Alternative Allocation Designs for a Highly Stratified Establishment Survey December 2007

Ernest Lawley[1], Marie Stetster[1], Eduardas Valaitis[2]
[1]US Bureau of Labor Statistics, 2 Massachusetts Ave NE, Suite 4985, Washington, DC, 20212
[2]American University, 4400 Massachusetts Avenue NW, Washington, DC, 20016

Abstract

The primary objective of Occupational Employment Statistics Survey, conducted by U.S. Bureau of Labor Statistics in partnership with the 50 States and District of Columbia, is to measure occupational employment and wages at the very detailed level of Metropolitan Statistical Areas (MSA) crossed by over 300 industries. That is, how many people are employed in one of the 800 Standard Occupational Codes (SOC) and what are the mean occupational wages for each industry by MSA. A given sampling frame contains about 175,000 non-empty MSA-by-industry cells. The occupational employment and wage estimates are also required at various aggregated levels of geography and industry. This study examines alternative sample allocation designs for a highly stratified population that deals with multiple issues such as establishment employment size and occupational diversity and variability.

KEY WORDS: Neyman allocation, probability proportional to estimated size (PPES), power allocation

1. INTRODUCTION

The primary objective of Occupational Employment Statistics Survey (OES), conducted by U.S. Bureau of Labor Statistics[1] (BLS) in partnership with the 50 States and four territories (District of Columbia, Guam, Puerto Rico, and the US Virgin Islands) is to measure occupational employment and wages at the very detailed level of Census-defined Metropolitan Statistical Areas (MSA) crossed by over 300 industries. That is, how many people are employed in each of the 801 Standard Occupational Codes [2] (SOC) and what are the mean occupational wages for each industry by MSA. A given sampling frame contains about 175,000 non-empty MSA-by-industry strata. The occupational employment and wage estimates are also required at various aggregated levels of geography and industry.

From the frame, the probability-proportional-to-size (PPS) allocation design generally allocates larger sample to those strata that have large populations and allocates smaller sample to those strata that have smaller populations. However, there are disadvantages to the PPS allocation design in view of an employment survey; notably, that occupational variability within each stratum is disregarded and that allocation is heavily focused on strata (particularly MSAs) with large populations.

An advantageous allocation design will allow for the balancing of particular variables within each stratum, specifically: stratum population, stratum occupational variability, and a desirable "spread" of sample allocation between all strata.

2. DEFINING STRATA

Most of the OES Frame is obtained from what is known as the BLS Longitudinal Database (LDB). Each business establishment reports employment and aggregated wage data to its respective State Unemployment Insurance (UI) program. BLS then aggregates all of these reporting business establishments (each containing data such as business name, business address, business location, past 12 monthly employment counts, and total aggregated wages paid) into the LDB. Reporting data are also collected from Guam (which is not included on the LDB) and all business establishments in the Railroad Industry (which may contain incomplete data on the LDB). These three file inputs are then all concatenated to create the OES Frame containing approximately 6.5 million business establishments.

After the OES Frame is created, it is stratified by geography, then by industry. Geographical stratification is done by State then by MSA. MSAs that cross States boundaries (such as Kansas City) are split into two or more parts, depending on how many States the MSA encompasses. Areas within each State that are not encompassed by an MSA are defined as Balance of State (BOS); each State generally has up to four BOS areas. BOS areas can be viewed as "rural" areas within each State. BOS divisions are determined by each individual State. There are approximately 600 State|MSA-BOS areas nationwide. The frame is then further stratified by Industry. Industries are defined by the North American Industry Classification System (NAICS); OES utilizes NAICS at the 4-digit level (and selected 5-digit level).

[1] Views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the Bureau of Labor Statistics.

[2] To define occupations, OES uses the Office of Management and Budget (OMB) Standard Occupation Classification (SOC) system; the SOC system is required to be used by federal agencies. OES uses 22 major occupations groups from the SOC to categorize each worker into 1 of 801 detailed occupations. Workers who have duties encompassing multiple occupations are categorized into the single occupation that contains the greatest preponderance of occupational duties.

