Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 177–187
https://www.ssca.org.in/journal

Sample Size Determination for Clinical Studies when Crisp Inputs are not available

Sai Sarada Vedururu[1], R.Vishnu Vardhan[2] and K.V.S. Sarma[3]
[1] *Department of Mathematics GITAM (Deemed to be) University, Hyderabad.*
[2] *Department of Statistics, Pondicherry University, Puducherry*
[3] *Former Professor of Statistics, Sri Venkeswara University, Tirupati*

Received: 04 November 2022; Revised: 25 November 2023; Accepted: 01 May 2024

Abstract

This paper deals with the classical problem of determining the minimum sample size (n) required in clinical studies for estimating the population prevalence p' of a characteristic. The popularly used formula for n requires prior knowledge on p' which may not be known as a crisp value. It can be estimated from a pilot study or specified as a range of values reflecting some uncertainty. In the first part we characterise n as a random variable whose values depend on the uncertainty in the anticipated p' modelled by a Beta distribution and thereby determine the expected sample size and its variance. In the second part of this paper we also propose a novel method to improve the formula by considering a triplet where a,b and c denote the minimum, most likely and maximum values of p', derive a new formula and show that it is more consistent than the classical method. We demonstrate the utility of the formula with illustrations and compare them with alternative ways of presenting the inputs.

Key words: Sample size; Triangular distribution; Triplet estimates; PERT.

AMS Subject Classifications: 0000

1. Introduction

Sample size determination is a basic requirement in the design and analysis of clinical studies including community trails. If enough subjects are not included in the study, the real effect or situation prevailing in the target group cannot be estimated correctly. A large sample needs more resources to achieve the desired precision than a small sample. Further, non-sampling errors erupt while executing a large study. Statistical methods offer a scientific approach to determine the minimum sample size such that sample-related risks of incorrect decisions are minimized. We discuss some interesting issues on sample size determination with clinical objectives as background, but the arguments apply to other areas too. We can broadly divide clinical studies into two viz.,

a) *Descriptive studies* in which the chief objective is to describe a population by estimating the characteristics from sample data and

Corresponding Author: R.Vishnu Vardhan
Email: vrstatsguru@gmail.com

b) *Comparative studies* in which comparison of outcomes among two or more groups (like mean or percentage) is the main objective.

There are two broad ways of summarizing any data depending on whether the outcome is a count or a measurement. In case of dichotomous categorical data, the proportion of the outcome of interest is the summary measure, denoted by p' which is called prevalence when observed over a period of time. If the events are observed in new cases, p' is called incidence or occurrence rate. An estimate of p' is $p = \frac{k}{n}$ where k subjects out of n are found to have the outcome of interest and n denotes the sample size. The true proportion in the population is however unknown unless the entire population is studied. The behaviour of p is modelled by Binomial distribution.

On the other hand, when the outcome in the sample is a measured characteristic like blood glucose level, it is summarized as the arithmetic mean (m) (or median in some cases) of the sample values along with standard deviation(s) of the values. The behaviour of m, over different samples of the population is modelled by normal distribution.

In both situations, a $100(1-\alpha)\%$ Confidence Interval (CI) can be provided around the sample estimate such that the true mean/proportion falls in this interval with desired confidence usually 95%. According to the Central Limit Theorem in statistics, the sampling distribution of both p or m tend to be normal when n is large. Hence the 95% CI for p or m are constructed making use of the standardized normal variable (Z).

1.1. Sample size for *p* based on margin of error

We briefly outline the method of determining the sample size with desired *margin of error* (d) when the objective is to estimate (a) the proportion of dichotomous outcomes or (b) the mean of a characteristic. This method is known as precision-based method because smaller margin of error leads to higher precision.

Let p be the estimate of p'obtained from a random sample of size n drawn from the population. Then the $100(1-\alpha)\%$ CI for the population prevalence is given by

$$\left\{p \text{ - } Z_{1-\alpha/2}\sqrt{\frac{p\ (1-p)}{n}}\ ,\ p + Z_{1-\alpha/2}\sqrt{\frac{p\ (1-p)}{n}}\right\} \tag{1}$$

The quantity $d = (Z_{1-\alpha/2}\sqrt{\frac{p\ (1-p)}{n}})$denotes the margin of error and $(Z_{1-\alpha/2})$is the inverse of the cumulative standard normal distribution corresponding to the chosen α(like 0.05).

