

СЕРГІЙ ЛІТВИНОВ,

кандидат соціологічних наук, провідний спеціаліст ДР маркетингового агентства IRS

Застосування поняття відносної похибки оцінювання для розрахунку вибірки із генеральної сукупності з малою часткою ознаки

Abstract

The paper elucidates how to determine a sample size from the general population if the analyzed variable's value is small. The Ukrainian sociologist Mykola Churylov was the first who highlighted this problem in the national scientific literature. The author argues that it is necessary to calculate both absolute and relative error. He also points to the sample size from the general population (in case of small variable's value) that should be larger than while using traditional formulas. A new enhanced approach presented in the paper includes using within a representative random sample only a priori characteristics.

Важливу роль в емпіричній соціології відіграє вибірковий метод, ґрунтований на теоретико-ймовірнісному підході щодо оцінювання характеристик генеральної сукупності (ГС) за характеристиками вибіркової сукупності (ВС). Специфіка предметної сфери і соціальної ситуації “звичайного” соціологічного дослідження (обмеженість засобів, технічних і людських ресурсів, вимога оперативності, важкодоступність елементів емпіричного об’єкта дослідження) стимулюють пошук методу, який уможливить обґрунтовані висновки щодо всієї сукупності на підставі даних обстеження її частини. Саме це має на меті вибірковий метод. Він застосовується скрізь, де доводиться виходити із зазначених вище умов — у дослідженнях бюджетів домогосподарств, структури вільного часу громадян, у контролі якості продукції та послуг, у статистичних обстеженнях населення, у масових соціологічних опитуваннях.

Вибірковий метод охоплює всю проблематику, пов'язану з добором одиниць, визначенням характеристик вибірки, а також із формулюванням суджень стосовно кількісних характеристик сукупності, на підставі якої побудовано вибірку [1]. Кожна інтегральна статистична характеристика вибірки (точніше, її математичне очікування), наприклад частка ознаки, середнє значення певної величини, її дисперсія, є оцінкою відповідної характеристики ГС. Характеристики ГС зазвичай називають параметрами. Характеристики ВС, розраховані за тим самим правилом, суть оцінки параметрів ГС. Оскільки ВС, за визначенням, не тотожна ГС, вибіркові оцінки відрізняються від дійсних значень параметрів. Кількісну міру їхньої розбіжності називають похибкою вибірки. Залежно від джерела виникнення й властивостей похибки репрезентативності розподіляють на систематичні та випадкові. Систематичні похибки виникають через вади планування вибірки, не випадковість добору одиниць, невалідність процедури обстеження й інструментарію вимірювання первинних характеристик тощо. Випадкові похибки пов'язані зі стохастичністю відхилень вибіркових оцінок від генеральних параметрів унаслідок недетермінованості добору одних елементів ГС і виключення інших (під випадковим ми розуміємо недетерміноване, стохастичне, а під імовірнісним — особливий різновид причинної детермінації — імовірнісну детермінацію). В тих випадках, коли йдеться про помилки репрезентативності, ми матимемо на увазі саме граничні випадкові похибки, котрі, на відміну від систематичних, можуть бути статистично оцінені до здійснення процедури добору одиниць ВС. Якщо обчислена випадкова похибка репрезентативності не перевищує прийнятних для дослідника меж, тоді вибірку вважають репрезентативною. У протилежному разі вона має вважатися нерепрезентативною.

Похибка репрезентативності завжди має конкретний зміст і характеризує не якість вибірки загалом, а точність вибіркової оцінки певного параметра генеральної сукупності. Вибірка, достатньою мірою репрезентативна стосовно одного параметра, наприклад частки деякої ознаки, може бути недостатньо репрезентативною стосовно частки іншої ознаки [2]. На практиці досяжна репрезентативність за двома-трьома, у ліпшому разі — ще за кількома ознаками одночасно. Парадокс вибіркового методу полягає в тому, що репрезентативність ВС за досліджуваною ознакою потребує вичерпного знання про її розподіл у генеральній сукупності, реконструкція якого саме й становить завдання вибіркового дослідження. У сучасній теорії вибіркового методу існує кілька можливостей виходу з цього замкненого кола.

