

WELCH CORRECTION ON STUDENT T TEST

Reliability and power analysis for Welch correction on Student T test in R

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \quad t = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2 \cdot \nu_1} + \frac{s_2^4}{N_2^2 \cdot \nu_2}}$$

I. Introduction et généralités

Les statistiques sont omniprésentes en science. Ouvrez n'importe quelle revue scientifique et vous en trouverez. Elles offrent aux scientifiques des outils extraordinaires pour l'analyse de données. En revanche l'étude de ces outils de leur manipulation attire bien moins ces mêmes scientifiques. Comme le souligne les livres de Darrell Huff¹ et Alex Reinhart², il n'est pas rare de voir des mauvaises utilisations des statistiques, que cela soit de manière malicieuse ou non.

Ainsi, selon John P. A. Ioannidis³, plus de la moitié des articles scientifiques présenteraient des résultats faux. Par ces fiabilités incertaines, par des expériences menant à des tests aux puissances insuffisantes ou encore par des analyses flexibles. Non pas que tous ces scientifiques seraient mûs par de mauvaises intentions mais plutôt que les manuels ne leur auraient pas été fournis avec les outils. Essayez de construire un meuble IKEA sans manuel et vous vous retrouverez sûrement plus à une œuvre à la Numérobis ou aux constructions en LEGO de votre enfant.

Nous nous intéresseront ici au test T de Student, l'un des tests les plus utilisés, à l'analyse simple mais qui peut cependant mener à des surprises⁶. Ce test permet de formuler une hypothèse relative à une comparaison de moyennes. Cela va nous permettre de comparer la distribution de deux populations à trois conditions :

- ◆ l'indépendance des individus
- ◆ les populations doivent suivre une distribution normale
- ◆ Il doit y avoir homogénéité des variances des échantillons

Au test T de Student, nous pouvons appliquer la correction de Welch. En effet, celle-ci nous permet de corriger l'absence d'homogénéité des variances de nos échantillons. Cela ne corrige cependant pas tout. Il faudrait alors vérifier préalablement l'homogénéité des variances grâce à un test de Snedecor. Cela montre alors une certaine fragilité de la fiabilité et de la puissance de la correction de Welch.

Parmi les erreurs pouvant survenir pendant une analyse statistique nous nous intéresserons aux erreurs inhérentes au test et non à la formulation des hypothèses ou à l'analyse des résultats. Lors d'une prise de décision concernant l'hypothèse nulle, nous pouvons être confronté à deux types d'erreurs :

- ◆ rejeter l'hypothèse nulle alors qu'elle est vraie - erreur de type I (notée α)
- ◆ rejeter l'hypothèse alternative alors qu'elle est vraie - erreur de type II (notée β)





		Réalité	
		L'hypothèse nulle est vraie	L'hypothèse alternative est vraie
Décision	L'hypothèse nulle est vraie	Juste $1 - \beta$ 	Erreur de type II β 
	L'hypothèse alternative est vraie	Erreur de type I α 	Juste $1 - \alpha$ 

Tableau 1 : Les types d'erreur - depuis un tableau de [datasciencedojo](https://datasciencedojo.github.io/)

L'erreur α correspond au degré de signification, expérimentateur peut agir dessus car c'est ce degré qui va déterminer l'analyse de la p-value. Si par exemple, on retient un α de 0.05, il y aura alors 5 chances sur 100 pour que l'on rejette l'hypothèse nulle alors que l'on aurait pas dû. Lors d'une prise de décision on est alors confiant à 95% ($1 - \alpha$).

L'erreur β correspond quant à elle à la puissance du test ($1 - \beta$), son estimation est plus complexe est bien souvent oubliée. Si par exemple, le test utilisé et les conditions expérimentales conduisent à une puissance de 80% alors il y aura 20 chances sur 100 pour que l'on n'arrive pas à rejeter l'hypothèse nulle quand on aurait dû.

Nous nous demanderons alors comment une mauvaise manipulation et un manque de connaissances en statistiques peuvent mener à des conclusions fallacieuses lors de test T de Student et si la correction de Welch permet de corriger certaines erreurs, et ça pour des échantillons où $n \leq 30$.

