Journal of Data Science **1**(2003), 261-274

Estimation of the Proportion of Sterile Couples Using the Negative Binomial Distribution

Mohammad Fraiwan Al-Saleh and Fatima Khalid AL-Batainah
Yarmouk Unuversity

Abstract: A Sterile family is a couple who has no children by their deliberate choice or because they are biologically infertile. Couples who are childless by chance are not considered to be sterile. The object is to estimate the proportion of sterile couples in Jordan indirectly based on the 1994 population census, by separating the two types of childless couples into sterile and fertile couples. Three methods of fitting a negative binomial distribution to the completed family size data obtained from 1994-population census are investigated. It appeared that the third method gives the best fit. Based on the fitted distribution, the proportion of sterile couples is estimated at 6.1% of all couples. This estimate is much lower than the corresponding estimate of sterile couples in the USA, which was estimated at 11%. The difference between the two can be due to some socio-cultural factors influencing the deliberate choice of couples to have no children. The method of estimation can be applied on other populations.

Key words: Family Size, fertile Couple, negative binomial distribution sterile couple.

1. Introduction

The distribution of completed family size (or sipship size) has been a subject of interest for human biologists, geneticists, demographers, and social scientists. Since the variance of the family size distribution is much larger than its mean, a Poisson distribution is unlikely to fit. However, there is a good empirical evidence that the distribution is nearly that of a

negative binomial (Kojima and Kelleher, 1962). Waller *et al.* (1973) in their paper entitled Heterogeneity of Childless Families noticed that, the number of childless families is much greater than the expected number of childless families when they fit the negative binomial distribution to the observed frequencies of completed family size from various sources, although the fit is good for the rest of the distribution. This led them to suggest that the childless family is a mixture of two types of families. The first type is biologically fertile and could have children, but by chance, didn't. This type of families should be a part of the general negative binomial distribution of family size. The second type is either biologically or electively not fertile (sterile) and thus has no children. This should not be a part of the general negative binomial distribution of family size. The proportion of this type of families is expected to vary among populations studied, due to sociocultural factors influencing the deliberate choice to have no children. They discussed the theoretical considerations that justify the use of a negative binomial distribution. One of these considerations is that a birth process leads to a negative binomial distribution. The negative binomial random variable X is a non-negative discrete random variable with probability function:

$$f(x,p,k) = \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k+1)} p^k q^x, \qquad x = 0,1,2,\cdots \tag{1}$$

where $q = 1-p, p \in (0,1), k > 0$.

The mean of X is $\mu = kq/p$ and the variance is $\sigma^2 = kq/p^2$. Note that the variance is always greater than the mean. Of particular interest to the problem under consideration is the first term of the distribution. For childless fertile families, $X = 0$ and $Pr[X = 0] = p_0 = p^k$, which is the theoretical proportion of childless fertile couples. Hence to estimate this proportion we need to obtain estimates for p and k.

This distribution has been found to provide useful representation in many fields. Its applicability in birth and death process has been shown by Furry (1937) and Kendall (1949). It was used to model family size by Rao *et al.* (1973). Wilson *et al.* (1983) and Binns (1986) have used it for modeling entomological data. Kault (1996) have used this distribution for modeling the number of sexual partners. This distribution has been used extensively on biological data and consequently, there has been some attempt

to give an ecological meaning to the mean μ and the shape parameter k. The mean μ has been thought of as the density of organisms in the area of interest, because an increase in μ results when the population increases in size or become more dense, see Wilson *et al.* (1984). Anscombe (1949), noted that there is a theoretical evidence that k depends on the intrinsic power of a species to reproduce it self and Waters (1959) suggested that k is a measure of aggregation. With this in mind, three methods of fitting this distribution, to the family size data obtained from the general census of the Jordan population, (1994), are employed. The best fitted distribution is used to estimate (or approximate) the proportion of sterile couples in Jordan population. To the best of our knowledge, there has not been any attempt to estimate this quantity directly; this could be due to the sensitivity of the issue.