There are approximately 300 of these industry groups. Each business establishment in the OES Frame belongs to one of approximately 175,000 non-empty cells stratified by geography and industry. After each and every business establishment has been assigned to a stratum, a sample of approximately 1.2 million establishments is allocated. Each of the 54 States and Territories are given a fixed sample size to allocate. These values were implemented in 1996 and were based on employment population. Allocation for each State and Territory must come within ½% of each State's or Territory's respective fixed value. These values, when summed, equal approximately 1.2 million.

3. METHODOLOGY

There are various methods to allocate a sample. The most basic method is a simple random sample model, which encompasses randomly choosing sample subjects from a "universe" of data. In this context, though, it is assumed that all subjects in the universe need not be stratified. By not stratifying data, it is highly probable that a simple random sample model will yield results that contain large differences between sample estimates and actual true values (known as sampling error). Thus, a more complex stratified model should be utilized once data has been stratified (or "grouped"). *Model Assisted Survey Sampling* by Sarndal, Swenson, and Wretman (1992) states that "in stratified sampling, the population is divided into non-overlapping subpopulations called strata. A probability sample is selected in each stratum. The selections in the different strata are independent. Stratified sampling is a powerful and flexible method that is widely used in practice."

A common stratified sample allocation design is the probability proportional to size (PPS) model. In our model, we use employment for the size value and establishment for the sample unit.
Given a state:

$$n_h = n \bullet \frac{EMP_h}{\sum EMP_h} \qquad (3.1)$$

n_h = number of establishments allocated to stratum h
n = fixed establishment allocation for each state
EMP_h = total number of employees at stratum h (obtained from the LDB)
ΣEMP_h = total number of State (or Territory) employees (obtained from the LDB); summed for all strata within a State (or Territory)[3]

[3] OES defines employment count for each observation as the maximum value of the 12 monthly employment counts in each business establishment. Maximum employment is useful in helping define the size of an establishment; using maximum employment eliminates having to make adjustments due to seasonality and minimizes sample weights.

Although it is nice that this model is simple, it is apparent that the PPS allocation design is not an adequate design for an employment survey that produces detailed estimates for local areas where strata include occupational variability. The PPS model contains no measures for variability. As a consequence, it is a model that cannot adequately allocate for anything beyond overall population values. That is, it only allocates proportional to stratum population without considering other factors within the stratum, notably occupational variability. Additionally, the PPS design may be good for national estimates, but it causes problems when local area estimates of approximately equal reliability are desired. *Model Assisted Survey Sampling* states that a more favorable allocation for a stratified population is the *Neyman Allocation.*

$$n_h = n \bullet \frac{EMP_h \bullet S_h}{\sum \left(EMP_h \bullet S_h \right)} \qquad (3.2)$$

n_h = number of establishments allocated to stratum h
n = fixed establishment allocation for each state
EMP_h = total number of stratum employees (obtained from the LDB)
S_h = a measure of occupational variability

Ideally, a sample should allocate most heavily to those strata where the least amount of certainty exists. In a survey that produces employment and wages for 801 different occupations, the challenge is to develop a measure that produces the best estimates for as many occupations in each geographic area as is practical. As a result, we considered that overall variation in occupational employment with each stratum as a measure for variability of the stratum. If some value that represents the overall occupational variability can be formed, then it is advantageous (in the sense of minimizing sampling variance) to allocate sample to each stratum utilizing this measure. The greater a stratum's occupational variability, the more sample points, relatively, can be taken from the stratum. A disadvantage to a design that considers only occupational variability is that it may reduce sample in those strata that have large populations and small occupational variability. This would result in large sample weights (defined as the number of units in the frame that the sample case represents); over-reliance in an individual sample case may result in "over-inflating" the value of a sampled business establishment within a stratum. If non-sampling error (inaccuracies that are due to reporting imperfections by respondents and interviewers or errors made in coding and processing data) occurs, a high sample weight may magnify these errors; there may not be other sample cases within the stratum to assist in "off-setting" non-sampling errors.

