Minimum Sample Size Method Based on Survey Scales

Louanglath, P.I.★

About the author(s)

★Louangrath, P.I. is an Assistant Professor in Business Administration at Bangkok University, Bangkok, Thailand. He could be reached by email at: Lecturepedia@gmail.com

ABSTRACT

The objective of this paper is to introduce a new sample size calculation method based on the type of response scale used surveys. The current literature on sample size calculation focuses data attributes and distribution. There is no prior research using response scale as the basis for minimum sample size calculation. This paper fills that gap in the literature. We introduced a new minimum sample size calculation method called *n* (n-Star)* by using the Monte Carlo iteration as the basis to find asymptotic normality in the survey response scale. This new method allows us to achieve up to 95% accuracy in the sample-population inference. The data used in this study came from the numerical elements of the survey scales. Three Likert and one non-Likert scales were used to determine minimum sample size. Through Monte Carlo simulation and NK landscape optimization, we found that minimum sample size according to survey scales in all cases is $n^* = 31.61 \pm 2.33$ ($p < 0.05$). We combined four scales to test for validity and reliable of the new sample size. Validity was tested by NK landscape optimization method resulted in error of $F(z^*) = 0.001$ compared to the theoretical value for the center of the distribution curve at $F(z) = 0.00$. Reliability was tested by using Weibull system analysis method. It was found that the system drift tendency is $L = 0.00$ and system reliability $R = 1.00$.

Keywords: Sample size, Monte Carlo, NK landscape

JEL Code: B12, B13, C10, F63

CITATION:

Louangrath, P.I. (2017). "Minimum Sample Size Method Based on Survey Scales." *Inter. J. Res. Methodol. Soc. Sci.*, Vol., 3, No. 3: pp. 44–52. (Jul. – Sep. 2017); ISSN: 2415-0371.

1. INTRODUCTION

Sample size calculation is a fundamental requirement for research. The function of minimum sample size is to produce (i) fair representation of the population where sample statistics could be used for population inference, and (ii) efficient means for population studies. The first requirement helps minimize bias. The second requirements helps in resource conservation. Minimum sample size answers the questions of "how much data is enough?" (Boen and Zahn 1982, pp. 120–121). Inadequate sample may evidence inferential errors of both Type I and II (Freiman et al., 1986; and Thornley and Adams, 1998).

The objective of this paper is to provide an improved method for minimum sample size determination. Existing methods for sample size calculation remain inefficient due to higher and variable in size of the required sample. The variation may come from the pilot sample (Taylor and Muller, 1995; Muller and Benignus (1992). For example, the Yamane sample table ranges from 83–400 (Yamane, 1967). Kish (1965) recommends 30-200 samples. Sudman (1976) recommends 100 samples. These variations in sample size requirement came from data attributes (Wright, 1997). Data with different attributes requires different sample size.

This paper looks for a unified minimum sample size based on the type of survey scales. The new method proposed by this paper consists of two steps: (i) use the numerical elements of the survey scale to run Monte Carlo simulation and obtain *N* iterations, (ii) calculate minimum sample size through using the log of the Monte Carlo iteration adjusted for NK landscape optimization (Kaufman and Levin, 1987). This approach is simple and efficient in comparison to other methods available in the literature. In prior literature, minimum sample size had been calculated based on data attributes and distribution. In this paper, we calculated minimum sample size by using the survey scale. This approach is a new contribution to the literature.

2. LITERATURE REVIEW

According to the literature, minimum sample size for social science research should be 30 – 200 (Kish, 1965, p. 17). Three criteria are considered when determining sample size: the level of precision, the level of confidence or risk, and the degree of variability in the attributes being measured (Miaoulis and Michener, 1976). There are several approaches to calculating sample size.

One approach to minimum sample size determination is to categorize the type of data into continuous and non-continuous probability. For continuous distribution data, the minimum sample size may be determined by:

$$n = \frac{4\sigma^2}{\alpha^2} \tag{1}$$

where 4 = $(1.96)^2$ for 95% confidence interval; σ = estimate standard deviation obtained by $\sigma = [(\bar{X} - \mu)/Z]\sqrt{n}$ from which $\mu = (\bar{X} - T(\mathrm{S}/\sqrt{n}))$; and α = error level, i.e. for 95% CI, the error level is $\alpha = 0.05$ (Cochran, 1963).

