RHODE ISLAND DEPARTMENT OF ENVIRONMENTAL MANAGEMENT
OFFICE OF TECHNICAL AND CUSTOMER ASSISTANCE
235 Promenade Street, Providence, RI 02908

AUTOMOTIVE REFINISHING INDUSTRY
ENVIRONMENTAL RESULTS PROGRAM
Total Sample Size Determination

DRAFT
Oct. '02

Introduction

In 2002, the Rhode Island automotive refinishing industry sector consisted of 371 licensed auto body shops. A complete database of all licensed shops was obtained from the Rhode Island Department of Business Regulation. The database was converted to Microsoft Access for the random selection of shops (using a random number generator) for baseline compliance audits. In total, approximately 10% of the facilities were audited for compliance with RCRA, air pollution, wastewater discharge, and occupational health and safety regulatory requirements, as well as, the adoption of pollution prevention measures. Baseline facility audits were conducted by the Office of Technical & Customer Assistance (OTCA), during the spring/summer of 2002, in consultation with DEM regulatory divisions and the Rhode Island Department of Health (DOH). Several years prior to conducting the audits, a considerable amount of survey and site-specific research had been performed in this sector in order to better understand industry practice and multi-media pollutant releases. Findings from these efforts were published in the peer-reviewed literature and incorporated into major components of the Auto Body Environmental Results Program (ERP) – e.g., the certification workbook, media-specific checklists, and fact sheets. Checklists used in the baseline audits were reviewed and approved by all media programs and DOH; these are the same checklists that will be used by the auto body industry to document compliance with certification program requirements.

Statistical Approach

The number of baseline and post-implementation compliance audits to be conducted was determined in consultation with Dr. Choudary Hanumara (Department of Computer Science and Statistics, University of Rhode Island) following the statistical methodology (for binomial proportions) employed by the Massachusetts Department of Environmental Protection (MA DEP); binomial simply refers to a two outcome event. The statistical method calculates sample size based on the difference between two proportions. The proportions of interest are the compliance rates before (p_1) and after (p_2) ERP implementation. More specifically:

$$p_1 = \text{baseline compliance rate expressed as a decimal fraction} = \frac{\text{no. of shops in compliance with regulatory standard}}{\text{total number of shops in the population}}$$

$$p_2 = \text{compliance rate post-ERP implementation} = \frac{\text{no. of shops in compliance with regulatory standard}}{\text{total number of shops in the population}}$$

The idea behind the calculation is to obtain two samples (one before and one after ERP implementation) of equal and sufficient size that would allow one to detect a difference (i.e., an improvement) in compliance rates if one truly exists. The calculation is based on the assumption that compliance rates are approximately normally distributed throughout the industry sector (population). The ability to detect this difference is not only a function of sample size, but is also dependent upon the *significance level* and *power* assumed and used in the calculation.

Definitions for significance level and power can be found in basic statistical handbooks and are described by Smoller (1990) as:

Significance level = magnitude of a Type I error (denoted by alpha, α)
= probability of finding an effect when there really isn't one

Power = probability of finding a real effect (of the size that you think is important)
= 1- beta (β)

where:
β = magnitude of a Type II error
= probability of failing to find an effect when in fact there really is one

Kriebel et al. (2001) provide a summary description of Type I and Type II errors within the context of a scientific investigation: "A Type I error is the mistake of concluding that a phenomenon or association exists when in truth it does not. (Technically, the Type I error is rejecting the null hypothesis when it is really true.) By convention, Type I (or alpha) errors are guarded against by setting that error rate low, usually at 5%. In other words, the finding must be so strong that there is less than a 5% probability that this result would have been seen by chance alone in a world in which no such phenomenon actually exists. In this case, the result is called statistically significant (with the clear implication that one is suppose to believe it). The Type II error, failing to detect something that actually does exist, is, by convention, often set at 20% (although practical limitations of sample size often result in a substantially higher or lower Type II error). Twenty percent of the time, a real phenomenon will be missed because the data were not strong enough to convincingly demonstrate its existence. There is an implicit bias here: the test is set up to be more cautious about falsely detecting something than about failing to detect something."