In order to minimize high sample weight situations, a sort of "smoothing" or "spreading out" of the sample should be done. *Power Allocations: Determining Sample Sizes for Subnational Areas* by Bankier (1988) uses a method referred to as Power Allocation. The measure of size in the variance component of the allocation is taken to a power "q" between 0 and 1. That is, applying an exponential value q, where 0<q<1, to the stratum population value (EMP_h) in the Neyman Allocation formula will allow a sufficient spread of the sample allocation. For expediency, q=0.5 was selected. In this case, employment was the measure of size and the square root of employment was taken in both the numerator and denominator of the ratio described for the Neyman Allocation (the ratio is the "variance component"). Further analysis will be conducted in the future to fine-tune the value of q to match the OES Program's preference for reliability across domain sizes (in particular, areas).

The Power Allocation (and working formula for this study) looks as such:

$$n_h = n \bullet \frac{\sqrt{EMP_h} \bullet S_h}{\sum \left(\sqrt{EMP_h} \bullet S_h \right)} \qquad (3.3)$$

Adding the power allocation component to the formula provides a balance between the need for quality national estimates and detailed estimates by local area. Medium and small size areas are still afforded enough sample to provide detailed estimates of occupations important to the local economies of these areas, while still adequately covering the largest areas that are important to reliable national occupational estimates.

3.1 Occupational Homogeneity and Heterogeneity

Consider an industry-only stratum. Each industry contains workers within a set of occupations; in general, industries are defined by the occupations in which they entail. Some industries entail a great number of occupations, while other industries entail a smaller number of occupations. Industries that entail a great number of occupations tend to have large occupational variability when comparing establishments within the industry. Industries that entail a smaller number of occupations tend to have smaller occupational variability when comparing establishments within the industry.

For example, the Accommodations/Food Services industry employs more than twice as many workers as the Wholesale Trade industry, yet contains roughly only half as many occupations (refer to Table 1). When comparing these two industries, it is evident that the Accommodations/Food Service Industry is more occupationally homogeneous (occupational staffing patterns are less variable when comparing industry business establishments) and that the Wholesale Trade Industry is more occupationally heterogeneous (occupational staffing patterns are more variable when comparing industry business establishments).

Table 1: Example of Occupational Variability

	Accom/Food Services Industry	Wholesale Trade Industry
National Empl Count	12.8 million	6.1 million
90^{th}-pctile Occ Count	88	175

3.2 Calculating Coefficient of Variation (CV)

Each industry contains workers within a set of defined occupations; OES obtains a list of these occupations for each industry using the most recent estimates file (using weighted data). A "coefficient of variation" (CV) is calculated for each occupation within each industry stratum. *Sample Survey Methods and Theory, Volume 1* by Hanson, Hurwitz, and Madow (1953) describes the CV as a measure of average relative error obtained by calculating the standard deviation of the distribution of the relative errors. The Relative Variance (CV^2) for an original variate Y_i is calculated as follows:

$$CV_y^2 = \frac{S^2{}_y}{\bar{y}^2} = \frac{\sum_i^N (y_i - \bar{y})^2}{(N-1) \cdot \bar{y}^2} \qquad (3.4)$$

$S^2{}_y$ = variance of variate y_i
$\bar{y}$ = mean of variate y_i
N = number of observations of variate y_i

$$CV_y = \sqrt{CV_y^2}$$

Using the following definitions:
y_i as the occupational employment for establishment i,
x_i as the total employment for establishment i,
w_i as the weight assigned to establishment i (number of business establishments that i represents on the frame),
The following formula can be defined as the ratio of occupational employment to total employment:
Given an occupation,

$$R = \frac{\sum_i y_i}{\sum_i x_i} \quad (3.5) \qquad R_w = \frac{\sum_i (w_i \cdot y_i)}{\sum_i (w_i \cdot x_i)} \quad (3.6)$$

(unweighted data) (weighted data)

R_w is a value that will be less than or equal to 1.

CV can then be defined (using unweighted data):

$$CV_y = \sqrt{\frac{S^2{}_y}{\bar{y}^2}} = \frac{S_y}{\bar{y}} = \frac{\sqrt{Var(y)}}{E(y)} \qquad (3.7)$$

$$CV_y = \frac{S_y}{\bar{y}} = \frac{S_y}{R \cdot \bar{x}} = \frac{\frac{1}{\bar{x}} \cdot S_y}{R} = \frac{\frac{1}{\bar{x}} \cdot \sqrt{\frac{\sum_{i=1}^{N} (y_i - R \cdot x_i)^2}{N-1}}}{R}$$

Since w_i (the sampling weight) represents the number of establishments each sample case represents, the sum of w_i's represents the total number of establishments within each stratum (N).

