Zeitschrift: Archives des sciences et compte rendu des séances de la Société
Band: 48 (1995)
Heft: 2: Archives des Sciences

Artikel: Propriétés de quelques test d'ajustement pour données groupées : un chapitre choisi de la méthodologie statistique
Autor: Driancourt, Catherine / Streit, Franz
DOI: https://doi.org/10.5169/seals-740255

Archs Sci. Genève	Vol. 48	Fasc. 2	pp. 185-196	Septembre 1995

Communication présentée à la séance du 2 février 1995

PROPRIÉTÉS DE QUELQUES TESTS D'AJUSTEMENT POUR DONNÉES GROUPÉES. UN CHAPITRE CHOISI DE LA MÉTHODOLOGIE STATISTIQUE

PAR

Catherine DRIANCOURT* & Franz STREIT**

ABSTRACT

Properties of some goodness-of-fit tests for grouped data. - After a short review of the aims of statistics and of some notions concerning statistical tests and in particular goodness-of-fit tests we propose a general class of such goodness-of-fit tests and show an optimality property of these tests. We determine the mean, the variance and the asymptotic distributions of their test statistics. We study special cases in calculating the actual level of significance and the power function and by indicating the alternative hypotheses against which the tests are unbiased.

§ 1. Types d'expériences statistiques et leurs buts

La statistique s'occupe des phénomènes aléatoires, et plus particulièrement du cas où on analyse de tels phénomènes aléatoires, en observant les valeurs que prennent des variables aléatoires (v.a.) liées à ce phénomène aléatoire. Observer une telle v.a. équivaut à effectuer une expérience à issue incertaine, qui est liée à ce phénomène aléatoire et dont le résultat est un nombre réel. Si l'on choisit habilement les v.a. ainsi observées, on arrive à décrire beaucoup de phénomènes aléatoires, même des phénomènes aléatoires assez complexes et des phénomènes aléatoires à composantes qualitatives (cela, en utilisant des codages appropriés).

Les analyses statistiques se distinguent par leur niveau de complexité:

1) Il se peut que l'on veuille uniquement donner une description succincte des observations que l'on a prises du phénomène aléatoire en question. Dans ce cas, on fait de l'analyse des données d'une situation expérimentale concrète particulière ou de la statistique descriptive. On résume les observations en regroupant les données dans des tableaux, en visualisant l'ensemble de ces données graphiquement et en caractérisant les aspects typiques et la structure de cet ensemble par des indices ou

* Ecole Supérieure de Commerce, 14, avenue Trembley, case postale 118, CH-1211 Genève 28.
** Section de Mathématiques, 2-4, rue du Lièvre, case postale 240, CH-1211 Genève 24.

coefficients. On arrive ainsi à un abstrait précis des résultats des expériences faites dans le cas concret analysé.

2) Pour les besoins de la recherche scientifique actuelle, des analyses statistiques du type statistique descriptive ne suffisent d'habitude pas, parce qu'elles expriment uniquement des résultats de cas particuliers. Sans considérations additionnelles qui permettent de modéliser le phénomène aléatoire mathématiquement, et de cette façon d'expliquer le mécanisme aléatoire, on n'arrive pas à des résultats applicables à tous les phénomènes aléatoires semblables à celui que l'on a examiné. C'est là où les modèles stochastiques de la théorie des probabilités interviennent. Ces modèles résument les tendances systématiques que l'on pense reconnaître dans les données observées et les combinent d'une façon précise avec des éléments purement stochastiques, qui représentent les fluctuations aléatoires (les déviations non-prédictibles des données individuelles de ces tendances). Chacun de ces modèles représente une interprétation possible des données, un modèle idéalisé de la réalité. En se demandant lesquels des modèles stochastiques sont aptes à expliquer un phénomène aléatoire, on va a priori prendre en considération plusieurs modèles susceptibles de décrire le mécanisme aléatoire. Il faut alors effectuer un choix parmi ces possibilités. En utilisant les méthodes de la statistique inférentielle ou inductive, on arrive à faire ce choix d'une façon aussi optimale que possible (du point de vue minimisation des effets néfastes d'un choix incorrect).

