A SIMPLE APPROXIMATION FOR THE HYPERGEOMETRIC PROBABILITY: CASE (0|N,n,k).

Irving Gedanken, Board of Governors of the Federal Reserve System.

The hypergeometric probability distribution is given for exactly x occurrences in a sample of n items without replacement.

$$p(x) = p(x|N,n,k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} =$$

$$\frac{k!n!(N-k)!(N-n)!}{(k-x)!(n-x)!x!N!(N-n-k+x)!} \quad . \qquad (1)$$

Computations involving the hypergeometric probability distribution have been notoriously tedious. Most authors present the formula, give some simple examples and then advise the reader to approximate, using the binomial, Poisson or normal distributions. Unfortunately, for simple solutions of the hypergeometric, the user generally would prefer equation (1) solved for n. The problem becomes more complex when it is not the probability p(x) that needs estimating but rather the sample size n, given the other parameters of the distribution.

This paper presents an easy solution for sample size n for that special case when none with the characteristic in question are found in the sample, that is x = 0 in equation (1). What size sample (n) need one take from the population (N), so that if a characteristic is absent from the sample, one can feel 1-p(0) per cent confident that there are no more than k such events in the population?

It is in the very nature of sampling that one can only surmise from an absence in a sample, x = 0, a similar absence in the population.

Having recorded the sunrise each day, from a size n sampling of recorded history, if no other data were available concerning the movements of the heavenly spheres, one could only hypothesize as follows,--in all of recorded history it is highly unlikely that the failure of the sun to rise occurred more than k times. I cite this example to emphasize that in most instances zero in the sample will be because of zero in the population. However, it is comforting on a cold winter night to be able to set some outside limits to the possibility that there may be events in the population missed by the sample--especially when one has other bases for questioning the zero results obtained from the sample.

The hypergeometric formula stated previously provides us with the exact probabilities of 0, 1, 2, etc., of the k events in the population showing up in a sample of size n. If the chance is p(0) of zero events in the sample, when there are k events in the population, then we can be 1- p(0) per cent confident that true k is at least no more than hypothesized k.

The probability of exactly zero events in the sample when there are k events in the population is

$$p(0) = p(0|N,n,k) = \frac{(N-k)!(N-n)!}{N!(N-n-k)!} \quad . \qquad (2)$$

Herbert Arkin in his "Handbook of Sampling for Auditing and Accounting"* using only selected values of N, n, and k, has over 25 pages of tables covering the condition x = 0. This paper will now derive an approximation which will simplify sample size determination for the zero condition. This condition is appropriate for zero acceptance sampling and for discovery or exploratory sampling.

Formula (2) can be expanded as shown in formula (3) as the product of k terms in both the numerator and the denominator. Each term is one less than the preceding term.

$$p(0) = \frac{(N-k)!(N-n)!}{N!(N-n-k)!} =$$

$$\frac{(N-k)!(N-n)(N-n-1)\ldots(N-n[k-1])(N-k)!}{N(n-1)\ldots(N-[k-1])(N-k)!(N-n-k)!} =$$

$$\frac{(N-n)(N-n-1)\ldots(N-n-k+1)}{N(n-1)\ldots(N-k+1)} \quad . \qquad (3)$$

One approach, as our chairman Mr. Raff pointed out, is that a product of k equally spaced positive numbers may be approximated by the kth power of their arithmetic mean if the spread of the numbers is not too great.

If you recall, the sum of an arithmetic progression is

$$S = \frac{n}{2}(A + L)$$

and

$$\text{Arith. Mean} = \frac{S}{n} = \frac{A + L}{2} \quad .$$

---

* Volume I - Methods, McGraw-Hill, 1963, pp. 613.

Therefore, taking the first and last terms,

$$p(0) \leq \left(\frac{N-n-\frac{k-1}{2}}{N-\frac{k-1}{2}} \right)^k \leq \left(1-\frac{n}{N-\frac{k-1}{2}} \right)^k , \quad (4)$$

where p(0) is the kth power of the geometric mean of the k terms in both numerator and denominator.

