独立性の検定に於ける χ^2-統計量の解釈

塩　谷　　実

(1954 年 12 月受付)

The Interpretation of χ^2-Statistic in the Test of Independence

Minoru SIOTANI

ABSTRACT: This paper gives a few remarks on the interpretation of χ^2-statistic in the test of independence. The partition of χ^2 in the two-way contingency table was given by H. O. Lancaster (1949 [1]). With his partition and the extension of it to three or more way contingency table, we can obtain the conspicuous interpretation of χ^2.

Institute of Statistical Mathematics.

1. 緒　言

χ^2-統計量は観測結果と仮設との喰い違いの度合を測るものであるが，それは全体としての喰い違いを意味して居る．故に若し χ^2 が有意に大きい値を取る場合には，その依つて来る原因を調べることが重要になつて来る．此の線に沿つて χ^2-統計量を成分に分割することが種々取扱われて居る（例えば [1],[2],[3]）．特に［1］に於いて，2つの属性の独立性を検定する時に屡々必要となる分割表に対して，χ^2-統計量が下の (2) の如く分割出来ることが示されている．此の分割は吾々が普通に使用する χ^2 の意味をはつきりさせて呉れる．更に［1］の結果を，3つ以上の属性の独立性を検定する時の χ^2 に対して拡張してやれば，一層はつきりした解釈を χ^2 に与える事が出来る．

2. χ^2-統計量の分割の例

a) 属性が2つの場合　　此の時の分割表に於ける i 行 j 列の要素を n_{ij} で表わし，周辺和を

$$n_{i.}=\sum_{j=1}^{s} n_{ij}\ (i=1,2,\cdots,r),\quad n_{.j}=\sum_{i=1}^{r} n_{ij}\ (j=1,2,\cdots,s),\quad n=\sum_{i=1}^{r} n_{i.}=\sum_{j=1}^{s} n_{.j}=\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$$

で表わす．(i,j) 階級の確率を p_{ij} とすれば吾々の検定すべき独立性の仮設は

$$H_1:\quad p_{ij}=p_{i.}p_{.j}\qquad i=1.\cdots,r;\ j=1,\cdots,s \tag{1}$$

である．玆に $p_{i.}=\sum_{j=1}^{s} p_{ij}$, $p_{.j}=\sum_{i=1}^{r} p_{ij}$, $\sum_{i=1}^{r} p_{i.}=\sum_{j=1}^{s} p_{.j}=\sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij}=1$. 仮設 H_1 の下に於ける χ^2 は次の様に分割される．

$$\begin{aligned}\chi^2{}_T&=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-np_{i.}p_{.j})^2}{np_{i.}p_{.j}}\\&=\sum_{i=1}^{r}\frac{(n_{i.}-np_{i.})^2}{np_{i.}}+\sum_{j=1}^{s}\frac{(n_{.j}-np_{.j})^2}{np_{j.}}+\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n_{i.}p_{.j}-n_{.j}p_{i.}+np_{i.}p_{.j})^2}{np_{i.}p_{.j}}\qquad (2)\\&=\chi^2{}_A+\chi^2{}_B+\chi^2{}_{AB}\end{aligned}$$

此の分割は［1］に於いて求められたものであるが，式を導く丈なら形式的に分散分析の分解方法を真似てやれば容易に得られる．［1］に依つて χ_A^2, χ_B^2, χ_{AB}^2 の3つの成分は，漸近的に夫々自由度 $(r-1)$, $(s-1)$, $(r-1)(s-1)$ をもつて互に独立に χ^2–分布に従う．

b) 属性が3つ以上の場合　属性が3つの場合を具体的に求めれば，それ以上の時も容易に理解されるであろう．a）の時と同様の記号を用いる．検定すべき仮設を

$$H_2:\ p_{ijk}=p_{i..}\,p_{.j.}\,p_{..k}\qquad i=1,\cdots,r;\ j=1,\cdots,s;\ k=1,\cdots,t \tag{3}$$

で表わす．但し

$$p_{i..}=\sum_{j=1}^{s}\sum_{k=1}^{t}p_{ijk},\qquad p_{.j.}=\sum_{i=1}^{r}\sum_{k=1}^{t}p_{ijk},\qquad p_{..k}=\sum_{i=1}^{r}\sum_{j=1}^{s}p_{ijk}$$