§ 2. Tests d'ajustement

Ci-après, on ne va pas considérer des phénomènes aléatoires du type échantillon aléatoire, c'est-à-dire on considère n v.a. $X_1, \ldots, X_n$ et on suppose qu'elles sont engendrées par répétition d'une expérience aléatoire du même type, ainsi que les chances de prendre les différentes valeurs sont les mêmes pour chaque X_j et que les X_j ne s'influencent pas mutuellement. Il suffit donc d'indiquer le modèle probabiliste commun pour tous les X_j. Cela se fait par spécification de la fonction de répartition F_X, la même pour tous les X_j [$j = 1, \ldots, n$]. F_X est la fonction qui indique pour chaque valeur réelle x la probabilité de l'événement $X_j \leq x$, donc $F_X(x) = P(X \leq x)$. A l'aide de la fonction de répartition, on peut calculer la probabilité de n'importe quel événement aléatoire (donc n'importe quelle condition probabilisable) se rapportant à X_j. On a que

(*) $F_X(-\infty) = 0$, $F_X(+\infty) = 1$ et que F_X est non-décroissante et continue à droite.

Il reste donc un grand choix de modèles stochastiques a priori possibles. Chaque fonction F_X satisfaisant aux conditions (*) en représente une possibilité. Comme les X_j ne s'influencent pas mutuellement, on a

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = F_X(x_1) \cdot \ldots \cdot F_X(x_n) .$$

Le but de l'inférence statistique dans ce contexte est de connaître la vraie fonction de répartition F_X.

Parfois la situation expérimentale suggère un choix privilégié F_0^* de F_X, qui est justifié par une théorie potentiellement correcte. La question à résoudre prend alors la forme suivante, caractéristique pour les tests d'ajustement : Décider laquelle des hypothèses

$H_0^{**}: F_X \equiv F_0^*$ (hypothèse nulle)

$H_1^{**}: F_X \not\equiv F_0^*$ (Hypothèse alternative)

est correcte. On peut montrer que tester H_0^{**} versus H_1^{**} revient pour F_0^* continue à la même chose que de tester

$H_0^*: F_X(x) = F_0(x) := x \quad [0 \leq x \leq 1]$ versus

$H_1^*: F_X \not\equiv F_0$

pour les variables transformées $F_0^*(X_1), \ldots, F_0^*(X_n)$. (Witting et al., 1970). En ce qui suit, nous allons donc supposer sans perte de généralité que la fonction de répartition de $X_1, \ldots, X_n$ à approuver ou rejeter est F_0, comme définie sous $H_0^\star$ (version standard du test d'ajustement). Il est alors raisonnable de limiter le choix de l'alternative à des fonctions de répartition satisfaisant $F_X(0) = 0$ et $F_X(1) = 1$.

La situation de test décrite jusqu'à présent se réfère au cas où l'on dispose de données observées exactes à plusieurs décimales. Mais souvent les données que l'on obtient sont déjà groupées, c'est-à-dire est seulement indiqué pour chaque observation l'intervalle dans lequel elle se trouve, et non pas sa valeur précise. On analyse dans la suite la subdivision de (0, 1] dans k sous-intervalles de longueurs égales, en posant $I_i = (k^{-1}(i-1), k^{-1}i]$ pour le i-ème sous-intervalle.

On écrira

$$p_i = P(X_j \in I_i) = F_X(\mathrm{k}^{-1}i) - F_X(\mathrm{k}^{-1}(i-1))$$

$$[i = 1, \ldots, k, j = 1, \ldots, n]$$

pour la probabilité que la j-ème observation tombe dans le i-ème intervalle et

$N_i = N_i(n, k)$ pour le nombre des données observées à valeur dans I_i.

