REVUE DE STATISTIQUE APPLIQUÉE

JOHN WALSH

Sur l'emploi de résultats de la théorie des échantillons de lois normales pour des observations non indépendantes, provenant de populations non normales

Revue de statistique appliquée, tome 20, n° 4 (1972), p. 5-12

<http://www.numdam.org/item?id=RSA_1972__20_4_5_0>

SUR L'EMPLOI DE RÉSULTATS DE LA THÉORIE DES ÉCHANTILLONS DE LOIS NORMALES POUR DES OBSERVATIONS NON INDÉPENDANTES, PROVENANT DE POPULATIONS NON NORMALES

John WALSH

Southern Methodist University [1] [2]

RESUME

Certains résultats statistiques bien connus reposent sur l'hypothèse d'un échantillon aléatoire d'une population normale, ou encore pour l'analyse de variance simple (one way ANOVA) l'hypothèse d'échantillons indépendants de populations normales. Un aspect intéressant de ces résultats, c'est qu'ils se réfèrent à des valeurs exactement connues de certaines probabilités. Le présent article montre qu'une grande partie de ces tests, régions de confiance, etc..., possèdent encore les mêmes propriétés (exactement et non asymptotiquement) sous des hypothèses moins restrictives et en interprétant convenablement les recherches. Plus précisément, on va étendre le concept d'échantillon d'une population normale et le concept d'échantillons indépendants de populations normales. Ces modèles généralisés n'exigent plus l'indépendance entre toutes les observations, ils ne demandent plus à chaque observation de provenir d'une même loi et à cette loi d'être normale. On peut obtenir des résultats exacts quand la ou les statistiques employées s'expriment exclusivement en fonction des différences entre observations. Dans le cas de l'étude d'un échantillon unique, de telles statistiques sont celles qu'on emploie pour estimer sans biais la variance ou pour rejeter de l'échantillon les données aberrantes (outliers). Certaines statistiques de même nature se rencontrent dans le cas d'une analyse de variance simple (one way ANOVA), comportant souvent l'emploi du F de Fisher-Snedecor pour tester si des échantillons indépendants proviennent bien d'une même population normale. On donnera des exemples numériques de ces deux cas.

INTRODUCTION ET RESULTATS

On a fait couramment l'hypothèse que les données sont un échantillon aléatoire provenant d'une population normale ou encore, pour l'analyse de variance simple (ANOVA) qu'elles sont des échantillons indépendants de populations normales. De façon plus précise, dans le cas l'échantillon unique, on peut mettre les observations sous la forme :

$$x_i = \mu + e_i, \qquad (i = 1, \ldots, n),$$

---

(1) Recherche en partie subventionnée par contrat AFOSR F 33615-71-C-1178 de l'Air Force et par Mobil Research and Development Corporation. Recherche également reliée au contrat ONR N 000 14-68-A-0515 et à la subvention NGR 44-007-028 de la NASA.

(2) L'auteur de cet article est mort brusquement, le 24 août dernier, alors qu'il participait à Dublin à la Sixième Conférence Internationale de Recherche Opérationnelle. Cette publication est ainsi un hommage à l'éminent statisticien qu'était le Professeur John Walsh. (NDLR)

où les e_i sont un échantillon aléatoire extrait d'une population normale de moyenne zéro et de variance σ^2 (en général inconnue). Dans le cas de l'analyse de variance simple, on peut mettre les observations sous la forme

$$x_{ij} = \mu_j + e_{ij}, \qquad [i = 1, \ldots, n(j) \; ; \; j = 1, \ldots, m],$$

où, pour un indice j donné, les e_{ij} sont un échantillon aléatoire de taille n(j) provenant d'une population normale de moyenne 0 et de variance σ_j^2 (en général inconnue). De plus les e_{ij} sont mutuellement indépendants.

Ces hypothèses sont la base de quelques résultats bien connus, concernant des tests de signification, régions de confiance, estimations et procédures de décisions, telles que celles permettant de rejeter les données aberrantes (outliers). Une propriété fort recherchée que présentent ces résultats, c'est que les seuils de signification, coefficients de confiance, etc... sont déterminés avec précision. De plus, les seuils de signification et coefficients de confiance peuvent prendre toute valeur demandée comprise entre zéro et un.

