International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Correlation Coefficient between Two types of Children in a Family

Talawar A. S.[1], Sarvade R. M.[2]

[1, 2]Department of Statistics, Karnataka University, Dharwad, Karnataka State, India

Abstract: ***Results show that both the modified negative binomial and negative binomial distribution give good fit. Comparing with Chi-square values, the family size fits better with a modified negative binomial distribution than with a negative binomial distribution. The observed correlation co-efficient between the two types of children is 0.2403 and the estimated correlation co-efficient between boys and girls , in case of negative binomial distribution is 0.216, when p≠q and 0.2164, when p=q=1/2 and in case of modified negative binomial distribution is 0.2244, when p≠q and 0.2249, when p=q=1/2.***

Keywords: probability model, truncated model, family size, correlation co-efficient, Chi-square test

1. Introduction

Statisticians, demographers and other scientists have been concerned about the family size and the distribution of boys and girls within families of different sizes and also the correlation between the two types of children. Rao et al. (1973) considered family size as a random variable following negative binomial distribution. They developed the more general expression for the correlation between two types of children in a family. Hamdan (1975) obtained the correlation between the two types of children in a family using truncated negative binomial distribution excluding the childless families. Gupta (1976) obtained the correlation co-efficient for the modified power series distribution (MPSD). Rao (1979) obtained the correlation between the numbers of two types of children in a family using modified power series distribution. Janardan (1982) obtained the correlation between the numbers of two types of children in a family using the Markov-polya model. In the present paper we have correlation co-efficient for some of the probability models. The purpose of this paper is to give a good fit to the distribution of births considering negative binomial distribution, modified negative binomial and truncated negative binomial and check their adequacy to the data and also to obtain the correlation coefficient between the two types of children in a family using the above distribution. For this purpose we have used cross-sectional data collected from randomly chosen 470 families (with completed fertility), out of which 32 are childless families.

2. Distribution of a Family Size Considering the Families with Childless Families

Let the family size N, be a random variable formed from the two types of children boys (B) and girls (G), that is N=B+G.

(a) Negative Binomial Distribution

Let the family size N has a negative binomial distribution with the probability function

$$P(N=k)=\binom{r+k-1}{k}\theta^r(1-\theta)^k, k=0,1,2,\ldots \quad (2.1)$$

Where r>0 and 0<θ<1. The mean and variance of the family size, N are given by

$$E(N)=\frac{r(1-\theta)}{\theta} \text{ and } V(N)=\frac{r(1-\theta)}{\theta^2}$$

The parameters are estimated using the method of maximum likelihood estimation.

(b) Modified Negative Binomial Distribution

Let the family size N has a modified negative binomial distribution with probability function

$$P(N=0) = P_0 = 1 - \alpha + \alpha\theta^r \quad (2.2)$$

$$P(N=k)=\alpha\binom{r+k-1}{k}\theta^r(1-\theta)^k, k=1,2,\ldots$$

Where 0<θ<1, 0<α<1 and α is proportion of fecundability (Sharma, 1995). The mean and variance of the family size N are given by

$$E(N)=\frac{\alpha r(1-\theta)}{\theta} \text{ and } V(N)=\frac{\alpha r(1-\theta)}{\theta^2}[1+(1-\alpha)r(1-\theta)]$$

The parameters are estimated using the method of maximum likelihood estimation.

2.1 Distribution of a family size without childless families

Here the childless families are excluded from data, because the childless families do not contribute to the distribution of births and hence are excluded.