Pour cette manière de rapporter les résultats de l'observation, le problème du test d'ajustement se modifie un peu, dû au fait qu'il est impossible de contrôler les valeurs

de F_X pour tout argument x. Seules les hypothèses concernant les probabilités p_i attribuées par F_X aux intervalles I_i peuvent être contrôlées à cause du regroupement des données. Le problème du test d'ajustement pour données groupées dans sa version standard se réfère donc aux hypothèses

$H_0 : p_i = k^{-1} \quad [i = 1, \ldots, k]$

donc $\underset{\sim}{p} = (p_1, \ldots, p_k)' = k^{-1}\underset{\sim}{1}$ où $\underset{\sim}{1} = (1, \ldots, 1)'$ e $\mathbb{R}^k$ et $'$ désigne la transposition d'un vecteur et

H_1 : il existe un i_0 tel que $p_{i\,0} \neq k^{-1}[i_0 \in \{1, \ldots, k\}]$.

Un test est une règle de décision qui nous indique pour chaque réalisation de l'échantillon aléatoire, donc pour chaque n-tuple de nombres réels admissibles comme résultat s'il faut accepter H_0 ou H_1. D'habitude on spécifie un test en spécifiant une statistique (donc une fonction t de $X_1, \ldots, X_n$ qui prend des valeurs réelles et pour laquelle on est capable de déterminer la fonction de répartition $F_t(X_1, \ldots, X_n)$ à partir de F_X) et on spécifie une valeur (critique) c. Dans la version standard du test, on va comparer la valeur obtenue pour t, à savoir $t(x_1, \ldots, x_n)$, à c et rejeter H_0 en faveur de H_1 si $t(x_1, \ldots, x_n)$ dépasse c. Naturellement la qualité du test dépend du choix de t et c. Cette qualité peut être mesurée par deux types de probabilités : La probabilité de commettre une erreur de première espèce en acceptant H_1 quand effectivement H_0 est correcte et la probabilité de commettre une erreur de seconde espèce en acceptant H_0 quand effectivement H_1 est correcte. La dernière probabilité mentionnée dépend du vecteur $\underset{\sim}{p} = (p_1, \ldots, p_k)'$. On a $\alpha_{eff} = P(t\,(X_1, \ldots, X_n) > c : k^{-1}\underset{\sim}{1})$ (probabilité d'accepter H_1 quand H_0 est correcte; α_{eff} est appelée niveau de signification effectif du test) et plus généralement on écrit $\mathrm{Pui}((p_1, \ldots, p_k)') = P(t(X_1, \ldots, X_n) > c: (p_1, \ldots, p_k)')$ pour la probabilité d'accepter H_1 si $\underset{\sim}{p} = (p_1, \ldots, p_k)'$ (Pui est appelée fonction de puissance). Un test (uniformément) optimal va avoir pour α prescrit et proche de 0 – on appelle α le niveau de signification (nominal) – l'inégalité $\alpha_{eff} \leq \alpha$ satisfaite et sous cette restriction les valeurs de sa fonction de puissance vont être aussi élevées que possible pour $p \neq k^{-1}\underset{\sim}{1}$. Dans notre problème modifié du test d'ajustement où $\underset{\sim}{X} = (X_1, \ldots, X_n)'$ est réduit à $\underset{\sim}{N} = (N_1, \ldots, N_k)'$ dû au regroupement des observations, le résultat fondamental de la théorie des tests (Lehmann, 1986, Chapitre 3) nous indique qu'un procédé optimal pour tester H_0 versus H_1 $(p^*) : \underset{\sim}{p} = \underset{\sim}{p}^*$, où $\underset{\sim}{p}^*$ est fixé, existe; comme statistique du test de ce procédé, on peut choisir $\tilde{t}(\underset{\sim}{N})$ aux valeurs

$$t(\underset{\sim}{x}) = \tilde{t}(\underset{\sim}{n}) = P(N_1 = n_1, \ldots, N_k = n_k : \underset{\sim}{p}^*)[P(N_1 = n_1, \ldots, N_k = n_k : \underset{\sim}{p} = k^{-1}\underset{\sim}{1}]^{-1}$$

$$= k^n p_1^{*n_1} \cdot \ldots \cdot p_k^{*n_k}$$

$\tilde{t}(\underset{\sim}{N})$ suit une distribution discrète et pour une borne α prescrit $c = c(\alpha)$ doit être choisi de telle façon que $P(t(X_1, \ldots, X_n) \geq c: k^{-1}\underset{\sim}{1}) \leq \alpha$ et que cette inégalité soit aussi proche que possible d'une égalité.

