Statistique et analyse des données

Christian Lavergne

Estimation et tests sur la variance dans un modèle à un facteur aléatoire pour des données qualitatives

Statistique et analyse des données, tome 13, n° 3 (1988), p. 33-43

<http://www.numdam.org/item?id=SAD_1988__13_3_33_0>

Statistique et Analyse des Données
1988 - Vol. 13 n° 3 p.33-43

ESTIMATION ET TESTS SUR LA VARIANCE DANS UN MODELE A UN FACTEUR ALEATOIRE POUR DES DONNEES QUALITATIVES

Christian Lavergne

Laboratoire de Statistique et Probabilités
Université Paul Sabatier, U.A.-C.N.R.S. 745
118, route de Narbonne, 31062 Toulouse cedex

Résumé: *On construit un modèle à un facteur aléatoire adapté à l'analyse de données qualitatives. On considère un estimateur sans biais de la variance du facteur aléatoire introduit; on présente un résultat de convergence le concernant. Un test d'hypothèse est alors introduit et illustré par des résultats numériques de simulation.*

Abstract: *A model with random factors adapted to qualitative data analysis is constructed. A variance unbaised estimator of the introduced random factor is constructed; a convergency result is presented. A hypothesis test is then introduced and illustred by simulated numerical results.*

Mots clés: *Modèle à effets aléatoires, données qualitatives, simulation.*

Indices de classification STMA: *07-090, 08-020.*

I- INTRODUCTION

Il est intéressant d'étendre à l'analyse des données qualitatives, groupées sous forme de table de contingence, les modèles d'analyse de variance à effets aléatoires, en considérant les niveaux de certains facteurs comme échantillon aléatoire d'une population de niveaux.

Nous observons une variable réponse dichotomique sur les niveaux d'un facteur. Dans l'analyse classique, on étudie l'influence du facteur sur la probabilité de réponse, les paramètres représentant les effets du facteur sont supposés constants. Supposons maintenant que ce facteur ait un effet aléatoire sur la distribution de la variable réponse. L'introduction de ce facteur aléatoire induit une variable aléatoire non observable P qui traduit la variabilité des probabilités des cellules sur les différents niveaux de ce facteur.

Manuscrit reçu le 27 avril 1987
Révisé le 11 janvier 1989

Certains auteurs ont étudié ce plan d'expérience en développant d'un certain point de vue des méthodes bayésiennes pour l'appliquer dans divers domaines: toxicologie, génétique. ([1],[2A],[2B],[4A],[4B],[5],[6])

Ici, comme dans [3] et [8], nous nous plaçons dans le cas où on ne fixe pas de loi à priori sur la distribution des effets du facteur. Ceci nous amène à introduire deux paramètres qui sont la moyenne et la variance de la variable aléatoire P. Il nous a paru ([7]) important d'estimer non seulement la moyenne mais également la variance, de donner une loi limite de l'estimateur de ce dernier paramètre afin de construire des tests d'hypothèses le concernant.

Dans un premier temps on pourra s'assurer que l'hypothèse nulle "variance = 0" est bien rejetée, ce qui revient à s'assurer de l'aléas du facteur considéré. Dans un second temps, lorsque l'on disposera de différents groupes on pourra tester la différence de variabilité du facteur au sein des groupes. ([7])

1) Description

Considérons n individus ou unités statistiques (I_i ,i=1,...,n). Chacune de ces unités I_i donne lieu à m_i réponses 0 ou 1. (Exemple: une famille de n taureaux dont les descendants sont classés en 2 catégories: vivants ou morts à un âge fixé).

Si on ne s'intéressse qu'à ces n unités, le facteur est à effets fixes: on peut chercher à comparer ces n unités. Le ième niveau est formé des réponses de l'unité i; m_i étant le nombre total de réponses, on dénombre x_i réponses 1. Si on admet que les réponses sont indépendantes les unes des autres, la loi de la variable aléatoire X_i "nombre de réponse 1" est binomiale de paramètres (m_i,p_i) où p_i appartient à [0,1]. La proportion de "réponse 1" de l'individu i est p_i; la variation du nombre de "réponse 1" est mesurée par la variance: $varX_i = m_i p_i(1-p_i)$.

