ЕДИН ПРАКТИЧЕСКИ ПОДХОД ЗА СТРАТИФИКАЦИЯТА НА СЪВКУПНОСТИ[1]

*Стефан Цонев**

В статията е описан начин за оптимално стратифициране на съвкупността чрез коефициент на вариация с цел излъчване на извадки при едновременно зададен обем и максимална грешка.

В сравнение с простия случаен подбор при еднакъв обем на извадката, стратифицирането на съвкупността дава възможност да се подобри точността на оценката или обратното - същата точност да се постигне с по-малка извадка.

Това е така, защото при стратифициране обемът на извадката не зависи от общото разсейване в съвкупността, а от разсейването в стратите. Съответно, колкото по-хомогенни са стратите, толкова по-малък е обемът на извадката. Но това изискване поражда проблем.

Разделянето на съвкупността на страти, когато се извършва по субективната преценка на изследователя, неизбежно повдига въпроса доколко добре са избрани границите на стратите и дали има и друг вариант, който да е по-подходящ с оглед на поставените цели.

Преди години бях изправен пред въпроса **какви граници на стратите да се използват** при извадковите изследвания на НСИ за вътрешната търговия. На разположение беше списък с около 40 000 търговски обекта с отчетените от тях обороти за предходната година.

При търсенето на подходящ размер на стратите пролича целият субективизъм и трудоемкост на подхода, изискващ проиграването на различни варианти по пътя на пробите и грешките.

Използването на стандартното отклонение не беше добър показател за стратифициране, тъй като разсейване от 10 000 хил. лв. можеше да е твърде голямо за групите магазини с малък оборот, но съвсем приемливо за групите на големите магазини.

---

* Началник на отдел „Статистика на околната среда и енергетиката”, НСИ; e-mail: stzonev@nsi.bg.

[1] Авторът споделя практическия си опит и няма никакви претенции относно оригиналността на описания подход. Поради това не са цитирани и литературни източници. Допълнителна причина е, че изчислителната техника става масово разпространена в относително по-нови времена.

Оказа се, че проблемът може да се избегне чрез използването на еднакъв коефициент на вариация (отношението на стандартното отклонение към средната аритметична) като критерий за формиране на стратите.

На практика задачата се решава, като списъкът от единиците в съвкупността се сортира възходящо по признака[2], по който ще се излъчва извадката. След това, започвайки от началото на списъка с добавянето на всяка нова единица, се изчислява коефициент на вариация. При достигане на предварително зададена стойност се формира страта. Следващата единица е първа за новата страта и процедурата се повтаря до формиране на нова страта или до изчерпване на съвкупността.

При наличните сега технологии подходът може да се онагледи чрез използването на Excel или друга подобна програма. За целта е необходим файл, съдържащ колона с единиците от изучаваната съвкупност, сортирани възходящо по признака, който ще се използва за стратификацията.

Създава се съседна колона, където на всеки ред се изчислява коефициент на вариация на базата на стойностите от предходните редове (от началото на таблицата). Когато коефициентът достигне предварително определения размер, се създава страта. Следващият ред става първи за новата страта и изчисленията се повтарят. Отново, след достигане на определения размер на коефициента на вариация, се формира нова страта и така до края на записите.

Най-вероятно съвкупността ще се изчерпи преди да е достигната определената граница на коефициента на вариация - т.е. последната страта, тази с най-големите единици, ще бъде с коефициент по-малък от указания.

При търсенето на отговор на въпроса какъв размер на коефициента на вариация да се използва за стратифициране проличава възможността да се работи едновременно със зададен обем на извадката и размер на максималната грешка.

Това се вижда, ако от следната формула за обема на извадката при **прост случаен подбор:**

$$n = \frac{t^2.(V\%)^2}{(\Delta\%)^2 + \frac{t^2.(V\%)^2}{N}} \qquad (1)$$

се изведе коефициентът на вариация от едната страна на равенството. Така се получава приблизителна оценка за търсения коефициент на вариация, съответстващ на зададените обем на извадката и максимална грешка:

$$V\% = \sqrt{\frac{(\Delta\%)^2}{t^2.(\frac{1}{n} - \frac{1}{N})}} \qquad (2)$$

където:

V% - търсеният коефициент на вариацията в % (*V%* = 100 * стандартното отклонение/средната аритметична);

Δ% - зададен размер на максималната грешка на оценката в %;

n - зададен брой на единиците в извадката;

N - брой на единиците в изучавана съвкупност;

[2] Относно избора по кой от изучаваните признаци да се излъчи извадката има достатъчно литература, поради което той не е предмет на настоящата статия.

t - множител, зависещ от зададеното равнище на значимост -1.96 за 0.05 (5%) и 2.54 за 0.99 (1%) вероятност действителната стойност на оценяваната средна аритметична да бъде извън интервала +/- максималната грешка.

