The Korean Journal of Applied Statistics (2008)
21(5), 889–900

A Child Labour Estimator for Lahore Based on Literacy and Poverty Variables

Ahmed F. Siddiqi[1]

[1]Dept. of Business and Economics, University of Management and Technology Lahore

(Received June 2008; accepted August 2008)

Abstract

Child labour is a disturbing issue for any society. It is attempted here in this article to develop an estimator to assess the numerical strength of this menace in Lahore division. A Horvitz and Thompson (1952) type of estimator is developed where weights are calculated on the basis of poverty and illiteracy to increase the sampling efficiency. Different characteristic features of this estimator, like its unbiasedness, variance, probability distribution, confidence intervals are also developed for its study from different angles.

Keywords: Child labour, Horvitz & Thompson estimator, illiteracy, Lahore, Pakistan, poverty.

1. Introduction

Child labour is growing like mushroom in all developing societies. While hardly a new phenomenon, the issue of child labour has attracted increasing attention in the past decade from policy makers, advocates and researchers. Quite persistent in nature (as Grootaert and Kanbur (1995) curse it), it seems to be omnipresent everywhere. It is present in developed, not-so-poor, societies of the west (ILO, 2004) as well as in, recognized poor, societies of the African and Asian countries (Cockburn, 2002). It is there in democratic states where people are making their own decisions for the lives as well as in states where people are subjugated by foreign decisions (Jayshi, 2005). It has made its presence felt in communities where people are living in smaller households with relatively lesser social intimacy as well as where people lives in joint families in larger households (see Edmonds and Pavcnik, 2002). No doubt, its gravity is decreasing in some parts of the developed countries, as enunciated in reports and surveys surfaced recently in connection with World Children day, falling on November 20, but the Asian and African developing countries are still facing the menace in its full potency.

Literature on child labour is replete with myriad of surveys, case studies, reports, news reports and articles exploring the phenomenon from different angles. Numerical strength of the menace, important causes that give birth to this problem, critical differentials present, either in society or in the household that fuel the problem, abominable consequences of this menace on the society are all well highlighted research subjects. However, academic literature is not discussing much about

[1]Assistant Professor, School of Business and Economics University of Management and Technology Lahore, Pakistan. E-mail: ahmed.siddiqi@umt.edu.pk

methods to assess numerical strength of this issue (Grootaert and Kanbur, 1995) and it is considered more of a state launched activity. Bequele and Boyden (1988) criticized on the surveys that those, in most cases, are based upon fairly small samples and carried out at the place of work, enabling authors to assess the working conditions and atmosphere instead of their households. ILO/SIMPOC and UNESCO has been vocal in this regard. ILO/SIMPOC has conducted thorough investigations on the availability of quality of statistical data relating to working kids in more than 200 countries and territories and found that such data did not exist in the majority of cases and much of the official statistics, if available at all, were deficient in many ways.

As a matter of fact, mostly child labour surveys have been conducted using design based sampling techniques (see Hanif (2000) for an excellent discussion over the debacle between design based and model based sampling techniques). However, the design based techniques are deficient in capturing a pocketed phenomenon (Basu, 1999) like child labour. It is quite logical to use some important determinant of the phenomenon, if they exist, as steering variables to capture the population in its real sense and to avoid non-responses. Such usage may lead to some model based sampling techniques, or at least in a Horvitz and Thompson (1952), or Hansen and Hurwitz (1943) type of weighted sampling schemes. It may be a hybrid scheme which is combining both of these philosophies but at different stages.

The present article is an attempt to develop an estimator for assessing the numerical strength of the child labour in Lahore division by making use of available important determinants of child labour. It has been assumed that poverty and illiteracy are two crucial variants in the dynamics of child labour so they can easily be used to increase the precision of the estimate. However, Horvitz and Thompson (1952) estimators are used instead of some congruent model in the sampling scheme. A multi stage probability proportional sampling scheme is proposed here which is based on weights calculated on the basis of poverty and illiteracy statuses and on systematic selection of clusters of households where each cluster is composed of 12 households. The details are available in the coming Section 2 which is showing different stages of the development of such an estimator for Lahore division which is the biggest division in the biggest province, Punjab, of Pakistan. Section 3 is expounding on different characteristic features of this estimator including its probability distribution, unbiasedness, variance and confidence interval.