Thus, the CV can be approximated (after applying weights):

$$CV_{y_R} \approx \frac{\frac{1}{\bar{x}} \cdot \sqrt{\frac{\sum_{i=1}^{n} [w_i y_i - R_w \cdot w_i x_i]^2}{\sum_i w_i - 1}}}{R_w} \qquad (3.8)$$

Note: $\bar{x}$ is a weighted average:

$$\bar{x} = \frac{\sum_{i}^{n} w_i x_i}{\sum_{1}^{n} w_i} \qquad (3.9).$$

The CV formula (3.8) is used to calculate a CV value for each occupation within each industry stratum. The smaller the CV, the less diverse the occupation is within the (industry) stratum.

3.3 Calculating Occupational Variability Value (S_h)

S_h can be defined as a measure used to estimate the occupational homogeneity (or heterogeneity) of a stratum. A CV value is calculated for each occupation within each (industry) stratum. These CVs are then aggregated to a single value S_h by taking a weighted mean (by employment size of each occupation) of the 90th-percentile of occupations within each (industry) stratum. That is, occupational proportions for each (industry) stratum are calculated; these values are then multiplied by the respective occupational CV values. Afterwards, the 90th-percentile of occupations within each (industry) stratum is summed. Summing over the 90th-percentile is done to eliminate less prevalent occupations (bottom 10%) from each (industry) stratum; atypical occupations within an industry tend to have large CV values which may lead to an inaccurate estimate of the stratum's overall occupational variability measure. Refer to the numerical example in Section 4 of this document.

3.4 Stratification Level of S_h

Is it appropriate to determine the level of stratification for S_h (including list of occupations) at the national level? Many cells within a given geographic|industry stratum have just a small handful of business establishments (observations); thus, estimates of CVs and S_h at this level may not be reliable. However, there are a large number of observations for each industry at the national level (aggregated geographically); thus, estimates of CVs and S_h at this level will be substantially more reliable. A major concern of using nationally aggregated values for S_h is an assumption that occupational staffing patterns for a given industry remain similar from each MSA (or BOS) when compared to the nation. For instance, can it be assumed that the national occupational staffing pattern for the educational industry is similar to the occupational staffing pattern in Los Angeles (or any other MSA or BOS) for the educational industry? In order to determine this, a proportional occupational distribution was calculated in each (4-digit and selected 5-digit) NAICS industry for the nation as well as each MSA/BOS area.

For example, it was determined that in the "Elementary and Secondary Schools" industry, nationwide—17.9% of workers in this industry were Elementary School Teachers, 12.3% were Secondary School Teachers, 12.0% were Teacher Assistants, and so forth. In the Los Angeles-Long Beach-Glendale, CA MSA—21.3% of workers in this industry were Elementary School Teachers, 13.5% were Secondary School Teachers, 15.8% were Teacher Assistants, and so forth. Are these percentages (distributions) for occupations in the "Elementary and Secondary School" industry in the Los Angeles area close enough to the nationwide percentages in the same industry? A two-tailed test was used to determine this with an alpha-level of 10%—results were 11.4% of MSA industry occupational staffing patterns differed from their national counterparts. That is to say that 88.6% of the time, national aggregations of occupational staffing patterns for each industry was justified. Of course, in larger states, such as California and New York, justification was much stronger (differences of only 9.1% and 9.4%, respectively) because these states contributed greatly to the nation. Smaller states, such as Idaho and Montana (differences of 17.8% and 17.3%, respectively) did not individually contribute greatly to the nation, but even in these states national aggregation of occupational staffing patterns of industry was justified 82-83% of the time. If these differences are accepted (as they were), then aggregations of

occupational staffing pattern at the national level is satisfactory.