Трябва да се има предвид, че горната формула, използваща само информация за броя на единиците в съвкупността и размера на извадката, дава ориентировъчна стойност за *V%,* тъй като:

- изчисленият *V%* показва каква вариация би трябвало да има една хипотетична съвкупност, за да се получат зададеният обем на извадката и максималната грешка, и то **при прост случаен подбор[3];**

- в процеса на изчисленията действителната стойност, при която се формира стратата, е приближение на зададения *V%* (само по изключение може да е равна);

- последната страта като правило е с по-малък *V%* от зададения, тъй като съвкупността се изчерпва преди достигане на търсената стойност.

Практиката показа, че като правило, при прилагане на оптимален подбор (Jerzy Neyman) полученият по формулата коефициент на вариация води до **по-малък обем** на извадката от този, заложен в изчисленията по формула (2).

Причината е, че **формула (2) не отчита ефекта от стратификацията,** въпреки че го подобрява, особено при оптималния подбор, при който обемът на подизвадките е пропорционален на разсейването в стратите (за разлика от пропорционалния подбор, където обемът на подизвадките е пропорционален на броя на единиците от генералната съвкупност, които са попаднали в съответната страта).

В този случай се налага да се зададе нарастване на коефициента на вариация с някаква стъпка (5 - 10%) и да се повторят изчисленията. Итерациите се повтарят до получаване на търсения обем на извадката. Може да се разработи софтуер, който да автоматизира процеса.

Въобще възможността за получаване на малки извадки при запазена точност може да изглежда привлекателна, но е добре да се има предвид следното:

- обикновено намаляването на обема на извадката (при равни други условия) е съпроводено с нарастване на броя на стратите;

- относително малкият обем на извадката и (или) големият брой на интервалите могат да доведат до увеличаване на очакваната грешка при други изучавани признаци в рамките на същата извадка;

- изгодата от малкия обем на извадката може да се окаже фиктивна поради допълнителните организационни изисквания при провеждане на наблюдението, дължащи се на многобройните групи;

- най-често данните, послужили за изчисленията, се отнасят за период, който е отдалечен от момента на изследването и ако са настъпили значими изменения от случаен характер при малък брой на единиците, то те трудно могат да се компенсират.

Пример: Проведеното през 2007 г. от НСИ пилотно изследване за използването на препарати за растителна защита при производството на пшеница може да илюстрира описания вече метод.

За целта беше използван списък от 14 241 земеделски производители. Зададеният обем на извадката беше общо 500 стопанства, от които 105 с над 1 000 хектара се предвиждаше да се изучат изчерпателно, а останалите 14 136 - с извадка от 395 единици.

---

[3] Ако допуснем, че е възможно да има две съвкупности с приблизително еднакъв брой единици, средна аритметична и стандартно отклонение, това не означава, че броят и големината на стратите ще са еднакви. Причината е, че **границата на стратите зависи от размера на отделната единица.** Съответно използваните във формулата параметри не са достатъчни за определяне на такъв коефициент на вариация, при който да се получи точно търсеният обем на извадката чрез стратифициран подбор.

1. Общи характеристики на съвкупността от стопанства, производители на пшеница[4]

Общ брой на стопанствата	14136
Среден размер на стопанството - ха	56.990
Стандартно отклонение - ха	135.552
Коефициент на вариация - %	237.851

При горните характеристики обемът на проста случайна извадка при зададени максимална грешка 5% и равнище на значимост 0.05 ще бъде 5 383 стопанства.

След заместване във формула (2) със съответните стойности за размер на съвкупността от 14 136 единици, извадка от 395 единици, максимална грешка 5% и множител 1.96 получаваме търсения *V%* = 51.4, или около четири пъти по-малък от фактическия (237.851%).