H_2 の下に於ける普通に使用される χ^2 の形は

$$\chi^2{}_T=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\frac{(n_{ijk}-np_{i..}\,p_{.j.}\,p_{..k})^2}{np_{i..}\,p_{.j.}\,p_{..k}} \tag{4}$$

であるが，此れは次の様に分割する事が出来る．

$$\begin{aligned}\chi^2{}_T=&\sum_{i=1}^{r}\frac{(n_{i..}-np_{i..})^2}{np_{i..}}+\sum_{j=1}^{s}\frac{(n_{.j.}-np_{.j.})^2}{np_{.j.}}+\sum_{k=1}^{t}\frac{(n_{..k}-np_{..k})^2}{np_{..k}}\\&+\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij.}-n_{i..}\,p_{.j.}-n_{.j.}\,p_{i..}+np_{i..}\,p_{.j.})^2}{np_{i..}\,p_{.j.}}\\&+\sum_{j=1}^{s}\sum_{k=1}^{t}\frac{(n_{.jk}-n_{.j.}\,p_{..k}-n_{..k}\,p_{.j.}+np_{.j.}\,p_{..k})^2}{np_{.j.}\,p_{..k}}\\&+\sum_{i=1}^{r}\sum_{k=1}^{t}\frac{(n_{i.k}-n_{i..}\,p_{..k}-n_{..k}\,p_{i..}+np_{i..}\,p_{..k})^2}{np_{i..}\,p_{..k}}\\&+\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\frac{(n_{ijk}-n_{ij.}\,p_{..k}-n_{.jk}\,p_{i..}-n_{i.k}\,p_{.j.}+n_{i..}\,p_{.i.}\,p_{..k}+n_{.j.}\,p_{i..}\,p_{..k}+n_{..k}\,p_{i..}\,p_{.j.}-np_{i..}\,p_{.j.}\,p_{..k})^2}{np_{i..}\,p_{.j.}\,p_{..k}}\\=&\chi^2_A+\chi^2_B+\chi^2_C+\chi^2_{AB}+\chi^2_{BC}+\chi^2_{CA}+\chi^2_{ABC}\end{aligned} \tag{5}$$

各成分の自由度は夫々 $(r-1)$, $(s-1)$, $(t-1)$, $(r-1)(s-1)$, $(s-1)(t-1)$, $(t-1)(r-1)$, $(r-1)(s-1)(t-1)$ である．

3. 解　　釈

（2）の分解に於いて，$\chi^2{}_A$ は行（属性 A）に於ける当てはめに依る喰い違いを測るものであり，$\chi^2{}_B$ は列（属性 B）に対する同様の意味の測度である．$\chi^2{}_{AB}$ は全体の喰い違いから，此等単に当てはめのみに貢献する成分を取り除いたものであり，此れこそ A, B の交互作用即ち関聯の度合を表わすものである．即ち独立性の仮設からの距りを表わす $\chi^2{}_T$ は，直ちに関聯度を表わすのではなく，その中に周辺に於ける当てはめに依る喰い違いを測る成分を含んで居るのである．

若し $p_{i.}$, $p_{.j}$ が未知で観測結果に基いて推定しなければならない時には，周知の如く．修正された最小 χ^2 法[4]即ち最尤法に依り $\hat{p}_{i.}=\dfrac{n_{i.}}{n}$, $\hat{p}_{.j}=\dfrac{n_{.j}}{n}$ を用いる．此れは（2）に依り $\chi^2{}_A$, $\chi^2{}_B$ を0にする．換言すれば，周辺に於ける当てはめを完全にする様なものである．此の時には，従つて，自由度は $\chi^2{}_A$, $\chi^2{}_B$ の所で失われ，$\chi^2{}_{AB}$ では不変である．吾々が上の様な推定値を採用する限り，推定値を入れた後の $\hat{\chi}^2{}_T$ と $\hat{\chi}^2{}_{AB}$ は一致し，普通に用いられて居る統計量

$$\hat{\chi}^2{}_T\hat{\chi}^2{}_{AB}=n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{ij}-\dfrac{n_{i.}\,n_{.j}}{n}\right)^2}{n_{i.}\,n_{.j}} \tag{6}$$