По-перше, можна міркувати в такий спосіб. Нехай нам потрібно кількісно оцінити параметри невідомого генерального розподілу величини A , водночас нам відомий розподіл величини B , пов'язаної з A . Тоді можна здійснити вибірку, репрезентативну стосовно ознаки B , припустивши, що чим більш репрезентативною буде вибірка щодо B , тим вище буде її репрезентативність щодо A . Зв'язок між змінними може бути обґрунтований теоретично або встановлений емпірично, скажімо, на основі наявності кореляції (коваріації) ознак. Розглянемо, наприклад, такі ознаки працівників ве-

ликого підприємства, як “стаж роботи”, “досвідченість” (рівень засвоєних фахових умінь і навичок) та “вік”. Найочевиднішим є зв’язок між першою і другою ознаками — працівники з більшим стажем роботи є досвідченішими. Менш очевидною є залежність між ознаками “вік” і “стаж роботи”. Як правило, більшість працівників старшого віку мають більший стаж роботи, оскільки для них характерна схожа періодичність життєвого шляху: в нашому суспільстві приблизно в одному віці розпочинають і завершують навчання, в одному віці розпочинається також трудова діяльність. Натомість стереотипна очевидність зв’язку між віком працівника та його досвідченістю оманна: на підприємстві за певних обставин, скажімо, за умов перепрофілювання виробництва, можуть переважати старші працівники з меншим стажем роботи в даній сфері, а відтак менш досвідчені, ніж молодші, які опанували новий технологічний процес. У такому разі наявність зв’язку можна встановити лише емпірично, виходячи із попередніх соціологічних досліджень чи статистичних обстежень.

Надто часто доводиться стикатися із ситуацією, відмінною від описаної вище, коли будь-яка надійна основа вибірки відсутня. Якщо це так, тоді можна: 1) виходячи з теоретичної моделі об’єкта дослідження, зробити припущення стосовно виду генерального розподілу, але в соціології такий підхід радше є винятком, ніж правилом; 2) скористатися висновками центральної граничної теореми теорії імовірностей (ЦГТ). Цей підхід став канонічним при побудові вибірки [3].

Назагал для ідеального дотримання умов придатності ЦГТ необхідно здійснити серію подібних вибірок, але за відомих допущень обходяться однією. Для оцінювання параметрів ГС у цьому випадку застосовують точкові й інтервальні наближення. Звісно, внаслідок самої сутності вибіркового методу точкове оцінювання відрізнятиметься від справжнього значення параметра на деяку стохастичну величину. Оскільки дійсне значення невідоме (його треба оцінити), то можна припустити, що похибка оцінювання (відхилення) не перевищуватиме визначеної заздалегідь заданої нами величини. Відхилення є статистикою, тобто кожному значенню відхилення відповідає імовірність, з якою відхилення реалізуються у тривалій серії випробувань (при великих обсягах вибірки). Це означає, що відхилення $|\hat{X} - X|$ точкового вибіркового оцінювання \hat{X} від “істинного” значення X параметра не перевищує заданої похибки Δ з імовірністю P . Щоб здійснити інтервальне оцінювання X , необхідно і достатньо побудувати довірчий інтервал для X . Довірчим інтервалом називають інтервал значень параметра X , обчислений за вибірковою оцінкою \hat{X} , в яких перебуває його справжнє значення з довірчою ймовірністю P . Довірчий інтервал має вигляд

$$\hat{X} - \Delta \leq X \leq \hat{X} + \Delta. \quad (1)$$

В окремих випадках перед дослідником стоїть завдання порівняння точності вибірових оцінок. Кількісною мірою точності в статистиці заведено вважати відносну похибку вибірки [2], котра показує, на яку частину своєї величини точкова оцінка відрізняється від справжнього зна-

чення параметра в межах одиничного довірчого інтервалу (при $t = 1$ і $P(|\hat{X} - X| < \sigma) \approx 0,683$):

$$E_{\mu} = \mu_{\hat{X}} / \hat{X}, \quad (2)$$

$\mu_{\hat{X}}$ — середнє квадратичне відхилення вибіркової оцінки генерального параметра.

Точність оцінки слід розглядати як найважливіший критерій репрезентативності вибірки. Разом із тим соціологи необґрунтовано користуються поняттям абсолютної похибки як єдиної характеристики репрезентативності. Для вибірок із генеральної сукупності з малою часткою ознаки це призводить до колізії, на яку звернув увагу М. Чурилов [4]. Використання традиційних формул веде тут до хибних висновків про невеликий обсяг репрезентативної вибірки. Так, для ознаки, частка якої в квазінескінченній ГС становить 10%, обсяг репрезентативної вибірки за похибки у 5% (візьмемо довірчу ймовірність як таку, що дорівнює 0,95) дорівнює 138 одиницям, а отже, у вибірці виявиться ... 14 (± 7 з довірчою ймовірністю 95%) одиниць, наділених досліджуваною ознакою, чого явно недостатньо для аналізу. Цю обставину можна пояснити великою відносною похибкою вибірки, яка в наведеному прикладі наближається до 50%.