4. A NUMERICAL EXAMPLE

An illustrative numerical example will show how S_h is calculated as defined. A hypothetical stratum with two business establishments will be used. The hypothetical stratum consists of two restaurants, one called "ABC" and the other called "XYZ". ABC will represent five business establishments, thus will have a sample weight of 5. XYZ will represent only itself, thus will have a sample weight of 1. Tables 2a and 2b show the individual occupational staffing patterns of ABC and XYZ.

Table 2a: ABC's Staffing Pattern	
Occupation	**# employed**
Wait Staff	8
Cook	4
Dishwasher	2
Janitor	1
Manager	1
TOTAL	16

Table 2b: XYZ's Staffing Pattern	
Occupation	**# employed**
Wait Staff	32
Cook	15
Dishwasher	10
Manager	3
TOTAL	60

Tables 3a and 3b show weighted calculations of ABC and XYZ.

Table 3a: Weighted Calculations for ABC				
Wait Staff	**Cook**	**Dishwasher**	**Janitor**	**Manager**
$y_i=8$	$y_i=4$	$y_i=2$	$y_i=1$	$y_i=1$
$w_iy_i=5\cdot8=40$	$w_iy_i=5\cdot4=20$	$w_iy_i=5\cdot2=10$	$w_iy_i=5\cdot1=5$	$w_iy_i=5\cdot1=5$
$x_i=16$	$x_i=16$	$x_i=16$	$x_i=16$	$x_i=16$
$w_ix_i=5\cdot16=80$	$w_ix_i=80$	$w_ix_i=80$	$w_ix_i=80$	$w_ix_i=80$

Table 3b: Weighted Calculations for XYZ			
Wait Staff	**Cook**	**Dishwasher**	**Manager**
$y_i=32$	$y_i=15$	$y_i=10$	$y_i=3$
$w_iy_i=1\cdot32=32$	$w_iy_i=1\cdot15=15$	$w_iy_i=1\cdot10=10$	$w_iy_i=1\cdot3=3$
$x_i=60$	$x_i=60$	$x_i=60$	$x_i=60$
$w_ix_i=1\cdot60=60$	$w_ix_i=60$	$w_ix_i=60$	$w_ix_i=60$

Assuming that these are the only two business establishments in this (imaginary) stratum, the following CVs can be calculated for each occupation, knowing that $\Sigma w_i=5+1=6$ and $\Sigma(w_i\cdot x_i)=80+60=140$.

Wait Staff

$$CV_{y_R} \approx \frac{\frac{1}{(140/6)}\cdot\sqrt{\frac{\left[40-80\cdot\frac{(40+32)}{140}\right]^2+\left[32-60\cdot\frac{(40+32)}{140}\right]^2}{6-1}}}{\frac{(40+32)}{140}} \approx 0.060$$

(Using formula 3.8)

Using formula 3.8, the following values can be calculated for each occupation:

Table 4: CV Values	
Occupation	**CV**
Wait Staff	0.060
Cook	0.000
Dishwasher	0.271
Janitor	1.626
Manager	0.203

As stated earlier, smaller CVs indicate less diverse occupations within stratum.

Now that CVs have been calculated for all occupations within this stratum, S_h can be calculated. Remember, S_h is the weighted mean of the 90th-percentile occupations within stratum.

Table 5: Weighted Mean of CVs of All Occupations			
Occupation	**CV**	**Proportion**	**Product**
Wait Staff	0.060	72/140≈0.51	0.060*0.514≈0.03
Cook	0	35/140=0.25	0*0.25≈0
Dishwasher	0.271	20/140≈0.14	0.271*0.143≈0.04
Janitor	1.626	5/140≈0.04	1.626*0.036≈0.06
Manager	0.203	8/140≈0.06	0.203*0.057≈0.01

Looking at Table 5, the 90th-percentile of occupations are Wait Staff, Cook, and Dishwasher (sum of the "Proportion" column of these three occupations = 0.90). Summing their weighted products yields a weighted mean of CVs: $S_h=0.03+0+0.04=0.07$.

This value for S_h is what is used in formula 3.3 when allocating the stratified sample for all business establishments chosen in the same industry stratum.

