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.

Tests d'écarts types (2 échantillons ou plus)

Généralités

L'Assistant Minitab comprend deux analyses permettant de comparer des échantillons indépendants afin de déterminer si leur variabilité diffère de façon significative. Le test d'écart type à 2 échantillons compare les écarts types de deux échantillons ; le test des écarts types, ceux de plus de 2 échantillons. Dans cette étude, nous appelons plans à 2 échantillons les plans à k échantillons pour lesquels k = 2, et plans à échantillons multiples les plans à k échantillons pour lesquels k > 2. Ces deux types de plans sont généralement étudiés séparément (voir l'Annexe A).

L'écart type étant la racine carrée de la variance, un test d'hypothèse comparant les écarts types équivaut à un test d'hypothèse qui compare les variances. De nombreuses méthodes statistiques ont été développées pour comparer les variances de deux populations ou plus. Parmi ces tests, celui de Levene/Brown-Forsythe est l'un des plus robustes et des plus fréquemment utilisés. Toutefois, la puissance du test de Levene/Brown-Forsythe n'est pas aussi satisfaisante que ses propriétés de calcul du taux d'erreur de 1ère espèce dans des plans à 2 échantillons. Pan (1999) démontre que pour certaines populations, notamment la population normale, la puissance de ce test dans des plans à 2 échantillons présente une borne supérieure pouvant être bien inférieure à 1, quelle que soit la valeur de la différence entre les écarts types. En d'autres termes, pour ces types de données, le test est plus à même de conclure qu'il n'existe pas de différence entre les écarts types, quelle que soit l'importance de la différence. C'est pourquoi l'Assistant utilise un test plus récent, le test de Bonett, pour le test d'écarts types à 2 échantillons. Pour le test des écarts types avec des plans à échantillons multiples, l'Assistant utilise une procédure de comparaisons multiples.

Le test de Bonett (2006) est une version modifiée du test d'égalité de deux variances de Layard (1978) et permet d'obtenir de meilleurs résultats avec de faibles échantillons. Banga et Fox (2013 A) dérivent les intervalles de confiance associés au test de Bonett et observent

que ces derniers sont aussi précis que ceux associés au test de Levene/Brown-Forsythe et offrent même un degré d'exactitude supérieur pour la plupart des lois. En outre, Banga et Fox (2013 A) ont établi que le test de Bonett est aussi robuste que celui de Levene/Brown-Forsythe, et qu'il est plus puissant pour la plupart des lois.

La procédure de comparaisons multiples comprend un test global d'homogénéité, ou d'égalité, des écarts types (ou des variances) pour plusieurs échantillons ; ce test s'appuie sur les intervalles de comparaison de chaque paire d'écarts types. Les intervalles de comparaison sont calculés de telle sorte que les résultats du test de comparaisons multiples sont significatifs si, et seulement si, au moins deux d'intervalles de comparaison ne se chevauchent pas. Banga et Fox (2013 B) démontrent que les propriétés du test de comparaisons multiples pour le calcul des taux d'erreur de 1ère et 2ème espèce sont semblables à celles du test de Levene/Brown-Forsythe pour la plupart des lois. L'un des grands avantages du test de comparaisons multiples est qu'il permet de visualiser graphiquement les intervalles de comparaison, ce qui permet d'identifier efficacement les échantillons qui présentent des écarts types différents. Pour les plans à 2 échantillons, le test de comparaisons multiples équivaut au test de Bonett.

Dans cette étude, nous évaluons la validité du test de Bonett et du test de comparaisons multiples pour différentes lois de distribution des données et différents effectifs d'échantillons. Nous étudions également les résultats de l'analyse de puissance et d'effectif d'échantillon du test de Bonett, qui est fondée sur une méthode d'approximation pour grands échantillons. Pour tenir compte de ces facteurs, nous avons développé les vérifications suivantes, que l'Assistant applique automatiquement à vos données et dont il affiche ensuite les résultats dans le rapport :

- Données aberrantes
- Normalité
- Validité du test
- Effectif d'échantillon (test d'écart type à 2 échantillons uniquement)

Méthodes de test des écarts types

Validité du test de Bonett et du test de comparaisons multiples

Dans leur étude comparative des tests d'égalité des variances, Conover et al. (1981) ont établi que le test de Levene/Brown-Forsythe était celui qui fournissait les meilleurs résultats pour l'évaluation des taux d'erreur de 1ère et 2ème espèce. Depuis, d'autres méthodes ont été proposées pour tester l'égalité des variances dans des plans à 2 échantillons ou à échantillons multiples (Pan, 1999 ; Shoemaker, 2003 ; Bonett, 2006). Pan démontre notamment que le test de Levene/Brown-Forsythe, malgré sa robustesse et sa facilité d'interprétation, n'est pas suffisamment puissant pour détecter d'importantes différences entre deux écarts types lorsque les échantillons proviennent de certains types de populations, dont la population normale. En raison de cette limitation critique, l'Assistant utilise le test de Bonett pour le test d'écart type à 2 échantillons (voir l'Annexe A ou Banga et Fox, 2013 A). Pour le test des écarts types avec plus de 2 échantillons, l'Assistant utilise une procédure de comparaisons multiples, avec des intervalles de comparaison permettant d'identifier visuellement les échantillons qui présentent des écarts types différents, lorsque le test de comparaisons multiples est significatif (voir l'annexe A et Banga et Fox, 2013 B).

Objectif

Nous souhaitions d'une part évaluer les résultats du test de Bonett lors de la comparaison des écarts types de deux populations. D'autre part, nous souhaitions évaluer les performances du test de comparaisons multiples lors de la comparaison des écarts types de plus de deux populations. En particulier, nous souhaitions évaluer la validité de ces tests lorsqu'ils sont réalisés sur des échantillons ayant des effectifs différents et obéissant à des types de lois distincts.

Méthode

Les méthodes statistiques utilisées pour le test de Bonett et le test de comparaisons multiples sont définies dans l'Annexe A. Pour évaluer la validité des tests, nous devions vérifier si les taux d'erreurs de 1ère espèce calculés étaient proches du seuil de signification cible (valeur alpha) sous différentes conditions. Pour cela, nous avons réalisé une série de simulations pour évaluer la validité du test de Bonett lors de la comparaison des écarts types de deux échantillons indépendants, puis d'autres séries de simulations pour déterminer la validité du test de comparaisons multiples lors de la comparaison des écarts types de plusieurs (k) échantillons indépendants, lorsque k > 2.

Nous avons créé 10 000 paires ou ensembles de k échantillons aléatoires, présentant différents effectifs et obéissant à différentes lois au sein de plans équilibrés et non équilibrés. Nous avons ensuite réalisé un test bilatéral de Bonett pour comparer les écarts types des deux échantillons, dans un cas, ou, dans le second cas, effectué un test de comparaisons multiples pour comparer les écarts types des k échantillons de chaque expérience avec un seuil de signification cible de $\alpha = 0{,}05$. Nous avons compté le nombre de fois, sur

les 10 000 répliques, où le test rejetait l'hypothèse nulle (alors que les écarts types réels étaient égaux), puis comparé cette proportion (appelée seuil de signification simulé) au seuil de signification cible. Si le test est efficace, le seuil de signification simulé, qui représente le taux d'erreur de 1ère espèce réel, devrait être très proche du seuil de signification cible. Pour plus de détails sur les méthodes spécifiques utilisées dans les simulations à 2 et k échantillons, reportez-vous à l'Annexe B.

Les résultats

Pour les comparaisons de 2 échantillons, les taux d'erreur de 1ère espèce simulés du test de Bonett sont proches du seuil de signification cible avec des effectifs d'échantillons modérés ou élevés, indépendamment de la loi et du type de plan (équilibré ou non équilibré). Toutefois, lorsque de faibles échantillons sont créés à partir de populations particulièrement asymétriques, le test de Bonett est généralement prudent et présente des taux d'erreurs de 1ère espèce légèrement inférieurs au seuil de signification cible (c'est-à-dire le taux d'erreur de 1ère espèce cible).