2. Development of an Estimator

Believing a highly pocketed nature of the phenomenon, a multi-stage probability proportional sampling scheme is proposed here where special weights are to be calculated on the basis of important determinants of child labour in any society. Poverty (Basu, 1999) and literacy level (Ray and Lancaster, 2004; Ray, 2000) can easily be regarded as such two determinants. The literacy level at any administrative hierarchical level may easily be gauged through the school enrolment statistics widely available in different government statistics. Poverty level is, however, somewhat difficult to calculate, especially at lower administrative hierarchical levels. An indirect way to gauge poverty is through calculating economic activity going on. Different marketing research companies, like AC Nielsen, Aftab Associates, Synovate, *etc.* are conducting retail-shops censuses, quite on regular basis, enumerating number of retail shops in all areas. These statistics are a good indicator of economic activity in that area; poverty status of an area is inversely proportional to number to shops in that area. A single joint index can, thus, be formulated by the help of these two.

If Y is the total number of child labourers in Lahore division and $\hat{Y}_h$ denotes its sample estimate,

Table 2.1

Districts	l	ω	$\Phi = \min(l, \omega)$
Lahore	0.11599	0.2314	0.11599
Kasur	0.12135	0.1387	0.12135
Sheikhupura	0.11847	0.1562	0.11847
Okara	0.11881	0.1645	0.11881

then Horvitz and Thompson (1952) estimator for the situation is given as

$$\hat{Y}_h = \sum_{i}^{d_h} \frac{\hat{Y}_{ih}}{\pi_i},$$

where d_h is the number of districts and π_i denotes the probability of selection associated with i^{th} district. This estimator does not depend on the number of times a unit may be selected and each distinct unit is utilized only once. Hansen and Hurwitz (1943) estimators are also available for these situations but they, generally, do better in case of sampling with replacement (Mohammad, 1999). Gupta *et al.* (1982) also present estimator for these situation by using a scheme based on certain combinatorial properties of balanced block designs and are available for any sample size. Deville and Tille (1998) propose a method consisting of splitting the inclusion probability vector into several new inclusion vectors. The method is giving us more efficient estimators but unfortunately there are a number of restrictions, like Sen-Yates-Gundy conditions (see Yates, 1960; Hansen, 1953) or Gablers' sufficient conditions (see Gabler, 1984) which these estimators have to follow. These restrictions are making the applicability of Deville and Tille's estimator confined to certain exclusive situations. The probability of selection is calculated through the special weights assigned with respect to the literacy and poverty statuses of the districts. Horvitz and Thompson (1952) restricted their use to the satisfaction of these conditions:

$$\begin{cases} \sum_{i}^{D_h} \pi_i = d_h, \\ \sum_{j\neq i}^{D_h} \pi_{ij} = (d_h - 1)\pi_i, \\ \sum_{i}^{D_h} \sum_{j>i}^{D_h} \pi_{ij} = \frac{1}{2} d_h (d_h - 1). \end{cases}$$

(Lohr, 1999, pp. 205) gives explicit discussions and proofs for these restrictions.

The administrative structure of Lahore division splits it into four districts which are further subdivided into tehsils. A literacy index, l, is the ratio of all students in the age bracket (5–15) to all kids in the same age bracket. A higher score on l would lower the chances of selection. A poverty index, ω, is the ratio of all shops in the area to all shops in Lahore. A higher score on l would increase the chances of selection. The single joint index, Φ, is the minimum of these two. Table 2.1 is giving these indices for all the four districts in Lahore.

If $1/3^{rd}$ districts are to be selected, going in line of Gallup Surveys (see Davis, 1989; Sudman, 1997) only Lahore district would be selected. In the next stage, tehsils are to be selected. Lahore district has two tehsils. Going through the same arguments, we are to select only one tehsil where its probability of selection, φ, is based on its literacy and poverty statuses. The sample estimate of the total number of child labourers in Lahore is thus modified as

$$\hat{Y}_h = \sum_i^{d_h} \frac{1}{\pi_i} \sum_j^{t_{ih}} \frac{\hat{Y}_{jih}}{\varphi_{ji}},$$

