# SAMPLE DESIGN FOR THE TERRORISM RISK INSURANCE PROGRAM SURVEY

**G. Hussain Choudhry, Westat; Mats Nyfjäll, Statisticon; and Marianne Winglee, Westat**
**G. Hussain Choudhry, Westat, 1650 Research Boulevard, Rockville, Maryland 20850**

**Key Words:** Stratified Sample Design, Optimum Sample Allocation, Non-linear Programming, Systematic Sampling, Composite Selection Probabilities

We describe in this paper the sample design for the terrorism risk insurance program survey that Westat conducted for the U.S. Department of Treasury to estimate at the national level and for a number of domains the uptake rate and the average premium paid for the terrorism risk insurance. The sampling frame for the private sector was constructed from the Dun and Bradstreet listing of businesses, and that for the state and local governments and special districts was compiled from the 2002 Census of Governments. The sample design was a stratified single-stage design with systematic sampling of business entities. We used a non-linear programming technique to determine the optimum sample allocation to minimize the total sample while achieving the required precision levels for the survey estimates. We determined the composite selection probabilities for the systematic sampling procedure that at least one member (headquarters or a subsidiary) of the business was selected, and constructed the sampling weights based on the composite selection probabilities.

## 1. Introduction

The terrorism risk insurance program survey was conducted in 2003 and 2004 for the U.S. Department of Treasury to estimate at the national level and for a number of domains the uptake rate and the average premium paid for terrorism risk insurance. The Terrorism Risk Insurance Act (TRIA) of 2002 mandates several studies by the Treasury Department as administrator of the Terrorism Risk Insurance Program. The Program was established

- To protect consumers by addressing market disruptions and ensure the continued widespread availability and affordability of property and casualty insurance for terrorism risk, and
- To allow for a transition period for the private markets to stabilize, resume pricing of such insurance, and build capacity to absorb any future losses.

To fulfill the requirements of the TRIA, the study was designed to assess the effectiveness of the program, and the likely capacity of the property and casualty insurance industry to offer terrorism risk insurance in workers compensation, other casualty, and property insurance lines after the program sunsets, by law, on December 31, 2005. Data collection involved three surveys: a demand-side survey of businesses; a supply-side survey of insurance companies (insurers); and a separate supply-side survey of re-insurers, i.e., the companies that insure the insurance companies.

We discuss in this paper the sample design and weighting methodology of the demand-side survey of businesses that will be representative of the entire industrial and governmental composition of the U.S. economy.

## 2. Sampling Frame

The target population for the demand-side survey of insurance purchasers consists of all private sector businesses, and state and local governments with 10 or more employees. The frame for the private sector included headquarters and subsidiaries of domestic businesses, and the subsidiaries located in the U.S. of foreign owned businesses with headquarters outside the United States. The description of the sampling frames for the private sector and for the governments follows.

### 2.1 Creation of Sampling Frame for the Private Sector

We constructed the private sector sampling frame using the Dun and Bradstreet (D&B) business directory. We included all business headquarters and subsidiaries located in the United States, including subsidiaries of foreign businesses with headquarters outside the U.S. Branches of businesses and businesses with less than 10 employees were excluded from the frame. The initial frame consisted of 1,523,635 businesses. We removed as much as possible public sector records in the D&B frame before combining with the Census of Government (CoG) frame. After removing the public sector records, the final private sector sampling frame consisted of 1,476,746 businesses entities.

The variables, geographic location, industry classes, and size, were the stratification variables. We defined 15 geographic location strata as follows. The first seven geographic locations were the seven high-risk cities identified by the Treasury. The businesses that were not in the high-risk cities were assigned to eight geographic locations defined by region by urban/non-urban status. We followed the Census Bureau definitions to define four regions, and used the

Census 2000 city population to define urban/non-urban status. Businesses located in cities with a population size of at least 350,000 were assigned urban status and businesses located in cities or places with fewer than 350,000 people were assigned non-urban status.

