ARC Journal of Clinical Case Reports
Volume 5, Issue 1, 2019, PP 13-19
ISSN No. (Online) 2455-9806
DOI: http://dx.doi.org/10.20431/2455-9806.0501004
www.arcjournals.org

On Determining the Volume of Repeated and Non-Repeated Sampling in the Preparation of Clinical Studies

Olga S. Kozhyna

Kharkiv National Medical University, Kharkov, Ukraine

Oleg M. Pihnastyi*

National Technical University "Kharkiv Polytechnic Institute", Kharkov, Ukraine

***Corresponding Author:** ***Oleg M. Pihnastyi**, National Technical University "Kharkiv Polytechnic Institute", Kharkov, Ukraine, **Email:** pihnastyi@gmail.com*

Abstract: *In this article non-repeatable and repeatable sampling methods were analyzed. Main objectives connected with the use of sampling research method were determined. The problem of non-repeatable sampling amount defining at given rate of bronchial asthma prevalence was analyzed. Key preconditions and assumptions used when constructing the expression to determine the non-repeatable and repeatable sample amount were depicted.*

Keywords: *non-repeatable sample amount, repeatable sample amount, bronchial asthma, ISAAC, research methods, statistical population feature*

1. INTRODUCTION

Bronchial asthma is one of the most common diseases of childhood [1, 2]. According to estimates of the World Health Organization, up to 180K of patients die from bronchial asthma in the world every year [7]. Numerous studies have significantly increased the understanding of bronchial asthma pathogenesis [3], use of new medications allowed to regulate the disease symptoms more effectively as well as to maintain the patients' physical activity, but it is still impossible to solve the recovery problem fundamentally [4, 5, 6].

Over the last years incidence rates of bronchial asthma, allergic rhinitis and atopic eczema have significantly increased [8] and thus have given occasion to start up and implement the international study in accordance with standardized methods in order to improve diagnostics early in disease. Upon the recommendation of the World Health Organization the international ISAAC (International Study of Asthma and Allergy in Childhood) program, divided into IV phases [9], has been realized since 1991. The ISAAC program includes research centers in 105 countries of the world and involves two millions of children [10, 11].

In Ukraine the ISAAC research center was founded at the premises of Kharkiv National Medical University in 1997. In 1997-2002 phases I, II and III [12] have been done. Phase IV of the study is currently underway, it involves development and expansion of the ISAAC scope of application, the use of various resources to determine etiological and pathogenetic mechanisms of bronchial asthma in order to reduce the disease severity and prevalence [13, 14]. Implementation of ISAAC's phase IV is relevant for Eastern Europe:

- estimate the current prevalence and severity of bronchial asthma respiratory symptoms in children;
- explore the dynamics of bronchial asthma symptomatic manifestations in children since 1998 [15, 16].

Characteristic of the Y feature of statistical population on the base of the sampling data in the amount of n [17, 18] is the final objective of the sampling observation. Let's study the statistical population relative to quantitative feature Y in the amount of N. There is a variety of sample extraction ways. It is possible to eyeball estimate the sample amount for each way taking into account the desired accuracy

level. Besides, when choosing the selection way, it is necessary to take into account relative value and time, required to implement any given sample method.

2. Characteristics of Non-Repeatable and Repeatable Sample Methods

If values of Y feature are $\{y_1, y_2 \ldots\ldots, y_{N-1}, y_N\}$ correspondingly, then population mean $\bar{y}$ is arithmetic mean of statistical population feature values [1, p.198]

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad (1)$$

Let's assume that, in order to study the statistical population relative to quantitative feature Y, sample in the amount of n was extracted. Suppose the values of quantitative feature X_w of sampling population are $\{x_{w,1}, x_{w,2} \ldots\ldots, x_{w,(n-1)}, x_{w,n}\}$ corresponding ly [1]

$$x_w = \frac{1}{n}\sum_{i=1}^{n} x_{w,i}, \qquad w = 1..W \qquad (2)$$

The sample mean x_w, obtained from one sample data, should be considered as a random value X_w. Thus, we may discuss the sample mean distribution as well as numeric characteristics of this distribution. Indeed, the sample can be made in a number of different ways where the quantitative feature X_w of sampling population will get values $\{x_1, x_2 \ldots\ldots, x_{(W-1)}, x_W\}$, here W is the possible number of sample variations realization.