To elucidate:

- S_h values differ only by industry (approximately 300 unique values)
- EMP_h values differ between geography|industry strata (approximately 175,000 unique values)
- n is the establishment fixed allocation value for each state or territory (54 unique values); all values sum to approximately 1.2 million

5. GRAPHICAL ILLUSTRATIONS

The goal of the new OES design was to ensure that sample was fairly allocated throughout strata in each state with consideration to population (EMP_h=frame employment) and variance (S_h=occupational variability). Moving from a PPS model to a Neyman model (and using aggregated occupational variability as defined as a proxy for S_h) allowed us to spread out the sample by industry (one of the stratification variables).

Refer to the earlier example on Table 1 (Wholesale Trade industry population and variance compared with Accommodations/Food Services industry population and variance). Looking at Graph 1 (on the next page), note that Wholesale Trade (sixth set of bars from the left) showed an increase in sample allocation with the Neyman design (due to greater occupational variability) while Accommodations/Food Services (third set of bars from the right) showed a decrease in sample allocation with the Neyman design (due to lower occupational variability).

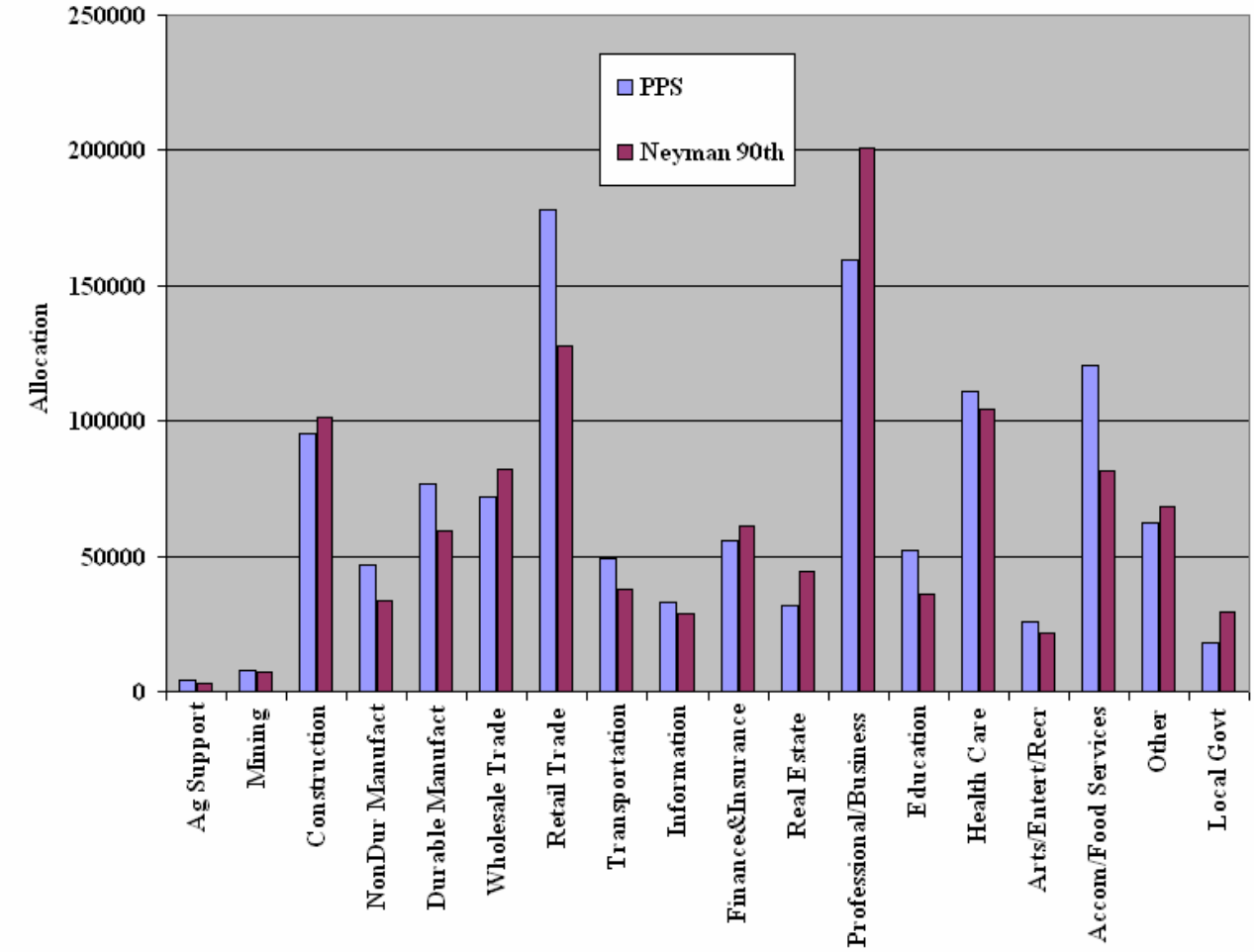
Incorporating a Power Allocation of 0.5 (q=0.5) into the Neyman design allowed us to spread out the sample by area (one of the stratification variables), particularly by MSA/BOS within each state.