We defined the industry classes following the 1997 North American Industry Classification System (NAICS) codes. The D&B businesses were classified by detailed Standard Industrial Classification codes (the SIC 2+2 Codes). To construct the industry groups by NAICS codes, we used conversion tables provided by the Census Bureau for mapping SIC codes to the NAICS codes. We mapped most of the businesses into NAICS groups by applying the conversion table. For a small number of remaining businesses, we resolved the mapping manually by looking up the 8-digit SIC codes on the D&B frame and the corresponding descriptions from the U.S. Bureau of the Census.

We defined the four size categories based on total assets (< $10 million; $10 – 100 million; $100 million – $1 billion; and > $1 billion). This classification was approximate because total assets were missing for about 78 percent of the businesses in the D&B frame. To circumvent this problem, we used the distribution of assets on the 2000 corporate tax returns to obtain number of businesses with assets between $10 and $100 million. The number of businesses with assets above $100 million on the corporate tax returns was divided into the categories: $100 million to $1 billion, and above $1 billion by using the count for number of businesses with assets over 1 billion from the D&B list and the distribution of employment (number of employees) to obtain the cut-points to match the asset distribution from the corporate tax returns.

### 2.2 Creation of Sampling Frame for Governments

All data files used to create a sampling frame for the governments were downloaded from the U.S. Bureau of the Census home page. Table 2-1 lists all the files used in the creation of the sampling frame for the governmental entities. Note that the government frame included only the *independent* governmental units whereas the *dependent* governmental units were excluded. The number of employees (full time equivalent), also obtained from the Census Bureau home page, was then merged onto the files in Table 2-1. However, employment data was missing for some governments. We imputed the employment figures for the missing cases since employment is used in the stratification. We defined the four size categories based on employment. The four size categories are: less than 150; 150-699; 700-3,999; and greater than or equal to 4,000.

Table 2-1. Files downloaded from the U.S. Census Bureau home page

| No. | File name | No. of records | Type of government |
|---|---|---|---|
| 1 | 2002GID_Counties | 3,034 | County Governments |
| 2 | 2002GID_Cities | 19,429 | Municipal (or City) Governments |
| 3 | 2002GID_Towns | 16,504 | Township Governments |
| 4 | 2002GID_Special_Districts | 35,052 | Special District Governments |
| 5 | 2002GID_Schools | 13,506 | Independent School Districts |

None of the files in Table 2-1 contains the 50 state governments, which were also eligible for the study. We created the final file of Census of Government (CoG) frame consisting of 40,375 records by processing the files 1 through 5 and adding the state governments.

## 3. Sample Design

The sample design for the demand-side survey is a single-stage stratified sample of businesses with systematic sampling of businesses within strata. The sample was allocated to the strata to minimize the total sample size while satisfying the required precision levels for the national and domain level estimates of uptake rates and total employment. The stratification and sample allocation are discussed in this section.

### 3.1 Stratification

First, we created two special strata with 100 percent sampling (certainty strata). One of these special strata was the 50 state governments, and the other was the businesses with more than 300,000 employees. There were 12 such businesses (i.e., with more than 300,000 employees), and one of these was a state government. Therefore, the two certainty strata contained 61 entities. In addition, a number of businesses that are owners of high-risk buildings were sampled with certainty.

The sampling strata for the remainder of the frame were defined by cross-classification of three categorical variables: Industry (10 categories), Geography (15 categories), and Size (4 categories). Among the 10 industry categories there were five high-risk industries, and five categories for the remainder of the industries. Among the 15 geography categories, there were seven high-risk cities, and eight categories defined as region by urban/non-urban without the seven high-risk cities. There were 522 non-empty strata, and

the number of sampling entities within strata varied from 1 to 149,605.

### 3.2 Sample Allocation

The sample allocations were determined for the combined frame constructed from the Dun and Bradstreet (D&B) listing and the Census of Governments (CoG). The combined D&B and CoG frame contained 1,517,121 business entities. The survey estimates were required for 28 domains including the national level estimates. The 28 domains of interest are: 4 Census Regions, 7 High Risk Cities, 5 High Risk Industries, 5 Major Industries, Urban/Non-Urban locations, 4 Size Categories, and the National level.