Let there be given the range of statistical population of Y (the amount $N = 3$) with the values of the feature $\{y_1 = 0, y_2 = 1, y_3 = 0\}$. If the sample is non-repeatable, the possible number of sample variations realization is

$$C_N^n = \frac{N!}{(N-n)!n!} \qquad (3)$$

If the sample amount $n = 2$ from the statistical population in the amount of $N = 3$ we obtain

$$C_3^2 = 3!/(1!2!) = 3,$$

$$\{x_{1,1} = y_1 = 0, x_{1,2} = y_2 = 1\}, \; x_1 = 0.5,$$

$$\{x_{2,1} = y_1 = 0, x_{2,2} = y_3 = 0\}, \; x_2 = 0.0,$$

$$\{x_{3,1} = y_2 = 1, x_{3,2} = y_3 = 0\}, \; x_3 = 0.5.$$

Thus, the random variate X_w for the 1-st, 2-nd and 3-rd sample takes on values in accordance with the expression (2), correspondingly equal to $x_1 = 0.5$, $x_2 = 0.0$, $x_3 = 0.5$.

If the sample method is repeatable (when y_i is extracted, it's value is fixed and got back with the possibility of another extraction) the possible number of combinations

$$n^2 = 9$$

Thus, the selection method and the sample amount may substantially affect the study results. For this reason the selection method and the sample amount should be well-grounded. It is being understood that the sample is representative, i.e. it completely and properly represents the properties of statistical population.

Value of quantitative feature Y of statistical population is connected with the value of quantitative feature X_w of sample population by means of relation [1]:

$$P(|X_w - \bar{y}| < \Delta) = 2\Phi\left(\frac{\Delta}{\sigma(X_w)}\right) = 2\Phi(t), \; 2\Phi(t) = \gamma \qquad (4)$$

$$\Delta = t\sigma(X_w), \qquad (5)$$

$$\Phi(t) = \frac{1}{\sqrt{2\pi}}\int_0^t e^{-\frac{t^2}{2}}\,dt, \qquad (6)$$

here $\Phi(t)$ is Laplace's function whose value may be taken from table [2]. Expression $|X_w - \bar{y}| < \Delta$ means that deviation of the random value X_w, obtained via expression (2) from arithmetical mean of the statistical population feature, does not exceed the error which is equal to Δ. This inequality can be put down as follows

$$x_w - \Delta < \bar{y} < x_w + \Delta \qquad (7)$$

here inequality (7) has the next meaning: if the sample mean x_w (2) is known, it is fair to say that, with a probability of γ, true value (1) of the statistical population quantitative feature Y falls within the limits from $(x_w - \Delta)$ to $(x_w + \Delta)$. Variate $\sigma(X_w)$ is a mean square deviation of the sample population quantitative feature X_w.

Variate $\sigma(X_w)$ normally depends on the sample amount n

$$\sigma(X_w) \approx f_\sigma(n) \tag{8}$$

3. Main Objectives of Sample Method Use

Equation system (4) can be represented as two equations:

$$\frac{\Delta}{\sigma(X_w)} = t, \tag{9}$$

$$2\Phi(t) = \gamma \tag{10}$$

with four variables: Δ, t, $\sigma(X_w)$, γ. It is necessary to define two additional equations to resolve it. In this regard, there are three basic objectives appearing during the use of sampling method [2]:

Objective №1. Define the sample amount n, needed to obtain the results with required accuracy Δ at given probability γ. It is assumed that sample mean x_w (2) is defined as a result of the sample implemented, while the sample population with the values $\{x_{w,1}, x_{w,2} \ldots\ldots, x_{w,(n-1)}, x_{w,n}\}$ is to be obtained. Probability γ, needed the quantitative feature Y to comply the inequality (7) with the given error Δ, is known. Therewith the fact that if the given probability is $\gamma = 0.95$, then in 50 cases of 1000 the inequality (7) will not be met should be taken into account. Hence, in the first objective, among four variables Δ, t, $\sigma(X_w)$, γ, the given ones are Δ, γ (and correspondingly t) and $\sigma(X_w)$ is to be defined and allowing us to obtain the sample amount n.

Objective №2. Define the possible non-sampling error limit, ensuring the results with the given probability, and compare it with the acceptable error. It is assumed that the sample amount n and the probability γ, meeting the relation (7) requirements, are given. Error Δ should be defined and compared with acceptable error to perform experiments. When setting up the problem the probability γ (or t), sample amount n ($\sigma(X_w)$ accordingly) are considered as given and the error value Δ should be defined.

