

ASSISTANT MINITAB - LIVRE BLANC

Ce livre blanc fait partie d'une série de documents qui expliquent les recherches menées par les statisticiens de Minitab pour développer les méthodes et les outils de vérification des données utilisés dans l'Assistant de Minitab Statistical Software.

ANOVA à 1 facteur contrôlé

Généralités

L'ANOVA à un facteur contrôlé permet de comparer les moyennes de trois groupes ou plus, afin de déterminer si elles diffèrent de manière significative les unes des autres. Une autre fonction importante de l'ANOVA à un facteur contrôlé est d'estimer les différences entre des groupes spécifiques.

La méthode la plus courante pour détecter des différences entre des groupes dans une ANOVA à un facteur contrôlé est

le test F, qui s'appuie sur l'hypothèse selon laquelle les populations de tous les échantillons partagent un écart type commun, mais inconnu. En pratique, nous sommes conscients que les échantillons présentent souvent des écarts types différents. C'est pourquoi nous souhaitons étudier les résultats de la méthode de Welch, une méthode alternative permettant de traiter des écarts types inégaux. En outre, nous souhaitons développer une méthode de calcul de comparaisons multiples qui permette de repérer les échantillons présentant des écarts types inégaux. Cette méthode permet de représenter graphiquement des intervalles individuels afin d'identifier facilement les groupes qui diffèrent les uns des autres.

Dans cet article, nous décrivons la façon dont nous avons développé les méthodes utilisées dans la procédure d'ANOVA à un facteur contrôlé de l'Assistant Minitab pour :

- Le test de Welch
- Les intervalles de comparaisons multiples

En outre, nous étudions des conditions pouvant avoir une incidence sur la validité des résultats de l'ANOVA à un facteur contrôlé, notamment la présence de données aberrantes, l'effectif d'échantillon et la puissance du test, ainsi que la normalité des données. Pour ces conditions, l'Assistant effectue automatiquement les vérifications suivantes sur vos données et présente les résultats dans le rapport :

- Donnée aberrantes
- Effectif d'échantillon

- Normalité des données

Dans ce document, nous étudions la façon dont ces conditions influent sur l'ANOVA à un facteur contrôlé dans la pratique et décrivons comment nous avons établi nos méthodes de vérification de ces conditions.

Méthodes d'ANOVA à un facteur contrôlé

Test F contre test de Welch

Le test F, couramment utilisé dans une ANOVA à un facteur contrôlé, part de l'hypothèse selon laquelle tous les groupes partagent un écart type commun, mais inconnu (σ). Dans la pratique, cette hypothèse est rarement vérifiée, ce qui rend difficile le contrôle du taux d'erreur de 1^{ère} espèce. L'erreur de 1^{ère} espèce est la probabilité de rejet à tort de l'hypothèse nulle (amenant à conclure que des échantillons sont significativement différents alors qu'ils ne le sont pas). Lorsque les échantillons présentent des écarts types différents, il y a une plus forte probabilité pour que le test parvienne à une conclusion erronée. Pour résoudre ce problème, une méthode alternative a été développée, le test de Welch (Welch, 1951).

Objectif

Nous souhaitons déterminer s'il était préférable d'utiliser le test F ou le test de Welch pour la procédure d'ANOVA à un facteur contrôlé de l'Assistant. Pour ce faire, nous devons évaluer dans quelle mesure les résultats réels du test F et du test de Welch étaient proches du seuil de signification cible (alpha ou taux d'erreur de 1^{ère} espèce) ; en d'autres termes, il s'agissait de déterminer si le test rejetait à tort l'hypothèse nulle plus ou moins souvent que voulu, avec des effectifs et des écarts types d'échantillons différents.

Méthode

Pour comparer le test F et le test de Welch, nous avons effectué plusieurs simulations en faisant varier le nombre d'échantillons, ainsi que l'effectif et l'écart type des échantillons. Pour chaque condition, nous avons effectué 10 000 tests ANOVA en utilisant le test F et la méthode de Welch. Nous avons généré des données aléatoires de telle sorte que les moyennes des échantillons soient identiques et que, par conséquent, l'hypothèse nulle soit vraie pour chaque test. Nous avons ensuite effectué les tests en utilisant des seuils de signification cibles de 0,05 et 0,01. Nous avons compté le nombre de fois sur les 10 000 tests où le test F et le test de Welch avaient rejeté l'hypothèse nulle, puis nous avons comparé cette proportion au seuil de signification cible. Si le test est adéquat, le taux d'erreur de 1^{ère} espèce estimé doit être très proche du seuil de signification cible.

Les résultats

Nous avons constaté que la méthode de Welch permettait d'obtenir des résultats équivalents à ceux du test F ou meilleurs pour toutes les conditions testées. Par exemple, lorsque nous comparons 5 échantillons à l'aide du test de Welch, les taux d'erreur de 1^{ère} espèce se situent entre 0,0460 et 0,0540 ; ils sont donc très proches du seuil de signification cible de 0,05. Cela indique que le taux d'erreur de 1^{ère} espèce obtenu avec la méthode de Welch correspond à la valeur cible, même lorsque l'effectif et l'écart type varient d'un échantillon à l'autre.

Par comparaison, les taux d'erreur de 1^{ère} espèce pour le test F se situent entre 0,0273 et 0,2277. Le test F a notamment produit des résultats erronés dans les conditions suivantes :

- Les taux d'erreur de 1^{ère} espèce étaient inférieurs à 0,05 lorsque le plus large échantillon présentait également l'écart type le plus important. Cette condition produit des résultats plus prudents et montre qu'augmenter l'effectif d'échantillon n'est pas une solution viable lorsque les écarts types des échantillons ne sont pas égaux.
- Les taux d'erreur de 1^{ère} espèce étaient supérieurs à 0,05 lorsque les échantillons étaient égaux mais que les écarts types différaient. Les taux étaient également supérieurs à 0,05 lorsque l'échantillon qui présentait l'écart type le plus important avait un effectif inférieur à celui des autres. De manière générale, lorsque de petits échantillons présentent des écarts types importants, le risque que ce test rejette à tort l'hypothèse nulle augmente.

Pour plus d'informations sur les résultats de la simulation et la méthodologie utilisée, reportez-vous à l'annexe A.

La méthode de Welch ayant obtenu de bons résultats avec des écarts types et des effectifs d'échantillons inégaux, c'est elle que nous utilisons pour la procédure d'ANOVA à un facteur contrôlé de l'Assistant.

Intervalles de comparaison

Lorsqu'un test ANOVA est statistiquement significatif et indique qu'au moins une des moyennes de l'échantillon diffère des autres, l'étape suivante de l'analyse consiste à identifier les échantillons qui sont statistiquement différents. Cette comparaison pourrait être effectuée de façon intuitive en traçant le graphique des intervalles de confiance et en cherchant les échantillons dont les intervalles ne se chevauchent pas. Toutefois, les conclusions tirées de ce graphique peuvent ne pas correspondre aux résultats des tests, car les intervalles de confiance individuels n'ont pas vocation à être comparés. Si une méthode de comparaisons multiples a bien été publiée pour les échantillons ayant des écarts types égaux, il était nécessaire de l'étendre pour pouvoir l'appliquer à des échantillons qui présentent des écarts types inégaux, ce que nous avons fait.

Objectif

Nous souhaitons développer une méthode pour calculer des intervalles de comparaisons individuels qui permettent de comparer des échantillons en offrant des résultats aussi proches que possible de ceux des tests. En outre, nous souhaitons fournir une méthode permettant de repérer visuellement les échantillons statistiquement différents des autres.

Méthode

Les méthodes de comparaisons multiples standard (Hsu, 1996) calculent l'intervalle contenant la différence entre chaque paire de moyennes tout en contrôlant l'augmentation du taux d'erreur associée aux comparaisons multiples. Dans le cas particulier d'effectifs d'échantillons égaux et partant de l'hypothèse de l'égalité des écarts types, il est possible d'afficher les intervalles individuels de chaque moyenne d'une façon qui corresponde

exactement aux intervalles des différences entre toutes les paires. Pour le cas d'effectifs d'échantillons inégaux et partant de l'hypothèse d'écarts types égaux, Hochberg, Weiss et Hart (1982) ont développé des intervalles individuels approximativement équivalents aux intervalles des différences entre les paires, en se fondant sur la méthode des comparaisons multiples de Tukey-Kramer. Dans l'Assistant, nous appliquons la même approche à la méthode de comparaisons multiples de Games-Howell, qui ne part pas de l'hypothèse d'écarts types égaux. L'approche adoptée dans l'Assistant dans la version 16 de Minitab est similaire dans son concept, mais ne repose pas directement sur l'approche de Games-Howell. Pour plus de détails, reportez-vous à l'Annexe B.

Les résultats

L'Assistant affiche les intervalles de comparaison dans le tableau comparatif des moyennes du rapport récapitulatif de l'ANOVA à un facteur contrôlé. Lorsque le test ANOVA est statistiquement significatif, tout intervalle de comparaison qui ne chevauche pas au moins un autre intervalle apparaît en rouge. Il est possible que le test et les intervalles de comparaison ne concordent pas, mais ce résultat est peu fréquent, car les deux méthodes ont la même probabilité de rejeter l'hypothèse nulle à tort. Si le test ANOVA est significatif mais que tous les intervalles se chevauchent, la paire présentant l'étendue de chevauchement la plus réduite apparaît en rouge. Si le test ANOVA n'est pas statistiquement significatif, aucun intervalle n'apparaît en rouge, même si certains d'entre eux ne se chevauchent pas.

Vérifications des données

Donnée aberrantes

Les données aberrantes sont des valeurs extrêmement grandes ou extrêmement petites, également connues sous le nom de valeurs aberrantes. Les données aberrantes peuvent avoir une forte influence sur les résultats de l'analyse et peuvent compromettre la possibilité de trouver des résultats statistiquement significatifs, notamment avec de petits échantillons. Les données aberrantes peuvent venir de problèmes de collecte de données ou être dues à un comportement inhabituel du procédé étudié. 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 souhaitons développer une méthode pour vérifier 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 les données aberrantes, inspirée de la méthode décrite par Hoaglin, Iglewicz et Tukey (1986), qui permet d'identifier les valeurs aberrantes dans les boîtes à moustache.

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 25^{ème} et 75^{ème} 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.

Lors du test des données aberrantes, l'Assistant affiche les indicateurs d'état suivants dans le rapport :

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

Effectif de l'échantillon

La puissance est une propriété importante de tout test d'hypothèse, car elle indique la probabilité de détecter une différence ou un effet significatif existant. La puissance est la probabilité de rejet de l'hypothèse nulle en faveur de l'hypothèse alternative. La manière la plus simple d'augmenter la puissance d'un test est souvent d'augmenter l'effectif de l'échantillon. Dans l'Assistant, pour des tests à faible puissance, nous indiquons l'effectif d'échantillon requis pour détecter la différence que vous avez spécifiée. Si aucune différence n'est spécifiée, nous indiquons la différence que vous pouvez détecter avec une puissance adéquate. Pour fournir cette information, il nous fallait développer une méthode de calcul de la puissance, étant donné que l'Assistant utilise la méthode de Welch et que celle-ci ne dispose pas d'une formule exacte pour calculer la puissance.

Objectif

Pour développer une méthodologie pour le calcul de la puissance, nous devons résoudre deux problèmes. D'une part, l'Assistant ne demande pas aux utilisateurs de saisir un ensemble complet de moyennes, mais seulement une différence de moyenne susceptible d'avoir des conséquences d'un point de vue pratique. Or, le nombre de configurations de moyennes permettant d'aboutir à une différence donnée est infinie. Par conséquent, nous devons développer une approche permettant de déterminer les moyennes à utiliser dans le calcul de la puissance, dans la mesure où nous ne pouvons pas calculer la puissance pour toutes les configurations de moyennes possibles. D'autre part, nous devons développer une nouvelle méthode pour calculer la puissance, car l'Assistant s'appuie sur la méthode de Welch, qui prend en compte des effectifs d'échantillons et des écarts types inégaux.

Méthode

Pour résoudre le problème lié au nombre infini de configurations de moyennes possibles, nous avons développé une méthode fondée sur l'approche utilisée dans la procédure d'ANOVA à un facteur contrôlé standard de Minitab (**Stat > ANOVA > A un facteur**). Nous nous sommes penchés sur les cas où seules deux des moyennes sont différentes (avec une différence donnée) et les autres sont égales (et sont définies sur la moyenne globale pondérée). Comme nous supposons que seules deux moyennes (et pas plus de deux) diffèrent de la moyenne globale, l'approche fournit une estimation prudente de la puissance. Toutefois, étant donné que les effectifs d'échantillons ou les écarts types peuvent différer, il est toujours nécessaire d'identifier les moyennes qui diffèrent pour pouvoir calculer la puissance.

