Demography India Vol. 26, No. 1
(1997), pp. 1-7

D. V. Gokhale* and Sudhakar Kunte**

Probability of a Male Child: Is it Constant Over the Population of Couples?

Introduction

ONE of the interesting questions while suggesting statistical models for human population is the following: Is it reasonable to assume that the probability of a male birth remains constant across the population of couples and also across the successive births for the same couple (see also Conclusions)? Suppose that this probability is a constant p (the population parameter) and that sex outcomes of children at successive birth are independent. Under these assumptions, X, the number of male children born to a couple with n children, would have a Binomial distribution (5-model) with parameters n and p. This hypothesis for the distribution of X is tested by Fisher (1948). He has reported data due to Geissler (1889), on 53,680 families with eight children classified according to the number of male children. These data being fairly old, it is reasonable to assume that they come from a natural population, unaffected by the present day family planning or pre-natal sex determination methods. Observed frequencies and the expected frequencies under the *B*-model are given in the columns (2) and (3) of Table 1 respectively. The χ^2-value for the best Binomial fit to the data is 91.869. For seven degrees of freedom this χ^2-value is quite high, suggesting that the Binomial model is not the right model for X. Fisher (1948: 66) attributes the lack of fit to "a tendency on the part of some parents to produce males or females" but does not propose a model to account for

* Department of Statistics, University of California, Riverside, California 92521, U.S.A.
** Department of Statistics, University of Pune, Pune-411 007, India.

it. He also discusses later (p. 68) the effect of multiple births, but concludes that only a fraction of the observed excess variation in the data can be reasonably ascribed to multiple births. In the present paper we formalize his earlier remark, given above, by proposing models which account for the variation in probability P. Further details about the methodology and the software packages used in the analysis may be obtained from the authors.

TABLE 1

No. of Boys	*Observed Frequencies*	*Expected Frequencies*		
		B Model	*BB Model*	*MBB Model*
	(2)	*(3)*	(4)	*(5)*
0	215	165	191	196
1	1485	1402	1513	1514
2	5331	5202	5335	5256
3	10649	11035	10961	10701
4	14959	14627	14348	14959
5	11929	12410	12253	12000
6	6678	6580	6667	6608
7	2092	1994	2114	2136
8	342	264	298	310
Total	53680	53680	53680	53680
χ^2-value		91.869	53.747	9.0914

Binomial *(B* Model): Estimated $\hat{P} = 5143$
^ Beta-Binomial *(BB* Model): Estimated x = 103.5, $\hat{\beta} = 97.7$
Mixed Beta Binomial (MBB Model): Estimated Parameters:

$\hat{\lambda} = .01824$, $\hat{\alpha} = 74.85$, $\hat{\beta} = 70.49$

The Beta-Binomial Model

Sex of a newborn baby depends upon whether the ovum is fertilized by the X-sperm or the Y-sperm. In deciding the sex of fetus, the quality and the quantity of the X and Y sperms in the semen is the male contribution, the environment through which the sperms have to travel before fertilization is the female contribution. Thus it seems reasonable to assume that p is the property of a couple and would remain constant over successive births to the same couple. However it seems possible that

***p* may** vary from couple to couple according to some probability distribution. Thus **we** propose the following model for *X:*

Theprobabilityp ofamale child varies overthe population of couples according to a Beta distribution with parameters *a* and *b.* Further for a given *p, X* follows the Binomial distribution with parameters *n* and *p.*

This is the Beta-Binomial model *(BB* model) for *X;*

$$P[X=x]= \frac{\Gamma(n+1)}{\Gamma(x+1)\,\Gamma(n-x+1)} \cdot \frac{\Gamma(x+a)\,\Gamma(n-x+b)}{\Gamma(n+a+b)}, \quad x = 0, 1, 2,..., n,$$

using the notation of Gamma functions.

Fitting the Beta-Binomial distribution to Geissler's data, estimated values of a and b are $\hat{a}$ = 103.5 and $\hat{b}$ = 97.7. These estimates are obtained by using method of moments. The likelihood surface is nearly flat around these estimates, so computation of maximum likelihood estimates is extremely difficult. The expected frequencies are given in column (4) of Table 1. The χ^2-value is 53.74. Although there is a substantial reduction in the χ^2-value from that for the Binomial model, this χ^2 value of 53.74 for six degrees of freedom is still quite high. One possible explanation for this high χ^2-value is that it is based on a very large sample of size 53,680. Even a slight departure from the assumed model would be significantly detected by such a large data. In this connection we refer to Kunte (1992) and Kunte and Gore (1992). The same model was also fitted to Geissler's data by Ishii and Hayakawa (1960). However they did not give any reasons as to why the model may be considered a suitable model for the data at hand.