For discrete data, the minimum sample size determination may be obtained by:

$$n = \frac{4pq}{\alpha^2} \tag{2}$$

where p = Laplace Rule of Succession or $p = (s+1)/(n+2)$, and q = probability of failure or $q = 1 - p$ (Cochran, 1963).

A second approach to minimum sample size determination is based on the population size. The population may be finite or non-finite. For finite or known population size, the Yamane method may be used (Yamane, 1967, p. 886). The Yamane equation is given by:

$$n_y = \frac{N}{1 + N(\alpha^2)} \tag{3}$$

where N = population size, and α = error level. This approach may not be appropriate where the population size is unknown or unstable. A second method for distribution-based sample size determination is known as non-finite population approach. The Yamane method had been criticized for its unreasonable assumption of normality because some data may not be normally distributed, i.e. stock market data (Officer, 1972; Fama, 1965; and Teichmoeller, 1971).

The non-finite population approach for sample size determination is given by:

$$n = \frac{Z^2 \sigma^2}{E^2} \tag{4}$$

where Z = critical value at a specified confidence interval, i.e. 1.96 for 95% CI; σ = estimated standard deviation; and E = standard error defined by $E = \sigma / \sqrt{n}$ (Cochran, 1963, p.75). This method requires the taking of pilot sample. The size of n increases as the size of the pilot sample increases. For instance, where the pilot sample size is 10, the minimum sample size is 13; if this test sample size is 20, the required minimum sample size is 27. As the test sample increases to 30, the minimum sample size may be as high as 80. Due to the instability of the minimum sample size result at various pilot sample sizes, this approach is not reliable.

In our earlier writing, we introduced the n-Omega method where the minimum sample size for non-finite population is obtained by:

$$n_\omega = \sqrt{\frac{(n_1 / 0.01) - (n_1 / 0.01)}{2}} \tag{5}$$

where $n_1 = \left((\sigma^2 n) / S^2\right) \div df(\alpha)$ and n = pilot sample. This non-finite population method produces a minimum sample size of approximately 30 counts (Louangrath, 2014; pp. 141-152).

This paper introduces a new method for calculating minimum sample size based on the type of survey response scale. This approach differs from what had been discussed in the literature because we focus on the survey scale as the basis for calculating minimum sample size. This approach is of interests to researchers in social science because it is simpler and more practical. It is simpler because it relies on the type of scale used in the survey; thus, there is no need outside pilot sample. By using the survey scale as the basis for the calculation, objectivity is ensured. The new method is practical because it is easy to understand.

3. DATA

The data used in this paper comes directly from the content of each scale. No outside data is not necessary. There are four scales selected for the study. These scales are categorized into two types: Likert and non-Likert. Likert scales include (1,2,3,4,5), (1,2,3,4,5,6,7) and (1,2,3,4,5,6,7,8,9,10) (Likert, 1932). Non-Likert scale is (0,1,2,3).

Despite disagreement in the literature over the question of whether Likert scale is quantitative or qualitative (Jamieson, 2004; and Norman, 2010), we treat Likert and non-Likert as quantitative data. Likert scales do not contain zero. Thus, they could only be subjected to continuous distribution testing (Abramowitz and Stegun, 1972). Non-Likert scale, on the other hand, contains zero. Thus, discrete and continuous distribution may be used for testing non-Likert scale or scale containing zero value. The ability of the data set to allow the type of probability analysis is not a small issue for purposes of hypothesis testing. A data set that allows both discrete and continuous distribution testing affords the researcher the flexibility and varied tools for hypothesis testing. The information obtained from such a data set is more extensive compared to the inflexible data set allowing only continuous distribution testing.

Table 1. Data characteristics of response scale commonly used in survey

Scale Type	Randomness	Skewness	Kurtosis
(0,1,2,3)	0.40	0.00	(6.12)
(1,2,3,4,5)	0.40	0.00	(3.47)
(1,2,3,4,5,6,7)	0.27	0.81	(3.81)
(1,2,3,4,5,6,7,8,9,10)	0.14	0.67	(2.96)

4. METHODOLOGY

Four response scales are used for our synthesis of minimum sample size calculation. The rationale for using these scales are that (i) they are common scales used in social science research, and (ii) the scale becomes the basis for statistical analysis in summarizing the data. The four types of scales used in this paper are Likert and non-Likert types. The Likert scale consists of (1,2,3,4,5), (1,2,3,4,5,6,7) and (1,2,3,4,5,6,7,8,9,10). The non-Likert scale is (0,1,2,3). The elements of each scale is subjected to Monte Carlo simulation to find the number of iterations in order to see normal distribution.

