

Life Table Analysis using Weighted Survey Data

James G. Booth* and Thomas A. Hirschl†

June 2005

Abstract

Formulas for constructing valid pointwise confidence bands for survival distributions, estimated using unequal probability samples, are derived. In addition, a test statistic for comparing estimated survival distributions from several groups is proposed. The methodology can be applied when survival times are subject to an independent censoring mechanism. The methods are illustrated using data from the Panel Survey on Income Dynamics on time until first home purchase, broken down by race and education level.

1 Introduction

Survey data is often collected, either intentionally or unintentionally, using an unequal probability sampling scheme. Individuals with certain characteristics are more or less likely to be included in the sample than others. In practice the selection probabilities for various subgroups may be unknown in advance, but may be estimated after the fact; for example, by calibrating post-stratification rates to known population proportions from the complete census. Information about the unequal selection probabilities is often provided to the public in the form of

*Department of Biological Statistics and Computational Biology, Warren Hall, Cornell University, NY 14853. Email: jb383@cornell.edu

†Department of Development Sociology, Cornell University

“sampling weights” which must be taken into account when estimating marginal properties of the population.

Throughout this paper we assume the following, idealized, probability sampling model. Suppose that the population of interest is divided into S strata, with a proportion π_s coming from stratum s . Individuals are selected by first choosing a stratum using probabilities τ_1, \dots, τ_S , and then selecting an individual at random from the chosen stratum. We suppose that the population is large enough that the difference between with and without replacement sampling is negligible. The sampling weights associated with individuals from the different strata are then proportional to the ratios, $\pi_s/\tau_s, g = 1, \dots, S$.

As noted above the selection probabilities must be taken into account when estimating marginal characteristics. For example, suppose we want to estimate the proportion of the population to whom a certain “event” has happened. Let E_i be an event indicator associated with the i th sampled individual. Then, the population characteristic of interest is $P = Pr(E_i = 1)$. It is straightforward to verify that an unbiased estimate of P based on a sample of size n is

$$\hat{P} = \frac{1}{n} \sum_{i=1}^n \frac{\pi_{s_i}}{\tau_{s_i}} E_i, \quad (1)$$

where s_i is the stratum from which the i th individual is selected. This is, in fact, the infinite population version of the famous Horwitz-Thompson estimator (see e.g. Cochran, 1977, Section 9.A.7), and is consistent for P as the sample size increases. If the probability ratios in (1) are unknown, but proportional sampling weights, $w_i \propto \pi_{s_i}/\tau_{s_i}, i = 1, \dots, n$, are available for sampled individuals, a consistent estimate of P is obtained as a ratio of two means,

$$\hat{P} = \frac{\frac{1}{n} \sum_{i=1}^n w_i E_i}{\frac{1}{n} \sum_{i=1}^n w_i}.$$

In this paper we develop methods for incorporating sampling weights into the analysis of discrete time-to-event data. Specifically, in Section 2 we modify the

standard survival function estimate based on a life table to account for individual sampling weights. In Section 3 we use the delta method to derive standard error formulas which may be used to construct pointwise confidence bands for the survival function, and in Section 4 we propose a test-statistic for comparing several survival functions based on data from different groups. We illustrate the proposed methodology in Section 5 using data on homeownership taken from the Panel Study of Income Dynamics (Hill, 1992). We conclude with some discussion of the appropriateness of our sampling model in practice.

2 Application to Life Table Analysis

In longitudinal surveys it is often of interest to estimate a “survival” curve for the population. What proportion of the population survive beyond the j th time interval without a particular event happening? This probability can be represented as

$$Q_j = q_1 q_2 \cdots q_j, \quad (2)$$

where q_r is the probability of surviving the r th interval, given that you survived until the end of the previous one. The distribution function for the time-to-event, corresponding to (2), is given by $P_j = 1 - Q_j, j = 1, 2, \dots$

Let R_{ij} be the indicator that the i individual is “at risk” at the beginning of interval j , and let E_{ij} be the indicator that the event occurs in interval j . Let R_j denote the set of individuals at risk (the risk set) at the beginning of interval j . Then, $q_j = 1 - p_j$, where

$$p_j = Pr(E_{ij} = 1 | i \in R_j) = \frac{Pr(E_{ij} = 1, i \in R_j)}{Pr(i \in R_j)} = \frac{Pr(E_{ij} = 1)}{Pr(i \in R_j)}. \quad (3)$$

