparons la méthode t modifiée de Welch et la procédure standard du test t à 2 échantillons afin de déterminer le test le plus fiable. Nous décrivons également les vérifications des données, qui sont automatiquement effectuées et affichées dans le rapport de l'Assistant, et nous expliquons comment elles influent sur les résultats de l'analyse :

- Normalité
- Donnée aberrantes
- Effectif de l'échantillon

Méthode du test t à 2 échantillons

Test t à 2 échantillons classique contre test de Welch

Lorsque des données proviennent de deux populations normales ayant des variances identiques, le test t à 2 échantillons classique est au moins aussi puissant que le test t de Welch. L'hypothèse de normalité n'est pas critique pour la procédure standard (Pearson, 1931 ; Barlett, 1935, Geary, 1947), mais l'hypothèse d'égalité des variances est nécessaire pour garantir la validité des résultats. Plus précisément, la procédure standard est sensible à l'hypothèse d'égalité des variances lorsque les effectifs des échantillons diffèrent, et ce, quelle que soit leur importance respective (Welch, 1937 ; Horsnell, 1953). Toutefois, en pratique, l'hypothèse d'égalité des variances prévaut rarement, ce qui peut entraîner des taux d'erreur de 1ère espèce plus élevés. Par conséquent, si le test t à 2 échantillons classique est utilisé avec deux échantillons qui présentent des variances différentes, il aura plus de risques de fournir des résultats incorrects.

Le test t de Welch est une alternative intéressante au test t classique, car il ne suppose pas l'égalité des variances. Par conséquent, quels que soient les effectifs des échantillons, la présence de variances différentes n'a pas d'influence sur ses résultats. Toutefois, le test t de Welch est basé sur une procédure d'approximation et ses résultats peuvent être discutables lorsque les échantillons sont petits. Nous souhaitions déterminer quelle méthode, du test t à 2 échantillons classique ou du test t de Welch, est la plus fiable et la plus pratique à utiliser dans l'Assistant.

Objectif

Nous souhaitions déterminer le test le plus fiable par le biais d'études de simulation et de considérations théoriques. Plus précisément, nous souhaitons examiner les éléments suivants :

- Les taux d'erreurs de 1ère et 2ème espèce du test t à 2 échantillons classique et du test t de Welch avec différents effectifs d'échantillons lorsque les données sont distribuées normalement et que les variances sont égales.
- Le taux d'erreur de 1ère et 2ème espèce du test t de Welch dans des plans non équilibrés et présentant des variances inégales, pour lesquels le test t à 2 échantillons classique n'est pas adapté.

Méthode

Nous avons effectué trois types de simulation :

- Nous avons comparé les résultats de simulation du test t à 2 échantillons classique et du test t de Welch sous différentes hypothèses de modèle, à savoir les hypothèses de normalité, de non-normalité, d'égalité des variances, d'inégalité des variances, de plans équilibrés et non équilibrés. Pour plus de détails, reportez-vous à l'Annexe A.

- Nous avons dérivé la fonction puissance du test t de Welch, puis nous l'avons comparée à la fonction puissance du test t à 2 échantillons classique. Pour plus de détails, reportez-vous à l'Annexe B.
- Nous avons étudié l'influence de la non-normalité sur la fonction puissance théorique du test t de Welch.

Les résultats

Lorsque les hypothèses nécessaires au modèle t à 2 échantillons classique sont satisfaites, le test t de Welch obtient des résultats aussi bons ou presque aussi bons que ceux du test t à 2 échantillons classique, excepté pour les plans non équilibrés de petites dimensions. Toutefois, le test t à 2 échantillons classique peut également fournir de mauvais résultats avec des plans non équilibrés de petites dimensions, en raison de sa sensibilité à l'hypothèse d'égalité des variances. En outre, en pratique, il est difficile d'être absolument certain que deux populations ont exactement la même variance. Par conséquent, la supériorité théorique du test t à 2 échantillons sur le test t de Welch n'a que peu de valeur pratique, voire aucune. C'est pourquoi l'Assistant utilise le test t de Welch pour comparer les moyennes de deux populations. Pour voir les résultats détaillés des simulations, reportez-vous aux annexes A, B et C.

Vérifications des données

Normalité

Le test t de Welch, qui est la méthode utilisée dans l'Assistant pour comparer les moyennes de deux populations indépendantes, suppose que les populations sont distribuées normalement. De plus et par chance, même lorsque les données ne sont pas distribuées normalement, le test t de Welch offre de bons résultats si les échantillons sont suffisamment grands.

Objectif

Nous souhaitions déterminer dans quelle mesure les seuils de signification simulés pour la méthode de Welch et le test t à 2 échantillons classique correspondent au seuil de signification (taux d'erreur de 1ère espèce) cible de 0,05.

Méthode

Nous avons effectué des simulations du test t de Welch et du test t à 2 échantillons classique sur 10 000 paires d'échantillons indépendants issus de populations normales, asymétriques et normales contaminées (variances égales et inégales). Nous avons utilisé différents effectifs d'échantillons. La population normale sert de population de contrôle, à des fins de comparaison. Pour chaque condition, nous avons calculé les seuils de signification simulés, puis nous les avons comparés au seuil de signification cible (ou nominal) de 0,05. Si le test est adapté, le seuil de signification simulé doit être proche de 0,05.

Les résultats

Pour des échantillons moyens ou importants, le test t de Welch présente le même taux d'erreur de 1ère espèce avec des données normales et non normales. Lorsque les effectifs des deux échantillons sont d'au moins 15, les seuils de signification simulés sont proches du seuil de signification cible. Pour plus de détails, reportez-vous à l'Annexe A.

Etant donné que le test obtient de bons résultats avec des échantillons relativement petits, l'Assistant ne teste pas la normalité des données. Il vérifie l'effectif des échantillons et affiche les indicateurs d'état suivants dans le rapport :

Etat	Condition
	Les deux effectifs d'échantillons sont d'au moins 15, la normalité n'est donc pas un problème.
	Au moins un des effectifs d'échantillons est inférieur à 15, ce qui peut poser un problème de normalité.

Données aberrantes

Les données aberrantes sont des valeurs extrêmement grandes ou extrêmement petites, également connues sous le nom de valeurs aberrantes. Ces données aberrantes peuvent avoir une grande influence sur les résultats de l'analyse. Lorsque l'échantillon est faible, elles peuvent réduire les chances d'obtenir des résultats statistiquement significatifs. Les données aberrantes peuvent indiquer des problèmes de collecte de données ou un comportement inhabituel d'un procédé. Ainsi, il vaut souvent la peine d'examiner ces points de données plus en profondeur et de les corriger lorsque cela est possible.

Objectif

Nous souhaitions développer une méthode pour analyser les valeurs très grandes ou très petites par rapport à l'échantillon global et susceptibles d'influer sur les résultats de l'analyse.

Méthode

Nous avons développé une méthode pour vérifier la présence de données aberrantes, inspirée de la méthode décrite par Hoaglin, Iglewicz et Tukey (1986), qui permet d'identifier les valeurs aberrantes dans les boîtes à moustaches.

Les résultats

L'Assistant identifie un point de données comme aberrant s'il se trouve à une distance 1,5 fois supérieure à l'étendue interquartile au-delà du quartile inférieur ou supérieur de la distribution. Les quartiles inférieur et supérieur sont les 25ème et 75ème percentiles des données. L'étendue interquartile représente la différence entre les deux quartiles. Cette méthode donne de bons résultats même lorsqu'il existe plusieurs valeurs aberrantes car elle permet de détecter chaque valeur aberrante spécifique.

Les valeurs aberrantes n'influent généralement sur la fonction puissance que lorsque les effectifs d'échantillons sont très faibles. En général, en présence de valeurs aberrantes, les valeurs de puissance observée ont tendance à être légèrement supérieures aux valeurs de puissance théorique cible. Cette tendance est visible sur la figure 10 de l'Annexe C : les courbes de puissance simulée et théorique ne sont raisonnablement proches que lorsque l'effectif minimal d'échantillon est d'au moins 15.

Lors du test des données aberrantes, le rapport de l'Assistant sur le test t à 2 échantillons affiche les indicateurs d'état suivants :

Etat	Condition
	Il n'existe aucun point de données aberrant.
	Au moins un point de données est aberrant et peut avoir une influence sur les résultats du test.

Effectif d'échantillon

Par définition, un test d'hypothèse vise à collecter des preuves permettant de rejeter l'hypothèse nulle de "non-différence". Lorsque les échantillons sont trop petits, la puissance du test peut ne pas être adaptée pour détecter une différence existante entre les moyenne, ce qui entraîne une erreur de 2ème espèce. Il est donc essentiel de s'assurer que les effectifs d'échantillons sont suffisamment grands pour détecter des différences importantes en pratique avec une probabilité élevée.

Objectif

Si les données actuelles ne permettent pas de rejeter l'hypothèse nulle, il nous faut déterminer si les effectifs d'échantillons sont suffisamment grands pour que le test détecte des différences pratiques avec une probabilité élevée. Même si la planification des effectifs d'échantillons vise à garantir que les échantillons sont suffisamment grands pour détecter d'importantes différences avec une probabilité élevée, ces échantillons ne doivent pas être grands au point que des différences sans importance deviennent statistiquement significatives avec une probabilité élevée.

Méthode

L'analyse de puissance et d'effectif d'échantillon s'appuie sur la fonction puissance théorique propre au test utilisé pour effectuer l'analyse statistique. Pour le test t de Welch, cette fonction puissance dépend de l'effectif des échantillons, de la différence entre les moyennes des deux populations, ainsi que des variances réelles des deux populations. Pour plus de détails, reportez-vous à l'Annexe B.

Les résultats

Lorsque les données ne fournissent pas suffisamment de preuves invalidant l'hypothèse nulle, l'Assistant calcule les différences pratiques pouvant être détectées avec une probabilité de 80 % et de 90 % pour les effectifs d'échantillons donnés. De plus, dans le cas où l'utilisateur indique une différence pratique à tester, l'Assistant calcule les effectifs d'échantillons qui offrent une probabilité de 80 % et de 90 % de détecter cette différence.