II. Fiabilité

A. Objectifs

Nous nous intéresserons dans un premier temps au cas de l'erreur de type I. Si nous prenons le cas d'un sirop antitussif ce type d'erreur reviendrait à dire que le sirop a un effet alors qu'il n'en a pas. On pourrait mettre sur le marché un sirop sans effet. Cette erreur est bien sûr influencée par les conditions expérimentales mais il est aussi possible de la diminuer en diminuant l' α , le seuil d'acceptation de la p-value, habituellement à 0.05. C'est pourquoi on entend souvent parler de cet α dans les articles.

Nous avons mis ici en place une simulation d'analyse de données où des populations normales sont générées aléatoirement. La moyenne reste toujours la même ($\mu=0$) entre les deux populations et entre chaque expérimentation. La taille (n) de ces populations varie entre chaque expérimentation et au sein de chacune l'écart-type des deux populations varie. L'une des deux populations a un écart-type stable de 1 et l'autre verra son écart-type de 1 multiplié par un facteur compris entre 1 et 4. Nous générons ensuite pour chacun de ces paramètres 10000 populations et testons si les résultats obtenus sont significatifs (p-value < 0.05) et cela avec ou sans la correction de Welch.

B. Résultats

A travers les résultats [Fig1](#), nous remarquons que l'on reste autour de 5% de cas significatif malgré des écarts plus prononcés sans correction de Welch, notamment dans les populations les plus petites. Ainsi, plus n augmente, plus les courbes sans et avec correction de Welch se rapprochent. Par exemple avec $n = \{2, 8\}$, nous voyons un écart important entre les deux courbes, tandis que qu'avec $n = 30$, nous remarquons un écart beaucoup moins important. Quelque soit la situation on retrouve plus ou moins l' α annoncé.

Avec la [Fig2](#) nous observons la variation de p-values mesurée avec différentes tailles d'échantillons et différents écarts de moyennes (représentatifs de l'effet du sirop). Ainsi on voit qu'ici la p-value est autant influencée par la taille des échantillons que par l'écart de moyenne. On observe aussi que peu sont significatives, ainsi avec des échantillons aussi petits, l'écart doit être prononcé pour l'observer.

Le graphique utilisé ici et les données utilisées proviennent des fichiers R (test_reliability.R) et Python (plotting.py et pvalue_plot.py). Les données sont aussi disponibles dans le classeur ReliabilityR.xlsx [c](#).

C. Interprétation

Dans le cas d'échantillons aux tailles réduites la correction de Welch se révèle particulièrement efficace. Tout en ayant un effet moins important lorsque n se rapproche de 30, en effet c'est pour cela qu'en augmentant en taille on peut assumer que la distribution se rapproche d'une distribution normale. La correction de Welch permet ainsi de conserver une fiabilité proche de l' α . En permettant de rectifier une certaine non homogénéité des variances il paraît alors normal qu'elle prouve son efficacité dans ce cadre là.

Dans le deuxième graphique nous retrouvons l'idée que la p-value n'est pas un reflet de l'effet mesuré. Ainsi dans le cas du sirop, une p-value faible ne veut pas dire que le sirop a un grand effet. La p-value n'est pas non plus un reflet de la certitude que nous devons avoir à rejeter l'hypothèse nulle mais plutôt notre surprise si jamais il n'y avait en réalité aucun effet.

III. Puissance

A. Objectifs

La puissance d'un test est la probabilité qu'elle distingue un effet d'une certaine taille de la chance pure. On cherche généralement une puissance statistique supérieure à 0,8, ce qui correspond à une chance de 80% de conclure qu'il y a un effet réel. Cependant, peu de scientifiques effectuent ce calcul, et ainsi très peu d'articles ne mentionnent la puissance statistique de leurs tests. C'est pourtant une valeur utile pour appuyer l'analyse des résultats obtenus.

Le protocole [4.5](#) est similaire au précédent à la différence qu'ici c'est la moyenne qui varie et l'écart-type qui reste constant. Ainsi l'une des populations garde une moyenne $\mu=0$ lorsque l'autre a une distance Δ comprise entre 0.2 et 2 ajoutée à cette moyenne.

B. Résultats

Avec les résultats [Fig3](#) [Fig4](#), nous remarquons que seules les courbes à $n = 2$ nous permettent d'observer une réelle différence entre sans et avec correction de Welch. A partir de $n = 6$ les deux courbes se confondent. En plus de cela on observe que les tests peinent à montrer une puissance de 80% dans des cas aux effectifs réduits..