Objective №3. Define the probability that the sample error will not exceed acceptable error. In this case the sample amount n and error Δ are given. The probability that the sample error will not exceed acceptable error is to be defined. When setting up the problem the probability γ (or compliant with it t parameter), sample amount n ($\sigma(X_w)$ accordingly) are considered to be known. It is necessary to define the probability γ that the sample error will not exceed acceptable error Δ.

4. Theoretical Foundations of Non-Repeatable and Repeatable Sample Amount Calculations in the Preparation of Clinical Studies

In the case of non-repeatable method of amount n elements selection, taken from the general population N for checkup, the general number of possible samples can be defined by combinatorial formula (3). Let us assume that the method of amount n elements extraction is such that each sample from the general number (3) has equal probability of being selected. This is a random sample method. As it was already mentioned the selected element is not to get back into the statistical population.

Sampling mean $\bar{x}_w$

$$\bar{x}_w = \frac{\sum_{w=1}^{C_N^n} x_w}{C_N^n} \tag{11}$$

of a random value X_w is an unbiased estimator of mean value $\bar{y}$ (1) for population Y:

$$M[X_w] = \bar{x}_w = \frac{\sum_{w=1}^{C_N^n} x_w}{C_N^n} = \frac{\sum_{w=1}^{C_N^n}\sum_{i=1}^{n} x_{w,i}}{nC_N^n} = \frac{\sum_{w=1}^{C_N^n}\sum_{i=1}^{n} y_i}{n\,N!/((N-n)!n!)} = \frac{\sum_{i=1}^{N} y_i}{N} \tag{12}$$

Since

$$M[Y] = \frac{1}{N}\sum_{i=1}^{N} y_i, \tag{13}$$

it follows that

$$M[Y] = M[X_w] \tag{14}$$

Variance of random variable X_w can be defined via expression

$$M\left[(X_w - \bar{y})^2\right] = \frac{\sum_{w=1}^{C_N^n}(x_w - \bar{y})^2}{C_N^n} \tag{15}$$

Substituting x_w into (15), let us put down

$$\sigma^2(X_w) = M\left[(X_w - \bar{y})^2\right] = \frac{N-n}{nN(N-1)}\sum_{i=1}^{N}(y_i - \bar{Y})^2 =$$

$$\frac{N-n}{n(N-1)}D(Y) = \frac{N-n}{n(N-1)}\sigma^2(Y) \qquad (16)$$

here

$$D(Y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2, \qquad (17)$$

Formula (16) can be used to calculate mean square deviation of random variable X_w

$$\sigma^2(X_w) = \frac{N-n}{n(N-1)}\sigma^2(Y) \qquad (18)$$

Taking into account formulae (4–6) let us obtain relationship of error Δ, probability γ and sample amount n

$$\Delta = t\sigma(X_w) = t\sigma(Y)\sqrt{\frac{N-n}{n(N-1)}} \qquad (19)$$

Let us put the last equality to the next form

$$\Delta^2 = t^2\sigma^2(Y)\frac{N-n}{n(N-1)}, \qquad (20)$$

$$\Delta^2 n(N-1) = t^2\sigma^2(Y)N - t^2\sigma^2(Y)n, \qquad (21)$$

$$n\cdot\left(\Delta^2(N-1) + t^2\sigma^2(Y)\right) = t^2\sigma^2(Y)N, \qquad (22)$$

allowing to obtain sample amount n through error Δ and t parameter, compliant with probability γ

$$n = \frac{t^2\sigma^2(Y)N}{\Delta^2(N-1) + t^2\sigma^2(Y)}N \qquad (23)$$

If a random value Y has binomial distribution with expectation function

$$M[Y] = p \qquad (24)$$

and deviation

$$D(Y) = \sigma^2(Y) = pq, \qquad (25)$$

$$p + q = 1, \qquad (26)$$

then formula (23) looks like:

$$n = \frac{t^2Npq}{\Delta^2(N-1) + t^2pq} \qquad (27)$$

In the most of real-world cases amount of statistical population is much greater than one: $N >> 1$. This helps us to obtain the final equation

$$n = \frac{t^2Npq}{\Delta^2N + t^2pq}, \qquad N >> 1. \qquad (28)$$

When $\Delta^2N >> t^2pq$ this equation is simplified to

$$n = \frac{t^2pq}{\Delta^2}, \quad \Delta^2N >> t^2pq \qquad (29)$$

In medical research it is comfortable to use alternative values of $p^*, q^* = 1000 - p^*, \Delta^*$, expressed in permille and related as follows:

$$p^* = 1000\cdot p, \quad q^* = 1000\cdot q, \quad \Delta^* = 1000\cdot\Delta,$$

$$p^* + q^* = 1000 \qquad (30)$$

Let's multiply the numerator and denominator of the equation (28) by 10^6 to obtain

$$n = \frac{t^2Npq\cdot10^6}{\Delta^2\cdot10^6N + t^2pq\cdot10^6} = \frac{t^2Np^*q^*}{\Delta^{*2}N + t^2p^*q^*}, \quad N >> 1 \qquad (31)$$

From now forth let us suppress the symbol (*) during calculations assuming that the corresponding dimension is given.

5. Definition of Sample Amount n in Non-Repeatable and Repeatable Sample

Let us estimate the amount of sample necessary for phase IV of ISAAC study implementation in Ukraine (Eastern Europe). Statistical population relative to quantitative feature of asthma incidence rate Y in the amount of $N = 89736$ is being studied. The rate of bronchial asthma prevalence in Ukraine in accordance with official data is

$$p = \bar{p}_M{}^{\pm\Delta\ 0/00}, \ \bar{p}_M = 111{,}2, \ \Delta = 2.7\ {}^{0}/_{00} \qquad (32)$$

here $\bar{p}_M$ is the mean prevalence of asthma incidence in Ukraine in permille. Equation (32) is possible to be put down in another form (fig 1):

$$|p_m - \bar{p}_M| < \Delta, \qquad (33)$$

$$p_1 < p < p_2, \ p_1 = \bar{p}_M - \Delta, \ p_2 = \bar{p}_M + \Delta \qquad (34)$$

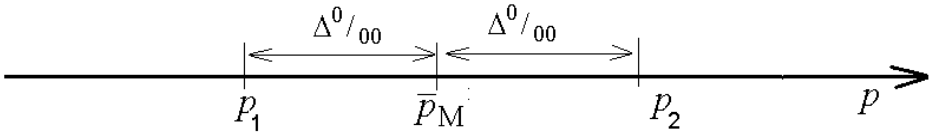

Figure1. *Variation interval for the probability p*

here p is the rate of bronchial asthma prevalence in Ukraine. It should be mentioned that this variate is unknown. However, we can affirm that it can be defined via inequality (34) with reliability $2\Phi(t)=\gamma=0.95$. The point of given reliability is in the fact that if sufficient number of samples is implemented then 95% of the samples will define sample's confidential intervals p_{1W}, p_{2W}, including the rate of bronchial asthma prevalence in Ukraine p

$$p_{1W} < p < p_{2W}$$

Only in 5% of cases the rate of bronchial asthma prevalence p can exceed the sample's confidential intervals p_{1W}, p_{2W}.

In view of the fact that the rate of bronchial asthma prevalence p is unknown, and we know only that it is between p_1 and p_2 (34), we can assume that the required amount of non-repeatable sample n is also in the interval

$$n_{\min} < n < n_{\max}$$

with the limits of $n_{\min}$, $n_{\max}$ defined via relation

$$n_{\min} = \min\{n_1, n_2\}, \qquad n_{\max} = \max\{n_1, n_2\}$$

$$n_1 = \frac{t^2 N p_1 q_1}{\Delta_1{}^2 N + t^2 p_1 q_1}, \qquad n_2 = \frac{t^2 N p_2 q_2}{\Delta_2{}^2 N + t^2 p_2 q_2}.$$

Let us take maximum value of $n = n_{max}$ as the sample amount value. Let us get sample amount for the data in (32):

$$p_1 = 111.2 - 2.7 = 108.5,$$
$$q_1 = 1000 - 108.5 = 891.5, \quad \Delta_1 = 2 \cdot \Delta = 5.4,$$
$$n_1 = \frac{t^2 N p_1 q_1}{\Delta_1{}^2 N + t^2 p_1 q_1} = \frac{1.96^2 \cdot 89736 \cdot 108.5 \cdot 891.5}{(2.7 \cdot 2)^2 \cdot 89736 + 1.96^2 \cdot 108.5 \cdot 891.5} = 1115853 \quad (35)$$

$$p_2 = 111.2 + 2.7 = 113.9,$$
$$q_2 = 1000 - 113.9 = 886.1, \quad \Delta_2 = 2 \cdot \Delta = 5.4,$$
$$n_2 = \frac{t^2 N p_2 q_2}{\Delta_2{}^2 N + t^2 p_2 q_2} = \frac{1.96^2 \cdot 89736 \cdot 113.9 \cdot 886.1}{(2.7 \cdot 2)^2 \cdot 89736 + 1.96^2 \cdot 113.9 \cdot 886.1} \approx 11580 \quad (36)$$