Pour les comparaisons d'échantillons multiples, les taux d'erreurs de 1ère espèce du test de comparaisons multiples sont proches du seuil de signification cible avec des effectifs d'échantillons modérés ou élevés, indépendamment de la loi et du type de plan (équilibré ou non équilibré). Toutefois, pour des échantillons faibles et extrêmement asymétriques, le test est généralement moins prudent et présente des taux d'erreurs de 1ère espèce supérieurs au seuil de signification cible lorsque le nombre d'échantillons du plan est élevé.

Les résultats de nos études corroborent ceux de Banga et Fox (2013 A) et (2013 B). Nous avons conclu que le test de Bonett et le test de comparaisons multiples obtenaient de bons résultats lorsque l'effectif d'échantillon le plus faible est d'au moins 20. Par conséquent, nous utilisons cet effectif minimal comme condition pour vérifier la validité du test dans le rapport de l'Assistant (voir la section Vérification des données).

Intervalles de comparaison

Lorsqu'un test de comparaison de deux écarts types ou plus est statistiquement significatif et indique qu'au moins un des écarts types 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 de chaque échantillon 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.

Objectif

Nous souhaitions développer une méthode pour calculer des intervalles de comparaison individuels pouvant être utilisés pour effectuer un test global de l'homogénéité des variances et pour identifier les échantillons qui présentent des variances différentes lorsque le test global est significatif. La procédure de comparaisons multiples s'appuie sur une condition essentielle, qui veut que le test global ne soit significatif que si, et seulement si, au moins

deux intervalles de comparaison ne se chevauchent pas, ce qui indique que les écarts types d'au moins deux échantillons diffèrent.

Méthode

La procédure de comparaisons multiples que nous utilisons pour comparer plusieurs écarts types est dérivée des comparaisons multiples deux à deux. Chaque paire d'échantillons est comparée à l'aide du test d'égalité des écarts types de Bonett (2006) pour deux populations. Les comparaisons deux à deux utilisent une méthode de correction de multiplicité qui s'appuie sur une approximation pour grands échantillons décrite dans Nayakama (2009). Cette approximation pour grands échantillons est préférable à la correction de Bonferroni fréquemment utilisée, car cette dernière tend à produire des résultats plus prudents à mesure que le nombre d'échantillons augmente. Enfin, les intervalles de comparaison sont calculés à partir des comparaisons deux à deux utilisés dans la procédure de meilleure d'approximation de Hochberg et al. (1982). Pour plus de détails, reportez-vous à l'Annexe A.

Les résultats

La procédure de comparaisons multiples satisfait l'exigence qui veut que le test global de l'égalité des écarts types ne soit significatif que si, et seulement si, au moins deux intervalles de comparaison ne se chevauchent pas. Si le test global n'est pas significatif, tous les intervalles de comparaison doivent se chevaucher.

L'Assistant affiche les intervalles de comparaison dans le tableau comparatif des écarts types du rapport récapitulatif. Conjointement au graphique, l'Assistant affiche la valeur de p du test de comparaisons multiples, qui est le test global de l'homogénéité des écarts types . Lorsque le test des écarts types est statistiquement significatif, tout intervalle de comparaison qui ne chevauche pas au moins un autre intervalle apparaît en rouge. Si le test des écarts types n'est pas statistiquement significatif, aucun intervalle n'apparaît en rouge.

Efficacité de la puissance théorique (plans à 2 échantillons uniquement)

Les fonctions puissance théorique des tests de Bonett et des tests de comparaisons multiples sont nécessaires pour la planification des effectifs d'échantillons. Pour les plans à 2 échantillons, il est possible de dériver une fonction de puissance théorique par approximation à partir de méthodes reposant sur la théorie des grands échantillons. Etant donné que cette fonction est le résultat de méthodes d'approximation pour grands échantillons, nous devons en évaluer les propriétés lorsque le test est réalisé avec de petits échantillons générés à partir de lois normales et non normales. Toutefois, lors de la comparaison des écarts types de plus de deux groupes, il est difficile d'obtenir la fonction puissance théorique du test de comparaisons multiples.

Objectif

Nous souhaitions déterminer si nous pouvions utiliser cette fonction puissance théorique dérivée de méthodes d'approximation pour grands échantillons afin d'évaluer la puissance et l'effectif d'échantillon requis pour le test d'écart type à 2 échantillons dans l'Assistant. Pour ce faire, il nous fallait déterminer si la fonction puissance théorique par approximation

reflétait avec précision la puissance réelle du test de Bonett lorsque celui-ci était réalisé sur des données obéissant à plusieurs types de lois, aussi bien normales que non normales.

Méthode

La fonction puissance théorique par approximation du test de Bonett pour les plans à 2 échantillons est dérivée dans l'Annexe C.

Nous avons réalisé des simulations afin d'estimer les niveaux de puissance réels (que nous appelons niveaux de puissance simulés) obtenus avec le test de Bonett. Dans un premier temps, nous avons créé des paires d'échantillons aléatoires présentant des effectifs variés et obéissant à des lois différentes, aussi bien normales que non normales. Pour chaque loi, nous avons effectué le test de Bonett sur chacune des 10 000 paires de répliques d'échantillons. Pour chaque paire d'effectifs d'échantillons, nous avons calculé la puissance simulée du test pour la détection d'une différence donnée, exprimée comme la fraction des 10 000 paires d'échantillons pour lesquelles le test est significatif. Pour effectuer la comparaison, nous avons également calculé le niveau de puissance correspondant obtenu avec la fonction puissance théorique par approximation du test. Si l'approximation est bonne, les niveaux de puissance théoriques et simulés doivent être proches. Pour plus de détails, reportez-vous à l'Annexe D.

Les résultats

Nos simulations ont montré que, pour la plupart des lois, les résultats des fonctions puissance théorique et simulée du test de Bonett sont quasiment égaux avec de faibles échantillons et tendent à se rapprocher lorsque l'effectif d'échantillon minimal est d'au moins 20. Pour les lois symétriques et presque symétriques à queues légères ou modérées, les niveaux de puissance théoriques sont légèrement supérieurs aux niveaux de puissance simulés (réels). En revanche, pour les lois asymétriques et à queues lourdes, ils sont inférieurs aux niveaux de puissance simulés (réels). Pour plus de détails, reportez-vous à l'Annexe D.

Dans l'ensemble, nos résultats indiquent que la fonction puissance théorique fournit une base solide pour la planification des effectifs d'échantillons.

Vérification des données

Données 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. Les études par simulation montrent que lorsque les données contiennent des valeurs aberrantes, les tests de Bonett et de comparaisons multiples sont prudents (voir l'Annexe B). Les seuils de signification réels de ces tests sont sensiblement inférieurs au seuil cible, en particulier lorsque l'analyse est réalisée avec de faibles échantillons.

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.

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.

Normalité

Contrairement à la plupart des tests d'égalité des variances, qui s'appuient sur l'hypothèse de normalité, les tests de Bonett et de comparaisons multiples sur l'égalité des écarts types ne formulent pas d'hypothèse sur la loi de distribution spécifique des données.

Objectif

Bien que les tests de Bonett et de comparaisons multiples s'appuient sur des méthodes d'approximation pour grands échantillons, nous souhaitions vérifier qu'ils permettaient d'obtenir de bons résultats avec des données normales et non normales dans de petits échantillons. En outre, nous souhaitions expliquer aux utilisateur de quelle façon la normalité des données influait sur les résultats des tests des écarts types.

Méthode

Pour évaluer la validité des tests sous différentes conditions, nous avons effectué des simulations pour examiner le taux d'erreur de 1ère espèce obtenu avec les tests de Bonett et de comparaisons multiples avec des données normales et non normales pour différents effectifs d'échantillons. Pour plus de détails, reportez-vous à la section Méthodes de test des écarts types et à l'Annexe B.

Les résultats

Nos simulations indiquent que la distribution des données n'a pas d'effet majeur sur les propriétés de calcul du taux d'erreur de 1ère espèce des tests de Bonett ou de comparaison multiple pour des échantillons suffisamment grands (effectif d'échantillon minimal ≥ 20). Les tests produisent des taux d'erreur de 1ère espèce invariablement proches du taux d'erreur cible pour les données normales et non normales.

Conformément à ces résultats sur le taux d'erreur de 1ère espèce, l'Assistant affiche les informations relatives à la normalité dans le rapport.