MDI Analysis

We next consider the application of Minimum Discrimination Information (MDI) approach (Gokhale and Kullback, 1978) for these data, using the analysis of Internal Constraints Problems (ICP). Stated briefly, an ICP is a model building problem, in which a model as close to a reference distribution is constructed such that the moments of the model distribution coincide with the corresponding observed moments (see Gokhale and Kullback, 1978, for details). For the present data we take the Binomial distribution with $n = 8$ and $p = 0.5$ as the reference distribution as this is the simplest possible model. The first internal constraint used is that the new model estimate should have the same mean as the observed sample mean. Since the ICP gives rise to exponential families, it is clear that the model estimate so

obtained is the same as the Binomial model, with $n = 8$ and the estimated p being the observed proportion. The estimated frequencies given in Column (3) of Table 2 coincide with those in Column (3) of Table 1. Column (4) of Table 2 gives what are called "OUTLIER" values. An OUTLIER value is a lower bound to the contribution to the MDI statistics due to that cell (Gokhale and Kullback, 1978:
62). There are many OUTLIER values in Column (4), showing that the model needs to be improved by adding more internal constraints. We therefore, add the constraint that the second momentsofthe model estimate and the sample be equal. The model so obtained is called the "Quadratic Binomial *(QB)"* model. To fit such models, the computer programmes given in Gokhale and Kullback (1978: Chapter 6) may be used. A newer version of such a computer programme is available with the first author. Estimated frequencies and corresponding OUTLIER values under the *QB* model are given as Columns (2) and (3) of Table 3.

The value of the MDI statistic is 52.73, which is highly significant, indicating a poor fit. Since the sample size is extremely large, it is more appropriate to consider the ratio of the MDI statistic under the new model compared to that under the base model. Such a ratio is independent of the sample size. In our example, this ratio is 52.73/88.74 = 0.5942, indicating that only 40.58 per cent reduction has been achieved by incorporating the constraint on the second moment.

Looking at the OUTLIER values in Column (3) of Table 3, we see that a substantial departure from the *QB* model is due to the cell corresponding to $X = 4$. Treating this cell as "unusual" we wish to re-estimate the *QB* model by fitting it to the remaining cells. This is achieved by adding one more constraint to the QB model in which the estimated and the observed cell frequencies *forX*= 4 are set equal. We call this the "Modified Quadratic Binomial *(MQB)"* model.

Estimated frequencies and OUTLIER values under *MQB* model are given in Columns (4) and (5) Table 3. It is seen that (a) the MDI statistic 7.7057 is not significant, (b) the A/g5 model gives a 100 x(88.74-7.7057)/88.74=91.31 percent reduction of the "disparity" between the base model and the data (this index is free of the sample size) and (c) the OUTLIER values in Column (5) are quite small. Incidentally, if the OUTLIER analysis of the *QB* model is ignored and a third constraint, setting equal the third moments of the model and the data, is used the MDI statistic under this new "Cubic Binomial" model comes outto be 52.3 0, showing very little reduction in the MDI statistic due the additional constraint on the third moment over the *QB* model. This further confirms our selection of the *MQB* model. The estimated *MQB* model is given by

$$\ln \frac{p(x)}{\pi(x)} = -0.0398 - 0.0640\,x + 0.015 \setminus x^2 + 0.07321_{[x=4]'}$$

where $p(x) = P[X = x]$, $\pi(x) = \binom{8}{x}(0.5)^8$ and $I_{[X = 4]}$ Equals one if X = 4 and equals zero otherwise.

TABLE 2

x	*Observed Frequency*	*Expected Frequency (SB Model)*	*Outlier Values*
	(2)	*(3)*	*(4)*
0	215	165.22	13.73
1	1485	1401.69	4.99
2	5331	5202.65	3.48
3	10649	11034.65	17.16
4	14959	14627.60	10.27
5	11929	12409.87	24.46
6	6678	6580.24	1.65
7	2092	1993.78	4.95
8	342	264.30	21.00
MDI Statistic		88.74	