Тож як оптимізувати обчислення обсягу вибірки на підставі наявної апріорної інформації про частку біноміальної ознаки? М. Чурилов з цією метою замість традиційних формул пропонує оцінювати обсяг ВС на основі коефіцієнта варіації вибіркової оцінки (фактично, це стандартна відносна похибка вибірки) [4]:

$$n = \frac{N}{(NpE_p^2 / (1-p)) + 1} = \frac{1}{(pE_p^2 / (1-p)) + 1/N}.$$

Якщо p мала і $pE_p^2 \sim 1/N$ (обсяг вибірки значно менший за обсяг генеральної сукупності), $n \approx 1/pE_p^2$. (3)

E_p — стандартна відносна похибка оцінки генеральної частки за вибірковою ($E_p = E_{\mu}$ для параметра “частка ознаки” p)¹.

Але частка ознаки в генеральній сукупності нам, як правило, заздалегідь не відома. Ще одна вада формули (3) полягає в тому, що до неї входять як генеральний параметр (апріорна інформація), так і величина, похідна від вибіркової статистики (апостеріорна інформація).

Щоб подолати цю ваду, треба перейти від використання абсолютних похибок до відносних величин як показників репрезентативності вибіркової сукупності. Один із можливих підходів такий.

Неоднозначність використання граничної абсолютної похибки при плануванні вибіркової сукупності полягає в тому, що за відсутності інформації про частку ознаки p її прирівнюють до 0,5, і стандартна відносна похибка оцінювання частки E_p виходить **мінімальною** (див. формулу (13) далі). Це призводить до заниження обсягу репрезентативної для частки p вибірки, особливо суттєвого у випадку малої p . Для будь-якого іншого значення

¹ Всі умовні позначення по статті розшифровані у Додатку.

частки біноміальної ознаки, такої, що $p < 0,5$, похибка E_p буде більшою за мінімальну. Тому для адекватної репрезентації частки досліджуваної ознаки або іншої, пов'язаної з досліджуваною ознакою, розподіл якої відомий, бажано використовувати приблизну апіорну інформацію про частку, а вибірку обчислювати, виходячи з оцінки відносної похибки.

Виразимо відносну похибку вибірки через абсолютну похибку Δ і в результаті матимемо оцінку відносної похибки оцінювання:

$$V_{\Delta} = \frac{\Delta}{\hat{X}} = \frac{t\mu_{\hat{X}}}{\hat{X}} = tV_{\mu}. \quad (4)$$

$$\text{Але } n \approx \frac{\hat{\sigma}^2 t^2}{\Delta^2}, \text{ тоді } n \approx \frac{\hat{\sigma}^2 t^2}{\Delta^2} \approx \frac{\hat{V}_X^2 \hat{X}^2 t^2}{E_{\Delta}^2 \hat{X}^2} = \frac{\hat{V}_X^2 t^2}{t^2 E_{\mu}^2} = \frac{\hat{V}_X^2}{E_{\mu}^2},$$

$$n \approx (\hat{V}_X / E_{\mu})^2, \quad (5)$$

де $E_{\Delta} = \Delta / \hat{X} = t\mu_{\hat{X}} / \hat{X} = tE_{\mu}$.

Співвідношення (5) фіксує той факт, що обсяг вибірки прямо пропорційний квадрату величини, яка показує, у скільки разів відносна похибка оцінювання менша за коефіцієнт варіації відповідного параметра (або величини, рівної відношенню стандартного відхилення параметра в ГС до абсолютної похибки вибіркової оцінки параметра). Щоб точність підвищилася в r разів (в r разів зменшилася *відносна* похибка оцінювання), потрібно збільшити обсяг ВС у r^2 раз.