Si les n unités ont les mêmes caractéristiques a priori, elle peuvent être considérées comme un échantillon tiré au hasard dans une population homogène. Les probabilités $p_1,p_2,\ldots,p_n$ sont des réalisations respectives de variables aléatoires $P_1,P_2,\ldots,P_n$ indépendantes et suivant une même loi; la variabilité entre les unités est mesurée par la variance commune des variables aléatoires P_i.

Il y a corrélation entre les réponses d'un même individu. On introduit alors n couples aléatoires indépendants (X_i,P_i); X_i est donc distribuée conditionnellement à $P_i = p_i$ comme une loi binomiale de paramètre (m_i,p_i).
Les n variables aléatoires P_i sont distribuées comme une variable aléatoire P de moyenne μ et de variance σ^2 , (μ et σ^2 sont des paramètres inconnus, $\sigma^2 > 0$).

On pose $X_i = \sum_{k=1}^{m_i} Z_{ik}$, où les Z_{ik} ($k = 1,\ldots,m_i$; $m_i \neq 0$) sont des variables aléatoires

à valeur dans $\{0,1\}$. (ce sont les m_i réponses de chaque unité I_i)
On a alors pour un niveau i du facteur:

$$E(Z_{ik}) = E(E(Z_{ik}/P_i)) = E(P_i) = \mu$$

et $\text{Var } Z_{ik} = \mu(1-\mu)$, car $Z_{ik}^2 = Z_{ik}$ (pour $k = 1,\ldots,m_i$)

De plus pour $k \neq k'$ $\text{cov}(Z_{ik}, Z_{ik'}) = \sigma^2$

On pose $Y_i = X_i/m_i$ et on a $E(Y_i) = \mu$ et $\text{var } Y_i = \mu(1-\mu)/m_i + (1-1/m_i)\,\sigma^2$

On peut remarquer que Var Y_i (fonction de μ et σ^2) ne dépend de σ^2 que pour $m_i > 1$. D'autre part si $m_i = 1$, Y_i prend les seules valeurs 0 ou 1 avec les probabilités $1-\mu$ et μ ce qui rend la loi des Y_i indépendante de σ^2.
On se placera dans le cas où pour tout i ($i = 1,\ldots,n$) $m_i > 1$.

2) Modèlisation

En résumé, considèrons deux groupes de variables aléatoires à réalisations dans [0,1]:

$(P_i, (i = 1,\ldots,n))$ sont des variables non observables;

$(Y_i, (i = 1,\ldots,n))$ sont des variables observées;

que l'on écrit sous la forme:

$$P_i = \mu + E_i \quad \text{et} \quad Y_i = \mu + E_i' \quad \text{où } \mu \text{ appartient à } [0,1]$$

Les E_i (resp E_i') sont des variables aléatoires indépendantes d'espérance mathématique nulle, de variance σ^2; σ^2 appartient à $[0, \mu(1-\mu)]$ (resp $f_i(\mu, \sigma^2)$; fonction des paramètres inconnus μ et σ^2).

Nous nous intéressons au cas où; m_i étant un entier observé strictement supérieur à 1; sachant $P_i = p_i$, $m_i Y_i$ est binomiale de paramètres m_i et p_i.
Ce qui implique: $f_i(\mu, \sigma^2) = \mu(1-\mu)/m_i + (1-1/m_i)\,\sigma^2$.

II– ESTIMATION

Nous allons introduire un estimateur sans biais du paramètre de variance σ^2 et étudier asymptotiquement la statistique envisagée.

On pose:

$$\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i \quad \text{et} \quad S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2$$

On a: $E(\overline{Y}_n) = \mu$ et $E(S_n^2) = \frac{1}{n} \sum_{i=1}^{n} \text{var}Y_i$

On introduit aussi: $\overline{u}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i}$

On peut remarquer que $\sigma^2 = [\frac{1}{n} \sum_{i=1}^{n} \text{var}Y_i - \overline{u}_n \mu(1-\mu)] / (1-\overline{u}_n)$

Considérons alors commme estimateur de σ^2 la statistique
$[S_n^2 - \overline{u}_n \overline{Y}_n(1-\overline{Y}_n)] / (1-\overline{u}_n)$.

