

MATERIAL DIDACTIC

EŞANTIONAREA ALEATORIE SIMPLĂ: SENSUL MATEMATIC AL FORMULEI DE CALCUL AL VOLUMULUI EŞANTIONULUI REPREZENTATIV

Leonid Margine – dr. în şt. med., conf. univ., Vladislav Badan – asist. univ.,
Dorin Spânu – student, Natalia Cujba – studentă,
IP USMF „Nicolae Testemiţanu”,
Catedra Medicină Socială şi Management Sanitar „Nicolae Testemiţanu”

Rezumat

Introducere. Specialiştii în biostatistică se confruntă tot mai des când cercetătorii ştiinţifici întâlnesc dificultăţi în calcularea corectă a volumului eşantionului reprezentativ ceea ce duce la comiterea unor erori în cercetări. Studii retrospective efectuate asupra studiilor publicate în revistele ştiinţifice demonstrează grave abateri de la normele statistice, în care majoritatea din greşelile comise se referă la eşantionarea incorectă. **Scopul lucrării.** Prezentarea legăturii dintre rezultatele ce urmează a fi obţinute şi formula de calcul a eşantionului, precum şi legăturile ce există între formulele de calcul ale eşantionului reprezentativ cel mai des folosite. **Material şi metode.** Studiu review. A fost cercetată literatura de specialitate şi selectat material în funcţie de relevanţă faţă de cuvintele-cheie. **Rezultate.** Prin metode matematice au fost obţinute formulele de calcul mai des folosite în practica biostatistică. **Concluzii:** Formulele de esantionare în biostatistica reprezintă o largă varietate în dependenţă de necesităţi.

Cuvinte-cheie: eşantionarea aleatorie simplă, eşantion reprezentativ, probabilitatea apariţiei evenimentului, eroarea maximă admisă

Summary. Simple random sampling: the mathematical meaning of the representative sample size calculation formula

Introduction. The experts in biostatistics are faced increasingly with the situation when scientific researchers have difficulty in correctly calculating the research sample size which, subsequently leads to errors in the process of studying and assessing statistical indicators leading to coming errors in their study. Retrospective studies conducted on studies published in scientific journals demonstrates serious violations of statistical rules, in which more than half of mistakes relate to incorrect sampling. **Objective of the study.** To provide a link between the results to be obtained and the sample calculation formula as well as the links that exist between the most commonly used representative sample formulas. **Material and methods.** Review study. Relevant scientific literature has been consulted and the material has been selected depending on its relevance to the key-words. **Results.** Using mathematical methods, the most used formulas to compute the representative sample size have been deduced. **Conclusions:** Sampling formulas in biostatistics are a large variety depending on the needs.

Key words: Simple random sampling, representative sample, the probability of the events, missing probability (counter-probability), the test of significance, MPE

Резюме. Простая случайная выборка: математический смысл расчета размера формулы репрезентативной выборки

Введение. Специалисты в области биостатистике сталкиваются все чаще с трудностями которые ученые испытывают в связи с расчетами которые ведут совершению ошибок в исследованиях. Ретроспективные исследования, проведенные на исследованиях, опубликованных в научных журналах показывают серьезные отклонений от статистических норм, большинство ошибок связано с неправильной выборки. **Точка.** Презентация связь между получаемые результаты и формулы для расчета образца, а также связи, которые существуют между репрезентативной выборки формул наиболее часто используемых. **Материал и методы.** Исследование обзор. Он исследовал литературу и материал, выбранный на основе имеющих отношение к ключевым словам. **Результаты.** Они были получены с помощью математических формул, обычно используемых в практике Biostatistics. **Выводы:** формулы выборки в биостатистике представляют широкий спектр в зависимости от потребностей.

Ключевые слова: простая случайная выборка, репрезентативная выборка, вероятность события, ПДВ

Formulele uzuale ale biostatisticii

În ultimii ani, tot mai multe studii folosesc metoda randomizată de selecție. Studiile efectuate asupra veridicității acestei metode au arătat că deciziile incorecte referitor la ipoteza nulă, în mare măsură, sunt generate în cazul unui volum insuficient al eșantionului [9, 11].