Comme $\tilde{t}(\underset{\sim}{N})$ et $ln(k^{-n}\tilde{t}(\underset{\sim}{N}))$ sont des statistiques équivalentes, on peut – en ajustant c – exprimer le test de manière équivalente à l'aide de la statistique

$$N_1 \ln(p_1^*) + \ldots + N_k \ln(p_k^*).$$

§ 3. Quelques propriétés d'une classe de tests

On est ainsi amené à considérer des tests basés sur l'utilisation des statistiques

$$V_{\underset{\sim}{C}}(n, k) = \tilde{V}(N_1, \ldots, N_k) = \sum_{i=1}^{k} c_i N_i\,(n, k),$$

où $c_1, \ldots, c_k$ sont des constantes non-positives réelles. Nous allons déduire quelques propriétés de $V_{\underset{\sim}{C}}(n, k)$.

A.) Calcul de l'espérance mathématique et de la variance de $V_{\underset{\sim}{C}}(n, k)$.

A cette fin, il est important de constater que $V_{\underset{\sim}{C}}(n, k)$ admet une représentation comme somme de contributions indépendantes provenant des variables aléatoires X_i individuelles. On a

$$V_{\underset{\sim}{C}}(n, k) = \sum_{j=1}^{n} V_{Xj}\,(1, k) \quad \text{où}$$

$$V_{Xj}\,(1, k) = \sum_{i=1}^{k} c_i 1_{\{I_i\}}(X_j) \quad \text{où } 1_{I_i}(u) = 1 \text{ pour } u \in I_i \text{ et } 0 \text{ pour } u \notin I_i.$$

Pour tout $\underset{\sim}{p} \in \mathbb{R}^k$ tel que $p_i \geq 0$ et $p_1 + \ldots + p_k = 1$, on trouve

$$E[V_{Xj}\,(1, k) : \underset{\sim}{p}\,] = \sum_{i=1}^{k} c_i\, p_i$$

et

$$\mathrm{Cov}[V_{Xj_1}(1, k), V_{Xj_2}(1, k) : p] = 0 \quad \text{si} \quad j_1 \neq j_2$$

(cela est dû à l'indépendance stochastique de X_{j_1} et X_{j_2} qui entraîne celle des variables $V_{Xj_1}\,(1, k)$ et $V_{Xj_2}\,(1, k)$) et d'autre part

$$\mathrm{Var}[V_{Xj}\,(1, k) : \underset{\sim}{p}\,] = \sum_{i=1}^{k} c_i^2\, \mathrm{Var}[1_{Ii}\,(X_j) : \underset{\sim}{p}\,]$$

$$+\sum_{i_1=1}^{k}\sum_{i_2=1}^{k} c_{i1}\, c_{i2}\, \mathrm{Cov}[\, 1_{Ii_1}(X_j),\, 1_{Ii_2}(X_j) : \underset{\sim}{p}\,]$$

$$i_1 \neq i_2$$

$$= \sum_{i=1}^{k} c_i^2 p_i(1-p_i) - \sum_{i_1=1}^{k}\sum_{\substack{i_2=1 \\ i_1 \neq i_2}}^{k} c_{i_1}\, c_{i_2}\, p_{i_1}\, p_{i_2}$$

$$= \sum_{i=1}^{k} c_i^2 p_i - \sum_{i_1=1}^{k}\sum_{i_2=1}^{k} c_{i_1}\, c_{i_2}\, p_{i_1}\, p_{i_2}$$

$$= (\sum_{i=1}^{k} c_i^2 p_i - (\sum_{i=1}^{k} c_i\, p_i)^2.$$

Notons que l'on obtient ce résultat également à partir de la formule développée pour la variance, à savoir en prenant comme point de départ la relation