を得る．更に此れが有意に大きい時に

$$f^2=\frac{\hat{\chi}^2{}_T}{n} \tag{7}$$

をもつて A, B の関聯度を測るものとする事の妥当性が出て来る．

併し実験遺伝学等に於いては，周辺確率 $p_{i\cdot}, p_{\cdot j}$ が過去の経験等から予め指定されて居ることも屢々である．此の時をも仮設に含めて

$$H_3:\quad p_{i\cdot}=p_{i\cdot}{}^0;\quad p_{\cdot j}=p_{\cdot j}{}^0;\quad p_{ij}=p_{i\cdot}\,p_{\cdot j}(=p_{i\cdot}{}^0 p_{\cdot j}{}^0) \tag{8}$$

を検定するものと考えて見る．若し $p_{i\cdot}, p_{\cdot j}$ の値に関する知識が，標本の取られた母集団に対して充分なものであれば，周辺に於ける当てはめは良いと考えられる．従つて $\chi^2{}_A$, $\chi^2{}_B$ の値は充分小さく出て来ることが予測され，此の時には χ^2_T に依る独立性の検定が意味を持つ．しかし周辺に於ける当てはめが充分なものでない時には，χ^2_T に依る検定は次に示す様に曖昧な意味しか持たなくなる．今仮設 H_3 に於いて，独立性の仮設は真であるが，周辺確率に関するものは真でなかつた時を考えて見る．即ち対立仮設

$$H_3^*:\quad p_{i\cdot}=p_{i\cdot}{}^*;\quad p_{\cdot j}=p_{\cdot j}{}^*;\quad p_{ij}=p_{i\cdot}\,p_{\cdot j}(=p_{i\cdot}{}^* p_{\cdot j}{}^*) \tag{9}$$

が真であつたとしよう．此の $H_3{}^*$ の下に（2）の各項の期待値を計算して見れば，容易に次の結果を得る．

$$\left.\begin{aligned}
E(\chi^2{}_A|H_{3*})&=(r-1)+(n-1)\left\{\sum_{i=1}^{r}\varDelta_i{}^2 p_{i\cdot}{}^2+\frac{1}{n-1}\sum_{i=1}^{r}\varDelta_i\right\}\\
E(\chi^2{}_B|H_3\ )&=(s-1)+(n-1)\left\{\sum_{j=1}^{s}\delta_j{}^2 p_{\cdot}{}^0{}_j+\frac{1}{n-1}\sum^{s}\delta_j\right\}\\
E(\chi^2{}_{AB}|H_3{}^*)&=(r-1)(s-1)+(n-1)\left\{\sum_{i=1}^{r}\varDelta_i{}^2 p_{i\cdot}{}^0\sum_{j=1}^{s}\delta_j{}^2 p_{\cdot j}{}^0+\frac{r-1}{n-1}\sum_{j=1}^{s}\delta_j+\frac{s-1}{n-1}\sum_{i=1}^{r}\varDelta_i\right\}
\end{aligned}\right\} \tag{10}$$

但し $\varDelta_i=\dfrac{p_{i\cdot}{}^*-p_{i\cdot}{}^0}{p_{i\cdot}{}^0}$, $\delta_j=\dfrac{p_{\cdot j}{}^*-p_{\cdot j}{}^0}{p_{\cdot j}{}^0}$ である．分散も同様の計算が可能であるがもつと複雑になる．此処で当てはめに依る喰い違いは $\varDelta_i$, δ_j として入つて来て居るが，此れが $\chi^2{}_{AB}$ の平均にも影響を与えて居る．各周辺の級の数 r, s は有限であるが，標本数 n は充分大きくとられるから，(10) の｛ ｝の中の第2項以下は省略出来る．若しすべての i, j に対して $\varDelta_i\approx 0$, $\delta_j\approx 0$ であれば，喰い違いは殆んどなく，実験の誤差の範囲内で $\chi^2{}_T\approx\chi^2{}_{AB}$ となるから普通の独立性の検定が意味を持つが，若し $\varDelta_i$, δ_j の中に可成りの値を持つものがある時には $\chi^2{}_A$, $\chi^2{}_B$ の平均だけでなく $\chi^2{}_{AB}$ の平均も可成り大きくなつて来る．従つて $H_3{}^*$ の下に於ける $\chi^2{}_A$, $\chi^2{}_B$, $\chi^2{}_{AB}$ の分布は，H_3 の時よりも右側に偏する様になり，普通の帰無仮設検定では棄却される可能性が増す．勿論 $\chi^2{}_T$ も有意に大きな値を示す確率は大きくなる．実際 $H_3{}^*$ の下に於ける（2）の各 χ^2–統計量の分布は最早や普通の Central χ^2–distribution では近似出来ないで．大きな平均を持つ Non-central χ^2-distribution に依る近似をしなければならないのである．以上の事から $p_{i\cdot}$, $p_{\cdot j}$ の値が予め指定されて居る時には，仮令独立性は成立して居ても，$\chi^2{}_{AB}$ 従つて $\chi^2{}_T$ の値が有意に大きな値を取る事が起るのであるから，$\chi^2{}_T$ の有意性から直ちに独立性の仮設が棄却されるのでもなく，$f^2=\chi^2{}_T/n$ 或いは $\chi^2{}_{AB}/n$ を以つて属性の関聯度を測るものとする事も不当である．此処に $\chi^2{}_T$ の分解の意義があるのである．$\chi^2{}_A, \chi^2{}_B$ の少くとも一方が有意な値を持つ時は，もつと深い考察から標本のとられた母集団に於ける $p_{i\cdot}, p_{\cdot j}$ に関して充分なる知識を求めるか．或いは，自由度を犠牲にしても標本からの推定値を用いて独立性の検定をやり直すことが必要となつて来る．

