Statistical Formulas and Citations for EPA's 2007 ERP Sample Planner and ERP Results Analyzer

Prepared by: Dr. A. Richard Bolstein, with Michael Crow
Under: EPA Contract Number EP-W-04-023, Work Assignment 3-61

May 12, 2008

I. SAMPLE PLANNER

A. FIND SAMPLE SIZE (1-SAMPLE)[1]

(1)
$$n_{\inf} = \left[\frac{p(1-p)}{2e^2} - 1\right] z^2 + \frac{z^2}{2e}\sqrt{1 - 2p(1-p) + \left[\frac{p(1-p)}{e}\right]^2}$$

(2)
$$n = \frac{n_{\inf}}{1 + \frac{n_{\inf}}{N}}$$

Definition of Symbols:

$n_{\inf}$ = **preliminary required sample size before adjustment for the finite population.**

p = 0.5 = value of true proportion that would require the largest sample size for a prescribed degree of statistical precision.

e = desired half-width for a confidence interval for the estimated proportion. For example, if e=0.05 (or 5%), the confidence interval will have a width of 10%.

z = 1.96 for 95% confidence (the standardized z-score or multiplier.)

z = 1.645 for 90% confidence

N = the number of facilities in the entire population being sampled.

n = final required sample size after adjustment for the finite population.

Example I.A.1: If N=1,000, e=.05, and confidence level desired is 95%, then

$$n_{\inf} = \left[\frac{.25}{2(.05)^2} - 1\right](1.96)^2 + \frac{(1.96)^2}{0.10}\sqrt{0.5 + \left[\frac{0.25)}{.05}\right]^2} = 188 + 194 = 382$$

$$n = \frac{382}{1 + \frac{382}{1000}} = 277$$ **rounded to the next largest integer.**

[1] Derived from: Agresti and Coull. 1998. "Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions." *The American Statistician*, v. 52, no. 2, 119-126.

B. FIND SAMPLE SIZE (2-SAMPLE)[2]

(3) $m = \left[p_1(1-p_1) + p_2(1-p_2) \right] \left[\frac{z}{e} \right]^2$ **= sample size for each round before finite population adjustment, where**

$$p_1 = p_2 = 0.5$$

are the values of the proportions in each round that would require the largest sample size to meet specified margin of error and confidence level.

(4) $n = \frac{m}{1 + \frac{m}{N}}$ **= sample size for each round of sampling. If the population size N is different in round 2 than in round 1, either calculate the two different sample sizes or use the largest of the two in both rounds.**

<u>Example I.B.1:</u> The desired margin of error is plus or minus 5% with 95% confidence. The population size is 1,000 in both rounds of sampling. Then

$$m = [0.25 + 0.25] \left[\frac{1.96}{.05} \right]^2 = 768$$

$n_1 = n_2 = \frac{768}{1 + \frac{768}{1000}} = 435$ **are the sample sizes for round 1 and round 2.**

[2] Derived from: Kish, Leslie. 1965. *Survey Sampling*. John Wiley & Sons, Inc. New York, NY. p. 41.

C. FIND MARGIN OF ERROR (1-SAMPLE)[3]

The Score interval is used in the Sample Planner rather than the usual Wald interval because it has better coverage characteristics when sample sizes are small, especially if proportions are near the extremes of zero or one. (By better 'coverage' we mean approximate 95% confidence intervals constructed using normal distribution theory really do contain the true proportion in nearly 95% of the possible samples.)

The observed proportion p is still used as the point estimate, but the center of the confidence interval is

(5)
$$\frac{p+\frac{z^2}{2n}}{1+\frac{z^2}{n}}=\frac{np+\frac{z^2}{2}}{n+z^2}$$

Where z = 1.96 for 95% confidence or 1.645 for 90% confidence.

If z = 2 (for a 95.44% confidence interval), the center of the Score confidence interval given by formula (5) simplifies to $\frac{np+2}{n+4}$**, which is equivalent to adding two successes (e.g., compliances) and two failures (e.g., non-compliances) to the data. Thus, the Score interval can be viewed as a regression of the proportion towards 0.5 to adjust for the small sample size. (For large n, or proportions near 0.5, the Score interval differs little from the traditional Wald interval.)**

The margin of error, or half-width, of the Score confidence interval is

(6)
$$\frac{z\sqrt{\left(1-\frac{n}{N}\right)\frac{p(1-p)}{n}+\frac{z^2}{4n}}}{1+\frac{z^2}{n}}$$

The lower and upper Score confidence limits are obtained by subtracting and adding (6) to (5), respectively.

<u>Example I.C.1:</u> Suppose n = 100, N = 1,000 and a 95% confidence interval is desired. Suppose that 50 facilities in the sample were in compliance with an item, so p=0.5.

