Random Sampling Notes
by Professor Emeritus John M. Bachar, Jr.
CSULB Mathematics Department
September 2006

Basic Theory.

Suppose that there is a large finite 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.

We construct a finite probability space as follows. Let M(n, p, m) consist of all ordered sequences 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 number of elements in M(n, p, m) is 2^n. The probability of each singleton subset, $\{(\epsilon_1,...,\epsilon_i,...,\epsilon_n)\}$, of M(n, p, m) is m$(\{(\epsilon_1,...,\epsilon_i,...,\epsilon_n)\}) = p^k(1-p)^{n-k}$ whenever k of the ϵ_i's are "S" and the others are "~S". Moreover, the probability of $E_{n,k,p}$ is given by

(1) $$\mathrm{m}(E_{n,k,p}) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k},$$

where $E_{n,k,p}$ is the set of all elements which contain k "S's" and n-k "~S's ". Furthermore, M(n, p, m) is partitioned by the collection $\{E_{n,0,p},...,E_{n,k,p},...,E_{n,n,p}\}$, that is,

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

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

(3) $$\mathrm{m}(\bigcup_{k=k_1}^{k_2} E_{n,k,p}) = \sum_{k=k_1}^{k_2}\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}.$$

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) $$\lim_{n\to\infty}(\mathrm{m}(\bigcup_{k=k_1}^{k_2} E_{n,k,p})) = \frac{1}{\sqrt{2\pi}}\int_{X_1}^{X_2} e^{-t^2/2}dt,$$

where

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

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 finite 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 (the sample preference) is "close to p" with a "high probability". More precisely, we wish to determine the smallest n such that

(6) $$\mathrm{m}(\bigcup E_{n,k,p} \mid \text{k satisfies } p\text{-}d \le k/n \le p\text{+}d) = P,$$

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 $\le$ k/n $\le$ p+d is true if and only if np-nd $\le$ k $\le$ 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) $$m\Big(\bigcup_{np\text{-}nd\le k\le np+nd} E_{n,k,p}\Big)$$

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

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, or 0.9974, according as X = 1, 2, or 3, respectively. Thus, if we choose our sample size n so that

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

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

Summary: Choose n = 1/4$(X/d)^2$ for X = 1, 2, or 3, respectively, and one gets:
m($\bigcup E_{n,k,p}$ | k satisfies pd $\le$ k/n $\le$ p+d) = 0.6826, 0.9544, or 0.9974, respectively.