We used non-linear programming to obtain the minimum sample size that would satisfy the CV requirement for the estimated employment and the 95 percent confidence interval half-width requirement for the estimated uptake rate for each of the 28 domains given above. We applied the additional constraint that the maximum sampling rate would be 40 percent because the assumed response rate would not exceed 40 percent.

The total sample was then minimized under the constraints that the CV requirements for the estimates of employment and 95 percent confidence interval half-widths of uptake rates would be satisfied for the 28 domains of interest.

$$\text{Min } f(\alpha) = \sum_{h=1}^{H} \alpha_h N_h \,, \qquad (3\text{-}1)$$

where $h$ denotes the sampling stratum. $N_h$ is the number of entities in stratum $h$, and $\alpha_h$ is the sampling rate for stratum $h$. We minimize (3-1) under the constraints that $0 < \alpha_h \le 0.40$, and the precision requirements are satisfied for the $D$ domains of interest.

We used 40 percent as the upper threshold for sampling fraction because we did not expect the response rate to be more than 40 percent. Thus, we would sample 2.5 times the required sample from each of the sampling strata.

We let $n_h = \alpha_h \times N_h$ denote the sample size for the sampling stratum $h$. We also define $W_h = \frac{1}{\alpha_h}$, where $W_h$ is the design weight for stratum $h$. Then equation (3-1) becomes

$$\text{Min } f(W) = \sum_{h=1}^{H} \frac{N_h}{W_h} \,, \qquad (3\text{-}2)$$

subject to the constraints that the weights $W_h \ge 2.5$, and the corresponding CV and 95 percent confidence interval half-width constraints for the $D$ domains of interest are satisfied. We computed the 95 percent confidence interval half-widths by assuming that the uptake rates would be 50 percent for the businesses in the largest size category, and 20 percent for the other 3 size categories. The CV constraints and the 95 percent confidence interval half-width constraints were obtained using the variances with finite population correction factors.

#### CV Constraints

The relative variance of the estimated total ${}_d\hat{Y}$ (or mean) that is the square of the coefficient of variation (CV) of the estimate is given as

$$\text{Rel.Var.}\left({}_d\hat{Y}\right) = \frac{Var\left({}_d\hat{Y}\right)}{\left({}_dY\right)^2} = \frac{\sum_{h \in d} (W_h - 1) N_h S_h^2}{\left(\sum_{h \in d} Y_h\right)^2}. \qquad (3\text{-}3)$$

Thus, the CV constraints can be expressed as linear constraints if the constraints are expressed in terms of the squares of the CVs as a function of the sampling weights instead of the sampling fractions.

#### 95 % Confidence Interval Half-Width Constraints

The estimate of a proportion for the domain of interest $d$ can be written as

$${}_d\hat{p} = \frac{\sum_{h \in d} W_h x_h}{\sum_{h \in d} W_h n_h} = \frac{\sum_{h \in d} W_h x_h}{\sum_{h \in d} N_h} \,, \qquad (3\text{-}4)$$

where $x_h$ is the number of observed cases from stratum $h$ that belong to the category "yes" (e.g., take terrorism risk insurance). Then, the variance of the estimated proportion is given by

$$Var\left({}_d\hat{p}\right) = \frac{\sum_{h \in d} (W_h - 1) N_h \left[p_h (p_h - 1)\right]}{\left(\sum_{h \in d} N_h\right)^2}, \qquad (3\text{-}5)$$

and the 95 percent Confidence Interval Half-Width $HW(_d\hat{p})$ is given by $2\sqrt{Var(_d\hat{p})}$. Thus, the squared 95 percent confidence interval half-width is also a linear function of the sampling weights. Thus, the both the CV and the 95 percent confidence interval half-width constraints become linear if expressed as the squared quantities in terms of sampling weights instead of sampling fractions. The vector of Gradients can also be computed analytically.

It should be noted that the sample allocation is a trade-off between obtaining smaller CVs or smaller confidence interval half-widths. The smaller CVs (of employment, assets, etc.) can be obtained by sampling at higher rates the larger businesses; but smaller half-widths of the confidence intervals require that the smaller businesses be sampled at higher rates.