Le but du présent article est de montrer que beaucoup de ces résultats, établis sur la base des hypothèses précédentes (échantillons normaux) restent vrais et conservent les mêmes propriétés quand on étend l'expression de x_i ou x_{ij} à des cas plus généraux.

Dans le cas de l'échantillon unique, on met les observations sous la forme

$$x_i = \mu + e_i + e', \qquad (i = 1, \ldots, n),$$

où les e_i ont les mêmes propriétés d'échantillon normal que ci-dessus, e' ayant alors une distribution arbitraire. En outre e' peut dépendre des e_i de façon quelconque et le signe de dépendance peut varier avec i.

Si les variances et covariances existent pour $(e', e_1, \ldots, e_n)$, la variance de e' étant supposée non nulle, et si l'on désigne par ρ_i la corrélation entre e' et e_i, la condition suivante :

$$\rho_1^2 + \ldots + \rho_n^1 \leqslant 1 \qquad (1)$$

devra en outre être satisfaite par les ρ_i.

On s'intéresse en outre aux propriétés de la population d'où les e_i sont extraits, alors que les propriétés concernant e' n'offrent aucun intérêt. La valeur e' représente une composante d'erreur imposée à toutes les observations par la situation expérimentale (composante qui serait nulle si les conditions étaient idéales) ; et si les e_i étaient autres, la valeur aléatoire de e' pourrait différer.

Les x_i ont des distributions continues (puisque les e_i ont des distributions continues) et ils ont même espérance mathématique (si l'espérance mathématique de e' existe bien). Puisque la distribution de $e_i + e'$ est affectée par la dépendance existant entre e_i et e', les x_i peuvent avoir des distributions différant sensiblement d'une loi normale.

En outre, toute paire de variables x_i présente une possibilité de dépendance, et cette dépendance peut être très forte. Un cas particulier est celui où les x_i ont une distribution multinormale et où on montre que la corré-

lation est la même pour tout couple des x_i. Dans ce dernier cas, les propriétés de quelques résultats bien connus sont examinées dans la référence 1. Comme on le verra (dans la dernière section du présent article), le modèle employé ici pour les x_i est équivalent à celui donné dans la référence 1, lorsque $(e', e_1, \ldots, e_n)$ ont une loi de distribution multinormale et que les ρ_i sont égaux. L'expression généralisée employée pour les x_i, de même que celle donnée plus loin pour les x_{ij}, sont des cas spéciaux de l'expression donnée dans la référence 2 pour l'analyse de variance d'un tableau à double entrée.

Lorsqu'une statistique peut s'exprimer en fonction des seules différences des x_i, la composante e' de x_i s'élimine ; et la statistique a les mêmes propriétés que lorsque les hypothèses d'échantillon normal sont valables. Par exemple, pour $n \geqslant 2$.

$$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{ , avec } \bar{x} = \sum_{i=1}^{n} x_i/n$$

est identique à

$$\sum_{i=1}^{n} (e_i - \bar{e})^2, \text{ avec } \bar{e} = \sum_{i=1}^{n} e_i/n.$$

Par suite s^2/σ^2 suit une loi de χ^2 à $(n - 1)$ degrés de liberté, et $s^2/(n - 1)$ est une estimation sans biais de σ^2.

De même les procédures permettant de rejeter les observations aberrantes (outliers) sont presque toujours basées sur les différences entre valeurs échantillons (par exemple, voir référence 2 et les références qui y sont données).

A présent, considérons l'expression suivante des observations, dans le cas d'une analyse de variance simple avec 2 échantillons et plus :

$$x_{ij} = \mu_j + e_{ij} + e'' \quad [i = 1, \ldots, n(j) ; j = 1, \ldots, m]$$

où les e_{ij} ont les mêmes propriétés que dans le cas des échantillons normaux, mais où e'' a une distribution arbitraire. En outre e'' peut dépendre des e_{ij} de façon quelconque, le degré de dépendance pouvant varier avec le couple (ij). Quand les variances et covariances de

$$(e'', e_{ij} \quad [i = 1, \ldots, n(j) ; j = 1, \ldots m])$$

existent, soit ρ_{ij} le coefficient de corrélation entre e'' et e ; alors la condition :

$$\sum_{j=1}^{m} \sum_{i=1}^{n(j)} \rho_{ij}^2 \leqslant 1 \qquad (2)$$

est nécessairement satisfaite par les ρ_{ij}.