Pour résoudre ce problème, nous identifions les deux paires de moyennes qui représentent le meilleur et le pire des cas. Le pire des cas se produit lorsque l'effectif de l'échantillon est faible par rapport à la variance de l'échantillon et que la puissance est donc minimale ; le meilleur des cas se produit lorsque l'effectif de l'échantillon est élevé par rapport à la variance de l'échantillon et que la puissance est donc maximale. Tous les calculs de puissance prennent en compte ces deux cas extrêmes, qui minimisent et maximisent la puissance, en partant de l'hypothèse selon laquelle deux moyennes, exactement, diffèrent de la moyenne pondérée globale.

Pour développer le calcul de la puissance, nous avons utilisé une méthode décrite dans Kulinskaya et al. (2003). Nous avons comparé les calculs de puissance obtenus dans le cadre notre simulation à ceux obtenus avec la méthode que nous avons développée pour résoudre le problème de la configuration des moyennes et avec celle décrite dans Kulinskaya et al. (2003). En outre, nous avons étudié les résultats d'une autre approximation de la puissance indiquant de façon plus claire la relation entre la puissance et la configuration des moyennes. Pour plus d'informations sur le calcul de la puissance, reportez-vous à l'Annexe C.




Les résultats



Notre comparaison indique que la méthode de Kulinskaya fournit une bonne approximation de la puissance et que notre méthode de traitement de la configuration des moyennes est adaptée.

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. En outre, si vous indiquez une différence pratique, l'Assistant calcule les valeurs de puissance minimale et maximale pour cette différence. Lorsque les valeurs de puissance sont inférieures à 90 %, l'Assistant calcule un effectif d'échantillon à partir de la différence spécifiée et des écarts types observés pour l'échantillon. Pour garantir que l'effectif d'échantillon offre des valeurs de puissance minimale et maximale toutes deux supérieures ou égales à 90 %, nous supposons que la différence indiquée est celle qui sépare les deux moyennes qui présentent la plus grande variabilité.

Si l'utilisateur n'indique pas de différence, l'Assistant indique la différence la plus élevée pour laquelle la valeur maximale de l'étendue des puissances est de 60 %. Cette valeur est indiquée à la limite des barres rouge et jaune du rapport de puissance, qui correspond à une puissance de 60 %. Nous recherchons également la différence la plus faible pour laquelle la valeur minimale de l'étendue des puissances est de 90 %. Cette valeur est indiquée à la limite des barres jaune et verte du rapport de puissance, qui correspond à une puissance de 90 %.

Lors de la vérification de la puissance et de l'effectif d'échantillon, l'Assistant affiche les indicateurs d'état suivants dans le rapport :

Etat	Condition
	Les données ne permettent pas de conclure qu'il existe des différences entre les moyennes. Aucune différence n'a été indiquée.
	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.
	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.

Etat	Condition
	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'échantillon nécessaires pour atteindre une puissance de 80 % et de 90 % sont indiqués.

Normalité

De nombreuses méthodes statistiques partent de l'hypothèse selon laquelle les données obéissent à une loi de distribution normale. Par chance, certaines méthodes s'appuyant sur l'hypothèse de normalité offrent de bons résultats même si les données ne sont pas distribuées normalement. Ceci s'explique en partie par le théorème de la limite centrale, selon lequel la distribution d'une moyenne d'échantillon est approximativement normale et se rapproche de la normale à mesure que l'effectif de l'échantillon augmente.

Objectif

Notre objectif était de déterminer l'effectif de l'échantillon à utiliser pour obtenir une distribution suffisamment proche de la loi normale. Nous souhaitions examiner les résultats du test de Welch et des intervalles de comparaison pour des effectifs d'échantillons faibles ou modérés obéissant à différents types de lois non normales. Nous souhaitions déterminer dans quelle mesure les résultats réels du test de Welch et des intervalles de comparaison correspondaient au seuil de signification du test (valeur alpha ou taux d'erreur de 1^{ère} espèce) ; en d'autres termes, il s'agissait de déterminer si le test rejetait à tort l'hypothèse nulle plus ou moins souvent que voulu, avec différents effectifs d'échantillons, nombres de niveaux et types de lois non normales.

Méthode

Pour estimer le taux d'erreur de 1^{ère} espèce, nous avons effectué plusieurs simulations en faisant varier le nombre d'échantillons, l'effectif des échantillons et la distribution des données. Des simulations ont été réalisées avec des lois asymétriques et à queues lourdes qui s'écartent sensiblement de la loi normale. Pour chaque test, l'effectif et l'écart type étaient constants d'un échantillon à l'autre.



Pour chaque condition, nous avons effectué 10 000 tests ANOVA en utilisant la méthode de Welch et les intervalles de comparaison. Nous avons généré des données aléatoires de telle sorte que les moyennes des échantillons soient identiques et que, par conséquent, l'hypothèse nulle soit vraie pour chaque test. Nous avons ensuite effectué les tests en utilisant un seuil de signification cible de 0,05. Nous avons compté le nombre de fois, sur 10 000, où les tests avaient rejeté l'hypothèse nulle, puis nous avons comparé cette proportion au seuil de signification cible. Pour les intervalles de comparaison, nous avons compté le nombre de fois, sur 10 000, où les intervalles avaient indiqué au moins une

différence. Si le test est adéquat, le taux d'erreur de 1^{ère} espèce doit être très proche du seuil de signification cible.

Les résultats

Dans l'ensemble, les tests et les intervalles de comparaison ont produit de très bons résultats pour toutes les conditions avec des effectifs d'échantillons faibles de 10 ou 15. Pour les tests à 9 niveaux ou moins, dans presque tous les cas, les résultats se trouvent à moins de 3 points de pourcentage du seuil de signification cible pour un effectif d'échantillon de 10, et de 2 points de pourcentage pour un effectif d'échantillon de 15. Pour les tests à 10 niveaux ou plus, dans la plupart des cas, les résultats se trouvent à moins de 3 points de pourcentage pour un effectif d'échantillon de 15 et de 2 points de pourcentage pour un effectif de 20. Pour plus d'informations, reportez-vous à l'Annexe D.

Etant donné que les tests obtiennent de bons résultats avec de petits échantillons, l'Assistant ne teste pas la normalité des données. En revanche, l'Assistant vérifie l'effectif des échantillons et signale les effectifs d'échantillons inférieurs à 15 quand le nombre de niveaux est compris entre 2 et 9 et ceux inférieurs à 20 quand le nombre de niveaux est compris entre 10 et 12. En fonction de ces résultats, l'Assistant affiche les indicateurs d'état suivants dans le rapport :

Etat	Condition
	Les effectifs d'échantillons sont d'au moins 15 ou 20, la normalité n'est donc pas un problème.
	Certains effectifs d'échantillons sont inférieurs à 15 ou 20, ce qui peut poser un problème de normalité.

Références

- Dunnett, C. W., (1980), Pairwise Multiple Comparisons in the Unequal Variance Case, *Journal of the American Statistical Association*, 75, 796-800.
- 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.
- Hochberg, Y., Weiss G. et Hart, S., (1982), On graphical procedures for multiple comparisons, *Journal of the American Statistical Association*, 77, 767-772.
- Hsu, J. (1996), *Multiple comparisons: Theory and methods*, Boca Raton, FL : Chapman & Hall.
- 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.
- Welch, B.L. (1947), The generalization of "Student's" problem when several different population variances are involved, *Biometrika*, 34, 28-35.
- Welch, B.L. (1951), On the comparison of several mean values: An alternative approach, *Biometrika*, 38, 330-336.

Annexe A : test F contre test de Welch

Le test F peut engendrer une augmentation du taux d'erreur de 1ère espèce lorsque l'hypothèse d'égalité des écarts types est contredite ; le test de Welch est conçu pour éviter ces problèmes.

Test de Welch

Nous étudions des échantillons randomisés d'effectifs n_1, \dots, n_k issus de k populations. Soit μ_1, \dots, μ_k les moyennes de populations et $\sigma_1^2, \dots, \sigma_k^2$ les variances de populations. Soit $\bar{x}_1, \dots, \bar{x}_k$ les moyennes d'échantillons et s_1^2, \dots, s_k^2 les variances d'échantillons. Nous souhaitons tester les hypothèses suivantes :

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1 : \mu_i \neq \mu_j \text{ pour des valeurs de } i \text{ et de } j.$$

Pour tester l'égalité de k moyennes, le test de Welch compare la statistique

$$W^* = \frac{\sum_{j=1}^k w_j (\bar{x}_j - \hat{\mu})^2 / (k-1)}{1 + [2(k-2)/(k^2-1)] \sum_{j=1}^k h_j}$$

à la loi $F(k-1, f)$, où

$$w_j = \frac{n_j}{s_j^2},$$

$$W = \sum_{j=1}^k w_j,$$

$$\hat{\mu} = \frac{\sum_{j=1}^k w_j \bar{x}_j}{W},$$

$$h_j = \frac{(1 - w_j/W)^2}{n_j - 1} \text{ et}$$

$$f = \frac{k^2 - 1}{3 \sum_{j=1}^k h_j}.$$

Le test de Welch rejette l'hypothèse nulle lorsque $W^* \geq F_{k-1, f, 1-\alpha}$, le percentile de la loi F dépassé avec une probabilité de α .

Écarts types inégaux

Dans cette section, nous démontrons la sensibilité du test F aux violations de l'hypothèse d'égalité des écarts types et le comparons au test de Welch.

Vous trouverez ci-dessous les résultats des tests ANOVA à un facteur contrôlé réalisés avec 5 échantillons de $N(0, \sigma^2)$. Chaque ligne correspond à 10 000 simulations effectuées à l'aide du test F et du test de Welch. Nous avons testé deux conditions de l'écart type en augmentant celui du cinquième échantillon, en le doublant et le quadruplant par rapport aux autres échantillons. Nous avons testé trois conditions différentes concernant l'effectif de

l'échantillon : les effectifs des échantillons sont égaux, le cinquième échantillon est plus grand que les autres et le cinquième échantillon est plus petit que les autres.

Tableau 1 Taux d'erreur de 1^{ère} espèce pour le test F et le test de Welch simulés avec 5 échantillons dont le seuil de signification cible est de $\alpha=0,05$.

Ecart type ($\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5$)	Effectif d'échantillon (n_1, n_2, n_3, n_4, n_5)	Test F	Test de Welch
1, 1, 1, 1, 2	10, 10, 10, 10, 20	0,0273	0,0524
1, 1, 1, 1, 2	20, 20, 20, 20, 20	0,0678	0,0462
1, 1, 1, 1, 2	20, 20, 20, 20, 10	0,1258	0,0540
1, 1, 1, 1, 4	10, 10, 10, 10, 20	0,0312	0,0460
1, 1, 1, 1, 4	20, 20, 20, 20, 20	0,1065	0,0533
1, 1, 1, 1, 4	20, 20, 20, 20, 10	0,2277	0,0503

Lorsque les échantillons ont des effectifs égaux (lignes 2 et 5), la probabilité que le test F rejette à tort l'hypothèse nulle est supérieure à la valeur cible de 0,05 et cette probabilité est plus forte lorsque l'inégalité entre les écarts types augmente. Le problème s'aggrave lorsque l'on réduit l'effectif de l'échantillon ayant le plus grand écart type. A l'inverse, augmenter l'effectif de l'échantillon ayant le plus grand écart type réduit la probabilité de rejet. Toutefois, une augmentation excessive de l'effectif d'échantillon entraîne une probabilité de rejet trop faible, ce qui non seulement rend le test plus prudent qu'il n'est nécessaire avec l'hypothèse nulle, mais réduit également la puissance du test avec l'hypothèse alternative. Comparez ces résultats avec ceux du test de Welch, qui correspondent bien au seuil de signification cible de 0,05 dans tous les cas.

Nous avons ensuite effectué une simulation pour $k = 7$ échantillons. Chaque ligne du tableau représente 10 000 tests F simulés. Nous avons fait varier les écarts types et les effectifs des échantillons. Les seuils de signification cibles sont de $\alpha = 0,05$ et de $\alpha = 0,01$. Comme dans la simulation précédente, nous pouvons constater des écarts très élevés par rapport aux valeurs cibles. L'utilisation d'un effectif d'échantillon plus petit lorsque la variabilité est forte entraîne des probabilités de taux d'erreur de 1^{ère} espèce très élevées, tandis que l'utilisation d'un effectif d'échantillon supérieur peut produire des résultats excessivement prudents. Les résultats sont présentés dans le tableau 2 ci-dessous.