After the sample allocation to the primary strata, we further stratified the sample on the basis of size in those primary strata for which the sample size was 50 or more in order to obtain additional stratification gains.

## 4. Sample Weighting

The sample design for the terrorism risk insurance program survey is a stratified single-stage sample of business entities. The probability of selecting a business is the composite probability that at least one of its component businesses is selected. The base weights were computed as reciprocal of the selection probabilities, and these weights were adjusted to account for the nonrespondent businesses.

### 4.1 Base Weights

The sampling unit is a subsidiary (or a headquarters) but the business may not always be able to report the insurance data separately for all of its subsidiaries. For example, a business may only report aggregate data at the ultimate headquarters level that will account for all the subsidiaries owned by the business. On the other hand, a business may report insurance data for a sub-set of its subsidiaries as a group for a number of such groups that will collectively account for the entire business. Thus, the headquarters level aggregate level data is a special case where the entire business will be a single group for reporting insurance data.

The weight assigned to the headquarters or a group of subsidiaries (hereafter referred to as the reporting units) will be based on the composite probability of selection of the group (or the reporting unit), which is the probability that at least one entity (headquarters or subsidiary) belonging to the reporting unit will be selected. Even if the headquarters was not selected and a subsidiary belonging to the reporting unit that contains the headquarters was selected data will be collected for the entire reporting unit.

We denote by $h$, $N_h$, and $n_h$ respectively the stratum, the number of business entities in the stratum, and the number sampled from the stratum. We selected the sample of business entities within each stratum with systematic sampling from a sorted list of these entities. For the sake of simplicity, the reporting unit (group of subsidiaries) will be called a business, which may or may not be reported by the ultimate headquarters. For example, a large corporation may own several businesses, and each is responsible for its own insurance, and the headquarters insures itself and any establishments directly reporting to the headquarters. We use the symbol $i$ to denote a business. For a business that is a single entity (i.e., with no subsidiaries) the probability of selection is given by $\pi_i = \frac{n_h}{N_h}; i \in h.$

For a business with multiple entities (headquarters and subsidiaries), we compute the composite probability that at least one of these entities will be selected. Suppose that the business (reporting unit) denoted by $i$ is in stratum $h$, and there are a number of entities (subsidiaries) both in stratum $h$ and in other strata for which data are reported through the business $i$ in stratum $h$. The number of entities (sampling units) in stratum $h$ is $N_h$. We denote by $N_h^{*(i)}$ the number of entities in stratum $h$ for which the data are reported through the business dented by $i$, where $0 \le N_h^{*(i)} \le N_h; \quad h = 1, 2, 3, ---, H.$

If $P_{hi}$ is the probability that at least one of the subsidiaries of business $i$ will be sampled from stratum $h$, then the composite probability of selecting the business $i$ (i.e., probability that at least one of the entities that is associated with business $i$ will be selected) is given by

$$\pi_i = \left\{1 - \prod_{h=1}^{H}(1 - P_{hi})\right\}; \quad i \in h, \qquad (4\text{-}1)$$

where $P_{hi} = 0$, if $N_h^{*(i)} = 0$ and $P_{hi} = \frac{n_h}{N_h}$, if $N_h^{*(i)} = 1$. The case when $N_h^{*(i)} \ge 2$ is discussed below.

**Case when $N_h^{*(i)} \geq 2$**

The $N_h$ sampling entities in stratum $h$ are labeled $1, 2, 3, ..., N_h$. Let $J_{h,k}^{(i)}$ be the index of the $k^{th}$ subsidiary of business $i$ in stratum $h$, where $k = 1, 2, ..., N_h^{*(i)}$. We sampled $n_h$ out of the $N_h$ business entities from stratum $h$ with systematic sampling procedure. We need to calculate the probability $P_{hi}$ that at least one of the entities (subsidiaries) of business $i$ will be sampled from stratum $h$. For calculating the probability $P_{hi}$, we define the intervals $\mathrm{I}_h$ and $I_{h,k}^{(i)}$ as follows.