It follows from (1) that a consistent estimate of p_j is given by

$$\hat{p}_j = \frac{\frac{1}{n} \sum_{i=1}^n \frac{\pi_{s_i}}{\tau_{s_i}} E_{ij}}{\frac{1}{n} \sum_{i=1}^n \frac{\pi_{s_i}}{\tau_{s_i}} R_{ij}} = \frac{\frac{1}{n} \sum_{i=1}^n w_i E_{ij}}{\frac{1}{n} \sum_{i=1}^n w_i R_{ij}}. \quad (4)$$

Clearly, the proportionality constant associated with the sampling weights cancels out in the calculation of \hat{p}_j in (4). Note that \hat{p}_j is a ratio of two sample means, $\hat{\mu}_j^E$ and $\hat{\mu}_j^R$, say, that are unbiased estimates of their corresponding population values μ_j^E and μ_j^R . Also, note that

$$\hat{q}_j = \frac{\sum_{i=1}^n w_i (R_{ij} - E_{ij})}{\sum_{i=1}^n w_i R_{ij}} = \frac{\sum_{i=1}^n w_i R_{i,j+1}}{\sum_{i=1}^n w_i R_{ij}}$$

implies

$$\hat{Q}_j = \frac{\sum_{i=1}^n w_i R_{i,j+1}}{\sum_{i=1}^n w_i}.$$

Thus, in the unweighted (unit weight) case, \hat{Q}_j is simply the proportion of individuals who survive until the beginning of the $(j + 1)$ st interval.

If some individuals are censored or withdraw during the j th interval, the estimate (4) will tend to underestimate p_j . Under the assumption that censoring times are independent of the times to the event of interest, we propose the following adjustment to \hat{p}_j ,

$$\hat{p}_j^* = \frac{\frac{1}{n} \sum_{i=1}^n w_i E_{ij}}{\frac{1}{n} \sum_{i=1}^n w_i R_{ij}^*} = \frac{\hat{\mu}_j^E}{\hat{\mu}_j^{R^*}},$$

where $R_{ij}^* = R_{ij} - C_{ij}/2$, and C_{ij} is a censoring/withdrawal indicator for the i th individual in interval j . This reduces to the standard actuarial adjustment in the unweighted case (Kalbfleisch & Prentice, 1980, Section 1.3). As in the unweighted case, \hat{p}_j^* is asymptotically biased as an estimator of p_j . However, as noted by Lawless (2003, Section 3.6.2), “the standard life table estimates are acceptable under random independent censorship provided that censoring is fairly evenly distributed across individual intervals and not too heavy.”

The survival function estimator discussed in this section can be computed using standard software by creating integer sampling weights, and treating them as observed frequencies (for example, using the frequency option in the SAS (2003)

lifetest procedure). This can be accomplished in practice by multiplying the sampling weights provided by a large positive constant and then rounding to the nearest integer. The problem with this approach is that it artificially inflates the sample size, with the result that standard errors and test statistics produced by the software are invalid.

3 Standard Errors and Confidence Bands

Since the estimate of Q_j is a product of ratios of means, standard errors for constructing confidence bands for the survival function can be estimated using the delta method (Stuart & Ord, 1994, Section 10.5-7). However, the application of the delta method with weighted survival data is slightly different from the unweighted case described in standard texts such as Kalbfleisch & Prentice (1980) or Lawless (2003). Specifically, since \hat{p}_j is a ratio of two means, an exact variance formula is not available. In contrast, in the unweighted case \hat{p}_j is (conditionally) a binomial proportion whose variance has a known form.