C'est un estimateur convergent sous la condition C_1:
"la suite $\overline{u}_n$ a une limite." (cette limite appartient nécessairement à [0,1/2]);
mais biaisé à distance finie. On peut supprimer le biais en appliquant une procédure de correction du biais. On obtient alors la statistique V_n d'espérance mathématique σ^2.

$$V_n = [S_n^2(\frac{n-\overline{u}_n}{n}) - \overline{u}_n \overline{Y}_n(1-\overline{Y}_n)] / (1-\overline{u}_n)$$

qui est de même un estimateur convergent sous la condition C_1.

A ce niveau il paraît intéressant de confronter l'estimateur V_n à la contrainte:
σ^2 appartient à [0, $\mu(1-\mu)$] où $\mu(1-\mu)$ appartient à [0,1/4].

Nous allons donc montrer que : $0 \leq \underline{\lim}\, V_n \leq \overline{\lim}\, V_n \leq 1/4$

En remarquant que l'on peut écrire la statistique V_n sous les 2 écritures suivantes:

$$V_n = S_n^2 - \frac{\overline{u}_n}{1-\overline{u}_n} \frac{1}{n} \sum_{i=1}^{n} Y_i (1-Y_i)$$

et $$V_n = -\frac{1}{n} [\frac{n}{n-1} + \frac{\overline{u}_n}{1-\overline{u}_n}] \sum_{i=1}^{n} Y_i (1-Y_i) + \frac{n}{n-1} \overline{Y}_n(1-\overline{Y}_n);$$

en remarquant aussi que $\frac{\bar{u}_n}{1-\bar{u}_n}$ appartient à [0,1] puisque $m_i > 1$ pour tout i;

que $\bar{Y}_n(1-\bar{Y}_n)$ appartient à [0,1/4]; nous avons les intervalles:

$$-1/4 \cdot \frac{1}{n} \frac{\bar{u}_n}{1-\bar{u}_n} \leq V_n \leq \frac{n}{n-1} \bar{Y}_n(1-\bar{Y}_n)$$

et $0 \leq \underline{\lim}\, V_n \leq \overline{\lim}\, V_n \leq 1/4$

Etudions maintenant les propriétés asymptotiques quand n tend vers + l'infini, de l'estimateur V_n:

Nous allons montrer que sous la condition C_1 déjà cité plus haut et sous une nouvelle condition C_2 qui sera explicitée au cours de la démonstration et qui permet de satisfaire aux conditions particulières du théorème de Liapounov, nous avons la propriété suivante (quand n tend vers + l'infini):

$[F_n(\bar{Y}_n)]^{-1} n^{1/2} (V_n - \sigma^2)$ converge en loi vers une loi normale N(0,1)

où $F_n(\bar{Y}_n)$ est la fonction de $\bar{Y}_n$ précisée par la suite.

Posons tout d'abord $Y_i' = \frac{1}{n}(Y_i - \mu)$ et décomposons V_n sous la forme suivante:

$$V_n = \frac{1}{n} \sum_{i=1}^{n} [g_n(Y_i) - \frac{2n^2}{n-1} Y_i' \sum_{j=1}^{i-1} Y_j']$$

où $g_n(Y_i) = Y_i^2 \frac{1}{1-\bar{u}_n} - Y_i (\frac{\bar{u}_n}{1-\bar{u}_n} + 2\mu) + \mu^2$

On remarque que $E(g_n(Y_i)) \neq \sigma^2$ mais $\frac{1}{n} \sum_{i=1}^{n} E(g_n(Y_i)) = \sigma^2$

On pose alors: $h_n(Y_i) = g_n(Y_i) - E(g_n(Y_i)) + \sigma^2$

On a $\sum_{i=1}^{n} h_n(Y_i) = \sum_{i=1}^{n} g_n(Y_i)$

où $h_n(Y_i)$ est une suite de variables aléatoires indépendantes, d'espérance mathématique σ^2 et uniformément bornée.

