

# Minimal Random Sample Size as a Function of the Confidence Probability and Margin of Error Associated with Repeated Independent Trials Processes, and Its Use in Estimating the Size of a Population

by **Professor John M. Bachar, Jr.**  
**CSULB Mathematics Department**

## Basic Theory.

Suppose that there is a large population of elements (persons, bacteria, whatever) and a subset of the population (those voting for option A on a ballot, or those bacteria having left-handed helical tails, etc.) whose fraction,  $p$ , we wish to determine. It is generally impractical or impossible to count directly the number of elements in this subset, and then divide by the number of elements in the whole population, in order to calculate the value of  $p$ . It turns out that by an appropriate choice of a random sample (meaning, every element in the population has an equal chance of being selected) of sufficient size selected from the whole population, one can estimate  $p$  as accurately as one pleases and with as high a confidence probability as one pleases. We now describe how this can be done, after which we will show how to use this result to estimate the size of a population without counting every element.

It is basic that for a sequence of  $n$  repeated independent trials, with probability  $p$  for one of exactly two outcomes (call the one "S", and the other, "~S" -- "not S") on any single trial, the probability of  $E_{n,k,p}$  is given by

$$(1) \quad \text{prob}(E_{n,k,p}) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k},$$

where  $E_{n,k,p}$  is the set of all outcomes which contain  $k$  "S's" and  $n-k$  "~S's". The space of all the individual outcomes of an  $n$  repeated independent trials process consists of  $2^n$  ordered  $n$ -tuples whose entries are either S or ~S. Furthermore,  $U$  is partitioned by the collection  $\{E_{n,0,p}, \dots, E_{n,k,p}, \dots, E_{n,n,p}\}$ , that is,

$$(2) \quad U = \bigcup_{k=0}^n E_{n,k,p} \text{ (disjoint union).}$$

If one has an  $n$  repeated independent trials process with probability  $p$  of getting S on any single trial, and if  $0 \leq k_1 \leq k_2 \leq n$ , then the probability of getting from  $k_1$  through  $k_2$  "S's" in  $n$  trials is

$$(3) \quad \text{prob}\left(\bigcup_{k=k_1}^{k_2} E_{n,k,p}\right) = \sum_{k=k_1}^{k_2} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}.$$

By the use of a deep theorem, (3) can be very accurately approximated for  $n$  "sufficiently large":

### DeMoivre-Laplace Limit Theorem. *The probability in (3) obeys the limit*

$$(4) \quad \lim_{n \rightarrow \infty} \left( \text{prob}\left(\bigcup_{k=k_1}^{k_2} E_{n,k,p}\right) \right) = \frac{1}{\sqrt{2\pi}} \int_{X_1}^{X_2} e^{-t^2/2} dt,$$

where

$$(5) \quad X_i = \frac{k_i - np}{\sqrt{np(1-p)}}, \quad i = 1, 2.$$

*The integral in (4) is the area under the normal curve,  $f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$  ( $-\infty < t < \infty$ ), from  $X_1$  to  $X_2$ .*

Next, if  $n$  denotes the size of a sample that is to be taken randomly (and one after another with replacement after each selection) from the whole population, if  $k$  denotes the number within the sample found to have property S, and if  $p$  is the actual fraction of the whole population having property S, then  $k/n$  is the fraction of the sample having property S. We wish to determine the smallest  $n$  such that  $k/n$  is "close to  $p$ " with a "high probability". More precisely, we wish to determine the smallest  $n$  such that

$$(6) \quad \text{prob}(p-d \leq k/n \leq p+d) = P,$$

where  $P$  is some desired probability (usually chosen to be near 1).  $P$  is called the "confidence probability", and  $d$  is called the "margin of error" (usually small, like .01, .02, .03, etc.).