TABLE 3

X	*Expected Freq. (QB Model)*	*Outlier Values*	*Expected Freq. (MQB Model)*	*Outlier Values*
(2)		(3)	*(4)*	*(5)*
0	190.46	3.04	201.50	0.88
1	1507.77	0.35	1535.06	1.70
2	5323.36	0.01	5273.13	0.70
3	10948.26	10.34	10668.04	0.04
4	14346.05	35.42	14959.00	0.00
5	12264.42	11.97	11950.64	0.05
6	6680.20	0.00	6617.32	0.63
7	2119.54	0.37	2157.98	2.12
8	299.93	5.67	317.32	1.88
MDI Statistic	52.73		7.7057	

The MDI analysis done up to this point has indicated that the cell $X = 4$ is unusual in the sense that it does not conform to the *QB* model. The estimated cell frequency underthe *QB* model is 14346.05 while the observed frequency is 14959.00 showing gross under-estimation. (It can also be interpreted as an over-reporting in the data, however, we have assumed that the data are correctly reported.) Note that the *MQB* model lacks the interpretation of the Beta-Binomial model in which the underlying probability of a male child is assumed to have the Beta distribution. Hence we use the OUTLIER value corresponding to $X = 4$ in the MDI analysis to modify the Beta-Binomial model as follows.

Final Model

We assume that the distribution of *X* the number of male children of a couple, is a *mixture* of a degenerate distribution at value 4 (with probability X) and a Beta-Binomial distribution (with probability 1 - *K).* This is called the "Mixed Beta-Binomial *(MBB)"* model.. It must be pointed out that the *MBB* model should be interpreted only as a device to adjust for the underestimation of the frequency at $X = 4$ in the usual Beta-Binomial model. In other words, it is hard to think of and justify a proportion *'k* of couples who are *destined to* have exactly four children. Parameters of this model i.e. \, *a* and *b* are estimated by using the minimum x^2 method. The estimators are difficult to obtain, in a closed form. However, good approximations can be obtained by putting a grid on values *of a, b* and adjusting λ *so* that the expected frequency *of* $X = 4$ matches with the observed frequency. This can be easily done on a computer. The optimal values of λ, *a* and *b* are given by $\hat{\lambda}$, = .01824, $\hat{a}$ = 74.85 and $\hat{b}$ = 70.49. The corresponding expected frequencies are given in column (5) of Table 1 and the observed χ^2-value is 9.0914. For five degrees of freedom, this observed value of$\hat{}^2$ is small enough to accept the goodness of fit of the model even at the conventional 5% level of significance.

Conclusions

The primary thesis of this paper is that the probability/? of a male child, which is an intrinsic characteristic of a couple, should be assumed to vary when data are aggregated over couples. To this end, the present paper examines the fit of the Beta-Binomial distribution to Geissler's data, as reported by Fisher (1948), and shows that except for the rather anomalous behaviour for $X = 4$, the Beta-Binomial fits quite well. The paper also shows how, in general, the MDI analysis can be used to supplement fitting of standard models to such data, using OUTLIER analysis.

The paper also makes a contribution towards consideration of a policy on family planning. In modem times, the probability/? of a male child can not only be thought of as an intrinsic quantity but also as a *controllable* quantity. For example, it is well known that for Asian populations in particular, probability *p* is affected by tec desirability of a male child. It may also be interesting to examine the effect, if any, on;?, due to the usage of contraceptives. As far as the authors know, there are no data to assess the effects of such "assignable causes" of variation in *p.* When such data become available, an assessment of the variation in *p,* (which could be measured by the estimated parameters of die Beta distribution, for example), would lead to implementation of different policies of family planning for different groups, such as countries, users or non-users of contraceptives, etc. The policy in each group may be based on the intrinsic "male or female proneness" for a child bearing couple.

References

Fisher, R. A., 1948, *Statistical Methods for Research Workers* (10th edn.). Hafner, New York. Geisslcr, 1889, Geitrage zur Frage des Geschlechts Verhaltnisses der Geborenen. *Zeitschrift des K. Sachsischen Statistischen Bureaus.* **Ookhale, D.** V. and Kullback, S., 1978, *The Information in Contingency Tables.* Marcel Dekker, New York.

Ishii, G. and Hayakawa, 1960, On compound binomial distribution. *Ann. Inst. Stal. Math,* 12 69-80.

Kunte, S.. 1992, Jeffreys-Lindley Paradox and a Related Problem. *Bayesian Analysis in Statistics and Econometrics. Lecture Notes in Statistics,* 75. Springer Verlag, New York.

Kunte, S. and Gore, A. P., 1992, The paradox of large samples. *Current Science,* 62: 393-395.