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

			Entries in table are the minimum random sample size, $n=1/4(X/d)^2$, for margin of error d and confidence probability P										
X	P	d	0.10%	0.50%	1.00%	1.50%	2.00%	2.50%	3.00%	3.50%	4.00%	4.50%	5.00%
0.50000	0.38292492		62,500	2,500	625	278	156	100	69	51	39	31	25
0.55000	0.41768063		75,625	3,025	756	336	189	121	84	62	47	37	30
0.60000	0.45149376		90,000	3,600	900	400	225	144	100	73	56	44	36
0.65000	0.48430778		105,625	4,225	1,056	469	264	169	117	86	66	52	42
0.70000	0.51607270		122,500	4,900	1,225	544	306	196	136	100	77	60	49
0.75000	0.54674530		140,625	5,625	1,406	625	352	225	156	115	88	69	56
0.80000	0.57628920		160,000	6,400	1,600	711	400	256	178	131	100	79	64
0.85000	0.60467491		180,625	7,225	1,806	803	452	289	201	147	113	89	72
0.90000	0.63187975		202,500	8,100	2,025	900	506	324	225	165	127	100	81
0.95000	0.65788775		225,625	9,025	2,256	1,003	564	361	251	184	141	111	90
1.00000	0.68268949		250,000	10,000	2,500	1,111	625	400	278	204	156	123	100
1.05000	0.70628189		275,625	11,025	2,756	1,225	689	441	306	225	172	136	110
1.10000	0.72866788		302,500	12,100	3,025	1,344	756	484	336	247	189	149	121
1.15000	0.74985613		330,625	13,225	3,306	1,469	827	529	367	270	207	163	132
1.20000	0.76986066		360,000	14,400	3,600	1,600	900	576	400	294	225	178	144
1.25000	0.78870045		390,625	15,625	3,906	1,736	977	625	434	319	244	193	156
1.30000	0.80639903		422,500	16,900	4,225	1,878	1,056	676	469	345	264	209	169
1.35000	0.82298402		455,625	18,225	4,556	2,025	1,139	729	506	372	285	225	182
1.40000	0.83848668		490,000	19,600	4,900	2,178	1,225	784	544	400	306	242	196
1.45000	0.85294148		525,625	21,025	5,256	2,336	1,314	841	584	429	329	260	210
1.50000	0.86638560		562,500	22,500	5,625	2,500	1,406	900	625	459	352	278	225
1.55000	0.87885848		600,625	24,025	6,006	2,669	1,502	961	667	490	375	297	240
1.60000	0.88474355		640,000	25,600	6,400	2,844	1,600	1,024	711	522	400	316	256
1.65000	0.89583744		680,625	27,225	6,806	3,025	1,702	1,089	756	556	425	336	272
1.70000	0.90606577		722,500	28,900	7,225	3,211	1,806	1,156	803	590	452	357	289
1.75000	0.91547253		765,625	30,625	7,656	3,403	1,914	1,225	851	625	479	378	306
1.80000	0.92410211		810,000	32,400	8,100	3,600	2,025	1,296	900	661	506	400	324
1.85000	0.93199897		855,625	34,225	8,556	3,803	2,139	1,369	951	698	535	423	342
1.90000	0.93920728		902,500	36,100	9,025	4,011	2,256	1,444	1,003	737	564	446	361
1.95000	0.94577064		950,625	38,025	9,506	4,225	2,377	1,521	1,056	776	594	469	380
2.00000	0.95173185		1,000,000	40,000	10,000	4,444	2,500	1,600	1,111	816	625	494	400
2.05000	0.95713263		1,050,625	42,025	10,506	4,669	2,627	1,681	1,167	858	657	519	420
2.10000	0.96201346		1,102,500	44,100	11,025	4,900	2,756	1,764	1,225	900	689	544	441
2.15000	0.96641339		1,155,625	46,225	11,556	5,136	2,889	1,849	1,284	943	722	571	462
2.20000	0.97036988		1,210,000	48,400	12,100	5,378	3,025	1,936	1,344	988	756	598	484
2.25000	0.97391876		1,265,625	50,625	12,656	5,625	3,164	2,025	1,406	1,033	791	625	506
2.30000	0.97709407		1,322,500	52,900	13,225	5,878	3,306	2,116	1,469	1,080	827	653	529
2.35000	0.97992804		1,380,625	55,225	13,806	6,136	3,452	2,209	1,534	1,127	863	682	552
2.40000	0.98245105		1,440,000	57,600	14,400	6,400	3,600	2,304	1,600	1,176	900	711	576
2.45000	0.98469161		1,500,625	60,025	15,006	6,669	3,752	2,401	1,667	1,225	938	741	600
2.50000	0.98667638		1,562,500	62,500	15,625	6,944	3,906	2,500	1,736	1,276	977	772	625
2.55000	0.98843017		1,625,625	65,025	16,256	7,225	4,064	2,601	1,806	1,327	1,016	803	650
2.60000	0.98997599		1,690,000	67,600	16,900	7,511	4,225	2,704	1,878	1,380	1,056	835	676
2.65000	0.99067762		1,755,625	70,225	17,556	7,803	4,389	2,809	1,951	1,433	1,097	867	702
2.70000	0.99195082		1,822,500	72,900	18,225	8,100	4,556	2,916	2,025	1,488	1,139	900	729
2.75000	0.99306605		1,890,625	75,625	18,906	8,403	4,727	3,025	2,101	1,543	1,182	934	756
2.80000	0.99404047		1,960,000	78,400	19,600	8,711	4,900	3,136	2,178	1,600	1,225	968	784
2.85000	0.99488974		2,030,625	81,225	20,306	9,025	5,077	3,249	2,256	1,658	1,269	1,003	812
2.90000	0.99562808		2,102,500	84,100	21,025	9,344	5,256	3,364	2,336	1,716	1,314	1,038	841
2.95000	0.99626837		2,175,625	87,025	21,756	9,669	5,439	3,481	2,417	1,776	1,360	1,074	870
3.00000	0.99682226		2,250,000	90,000	22,500	10,000	5,625	3,600	2,500	1,837	1,406	1,111	900

Sometimes one wishes to specify in advance the confidence probability P (this is equivalent to specifying X because P and X are in one-to-one correspondence -- use the above table) and the random sample size n. In this case, the margin of error d is given by

(10) $$d = (X/2)(1/\sqrt{n}).$$

Stratified Random Sampling

Instead of sampling the whole population as a single entity, it is useful to partition the population and then to randomly sample each set in the partition independently. We now describe how this can be done.