Now  $p-d \leq k/n \leq p+d$  is true if and only if  $np-nd \leq k \leq np+nd$  is true. This gives the range on  $k$ , the number of  $S$ 's within our sample. Thus, the left hand side of (6) is equal to

$$(7) \quad \text{prob}\left(\bigcup_{np+nd \leq k \leq np-nd} E_{n,k,p}\right)$$

which in turn is equal to (by use of the DeMoivre-Laplace Limit Theorem, with  $k_1=np-nd$  and  $k_2=np+nd$ )

$$(8) \quad \frac{1}{\sqrt{2\pi}} \int_{X-d}^{X+d} e^{-t^2/2} dt, \quad \text{where } X_{\pm d} = \frac{\pm d \sqrt{n}}{\sqrt{p(1-p)}}.$$

From tables of areas under the normal curve, it is known that  $\frac{1}{\sqrt{2\pi}} \int_{-X}^X e^{-t^2/2} dt = 0.6826, 0.9544, \text{ or } 0.9974$ , according as  $X = 1, 2, \text{ or } 3$ , respectively. Thus, if we choose our sample size  $n$  so that

$$(9) \quad X = \frac{d \sqrt{n}}{\sqrt{p(1-p)}}, \quad \text{where } X = 1, 2, \text{ or } 3, \text{ respectively,}$$

then we get the minimal sample size  $n = p(1-p)(X/d)^2$ , with  $X = 1, 2, \text{ or } 3$ , respectively. But since  $p(1-p) \leq 1/4$  for all values of  $p$ , we conclude that by choosing  $n \geq 1/4(X/d)^2$  (the latter is  $\geq p(1-p)(X/d)^2$  !!), we get the desired confidence level  $P = 0.6826, 0.9544, \text{ or } 0.9974$ , with margin of error  $d$ , by choosing  $n = 1/4(X/d)^2$ , for  $X = 1, 2, \text{ or } 3$ , respectively.

**Summary: Choose  $n = 1/4(X/d)^2$  for  $X = 1, 2, \text{ or } 3$ , respectively, and one gets  $\text{prob}(p-d \leq k/n \leq p+d) = 0.6826, 0.9544, \text{ or } 0.9974$ , respectively.**

Various other confidence probabilities can be used by choosing  $X$  according to the following table.

**TABLE OF VALUES OF CONFIDENCE PROBABILITIES,  $P (= \frac{1}{\sqrt{2\pi}} \int_{-X}^X e^{-t^2/2} dt)$ , FOR VARIOUS VALUES OF  $X$  :**

<b>X</b>	<b>0.25</b>	<b>0.50</b>	<b>0.75</b>	<b>1.00</b>	<b>1.25</b>	<b>1.50</b>	<b>1.75</b>	<b>2.00</b>	<b>2.25</b>	<b>2.50</b>	<b>2.75</b>	<b>3.00</b>
<b>P</b>	<b>.19741</b>	<b>.38292</b>	<b>.54674</b>	<b>.68269</b>	<b>.78870</b>	<b>.86639</b>	<b>.91988</b>	<b>.95450</b>	<b>.97555</b>	<b>.98758</b>	<b>.99404</b>	<b>.99730</b>
<b>X</b>	<b>1.598</b>	<b>1.645</b>	<b>1.695</b>	<b>1.751</b>	<b>1.812</b>	<b>1.881</b>	<b>1.960</b>	<b>2.054</b>	<b>2.170</b>	<b>2.326</b>	<b>2.576</b>	<b>2.807</b>
<b>P</b>	<b>.89000</b>	<b>.90000</b>	<b>.91000</b>	<b>.92000</b>	<b>.93000</b>	<b>.94000</b>	<b>.95000</b>	<b>.96000</b>	<b>.97000</b>	<b>.98000</b>	<b>.99000</b>	<b>.99500</b>

### **Estimating the Size of a Population by Random Sampling Instead of Counting Every Element.**

There are countless instances wherein one wishes to determine the size of a certain population of interest, but because of the nature of the population, it is impossible, or at best, extremely difficult, to count directly every element in the population. For example, in census taking, it is impossible to count every one by employing the usual method of door-to-door inquiry plus questionnaire mailings. Another example is the problem of determining the number of fish in a lake, or of determining the number of wolves in a given geographical region. One can think of a myriad of similar examples.

It turns out that one can determine the size of a population with as much accuracy as desired and with as high a confidence probability as desired by means of a process that incorporates the method of random sampling described above. We now describe this process.

We first select a subset (it need not be random) of size  $n_1$  from the population whose size,  $N$ , we wish to estimate. We then "tag" each element in the subset. The fraction of "tagged" elements in the population is given by  $p = n_1/N$ . Since  $n_1$  is known, it follows that the task of estimating  $N$  is equivalent to the task of estimating  $p$  (because  $N = n_1/p$ ). But the latter estimation task simply employs the random sampling method described above.

