Using Survival Analysis to Predict Sample Retention Rates

December 2006

Andy Sadler and Larry Lang
U.S. Bureau of Labor Statistics
2 Massachusetts Avenue NE, Washington DC 20212 U.S.A
Sadler.Andy@bls.gov

Key Words: Price index; Retention rates; Quote allocation; Life-Table estimate; Survivor function; Censored observations.

1. Introduction and Motivation

The International Price Program (IPP) publishes monthly price indexes which measure the average change over time in prices of internationally traded products and services. IPP's sampling methodology supports its published merchandise price indexes by attempting to maintain a minimum number of items over the two year life cycle [1] of each published stratum. Retention rates -- which measure the proportion of sampled items in a stratum which are still active (non-discontinued) after a certain amount of time past the sample fielding date -- are used to set the minimum stratum item allocation.

Stratum retention rates are currently calculated by taking an average of the number of items for which a response was received (active or non-discontinued items) divided by the number of fielded items around each period of interest. This approach calculates exact historical averages of the retention rates and does not make optimal use of the available information.

The major objective of this paper is to develop a general procedure to model stratum retention rates. We use statistical techniques for survival data to estimate stratum retention rates over time.[2] We calculate survival functions and hazard functions for the four different classification systems at which IPP publishes price indexes: Harmonized System (HS), Bureau of Economic Analysis End Use (BEA), North American Industrial Classification System (NAICS) and the Standard International Trade Classification System (SITC). We find that the probability of item survival decreases sharply immediately after the item has been fielded before leveling off and decreasing more slowly.

The paper proceeds as follows. We begin with a brief overview of the International Price Program (IPP) and the various indexes that we produce. Next, we discuss the sampling process used by IPP to select its import and export samples. We then describe the data used in our analysis and the statistical techniques employed in analyzing the data. Finally, we discuss the results of our study.

2. Background

The International Price Program (IPP) of the Bureau of Labor Statistics (BLS) produces two of the major price indexes for the United States: the Import Price Indexes and the Export Price Indexes. These indexes, along with the BLS's two other monthly price indexes -- the Producer Price Indexes and the Consumer Price Indexes (CPI) -- provide a complete description of price trends in the U.S. economy. Import and Export Price Indexes are used in a variety of ways including deflating U.S. trade data, measuring price changes and trends in the foreign trade sector of the U.S. economy, measuring international competitiveness, and measuring the economic effects of exchange rate movements. The IPP, as the primary source of data on price change in the foreign trade sector of the U.S. economy, publishes index estimates of price change for internationally traded goods using four different classification systems - Harmonized System (HS), Bureau of Economic Analysis End Use (BEA), North American Industrial Classification System (NAICS) and the Standard

[1] A stratum's life cycle describes a series of stages -- in terms of the number of non-discontinued items remaining within the stratum -- through which the stratum passes. The number of non-discontinued items for a given published stratum reaches a minimum at about 2.5 and 4.5 years after sample fielding. (At the 2.5 year and 4.5 year mark the number of items entering a stratum from newly fielded samples is greater than the number of items falling out of the stratum.) The number of items allocated in each publishable stratum is done in such a way as to ensure that at the low point in a stratum's life cycle, the number of items will not fall below 20.

[2] The concept of item survival is a useful means of describing the retention of items within strata. The survival rate for a set of items within a stratum is defined as that proportion of the items which remain active (non-discontinued) after a given number of months.

International Trade Classification System (SITC)[3]. IPP also publishes selected services indexes and goods indexes based upon the country or region of origin. This paper uses information on the Import and Export goods samples that IPP uses to collect its price data.

3. Sampling

IPP's import merchandise sampling frame is obtained from the U.S. Customs Service and our export merchandise sampling frame is a combination of data obtained from the Canadian customs service for exports to Canada and from the Bureau of the Census for exports to the rest of the world.

The frames contain information about all import or export transactions that were filed with the U.S. Customs Service during the reference year (or Canadian customs service for exports to Canada). The frame information available for each transaction includes a company identifier (usually the Employer Identification Number), the detailed product category (Harmonized Tariff number for Imports and the Schedule B number for Exports) of the goods that are being shipped and the corresponding dollar value of the shipped goods.

Starting in 1989, IPP divided the import and export merchandise universes into two halves referred to as panels. Samples for one import panel and one export panel are selected each year and sent to the field offices for collection, so both universes are fully re-sampled every two years. The sampled products are priced for approximately five years until they are replaced by a fresh sample from the same panel. As a result, each published index is based upon the price changes of items from up to three different samples[4].