Let U be a finite population and let $(U_1,..., U_N)$ be a partition of U. Let p be the probability that an element of U has property A, and let p_i be the probability that an element of U_i has property A, i = 1,..., N. Furthermore, let $U_{Ai} \subset U_i$ be the set of elements in U_i having property A, and $U_A \subset U$ be the set of elements in U having property A. Since $(U_1,..., U_N)$ is a partition,

(10₁) $$\#U_A = \sum_{i=1}^{N} \#U_{Ai}, \quad p_i = \frac{\#U_{Ai}}{\#U_i}, \text{ and } p = \frac{\#U_A}{\#U} = \sum_{i=1}^{N} \frac{\#U_{Ai}}{\#U} = \sum_{i=1}^{N} \left(\frac{\#U_{Ai}}{\#U_i}\right)\left(\frac{\#U_i}{\#U}\right) = \sum_{i=1}^{N} p_i w_i,$$

where $w_i = \frac{\#U_i}{\#U}$, i = 1,..., N.

Note that $\sum_{i=1}^{N} w_i = 1$.

Let us now consider the case where each U_i is large enough so that the DeMoivre-Laplace Limit Theorem may be applied to random samples taken from each U_i. Thus, let $s_i \subset U_i$ be a random sample of size n_i so that

(10_2) $$p_i - d_i \le \frac{k_i}{n_i} \le p_i + d_i \text{ (equivalently, } n_ip_i - n_id_i \le k_i \le n_ip_i + n_id_i\text{). } i = 1,..., N,$$ where d_i is the margin of error.

Multiplying by w_i and then summing from i = 1 to i = N leads to

(10_3) $$\sum_{i=1}^{N} w_i p_i - \sum_{i=1}^{N} w_i d_i \le \sum_{i=1}^{N} \frac{k_i}{n_i} w_i \le \sum_{i=1}^{N} w_i p_i + \sum_{i=1}^{N} w_i d_i, \text{ or (because of } (10_1))$$
$$p - \sum_{i=1}^{N} w_i d_i \le \sum_{i=1}^{N} \frac{k_i}{n_i} w_i \le p + \sum_{i=1}^{N} w_i d_i, \text{ or equivalently,}$$
$$\sum_{i=1}^{N} \frac{k_i}{n_i} w_i - \sum_{i=1}^{N} w_i d_i \le p \le \sum_{i=1}^{N} \frac{k_i}{n_i} w_i + \sum_{i=1}^{N} w_i d_i, \text{ or equivalently,}$$
$$\left| p - \sum_{i=1}^{N} \frac{k_i}{n_i} w_i \right| \le \sum_{i=1}^{N} w_i d_i, \text{ which, by use of (10), is equivalent to}$$
$$\left| p - \sum_{i=1}^{N} \frac{k_i}{n_i} w_i \right| \le \sum_{i=1}^{N} w_i \frac{X_i}{2} \frac{1}{\sqrt{n_i}} .$$

Notice that the last inequality gives a range on p, the population-at-large probability of having property A.

If we choose each d_i to be the same, say d, then the fourth inequality becomes

(10_4) $$\sum_{i=1}^{N} \frac{k_i}{n_i} w_i - d \le p \le \sum_{i=1}^{N} \frac{k_i}{n_i} w_i + d, \text{ or } \left| p - \sum_{i=1}^{N} \frac{k_i}{n_i} w_i \right| \le d .$$

Consider the collection of finite probability spaces, $M_i(n_i, p_i, m_i)$, i = 1,..., N, and form the product probability measure space

(10_5) $$M(n, m) = \prod_{i=1}^{N} M_i(n_i, p_i, m_i), \text{ where } n = \sum_{i=1}^{N} n_i .$$