This result was also developed by D. B. Owen, E. J. Gilbert, G. P. Steck and D. A. Young in "A Formula for Determining Sample Size in Hypergeometric Sampling When Zero Defectives are Observed in the Sample."* Solving for n in equation (4)

$$n \leq [1- p(0)^{1/k}][N- \frac{k-1}{2}]. \quad (5)$$

The key simplification is to go one step further and drop the reducing term $-(k-1)/2$. Hence

$$\frac{n}{N} \leq (1- p(0)^{1/k}) \quad (6)$$

or

$$p(0) \leq (1- \frac{n}{N})^k$$

$$p(0) = (1- \frac{n'}{N})^k \quad (7)$$

where $n' \geq n$

and $n' \leq N-k$.

Another approach which may, perhaps, be more evident to some, is to pair the terms of the numerator and denominator of equation (3).

$$p(0) = \left(\frac{N-n}{n}\right)\left(\frac{N-n-1}{N-1}\right)\left(\frac{N-n-2}{N-2}\right) \cdots \left(\frac{N-n-(k-1)}{N-(k-1)}\right)$$

$$= \left(1- \frac{n}{N}\right)\left(1- \frac{n}{N-1}\right)\left(1- \frac{n}{N-2}\right) \cdots \left(1- \frac{n}{N-(k-1)}\right). \quad (8)$$

Since each succeeding term on the right in equation (8) is less than its predecessor--

$$p(0) \leq (1- \frac{n}{N})^k$$

$$p(0) = (1- \frac{n'}{N})^k \quad \text{then} \quad (7)$$

$$\frac{\log p(0)}{k} = \log (1- \frac{n'}{N}). \quad (9)$$

Equation (9) provides a simple formula for solving for n' given N, k and the confidence level desired $(1- p(0))$.

On log log paper k, $\frac{n'}{N}$ and $1- p(0)$ show a simple linear relationship that is easy to plot and easy to read.

How much different is equation (7) from equation (4) or

$$\left(1- \frac{n'}{N}\right)^k \text{ from } \left(1- \frac{n}{N-\frac{k-1}{2}}\right)^k \; ?$$

Since both are functions of p(0):

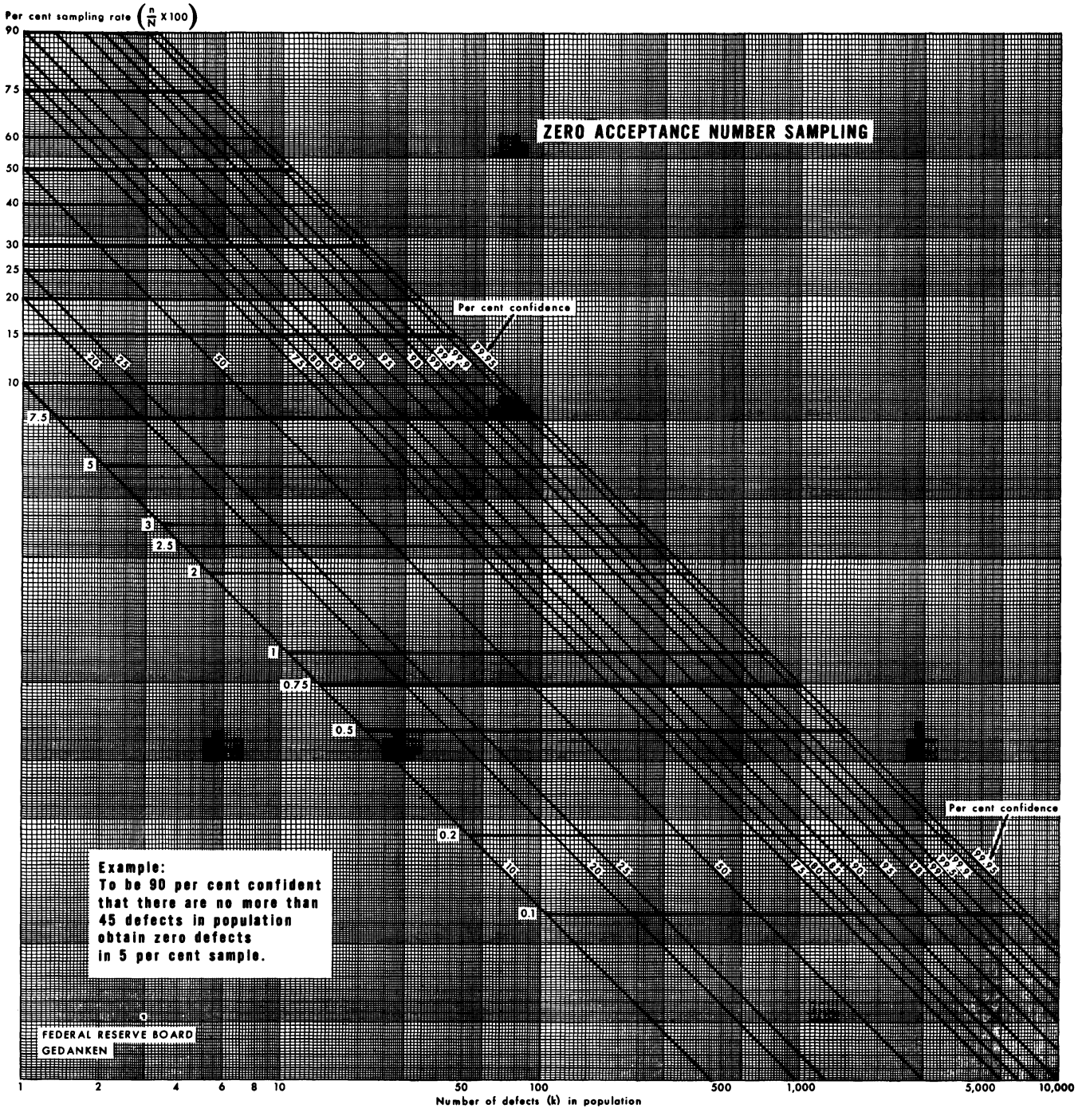
$$\frac{n'}{N} \leq \frac{n}{N-\frac{k-1}{2}} \quad (10)$$