Les graphiques utilisés ici et les données utilisées proviennent des fichiers R (test_power.R) et Python (plotting.py et power_plot.py). Les données sont aussi disponibles dans le classeur PowerR.xlsx [c](#).

C. Interprétation

La correction de Welch ne semble pas avoir de grande incidence sur la puissance du test T de Student. Effectivement elle permet seulement de corriger une non homogénéité des variances, c'est à dire des écarts entre les écarts types des populations (voir test de fiabilité).

En reprenant l'exemple précédent du sirop antitussif le type d'erreur décrit peut sembler à premier abord d'une moindre importance. Ainsi l'erreur de type II reviendrait ici à dire que le sirop n'a pas d'effet alors qu'il en a un. Au moins un sirop sans effet n'a pas été mis sur le marché comme précédemment. Cependant cette erreur entraînera des coûts de recherche supplémentaires qui auraient pu être évités. Mais au-delà de ça, nous aurions pu vouloir tester les possibles effets secondaires de ce sirop. Une erreur ici aurait alors montré aucun effet secondaire alors qu'il y en avait bien.

IV. Conclusion

Nous pouvons alors conclure que la correction de Welch a un effet sur la fiabilité du test, mais ne semble pas avoir un grand impact sur la puissance du test, avec des échantillons à $n \leq 30$. La correction de Welch n'est donc pas un ingrédient miracle et cela ne permet pas de régler tout les problèmes liés à la fiabilité et la puissance du test. De plus, nous avons vu que plus la taille des échantillons se rapproche de $n = 30$, moins la correction de Welch a un effet important.

Ainsi espérer régler certaines prises de liberté pendant les expérimentations pendant les tests statistiques peut paraître vain. Les p-values obtenues et les analyses en découlant auront un grand risque de ne pas refléter la réalité.