Pour les plans à 2 échantillons, l'Assistant affiche l'indicateur suivant :

Etat	Condition
	Cette analyse utilise le test de Bonett. Si les échantillons sont suffisamment grands, le test offre de bons résultats pour les données normales comme pour les données non normales.

Pour les plans à échantillons multiples, l'Assistant affiche l'indicateur suivant :

Etat	Condition
	Cette analyse utilise un test de comparaisons multiples. Si les échantillons sont suffisamment grands, le test offre de bons résultats pour les données normales comme pour les données non normales.

Validité du test

Dans la section Méthodes de test des écarts types, nous avons démontré que pour les comparaisons à 2 échantillons et à échantillons multiples (à k échantillons), les tests de Bonett et de comparaisons multiples produisent des taux d'erreur de 1ère espèce proches du taux d'erreur cible lorsque les effectifs d'échantillons sont modérés ou élevés, et ce, aussi bien avec des données normales qu'avec des données non normales, et dans des plans équilibrés et non équilibrés,. En revanche, lorsque les effectifs d'échantillons sont faibles, les tests de Bonett et de comparaisons multiples ne fournissent généralement pas de bons résultats.

Objectif

Nous souhaitions appliquer une règle permettant d'évaluer la validité des résultats du test d'écarts types pour 2 échantillons et pour des échantillons multiples (à k échantillons) en fonction des données de l'utilisateur.

Méthode

Pour évaluer la validité des tests sous différentes conditions, nous avons effectué des simulations pour examiner le taux d'erreur de 1ère espèce obtenu avec les tests de Bonett et de comparaisons multiples pour plusieurs lois, nombres d'échantillons et effectifs d'échantillons, comme décrit précédemment dans la section Méthodes de test des écarts types. Pour plus de détails, reportez-vous à l'Annexe B.

Les résultats

Le test de Bonett et le test de comparaisons multiples obtiennent de bons résultats lorsque l'effectif d'échantillon le plus faible est d'au moins 20. Par conséquent, l'Assistant affiche les indicateurs d'état suivants dans le rapport pour évaluer la validité des tests des écarts types :

Etat	Condition
	Les effectifs d'échantillons sont d'au moins 20, la valeur de p devrait donc être exacte.
	Certains effectifs d'échantillons sont inférieurs à 20, la valeur de p peut ne pas être exacte. Vous devriez utilisez des effectifs d'échantillons d'au moins 20.

Effectif d'échantillon (test d'écart type à 2 échantillons uniquement)

Par définition, un test d'hypothèse statistique vise à collecter des preuves permettant de rejeter l'hypothèse nulle de "non différence". Lorsque l'échantillon est trop faible, la puissance du test peut ne pas être adaptée pour détecter une différence existante, 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 dans la pratique avec une probabilité élevée.

Objectif

Si les données 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 effectifs d'échantillons sont suffisamment grands pour détecter d'importantes différences avec une probabilité élevée, ces effectifs 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 de l'échantillon pour le test d'écart type à 2 échantillons est fondée sur une approximation de la fonction puissance du test de Bonett, qui fournit généralement de bonnes estimations de la fonction puissance réelle du test (voir le récapitulatif des résultats de simulation dans la section Méthode de la partie Efficacité de la fonction puissance théorique).

Les résultats

Lorsque les données ne fournissent pas suffisamment de preuves invalidant l'hypothèse nulle, l'Assistant utilise la fonction puissance par approximation du test de Bonett pour calculer les différences pratiques pouvant être détectées avec une probabilité de 80 % et de 90 % pour l'effectif d'échantillon donné. De plus, si l'utilisateur indique une différence pratique spécifique présentant un intérêt particulier, l'Assistant utilise la fonction puissance du test d'approximation selon la loi normale pour calculer les effectifs d'échantillons qui offrent une probabilité de 80 % et de 90 % de détecter cette différence.

Pour faciliter l'interprétation des résultats, le rapport de l'Assistant pour le test d'écart type à 2 échantillons affiche les indicateurs d'état suivants lors du test de la puissance et de l'effectif d'échantillon :

Etat	Condition
	Le test détecte une différence entre les écarts types, par conséquent la puissance n'est pas un problème. OU La puissance est suffisante. Le test n'a pas détecté de différence entre les écarts types, 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 écarts types, 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 écarts types, 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.

Etat	Condition
	La puissance n'est pas suffisante. Le test n'a pas détecté de différence entre les écarts types, 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.
	Le test n'a pas détecté de différence entre les écarts types. Vous n'avez pas indiqué de différence pratique à détecter. Par conséquent, le rapport indique les différences ayant 80 et 90 % de chances d'être détectées avec les effectifs d'échantillons et la valeur alpha utilisés.

Références

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

Banga, S.J. et Fox, G.D. (2013 A), Sur l'intervalle de confiance robuste de Bonett pour le rapport des écarts types, *Livre blanc, Minitab Inc*.

Banga, S.J. et Fox, G.D. (2013 B), Procédure graphique de comparaisons multiples pour plusieurs écarts types, *Livre blanc, Minitab Inc*.

Bonett, D.G. (2006), Robust confidence interval for a ratio of standard deviations, *Applied Psychological Measurements*, 30, 432-439.

Brown, M.B. et Forsythe, A.B. (1974), Robust tests for the equality of variances, *Journal of the American Statistical Association*, 69, 364-367.

Conover, W.J., Johnson, M.E. et Johnson, M.M. (1981), A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data, *Technometrics*, 23, 351-361.

Gastwirth, J. L. (1982), Statistical properties of a measure of tax assessment uniformity, *Journal of Statistical Planning and Inference*, 6, 1-12.

Hochberg, Y., Weiss G. et Hart, S. (1982), On graphical procedures for multiple comparisons, *Journal of the American Statistical Association*, 77, 767-772.

Layard, M.W.J. (1973), Robust large-sample tests for homogeneity of variances, *Journal of the American Statistical Association*, 68, 195-198.

Levene, H. (1960), Robust tests for equality of variances, In I. Olkin (Ed.), *Probability and statistics* (278-292), Stanford University Press, Palo Alto, Californie.

Nakayama, M.K. (2009), Asymptotically valid single-stage multiple-comparison procedures, *Journal of Statistical Planning and Inference*, 139, 1348-1356.

Pan, G. (1999), On a Levene type test for equality of two variances, *Journal of Statistical Computation and Simulation*, 63, 59-71.

Shoemaker, L. H. (2003), Fixing the F test for equal variances, *The American Statistician*, 57 (2), 105-114.

Annexe A : méthode pour le test de Bonett et le test de comparaisons multiples

Les hypothèses sous-jacentes permettant de calculer des inférences sur les écarts types ou les variances à l'aide de la méthode de Bonett (plans à 2 échantillons) ou la procédure de comparaisons multiples (plans à échantillons multiples) peuvent être décrites comme suit. Soit $X_{11}, \ldots, X_{1n_1}, \ldots, X_{k1}, \ldots, X_{kn_k}$ k échantillons aléatoires indépendants ($k \geq 2$), chacun créé à partir d'une loi ayant respectivement une variance σ_i^2 et une moyenne μ_i inconnue pour $i = 1, \ldots, k$. Supposons que les lois parent des échantillons présentent un aplatissement fini $\gamma = E(Y - \mu)^4/\sigma^4 < \infty$ commun. Cette hypothèse est essentielle pour les calculs théoriques, mais pas pour la plupart des applications pratiques où les échantillons sont suffisamment grands (Banga et Fox, 2013 A).

Méthode A1 : test d'égalité de deux variances de Bonett

Le test de Bonett s'applique uniquement aux plans à 2 échantillons où l'on compare deux variances ou deux écarts types. Ce test est une version modifiée du test d'égalité des variances de Layard (1978) dans les plans à deux échantillons. Un test d'égalité de deux variances bilatéral de Bonett ayant un seuil de signification de α rejette l'hypothèse nulle d'égalité si, et seulement si,

$$\left|\ln(c\, S_1^2/S_2^2)\right| > z_{\alpha/2}\sqrt{\frac{\hat{\gamma}_P - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P - g_2}{n_2 - 1}}$$

où :

S_i est l'écart type de l'échantillon i

$g_i = (n_i - 3)/n_i, i = 1, 2$

$z_{\alpha/2}$ désigne le percentile supérieur $\alpha/2$ de la loi normale standard

$\hat{\gamma}_P$ est l'estimateur d'aplatissement regroupé, exprimé comme suit :

$$\hat{\gamma}_P = (n_1 + n_2)\frac{\sum_{j=1}^{n_1}(X_{1j} - m_1)^4 + \sum_{j=1}^{n_2}(X_{2j} - m_2)^4}{[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]^2}$$

Dans l'expression de l'estimateur d'aplatissement regroupé, m_i est la moyenne tronquée de l'échantillon i, avec la proportion de troncature $1/[2(n_i - 4)^{1/2}]$.