În acest capitol vor fi detaliate formulele cele mai uzuale din biostatistică, referitor la calcularea volumului eșantionului reprezentativ, pentru eșantionarea aleatorie simplă.

Aceste formule pot fi aplicate în cazuri generale, există însă numeroase situații particulare când, în aplicarea acestora, se va ține cont de un șir de indicații speciale.

În manualul său, W. G. Cochran propune următoarea formulă [1]:

$$\bullet \quad n = \frac{t^2 \times p(1-p)}{\Delta^2}$$

Taro Yamane propune o altă formulă, care nu ține cont de probabilitatea apariției evenimentului și de lipsa acestuia, ci numai de volumul totalității statistice (N) și de eroarea maxim admisă [8]:

$$\bullet \quad n = \frac{N}{1 + N \times \Delta^2}$$

Formula recomandată de P. Mureșan reprezintă o sinteză a ambelor formule, ținând cont atât de volumul totalității statistice, cât și de incidența și lipsa fenomenului [4]:

$$\bullet \quad n = \frac{t^2 \times p(1-p) \times N}{t^2 \times p(1-p) + \Delta^2 \times N}$$

(unde: n – volumul eșantionului reprezentativ;
t – testul de semnificație;
p – probabilitatea apariției evenimentului;
N – totalitatea statistică;
Δ – eroarea maxim admisă,
1 – p – contraprobabilitatea.)

Deducerea matematică a formulelor de calcul al eșantionului reprezentativ

În urma analizei statistice a unui eșantion, se obțin indicatori statistici de interes: media, devierea standard, eroarea standard ș.a. În cadrul unui studiu științific, ideea eșantionării corecte prevede ca indicatorii sintetici (devierea standard, eroarea standard), care urmează a fi obținuți, să aparțină unor limite prestabilite, care asigură reprezentativitatea. Criteriile de reprezentativitate sunt: valoarea minimă a testului de semnificație t , și eroarea maxim admisă Δ [5, 6].

Indicatorul de bază obținut este media eșantionului. În studiile cu caracter biologic, media reprezintă rezultatul interacțiunii factorului analizat cu subiecții studiului. Natura interacțiunii dintre factor și subiecți

poate fi diferită. Dacă factorul are un impact real și asemănător asupra tuturor subiecților, media obținută va fi reprezentativă. Dacă factorul nu are efect asupra subiecților, sau dacă efectul observat variază în realitate considerabil de la un subiect la altul, variațiile pot depinde de un factor cu totul diferit. În acest caz media obținută poate să nu fie reprezentativă [5, 6].

Pentru a demonstra că media este reprezentativă, se calculează și se analizează devierea standard. Devierea standard (SD) reprezintă variația totală a unei caracteristici studiate, induse de cauze atât esențiale, cât și întâmplătoare, care permite evaluarea gradului de omogenitate al variabilelor din seria de variație.

$$SD = \pm \sqrt{\sigma^2} = \pm \sqrt{\frac{\sum d^2}{n}} = \pm \sqrt{\frac{\sum (Me - x_i)^2}{n}}$$

unde: Me – media eșantionului, x_i – variabilele seriei statistice analizate,

n – numărul de cazuri [8].