The Monte carlo simulation begins with a supposition that given a set of observed values: $X_i : (x_1, x_2, .., x_n)$ where n is the total observations. We are presented with a question: "if *n* is the initial observation set, what is the minimum sample size $E[n]$ in order for *n* to fairly represents the population *N?"* in order to answer this question, the following information is required: (i) maximum, minimum, and median, (ii) Monte Carlos' iteration counts, and (iii) pre-specified error level.

From the observation set $X_i : (x_1, x_2, .., x_n)$, it is necessary to determine the descriptive (maximum, minimum, $\bar{X}$ and S) and inferential statistics (μ and σ) of the set. We now have four pieces of information: maximum, minimum, median, arithmetic mean and estimated mean. These four items are designated as $W_j : (w_1, w_2, w_3 \text{ and } w_4)$. The set W_j is used to determine the number of iterations under the Monte Carlo method (Metropolis, 1987; pp. 125-130). The Monte Carlo number of iterations may be determined by:

$$N = \left(\frac{3\sigma_w}{E} \right)^2 \tag{6}$$

where $E = ((\max - \min)/2) \div 50$. The minimum sample size is obtained simply by:

$$n^* = 0.23 \left(\left| \frac{\ln(N)}{Z_{nk}} \right| \right) \tag{7}$$

where $N =$ Monte Carlo iterations, and Z_{nk} is the NK landscape simulation for optimization (Kaufman and Weinberger, 1989) obtained from:

$$Z_{nk} = \frac{F(X) - 0.50}{\sqrt{\frac{1}{12n}}} \tag{8}$$

The term $F(X)$ is the mean value of the cumulative distribution function (CDF) of $\ln(N)$ for each survey scale type.

5. FINDINGS & DISCUSSION

We examined four types of scales used in survey. Using the survey response choice as the basis for Monte Carlo simulation to obtain the number of iterations (N), the minimum sample size is about 31.61 ± 2.33 (Table 1). This number is consistent with the literature advocating minimum sample size to be 30 (Israel, 1992; Smith and Wells, 2006) where the properties of the central limit theorem are manifested (Agresti and Min, 2003)

Table 2. Minimum sample size under log Monte Carlo iteration method

Type of scale	N	$\ln(N)$	$F(Z)$	OPT	n
(0,1,2,3)	158,548	11.97	0.1251	(0.10)	29.08
(1,2,3,4,5)	281,864	12.55	0.2912	(0.10)	30.48
(1,2,3,4,5,6,7)	634,195	13.36	0.6410	(0.10)	32.45
(1,2,3,4,5,6,7,8,9,10)	1,426,938	14.17	0.8870	(0.10)	34.42
				Mean	31.61
				Standard deviation	2.33

One common method of determining minimum sample size based on the data distribution: $n = (Z^2\sigma^2)/SE^2$, produces greater variance for the final sample size. For the four scales, the minimum sample size ranges from 20.02 to 91.40 or 46.90 ± 31.99. This old method is both unreliable and inefficient (Table 2).

Table 3. Minimum sample size according to the type of survey

Type of scale	N iterations in Monte-Carlo Simulation	Minimum sample size	
		Log Monte-Carlo	$n = (Z^2\sigma^2)/SE^2$
(0,1,2,3)	158,548	29.08	20.02
(1,2,3,4,5)	281,864	30.48	27.78
(1,2,3,4,5,6,7)	634,195	32.45	48.40
(1,2,3,4,5,6,7,8,9,10)	1,426,938	34.42	91.40
		Mean	46.90
		Standard deviation	31.99

5.1 VALIDITY TEST

Validity is the test for precision. In order to pass validity test, the residual must be kept at minimal. In this study, the tolerance level for residual value was kept at 1% or requiring the proposed method to achieve 99% accuracy. The result of the calculation under the new minimum sample size calculation is the observed value. The expected value for the sample size is obtained through a log maximum likelihood method which is given in two steps: Step 1:

$$\ln L(X) = \sum\left[\left(\frac{X_i \ln F(Z)}{n-1}\right)+\left(\frac{(1-X_i)(1-F(Z))}{n^2-n-1}\right)\right] \tag{9}$$

With the estimate obtained in (9), the expected value for the minimum sample size is obtained through:

$$E[X] = \frac{1}{n-1}\sum\left[\left(\left(X_i - \ln L(X_i)\right) - \left|\frac{A}{n-1}\right|\right) \div n\right] \tag{10}$$

where $A = \ln L(X)$ in (7). The percentage probability for the precision under this method is: $Z = (31.61 - 27.98)/2.33 = 1.56$ with the corresponding $F(Z) = 0.941$ or 94.10%.

