HONAM MATHEMATICAL JOURNAL
Volume 5, Number 1, September, 1983

The Effect of Dependence and the Convolution in a Contingency Table

By Ju-Sun Ahn

1. Introduction

In investigating the contingency table of two dependent discrete (non-negative) random variables X and Y, what is the effect of dependence?
Despite the statistical interest in this question, we have few papers on it.

PETTITT and SISKIND investigated the effect of within-sample dependence and between-sample independence on the variance of the Mann-Whitney-Wilcoxon statistic, given that the marginal distributions are identical in each sample. For basic results, they considered two independent X_i, $i=1, \cdots\cdots, n$ and Y_j, $j=1, \cdots\cdots, m$, where the $X's$ and $Y's$ have some dependence within samples, and X_i and Y_j are assumed to have the same marginal distribution.

Also, the effect of dependence on Chi-squared tests of fit that assume *i. i. d.* observations was studied when the data formed a stationary stochastic process by MOORE. He conducted the general form of asymptotic distribution theory under the null hypothesis.

SUZUKI proposed the concepts of dependence convolution and dependent outlier.

The purpose of this paper is to see the properties of a contingency table in the special case that the sample space classified according to two discretes X and Y can be divided into m disjoint subsample spaces $R_1, \cdots, R_m$, where X and Y are independent within each R_k.

We examine covariances and convolutions of the two random variables X, Y with such structure. We will think of the difference of the dependent convolutions and independent convolutions as the effect of dependence.

And then we find that the limiting distribution of the statistic concerned with the effect of dependence of cells is a Chi-squared distribution under the

null hypothesis and that the effect of dependence is invariant under redividing subset R_k's.

2. Contingency tables

Suppose that the n individuals of a sample space R are classified according to two variable arguments X, Y in a tow-way table.

A table of this kind is known as a contingency table, and it is often required to test the hypothesis that the two variable arguments are independent.

It has been known to apply the Chi-square test to this problem.

In order to try to approach other examinations of the contingency table two assumptions are taken as follows :

(i) The X and Y are two discrete (non-negative) random variables with the marginal probability function f and g within R, respectively.

(ii) The R is divided into m disjoint subsets $R_1, \cdots\cdots, R_m$ so that the X and Y are independent within each R_k where $P_r(t\in R_k)=\lambda_k>0, \sum_{k=1}^{m}\lambda_k=1$.

Under the above assumptions, we make m sub-contingency tables with the two-independent discrete variables X and Y.

Denote by r_{kpq} $(k=1, \cdots, m,\ p=1, \cdots, r,\ q=1, 2, \cdots, s.)$ the relative frequency which belongs to the subset R_k the pth row and the qth column of that table and by f_k and g_k the marginal probability function of X and Y within the R_k, respectively. Then, we write

$$r_{kpq}=f_k(X=x_p)\cdot g_k(Y=y_q)$$

$$f_k(x_p)=P_r(X=x_p \mid t\in R_k),\quad \mu_k=\sum_{p=1}^{r}x_p f_k(x_p)$$

and $g_k(y_q)=P_r(Y=y_q \mid t\in R_k),\quad \nu_k=\sum_{q=1}^{s}y_q g_k(y_q)$

where the μ_k and ν_k are the average of X and Y within R_k, respectively.

It is easily seen that

$$f(x_p)=\sum_{k=1}^{m}\lambda_k f_k(x_p) \qquad \mu=\sum_{k=1}^{m}\lambda_k\mu_k$$

and $$g(y_q)=\sum_{k=1}^{m}\lambda_k g_k(y_q) \qquad \nu=\sum_{k=1}^{m}\lambda_k\nu_k$$

where the μ and ν are the mean of X and Y within R, respectively.

From the contingency table with the above properties, following results are obtained.