2. Methods of fitting

Three methods of fitting the negative binomial distribution to the observed data have been employed (see Waller *et al.*, 1973).

Method I: (Complete)

This method consists of approximating the mean (μ) and the variance (σ^2) of the negative binomial distribution directly from the observed data, and the parameters p, k are estimated using the formulas:

$$p = \frac{\mu}{\sigma^2}, \quad k = \frac{\mu^2}{\sigma^2 - \mu} \tag{2}$$

Method II: (truncated)

In this method the zero class is considered as missing and the parametersx k, p are estimated on the basis of the incomplete (truncated) distribution. Let

$$T_j = \sum_{x=0}^{\infty} x^j f_x$$

where f_x is the frequency of families of size x. All summations range from $x = 0$ to the maximum class value (family size). The estimates of the parameters for the negative binomial distribution are:

$$\hat{k} = \frac{2T_2^2 - T_1T_2 - T_1T_3}{T_1T_3 - T_1T_2 + T_1^2 - T_2^2} \quad (3)$$

$$\hat{p} = \frac{T_1T_2 - T_1^2}{T_1T_3 - T_2^2} \quad (4)$$

See Rider (1955).

Method III: (iteration)

This method consists of iterating method from the sets of initial trial values p, k obtained from methods I or II. T_0 is the total number of families in the data set with at least one child. In iterating from either set of values, the calculated new total number for the distribution is $\tilde{N} = T_0/(1 - \hat{p}^k)$, and the estimated number for the zero class is $\tilde{N}_0 = \hat{p}^k\tilde{N}$. Then we replace the observed total number with the calculated values of and calculate new values for p and k by using the formulas in Method I. The end point of this iteration process is reached when successive values of $\tilde{N}$ and $\tilde{N}_0$ do not change widely. See Waller *et al.* (1973).

3. Population of Jordan

Based on the general census of Jordan (1994), the observed family size distribution is shown in Table (1). Where x is the sipship size (number of children in the family) and f_x is the number of families which have x children (frequency). Now, we will apply the previous methods of fitting on this population.

Method I

Using Table 1, and using $x = 13$ for all families with size 13+, we find that $\mu = 4.32$, and $\sigma^2 = 9.19$. Hence, using formula (1) we have

$\hat{p} = 0.4703, \hat{k} = 3.837$. Then, the fitted negative binomial distribution is

$$f(x) = \frac{\Gamma(x + 3.837)}{\Gamma(x+1)\Gamma(3.847)}(0.4703)^{3.937}(0.5297)^{x}, \quad x = 0, 1, \cdots, 13 \qquad (5)$$

Table 1 contains the fitted family size distribution based on Method I. Column 1 contains the family size (x). Column 2 contains the observed number families f_x for each value of x, column 3 contains the relative frequency of each value of x, column 4 contains the theoretical probability of each value of x based on the negative binomial distribution with $p = 0.4703$ and $k = 3.837$. The last column contains the fitted family size.

Table 1: Family Size Distribution of Jordan population by Method I

Family size x	Observed number of family f_x	Relative frequency Y	$p_x =$ NB fitted probability	Expected number of families
0	59979	0.09419	0.05534	35240.3
1	64047	0.10058	0.11247	71620.5
2	78838	0.1238	0.14407	91743.2
3	82384	0.12937	0.14847	94545.1
4	77575	0.12182	0.13441	85593.7
5	68431	0.10746	0.11159	71060.1
6	57795	0.09076	0.08705	55433.1
7	46127	0.07244	0.06479	41258.0
8	35983	0.05065	0.04649	29604.7
9	25766	0.04046	0.03239	20623.3
10	18410	0.02891	0.02202	14022.3
11	8303	0.01304	0.01467	9341.8
12	5315	0.00835	0.00961	6119.6
13+	7843	0.01232	0.01663	10590.3
N	636796	1	1	636796