De plus $varh_n(Y_i)$ (= $varg_n(Y_i)$) ne tend pas vers 0 quand n tend vers + l'infini.

Comme $E(n^{-1/2} \frac{2n^2}{n-1} \sum_{i=1}^{n} Y_i' \sum_{j=1}^{i-1} Y_j')^2$ tend vers 0 quand n tend vers + l'infini,

on obtient la convergence en loi vers une loi normale N(0,1) de l'expression:

$$\left(\sum_{i=1}^{n} varg_n(Y_i) \right)^{-1/2} n \, (V_n - \sigma^2)$$

Nous sommes maintenant amené à chercher un estimateur convergent de $\frac{1}{n} \sum_{i=1}^{n} varg_n(Y_i)$

On pose pour cela:

$$\Phi(i,\mu) = g_n(Y_i) - \mu^2$$

$$\Phi(\mu) = \frac{1}{n} \sum_{i=1}^{n} \Phi(i,\mu)$$

et $$F_n^2(\mu) = \frac{1}{n-1} \sum_{i=1}^{n} (\Phi(i,\mu) - \Phi(\mu))^2$$

On introduit la condition C_2:

"la suite $\frac{1}{n} \sum_{i=1}^{n} var\ \Phi(i,\mu) = \frac{1}{n} \sum_{i=1}^{n} varg_n(Y_i)$ *a une limite strictement positive."*

Si les deux conditions (C_1 et C_2) sont satisfaites, on peut remarquer alors:

i) $(\Phi(i,\mu), i=1,\ldots,n)$ sont n variables aléatoires indépendantes et uniformément bornées.

ii) $F_n^2(\mu) - \frac{1}{n} \sum_{i=1}^{n} varg_n(Y_i)$ converge en probabilité vers o.

iii) $\overline{Y}_n$ converge en probabilité vers μ.

D'où $F_n^2(\overline{Y}_n) - F_n^2(\mu)$ converge en probabilité vers 0.

$F_n^2(\overline{Y}_n)$ est donc un estimateur convergent de $\frac{1}{n} \sum_{i=1}^{n} varg_n(Y_i)$

Ce qui achève la démonstration de la propriété.

III- SIMULATION

1) Introduction du test

De la propriété du II nous pouvons déduire un test qui permet d'éprouver l'hypothèse H_0: "$\sigma^2 = \sigma_0^2$"contre H_1: "$\sigma^2 \neq \sigma_0^2$".

Nous noterons: $T(\overline{Y}_n) = [F_n(\overline{Y}_n)]^{-1} n^{1/2} (V_n - \sigma_0^2)$

Pour un niveau asymptotique "α" on rejette l'hypothèse H_0contre l'hypothèse H_1 si $|T(\overline{Y}_n)| > b(\alpha)$ avec $P(N(0,1)>b(\alpha)) = \alpha/2$

Dans la suite des simulations, on testera l'hypothèse H_0 contre l'hypothèse H_1 pour un niveau asymptotique $\alpha = 0.05$; on a donc $b(\alpha) = 1.96$.

2) Principe et description de la simulation

Nous nous proposons, par simulation, d'étudier empiriquement à taille finie les propriétés de ce test. La démarche que nous avons suivie, consiste à fixer μ et σ^2 puis à simuler "n observations" $p_1,p_2,\ldots,p_n$ d'une variable aléatoire à valeur dans [0,1] de moyenne μ et de variance σ^2; et pour chaque p_i nous simulons m_i variables de Bernoulli de paramètres p_i. Nous notons cette situation expérimentale $SE(n,\mu,\sigma^2)$.

Pour faire une étude du niveau empirique du test, on simulera 100 fois une même $SE(n,\mu_0,\sigma_0^2)$ et on notera le nombre de rejets de l'hypothèse H_0.

Pour faire une étude de la puissance empirique du test, on simulera 100 fois une même $SE(n,\mu_0,\sigma_1^2)$ avec $\sigma_1^2 \neq \sigma_0^2$ et on notera le nombre de rejets de l'hypothèse H_0: "$\sigma^2 = \sigma_0^2$"; on effectuera ce calcul pour diverses valeurs de σ_1^2.