Each panel is sampled every other year using a three stage sample design. The first stage selects establishments independently proportional to size (dollar value) within each broad product category (stratum) identified within the harmonized classification system (HS).

The second stage selects detailed product categories (classification groups) within each establishment - stratum using a systematic probability proportional to size (PPS) design. The measure of size is the relative dollar value adjusted to ensure adequate coverage for all published strata across all classification systems (HS, BEA, SITC and NAICS) and known non-response factors (total company burden and frequency of trade within each classification group). Each establishment – classification group (or sampling group) can be sampled multiple times and the number of times each sampling group is selected is then referred to as the number of quotes requested.

In the third and final stage, the BLS Field Economist, with the cooperation of the company respondent, performs the selection of the actual item for use in the IPP indexes. Beginning with these entry-level classification groups and the list of items provided by the respondent to the field economist, further stages of sampling are completed until one item for each quote sampled in the classification group is selected. This process is called disaggregation. This process is done with replacement, so the same item can be selected more than once.

4. Methods

We are interested in studying the lengths of time between when an item is fielded, and when that item is either no longer being traded or the establishment trading the item has gone out of business -- events we will refer to as discontinuations. Our analysis time -- which measures the length of time that an item remains in the sample -- begins at sample fielding and ends when the item is discontinued or censored. Because we often do not know whether an item is no longer active due to discontinuation or for some other reason, we need to account for censoring in our analysis. We begin with a brief discussion of the way retention rates are currently calculated and used.

4.1 ***Current method for calculating and using stratum retention rates***

Currently, stratum retention rates are calculated in the following way:

$$Ret\ rate\ for\ stratum\ k = \frac{\left(\frac{\sum_{sample\ i} \sum_{month\ j} non-disc items_{ijk}}{3} \right)}{\sum_{sample\ i} items\ fielded_{ik}}$$

For the 2.5 year rate for each stratum, the number of non-discontinued items is calculated at 29, 30, and 31 months past the sample fielding date for each of the active samples containing the stratum. For the 4.5 year rate, the number of non-discontinued items is calculated at 53, 54, and 55 months past the sample fielding date for each of the active samples containing the stratum. The retention rate for each of these periods is calculated by dividing the number of non-discontinued items by

[3] The IPP will cease publishing SITC indexes in July, 2006.

[4] Indexes for published strata that cross panels, may be based upon items from up to six samples at any one time.

the number of items fielded within each stratum. The overall 2.5 year retention rate is calculated by taking the average retention rate across all 3 months (29, 30, 31) and across all active samples for each stratum. A similar procedure is used in calculating the 4.5 year retention rate.

Stratum retention rates are used in calculating minimum stratum item allocation levels prior to fielding the sample. Stratum item allocations are done using the results from the following formula:

$$\frac{20}{2.5\,year\,str\,ret\,rate + 4.5\,year\,str\,ret\,rate}.$$

We use 20 items in the numerator of the formula because it has been determined that this is the minimum number of items needed to maintain a stable index for each stratum. We use stratum retention rates calculated at 30 months (2.5 years) and 54 months (4.5 years) after sample fielding in the denominator of the formula because at these points the number of non-discontinued items within the stratum reaches a minimum in the sample rotation cycle. Using this formula to make stratum item allocations ensures that at the 2.5 year and 4.5 year marks (the low points for the total number of items in the stratum in the sample rotation cycle), the number of items in the stratum will not fall below 20.

4.2 *Proposed Methodology: Survival function calculations*

We now turn to models estimated using censored data. The survivor and hazard functions are usually of central interest when summarizing survival data.[5] The actual survival time of an item, t, can be regarded as the value of a random variable, T, which can take any non-negative value. T is the random variable associated with the survival time. The survivor function, S(t), is defined to be the probability that the survival time is greater than or equal to t. The survivor function can therefore be used to represent the probability that an item survives from the time origin to some time beyond t.

An important assumption of survival analysis is that censoring is non-informative, i.e., the censoring time of an item is independent of the survival time of the item. If we define C to be the time at which censoring of an item occurs, then an item is censored if T > C and uncensored if T < C.