As in the unweighted case, the estimators, $\hat{p}_1, \dots, \hat{p}_j$, are asymptotically uncorrelated. To see this, define $H_i = \{(w_i E_{ij}, w_i R_{ij}, w_i C_{ij}), j = 1, 2, \dots\}$, and notice that H_1, \dots, H_n are i.i.d.. Also, define $\mathcal{H}(j) = \{(i, H_i), i \notin R_j\}$. Thus, $\mathcal{H}(j)$ represents the “history” of the sampled individuals up until the beginning of the j th interval. Now, the linear approximation,

$$\hat{p}_j \approx p_j + \frac{1}{\mu_j^R} (\hat{\mu}_j^E - p_j \hat{\mu}_j^R) = p_j + \frac{1}{n\mu_j^R} \sum_{i=1}^n w_i (E_{ij} - p_j R_{ij}), \quad (5)$$

implies, for $j < k$,

$$\begin{aligned}
\text{cov}(\hat{p}_j, \hat{p}_k) &\approx \frac{1}{n\mu_j^R \mu_k^R} E \{w_i(E_{ij} - p_j R_{ij}) w_i(E_{ik} - p_k R_{ik})\} \\
&= \frac{1}{n\mu_j^R \mu_k^R} E [E \{w_i(E_{ij} - p_j R_{ij}) w_i(E_{ik} - p_k R_{ik}) \mid \mathcal{H}(k)\}] \\
&= \frac{1}{n\mu_j^R \mu_k^R} E [w_i(E_{ij} - p_j R_{ij}) E \{w_i(E_{ik} - p_k R_{ik}) \mid \mathcal{H}(k)\}] \\
&= 0.
\end{aligned}$$

(Technically, this argument shows that $\text{cov}(\hat{p}_j, \hat{p}_k) = o(1/n)$.) The same argument works when there is censoring, with R replaced by R^* , and p_j replaced by p_j^* , the asymptotic limit of \hat{p}_j^* .

The approximation (5) leads to a variance formula for \hat{p}_j given by

$$\text{var}(\hat{p}_j) \approx \frac{1}{n(\mu_j^R)^2} E \{w_i(E_{ij} - p_j R_{ij})\}^2, \quad (6)$$

and its estimator

$$\widehat{\text{var}}(\hat{p}_j) \approx \frac{1}{(n\hat{\mu}_j^R)^2} \sum_{i=1}^n \{w_i(E_{ij} - \hat{p}_j R_{ij})\}^2. \quad (7)$$

To obtain a variance formula for the survival probabilities we use the linear approximation

$$\hat{Q}_j = \prod_{i=1}^j (1 - \hat{p}_i) \approx Q_j - \sum_{i=1}^j \left(\prod_{k \neq i}^j q_k \right) (\hat{p}_i - p_i).$$

Hence, since the \hat{p}_i 's are asymptotically uncorrelated, the delta method implies the approximate variance formula,

$$\text{var}(\hat{Q}_j) \approx \sum_{i=1}^j \left(\prod_{k \neq i}^j q_k \right)^2 \text{var}(\hat{p}_i) = Q_j^2 \sum_{i=1}^j \frac{\text{var}(\hat{p}_i)}{q_i^2}. \quad (8)$$

An asymptotic confidence interval for Q_j based on using (8) directly is $\hat{Q}_j \pm z \times \widehat{\text{se}}(\hat{Q}_j)$, where “se” denotes standard error, and z is a quantile of the standard

normal distribution. It is well known that this direct method can yield nonsensical limits, outside of the unit interval, especially when \hat{Q}_j is close to zero or one. Hence, it is standard practice (Kalbfleisch & Prentice, 1980, p.14) to calculate the interval for the transformed parameter $V_j = \log(-\log Q_j)$, which has an unrestricted range, and then transform back. This results in confidence limits for Q_j given by

$$\hat{Q}_j^{\exp[\pm z \times \widehat{\text{se}}(\hat{V}_j)]},$$

where $\text{se}(\hat{V}_j) = \text{se}(\hat{Q}_j)/|Q_j \log Q_j|$. Even with this modification it is possible that the confidence bands for the survival function may not be monotone, especially in the tail of the distribution where data may be scarce. Thus, we also suggest following the standard practice of enforcing monotonicity (Anderson et al., 1992, pp. 215-216).

4 Comparison of Distribution Functions

Suppose the subjects in the sample can be separated into G groups, for example, based on years of education. In this section we propose a test statistic for contrasting the distribution function estimates from the different groups. We use the same notation introduced in the previous section, except that an additional subscript “ g ” is added to identify the group. Thus, for example, p_{gj} is the probability that the event of interest occurs during the j th time interval for individuals in group g , conditional on their being at risk at the beginning of the interval. An estimate of this probability is given by the ratio $\hat{\mu}_{gj}^E/\hat{\mu}_{gj}^R$ (or $\hat{\mu}_{gj}^E/\hat{\mu}_{gj}^{R*}$, if there is some censoring).