Alternatively, validity may be tested by using NK landscape simulation for optimization. The local optimum equation under NK Landscape simulation method:

$$OPT_{loc} = \frac{F(\mathrm{X}) - 0.50}{\sqrt{\frac{1}{12n}}} \tag{11}$$

where $F(X)$ = average of percentage probability ($\Phi(x_i)$) among observed values, and *n* is the test sample size. The decision rule is that the lowest optimum point is the most efficient. The observed optimum is compared to the expected value at:

$$\mu + \sigma\sqrt{\frac{2\ln(K+1)}{K+1}} \tag{12}$$

where μ = expected mean of $F_i\left(x_{i,k}\right)$, σ = expected standard deviation of $F_i\left(x_{i,k}\right)$, $K = N - 1$, and N = sample size. The theoretical optimum is:

$$OPT* = \frac{F(\hat{\mathrm{X}}) - 0.50}{\sqrt{\frac{1}{12n}}} \tag{13}$$

where $\hat{X} = \mu + \sigma\sqrt{\frac{2\ln(K+1)}{K+1}}$ for which $1 << K \leq N$ (Weinberger, 1991). The result of our calculation shows that $OPT_{loc} = 0.001$ and the theoretical value beyond which precision is loss: $OPT* = 0.40$. The target is 0.00 or the center of the distribution curve, the three scale combined as a unified method for sample size determination came near the target at 0.00 with an observed value of 0.001. Precision under this approach is 99.98%.

5.2 RELIABILITY TEST

Reliability test requires that the proposed method be reproducible. Reproducibility means that series of subsequent tests must show consistent results. Consistency is defined as minimal variance within the series. This reliability is achieved through:

$$R = 1 - CDF \tag{14}$$

where $CDF = 1 - \exp(-x/\eta)^{b}$ where x = point for evaluation; η = *exp(a)* or re-scaled of the y-intercept; and b = slope of the linear equation $Y = a + bX$ obtained through a time function QQ-plot (Weibull, 1951) with observations values: (29.08, 30.48, 32.45, 34.42). The stability of the stable is defined by the potential shift of the probability reading. This shift potential is obtained by lambda:

$$\lambda = \frac{CDF}{1 - PDF} \qquad (15)$$

Under this reliability test method, The most reliable scale is the non-Likert scale type (0,1,2,3). Among the various sample sizes obtained through the four different scales in this study (29.08, 30.48, 32.45, 34.42), lambda is 0.67 with system reliability of 0.37. Among the four scales, (0,1,2,3) is the most reliable. The corresponding minimum sample size is 29.08.

Table 4. System analysis for scale reliability

Scale Types	CDF	PDF	$\lambda = \frac{CDF}{1-PDF}$	$R = 1 - CDF$
(0,1,2,3)	0.00	0.12	0.00	1.00
(1,2,3,4,5)	0.34	0.11	0.39	0.66
(1,2,3,4,5,6,7)	0.47	0.11	0.53	0.53
(1,2,3,4,5,6,7,8,9,10)	0.58	0.11	0.66	0.42