We used the life-table method to estimate the survivor function because our event times are grouped into monthly intervals. For this analysis, we divided our data into 72 intervals each one month in length. We chose 72 months because this is the approximate time from when a sample is fielded to when the sample is completely phased out. We assume that the j^{th} interval, $j = 1, 2, \ldots 72$, extends from t'_j to t'_{j+1}, and let d_j and c_j denote the number of discontinued items and the number of censored survival times, respectively, in this time interval. Let n_j be the number of items which are still active (non-discontinued), and therefore at risk of discontinuation, at the start of the j^{th} interval. Also, let $n'_j = n_j - \frac{c_j}{2}$ be the average number of non-discontinued items which are at risk of discontinuation in the j^{th} interval.[6] The life-table estimate of the survivor function is given by

$$\hat{S}(t) = P(T \geq t) = \prod_{j=1}^{k} \left(\frac{n'_j - d_j}{n'_j} \right),$$

for $t'_k \leq t < t'_{k+1}$, $k = 1, 2, \ldots, 72$.

In the j^{th} interval, the probability of item discontinuation can be estimated by d_j / n'_j, so that the corresponding survival probability is $(n'_j - d_j)/n'_j$. The probability that an item survives beyond time t'_k, $k = 1, \ldots, 72$, is the product of the probabilities that the item survives beyond the start of the k^{th} interval and through each of the $k-1$ preceding intervals which gives the survival function. The estimated probability of surviving until the start of the first interval is one, while the estimated probability of surviving beyond the last interval is zero.

---

[5] Even though estimating the hazard function was not essential to our objective of modeling retention rates and calculating stratum item allocations, we've included it in our paper to (1) illustrate the relationship between the survivor function and the hazard function; and (2) show how the risk of item discontinuations change over the period of our study.

[6] The life-table method treats any cases censored within an interval as if they were censored at the midpoint of the interval. This treatment is equivalent to assuming that the distribution of censoring time is uniform within the interval. Since censored cases are at risk for only half of the interval, they count for only half in figuring the effective sample size.

The life-table estimate is robust to censoring and uses information from both censored and non-censored observations. The estimated variance of the life-table estimator is given by Greenwood's formula

$$\hat{Var}\left\{\hat{S}(t)\right\} \approx \left[\hat{S}(t)\right]^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}.$$ (Collett, 2003, p.25)

4.3 ***Hazard function calculations***[7]

The life table estimate of the hazard function is estimated by taking the ratio of discontinued items in a given month to the average time survived in that month to give an estimate of the rate of item discontinuations per month. Let the number of discontinued items in the j^{th} interval be d_j, j = 1, 2, …, 72 and let n'_j be the average number of non-discontinued items at risk of discontinuation in the interval. If we assume that the discontinuation rate is constant during the j^{th} month, the average time survived in that interval is $\left(n'_j - \frac{d_j}{2}\right)t_j$, where t_j is the length of the j^{th} time interval (one in this case). The life-table estimate of the hazard function in the j^{th} interval is then given by

$$\hat{h}(t) = \frac{d_j}{(n'_j - \frac{d_j}{2})t_j}, \text{ for } t'_j \le t < t'_{j+1}, j = 1,2,\ldots,72.$$

Because the hazard function measures the rate of item discontinuation per unit time, we expect that when the probability of item survival is low, i.e., when the survival function is decreasing rapidly, the hazard function should be increasing. Likewise, when the estimated survival function decreases more slowly or levels off, the hazard function will show a tendency to fall or rise less dramatically.

4.4 ***Sample retention rates and average sample retention rates***

For comparison purposes, we also calculated the sample retention rates and average sample retention rates to demonstrate the differences between the estimated survival probabilities and the retention rates calculated using a process similar to the current averaging method.

Sample retention rates for the k^{th} stratum in sample i are calculated as follows:

$$\frac{\sum_{Sample\,i} non-disc\,Items_{ik}}{\sum_{Sample\,i} items\ fielded_{ik}}$$

Average sample retention rates for the k^{th} stratum across n samples are calculated as follows:

$$\frac{\left(\frac{\sum_{Sample\,1} non-disc\,items_k}{\sum_{Sample\,1} items\ fielded_k} + \ldots + \frac{\sum_{Sample\,n} non-disc\,items_k}{\sum_{Sample\,n} items\ fielded_k}\right)}{n}$$

5. Data

Our analysis is based on U.S. import and export statistics as compiled by the International Price Program's Unified Database (UDB). Data on non-discontinued items from eight import and eight export samples -- which were fielded, and initiated during the period 1997 to 2004 -- were compiled and aggregated to the Harmonized System (HS), SITC, BEA and NAICS classification systems.