where φ_{ji} is the probability that j^{th} tehsil of the i^{th} district is in the sample, t_{ih} is the total number of tehsils be selected in the i^{th} district and $\hat{Y}_{jih}$ is the total number of child labourers in the j^{th} tehsil of i^{th} district of Lahore division. In each of the selected tehsils, clusters of households are to be selected systematically, in the next stage. This clustering is based on geographical proximity of the households while the selection is based on grouping with respect to the location of the union councils. If η_j denotes the total numbers of clusters be selected in the j^{th} tehsil, given by

$$\eta_j = \frac{H_{jih}}{\sum_i^{D_h} \sum_j^{T_{ih}} H_{jih}} \times n_h,$$

where H_{jih} denotes the number of clusters of households in j^{th} tehsil in the i^{th} district while D_h and T_{ih} denotes the total of districts and tehsils in Lahore division. The n_h denotes the sample size calculated for the study. Using Cochran (1977), $\hat{Y}_{jih}$ is given by

$$\hat{Y}_{jih} = H_{jih} \sum_k^{\eta_j} \frac{\hat{Y}_{kjih}}{\eta_j},$$

where $\hat{Y}_{kjih}$ is the sample estimate of the total number of child labourers in the k^{th} PSU in the j^{th} tehsil of the i^{th} district. So, the final estimator turns out to be

$$\hat{Y}_h = \sum_i^{d_h} \frac{1}{\pi_i} \sum_j^{t_{ih}} \frac{H_{jih}}{\varphi_{ji}} \sum_k^{\eta_j} \frac{\hat{Y}_{kjih}}{\eta_j}.$$

As every PSU is a cluster of 12 households, the final shape of the estimator is thus given by

$$\hat{Y}_h = \sum_i^{d_h} \frac{1}{\pi_i} \sum_j^{t_{ih}} \frac{H_{jih}}{\varphi_{ji}} \left(\frac{\sum_i^{D_h} \sum_j^{T_{ih}} H_{jih}}{n_h H_{jih}} \right) \sum_k^{\eta_j} \sum_l^{12} \hat{Y}_{lkjih},$$

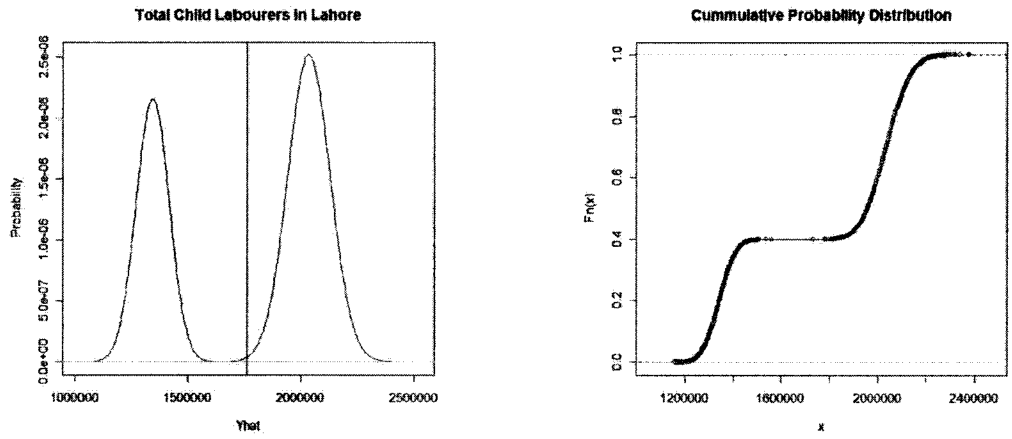
where $\hat{Y}_{lkjih}$ is giving the total number of labouring kids per households.

3. Properties of the Estimator

The statistical validity of the estimator is established by investigating its characteristic features, including its biasedness, variability, associated probability distribution, confidence interval, etc. In the coming sections, these properties are explored.

3.1. Probability distribution

A traditional and conventional method for identifying the probability distribution associated with this estimator stems from the identification of the distribution associated with $\hat{Y}_{lkjih}$, the sample