The objective is to determine n such that p is contained in (1). Since the precision of the estimate increases when d is small, one way of estimating n is to keep $d \leq d_0$ where d_0 is the desired upper limit, like 5%. Solving for n in d leads to $n \geq \frac{z^2_{1-\alpha/2}p(1-p)}{d_0^2}$. Hence for a fixed choice of *d,* the expression for the minimum sample size will be

$$n = \frac{z^2_{1-\alpha/2}p(1-p)}{d^2} \tag{2}$$

Formula (2) is known as Cochran's formula (Cochran, 1977) applicable for large populations. When the population size is finite, like the number of employees of a company, then n'$= \frac{n}{1+\frac{n}{N}}$

gives the sample size corrected for finite population size. The chief input for implementing (2) is p.

1.2. Sensitivity of n to changes in p

The formula in (2) requires p as input which is known only when a pilot study is carried out. When pilot study is not possible, we can get p from previous research reports or by a personal guess. For instance, if p is known as 0.9 it means that there is 90% chance that the desired condition occurs. Then for $\alpha = 0.05$ we get $(Z_{1-\alpha/2}) = 1.96$ and with $d = 0.05$ we get $n = 385$. This value quickly drops to 96 if d is taken as 0.10, keeping other parameters unchanged. Approximating $(Z_{1-\alpha/2})$ by 2, the constant appearing in (2) is approximately 1600. When $p = 0.5$ we get $n = 400$ while the actual value with 1.96 is 385. Hence the reliability of n depends on the precision with which p known. 1 shows the pattern of n against p which is concave reaching a maximum of 385 at $p = 0.50$. We see that n changes rapidly with d but symmetric around $p = 0.5$. The decrease in n for values of p away from 0.5 may be called the *gain due to information.* In section-2 we develop a methodology to formulate the distribution of n viewing p as a random variable using beta distribution. We also study the empirical distribution of n under the chosen model by estimating its parameters, instead of using a single p value. In section 3 we develop a new method of determining n when the input value of p is not precisely known but expressed as an *interval,* along with a middle value, which we call a *triplet.* The new estimate is proposed as a weighted average of the expected sample size at each of the three elements of the triplet. We call this *triplet estimation* and study the properties this new estimate.

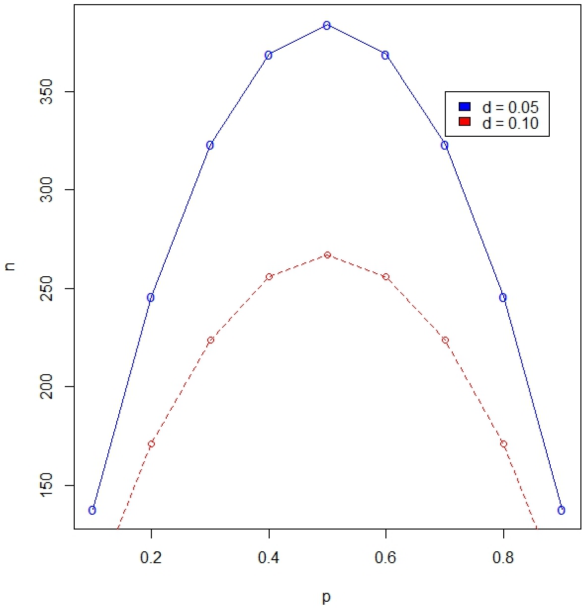

Figure 1: Sensitivity of n against p for different values of d

2. A model for the probability distribution of *n*

The formula given in (2) can be stated as $n = kp(1-p)$ where $k = \frac{z^2_{1-\alpha/2}}{d^2}$ is constant for pre-determined values of α and *d.* We wish to identify a probability distribution for n by viewing p as a continuous random variable (Y) in $[(0, 1)$ so that $n = kY(1-Y)$ and the type-1 beta distribution is natural choice for the distribution of Y specified by the density function

$$f(y) = \frac{1}{\beta(u,v)}\ y^{u-1}(1-y)^{v-1}\text{for}\quad 0<y<1, u,\ v>0 \tag{3}$$