Знайдемо апіорний вираз для стандартної відносної похибки. Із того, що вибірковий коефіцієнт варіації дорівнює $\hat{V}_X = \hat{\sigma}_X / \hat{X}$,

$$a\mu_X = \sqrt{\frac{N-n}{N(n-1)}} \hat{\sigma}_X^2 \approx \sqrt{\frac{1}{n-1} - \frac{1}{N}} \hat{\sigma}_X \text{ або } \mu_X \approx \frac{\hat{\sigma}_X}{\sqrt{n}}, \text{ можна отримати, що}$$

стандартна відносна похибка оцінки дорівнює з (2):

$$E_{\mu} = \frac{\hat{V}_X \hat{X} / \sqrt{n}}{\hat{X}} = \frac{\hat{V}_X}{\sqrt{n}}; \quad (6)$$

з поправкою на скінченність ГС і за умови, що ми плануємо вибірку більше кількох десятків одиниць:

$$E_{\mu} = \sqrt{\frac{N-n}{N(n-1)}} \hat{V}_X \hat{X}^2 \cdot \frac{1}{\hat{X}} = \frac{\hat{V}_X}{\sqrt{n-1}} \sqrt{1 - \frac{n}{N}} \approx \sqrt{\frac{1}{n} - \frac{1}{N}} \cdot \hat{V}_X. \quad (7)$$

Тоді із (7) отримуємо уточнення формули (5) із поправкою на скінченність ГС:

$$n = \frac{\hat{V}_X^2 + E_{\mu}^2}{E_{\mu}^2 + \hat{V}_X^2 / N}. \quad (8)$$

За заданою величиною відносної похибки можна обчислити d^{\wedge}_X , а на її підставі — довірчі інтервали, довірчі ймовірності та обсяги репрезентативних щодо частки ознаки вибірок.

$$\hat{V}_X = \hat{\sigma}_X / \hat{X} = E_\mu \sqrt{n-1}; \quad \hat{\sigma}_X = \hat{X} \hat{V}_X = E_\mu \hat{X} \sqrt{n-1}, \quad (9)$$

або

$$\hat{\sigma}_X = E_\mu \hat{X} \sqrt{n-1} / \sqrt{1 - \frac{n}{N}} = E_\mu \hat{X} \sqrt{\frac{N(n-1)}{N-n}}. \quad (10)$$

$$\text{У разі частки ознаки } E_\mu = \frac{\mu_p}{\hat{p}} = \frac{\sqrt{\frac{p(1-p)}{n-1}} \sqrt{1 - \frac{n}{N}}}{\hat{p}}. \quad (11)$$

Для великих n , виходячи з теореми Бернуллі, можна вважати, що у формулі (11) $\hat{p} / p \approx 1$, а

$$E_\mu = \sqrt{\frac{p(1-p)}{n-1}} \sqrt{1 - \frac{n}{N}}. \quad (12)$$

Стандартна відносна похибка оцінки частки біноміальної ознаки приблизно дорівнює

$$E_p \approx V_p = \sqrt{\frac{1-p}{np}}. \quad (13)$$

Для вибіркової оцінки частки ознаки із (12), з огляду на те, що $E_\Delta = t E_p$, отримуємо:

$$n \approx \frac{1-p}{p E_p^2} = t^2 \frac{1-p}{p E_\Delta^2}. \quad (14)$$

Для малих відомих p відносно похибку оцінювання слід брати мінімальною за абсолютної похибки, рівної Δ , інакше апостеріорна відносна похибка оцінювання з високою ймовірністю буде більшою, ніж прийнята нами апріорна відносна похибка E_Δ у формулі (14):

$$E_\Delta (\text{min}) = \frac{\Delta}{\hat{p}(\text{max})} = \frac{\Delta}{p+\Delta}. \quad (15)$$

Перетворимо (14) відповідно до (15) і отримуємо:

$$n_1 = t^2 \left(\frac{p}{\Delta} + 1 \right)^2 \left(\frac{1}{p} - 1 \right) = t^2 \left(\frac{1}{\varepsilon} + 1 \right)^2 \left(\frac{1}{p} - 1 \right), \quad (16)$$

$\varepsilon = \Delta / p$ — критерій точності вибірки, який за змістом аналогічний відносній похибці, але який містить не вибірку оцінку частки, а її дійсну величину в ГС.

Формула (16) видається більш задовільною, ніж (3), адже до неї входить лише апріорна інформація про генеральну сукупність, а не гіпотези щодо вибіркової статистики, що було б нелогічним.