In order to estimate  $p$  from a random sample of size  $n_2$  taken one after another (with replacement after each selection) from the population, we simply choose the desired confidence probability,  $P$ , together with its associated value of  $X$  (via the DeMoivre-Laplace Limit Theorem), and a margin or error,  $d$ . Thus, by the first sentence after (9),  $n_2 = p(1-p)(X/d)^2$ .

With confidence probability  $P$ , the estimate,  $p_s$ , of  $p$  that is obtained from the random sample will satisfy the inequality  $p-d \leq p_s \leq p+d$ . But if we make the substitution  $d=\delta p$ , where  $\delta$  is chosen in advance, then this inequality becomes

$$(10) \quad (1-\delta)p \leq p_s \leq (1+\delta)p$$

and  $n_2$  becomes

$$(11) \quad n_2 = [(1-p)/p](X/\delta)^2.$$

The estimate,  $N_s$ , of  $N$ , in view of (10) and the fact that  $N_s = n_1/p_s$  and  $N = n_1/p$ , satisfies

$$(12) \quad N/(1+\delta) \leq N_s \leq N/(1-\delta)$$

and

$$(13) \quad -\delta/(1-\delta) \leq (N - N_s)/N \leq \delta/(1+\delta) \leq \delta/(1-\delta),$$

and therefore,

$$(14) \quad |N - N_s|/N \leq \delta/(1-\delta) \quad \text{and} \quad |N - N_s| \leq [\delta/(1-\delta)]N.$$

If we define  $\beta = \delta/(1-\delta)$  (hence  $\delta = \beta/(1+\beta)$ ), then the relative error,  $|N - N_s|/N$ , satisfies

$$(15) \quad |N - N_s|/N \leq \beta$$

with confidence probability  $P$  provided  $n_2$  is chosen by (see (11) with  $\delta = \beta/(1+\beta)$ )

$$(16) \quad n_2 = (1/p - 1)(1 + 1/\beta)^2 X^2.$$

In practice, one does not know  $N$  in advance, of course, but one usually knows an upper bound for  $N$ , say  $N_m$ . Because  $p = n_1/N$ , this is equivalent to knowing a lower bound,  $p_1$ , for  $p$ . This implies  $1/p - 1 \leq 1/p_1 - 1$ . **Thus, if we choose  $n'_2 = (1/p_1 - 1)(1 + 1/\beta)^2 X^2$  (which is  $\geq n_2$ ), then with confidence probability  $P$ , the relative error in estimating  $N$  by using a random sample of size  $n'_2$  will satisfy  $|N - N_s|/N \leq \beta$ .**

As an example, suppose we wish to estimate the population size,  $N$ , of the U.S.A. (assume  $N$  is at most 270 million). Let us say we want a confidence probability of 0.9545 (so that  $X = 2$ ), and that the number  $n_1$  of "tagged" individuals is  $0.8N$  (about 216 million in the first survey) so that  $p = 0.8$ . If we wish the relative error to be 0.1% (so  $\beta = 0.001$  and  $|N - N_s| \leq 270,000$ ), then the random sample size must be  $n'_2 = 1,002,001$ ; if we wish the relative error to be 0.05% (so  $\beta = 0.0005$  and  $|N - N_s| \leq 135,000$ ), then  $n'_2 = 4,004,001$ . Finally, assume the number of "tagged" individuals is  $0.95N$  (=256.5 million, so at most 13.5 million are uncounted in the first survey), hence that  $p=0.95$  and  $\beta = 0.0001$  (i.e., a relative error of 0.01% and with  $|N - N_s| \leq 27,000$ ), then  $n'_2 = 8,882,000$ .

Note that the total number contacted in the two-stage census survey (= the number,  $n_1$ , of "tagged" individuals plus the number,  $n_2$ , in the random sample) is, respectively, 217 million (maximum uncertainty, 270,000, or 0.1%), 220 million (maximum uncertainty, 135,000, or 0.05%), and 265.4 million (maximum uncertainty, 27,000, or 0.01%). Thus, by employing this process, the total number of individuals contacted is less than the population size! Moreover, the estimation of  $N$  is more accurate (and, to boot, has a confidence probability of 0.9545) than the traditional "try-to-count-everyone-in-one-try" method!