Refer to Graph 2 (on the next page). In Illinois, using the Neyman formula without the Power Allocation allocated a majority (approximately 55%) of Illinois's sample to Chicago. This will inevitably lead to unreliable estimates in smaller areas within Illinois. When including the Power Allocation (q=0.5) with the Neyman formula, Illinois's allocation to Chicago dropped to approximately 33%, thus allowing more sample allocated to smaller areas in Illinois and improved reliability in those smaller areas. As Chicago's sample decreased, notice that all non-Chicago areas sample increased.

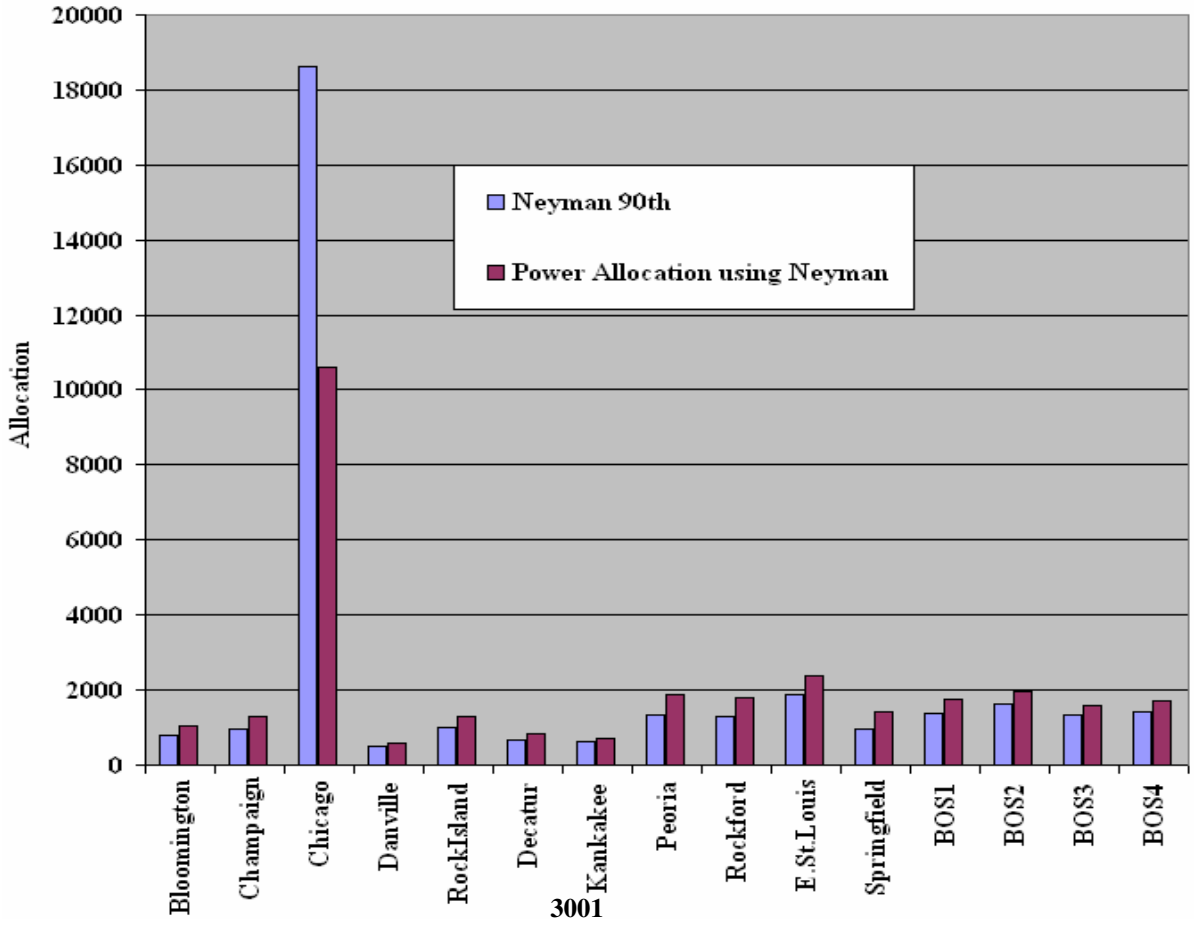
6. CONCLUDING REMARKS

In the past, OES used a probability-proportional-to-size sample allocation model, which produced samples that were not reliable due to possible oversampling in strata that had high populations, but small variability. The OES frame is stratified by geography and industry. The PPS model adjusted sample sizes based only on strata populations, without consideration to occupational variability. Therefore, adjustments had to be made to accommodate for variability in each stratification class. To accommodate for the industry stratification class, a Neyman Allocation model was adopted. In the Neyman model, the variance component, particularly S_h, was studied. In the end, S_h was determined a better measure for occupational variability within strata. A technique was devised to exploit S_h as an occupational variability measure, thus allowing the Neyman formula to adjust sample sizes based on strata populations and strata occupational variability. To accommodate for possible oversampling the geography stratification class, a Power Allocation model was adopted. The Power Allocation model allowed for adjustments of sample sizes based on geographic area, moving sample from highly concentrated geographic areas to smaller geographic areas, thus allowing for more reliability of estimates in smaller areas while maintaining reliability in large geographic areas.