Each element of the product space has the form $e = (e_1,...,e_i,...,e_N)$, where $e_i \in M_i(n_i, p_i, m_i)$, i = 1,..., N, and the product probability measure m, on singleton subsets, is given by $m(\{e\}) = \prod_{i=1}^{N} m_i(\{e_i\})$. (Recall that, from the definition of the probability of a singleton subset given above equation (1), $m_i(\{e_i\}) = p_i^{k_i}(1 - p_i)^{n_i - k_i}$.) Finally, if $E = (E_1,...,E_i,...,E_N)$, where $E_i \subset M_i(n_i, p_i, m_i)$, i = 1,..., N, is any subset of M(n, m), then $m(E) = \prod_{i=1}^{N} m_i(E_i)$.

Since $d_i = d$ for all i, (10_2) becomes $n_ip_i - n_id \le k_i \le n_ip_i + n_id$ for all i. Thus (7) becomes

(10_6) $$m_i\Big(\bigcup_{n_ip_i - n_id \le k_i \le n_ip_i + n_id} E_{n_i,k_i,p_i}\Big) \text{ for all i.}$$

By the DeMoivre-Laplace Limit Theorem (for each n_i sufficiently large), the latter is equal to (see (8))

(10_7) $$\frac{1}{\sqrt{2\pi}} \int_{X_{-i}}^{X_{+i}} e^{-t^2/2} dt, \quad where\ X_{\pm i} = \frac{\pm d\sqrt{n_i}}{\sqrt{p_i(1-p_i)}} \text{ for all i.}$$

Finally, the probability that (10_4) holds is given by

(10_8) $$\prod_{i=1}^{N} m_i\Big(\bigcup_{n_ip_i - n_id \le k_i \le n_ip_i + n_id} E_{n_i,k_i,p_i}\Big) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \int_{X_{-i}}^{X_{+i}} e^{-t^2/2} dt,$$
$$where\ X_{\pm i} = \frac{\pm d\sqrt{n_i}}{\sqrt{p_i(1-p_i)}} \text{ for all i.}$$

Concrete Examples of Stratified Sampling

(1) Let U be the set of voters in California who casted ballots in an election about Proposition A. Let N = 58 and let $(U_1,..., U_N)$ be the partition of U into California's 58 counties. Let d = 0.01 and let the confidence probability be P = 0.9518 for each of the counties. Thus, $\frac{1}{\sqrt{2\pi}} \int_{X_{-i}}^{X_{+i}} e^{-t^2/2} dt$ = 0.9518, and the probability that (10_4) holds is

(using (10_8)) $(0.9518)^{58} = 0.057$. The sample size in each county must be 10,000. Thus, if each county uses a random sample of 10,000 (which gives a margin of error of d = 0.01), then the confidence level for the entire state for (10_8) is only 0.057! On the other hand, if d = 0.01 and the confidence probability is P = 0.9974 for each of the counties, then the probability that (10_4) holds is $(0.9974)^{58} = 0.860$.

(2) Setting $n = \sum_{i=1}^{N} n_i$ and $k = \sum_{i=1}^{N} k_i$, and if we let $n_i = nw_i$, i = 1,..., N, then the last inequality in (10_3) becomes

$$\left| p - \frac{k}{n} \right| \leq \frac{1}{2\sqrt{n}} \sum_{i=1}^{N} X_i \sqrt{w_i} \,. \tag{10_9}$$

This is the case where the sample sizes are weighted the same as the size of each U_i is weighted to the size of U.

(3) *It is clear that the stratified method of estimating p for the whole state (with n=10,000, d = 0.01 and P = 0.9518 for each of the counties) yields a very small confidence probability, namely 0.057, while the method of taking a random sample from the whole state (say, 10,000, and with d = 0.01) yields a very high confidence probability, namely 0.9518!*

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 δ is chosen in advance, then this inequality becomes

$$(1-\delta)p \leq p_s \leq (1+\delta)p \tag{10}$$

and n_2 becomes

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

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

$$N/(1+\delta) \leq N_s \leq N/(1-\delta) \tag{12}$$

and

$$-\delta/(1-\delta) \leq (N - Ns)/N \leq \delta/(1+\delta) \leq \delta/(1-\delta), \tag{13}$$

and therefore,

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

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

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

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

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

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 USA (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 $|N - N_s| \leq 135{,}0000$), 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!