Practical Considerations

To calculate sample size, compliance rate proportions (p_1 and p_2), significance level, and power have to be specified. The values that OTCA used were based on prior knowledge of the industry, expectations for program success, and the level of certainty that we were willing to accept in order to demonstrate program success. Overall, our goal was to determine a sample of sufficient size (for baseline and post-implementation inspections) that would support the measurement component of the ERP, but not be so large as to defeat the purpose of the program which is to improve regulatory compliance with increased efficiency (e.g., fewer on-site multi-media inspections).

In agreement with conventional practice, we initially chose an alpha level of 5% and power of 80%; by comparison, clinical trials often use a power of 80%, as greater assurance in finding an effect can be more costly (C. Hanumara). From prior experience, we also felt that the industry's baseline performance, relative to environmental, health and safety regulatory requirements, would be low and estimated the overall compliance rate (p_1) to be about 40%; that is, on average body shops were expected to be in compliance with only about 40% of all regulatory requirements. We also projected that the ERP could improve compliance in this industry sector by more than 25% (effect size). We therefore, set p_2 at 65% (40% + 25%). This meant that we wanted to be 80% sure that we could detect a difference in compliance rates of 25 % (from 40 to 65%) at the .05 level. Considering available resources, we also felt that it would not be too burdensome for our small unit to conduct comprehensive multimedia audits of about 10% of the licensed auto body shop population (audits were expected to take 2-3 months to complete).

Finally, although an overall p_1 of .40 was specified, we realized that compliance with individual regulatory requirements could vary widely. To measure ERP program success, we selected 19 environmental business practice indicators (EBPIs) that were representative of a range of air, water, health/safety, pollution prevention, and hazardous waste management activities. Indicators included, for example, proper labeling of hazardous waste containers, the control of airborne sanding dust, whether wastewater was allowed to run off-site, and whether a worker respiratory protection plan had been developed. The fact that the compliance rate (p_1) for each individual regulatory requirement could be << .40 or >> .40, coupled with the nonapplicability of some regulatory requirements to certain facilities (for example, OSHA does not require one man shops to comply with their regulations), meant that the final power achieved for each indicator could be less than 80%.

Calculating the Sample Size

The formula used to calculate the total sample size (n) based on the difference between two proportions (Woolson and Clark) is given below:

$$n = 2\left(\frac{Z_{1-\beta}\ \theta\overline{p_1(1-p_1) + p_2(1-p_2)} + Z_{1-\alpha}\ \theta\overline{2\bar{p}(1-\bar{p})}}{✤}\right)^2$$

where: p_1 = baseline compliance rate
p_2 = post implementation compliance rate
✤ = $|\,p_1 - p_2\,|$
$Z_{1-\alpha}$ = significance level test statistic
$Z_{1-\beta}$ = power test statistic
$\bar{p} = (p_1 + p_2)/2$

Based on our original assumptions [where p_1 = .40, p_2 = .65, power = .80 ($Z_{0.80}$ = 0.842), $\alpha = .05$ ($Z_{0.95}$ = 1.645)] the <u>total</u> number (n) of body shops that would need to be audited (i.e., baseline + post-ERP implementation audits) is:

$$n = 2\left(\frac{.842 \times \theta\overline{.4(1-.4) + .65(1-.65)} + 1.645 \times \theta\overline{2 \times [(.4+.65)/2] \times [(1-(.4+.65)/2]}}{|\,.4 - .65\,|}\right)^2$$

$$= 2\left(\frac{(.842)\ \mathrm{SQRT}(.24 + .2275) + (1.645)\ \mathrm{SQRT}(2 \times .525 \times .475)}{.25}\right)^2$$

$$= 2\left(\frac{(.842 \times 0.6837) + (1.645 \times .7062)}{.25}\right)^2$$

$$= 2\,(6.9495)^2$$

= 97 (that is 97/2, or 49 baseline and 49 post implementation audits)

The effect of choosing different values for p_1, p_2, and ✤ are shown in Table 1. This table shows that, in general, in order to detect a small effect size (difference between proportions p_1 and p_2, which equals ✤), larger samples are needed. In Table 1, the significance levels, ✤$_1$ and ✤$_2$, and power are set at .05/.10 and .80, respectively. Increasing the alpha value from .05 to .10, effectively reduces the total number of samples (n_2) that are needed for a study.