Tableau 2 Taux d'erreur de 1^{ère} espèce pour les tests F simulés avec 7 échantillons

Ecart type ($\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \sigma_6, \sigma_7$)	Effectifs d'échantillons ($n_1, n_2, n_3, n_4, n_5, n_6, n_7$)	α cible = 0,05	α cible = 0,01
1,85, 1,85, 1,85, 1,85, 1,85, 1,85, 2,9	21, 21, 21, 21, 22, 22, 12	0,0795	0,0233
1,85, 1,85, 1,85, 1,85, 1,85, 1,85, 2,9	20, 21, 21, 21, 21, 24, 12	0,0785	0,0226
1,85, 1,85, 1,85, 1,85, 1,85, 1,85, 2,9	20, 21, 21, 21, 21, 21, 15	0,0712	0,0199

Ecart type ($\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \sigma_6, \sigma_7$)	Effectifs d'échantillons (n1, n2, n3, n4, n5, n6, n7)	α cible = 0,05	α cible = 0,01
1,85, 1,85, 1,85, 1,85, 1,85, 1,85, 2,9	20, 20, 20, 21, 21, 23, 15	0,0719	0,0172
1,85, 1,85, 1,85, 1,85, 1,85, 1,85, 2,9	20, 20, 20, 20, 21, 21, 18	0,0632	0,0166
1,85, 1,85, 1,85, 1,85, 1,85, 1,85, 2,9	20, 20, 20, 20, 20, 20, 20	0,0576	0,0138
1,85, 1,85, 1,85, 1,85, 1,85, 1,85, 2,9	18, 19, 19, 20, 20, 20, 24	0,0474	0,0133
1,85, 1,85, 1,85, 1,85, 1,85, 1,85, 2,9	18, 18, 18, 18, 18, 18, 32	0,0314	0,0057
1,85, 1,85, 1,85, 1,85, 1,85, 1,85, 2,9	15, 18, 18, 19, 20, 20, 30	0,0400	0,0085
1,85, 1,85, 1,85, 1,85, 1,85, 1,85, 2,9	12, 18, 18, 18, 19, 19, 36	0,0288	0,0064
1,85, 1,85, 1,85, 1,85, 1,85, 1,85, 2,9	15, 15, 15, 15, 15, 15, 50	0,0163	0,0025
1,85, 1,85, 1,85, 1,85, 1,85, 1,85, 2,9	12, 12, 12, 12, 12, 12, 68	0,0052	0,0002
1,75, 1,75, 1,75, 1,75, 1,75, 1,75, 3,5	21, 21, 21, 21, 22, 22, 12	0,1097	0,0436
1,75, 1,75, 1,75, 1,75, 1,75, 1,75, 3,5	20, 21, 21, 21, 21, 24, 12	0,1119	0,0452
1,75, 1,75, 1,75, 1,75, 1,75, 1,75, 3,5	20, 21, 21, 21, 21, 21, 15	0,0996	0,0376
1,75, 1,75, 1,75, 1,75, 1,75, 1,75, 3,5	20, 20, 20, 21, 21, 23, 15	0,0657	0,0345
1,75, 1,75, 1,75, 1,75, 1,75, 1,75, 3,5	20, 20, 20, 20, 21, 21, 18	0,0779	0,0283
1,75, 1,75, 1,75, 1,75, 1,75, 1,75, 3,5	20, 20, 20, 20, 20, 20, 20	0,0737	0,0264
1,75, 1,75, 1,75, 1,75, 1,75, 1,75, 3,5	18, 19, 19, 20, 20, 20, 24	0,0604	0,0204
1,75, 1,75, 1,75, 1,75, 1,75, 1,75, 3,5	18, 18, 18, 18, 18, 18, 32	0,0368	0,0122
1,75, 1,75, 1,75, 1,75, 1,75, 1,75, 3,5	15, 18, 18, 19, 20, 20, 30	0,0390	0,0117
1,75, 1,75, 1,75, 1,75, 1,75, 1,75, 3,5	12, 18, 18, 18, 19, 19, 36	0,0232	0,0046
1,75, 1,75, 1,75, 1,75, 1,75, 1,75, 3,5	15, 15, 15, 15, 15, 15, 50	0,0124	0,0026
1,75, 1,75, 1,75, 1,75, 1,75, 1,75, 3,5	12, 12, 12, 12, 12, 12, 68	0,0027	0,0004
1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 3,9	21, 21, 21, 21, 22, 22, 12	0,134	0,0630
1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 3,9	20, 21, 21, 21, 21, 24, 12	0,1329	0,0654
1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 3,9	20, 21, 21, 21, 21, 21, 15	0,1101	0,0484
1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 3,9	20, 20, 20, 21, 21, 23, 15	0,1121	0,0495

Ecart type ($\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \sigma_6, \sigma_7$)	Effectifs d'échantillons (n1, n2, n3, n4, n5, n6, n7)	α cible = 0,05	α cible = 0,01
1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 3,9	20, 20, 20, 20, 21, 21, 18	0,0876	0,0374
1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 3,9	20, 20, 20, 20, 20, 20, 20	0,0808	0,0317
1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 3,9	18, 19, 19, 20, 20, 20, 24	0,0606	0,0243
1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 3,9	18, 18, 18, 18, 18, 18, 32	0,0356	0,0119
1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 3,9	15, 18, 18, 19, 20, 20, 30	0,0412	0,0134
1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 3,9	12, 18, 18, 18, 19, 19, 36	0,0261	0,0068
1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 3,9	15, 15, 15, 15, 15, 15, 50	0,0100	0,0023
1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 1,68333, 3,9	12, 12, 12, 12, 12, 12, 68	0,0017	0,0003
1,55, 1,55, 1,55, 1,55, 1,55, 1,55, 4,7	21, 21, 21, 21, 22, 22, 12	0,1773	0,1006
1,55, 1,55, 1,55, 1,55, 1,55, 1,55, 4,7	20, 21, 21, 21, 21, 24, 12	0,1811	0,1040
1,55, 1,55, 1,55, 1,55, 1,55, 1,55, 4,7	20, 21, 21, 21, 21, 21, 15	0,1445	0,0760
1,55, 1,55, 1,55, 1,55, 1,55, 1,55, 4,7	20, 20, 20, 21, 21, 23, 15	0,1448	0,0786
1,55, 1,55, 1,55, 1,55, 1,55, 1,55, 4,7	20, 20, 20, 20, 21, 21, 18	0,1164	0,0572
1,55, 1,55, 1,55, 1,55, 1,55, 1,55, 4,7	20, 20, 20, 20, 20, 20, 20	0,1020	0,0503
1,55, 1,55, 1,55, 1,55, 1,55, 1,55, 4,7	18, 19, 19, 20, 20, 20, 24	0,0834	0,0369
1,55, 1,55, 1,55, 1,55, 1,55, 1,55, 4,7	18, 18, 18, 18, 18, 18, 32	0,0425	0,0159
1,55, 1,55, 1,55, 1,55, 1,55, 1,55, 4,7	15, 18, 18, 19, 20, 20, 30	0,0463	0,0168
1,55, 1,55, 1,55, 1,55, 1,55, 1,55, 4,7	12, 18, 18, 18, 19, 19, 36	0,0305	0,0103
1,55, 1,55, 1,55, 1,55, 1,55, 1,55, 4,7	15, 15, 15, 15, 15, 15, 50	0,0082	0,0021
1,55, 1,55, 1,55, 1,55, 1,55, 1,55, 4,7	12, 12, 12, 12, 12, 12, 68	0,0013	0,0001

Annexe B : intervalles de comparaison

Le tableau comparatif des moyennes permet d'évaluer la signification statistique des différences entre les moyennes des populations.

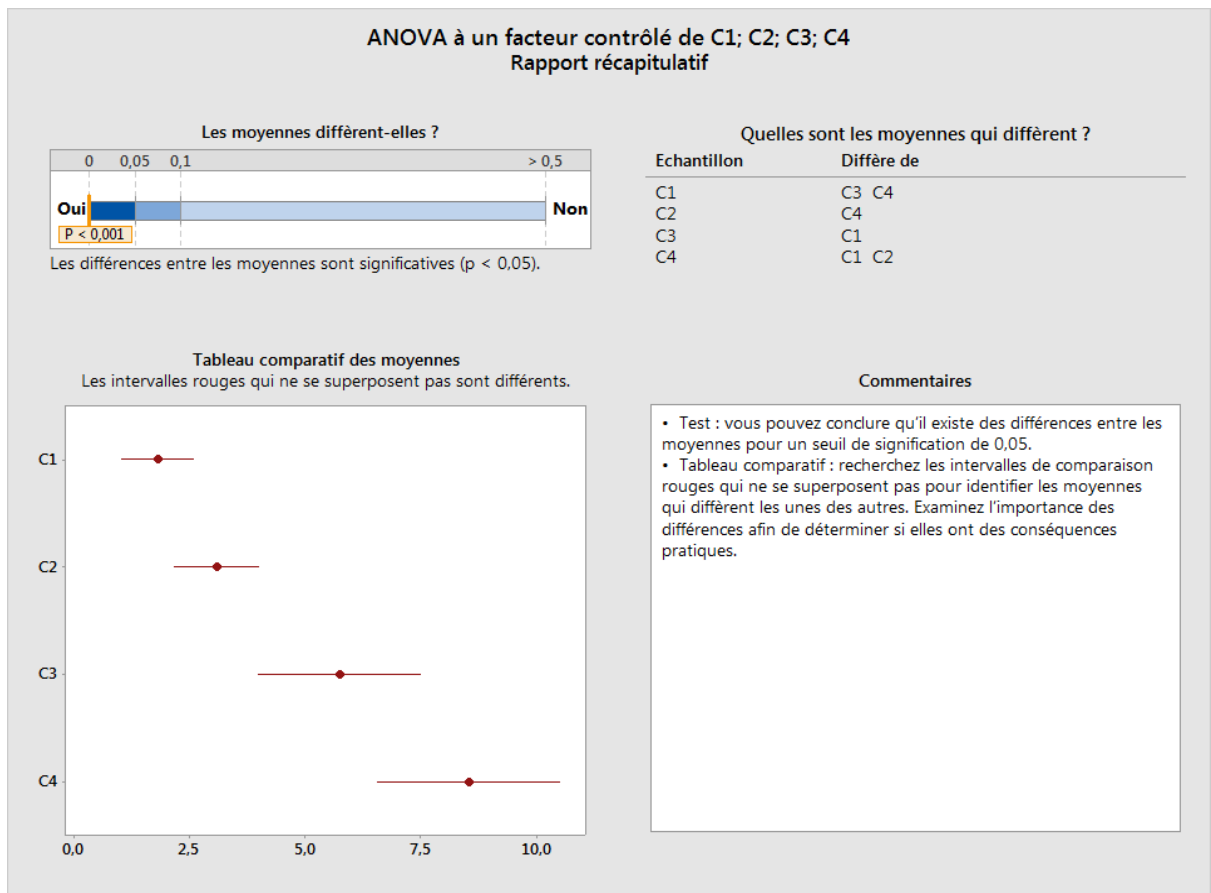
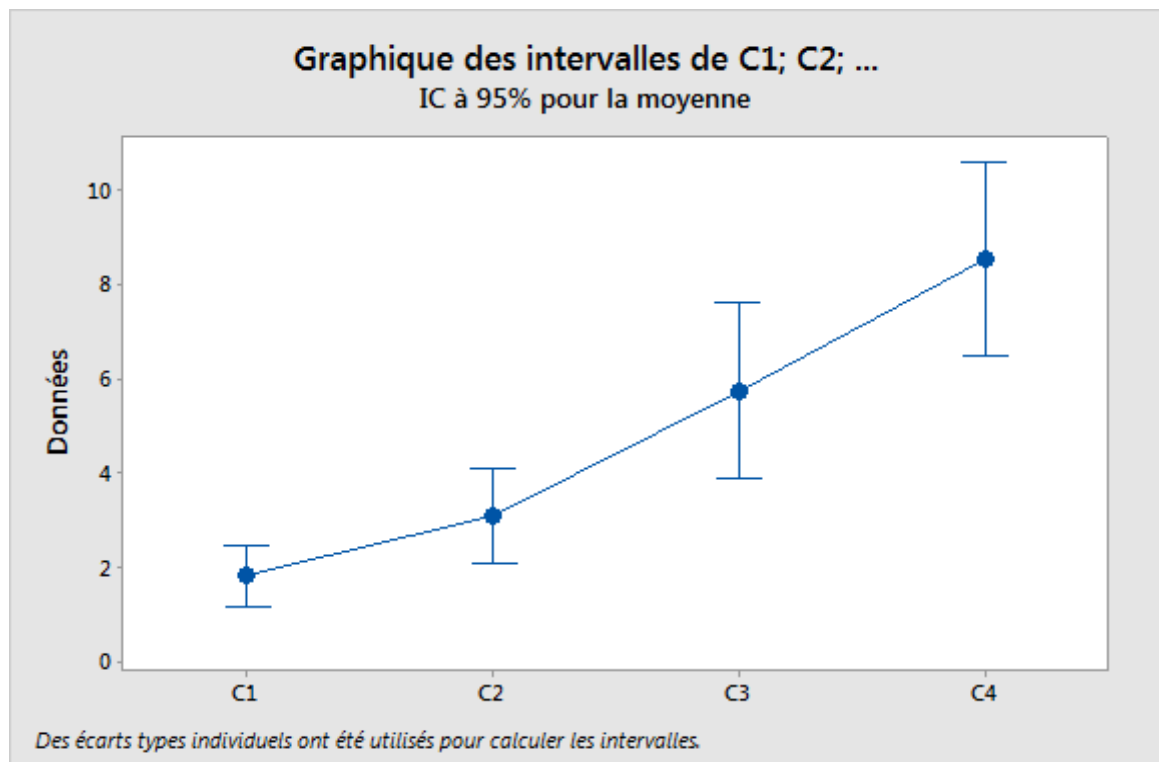


Figure 1 Tableau comparatif des moyennes du rapport de l'ANOVA à un facteur contrôlé de l'Assistant

Un ensemble d'intervalles semblable apparaît dans les résultats de la procédure d'ANOVA à un facteur contrôlé standard de Minitab (Stat > ANOVA > A un facteur) :



Toutefois, notez que les intervalles ci-dessus ne sont que les intervalles de confiance des moyennes. Lorsque le test ANOVA (F ou Welch) conclut que certaines moyennes sont différentes, nous tendons naturellement à chercher des intervalles qui ne se chevauchent pas et à en déduire que les moyennes correspondantes diffèrent. Cette analyse informelle des intervalles de confiance individuels permet souvent d'obtenir des conclusions correctes, mais elle ne prend pas en compte la probabilité d'erreur de la même façon que le test ANOVA. Selon le nombre de populations, les intervalles auront plus ou moins tendance à conclure à l'existence de différences que le test, parfois dans des proportions importantes. Par conséquent, les deux méthodes peuvent facilement parvenir à des conclusions contradictoires. Le tableau comparatif permet d'obtenir des résultats plus proches de ceux du test de Welch lorsque vous effectuez des comparaisons multiples, mais il ne garantit pas toujours une cohérence absolue.

