

Analysis Of Errors In Estimating The Fraction Of A Population Having A Certain Characteristic That Arise From Using Samples That Are Too Small

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

Basic Theory.

Suppose that there is a population of elements (persons, bacteria, animals, etc.) and a subset of the population (those persons having chromosome XXX, or those bacteria having left-handed helical tails, or those mice having Alzheimer's disease, 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 analyze the errors that occur in estimating p when samples are chosen to be too small.

Research journals in many areas (most notably in the life sciences - medicine, biology, neurology, to name a few) contain research summaries or conclusions based on statistical sampling that are frequently misleading, erroneous, blatantly false, or otherwise invalid because the researchers are ignorant of the fundamental mathematical facts about sampling, especially the inevitable errors that arise from using samples that are too small. The purpose of this paper is to present a comprehensive explanation of the salient mathematical facts about sampling in order that such errors can be avoided. A list of tables is given in the section below entitled "Analysis of Errors Arising from Using Small Samples" which contain quantitative information about the errors that are made when small samples are used. For example, if a random sample of 10 persons is used to test for a certain characteristic of the population from which the sample was taken, and if this characteristic is truly possessed by 20% of the population, then the probability that exactly 20% (=2) of the sample is found to have this characteristic is only 30.2% and the probability is 69.8% that the characteristic held by the sample is 10% or more away from the true 20%. In the absence of this data, one has no idea of the meaning of a statement such as: "Based on a study of 10 randomly selected men from the population of men who are 60 years and older, we find that 40% of this population have prostate cancer."

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 U 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,

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

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

$$(4) \quad \lim_{n \rightarrow \infty} \text{prob}\left(\bigcup_{k=k_1}^{k_2} E_{n,k,p}\right) = A(X_2) - A(X_1),$$

where

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

and where $A(X)$ is the area from minus infinity to X under the normal curve

$$(6) \quad f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \quad -\infty < t < \infty.$$

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 $p_s = k/n$ 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

$$(7) \quad \text{prob}(p - d \leq k/n \leq p + d) = \text{prob}(p - d \leq p_s \leq p + d) = \text{prob}(p_s - d \leq p_s \leq p_s + d) = P,$$

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.). **Note that the probability that p_s is within d of p is the same as the probability that p is within d of p_s !**

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 (7) is equal to

$$(8) \quad \text{prob}\left(\bigcup_{k=np-nd}^{np+nd} E_{n,k,p}\right),$$

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

$$(9) \quad A\left(\frac{d\sqrt{n}}{\sqrt{p(1-p)}}\right) - A\left(\frac{-d\sqrt{n}}{\sqrt{p(1-p)}}\right).$$

From tables of areas under the normal curve, it is known that $A(X) - A(-X) = 0.6826, 0.9544, \text{ or } 0.9974$, according as $X = 1, 2, \text{ or } 3$, respectively. Thus, if we choose our sample size n so that

$$(10) \quad \frac{d\sqrt{n}}{\sqrt{p(1-p)}} = X, \text{ where } X = 1, 2, \text{ or } 3, \text{ respectively.}$$

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 tolerance d , by choosing $n \geq 1/4(X/d)^2$, for $X = 1, 2, \text{ or } 3$, respectively, for all values of p .

SUMMARY: No matter what the true value of p is, choose $n \geq 1/4(X/d)^2$ for $X = 1, 2, \text{ or } 3$, respectively, and one gets
 $\text{prob}(p - d \leq k/n \leq p + d) = \text{prob}(p_s - d \leq p \leq p_s + d) = 0.68268, 0.95450, \text{ or } 0.99730$, respectively. For $X = 1.645$ or 2.327 , respectively,
 $\text{prob}(p - d \leq k/n \leq p + d) = \text{prob}(p_s - d \leq p \leq p_s + d) = 0.95000$ or 0.99000 , respectively.

Various other confidence probabilities, P , can be used by choosing X according to the following table, and then calculating $n = 1/4(X/d)^2$ using the corresponding value of X .