Табулюємо n_1 для різних p і Δ , враховуючи, що $\Delta \leq p$ і $t = 2$ ($P = 0,95$):

Таблиця 1

Обсяг n , простої імовірнісної вибірки, придатної для репрезентації частки ознаки p у ГС за вибірковою оцінкою зі статистичною похибкою Δ^1

$p \setminus \Delta$	0,01	0,05
0,01	1584	–
0,05	2736	304
0,10	4356	324
0,15	5803	363
0,20	7056	400
0,25	8112	432
0,30	8969	457
0,35	9627	475
0,40	10086	486
0,45	10345	489
0,50	10404	484

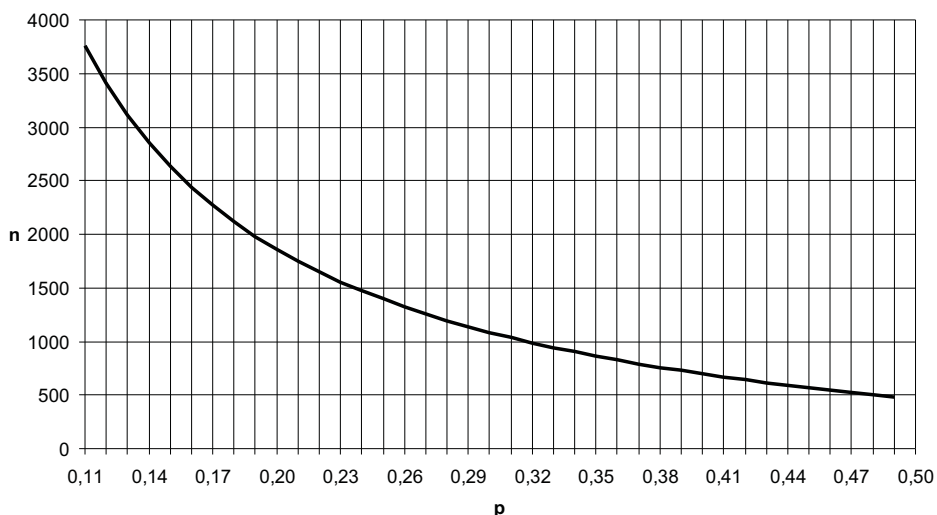


Рисунок. Залежність оптимального обсягу $n = t^2 (1/\varepsilon + 1)^2 (1/p - 1)$ простої імовірнісної вибірки, придатної для репрезентації частки ознаки p (інтервал від 0,11 до 0,5) у ГС за вибірковою оцінкою зі статистичною похибкою не більш як $\Delta = 0,1 p$ ($P = 0,95$)

Корисно також мати таблицю обсягу ВС $n_2 \approx (V_X / E_\mu)^2 = t^2 (V_X / E_\Delta)^2, t = 2$ ($P = 0,95$):

¹ Див. рисунок.

Обсяг n_2 простої імовірнісної вибірки, придатної для репрезентації параметра ГС із коефіцієнтом варіації V_x за вибірковою оцінкою з відносною статистичною похибкою E_Δ

$E_\Delta \backslash V_x$	0,01	0,05	0,1	0,2	0,3
0,05	100	–	–	–	–
0,1	400	–	–	–	–
0,2	1600	–	–	–	–
0,3	3600	144	–	–	–
0,4	6400	264	–	–	–
0,5	10000	400	100	–	–
0,6	–	584	144	–	–
0,7	–	784	196	–	–
0,8	–	1024	256	–	–
0,9	–	1296	324	–	–
1,0	–	1600	400	100	–

* *Примітка.* Позначку “–” використано щодо значень, які не відповідають умовам апроксимації генеральної частки нормальним розподілом вибірових оцінок.

Розглянемо відмінності між методами обчислення вибіркової сукупності, про які йшлося вище, на прикладі. Нехай необхідно здійснити масштабне електоральне дослідження, спрямоване, зокрема, на вивчення характеристик електорату політичних партій, які здолають тривідсотковий прохідний бар'єр на виборах до парламенту¹.

(а) Відповідно до традиційного підходу, число одиниць ВС (генеральну сукупність можна вважати квазінескінченною) дорівнює: $n \approx \sigma^2 t^2 / \Delta^2$, $\sigma^2 = p(1-p)$, $p = 0,03$ – частка ознаки в ГС. Візьмемо як критерій задовільної точності вибірки $\varepsilon = \Delta / p = 0,2$, $t = 1,96$.

$$\text{Тоді } n = \lim_{p \rightarrow 0,03} \frac{p(1-p) \cdot t^2}{\Delta^2} \approx \frac{t^2}{\varepsilon \Delta} = \frac{t^2}{\varepsilon^2 p}$$

$$n \approx 1,96^2 / (0,2^2 \cdot 0,03) \approx 3201$$

(б) Згідно із підходом М.Чурилова, $n = 1 / (p E_p^2)$.

$$E_p = E_\Delta / t \approx \varepsilon / t$$

$$\text{Звідки: } n = \frac{1}{p} \cdot \frac{1}{(\varepsilon/t)^2} = \frac{t^2}{\varepsilon^2 p}, \quad n \approx 1,96^2 / (0,2^2 \cdot 0,03) \approx 3201$$

У даному випадку підхід (б) є лише коректним формулюванням традиційного підходу (а).