$$\mathrm{I}_h = \left(0, \frac{N_h}{n_h}\right) \tag{4-2}$$

and

$$\mathrm{I}_{h,k}^{(i)} \begin{cases} = \left(L_{h,k}^{(i)}, L_{h,k}^{(i)} + 1\right); \\ \text{if } L_{h,k}^{(i)} + 1 \leq \dfrac{N_h}{n_h} \\ = \left(L_{h,k}^{(i)}, \dfrac{N_h}{n_h}\right) \cup \left(0, L_{h,k}^{(i)} + 1 - \dfrac{N_h}{n_h}\right); \\ \text{otherwise} \end{cases}, \tag{4-3}$$

where $L_{h,k}^{(i)} = \left(J_{h,k}^{(i)} - 1\right) \bmod \left(\frac{N_h}{n_h}\right)$. We compute the $N_h^{*(i)}$ intervals corresponding to the $N_h^{*(i)}$ subsidiaries of the business $i$ that are in stratum $h$ where $h = 1, 2, 3, ..., H$. Next, we compute the union of the $N_h^{*(i)}$ intervals that correspond to the $N_h^{*(i)}$ subsidiaries of business $i$ that are in stratum $h$ and denote the composite interval by

$$\mathrm{I}_h^{(i)} = \bigcup_{k=1}^{N_h^{*(i)}} \mathrm{I}_{h,k}^{(i)} . \tag{4-4}$$

The composite interval $\mathrm{I}_h^{(i)}$ is the interval that corresponds to the composite probability that at least one of the $N_h^{*(i)}$ subsidiaries of business $i$ will be sampled from stratum $h$. The corresponding probability $P_{hi}$ is then given by

$$P_{hi} = \frac{\left\|\mathrm{I}_h^{(i)}\right\|}{\left\|\mathrm{I}_h\right\|} \tag{4-5}$$

where $\|.\|$ denotes the length of the interval. Note that $\left\|\mathrm{I}_h\right\| = \frac{N_h}{n_h}$ is the sampling interval for stratum $h$. We also note that $P_{hi} = \frac{\left\|\mathrm{I}_h^{(i)}\right\|}{\left\|\mathrm{I}_h\right\|}$ reduces to $\frac{n_h}{N_h}$ when $N_h^{*(i)} = 1$ because the length of the interval $\mathrm{I}_h^{(i)} = 1$ for $N_h^{*(i)} = 1$, i.e. there is only one subsidiary of business $i$ in stratum $h$. The composite probability that at least one entity belonging to the business $i$ will be selected is then given by

$$\pi_i = \left\{1 - \prod_{h=1}^{H} \left(1 - P_{hi}\right)\right\}; \quad i \in h , \tag{4-6}$$

where $P_{hi} = 0$, if $N_h^{*(i)} = 0$, and $P_{hi}$ is given by (4-1) for $N_h^{*(i)} \geq 1$. The base weight assigned to business $i$ will be the reciprocal of the probability of selection, i.e., $w_i = \frac{1}{\pi_i}; \quad i \in h.$

### 4.2 Sampling Weights for the High Risk Buildings

After the sample had been selected, Westat identified more than 200 high risk buildings and the businesses that owned these high risk buildings were selected with certainty. Some of these businesses were in the original sample and those not already in the sample were included in the sample. Thus, the sample size becomes random. We considered both the conditional and unconditional approaches for constructing the sampling weights that account for the supplementary sample of certainty businesses. These two approaches are discussed below.

**Conditional Approach**

We condition on the achieved sample size and the randomization is over all possible samples of size equal to the achieved sample (Särndal and Hidiroglou, 1989). If $M_h$ out of the $N_h$ entities in stratum $h$ are high risk (*HR*) entities, and $m_h$ of the high risk entities are in the initial sample then the achieved Sample size from stratum $h$ is $n_h - m_h + M_h$ business entities

including the high risk businesses. The weighting under the conditional approach is conditionally unbiased, and hence it is unconditionally unbiased as well.

**Unconditional Approach**

Under the unconditional approach the randomization is over all possible samples that could be selected irrespective of the achieved sample size. Therefore, the selection probabilities of the businesses that are not at high risk do not depend on the achieved sample size. It should be noted that the weighting under the unconditional approach are conditionally biased.