Nous ne pouvons pas vous présenter de résultat général, car les résultats dépendent des échantillons spécifiques de l'utilisateur. Toutefois, vous pouvez vous reporter aux annexes B et C pour obtenir plus d'informations sur la fonction puissance du test de Welch.

Lors du test de puissance et d'effectif d'échantillon, le rapport de l'Assistant sur le test t à 2 échantillons affiche les indicateurs d'état suivants :

Etat	Condition
	Le test détecte une différence entre les moyennes, par conséquent la puissance n'est pas un problème. OU La puissance est suffisante. Le test n'a pas détecté de différence entre les moyennes, mais l'échantillon est suffisamment grand pour fournir une probabilité d'au moins 90 % de détecter la différence donnée.

Etat	Condition
	Il se peut que la puissance soit suffisante. Le test n'a pas détecté de différence entre les moyennes, mais l'échantillon est suffisamment grand pour fournir une probabilité de 80 % à 90 % de détecter la différence donnée. L'effectif d'échantillon nécessaire pour atteindre une puissance de 90 % est indiqué.
	Il se peut que la puissance ne soit pas suffisante. Le test n'a pas détecté de différence entre les moyennes, mais l'échantillon est suffisamment grand pour fournir une probabilité de 80 % à 90 % de détecter la différence donnée. Les effectifs d'échantillon nécessaires pour atteindre une puissance de 80 % et de 90 % sont indiqués.
	La puissance n'est pas suffisante. Le test n'a pas détecté de différence entre les moyennes, et l'échantillon n'est pas suffisamment grand pour fournir une probabilité d'au moins 60 % de détecter la différence donnée. Les effectifs d'échantillons nécessaires pour atteindre une puissance de 80 % et de 90 % sont indiqués.
	Le test n'a pas détecté de différence entre les moyennes. Vous n'avez pas indiqué de différence pratique à détecter entre les moyennes. Par conséquent, le rapport indique les différences ayant 80 et 90 % de chances d'être détectées avec les effectifs d'échantillons, les écarts types et la valeur alpha utilisés.

Références

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. et 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. et 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. et 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. et 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 (*4ème éd.), Champaign, IL : Wolfram Media/Cambridge University Press.

Annexe A : Impact de la non-normalité et de l'hétérogénéité sur le test t à 2 échantillons classique et le test t de Welch

Nous avons réalisé plusieurs études de simulation visant à comparer le test t à 2 échantillons classique et le test t de Welch avec différentes hypothèses de modèle.

Etude par simulation A

Nous avons décomposé l'étude en trois étapes :

- Dans la première partie de l'étude, nous avons évalué la sensibilité du test t à 2 échantillons classique et du test t de Welch à l'hypothèse d'égalité des variances lorsque l'hypothèse de normalité est vraie. Deux échantillons ont été créés à partir de deux populations normales indépendantes. Le premier échantillon, l'échantillon de base, a été créé à partir d'une population normale ayant pour moyenne 0 et pour écart type $\sigma_1 = 2$, $N(0, 2)$. Le second échantillon a également été créé à partir d'une population ayant une moyenne de 0, mais avec un écart type de σ_2, choisi de telle sorte que le rapport $\rho = \sigma_2/\sigma_1$ soit égal à 0,5, 1,0, 1,5 et 2. En d'autres termes, les seconds échantillons ont été créés à partir des populations $N(0, 1)$, $N(0, 2)$, $N(0, 3)$ et $N(0, 4)$, respectivement. En outre, pour chaque cas, l'effectif de l'échantillon de base a été défini sur $n_1 = 5, 10, 15, 20$ et pour chaque valeur de n_1 donnée, le second effectif de l'échantillon, n_2, a été choisi de telle sorte que le rapport des effectifs d'échantillons, $r = n_2/n_1$, soit approximativement égal à 0,5, 1, 1,5 et 2,0.

  Pour chacun de ces plans à 2 échantillons, nous avons créé 10 000 paires d'échantillons indépendants à partir des populations étudiées. Nous avons ensuite réalisé le test t à 2 échantillons classique et le test t de Welch sur chacune des 10 000 paires d'échantillons pour tester l'hypothèse nulle selon laquelle il n'existe aucune différence entre les moyennes. Etant donné que la différence réelle entre les moyennes est nulle, la fraction des 10 000 répliques pour laquelle l'hypothèse nulle est rejetée représente le seuil de signification simulé du test. Etant donné que le seuil de signification cible de chacun des tests est de $\alpha = 0{,}05$, l'erreur de simulation associée à chaque test et chaque expérience est d'environ 0,2 %.

- Dans la deuxième partie, nous avons étudié l'impact de la non-normalité, et plus précisément de l'asymétrie, sur les seuils de signification simulés des deux tests. Cette simulation a été effectuée de la même façon que la précédente. Toutefois, l'échantillon de base a été créé en utilisant une loi du Khi deux à 2 degrés de liberté, Khi(2), et les seconds échantillons ont été créés en utilisant d'autres lois du Khi deux, de sorte que $\rho = \sigma_2/\sigma_1$ soit égal à 0,5, 1,0, 1,5 et 2. La différence hypothétisée entre

les moyennes a été configurée pour constituer la différence réelle entre les moyennes des populations parent.

- Dans la troisième partie, nous avons examiné l'effet des valeurs aberrantes sur les résultats des deux tests t. C'est pourquoi les deux échantillons ont été créés à partir de lois normales contaminées. Une population normale contaminée $\mathrm{CN}(\mathrm{p}, \sigma)$ est un mélange de deux populations normales : la population $\mathrm{N}(0,1)$ et la population normale $\mathrm{N}(0, \sigma)$. Nous définissons une loi normale contaminée de la façon suivante :

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

où p est le paramètre de mélange et $1 - \mathrm{p}$ la proportion de contamination (ou la proportion de valeurs aberrantes). Il est simple de démontrer que, lorsque X obéit à la loi $CN(p, \sigma)$, sa moyenne est de $\mu_X = 0$ et son écart type est égal à $\sigma_X = \sqrt{p + (1-p)\sigma^2}$.

L'échantillon de base a été créé à partir de la loi $CN(0{,}8, 4)$ et le second à partir de la loi normale contaminée $CN(0{,}8, \sigma)$. Le paramètre σ a été choisi de telle sorte que le rapport des écarts types des deux populations (contaminées), $\rho = \sigma_2/\sigma_1$, soit égal à 0,5, 1,0, 1,5 et 2, comme dans les parties I et II. Comme $\sigma_1 = \sqrt{0{,}8 + (1-0{,}8) * 16} = 2{,}0$, la valeur d'écart type choisie sera $\sigma = 1, 4, 6.40, 8.72$, respectivement. En d'autres termes, les seconds échantillons ont été créés en utilisant les lois $CN(0{,}8, 1)$, $CN(0{,}8, 4)$, $CN(0{,}8, 6{,}4)$ et $CN(0{,}8, 8{,}72)$. Nous avons ensuite réalisé les simulations comme décrit dans la partie I.

Les résultats de l'étude sont présentés dans le tableau 1 et modélisés dans les figures 1, 2 et 3.

Résultats et récapitulatif

En général, les résultats de la simulation corroborent les résultats théoriques selon lesquels, partant des hypothèses de normalité et d'égalité des variances, le test t à 2 échantillons classique offre des seuils de signification proches du seuil cible, même avec de faibles échantillons. Dans la figure 1, la deuxième colonne des diagrammes représente les seuils de signification simulés dans des plans où les variances des deux populations normales sont égales. Les courbes des seuils de signification simulés obtenues avec le test t à 2 échantillons classique sont indissociables des lignes de seuil cible.

Les tableaux ci-dessous indiquent les seuils de signification simulés de tests bilatéraux réalisés pour le test t à 2 échantillons classique et le test t de Welch, avec $\alpha = 0{,}05$ pour des paires d'échantillons créées à partir de populations normales, asymétriques (Khi deux) et normales contaminées. Les paires d'échantillons obéissent à la même famille de lois, mais les variances des populations parent respectives ne sont pas nécessairement égales.

Tableau 1 : Seuils de signification simulés de tests bilatéraux (test t à 2 échantillons classique et test t de Welch, tous deux avec $\alpha = 0{,}05$) pour n = 5.

			Pop. de base : N(0,2) 2ème pop. : N(0, σ_2)				**Pop. de base : Khi(2) 2ème pop. : Khi deux**				**Pop. de base : CN(0,8, 4) 2ème 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	$\frac{n_2}{n_1}$	**Méth.**	$n_1 = 5$				$n_1 = 5$				$n_1 = 5$			
3	**0,6**	**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

Tableau 2 : Seuils de signification simulés de tests bilatéraux (test t à 2 échantillons classique et test t de Welch, tous deux avec $\alpha = 0{,}05$) pour n = 10.

			Pop. de base : N(0,2) 2ème pop. : N(0, σ_2)				**Pop. de base : Khi(2) 2ème pop. : Khi deux**				**Pop. de base : CN(0,8, 4) 2ème 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	$\frac{n_2}{n_1}$	**Méth.**	$n_1 = 10$				$n_1 = 10$				$n_1 = 10$			
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

Tableau 3 : Seuils de signification simulés de tests bilatéraux (test t à 2 échantillons classique et test t de Welch, tous deux avec $\alpha = 0{,}05$) pour n = 15.

			Pop. de base : N(0,2) **2ème pop. : N(0, σ_2)**				**Pop. de base : Khi(2)** **2ème pop. : Khi deux**				**Pop. de base : CN(0,8, 4)** **2ème 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	$\frac{n_2}{n_1}$	**Méth.**	$n_1 = 15$				$n_1 = 15$				$n_1 = 15$			
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

Tableau 4 : Seuils de signification simulés de tests bilatéraux (test t à 2 échantillons classique et test t de Welch, tous deux avec $\alpha = 0{,}05$) pour n = 20.