3. Distribution of $X+Y$

Theorem 3.1 : The covariance of X and Y is given by

$$\operatorname{cov}(X, Y) = \tfrac{1}{2}\underset{\sim}{\lambda}' M\underset{\sim}{\lambda}$$

where $\underset{\sim}{\lambda} = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_m \end{pmatrix}$, $\sum_{k=1}^{m}\lambda_k = 1$ and $M = \begin{pmatrix} c_{11}\ c_{12}\cdots\cdots c_{1m} \\ \vdots \\ c_{m1}\ c_{m2}\cdots\cdots c_{mm} \end{pmatrix}$

$$c_{ij} = (\mu_i - \mu_j)\ (\nu_i - \nu_j)$$

Proof : The Section 2 enables to write $E(XY) = \sum_{k=1}^{m}\lambda_k\mu_k\nu_k$, $E(X) = \sum_{k=1}^{m}\lambda_k\mu_k$ and $E(Y) = \sum_{k=1}^{m}\lambda_k\nu_k$ (see Appendix).

Therefore, combining these three formulas, we see that

$$\begin{aligned}\operatorname{cov}(X, Y) &= E(XY) - E(X)E(Y) \\ &= \sum_{k=1}^{m}\lambda_k(1-\lambda_k)\mu_k\nu_k - \sum_{\substack{i=1 \\ i\neq j}}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j\mu_i\nu_j\end{aligned}$$

Since $\sum_{k=1}^{m}\lambda_k = 1$,

$$\begin{aligned}\operatorname{cov}(X, Y) &= \sum_{\substack{i=1 \\ i\neq j}}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j\mu_i\ \nu_i - \sum_{\substack{i=1 \\ i\neq j}}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j\mu_i\nu_j \\ &= \tfrac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j(\mu_i - \mu_j)(\nu_i - \nu_j)\end{aligned}$$

Hence, $\operatorname{cov}(X, Y) = \frac{1}{2}\underset{\sim}{\lambda}'\ M\underset{\sim}{\lambda}$. Q. E. D.

From the above theorem 3.1, we see that X and Y are uncorrelated if and only if $\mu_i = \mu_j$ and/or $\nu_i = \nu_j$ for all i, j. But it does not mean that X and Y are independent.

Theorem 3.2 : If X and Y are the two discrete random variables which are given by Section (2) and $Z = X + Y$, then

$$h_z(z) = h_{x+y}(z) = f\circledast g(z) = f * g(z) + \tfrac{1}{2}\underset{\sim}{\lambda}' D(z)\underset{\sim}{\lambda}$$

where $f\circledast g$ and $f * g$ are convolutions of two random variables which are dependent and independent, respectively, and

$$D(z) = \begin{pmatrix} d_{11}(z)\cdots\cdots d_{1m}(z) \\ \vdots \qquad\qquad \vdots \\ d_{m1}(z)\cdots\cdots d_{mm}(z) \end{pmatrix}, \quad d_{ij}(z) = d_{ji}(z) = (f_i - f_j) * (g_i - g_j)(z)$$

Proof : Note that $f\circledast g(z) = \sum_{x+y=z}\sum_{k=1}^{m}\lambda_k f_k(x)\, g_k(y)$ where $\sum_{k=1}^{m}\lambda_k = 1$

and $(f_i - f_j) * (g_i - g_j)(z) = \sum_{x+y=z}(f_i(x) - f_j(x))(g_i(y) - g_j(y))$.

Thus, we see that

$$\begin{aligned}h_z(z) = f\circledast g(z) &= \sum_{x+y=z}\sum_{k=1}^{m}\lambda_k(\lambda_1 + \cdots + \lambda_m) f_k(x)\, g_k(y) \\ &= \sum_{x+y=z}\{\sum_{k=1}^{m}\lambda_k f_k(x)\sum_{k=1}^{m}\lambda_k g_k(y) - \sum_{\substack{i=1 \\ i\neq j}}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j f_i(x)\, g_j(y)\end{aligned}$$

$$+\sum_{i=1}^{m}\sum_{\substack{j=1\\ i\neq j}}^{m}\lambda_i\lambda_j f_i(x)\ g_i(y)$$

$$=\sum_{x+Y=z}\{f(x)\,g(y)+\tfrac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j(f_i(x)-f_j(x))\,(g_i(y)-g_j(y))\}.$$

Hence, $h_z(z)=f*g(z)+\frac{1}{2}\underset{\sim}{\lambda}'\ D(z)\underset{\sim}{\lambda}$. Q. E. D.