Dans la formule initiale, la constante c permet d'ajuster légèrement l'échantillon pour réduire l'effet de queues inégales sur le calcul des probabilités d'erreur dans les plans non équilibrés. Cette constante est exprimée de la façon suivante : $c = c_1/c_2$, où

$$c_{\mathrm{i}} = \frac{n_{\mathrm{i}}}{n_{\mathrm{i}} - z_{\alpha/2}}, i = 1, 2$$

Si le plan est équilibré, c'est-à-dire si $n_1 = n_2$, la valeur de p du test est calculée de la façon suivante :

$$\mathrm{P} = 2\,\Pr(Z > z)$$

où Z est une variable aléatoire dont la distribution suit une loi normale standard et z est la valeur observée pour les statistiques suivantes, à partir des données disponibles. La statistique est

$$Z = \frac{\ln(C\ S_1^2/S_2^2)}{es}$$

où

$$es = \sqrt{\frac{\hat{\gamma}_P - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P - g_2}{n_2 - 1}}$$

En revanche, si le plan est non équilibré, la valeur de p du test s'obtient comme suit :

$$P = 2\min(\alpha_L, \alpha_U)$$

où $\alpha_L = \Pr(Z > z_L)$ et $\alpha_U = \Pr(Z > z_U)$. La variable z_L est le plus petit zéro (point d'annulation) de la fonction

$$L(z, S_1, S_2, n_1, n_2) = \ln\frac{n_1}{n_2} + \ln\frac{n_2 - z}{n_1 - z} - z\ es + \ln\frac{S_1^2}{S_2^2} - \ln\rho_o^2\,, z < \min(n_1, n_2)$$

et z_U est le plus petit zéro de la fonction $L(z, S_2, S_1, n_2, n_1)$.

Méthode A2 : test de comparaisons multiples et intervalles de comparaison

Soit k groupes ou échantillons indépendants ($k \geq 2$). Notre objectif était de créer un système à k intervalles pour les écarts types de la population, de sorte que le test d'égalité des écarts types soit significatif si, et seulement si, au moins deux des k intervalles ne se chevauchent pas. Ces intervalles sont appelés intervalles de comparaison. Cette méthode de comparaison est semblable aux procédures de comparaisons multiples des moyennes dans les modèles d'ANOVA à un facteur contrôlé, qui ont initialement été développées par Tukey-Kramer, avant d'être généralisées par Hochberg et al. (1982).

Comparaison de deux écarts types

Pour les plans à 2 échantillons, il est possible de calculer directement les intervalles de confiance du rapport des écarts types obtenu avec le test de Bonett pour évaluer la différence entre les écarts types (Banga et Fox, 2013 A). C'est d'ailleurs la méthode que nous utilisons dans la version 17 de Minitab (Stat > Statistiques élémentaires > 2 variances). Toutefois, dans l'Assistant, nous souhaitions fournir des intervalles de comparaison plus faciles à interpréter que l'intervalle de confiance du rapport des écarts types. Pour cela, nous avons utilisé la procédure de Bonett décrite dans la Méthode A1 pour déterminer les intervalles de comparaison de deux échantillons.

Lorsqu'il n'y a que deux échantillons, le test d'égalité des variances de Bonett est significatif si, et seulement si, l'intervalle d'acceptation suivant associé au test ne contient pas la valeur 0.

$$\ln(c_1 S_1^2) - \ln(c_2 S_2^2) \pm z_{\alpha/2} \sqrt{\frac{\hat{\gamma}_P - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P - g_2}{n_2 - 1}}$$

où les estimateurs d'aplatissement regroupé $\hat{\gamma}_P$ et $g_i, i = 1, 2$ sont exprimés de la même façon que précédemment.

Nous déduisons de cet intervalle les deux intervalles de comparaison suivants, de sorte que le test d'égalité des variances ou d'écart type ne soit significatif que si, et seulement si, ils ne se chevauchent pas. Ces deux intervalles sont

$$\left[S_i \sqrt{\mathrm{C_i} \exp(-z_{\alpha/2} V_i)}, S_i \sqrt{C_i \exp(z_{\alpha/2} V_i)} \right], i = 1, 2$$

où

$$V_i = \frac{\sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1}}}{\sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1}} + \sqrt{\frac{\hat{\gamma}_P - g_j}{n_j - 1}}} \sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1} + \frac{\hat{\gamma}_P - g_j}{n_j - 1}}, i = 1, 2; j = 1, 2; i \neq j$$

L'utilisation de ces intervalles pour tester d'égalité des écarts types revient à effectuer un test d'égalité des écarts types de Bonett. Plus précisément, les intervalles ne se chevauchent pas si, et seulement si, le test de Bonett d'égalité de l'écart type est significatif. Toutefois, notez que ces intervalles ne sont pas des intervalles de confiance des écarts types et qu'ils ne sont appropriés que pour les comparaisons multiples d'écarts types. Pour cette raison, Hochberg et al. désignent ce genre d'intervalles servant à comparer des moyennes sous le terme "intervalles d'incertitude". Nous les appelons intervalles de comparaison.

Cette procédure d'intervalles de comparaison étant équivalente au test d'égalité des écarts types de Bonett, la valeur de p utilisée pour cet intervalle de comparaison est identique à celle du test d'égalité de deux écarts types de Bonett décrit précédemment.

Comparaison de plusieurs écarts types

Lorsqu'il existe plus de deux groupes ou échantillons, les k intervalles de comparaison sont déduits sur la base de $k(k-1)/2$ tests d'égalité des écarts types, effectués simultanément, deux à deux, avec un seuil de signification par famille de α. Plus précisément, soit $X_{i1}, \ldots, X_{in_i}$ et $X_{j1}, \ldots, X_{jn_j}$ les données échantillons pour toute paire d'échantillons (i, j). Comme dans les plans à 2 échantillons, le test d'égalité des écarts types pour la paire d'échantillons (i, j) est significatif à un niveau α' déterminé si, et seulement si, l'intervalle

$$\ln(c_i S_i^2) - \ln(c_j S_j^2) \pm z_{\alpha'/2} \sqrt{\frac{\hat{\gamma}_{ij} - g_i}{n_i - 1} + \frac{\hat{\gamma}_{ij} - g_j}{n_j - 1}}$$

ne contient pas la valeur 0. Dans la formule ci-dessus, $\hat{\gamma}_{ij}$ est l'estimateur d'aplatissement regroupé de la paire d'échantillons (i, j), exprimé comme suit :

$$\hat{\gamma}_{ij} = (n_i + n_j)\frac{\sum_{l=1}^{n_i}(X_{il} - m_i)^4 + \sum_{l=1}^{n_j}(X_{jl} - m_j)^4}{\left[(n_i - 1)S_i^2 + (n_j - 1)S_j^2\right]^2}$$

En outre, comme défini précédemment, m_i est la moyenne tronquée de l'échantillon i, avec la proportion de troncature $1/[2(n_i - 4)^{1/2}]$ et

$$g_i = \frac{n_i - 3}{n_i}, g_j = \frac{n_j - 3}{n_j}, c_i = \frac{n_i}{n_i - z_{\alpha'/2}}, c_j = \frac{n_j}{n_j - z_{\alpha'/2}}$$

Etant donné qu'il y a $k(k-1)/2$ tests deux à deux simultanés, le niveau α' doit être choisi de telle sorte que le taux d'erreur réel par famille soit proche du seuil de signification cible α. Il existe un ajustement possible, fondé sur l'approximation de Bonferroni. Toutefois, les corrections de Bonferroni sont connues pour accroître la prudence des résultats à mesure que le nombre d'échantillons du plan augmente. L'approximation selon la loi normale fournie par Nakayama (2008) constitue une meilleure méthode. Avec cette approche, il nous suffit de remplacer $z_{\alpha'/2}$ par $q_{\alpha,k}/\sqrt{2}$, où $q_{\alpha,k}$ constitue le point α le plus élevé de l'étendue de k variables aléatoires indépendantes et distribuées de façon identique selon loi normale standard ; c'est-à-dire

$$\Pr\left(\max_{1\le i<j\le k}\left|Z_i - Z_j\right| \le q_{\alpha,k}\right) = 1 - \alpha$$

où $Z_1, \ldots, Z_k$ sont des variables aléatoires indépendantes, distribuées de façon identique selon loi normale standard.