$$n'(1- \frac{k-1}{2N}) \leq n. \quad (11)$$

It is evident that $n' = n$ when $k = 1$, and that if k is substantial, the reducing factor is approximately $1- \frac{1}{2} k/N$. It would be highly unlikely that this type of sampling would be used if k exceeded .2N or even .1N. However, even at .2N, the sample size n', as read from the chart or equation (10), would be only 10 per cent too large. In most situations of practical interest, the two estimates would be approximately equal. In any event, it is worth noting that this is an upper-bound or a "fail-safe" approximation. If it errs at all in estimating sample size it errs conservatively by requiring a size larger than is necessary. Of course, if the elegance of refinement is necessary one can always multiply the n' estimate by $(1- \frac{k-1}{2N})$.

---

* Sandia Corporation Technical Memorandum, SCTM 178-59 (51). Available from the clearinghouse for Federal Scientific and Technical Information, U. S. Department of Commerce, Washington, D. C.



Copies of this chart may be obtained from the author at the Board of Governors of the Federal Reserve System, 20th and Constitution Avenue, N. W., Washington, D. C. 20551

For the user not interested in derivations, an understanding of the chart would be worthwhile. What would a one per cent sample which produces zero defects give you 95 per cent confidence in--(looking at the chart)--that there are no more than 300 defects in the population.

Example: $N = 100,000$
$n = 1,000$
$x = 0$
$k \leq 300$ or $\frac{3}{10}$ of one per cent.

Another example is indicated on the right hand side of the chart. This is worded the way the problem is usually raised. What size sample do I need to be 90 per cent confident that there are less than $\underline{k}$ occurrences in the population when no such events occur in the sample? Answer: a 5 per cent sample.

One might also refer to the p(0) values associated with the diagonals as the producer risk that a defect will show up in the sample when there are only $\underline{k}$ defects in the population. One minus the p(0) value indicates the risk the consumer takes that there are $\underline{k}$ defects in the population when none show up in the sample.

Consider again the 5 per cent sample--the consumer risk is just about 1/2 of one per cent that there are more than 100 defects when none are found in the sample. The producer risk on the other hand is that, even if there are as few as 4 defects, the chances are 20 per cent that one will be found in a 5 per cent sample.

You will note that the sampling rate is identically the chance of picking up the one defect when there is only one defect in the population or lot. In a 5 per cent sample, the producer risk is 5 per cent of failing lots with only one defect.

As I mentioned earlier, this chart should be a useful adjunct to zero acceptance sampling and to discovery or exploratory sampling.