5.3 EFFICIENCY TEST

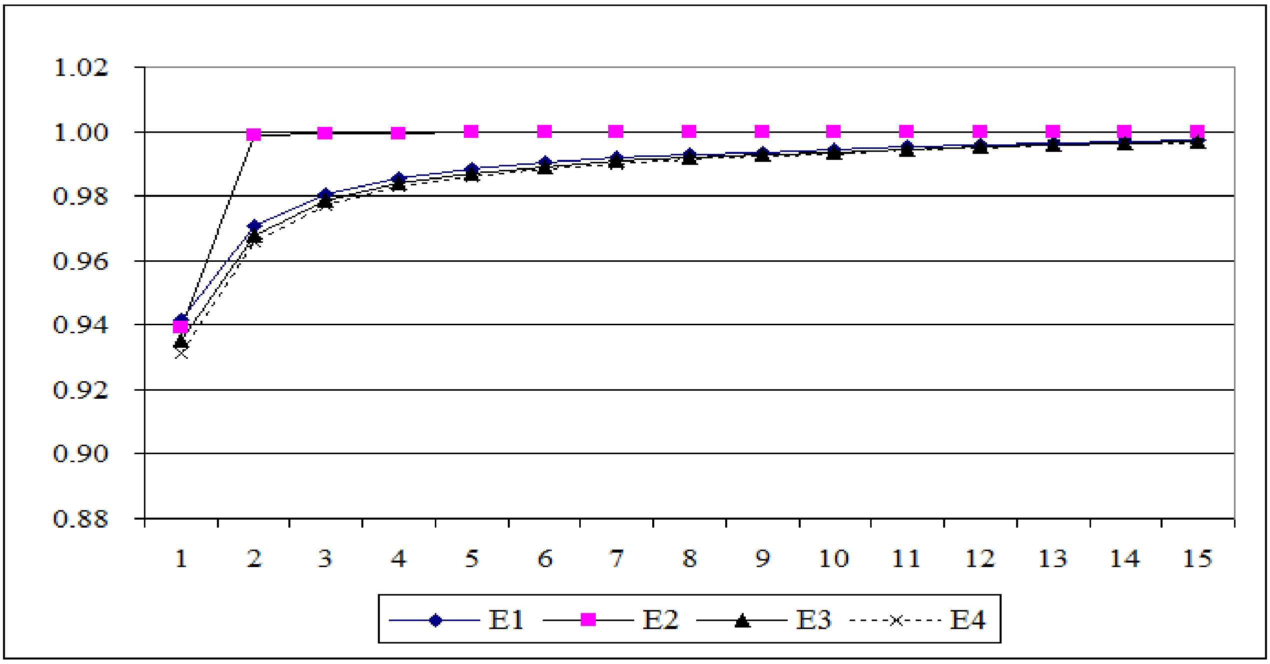
In sample size determination, efficiency is defined as achieving minimum sample size with the least resource utilization, i.e. minimal calculation steps, and shorter time. We define efficiency as "smaller sample" size in achieving the objective. Our objective is to provide a means to obtain minimum sample size that allows fair representation of the population. Efficiency is calculated as: $E = 1 - (I / O)$ where I is input or the sample size under specific type of survey scale and O is out put or the population size N.

Table 5. Efficiency of sample size relative to population size

Population: N	E1: (0,1,2,3)	E2: (1,2,3,4,5)	E3: (1,2,3,4,5,6,7)	E4: (1,2,3,4,5,6,7,8,910)
500	0.94	0.94	0.94	0.93
1,000	0.97	1.00	0.97	0.97
1,500	0.98	1.00	0.98	0.98
2,000	0.99	1.00	0.98	0.98
2,500	0.99	1.00	0.99	0.99
3,000	0.99	1.00	0.99	0.99
3,500	0.99	1.00	0.99	0.99
4,000	0.99	1.00	0.99	0.99
4,500	0.99	1.00	0.99	0.99
5,000	0.99	1.00	0.99	0.99
6,000	1.00	1.00	0.99	0.99
7,000	1.00	1.00	1.00	1.00
8,000	1.00	1.00	1.00	1.00
9,000	1.00	1.00	1.00	1.00
10,000	1.00	1.00	1.00	1.00

The asymptotic tendency of the efficiency of the sample size as the population increases is shown in the figure below.

Fig. 1: Efficiency trend for each sample size determined by various scales



6. CONCLUSION

The new minimum sample size determination method proposed by this paper could obtain an efficient size and could overcome potential bias. Efficiency was achieved through Monte Carlo simulation method. Where other method based the sample size on the error level, this new method based the sample size on the iteration counts under Monte Carlo simulation. It is efficient because the new method has shorter procedure to calculate and, thus, it saves time and resources. The new method overcomes potential bias by maintaining asymptotic normality through the used of the Monte Carlo iteration, which represents the number of repeated measurement to achieve asymptotic normal of a given data set. This new discovery is a contribution to the field because it is an efficient and reliable means for sample size determination.

REFERENCES

Agresti, Alan and Min, Yongyi (2002). "On Sample Size Guidelines for Teaching Inference about the Binomial Parameter in Introductory Statistics." http://www.stat.ufl.edu/~aa/articles/ci_proportion.pdf

Abramowitz, M. and Stegun, I. A. (Eds.) (1972). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing. New York: Dover, pp. 927 and 930.

Boen, J. R. and Zahn, D. A. (1982), The Human Side of Statistical Consulting, pp. 120-121. Lifetime Learning Publications, Belmont, CA.