The combined estimate of the event occurring in time interval j , based on the data from all the groups, is given by

$$\hat{p}_j = \frac{\sum_{g=1}^G n_g \hat{\mu}_{gj}^E}{\sum_{g=1}^G n_g \hat{\mu}_{gj}^R}. \quad (9)$$

A measure of the discrepancy between the distribution in a particular group and the combined distribution is given by the quantity

$$\Delta_g = \sum_j n_g (\hat{\mu}_{gj}^E - \hat{p}_j \hat{\mu}_{gj}^R). \quad (10)$$

Note that (9) implies that $\sum_g \Delta_g = 0$. Let $\Delta = (\Delta_1, \dots, \Delta_G)'$, and let Σ denote the estimate of the variance-covariance matrix of Δ , obtained using the delta method. (An asymptotic formula for Σ is given in the Appendix.) Then, a test of the hypothesis of no difference between groups (homogeneity) is obtained by comparing the statistic,

$$\chi^2 = \Delta' \hat{\Sigma}^- \Delta, \quad (11)$$

where Σ^- is a generalized inverse of Σ , to a chi-squared distribution with $G - 1$ degrees of freedom.

5 Application to Homeownership

The Panel Study of Income Dynamics (PSID) is a nationally representative, longitudinal sample of households and families interviewed since 1968 (Hill, 1992). It constitutes the longest running panel data set in the United States, and was specifically designed to track social dynamics over time. Since the PSID employs an unequal probability sampling scheme, each sampled individual is assigned a weight to correct for sampling bias.

In this section we use a subset of the PSID database to illustrate the methods described in this paper. Specifically, we focus on the distribution of the number of years to first home ownership among individuals who were renters at age 25, and how this distribution depends upon race (Black or White) and education (< 12, 12, or > 12 years). Summary statistics for the sampling weights in the six race/education categories are provided in Table 1. The overall sample sizes clearly

indicate that blacks were oversampled, and this is reflected in their lower average sampling weights. Also note that there is considerable amount of variability in the sampling weights within each race/education category.

	<i>Education Level</i>			Overall
	< 12	12	> 12	
<i>Blacks</i>				
Sample size	734	1017	498	2249
Mean weight	6.00	6.14	8.23	6.56
Std.Dev.	9.42	9.36	12.93	10.31
<i>Whites</i>				
Sample size	768	1375	1242	3385
Mean weight	26.46	27.90	30.46	28.51
Std.Dev.	9.83	9.22	8.34	9.19

Table 1: Summary statistics for individual sampling weights by race and years of education.

The upper panels in Figure 1 show the estimated distributions of time to first homeownership among whites and blacks along with 95% pointwise confidence bands. The distributions are clearly different, with blacks tending to take longer to purchase their first home. For example, about 80% of whites have been homeowners by age 35, compared with about 50% of blacks. The confidence bands are wider for the black distribution, in part because of the smaller sample size, but also because of the higher censoring rate (1088 blacks censored compared with 586 whites).

The lower panels in Figure 1 show the distribution function estimates broken down by education. Among whites, the distributions appear to be very consistent across education levels. The chisquared statistic discussed in Section 4 is 13.1 with 2 degrees of freedom in this case, indicating statistically significant differences among the education categories. This is not surprising given the relatively large sample sizes, and consequential high power of the test. There are much

larger differences in the distributions among blacks by education, resulting in a larger chisquared statistic of 19.9 with 2 degrees of freedom. Here the distributions appear to be stochastically ordered with the probability of homeownership increasing with education level, particularly in the 35 to 45 age range.

6 Discussion

We have proposed methods for utilizing sampling weights when analyzing time-to-event data, based on a simple unequal probability sampling model. While this model may not exactly reflect the sampling process, we contend that it serves as a reasonable probabilistic approximation under a variety of realistic sampling schemes. It is clear, for example, that the methods are valid for a stratified random sample. Suppose there are known to be N_s members of stratum s , and it was possible to identify members of the various strata in advance. If a predetermined number, n_s , of individuals from stratum s are selected, then the estimate of \hat{p}_j in (4), and its asymptotic variance given in (6), remain unchanged with π_s and τ_s replaced by the corresponding proportions, N_s/N and n_s/n .