TABLE OF VALUES OF CONFIDENCE PROBABILITIES, $P (= A(X) - A(-X))$, FOR VARIOUS VALUES OF X

X	0.250	0.500	0.750	1.000	1.250	1.500	1.598	1.645	1.695	1.750	1.751	1.812
P	0.19741	0.38292	0.54674	0.68269	0.78870	0.86639	0.89000	0.90000	0.91000	0.91988	0.92000	0.93000
X	1.881	1.960	2.000	2.054	2.170	2.250	2.326	2.500	2.576	2.750	2.807	3.000
P	0.94000	0.95000	0.95450	0.96000	0.97000	0.97555	0.98000	0.98758	0.99000	0.99404	0.99500	0.99730

TABLE OF MINIMUM RANDOM SAMPLE SIZES, $n = 1/4(X/d)^2$

The entries are the random sample size such that the probability is P that the sample estimate is within d of the true p for the whole population						
		Values for d , the margin of error				
X	P	0.01	0.02	0.03	0.04	0.05
1.000	0.68269	2500	625	278	156	100
1.250	0.78870	3906	977	434	244	156
1.500	0.86639	5625	1406	625	352	225
1.598	0.89000	6384	1596	709	399	255
1.645	0.90000	6765	1691	752	423	271
1.695	0.91000	7183	1796	798	449	287
1.750	0.91988	7656	1914	851	479	306
1.751	0.92000	7665	1916	852	479	307
1.812	0.93000	8208	2052	912	513	328
1.881	0.94000	8845	2211	983	553	354
1.960	0.95000	9604	2401	1067	600	384
2.000	0.95450	10000	2500	1111	625	400
2.054	0.96000	10547	2637	1172	659	422
2.170	0.97000	11772	2943	1308	736	471
2.250	0.97555	12656	3164	1406	791	506
2.326	0.98000	13526	3381	1503	845	541
2.500	0.98758	15625	3906	1736	977	625
2.576	0.99000	16589	4147	1843	1037	664
2.750	0.99404	18906	4727	2101	1182	756
2.807	0.99500	19698	4925	2189	1231	788
3.000	0.99730	22500	5625	2500	1406	900

Analysis of Errors Arising from Using Small Samples.

The following tables give the probabilities that the sample estimate differs from the true p (for the whole population under consideration) for random sample sizes of 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100, and for values of p equal to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9.

As an example of how these results should be used when a small sample is employed by a researcher, let us suppose that a random sample of size 20 is taken from a certain population. Of course, the value of p for this population is unknown in the absence of testing the entire population, and so the researcher must compare the sample estimate with the true p for various possible values of p (here chosen to be possibly 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, or 0.9). The researcher should then state the sample estimate, p_s , found in the study, and include the following table of probabilities that p_s differs from the true p for the possible range of values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 for the true p :

n=20								
p =	p =	p =	p =	p =	p =	p =	p =	p =
0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
28.52%	21.82%	19.16%	17.97%	17.62%	17.97%	19.16%	21.82%	28.52%
74.55%	59.81%	53.48%	50.53%	49.66%	50.53%	53.48%	59.81%	74.55%
25.45%	40.19%	46.52%	49.47%	50.34%	49.47%	46.52%	40.19%	25.45%
The first row is the probability that the sample estimate equals p								
The second row is the probability that the sample estimate is within 5% of p								
The third row is the probability that the sample estimate exceeds p by at least 10%								

The complete set of these tables is now listed.

n=10								
p = 0.1	p = 0.2	p = 0.3	p = 0.4	p = 0.5	p = 0.6	p = 0.7	p = 0.8	p = 0.9
38.74%	30.20%	26.68%	25.08%	24.61%	25.08%	26.68%	30.20%	38.74%
61.26%	69.80%	73.32%	74.92%	75.39%	74.92%	73.32%	69.80%	61.26%

The first row is the probability that the sample estimate equals p

The second row is the probability that the sample estimate exceeds p by at least 10%

n=20								
p = 0.1	p = 0.2	p = 0.3	p = 0.4	p = 0.5	p = 0.6	p = 0.7	p = 0.8	p = 0.9
28.52%	21.82%	19.16%	17.97%	17.62%	17.97%	19.16%	21.82%	28.52%
74.55%	59.81%	53.48%	50.53%	49.66%	50.53%	53.48%	59.81%	74.55%
25.45%	40.19%	46.52%	49.47%	50.34%	49.47%	46.52%	40.19%	25.45%

The first row is the probability that the sample estimate equals p

The second row is the probability that the sample estimate is within 5% of p

The third row is the probability that the sample estimate exceeds p by at least 10%

n=30								
p = 0.1	p = 0.2	p = 0.3	p = 0.4	p = 0.5	p = 0.6	p = 0.7	p = 0.8	p = 0.9
23.61%	17.95%	15.73%	14.74%	14.45%	14.74%	15.73%	17.95%	23.61%
64.08%	50.56%	44.90%	42.30%	41.53%	42.30%	44.90%	50.56%	64.08%
35.92%	49.44%	55.10%	57.70%	58.47%	57.70%	55.10%	49.44%	35.92%

