Modern Ceylon Studies, volume 5 : 1, *January* 1974

A Statistical Model for the School Leaving Process*

H. D. SUMANASEKERA

It is well known that once pupils enter the first grade of schools all of them do not progress continuously through each successive grade till their school leaving age is reached. Apart from those who repeat each grade, there is a considerable number who leave the school earlier or later. In this paper, the pattern of drop-outs in schools in Sri Lanka for a given cohort of pupils has been studied in statistical terms with the aim of describing the school leaving process by a theoretical distribution. In the first section, the passage rates and the retention rates have been estimated using the past data. The second while observing the close analogy between the pupils' leaving process and the labour wastage in an institution, develops the argument that the lognormal distribution adequately describes the pattern of the pupils' leaving process within the compulsory school attendance age. The third section formulates some summary measures of wastage which can be use to compare different cohorts. Finally, some possible improvements to the model have been suggested in the concluding section.

1. Estimation of the Passage and Retention Rates

The data used in this study[1] consist of the island-wide pupil population in each grade for twelve successive years from 1954 to 1965. Recalling that during this period the school system comprised a lower kindergarten and an upper kindergarten, let us denote the grades by x_0, x_1, x_2........ x_9 where x_0 denotes the lower kindergarten, x_1 the upper kindergarten, x_2 the second grade and so on. Here we are concerned only with grades up to and including grade 9 because the compulsory school attendance age ended at the end of that grade during the above period. Let $N_{i,t}$ $(i=0,1........, 9; t=1,2, ...,T)$ denote the total number of pupils in the i-th grade in year t.

We further define the following symbols:

p_i $(i = 1, 2,......,9)$ the passage rate from (i-l)-th grade to i-th grade: the fraction of pupils attending i-th grade among those who attended (i-l)-th grade.

$G_i = \prod_{j=1}^{i} p_j$ the retention rate of the i-th grade; the proportion of a given cohort reaching i-th grade.

* This essay is drawn from H. D. Sumanasekera, *Statistical Models for Educational Planning in Ceylon,* University of Sheffield: Unpublished Ph.D. Thesis, 1972.

[1] Sources: (a) 1954-1964 — UNESCO, 1969 pp. 169-182.
(b) 1965 — The author is grateful to the Director, Department of Census and Statistics, Colombo for making available these data.

The first task is the efficient estimation of p_i's. If $\hat{p}_i$ indicates the estimate of p_i, it can be obtained either by,

$$\text{(a)} \quad \hat{p}_i = \sum_{t=2}^{T} N_{i,t} \div \sum_{t=2}^{T} N_{i-1,t-1} \qquad (1)$$

$$\text{(b)} \quad \hat{p}_i = \frac{1}{T-1} \sum_{t=2}^{T} \left(N_{i,t} \div N_{i-1,t-1}\right) \qquad (2)$$

or by (c) estimating the parameter of the regression model.

$$N_{i,t} = p_i N_{i-1,t-1} + \epsilon_{t-1}, \quad t = 2, 3 \ldots\ldots\ldots T \qquad (3)$$

where ϵ_{t-1} is a random normal error. It can be shown[2] that a constant term does not enter the regression model for all i and hence for the choice of the form (3).

Further it can be shown that the random errors in the model (3), have constant variance.[3] According to the theory of regression, this implies that the estimates of the p_i's obtained using the model (3) are most efficient. Hence the model (3) has been used for the estimation of p_i's. The estimates of p_i's, their standard errors and the retention rates are given in Table 1.

TABLE 1

Passage Rates, their Standard Errors and Retention Rates

Grade	*Passage rate*	*Standard error*	*Retention rate*
x_i	$\hat{p}_i$	*of* $\hat{p}_i$	G_i
0	—	—	1.000
1	0.793	0.0185	0.793
2	0.923	0.0132	0.732
3	0.914	0.0113	0.669
4	0.887	0.0062	0.593
5	0.875	0.0083	0.519
6	0.869	0.0110	0.451
7	0.854	0.0059	0.385
8	0.903	0.0121	0.348
9	0.920	0.0186	0.320

[2] For example, considering grades 6 and 7 it is observed that,

$$N_7 = \underset{(3318.85)}{-6558.54} + \underset{(0.0237)}{0.8998}\, N_6$$

't-statistic' 1.98 37.94 (d.f. = 9)

The figures in parentheses are the standard errors of the estimates. Using the 't-statistic', it is clear that the constant term is not significant at the 5% level of significance and thus it can be deleted to obtain the following:

$$N_7 = \underset{(0.0059)}{0.8540}\, N_6 \qquad (4)$$

't-statistic' 144.8 (d.f. = 10)

This is observed to be true for all cases.