We constructed the sampling weights under the conditional approach because the estimates are conditionally unbiased and hence unconditionally unbiased as well. Moreover, the conditional variance is also unconditionally unbiased. Kalton (2002) also recommended that if a subset of sampling distribution in which the estimator is conditionally approximately unbiased can be identified, then a conditional analysis should be employed in analyzing an actual sample.

### 4.3 Final Sample Weights

The final sample weights were constructed by applying adjustment to account for nonresponse (Elliot, 1991). The sample cases can be divided into respondents and nonrespondents. Further, the respondents can be either eligible or ineligible (out of scope) for the survey. The eligibility of the nonrespondent businesses could not always be determined. For example, a sampled business that did not cooperate could be very small (less than 10 employees) and hence ineligible for the survey. Therefore, the nonrespondent businesses were classified into two categories: (1) eligible nonrespondents and (2) nonrespondents with unknown eligibility. In order to apply the adjustments for unknown eligibility and nonresponse, the sample cases were grouped into four response status categories: 1. Eligible Respondents; 2. Ineligible or Out of Scope; 3. Eligible Nonrespondents; and 4. Unknown Eligibility.

In a typical application, the nonresponse adjustment can be carried out in two stages. At the first stage the base weights of those with unknown eligibility (Category 4) are allocated proportionally to those whose eligibility is known (Categories 1, 2, and 3) and the weights of those with unknown eligibility are set to zero. In the second stage the adjusted weights of eligible non-respondents (Category 3) is redistributed among the respondents (Category 1). Since additional information on the activity status (active versus inactive) is available on a subset of those with unknown eligibility, this information can be used for making the adjustment for unknown eligibility. As suggested by Nathan (2003), the unknown eligibility adjustment can itself be made in two stages by making use of this information. In the first stage we allocated the weight of those for whom the activity status was unknown between those whose activity status was known, and the weight of those with unknown activity status was set to zero. In the second stage the weights of those who were known to be active with unknown size (number of employees) were allocated between those who were known to be active with known sizes, and the weights of those with unknown size were set to zero. Choudhry et al. (2002) also implemented a similar strategy for constructing weights for the Random Digit Dial (RDD) sample for the national survey of veterans where adjustment for unknown eligibility was made in two stages. First, unknown eligibility adjustment was made for those telephone numbers for which residential status (residential versus nonresidential) was not known. Second, the unknown eligibility adjustment was made for those known to be residential but it was not known whether there was a veteran in the household.

In all the non-response adjustments described above, the adjustments were carried out within weighting classes defined by the combination of the stratification variables: Geo-location, Size and NAICS code. Since the cross-classification by all three stratification variables would have resulted in a sparse table with too many cells, standard *CHAID* (Chi-square Hierarchical Automatic Interaction Detector) analysis was used to form the weighting classes by combining classes without significant differences in response or eligibility propensities (Kass, 1980). The *CHAID* analysis was done separately for each nonresponse adjustment type, i.e., first stage of unknown eligibility adjustment, second stage of unknown eligibility adjustment, and the nonresponse adjustment for eligible nonrespondents.

The final survey weight was defined as the product of the base weight and the nonresponse adjustment factors as described above. These weights were used to obtain survey estimates at the national level and for the domains of interest.

## 5. References

Choudhry, G.H., Park, I., Kudela, M.S., and Helmick, J.C. (2002). *2001 National Survey of Veterans Design and Methodology – Final Report*. Westat, Rockville, Maryland.

Elliot, D. (1991). *Weighting for Nonresponse: A Survey Researcher's Guide*. Office of Population Censuses and Surveys, Social Surveys Division, London.

Kalton, G. (2002). Models in the Practice of Survey Sampling (Revisited). *Journal of Official Statistics,* **18**, 129-154.

Kass, G. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics,* **29**, 119-127.

Nathan, G. (2003). *Nonresponse Adjustment for the TRIP Establishment Sample*. Westat Memorandum Number S-34 dated December 13, 2003.

Särndal, C.E., and Hidiroglou, M.A. (1989). Small Domain Estimation: A Conditional Approach. *Journal of the American Statistical Association*, **84**, 266-275.