Cochran, W. G. (1963). Sampling Techniques, 2nd Ed., p. 75. New York: John Wiley and Sons, Inc.

Fama, E.F. (1965). "The Behavior of Stock Market Prices." *Journal of Business*, **37** (January 1965), 34-105.

Freiman, J. A., Chalmers, T. C., Smith, Jr., H., and Kuebler, R. R. (1986). "The Importance of Beta, the Type II Error, and Sample Size in the Design and Interpretation of the Randomized Controlled Trial: Survey of 71 "Negative" Trials," in Medical Uses of Statistics, eds. J. C. Bailar III and F. Mosteller, chap. 14, pp. 289–304, NEJM Books, Waltham, Mass.

Jamieson, Susan (2004). "Likert Scales: How to (Ab)use Them." *Medical Education*, **38**(12), pp.1217-1218.

Kauffman, S.; Levin, S. (1987). "Towards a general theory of adaptive walks on rugged landscapes." *Journal of Theoretical Biology*. **128**(1): 11–45. doi:10.1016/s0022-5193(87)80029-2.
http://www.sciencedirect.com/science/article/pii/S0022519387800292?via%3Dihub

Kauffman, S.; Weinberger, E. (1989). "The NK Model of rugged fitness landscapes and its application to the maturation of the immune response." Journal of Theoretical Biology. 141 (2): 211–245. doi:10.1016/s0022-5193(89)80019-0.
http://www.sciencedirect.com/science/article/pii/S0022519389800190?via%3Dihub

Louangrath, P. (2014). "Sample Size Determination for Non-Finite Population." *Southeast-Asian J. of Sciences* Vol. 3, No. 2 (2014) pp. 141-152

Kish, Leslie (1965). Survey Sampling, p. 17. New York: John Wiley and Sons, Inc.

Metropolis, N. (1987). "The beginning of the Monte Carlo method." *Los Alamos Science* (1987 Special Issue dedicated to Stanislaw Ulam): 125–130.
http://library.lanl.gov/cgi-bin/getfile?00326866.pdf

Miaoulis, George, and Michener, R.D. (1976). An Introduction to Sampling. Dubuque, Iowa: Kendall/Hunt Publishing Company.

Muller, K. E. and Benignus, V. A. (1992), "Increasing scientific power with statistical power." *Neurotoxicology and Teratology*, **14**, 211–219.

Officer, R.R. (1972). "The Distribution of Stock Returns." *Journal of the American Statistical Association*, **67**, 340. (Dec., 1972), pp. 807-812.
http://links.jstor.org/sici?sici=0162-1459%28197212%2967%3A340%3C807%3ATDOSR%3E2.0.CO%3B2-Y

Smith, Zachary R. and Wells , Craig S. (2006). "Central Limit Theorem and Sample Size." Paper presented at the annual meeting of the Northeastern Educational Research Association, Kerhonkson, New York, October 18-20, 2006.
https://www.umass.edu/remp/Papers/Smith&Wells_NERA06.pdf

Taylor, D. J. and Muller, K. E. (1995), "Computing Confidence Bounds for Power and Sample Size of the General Linear Univariate Model." *The American Statistician*, **49**, 43–47.

Teichmoeller, J. (1971). "Distribution of Stock Price Changes." Journal of the *American Statistical Association*, **66** (June 1971), 282-4.

Thornley, B. and Adams, C. (1998). "Content and quality of 2000 controlled trials in schizophrenia over 50 years." *British Medical Journal*, **317**, 1181–1184.

Weibull, W. (1951). "A statistical distribution function of wide applicability." (PDF), *J. Appl. Mech.-Trans. ASME*, **18** (3): 293–297.
http://www.barringer1.com/wa_files/Weibull-ASME-Paper-1951.pdf

Weinberger, Edward (November 15, 1991). "Local properties of Kauffman's N-k model: A tunably rugged energy landscape." *Physical Review* **A. 10**. 44: 6399–6413. doi:10.1103/physreva.44.6399.
https://journals.aps.org/pra/abstract/10.1103/PhysRevA.44.6399

Wright, T. (1997), "A simple algorithm for tighter exact upper confidence bounds with rare attributes infinite universes." *Statistics and Probability Letters*, **36**, 59–67.

Yamane, Taro. (1967). Statistics, An Introductory Analysis, 2nd Ed., p. 886. New York: Harper and Row.