

REVUE DE STATISTIQUE APPLIQUÉE

JOHN E. WALSH

La distribution d'une statistique d'ordre établie sur un ensemble d'observations qui ne forment pas un échantillon

Revue de statistique appliquée, tome 18, n° 3 (1970), p. 65-69

http://www.numdam.org/item?id=RSA_1970__18_3_65_0

© Société française de statistique, 1970, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

LA DISTRIBUTION D'UNE STATISTIQUE D'ORDRE ÉTABLIE SUR UN ENSEMBLE D'OBSERVATIONS QUI NE FORMENT PAS UN ÉCHANTILLON

John E. WALSH

Southern Methodist University (1)

Considérons n observations unidimensionnelles ayant une distribution arbitraire; on montre qu'en général, la distribution de toute statistique d'ordre établie à partir de ces observations est la même que celle de cette même statistique d'ordre établie pour un échantillon au hasard d'effectif n (provenant d'une distribution dépendant de celle des n observations initiales). Ainsi, les statistiques d'ordre peuvent être regardées individuellement comme provenant d'échantillons. Cependant, la distribution "échantillonnée" peut dépendre beaucoup de la statistique d'ordre considérée. Pour des valeurs de n grandes ou petites, ces résultats sont utiles pour définir des coefficients de régression pour un type de régression ayant un large domaine d'application (régression non linéaire linéarisée). Ils sont aussi utiles pour la détermination des propriétés asymptotiques des distributions des extrêmes et des percentiles.

Pour n donné, assez grand, une distribution asymptotique peut être utilisée dans des situations plus générales, si la distribution "échantillonnée" appartient à une classe de distributions assez large. Ainsi, dans le cas continu, les percentiles observés ont asymptotiquement des distributions normales sous des conditions assez générales. De même, les distributions asymptotiques établies pour les valeurs extrêmes d'échantillons pourraient souvent être utilisées pour des situations continues. On discute de quelques applications de ces résultats asymptotiques pour la prévision.

INTRODUCTION ET DISCUSSION

L'hypothèse d'un échantillon au hasard est très commode lorsqu'on établit les propriétés de la distribution d'une statistique d'ordre d'un ensemble d'observations unidimensionnelles. Aussi, dans la pratique, cette hypothèse est souvent retenue alors qu'elle est largement violée (par exemple en ceci qu'il existe une forte liaison entre certaines observations). En dépit de ces

(1) Recherche subventionnée en partie par Mobil Research and Development Corp.,
NASA Grant NGR 44 077-028 et un contrat ONR N 00014-68-A-0515

réserve, l'évidence empirique montre qu'au moins quelques unes des propriétés sont les mêmes que s'il s'agissait d'échantillons. Par exemple, une forme de distribution asymptotique pour la plus grande valeur d'un échantillon semble satisfaisante dans certaines situations pratiques où la "condition d'échantillon" semble être fortement violée (Gumbel, 1958).

Quand les observations sont indépendantes, on peut donner une explication à la possibilité d'appliquer au moins quelques uns des résultats établis pour les échantillons. Dans des circonstances assez générales, une statistique d'ordre se comporte à peu près comme si elle provenait d'un échantillon au hasard (de même effectif) à partir d'une distribution égale à la moyenne arithmétique des distributions des n observations (Walsh, 1959, 1964). Des intervalles de confiance approchés, unilatéraux et bilatéraux, et des tests du genre test de signe peuvent être obtenus pour les percentiles de cette distribution moyenne. De même, pour une large classe de situations, des résultats asymptotiques basés sur un échantillon de la distribution moyenne peuvent être appliqués pour les valeurs extrêmes.

L'établissement et l'emploi de propriétés des distributions sont beaucoup plus compliqués lorsqu'intervient une dépendance non triviale. Cependant, nous montrons dans cet article que toute statistique d'ordre peut être traitée comme si elle provenait d'un échantillon au hasard d'effectif n . Malheureusement, à cause des diverses dépendances, la distribution "échantillonnée" peut être très différente pour les diverses statistiques d'ordre, et il ne semble pas exister de relation utile en général entre ces distributions. En fait, sauf pour la plus grande et la plus petite observation, la distribution "échantillonnée" n'est pas bien identifiée en termes de distribution individuelle (conditionnelle ou non conditionnelle) pour les observations individuelles.

Ces propriétés montrent qu'une vérification empirique du caractère d'échantillon sur des statistiques d'ordre n'est pas, même de façon approchée, une vérification de l'hypothèse d'échantillon. Des distributions de cette nature peuvent se présenter (exactement) pour toute sorte de situations où l'on n'a pas affaire à un échantillon.