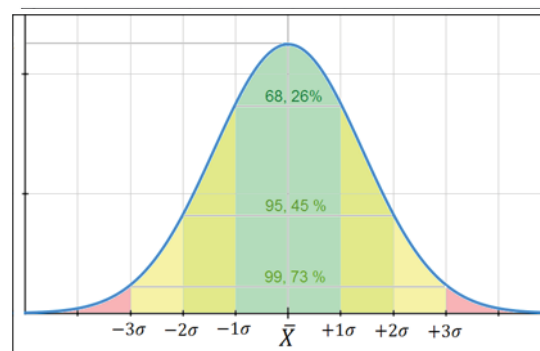


Figura 1. Graficul distribuției normale a variabilelor într-o serie de variație

Prin intermediul devierii standard, putem calcula eroarea standard a mediei (Es). Anume acest parametru statistic reprezintă acel indice pe lângă media statistică, care oferă validitate studiului efectuat, deoarece pe baza erorii standard a mediei putem calcula intervalul de încredere [5]. Eroarea standard arată cu cât poate greși cercetătorul calculând valoarea medie pe un eșantion, în comparație cu media totalității statistice generale, întrucât este puțin probabil ca media pe eșantion să fie egală cu media pe întreaga totalitate statistică. Această diferență nu poate fi calculată decât dacă s-ar putea calcula ambele medii, deaceia se setează o valoare-limită a acestei diferențe – eroarea maximă admisă. Deci, este mai rațional să se ia un interval care ar include media generală, decât să se ia o valoare exactă. Intervalul de încredere reprezintă acel interval în jurul mediei eșantionului, unde se presupune că s-ar afla media generală. Fiecărui interval de încredere îi este specific un anumit nivel de semnificație α , care constituie valorile considerate nereprezentative [9].

Intervalul de încredere este descris de formula:

$$[Me-\Delta; Me+\Delta]$$

unde: Me – media eșantionului,

Δ – eroarea limită admisă.

Eroarea limită admisă se calculează după următoarea formulă:

$$\Delta = Es \times t,$$

unde: Es – eroarea standard a mediei;

t – testul de semnificație [7].

Eroarea standard a mediei (Es) se calculează pe baza indicatorilor obținuți în urma analizei statistice al eșantionului, adică depinde de seria de variație obținută în cadrul eșantionului și deci, este o caracteristică a acestuia.

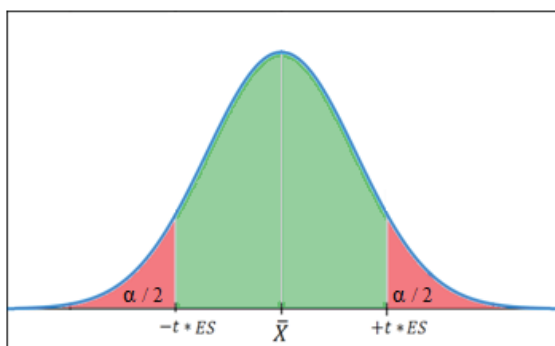


Figura 2. Graficul dependenței testului de semnificație t de nivelul de semnificație α

Testul de semnificație (t) depinde de valoarea nivelului de semnificație α , iar legătura dintre acestea este exprimată de curba distribuției normale Gauss-Laplace. În general, fiecărui nivel de semnificație α îi corespunde o anumită valoare a coeficientului testului de semnificație t . În studiile științifice, valoarea maximă acceptată a lui α (adică numărul relativ al valorilor nereprezentative) este de 5%, ceea ce corespunde unui coeficient de $t = 1,96$, pentru un eșantion cu un număr mai mare decât 120 de unități statistice, ceea ce ilustrează figura de (fig. 2) [5, 6, 7, 9].

Eroarea maximă admisă (Δ), după cum îi sugerează numele, este o valoare arbitrară setată de către cercetător. În studiile acceptate de comunitatea științifică, Δ trebuie să nu depășească 5%.

Deci, un eșantion reprezentativ ar fi acela care, la un nivel de semnificație $\alpha=0,05$ (ceea ce ar corespunde unui nivel al testului de semnificație $t = 1,96$), va admite o eroare maximă admisă de $\Delta = 0,05$. Dacă le-am atribui valori constante acestor 2 indicatori, care ar asigura reprezentativitatea, atunci am putea afla valoarea limită reprezentativă a erorii standard.

Deducerea formulei Cochran

După cum am menționat anterior, $\Delta = t \times Es$. Calcularea erorii standard se face conform următoarelor formule:

Pentru variabile cantitative:

$$Es = \pm \sqrt{\frac{\sigma^2}{n}},$$

unde: σ – devierea standard,

n – volumul eșantionului.

Pentru variabile calitative

$$Es = \pm \sqrt{\frac{p(1-p)}{n}},$$

unde: p – incidența fenomenului;

n – volumul eșantionului.