Dans ce qui suit, nous simulerons des $SE(n,1/2,\sigma^2)$: c'est en effet pour la valeur $\mu_0 = 1/2$ que l'on peut avoir une variance "la plus variable possible" (dans ce cas σ^2 appartient à [0,1/4]).

2.1) Simulation des p_i

Dans un premier temps, en utilisant un générateur de loi uniforme sur [0,1], nous obtenons des nombres $(v_1,\ldots, v_n)$ appartenant à [0,1].

Dans un second temps, nous introduisons une transformation qui permet de faire varier la variance "σ^2" tout en la controlant.

Pour k fixé, >0, on pose:

$$p(i,k) = 1/2\ [1 - (1-2v_i)^{1/k}] \quad \text{si } v_i \leq 1/2$$

$$p(i,k) = 1/2\ [1 + (2v_i-1)^{1/k}] \quad \text{si } v_i \geq 1/2$$

Les p(i,k) sont alors considérés comme des réalisations de variables aléatoires P(i,k) pour i=1,...,n ; indépendantes et identiquement distribuées comme une variable aléatoire P(k) de moyenne 1/2 et de variance $\sigma^2(k) = 1/4.k/(k+2)$ (si k=1 on retrouve la loi uniforme; on obtient une variable aléatoire de plus petite variance que celle de la loi uniforme si k<1 et de plus grande variance si k>1).

2.2) Simulation des m_i (m_i est un entier > 1)

Nous avons choisi d'utiliser une loi de khi-deux de paramètres λ fixé (λ est le degré de liberté de la loi de khi-deux) en prenant pour les m_i (i=1,...,n) les parties entières des nombres générés.

Nous avons choisi $\lambda=20$ (ayant envisagé des $SE(n,1/2,\sigma^2(k))$ pour des valeurs de λ fixées à 20, 40, 60, 80; les résultats numériques ne semblaient pas être influencés de façon notable par ces différentes valeurs de λ).

3) Résultats numériques

3.1) Niveau empirique du test

Nous avons considéré les $SE(n,1/2,\sigma^2(k_0))$ où n prend les valeurs 10, 20, 30, 40, 50, et k_0 prend ses valeurs dans K avec K = {1/7, 1/6, 1/5, 1/4, 1/3, 1/2, 1, 2, 3, 4, 5, 6, 7}. (la variance σ^2 varie de $\sigma^2(1/7) = 0.0167$ à $\sigma^2(7) = 0.1944$).

Dans chaque cas nous avons fait 100 simulations et noté le nombre de rejets. Les résultats figurant dans le tableau 1; nous pouvons faire les commentaires suivants:

c_1) Il semble que le niveau empirique et le niveau asymptotique (ici $\alpha = 0.05$) sont du même ordre pour $n \geq 20$ et ceci quel que soit k_0.
Les résultats asymtotiques semblent donc applicables à condition que le nombre d'individus soit supérieur à 20.

c_2) Plus k_0 est petit, plus l'influence de n semble grande (la variance $\sigma^2(k_0)$ qui mesure l'effet du facteur est d'autant plus petite que k_0 est petit).
Ces résultats semblent donc indiquer que moins le facteur a d'effets, plus les performances du test sont sensibles au nombre de niveaux considérés.

Tableau I:

$\sigma^2(k_0)$	0.0167	0.0192	0.0227	0.0278	0.0357	0.05	0.0833	0.1250	0.150	0.1667	0.1785	0.1875	0.1944
k_0	1/7	1/6	1/5	1/4	1/3	1/2	1	2	3	4	5	6	7
n=10	26	24	27	11	21	14	17	9	6	5	12	4	11
n=20	10	8	20	15	11	9	8	6	0	6	4	13	5
n=30	9	9	7	13	11	12	10	7	5	9	6	6	5
n=40	14	10	11	12	5	7	7	9	5	2	4	2	4
n=50	11	4	17	6	8	6	6	5	6	2	7	3	5

Ex: pour n=20 et k=1/3, on a obtenu 11 rejets sur 100 de l'hypothèse H_0: "$\sigma^2 = 0.0357$".