Des méthodes de comparaisons multiples, telles que les comparaisons de Tukey-Kramer et Games-Howell dans Minitab (Stat > ANOVA > A un facteur), vous permettent de tirer des conclusions statistiquement valides sur les différences entre des moyennes individuelles. Ces deux méthodes effectuent des comparaisons deux à deux et fournissent un intervalle pour la différence entre chaque paire de moyennes. La probabilité que tous les intervalles contiennent simultanément les différences qu'ils estiment est d'au moins $1 - \alpha$. La méthode de Tukey-Kramer dépend de l'hypothèse d'égalité des variances, tandis que la méthode Games-Howell n'exige pas que les variances soient égales. Si l'hypothèse nulle d'égalité des moyennes est vraie, il n'existe aucune différence et la probabilité que l'un des intervalles de Games-Howell ne contienne pas la valeur zéro est au maximum de α . Par conséquent, nous pouvons utiliser les intervalles pour réaliser un test d'hypothèse avec un seuil de signification

de α . Nous utilisons des intervalles de Games-Howell comme points de départ pour obtenir les intervalles du tableau comparatif de l'Assistant.

Pour un ensemble d'intervalles $[L_{ij}, U_{ij}]$ contenant toutes les différences $\mu_i - \mu_j$, $1 \leq i < j \leq k$, nous souhaitons déterminer un ensemble d'intervalles $[L_i, U_i]$ des moyennes individuelles μ_i , $1 \leq i \leq k$, qui fournisse les mêmes informations. Pour cela, il est nécessaire que toute différence d soit comprise dans l'intervalle $[L_{ij}, U_{ij}]$, mais uniquement à la condition qu'il existe une valeur $\mu_i \in [L_i, U_i]$ et une valeur $\mu_j \in [L_j, U_j]$ telles que $\mu_i - \mu_j = d$. Les bornes des intervalles doivent être liées par les équations

$$U_i - L_j = U_{ij} \text{ et} \\ L_i - U_j = L_{ij}.$$

Pour $k = 2$, il n'existe qu'une différence, mais deux intervalles individuels, il est donc possible d'obtenir des intervalles de comparaison exacts. Il existe même une certaine souplesse dans la largeur des intervalles qui satisfont cette condition. Pour $k = 3$, nous avons trois différences et trois intervalles individuels. Là encore, il est donc possible de satisfaire à la condition, mais cette fois sans disposer de la même flexibilité pour définir la largeur des intervalles. Pour $k = 4$, il y a six différences, mais seulement quatre intervalles individuels. Les intervalles de comparaison doivent tâcher de transmettre les mêmes informations en utilisant moins d'intervalles. De manière générale, si $k \geq 4$, il existe plus de différences que de moyennes individuelles. Il n'existe donc pas de solution exacte, sauf si des conditions supplémentaires sont imposées aux intervalles de différences, comme des largeurs égales.

Les intervalles de Tukey-Kramer ne possèdent des largeurs égales que si tous les effectifs d'échantillons sont identiques. Des largeurs égales peuvent aussi être obtenues si l'on suppose une égalité des variances. Les intervalles de Games-Howell ne supposent pas l'égalité des variances et n'ont donc pas de largeurs égales. Dans l'Assistant, il nous faudra donc recourir à des méthodes par approximation pour définir des intervalles de comparaison.

L'intervalle de Games-Howell pour $\mu_i - \mu_j$ est égal à

$$\bar{x}_i - \bar{x}_j \pm |q^*(k, \hat{\nu}_{ij})| \sqrt{s_i^2/n_i + s_j^2/n_j}$$

où $q^*(k, \hat{\nu}_{ij})$ est le percentile concerné de la loi de l'étendue studentisée, lequel dépend de k , le nombre de moyennes comparées, et de

ν_{ij} , les degrés de liberté associés à la paire (i, j) :

$$\hat{\nu}_{ij} = \frac{\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}\right)^2}{\left(\frac{s_i^2}{n_i}\right)^2 \frac{1}{n_i - 1} + \left(\frac{s_j^2}{n_j}\right)^2 \frac{1}{n_j - 1}}.$$

Hochberg, Weiss et Hart (1982) ont obtenu des intervalles individuels approximativement équivalents à ces comparaisons deux à deux en utilisant :

$$\bar{x}_i \pm |q^*(k, \nu)| s_p X_i.$$

Les valeurs X_i sont sélectionnées pour minimiser

$$\sum \sum_{i \neq j} (X_i + X_j - a_{ij})^2,$$

où :

$$a_{ij} = \sqrt{1/n_i + 1/n_j}.$$

Nous adaptons cette approche au cas des variances inégales en dérivant les intervalles de comparaison de Games-Howell sur la forme

$$\bar{x}_i \pm d_i.$$

Les valeurs d_i sont sélectionnées pour minimiser

$$\sum \sum_{i \neq j} (d_i + d_j - b_{ij})^2,$$

où :

$$b_{ij} = |q^*(k, \hat{v}_{ij})| \sqrt{s_i^2/n_i + s_j^2/n_j}.$$

La solution est

$$d_i = \frac{1}{k-1} \sum_{j \neq i} b_{ij} - \frac{1}{(k-1)(k-2)} \sum_{j \neq i, l \neq i, j < l} b_{jl}.$$

Les graphiques ci-dessous comparent les résultats de simulation du test de Welch à ceux des intervalles de comparaison obtenus à l'aide de deux méthodes : la méthode adaptée de Games-Howell que nous utilisons actuellement et la méthode utilisée dans la version 16 de Minitab, qui s'appuie sur la moyenne des degrés de liberté. L'axe vertical représente le nombre de fois, sur 10 000 simulations, que le test de Welch rejette à tort l'hypothèse nulle ou que tous les intervalles de comparaison ne se chevauchent pas. Dans ces exemples, la valeur alpha cible est de $\alpha = 0,05$. Ces simulations couvrent divers cas d'écart types et d'effectifs d'échantillons inégaux ; chaque position en abscisse représente un cas différent.

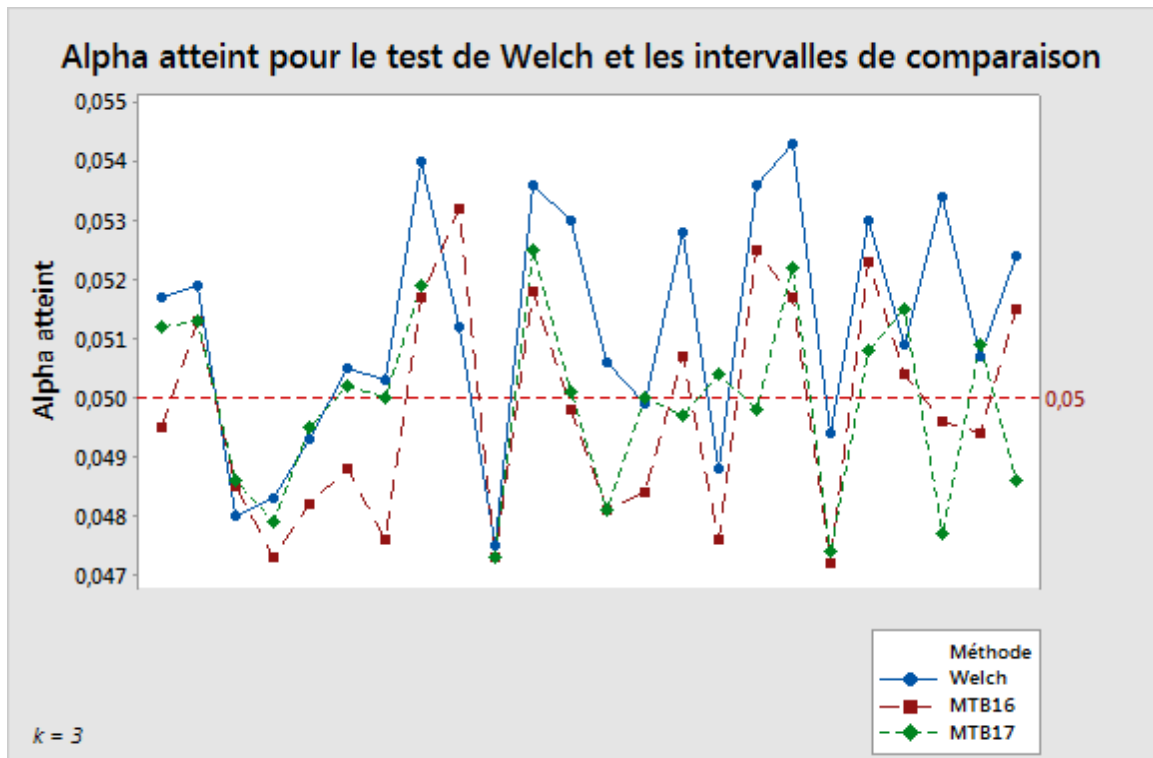


Figure 2 Comparaison du test de Welch à deux méthodes de calcul d'intervalles de comparaison pour 3 échantillons

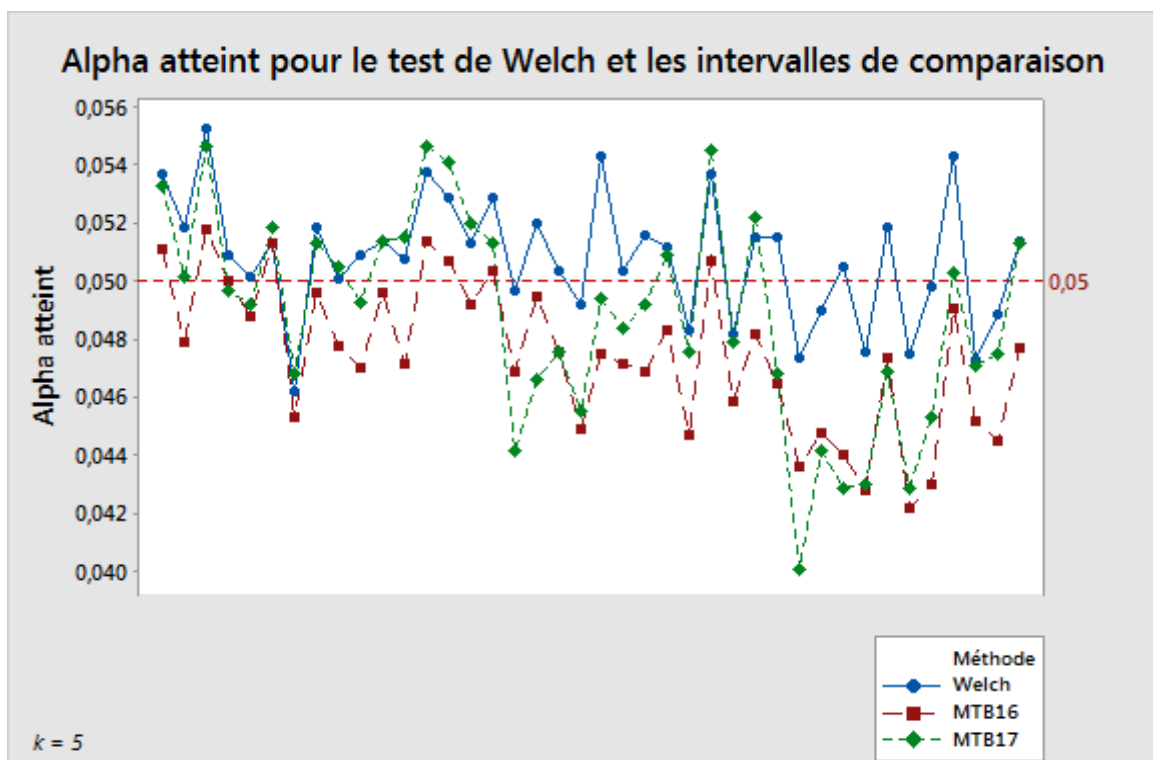


Figure 3 Comparaison du test de Welch à deux méthodes de calcul d'intervalles de comparaison pour 5 échantillons