En outre, avec une méthode semblable à celle de Hochberg et al. (1982), la procédure qui se rapproche le plus de la procédure deux à deux décrite ci-dessus, l'hypothèse nulle d'égalité des écarts types n'est rejetée que si, et seulement si, pour une paire de (i, j) échantillons,

$$\left|\ln(c_i S_i^2) - \ln(c_j S_j^2)\right| > q_{\alpha,k}(V_i + V_j)/\sqrt{2}$$

où V_i est choisi dans le but de minimiser la quantité

$$\sum_{i\ne j}\sum\left(V_i + V_j - b_{ij}\right)^2$$

avec

$$b_{ij} = \sqrt{\frac{\hat{\gamma}_{ij} - g_i}{n_i - 1} + \frac{\hat{\gamma}_{ij} - g_j}{n_j - 1}}$$

Comme l'ont illustré Hochberg et al. (1982), la solution à ce problème consiste à utiliser

$$V_i = \frac{(k-1)\sum_{j\ne i} b_{ij} - \sum\sum_{1\le j<l\le k} b_{jl}}{(k-1)(k-2)}$$

Par conséquent, le test fondé sur la procédure d'approximation est significatif si, et seulement si, au moins une paire des k intervalles suivants ne se chevauchent pas.

$$\left[S_i\sqrt{\mathrm{C_i}\exp(-q_{\alpha,k}V_i/\sqrt{2})}, S_i\sqrt{C_i\exp(q_{\alpha,k}V_i/\sqrt{2})}\right], i = 1, \ldots, k$$

Pour calculer la valeur de p globale associée au test de comparaisons multiples, nous appelons P_{ij} la valeur de p associée à tout paire d'échantillons (i, j). Ainsi, la valeur de p globale associée au test de comparaisons multiples est égale à

$$P = \min\{ P_{ij}, 1 \le i < j \le k\}$$

Pour calculer P_{ij}, nous utilisons l'algorithme fourni pour le plan à 2 échantillons dans la Méthode A1, avec

$$es = V_i + V_j$$

où V_i est exprimé de la façon décrite précédemment.

Plus spécifiquement, si $n_i \neq n_j$

$$P_{ij} = \min(\alpha_L, \alpha_U)$$

où $\alpha_L = \Pr(Q > z_L\sqrt{2})$, $\alpha_U = \Pr(Q > z_U\sqrt{2})$, la variable z_L est le plus petit zéro de la fonction $L(z, S_i, S_j, n_i, n_j)$, la variable z_U est le plus petit zéro de la fonction $L(z, S_j, S_i, n_j, n_i)$ et Q est une variable aléatoire qui présente la loi d'étendue définie précédemment.

Si $n_i = n_j$, alors $P_{ij} = \Pr(Q > |z_o|\sqrt{2})$ où

$$z_o = \frac{\ln S_i^2 - \ln S_j^2}{V_i + V_j}$$

Annexe B : Validité du test de Bonett et du test de comparaisons multiples

Simulation B1 : validité du test de Bonett (modèles à 2 échantillons, plans équilibrés et non équilibrés)

Nous avons créé des paires d'échantillons aléatoires dont les effectifs sont faibles ou modérés et qui obéissent à des lois ayant des propriétés différentes. Ces lois étaient les suivantes :

- La loi normale standard (N(0, 1))
- Des lois de distribution symétriques à queues légères, à savoir la loi de distribution uniforme (U(0,1)) et la loi bêta, avec les deux paramètres définis sur 3 (B(3, 3))
- Des lois de distribution symétriques à queues lourdes, à savoir des lois T à 5 et 10 degrés de liberté (T(5)),T(10)) et la loi de Laplace avec un emplacement 0 et une échelle 1 (Lpl))
- Des lois de distribution asymétriques à queues lourdes, à savoir la loi exponentielle avec une échelle 1 (Exp) et des lois du Khi deux à 5 et 10 degrés de liberté (Khi(5), Khi(10))
- Une loi de distribution à queues lourdes et présentant une asymétrie vers la gauche, plus précisément une loi bêta avec des paramètres définis sur 8 et 1, respectivement (B(8,1))

En outre, pour évaluer l'effet direct des valeurs aberrantes, nous avons généré des paires d'échantillons en utilisant des lois normales contaminées, telles que

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

où p est le paramètre de mélange et $1 - p$ la proportion de contamination (égale à la proportion de valeurs aberrantes). Nous avons sélectionné deux populations normales contaminées pour l'étude : $CN(0{,}9,\ 3)$, où 10 % de la population est constituée de valeurs aberrantes, et $CN(0{,}8,\ 3)$, où 20 % de la population est constituée de valeurs aberrantes. Ces deux distributions sont symétriques et présentent de longues queues à cause des valeurs aberrantes.

Nous avons réalisé un test de Bonett bilatéral avec un seuil de signification cible de $\alpha = 0{,}05$ pour chacune des paires d'échantillons générées avec chaque loi. Etant donné que, dans chaque cas, les seuils de signification simulés s'appuyaient sur 10 000 paires de répliques d'échantillons et que nous avons utilisé un seuil de signification cible de 5 %, le taux d'erreur de la simulation est de : $\sqrt{0{,}95(0{,}05)/10\,000} = 0{,}2\%$.

Les résultats de la simulation sont récapitulés dans le Tableau 1 ci-dessous.

Tableau 1 Seuils de signification simulés pour un test de Bonett bilatéral dans des plans à 2 échantillons
équilibrés et non équilibrés. Le seuil de signification cible est de 0,05.

Loi de distribution	n_1, n_2	Seuil simulé	Loi de distribution	n_1, n_2	Seuil simulé
N(0,1)	10, 10	0,038	Exp	10, 10	0,052
	20, 10	0,043		20, 10	0,051
	20, 20	0,045		20, 20	0,049
	30, 10	0,044		30, 10	0,044
	30, 20	0,046		30, 20	0,042
	25, 25	0,048		25, 25	0,043
	30, 30	0,048		30, 30	0,042
	40, 40	0,051		40, 40	0,042
	50, 50	0,047		50, 50	0,039
T(5)	10, 10	0,044	Khi(5)	10, 10	0,040
	20, 10	0,042		20, 10	0,043
	20, 20	0,046		20, 20	0,040
	30, 10	0,041		30, 10	0,039
	30, 20	0,046		30, 20	0,043
	25, 25	0,048		25, 25	0,042
	30, 30	0,043		30, 30	0,043
	40, 40	0,046		40, 40	0,040
	50, 50	0,050		50, 50	0,039

Loi de distribution	n_1, n_2	Seuil simulé	Loi de distribution	n_1, n_2	Seuil simulé
T(10)	10, 10	0,041	Khi(10)	10, 10	0,044
	20, 10	0,040		20, 10	0,042
	20, 20	0,045		20, 20	0,041
	30, 10	0,046		30, 10	0,043
	30, 20	0,045		30, 20	0,045
	25, 25	0,046		25, 25	0,046
	30, 30	0,048		30, 30	0,038
	40, 40	0,045		40, 40	0,042
	50, 50	0,051		50, 50	0,049
Lpl	10, 10	0,054	B(8,1)	10, 10	0,053
	20, 10	0,056		20, 10	0,045
	20, 20	0,055		20, 20	0,048
	30, 10	0,057		30, 10	0,042
	30, 20	0,058		30, 20	0,047
	25, 25	0,057		25, 25	0,041
	30, 30	0,053		30, 30	0,040
	40, 40	0,047		40, 40	0,042
	50, 50	0,048		50, 50	0,038
B(3,3)	10, 10	0,032	CN(0,9, 3)	10, 10	0,024
	20, 10	0,037		20, 10	0,022
	20, 20	0,042		20, 20	0,018
	30, 10	0,039		30, 10	0,019
	30, 20	0,038		30, 20	0,020
	25, 25	0,039		25, 25	0,019
	30, 30	0,041		30, 30	0,015
	40, 40	0,044		40, 40	0,020
	50, 50	0,046		50, 50	0,017