Table 1. Sample size determination for comparison of proportions (α = .05, power = .80)

p_1	p_2	q_1	q_2	$\bar{p}$	Δ	α	power	n
0.20	0.40	0.80	0.60	0.30	0.20	0.05	0.80	128
0.20	0.45	0.80	0.55	0.33	0.25	0.05	0.80	85
0.20	0.50	0.80	0.50	0.35	0.30	0.05	0.80	60
0.20	0.60	0.80	0.40	0.40	0.40	0.05	0.80	35
0.25	0.45	0.75	0.55	0.35	0.20	0.05	0.80	139
0.25	0.50	0.75	0.50	0.38	0.25	0.05	0.80	91
0.25	0.55	0.75	0.45	0.40	0.30	0.05	0.80	64
0.25	0.65	0.75	0.35	0.45	0.40	0.05	0.80	36
0.30	0.50	0.70	0.50	0.40	0.20	0.05	0.80	146
0.30	0.55	0.70	0.45	0.43	0.25	0.05	0.80	95
0.30	0.60	0.70	0.40	0.45	0.30	0.05	0.80	66
0.30	0.70	0.70	0.30	0.50	0.40	0.05	0.80	37
0.35	0.55	0.65	0.45	0.45	0.20	0.05	0.80	151
0.35	0.60	0.65	0.40	0.48	0.25	0.05	0.80	97
0.35	0.65	0.65	0.35	0.50	0.30	0.05	0.80	67
0.35	0.75	0.65	0.25	0.55	0.40	0.05	0.80	36
0.40	0.60	0.60	0.40	0.50	0.20	0.05	0.80	153
0.40	0.65	0.60	0.35	0.53	0.25	0.05	0.80	97
0.40	0.70	0.60	0.30	0.55	0.30	0.05	0.80	66
0.40	0.80	0.60	0.20	0.60	0.40	0.05	0.80	35
0.45	0.65	0.55	0.35	0.55	0.20	0.05	0.80	151
0.45	0.70	0.55	0.30	0.58	0.25	0.05	0.80	95
0.45	0.75	0.55	0.25	0.60	0.30	0.05	0.80	64
0.45	0.85	0.55	0.15	0.65	0.40	0.05	0.80	33
0.50	0.70	0.50	0.30	0.60	0.20	0.05	0.80	146
0.50	0.75	0.50	0.25	0.63	0.25	0.05	0.80	91
0.50	0.80	0.50	0.20	0.65	0.30	0.05	0.80	60
0.50	0.90	0.50	0.10	0.70	0.40	0.05	0.80	30
0.55	0.75	0.45	0.25	0.65	0.20	0.05	0.80	139
0.55	0.80	0.45	0.20	0.68	0.25	0.05	0.80	85
0.55	0.85	0.45	0.15	0.70	0.30	0.05	0.80	56
0.55	0.95	0.45	0.05	0.75	0.40	0.05	0.80	27
0.60	0.80	0.40	0.20	0.70	0.20	0.05	0.80	128
0.60	0.85	0.40	0.15	0.73	0.25	0.05	0.80	77
0.60	0.90	0.40	0.10	0.75	0.30	0.05	0.80	49
0.60	0.95	0.40	0.05	0.78	0.35	0.05	0.80	33
0.65	0.85	0.35	0.15	0.75	0.20	0.05	0.80	114
0.65	0.90	0.35	0.10	0.78	0.25	0.05	0.80	67
0.65	0.95	0.35	0.05	0.80	0.30	0.05	0.80	42

$\mathbf{n} = 2\,[(Z_{1\text{-}\beta}\ \mathrm{SQRT}\ (p_1 q_1 + p_2 q_2)) + (Z_{1\text{-}\alpha}\ \mathrm{SQRT}\ (2p\bar{q}')) / \Delta)]^2$

Where:

$q_1 = 1\text{-}p_1$ $\quad q' = 1 - p'$ $\quad Z_{1\text{-}\alpha} = 1.645$

$q_2 = 1\text{-}p_2$ $\quad \Delta$ = absolute value $p_1 - p_2$

$\bar{p} = (p1 + p2)/2$ $\quad Z_{1\text{-}\beta} = 0.842$

Excel syntax example (row 1, n=128): 2*(((0.842*SQRT(A1*C1+(B1*D1)))+
1.645*SQRT(2*E1*(1-E1)))/F1)^2

Table 2. Autobody Sector Environmental Business Practice Indicators[1]

Hazardous Waste
Manifest tracking (n=40; p_1=.55)
[2]Accumulation label (n=40; p_1=.25)
<u>90-Day Storage</u>
Secure area/storm water protection (n=16; p_1=.75)
Secondary containment (n=16; p_1=.63)
Container inspections documented (n=16; p_1=.00)
Written contingency plan (n=16; p_1=.00)
Written training records (n=16; p_1=.06)

Air/Pollution Prevention
Sanding dust control (n=40; p_1=.30)
[3]Methylene chloride usage (n=40; p_1=.67)
[4]Compliant coatings (n=40; p_1=1.00)
[4]HVLP spray guns (n=40; p_1=1.00)

Air/Pollution Prevention (Cont'd)
Rag storage (n=38; p_1=.76)

Wastewater
Unpermitted flr drains (n=17; p_1=.24)
Wash water runoff (n=40; p_1=.30)
Discharge signage (n=40; p_1=.00)

OSHA
Safety/health poster (n=32; p_1=.43)
Haz Com program (n=32; p_1=.09)
PPE program (n=32; p_1=.18)
Lockout/tagout (n=32; p_1=.12)
Resp program est. (n=32; p_1=.25)
Emergency action plan (n=32; p_1=.72)

[1] n = the number of shops to which the specified regulatory requirement was found to apply as the result of conducting 40 baseline audits.
[2] Proportion of 90-dy storage (n=16) plus satellite accumulation generators (n=24) in compliance w/ labeling requirements (including accumulation start date).
[3] For this indicator, p_1 equals the number of shops that do not use MeCl and are therefore practicing pollution prevention.
[4] All shops in the sample were found to be in compliance with this air regulation during baseline audits. These performance data are retained for future reference.

Sample Size Determination for RI Auto body ERP

From Table 1, to detect an effect size of 25% (at ✽ = .05 and power = .80) for p_1 values that range from .20 to .65, the total number of compliance audits (n_1) that would be needed for a statistically valid sample ranges from 67 (i.e., 34 baseline + 34 post-implementation audits) to 97 (i.e., 49 baseline + 49 post-implementation audits); to detect a 30% effect size, "n_1" ranges from 42 to 67. By comparison, if one is willing to accept a significance level of .10 in place of .05, then the range of values for "n_2" for a 25% effect size can be reduced – i.e., 49 (i.e., 25 baseline + 25 post-implementation audits) to 70 (i.e., 35 baseline + 35 post-implementation audits) for p_1 = .20 to p_1 = .65.

As previously discussed, given the available resources, we made an initial determination that a 10% sample size (a total sample size where "n" = 80) was achievable. By direct comparison with Tables 1, the value n=80 is appropriate for 1) studies with an alpha = .05 and power of .80 when one is trying to detect an effect size of 30% or more, or 2) for studies where the minimum effect size being measured is 25% at a .10 significance level. With this as a guide, the final approach we decided to take was to conduct 40 baseline audits (10% of the population), analyze the resultant data, and estimate the actual p_1 values for selected Environmental Business Practice Indicators (EBPIs). This was viewed as the most prudent approach, since it was anticipated that p_1 values would vary widely for the range of indicators.

EBPIs and Measurement Summary

MA DEP defines an EBPI as an "industry-specific measure designed to give a snapshot of a facility's environmental performance." For the auto body industry sector, we selected 19 EBPIs (not including HVLP spray guns and compliant coatings, Table 2) on which to measure performance in air, water, RCRA, pollution prevention, and occupational health and safety. EBPIs were selected based on best professional judgement considering the potential for release prevention, emissions reduction, and human health/environmental protection.