The parameters ($u,\ v$) are related to E(Y) and V(Y) and given as

$$u = \ E(Y)\left\{\frac{E(Y)\,(1-E(Y))}{V(Y)} -\ 1\right\}\text{and } v = \ \{1-E(Y)\}\left\{\frac{E(Y)\,(1-E(Y))}{V(Y)} -\ 1\right\} \tag{4}$$

In fact p is the anticipated point-mass on the Bernoulli distribution which varies with the discretion of the researcher. When p is specified a fixed value there exists a single unique value of n from (3). Instead, we assume a probability distribution in the domain (0,1) with peak density at p so that we can account for the uncertainty in p and thereby determine the theoretical mean and variance of Y. The *triangular distribution* $(0,\ p, 1)$ is one choice for distribution of Y which help to obtain adhoc estimates of E(Y) and V(Y), while the beta-PERT distribution on *(a,b,c)* where *a = 0, b =* p and *c =* 1 is another. We use the Triangular distribution only to summarize the Bernoulli p since the truncation limits for the distribution of Y are not known at this stage. For the triangular distribution we have

$$E(Y) = \frac{\{1+p\}}{3}) \ \text{ and } \quad V(Y) = (\frac{\{p^2-p+1\}}{18} \tag{5}$$

Thus we have transformed the single anticipated p into a probability distribution and captured its mean and variance as summary. As a result, for each value of p we can uniquely identify a $\beta(u,v)$ distribution and estimate the parameters using (3).

Remark: If we use PERT (0,p,1) distribution instead of triangular distribution to estimate E(Y) and V(Y) we get u = 1+4p and v = 1+4(1-p) but (u+v) = 6 which is irrespective of u and v, which is a constraint on the parameters, not defined for the beta distribution. Hence we use triangular distribution to supply primary inputs to estimate u and v. Consider the following proposition.

Proposition-1: With Y $\sim$ Beta(u,v) the empirical distribution of n is proportional to that of Y by a constant k.

The empirical distribution of n can be obtained by simulating random deviates from $\beta(u, v)$. Table (1) gives summary of the empirical distribution of n for selected values of p, taking 95% confidence level and *d = 0.05.* This gives k = 1536.584 and the value of n rounded to the upper integer.

We observe the following from Table (1):

(a) The values of variance of n are much larger than the corresponding mean, due the fact the mean and variance of Y(1-Y) are multiplied by k and k^2 respectively.

Table 1: Empirical distribution of n with 1000 simulations.

p	n at p	E(Y)	V(Y)	(u,v)	E(n)	V(n)	Empirical Mean of n	Empirical variance of n
0.25	288	0.4166	0.0451	(1.826, 2.557)	373	4810.81	304	7548.13
0.35	350	0.4500	0.0429	(2.145, 2.621)	380	4349.40	314	5820.16
0.5	384	0.5000	0.0416	(2.500, 2.500)	384	4098.56	323	4991.42
0.65	350	0.5500	0.0429	(2.621, 2.145)	380	4348.08	312	6198.41
0.75	288	0.5833	0.0451	(2.557, 1.826)	373	4810.81	306	7059.36

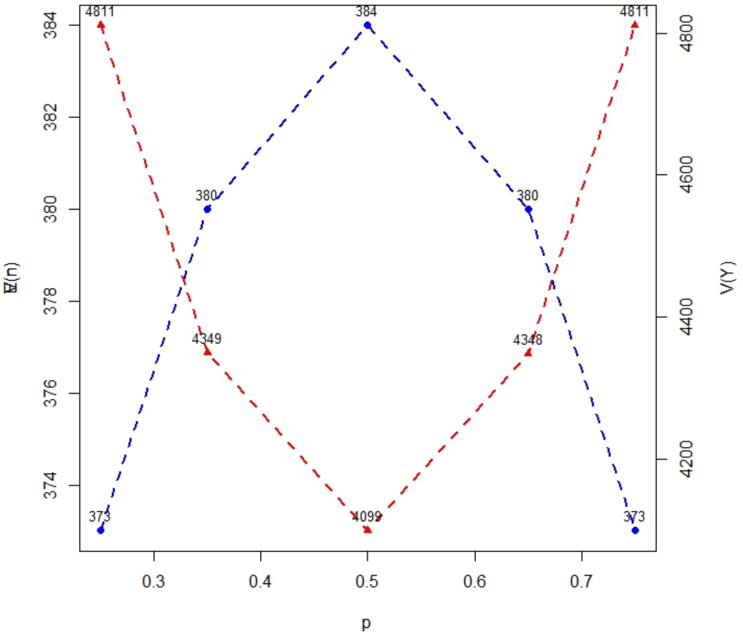

Figure 2: E(n) and V(n) as a function of p.