Loi de distribution	n_1, n_2	Seuil simulé	Loi de distribution	n_1, n_2	Seuil simulé
U(0,1)	10, 10	0,030	CN(0,8, 3)	10, 10	0,022
	20, 10	0,032		20, 10	0,019
	20, 20	0,031		20, 20	0,020
	30, 10	0,034		30, 10	0,017
	30, 20	0,034		30, 20	0,020
	25, 25	0,034		25, 25	0,021
	30, 30	0,037		30, 30	0,017
	40, 40	0,043		40, 40	0,023
	50, 50	0,043		50, 50	0,020

Comme l'indique le Tableau 1, lorsque les effectifs d'échantillons sont faibles, les seuils de signification simulés du test de Bonett sont inférieurs au seuil de signification cible (0,05) pour les lois symétriques ou presque symétriques à queues légères ou modérées. En revanche, les seuils simulés ont tendance à être supérieurs au seuil cible lorsque l'on analyse des effectifs d'échantillons faibles, créés à partir de lois très asymétriques.

Lorsque les effectifs d'échantillons sont modérés ou élevés, les seuils de signification simulés sont proches du seuil cible avec toutes les lois. En réalité, le test obtient des résultats raisonnablement bons même pour les lois très asymétriques, telles que la loi exponentielle et la loi bêta(8, 1).

En outre, les valeurs aberrantes semblent avoir une influence plus grande sur les faibles échantillons que sur les échantillons élevés. Lorsque l'effectif d'échantillon minimal est d'au moins 20, les seuils de signification simulés des populations normales contaminées se stabilisent à 0,020 approximativement.

Lorsque l'effectif minimal des deux échantillons est de 20, les seuils de signification simulés sont invariablement compris dans l'intervalle [0,038, 0,058], sauf pour la loi uniforme continue et pour les lois normales contaminées. Bien qu'un seuil de signification simulé de 0,040 soit légèrement prudent pour un seuil cible de 0,05, ce taux d'erreur de 1ère espèce peut être acceptable dans un contexte plus pratique. Par conséquent, nous pouvons conclure que le test de Bonett est valide lorsque l'effectif minimal des deux échantillons est d'au moins 20.

Simulation B2 : validité du test de comparaisons multiples (modèles à échantillons multiples)

Partie I : plans équilibrés

Nous avons effectué une simulation pour examiner les performances du test de comparaisons multiples dans les modèles à échantillons multiples avec des plans équilibrés. Nous avons créé k échantillons d'effectif égal et obéissant à la même loi, pour l'ensemble de lois utilisées dans la Simulation B1. Les nombres d'échantillons dans un même plan sont $k = 3$, $k = 4$ et $k = 6$. Nous avons défini l'effectif des k échantillons de chaque expérience sur 10, 15, 20, 25, 50 et 100.

Nous avons effectué un test de comparaisons multiples bilatéral avec un seuil de signification cible de $\alpha = 0{,}05$ sur les mêmes échantillons pour chaque plan. Etant donné que, dans chaque cas, les seuils de signification simulés s'appuyaient sur 10 000 paires de répliques d'échantillons et que nous avons utilisé un seuil de signification cible de 5 %, le taux d'erreur de la simulation est de : $\sqrt{0{,}95(0{,}05)/10\,000} = 0{,}2\%$.

Les résultats de la simulation sont récapitulés dans les Tableaux 2a et 2b ci-dessous.

Tableau 2a Seuils de signification simulés pour un test de comparaisons multiples bilatéral dans des plans équilibrés à échantillons multiples. Le seuil de signification cible du test est de 0,05.

Loi de distribution	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Seuil simulé	n_i	Seuil simulé	n_i	Seuil simulé
N(0,1)	10	0,038	10	0,038	10	0,036
	15	0,040	15	0,041	15	0,039
	20	0,039	20	0,040	20	0,041
	25	0,045	25	0,047	25	0,047
	50	0,046	50	0,046	50	0,052
	100	0,049	100	0,049	100	0,052
T(5)	10	0,042	10	0,044	10	0,042
	15	0,041	15	0,044	15	0,046
	20	0,043	20	0,045	20	0,045
	25	0,046	25	0,048	25	0,046
	50	0,040	50	0,039	50	0,038
	100	0,038	100	0,040	100	0,040

Loi de distribution	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Seuil simulé	n_i	Seuil simulé	n_i	Seuil simulé
T(10)	10	0,033	10	0,037	10	0,038
	15	0,040	15	0,042	15	0,041
	20	0,042	20	0,043	20	0,043
	25	0,041	25	0,042	25	0,045
	50	0,047	50	0,044	50	0,047
	100	0,048	100	0,046	100	0,047
Lpl	10	0,056	10	0,063	10	0,071
	15	0,056	15	0,061	15	0,063
	20	0,054	20	0,058	20	0,059
	25	0,051	25	0,056	25	0,58
	50	0,045	50	0,051	50	0,049
	100	0,044	100	0,047	100	0,050
B(3,3)	10	0,031	10	0,031	10	0,031
	15	0,037	15	0,036	15	0,034
	20	0,035	20	0,036	20	0,037
	25	0,039	25	0,038	25	0,040
	50	0,044	50	0,044	50	0,044
	100	0,044	100	0,046	100	0,043
U(0,1)	10	0,029	10	0,025	10	0,023
	15	0,026	15	0,027	15	0,026
	20	0,028	20	0,030	20	0,028
	25	0,034	25	0,033	25	0,032
	50	0,041	50	0,036	50	0,036
	100	0,048	100	0,047	100	0,045

Loi de distribution	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Seuil simulé	n_i	Seuil simulé	n_i	Seuil simulé
Exp	10	0,063	10	0,073	10	0,076
	15	0,056	15	0,058	15	0,064
	20	0,051	20	0,053	20	0,057
	25	0,043	25	0,045	25	0,050
	50	0,033	50	0,037	50	0,038
	100	0,033	100	0,035	100	0,035

Tableau 2b Seuils de signification simulés d'un test de comparaisons multiples bilatéral dans des plans équilibrés à échantillons multiples. Le seuil de signification cible du test est de 0,05.

Loi de distribution	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Seuil simulé	n_i	Seuil simulé	n_i	Seuil simulé
Khi(5)	10	0,040	10	0,046	10	0,048
	15	0,043	15	0,046	15	0,049
	20	0,040	20	0,040	20	0,042
	25	0,040	25	0,045	25	0,042
	50	0,037	50	0,038	50	0,040
	100	0,036	100	0,037	100	0,038
Khi(10)	10	0,042	10	0,045	10	0,045
	15	0,038	15	0,044	15	0,047
	20	0,036	20	0,039	20	0,040
	25	0,043	25	0,044	25	0,045
	50	0,041	50	0,040	50	0,042
	100	0,038	100	0,040	100	0,042
B(8,1)	10	0,058	10	0,060	10	0,066
	15	0,057	15	0,061	15	0,064
	20	0,049	20	0,051	20	0,055

Loi de distribution	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \cdots = n_6$	
	n_i	Seuil simulé	n_i	Seuil simulé	n_i	Seuil simulé
	25	0,044	25	0,046	25	0,050
	50	0,037	50	0,037	50	0,039
	100	0,037	100	0,038	100	0,039
CN(0,9, 3)	10	0,020	10	0,018	10	0,016
	15	0,022	15	0,020	15	0,017
	20	0,014	20	0,012	20	0,008
	25	0,011	25	0,011	25	0,008
	50	0,009	50	0,007	50	0,006
	100	0,010	100	0,008	100	0,008
CN(0,8, 3)	10	0,017	10	0,015	10	0,011
	15	0,013	15	0,011	15	0,008
	20	0,012	20	0,012	20	0,009
	25	0,013	25	0,010	25	0,009
	50	0,011	50	0,011	50	0,009
	100	0,014	100	0,012	100	0,010

Comme l'indiquent les Tableaux 2a et 2b, lorsque l'effectif de l'échantillon est faible, le test de comparaisons multiples est généralement prudent pour les lois symétriques et presque symétriques dans les plans équilibrés. En revanche, le test est libéral pour les échantillons obtenus à partir de lois de distribution très asymétriques, telles que la loi exponentielle et la loi bêta(8, 1). Toutefois, plus l'effectif de l'échantillon augmente, plus les seuils de signification simulés se rapprochent du seuil de signification cible (0,05). En outre, le nombre d'échantillons ne semble pas avoir d'effet important sur les performances du test pour les échantillons dont l'effectif est modéré. Toutefois, la présence de valeurs aberrantes dans les données a un effet important sur les résultats. Lorsque des valeurs aberrantes figurent dans les données, le test est invariablement et excessivement prudent.