The first row is the probability that the sample estimate equals p

The second row is the probability that the sample estimate is within 3.33% of p

The third row is the probability that the sample estimate exceeds p by at least 6.67%

n=40								
p = 0.1	p = 0.2	p = 0.3	p = 0.4	p = 0.5	p = 0.6	p = 0.7	p = 0.8	p = 0.9
20.59%	15.60%	13.66%	12.79%	12.54%	12.79%	13.66%	15.60%	20.59%
57.09%	44.59%	39.45%	37.11%	36.42%	37.11%	39.45%	44.59%	57.09%
42.91%	55.41%	60.55%	62.89%	63.58%	62.89%	60.55%	55.41%	42.91%

The first row is the probability that the sample estimate equals p

The second row is the probability that the sample estimate is within 3.33% of p

The third row is the probability that the sample estimate exceeds p by at least 6.67%

n=50								
p = 0.1	p = 0.2	p = 0.3	p = 0.4	p = 0.5	p = 0.6	p = 0.7	p = 0.8	p = 0.9
18.49%	13.98%	12.23%	11.46%	11.23%	11.46%	12.23%	13.98%	18.49%
51.99%	40.33%	35.60%	33.45%	32.82%	33.45%	35.60%	40.33%	51.99%
48.01%	59.67%	64.40%	66.55%	67.18%	66.55%	64.40%	59.67%	48.01%

The first row is the probability that the sample estimate equals p

The second row is the probability that the sample estimate is within 2% of p

The third row is the probability that the sample estimate exceeds p by at least 4%

n=60								
p = 0.1	p = 0.2	p = 0.3	p = 0.4	p = 0.5	p = 0.6	p = 0.7	p = 0.8	p = 0.9
16.93%	12.78%	11.18%	10.47%	10.26%	10.47%	11.18%	12.78%	16.93%
48.06%	37.10%	32.69%	30.70%	30.11%	30.70%	32.69%	37.10%	48.06%
51.94%	62.90%	67.31%	69.30%	69.89%	69.30%	67.31%	62.90%	51.94%
72.10%	58.02%	51.84%	48.96%	48.10%	48.96%	51.84%	58.02%	72.10%
27.90%	41.98%	48.16%	51.04%	51.90%	51.04%	48.16%	41.98%	27.90%

The first row is the probability that the sample estimate equals p

The second row is the probability that the sample estimate is within 1.67% of p

The third row is the probability that the sample estimate exceeds p by at least 3.33%

The fourth row is the probability that the sample estimate is within 3.33% of p

The fifth row is the probability that the sample estimate exceeds p by at least 5%

n=70								
p = 0.1	p = 0.2	p = 0.3	p = 0.4	p = 0.5	p = 0.6	p = 0.7	p = 0.8	p = 0.9
15.70%	11.85%	10.36%	9.70%	9.50%	9.70%	10.36%	11.85%	15.70%
44.90%	34.54%	30.40%	28.53%	27.98%	28.53%	30.40%	34.54%	44.90%
55.10%	65.46%	69.60%	71.47%	72.02%	71.47%	69.60%	65.46%	55.10%
68.26%	54.48%	48.53%	45.78%	44.96%	45.78%	48.53%	54.48%	68.26%
31.74%	45.52%	51.47%	54.22%	55.04%	54.22%	51.47%	45.52%	31.74%

The first row is the probability that the sample estimate equals p

The second row is the probability that the sample estimate is within 1.43% of p

The third row is the probability that the sample estimate exceeds p by at least 2.86%

The fourth row is the probability that the sample estimate is within 2.86% of p

The fifth row is the probability that the sample estimate exceeds p by at least 4.29%