Then the center of the Score interval, formula (5), also equals 0.5, so in this case the Score confidence interval is symmetric about p. The half-width is

---

[3] Agresti and Coull. 1998. "Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions." *The American Statistician*, v. 52, no. 2, 119-126.

$$\frac{1.96\sqrt{\frac{\left(1-\frac{100}{1000}\right).25+\frac{3.84}{400}}{100}}}{1+\frac{3.84}{100}} = .091 = 9.1\%$$

So the 95% Score confidence interval would be 40.9% to 59.1%.

<u>Example I.C.2:</u> Suppose again that n = 100, N = 1,000 and a 95% confidence interval is desired, but only 20 facilities in the sample were in compliance with an item, so p=0.2.

In this case the center of the Score interval given by (5) is 0.211 = 21.1%, and the half-width given by (6) is .074 = 7.4%, so the Score confidence interval is 13.7% to 28.5%.

D. FIND MARGIN OF ERROR (2-SAMPLE)[4]

The standard error of the difference between two proportions is

$$\textbf{(7)} \qquad SE(p_1 - p_2) = \sqrt{\left(1 - \frac{n_1}{N_1}\right)\frac{p_1(1-p_1)}{n_1 - 1} + \left(1 - \frac{n_2}{N_2}\right)\frac{p_2(1-p_2)}{n_2 - 1}}$$

The subscripts refer to the two different samples. In the Sample Planner, the size of the population N_1 is known but N_2 is unknown since the population may change by the time the second round of inspections is to be chosen. Therefore, in the Sample Planner we assume $N_2 = N_1$. Likewise, we plan for sample sizes to be the same in both rounds, $n_2 = n_1$.

The margin of error of the estimated difference p1-p2 is formula (7) multiplied by the appropriate z-factor. The margin of error will be largest when both proportions equal 0.5, so this would provide an upper bound for the error.

Example I.D.1: Assume the population and sample sizes are 1,000 and 100 in the initial survey, and it is anticipated they will remain the same in the second survey. Inserting 0.5 for the values of p1 and p2 in formula (7) produces a standard error of 6.74%. If a 95% confidence interval is required, the margin of error is (1.96)(6.74%) = 13.2%. This is the maximum margin of error (half-width of confidence interval) that can occur with these sample sizes.

[4] Derived from: Kish, Leslie. 1965. *Survey Sampling*. John Wiley & Sons, Inc. New York, NY. p. 41.

II. RESULTS ANALYZER

A. CONFIDENCE INTERVAL FOR A PROPORTION (1-SAMPLE)[5]

The lower endpoint of the Score confidence interval is obtained by subtracting formula (6) from (5). The upper endpoint is obtained by adding (6) to (5). (See section I.C for more details on the Score confidence interval.)

The Score interval is the default interval calculated in the Results Analyzer. The standard Wald confidence interval is offered in the section of the Results Analyzer labeled 'For Advanced Users'. The formula for the half-width of the Wald interval is

(8) $$z\sqrt{\left(1-\frac{n}{N}\right)\frac{p(1-p)}{n-1}}+\frac{1}{2n}$$

(The last term in (8) is called the 'continuity correction factor'.) The lower endpoint of the Wald confidence interval is obtained by subtracting (8) from the observed proportion p. The upper endpoint is obtained by adding (8) to the observed proportion p.

Example II.A.1: As in Example I.C.1, suppose n = 100, N = 1,000 and a 95% confidence interval is desired. Assume that 50 facilities in the sample were in compliance with an item, so p=0.5.

The Score confidence interval has a half-width of 9.1%, as shown in Example I.C.1 in section I.C of this document.

Using (8), the Wald interval has a half-width of 9.8%, since

$$(1.96)\sqrt{\left(1-\frac{100}{1000}\right)\frac{.25}{99}}+\frac{1}{200}=.098$$

This illustrates that the Score interval is tighter and hence preferable to the Wald.

Example II.A.2: As in Example I.C.2, suppose n = 100, N = 1,000, but this time only 20 facilities are in compliance, so p = 0.2.

The Score confidence interval has a half-width of 7.4% (from Example I.C.2, section I.C of this document), whereas the Wald interval has a half-width of 8.0%.

The Score confidence interval is 13.7% to 28.5% and is not symmetric about the observed proportion p = 20%. The Wald interval is 20.0% ± 8.0% (12% to 28%). It is symmetric about the observed proportion but is larger than the Score interval.

---

[5] (1) Agresti and Coull. 1998. "Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions." *The American Statistician*, v. 52, no. 2, 119-126. (2) Cochran, W. G. 1977. *Sampling Techniques*, third edition. p. 57.