V. Annexe

Figure 1 : Analyse de Fiabilité \uparrow

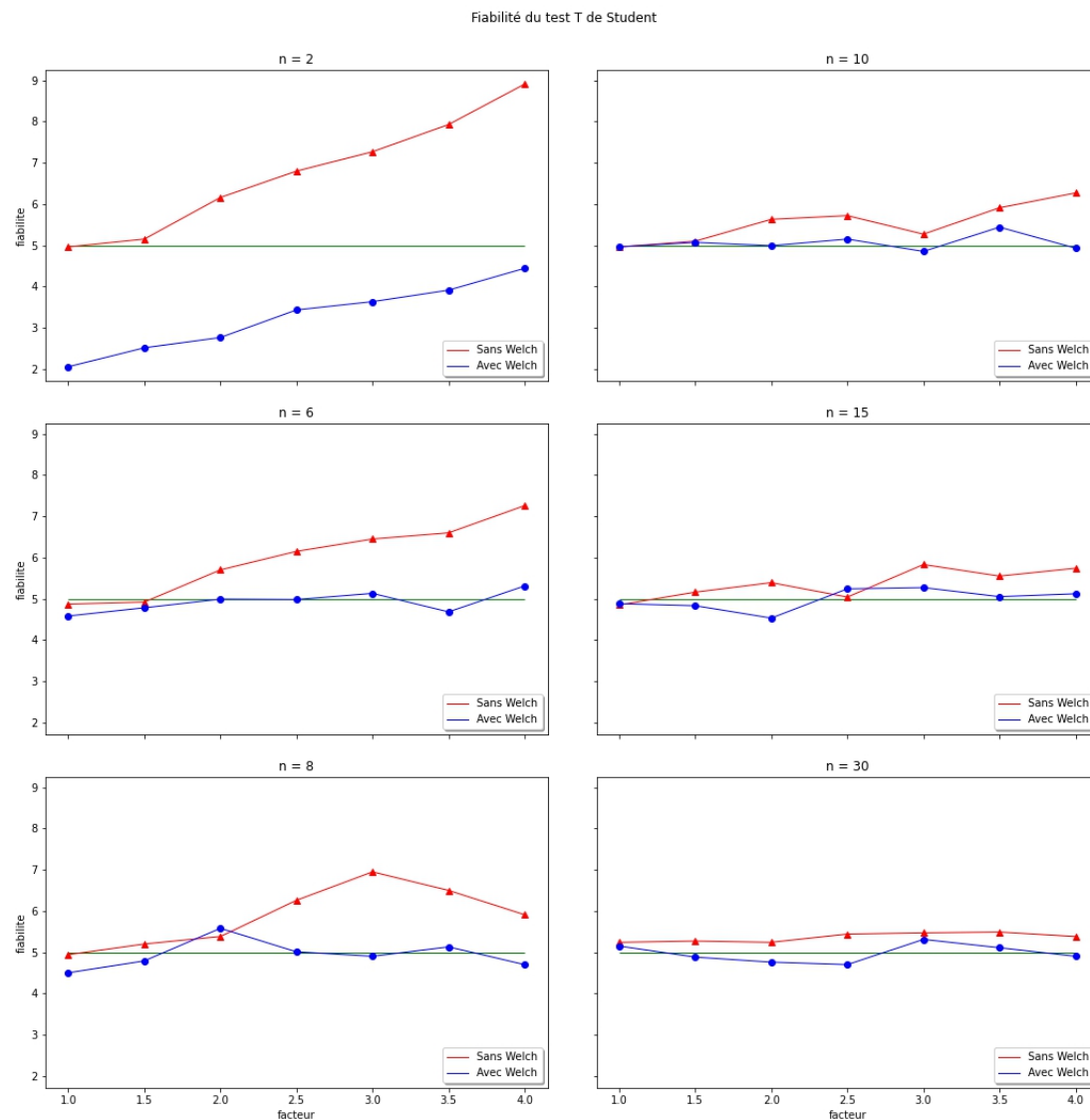


Figure 2 : représentation 3D de la variation des p-values en fonction de la taille de l'échantillon (n) et de l'écart de moyenne (δ) \uparrow