Les propriétés semblables à celles d'un échantillon ont des applications directes pour des valeurs quelconques de n et pour n grand (cas asymptotique). Pour n quelconque, la considération de la distribution "échantillonnée" peut conduire aux bonnes définitions des paramètres statistiques. Un emploi de cette méthode concerne l'extension de la méthode de régression non linéaire linéarisée (introduite dans Walsh, 1963). Cette extension utilise la distribution "échantillonnée" pour obtenir la médiane de certaines observations comme base de définition d'une fonction linéaire particulière des coefficients de régression. Pour des données continues et pour ce type de définition, on obtient des estimations de la médiane exacte pour des fonctions linéaires spécifiées des coefficients de régression (Walsh et Kelleher, 1969).

Les propriétés semblables à celles d'un échantillon ont quelques implications de nature non paramétrique dans les cas asymptotiques. Il existe de très larges classes de distributions telles que les statistiques d'ordre établies sur des échantillons suffisamment larges de toute distribution d'une classe ont (approximativement) une forme de distribution qui est complètement déterminée sauf pour les valeurs d'un petit nombre de paramètres. En particulier, cette situation se présente dans le cas continu et pour les va-

leurs extrêmes ou les percentiles expérimentaux. Considérons maintenant le cas général de n observations et d'une statistique d'ordre donnée. Si n est suffisamment grand et si (pour cette valeur de n) la distribution "échantillonnée" pour la statistique d'ordre appartient à la classe très large, alors la forme paramétrique de la distribution asymptotique peut être utilisée. Ceci simplifie le problème statistique en le réduisant à la recherche d'un petit nombre de paramètres.

D'une façon plus précise, considérons tout d'abord le percentile observé dans le cas où les n observations sont une observation d'une population n -dimensionnelle qui est continue. Pour n grand, la distribution d'un percentile observé est à peu près normale si la distribution "échantillonnée" appartient à une classe extrêmement large de distributions. Ainsi, pour n très grand, la distribution d'un percentile observé doit-elle être souvent à peu près normale. Des considérations semblables s'appliquent aux distributions asymptotiques des valeurs extrêmes. Ici, cependant, la classe des distributions conduisant aux distributions asymptotiques paramétriques ("asymptotes", voir Gumbel, 1958) est seulement modérément large.

Considérons maintenant quelques applications pratiques. Une utilisation intéressante des propriétés semblables à celles d'un échantillon concerne la prédiction dans un type particulier de situation. Ici, un nouvel ensemble d'observations unidimensionnelles d'effectif n doit être tiré indépendamment d'une loi qui est à peu près la même que plusieurs lois qui ont déjà fourni des ensembles indépendants d'effectif n . C'est-à-dire qu'une nouvelle observation multidimensionnelle avec n composantes doit être obtenue indépendamment à partir d'une distribution multidimensionnelle qui est à peu près la même que plusieurs distributions multidimensionnelles indépendantes avec n composantes. Alors, pour une statistique d'ordre donnée des valeurs des composantes, la distribution "échantillonnée" doit être à peu près la même pour toutes ces observations. Le problème de prédiction consiste à estimer les propriétés probabilistes d'une statistique d'ordre donnée de la nouvelle observation à partir des valeurs de cette statistique d'ordre pour les observations déjà obtenues.

Une application concerne la régression non linéaire linéarisée (voir Walsh et Kelleher, 1969). Quelques autres applications se présentent pour les cas asymptotiques. En effet, les distributions asymptotiques, qui sont complètement spécifiées, sauf en ce qui concerne les valeurs d'un petit nombre de paramètres, peuvent souvent être retenues. Les statistiques d'ordre correspondantes à partir de plusieurs observations multidimensionnelles passées fournissent le matériel de base pour estimer ces paramètres. Des propriétés probabilistes pour la nouvelle statistique d'ordre peuvent être estimées en utilisant la distribution asymptotique avec ces estimations des paramètres. Par exemple, cette méthode peut être utile conjointement avec les distributions asymptotiques pour les valeurs extrêmes quand on a affaire à des variables continues.

Il existe une méthode plus précise quand on considère un percentile des coordonnées d'une observation multidimensionnelle dans le cas continu. La distribution asymptotique d'un tel percentile doit souvent être approximativement normale. Ainsi, les percentiles observés pour les observations passées forment approximativement un échantillon au hasard d'une population normale. Le nouveau percentile est une valeur indépendante issue approxi-

mativement de la même population. Une statistique $t^{(1)}$ peut être utilisée pour estimer la probabilité que le nouveau percentile dépasse une valeur donnée d'avance. On peut tester également si le nouveau percentile provient approximativement de la même population que les percentiles déjà obtenus. En fait, tout procédé pour étudier une nouvelle observation sur la base d'un échantillon indépendant de la même population normale peut être employé comme approximation.