(a) Zero-truncated Negative Binomial Distribution (TNBD)

Assuming that the family size N has a zero-truncated negative binomial distribution with probability function

$$P[N=k]=\frac{(r+k-1)!}{(r-1)!k!}\frac{\theta^r(1-\theta)^k}{1-\theta^r}, k=1,2,3 \ldots.(3.1)$$

The first two non-central moments of the TNBD are

$$\mu_1' = E(N) = \frac{\alpha r(1-\theta)}{\theta(1-\theta^r)}$$

$$\mu_2' = E(N^2) = \frac{r(1-\theta)(1+r(1-\theta))}{\theta^2(1-\theta^r)}$$

And therefore the variance of the family size is thus obtained

$$V(N) = \mu_2' - \mu_1'^2$$

$$= \frac{r(1-\theta)}{\theta^2(1-\theta^r)^2}\left(1-\theta^r\left(1+r(1-\theta)\right)\right)$$

The parameters θ and r of (3.1) are estimated using the

methodology provided by Brass (1958a and 1958b) and Sampford (1955). We use the methodology provided by Sampford (1955) . Therefore

We have a simplified equation

$$\varphi(w) = m \exp\left\{-\left(m + \frac{s^2}{m} - 1\right)\phi(w)\right\} + \frac{s^2}{m}w = 1$$

Taking the initial values for

$$w = \frac{1}{s^2/m}$$

$$\varphi'(w) = \frac{s^2}{m} - m\left(m + \frac{s^2}{m} - 1\right)\left[\frac{\varphi(w) - w}{w(1-w)}\right]\exp\left\{-\left(m + \frac{s^2}{m} - 1\right)\phi(w)\right\}$$

Thus from Newton-Rapson method, for the i^{th} iteration

$$w_i = w_{i-1} - \left.\frac{\varphi(w)}{\varphi'(w)}\right|_{w=w_{i-1}}$$

(b) Logarithmic Series Distribution

Let the random variable N, the family size, has a logarithmic series distribution with probability function

$$P(N=k) = \frac{\alpha\theta^k}{\theta}, k = 1,2,3,\ldots, 0 < \theta < 1 \quad (3.2)$$

where $\alpha = -[\log(1-\theta)]^{-1}$

The individual probabilities are the terms in the series expansion of $-\alpha\log(1-\theta)$. The mean and variance of N are given by

$$E(N) = \frac{\alpha\theta}{1-\theta} \text{ and } V(N) = \frac{\alpha\theta(1-\alpha\theta)}{(1-\theta)^2}$$

The parameters are obtained using the likelihood equation

$$\bar{X} = \frac{\alpha\theta}{(1-\theta)}$$

and the table provided by Patil (1962).

(c) Beta-Geometric Distribution:

If we assume that the family size N has a beta-geometric distribution with the following form of probability function

$$P(N = k) = \mu\frac{\prod_{i=1}^{k-1}[1-\mu+(i-1)\theta]}{\prod_{i=1}^{k}[1+(i-1)\theta]}, k = 1, 2, \ldots \quad (3.3)$$

The mean and variance of model (3.3) are given by

$$E[N] = \left[\frac{1-\theta}{\mu-\theta}\right] \text{ and } Var[N] = \frac{\mu(1-\mu)(1-\theta)}{(\mu-\theta)^2(\mu-2\theta)}$$

The parameters μ and θ are estimated by maximum likelihood method (Weinberg and Gladen , 1986).

2. General formula for correlation coefficient:

Let the non-negative integer-valued random variable N denotes for the size of a family and follows any distribution with mean, E(N) and variance, V(N) both finite. Let. We assume that, N=B+G, where B is the number of boys and G is the number of girls in a family for any given size, B is a binomial variable (N, p) where p is the probability of a child being a boy and q=1-p is probability of a child being a girl.

The general formula the correlation coefficient between the two types of children B and G due to Rao et al. (1973), is given by

$$\rho_{BG} = \frac{(pq)^{\frac{1}{2}}[V(N) - E(N)]}{[pV(N)+qE(N)]^{\frac{1}{2}}[qV(N)+pE(N)]^{\frac{1}{2}}}, \text{if } p \neq q$$

$$= \frac{V(N) - E(N)}{V(N)+E(N)}, \text{if } p = q = \frac{1}{2} \quad (4.1)$$

Where p and q are respectively, the proportions of boys and girls in a family.

