Communications on Stochastic Analysis
Vol. 14, No. 3-4 (2020)

NOTE ON THE EFFECTIVE SAMPLE SIZE WHEN POPULATION DEFECT RATE IS ONE PART PER MILLION

SEOK HO CHANG[1*]

ABSTRACT. In this paper, we examine the effective sample size to obtain a reliable conclusion when the population defect rate is one part per million (PPM). Although such problems are of great importance in the context of quality management, classical textbooks of statistics only discuss the situation in which the defective rate is fairly high compared to one PPM or suggest approximate methods. In this paper, we first summarize previous formulae for sample size and present their limitations in handling situations in which the defective rate is one PPM. We then present two formulae for sample size that are effective in handling situations in which the population defect rate is one PPM. Based on our formulae, we present the sample size numerically by varying the desired level of accuracy and confidence coefficient. Based on our results, we also offer practical suggestions for determining sample sizes in low-defect cases.

KEYWORDS sample size, one part per million defective, fraction defective, confidence interval, estimation, precision and accuracy.

1. Introduction

Consumer demands for products and services are changing rapidly in the dynamic global economic environment of the early 21st century. In a situation marked by such high levels of change and competition, there is an urgent need to improve the quality of both products and services.

For products that damage people's lives in the event of failure, the sales market demands a fairly high level (e.g., one part per million [PPM] defective) of quality for each component. This can be illustrated by the following simple example: suppose there is a product that consists of an assembly of 1,200 components, and all 1,200 components must be nondefective for the product function satisfactorily. If each component is under three-sigma quality, the probability that any specific unit of the product is nondefective is $0.9973^{1200} = 0.03899265869$, which means that about 3.9% of products under three-sigma quality will be nondefective. This situation is not desirable considering that

Received 2019-8-26; Accepted 2019-10-14;
[1] College of Business, Hankuk University of Foreign Studies, Republic of Korea; *seokhochangaaa@gmail.com*
[*] Correspondence: *seokhochangaaa@gmail.com*
Acknowledgement: This research was partially supported by Hankuk University of Foreign Studies Research Fund.

the products damage people's lives in the event of failure. If we use six-sigma quality for each component, the situation is much better: the probability that any specific unit of the product is non-defective is 0.999999998^{1200}=0.9999976, which means that about 99.99976% of the product will be nondefective.

This article considers the following situation: for a product manufacturer (such as airplanes or ships) requiring several million components, vendors want to deliver parts for manufacture. Suppose that the defective rate of the lot that is a large set of components currently used in the company is one PPM. Suppose that one supplier claims that the defective rate of the lot provided is less than one PPM. Sampling inspection can be performed to determine whether their claims are valid or not. In this case, an important question is how many samples are necessary to draw a reliable conclusion. Increasing the sample size increases the accuracy of the estimate, but can increase sampling costs and data processing time. Predetermining the size of the sample needed to meet a given accuracy is an important issue for decision-makers. For products that damage people's lives in the event of failure, highly accurate estimates should be reflected in decision making.

This paper seeks to resolve the difficulty and determine the effective sample size to verify whether a population defect rate is one PPM. Although such problems are of great importance, classical statistics textbooks discuss cases in which the defective rate is fairly high (such as 0.1, 0.01 or 0.001) compared to one PPM or suggest approximate methods ([3,5,6], and the references therein). We first summarize previous formulae to determine sample size and present their limitations in handling cases in which the defective rate is one PPM. We then present two formulae to determine sample size that are effective in handling such situations. We present numerical sample sizes based on our formulae by varying the desired level of accuracy and confidence coefficient. Based on our results, we also offer practical suggestions for determining sample sizes in low-defect cases.

The rest of this article is organized as follow. In section 2, we summarize previous formulae for the sample size, and present their limitations in handling the case where the defective rate is one PPM. In section 3, we present two formulae for the sample size, which are effective in handling the case when the population defect rate is one PPM. Based on our formulae, we present the sample size numerically by varying the desired value of accuracy and confidence coefficient. In section 4, we conclude this paper, and suggest practical suggestion for determining sample sizes in low-defect cases.

2. The limitations of previous approaches

In this section, we summarize previous formulae for sample size, and present their limitations in handling situations in which the defective rate is one PPM.

Sample size formula based on normal approximation interval

A commonly used formula for a confidence interval for unknown fraction defective of a population relies on approximating the binomial distribution with a normal distribution, which is based on the central limit theorem [1,3,5,6]. The normal approximation to binomial is known to be satisfactory for the unknown defective rate (or fraction defective) P of approximation is 1/2 and sample size n >10 [1,3,5,6]. For other values of P, larger values of sample size n is required [1,3,5,6].

