Determination of the experiment size for the optimum reasoning on success probability in binominal distribution for experiments in agricultural engineering

Andrzej Kornacki

Department of Applied Mathematics and Computer-Science, University of Life Sciences in Lublin
Akademicka 15, 20-950 Lublin Poland, e-mail:Andrzej.kornacki@up.lublin.pl

Received: March 09, 2018; Accepted: April 12, 2018

Abstract. In the study the method of determination of the experiment size is provided so that the optimum reasoning could be applied on the basis of experiment results. The parameter estimation is examined here at the set absolute or relative error as well as hypothesis testing at pre-set (high) power. The considerations contained in the paper refer to success probability in binominal distribution. Such a distribution frequently occurs in the research in agricultural engineering.

Key words: Test power, sample size, binominal distribution, success probability.

INTRODUCTION

The issue of sample size determination is of crucial importance in these areas where the basis is the experiment (agricultural, technical, medical sciences and others). On one hand the researchers would like to limit experiment costs by decreasing the number of personnel conducting the research, cutting down on expensive equipment etc. and , on the other one, the effective statistical analysis of experiment results requires a large sample size. In practice some compromise must be reached between these two opposing trends.

Determination of the sample size that would guarantee an effective statistical analysis of experiment results goes in three directions:

1. In parameter estimation problems where the sample comes from the known probability distribution, the sample size is determined by controlling the absolute or relative error.
2. When the sample comes from finite population its size can be determined by controlling the estimator variance.
3. In hypothesis testing procedures the sample size is determined by controlling the test power at the determined significance level for alternatives.

In this study we will deal with items 1 and 3 (i.e. parameter estimation and hypothesis testing) as regards success probability p in binominal distribution. This distribution frequently occurs in agricultural engineering research. Quality control assesses the product defectiveness (e.g. defects of bearing, lamps, machinery etc. (Bobrowski [2], Stark and Nicholis [10]. In plant evaluation centres seed germination capacity is assessed (Niedokos [8]). All these values are success probability values in binominal distribution.

OBJECTIVES

The aim of the study was to introduce formulae for sample size that could guarantee the effective success probability p reasoning in binominal distribution. We focussed on the success probability p evaluation with beforehand controlled error and on testing probability p hypothesis at the determined significance level α characterised by beforehand given (adequately high) power.

MATERIALS AND METHODS

I Determination of the required sample size during evaluation of success probability p in binominal distribution.

Let X be a discrete zero-one random variable i.e. a random variable with probability function described by the formula:

$$P(x,p) == p^x(1-p)^{1-x}\ , \quad x = 0,1\,. \tag{1}$$

Let it denote the random sample $X_1, X_2, \dots, X_n$ from distribution X. Then $Y = \sum_{i=1}^{n} X_i$ is known as a random variable with binominal distribution (Fisz [4], Benjamin, Cornell [1]) with parameters p and n with probability function:

$$f(y; n, p) = \binom{n}{y} p^y (1-p)^{n-y} \quad y=0,1,\dots,n. \qquad (2)$$

The expected value Y and its variance are equal respectively np and np(1-p) (Feller [3]).

Thus $\hat{p} = \frac{Y}{n}$ is an unbiased estimator of parameter p. Moreover, we have:

$$E(\frac{Y}{n}) = p \;\; i \quad Var(\frac{Y}{n}) = \frac{p(1-p)}{n}. \qquad (3)$$

In order to determine the sample size necessary to estimate p by $\hat{p} = \frac{Y}{n}$ controlling the absolute error with a great probability, we need to choose n in such a way as to fulfil the condition:

$$P(\left|\frac{Y}{n} - p\right| \le d) \ge 1 - \alpha. \qquad (4)$$

for set constant values d and α.

The left side of the formula (4) after the transformation is equal to:

$$P([n(p-d)] + 1 \le Y \le [n(p-d)] =$$

$$= P(y_1 \le Y \le y_2) = \sum_{y=y_1}^{y_2} \binom{n}{y} p^y (1-p)^{n-y}. \qquad (5)$$

In order to obtain a quick solution of the equation (4) the variable Y/n can be approximated.

From the central limit theorem we get: (Rao [9]):

$$\frac{\frac{Y}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1). \qquad (6)$$

Using this fact we get from (4):

$$\Phi(\frac{d}{\sqrt{\frac{p(1-p)}{n}}}) - \Phi(\frac{-d}{\sqrt{\frac{p(1-p)}{n}}}) \ge 1 - \alpha. \qquad (7)$$

In formula (7) Φ denotes the cumulative distribution function of an normal zero-one distribution. Condition (7) is transformed to:

$$\Phi(\frac{d}{\sqrt{\frac{p(1-p)}{n}}}) \ge 1 - \frac{\alpha}{2}. \qquad (8)$$

Therefore, n must satisfy the condition:

$$n \ge \frac{p(1-p) z_{\frac{\alpha}{2}}^2}{d^2}. \qquad (9)$$

Because p is unknown and p(1-p) reaches the greatest value for p=0.5 we can use (redundancy) estimator on n form: [n*]+1 where:

$$n^* = \frac{z_{\frac{\alpha}{2}}^2}{4d^2}. \qquad (10)$$

In formula (10) symbol z_α denotes the upper critical value of the standard normal variable i.e. such a value that:

$$P(Z > z_\alpha) = \alpha. \qquad (11)$$

and notation [a] denotes the highest integer that does not exceed a. To sum up it was demonstrated that:

The approximate sample size necessary to asses success probability p in binominal distribution by $\hat{p} = \frac{Y}{n}$ fulfilling condition (4) is [n*]+1 where n* is given by formula (10).