Както беше посочено, изчисленият *V%* е само ориентировъчен и полученият след изчисленията за оптимален подбор обем на извадката се оказа по-малък от търсения - 233 срещу 395.

Търсеният обем на извадката се получи след 6 итерации, изискващи нищожно изчислително време. Резултатите са показани в табл. 2.

2. Последователност на изчисленията до получаване на необходимия обем на извадката

№ на итерацията	Зададен коефициент на вариация %	Брой страти	Обем на извадката
1	51.4	6	233
2	61.4	4	315
3	66.4	4	339
4	68.0	4	385
5	69.0	4	409
6	68.5	4	394

В резултат на изчисленията с *V%* = 68.5 се получи следното разпределение на съвкупността по страти, което е показано в табл. 3.

[4] Всички последващи изчисления са направени с написана от автора на статията компютърна програма.

3. Данни за генералната съвкупност и извадката, получена при стратификация със зададен коефициент на вариация 68.5%

№	Групи до (ха)	Данни за съвкупността преди провеждане на наблюдението					
		брой стопанства в генералната съвкупност	земя в групата от генералната съвкупност (ха)	среден размер земя на ферма от генералната съвкупност (ха)	стандартно отклонение в стратата от генералната съвкупност	коефициент на вариация (%) в групата от генералната съвкупност	брой стопанства в извадката
1	4.717	7087	12083	1.70	1.17	68.50	8
2	45.000	4026	61811	15.35	10.56	68.80	39
3	527.730	2679	492775	183.94	125.94	68.47	309
4	998.640	344	238941	694.60	122.18	17.59	38
	Общо за извадката	**14136**	**805611**	**56.99**	**135.55**	**237.85**	**394**
5	Наблюдавани изчерпателно >1 000 ха	105	216216	2059.20	4548.87	220.90	105
	Общо	**14241**	**1021827**	**71.75**	**445.76**	**621.25**	**499**

След приключване на изследването и обработката на данните, се получиха резултати, показващи разликите между действителните характеристики на съвкупността и тези, използвани за определяне на обема на извадката (табл. 4).

4. Характеристики на изучаваната съвкупност след провеждане на изследването

№	Групи до (ха)	Данни за съвкупността, получени от извадката, след провеждане на наблюдението					
		брой стопанства в генералната съвкупност	брой анкетирани в извадката	общо земя в групата от извадката (ха)	среден размер на земята (ха)	стандартно отклонение	коефициент на вариация (%) в групата от извадката
1	4.717	7087	10	15.29	1.53	0.82	53.82
2	45.000	4026	37	655.02	17.70	11.40	64.40
3	527.730	2679	307	59171.78	192.74	129.73	67.31
4	998.640	344	37	26280.90	710.29	129.00	18.16
	Общо за извадкта	**14136**	**391**	**86123**	**59.62**		
5	Наблюдавани изчерпателно >1 000 ха	94	94	132347	1407.95	506.56	35.98
	Общо	**14230**	**485**	**218470**	**68.53**		

По причини, които не са предмет на коментар в статията, характеристиките на съвкупността след провеждане на наблюдението се оказаха различни от тези, използвани във фазата на планиране.

Независимо от това, както се вижда от приложените таблици, общите резултати са в рамките на планираните грешки по отношение на земята и силно корелирания с нея добив на пшеница - под 5% (табл. 5).

5. Основни резултати от проведеното изследване

(Извадка+изчерпателно)

	Оценка	Максимална грешка - %
Среден размер на земята - ха	68.53	**4.30**
Среден добив на пшеница - кг/ха	2106.28	**4.45**
Употребени фунгициди средно - кг/ха	7.46	17.79
Употребени хербициди средно - кг/ха	8.94	26.45
Употребени инсектициди средно - кг/ха	1.87	48.96

Грешките на оценките на употребените препарати за растителна защита са значително по-големи, тъй като техните количества (измерени в активно вещество на хектар) варират много повече от добивите на пшеница на хектар. За сравнение, съгласно данни на Министерството на земеделието средният добив на пшеница за същия период (2007 г.) е 2 197 кг/ха.

При грешките на оценките за площите са използвани формулите за **оптимален** подбор, докато за останалите показатели, по които не е извършена стратификация, са използвани формулите за **пропорционален** подбор и грешка на частно.