Using the normal approximation, 100×(1-α)% confidence interval for the unknown defective rate P of a population is

$$[\hat{P} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}, \quad \hat{P} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}], \tag{2-1}$$

where $\hat{P}$ denotes the sample fraction defective (or sample defective rate), n denotes the sample size, and $z_{\frac{\alpha}{2}}$ denotes the (1- $\frac{\alpha}{2}$) quantile of standard normal distribution corresponding to target error rate α. For a 95% confidence interval, the error α= 1-0.95=0.05, so (1- $\frac{\alpha}{2}$) = 0.975, and $z_{\frac{\alpha}{2}} = 1.96$.

Based on the confidence interval, we can calculate the minimum sample size for the unknown defective rate (fraction defective) of a population using the following formula:

$$n = \left(\frac{z_{\frac{\alpha}{2}}\sqrt{\hat{P}(1-\hat{P})}}{B}\right)^2, \tag{2-2}$$

where B denotes the desired level of precision [3].

A common mistake in using sample size formula based on normal approximation

A common mistake that can be made in using the formula (2-1) and (2-2) is that one can choose the sample size n that satisfies $n > 30$ in handling situations in which the population defective rate is one PPM. This mistake is due to the argument that the normal approximation to binomial is based on central limit theorem [1,3,5,6] which usually hold for $n > 30$.

Suppose we choose the sample size $n = 40$ since it satisfies the inequality $n > 30$. Assume that the defective rate of a company's lot is known to be one PPM. A random sampling (sampling with replacement) is used to estimate the unknown defective rate. For convenience, we assume an infinite population. If the defective rate of the lot is one PPM, the probability that all the parts contained in the sample of size 40 are not defective is equal to $(1-0.000001)^{40}$ = 0.99996000078, which is close to 1. Then, sample fraction defective is equal to 1-0.99996000078=0.00003999922, which is close to 0. If we roughly plug $\hat{P} = 0$ into the formulae (2-1) and (2-2), we get the following results:

100×(1-α)% confidence interval for the unknown defective rate P of a population is $[0\ ,\ 0]$, (2-3)

the minimum sample size for the unknown defective rate (fraction defective) of a population is equal to 0. (2-4)

(2-3) and (2-4) suggest that the formulae (2-1) and (2-2) are not effective in handling situations in which the population defective rate is one PPM. For readers reference, we summarize the probability that all the parts contained in the sample are not defective for various values of the sample size in Appendix 1. As you can see Appendix 1, the probability that all the parts contained in the sample (of size 10,000) are not defective is 0.99004982879, which is close to 1. If the sample size is 1,000,000, the probability that all the parts contained in the sample are not defective becomes 0.36787925722.

Wilson's approximate formula

Wilson [3,4] presents the approximate confidence interval for the event where the defects are relatively rare:

100×(1-α)% confidence interval for the unknown defective rate P of a population is

$$[\tilde{P} - z_{\frac{\alpha}{2}}\sqrt{\frac{\tilde{P}(1-\tilde{P})}{n+4}}\ ,\quad \tilde{P} + z_{\frac{\alpha}{2}}\sqrt{\frac{\tilde{P}(1-\tilde{P})}{n+4}}\ ], \tag{2-5}$$

where $\tilde{P}$ denotes the estimate of the fraction defective, and it satisfies

$$\tilde{P} = \frac{x+2}{n+4}, \tag{2-6}$$

where n denotes the sample size, and x denotes the number of defects in the sample size.

A limitation in Wilson's formulae

Suppose we choose the sample size $n = 40$ since it satisfies the inequality $n > 30$. If the defective rate of the lot is one PPM, the probability that all the parts contained in the sample of size 40 are not defective is equal to $(1-0.000001)^{40}$ = 0.99996000078, which is close to 1. Then, sample fraction defective is equal to 1-0.99996000078=0.00003999922, which is close to 0. If we plug $x = 0$ and $n = 40$, into the formula (2-6), we get the following result:

the estimate of the fraction defective $\tilde{P}$ satisfies

$$\tilde{P} = \frac{0+2}{40+4} = \frac{1}{22} = 0.04545454545. \tag{2-7}$$

The estimate of the fraction defective $\tilde{P}$ is 0.04545, which is much higher than true fraction defective (one PPM). Improved methods are needed to make more accurate, precise and reasonable estimate.

Duncan's approximate formula

Duncan [2] presents the approximate confidence interval and sample size formula using the Poisson approximation. Readers are referred to [2,3,5,6] and the references therein for other approximate methods.