			Pop. de base : N(0,2) **2ème pop. : N(0, σ_2)**				**Pop. de base : Khi(2)** **2ème pop. : Khi deux**				**Pop. de base : CN(0,8, 4)** **2ème 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	$\frac{n_2}{n_1}$	**Méth.**	$n_1 = 20$				$n_1 = 20$				$n_1 = 20$			
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

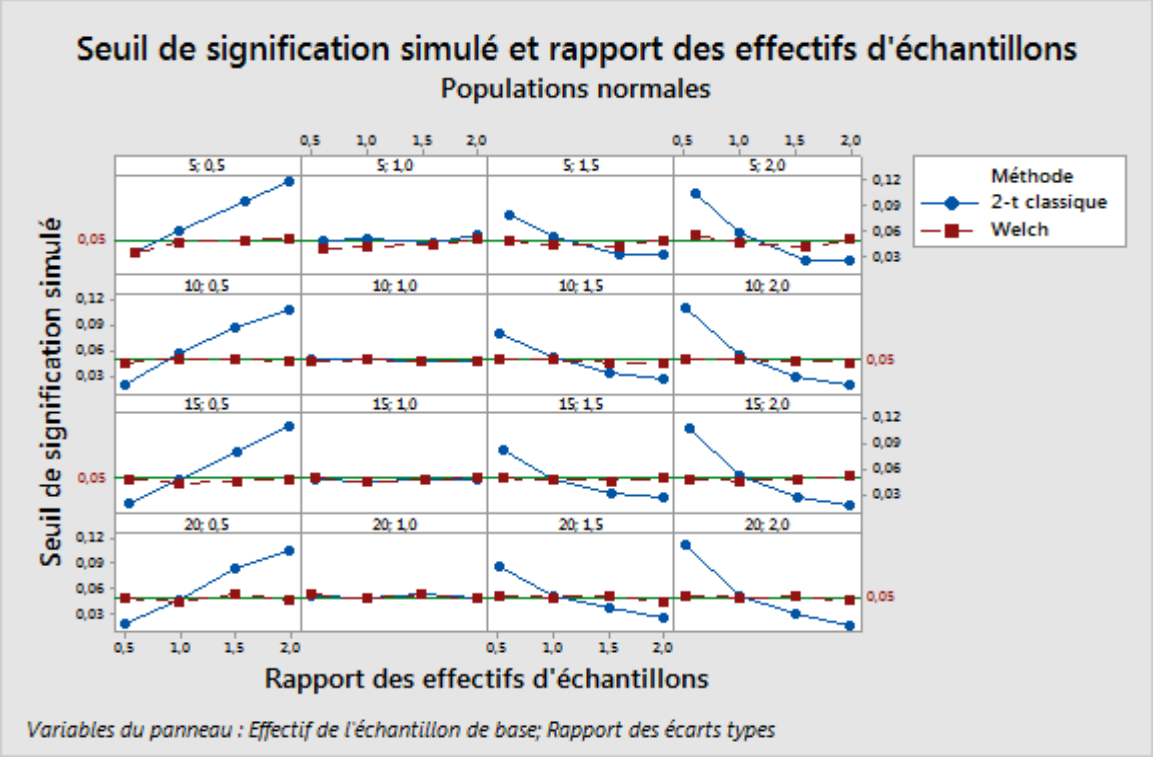

Figure 1 : Seuils de signification simulés de tests bilatéraux (test t à 2 échantillons classique et test t de Welch, tous deux avec $\alpha = 0{,}05$) pour des paires d'échantillons créées à partir de deux populations normales avec des variances égales ou inégales, en fonction du rapport des effectifs d'échantillons.

Les résultats de la simulation indiquent que, pour des échantillons relativement faibles, le test t à 2 échantillons classique est robuste à la non-normalité mais sensible à l'hypothèse d'égalité des variances, excepté lorsque le plan à 2 échantillons est pratiquement équilibré. Les graphiques des figures 1, 2 et 3 illustrent ces conclusions. Les courbes du seuil de signification simulé pour le test t à 2 échantillons classique coupent la ligne de seuil cible au point où le rapport des effectifs d'échantillons est de 1,0, même avec des variances très différentes. Pour les trois familles de lois (populations normale, Khi deux et normale contaminée), lorsque les effectifs d'échantillons sont différents, les seuils de signification simulés du test t à 2 échantillons classique ne sont proches du seuil cible que lorsque les variances sont égales. La deuxième colonne de graphiques des figures 1, 2 et 3 illustre cette observation.

Les résultats du test t classique ne sont pas satisfaisants avec un plan non équilibré et des variances inégales. Des différences entre les variances, même mineures, posent problème. Pour ces plans non équilibrés présentant des variances inégales, les seuils de signification simulés restent non satisfaisants, même lorsque les données sont normalement distribuées. En fait, plus les effectifs d'échantillons augmentent, plus les seuils de signification simulés s'écartent du seuil cible, quelle que soit la population parent. Lorsque l'échantillon le plus grand est issu de la population ayant la variance la plus élevée, les seuils de signification simulés sont inférieurs au seuil cible. Lorsque les échantillons les plus grands sont issus de la population ayant la variance la plus faible, les seuils simulés sont supérieurs aux seuils cible.

Arnold (1990, p. 372) formule une observation similaire à propos de la loi asymptotique de la statistique du test t à 2 échantillons classique sous l'hypothèse d'inégalité des variances.

Par ailleurs, le test t à 2 échantillons de Welch n'est pas sensible à des écarts par rapport à l'hypothèse d'égalité des variances, comme le montrent les figures 1, 2 et 3. Cela n'est pas étonnant, car le test t de Welch ne suppose pas l'égalité des variances. L'hypothèse normale sur laquelle s'appuie le test t de Welch semble n'avoir d'importance que lorsque l'effectif minimal des deux d'échantillons est très faible. Pour des échantillons plus grands, en revanche, le test n'est pas sensible à des écarts par rapport à l'hypothèse de normalité. Les figures 2 et 3 illustrent cette observation : les seuils de signification simulés restent invariablement proches du seuil cible lorsque l'effectif minimal des deux échantillons est de 15. Lorsque les deux échantillons obéissent à une loi du Khi deux à 2 degrés de liberté et que l'effectif des deux échantillons est de 15, le seuil de signification simulé est de 0,042 (voir tableau 3).

Les valeurs aberrantes ne semblent pas non plus avoir de conséquences sur les résultats du test t de Welch lorsque l'effectif minimal des deux échantillons est suffisamment élevé. Le tableau 3 et la figure 3 indiquent que, lorsque l'effectif minimal des deux échantillons est d'au moins 15, les seuils de signification simulés sont proches du seuil cible (les seuils de signification simulés sont de 0,045, 0,045, 0,041 et 0,037 lorsque le rapport des écarts types est de 0,5, 1,0, 1,5 et 2,0 respectivement).

Ces résultats indiquent que, pour la plupart des applications pratiques, le test t à 2 échantillons de Welch offre de meilleurs résultats que le test t à 2 échantillons classique en ce qui concerne le seuil de signification simulé ou le taux d'erreur de 1ère espèce.

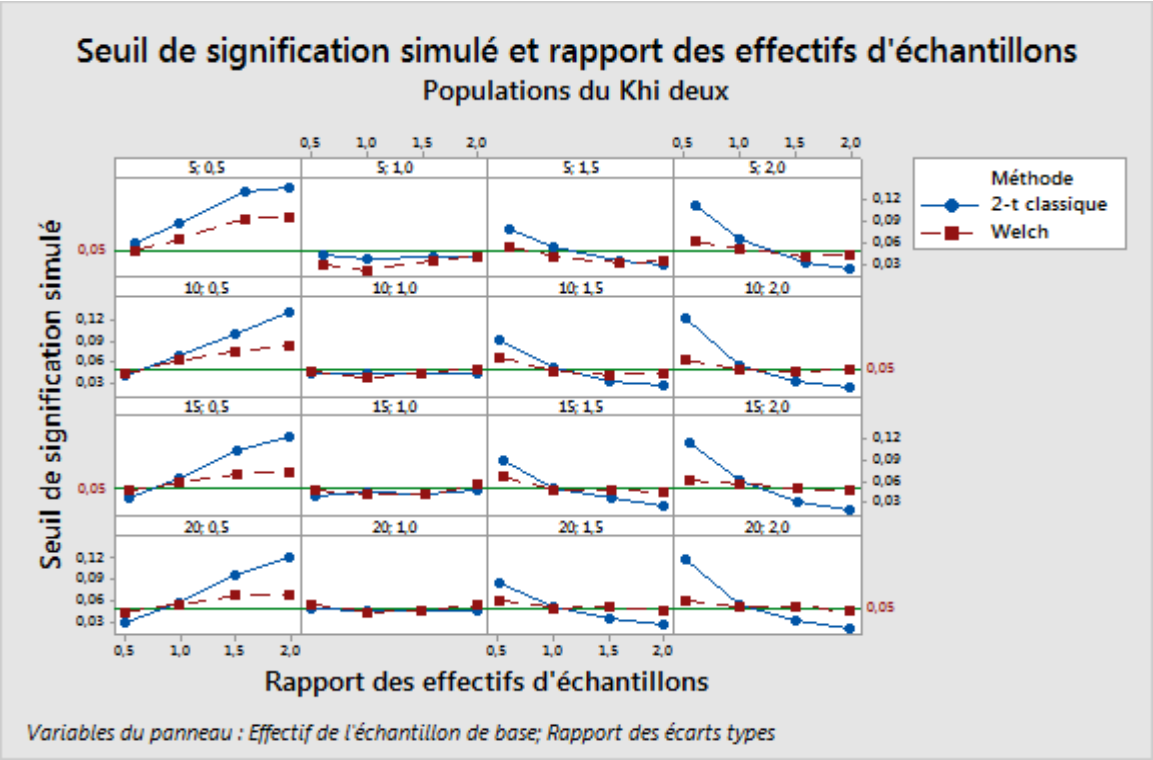

Figure 2 : Seuils de signification simulés de tests bilatéraux (test t à 2 échantillons classique et test t de Welch) pour des paires d'échantillons créées à partir de deux populations

normales avec des variances égales ou inégales, en fonction du rapport des effectifs d'échantillons.