Probability Distribution | Cumulative Probability Distribution

Figure 3.1. Empirical probability distribution for the child labour estimator for Lahore division

estimate of the number of child labourers per household. Statistical literature suggests it to be Poisson distribution while the Punjab statistics suggests its parametric value to be 4 (PCO, 1998). Sex ratio, in Punjab, is 52.37%. Since, the actual selection is guided and based on economic and literacy statuses, so the chances that the selected household has labouring kids is significant. Further, using results from Basu and Chau (2003) and Beegle *et al.* (2004), if one kid is working in a household the chances that all kids are working in that households is also significant. A simulation experiment, however, conducted using *R* with its special routine PPS to have an idea about the probability distribution for the estimator. Number of kids per households is set to follow Poisson distribution with mean 4 which are further divided into males and females according to sex ratio. The code, written in *R*, is available in appendix for ready reference. Simulation is done for 10,000 times and it took 35 minutes on a Pentium IV D 3.4GHz with 2GB RAM machine. The code is using real data except for the number of child labourers per household for which it is using Poisson (4) distribution to generate it. The PPS routine is responsible for selecting household on the basis of joint index defined on the twin basis of literacy and economic indices. It is asked to produce results in the form of probability density graph along with its associated ogive shown here in Figure 3.1: left panel is showing the empirical probability distribution while the right panel is showing the corresponding ogive.

The curve in Figure 3.1 (left panel) is a, kind of, bi-modal curve except the left sided hump is of lesser height. Repeated simulations do not have any effect on this bi-modal-ness. One possible reason for this bi-modal-ness is a different distribution for male and female kids per households. One may suspect two different selection probabilities; one for district selection whiles the other for tehsil as the cause of this bi-modal-ness. Apparently, the curve does not look like any known probability modal. Even, the ogive, in the right panel, is adding in the confusion. However, no decision can be made regarding the probability distribution of the estimator from here and it needs rigorous testing.

3.2. Calculating probability

The same empirical probability distribution, in Figure 3.1 may also be used to calculate probabilities statements. If $f(\hat{Y})$ denotes the probability distribution associated with $\hat{Y}$, then

$$P(X \geq T) = \int_T^\infty f\left(\hat{Y}\right) d\left(\hat{Y}\right).$$

In the absence of a density function, we have to rely on the empirical distribution of the estimator. 10,000 observations are quite enough to be used for probability calculations. Here is a small function written in R to calculate this probability at any given point x.

```
prob<-function(x){
p<-0
for (i in 1:length(ys)){
      if (ys[i]>x) p<-p+1}
      p/length(ys)
}
```

As an illustration of this code,

```
prob(mean(data))=0.6013
```

3.3. Unbiasedness

Another important characteristic of the estimator is its unbiasedness. The unbiasedness of the estimator would establish the sanctity of the estimator against different sampling techniques. The multistage nature of the estimator asks for a separate process(expectation to be exact) at each step. Equivalently,

$$\begin{aligned}
E\left(\hat{Y}_h\right) &= E\left[\sum_i^{d_h} \frac{1}{\pi_i} \sum_j^{t_{ih}} \frac{H_{jih}}{\varphi_{ji}} \sum_k^{\eta_j} \frac{1}{\eta_j} \sum_l^{12} \hat{Y}_{lkjih}\right] \\
&= E_1\left[\sum_i^{d_h} \frac{1}{\pi_i} E_2\left(\sum_j^{t_{ih}} \frac{1}{\varphi_{ji}} E_3\left(H_{jih} \sum_k^{\eta_j} \frac{1}{\eta_j} E_4\left(\sum_l^{12} E_5\left(\hat{Y}_{lkjih}\right)\right)\right)\right)\right] \\
&= E_1\left[\sum_i^{d_h} \frac{1}{\pi_i} E_2\left(\sum_j^{t_{ih}} \frac{1}{\varphi_{ji}} E_3\left(H_{jih} \sum_k^{\eta_j} \frac{1}{\eta_j} E_4\left(\sum_l^{12} \hat{Y}_{lkjih}\right)\right)\right)\right] \quad \text{as } \hat{Y}_{lkjih} \text{ in unbiased} \\
&= E_1\left[\sum_i^{d_h} \frac{1}{\pi_i} E_2\left(\sum_j^{t_{ih}} \frac{1}{\varphi_{ji}} E_3\left(H_{jih} \sum_k^{\eta_j} \frac{1}{\eta_j} \sum_l^{12} \hat{Y}_{lkjih}\right)\right)\right] \\
&= E_1\left[\sum_i^{d_h} \frac{1}{\pi_i} E_2\left(\sum_j^{t_{ih}} \frac{1}{\varphi_{ji}} \sum_k^{\eta_j} \frac{H_{jih}}{\eta_j} \sum_l^{12} \hat{Y}_{lkjih}\right)\right] \\
&= E_1\left[\sum_i^{d_h} \frac{1}{\pi_i} E_2\left(\sum_j^{t_{ih}} \frac{1}{\varphi_{ji}} \sum_k^{\eta_j} \hat{Y}_{kjih}\right)\right].
\end{aligned}$$