Seules importent les propriétés des populations dont sont extraits les e_{ij}. La valeur e'' représente une composante de l'erreur qui, dans des conditions idéales, serait nulle. L'hypothèse zéro H° considérée ici est la même que dans le cas des échantillons normaux. C'est-à-dire que selon H°, les $e_{ij} + \mu_j$ proviennent de la même population normale. En conséquence de quoi, envisager le cas où toutes les observations sont affectées de l'erreur sup-

plémentaire e'' paraît bien être une généralisation convenable de l'expression supposée des observations dans l'analyse de variance simple courante.

Les distributions des x_{ij} sont continues mais peuvent ne pas être normales et même différer notablement d'une distribution normale. Il existe une dépendance entre tout couple des x_{ij}, et cette dépendance peut même être très forte.

La composante e'' s'élimine de toute statistique dépendant seulement des différences entre les x_{ij}. En particulier supposons tous les n(j) $\geqslant 2$ et considérons la statistique F de Fisher-Snedecor souvent utilisée pour tester si des échantillons indépendants proviennent de la même population normale. Sous l'hypothèse H°, les $\mu_j + e_{ij}$ sont bien issus de la même population normale, de sorte que :

$$S_E^2 = \sum_{j=1}^{m} \sum_{i=1}^{n(j)} (x_{ij} - \bar{x})^2 = \sum_{j=1}^{m} \sum_{i=1}^{n(j)} (e_{ij} - \bar{e})^2$$

$$S_B^2 = \sum_{j=1}^{m} n(j) \, (\bar{x}_{.j} - \bar{x})^2 = \sum_{j=1}^{m} n(j) \, (\bar{e}_{.j} - \bar{e})^2$$

où

$$\bar{x} = \sum_{j=1}^{m} \sum_{i=1}^{n(j)} x_{ij}/N \ , \ \bar{e} = \sum_{j=1}^{m} \sum_{i=1}^{n(j)} e_{ij}/N$$

$$\bar{x}_{ij} = \sum_{i=1}^{n(j)} x_{ij}/n(j) \ , \ \bar{e}_{ij} = \sum_{i=1}^{n(j)} e_{ij}/n(j)$$

et

$$N = \sum_{j=1}^{m} n(j)$$

Ainsi la statistique $(N - m) \, S_B^2/(m - 1)S_E^2$ est distribuée comme le F de Fisher-Snedecor avec respectivement n - 1 et N - m degrés de liberté quand l'hypothèse H° est vérifiée. Plus généralement la distribution de cette statistique est la même que dans le cas des échantillons normaux de l'analyse de variance simple.

La section qui suit sera consacrée à la vérification des relations (1) et (2) et à la justification de nos dires, comme quoi le modèle utilisé ici dans le cas de l'échantillon unique est bien équivalent au modèle employé dans la référence 1 quand (e', e_1, ..., e_n) a une distribution multinormale avec des ρ_i égaux entre eux.

Dans la section finale on trouvera un exemple numérique d'extension du modèle au cas d'un échantillon unique, et un exemple numérique d'extension de l'analyse de variance simple.

VERIFICATIONS

1 - En premier lieu, proposons-nous de vérifier la relation (1) dans le cas de l'échantillon unique, quand toutes les variances et covariances de (e', e_1, ..., e_n) existent et aucune variance n'est nulle. Le déterminant de la matrice des variances et covariances peut s'écrire:

$$\begin{vmatrix} \sigma_o^2 & \rho_1\sigma_o\sigma & \rho_2\sigma_o\sigma & \cdots\cdots\cdots & \rho_?\sigma_o\sigma \\ \rho_1\sigma_o\sigma & \sigma^2 & 0 & \cdots\cdots\cdots & 0 \\ \rho_2\sigma_o\sigma & 0 & \sigma^2 & \cdots\cdots\cdots & 0 \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \rho_n\sigma_o\sigma & 0 \ldots\ldots & 0 \ldots\ldots\ldots\ldots & & \sigma^2 \end{vmatrix}$$

où σ_o^2 désigne la variance de e'. En développant ce déterminant on le trouve égal à :

$$\sigma_o^2\ \sigma^{2n}\ (1 - \rho_1^2 - \rho_2^2 \ldots\ldots - \rho_n^2)$$

et il est nécessairement non négatif. D'où la relation (1).