(b) The expected n and its variance are both symmetric around $p = 0.5$ and the empirical values also exhibit a similar pattern. When compared to the true n obtainable from (2) using the single value of p, the values of E(n) are higher and this can be because the former does not account for the impreciseness in p but E(n) takes into account a background *triangular* model to determine the mean.

(c) The empirical distribution has a shape that is similar to a beta distribution.

Figure (2) shows the pattern of E(n) and V(n) against values of p. The variance of n decreases symmetrically as p increases and reaches a minimum at $p = 0.5$ while E(n) moves in the opposite direction and reaches a maximum at the same p. In the following discussion we propose a method of summarizing the distribution of Y(1-Y) the moments of beta distribution. The empirical distribution of Y(1-Y) is shown in Figure-3.

Proposition-2: If we write T = Y(1-Y) with Y$\sim$ Beta(u,v) then the mean and variance of

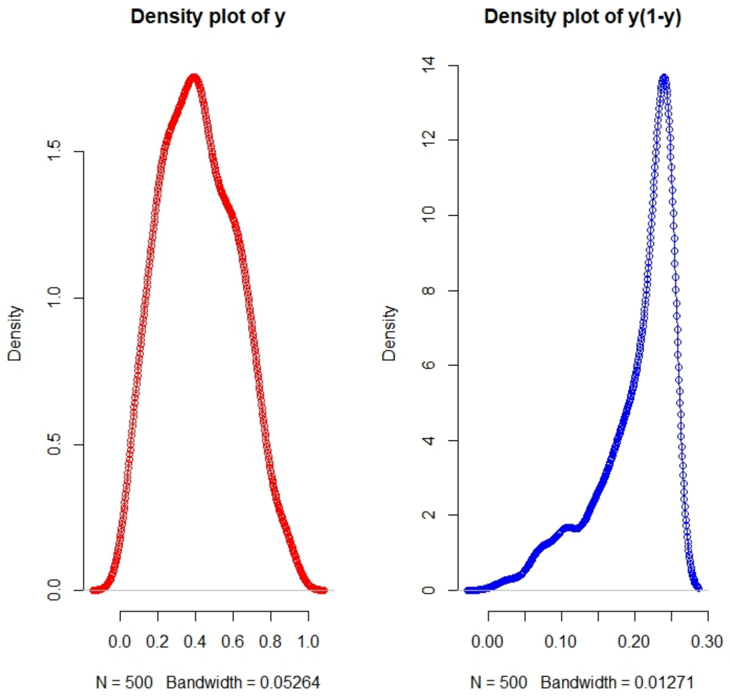

Figure 3: Empirical distribution of Y and Y(1−Y) for *p = 0.25*

T can be obtained as E(T) = E(Y){1-E(Y)} and V(T) = V(Y){1-V(Y)} which reduce to

$$E(T) = \frac{uv}{(u+v)^2} \quad \text{and} \quad V(T) = \frac{(uv)^2}{(u+v)^4(1+u+v)^2} \tag{6}$$

Further the expected sample size is

$$E(n) = k\frac{uv}{(u+v)^2} \quad \text{and} \quad V(n) = k^2\frac{(uv)^2}{(u+v)^4(1+u+v)^2} \tag{7}$$

Proof: The results follow by replacing Y and (1-Y) with their expected values and noting that $E(Y) = \frac{u}{(u+v)}$ and $E(1-Y) = \frac{v}{(u+v)}$. Similarly V(T) follows by noting that $V(Y) = \frac{uv}{(u+v)^2(1+u+v)}$ and V(1-Y) is the same as V(Y). Finally $E(n) = kE(T)$ and $V(n) = k^2V(T)$ which lead to (6) and (7). Hence the proof. With this background, we develop a new estimate of n as (i) a weighted mean of the n values obtainable at the triplet values under the beta distribution model and (ii) using PERT summary as a single input in (2).