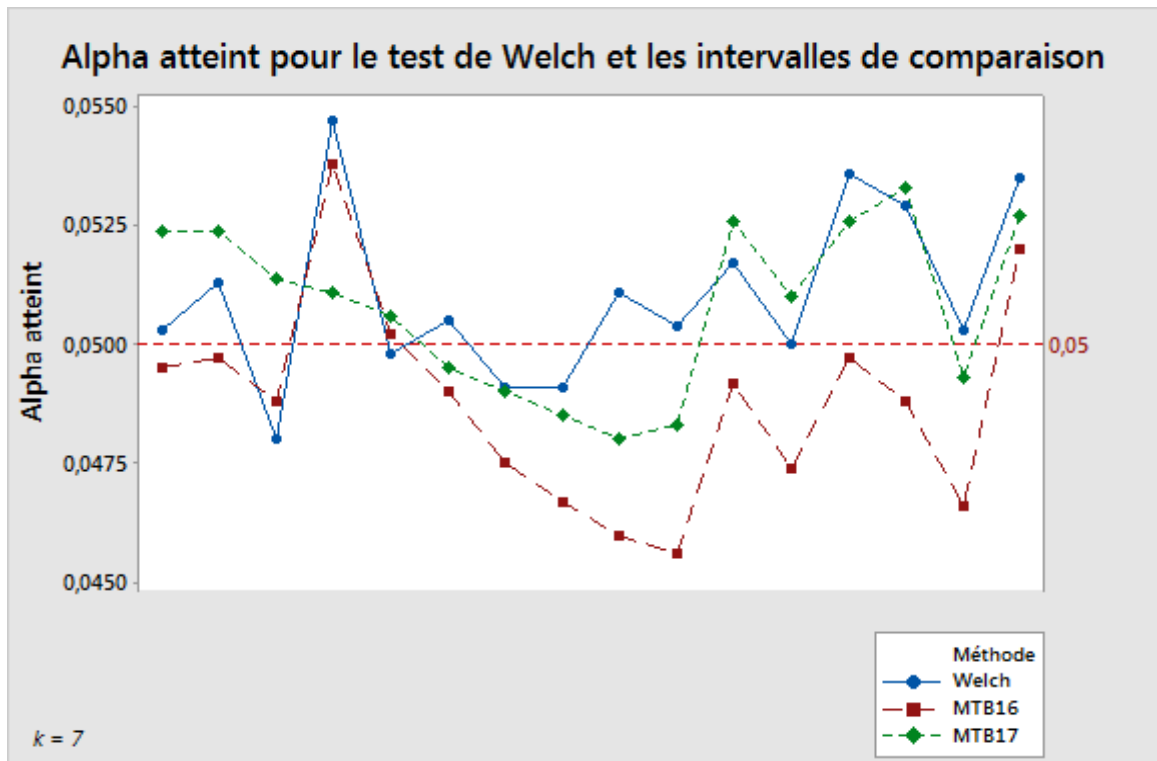


Figure 4 Comparaison du test de Welch à deux méthodes de calcul d'intervalles de comparaison pour 7 échantillons

Ces résultats donnent des valeurs alpha simulées comprises dans une étendue étroite entourant la valeur cible de 0,05. En outre, les résultats obtenus à l'aide de la méthode de Games-Howell mise en oeuvre dans la version 17 de Minitab sont sans doute plus étroitement alignés sur ceux du test de Welch que ceux obtenus avec la méthode utilisée dans la version 16 de Minitab.

Les données montrent que la probabilité de couverture des intervalles peut être sensible aux écarts types inégaux. Toutefois, cette sensibilité n'est pas aussi extrême que celle du test F. Le graphique ci-dessous illustre cette dépendance dans le cas où $k = 5$.

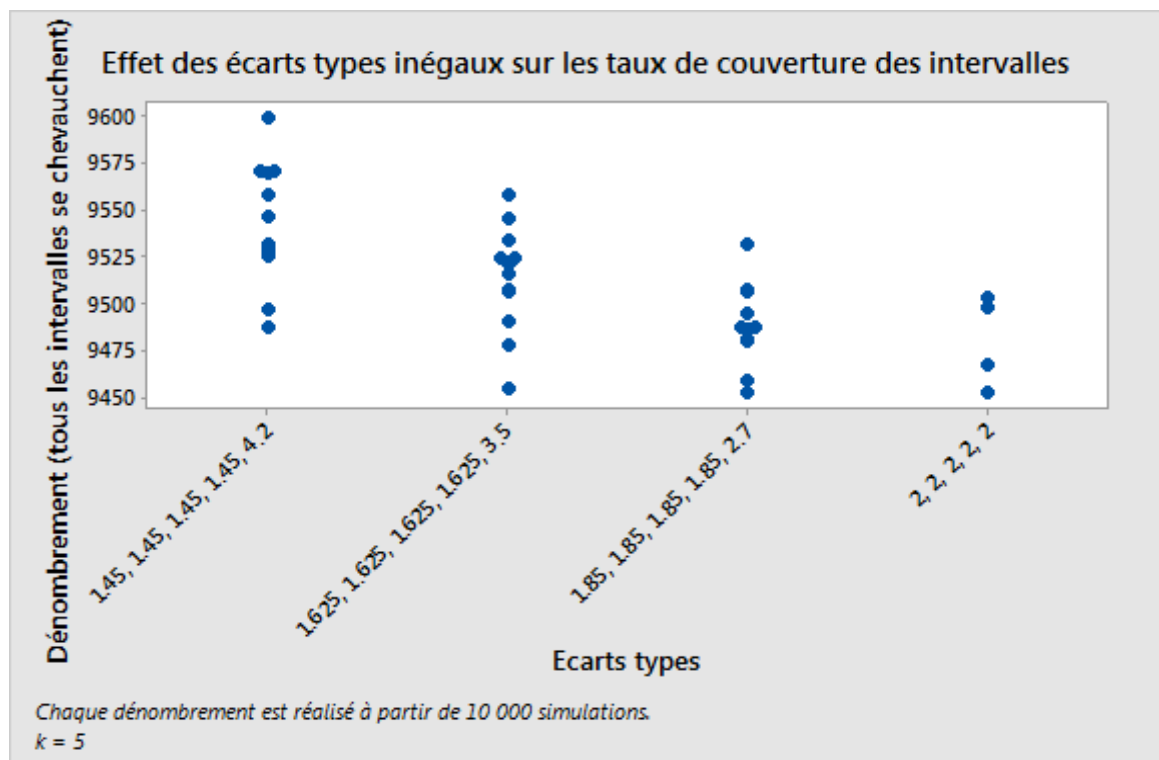


Figure 5 Résultats de la simulation avec des écarts types inégaux.

Utilisation conjointe du test d'hypothèse et d'intervalles de comparaison

Dans de rares cas, il est possible que le test d'hypothèse et la comparaison ne concordent pas sur le rejet de l'hypothèse nulle. Le test peut rejeter l'hypothèse nulle même si les intervalles de comparaison se chevauchent tous. A l'inverse, il est possible que le test ne rejette pas l'hypothèse nulle, même si certains intervalles ne se chevauchent pas. Ces désaccords sont rares, car la probabilité de rejet à tort de l'hypothèse nulle est la même pour les deux méthodes.

Lorsque cela se produit, nous examinons d'abord les résultats du test, puis, si le résultat est significatif, nous utilisons les comparaisons pour approfondir l'analyse. Si le test rejette l'hypothèse nulle au seuil de signification α , alors tout intervalle de comparaison qui ne chevauche pas au moins un autre intervalle apparaît en rouge. Cette indication visuelle signale que la moyenne du groupe correspondant diffère d'une autre moyenne au moins. Même si tous les intervalles se chevauchent, la paire présentant l'étendue de chevauchement la plus réduite apparaît en rouge si le test est significatif, pour indiquer la différence "la plus probable" (voir la figure 6 ci-dessous). Il s'agit d'un choix quelque peu arbitraire, surtout si d'autres paires présentent de faibles chevauchements. Toutefois, aucune autre paire ne présente de différence dont la borne soit plus proche de zéro.

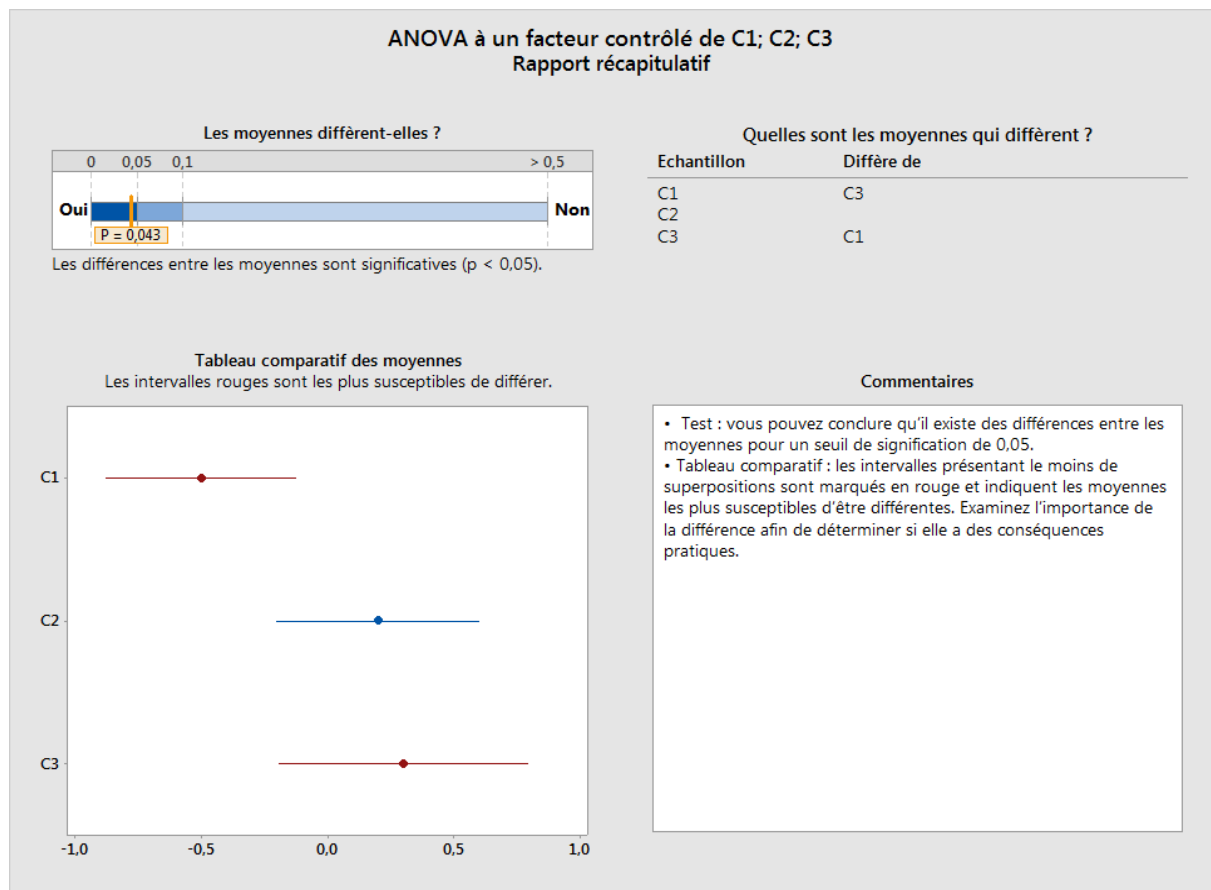


Figure 6 Test significatif : les intervalles apparaissent en rouge même lorsqu'ils se chevauchent

Si le test ne rejette pas l'hypothèse nulle, aucun intervalle n'apparaît en rouge, même si certains d'entre eux ne se chevauchent pas (voir figure 7 ci-dessous). Bien que les intervalles impliquent qu'il existe des différences entre les moyennes, gardez à l'esprit que le non-rejet de l'hypothèse nulle ne revient pas à conclure que l'hypothèse nulle est vraie. Cela indique seulement que les différences observées ne sont pas assez importantes pour exclure la possibilité qu'elles soient dues au hasard. Il convient aussi de souligner que l'écart entre les intervalles qui ne se chevauchent pas sera généralement très faible dans ce cas. Par conséquent, de petites différences restent cohérentes avec les intervalles et n'indiquent pas nécessairement de différence susceptible d'avoir des conséquences pratiques.