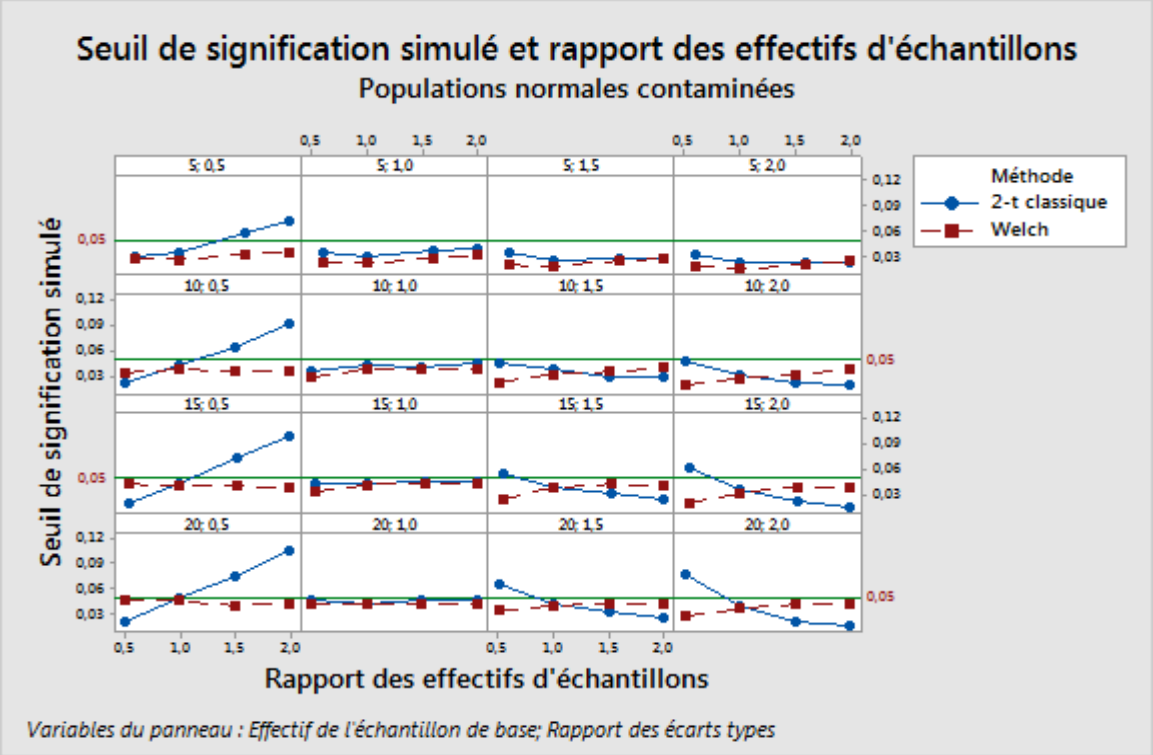

Figure 3 : Seuils de signification simulés de tests bilatéraux (test t à 2 échantillons classique et test t de Welch) pour des paires d'échantillons créées à partir de deux populations normales avec des variances égales ou inégales, en fonction du rapport des effectifs d'échantillons.

Annexe B : comparaison des fonctions puissance des deux tests

Nous souhaitions déterminer les conditions dans lesquelles les résultats de la fonction puissance du test t de Welch pouvaient être plus ou moins équivalents à ceux du test t à 2 échantillons classique.

De manière générale, les fonctions puissance des tests t (à 1 ou 2 échantillons) ont été largement étudiées et sont abordées dans de nombreuses publications (Pearson et Hartley, 1952 ; Neyman et al., 1935 ; Srivastava, 1958). Le théorème suivant indique la fonction puissance de chacune des trois hypothèses alternatives des deux plans à 2 échantillons.

THEOREME B1

Sous les hypothèses de normalité et d'égalité des variances, la fonction puissance d'un test t bilatéral à 2 échantillons de taille nominale α peut être exprimée comme fonction des effectifs d'échantillons et de la différence $\delta = \mu_1 - \mu_2$, tel que

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

où $F_{d_C,\lambda}(.)$ est la fonction de distribution cumulative (DCF) de la loi T à $d_C = n_1 + n_2 - 2$ degrés de liberté avec le paramètre de non-centralité

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

En outre, la fonction puissance associée à l'hypothèse alternative $\mu_1 > \mu_2$ est exprimée comme suit :

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

Par ailleurs, lors du test de l'hypothèse alternative $\mu_1 < \mu_2$, la puissance est exprimée comme suit :

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

Bien que le résultat du théorème ci-dessus soit connu, aucune publication n'aborde spécifiquement la fonction puissance du test t de Welch modifié. Il est possible de déduire une approximation en utilisant la fonction puissance par approximation du modèle de l'ANOVA à un facteur contrôlé (voir Kulinskaya et al., 2003). Malheureusement, cette fonction de puissance est uniquement applicable aux alternatives bilatérales. Cela dit, le plan à 2 échantillons est un cas très particulier et une méthode différente peut être adoptée pour obtenir la fonction puissance exacte du test t de Welch pour chacune des trois hypothèses alternatives. Ces fonctions sont fournies dans le théorème suivant.

THEOREME B2

Partant de l'hypothèse selon laquelle les populations sont distribuées normalement (mais pas nécessairement avec la même variance), la fonction puissance du test t bilatéral de Welch de taille nominale α peut être exprimée comme fonction des effectifs d'échantillons et de la différence $\delta = \mu_1 - \mu_2$, tel que

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

où $G_{d,\lambda}(.)$ est la fonction de distribution cumulative (DCF) de la loi T à d_W degrés de liberté tel que

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

et dotée du paramètre de non-centralité

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

Pour les alternatives unilatérales, les fonctions puissance sont exprimées de la façon suivante :

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

et

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

pour le test de l'hypothèse nulle contre les hypothèses alternatives $\mu_1 > \mu_2$ et $\mu_1 < \mu_2$, respectivement.

Le résultat est démontré dans l'Annexe D.

Avant que nous ne comparions ces deux fonctions puissance, notez que le test t à 2 échantillons classique s'appuie sur l'hypothèse supplémentaire selon laquelle les variances des populations sont égales et que, par conséquent, les fonctions puissance théorique de ces deux tests doivent être comparées lorsque cette seconde hypothèse est vérifiée pour le test t de Welch.

En théorie, nous savons que lorsque les hypothèses de normalité et d'égalité des variances, sont vérifiées,

$\pi(n_1, n_2, \delta) \geq \pi_W(n_1, n_2, \delta)$ pour toutes les valeurs n_1, n_2, δ

Le résultat suivant indique les conditions sous lesquelles les deux fonctions sont (approximativement) égales.

THEOREME B3

Lorsque les hypothèses de normalité et d'égalité des variances sont vérifiées, nous savons que :

1. Si $n_1 \sim n_2$, alors $\pi(n_1, n_2, \delta) \sim \pi_W(n_1, n_2, \delta)$ pour chaque différence δ. En particulier, si $n_1 = n_2$, alors $\pi(n_1, n_2, \delta) = \pi_W(n_1, n_2, \delta)$ pour chaque différence δ et par conséquent le test t de Welch est aussi puissant que le test t à 2 échantillons classique.
2. Si n_1 et n_2 sont petits et que $n_1 \neq n_2$, alors le test t de Welch est moins puissant que le test t à 2 échantillons classique. En revanche, si n_1 et n_2 sont élevés, alors $\pi(n_1, n_2, \delta) \sim \pi_W(n_1, n_2, \delta)$ (quelle que soit la différence entre les effectifs d'échantillons).

Ce résultat est démontré dans l'Annexe E.

Sous l'hypothèse d'égalité des variances, les paramètres de non-centralité associés aux fonctions de puissance des deux tests sont identiques. La différence entre les fonctions puissance ne peut être attribuée qu'à la différence entre leurs degrés de liberté respectifs. Nous savons qu'en théorie, sous les hypothèses énoncées, le test t classique est UPP (Uniformément le plus puissant) et que par conséquent, il présente des degrés de liberté supérieurs. Cependant, les résultats ci-dessus montrent que lorsque le plan est équilibré ou à peu près équilibré, les fonctions puissance sont identiques ou approximativement identiques. La puissance du test t classique n'est sensiblement supérieure à celle du test t de Welch que lorsque le plan est particulièrement non équilibré et que les effectifs d'échantillons sont faibles. Malheureusement, ces conditions sont également celles dans lesquelles le test t à 2 échantillons classique est particulièrement sensible à l'hypothèse d'égalité des variances, comme indiqué dans l'Annexe A. Par conséquent, la fonction puissance du test t de Welch est la plus fiable en pratique.

Les résultats du théorème B3 sont illustrés dans l'exemple suivant, où les deux populations normales ont un même écart type de 3. Les valeurs de puissance obtenues à l'aide des fonctions puissance (bilatérales) des théorèmes B1 et B2 sont calculées pour les quatre scénarios suivants :

1. Les deux échantillons sont faibles, mais ont le même effectif ($n_1 = n_2 = 10$).
2. Les deux échantillons sont faibles, mais l'un d'eux est deux fois plus grand que l'autre ($n_1 = 10, n_2 = 20$).
3. Un échantillon est faible et l'autre d'effectif modéré, mais ce dernier est quatre fois plus grand que le premier ($n_1 = 10, n_2 = 40$).
4. Un échantillon est de taille modérée et l'autre est grand, mais ce dernier est quatre fois plus grand que le premier ($n_1 = 50, n_2 = 200$).

En supposant que pour les deux tests $\alpha = 0{,}05$, les fonctions puissance sont évaluées pour chaque scénario avec la différence $\delta = 0{,}0, 0{,}5, 1{,}0, 1{,}5, 2{,}0, \ldots 5{,}0$. Les résultats apparaissent dans le tableau 5 et les fonctions sont représentées sur la figure 4.