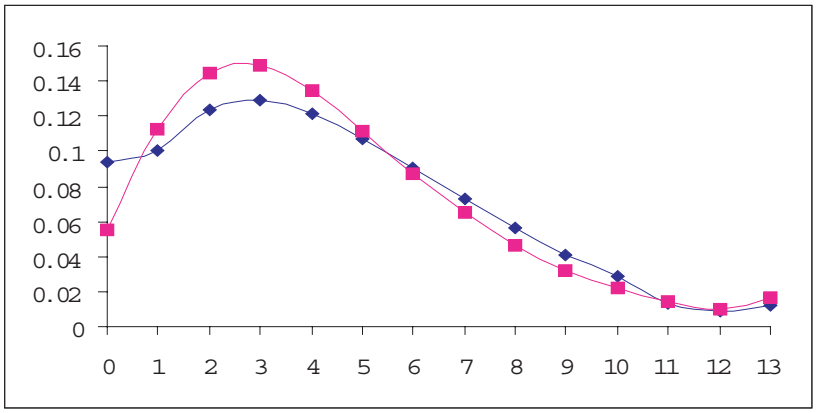

Figure 1: Observed and fitted curve of family size using method I, bold = observed; smokey = fitted

It can be seen from this table and Figure 1 that the fitted curve and the actual curve have a similar general shape, but with a large gap at $x = 0$. Furthermore, both curves are skewed to the right with $x = 3$ being the mode as well as the median of both distribution. The large gap between the two curves at $x = 0$ suggests that there is a heterogeneity among the families of this class; some may have no children subject to chance alone while some are biologically or deliberately sterile and hence should not be part of the distribution. In method II and III below, this second type of families is isolated.

Method II

Here we deal with the zero class as missing values. We use Rider Method (1955) to estimate the parameters on the basis of the incomplete distribution. Let

$$T_i = \sum_{x=1}^{13} x^i f_x, \quad i = 0, 1, 2, 3$$

Then we have

$$T_0 = 576817, \quad T_1 = 2751895, \quad T_2 = 17740121, \quad T_3 = 137826469$$

Table 2: Family Size Distribution of Jordan Population by Method II

Family size x	Observed number of family f_x	Relative frequency Y	p_x = NB fitted probability	Expected number of families
0	12411*	0.021063	0.021063	12411
1	64047	0.108696	0.065504	38596.8
2	78838	0.133799	0.113704	66997.6
3	82384	0.139817	0.145292	85610.1
4	77575	0.131655	0.152382	89787.7
5	68431	0.116136	0.138880	81832.0
6	57795	0.098089	0.113852	67084.8
7	46127	0.078284	0.085885	50605.8
8	35983	0.061068	0.060573	35691.3
9	25766	0.043728	0.040409	23810.1
10	18410	0.031244	0.025724	15157.3
11	8303	0.014091	0.015733	9270.3
12	5315	0.009203	0.009295	5476.9
13+	7843	0.013311	0.011704	6896.3
Total	589228	1	1	589228

Then,

$$\hat{k} = \frac{2T_2^2 - T_1T_2 - T_1T_3}{T_1T_3 - T_1T_2 + T_1^2 - T_2^2} = 8.5965$$
$$\hat{p} = \frac{T_1T_2 - T_1^2}{T_1T_3 - T_2^2} = 0.6382$$
$$\hat{p}^{\hat{k}} = 0.021063$$