$\mathrm{Var}[V_{Xj}\,(1, k) : \mathrm{p}\,] = E[\, V_{Xj}\,(1, k)^2 : \underset{\sim}{p}\,] - (E[\, V_{Xj}\,(1, k) : \underset{\sim}{p}\,])^2$. A l'aide de ces considérations, on obtient donc

$$E[\, \underset{\sim}{V_c}(\mathrm{n}, k) : \underset{\sim}{p}\,] = \sum_{i=1}^{n} E[\, V_{Xi}\,(1, k) : \underset{\sim}{p}\,] = n \sum_{i=1}^{k} c_i\, p_i\,.$$

$$\mathrm{Var}[\, \underset{\sim}{V_c}(n, k) : \underset{\sim}{p}\,] = \sum_{i=1}^{n} \mathrm{Var}[\, V_{Xi}\,(1, k) : \underset{\sim}{p}\,] = n\,(\sum_{i=1}^{k} c_i^2 p_i - (\sum_{i=1}^{k} c_i\, p_i)^2).$$

B.) Calcul de la distribution asymptotique de la $\underset{\sim}{V_c}\,(n, k)$.

Nous avons constaté sous A.) que $V_c\,(n, k)$ admet la représentation comme somme de v.a. indépendantes et identiquement distribuées à variance positive, pourvu qu'au moins deux composantes de p soient positives et que les coefficients c_i correspondants soient différents et différents de 0. Sous cette condition, on peut donc appliquer le théorème limite central sous sa forme la plus classique (Chung, 1974) et on trouve pour le cas où la taille n de l'échantillon aléatoire tend vers ∞ que

$$\underset{\sim}{V_c}(n, k) \xrightarrow{L} N\,(n \sum_{i=1}^{k} c_i\, p_i\,,\, n\,[\sum_{i=1}^{k} c_i^2 p_i - (\sum_{i=1}^{k} c_i\, p_i)^2])$$

c'est-à-dire $V_c\,(n, k)$ tend vers une distribution normale quand le nombre de v.a. observées augmente vers l'infini. La relation indiquée signifie que

$$\lim_{\mathrm{n} \to \infty} F_{\underset{\sim}{V_c}\,(n, k)}\,(x) = F_W(x) \text{ où } x \in \mathbb{R} \qquad \text{et } W$$

est une v.a. Gaussienne à distribution normale N de paramètres comme indiqués. Cela

entraîne que pour un échantillon aléatoire de grande taille (n au moins 40 est recommandé), $V_c(n, k)$ se comporte approximativement comme une v.a. normale, et par conséquent une table de la distribution normale centrée réduite peut être utilisée pour déterminer d'une manière approximative les probabilités des événements aléatoires concernant $V_c(n, k)$ sans commettre des erreurs considérables.

§ 4. Cas spéciaux

I.) $c_i = -i\ ln\ (d_k)\ \ [i = 1, \ldots, k]$.

On a $V_c(n, k) = -ln(d_k) \sum_{i=1}^{k} i\ N_i(n, k) = -ln(d_k) R^*(n, k)$.

a.) Propriété d'optimalité

Le test est optimal pour discerner les hypothèses H_0 et H_1: $p = (d_k^{-1}, \ldots, d_k^{-k})'$.

En effet $e^{ci} = d_k^{-i} = p_i^*$. Il s'agit donc d'un test optimal de H_0 versus une situation spécifique où les p_i^* décroissent pour i croissant d'une manière géométrique.

b.) Détermination de d_k.

De la relation $\sum_{l=1}^{k} d_k^{-l} = 1$ découle

$$2d_k^{-1} - d_k^{-(k+1)} - 1 \equiv 0.$$

Cette équation admet une solution unique du problème considéré. On trouve

pour k = 2 $\quad d_2 = 1{,}62$

k = 3 $\quad d_3 = 1{,}83$

k$\to \infty$ $\quad d_\infty = 2$.

Selon le choix de k d_k prend différentes valeurs dans l'intervalle [1, 62 ; 2].

c.) Etude de la statistique $R^*(n, k)$ équivalente à $V_c(n, k)$ dans le cas I.