Let's select the maximum value n=1158042≅11581 of two values n_1, n_2. During calculations it was assumed that the sample we used had the rate of bronchial asthma prevalence p_w within the limits (34) of $p_1 < p_w < p_2$. Then, assuming $p_w = p_1$, maximum possible deviation between the sample's mean p_w and the rate of bronchial asthma prevalence p if $p = p_2$ is equal to

$$|p_w - p| = |p_1 - p_2| = 2 \cdot \Delta = \Delta_1$$

This value was used in calculations (fig.2).

The same steps help us to obtain

$$|p_w - p| = |p_2 - p_1| = 2 \cdot \Delta = \Delta_2,$$

if $p_w = p_2$ (fig.3).

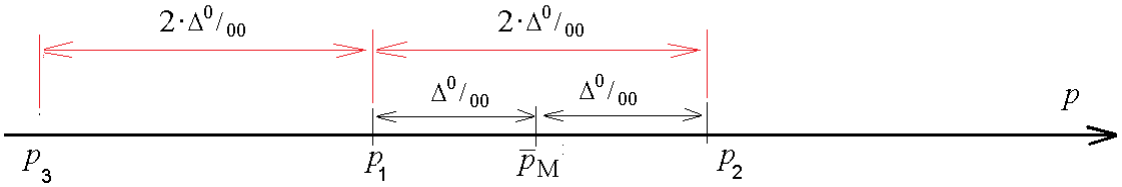

Figure2. *Variation interval for the probability p relative to* p_1

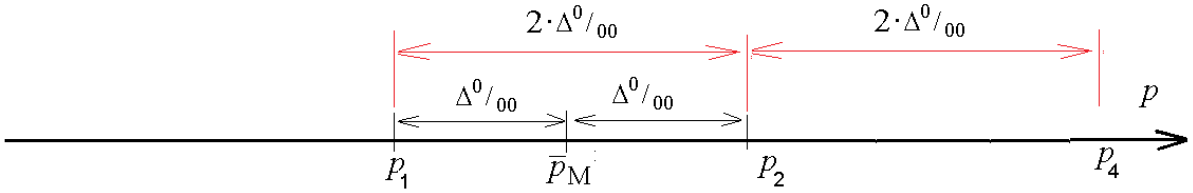

Figure3. *Variation interval for the probability p relative to* p_2

In calculations of the sample amount the error $\Delta_1 = 2 \cdot \Delta$ was used. It twice exceeds the error for the rate of bronchial asthma prevalence in Ukraine. Its structure may look like

$$\Delta_1 = \Delta_w + \Delta, \qquad \Delta_2 = \Delta_w + \Delta$$

here resultant errors Δ_1, Δ_2 contain sample error Δ_w and error Δ, used in obtaining the rate of bronchial asthma prevalence (32). For this purpose we assumed that both errors were equal $\Delta_w = \Delta$.

6. Conclusions

Calculation of sample population amount for representativeness of received data is one of the first and the key moments of the studies oriented to true characteristics of investigated pathology definition and its analysis. In this article repeatable and non-repeatable sample methods were analyzed to reveal reliable information on prevalence of bronchial asthma respiratory symptoms during questioning in accordance with standardized methods of ISAAC. The main objectives connected with the use of sampling research method were investigated. Theoretical foundations were demonstrated and dependence of required sample amount on statistical population value N, required accuracy Δ of the received results with the given probability γ was explained. Objective of non-repeatable sample amount definition with the given rate of bronchial asthma prevalence was analyzed. Key preconditions and assumptions used when constructing the expression to determine the sample amount were depicted.

References

[1] Global Initiative for Asthma (GINA): Global strategy for asthma management and prevention /Update 2014 and Online Appendix. –2014. – available: http:// www. ginasthma.org.