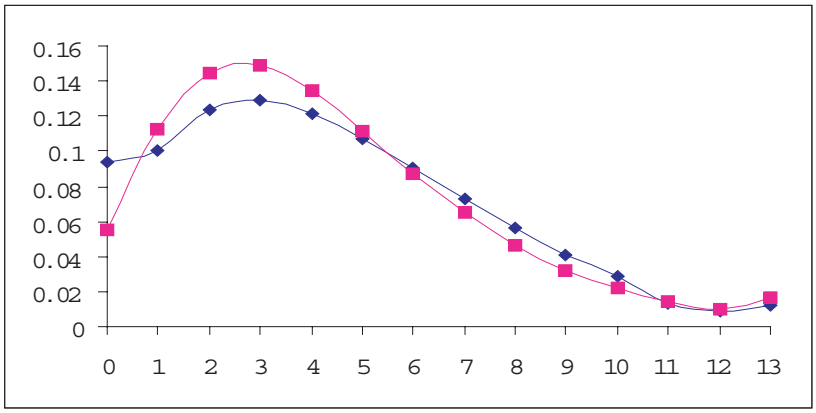

Figure 2: Observed and fitted curve of family size using method II, bold = adjusted value, smoky = fitted value.

$$\tilde{N} = \frac{T_0}{1-\hat{p}^{\hat{k}}} = 589227.9$$
$$\tilde{N}_0 = \hat{p}^{\hat{k}}\tilde{N} = 12410.0$$

The fitted distribution is:

$$f(x) = \frac{\Gamma(x+8.5965)}{\Gamma(x+1)\Gamma(8.5865)}(0.6382)^{8.5965}(0.3618)^x, \quad x = 0, 1, \cdots, 13 \qquad (6)$$

Table 2 contains the family size distribution based on Method II. Column 1 contains the family size (x). Column 2 contains the observed number families f_x for each value of x, with the zero class being adjusted. Column 3 contains the relative frequency of each value of x, column 4 contains the theoretical probability of each value of x based on the negative binomial distribution with $p = 0.6382$ and $k = 8.5965$. The last column contains the fitted family size.

It can be seen from this table and Figure 2 that the fitting curve and the actual curve have a very similar general shape with some discrepancies in the empirical and the theoretical probabilities. The total number of electively or

Table 3: Results of Iteration for the Jordanian Families Population

	μ	σ^2	p	k	$\tilde{N}_0$	$\tilde{N}$
1	4.58052	5.46960	0.535923	5.28965	22170.1	598918
2	4.59429	8.50939	0.539908	5.39130	21591.0	598386
3	4.59873	8.49718	0.541207	5.42481	21406.4	598217
4	4.60015	8.49327	0.541623	5.43560	21347.6	598162
5	4.60060	8.49203	0.547756	5.43903	21329.0	598145
6	4.60075	8.49163	0.541798	5.44013	21323.0	598140
7	4.60079	8.49150	0.541812	5.44048	21321.1	598138
8	4.60081	8.49146	0.541816	5.44059	21320.5	598138
9	4.60081	8.49145	0.541817	5.44062	21320.4	598137
10	4.60081	8.49145	0.541817	5.44063	21320.3	598137
11	4.60081	8.49145	0.541818	5.44063	21320.3	598137

biologically sterile families is approximated at $59979-12411 = 47568(7.5\%)$. In other words, based on this method of fitting about 79.3% of all childless families are electively or biologically sterile.

Method III

We do method III on the computer, beginning with values of p, k from the first method. The algorithm is:

1. Find mean (μ) and variance (σ^2) from the frequency distribution.
2. Obtain p and k of the distribution using $p = \mu/\sigma^2, k = \mu^2/(\sigma^2 - \mu)$
3. Obtain $\tilde{N}$ (total) and N_0 (total number of zero's) using $\tilde{N} = T_0/(1-p^k)$, where $T_0 = 576817$, and $\tilde{N}_0 = \tilde{N}_0 p^k$.
4. Replace f_0 by $\tilde{N}_0$, and N by $\tilde{N}$.
5. Repeat steps 1, 2, 3 and 4 many time (L times) until the values of $\tilde{N}$ and $\tilde{N}_0$ converge.

Table 3 contains the results of the first 11 iterations.