2 - En second lieu, vérifions la relation (2) dans le cas de l'analyse de variance simple, quand toutes les variances et covariances de

$$[e'',\ e_{ij}\ ;\ i = 1,\ \ldots\ n(j)\ ;\ j = 1,\ \ldots\ m]$$

existent et aucune variance n'est nulle. Le déterminant de la matrice des variances et covariances de ce vecteur aléatoire multinormal à N + 1 composantes, est :

$$\begin{vmatrix} \sigma_{oo}^2 & \rho_{11}\sigma_{oo}\sigma_1 & \cdots\cdots\cdots & \rho_{n(m)m}\sigma_{oo}\sigma_m \\ \rho_{11}\sigma_{oo}\sigma_1 & \sigma_1^2 & \cdots\cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \rho_{n(m)m}\sigma_{oo}\sigma_m & 0 \ldots\ldots & \ldots\ldots\ldots\ldots & \sigma_m^2 \end{vmatrix}$$

où σ_{oo}^2 désigne la variance de e''. La valeur de ce déterminant est :

$$\sigma_{oo}^2\ \sigma^{2n(1)} \ldots\ldots\ \sigma_m^{2n(m)}\ [1 - \sum_{j=1}^{m} \sum_{i=1}^{n(j)} \rho_{ij}^2\ ]$$

et elle est nécessairement non négative. D'où la relation (2).

3 - Finalement supposons que $(e', e_1, \ldots, e_n)$ ait une distribution (n + 1) normale avec σ_o^2 et σ^2 positives et

$$\rho_1 = \ldots\ \rho_n = \rho.$$

Alors la corrélation entre toute paire de x est la même et peut s'écrire :

$$(\sigma_o^2 + \rho\sigma_o\sigma)\ \ /\ (\rho_o^2 + \sigma^2 + 2\ \rho\sigma_o\sigma) \qquad (3)$$

Le modèle donné ici pour le cas de l'échantillon unique est équivalent au modèle donné dans la référence 1, si l'on peut montrer que la valeur de (3) peut prendre (comme une corrélation intraclasse) toute valeur comprise entre $-1/(n-1)$ et 1. Puisque (3) est fonction continue de σ_o/σ et de ρ, il suffit de montrer qu'on peut bien atteindre les valeurs $-1/(n-1)$ et 1 (au moins à la limite). D'une part l'expression (3) tend bien vers 1 quand σ_o/σ tend vers l'infini. D'autre part, il résulte de la relation (1) qu'on a :

$$-1/\sqrt{n} \leqslant \rho \leqslant +1/\sqrt{n}$$

Faisons $\rho = -1/\sqrt{n}$ et $\sigma_o/\sigma = \sqrt{n}$, la valeur de (3) est alors $-1/(n-1)$. On en conclut que les deux modèles sont bien équivalents.

APPLICATIONS NUMERIQUES

Pour simplifier, on considère seulement dans les deux exemples ci-après des variables aléatoires multinormales. Autrement dit, e', e_1, ..., e_n auront une distribution multinormale, de façon à étendre le cas de l'échantillon unique considéré. De même, e'' et les

$$e_{ij} \quad [i = 1, \ldots, n(j) \; ; \; j = 1, \ldots, m]$$

auront une distribution multinormale pour étendre le cas de l'analyse de variance simple.

Il importe de définir clairement quel modèle est employé, ce qu'on va étudier, enfin quelles statistiques sont employées. Dans le premier cas (échantillon unique) le modèle sera :

$$x_i = \mu + e_i + e' \qquad (i = 1, \ldots, n),$$

sous l'hypothèse particulière que les x_i ont une distribution multinormale. La propriété étudiée est l'existence d'une variance commune σ^2 pour les e_i ; la statistique employée est s^2. Ce modèle peut servir à étudier l'existence de σ^2 si les corrélations entre e' et les e_i satisfont à l'inégalité (1).