Partie II : plans non équilibrés

Nous avons effectué une simulation pour examiner les performances du test de comparaisons multiples dans les plans non équilibrés. Nous avons créé 3 échantillons obéissant à la même loi, pour l'ensemble de lois utilisées dans la Simulation B1. Dans la première série d'expériences, l'effectif des deux premiers échantillons était de $n_1 = n_2 = 10$ et celui du troisième de $n_3 = 15, 20, 25, 50, 100$. Dans la deuxième série d'expériences,

l'effectif des deux premiers échantillons était de $n_1 = n_2 = 15$ et celui du troisième ensemble d'échantillons de $n_3 = 20, 25, 30, 50, 100$. Dans la troisième série d'expériences, nous avons défini l'effectif minimal de l'échantillon sur 20, l'effectif des deux premiers échantillons sur $n_1 = n_2 = 20$ et celui du troisième échantillon sur $n_3 = 25, 30, 40, 50, 100$.

Nous avons effectué un test de comparaisons multiples bilatéral avec un seuil de signification cible de $\alpha = 0{,}05$ sur les mêmes échantillons pour chaque plan. Etant donné que, dans chaque cas, les seuils de signification simulés s'appuyaient sur 10 000 paires de répliques d'échantillons et que nous avons utilisé un seuil de signification cible de 5 %, le taux d'erreur de la simulation est de : $\sqrt{0{,}95(0{,}05)/10\,000} = 0{,}2\%$.

Les résultats de la simulation sont récapitulés dans les Tableaux 3a et 3b ci-dessous.

Tableau 3a Seuils de signification simulés pour un test de comparaisons multiples dans des plans non équilibrés à échantillons multiples. Le seuil de signification cible du test est de 0,05.

Loi de distribution	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Seuil simulé	n_3	Seuil simulé	n_3	Seuil simulé
N(0,1)	15	0,032	20	0,040	25	0,045
	20	0,037	25	0,039	30	0,041
	25	0,038	30	0,037	40	0,043
	50	0,041	50	0,044	50	0,041
	100	0,042	100	0,042	100	0,044
T(5)	15	0,040	20	0,042	25	0,043
	20	0,036	25	0,040	30	0,037
	25	0,044	30	0,036	40	0,038
	50	0,033	50	0,036	50	0,035
	100	0,032	100	0,031	100	0,032
T(10)	15	0,039	20	0,042	25	0,042
	20	0,038	25	0,041	30	0,040
	25	0,040	30	0,041	40	0,041
	50	0,037	50	0,043	50	0,042
	100	0,036	100	0,039	100	0,040
Lpl	15	0,059	20	0,060	25	0,054
	20	0,057	25	0,054	30	0,051
	25	0,056	30	0,051	40	0,050

Loi de distribution	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Seuil simulé	n_3	Seuil simulé	n_3	Seuil simulé
	50	0,049	50	0,051	50	0,050
	100	0,048	100	0,047	100	0,046
B(3,3)	15	0,034	20	0,033	25	0,037
	20	0,031	25	0,035	30	0,039
	25	0,031	30	0,034	40	0,039
	50	0,036	50	0,039	50	0,038
	100	0,035	100	0,039	100	0,039
U(0,1)	15	0,027	20	0,030	25	0,032
	20	0,030	25	0,030	30	0,031
	25	0,028	30	0,032	40	0,036
	50	0,039	50	0,034	50	0,037
	100	0,042	100	0,038	100	0,042
Exp	15	0,061	20	0,053	25	0,042
	20	0,060	25	0,052	30	0,047
	25	0,054	30	0,049	40	0,043
	50	0,050	50	0,046	50	0,041
	100	0,044	100	0,040	100	0,040

Tableau 3b Seuils de signification simulés pour un test de comparaisons multiples dans des plans non équilibrés à échantillons multiples. Le seuil de signification cible du test est de 0,05.

Loi de distribution	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Seuil simulé	n_3	Seuil simulé	n_3	Seuil simulé
Khi(5)	15	0,047	20	0,045	25	0,041
	20	0,043	25	0,042	30	0,039
	25	0,043	30	0,039	40	0,040
	50	0,039	50	0,037	50	0,040
	100	0,034	100	0,035	100	0,034

Loi de distribution	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Seuil simulé	n_3	Seuil simulé	n_3	Seuil simulé
Khi(10)	15	0,043	20	0,042	25	0,042
	20	0,039	25	0,038	30	0,041
	25	0,040	30	0,041	40	0,038
	50	0,038	50	0,041	50	0,042
	100	0,035	100	0,034	100	0,035
B(8,1)	15	0,056	20	0,052	25	0,048
	20	0,054	25	0,046	30	0,044
	25	0,050	30	0,047	40	0,046
	50	0,046	50	0,043	50	0,043
	100	0,043	100	0,042	100	0,044
CN(0,9, 3)	15	0,017	20	0,020	25	0,017
	20	0,020	25	0,019	30	0,012
	25	0,017	30	0,016	40	0,013
	50	0,019	50	0,016	50	0,012
	100	0,014	100	0,016	100	0,010
CN(0,8, 3)	15	0,012	20	0,013	25	0,013
	20	0,016	25	0,012	30	0,012
	25	0,014	30	0,010	40	0,010
	50	0,015	50	0,010	50	0,013
	100	0,012	100	0,011	100	0,010

Les seuils de signification simulés présentés dans les Tableaux 3a et 3b sont cohérents avec ceux signalés précédemment pour des plans équilibrés à échantillons multiples. Par conséquent, les plans non équilibrés ne semblent pas avoir d'impact sur la performance du test de comparaisons multiples. En outre, lorsque l'effectif d'échantillon minimal est d'au moins 20, les seuils de signification simulés se rapprochent du seuil cible, sauf pour les données contaminées.

En conclusion, lorsque l'effectif d'échantillon le plus faible est d'au moins 20, le test de comparaisons multiples fournit de bons résultats pour des plans équilibrés et non équilibrés à échantillons multiples (à k échantillons). Toutefois, avec des effectifs d'échantillons faibles,

le test fournit des résultats prudents pour les données symétriques et presque symétriques, et des résultats libéraux pour des données très asymétriques.

Annexe C : fonction puissance théorique

Il n'existe pas de fonction exacte pour le calcul de la puissance théorique du test de comparaisons multiples. Toutefois, pour des plans à 2 échantillons, il est possible d'obtenir une fonction de puissance par approximation à partir de méthodes reposant sur la théorie des grands échantillons. Les plans à échantillons multiples requièrent plus de recherches pour obtenir une approximation similaire.

Pour des plans à 2 échantillons en revanche, il est possible d'obtenir la fonction puissance théorique du test de Bonett à partir de méthodes reposant sur la théorie des grands échantillons. Plus concrètement, la statistique du test T donnée ci-dessous est distribuée asymptotiquement comme une loi du Khi deux à 1 degré de liberté :

$$T = \frac{(\ln \hat{\rho}^2 - \ln \rho^2)^2}{\frac{\gamma - g_1}{n_1 - 1} + \frac{\gamma - g_2}{n_2 - 1}}$$

Dans l'expression de T, $\hat{\rho} = S_1/S_2$, $\rho = \sigma_1/\sigma_2$, $g_i = (n_i - 3)/n_i$ et γ est l'aplatissement inconnu commun aux deux populations.