3. The triplet estimate to handle imprecise estimates

When a single precise value of p is not available it is customary to specify the same as a triplet (p_1, p_2, p_3) where p_2 is the *most likely value* and (p_1, p_3) are the *lower* and *upper* values of p such that $p_1 < p_2 < p_3$. This approach is used in project management studies to describe the activity durations and latter summarized into mean and SD using beta distribution. Malcolm *et al.* (1959) and Clark (1962) used this approach to summarise the activity durations in project management and to estimate the time to completion the project. Books on Operations Research widely discuss this method (Taha, 2013).

Applying this logic to $(p_1,\ p_2,\ p_3)$ we obtain $p_0 = (p_1+4p_2+p_3)/6$ as the mean prevalence. If we use this p_0 in (2) we get a single value of n denoted by $n_{0.}$ Our new approach is to evaluate n at each of the three values of the triplet and summarize them as a weighted average to get a new crisp value.

We now use the method of triplet inputs to determine the sample size for estimating the population prevalence. Here is another proposition.

We now use the method of triplet inputs to determine the sample size for estimating the population prevalence. Here is another proposition.

Proposition-3: Let n_i be the sample size when the anticipated prevalence is p_i for i = 1,2,3. Then E(n_i) = $k\frac{u_i v_i}{(u_i+v_i)^2}$ and V(n_i) = $k^2\frac{(u_i v_i)^2}{(u_i+v_i)^4(1+u_i+v_i)^2}$ where (u_i, v_i) denote the parameters of the underlying beta distribution for i = 1,2,3 and k is the constant by design. Then the new estimate of n will be

$$n_{cap} = \sum_{i=1}^{3} w_i E(n_i\ ) \tag{8}$$

where $w_\mathrm{i} \geq 0$ and $w_1+w_2+w_3 = 1$. We call this the *triplet estimate* of n and $V\,(n_{cap}\ ) = \sum_{i=1}^{3} w_i^2 V(n_i\ )$. It also true that $V\,(n_{cap}\ ) \leq V(n_i\ )$ for i = 1,2,3.

One way of assigning weights is to take $w_2 = 0.5$ and $w_1 = w_3 = 0.25$ so that $E\,(n_2)$ receives more weight than the other two because p_2 is more likely valid than the other two values of the triplet. Another set of weights is {1/6. 4/6, 1/6} corresponding to {$w_{1,}$ $w_{2,}$ w_3} which are the weights used in PERT calculations.

Vardhan and Sarma (2010) have used the triplet method in the context of ROC curve analysis. Sarada *et al.* (2018); Vedururu *et al.* (2019) used this method in the context of measuring the process capability index in quality control.Venkatesu *et al.* (2019) have applied this method to redesign a control chart. In all these applications, it was found that the new estimator has lower SE than the classical point estimator.

Instead of pre-defined fixed weights, an objective way is to define weights which reflect the uncertainty in the specification of p (in terms of a triplet). We propose the following weights.

Proposition-4: The weight w_i for E(n_i) will be the ordinate of the $\beta(u,v)$ distribution at p_i for i = 1,2,3 and normalized to make the sum equal to unity.

This method allots weight as a function of p_i and hence accounts for the anticipated uncertainty in specifying p. We cannot determine the weights with PERT distribution, since the density of vanishes at p_2 and p_3 (truncation limits) and hence n_{cap} cannot be evaluated. Hence the full beta distribution without truncation will be used. Here is an illustration.

Illustration-3

Let us take $p_1 = 0.25$, $p_2 = 0.35$, $p_3 = 0.5$. From the intermediate results from Table-1 we see that the vector of means as (373, 380, 384) and the corresponding variance vector is (4810.76, 4348.74, 4099.11). The vector of weights from beta distribution with corresponding ($u_{i,}$ v_i) becomes w = (1.5492, 1.6330, 1.6976). Dividing each weight by the sum of weights and applying (8) gives n_{cap} = 380 and V(n_{cap}) = 1467.96 which smaller than the minimum of the three variances.

Table (2) shows some experimental results comparing n_{cap} with the n obtainable when we use only a single value p_2 as the input in (2).