Survival analysis methods apply whenever we are interested in and examining the time to occurrence of an event. The main feature of survival data is the presence of *censored* observations. Censoring occurs when an item is observed for some period of time without the event of interest (item being discontinued) occurring. In compiling the survival data used in the analysis, censoring occurred for:[8] Items that were

[7] In calculating the hazard functions, we decided to use a local averaging technique -- called smoothing -- to summarize the trend of the hazard rate. This technique was implemented using a SAS macro contained in Paul D Allison's "Survival Analysis Using SAS" manual. The macro employs a kernel smoothing method described by Henrik Ramlau-Hansen (1983), "Smoothing Counting Process Intensities by Means of Kernel Functions," The Annals of Statistics 11, 453-466. The smoothed hazard function produces an estimate of the trend that is less jagged than the hazard rate itself, is more visually appealing and helps us to pick out the trend in the plot. Our calculations used a bandwidth of 7 months arbitrarily chosen. This means that data points more than 7 months away from the point being estimated are not used in the smoothing function.

[8] There are three types of possible censoring schemes -- right censored, interval censored and left censored. The

phased out[9] and items which had not been discontinued at the time the analysis was being done.

6. Results[10]

Survival function and smoothed hazard function results for HS import strata P18-9032 and P10-49

Figures 1 and 3 display estimates of the survival probabilities with 95% confidence intervals, and average sample retention rates for HS import strata P18-9032 (Parts of automatic regulating or controlling instruments), and P10-49 (Printed matter). Figures 2 and 4 display the estimated smoothed hazard functions for HS import strata P18-9032 and P10-49.

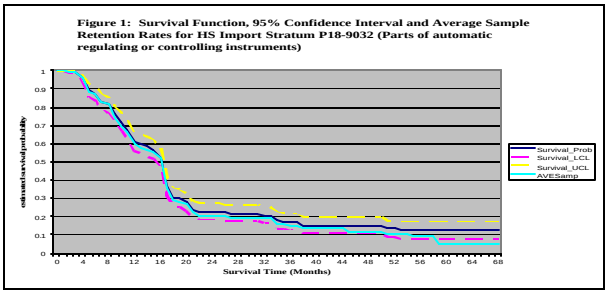
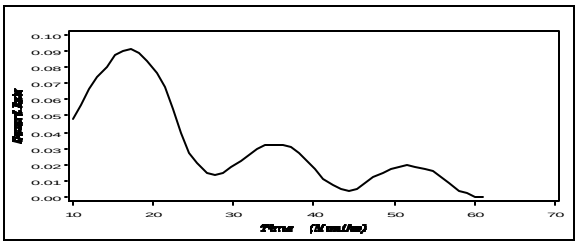


Figure 2: Smoothed Hazard Function Estimate for HS Import Stratum P18-9032



The estimated survival function in Figure 1 for HS import stratum P18-9032 decreases sharply for the first 18 months before leveling off and decreasing at a slower rate. The trend of the estimated survival function in Figure 1 suggests that the hazard rate should be an increasing function during the first 18 months and this is confirmed in Figure 2. The failure rate as shown in Figure 2 is quite high in the first 18 months after fielding but decreases appreciably thereafter. Items in HS import stratum P18-9032 are less likely to drop out of the sample once they have survived the first 18 months. The average sample retention rate for HS import stratum P18-9032 is contained in the 95% confidence interval up to 59 months after fielding. Thereafter, the average sample retention rate curve falls below the 95 % confidence band. This occurs because items are censored at these points and are not used in calculating the average sample retention rate.

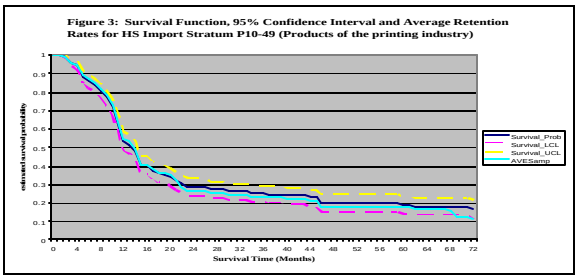
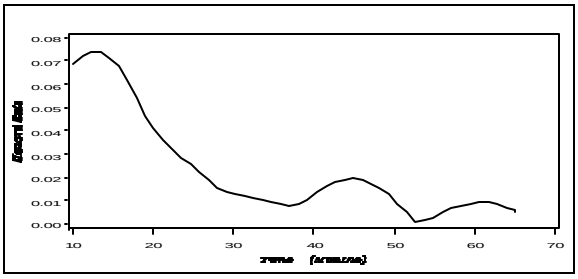