Considérons maintenant la statistique

$$R^*(n, k) = \sum_{i=1}^{k} i\ N_i(n, k).$$

Elle ne se distingue de $V_c\ (n, k)$ avec $c_i = -i\ ln(d_k)$ que par une constante de proportionnalité et $R^*(n, k)$ est donc une statistique de test équivalent à $V_c(n, k)$. Evidemment il faut ajuster la région critique du test si l'on teste H_0 versus H_1 comme indiqué sous a.), la région critique de ce test a maintenant la forme $R^*(n, k) < c$, donc elle est unilatérale du côté gauche. Notons que la statistique $R^*(n, k)$ est liée à la statistique $R(n, k)$ mentionnée dans (Maag *et al.*, 1973) et sa forme spécialisée $R(n, n) = S^*(n)$ mentionnée dans (Riedwyl, 1967) par la relation

$$R(n, k) = n^{-1}\, R^*(n, k) - 2^{-1}(k + 1).$$

Les formules pour l'espérance mathématique et la variance $R^*(n, k)$ sont

$$E[\, R^*(n, k) : \underset{\sim}{p}\,] = n \sum_{i=1}^{k} i\, p_i \quad \text{et}$$

$$\mathrm{Var}[\, R^*(n, k) : \underset{\sim}{p}\,] = n\Big(\sum_{i=1}^{k} i^2 p_i - \Big(\sum_{i=1}^{k} i\, p_i\Big)^2\Big).$$

En particulier on trouve

$$E[\, R^*(n, k) : H_0] = n\, 2^{-1}(k + 1) \quad \text{et}$$

$$\mathrm{Var}[\, R^*(n, k) : H_0\,] = n(12)^{-1}\, (k^2 - 1).$$

Ces moments servent à déterminer les distributions asymptotiques normales pour n grand.

d.) Détermination du niveau de signification effectif et de la fonction de puissance.

Déterminons maintenant le niveau de signification exact et la puissance exacte du test à région critique $R^*(n, k) < c$ pour n finie. En tenant compte du fait que $\underset{\sim}{N} = (N_1, \ldots, N_k)'$ suit une distribution hypergéométrique

$$\mathrm{Pui}(\underset{\sim}{p}) = \sum_{\Delta} n![\, n_1\, ! \ldots n_k!]^{-1}\, p_1^{n_1} \ldots p_k^{n_k}$$

où $\Delta = \{(n_1, \ldots, n_k) \in \{0, 1, 2, \ldots\}^k$ tels que:
$n_1 + n_2 + \ldots + n_k = n$ et $n_1 + 2n_2 + \ldots + kn_k < c\}$.

Un calcul simple pour n fixé tenant compte de la formule

$$(p_1 x + p_2 x^2 + \ldots + p_k\, x^k)^n = x^n\, (p_1 + p_2 x + \ldots + p_k\, x^{k-1})^n$$

$$= \sum_{m=n}^{nk} P(R^*(n, k) = m) x^m = \sum_{m=n}^{nk} \gamma_m(\underset{\sim}{p}) x^m$$

et de la récursion

où $\gamma_n(\underset{\sim}{p}) = p_1^n$

et $\gamma_{n+j}(\underset{\sim}{p}) = \dfrac{1}{jp_1} \sum_{v=1}^{\min(j,\,k-1)} (vn - j + v) p_{v+1} \gamma_{n+j-v}(\underset{\sim}{p})$

$$[j = 1, 2, \ldots, (k-1)n]$$

conduit à

$$\mathrm{Pui}(\underset{\sim}{p}) = \sum_{m=n}^{[c-0]} \gamma_m(\underset{\sim}{p})$$

où $[c-0] := l - 1$ si $l - 1 < c \leq l$ et $l \in \{0, 1, 2, \ldots\}$.

α_{eff} s'obtient en substituant $p_i = k^{-1}$, donc $\underset{\sim}{p} = k^{-1}\underset{\sim}{1}$, dans les formules précédentes. Nous écrivons $\tilde{\gamma}_m$ comme abréviation de $\gamma_m(k^{-1}\underset{\sim}{1})$.

e.) Détermination des alternatives pour lesquelles le test est sans biais.