Table 2: Triplet estimate of *n* with arbitrary window around the middle.

p_2	Triplet $(p_1,\ p_2,\ p_3)$	n_2 (at p_2)	V(n_2)	n_{cap}	$V(n_{cap})$
0.25	(0.15, 0.25, 0.35)	373	4810.76	370	1682.01
0.35	(0.25, 0.35, 0. 50)	380	4348.74	380	1467.95
0.50	(0.30, 0.50, 0.65)	384	4099.11	381	1441.74
0.65	(0.45, 0.65, 0.75)	380	4348.74	380	1471.39
0.75	(0. 50, 0.70, 0.75)	377	4548.01	379	1489.95

Suppose we take fixed weights instead of deriving from beta density. We consider two types of fixed weights and compare the resulting n_{cap} and its variance.

Table 3: Triplet estimate of *n* under different schemes of weights.

Triplet $(p_1,\ p_2,\ p_3)$	$w = \{1/6, 4/6, 1/6\}$		$w = \{1/3, 1/3, 1/3\}$	
	n_{cap}	V(n_{cap})	n_{cap}	V(n_{cap})
(0.15, 0.25, 0.35)	373	2413.01	372	1634.11
(0.25, 0.35, 0. 50)	380	2180.27	379	1473.18
(0.30, 0.50, 0.65)	383	2068.96	381	1443.98
(0.45, 0.65, 0.75)	380	2181.03	379	1476.22
(0. 50, 0.70, 0.75)	378	2268.83	378	1495.32

4. Stepwise procedure

The following is a stepwise procedure to handle the calculations.

1. Obtain the anticipated prevalence as a triplet(p_1,p_2,p_3) margin of error as d and level of significance as α For each i = 1,2,3 calculate the following.

2. Transform each p_i into as a point on triangular (0,1) distribution

3. Evaluate the trial values of mean and variance as $\mu_{1\mathrm{i}}$ and ${\sigma_{1\mathrm{i}}}^2$ respectively.

4. Identify a Beta distribution on (0,1) and estimate is parameters $(u_i,\ v_{\mathrm{i}})$ Using $\mu_{1\mathrm{i}}$ and ${\sigma_{1\mathrm{i}}}^2$ calculate $E(n_i) = k\dfrac{u_i v_i}{(\mu_i + \nu_i)^2}$ and $V(n_i) = k^2 \dfrac{(u_i v_i)^2}{(u_i + v_i)^4\,(1 + u_i + v_i)^2}$.

5. Find $w_i = \dfrac{y_i}{\sum_{i=1}^{3} y_i}$ where y_i denotes the ordinate of the Beta distribution corresponding to p_i

6. Evaluate $n_{cap} = \sum_{i=1}^{3} w_i E(n_i\ )$ is the new triplet estimate of n and $V(n_{cap}\ ) = \sum_{i=1}^{3} w_i^2 V(n_i\ )$

5. Alternative way of summarising the triplet

The approach used to derive n_{cap} may be called *evaluate and summarize* method because we evaluate E(n) at each component of the triplet and then summarized them as a weighted average. The variance of n_{cap} was also obtained with this logic.

Alternatively, we may summarise the triplet and then evaluate as a single value from which we can obtain E(n) and V(n) In this method we use $p_0 = (p_1{+}4p_2{+}p_3)/6$ basing on the PERT weights.

Again with given p_0 we again identify a triangular distribution with p_0 at the peak and obtain

$$E(Y) = \frac{\{1+p_0\}}{3} \text{ and } V(Y) = \frac{\{p_0^2 - p_0 + 1\}}{18} \tag{9}$$

With these values we can identify a beta distribution with parameters say (u_0, v_0) and evaluate

$$E\left(T_0\right) = \frac{u_0 v_0}{\left(u_0 + v_0\right)^2} \text{and} V\left(T_0\right) = \frac{\left(u_0 v_0\right)^2}{\left(u_0 + v_0\right)^4 \left(1 + u_0 + v_0\right)^2} \tag{10}$$

where T_0 denotes the quantity Y(1-Y) under this method. If we call this resulting n as n_0 we get E(n_0) = k $E\left(T_0\right)$ and V(n_0) = k^2 $V\left(T_0\right)$. Here is an illustration.