Let ν_{ji} $(j = 1, 2, \ldots, t_{ih})$ be a random variable that takes the value 1 if the j^{th} tehsil is drawn and zero otherwise. Then ν_{ji} follows the Bernoulli distribution, with probability φ_{ji}. Thus the mean and the variance of this random variable are given by φ_{ji} and $\varphi_{ji}(1-\varphi_{ji})$ respectively (see Ross, 2002). Using this scheme of thought to tackle E_2.

$$\begin{aligned} E\left(\hat{Y}_h\right) &= E_1\left[\sum_i^{d_h}\frac{1}{\pi_i}E_2\left(\sum_j^{t_{ih}}\frac{1}{\varphi_{ji}}\sum_k^{\eta_j}\hat{Y}_{kjih}\right)\right] \\ &= E_1\left[\sum_i^{d_h}\frac{1}{\pi_i}E_2\left(\sum_j^{t_{ih}}\frac{\nu_{ji}}{\varphi_{ji}}\sum_k^{\eta_j}\hat{Y}_{kjih}\right)\right] \\ &= E_1\left[\sum_i^{d_h}\frac{1}{\pi_i}\sum_j^{t_{ih}}\sum_k^{\eta_j}\hat{Y}_{kjih}\right] \\ &= E_1\left[\sum_i^{d_h}\frac{1}{\pi_i}\sum_j^{t_{ih}}\hat{Y}_{jih}\right]. \end{aligned}$$

Using again the same logic. let w_i $(i = 1, 2, \ldots, d_h)$ be a random variable that takes the value 1 if the i^{th} district is drawn and zero otherwise. Then w_i follows the Bernoulli distribution with mean and the variance are given by π_i and $\pi_i(1-\pi_i)$ respectively. So,

$$\begin{aligned} E\left(\hat{Y}_h\right) &= E_1\left(\sum_i^{d_h}\frac{W_i}{\pi_i}\sum_j^{t_{ih}}\hat{Y}_{jih}\right) \\ &= \sum_i^{d_h}\sum_j^{t_{ih}}\hat{Y}_{jih} \\ &= Y, \end{aligned}$$

which proves the unbiasedness of the estimator.

3.4. Variance of the estimator

The precision of the estimator needs variance to be calculated which is derived as

$$V\left(\hat{Y}\right) = V\left(\sum_h^{3}\sum_i^{d_h}\frac{1}{\pi_i}\sum_j^{t_{ih}}\frac{H_{jih}}{\varphi_{ji}}\sum_k^{\eta_j}\frac{1}{\eta_j}\sum_l^{50}\hat{Y}_{lkjih}\right).$$

Using some established results from sampling theory

1. Available in Lohr (1999) that $V(\hat{Y}) = \sum_h^L N_h(N_h - n_h)S_h^2/n_h$.
2. Available in Raj (1966) that for all known w_{is}, $V(\sum_i^n w_{is}\hat{Y}_i) = V(\sum_i^n w_{is}Y_i) + \sum_i^N E_1({w'_{is}}^2)\sigma_{2i}^2$ where is w'_{is} a random variable that equals wis if ith unit is drawn and equals zero otherwise.
3. Available in Gabler (1981) that for $(\pi_i > 0)$ $V(\sum_i^n(y_i/\pi_i)) = \sum_i^N\sum_{j>i}^N(\pi_i\pi_j - \pi_{ij})(y_i/\pi_i - y_j/\pi_j)^2$.