Tableau 5 : Comparaison des fonctions puissance théorique de tests t à 2 échantillons classiques bilatéraux et de tests t de Welch bilatéraux. Les effectifs d'échantillons, n_1 et n_2, sont fixes et les fonctions puissance sont évaluées pour des différences δ allant de 0,0 à 5,0.

δ	**0,0**	**0,5**	**1,0**	**1,5**	**2,0**	**2,5**	**3**	**3,5**	**4**	**4,5**	**5,0**
$n_1 = n_2 = 10$											
$\pi(n_1, n_2, \delta)$	0,05	0,064	0,109	0,185	0,292	0,422	0,562	0,694	0,805	0,887	0,941
$\pi_W(n_1, n_2, \delta)$	0,05	0,064	0,109	0,185	0,292	0,422	0,562	0,694	0,805	0,887	0,941
$n_1 = 10, n_2 = 20$											
$\pi(n_1, n_2, \delta)$	0,05	0,070	0,132	0,239	0,383	0,547	0,703	0,828	0,913	0,962	0,986
$\pi_W(n_1, n_2, \delta)$	0,05	0,070	0,129	0,231	0,371	0,531	0,686	0,813	0,902	0,955	0,982

δ	0,0	0,5	1,0	1,5	2,0	2,5	3	3,5	4	4,5	5,0
$n_1 = 10, n_2 = 40$											
$\pi(n_1, n_2, \delta)$	0,05	0,075	0,152	0,283	0,455	0,637	0,791	0,899	0,959	0,986	0,996
$\pi_W(n_1, n_2, \delta)$	0,05	0,072	0,142	0,261	0,419	0,592	0,748	0,865	0,938	0,976	0,992
$n_1 = 50, n_2 = 200$											
$\pi(n_1, n_2, \delta)$	0,05	0,182	0,556	0,883	0,987	0,999	1,0	1,0	1,0	1,0	1,0
$\pi_W(n_1, n_2, \delta)$	0,05	0,180	0,548	0,877	0,986	0,999	1,0	1,0	1,0	1,0	1,0

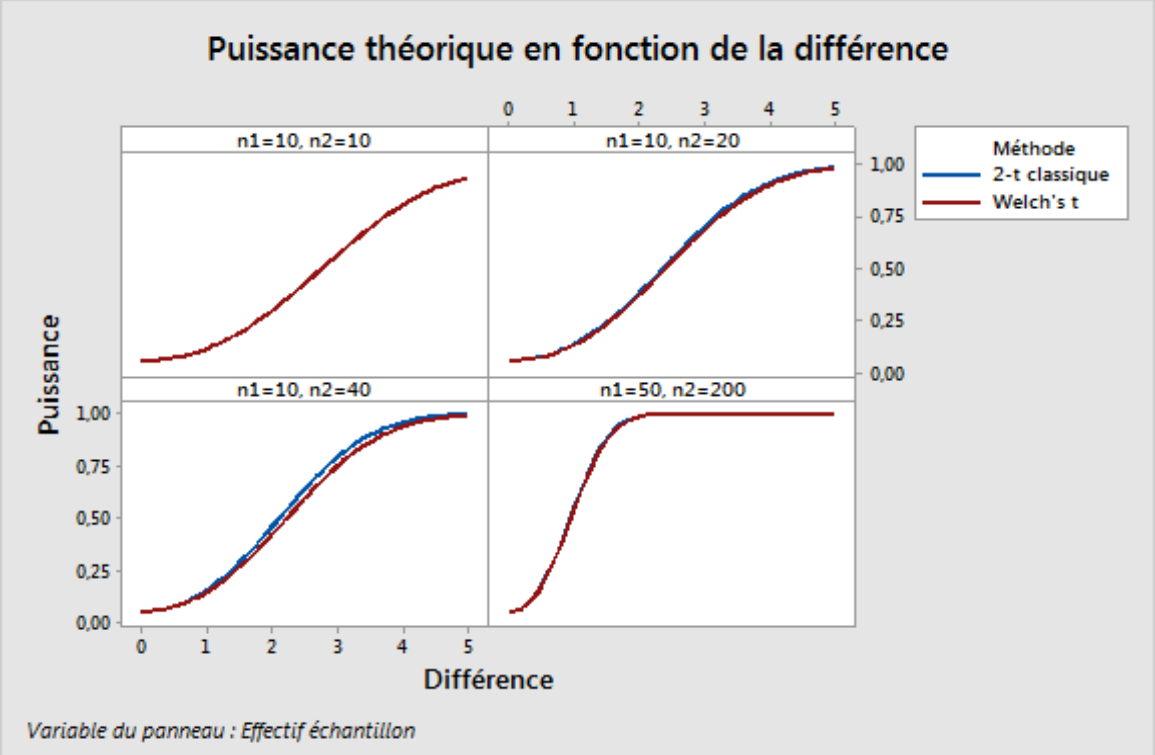

Figure 4 : Graphiques de fonctions puissance théorique de tests t à 2 échantillons classiques bilatéraux et de tests t de Welch bilatéraux en fonction de la différence δ à détecter entre les moyennes. Les deux tests utilisent $\alpha = 0{,}05$. Les populations supposées sont normales, avec un écart type de 3.

Etude par simulation B

Le but de cette étude par simulation est de comparer les niveaux de puissance associés au test t à 2 échantillons classique et ceux du test t de Welch dans des plans équilibrés où les variances sont supposées inégales. Les expériences réalisées pour ces études sont semblables à celles décrites dans l'Annexe A.

Dans le premier groupe d'expériences, nous avons créé des paires d'échantillons d'effectifs égaux, issues de populations normales avec des variances inégales. La population de base a

été définie comme $N(0,2)$ et les secondes populations normales ont été choisies de telle sorte que le rapport des écarts types $\rho = \sigma_2/\sigma_1$ soit égal à 0,5, 1,5 et 2. De la même façon, dans un deuxième groupe, les deux échantillons sont issus de populations obéissant à des lois du Khi deux avec des variances inégales (la population de base est Khi(2)). Dans la dernière série d'expériences réalisées, les paires d'échantillons sont issues de populations normales contaminées (population de base CN(0,8, 4)), comme définies précédemment dans l'Annexe A.

Pour chaque série d'expériences, nous avons calculé les niveaux de puissance simulée (avec une différence détectable δ donnée) associés à chaque test pour les effectifs d'échantillons $n = n_1 = n_2 = 5, 10, 15, 20, 25, 30$. Dans chaque expérience, nous avons calculé le niveau de puissance simulé comme la proportion de rejet de l'hypothèse nulle quand celle-ci est fausse. Pour toutes les expériences, la différence entre les moyennes a été fixée comme une unité de l'écart type de la population de base (le premier des deux échantillons). Plus précisément, nous avons défini $\delta = 1{,}0 \times \sigma_1 = 2{,}0$ car il s'agit d'une différence relativement faible pour les trois familles de lois de cette étude. Les résultats des simulations sont présentés dans le tableau 2.2 et sur les figures 2.2a, 2.2b et 2.2c.

Résultats et récapitulatif

Les résultats du tableau 6 et de la figure 4 montrent que, sous l'hypothèse d'égalité des variances, les fonctions puissance théorique sont identiques dans des plans équilibrés, comme indiqué dans le théorème 2.3. En outre, lorsque les effectifs d'échantillons sont relativement faibles mais approximativement de la même taille, les deux fonctions donnent des valeurs de puissance pratiquement égales. Ce n'est que lorsque les échantillons sont relativement faibles et que l'un des deux est environ quatre fois plus grand que l'autre que des différences notables entre les fonctions puissance commencent à apparaître (par exemple, lorsque $n_1 = 10, n_2 = 40$). Même dans ce cas, les valeurs de puissance théorique obtenues pour le test t à 2 échantillons classique ne sont que légèrement plus élevées que celles calculées pour le test t de Welch. Enfin, lorsque les plans sont particulièrement non équilibrés mais que les échantillons sont (relativement) grands, les deux fonctions puissance sont essentiellement identiques, comme indiqué dans le théorème B3.

De plus, dans des plans équilibrés présentant des variances inégales, les deux tests donnent des valeurs de puissance pratiquement identiques. Toutefois, avec de très faibles échantillons ($n < 10$), le test t à 2 échantillons classique offre des résultats légèrement meilleurs.

Tableau 6 : Comparaison des niveaux de puissance simulée du test t à 2 échantillons classique et du test t de Welch dans des plans équilibrés à variances inégales.