Figure 4: Smoothed Hazard Function Estimate for HS Import Stratum P10-49



The estimated survival function in Figure 3 for HS import stratum P10-49 decreases sharply for the first 14 months before leveling off and decreasing at a slower rate. The trend of the smoothed hazard function for HS import stratum P10-49 as shown in Figure 4 shows an increasing slope up to the 14 month mark before decreasing sharply. Item failure rate for HS import stratum P10-49 is quite high in the first 14 months after fielding but decreases appreciably thereafter. Items in HS import stratum P10-49 are less likely to drop out of the sample once they have survived the first 14 months. The average sample retention rate for HS import stratum P10-49 is contained in the 95% confidence interval up to 69 months after fielding. Thereafter, the average sample retention rate curve falls below the 95 % confidence band because censored data are not used in the calculations.

---

type of censoring observed in our data is right censoring. Right censoring implies that the event of interest (*i.e.* the time-to-item discontinuation) is to the right of the item's last known survival time.

[9] Phaseout is a term used to indicate the month and year when an item is scheduled to be discontinued from repricing.

[10] We've limited the discussion of our results to four Harmonized Classification System (HS) and BEA Classification System strata. These strata were chosen to highlight the salient features of our findings.

Survival function and hazard function results for BEA export strata Q21500 and Q21150

Figures 5 and 7 display estimates of the survival probabilities with 95% confidence intervals, and average sample retention rates for BEA export strata Q21500 (Business machinery and equipment), and Q21150 (Pulp and paper machinery). Figures 6 and 8 display the estimated smoothed hazard functions for BEA export strata Q21500 and Q21150.

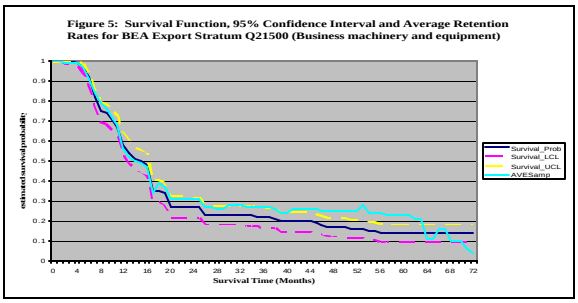
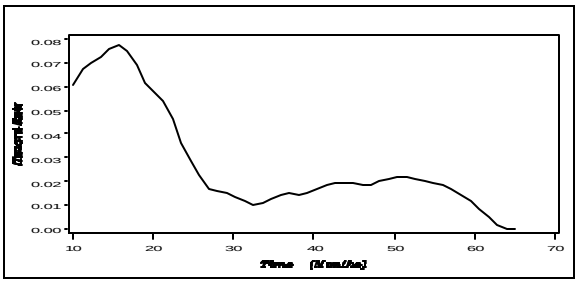


Figure 6: Smoothed Hazard Function Estimate for BEA Export Stratum Q21500



The estimated survival function in Figure 5 for BEA export stratum Q21500 decreases sharply for the first 16 months before leveling off and decreasing at a slower rate. The trend of the estimated survival function for BEA export stratum Q21500 suggests that the hazard rate should be an increasing function during the first 16 months and this is confirmed in Figure 6. The failure rate for BEA export stratum Q21500 is quite high in the first 16 months after fielding but decreases appreciably thereafter. Items in BEA export stratum Q21500 are less likely to drop out of the sample once they have survived the first 16 months. The average sample retention rate for BEA export stratum Q21500 is contained in the 95% confidence interval up to 26 months after fielding. Thereafter, the average sample retention rate curve goes intermittently above and below the 95 % confidence band when items are censored and are not used in the calculations.