(a) Negative binomial distribution

Substituting mean and variance of N for the negative binomial distribution in (4.1) we get

$$\rho_{BG} = \frac{(pq)^{\frac{1}{2}}(1-\theta)}{[(p+q\theta)(q+p\theta)]^{\frac{1}{2}}}, \text{if } p \neq q$$

$$= \frac{1-\theta}{1+\theta}, \text{if } p = q = \frac{1}{2} \quad (4.2)$$

Note that the correlation coefficient between B and G (ρ_{BG}) is independent of the parameter r. We get the same ρ_{BG}, when the size of family, N has a geometric distribution.

That is, if N has a geometric distribution with probability function

$P(N=k) = \theta(1-\theta)^k$, k=0, 1, 2 . . .

Mean and variance of N are

$$E(N) = \frac{1-\theta}{\theta} \text{ and } V(N) = \frac{1-\theta}{\theta^2}$$

Therefore, the correlation coefficient between the two types of children, B and G is

$$\rho_{BG} = \frac{(pq)^{\frac{1}{2}}(1-\theta)}{\sqrt{(p+q\theta)(q+p\theta)}}, \text{if } p \neq q$$

$$= \frac{1-\theta}{1+\theta}, \text{if } p = q = \frac{1}{2}$$

(b) Modified Negative binomial distribution

Substituting the mean and variance of the modified negative binomial distribution in (4.1), we get

$$\rho_{BG} = \frac{(pq)^{\frac{1}{2}}\left[\frac{\alpha r(1-\theta)}{\theta^2}(1+(1-\alpha)r(1-\theta)) - \frac{\alpha r(1-\theta)}{\theta}\right]}{\left[\frac{p\alpha r(1-\theta)}{\theta^2}(1+(1-\alpha)r(1-\theta)) + \frac{q\alpha r(1-\theta)}{\theta}\right]^{\frac{1}{2}}\left[\frac{q\alpha r(1-\theta)}{\theta^2}(1+(1-\alpha)r(1-\theta)) + \frac{p\alpha r(1-\theta)}{\theta}\right]^{\frac{1}{2}}}$$

$$\rho_{BG} = \frac{(pq)^{\frac{1}{2}}\left[\frac{\alpha r(1-\theta)}{\theta^2}(1+(1-\alpha)r(1-\theta)) - \theta\right]}{\frac{\alpha r(1-\theta)}{\theta^2}[p(1+(1-\alpha)r(1-\theta)) + q\theta]^{\frac{1}{2}}[q(1+(1-\alpha)r(1-\theta)) + p\theta]^{\frac{1}{2}}}$$

$$\rho_{BG} = \frac{(pq)^{\frac{1}{2}}[1+(1-\alpha)r(1-\theta) - \theta]}{[p(1+(1-\alpha)r(1-\theta) + q\theta]^{\frac{1}{2}}[q(1+(1-\alpha)r(1-\theta)) + p\theta]^{\frac{1}{2}}},$$

$$= \frac{1+(1-\alpha)r(1-\theta)-\theta}{1+(1-\alpha)r(1-\theta)+\theta}, \text{if } p \neq q$$
$$= \frac{1}{2}, \text{if } p = q = \frac{1}{2} \quad (4.3)$$

(c) Truncated Negative binomial distribution:

The correlation coefficient between the two kinds of children, B and G is

$$\rho_{BG} = \frac{(pq)^{\frac{1}{2}}\left[1-\theta^r\left(1+r(1-\theta)\right)-\theta(1-\theta^r)\right]}{\left[p\left(1-\theta^r\left(1+r(1-\theta)\right)\right)+q\theta(1-\theta^r)\right]^{\frac{1}{2}}\left[q\left(1-\theta^r\left(1+r(1-\theta)\right)\right)+p\right]^{\frac{1}{2}}}$$