Substituind în formula de mai sus obținem:

$$\Delta = t \times Es = t \times \pm \sqrt{\frac{\sigma^2}{n}},$$

$$\Delta = t \times Es = t \times \left(\pm \sqrt{\frac{p(1-p)}{n}} \right).$$

Să presupunem că este cunoscută distribuția valorilor în populația statistică în funcție de factorul studiat, adică sunt cunoscute valorile σ (pentru variabilele cantitative) și a lui p (pentru variabilele calitative). Așa cum am atribuit valori constante celorlalți indicatori pentru a asigura reprezentativitatea, pe baza formulelor obținute mai sus, la fel putem afla valoarea minimă a volumului eșantionului reprezentativ n . Pentru aceasta, efectuăm următoarele transformări:

Ridicând ambele părți la pătrat, obținem:

$$\Delta^2 = \left[t \times \left(\pm \sqrt{\frac{\sigma^2}{n}} \right) \right]^2 = t^2 \times \frac{\sigma^2}{n}$$

$$\Delta^2 = \left[t \times \left(\pm \sqrt{\frac{p(1-p)}{n}} \right) \right]^2 = t^2 \times \frac{p(1-p)}{n}$$

Extragem n :

$$n = \frac{t^2 \times \sigma^2}{\Delta^2}$$

$$n = \frac{t^2 \times p(1-p)}{\Delta^2}$$

Am obținut formula Cochran.

Se poate constata că valoarea n conform formulei Cochran, nu depinde de volumul totalității statistice (N), ci numai de particularitățile acestei (valoarea p – probabilitatea apariției fenomenului). Astfel, această formulă este utilă în practică atunci, când volumul totalității statistice nu este cunoscut, iar pentru calcularea volumului eșantionului reprezentativ este necesar să se cunoască doar incidența fenomenului (cum se va demonstra mai jos, acest criteriu poate fi de asemenea exclus), nivelului admis de semnificație α (cărui îi corespunde un coeficient al testului de semnificație t specific) și eroarea maxim admisă Δ [6, 7].

Deducerea formulei Mureșan

Formula recomandată de P. Muresan se bazează pe legea cifrelor mari (P. Cebîșev), conform căreia, pentru un număr suficient de cazuri examinate într-o totalitate parțială, rezultatul obținut se apropie cu o exactitate suficientă de rezultatele care s-ar fi observat la examinarea totalității statistice [12].

Eroarea standard în formula Cochran este caracteristică eșantionului și independentă de volumul totalității statistice din care acesta a fost extras. Pentru a obține o eroare standard care ia în considerare și volumul totalității statistice, se aplică factorul de corecție pentru populații finite (FPC – finite population correction) [2].

$$FPC = \sqrt{\frac{N-n}{N-1}}$$

unde: N – volumul totalității statistice, n – volumul eșantionului.

Astfel, formula erorii standard ia forma [14]:

$$Es = \pm \sqrt{\frac{p(1-p)}{n} \times \frac{N-n}{N-1}}$$

Prin transformări asemănătoare celor de mai sus, se calculează volumul eșantionului reprezentativ, considerându-se și volumul totalității statistice:

$$\Delta = t \times Es = t \times \pm \sqrt{\frac{p(1-p)}{n} \times \frac{N-n}{N-1}}$$

Ridicând la pătrat, obținem:

$$\Delta^2 = t^2 \times \frac{p(1-p)}{n} \times \frac{N-n}{N-1}$$

Prin următoarele transformări, ajungem la volumul eșantionului reprezentativ:

$$\begin{aligned} \Delta^2 n &= t^2 \times p(1-p) \times \frac{N-n}{N-1} \Leftrightarrow \Delta^2 \times n \times (N-1) = t^2 \times p(1-p) \times (N-n) \Leftrightarrow \\ &\Leftrightarrow \Delta^2 \times n \times (N-1) = t^2 \times p(1-p) \times N - t^2 \times p(1-p) \times n \Leftrightarrow \\ &\Leftrightarrow \Delta^2 \times n \times (N-1) + t^2 \times p(1-p) \times n = t^2 \times p(1-p) \times N \Leftrightarrow \\ &\Leftrightarrow n \times [\Delta^2 \times (N-1) + t^2 \times p(1-p)] = t^2 \times p(1-p) \times N \Leftrightarrow \\ &\Leftrightarrow n = \frac{t^2 \times p(1-p) \times N}{\Delta^2 \times (N-1) + t^2 \times p(1-p)} \end{aligned}$$