Illustration-4

Let us consider the triplet $(0.25, 0.35, 0.50)$. We get $p_0 = 0.675, E(Y) = 0.5583, V(Y) = 0.0433, u_0 = 2.6164, v_0 = 2.0697$. Using the k value from normal distribution with (1-α) = 0.95 and 5%margin of error (d), we get E(n_0) = 379 and V(n_0) = 4440.71. With different triplets used in Illustration-3 we get the expected sample size and variance under this method of 'summarize and evaluate' are shown in Table (4).

Table 4: Estimated sample size with a pre-summarized triplet.

Triplet (p_1, p_2, p_3)	n_0	V(n_0)
(0.15, 0.25, 0.35)	373	4810.76
(0.25, 0.35, 0. 50)	381	4321.43
(0.30, 0.50, 0.65)	384	4099.87
(0.45, 0.65, 0.75)	381	4295.74
(0. 50, 0.70, 0.75)	379	4440.71

We observe that sample size exhibits higher variance by this method when compared with the method of evaluating three n values and summarizing them with beta density as weights.

6. Conclusion

The problem of finding the minimum sample size to estimate a proportion is better explained with a statistical model instead of simply evaluating the available formula with a single anticipated value of the population proportion (p'). The triangular distribution plays a key role in transforming the single p into random variable so that its mean and variance can be used to determine the parameters of the beta distribution, which has better shape and properties than the triangular distribution. The uncertainty about p' can be handled by a beta distribution leading to a statistically summarised estimate of n. It also helps in estimating the variance of n while the classical formula gives only single value. With this logic we have proposed a new estimate of n basing on a triplet of input values for p and summarised them as a weighted average. It is shown that the new estimate (n_{cap}) has smaller variance than the variance obtainable at each of the three p values. We have used the weights from the density of beta distribution at the triplet values, so that they reflect the baseline uncertainty in the inputs and normalized them. It is also established that this method is more objective than using other methods of fixed weights, in terms of variance of n. We conclude with the observation that sample size formula greatly depends on the accuracy of the inputs given and the often found attitude among users, *to adjust the inputs until a comfortable number is reached* should be avoided.

Acknowledgements

I am indeed grateful to the Editors for their guidance and counsel. I am very grateful to the reviewer for valuable comments and suggestions of generously listing many useful references.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

(2022). Pert distribution. `https://en.wikipedia.org/wiki/PERT_distribution`.

Clark, C. E. (1962). The pert model for the distribution of an activity time. *Operations Research*, **10**, 405–406.

Cochran, W. G. (1977). *Sampling Techniques.* John Wiley & Sons, Inc., New York.

Cohen, J. (2016). A power primer. *American Psychological Association*, **112**, 155–159.

Daniel, W. W. and Cross, C. L. (2018). *Biostatistics.* John Wiley & Sons, New York.

Indrayan, A. and Malhotra, R. K. (2017). *Medical Miostatistics.* CRC Press.

Malcolm, D. G., Roseboom, J. H., Clark, C. E., and Fazar, W. (1959). Application of a technique for research and development program evaluation. *Operations Research*, **7**, 646–669.

Sarada, V. S., Subbarayudu, M., and Sarma, K. (2018). Estimation of process capability index using confidence intervals of process parameters. *Research & Reviews: Journal of Statistics*, **7**, 49–57.

Taha, H. A. (2013). *Operations Research: An Introduction.* Pearson Education India.

Vardhan, R. V. and Sarma, K. (2010). Estimation of the area under the roc curve using confidence intervals of mean. *ANU Journal of Physical Sciences*, **2**, 29–39.

Vedururu, S. S., Subbarayudu, M., and Sarma, K. (2019). A new method of estimating the process capability index with exponential distribution using interval estimate of the parameter. *Stochastics and Quality Control*, **34**, 95–102.

Venkatesu, B., Abbaiah, R., and Sarada, V. S. (2018). A new method of estimating the process spread using confidence interval of sample range. *Research Review International Journal of Multidisciplinary*, **8**.

Venkatesu, B., Abbaiah, R., and Sarada, V. S. (2019). On the estimating the operating characteristic of shewart control chart for means using interval estimates of process mean and spread.

Yamane, T. (1973). *Statistics: An Introductory Analysis.* Harper & Row New York.