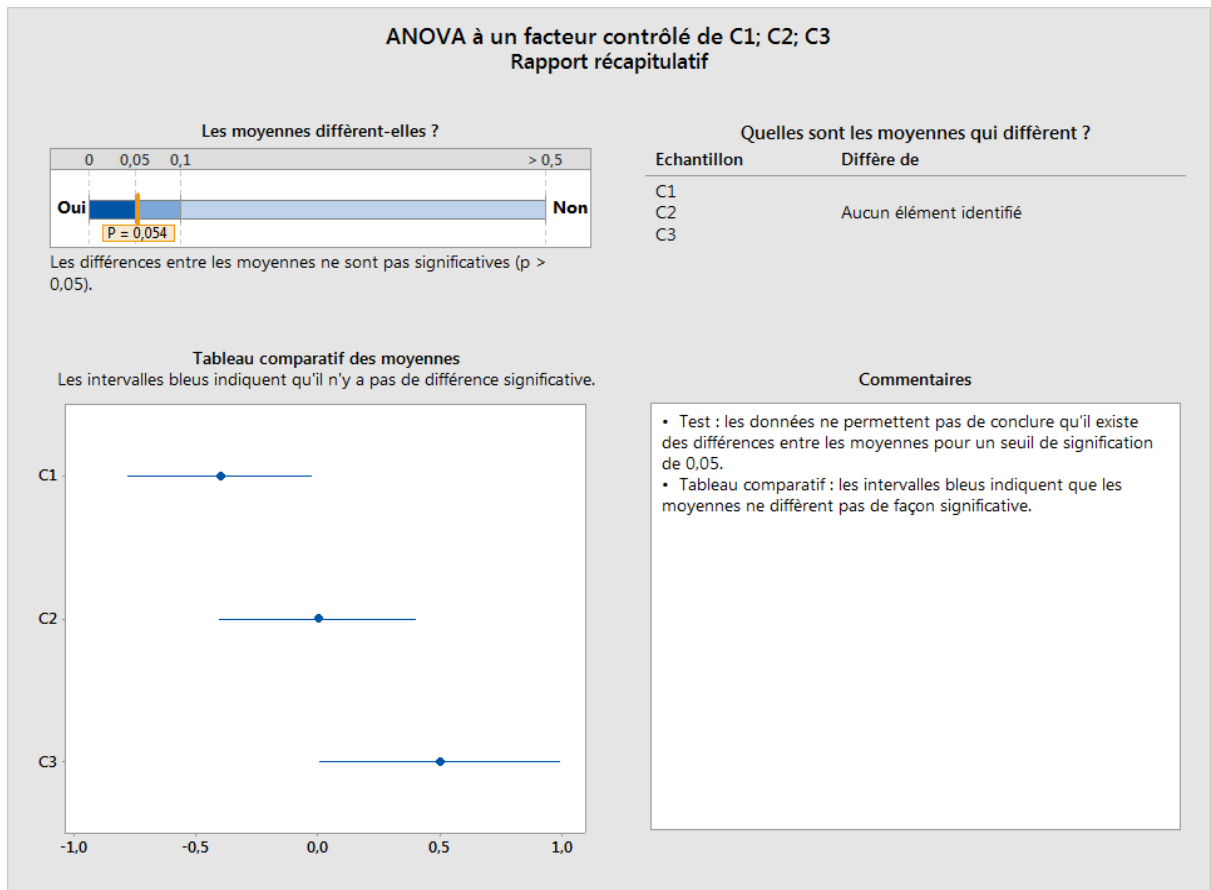


Figure 7 Echec du test : aucun intervalle n'apparaît en rouge, même lorsque des intervalles d'échantillons ne se chevauchent pas

Annexe C : effectif d'échantillon

Dans l'ANOVA à un facteur contrôlé, les paramètres testés sont les moyennes de populations $\mu_1, \mu_2, \dots, \mu_k$ des différents groupes ou populations. Lorsqu'ils sont tous égaux, les paramètres satisfont l'hypothèse nulle. S'il existe des différences entre les moyennes, ils satisfont l'hypothèse alternative. La probabilité de rejet de l'hypothèse nulle ne doit pas être supérieure à α pour les moyennes qui satisfont l'hypothèse nulle. Les probabilités réelles dépendent de l'écart type des lois de distribution, ainsi que des effectifs d'échantillons. La puissance de détection d'un écart par rapport à l'hypothèse nulle augmente lorsque les écarts types sont faibles ou lorsque les échantillons sont importants.

Nous pouvons calculer la puissance du test F pour l'hypothèse de normalité avec des écarts types égaux en utilisant une loi F non-centrale. Le paramètre de non-centralité est :

$$\theta_F = \sum_{i=1}^k n_i (\mu_i - \mu)^2 / \sigma^2$$

où μ représente la moyenne pondérée des moyennes

$$\mu = \sum_{i=1}^k n_i \mu_i / \sum_{i=1}^k n_i$$

et σ est l'écart type, qui est supposé constant. Tous les autres éléments étant égaux, la puissance augmente avec θ_F . C'est précisément la raison pour laquelle la puissance augmente à mesure que les moyennes s'écartent de l'hypothèse nulle.

Contrairement au test F, le test de Welch ne possède pas de formule de puissance exacte. Toutefois, nous allons nous pencher sur deux formules par approximation qui offrent de relativement bons résultats. La première utilise une loi F non centrale, comme pour le calcul de la puissance du test F. Le paramètre de non-centralité que nous utilisons reste de la forme suivante :

$$\theta_W = \sum_{i=1}^k w_i (\mu_i - \mu)^2$$

où μ représente la moyenne pondérée :

$$\mu = \sum_{i=1}^k w_i \mu_i / \sum_{j=1}^k w_j$$

mais où les pondérations dépendent des écarts types et des effectifs d'échantillons, à savoir $w_i = n_i / \sigma_i^2$ ou $w_i = n_i / s_i^2$, selon que nous simulons les résultats pour des écarts types σ_i^2 connus ou que nous estimions la puissance en fonction des écarts types des échantillons s_i^2 . La puissance par approximation est alors calculée comme suit :

$$P(F_{k-1, f, \theta_W} \geq F_{k-1, f, 1-\alpha})$$

ou les degrés de liberté du dénominateur sont

$$f = \frac{k^2 - 1}{3 \sum_{i=1}^k (1 - w_i / \sum_{j=1}^k w_j) / (n_i - 1)}$$

Comme indiqué ci-dessous, cette méthode fournit des approximations relativement correctes de la puissance observée dans les simulations. Ainsi, bien que nous utilisions une approximation différente pour le calcul de la puissance dans le menu Assistant, la méthode précédente fournit des informations utiles et nous permet de sélectionner la configuration de moyennes avec laquelle nous calculons la puissance dans le menu Assistant.

Configuration de moyennes

Conformément à l'approche utilisée pour calculer la puissance et l'effectif d'échantillon dans Minitab (Stat > ANOVA > A un facteur), l'Assistant ne demande pas à l'utilisateur un ensemble complet de moyennes pour évaluer la puissance. Il lui demande d'indiquer une différence de moyenne susceptible d'avoir des conséquences d'un point de vue pratique. Or, il existe un nombre infini de configurations de moyennes pour lesquelles la différence entre la plus grande et la plus petite moyenne correspond à cette valeur donnée. Par exemple, tous les ensembles de moyennes suivants présentent une différence maximale de 10 :

$$\mu_1 = 0, \mu_2 = 5, \mu_3 = 5, \mu_4 = 5, \mu_5 = 10 ;$$

$$\mu_1 = 5, \mu_2 = 0, \mu_3 = 10, \mu_4 = 10, \mu_5 = 0 ;$$

$$\mu_1 = 0, \mu_2 = 10, \mu_3 = 0, \mu_4 = 0, \mu_5 = 0 ;$$

et les possibilités sont infinies.

Nous suivons la méthode utilisée pour calculer la puissance et l'effectif d'échantillon dans Minitab (Stat > Puissance et effectif de l'échantillon > ANOVA à un facteur contrôlé). En d'autres termes, nous choisissons un cas où toutes les moyennes sauf deux correspondent à la moyenne pondérée des moyennes et où les deux moyennes restantes présentent une différence égale à la valeur indiquée. Toutefois, étant donné la possibilité de variances et d'effectifs d'échantillons inégaux, le paramètre de non-centralité (et donc la puissance) dépend encore des deux moyennes supposées différer.

Imaginons une configuration de moyennes μ_1, \dots, μ_k , où toutes les moyennes sauf deux sont égales à la moyenne globale pondérée μ , et deux moyennes, mettons $\mu_i > \mu_j$, diffèrent l'une de l'autre et de la moyenne globale. Soit $\Delta = \mu_i - \mu_j$ la différence entre les moyennes. Soit $\Delta_i = \mu_i - \mu$ et $\Delta_j = \mu - \mu_j$. Alors $\Delta = \Delta_i + \Delta_j$. Etant donné que μ représente la moyenne pondérée de l'ensemble des k moyennes et que $(k - 2)$ des moyennes sont supposées égales à μ :

$$\mu = \left[\sum_{l \neq i, j} w_l \mu_l + w_i(\mu + \Delta_i) + w_j(\mu - \Delta_j) \right] / \sum_{l=1}^k w_l = \mu + (w_i \Delta_i - w_j \Delta_j) / \sum_{l=1}^k w_l .$$

Alors :

$$w_i \Delta_i = w_j \Delta_j = w_j (\Delta - \Delta_i) ,$$

et donc,

$$\Delta_i = \frac{w_j}{w_i + w_j} \Delta$$

$$\Delta_j = \frac{w_i}{w_i + w_j} \Delta$$

Pour cette configuration de moyennes particulière, nous pouvons calculer le paramètre de non-centralité lié au test de Welch :

$$\begin{aligned}\theta_W &= w_i(\mu_i - \mu)^2 + w_j(\mu_j - \mu)^2 \\ &= \frac{w_i w_j^2 \Delta^2 + w_j w_i^2 \Delta^2}{(w_i + w_j)^2} = \frac{w_i w_j \Delta^2}{w_i + w_j}\end{aligned}$$

Cette quantité augmente avec w_i pour une valeur de w_j fixe, et vice versa. Par conséquent, cette valeur est maximisée avec la paire (i, j) présentant les deux pondérations les plus élevées et minimisée avec la paire présentant les pondérations les plus faibles. Tous les calculs de puissance prennent en compte ces deux cas extrêmes, qui minimisent et maximisent la puissance, en partant de l'hypothèse selon laquelle deux moyennes, exactement, diffèrent de la moyenne pondérée globale.

Si vous indiquez une différence pour le test, les valeurs de puissance minimale et maximale sont évaluées pour cette différence. Sur les rapports, l'étendue de ces puissances est évaluée en fonction d'une barre utilisant un code couleur, les puissances égales ou inférieures à 60 % apparaissant en rouge, celles égales ou supérieures à 90 % apparaissant en vert et celles comprises entre 60 % et 90 % apparaissant en jaune. Les résultats du rapport dépendent de l'endroit où se situe l'étendue des puissances sur cette échelle de couleurs. Si l'ensemble de l'étendue se situe dans la zone rouge, la puissance est inférieure ou égale à 60 %, quelle que soit la paire de groupes ; dans ce cas, l'icône rouge apparaît sur le rapport, indiquant une puissance insuffisante. Si l'ensemble de l'étendue se situe dans la zone verte, la puissance est d'au moins 90 %, quelle que soit la paire de groupes ; dans ce cas, l'icône verte apparaît sur le rapport, indiquant que la puissance est suffisante. Toutes les autres conditions sont traitées comme des situations intermédiaires et sont indiquées par une icône jaune sur le rapport.

Lorsque la condition "verte" n'est pas remplie, l'Assistant calcule un effectif d'échantillon permettant de satisfaire cette condition en fonction de la différence indiquée par l'utilisateur et des écarts types observés dans l'échantillon. La puissance estimée dépend des effectifs d'échantillons via les pondérations, étant donné que $w_i = n_i/s_i^2$. Si tous les échantillons sont supposés avoir le même effectif, les deux pondérations les plus faibles correspondent aux deux groupes qui présentent les écarts types d'échantillon les plus importants. L'Assistant calcule l'effectif d'échantillon permettant d'obtenir une puissance d'au moins 90 % si la différence indiquée est celle qui sépare les deux groupes ayant la plus grande variabilité. Par conséquent, si l'on adopte un effectif d'échantillon de cette taille pour tous les groupes, l'étendue complète des valeurs de puissance serait au minimum égale à 90 %, ce qui satisfait la condition "verte".

Si l'utilisateur n'indique pas de différence pour le calcul de la puissance, l'Assistant recherche la plus grande différence pour laquelle la valeur maximale de l'étendue des puissances calculées serait de 60 %. Cette valeur est indiquée à la limite des sections rouge et jaune de la barre, qui correspond à une puissance de 60 %. L'Assistant recherche également la plus petite différence pour laquelle la valeur minimale de l'étendue des puissances calculées serait de 90 %. Cette valeur est indiquée à la limite des sections jaune et verte de la barre, qui correspond à une puissance de 90 %.

Calcul de la puissance

La puissance est calculée à l'aide de l'approximation de Kulinskaya et al. (2003) :

Soit :

$$\lambda = \sum_{i=1}^k w_i (\mu_i - \mu)^2 ,$$

$$A = \sum_{i=1}^k h_i ,$$

$$B = \sum_{i=1}^k w_i (\mu_i - \mu)^2 (1 - w_i/W) / (n_i - 1) ,$$

$$D = \sum_{i=1}^k w_i^2 (\mu_i - \mu)^4 / (n_i - 1) ,$$

$$E = \sum_{i=1}^k w_i^3 (\mu_i - \mu)^6 / (n_i - 1)^2 .$$

Les trois premiers cumulants du numérateur $\sum_{i=1}^k w_i (\bar{x}_i - \hat{\mu})^2$ de la statistique de Welch peuvent être estimés comme suit :

$$\kappa_1 = k - 1 + \lambda + 2A + 2B ,$$

$$\kappa_2 = 2(k - 1 + 2\lambda + 7A + 14B + D) ,$$

$$\kappa_3 = 8(k - 1 + 3\lambda + 15A + 45B + 6D + 2E) .$$

Soit $F_{k-1, f, 1-\alpha}$ le quantile $(1 - \alpha)$ de la loi $F(k - 1, f)$. Gardez à l'esprit que $W^* \geq F_{k-1, f, 1-\alpha}$ est le critère de rejet de l'hypothèse nulle pour un test de Welch d'effectif α .