3. Main results

In this section, we present main results. Readers are referred to Appendix 2 for the proofs of main results.

Assumptions

Assume that the defective rate of a company's lot is known to be one PPM. Suppose that one supplier claims that the defective rate of the lot provided is less than one PPM. To verify his/her claim, we consider the following hypotheses:

H_0: the defective rate of the lot = one PPM.

H_1: the defective rate of the lot < one PPM.

Assume that the random sampling (i.e., the sampling with replacement) is used to verify the claim of the supplier. For convenience, we assume an infinite population. For this type of hypothesis testing, we present the effective sample size to draw a reliable conclusion in the following Lemma 1 and Theorem 1.

Lemma 1. The effective sample size for estimating the unknown defective rate (fraction defective) of a population is greater than $1/P$, where P denotes unknown defective rate of a population.

Remark 1. See Appendix 2 for the proof of Lemma 1. As you can see Lemma 1, we need an estimate of P. We can replace P by the value (one PPM) specified in the null hypothesis H_0 in the hypothesis testing procedure. We then have the following result:

The effective sample size for estimating the unknown defective rate (fraction defective) of a population is greater than 1,000,000 (one million). (3-1)

Remark 2. When the population defective rate is one PPM, Lemma 1 is more reliable than the formula based on normal approximation to binomial and Wilson's formula. However, there is still a room for improvement in Lemma 1. Although Lemma 1 is simple, easy to use, and produces more reliable results in handling the situations in which the population defect rate is one PPM than previous approaches, it does not use the information of desired level of accuracy and confidence coefficient. To get a more precise and reliable conclusion, we need the following Theorem 1.

We define the following notation:

Let a denotes desired level of accuracy in the estimation of an unknown defective rate.

Remark 3. Note that a is a term related to the length of the confidence interval of the unknown population defective rate. Under the assumption that the confidence coefficient is fixed, the shorter the length of the confidence interval, the higher the accuracy of the estimate [1].

Theorem 1.

If $0 < a \leq 2\,\hat{P}$,

$$\text{The effective sample size } n \geq \frac{4z_{\frac{\alpha}{2}}^2\hat{P}(1-\hat{P})}{a^2} \tag{3-2}$$

Otherwise (i.e., $a > 2\,\hat{P}$),

$$\text{The effective sample size } n \geq \frac{z_{\frac{\alpha}{2}}^2\hat{P}(1-\hat{P})}{a^2} \tag{3-3}$$

where the definitions of $\hat{P}$ and $z_{\frac{\alpha}{2}}$ are in (2-1).

Remark 4. For the proof of Theorem 1, see Appendix 2. In Theorem 1, $\hat{P} =$ one PPM. Theorem 1 is more reliable than Lemma 1 in the sense that it contains the information of the desired level of accuracy (a) and confidence coefficient in the estimation. As you can see Theorem 1, we need to specify the desired level of accuracy (a) and the target error rate α to use Theorem 1.

Remark 5. When the population defect rate is one PPM, it is desirable that users are required to specify the desired level of accuracy (a) in a fairly low level (e.g., one PPM or two PPM) to get an accurate and reasonable sample size.

Assume that the desired level of accuracy is specified by two PPM. We also assume 95 % confidence interval (i.e., $z_{\frac{\alpha}{2}} = 1.96$). For this case, the effective sample size can be calculated using (3-2):

$$\text{The effective sample size } n \geq 3{,}841{,}596.1584 \tag{3-4}$$

That is, minimum sample size is 3,841,596 to obtain a reliable conclusion.

We present numerical sample sizes based on our formulae by varying the desired level of accuracy and confidence coefficient in the following Table 1:

Table 1: The effective sample size for several different values of desired value of accuracy (a) and confidence level

a	Confidence level = 90%	Confidence level =95%	Confidence level = 99%
1 ppm	10,824,089	15,366,385	26,543,078
⋮	⋮	⋮	⋮
2 ppm	2,706,022	3,841,596	6,635,769
⋮	⋮	⋮	⋮
3 ppm	300,669	426,844	737,308
⋮	⋮	⋮	⋮
10 ppm	108,241	153,664	265,431

4. Conclusion and suggestion

In this paper, we examine the effective sample size to obtain a reliable conclusion when the population defect rate is one part per million (PPM). We first summarize previous formulae for sample size and present their limitations in handling situations in which the defective rate is one PPM. We then present two formulae for sample size that are effective in handling situations in which the population defect rate is one PPM. Based on our formulae, we present the sample size numerically by varying the desired level of accuracy and confidence coefficient.