From the above theorem (3.2), it is seen that X and Y are independent if and only if $f_k(x)$'s and/or $g_k(y)$'s are same for all k.

If the X and Y are dependent, the second term $\frac{1}{2}\underset{\sim}{\lambda}'\mathrm{D}(z)\underset{\sim}{\lambda}$ (denote it by ED) does not vanish.

Thus it may be considered that the *ED* is generated by dependency but the *ED* gives no direct information about the degree of dependence between X and Y. Therefore, we call the *ED* the effect of dependence.

Now, denote $\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j(f_i(x_p)-f_j(x_p))\,(g_i(y_q)-g_j(y_q))$

which is the *ED* of a (p,q) cell by $ED(p,q)$.

By theorem (3.2), since $ED(p,q)=\sum_{k=1}^{m}\lambda_k f_k(x_p)\,g(y_q)-f(x_p)\,g(y_p)=r_{pq}-f(x_p)\,g(y_q)$,

$\dfrac{n\sum_{p=1}^{r}\sum_{q=1}^{s}\{ED(p,q)\}^2}{f(x_p)\,g(y_q)}$ is for large n approximately distributed in a Chi-squared distribution with $(r-1)(s-1)$ degree of freedom under the null hypothesis. (see (2))

Furthermore, it is easily seen that the *ED* is invariant under redividing R_k's into R_m's where X and Y are independent within R_m.

〈**Appendix**〉

(i) $f(x_p)=\sum_t P_r(X=x_p,\ t\in R_k)=\sum_t P_r(X=x_p\mid t\in R_k)\,P_r(t\in R_k)=\sum_{k=1}^{m}\lambda_k f_k(x_p)$

(ii) $\mu\quad=E(X)=\sum_{p=1}^{r}x_p f(x_p)=\sum_{p=1}^{r}x_p\sum_{k=1}^{m}\lambda_k f_k(x_p)=\sum_{k=1}^{m}\lambda_k\mu_k$

(iii) $E(XY)=\sum_u uP_r(XY=u)=\sum_u u\ \sum_t P_r(XY=u,\ t\in R_k)$

$=\sum_t\sum_u u\ P_r(XY=u\mid t\in R_k)\,P_r(t\in R_k)=\sum_t E(XY\mid t\in R_k)\,P_r(t\in R_k)$

$=\sum_{k=1}^{m}\lambda_k E(X\mid t\in R_k)\ E(Y\mid t\in R_k)=\sum_{k=1}^{m}\lambda_k\mu_k\nu_k.$

References

[1] Aitkin M. (1979) : *A simultaneous test procedure for contingency table models*, Appl. Statist. 28 No. 3 233-242.

[2] Cramer H. (1974) : *Mathematical methods of statistics*, Princeton.

[3] Lagakos S.W. and Reid N. (1981) : *Estimating convolutions from partially censord data*, Biometrika 68, 1. 437-41

[4] Moore D.S. (1982) : *the effect of dependence on chi-squared tests of fit*, The annals of statistics Vol. 10, No. 4, 1163-1171

[5] Pettitt A N. and Siskind V. (1981) : *Effect of within sample dependence on the Mann-Whitney-Wilcoxon statistic*, Biometrika 68, 2, 437-41

[6] Suzuki G. (1982) : *On some concepts of a clependent convolution and a dependent outlier*, pacific Area Statistical Conference