Soit

$$q = (k - 1) \left[1 + \frac{2(k-2)A}{k^2 - 1} \right] F_{k-1, f, 1-\alpha} ,$$

$$b = \kappa_1 - 2\kappa_2^2 / \kappa_3 ,$$

$$c = \kappa_3 / (4\kappa_2) \text{ [Remarque : l'expression de } c \text{ est indiquée sans parenthèses dans Kulinskaya et al. (2003).]}$$

$$v = 8\kappa_2^3 / \kappa_3^2 .$$

La puissance du test de Welch estimée par approximation est égale à :

$$P(\chi_v^2 \geq \frac{q - b}{c})$$

où χ_v^2 représente une variable aléatoire de Khi deux à v degrés de liberté.

Les résultats suivants comparent la puissance obtenue avec les deux méthodes d'approximation à la puissance simulée pour plusieurs exemples, à partir de 10 000 simulations.

Tableau 3 Comparaison des calculs de puissance des deux méthodes d'approximation par rapport à la puissance simulée

Exemple	alpha	Puissance simulée	Loi F non centrale	Kulinskaya et al.
$\mu : 0, 0, 0, -0.1724, 0.8276$ $\sigma : 2, 2, 2, 2, 4$ $n : 12, 12, 12, 12, 10$	0,10 0,05 0,01	0,1372 0,0739 0,0195	0,135702 0,072563 0,016587	0,135795 0,069512 0,012538
$\mu : 0, 0, 0, -0.3448, 1.6552$ $\sigma : 2, 2, 2, 2, 4$ $n : 12, 12, 12, 12, 10$	0,10 0,05 0,01	0,2498 0,1574 0,0541	0,251064 0,153128 0,045211	0,257455 0,156215 0,042195
$\mu : 0, 0, 0, -0.5172, 2.4828$ $\sigma : 2, 2, 2, 2, 4$ $n : 12, 12, 12, 12, 10$	0,10 0,05 0,01	0,4534 0,3211 0,1273	0,44557 0,311994 0,121225	0,453506 0,321575 0,125065
$\mu : 0, 0, 0, -0.6896, 3.3104$ $\sigma : 2, 2, 2, 2, 4$ $n : 12, 12, 12, 12, 10$	0,10 0,05 0,01	0,662 0,5219 0,2842	0,671317 0,533819 0,271316	0,670296 0,538617 0,282759
$\mu : 0, 0, 0, -0.8620, 4.1380$ $\sigma : 2, 2, 2, 2, 4$ $n : 12, 12, 12, 12, 10$	0,10 0,05 0,01	0,8417 0,7382 0,4883	0,852589 0,752173 0,487601	0,846697 0,746121 0,49323
$\mu : 0, 0, 0, -1.0344, 4.9656$ $\sigma : 2, 2, 2, 2, 4$ $n : 12, 12, 12, 12, 10$	0,10 0,05 0,01	0,9429 0,8866 0,691	0,952077 0,901485 0,711055	0,954929 0,897937 0,703379
$\mu : 0, 0, 0, 0, 0, -0.148148, 1.85185$ $\sigma : 2, 2, 2, 2, 2, 2, 5$ $n : 20, 20, 20, 20, 20, 20, 10$	0,10 0,05 0,01	0,2011 0,1201 0,0385	0,189392 0,108986 0,028986	0,200114 0,11742 0,031456
$\mu : 0, 0, 0, 0, 0, -0.296296, 3.70370$ $\sigma : 2, 2, 2, 2, 2, 2, 5$ $n : 20, 20, 20, 20, 20, 20, 10$	0,10 0,05 0,01	0,4942 0,3677 0,177	0,485917 0,351593 0,149041	0,500143 0,375296 0,177189
$\mu : 0, 0, 0, 0, 0, -0.444444, 5.55556$ $\sigma : 2, 2, 2, 2, 2, 2, 5$ $n : 20, 20, 20, 20, 20, 20, 10$	0,10 0,05 0,01	0,8125 0,7131 0,4876	0,829702 0,727384 0,474291	0,819542 0,720807 0,49469
$\mu : 0, 0, 0, 0, 0, -0.592593, 7.40741$ $\sigma : 2, 2, 2, 2, 2, 2, 5$ $n : 20, 20, 20, 20, 20, 20, 10$	0,10 0,05 0,01	0,9645 0,9286 0,7938	0,977211 0,949997 0,831174	0,984213 0,949239 0,814067

Exemple	alpha	Puissance simulée	Loi F non centrale	Kulinskaya et al.
$\mu : 0, 0, 0, 0, 0, -0,740741, 9,25926$ $\sigma : 2, 2, 2, 2, 2, 2, 5$ $n : 20, 20, 20, 20, 20, 20, 10$	0,10 0,05 0,01	0,9961 0,9895 0,9528	0,998947 0,996653 0,977536	1,00 1,00 0,98705
$\mu : 0, 0, 0, 0, 0, -0,888889, 11,1111$ $\sigma : 2, 2, 2, 2, 2, 2, 5$ $n : 20, 20, 20, 20, 20, 20, 10$	0,10 0,05 0,01	0,9999 0,9995 0,9943	0,999985 0,999926 0,99891	1,00 1,00 1,00
$\mu : 0, 0, 0, 0, 0, -0,518519, 6,48148$ $\sigma : 2, 2, 2, 2, 2, 2, 5$ $n : 20, 20, 20, 20, 20, 20, 10$	0,10 0,05 0,01	0,9059 0,8403 0,6511	0,929392 0,868721 0,67121	0,924696 0,85672 0,66652
$\mu : 0, 0, 0, 0, 0, -0,5, 0,5$ $\sigma : 2, 2, 2, 2, 2, 2, 2$ $n : 12, 12, 12, 12, 12, 12, 12$	0,10 0,05 0,01	0,187 0,1098 0,0315	0,186658 0,106600 0,027773	0,18329 0,100189 0,021332
$\mu : 0, 0, 0, 0, 0, -1, 1$ $\sigma : 2, 2, 2, 2, 2, 2, 2$ $n : 12, 12, 12, 12, 12, 12, 12$	0,10 0,05 0,01	0,4734 0,3394 0,1378	0,474736 0,338655 0,137788	0,472469 0,33443 0,128693
$\mu : 0, 0, 0, 0, 0, -1,5, 1,5$ $\sigma : 2, 2, 2, 2, 2, 2, 2$ $n : 12, 12, 12, 12, 12, 12, 12$	0,10 0,05 0,01	0,8228 0,7112 0,4391	0,817355 0,707319 0,441154	0,810181 0,698461 0,431868
$\mu : 0, 0, 0, 0, 0, -2, 2$ $\sigma : 2, 2, 2, 2, 2, 2, 2$ $n : 12, 12, 12, 12, 12, 12, 12$	0,10 0,05 0,01	0,9691 0,9312 0,7817	0,973246 0,940585 0,799339	0,973319 0,936546 0,785099
$\mu : 0, 0, 0, 0, 0, -2,5, 2,5$ $\sigma : 2, 2, 2, 2, 2, 2, 2$ $n : 12, 12, 12, 12, 12, 12, 12$	0,10 0,05 0,01	0,9984 0,9936 0,9587	0,998579 0,99533 0,967674	0,999763 0,997481 0,966249
$\mu : 0, 0, 0, 0, 0, -3, 3$ $\sigma : 2, 2, 2, 2, 2, 2, 2$ $n : 12, 12, 12, 12, 12, 12, 12$	0,10 0,05 0,01	1,00 0,9997 0,9959	0,999975 0,99987 0,997927	1,00 1,00 0,99961
$\mu : 0, 0, 0, 0, 0, -3,5, 3,5$ $\sigma : 2, 2, 2, 2, 2, 2, 2$ $n : 12, 12, 12, 12, 12, 12, 12$	0,10 0,05 0,01	1,00 1,00 0,99998	1,00 1,00 0,99995	1,00 1,00 1,00
$\mu : 0, 0, 0, 0, 0, -1,75, 1,75$ $\sigma : 2, 2, 2, 2, 2, 2, 2$ $n : 12, 12, 12, 12, 12, 12, 12$	0,10 0,05 0,01	0,914 0,8418 0,619	0,921225 0,852755 0,633815	0,916652 0,843856 0,620704

Exemple	alpha	Puissance simulée	Loi F non centrale	Kulinskaya et al.
$\mu : 0, -0,5, 0,5$	0,10	0,2548	0,259249	0,257149
$\sigma : 2, 2, 2$	0,05	0,1549	0,160861	0,156251
$n : 12, 12, 12$	0,01	0,0470	0,049045	0,042292
$\mu : 0, -1, 1$	0,10	0,654	0,659073	0,654105
$\sigma : 2, 2, 2$	0,05	0,5205	0,522885	0,515816
$n : 12, 12, 12$	0,01	0,2612	0,26355	0,252469
$\mu : 0, -1,5, 1,5$	0,10	0,9364	0,935939	0,937768
$\sigma : 2, 2, 2$	0,05	0,8747	0,87562	0,872608
$n : 12, 12, 12$	0,01	0,6614	0,664478	0,652563
$\mu : 0, -1,75, 1,75$	0,10	0,981	0,981434	0,986815
$\sigma : 2, 2, 2$	0,05	0,9522	0,9561	0,959796
$n : 12, 12, 12$	0,01	0,8251	0,830726	0,823624
$\mu : 0, -2, 2$	0,10	0,9953	0,995969	0,999332
$\sigma : 2, 2, 2$	0,05	0,9878	0,988175	0,993705
$n : 12, 12, 12$	0,01	0,9308	0,931922	0,933446
$\mu : 0, -2,5, 2,5$	0,10	0,9999	0,999923	1,00
$\sigma : 2, 2, 2$	0,05	0,9997	0,999634	1,00
$n : 12, 12, 12$	0,01	0,9949	0,994725	0,99909
$\mu : 0, -3, 3$	0,10	1,00	1,00	1,00
$\sigma : 2, 2, 2$	0,05	1,00	1,00	1,00
$n : 12, 12, 12$	0,01	0,9999	0,99985	1,00
$\mu : 0, -3,5, 3,5$	0,10	1,00	1,00	1,00
$\sigma : 2, 2, 2$	0,05	1,00	1,00	1,00
$n : 12, 12, 12$	0,01	0,9999	1,00	1,00
$\mu : 0, -0,142857, 0,857143$	0,10	0,1452	0,143156	0,146824
$\sigma : 2, 2, 4$	0,05	0,0790	0,077699	0,077538
$n : 14, 12, 8$	0,01	0,0223	0,018200	0,014338
$\mu : 0, -0,285714, 1,71429$	0,10	0,2765	0,27424	0,286222
$\sigma : 2, 2, 4$	0,05	0,1787	0,170628	0,179469
$n : 14, 12, 8$	0,01	0,0624	0,051588	0,050335
$\mu : 0, -0,428571, 2,57143$	0,10	0,4861	0,476925	0,490018
$\sigma : 2, 2, 4$	0,05	0,3487	0,338626	0,355743
$n : 14, 12, 8$	0,01	0,1467	0,132405	0,141352

Exemple	alpha	Puissance simulée	Loi F non centrale	Kulinskaya et al.
$\mu : 0, -0,50000, 3$	0,10	0,5846	0,588533	0,596795
$\sigma : 2, 2, 4$	0,05	0,4425	0,444491	0,460707
$n : 14, 12, 8$	0,01	0,2107	0,19729	0,212798
$\mu : 0, -0,571429, 3,42857$	0,10	0,6933	0,694684	0,696773
$\sigma : 2, 2, 4$	0,05	0,5631	0,555731	0,567129
$n : 14, 12, 8$	0,01	0,3052	0,279131	0,299302
$\mu : 0, -0,714286, 4,28571$	0,10	0,848	0,861469	0,859329
$\sigma : 2, 2, 4$	0,05	0,7402	0,759703	0,759762
$n : 14, 12, 8$	0,01	0,4871	0,480052	0,497421
$\mu : 0, -0,857143, 5,14286$	0,10	0,9434	0,952562	0,961913
$\sigma : 2, 2, 4$	0,05	0,8869	0,898817	0,902716
$n : 14, 12, 8$	0,01	0,6649	0,687058	0,692591
$\mu : 0, -1, 6$	0,10	0,9849	0,987981	0,999989
$\sigma : 2, 2, 4$	0,05	0,9609	0,967589	0,985049
$n : 14, 12, 8$	0,01	0,8294	0,847436	0,853787
$\mu : 0, -1,14286, 6,85714$	0,10	0,9976	0,997776	1,00
$\sigma : 2, 2, 4$	0,05	0,989	0,99222	1,00
$n : 14, 12, 8$	0,01	0,9222	0,940972	0,96383
$\mu : 1, 2, 3$	0,10	0,8838	0,882194	0,884649
$\sigma : 0,3, 2,4, 3,6$	0,05	0,7995	0,797869	0,802137
$n : 13, 19, 25$	0,01	0,5632	0,556486	0,563208
$\mu : 1, 2, 3$	0,10	0,5649	0,566831	0,565141
$\sigma : 2,77489, 2,77489, 2,77489$	0,05	0,4305	0,431302	0,428126
$n : 13, 19, 25$	0,01	0,1994	0,201329	0,195734

Les résultats ci-dessus sont résumés dans le diagramme suivant, qui indique les écarts entre chaque approximation et la valeur de puissance estimée par simulation.