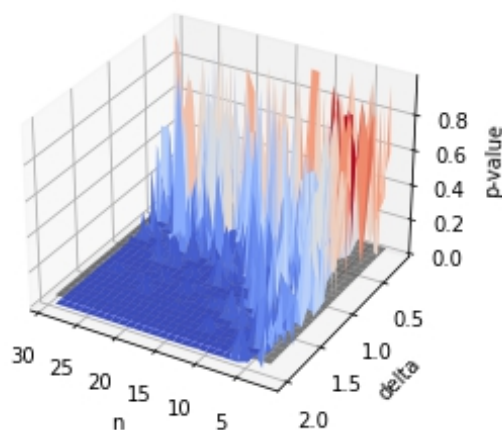


Figure 3 : Analyse de Puissance ↗

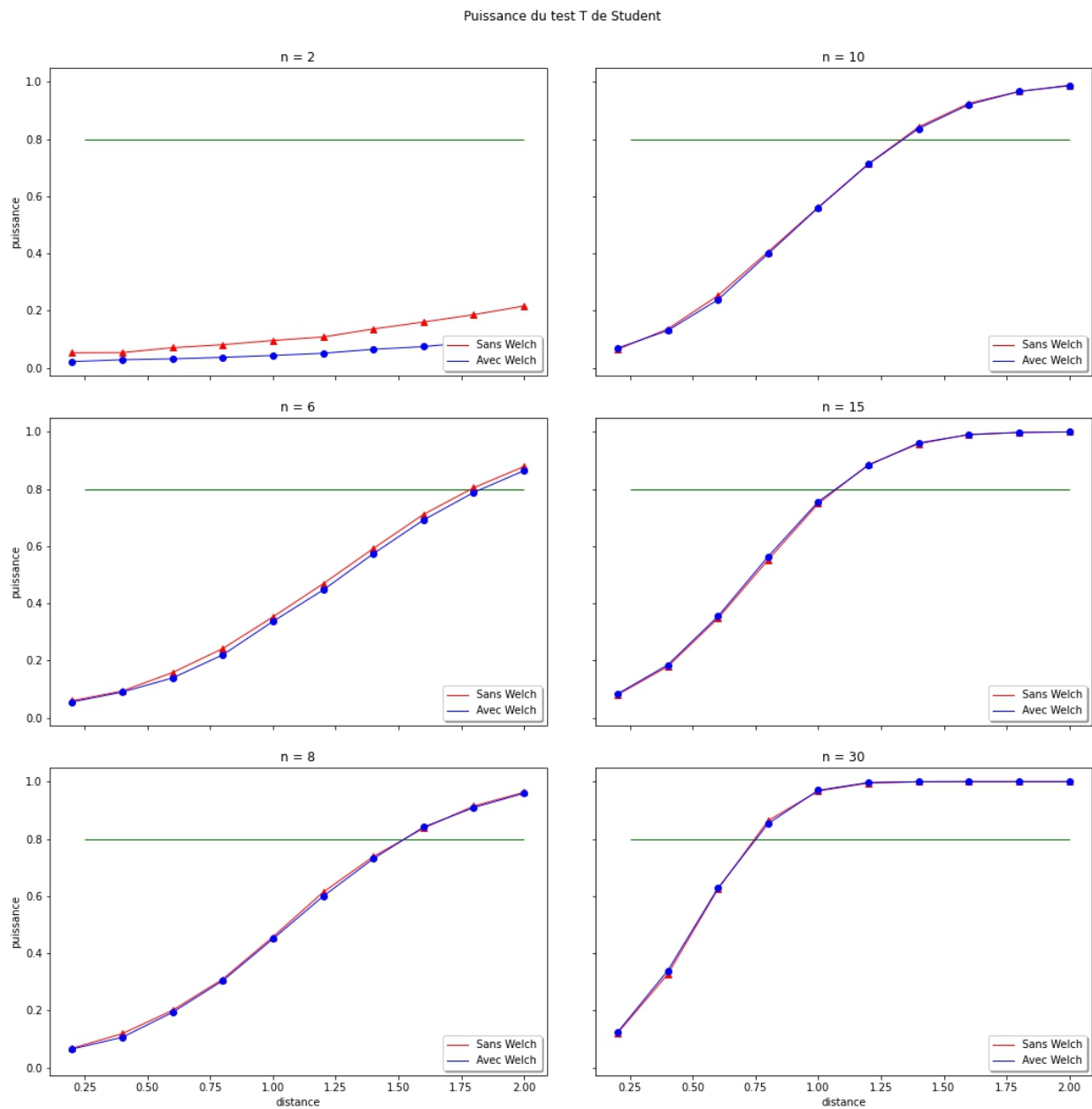
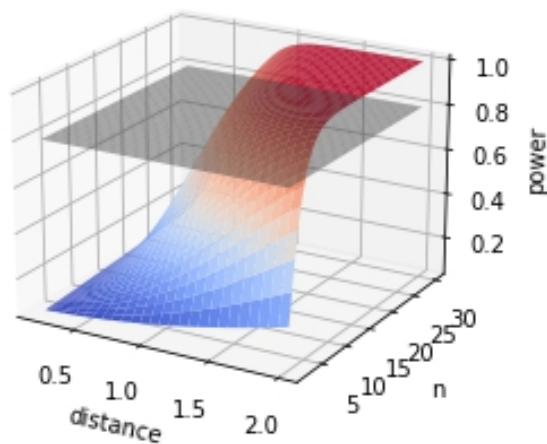


Figure 4 : Representation 3D de l'analyse de puissance ↗



VI. Ressources

^a ↑ Formule du test T de Student - LATEX :

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

^b ↑ Formule de la correction de Welch appliquée au test T de Student - LATEX :

$$t = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2} \right)^2}{\frac{s_1^4}{N_1^2 \cdot \nu_1} + \frac{s_2^4}{N_2^2 \cdot \nu_2}}$$

^c ↑ GitHub repository : https://github.com/nobabar/Welch_Power-and-Reliability

VII. Bibliographie

¹ ↑ Darrell Huff, 1954. *How to Lie with Statistics*. Norton, New York, ISBN 0-393-31072-8

² ↑ Alex Reinhart, March 2015. *Statistics done wrong*. ISBN-13: 978-1-59327-620-1.

statisticsdonewrong.com

³ ↑ Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLOS Medicine 2(8): e124. <https://doi.org/10.1371/journal.pmed.0020124>

⁴ ↑ Bausell, R., & Li, Y. (2002). *Power Analysis for Experimental Research: A Practical Guide for the Biological, Medical and Social Sciences*. Cambridge: Cambridge University Press.
doi:10.1017/CBO9780511541933

⁵ ↑ Fabio Veronesi in R bloggers. July 21, 2017. Power analysis and sample size calculation for Agriculture. r-bloggers.com/2017/07/power-analysis-and-sample-size-calculation-for-agriculture/

⁶ ↑ <https://xkcd.com/882/>