更に $p_{i\cdot}$, $p_{\cdot j}$ が $p_{i\cdot}(\theta_1,\cdots,\theta_\nu)$, $p_{\cdot j}(\theta_1',\cdots,\theta_\mu')$ の如く未知のパラメータを含み，此等が観測結果から推定される場合にも上に述べた事情が考えられねばならない．勿論推定の基礎になる測

度は χ^2_A, χ^2_B である．

今迄属性が2つの時の χ^2-統計量に就いて述べて来たが．属性が3つ以上ある時を考えればもつとはつきりして来る．仮設 H_2 の下に（5）に於ける各成分は，漸進的に夫々の自由度をもつて，しかも互に独立に χ^2-分布に従うが，検定の結果 χ^2_T が有意に大きいことがわかつても，直ちに．3つの属性 A, B, C の間の独立性の仮設が破れる原因が何処にあるかはわからないのである．χ^2_A, χ^2_B, χ^2_C は各属性に対する当てはめに依る喰い違いを測り，χ^2_{AB}, χ^2_{BC}, χ^2_{CA} は夫々2つの属性の独立性からの距りを表わして居る．一般に3つの属性の任意の2つが独立であつても，3つの属性の間の独立性は必ずしも言えないから，χ^2_{ABC} が A, B, C の間の独立性からの距りを表わして居ると解釈される．$p_{i\cdot\cdot}$, $p_{\cdot j\cdot}$, $p_{\cdot\cdot k}$ を観測結果から推定する時には，前と同様に周辺に於ける当てはめを完全にする様なもの，即ち $\chi^2_A, \chi^2_B, \chi^2_C$ を0にする様な推定値 $\hat{p}_{i\cdot\cdot}=\frac{n_{i\cdot\cdot}}{n}$, $\hat{p}_{\cdot j\cdot}=\frac{n_{\cdot j\cdot}}{n}$, $\hat{p}_{\cdot\cdot k}=\frac{n_{\cdot\cdot k}}{n}$ を得るが，今度は χ^2_{AB}, χ^2_{BC}, χ^2_{CA} が各ペヤーの間の独立性を検定する時に現われる（6）式の形で矢張り成分として残る．推定値を代入した時の χ^2_{ABC} は次の形になる．

$$\hat{\chi}^2_{ABC}=n^2\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\frac{1}{n_{i\cdot\cdot}n_{\cdot j\cdot}n_{\cdot\cdot k}}\left(n_{ijk}-\frac{n_{ij\cdot}n_{\cdot\cdot k}}{n}-\frac{n_{\cdot jk}n_{i\cdot\cdot}}{n}-\frac{n_{i\cdot k}n_{\cdot j\cdot}}{n}+2\frac{n_{i\cdot\cdot}n_{\cdot j\cdot}n_{\cdot\cdot k}}{n^2}\right)^2 \tag{11}$$