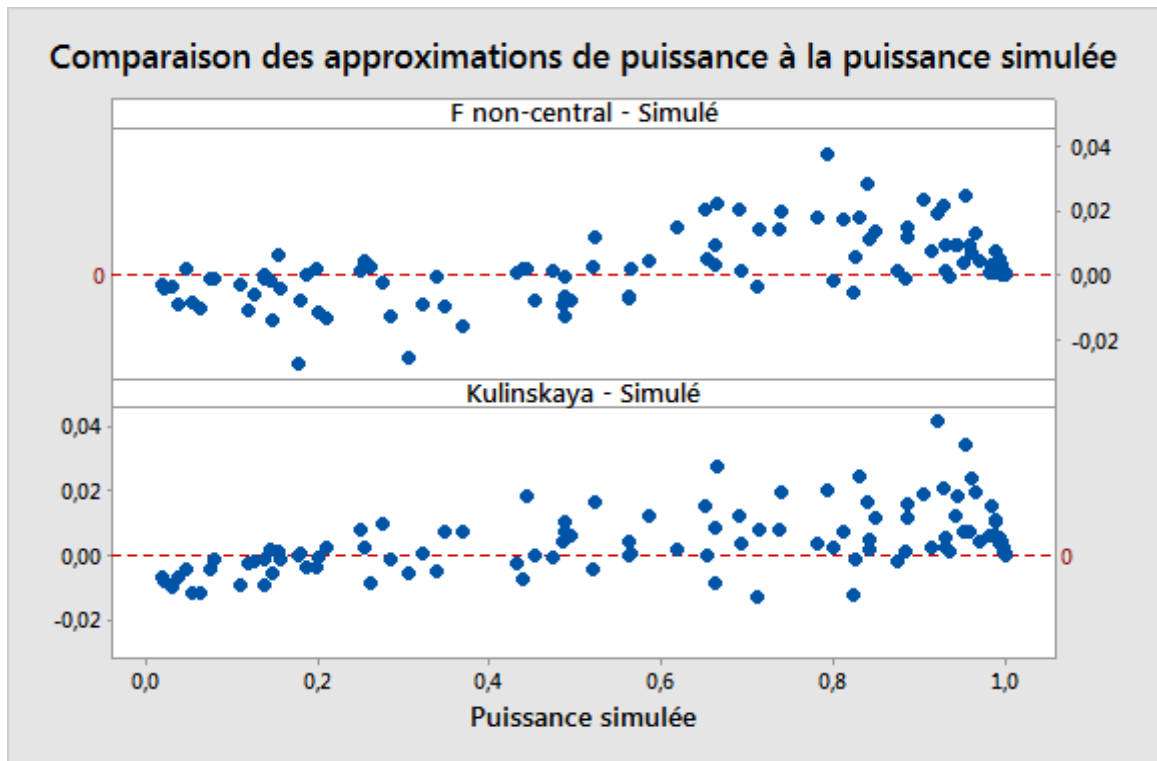


Figure 8 Comparaison des deux approximations de puissance et de la puissance estimée par la simulation.

Annexe D : normalité

La section suivante présente les simulations effectuées pour évaluer la performance du test de Welch et des intervalles de comparaison avec des échantillons petits ou moyens obéissant à plusieurs lois non normales.

Les tableaux ci-dessous regroupent les résultats de simulations effectuées avec différents types de lois pour l'hypothèse nulle d'égalité des moyennes. Dans ces exemples, les écarts types sont eux aussi tous égaux et tous les échantillons ont le même effectif. Le nombre d'échantillons est de $k = 3, 5$ ou 7 .

Chaque cellule indique l'estimation du taux d'erreur de 1ère espèce pour 10 000 simulations. Le seuil de signification cible (valeur α cible) est 0,05.

Tableau 4 Résultats de simulation du test de Welch avec des moyennes égales pour différentes lois

Loi de distribution	Effectif d'échantillon n = 10			Effectif d'échantillon n = 15		
	k = 3	k = 5	k = 7	k = 3	k = 5	k = 7
N(0, 1)	0,0490	0,0486	0,0512	0,0534	0,0522	0,0550
T(3)	0,0371	0,0361	0,0348	0,0353	0,0385	0,0365
T(5)	0,0440	0,0425	0,0439	0,0435	0,0428	0,0428
Laplace(0, 1)	0,0433	0,0354	0,0345	0,0445	0,0397	0,0407
Uniforme(-1, 1)	0,0544	0,0640	0,0718	0,0517	0,0573	0,0585
Bêta(3, 3)	0,0504	0,0577	0,0622	0,0501	0,0538	0,0564
Exponentielle	0,0508	0,0621	0,0748	0,0483	0,0633	0,0779
Khi deux(3)	0,0473	0,0579	0,0753	0,0499	0,0588	0,0703
Khi deux(5)	0,0458	0,0594	0,0643	0,0504	0,0606	0,0679
Khi deux(10)	0,0463	0,0510	0,0585	0,0463	0,0552	0,0567
Bêta(8, 1)	0,0500	0,0622	0,0775	0,0549	0,0653	0,0760

L'ensemble des taux d'erreur de 1ère espèce sont situés à moins de 3 points de pourcentage de la valeur α cible, même avec des effectifs d'échantillons de 10. Nous pouvons constater des écarts plus larges lorsque le nombre de groupes est plus important et lorsque les lois de distribution sont éloignées de la normalité. Avec des effectifs d'échantillons de 10, la probabilité d'acceptation ne s'écarte de plus de 2 points de pourcentage que lorsque $k = 7$. Cela se produit avec la loi uniforme (dont les queues sont beaucoup plus courtes que la normale) et avec les lois exponentielle, du Khi deux(3) et bêta(8, 1), qui sont extrêmement asymétriques. L'augmentation de l'effectif des échantillons à 15 améliore considérablement les résultats pour la loi uniforme, mais pas pour les lois fortement asymétriques.

Nous avons effectué une simulation similaire pour les intervalles de comparaison. Ici, la valeur α simulée représente le nombre de simulations, sur 10 000, pour lesquelles certains intervalles ne se chevauchent pas. La valeur cible est $\alpha = 0,05$.

Tableau 5 Résultats de simulation des intervalles de comparaison avec des moyennes égales pour différentes lois

Loi de distribution	Effectif d'échantillon n = 10			Effectif d'échantillon n = 15		
	k = 3	k = 5	k = 7	k = 3	k = 5	k = 7
N(0, 1)	0,0493	0,0494	0,0469	0,0538	0,0518	0,0561
T(3)	0,0378	0,0321	0,0254	0,0347	0,0343	0,0289
T(5)	0,0449	0,0399	0,0361	0,0447	0,0444	0,0412
Laplace(0, 1)	0,0438	0,0305	0,0246	0,0456	0,0366	0,0348
Uniforme(-1, 1)	0,0559	0,0605	0,0699	0,0534	0,0607	0,0590
Bêta(3, 3)	0,0515	0,0569	0,0615	0,0510	0,0553	0,0568
Exponentielle	0,0353	0,0254	0,0207	0,0346	0,0310	0,0275
Khi deux(3)	0,0375	0,0305	0,0296	0,0384	0,0359	0,0339
Khi deux(5)	0,0405	0,0390	0,0353	0,0417	0,0433	0,0416
Khi deux(10)	0,0425	0,0428	0,0447	0,0435	0,0476	0,0464
Bêta(8, 1)	0,0381	0,0352	0,0287	0,0459	0,0428	0,0403

Comme pour le test de Welch, l'ensemble des taux d'erreur de 1ère espèce sont à moins de 3 points de pourcentage de la cible α , même avec des effectifs d'échantillons de 10. Nous pouvons constater des écarts plus larges lorsque le nombre d'échantillons est plus important et lorsque les lois de distribution sont éloignées de la normalité. Avec un effectif d'échantillons de 10, les taux d'erreur de 1ère espèce sont parfois éloignés de la cible de plus de 2 points de pourcentage lorsque $k = 7$ (et, dans un cas, pour $k = 5$). Cela se produit avec la loi T à 3 degrés de liberté, dont les queues sont extrêmement lourdes, la loi de Laplace et les lois à forte asymétrie que sont la loi exponentielle et la loi du Khi deux(3). L'augmentation de l'effectif des échantillons à 15 améliore les résultats, puisque seules la loi T(3) et les lois exponentielles présentent encore des valeurs simulées α hors-cible de plus de 2 points de pourcentage. Notez que, contrairement aux résultats du test de Welch, les écarts des intervalles de comparaison les plus importants vont dans le sens d'une estimation plus prudente.

La fonction d'ANOVA à un facteur contrôlé de l'Assistant permet d'utiliser jusqu'à $k = 12$ échantillons. Aussi, nous allons maintenant évaluer les résultats pour plus de 7 échantillons. Le tableau ci-dessous indique les taux d'erreur de 1ère espèce obtenus à l'aide du test de Welch pour des données non normales issues de $k = 9$ groupes. Là encore, la valeur alpha cible est de $\alpha = 0,05$.

Tableau 6 Résultats de simulation du test de Welch pour différentes lois avec 9 échantillons

Loi de distribution	k = 9
T(3)	0,0362
T(5)	0,0426
Laplace(0, 1)	0,0402
Uniforme(-1, 1)	0,0625
Bêta(3, 3)	0,0584
Exponentielle	0,0885
Khi deux(3)	0,0774
Khi deux(5)	0,0686
Khi deux(10)	0,0581
Bêta(8, 1)	0,0863

Comme prévu, les lois très asymétriques sont celles qui présentent les écarts les plus importants par rapport à la valeur cible α . Toutefois, aucun taux d'erreur ne s'écarte de la cible de plus de 4 points de pourcentage, même si l'écart correspondant à la loi exponentielle en est proche. Le rapport considère un effectif d'échantillons de 15 comme suffisant pour ne pas signaler un problème de données non normales, car tous les résultats sont proches de la valeur cible α .

Les résultats obtenus avec un effectif d'échantillon de $n = 15$ sont moins concluants lorsque l'on utilise $k = 12$ échantillons. Nous examinons ci-dessous les résultats simulés du test de Welch pour différents effectifs d'échantillons obéissant à des lois s'écartant très fortement de la normale, ce qui nous permettra d'élaborer un critère raisonnable pour évaluer l'effectif d'échantillon.

Tableau 7 Résultats de simulation du test de Welch pour différentes lois avec 12 échantillons

n	T(3)	Uniforme	Khi deux(5)
10	0,0397	0,0918	0,0792
15	0,0351	0,0695	0,0717
20	0,0362	0,0622	0,0671
30	0,0408	0,0573	0,0657

Avec ces lois, un échantillon de $n = 15$ est acceptable si nous sommes prêts à accepter un écart légèrement supérieur à 2 points de pourcentage par rapport à la valeur cible α . Pour que l'écart reste toujours inférieur à 2 points de pourcentage, l'effectif de l'échantillon doit être de 20. A présent, examinons les résultats obtenus avec la loi du Khi deux(3) et la loi exponentielle, qui sont les plus asymétriques.

Tableau 8 . Résultats de simulation du test de Welch avec la loi du Khi deux(3) et la loi exponentielle avec 12 échantillons

n	Khi deux(3)	Exponentielle
10	0,1013	0,1064
15	0,0854	0,1079
20	0,0850	0,0951
30	0,0746	0,0829
40	0,0727	0,0735
50	0,0675	0,0694

Ces lois très asymétriques représentent une plus grande difficulté. Si nous sommes prêts à accepter un écart bien supérieur à 3 points de pourcentage par rapport à la cible, $n = 15$ peut être considéré comme un effectif suffisant, même pour la loi du Khi deux(3). En revanche, la loi exponentielle requiert une valeur se rapprochant de $n = 30$. Bien que la définition d'un critère d'effectif d'échantillon soit quelque peu arbitraire et qu'une valeur de $n = 20$ offre de très bons résultats pour un grand nombre de lois et des résultats plus faibles pour des lois très asymétriques, nous utilisons $n = 20$ comme effectif d'échantillon minimal recommandé pour un nombre d'échantillons compris entre 10 et 12. Cependant, si un écart faible est requis même pour des lois très asymétriques, il est clairement préférable d'utiliser de plus grands échantillons.