Având în vedere că N are de obicei valori suficient de mari, procedeul de scădere de la numitor poate fi neglijat, iar formula ia următoarea formă:

$$n = \frac{t^2 \times p(1-p) \times N}{\Delta^2 \times N + t^2 \times p(1-p)}$$

Astfel, am obținut formula Mureșan.

Putem observa că formula Taro Yamane reprezintă un caz particular al formulei recomandate de P. Mureșan, în care $t^2 \times p(1-p) = 1$, adică $t = 1,96$, $p = 0,5$ și $q = 0,5$, de unde $1,96^2 \times 0,5 \times 0,5 \cong 1$.

$$n = \frac{N \times 1}{1 + N \times \Delta^2} = \frac{N}{1 + N \times \Delta^2}$$

Aplicarea formulelor Taro-Yamane și Mureșan este posibilă doar în cazul cunoașterii volumului totalității statistice. Astfel, se poate obține un volum su-

ficient, dar mai restrâns al eșantionului reprezentativ, eliminându-se necesitatea studierii unor cazuri suplimentare, ceea ce permite să se economisească timp și resurse, lucruri atât de prețioase pentru clinicianul cercetător.

Concluzie:

• Formulele de bază folosite în biostatistică pentru calcularea volumului eșantionului reprezentativ pentru eșantionare aleatorie simplă rezultă una din alta și, pot fi folosite prioritar în dependență de particularitățile obiectului de studiu.

• Când cunoaștem probabilitatea apariției evenimentului în totalitatea statistică, dar nu cunoaștem volumul ei, este rezonabil să folosim formula lui Cochran.

• Când cunoaștem volumul totalității statistice și, acesta este mai mic de 100 000 unități, e rezonabil să folosim formula recomandată de P. Muresan. În cazul în care nu se cunoaște probabilitatea apariției evenimentului cercetat se folosește valoarea $p = 0.5$ și $q = 0.5$.

• Este important ca cercetătorul să cunoască în prealabil datele bibliografice ale literaturii de specialitate privind probabilitatea apariției evenimentului studiat prin calcularea unei valori mai mici a volumului eșantionului reprezentativ, pentru a optimiza cheltuielile de timp și resurse într-un studiu științific.

Bibliografie

1. Cochran W. G. *Sampling Techniques*, New York, John Wiley and Sons, 1963.
2. Isserlis L. *On the value of a mean as calculated from a sample*. Journal of the Royal Statistical Society (Blackwell Publishing) 81 (1), pp. 75–81.
3. Mureșan P. *Statistica medico-sanitară*. București, 1959.
4. Muresan P. *Manual de metode matematice în analiza stării de sănătate*. București, 1989.
5. Sullivan L. M. *Essentials of Biostatistics in Public Health*, 2011.
6. Tintiuc D., Grosu Iu. *Sănătate publică și management*. Chișinău, 2007.
7. Tintiuc D. *Biostatistica și metodologia cercetării științifice*. Chișinău, 2011.
8. Yamane T. *Statistics: An Introductory Analysis*. New York, Harper and Row, 1967.
9. Всемирная Организация Здравоохранения, „Обучение медицинской статистики”, 1989.

Webografie

10. <http://www.seap.usv.ro/~valentin/capitol%203.pdf>
11. http://www.marketingsondaje.ro/04determinare_volum.php
12. http://ro.wikipedia.org/wiki/Legea_numerelor_mari
13. <http://www.amstat.org/about/statisticiansinhistory/index.cfm?fuseaction=biosinfo&BioID=21>
14. http://courses.wcupa.edu/rbove/Berenson/10th%20ed%20CD-ROM%20topics/section7_3.pdf