La section qui suit contient la démonstration que les statistiques d'ordre pour une situation quelconque où l'on n'est pas en présence d'un échantillon ont des distributions semblables à celles d'un échantillon. La distribution "échantillonnée" pour une statistique d'ordre donnée est déterminée par la valeur de n et par la distribution conjointe des observations. Elle peut s'exprimer à l'aide de distributions de probabilité conditionnelle et non-conditionnelle. Cependant, les expressions sont fort compliquées sauf pour les cas de la plus grande observation et de la plus petite observation.

VERIFICATION

Soit X_t la $t^{\text{ième}}$ statistique d'ordre dans un ensemble de n observations qui peuvent avoir n'importe quelle distribution conjointe ($t = 1$ correspond à la plus petite observation, etc...). La fonction de répartition de X_t pour cette situation est déterminée par la distribution conjointe et sera notée $F_t(x; n)$. Soit $G_t(x; n)$ définie par $0 \leq G_t(x; n) \leq 1$ et

$$\sum_{i=t}^n \binom{n}{i} G_t(x; n)^i [1 - G_t(x; n)]^{n-i} = F_t(x; n) \quad (1)$$

La fonction $G_t(x; n)$ a les propriétés d'une fonction de répartition (c'est-à-dire que $G_t(x; n)$ est monotone croissante, $G_t(-\infty, n) = 0$, $G_t(\infty, n) = 1$, etc...). Ceci résulte immédiatement du fait que le membre de gauche de (1) est une fonction continue et strictement croissante de $G_t(x; n)$ pour $0 \leq G_t(x; n) \leq 1$; d'autre part sa valeur est zéro pour $G_t(x; n) = 0$ et l'unité pour $G_t(x; n) = 1$.

Considérons la $t^{\text{ième}}$ statistique d'ordre dans un échantillon au hasard d'effectif n provenant de la fonction de répartition $G_t(x; n)$. Sa distribution est donnée par le membre de gauche de l'équation (1). Ainsi X_t a la même distribution que la $t^{\text{ième}}$ statistique d'ordre dans un échantillon au hasard d'effectif n obtenu à partir de la fonction de répartition $G_t(x; n)$.

Pour $t = 1$ et n , la fonction $G_t(x; n)$ s'exprime aisément à l'aide de distributions non conditionnelle et conditionnelle des observations individuelles. On peut également l'exprimer explicitement en fonction de $F_t(x; n)$. On trouve :

$$G_n(x; n) = [F_n(x; n)]^{1/n}$$

$$G_1(x; n) = 1 - [F_1(x; n)]^{1/n}$$

Soit $H_j(x; s_{j-1})$ la probabilité conditionnelle que la $j^{\text{ième}}$ observation unidimensionnelle ait une valeur au plus égale à x étant donné que les première, seconde, ..., $(j-1)^{\text{ième}}$ observations ont des valeurs $\leq x$. Ici, s_0

(1) statistique t de Student.

n'indique aucune condition. Alors, $F_n(x ; n) = \prod_{j=1}^n H_j(x ; s_{j-1})$, ce qui exprime $G_n(x ; n)$ à l'aide d'une fonction de répartition non conditionnelle et de $n - 1$ conditionnelles. Cependant, l'indice j pourrait être affecté aux observations de n'importe quelle manière. Ainsi, il y a $n!$ façons d'exprimer $F_n(x ; n)$ à l'aide d'une fonction de répartition non conditionnelle et de $n - 1$ conditionnelles, plus un nombre illimité de combinaisons de ces $n!$ expressions.

Finalement soit $I_j(x ; s'_{j-1})$ la probabilité conditionnelle que la $j^{\text{ième}}$ observation ait une valeur $\leq x$ sachant que les première, seconde, ..., $(j - 1)^{\text{ième}}$ observations ont des valeurs $> x$. Ici, s'_0 n'indique aucune condition. Alors,

$$F_1(x ; n) = 1 - \prod_{j=1}^n [1 - I_j(x ; s'_{j-1})]$$

et $F_1(x ; n)$ s'exprime à l'aide d'une fonction de répartition non conditionnelle et de $n - 1$ conditionnelles.

REFERENCES

- GUMBEL E.J. - Statistics of extremes, Columbia Univ. Press, 1958
- WALSH J.E. - "Definition and use of generalized percentage points", Sankhyā, Vol. 21 (1959) p. 281/88
- WALSH J.E. - "Use of linearized nonlinear regression for simulations involving Monte-Carlo", Operations Research, Vol. 11 (1963, p. 228/235
- WALSH J.E. - "Approximate distribution of extremes for nonsample cases" Journal Amer. Stat. Assoc. Vol 59 (1964) p. 429/436
- WALSH J.E. and KELLEHER G.J. - Extended uses of linearized nonlinear regression for random-nature simulations", Proceedings of the 5th International Conference on Operational Research, 1969 (in press).