The variance is given by

$$V\left(\hat{Y}\right) = \sum_i^{D_h}\sum_{g>i}^{D_h}(\pi_i\pi_g - \pi_{ig})\left(\frac{Y_i}{\pi_i} - \frac{Y_g}{\pi_g}\right)^2 + \sum_i^{D_h}\frac{T_{ih}(T_{ih} - t_{ih})}{t_{ih}}\left(\frac{S_{2i}^2}{\pi_i}\right),$$

where S_{2i}^2 is the second stage variance, *i.e.* the variance within a selected district. Within a district, tehsils are to be selected. This selection is again a PPS. Using afore mentioned results again to have

$$S_{2i}^2 = \sum_{j}^{T_{ih}} \sum_{f>j}^{T_{ih}} (\varphi_j \varphi_f - \varphi_{jf}) \left(\frac{Y_j}{\varphi_j} - \frac{Y_f}{\varphi_f} \right)^2 + \sum_{j}^{T_{ih}} \frac{H_{jih}(H_{jih} - \eta_j)}{\eta_j} \left(\frac{S_{3j}^2}{\varphi_j} \right),$$

where S_{3i}^2 is the third stage variance, *i.e.* the variance within a selected tehsil. Within a tehsil, clusters of households, PSUs, are to be selected. This selection is done using systematic sampling. Using results in Cochran (1977), this variance is given by

$$S_{3j}^2 = H_{jih}(H_{jih} - 1)S_H - H_{jih}q(\eta_j - 1)S_{4k}^2,$$

where S_H is the total variance of all PSUs within selected j^{th} tehsil, S_{4k}^2 is the variance within selected PSU, and q are the number of systematic groups to select n_{jih} given by ratio of H_{jih} to sample size in the locale, n_{jih}. S_{4j}^2 is the sum total of variance per household. If S_{5k}^2 is the variance per household, the variance of the estimator is modified as

$$\begin{aligned} V\left(\hat{Y}\right) = & \sum_{i}^{D_h} \sum_{g>i}^{D_h} (\pi_i \pi_g - \pi_{ig}) \left(\frac{Y_i}{\pi_i} - \frac{Y_g}{\pi_g} \right)^2 \\ & + \sum_{i}^{D_h} \frac{T_{ih}(T_{ih} - t_{ih})}{t_{ih}\pi_i} \left[\sum_{j}^{T_{ih}} \sum_{f>j}^{T_{ih}} (\varphi_j \varphi_f - \varphi_{jf}) \left(\frac{Y_j}{\varphi_j} - \frac{Y_f}{\varphi_f} \right)^2 \right] \\ & + \sum_{i}^{D_h} \frac{T_{ih}(T_{ih} - t_{ih})}{t_{ih}\pi_i} \left[\sum_{j}^{T_{ih}} \frac{H_{jih}(H_{jih} - \eta_j)}{\eta_j} \left(\frac{H_{jih}(H_{jih} - 1)S_H - H_{jih}q(\eta_j - 1)\sum_k^{\eta_j} S_{5k}^2}{\varphi_j} \right) \right], \end{aligned}$$

which is the variance associated with the estimator. Corresponding standard error, relative precision may easily be calculated.

3.5. Numerical estimate of the child labour in Lahore

In the absence of actual household data, the empirical distribution, in Figure 3.1, may be used to give an idea of the numerical magnitude of the child labour in Lahore division. This would also establish the veracity of the estimator. Summary statistics is given as follows,

Min	1^{st} Quartile	Median	Mean	3^{rd} Quartile	Max	S.D.	S.E.	C.V.	Relative Precision
1156000	1361000	1960000	1761000	2053000	2380000	347285.24	137.82	19.7%	3.21%

In other words, there are 1.76 million kids are working in Lahore division detrimental to their physical, mental or psychological health. This amount to 1 in every 5 kids in Lahore is a child labourer. These statistics seems quite exaggerated in comparison to the figures revealed in the only child labour survey conducted in Pakistan, in 1996 (FBS, 1996) but exactly in line of figures generated by other non-governmental organizations, like SPARC (2005).

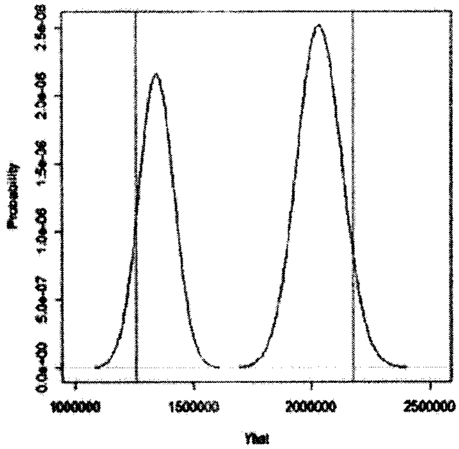


Figure 3.2. 95% confidence limits for child labour estimate in Lahore division

3.6. Confidence interval

Typically, a two sided $(1-\alpha)100\%$ confidence interval for Y, as Lehmann (1998), Chapter5, Section 6, put it, is given by

$$P(X_l \le Y \le X_u) \ge 1-\alpha,$$

where X_l and X_u are respectively the lower and upper confidence bounds. In other words,

$$\int_{\hat{Y}}^{X_u} f\left(\hat{Y}\right) d\left(\hat{Y}\right) = 1 - \frac{\alpha}{2} = \int_{\hat{Y}}^{X_l} f\left(\hat{Y}\right) d\left(\hat{Y}\right).$$