[3] For example the time ordered plot of residuals of the model (4) reveals the impresison of a 'horizontal band' if one performs a 'step back' from it. Such an impression, according to Draper and Smith (1966, pp. 88–89) accounts for the presence of constant variance among the random errors.

2. Pupils' Leaving Process and the Theory of Labour Wastage

There is a close analogy between the drop-out of pupils in a school system and the labour wastage in an institution. Enrolments and drop-outs in the former correspond to recruitment and wastage in the latter. This analogy helps to make use of some of the well-established techniques available in the analysis of labour wastage for the study of the pupils' leaving process. In recent years the theory of labour wastage has been widely studied in statistical terms by several authors.[4] Almost all these studies were concerned with the leaving characteristics of a homogeneous group of individuals. These studies have shown that the individual's propensity to leave an institution depends on many factors of which length of service is the most important. In developing the theory in continuous time, three basic functions specifying the leaving process have been defined by Bartholomew (1971). In the first function the dependence of propensity to leave on length of service is expressed by the force of separation ϕ (x) defined as follows:

$$\Pr\left\{\begin{matrix}\text{man leaves with length of service in} \\ (x, x+\delta x)\text{ given that he survives to }x\end{matrix}\right\} = \phi(x)\,\delta x$$

for small δx. The second function is the survivor function G(x) which is the probability that an individual survives for a time x. Its complement F(x) is the distribution function of completed length of service and the corresponding density function is denoted by f(x). This is the third of the three basic functions referred to above. The equivalence of these functions is almost evident from their definitions. Thus,

$$f(x) = dF(x)/dx = -dG(x)/dx,$$

and

$$f(x)dx = G(x)\phi(x)dx.$$

Some of the authors mentioned have been able to fit simple parametric functions to the data on the leaving process, and have thus made empirical estimates of these basic functions. Of several parametric functions that have been successfully employed, the most interesting here is the lognormal distribution suggested and extensively applied by Lane and Andrew (1955).

The retention rates given in the last column of Table 1 represent the proportions reaching various grades. These can be considered as probabilities with which pupils reach (or survive to) various grades. The retention rates are thus equivalent to the survivor function G(x) defined above, with the constant value G_i as a stepwise approximation to the continuous $G(x_i)$. Henceforth we refer to the retention rates as the survivor function and denote its discrete values by G_i.

Now, to indicate roughly the fit of a given function to a lognormal distribution, we plot it on logarithmic probability graph paper and examine whether the points lie on or close to a straight line. Figure 1 depicts the plot of the

[4] Some of these are: Lane and Andrew (1955), Bartholomew (1959, 1967 and 1971), Forbes (1971) and Young (1971).

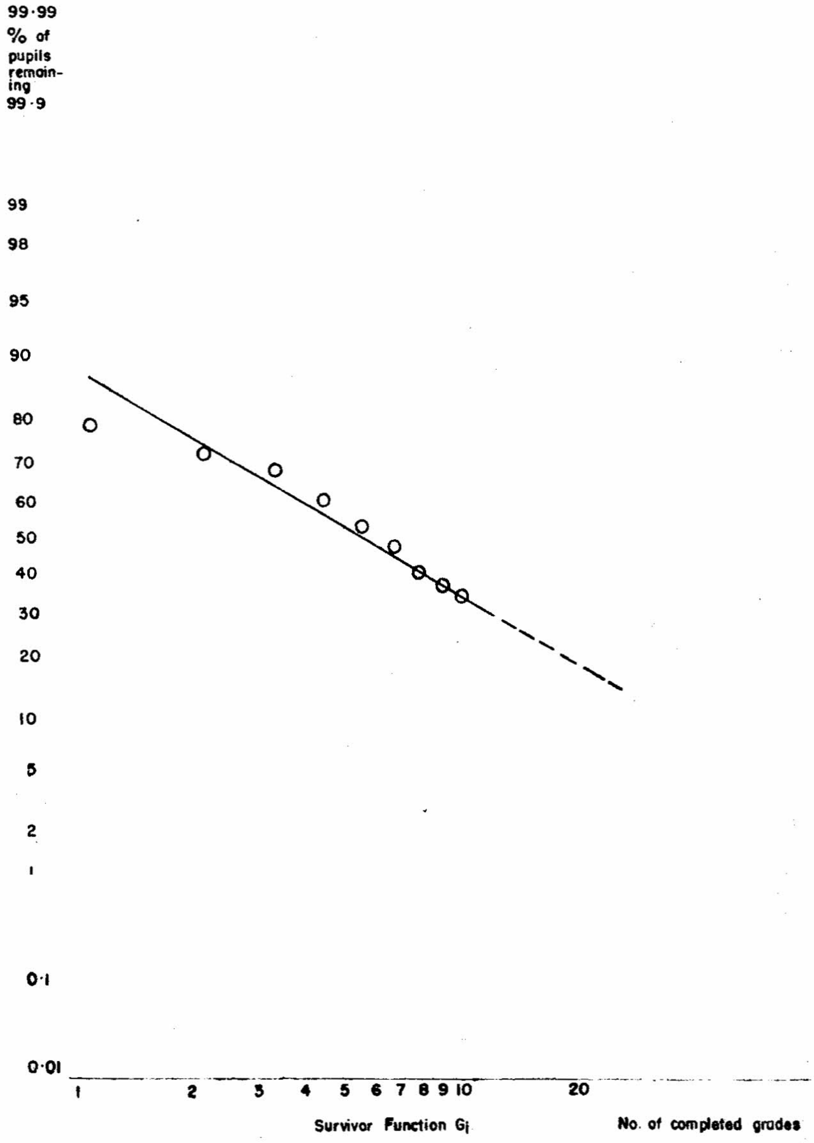
99·99
% of pupils remaining
99·9
99
98
95
90
80
70
60
50
40
30
20
10
5
2
1
0·1
0·01
1
2
3
4
5
6
7
8
9
10
20
Survivor Function Gj
No. of completed grades