The quality control department wants to evaluate the defectiveness of manufactured bearings with the error d=5% and probability of at least 95%. Due to (10) we get:

$$n^* = \frac{(1{,}96)^2}{4(0{,}95)^2} = 384{,}16.$$

Therefore, the required sample size is **n=385**. If we allowed a larger error, for example d = 10%, we would get it:

$$n^* = \frac{(1{,}96)^2}{4(0{,}1)^2} = 96{,}04.$$

Thus, the required sample size is equal **n=97**.

II Determination of the necessary sample size to test success probability p hypothesis in binominal distribution.

Let us consider a standard hypothesis p in binominal distribution in the form:

H_0: p=p_0 versus alternative H_A: p>p_0. The critical area for the test at significance level α has a form: Y> r where r fulfils the condition:

$$P\left[Y > r \middle| p = p_0\right] = \alpha, \tag{12}$$

or equivalently (Johnson and Kotz [5]):

$$I_{p_0}\left[r+1, n-r\right] = \alpha. \tag{13}$$

$I_x(a,b)$ denotes in (13) an incomplete ordinary beta function (Kapel et al.[6]) in the form:

$$I_x(a,b) = \frac{\int_0^x t^{a-1}(1-t)^{b-1}\,dt}{\int_0^1 t^{a-1}(1-t)^{b-1}\,dt}. \tag{14}$$

Demanding that the test power be equal 1-β for p=p_1 (>p_0) we get:

$$P\left[Y > r \middle| p = p_1\right] = 1-\beta. \tag{15}$$

And from this size n can be determined, Therefore, n and r must satisfy the 2 following equations:

$$\begin{cases} I_{p_0}(r+1, n-r) = \alpha \\ I_{p_1}(r+1, n-r) = \beta \end{cases}. \tag{16}$$

The approximate solution due to n size can be found by doing a transformation arcsin on Y/n and by expressing probability values (12) and (15) in cumulative distribution function terms of standard normal distribution. It is known (Rao 2003, Krysicki et al.1998) that:

$$Z = 2\sqrt{n}(\arcsin\sqrt{\hat{p}} - \arcsin\sqrt{p}) \sim N(0,1). \tag{17}$$

Let us note that fuction arcsin() must be given in radians. Thus n fulfils the conditions:

$$\begin{cases} 1-\Phi\left[2\sqrt{n}(\arcsin\sqrt{\frac{r}{n}} - \arcsin\sqrt{p_0})\right] = \alpha \\ \left\{1-\Phi\left[2\sqrt{n}(\arcsin\sqrt{\frac{r}{n}} - \arcsin\sqrt{p_1})\right] = 1-\beta\right. \end{cases}. \tag{18}$$

While solving the system of equations (18) due to n we get the following formula for the required sample size:

$$n* = \left\{\frac{z_\alpha + z_\beta}{2\left(\arcsin\sqrt{p_1} - \arcsin\sqrt{p_0}\right)}\right\}^2. \tag{19}$$

Therefore, the following result has been proved:

The approximate sample size that is required to obtain a one-sided test of significance α of the set power 1-β for hypothesis H_0:p=p_0 versus alternative H_1:p=p_1(>p_0) is [n^*]+1 where n^* is determined by the formula 19).

It is assumed that the grain germination capacity of a new plant variety is higher than 90%. How many grains should be planted in the material assessment laboratory in order to check this assumption with a power test (1-β)=0,9 at significance level α=0,05. Putting p_1=0,95 we get z (19):

$$n^* = \left(\frac{1{,}645 + 1{,}282}{2{,}6906 - 2{,}4981}\right)^2 = 231{,}2.$$

Therefore, the required number of grains is **n=232**.

REFERENCES

1. **Benjamin J.R, Cornell C.A. 1977.** Rachunek prawdopodobieństwa statystyka matematyczna i teoria decyzji dla inżynierów. WNT Warszawa.(In Polish).
2. **Bobrowski D. 1986.** Probabilistyka w zastosowaniach technicznych. WNT Warszawa.(in Polish).
3. **Feller W. 1981.** Introduction to Probability Theory and Its Application. vol II.
4. **Fisz M. 1967.** Rachunek prawdopodobieństwa i statystyka matematyczna. PWN.(in Polish).

5. **Johnson N.L, Kotz S. 1969.** Discrete distributions. John Wiley & Sons.
6. **Kapel J.K, Kapadia C.H, Owen D.B. 1976**. Handbook of Statistical Distributions. Marcel Dekker.
7. **Krysicki W., Bartos J., Dyczka W., Krolikowska K., Wasilewski M. 1998**. Rachunek prawdopodobieństwa i statystyka matematyczna w zadaniach cz II PWN.(in Polish).
8. **Niedokos E. 1995.** Zastosowanie rachunku prawdopodobieństwa i statystyki matematycznej. Wyd AR Lublin.(in Polish).
9. **Rao C.R. 2003.** Linear Statistical Inference and its Applications John Wiley and Sons.
10. **Stark R.M., Nicholis R.L 1979.** Matematyczne podstawy projektowania inżynierskiego. PWN Warszawa.(in Polish).