Par exemple soit $n = 5$ et une matrice des variances et covariances des $x_1 \ldots x_n$ égale (à un facteur positif constant près) à :

$$\begin{vmatrix} 6,6 & 2,4 & 0,6 & 1,6 & 2,8 \\ 2,4 & 6,2 & 0,4 & 1,4 & 2,6 \\ 0,6 & 0,4 & 2,6 & -0,4 & 0,8 \\ 1,6 & 1,4 & -0,4 & 4,6 & 1,8 \\ 2,8 & 2,6 & 0,8 & 1,8 & 7,0 \end{vmatrix}$$

L'étude de cette matrice montre qu'elle correspond au modèle où l'on saurait :

$$\sigma/\sigma_o = 2, \; \rho_1 = 0,4, \; \rho_2 = 0,3, \; \rho_3 = -0,6, \; \rho_4 = -0,1, \; \rho_5 = 0,5.$$

Le modèle est utilisable car ces valeurs des ρ_i satisfont à (1). Aussi, du fait de la distribution multinormale attribuée à e', e_1, ... e_5, le modèle représente n'importe quels x_1, ..., x_5 ayant la matrice ci-dessus des variances et covariances et une distribution multinormale.

Dans le second cas (analyse de variance simple) le modèle est :

$$x_{ij} = \mu_j + e_{ij} + e'' \quad [i = 1, 2, \ldots n(j) ; j = 1, \ldots m]$$

avec cette hypothèse supplémentaire que les x_{ij} ont une distribution multinormale. La propriété étudiée est celle de savoir si les $\mu_j + e_{ij}$ proviennent bien de la même distribution, et on emploie à cet effet la statistique F

$$F = (N - m)\, S_B^2 / (m - 1)\, S_E^2$$

Ce modèle s'applique si les corrélations entre e'' et les e_{ij} satisfont à la condition (2).

Comme exemple d'analyse de variance simple, posons m = 3 et n(j) = 2, j = 1, 2. La matrice des variances et covariances entre x_{11}, x_{12}, x_{13}, x_{21}, x_{22} et x_{23} est (à un facteur positif constant près) :

$$\begin{vmatrix} 3,4 & 0,8 & 0,6 & 0,7 & 0,6 & -0,4 \\ 0,8 & 6,2 & 2,0 & 2,1 & 1,4 & 1,0 \\ 0,6 & 2,0 & 5,8 & 1,7 & 1,0 & 0,6 \\ 0,7 & 2,1 & 1,7 & 3,0 & 1,3 & 0,9 \\ 0 & 1,4 & 1,0 & 1,3 & 1,6 & 0,2 \\ -0,4 & 1,0 & 0,6 & 0,9 & 0,2 & 0,8 \end{vmatrix}$$

Cette matrice correspond au modèle avec :

$$\sigma_1/\sigma_{oo} = 2, \quad \sigma_1/\sigma_2 = 2 \; ; \; \rho_{11} = -0,4 \; ; \; \rho_{12} = 0,3 \; ; \; \rho_{13} = 0,2$$
$$\rho_{21} = 0,5 \; ; \; \rho_{22} = -0,2 \; ; \; \rho_{23} = -0,6$$

Ce modèle est utilisable car ces valeurs des ρ_{ij} satisfont à la condition (2).

Aussi, en raison de la distribution supposée multinormale de e'' et des e_{ij}, le modèle représente n'importe quelles variables

$$x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}$$

ayant pour distribution une loi multinormale dont la matrice des variances et covariances est celle qui précède.

REFERENCES

[1] John E. WALSH. - Concerning the effect of intraclass correlation on significance tests - Annals of Mathematical Statistics, Vol. 18 (1947) p. 88/96.

[2] Irwin GUTTMANN & Dennis E. SMITH. - Investigation of rules for dealing with outliers in small samples from the normal distribution I : Estimation of the mean. Technometrics Vol. 11 (1969) p. 527/550

[3] John E. WALSH. - Handbook of Nonparametric Statistics, III : Analysis of Variance D. Van Nostrand Co. Inc. Princeton, N.J., 1968, 771 p.