Nous disons que le test statistique à région critique $R^*(n, k) < c$ est sans biais pour tester $H_0 : \underset{\sim}{p} = k^{-1}\underset{\sim}{1}$ versus $H_1(p^*) : \underset{\sim}{p} = \underset{\sim}{p}^*$ quand $\mathrm{Pui}(\underset{\sim}{p}^*) \geq \alpha_{eff}$. Cette propriété désirable d'un test nous garantit que le test ne rejette H_0 pas plus fréquemment dans le cas où H_0 est correcte que dans le cas où H_0 est incorrecte. Selon le résultat d.) le test considéré est sans biais si et seulement si

$$\sum_{m=n}^{[c-0]} \gamma_m(\underset{\sim}{p}^*) \geq \sum_{m=n}^{[c-0]} \tilde{\gamma}_m.$$

Ce critère peut être utilisé pour délimiter la classe de toutes les hypothèses alternatives uniponctuelles par rapport auxquelles le $R^*(n, k)$-test est sans biais. Nous illustrons la technique à l'aide de l'exemple suivant qui traite le cas $n = 3$ et $k = 4$.

Table des valeurs possibles de $R^*(n, k)$ désignées par m, de $\gamma_m(\underset{\sim}{p}^*)$ et de $\tilde{\gamma}_m$.

m	$\gamma_m(\underset{\sim}{p}^*)$	$\tilde{\gamma}_m$
3	p_1^{*3}	1/64
4	$3p_1^{*2}\, p_2^*$	3/64

5	$3\,p_1^* p_2^{*2} + 3p_1^{*2}p_3^*$	6/64
6	$6p_1^* p_2^* p_3^* + 3p_1^{*2}p_4 + p_2^{*3}$	10/64
7	$6p_1^* p_2^* p_4^* + 3p_1^* p_3^{*2} + 3p_2^{*2} p_3^*$	12/64
8	$6p_1^* p_3^* p_4^* + 3p_2^* p_3^{*2} + 3p_2^{*2} p_4^*$	12/64
9	$6p_2^* p_3^* p_4^* + 3p_1^* p_4^{*2} + p_3^{*3}$	10/64
10	$3p_2^* p_4^{*2} + 3p_3^{*2} p_4^*$	6/64
11	$3p_3^* p_4^{*2}$	3/64
12	p_4^{*3}	1/64

Supposons que l'on désire que $\alpha = \alpha_{eff} = 5\%$. On constate

$$P(R^*(3, 4) < 4 : \underset{\sim}{p} = k^{-1}\,\underset{\sim}{1}) = 1/64 < 0{,}05$$

$$P(R^*(3, 4) < 5 : \underset{\sim}{p} = k^{-1}\,\underset{\sim}{1}) = 1/16 > 0{,}05.$$

On peut cependant réaliser $\alpha_{eff} = 0{,}05$ exactement en choisissant le test de région critique $r^*(3, 4) < 4$ qui rejette H_0 avec probabilité 11/15 quand $r^*(3, 4) = 4$ (test randomisé).

La puissance de ce test vaut

$$\text{Pui}\,(\underset{\sim}{p}^*) = P(R^*(3, 4) < 4;\ \underset{\sim}{p} = \underset{\sim}{p}^*) + (11/15)\,P(R^*(3, 4) = 4 : \underset{\sim}{p} = \underset{\sim}{p}^*) = p_1^{*3} + 11/5)p_1^{*2}\,p_2^*.$$

Le test est sans biais si et seulement si

$$p_1^{*3} + (11/5)\,p_1^{*2}\,p_2^* \geq 0{,}05.$$

On peut donc faire les remarques suivantes:

1.) Quand $p_1^* \in [0;\ 0,\ 146]$, l'inégalité indiquée n'est pas satisfaite et donc le test est biaisé.

2.) Le test est sans biais quand

$$p_1^* \in (0,\ 146;\ 0,\ 368] \quad \text{et} \quad p_2^* \geq 5(0{,}05 - p_1^{*3})p_1^{*-2}/11$$

avec $p_2^*, p_3^*, p_4^* \in [0, 1]$ tels que $\sum_{i=1}^{4} p_i^* = 1$. Il a du biais quand la condition indiquée pour p_1^* est satisfaite, mais celle pour p_2^* ne l'est pas.