		Population de base : N(0,2)			**Population de base : Khi(2)**			**Population 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

		Population de base : N(0,2)			Population de base : Khi(2)			Population de base : CN(0,8, 4)		
	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
	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

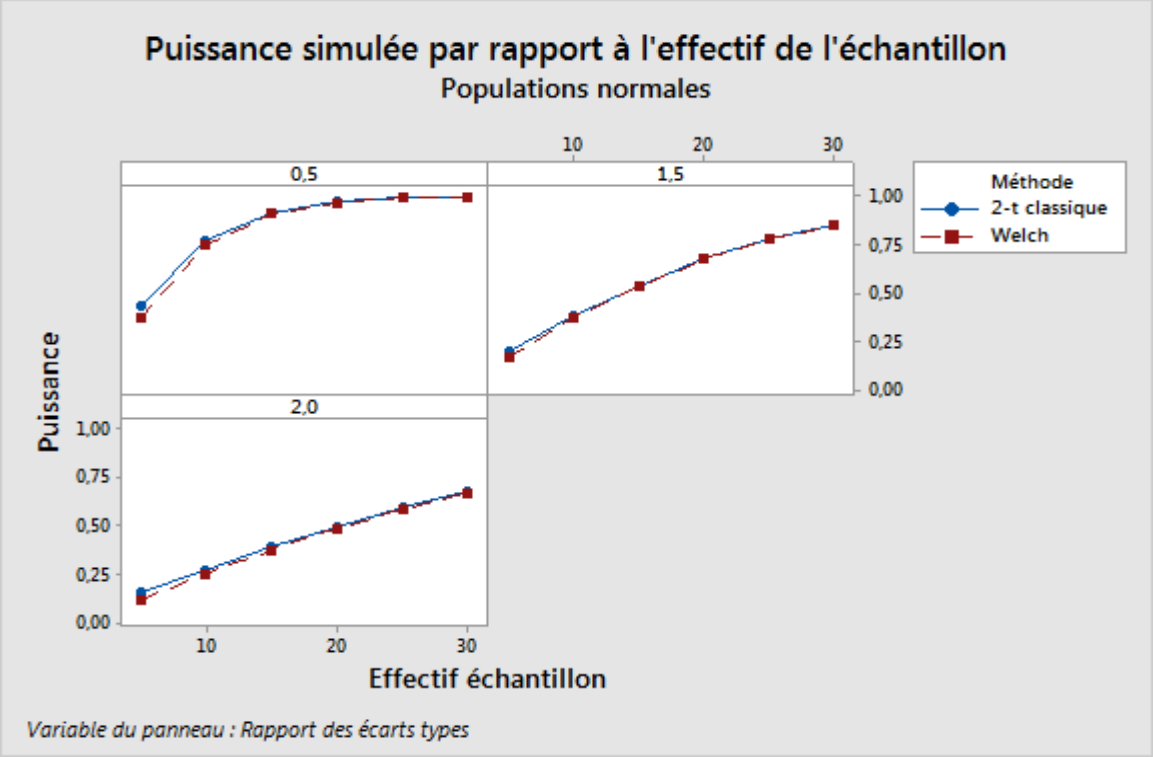

Figure 5 : Comparaison des niveaux de puissance simulée du test t à 2 échantillons classique et du test t de Welch dans des plans équilibrés à variances inégales. Les échantillons ont été créés à partir de populations normales à variances inégales, de telle sorte que les rapports des écarts types soient de 0,5, 1,5 et 2,0.

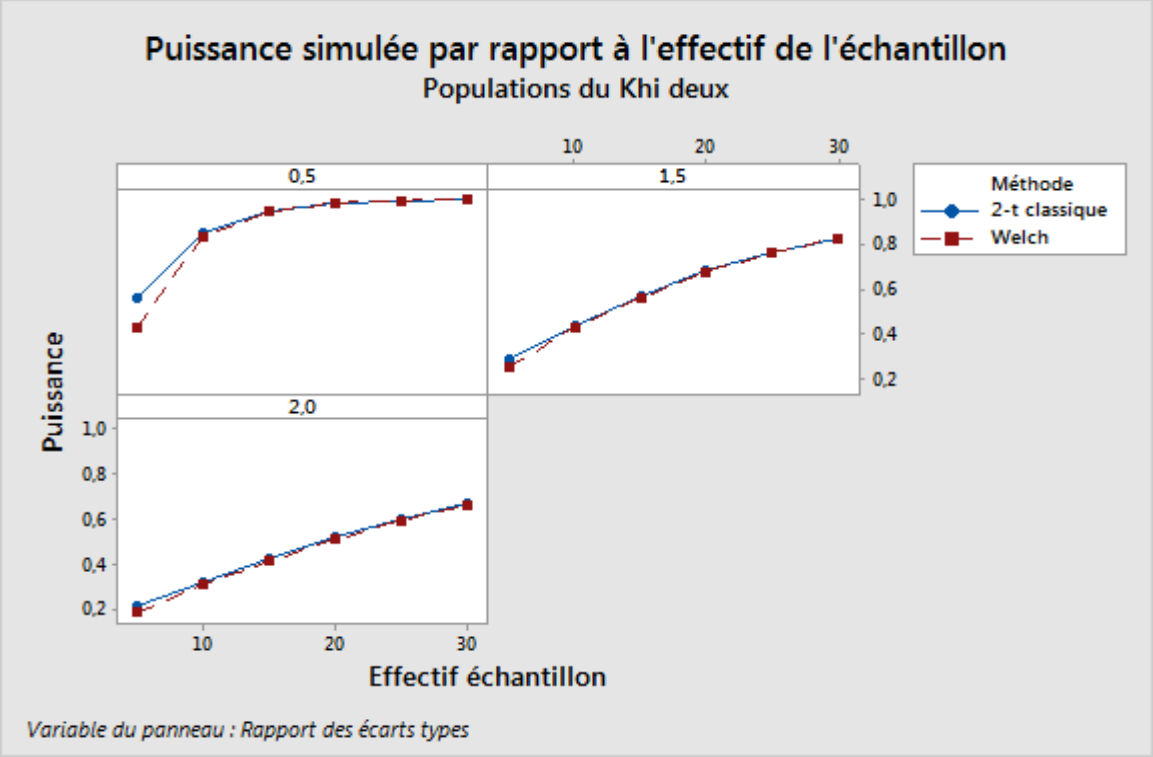

Figure 6 : Comparaison des niveaux de puissance simulée du test t à 2 échantillons classique et du test t de Welch dans des plans équilibrés à variances inégales. Les échantillons ont été créés à partir de populations obéissant à une loi du Khi deux, avec des variances inégales, de telle sorte que les rapports des écarts types soient de 0,5, 1,5 et 2,0.

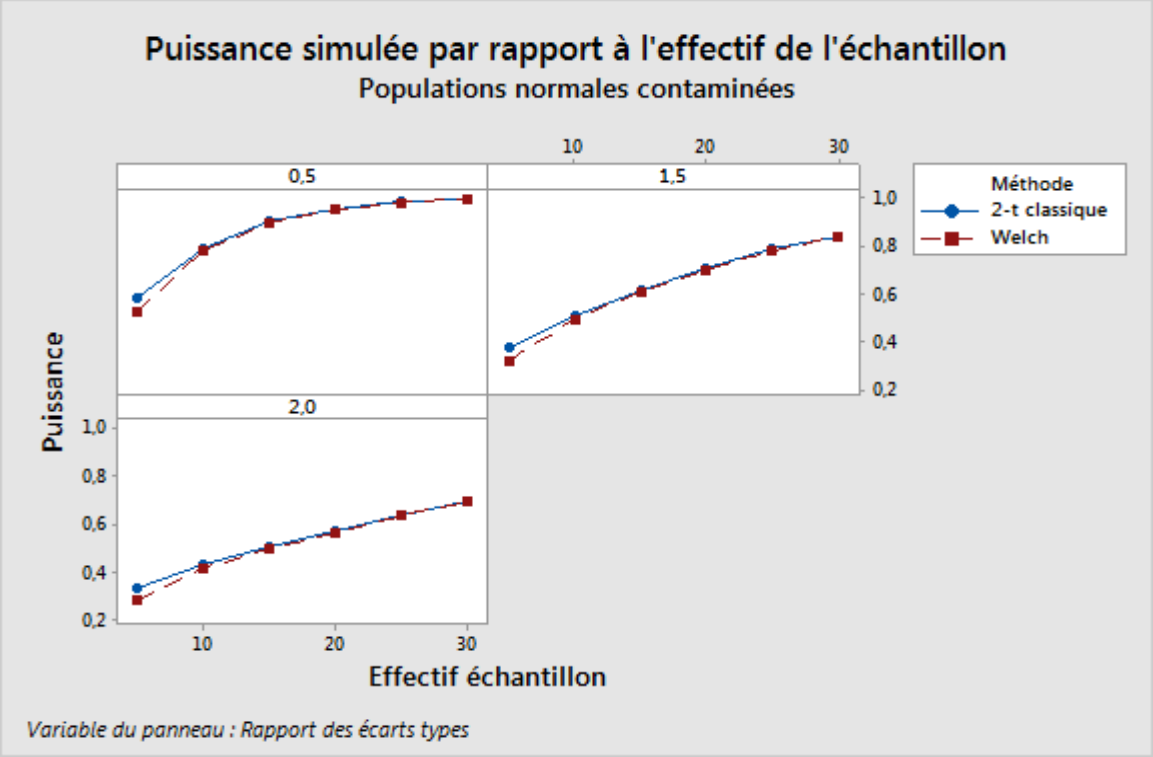

Figure 7 : Comparaison des niveaux de puissance simulée du test t à 2 échantillons classique et du test t de Welch dans des plans équilibrés à variances inégales. Les échantillons ont été créés à partir de populations normales contaminées à variances inégales, de telle sorte que les rapports des écarts types soient de 0,5, 1,5 et 2,0.

Annexe C : puissance et effectif d'échantillon, sensibilité à la normalité

Dans l'Assistant, l'analyse de la puissance utilisée pour comparer les moyennes de deux populations a recours à la fonction puissance du test t de Welch. Si cette fonction est sensible à l'hypothèse de normalité sous laquelle elle est établie, l'analyse de la puissance peut amener à des conclusions erronées. C'est pourquoi nous avons mené une étude par simulation pour examiner la sensibilité de cette fonction à l'hypothèse normale. La sensibilité est évaluée en étudiant la cohérence entre les niveaux de puissance simulée et ceux calculés à l'aide de la fonction puissance théorique, lorsque les échantillons sont créés à partir de lois non normales. La loi normale sert de population de contrôle, car, d'après le théorème B2, les niveaux de puissance simulée et théorique sont les plus proches lorsque les échantillons sont créés à partir de lois normales.

Etude par simulation C

L'étude est menée en trois parties avec trois lois différentes : la loi normale, la loi du Khi deux et la loi normale contaminée. Pour plus de détails, reportez-vous à l'Annexe A. Pour chaque partie de l'étude, nous calculons la puissance simulée (celle obtenue avec les effectifs d'échantillons n_1 et n_2 pour une différence détectable δ donnée) comme la proportion de rejet de l'hypothèse nulle quand celle-ci est fausse. Dans tous les cas, la différence à détecter est fixée comme une unité de l'écart type de la population de base. Ainsi $\delta = 1{,}0 \times \sigma_1 = 2{,}0$ pour les trois familles de lois de cette étude. Nous calculons également les valeurs de puissance théorique à partir du test t de Welch pour effectuer la comparaison.