Put $1-\theta^r = A$

$$1-\theta^r\left(1+r(1-\theta)\right)-\theta(1-\theta^r) = 1-\theta^r-\theta^r r(1-\theta)-\theta(1-\theta^r)$$
$$= A-r\theta^r(1-\theta)-\theta A$$
$$= (1-\theta)A-r\theta^r(1-\theta)$$
$$= (1-\theta)(A-r\theta^r)$$

we get

$$\rho_{BG} = \frac{(pq)^{\frac{1}{2}}(1-\theta)(A-r\theta^r)}{[(p+q\theta)A-pr(1-\theta)\theta^r]^{\frac{1}{2}}[(q+p\theta)A-qr(1-\theta)\theta^r]^{\frac{1}{2}}}, p \neq q$$

Which is similar to Hamdan (1975) (equation (2.2)).

$$\rho_{BG} = \frac{(1-\theta)(A-r\theta^r)}{[(1+\theta)A-r(1-\theta)\theta^r]}, p=q=\frac{1}{2} \quad (4.4)$$

(d) Logarithmic series distribution:

The correlation coefficient between the two types of children is

$$\rho_{BG} = \frac{(pq)^{\frac{1}{2}}\left[\frac{\alpha\theta(1-\theta)}{(1-\theta)^2}-\frac{\alpha\theta}{(1-\theta)}\right]}{\left[\frac{p\alpha\theta(1-\alpha\theta)}{(1-\theta)^2}+\frac{q\alpha\theta}{(1-\theta)}\right]^{\frac{1}{2}}\left[\frac{q\alpha\theta(1-\alpha\theta)}{(1-\theta)^2}+\frac{p\alpha\theta}{(1-\theta)}\right]^{\frac{1}{2}}}$$
$$= \frac{(pq)^{\frac{1}{2}}\theta(1-\alpha)}{[p-p\alpha\theta+q-q\theta]^{\frac{1}{2}}[q-q\alpha\theta+p-p\theta]^{\frac{1}{2}}}$$
$$\rho_{BG} = \frac{(pq)^{\frac{1}{2}}\theta(1-\alpha)}{[1-\theta(p\alpha+q)]^{\frac{1}{2}}[1-\theta(q\alpha+p)]^{\frac{1}{2}}}, p \neq q, \text{ since } p+q=1$$
$$= \frac{\theta(1-\alpha)}{2-\theta(1+\alpha)}, p = q = \frac{1}{2} \quad (4.5)$$

(e) Beta-Geometric Distribution:

The correlation coefficient between the two types of children, B and G is

$$\rho_{BG} = \frac{(pq)^{\frac{1}{2}}\left[\frac{\mu(1-\mu)(1-\theta)}{(\mu-\theta)^2(\mu-2\theta)}-\frac{(1-\theta)}{(\mu-\theta)}\right]}{\left[\frac{p\mu(1-\mu)(1-\theta)}{(\mu-\theta)^2(\mu-2\theta)}+\frac{q(1-\theta)}{(\mu-\theta)}\right]^{\frac{1}{2}}\left[\frac{q\mu(1-\mu)(1-\theta)}{(\mu-\theta)^2(\mu-2\theta)}+\frac{p(1-\theta)}{(\mu-\theta)}\right]^{\frac{1}{2}}}$$
$$\rho_{BG} = \frac{(pq)^{\frac{1}{2}}[\mu(1-\mu)-(\mu-\theta)(\mu-2\theta)]}{[p\mu(1-\mu)+q(\mu-\theta)(\mu-2\theta)]^{\frac{1}{2}}[q\mu(1-\mu)+p(\mu-\theta)(\mu-2\theta)]^{\frac{1}{2}}}$$

if $p \neq q$

$$= \frac{\mu(1-\mu)-(\mu-\theta)(\mu-2\theta)}{\mu(1-\mu)+(\mu-\theta)(\mu-2\theta)}, \text{if } p = q = \frac{1}{2} \quad (4.6)$$

3. Application of the Models

The following **Table 1** gives the observed numbers of males and females offspring among the 470 families. Total number of boys is 899 and total number of girls is 785. Therefore the total number of children in 470 families is 1684. The proportion of males (p) is 0.534 and the proportion of females (q) is 0.466. Figures in parenthesis give the percentage values of particular combinations of males and females.