全体としての $\hat{\chi}^2_T$ が有意でなければ各成分とも有意でないであろうが（絶対的ではない！），若し χ^2_T が有意に大きい時には，その原因には種々の場合が考えられる．例えば $\hat{\chi}^2_{AB}$ 丈が大きい．即ち A, B の独立性からの距りが大きいのが原因となつて居て，C は A, B と独立であるといつた場合が起り得る．$\hat{\chi}^2_{AB}$, $\hat{\chi}^2_{BC}$, $\hat{\chi}^2_{CA}$ は偶然変動の範囲内にあるが．$\hat{\chi}^2_{ABC}$ が有意に大きい為に $\hat{\chi}^2_T$ を大きくして居たという場合も考えられる．従つて属性の間の関聯度は $\hat{\chi}^2_T/n$ で表わすのでは漠然としたもので，3 属性の間の関聯度であるのか，2 属性の間の関聯であるのかはつきりさせる必要がある．3 属性としての関聯度は

$$f^2=\frac{\hat{\chi}^2_{ABC}}{n}=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\frac{1}{\frac{n_{i\cdot\cdot}}{n}\frac{n_{\cdot j\cdot}}{n}\frac{n_{\cdot\cdot k}}{n}}\left(\frac{n_{ijk}}{n}-\frac{n_{ij\cdot}}{n}\frac{n_{\cdot\cdot k}}{n}-\frac{n_{\cdot jk}}{n}\frac{n_{i\cdot\cdot}}{n}-\frac{n_{i\cdot k}}{n}\frac{n_{\cdot j\cdot}}{n}+2\frac{n_{i\cdot\cdot}}{n}\frac{n_{\cdot j\cdot}}{n}\frac{n_{\cdot\cdot k}}{n}\right)^2 \tag{12}$$

で推定されるべきであろう．

$p_{i\cdot\cdot}$, $p_{\cdot j\cdot}$, $p_{\cdot\cdot k}$ が予め指定されて居る場合には，2 属性の時と同様に，周辺の当てはめに依る喰い違いが2属性，3属性に対する成分に影響を及ぼし，仮令独立性の仮設が真であつても，此等の成分を有意に大きな値にしてしまう事が生ずる．総じて分割表を用いて独立性の仮設を検定するには，何よりも周辺に対する当てはめを出来る丈よくしておく事が肝要であると言える．

（5）に於ける分割に於いて例えば（BとC），（CとA）の独立性が予めわかつて居る場合には，此の部分を特に取出す必要はなく，分割を次の如く縮める事が出来る．

$$\begin{aligned}\chi_T^2&=\sum_{i=1}^{r}\frac{(n_{i\cdot\cdot}-np_{i\cdot\cdot})^2}{np_{i\cdot\cdot}}+\sum_{j=1}^{s}\frac{(n_{\cdot j\cdot}-np_{\cdot j\cdot})^2}{np_{\cdot j\cdot}}+\sum_{k=1}^{t}\frac{(n_{\cdot\cdot k}-np_{\cdot\cdot k})^2}{np_{\cdot\cdot k}}\\&+\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij\cdot}-n_{i\cdot\cdot}p_{\cdot j\cdot}-n_{\cdot j\cdot}p_{i\cdot\cdot}+np_{i\cdot\cdot}p_{\cdot j\cdot}p_{\cdot\cdot k})^2}{np_{i\cdot\cdot}p_{\cdot j\cdot}}\\&+\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}\frac{(n_{ijk}-n_{ij\cdot}p_{\cdot\cdot k}-n_{\cdot\cdot k}p_{i\cdot\cdot}p_{\cdot j\cdot}+np_{i\cdot\cdot}p_{\cdot j\cdot}p_{\cdot\cdot k})^2}{np_{i\cdot\cdot}p_{\cdot j\cdot}p_{\cdot\cdot k}}\\&=\chi^2_A+\chi^2_B+\chi^2_C+\chi^2_{AB}+\chi^2_{ABC}\end{aligned} \tag{13}$$

（統計数理研究所）

参　考　文　献

[1] LANCASTER, H. O. (1949), "The derivation and partition of χ^2 in certain discrete distributions" Biometrika, Vol. 36. pp. 117-129.
[2] FISHER. R. A. (1948), Statistical Methods for Research Workers, 10th. edn. Oliver and Boyd. Edinburgh. (pp. 101-111).
[3] 小川潤次郎 (1954), 近代数理統計学序説, 惠文堂, (pp. 286-289).
[4] CRAMÉR, H. (1946), Mathematical Methods of Statistics, Princeton University Press. (pp. 424 -450).