Figure - 1

survivor function G_i on 2-cycle logarithmic probability graph paper, and shows the straight line fitted 'by eye' to the points. It is interesting to observe that except for the first point, the points lie on or very close to the straight line. Thus the survivor function appears to be approximately lognormal.

To obtain a goodness of fit test statistic, denote the expected number of leavers, as given by the fitted straight line, by E_i and the observed numbers by O_i,$(i=1,2,\ldots,9)$. The usual Pearson X^2 statistic may now be calculated,

$$X^2 = \sum_{i=1}^{9} (O_i - E_i)^2 / E_i$$

This has a chi-square distribution with 8 degrees of freedom on the assumption that the observed values arise from a distribution with expected values E_i. Note that in the test statistic, the number of leavers are used rather than the numbers of survivors (in grades). This is because the numbers of leavers are multinomially distributed[5] over the grades; this should be the usual situation for X^2 goodness of fit tests.

For a given cohort of 100 pupils Table 2 calculates the X^2 statistic, which is not significant at the 5% level of significance with 8 degrees of freedom. Hence it follows that the lognormal distribution adequately fits the data on the survivor function of the pupils.

TABLE 2

Calculation of the Chi-square Statistic

Grade	*No. of pupils in grade* i		*No. of leavers*		$(O_i - E_i)^2$
x_i	*Observed*	*Expected*	*Observed* O_i	*Expected* E_i	E_i
0	100	100	—	—	
1	79	88	21	12	6.75
2	73	74	6	14	4.57
3	67	64	6	10	1.60
4	59	56	8	8	0
5	52	49	7	7	0
6	45	43	7	6	0.17
7	39	39	6	4	1.00
8	35	35	4	4	0
9	32	32	3	3	0
Total			68	68	14.09 (d.f = 8)

3. Lognormal Hypothesis for the School Leaving Process and Summary Measures of Wastage

The argument advocated so far is that the lognormal hypothesis adequately describes the pattern of the pupil's leaving process within the compulsory school attendance age. This hypothesis simply states that the proportion of

[5] Chiang, 1968, p. 225.

pupils in a given cohort who leave the school after remaining a number of grades follows a certain distribution, i.e. the lognormal. The proportion leaving increases to a peak early in the school life and then decreases. This seems a reasonable explanation of the school leaving process. It can be postulated that a new entrant to the school spends the first one or two grades in preparing to adapt to the school life. However, factors such as personal handicaps, the weak financial position of the parents, the lack of education of parents, the absence of schools within a reasonable distance from home, and so forth, can result in early discontinuation of a child's school career. Provided the child has managed to withstand these forces in the early period of schooling, he is most likely to continue his school career. Thus, an important practical consequence is that if care is taken to ensure satisfaction during a child's first few years of schooling, then he is most likely to continue till the end of the compulsory school attendance age.

One of the aims of describing the school leaving process by a theoretical distribution is to search for suitable summary measures of wastage; these can then be used to compare different cohorts. One possible such quantity is the mean length of stay in school, denoted by M, as a measure of stability. Stability may be thought of as the inverse of wastage: a measure of one will serve as a measure of the other. To estimate the mean M of the fitted lognormal distribution an adjustment needs to be made at the upper tail since the lognormal distribution under study has been truncated at the end of the compulsory school attendance age.