As you can see (3-1) and (3-4), the effective sample size is more than 1,000,000 (one million) or more than 3,841,596 to obtain a reliable conclusion in situations in which the population defective rate is one PPM. Our numerical results suggest that the number of sample at the full inspection level is required to obtain a reliable conclusion. For products that damage people's lives in the event of failure, our results based on highly accurate estimates could be used as a reference in decision making. In practice, increasing the accuracy obtained by reducing the length of the confidence interval often does not compensate for the increase in the cost of sampling and non-sampling errors. For this kind of problems, it is suggested that we focus on the process capability index which is a statistical tool to measure the ability of a process to produce output within specification limits, and control the defect rate of a population.

References

1. Chang, S.H. 2018. *Summary Note for Management Statistics & Data Interpretations*, G-World Publisher Corp. (in Korean), Republic of Korea.
2. Duncan, A.J. 1980, *Quality Control and Industrial Statistics*, 5th edi. Irwin Home, IL.
3. Keller, G., 2014, *Statistics for Management and Economics*, 10th edi, Cengage Learning, 2014.
4. Wilson, E.B., 1927, Probable inference, the law of succession, and statistical inference, *Journal of the American Statistical Association* 22 (158), 209-212.
5. Robert S.W. and John S. W. *Statistics*, Willey, 2013.
6. Robert S. and Dean F., *Statistics for Business: Decision Making and Analysis*, Prentice new international edition, 2014.

Appendix 1. The probability that all the parts contained in the sample are not defective for several different sample sizes

Assume that the unknown defective rate of a company's lot is known to be one PPM. Suppose that the random sampling (i.e., the sampling with replacement) is used to estimate the unknown defective rate P of a population. For convenience, we assume an infinite population.

Let W denotes the probability that all the parts contained in the sample are not defective.

We provide W for several different sample sizes in the following tables:

Table A. The effect of the sample size on the probability that all the parts contained in the sample are not defective

Sample size	W
40	0.99996000078
50	0.99995000122
⋮	⋮
100	0.99990000495
200	0.99980001989
⋮	⋮
1000	0.99900049933
10,000	0.99004982879
⋮	⋮
100,000	0.90483737279
200,000	0.8187306712
⋮	⋮
1,000,000	0.36787925722
2,000,000	0.13533514789

Remark A-1. We use double precision in presenting the results in Table A.

Appendix 2. Proofs of Lemma 1 and Theorem 1

Proof of Lemma 1.

Suppose that the null hypothesis (the defective rate of the lot = one PPM) is true. If we recall the definition of sampling with replacement, the number of independent inspections to get the result of first failure follows geometric distribution with unknown parameter (probability of failure) given by one PPM. The mean of the geometric distribution is simply the reciprocal of the unknown parameter (probability of failure), and it is given by one million. Therefore, it is reasonable to assume that the effective sample size for estimating the unknown defective rate of a population is greater than the mean of the geometric distribution.

Proof of Theorem 1.

Using (2-1), it can be seen that the length of 100×(1- α)% confidence interval (CI) for the unknown defective rate P of a population is

If $0 < a \leq 2\,\hat{P}$,

$$\textit{The length of CI} = 2z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

Otherwise (i.e., $a > 2\,\hat{P}$),

$$\textit{The length of CI} = z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

Remark A-1. In the above inequality, $\hat{P} = 0.000001$ (one PPM). If $a > 2\,\hat{P}$, the lower bound of the confidence interval becomes negative. We replace the negative lower bound by 0. Then it can be seen that the length of the confidence interval become $z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$.

We want to get the sample size so that it satisfies the following inequality:

If $0 < a \leq 2\,\hat{P}$,

$$2z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq a$$

Otherwise (i.e., $a > 2\,\hat{P}$),

$$z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq a$$

Squaring both sides of the equation, we get the following inequality:

If $0 < a \leq 2\,\hat{P}$,

$$4z_{\frac{\alpha}{2}}^{2}\frac{\hat{P}(1-\hat{P})}{n} \leq a^2$$

Otherwise (i.e., $a > 2\,\hat{P}$),

$$z_{\frac{\alpha}{2}}^{2}\frac{\hat{P}\left(1-\hat{P}\right)}{n}\leq a^{2}$$

from which we get the following inequality:

If $0 < a \leq 2\,\hat{P}$,

The effective sample size $n \geq \frac{4z_{\frac{\alpha}{2}}^{2}\hat{P}(1-\hat{P})}{a^{2}}$

Otherwise (i.e., $a > 2\,\hat{P}$),

The effective sample size $n \geq \frac{z_{\frac{\alpha}{2}}^{2}\hat{P}(1-\hat{P})}{a^{2}}$

This completes the proof.