In practice, separate sampling frames for the different strata will typically not be available. A more realistic sampling scheme might involve taking 1 in k_d samples (of say households) in a collection of districts or regions, $d = 1, \dots, D$. The value of k_d may be varied in order to oversample certain groups based on known demographics for the different districts. With this scheme the sample sizes, n_s , are unknown in advance, and so the marginal selection probabilities, τ_s , $s = 1, \dots, S$, must be estimated based on post-stratification rates. Methods that account for the estimation of the sampling weights would be considerably more complicated than those discussed in this paper. It is not clear to what extent such an extension is important, but it is an interesting topic worthy of further study.

Appendix

An asymptotic formula for the variance-covariance matrix, Σ , defined in Section 4 can be obtained using the delta method. Note that $\Delta_g = \sum_j \Delta_{gj}$, where

$$\Delta_{gj} = n_g(\hat{\mu}_{gj}^E - \hat{p}_j \hat{\mu}_{gj}^R) \approx n_g(\hat{\mu}_{gj}^E - p_j \hat{\mu}_{gj}^R - \hat{p}_j \mu_{gj}^R) + \text{constant},$$

and that

$$\begin{aligned} \hat{p}_j &= \frac{\hat{\mu}_j^E}{\hat{\mu}_j^R} \approx p_j + \frac{1}{\mu_j^R}(\hat{\mu}_j^E - p_j \hat{\mu}_j^R) + \frac{1}{\mu_j^R}(\mu_j^E - p_j \mu_j^R) \\ &= p_j + \frac{1}{\mu_j^R} \sum_{g=1}^G \frac{n_g}{n} (\hat{\mu}_{gj}^E - p_j \hat{\mu}_{gj}^R) + \text{constant}. \end{aligned}$$

It follows that,

$$\Delta_{gj} \approx n_g \left\{ (\hat{\mu}_{gj}^E - p_j \hat{\mu}_{gj}^R) - \frac{\mu_{gj}^R}{\mu_j^R} \sum_{k=1}^G \frac{n_k}{n} (\hat{\mu}_{kj}^E - p_j \hat{\mu}_{kj}^R) \right\} + \text{constant}. \quad (\text{A.1})$$

The arguments in Section 3 imply that

$$\text{cov}(\hat{\mu}_{gj}^E - p_j \hat{\mu}_{gj}^R, \hat{\mu}_{gk}^E - p_k \hat{\mu}_{gk}^R) \approx 0.$$

Hence, equation A.1 implies $\text{var}(\Delta_g) \approx \sum_j \text{var}(\Delta_{gj})$ and $\text{cov}(\Delta_g, \Delta_h) \approx \sum_j \text{cov}(\Delta_{gj}, \Delta_{hj})$, where

$$\begin{aligned} \text{var}(\Delta_{gj}) &= n_g^2 \left\{ \left(1 - \frac{n_g \mu_{gj}^R}{n \mu_j^R} \right)^2 \text{var}(\hat{\mu}_{gj}^E - p_j \hat{\mu}_{gj}^R) \right. \\ &\quad \left. + \left(\frac{\mu_{gj}^R}{\mu_j^R} \right)^2 \sum_{k \neq g} \left(\frac{n_k}{n} \right)^2 \text{var}(\hat{\mu}_{kj}^E - p_j \hat{\mu}_{kj}^R) \right\}, \end{aligned}$$

and

$$\begin{aligned} \text{cov}(\Delta_{gj}, \Delta_{hj}) &= n_g n_h \left\{ \frac{\mu_{gj}^R \mu_{hj}^R}{(\mu_j^R)^2} \sum_{k=1}^G \left(\frac{n_k}{n} \right)^2 \text{var}(\hat{\mu}_{kj}^E - p_j \hat{\mu}_{kj}^R) \right. \\ &\quad \left. - \frac{n_g \mu_{gj}^R}{n \mu_j^R} \text{var}(\hat{\mu}_{gj}^E - p_j \hat{\mu}_{gj}^R) - \frac{n_h \mu_{hj}^R}{n \mu_j^R} \text{var}(\hat{\mu}_{hj}^E - p_j \hat{\mu}_{hj}^R) \right\}. \end{aligned}$$