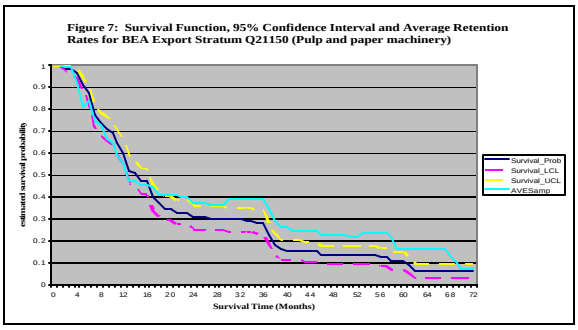
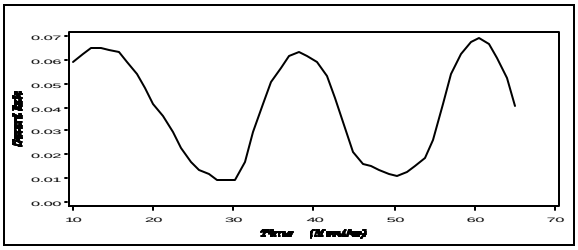


Figure 8: Smoothed Hazard Function Estimate for BEA Export Stratum Q21150



The trend of the estimated survival function for BEA import stratum Q21150 is markedly different from the other strata studied so far. The estimated survival function shown in Figure 7 displays a clear tendency to decrease rapidly during certain periods. We see that the estimated survival function for BEA export stratum Q21150 experienced sharp declines for periods corresponding to: fielding to 13 months; the period spanning months 31 to 38 after fielding; and 50 months to 60 months after fielding. These cyclical trends are reflected in the shape of the smoothed hazard function shown in Figure 8 which increases and decreases during periods of rapid or slowing retention. The average sample retention rate for BEA export stratum Q21150 is contained in the 95% confidence interval up to 21 months after fielding. Thereafter, the trend of the average sample retention rate curve goes above the 95 % confidence band before falling back into the band after 69 months.

6.1 Allocations

Minimum stratum item allocations were calculated using both the averaging process and also the estimated survival probabilities from the lower bound on the 95% confidence interval. This table compares minimum stratum item allocations calculated using these two methods. The results for the four strata discussed are contained in the table below.

The average 2.5 year and 4.5 year retention rates were calculated using the formulas discussed in the *Methods* section of this paper. The calculations for the minimum stratum item allocations were done using the formula for item allocations discussed previously. The

allocations displayed in the column labeled "*Item allocation based on stratum dollar value*" are the actual allocations which were made on the basis of the stratum's dollar value of trade in the most recently fielded sample. In cases where the number of items allocated to a stratum -- based on that stratum's dollar value of trade -- is less than the minimum allocation calculated from the item allocation formula we increase that stratum's item allocation to match the number calculated from the allocation formula.

We see from the table that item allocations based on the estimated survival probabilities are always greater than those determined from the averaging process. This singular fact highlights the benefit of using the estimated survival probabilities: larger minimum item allocations will ensure that rapidly deteriorating strata will receive more items thus maintaining their publishability status.

7. Conclusion

There are three important conclusions to draw based on our results. First, most items within our samples are discontinued during the early stage of the repricing process. Once an item has survived for the first 15 - 20 months it has a greater chance of remaining in the sample until phaseout.

Second, the results show that the survival model provides a superior fit to the data than using the average rates which do not utilize information provided by censored cases. This important finding allows us to allocate items to strata more effectively especially in strata which experience high retention rates.

Third, the survival model provides us with the option of using the Lower Confidence Limit (LCL) which gives a more conservative estimate of the survival rates. Using the LCL may guard against unexpected results -- for example samples being pre-maturely phased out -- or results that do not match historical retention patterns.

Stratum	2.5 year retention rate		4.5 year retention rate		Minimum stratum item allocation		Item allocation based on stratum dollar value[11]	Actual stratum allocation	
	ave rate	surv prob	ave rate	surv prob	ave alloc	surv alloc		ave alloc	surv alloc
P18-9032	.1959	.1729	.1011	.0813	67	78	81	81	81
P10-49*	.2437	.2138	.1839	.1533	47	54	50	50	54
Q21500*	.2815	.1818	.245	.108	38	69	65	65	69
Q21150	.3919	.2420	.2388	.0899	32	60	70	70	70

* shows strata whose item allocations would increase using the estimated survival probabilities to calculate allocation levels.

[11] Item's are generally allocated to strata based on trade dollar values. Larger strata are allocated more items than smaller strata.

References

Bureau of Labor Statistics (1997). BLS Handbook of methods. Washington, DC: U.S. Department of Labor.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. Chapman & Hall.

Allison, Paul D (1995). *Survival Analysis Using SAS: A Practical Guide*. SAS Publishing.

The IPP Lower Level Weights Team (2004), Lower Level Weights Proposal.

Acknowledgements

The views expressed in this paper are those of the authors and do not reflect the policies of the U.S. Bureau of Labor Statistics.