We may use the same empirical distribution in Figure 3.1, instead of theoretical distribution, to calculate these limits. As a matter of fact, the same routine, appearing in Section 3.2, may be used to find these limits. Here is a small routine, in R, to find these confidence limits for any given confidence level.

```
conf.int<-function(x) quantile(ys,c((100-x)/200,(100+x)/200))
```

The code is asking for confidence level (in percentage but without its sign). For example `conf.int(95)` yields 1257580, 2179897 as the lower and upper bounds respectively. Figure 3.2 is showing these confidence limits graphically on the graph of its empirical distribution.

4. Epilogue

The article is attempting an estimator for assessing numerical magnitude of the child labour menace in the Lahore division, Pakistan. The estimator is using poverty and illiteracy as steering variables in devising a sampling scheme and used these statuses as weighting constants in its development. The empirical distribution ascribed to this estimator is a kind of bi-modal graph. The numerical results from this empirical distribution are quite in consonance with many non-governmental organizations. The design is applied, here in this article, in Lahore division but it is powerful enough to be used anywhere.

The child labour problem is an endemic problem seems to be knitted with socio economic structure of a society and needs special and exclusive efforts to have a hold over. The very first step, academicians should, and can, take is to provide academic base so that practitioners may use it to eliminate this menace. An assessment of numerical strength is giving such a base but in numbers which are then used by policy makers and others.

Appendix

```
sim<-10000
ys<-array(0,sim)
for (a in 1:sim){

fs<-7
pcl<-0.11575
cluster<-12
teh<-matrix(0,2,1)
HH<-matrix(0,2,2)
yklj<-array(0,2)

hhts<-rbind(c(531277,350431,0,0,0,0),
      c(106074,114380,47586,28413,57903,0))/cluster

vphi<-matrix(0,2,6)
phi<-matrix(0,2,6)
sizest<-rbind(c(0.823,0.546,0,0,0,0),
      c(0.65763,0.834,0.744,0.7112,0.6781,0))
vphi[1,]<-(sizest[1,])/sum(sizest[1,])
vphi[2,]<-(sizest[2,])/sum(sizest[2,])
phi[1,]<-c(max(vphi[1,]),max(vphi[1,]),
    max(vphi[1,]),max(vphi[1,]),
    max(vphi[1,]),max(vphi[1,]))
phi[2,]<-diag(sampfordpi(vphi[2,],2))

pi<-array(0,2)
Phi<-rbind(c(0.11599,0.12135,0.11847,0.11881),
      c(0.1244,0.1103,0.4523,0))
pi[1]<-max((1-Phi[1,])/sum(1-Phi[1,]))
pi[2]<-max((1-Phi[2,])/sum(1-Phi[2,]))

teh[1,1]<-1
teh[2,1]<-2
HH[1,]<-c(hhts[1,ppss(hhts[1,],1)],0)
f<-sampford(hhts[2,],2)
HH[2,]<-c(hhts[2,f[1]],hhts[2,f[2]])
```

```
et<-hhts/sum(hhts)
eta<-matrix(0,2,6)
eta[1,]<-et[1,]*205
eta[2,]<-et[2,]*104

for (h in 1:2){
      for (j in 1: teh[h,1]){
            ykl<-0
            for (k in 1:eta[h,j]){
                  yl<-sum(rpois(cluster,(fs-2)*pcl))
                  ykl<-ykl+yl
            }
            ykleta<-ykl/eta[h,j]
            ykletaH<-ykleta*HH[h,j]/phi[h,f[j]]
            yklj[h]<-yklj[h]+ykletaH
      }
}
yd<-yklj/pi
ys[a]<-yd[1]
}
# Density Plot for Results
plot(density(ys),
main="Total Child Labourers in Lahore",
xlab="Yhat",
ylab="Probability")
abline(v=mean(ys))
```

References

Basu, A. and Chau, N. H. (2003). Targeting child labor in debt bondage: Evidence, theory and policy implications, *World Bank Economic Review*, **17**, 225–281.