Graph 1: Comparison of PPS Design to Neyman Design



Graph 2: Power Allocation Illustration—Total Allocation for Illinois



ACKNOWLEDGEMENTS

George D. Stamas, Bureau of Labor Statistics
Meghan O'Malley, Bureau of Labor Statistics
Jeffrey T. Willingham, Market Force Information
Laura Train, Bureau of Labor Statistics
Teresa Hesley, Bureau of Labor Statistics
Martha Duff, Bureau of Labor Statistics
Edwin L. Robison, Bureau of Labor Statistics
Kenneth W. Robertson, Bureau of Labor Statistics
Albert Tou, Bureau of Labor Statistics

REFERENCES

Bankier, Michael D. (1988). Power Allocations: Determining Sample Sizes for Subnational Areas. *American Statistician*, Vol. 42, pp. 174-177.

Cochran, William G. (1977). *Sampling Techniques*, 3rd ed. John Wiley and Sons, New York.

Gilliland, Phil D. (1981) *Sample Design for Occupational Statistics Survey*. Bureau of Labor Statistics. Washington.

Hansen, Morris H., William N. Hurwitz, and William G. Madow (1953). *Sample Survey Methods and Theory*, Vol. 1. John Wiley and Sons, New York.

Hogg, Robert V. and Allen T. Craig (1995). *Introduction to Mathematical Statistics*, 5th ed. Prentice-Hall, Englewood Cliffs, NJ.

Koti, Kallappa M. (1988). *Optimum Stratified Sampling Using Prior Information*. Ph.D. dissertation submitted to Texas Tech University

Robertson, Kenneth W. (2001). *Occupational Employment Statistics, Statistical Methods Documentation*, pp. 13-18

Sarndal, Carl-Erik, Bengt Swensson, and Jan Wretman. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Triplett, Jack E. (1994) *The Use of OES CV Data for Measuring Occupational Homogeneity in SICs*. Bureau of Labor Statistics. Washington.

United States. Executive Office of the President. Office of Management and Budget. (2001) *Standard Occupational Classification Manual.* Bernan Associates, Lanham, MD.

United States. Executive Office of the President. Office of Management and Budget. (2002) *North American Industry Classification System*. Bernan Associates, Lanham, MD.