Résultats de la simulation et récapitulatif

Les résultats indiquent que, pour des échantillons relativement faibles, la fonction puissance du test t de Welch est robuste à l'hypothèse de normalité. De manière générale, lorsque l'effectif minimal des deux échantillons est aussi faible que 15, les valeurs de puissance simulée sont proches du niveau théorique cible correspondant (voir les tableaux 7 à 10 et les figures 8 à 10).

Les tableaux 7 à 10 indiquent les niveaux de puissance simulée d'un test t de Welch bilatéral avec $\alpha = 0{,}05$ pour des paires d'échantillons issues de populations normale, asymétrique (Khi deux) et normale contaminée. Les paires d'échantillons obéissent à la même famille de loi, mais les variances des populations parent ne sont pas nécessairement égales. Nous avons calculé les valeurs de puissance théorique pour effectuer la comparaison.

Tableau 7 : Niveaux de puissance simulée d'un test t de Welch bilatéral avec $\alpha = 0{,}05$ pour n=5.

			Population de base : N(0,2)				**Population de base : Khi(2)**				**Population 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.	,288	,158	,113	,091	,432	,305	,211	,149	,361	,257	,234	,220
		Cible	,353	,192	,116	,092	,353	,192	,116	,092	,353	,192	,116	,092
5	1,0	Obs.	,370	,252	,169	,121	,427	,334	,248	,189	,522	,380	,319	,284
		Cible	,389	,286	,190	,137	,389	,286	,190	,137	,389	,286	,190	,137
8	1,6	Obs.	,387	,326	,242	,179	,427	,364	,286	,225	,573	,453	,374	,319
		Cible	,400	,345	,260	,193	,400	,345	,260	,193	,400	,345	,260	,193
10	2,0	Obs.	,390	,351	,272	,208	,421	,373	,296	,235	,590	,483	,394	,336
		Cible	,402	,364	,291	,223	,402	,364	,291	,223	,402	,364	,291	,223

Tableau 8 : Niveaux de puissance simulée d'un test t de Welch bilatéral avec $\alpha = 0{,}05$ pour n=10

			Population de base : N(0,2)				**Population de base : Khi(2)**				**Population 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.	,651	,346	,197	,131	,768	,493	,320	,221	,689	,484	,404	,358
		Cible	,666	,364	,206	,139	,666	,364	,206	,139	,666	,364	,206	,139
10	1,0	Obs.	,742	,556	,369	,254	,831	,612	,430	,308	,776	,619	,496	,419
		Cible	,745	,562	,337	,259	,745	,562	,337	,259	,745	,562	,337	,259
15	1,5	Obs.	,765	,641	,483	,358	,865	,679	,511	,377	,792	,679	,547	,456
		Cible	,767	,643	,483	,352	,767	,643	,483	,352	,767	,643	,483	,352
20	2	Obs.	,774	,683	,549	,417	,898	,737	,565	,448	,797	,716	,596	,490
		Cible	,777	,686	,551	,422	,777	,686	,551	,422	,777	,686	,551	,422

Tableau 9 : Niveaux de puissance simulée d'un test t de Welch bilatéral avec $\alpha = 0{,}05$ pour n=15.

			Population de base : N(0,2)				**Population de base : Khi(2)**				**Population 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	0,53	Obs.	,857	,569	,342	,229	,871	,651	,421	,293	,853	,632	,505	,428
		Cible	,861	,568	,338	,221	,861	,568	,338	,221	,861	,568	,338	,221
15	1,0	Obs.	,906	,745	,535	,368	,942	,763	,563	,415	,891	,760	,611	,500
		Cible	,910	,753	,541	,379	,910	,753	,541	,379	,910	,753	,541	,379
23	1,53	Obs.	,928	,831	,667	,502	,975	,858	,676	,517	,898	,825	,698	,572
		Cible	,925	,830	,670	,509	,925	,830	,670	,509	,925	,830	,670	,509
30	2,0	Obs.	,933	,861	,737	,589	,984	,903	,750	,598	,902	,847	,742	,619
		Cible	,931	,863	,736	,589	,931	,863	,736	,589	,931	,863	,736	,589

Tableau 10 : Niveaux de puissance simulée d'un test t de Welch bilatéral avec $\alpha = 0{,}05$ pour n=20

			Population de base : N(0,2)				**Population de base : Khi(2)**				**Population 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.	,938	,687	,426	,275	,920	,698	,486	,333	,923	,716	,568	,476
		Cible	,941	,686	,424	,277	,941	,686	,424	,277	,941	,686	,424	,277
20	1,0	Obs.	,971	,866	,672	,485	,981	,858	,670	,506	,952	,856	,696	,567
		Cible	,971	,869	,673	,489	,971	,869	,673	,489	,971	,869	,673	,489
30	1,5	Obs.	,977	,923	,791	,629	,995	,932	,785	,631	,960	,908	,798	,662
		Cible	,978	,922	,791	,628	,978	,922	,791	,628	,978	,922	,791	,628
40	2,0	Obs.	,983	,950	,858	,724	,998	,966	,864	,726	,958	,929	,845	,725
		Cible	,981	,945	,854	,719	,981	,945	,854	,719	,981	,945	,854	,719

Lorsque les deux échantillons sont issus de populations normales, les valeurs de puissance simulée concordent avec les valeurs de puissance théorique, même avec de très faibles échantillons. Comme indiqué dans la figure 7, les courbes de puissance théorique et simulée sont pratiquement indissociables. Ces résultats corroborent le théorème B2.

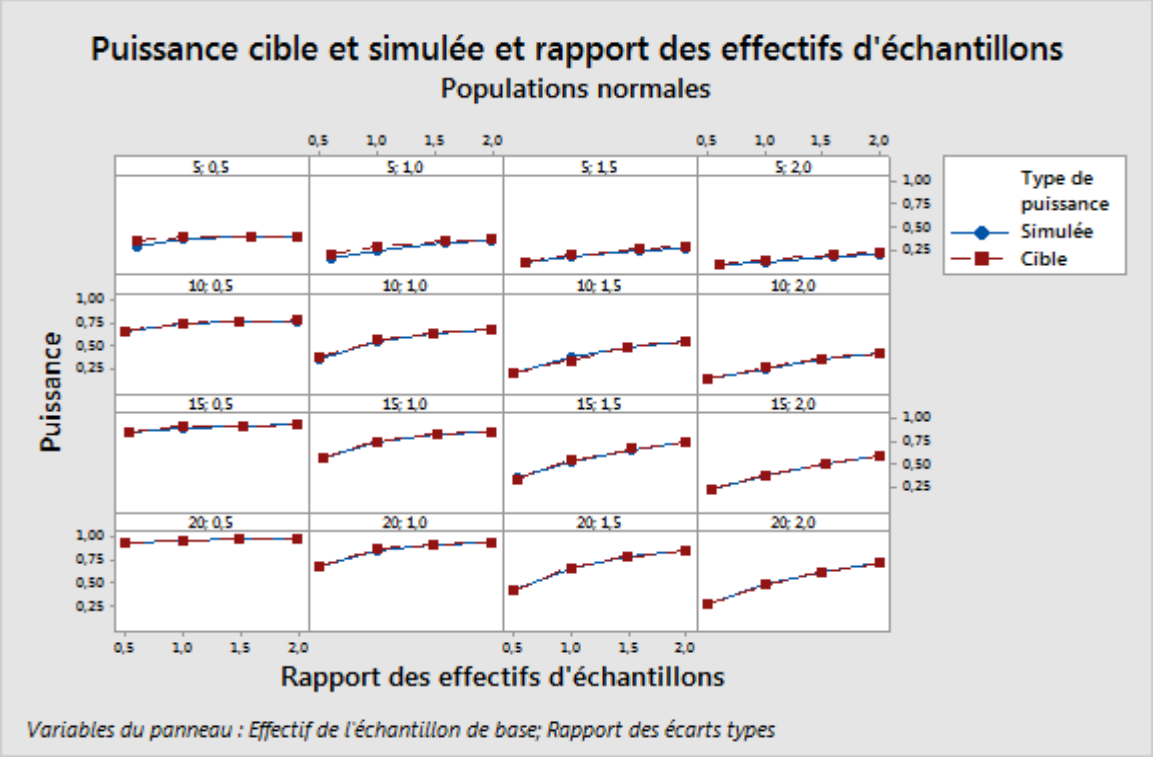

Figure 8 : Niveaux de puissance simulée et théorique cible d'un test t à 2 échantillons de Welch bilatéral avec $\alpha = 0{,}05$ pour des paires d'échantillons issues de deux populations normales à variances égales ou inégales, en fonction du rapport des effectifs d'échantillons.

Lorsque les échantillons sont créés à partir de lois du Khi deux asymétriques, les valeurs de puissance simulée sont plus élevées que celles de puissance théorique pour de petits échantillons ; toutefois, plus les effectifs d'échantillons augmentent, plus les valeurs de puissance se rapprochent. La figure 9 indique que les courbes de puissance théorique cible et de puissance simulée sont toujours proches lorsque l'effectif minimal des deux échantillons est d'au moins 10. Cela montre bien que les données asymétriques n'ont pas d'effet notable sur la fonction puissance du test t de Welch, même lorsque les échantillons sont relativement faibles.