Fitting of Probability Models with zeroes and without zeroes

Empirical statistics shows that both the modified negative binomial and negative binomial distribution give good fit. Comparing with Chi-square values, the human family size fits better with a modified negative binomial distribution than with a negative binomial distribution (**Table 2**). The observed correlation co-efficient between the two types of children is 0.2403 and the estimated correlation co-efficient between B and G , in case of negative binomial distribution is 0.216, when p≠q and 0.2164, when p=q=1/2 and in case of modified negative binomial distribution is 0.2244, when p≠q and 0.2249, when p=q=1/2.

Table 1: Observed Numbers of Males and Females offspring among 470 Families

Number of male offspring	Number of female offspring									Total number of families
	0	1	2	3	4	5	6	7	8	
9				1						1
8										0
7			1	1						2
6		7	1	2	1	1				12
5	1	3	4	4	2	2				16
4	4	9	10	5	3	1				32
3	8	18	18	12	7	4	1	1	1	70
2	25	46	29	19	7	1	1	1		129
1	34	37	33	9	11	1	1	2		128
0	32	24	15	7	2					80
Total number of families	104	144	111	60	33	10	3	4	1	470

Table 2: Fitting of Probability Models

Size of family	Observed No. of families	Expected Number of Families		
		Negative Binomial $\hat{r}$=6.476 & $\hat{\theta}$=0.6442	Modified NBD $\hat{r}$=6.84 $\hat{\theta}$=0.654 & $\hat{\alpha}$=0.9858	TNBD $\hat{r}$=7.94 & $\hat{\theta}$=0.685
0	32	27	32	-
1	58	63	60	58
2	77	83	81	77
3	94	85	83	94
4	62	71	71	62
5	58	53	53	58
6	33	36	36	33
7	25	23	23	25
8	15	14	14	15
9	8	8	8	8
10	5	4	4	5
11	2	2	2	2
12	1	1	1	1
	470	470	470	438
		$\chi^2 = 4.96^{ns}$	$\chi^2 = 3.86^{ns}$	$\chi^2 = 3.78^{ns}$

^{ns} indicates significant at 5% level of significance

TNBD gives good fit to the size of a family when the childless families are excluded.

References

[1] **Brass, W. (1958a)** Simplified methods of fitting the truncated negative binomial distribution, *Biometrika,* 15, 237-250.

[2] **Brass, W. (1958b)** The distribution of births in human population, *population studies*, 12, 57-72.

[3] **Gupta R. C. (1976)** Application of Modified Power Series Distribution in Genetics, *Sankhya*: The Indian J. of Statistics, Series B, Vol. 38, No2, pp 187-191.

[4] **Hamdan, M. A. (1975)** Correlation between the numbers of two types of children when the family size distribution is zero-truncated negative binomial, *Biometrics* 31, 765-766.

[5] **Janardan, K.G., (1982)** Correlation between the numbers of two types of children in a family, using the

Markov-Polya Model, *Mathematical Biosciences*, 62(1), 123-136.

[6] **Patil, G.P. (1962)** Some methods of estimation for the logarithmic series distribution, *Biometrics*, 18, 68-75.

[7] **Rao, B. R., (1981)** Correlation between the numbers of two types of children in a family with the mpsd for the family size, *communications in statistics – Theory and methods*, 10(3), 1981.

[8] **Rao, B. R., Mazumdar, S., Walter, H.J., and Li, C.C., (1973)** Correlation between the numbers of two types of children in a family, *Biometrics,* 29, 271-279.

[9] **Sharma, H.L., (1995)** Estimation of parameters involved in the distribution of numbers of two types of children in a family, *Journal of ISPS*, Vol 2, 1-14.

Author Profile

Dr. A. S. Talawar is M.Sc., M.Phil., Ph.D. having experience of 19 years. Presently working as Associate Professor and has published 18 papers in national and international journal

R. M. Sarvade is Ph.D. Student who is about to submit his Thesis shortly. He has three papers published in his name