(в) Застосуємо для обчислення ВС формулу (16), що ґрунтується на припущенні про максимальну можливу відносну похибку:

$$n = t^2 (1/\varepsilon + 1)^2 (1/p - 1), \quad n = 1,96^2 (5 + 1)^2 (1/0,03 - 1) \approx 4472.$$

¹ Приклад вигаданий, оскільки в реальності для вказаної мети випадковий добір респондентів є малоефективним.

Як бачимо, вибірка із генеральної сукупності з малою часткою ознаки має бути більшою за обсягом, ніж можна судити, виходячи зі звичних формул (див. рис.). Більший обсяг вибірки зумовлений тим, що застосування принципу мінімізації відносної помилки вможливорює уникнення заниження дисперсії біноміальної ознаки з малою часткою на етапі планування вибірки. За використання традиційного підходу апріорне заниження дисперсії дуже ймовірне, і тому вибіркові оцінки частки виявляються неточними, із невикористано широкими довірчими межами.

Зауважимо, що майже всі наведені способи оцінювання обсягу вибіркової сукупності застосовні лише: (1) для великих вибірок ($n > 100$); (2) для незсунутих вибірок або таких, для яких величиною зсуву B можна знехтувати, коли $B/\hat{\sigma} < 0,1$ [5]; (3) для генеральної сукупності із коефіцієнтом варіації досліджуваної ознаки $\sigma/X < \sqrt{n}/3$, де n – обсяг вибіркової сукупності. Якщо коефіцієнт варіації більший, то за великого обсягу ВС (ГС квазінескінченна і $\hat{V}_x \approx V_x$) нижня межа довірчого інтервалу параметра X з імовірністю $P(t = 3) \approx 0,997$ наближається до $(X_{\min} \approx \hat{X} - 3\hat{\sigma}/\sqrt{n}) < 0$, що не має сенсу, тож слід використовувати скориговані формули обчислення граничної похибки вибірки. Можна припустити, що в останньому випадку до розподілу ймовірностей похибки оцінювання ЦГТ (внаслідок недотримання умов застосовності теореми Ляпунова) незастосовна, тому його не можна вважати нормальним. Відповідно, втрачають сенс традиційні формули побудови інтервальних оцінок, а самі інтервали не будуть симетричними [1; 6]. Розподіл похибок випадкових вибірок із генеральної сукупності з малою часткою ознаки апроксимується не нормальним розподілом, а розподілом Пуасона чи подібним до нього скошеним розподілом.

Таким чином, вибірки з генеральних сукупностей із малою часткою ознаки треба обчислювати на інших статистичних засадах, ніж звичайні, або ж їх варто замінити одним із методів спрямованого добору, монографічним дослідженням, використовувати процедури бустінга.

ДОДАТОК

Абревіатури та основні умовні позначення

X – параметр розподілу статистичної величини в генеральній сукупності

\hat{X} – вибіркова оцінка параметра розподілу статистичної величини в ГС

σ – середньоквадратичне відхилення ознаки в ГС

t – квантиль нормального розподілу оцінок в інтервальному оцінюванні характеристик ГС

$\hat{\sigma}$ – середньоквадратичне відхилення ознаки у ВС

μ_x – середнє квадратичне відхилення вибіркової оцінки генерального параметра від “істинного” значення

V_X — генеральний коефіцієнт варіації ознаки

\hat{V}_X — коефіцієнт варіації ознаки у ВС

$E_\mu = \mu_X / \hat{X}$ — стандартна відносна похибка вибірки (вибіркової оцінки параметра \hat{X})

N — обсяг ГС

n — обсяг ВС

Література

1. *Шварц Г.* Выборочный метод. Руководство по применению статистических методов оценивания. — М., 1978.
2. Общая теория статистики / Под ред. А.Я.Боярского, Г.Л.Громько. — М., 1985.
3. *Гнеденко Б.В.* Курс теории вероятностей. — М., 1988.
4. Оперативные социологические исследования : Учебное пособие. — Минск, 1997.
5. *Кокрен У.* Методы выборочного исследования. — М., 1976.
6. *Орлов А.И.* Социология: методология, методы, математические модели. — М., 1992. — С. 28–50.