References

- ANDERSON, P., BORGAN, O., GILL, R. & KIEDING, N. (1992). *Statistical Models Based on Counting Processes*. Springer.
- COCHRAN, W. G. (1977). *Sampling Techniques*. Wiley, 3rd ed.
- HILL, M. S. (1992). *The Panel Study of Income Dynamics: A Users Guide*. Newbury Park, CA: Sage Publications.
- KALBFLEISCH, J. D. & PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley.
- LAWLESS, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley, 2nd ed.
- SAS (2003). Lifestest procedure. Cary, NC: SAS Institute Inc.
- STUART, A. & ORD, K. (1994). *Kendall's Advanced Theory of Statistics*, vol. 1. Edward Arnold, 6th ed.

Distribution of years to first homeownership

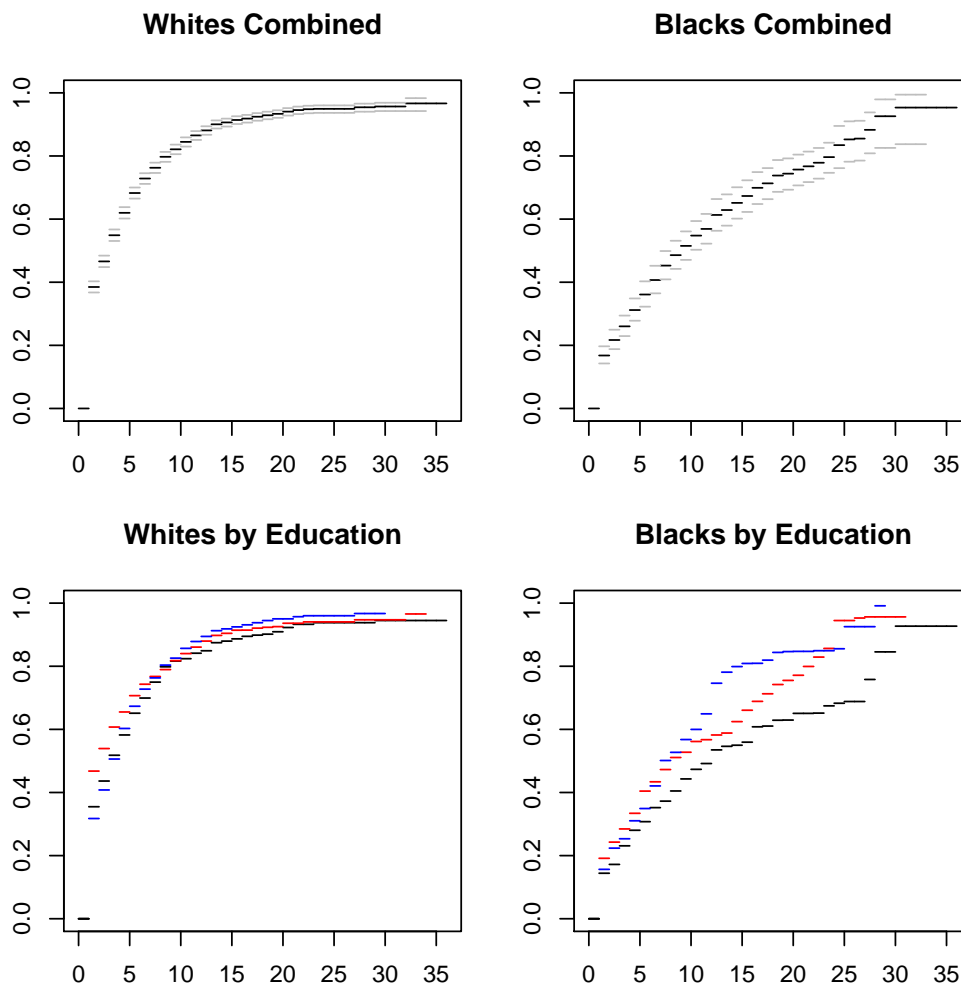


Figure 1: Upper panels show estimated distribution functions with pointwise 95% confidence limits. Lower panels show estimates for three education categories: highschool dropout (black), highschool graduate (red), and some college (blue)