Basu, K. (1999). Child labor: Cause, consequence and cure, with remarks on international labor standards, *Journal of Economic Literature*, **37**, 1083–1119.

Beegle, K., Dehejia, R. and Gatti, R. (2004). Why should we care about child labor? The education, labor market and health consequences of child labor, *National Bureau of Economic Research*, Cambridge.

Bequele, A. and Boyden, J. (1988). Working children: Current trends and policy responses, *International Labour Review*, **127**, 153–172.

Cochran, W. G. (1977). *Sampling Techniques*, John Wiley & Sons, New York.

Cockburn, J. (2002). Child work and poverty in developing countries, University of Oxford, Dept. of Economics: vi, 218 leaves.

Davis, J. A. (1989). Survey research in the United States: Roots and emergence 1890–1960, *Public Opinion Quarterly*, **53**, 136–138.

Deville, J. C. and Tille, Y. (1998). Unequal probability sampling without replacement through a splitting method, *Biometrika*, **85**, 89–101.

Edmonds, E. and Pavcnik, N. (2002). Does globalization increase child labor? Evidence from Vietnam, *National Bureau of Economic Research*, Working Paper Series, **8760**, 1–49.

FBS (1996). Summary results of child labour survey in Pakistan. M. o. L. FBS, Manpower and Overseas Pakistanis, Government of Pakistan.

Gabler, S. (1981). A comparison of Sampford's sampling procedure versus unequal probability sampling with replacement, *Biometrika*, **68**, 725–727.

Gabler, S. (1984). On unequal probability sampling: Sufficient conditions for the superiority of sampling without replacement, *Biometrika*, **71**, 171–175.

Grootaert, C. and Kanbur, R. (1995). Child labour: An economic perspective, *International Labour Review*, **134**, 187–203.

Grootaert, C. and Kanbur, R. (1995). Child labour: A review, *World Bank Office of the Vice President Development Economics*, Washington DC.

Gupta, V. K., Nigam, A. K. and Kumar, P. (1982). On a family of sampling schemes with inclusion probability proportional to size, *Biometrika*, **69**, 191–196.

Hanif, M. (2000). Design- and model- based sampling inference, *Pakistan Journal of Statistics*, **16**, 229–246.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations, *The Annals of Mathematical Statistics*, **14**, 333–362.

Hansen, M. H. (1953). *Sample Survey Methods and Theory*, John Wiley & Sons, New York.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663–685.

ILO (2004). *Child Labour: A Textbook for University Students*, International Labour Organization, Geneva.

Jayshi, D. (2005). *Conflict Pushes More Kids to Work in Nepal*, Dawn, Karachi.

Lehmann, E. L. (1998). *Testing Statistical Hypotheses*, Springer, Berlin.

Lohr, S. L. (1999). *Sampling: Design and Analysis*, Duxbury, New York.

Mohammad, S. M. (1999). Rao-Blackwell versions of the Horvitz-Thompson and Hansen-Hurwitz in adaptive cluster sampling, *Environmental and Ecological Statistics*, **6**, 183–195.

PCO. (1998). Population size and growth of major cities, *Pakistan Population*, from http://www.statpak.gov.pk/depts/pco/statistics/pop_major_cities/pop_major_cities.html.

Raj, D. (1966). Some remarks on a simple procedure of sampling with replacement, *Journal of the American Statistical Association*, **61**, 391–396.

Ray, R. (2000). Child labor, child schooling, and their interaction with adult labor: Empirical evidence for Peru and Pakistan, *World Bank Economic Review*, **14**, 347–367.

Ray, R. and Lancaster, G. (2004). Does child labour affect school attendence and school performance? Multi country evidence on SIMPOC data, *ILO*, Geneva.

Ross, S. (2002). *A First Course in Probability*, Prentice Hall.

SPARC (2005). *The State of Pakistan's Children 2004*, SPARC, Islamabad.

Sudman, S. (1997). Where have we been: Survey research 1967–1997, *Survey Research*, **28**.

Yates, F. (1960). *Sampling Methods for Censuses and Surveys*, John Wiley & Sons, New York.