B. MEAN (1-SAMPLE)[6]

The margin of error of the sample mean $\bar{Y}$ as an estimator of the true population mean is given by

(9) $$t_{n-1}\sqrt{\left(1-\frac{n}{N}\right)s^2/n}$$

where

(10) $$s^2=\frac{1}{n-1}\sum_i\left\{y_i-\bar{Y}\right\}^2$$ **= sample variance**

(11) **t_{n-1} = Value of Student's t-distribution with (n-1) degrees of freedom such that the right-tail probability is α/2 if the desired confidence level is 1-α**

<u>Example II.B.1</u>: Assume N=1,000, n=100, and a 95% confidence interval is desired. Then t_{n-1} = 1.9842. If the sample mean and variance were separately calculated as $\bar{Y}$ = 85 and s^2 = 225, then (9) gives the margin of error as 2.8.

A 95% confidence interval is therefore 82.2 to 87.8.

[6] (1) Cochran, W. G. 1977. *Sampling Techniques*, third edition. p. 27. (2) Snedecor, G. W. and W. G. Cochran. 1989. *Statistical Methods*, eighth edition. Iowa State University Press. Ames, Iowa. p. 55.

C. DIFFERENCE BETWEEN PROPORTIONS (2-SAMPLE)[7]

Formula (7) above is the standard error of the difference p_1-p_2 between the proportions of compliant facilities in the two rounds. If z is the appropriate multiplier for the desired confidence level, then the confidence interval for the difference is given by

$$(12) \qquad (p_1 - p_2) \pm z\sqrt{\left(1-\frac{n_1}{N_1}\right)\frac{p_1(1-p_1)}{n_1-1}+\left(1-\frac{n_2}{N_2}\right)\frac{p_2(1-p_2)}{n_2-1}}$$

Example II.C.1: Assume the population and sample sizes were 1,000 and 100 in each round, and the number of instances of compliance improved from 50 in the first round to 80 in the second. Then p_1 = 50% and p_2 = 80%, so the absolute value of the difference p_1-p_2 is 30%.

The standard error (square root term) equals 6.1%, so the margin of error with 95% confidence is (1.96)(6.1%) = 12.0%.

Therefore, a 95% confidence interval for the absolute difference in compliance rates from round 1 to round 2 is 30% ± 12% or 18% to 42%. Since this confidence interval does not contain zero, we can conclude that compliance rates are definitely different in the two rounds, and that the rate is higher in round 2 than round 1.

[7] Kish, Leslie. 1965. *Survey Sampling*. John Wiley & Sons, Inc. New York, NY. p. 41.

D. DIFFERENCE BETWEEN MEANS (2-SAMPLE)[8]

The margin of error of the estimated difference $\bar{Y}$ between means of independent samples in rounds 1 and 2 is

(13) $$t_{1-\alpha,\upsilon}\sqrt{\left(1-\frac{n_1}{N_1}\right)\frac{s_1^2}{n_1}+\left(1-\frac{n_2}{N_2}\right)\frac{s_2^2}{n_2}}$$

where s_1^2 and s_2^2 are the sample variances from rounds 1 and 2 as calculated from formula (10). Also,

(14) $t_{1-\alpha,\upsilon}$ **= Value of Student's t-distribution with ν degrees of freedom such that the right-tail probability is α/2 if the desired confidence level is 1-α.**

(15) $$\nu=\frac{\left[\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right]^2}{\frac{1}{n_1-1}\left[\frac{s_1^2}{n_1}\right]^2+\frac{1}{n_2-1}\left[\frac{s_2^2}{n_2}\right]^2}$$ **= degrees of freedom (rounded to the nearest integer.)**

<u>Example II.D.1:</u> Assume that the sample sizes and population sizes are 100 and 1,000 respectively in each round of sampling. Assume also that the mean and standard deviation calculated separately in each round are as follows:

$\bar{Y}_1=85 \qquad \bar{Y}_2=65 \qquad s_1=75 \qquad s_2=50$

The degrees of freedom for the t-distribution multiplier is, after rounding,

$$\nu=\frac{\left[\frac{5625}{100}+\frac{2500}{100}\right]^2}{\frac{1}{99}\left[\frac{5625}{100}\right]^2+\frac{1}{99}\left[\frac{2500}{100}\right]^2}=\frac{6601.5625}{38.2734}=172$$

$t_{..05,172}=1.974$

From (13), the margin of error at the 95% confidence level is 16.9. Therefore, a 95% confidence interval for the absolute difference in means is 20 ± 16.9. Since this interval does not contain zero, we can conclude the round 2 population mean is greater than the round 1 population mean.

[8] Snedecor, G. W. and W. G. Cochran. 1989. *Statistical Methods*, eighth edition. Iowa State University Press. Ames, Iowa. Page 97.