Assume that the compulsory school attendance period extends for k years so that the density function of the lognormal distribution becomes,

$$f(x) = \frac{1}{\sigma x \sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2} (\log_e x - \mu)^2 \right\} \qquad 0 \leqslant x < k$$

$$= 0 \qquad x > k$$

$$\text{and, } \Pr(x > k) = \int_k^\infty \frac{1}{\sigma x \sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2} (\log_e x - \mu)^2 \right\} dx.$$

Suppose the parameters μ and σ^2 are estimated from the fitted straight line in Figure 1. Then the quantiles of order 16, 50 and 84% are,

$$x_{16\%} = 18.25, \quad x_{50\%} = 4.80 \text{ and } x_{84\%} = 1.275.$$

Let m and s^2 denote the estimates of μ and σ^2, then it follows that,

$$e^{m+s} = 18.25, \quad e^{m} = 4.80 \text{ and } e^{m-s} = 1.275.$$

Therefore, $m = 1.57$ and $s = 1.33$ or $s^2 = 1.77$.

The mean length of stay in school is given by [6],

$$M = \alpha \phi\left(\frac{\log_e k - \mu - \sigma^2}{\sigma}\right) + k\left\{1 - \phi\left(\frac{\log_e k - \mu}{\sigma}\right)\right\}$$

where $\alpha = \exp\left(\mu + \frac{\sigma^2}{2}\right)$, the mean of the lognormal distribution.

and $\phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \exp\left(-\frac{z^2}{2}\right) dz$, which is tabulated.

Using the estimates m and s^2 of μ and σ^2 respectively, we obtain,

$$\alpha = 11.65 \text{ years}$$

and hence, $M = 5.45$ years

taking $k = 10$ years because the compulsory school attendance age extends from 5 to 14 years. This means that at the time of enrolment the expected mean length of stay in school is 5.45 years.

An alternative measure to the mean length of stay is the half-life or median of the survivor function G(x). This is simply the time taken for half of a given cohort to leave. A large half-life indicates high stability and low wastage.

In the present case the half-life is simply,

$$x_{50\%} = e^{m} = 4.80 \text{ years},$$

This shows that roughly at the end of the fourth grade half of the cohort has left the school.

4. Possible Improvements to the Model

One strong assumption we have made during the construction of the model is that the island-wide 'entry cohort' of pupils is homogeneous with respect to their ability to remain in school. It is generally accepted that all the districts in the island are not homogeneous in educational and socio-economic levels. As a result, one can argue that the effectiveness of the reasons listed in section 3 for discontinuing pupils' schooling may vary from district to district or from one homogeneous group of district to another. To allow for this, separate models have to be constructed for each district or homogeneous group of districts. Since district-wise data for the past years are not available, improvement to the model in this direction has not been possible.

Another assumption in the model is that there is no temporal variation in the passage rates. This is justifiable because of the very small standard errors accompanying the estimates of the passage rates (see Table 1). However, this assumption can be relaxed if one is willing to fit lognormal distributions to data for each year.

[6] Lane and Andrew, 1955, p. 307.

Finally, note that during the estimation of passage rates we assumed that all the pupils in a given grade take the promotion test at the end of the year for the first time and the majority of them get promoted to the next grade while the rest leave the school. Since the available data for each grade do not distinguish between those attending for the first time and those repeating, we are forced to make this assumption. Suppose there has been a considerable number of repeaters in each grade and if it is reasonable to assume a higher passage rate for them, then under the above assumption we have over-estimated the passage rates. Unless more refined data are available relaxation of this assumption too is difficult.

References

Bartholomew, D. J.
1959 : "Note on the Measurement and Prediction of Labour Turnover," *Journal of the Royal Statistical Society*, Vol. 122(A), pp. 232–39.
1967 : *Stochastic Models for Social Processes*, New York: John Wiley & Sons.
1971 : "The Statistical Approach to Manpower Planning," *The Statistician*, Vol. 20, pp. 3–36.

Chiang, C. L.
1968 : *Introduction to Stochastic Processes in Biostatistics*, New York: John Wiley & Sons.

Draper, N. R. and Smith, H.
1966 : *Applied Regression Analysis*, New York: John Wiley & Sons.

Forbes, A. F.
1971 : "Non-parametric Methods of Estimating the Survivor Function," *The Statistician*, Vol. 20, pp. 27–53.

Lane, K. F. and Andrew, J. E.
1955 : "A Method of Labour Turn-over Analysis," *Journal of the Royal Statistical Society*, Vol. 118(A), pp. 296–323.

Sumanasekera, H. D.
1972 : *Statistical Models for Educational Planning in Ceylon*, Unpublished Ph.D. thesis, University of Sheffield.

UNESCO
1969 : *Progress of Education in the Asian Region: A Statistical Review*, Bangkok: UNESCO Regional Office for Education in Asia.

Young, A.
1971 : "Demographic and Ecological Models for Manpower Planning," in D. J. Bartholomew and B. R. Morris (eds.), *Aspects of Manpower Planning*, London: The English Universities Press, pp. 75–97.