3.2 Puissance empirique du test

Etant donné 100 simulations d'une $SE(n,1/2,\sigma^2(k_1))$ fixée par le choix de k_1, on relève le nombre de rejets de l'hypothèse H_0:"$\sigma^2 = \sigma^2(k_0)$" contre H_1: "$\sigma^2 \neq \sigma^2(k_0)$" pour k_0 donné.

si $k_0 = k_1$: c'est le nombre de rejets de l'hypothèse H_0 alors qu'elle est vraie; c'est à dire 100 fois le niveau empirique.

si $k_0 \neq k_1$: c'est le nombre de rejets de l'hypothèse H_0 alors qu'elle est fausse; c'est à dire 100 fois la puissance empirique.

Prenons $k_0 = 1$ $[\sigma^2(k_0) = 0.0833]$
Nous avons simulé des $SE(n,1/2,\sigma^2(k_1))$ pour n = 15, 25, 50; k_1 variant de 0.1 à 2 par pas de 0.1; et on note le nombre de rejets de l'hypothèse H_0 : "$\sigma^2 = 0.0833$" sur 100 simulations d'une $SE(n,1/2,\sigma^2(k_1))$ donnée.

Tableau II

k_1	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2
n=15	97	89	68	62	48	36	36	18	14	9	10	12	8	14	19	30	32	24	38	32
n=25	100	97	85	78	58	42	28	20	12	4	5	5	16	18	29	40	33	44	48	57
n=50	100	100	100	96	87	64	36	24	11	7	5	15	24	32	41	52	67	74	86	89

Ex: n=25 et k_1=0.5 . Etant donné 100 simulations de la SE(25, 1/2, $\sigma^2(0.5)$) on a rejeté 58 fois l'hypothèse H_0: "$\sigma^2 = 0.0833$".

Nous observons empiriquement sur le tableau I le caractère sans biais du test considéré.

Nous observons de plus (tableau II) l'influence du nombre de niveaux "n" sur la puissance du test: il semble que la puissance empirique tend vers "1" plus rapidement quand n est grand. C'est à dire que le test est d'autant plus puissant que n est grand.

BIBLIOGRAPHIE

[1] CHATFIELD, C. and GOODHARDT, G.J. (1970) :
The beta−binomial model for consumer purchasing behaviour.
J. Royal Statist. Soc. C 19, 240−250.

[2A] CROWDER, M.J. (1978) :
Beta binomial Anova for proportions.
J. Royal Statist. Soc. C 27, 34−37.

[2B] CROWDER, M.J. (1979) :
Inference about the intraclass correlation coefficient in the beta binomial Anova for proportions.
J. Royal Statist. Soc. B 41, 230−234.

[3] GLADEN, B. (1979) :
The use of Jackknife to estimate proportions from toxicological data in the presence of litter effects.
J. Amer. Stat. Soc. 74, 278−283.

[4A] HASEMAN, J.K. and KUPER, J.K. (1978) :
The use of a correlated binomial model for the analysis of a certain toxicological experiments.
Biometrics 34, 69−76.

[4B] HASEMAN, J.K. and KUPER, J.K. (1979) :
Analysis of dichotomous response data from certain toxicological experiments.
Biometrics 35, 281−294.

[5] IM, S. (1982) :
Contribution à l'étude des tables de contingences à paramètres aléatoires. Utilisation en Biométrie.
Thèse de 3ième cycle, Université Paul Sabatier, Toulouse.

[6] KLEIMANN, J.C. (1973) :
Proportions with extraneous variance: single and independant samples.
J. Amer. Stat. Soc. 68, 157−173.

[7] LAVERGNE, C. (1984) :
Contribution à l'étude des modèles à effets aléatoire dans l'analyse des données qualitatives.
Thèse de 3ième cycle, Université Paul Sabatier, Toulouse.

[8] VAN RYZIN, J. (11075) :
Estimating the mean of a random binomial parameter with trial size random.
Sankhya B 37, 10−27.