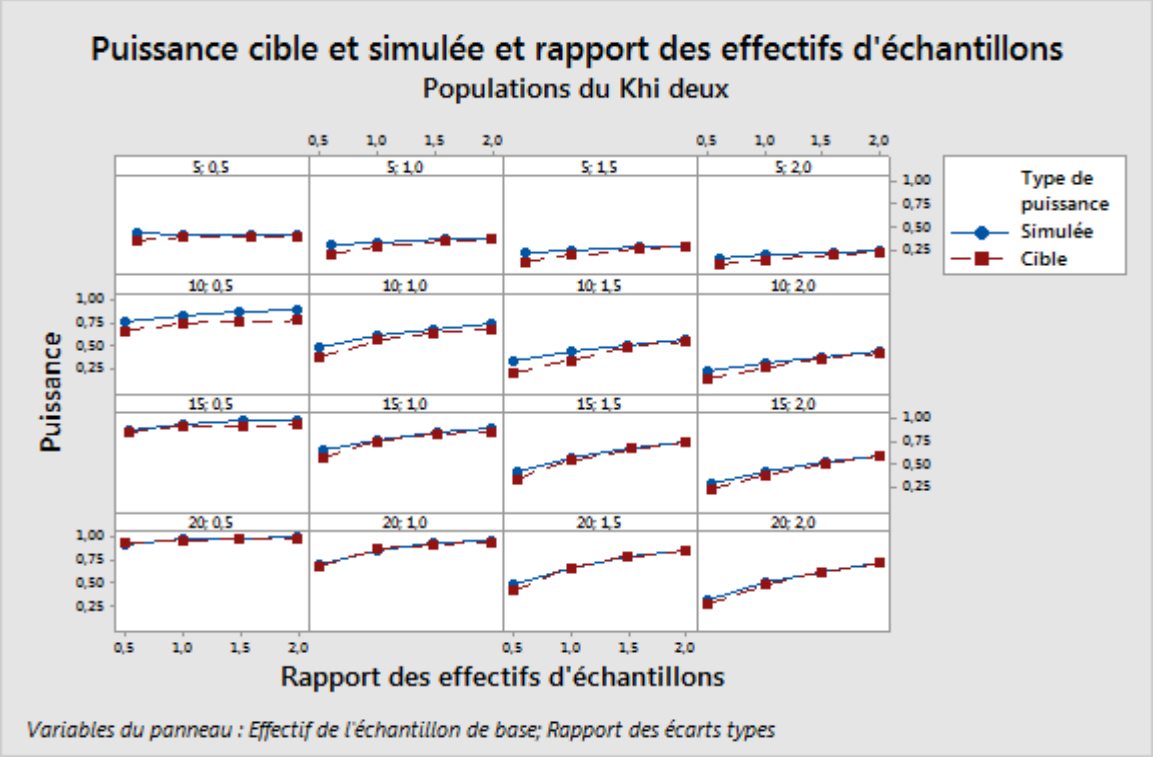

Figure 9 : Niveaux de puissance simulée et théorique cible d'un test t à 2 échantillons de Welch bilatéral avec $\alpha = 0{,}05$ pour des paires d'échantillons issues de deux populations normales à variances égales ou inégales, en fonction du rapport des effectifs d'échantillons.

En outre, les valeurs aberrantes n'influent généralement sur la fonction puissance que lorsque les effectifs d'échantillons sont très faibles. En général, en présence de valeurs aberrantes, les valeurs de puissance simulée ont tendance à être légèrement supérieures aux valeurs de puissance théorique cible. C'est ce que montre la figure 10 : les courbes de puissance simulée et théorique ne sont raisonnablement proches que lorsque l'effectif minimal d'échantillon est d'au moins 15.

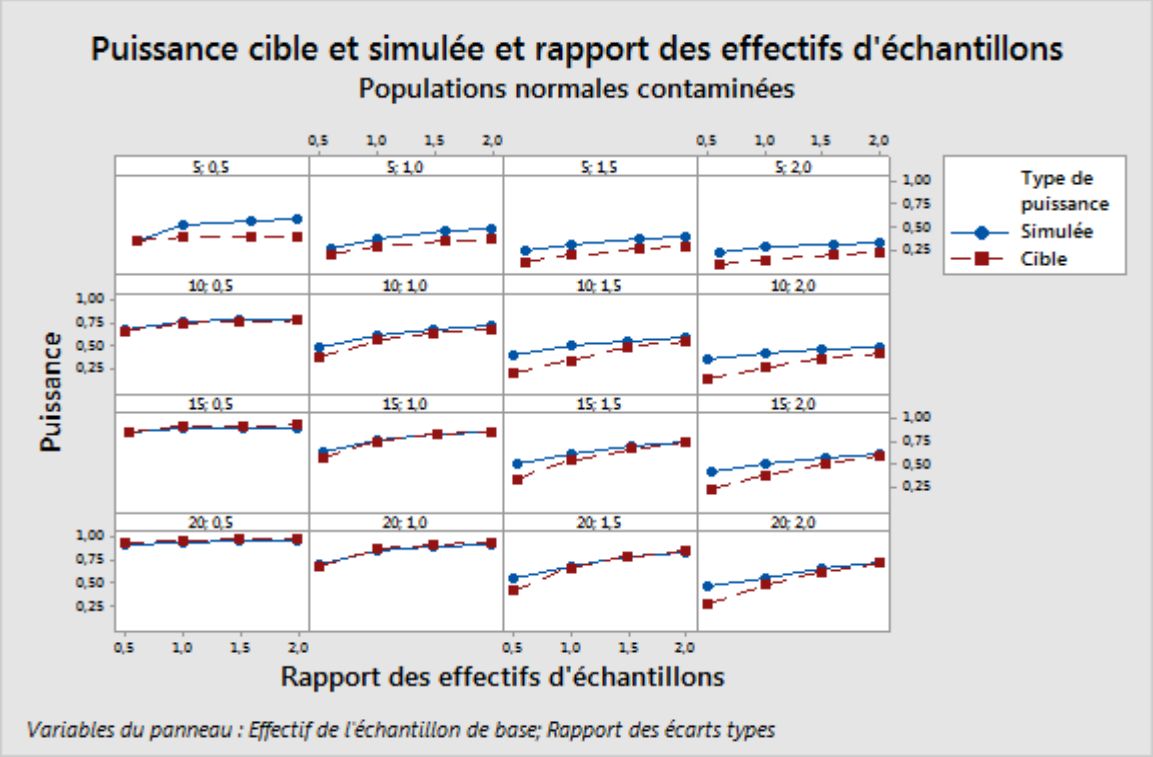

Figure 10 : Niveaux de puissance simulée et théorique cible d'un test t à 2 échantillons de Welch bilatéral avec $\alpha = 0{,}05$ pour des paires d'échantillons issues de deux populations normales à variances égales ou inégales, en fonction du rapport des effectifs d'échantillons.

Annexe D : démonstration du théorème B2

Pour le modèle à 2 échantillons, la méthode de Welch utilisée pour calculer la loi de la statistique de test

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

sous l'hypothèse nulle s'appuie sur une approximation de la loi

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

comme étant proportionnelle à une loi du Khi deux. Plus spécifiquement,

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

est approximativement distribué comme une loi du Khi deux à d_W degrés de liberté, où

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

(Notez que dans une configuration à 1 échantillon, cela se réduit au résultat classique et bien connu selon lequel $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$)

Imaginons que l'on teste l'hypothèse nulle H_o: $\mu_1 = \mu_2$ (ce qui revient à dire que $\delta = 0$) contre l'hypothèse alternative H_A: $\mu_1 \neq \mu_2$ (ce qui revient à dire que $\delta \neq 0$)

Sous l'hypothèse nulle, la fonction puissance

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

où t_d^α désigne le percentile supérieur 100 α de la loi T à d degrés de liberté.

Sous l'hypothèse alternative,

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

obéit approximativement à une loi T non centrée à d_W degrés de liberté et a le paramètre de non-centralité

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

car, comme indiqué précédemment

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

est approximativement distribué comme une loi du Khi deux à d_W degrés de liberté, et

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

est distribué comme une loi normale standard.

Par conséquent, sous l'hypothèse alternative,

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

où $G_{d_W,\lambda}(.)$ est la fonction de distribution cumulative (DCF) de la loi T à d_W degrés de liberté dotée du paramètre de non-centralité λ, tel qu'exprimé auparavant.

Annexe E : démonstration du théorème B3

Tout d'abord, notez que d_W peut aussi être exprimé de la façon suivante :

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

où $\rho = \sigma_1/\sigma_2$.

De même, le paramètre de non-centralité associé à la fonction puissance du test t de Welch peut être réécrit comme suit :

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

Sous l'hypothèse d'égalité des variances, les paramètres de non-centralité associés aux fonctions puissance du test t à 2 échantillons classique et du test t de Welch coïncident. En d'autres termes

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

où σ est la variance commune aux deux populations. Par conséquent, la seule différence entre les fonctions puissance des deux tests tient à leurs degrés de liberté respectifs. Cependant, sous l'hypothèse d'égalité des variances, la valeur des degrés de liberté associés à la fonction puissance du test t de Welch devient

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

D'après le théorème 1, la valeur des degrés de liberté liés à la fonction puissance du test t à 2 échantillons classique est $d_C = n_1 + n_2 - 2$. Après quelques manipulations algébriques, nous obtenons

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

Le fait que $d - d_W \geq 0$ n'est pas surprenant, car nous savons que sous l'hypothèse d'égalité des variances, le test t à 2 échantillons classique est uniformément le plus puissant (UPP) ; il est donc logique que le nombre de degrés de liberté associés à sa fonction puissance soit plus élevé.

Ainsi, si $n_1 \sim n_2$, alors $d \sim d_W$ et par conséquent, les fonctions de puissance présentent le même ordre de grandeur. En particulier, les fonctions puissance des deux tests sont identiques lorsque $n_1 = n_2$. Cela démontre la première partie du théorème 2.3.

Si $n_1 \neq n_2$, alors $d_C - d_W > 0$, ce qui signifie que le test t de Welch est moins puissant que le test t à 2 échantillons classique.

En outre, si les échantillons sont élevés, en d'autres termes si $n_1 \to \infty$ et $n_2 \to \infty$, alors $d_C \to \infty$ et $d_W \to \infty$, de sorte que la loi asymptotique de la statistique de test associée aux deux tests est la loi normale standard. Par conséquent, les tests sont asymptotiquement équivalents et ont la même fonction puissance asymptotique.