Table 4: Family Size Distribution of Jordan Population by Method III

Family size x	Observed number of family f_x	Relative frequency Y	$p_x =$ NB fitted probability	Expected number of families
0	21320*	0.033645	0.035645	21320
1	64047	0.107071	0.088856	53148
2	78838	0.131806	0.131106	78419
3	82384	0.137734	0.148987	89115
4	77575	0.129694	0.144046	86159
5	68431	0.114407	0.124615	74537
6	57795	0.096630	0.099354	59427
7	46127	0.077118	0.074400	44502
8	35983	0.060158	0.053011	31708
9	25766	0.043077	0.036273	21696
10	18410	0.030779	0.024000	14355
11	8303	0.013881	0.015435	9233
12	5315	0.008886	0.009689	5795
13+	7843	0.013112	0.014583	8723
N	598137	1	1	598137

* Adjusted value

The results after 11 iterations are:

$$\hat{p} = 0.541818, \hat{k} = 5.44063, \tilde{N}_0 = 21320.3, \tilde{N} = 598137$$

and hence,

$$f(x) = \frac{\Gamma(x + 5.4406)}{\Gamma(x+1)\Gamma(5.4406)}(0.54406)^{5.4406}(0.4582)^x, \quad x = 0, 1, \cdots, 13 \quad (7)$$

Table 4 contains the family size distribution based on Method III. column 4 contains the theoretical probability of each value of x based on the negative

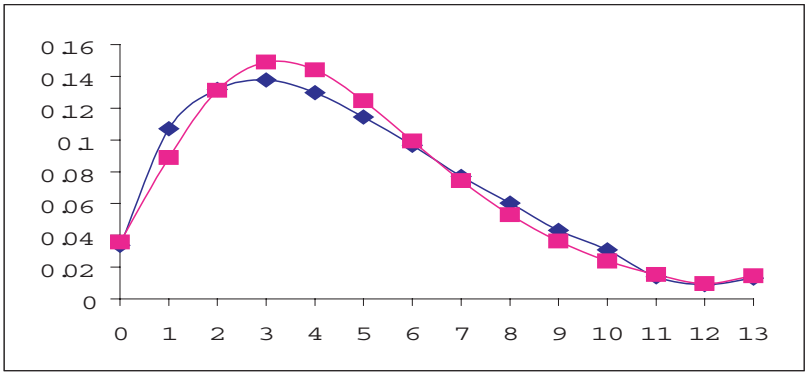

Figure 3: Observed and fitted curve of family size using method III, bold = adjusted value, smoky = fitted value.

binomial distribution with $p = 0.5418$ and $k = 5.4406$. The last column contains the fitted family size.

It can be seen from this table and Figure 3 that the fitting curve and the actual curve are very close.

Table 5 contains the fitted frequencies for the three methods along with the actual frequencies. Table 6 contains a summary of the estimated $\mu, \sigma^2 p$ and k for the three methods.

4. Results

It can be seen from the previous tables and graphs that the last method is the most appropriate method of fitting. So we adopt this method. Based on this method, the number of childless fertile couples is approximated by (213200) families, and the number of childless sterile (biologically or electively) families are (38659) family. Hence, we can approximate the proportion of childless sterile families in Jordan at (6.1%). To the best of our knowledge there was no reported value of the proportion of sterile families in Jordan.

Table 5: Observed and Fitted Number of Families of Various sizes

Family size x	Observed number of families	Fitted Number		
		I Complete	II Truncaated	III Iteration
0	59979	35240.3	12410.9	21320.30
1	64047	71620.5	38596.8	53147.95
2	78838	91743.2	66997.6	78419.45
3	82384	94545.1	85610.1	89114.78
4	77575	85593.7	89787.7	86159.26
5	68431	71060.1	81832.0	74536.82
6	57795	55433.1	67084.8	59427.09
7	46127	41258.0	50605.8	44501.63
8	35983	29604.7	35691.3	31707.65
9	25766	20623.3	23810.1	21696.11
10	18410	14022.3	15157.3	14355.17
11	8303	9341.8	9270.3	9232.49
12	5315	6119.6	5476.9	5795.39
13+	7843	10590.3	6896.3	8722.63
total	636796	636796	589227.9	598137
sterile	–	-	47568.1	38659
Grand total	636796	636796	636796	636796