[2] Papadopoulos N.G., Arakawa H., Carlsen K. H. [et al.] / International consensus on (ICON) pediatric asthma // Allergy. – 2012. –Vol. 67, №8.–P.976–997. https://doi.org/10.1111/j.1398-9995.2012.02 865.x

[3] Holgate S. T. Mechanisms of asthma and implications for its prevention and treatment: a personal journey // Allergy Asthma Immunol. Res. – 2013. – Posted online 2013. https:// doi. org/10.4168/aair.2013.5.6.343

[4] GINA Report. / Global Strategy for Asthma Management and Prevention. - 2018. – available: https://ginasthma.org/2018-gina-report-global-strategy-for-asthma-management-and-prevention/ Global atlas of asthma.

[5] Nalina N., Chandra M.R. / Assessment of quality of life in bronchial asthma patients // International Journal of Medicine and Public Health. – 2015. – Vol. 5, N 1. – P.93-97. – https://doi.org/10.4103/2230-8598.151270

[6] Hossny, E. Severe asthma and quality of life / E. Hossny, L. Caraballo, T. Casale, Y. El-Gamal, L. Rosenwasser // World Allergy Organ. -2017; 10(1): 28. - Published online 2017 Aug 21. https://doi.org/ 10.1186/s40413-017-0159-y

[7] World Health Organization -2018. –available: http://www.who.int/ru/news-room/fact-sheets /detail/asthma/

[8] Hansen T. E., Evjenth B., Holt J. Increasing prevalence of asthma, allergic rhino conjunctivitis and eczema among school children: three surveys during the period 1985–2008 // Foundation Acta Paediatrica. – 2013. – Vol. 102. – P. 47–52. –https://doi.org//10.1111 /apa.12030

[9] ISAAC. The International Study of Asthma and Allergies in Childhood /New Zealand: The University of Auckland. – 2012. –available: http://isaac.auckland.ac.nz/.

[10] Lai C. K .W., Beasley R., Crane J., Foliaki S., Shah J., Weiland S., the ISAAC Phase Three Study Group/Global variation in the prevalence and severity of asthma symptoms: Phase Three of the International Study of Asthma and Allergies in Childhood (ISAAC) // Thorax . – 2009 March. -Vol. 64, N 6. https://doi.org//10. 1136/thx.2008.106609

[11] Pearce N., Beasley R., Mallol J., Keil U., Mitchell E., Robertson and the ISAAC Phase Three Study Group / Worldwide trends in the prevalence of asthma symptoms: phase III of the International Study of Asthma and Allergies in Childhood (ISAAC) // Thorax. - 2007 – Vol. 62. – P.758–766. https://doi.org//10.1136/thx. 2006.070169

[12] Ognev V.A. Asthma and Allergy Epidemiology in Children// Kharkiv: Shhedra Usad'ba Pljus. – 2015 - 43.

[13] Tischer C.G., Hohmann C., Thiering E., Herbarth O.,Muller A., Henderson J., et al. Meta-analysis of mould and dampness exposure on asthma and allergy in eight European birth cohorts: an ENRIECO initiative // Allergy. – 2011 –Vol. 66, №12 – P.1570–1579. https:// doi.org//10.1111/j.1398-9995.2011. 02712.x.

[14] Wennergren G., Ekerljung L., Alm B., Eriksson J., Lotvall J., Lundback B. Asthma in late adolescence--farm childhood is protective and the prevalence increase has levelled off // Pediatr Allergy Immunol. –2010 –Vol. 21, –№5 –P. 806–813. https://doi.org//10.1111 / j.1399-3038.2010.0107.7.x

[15] Klymenko V.A., Karpushenko Y.V., Kozhyna O.S. Time course of symptomatic manifestations in children with bronchial asthma residing in Kharkiv region according to the ISAAC study // Inter Collegas.- 2018 – Vol.5, №2 – P.69-72

[16] Kozhyna O.S. Effect of Ecological Factors on Respiratory Diseases Manifestation // Child's Health. –Vol.13, –№5, –p.88–92. https://doi.org//10.22141/2224-0551.13.5.2018.141561

[17] Gnurman V.E. Theory of Probability and Mathematical Statistics– Moscow: Vysshaja shkola, 1972. – 368 p.

[18] Vencel E.S. Theory of Probability and its Engineering Application –Москва: Moscow: Vysshaja shkola, 2000. –480 p.

Citation: *Olga S. Kozhyna, Oleg M. Pihnastyi, On Determining the Volume of Repeated and Non-Repeated Sampling in the Preparation of Clinical Studies. ARC Journal of Clinical Case Reports. 2019; 5(1): 13-19. doi:dx.doi.org/ 10.20431/2455-9806.0501004.*