Par conséquent, la fonction puissance théorique d'un test bilatéral d'égalité des variances de Bonett avec un seuil de signification par approximation de α peut être exprimée comme suit :

$$\pi(n_1, n_2, \rho) = 1 - \Phi\left(z_{\alpha/2} - \frac{\ln \rho^2}{es}\right) + \Phi\left(-z_{\alpha/2} - \frac{\ln \rho^2}{es}\right)$$

où

$$es = \sqrt{\frac{\gamma - g_1}{n_1 - 1} + \frac{\gamma - g_2}{n_2 - 1}}$$

Pour les tests unilatéraux, la fonction de puissance par approximation lors de la réalisation du test pour $\sigma_1 > \sigma_2$ est

$$\pi(n_1, n_2, \rho) = 1 - \Phi\left(z_{\alpha/2} - \frac{\ln \rho^2}{es}\right)$$

et pour $\sigma_1 < \sigma_2$, la fonction puissance par approximation est

$$\pi(n_1, n_2, \rho) = \Phi\left(-z_{\alpha/2} - \frac{\ln \rho^2}{es}\right)$$

Notez que lors de la phase de planification de l'effectif d'échantillon pour l'analyse des données, l'aplatissement commun aux populations, γ, est inconnu. Par conséquent, la personne chargée des recherches doit généralement s'appuyer sur l'opinion d'experts ou sur les résultats d'expériences précédentes afin d'obtenir une valeur de γ pour la planification. Si cette information n'est pas disponible, il est conseillé de mener une petite étude pilote afin de développer les plans pour la véritable étude. Les échantillons de l'étude pilote permettent de calculer une valeur d'aplatissement regroupé γ pour la planification, exprimée de la façon suivante

$$\hat{\gamma}_P = (n_1 + n_2)\frac{\sum_{j=1}^{n_1}(X_{1j} - m_1)^4 + \sum_{j=1}^{n_2}(X_{2j} - m_2)^4}{[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]^2}$$

Dans le menu Assistant, la valeur estimée de γ pour la planification est obtenue a posteriori, à partir des données dont dispose l'utilisateur.

Annexe D : comparaison des puissances théorique et simulée

Simulation D1 : puissance simulée (réelle) du test de Bonett

Nous avons effectué une simulation pour comparer les niveaux de puissance simulés du test de Bonett et ceux obtenus à l'aide de la fonction puissance par approximation dérivée dans l'Annexe C.

Nous avons créé 10 000 paires d'échantillons pour chacune des lois décrites précédemment (voir Simulation B1). En général, les effectifs d'échantillons sélectionnés étaient suffisamment élevés pour que le seuil de signification simulé du test se rapproche du seuil de signification cible, selon les résultats de la simulation B1.

Pour évaluer les niveaux de puissance simulés avec un rapport d'écarts types de $\rho = \sigma_1/\sigma_2 = 1/2$, nous avons multiplié le second échantillon de chaque paire par la constante 2. Ainsi, pour une loi et pour des effectifs d'échantillons n_1 et n_2 donnés, le niveau de puissance simulé est calculé comme la fraction des 10 000 paires de répliques d'échantillons pour lesquels le test bilatéral de Bonett était significatif. Le seuil de signification cible du test a été fixé à $\alpha = 0{,}05$. Pour comparaison, nous avons ensuite calculé les niveaux de puissance théoriques correspondants en utilisant la fonction puissance par approximation calculée dans l'Annexe C.

Les résultats sont présentés dans le Tableau 4 ci-dessous.

Tableau 4 Comparaison des niveaux de puissance simulée et par approximation d'un test bilatéral de Bonett. Le seuil de signification cible est de 0,05.

Loi de distribution	n_1, n_2	Puissance par app.	Puissance simulée	Loi de distribution	n_1, n_2	Puissance par app.	Puissance simulée
N(0, 1)	20, 10	0,627	0,527	Exp	20, 10	0,222	0,227
	20, 20	0,83	0,765		20, 20	0,322	0,368
	20, 30	0,896	0,846		20, 30	0,377	0,434
	20, 40	0,925	0,886		20, 40	0,412	0,475
	30, 15	0,825	0,771		30, 15	0,32	0,307
	30, 30	0,954	0,925		30, 30	0,458	0,50
	30, 45	0,98	0,97		30, 45	0,531	0,579
	30, 60	0,989	0,984		30, 60	0,575	0,622

Loi de distribution	n_1, n_2	Puissance par app.	Puissance simulée	Loi de distribution	n_1, n_2	Puissance par app.	Puissance simulée
T(5)	20, 10	0,222	0,379	Khi(5)	20, 10	0,355	0,347
	20, 20	0,322	0,569		20, 20	0,517	0,53
	20, 30	0,377	0,637		20, 30	0,597	0,616
	20, 40	0,412	0,69		20, 40	0,644	0,661
	30, 15	0,32	0,545		30, 15	0,513	0,51
	30, 30	0,458	0,733		30, 30	0,701	0,711
	30, 45	0,531	0,795		30, 45	0,781	0,793
	30, 60	0,575	0,828		30, 60	0,823	0,833
T(10)	20, 10	0,476	0,45	Khi(10)	20, 10	0,454	0,414
	20, 20	0,673	0,673		20, 20	0,646	0,631
	20, 30	0,756	0,749		20, 30	0,73	0,717
	20, 40	0,80	0,803		20, 40	0,776	0,771
	30, 15	0,668	0,659		30, 15	0,641	0,618
	30, 30	0,85	0,852		30, 30	0,828	0,819
	30, 45	0,91	0,911		30, 45	0,892	0,882
	30, 60	0,936	0,937		30, 60	0,921	0,912
Lpl	20, 10	0,321	0,33	B(8, 1)	20, 10	0,363	0,278
	20, 20	0,469	0,519		20, 20	0,528	0,463
	20, 30	0,545	0,585		20, 30	0,609	0,549
	20, 40	0,59	0,632		20, 40	0,655	0,60
	30, 15	0,466	0,475		30, 15	0,524	0,419
	30, 30	0,647	0,673		30, 30	0,713	0,634
	30, 45	0,729	0,758		30, 45	0,792	0,737
	30, 60	0,773	0,80		30, 60	0,833	0,777
B(3, 3)	20, 10	0,777	0,628	CN(0,9, 3)	20, 10	0,238	0,284
	20, 20	0,939	0,869		20, 20	0,346	0,452
	20, 30	0,973	0,936		20, 30	0,405	0,517

Loi de distribution	n_1, n_2	Puissance par app.	Puissance simulée	Loi de distribution	n_1, n_2	Puissance par app.	Puissance simulée
	20, 40	0,984	0,964		20, 40	0,442	0,561
	30, 15	0,935	0,871		30, 15	0,343	0,374
	30, 30	0,993	0,98		30, 30	0,491	0,598
	30, 45	0,998	0,995		30, 45	0,567	0,70
	30, 60	0,999	0,999		30, 60	0,612	0,719
U(0, 1)	20, 10	0,916	0,74	CN(0,8, 3)	20, 10	0,26	0,223
	20, 20	0,992	0,95		20, 20	0,379	0,396
	20, 30	0,998	0,985		20, 30	0,444	0,467
	20, 40	0,999	0,995		20, 40	0,484	0,52
	30, 15	0,991	0,941		30, 15	0,376	0,354
	30, 30	1,0	0,996		30, 30	0,535	0,549
	30, 45	1,0	1,0		30, 45	0,614	0,65
	30, 60	1,0	1,0		30, 60	0,661	0,706

Les résultats indiquent que, de manière générale, les niveaux de puissance par approximation et simulés sont proches. Ils se rapprochent à mesure que les effectifs d'échantillons augmentent. Généralement, les niveaux obtenus par approximation sont légèrement supérieurs aux niveaux simulés pour les lois symétriques et presque symétriques à queues modérées ou légères. Toutefois, ils sont légèrement inférieurs aux niveaux de puissance simulés pour les lois symétriques à queues lourdes ou pour les lois fortement asymétriques. La différence entre les deux fonctions puissance n'est généralement pas importante, sauf lorsque les échantillons sont créés à partir de la loi T à 5 degrés de liberté.

Dans l'ensemble, lorsque l'effectif minimal de l'échantillon est d'au moins 20, les niveaux de puissance obtenus par approximation et simulés sont particulièrement proches. Par conséquent, la planification des effectifs d'échantillons peut être effectuée à l'aide des fonctions puissance par approximation.