The data in Table 2 show that although 40 baseline audits were conducted, the number (n) of shops that EBPI's actually applied to were often <<40. For example, RCRA hazardous waste generator requirements for 90-dy storage do not apply to shops that only engage in satellite accumulation; also, Rhode Island regulations do not contain a SQG exclusion. Similarly, certain OSHA requirements do not apply to one person owner/operated shops. Regarding wastewater discharge, only 17 of 40 shops were found to have floor drains (13 of which were not permitted) in production areas. In order to achieve n=40 for all indicators, therefore, a much larger number of baseline audits would need to be conducted.

To illustrate, Table 3 lists the range of proportions (p_1) calculated from baseline audits, and shown in Table 2, in ascending order. The sample size calculation using the p_1 values determined from baseline audits and an alpha level of .10 with a power of .80 resulted in "n" values (shaded column) that could be achieved with the resources at hand. [Sample calculation – "container inspections documented" EBPI in Table 2. Sixteen out of 40 shops that received baseline audits, were required to inspect their hazardous waste containers on a weekly basis (all others were subject to satellite accumulation standards only). Of the 16 shops to which the regulation applied, none (p_1=.00) were in compliance with the inspection requirements. In order to measure an improvement in compliance (p_2) of 25% or more, the total number of shops that must be inspected for compliance with this EBPI is 30 (15 baseline audits + 15 post-ERP implementation audits) as shown in Table 3.

In sum, total sample size and p_1 values were found to vary by indicator upon analysis of the baseline audit data. Therefore, even though measurable improvements in industry performance can be tracked over time, with the submittal of return compliance forms and upon completion of a second round of post-ERP implementation audits, some of the observed improvements will not achieve the initially desired level of statistical confidence; i.e., the ERP team will have to accept a lower alpha level for the affected EBPI and/or limited ability to measure only relatively large effect sizes (e.g., 30-40%).

Table 3. Sample size determination (actual p_1 field rates)

p_1	p_2	q_1	q_2	p'	Δ	α	power	n
0.00	0.25	1.00	0.75	0.13	0.25	0.10	0.80	30
0.06	0.36	0.94	0.64	0.21	0.30	0.10	0.80	31
0.09	0.31	0.91	0.69	0.20	0.22	0.10	0.80	58
0.12	0.34	0.88	0.66	0.23	0.22	0.10	0.80	64
0.18	0.43	0.82	0.57	0.31	0.25	0.10	0.80	59
0.24	0.59	0.76	0.41	0.42	0.35	0.10	0.80	34
0.25	0.55	0.75	0.45	0.40	0.30	0.10	0.80	46
0.25	0.50	0.75	0.50	0.38	0.25	0.10	0.80	66
0.30	0.55	0.70	0.45	0.43	0.25	0.10	0.80	69
0.43	0.73	0.57	0.27	0.58	0.30	0.10	0.80	47
0.55	0.80	0.45	0.20	0.68	0.25	0.10	0.80	62
0.63	0.93	0.37	0.07	0.78	0.30	0.10	0.80	33
0.67	0.87	0.33	0.13	0.77	0.20	0.10	0.80	78
0.72	0.92	0.28	0.08	0.82	0.20	0.10	0.80	65
0.75	1.00	0.25	0.00	0.88	0.25	0.10	0.80	30
0.76	0.96	0.24	0.04	0.86	0.20	0.10	0.80	53

Where: $Z_{1-\alpha} = 1.282$

References

Kriebel, D. et al. The Precautionary Principle in Environmental Science. Environ Health Perspect 109:871-876 (2001).

MADEP. Printer's Sampling Design Interoffice Memo. M. Hutcheson. September (1995).

Wassertheil-Smoller, S. Biostatistics and Epidemiology. 2nd ed. Springer-Verlag New York Inc. (1990).

Woolson, R.F. and W.R. Clarke. Statistical Methods for the Analysis of Biomedical Data. 2nd ed. Wiley (2002).