n=80								
p = 0.1	p = 0.2	p = 0.3	p = 0.4	p = 0.5	p = 0.6	p = 0.7	p = 0.8	p = 0.9
10.32%	11.09%	9.70%	9.07%	8.89%	9.07%	9.70%	11.09%	10.32%
42.30%	32.45%	28.53%	26.76%	26.24%	26.76%	28.53%	32.45%	42.30%
57.70%	67.55%	71.47%	73.24%	73.76%	73.24%	71.47%	67.55%	57.70%
64.97%	51.51%	45.78%	43.14%	42.36%	43.14%	45.78%	51.51%	64.97%
35.03%	48.49%	54.22%	56.86%	57.64%	56.86%	54.22%	48.49%	35.03%
The first row is the probability that the sample estimate equals p								
The second row is the probability that the sample estimate is within 1.25% of p								
The third row is the probability that the sample estimate exceeds p by at least 2.50%								
The fourth row is the probability that the sample estimate is within 2.50% of p								
The fifth row is the probability that the sample estimate exceeds p by at least 3.75%								
n=90								
p = 0.1	p = 0.2	p = 0.3	p = 0.4	p = 0.5	p = 0.6	p = 0.7	p = 0.8	p = 0.9
13.89%	10.46%	9.14%	8.56%	8.39%	8.56%	9.14%	10.46%	13.89%
40.10%	30.69%	26.96%	25.29%	24.80%	25.29%	26.96%	30.69%	40.10%
59.90%	69.31%	73.04%	74.71%	75.20%	74.71%	73.04%	69.31%	59.90%
62.10%	48.97%	43.45%	40.91%	40.16%	40.91%	43.45%	48.97%	62.10%
37.90%	51.03%	56.55%	59.09%	59.84%	59.09%	56.55%	51.03%	37.90%
The first row is the probability that the sample estimate equals p								
The second row is the probability that the sample estimate is within 1.11% of p								
The third row is the probability that the sample estimate exceeds p by at least 2.22%								
The fourth row is the probability that the sample estimate is within 2.22% of p								
The fifth row is the probability that the sample estimate exceeds p by at least 3.33%								
n=100								
p = 0.1	p = 0.2	p = 0.3	p = 0.4	p = 0.5	p = 0.6	p = 0.7	p = 0.8	p = 0.9
13.19%	9.93%	8.68%	8.12%	7.96%	8.12%	8.68%	9.93%	13.19%
38.22%	29.19%	25.63%	24.03%	23.56%	24.03%	25.63%	29.19%	38.22%
61.78%	70.81%	74.37%	75.97%	76.44%	75.97%	74.37%	70.81%	61.78%
59.58%	46.77%	41.44%	38.99%	38.27%	38.99%	41.44%	46.77%	59.58%
40.42%	53.23%	58.56%	61.01%	61.73%	61.01%	58.56%	53.23%	40.42%
75.90%	61.86%	55.49%	52.49%	51.59%	52.49%	55.49%	61.86%	75.90%
24.10%	38.14%	44.51%	47.51%	48.41%	47.51%	44.51%	38.14%	24.10%
86.99%	74.01%	67.40%	64.16%	63.18%	64.16%	67.40%	74.01%	86.99%
13.01%	25.99%	32.60%	35.84%	36.82%	35.84%	32.60%	25.99%	13.01%
93.64%	83.21%	77.04%	73.86%	72.87%	73.86%	77.04%	83.21%	93.64%
6.36%	16.79%	22.96%	26.14%	27.13%	26.14%	22.96%	16.79%	6.36%
The first row is the probability that the sample estimate equals p								
The second row is the probability that the sample estimate is within 1% of p								
The third row is the probability that the sample estimate exceeds p by at least 2%								
The fourth row is the probability that the sample estimate is within 2% of p								
The fifth row is the probability that the sample estimate exceeds p by at least 3%								
The sixth row is the probability that the sample estimate is within 3% of p								
The seventh row is the probability that the sample estimate exceeds p by at least 4%								
The eighth row is the probability that the sample estimate is within 4% of p								
The ninth row is the probability that the sample estimate exceeds p by at least 5%								
The tenth row is the probability that the sample estimate is within 5% of p								
The eleventh row is the probability that the sample estimate exceeds p by at least 6%								

Of course, in order to avoid the serious pitfalls of using small samples, the optimum thing to do is to use a larger sample whose size is given in the "TABLE OF MINIMUM RANDOM SAMPLE SIZES, $n = 1/4(X/d)^2$ ". Usually, one wishes to have a confidence probability, P, of 95% or 99%, and a margin of error of 1%, 2%, or 3%. From this table, one sees that the random sample size must be 9604, 2401, or 1067, respectively, for P = 95% and a margin of error 1%, 2%, or 3%, respectively, and 16589, 4147, or 1843 respectively, for P = 99% and a margin of error 1%, 2%, or 3%, respectively.

In this paper, the values of $\text{prob}(E_{n,k,p})$ were calculated to 16 decimal place accuracy by the recursion relation $\text{prob}(E_{n,k+1,p}) = [(n - k)p / (k + 1)(1 - p)] \text{prob}(E_{n,k,p})$, $k = 0, 1, \dots, n - 1$.