5. Conclusions

The methods of fitting the negative binomial distribution to the population data give us clues to how to estimate the proportion of sterile (infertile) families in Jordan population (π). It should be noted that not all-childless families are sterile families. Thus estimating the proportion of sterile families based on all childless families (Crude estimate) would give a number that is higher than the actual number. To approximate π, we consider the childless family as being a mixture of two types of families: (1) One

Table 6: Estimates of the parameters of the fitted negative binomial distribution for the Jordanian population for the three methods.

Parameters	Complete I	Truncated II	Iteration III
Mean μ	4.32151	4.87267	4.60081
Variance σ^2	9.187	7.6346	8.49145
$\hat{p} = \mu/\sigma^2$	0.47034	0.638235	0.541818
$\hat{k} = \mu^2/(\sigma^2 - \mu)$	3.837	8.596491	5.44063
Estimate Total	636796	589227.9	598137
Estimated proportion of sterile families	0.0942 (crude estimate)	0.0747	0.0607

biologically fertile and could have children but didn't; this type of family should be a part of the general negative binomial distribution of family size. (2) Another type is either biologically or electively not fertile and thus has no children. Since method III gives the best fit, we may conclude that $\hat{\pi} = 0.0607$ is a good estimate of π. Hence, the percentage of electively or biologically sterile couples in Jordan is about 6.1

It can be seen from previous tables that about 64% of the childless families are infertile. Also by inspecting table (5), and figure (3), it can be seen that there is some indication that the number of 3-child, and 4-child families are also in access. I.e. some of those families may have been become sterile (not fertile either electively or biologically).

References

Anscombe, F.J . (1948). The Transformation of Poisson, Binomial and Negative Binomial Data. *Biometrika* , **35**, 246-254.

Binns, M. R. (1986). Behavioral Dynamics and the Negative Binomial Distribution. *Oikos*, **47**, 315-318.

Furry, W.H. (1937). On Fluctuation Phenomena in the Passage of High Energy Electron Through leads. *Physical Review*, **52**, 569-581.

General Department of Statistics. (1997). Results of the General Census of Population and Housing of Jordan 1994. *Population Characteristics* 1997, 2.

Kault, D. (1996). The Shape of the Distribution of the Number of Sexual Partners. *Statistics in Medicine*, **15**, 221-230.

Kendaal, M.G. (1949). Stochastic Process and Population Growth. *Journal of Royal Statistical Society, Series B*, **11**, 230-282.

Kojima, K.I. and Kelleher,T.M. (1962) Survival of Mutant Genes. *Amer. Nature*, 329-326.

Rao, B.R., Mazumdar, S., Waller, T.H, and Li, C.C. (11973). Correlation Between the Number of Two Types of Children in a Family. *Biometrics*, **29**, 271-279.

Waller, T.H, Rao, B.R, and Li, C.C. (1973). Heterogeneity of Childless Families. *Social Biology*, **20**, 133-138.

Wilson, L.J., Fulks, J.L., and Young, J.H. (1984). Multistage estimation compared with fixed sample Size Estimation of The negative Binomial Parameter k. *Biometrics*, **40**, 109-117.

Wilson, L.T.,and Room, P.M. (1983). Clumping Patterns of Fruit and Arthropods in Cotton with Implications for Binomial Sampling. *Environmental Entomology*, **12**, 50-54.

Received July 24, 2002; accepted November 16, 2002

Mohammad Fraiwan Al-Saleh
Department of Statistics
Yarmouk Unuversity
Irbid-Jordan
E-mail: m-saleh@yu.edu.jo

Fatima Khalid AL-Batainah
Department of Statistics
Yarmouk Unuversity
Irbid-Jordan