3.) Le test est sans biais quand $p_1^* \in (0, 368; 1]$ et $p_2^*, p_3^*, p_4^* \in [0, 1]$ tels que $\sum_{i=1}^{4} p_i^* = 1$.

II.) $c_i = ln\ (id_k) \quad [i = 1, \ldots, k]$.

On a $V_{\underset{\sim}{c}}(n, k) = \sum_{i=1}^{k} ln\ (i)\ N_i(n, k) + nln(d_k)$

$= L^*(n, k) + nln\ (d_k)$.

a.) Propriété d'optimalité.

Le test est optimal pour discerner les hypothèses H_0 et H_1 : $\underset{\sim}{p} = (d_k, 2\ d_k, \ldots, kd_k)$. En effet $e^{ci} = id_k = p_i^*$.

Il s'agit donc d'un test optimal pour discerner les hypothèses H_0 et $H_1 : p_i^* = id_k$ où les p_i^* croissent avec i de la façon $p_i^* = ip_1^*$.

b.) Détermination de d_k.

De la relation $\sum_{i=1}^{k} p_i^* = 1$ découle $d_k = 2\ [k\ (k + 1)]^{-1}$.

c.) Etude de la statistique $L^*(n, k)$ équivalente à $V_{\underset{\sim}{c}}(n, k)$ dans le cas II.

Pour l'espérance mathématique et la variance, on trouve

$E[\ L^*(n, k) : \underset{\sim}{p}] = n \sum_{i=1}^{k} ln(i) p_i$ et

$\text{Var}\ [\ L^*(n, k) : \underset{\sim}{p}] = n\ (\sum_{i=1}^{k} p_i\ ln^2(i) - (\sum_{i=1}^{k} p_i\ ln(i))^2)$.

En particulier, on obtient pour $\underset{\sim}{p} = k^{-1}\ \underset{\sim}{1}$

$E[\ L^*(n, k) : H_0] = k^{-1}\ nln(k!)$ et

$\text{Var}[\ L^*(n, k) : H_0] = n(k^{-1} \sum_{i=1}^{k} ln^2\ (i) - k^{-2}\ ln^2(k!))$.

Pour n grand, les distributions asymptotiques de $L^*(n, k)$ sont Gaussiennes et déterminées par ces moments.

Une étude plus approfondie de quelques propriétés du R*(n, k)-test et de quelques autres tests d'ajustement se trouve dans la thèse du premier auteur (Driancourt, 1994).

RÉSUMÉ

Après un bref rappel des buts de la statistique et de quelques notions relatives aux tests statistiques, et en particulier aux tests d'ajustement, nous proposons une classe générale de tests d'ajustement et indiquons une propriété d'optimalité de ces tests. Nous étudions l'espérance mathématique, la variance et les distributions asymptotiques des statistiques concernées et traitons des cas particuliers en abordant la question de la détermination du niveau de signification effectif et de la fonction de puissance, et en analysant pour lesquelles des hypothèses alternatives les tests sont sans biais.

RÉFÉRENCES BIBLIOGRAPHIQUES

CHUNG, K.L. (1974). *A course in probability theory*. Academic Press, New York.

DRIANCOURT, C. (1994). *Analyse concrète de propriétés relatives à des tests d'ajustement*. Thèse. Université de Genève.

LEHMANN, E.L. (1986). *Testing statistical hypotheses*. J. Wiley, New York.

MAAG, U., STREIT, F. & DROULLY, P. (1973). Goodness-of-fit tests for grouped data. *J. Americ. Statist. Assoc.* 68: 462-465.

RIEDWYL, H. (1967). Goodness of fit. *J. Americ. Statist. Assoc.* 62: 390-398.

WITTING, H. & NÖLLE, G. (1970). *Angewandte Mathematische Statistik*, Teubner, Stuttgart, p. 106 (théorème 3.5).